Data Science para Gestores
Parte I

Hitoshi Nagano, Ph.D.

FGV IDE
In Company Management Online

Data Science para Gestores

FGV IDE In Company Management Online

PARTE I

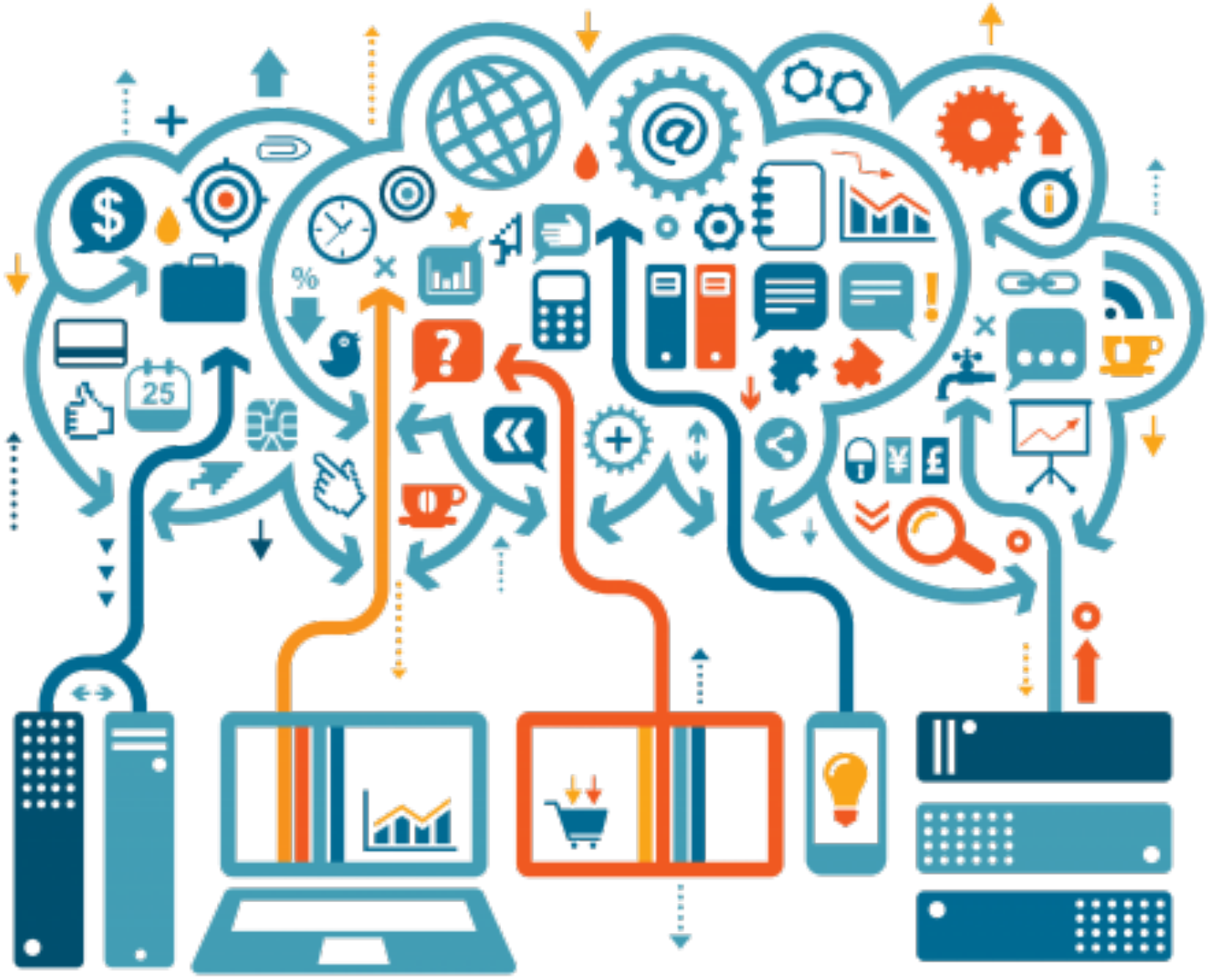
- Por que data science?
- Metodo Cientifico
- Perfil do profissional
- Projetos de data science
- Mercado de trabalho
- Fatos & piadas sobre data science
- Data science para voce

PARTE II

- A equipe
- Montando a equipe
  - onde/como encontrar?
  - perfis interessantes
  - sinais importantes
  - checklist técnico
  - quanto custa?
- O gestor dos DS's/DE's
- Modelos organizacionais
- Bugs & debitos técnicos em Ciencia de Dados

PARTE III

- Aspectos Técnicos
  - tipos de problemas
  - métricas
  - validação cruzada
- Workflows
  - pipeline de modelagem
  - pipeline de produção
- Principais frameworks
- Cases & Demos



POR QUE DATA SCIENCE?

FGV IDE
In Company Management Online

Quantas variáveis conseguimos processar?

- 3?
- 4?
- 10?
- https://www.psychologicalscience.org/pdf/ps/mind_variables.pdf

o que é aprendizagem de máquina?

- Através de dados reais
- a máquina (computador) entende os padrões…
- …e ajuda a predizer um resultado

  (classe ou valor numérico)
- predição é usada para tomada de uma decisão

  de negócios

Forbes INSIGHTS

BLAZING THE TRAIL FROM DATA TO INSIGHT TO ACTION

EXECUTIVE BRIEFING

The enterprises leading the way in data analytics are demonstrating an enormous capability to capture, process, scale and make available data to their organizations and across their networks of customers, partners and suppliers. But is this enough? The challenge is to develop this data into insights that can be applied to increase business value—moving from data to insight to action. Executives say there is a direct correlation between well-designed analytics programs and success in the marketplace. However, while many executives are embracing customer data analytics to guide their businesses, the infrastructures and processes they require to support and sustain such efforts still lag.

These are the findings of a new survey of 105 executives of large global organizations, conducted by Forbes Insights in partnership with SAS, which explores the depth of their embrace of data analytics. The survey finds that while half of large enterprises (with $500 million or more in annual revenue) have tightly integrated customer data analytics into their key processes, there are many areas that are still works in progress.

KEY FINDINGS FROM THIS SURVEY INCLUDE THE FOLLOWING:

- **Half of large enterprises are still in the learning stages of customer analytics, and are in the process of applying these analytics to enhance their customer experience. The results of such efforts, no**

Fonte: "BLAZING THE TRAIL FROM DATA TO INSIGHT TO ACTION", Forbes

FIGURE 1

IN YOUR EFFORTS TO BECOME A MORE CUSTOMER-CENTERED BUSINESS, WHICH OF THE FOLLOWING STATEMENTS HOLDS TRUE FOR YOUR ORGANIZATION?

- We provide a superior customer experience through a well-developed and designed enterprise data analytics effort
- We are rapidly and systematically developing data analytics capabilities to improve our customer experience
- We have data analytics capabilities in many parts of the organization, and are beginning to apply these analytics to improve the customer experience
- We are not using analytics for customer experience
- We are still working on developing data analytics and have difficulty providing consistent customer experience

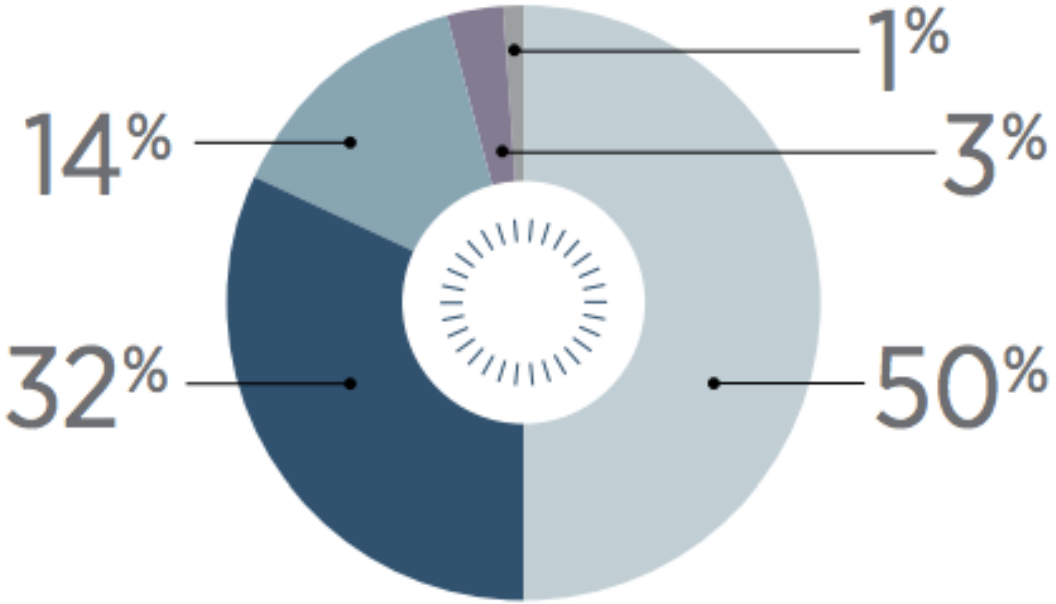


FIGURE 2 PROVIDE SUPERIOR CUSTOMER EXPERIENCE THROUGH ANALYTICS– BY INDUSTRY GROUP

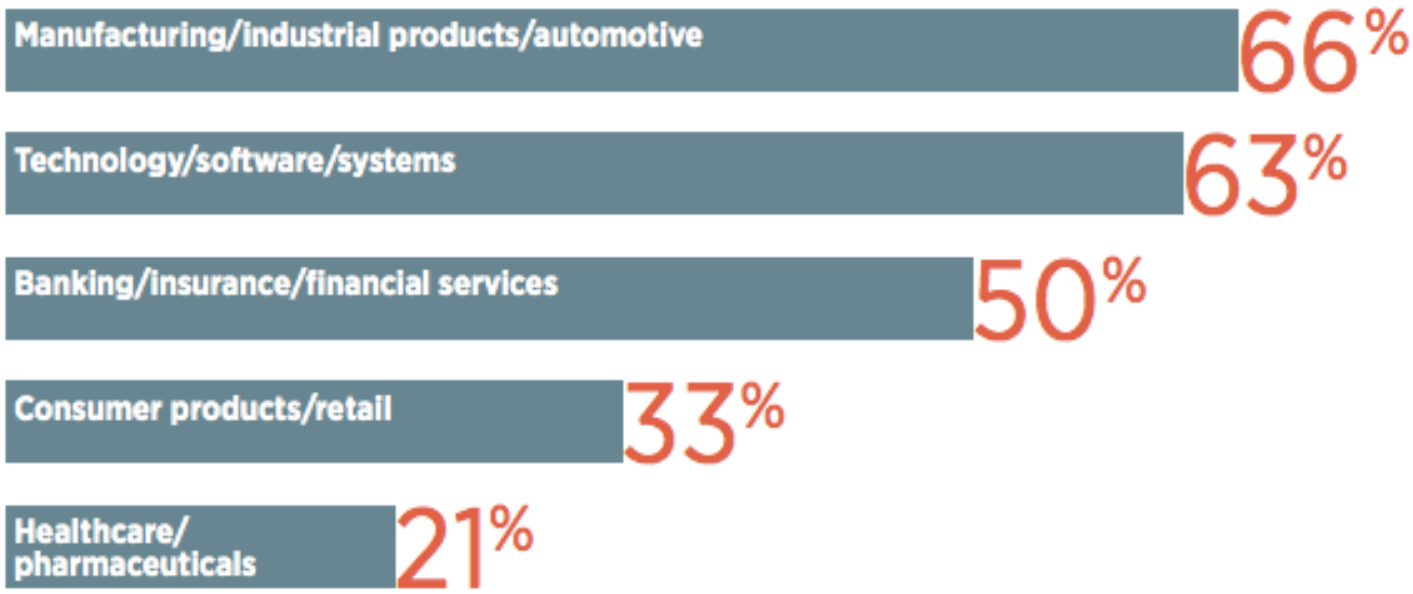


FIGURE 3

WHAT ROLE HAS DATA ANALYTICS PLAYED IN YOUR ORGANIZATION'S ABILITY TO DELIVER A SUPERIOR CUSTOMER EXPERIENCE?

- Data analytics has had an extremely positive impact on our customer experience
- Data analytics has pushed the needle in a positive direction
- Data analytics has shown potential in areas where it has been applied
- Data analytics has not played a significant role yet

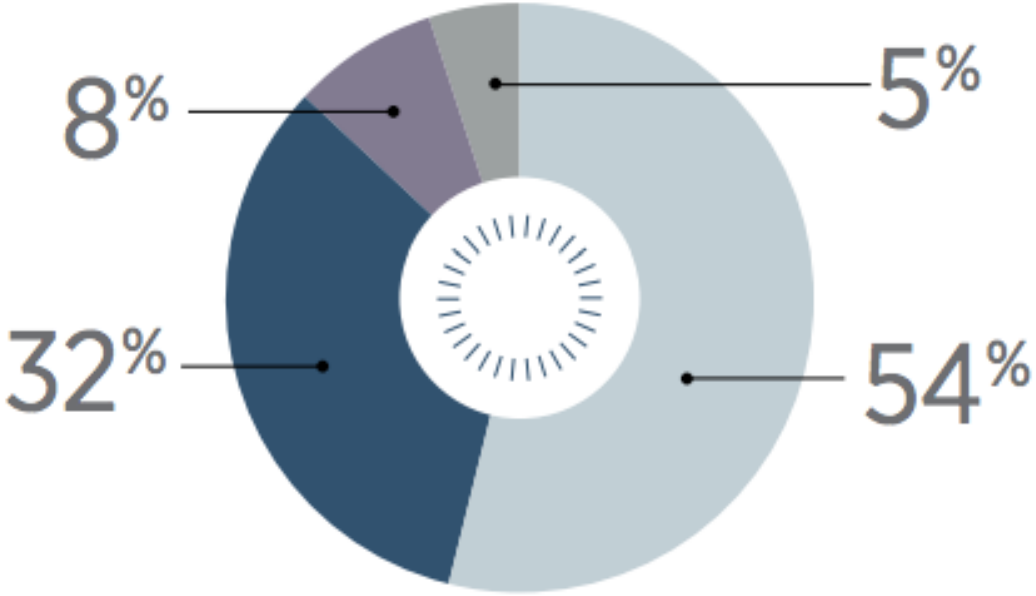
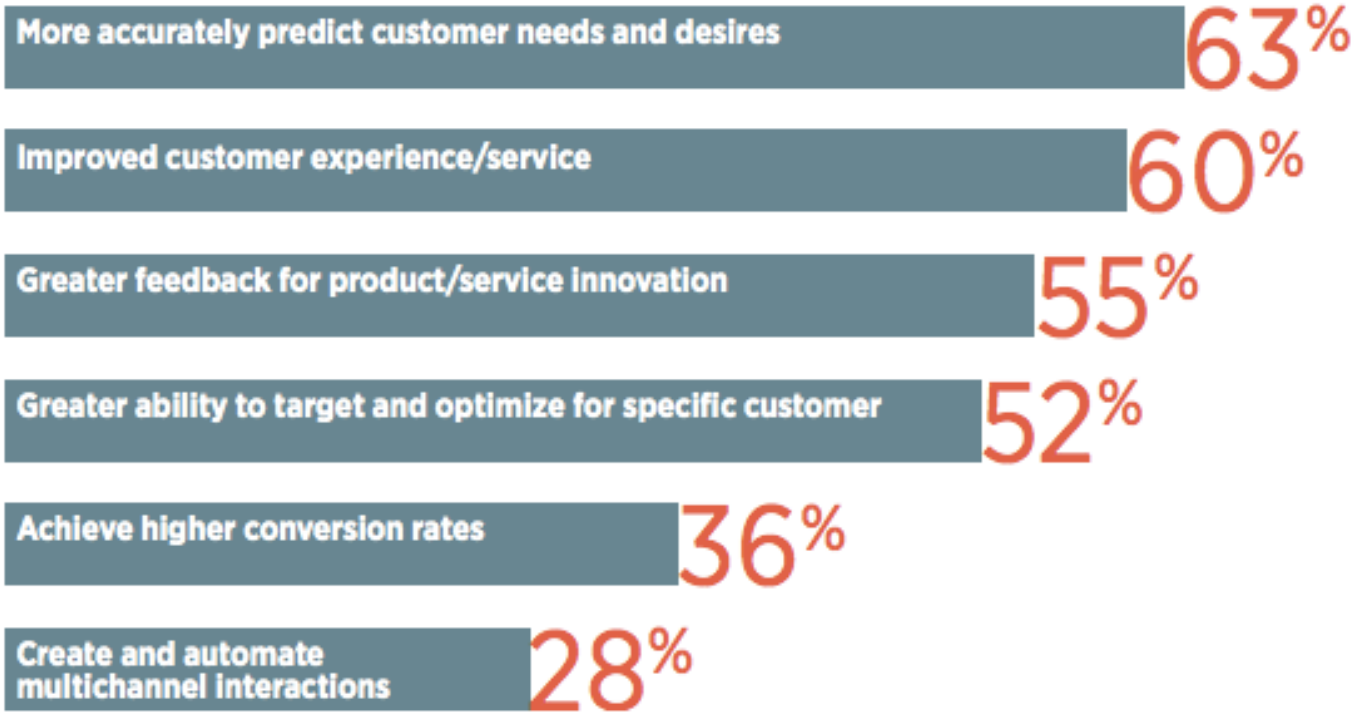


FIGURE 9 WHAT BENEFITS CAN BE GAINED FROM ACHIEVING A MORE COMPLETE OR UNIFIED VIEW OF THE CUSTOMER?



Frases

"contra fatos não há argumentos"

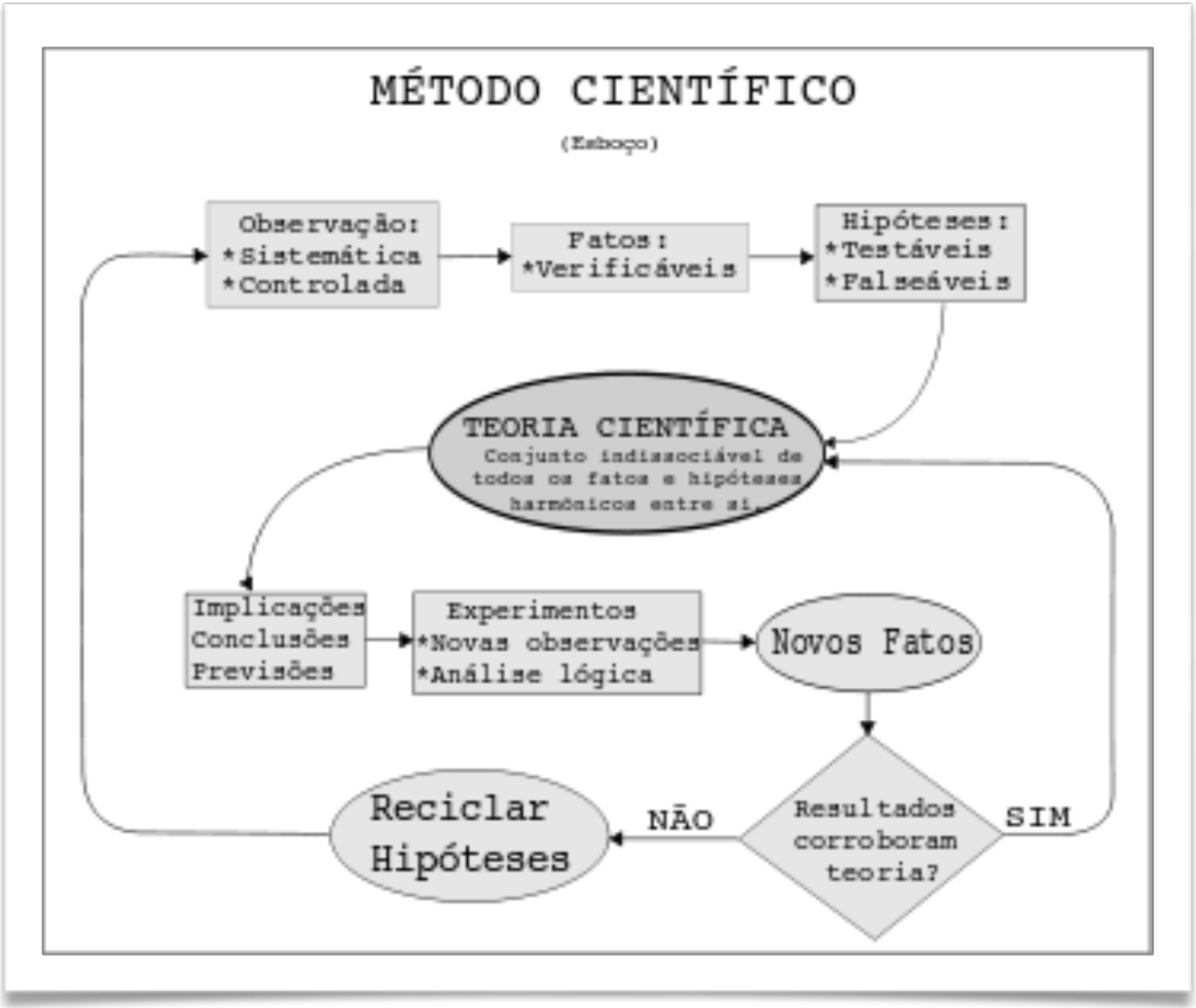
"números não mentem"

"If you torture the data enough, nature will always confess"

"In God we trust, all others must bring data"

Por que ciência?

- ciência em “ciência de dados”
- método científico



Fonte: https://pt.wikipedia.org/wiki/Teoria

PERFIL DO CIENTISTA DE DADOS

FGV IDE
In Company Management Online

Cientista de Dados

- estatística + matemática
- computação + ferramentas
- perspicácia comercial + intuição

  (ex: domain knowledge)
- explorador + inquisitivo
- arte e estética, story telling

Principais conhecimentos

FGV IDE
In Company Management Online

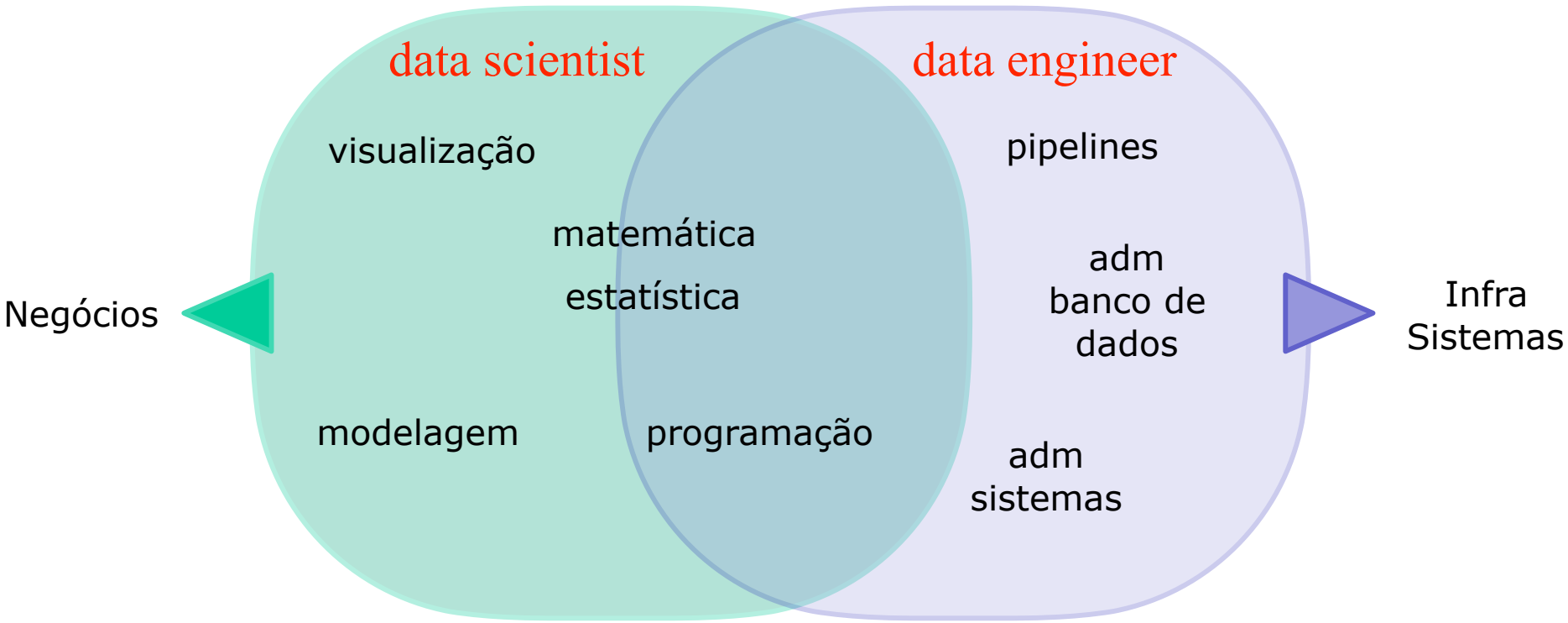
```
def f(n):
  if n == 1:
    return 1
  else:
    return n*f(n-1)
```

$$lim_{z \to -\infty} \frac{1}{1 + e^{-z}}$$

$$lim_{z \to \infty} \frac{1}{1 + e^{-z}}$$

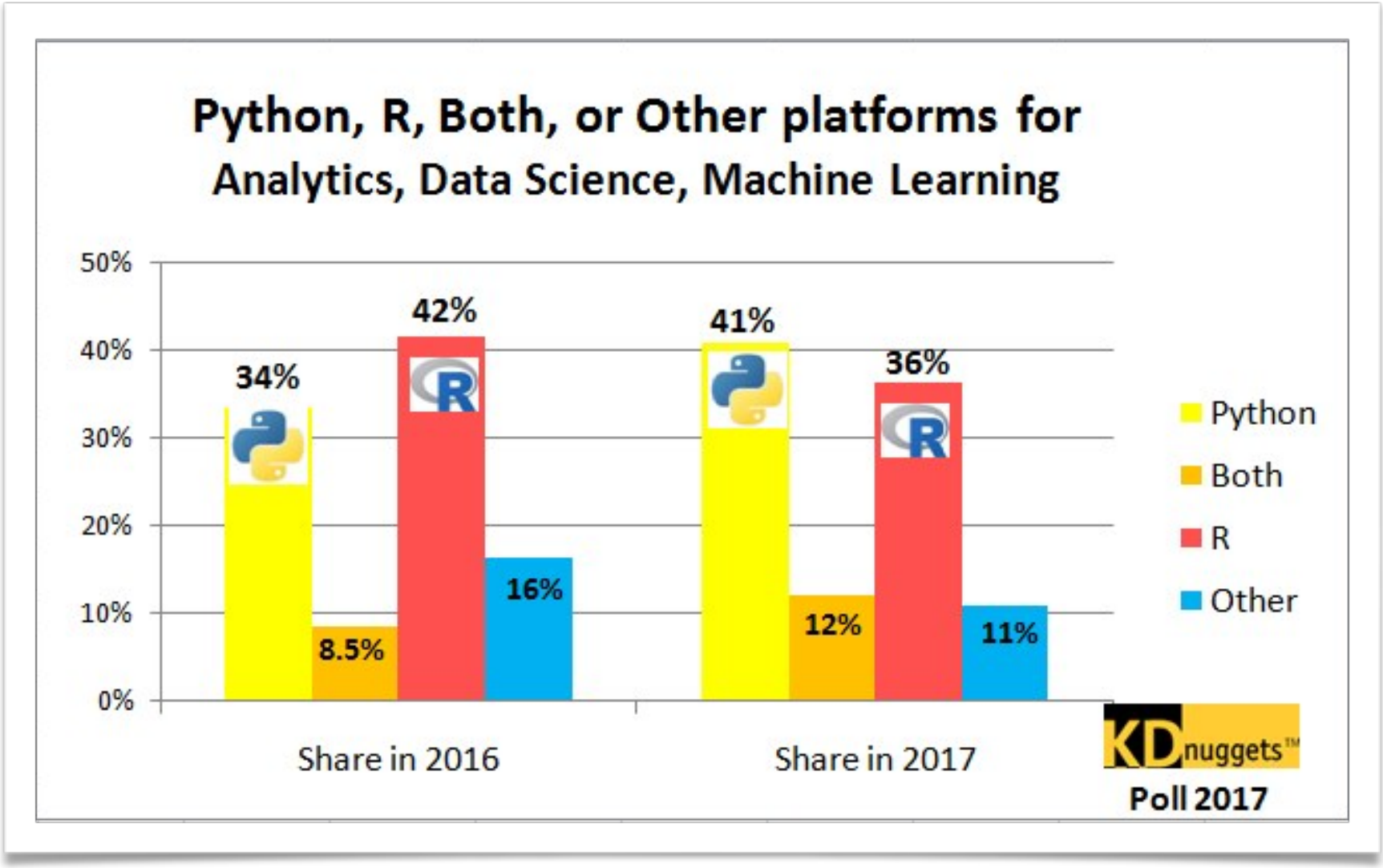
churn, up-sell & cross-sell, lifetime? RFV? NBO? CR, CTR, CPA, CPC?

Equipe de ciencia de dados



Inspirado em: http://101.datascience.community/2014/07/08/data-scientist-vs-data-engineer/

linguagens...



https://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html

PROJETOS DE DATA SCIENCE

FGV IDE
In Company Management Online

- existe um problema de negócios?
- a empresa precisa da análise?
- tenho dados?
- consigo extrair boas features?
- existe um padrão?

exemplos de projetos

FGV IDE
In Company Management Online

- predição de risco de churn
- clusterização de clientes
- extração de tópicos
- recomendação de produtos
- score de clientes
- reordernação de inadimplentes na fila de cobrança
- recomendação de ordem de serviço
- modelagem de funil de conversão
- predição de severidade de falhas em redes
- predição de convulsões epilépticas
- classificação de rotas para marketing geo-referenciado

MERCADO DE TRABALHO DATA SCIENCE

FGV IDE
In Company Management Online

Quem contrata cientista de dados?

FGV IDE In Company Management Online

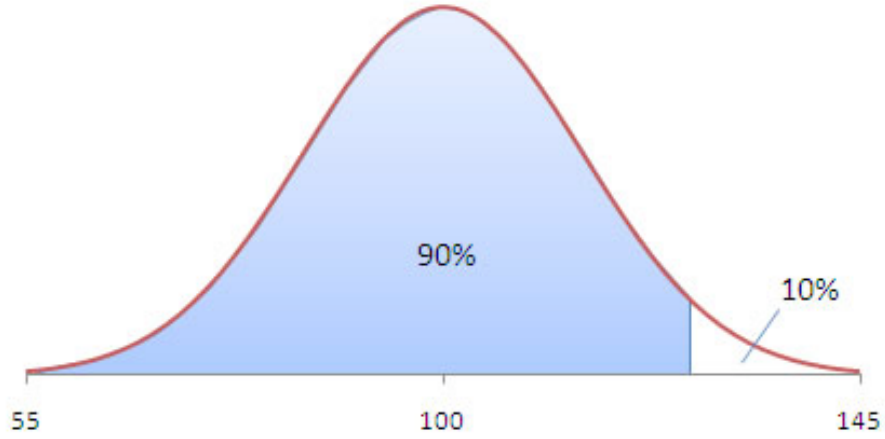
- start-ups
- consultorias
- telecom, varejo, e-commerce, bancos, governo, farmaceuticas, …
- http://corporate.canaltech.com.br/noticia/vivo/vivo-cria-nucleo-de-bi-e-big-data-em-departamento-com-120-profissionais-74597/

Quantos cientistas de dados existem no mundo?

- Forbes 2014
  (https://www.forbes.com/sites/gilpress/2015/10/21/the-number-of-data-scientists-has-doubled-over-the-last-4-years/#5027125538d4)
  Quantidade: **11430 profissionais**

- Stitchdata 2015
  (https://www.stitchdata.com/resources/reports/the-state-of-data-science/)
  Quantidade: **11400 funcionários em empresas**

- KDnuggets 2014
  (http://www.kdnuggets.com/2014/03/how-many-data-scientists-are-there.html)
  Quantidade: **150K a 250K** pessoas tecnicamente capacitadas (perfil LinkedIn)



Quantos cientistas de dados existem no mundo?

>6M Users

https://github.com/jupyter/design/blob/master/surveys/2015-notebook-ux/analysis/report_dashboard.ipynb

cenario 2016

THE DATA SCIENCE / ANALYTICS LANDSCAPE

2,350,000
DSA job listings in 2015

By 2020, DSA job openings are projected to grow
15%

364,000
Additional job listings projected in 2020

Demand for both Data Scientists and Data Engineers is projected to grow
39%

DSA jobs remain open
5 days
longer than average

DSA jobs advertise average salaries of
$80,265
With a premium over all BA+ jobs of
$8,736

81%
Of DSA jobs require workers with 3-5 years of experience or more

OUTROS FATOS E PIADAS SOBRE DATA SCIENCE

FGV IDE
In Company Management Online

- http://priceonomics.com/whats-the-difference-between-data-science-and/
- https://www.analyticsvidhya.com/blog/2015/10/job-comparison-data-scientist-data-engineer-statistician/
- http://blog.revolutionanalytics.com/2013/05/statistics-vs-data-science-vs-bi.html

Analysis Tool	Similar Superhero	Super Powers in Common
R 	Batman 	• Detective Work • Intelligence • Cunning • Usage of Tools • More Brain than Muscles
Python 	Superman 	• Muscle Power • Super Strength • Elegance • Wide Range • More Muscles than Brain

http://i2.wp.com/ucanalytics.com/blogs/wp-content/uploads/2015/10/R.jpeg?resize=198%2C150

- …machine learning faz com que nossa ação seja mais mais assertiva
- A gestão data-driven nos leva a decisões mais assertivas
- https://www.catho.com.br/carreira-sucesso/colunistas/carlos-hilsdorf/o-que-e-assertividade

https://pt.wikipedia.org/wiki/Assertividade

Como tomar decisões mais assertivas na sua empresa

Quando surge aquele pepino, como conduzir a tomada de decisão? Veja algumas dicas para resolver

Da Endeavor Brasil - 02/10/2015

BUSCA DE FRANQUIAS

situações típicas (reais)

- solicitei os dados, e vão me entregar em uma semana
- montei um data lake, gastei uma 💰 💰 💰 , e agora?
- hadoop para 1M de clientes

  https://cloud.google.com/products/calculator/

  https://www.surveysystem.com/sscalc.htm
- legal, essas são as propensões. qual ação devo tomar?

join_us.py

```
> import Analytics_Center_of_Excellence as ACE
> print ACE.vagas.title
[1] "Cientistas de Dados"
> print ACE.vagas.to_know
[1] "O cientista de dados do ACE irá atuar em projetos atendendo diversas áreas do banco (CRM, Crédito, Cobrança, Fraudes, ...) e terá como objetivo atacar problemas desafiadores em ciência de dados.
As soluções envolvem profundo conhecimento teórico e prático das principais técnicas de machine learning e ferramentas do mundo Big Data. Venha para o ACE!"
> print ACE.tecnologias.to_know
```


```
> help ACE.vagas
[1] "Mande e-mail para
    cienciadedados@itau-unibanco.com.br
com o assunto sendo o resultado de:

    sum_hex([x.Unicode\
        .replace("U+","") for\
        x in [🏦, 🎲, 💻, 📚, 🚀]])

Atenção: sintaxe aproximadamente correta,
use apenas como motivação."
```

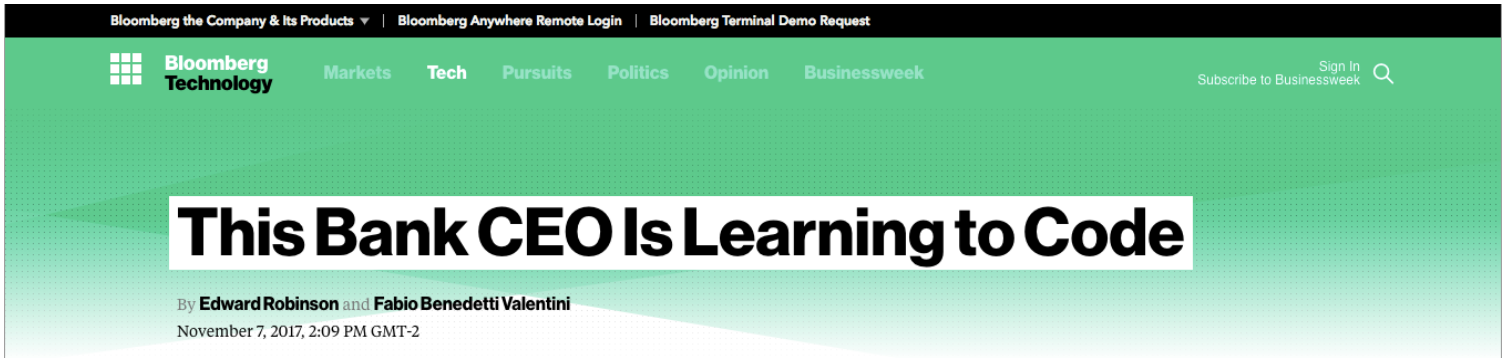
empresa contratando

cientistas de dados

DATA SCIENCE
PARA VOCE...
... E PARA OS CEO'S

FGV IDE
In Company Management Online

Python é para voce… e para os CEO's também



- **Frederic Oudea** is doing everything he can to keep up with the technological changes roiling the European banking industry. The chief executive officer of Societe Generale SA has collaborated with fintech startups, backed accelerator programs to nurture innovation, and invested heavily in its French mobile-banking unit as well as in hundreds of apps.

  Now **he's even taken up writing software code himself**.

  "It was important for me also **to understand exactly what coding means**, so I spend a **few hours coding in Python, which is one of the two languages for data**," Oudea said in an interview at Web Summit 2017, a tech-industry conference in Lisbon.

- "We are taking the challenge seriously and understanding that there is a need to change the model and culturally embrace new technologies," said Oudea, 54.

  By offering alternative methods for making payments, newcomers are trying to take customers away from traditional banks for other services, such as lending. **Societe Generale, the third-biggest French bank by market value**, will be locked in this battle for the next few years, Oudea said.

https://www.bloomberg.com/news/articles/2017-11-07/oudea-learns-to-code-as-societe-generale-fends-off-tech-rivals

Python é para voce… e para os CEO's tambem

Societe Generale CEO Frederic Oudea Oudea speaks to Bloomberg's Ed Robinson from the Web Summit in Lisbon.

Microsoft Closes Acquisition of Revolution Analytics

April 6, 2015 by Cortana Intelligence and ML Blog Team // 10 Comments

This blog post is authored by Joseph Sirosh, Corporate Vice President of Information Management & Machine Learning at Microsoft.

Earlier this year we announced our intent to acquire Revolution Analytics and today I'm happy to say we have closed the acquisition agreement.

It is my pleasure to welcome the Revolution team to Microsoft. Together we will help unlock the power of the R language for advanced analytics on big data.

REVOLUTION ANALYTICS + Microsoft

R is the world's most popular programming language for statistical computing and predictive analytics, used by more than 2 million people worldwide. Revolution has made R enterprise-ready with speed and scalability for the largest data warehouses and Hadoop systems. For example, by leveraging Intel's Math Kernel Library (MKL), the freely available Revolution R Open executes a typical R benchmark 2.5 times faster than the standard R distribution and some functions, such as linear regression, run up to 20 times faster. With its unique parallel external memory algorithms, Revolution R Enterprise is able to deliver speeds 42 times faster than competing technology from SAS.

Anaconda and Microsoft Partner to Offer Python and R for Powerful Machine Learning

October 26, 2017 by Cortana Intelligence and ML Blog Team // 1 Comments

This post was authored by Nagesh Pabbisetty, Partner Director of Program Management, Microsoft Machine Learning Services.

Recently, at Strata Data Conference in New York City, Microsoft and Anaconda announced an exciting partnership to make Anaconda Python distribution into SQL Server, Machine Learning Server, Azure Machine Learning, and Visual Studio to deliver real-time insights. In addition, Anaconda will be distributing Microsoft R. Let's take a deeper look at this exciting new partnership.

Microsoft is committed to helping developers build AI powered applications by enabling them to do machine learning and AI wherever their data is. SQL Server 2017 includes Machine Learning Services — enterprise grade in-database machine learning capabilities with R and Python languages. Machine Learning Server enables customers to do scalable machine learning using R or Python on standalone Windows and Linux servers, Hadoop clusters and Azure data platforms.

Anaconda is the leading distribution of Python leveraged by millions of users today. A strong partnership with this popular Python distribution for data science further strengthens Microsoft's goal of building tools to empower every organization to build their own AI capabilities.

Microsoft and Anaconda built a customized Anaconda distribution – *Anaconda for Microsoft* for doing machine learning with Microsoft products and services. Packages from this distribution will initially be included in SQL Server 2017, Machine Learning Server and Azure Machine Learning.