

Machine Learning and the need for reproducibility

Prof. Luiz Eduardo S. Oliveira, Ph.D.
Department of Informatics, UFPR

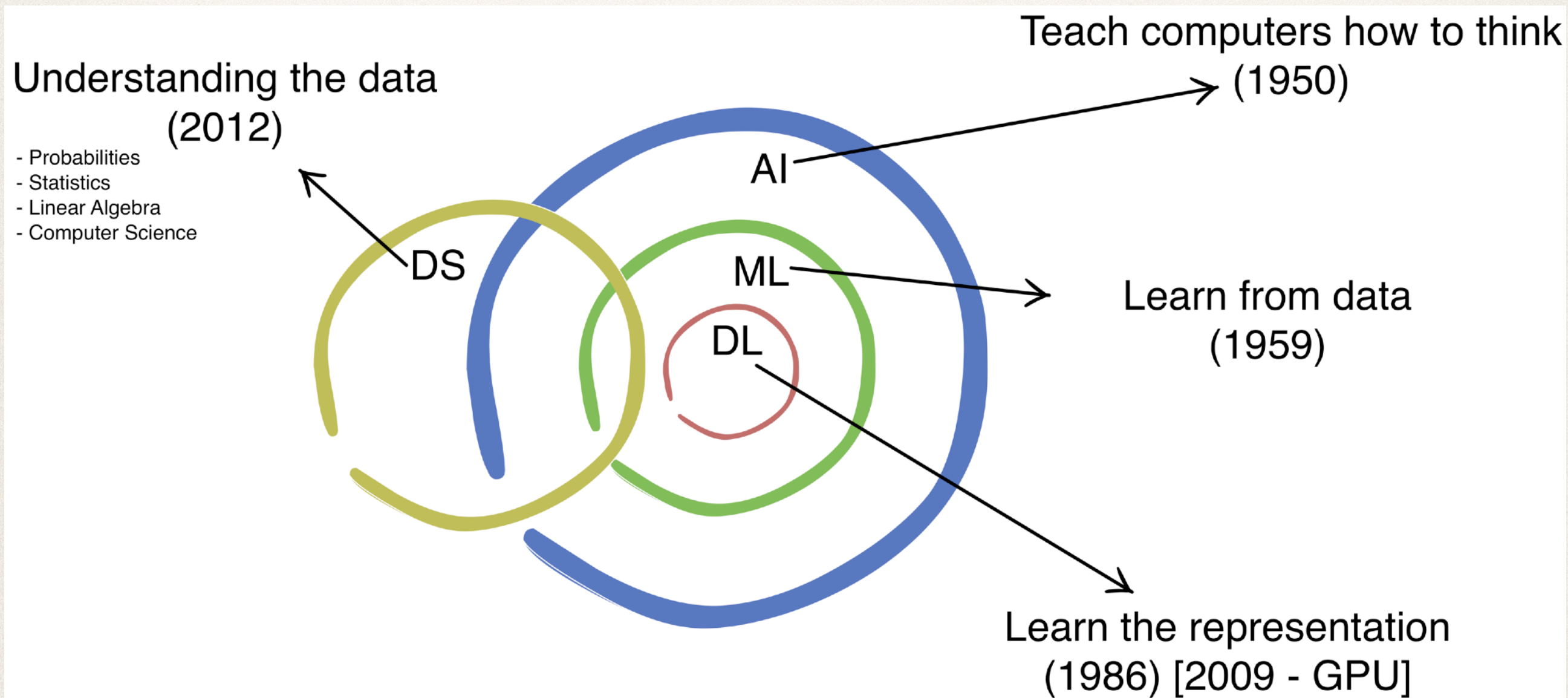
About me

- ❖ BSc, MSc, and PhD in Computer Science (95,98,03)
- ❖ Professor at PUCPR [2004-2009]
- ❖ Head of Research and Development at Invisys Computer Vision Systems [2004-2009]
- ❖ Professor at DInf, UFPR [2009-]
 - ❖ Head of the Graduate Program in Informatics [2012-2014, 2016-2018]
 - ❖ Member of the VRI and DSBD research groups

Agenda

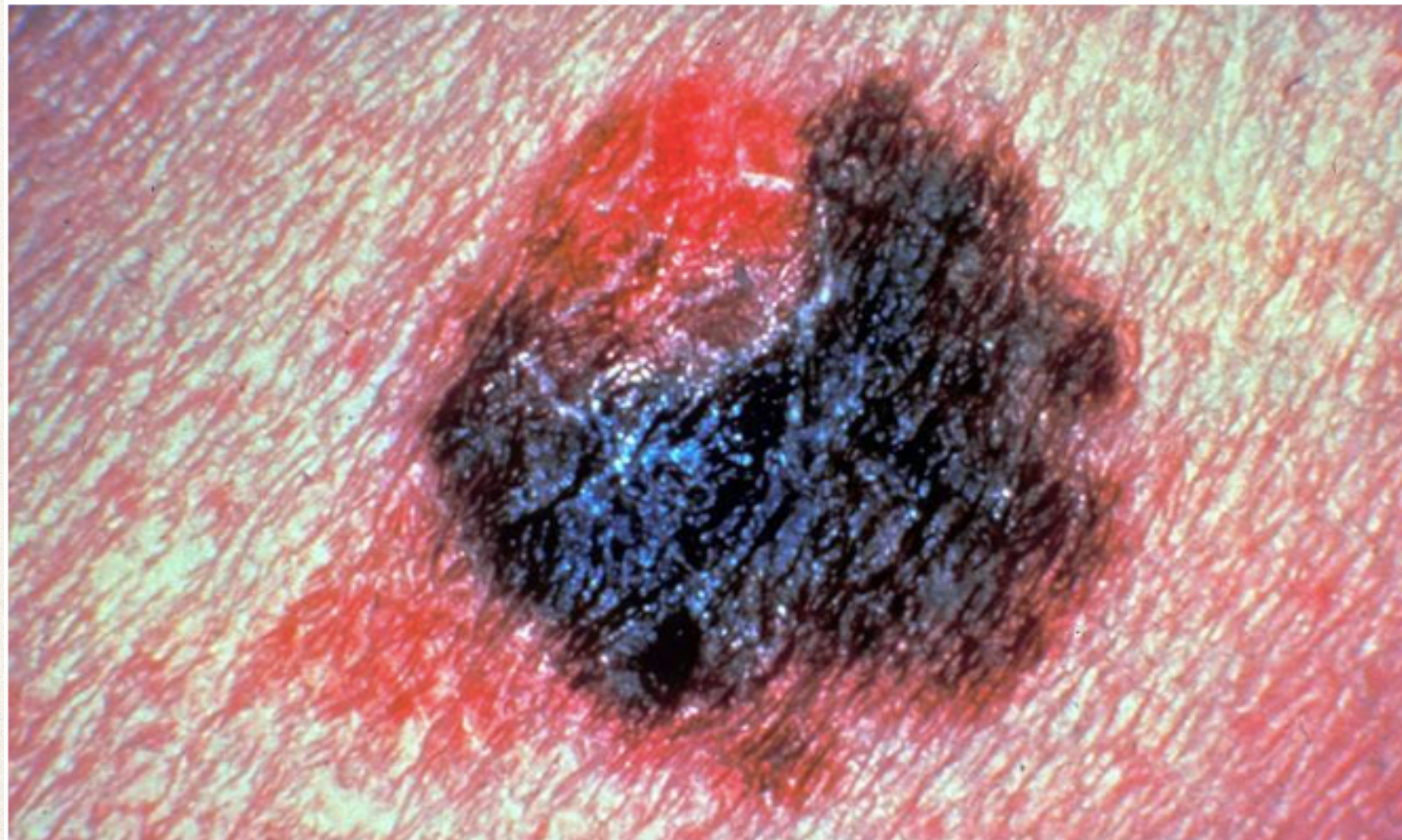
- ❖ Machine Learning
 - ❖ What is it?
 - ❖ Why now?
- ❖ Reproducibility Crisis
- ❖ What can we do?

What is Machine Learning



This AI can spot skin cancer as accurately as a doctor

The artificial intelligence was trained on an image database of 129,000 images and performed as well as trained medical professionals

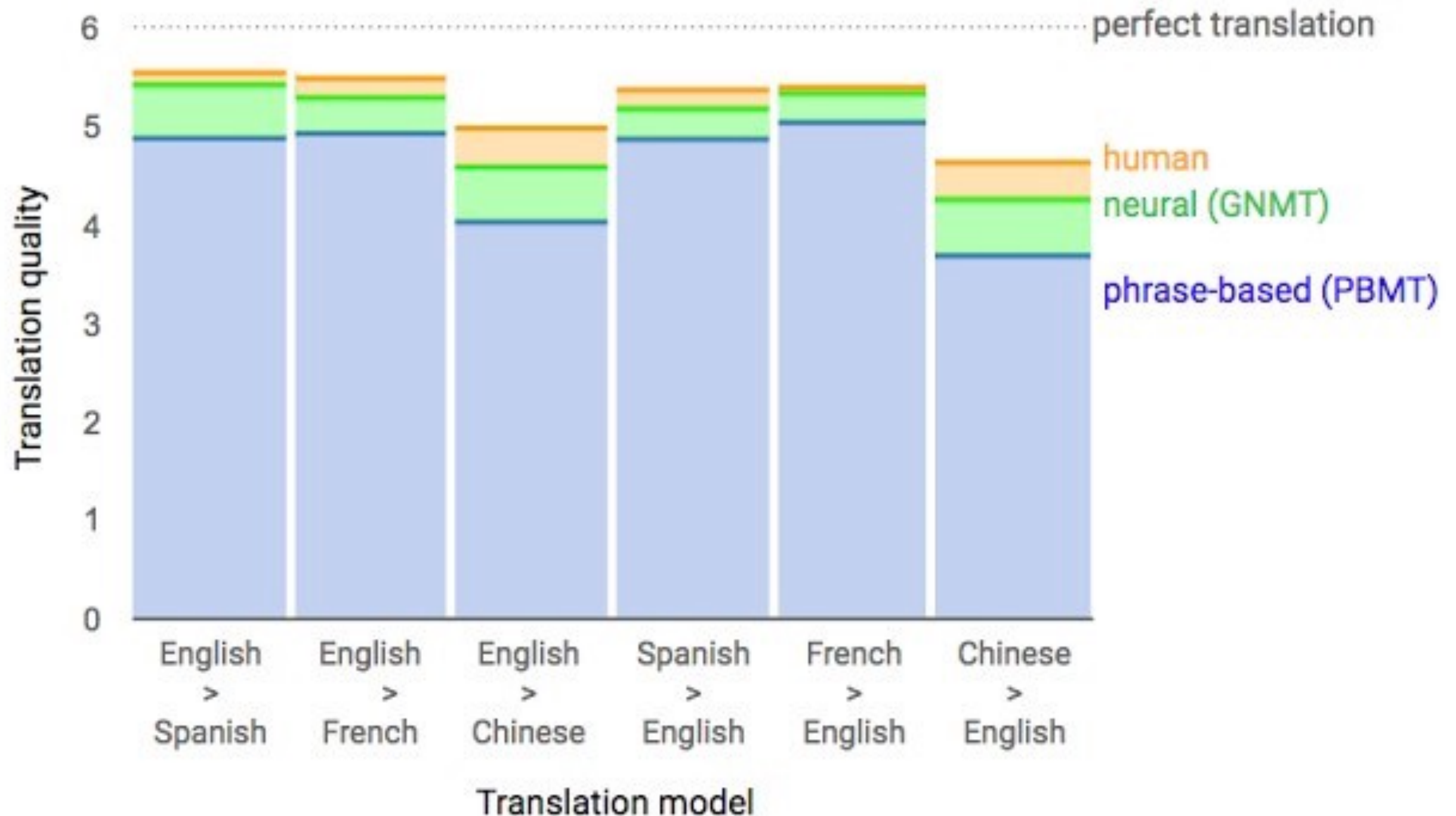


Google's Go-playing AI still undefeated with victory over world number one

AlphaGo has won its second game against China's Ke Jie, sealing the three-game match in its favour



Google Neural Machine Translation



A black Amazon Echo smart speaker is centered on a white shelf. The speaker has a glowing blue light ring at the top. To the left of the speaker, there are some blue books and a silver paperclip. To the right, there is a small gold-colored decorative object. In the background, there is a light gray wall and some green leaves of a plant hanging from the top right.

INDY/TECH

AMAZON ECHO: HOW IT WILL BRING ARTIFICIAL INTELLIGENCE INTO OUR HOMES MUCH SOONER THAN EXPECTED

Lip Reading



Chung et al, Lip Reading Sentences in the Wild, <https://arxiv.org/pdf/1611.05358v1.pdf>

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules





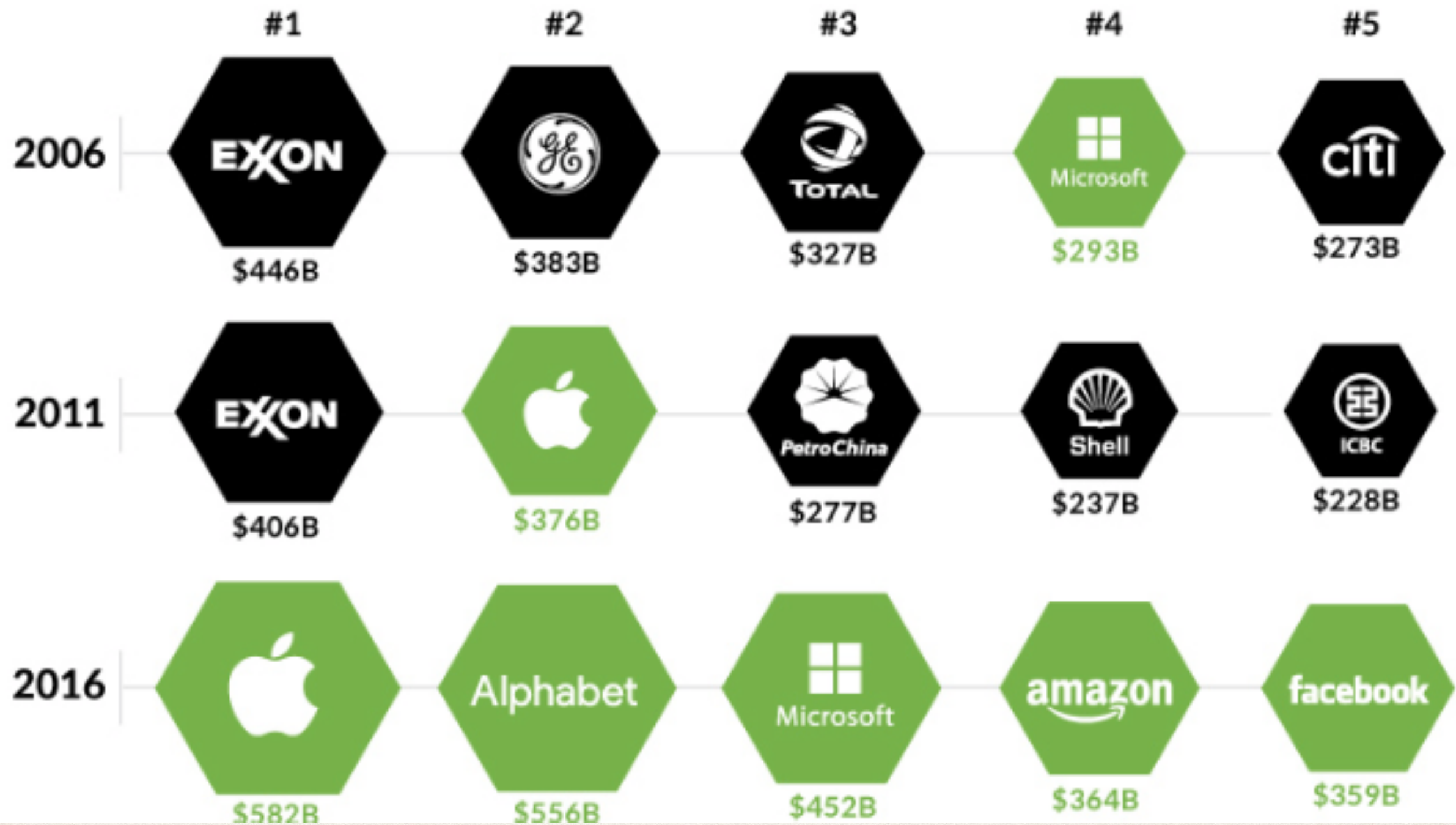
Top 5 Publicly Traded Companies (by Market Cap)



Tech



Other



What is Machine Learning

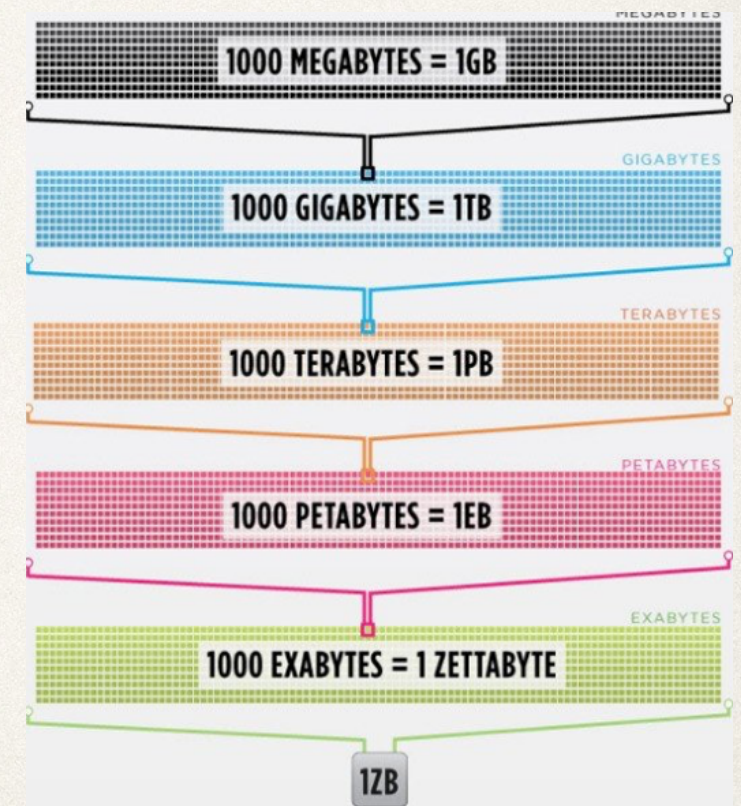
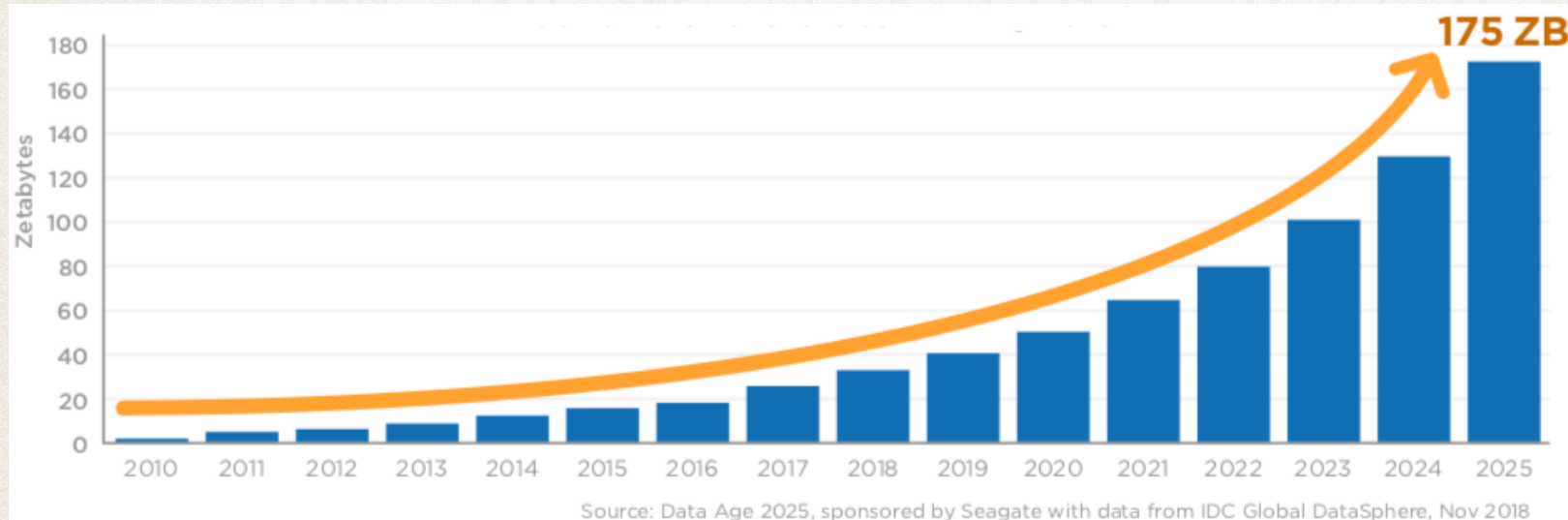
- ❖ Machine Learning at its most basic is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world.” – *Nvidia*
- ❖ Machine learning is the science of getting computers to act without being explicitly programmed.” – *Stanford*
- ❖ Machine learning is based on algorithms that can learn from data without relying on rules-based programming.” - *McKinsey & Co*
- ❖ Machine learning algorithms can figure out how to perform important tasks by generalizing from examples.” – *University of Washington*
- ❖ Machine learning research is part of research on [artificial intelligence](#), seeking to provide knowledge to computers through data, observations and interacting with the world. That acquired knowledge allows computers to correctly generalize to new settings - *Université de Montreal*

Why now

- ❖ Data available
- ❖ Computational power
- ❖ Frameworks

Data

- ❖ IDC predicts that the data we create, capture, or replicate (DataSphere) will grow significantly over the next decade

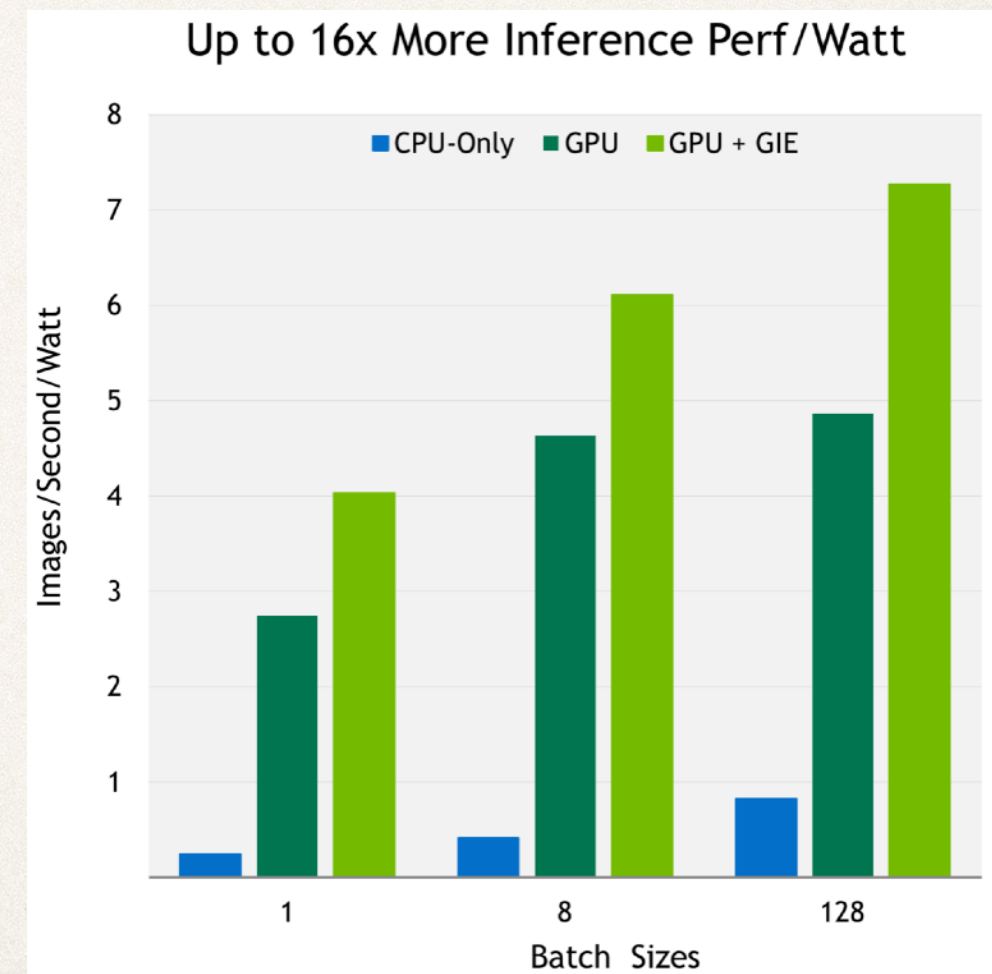
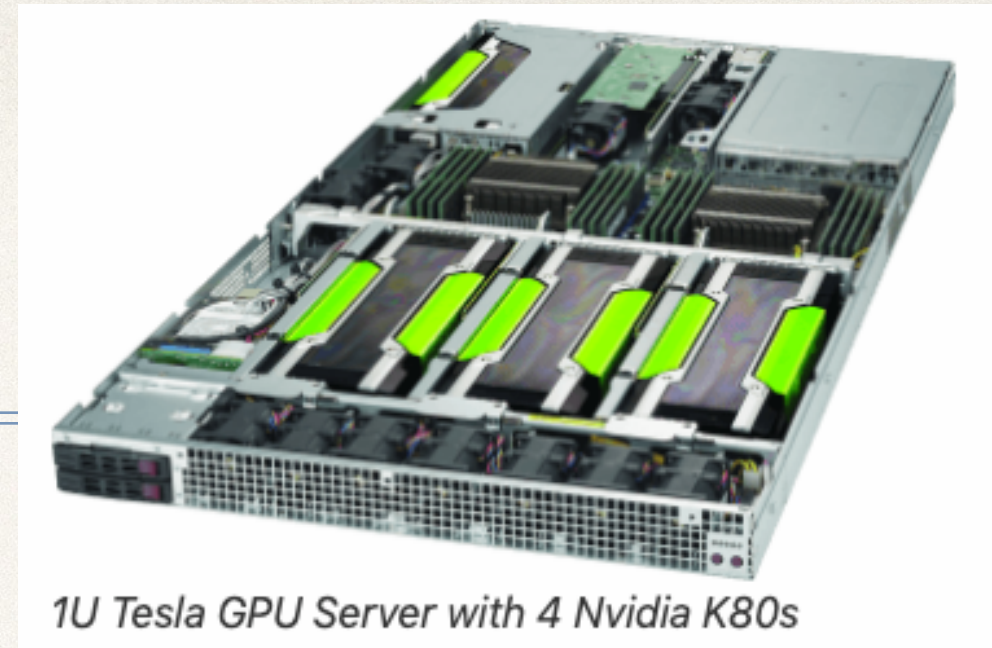


Data Sources



Hardware

- ❖ GPU popularised by NVIDIA
- ❖ Hardware designed for games
- ❖ GIE (GPU Inference Engine): Inference optimizer and runtime that delivers low latency and high-throughput for deep learning inference applications.



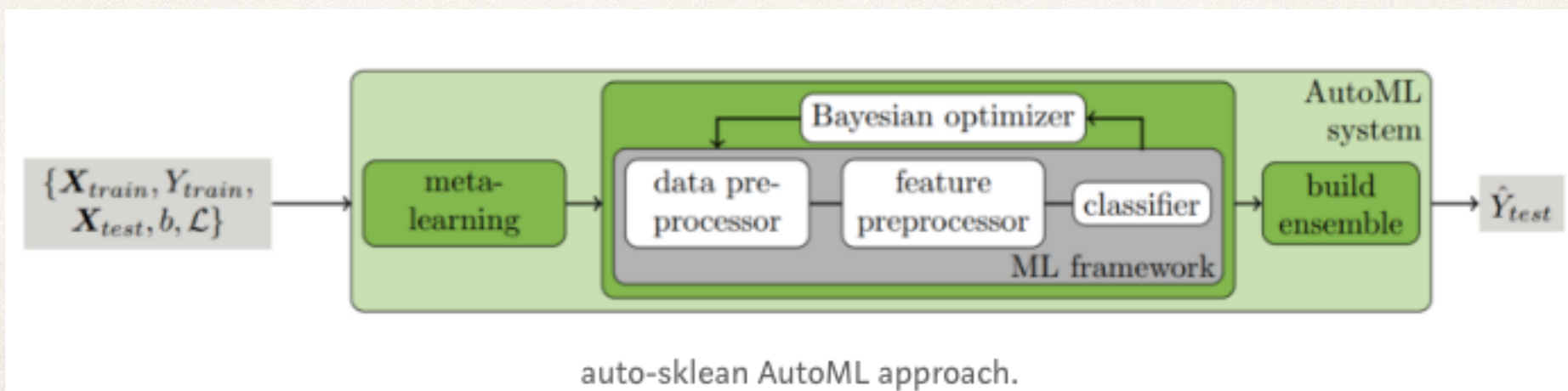
Frameworks

Machine Learning Frameworks

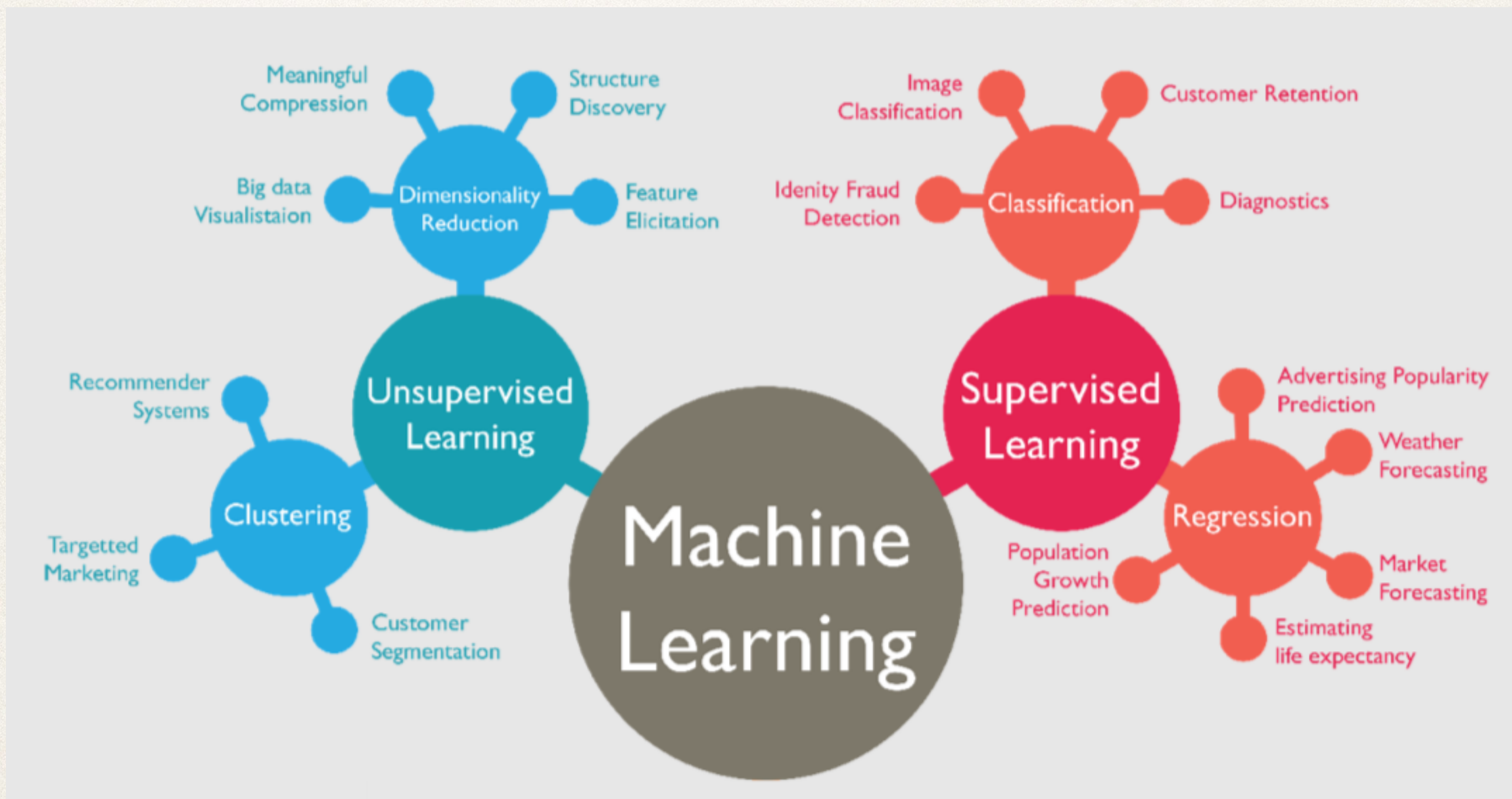


Frameworks

- ❖ AutoML
- ❖ Google, Microsoft, SKLearn, Keras
- ❖ ML algorithm to decide which algorithm you should use for a given dataset



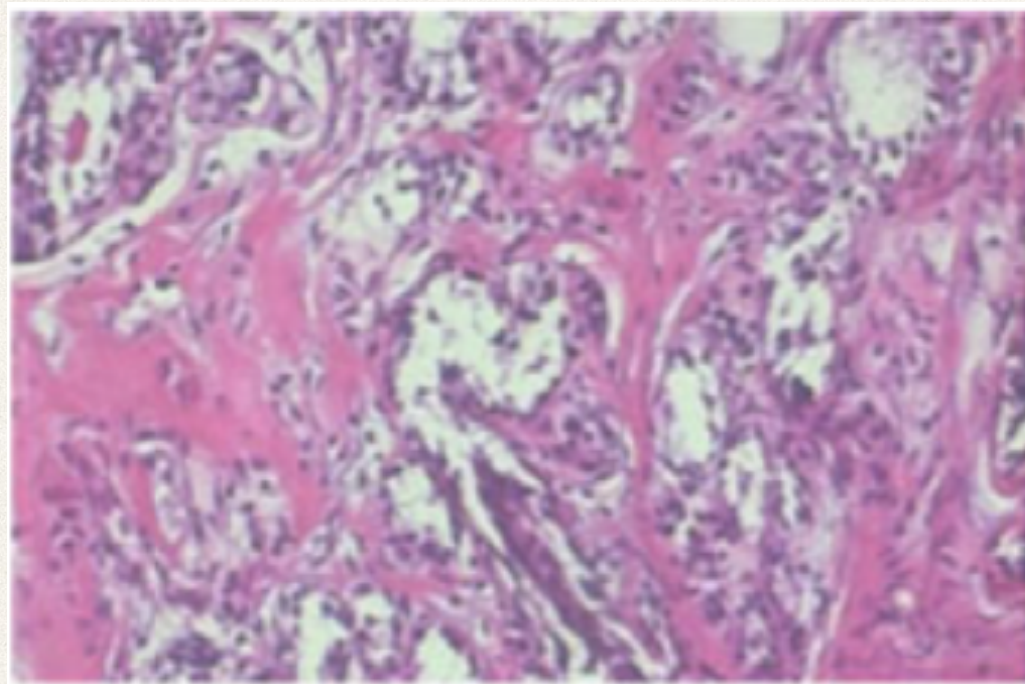
Where?



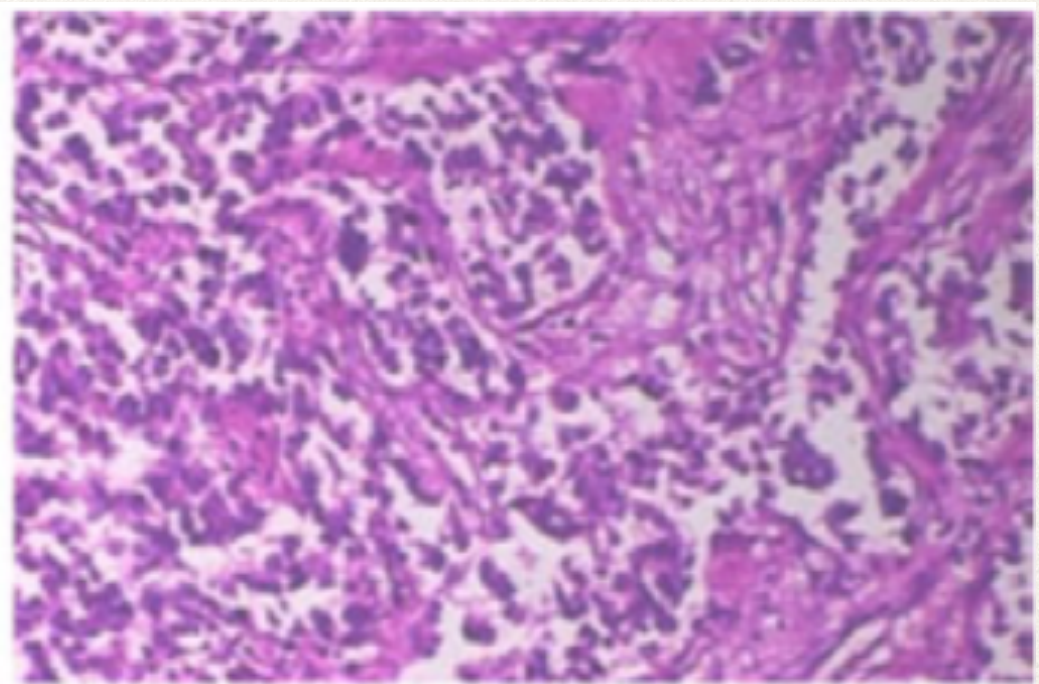
5512

Supervised Learning - Classification

- ❖ To build a classification system to discriminate between benign and malignant tumor on histopathological images

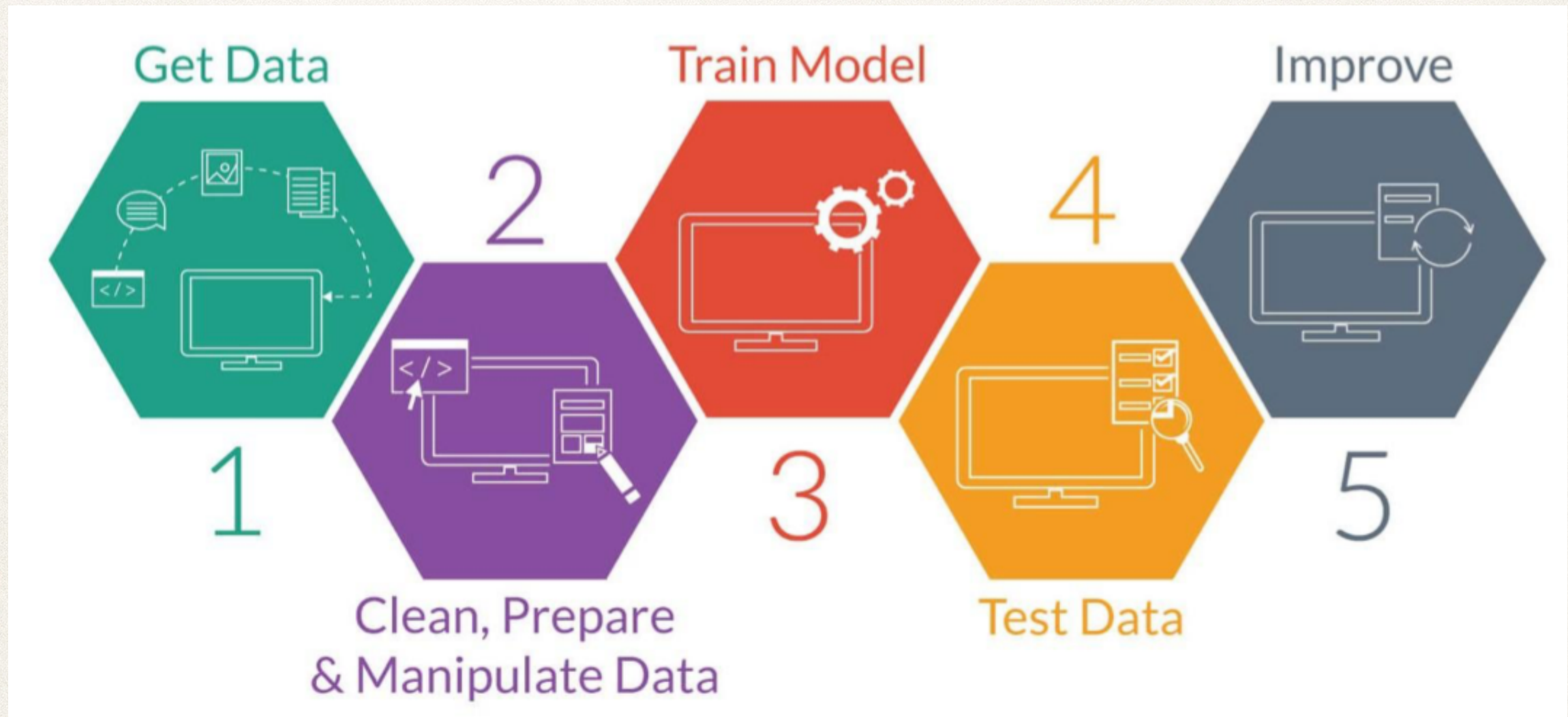


Benign



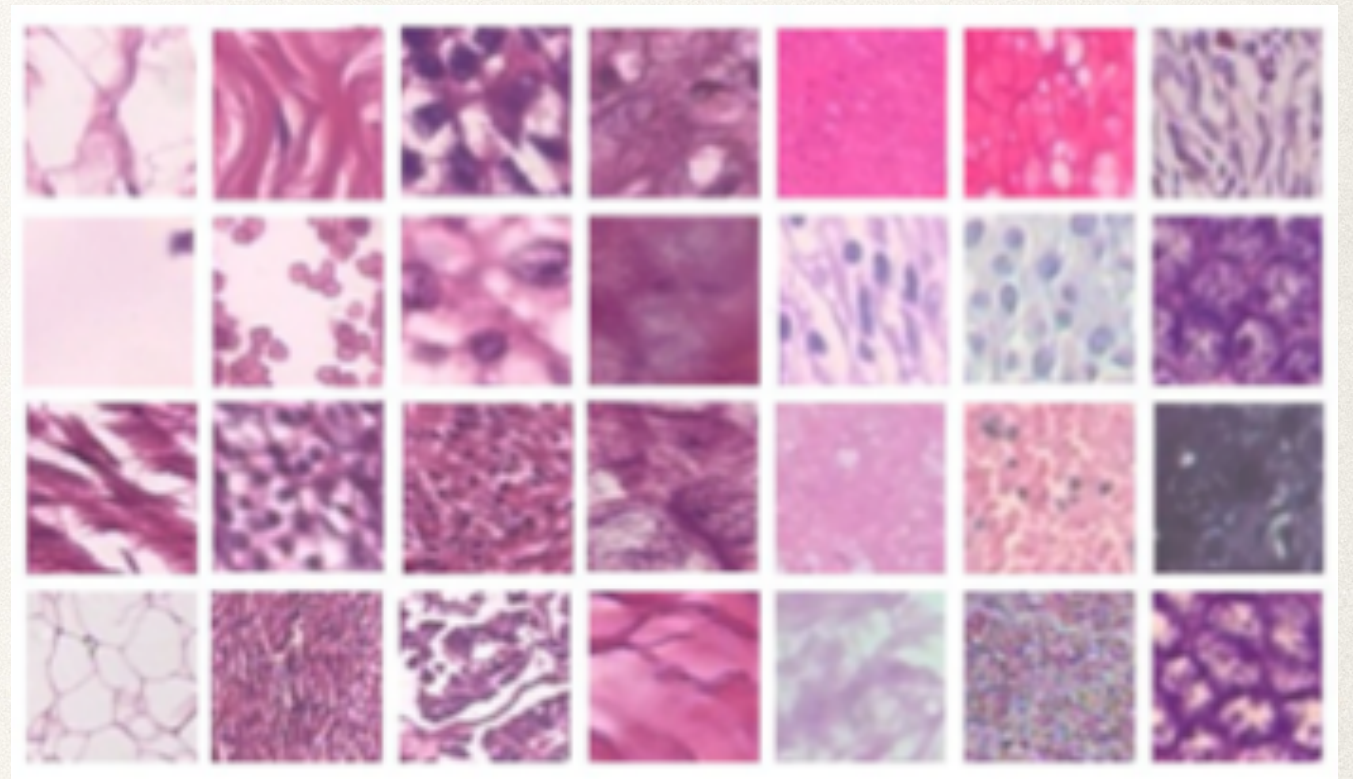
Malignant

The ML Pipeline



Get Data

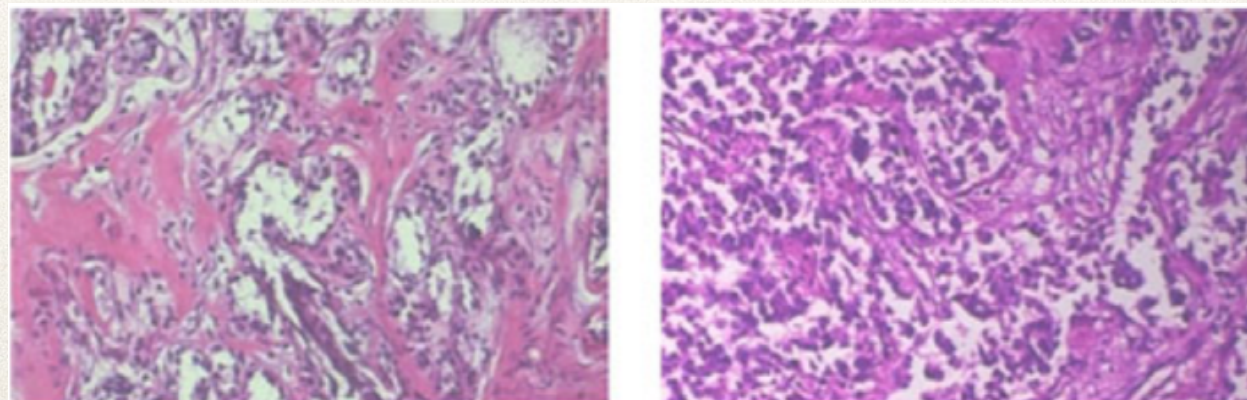
- ❖ Dataset of labeled data (classification is a supervised task)
- ❖ The dataset is usually divided into, training, validation, and testing.



Data variability

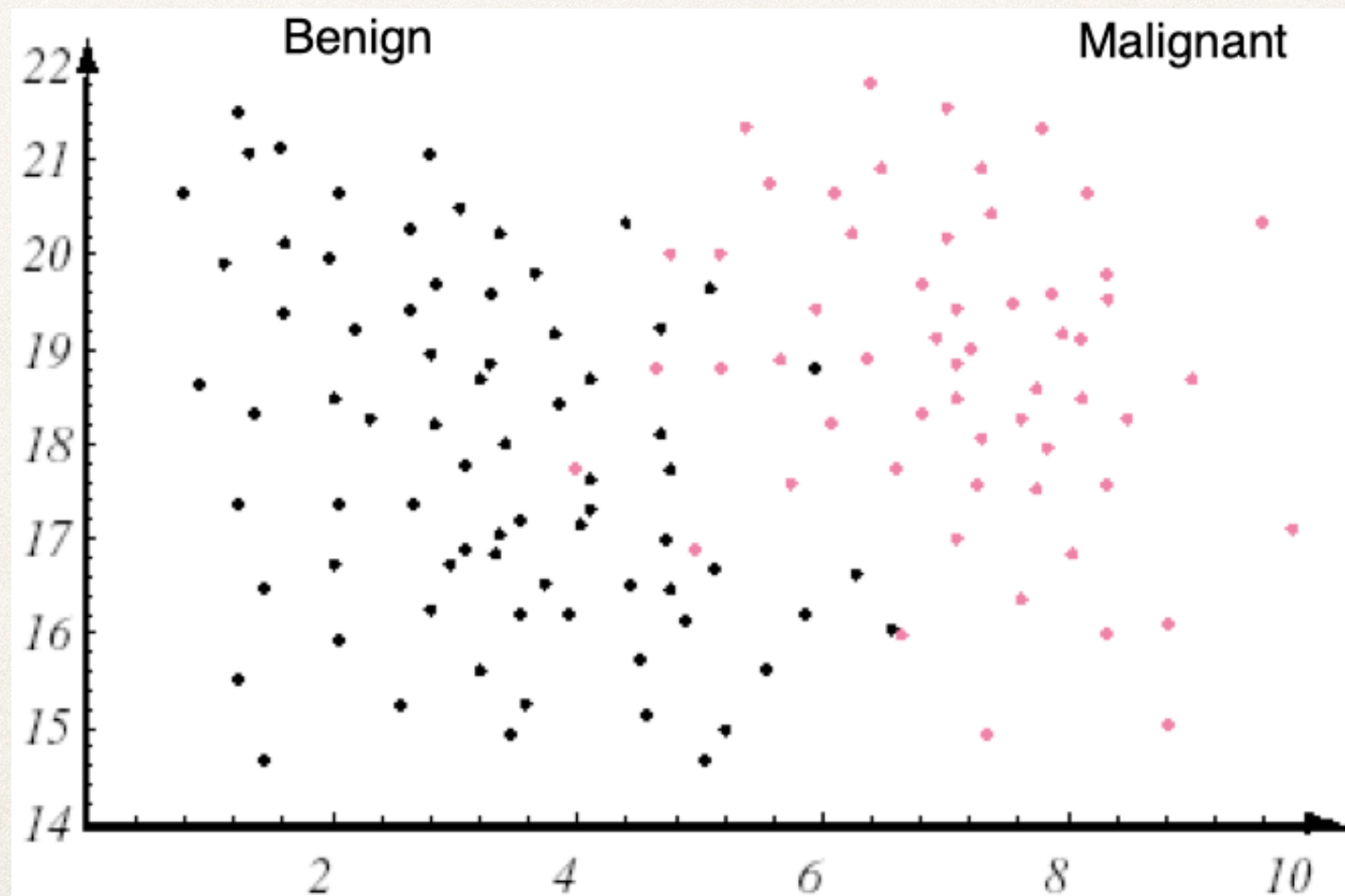
Prepare the Data

- ❖ Clean the data (noise reduction)
 - ❖ In some problems, deal with missing data
- ❖ Find a **representation**, i.e, what is the main difference between these two classes of images?
 - ❖ This process is also known as feature engineering.
 - ❖ To convert the image into a numerical feature vector.



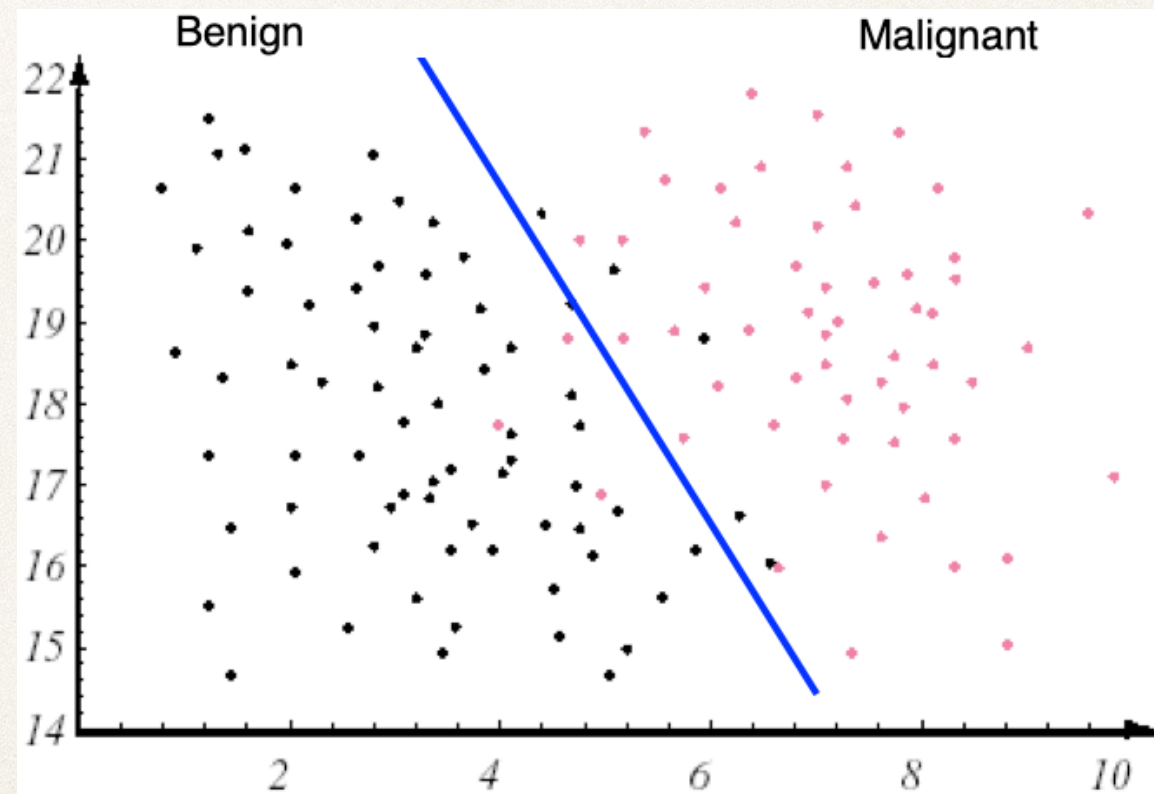
Representation

- ❖ Two hypothetical features that discriminate the two classes.



Supervised Learning: Train the model

- ❖ Com base nos dados de treinamento, treinar um modelo que separe as duas classes



Separação Linear: $y = ax + b$

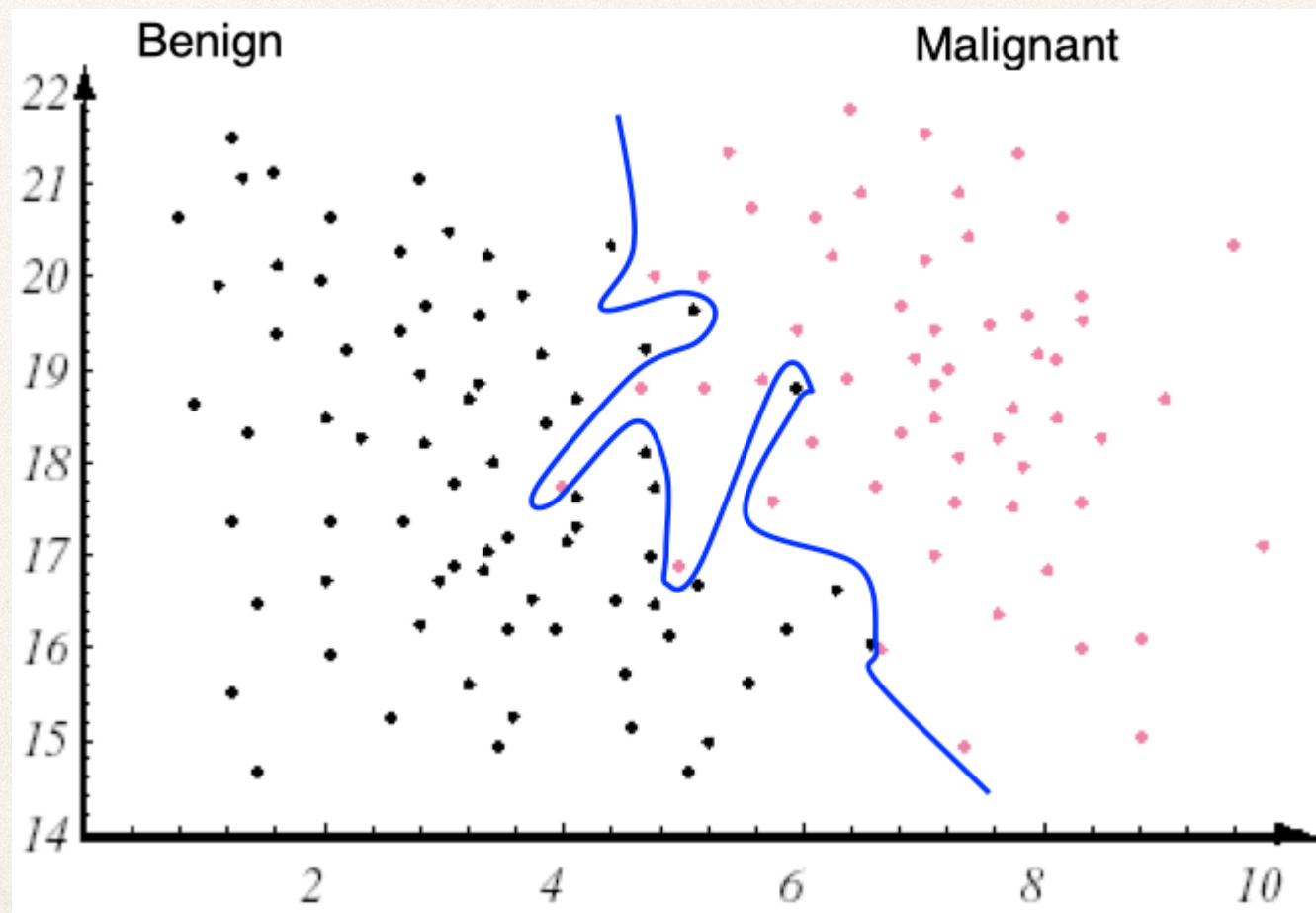
Funções Discriminantes Lineares (Perceptron, SVM)

Test and Improve the Model

- ❖ A linear model can be improved
 - ❖ Other models
- ❖ More features
 - ❖ Usually means more data
 - ❖ Curse of dimensionality

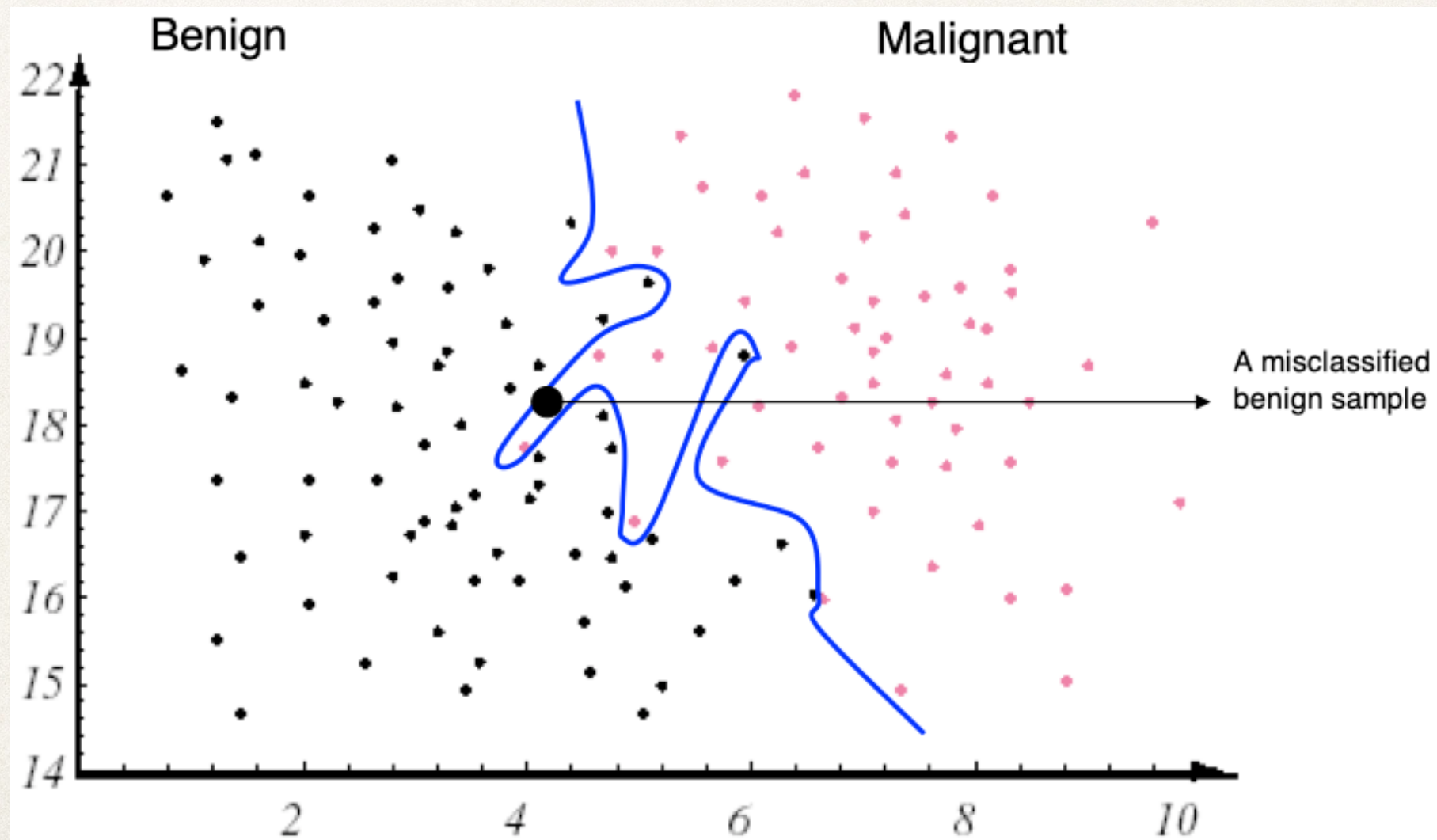
Improve the Model

- ✿ A more complex model. Is it better?



Improve the Model

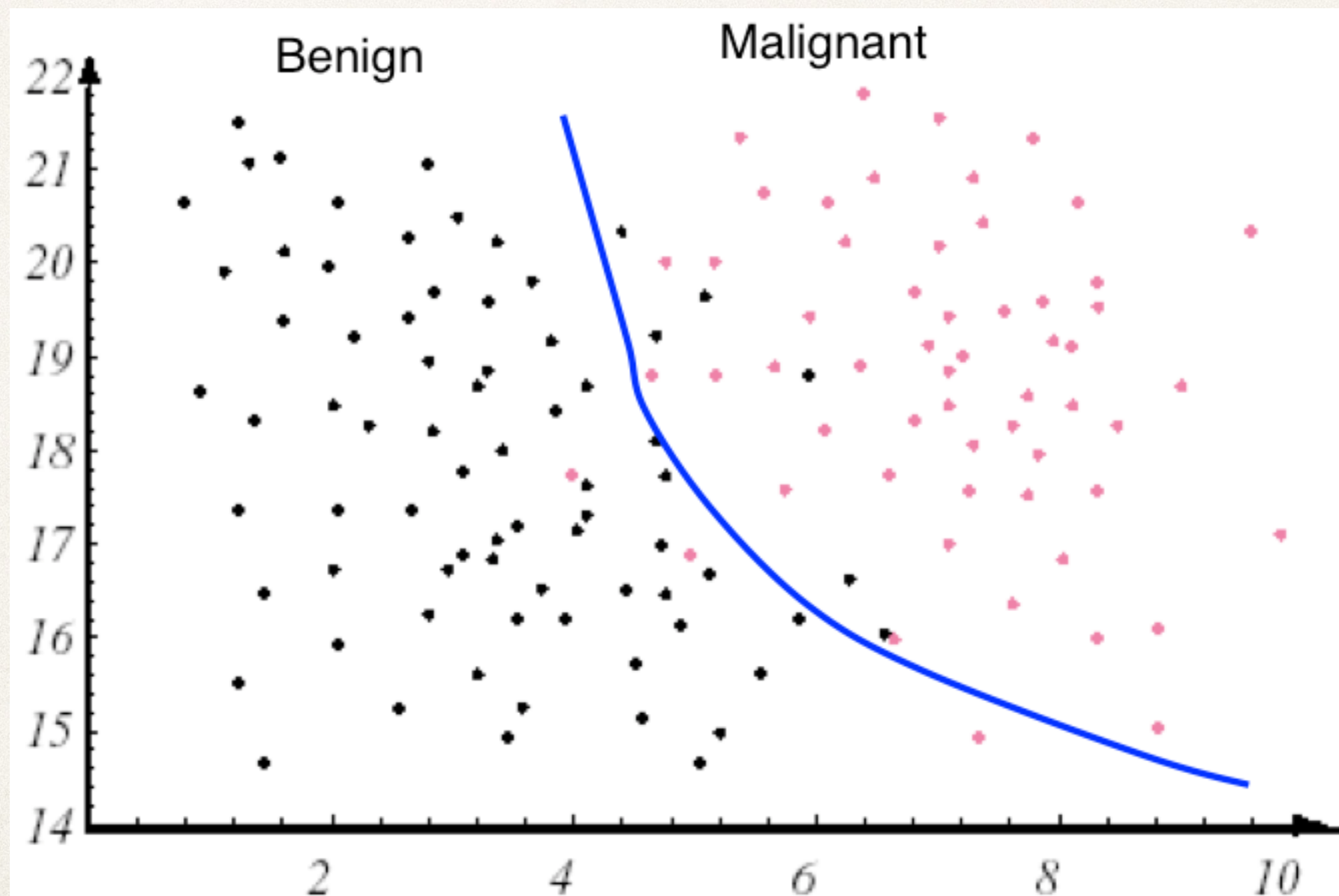
- ❖ Over-fitting, i.e., lack of generalization on unseen data.



Improve the Model

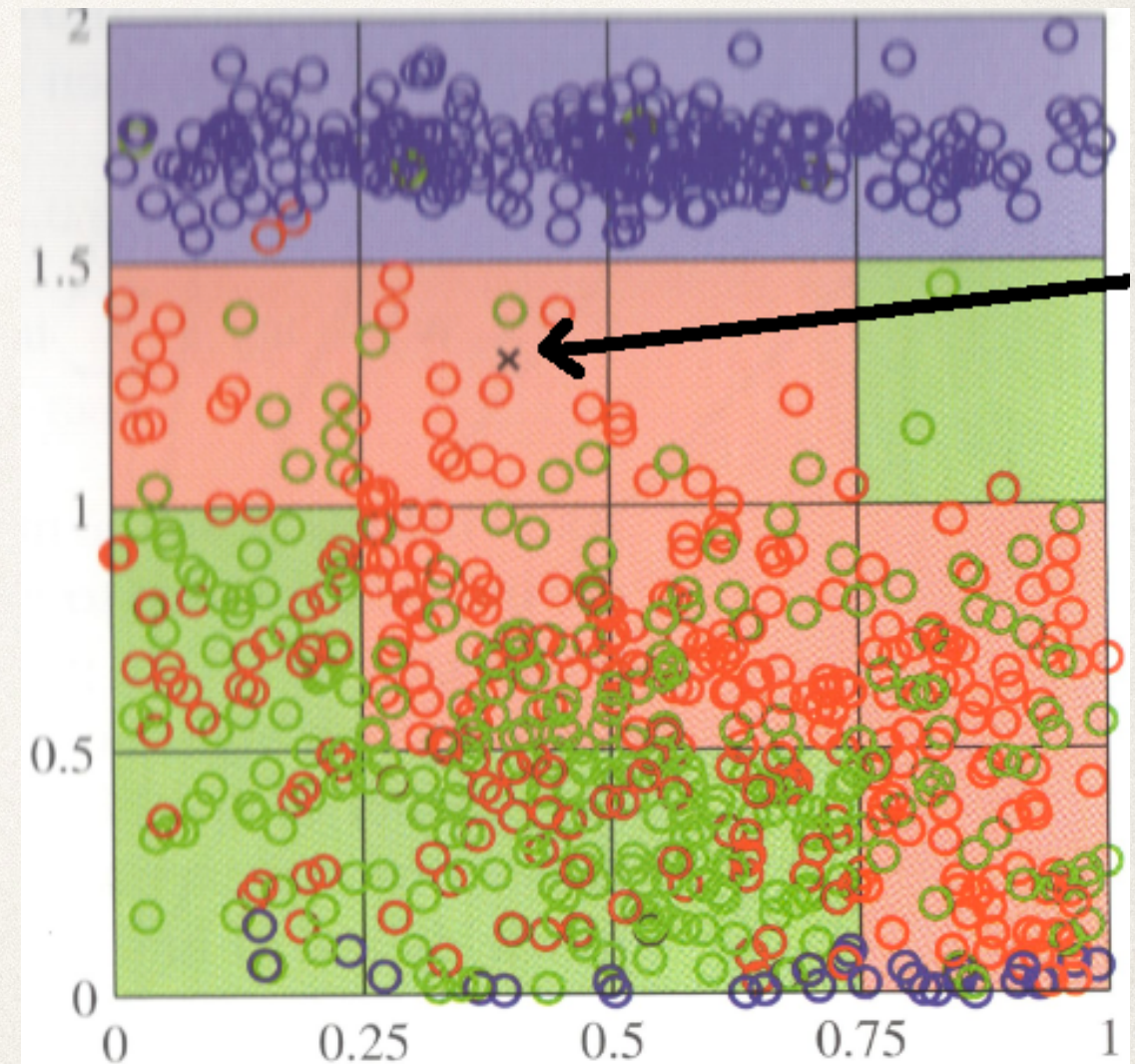
- ❖ Under-fitting: refers to a model that can neither model the training data nor generalize to new data.
- ❖ Over-fitting: refers to a model that models the training data too well.

Improve the Model



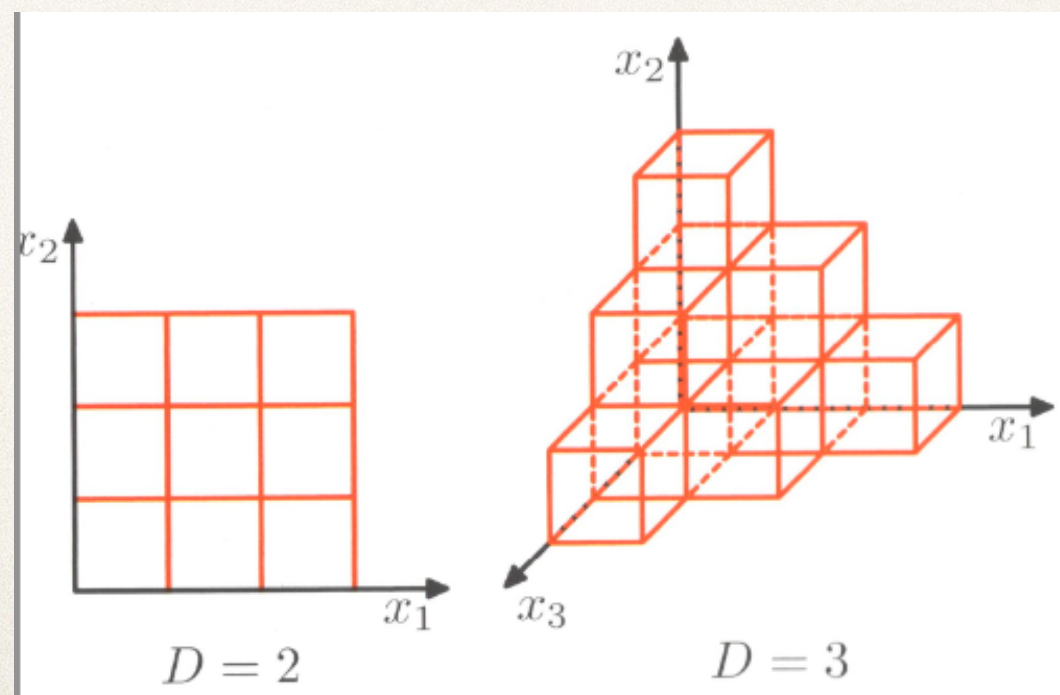
Curse of Dimensionality

- ❖ A common practice is to add more features when looking for better machine learning models
- ❖ Consider the following problem in the 2D space
- ❖ To classify the input pattern x , we divide the feature space and assign x to the of its closest neighbor



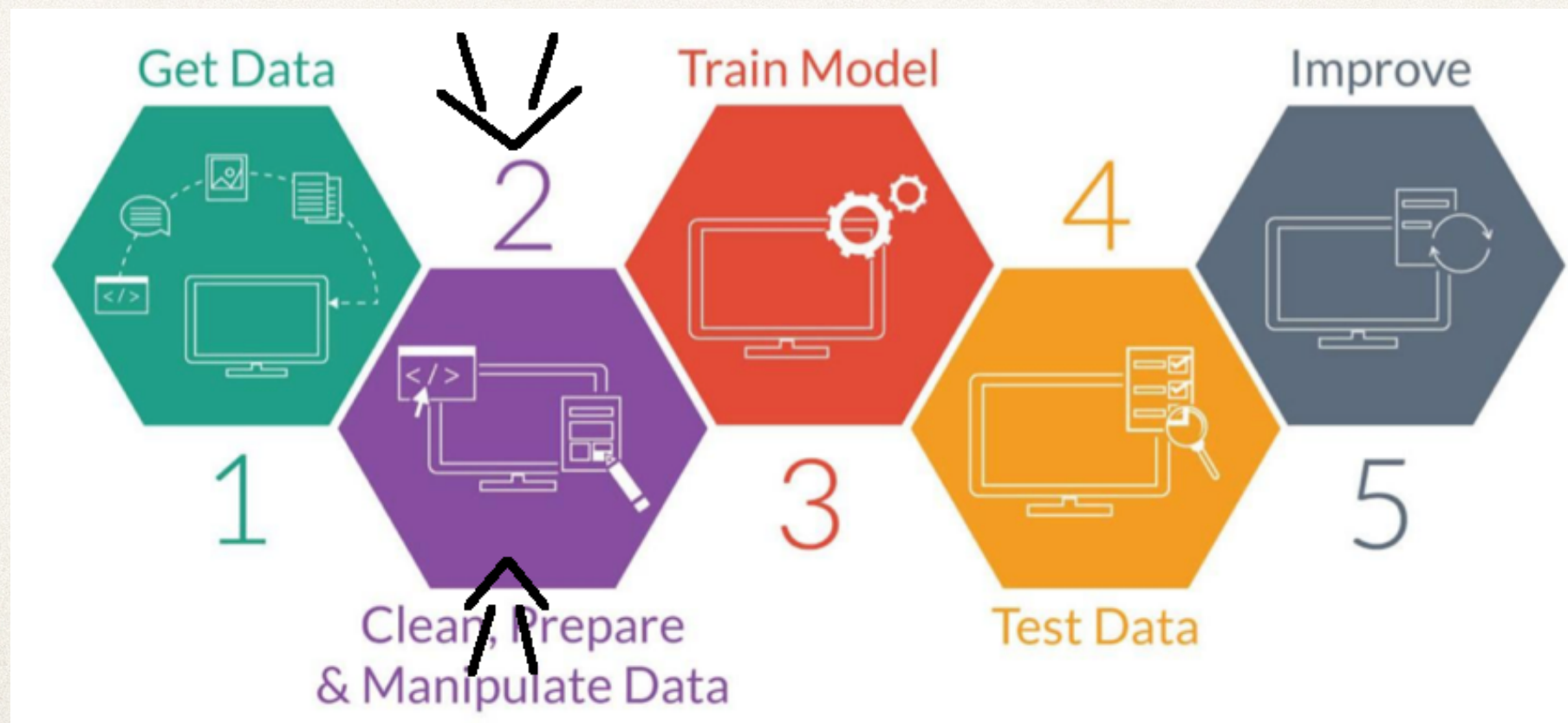
Curse of Dimensionality

- ❖ As we add more features, the number of cells grows exponentially
- ❖ Hence, more data is necessary to fill all those cells
- ❖ Usually the amount of training data is limited



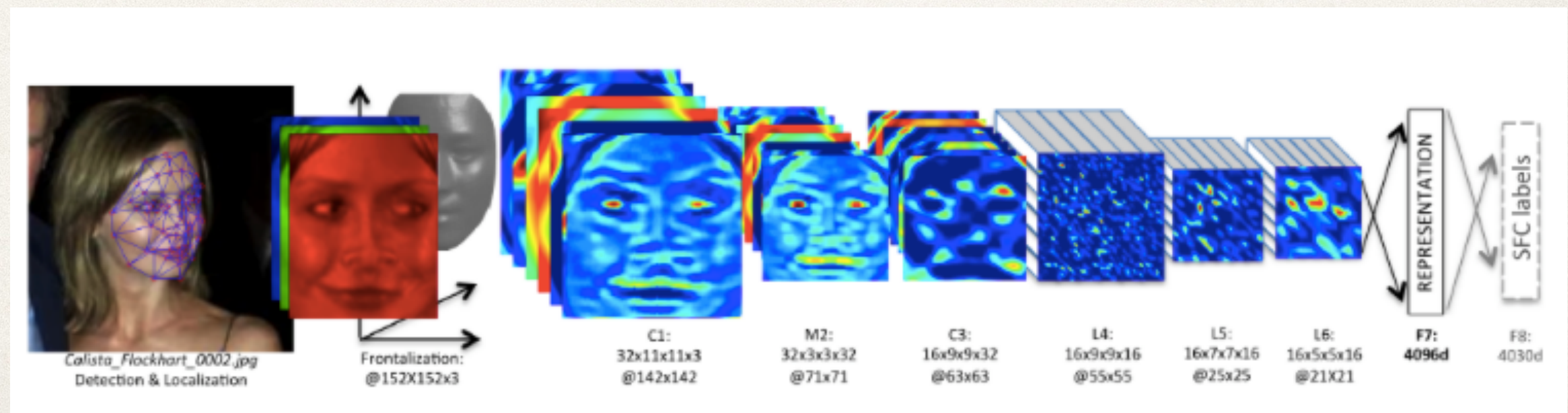
Supervised Learning: Going Deep

- ❖ The weakest link
 - ❖ The representation defined by the human being.



Deep Learning

- ❖ aka Representation Learning
 - ❖ Learning the representation from data
 - ❖ Huge amount of parameters
 - ❖ Lots of data!

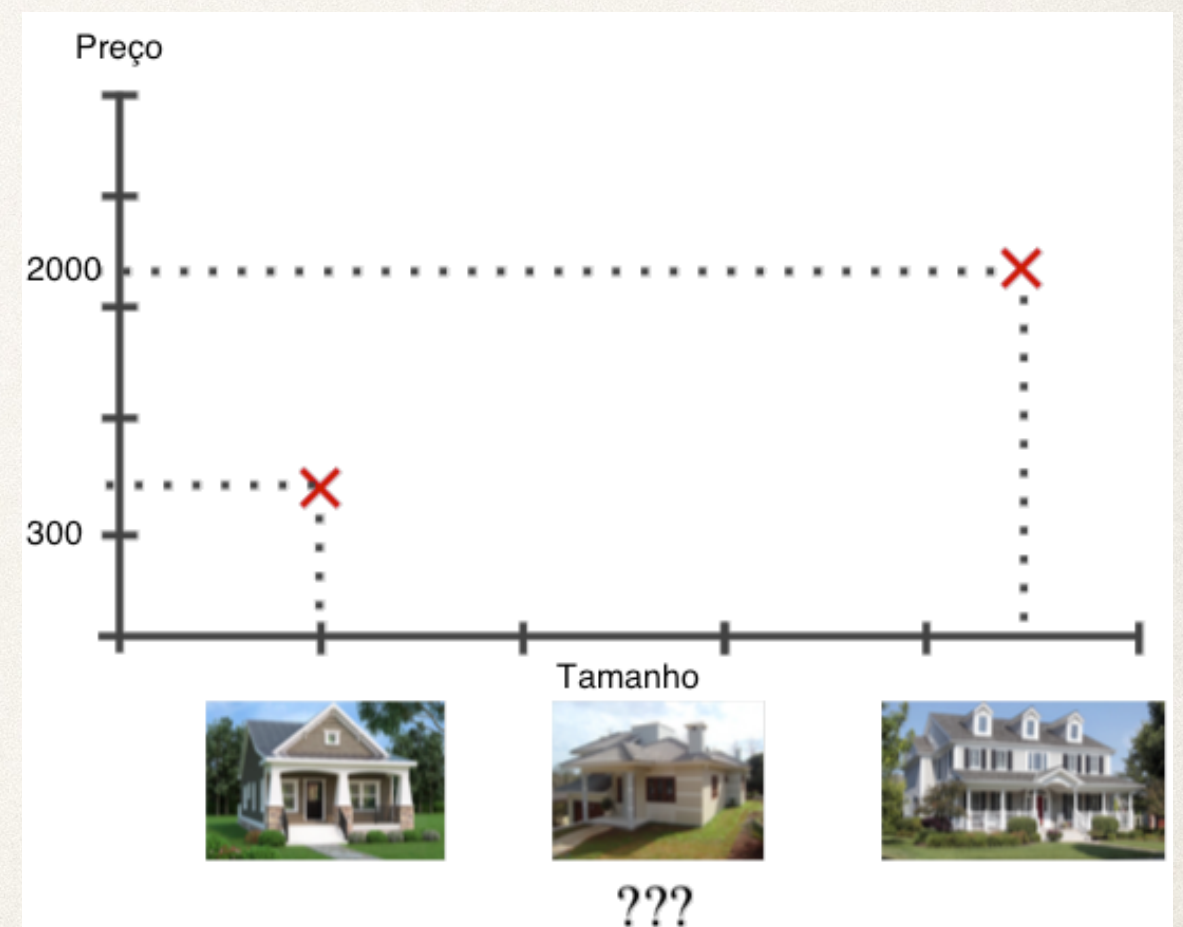


Supervised Learning: Regression



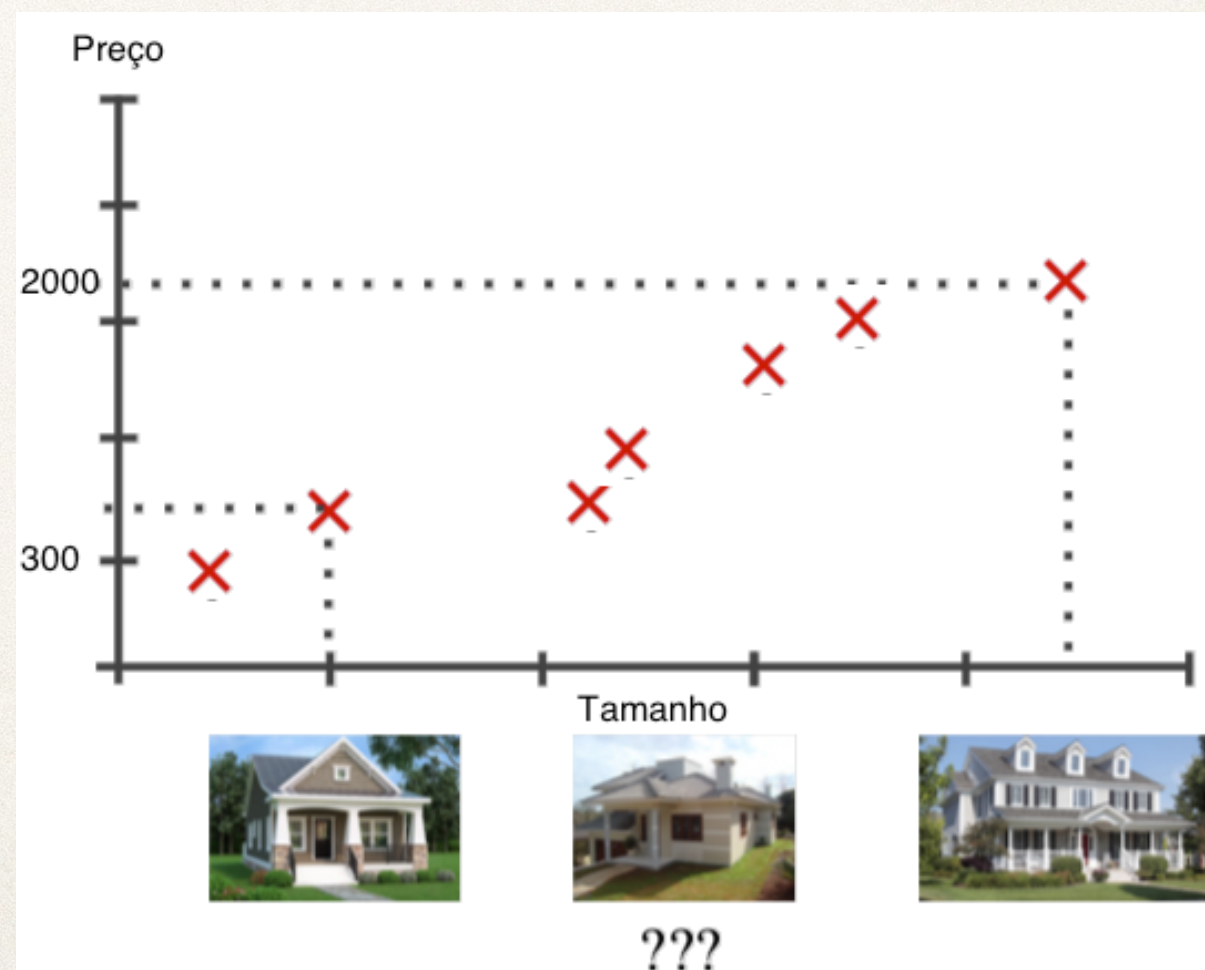
Supervised Learning: Regression

- ❖ Based on the price of several houses, the regression task consists in estimating the price of a given house
- ❖ Differently from the classification, in this case the output is a real value.



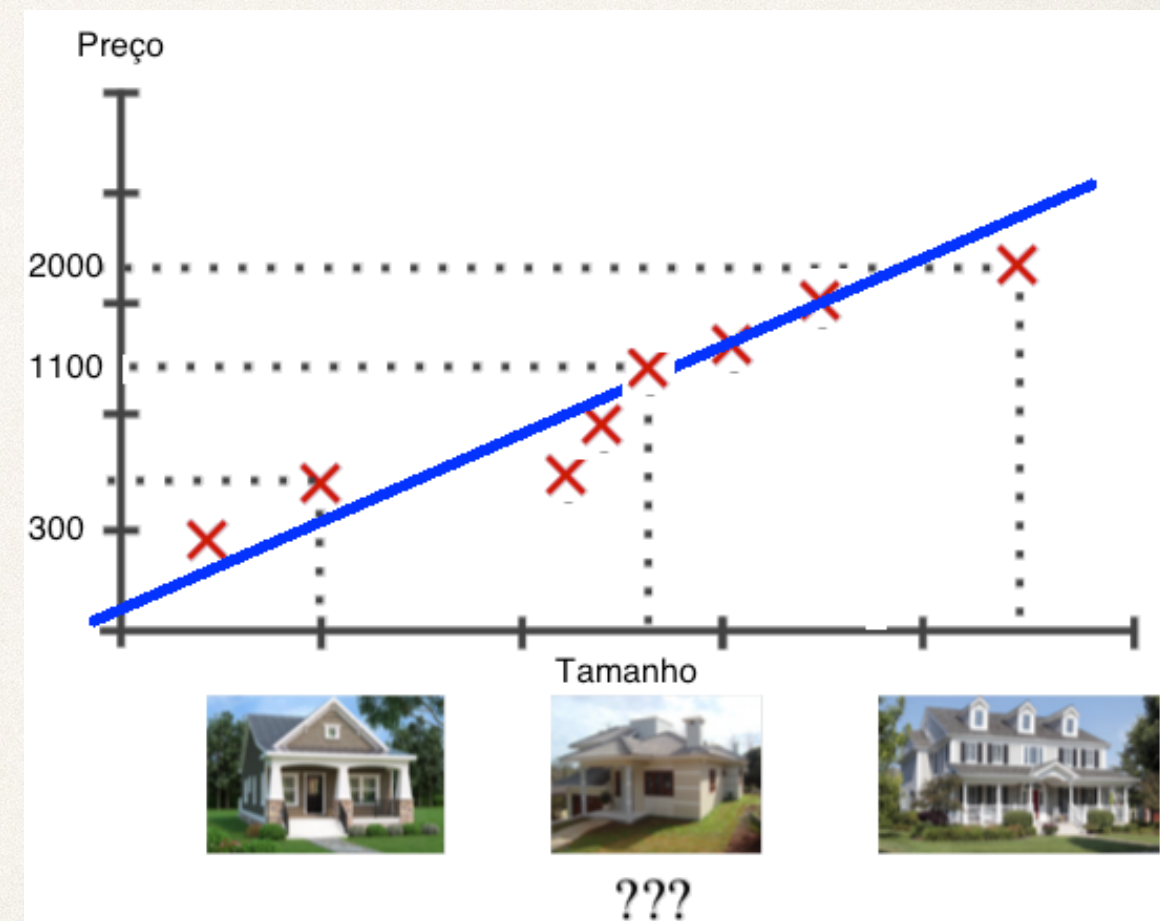
Supervised Learning: Regression

- ❖ Base on this small dataset, what would be the price for this house?

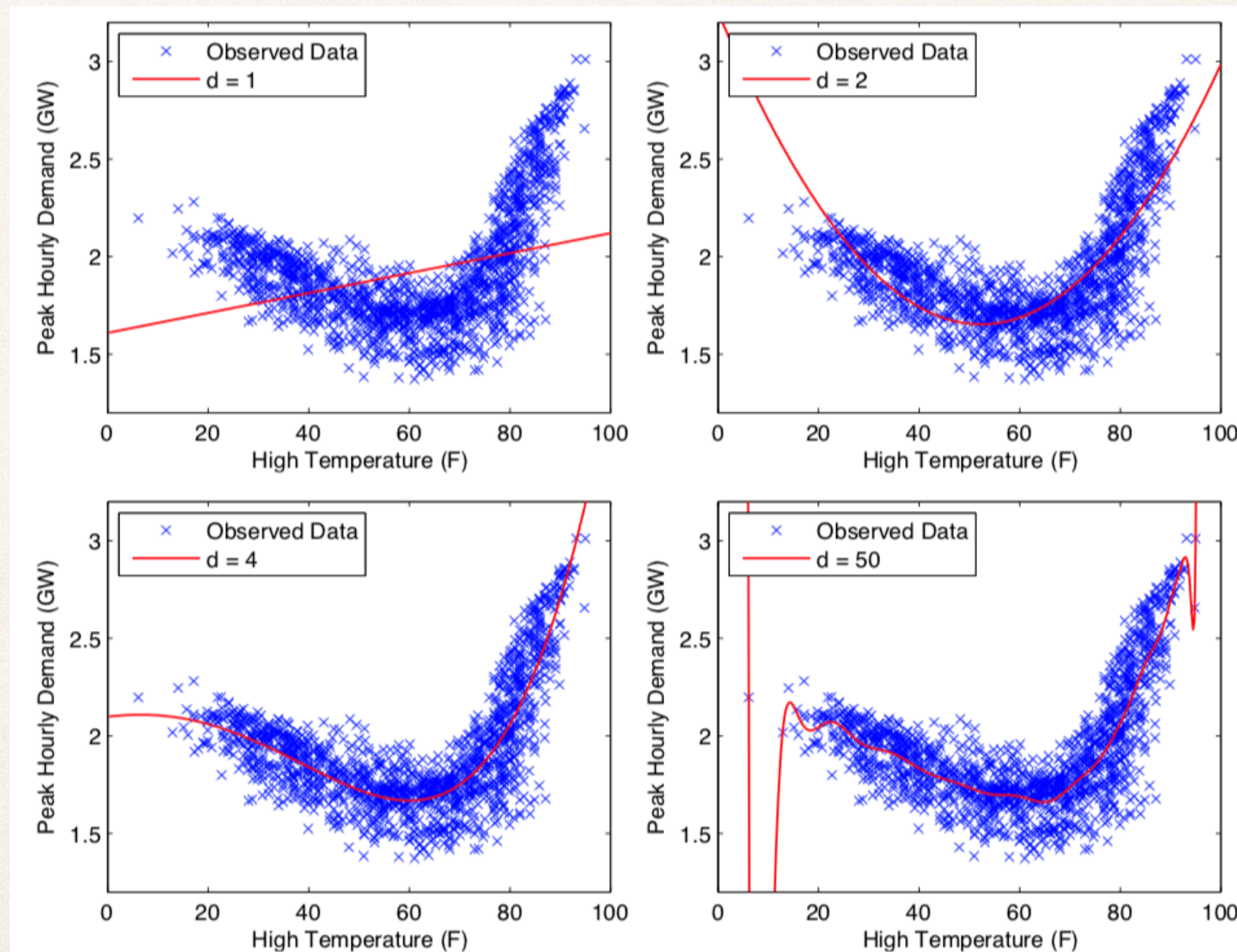


Supervised Learning: Regression

- ❖ Given a labeled dataset, we can find a model to estimate the price of other houses
- ❖ A very simple model would be the linear regression
- ❖ One must take care to not extrapolate the data



Supervised Learning: Regression



Under vs overfitting

Unsupervised Learning

- ❖ In supervised Learning we have $[X, y]$, where X and y stand for the feature vector and label, respectively.
- ❖ In Unsupervised Learning, on the other hand, we have only X .
- ❖ Clustering data
 - ❖ Market segmentation
 - ❖ Recommending systems
 - ❖ Clustering patients | clients | documents, etc...

Unsupervised Learning

❖ How many groups?



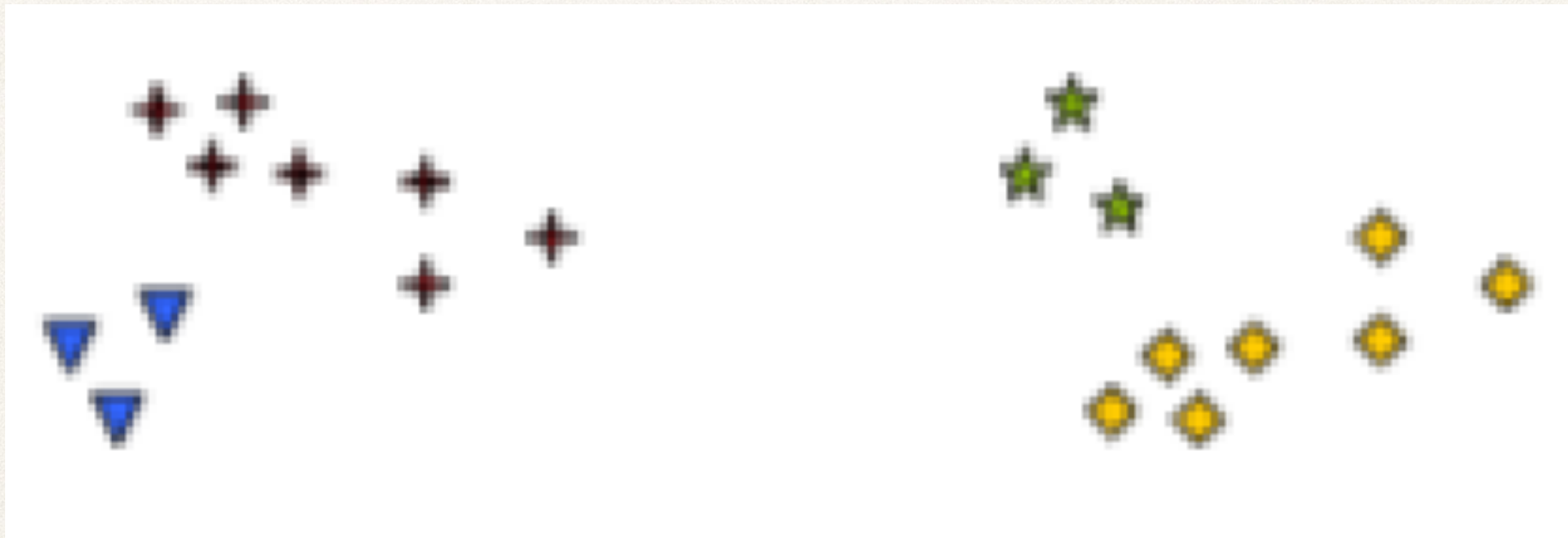
Unsupervised Learning

❖ How many groups? Two



Unsupervised Learning

- ❖ How many groups? Four



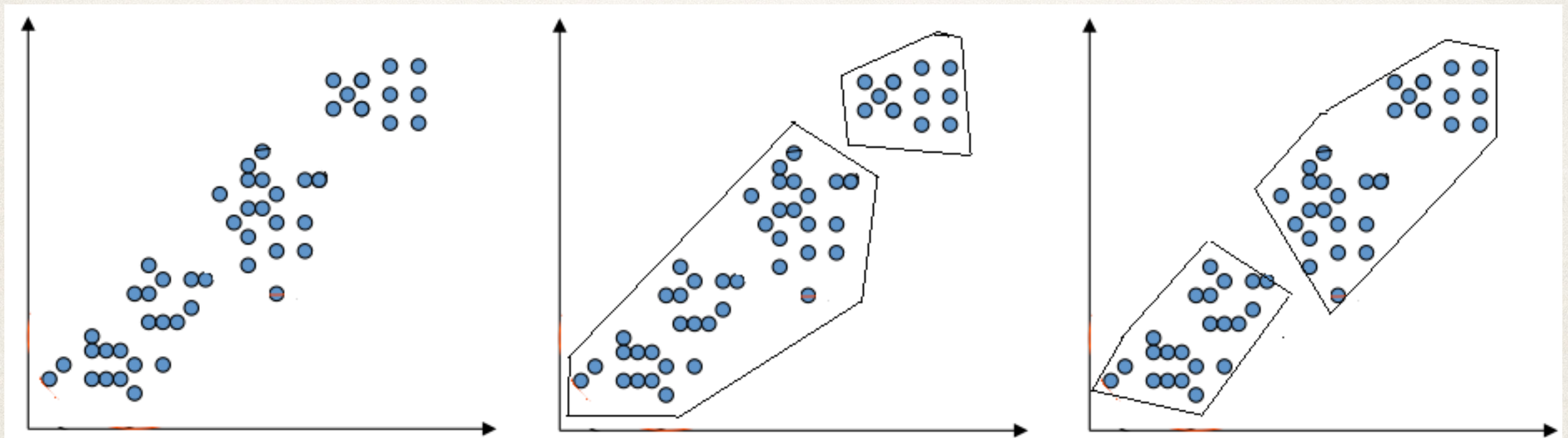
Unsupervised Learning

❖ How many groups? Six



Unsupervised Learning

- ❖ What is a good cluster?
 - ❖ Compact and far from the others.
 - ❖ How to measure that? Cluster indexes.



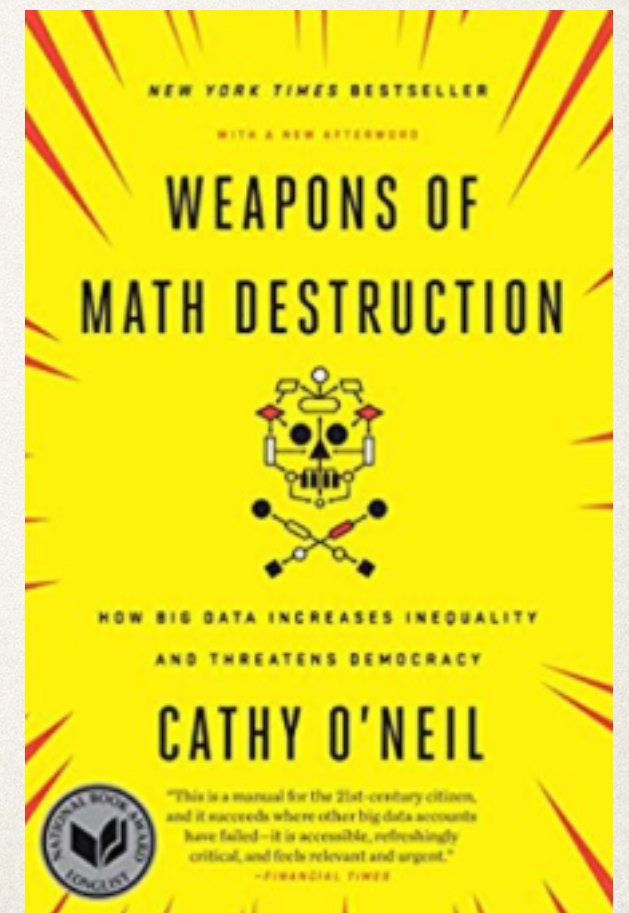
Where?



“It is not a bed of roses”

Ethics & Fairness of ML Systems

- ❖ With the alarming rate of technological progress threatening to eclipse our own understanding of our creations
- ❖ Actions to ensure benevolent AI
- ❖ Drop our blind faith in big data
 - ❖ Data may be biased





Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Ethics & Fairness of ML Systems

Online Ads for High-Paying Jobs Are Targeting Men More Than Women New study uncovers gender bias

When Algorithms Discriminate

The online world is shaped by forces beyond our control, the stories we read on Facebook, the people we meet on OkCupid, the search results we see on Google. Big data is used to make decisions in health care, employment, housing, education and policing.

Amazon just showed us that 'unbiased' algorithms can be inadvertently racist

Technology

Google apologises for Photos app's racist blunder

© 1 July 2016 Technology



TOM SIMONITE BUSINESS 08.21.17 09:00 AM

MACHINES TAUGHT BY PHOTOS LEARN A SEXIST VIEW OF WOMEN

Do Google's 'unprofessional hair' results show it is racist?
Leigh Alexander

... of black women, which some say is just reflecting the wider social landscape



... search for 'unprofessional hair for work' 'DfP' and 'unprofessional hair for work' 'negative' tags

When it Comes to Policing, Data Is Not Benign

The New York Times <https://nyti.ms/2B9YoW>

Opinion | OP-ED CONTRIBUTOR

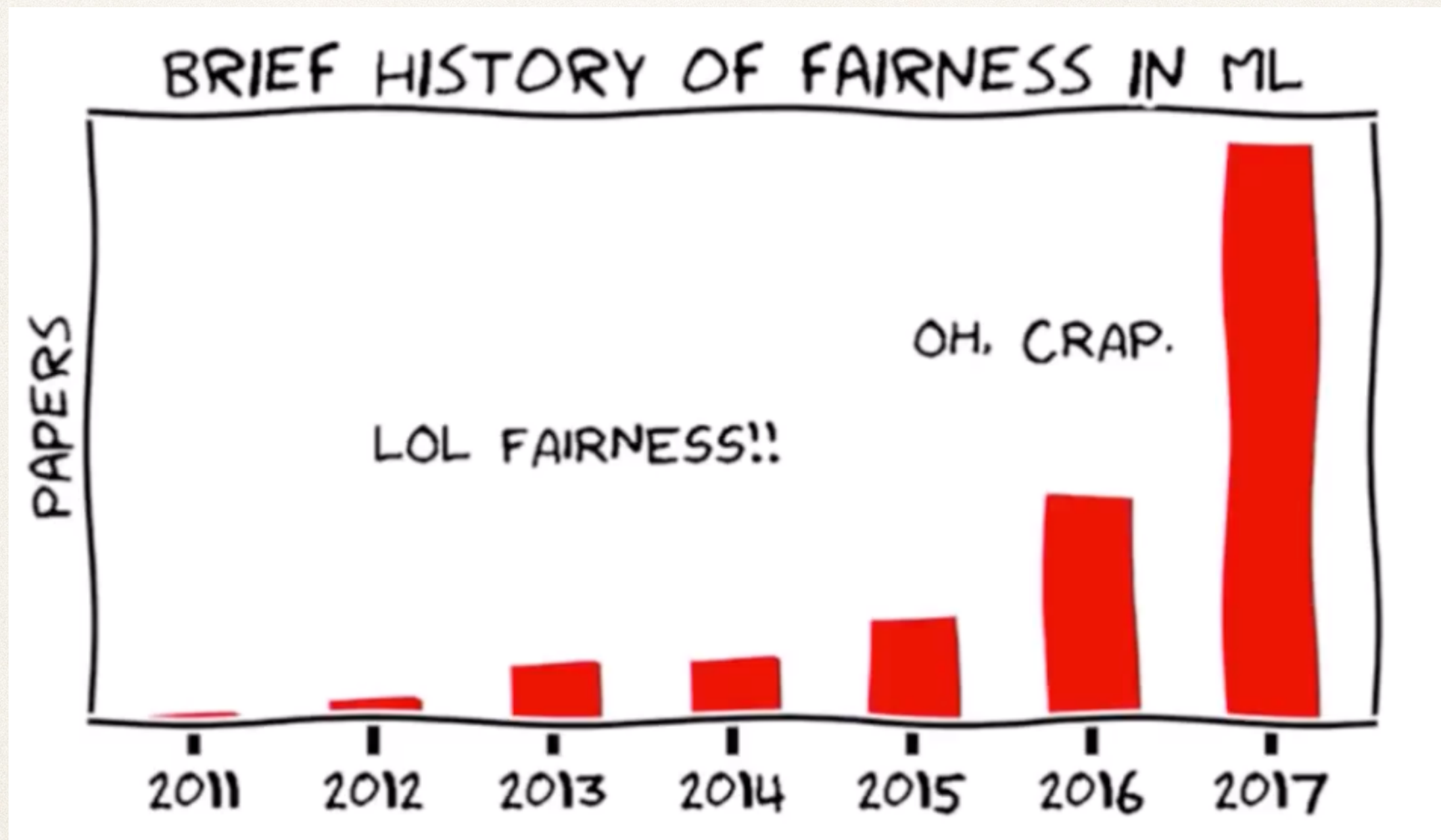
When an Algorithm Helps Send You to Prison

By ELLORA THADANEY ISRANI OCT. 26, 2017

Amazon Prime and the racist algorithms

https://www.youtube.com/watch?v=fMym_BKWQzk&t=6s

Ethics & Fairness of ML Systems



Reproducibility

- ❖ Replication

- ❖ The ultimate standard for strengthening scientific evidence is replication of findings and conducting studies with independent investigators, data, lab, etc.
- ❖ If we are able to replicate the study and produce the same results, then one can say that the original study is true.

Reproducibility

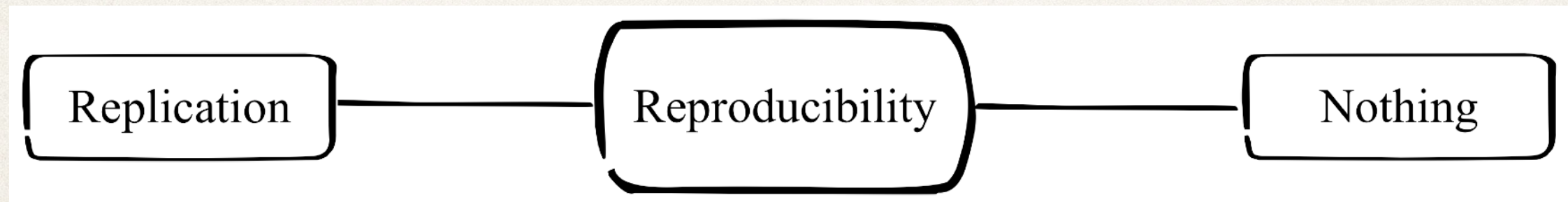
- ❖ What is the problem with replication?
 - ❖ Studies are getting bigger and bigger.
 - ❖ Collecting labelled data is expensive.
 - ❖ Some machine learning models require very expensive hardware for training.
 - ❖ In summary, sometimes replicating an entire research is not feasible.

Replication

Nothing

Reproducibility

- ❖ It is something in between the gold standard (replication) and nothing.
- ❖ Make data and code available so that others may reproduce your findings.



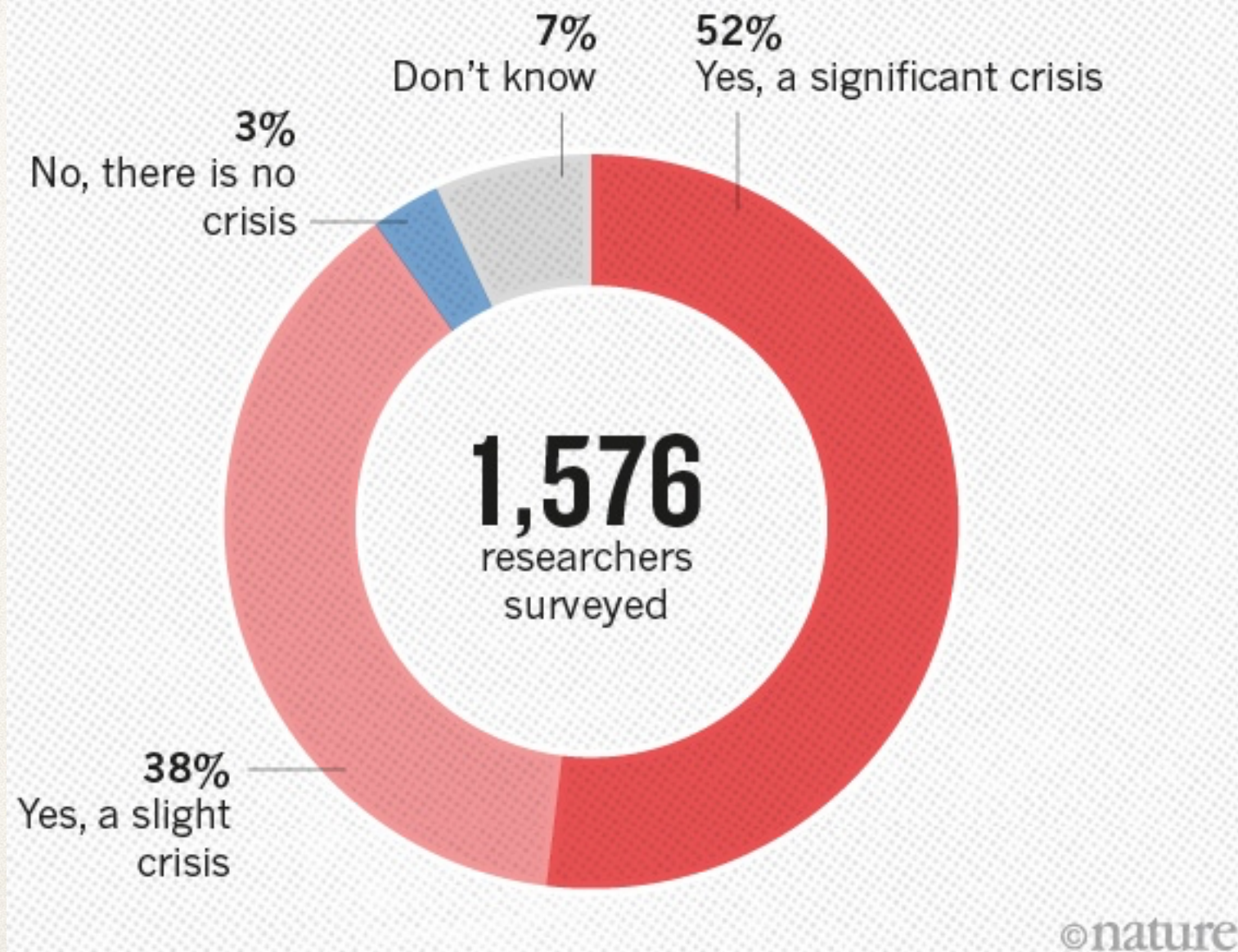
Reproducibility Crisis

- ❖ Several studies point out that it is increasingly recognized that computational science is facing a credibility crisis.
- ❖ It is impossible to verify most of the computational results presented at conferences and papers today.
- ❖ In biomedical research, it is estimated that 85% of research investment is wasted*. Findings are not reproducible.

*Macleod et al., Biomedical Research: Increasing value, reducing waste. The Lancet, 383, 2014.

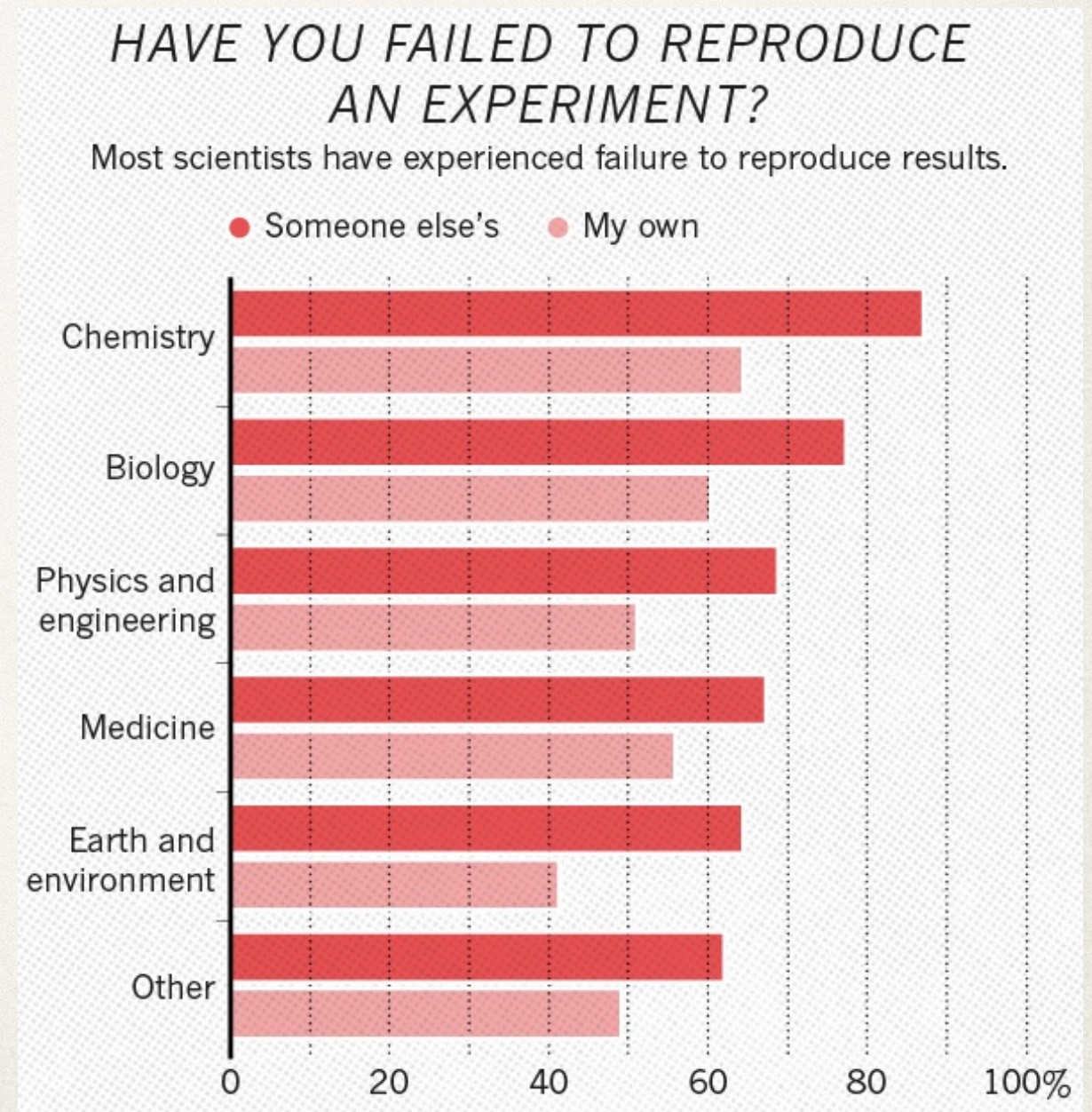
Reproducibility Crisis

IS THERE A REPRODUCIBILITY CRISIS?



Reproducibility Crisis

- ❖ Parameter values, function invocation sequences, and other computational details are typically omitted from published articles but are critical for replicating results or reconciling sets of independently generated results.



Corrective Measures

- ❖ Addressing this credibility crisis requires a change in the culture of scientific publishing.
- ❖ In computer science, sharing data, code, and trained models.

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. *The actual scholarship is the complete software development environment and the complete set of instructions which generated the tables figures.*

Tools

- ❖ Sharing code: Github, Gitlab, BitBucket
- ❖ Sharing data:
 - ❖ Kaggle: Focused on competitions
 - ❖ Research Labs (<https://web.inf.ufpr.br/vri/databases/>)
 - ❖ Google Dataset Search

9 resultados encontrados

 Breast Histopathology Images

www.kaggle.com

Atualizado em Dec 19, 2017

 **PLOS** A summary of all features extracted from breast cancer...

plos.figshare.com

Atualizado em Dec 3, 2015



Breast Cancer Cell Segmentation

academictorrents.com



Data from: Accurate and reproducible invasive breast...

explore.openaire.eu

Criado em Apr 1, 2017

 **PLOS** Performance Analysis of Proposed Method.

plos.figshare.com

Atualizado em Sep 21, 2016

 **PLOS** Performance evaluation and comparison of the proposed...

plos.figshare.com



Breast Histopathology Images

198,738 IDC(-) image patches; 78,786 IDC(+) image patches

175 artigos acadêmicos citam este conjunto de dados ([ver no Google Acadêmico](#))

Conjunto de dados atualizado Dec 19, 2017

Autores

Paul Mooney

Licença

[CC0: Public Domain](#)

Formatos de download disponibilizados pelos provedores

zip (1598158593 bytes), zip (1644892042 bytes)

Descrição

Context

Invasive Ductal Carcinoma (IDC) is the most common subtype of all breast cancers. To assign an aggressiveness grade to a whole mount sample, pathologists typically focus on the regions which contain the IDC. As a result, one of the common pre-processing steps for automatic aggressiveness grading is to delineate the exact regions of IDC inside of a whole mount slide.

Content

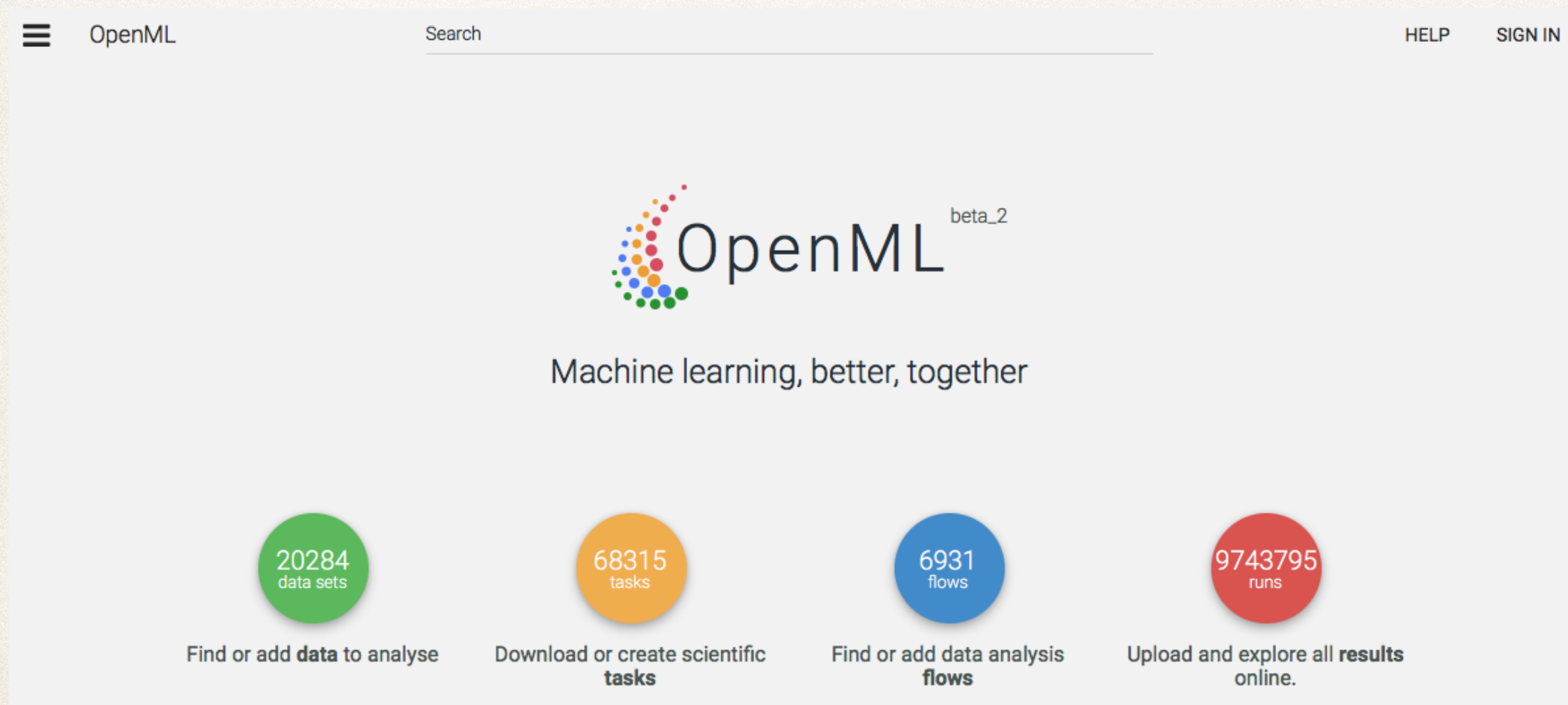
The original dataset consisted of 162 whole mount slide images of Breast Cancer (BCa) specimens scanned at 40x. From that, 277,524 patches of size 50 x 50 were extracted (198,738 IDC negative and 78,786 IDC positive). Each patch's file name is of the format: u_xX_yY_classC.png —> example 10253_idx5_x1351_y1101_class0.png . Where u is the patient ID (10253_idx5), X is the x-coordinate of where this patch was cropped from, Y is the y-coordinate of where this patch was cropped from, and C indicates the class where 0 is non-IDC and 1 is IDC.

Acknowledgements

You need to submit a markup file to index your data

Tools

- ❖ OpenML*: Collaboration platform to share, organize, and discuss ML experiments, data, and algorithms.



* <https://arxiv.org/pdf/1407.7722.pdf>

Thanks for your attention!

web.inf.ufpr.br/luizoliveira
luiz.oliveira@ufpr.br

5512