ManagementSolutions
Making things happen

Data science
e a transformação do setor financeiro

Instituições Financeiras

www.managementsolutions.com

Design e diagramação
Departamento de Marketing e Comunicação
Management Solutions - Espanha

Fotografías
Arquivo fotográfico da Management Solutions
iStock

Índice

Introdução

Sem dados, você é apenas mais uma pessoa com opinião.
W. Edwards Deming[1]

O mundo está se transformando em alta velocidade. Estamos sendo testemunhas de uma revolução tecnológica de dimensões nunca antes vistas.

Não se trata de um fato conjuntural. O índice de mudança de paradigma (a velocidade de adoção de novas idéias) está duplicando a cada década: enquanto demorou quase meio século para adotarmos o telefone e várias décadas para aceitarmos a televisão e o rádio, a instauração do computador, da Internet e dos telefones celulares ocorreu em menos de 10 anos[2]. Em 2014, o número de telefones celulares já se equiparava ao número de pessoas no mundo, 7 bilhões, um terço dos quais smartphones; e o número de usuários de Internet alcançou quase os 3 bilhões[3].

As tecnologias da informação duplicam a cada ano sua capacidade e a relação qualidade/preço, como previu a Lei de Moore[4], que tem sido cumprida até hoje (Fig. 1). A consequência disso é um crescimento exponencial na disponibilidade da tecnologia e uma redução equivalente no seu custo, indiferente às crises vividas nos últimos anos, que previsivelmente continuará a evoluir nas próximas décadas.

Mas esta revolução tecnológica está adquirindo uma nova dimensão nos últimos anos: ao aumentar as funcionalidades técnicas, também está aumentando a capacidade de criar, armazenar e processar informação, com uma taxa também exponencial, o que resultou na designação do fenômeno «big data». Algumas evidências a esse respeito:

- O volume total de dados no mundo duplica a cada 18 meses[5,6].
- Mais de 90% dos dados que existem hoje foram gerados nos dois últimos anos[6].
- A capacidade per capita de armazenar informação duplicou a cada 40 meses desde 1980, e o custo foi reduzido em mais de 90%[6].
- A capacidade de processamento foi multiplicada por 300 desde o ano 2000, permitindo o processamento de milhões de transações por minuto[6].

O impacto desta transformação tecnológica está sendo especialmente relevante no setor financeiro, que soma-se a outras quatro grandes tendências que estão marcando sua evolução:

1. Uma conjuntura macroeconômica caracterizada por um crescimento fraco, taxas de inflação baixas e taxas de juros reduzidas, que penalizou as margens de lucro do setor bancário nas economias maduras durante um prolongado período de tempo; e um comportamento díspar nos países emergentes, com uma tendência para a desaceleração do crescimento e o aumento da inadimplência.

2. Um ambiente normativo mais exigente e intrusivo, onde a regulação adquire um caráter global em termos de governança corporativa, solvência, liquidez, limitação do resgate financeiro de instituições, proteção do consumidor, prevenção da fraude e requisitos de informação e reporte, entre outros.

3. Uma mudança profunda no comportamento do cliente, que agora tem maior cultura financeira, espera e exige excelência nos serviços, enquanto manifesta uma crescente confusão perante a complexidade e disparidade da oferta, o que o torna mais dependente dos líderes de opinião.

4. A entrada de novos concorrentes no mercado financeiro, alguns deles com novos modelos de negócio que afetam o status quo.

[1] William Edwards Deming (1900-1993). Estatístico norte-americano, professor universitário, autor, consultor e difusor do conceito de qualidade total, notável pelo seu trabalho no desenvolvimento e crescimento do Japão após a Segunda Guerra Mundial.
[2] Kurzweil [Diretor de Engenharia da Google] (2014).
[3] International Telecommunication Union (2014).
[4] Observação de Gordon Moore, cofundador da Intel, em 1965, de que a tecnologia evolui de modo que o número de transístores em um circuito integrado duplica a cada dois anos aproximadamente. Moore (1965).
[5] Estima-se que, em 2012, todos os dias, foram produzidos 2,5 exabytes de dados, um volume de informação que equivale a 12 vezes todos os livros impressos que existem no mundo.
[6] Federal Big Data Commission (2014).

O efeito combinado destes quatro fatores, juntamente com a transformação tecnológica, está fazendo com que, entre outras questões, nos enfoquemos no uso eficiente da informação, possibilitando, assim, a entrada de uma disciplina no setor financeiro até agora mais enfocada no setor tecnológico: data science.

Data science, ou ciência dos dados, é o estudo da extração generalizável de conhecimento a partir dos dados por meio do uso combinado de técnicas de aprendizagem automática, inteligência artificial, matemática, estatística, bancos de dados e otimização, em conjunto com uma compreensão profunda do contexto de negócios[7].

Todas estas questões já eram empregadas no âmbito financeiro em graus distintos, mas esta disciplina tem características que a tornam indispensável para enfrentar a transformação do setor que já está ocorrendo.

Especificamente, todos os elementos do complexo contexto enfrentado pelo setor financeiro mencionado anteriormente requerem dados abundantes e técnicas analíticas complexas para que sejam enfrentados, que é exatamente o campo de especialidade da data science. Além disso, data science é uma disciplina potencializada como consequência do fenômeno big data e, portanto, os data scientists são profissionais qualificados para lidar com quantidades massivas de dados desestruturados (como por exemplo, os provenientes de redes sociais), cada vez mais relevantes para as instituições.

Por outro lado, esta explosão na geração, no acesso, no processamento e no armazenamento dos dados, e na tomada de decisões baseadas nos mesmos, somada a outros fatores conjunturais descritos, não passou despercebida por parte dos reguladores. De fato, há uma tendência global sustentada, entre outros, pelo Comitê de Supervisão Bancária de Basileia (por meio da norma BCBS 239), para a exigência de um framework robusto de governança de dados, que garanta sua qualidade, integridade, rastreabilidade, consistência e replicabilidade para a tomada de decisões, especialmente (mas não apenas) no âmbito de Riscos.

Esta tendência é complementada com a promovida pela Federal Reserve e a OCC dos EUA[8], que exige das instituições um framework robusto de governança dos modelos, para controlar e mitigar o risco derivado de sua utilização, conhecido como «risco de modelo»[9].

As instituições financeiras estão avançando de forma decisiva no desenvolvimento destes frameworks de governança (dados e modelos) que, conjuntamente, formam a governança das capacidades da data science.

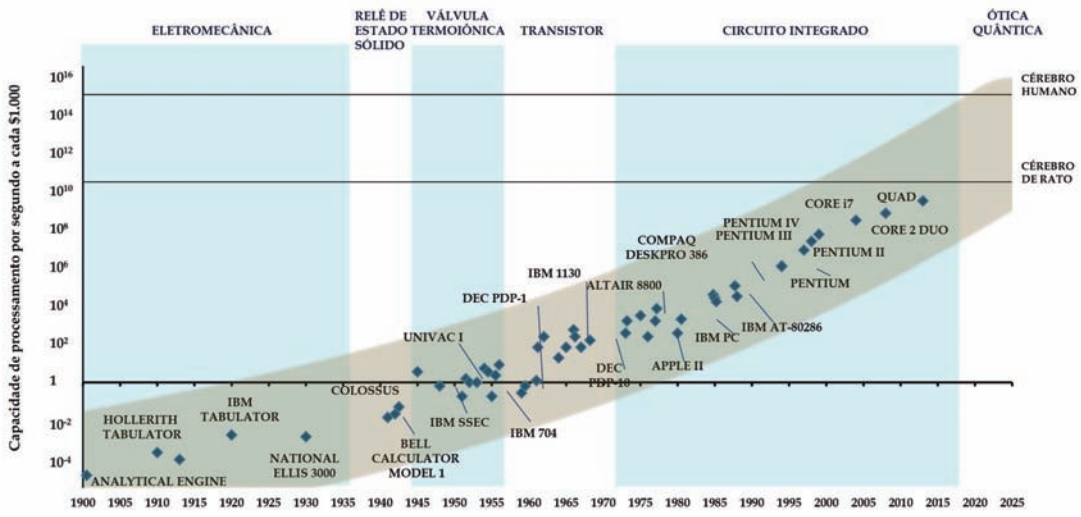
Diante deste ambiente em constante mudança, a transformação das instituições financeiras não é uma possibilidade; é uma necessidade para assegurar a sobrevivência. Uma transformação intimamente ligada à inteligência, que, em suma, é a capacidade de receber, processar e armazenar informação para resolver problemas.

Neste contexto, o presente estudo pretende descrever, de forma prática, o papel desempenhado pela disciplina de data science, mais especificamente, no setor financeiro. Para tal, o documento está estruturado em três seções, que respondem a três objetivos:

- Descrever a revolução tecnológica em que o setor financeiro está imerso e suas consequências.
- Introduzir a disciplina de data science, descrever as características do data scientist e analisar as tendências observadas a esse respeito, assim como seu impacto nos frameworks de governança dos dados e dos modelos nas instituições financeiras.
- Expor um estudo de caso para ilustrar a aplicação da data science no setor financeiro, que consiste no desenvolvimento de um modelo de scoring de crédito para pessoas físicas utilizando dados extraídos de redes sociais.

[7] Dhar [Center for Data Science, New York University] (2013).
[8] OCC/Fed (2011).
[9] Esta tendência foi analisada em profundidade pela Management Solutions (2014).

Fig. 1. Lei de Moore: crescimento exponencial da capacidade de processamento por segundo e 1.000 dólares.

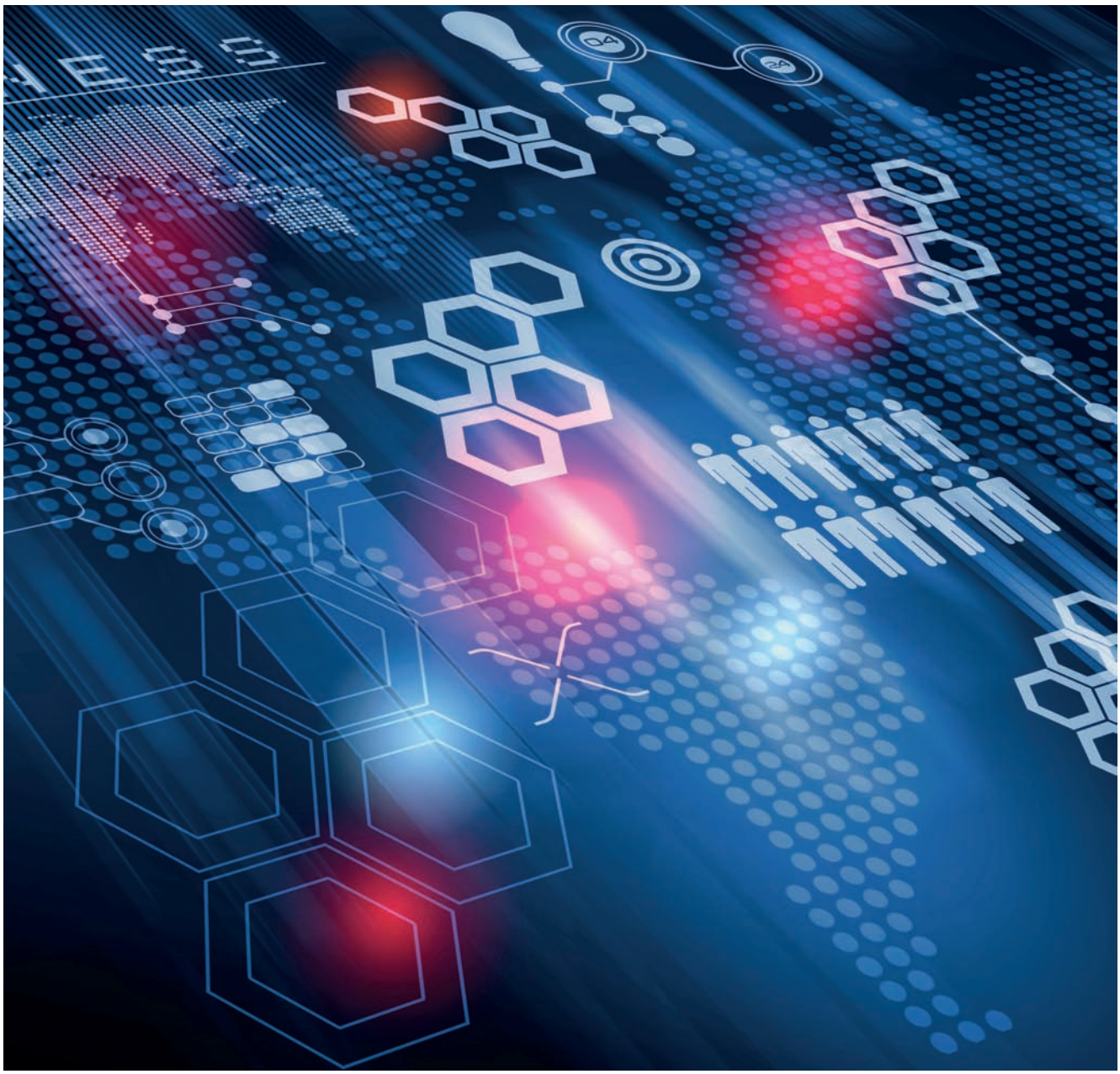


Fonte: Kurzweil (2014).

Resumo executivo

Se você não consegue explicar de forma simples,
é porque não compreende suficientemente bem.

Albert Einstein[10]



Um setor financeiro em transformação

1. As instituições financeiras enfrentam uma revolução tecnológica sem precedentes, que está transformando o mundo em relação ao que se pode fazer e ao custo a que se pode fazer, e que, consequentemente, impacta sua atividade de forma substancial. Esta revolução manifesta-se tanto na geração e acesso à informação como no seu armazenamento, processamento e modelagem.

    - A velocidade em que a informação é gerada está aumentando de forma vertiginosa: estima-se que o volume total de dados no mundo duplica-se a cada 18 meses[11]. Há anos, esses dados são na maioria digitais, sendo que 50% deles já podem ser acessados pela Internet, e 80% são desestruturados (vídeos, imagens, e-mails, etc.)[11]. Além disso, grande parte desses dados provém de fontes novas: redes sociais, logs de atividade, etc. Consequentemente, o fenômeno big data está se caracterizando por uma explosão dos «três V's»: volume, variedade de fontes e velocidade de geração de dados.

    - Observa-se também uma explosão do acesso à informação por meio de dispositivos móveis: o mercado global de telefones celulares está cada vez mais próximo do ponto de saturação e o de smartphones alcançará uma participação de 51% em 2016[12]. Embora ainda haja um longo percurso, o crescimento dos smartphones também está se desacelerando, o que aponta para uma nova tecnologia de substituição no futuro próximo, que possivelmente passe pela «Internet das coisas»[13] e por dispositivos wearables[14], tais como óculos, relógios, etc., com tecnologia móvel integrada.

    - A capacidade de armazenamento também cresce de forma exponencial e o custo unitário diminui ao mesmo ritmo: armazenar 1 GB de dados em 1980 custava 10 milhões de dólares; hoje mal chega a dez centavos de dólar[15]. Isso fez com que a quantidade de informação armazenada no mundo seja massiva: em 2015, há 8 ZB de dados; o triplo em relação a 2012[16,17].

    - O mesmo fenômeno ocorre no processamento: a capacidade de executar instruções por segundo a cada mil dólares de processador foi multiplicada por quase 300 desde o ano 2000[18]. Além disso, o desenvolvimento da computação distribuída permite paralelizar as operações em diversos cores e, suportada por gigantes tecnológicos e varejistas como a Google ou a Amazon, é apresentada como o futuro do processamento. No setor financeiro, os bancos digitais e os requisitos do mundo informático fazem com que as instituições necessitem de maiores capacidades de processamento e, por isso, já estão adotando máquinas de alto rendimento e computação distribuída.

    - Por último, as capacidades de modelagem estão evoluindo rapidamente, impulsionadas pelas novas tecnologias e pela disponibilidade de informação, que abrem um horizonte de possibilidades antes impensável[19]. O número de decisões que são tomadas de forma automática utilizando modelos nas instituições financeiras é multiplicado a cada ano, o

[10] Albert Einstein (1879-1955). Físico alemão (posteriormente naturalizado suíço e norte-americano), autor da teoria da relatividade, um dos dois pilares da Física moderna.
[11] Federal Big Data Commission (2014).
[12] International Telecommunication Union (2014).
[13] Interconexão digital de objetos quotidianos com a Internet. Estima-se que, em 2020, haja 26 bilhões de dispositivos ligados à Internet das coisas. (Gartner, 2013).
[14] Roupa ou acessórios que incorporam tecnologia eletrônica avançada.
[15] McCallum (2014).
[16] SiliconAngle (2014).
[17] O setor financeiro também segue esta tendência ascendente: cada instituição manipula, em média, 1,9 PB em seus sistemas de informática (DeZyre, 2014), o que está impulsionando o uso de plataformas de armazenamento distribuído.
[18] Kurzweil [Diretor de Engenharia da Google] (2014).
[19] Por exemplo, a proliferação da «aprendizagem automática», que permite uma modelagem mais ágil (5.000 modelos/ano com 4 analistas versus 75 modelos hoje).

que, sem dúvida, produz benefícios (eficiência, objetividade, automatização), mas também implica riscos.

2. Tais fatos estão convertendo os dados em uma nova commodity: são gerados, armazenados e processados a um custo muito reduzido, são fungíveis (perdem vigência) e, convenientemente transformados, têm o potencial de proporcionar um enorme valor. Isso requer profissionais e ferramentas especializadas; em outras palavras: data science.

3. O setor financeiro deveria ser um dos mais beneficiados pela adoção da data science. Não é por acaso que é o setor que manipula a maior quantidade e qualidade de informação de seus clientes (atuais e potenciais) para extrair conhecimento e incorporá-lo em sua proposta de valor (entendimento de suas necessidades, personalização da oferta, adequação do modelo de relacionamento multicanal, etc.).

4. No entanto, no setor financeiro esta revolução tecnológica soma-se a um contexto especialmente complexo, que combina elementos conjunturais com uma forte pressão regulatória e com alterações no comportamento dos clientes.

5. Nas economias desenvolvidas, a conjuntura macroeconômica é caracterizada por um período prolongado de baixas taxas de juros, um crescimento fraco e baixas taxas de inflação, o que penaliza as margens de lucro dos bancos. Nas economias emergentes, com uma elevada dependência do investimento público e de políticas fiscais expansivas, está ocorrendo uma certa desaceleração do crédito bancário, uma desaceleração do crescimento e o aumento da inadimplência. Esses fatores fazem com que seja necessário gerir a conta de resultados com maior intensidade.

6. No âmbito normativo, está ocorrendo um «tsunami regulatório» caracterizado pela proliferação, harmonização, endurecimento e pelo caráter transnacional e mais intrusivo das normas em vários âmbitos: (1) capital e liquidez: buffers de capital, índices de liquidez e alavancagem, revisão de requerimentos por risco de crédito, mercado e operacional; (2) supervisão prudencial: reforço do SREP, ICAAP e ILAAP[20] , testes de estresse supervisores; (3) limitação do apoio público: planos de recuperação e resolução, TLAC e MREL, ring-fencing; (4) governança corporativa: maiores exigências ao Conselho e à Alta Administração, novas figuras (CRO, CDO, CCO[21], etc.); (5) proteção dos consumidores: função de Compliance, controle de qualidade, gestão de reclamações, risco de conduta; (6) combate à fraude e os paraísos fiscais: FATCA, endurecimento de sanções por lavagem de dinheiro; (7) cibersegurança e segurança da informação: FISMA, Convênio de Cibercrime de Budapeste, Diretriz de Segurança nas Redes, ISO 27032; e (8) informação e reporte: RDA&RRF, COREP, FINREP, FR Y-14, FR Y-9C, etc.

7. Este processo dispendioso de adequação normativa implica, não obstante, um elemento diferencial para os clientes bancários, ao permitir que acessem os processos regulados mais seguros e supervisados, aspecto que as instituições financeiras acabarão por valorizar face aos demais novos concorrentes.

8. Por outro lado, o cliente bancário ficou mais exigente, está permanentemente conectado (usa o telefone celular 110 vezes por dia[22]) e consulta as redes sociais antes de comprar. Além disso, já não vê os bancos como os únicos fornecedores de serviços financeiros nem as agências como o canal básico de relacionamento, e o cliente se habituou à

[20] SREP: Supervisory Review and Evaluation Process; ICAAP: Internal Capital Adequacy Assessment Process; ILAAP: Internal Liquidity Adequacy Assessment Process.
[21] CRO: Chief Risk Officer; CDO: Chief Data Officer; CCO: Chief Compliance Officer.
[22] KPCB (2014).

personalização na oferta de serviços. Mas, ao mesmo tempo, dá sinais de confusão perante a abundância e a complexidade da oferta, o que favorece o poder de recomendação que os líderes de opinião exercem. Esta mudança está obrigando as instituições a reformular sua oferta de serviços e de seus canais, e a adotar uma visão mais focada no cliente, com impactos relevantes em todos os âmbitos.

9. Além disso, novos concorrentes afloram no setor financeiro com novos modelos de negócio, provenientes de setores não sujeitos à estrita regulação bancária, mas que têm uma imagem de marca que os clientes recebem de forma bastante favorável.

Data science: uma disciplina emergente

10. O contexto descrito está favorecendo a adoção no setor financeiro de uma disciplina emergente, proveniente, em grande parte, do setor tecnológico, e necessária para abordar a transformação enfrentada pelas instituições: a data science.

11. A característica essencial da data science[23] é o caráter de ciência; aproxima-se à extração de valor dos dados por meio de um método científico, o «processo data science»: a formulação de uma pergunta ou hipótese; a obtenção de informação de diversas fontes de dados massivos e possivelmente desestruturados para respondê-la; a exploração dos dados por meio de estatística descritiva; a modelagem do fenômeno com os dados disponíveis; e a visualização e comunicação dos resultados, que confirmarão ou refutarão a hipótese ou pergunta formulada.

12. A data science implica, portanto, a evolução da modelagem tradicional, em grande parte aprimorada pelo contexto big data, e emprega ferramentas e novas técnicas[24] que permitem o self-service de dados, o acesso móvel, a fusão de dados de várias fontes, a conectividade não relacional, a utilização da nuvem e a visualização interativa de dados[25] .

13. Graças a estas capacidades, a adoção da data science permite às instituições formular e responder perguntas antes impensáveis em todos os âmbitos (riscos, marketing, finanças, operações, etc.), sobre os clientes e seu contexto e, inclusive, sobre a própria organização.

14. Como exemplo, já é possível enriquecer os modelos de classificação de crédito com informação de redes sociais e do digital footprint, melhorar os modelos de estimativa de rendas com dados publicamente disponíveis na rede cruzados com geolocalização, prevenir a fuga de clientes analisando as gravações de call centers por meio do processamento da linguagem natural ou detectar a fraude e a lavagem de dinheiro com a detecção de padrões de comportamento nos logs de atividade, entre muitas outras possibilidades.

15. No entanto, a evolução para estas capacidades, não está isenta de desafios: o custo e a dificuldade de manipular volumes massivos de dados, os aspectos de privacidade, ética e segurança na manipulação dos dados, a captação e formação do talento em data science, o risco de que muitas decisões relevantes dependam de modelos automáticos, e a governança dos dados e dos modelos.

[23] Além da definição de data science descrita na introdução, a maioria dos estudos analisa as competências e conhecimentos que o data scientist necessita: (1) formação em Matemática, Física, Estatística, etc., e aprendizagem automática, algoritmia, otimização, simulação ou séries temporais, entre outros; (2) competências tecnológicas, domínio de linguagens estatísticas, manipulação de bancos de dados relacionais e não relacionais; e (3) um conhecimento profundo do negócio, que é a chave para o sucesso dos modelos.

[24] Novos dispositivos e canais de relacionamento com o cliente, novas plataformas, novos meios de pagamento, soluções BPM (business process management), redes sociais, tais como canal de contratação, acompanhamento de marca e atendimento de reclamações, sistemas distribuídos e escaláveis horizontalmente, infraestrutura como serviço, novos bancos de dados (NoSQL e in-memory), captura e processamento de dados em tempo real, novas ETL e motores de consulta para dados desestruturados, ferramentas de data discovery e novas linguagens de programação, além de novas ferramentas de visualização e de exploração de dados on-line.

[25] Para avaliar a importância deste perfil, lembramos que o presidente Barack Obama criou, em fevereiro de 2015, o cargo de Chief Data Scientist e nomeou pessoalmente Dhanurjay 'DJ' Patil.

16. A esse respeito, as instituições financeiras foram se adaptando ao fenômeno descrito, transformando seus processos de geração de dados e reporte, embora, em muitos casos, de forma desestruturada e como consequência de pedidos incrementais dos supervisores e reguladores, de necessidades de gestão não planejadas ou de processos de integração de instituições.

17. Os reguladores indicaram as carências na informação como uma das causas da crise financeira, o que levou à publicação de normas específicas com fortes requerimentos em qualidade, consistência, integridade, rastreabilidade e replicabilidade dos dados (especialmente no caso de riscos[26]). Tais normas conduzem à necessidade de revisão dos frameworks de governança dos dados das instituições.

18. O esquema de governança dos dados deveria ser concretizado em um framework de gestão, que descrevesse os princípios, os intervenientes (com novas figuras como o CDO[27]), a estrutura de comitês, os processos críticos relacionados a dados e informação, as ferramentas (dicionário de dados, arquitetura de datawarehouses, soluções de exploração, etc.), e o controle da qualidade dos dados.

19. A governança dos dados apresenta vários desafios, entre os quais sem destacam o envolvimento da Alta Administração, a definição do escopo de dados objeto do framework de governança, os aspectos de privacidade no uso dos dados e cibersegurança (que inclui a proteção contra o «hacktivismo», os ciberdelitos financeiros, a espionagem e o roubo de informação) ou a adaptação às novas arquiteturas de armazanamento, como los data lakes[28], entre outros. Mas é um fato a relevância que essa governança adquiriu na

[26] Risk Data Aggregation and Risk Reporting Framework; ver BCBS (2013).
[27] Chief Data Officer.
[28] Repositórios massivos de dados por transformar (no formato origem).

gestão das instituições, convertendo-se em condição necessária para o aprovisionamento correto dos dados, a informação disponível e, em suma, em um pilar estratégico das instituições.

20. No caso dos modelos, as normas também incidem na necessidade de dispor de um framework de governança dos mesmos[29]. Os elementos deste framework foram abordados em detalhe em publicações anteriores da Management Solutions[30].

21. As instituições mais avançadas neste assunto já desenvolveram frameworks de gestão do risco de modelo, que regulam o inventário e a classificação dos modelos, sua documentação e um esquema de acompanhamento dos mesmos.

22. De qualquer forma, a governança dos modelos também apresenta desafios, entre os quais, o envolvimento da Alta Administração, a reflexão sobre o escopo (o que é um modelo e que modelos devem se submeter a essa governança), a segregação de funções (ownership, controle e compliance[31]), o effective challenge ou a disposição de ferramentas de inventário e workflow de modelos, entre outros. Mas é inquestionável que a governança dos modelos é uma questão que requer um envolvimento no primeiro nível, porque dele depende a correta tomada de decisões nas instituições.

23. Em síntese, a governança dos dados e dos modelos constitui um elemento estratégico para as instituições financeiras, impulsionado pelas normas e como resposta ao fenômeno big data. Esta questão tem impacto em vários âmbitos de uma instituição, desde a organização, passando pelas políticas e procedimentos, até as ferramentas e os sistemas de informação, e é formada como um pilar-chave de atuação nos próximos anos.

Estudo de caso: redes sociais e credit scoring

24. Para ilustrar alguns dos conceitos descritos, e como estudo de caso, apresentamos um modelo de scoring de crédito que utiliza dados de redes sociais integrado com um modelo tradicional. Também analisamos em que medida a inclusão destes dados melhora o poder de previsão do modelo tradicional.

25. Entre as conclusões, destacamos: (1) os dados nas redes sociais têm uma qualidade muito inferior aos internos; menos da metade dos clientes tem dados, e desses, um número reduzido está completo; (2) existe um problema de desambiguação[32] para identificar cada cliente de forma inequívoca; (3) apesar disso, o poder de previsão do modelo baseado em redes sociais é equiparado ao do modelo tradicional e, ao combiná-los, o poder de previsão resultante é aumentado[33]; (4) para isso, foram utilizadas variáveis da formação regulada e não regulada do cliente, sua experiência profissional, a localização geográfica e outros dados relativos a gostos e interesses.

26. O estudo demonstra o potencial da aplicação da data science no âmbito de riscos, e é extensível a outros tipos de modelos (avaliação de garantias, fidelização, renda, abandono, propensão à compra, etc.) e fontes de informação (logs internos, bancos de dados públicos, informação web, etc.).

27. A data science figura como uma matéria multidisciplinar emergente que abre um novo campo de possibilidades para o setor financeiro, ao aplicar uma abordagem científica sobre o fenômeno big data, aproveitando a explosão de informação e de capacidades tecnológicas para aumentar a inteligência das instituições[34].

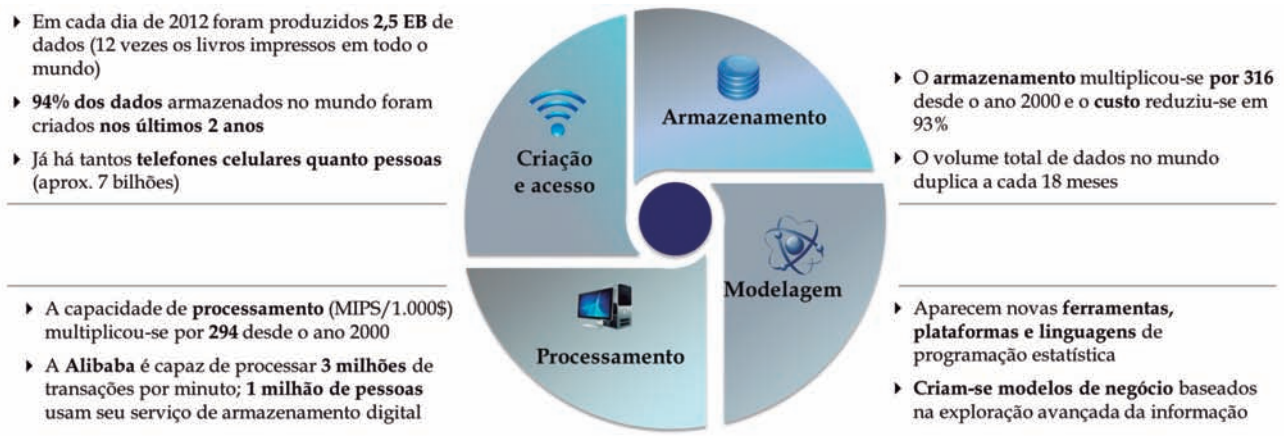
[29] OCC/Fed (2011-12).
[30] Ver Management Solutions (2014): Model Risk Management: aspectos quantitativos e qualitativos da gestão do risco de modelo.
[31] O model owner define os requisitos do modelo e costuma ser seu usuário final. O Controle inclui a mensuração do risco de modelo, o estabelecimento de limites e o acompanhamento, assim como a validação independente. O Compliance abrange os processos que assegurem que as funções do model owner e de controle sejam desempenhadas de acordo com as políticas estabelecidas.
[32] Técnica que permite identificar um cliente de forma unívoca entre vários que compartilham o mesmo nome e outras características (localização, idade, etc.).
[33] Medido com a ROC (Receiver Operating Characteristic), métrica de poder de previsão de um modelo de classificação binária.
[34] Capacidade de receber, armazenar e processar informação para resolver problemas.

Um setor financeiro em transformação

A tecnologia e a infraestrutura necessárias para conectar as pessoas não têm precedentes, e acreditamos que este é o maior problema em que devemos nos concentrar.

Mark Zuckerberg[35]

Fig. 2. A revolução tecnológica: um instantâneo do momento atual.



A revolução tecnológica

O setor financeiro enfrenta uma revolução tecnológica de uma dimensão e uma extensão sem precedentes, que está transformando sua atividade de forma substancial.

A principal característica dessa revolução é a aceleração. Como prevê a Lei de Moore, a potência tecnológica, medida em várias dimensões, está crescendo de forma exponencial, e as previsões apontam que este fenômeno irá continuar. Esta revolução ocorre em quatro dimensões: geração e acesso à informação (incluindo mobilidade), armazenamento, processamento e modelagem.

Para compreender este fenômeno e suas implicações, é necessário observá-lo em quatro dimensões: geração e acesso à informação, armazenamento, processamento e modelagem (Fig. 2).

Fig. 3. Alguns números sobre geração e intercâmbio de informação.

Fonte: Pethuru (2014)

Nesta seção, serão analisados os três primeiros; os detalhes sobre modelagem e data science serão abordados na próxima seção.

Geração e acesso à informação

A primeira faceta deste fenômeno é o aumento da velocidade de geração dos dados digitais. As estimativas da velocidade e da aceleração da geração de dados digitais variam segundo os analistas, mas todos concordam que se trata de uma aceleração exponencial e em todos os âmbitos; para citar alguns exemplos[36,37](Fig. 3):

- Em 2012, foram gerados 2,5 exabytes de dados todos os dias; esta taxa continuou aumentando.
- Mais de 90% de todos os dados que existem hoje foram gerados nos dois últimos anos.
- A cada minuto são carregadas 12 horas de vídeo no YouTube.
- Todos os dias são gerados 12 terabytes de tweets no Twitter.
- São produzidos 5 bilhões de transações financeiras por dia.
- Em 2012, havia 2,4 bilhões de usuários de Internet no mundo, quase metade deles na Ásia…
- … que trocaram 144 bilhões de e-mails por dia, dos quais 69% eram spam.
- Também em 2012, o Facebook ultrapassou 1 bilhão de usuários, e estima-se que em 2016 tenha mais usuários que a população da China.

[35]Mark Elliot Zuckerberg (n. 1984). Cofundador e diretor-geral do Facebook.
[36]IBM (2014a).
[37]Pingdom (2015).

Fig. 4. Quanto é um zettabyte? Algumas magnitudes orientativas.

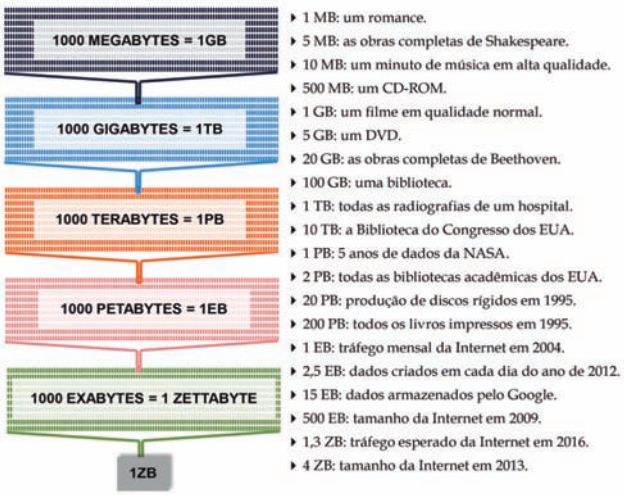


Fig. 5. Volume, variedade e velocidade de geração de dados no âmbito comercial.

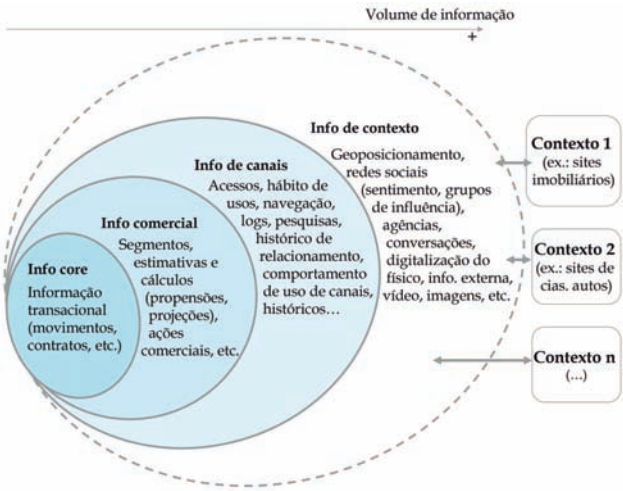
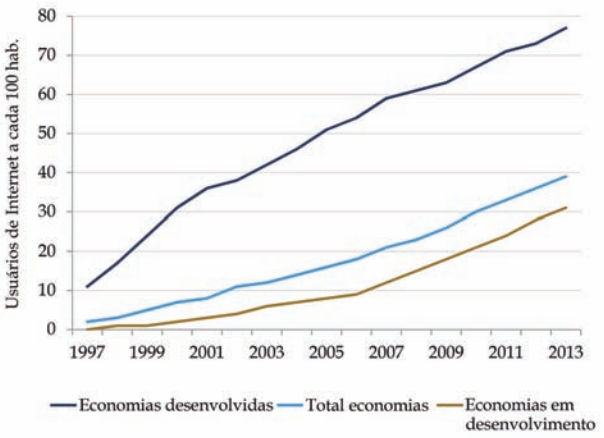


Fig. 6. Usuários de Internet a cada 100 habitantes.



Fonte: International Telecommunications Union (2012).

Nas palavras da Federal Big Data Commission, à qual o Governo dos Estados Unidos encomendou a missão de compreender o fenômeno big data nas agências governamentais[38]:

> *Nos últimos anos, os Governos federais, estaduais e municipais enfrentam um maremoto de mudanças como resultado do aumento drástico no volume, na variedade e na velocidade dos dados em seus próprios entornos e em todo o ecossistema governamental. […]*
>
> *Desde o ano 2000, a quantidade de informação que o Governo federal captura aumentou exponencialmente. Em 2009, o Governo dos Estados Unidos produziu 848 petabytes de dados, e só o sistema de Saúde alcançou os 150 exabytes. Cinco exabytes (10^{18} gigabytes) de dados conteriam todas as palavras pronunciadas por todos os seres humanos. A esta taxa, o fenômeno big data na área da Saúde dos EUA rapidamente alcançará a escala dos zettabytes (10^{21} gigabytes) e, pouco depois, os yottabytes (10^{24} gigabytes).*

Estes números vertiginosos são difíceis de imaginar; ver a Fig. 4 para uma referência orientativa de cada um destas magnitudes.

Por outro lado, esses dados já não são gerados apenas de forma estruturada; pelo contrário, 80% dos dados que são gerados a cada dia são desestruturados: vídeos, imagens, e-mails, etc., e provêm de uma ampla variedade de fontes novas: redes sociais, sensores, registros de navegação por Internet, logs de atividade, registros de chamadas, transações, etc.

Em outras palavras, o fenômeno big data é uma explosão no volume, na variedade e na velocidade de geração de dados, o que foi chamado «os três V's do big data» (aos quais alguns autores acrescentam o quarto V, de «veracidade»). Como exemplo, no âmbito comercial das instituições financeiras ocorre uma expansão nestas três dimensões, que foi crescendo desde a informação core transacional, passando por dados comerciais e provenientes de canais, até chegar à informação proveniente do contexto do cliente, de grande riqueza, variedade e heterogeneidade (Fig. 5).

Em relação ao acesso à informação, embora subsistam diferenças entre as economias desenvolvidas e as emergentes e alguns países mal tenham acesso à Internet, a tendência é clara, e estima-se que, em poucos anos, será alcançado o pleno uso da Internet em quase todo o mundo (Fig. 6).

Igualmente, observamos uma explosão no acesso à informação por meio de dispositivos móveis. O número de telefones celulares já se equipara ao número de habitantes e nas economias desenvolvidas ultrapassa em 21%[39], enquanto as economias em desenvolvimento se aproximam da paridade, com nove linhas de telefones celulares a cada dez habitantes.

[38] Federal Big Data Commission (2014).
[39] International Telecommunication Union (2014).

As instituições financeiras não ficaram alheias a esta proliferação de dispositivos e, nos países desenvolvidos, observa-se uma evolução constante do banco digital (Fig. 7), que, em grande parte, é impulsionado pelo próprio cliente, contém o banco móvel e vai mais além. Partindo do e-banking e da integração multicanal da primeira década do século XXI e passando pelo fenômeno da omnicanalidade que se desenvolveu nos últimos anos, para o futuro, a tendência fica marcada pela chamada «Internet das coisas»: o predomínio e a ubiquidade dos dispositivos inteligentes, com capacidade para gerar informação digital sobre a sua utilização.

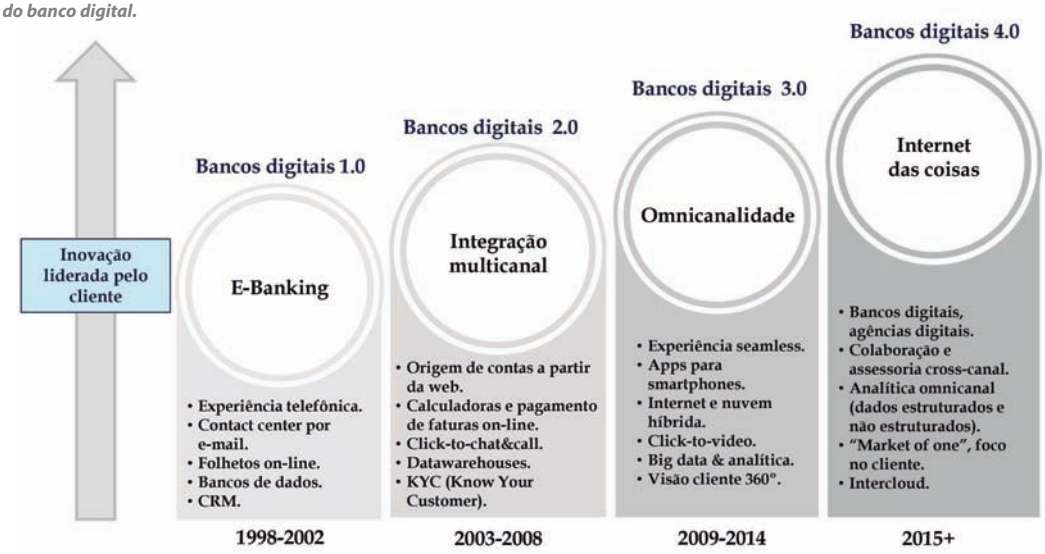
Nesse sentido, nos Estados Unidos quase 60% dos clientes já operam apenas por canais digitais, passam 400 vezes mais tempo nestes canais do que na agência e apenas 1% das transações ocorrem na agência. Apesar disso, 90% das vendas continuam ocorrendo nas agências (Fig. 8).

A evolução, portanto, é heterogênea, e grande parte das instituições está ainda na fase de adaptação ao banco móvel; a maioria desenvolveu aplicativos para dar acesso móvel aos principais serviços. De acordo com o estudo realizado pela Federal Reserve dos Estados Unidos[40], neste país, mais de 50% dos usuários de um smartphone utilizaram o banco móvel nos últimos 12 meses, e os usos mais comuns (Fig. 9) são consultar o saldo ou transações recentes (93%) e realizar transferências entre contas próprias (57%).

A tendência é clara: o banco digital está ganhando participação e, como os clientes com idade inferior a 29 anos utilizam os canais digitais quase quatro vezes mais do que os tradicionais, é previsível que esta expansão continue nos próximos anos. Contudo, a monetarização deste fenômeno apresenta evidentes oportunidades de melhoria.

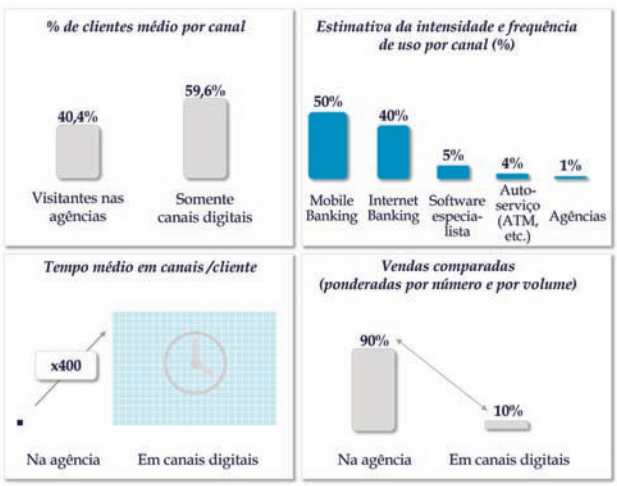
[40]Fed (2014).

Fig. 7. Evolução do banco digital.



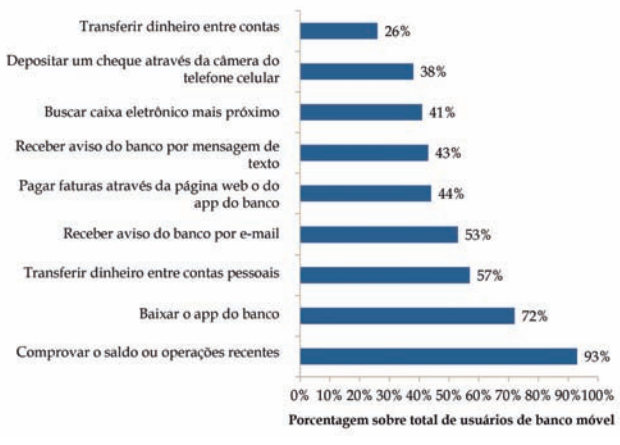
Fonte: CISCO.

Fig. 8. Transformação dos canais nas instituições financeiras.



Fonte: Digital Leadership GmbH (2014).

Fig. 9. Usos do banco móvel.



Fonte: Fed (2014).

Armazenamento

De forma paralela à geração e ao acesso à informação, a capacidade de armazenamento também está crescendo de forma exponencial, segundo a Lei de Moore, e seu custo unitário continua diminuindo no mesmo ritmo (Fig. 10). Enquanto em 1980 armazenar um gigabyte requeria dispositivos no valor de 10 milhões de dólares, hoje custa apenas dez centavos de dólar em uma diminuta fração de um disco rígido SSD.

Os dispositivos de armazenamento evoluíram de forma acelerada, desde a fita magnética de 1920, passando pelos tubos de raios catódicos de 1940, o primeiro disco rígido de 1956 (Fig. 11), a fita cassete de 1963, a memória DRAM de 1966, os disquetes (floppy disks) da década de 1970, os CDs de 1980, os zips e DVDs de 1994 e 1995, os cartões flash de 1995, os cartões MMC de 1997, os pen drives de 1999, os cartões SD de 2000, o Blu-Ray de 2003, até a memória sólida e o armazenamento na nuvem da década de 2010. A quantidade de novos formatos por década aumentou de forma exponencial, como também ocorreu na capacidade dos dispositivos.

Tais evoluções fizeram com que a quantidade de informação armazenada no mundo tenha crescido de forma massiva nos últimos anos. Estima-se que, em 2012, havia um total de 2,75 zettabytes de dados armazenados digitalmente, e que esta cifra chegue aos 8 zettabytes em 2015 e continue sua trajetória ascendente[41].

No setor financeiro, a evolução dos sistemas de armazenamento ocorreu em paralelo com a necessidade de compilar e gerir grandes quantidades de informação. No final da década de 1970, as instituições começaram a implantar servidores host, entornos tecnológicos onde é recebida, processada e armazenada toda a informação gerada pela gestão de transações, que é seu objetivo fundamental. Posteriormente, a incorporação de sistemas de informação permitiu o desacoplamento da consulta massiva de informação dos processos operacionais, abrindo as portas para quantidades muito superiores de dados e sua historização.

Esta tendência concretizou-se com a criação de datawarehouses, repositórios de informação estruturada com quatro características principais:

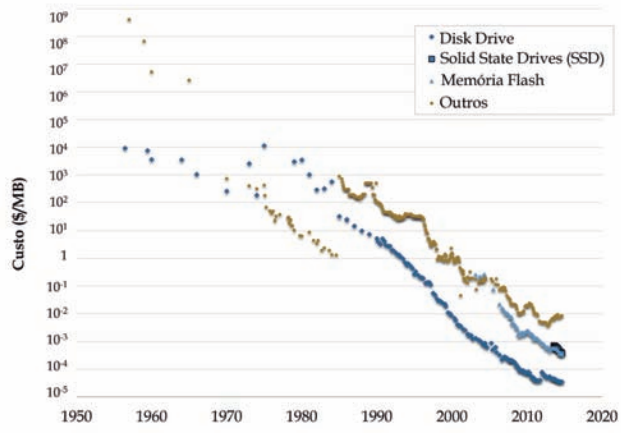
- Orientado: os dados armazenados são estruturados e agrupados por temáticas.
- Atualizado: permitem incorporar nova informação ao longo do tempo.
- Histórico: mantém um registro de dados históricos, não é necessário eliminá-los.
- Integrado: asseguram a consistência dos dados registrados por diferentes sistemas.

No início do século XXI, o setor financeiro incorporou nova tecnologia para modernizar e otimizar o armazenamento de informação com grandes sistemas dedicados, que se caracterizam por utilizar software e hardware concebidos especificamente para a exploração de informações. Paralelamente, foi iniciada a implantação de software de consulta mais sofisticado, que permite maior liberdade ao usuário (ferramentas OLAP, query & reporting, etc.).

Atualmente, as instituições financeiras já manipulam cerca de 1,9 petabytes em seus sistemas de informações[42], representando um desafio para a arquitetura implantada. Consequentemente, estamos vivendo uma nova revolução nos sistemas de armazenamento de informação: as plataformas de armazenamento distribuído. Esta tecnologia permite o armazenamento massivo de dados desestruturados e sua gestão por meio de uma arquitetura nodular.

[41] SiliconAngle (2014).
[42] DeZyre (2014).

Fig. 10. Custo de armazenamento, em dólares por megabyte (observe que o eixo Y é logarítmico).



Fonte: McCallum (2014).

Fig. 11. Disco rígido IBM 350, de 1956, com uma capacidade de 5 megabytes.

De fato, algumas das principais instituições financeiras, e nos Estados Unidos, pelo menos três das cinco maiores[43], já adotaram plataformas de armazenamento distribuído e começaram a explorar seu potencial de armazenamento e processamento de dados, ainda que de forma limitada.

Processamento

Assim como a geração e o armazenamento da informação, a potência de processamento também está tendo o mesmo crescimento acelerado[44]. Levando em conta o custo, a capacidade de processar instruções por segundo a cada mil dólares de processador foi multiplicada por quase 300 desde o ano 2000. Isso permite que alguns varejistas sejam capazes de processar milhões de transações comerciais por minuto, o que está na essência de seu modelo de negócio.

Por outro lado, a aparição da computação distribuída permitiu combinar as capacidades de vários processadores (cores) para executar operações de forma paralela. Em 2012, estimava-se[45] que a Google dispunha de cerca de 7,2 milhões de cores em mais de 1,8 milhão de máquinas e que, com sua potência combinada, era capaz de executar cerca de 43.000 trilhões de operações por segundo, quatro vezes mais do que a máquina mais potente do mundo (Fujitsu K). O mesmo estudo calcula que a Amazon podia alcançar os 240 trilhões de operações por segundo.

Os principais players tecnológicos, conscientes de que sua capacidade de processamento distribuído é em si um serviço valioso, comercializam o acesso à computação distribuída em suas respectivas nuvens; a superabundância desta capacidade possibilita o aluguel a custos inferiores a um dólar por hora[46].

No setor financeiro, por sua vez, a exigência de processamento da atividade tradicional bancária (nas agências ou nos terminais ponto de venda) é suportada pela tecnologia implantada atualmente de forma razoavelmente satisfatória. No entanto, os novos canais como os bancos digitais, requerem um aumento da capacidade de processamento transacional em paralelo.

No mundo informacional, por sua vez, as sofisticadas técnicas de modelagem fazem com que as instituições requeiram uma maior capacidade de cálculo e processamento massivo de informação. Para tal, a adoção de técnicas de computação paralela ou distribuída, bastante ligadas à própria estrutura de armazenamento, permite às instituições tirar o máximo proveito da informação diminuindo o tempo de processamento.

Uma nova commodity: os dados

Esta explosão na capacidade de gerar, armazenar e processar informação, e acessá-la a qualquer momento e local por meio de dispositivos móveis, está causando um novo fenômeno: os dados são agora uma nova commodity. De fato, os dados são gerados, armazenados e processados com um custo muito reduzido, são fungíveis porque perdem vigência com rapidez, e são a matéria-prima que, transformada, permite a geração de serviços de todos os tipos.

No entanto, esta nova commodity tem duas características particulares: assim como a energia, tornou-se indispensável para o funcionamento da maioria dos negócios, incluindo os serviços financeiros; e, como todas as commodities, requer profissionais e ferramentas especializadas para seu processamento. Este é precisamente o campo da data science.

[43]Goldman (2014).
[44]Ver gráfico na introdução.
[45]Pearn (2012).
[46]Para 8 cores com 30 GB de RAM na Amazon, Google e Windows, por exemplo.

Fig. 12. Conjuntura macroeconômica e fatores de risco.

Contexto do setor financeiro

Embora esta revolução tecnológica tenha um impacto relevante em todos os setores, é provável que o setor financeiro seja um dos mais beneficiados pela adoção da data science como pilar estratégico. É o setor que manipula a maior quantidade e qualidade de informação de seus clientes e targets e, portanto, tem um enorme potencial para extrair conhecimento e incorporá-lo à sua proposta de valor, o que o diferencia de outros setores.

No entanto, no setor financeiro esta revolução tecnológica ocorre em um contexto singular, caracterizado por uma difícil conjuntura macroeconômica, um contexto regulatório exigente e uma alteração no padrão de comportamento dos clientes, fatores que não estão tendo impacto do mesmo modo em outros setores.

Conjuntura macroeconômica

No aspecto macroeconômico (Fig. 12), mantém-se o caráter dual da economia mundial (países desenvolvidos versus emergentes) em termos de crescimento, pressões inflacionárias e fluxos de investimento, de forma que certos padrões continuam nos principais indicadores macro (Fig. 13) que afetam a evolução do negócio bancário tanto em suas fontes de financiamento como em seus investimentos e margens financeiras.

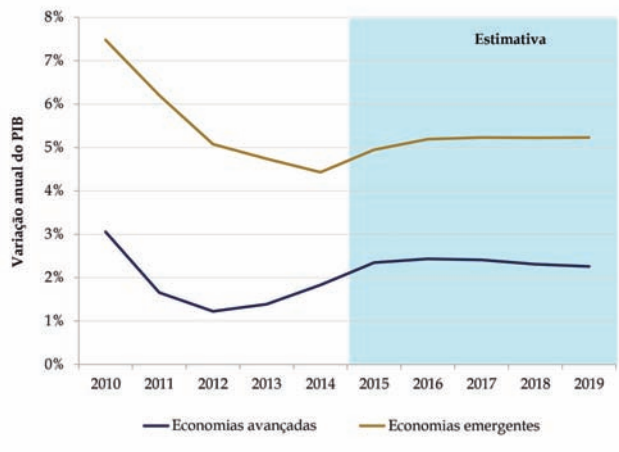
No caso das economias avançadas, o cenário prolongado de taxas de juros baixas gerou um aumento relativo do preço das ações, a compressão dos spreads e uma redução generalizada da volatilidade até voltar a níveis anteriores à crise. No entanto, isso não gerou uma retomada do investimento, o que contrasta com a evolução da poupança, que aumentou, provocando uma fraqueza maior da demanda privada. Dessa forma, espera-se que, nas economias avançadas, este estancamento (que foi classificado como «estancamento secular»[47]) prossiga durante vários anos, embora de forma heterogênea entre países.

Além disso, cada vez mais existe maior evidência de que o crescimento potencial das economias avançadas começou a diminuir antes do início da crise financeira como consequência do envelhecimento da população ativa e do fraco crescimento da produtividade dos fatores[48].

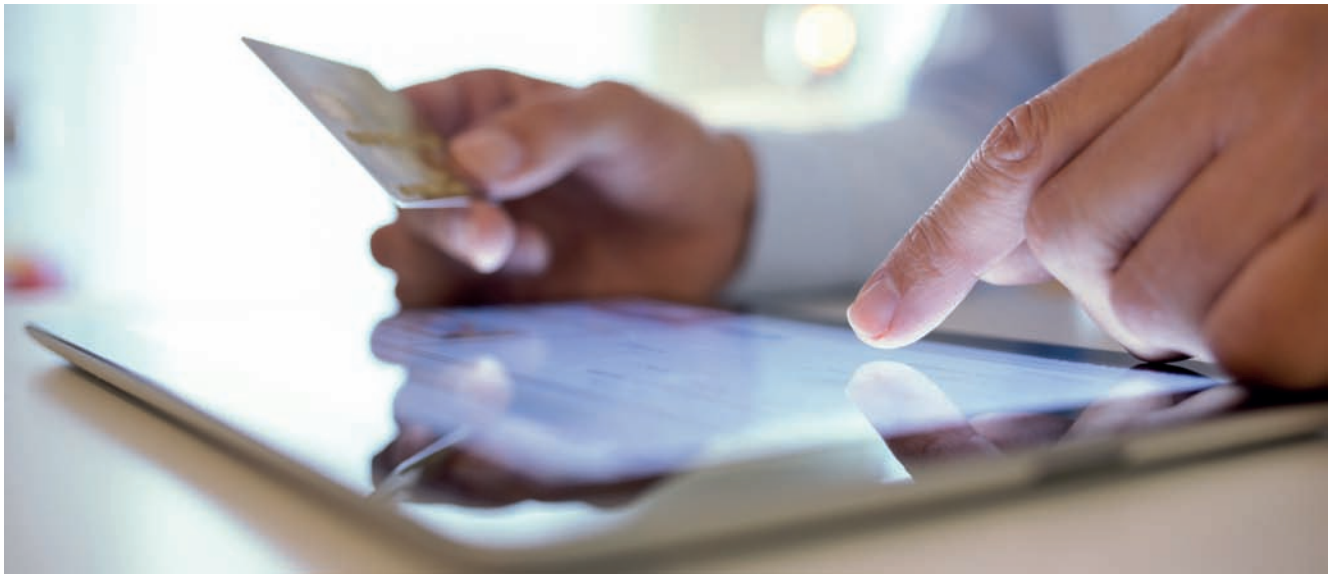
Em algumas economias avançadas, a inflação também é baixa ou existem indícios de deflação e, assim, as taxas de referência não têm margem de redução. Tais fatos têm um impacto negativo na evolução do crédito e nas margens das instituições financeiras.

[47] FMI (2014).
[48] Para o caso dos Estados Unidos, ver, por exemplo, Fernald (2014), Gordon (2014) e Hall (2014).

Fig. 13. Crescimento do PIB em economias emergentes versus avançadas.



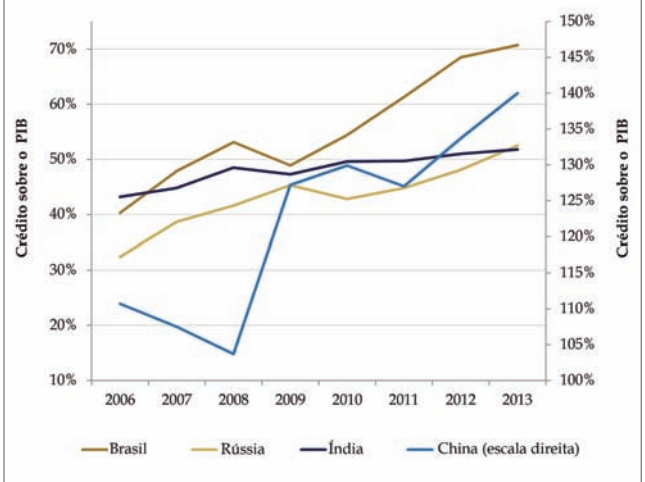
Fonte: IMF (2014).



No caso das economias emergentes, o crescimento desacelerou, embora ainda se mantenham níveis relativamente elevados em relação às economias avançadas. O consumo privado contribui em grande parte para este crescimento, embora também exista uma dependência elevada do investimento público e das políticas fiscais expansivas.

Embora a expansão do crédito bancário desacelere em alguns mercados emergentes (Brasil, Índia, Rússia), mantêm-se taxas de crescimento de dois dígitos (Fig. 14). Por outro lado, a taxa de inadimplência está aumentando de forma generalizada nas economias emergentes devido a várias causas, entre as quais os problemas em determinados setores (como o setor de mineração no Peru ou o setor público no Chile e no Brasil[49]) e a incorporação de novos clientes antes não bancarizados e com um pior perfil de crédito.

Fig. 14. Crédito sobre o PIB em economias emergentes.



Fonte: IMF (2014).

Por último, coexistem dois elementos que geram incerteza neste contexto. Por um lado, a dependência econômica da China, cuja desaceleração provocaria uma contração massiva das exportações no restante do mundo, a redução do preço das matérias-primas e a queda dos índices de confiança de consumidores e empresas. A esse respeito, na China existem riscos sobre o crescimento devido ao excesso de capacidade de produção e uma superabundância de crédito, que são os principais impulsionadores de seu crescimento.

Por outro lado, a normalização da política monetária nos Estados Unidos, Japão e União Europeia, que nos últimos anos se expandiu para países como Chile, México e Peru, representa um risco pelo possível efeito deflacionário e pela atração de fluxos de investimento dos países emergentes.

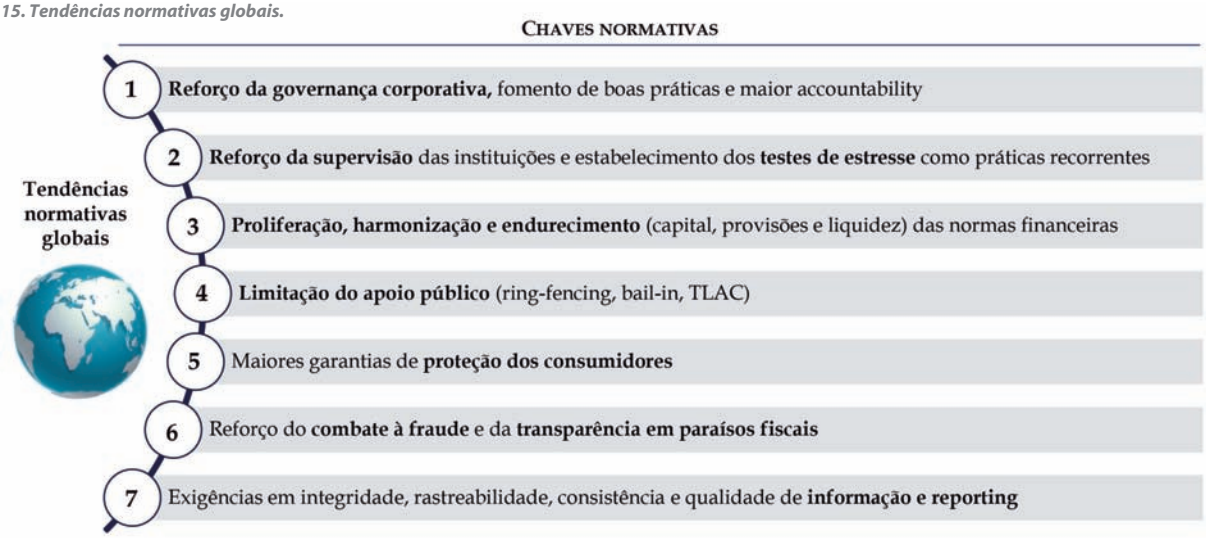
Esta conjuntura macroeconômica está gerando um estreitamento das margens no setor financeiro, mas também coloca pressão sobre o capital e a liquidez. A consequência é que as instituições intensificaram a gestão da rentabilidade, do capital e da estrutura de balanço, provendo-as de maior inteligência analítica e uma visão de riscos, com atenção às previsões econômicas e seu potencial impacto.

Contexto regulatório

O setor financeiro está vivenciando uma proliferação notável, que cabe classificar como «tsunami», tanto de regulação supranacional como de normas locais nos âmbitos financeiros que tiveram maior influência na crise iniciada em 2007: contábil, de supervisão prudencial, de conduta e compliance, de governança corporativa, de proteção ao consumidor e de riscos, em um sentido amplo.

[49]BBVA Research (2014).

Fig. 15. Tendências normativas globais.



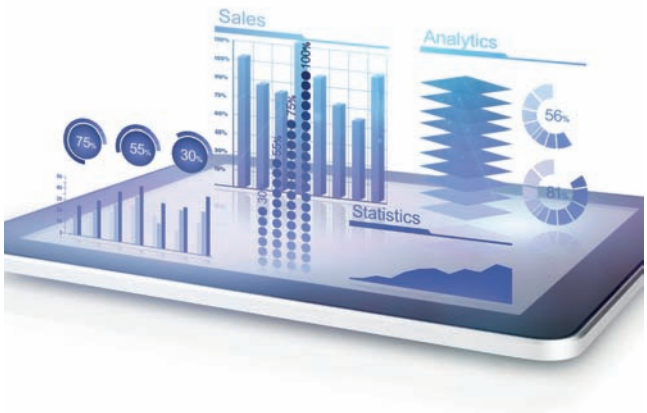
Igualmente, observa-se uma tendência de harmonização das normas entre os diferentes países, o que contribui de forma decisiva para a constituição de órgãos reguladores e supervisores supranacionais, como a Autoridade Bancária Europeia (EBA), o Mecanismo Único de Supervisão (SSM) do Banco Central Europeu ou o Financial Stability Board (FSB), entre outros.

Ao mesmo tempo, o caráter das normas está passando a ser mais intrusivo e mais prescritivo e deixa menos espaço para adaptações e interpretações. Como exemplo, na União Europeia foi adotado o Basileia III como uma combinação de um Regulamento (portanto, de aplicação imediata em todos os países da União) e uma Diretiva (que deve ser transposta para as normas locais); enquanto o Basileia II foi adotado apenas na forma de uma Diretiva.

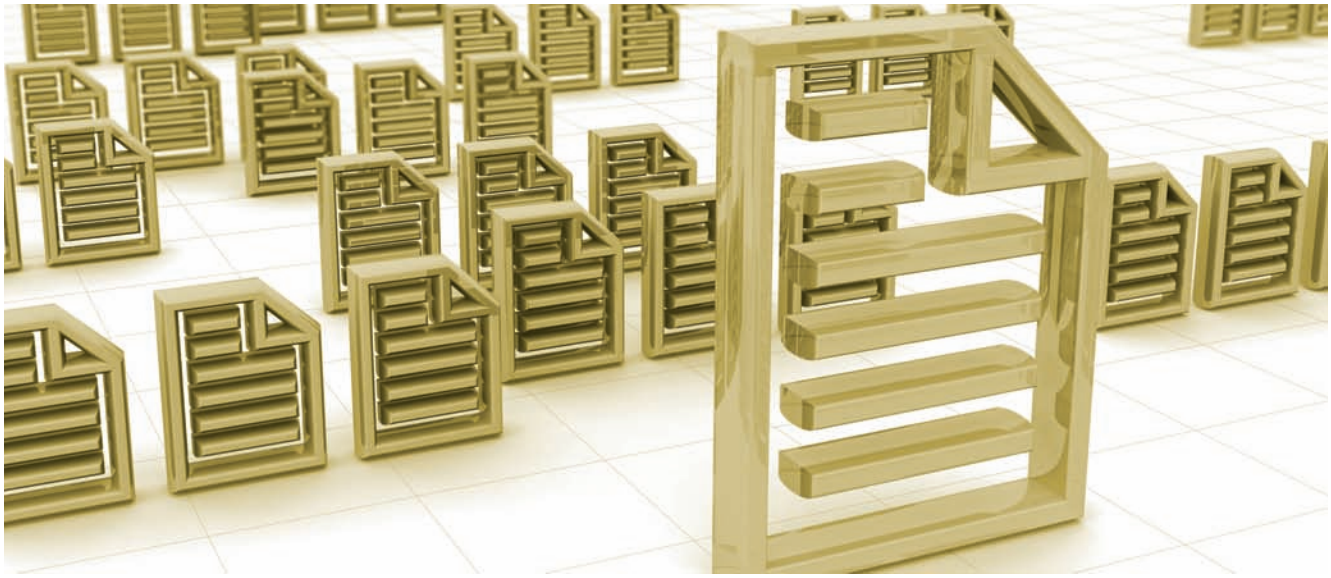
Especificamente, esta proliferação e harmonização de normas financeiras está sendo concretizada em mais normas (e de caráter mais restritivo) em vários âmbitos, entre os quais destacamos (Fig. 15):

- Capital e liquidez: como consequência do Basileia III, aparecem maiores requerimentos de capital (tanto em quantidade como em qualidade), um novo índice de alavancagem e dois índices de liquidez (a curto e a longo prazo[50]). Além disso, os requerimentos de capital por risco de crédito, de mercado e operacional são revistos e simplificados[51].

- Reforço da supervisão prudencial: são estabelecidas diretrizes comuns[52] para a supervisão das instituições, que reforçam os processos SREP, ICAAP e ILAAP[53] (especialmente na Europa, com a entrada em vigor do Mecanismo Único de Supervisão em novembro de 2014). Além disso, os testes de estresse supervisores são solidificados e estabelecidos como práticas recorrentes[54].



[50] Management Solutions (2012).
[51] A esse respeito, o Comitê de Supervisão Bancária de Basileia emitiu, em 2014, vários documentos que revisam o método padronizado de risco de crédito, simplificam o cálculo de capital por risco operacional, estabelecem mínimos de capital e revisam o cálculo de capital na carteira de negociação, entre outros.
[52] EBA (2014) e ECB (2014).
[53] SREP: Supervisory Review and Evaluation Process; ICAAP: Internal Capital Adequacy Assessment Process; ILAAP: Internal Liquidity Adequacy Assessment Process.
[54] Management Solutions (2013).



- Limitação do apoio público: especialmente nas economias avançadas, pretende-se que em nenhum caso o Estado tenha que voltar a resgatar instituições com fundos públicos por serem sistêmicas, o que se conhece como o fim do «too big to fail». Para isso, as instituições são obrigadas a ter planos de recuperação e resolução[55], e a dispor de passivos suficientes com capacidade de absorção de perdas (TLAC, MREL); e, na União Europeia, cria-se uma autoridade para gerir a resolução das instituições inviáveis (Single Resolution Board). Por outro lado, nos Estados Unidos e no Reino Unido, e posteriormente no restante da União Europeia, a regulação sobre ring-fencing é fortalecida, que obriga a separação jurídica entre as atividades de atacado e de bancos tradicionais.

- Reforço da governança corporativa: impõem-se maiores exigências ao Conselho de Administração e à Alta Administração sobre a aprovação e a supervisão do cumprimento da estratégia de negócio, o apetite ao risco e o framework de gestão de riscos, e criam-se novas figuras-chave (CRO, CDO, CCO[56], funções corporativas de Risk MI, etc.).

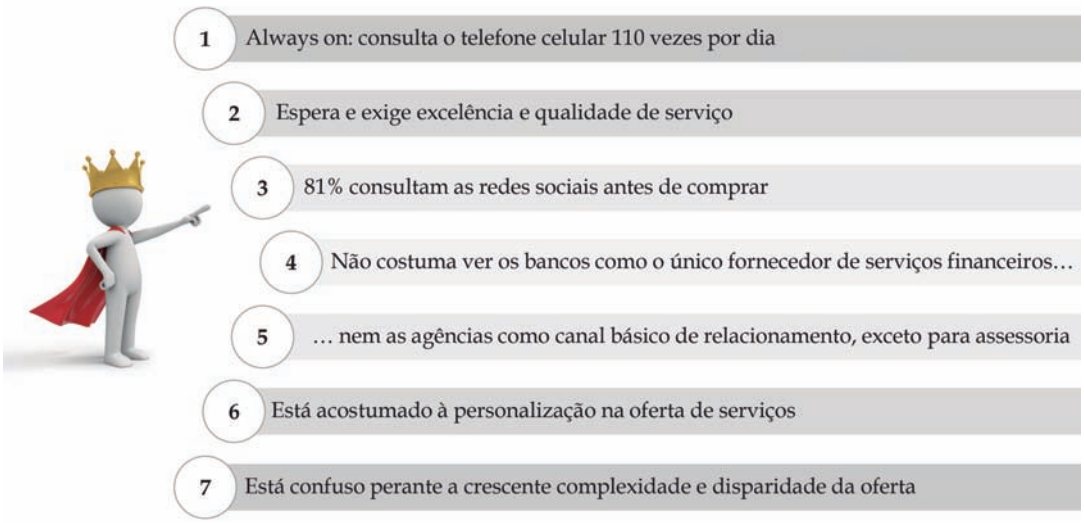
- Proteção dos consumidores: como consequência dos escândalos no setor financeiro associados a produtos, canais de distribuição, tecnologia de pagamentos, abuso de mercado e lavagem de dinheiro, aparece uma regulação mais intensiva e prescritiva[57] que exige o reforço da função de Compliance (recursos, meios, capacidades e linhas de reporte), de controle de qualidade (mystery shopping) e da política de gestão de reclamações, assim como o reforço da mensuração, gestão, controle, supervisão e reporte do risco de conduta para mercados e clientes. Esta tendência é liderada pelo Reino Unido, com a criação de uma autoridade específica, a Financial Conduct Authority, que apenas entre 2013 e 2014 aplicou quase 2 bilhões de libras em multas por questões de conduta[58].

- Combate à fraude e os paraísos fiscais: devido ao aumento da fraude pelo uso intensivo dos canais eletrônicos e pelas constantes mudanças nas organizações, observa-se a necessidade de contar com um controle intensivo de fraude interna e externa. Aparecem políticas agressivas de alguns países para evitar fraude fiscal dos seus cidadãos (por exemplo, FATCA[59]). Aumentam as exigências na combate à lavagem de dinheiro (sanções elevadas[60]), o que exige adaptações importantes nos processos e sistemas das instituições.

- Cibersegurança: normas específicas para combater o aumento dos ataques à segurança das instituições é desenvolvida («hacktivismo», ciberdelitos financeiros, espionagem, roubo de informação, etc.): nos Estados Unidos, o Federal Information Security Management Act (FISMA), entre outros; na Europa, o Convênio contra o Cibercrime de Budapeste ou a Diretiva de Segurança nas Redes (SRI); e em âmbito global, a norma ISO 27032, que proporciona diretrizes específicas sobre cibersegurança.

[55] Na União Europeia, por meio da Diretiva de Recuperação e Resolução de Instituições Bancárias (BRRD), resumida na European Commission (2014).
[56] CRO: Chief Risk Officer; CDO: Chief Data Officer; CCO: Chief Compliance Officer; MI: Management Information.
[57] Mortgage Market Review e Retail Distribution Review no Reino Unido, Diretiva sobre Abuso de Mercado e sobre Lavagem de Dinheiro e Reporte sobre Tendências nos Consumidores na União Europeia, entre outros.
[58] FCA (2015).
[59] A Foreign Account Tax Compliance, lei federal dos EUA que exige que instituições financeiras de todo o mundo reportem à agência fiscal dos EUA as contas que os cidadãos norte-americanos têm no estrangeiro.
[60] Como a multa de 1,9 bilhão de dólares, aplicada ao HSBC por falhas em seus controles contra a lavagem de dinheiro; ver Bloomberg (2013).

Fig. 16. O cliente de hoje.



- Informação e reporte: os processos de geração de informação e reporte de riscos foram perdendo eficácia por diversas razões, o que foi indicado como uma das causas da crise, o que levou os reguladores a emitir normas[61] que obrigam uma revisão completa dos dados e o reporte de riscos para garantir sua qualidade, integridade, rastreabilidade e consistência, conhecida como «RDA&RRF[62]». O objetivo é reforçar as capacidades de agregação de dados de riscos e as práticas de apresentação de relatórios, para assim melhorar a gestão e a tomada de decisões. Por outro lado, foram unificados os critérios de reporte de capital, liquidez e informação financeira (COREP, FINREP). Tais fatos obrigam as instituições a realizarem uma revisão profunda dos sistemas e processos de geração de informação e reportes.

Enfrentar este «tsunami regulatório» representa um enorme custo para as instituições e as obriga a desenvolver ambiciosos processos de transformação. Não obstante, esta transformação é um elemento claramente diferencial das instituições, porque permite oferecer aos clientes a segurança de dispor dos processos mais seguros, regulados e supervisionados de todos os setores digitais. É um aspecto-chave que as instituições acabarão por valorizar face aos demais novos concorrentes.

Comportamento do cliente

Em um prazo de poucos anos, o setor financeiro assistiu à transformação do comportamento de seus clientes: estão mais informados, mais conectados, têm mais cultura financeira e mostram uma demanda mais específica (Fig. 16). Exigem um serviço que proporcione comodidade, velocidade, personalização e tratamento justo, além de acesso a partir de dispositivos móveis.

O cliente é caracterizado por suas elevadas expectativas; compara a qualidade de serviço proporcionada pela instituição financeira com a de fornecedores de outros setores (tecnológico, varejistas, etc.) e espera um nível similar de funcionalidades e de resposta em tempo real.

Também é competente e ativo no uso das redes sociais, que utiliza tanto para comparar informação (81% consulta as redes sociais antes de comprar) como para difundir sua inconformidade perante uma experiência deficiente.

Os estudos[63] demonstram que a experiência do cliente está positivamente correlacionada com a retenção. Apesar disso, revelam que, embora se observe uma melhoria gradual da qualidade da experiência do cliente com as instituições financeiras, ainda não é suficiente: mais de 50% dos clientes manifestam a intenção de mudar de banco antes de seis meses.

Mais ainda, o cliente já não contempla os bancos como os únicos fornecedores de serviços financeiros nem as agências como o canal básico de relacionamento (salvo para assessoria). Tais fatos estão obrigando as instituições financeiras a uma reformulação total da sua oferta de serviços e dos seus canais e a adotar, em suma, uma visão centrada no cliente ou de «360°», que tem impacto em todos os âmbitos, desde os processos aos sistemas, passando pela organização, pelo controle de riscos ou pelo planejamento de negócio.

No entanto, também constatamos nos clientes uma crescente confusão perante a complexidade e a diversidade da oferta. Isso está levando as instituições a adotarem uma visão comercial orientada para a simplificação da oferta, adequando-a às necessidades do cliente (revendo, assim, seu catálogo de produtos e serviços).

De forma simultânea e muito relacionada à mudança do perfil do cliente, observamos um efeito transformador e novo: a

[61] BCBS 239 (2013).
[62] Risk Data Aggregation and Risk Reporting Framework.
[63] EFMA (2013).

entrada de novos concorrentes no setor financeiro, alguns provenientes de outros setores (Fig. 17), que satisfazem necessidades não totalmente cobertas pelos bancos tradicionais.

Esta nova concorrência pode ser classificada em três famílias:

- Concorrência conhecida que oferece novos serviços (tais como bancos 100% móveis).
- Novos players financeiros, que antes não existiam no mercado e que cobrem nichos não atendidos.
- Novos modelos de negócio, que provêm de outros setores; principalmente, tecnologia, venda varejista e telecomunicações.

No caso da concorrência proveniente de outros setores, esta ameaça é particularmente lesiva por várias razões: por um lado, os novos concorrentes não estão sujeitos à estrita regulação bancária; por outro, têm modelos de negócio bastante eficientes, com custos muito reduzidos; em terceiro lugar, dispõem de «ecossistemas» que agrupam muitas necessidades do cliente: dispositivos físicos, contexto de trabalho, música, filmes, livros, revistas, etc., onde é natural integrar os serviços financeiros; e, por último, têm imagens de marca que são percebidas de forma muito positiva pelos clientes.

O setor bancário ainda vê estes concorrentes como uma ameaça moderada (Fig. 18) e é certo que, por enquanto, estão se concentrando em determinados nichos, tais como os meios de pagamento (por exemplo, PayPal, Google Wallet ou Apple Pay), e que as barreiras de entrada regulatórias são elevadas para acessar os serviços core de depósito e crédito. No entanto, pelo tamanho e influência, estes concorrentes têm o potencial de superar as barreiras e alterar o mercado de forma significativa em um futuro próximo.

Fig. 17. Nova concorrência no setor financeiro.

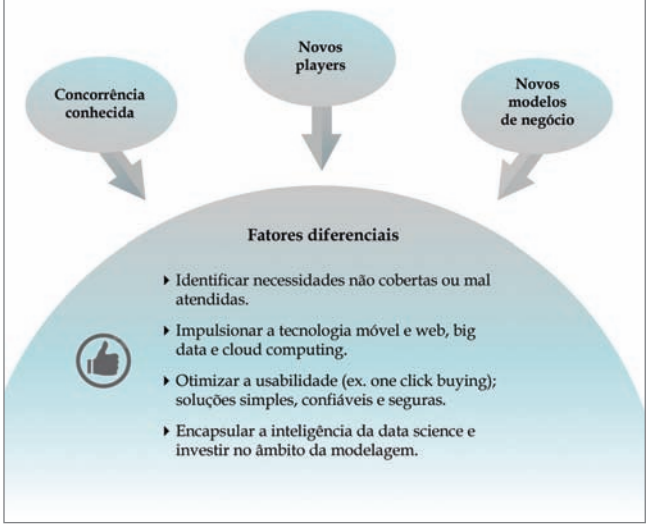
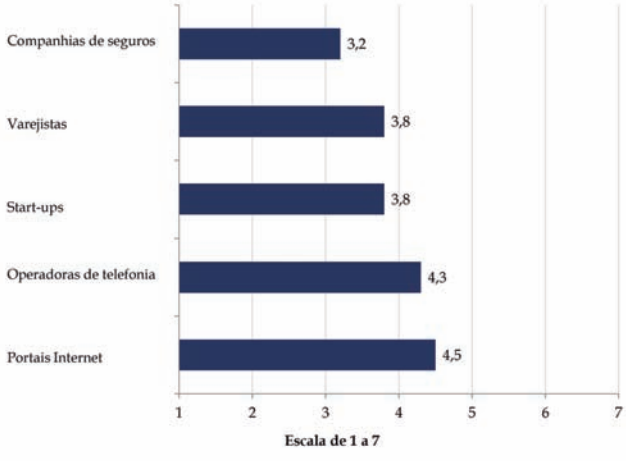


Fig. 18. Percepção de ameaças por parte das instituições financeiras.



Fonte: EFMA (2013).

Data science: uma disciplina emergente

Todas as empresas têm o big data em seu futuro e todas as empresas estarão no negócio dos dados mais cedo ou mais tarde.

Thomas H. Davenport[64]

O que é data science?

A comoditização dos dados e a governança dos dados e dos modelos que se impõe como consequência implica, como acontece com qualquer matéria-prima, a aparição de novas ferramentas e técnicas para seu processamento. O conjunto destas ferramentas e técnicas forma uma disciplina que, embora não seja nova, tem um caráter emergente e está recebendo uma atenção crescente em todos os setores, incluindo o financeiro: data science.

Definição

A natureza de novidade do interesse por esta disciplina, em conjunto com o seu caráter inovador e ligado às tecnologias de big data, faz com que não exista uma definição formal e geralmente aceita de data science. O Center for Data Science da New York University aborda o termo da seguinte forma[65]:

> *Data science, ou a ciência dos dados, é o estudo da extração generalizável de conhecimento a partir dos dados pelo uso combinado de técnicas de aprendizagem automática, inteligência artificial, matemática, estatística, bancos de dados e otimização, em conjunto com uma compreensão profunda do contexto de negócio.*

No entanto, a maior parte das aproximações ao conceito de data science descrevem mais as competências e conhecimentos que o profissional necessita para ser considerado um data scientist:

> *Um profissional com formação e curiosidade para tomar decisões no mundo do big data. […] Realiza descobertas, estando completamente submerso em dados e é capaz de estruturar grandes quantidades de dados sem formato, assim como identificar fontes de informação ricas, embora incompletas, e combiná-las para obter conjuntos de dados muito mais completos e de grande valor.*[66]

> *Um profissional com um conhecimento profundo de dados que pode trabalhar de forma eficaz com dados de uma maneira escalável.*[67]

> *Uma evolução do analista de negócios: tem uma base sólida em computação e suas aplicações, matemática, modelagem, estatística e análise de dados. No entanto, o data scientist se destaca pela acuidade do seu sentido de negócio do setor e sua capacidade de comunicação.*[68]

Como pode-se observar, o data scientist é um profissional com um perfil multidisciplinar e, especificamente, combina pelo menos três características:

- Uma formação de base em alguma ciência ou disciplina quantitativa, que inclua conhecimentos de aprendizagem automática, algoritmia, otimização, simulação, séries temporais e modelos de associação e classificação, entre outros.

- Competências tecnológicas avançadas, que incluem o domínio de linguagens de programação estatística, mas também conhecimentos técnicos para a extração eficiente, o uso e o armazenamento de informação, a manipulação de bancos de dados relacionais e não relacionais e a capacidade para extrair dados da Internet e processar grandes quantidades de informação.

[64]Thomas Hayes Davenport (n. 1954). Acadêmico norte-americano especialista em gestão do conhecimento e inovação de processos de negócio. Foi nomeado um dos três melhores analistas de negócio do mundo em 2005 pela revista Optimize.
[65]Dhar [Center for Data Science, New York University] (2013).
[66]Harvard Business Review (2012).
[67]Berkeley (2015).
[68]IBM (2014b).

- E, o que possivelmente estabeleça uma diferença maior em relação a outros perfis similares, um conhecimento profundo do negócio em que desenvolvem seu trabalho como data scientists.

O terceiro aspecto, a experiência no negócio, é particularmente relevante porque aproxima de forma definitiva as capacidades analíticas, no caso do setor bancário, ao conhecimento financeiro; isto é essencial para que os modelos se integrem de forma plena na gestão, o que é uma condição indispensável para o sucesso e uso adequado (Fig. 19).

Para avaliar a importância deste perfil, destacamos que, nos Estados Unidos, o presidente Barack Obama criou, em fevereiro de 2015, o cargo de Chief Data Scientist, e nomeou pessoalmente Dhanurjay "DJ" Patil[69] para esta função, com a missão de impulsionar novas aplicações do big data em todas as áreas do Governo[70].

O processo data science

Por outro lado, como indicam alguns autores[71], a característica mais relevante da data science é precisamente sua qualidade de ciência: perante a quantidade massiva de dados que enfrenta uma instituição, a abordagem de um data scientist é formular uma teoria (uma pergunta ou uma hipótese) proveniente da realidade do negócio, e aplicar seus conhecimentos e competências nos dados para verificá-la ou descartá-la. Isso é o que é conhecido como o «processo data science», que é composto por cinco etapas (Fig. 20):

1. **Formulação:** é apresentada uma pergunta relevante para o negócio, que deverá ser respondida através dos dados e das técnicas disponíveis. Uma das alterações essenciais que aporta a disciplina de data science é precisamente a formulação de questões ou hipóteses que antes eram impossíveis de serem verificadas e que, com a abundância atual de dados, ferramentas e técnicas, abre novas possibilidades. Por exemplo, «a julgar pelos seus comentários nas últimas ligações para o call center, qual é a probabilidade de que cada um dos meus clientes mude de banco nos próximos seis meses, e o que posso fazer para evitá-lo?».

2. **Obtenção de dados:** são localizadas todas as fontes disponíveis, incluindo fontes estruturadas (datawarehouses, datamarts, etc.) e não estruturadas (logs de atividade, redes sociais , etc.). A quantidade massiva de dados e, se for o caso, sua natureza desestruturada é o núcleo do desafio computacional de todo o processo. Nesta fase, também são abordados os aspectos legais, tais como a proteção de dados, a confidencialidade ou as cláusulas de restrição de uso.

3. **Exploração de dados:** são aplicadas técnicas de estatística descritiva para realizar uma primeira análise exploratória. Neste ponto, a data science traz novas técnicas de exploração que facilitam o trabalho e, dado o potencial de paralelização nestas tarefas, aproveita as plataformas de computação distribuída.

4. **Modelagem:** a construção e a validação tradicional dos modelos são enriquecidas por algoritmos de alto rendimento desenvolvidos ad hoc para grandes volumes de informação, assim como por tipos de modelos alternativos aos clássicos, que trazem melhorias em termos de estabilidade, robustez e aproveitamento da riqueza da informação, como os random forests e as support vector machines, entre outros (Fig. 21). Para isso, os fornecedores tradicionais de ferramentas analíticas estão completando suas suítes de produtos, e surgem novas linguagens de programação estatística, muitas delas em código aberto.

[69]Data scientist de grande reputação que trabalhou na LinkedIn, na eBay, na PayPal e na Skype, entre outras, e a quem se atribui a criação do termo «data scientist».
[70]Wired (2015).
[71]O'Neil e Schutt (2013).

Fig. 19. Características do data scientist.

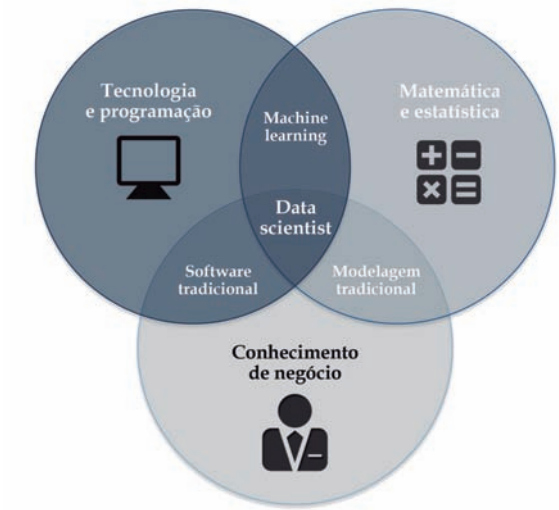
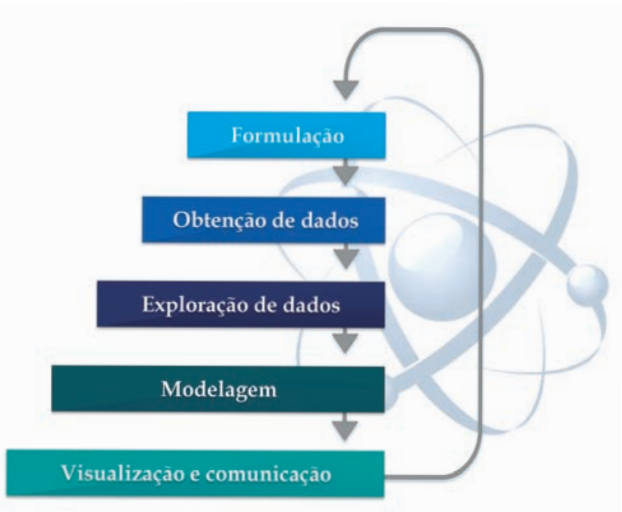
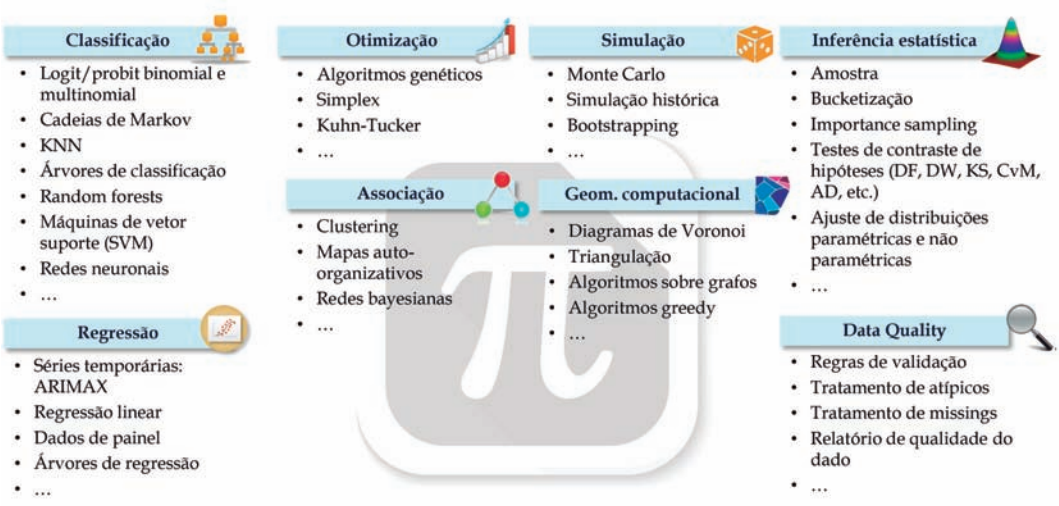


Fig. 20. O processo data science.



Fonte: adaptado de Harvard (2014).

Fig. 21. Técnicas de data science.



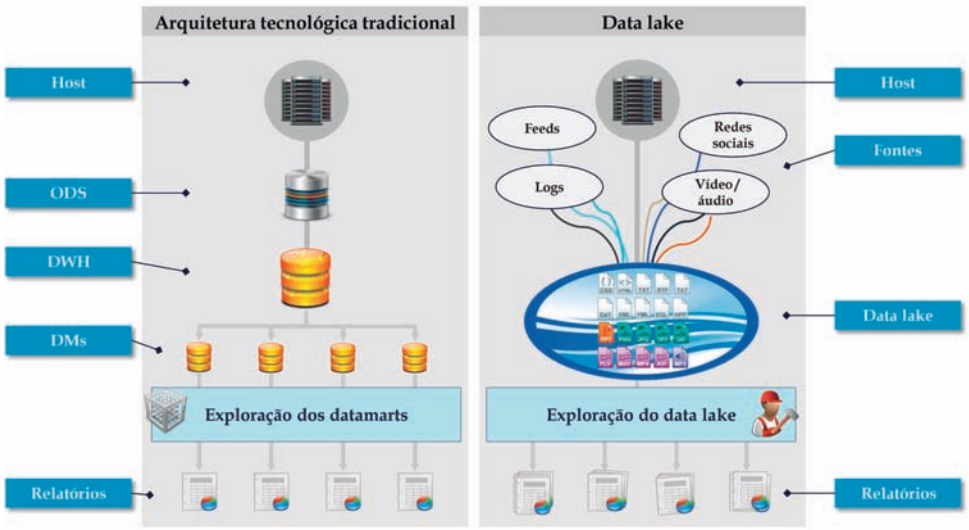
Data lake: uma nova arquitetura informacional

Perante o volume crescente de informação e a natureza heterogênea das fontes, é necessário contar com novas técnicas e tecnologias capazes de armazenar a informação de forma otimizada. Neste cenário, surge o conceito de «data lake» ou «data pool», um imenso repositório de dados no seu formato de origem, com as seguintes características:

- ***Integridade.*** É um único repositório onde se armazena toda a informação, assegurando a rastreabilidade da informação.
- ***Flexibilidade.*** O data lake armazena qualquer informação de interesse, independentemente do formato, eliminando qualquer padrão específico sobre captura e armazenamento de dados.
- ***Independência.*** Elimina, em grande parte, a dependência dos departamentos de Tecnologia da Informação, dado que a carga de informação é flexível e o usuário pode realizar as extrações necessárias diretamente a partir do data lake.

O uso e a exploração de data lakes e as técnicas de data science associadas permitem aos data scientists trabalharem com a informação sem manipulação, em seu formato original; permite escalabilidade horizontal e elimina os limites em manipulação de grandes volumes de informação sem estruturação. Esta arquitetura não substitui necessariamente a tradicional. Pelo contrário, em geral, é complementar aos datawarehouses e datamarts.

Data lake versus a arquitetura tradicional.



5. **Visualização e comunicação:** dois dos aspectos que mais receberam atenção pela data science, a visualização dos resultados e sua comunicação de forma inteligível a terceiros, são duas qualidades que se esperam de um data scientist, e também são potencializadas por novas ferramentas que integram o código com a documentação de forma intuitiva e natural.

Embora possa parecer um enfoque de sentido comum, esta aproximação aos dados por meio de um método científico implica uma mudança de metodologia de trabalho: os analistas abordam com frequência o problema de forma inversa (lançar modelos ao acaso sobre uma quantidade massiva de dados em busca de relações ocultas), o que pode representar um consumo elevado de recursos sem um objetivo claramente estabelecido nem uma hipótese a verificar.

Em suma, a data science representa a evolução da modelagem tradicional no contexto big data e abre novas possibilidades antes impensáveis, que chegam a transformar os modelos de negócio estabelecidos. A adoção da data science como um pilar estratégico de desenvolvimento é uma prioridade para o setor tecnológico e, como veremos, também começa a ser uma prioridade para o setor financeiro.

Ferramentas de data science

A comoditização dos dados também está favorecendo a evolução e a aparição de novas ferramentas tecnológicas de data science que facilitam seu processamento, análise e visualização. Todos os fornecedores tradicionais estão impulsionando ecossistemas analíticos e, continuamente, aparecem start-ups com novas propostas, assim como ferramentas e linguagens open source (Fig. 22), o que converte este mercado em um foco de concorrência e desenvolvimento acelerado.

Estas ferramentas permitem superar as limitações dos sistemas tradicionais, que eram insuficientes face à heterogeneidade dos dados (não podiam analisar informação estruturada e não estruturada conjuntamente), sua desfragmentação (a informação estava distribuída em silos diferentes sob modelagens imprecisas), a dependência da Tecnologia da Informação (os usuários de negócio tinham que delegar a áreas de Sistemas a tarefa de compilar e organizar a informação em datawarehouses, implicando um tempo excessivo de preparação de dados) e, em geral, a falta de adaptação às fontes de dados atuais (os sistemas tradicionais não se integravam com redes sociais, call centers, sensores, posicionamento geográfico, etc., nem eram adequados para lidar com o volume de informação criada por eles).

Assim, entre as principais contribuições que estas ferramentas incorporam, vale destacar[72]:

- Self-service: no esquema tradicional, apenas alguns profissionais da instituição, muito especializados, tinham acesso aos dados e às ferramentas analíticas. No esquema de data science, aparecem ferramentas de maior simplicidade, que permitem a mais profissionais explorar, analisar e visualizar dados. Este fenômeno ocorre em todos os setores e contribui para potencializar as capacidades analíticas dos profissionais.

- Fusão de dados: a fusão de dados está relacionada à combinação de informação proveniente de diversas fontes e em formatos diferentes. No esquema tradicional, isso ocorre por meio de processos de ETL e a implementação de modelos de dados que podem se tornar muito complexos. No esquema de data science mais avançado, os dados são carregados em um data lake comum, bem documentado com um dicionário de dados, e as ferramentas são capazes de capturar os arquivos e uni-los em pouco tempo.

[72] Adaptado de Gigaom Research (2014).

Fig. 22. Quem é quem em data science? Algumas das principais ferramentas e linguagens open source.

hadoop	**Hadoop:** infraestrutura de programação de código aberto que permite armazenar, processar e analisar grandes volumes de dados, disseminando-os em grandes clusters de servidores e que se processam em paralelo.
HIVE	**Hive:** sistema de datawarehouse sobre Hadoop, desenvolvido pelo Facebook. É utilizado para lançar queries e manipular grandes quantidades de dados em um armazenamento distribuído. Utiliza uma linguagem similar a SQL denominado HiveQL (HQL), o que evita aos usuários a necessidade de usar Java ou APIs do Hadoop.
	Pig: aplicativo open source criado pelo Yahoo e construído sobre o Hadoop, focado no processamento de grandes volumes de dados (estruturados e semiestruturados) em modo batch.
	Impala: plataforma open source que proporciona consultas SQL (processamento e análise de dados) em tempo real. Utiliza os componentes do Hadoop.
Lucene	**Lucene:** API open source que serve como motor de pesquisa. Utiliza-se para indexar e pesquisar dados sobre datasets delimitados. Foi implementado nas web apps do Twitter, LinkedIn, Apple, AOL, Eclipse, etc.
mahout	**Mahout:** projeto open source para implementar algoritmos escaláveis e distribuídos de machine learning. É um framework de Java e utiliza Hadoop. Utilizados por Adobe, AOL, Intel, LinkedIn, Twitter e Yahoo, entre outros.
R	**R:** Linguagem e ambiente de programação estatístico muito versátil devido à sua modularidade (pacotes de funcionalidades avançadas já criadas por outros programadores podem ser instalados) e à sua natureza open source.
python	**Python:** linguagem de programação de propósito geral que foca a legibilidade e a intuição do código. Em data science, é especialmente indicado para a captura de dados de fontes on-line.

Novas capacidades tecnológicas

O fenômeno big data proporciona novas capacidades tecnológicas, que são estruturadas em quatro camadas:

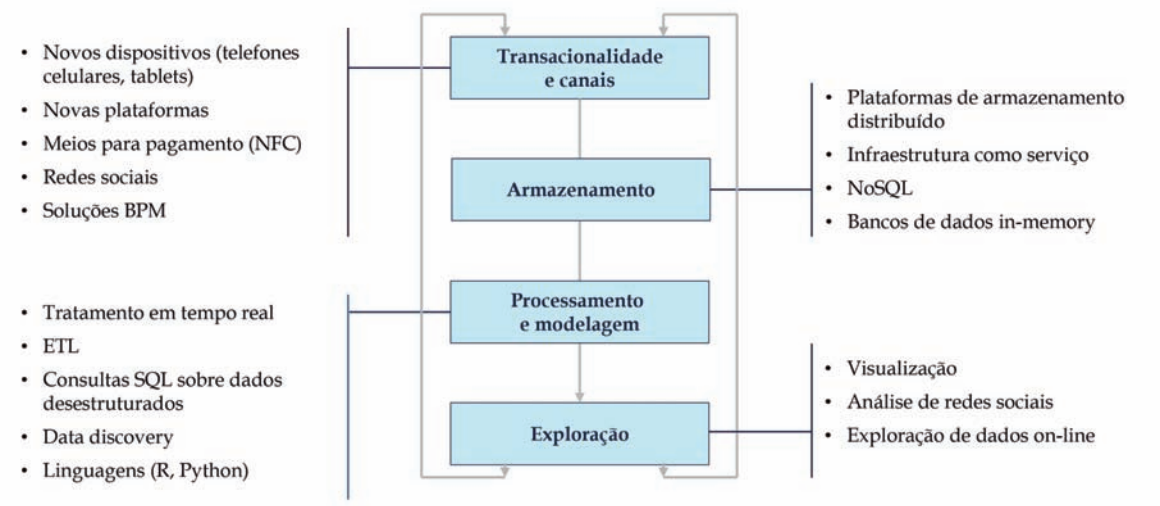
- Na camada **transacional**, aparecem novos dispositivos e canais de relacionamento com os clientes (telefone celular, tablet, etc.), novas plataformas para o desenvolvimento de aplicativos e a prestação de serviços (CRM, apps para telefones celulares para cobrir novas necessidades financeiras e operacionais do cliente, etc.), novas tecnologias de meios de pagamento, como o pagamento por telefone celular (como NFC) e soluções BPM (Business Process Management) para a integração de plataformas e a automação de processos, como a contratação on-line ou a gestão de documentos. Além disso, as redes sociais são potencializadas como um novo canal de relacionamento com o cliente. As redes sociais se apresentam como potencial canal de contratação, por meio das quais são realizadas análises de sentimento de marca e atendimento de queixas e reclamações.

- Na camada de **armazenamento**, aparecem novos sistemas de armazenamento concebidos para serem executados em hardware de baixo custo, oferecendo alta disponibilidade, tolerância a falhas (com dados replicados em vários nós), com escalabilidade horizontal e que permitem o processamento massivo de dados. Surge a infraestrutura como serviço nas modalidades de cloud público, cloud privado ou cloud híbrido. Aparecem novos bancos de dados (NoSQL) orientados ao processamento em batch de grandes volumes de informação e novas estruturas de dados: bancos de dados colunares e documentais e também novos bancos de dados in-memory para o processamento de informação em memória com uma alta velocidade de resposta a consultas.

- Na camada de **processamento e modelagem**, surgem ferramentas para a captura e processamento de informação em tempo real, assim como novas ETL para transformação de dados desestruturados, como o Pig, e novos motores de consulta de dados desestruturados em linguagem SQL. Também aparecem ferramentas para a implementação de mecanismos que garantam a governança dos dados: catalogação, transformação, rastreabilidade, qualidade, consistência e controle de acesso, e ferramentas de data discovery para a extração de conhecimento de forma livre a partir de fontes diversas, estruturadas e não estruturadas. E, por último, aparecem novas técnicas, algoritmos matemáticos e linguagens para o reconhecimento de padrões nos dados, a análise preditiva, a implementação dos modelos e a aprendizagem automática (machine learning).

- Por último, na camada de **exploração** aparecem novas ferramentas de análise multidimensional e reporting com capacidade de acesso a grandes volumes de informação em memória, soluções específicas para a análise da informação proveniente de redes sociais e para a exploração de fluxos de dados on-line para a tomada de decisões e desencadeamento de eventos em tempo real, como a detecção de fraude, a detecção e o lançamento de eventos comerciais ou os scorings de riscos, entre muitos outros usos.

Novas capacidades por camada de arquitetura tecnológica.



- Conectividade não relacional: ao contrário das ferramentas tradicionais, que previam apenas a conexão com bancos de dados, as ferramentas de data science permitem a conexão com outras fontes de informação: NoSQL, plataformas de computação distribuída, e informação de redes sociais, na nuvem ou em sistemas de software as a service, de importância crescente para as instituições.

- A nuvem: uma das novidades mais relevantes é a utilização da nuvem no âmbito analítico, que proporciona a funcionalidade de armazenamento de dados na versão mais básica, permitindo desligar o trabalho do data scientist de uma localização ou um servidor específicos, facilitando o trabalho a partir de diferentes localidades geográficas. Em alguns casos, integra também os serviços de ETL, visualização de dados e implementação em dispositivos móveis, criando um ecossistema analítico completo, que simplifica o trabalho de análise.

- Visualização de dados: uma das características diferenciadoras da disciplina de data science, ligada ao conhecimento de negócio, é a visualização dos dados. Algumas ferramentas vão muito além da criação de gráficos com anotações e já são capazes de produzir, de forma automática, dashboards e apresentações interativas que permitem ao usuário aprofundar-se nas análises de forma dinâmica.

Data science no setor financeiro

O setor financeiro está registrando a mesma explosão em termos de geração e necessidade de armazenamento de dados que outros setores, e tem o potencial de extrair um conhecimento profundo tanto de seus clientes como de seu contexto (concorrência, atividade econômica setorial, geolocalização, etc.) cujo acesso era impossível anteriormente.

Para isso, as instituições estão desenvolvendo uma série de capacidades tecnológicas e metodológicas, que estão abrindo um horizonte de novas possibilidades no setor, embora apresentem uma série de desafios.

Novas oportunidades no setor financeiro

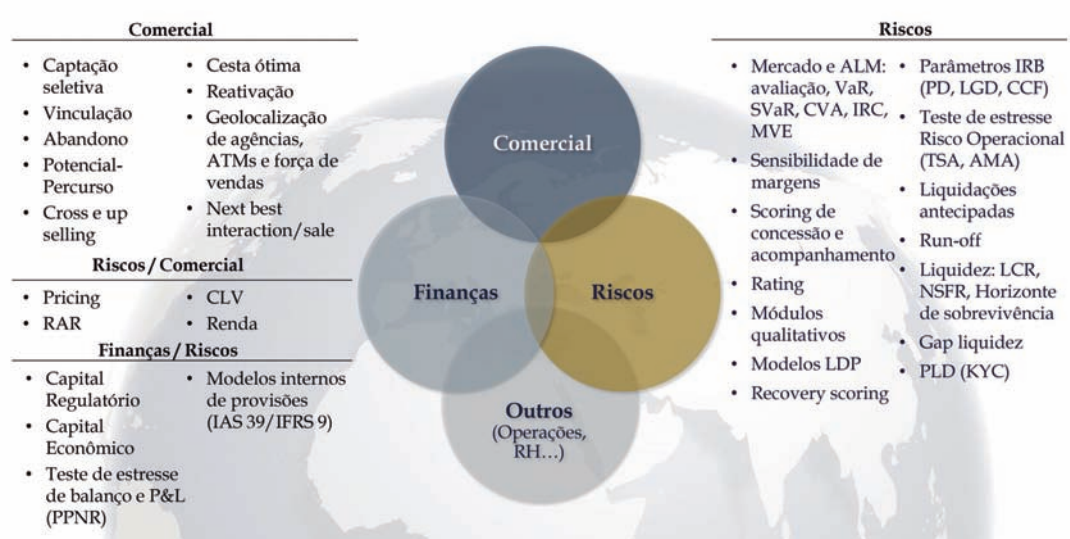
Graças a estas novas capacidades, os modelos tradicionais são ampliados e enriquecidos em todos os âmbitos das instituições financeiras: riscos, marketing, finanças, operações, etc. (Fig. 23), contribuindo para o aproveitamento de toda a informação disponível para melhorar a tomada de decisões nesses âmbitos e, em alguns casos, automatizar alguns dos processos.

Assim, como exemplo, podemos citar algumas aplicações emergentes da data science no setor, baseadas em dados de redes sociais, geolocalização, multimídia ou logs, entre outros, que antes não recebiam atenção:

- Credit scoring com digital footprint: a classificação de crédito de pessoas físicas, geralmente baseada em poucas variáveis (entre 5 e 20, dependendo da carteira e da informação disponível), é enriquecida e ampliada com a informação presente nas redes sociais e na Internet em geral, o que se conhece como o «digital footprint» ou «impressão digital». São construídos modelos baseados nesta informação, que melhoram substancialmente o poder de previsão e, portanto, o controle da inadimplência, especialmente na população de não clientes, tradicionalmente classificada de forma inadequada pelas limitações na disponibilidade de dados.

- Prevenção do abandono pelo processamento da linguagem natural (NLP): as gravações dos call centers, que costumavam utilizar-se quase exclusivamente para o controle interno de qualidade, são uma fonte valiosa de prevenção da fuga de clientes. Por meio de modelos de reconhecimento da fala, são transcritas de forma automática todas as conversações com os clientes e, sobre os textos resultantes, são aplicadas técnicas de text mining

Fig. 23. Principais modelos nas instituições financeiras.



e linguística computacional para identificar a probabilidade de que um cliente em particular decida mudar de instituição nas próximas semanas. Para isso, primeiramente, é realizada uma análise lexicográfica (a detecção de certas palavras que se associam à intenção de mudança), mas de forma experimental também se avança para um nível semântico, onde o modelo inclui padrões mais complexos de significado no discurso do cliente.

- Modelos de renda e propensão baseados em redes sociais cruzados com geolocalização: a informação disponível de um cliente nas redes sociais é cruzada com dados de censos, imobiliários, do Google Maps e de outras fontes e, com isso, são realizadas estimativas melhoradas de seu nível de renda, sua capacidade de poupança, suas necessidades de produtos financeiros, o valor do imóvel onde reside (utilizado também para avaliar o subjacente em uma securitização), etc. Tais dados complementam a informação disponível e contribuem para a melhoria das ações comerciais sobre os clientes.

- Personalização de promoções para reduzir os custos de aquisição: é feita a compilação e cruzamento de toda a informação disponível sobre cada cliente em suas transações por todos os canais e dados de redes sociais, obtendo, assim, uma visão 360° do cliente. Com isso, reduz-se ao máximo o nicho objetivo de cada promoção e, consequentemente, aumenta-se a proporção de captação de clientes, além de diminuir os custos de aquisição.

- Campanhas de bonificação por meio de análises de transações de cartões: partindo dos movimentos dos cartões, é possível conhecer os costumes dos titulares, os momentos em que realizam compras, suas viagens, lojas habituais, etc. e propor campanhas de bonificação no momento adequado, com maior probabilidade de sucesso.

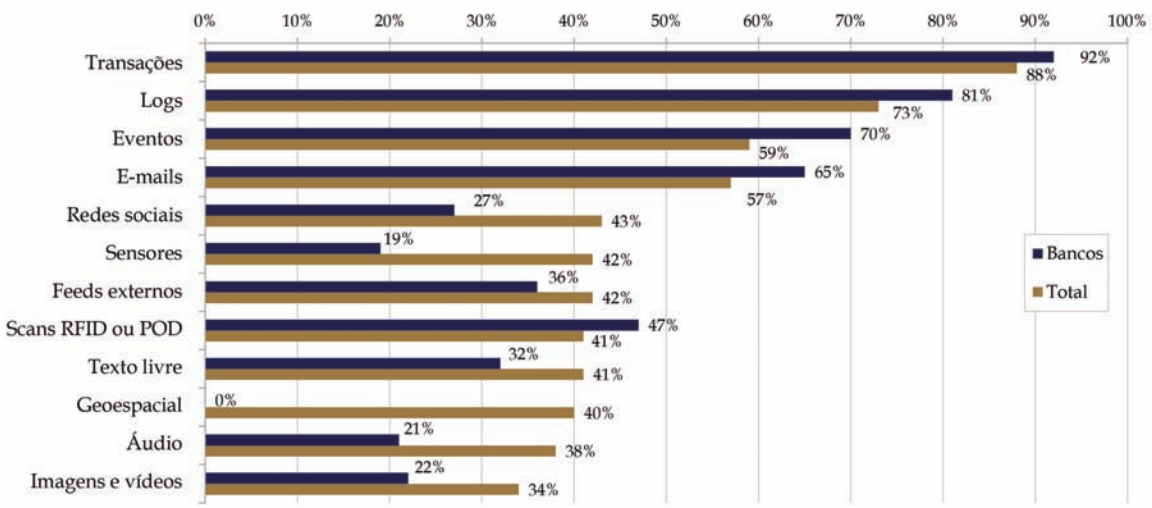
- Detecção de fraude e de lavagem de dinheiro e melhoria do controle de qualidade por meio de logs: os registros de atividade ou logs são grandes arquivos pouco estruturados onde constam todas as ações que um cliente ou um empregado realiza em uma plataforma digital (computador, dispositivo móvel, caixa automático, etc.). A detecção de padrões de comportamento nos logs é complexa, porque requer um processamento de informação especialmente massivo, mas pode servir para identificar tentativas de fraude (tanto internas como externas) e de lavagem de dinheiro. Igualmente, é a base de uma modalidade nova de controle de qualidade, potencialmente muito extensa, que pode abarcar desde os tempos de resposta em uma agência até a dificuldade para utilizar um novo aplicativo, passando pela preferência dos clientes por um ou outro canal para cada tipo de transação, entre muitos outros.

Estes são apenas alguns exemplos; as oportunidades são tantas como as perguntas passíveis de formulação, considerando a proliferação de fontes de dados nas instituições financeiras e as crescentes capacidades em data science (talento e ferramentas) que já estão sendo contempladas.

Nesse sentido, a automação de processos e a melhoria dos modelos utilizados no setor financeiro estão estreitamente ligadas à capacidade das instituições de capturar informação relevante de seus clientes, processos, produtos, etc. e, posteriormente, utilizando ferramentas de data science, proceder ao seu armazenamento, processamento e exploração.

Atualmente, as fontes de informação disponíveis para serem exploradas pelas instituições são praticamente ilimitadas; isso evidencia que qualquer informação, independentemente da procedência (interna ou externa) e seu caráter (estruturado ou desestruturado), é potencialmente relevante para a tomada de decisões. Mais ainda, estima-se que a maioria dos dados massivos nos bancos provém das transações, dos logs, dos eventos e dos e-mails, e menos de outras fontes (Fig. 24).

Fig. 24. Fontes de big data no setor bancário e em outros setores.



Fonte: IBM & University of Oxford (2015).

Desafios perante a adoção de data science

Como fica evidente, as possibilidades abertas pela disciplina de data science no setor financeiro são numerosas e têm o potencial de melhorar substancialmente as métricas de desempenho em todos os âmbitos: a experiência do cliente, a eficiência, o controle do risco, a eficácia das ações comerciais, etc.

No entanto, a evolução para estas capacidades não é simples nem imediata; as instituições enfrentam uma série de desafios (Fig. 25), que foram relacionados aos desafios do big data em pesquisas realizadas no setor (Fig. 26), entre os quais destacamos:

- Volume e compartilhamento de dados: as mesmas quantidades massivas de dados que possibilitam a existência de data science representam um desafio em termos de dimensionamento dos bancos de dados, da arquitetura de armazenamento, do custo da capacidade de processamento, do tempo de computação e da necessidade de algoritmos otimizados, o que é solucionado em parte com as novas ferramentas e plataformas, mas que requer um cuidadoso planejamento tecnológico. Além disso, o compartilhamento dos dados entre as áreas das instituições representa um desafio tecnológico e organizacional complexo para resolver.

- Restrições orçamentárias: vinculada ao item anterior, a evolução para a data science tem uma necessidade associada de investimento em infraestrutura e em talento que, no contexto de margens pressionadas e abundante regulação, é necessário articular com as restrições orçamentárias.

- Privacidade, ética e segurança: a utilização de informação abundante sobre os clientes levanta questões sobre a privacidade dos dados e a ética do seu uso. O serviço é muito mais personalizado quanto mais informação for utilizada, dado que a regulação ainda é incipiente sobre muitas das nuances da privacidade, e é complicado que avance na mesma velocidade do fenômeno big data. Para encontrar o equilíbrio entre privacidade e experiência de cliente é necessária a intervenção das áreas de assessoria jurídica e escutar o próprio cliente. Neste ponto, também englobam-se os aspectos de segurança da informação e da garantia de que os dados compilados de diversas fontes estejam a salvo de hackers, roubos de identidade e venda indevida a terceiros.

- Captação e formação do talento: os data scientists constituem ainda um perfil relativamente escasso e a demanda destes profissionais supera largamente a oferta. Por isso, um dos principais desafios é a captação de data scientists no mercado e a formação dos metodologistas tradicionais para adquirir novas capacidades. Alguns

Fig. 25. Desafios na evolução para a data science nas instituições financeiras.

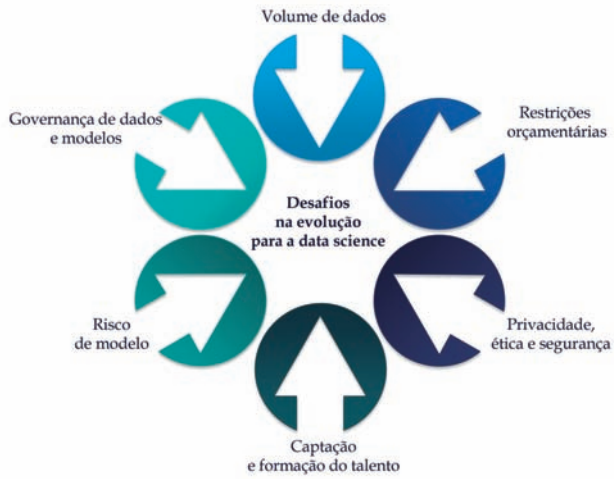
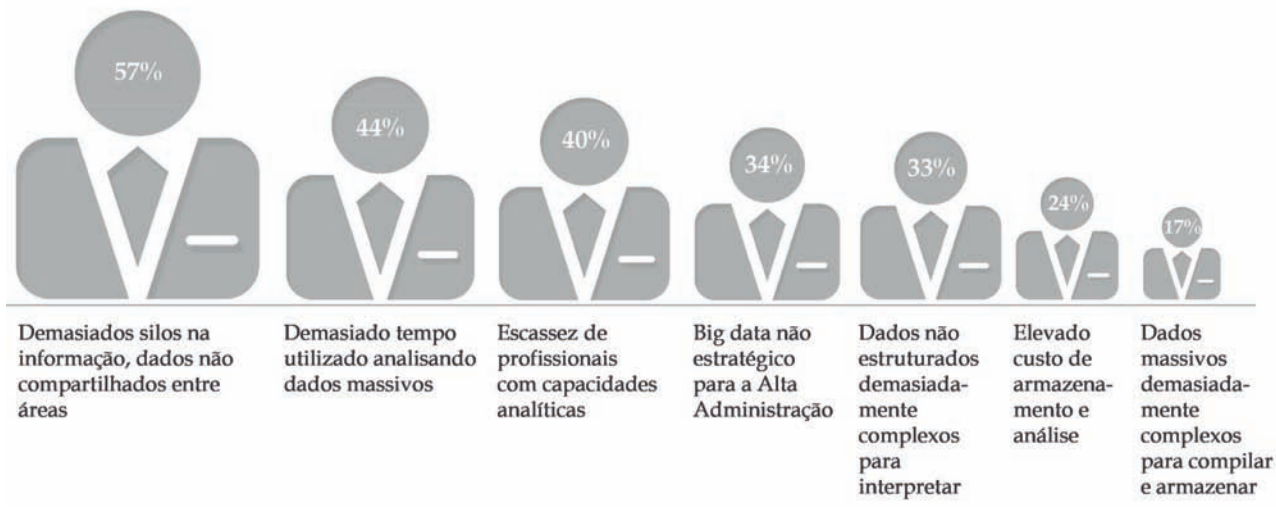


Fig. 26. Resultados da pesquisa sobre «maiores obstáculos nos bancos para o sucesso em big data» (porcentagens sobre o total de respostas).



Fonte: adaptado da Inetco (2015).

estudos[73] apontam para uma escassez de mais de 140.000 data scientists em 2018 só nos Estados Unidos.

- Risco do modelo: a utilização de modelos implica riscos, que emanam dos dados que utilizam, mas também da própria estimativa do modelo e de seu potencial uso indevido. A gestão e o controle do risco de modelo são um foco de atenção crescente perante o aparecimento da data science[74].

- Governança dos dados e dos modelos: por último, a organização e a governança das estruturas internas necessárias para gerenciar adequadamente os dados e os modelos de uma instituição financeira são um elemento-chave para o sucesso da evolução para a data science na organização, conforme indicado a seguir.

Impactos na governança de dados e modelos

Embora a aceleração na geração e no acesso aos dados e as possibilidades que isso oferece constituam uma certa novidade no setor financeiro, a realidade é que nem as instituições financeiras nem os reguladores e supervisores ficam alheios ao fenômeno descrito.

Pelo contrário, as instituições vêm realizando transformações sobre seus sistemas e processos de geração de informação e de reporte, especialmente nos âmbitos de riscos, financeiro e comercial. No entanto, em muitos casos, estas transformações foram realizadas de forma pouco estruturada e com uma perspectiva limitada, como consequência de pedidos adicionais dos reguladores e supervisores, de necessidades de gestão não planejadas com visão global e, em muitos casos, das migrações de dados ocasionadas por fusões e aquisições. Isso fez com que os processos de geração de informação e reporte tenham perdido efetividade e, às vezes, a consistência dos dados não é garantida.

Mais ainda, os reguladores indicaram as carências nos dados como uma das causas da crise financeira iniciada em 2007:

> *Uma das principais lições da crise financeira mundial iniciada em 2007 foi que a inadequação das tecnologias da informação (TI) e das arquiteturas de dados dos bancos impediram a realização de uma gestão integrada dos riscos financeiros. […] Em alguns bancos, a incapacidade para gerir adequadamente os riscos respondia a carências na agregação de dados sobre riscos e nas práticas de apresentação dos relatórios correspondentes. Isso teve consequências graves para os próprios bancos e para a estabilidade do sistema financeiro no conjunto.*[75]

Por outro lado, a utilização de modelos para a tomada de decisões é um fenômeno que também está se proliferando com grande rapidez, o que traz indubitáveis benefícios, como a melhoria da objetividade, a automação e a eficiência. No entanto, seu uso também implica um «risco de modelo», entendido como os potenciais prejuízos (econômicos, de reputação, etc.) provocados por decisões baseadas em modelos errados ou utilizados de forma inapropriada[76].

Tanto a regulação como os aspectos de gestão derivados da abundância de informação e sua utilização em modelos para a tomada de decisões implicam a necessidade de estabelecer um novo framework para governar os dados e os modelos de forma apropriada em cada instituição financeira. Nessa seção, serão analisadas as práticas no setor em relação a estes frameworks de governança.

Governança dos dados

O estabelecimento de mecanismos de governança dos dados é uma necessidade estratégica complexa nas instituições, que torna-se especialmente urgente na linha do fenômeno big data e é uma condição imprescindível para a obtenção do máximo aproveitamento da informação.

[73] McKinsey (2011).
[74] Management Solutions (2014).
[75] BCBS (2013).
[76] Management Solutions (2014).

Este aspecto não passou despercebido dos supervisores que, em alguns âmbitos, emitiram normas específicas relativas à governança dos dados e da informação, destacando especialmente, no que se refere a riscos, os *Princípios para uma agregação eficaz de dados sobre riscos e apresentação de relatórios de riscos* (BCBS, 2013), conhecidos como «RDA&RRF»[77], com requerimentos em qualidade, consistência, integridade, rastreabilidade e replicabilidade dos dados. Essa norma é vinculante para as instituições globalmente sistêmicas e suas filiais e, no futuro, para as localmente sistêmicas.

Além disso, a importância desses aspectos se manifesta no fato de que certas instituições já articularam um envolvimento máximo do Conselho de Administração e da Alta Administração nos aspectos relativos à informação e reporting, concretizada, em alguns casos, na criação de comitês delegados do Conselho para a gestão de dados. Por outro lado, as instituições estão abordando de forma generalizada a criação de figuras organizacionais como o Chief Data Officer (CDO) ou os responsáveis por Risk Management Information (RMI), e estão abordando iniciativas estratégicas para fortalecer a infraestrutura de suporte aos processos de geração de informação.

Os benefícios de uma governança sólida dos dados são claros: permite a obtenção de um reporte homogêneo e consistente, utilizando conceitos uniformes em toda a organização e alinhando a matriz e as filiais no caso de grupos internacionais; garantia da consistência entre o reporting regulatório e o gerencial; obtenção de uma maior eficiência (maior automação, menores redundâncias), melhoria do time-to-market e torna mais flexível a geração de reporting; e, no caso de riscos, facilita o conhecimento preciso dos riscos pela Alta Administração e, em suma, contribui para melhorar a gestão e o controle de riscos da instituição.

Elementos de um framework de informação e governança dos dados

Portanto, as instituições mais avançadas na questão dispõem de um framework de governança da informação e dos dados que detalha os princípios básicos, os intervenientes e suas funções, a estrutura de governança e os elementos (processos e ferramentas) de suporte em relação à gestão dos dados e à geração de informação.

Em relação aos princípios, o framework deve identificar as diretrizes básicas que regem a governança da informação e dos dados, incluindo o escopo do framework e seu âmbito de aplicação, a propriedade dos dados, a consistência entre os diferentes âmbitos ou os mecanismos implementados para garantir a qualidade da informação.

Em relação à organização e à governança, o framework identifica os intervenientes no processo e suas funções, incluindo, em particular, os responsáveis pela geração da informação, garantia da qualidade e repositórios de informação. Se destacam figuras-chave como o Chief Data Officer (CDO), responsável por assegurar a qualidade e a rastreabilidade end-to-end dos dados nos relatórios para a Alta Administração e a consistência da informação (que se apoia em ferramentas como o dicionário de dados), os responsáveis pelos dados (por âmbito) ou, no âmbito de Riscos, o responsável de Risk Management Information (RMI), entre outros.

Além disso, o framework contempla a definição dos órgãos de governança responsáveis pela informação e pelos dados, cujas atribuições incluem a promoção da elaboração e implantação efetiva do modelo da governança dos dados, a revisão e aprovação de modificações relevantes no processo de geração de informação, a aprovação dos objetivos de qualidade dos

[77] Risk Data Aggregation and Risk Reporting Framework.

dados e, em geral, a definição da estratégia de gestão de dados. Nestes comitês, devem estar representados todos os usuários dos dados, incluindo habitualmente as divisões de Negócio, Riscos e Finanças, assim como os responsáveis pela área de Sistemas, o CDO e os responsáveis pela geração de informação.

Nos comitês, é possível distinguir vários níveis, incluindo comitês técnicos responsáveis pelo suporte aos comitês de categoria mais elevada e resolver eventuais conflitos sobre dados que afetem vários âmbitos ou áreas geográficas. Além disso, em instituições internacionais, é preciso garantir a extensão da governança dos dados de forma consistente a todas as áreas geográficas, sendo necessária a constituição dos comitês pertinentes nos diferentes países e o estabelecimento dos mecanismos de reporting adequados e a escalação aos comitês de âmbito corporativo.

A governança dos dados requer uma série de elementos para a sua correta articulação, que facilitem o cumprimento dos princípios de qualidade, rastreabilidade, consistência e granularidade requeridos pelos reguladores[78], entre os quais destacamos:

- Dicionário de dados: é um inventário unificado de métricas, dimensões e componentes associados aos relatórios, com definições funcionais claras e unificadas para toda a instituição.

- Metadados: é a informação específica sobre cada dado, contida no dicionário de dados, que permite sua catalogação e condiciona sua utilização. A tendência é enriquecer os metadados atuais (de negócio e técnicos), completando-os com metadados adicionais sobre a origem dos dados, sua transformação e sua qualidade, que condicionam a decisão sobre o uso que se pode dar a cada dado.

- Datawarehouses e data lakes: são os bancos de dados e outras fontes de informação que reúnem a qualidade suficiente para construir as métricas candidatas a serem incluídas nos relatórios.

- Ferramentas de exploração: são as ferramentas analíticas de processamento e visualização da informação; entre as que desempenham um papel essencial são as que têm capacidades de processamento de big data.

Por fim, é fundamental assegurar a qualidade dos dados utilizados. Para isso, as melhores práticas consideram:

- O estabelecimento de um modelo de controle dos dados que inclua o monitoramento dos processos de geração dos relatórios para a Alta Administração e a definição dos níveis de tolerância de qualidade que devem ser aplicados aos dados que serão reportados.

- A identificação, definição e implantação de KPIs que permitam mensurar o grau de qualidade da informação em múltiplos níveis no ciclo de vida dos dados, assim como a definição de ferramentas (dashboards) de agregação e acompanhamento dos níveis de qualidade dos dados nos relatórios para a Alta Administração.

- A execução de planos de qualidade dos dados em diferentes níveis (sistemas operacionais, repositórios de informação e relatórios), que são concretizados em iniciativas de depuração de dados históricos (que habitualmente são abordados com planos de choque) e de melhoria da nova produção de informação (por meio de modificações nos processos).

[78]Por exemplo, o Comitê de Supervisão Bancária do Basileia, a Federal Reserve e a OCC.

Desafios na governança dos dados

O desenvolvimento sólido de uma governança dos dados implica, no entanto, uma série de desafios que as instituições devem enfrentar, entre os quais se destacam:

- Garantir o envolvimento da Alta Administração na governança da informação, dos dados e de sua qualidade.
- Definir o escopo de dados sujeito ao modelo de governança (em especial, considerando o crescimento exponencial quanto à diversidade e volume) e assegurar que o modelo seja operacional e garanta os níveis adequados de qualidade, rastreabilidade e consistência dos dados sem implicar um descrédito das capacidades da organização para otimizar seu uso.
- Resolver os aspectos relativos à privacidade e segurança da informação e a garantia de que os dados estão a salvo de usos fraudulentos.
- Reforçar a cibersegurança, que inclui a proteção contra o «hacktivismo» (os ataques contra as instituições por motivos ideológicos através de vírus, malware, etc.), o uso fraudulento dos dados, os ciberdelitos financeiros, a espionagem e o roubo de informação. (Cabe mencionar que em 2014 se o risco de ciberataques foi incorporado aos top 5 riscos globais do Fórum Econômico Mundial).
- Identificar e implementar ferramentas que facilitem a governança dos dados e adaptar os mecanismos de governança dos dados no caso de novas arquiteturas, como os data lakes.
- Implantar um dicionário único de conceitos, que permita uma homogeneidade de entendimento em toda a instituição e, se for o caso, nas filiais.
- Envolver as diversas filiais, no caso de um grupo financeiro, na governança conjunta de dados.

Governança dos modelos

De acordo com a Federal Reserve e a OCC, o termo «modelo» se refere a «um método quantitativo, sistema ou estratégia que aplica teorias, técnicas e hipóteses estatísticas, econômicas, financeiras ou matemáticas para processar dados e obter estimativas quantitativas»[79].

Até hoje, existe pouca legislação que regule de forma específica o risco de modelo e tende a ser pouco específica tanto em sua delimitação como no tratamento esperado. A exceção é a Supervisory Guidance on Model Risk Management publicada em 2011-12 pela OCC e pela Federal Reserve dos EUA.

Nesta publicação, é definido, pela primeira vez, o risco de modelo como «o conjunto de eventuais consequências adversas decorrentes de decisões baseadas em resultados e relatórios incorretos de modelos ou do seu uso inapropriado» e foi estabelecida, por meio de diretrizes, a necessidade de que as instituições disponham de um framework para identificar e gerir este risco, aprovado por seus conselhos de administração.

Estas diretrizes abrangem todas as fases do ciclo de vida de um modelo: desenvolvimento e implantação, uso, validação, governança, políticas, controle e documentação por parte de todos os intervenientes. Entre os principais aspectos requeridos está a necessidade de tratar o risco de modelo com o mesmo rigor que qualquer outro risco, com a particularidade de que não pode ser eliminado, apenas mitigado por com um questionamento efetivo («effective challenge»).

[79] OCC/Fed (2011-12).

Elementos de um framework objetivo de MRM

As instituições mais avançadas nesse assunto dispõem de um framework de gestão do risco de modelo (MRM) substanciado em um documento aprovado pelo Conselho de Administração e que indica aspectos relativos à organização e governança, gestão de modelos, etc.

Em relação à organização e governança, o framework de MRM é caracterizado pela transversalidade (envolve várias áreas, como as linhas de Negócio, Riscos, Auditoria Interna, Tecnologia, Finanças, etc.), a definição explícita das três funções que o regulador exige (ownership, controle e compliance[80]) e sua atribuição a funções específicas da organização e, sobretudo, o estabelecimento de uma função de Gestão de Risco de Modelo, cuja responsabilidade seja criar e manter o framework de MRM.

No que se refere à gestão de modelos, o framework de MRM inclui aspectos tais como: (a) o inventário de modelos, registrando todos os modelos da instituição em todos os âmbitos (riscos, comercial, finanças, etc.), suportado normalmente em uma ferramenta tecnológica apropriada que guarde informação de todas as alterações e versões; (b) um sistema de classificação ou tiering dos modelos, segundo o risco que impliquem para a instituição, do qual depende o nível de exaustividade no acompanhamento, na validação e a documentação dos modelos; (c) uma documentação completa e detalhada de cada modelo, que permita a réplica por parte de terceiros e a passagem para um novo modelador sem perda de conhecimento; e (d) um esquema de acompanhamento dos modelos que permita detectar, antecipadamente, desvios do desempenho do modelo em relação ao previsto, assim como usos inadequados, para a tomada das devidas ações.

A validação dos modelos é um elemento central para a gestão do risco de modelo, e deve tomar como princípio fundamental o questionamento (challenge) crítico, efetivo e independente de todas as decisões tomadas no desenvolvimento, acompanhamento e uso do modelo. A periodicidade e a intensidade da validação de cada modelo devem ser proporcionais ao seu risco, mensurado por meio de seu tier e o processo e o resultado da validação devem ser exaustivamente documentados.

Desafios na governança de modelos

Assim, surge a necessidade de definir uma governança de modelos robusta e estável, o que representa uma série de desafios às instituições financeiras, entre os quais cabe mencionar:

- A reflexão sobre o que é um modelo e que modelos devem submeter-se a estes procedimentos (possivelmente dependendo do tipo de modelo e sua classificação ou tiering) e como compatibilizar esta necessidade de governança dos modelos com um maior uso dos mesmos para fins múltiplos.
- Resolver as dificuldades apresentadas pela existência de maiores volumes e tipos de dados (nem todos submetidos aos mesmos controles de qualidade) utilizados no processo de modelagem.
- Obter o envolvimento da Alta Administração na governança dos modelos e, especificamente, definir e aprovar o framework de risco de modelo ao mais alto nível.
- Definir o esquema organizacional da função (ou funções) da data science em termos de centralização ou descentralização tanto geográfica como entre as áreas da instituição, e delimitar as responsabilidades entre as áreas corporativas e locais, no caso de grupos internacionais.
- Construir ou reforçar os mecanismos de governança em torno de cada um dos processos associados à função analítica.

Em suma, governar os dados e sua transformação em conhecimento que, por sua vez, implica governar os modelos que articulam esta transformação, passou a ser um pilar estratégico de atuação para qualquer organização e, em particular, para as instituições financeiras. Consequentemente, a tendência indubitável nos próximos anos será o impulso decisivo dos respectivos frameworks de governança.

[80] O model owner define os requisitos do modelo e costuma ser seu usuário final. O Controle inclui a mensuração do risco de modelo, o estabelecimento de limites e o acompanhamento, assim como a validação independente. O Compliance abrange os processos que asseguram que as funções do model owner e de controle são desempenhadas de acordo com as políticas estabelecidas.

Estudo de caso: redes sociais e credit scoring

É um erro capital teorizar antes de ter dados. Sem se dar conta, uma pessoa começa a deformar os fatos para que se ajustem às teorias, ao invés de ajustar as teorias aos fatos.

Sir Arthur Conan Doyle[81]

Objetivo

Com o propósito de ilustrar de forma direta a aplicação da disciplina de data science no setor financeiro, considerou-se de interesse a realização de um exercício quantitativo utilizando algumas das ferramentas descritas para um uso específico em uma instituição financeira.

Especificamente, o objetivo do estudo é desenvolver um modelo de scoring de crédito para pessoas físicas utilizando dados extraídos de redes sociais, fazer sua integração com um modelo tradicional de empréstimos pessoais e comprovar o grau de melhoria do poder de previsão.

Dados do estudo

O estudo foi realizado utilizando os seguintes dados e modelos:

- Uma amostra real de construção de um modelo de scoring de empréstimos a pessoas físicas, composta por aproximadamente 75.000 registros, com uma taxa de default de cerca de 12%.
- Variáveis adicionais sobre os clientes da amostra, que permitem a pesquisa nas redes sociais.
- Um modelo de scoring construído sobre a amostra anterior, que utiliza 12 variáveis e tem um poder de previsão médio (ROC[82] de cerca de 73%).

Principais conclusões

As principais conclusões do estudo são as seguintes:

- A quantidade e a qualidade da informação disponível nas redes sociais são notavelmente inferiores às dos dados internos do banco: apenas 24% dos clientes têm dados, e destes, apenas 19% têm informação completa ou quase completa.
- Além disso, a extração de dados de redes sociais é caracterizada por um problema de desambiguação: as pessoas físicas não se identificam de forma inequívoca com um documento de identidade na rede e, portanto, existe uma probabilidade de erro na identificação de cada cliente com o seu perfil nas redes sociais. Para este estudo, foram descartados os clientes em que esta probabilidade foi estimada em percentual superior a 25%.
- As variáveis extraídas das redes sociais também são, na maioria, qualitativas e podem tomar uma grande quantidade de valores, o que dificulta o processamento, mas, em contrapartida, permite construir variáveis de grande riqueza.
- O poder de previsão do modelo de scoring baseado em redes sociais utiliza 9 variáveis numéricas e categóricas (algumas convertidas em variáveis discretas) que cobrem vários aspectos do perfil profissional do cliente (em especial o histórico profissional, mas também o setor, os estudos e os idiomas) e alcança um poder de previsão equiparável ao do modelo original, com uma ROC de 72%.
- A combinação de ambos os modelos, no entanto, eleva substancialmente o poder de previsão, até alcançar 79%.

[81] Sir Arthur Ignatius Conan Doyle (1859-1930). Escritor e médico escocês, célebre pelo seu personagem Sherlock Holmes.
[82] Receiver operating characteristic, medida do poder de previsão de um modelo de resposta binária.

Em síntese, este estudo revela que a informação das redes sociais proporciona uma informação substancialmente diferente, que complementa e enriquece significativamente o modelo de scoring tradicional. Não obstante, subsistem certas dificuldades inerentes à sua utilização que, no futuro, previsivelmente, serão resolvidas em grande parte com a coleta ordenada dos dados por parte das instituições como parte do seu processo de concessão e monitoramento do crédito.

Descrição do estudo

O estudo foi abordado em quatro fases: extração de dados das redes sociais, limpeza e processamento, construção do «módulo social» e integração com o «módulo tradicional».

Para a extração de dados das redes sociais, foi utilizada uma combinação de uma ferramenta especificamente concebida na Python, que é ligada por intermédio de APIs nativas das próprias redes sociais, com um módulo em VBA que extrai e ordena a informação em um formato acessível.

A informação das redes sociais é muito irregular em sua completitude e qualidade, costuma ser composta por variáveis qualitativas e não se adapta a listas de valores preestabelecidos. Além disso, não é extraída em um esquema relacional clássico, mas em registros que requerem um processo de análise ou parsing para convertê-los em informação utilizável.

Para este estudo, foram encontrados dados de aproximadamente 18.000 dos 75.000 clientes, dos quais a informação estava razoavelmente completa, em cerca de 4.000. Não obstante, foi efetuado um processamento exaustivo dos valores ausentes ou missings, de modo que a quantidade de registros finalmente utilizável foi maior.

Para esses clientes, foram extraídas 30 variáveis originais, com diferentes níveis de falta de informação. As variáveis abarcavam vários âmbitos do perfil profissional e pessoal do cliente: sua formação acadêmica formal e não formal, sua experiência profissional, sua localização geográfica e outros dados relativos a gostos, interesses, etc.

O processamento dos dados compreendeu também as transformações necessárias para manipular a informação, a construção de variáveis compostas a partir das originais. Assim, a partir das 30 variáveis extraídas foram criados mais de 120 campos, na maioria, categorizações baseadas em análises univariantes, bivariantes e análises temporais do histórico profissional de cada cliente.

A fase de construção do modelo é muito similar à de qualquer modelo de scoring. Combinou-se a análise especialista com um processo stepwise de seleção de variáveis, eliminaram-se as variáveis redundantes do ponto de vista de negócio e aplicou-se um critério de aceitação de 95% de confiança (p-valor inferior a 0,05).

O algoritmo utilizado foi uma combinação de uma árvore de decisão (reforçado com um algoritmo de corte para reduzir a entropia e, portanto, melhorar a robustez e a estabilidade) e uma regressão logística binomial:

$$P(Y=1)=\frac{1}{1+\exp(-\beta_0-\beta_1 x_1-\cdots-\beta_n x_n)}$$

Como resultado, obteve-se um modelo de 9 variáveis, que está resumido na Fig. 27. A curva ROC, que mede o poder de previsão, é apresentada na Fig. 28.

Como podemos observar, é um modelo com um poder de previsão médio, mas equiparável ao do modelo tradicional, mesmo utilizando apenas 9 variáveis; isso atribui-se, em certa medida, à presença de variáveis qualitativas. Os demais dados estatísticos mostram que o modelo é robusto e tem boas propriedades estatísticas.

A última fase é a integração do módulo social com o módulo tradicional do scoring. Para tal, foi construída uma nova regressão logística que toma como variáveis independentes as pontuações ou scores de cada um dos módulos:

$$P(Y=1)=\frac{1}{1+\exp(-\delta_0-\delta_1 score\_trad-\delta_2 score\_social)}$$

Ambos os scores são significativos no nível de confiança em 95% e, como se pode verificar na Fig. 29, o modelo final tem uma área sob a curva ROC de 79%, o que melhora substancialmente o poder de previsão de cada modelo separadamente.

O presente estudo foi enfocado no caso de um modelo de scoring de crédito e a incorporação de dados provenientes de redes sociais, mas o exercício é extensível a outros tipos de modelos (avaliação de garantias, fidelização, renda, abandono, propensão à compra, etc.) e fontes de informação (logs internos, bancos de dados públicos, informação web, etc.).

Concluindo, como foi demonstrado, a incorporação de variáveis provenientes de outras fontes tem o potencial de aumentar significativamente a capacidade discriminante dos modelos tradicionais.

Fig. 27. Variáveis do módulo social do scoring de crédito.

Variável	Descrição	Peso relativo
Duração no cargo atual	Duração em meses do cargo atual	18%
Tempo de carreira profissional	Tempo de carreira profissional em meses	15%
Duração mínima em um cargo	Mínima duração em um cargo em sua trajetória profissional	15%
Duração máxima em um cargo	Máxima duração em um cargo em sua trajetória profissional	13%
Setor de atividade	Código de categoria do setor profissional	12%
Número de empregos	Número de empregos atuais e históricos	9%
Tempo sem estudar	Tempo transcorrido desde seus últimos estudos	7%
Idiomas	Número de idiomas falados	7%
Índice cargos/anos	Número de cargos/número de anos de trajetória profissional	4%

Fig. 28. Curva ROC do módulo social do scoring de crédito.

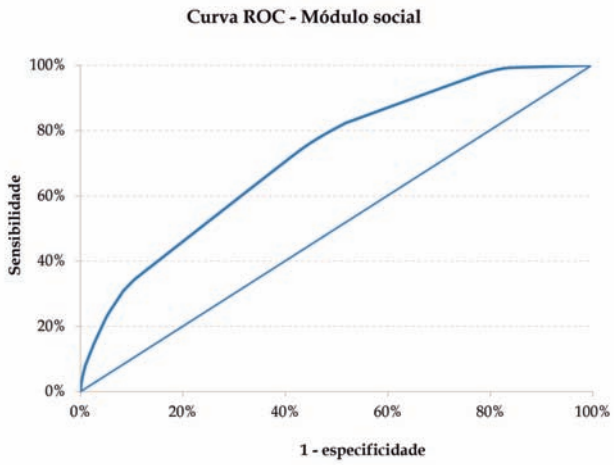
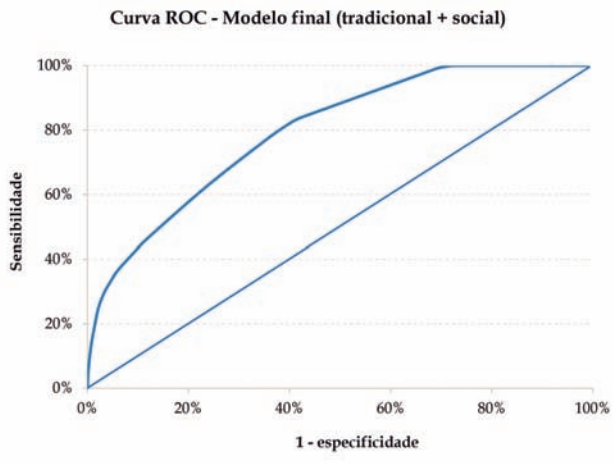


Fig. 29. Curva ROC do modelo final, que integra os módulos tradicional e social.



Referências

Basel Committee on Banking Supervision 239 (2013). *Principles for effective risk data aggregation and risk reporting.*

BBVA Research (2014). *Latam Economic Outlook: Third Quarter 2014.*

Berkeley (2015). http://datascience.berkeley.edu/about/what-is-data-science/

Bloomberg (2013). *HSBC Judge Approves $1.9B Drug-Money Laundering Accord.*

DeZyre (2014). *Hadoop in Financial Sector.*

Dhar, V. (2013). *Data Science and Prediction, Association for Computer Machinery.*

Digital Leadership GmbH (2014). *What banks will have to work on over the next couple of years.*

EFMA (2013). *World Banking Report 2013.*

European Banking Authority (2014). *Guidelines on common procedures and methodologies for the supervisory review and evaluation process (SREP).*

European Central Bank (2014). *ECB Banking Supervision.*

European Commission (2014). *EU Bank Recovery and Resolution Directive (BRRD): Frequently Asked Questions.*

Evans, P. (2014). *How data will transform business. TED.*

Federal Big Data Commission (2014). *Demistifying big data, a practical guide to transforming the business of Government.*

Federal Reserve (2014). *Consumer and Mobile Financial Services 2014.*

Fernald, J. (2014). *Productivity and Potential Output before, during, and after the Great Recession. NBER Working Paper 20248,* National Bureau of Economic Research, Cambridge, Massachusetts.

Financial Conduct Authority (2015). fca.org.uk/firms/being-regulated/enforcement/fines

Gartner (2013). *Gartner Says the Internet of Things Installed Base Will Grow to 26 Billion Units By 2020.*

Gigaom Research (2014). *Sector RoadMapTM: data discovery in 2014.*

Goldman, R. (2014). *Big Data, Risk Management, and Full-Fidelity Analytic.* Cloudera.

Gordon, R. (2014). *The Demise of U.S. Economic Growth: Restatement, Rebuttal, and Reflections.* NBER Working Paper 19895, National Bureau of Economic Research, Cambridge, Massachusetts.

Hall, R. (2014). *Quantifying the Lasting Harm to the U.S. Economy from the Financial Crisis.* NBER Working Paper 20183, National Bureau of Economic Research, Cambridge, Massachusetts.

Harvard Business Review (2012). *Data Scientist: The Sexiest Job of the 21st Century.*

Harvard (2014). *CS109 Data Science.*

KPCB (2014). *Internet trends 2014.*

IBM (2014a). *Demystifying Big Data: Decoding The Big Data Commission Report.*

IBM (2014b). *What is a Data Scientist.*

IBM & University of Oxford (2015). *The real world of Big Data.*

Inetco (2015). *Driving Banking Engagement with Customer Analytics.*

International Monetary Fund (2014). *World Economic Outlook, oct. 2014.*

International Telecommunication Union (2014). *The World in 2014, facts and figures.*

Kurzweil, R. (2014). *The accelerating power of technology.*

Management Solutions (2012). *Risco de liquidez: marco regulatório e impacto na gestão.*

Management Solutions (2013). *Análise de impacto de testes de estresse do sistema financeiro.*

Management Solutions (2014). *Model Risk Management: aspectos quantitativos e qualitativos da gestão do risco de modelo.*

McCallum (2014). *Disk Drive Prices (1955-2014),* jcmit.com.

McKinsey (2011). *Big data: The next frontier for innovation, competition and productivity.*

Moore, G. (1965). *Cramming more components onto integrated circuits.* Electronics Magazine. p. 4.

Office of the Comptroller of the Currency e Board of Governors of the Federal Reserve System (2011-12). *Supervisory Guidance on Model Risk Management.*

Office of the Comptroller of the Currency (2014). *OCC Guidelines Establishing Heightened Standards for Certain Large Insured National Banks, Insured Federal Savings Associations, and Insured Federal Branches; Integration of Regulations.*

O'Neil, C. e Schutt, R. (2013). *Doing Data Science.* O'Reilly.

Pearn, J. (2012). *What is Google's total computational capacity?*

Pethuru, R. (2014). *Handbook of Research on Cloud Infrastructures for Big Data Analytics.* IGI Global.

Pingdom: royal.pingdom.com

Portio Research (2013). *Portio Research Mobile Factbook 2013.*

SiliconAngle (2014). *When Will the World Reach 8 Zettabytes of Stored Data?*

Wired (2015). *White House Names DJ Patil as the First US Chief Data Scientist.*

World Bank, Sabbata, S. e Graham, M. (2013). *Internet Population 2011* – DeSabbata Graham OII.

Glossário

Bail-in: resgate de uma instituição com recursos de seus acionistas e credores.

Bail-out: resgate de uma instituição com recursos de fundos públicos.

Buffer de capital: adicional de capital, cujo objetivo é garantir que uma instituição seja capaz de absorver as perdas decorrentes da atividade em períodos de estresse.

Comitê de Supervisão Bancária de Basileia (BCBS): órgão supranacional para a regulação prudencial dos bancos. Seu objetivo é melhorar a qualidade e promover a homogeneização da supervisão do sistema financeiro.

COREP (Common Reporting): framework normativo de reporting definido pela EBA que padroniza a apresentação de relatórios de solvência.

Curva ROC (Receiver Operating Characteristic): curva utilizada para analisar o poder de previsão de um modelo de saída binária. Representa a relação entre o erro de tipo 1 (classificar incorretamente acontecimentos adversos) e o erro de tipo 2 (classificar incorretamente acontecimentos favoráveis).

EBA (European Banking Authority): autoridade independente da União Europeia, cujo objetivo principal é manter a estabilidade financeira dentro da União e salvaguardar a integridade, eficiência e funcionamento do setor bancário. Foi estabelecida em 1 de janeiro de 2011 como parte do Sistema Europeu para a Supervisão Financeira (ESFS) e absorveu o anterior Comitê Europeu de Supervisores Bancários (CEBS).

FATCA (Foreign Account Tax Compliance Act): lei federal que exige que as instituições financeiras de todo o mundo reportem à agência fiscal dos Estados Unidos as contas dos cidadãos norte-americanos no exterior. O objetivo é promover a transparência fiscal.

Fed (Federal Reserve System): banco central dos Estados Unidos, fundado em 1913, com o objetivo de proporcionar à nação um sistema monetário e financeiro mais seguro, flexível e estável. Com o passar dos anos, seu papel no setor bancário e econômico expandiu-se, incluindo atividades como a gestão da política monetária nacional, a supervisão e regulação das instituições bancárias ou o fornecimento de serviços financeiros a instituições depositárias.

Federal Big Data Commission: comissão federal cujo objetivo é proporcionar assessoria ao Governo dos Estados Unidos sobre como utilizar os dados para aumentar a eficiência e reduzir seus custos.

Financial Stability Board (FSB): órgão supranacional que pretende aumentar a estabilidade do sistema financeiro global por meio de uma coordenação maior entre as autoridades financeiras nacionais.

FINREP (Financial Reporting): framework normativo de reporte definido pela EBA que padroniza a apresentação das demonstrações financeiras.

IAS 39 e IFRS 9: normas contábeis relativas à contabilidade de instrumentos financeiros e que, entre outras medidas, requerem o cálculo de provisões com a utilização de modelos internos.

ICAAP (Internal Capital Adequacy Assessment Process): processo interno de autoavaliação da adequação do capital no setor bancário.

ILAAP (Internal Liquidity Adequacy Assessment Process): processo interno de autoavaliação da adequação da liquidez no setor bancário.

Internet das coisas: interconexão dos objetos de uso quotidiano por intermédio da Internet. Segundo o Gartner, em 2020 haverá no mundo 26 bilhões de objetos conectados.

IRB (Internal Rating Based): método avançado de estimativa de capital regulatório baseado em modelos de rating internos. Para

acessar esse método, as instituições devem cumprir um conjunto de requisitos e obter autorização do supervisor.

KYC (Know Your Customer): informação relevante de clientes obtida com diversos objetivos, como o cumprimento regulatório em relação à fraude, lavagem de dinheiro, financiamento do terrorismo ou corrupção.

Mecanismo Único de Supervisão (SSM): mecanismo criado em 2014 que assume as competências de supervisão das instituições financeiras europeias. É formado pelo Banco Central Europeu e pelas autoridades nacionais competentes de supervisão dos países da zona do euro. Os principais objetivos são assegurar a solidez do sistema bancário europeu e aumentar a integração e a segurança financeiras na Europa. Realiza a supervisão direta das 120 instituições mais significativas e a indireta das aproximadamente 3.000 menos significativas.

Modelo de scoring de crédito (credit scoring): sistema de classificação automática do nível de risco de um crédito. Utiliza-se, entre outros usos, para o cálculo da probabilidade de default e para decidir sua concessão de forma automática.

MREL (Minimum Requirement for Own Funds and Eligible Liabilities): requerimento mínimo de fundos próprios e passivos elegíveis para o bail-in.

NFC (Near Field Communication): tecnologia sem fio que permite enviar e receber dados em alta velocidade e a curta distância. Utiliza-se, entre outros usos, para realizar pagamentos com o telefone celular.

NLP (Natural Language Processing): processamento da linguagem natural; estudo das interações entre máquinas e a linguagem humana por meio da análise das construções sintáticas e o nível léxico, entre outros elementos.

OCC (Office of the Comptroller of the Currency): agência federal norte-americana responsável pela regulação e supervisão de bancos nacionais, gabinetes federais e agências de bancos estrangeiros. Seu objetivo principal é garantir que operem de forma segura e sólida, assim como o cumprimento regulatório, incluindo o tratamento justo e imparcial de clientes e seu acesso ao mercado financeiro.

PPNR (Pre-Provision Net Revenue): receita líquida antes do ajuste por provisões.

***Ring-fencing*:** divisão financeira dos ativos de uma empresa feita geralmente por motivos fiscais, normativos ou de segurança. No setor financeiro, alude à separação jurídica entre as atividades de atacado e de banco tradicional, como medida de proteção dos depositantes.

Single Resolution Board: autoridade do mecanismo único de resolução que opera desde 1 de janeiro de 2015, responsável pela tomada de medidas perante a inviabilidade de uma instituição de crédito.

SREP (Supervisory Review and Evaluation Process): processo de revisão e avaliação supervisora. Seu objetivo é assegurar que as instituições financeiras contam com os processos, o capital e a liquidez adequados para garantir uma gestão sólida dos riscos e uma cobertura adequada dos mesmos.

Teste de estresse: técnica de simulação utilizada para determinar a resistência de uma instituição perante uma situação financeira adversa. Em um sentido mais amplo, refere-se a qualquer técnica para avaliar a capacidade para suportar condições extremas, e é aplicável a instituições, carteiras, modelos, etc.

TLAC (Total Loss Absorbing Capacity): requerimento de capacidade total de absorção de perdas, cujo objetivo é garantir que as instituições de importância sistêmica internacional (G-SIBs) tenham a capacidade necessária para assegurar que, em caso de resolução e imediatamente depois, as funções críticas sejam mantidas sem colocar em risco os fundos dos contribuintes nem a estabilidade financeira.

Nosso objetivo é superar as expectativas de nossos clientes, nos convertendo em parceiros de confiança

A Management Solutions é uma empresa internacional de serviços de consultoria com foco em assessoria de negócios, riscos, organização e processos, tanto sobre seus componentes funcionais como na implementação de tecnologias relacionadas.

Com uma equipe multidisciplinar (funcionais, matemáticos, técnicos, etc.) com mais de 1.400 profissionais, a Management Solutions desenvolve suas atividades em 18 escritórios (9 na Europa, oito nas Américas e um na Ásia).

Para atender às necessidades de seus clientes, a Management Solutions estruturou suas práticas por setores (Instituições Financeiras, Energia e Telecomunicações) e por linha de negócio (FCRC, RBC, NT), reunindo uma ampla gama de competências de Estratégia, Gestão Comercial e Marketing, Organização e Processos, Gerenciamento e Controle de Riscos, Informação Gerencial e Financeira e Tecnologias Aplicadas.

No setor financeiro, a Management Solutions oferece serviços para todos os tipos de empresas, bancos, seguradoras, empresas de investimento, financeiras, etc. - tanto para organizações globais, como para instituições locais e órgãos públicos.

Juan Fabios
Sócio
juan.fabios@msbrazil.com

Marcos Izena
Sócio
marcos.izena@msbrazil.com

Pedro Martínez
Sócio
pedro.martinez.ojeda@msspain.com

Javier Calvo
Diretor de P&D
javier.calvo.martin@msspain.com

Design e diagramação
Departamento de Marketing e Comunicação
Management Solutions - Espanha