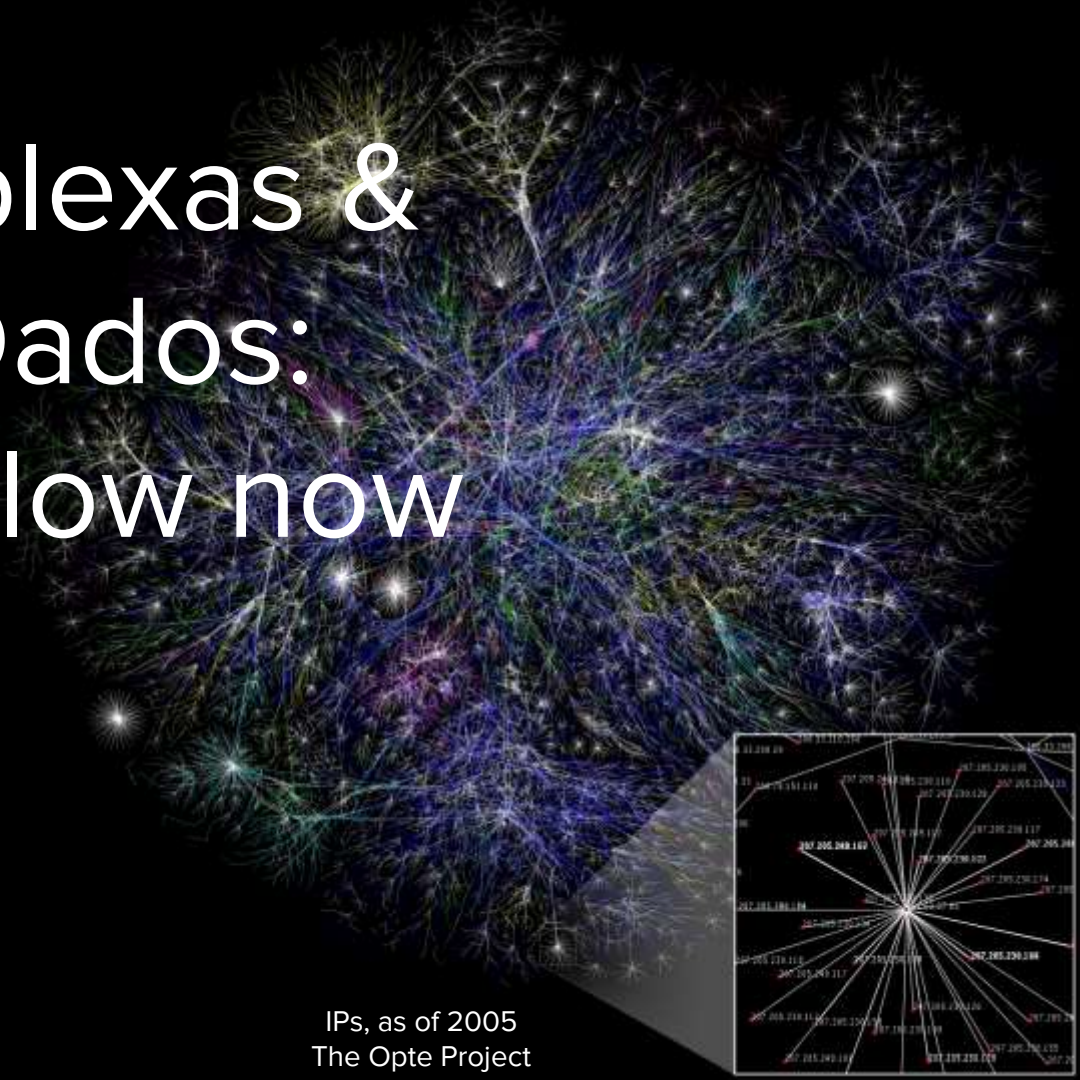
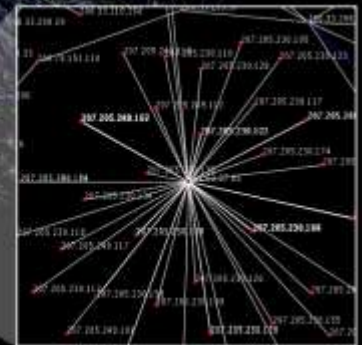


Redes Complexas & Ciência de Dados: juntos e shallow now

UFABC
19.06.2019



IPs, as of 2005
The Opte Project



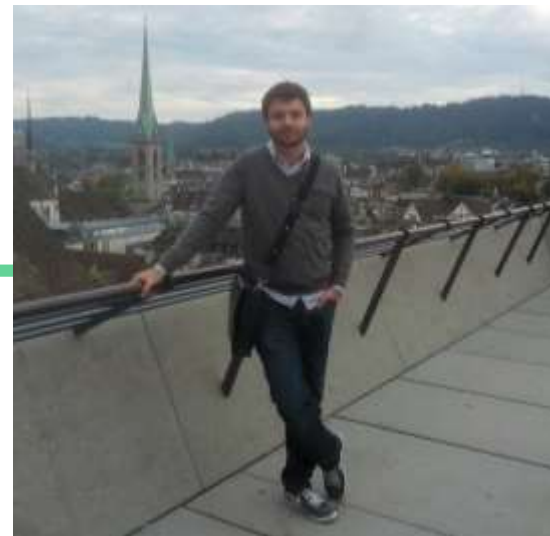
Sobre o Palestrante



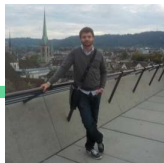
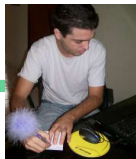
Unicamp
2005-2008
Matemática Aplicada e Computacional



USP
2009-2010
Mestrado em Bioinformática



ETH Zurich
2011-2015
Doutorado em Física
Computacional



2010-2011
Validação de modelos


SYSTEMMORPH®
2015
Business Analyst/Dev


2016 (3M!)
Data Scientist

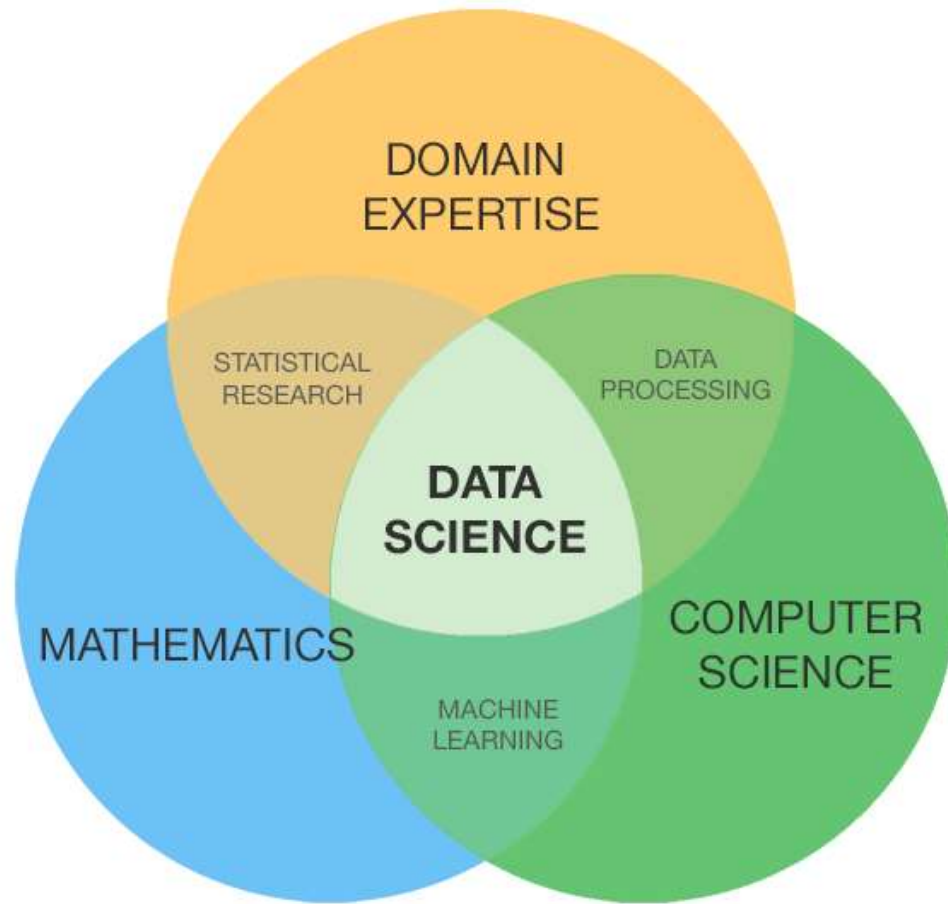


2016-
Cientista de Dados

CD Sr
Coordenador de CD
Especialista em CD

Ciência de Dados

“muito mais que um monte de dados”



*Source: Palmer, Shelly. Data Science for the C-Suite.
New York: Digital Living Press, 2015. Print.*



Backblaze Average Cost per GB for Hard Drives

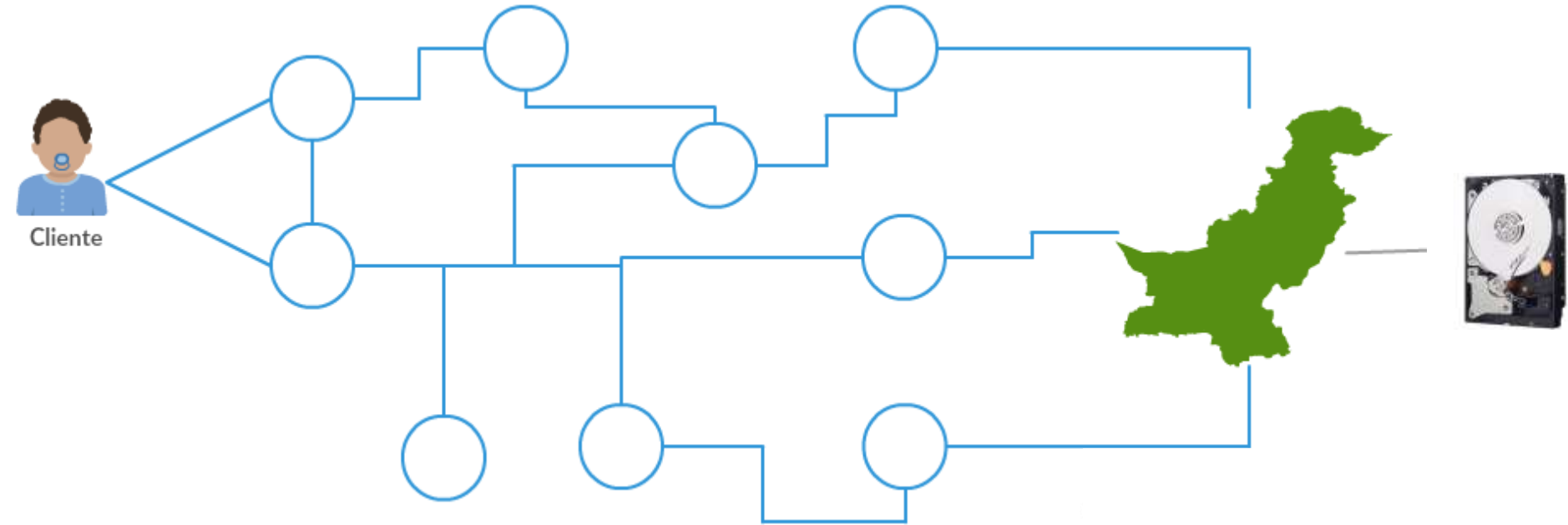
By Quarter: Q1 2009 - Q2 2017



Nós vivemos em um mundo interconectado...



... que depende de alguns pontos principais.



Uma rede complexa de interações econômicas!

Thailand's devastating floods are hitting PC hard drive supplies, warn analysts

Country is centre of world hard drive manufacture but factories have been seriously affected by floods, which could affect supplies early in 2012

Charles Arthur and
agencies

✉ @charlesarthur

Tue 25 Oct 2011 06:41 BST



▲ Thai residents make their way through a flooded street in Pathum Thani, Thailand. Photograph: Daniel Berehulak/Getty Images

The floods in Thailand which have caused widespread chaos in the country, leading to the deaths of more than 350 people after months of unusually heavy rainfall, will also have effects further afield, in a sign of the connectedness of the world economy.

Thailand's devastating floods are hitting



Personal Backup

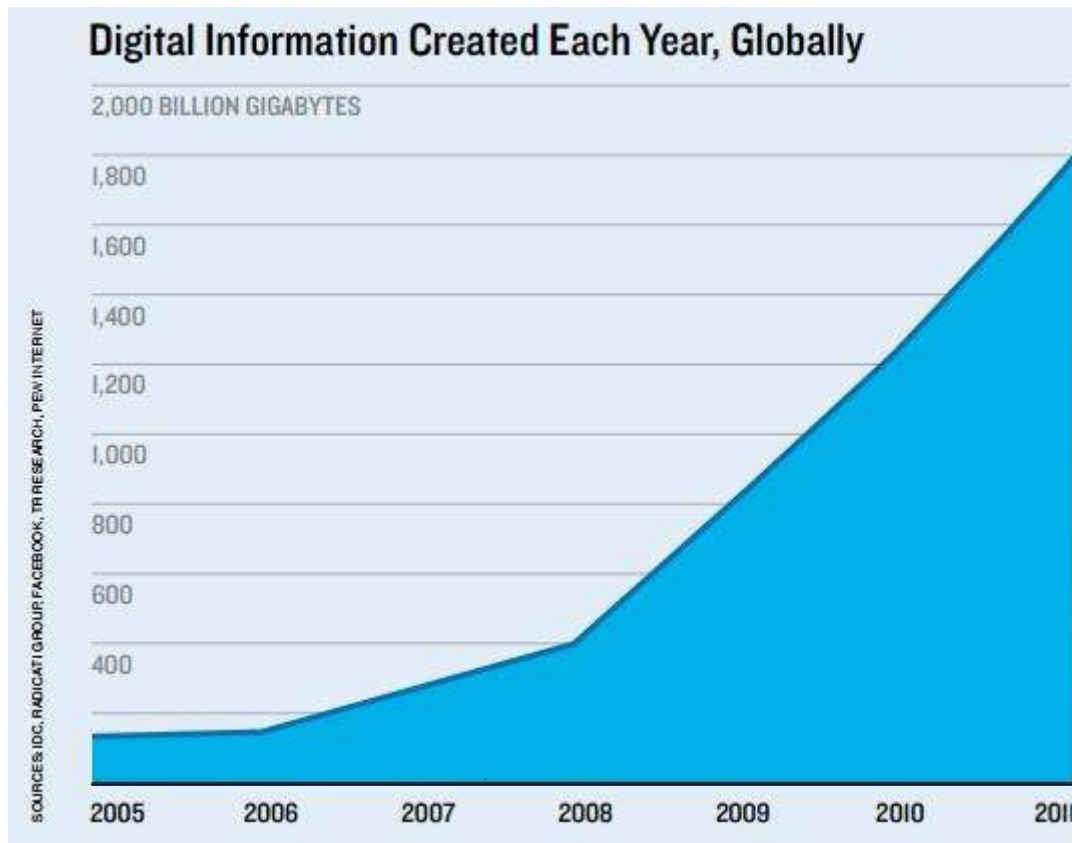
Busin

We didn't plan for rain in Thailand

We started drive farming in November 2011. The reason was simple, the supply of the 3TB hard drives used in our [Storage Pods](#) had dried up – or more correctly was under water. The tragic flooding in Thailand began in August 2011 and by early-October had submerged houses, schools and factories. Over 800 people died and many more were homeless and hungry, with over 1 million people thrown out of work. As the water receded, the human cost of the flooding was obvious and the economic impact was slowly coming into focus. In late-October, it was estimated that up to 50% of the worldwide hard drive manufacturing capacity was lost or damaged. The impact of the lost capacity was immediate as hard drive prices nearly tripled overnight.

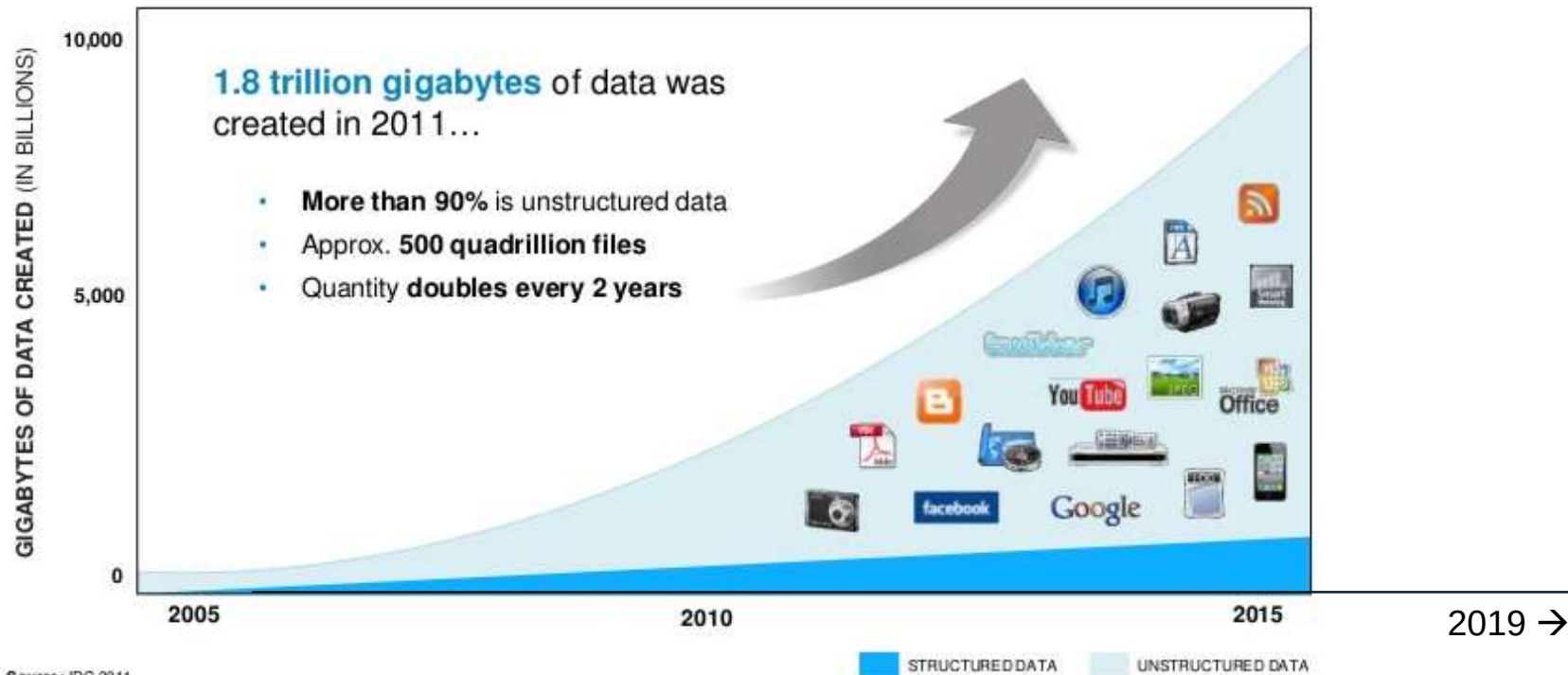
The floods in Thailand, which have caused widespread chaos in the country, leading to the deaths of more than 350 people after months of unusually heavy rainfall, will also have effects further afield, in a sign of the connectedness of the world economy.

Temos muitos dados

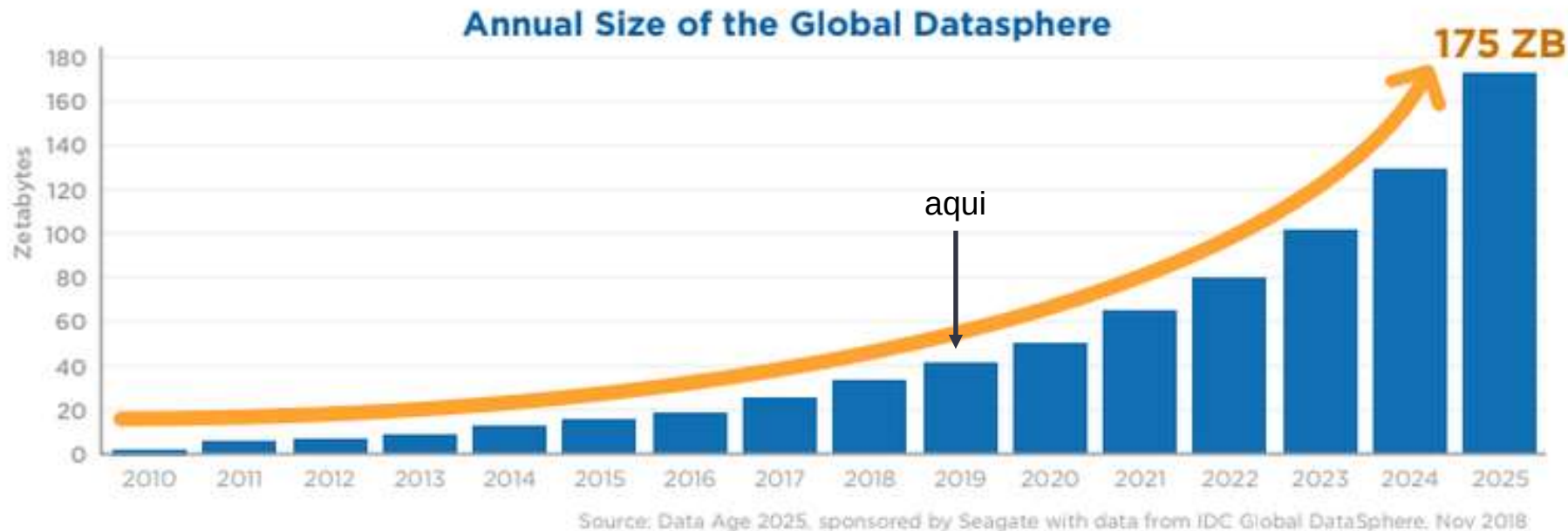


2019 →

Na verdade, muito mais

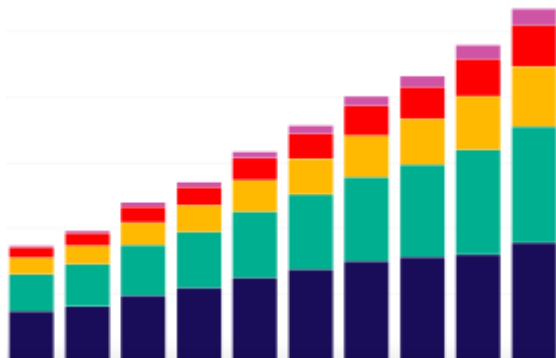


E a quantidade está só aumentando

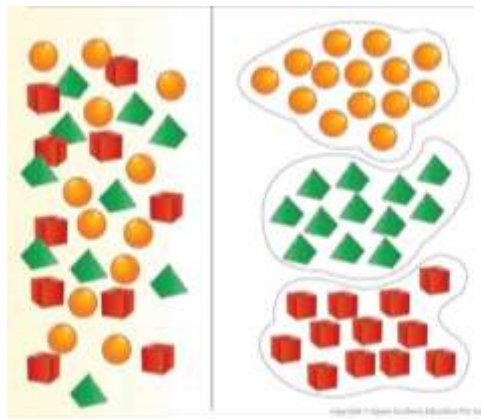


Principais abordagens da Ciência de Dados

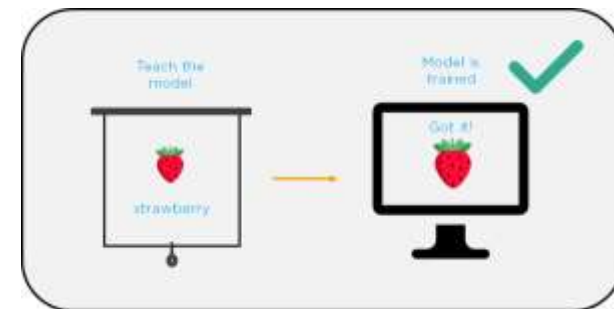
Exploratória



Não-supervisionado

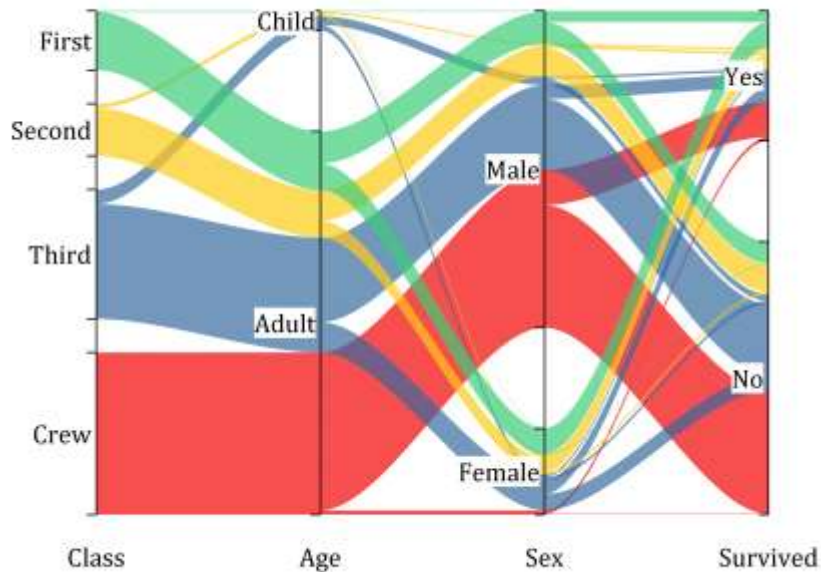


Supervisionado

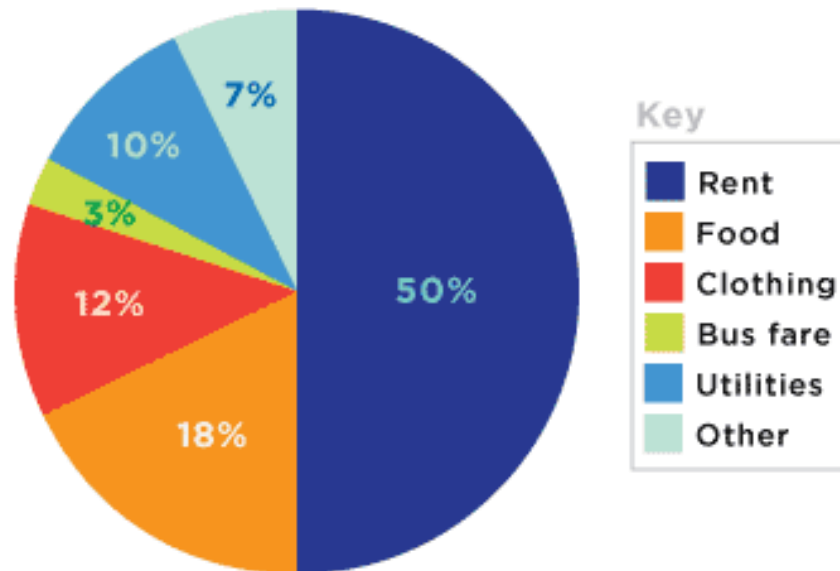


Análise Exploratória de Dados

A arte de fazer gráficos e sumarizações para entender o que está acontecendo...



Tina's Monthly Living Expenses



Saber SQL é muito importante!

```
1 SELECT COUNT(*) as nbr,  
2       dest  
3 FROM flight_delays  
4 GROUP BY dest  
5 HAVING COUNT(*) > 10000;
```

Result 1 Messages

nbr	dest
19272	ANC
21588	ELP
10273	CAE
60050	BNA
10707	SGF
15901	KOA
10660	CID
57537	SMF
27522	OMA
14025	LGB
27925	TUS
21810	SDF
25275	OGG

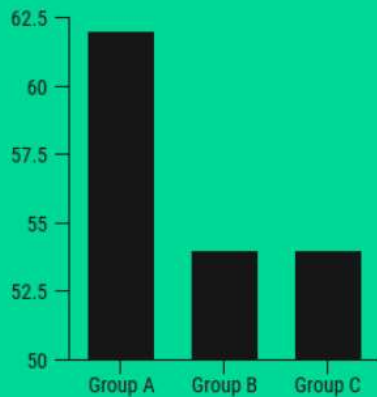
<https://www.khanacademy.org/computing/computer-programming/sql>

ou não

1

OMITTING THE BASELINE

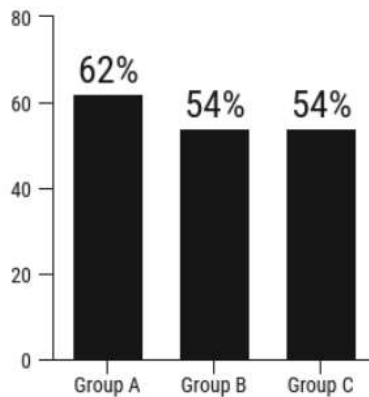
In most cases, the baseline for a graph is 0. But writers can skew how data is perceived by making the baseline a different number. This is known as a “truncated graph”.



MISLEADING

- Starting the vertical axis at 50 makes a small difference between groups seem massive
- Group A looks much larger than Groups B and C

VS



ACCURATE



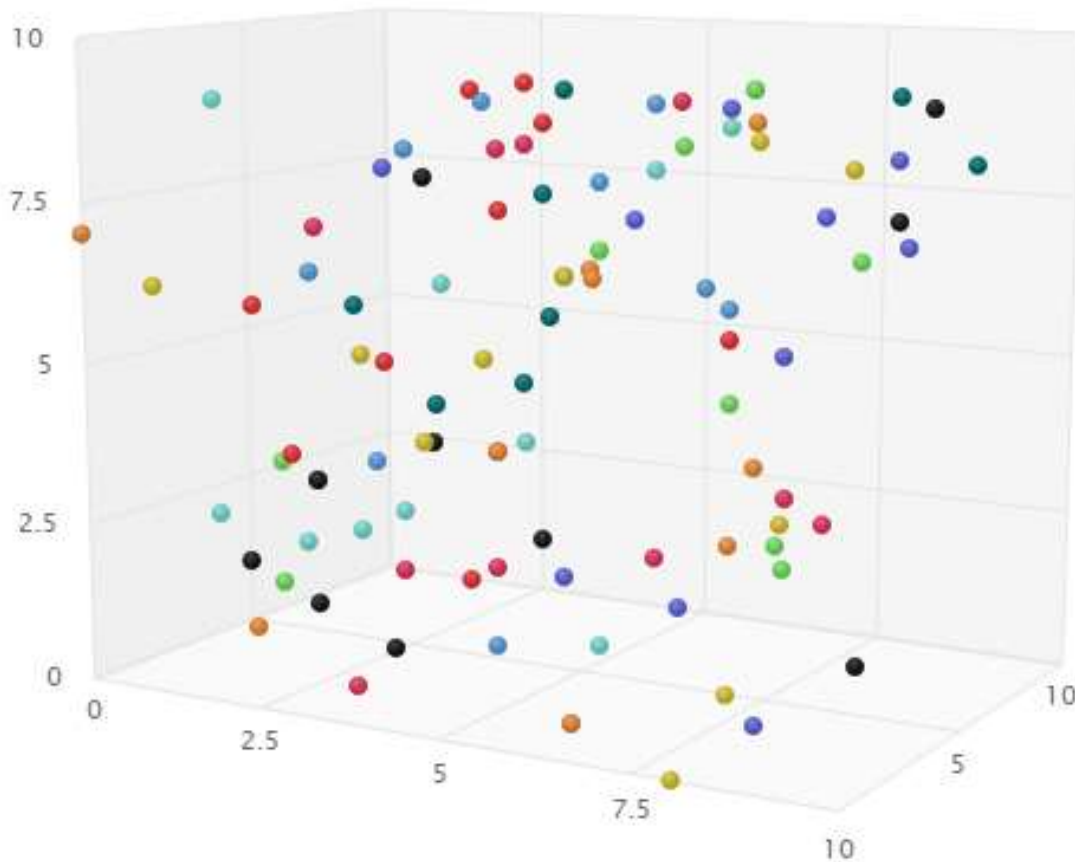
- Starting the vertical axis at 0 offers a more accurate depiction of the data
- The difference between the groups does not seem as dramatic

Aprendizado Não-supervisionado

Encontrar grupos/padrões nos dados

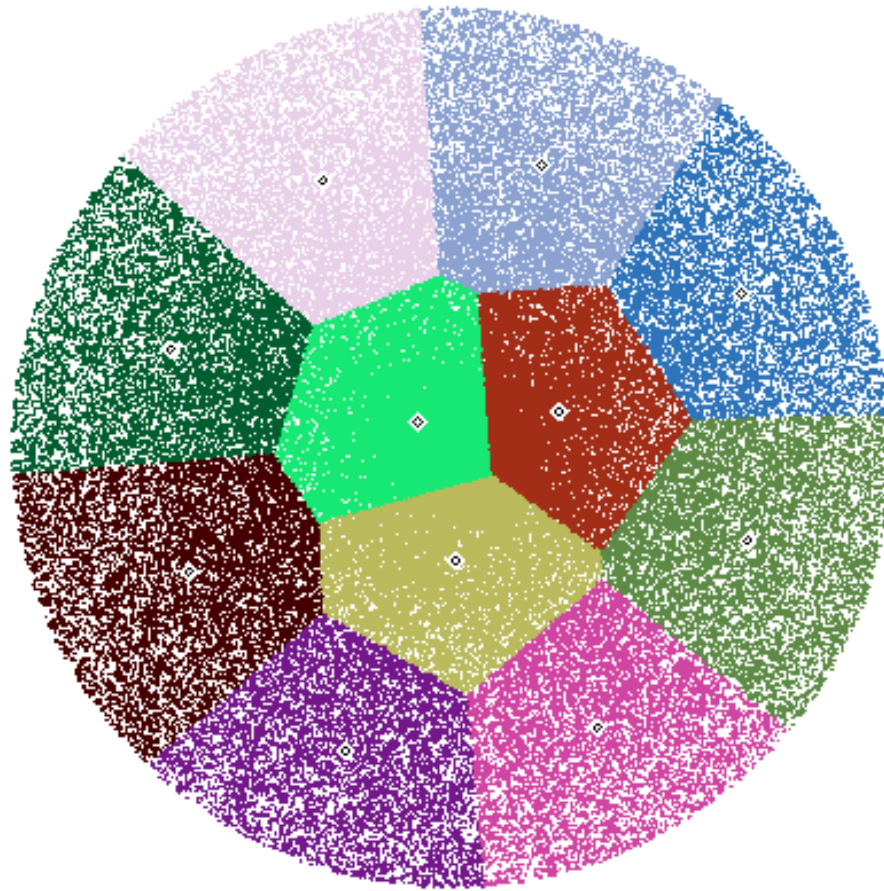
Name	Gender	Age	...	Kills
Tyrion	M	30		2
Jon	M	18		7
Robb	M	17		300
Sansa	F	16		5
...				

1) Colete seus dados



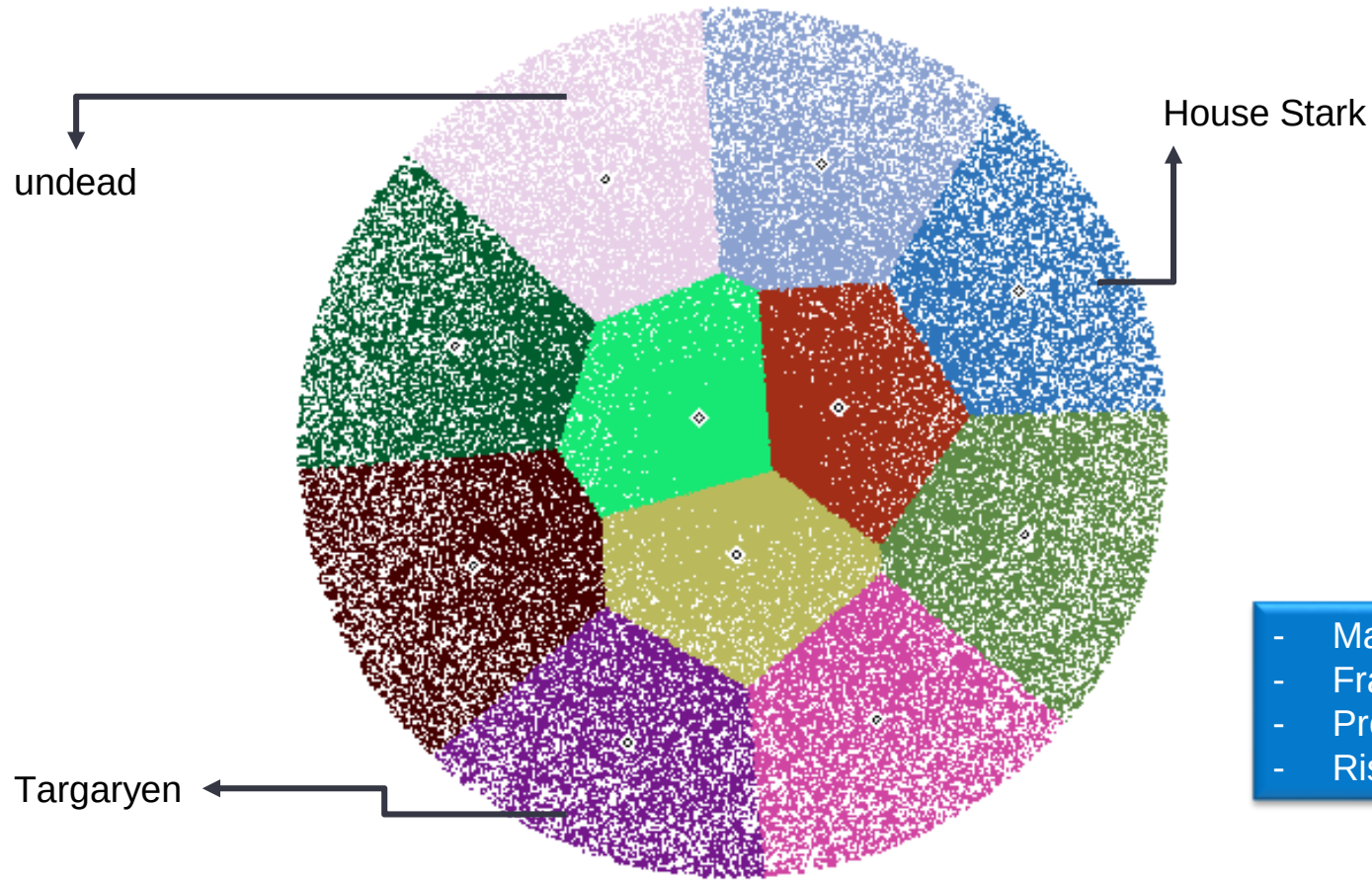
- One Hot Encoding
- Normalização
- PCA
- tSNE

2) Represente os dados na sua forma N-dimensional preferida



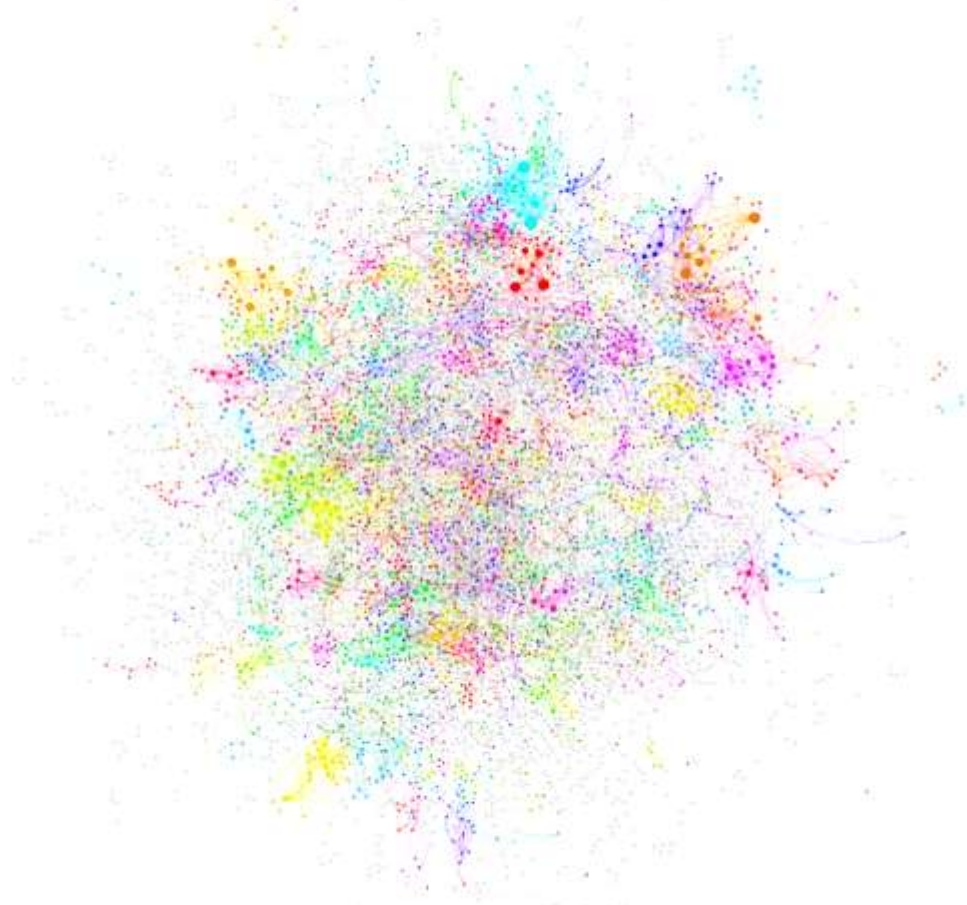
- K-means
- K-medians
- DBSCAN
- GMM

3) Encontre clusters com seu algoritmo preferido



- Marketing
- Fraude
- Produtos
- Riscos

4) Interprete seus clusters com seu parceiro de negócio favorito



5) Às vezes as coisas dão errado

6) Não desanime e tente outra vez!

- ~~One Hot Encoding~~
- Normalização
- PCA
- tSNE

- K-means
- ~~K-medians~~
- DBSCAN
- GMM

- Marketing
- Fraude
- ~~Produtos~~
- Riscos

Aprendizado Supervisionado

Construir uma máquina preditiva

Name	Gender	Age	...	Kills	Morre?
Tyrion	M	30		2	SIM
Jon	M	18		7	SIM
Robb	M	17		300	NÃO
Sansa	F	16		5	NÃO
...					...

1) Colete seus dados, juntamente com algo que você quer prever

Name	Gender	Age	...	Kills	Morre?
Tyrion	M	30		2	SIM
Jon	M	18		7	SIM
Robb	M	17		300	NÃO
Sansa	F	16		5	NÃO
...					...



2) Tenha vários tipos de colunas!

Name	Gender	Age	...	Kills
Tyrion	M	30		2
Jon	M	18		7
Robb	M	17		300
Sansa	F	16		5
...				

Morre?
SIM
SIM
NÃO
NÃO
...



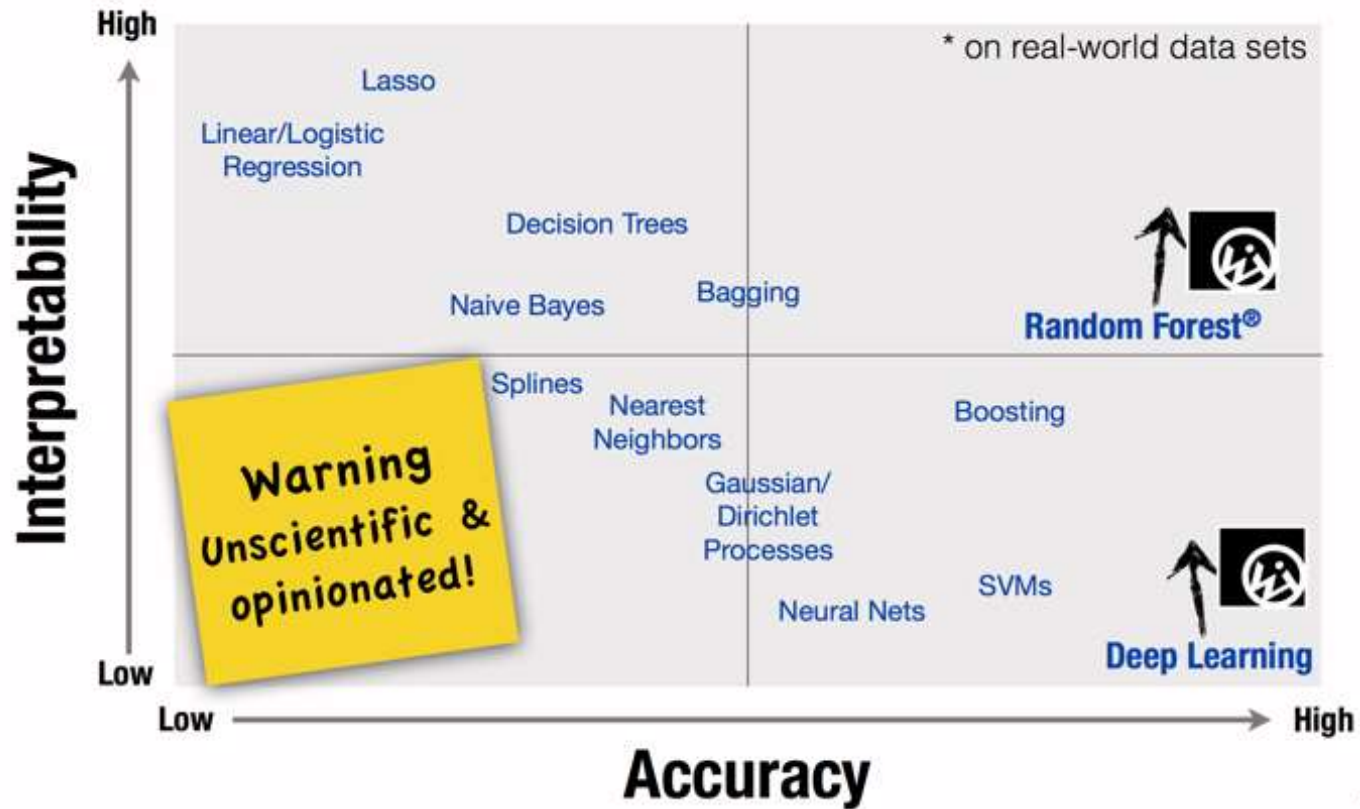
3) Tenha variabilidade na Resposta!



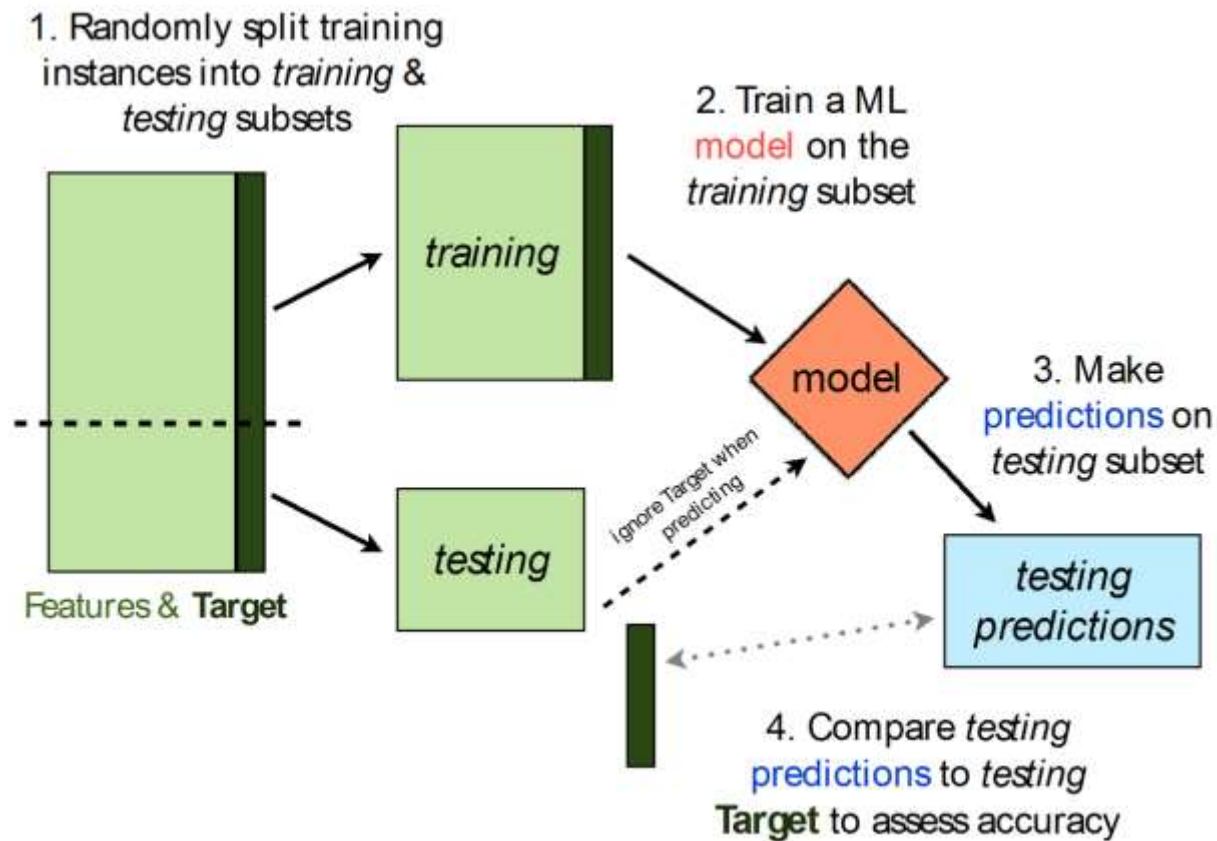
Name	Gender	Age	...	Kills
Tyrion	M	30		2
Jon	M	18		7
Robb	M	17		300
Sansa	F	16		5
...				

Morre?
SIM
SIM
NÃO
NÃO
...

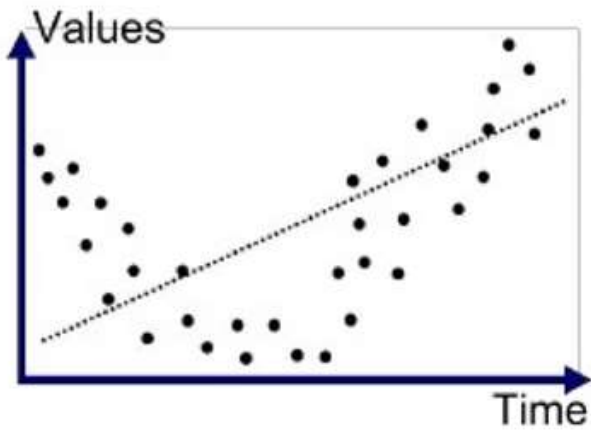
4) Tenha volume de dados!



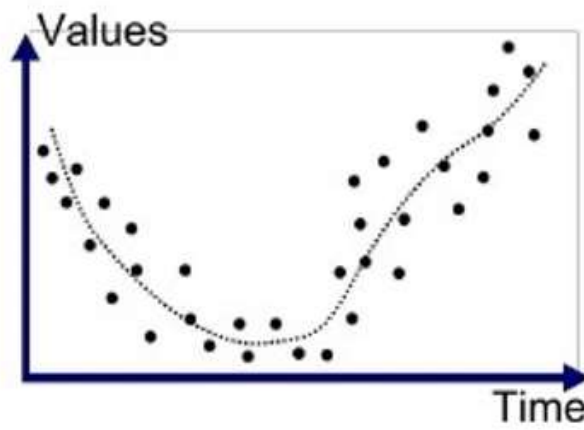
5) Treine seu algoritmo preditivo favorito nos dados!



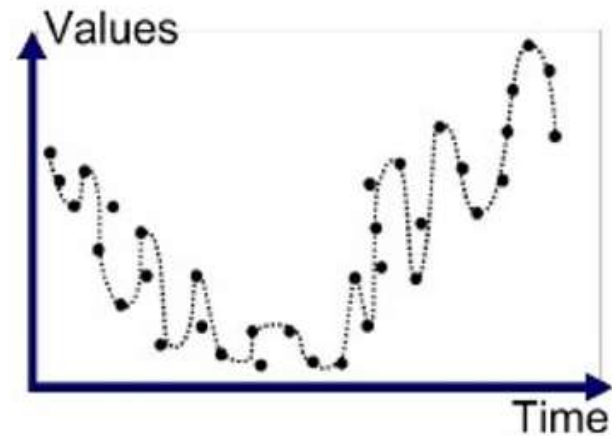
6) Mas acompanhe a performance em outros dados



Underfitted



Good Fit/Robust



Overfitted

7) Às vezes as coisas dão errado


learn

[Home](#)
[Installation](#)
[Documentation](#)
[Examples](#)

[Previous version](#)
[Next version](#)
[API](#)

scikit-learn v0.21.2
Other versions

Please **DO US** if you use the software.

1. Supervised learning

1. Supervised learning

1.1. Generalized Linear Models

- 1.1.1. Ordinary Least Squares
 - 1.1.1.1. Ordinary Least Squares Complexity
- 1.1.2. Ridge Regression
 - 1.1.2.1. Ridge Complexity
 - 1.1.2.2. Setting the regularization parameter: generalized Cross-Validation
- 1.1.3. Lasso
 - 1.1.3.1. Setting regularization parameter
 - 1.1.3.1.1. Using cross-validation
 - 1.1.3.1.2. Information-criteria based model selection
 - 1.1.3.1.3. Comparison with the regularization parameter of SVM
- 1.1.4. Multi-task Lasso
- 1.1.5. Elastic-Net
- 1.1.6. Multi-task Elastic-Net
- 1.1.7. Least Angle Regression
- 1.1.8. LARS Lasso
 - 1.1.8.1. Mathematical formulation
- 1.1.9. Orthogonal Matching Pursuit (OMP)
- 1.1.10. Bayesian Regression
 - 1.1.10.1. Bayesian Ridge Regression
 - 1.1.10.2. Automatic Relevance Determination - ARD
- 1.1.11. Logistic regression
- 1.1.12. Stochastic Gradient Descent - SGD
- 1.1.13. Perceptron
- 1.1.14. Passive Aggressive Algorithms
- 1.1.15. Robustness regression: outliers and modeling errors
 - 1.1.15.1. Different scenario and useful concepts
 - 1.1.15.2. RANSAC: Random Sample Consensus
 - 1.1.15.2.1. Details of the algorithm
 - 1.1.15.3. Theil-Sen estimator: generalized-median-based estimator
 - 1.1.15.3.1. Theoretical considerations
 - 1.1.15.4. Huber Regression
 - 1.1.15.5. Notes
- 1.1.16. Polynomial regression: extending linear models with basis functions

1.2. Linear and Quadratic Discriminant Analysis

- 1.2.1. Dimensionality reduction us
- 1.2.2. Mathematical formulation of
- 1.2.3. Mathematical formulation of
- 1.2.4. Shrinkage
- 1.2.5. Estimation algorithms

1.3. Kernel ridge regression

1.4. Support Vector Machines

- 1.4.1. Classification
 - 1.4.1.1. Multi-class classification
 - 1.4.1.2. Kernel support vector machines

1.3. Kernel ridge regression

1.4. Support Vector Machines

- 1.4.1. Classification
 - 1.4.1.1. Multi-class classification
 - 1.4.1.2. Scores and probabilities
 - 1.4.1.3. Unbalanced problems
- 1.4.2. Regression
- 1.4.3. Density estimation, novelty detection
- 1.4.4. Complexity
- 1.4.5. Tips on Practical Use
- 1.4.6. Kernel functions
 - 1.4.6.1. Custom Kernels
 - 1.4.6.1.1. Using Python functions as kernels
 - 1.4.6.1.2. Using the Gram matrix
 - 1.4.6.1.3. Parameters of the RBF Kernel
- 1.4.7. Mathematical formulation
 - 1.4.7.1. SVC
 - 1.4.7.2. NuSVC
 - 1.4.7.3. SVR
- 1.4.8. Implementation details

1.5. Stochastic Gradient Descent

- 1.5.1. Classification
- 1.5.2. Regression
- 1.5.3. Stochastic Gradient Descent for sparse data
- 1.5.4. Complexity
- 1.5.5. Stopping criterion
- 1.5.6. Tips on Practical Use
- 1.5.7. Mathematical formulation
 - 1.5.7.1. SGD
- 1.5.8. Implementation details

1.6. Nearest Neighbors

- 1.6.1. Unsupervised Nearest Neighbors
 - 1.6.1.1. Finding the Nearest Neighbors
 - 1.6.1.2. KDTree and BallTree Classes
- 1.6.2. Nearest Neighbors Classification
- 1.6.3. Nearest Neighbors Regression
- 1.6.4. Nearest Neighbor Algorithms

- 1.6.5. Nearest Centroid Classifier
 - 1.6.5.1. Nearest Shrunken Centroid
- 1.6.6. Neighborhood Components Analysis
 - 1.6.6.1. Classification
 - 1.6.6.2. Dimensionality reduction
 - 1.6.6.3. Mathematical formulation
 - 1.6.6.3.1. Mahalanobis distance
 - 1.6.6.4. Implementation
 - 1.6.6.5. Complexity

- 1.11.4.8. Multiclass Classifier
 - 1.11.4.8.1. Stochastic
 - 1.11.4.8.2. Subsampling
- 1.11.4.7. Interpretation
 - 1.11.4.7.1. Feature importance

- 1.11.5. Voting Classifier
 - 1.11.5.1. Majority Class Labels (Majority/Rand Voting)
 - 1.11.5.1.1. Usage
 - 1.11.5.2. Weighted Average Probabilities (Soft Voting)
 - 1.11.5.3. Using the VotingClassifier with GridSearchCV
 - 1.11.5.3.1. Usage

- 1.11.6. Voting Regressor

1.12. Multiclass and multilabel algorithms

- 1.12.1. Multilabel classification format
- 1.12.2. One-Vs-The-Rest
 - 1.12.2.1. Multiclass learning
 - 1.12.2.2. Multilabel learning
- 1.12.3. One-Vs-One
 - 1.12.3.1. Multiclass learning
- 1.12.4. Error-Correcting Output-Codes
 - 1.12.4.1. Multiclass learning
- 1.12.5. Multioutput regression
- 1.12.6. Multioutput classification
- 1.12.7. Classifier Chain
- 1.12.8. Regressor Chain

1.13. Feature selection

- 1.13.1. Removing features with low variance
- 1.13.2. Univariate feature selection
- 1.13.3. Recursive feature elimination
- 1.13.4. Feature selection using SelectFromModel
 - 1.13.4.1. L1-based feature selection
 - 1.13.4.2. Tree-based feature selection
- 1.13.5. Feature selection as part of a pipeline

1.14. Semi-Supervised

- 1.14.1. Label Propagation

1.15. Isotonic regression

1.16. Probability calibration

models (supervised)
neception

7) Não desanime e tente outra vez!

- 1.17.3. Regression
- 1.17.4. Regularization
- 1.17.5. Algorithms
- 1.17.6. Complexity
- 1.17.7. Mathematical formulation
- 1.17.8. Tips on Practical Use
- 1.17.9. More control with warm_start

Case 1 – Validação de Cheques

EXPERIÊNCIAS DIGITAIS
digital de ponta a ponta

DEPÓSITO DE CHEQUES

NO PASSADO

Depósito e verificação manuais



HOJE

Depósito pelo app



EM BREVE

Verificação utilizando
inteligência artificial

Próximo passo
para uma jornada
100% DIGITAL



Machine learning
para validação
formalística

Deep learning para
reconhecimento
de assinatura

Verificação de **100%** dos elementos
em **100% dos cheques**
Humano avalia somente dúvidas no modelo

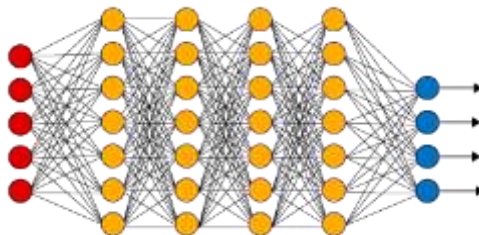
Case 1 – Validação de Cheques

Dados



Assinaturas
Verdadeiras e
Falsas

Algoritmo



Deep Learning

Objetivo



Marcar a
assinatura como
verdadeira ou
falsa no cheque

Realidade

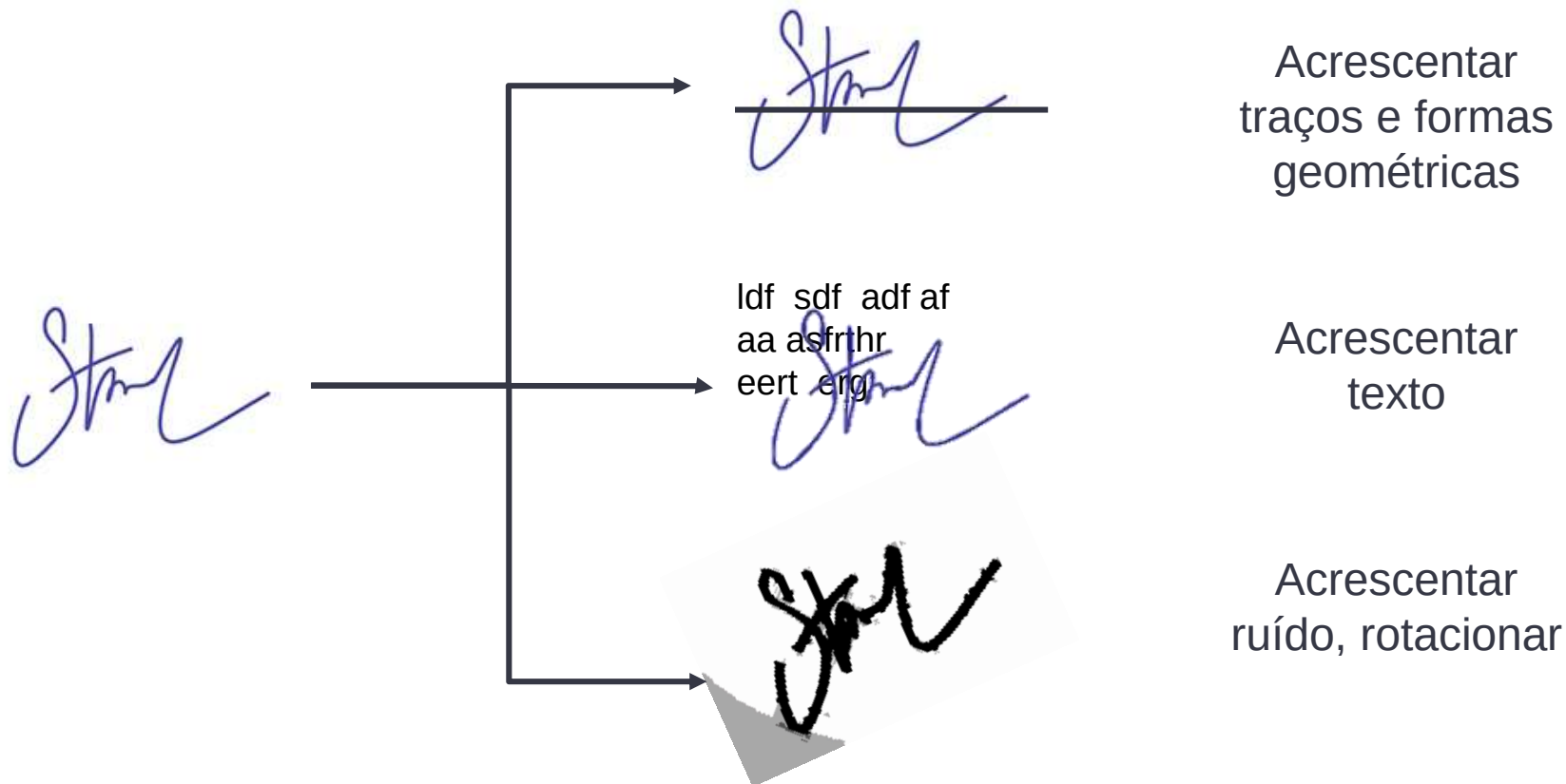
EXECUTED at Los Angeles, California, on the
10th day of October, 19 83.

PRODUCER
OPTIMUM PRODUCTIONS
Name of Company

By [Signature]
Name and Title: Michael Jackson, President

SCREEN ACTORS GUILD
By [Signature]
Assistant National Executive Secretary
Executed at Los Angeles, Calif. on the 12th day of December, 1983

O pulo do gato



Modificar os dados de treinamento para se aproximar da realidade

The usage of U-Net for pre-processing document images

Diandra A. A. Kubo, Tiago S. de Nazare, Priscila L. R. Aguirre, Bruno D. Oliveira, Felipe S. L. G. Duarte

Data Science Team Itaú Unibanco

{diandra.kubo, tiago.nazare, priscila.aguirre, bruno.domingues-oliveira, felipe.duarte}@itau-unibanco.com.br

Abstract—When processing documents in real-world scenarios, it is common to deal with artifacts that may hamper document analysis, such as stamps, noise and strange backgrounds. Aiming to mitigate these problems, we propose the use of U-Net, a very successful biomedical image segmentation network, for handwritten and machine text segmentation. In order to do so, we trained a model for each type of text. One of the main advantages presented is that the models are trained on artificial data, avoiding the wearisome task of data labeling. For the machine text segmentation model, we test its impacts on both word and character recognition when combined with the Tesseract OCR model. For the handwritten segmentation model, we present qualitative results. Initial experiments indicate that both models are able to improve results in their respective applications.

I. INTRODUCTION

Recent advances in machine vision techniques allowed the usage of models to solve real-world problems such as image classification [1], object detection [2], signature verification [3] and optical character recognition (OCR) [4]. Despite such progress, in most cases, the impacts that image quality can have on performance is overlooked, therefore models are still limited with regards to their applicability in practical situations where images obtained in unconstrained scenarios [5].

When a document is scanned by a sensor and converted to a digital image, some noise can be inherited in this process. The

and post-processing are the steps to be tackled when considering performance boosting for such applications. The major issue when using OCR techniques is the high variability in page layout and design, due to different document production processes [6].

When dealing with documents images, another recurring application is detecting the number of signatures in a document, finding their relative position on the page and who they belong to. This step depends on the correct extraction of possible signatures [8], detecting its locations and cleaning any noise surrounding them. On real-word applications, this is a complex problem [9], due to the variety of background and artifacts a document may present, such as stamps that might overlap the signature to be recognized.

Motivated by those gaps, in this paper, we propose a new framework for pre-processing document images using U-Net in two scenarios: (i) segmentation of machine-made text to improve OCR performance and (ii) extraction of handwritten text to help in applications such as signature verification. Our initial results show that U-Net can greatly improve in both character and word recognition when combined with Tesseract for the OCR task. Also, we present some images that indicate that this method can also be helpful when segmenting signatures (and other handwritten texts).

Case 2 – Prevendo quem morre no Titanic



Getting Started Prediction Competition

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics



Kaggle · 11,172 teams · Ongoing

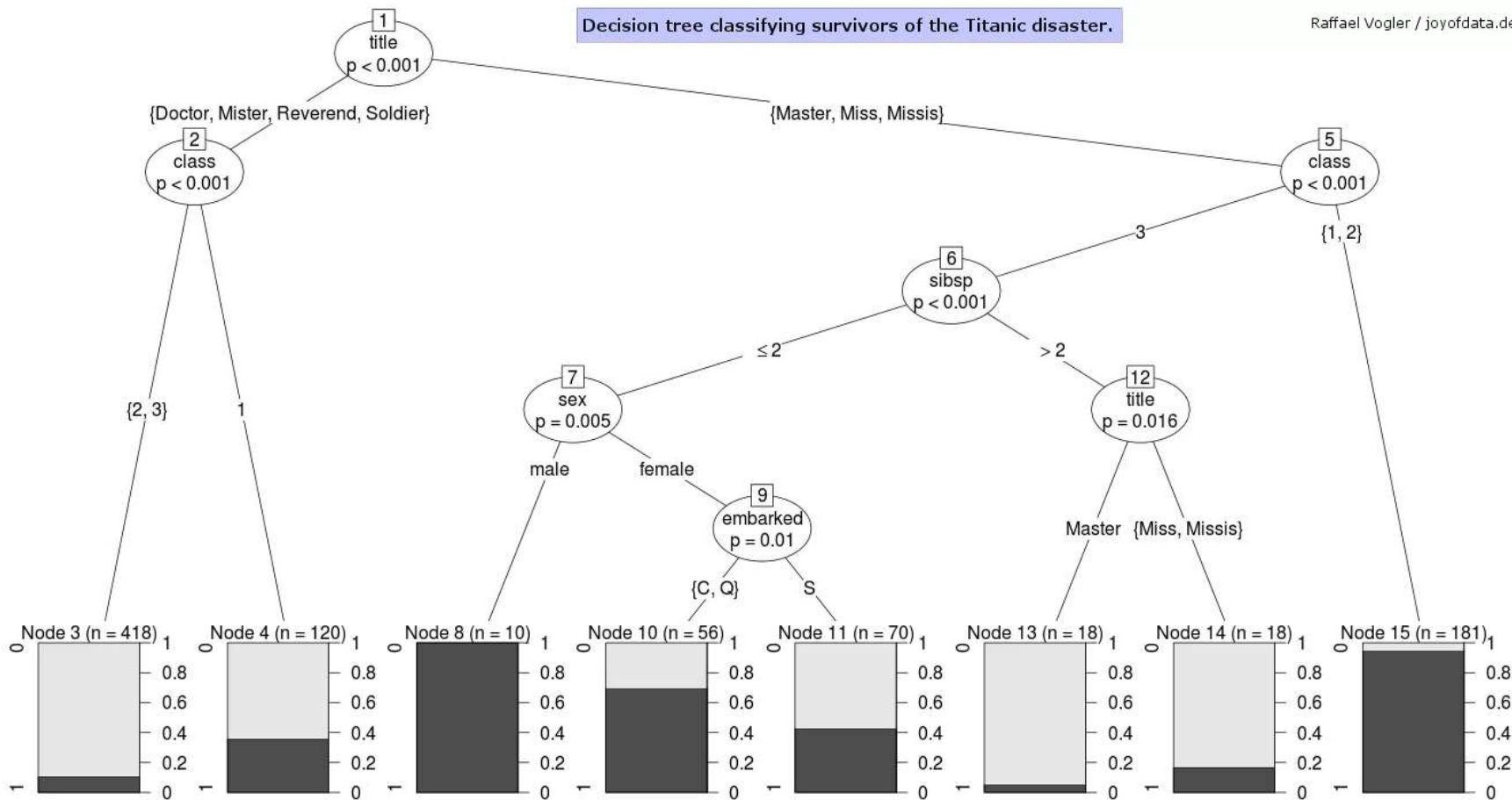
Dados

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	

Um possível algoritmo

Decision tree classifying survivors of the Titanic disaster.

Raffael Vogler / joyofdata.de



Para quem quiser aprender mais



Be in Demand

Learn the latest tech skills
to advance your career.

Start your future today with a Nanodegree program
in Data Science.



Explorar ▾

O que você deseja aprender? 

Navegar > Ciência de Dados > Análise de Dados

Programa de cursos integrados Ciência de Dados

Inicie uma carreira em Ciência de Dados. A ten-course introduction to data science, developed and taught by leading professors.

Cadastre-se gratuitamente
Inicia em jun 16

Auxílio financeiro disponível

245.380 já inscritos!

Data science courses on edX

Data science is a multifaceted field used to gain insights from complex data.

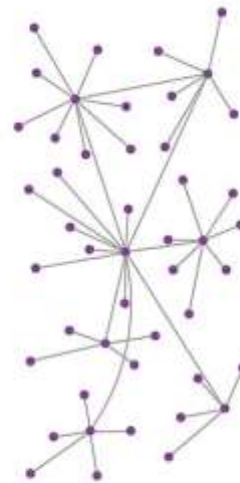
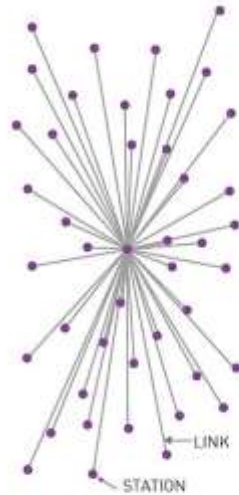
E para quem quiser se arriscar, kaggle.com

15 Active Competitions		
 TWO SIGMA	Two Sigma: Using News to Predict Stock Movements Use news analytics to predict stock price performance <i>Featured</i> - Kernels Competition - a month to go - news agencies, time series, finance , money	\$100,000 2,927 teams
	Jigsaw Unintended Bias in Toxicity Classification Detect toxicity across a diverse range of conversations <i>Featured</i> - Kernels Competition - 10 days to go - biases, nlp, text data	\$65,000 2,924 teams
	Predicting Molecular Properties Can you measure the magnetic interactions between a pair of atoms? <i>Featured</i> - 2 months to go - chemistry, tabular data, regression	\$30,000 934 teams
	Open Images 2019 - Object Detection Detect objects in varied and complex images <i>Research</i> - 4 months to go - image processing, image data	\$25,000 77 teams
	Open Images 2019 - Visual Relationship Detect pairs of objects in particular relationships <i>Research</i> - 4 months to go - image processing, image data	\$25,000 37 teams
	Data Science for Good: City of Los Angeles Help the City of Los Angeles to structure and analyze its job descriptions <i>Analytics</i> - 5 days to go - employment, nlp, image data, text data	\$15,000

Aviso: problemas do kaggle diferem muito da vida real

Redes Complexas

“muito mais que grafos complicados”



This map illustrates the initial air attacks on Iraq and Kuwait during the early morning of January 17, 1991. The map shows the geographical context, including Turkey, Syria, Saudi Arabia, and the Persian Gulf. Key locations marked include Baghdad, Kuwait City, and various military bases and targets. The legend identifies the types of attacks: Strategic Air Campaign Targets (red dots), F-117 Attacks (pink triangles), and Tomahawk Missile Attacks (blue T-shaped symbols). Green arrows indicate the movement of various aircraft, including B-52s, F-15s, F-16s, F-111s, and F-117s, as well as Tomahawk missiles. The map also shows the deployment of AWACS (Air Warning and Control System) aircraft and the use of PAVE LOW and APACHE helicopters. The map is titled "INITIAL AIR ATTACKS Early Morning, January 17" and includes a scale bar from 0 to 200 miles.

January 17, 1991

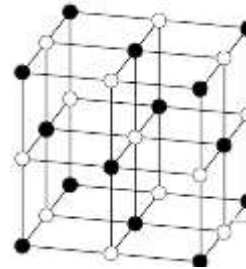
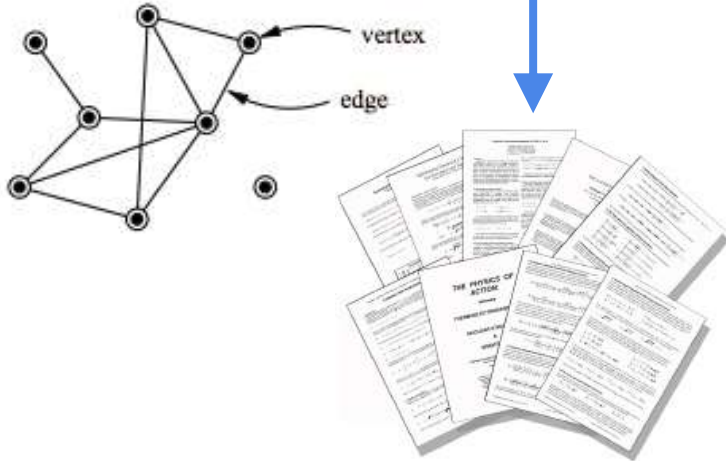
Speaking away from television cameras on the understanding that they not be quoted by name, other American officials acknowledged that the early air strikes had not entirely achieved their mission of severing President Saddam Hussein's command and communications networks. A

[illegible]

It's nice to know that it works, although perhaps this was not the most opportune way to find out."

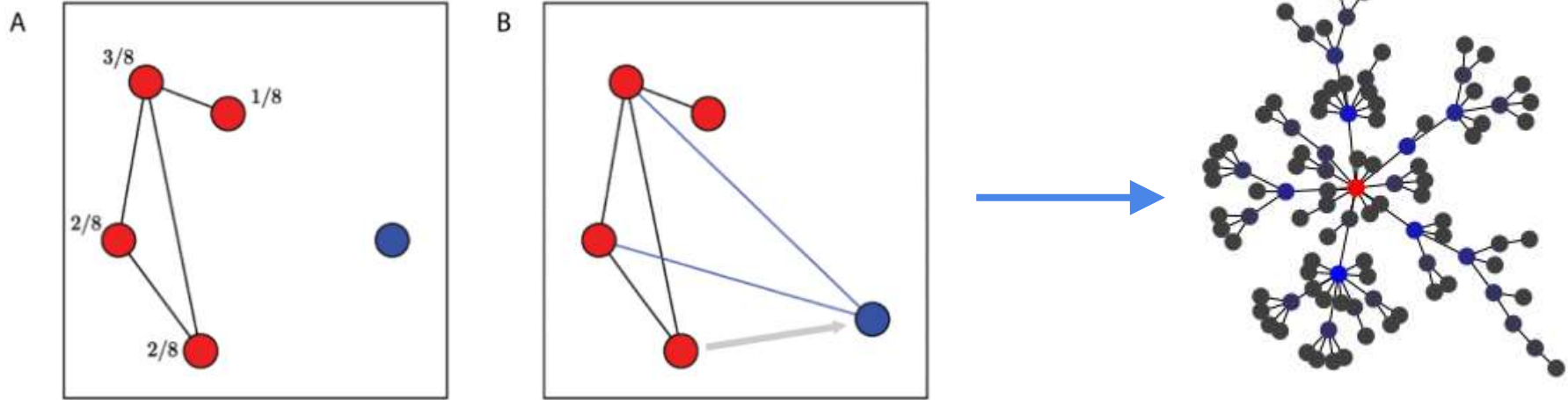
Complex Network

“a **graph** with **non-trivial** features - features that do not occur in simple networks such as **lattices** or random graphs but often occur in graphs modelling **real systems**”

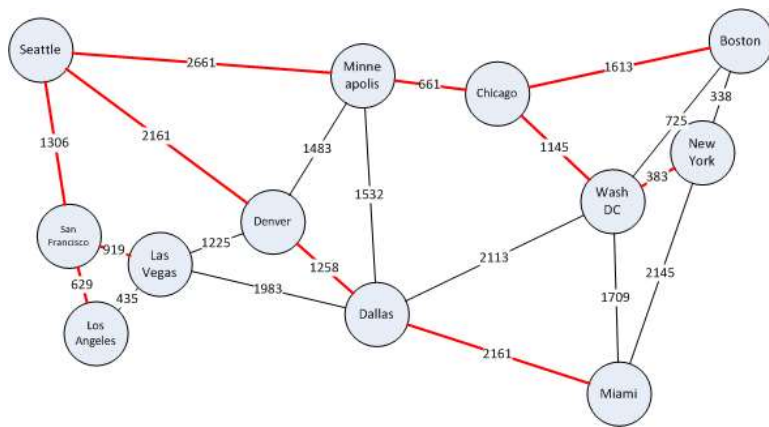


Preferential Attachment

In real networks new nodes tend to link to the more connected nodes.



Typical Graph Problem



Focus on efficiently **computing** something of the graph and efficiently **storing** it;

“How to calculate this optimal path?”

“How can I find the friends of my college friends who are single and know R?”



Typical Complex Network Problem



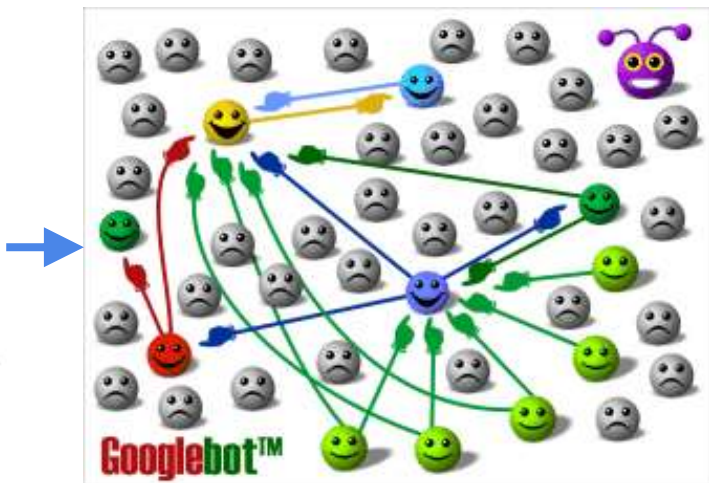
Focus on discovering interesting **features** of the nodes, vertices, or entire graph;

“Is this a resilient military communication network?”

“Who is the most influential person in my company?”



Case I – Google & PageRank



Finding something online was hard

Larry Page and Sergey Brin
publish the PageRank algorithm
(1998)

Profit
(????)

Case I – Google & PageRank

Nowadays, there are billions of websites and links to calculate the PageRank.

How Do They Do it?

I don't know! But,



PageRank in SQL

03 December 2013

One of the most profound ideas in the last 20 years is [PageRank](#), the original algorithm of Google's search engine.

PageRank starts with a simple and clever idea: the importance of a page is determined by how many pages link to it, and how important they are. It's a link analysis algorithm, and it ranks pages by how important they are to the entire collection.

In other words, when a source page (like this blog) links to another, target page (like a Wikipedia article), some of the source's importance should transfer over to the target.

The only things you need are pages and their links. It's a [graph structure](#).



SQL and Graphs

A graph can be stored in SQL using 2 tables, **Nodes** and **Edges**

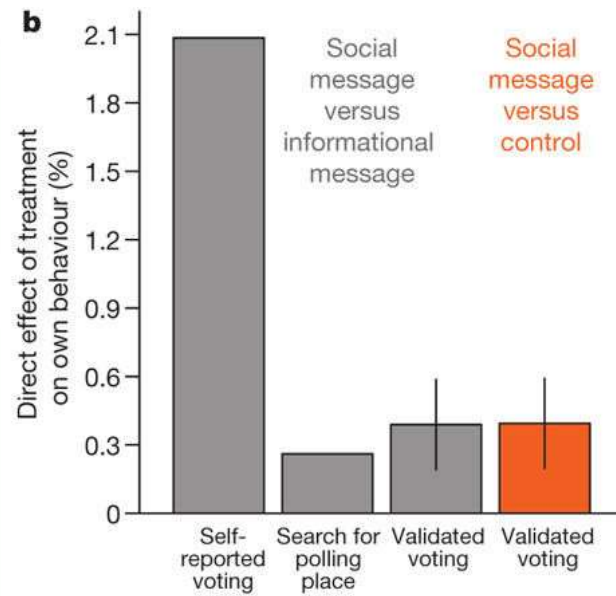
```
CREATE TABLE Nodes
(NodeId int not null
,NodeWeight decimal(10,3) not null
,NodeCount int not null default(0)
,NodeInDegree int not null default(0)
,constraint NodesPK primary key clustered (NodeId))
)

CREATE TABLE Edges
(SourceNodeId int not null
,TargetNodeId int not null
,constraint EdgesPK primary key clustered (SourceNodeId, TargetNodeId)
,constraint EdgesChk check SourceNodeId <> TargetNodeId --ignore self references
)
```

Case II – Facebook & Node Features

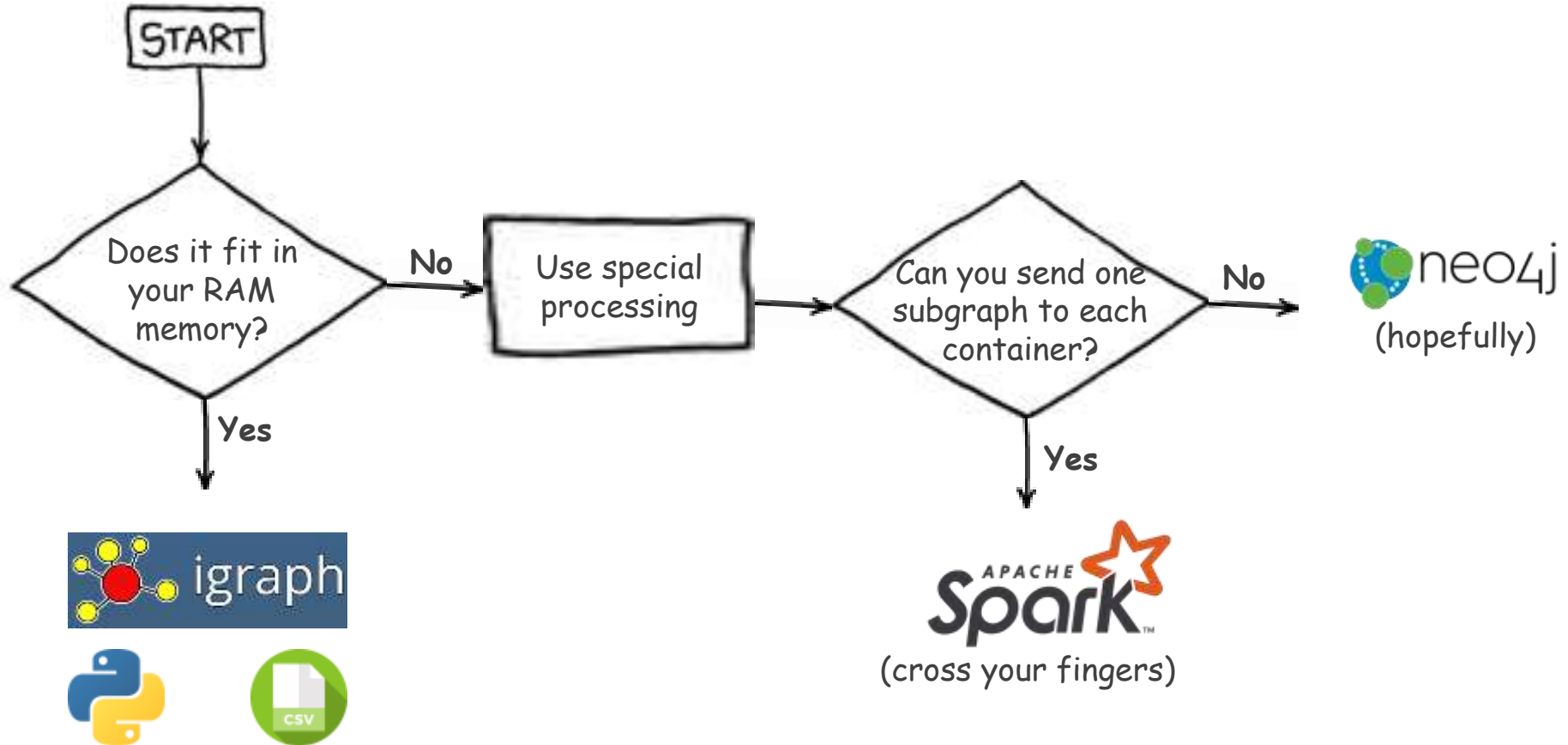


Voting is boring



Showing a picture of your friends saying “I voted” increases turnout: more than **60,000** voted because of the **experiment**

Como Processar Redes Complexas



Redes para se arriscar

<https://snap.stanford.edu/data/>

By Jure Leskovec

STANFORD
UNIVERSITY



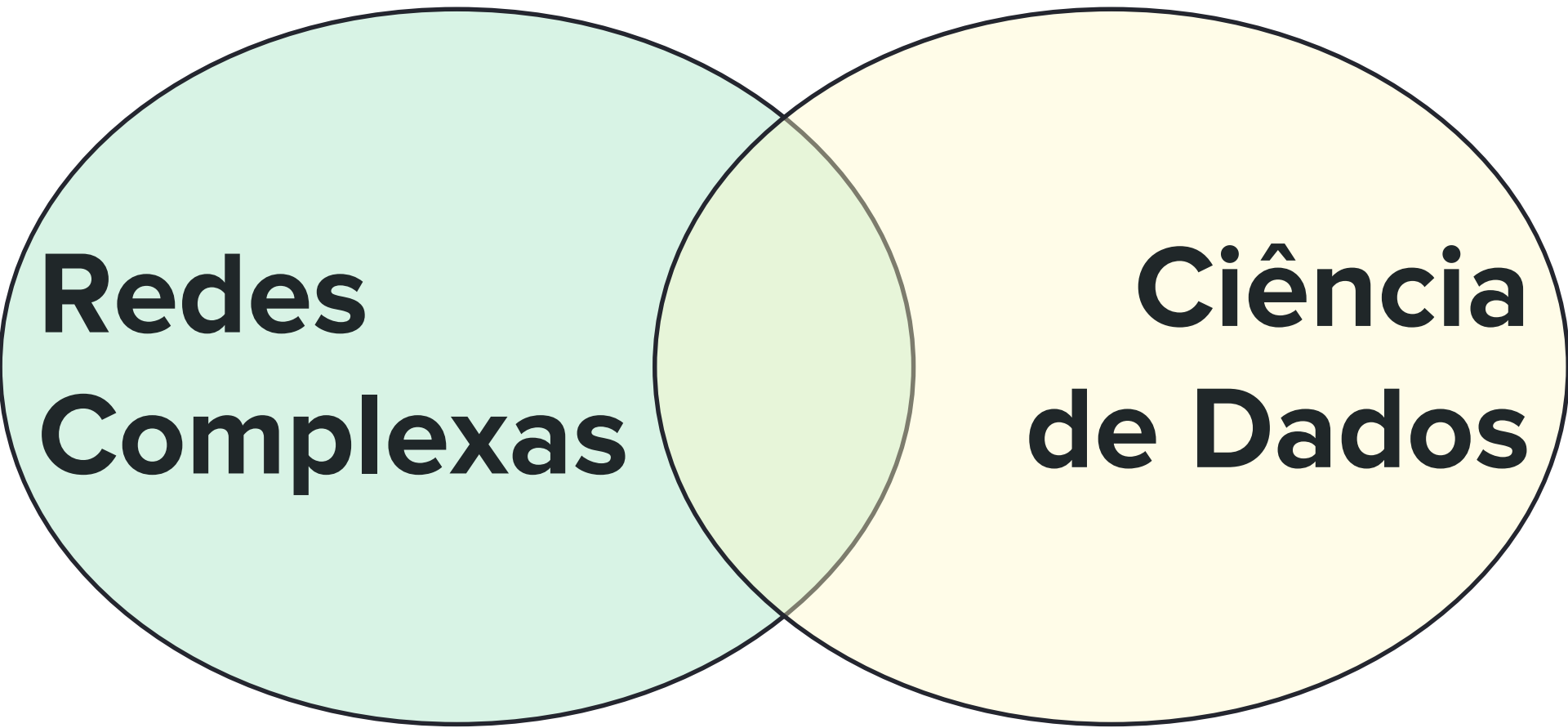
- SNAP for C++ ▶
- SNAP for Python ▶
- SNAP Datasets ▶
- BIOSNAP Datasets
- What's new
- People
- Papers
- Projects ▶
- Citing SNAP
- Links
- About
- Contact us



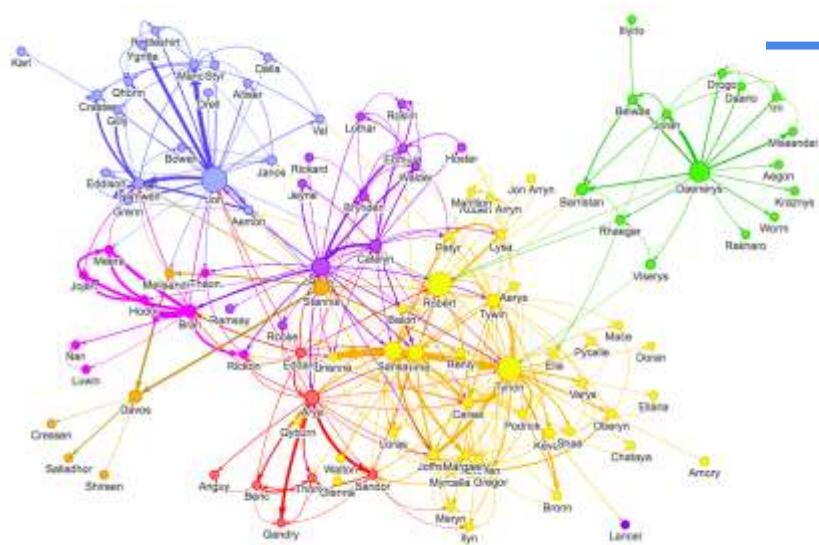
Stanford Large Network Dataset Collection

- **Social networks** : online social networks, edges represent interactions between people
- **Networks with ground-truth communities** : ground-truth network communities in social and information networks
- **Communication networks** : email communication networks with edges representing communication
- **Citation networks** : nodes represent papers, edges represent citations
- **Collaboration networks** : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- **Web graphs** : nodes represent webpages and edges are hyperlinks
- **Amazon networks** : nodes represent products and edges link commonly co-purchased products
- **Internet networks** : nodes represent computers and edges communication
- **Road networks** : nodes represent intersections and edges roads connecting the intersections
- **Autonomous systems** : graphs of the internet
- **Signed networks** : networks with positive and negative edges (friend/foe, trust/distrust)
- **Location-based online social networks** : Social networks with geographic check-ins
- **Wikipedia networks, articles, and metadata** : Talk, editing, voting, and article data from Wikipedia
- **Temporal networks** : networks where edges have timestamps
- **Twitter and Memetracker** : Memetracker phrases, links and 467 million Tweets
- **Online communities** : Data from online communities such as Reddit and Flickr
- **Online reviews** : Data from online review systems such as BeerAdvocate and Amazon

SNAP networks are also available from [SuiteSparse Matrix Collection](#) by [Tim Davis](#).



1 – Features dos nós



Name	Indegree	Outdegree	...	Pagerank
Tyrion	4	20		0.042884981
Jon	3	17		0.035828696
Robb	2	12		0.030171146
Sansa	2	10		0.030009716
...				

Se seus dados possuem algum tipo de relacionamendo (personagens, cenas em comum) **descreva** eles como um grafo

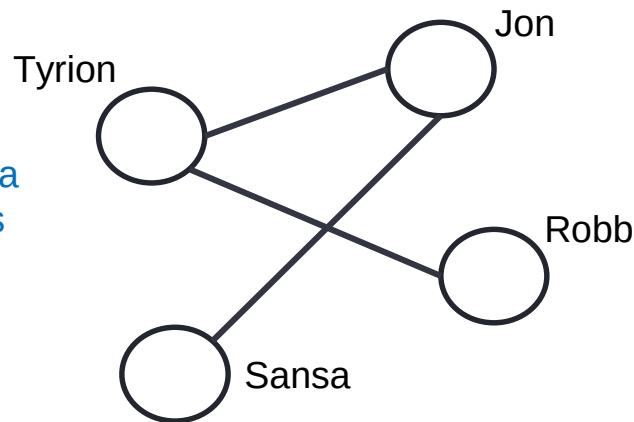


2 – Linhas como nós

Name	Gender	Age	...	Kills
Tyrion	M	30		2
Jon	M	18		7
Robb	M	17		300
Sansa	F	16		5
...				

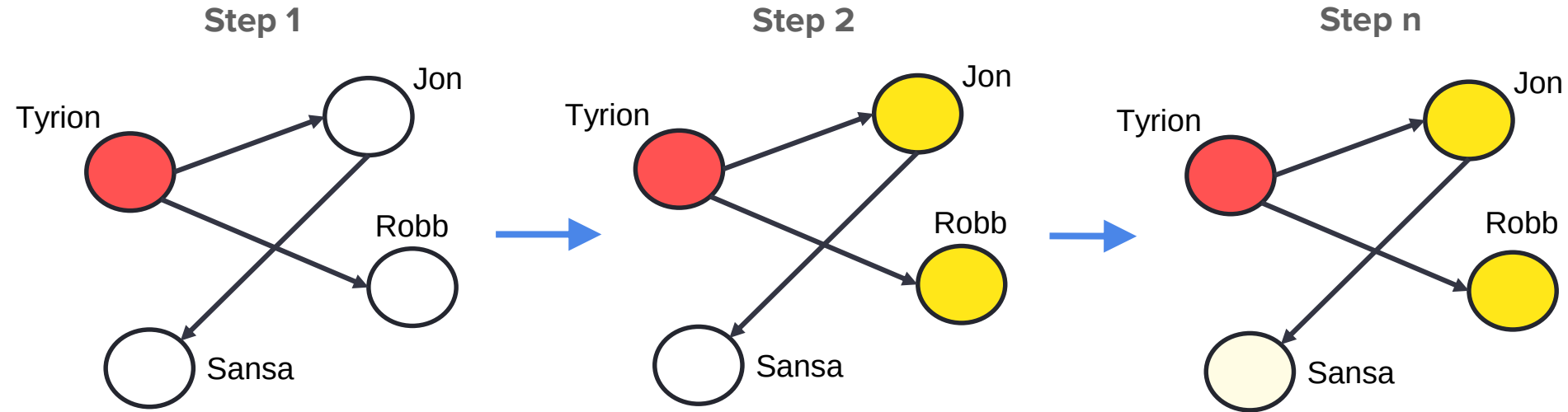
Crie um grafo em que as linhas são os nós e os links são uma medida de **similaridade**

Compartilha
+2 features



Knn ou outro
algoritmo baseado
em distância

3 – Simulação



↓
Predição

Simule o efeito de uma **regra de propagação** de um evento pelo grafo. Use o resultado final como predição.

Contact me!

- [hugolouzada --em-- gmail.com](mailto:hugolouzada@gmail.com)
 - <https://www.linkedin.com/in/vhlouzada>
-

Extra stuff!

- node2vec: Scalable Feature Learning for Networks, <http://snap.stanford.edu/node2vec/>
- Fraud Prevention with Spark and Graph Analysis, <https://youtu.be/0VO-ts0dsbl>
- GraphFrames: Spark package for graphs, <https://graphframes.github.io>
- Robustness of airport networks, <https://doi.org/10.1371/journal.pone.0118635>

Obrigado!

We are hiring!

Mande seu CV para atratividade@itau-unibanco.com.br