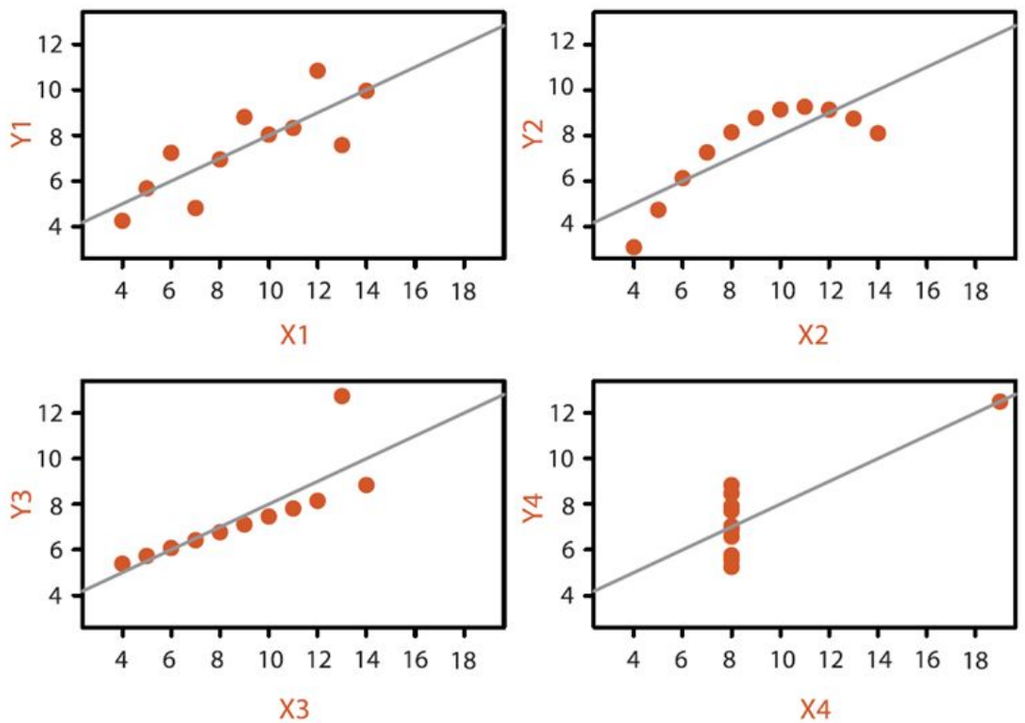
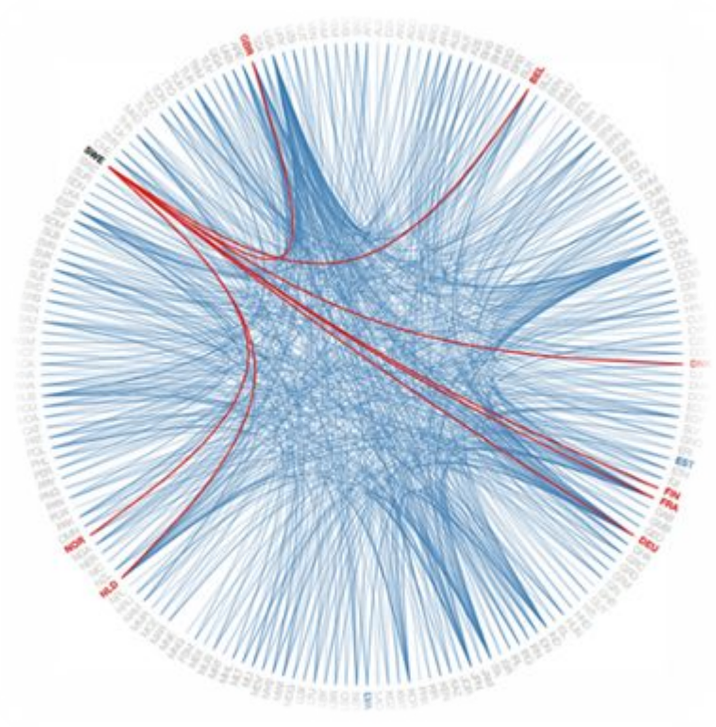




Data Science & Visualization

Introduction to Data Science

Gabriela Molina León
molina@uni-bremen.de
Institute for Information Management Bremen
Information Management Group (AGIM)

University of Bremen

Today

Date	Lecture (10:00-11:45)	Practical (12:00-13:15)
08.04.24	Introduction to Data Science	*Python Introduction*

01.07.24 Last session

19.07.24 Deadline Final Report

Note that for the slots marked red, you are expected to prepare presentations.

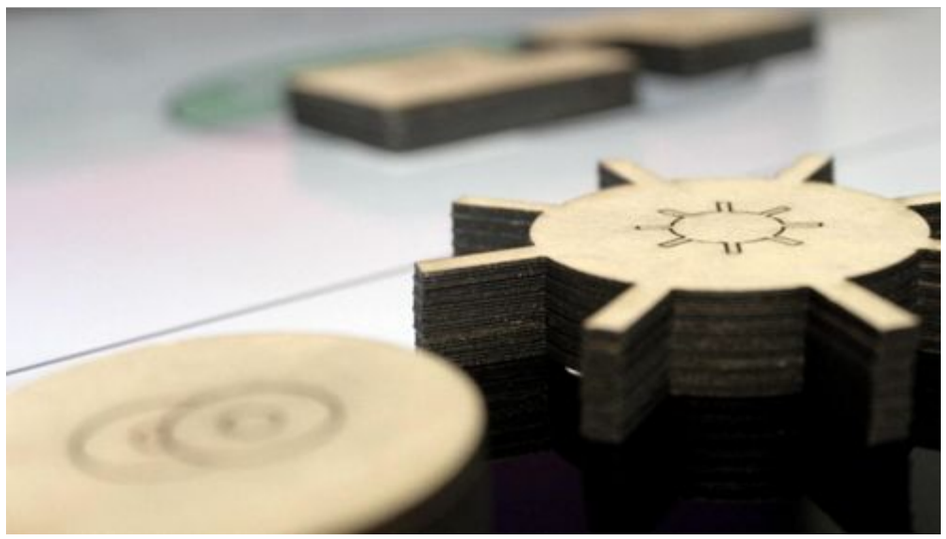
For the slots marked blue, you are expected to bring a computer.

About me

- Computer Engineer by training (Simón Bolívar University, Venezuela)
- M. Sc. in Human-Computer Interaction (Bauhaus-Universität Weimar, Germany)

- Ph.D. Student at the Information Management Research Group (AGIM)
- Contact: molina@uni-bremen.de

Simón Bolívar University



pART bench project @ Weimar

What about you?

1. Who are you? (major, previous experience)

2. Why are you taking this course?

3. What skills do you want to improve by taking this course?

Let’s decide on a schedule

Date	Lecture (10:00-11:45)	Practical (12:00-13:15)
08.04.24	Introduction to Data Science	*Python Introduction*

Options:

A. Standard (lecture 10:15-11:45, tutorial 12:15-13:45)
B. Compact (lecture 10:15-11:45, tutorial 12:00-13:30)
C. Compact & early (lecture 10:00-11:30, tutorial 11:45-13:15)

19.07.24 Deadline Final Report

Note that for the slots marked red, you are expected to prepare presentations.

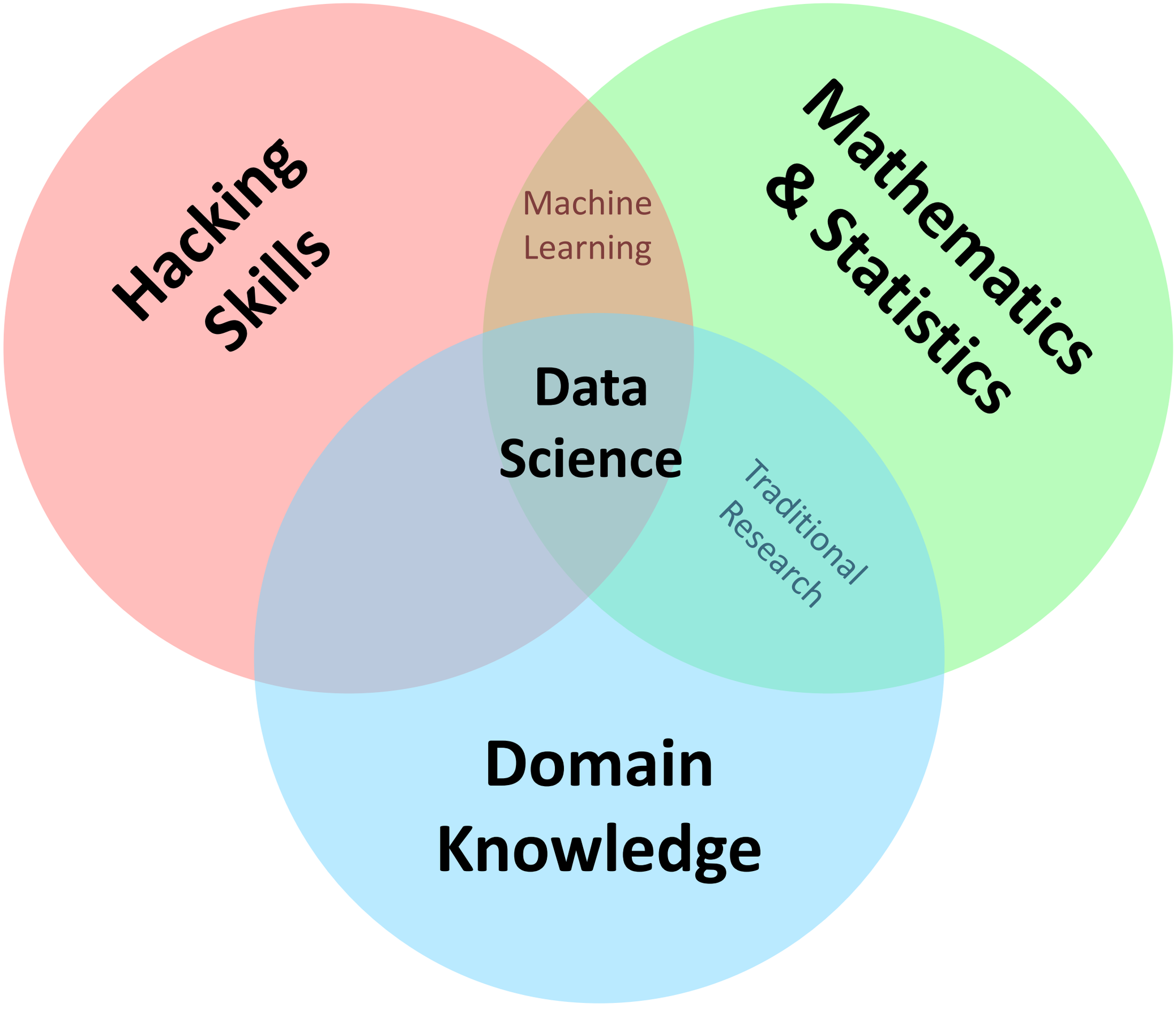
For the slots marked blue, you are expected to bring a computer.

Disclaimer

This course is based on the
“Data Science: Applied Machine Learning”
course previously taught by Prof. Dr. Hendrik Heuer.

Main difference: I will focus on **data visualization** instead of **machine learning**.

We will learn about both though.


Hacking Skills
Machine Learning
Mathematics & Statistics
Data Science
Traditional Research
Domain Knowledge

Private traits and attributes are predictable from digital records of human behaviour

MICHAL KOSINSKI
DAVID STILLWELL
THORE GRAEPEL

Private traits and attributes are predictable from digital records of human behaviour

MICHAL KOSINSKI
DAVID STILLWELL
THORE GRAEPEL

What does this predict?

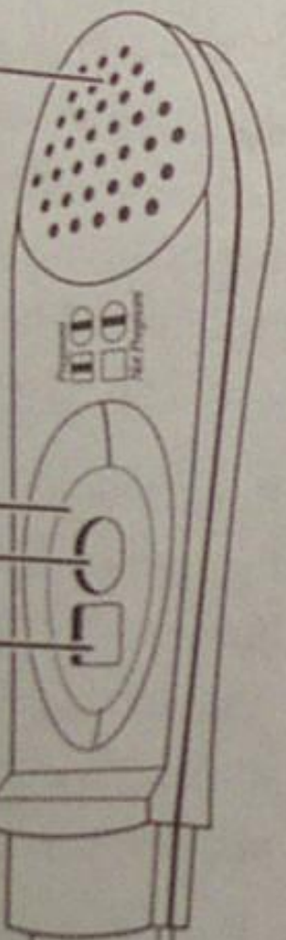
What does this predict?

What does this predict?

58000 volunteers provided Facebook likes

caucasian or african-american: 95%
male or female: 93%
hetero or homosexual: 88%
democrat or republican: 85%
christian or muslim: 82%
substance use: 65% - 73%
parents separated before 21st bday: 60%

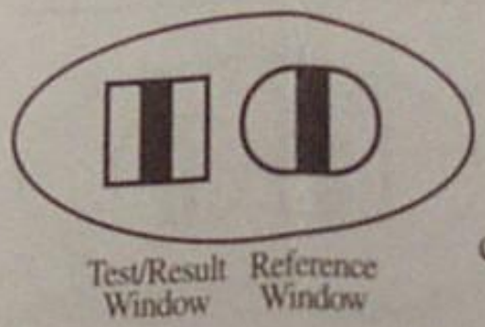
Discrimination?



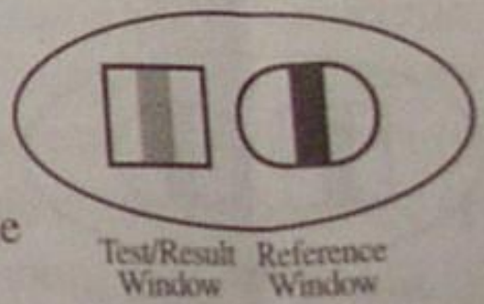
been applied. A positive result may be observed in as early as 1 minute. To verify a negative result, allow the test to run for the full 3 minutes.

DO NOT interpret your result after 10 minutes.

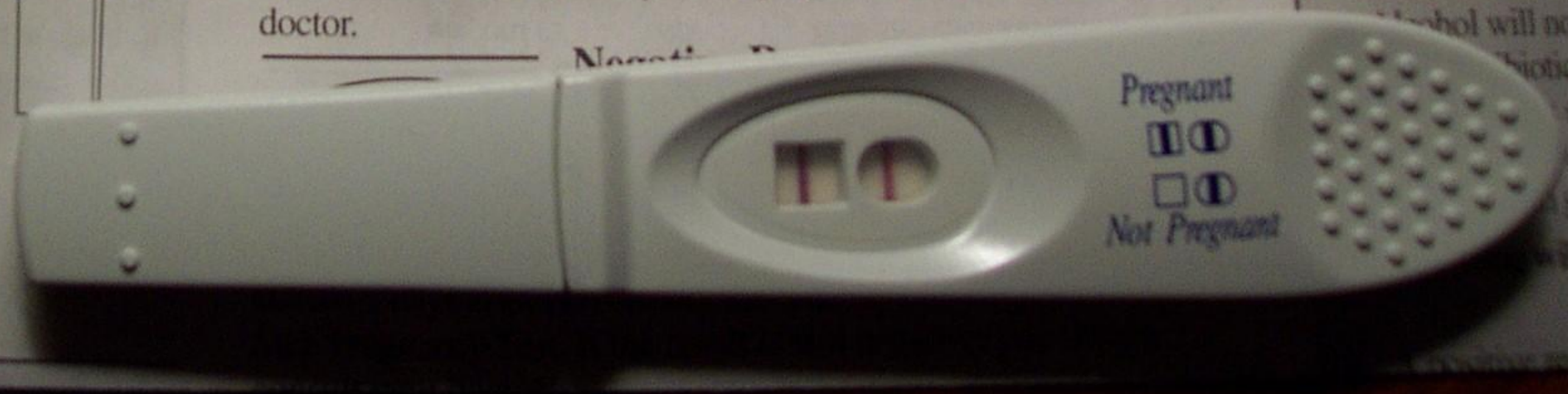
Positive Result



The Test Line can vary in shades of color compared to the Reference Line

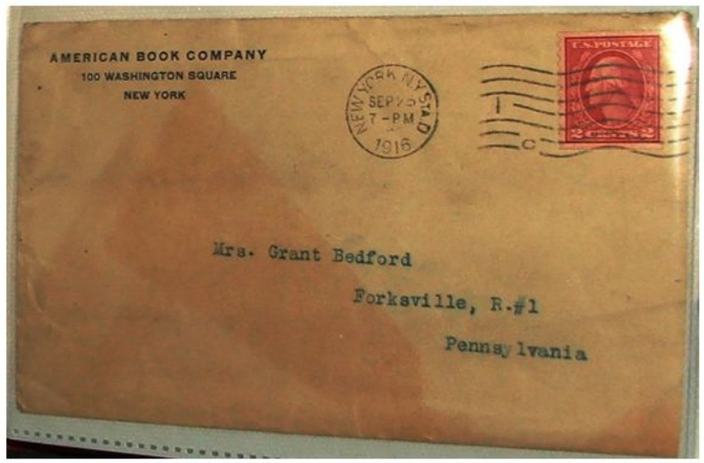


A positive result is observed when there is a purple-pink line in the oval Reference Window and the same color line in the square Test/Result Window. The line in the square Test/Result Window may be a lighter shade than the oval Reference Window line. This still means that the pregnancy hormone has been detected in your urine. You should assume you are pregnant and consult your doctor.



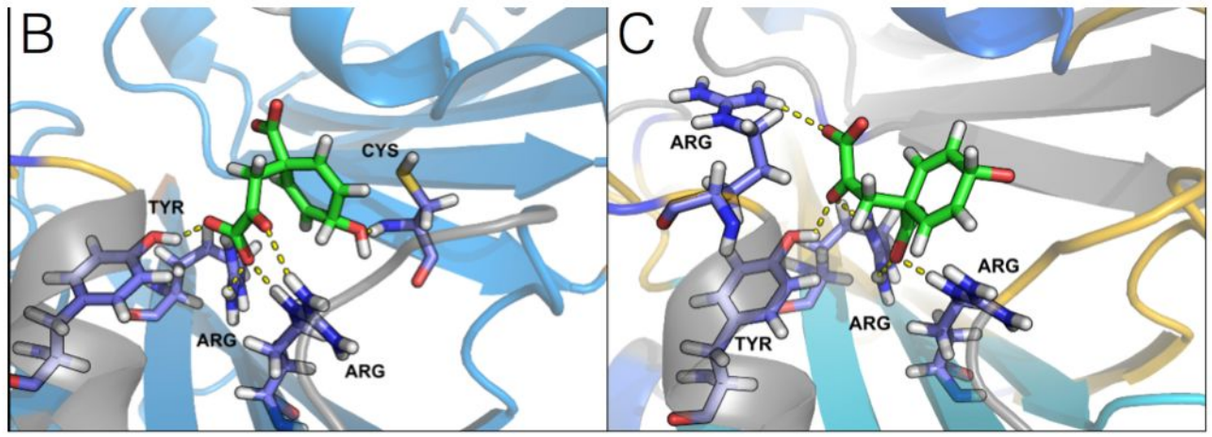
Machine Learning Examples

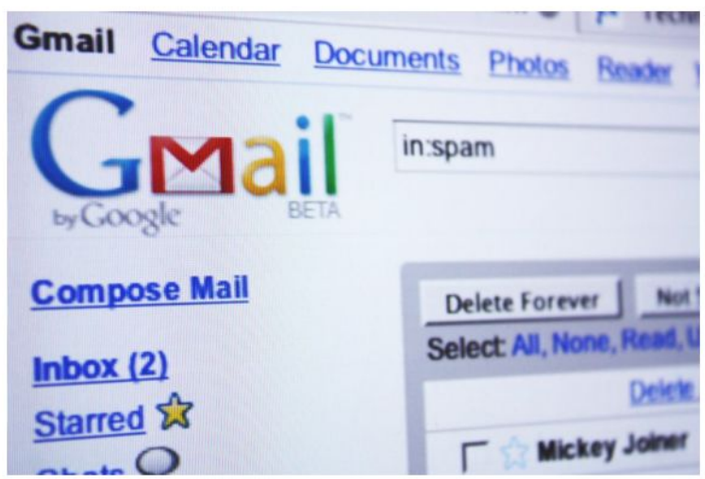
Text Recognition



http://commons.wikimedia.org/wiki/File:American_book_company_1916._letter_envelope-2.JPG#filelinks
[public domain]

Biology



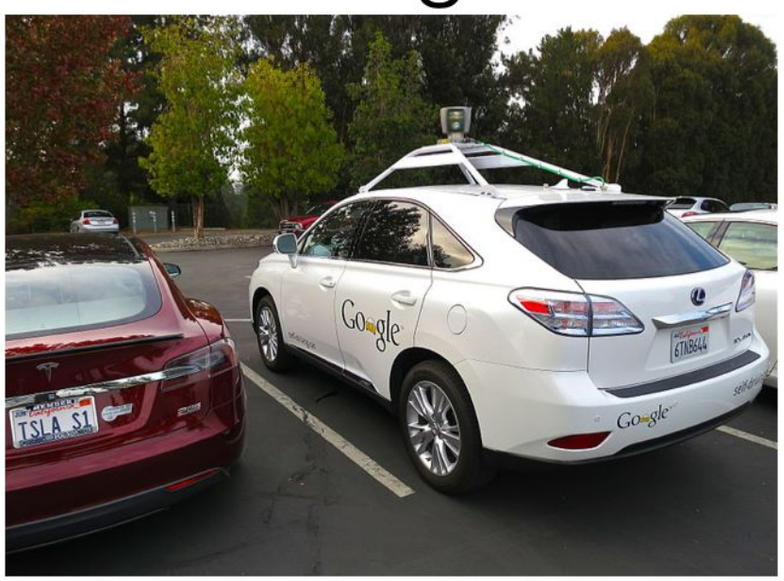


https://flic.kr/p/5BLW6G [CC BY 2.0]

Spam Filtering

Machine Learning Examples

Self-driving cars



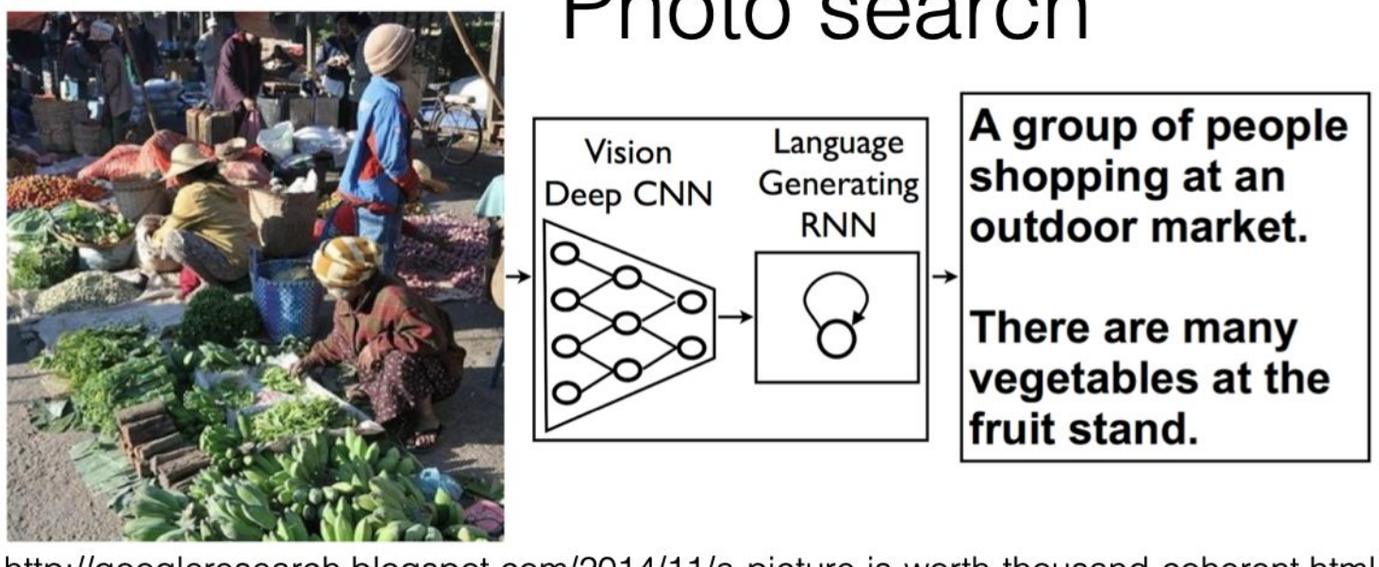
By Steve Jurvetson [CC BY 2.0]

Recommendation systems

http://commons.wikimedia.org/wiki/File:Netflix_logo.svg [public domain]

and many, many more ...

Photo search



http://googleresearch.blogspot.com/2014/11/a-picture-is-worth-thousand-coherent.html

Car Detection for Autonomous Driving

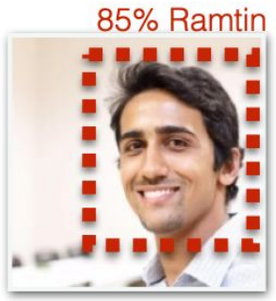
[Deep Learning Specialization]

Kian Katanforoosh

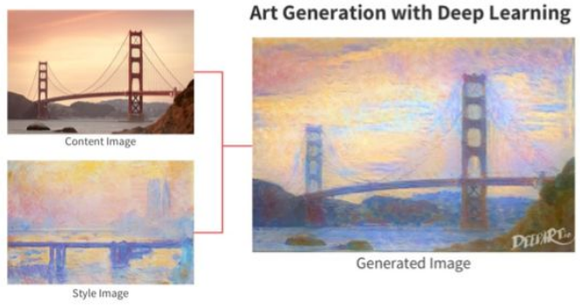
Examples

Optimal goalkeeper shoot prediction

Car detection



Face recognition



Art generation

Music generation

Text generation

“I love you”

Emojifier

Machine translation

Trigger word detection

Definition: Data Science

The application of computational and statistical techniques to address or gain insight into some real-world problem.

Source: J. Zico Kolter, 15-388/688, Practical Data Science

Difference to machine learning

- Machine learning is focused on fancy algorithms
- Data science takes domain knowledge into account
- Data science is applied machine learning (but not only)

Definition: Machine Learning

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

- Arthur Samuel (1959)

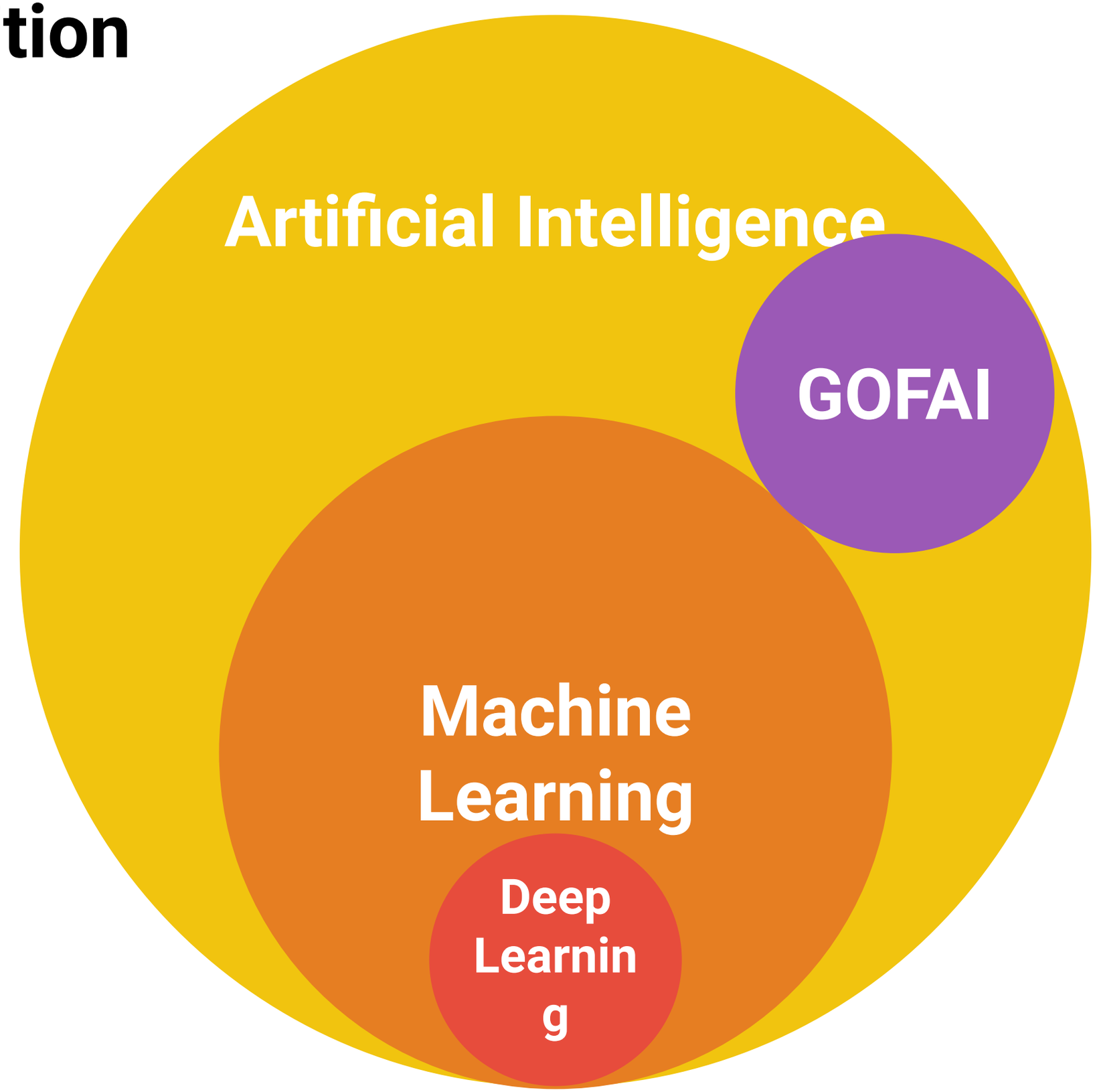
Definition: Machine Learning

"A computer program is said to learn from **experience E** with respect to some class of **tasks T** and **performance measure P** if its performance at tasks in T, as measured by P, **improves with experience E.**"

- Tom M. Mitchell (1997)

His book is a ML classic!

Definition



Goodfellow et al., 2016. Deep Learning.

Difference to statistics

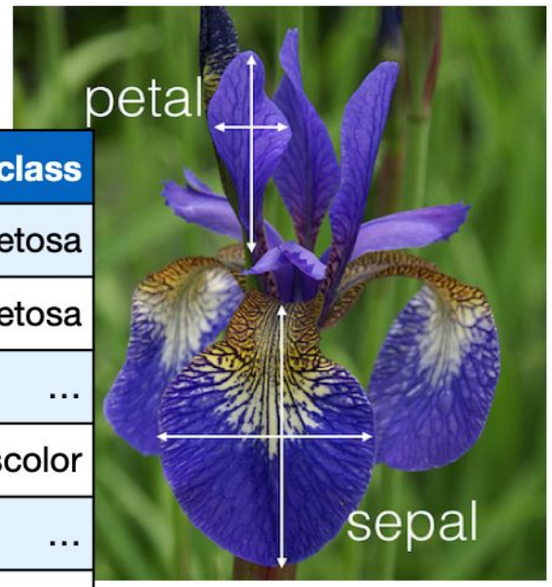
- statistics is about explanation
- more theoretical and mathematical
- data science: more focus on practical problems
- data scraping, data transformation

Nomenclature

https://archive.ics.uci.edu/ml/datasets/Iris

Instances (samples, observations)

	sepal_length	sepal_width	petal_length	petal_width	class
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
...	...	...	...	...	...
50	6.4	3.2	4.5	1.5	veriscolor
...	...	...	...	...	...
150	5.9	3.0	5.1	1.8	virginica



Features (attributes, dimensions)

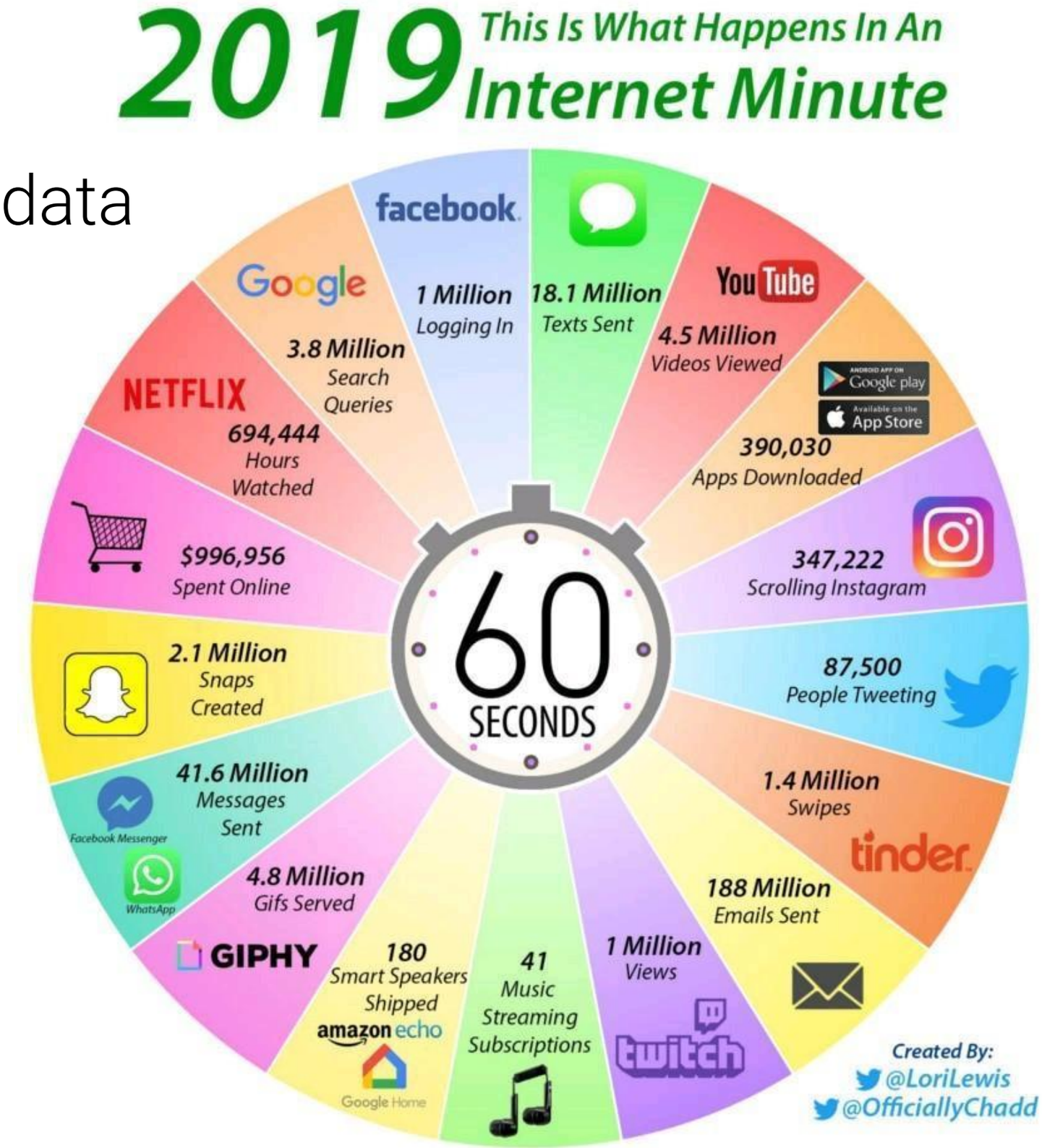
Classes (targets)

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ x_{31} & x_{32} & \cdots & x_{3D} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{bmatrix}$$

$$\mathbf{y} = [y_1, y_2, y_3, \cdots y_N]$$

Difference to big data

- we benefit from big data
- but it is not strictly necessary

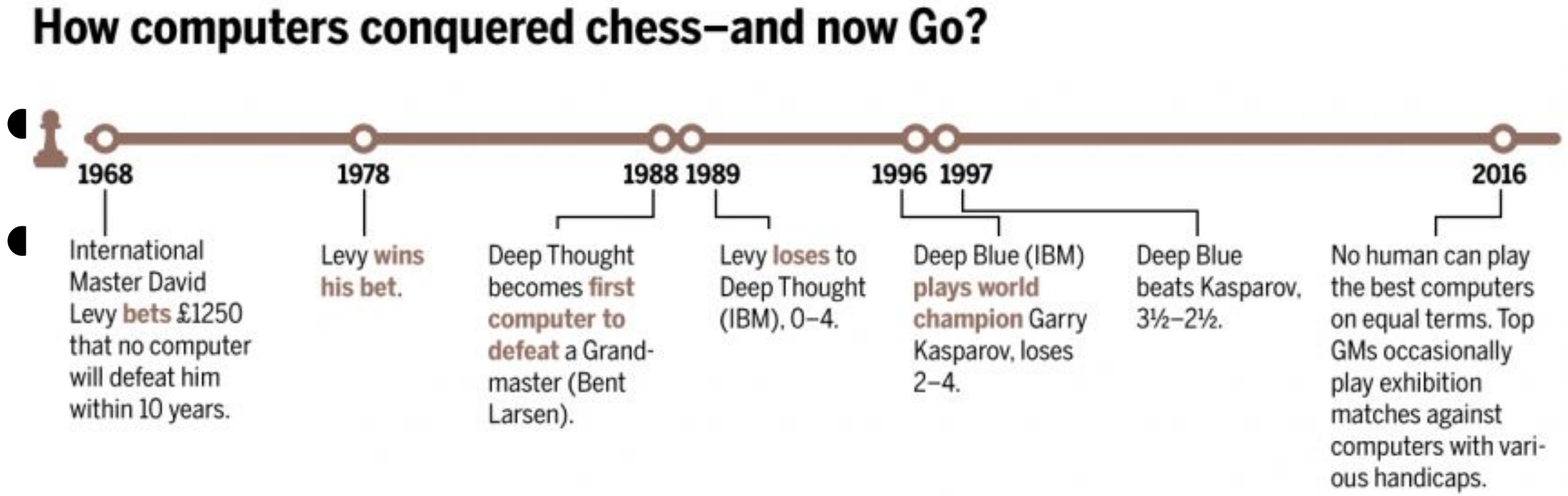


Games

- 1997 Chess, IBM
- 2016 Go, Google

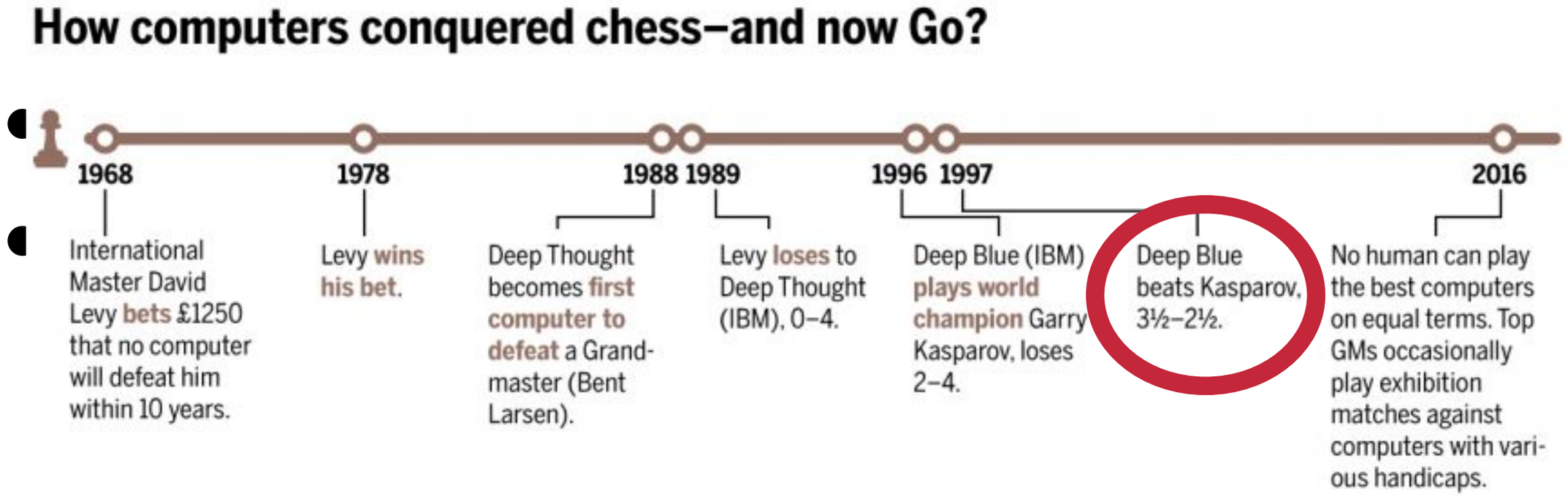
Science Magazine:
http://www.sciencemag.org/news/2016/03/update-why-week-s-man-versus-machine-go-match-doesn-t-matter-and-what-does

Games



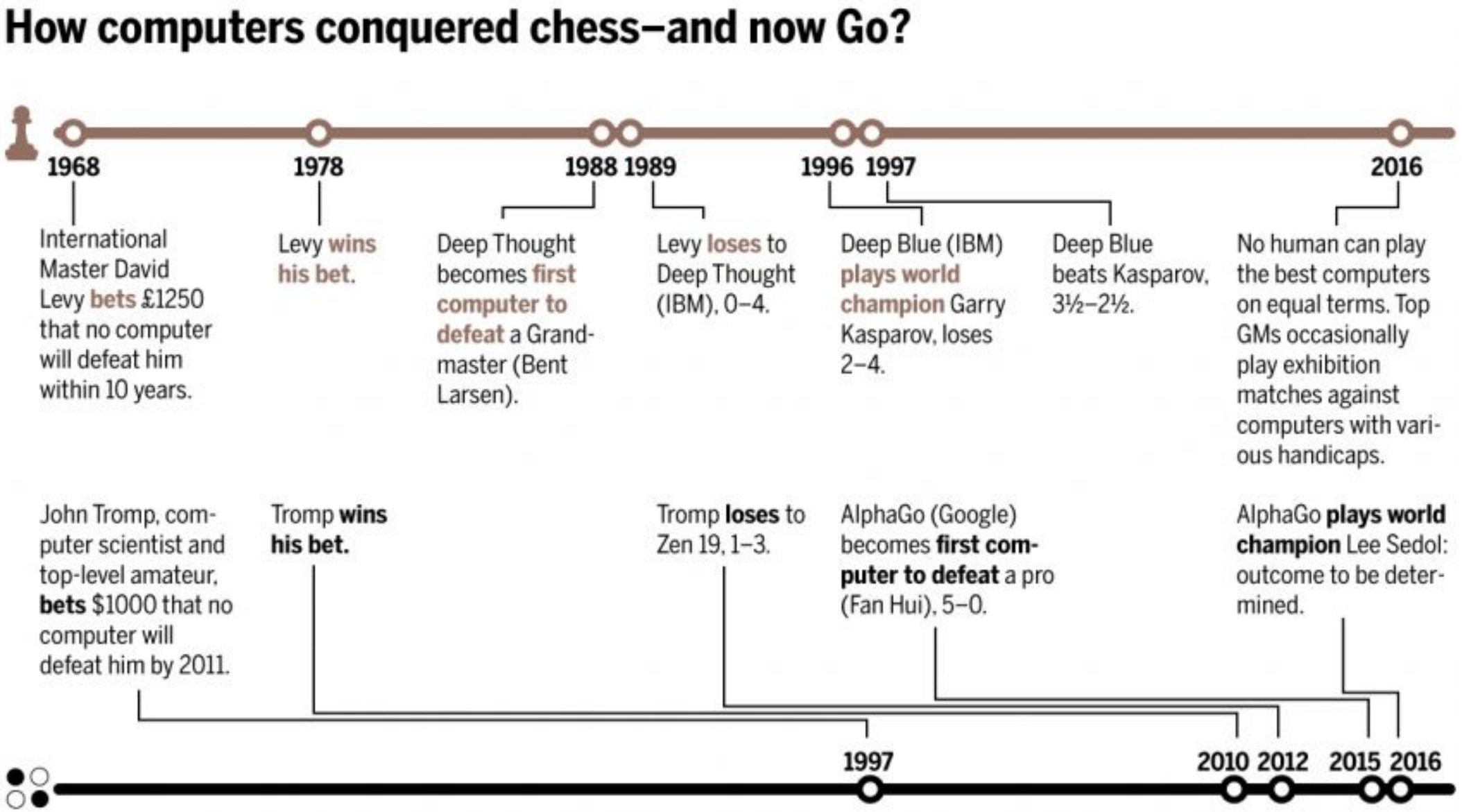
Science Magazine:
http://www.sciencemag.org/news/2016/03/update-why-week-s-man-versus-machine-go-match-doesn-t-matter-and-what-does

Games



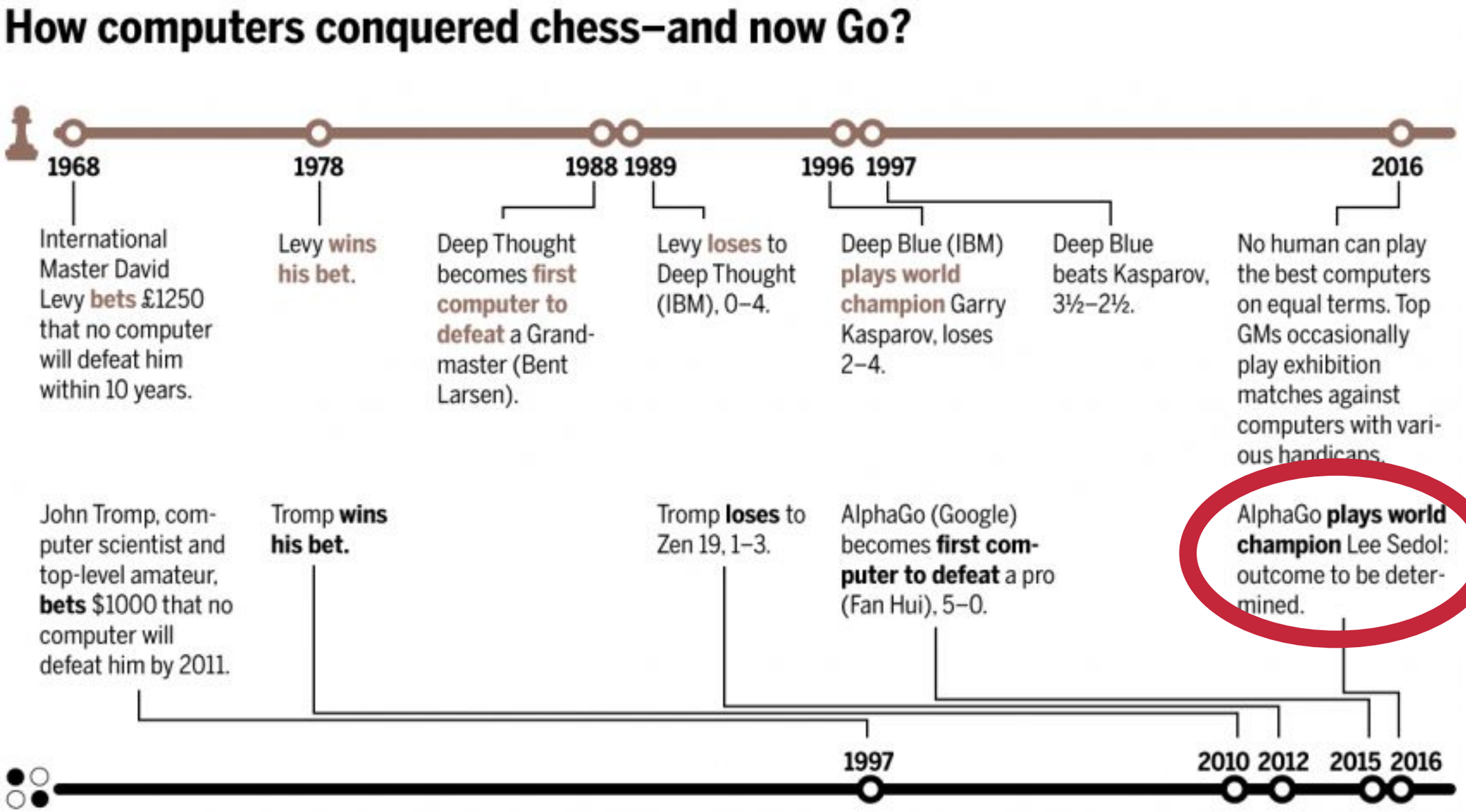
Science Magazine:
http://www.sciencemag.org/news/2016/03/update-why-week-s-man-versus-machine-go-match-doesn-t-matter-and-what-does

Games



Science Magazine:
http://www.sciencemag.org/news/2016/03/update-why-week-s-man-versus-machine-go-match-doesn-t-matter-and-what-does

Games

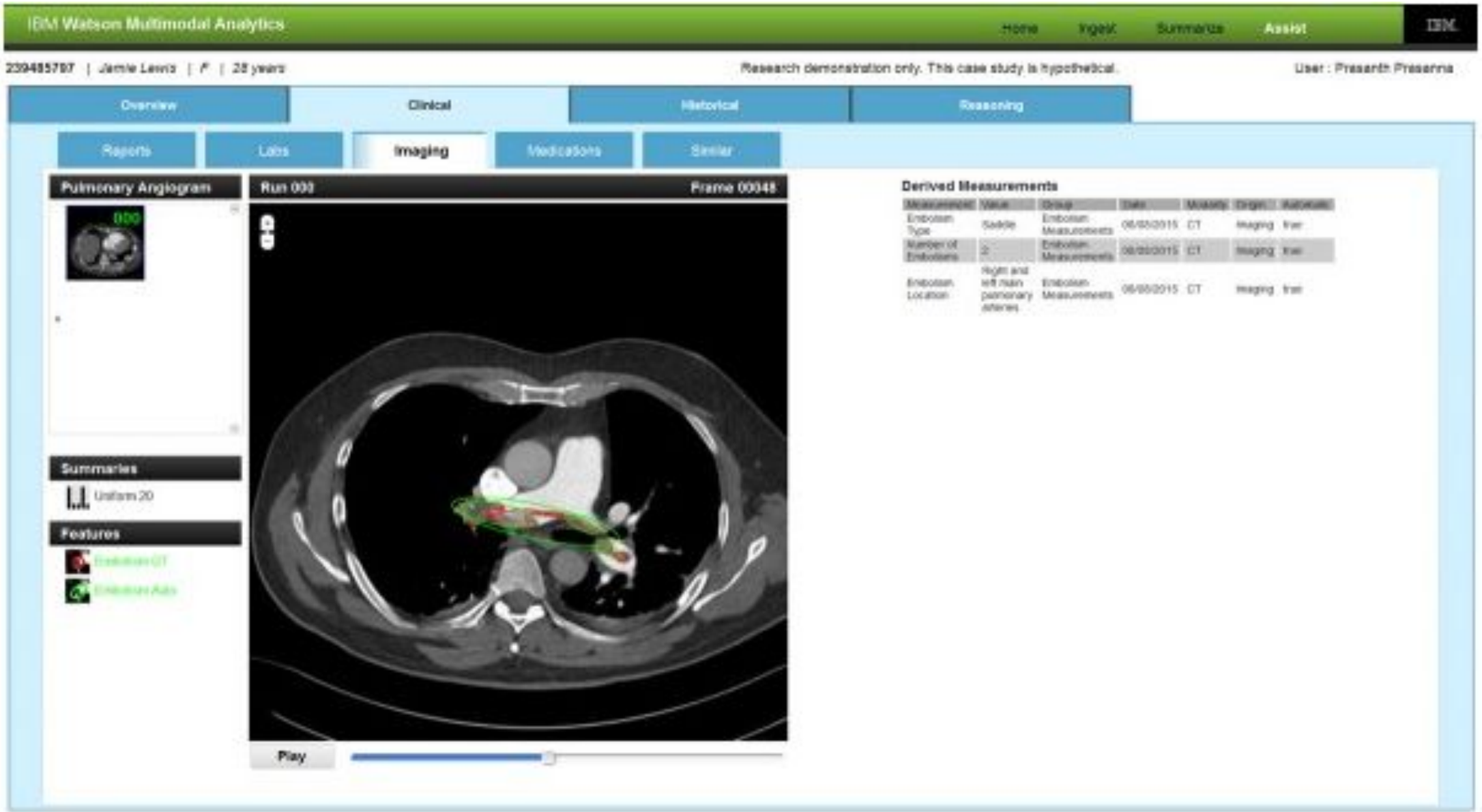


Science Magazine:
http://www.sciencemag.org/news/2016/03/update-why-week-s-man-versus-machine-go-match-doesn-t-matter-and-what-does

Self Driving (autonomous) vehicles

Junior, a robotic Volkswagen Passat, in a parking lot at Stanford University, 24 October 2009,
By: Steve Jurvetson

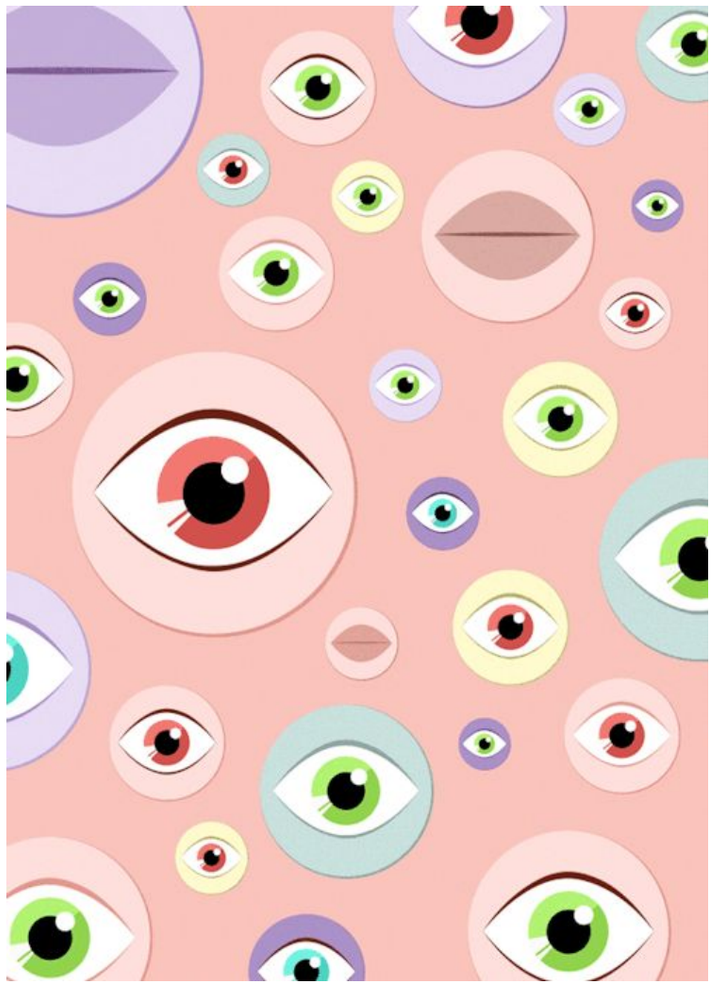
Medicine: Radiology



https://www.technologyreview.com/s/600706/ibms-automated-radiologist-can-read-images-and-medical-records/

Medicine

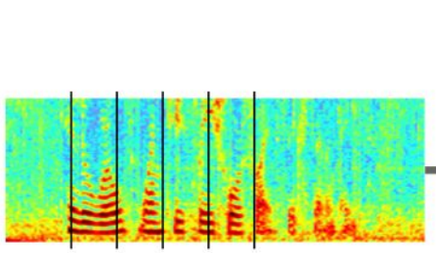
- Glaucoma is the second leading cause of blindness worldwide
- 50% of cases go undetected
- 88,000 retina images



https://twitter.com/IBMWatson/status/844545761740292096

Chatbots

Speech Recognition



Acoustic Input

Deep
Recurrent
Neural Network

"How cold is
it outside?"

Text Output

Reduced word errors by more than 30%

Google Research Blog - August 2012, August 2015

Research at Google

Google Photos Search

Your Photo

Deep Convolutional Neural Network

"ocean"

Automatic Tag

Search personal photos without tags.

Google Research Blog - June 2013

Research at Google

Inbox by Gmail

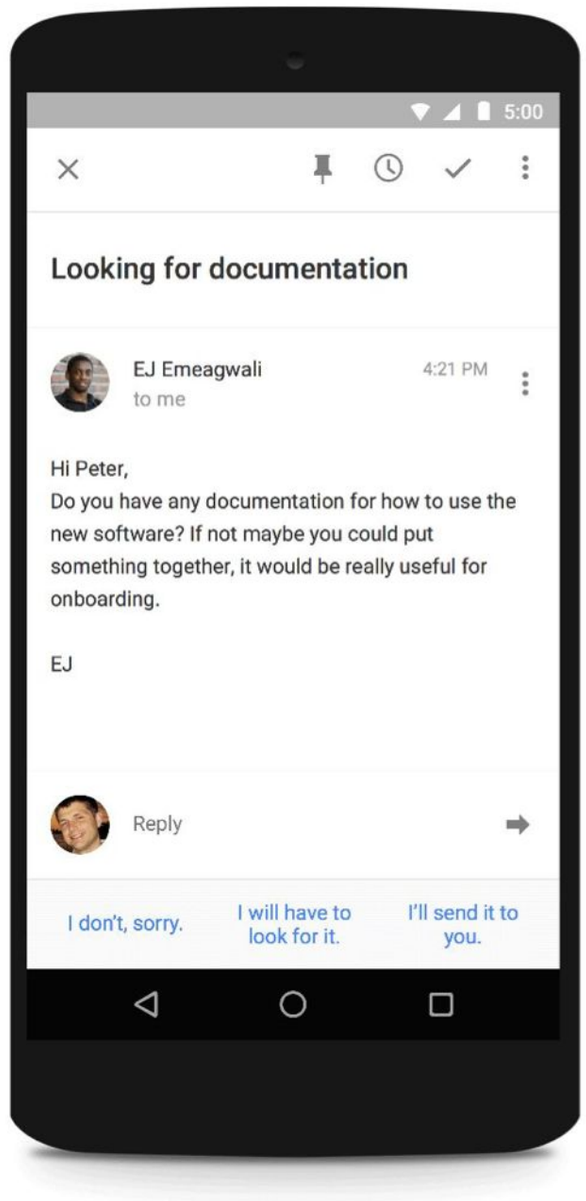
Smart Reply

April 1, 2009: **April Fool's Day joke**

Nov 5, 2015: **Launched Real Product**

Feb 1, 2016: **>10% of mobile Inbox replies**

Research at Google



Discriminative

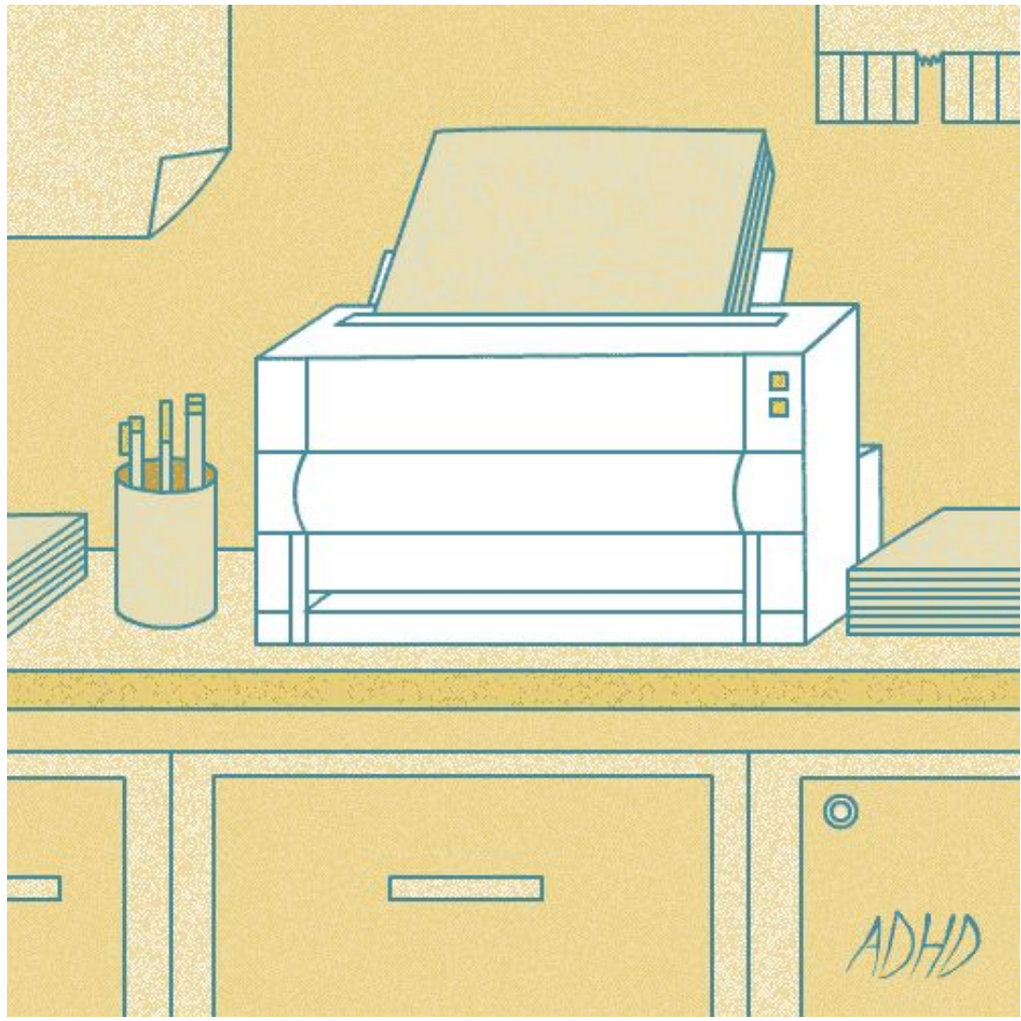
- discriminative models focus on tasks (like sorting examples)

Generative

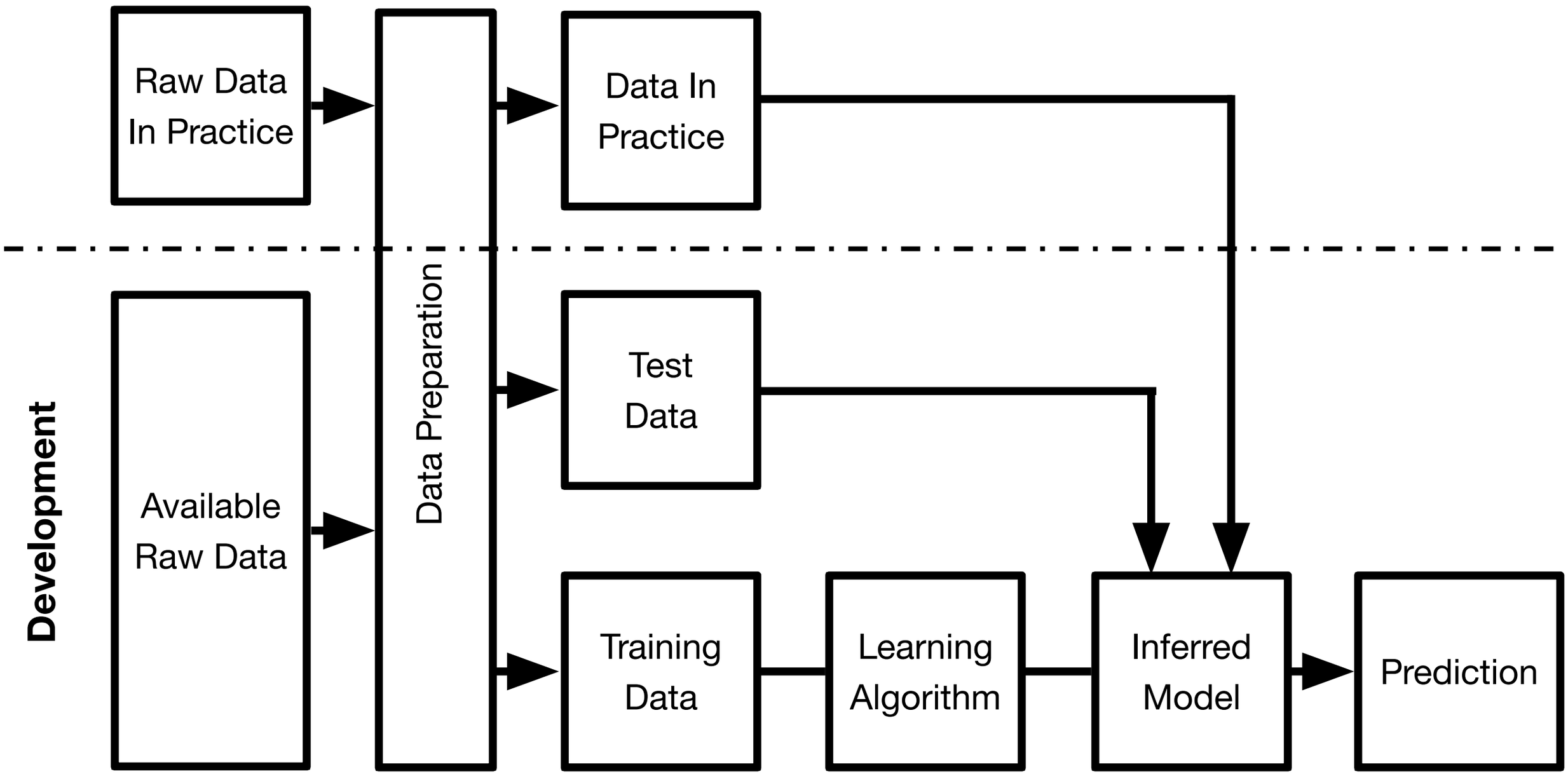
- tell a mythical story to explain the data

No Free Lunch

- “No one model works best for all possible situations.”
- Our model is a simplification of reality
  - Simplification is based on assumptions (model bias)
    - Assumptions fail in certain situations



Technical Model of a Machine Learning System



For ML, we will focus on scikit-learn

scikit learn Install User Guide API Examples Community More Go

scikit-learn

Machine Learning in Python

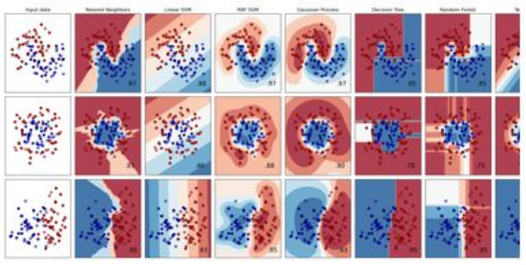
Getting Started | Release Highlights for 1.4 | GitHub

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.
Algorithms: Gradient boosting, nearest neighbors, random forest, logistic regression, and more...

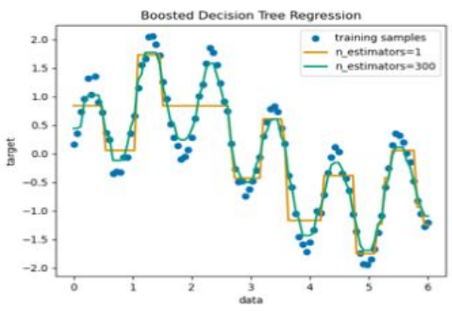


Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: Gradient boosting, nearest neighbors, random forest, ridge, and more...

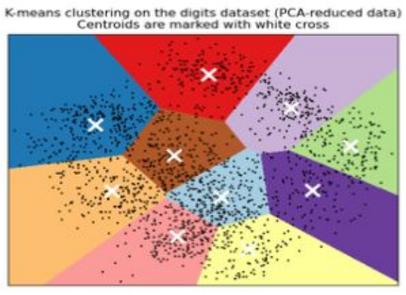


Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, HDBSCAN, hierarchical clustering, and more...



Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Model selection

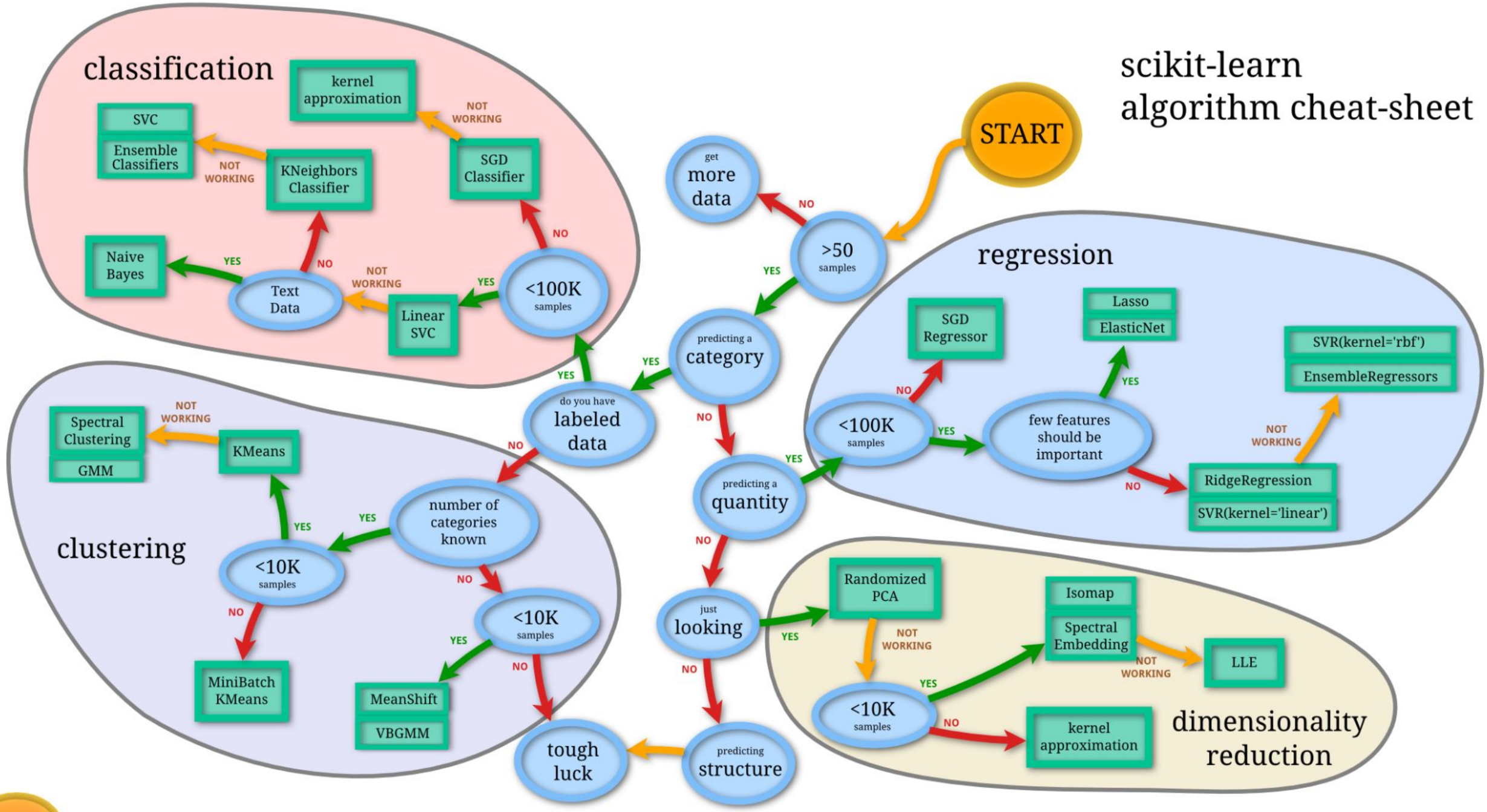
Comparing, validating and choosing parameters and models.

Preprocessing

Feature extraction and normalization.

Applications: Transforming input data such as text

scikit-learn algorithm cheat-sheet



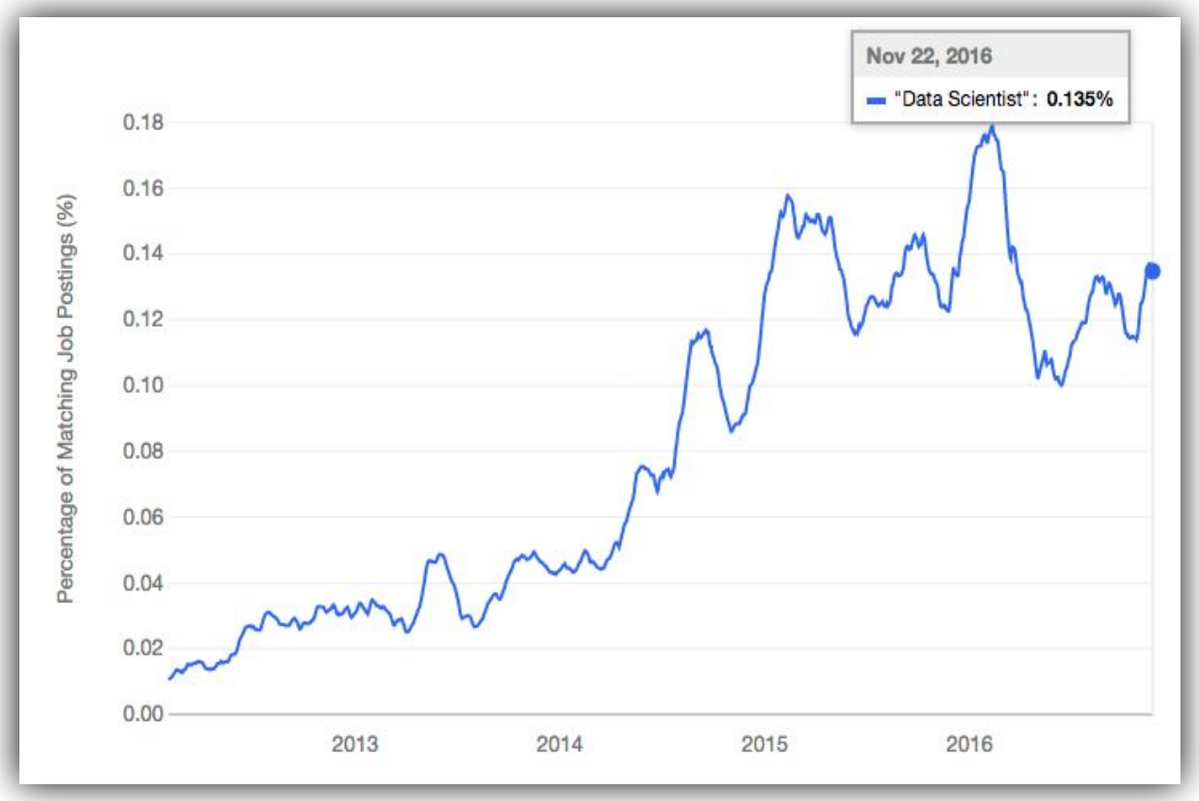
Data Scientists in the Industry

Companies that employ Data Scientists:

- Google, Spotify, Uber, New York Times

Industries that employ Data Scientists:

- Logistics
- Commerce
- Manufacturing
- Finance
- Journalism
- Smart Cities
- Sport Analytics



Job Market

About this course

- Course No.: 03-BE-802.98a
- CP (ECTS): 6
- Module: Depends on your study program, please check on Stud.IP
- Prep course for the WelfareComp Master’s project
- Lectures: Mon 10-12h (MZH 5600)
- Tutorials: Mon 12-14h (MZH 5600)

Credits (ECTS)

6 ECTS == 180 hours

- Lectures: 24h
- Tutorials: 24h
- Learning Python: 6h
- Data Science & Vis Paper Presentation: 12h
- Exposé: 18h
- Progress Presentation: 12h
- Final Presentation: 12h
- Final Report: 24h
- Final Project: 48h

Grading Criteria

- Data science & Vis paper presentation (20%)
- Exposé & exposé presentation (required)
- Interims presentation (required)
- Final presentation (30%)
- Final report (50%)

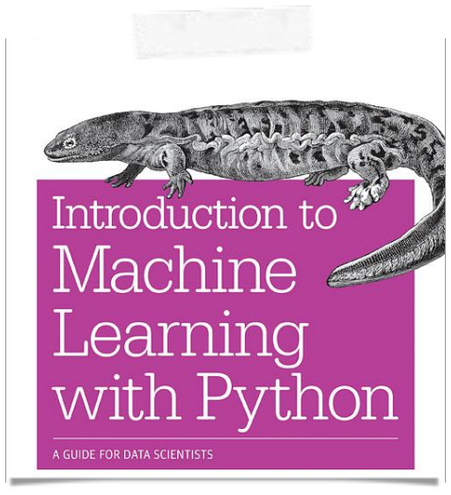
Sessions & Deliverables (draft)

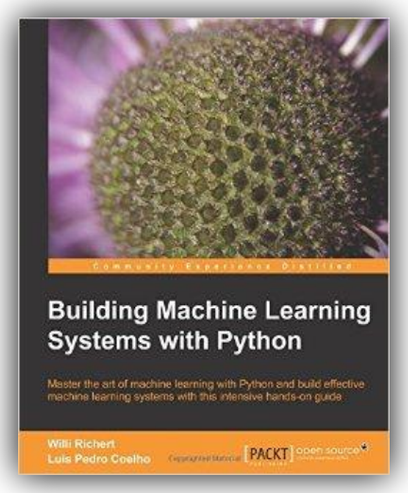
Date	Lecture (10:15-11:45)	Practical (12:15-13:45)
08.04.24	Introduction to Data Science	*Python Introduction*
15.04.24	Basic Statistics & Supervised Learning	*Practical Supervised*
22.04.24	Introduction to Data Visualization	*Practical Visualization*
29.04.24	Exploratory Data Analysis	Text Mining
06.05.24	Unsupervised Learning	**Data Science & Vis Presentation**

19.07.24 Deadline Final Report

Note that for the slots marked red, you are expected to prepare presentations.
For the slots marked blue, you are expected to bring a computer.

Books





Doing Data Science

- Cathy O'Neil & Rachel Schutt
- One of the first books on Data Science
- Schutt was a data scientist with Google

Introduction to Machine Learning with Python

- Andreas C. Müller & Sarah Guido
- From the creator of scikit-learn
- Practical introduction to Python and ML

Building Machine Learning Systems with Python

- Luis Pedro Coelho & Willi Richert
- Practical book on advanced topics

Data Science & Vis Presentation

- In groups of 3,
  you will present a data science & vis paper
- Goal: get a shared understanding of data science & vis
- You have to present the content of a paper
- Argue why it is relevant to data science & vis
- Argue why it is relevant to you
- You have 5 minutes

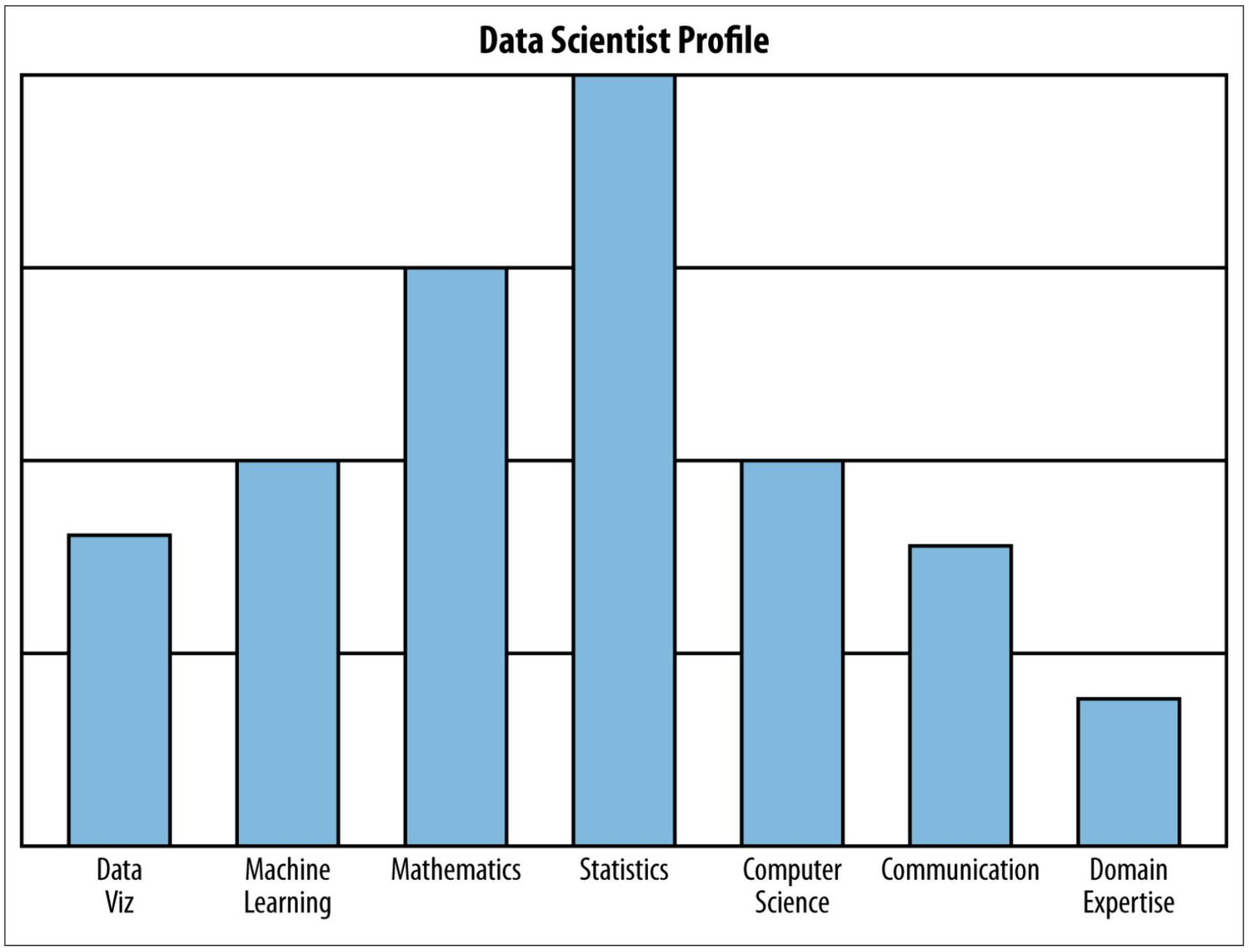
Data Science & Vis Presentation

- **Deadline for topic: 22.04.2024 23:59**
- **Date of the presentation: 06.05.2024**
- Submit your paper choice via e-mail to me (**molina@uni-bremen.de**)
- If you have another paper in mind that you would like to present, please let us know - if we approve it, you can present it

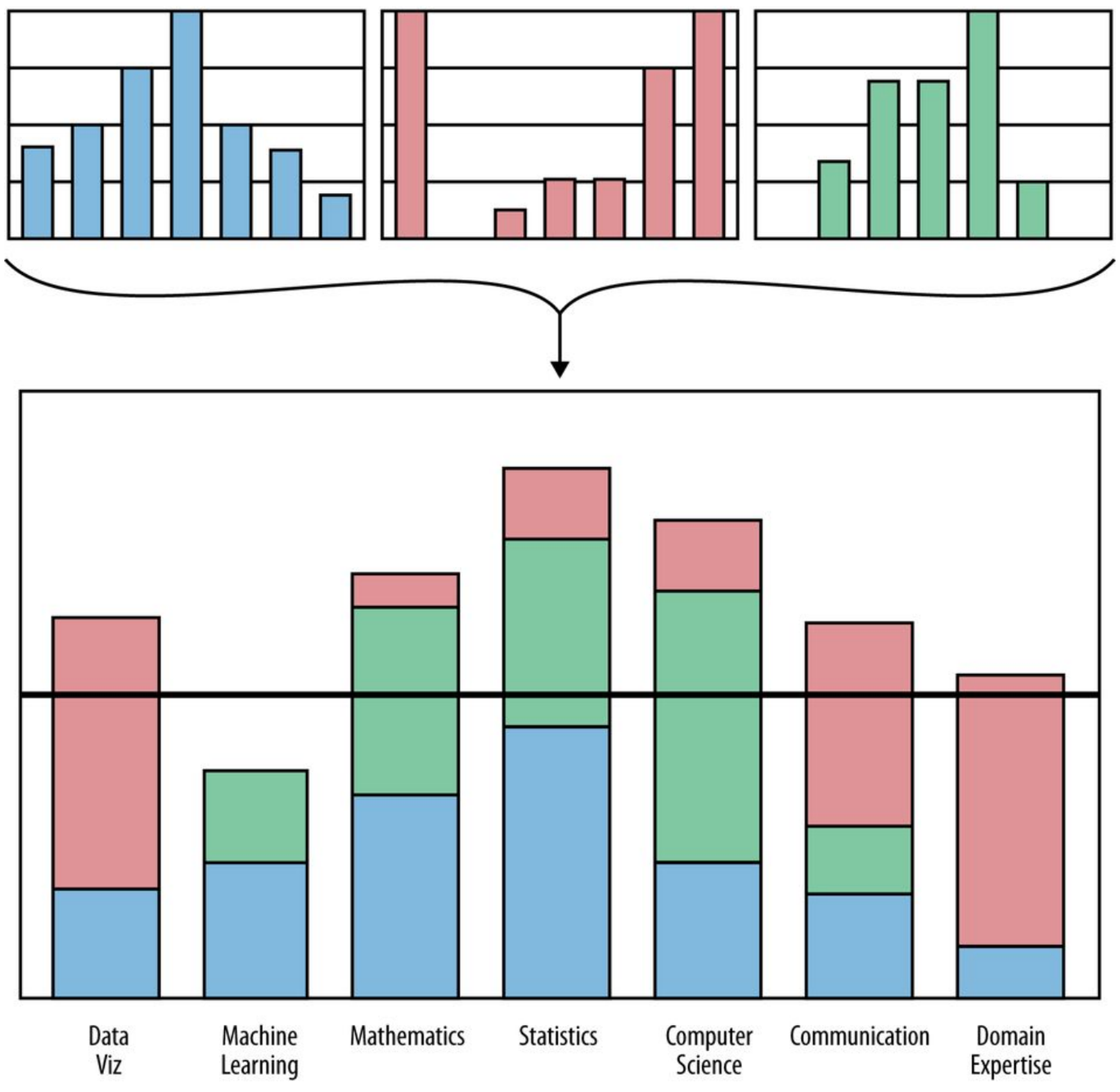
Group Project

- As groups of three people, you will work on a project throughout the course
- Pick a dataset
- Pick a research question
- Pick a suitable method
- Find the best* **analysis and visualization techniques** for your dataset, question, and method
- Write a report on your findings and motivate your choices

Data Science Profile



About your group



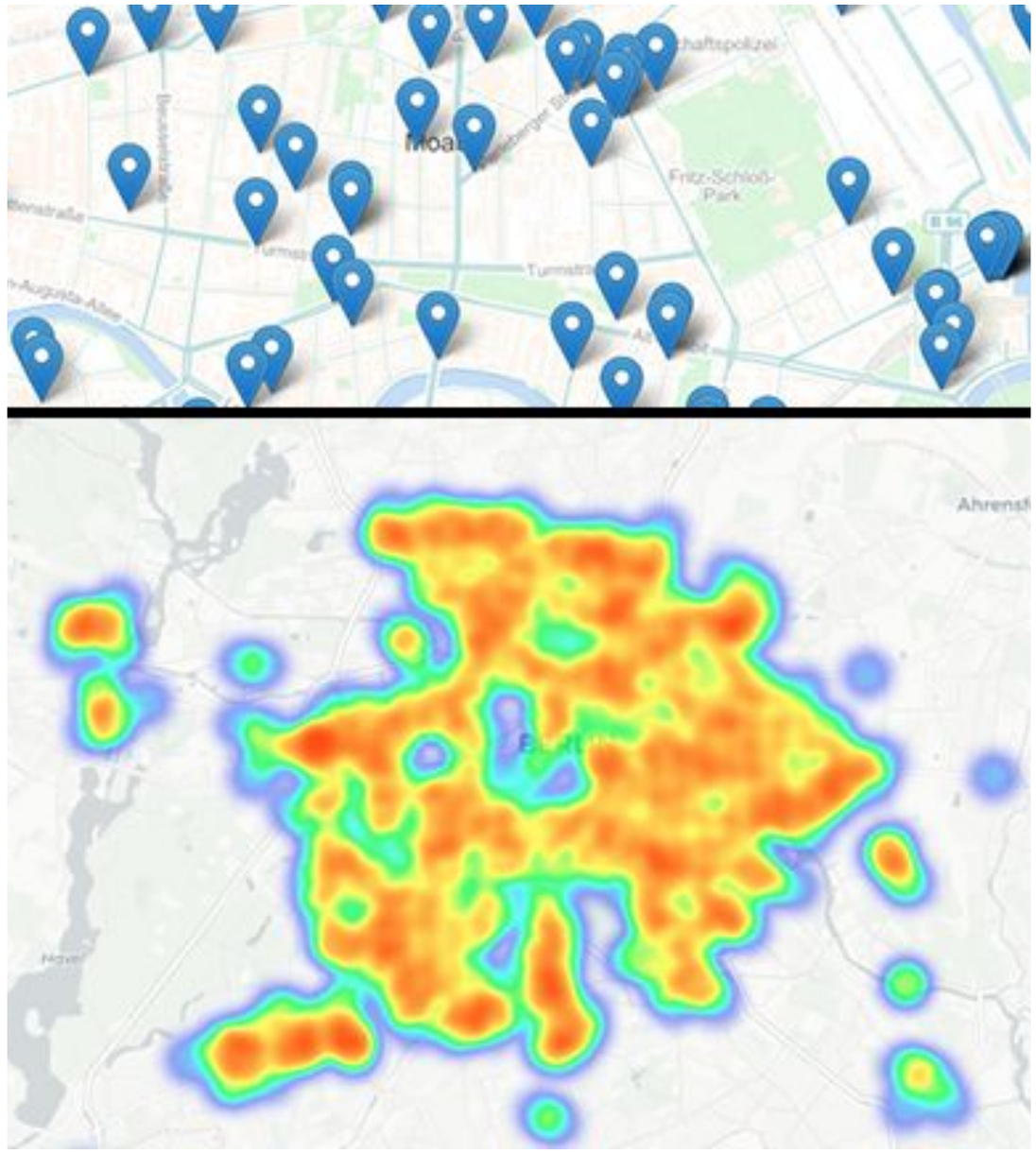
Research-based learning

Past projects in this course

- Exploring Recipes Depending On The Cuisine and The Proximity of Ingredients
- Beach or not beach - Binary Scene Classification
- Understanding The Prevalent Issues of Airline Passengers Based on Customer Reviews
- Using Machine Learning to Help Users Select Emojis

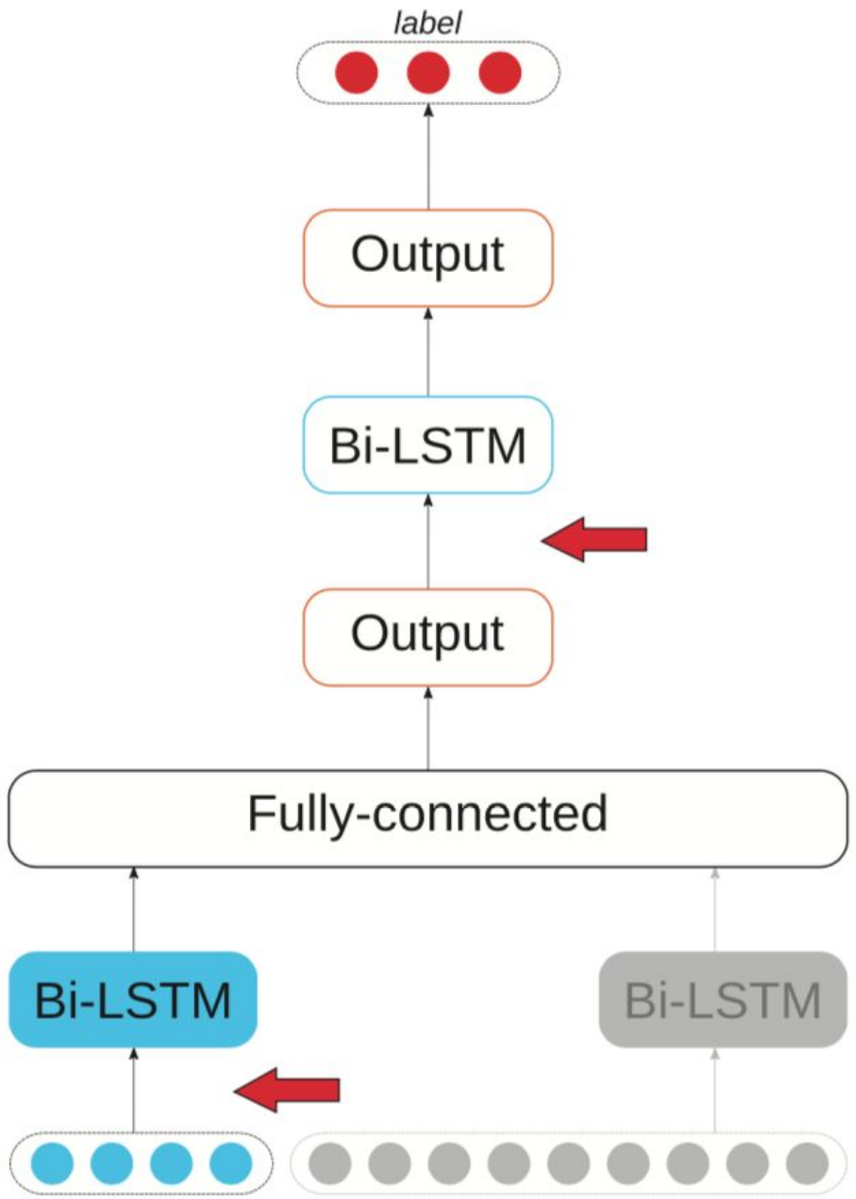
Bachelor's project AViDa

- 12 students build an online information system to predict the behavior of car sharing users
- investigated how date and time, as well as other factors such as weather and major events (e.g., concerts and football matches), affect the use of such services



Master's project CoVis

- 9 students use of machine learning and natural language processing to support social scientists in the analysis of labour laws
  - track changes in legal texts and regulations
  - recognize arguments in unstructured texts
  - discover semantically similar text passages
  - developing an interactive user interface to make the internal processes of the machine learning models transparent and understandable

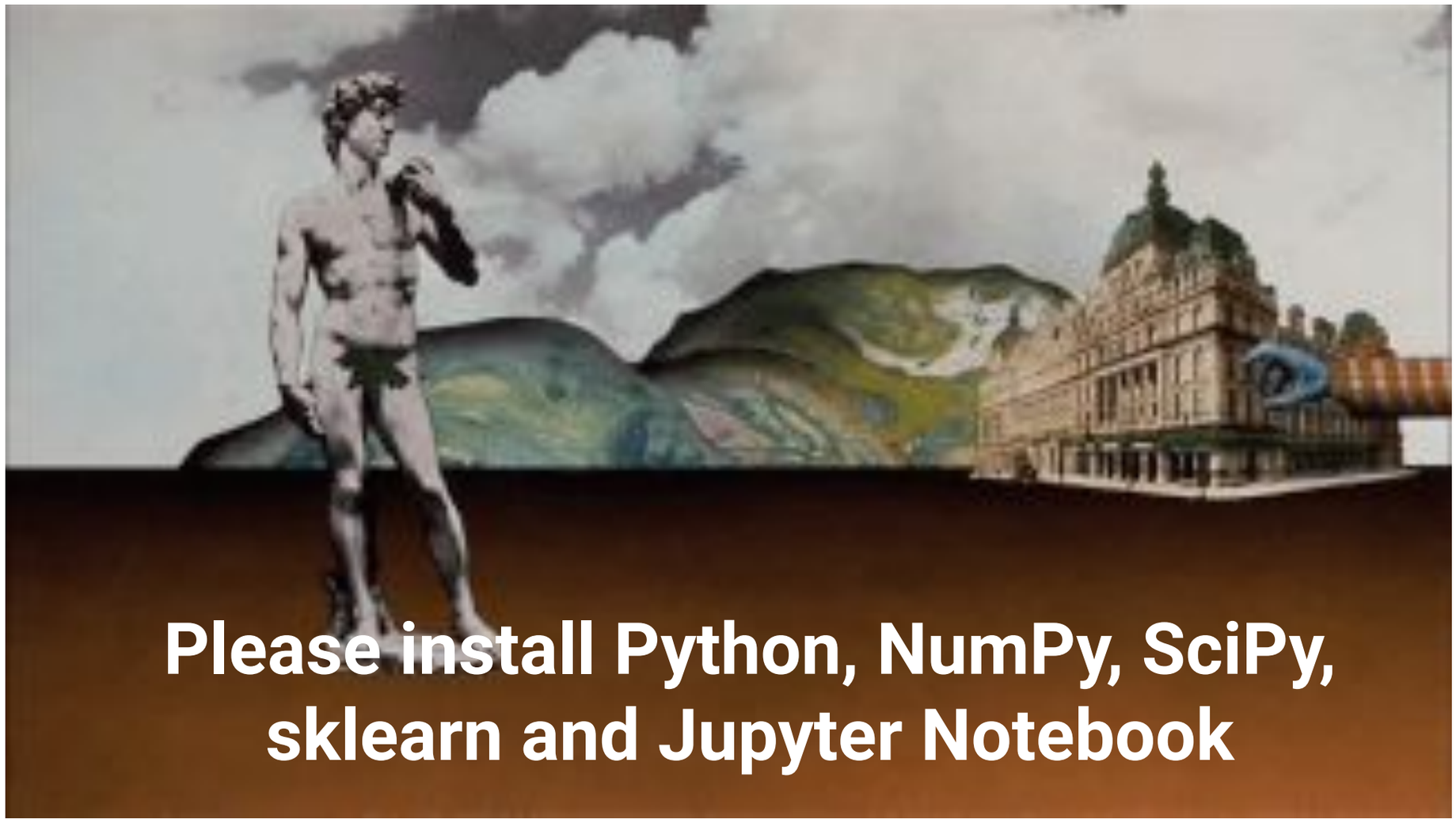


Sessions & Deliverables (draft)

Date	Lecture (10:15-11:45)	Practical (12:15-13:45)
08.04.24	Introduction to Data Science	*Python Introduction*
15.04.24	Basic Statistics & Supervised Learning	*Practical Supervised*
22.04.24	Introduction to Data Visualization	*Practical Visualization*
29.04.24	Exploratory Data Analysis	Text Mining
06.05.24	Unsupervised Learning	**Data Science & Vis Presentation**

19.07.24 Deadline Final Report

Note that for the slots marked red, you are expected to prepare presentations.
For the slots marked blue, you are expected to bring a computer.



Use ANACONDA (it has everything you need)
https://www.anaconda.com/

You can also install the packages individually, different option on each OS
pip, conda (everywhere)
apt-get (Debian, Ubuntu)
brew (MacOSX)