



Machine Learning and Pattern Recognition

A High Level Overview

Prof. Sandra Avila

Institute of Computing (IC/Unicamp)

The Hype

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

The data economy demands a new approach to antitrust rules



CADE METZ BUSINESS 03.08.16 07:00 AM

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





all sections

[news](#) / [opinion](#) / [sport](#) / [arts](#) / [life](#)[tech](#) / [world](#) / [UK](#) / [science](#) / [cities](#) / [global development](#) / [business](#) / [environment](#) / [obituaries](#)**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

JACK STEWART TRANSPORTATION 03.29.17 3:48 PM

SHARE



SHARE

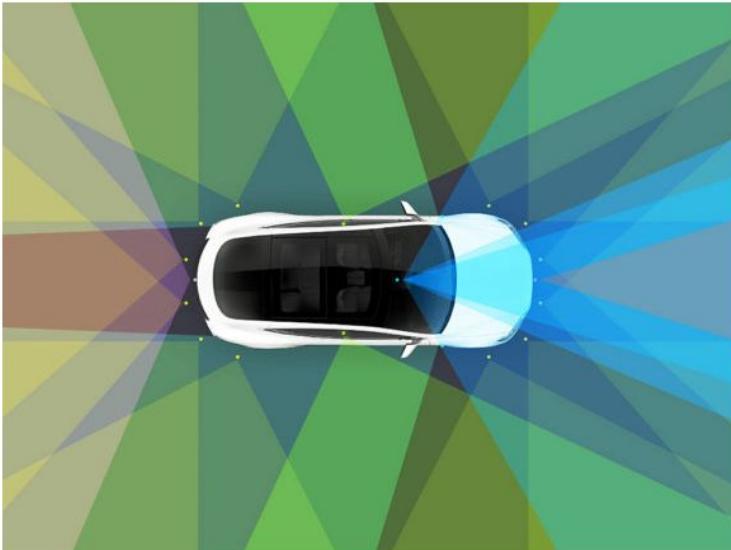


TWEET



EMAIL

TESLA FINALLY MAKES ITS NEW AUTOPILOT AS GOOD AS THE OLD ONE





INDY/TECH

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



Got a tip? [Let us know.](#)

Follow Us [f](#) [i](#) [t](#) [y](#) [F](#) [in](#) [g+](#) [r](#)

News ▾ Video ▾ Events ▾ Crunchbase

[Message Us](#)

[Search](#)



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 Coldevey](#)



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

<https://techcrunch.com/2017/08/03/high-schooler-makes-3d-printed-machine-learning-powered-eye-disease-diagnosis-system/>

Viome raises \$15M to manage the microbiome with machine learning

By **Bernie Monegain** | August 07, 2017



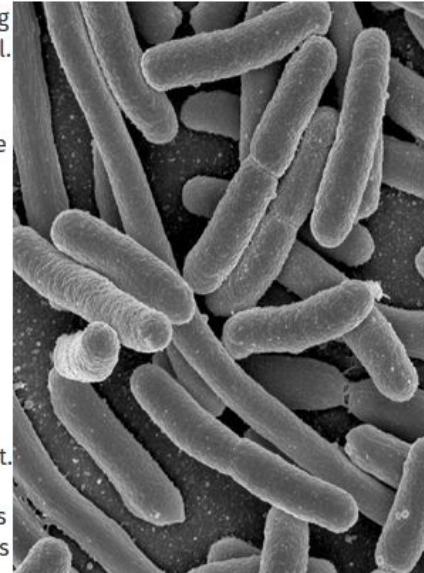
Viome, which has developed technology aimed at balancing microorganisms in the gut, has landed \$15 million in capital.

The funds will go towards the launch of the company's at-home health kit, which targets the microbiome. Microbiome refers to microorganisms, such as bacteria, fungi, and viruses, in the human body.

Viome is led by CEO Naveen Jain, innovator, philanthropist and founder of Moon Express, Intelius, TalentWise and InfoSpace. It was founded by Jain and a group of entrepreneurs.

The company is currently operating as an early beta program with several thousand customers using the product.

Viome uses proprietary technology licensed through the Los Alamos National Laboratory. Viome identifies and quantifies all microorganisms in the gut – and analyzes what they are



Why now?



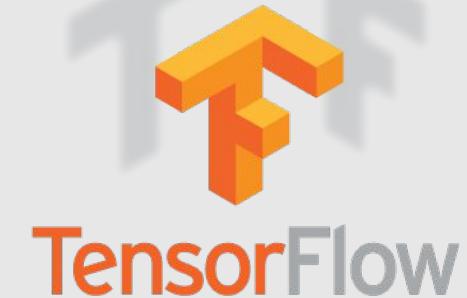
www.image-net.org

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



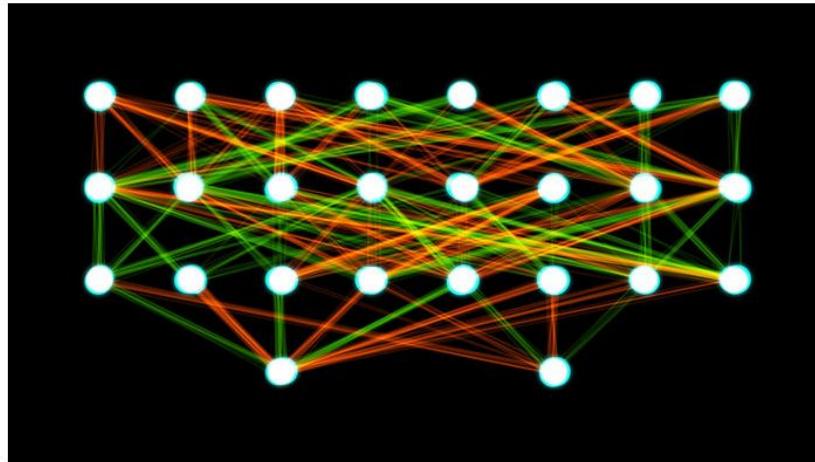
Machine Learning Frameworks





“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.”

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 Definition

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

[Aurélien Géron, 2017]

Machine Learning Definition

“Field of study that gives computers the ability to **learn** without being explicitly programmed.”

[Arthur Samuel, 1959]

Machine Learning Definition

“A computer program is said to **learn from experience** E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E . ”

[Tom Mitchell, 1997]

ARTIFICIAL INTELLIGENCE



1950

1960

1970

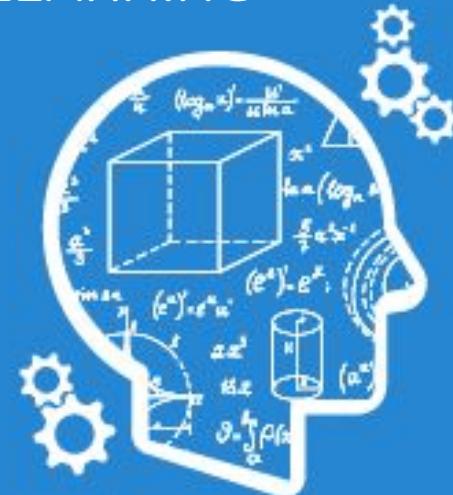
1980

1990

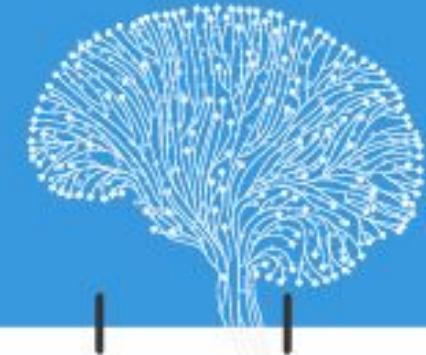
2000

2010

MACHINE LEARNING



DEEP LEARNING



Why is this
so important?

MACHINE LEARNING



MACHINE LEARNING EVERYWHERE

mator.net

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

- **Supervised Learning**
 - Specific target signal to predict
 - Training data have known target values
- **Unsupervised Learning**
 - No given target value; looking for structure

Supervised Learning

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

Regression is used to predict continuous values.

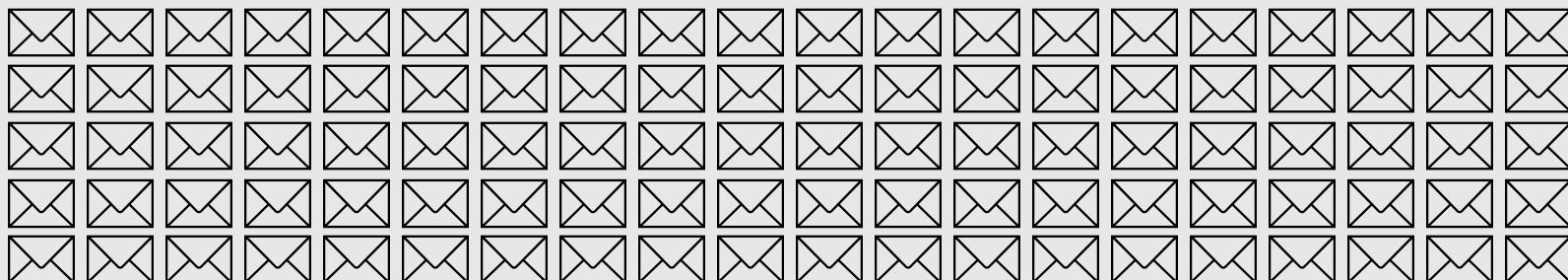
Spam Filtering



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

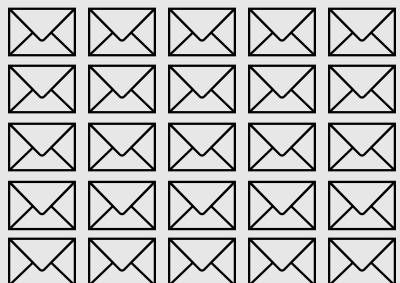


100 emails

Spam Filtering

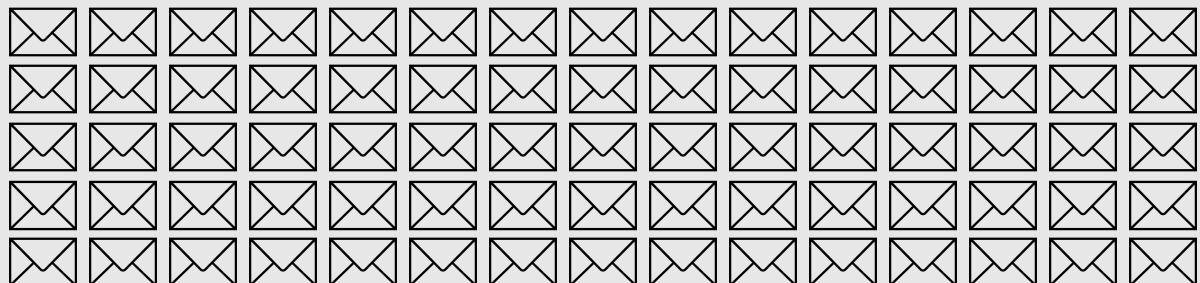


Spam



25 emails

Non-spam

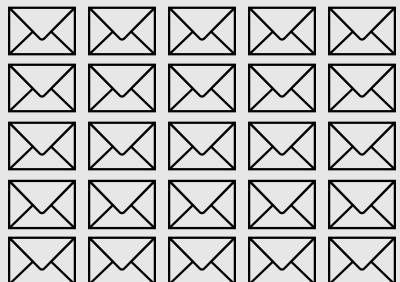


75 emails

Spam Filtering

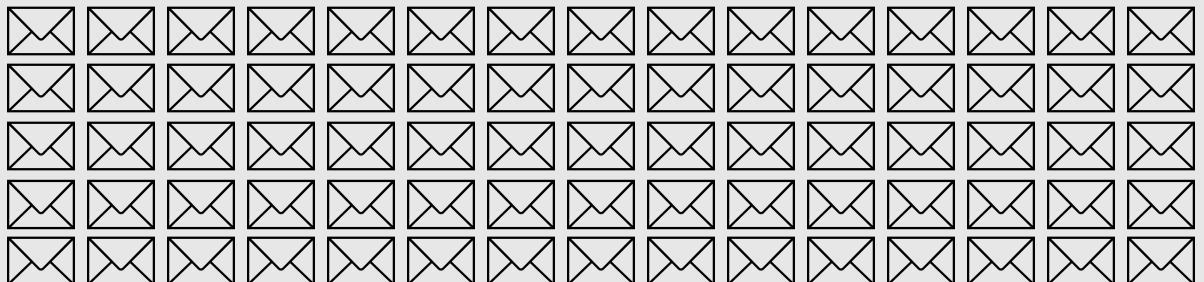
! “Cheap”

Spam



25 emails

Non-spam

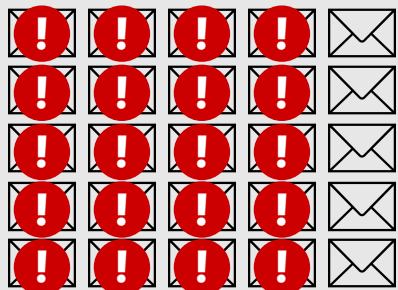


75 emails

Spam Filtering

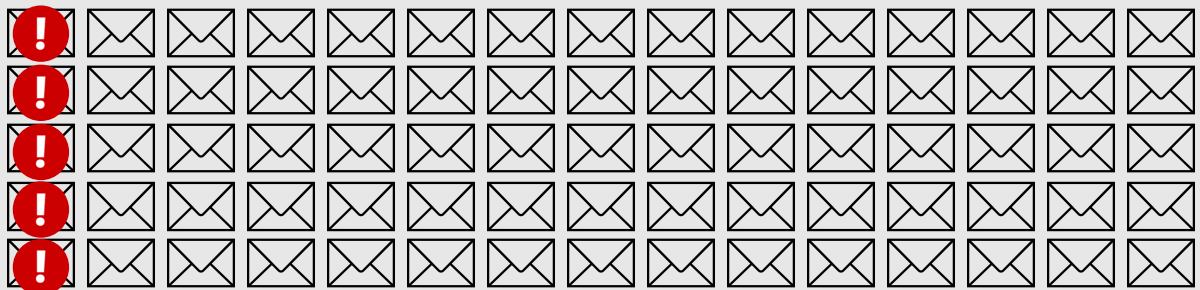
! “Cheap”

Spam



25 emails

Non-spam

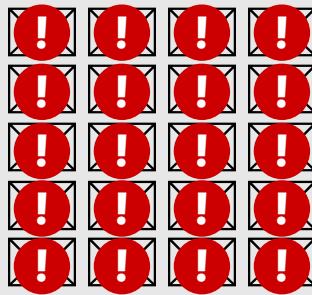


75 emails

Spam Filtering

! “Cheap”

Spam



Non-spam



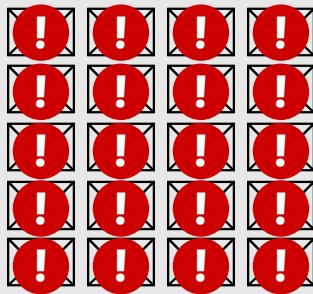
If an email contains the word “cheap”, what is the probability of it being spam?

- 40%
- 60%
- 80%

Spam Filtering

! “Cheap”

Spam



20

Non-spam



5

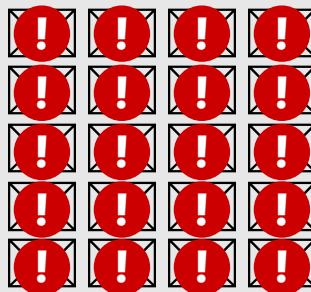
If an email contains the word “cheap”, what is the probability of it being spam?

- 40%
- 60%
- 80%

Spam Filtering

! “Cheap”

Spam



80%

Non-spam



20%

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

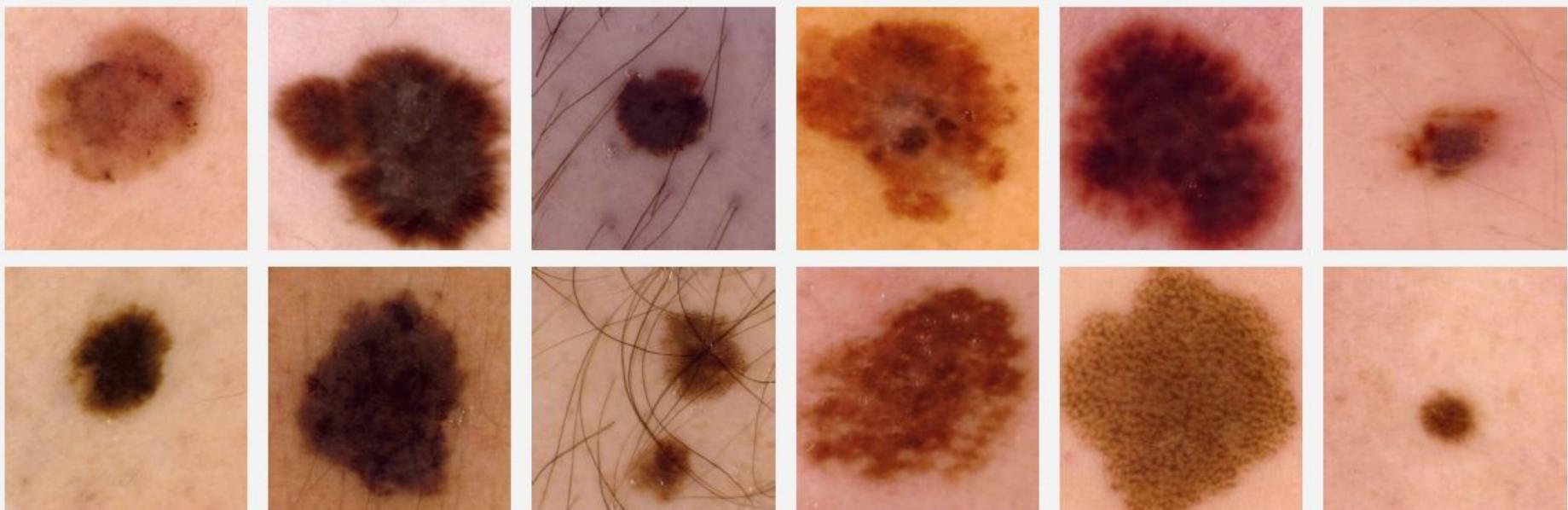
- ! “Cheap” → 80%
- ! Spelling mistake → 70%
- ! Missing title → 95%
- ! etc ...

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



I Autor Divulgação laboratório RECOD

I Fotos Mijail Vidal

I 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,



Sistema garante proteção de crianças | IMAGE SOURCE/POLHA PRESS

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.

"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.



HIDAIANA
ROSA

METRO CAMPINAS

House Price Prediction (Regression)



\$ 70 000

House Price Prediction

(Regression)



\$ 160 000

House Price Prediction (Regression)

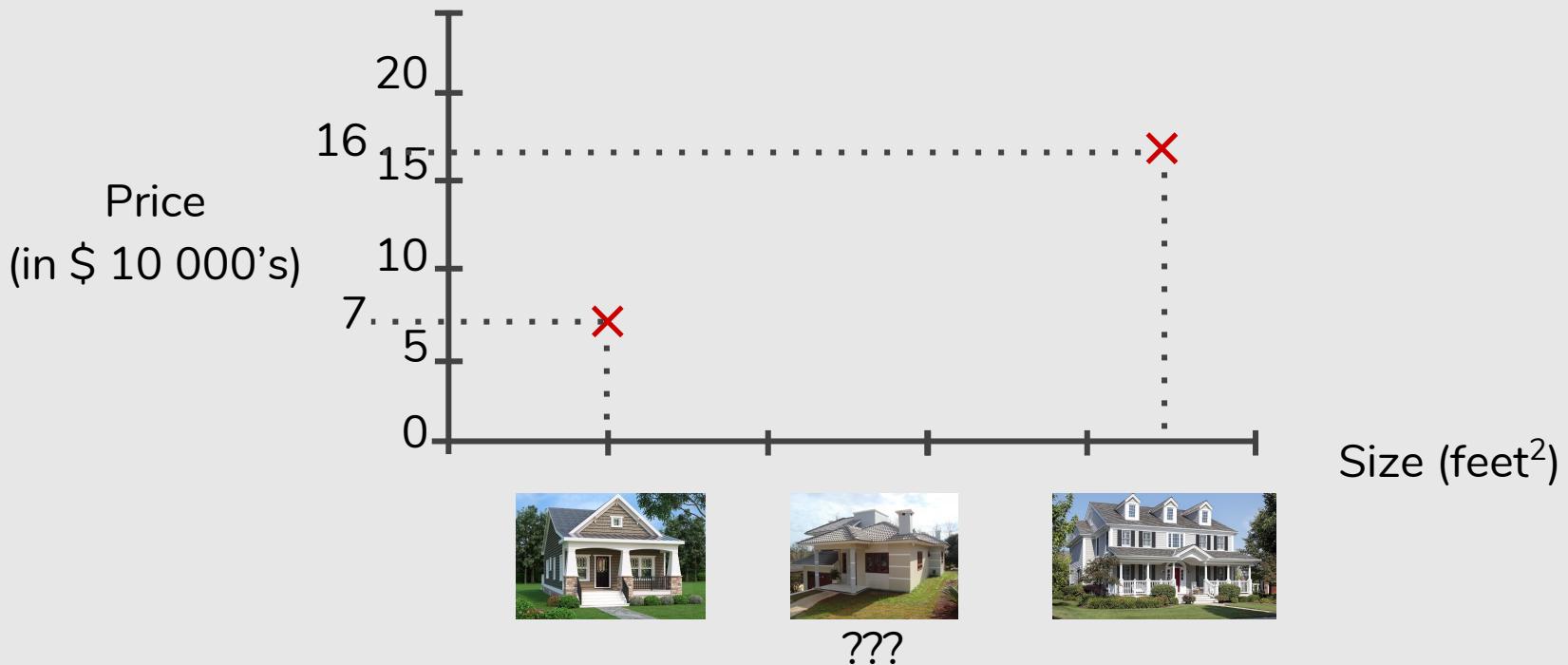


???

House Price Prediction (Regression)

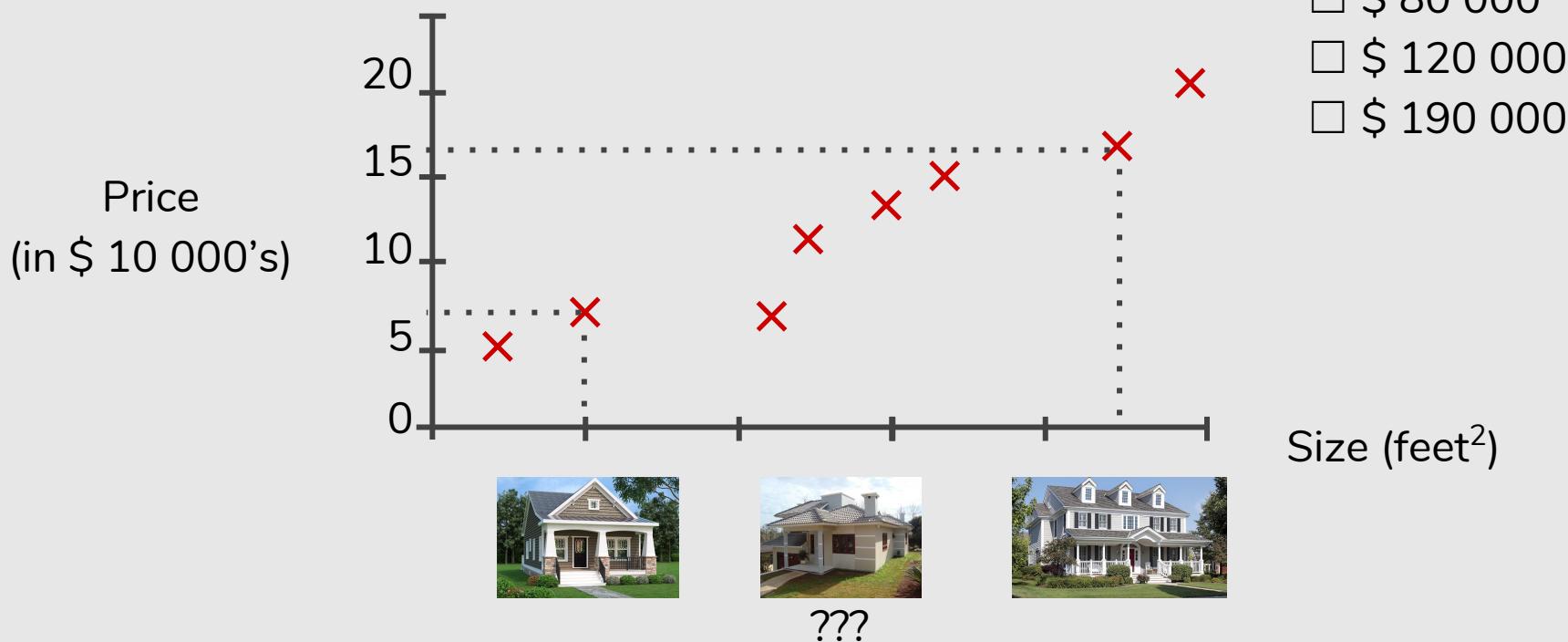


House Price Prediction (Regression)



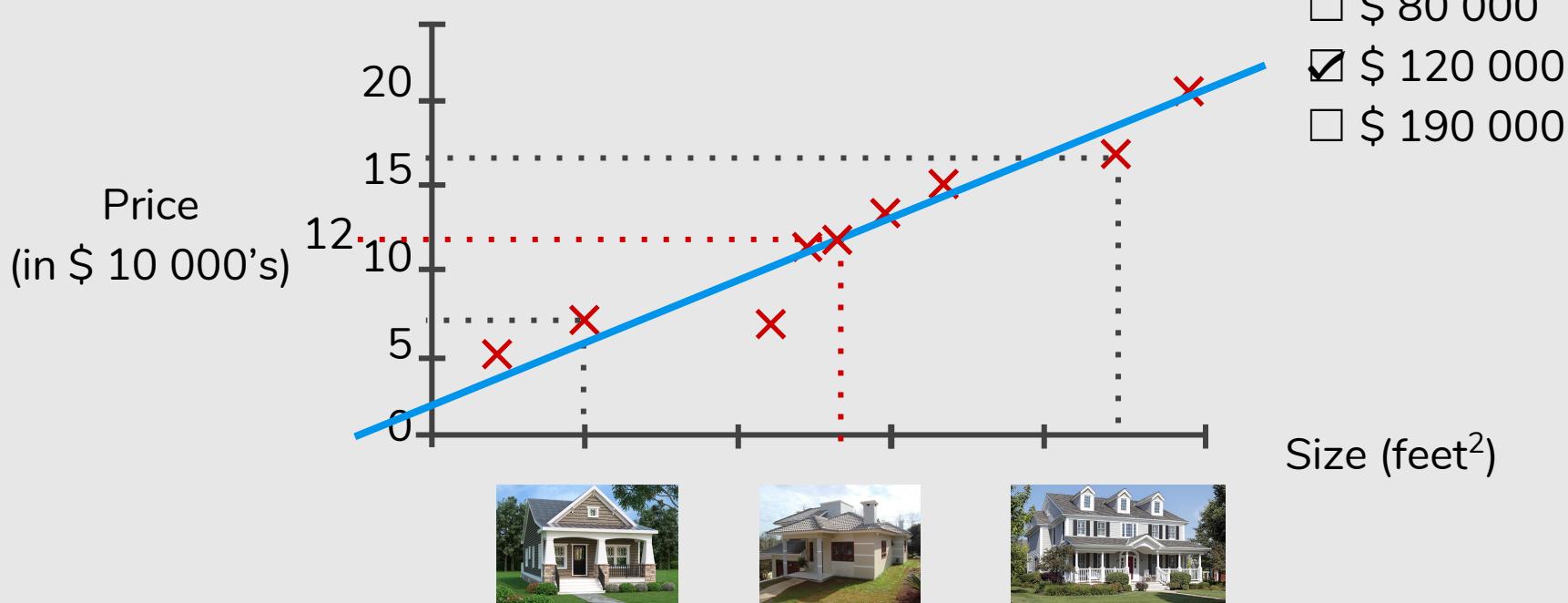
House Price Prediction (Regression)

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



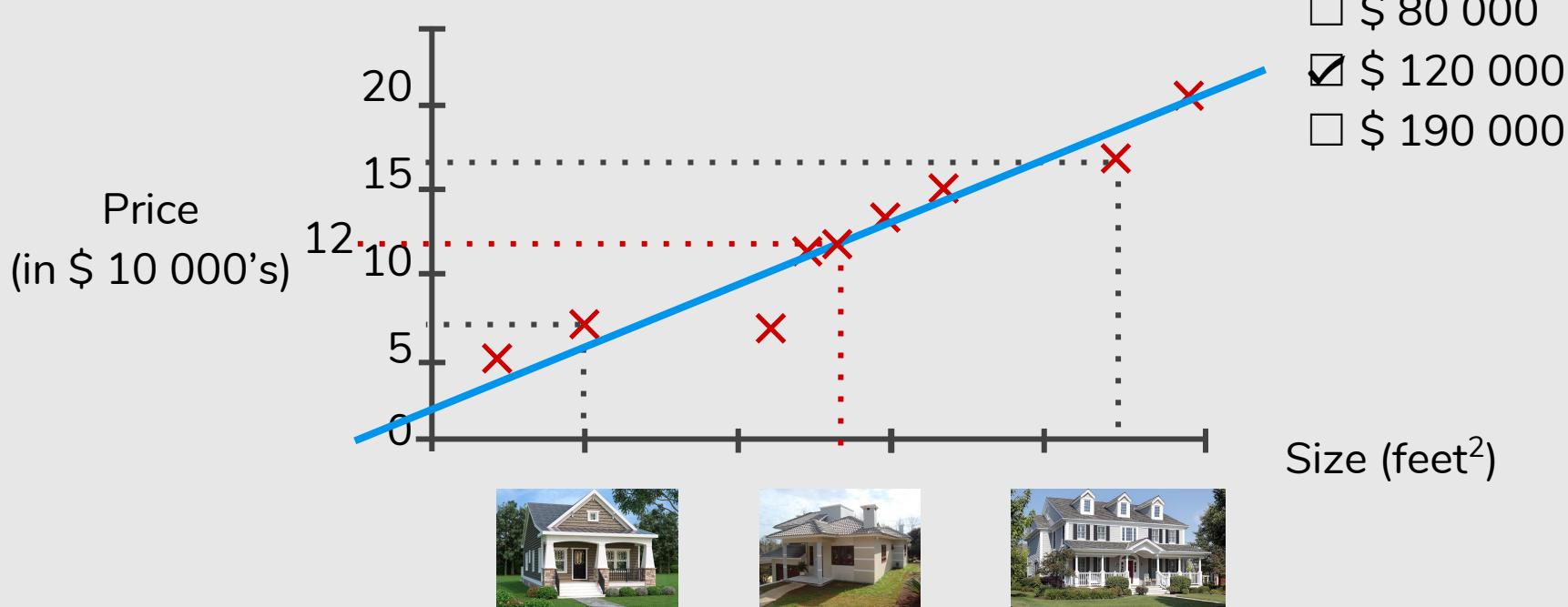
House Price Prediction (Regression)

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



Linear Regression

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



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 loosing too much information.

— — —

Did anyone say pizza?



<https://www.youtube.com/watch?v=lpGxLWOIzy4>

Map data ©2017 Google Brasil Terms Send feedback 100 m

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?



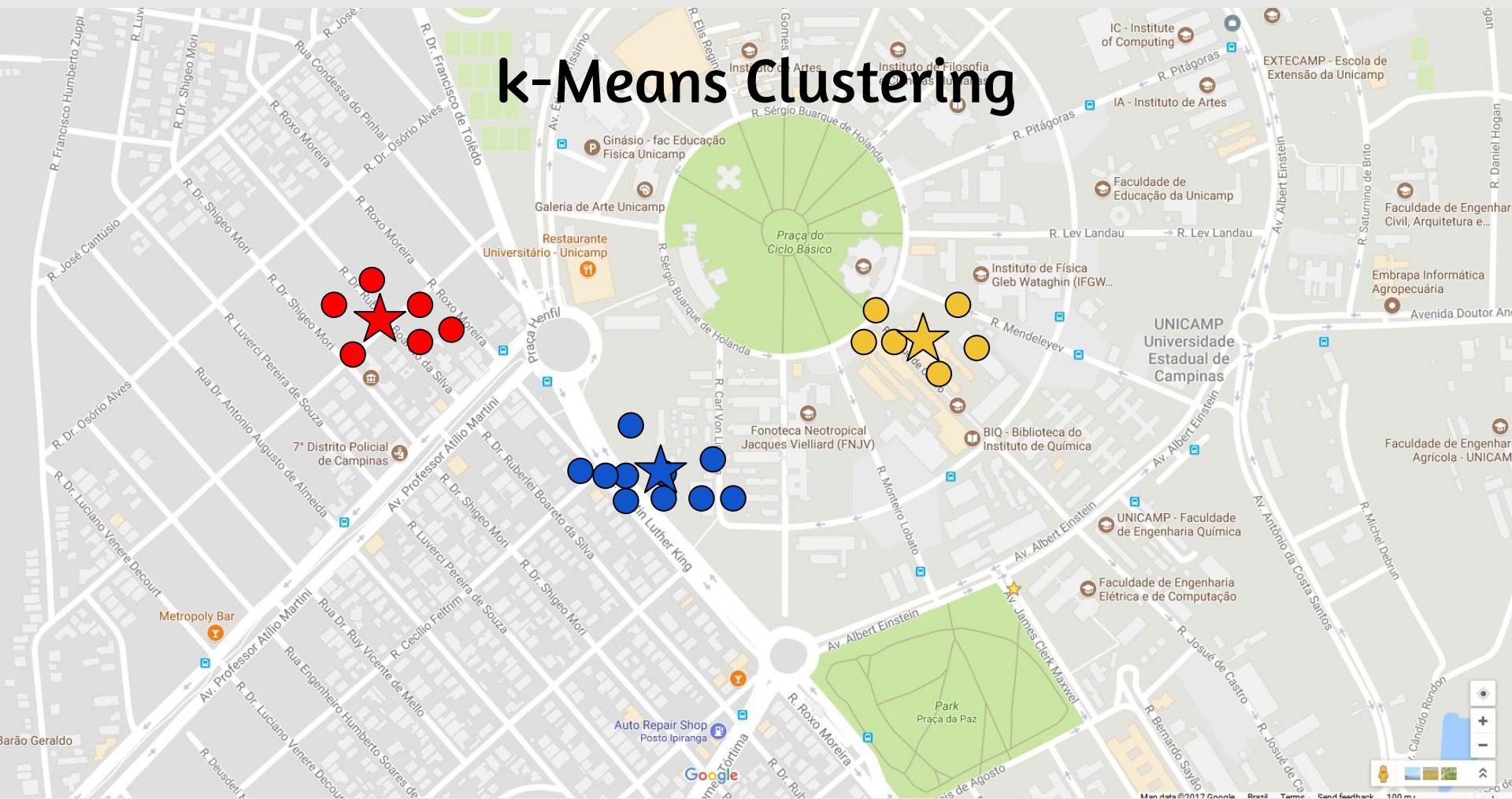
Did anyone say pizza?



Did anyone say pizza?

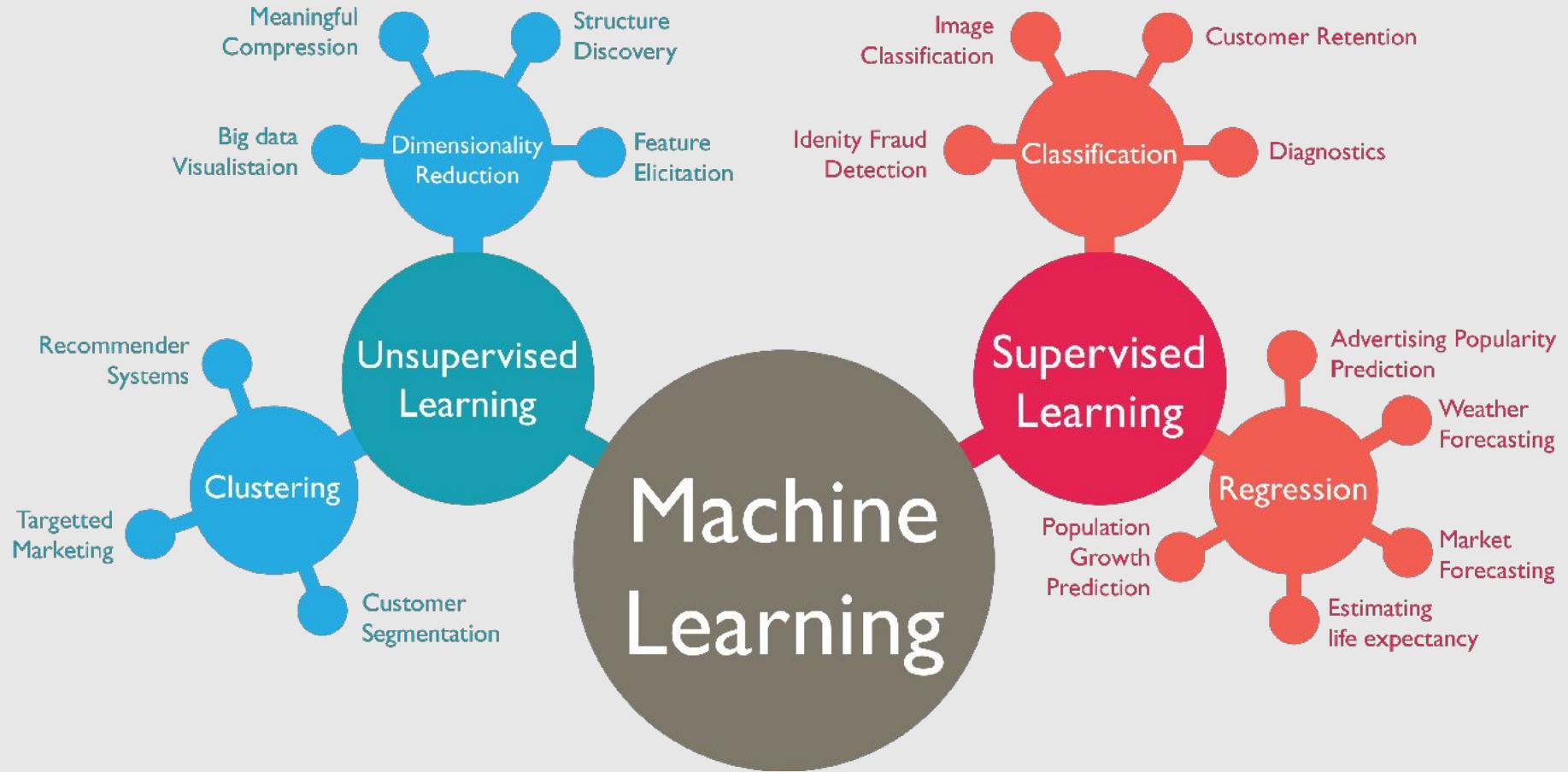


k-Means Clustering



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)



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
-
- Overfitting the training data
 - Underfitting the training data



“Bad data”

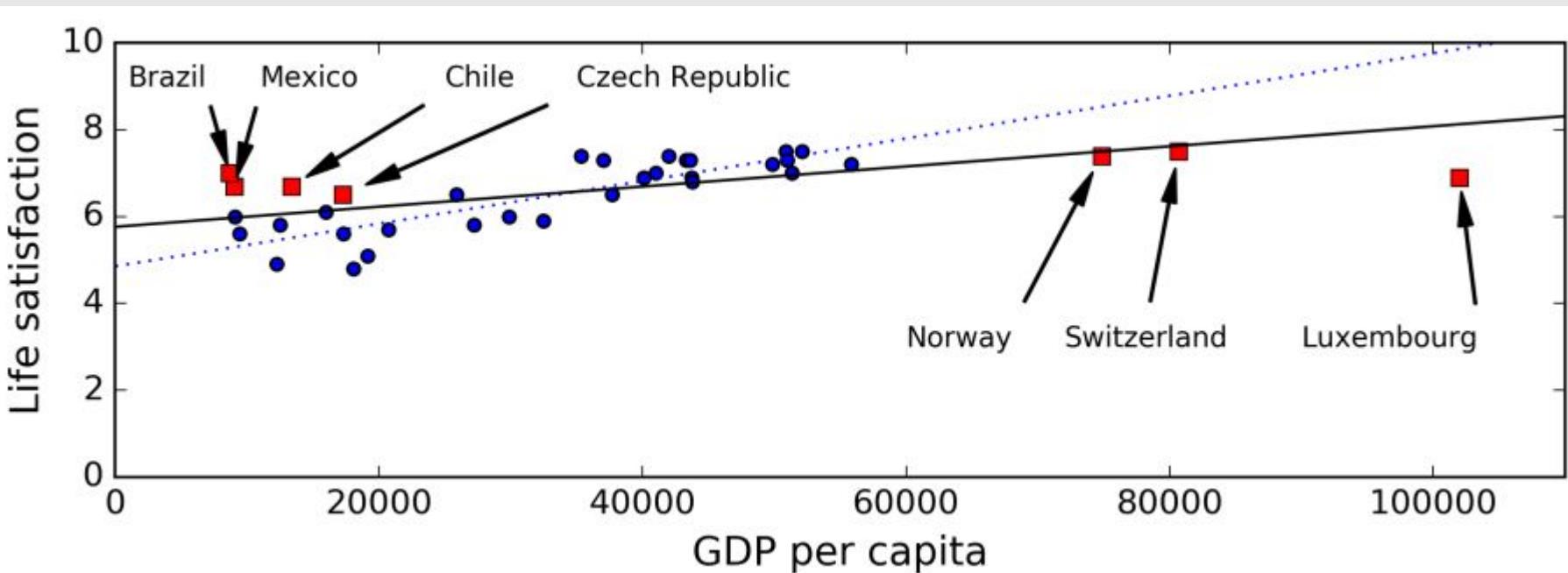


“Bad algorithm”

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.

Non Representative Training Data



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**. This involves:

- 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
-
- Overfitting the training data
 - Underfitting the training data



“Bad data”



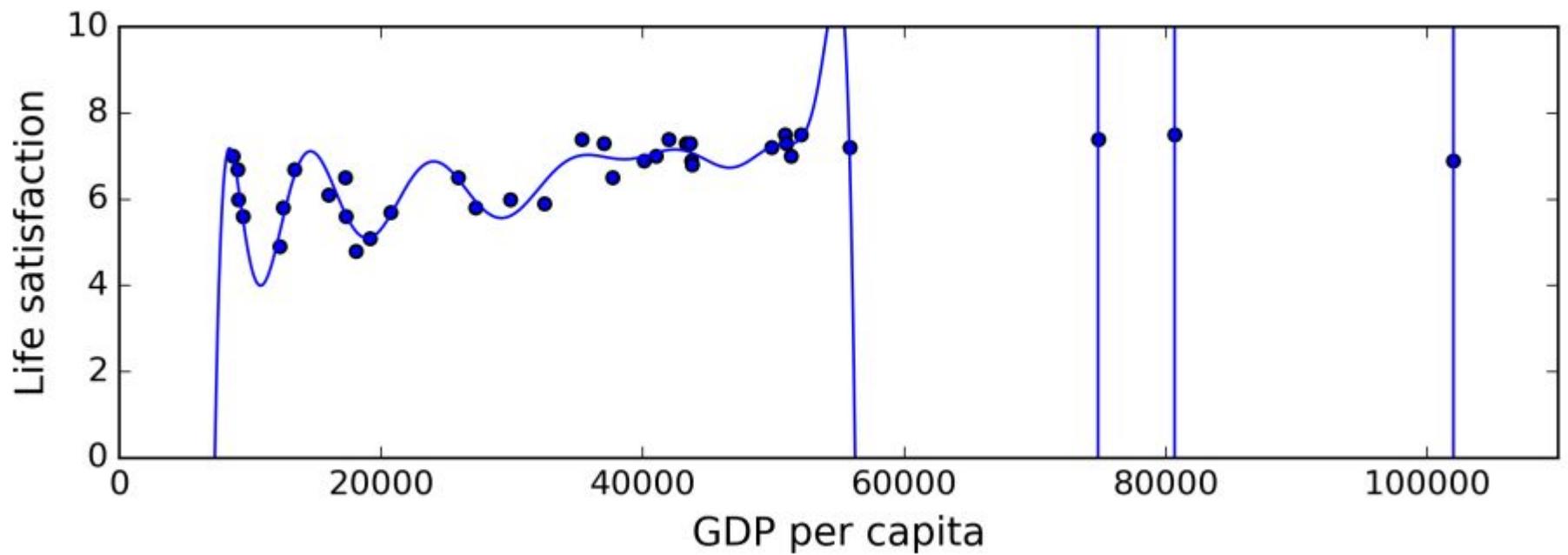
“Bad algorithm”

Overfitting the Training Data

Over-generalizing is something that we humans do all too often, and unfortunately machines can fall into the same trap if we are not careful.

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. The possible solutions are:

- to simplify the model by selecting one with less parameters, by reducing the number of attributes in the training data or by constraining the model,
- to gather more training data,
- to reduce the noise in 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.

The main options to fix this problem are:

- selecting a more powerful model, with more parameters,
- feeding better features to the learning algorithm (feature engineering),
- reducing the constraints on the model.

Main Challenges of Machine Learning

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

}

“Bad data”

}

“Bad algorithm”

Testing and Validating

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



Data

Data



Training

Test

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.

Data



Training

Test

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

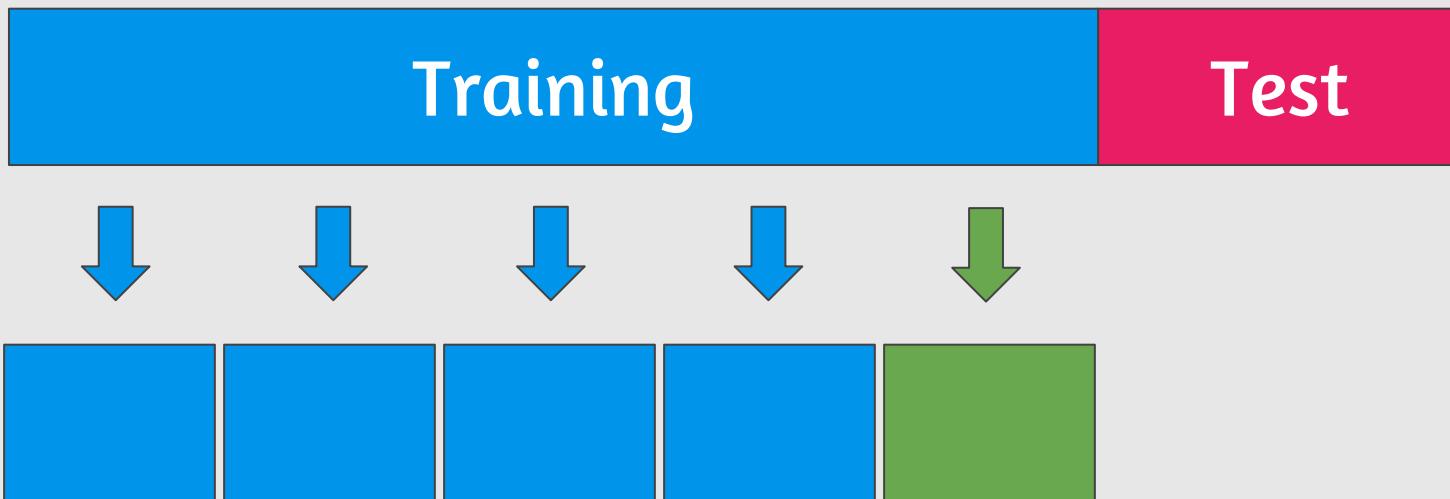


Training

Validation

Test

Cross Validation

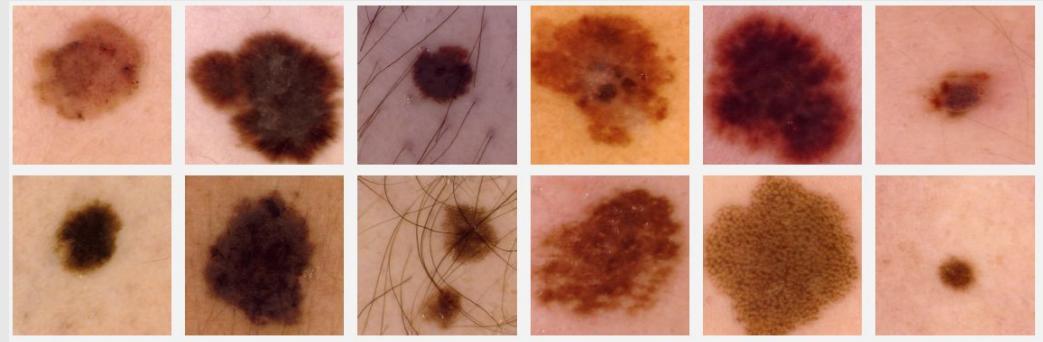


Training

Test

Skin Cancer Classification

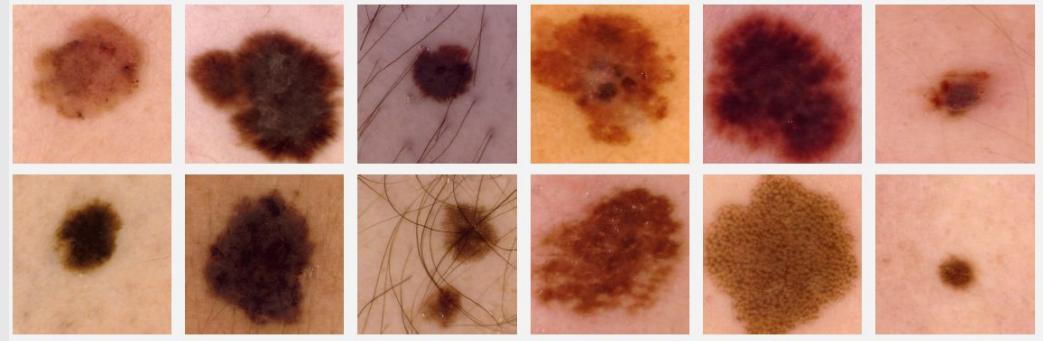
ISBI Challenge 2017



Training data & Validation data & Test data
2 000 images 150 images 600 images

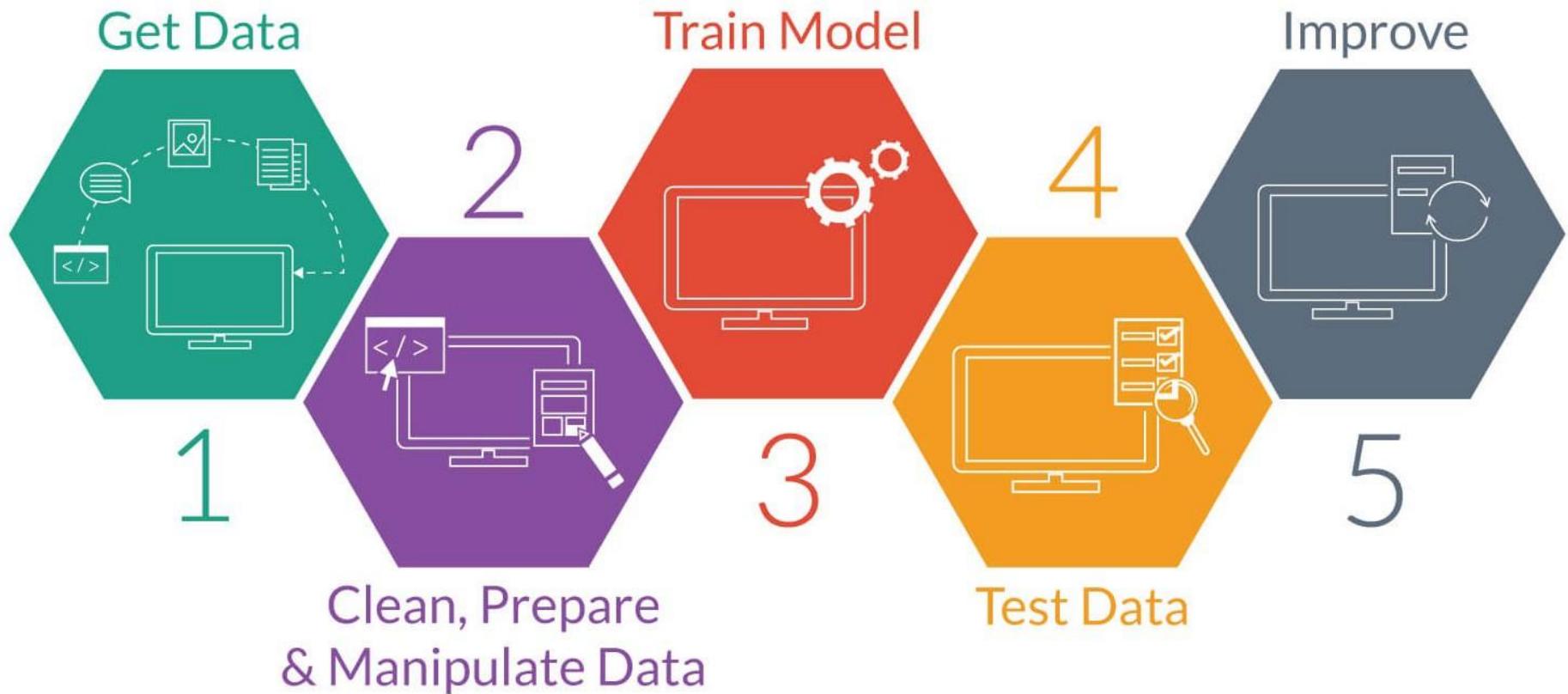
Skin Cancer Classification

ISBI Challenge 2017



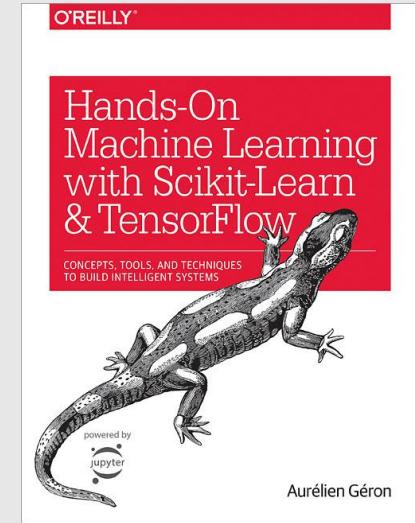
Training data & Validation data & Test data
2 000 images 150 images 600 images
95.1% **90.8%** **87.4%**
(internal validation)

Summary



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
- Biggest doubt in class (individual) **5%**
- 3 practical assignments (group of 2 people) **$3 \times 20\%$**
- 1 larger project (group of 4 or 5 people) **$1 \times 35\%$**
- Academic infraction ⇒ Zero



Prerequisites

- Some Python programming experience (R / Octave)
 - <http://learnpython.org/>
- Calculus, Linear algebra, Probabilities and Statistics

Syllabus

<http://www.ggte.unicamp.br/moodle/>

That's all!



Credit: <http://carlvondrick.com/ihog/>