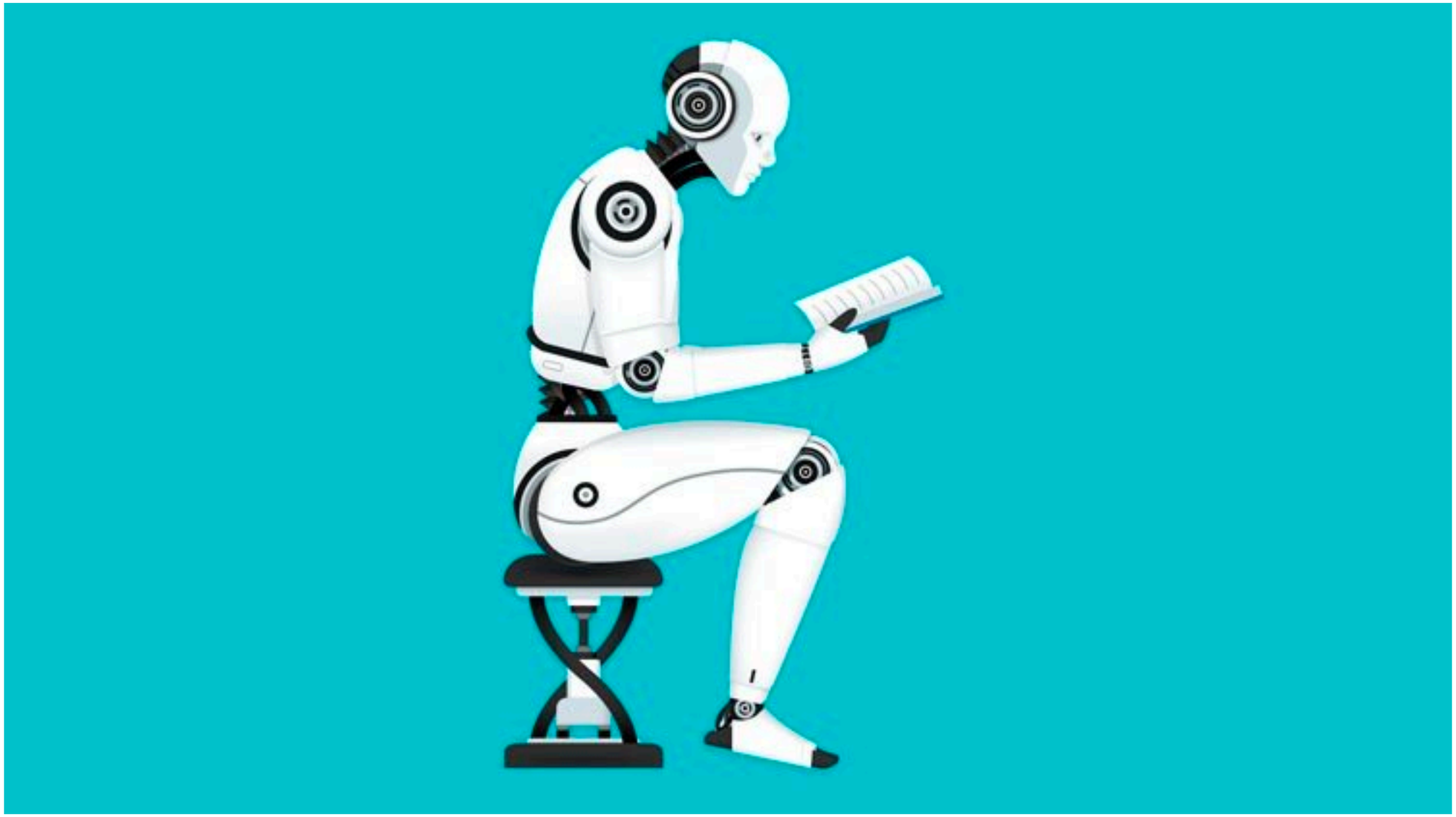
UNIVERSITAT DE BARCELONA



# Presentació Curs

**Machine Learning | Enginyeria Informàtica**

Santi Seguí | 2020-2021

# 1. Objectius del curs

## Objectius del curs

- Introducció descriptiva a un conjunt de tècniques i mètodes basats amb l'aprenentatge automàtic (Machine Learning).
- Coneixement dels principis en que es bases algunes d'aquestes tècniques.
- Coneixement dels principis d'avaluació dels mètodes d'aprenentatge automàtic.
- Contacte pràctic amb exemples representatius amb diversitat de dades

# 2. Prerequisits

## Prerequisits

- Conceptes dels cursos de Càlcul i Àlgebra
- Algunes idees generals del curs de Probabilitat i Estadística
- Curiositat per la **Intel·ligència Artificial**

# 3. Organització

## Coordinador:

- **Santi Seguí** **Email:** santi.segui@ub.edu

## **Teoria Online** (Divendres 12.30–13.30h)

- **Santi Seguí** **Email:** santi.segui@ub.edu

## **Teoricopràctica** (Dijous 15-16.30h)

- **Santi Seguí** **Email:** santi.segui@ub.edu

## **Laboratoris** (Dimarts 18h-19h)

- **Josep Fortiana** **Email:** fortiana@ub.edu

## Com s'organitza l'assignatura?

- L'assignatura s'imparteix en classes teòriques i pràctiques. L'assignatura es coordinarà mitjançat una eina electrònica (basada en Moodle 2) que s'anomena Campus Virtual i que és accessible a través de la web. A través d'aquest entorn tindreu: anuncis, apunts, notes, fòrum, calendari, enllaços a la bibliografia, etc.
  - https://campusvirtual.ub.edu/

## • Com seran les classes teòriques? (1 hora a la setmana)

- Contingut online acotat a una hora de dedicació

## Com seran les classes teòricopràctiques? (1.5 hores a la setmana)

- Classes presencials on es reforçaran conceptes i resoldran problemes

## • Com seran les classes pràctiques? (1 hores a la setmana)

- Sessió online. Les pràctiques es realitzen de forma individual o amb parelles.

## Llenguatge de programació?

- Python & R

## ENGINYERIA INFORMÀTICA

### Curs: 2020-2021 Assignatures - Horaris

### APRENENTATGE AUTOMÀTIC

Codi: **365828**

Pla docent

| Tipus | Impartició | Crédits | Curs/Semestre | Unitat Acadèmica |
|---|---|---|---|---|
| Optativa del grau | Quadrimestral | 6 | 4 / | Dept. Matemàtiques i Informàtica |

**Programació de l'oferta docent del Primer semestre**

**Activitat**

| Grup | Dies | Horari | Professorat | Aula | Idioma |
|---|---|---|---|---|---|
| **Teoria [No presencial]** | | | | | |
| M0 | dl. dt. dc. dj. **dv.** | 1r sem. 12.30-13.30 | Segui Mesquida, Santiago | Aula virtual - Matemàtiques i Informàtica | Català |
| **Teoricopràctica [Presencial]** | | | | | |
| MA | dl. dt. dc. **dj.** dv. | 1r sem. 15.00-16.30 | Segui Mesquida, Santiago | Aula B7 | |
| **Pràctiques de laboratori [No presencial]** | | | | | |
| Ma | dl. dt. dc. **dj.** dv. | 1r sem. 18.00-19.00 | Fortiana Gregori, Jose | Aula virtual - Matemàtiques i Informàtica | Català |
| **Exàmens : 1r parcial [Presencial]** | | | | | |
| G1 | | | | | - |
| **Exàmens : Final [Presencial]** | | | | | |
| G1 | | | | | - |
| **Exàmens : Reavaluació [Presencial]** | | | | | |
| G1 | | | | | - |

| | Theory (1 hora) | Presencial 1.5h | Pràctiques |
|---|---|---|---|
| 1-Oct | | Introduction | Introducció R |
| 8-Oct | A typical Machine Learning project | Your first DS problem | Regression |
| 15-Oct | Regression | A typical Machine Learning project | Regression |
| 22-Oct | Classification | Regression | Regression |
| 29-Oct | Trainining Models | Classification | Classification |
| 5-Nov | Support Vector Machines | Training Models | Classification |
| 12-Nov | Tree Based Methods | Examen Parcial | Classification |
| 19-Nov | Boosting & Bagging - Ensembles | Support Vector Machines & Tree Based Methods | Unsupervised Learning |
| 26-Nov | Dimensionality Reduction | Boosting & Bagging - Ensembles | Unsupervised Learning |
| 3-Dec | Unsupervised Learning | Dimensionality Reduction | Unsupervised Learning |
| 9-Dec | Neural Networks | Unsupervised Learning | NN |
| 17-Dec | Neural Networks | Neural Networks | NN |

# 4. Avaluació

# Avaluació Continuada

## Dues modalitats:

- Basada amb projectes
- Basada amb exàmens

## Avaluació Continuada
## Basada amb Exàmens

### Com s'avaluarà l'assignatura?

- **participació i entrega dels projectes**
- **lliurament de pràctiques**

### Proves presencials:

- Durant el curs es presentaran diversos projectes ( >=3).
- Cadascun d'aquests projectes tindrà una puntuació associada.
- La nota mínima final obtinguda ha de ser de 4 punts
- La nota màxima final que podrà obtenir l'alumne es de 10 punts
- L'alumne haurà de defensar el projecte i demostrar la seva autoria (a les sessions presencials o mitjançant sessions online específiques).

### Lliurament de pràctiques:

- Lliurament de pràctiques: Cada un dels lliuraments de pràctiques serà avaluat pel professor amb una nota que pot anar de 0 (nota mínima) a 10 (nota màxima). Si l'estudiant no lliura les pràctiques dins del període assenyalat, obtindrà un 0.
- La nota final (NP) de la part de pràctiques és la mitjana de tots els lliuraments (3 en total).

**IMPORTANT:** La nota final de teoria (**NT**) i la nota final de pràctiques (**NP**) han de tenir una nota mínima de 4.5 per fer mitja.

## Avaluació Continuada
## Basada amb Projectes

### Com s'avaluarà l'assignatura?

- L'assignatura seguirà un esquema d'avaluació continuada, amb dos elements principals: **proves presencials** i **lliurament de pràctiques**

### Proves presencials:

- Durant el curs, l'estudiant ha de fer 2 proves escrites sobre la teoria:
  - Parcial (NTP) + Final (NTF)
- **La nota final teoria NT = NTP/2 + NTF/2**

### Lliurament de pràctiques:

- Lliurament de pràctiques: Cada un dels lliuraments de pràctiques serà avaluat pel professor amb una nota que pot anar de 0 (nota mínima) a 10 (nota màxima). Si l'estudiant no lliura les pràctiques dins del període assenyalat, obtindrà un 0.
- La nota final (NP) de la part de pràctiques és la mitjana de tots els lliuraments (3 en total).

**IMPORTANT:** La nota final de teoria (**NT**) i la nota final de pràctiques (**NP**) han de tenir una nota mínima de 4.5 per fer mitja.

## Avaluació Única

- L'estudiant que es vulgui acollir a l'avaluació única ho ha de sol·licitar a la Secretaria de la Facultat dins del termini establert en cada curs acadèmic.

- Hi ha un examen final de teoria i un examen final de pràctiques de laboratori. Anomenem ***NT*** i ***NP***, respectivament, les notes obtingudes en aquests exàmens.
- Es requereix la presentació oral i escrita d'un treball de curs, prèviament acordat amb el professor. Anomenem ***NPTC*** la qualificació d'aquest treball.

- La nota final de l'assignatura (Nota_Final) es calcula mitjançant la fórmula següent: Nota_Final = 0,5 * NPTC + 0,2 * NT + 0,3 * NP.
- Per poder calcular la nota final és imprescindible una puntuació igual o superior a 3 en tots tres components.

# 5. Recursos

## GITHUB / Campus Virtual

*https://github.com/ssegui/ml_ub*

- Làmines de les sessions de l'aula
- Guions de les pràctiques
- Entregues
- Documentació i informació complementària

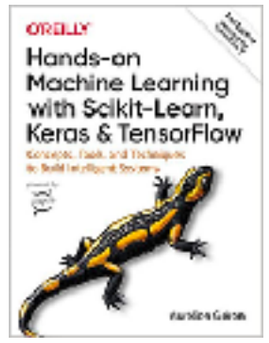
## Programari

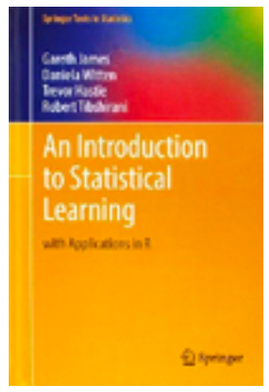
- Python & R

## Bibliografia

- Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. Aurelien Geron
- An Introduction to Statistical Learning: with Applications in R

PDF Online Gratuit: http://faculty.marshall.usc.edu/gareth-james/




Kahoot!

# 6. Delimitar els continguts de l'assignatura

# What is Machine Learning:

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

A more general definition: *Machine Learning is the filed of study that gives computers the ability to learn without being explicitly programmed.*

*— Arthur Samuel, 1959*

# What is Machine Learning:

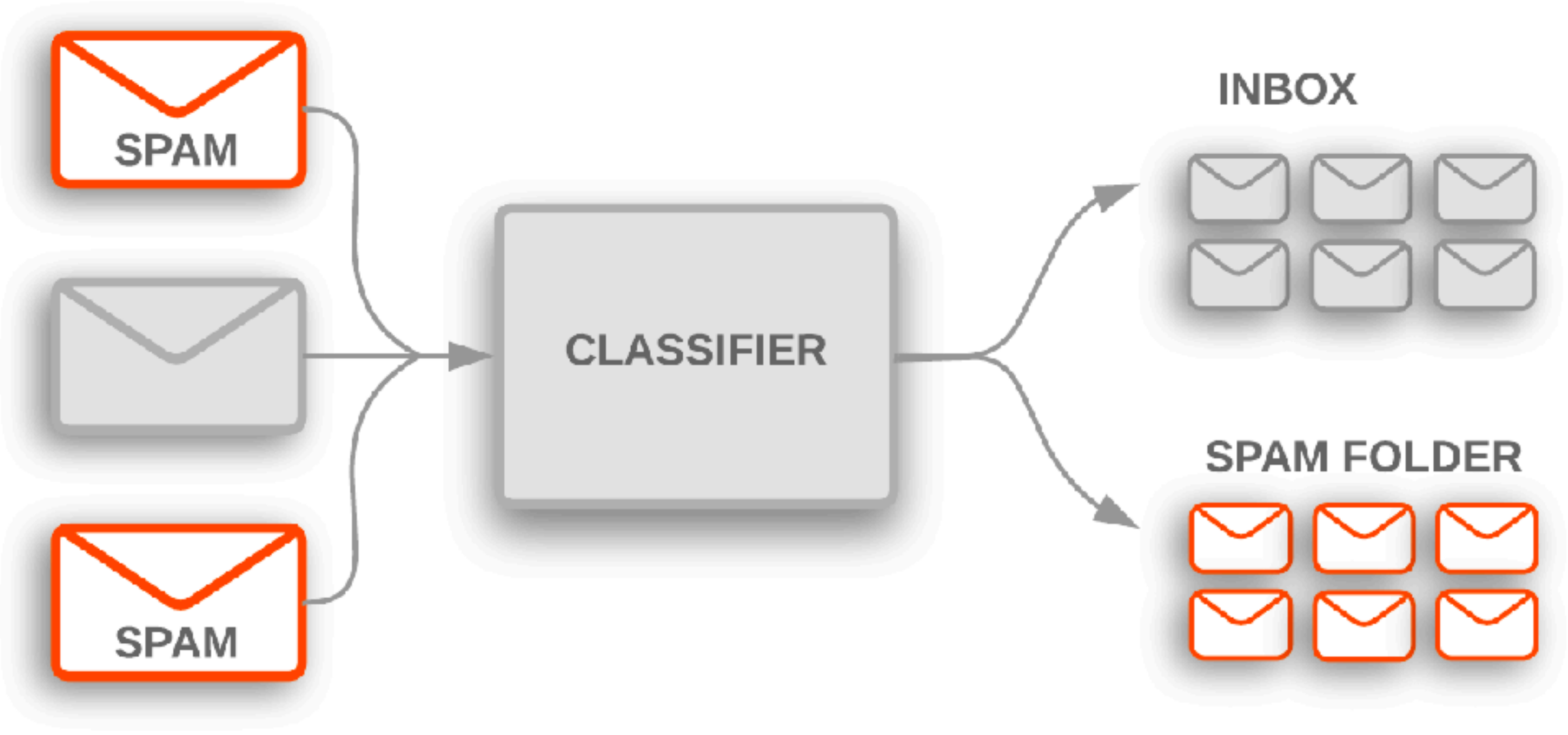
Study of algorithms that:

- improve their **performance** P
- at some **task** T
- with **experience** E

Well-defined learning task: <P,T,E>

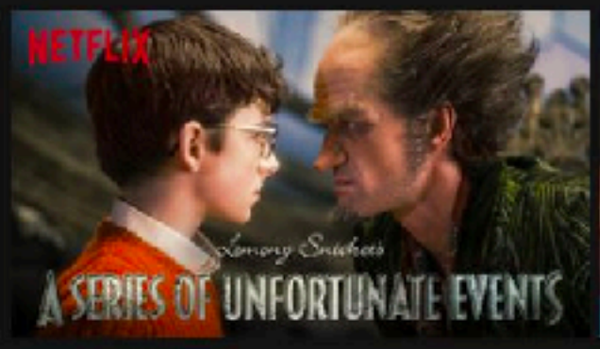
— *Tom Michell, 1997*

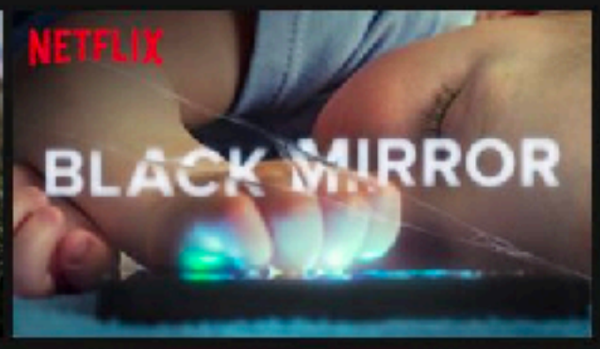
**Machine learning** is the **“best” solution** for tasks such as:


SPAM
SPAM
CLASSIFIER
INBOX
SPAM FOLDER

NETFLIX Browse DVD

## Because you watched Stranger Things

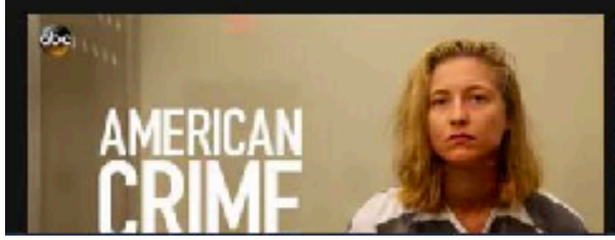


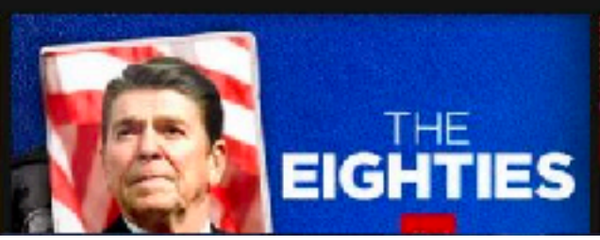
## Because you watched The Crown

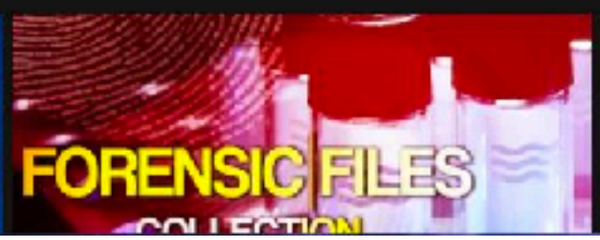


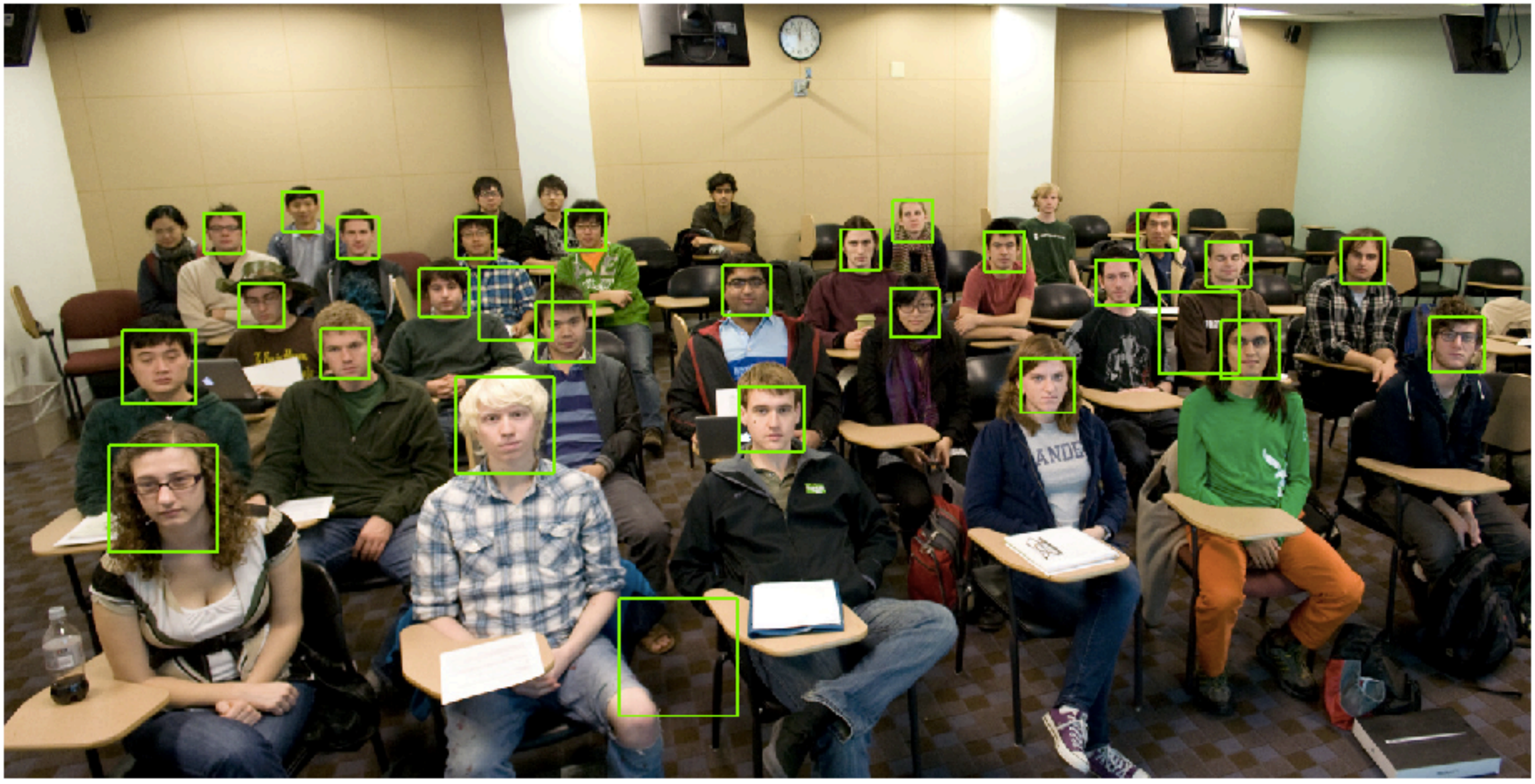


## Because you watched American Crime Story: The People v. O.J. Simpson



Terminal check

*Correct Pin?*
*Sufficient Balance ?*
*Blocked Card?*

Rejected

Predictive model

Fraud score

Investigators

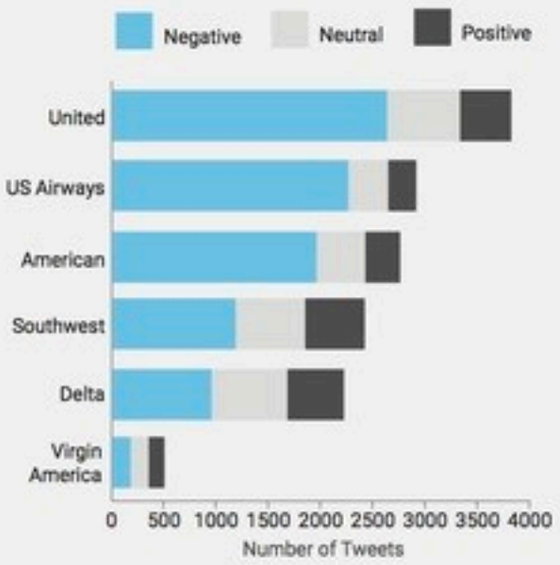

Feedback

# U.S. Airline Twitter Sentiment

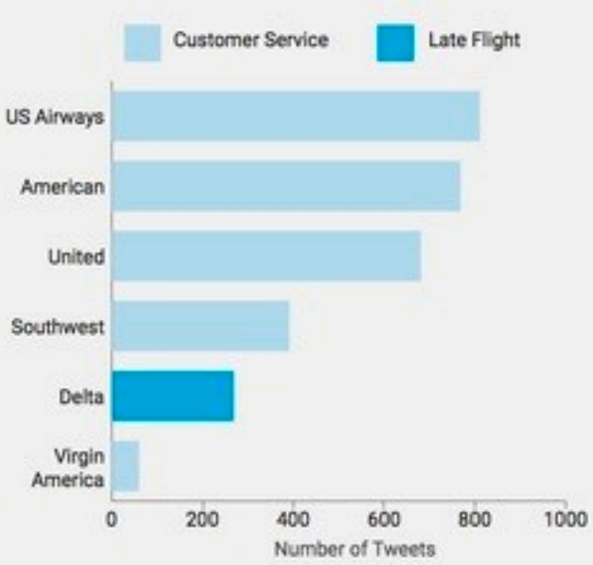
## Analysis of traveler's Tweets from a week in February 2015

Cathy Liewen, Heidi Slojewski
HCI 512 | Winter 2016

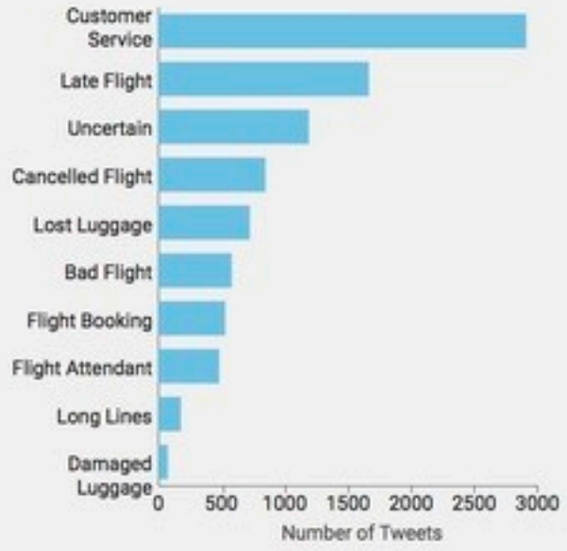
### Sentiment by Airline




### Airlines' Top Reasons for Negative Sentiment




### Most Common Reasons for Negative Sentiment

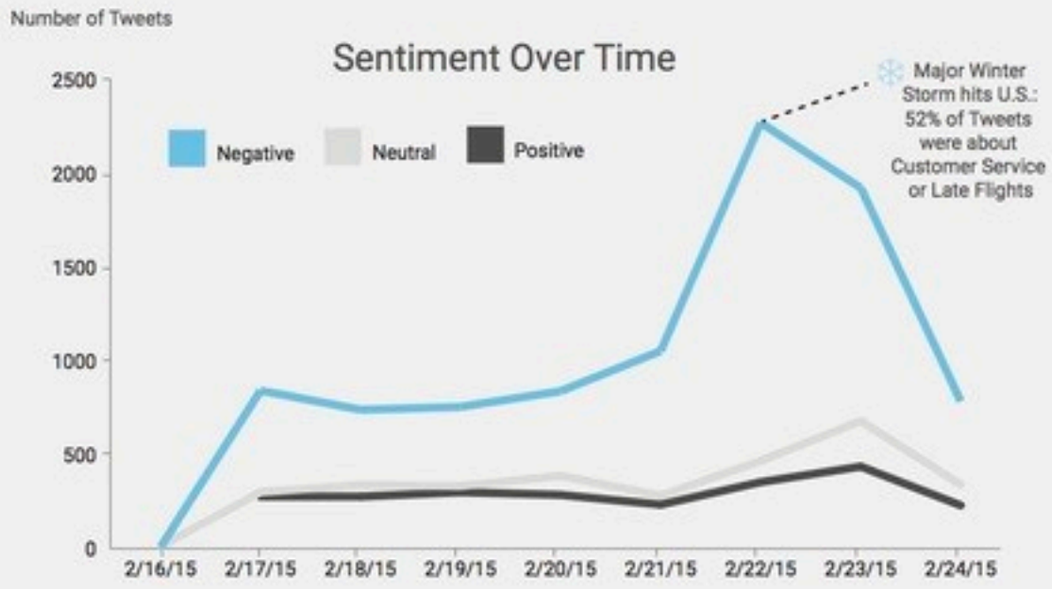



### Most Retweeted Negative Sentiments

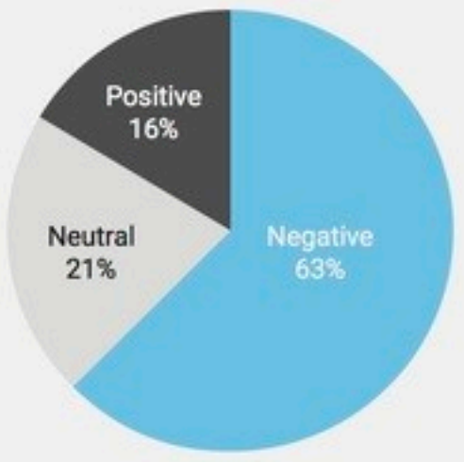
"@USAirways 5 hr flight delay and a delay when we land . Is that even real life ? Get me off this plane , I wanna go home"
-OBJ_3

"@USAirways of course never again tho. Thanks for tweetin ur concern but not Doin anythin to fix what happened. I'll choose wiser next time"
-OBJ_3

### Sentiment Over Time




### Sentiment Breakdown




### Most Common Words from Negative Tweets

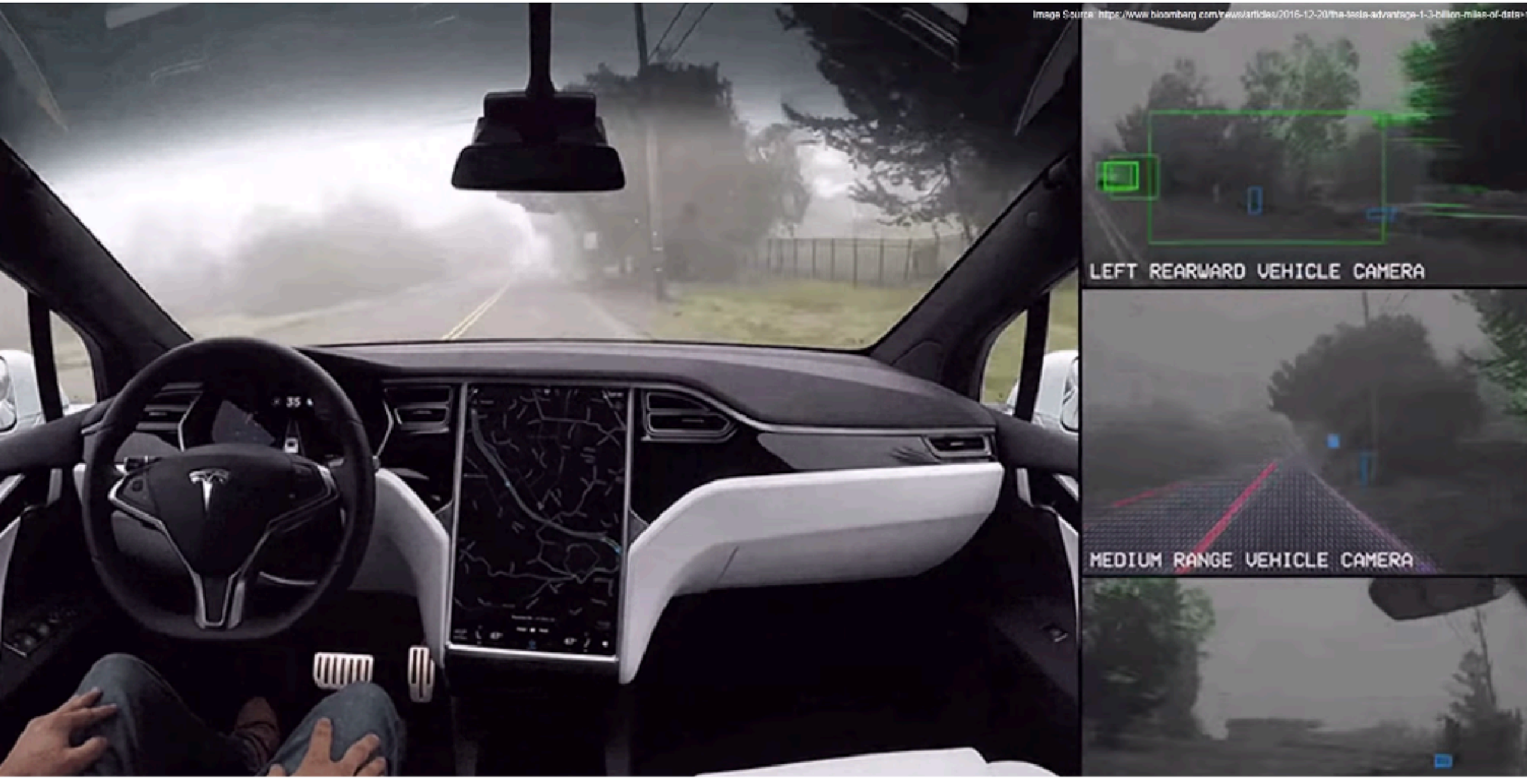
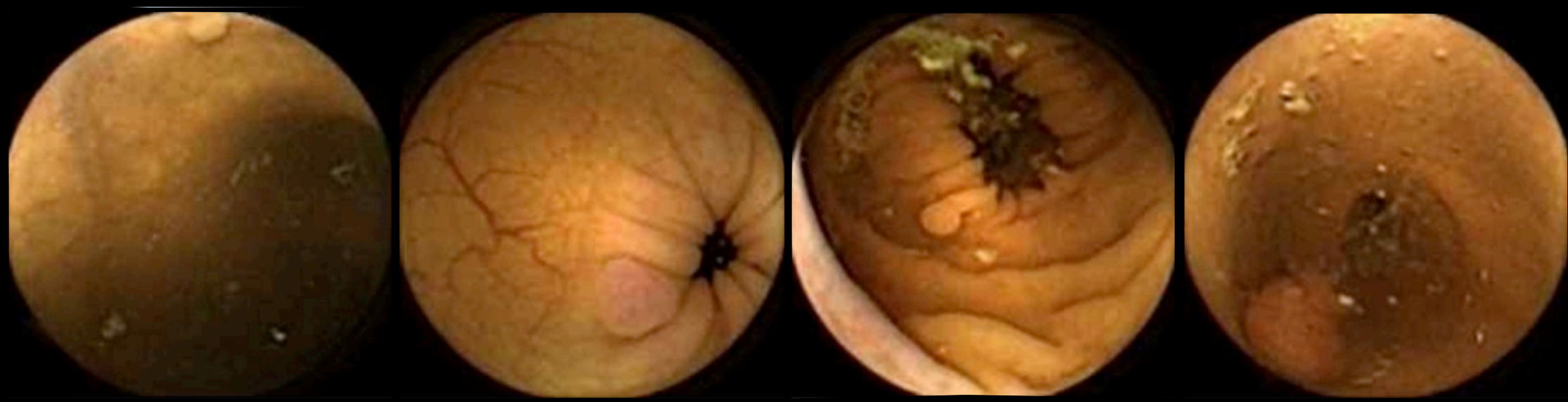
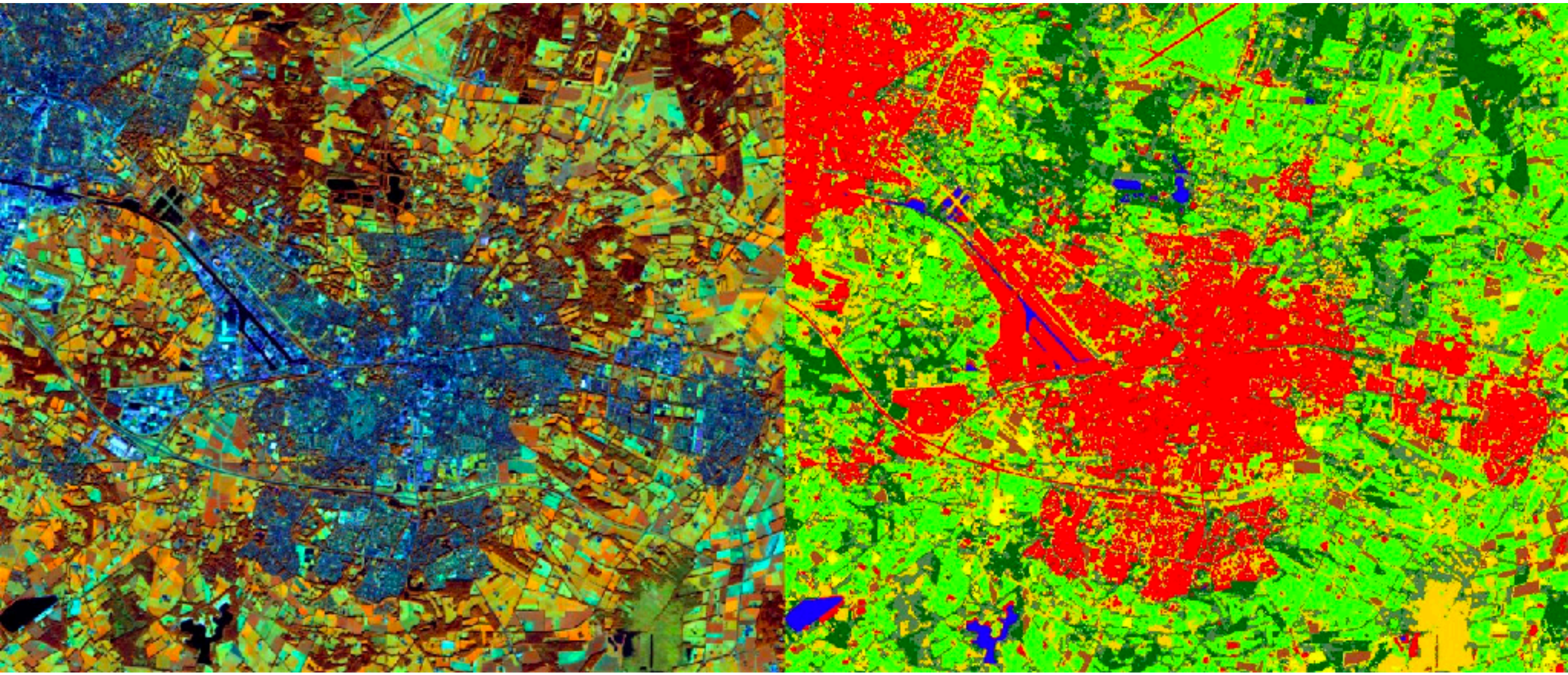


Image Source: https://www.bloomberg.com/news/articles/2016-12-20/the-tesla-advantage-1-3-billion-miles-of-data
LEFT REARWARD VEHICLE CAMERA
MEDIUM RANGE VEHICLE CAMERA

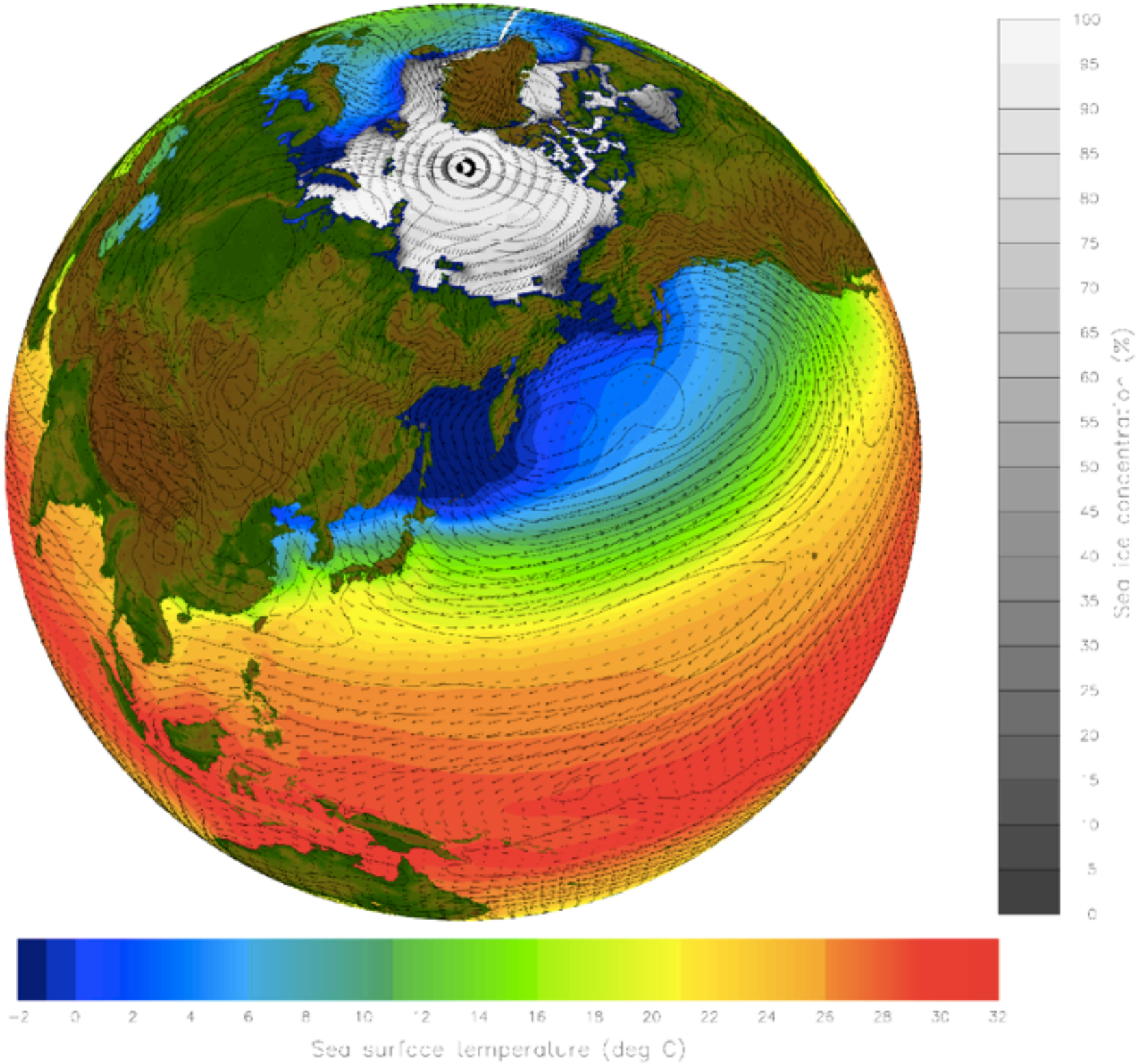


UNIVERSITAT DE BARCELONA

DATA SCIENCE @ UNIVERSITAT DE BARCELONA

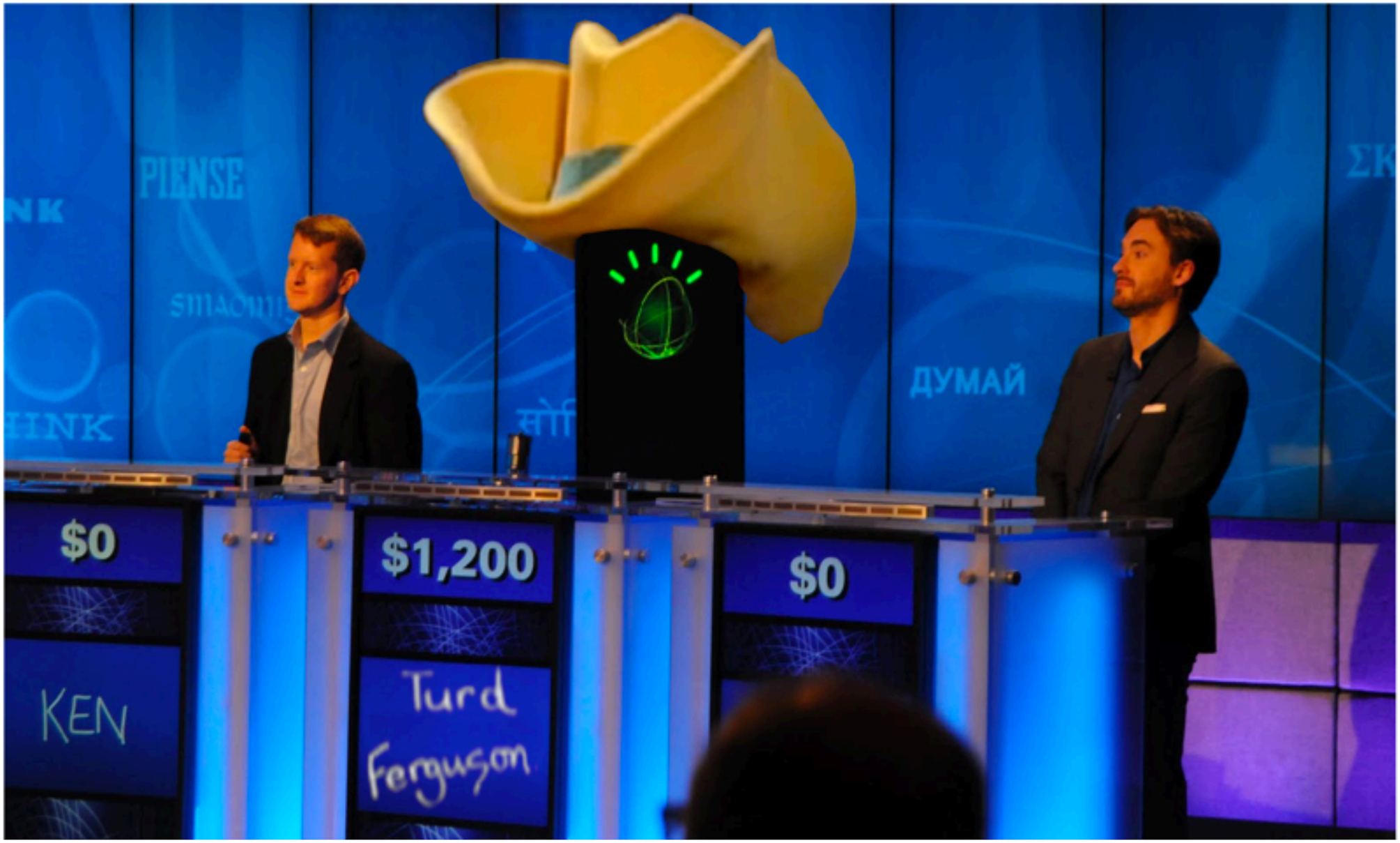


# Climate change




This call may be recorded to improve translations
skype
Hi grandma. Happy birthday. Are you having a great day?
Hola abuela. Feliz cumpleaños. ¿Tienes un buen día?
Hi Dylan. Yes, a great day thanks. And the flowers that you have sent are precious.
Hola Dylan. Sí, un día estupendo gracias. Y las flores que has enviado son preciosas.
I'm glad they arrived.
Me alegra que llegaron.
They arrived first thing in the morning.
Llegaron a primera hora de la mañana.
Type a message here

How IBM build Watson, its Jeopardy-playing supercomputer (2011).

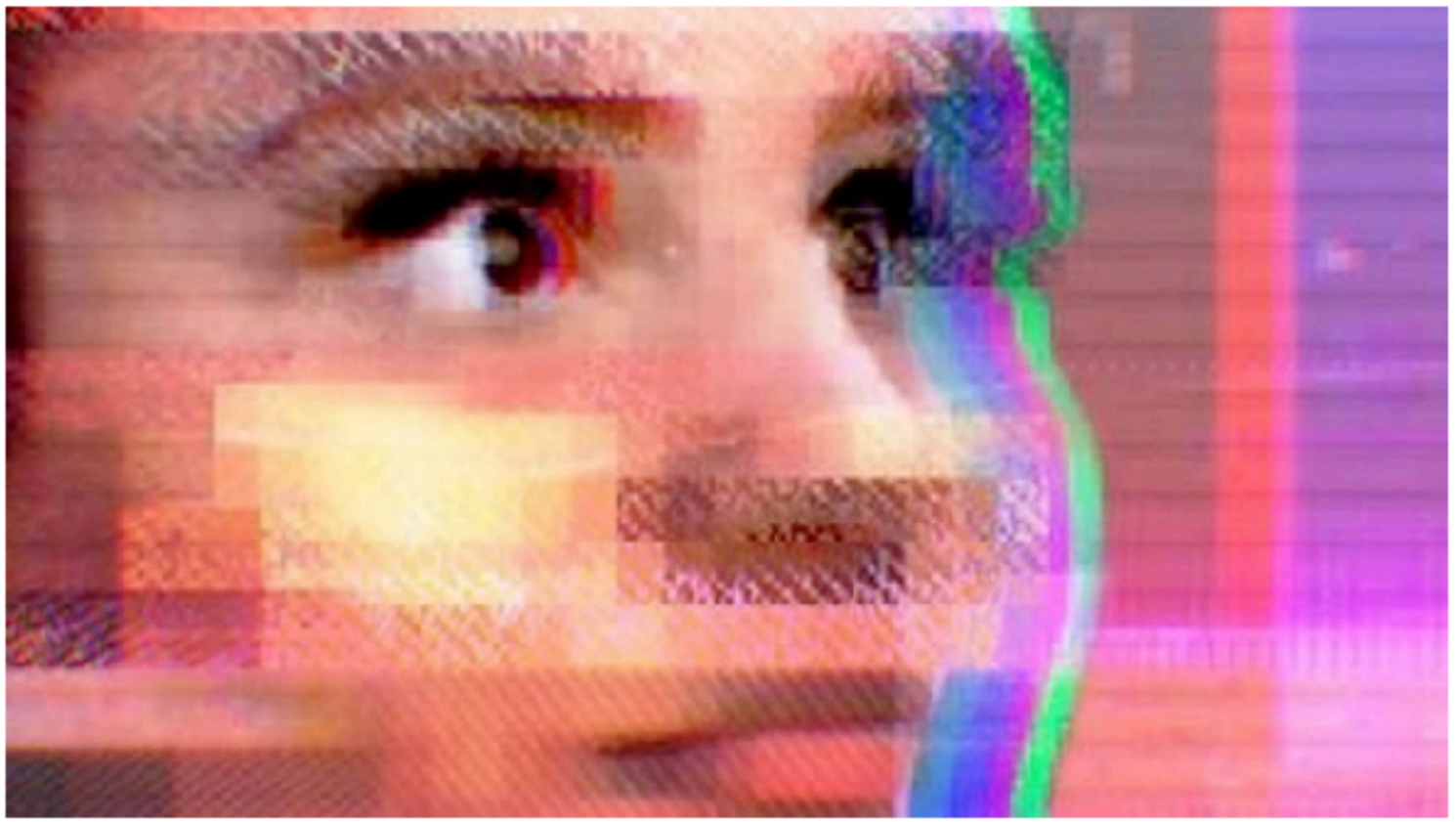


**Machine learning is the “best” solution for tasks such as:**

- Voice Recognition
- Product recommendation
- Search engines
- Detection of fraudulent cards
- Biometric identification
- Character recognition
- Face recognition
- Medical diagnosis of some pathologies
- …

REDES SOCIALES • Microsoft ha pedido disculpas y la ha desactivado

# Una inteligencia artificial se vuelve racista, antisemita y homófoba en menos de un día en Twitter

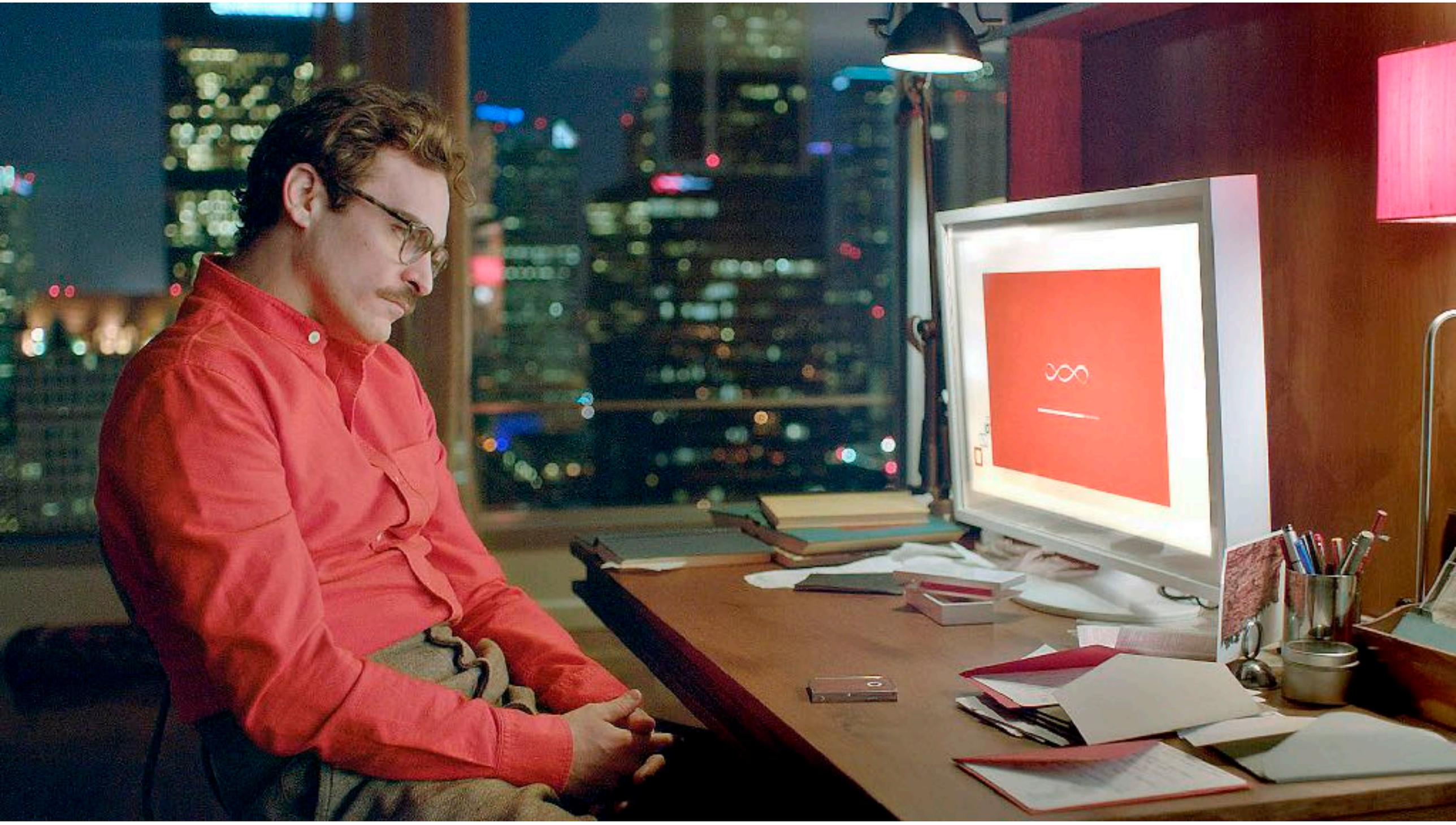


Compartido 3

Comentar noticia

· En algunos de sus 'tweets', dijo que Hitler tenía razón. También deseó que las feministas ardieran en el infierno.

# Science Fiction or Future?



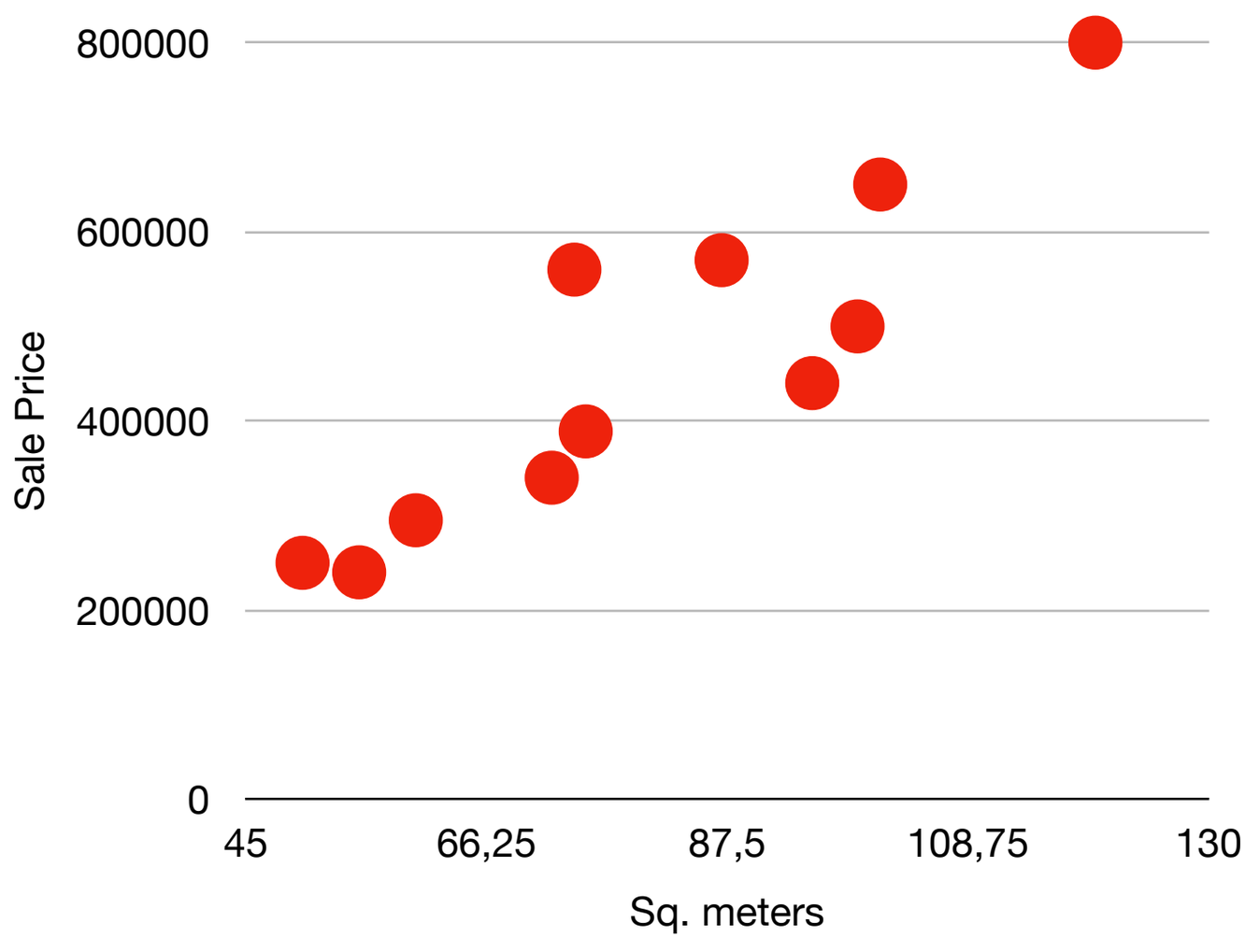
UNIVERSITAT DE BARCELONA

$$\hat{Y} = bX + a$$

# Example of Machine Learning

**Task: Predict sale price**

| Square meters | Sale Price |
|---|---|
| 50 | 250.000 |
| 75 | 389.000 |
| 72 | 340.000 |
| 60 | 295.000 |
| 95 | 440.000 |
| 55 | 240.000 |
| 120 | 800.000 |
| 87 | 570.000 |

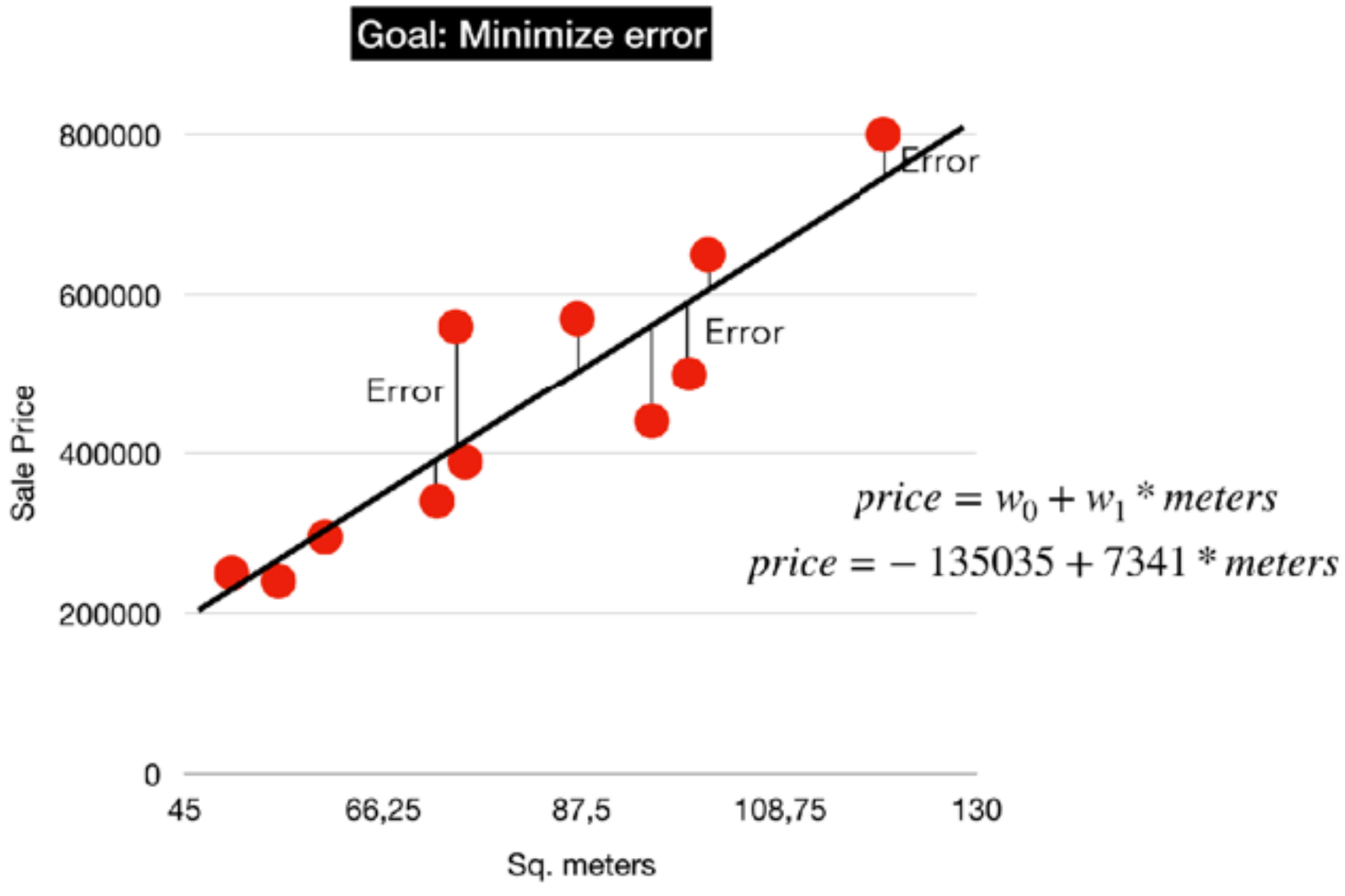


# Example of Machine Learning

| Sq. meters | Sale Price |
|---|---|
| 50 | 250.000 |
| 75 | 389.000 |
| 72 | 340.000 |
| 60 | 295.000 |
| 95 | 440.000 |
| 55 | 240.000 |
| 120 | 800.000 |
| 87 | 570.000 |

# Example of Machine Learning

| Sq. meters | Sale Price | Prediction |
|---|---|---|
| 50 | 250.000 | 232.015 |
| 75 | 389.000 | 415.540 |
| 72 | 340.000 | 393.517 |
| 60 | 295.000 | 305.425 |
| 95 | 440.000 | 562.360 |
| 55 | 240.000 | 268.720 |
| 120 | 800.000 | 745.885 |
| 87 | 570.000 | 503.632 |

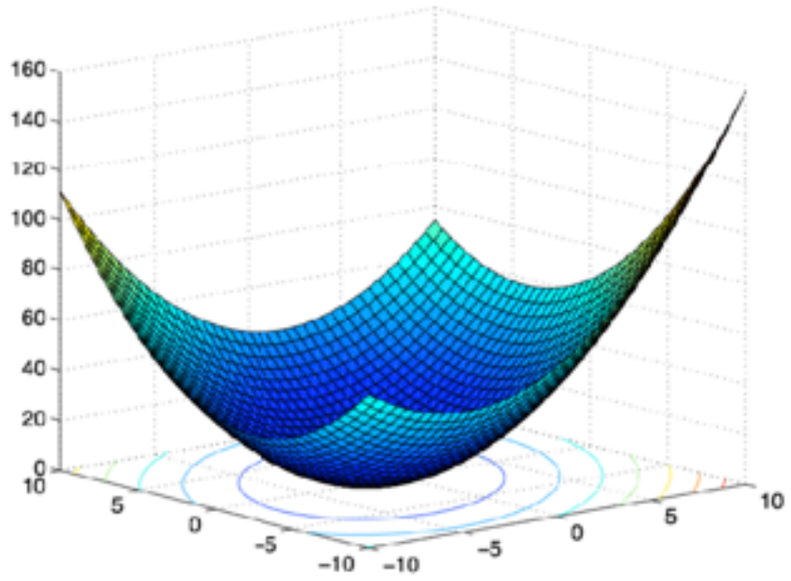


# Example of Machine Learning

We will have to define a cost function, as for instance:

$$cost = \frac{\sum_i^N (y_i - \hat{y}_i)^2}{N}$$

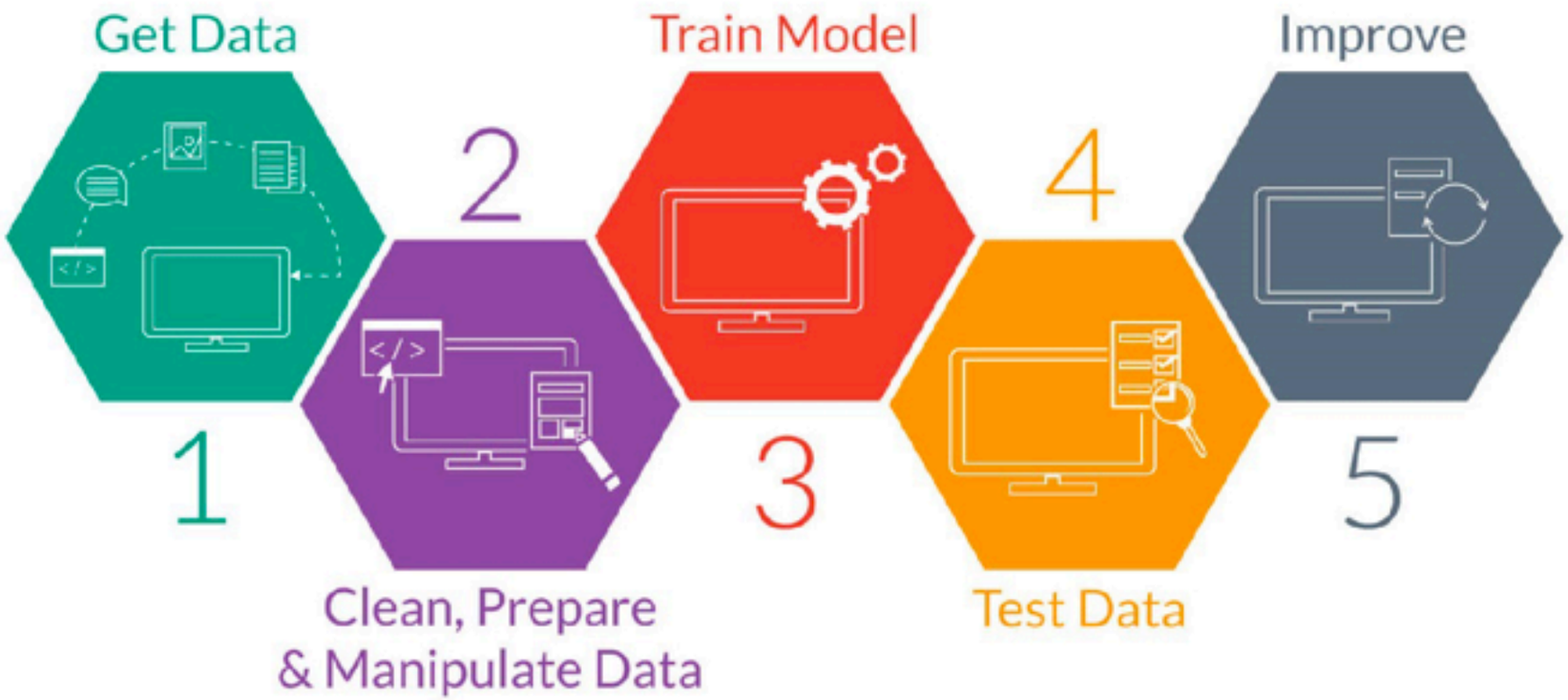
and minimize it using the training data



# Type of Machine Learning

| Unsupervised Learning | Supervised Learning | Reinforcement Learning |
|---|---|---|
| Clustering | Classification<br>Regression | Learn from mistakes |

## The core steps of typical machine learning workflow



BIG DATA

Fat Data

Dirty Data

Data

Science

Data Mining

Clustering

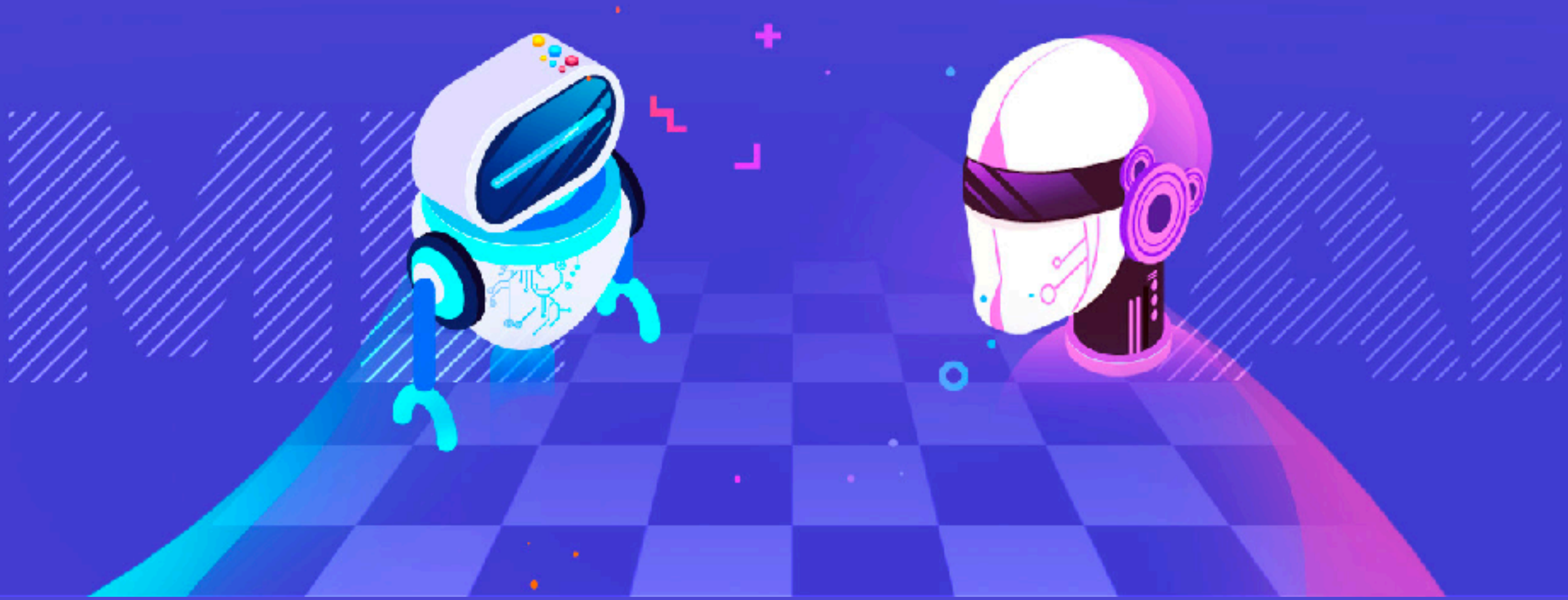
Artificial Intellingence

# Machine Learning

Deep Learning

Reinforcement Learning

# Machine Learning vs Artificial Intelligence



# Machine Learning vs Artificial Intelligence

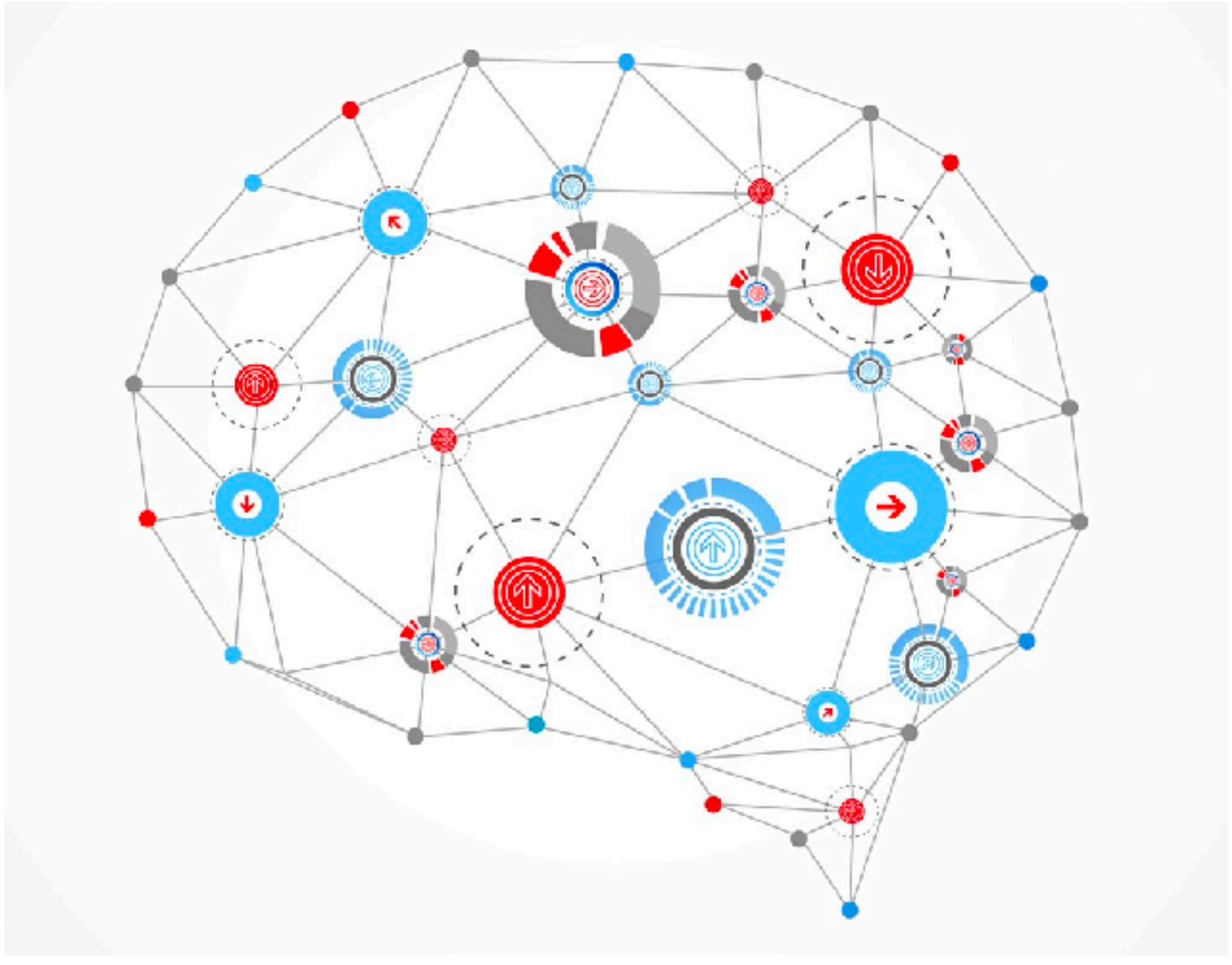
**Artificial Intelligence** is an academic discipline devoted to the theory and development of computer systems able to perform tasks normaly requiring human intelligence, such as visual perception, language recognition, decision-making, planning, reasoning, etc.

Artificial Intelligence is classified into two parts, **General AI** and **Narrow AI**. General AI refers to making intelligent in a wide array of activities that involve thinking and reasoning. Narrow AI, on the other hand, involves the use of artificial intelligence for a very specific task.

**Machine learning** is a subset of artifi cial intelligence that uses algorithms to learn from data (inductive behavior).

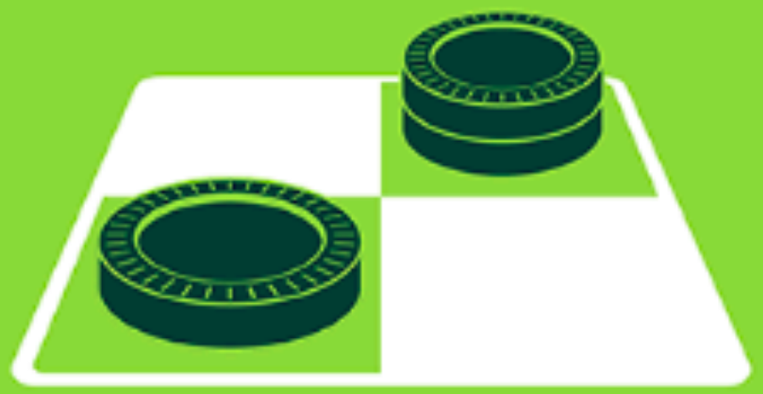
"**Machine learning** is a subset of **artificial intelligence** that uses algorithms to **learn from data** and enables machines to improve with **experience**"

# Deep learning (DL) is ML that uses a particular class of algorithms (neural networks)



# ARTIFICIAL INTELLIGENCE

**Early artificial intelligence stirs excitement.**



# MACHINE LEARNING

**Machine learning begins to flourish.**

# DEEP LEARNING

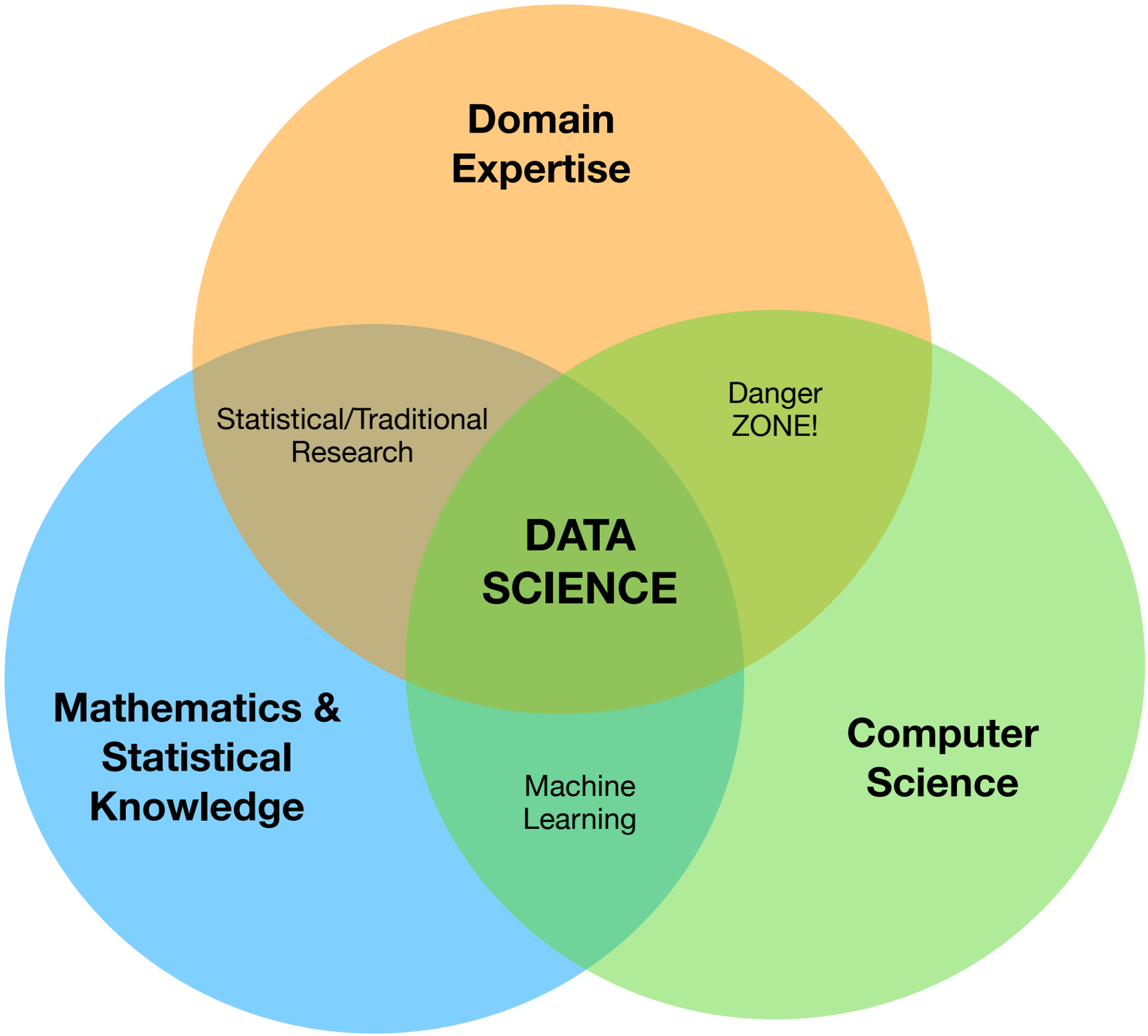
**Deep learning breakthroughs drive AI boom.**

1950's 1960's 1970's 1980's 1990's 2000's 2010's

# Data Science

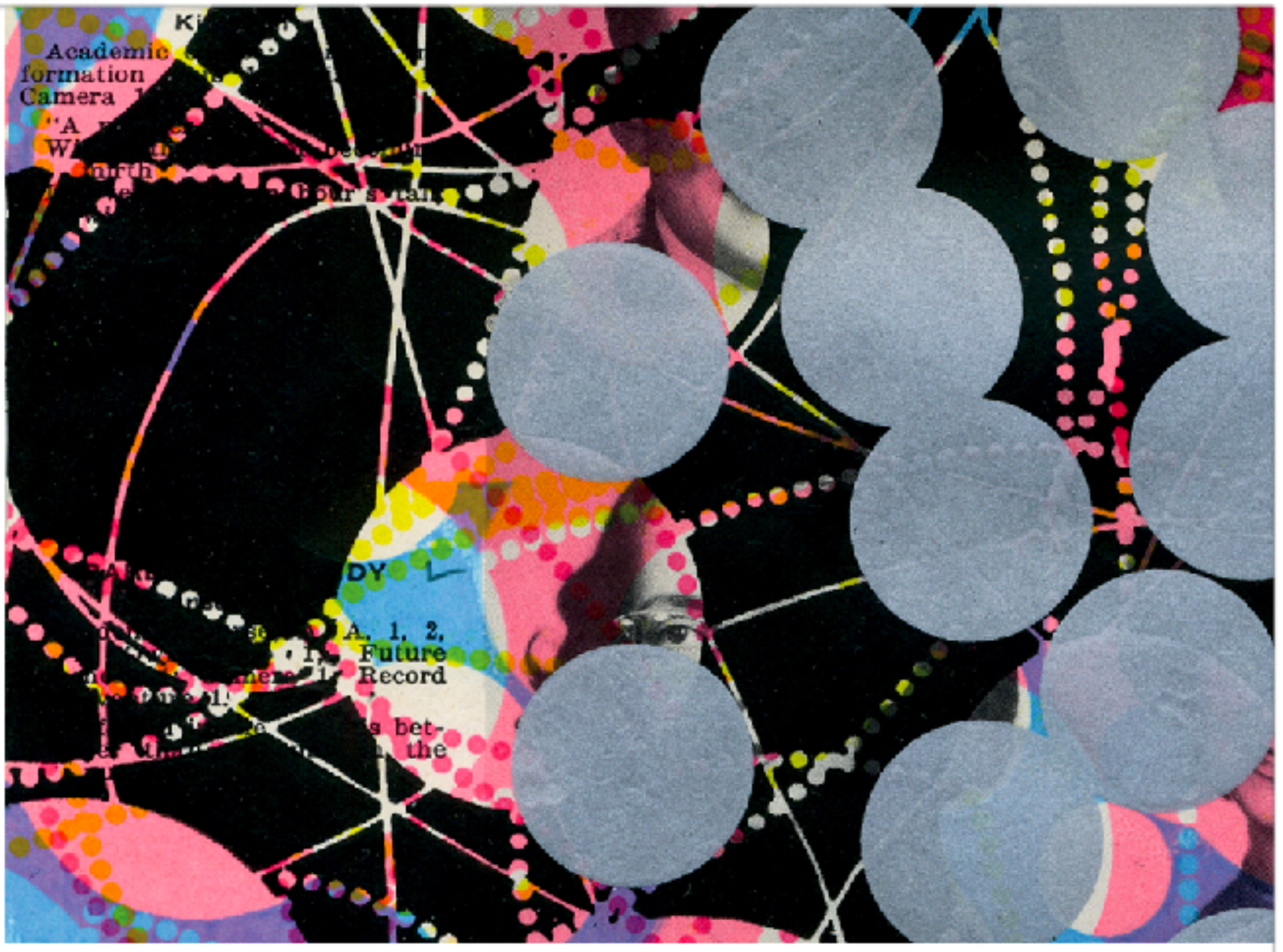
is a **multidiciplinary methodology**

to help to define what we want to do with data, how to evaluate our alogorithms, what decisions can be grounded on dada, how do we combine evidences from several sources, etc..




**Drew Conway's Data Science Venn Diagram**

Harvard
Business
Review




DATA

# Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

# THE DATA SCIENCE **HIERARCHY OF NEEDS**

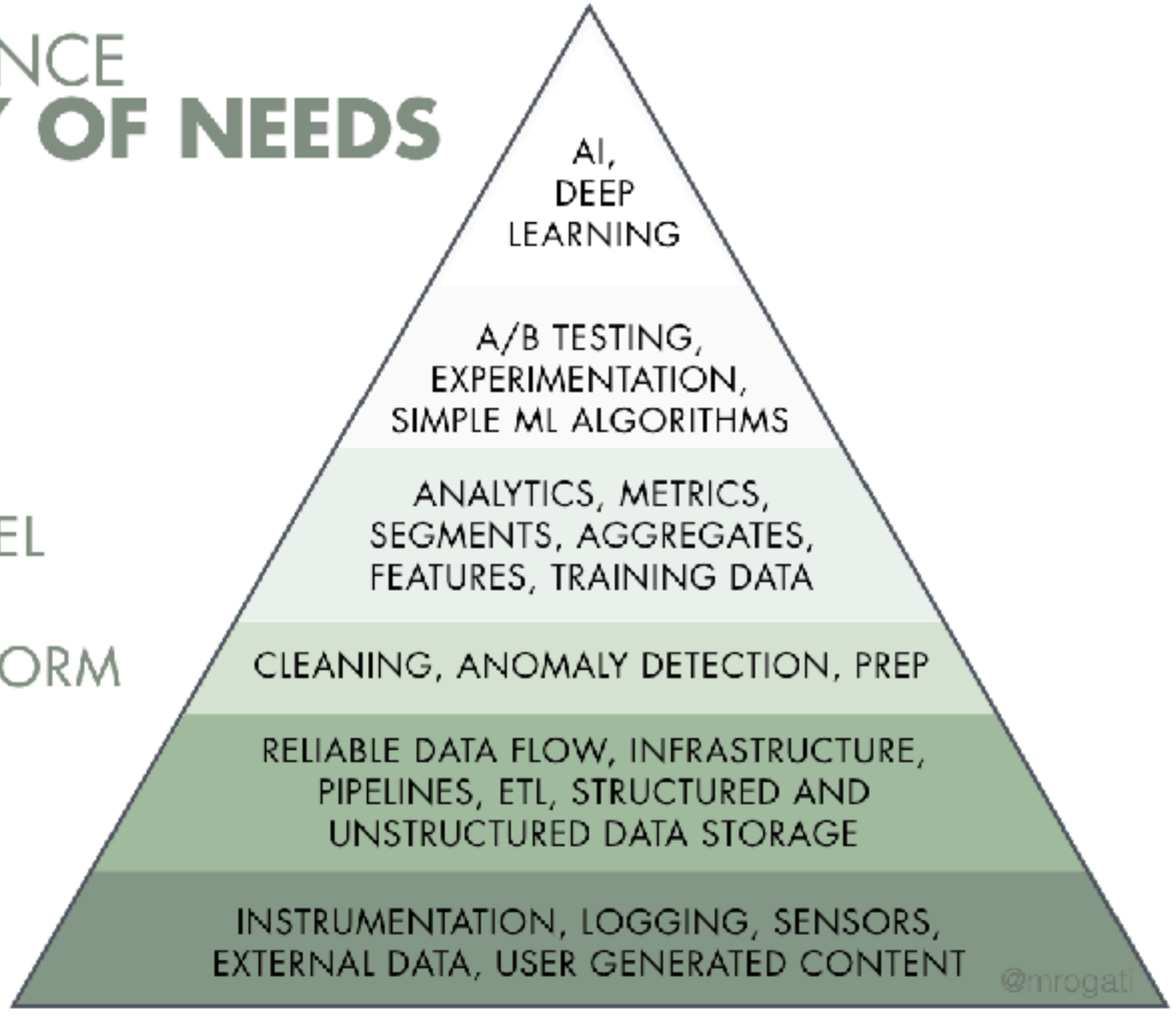
LEARN/OPTIMIZE

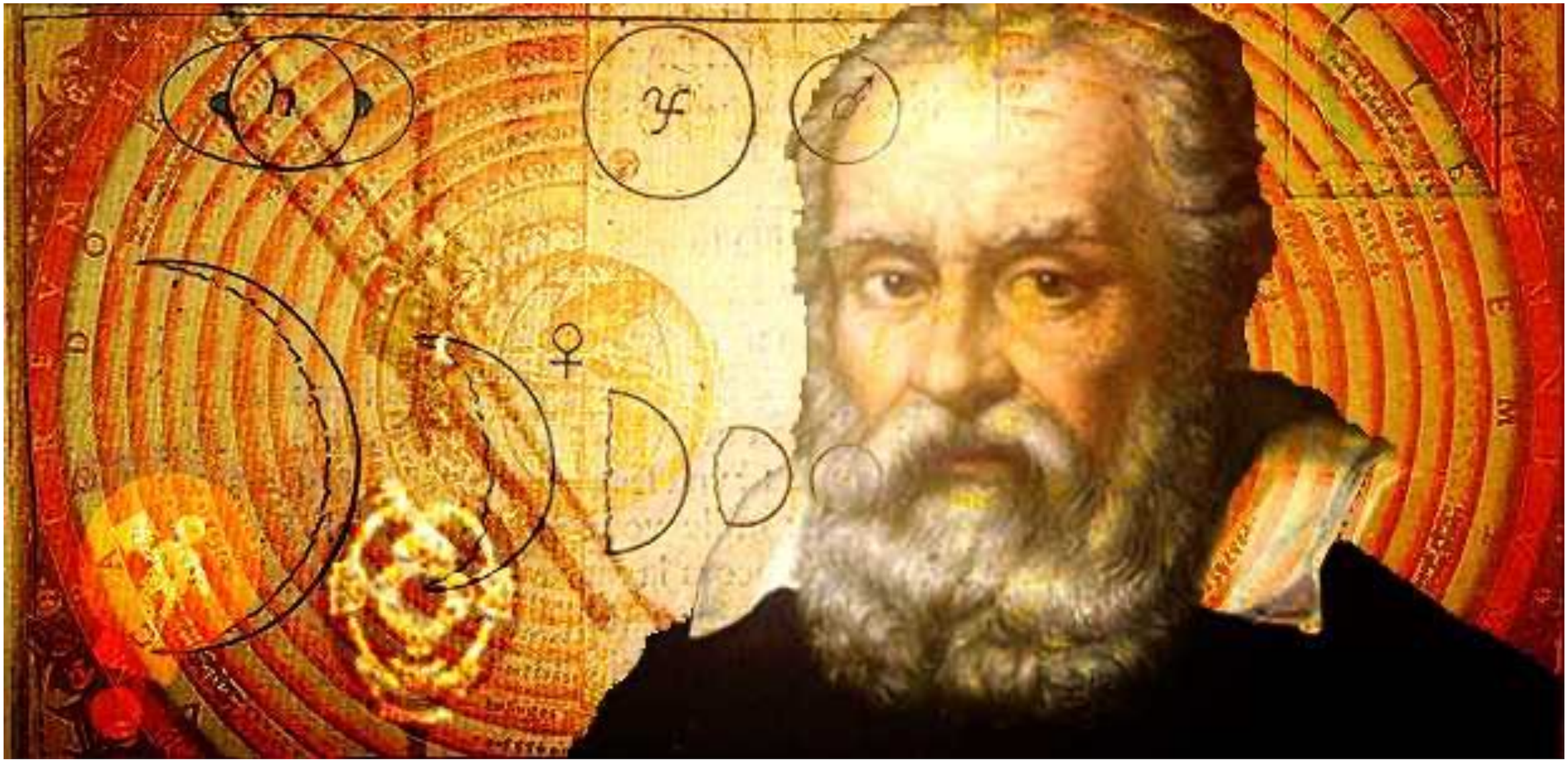
AGGREGATE/LABEL

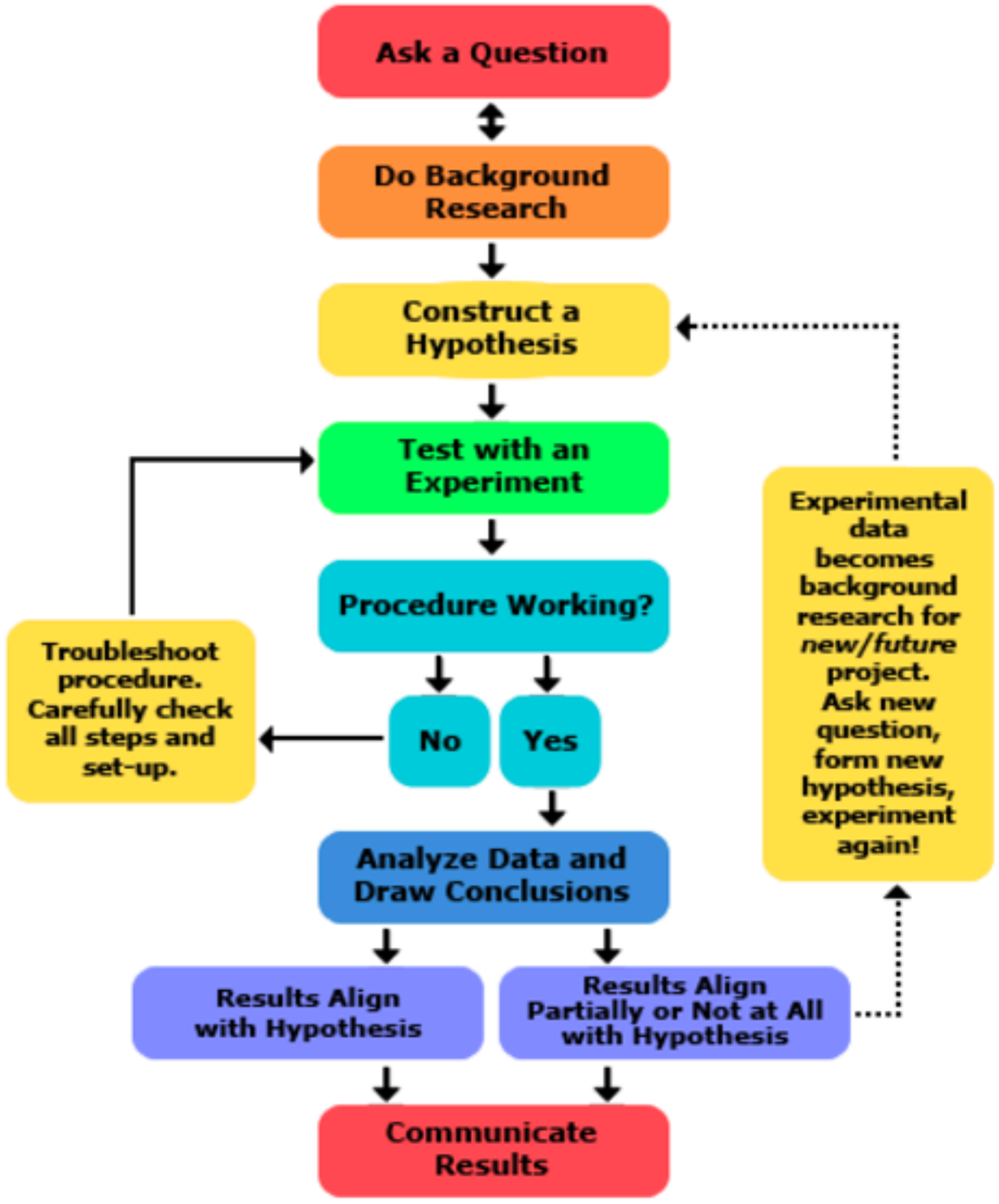
EXPLORE/TRANSFORM

MOVE/STORE

COLLECT






Ask a Question
Do Background Research
Construct a Hypothesis
Test with an Experiment
Procedure Working?
No
Yes
Troubleshoot procedure. Carefully check all steps and set-up.
Analyze Data and Draw Conclusions
Results Align with Hypothesis
Results Align Partially or Not at All with Hypothesis
Communicate Results
Experimental data becomes background research for new/future project. Ask new question, form new hypothesis, experiment again!

# Data Science Path

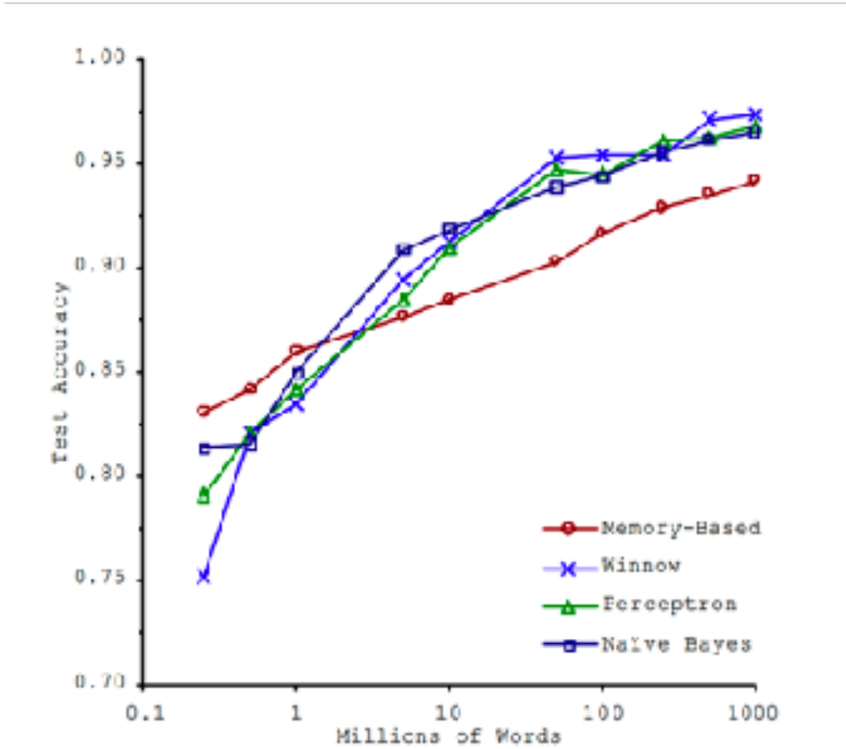
| | |
|---|---|
| What do I want? Does it have sense? | **Question** |
| What are my data sources? How reliable are they? | **Acquire** |
| How do I develop an understanding of the content of my data? | **Describe** |
| What are the key relationships in my data? | **Discover** |
| How do I develop an understanding of the content of my data? | **Analyze** |
| What are the likely future outcomes? | **Predict** |
| Are my expectations fulfilled? | **Evaluate** |

# Main Challenges of Machine learning

Insufficient Quantity of Data
Nonrepresentative Training Data
Poor-Quality Data
Irrelevant Features
Overfitting the Training Data
Underfitting the Training Data

# Insufficient Quantity of Data

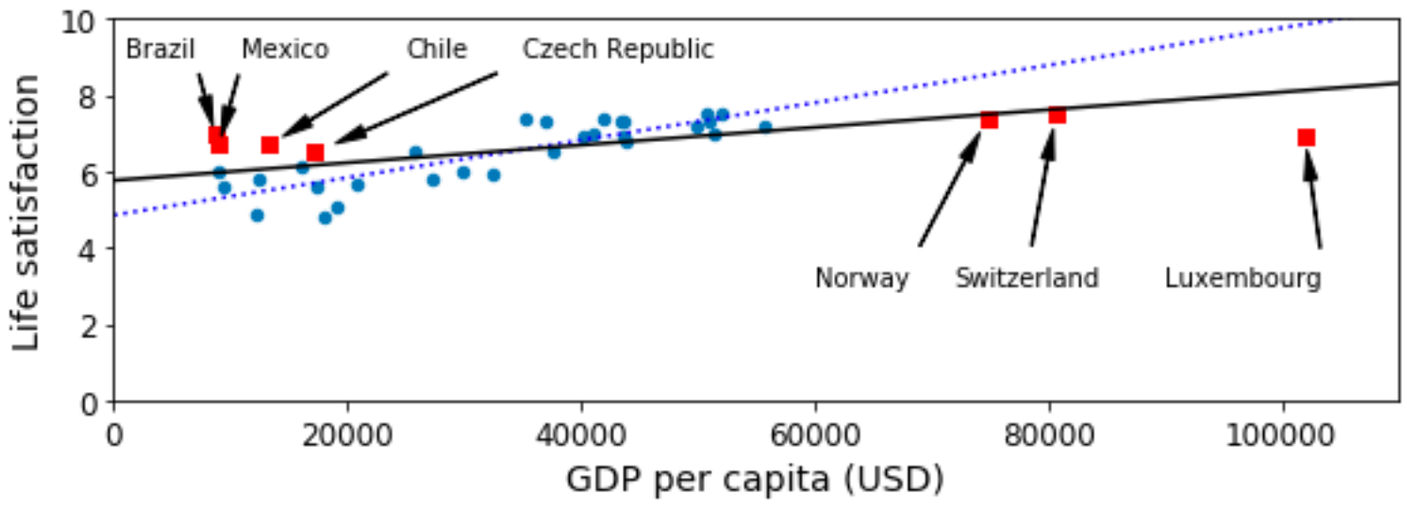
- The importance of data versus algorithms
- What is best? more data or a better model?
    - Several studies shows that very different algorithmes, including fairly simples ones, perform almost identically when enugh data is provided




- Peter Norvig in his paper titled “The Unreasonable Effectiveness of data”  popularized the idea that data matters more than algorithms.
    - However, small- and medium -sized datasets are still very common. In many cases data is really expensive

# Nonrepresentative 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.




**Careful about the sampling BIAS**

# Poor-Quality of data

- Our training data can be full of errors, outliers, and noise.
    - If there are outliers in the training set perhaps we should simply discard them.
    - Missing features from some instances. **What we can do?**
        - Remove those instances
        - Remove those features
        - Impute those features to those instances

# Irrelevant Features

- Your system will be only capable to learn if the training data contains **enough relevant** features and **not too many irrelevant ones.**

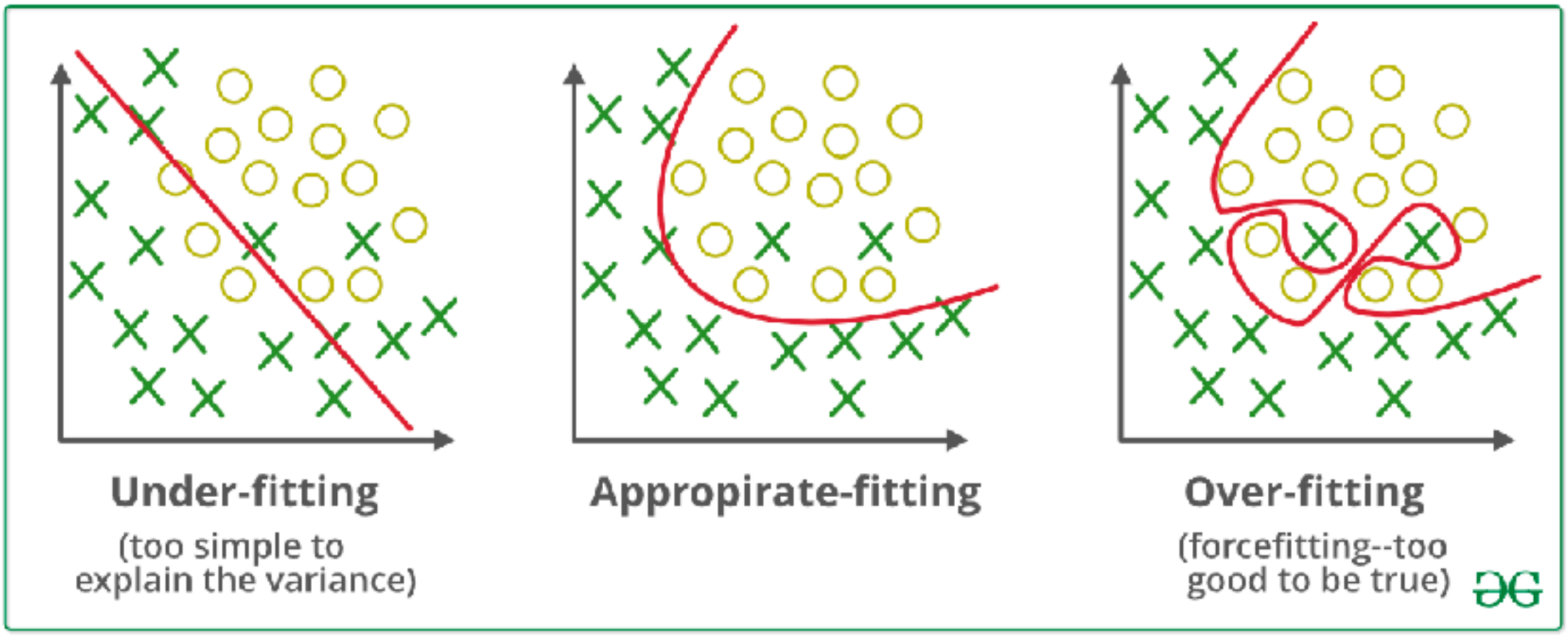
- The process called *feature* engineering aims to come up with a good set of features to train with. The process involves the following steps:

  - Feature Selection

  - Feature Extraction

  - Creating new features by gathering new data

# **Overfitting/Underfitting** the Training Data



**Next Session**

https://github.com/ssegui/ml-ub/tree/master/notebooks/Session1.ipynb

**First Project:**

https://www.kaggle.com/t/afe495cd9e90462baa84b5ce320791dd

**Score:**

1 point + 0.25 for the winner