

# Data Mining and Machine Learning

Fall 2019

Michalis Vlachos



**UNIL** | Université de Lausanne

# Introduction

# What is Data Mining?

# Tools for Data Economy

- We live in an Information (Data) Economy
- Every physical or digital transaction generates some data. This trend will continue to grow.
- Especially with the growth of social networks, we can learn much more about what our clients do, like, and connect with.

# How much data exists today?

- By 2020 there will be 40 zettabytes of data.
  - $1 \text{ ZB} = 10^9 \text{ Terabytes}$
- If printed in textbooks, the stack of textbooks would reach to Pluto and back 20,000 times.

# Data Mining – various definitions

Data Mining (DM) = Use of mathematical methods (regression, clustering, classification, deep learning) to extract patterns and **actionable** information from large amount of data (=Big Data)

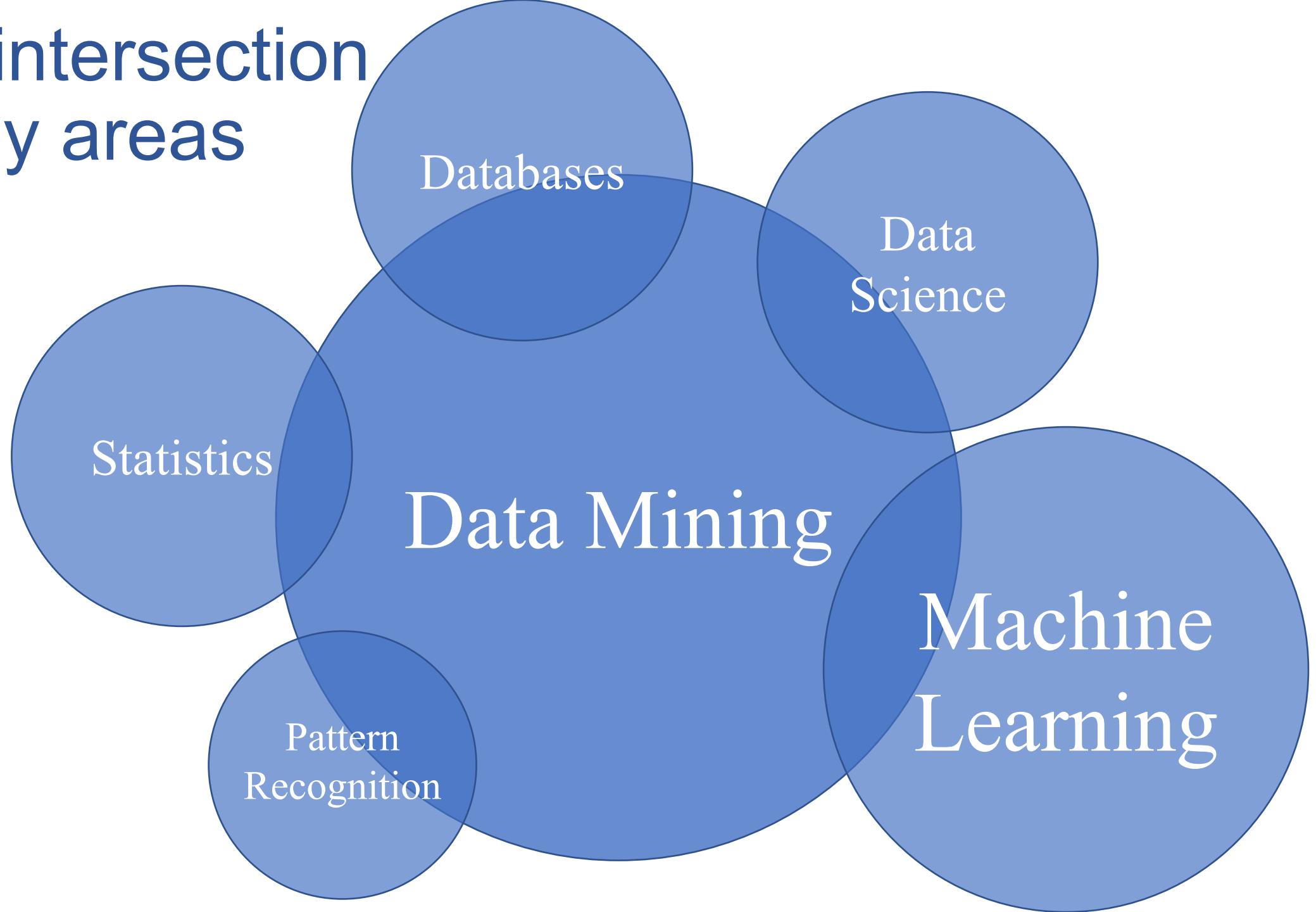
Data mining is the nontrivial extraction of **implicit**, **previously unknown**, and **potentially useful** knowledge from data (Frawley et al., 1992)

Data science + machine learning + visualization = Data Mining

my definition

One thing is for sure, DM the first field that understood the big data trend

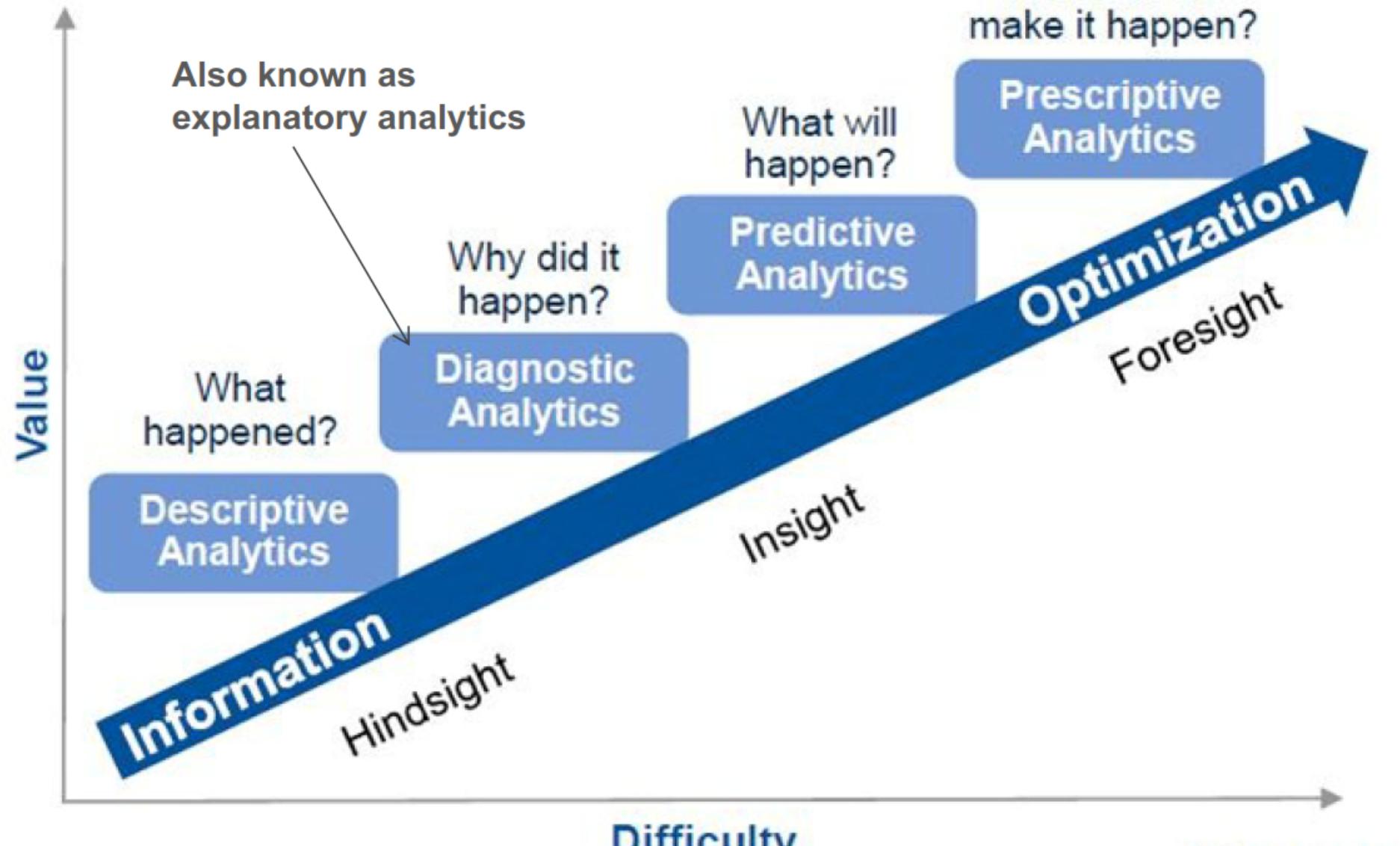
At the intersection  
of many areas



# Goals of Data Mining

## Goals:

- **Improve processes**
  - A hospital realizes patients in emergencies take too long to process
- **Understand your clients**
  - Do we serve our customer needs with our products?
- **Increase productivity**
- **Improve efficiency**



Gartner

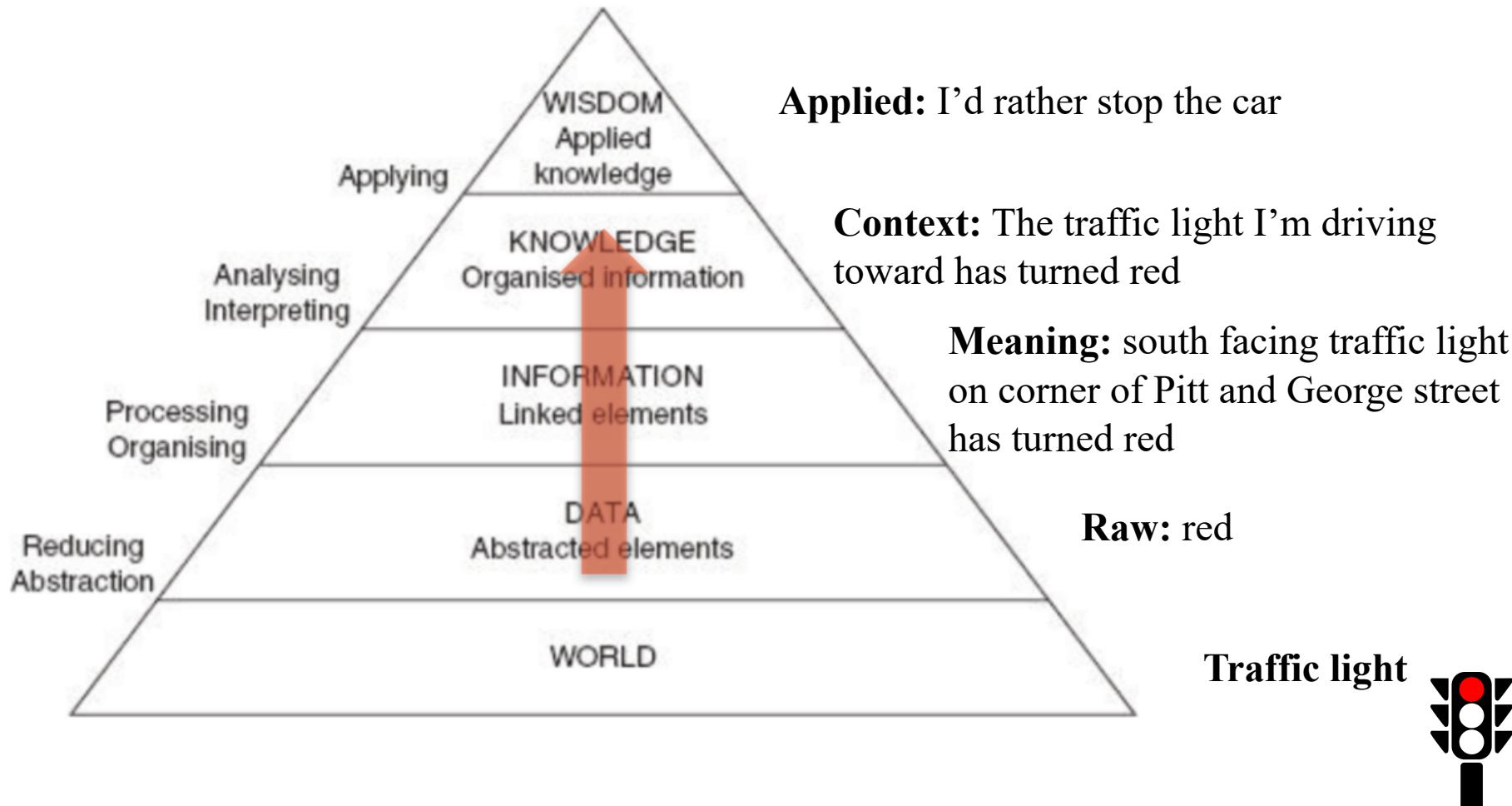
# Data = knowledge?

- **Data** = simple, isolated facts (e.g. values in a database table)
- **Information** = data in context
- **Knowledge** = interpreted information
- **Intelligence** = use of knowledge to choose between alternatives
- **Wisdom** = intelligence + experience (guided by values and commitment)

**data < information < knowledge**

- without data, there can be no information and
- without information, there can be no knowledge

# From Data To Wisdom - Example



There exist many tools to perform analytics and data mining. We will be focusing on Python in this class.

1. We will build some techniques from scratch.

2. We will learn how to use existing libraries.

**ANALYTICS**

**DATA ANALYST PLATFORMS**

Microsoft, pentaho A Hitachi Group Company, alteryx, Digital Reasoning, GUAVUS, AYASDI, ATTIV/O, Datameer, Quid, incorta, interana, ClearStory Data, Origami logic, ASCEND.io, ENDOR, MODE, Bottlenose, switchboard

**DATA SCIENCE PLATFORMS**

IBM, KNIME Open for Innovation, dataiku, DOMINO, rapidminer, CONTINUUM ANALYTICS, ALGORITHMIA, DATAWATCH ANGOSST, SAS

**BI PLATFORMS**

Microsoft, AWS, tableau,波形图, Wave Analytics, looker, THOUGHTSPOT, AT SCALE, ARCADIA DATA, GoodData, Information Builders, SISENSE, MicroStrategy, burst, ZEPL, CHARTIO, Qlik, Q, Google Cloud, celonis, Periscope Data, ZOOMDATA, plotly, VISENZE, TOUCAN TOCO

**VISUALIZATION**

SAP Lumira, SAP, celonis, Qlik, Q, Google Cloud, Periscope Data, ZEPL, ZOOMDATA, plotly, CHARTIO, TOUCAN TOCO

**MACHINE LEARNING**

Azure Machine Learning, AWS, Google Cloud, H2O.ai, DataRobot, gamalon, ELEMENT AI, VISENZE, deepsense.io, bonsai

**COMPUTER VISION**

Microsoft Azure, Amazon Rekognition, clarifai, Cloud Vision API, EVER AI, deepomatic, twentybn, neurala

**HORIZONTAL AI**

IBM Watson, Cortana, Face++, 旷视, sentient technologies, Voyager Labs, vicarious, Affactiva, PROPHESEE, METAVISION FOR MACHINES, CognitiveScale, Numenta, PETUM, Scaled Inference, naraLogics, THE CURIOUS AI COMPANY, OSARO, BLUE VISION

**SPEECH & NLP**

Google Cloud, twilio, amazon alexa, narrative science, semanticmachines, WolframAlpha, Mobvoi, EigenTechnologies, SoundHound Inc., PRIMER, MindMeld, voicera, cortical.io, NUANCE, maluuba, Gridspace, snips, yseop

**SEARCH**

ORACLE ENDECA, elasticsearch, EXALEAD, COVEO, Lucidworks, ATTIV/O, swiftype, algolia, alphasense, MAANA, omni:us, SINEQUA

**LOG ANALYTICS**

splunk, sumologic, loggly, TIMBER, kibana, logz.io

**SOCIAL ANALYTICS**

Hootsuite, sprinklr, NETBASE, synethesio, trax, simplereach, bitly, predata, SimilarWeb

**WEB / MOBILE / COMMERCE ANALYTICS**

Google Analytics, mixpanel, AMPLITUDE, sumall, Airtable, RESCI, SIGOPT, granify, custora

# Question

Why should you take a data scientist with you in the jungle?

Answer: Because they take care of the python problems.

# What is Machine Learning?

# What is ML?

It is a branch of Artificial Intelligence.

**“Machine learning** is the science of making computers learn and act like humans by feeding data and information without explicitly being programmed”

# What does the previous definition mean?

	customerID	TotalCharges
0	7590-VHVEG	29.85
1	5575-GNVDE	1889.5
2	3668-QPYBK	108.15
3	7795-CFOCW	1840.75
4	9237-HQITU	151.65

If I give you this table, do you know how to write a program (i.e., rules, flow) to compute the **total charges**?

# What does the previous definition mean?



If I give you some images, can you write code that recognizes if something is a car?

This is much more complicated.

# What does the previous definition mean?



car



car



car



car



car



car

Instead, you write a very generic program that “learns by examples”.

So you can apply this to recognize or “predict” anything (cars, apples, ..., stock price will increase,...)

# Example of ML: autocomplete in search engines

The image shows a screenshot of a Google search results page. At the top is the Google logo. Below it is a search bar containing the text "data mining". A dropdown menu displays a list of autocomplete suggestions: "data mining", "data mining eth", "data mining vs machine learning", "data mining techniques", "data mining concepts and techniques", "data mining tools", "data mining cup", "data mining deutsch", "data mining définition", and "data mining process". At the bottom of the search interface are two buttons: "Google Search" and "I'm Feeling Lucky". Below the search bar, there is a link to report inappropriate predictions and a "Learn more" button.

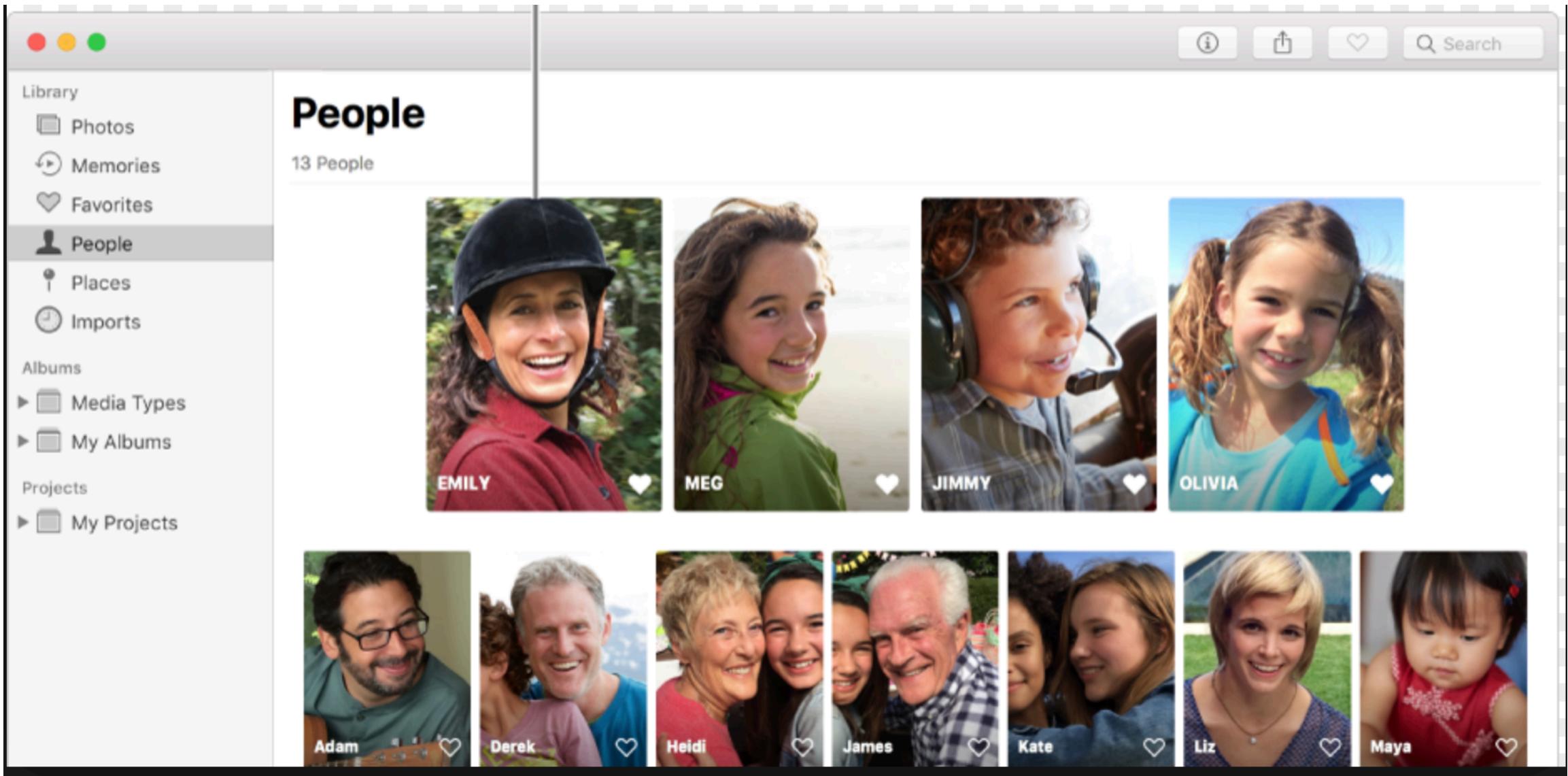
data mining

data mining  
data mining eth  
data mining vs machine learning  
data mining techniques  
data mining concepts and techniques  
data mining tools  
data mining cup  
data mining deutsch  
data mining définition  
data mining process

Google Search I'm Feeling Lucky

Report inappropriate predictions  
Learn more

# Example of ML: person detection in iPhoto



# Work in Two: Travallier à deux

- [3mins] Work with your neighbor.
- Do you think Machine Learning will match at some point in the future the human intelligence? When?
- [3mins] We discuss.

What is the difference  
between Data Mining  
and Machine Learning?

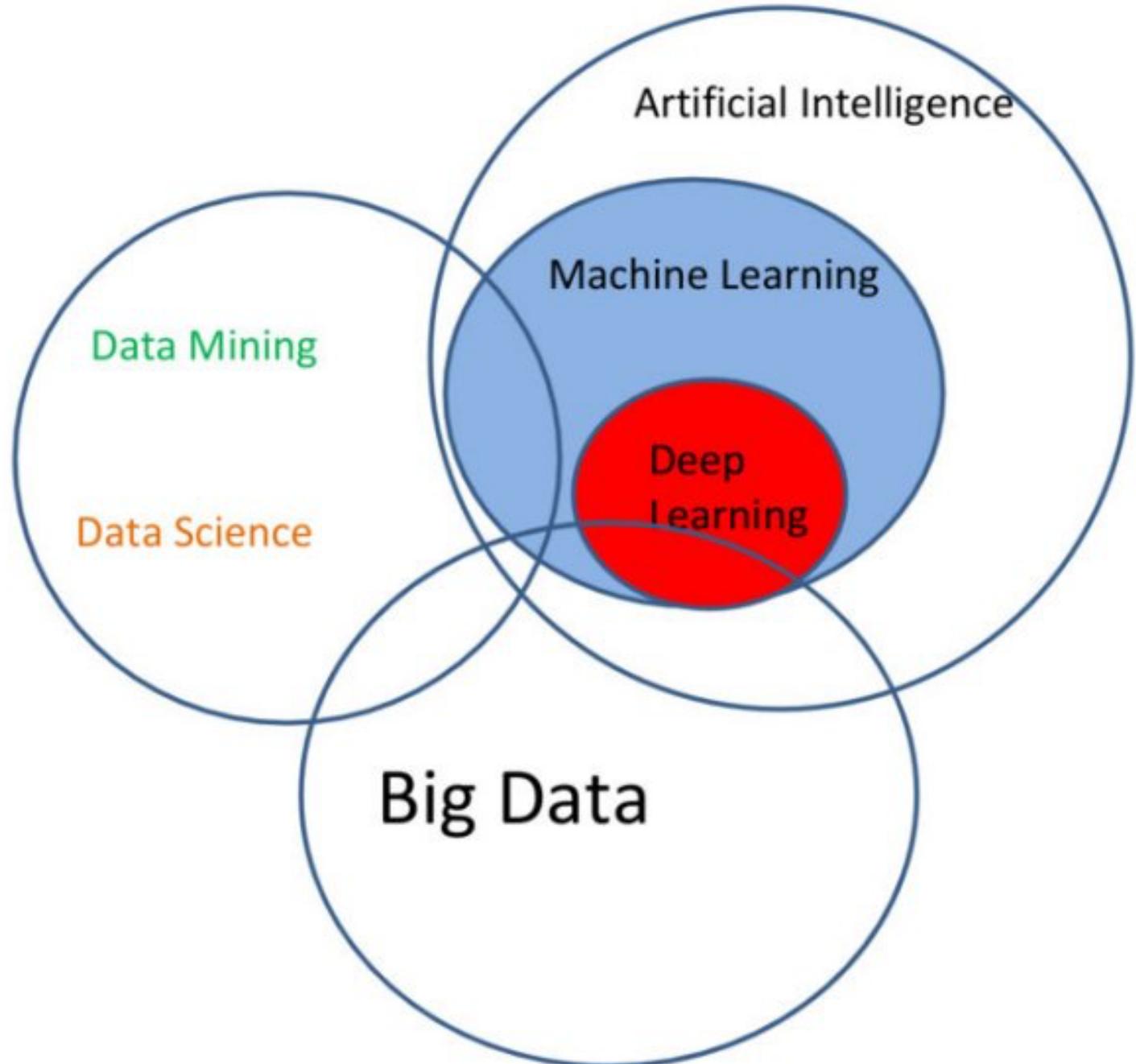
# Historical Differences

- In the '80s, '90s, 2000 Machine Learning was only interested in the “learning” part. How to make a better model that gives better **accuracy**: e.g., detects images of cars more accurately.
- Data mining in the '90s was more focused on the **data-scale** part. How to make existing machine learning algorithm, e.g., clustering, applicable for large amount of data through sampling, summarization, etc.

# Today's Differences

- Today, these differences are more nuanced as the two fields converge into one.
- Today, you can say data mining deals with the search, cleaning, processing, storage of the data. Machine learning deals with the creation and tuning of the proper learning model.

# The big picture



# your expectations

1. Why you enrolled in the class?
2. What do you expect to learn?

# Take the class survey

<https://tinyurl.com/DM-survey2019>

# Why follow this class

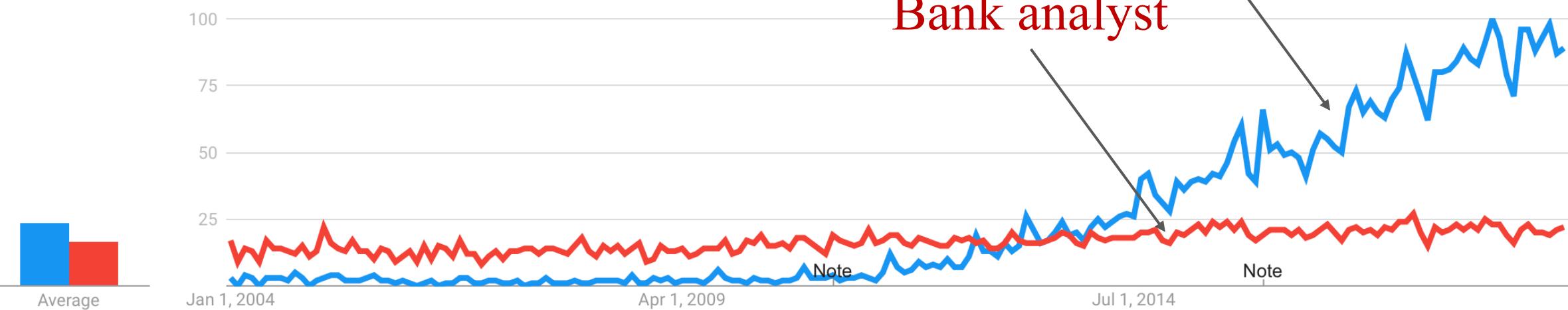
- You want to land a job as a data scientist.
- You want to enhance your coding skills in Python.
- You want to understand what the whole AI buzz is about.
- You want to publish papers in the area and receive a PhD.

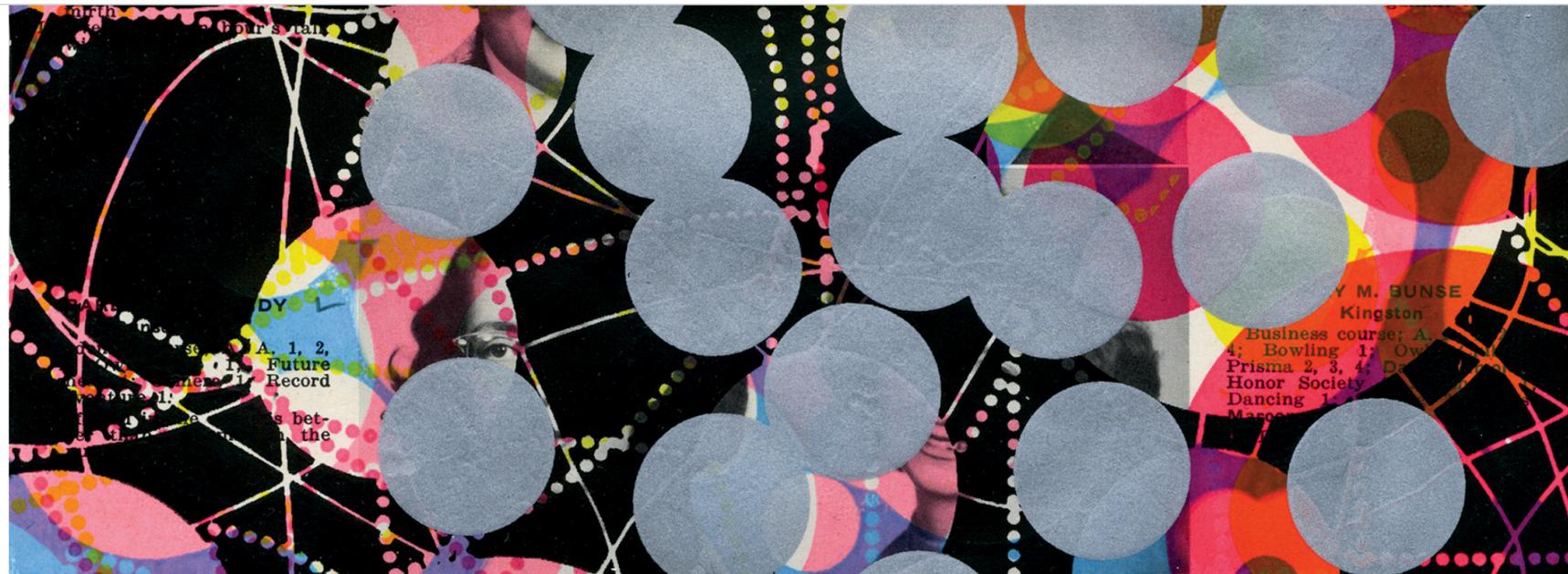
Interest over time [?](#)

Data scientist



Bank analyst





ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

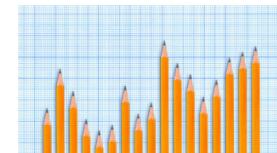
# Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

[Summary](#) [Save](#) [Share](#) [Comment 16](#) [Text Size](#) [Print](#) [\\$8.95 Buy Copies](#)

## WHAT TO READ NEXT



**What Data Scientists Really Do, According to 35 Data Scientists**

<https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century?>

# Class goals

31

1. Approach problems data-analytically
  - Think carefully & systematically about whether and how data can improve performance
2. Be able to interact competently on the topic of data mining analytics
  - Know the basics of data mining processes, techniques, and concepts well enough
3. Receive hands-on experience mining data
  - You should be able to follow up on ideas or opportunities that present themselves

# Course Logistics and Expectations

# Requirements

- You should be comfortable using Python. If you don't, most likely you shouldn't be taking this class.
- Basic statistics and math
- You **have to** bring your laptop in-class (but not check emails, browse the internet, etc. Tough...I know!)

An important piece of advice: The class can be quite demanding. You will be expected to spend at least 6-8 hours per week on it. Most of it will be coding and also understanding and applying analytics concepts.

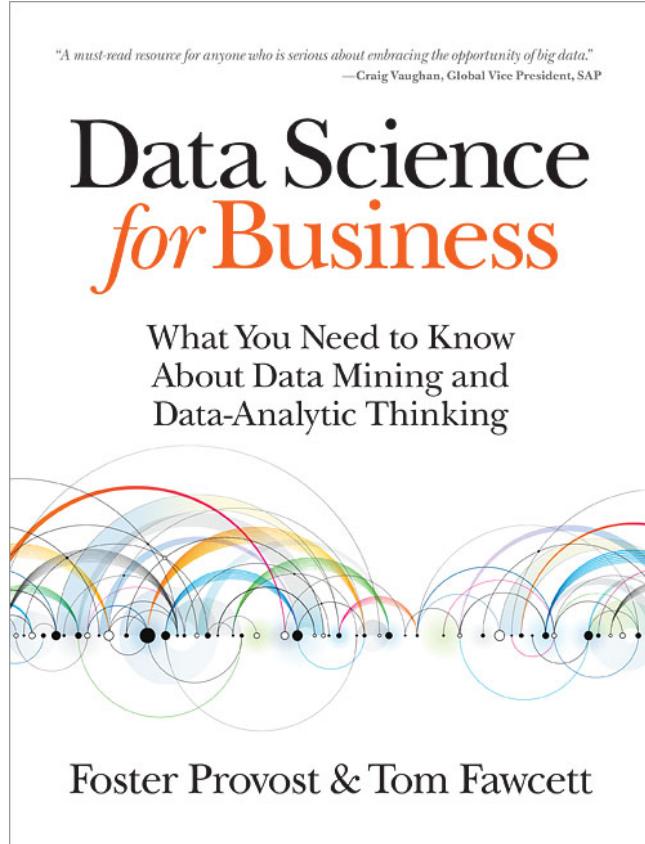
# Grades and Course Load

- **30%**: 4 in-class quizzes (20mins) – Lowest grade will be dropped.
- **30%**: Personal assignments. Report + Python + video (several days)
  - No late policies. Once submission is closed, it's closed.
- **30%**: One group project (several weeks)
- **10%**: Class participation: at least one post and one reaction/response to someone else's post on class slack channel **for 10 weeks**.
  - Example Post: You post a link to an interesting related article related to the week's topic you found in Times, the Economist, blogpost, etc and you explain how it relates to our material.
  - Example of good response: “The article here contradicts your statement, but still liked your post. I also believe that ...”
  - Example of bad response: “You don't know what you are talking abo

Be civil and  
respect your  
fellow classmates.

# Recommended Reading

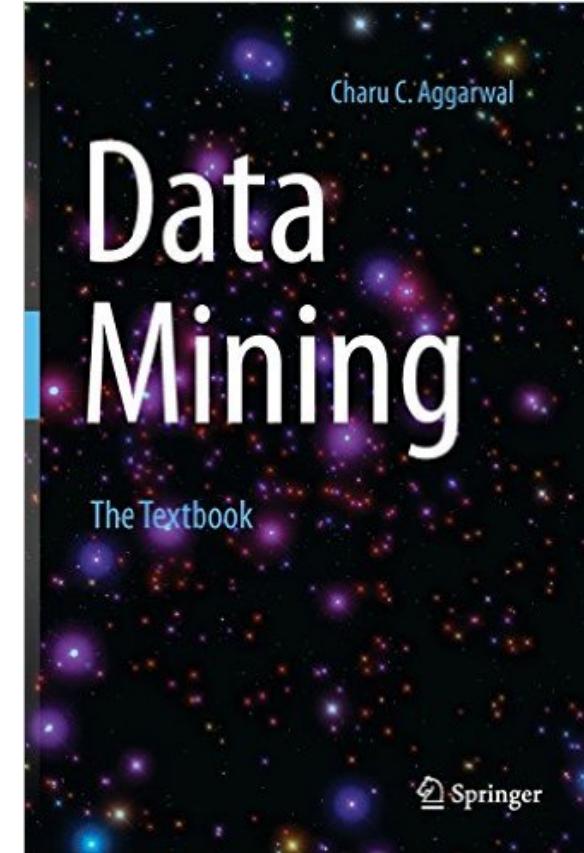
Good for business focus



Tom Fawcett, Foster Provost,  
“Data Science for Business”,  
O’Reilly Media, 2013  
(ISBN 1449361323)

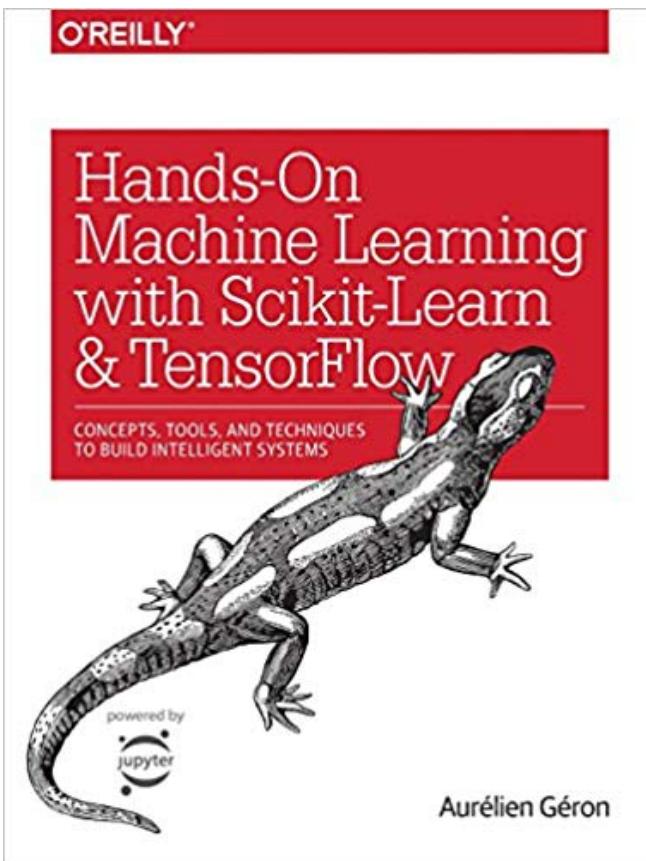
See the errata for the  
book [here](#).

Purchasing any of those is neither  
necessary nor compulsory

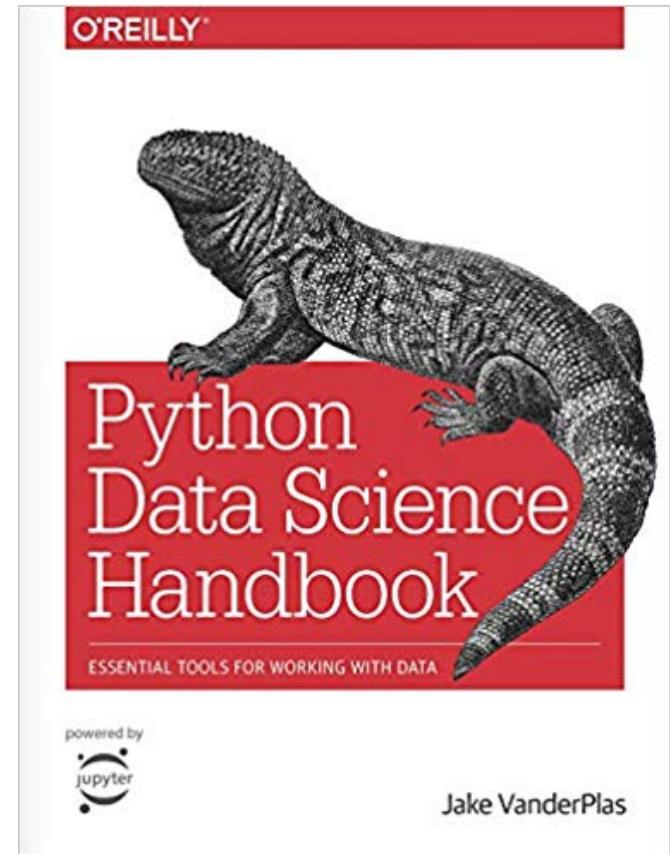


**Data Mining: The  
Textbook** by Charu  
C. Aggarwal.

# Recommended Reading



More hands-on introduction  
with Python code.



# If you have a question

- Ask your neighbor classmate
- Ask Google
- Post the question at the class slack channel
  - <https://bit.ly/33tNiPR> (click to join, but **use your UNIL email**)
  - The channel will be monitored daily, and it is **also your responsibility** as a student to visit it at least a few times per week and interact with your classmates. You get better, by helping others...
- Email me directly only for personal matters

# Cheating Policy

- Respect academic honesty.
- Don't cheat. As simple as that! If you are caught cheating (e.g., blatantly copying something that exists in the Internet) you will not receive any grade on your assignment and have further repercussions for your studies.
- Our goal is to learn in this class.
- If for an assignment you discussed with another classmate, this is OK. Just write it explicitly in your writeup: "I discussed question 5 with my friend John".

# Classroom behavior

- Put your phones in **silent mode** during class/lab. Imagine if everyone's phone was ringing at random times! In case you are a repeat offender (> 2 times), come and see me after the session to explain.
- You are encouraged to ask questions and initiate discussions during the class.



# Applications of Data Mining and Machine Learning

# Promotions/discounts at Supermarkets



# Sending out selected brochures/emails

Don't just send brochures/emails to all clients. Only to those "most likely" to buy something.

The image displays two magazine covers side-by-side, illustrating how to target specific audiences through direct mail.

**Left Cover: Coop Coopération**

This yellow-toned magazine cover features a woman in a tan leather jacket holding a large glass trophy. The headline reads: "Virginie Fajne Elle vise une médaille aux JO". A callout bubble says: "Notre supplément Tout sur les consoles de jeux". The logo "coop" is at the top left, and "Coopération N° 49 du 3 décembre 2013" is at the bottom left. The right side has a blue background with a woman's face and text: "Zoom sur l'égalité romaine". A small vertical column on the left edge says "JAA 1052 SUR RÉPONSE CENTRAL".

**Right Cover: MIGROS MAGAZINE**

This blue-toned magazine cover features a woman wearing yellow rubber gloves, smiling. The headline reads: "La plainte des gorges du Taubenloch" and "Le lynx: sa conquête du territoire à pas feutrés". Other visible text includes "Art-thérapie p. 14 | Fraises p. 43 | Résilience p. 88 | Photo p. 102", "Titou Lecoq", "L'égalité, ça passe aussi par les tâches ménagères", "Page 8", and "Page 18". The logo "MIGROS MAGAZINE MM16, 16.4.2018 www.migrosmagazine.ch" is at the top left. A small vertical column on the left edge says "Photo: Alain Bernhardou".

# Recommendations at Netflix

Would you have bought as many products, without the recommendations?



< Back to results

**BLACK+DECKER LDX120C 20V MAX Lithium**  
by BLACK+DECKER  
★ ★ ★ ★ ★ 5,112 customer reviews | 976 answered questions

Price: \$49.00  
Free Amazon Tech support included

Style: 20V MAX\* Drill/Driver

Drill w/ 10 Piece Drill Bit Set	20V MAX* Drill/Driver
\$55.24	<b>\$49.00</b>
Drill w/ Circular Saw	Drill w/ Extra Battery
\$91.36	--

- Ideal for drilling into wood, metal plastic and all screw driving tasks
- Chuck Size: 3/8 in; Clutch Setting: 11; Included Components: (1) LD
- Double Ended Bit; Power Source: Cordless
- Lithium Ion Technology & 20V MAX: Lighter, more compact, no men
- 11 Position Clutch: Provides precise control for drilling into wood, m
- Compact and Lightweight: Less fatigue and allows users to drill / sci Demands Attention
- Variable Speed: Allows countersinking without damaging material

[See more product details](#)

[Compare with similar items](#)

[Used & new \(28\) from \\$49.00 Details](#)

Click image to open expanded view

## Frequently bought together



# Detecting Fake News

- Fake news = false, often sensational, information disseminated under the guise of news reporting

Pope Francis endorses Trump for president

All Images News Videos Shopping More Settings To

About 1'600'000 results (0.31 seconds)

'Fake news' went viral in 2016. This expert studied who clicked.

<https://www.nbcnews.com/.../fake-news-went-viral-2016-expert-studied-who-clicked-...> ▾

Jan 14, 2018 - But before "fake news" became President Donald Trump's favorite ... articles like "Pope Francis Shocks World, Endorses Donald Trump For ..."

[Pope Francis Shocks World, Endorses Donald Trump for President ...](#)

[endingthefed.com/pope-francis-shocks-world-endorses-donald-trump-for-president-rele...](http://endingthefed.com/pope-francis-shocks-world-endorses-donald-trump-for-president-rele...)

No information is available for this page.

[Learn why](#)

[Pope Francis Shocks World, Endorses Hillary Clinton for President](#)

<https://www.snopes.com/fact-check/junk-news/> ▾

Claim. Pope Francis has endorsed Hillary Clinton for President. ... reversed himself yet again and endorsed Trump's rival in the presidential race, Hillary Clinton:





**Stephen King**   
@StephenKing

 Follow

The news is real. The president is fake.

6:38 PM - 9 Jul 2017



59,919



187,721



# Launching a new product with Big Data

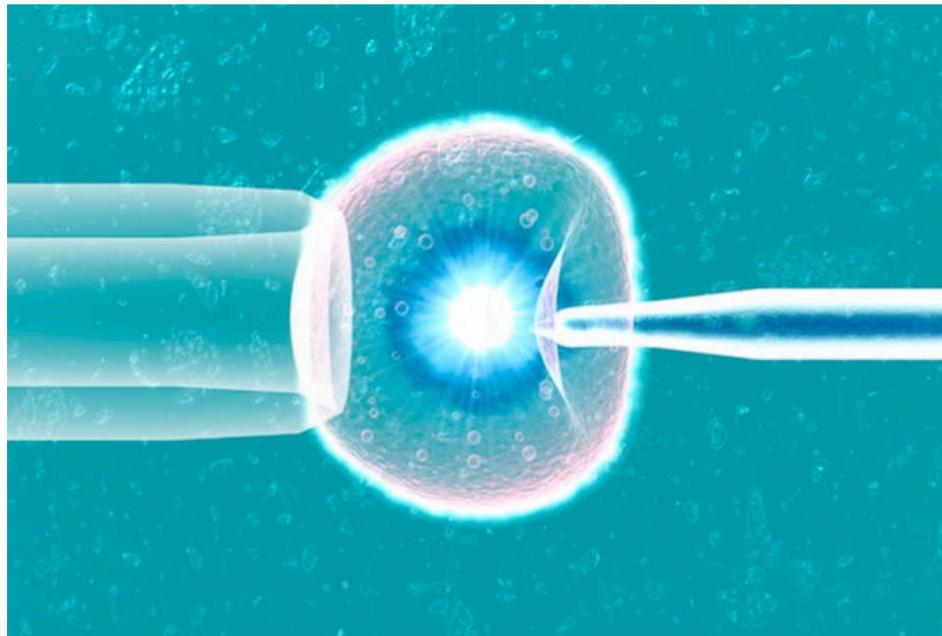
Starbucks was introducing a new coffee product but was concerned that customers would find its taste too strong. The morning that the coffee was rolled out, Starbucks monitored blogs, Twitter, and niche coffee forum discussion groups to assess customers' reactions. By mid-morning, Starbucks discovered that although people liked the taste of the coffee, they thought that it was too expensive. Starbucks lowered the price, and by the end of the day all of the negative comments had disappeared.



# Which egg to use? What AI Can Do for IVF

Up to two thirds of in vitro fertilization patients experience failed cycles—but AI systems might be able to flag the most viable embryos far better than humans can

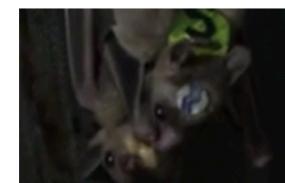
By Carol Lynn Curchoe, Charles L. Bormann on January 10, 2018



## LATEST NEWS



Mucus: The Body's Unsung Hero



Male Bats Up Mating Odds with Mouth Morsels



New NASA Mission Will Fly Titan's Frigid Skies to Search for Life's Beginnings

# Detecting a fall with Apple Watch



Fall

Trip

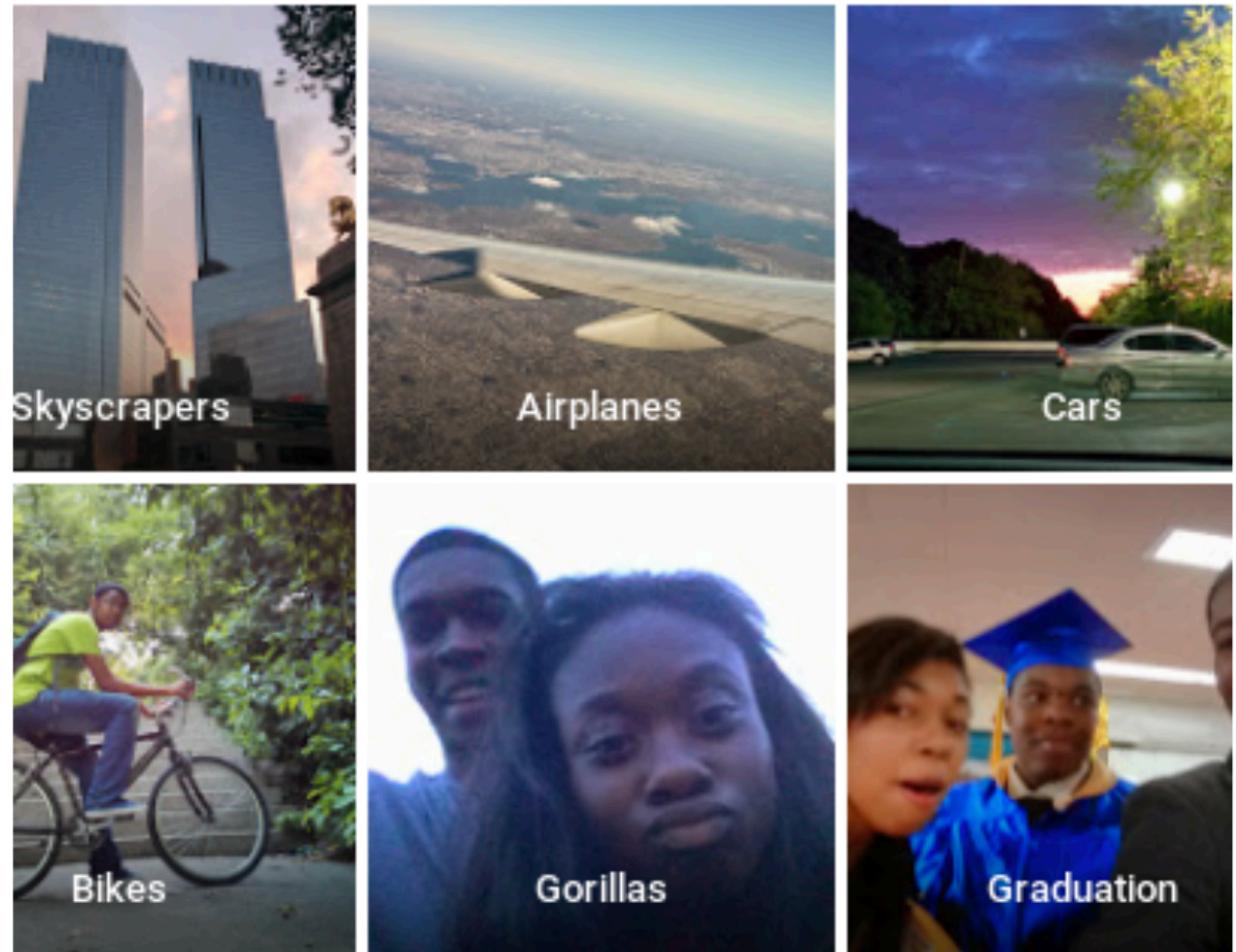
Slip

# How about pitfalls?

- The previous were all success stories. Has there been any pitfalls?
- Algorithmic Bias

(read the above article at  
a later time and discuss  
in the class slack channel)

<https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai>



# There are numerous mining or “learning” tasks

The majority of the previous examples were **classification** tasks, but there are many more

- Regression
- Clustering
- Outlier Detection
- Association Rules
- Recommendations
- Similarity Search

# Data Mining – Why now?



Arthur C Clarke  
~ 1974

Why? Why now? Confluence of 4 technical advances:

## 1. Storage

- Disk densities have been doubling each year
- A \$100 disk today has 1,000,000× more capacity/\$ than the disks of 30 years ago

## 2. Networking / Internet

- Data can be transferred easily between collection, storage, and use
- eBusiness systems do it as a matter of course (w/ fewer data errors)

## 3. Algorithms

- Advanced algorithms from machine learning, pattern recognition, and applied statistics have become mature enough for mainstream use

## 4. Computing power

- Processing power has been doubling every 1.5 years or so (Moore's law)
- Laptops are more powerful than the supercomputers of yesteryear

All four of these are essential for effective, successful data mining

# Data-mining

## 5. Data

We never generated and collected more data than now

“More data was created in the past two years than in the entire previous history of mankind.”

# “Big data” is “data mining”

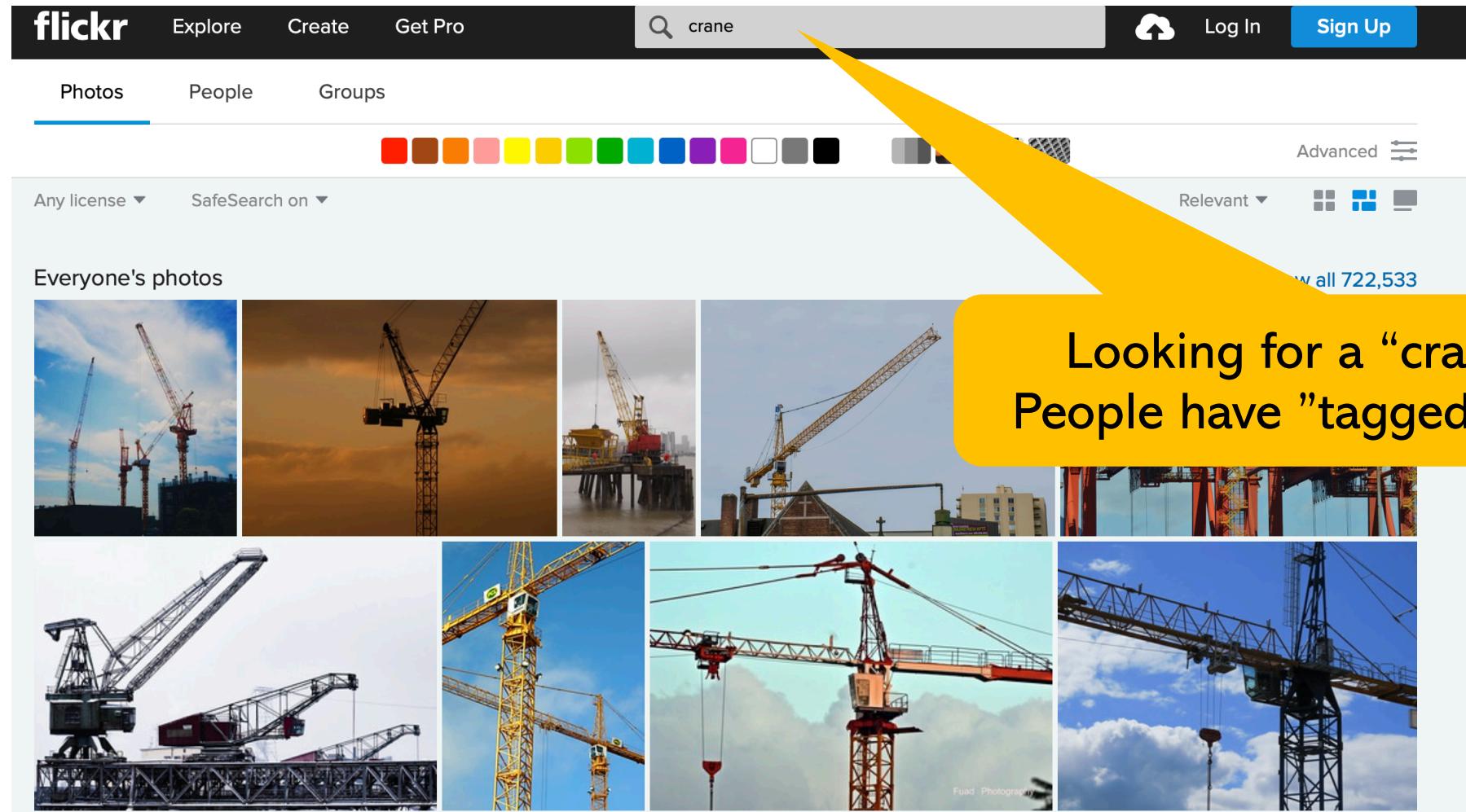


# Data-mining

54

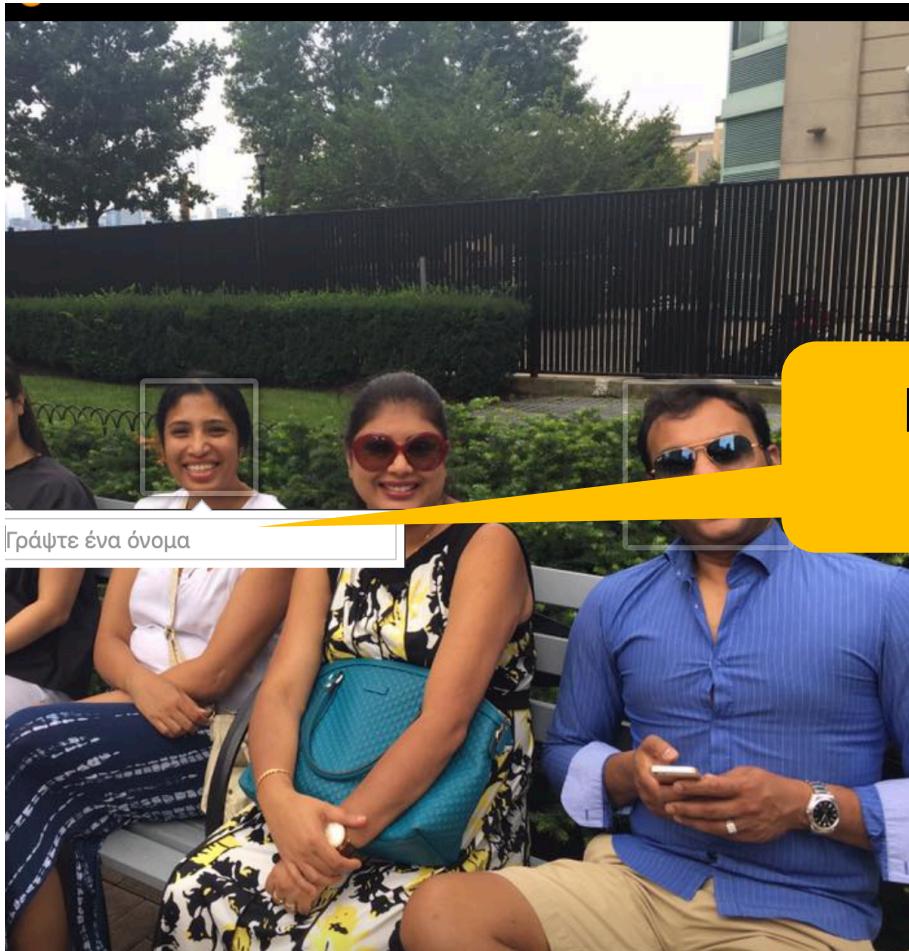
## 6. Users

Learning algorithms need training (“labeled”) data. We/you provide those!



## 6. Users

Learning algorithms need training (“labeled”) data. We/you provide those!



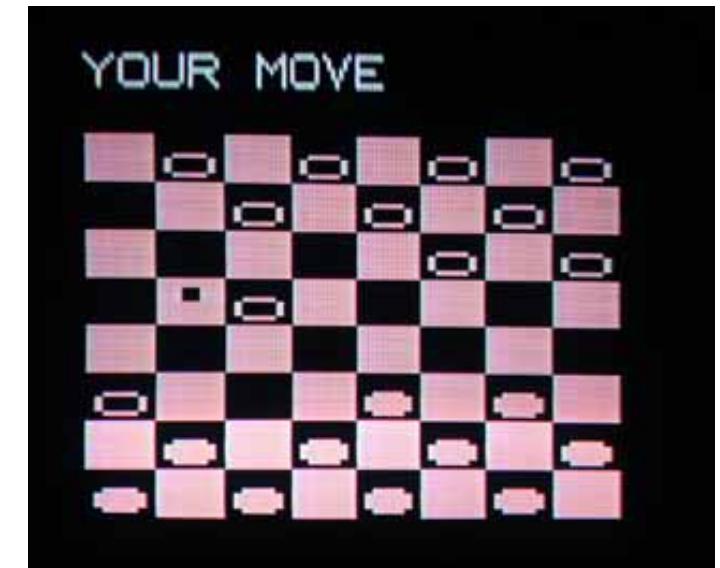
Facebook urges you to “write a name” and label the face!

# A Brief History of Data Mining and Machine Learning

- 1700s: Bayes theorem: understand complex realities based on probabilities. It is to the theory of probability what the Pythagorean theorem is to geometry.
- 1800s: Regression analysis.
- 1950: “[Computing Machinery and Intelligence](#)”, aka Turing’s Test
- 1956: Term “Artificial Intelligence” introduced at Dartmouth Conference
- 50s-60s: Golden years
  - 50s: neural networks, genetic algorithms, clustering
  - 60s: decision trees

# Samuel's checkers program

- 1959: [Arthur Samuel's](#) computer program learns to play checkers better than humans, given past games as examples.



- 1700s: Bayes theorem: understand complex realities based on probabilities. It is to the theory of probability what the Pythagorean theorem is to geometry
- 1800s: Regression analysis
- 1950: “Computing Machinery and Intelligence”, aka Turing’s Test
- 1956: Artificial Intelligence term introduced at Dartmouth Conference
- 50s-60s: Golden years
  - 50s: neural networks, genetic algorithms, clustering
  - 60s: decision trees
- 70s-80s: AI winter

- 1988: “A statistical approach to Language Translation”
- 1989: First conference on Data Mining (KDD)
- 1991: The birth of Internet
- 1994: Association Rules
- 1997: Deep Blue defeats Garry Kasparov
- 2001: Product recommendations on Amazon
- 2011: IBM Watson wins human at TV game Jeopardy!
- 2012: Power of deep Learning
- 2014: Neural Machine Translation
- 2015: Machines “see” better than humans
- 2016: Google’s AlphaGo (deep learning, reinforcement learning)
- 2018: Self-driving taxis

Google's DeepMind project "AlphaGO", a computer program that plays the board game 'GO' has defeated the world's number one Go player Ke Jie



## THE ULTIMATE GO CHALLENGE

GAME 3 OF 3

27 MAY 2017



vs



AlphaGo

*Winner of Match 3*

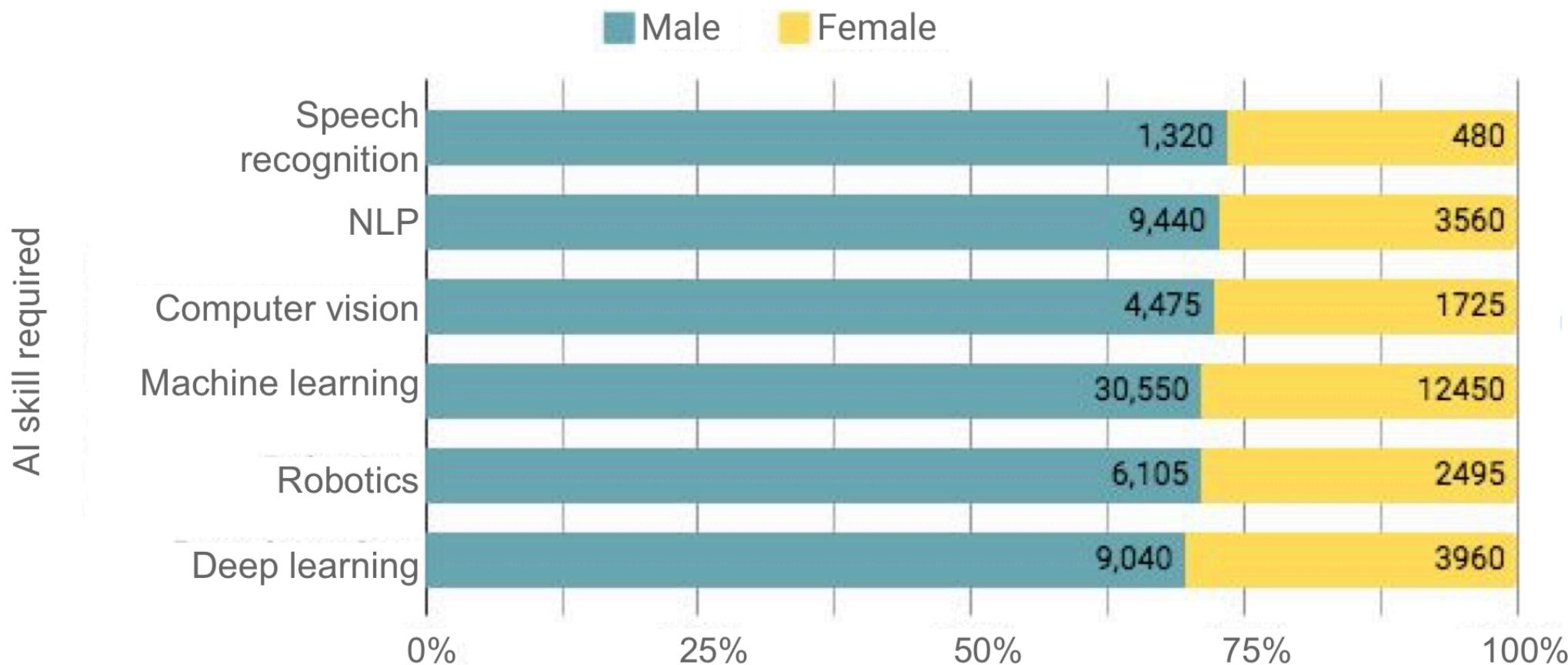
Ke Jie

**RESULT B + Res**

- 1988: “A statistical approach to Language Translation”
- 1989: First conference on Data Mining (KDD)
- 1991: The birth of Internet
- 1994: Association Rules
- 1997: Deep Blue defeats Garry Kasparov
- 2001: Product recommendations on Amazon
- 2011: IBM Watson wins human at jeopardy
- 2012: Power of deep Learning
- 2014: Neural Machine Translation
- 2015: Machines “see” better than humans
- 2016: Google’s AlphaGo (deep learning, reinforcement learning)
- 2018: Self-driving taxis
- **2019: Class on “Data Mining and ML” (your personal milestone)**

# Current state of ML/AI

- If you want to see where Machine Learning stands today and to understand its growth, you can visit [aiindex.org](http://aiindex.org)



# Techniques

Regression, Classification, Outlier/Anomaly Detection,  
Association Rules, Recommendation

# Finding the price of a house

CHF 500'000



?



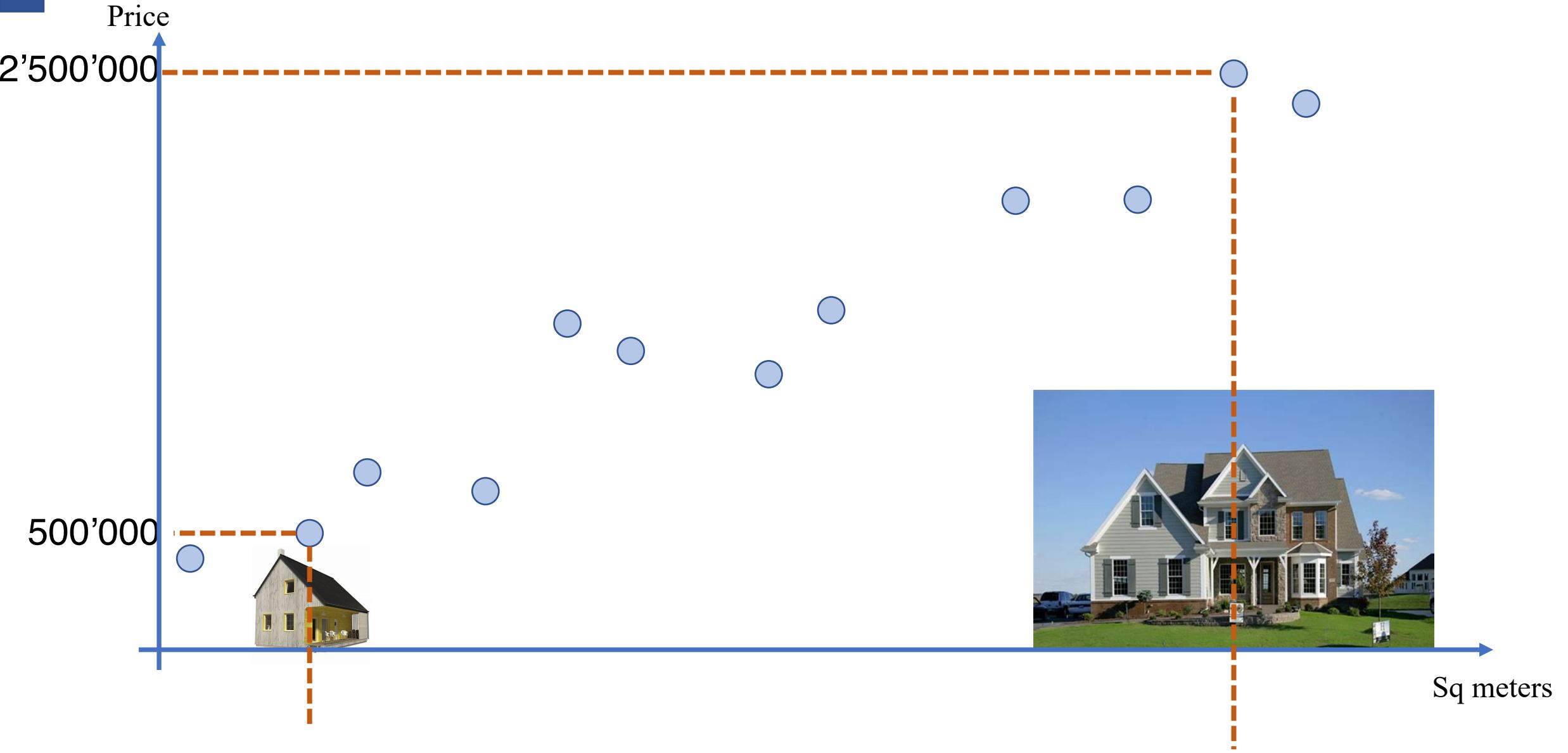
CHF 2'500'000



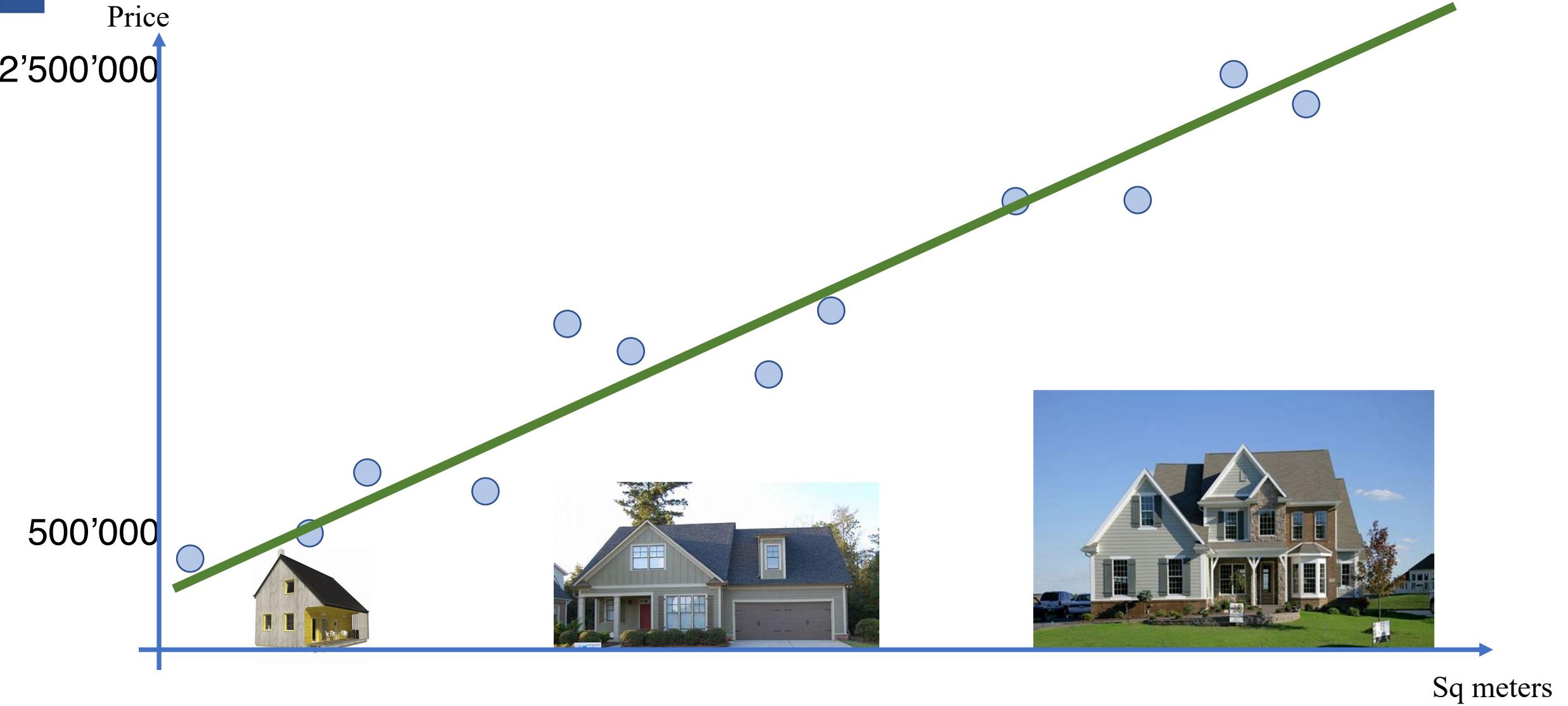
# Finding the price of a house



# Finding the price of a house

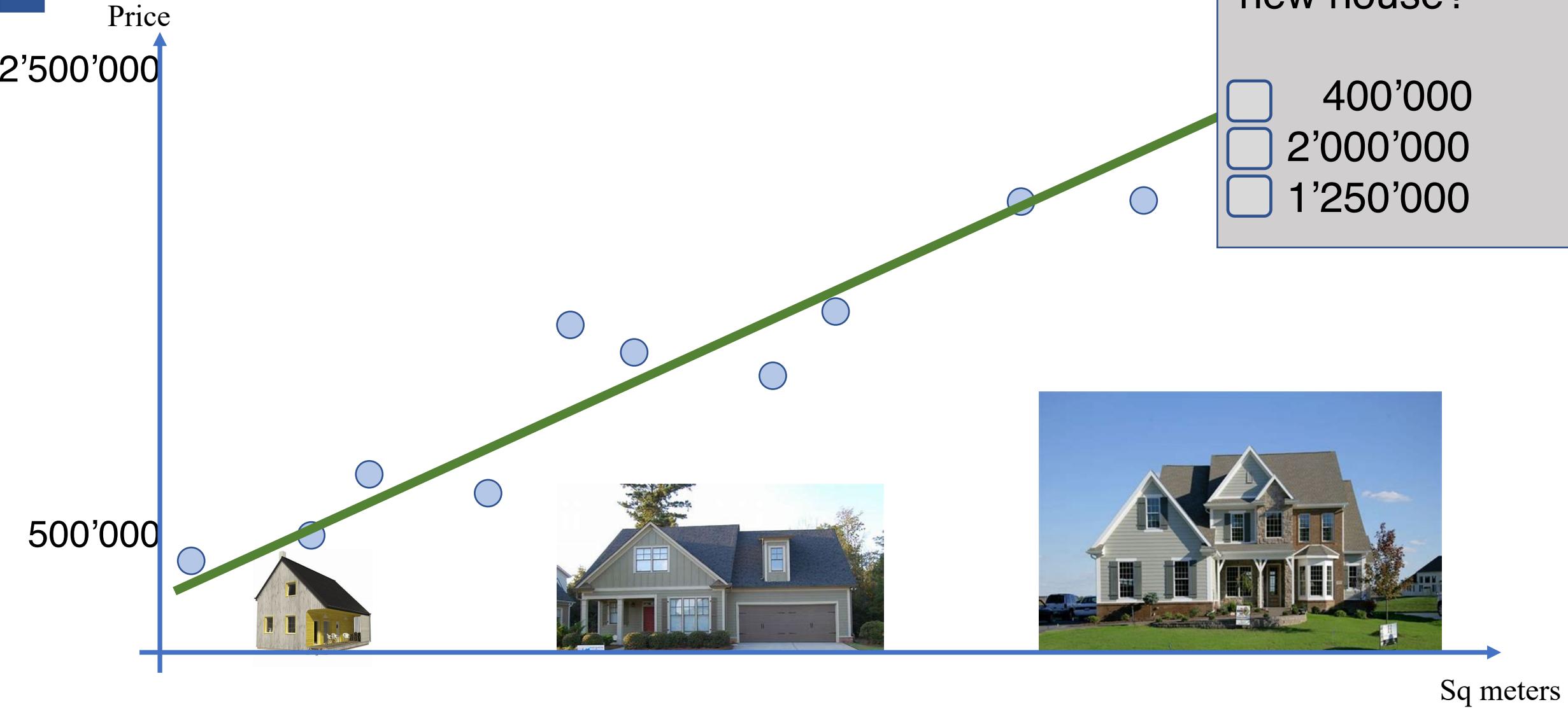


# Finding the price of a house

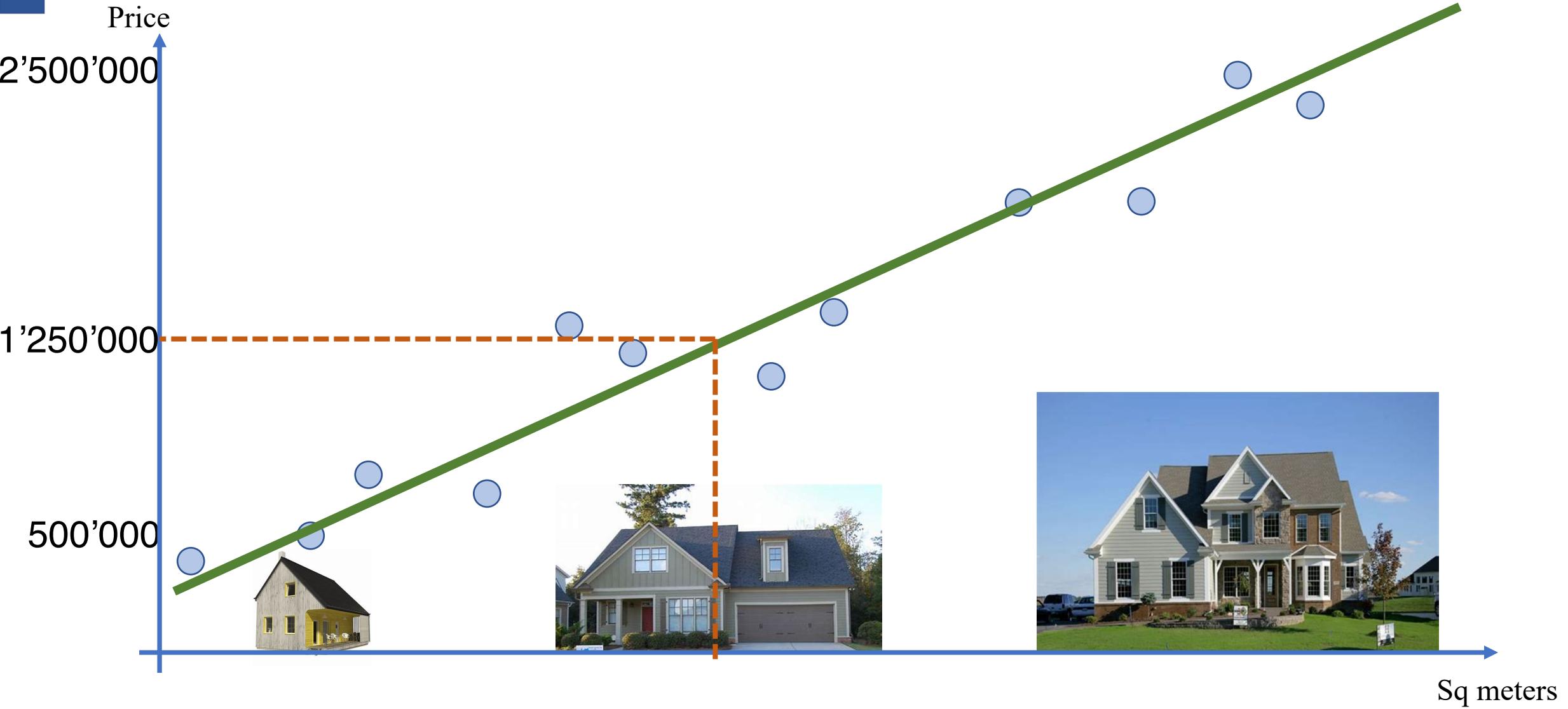


# Finding the price of a house

What is your estimate for the new house?



# Finding the price of a house



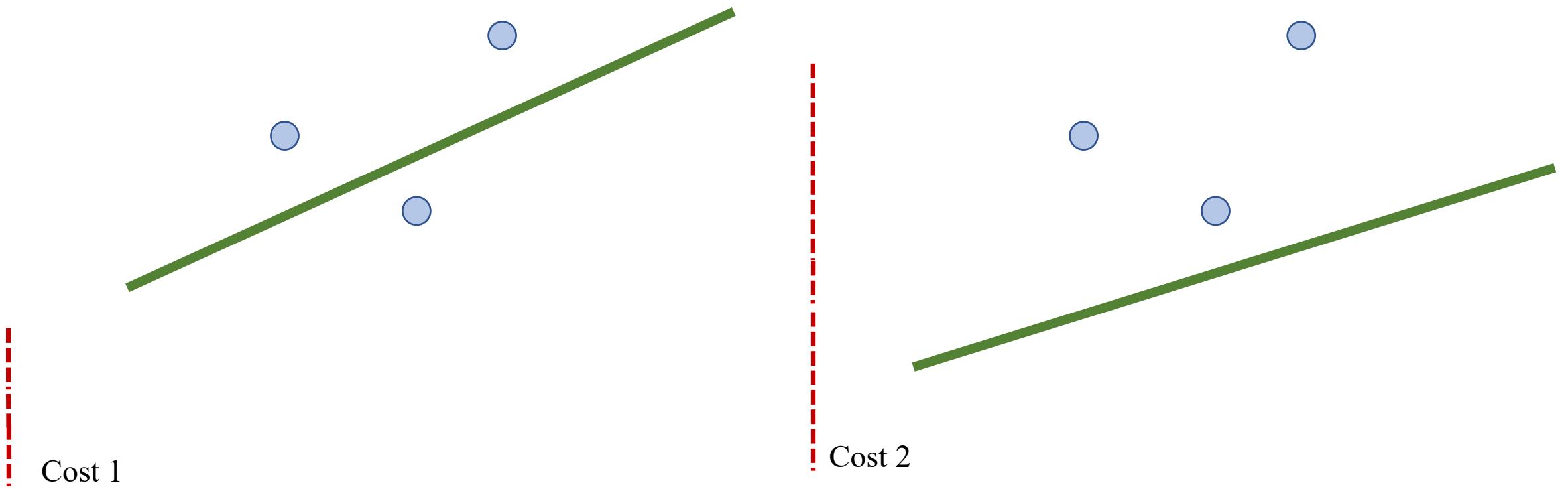
# How do we find this best line?

- We try different lines and keep the one that minimizes some cost function.



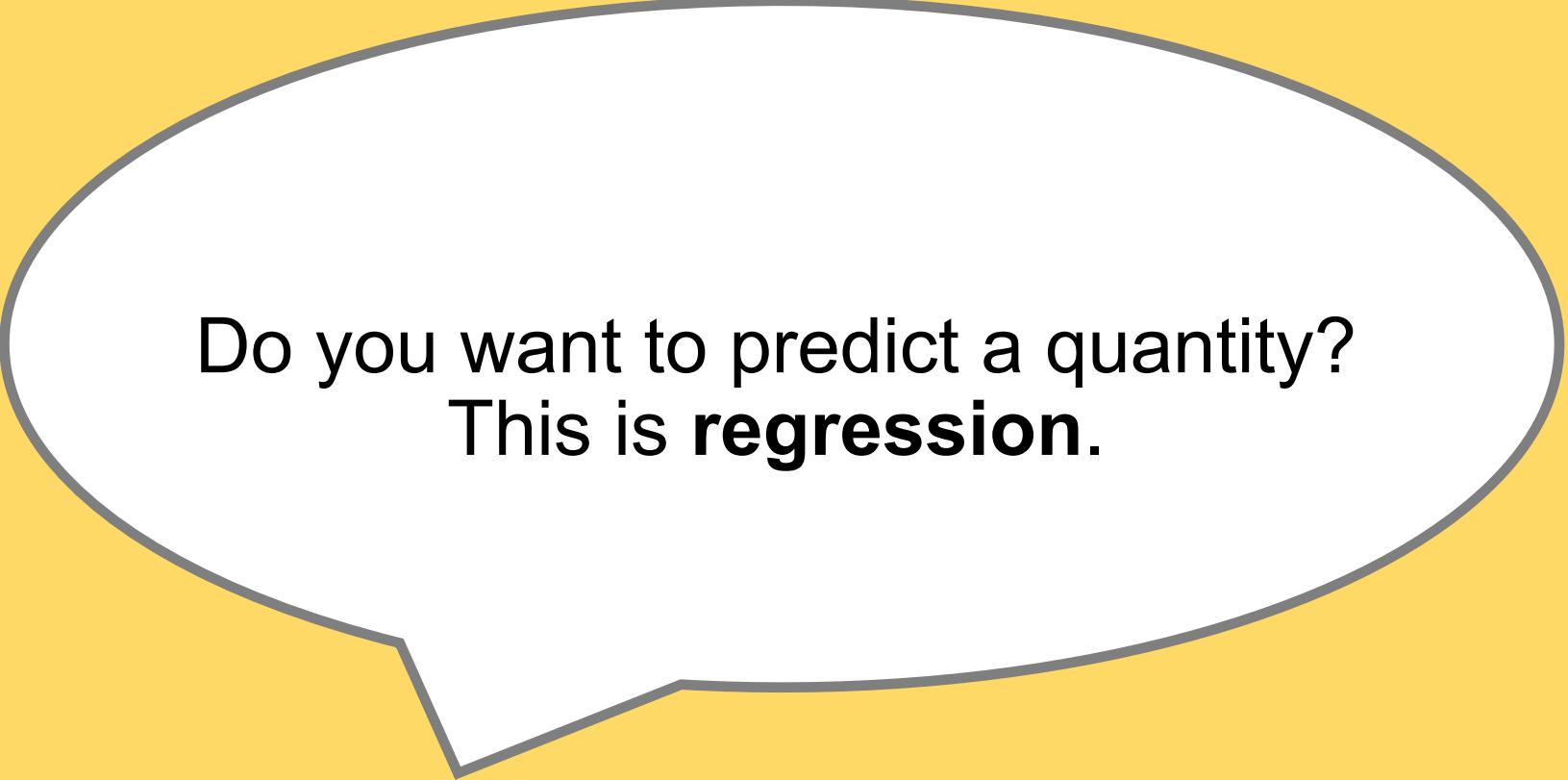
# How do we find this best line?

- We try different lines and keep the one that minimizes some cost function.



# How do we find this best line?

- We don't just try them randomly.
- We use an approach called “gradient descent” that steers the line towards the most promising directions.



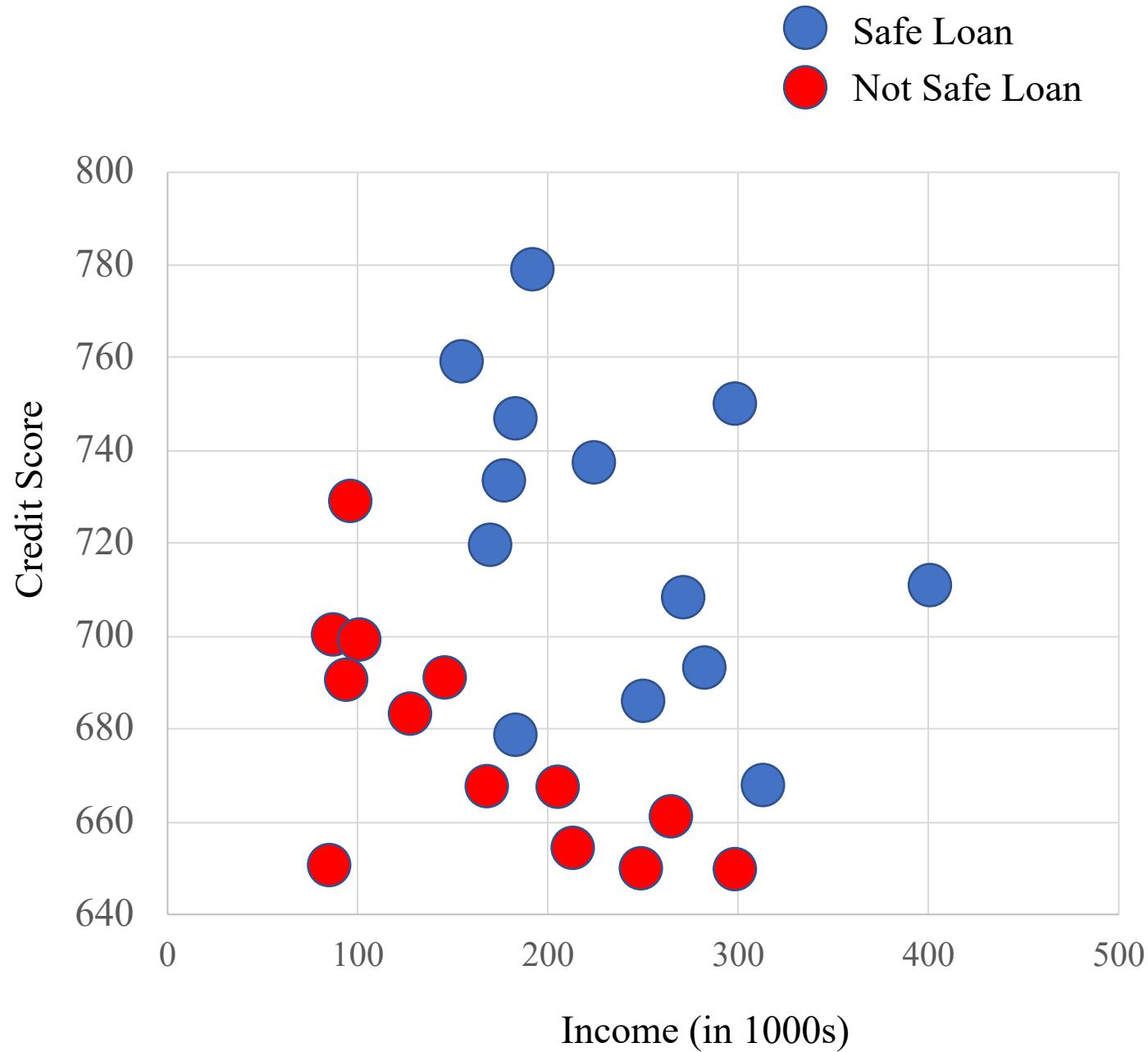
Do you want to predict a quantity?  
This is **regression**.

# Give a loan or not?

Your company a bank gives you data of past clients in the form:

(credit score, income)  
→ Safe Loan or Not Safe

Can you use such data to find if a new customer should be given a loan or not?



# Give a loan or not?

Your company a bank gives you data of past clients in the form:

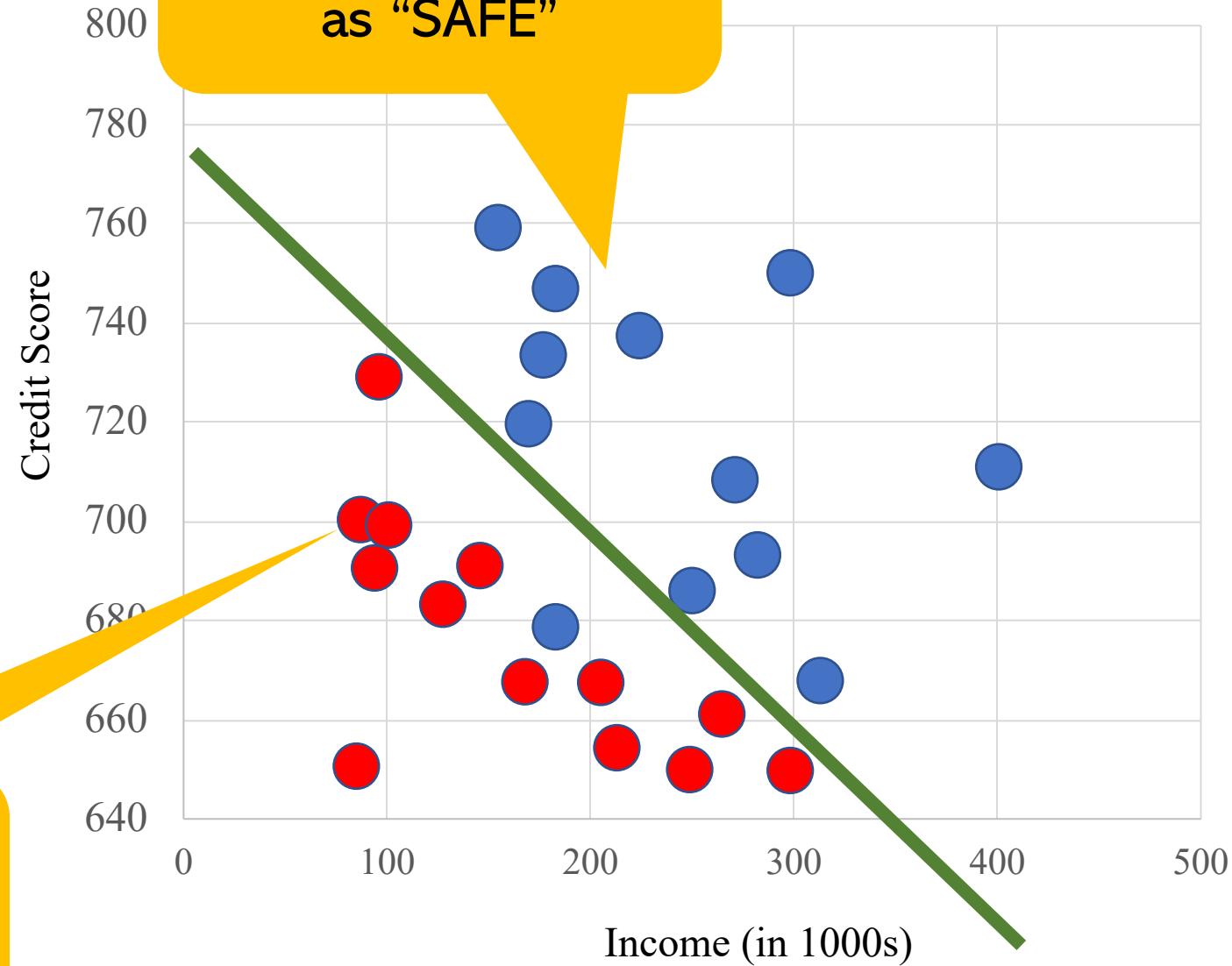
(credit score, income)

→ Safe Loan or Not Safe

Can you use such data to find if a new customer should be given a loan or not?

Anything below the line is categorized as “NOT SAFE”

Anything above the line is categorized as “SAFE”



# Give a loan or not?

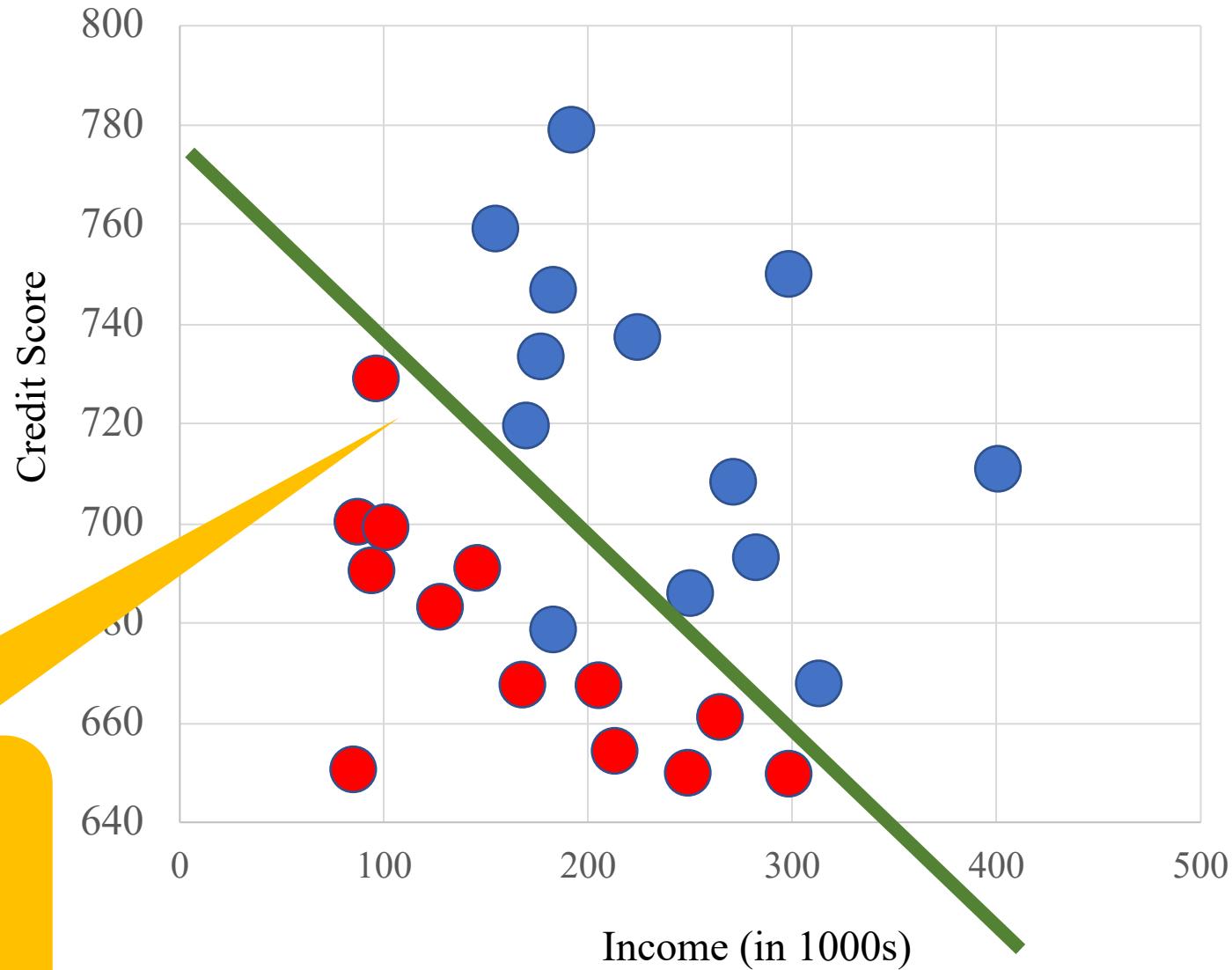
Your company a bank gives you data of past clients in the form:

(credit score, income)

→ Safe Loan or Not Safe

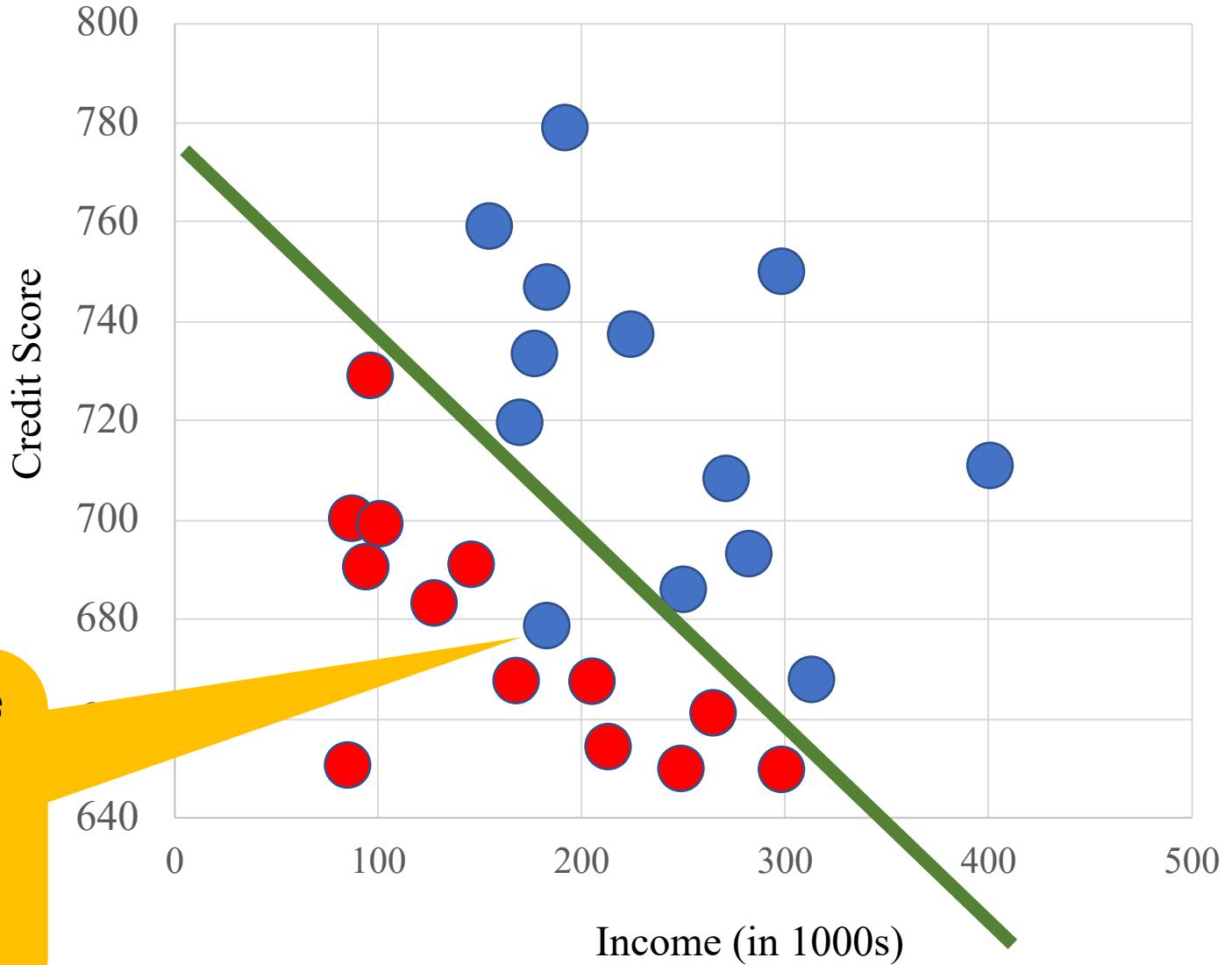
Can you use such data to find if a new customer should be given a loan or not?

This is an example  
of a “Logistic  
Regression”



# Give a loan or not?

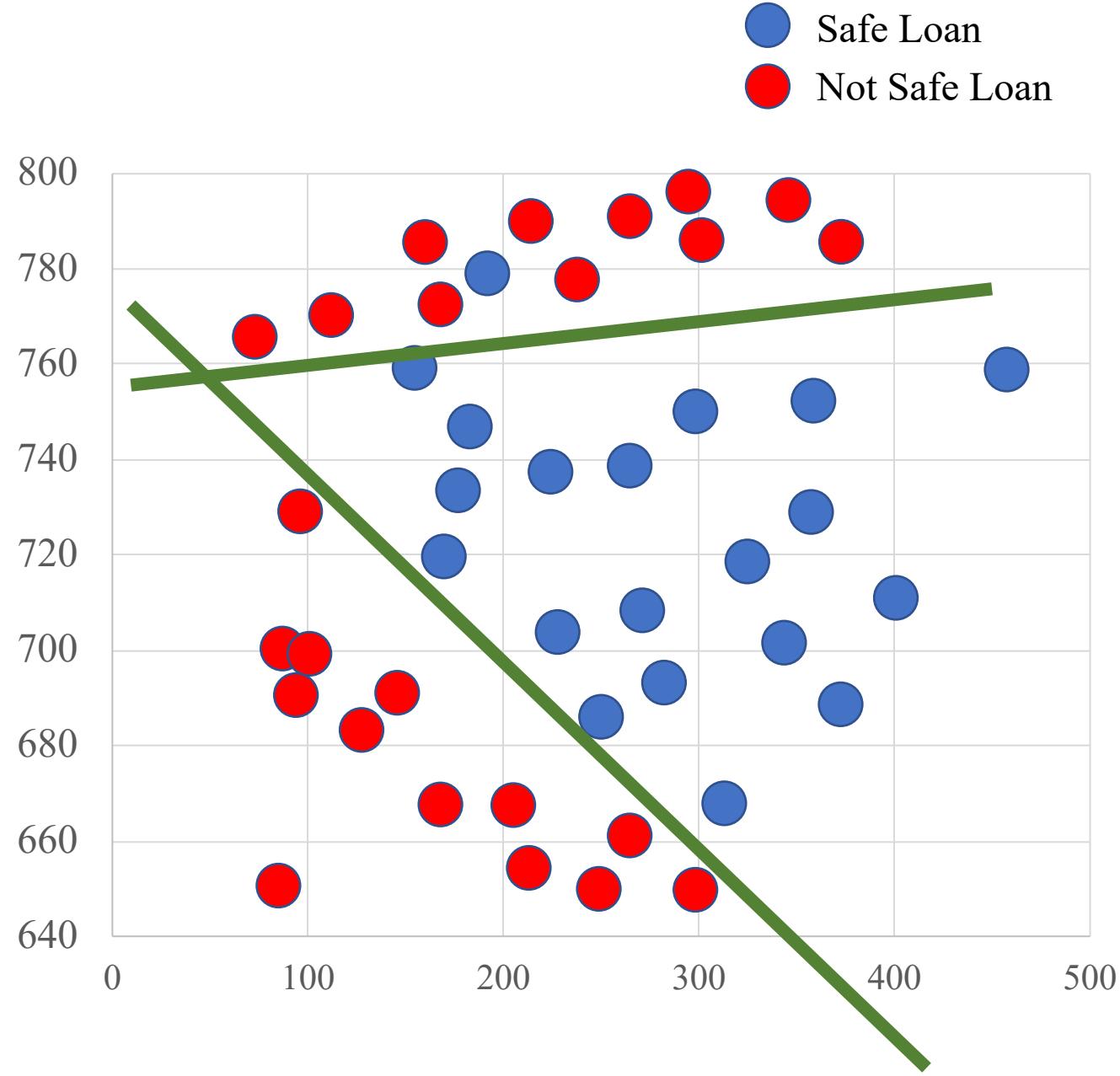
Safe Loan  
Not Safe Loan



You may have some “misclassifications” or errors. The goal is to find a line that has a fewest errors.

# Neural Networks

If we combine many of these functions, then we can solve more complex categorization problems, like the one on the right.



Do you want to predict a category  
(something discrete)?  
That's **classification**.

I recently received  
this email  
in my inbox.

Because I logged in  
to my account from  
a location I haven't  
previously.

## Microsoft account unusual sign-in activity ➤ Inbox ×

**Microsoft account team** <account-security-noreply@accountprotection.microsoft.com>  
to me ▾

Microsoft account

# Unusual sign-in activity

We detected something unusual about a recent sign-in to the Microsoft account [mi\\*\\*\\*\\*\\*@outlook.com](mailto:mi*****@outlook.com).

#### Sign-in details

Country/region: Macao SAR

IP address: 202.175.165.98

Date: 8/11/2019 2:57 PM (CET)

Platform: Mac OS

Browser: Safari

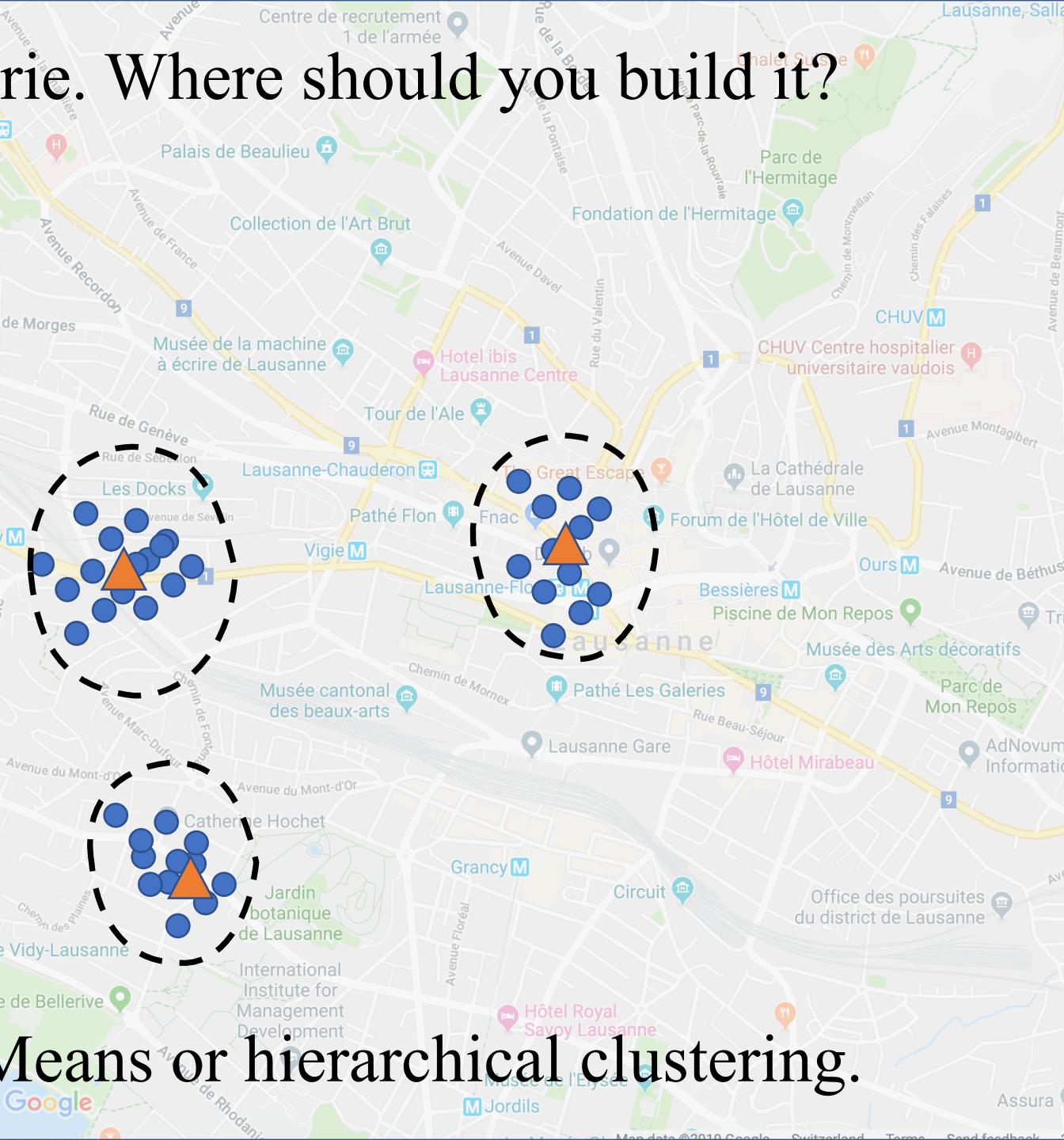
Please go to your recent activity page to let us know whether or not this was you. If this wasn't you, we'll help you prevent similar activity in the future.

[Review recent activity](#)

Do you want to predict something  
out of the ordinary? That's  
**anomaly or outlier detection.**

# You want to build a boulangerie. Where should you build it?

Circles show where people consume a lot of bread.

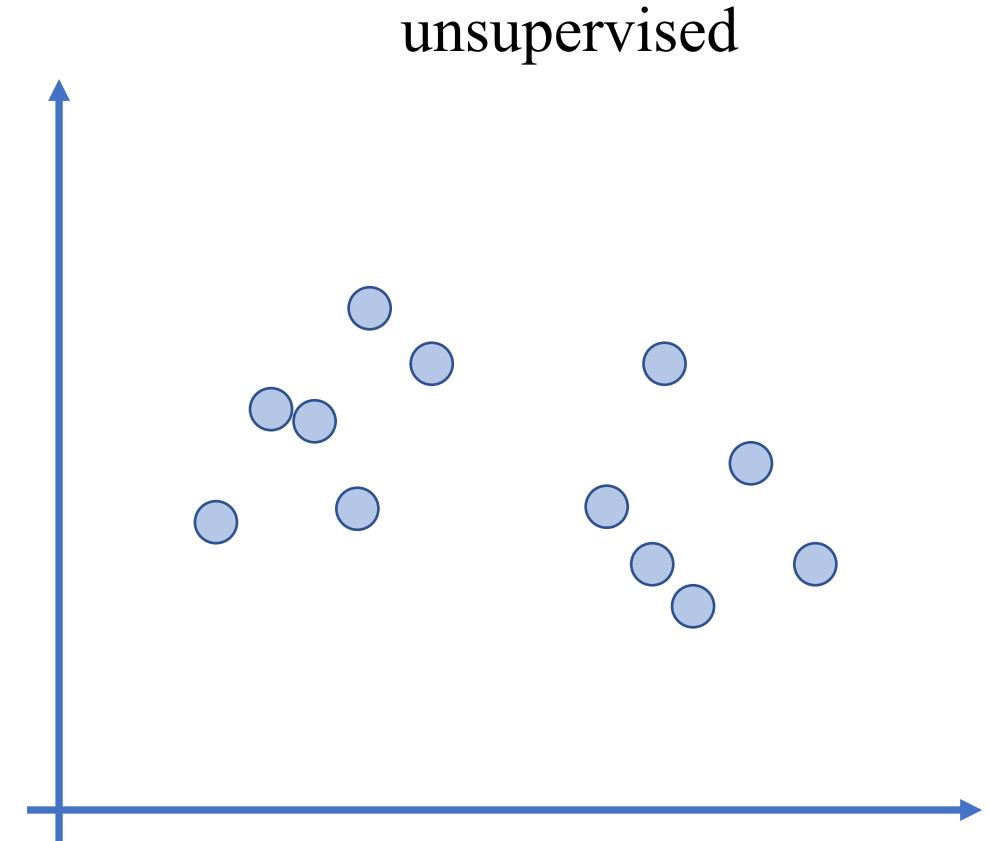
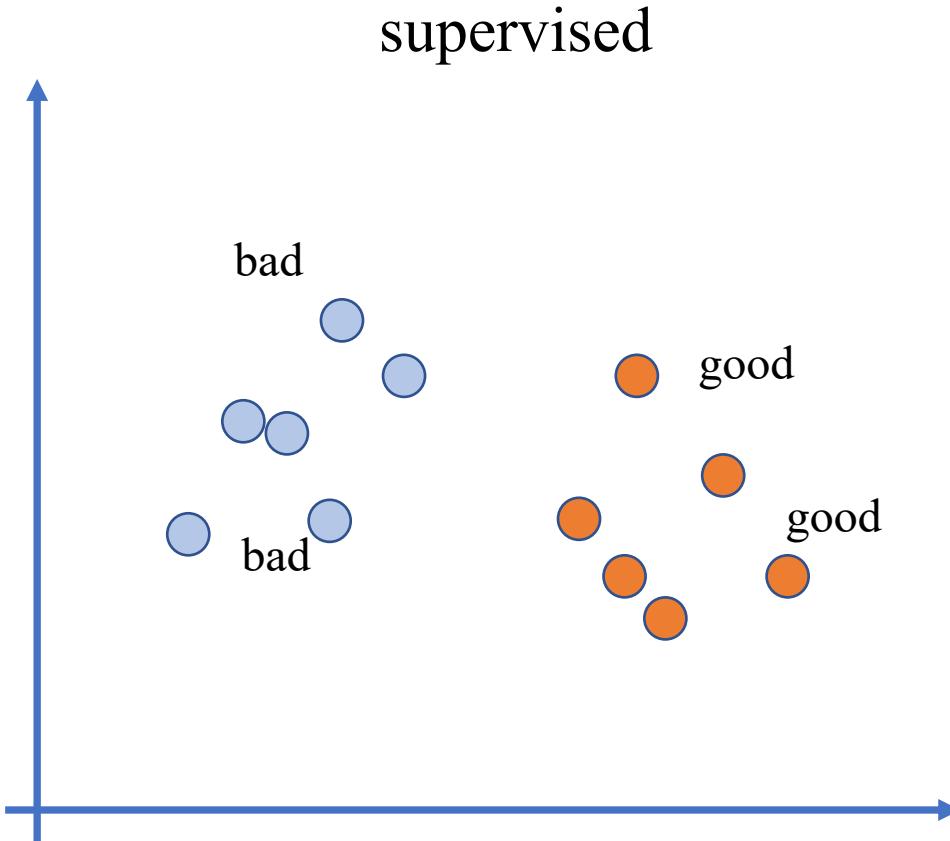


You can use algorithms such as k-Means or hierarchical clustering.

Do you want to discover structure  
in unexplored (unlabeled) data?  
That's **clustering**.

# Supervised vs Unsupervised Learning

- Find structure in “unlabeled” data
- We are not giving to the algorithm the “right” answer



# Example of unsupervised learning

≡ Google News



Search for topics, locations & sources

Top stories

For you

Favorites

Saved searches

U.S.

World

Business

Technology

Entertainment

## Headlines

More Headlines

### 9 killed in Ohio in second U.S. mass shooting within 24 hours

The Washington Post · 1 hour ago



- Special Report: Two Mass Shootings In 13 Hours Leave 29 People Dead | NBC News  
    ▶ NBC News · 3 hours ago
- Special Report: Nine people killed in Dayton, Ohio, shooting  
    ▶ CBS News · 4 hours ago
- Governor DeWine orders flags at half mast following mass shooting in Dayton  
    WJW FOX 8 News Cleveland · 4 hours ago · Local coverage
- Gov. DeWine orders flags at half-staff for victims of Dayton mass shooting  
    WLWT Cincinnati · 1 hour ago

View full coverage

# Example of unsupervised learning

Google News

Search for topics, locations & sources

Top stories

For you

Favorites

Saved searches

U.S.

World

Business

Technology

Entertainment

## Headlines

### 9 killed in Ohio in second U.S. mass shooting within 24 hours

The Washington Post · 1 hour ago

- Special Report: Two Mass Shootings In 13 Hours Leave 29 People Dead | NBC News  
NBC News · 3 hours ago
- Special Report: Nine people killed in Dayton, Ohio, shooting  
CBS News · 4 hours ago
- Governor DeWine orders flags at half mast following mass shooting in Dayton  
WJW FOX 8 News Cleveland · 4 hours ago · Local coverage
- Gov. DeWine orders flags at half-staff for victims of Dayton mass shooting  
WLWT Cincinnati · 1 hour ago

[View full coverage](#)

More Headlines



## Governor DeWine orders flags at half mast following mass shooting in Dayton

POSTED 6:55 AM, AUGUST 4, 2019, BY TALIA NAQUIN, UPDATED AT 11:03AM, AUGUST 4, 2019

[FACEBOOK](#)

[TWITTER](#)

[PINTEREST](#)

Governor DeWine orders flags at half mast following mass shoot...

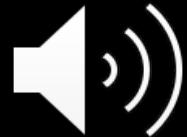
YouTube CH

CBS NEWS | SPECIAL REPORT

SHOOTINGS IN DAYTON AND EL PASO

0:12 / 1:46

QUIZ  
Time



# In-Class Exercise

7 mins

Find at least two more examples of a data analytics success story or a startup in the field of data mining/machine learning/data science

- Business Model: How do they make money?
- What do they do? What is special/innovative?

# The Data & the Model

# Typical data you should expect

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	OnlineSecurity	...	TotalCharges	Churn
0	7590-VHVEG	Female	0	Yes	No	1	No	No	...	29.85	No
1	5575-GNVDE	Male	0	No	No	34	Yes	Yes	...	1889.5	No
2	3668-QPYBK	Male	0	No	No	2	Yes	Yes	...	108.15	Yes
3	7795-CFOCW	Male	0	No	No	45	No	Yes	...	1840.75	No
4	9237-HQITU	Female	0	No	No	2	Yes	No	...	151.65	Yes

# Typical data you should expect

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	OnlineSecurity	...	TotalCharges	Churn
0	7590-VHVEG	Female	0	Yes	No	1	No	No	...	29.85	No
1	5575-GNVDE	Male	0	No	No	34	Yes	Yes	...	1889.5	No
2	3668-QPYBK	Male	0	No	No	2	Yes	Yes	...	108.15	Yes
3	7795-CFOCW	Male	0	No	No	45	No	Yes	...	1840.75	No
4	9237-HQITU	Female	0	No	No	2	No	No	...	151.65	Yes

or “training example”

Observation = one row

# Typical data you should expect

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	OnlineSecurity	...	TotalCharges	Churn
0	7590-VHVEG	Female	0	Yes	No	1	No	No	...	29.85	No
1	5575-GNVDE	Male	0	No	No	34	Yes	Yes	...	1889.5	No
2	3668-QPYBK	Male	0	No	No	2	Yes	Yes	...	108.15	Yes
3	7795-CFOCW	Male	0	No	No	45	No	Yes	...	1840.75	No
4	9237-HQITU	Female	0	No	2	Yes	No	...	151.65	Yes	

Feature x = column  
(independent variables or predictors)

# Typical data you should expect

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	OnlineSecurity	...	TotalCharges	Churn
0	7590-VHVEG	Female	0	Yes	No	1	No	No	...	29.85	No
1	5575-GNVDE	Male	0	No	No	34	Yes	Yes	...	1889.5	No
2	3668-QPYBK	Male	0	No	No	2	Yes	Yes	...	108.15	Yes
3	7795-CFOCW	Male	0	No	No	45	No	Yes	...	1840.75	No
4	9237-HQITU	Female	0	No		2	Yes	No	...	151.65	Yes

All the features → X

# Typical data you should expect

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	OnlineSecurity	...	TotalCharges	Churn
0	7590-VHVEG	Female	0	Yes	No	1	No	No	...	29.85	No
1	5575-GNVDE	Male	0	No	No	34	Yes	Yes	...	1889.5	No
2	3668-QPYBK	Male	0	No	No	2	Yes	Yes	...	108.15	Yes
3	7795-CFOCW	Male	0	No	No	45	No	Yes	...	1840.75	No
4	9237-HQITU	Female	0	No	No	2	Yes	No	...	151.65	Yes

“High-dimensional data” =  
lots of columns

Number of all features =  
data dimensionality

# Typical data you should expect

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	OnlineSecurity	...	TotalCharges	Churn
0	7590-VHVEG	Female	0	Yes	No	1	No	No	29.85	No
1	5575-GNVDE	Male	0	No	No	34	Yes	Yes	1889.5	No
2	3668-QPYBK	Male	0	No	No	2	Yes	Yes	108.15	Yes
3	7795-CFOCW	Male	0	No	No	45	No	Yes	1840.75	No
4	9237-HQITU	Female	0	No	No	2	Yes	No	151.65	Yes

Target variable  
(dependent variable)

Regression problem

# Typical data you should expect

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	OnlineSecurity	...	TotalCharges	Churn
0	7590-VHVEG	Female	0	Yes	No	1	No	No	...	29.85	No
1	5575-GNVDE	Male	0	No	No	34	Yes	Yes	...	1889.5	No
2	3668-QPYBK	Male	0	No	No	2	Yes	Yes	...	108.15	Yes
3	7795-CFOCW	Male	0	No	No	45	No	Yes	...	1840.75	No
4	9237-HQITU	Female	0	No	No	2	Yes	No	...	151.65	Yes

Target variable y,  
dependent variable

Classification problem

# Our goal is to build a **model** to answer the business question

Build a model = Learn a function  $f$  that given  $\mathbf{X}$ , and  $\mathbf{y}$

(predicted  $\mathbf{y}$ ) =  $f(\mathbf{X})$

**In Regression:**  $\mathbf{y}$  will give real values

**In Classification:**  $\mathbf{y}$  will give 0,1 (..2, 3)

The function  $f$  can also be seen as a **set of rules**.

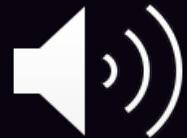
So, a model is nothing more than a set of rules, or a set of mathematical functions.

# Types of data mining tasks

- Similarity Matching
  - What other companies are like our best small business customers?
- Dimensionality Reduction
  - Compress/Summarize my data
- Description / Profiling
  - What does “normal behavior” look like?
- Clustering
  - Do my customers form natural groups?

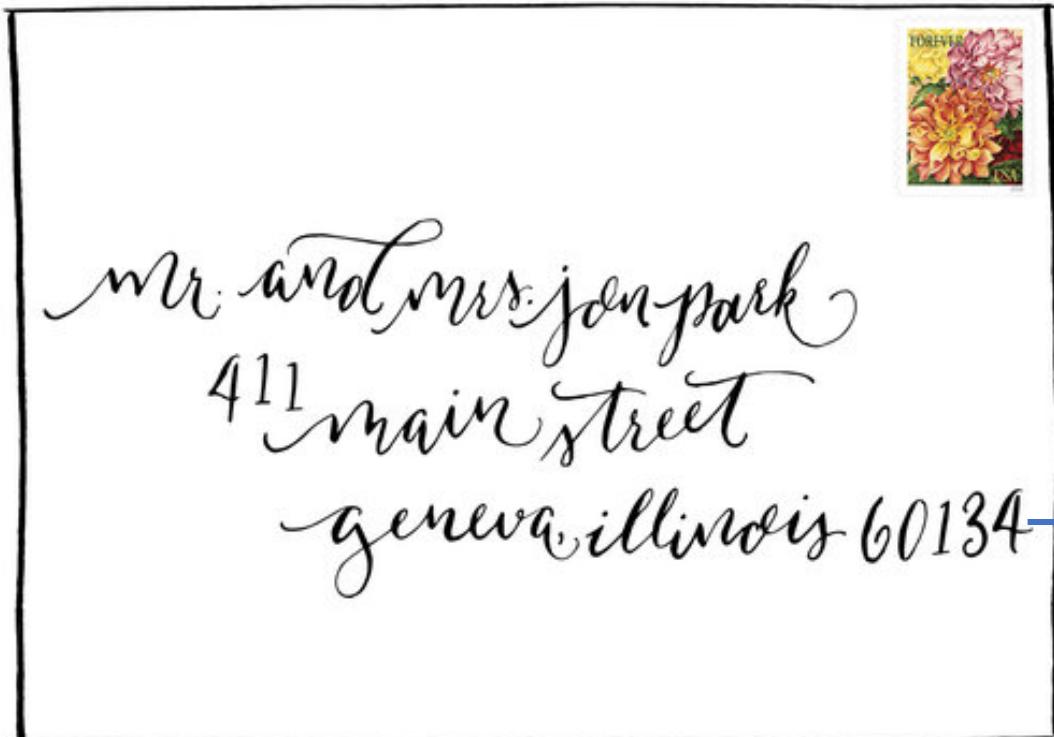
- 
- Predictive modelling
    - Will customer  $X$  churn next month / default on her loan?
    - How much would prospect  $X$  spend?
    - Will user  $X$  click on an ad / purchase a product / watch a movie?
    - Who might be good “friends” on our social networking site?
  - Regression (continuous output variables)
  - Classification (discrete output variables)

QUIZ  
Time



# Which problem is it?

Identifying the zipcode digits on handwritten envelopes.



6    0    1    3    4

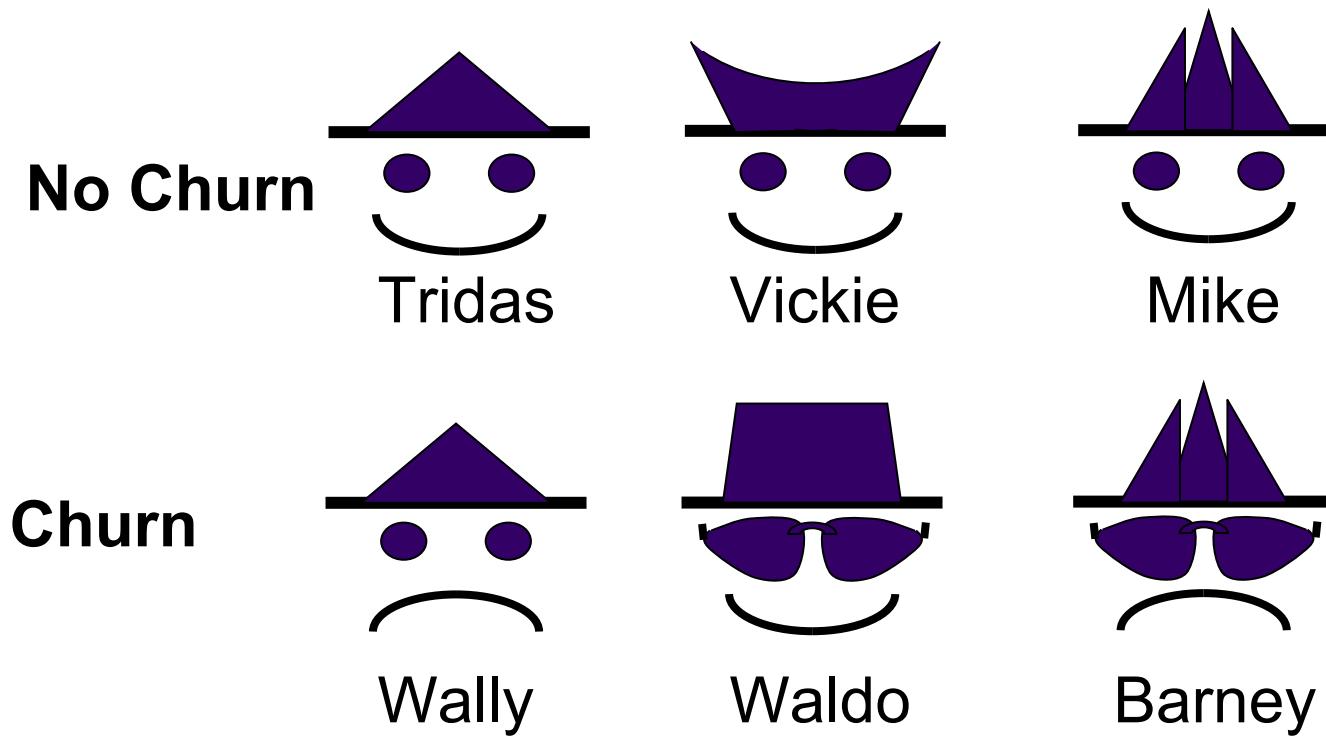
# Which problem is it?

Predicting someone's salary based on years of experience

# Which problem is it?

We wish to organize our company's pdf documents in different categories based on document title.

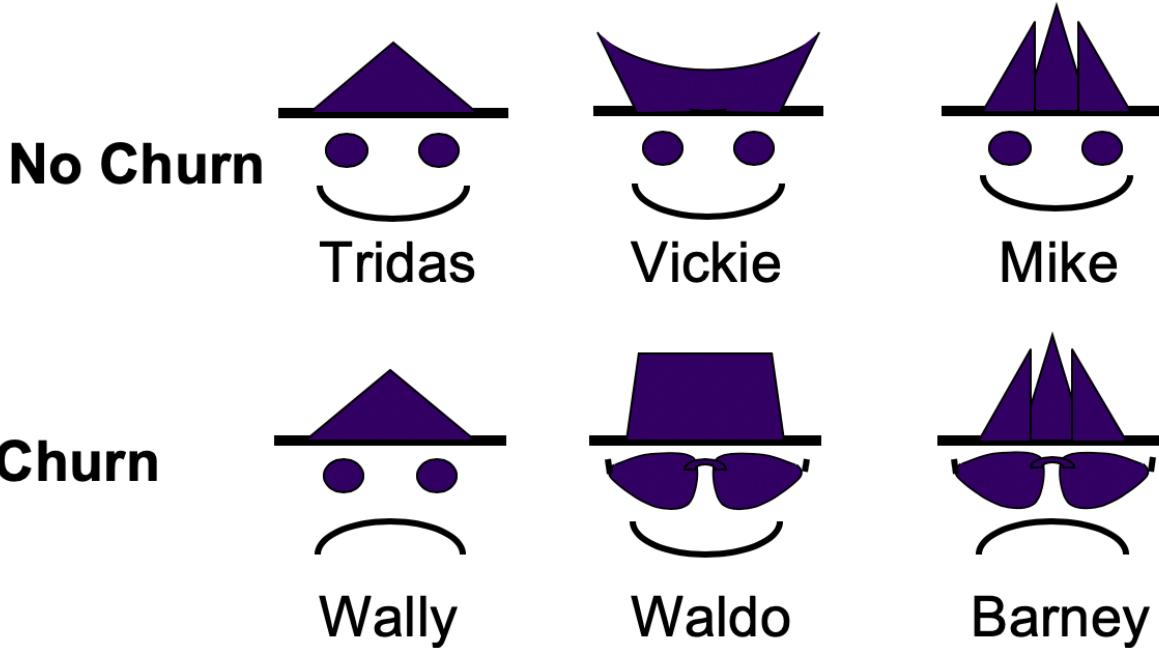
# What does it mean to build a model?



# What features to use here?

# “Manual” Data Mining

What can we learn from the training data?  
Are there rules or patterns or features?



*IF “no sun glasses” AND “smile” THEN “no churn”*

# Using the model/rule for prediction

*IF “no sun glasses” AND “smile” THEN “honest”*

*New example:  
Honest or crooked?*

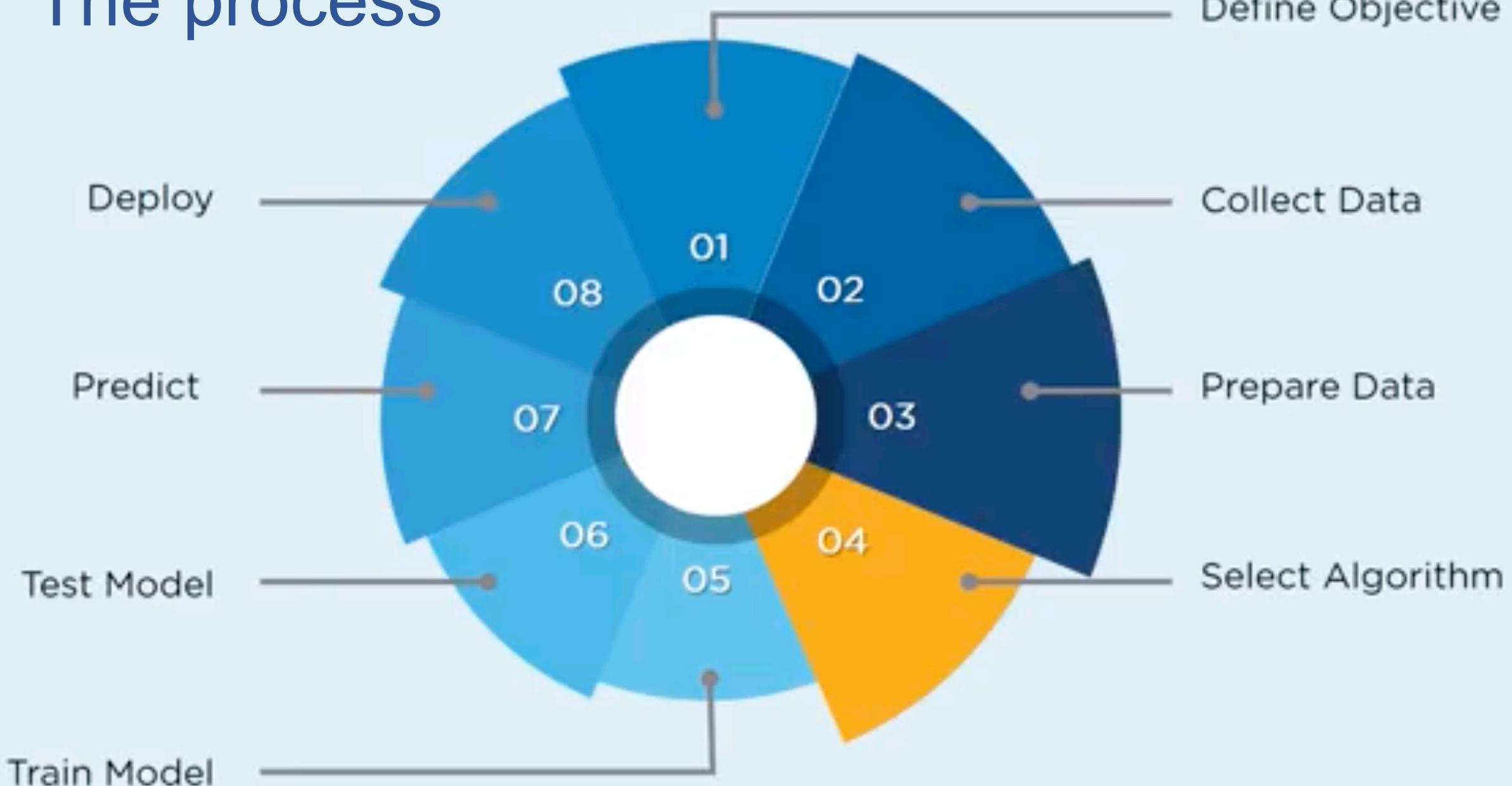


Honest!

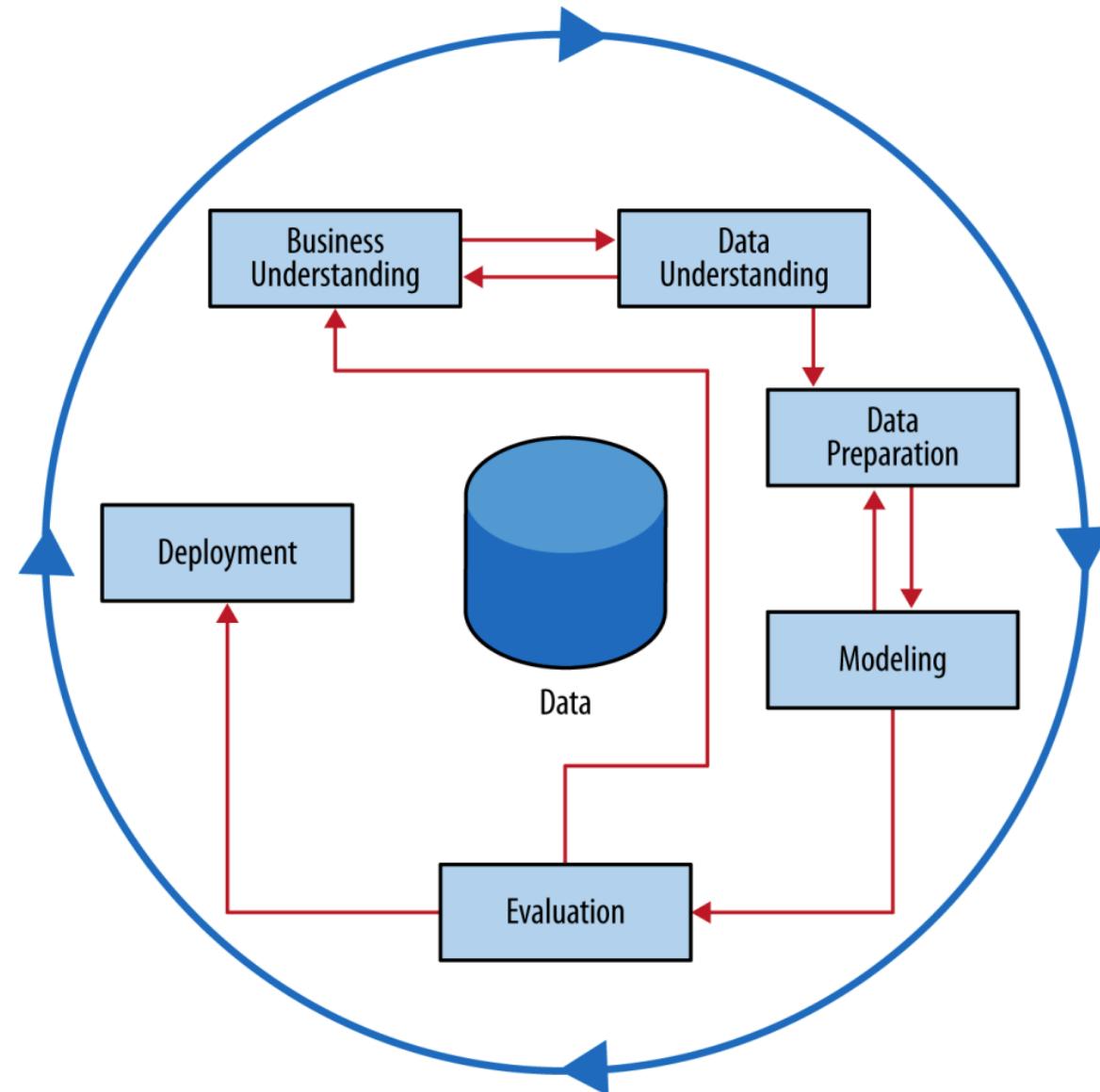
# Data Mining Cycle

Overview

# The process



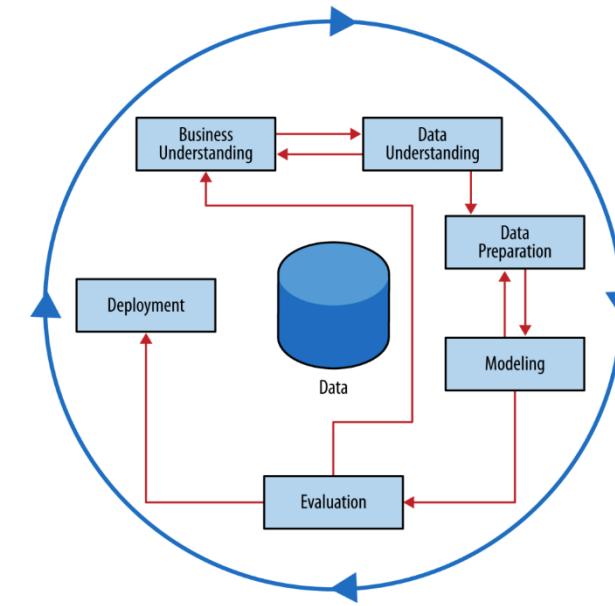
# Cross-industry standard process for data mining (CRISP-DM)



# CRISP-DM Steps

## Business Understanding

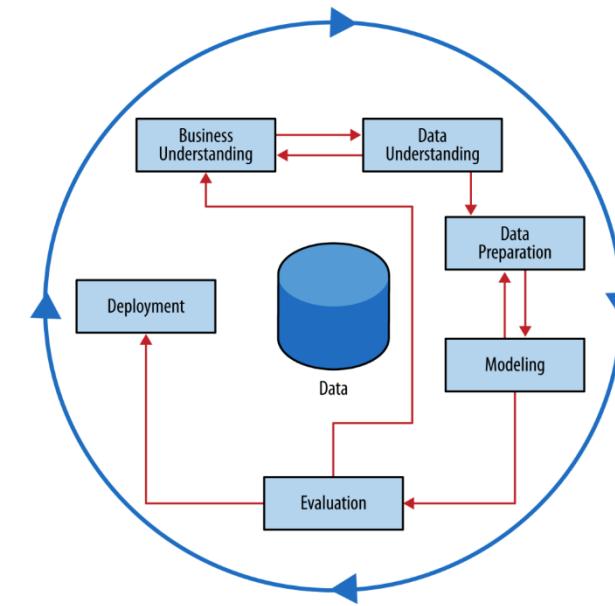
- Define business problem/question
- Restructure business problem as a data mining problem
  - Divide and conquer
    - Recursively split problem into smaller parts and solve them
    - Classification/regression/clustering/...
- Define success criteria
  - Eg. 90% correct classification rate



# CRISP-DM Steps

## Data Understanding

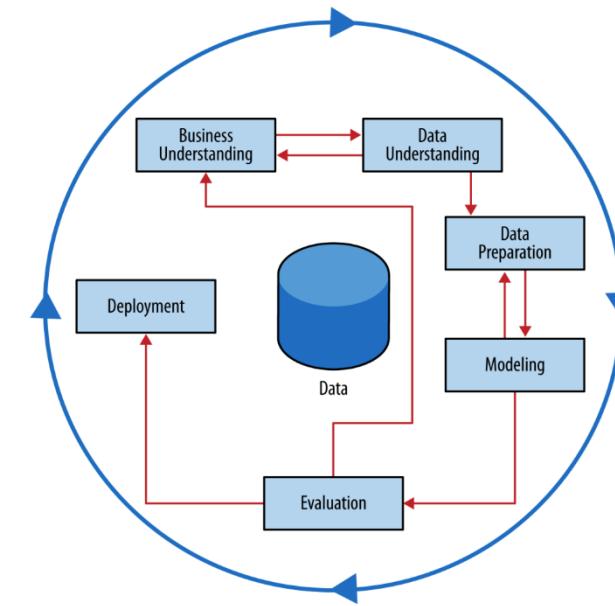
- Collect initial dataset
- Acquire additional data
- Understand data structures
- **Exploratory data analysis** to get first insights
- Assess data quality



# CRISP-DM Steps

## Data Preparation

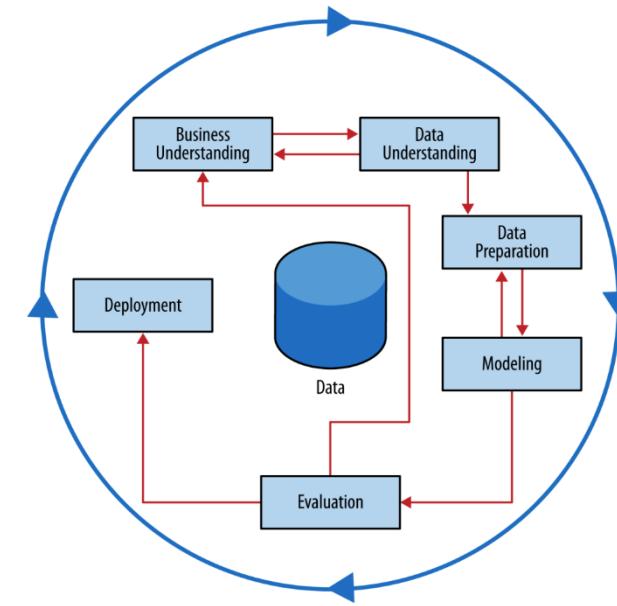
- Clean data
- Construct data
  - new variables
  - new records
  - choose attributes
- Integrate data
- Reformat data



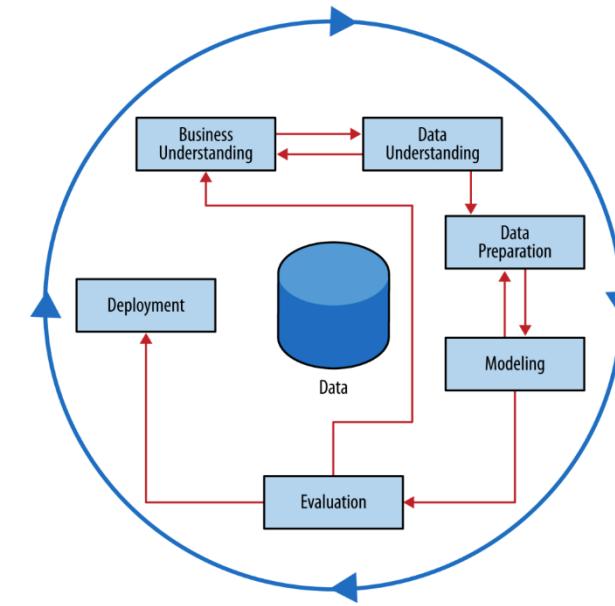
# CRISP-DM Steps

## Modeling

- Select algorithms
  - Generate training and test sets
  - Fit model to data
  - Assess model accuracy
- 
- Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.



# CRISP-DM Steps



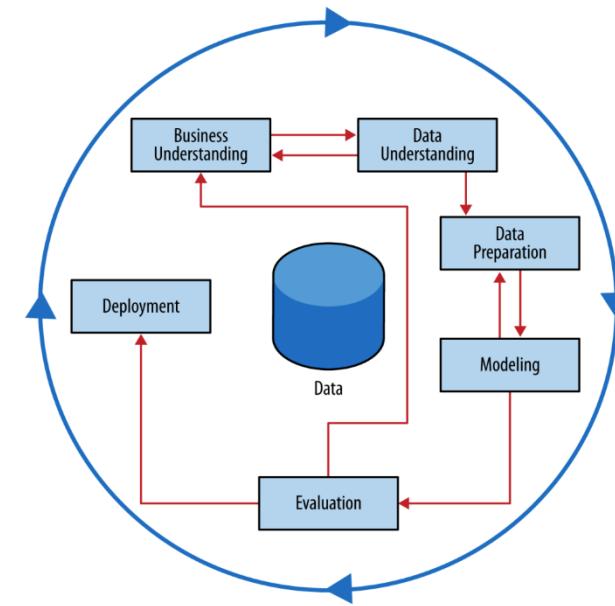
## Evaluation

- Evaluate data mining results with respect to business goals and success criteria
- Assess legal and ethical considerations
- “In vivo” testing
- “Sign off” from relevant stakeholders

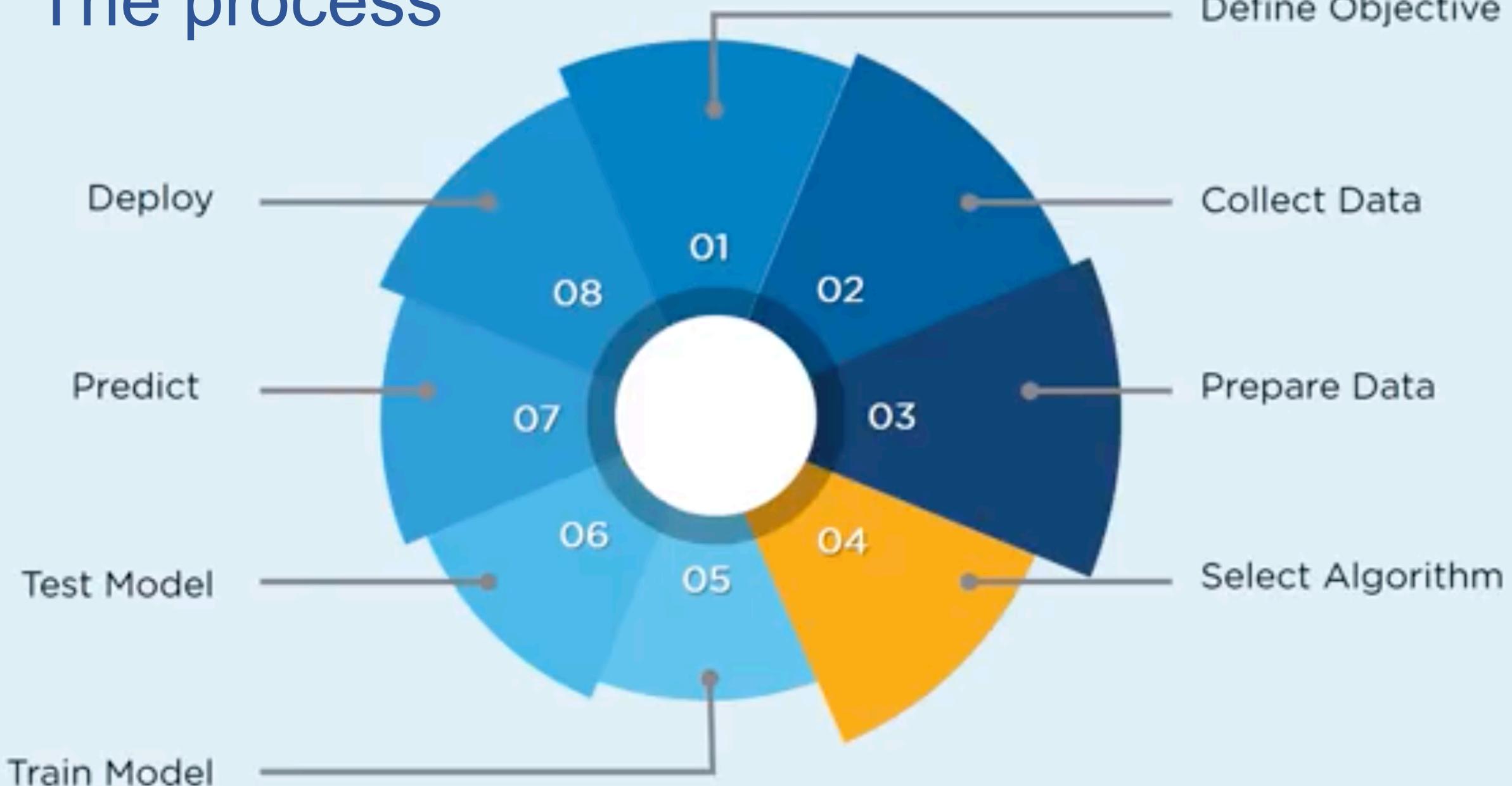
# CRISP-DM Steps

## Deployment

- Develop production system
- Train employees
- Continuous monitoring and maintenance
- Project review



# The process



# Exploratory Data Analysis

# Before we do anything else we start with EDA

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	OnlineSecurity	...	TotalCharges	Churn
0	7590-VHVEG	Female	0	Yes	No	1	No	No	...	29.85	No
1	5575-GNVDE	Male	0	No	No	34	Yes	Yes	...	1889.5	No
2	3668-QPYBK	Male	0	No	No	2	Yes	Yes	...	108.15	Yes
3	7795-CFOCW	Male	0	No	No	45	No	Yes	...	1840.75	No
4	9237-HQITU	Female	0	No	No	2	Yes	No	...	151.65	Yes

# Useful to do some Exploratory Data Analysis (EDA)

- Do you understand what this data is?
- How many rows and columns it has?

## About this file

### **Telcom Customer Churn**

Each row represents a customer, each column contains customer's attributes described on the column Metadata.

The raw data contains 7043 rows (customers) and 21 columns (features).

The "Churn" column is our target.

# EDA – what do your columns mean?

- Do you understand what the columns mean?

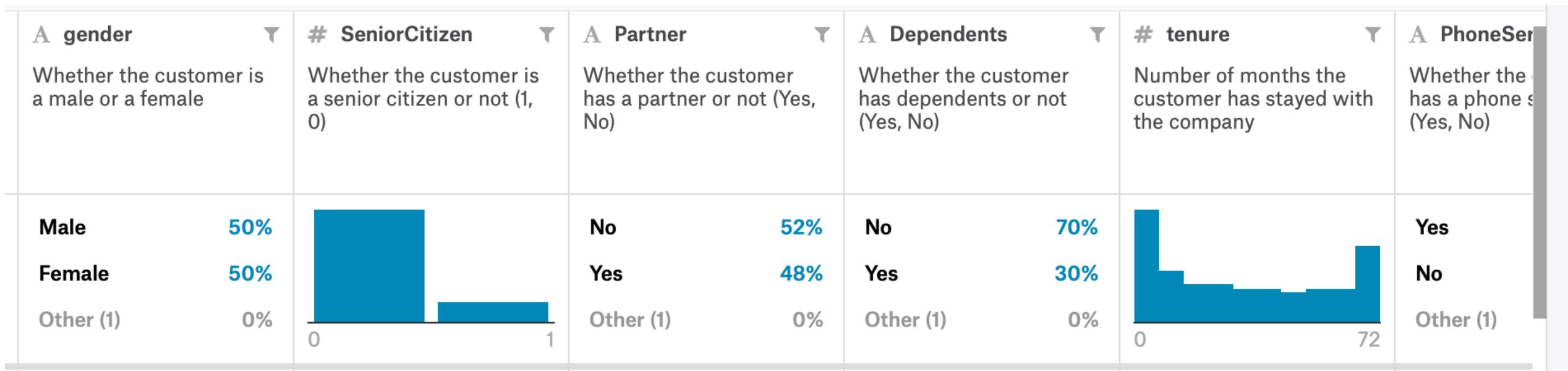
## Columns

- A **customerID** Customer ID
- A **gender** Whether the customer is a male or a female
- # **SeniorCitizen** Whether the customer is a senior citizen or not (1, 0)
- A **Partner** Whether the customer has a partner or not (Yes, No)
- A **Dependents** Whether the customer has dependents or not (Yes, No)

# EDA – do the features have the correct datatype?

- For example, there was a column with dates. Was this properly imported and interpreted as date, or is it just a string?

# EDA – do you understand the feature distributions?



<https://www.kaggle.com/blastchar/telco-customer-churn>

# EDA

- All the previous steps should help you understand whether the data need some “cleaning”, but for now we will assume that the data have all the values needed and they are correct
- In your class project, expect a big part of your effort to be allocated to the data cleaning and sanitization.

# Week 1 Goals

Terminology, Identify the problem

# Week 1 Terminology

- Supervised learning vs unsupervised learning
- Regression, Classification, Clustering, Anomaly detection
- What is feature, input variable, output variable, target variable, data dimensionality

# Lab1: Python and Notebooks

# Week 1 Lab work

- Install Anaconda, install Spyder (15mins) Hour 1
- Introduction to Python Notebooks (15mins)
- Exercise Notebook (15mins)
- Basic Intro to Python with Pandas (20mins) Hour 2
  - Load a dataset
  - Some basic data analysis
  - Python tutorial: <https://docs.python.org/3/> (on your own time)
- Describe Assignment 1 (15mins + questions)
  - Requirements
  - How to do a [screen recording](#).

# Question

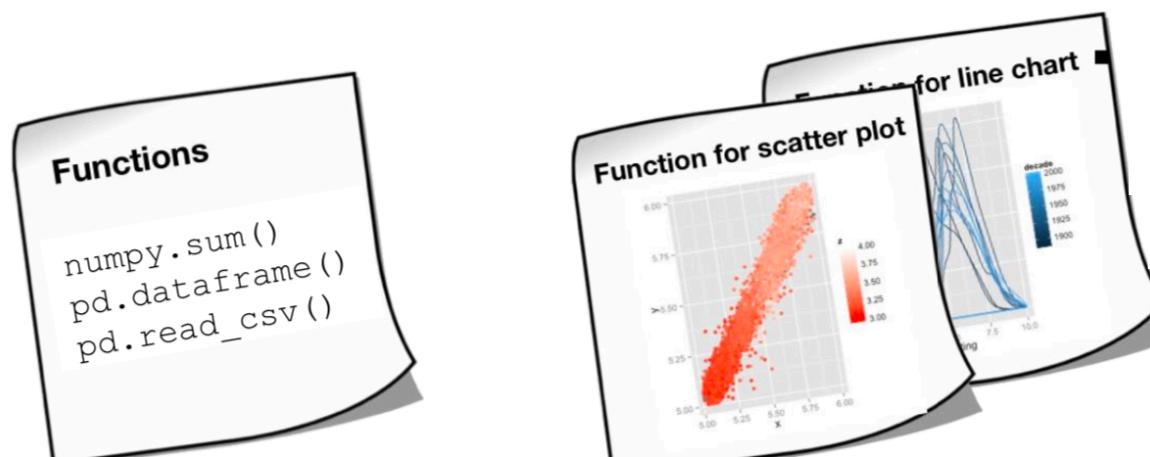
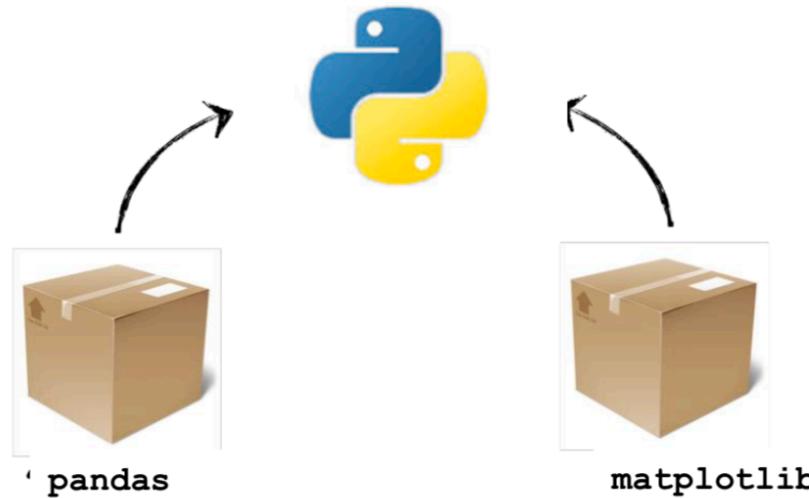
Why should you take a data scientist with you in the jungle?

# We will be using Python because

- It's **open source** and **free**,
- **simple syntax**; by looking at the code you can understand what it does
- ability to allow **rapid testing** of complex algorithms,
- access to comprehensive **libraries**
- Great **community** support in Stackoverflow

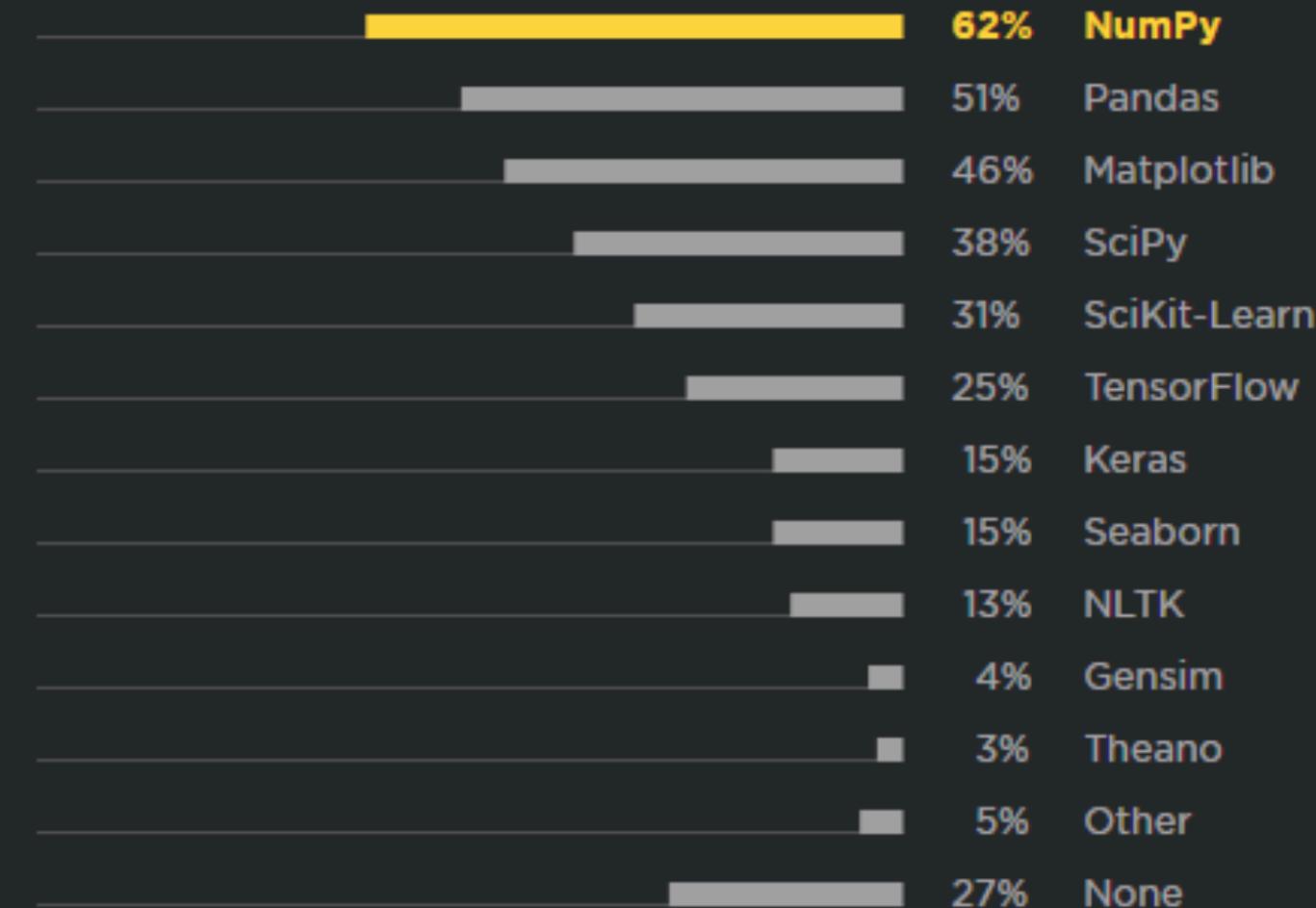
# Why is Python so popular?

- Many packages to extend its functionality.



Some popular frameworks in Python. We will cover some of those in the class.

## Data Science Frameworks and Libraries (multiple answers)



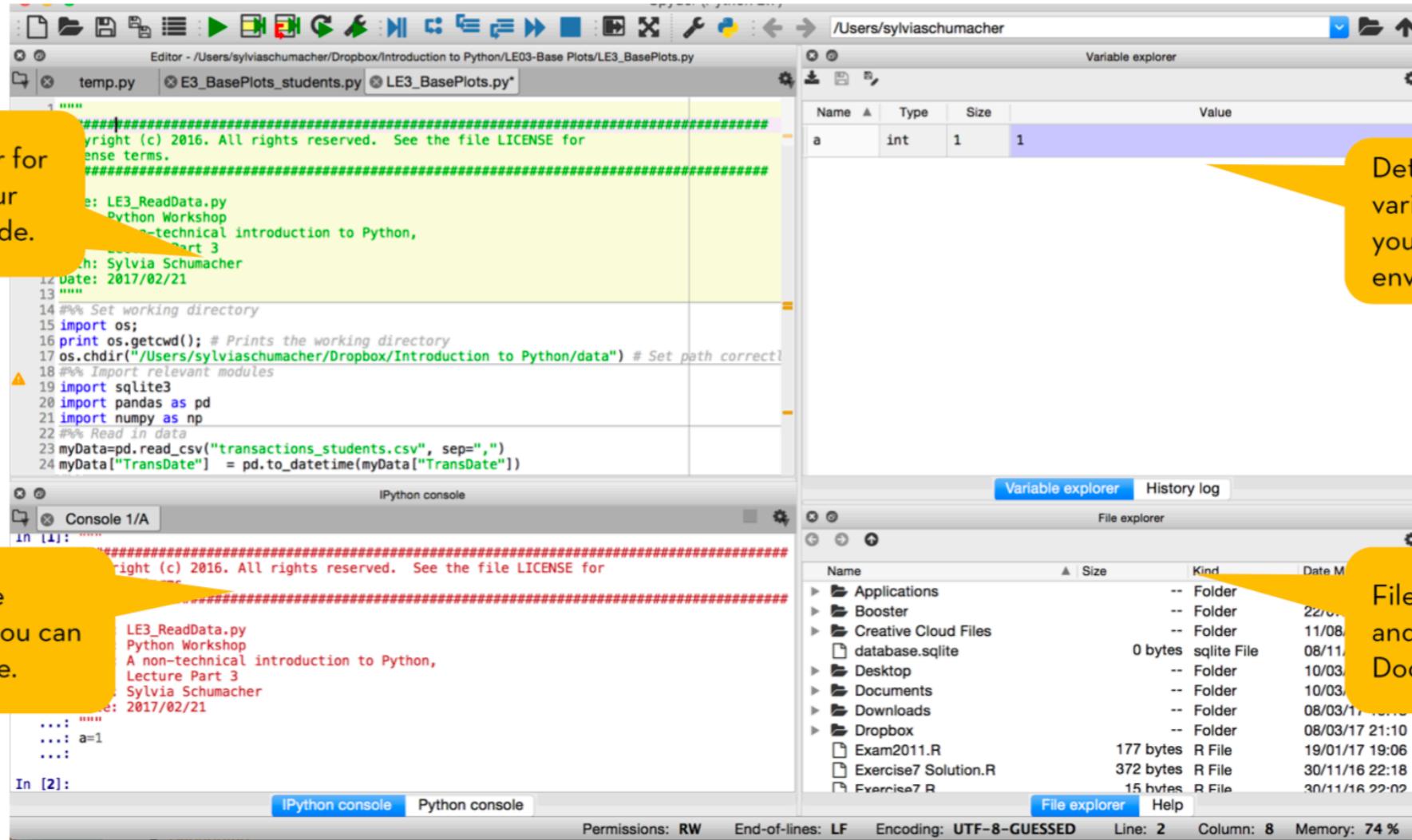
Source: <https://www.jetbrains.com/research/python-developers-survey-2018/>

# What is Spyder?

- Spyder is an open-source IDE for Python
- Available for all Operating Systems
- Provides
  - Console
  - Editor
  - Workspace manager



# Spyder Interface



# Alternative IDE's

- PyCharm
- On the cloud
  - [Google Colab](#) (also get GPU time!)
  - [Azure Notebooks](#)
  - [Jupyter Lab](#)

Microsoft Azure Notebooks Preview

Libraries FAQ/Support What's New

Here is a list of numbers:

In [2]: `let sampleNumbers = [ 0 .. 15 ]`

sampleNumbers

Out[2]: `[0; 1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12; 13; 14; 15]`

Next you use `let` to define a function that accepts and returns an integer. Parentheses accept a parameter name using `(argument: type)`. Documentation comments are added using `///`.

In [3]: `/// A function to compute a sample curve`

The screenshot shows the Google Colab interface. At the top, there's a toolbar with 'File', 'Edit', 'View', 'Insert', 'Runtime', 'Tools', and 'Help'. Below the toolbar are buttons for 'CODE', 'TEXT', 'CELL', and 'EDITING'. On the right side, there are status indicators for 'RAM' and 'Disk'. The main area displays a Jupyter notebook cell with the following code:

```
[1] import matplotlib.pyplot as plt
import pandas as pd
pd.__version__
'D: '0.22.0'

[2] # ri stands for Rhode Island
ri = pd.read_csv('police.csv').
```

Below the code cell, there's a section titled 'PyCon 2018: Using pandas for Better (and Worse) Data Science' with a link to GitHub (<https://github.com/justmarkham/pycon-2018-tutorial>). Another section titled 'Dataset: Stanford Open Policing Project ([video](#))' is also visible.

_name	driver_gender	driver_age_raw	driver_age	driver_race	violation
NaN	M	1985.0	20.0	White	Speeding
NaN	M	1965.0	40.0	White	Speeding
NaN	M	1972.0	33.0	White	Speeding
NaN	M	1986.0	19.0	White	Call for Service
NaN	F	1984.0	21.0	White	Speeding

# What is Anaconda?

- Anaconda is the leading data science platform powered by Python
- Includes over 100 of the most popular Python, R, Scala packages for data science.
- Part of Anaconda, is conda, a package, dependency and environment manager, which allows the easy installation of >700 important packages.

The Enterprise Data Science Platform for...



Data Scientists



IT Professionals



Business Leaders

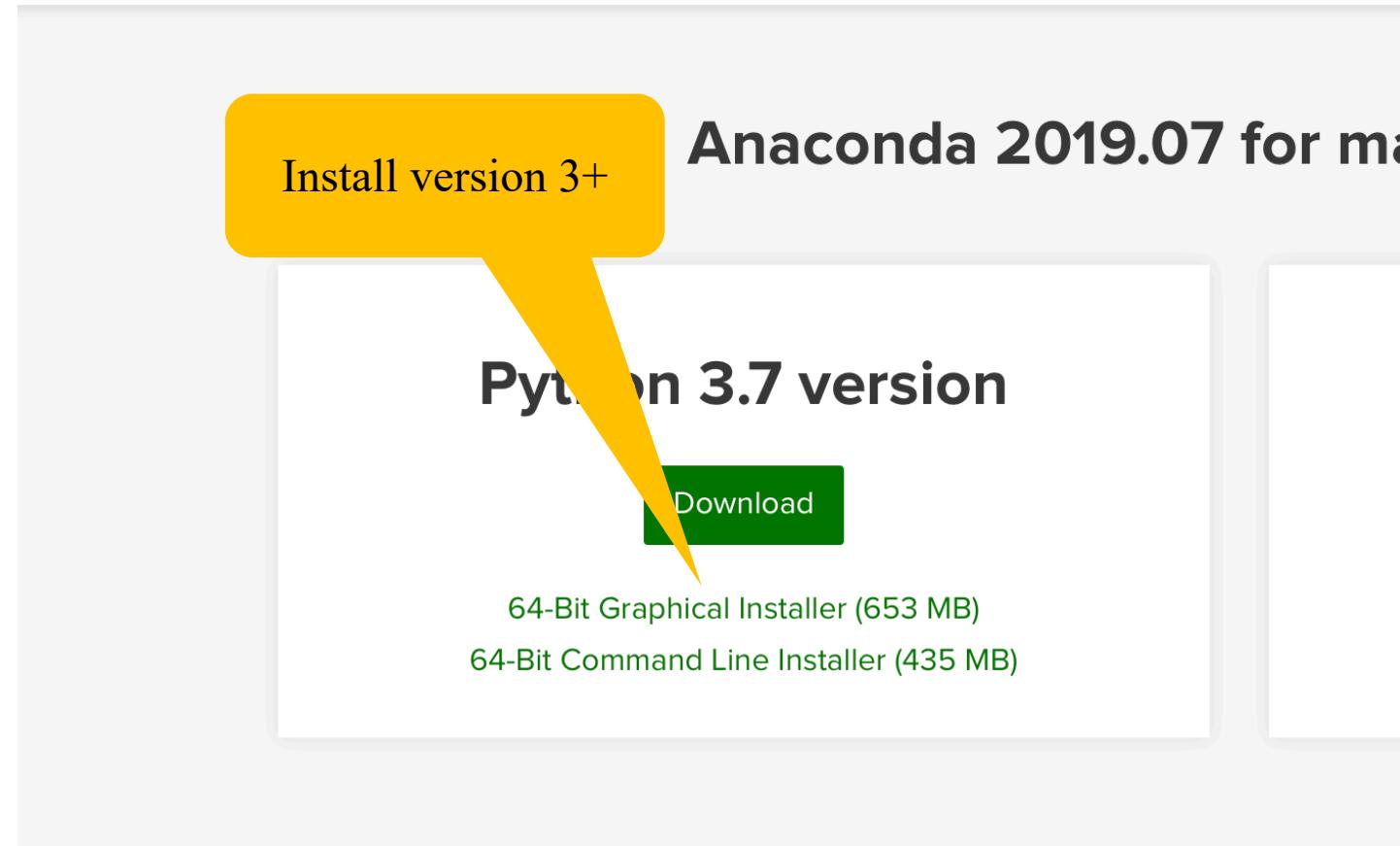
# How to install Anaconda



Go to:

<https://www.anaconda.com/distribution/>

And download the graphical install for Python version 3+



# How to install Spyder

Go to the “Anaconda Navigator”



Click launch or “install” if it hasn’t been already installed

The screenshot shows the Anaconda Navigator application window. On the left is a sidebar with icons for Home, Environments, Learning, and Community. The main area displays a grid of application cards. The Spyder card is highlighted with a yellow callout bubble pointing to its "Launch" button. Other visible applications include JupyterLab, Jupyter Notebook, Qt Console, Glueviz, and VS Code. Each card includes a brief description and either a "Launch" or "Install" button.

Application	Description	Action
JupyterLab	An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.	Launch
Jupyter Notebook	Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.	Launch
Qt Console	PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.	Launch
Spyder	Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features	Launch
Glueviz	Multidimensional data visualization files. Explore relationships within related datasets	Install
Omega3D	None	None
RStudio	None	None
VS Code	None	None

# What is Jupyter Notebooks?

- Build small chunks of code
- Builds narratives
- Share them for online use

The screenshot shows a Jupyter Notebook interface with the following features highlighted by yellow callout bubbles:

- Exporting your document will produce immediate results of your calculations.** (points to the top right corner of the notebook window)
- Different export formats, e.g. HTML or PDF.** (points to the top left of the notebook window)
- Easy text formatting with Markdown.** (points to the text area labeled "Exercise Part I")
- Integration of Python code and other languages, e.g. SQL and R** (points to the code cell "In [4]:" containing Python code and the resulting plot)
- Output of code chunks is directly displayed** (points to the bottom right of the notebook window)
- Put your work in a nice format.** (points to the bottom center of the notebook window)

**Exercise Part I**

**Orders in 2015**

Only certain business segments were considered:

- Food
- Beverages
- Hygiene

Other business segments were **not** considered:

- Clothes
- Accessories

**Include Plots**

In [4]:

```
import matplotlib.pyplot as plt
%matplotlib inline
x=[1,3,4,5,7,10,12]
plt.plot(x,x)
plt.title("Include plots inline in your notebook")
plt.xlabel("x")
plt.show()
```

include plots inline in your notebook

Put your work in a nice format.

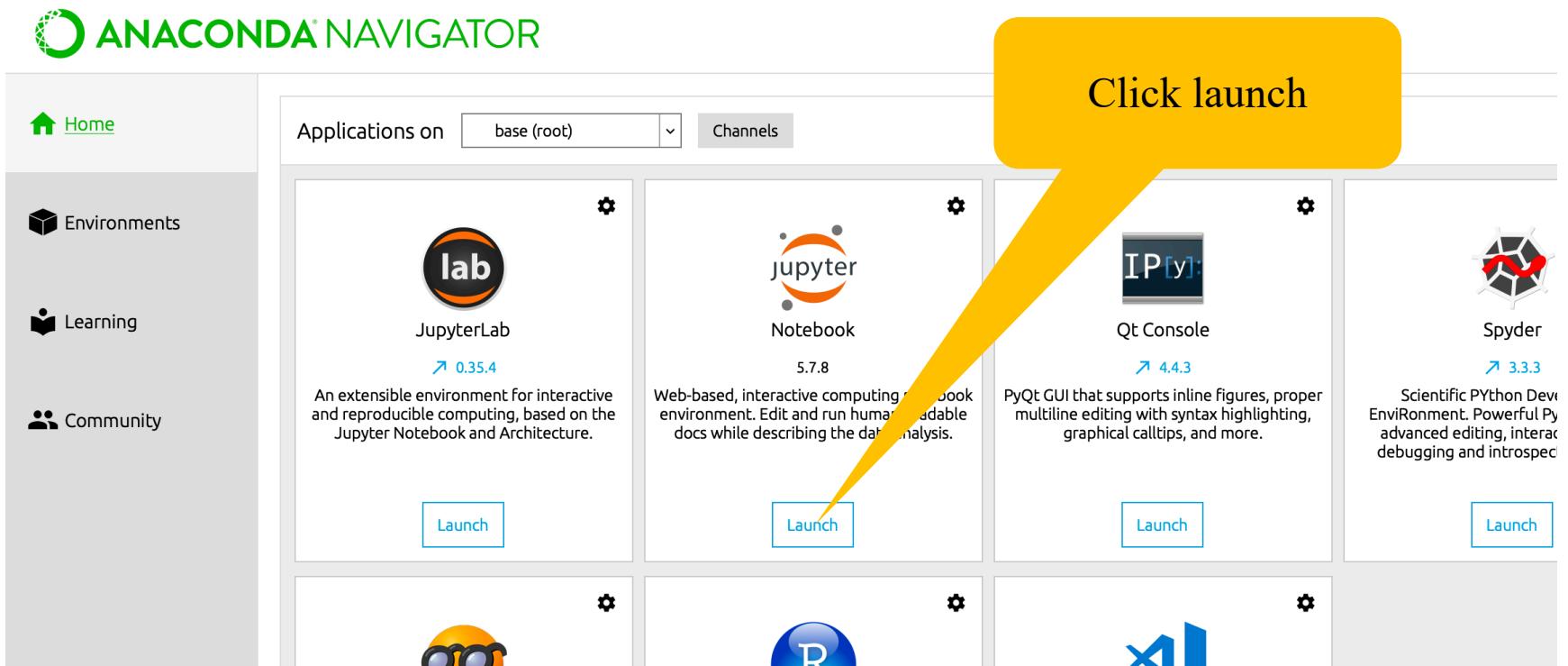
# Notebooks vs regular programming scripts

Activity	Scripts	Notebooks
Building a narrative	Comments are available to describe code.	Rich text is available to describe the process and conclusion
Manage output	Output is available in the console or environment.	Output is embedded in a single document. Code and output may be divided into separate <b>code chunks</b> .
Creating a final report	Creating a report is a separate and time-consuming step.	Instant report possible. Reports are publishable as HTML, PDF, ...)

# How to create your first Jupyter notebook

- Already installed with Anaconda
- Just go to command line and type: **jupyter notebook**

Just go to the “Anaconda Navigator”



# Main components of a Notebook

The image shows two screenshots of the Jupyter Notebook interface. On the left, the 'Dashboard' is displayed, showing a list of open notebooks. A yellow callout points to the 'Running' tab with the text 'View open Notebooks'. Another yellow callout points to the 'New' button with the text 'Create a new Notebook.'. A third yellow callout points to the bottom of the dashboard with the text 'Manage files in the “Dashboard”'. On the right, a 'Jupyter Notebook' window is shown with a title bar 'localhost Jupyter Notebook'. The main area contains sections like 'Exercise Part I' and 'Include Plots'. In the 'Include Plots' section, a code cell in Python is shown generating a plot of a linear function. A yellow callout points to this plot with the text 'Actual notebook'.

Jupyter Notebook runs in your browser.

View open Notebooks

Create a new Notebook.

Manage files in the “Dashboard”

Actual notebook

```
In [4]: import matplotlib.pyplot as plt
%matplotlib inline
x=[1,3,4,5,7,10,12]
plt.plot(x,x)
plt.title("Include plots inline in your notebook")
plt.xlabel("x")
plt.show()
```

Exercise Part I

Orders in 2015

Only certain business segments were considered:

- Food
- Beverages
- Hygiene

Other business segments were **not** considered:

- Clothes
- Accessories

Include Plots

In [4]: import matplotlib.pyplot as plt  
%matplotlib inline  
x=[1,3,4,5,7,10,12]  
plt.plot(x,x)  
plt.title("Include plots inline in your notebook")  
plt.xlabel("x")  
plt.show()

Include plots inline in your notebook

# Lab Exercise 1: Notebooks

# Notebook - Exercise

1. Re-create the Jupyter Notebook on the right.

The screenshot shows a Jupyter Notebook interface with the following elements:

- Toolbar:** Includes icons for file operations (New, Open, Save, etc.), cell selection, and a "Markdown" button.
- Title:** "LE 12 - Task 2: Some simple Statistics tasks".
- Section Headers:** "Head in size 1." (above the title) and "Bold font." (near the title).
- Table of Contents:** A list of topics:
  1. Arithmetic mean
  2. Mean Squared Error
    - A. Bias of an estimator
    - B. Variance of an estimator
- Text:** "Use a list."
- Section Header:** "1. The Arithmetic Mean".
- Equation:** 
$$Mean(x) = \sum_{i=1}^n x_i$$
- Text:** "Use a Latex formula."
- Text:** "Embedded a formatted code block:"
- Code:**

```
x=[1,3,9,4]
x_bar=1/4*sum(x)
```
- In [28]:**

```
#Actual Python Calculation:
x=[1,3,9,4]
x_bar=1/len(x)*sum(x)
x_bar
```
- Out[28]:** 4.25
- Text:** "Python code cell."
- Section Header:** "2. Mean Squared Error".
- Equation:** 
$$MSE(x) = bias^2(x) + Var^2(x)$$
- Text:** "Use a Latex formula."
- Text:** "For further information see: <http://www.statisticshowto.com/mean-squared-error/>"
- Text:** "Link."

# Join Slack Workspace

# Slack - Collaboration

Join the workspace <https://bit.ly/33tNiPR> (use your UNIL email!)

The screenshot shows a Slack workspace interface. On the left, there's a sidebar with a dark purple background containing the workspace name "UNIL2019 - Da...", a user icon for "Michalis Vlachos", and a bell icon. Below this are sections for "Threads", "Channels" (with a plus sign), and a list of channels: "# general", "# project", "# week1" (which is highlighted in blue), "# week2", "# week3", "# week4", "# week5", and "# week6". There are also buttons for "+ Add a channel" and "Direct Messages" (with a plus sign). Under "Direct Messages", it lists "Michalis Vlachos (you)" and "+ Invite people". At the bottom, there's an "Apps" section with a plus sign.

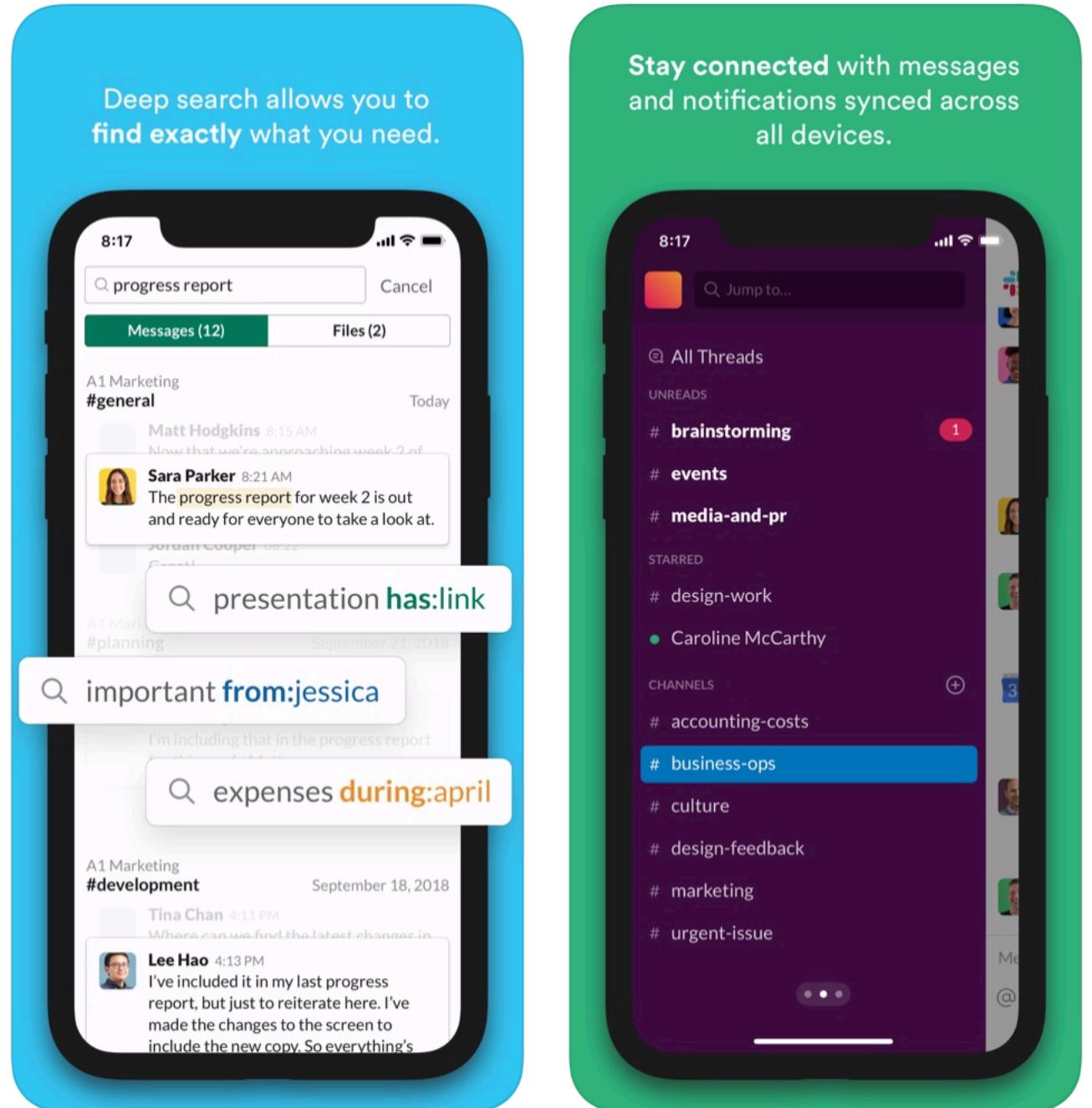
The main area shows the "#week1" channel. The header includes a phone icon, a help icon, a gear icon, and a bracket icon. It has statistics: 1 star, 1 message, 1 file, and a "Add a topic" button. Below the header, the channel purpose is displayed: "You created this channel on August 11th. This is the very beginning of the #week1 channel. Purpose: Here we post our comments for week 1 (edit)". There are buttons for "+ Add an app" and "Add people to this channel".

A timestamp "Sunday, August 11th" is shown above two messages from "Michalis Vlachos". The first message says "joined #week1." and the second says "set the channel purpose: Here we post our comments for week 1".

A horizontal line separates the past messages from the present. The word "Today" is centered above a pinned message by "Michalis Vlachos" at 9:59 AM. The message is titled "#TODO for week 1" and was last edited 4 hours ago. A small "Pinned by you" icon is next to the message.

# Slack clients

- It may be useful to download the app also for your phone.
- Check also a [short video](#) how to use slack.



# Assignment 1

**Goal:** Familiarize yourself with how to create a compelling Python notebook.

# Week 1 Assignment [personal]

- Find an interesting dataset from the Internet. It's better if the dataset is relatively new, e.g., after year 2000, and it should not contain only a few datapoints. You can find many in [Kaggle](#). If the data is about Switzerland even better! The dataset may be in CSV, JSON, XLS or other format. Prepare a **Python notebook** that should answer in some narrative form the following:
- Task 1:
  - Describe the dataset. Why did you choose it? Where did you find it? What it is about? What attributes it contains? Types of attributes, ranges, etc. (if you don't know how do all, do as many as you can).
  - To answer the above you have to read the readme file, and look at the data. For this assignment you can with Excel (or you can use Pandas if you are familiar with it).
- Task 2:
  - What kind of questions would one be interested to ask/answer about this dataset?

# Week 1 Assignment

- Task 3
  - Prepare a **short video** (1-2mins) with the following content:
    - Your name and student ID.
    - The dataset and where you found it.
    - What questions one can ask on it?
    - (optional) Any basic analysis you may have conducted on it (analysis in pandas, visualization).
  - The video should include voice-over and also the face of the narrator in the beginning.
  - To create the video, you may use Camtasia or Quicktime Player. **Upload** it in youtube and embed the link in the notebook (as in the Lab assignments).
- Task 4
  - In the [slack channel](#) of **week 1**, write your name and email, the name of the dataset, the link to your video.

# Week 1 Assignment

Task 5:

- Visit the [class slack channel](#), and comment about which video and/or dataset you liked the most (other than yours!). This will be credited to your class participation points.

In week 2, I will select a few of these videos to show in class.

# TODO for Week 1

1. Join class workspace at: <https://bit.ly/33tNiPR> (use your UNIL email!)

There is more in the next page!



Very IMPORTANT!  
If you want to get  
participation credit you  
should use your UNIL  
email!

# TODO for Week 1

1. Join class workspace at: <https://bit.ly/33tNiPR> (use your UNIL email!)
2. Read the article “[Data Scientist: The Sexiest Job of the 21st Century](#)” from HBR.
3. Prepare and submit assignment 1.
4.
  - a) Post the link to the video from assignment 1 on the “**week1**” channel in the workspace.
  - b) Comment on other’s people’s videos (class participation!).
5. Go over week1 slides.
6. Go over the Python notebooks (notebook basics, pandas basics) we did in class.