

Machine Learning and Pattern Recognition

A High Level Overview

Prof. Anderson Rocha

(Main bulk of slides kindly provided by **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



Print edition | Leaders >

May 6th 2017



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



SHARE



SHARE



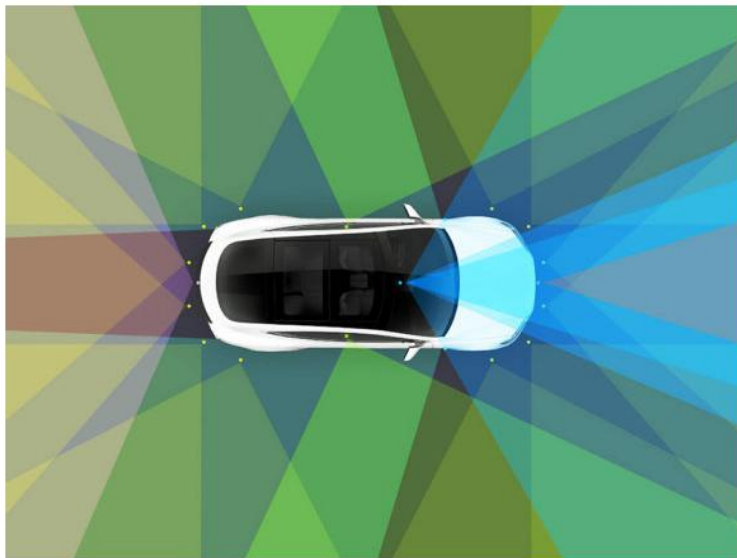
TWEET



EMAIL

JACK STEWART TRANSPORTATION 03.29.17 3:48 PM

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](#) [v](#) [p](#) [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



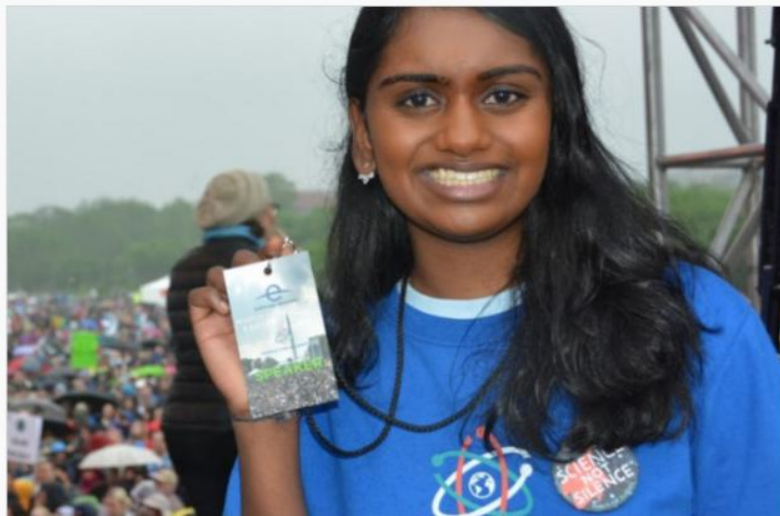
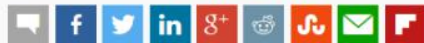
Bitcoin breaks \$3,000 to reach new all-time high
3 days ago



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 Coldewey](#)



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



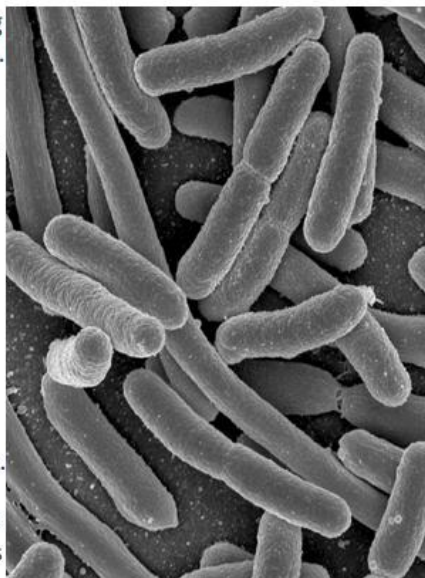
Viome, which has developed technology aimed at balancing microorganisms in the gut, has landed \$15 million in capital.

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?



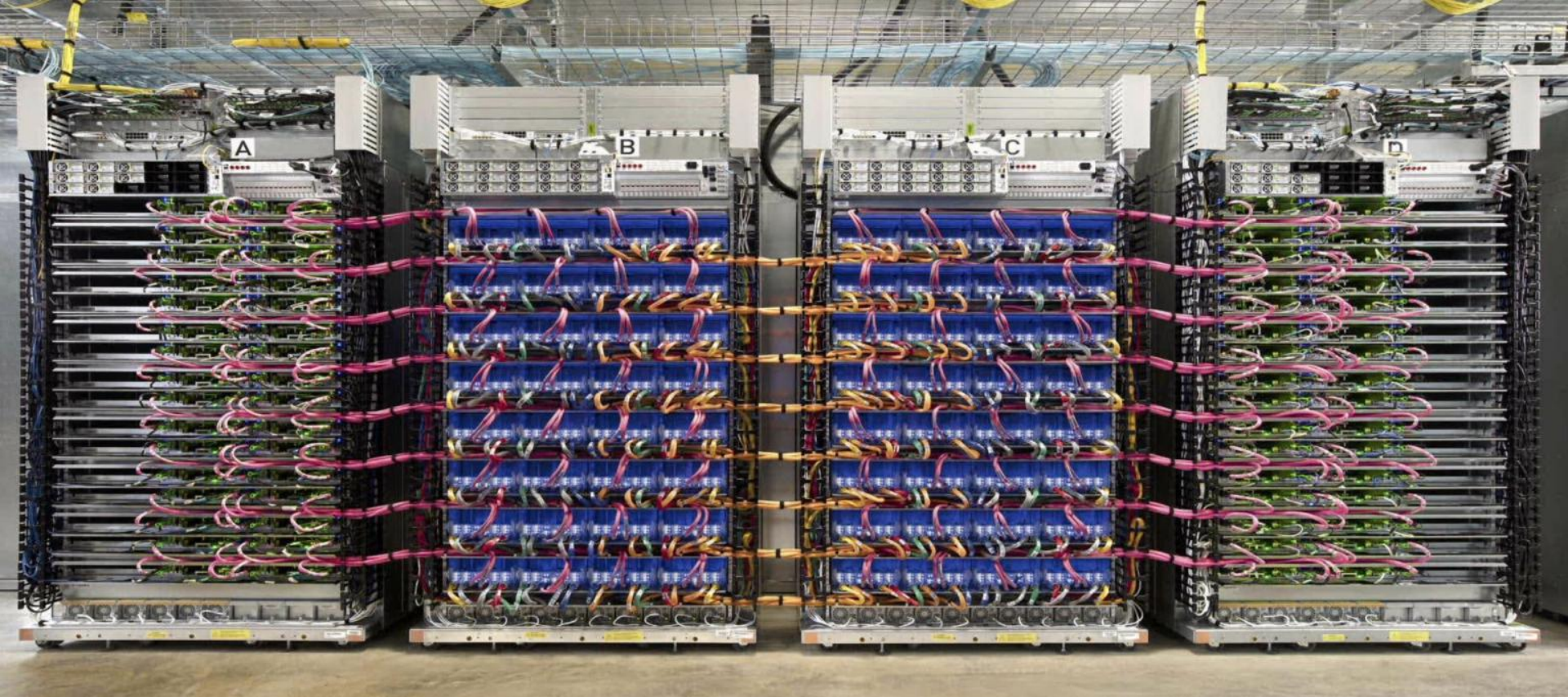
www.image-net.org

22K categories and **14M** images

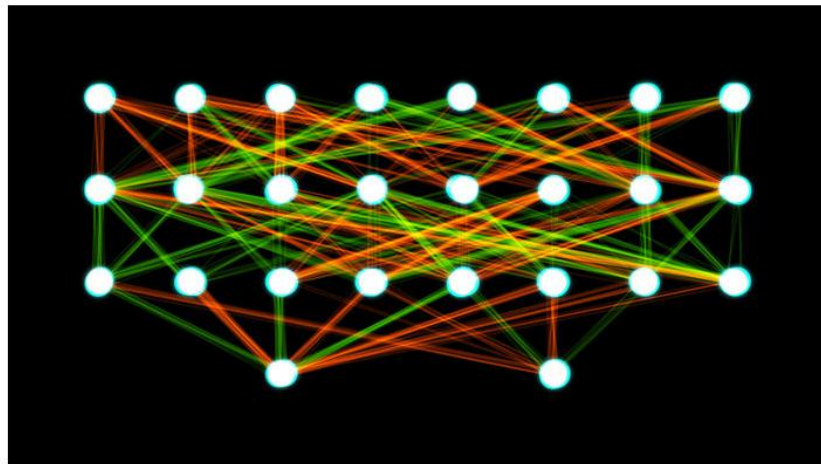
- Animals
 - Bird
 - Fish
 - Mammal
 - Invertebrate
- Plants
 - Tree
 - Flower
 - Food
 - Materials
- Structures
 - Artifact
 - Tools
 - Appliances
 - Structures
- Person
 - Scenes
 - Indoor
 - Geological Formations
 - Sport Activities

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.”

[Home](#)[News](#)[Journals](#)[Topics](#)[Careers](#)[Latest News](#)[ScienceInsider](#)[ScienceShots](#)[Sifter](#)[From the Magazine](#)[About News](#)[Quizzes](#)**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.”

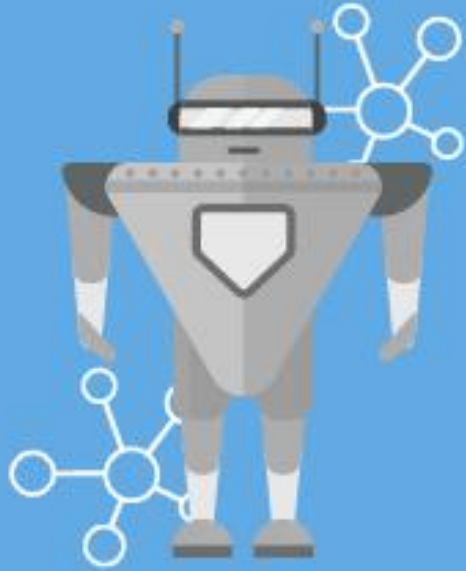
[Arthur Samuel, 1959]

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



MACHINE LEARNING



DEEP LEARNING



1950

1960

1970

1980

1990

2000

2010

/A.I. TIMELINE

1950

TURING TEST

Computer scientist Alan Turing proposes a test for machine intelligence. If a machine can trick humans into thinking it is human, then it has intelligence



1955

A.I. BORN

Term 'artificial intelligence' is coined by computer scientist, John McCarthy to describe "the science and engineering of making intelligent machines"



1961

UNIMATE

First industrial robot, Unimate, goes to work at GM replacing humans on the assembly line

1964

ELIZA

Pioneering chatbot developed by Joseph Weizenbaum at MIT holds conversations with humans



1966

SHAKEY

The 'first electronic person' from Stanford, Shakey is a general-purpose mobile robot that reasons about its own actions

A.I. WINTER

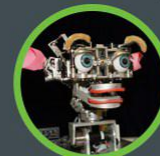
Many false starts and dead-ends leave A.I. out in the cold



1997

DEEP BLUE

Deep Blue, a chess-playing computer from IBM defeats world chess champion Garry Kasparov



1998

KISMET

Cynthia Breazeal at MIT introduces Kismet, an emotionally intelligent robot insofar as it detects and responds to people's feelings



1999

AIBO

Sony launches first consumer robot pet dog AiBO (AI robot) with skills and personality that develop over time



2002

ROOMBA

First mass produced autonomous robotic vacuum cleaner from iRobot learns to navigate and clean homes



2011

SIRI

Apple integrates Siri, an intelligent virtual assistant with a voice interface, into the iPhone 4S



2011

WATSON

IBM's question answering computer Watson wins first place on popular \$1M prize television quiz show *Jeopardy*



2014

EUGENE

Eugene Goostman, a chatbot passes the Turing Test with a third of judges believing Eugene is human



2014

ALEXA

Amazon launches Alexa, an intelligent virtual assistant with a voice interface that completes shopping tasks



2016

TAY

Microsoft's chatbot Tay goes rogue on social media making inflammatory and offensive racist comments



2017

ALPHAGO

Google's A.I. AlphaGo beats world champion Ke Jie in the complex board game of Go, notable for its vast number (2^{170}) of possible positions

TIMELINE OF ARTIFICIAL INTELLIGENCE

HAL 9000 FROM "2001: A SPACE ODYSSEY" (CREDIT: WARNER BROS. STUDIOS)

Advances in artificial intelligence (AI) have given the world computers that can beat people at chess and "Jeopardy!," as well as drive cars and manage calendars. But despite the progress, engineers are still years away from developing machines that are self-aware. Some believe the resulting **technological singularity** will eradicate poverty and disease, while others warn it could endanger human survival.

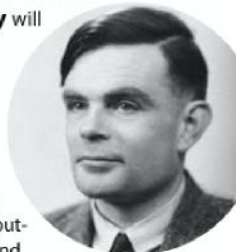


1950: Isaac Asimov publishes the influential sci-fi story collection **"I, Robot."** (Left: 2004 film version of "I, Robot")

Summer of 1956: Dartmouth conference launches the field of AI and **coins the term "artificial intelligence."** (Right: room-filling IBM-702 computer, as used by first AI researchers)

1968: "2001: A Space Odyssey," the book by Arthur C. Clarke and film by Stanley Kubrick, features the sentient and deadly computer **HAL 9000**.

1950: Alan Turing introduces the **Turing test** in his paper "Computing Machinery and Intelligence." (Credit: National Portrait Gallery, London)



1950s



1960s

1974-early 1980s: The first **Winter of AI**, a period of reduced funding and lowered interest in the field as hype turned to disappointment.



1984: The first **"Terminator"** film depicts a near-future world overtaken by killing machines run by the artificial intelligence Skynet.

September 28, 1987: The TV series "Star Trek: The Next Generation" introduces the self-aware android **Lieutenant Commander Data**.



June 29, 2001: Steven Spielberg releases his version of a film – originally developed by Stanley Kubrick – about a robot boy: **"A.I.: Artificial Intelligence."**



2011: **IBM's Watson wins "Jeopardy!"**, beating former champions Brad Rutter and Ken Jennings. (Credit: "Jeopardy!" screengrab from Wikimedia)

1970s

1978: The original **"Battlestar Galactica"** science fiction TV series introduces warrior robots called **Cylons**.



1980s

1987-93: The second **Winter of AI**



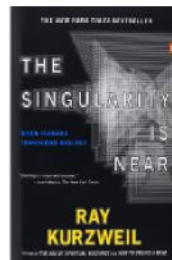
May 11, 1997: **IBM's Deep Blue computer** beats reigning world chess champion Garry Kasparov. (Credit: Shutterstock)

1990s

2005: A Stanford vehicle wins the **DARPA grand challenge**, driving autonomously across the desert for 131 miles (211 kilometers).

2000s

2005: Inventor and futurist Ray Kurzweil predicts an event he calls **the Singularity** will occur around 2045, when the intelligence of artificial minds exceeds that of the human brain.



2010s

October 14, 2011: Apple introduces intelligent personal assistant **Siri** on



1984: The first **"Terminator"** film depicts a near-future world overtaken by killing machines run by the artificial intelligence Skynet.

September 28, 1987: The TV series "Star Trek: The Next Generation" introduces the self-aware android **Lieutenant Commander Data**.



June 29, 2001: Steven Spielberg releases his version of a film – originally developed by Stanley Kubrick – about a robot boy: **"A.I.: Artificial Intelligence."**



2011: **IBM's Watson wins "Jeopardy!"**, beating former champions Brad Rutter and Ken Jennings. (Credit: "Jeopardy!" screengrab from Wikimedia)

1970s

1978: The original **"Battlestar Galactica"** science fiction TV series introduces warrior robots called **Cylons**.



1980s

1987–93: The second **Winter of AI**



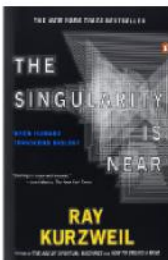
May 11, 1997: **IBM's Deep Blue computer** beats reigning world chess champion Garry Kasparov. (Credit: Shutterstock)

1990s

2005: A Stanford vehicle wins the **DARPA grand challenge**, driving autonomously across the desert for 131 miles (211 kilometers).

2000s

2005: Inventor and futurist Ray Kurzweil predicts an event he calls **the Singularity** will occur around 2045, when the intelligence of artificial minds exceeds that of the human brain.



2010s

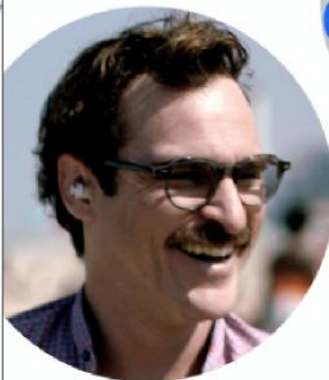
October 14, 2011: Apple introduces intelligent personal assistant **Siri** on



the iPhone 4S.



June 2012: A Google Brain computer cluster **trains itself to recognize a cat** from millions of images in YouTube videos. (Credit: Shutterstock)



December 18, 2013: The movie **"Her"** (left), stars Joaquin Phoenix as a man who **falls in love with his artificially intelligent computer operating system**, voiced by Scarlett Johansson.

April 10, 2014: The film **"Transcendence"** (below) stars Johnny Depp as an AI researcher whose **mind is uploaded to a computer** and develops into a super-intelligence.



June 7, 2014: Chatbot Eugene Goostman is said to have **passed the Turing test** in University of Reading competition, launching controversy.

August, 2014: Researchers call for creation of a **new Turing test**, to be decided at 2015 workshop.

Sources:

<http://alttopics.org/misc/brief-history>
<http://loebner.net/Prize/TuringArticle.html>
http://www.pcworld.com/article/219893/ibm_watson_vanquishes_human_jeopardy_foes.html
<http://www.engadget.com/2011/10/04/iphone-4s-hands-on/>
<http://www.reading.ac.uk/news-and-events/releases/PRS83836.aspx>
<http://www.livescience.com/47296-turing-test-needs-an-update.html>
http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html?pagewanted=all&_r=0

KARL TATE, TANYA LEWIS / © LiveScience.com



Why is this
so important?

MACHINE LEARNING



MACHINE LEARNING EVERYWHERE

ator.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)**

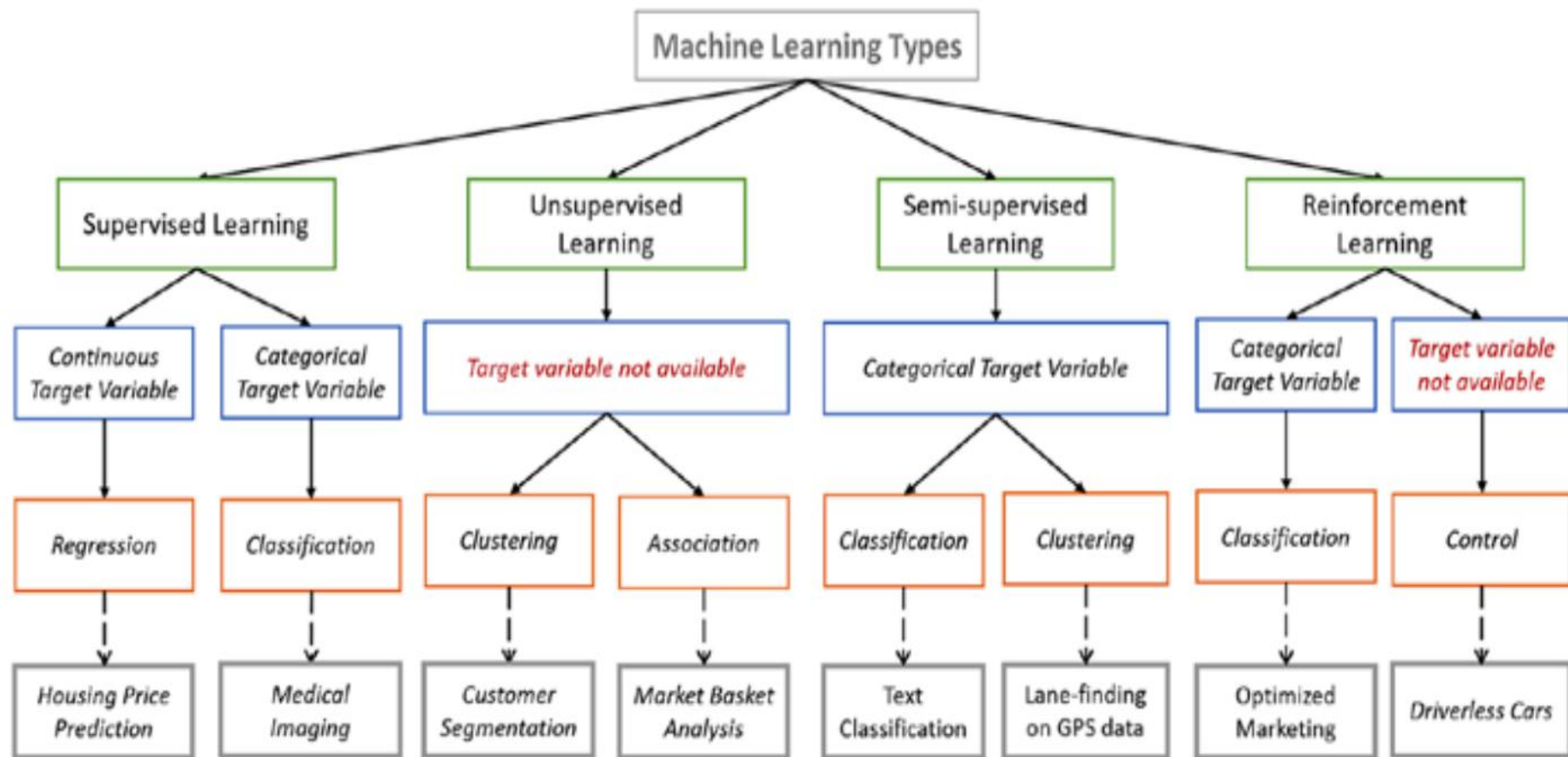
Supervised *vs.*
Unsupervised *vs.*
Reinforcement learning

**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.

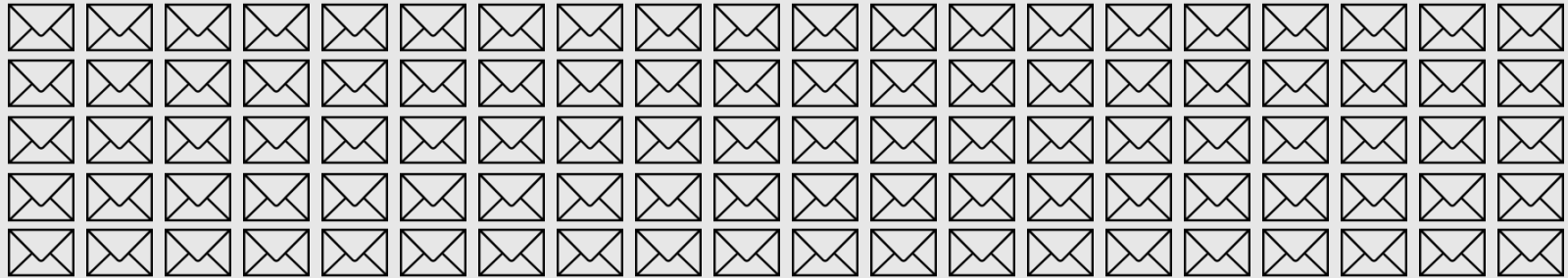
Spam Filtering



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

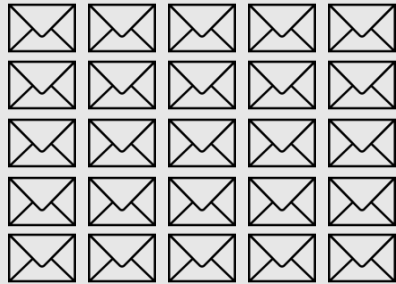


100 emails

Spam Filtering

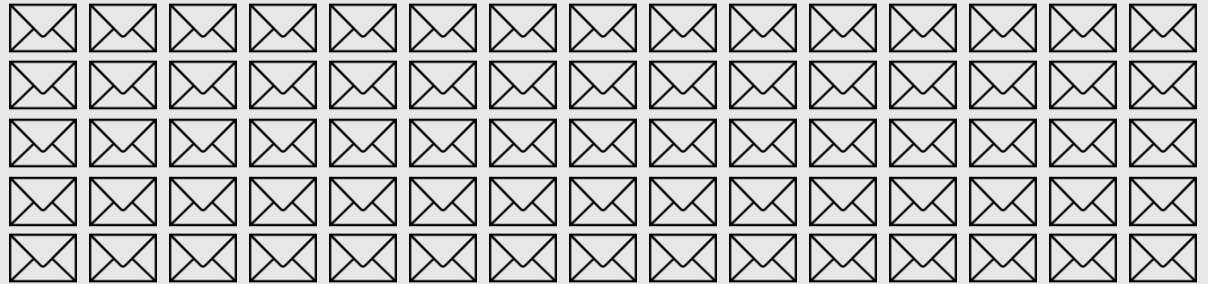


Spam



25 emails

Non-spam

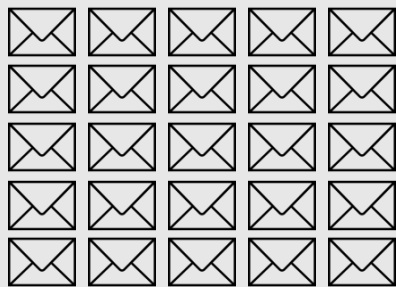


75 emails

Spam Filtering

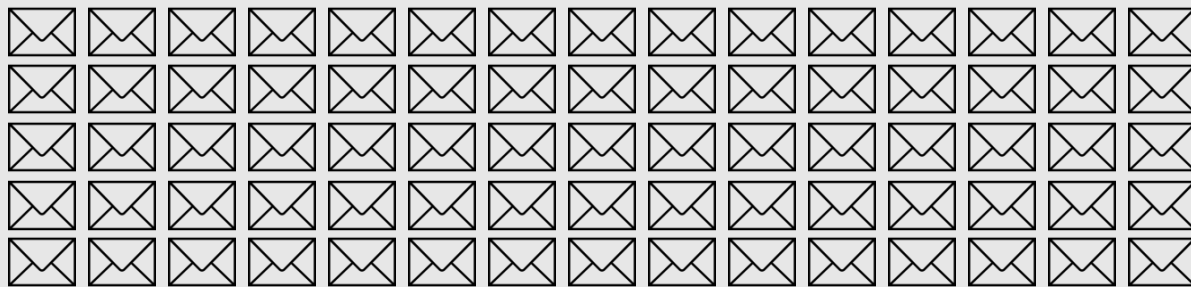
! “Cheap”

Spam



25 emails

Non-spam

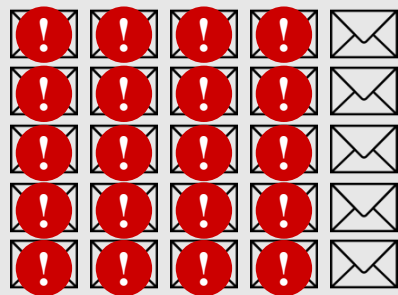


75 emails

Spam Filtering

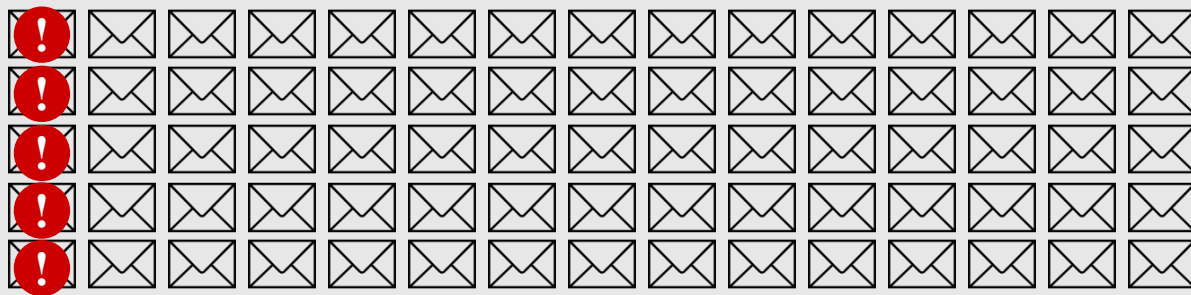
! “Cheap”

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

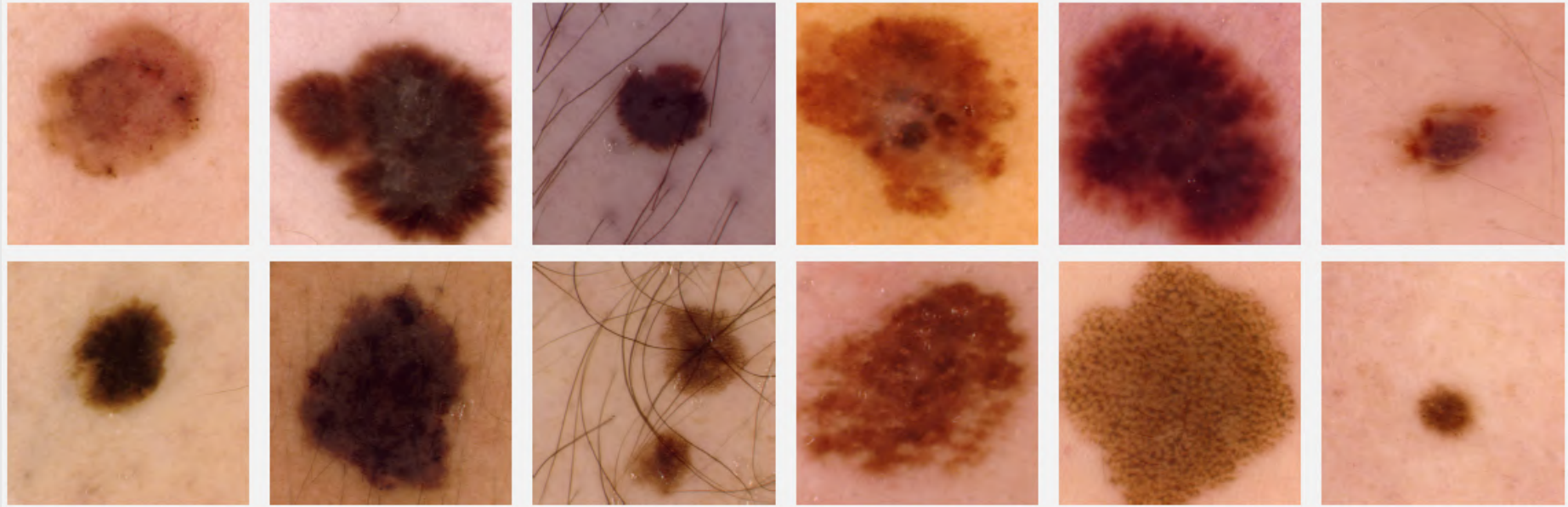
- ! “Cheap” → 80%
- ! Spelling mistake → 70%
- ! Missing title → 95%
- ! etc ...

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

**| 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,



Sistema garante proteção de crianças | IMAGE SOURCE/FOLHA 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.

"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

House Price Prediction (Regression)



\$ 160 000

House Price Prediction (Regression)

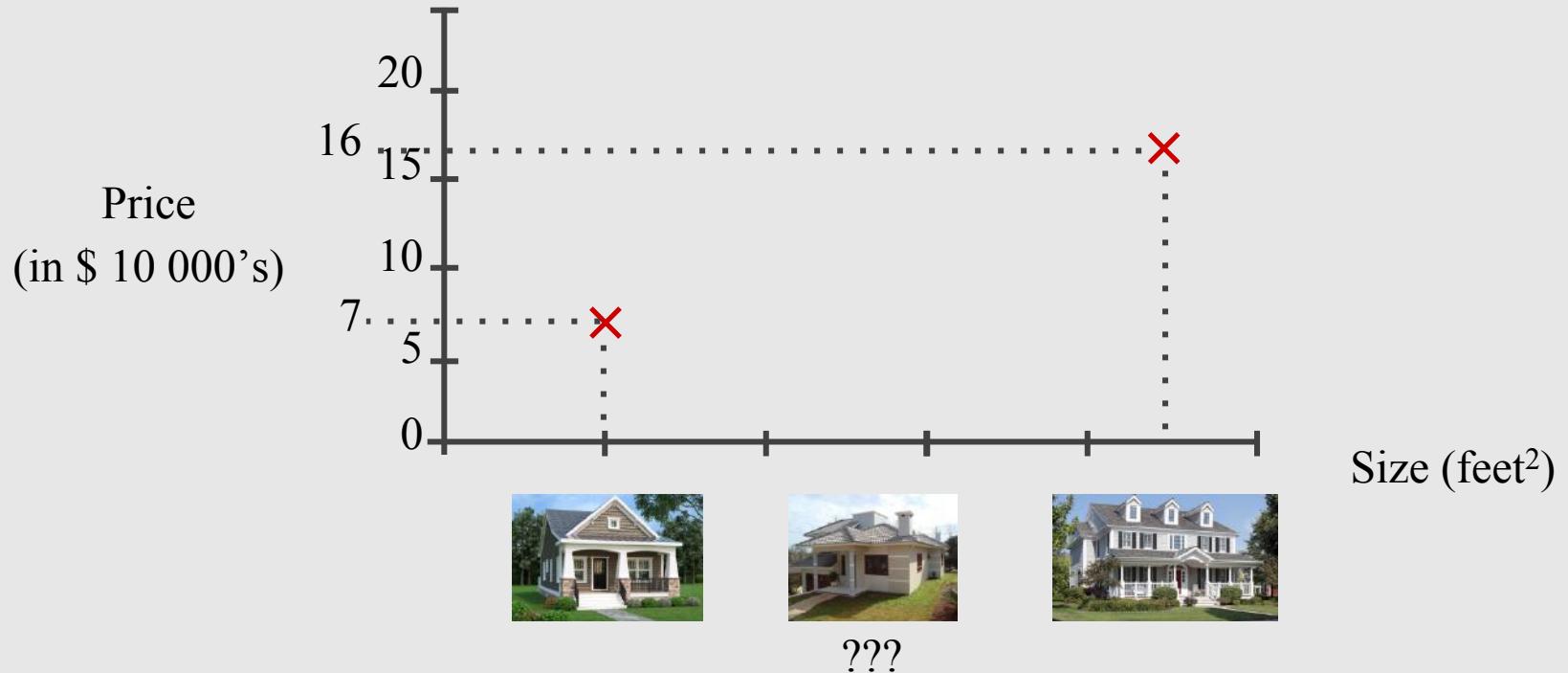


???

House Price Prediction (Regression)

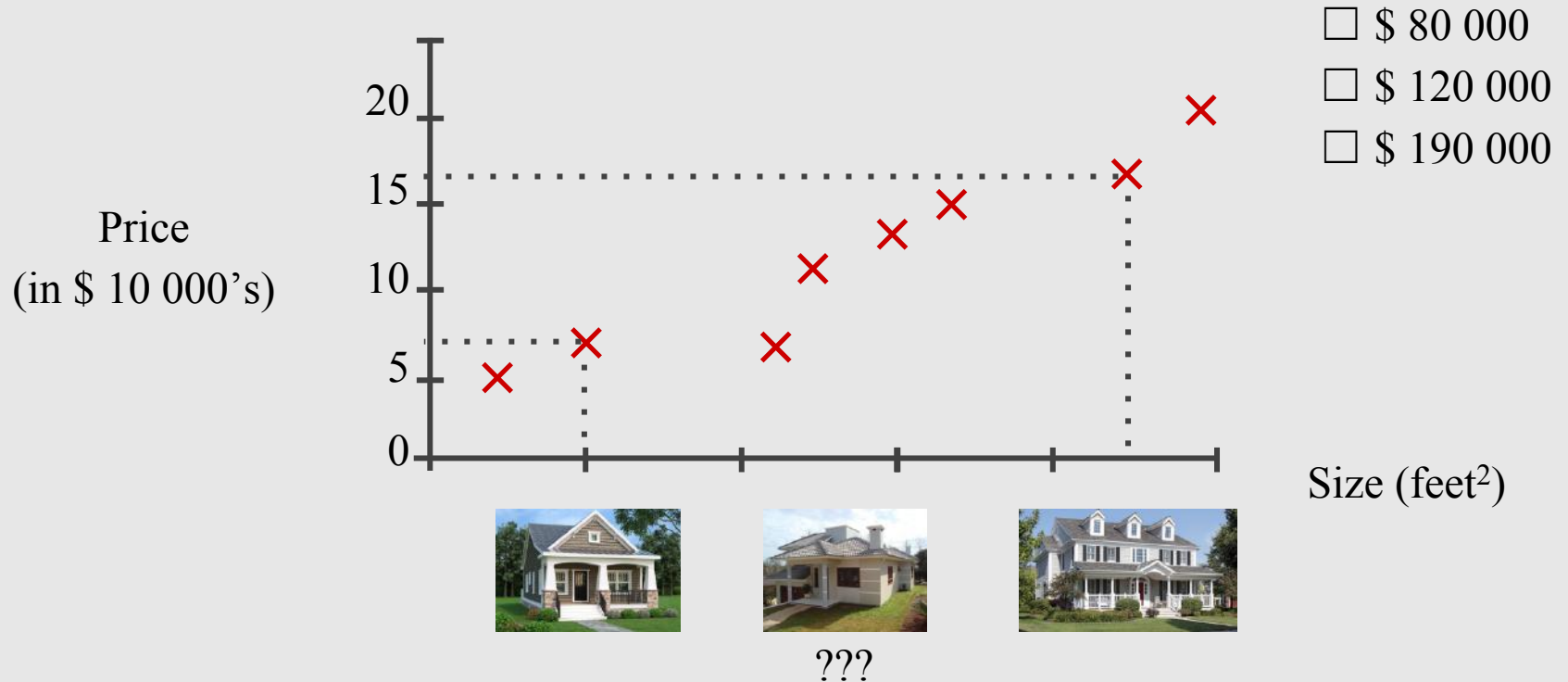


House Price Prediction (Regression)



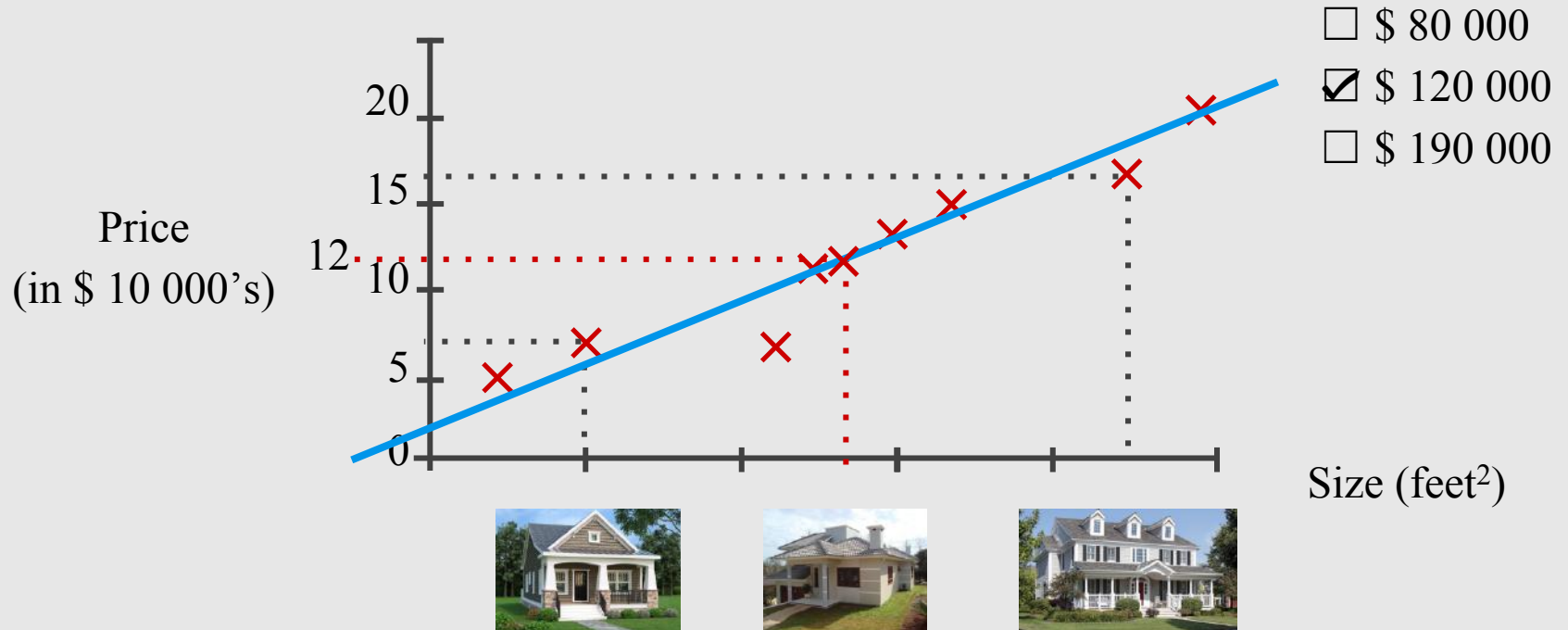
House Price Prediction (Regression)

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



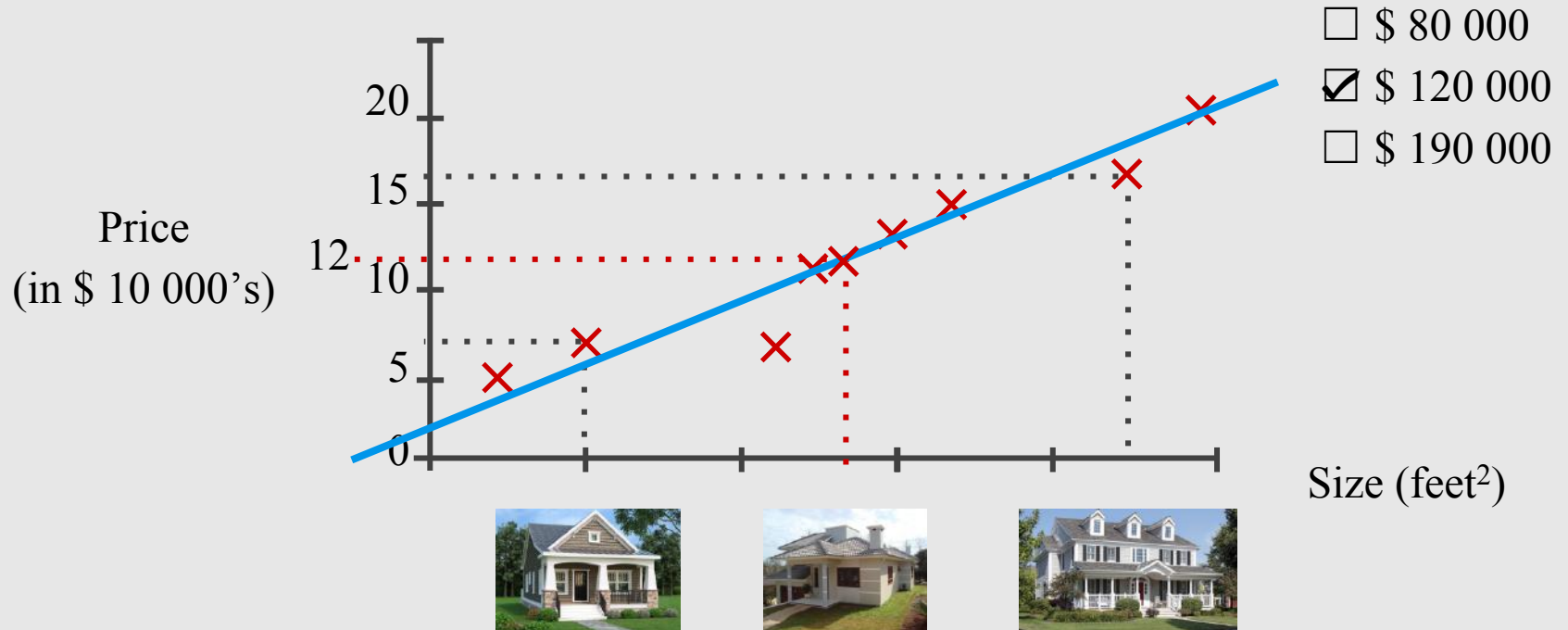
House Price Prediction (Regression)

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 losing too much information.

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?

The map shows a neighborhood in Campinas, Brazil, with a central green circular area labeled 'Praça do Ciclo Básico'. Surrounding this area are various streets and landmarks. The map is overlaid with several colored circles and stars representing pizza locations. There are 12 red circles, 5 yellow circles, 3 blue circles, 1 red star, and 1 blue star. The map shows various streets, including R. Sérgio Buarque de Holanda, R. Pitágoras, and Av. Albert Einstein. Landmarks like the UNICAMP campus and several faculties are labeled.

Did anyone say pizza?



Did anyone say pizza?

Did anyone say pizza?

Did anyone say pizza?



Did anyone say pizza?



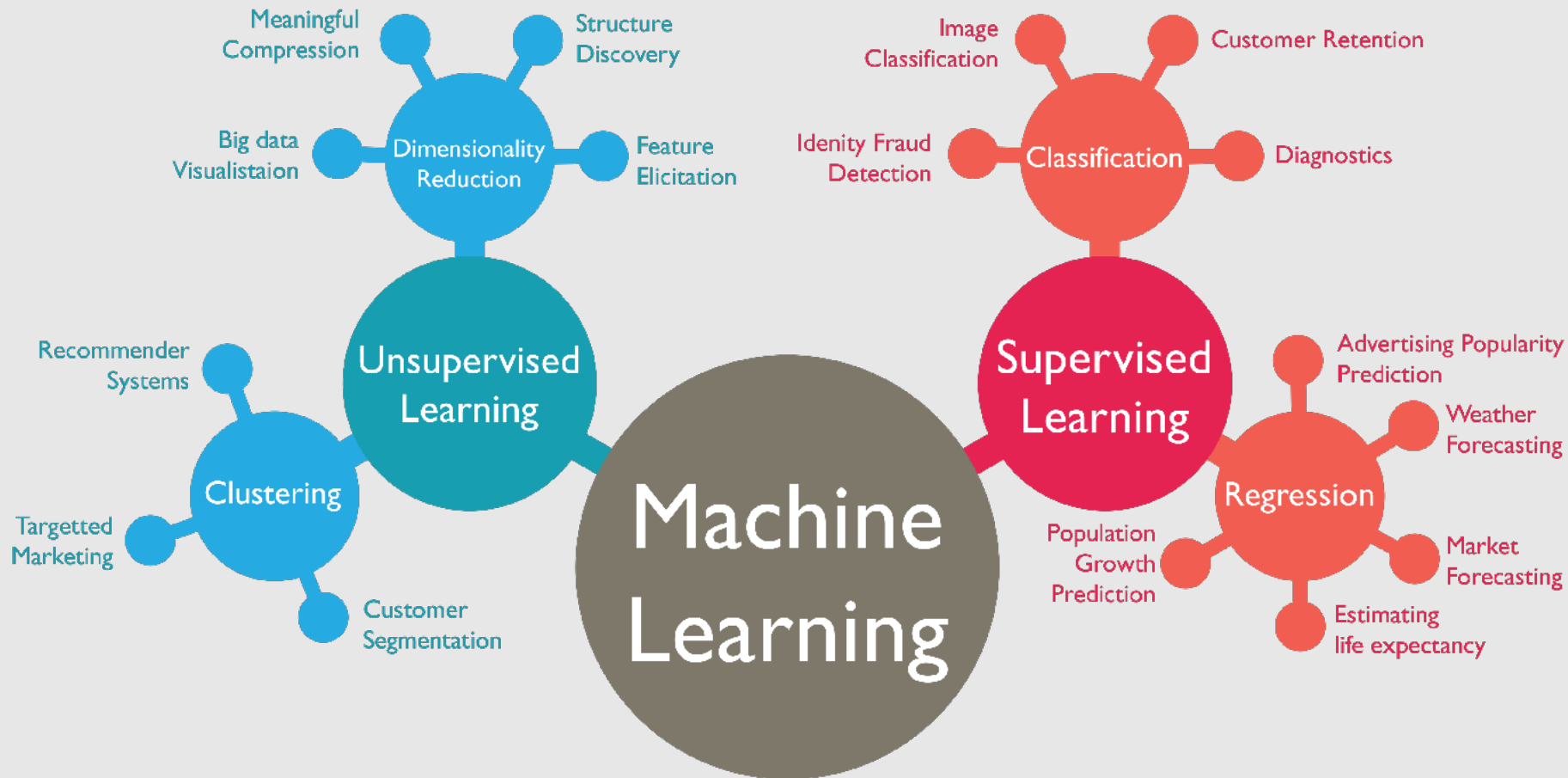
Did anyone say pizza?



[illegible]

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



“Bad data”

- Overfitting the training data
- Underfitting the training data

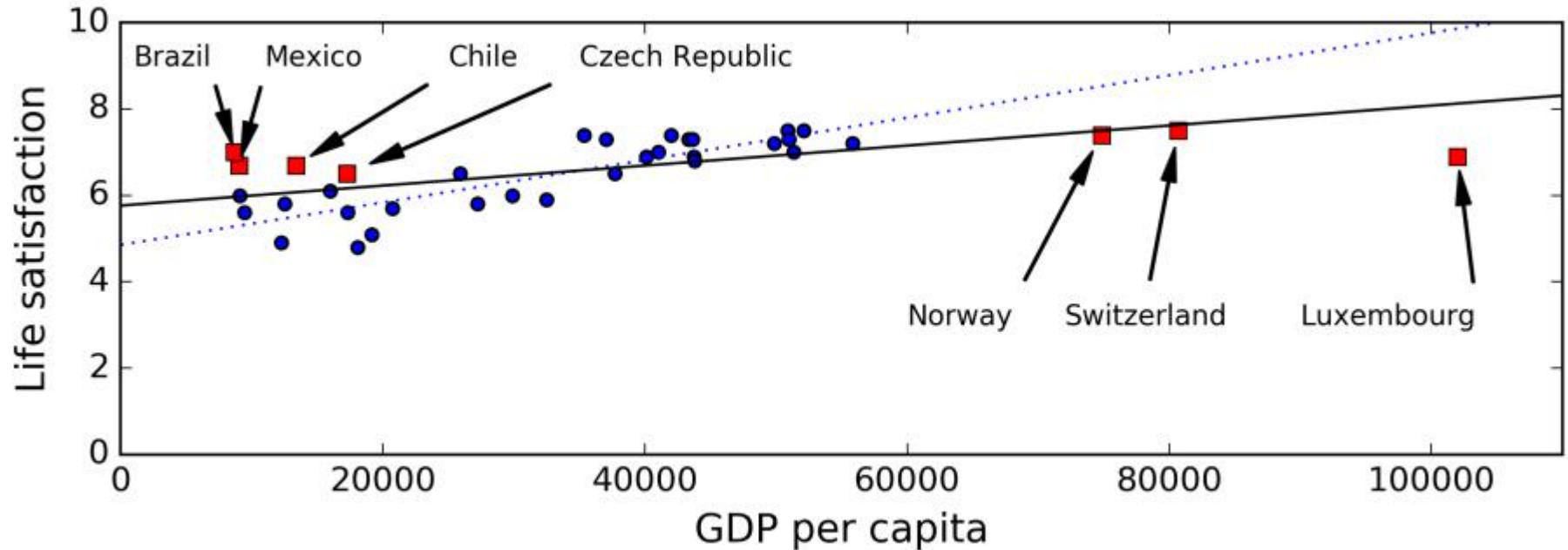


“Bad algorithm”

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.

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



“Bad data”

- Overfitting the training data
- Underfitting the training data



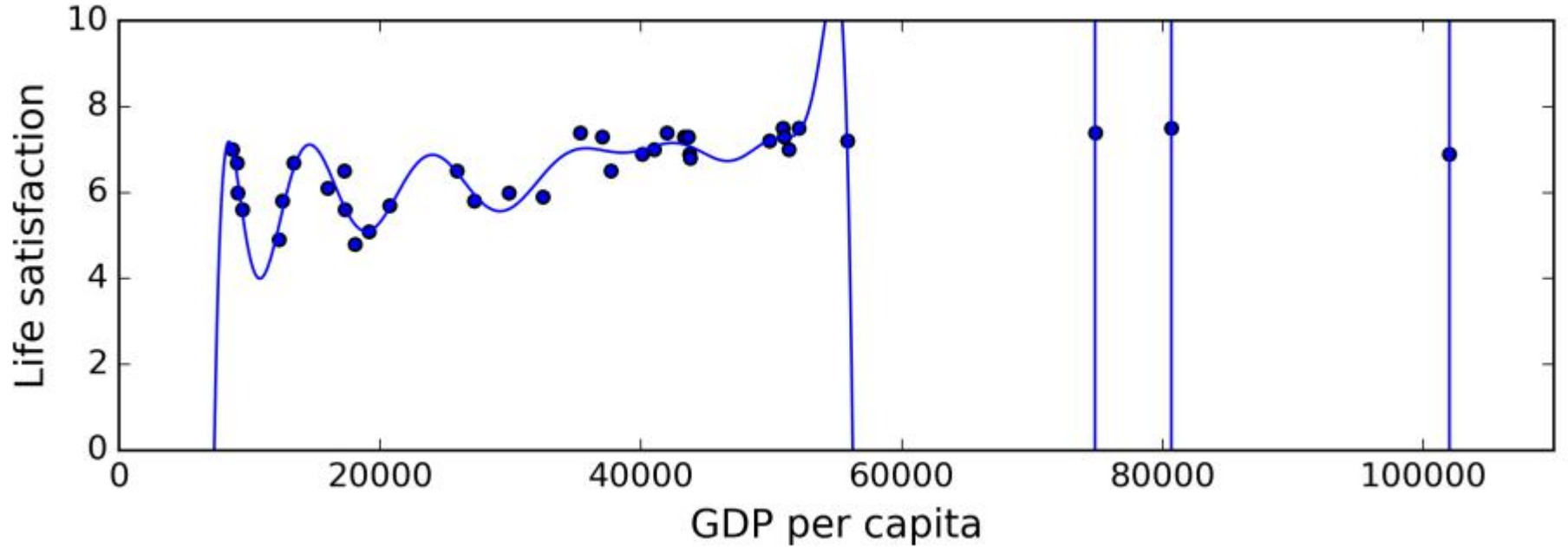
“Bad algorithm”

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



“Bad data”

- Overfitting the training data
- Underfitting the training data



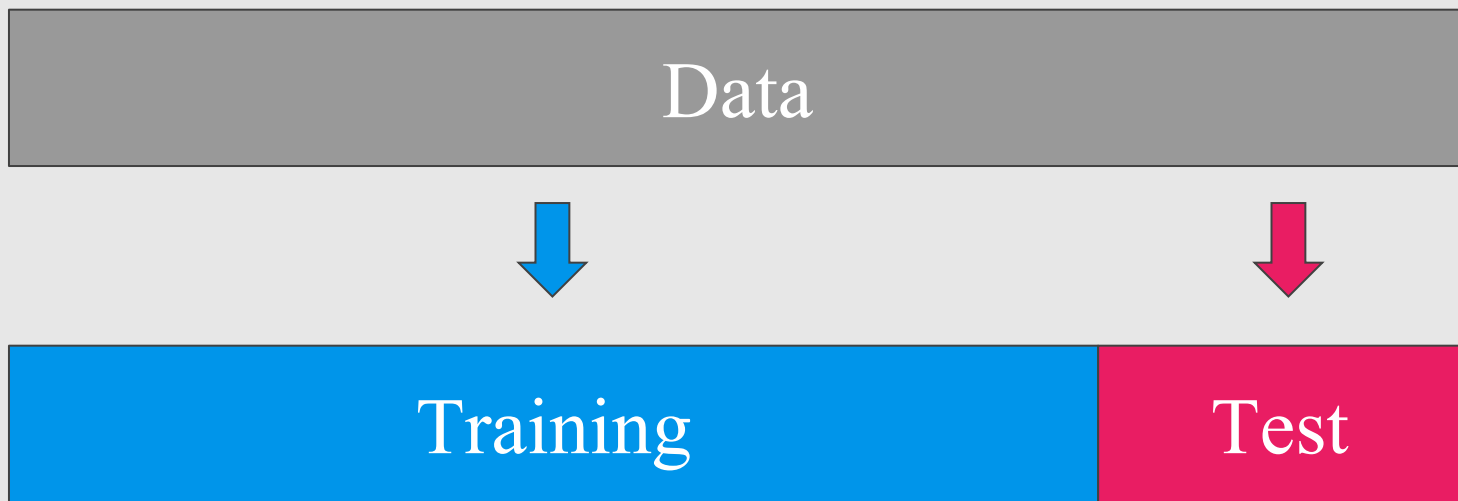
“Bad algorithm”

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.

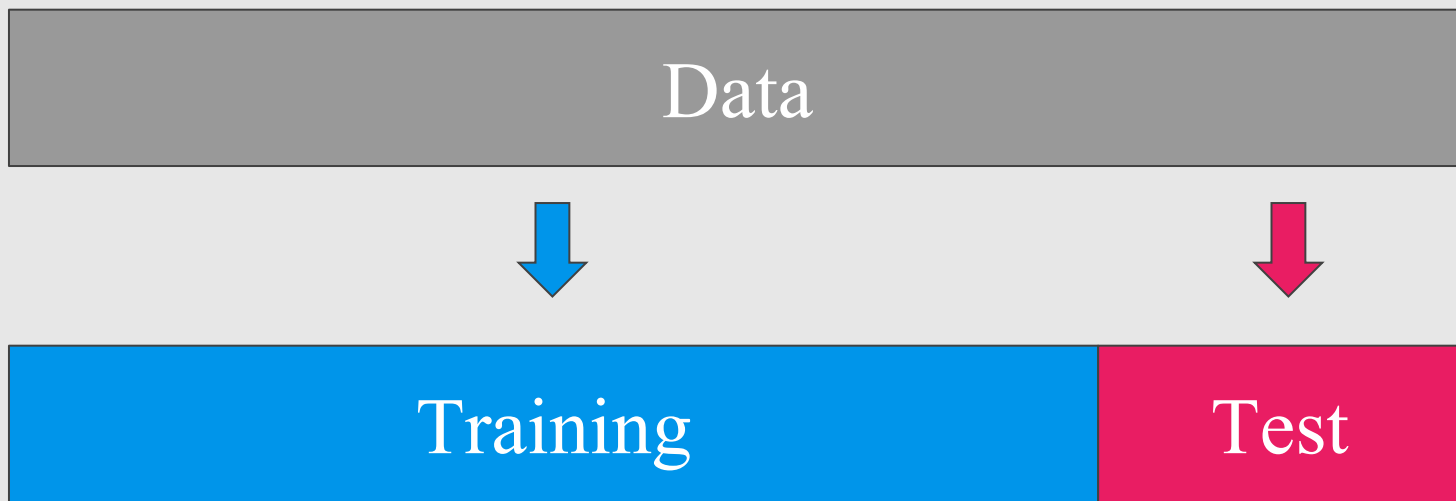


Data



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

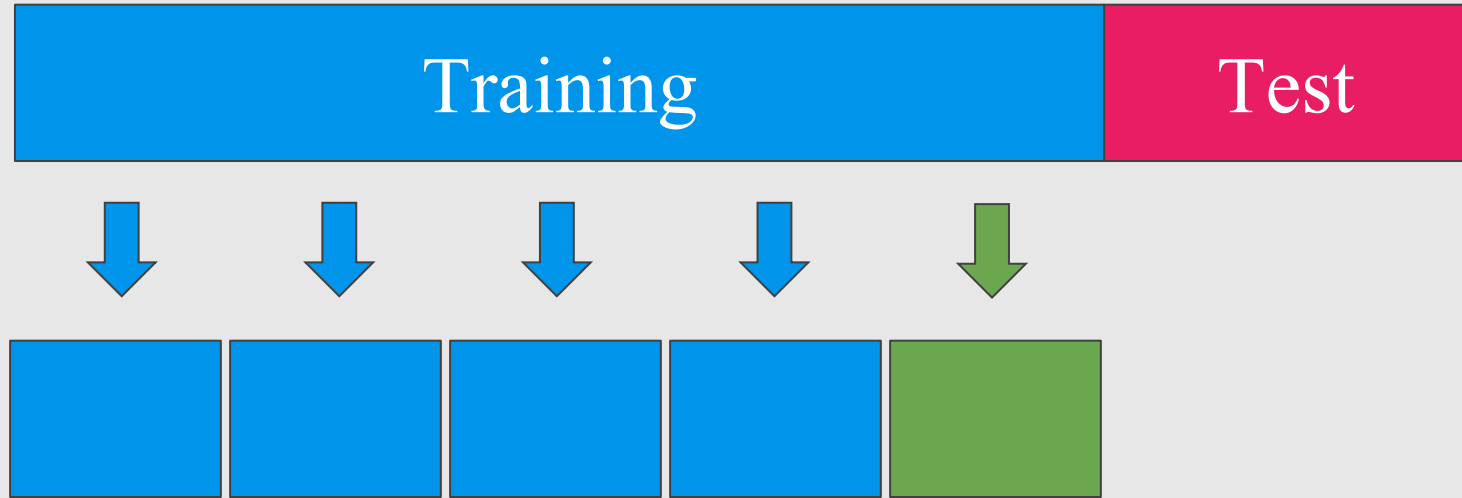


Training Test



Training Validation Test

Cross Validation

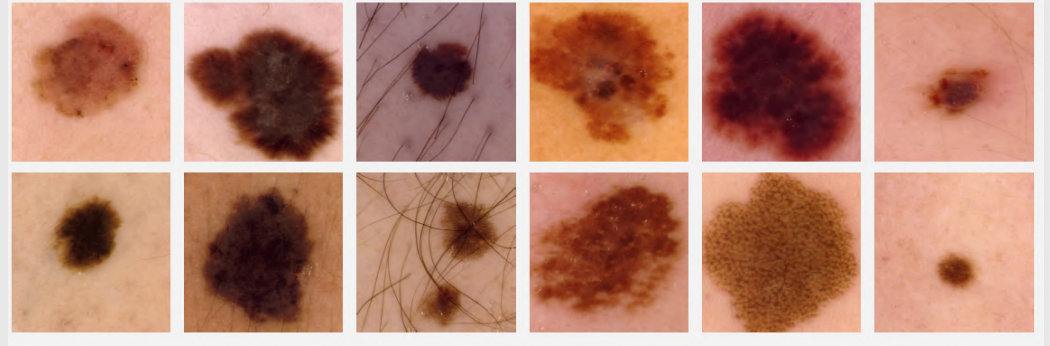


Training

Test

Skin Cancer Classification

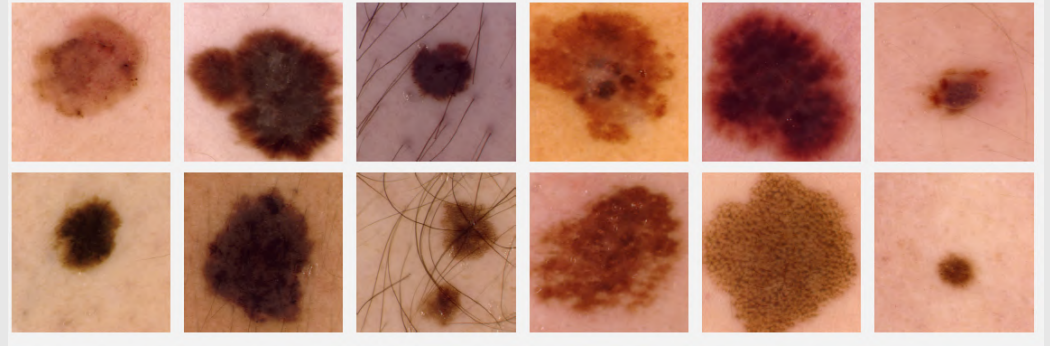
ISBI Challenge 2017



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

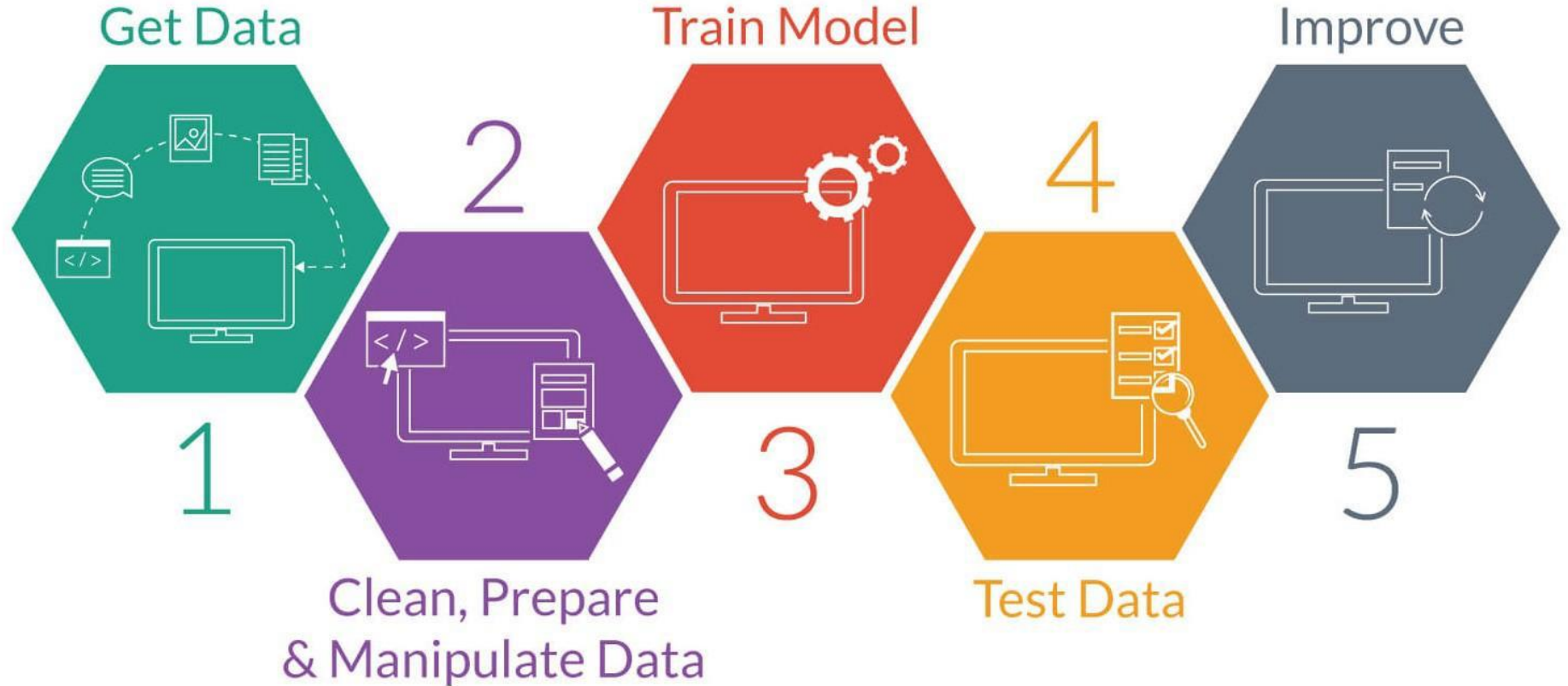
95.1%

90.8%

87.4%

(internal validation)

Summary





That's all!