CAP 5768: Introduction to Data Science

Introduction to Data Science

Instructor information

Dr. Hamzah Al-Najada

Assistant Adjunct Professor

Department of Electrical Engineering and Computer Science

Email: halnajada2014@fau.edu

Office: Virtual

Office hours: By Appointment via Zoom

TA: Tuan Vo

Email: tvo2019@fau.edu

Course structure and prerequisites

One two-hour and 40-minute lecture (two ten-minute breaks) every Thursday 4:00pm-6:50pm

All assignments posted and submitted to Canvas

Prerequisite: programming competency at the level of an online short course (*e*.*g*., Code Academy)

Course objectives

Apply the Python 3 and its associated packages to perform an array of data analysis techniques.

Effectively manipulate, curate, visualize, and explore data and draw conclusions from this data.

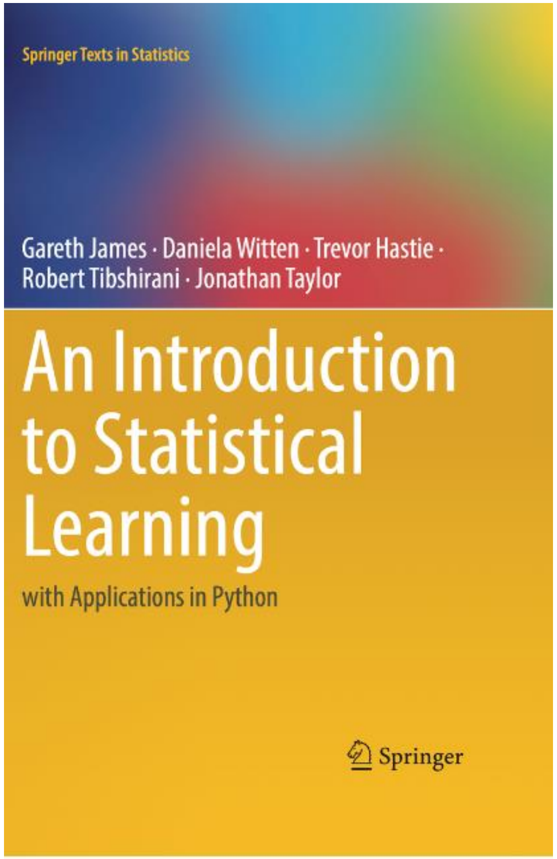
Identify appropriate statistical models to address diverse problems in data analytics.

Textbook

An Introduction to Statistical Learning with Applications in Python, by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Springer, 2023.

**** FREE ebook from author website**

https://statlearning.com



Springer Texts in Statistics

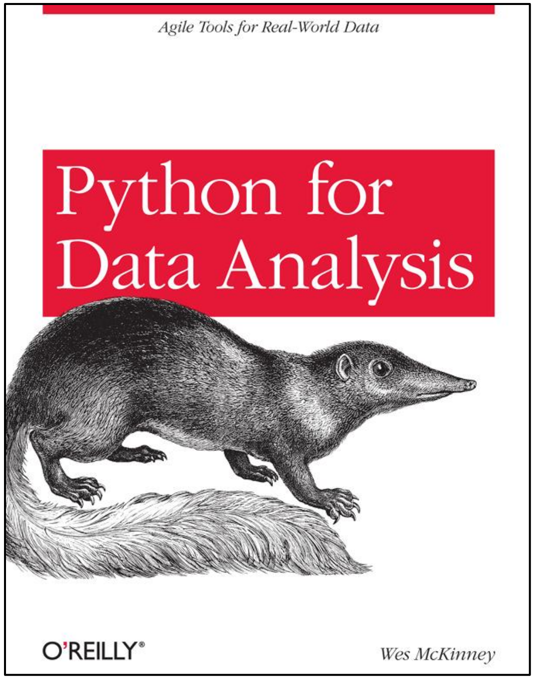
Gareth James · Daniela Witten · Trevor Hastie ·
Robert Tibshirani · Jonathan Taylor

An Introduction to Statistical Learning

with Applications in Python

Springer

Textbook - Python

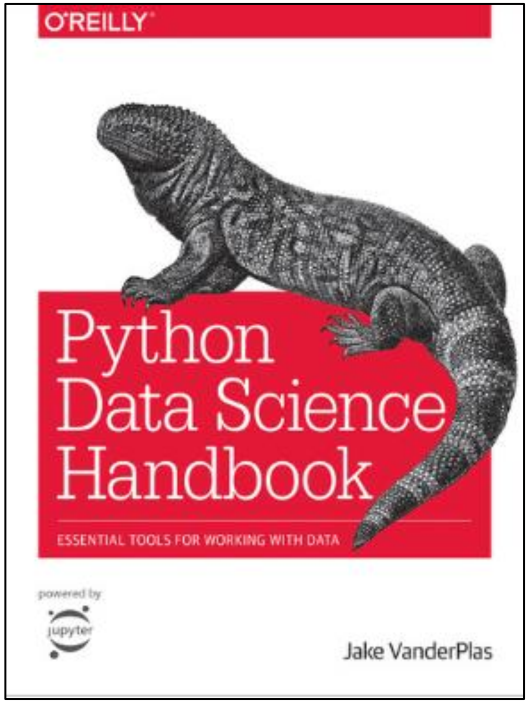


Agile Tools for Real-World Data

Python for Data Analysis

O'REILLY®

Wes McKinney



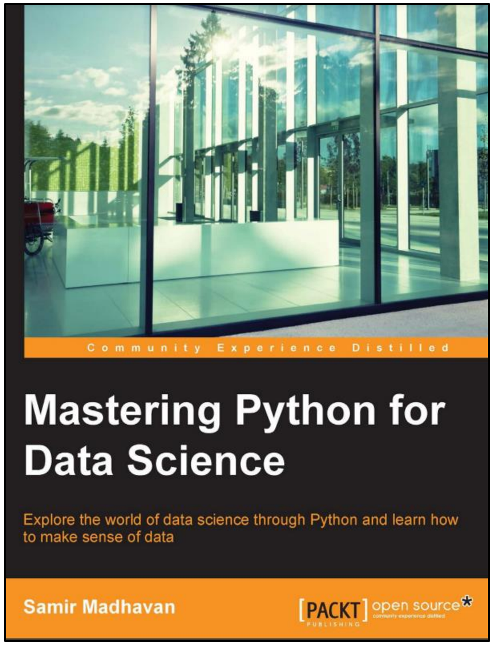
O'REILLY®

Python Data Science Handbook

ESSENTIAL TOOLS FOR WORKING WITH DATA

powered by jupyter

Jake VanderPlas



Community Experience Distilled

Mastering Python for Data Science

Explore the world of data science through Python and learn how to make sense of data

Samir Madhavan

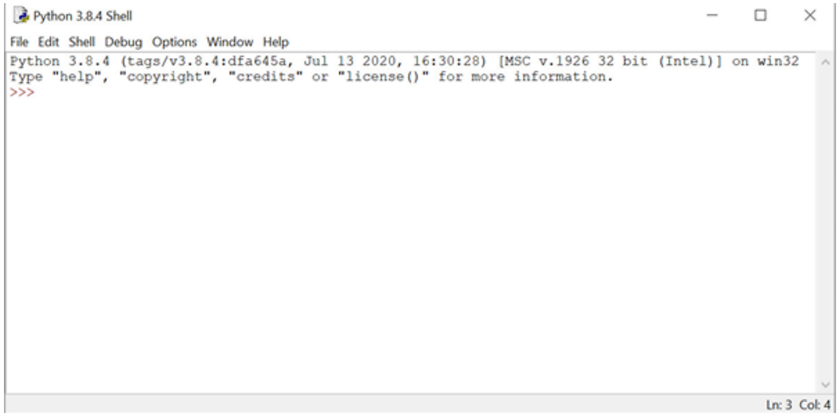
[PACKT] PUBLISHING open source*

Using Python

Option1:

Installing Python

https://www.python.org/downloads/windows/



```
Python 3.8.4 (tags/v3.8.4:dfa645a, Jul 13 2020, 16:30:28) [MSC v.1926 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
```

Then you can install the IDE of your comfort,

PyCharm, VS Code, Spyder …

Using Python

Option2:

Installing **Anaconda**

Anaconda is a distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment.

https://docs.anaconda.com/free/anaconda/install/index.html

Then you can install the IDE of your comfort,

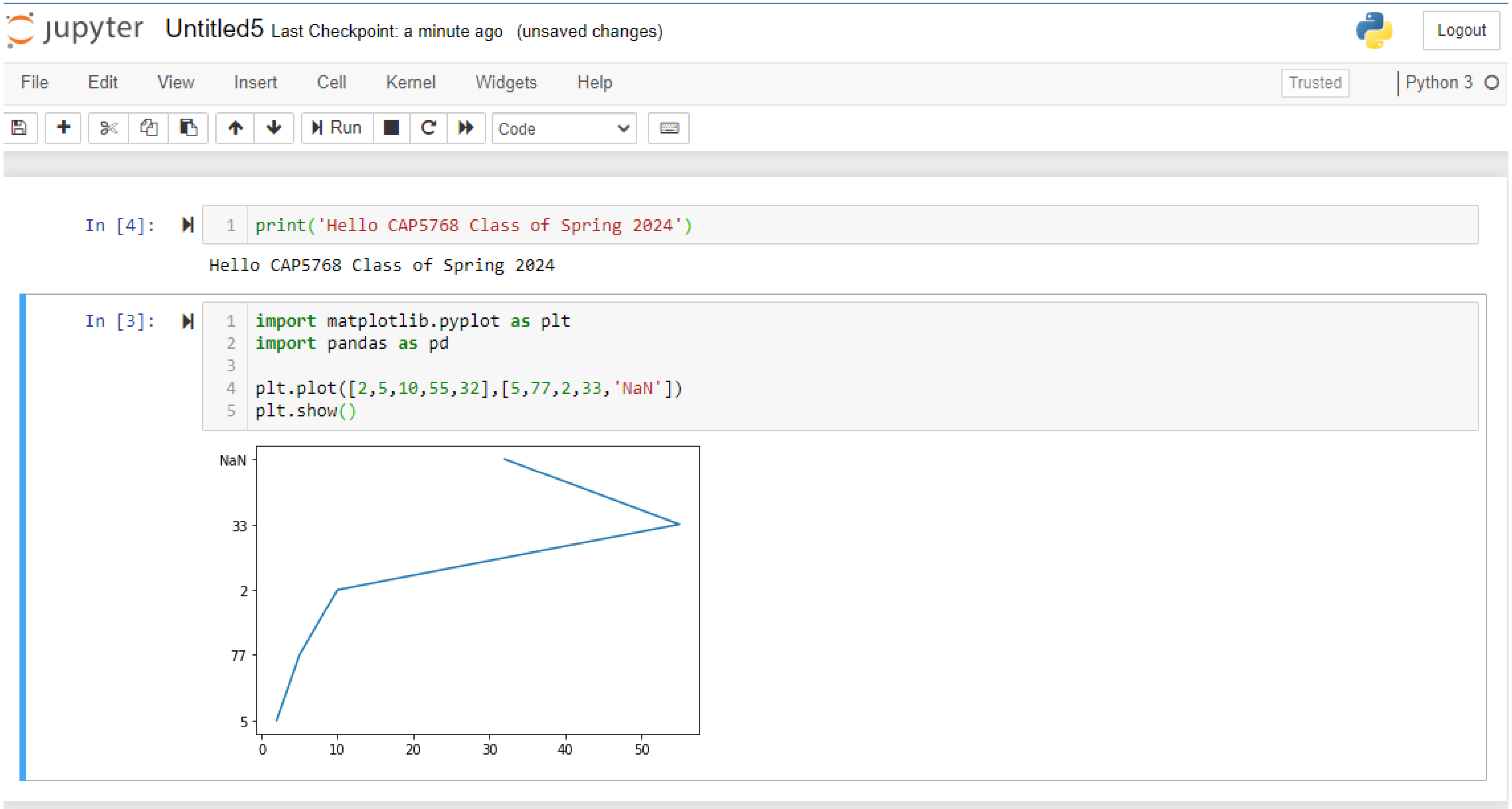
PyCharm, VS Code, Spyder …

Recommended to use Jupyter Notebook for

Interactive computing environment.

Jupyter Notebook

Open-source project to develop open-source software, open standards, and services for interactive computing across multiple programming languages.



In [4]:

```
print('Hello CAP5768 Class of Spring 2024')
```

```
Hello CAP5768 Class of Spring 2024
```

In [3]:

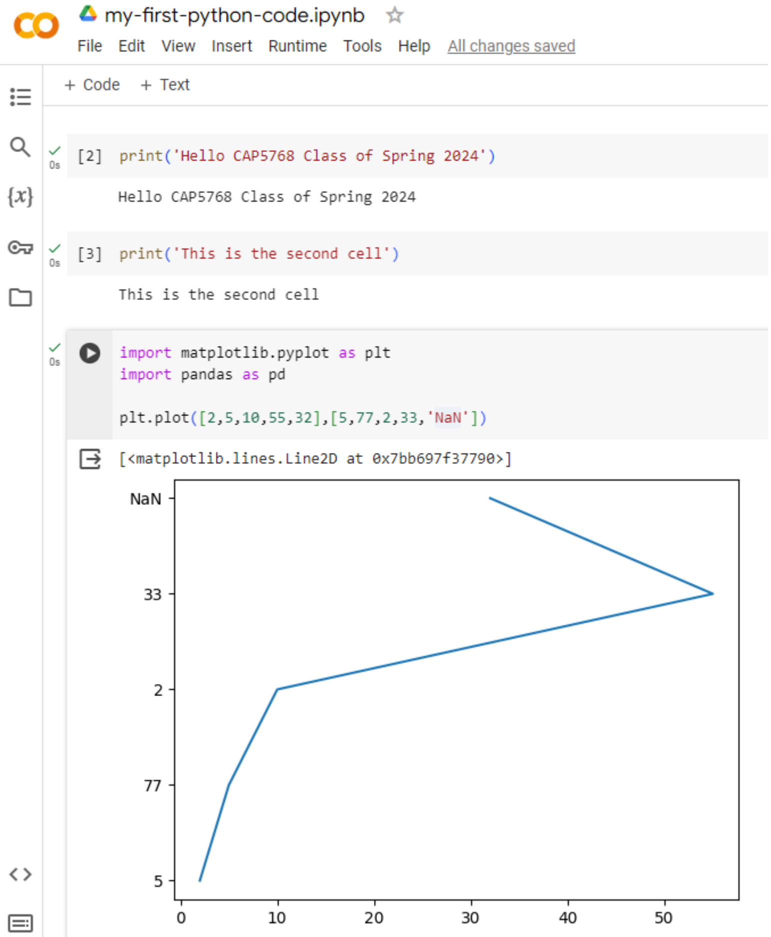
```
import matplotlib.pyplot as plt
import pandas as pd

plt.plot([2,5,10,55,32],[5,77,2,33,'NaN'])
plt.show()
```

Using Python

Option3:

Using the cloud Google Colab

https://colab.research.google.com/



```
[2]  print('Hello CAP5768 Class of Spring 2024')

     Hello CAP5768 Class of Spring 2024
```

```
[3]  print('This is the second cell')

     This is the second cell
```

```
import matplotlib.pyplot as plt
import pandas as pd

plt.plot([2,5,10,55,32],[5,77,2,33,'NaN'])
```

```
[<matplotlib.lines.Line2D at 0x7bb697f37790>]
```

Course evaluation

Your final grade will be based on the following weighted distribution:

Assessment	**Weight (%)**
Assignments and Quizzes	35%
Research Paper	10%
Term Project	30%
Final Exam	25%
Bonus	5%
Total	**105%**

Homework assignments will consist of data manipulation, curation, visualization, and analysis exercises using simulated and real datasets.

Course grading

A	[93%, 100%]
A-	[90%, 93%)
B+	[87%, 90%)
B	[83%, 87%)
B-	[80%, 83%)
C+	[77%, 80%)
C	[73%, 77%)
C-	[70%, 73%)
D+	[67%, 70%)
D	[63%, 67%)
D-	[60%, 63%)
F	[0%, 60%)

Square brackets indicate inclusive

Parentheses indicate exclusive

Grading policies

Late assignments will not be accepted.

Incomplete grades are given only if there is solid evidence of medical or otherwise serious emergency and the student is currently passing the class.

Academic integrity

Cheating is unfortunately a big issue currently.

All code and solutions for assignments must be your own work .

Resemblance of code and/or solutions to those of online resources or other students will not be tolerated.

These will be evaluated on a case-by-case basis and may be reported to the Department Chair, resulting in a zero on the assignment, F in the course, and potential further administrative action.

It’s not worth it!

List of topics

Topic 1: Introduction to data science

Topic 2: Introduction to Python and data visualization

Topic 3: Data transformations

Topic 4: Exploratory data analysis

Topic 5: Linear regression

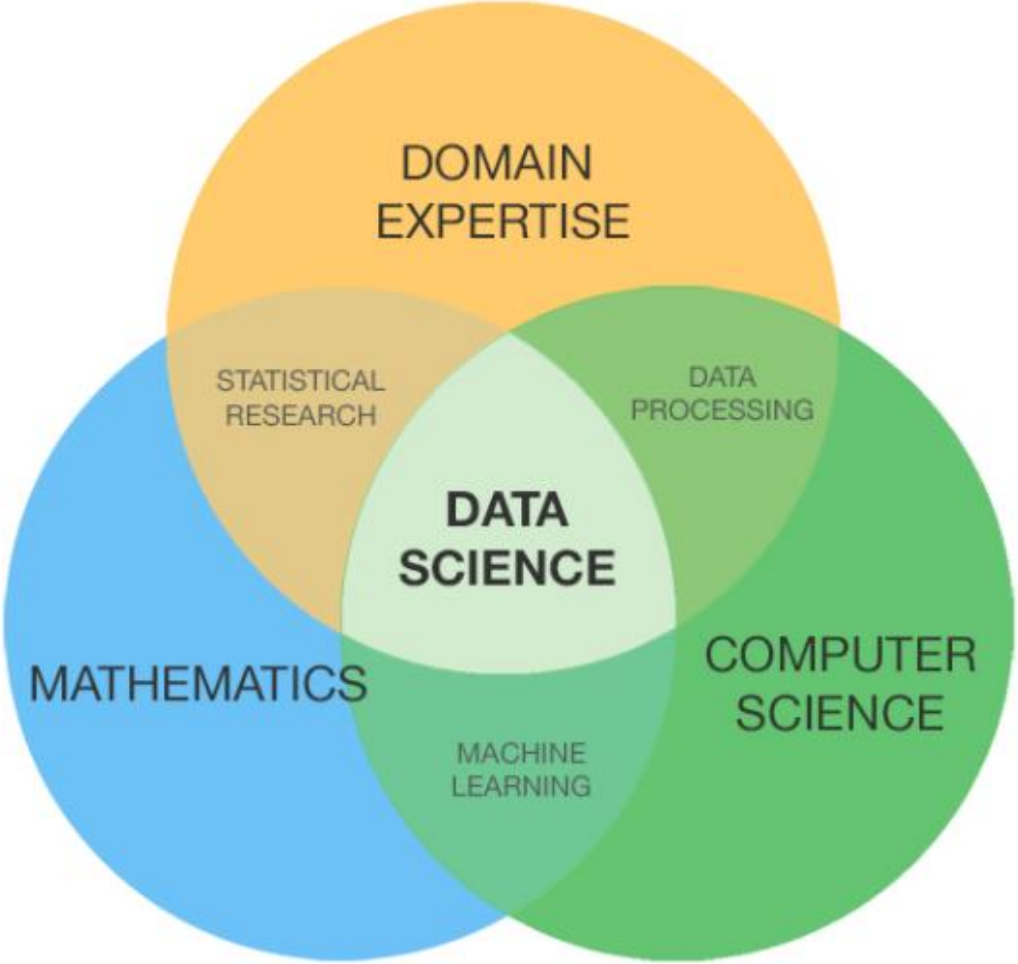
Topic 6: Classification with logistic regression

Topic 7: Model selection, feature selection, and regularization

Topic 8: Unsupervised learning

What is data science?

Data science represents the tools, methodologies, and theory for uncovering patterns in data combined with domain expertise to make actionable predictions and recommendations.



https://towardsdatascience.com/data-science-interview-guide-4ee9f5dc778

What is data science?

1.Data Science: An interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. It combines various fields including statistics, data analysis, machine learning, and computer science to analyze and interpret complex data.

2.Statistical Learning: A set of tools for modeling and understanding complex datasets. It is a subfield of statistics.

3. Artificial Intelligence (AI): AI refers to the development of computer systems that can perform tasks that normally require human intelligence.

4.Machine Learning (ML): A subset of AI that involves the use of statistical techniques to enable machines to improve at tasks with experience, with minimal human intervention.

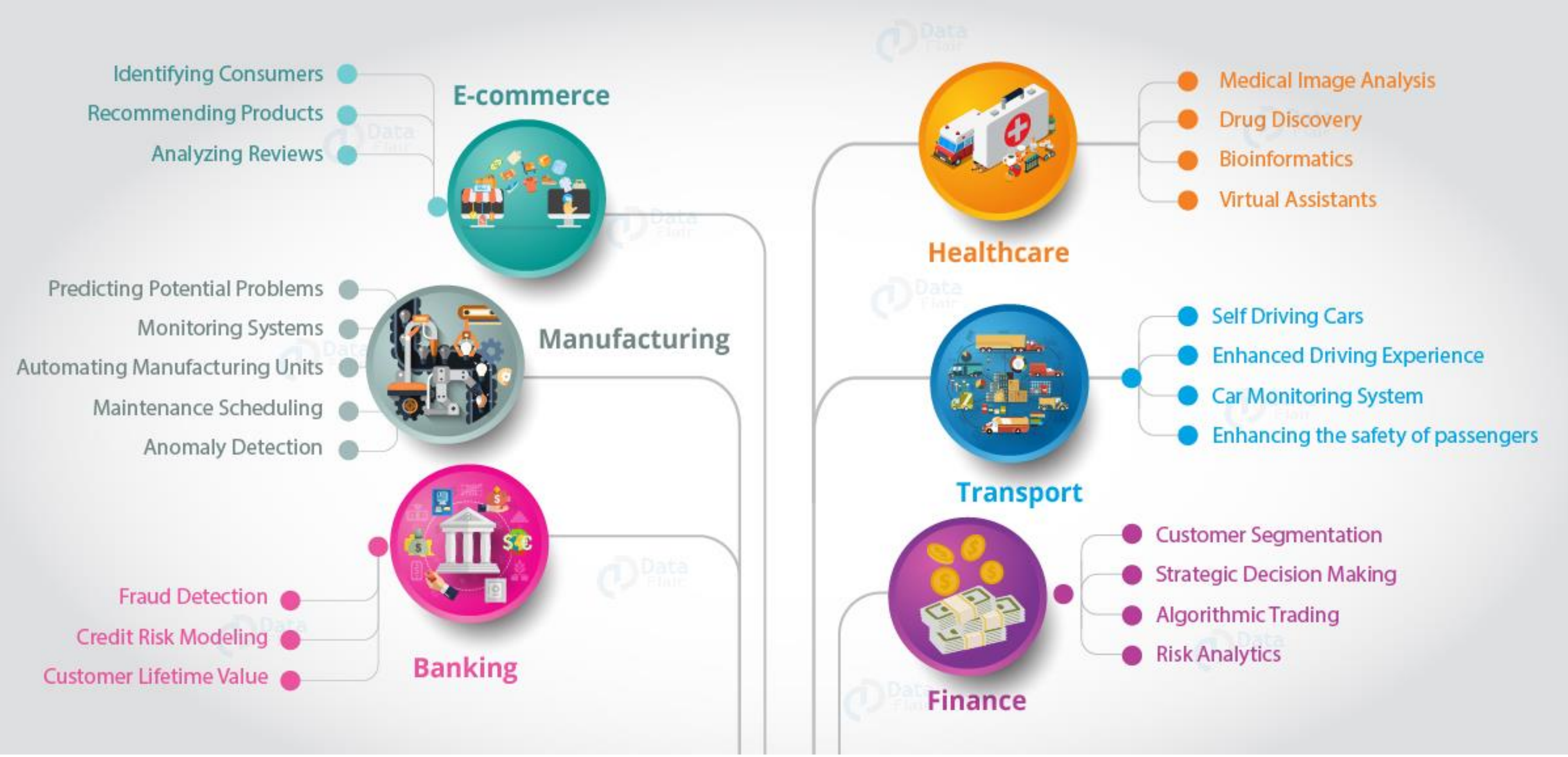
Why is data science important?

A number of driving forces, including:

- Big data, which encompasses enormous volumes of highly complex (dimension and structure) and dynamic data
- Explosive growth of data across many fields due to cheaper storage with expanded capacity, faster communication, and better management of databases
- Rapid increases in computing power

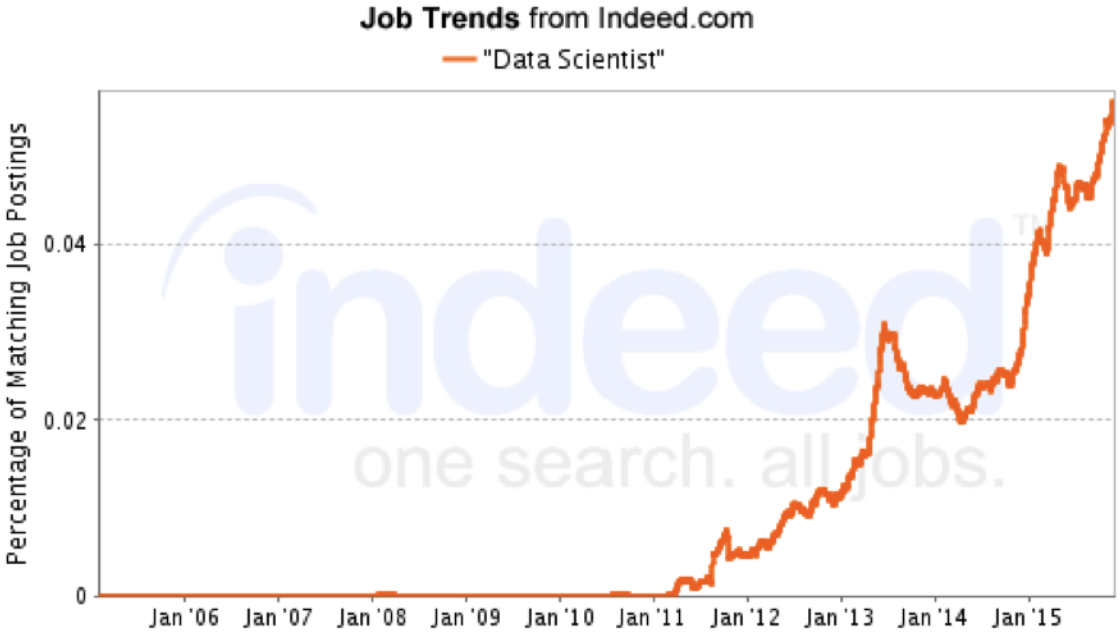
Where is data science applied?

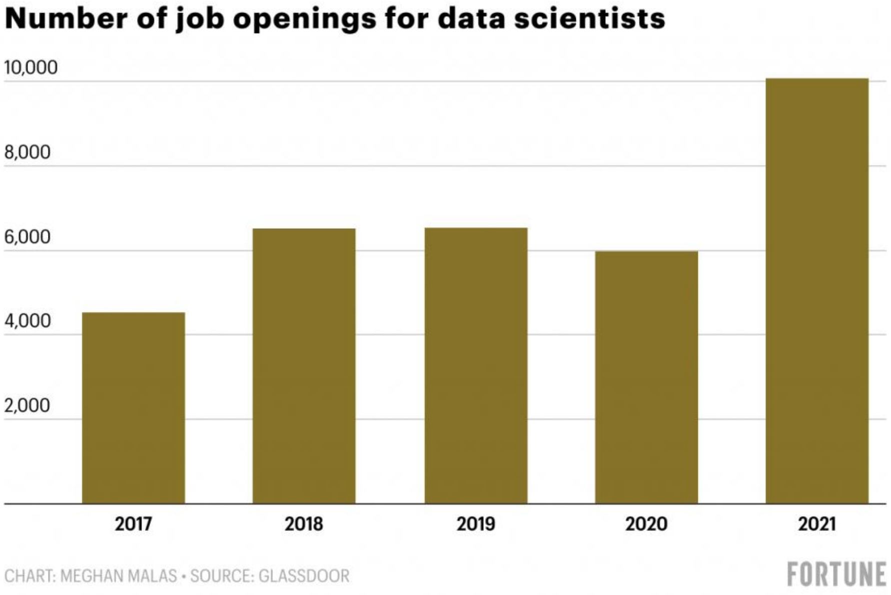
Everywhere!



https://data-flair.training/blogs/data-science-applications/

High demand for data scientists





https://towardsdatascience.com/how-to-choose-a-data-science-job-53007d7f195f

https://fortune.com/education/articles/glassdoors-no-3-best-job-in-the-u-s-has-seen-job-growth-surge-480/

High salaries for data scientists

The mean data scientist salary based on Glassdoor is $120,000.

The reason for such high salaries is that organizations are beginning realize the power of big data and wish to employ such data as an engine for making intelligent business decisions.

Moreover, because the field is not saturated, the demand is high for data professionals, leading to starting salaries upward of $95,000.

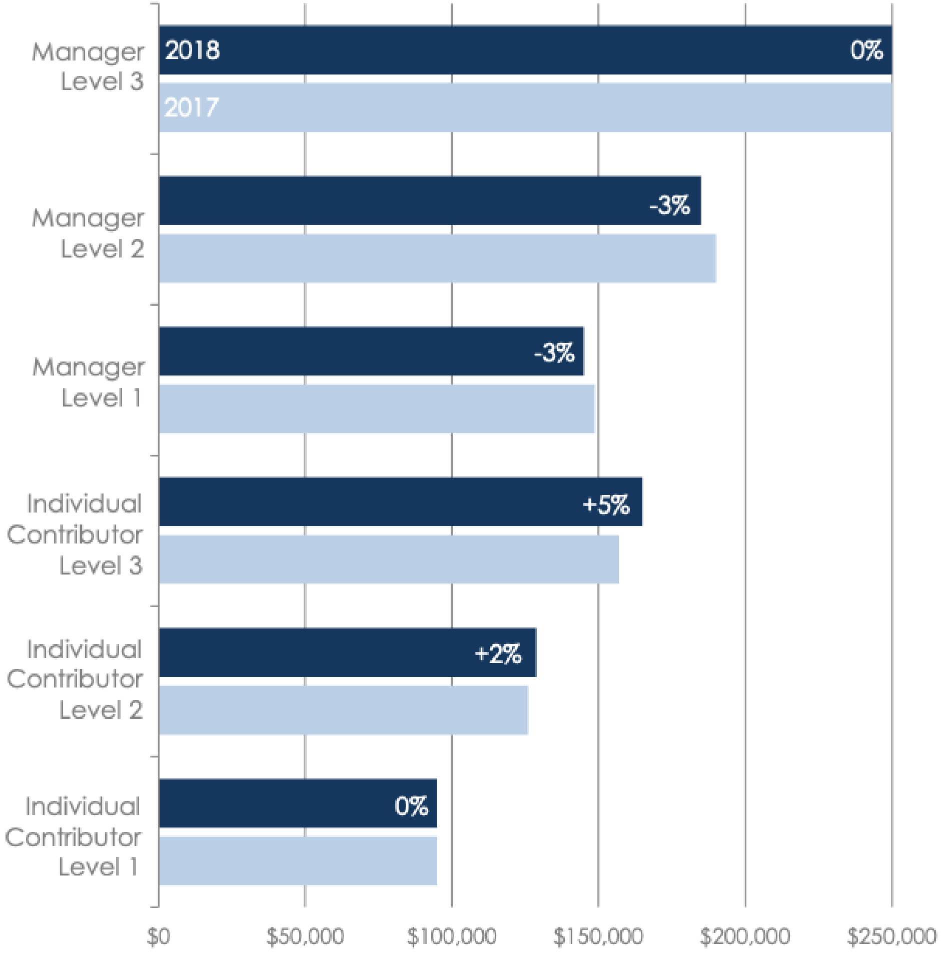
Salary, of course, depends on level of experience, job title, industry, company size, and geographic location.

https://datasciencedegree.wisconsin.edu/data-science/data-scientist-salary/

Median data scientist salary based on experience (2020)

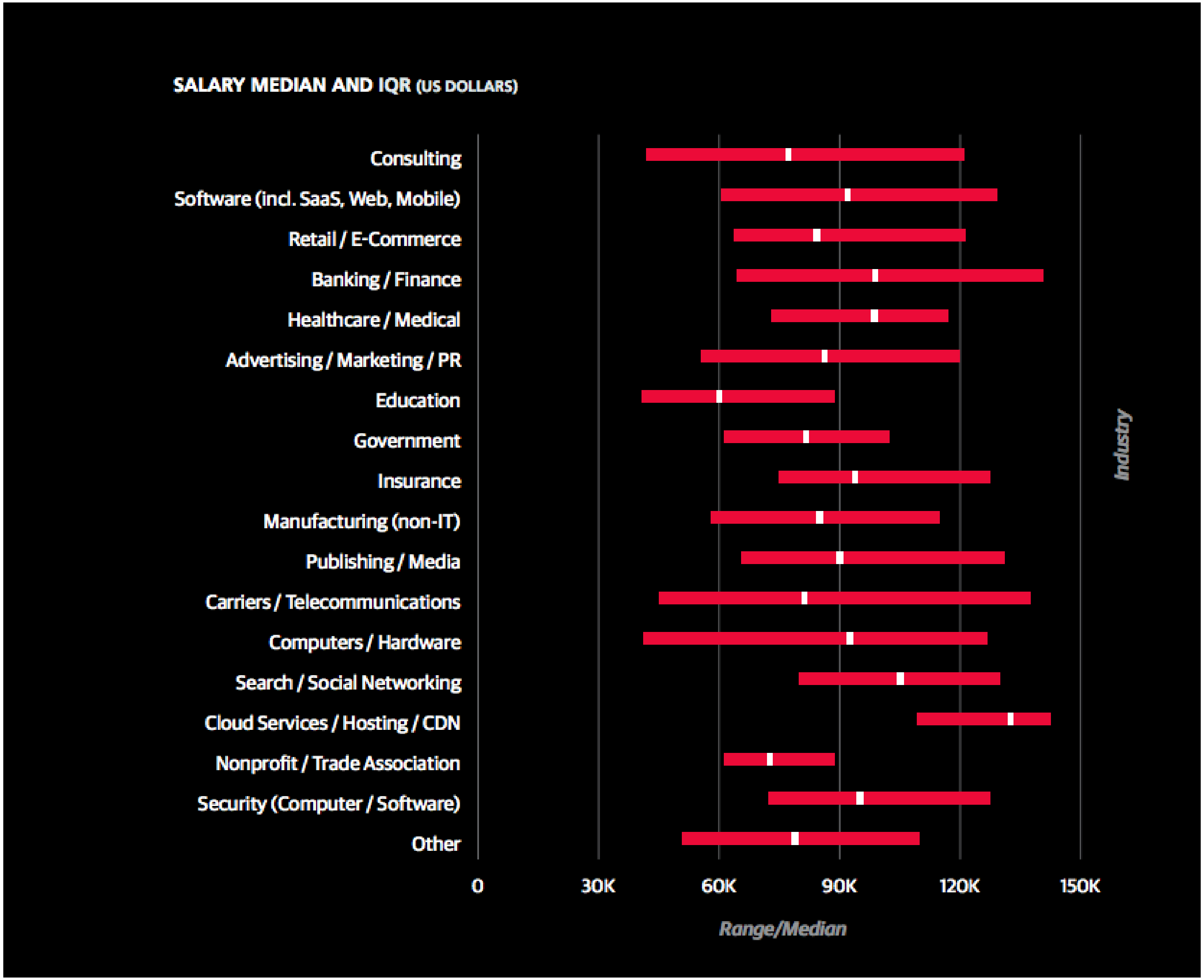
Entry-level	$95,000
Mid-level	$128,750
Mid-level (manager)	$185,000
Experienced	$165,000
Experienced (man.)	$250,000

Figure 1 Comparison of Data Scientists' Median Base Salaries by Job Category



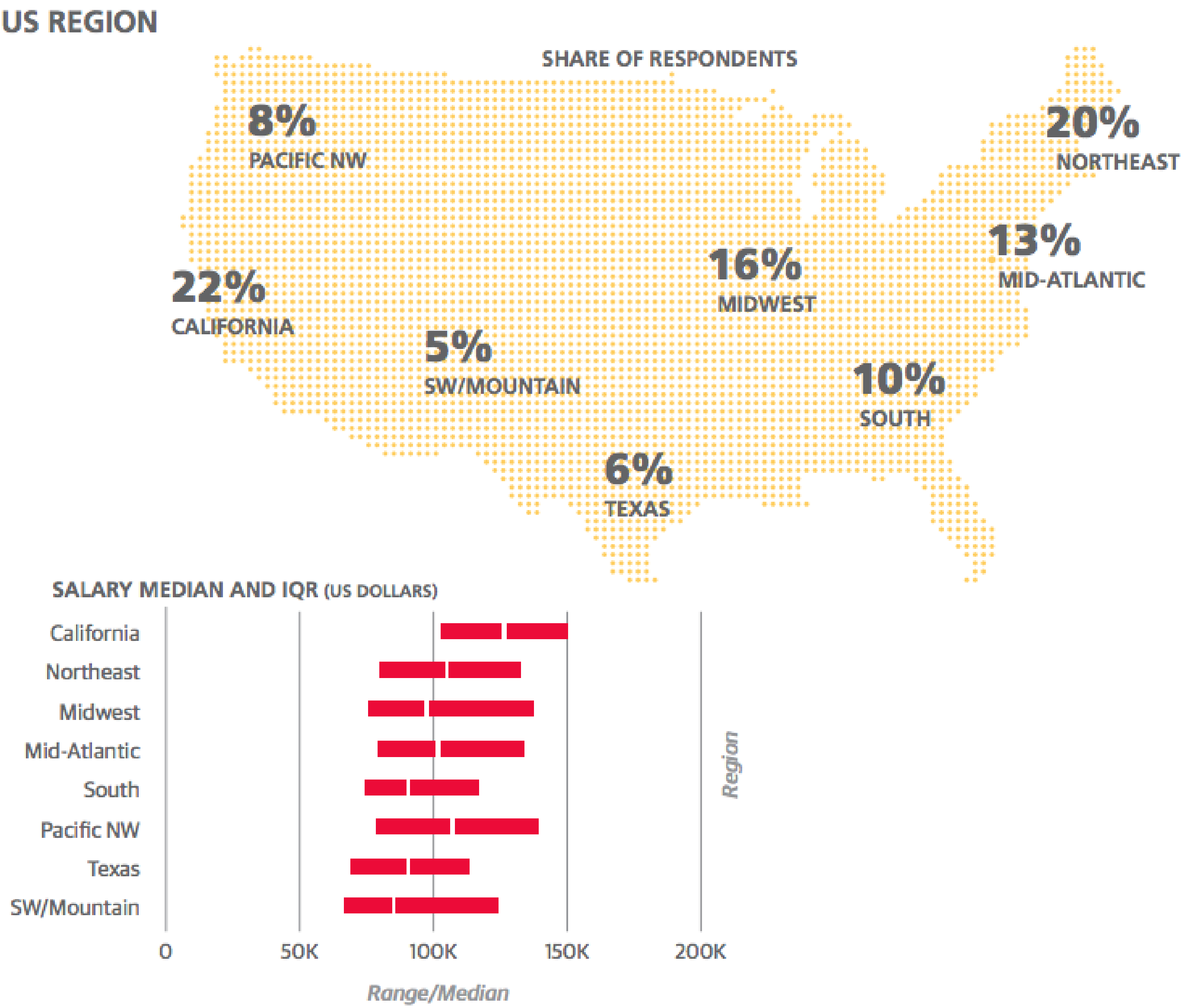
https://datasciencedegree.wisconsin.edu/data-science/data-scientist-salary/

Data scientist salary by industry (2016)



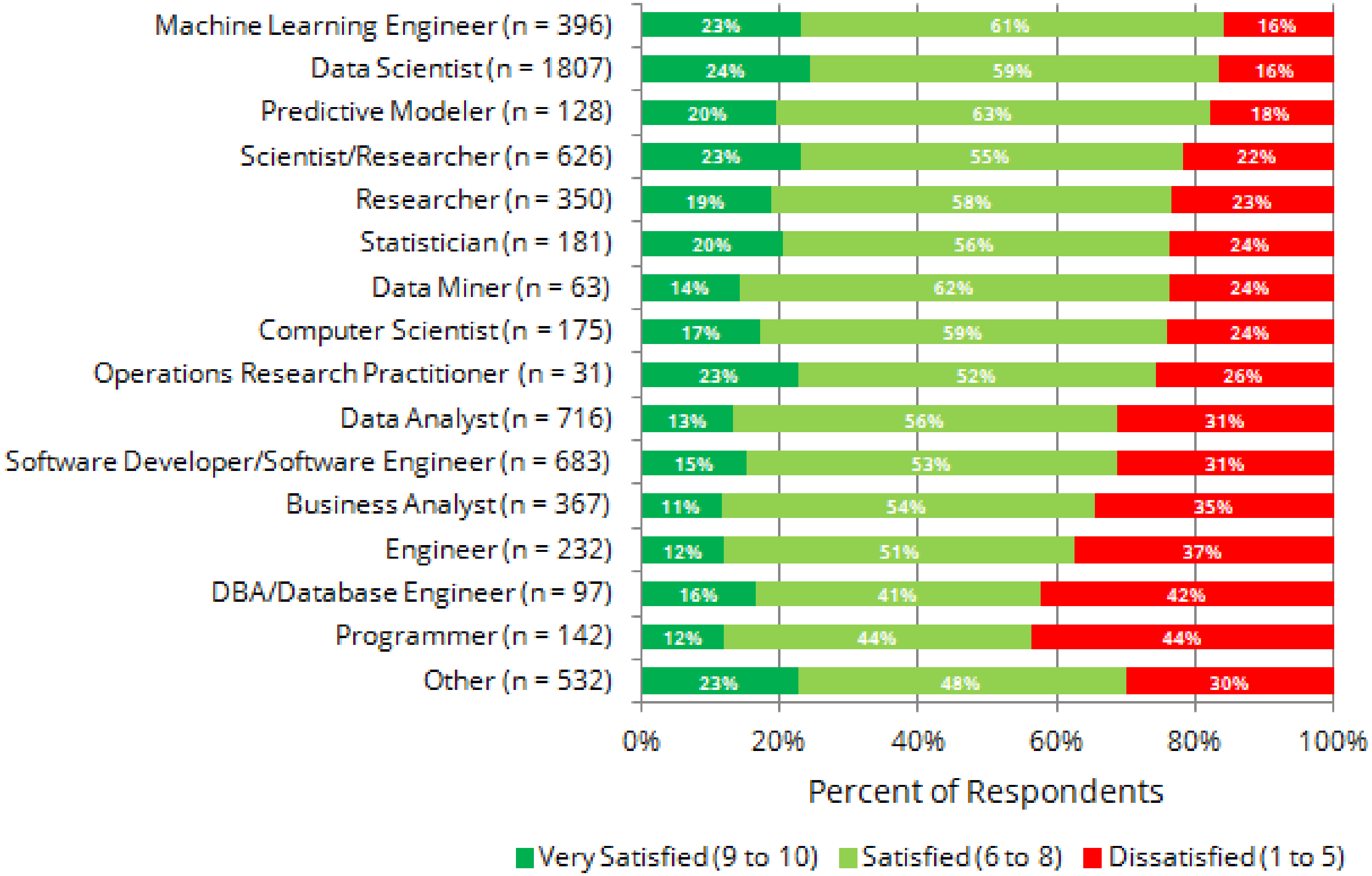
https://datasciencedegree.wisconsin.edu/data-science/data-scientist-salary/

Data scientist salary by region of United States (2016)



https://datasciencedegree.wisconsin.edu/data-science/data-scientist-salary/

High job satisfaction for data scientists (2017)



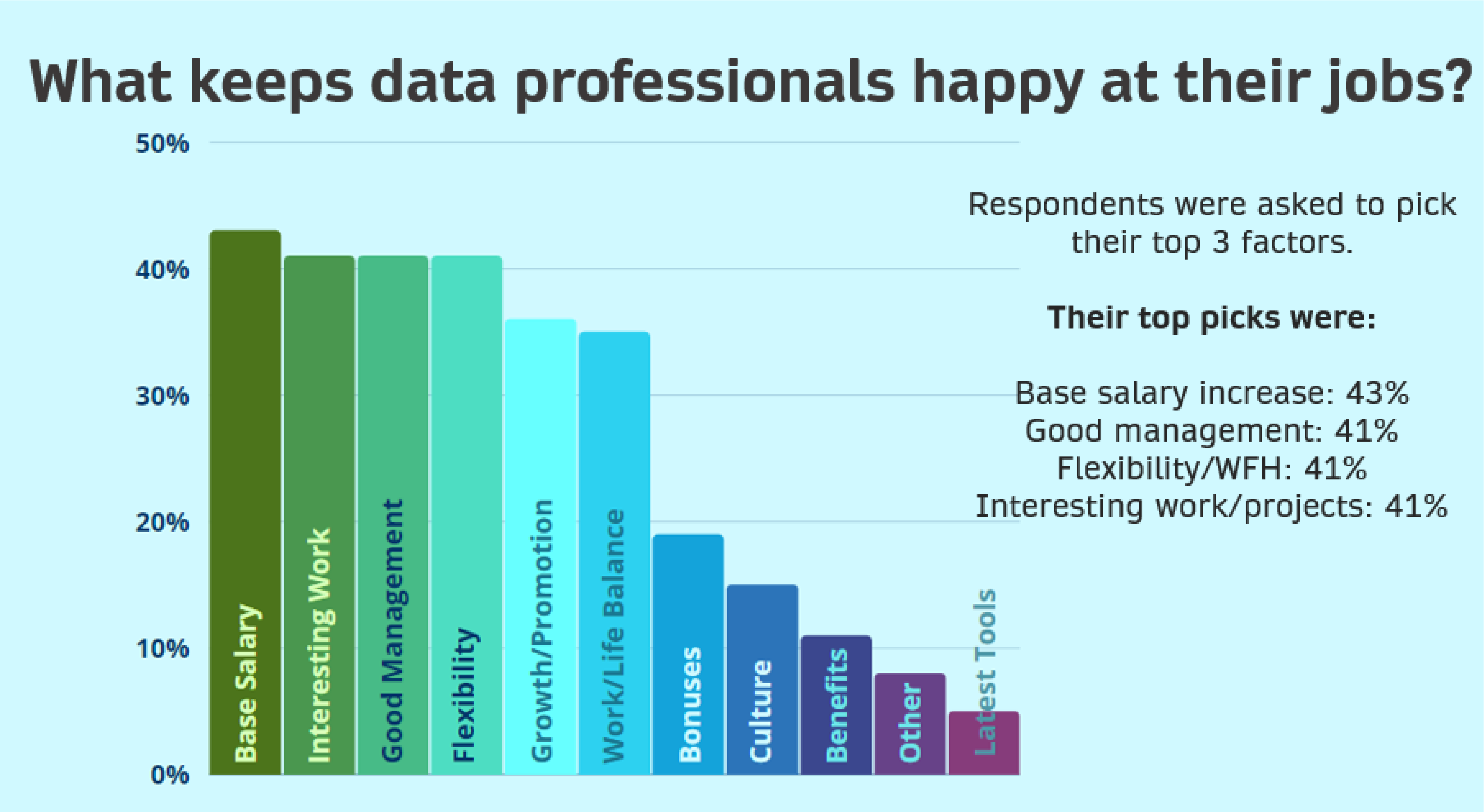
Data are from the Kaggle 2017 The State of Data Science and Machine Learning study. You can learn more about the study and download the data here: https://www.kaggle.com/surveys/2017.

Job Titles are ranked by percent of satisfied and very satisfied respondents.

Job satisfaction measured using question "On a scale from 1 (Highly Dissatisfied) - 10 (Highly Satisfied), how satisfied are you with your current job?" A total of 6529 respondents answered this question.

BUSINESS BROADWAY

Top pick for job satisfaction is interesting work (2021)



What keeps data professionals happy at their jobs?

Respondents were asked to pick their top 3 factors.

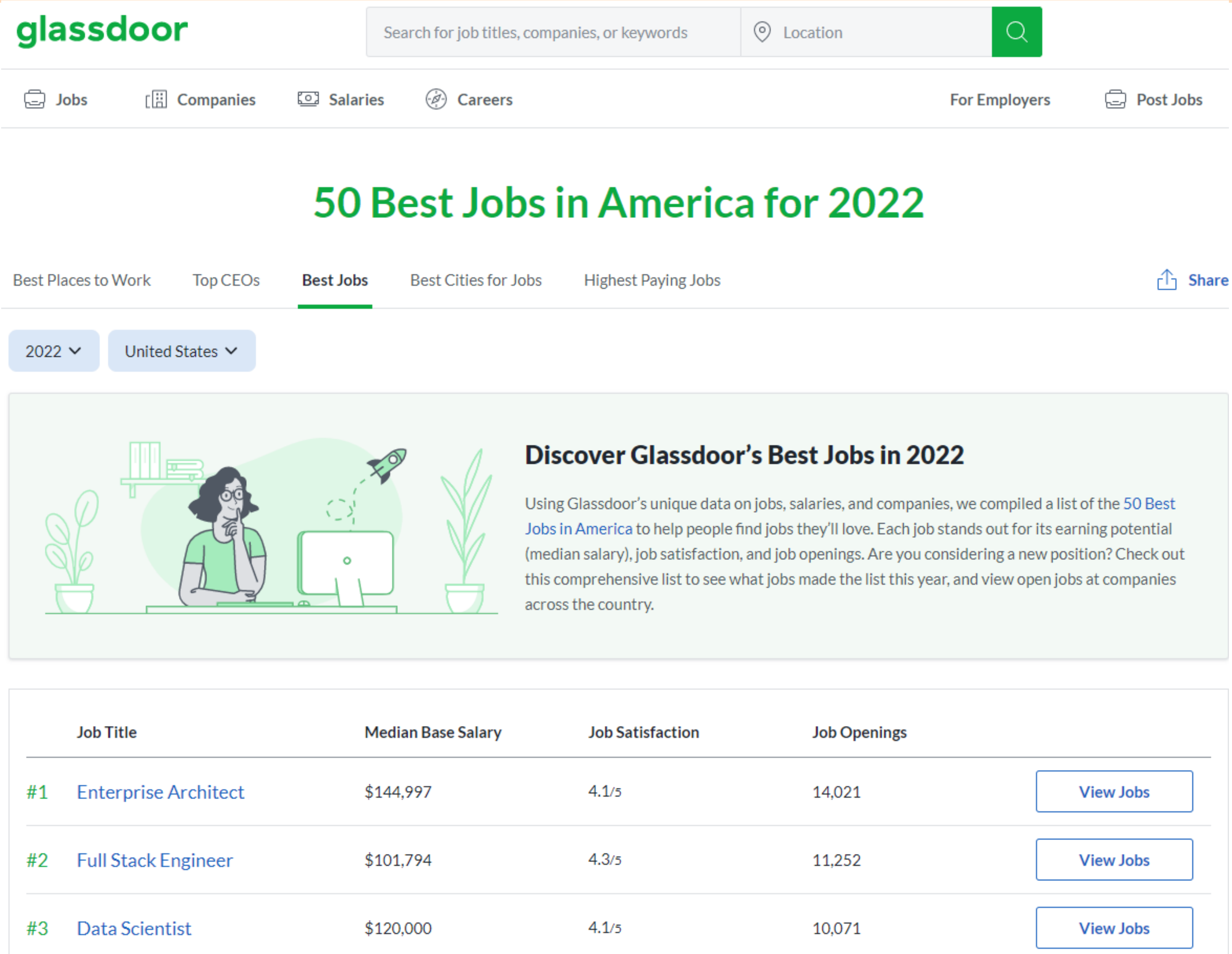
Their top picks were:

Base salary increase: 43%
Good management: 41%
Flexibility/WFH: 41%
Interesting work/projects: 41%

>350 respondents from data scientists, engineers, and analysts

https://www.burtchworks.com/2021/10/20/survey-results-what-keeps-data-scientists-data-engineers-happy-at-their-jobs/

Ranked #3 job in America by Glassdoor for 2022



50 Best Jobs in America for 2022

Discover Glassdoor's Best Jobs in 2022

Using Glassdoor's unique data on jobs, salaries, and companies, we compiled a list of the 50 Best Jobs in America to help people find jobs they'll love. Each job stands out for its earning potential (median salary), job satisfaction, and job openings. Are you considering a new position? Check out this comprehensive list to see what jobs made the list this year, and view open jobs at companies across the country.

	Job Title	Median Base Salary	Job Satisfaction	Job Openings	
#1	Enterprise Architect	$144,997	4.1/5	14,021	View Jobs
#2	Full Stack Engineer	$101,794	4.3/5	11,252	View Jobs
#3	Data Scientist	$120,000	4.1/5	10,071	View Jobs

https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

Most important reason to learn data science

It’s the future!

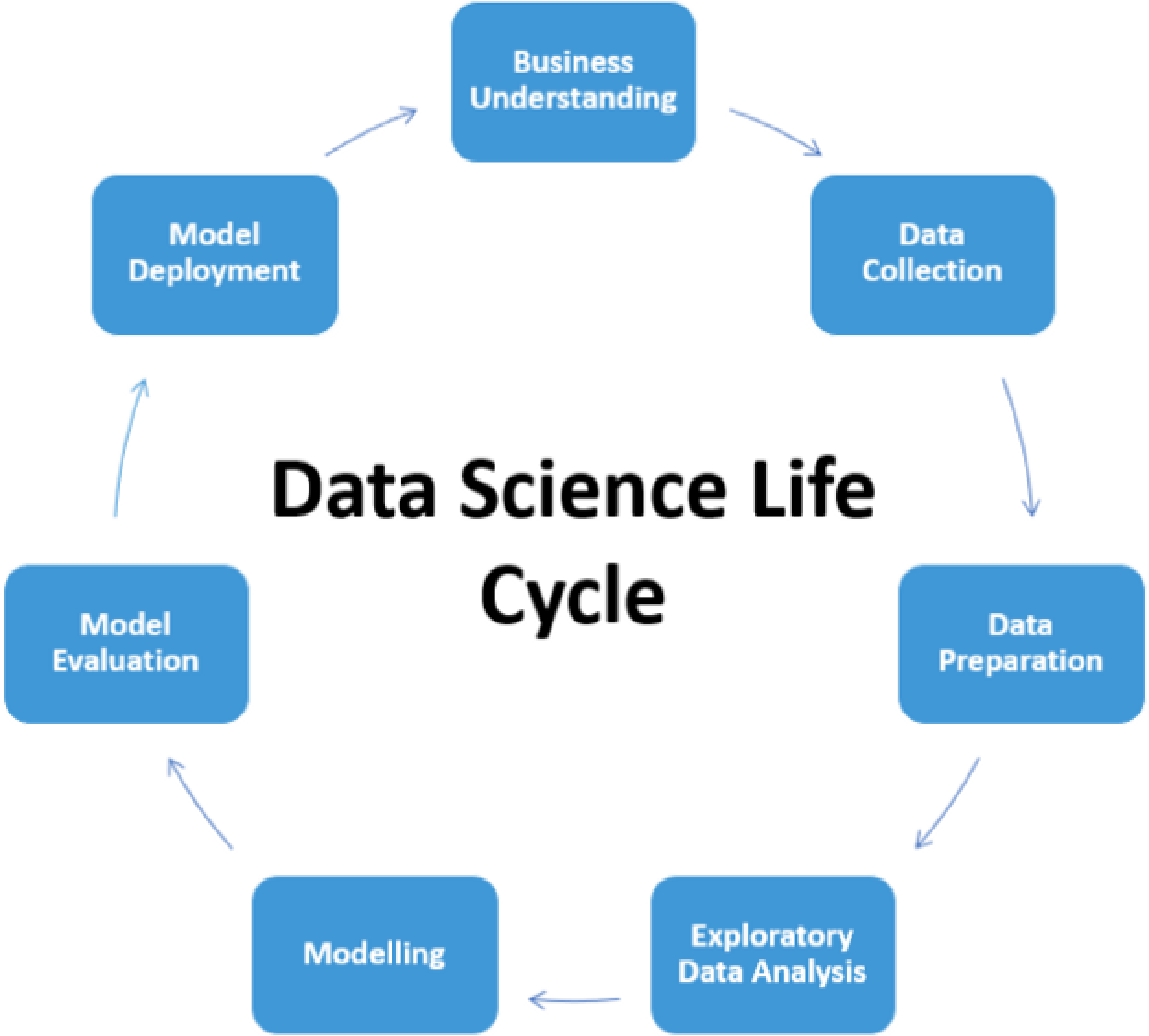
LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
TOM FISH BURNE
© marketoonist.com

DATA SCIENTIST
THE SEXIEST JOB OF THE 21ST CENTURY
15 MARCA 2013 WYDZIAŁ INFORMATYKI ZUT
Bigbit
BRAINS

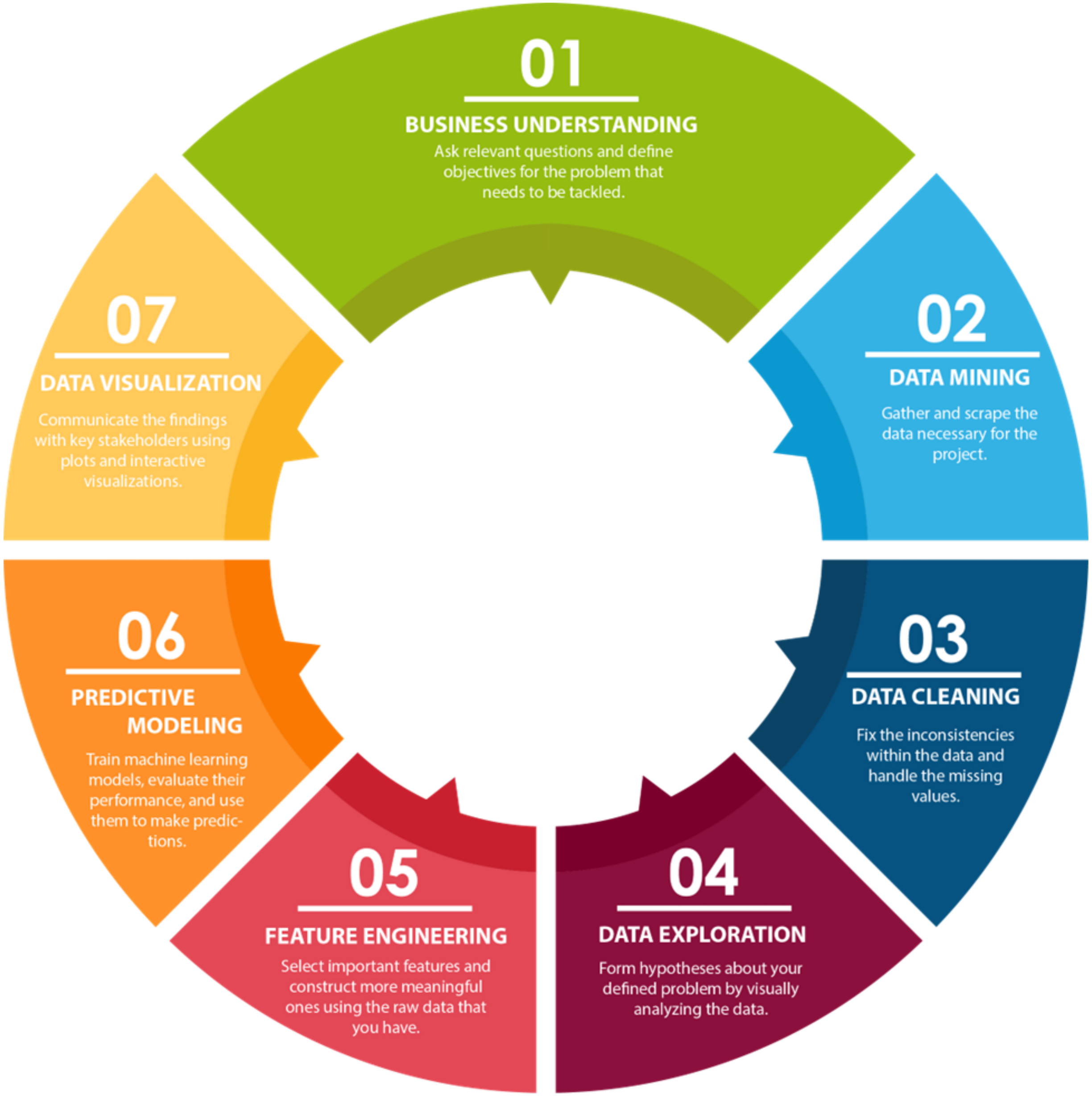
HBR.ORG
Harvard Business Review
OCTOBER 2012
46 The Big Idea
The True Measures Of Success
Michael J. Mauboussin
84 International Business
10 Rules for Managing Global Innovation
Keeley Wilson and Yves L. Doz
93 Leadership
What Ever Happened To Accountability?
Thomas E. Ricks
GETTING CONTROL OF
BIG DATA
How vast new streams of information are changing the art of management
PAGE 59

EXOPLANETS ✶ CITIZEN SCIENCE ✶ SUPERNOVAS
DARK MATTER ✶ GALAXIES ✶ DIAMOND WORLDS
SPECIAL ISSUE
AMERICAN Scientist
September–October 2018
www.americanscientist.org
BIG DATA
takes on
THE UNIVERSE
SIGMA XI

Data Science Lifecycle



Data Science Components

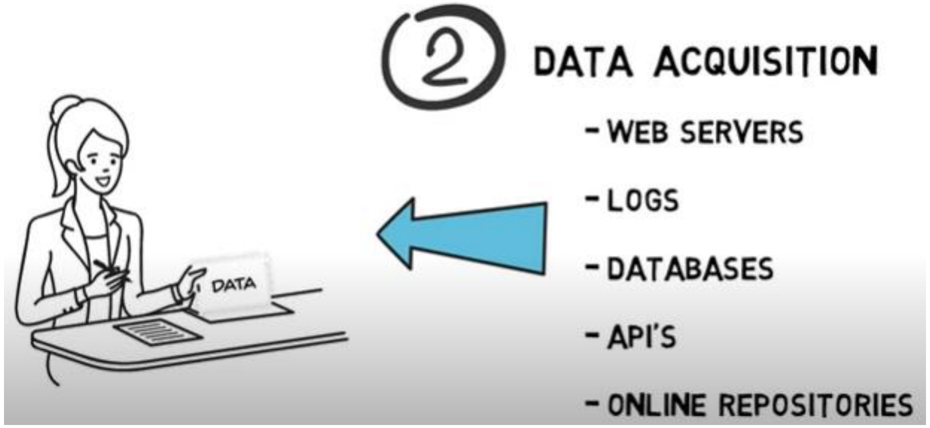

01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.

Important components of data science (Business Problem)

This may be the most important and uncertain step.

- What are the goals of the model?
- What's in the scope and outside the scope of the model?
- Asking the right question will determine what data to collect later.
- This also determines if the cost to collect the data can be justified by the impact of the model.
- Also, what are the risk factors known at the beginning of the process?

Important components of data science (Collection)

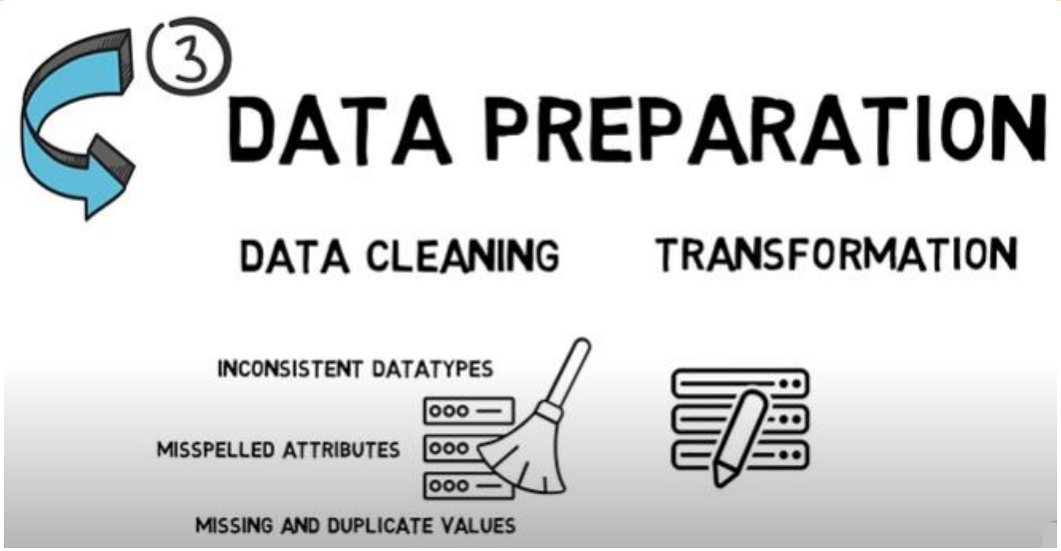


Loading, Import, Store

Given collected data, to begin any analysis, this data must first be **imported** into an analysis environment (in this class we will use **Python**).

That is, we will need to retrieve data from a stored file, database, or the web, and load it into a program for downstream analysis.

Important components of data science (Wrangling)

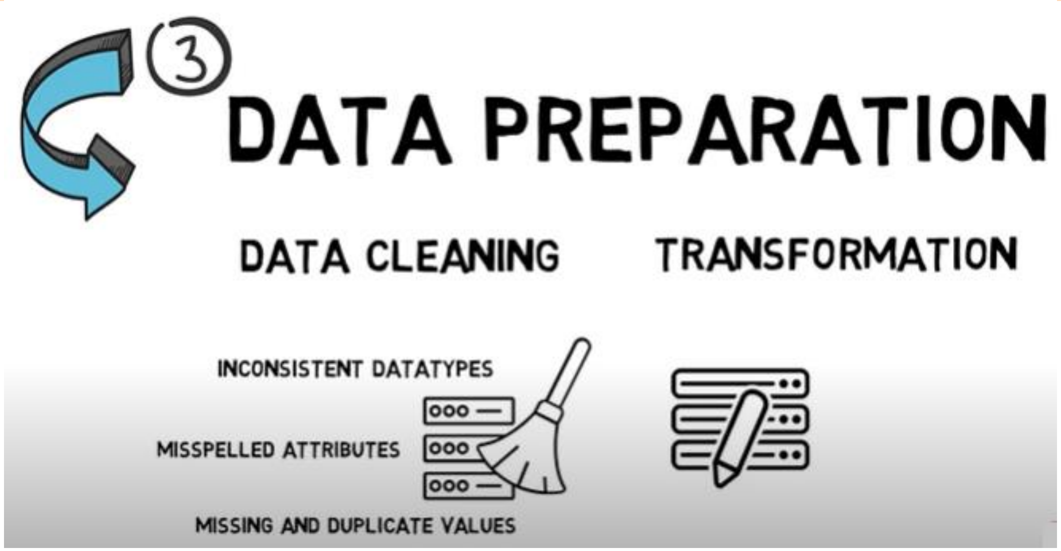


Clean, Transform, Merge, Reshape

Given the imported data, it is important to preprocess it it, and storing it in a format that is easy to parse and analyze.

That is, placing the data in a format that has each observation on a separate row and each variable (feature) in its own column.

Important components of data science (Transformation)

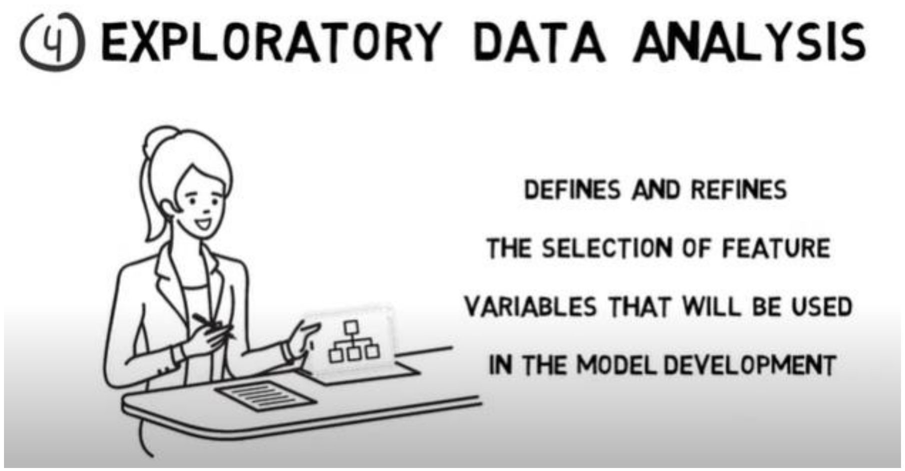


Replacing, Encoding, Standardization, Renaming

The transformed data may be easier to analyze if it is summarized (or **transformed**) as new variables.

For example, if you have observations on distance and time, then one may be able to transform it to a new variable (such as speed).

Important components of data science (EDA)



4 EXPLORATORY DATA ANALYSIS

DEFINES AND REFINES THE SELECTION OF FEATURE VARIABLES THAT WILL BE USED IN THE MODEL DEVELOPMENT

Exploratory Data Analysis refers to the critical process of performing initial investigations on data to:

- Discover patterns
- Spot anomalies
- Test hypothesis
- Check assumptions with the help of summary statistics and graphical representations.
- Huge portion of that is Visualization

Important components of data science (visualize)

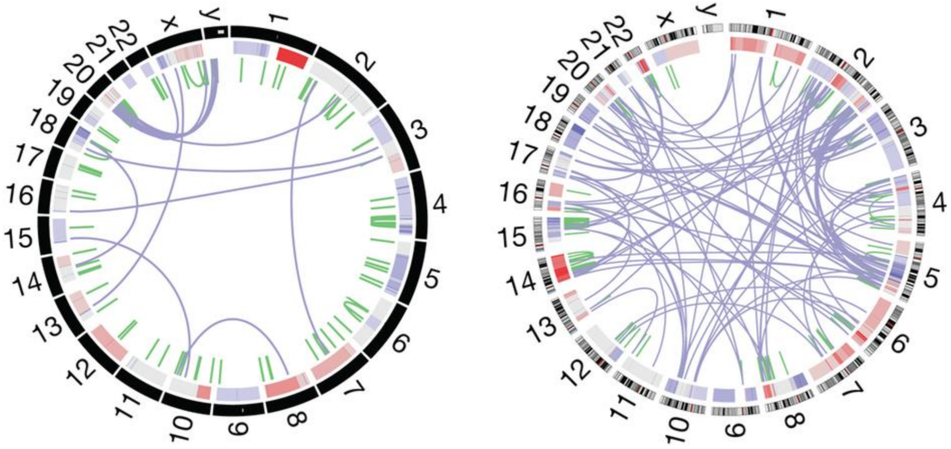
Visualize

Data **visualization** is extremely important for two reasons:

1) The human eye is adept at pattern recognition, making it faster and easier for us to spot trends in visual displays.
2) There are fewer barriers (language, grammar, etc.), enabling communication to a wider audience.
3) Data Visualization is important in most of the data science project milestones/steps.

Important components of data science (visualize)

“A picture is worth a thousand words”



Important components of data science (visualize)

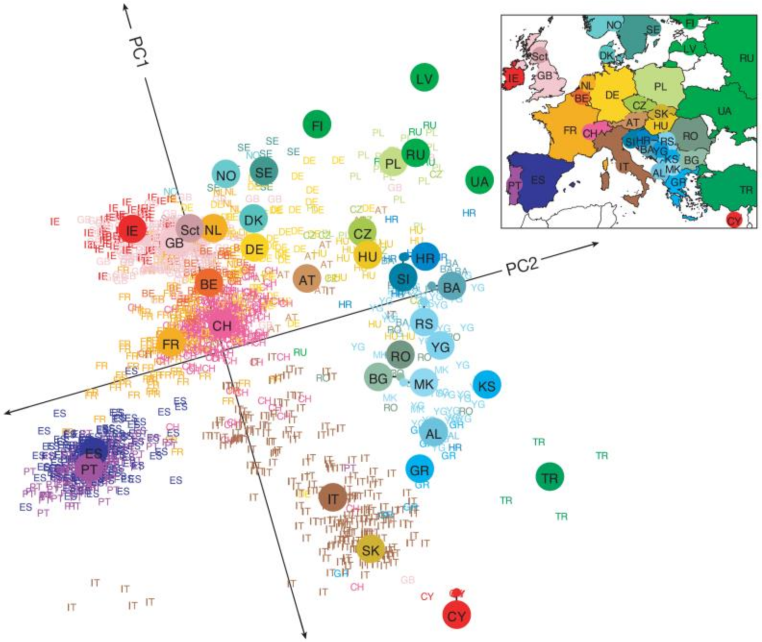
Data **visualization** is crucial in data science because it can help us:

1) Identify patterns and issues in data
2) Generate hypotheses or questions
3) Communicate findings to both experts and non-experts

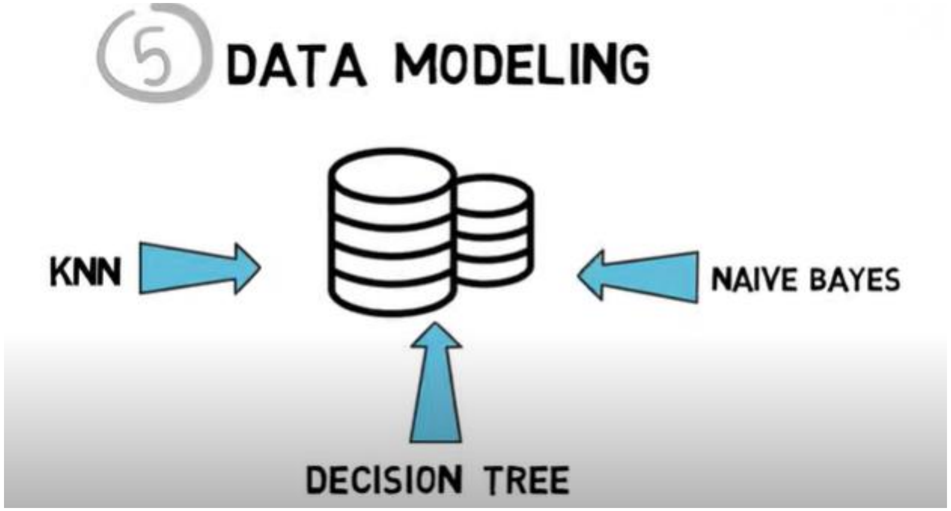
Important components of data science (visualize)

The graph on the right displays approximately 5,000 individuals (points) sampled across Europe, which would be a mess without color.

However, with color (based on country), we can see that individuals (points) sampled from the same country (color) cluster together, as they are genetically related.



Important components of data science (model)



Data **modeling** is complementary to visualization.

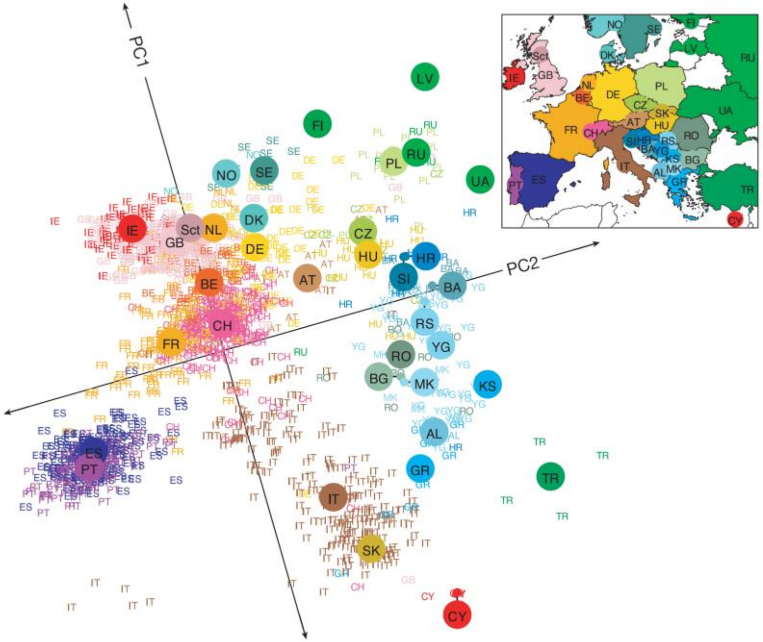
Visualization can aid in formulating precise questions, which can be answered through modeling.

Important components of data science (model)

Consider again the biology example.

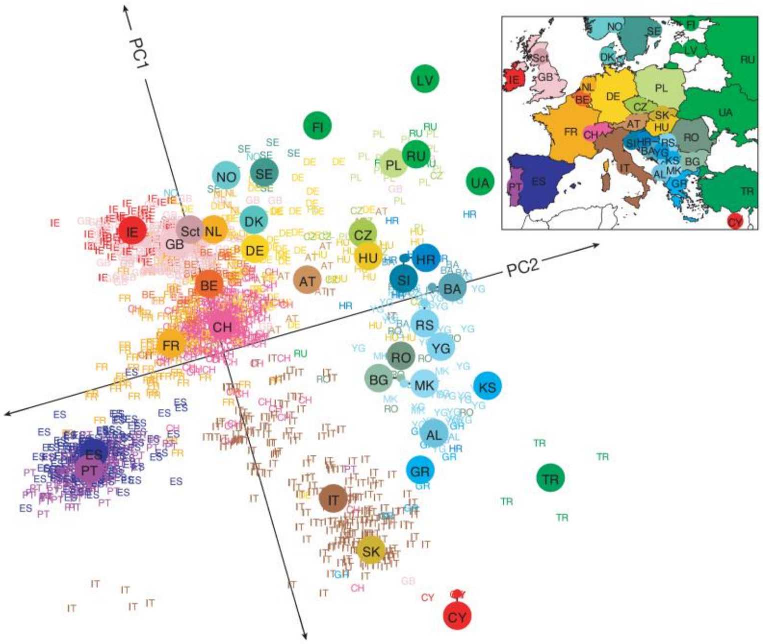
Hypothesis: From genetics, we can separate individuals by ancestry.

A natural question is then, given an individual’s genotypes at many genomic locations, what is their ancestry?



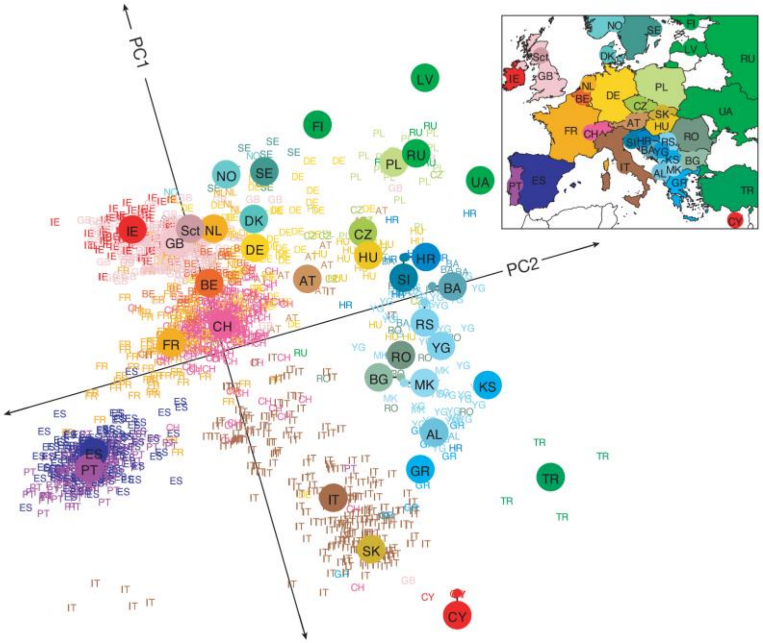
Important components of data science (model)

Companies like 23andMe and Ancestry.com utilize genetic information at close to a million genomic locations to accurately predict participant ancestry.

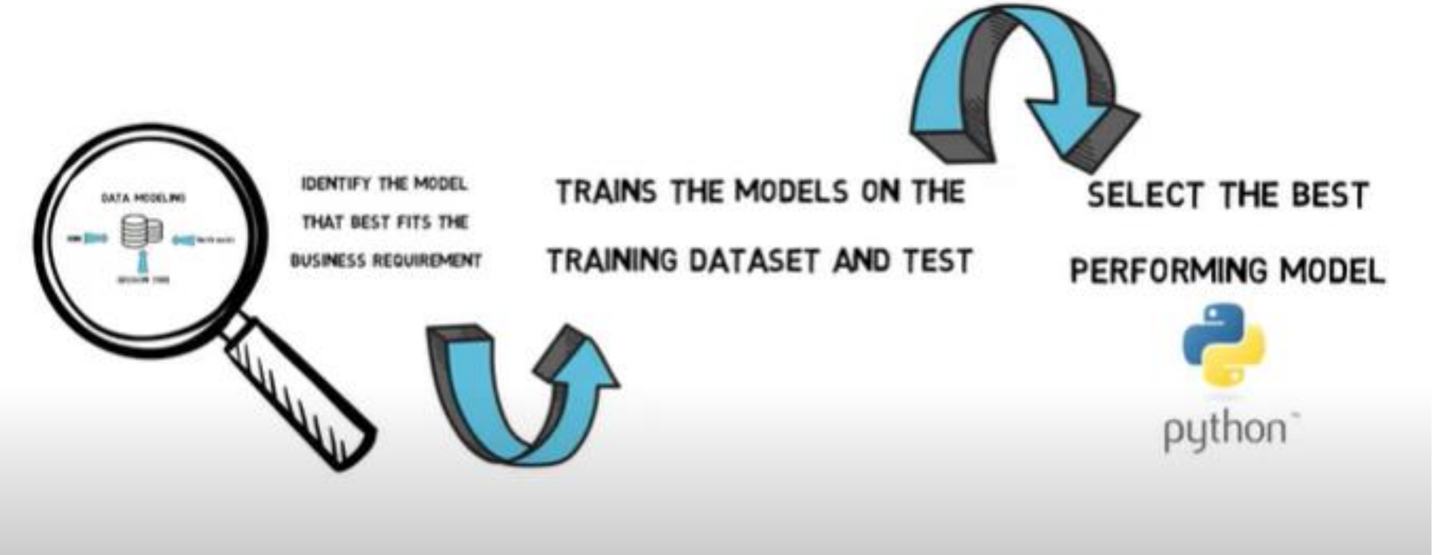


Important components of data science (model)

They do this utilizing a model, which they employ to compute the probability that the observed genotypes of an individual derive from a particular ancestry group based on reference individuals.

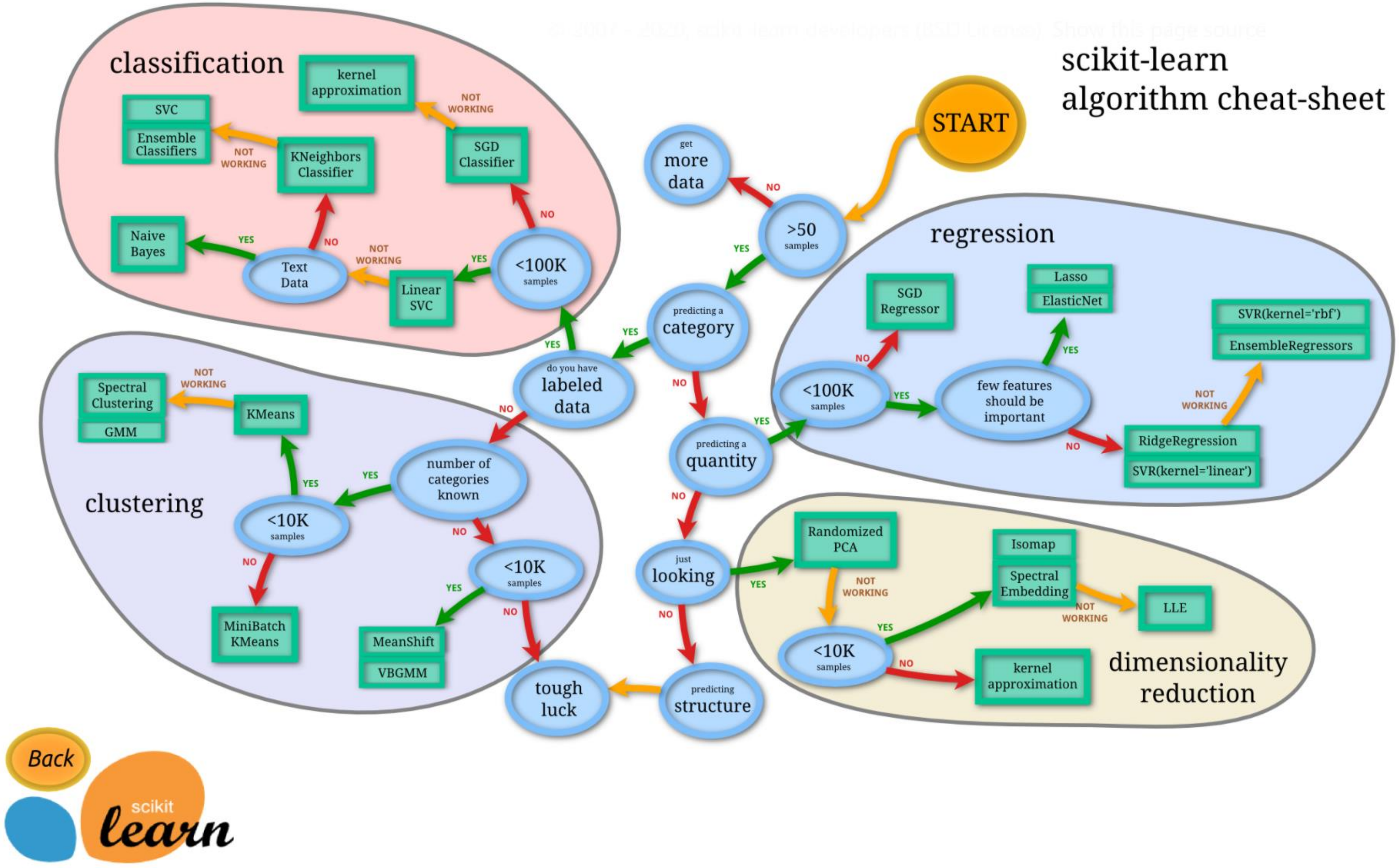


Important components of data science (model)


DATA MODELING
IDENTIFY THE MODEL THAT BEST FITS THE BUSINESS REQUIREMENT
TRAINS THE MODELS ON THE TRAINING DATASET AND TEST
SELECT THE BEST PERFORMING MODEL
python™

Important components of data science (model)

Choosing the Right Estimator



Important components of data science (communicate)

Communicate

Finally, the ability to **communicate** findings of your analysis is absolutely crucial to being a successful data scientist, regardless of how good your are a data visualization or modeling.

Important components of data science (understand)

Transform

Visualize

Model

Understanding = “statistical learning”

Terminology

Notation

Input X: X is often multidimensional. Each dimension of X is referred to as a feature, or independent variable

Output Y: The response, or dependent variable

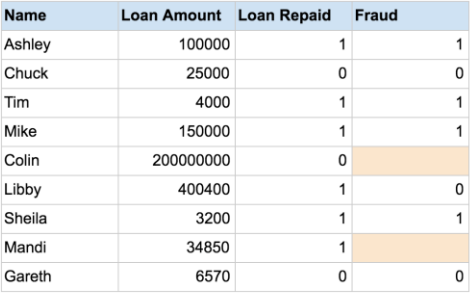
Categorization

- Supervised learning vs. unsupervised learning
  - Is Y available in the training data?
- Regression vs. classification
  - Is Y quantitative or qualitative?

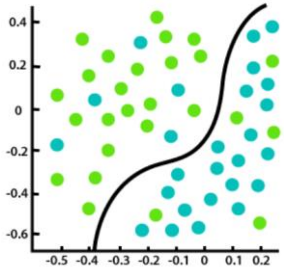
Terminology

What is Supervised Learning?

In supervised learning, the computer is taught by example. It learns from past data and applies the learning to present data to predict future events. In this case, both input and desired output data provide help to the prediction of future events.

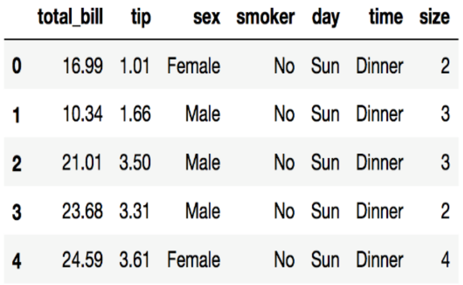


Name	Loan Amount	Loan Repaid	Fraud
Ashley	100000	1	1
Chuck	25000	0	0
Tim	4000	1	1
Mike	150000	1	1
Colin	200000000	0	
Libby	400400	1	0
Sheila	3200	1	1
Mandi	34850	1	
Gareth	6570	0	0

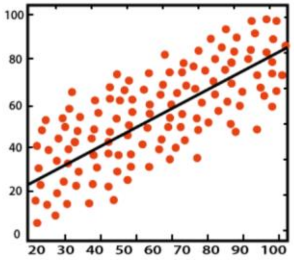


Classification

<u>1) Classification </u> - is used for problems where the output variable can be categorized, such as “Yes” or “No”, or “Pass” or “Fail.” Classification Models are used to predict the category of the data.



	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4



Regression

<u>2) Regression </u> – is used for problems where the output variable is a real value such as a unique number, dollars, salary, weight or pressure, for example. It is most often used to predict numerical values based on previous data observations. Some of the more familiar regression algorithms include linear regression.

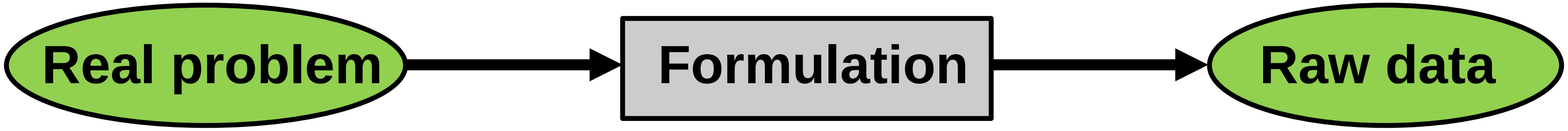
Statistical learning is simply function (f) estimation

Real problem

Statistical learning is simply function (f) estimation

Real problem
Formulation

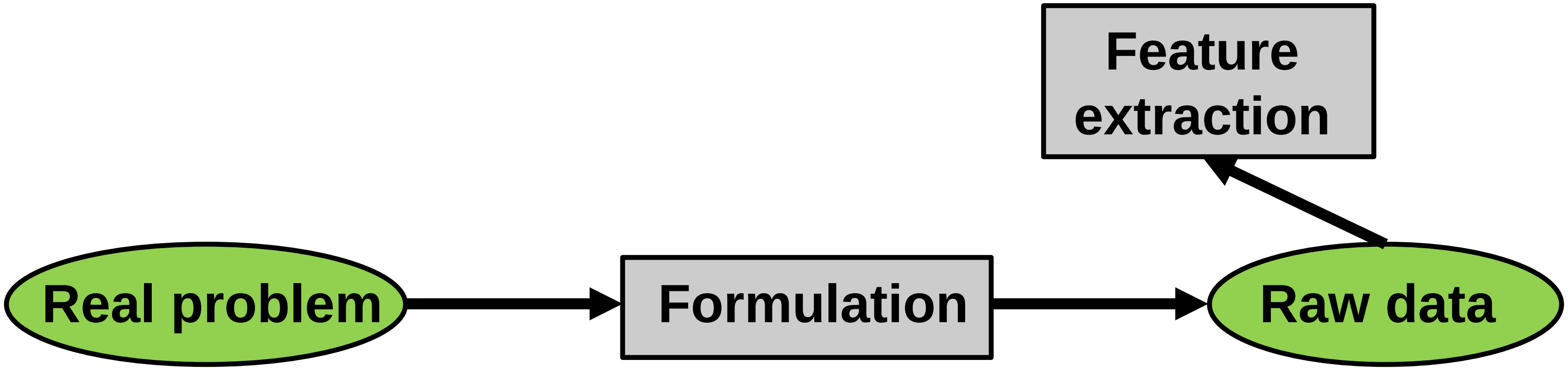
Statistical learning is simply function (f) estimation


Real problem

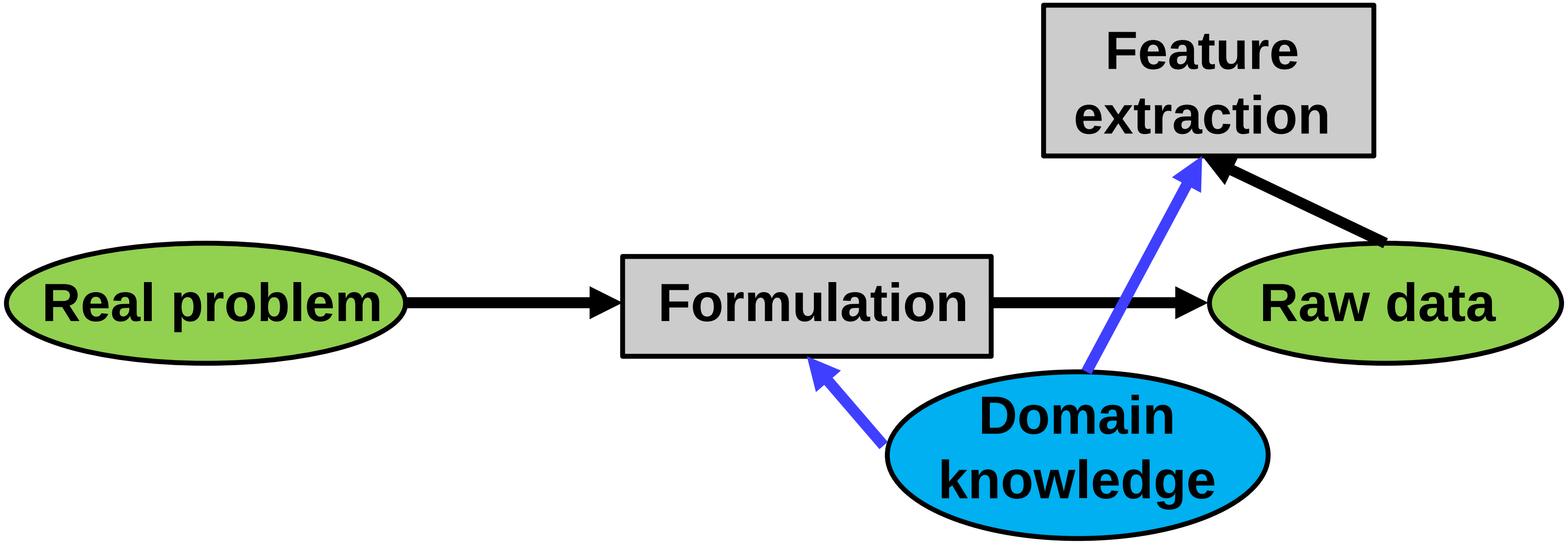
Formulation

Raw data

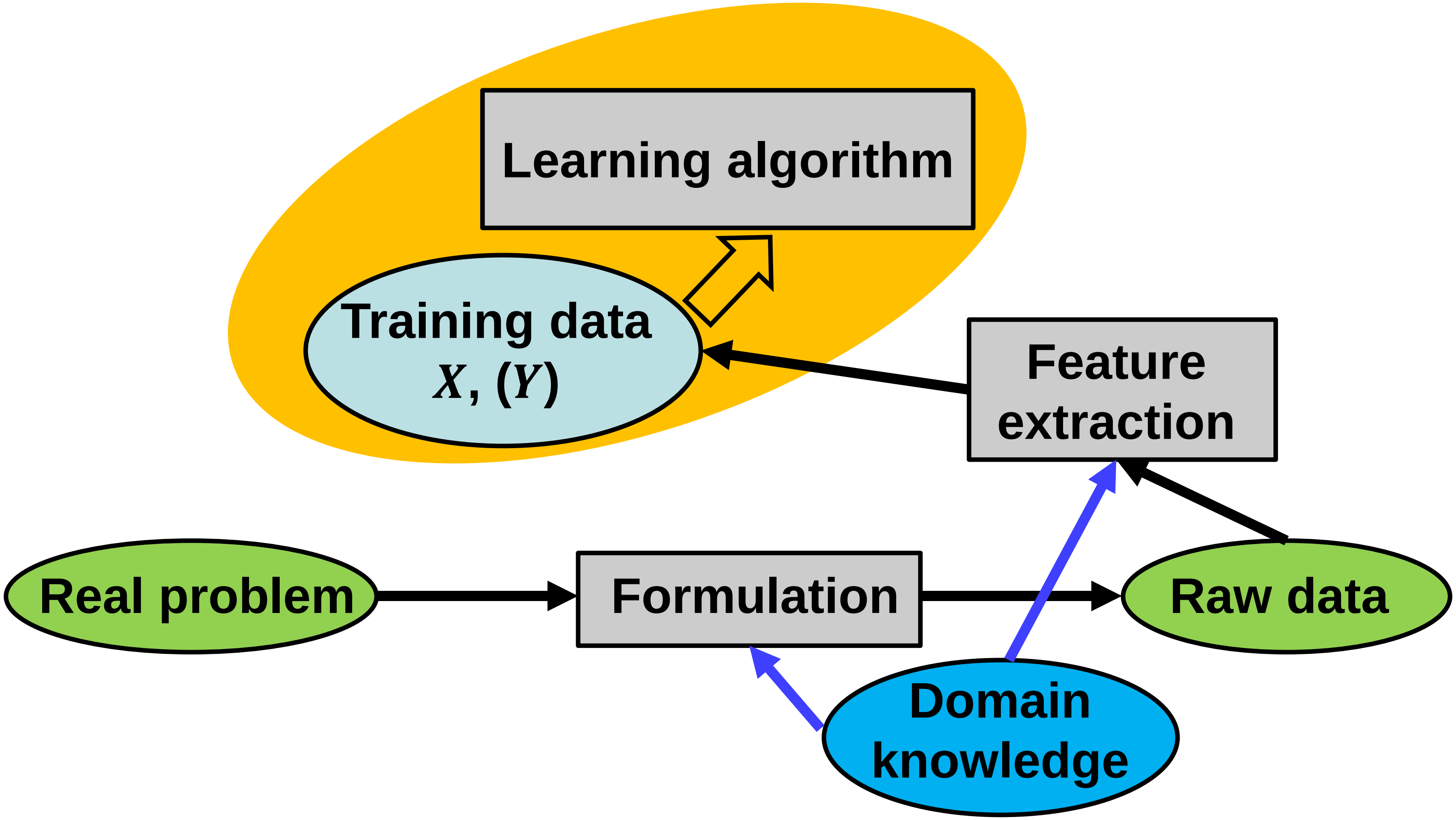
Statistical learning is simply function (f) estimation



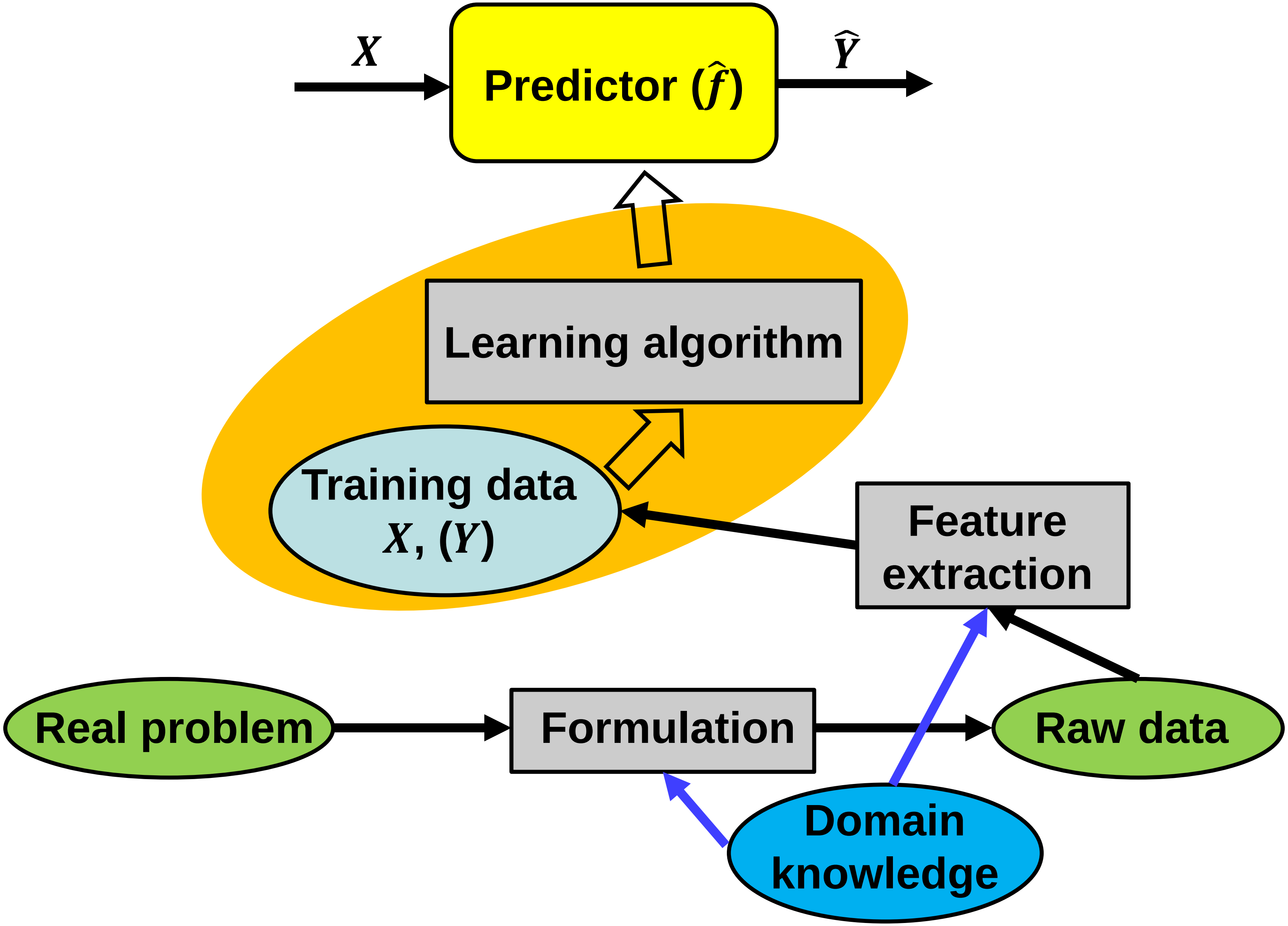
Statistical learning is simply function (f) estimation



Statistical learning is simply function (f) estimation



Statistical learning is simply function (f) estimation



Supervised learning (regression)

Example: wage prediction

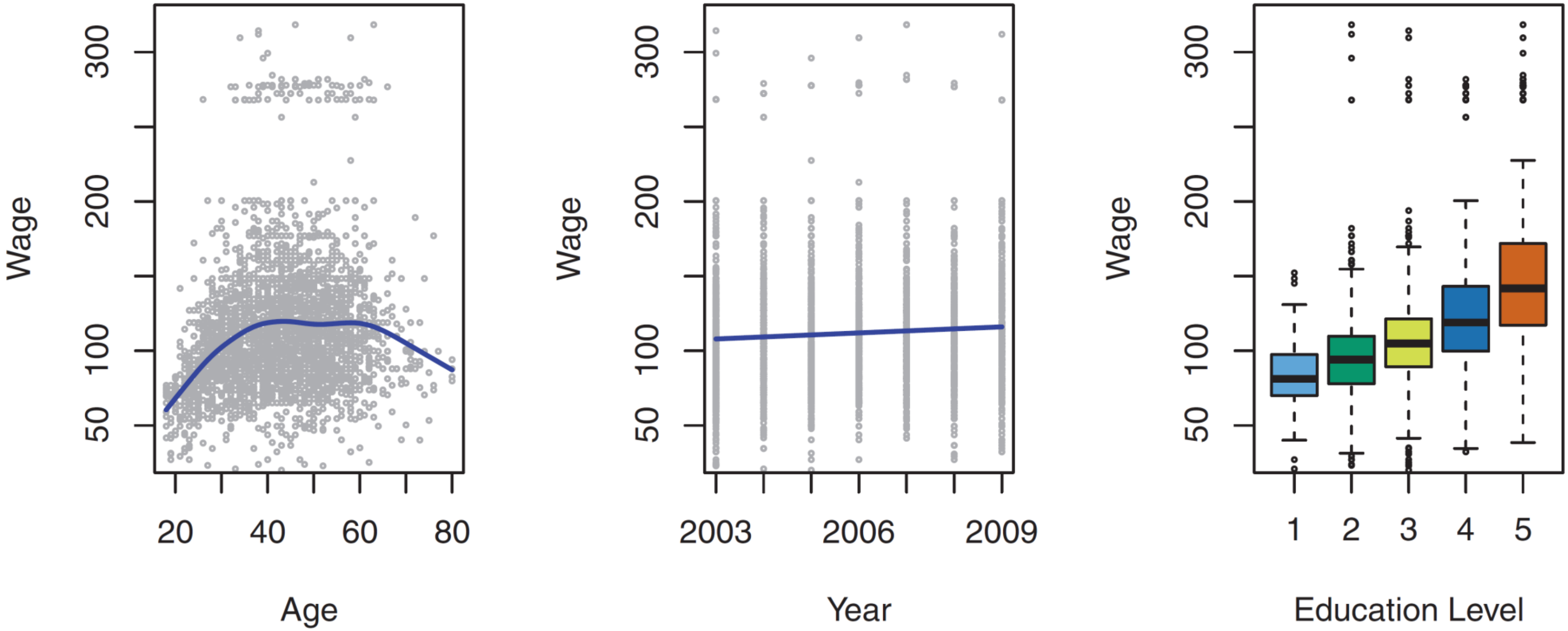
Goal: predict an individual’s wage, based on certain features, such as age, year, and education level.

Raw data: wage together with age of individual, calendar year, and categorical value representing the education level of the individual at the time the wage value was collected.

Input data: a three-dimensional vector.

Supervised learning (regression)

Predict an individual's wage, based on certain features, such as age, year, and education level.



Supervised learning (classification)

Example: Handwritten digit recognition

Goal: identify the digits (0, 1, …, 9) associated with handwritten numbers.

Raw data: images that are scaled segments from five-digit ZIP codes.

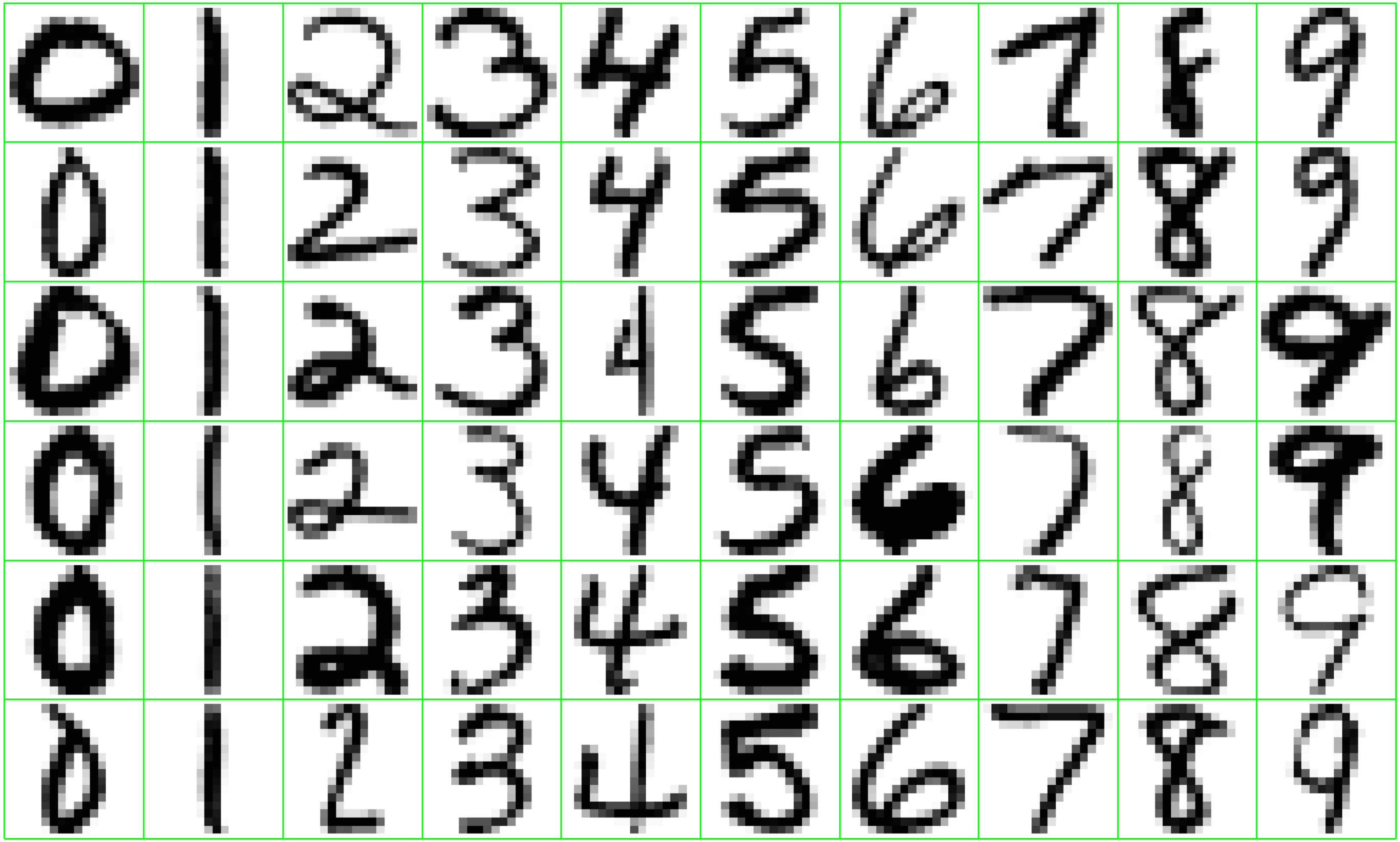
16 × 16 eight-bit grayscale maps

Pixel intensities range from 0 (black) to 1 (white)

Input data: a 256-dimensional vector, or feature vectors with lower dimensions.

Supervised learning (classification)

Identify the digits (0, 1, …, 9) associated with handwritten numbers

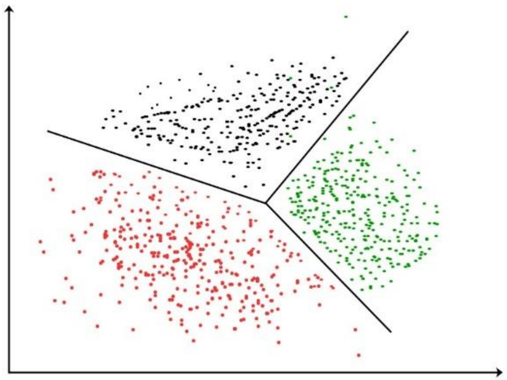


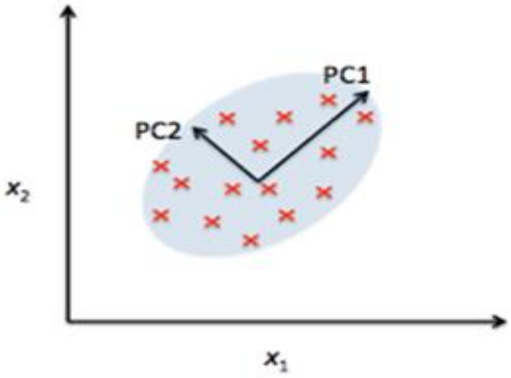
Supervised learning (classification)

What is Unsupervised Learning?

Unsupervised learning, on the other hand, is the method that trains machines to use data that is neither classified nor labeled. It means no training data can be provided and the machine is made to learn by itself. The machine must be able to classify the data without any prior information about the data.

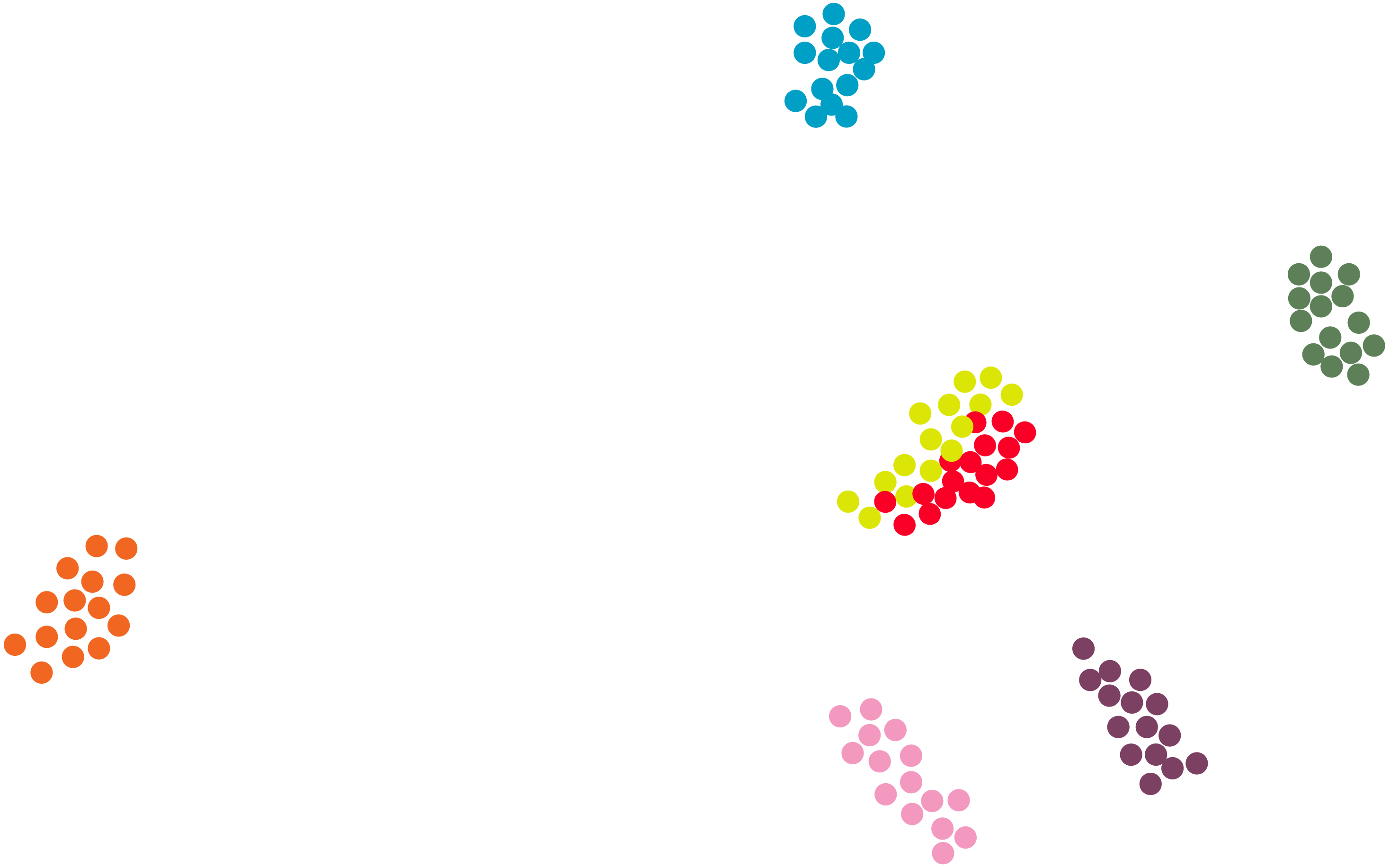
These models include tasks such as *clustering, association,* and *dimensionality reduction.*





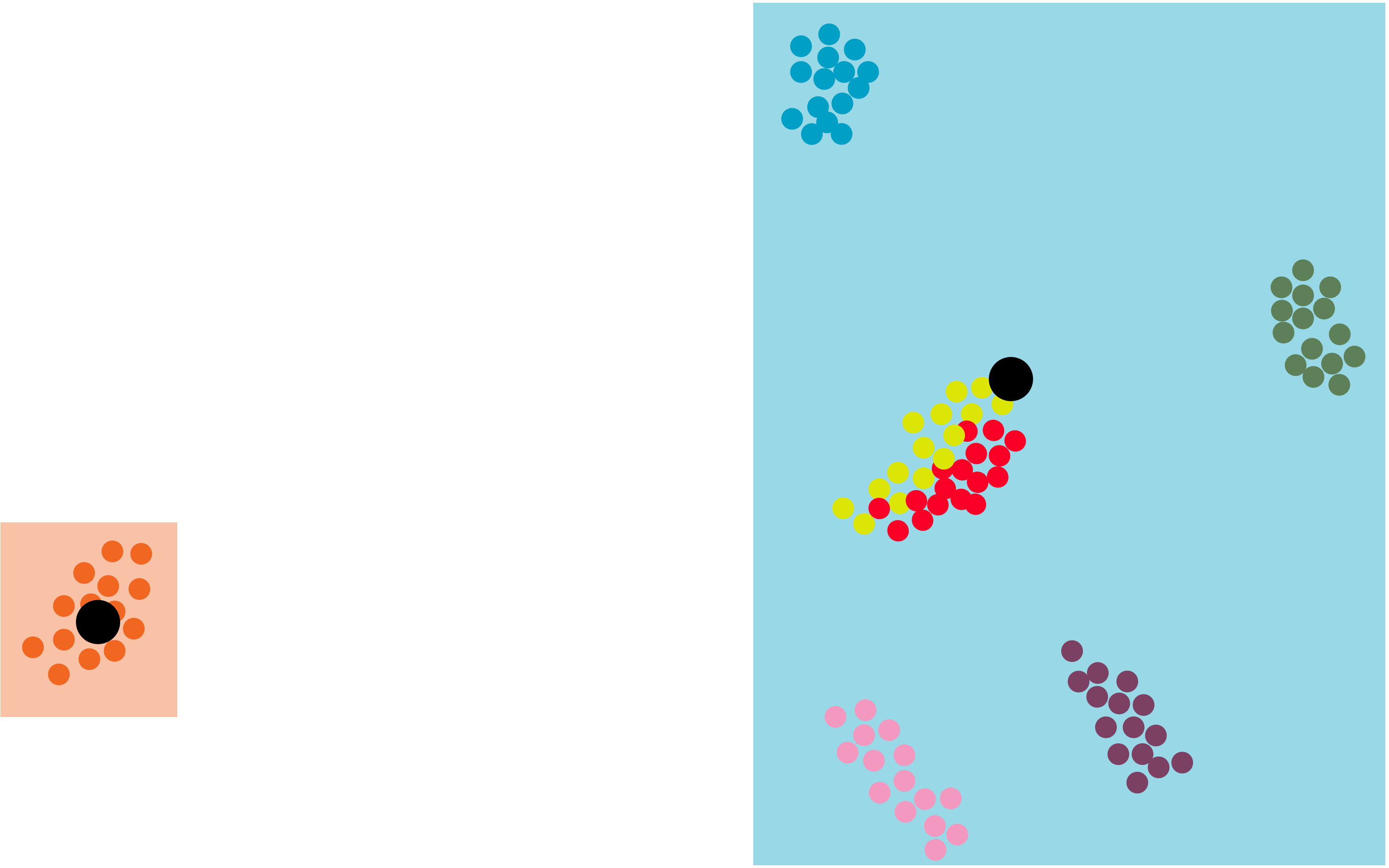
Unsupervised learning (prototype clustering)

Attempt to assign observations to a discrete number of *K* clusters



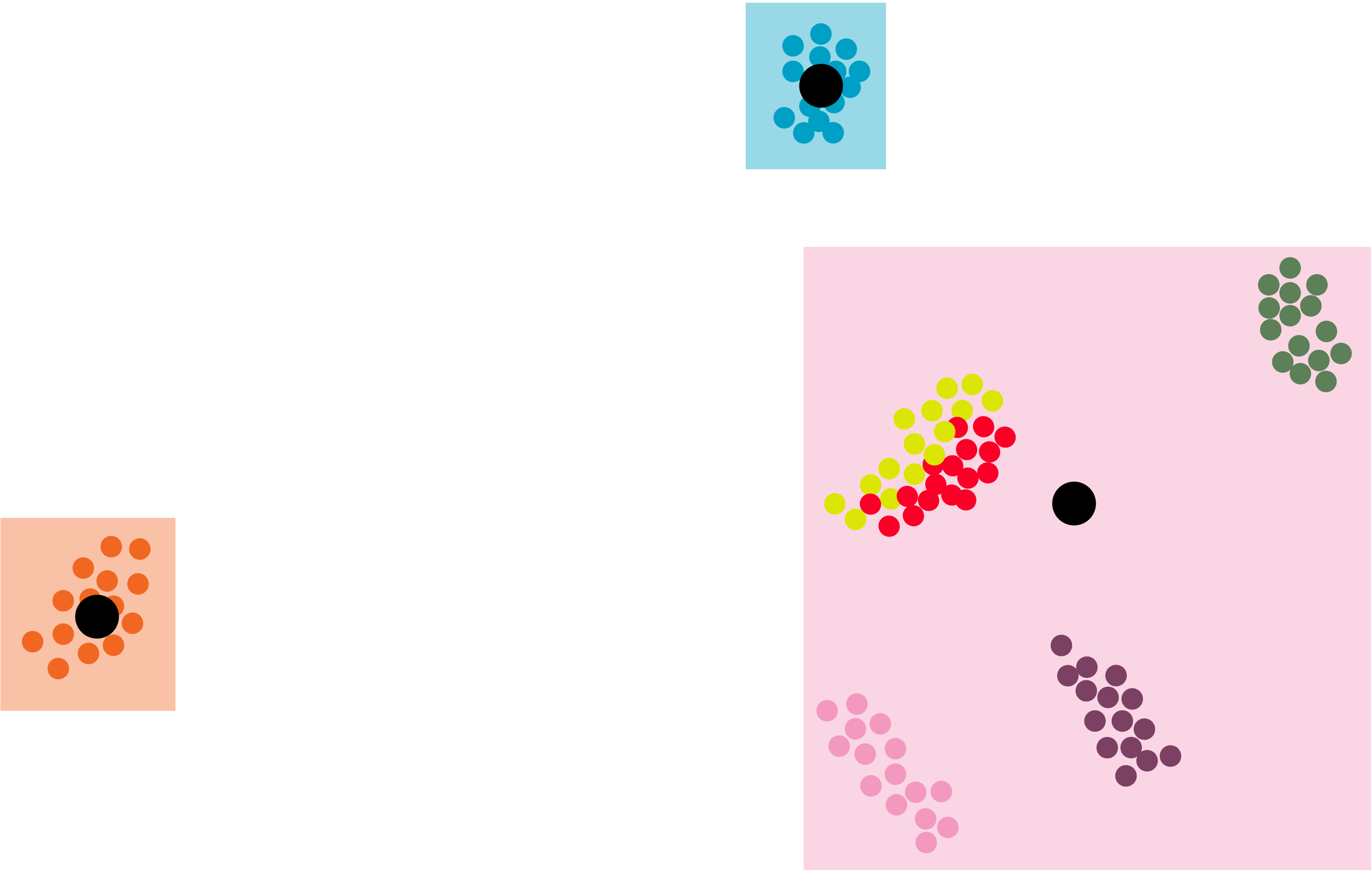
Unsupervised learning (prototype clustering)

K = 2 clusters



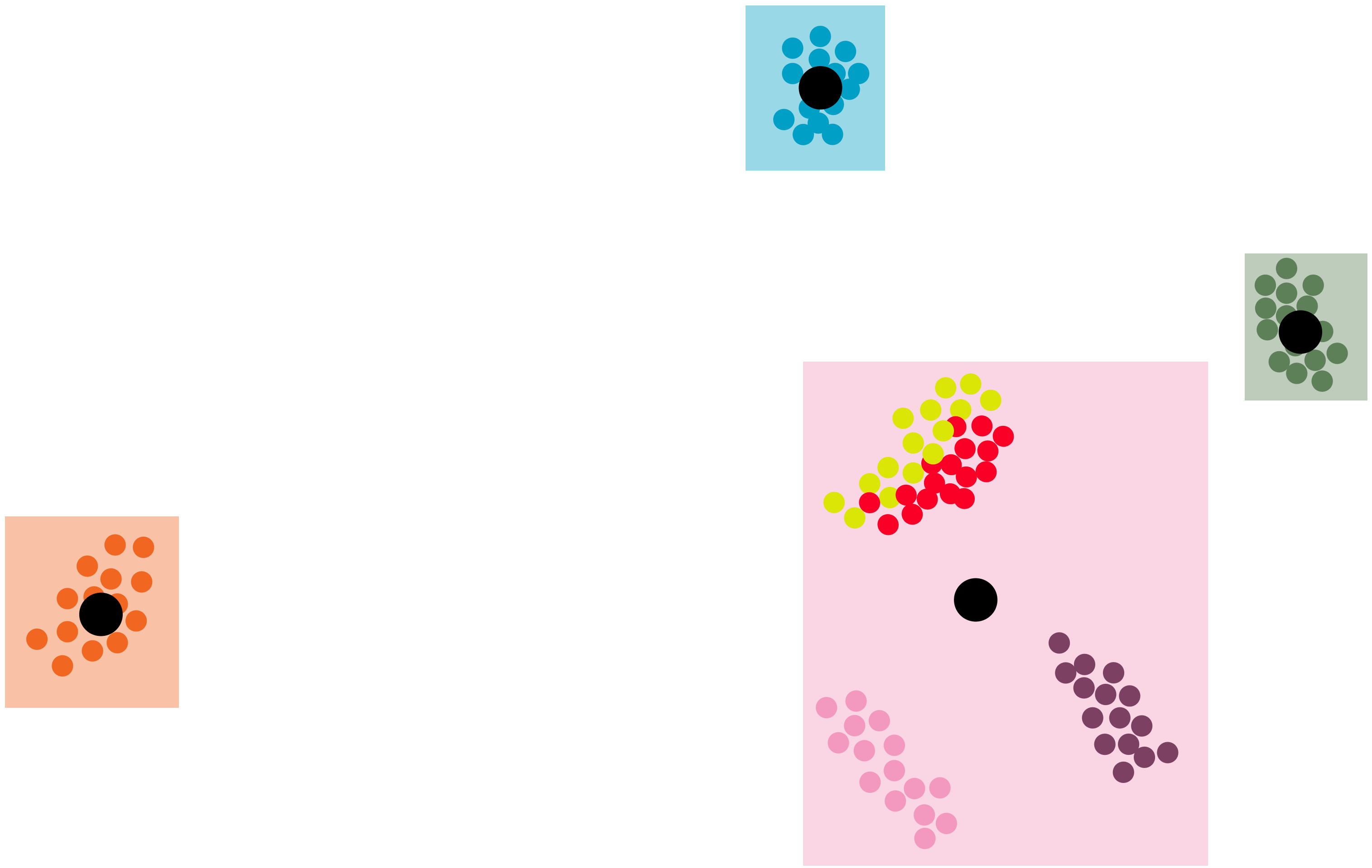
Unsupervised learning (prototype clustering)

K = 3 clusters



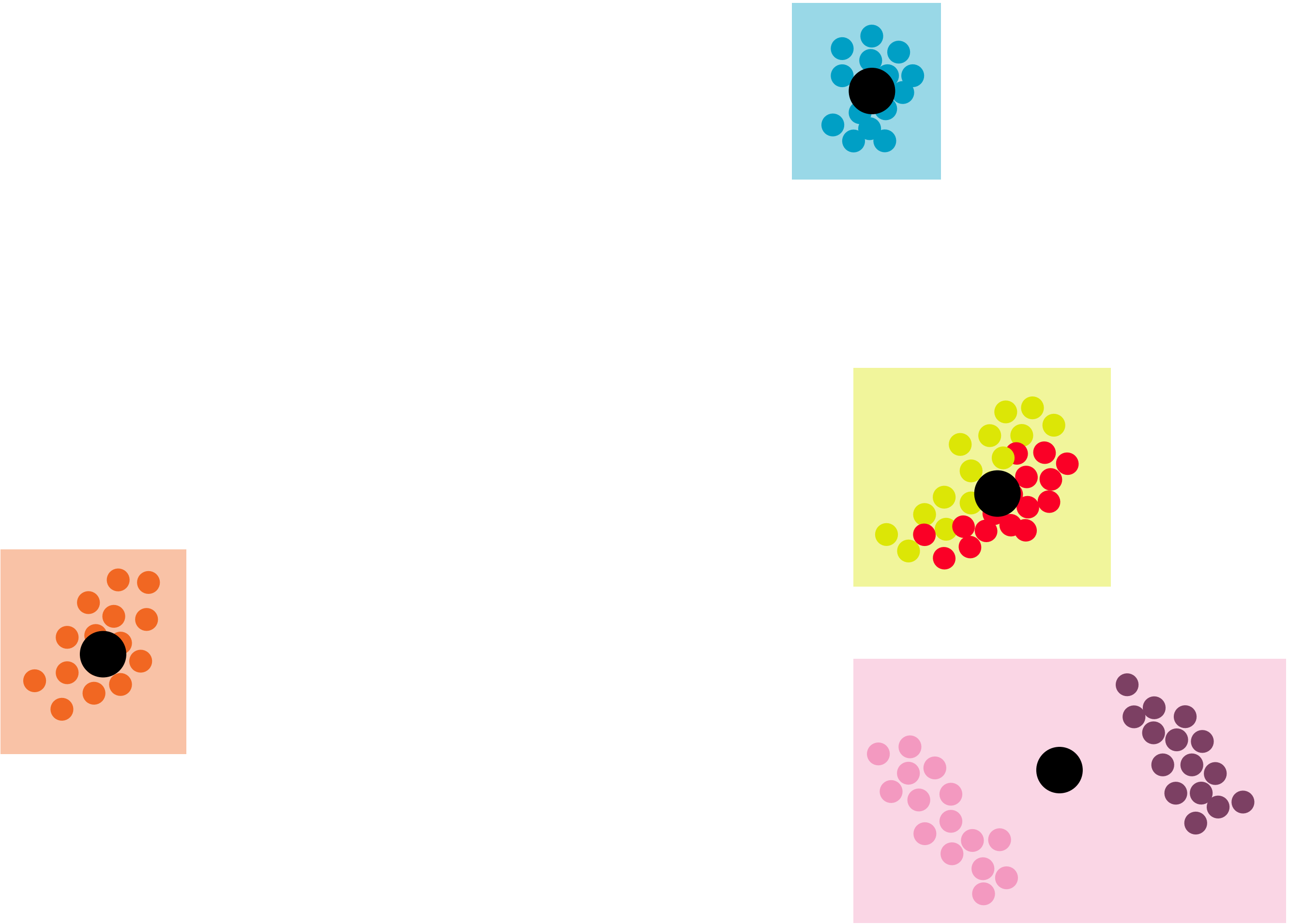
Unsupervised learning (prototype clustering)

K = 4 clusters



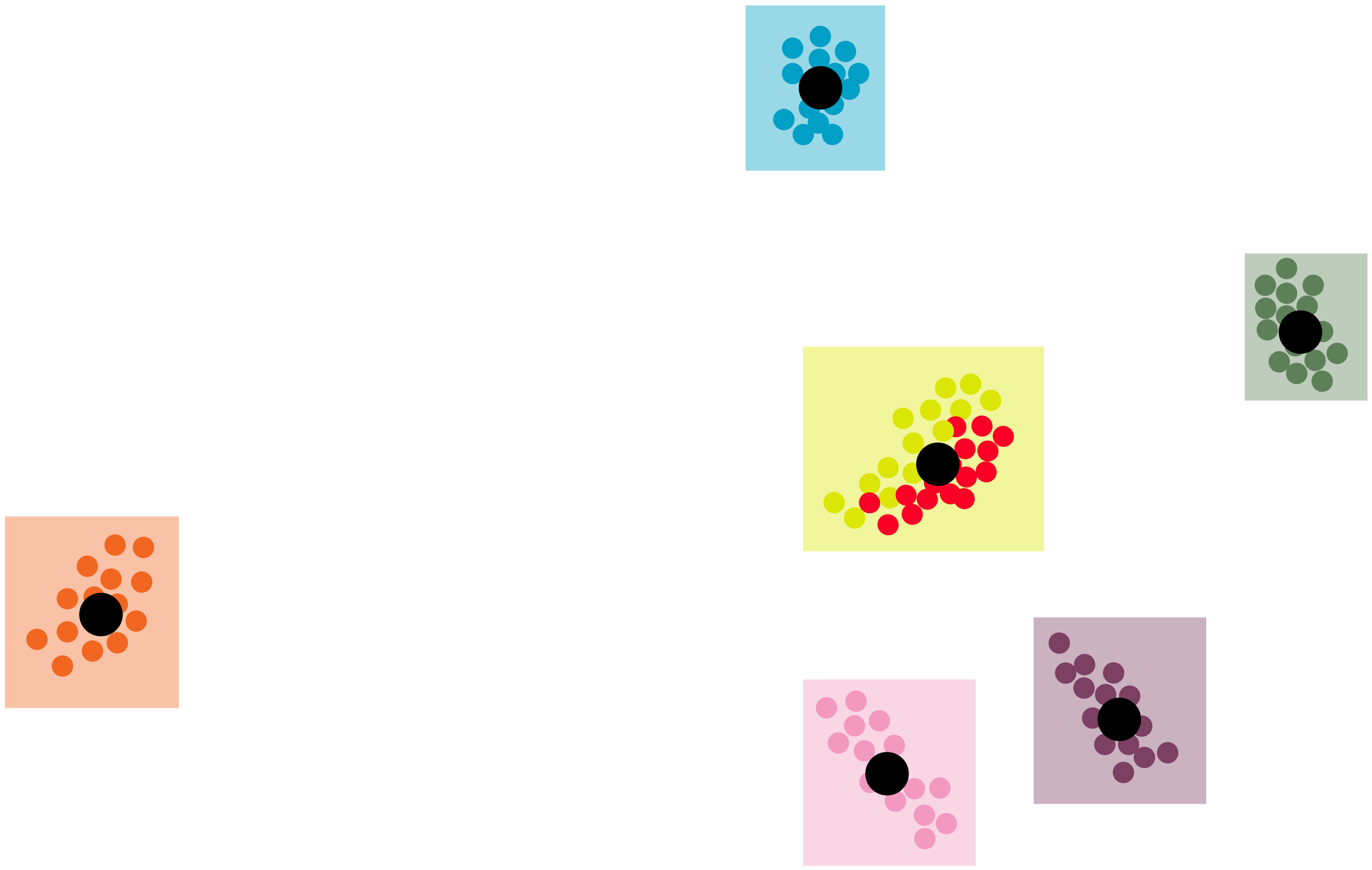
Unsupervised learning (prototype clustering)

K = 5 clusters



Unsupervised learning (prototype clustering)

K = 6 clusters



Example of prototype clustering

Example: identifying clusters of genetically similar individuals

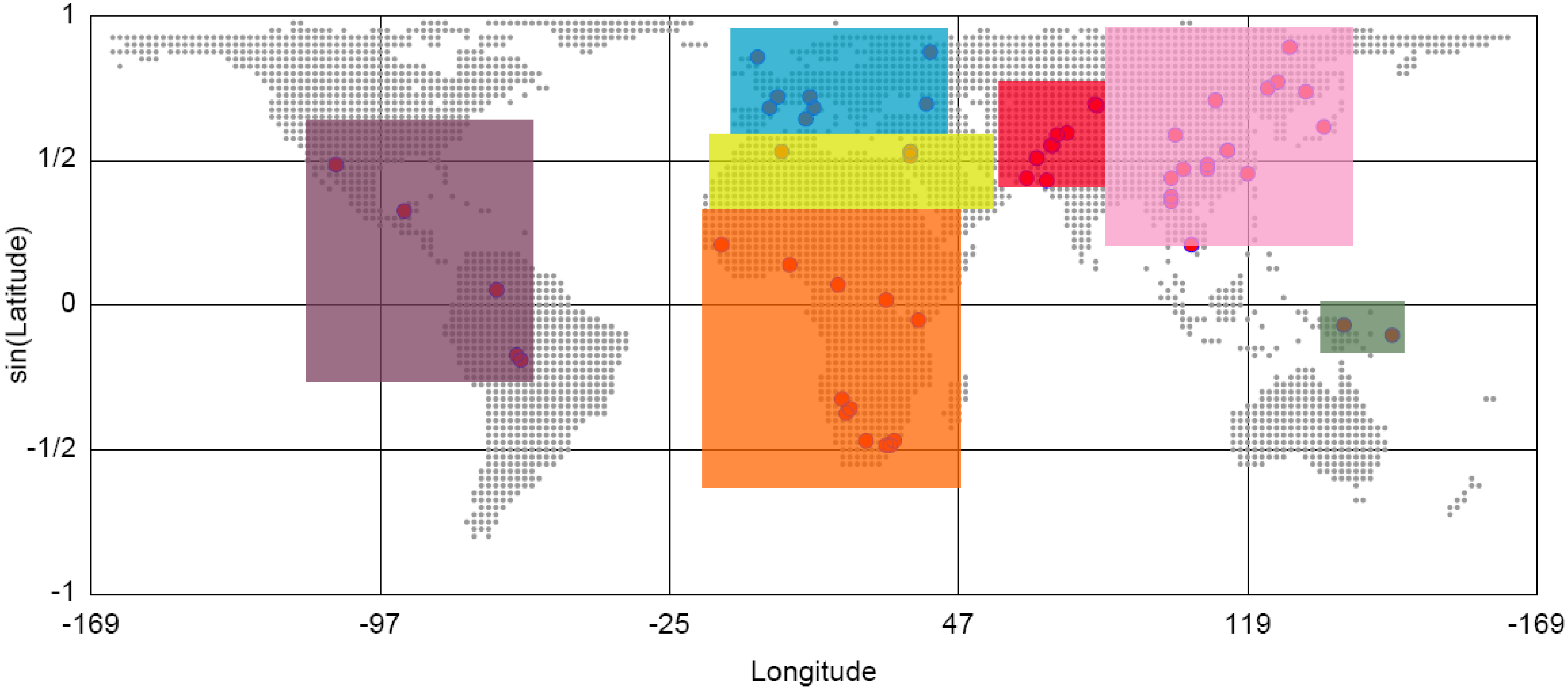
Goal: assign an individual to one of K clusters based on their genetic data.

Raw data: number of copies (0, 1, or 2), also known as a genotype, at many genomic locations (loci) across an individual’s genome.

Input data: vector with values 0, 1, or 2, of dimension equal to the number of genomic locations analyzed, and the number of clusters K.

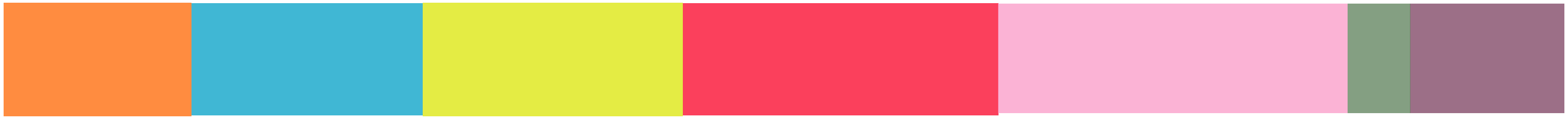
Rosenberg *et al*. (2002) *Science* 298:2381-285

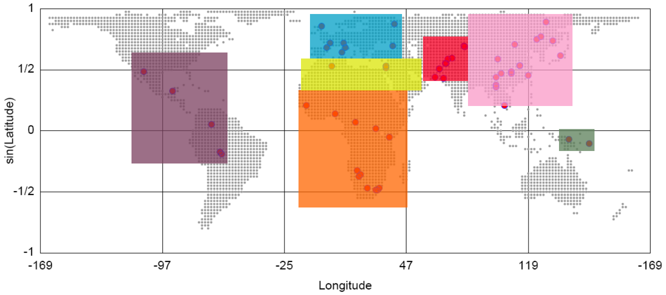
Geographic locations of sampled populations



Rosenberg *et al*. (2002) *Science* 298:2381-285

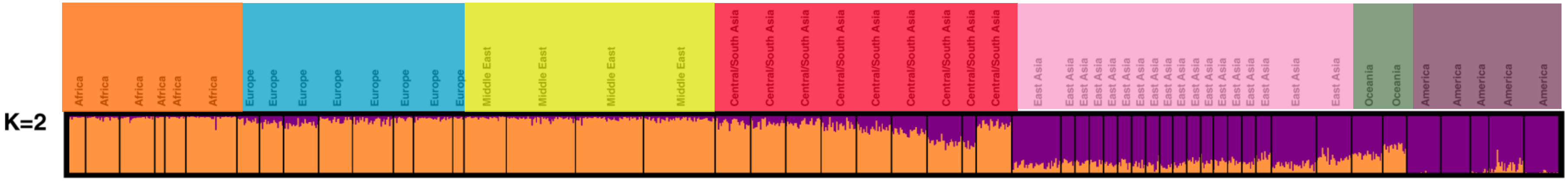
Example of prototype clustering

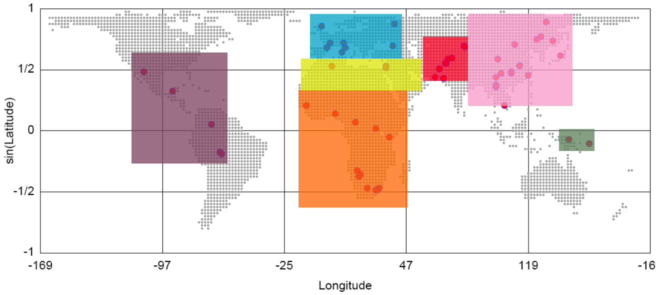




Rosenberg *et al*. (2002) *Science* 298:2381-285

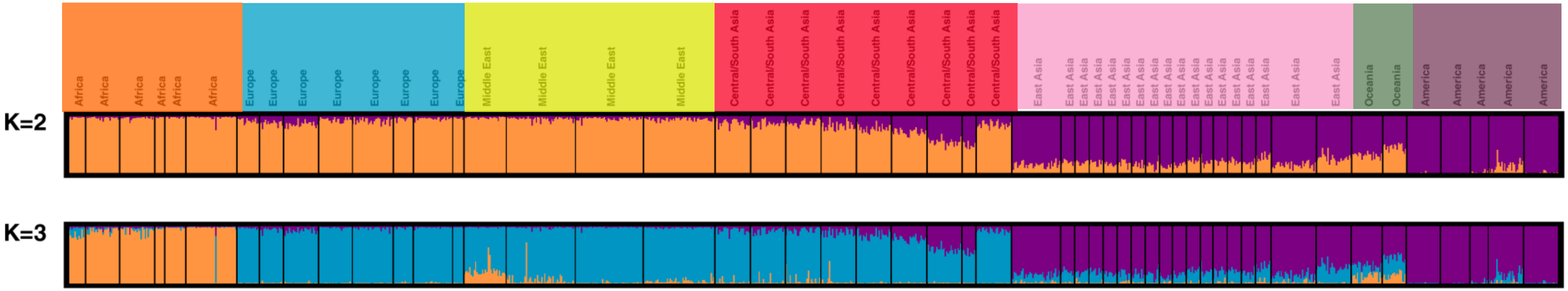
Example of prototype clustering (*K* = 2 clusters)

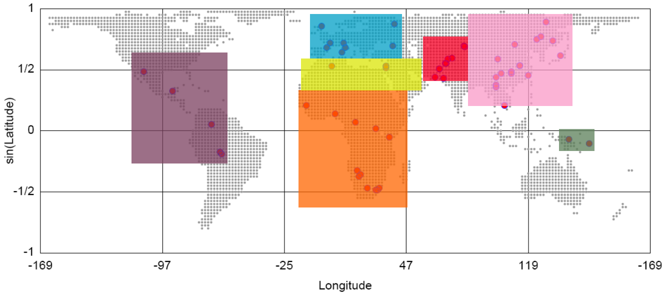




Rosenberg *et al*. (2002) *Science* 298:2381-285

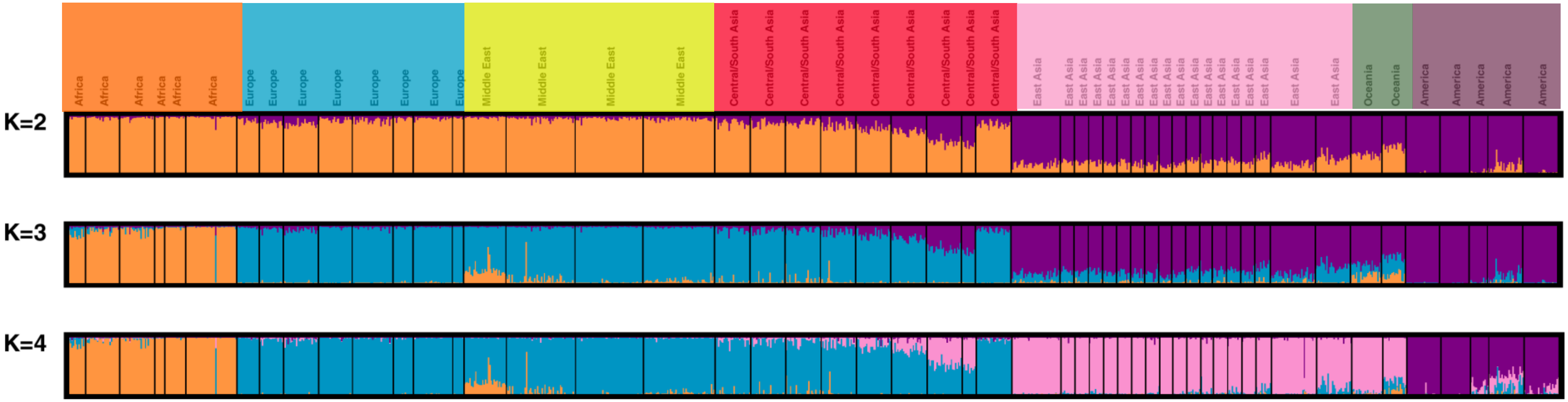
Example of prototype clustering (*K* = 3 clusters)

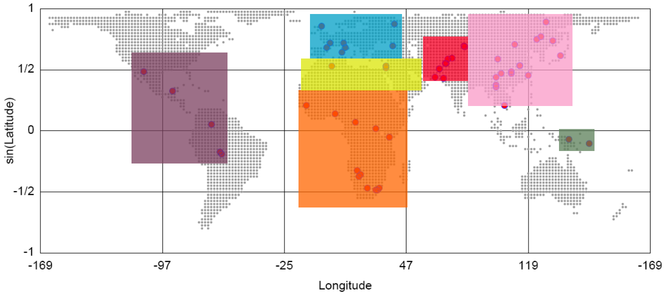




Rosenberg *et al*. (2002) *Science* 298:2381-285

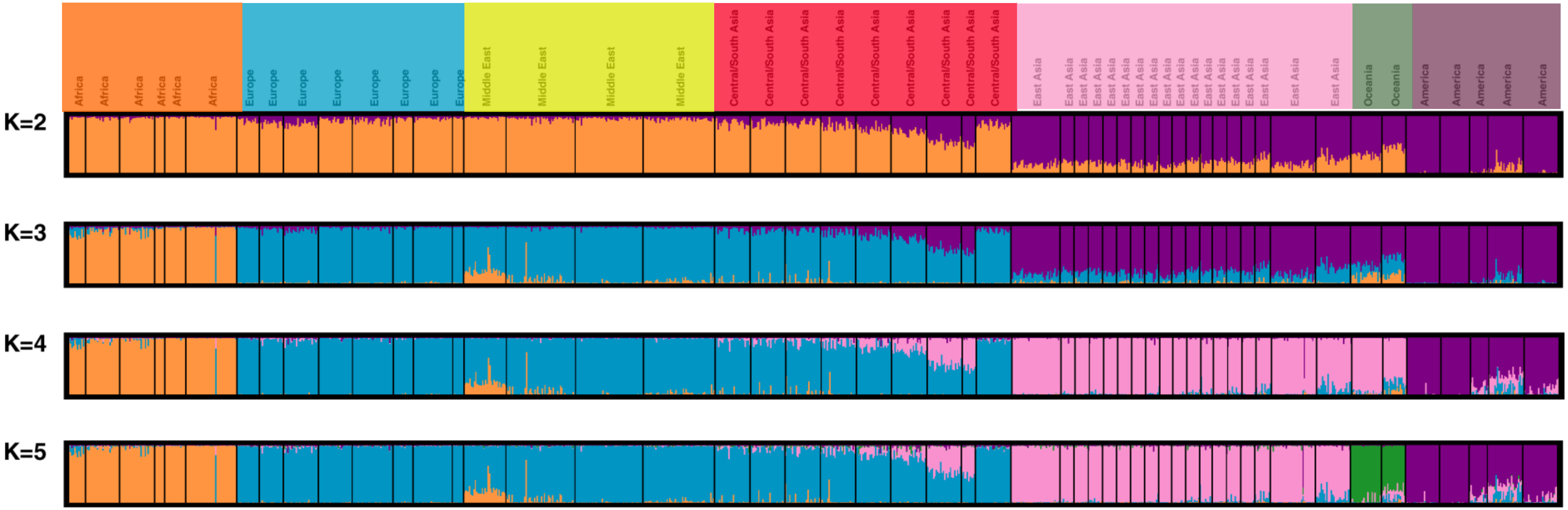
Example of prototype clustering (*K* = 4 clusters)

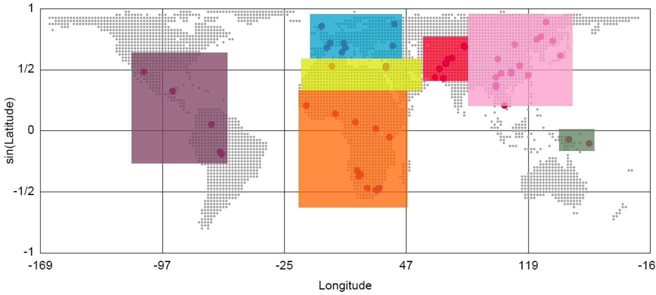




Rosenberg *et al*. (2002) *Science* 298:2381-285

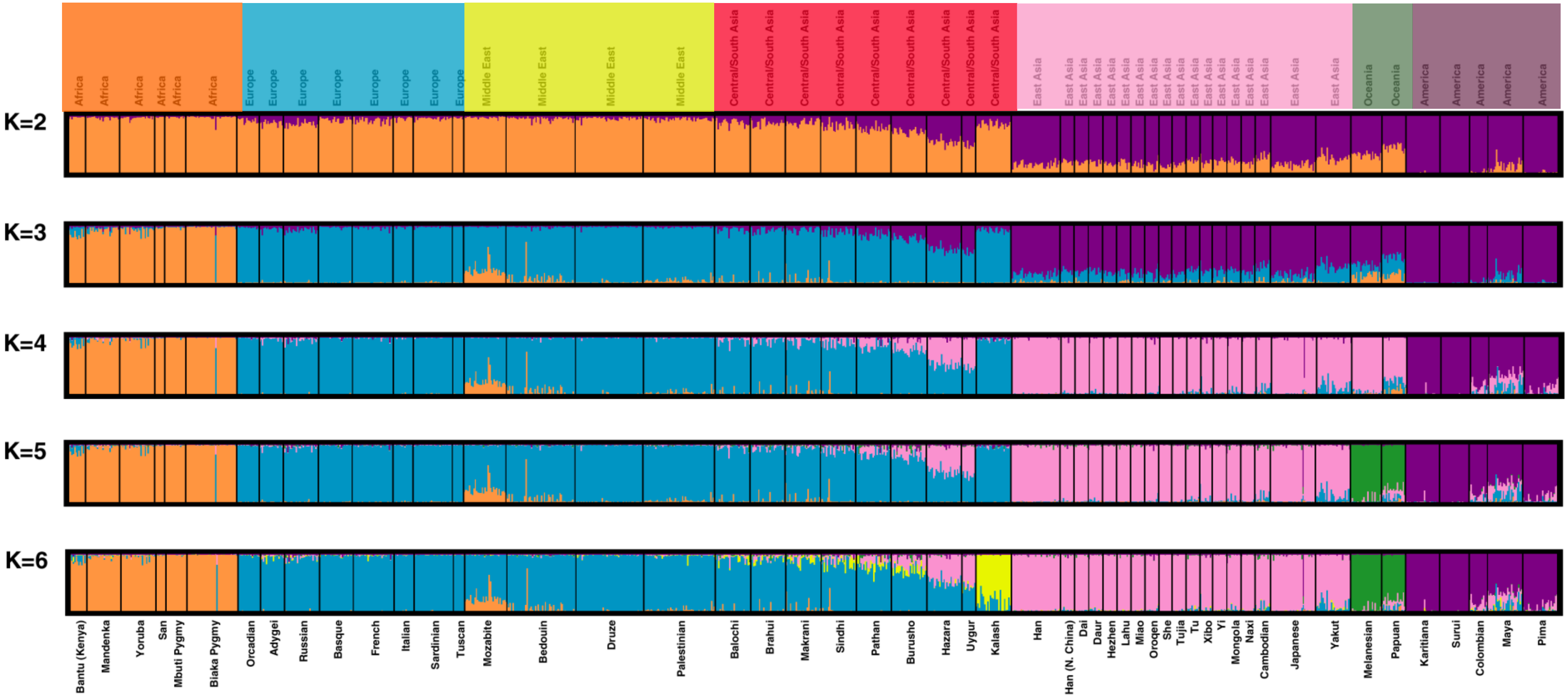
Example of prototype clustering (*K* = 5 clusters)

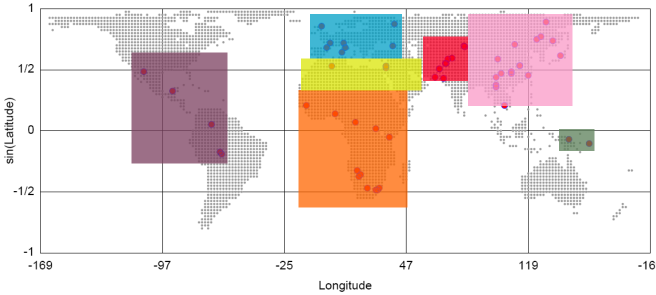




Rosenberg *et al*. (2002) *Science* 298:2381-285

Example of prototype clustering (*K* = 6 clusters)

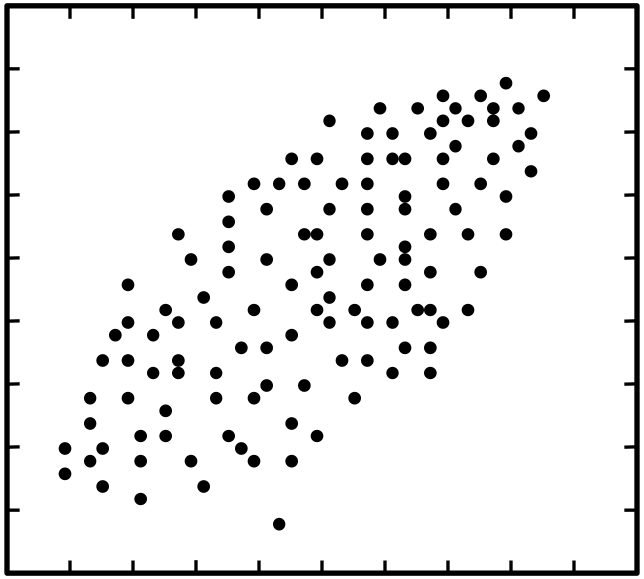




Rosenberg *et al*. (2002) *Science* 298:2381-285

Unsupervised learning (dimensionality reduction)

Principal components analysis (PCA)



- A popular technique for analyzing large datasets containing a high number of dimensions/features per observation.

- Increasing the interpretability of data while preserving the maximum amount of information

- Enabling the visualization of multidimensional data.

- Formally, PCA is a statistical technique for reducing the dimensionality of a dataset.

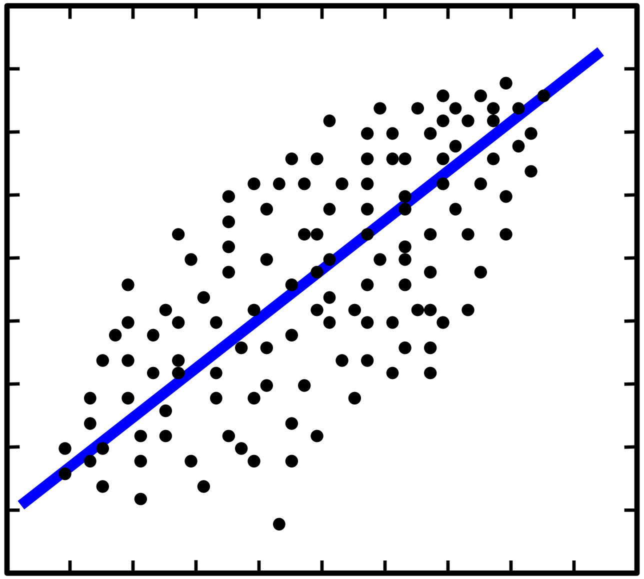
Solutions for high-dimensional data

•Feature Elimination

•Feature Extraction

Unsupervised learning (dimensionality reduction)

Principal components analysis (PCA)



When should I use PCA?

1. Do you want to reduce the number of variables, but aren't able to identify variables to completely remove from consideration?

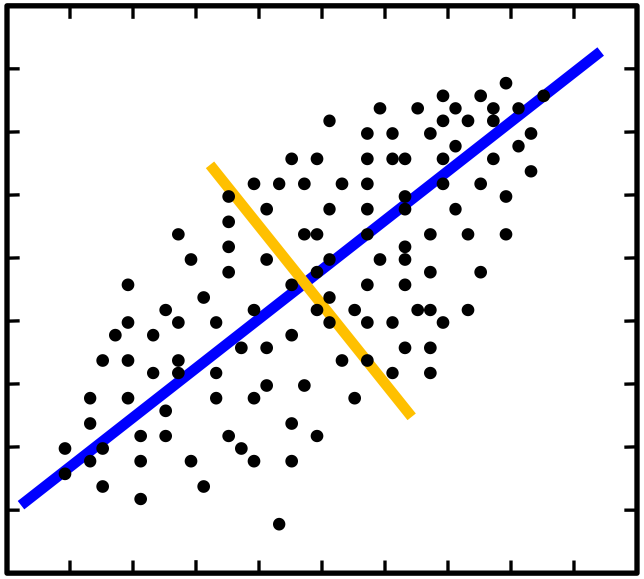
2. Do you want to ensure your variables are independent of one another?

3. Are you comfortable making your independent variables less interpretable?

https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

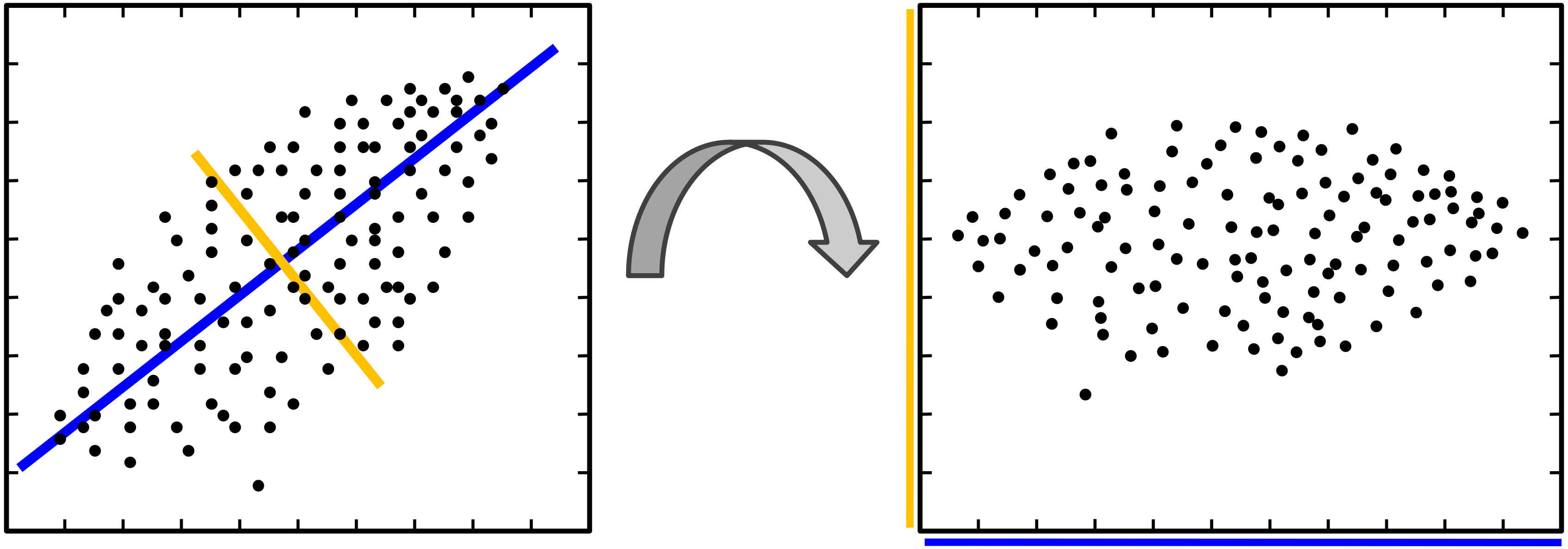
Unsupervised learning (dimensionality reduction)

Principal components analysis (PCA)



Unsupervised learning (dimensionality reduction)

Principal components analysis (PCA)



PCA on genetic data

Example: representing individuals in a low-dimensional space based on their genetic data

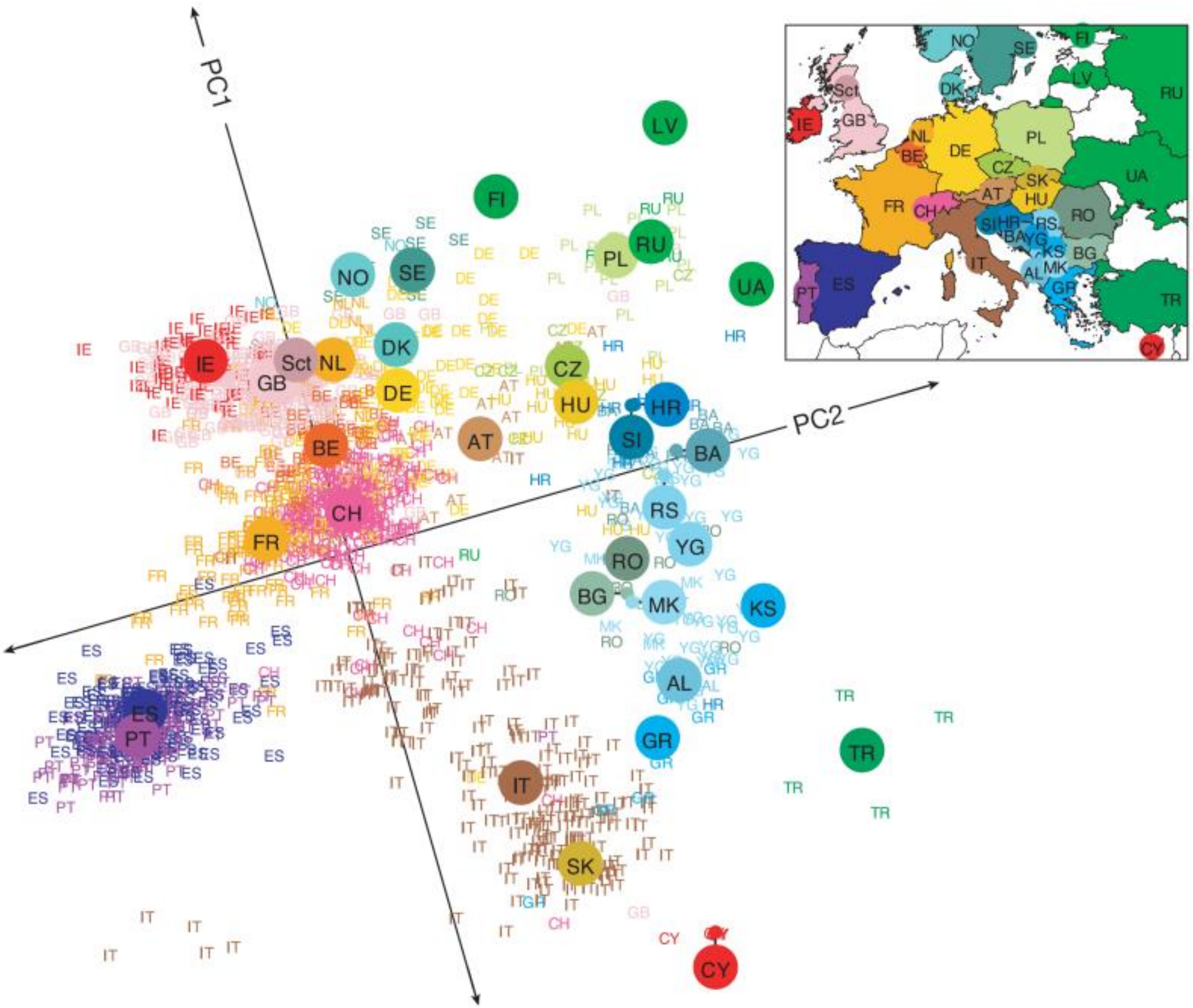
Goal: represent a set of individuals in a low-dimension (*e*.*g*., two dimensions) while retaining as much of the variability in the data as possible based on their genetic data.

Raw data: number of copies (0, 1, or 2), also known as a genotype, at many genomic locations (loci) across an individual's genome.

Input data: vector with values 0, 1, or 2, of dimension equal to the number of genomic locations analyzed.

Novembre *et al*. (2008) *Nature* 456:98-101

PCA reveals that genetics mirrors geography



Novembre *et al*. (2008) *Nature* 456:98-101

Assessing model accuracy

Predictions are useless if the model is wrong!

Ideal approach is to:

1) Use **training data** to train a method how to estimate the model

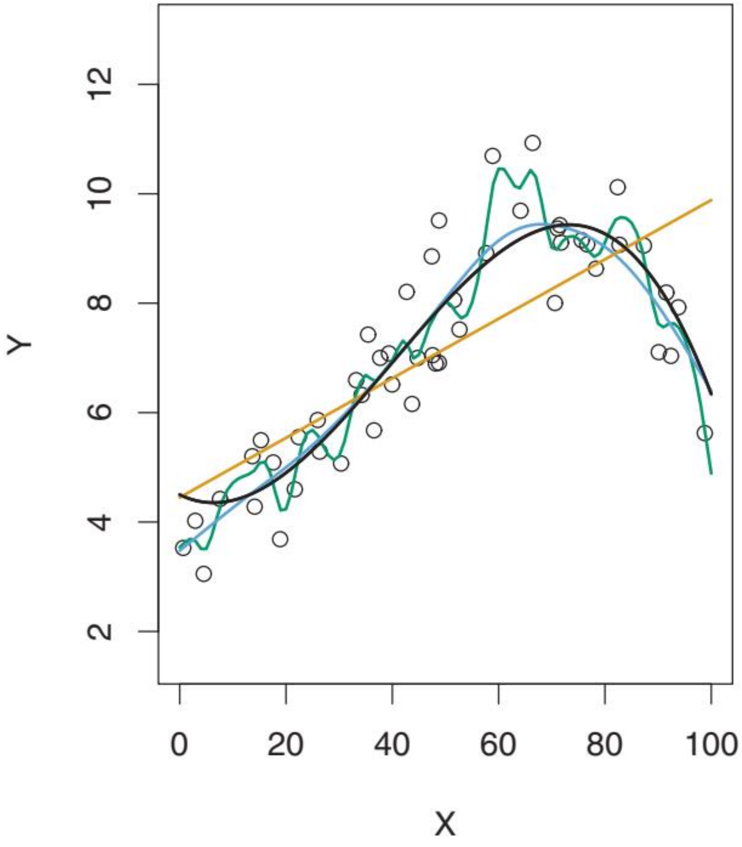
2) Apply the method to **test data** to make predictions

3) Evaluate performance of method on test data

How do we evaluate performance on test data?

How close is each predicted response to the true response?

Example: data simulated from f shown in black. Three estimates of f ($\hat{f}$) are depicted by orange, blue, and green curves.



In this regression setting, the most commonly-used measure is the **mean squared error (MSE)**, given by

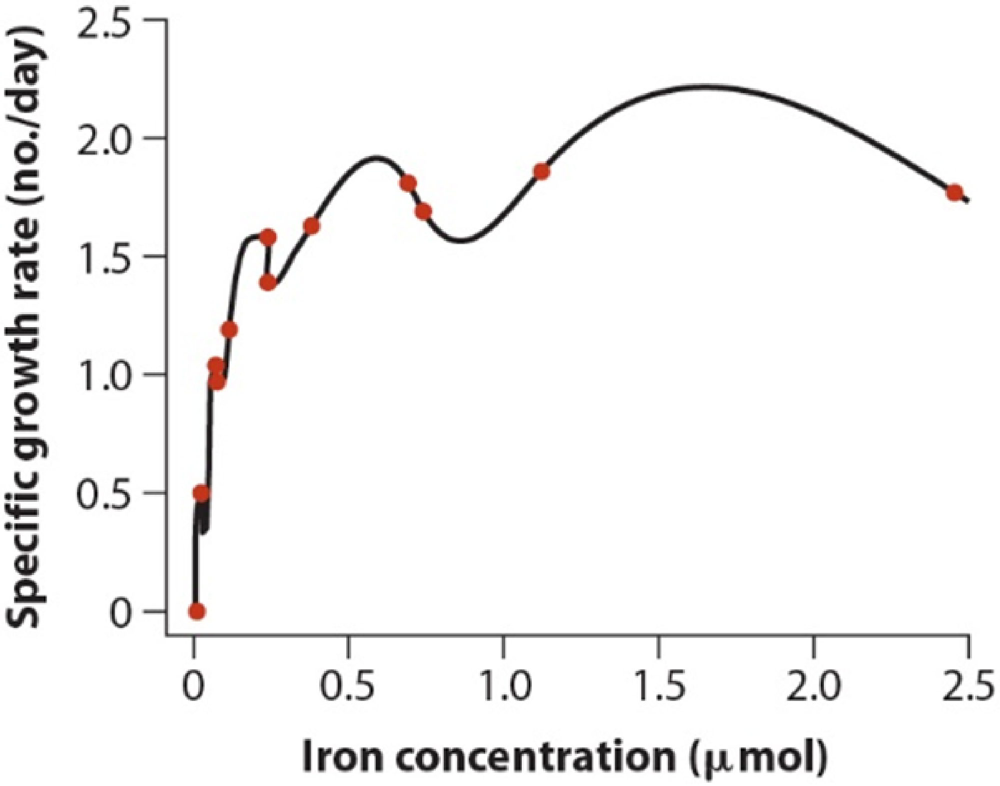
$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{f}(x_i)\right)^2$$

where $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for the ith observation.

Perfect accuracy is not always a good idea

How close is each predicted response to the true response?

Example: relationship between population growth rate of phytoplankton and the iron concentration in the medium



$\mathrm{MSE} = 0$, BUT:

How do we justify this?

Are all points important?

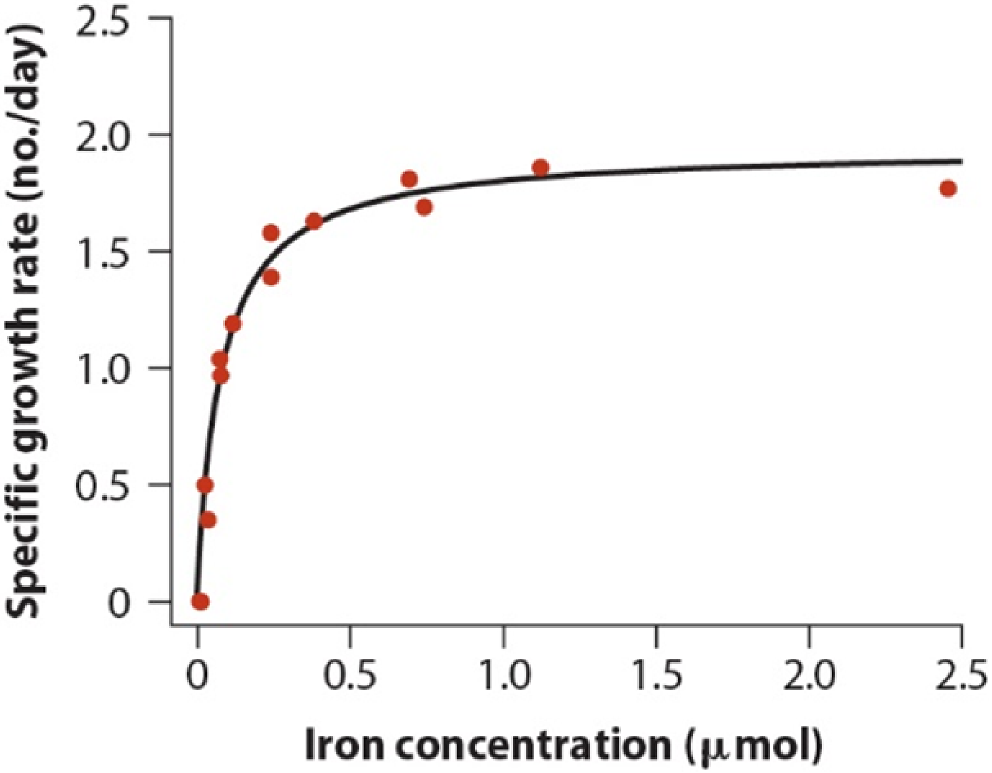
What will happen if there are new test data?

Sunda, Hunstman (1997) *Nature* 390:389-392.

Perfect accuracy is not always a good idea

How close is each predicted response to the true response?

Example: relationship between population growth rate of phytoplankton and the iron concentration in the medium



$\text{MSE} > 0$, but still small

The Michaelis-Menten equation is a much simpler function that captures the general trend of the data.

Sunda, Hunstman (1997) *Nature* 390:389-392.

Next …

Introduction To Python And Data Visualization