RECOD
reasoning for complex data

Machine Learning and Pattern Recognition

A High Level Overview

Prof. Sandra Avila
Institute of Computing (IC/Unicamp)

MC886/MO444, August 2, 2018

The Hype

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

The data economy demands a new approach to antitrust rules

Print edition | Leaders ›

May 6th 2017

Data is not the new oil

Jocelyn Goldfein, Ivy Nguyen Mar 27, 2018

https://techcrunch.com/2018/03/27/data-is-not-the-new-oil/

build software,

than ever to build

Jocelyn Goldfein
Contributor

CADE METZ BUSINESS 03.08.16 07:00 AM

GOOGLE'S AI IS ABOUT TO BATTLE A GO CHAMPION—BUT THIS IS NO GAME

DeepMind

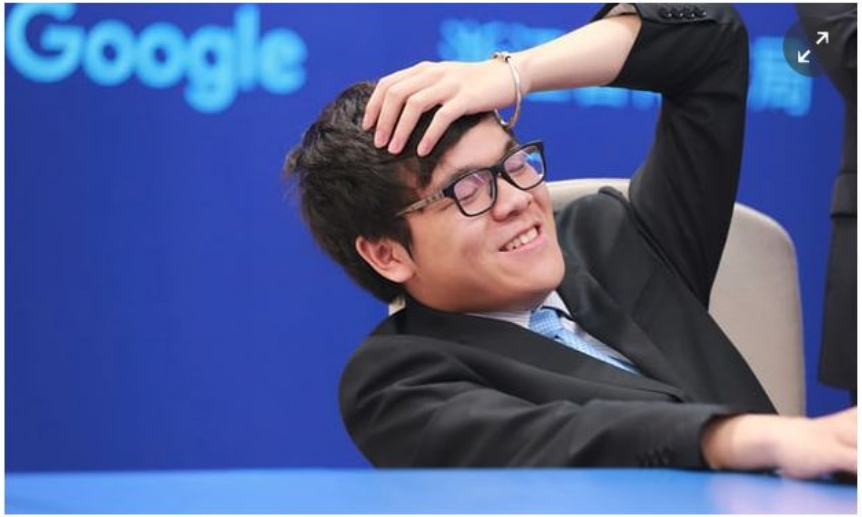
Google's Go-playing AI still undefeated with victory over world number one

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

Alex Hern

@alexhern

Thursday 25 May 2017
09.50 BST



Chinese Go player Ke Jie reacts during his second match against Deepmind's game-playing AI, AlphaGo. Photograph: China Stringer Network/Reuters

Google's Go-playing AI has won its second game against the world's best player of

INDY/TECH

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

3D Printing
Artificial Intelligence
Popular Posts

Tesla lowers Model X base price due to improved margins
2 days ago

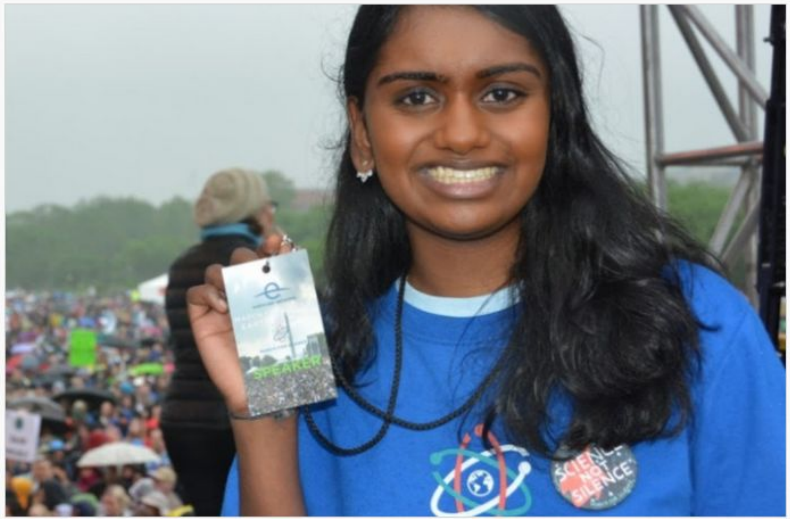
Elon Musk says Model Y small SUV will leverage Model 3 platform after all
5 days ago

Bitcoin breaks $3,000 to reach new all-time high
3 days ago

Tesla has completed its first ever Solar Roof product installations
5 days ago

High schooler makes 3D-printed, machine learning-powered eye disease diagnosis system

Posted Aug 3, 2017 by Devin Coldewey



If, like me, you're one of those people who worries that you haven't accomplished much in

NIPS: Conference on Neural Information Processing Systems

View Earlier Meetings » | Accepted Pape

The ENTIRE conference has sold out.

Announcements

- The ENTIRE conference has sold out. Tutorials, Conference and Workshops are sold out.
- If you are a **presenter** on a **tutorial**, **talk**, or **poster** you may still register. NIPS has held a number of tickets in reserve. When your presentation is visible in your profile, you will be able to register as you normally would using the green button above. If you don't see your presentation, verify that you used the same email address at NIPS.cc and CMT. See merge

Skin cancer

Computer learns to detect skin cancer more accurately than doctors

Artificial intelligence machine found 95% of melanomas in study compared to 86.6% for dermatologists

Agence France Presse

Tue 29 May 2018 03.06 BST

▲ An computer that was taught to distinguish malignant moles from benign ones outperformed dermatologists. Photograph: Dan Himbrechts/AAP

ANNALS OF ONCOLOGY

Article Contents

CORRECTED PROOF

Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists FREE

H A Haenssle ✉, C Fink, R Schneiderbauer, F Toberer, T Buhl, A Blum, A Kalloo, A Ben Hadj Hassen, L Thomas, A Enk L Uhlmann

Annals of Oncology, mdy166, https://doi.org/10.1093/annonc/mdy166
Published: 28 May 2018

Abstract

Background

Deep learning convolutional neural networks (CNN) may facilitate

ANNALS OF ONCOLOGY

ESM

Article Contents

Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists FREE

H A Haenssle ✉, C Fink, R Schneiderbauer, F Toberer, T Buhl, A Blum, A Kalloo, A Ben Hadj Hassen, L Thomas, A Enk L Uhlmann

A Ben Hadj Hassen
Faculty of Computer Science and Mathematics, University of Passau, Germany

L Uhlmann
Institute of Medical Biometry and Informatics, University of Heidelberg, Germany

Deep learning convolutional neural networks (CNN) may facilitate

Intelligent Machines

Machine learning predicts World Cup winner

Researchers have predicted the outcome after simulating the entire soccer tournament 100,000 times.

by Emerging Technology from the arXiv June 12, 2018

The 2018 soccer World Cup kicks off in Russia on Thursday and is likely to be one of the most widely viewed sporting events in history, more popular even than the Olympics. So the potential winners are of significant interest.

One way to gauge likely outcomes is to look at bookmakers' odds. These companies use professional statisticians to analyze extensive databases of results in a way that quantifies the probability of different outcomes of any possible match. In this way, bookmakers can offer odds on all the games that will kick off in the next few weeks, as well as odds on potential winners.

"According to the most probable tournament course, instead of the Spanish the **German team would win the World Cup**."

https://www.technologyreview.com/s/611397/machine-learning-predicts-world-cup-winner/

Why now?

IMAGENET

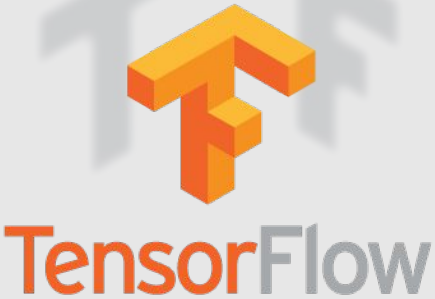
www.image-net.org

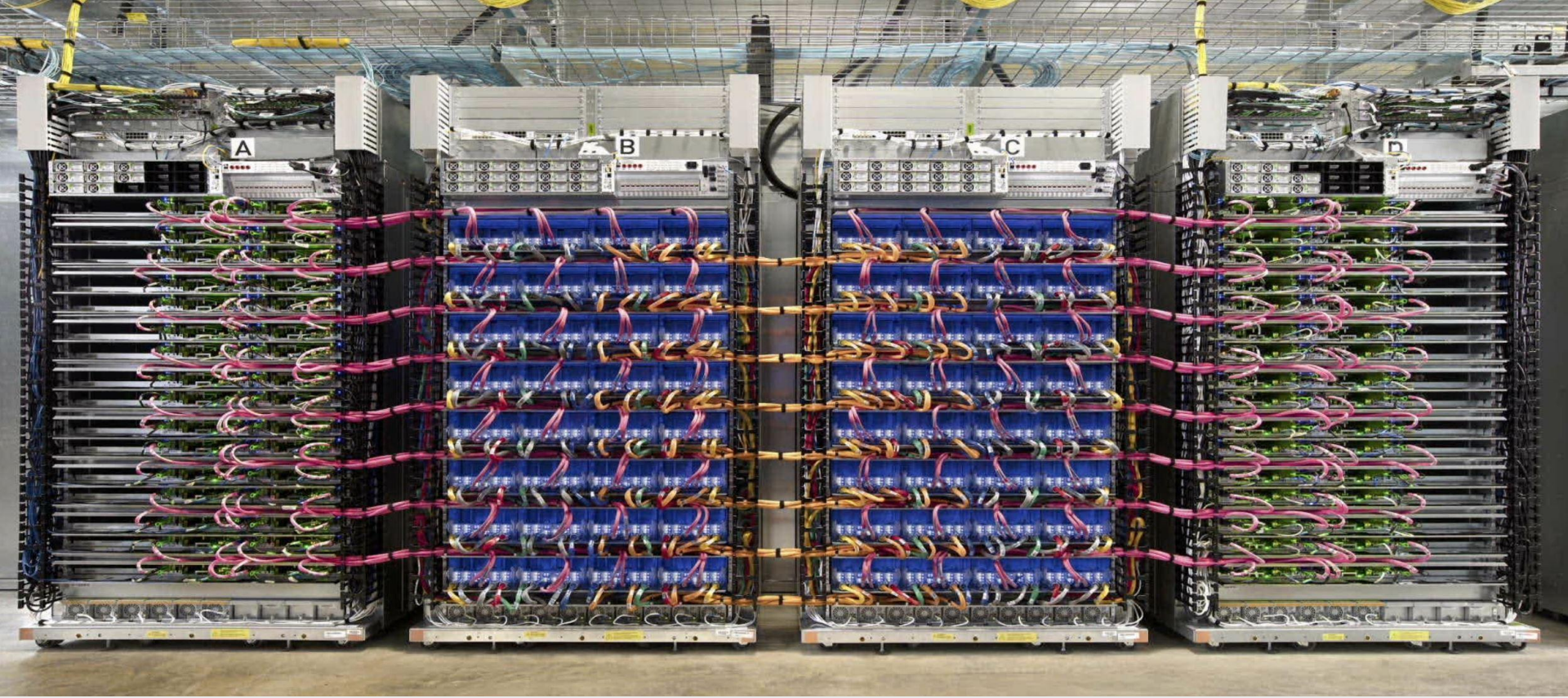
22K categories and 14M images

- Animals
  - Bird
  - Fish
  - Mammal
  - Invertebrate
- Plants
  - Tree
  - Flower
- Food
- Materials
- Structures
- Artifact
  - Tools
  - Appliances
  - Structures
- Person
- Scenes
  - Indoor
  - Geological Formations
- Sport Activities

Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009

Machine Learning Frameworks

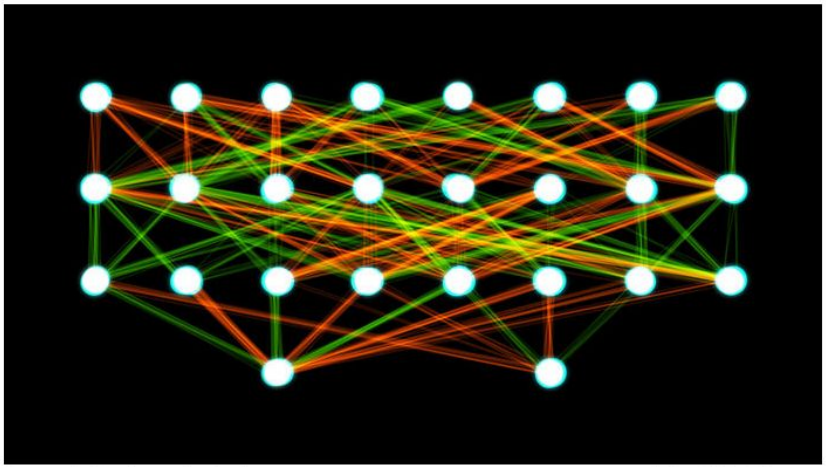




“To create the image and speech recognition algorithms designed by AutoML, Google reportedly let a cluster of **800 GPUs** iterate and crunch numbers for weeks.”

http://www.tomshardware.com/news/google-automl-aritifical-intelligence-ai,34533.html

SHARE



A representation of a neural network. Akritasa/Wikimedia Commons

Brainlike computers are a black box. Scientists are finally peering inside

By Jackie Snow | Mar. 7, 2017 , 3:15 PM

Last month, Facebook announced software that could simply look at a photo and tell, for example, whether it was a picture of a cat or a dog. A related program identifies cancerous

Today's Agenda

- What is Machine Learning?
- Why is this so Important?
- Types of Machine Learning Systems
- Main Challenges of Machine Learning
- Course Logistics

What is Machine Learning?

“Machine Learning is the science (and art) of programming computers so they can learn from data”.

[Aurélien Géron, 2017]

Artificial Intelligence

Machine Learning

Deep Learning

1950 1960 1970 1980 1990 2000 2010

https://www.linkedin.com/pulse/ai-machine-learning-evolution-differences-connections-kapil-tandon

Why is this so important?


MACHINE LEARNING
MACHINE LEARNING EVERYWHERE

Why is this so important?

- Data available at unprecedented scales
  - Petabyte, Exabyte, Zettabyte, Yottabyte scale computing ...
- Impossible for humans to deal with this information overflow
- Imagine the resources required to
  - look at every image in Flickr and categorize it
  - check every inch of Google earth for changes
  - look through all webpages for the interesting ones

Types of Machine Learning Systems

Types of Machine Learning Systems

Trained with human supervision (or not)

Supervised vs.
Unsupervised vs.
Reinforcement learning

Can learn incrementally on the fly (or not)

Online vs.
Batch Learning

How they generalize

Instance based vs.
Model based learning

Types of Machine Learning Systems

Trained with human supervision (or not)

Supervised vs.
Unsupervised vs.
Reinforcement learning

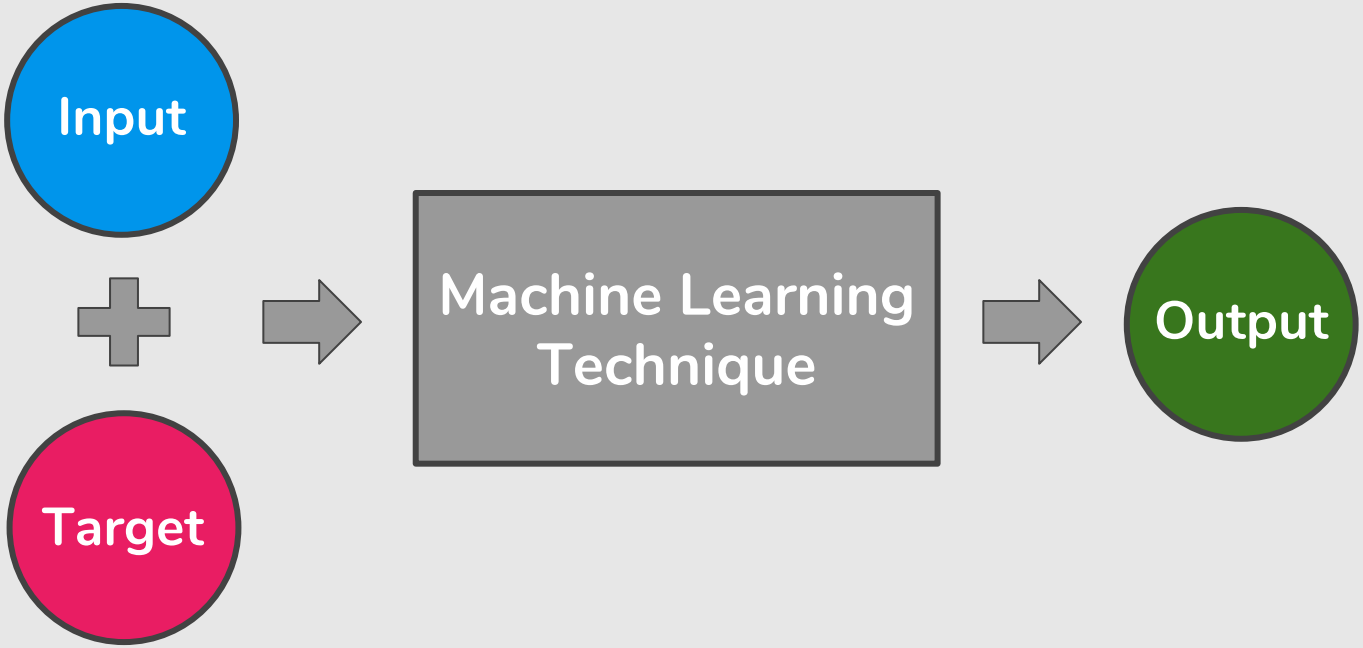
Can learn incrementally on the fly (or not)

Online vs.
Batch Learning

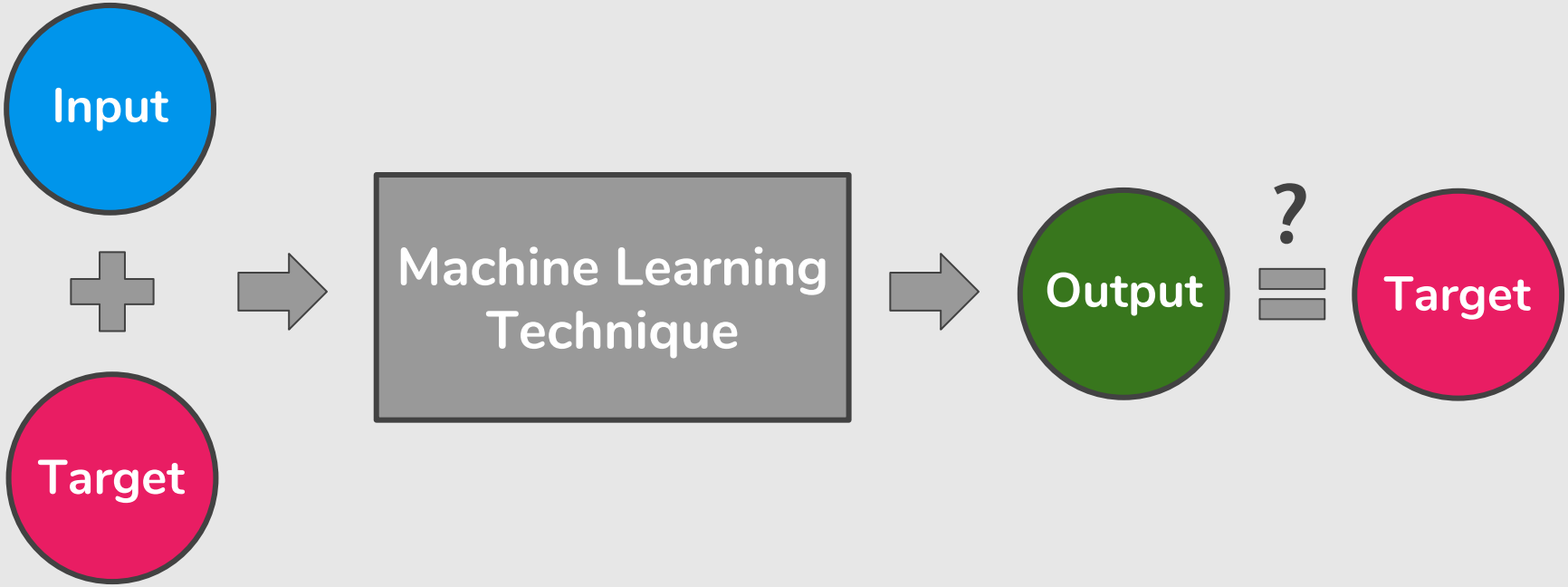
How they generalize

Instance based vs.
Model based learning

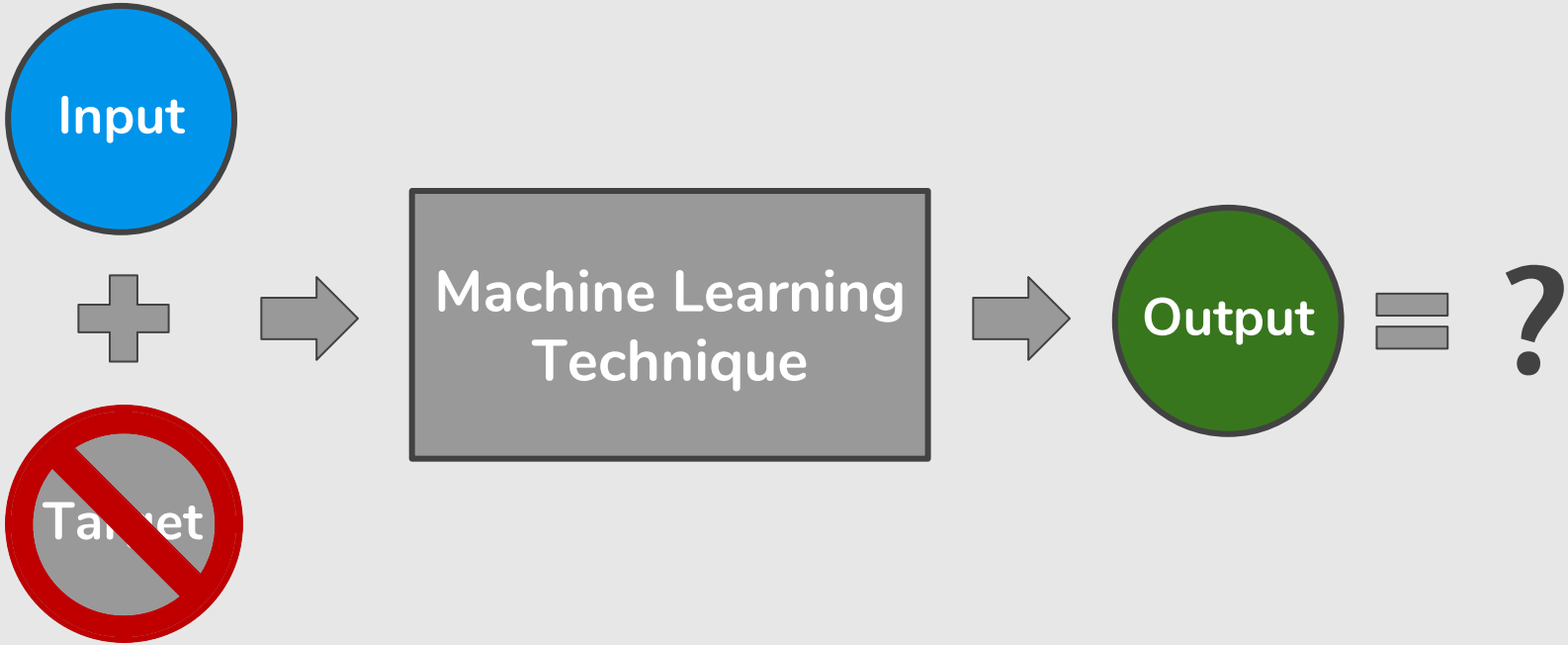
Supervised Learning



Supervised Learning



Unsupervised Learning



Unsupervised Learning

Supervised Learning

Classification is used to predict discrete values (class labels).

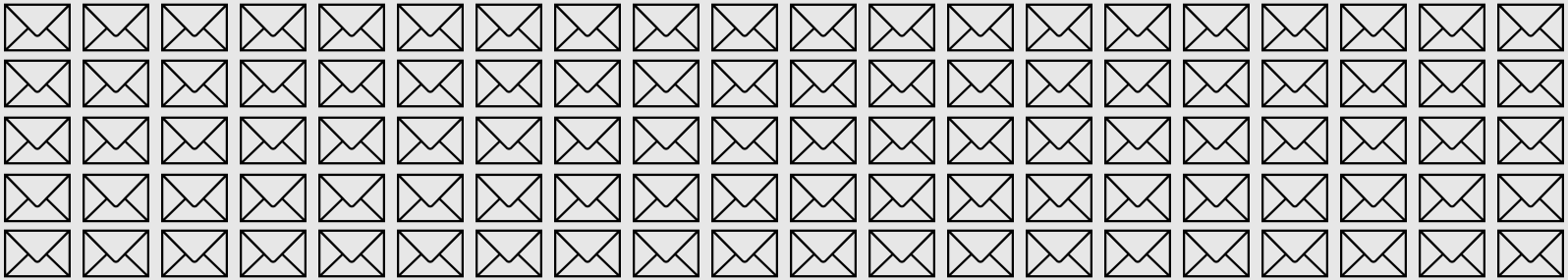
Regression is used to predict continuous values.

Spam Filtering

Bad Cures fast and effective! - Canadian *** Pharmacy #1 Internet Inline Drugstore Viagra Cheap Our price $1.99 …

Good Interested in your research on graphical models - Dear Prof., I have read some of your papers on probabilistic graphical models. Because I …

Spam Filtering

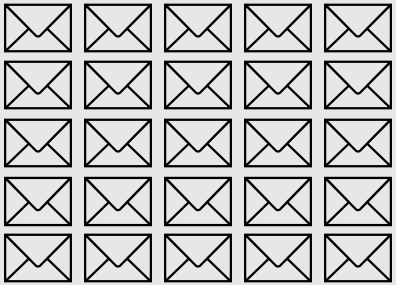


100 emails

https://www.youtube.com/watch?v=IpGxLWOIZy4

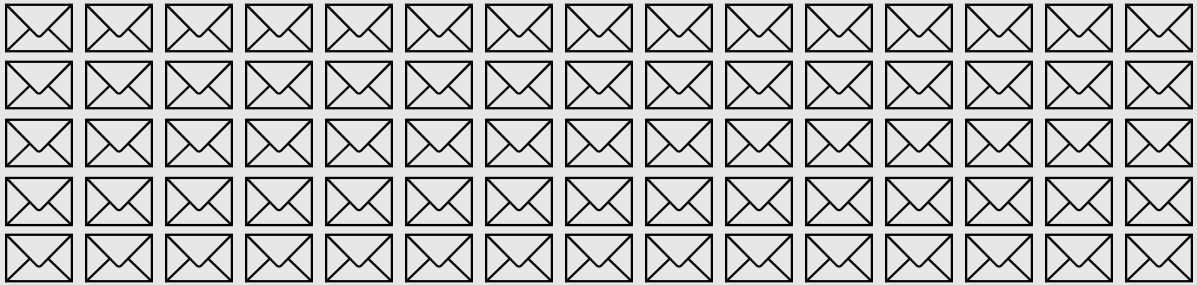
Spam Filtering

Spam



25 emails

Non-spam

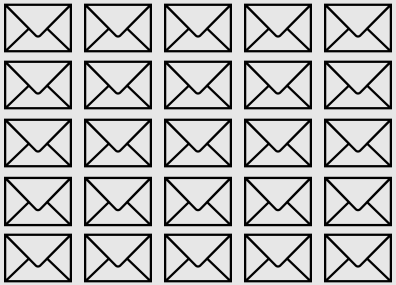


75 emails

Spam Filtering

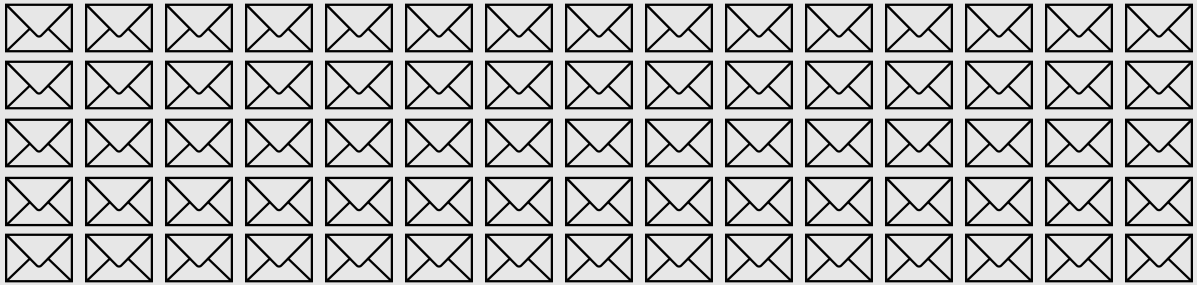
❗ “Cheap”

Spam



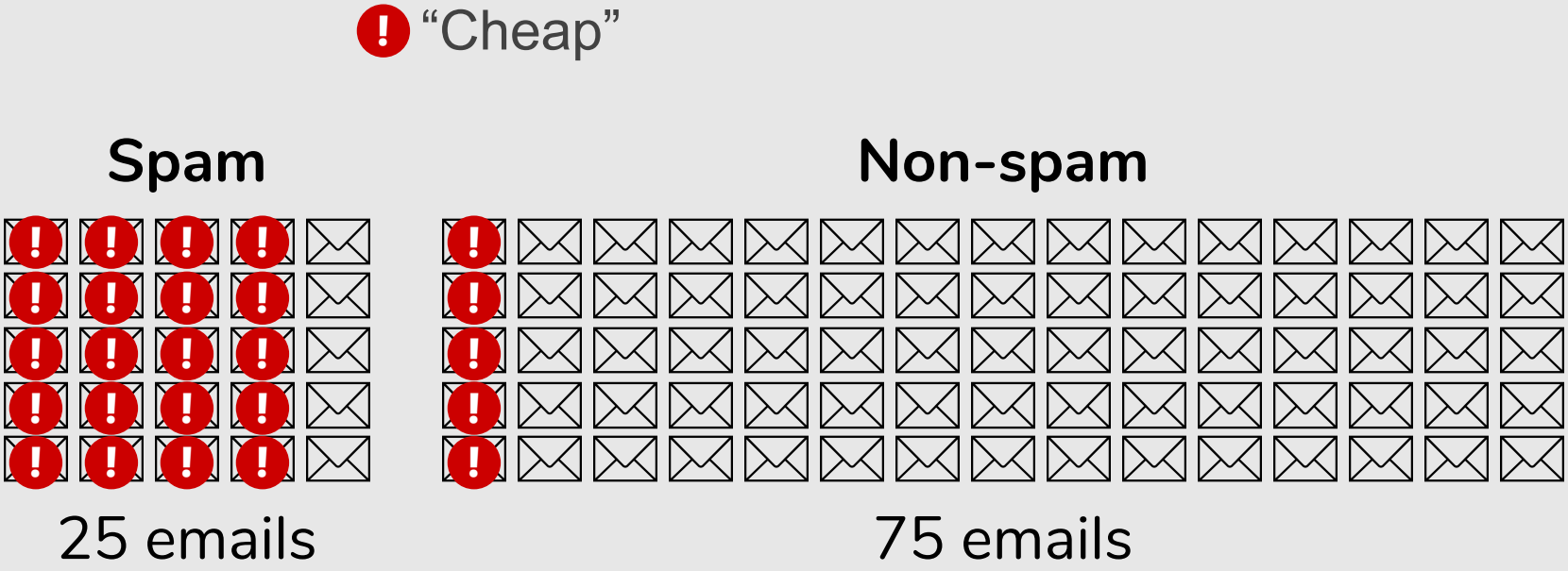
25 emails

Non-spam

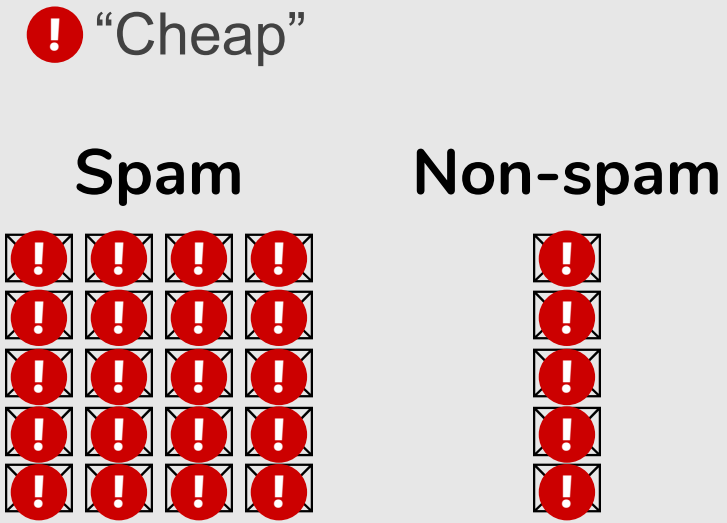


75 emails

Spam Filtering



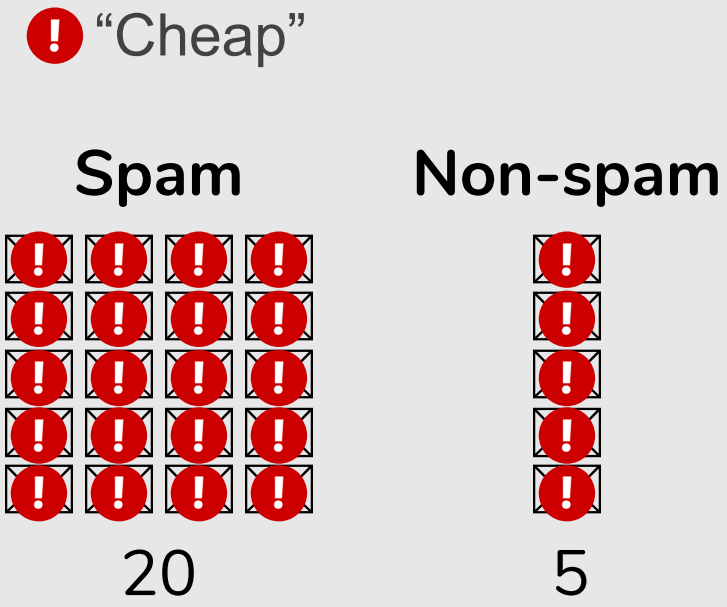
Spam Filtering



If an email contains the word “cheap”, what is the probability of it being spam?

- ☐ 40%
- ☐ 60%
- ☐ 80%

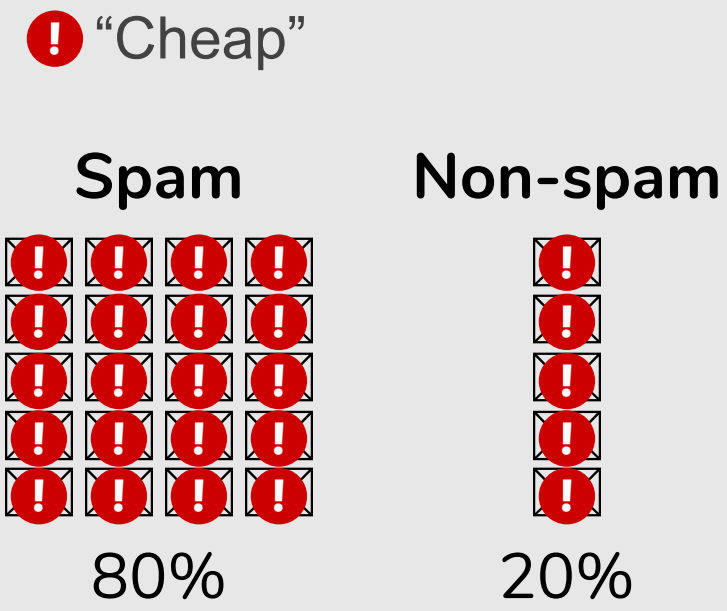
Spam Filtering



If an email contains the word “cheap”, what is the probability of it being spam?

- ☐ 40%
- ☐ 60%
- ☐ 80%

Spam Filtering



If an email contains the word “cheap”, what is the probability of it being spam?

- ☐ 40%
- ☐ 60%
- ☑ 80%

Spam Filtering

- “Cheap” → 80%
- Spelling mistake → 70%
- Missing title → 95%
- etc ...

If an email contains the word “cheap”, what is the probability of it being spam?

- ☐ 40%
- ☐ 60%
- ☑ 80%

Conclusion: If an email contains the word “cheap”, the probability of it being spam is 80%.

Naïve Bayes Algorithm

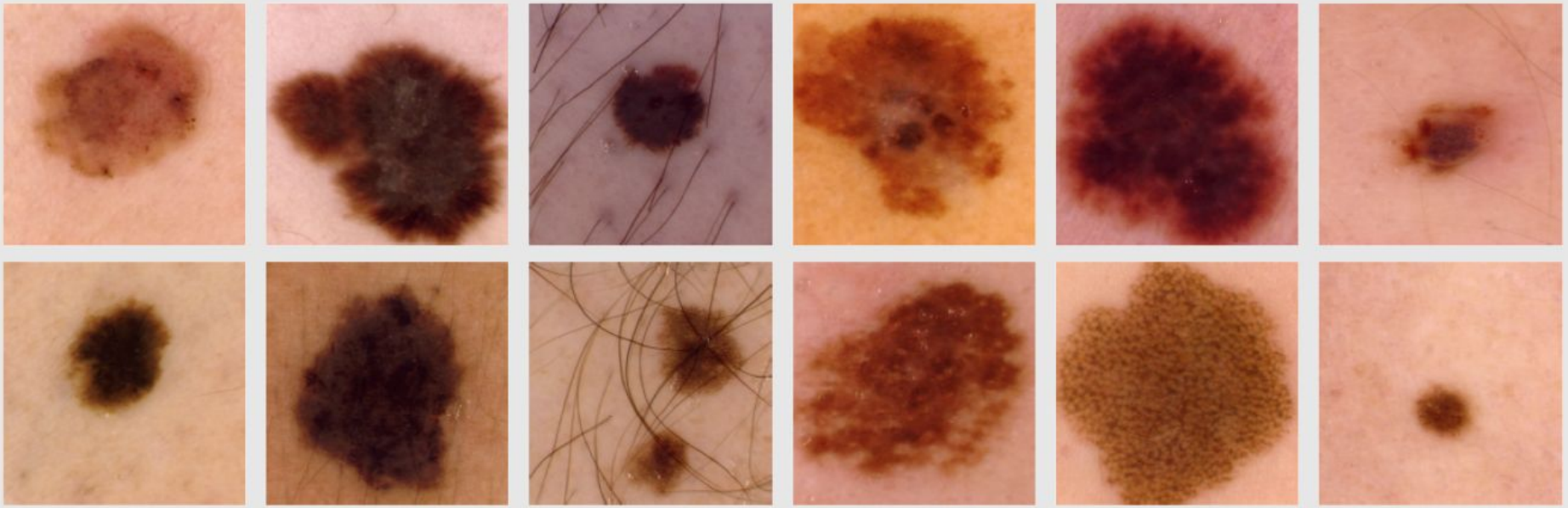
- “Cheap” → 80%
- Spelling mistake → 70%
- Missing title → 95%
- etc ...

If an email contains the word “cheap”, what is the probability of it being spam?

- ☐ 40%
- ☐ 60%
- ☑ 80%

Conclusion: If an email contains the word “cheap”, the probability of it being spam is 80%.

Skin Cancer Classification



Melanomas (top row) and **benign** skin lesions (bottom row)

23, MAR - 2017 | 09:00 | COMUNIDADE INTERNA

Equipe da Unicamp fica no topo de competição internacional de detecção automática de melanoma

Autor Divulgação laboratório RECOD | **Fotos** Mijail Vidal | **Edição de imagem** Paulo Cavalheri

Uma equipe de professores e pesquisadores da Unicamp obteve excelente resultado na segunda edição da Competição Internacional de Análise de Lesões de Pele, evento anual não-presencial organizado pela Colaboração Internacional para Imagens de Lesões de Pele (ISIC). *Os* organizadores disponibilizam

Sensitive Content Classification

Unicamp cria tecnologia para barrar pornografia e violência

Segurança. Pesquisadores lançaram método que identifica cerca de 97% do conteúdo impróprio em telas de celulares e computador

Em parceria com pesquisadores do Samsung Research Institute Brazil, o IC (Instituto de Computação) da Unicamp (Universidade Estadual de Campinas) desenvolveu um método capaz de filtrar 97% do conteúdo pornográfico e 80% do material de violência exibido em telas de celulares, computadores e tablets.

No novo método, os pesquisadores buscaram a combinação do uso de informações estáticas e de movimento com uma metodologia de aprendizado de máquina conhecida como deep learning ou "aprendizagem profunda". Com isso, a solução que o grupo desenvolveu extrai um quadro por segundo de cada vídeo que é acessado em tempo real em celular ou computador. Os quadros com as imagens estáticas são em seguida analisados aplicando-se o método de classificação de descrições do que é permitido e do que é pornográfico.

Ao mesmo tempo, a sequência de quadros analisados fornece os elementos para sequenciar os movimentos dos objetos e pessoas presentes na cena. Dependendo do tipo de movimento, o vídeo é bloqueado.

"Para a detecção de pornografia, os testes foram realizados em um conjunto de dados contendo aproximadamente 140 horas, sendo 1 mil vídeos pornográficos e 1 mil vídeos não pornográficos", explica a pesquisadora do IC da Unicamp, Sandra Avila, ao comentar sobre o processo de criação da tecnologia, que durou 27 meses.


Sistema garante proteção de crianças | IMAGE SOURCE/FOLHAPRESS

"Filtrar cenas de violência, por ser mais subjetivo, é um problema mais difícil comparado à pornografia. Devido a essa subjetividade e os diferentes conjuntos de dados, a eficácia da nossa solução para filtrar cenas de violência está em torno de 80%", conta Sandra.

Ainda segundo a representante da Unicamp, a tecnologia lançada em parceria com a Samsung pode ajudar as autoridades policiais.

"O método proposto para filtrar conteúdo pornográfico está sendo adaptado para outros tipos de conteúdo sensível. Por exemplo, em parceria com peritos da Polícia Federal, estamos desenvolvendo uma ferramenta para detectar pornografia infantil. Temos hoje uma solução que identifica 88% do conteúdo pornográfico infantil em imagens. Para dar uma ideia da importância do resultado, o melhor resultado alcançado pelas ferramentas forenses testadas foi 58%", relata a pesquisadora.

HIDAIANA ROSA
METRO CAMPINAS

House Price Prediction

(Regression)

$ 70 000

https://www.youtube.com/watch?v=IpGxLWOIZy4

House Price Prediction

(Regression)

$ 160 000

House Price Prediction

(Regression)

???

House Price Prediction
(Regression)

House Price Prediction
(Regression)

House Price Prediction
(Regression)

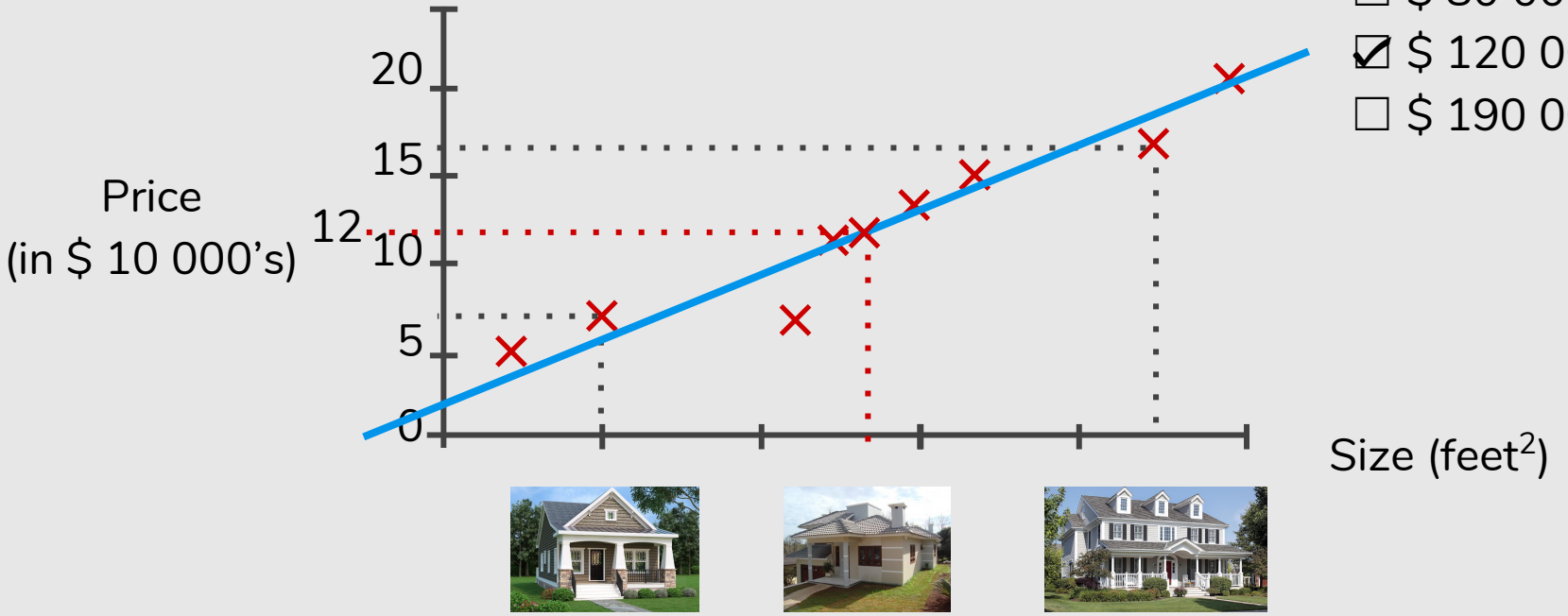
What's the best estimate for the price of the house?

- ☐ $ 80 000
- ☐ $ 120 000
- ☐ $ 190 000

House Price Prediction
(Regression)

What's the best estimate for the price of the house?

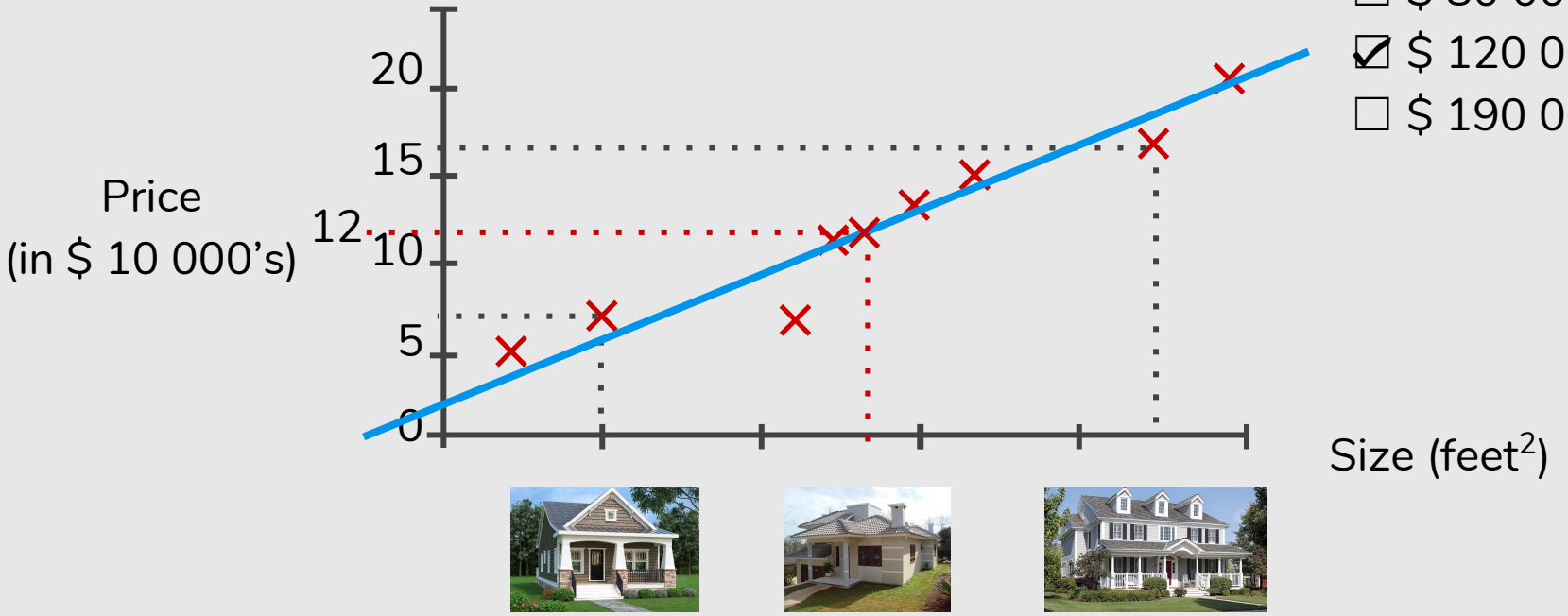
- ☐ $ 80 000
- ☑ $ 120 000
- ☐ $ 190 000



Linear Regression

What's the best estimate for the price of the house?

- ☐ $ 80 000
- ☑ $ 120 000
- ☐ $ 190 000



Important Supervised Learning Algorithms

- Linear Regression
- Logistic Regression
- k-Nearest Neighbors
- Support Vector Machines (SVMs)
- Neural Networks
- Decision Trees and Random Forests

Unsupervised Learning

Clustering algorithm tries to detect similar groups.

Dimensionality reduction tries to simplify the data without losing too much information.

Did anyone say pizza?

https://www.youtube.com/watch?v=IpGxLWOIZy4

Did anyone say pizza?

Did anyone say pizza?

Did anyone say pizza?

Did anyone say pizza?

Did anyone say pizza?

Did anyone say pizza?

Did anyone say pizza?

Did anyone say pizza?

Did anyone say pizza?

Did anyone say pizza?

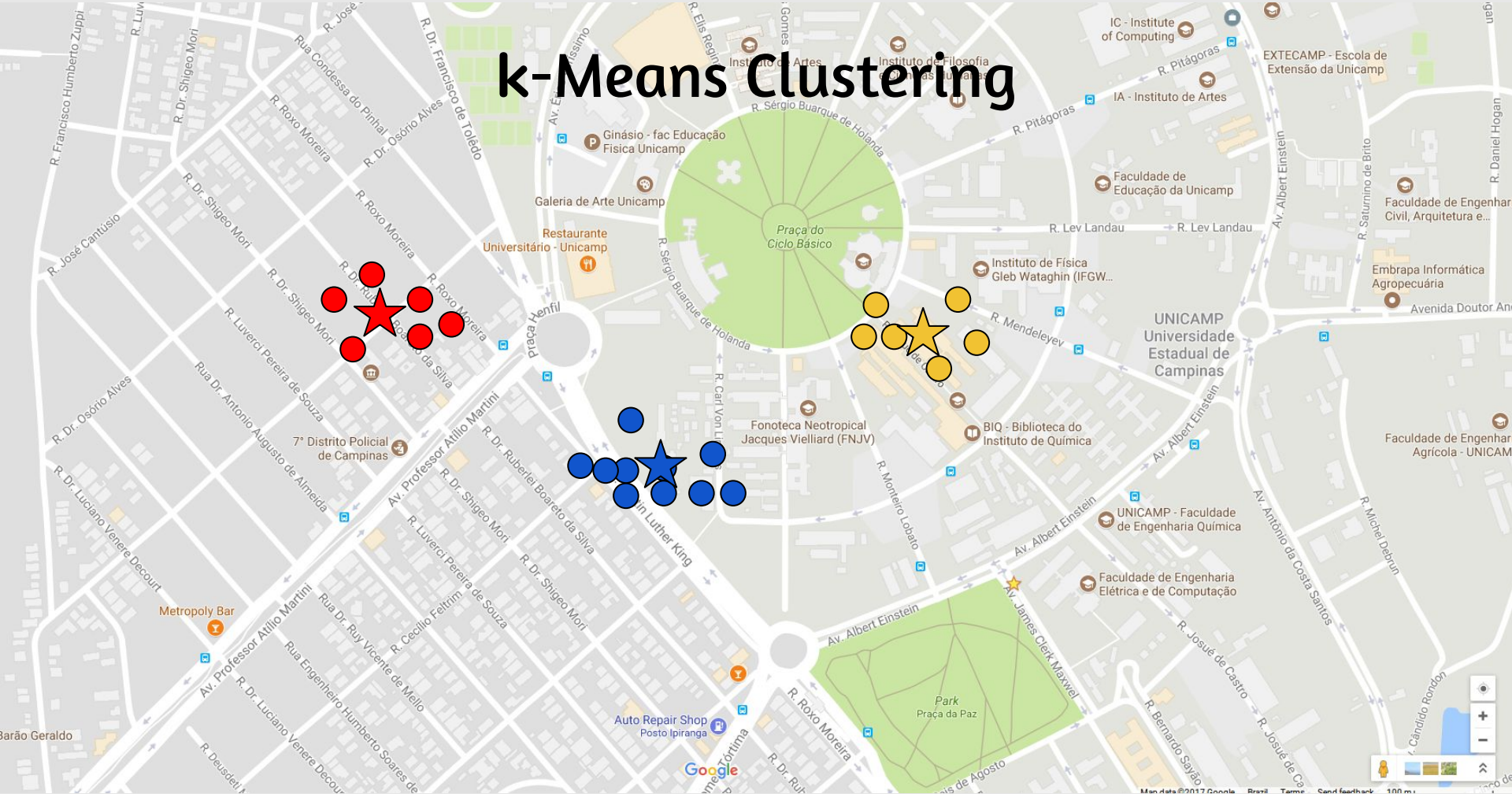
Did anyone say pizza?

Did anyone say pizza?

Did anyone say pizza?

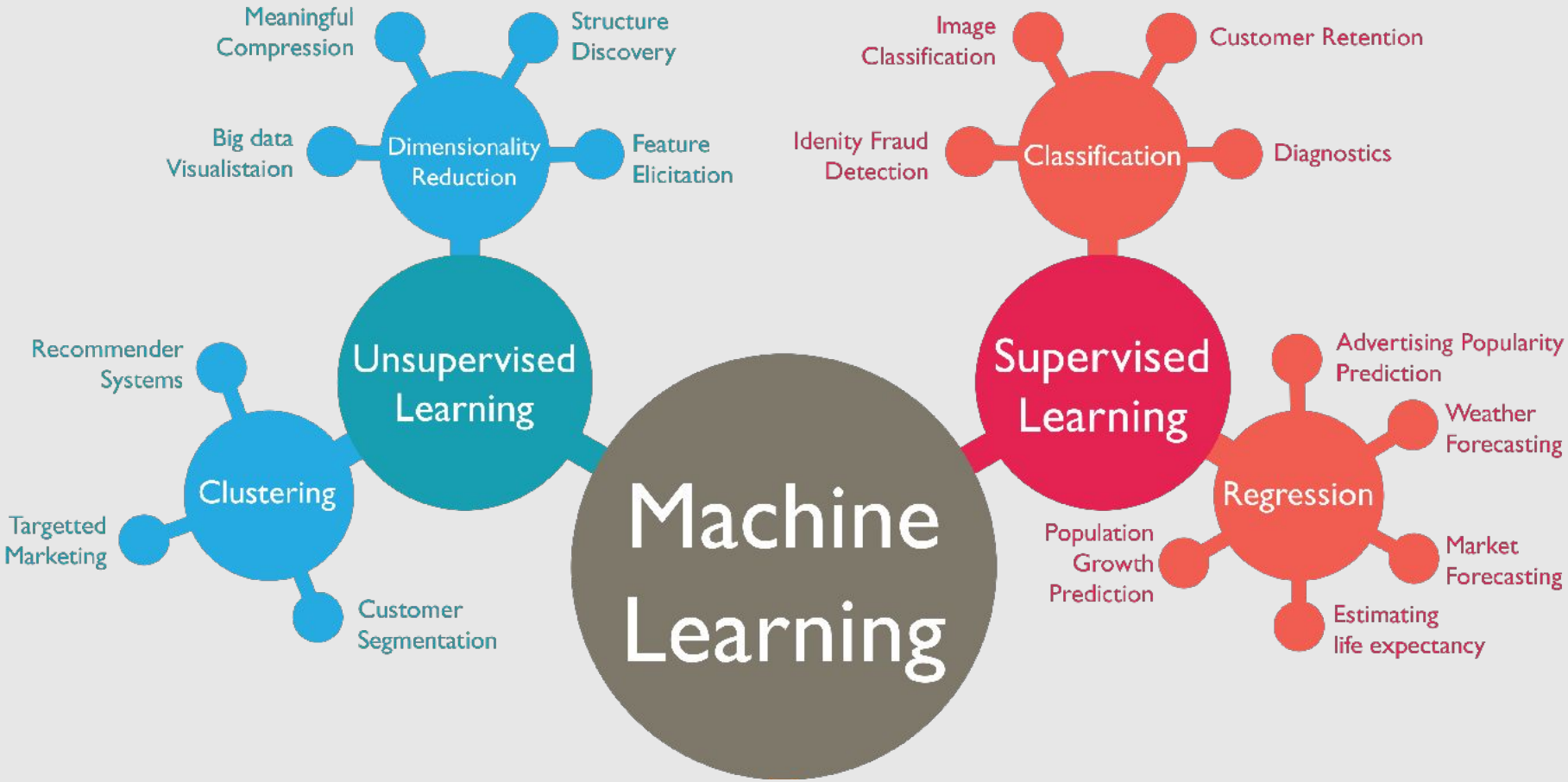
Did anyone say pizza?

k-Means Clustering



Important Unsupervised Learning Algorithms

- k-Means
- Hierarchical Cluster Analysis (HCA)
- Expectation Maximization
- Principal Component Analysis (PCA)
- Kernel PCA
- t-distributed Stochastic Neighbor Embedding (t-SNE)



http://www.isaziconsulting.co.za/machinelearning.html

Main Challenges of Machine Learning

I SEE BAD DATA

Main Challenges of Machine Learning

- Insufficient quantity of training data
- Non representative training data
- Poor quality data
- Irrelevant features

} **“Bad data”**

- Overfitting the training data
- Underfitting the training data

} **“Bad algorithm”**

Non Representative Training Data

In order to generalize well, it is crucial that your **training data be representative of the new cases** you want to generalize to.

Poor Quality Data

Obviously, **if your training data is full of errors, outliers and noise**, it will make it harder for the system to detect the underlying patterns, so your system is less likely to perform well.

Irrelevant Features

A critical part of the success of a Machine Learning project is coming up with a good set of features to train on. This is called **feature engineering**.

- **Feature Selection:** the process of selecting the most useful features to train on among existing features.
- **Feature Extraction:** combining existing features to produce a more useful one.

Main Challenges of Machine Learning

- Insufficient quantity of training data
- Non representative training data
- Poor quality data
- Irrelevant features

} **“Bad data”**

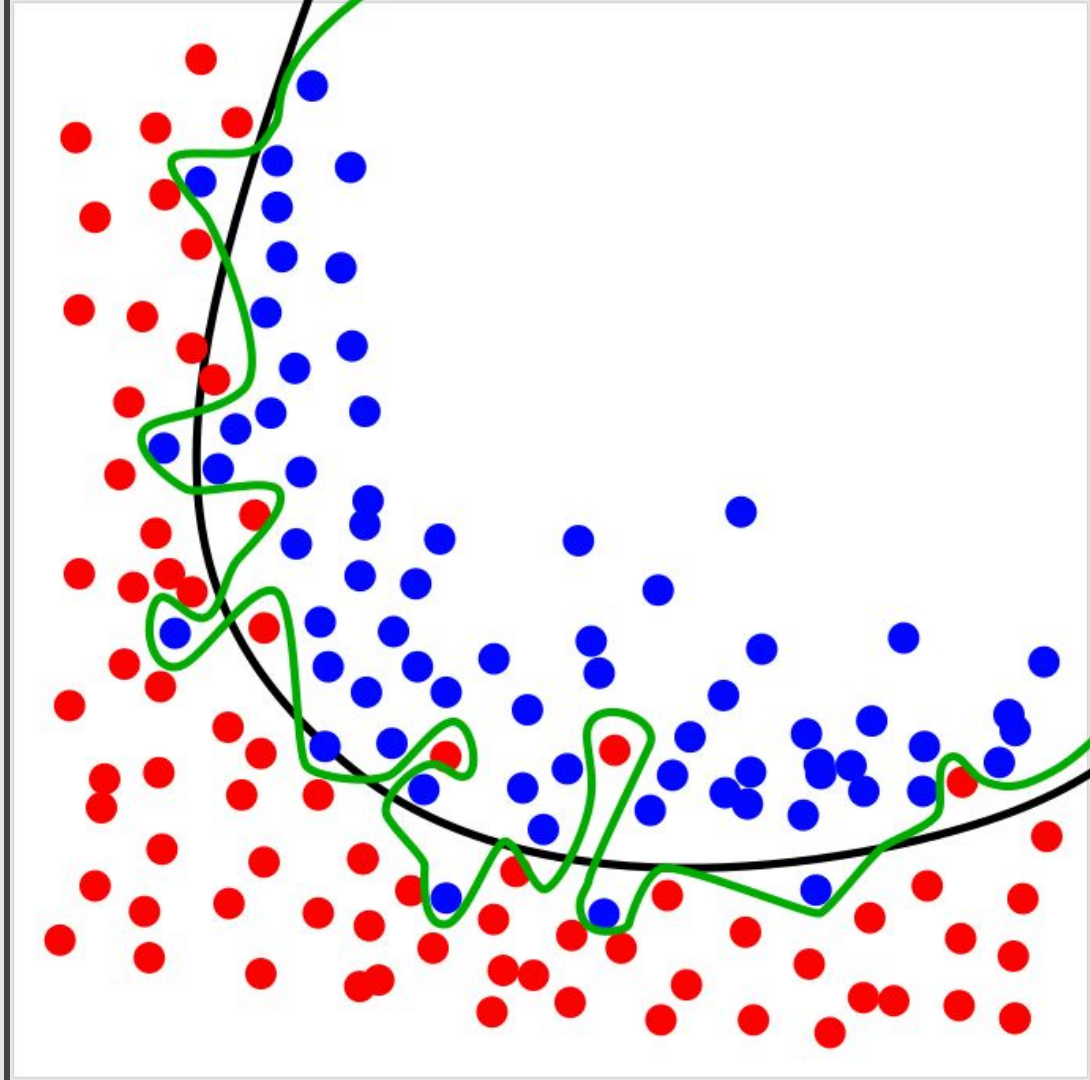
- Overfitting the training data
- Underfitting the training data

} **“Bad algorithm”**

Overfitting the Training Data

Overfitting means that the model performs well on the training data but **it does not generalize**.

Overfitting the Training Data



Overfitting the Training Data

Overfitting happens when the **model is too complex** relative to the amount and noisiness of the training data.

Underfitting the Training Data

Underfitting is the opposite of overfitting: it occurs when your **model is too simple** to learn the underlying structure of the data.

Main Challenges of Machine Learning

- Insufficient quantity of training data
- Non representative training data
- Poor quality data
- Irrelevant features

} **“Bad data”**

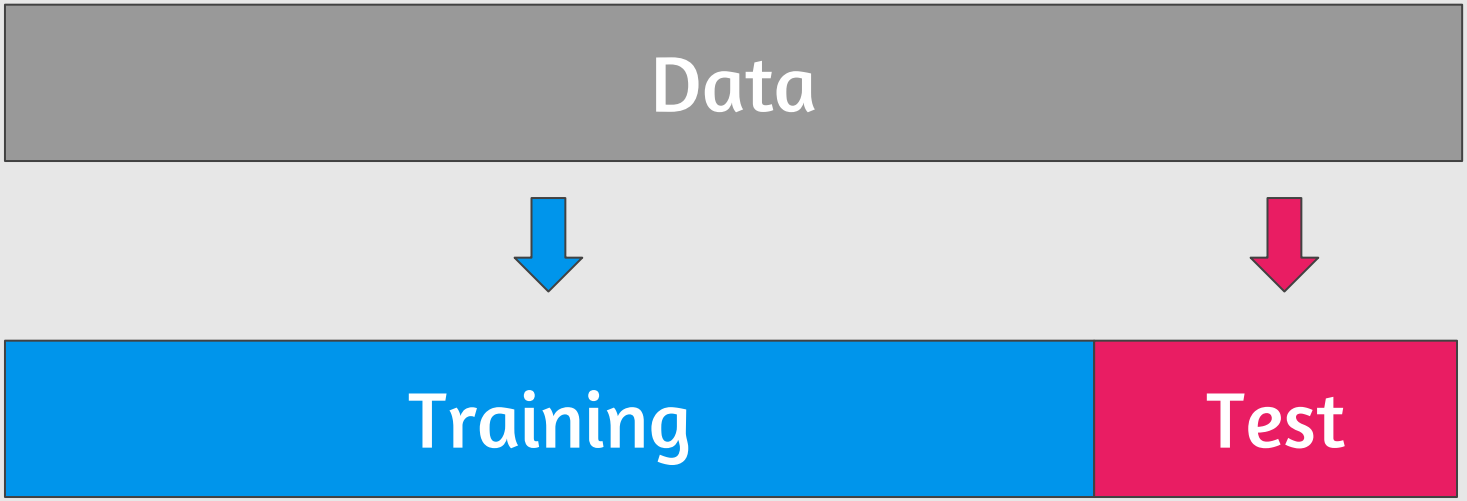
- Overfitting the training data
- Underfitting the training data

} **“Bad algorithm”**

Validating and Testing

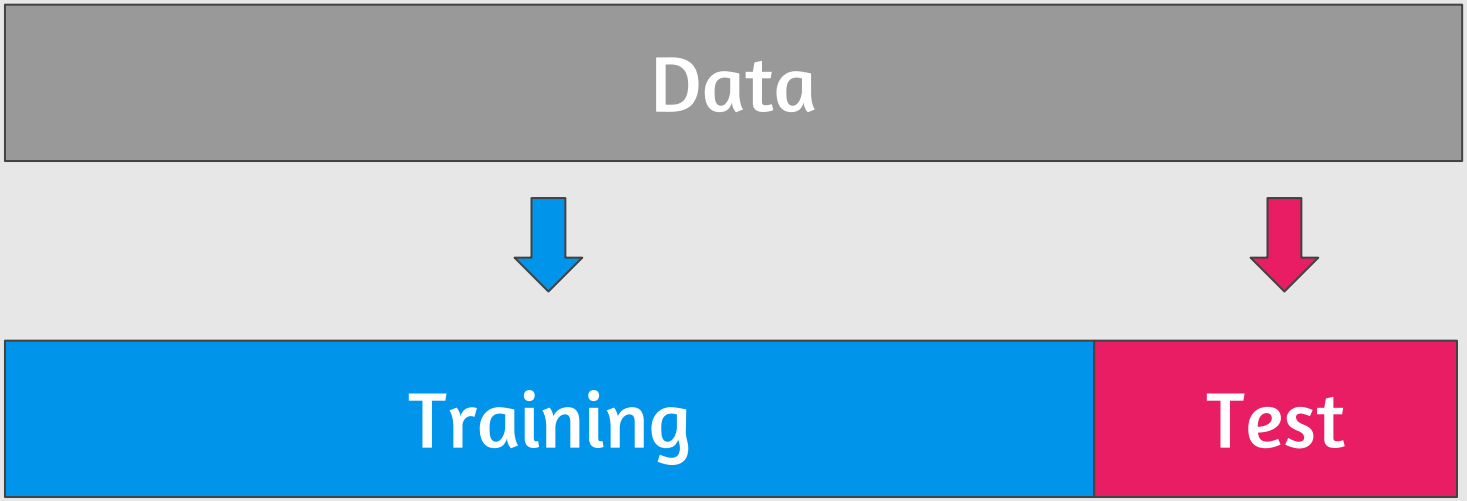
The only way to know how well a model will generalize to new cases is to actually try it out on new cases.

Data



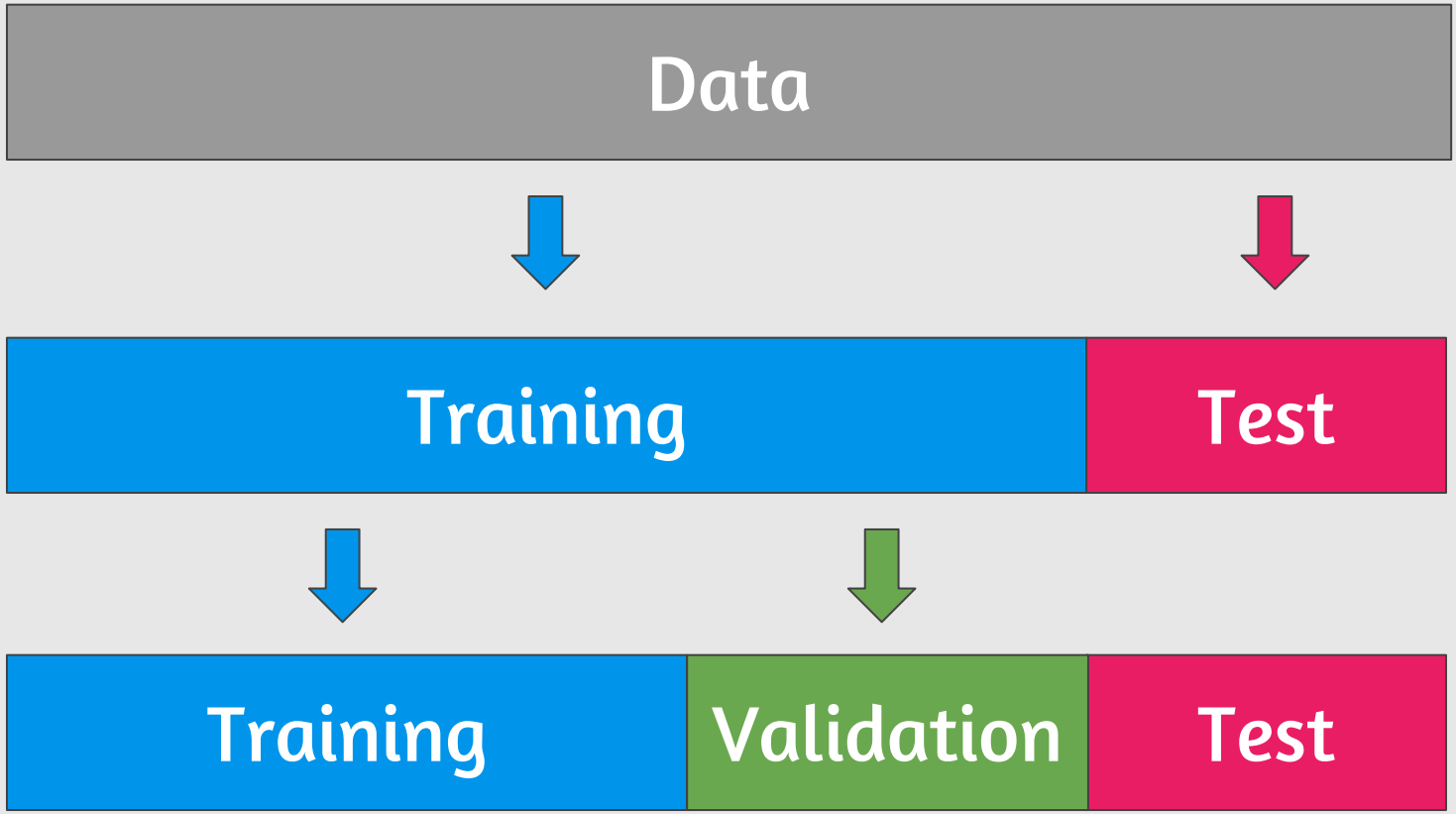
So evaluating a model is simple enough: just use a test set.

It is common to use 80% of the data for training and **hold out** 20% for testing.

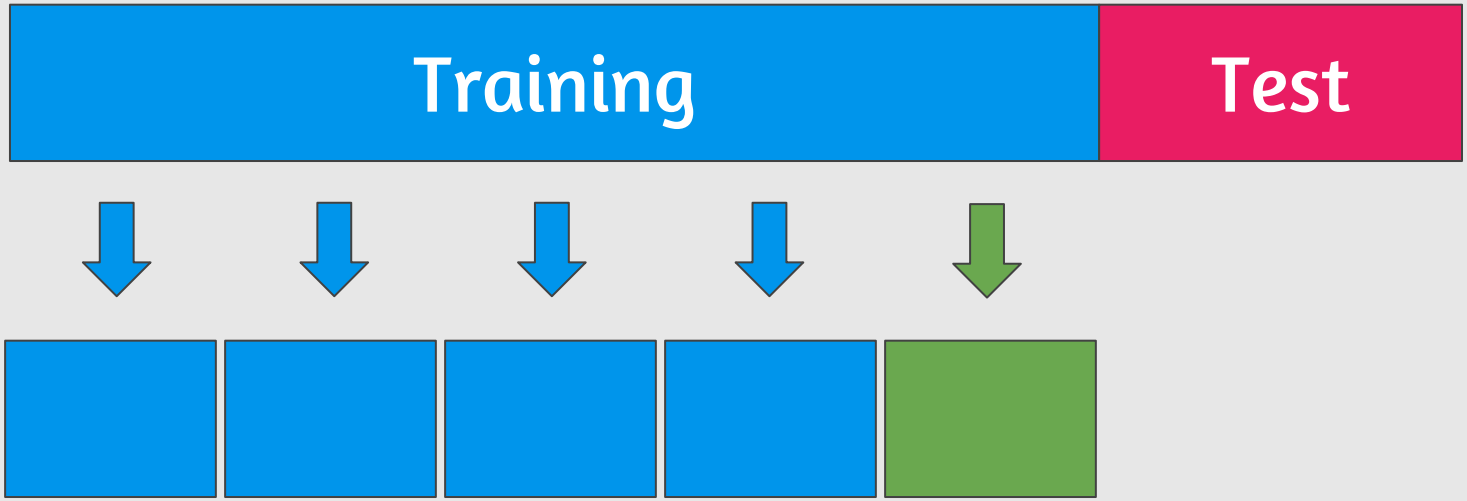


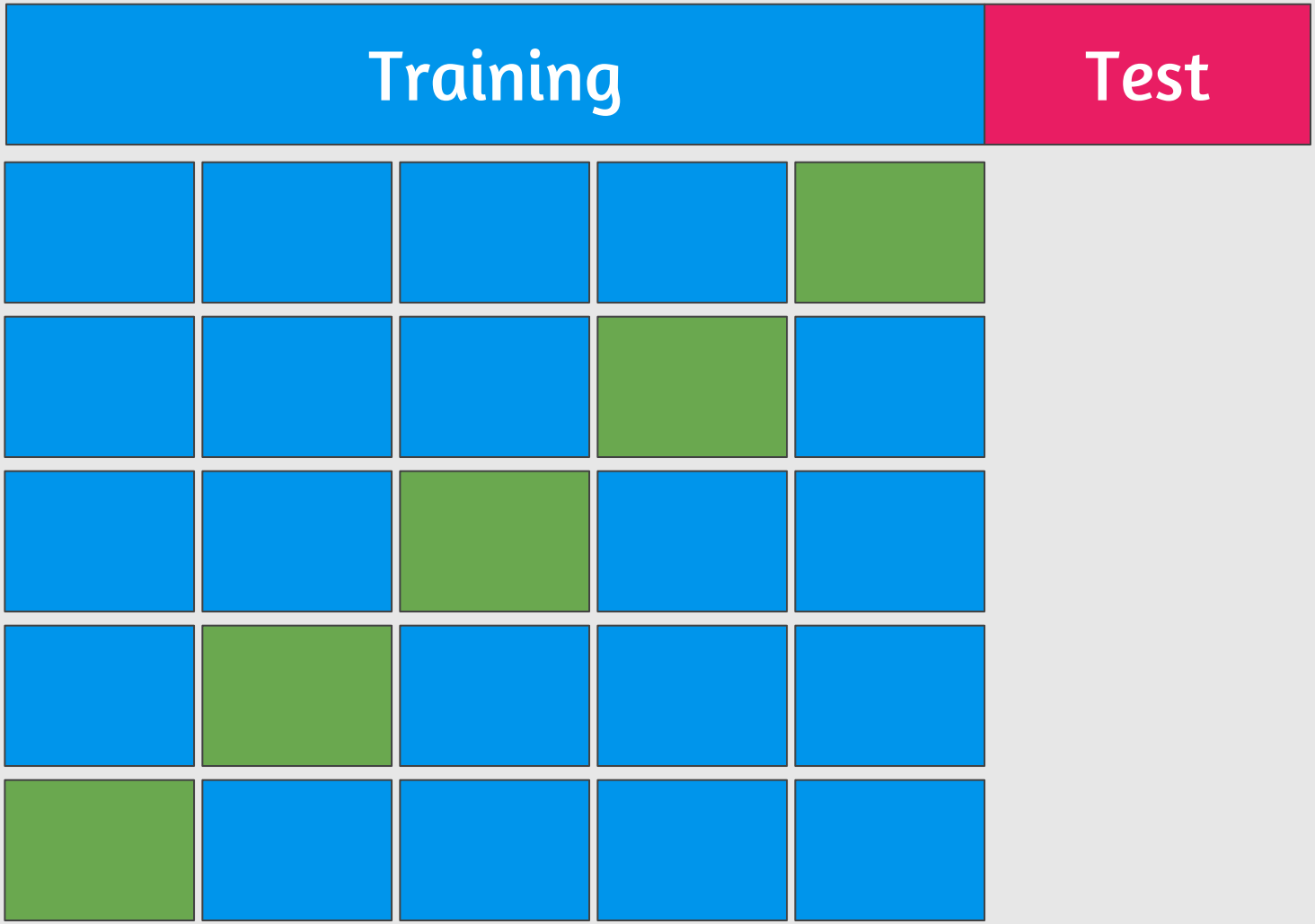
So evaluating a model is simple enough: just use a test set.

Now suppose you are hesitating between two models.
How can you decide?

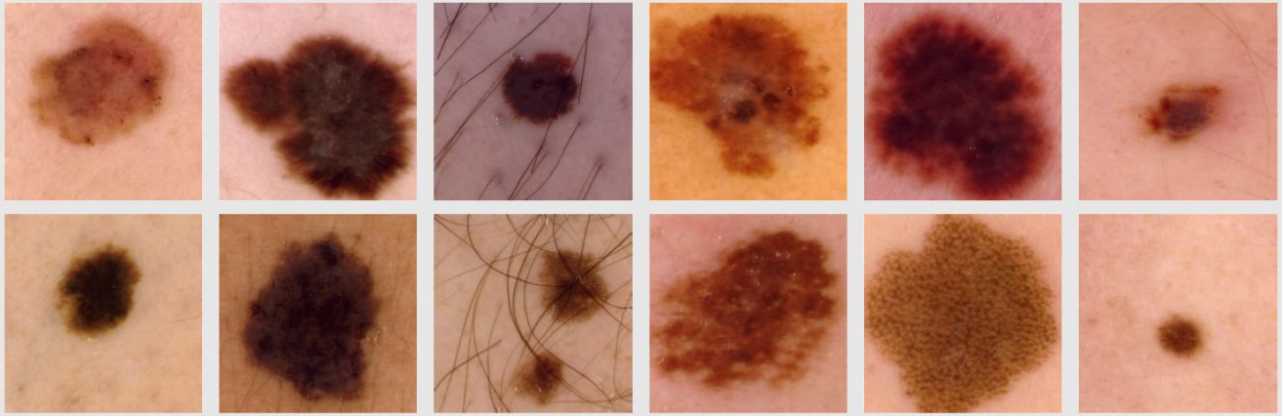

Data
Training
Test
Training
Validation
Test

Cross Validation




Training
Test

Skin Cancer Classification
ISIC Challenge 2017

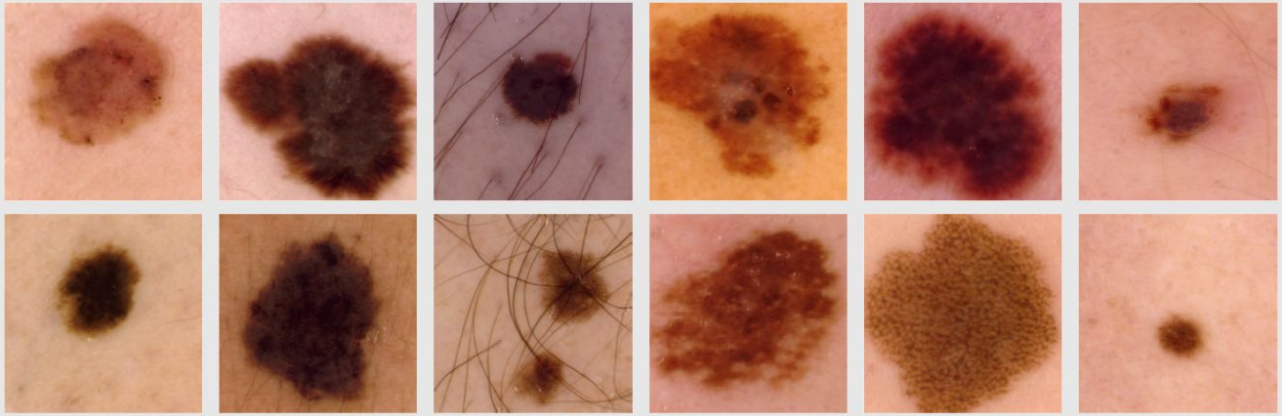


"RECOD Titans at ISIC Challenge 2017". A. Menegola, J. Tavares, M. Fornaciali, L.T. Li, S. Avila, E. Valle, arXiv preprint arXiv:1703.04819, 2017.

Training data 2000 images & **Validation data** 150 images & **Test data** 600 images

Skin Cancer Classification

ISIC Challenge 2017



Important

"RECOD Titans at ISIC Challenge 2017".
A. Menegola, J. Tavares, M. Fornaciali, L.T. Li, S. Avila, E. Valle, arXiv preprint arXiv:1703.04819, 2017.

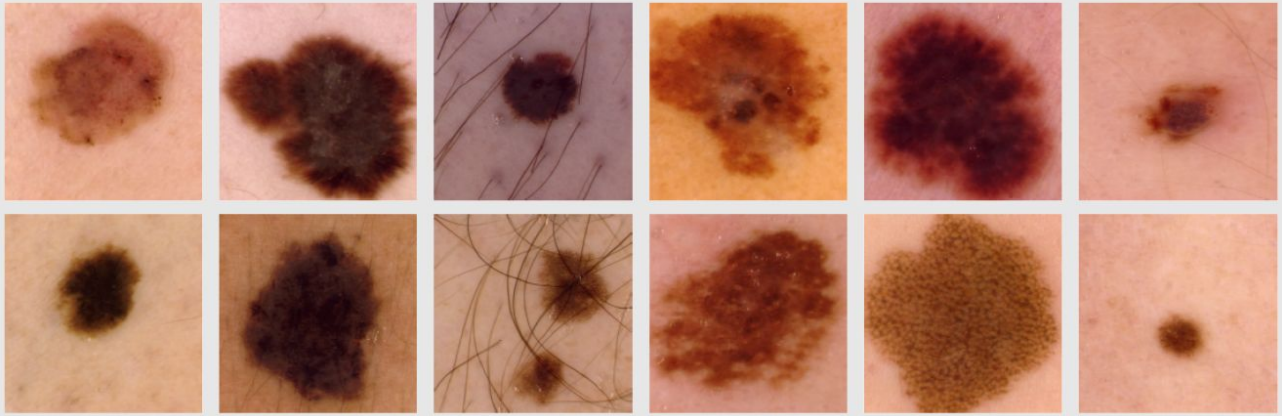
Training data 2000 images & **Validation data** 150 images & **Test data** 600 images

95.1%

(internal validation)

Skin Cancer Classification

ISIC Challenge 2017



"RECOD Titans at ISIC Challenge 2017".
A. Menegola, J. Tavares, M. Fornaciali, L.T. Li, S. Avila, E. Valle, arXiv preprint arXiv:1703.04819, 2017.

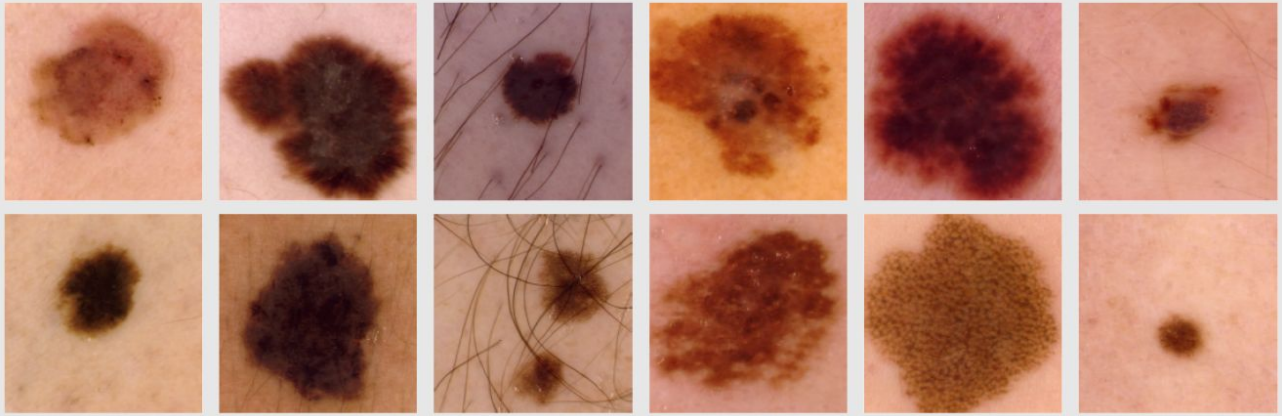
Training data & **Validation data** & **Test data**

2000 images & 150 images & 600 images

95.1% **90.8%**

(internal validation)

Skin Cancer Classification
ISIC Challenge 2017



"RECOD Titans at ISIC Challenge 2017". A. Menegola, J. Tavares, M. Fornaciali, L.T. Li, S. Avila, E. Valle, arXiv preprint arXiv:1703.04819, 2017.

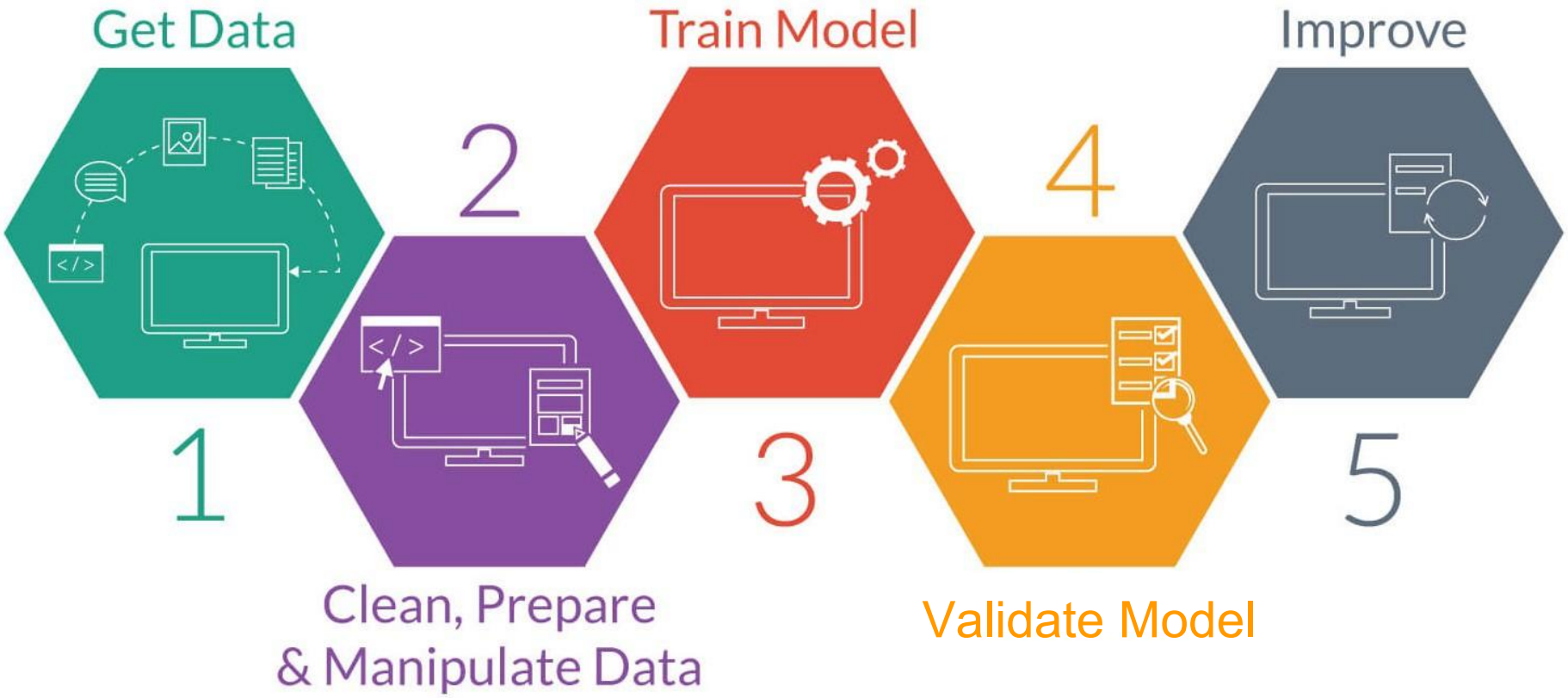
Training data & **Validation data** & **Test data**

2000 images & 150 images & 600 images

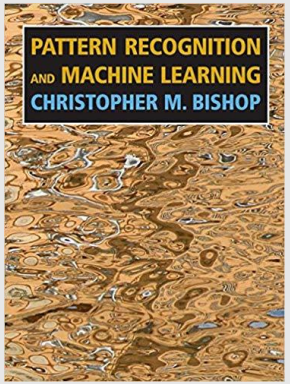
95.1% **90.8%** **87.4%**

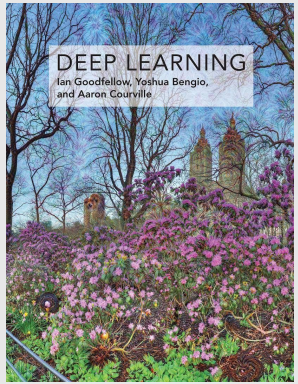
(*internal validation*)

Summary

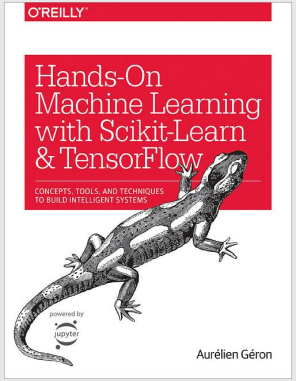

Get Data
Train Model
Improve
2
4
1
3
5
Clean, Prepare
& Manipulate Data
Validate Model

Course Logistics





- 4 credits (60 h/class)
- Material:
  - Books, blogs, online courses
  - Optional textbook:
    “Hands-On Machine Learning with Scikit-Learn and TensorFlow”, by Aurélien Géron, 2017



Grading Policy

- No written exam

Grading Policy

- What is your biggest doubt? (individual) **5%** (extra)
  - To send by Moodle
  - Until 3pm a day after the class

Grading Policy

- 4 practical assignments (2 people): Technical Report & Code

  10%: Linear Regression

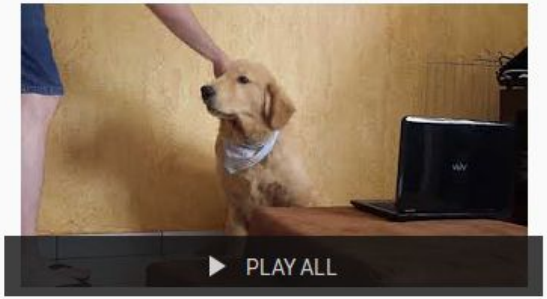
  20%: Logistic Regression & Neural Networks

  15%: Dimensionality Reduction & Clustering

  15%: Deep Learning

Grading Policy

- Final Project (4 or 5 people) **40%**
  - **5%** Proposal & Dataset
  - **5%** Baseline
  - **10%** Presentation (videos 4-minutes long)
  - **20%** Technical Report & Code



MC886/MO444 Machine Learning and Pattern Recognition IC/Unicamp 2017s2

23 videos • 1,555 views • Last updated on 17 Jan 2018

Sandra Avila

Final Projects: MC886 (undergraduate) / MO444 (graduate)
Institute of Computing (IC), University of Campinas (Unicamp), 2017
Prof. Sandra Avila (https://www.ic.unicamp.br/~sandra/), TA: Samuel G. Fadel

Presentations in English or in Portuguese.

1 

Machine learning age-gender recognition
terra0009

2 

Aprendizagem em dados Geofísicos
Lucas Carrilho Pessoa

3 

Understanding the Amazon From Space - MO444 - Group 7
Bárbara Benato

4 

[MO444] Evoluindo Redes Neurais para jogar Mega Man X
Alexandre Almeida

5 

Brazilian Coins - Projeto MC886
HeyHiHelloHard

6 

MC886 - Generating TV Script for Simpsons episode
Vitor Arrais

7 

16PF e Machine Learning
Giovani Nascimento Pereira

Grading Policy

- **Academic infraction ⇒ Zero**
  - Allowing another to copy from one's work.
  - Submitting the work of another as one's own.
  - Providing false or misleading information for the purpose of gaining an academic advantage.
  - etc.

Frequency

- The frequency must be greater than or equal to **75% for approval**.

Prerequisites

- Some Python programming experience
  - http://learnpython.org
- Calculus, Linear algebra, Probabilities and Statistics
  - Part I: Applied Math and Machine Learning Basics https://www.deeplearningbook.org

Syllabus

- **Submission**

  Moodle: www.ggte.unicamp.br/moodle

- **Information**

  www.ic.unicamp.br/~sandra/teaching/2018-2-mc886-mo444

- **Discussion**

  Slack workspace Machine Learning: ml-unicamp.slack.com

That’s all!


YOU KNOW, IF YOU GO IN THERE, THERE'S NO COMING BACK.
I... I HAVE TO DO IT. I HAVE TO KNOW WHAT HAPPENED TO MY FRIENDS.
Machine Learning
THE FOOL...
SHEN COMIX