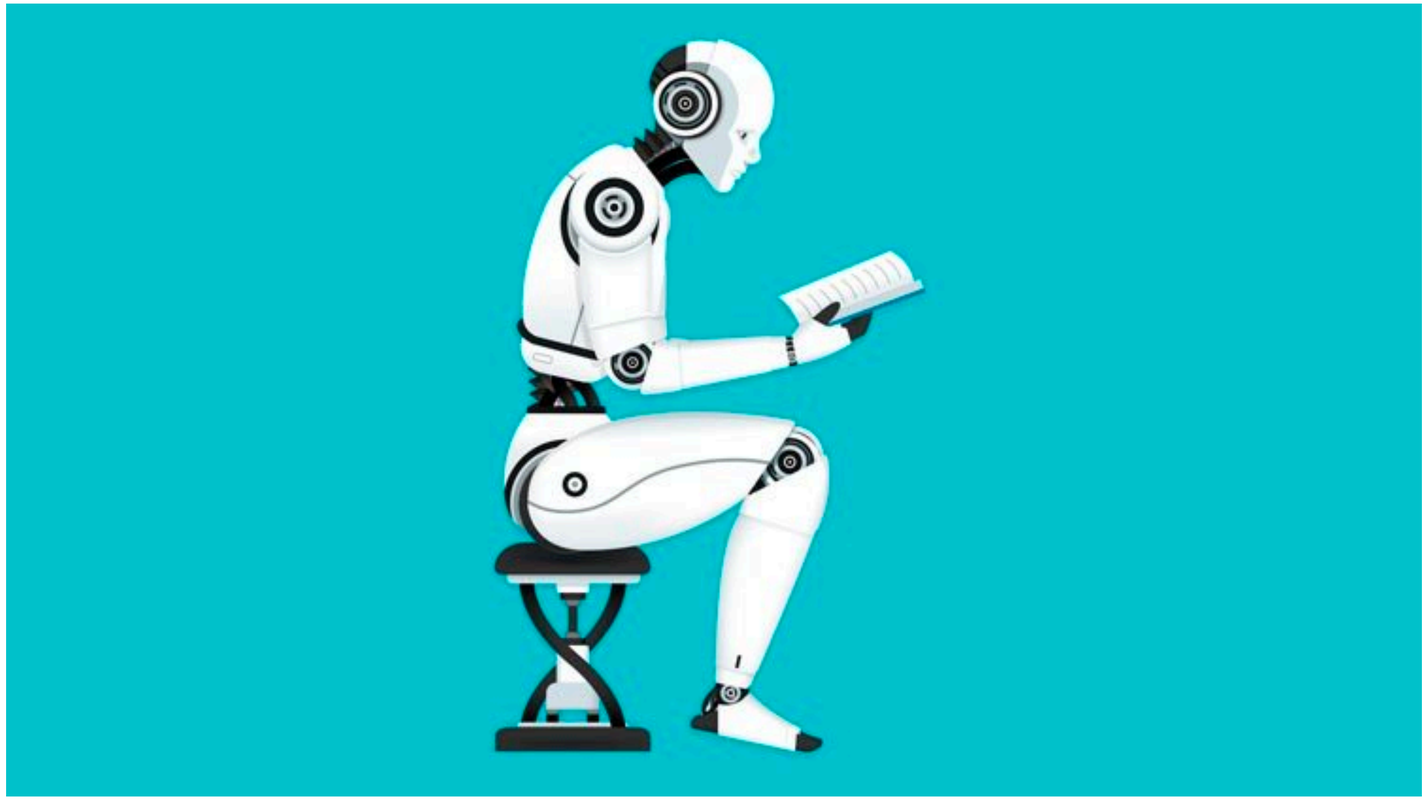
UNIVERSITAT DE BARCELONA



Presentació del Curs

Machine Learning | Enginyeria Informàtica

Santi Seguí | 2022-2023

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 (Dijous 17.00–19.00h)

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

Laboratoris (Dimarts 15h-17h)

- **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 el:
  - Campus Virtual. A través d'aquest entorn tindreu: anuncis, apunts, notes, fòrum, calendari, enllaços a la bibliografia, etc.
    - https://campusvirtual.ub.edu/
  - Github del curs:
    https://github.com/ssegui/ml_ub

• Com seran les classes teòriques? (2 hora a la setmana)

- Introduirem el conceptes teòric

• Com seran les classes pràctiques? (2 hores a la setmana)

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

Llenguatge de programació?

- Python & R

Programació de l'oferta docent del Primer semestre

Activitat

Grup	Dies	Horari	Professorat	Aula	Llengua
Teoria [Presencial]					
T1	dl. dt. dc. dj. dv. 1r sem.	17.00-19.00	Segui Mesquida, Santiago	Aula B7	Català
Pràctiques de laboratori d'ordinadors [Presencial]					
a00	dl. dt. dc. dj. dv. 1r sem.	15.00-17.00	Fortiana Gregori, Jose	Aula IG	Català
Exàmens : 1r parcial [Presencial]					
G1	8 de novembre de 2022.	18.00-21.00	Fortiana Gregori, Jose Segui Mesquida, Santiago		-
Exàmens : Final [Presencial]					
G1	12 de gener de 2023.	15.00-20.00	Fortiana Gregori, Jose Segui Mesquida, Santiago		-
Exàmens : Reavaluació [Presencial]					
G1	27 de gener de 2023.	15.00-20.00	Fortiana Gregori, Jose Segui Mesquida, Santiago		-

	Laboratori (Dimarts - 13h-15h)		Teoria (Dijous (17h - 19h)
13-Set	*No Class*	15-Set	Introduction
20-Set	Lab0	22-Set	A typical Machine Learning project
27-Set	Lab1	29-Set	Regression
4-Oct	Lab2	6-Oct	Classification
11-Oct	Lab3	13-Oct	Training Models
18-Oct	Lab4	20-Oct	Support Vector Machines
25-Oct	Lab5	27-Oct	Tree Based Models
1-Nov	*FESTIU*	3-Nov	Boosting - Bagging - Ensembles
8-Nov	*Exam - NO EXAM!!!*	10-Nov	*Exam - NO EXAM!!!*
15-Nov	Lab7	17-Nov	Neural Networks
22-Nov	Lab8	24-Nov	Convolutional Neural Networks
29-Nov	Lab9	1-Dec	Convolutional Neural Networks
6-Dec	*Festiu*	8-Dec	*Festiu*
13-Dec	Lab10	15-Dec	Unsupervised Learning
19-Dec	Lab11	22-Dec	*Free Class*

4. Avaluació

Avaluació basada en projectes

Avaluació Continuada
Basada en Projectes

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

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

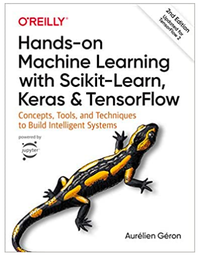
Programari

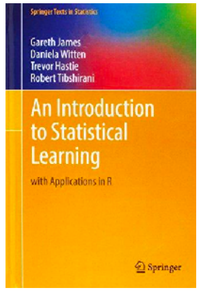
- Python & R

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/






Temas
Para jugar este cuestionario
1. Usa cualquier dispositivo para abrir
joinmyquiz.com
2. Ingrese el código de ingreso
166884
o compartir mediante...
INICIAR
JURASSIC SCHOOL BUS
uego copiado al portapapeles!

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 field of study that gives computers the ability to learn without being explicitly programmed.*

— Arthur Samuel, 1959

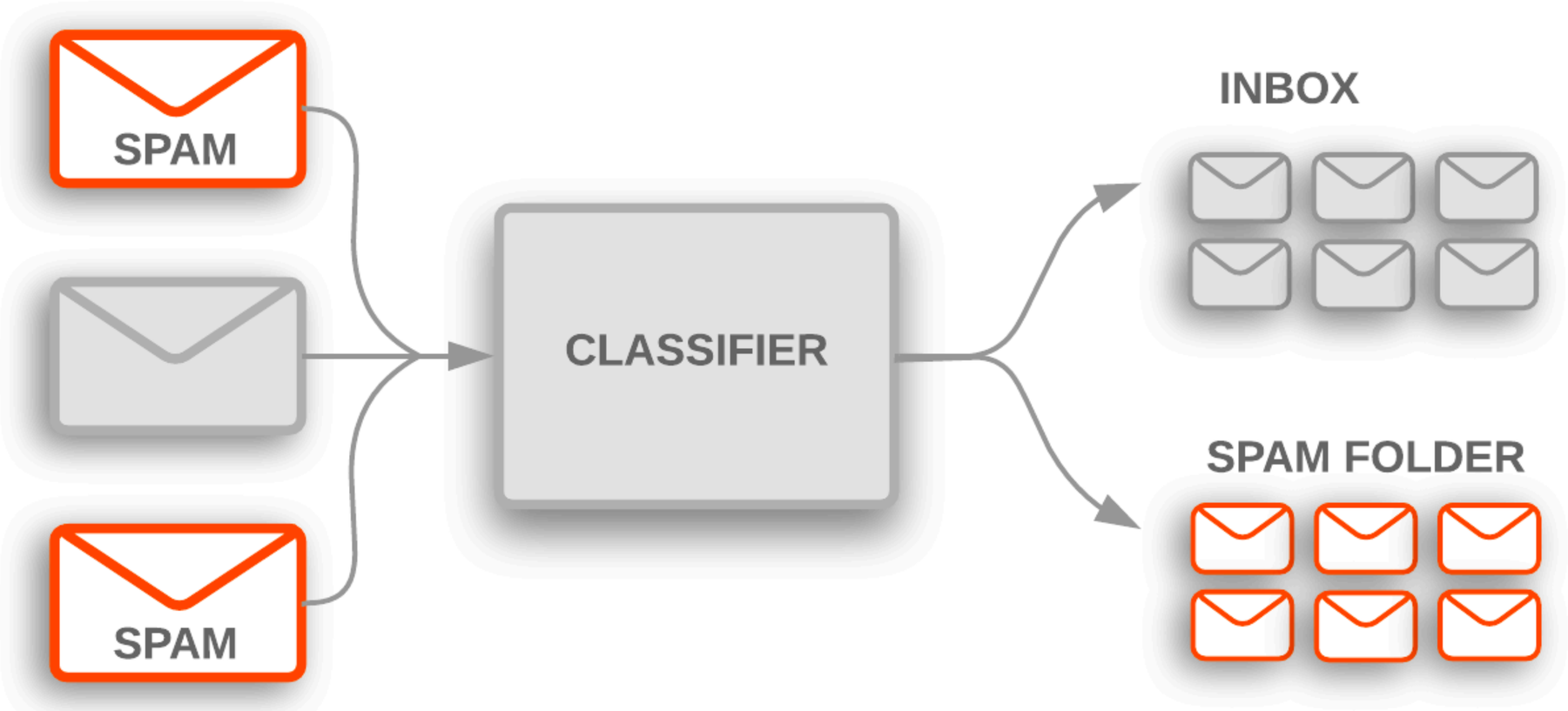
What is Machine Learning:

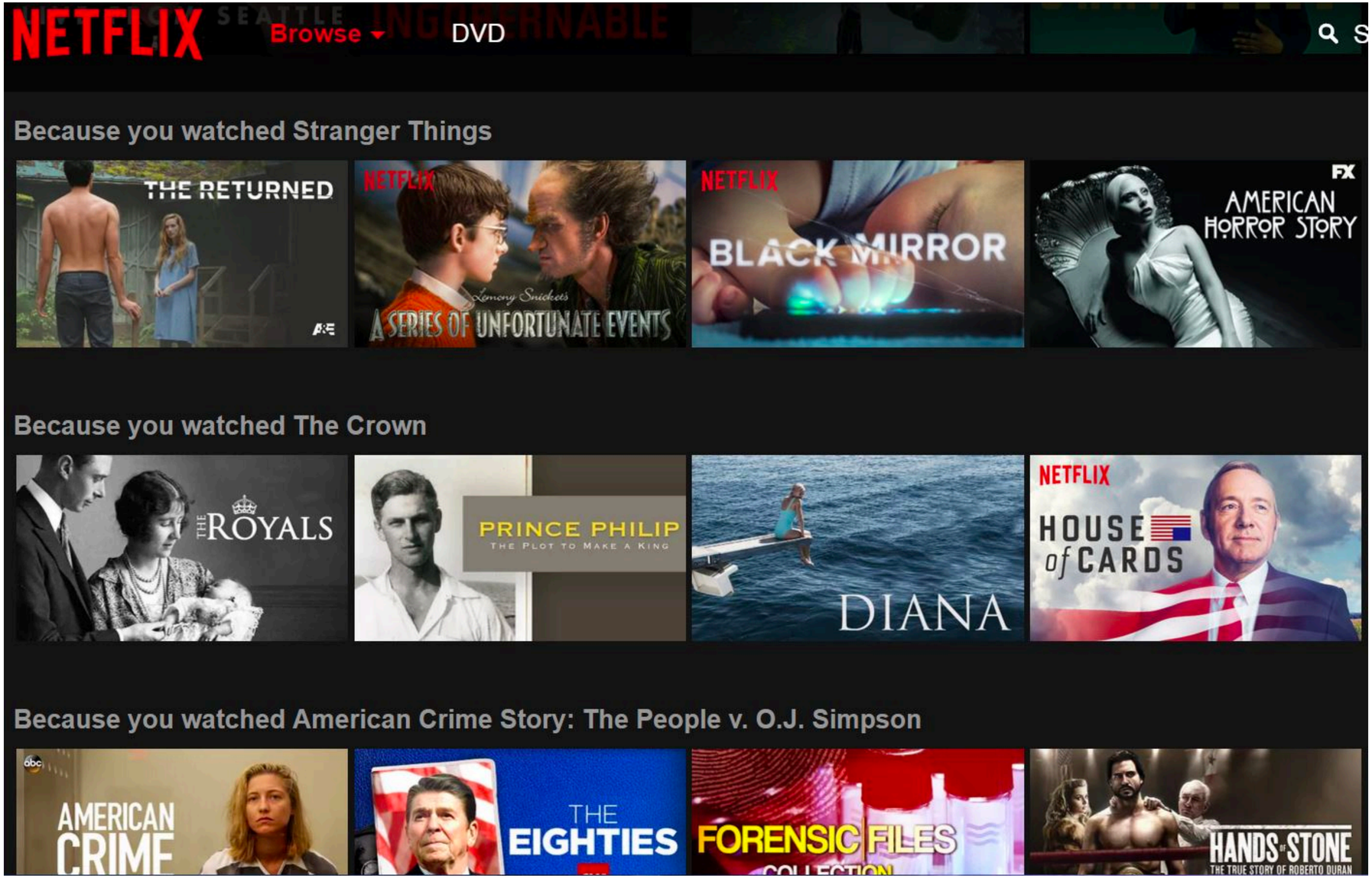
Study of algorithms that:

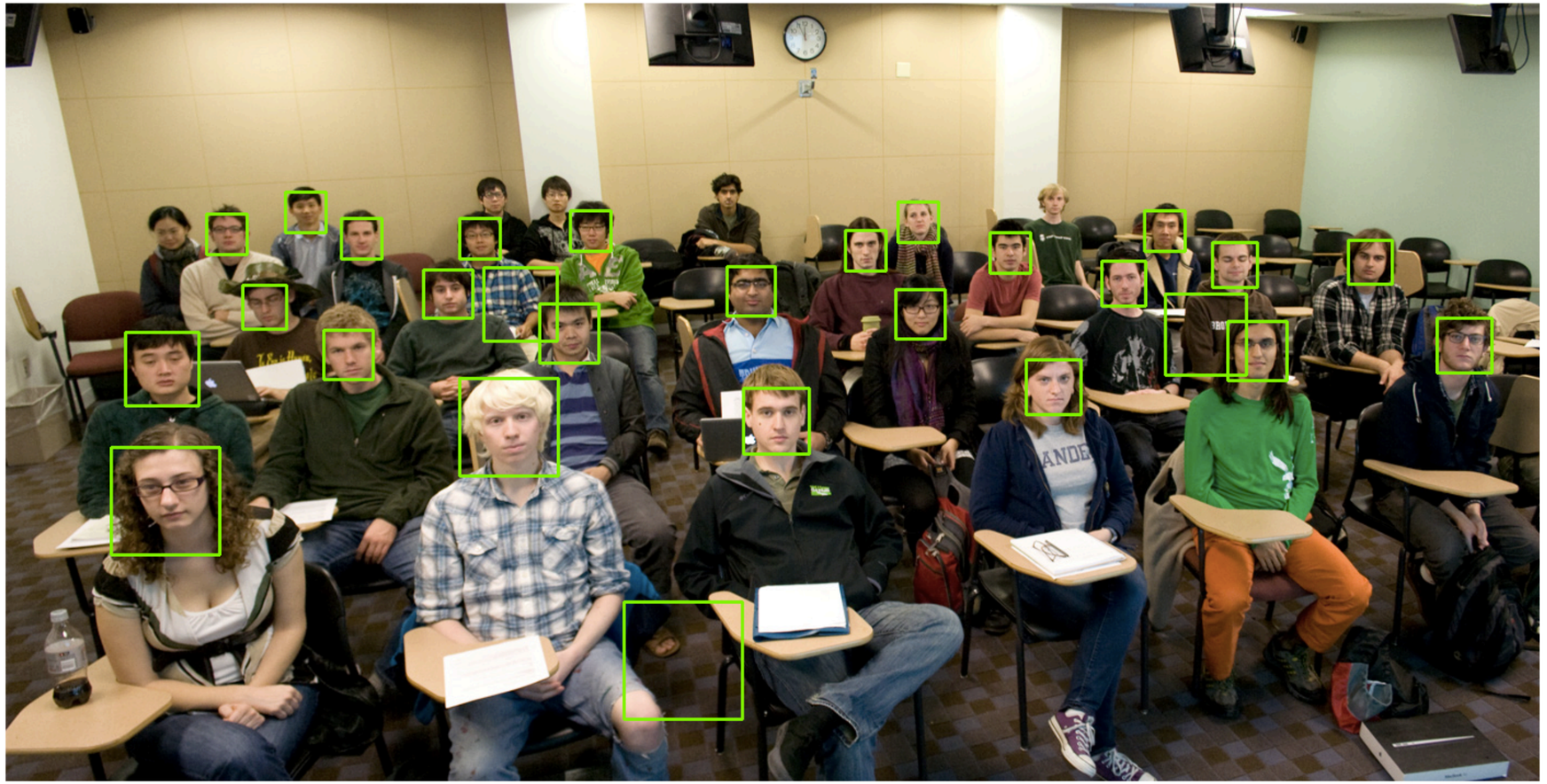
- improve their **performance** P
- at some **task** T
- with **experience** E

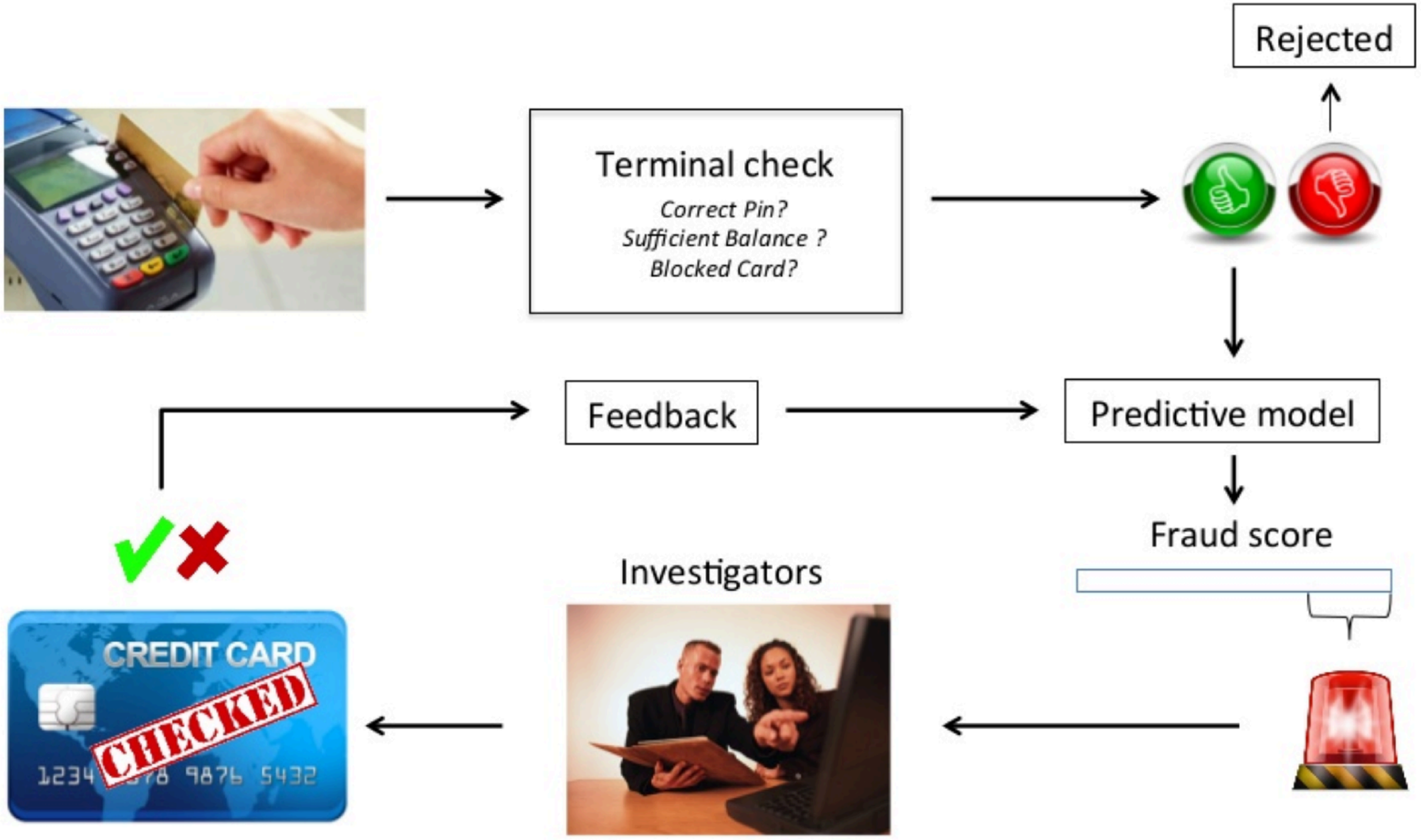
Well-defined learning task: <P,T,E>

— Tom Michell, 1997


SPAM
SPAM
CLASSIFIER
INBOX
SPAM FOLDER


NETFLIX
Browse
DVD
Because you watched Stranger Things
THE RETURNED
A&E
NETFLIX
Lemony Snicket's
A SERIES OF UNFORTUNATE EVENTS
NETFLIX
BLACK MIRROR
FX
AMERICAN HORROR STORY
Because you watched The Crown
THE ROYALS
PRINCE PHILIP
THE PLOT TO MAKE A KING
DIANA
NETFLIX
HOUSE of CARDS
Because you watched American Crime Story: The People v. O.J. Simpson
abc
AMERICAN CRIME
THE EIGHTIES
FORENSIC FILES
HANDS of STONE
THE TRUE STORY OF ROBERTO DURAN



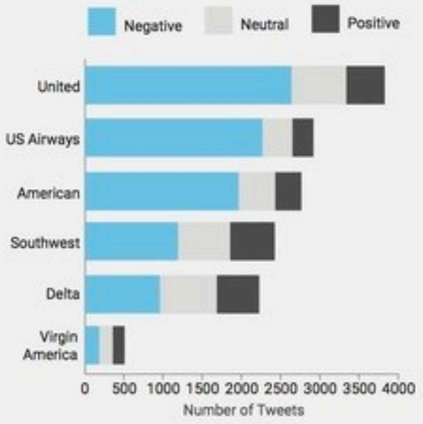

Rejected
Terminal check
Correct Pin?
Sufficient Balance ?
Blocked Card?
Feedback
Predictive model
Fraud score
Investigators
CREDIT CARD
CHECKED
1234 9876 5432

U.S. Airline Twitter Sentiment

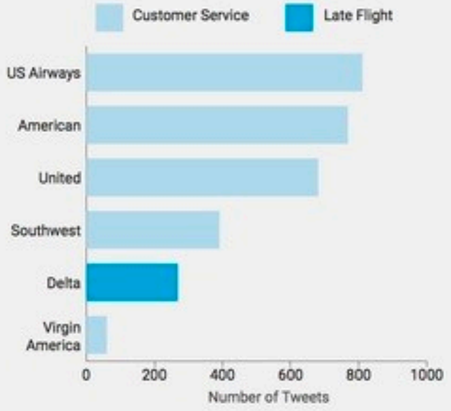
Analysis of traveler's Tweets from a week in February 2015

Cathy Liewen, Heidi Slojewski
HCI 512 | Winter 2016

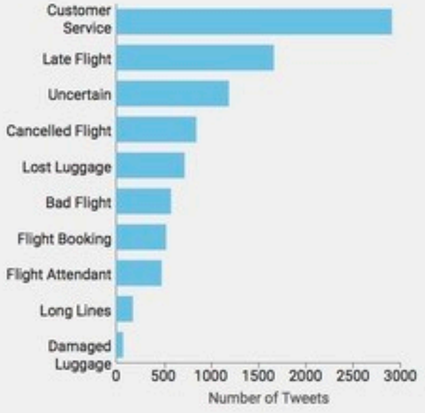
Sentiment by Airline



Airlines' Top Reasons for Negative Sentiment



Most Common Reasons for Negative Sentiment

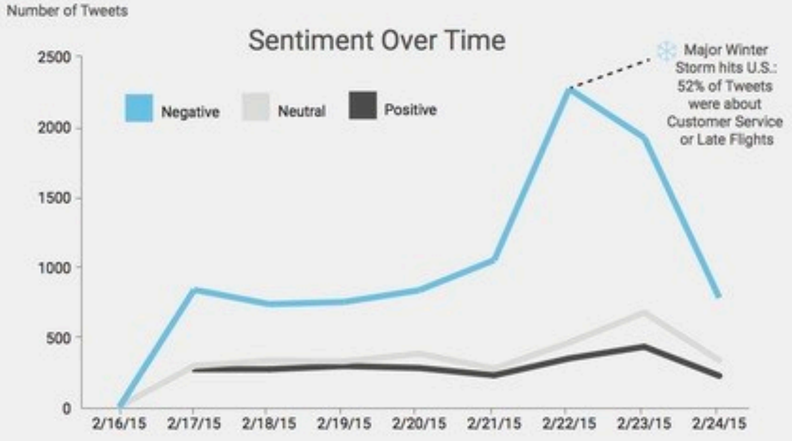


Most Retweeted Negative Sentiments

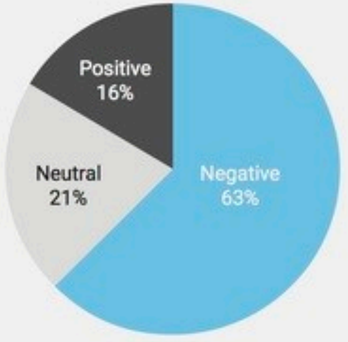
"**@USAirways** 5 hr flight delay and a delay when we land . Is that even real life ? Get me off this plane , I wanna go home"
-OBJ_3

"**@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

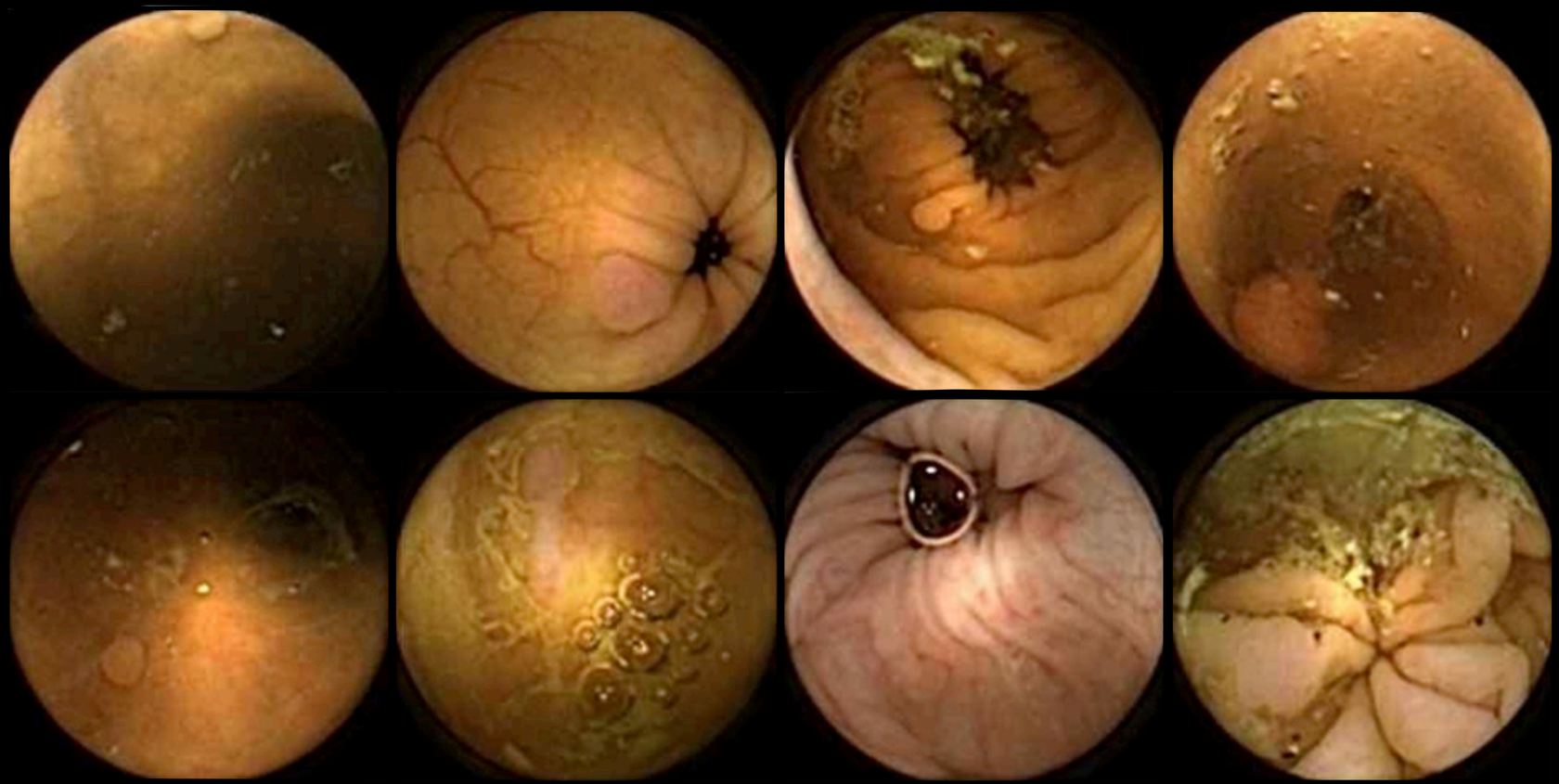


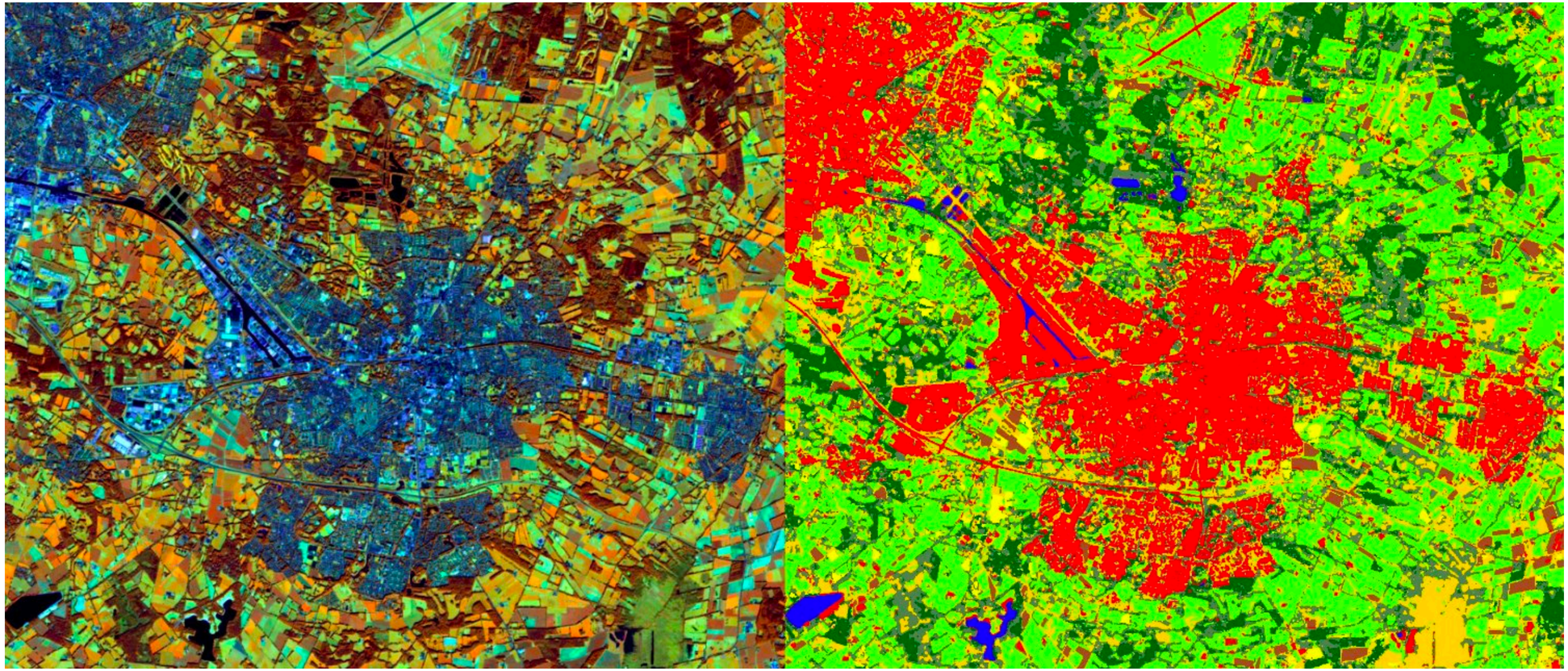
Sentiment Breakdown



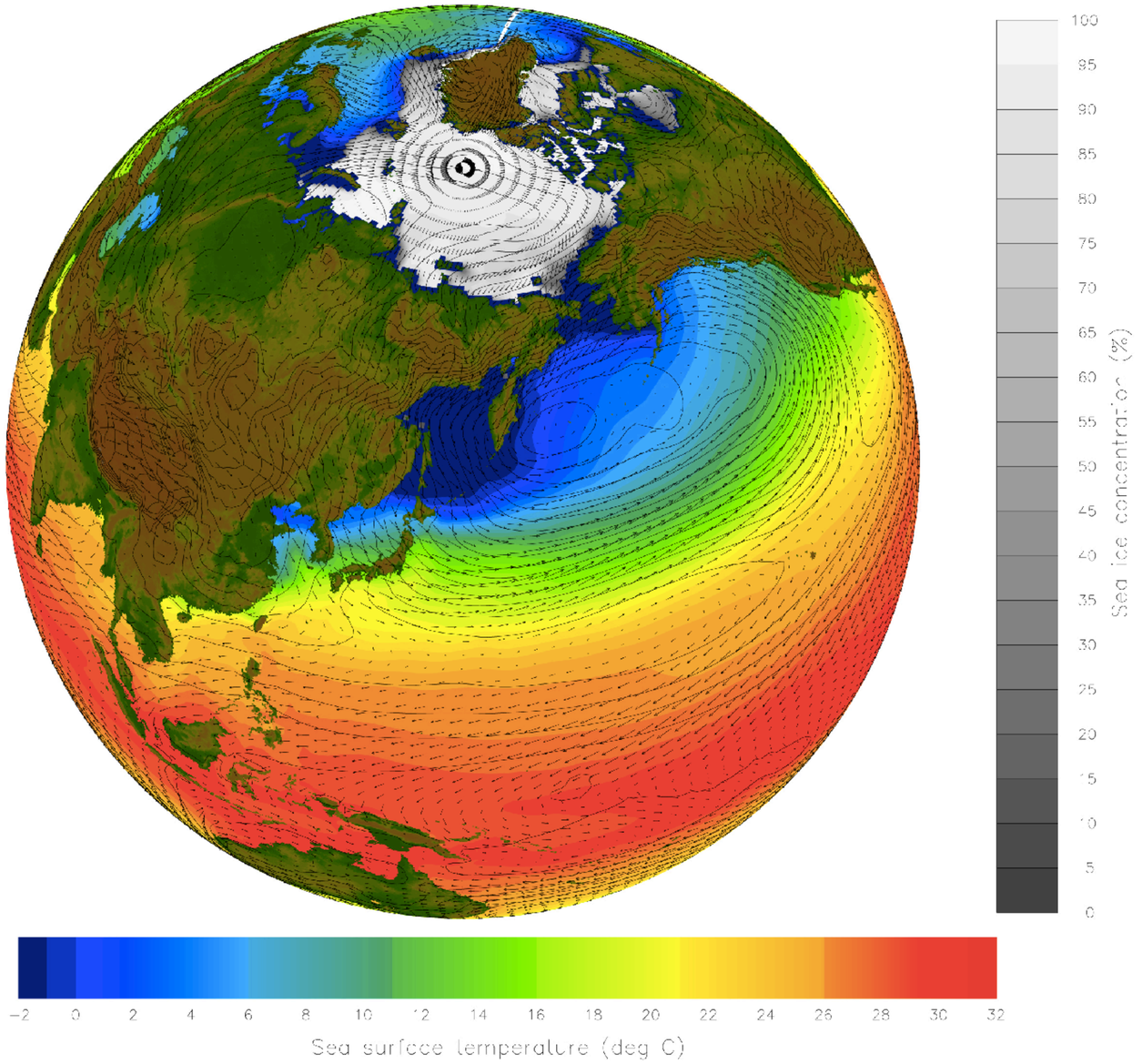
Most Common Words from Negative Tweets


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



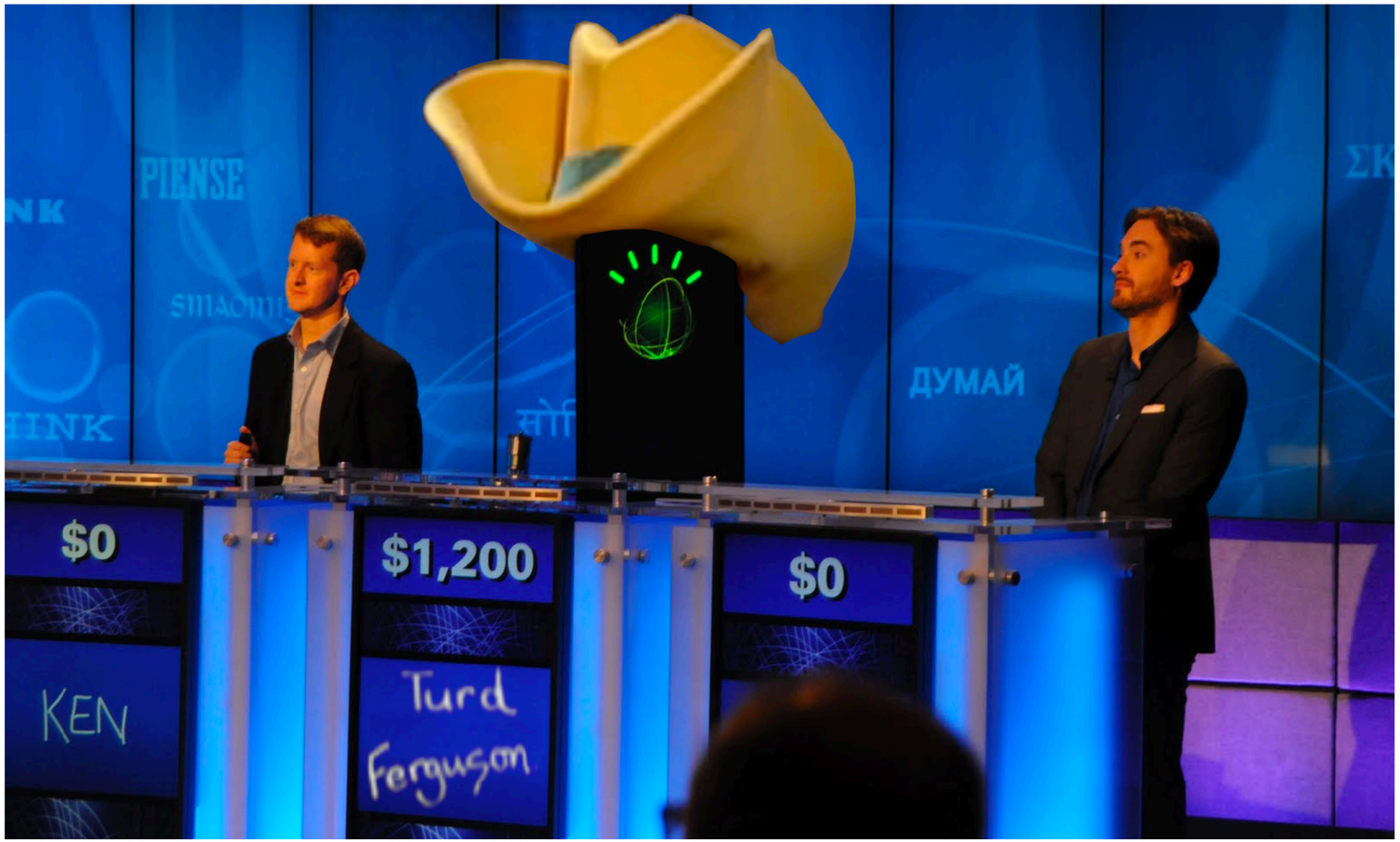


Climate change



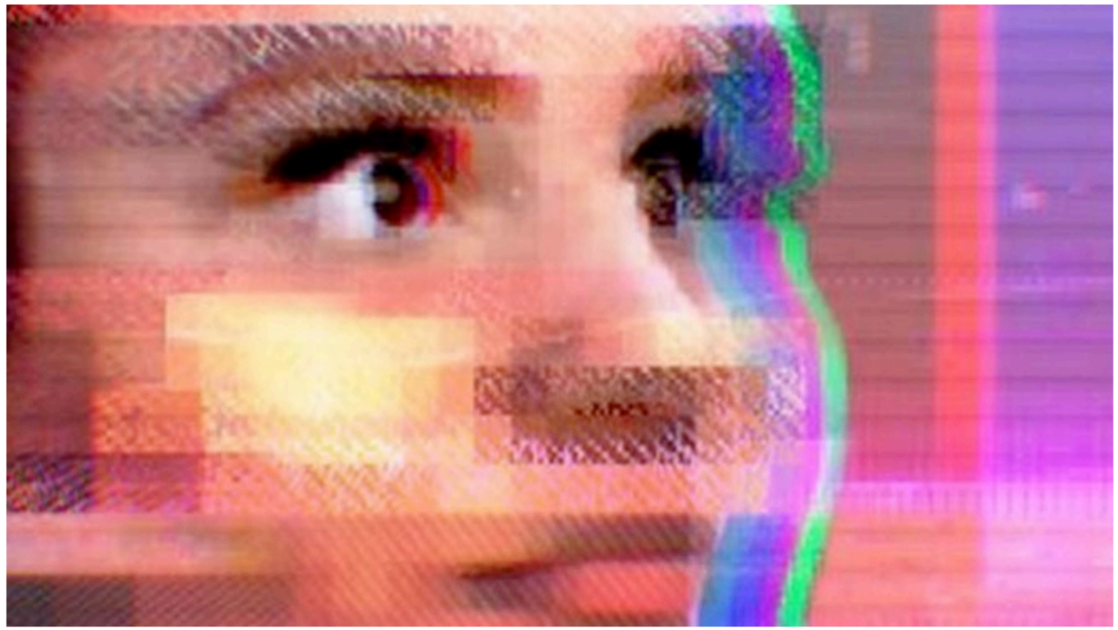

This call may be recorded to improve translation
skype
Hi grandma. Happy birthday. Are you having a great day?
Hola abuela. Feliz cumpleaños. ¿Tienes un buen día?
Hi Dylan. Yes, a great day thanks. And the flowers that you have sent are precious.
Hola Dylan. Sí, un día estupendo gracias. Y las flores 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).



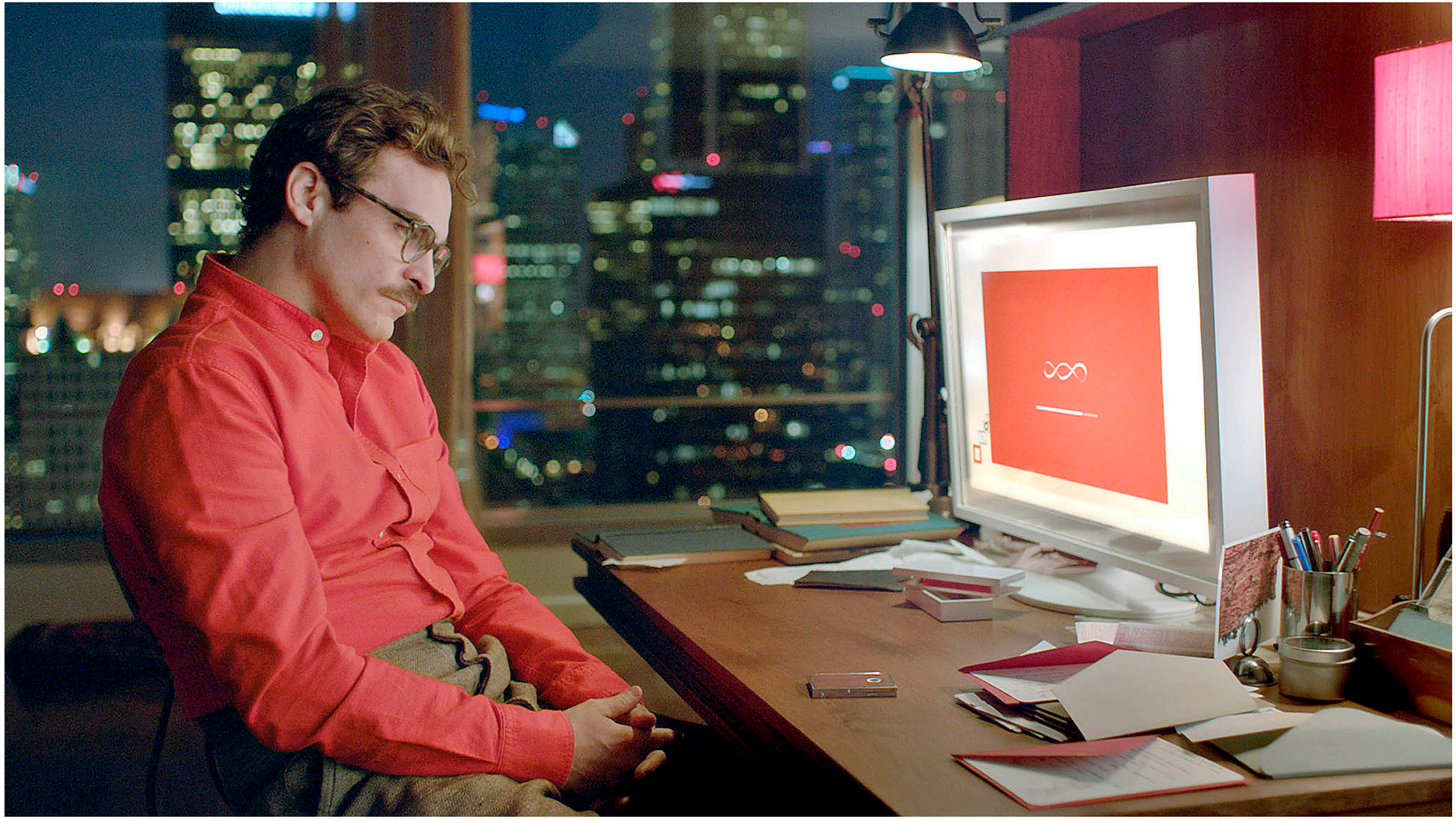
REDES SOCIALES • Microsoft ha pedido disculpas y la ha desactivado

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



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

Science Fiction or Future?

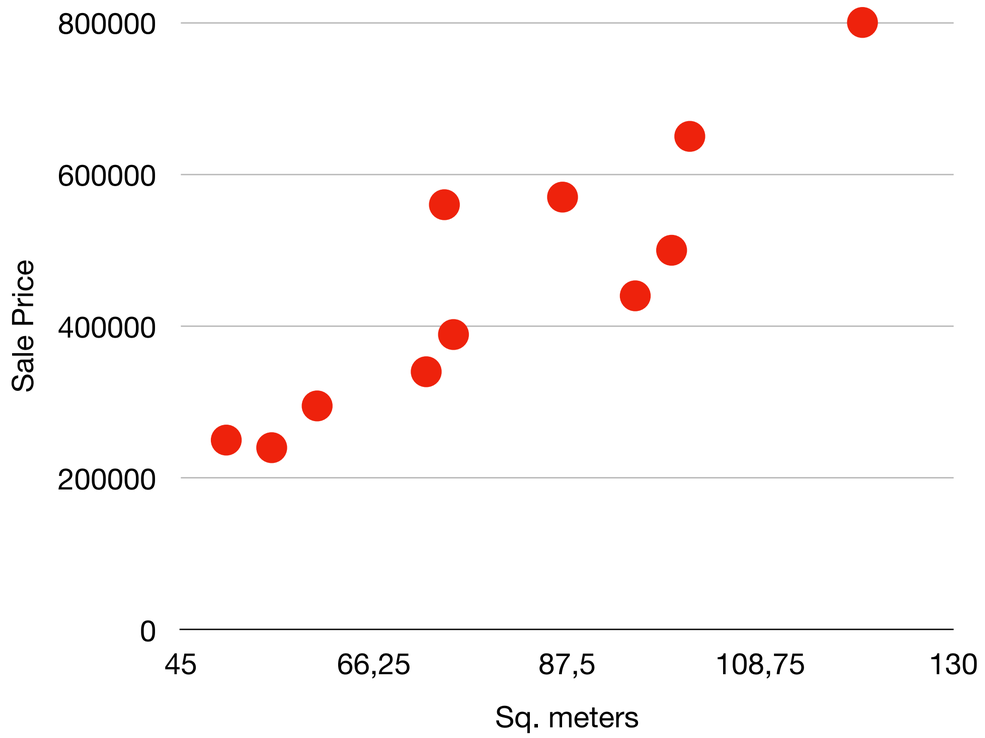


$$\hat{Y} = bX + a$$

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

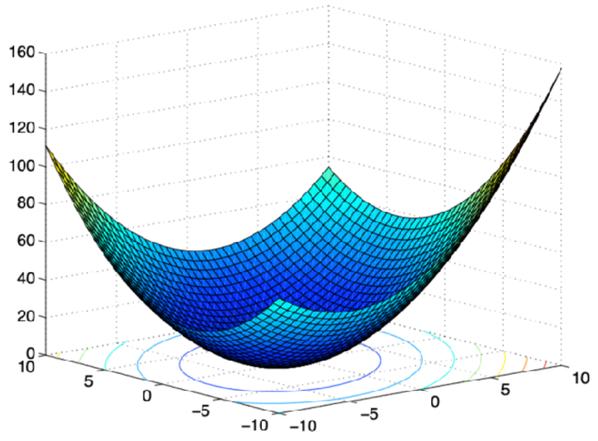
Sq. meters	Sale Price	Prediction
50	250.000	232.015
75	389.000	415.540
72	340.000	393.517
60	295.000	305.425
95	440.000	562.360
55	240.000	268.720
120	800.000	745.885
87	570.000	503.632

Optimization

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

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

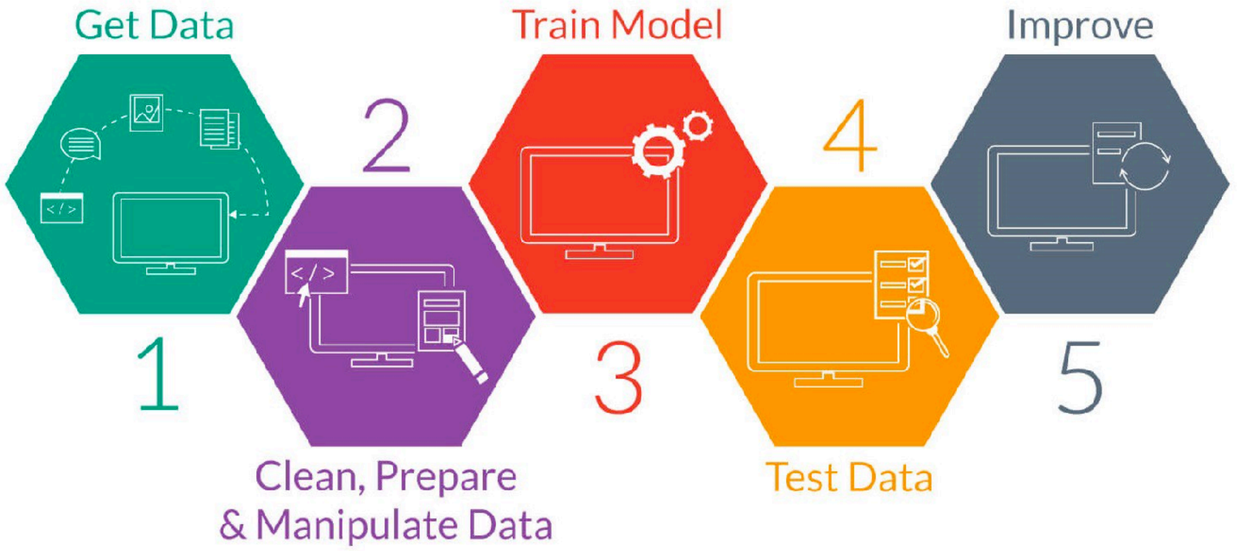
and minimize it using the training data



Type of Machine Learning

Unsupervised Learning	**Supervised Learning**	**Reinforcement Learning**
Clustering	Classification 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

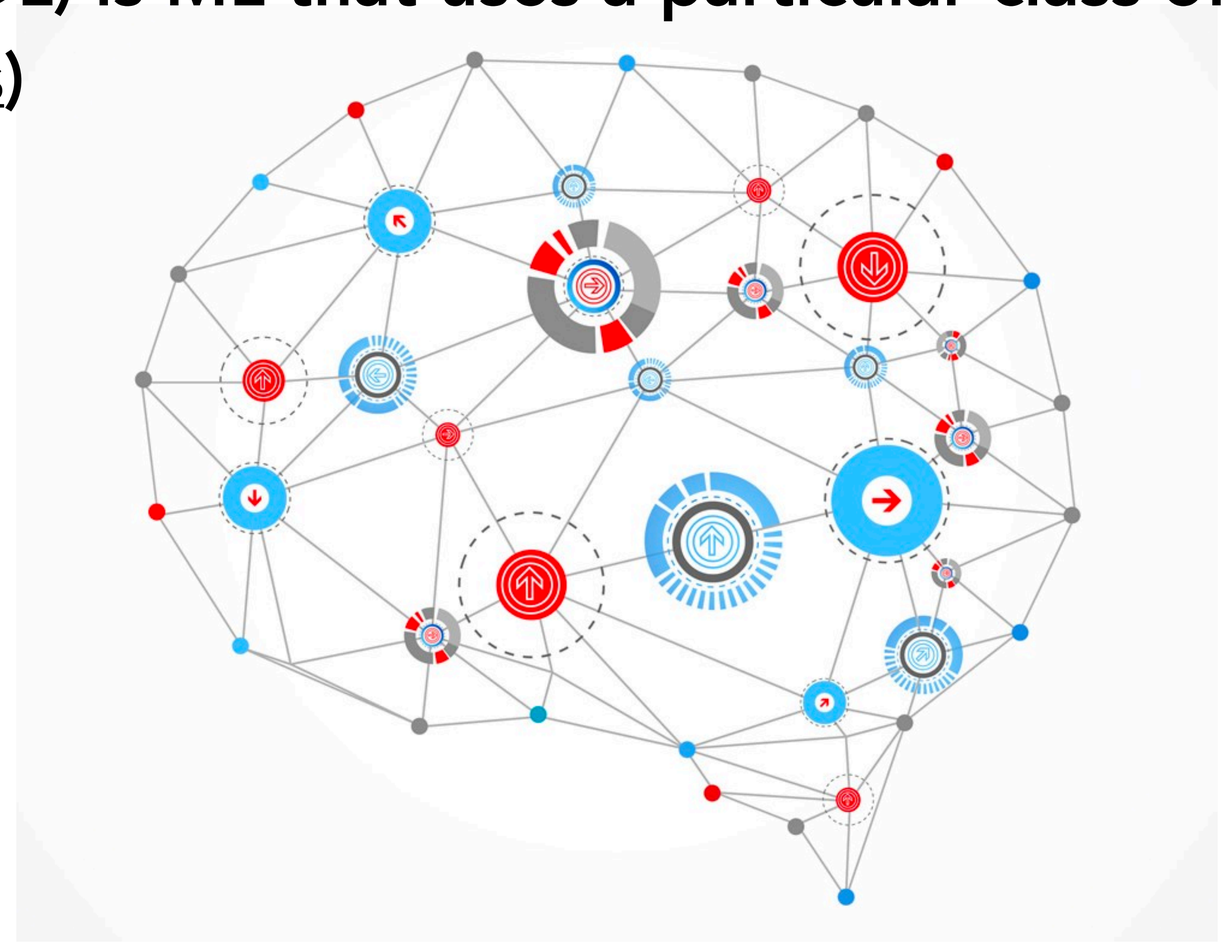
Artificial Intelligence is an academic discipline devoted to the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, language recognition, decision-making, planning, reasoning, etc.

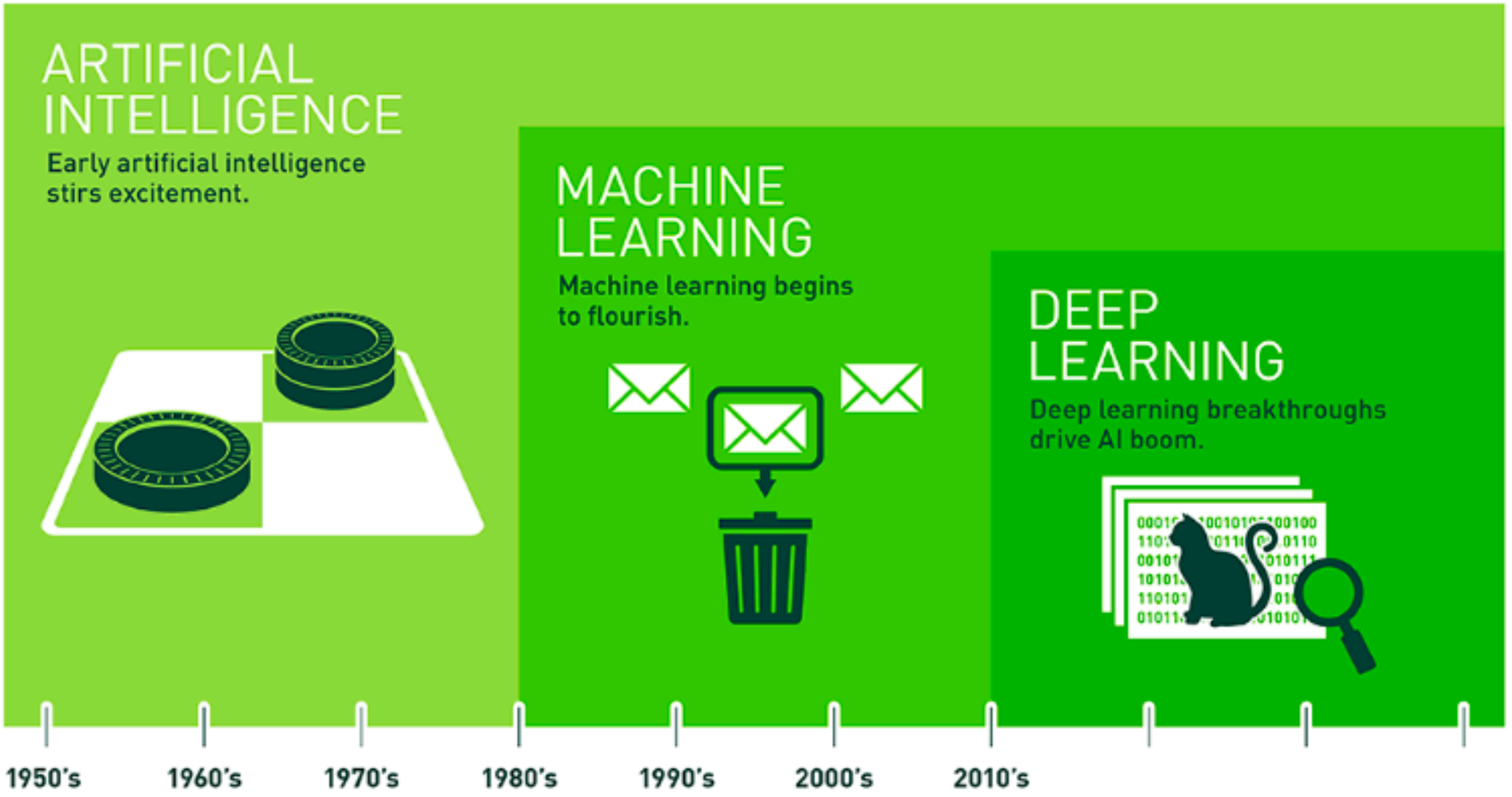
Artificial Intelligence is classified into two parts, **General AI** and **Narrow AI**. General AI refers to making intelligent in a wide array of activities that involve thinking and reasoning. Narrow AI, on the other hand, involves the use of artificial intelligence for a very specific task.

Machine learning is a subset of artificial 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)

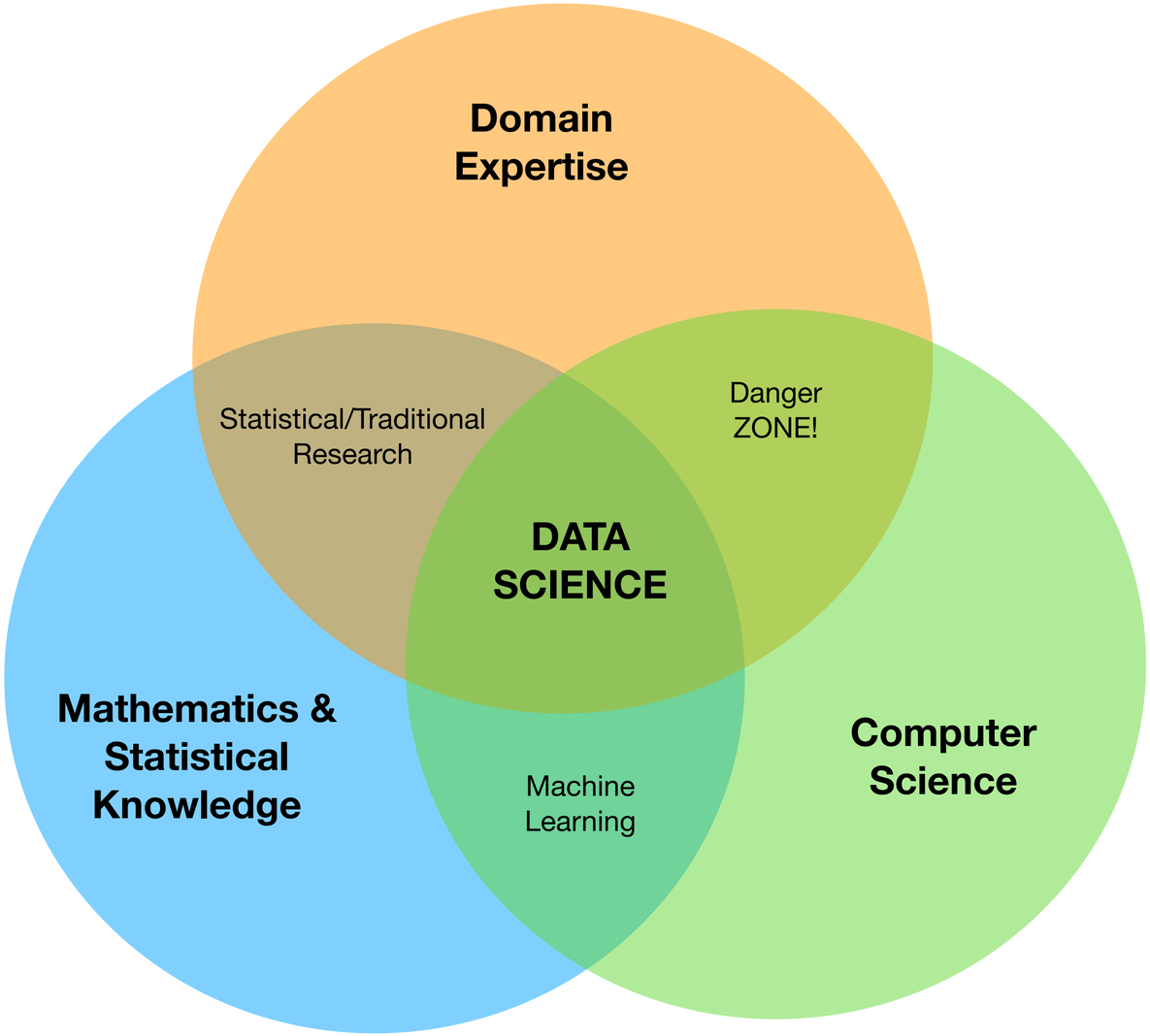



ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

Data Science

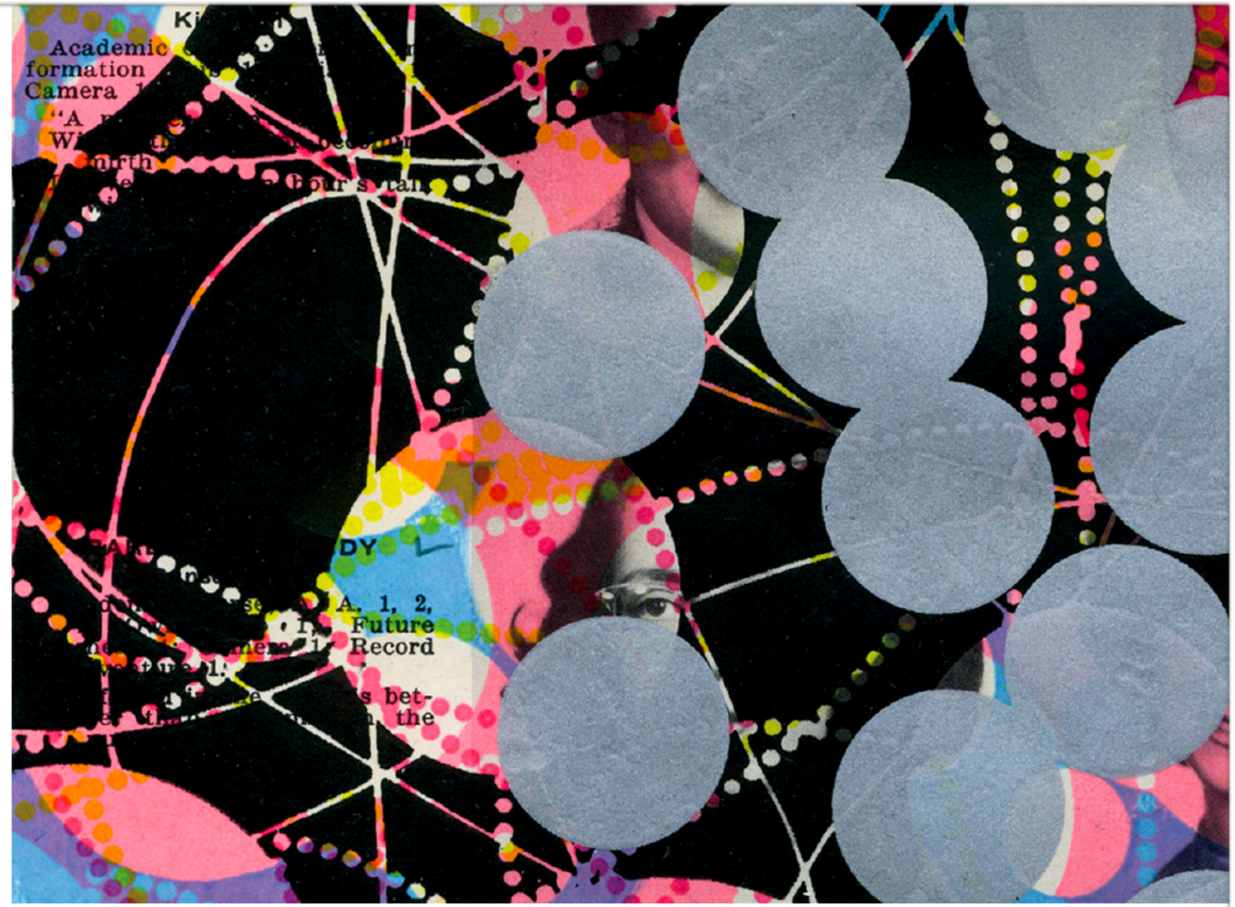
is a **multidiciplinary methodology**

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



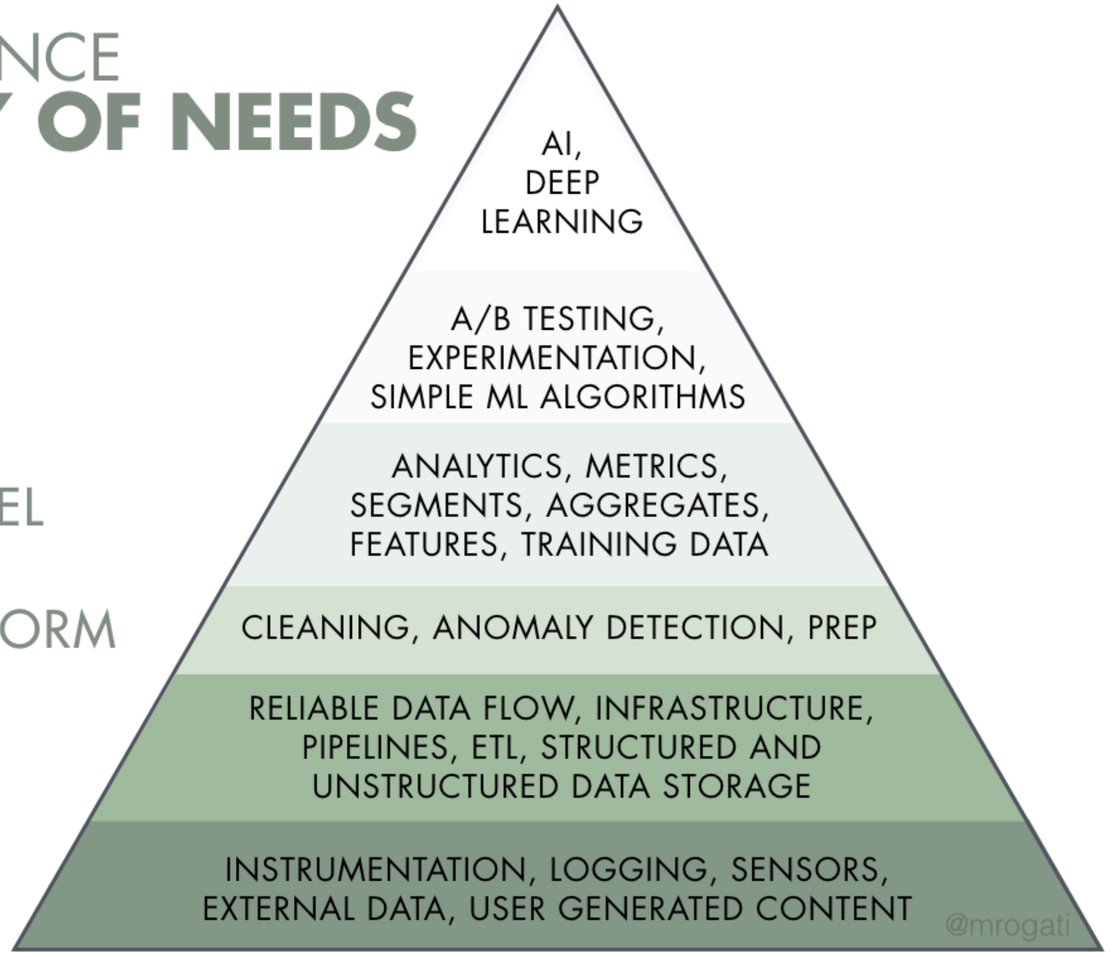
DATA

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



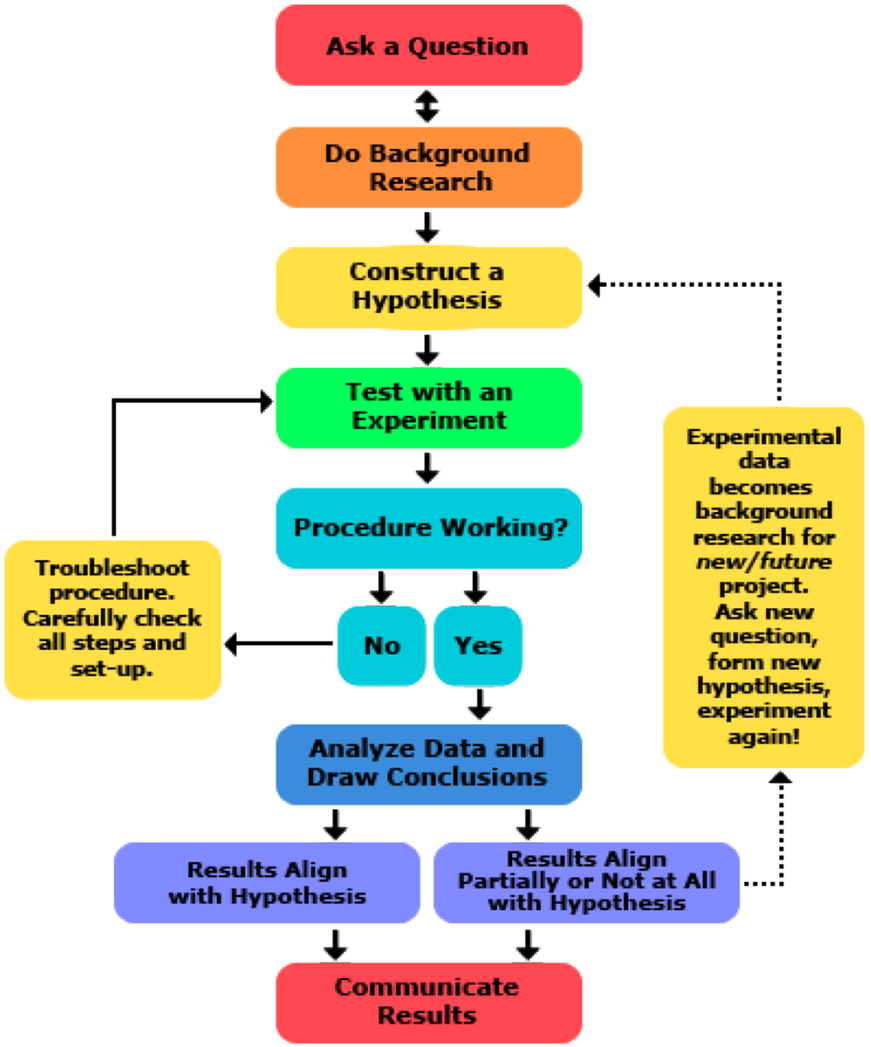
LEARN/OPTIMIZE

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

Non representative Training Data

Poor-Quality Data

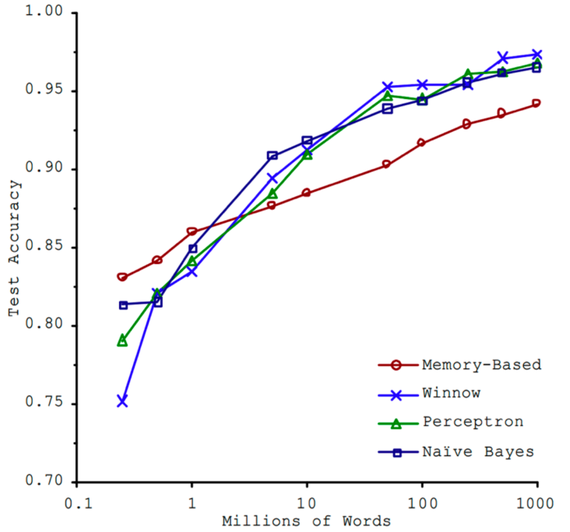
Irrelevant Features

Overfitting the Training Data

Underfitting the Training Data

Insufficient Quantity of Data

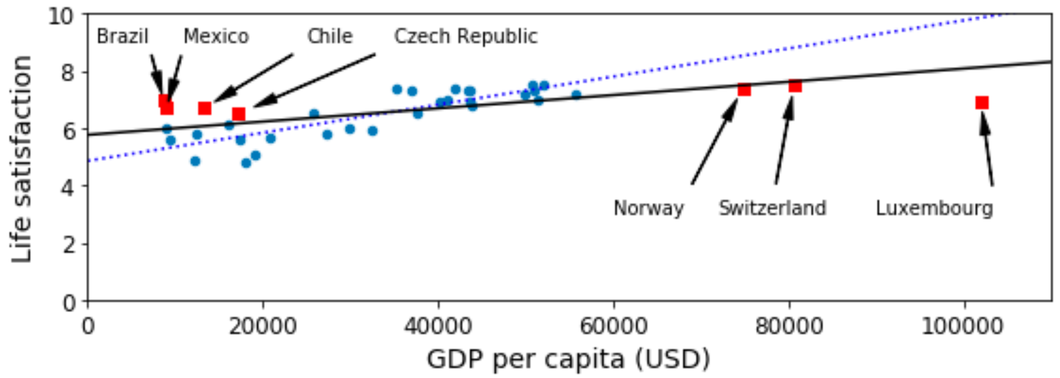
- The importance of data versus algorithms
- What is best? more data or a better model?
  - Several studies shows that very different algorithms, perform almost identically when enough data is provided



- 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

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.



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