

# Data Science para Gestores

## Parte II

Hitoshi Nagano, Ph.D.



## PARTE I

- Por que data science?
- Método Científico
- Perfil do profissional
- Projetos de data science
- Mercado de trabalho
- Fatos & piadas sobre data science
- Data science para você

## 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

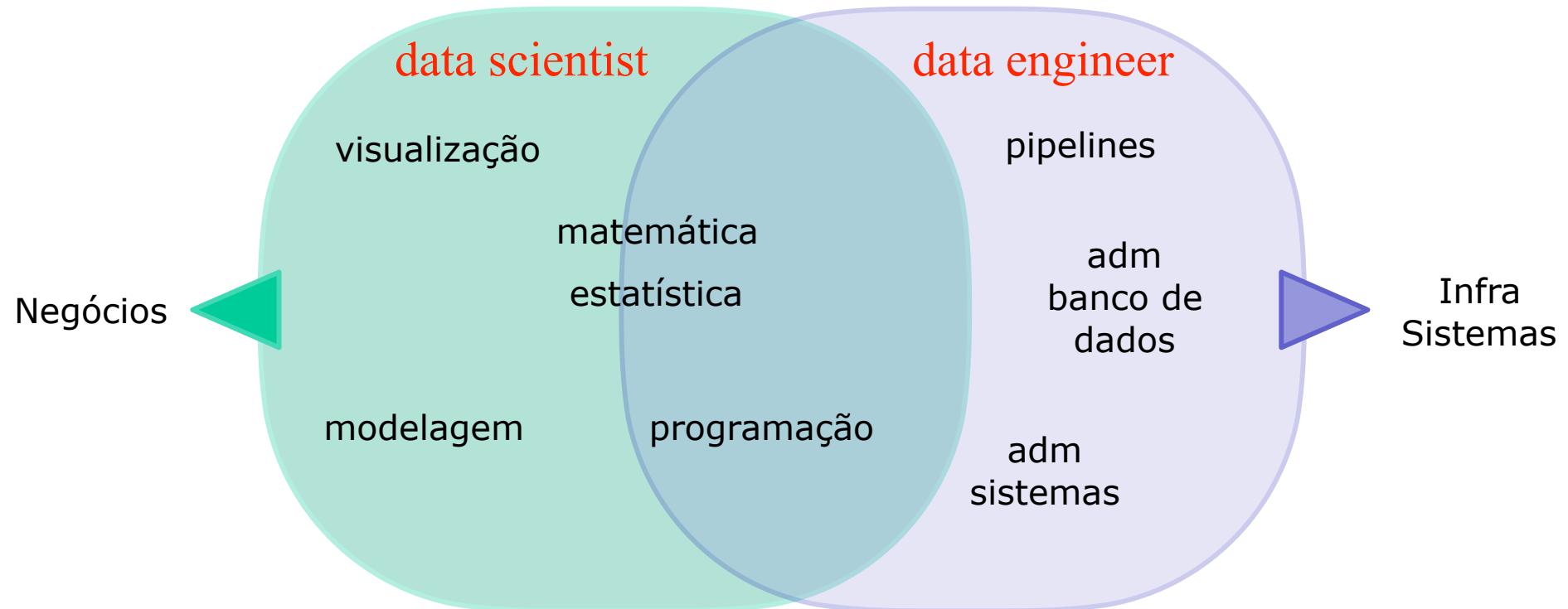
## PARTE III

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

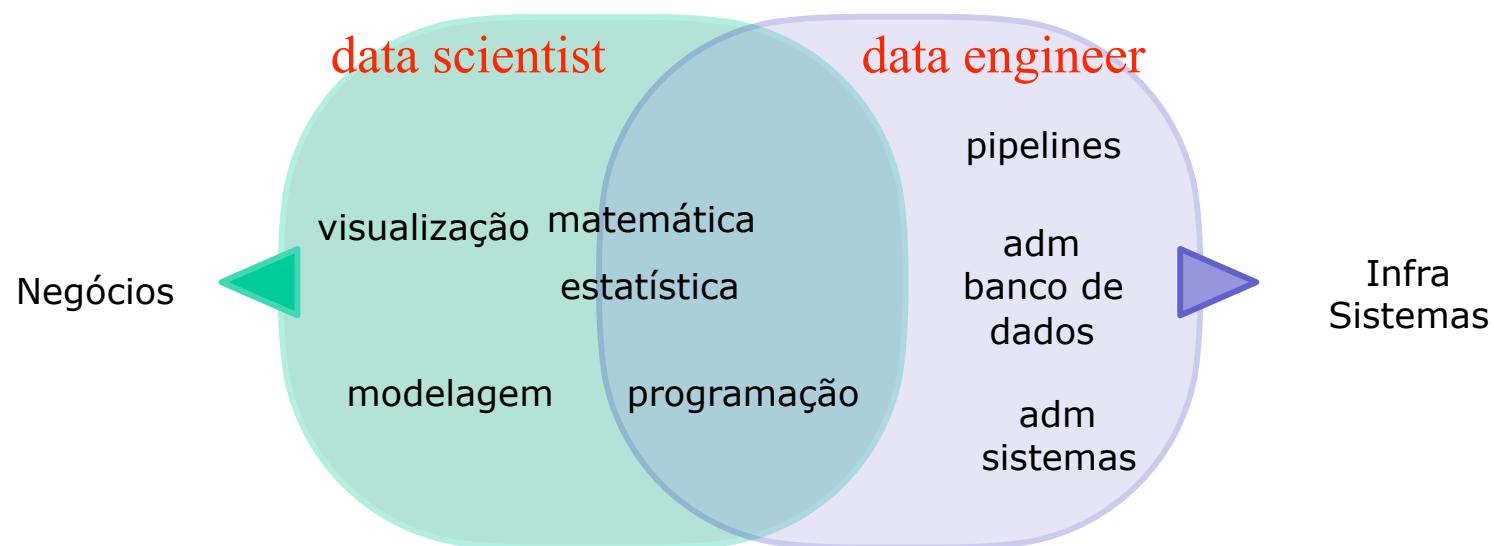
# A EQUIPE

- Trabalho em equipe
  - ➔ Engenheiro de dados
    - >>> governança: armazenamento & acesso & qualidade
    - >>> implementação e administração de ambientes
  - ➔ Cientista de dados
    - >>> algoritmos, programação



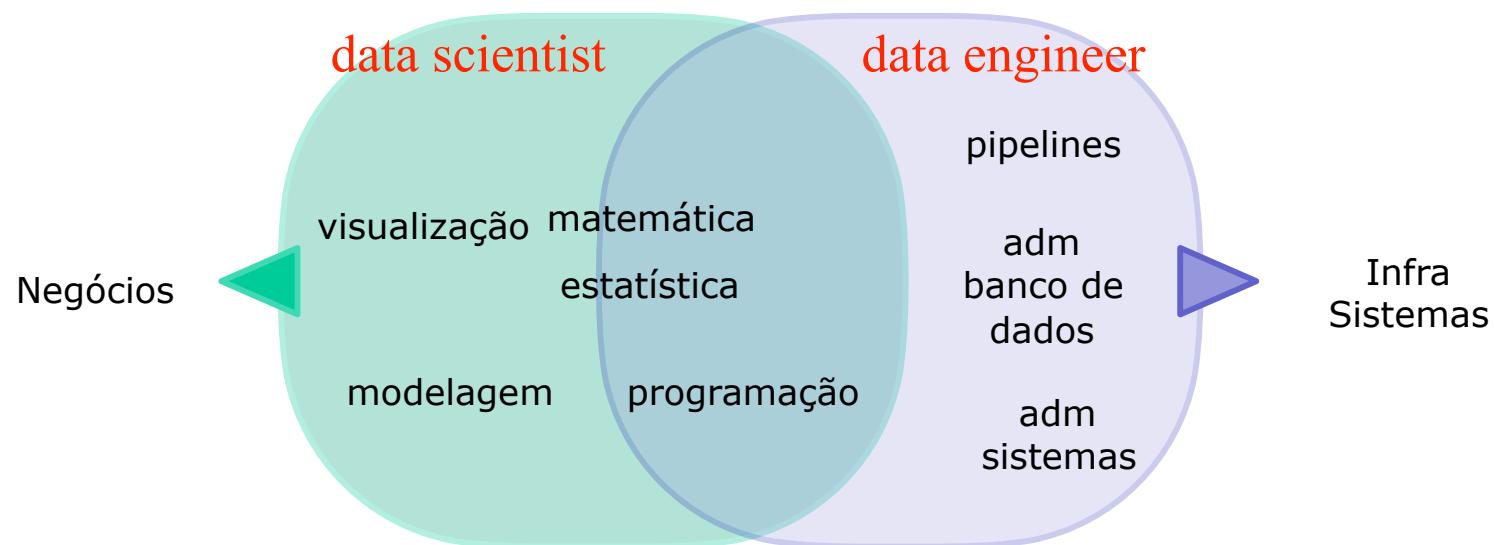


Inspirado em: <http://101.datascience.community/2014/07/08/data-scientist-vs-data-engineer/>



- Atribuições:
  - infra: HW/SW
  - arquitetura
  - implementação em produção
- Competências técnicas:
  - Escolher, montar e manter configurações de HW/SW
  - Banco de dados
- Escalabilidade & segurança
- Interações com outros sistemas, downstream
- No dia-a-dia:
  - Entender o trabalho do cientista de dados
  - Interação com cientistas de dados
  - Encontrar a melhor solução para a fluidez do processo fim-a-fim

Fonte: <http://101.datascience.community/2014/07/08/data-scientist-vs-data-engineer/>



- Atribuições:
  - Construção e validação de modelos
  - Comunicação de resultados
- Competências técnicas:
  - estatística/matematica/programação
  - Domain knowledge
- Story telling
- No dia-a-dia:
  - Entender o trabalho do data engineer
  - Interação com data engineers
  - Encontrar a melhor solução para a fluidez do processo fim-a-fim

Fonte: <http://101.datascience.community/2014/07/08/data-scientist-vs-data-engineer/>

- Data Scientists:

- ➡ Formações:

- Estatística, matemática
    - Ciencia da computação
    - Fisica, química, geofisica
    - Engenharia
    - Administração, economia

- ➡ Experiência profissional:

- Analistas de negócio
    - Desenvolvedores
    - Acadêmico
    - Onde: bancos, seguradoras, corretoras, varejistas, telecom, serviços online,...

- ➡ Comportamental:

- inventivo, incansável, inquisitivo
    - inquisitivo consigo mesmo
    - Aprender sempre!

- Data Engineers:

- ➡ Formações:

- Ciencia da Computação
    - Engenharia
    - Estatística e Matemática

- ➡ Experiência profissional:

- Desenvolvedores / DevOps
    - Administradores de sistema e banco de dados
    - Onde: vide Data Scientists

- ➡ Comportamental:

- disciplinado, organizado e calmo (sob stress)
    - Aprender sempre!

- Gestor

- Gestor do grupo, recrutamento,...
- Comunicação, interação com níveis superiores
- Venda interna das competências



# MONTANDO A EQUIPE

- Populares:
  - LinkedIn
  - Empresas de R&S
  - Redes de contatos
- Universidades:
  - professores
  - pesquisadores
  - alunos
- Sites de competições:
  - Kaggle
- Meetups:
  - machine learning:  
<https://www.meetup.com/machine-learning-big-data-engenharia/>
  - deep learning:  
<https://www.meetup.com/pt-BR/Deep-Learning-Sao-Paulo/>

- Patrocinar  
CURSOS

**Mini Curso na Poli - 2018**

**Introdução à Ciência de Dados**

**Prof. Hitoshi Nagano, Ph.D**

Engenheiro pelo ITA, Doutor em Computação pelo Nagoya Institute of Technology. Professor da FGV-SP

Datas: 18/01, 1/02 e 8/2  
Horário: 17 às 20 hs  
Local: Engenharia Elétrica – Poli-USP

Inscrição com Prof. Leopoldo Yoshioka  
[leopoldo.yoshioka@usp.br](mailto:leopoldo.yoshioka@usp.br)



- Buscar nas universidades

### Grupo Turing

Grupo de extensão da Escola politécnica da USP que se reúne para estudar tópicos de Inteligência Artificial e fazer projetos relacionados ao assunto.

Nota do Autor: Texto extraído do facebook do grupo em JAN/2017.



**.INF**  
INSTITUTO  
DE INFORMÁTICA  
UFRGS

INSTITUTIONAL    UNDERGRADUATE PROGRAMS    GRADUATE PROGRAMS    RESEARCH

PEOPLE    PUBLICATIONS

Search... 

**Research Groups | Artificial Intelligence**

The objective of the Group of Artificial Intelligence (AI) is the development of theoretical and applied research on methods and techniques for AI used in several parts of this broad area.



- Roteiro usual:
  - ➡ múltiplas avaliações:
    - gestores diretos,
    - R&S,
    - áreas de interação horizontal
  - ➡ análise de currículo: experiencia academica, profissional
  - ➡ projetos anteriores
  - ➡ teste técnico

- boa colocação no Kaggle
- perfil GitHub
- certificações e qualificações no Coursera, Udacity
- experiencias passadas
- MSc e PhD

- Psicólogos,
- Sociólogos,
- Historiadores,
- Linguistas
- ...podem ser cientistas de dados???
- **SIM!!!!**

- comportamento:
  - na atividade: inventivo, incansável, inquisitivo
  - na comunicação:  
assertivo com humildade, e  
articulado sem exagero
  - no relacionamento: confiável, solícito

- Checklist qualitativo: conversar sobre uma situação problema, e avaliar...
  - ➔ que dúvidas teve
  - ➔ como abordaria o problema
  - ➔ que tipo de soluções
  - ➔ pontos de atenção levantados

- **basic exploratory analysis:** understand the variables involved, distributions and correlations
- **outlier detection and imputation:** some of the columns (eg. age) contain outliers and missing values
- **feature encoding:** OHE (one-hot-encoding) or label-encoding are options for categorical columns
- innovative **feature engineering:** hopefully make the recruiter surprised by features created that were unthought of
- **validation strategy:** train/test split, K-Fold or more advanced techniques
- **metric:** explain the metric chosen
- **algorithm:** linear versus non-linear algorithm. understand the reasons for candidate's choice.
- **final test:** run the proposed algorithm on new data

## Data scientist recruiting interview

### A) Please do not send this notebook to the candidate - recruiter eyes only

The dataset 'df\_train.csv' is based on [Airbnb recruiting competition](#). In essence, the prediction objective is very similar, which is to predict country destination of user's first booking. However, this version of the dataset was simplified such that the candidate can perform an end-to-end datascience demonstration of skills after a 3-4 hours work.

#### Evaluation checklist:

1. basic exploratory analysis: understand the variables involved, distributions and correlations
2. outlier detection and imputation: some of the columns (eg. age) contain outliers and missing values
3. feature encoding: OHE (one-hot-encoding) or label-encoding are options for categorical columns
4. innovative feature engineering: hopefully make the recruiter surprised by features created that were unthought of
5. validation strategy: train/test split, K-Fold or more advanced techniques
6. metric: explain the metric chosen
7. algorithm: linear versus non-linear algorithm. understand the reasons for candidate's choice.
8. final test: run the proposed algorithm on new data

### B) Material to be sent to candidate:

#### email text for the candidate:

---

The objective of this exercise is to predict the outcome in the last column 'country\_destination'. This is multi-label classification problem with four labels:

- Python:  
pandas, numpy, matplotlib, seaborn  
scikit-learn  
xgboost, lightGBM  
tensorflow/Keras
- R  
dplyr, ggplot2, caret...
- SQL
- KNIME
- MATLAB, SAS, SPSS, Mathematica,...

# Quanto custa um profissional?

2017

| Cargo   | 2016                    | 2017                    | %     |
|---|-------------------------|-------------------------|-------|
| <b>BIG DATA (BIG DATA) (B)</b>                            |                         |                         |       |
| Engenheiro de Big Data - <i>Big Data Engineer</i>         |                         |                         |       |
| R\$ 15.000 - R\$ 35.000                                   | R\$ 15.000 - R\$ 40.000 | 10,0%                   |       |
| Gerente de Big Data - <i>Big Data Manager</i>             | R\$ 8.000 - R\$ 18.000  | R\$ 10.000 - R\$ 20.000 | 15,4% |
| Administrador de Big Data - <i>Big Data Administrator</i> | R\$ 6.000 - R\$ 13.000  | R\$ 6.100 - R\$ 13.000  | 0,5%  |
| Arquiteto de Dados - <i>Data Architect</i>                | R\$ 8.000 - R\$ 18.000  | R\$ 8.160 - R\$ 18.360  | 2,0%  |
| ★ Cientista de Dados - <i>Data Scientist</i>              | R\$ 12.000 - R\$ 25.000 | R\$ 12.000 - R\$ 28.000 | 8,1%  |
| Analista de BI - <i>Business Intelligence Analyst</i>     | R\$ 6.000 - R\$ 13.000  | R\$ 6.600 - R\$ 14.300  | 10,0% |

2018

| CARGO (JOB TITLE)   | 2017                    | 2018                    | %      |
|---|-------------------------|-------------------------|--------|
| <b>BIG DATA (BIG DATA) (C)</b>  |                         |                         |        |
| Especialista de Big Data / Cientista de dados - <i>Big Data / Data Scientist specialist</i> |                         |                         |        |
| R\$ 11.500 - R\$ 18.000   | R\$ 12.000 - R\$ 22.000 | 15,25%                  |        |
| ★ Analista de Big Data/ Cientista de dados - <i>Big Data / Data Scientist analyst</i>       | R\$ 5.300 - R\$ 10.500  | R\$ 5.500 - R\$ 12.500  | 13,92% |
| Especialista de BI - <i>Business Intelligence specialist</i>                                | R\$ 8.000 - R\$ 16.000  | R\$ 10.000 - R\$ 16.500 | 10,42% |
| Analista de BI - <i>Business Intelligence analyst</i>                                       | R\$ 4.000 - R\$ 10.000  | R\$ 4.500 - R\$ 11.000  | 10,71% |

2017

**(B) Aos salários pode ser acrescentado o percentual abaixo, de acordo com as habilidades específicas:**

|                                    |     |                      |    |
|------------------------------------|-----|----------------------|----|
| ETL .....                          | 6%  | Oracle database..... | 7% |
| Microsoft SQL Server database..... | 10% |                      |    |

2018

**(C) Aos salários pode ser acrescentado o percentual abaixo, de acordo com as habilidades específicas:**

|                    |    |                                    |    |
|--------------------|----|------------------------------------|----|
| ETL .....          | 8% | Power BI .....                     | 8% |
| SQL Server.....    | 5% | Microstrategy.....                 | 8% |
| Hadoop .....       | 8% | Oracle database.....               | 8% |
| R .....            | 8% | Python.....                        | 8% |
| Pig .....          | 8% | PowerCenter.....                   | 8% |
| Datawarehouse..... | 8% | BusinessObjects.....               | 8% |
| Teradata.....      | 8% | SSIS - Integrations Services ..... | 8% |
| SAS.....           | 8% | SSRS - Reports Services .....      | 8% |
| Basona.....        | 8% | SSAS - Analysis Services .....     | 8% |
| Tableau.....       | 8% |                                    |    |

## Data Scientist Salaries

2,951 Salaries Updated Feb 23, 2018

All Industries

All Company Sizes

All Years of Experience

Average Base Pay

**\$120,931 /yr**



Additional Cash Compensation [?](#)

Average \$11,772

Range \$4,006 - \$27,409

How much does a Data Scientist make?

The national average salary for a Data Scientist is \$120,931 in United States. Filter by location to see... [More](#)

## Data Engineer Salaries

1,170 Salaries Updated Feb 23, 2018

All Industries

All Company Sizes

All Years of Experience

Average Base Pay

**\$137,776 /yr**



Additional Cash Compensation [?](#)

Average \$10,000

Range \$xx,xxx

How much does a Data Engineer make?

The national average salary for a Data Engineer is \$137,776 in United States. Filter by location to see Data... [More](#)

Fonte: Glassdoor  
fev/2018

# O GESTOR

- Gestão:
  - da equipe: contratações, feedback, liderança ...
  - do processo, do workflow
  - das interfaces: horizontais e verticais
- Gestão:
  - da comunicação
  - das expectativas
- Competências:
  - Conhecimento técnico
  - Capacidade gerencial

- Gestor já botou a mão na massa, e assim continua:



Andrej Karpathy   
@karpathy

Follow

v

Excited to join Tesla as the Director of AI!



## Tesla hires deep learning expert Andrej Karpathy to lead Autopilot

Tesla has hired deep learning and computer vision expert Andrej Karpathy to lead its Autopilot role. Karpathy most recently held a role as a research scientist at Facebook's AI Research (FAIR) group.

[techcrunch.com](http://techcrunch.com)

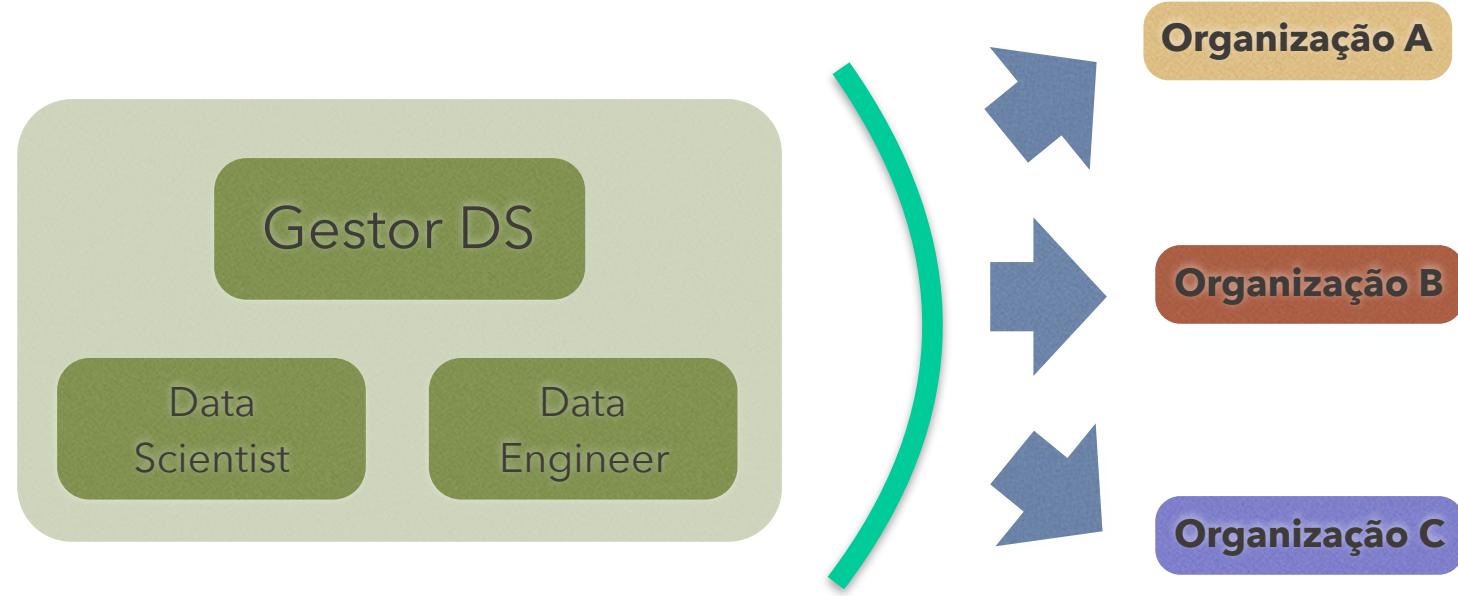


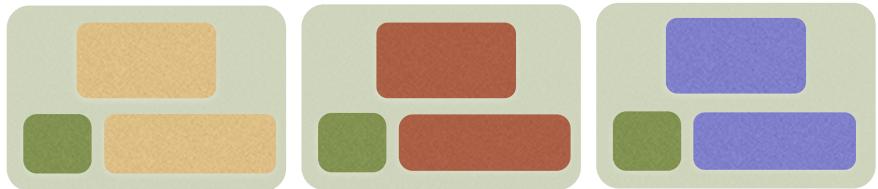
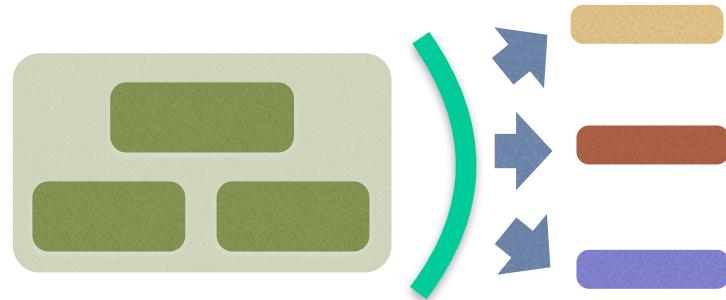
Facebook AI Research (FAIR)

I am Chief AI Scientist for Facebook AI Research (FAIR), joining Facebook in December 2013. I am also a Silver Professor at New York University on a part time basis, mainly affiliated with the NYU Center for Data Science, and the Courant Institute of Mathematical Science.

# **MODELOS ORGANIZACIONAIS**

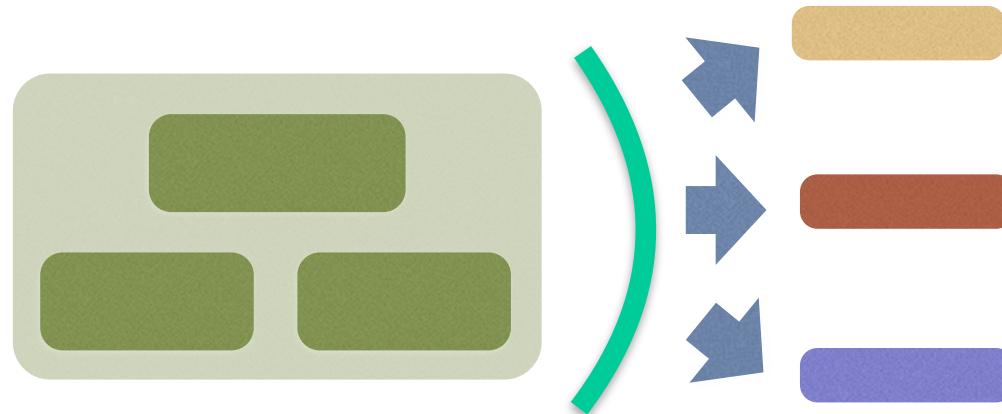
- Empresas pequenas:
  - Data joker:  $\frac{1}{2}$  Data engineer +  $\frac{1}{2}$  Data Scientist
- Empresas médias:
  - Um ou dois data engineers
  - Um ou dois data scientists
- Empresas grandes:
  - Dois ou mais data engineers
  - Dois ou mais data scientists
  - Um gestor





- Maior intercambio, troca de informações entre DS's e entre DS/ DE
- Uniformidade entre projetos
- Menor duplicação de esforços
- DS/DE com maior autonomia

- Maior vazão de projetos
- Maior visibilidade
- Cliente com melhor entendimento do potencial do DS



- Atendimento a demandas
- Engajamento em projetos
- Treinamento
- Ofertas pró-ativas

# GESTÃO NO DIA-A-DIA

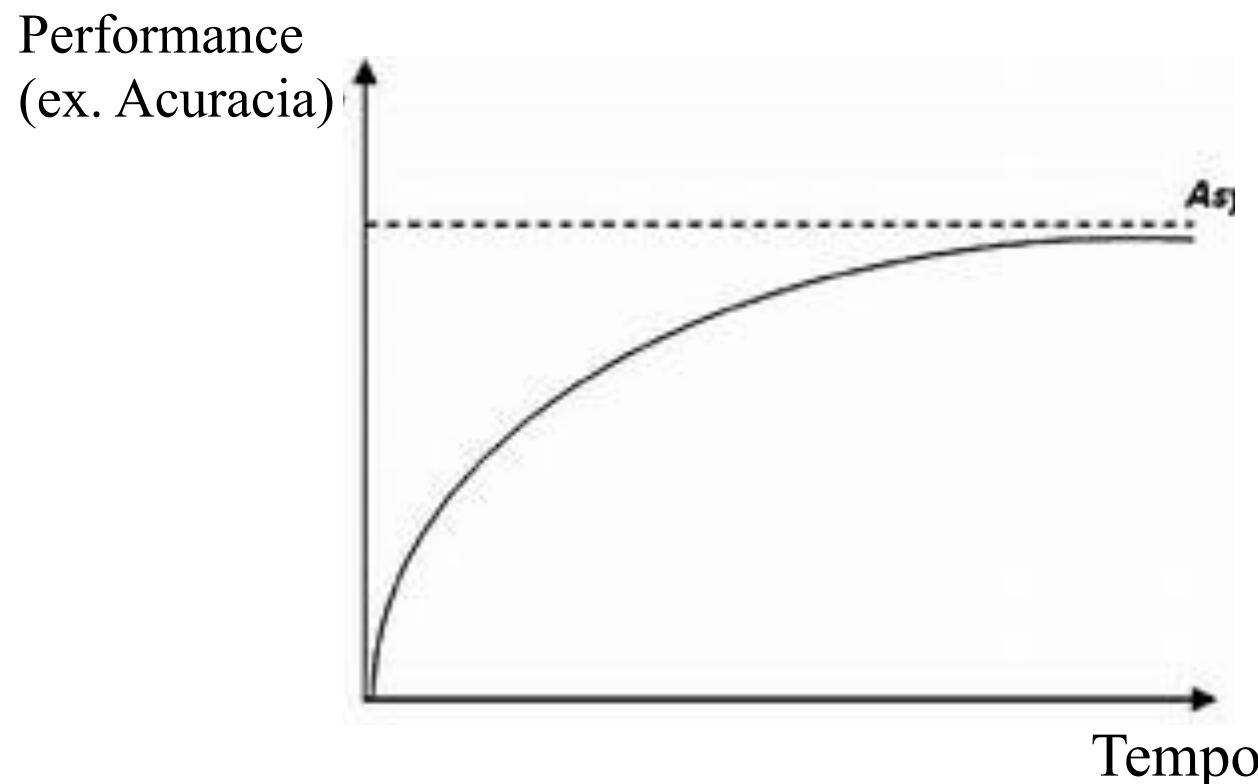
- reuniões semanais
- reuniões diárias
- reuniões
  - internas ao grupo
  - externas, por exemplo clientes internos
  - com outras empresas

- Dois ou mais projetos simultâneos por DS
- Dois ou mais projetos simultâneos por DE
- Tempo semanal para estudo, coisas novas p. ex:
  - Novas arquiteturas de redes neurais
  - Avanços em frameworks para pipelines
  - Métodos Bayesianos
  - Sistemas de Recomendação
  - Algoritmos de Clusterização
  - Produtização de algoritmos
  - ...

| SEG                                | TER    | QUA                                | QUI                        | SEX                        | SAB        |
|------------------------------------|--------|------------------------------------|----------------------------|----------------------------|------------|
| Preparação<br>de aulas /<br>Estudo |        | Prospecção<br>de novos<br>clientes |                            |                            | MBA<br>FGV |
| FGV                                | Estudo | FGV                                | Projetos<br>de<br>clientes | Projetos<br>de<br>clientes |            |

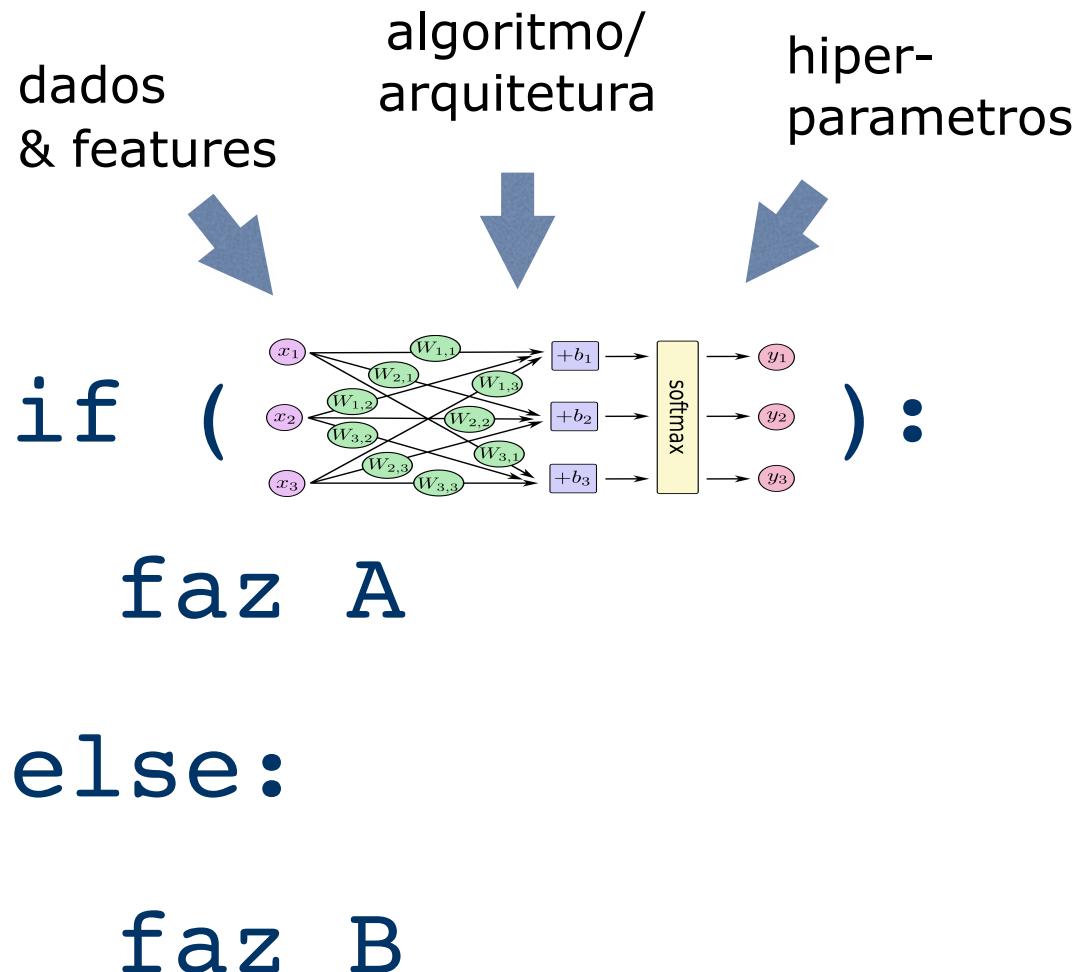
- Errar é ...
- ... necessário
- Ideias novas, algumas boas outras não...
- ... mas sempre  novas
- Compartilhar  com colegas, ouvir feedback

- Tradeoff complexidade/desempenho vs. tempo
  - Quanto mais tempo, mais testes realizo, maior desempenho consigo
  - Quanto mais tempo, menos tempo tenho para o algoritmo produza algo util.

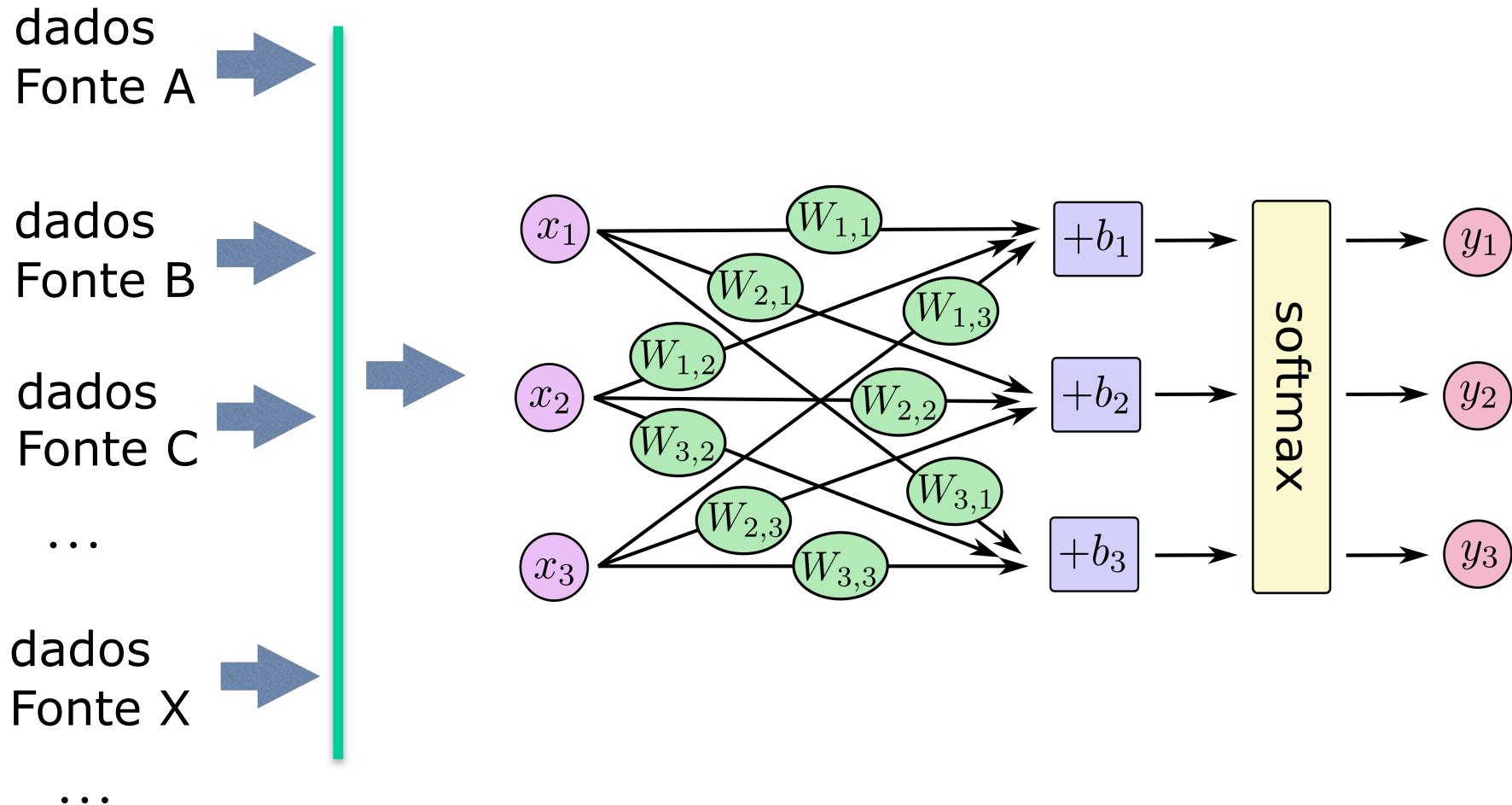


# **BUGS E DÉBITOS TÉCNICOS EM CIÊNCIA DE DADOS**

```
if condição:  
    faz A  
  
else:  
    faz B
```



Code = data



## Hidden Technical Debt in Machine Learning Systems

**D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips**  
{dsculley, gholt, dg, edavydov, toddphillips}@google.com  
Google, Inc.

**Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison**  
{ebner, vchaudhary, mwy, jfcrespo, dennison}@google.com  
Google, Inc.

### Abstract

Machine learning offers a fantastically powerful toolkit for building useful complex prediction systems quickly. This paper argues it is dangerous to think of these quick wins as coming for free. Using the software engineering framework of *technical debt*, we find it is common to incur massive ongoing maintenance costs in real-world ML systems. We explore several ML-specific risk factors to account for in system design. These include boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, configuration issues, changes in the external world, and a variety of system-level anti-patterns.

- Possíveis causas:
  - Testes incompletos
  - Complexidade do sistema
  - Dependencias, bola de espuague
  - Legibilidade, documentação deficiente
  - Design anti-patterns, smells
  - Pressão do deadline
- Possíveis ações:
  - Testar
  - Refatorar: código, interfaces e a nível de projeto
  - Redução da complexidade
  - Melhorar legibilidade e documentação

- Emaranhado por natureza:
  - mudanças isoladas podem afetar quase todo o resto
  - CACE: changing anything changes everything
  - Hiperparametros, dados novos, novo algoritmos, tudo pode melhorar algo e piorar outros.
- Hidden Feedback loops:
  - A ação de um sistema de IA acaba influenciando os dados de treinamento do próprio ou de outro sistema
  - Por bugs ou não

- Consumidores não declarados
  - Sistemas downstream “acostumados” com um padrão de outputs.
  - Uma melhoria no output de um sistema de ML pode causar problemas nesses sistemas downstream.
- Dependencia dos dados:
  - dados = código ...
  - Mais features → melhor modelo ... OK
  - Mais features → mais manutenção na produção ... OK?

- Calibração e correção de resultados
  - Calibrações forçadas segundo regras, com o objetivo de corrigir erros pontuais
  - Engessa os resultados (para essas previsões)
- Escolhas para pipeline:
  - SQL versus Python versus R ...
  - ... versus linguagem xyz
  - Cloud versus in-premises

- Consequências:
  - Glue code
  - Pipeline jungles: múltiplas linguagens, múltiplas fontes, múltiplos joins,

- simplificar pipelines, workflows
- testar: testes unitários, integração
- monitorar: estágios do pipeline, resultados
- Teste e monitoramento no **código e dados**

- bugs típicos:
  - travam o sistema, geram indisponibilidade
  - mensagens de erro
  - Bugs lógicos: resultados improváveis ou impossíveis
- fora esses, bugs em data science:
  - São mais sutis, e se manifestam sob ...
  - ... modelos sub-ótimos
  - ... discrepancia entre validação e resultados da produção

- Leak
- Bias variance
- Erros de amostragem
- Simpson's Paradox
- Qualidade do label
- Diferenças na extração entre validação e produção