The Data Science Landscape: foundations, tools, and practical applications

Oge Marques, PhD

Professor

College of Engineering and Computer Science

Florida Atlantic University

@ProfessorOge

Outline

1. The era of Data Science
2. Data Science concepts and terminology
3. Data Science workflow/ecosystem
4. Exploratory Data Analysis (EDA) (*)
5. Statistics and Data Science (*)
6. Using data to answer questions (*)
7. Machine Learning and Data Science (*)
8. Data Science beyond the code
9. Recommended books and resources

Jupyter notebooks (examples)

tinyurl.com / icmla2019

Part 1:

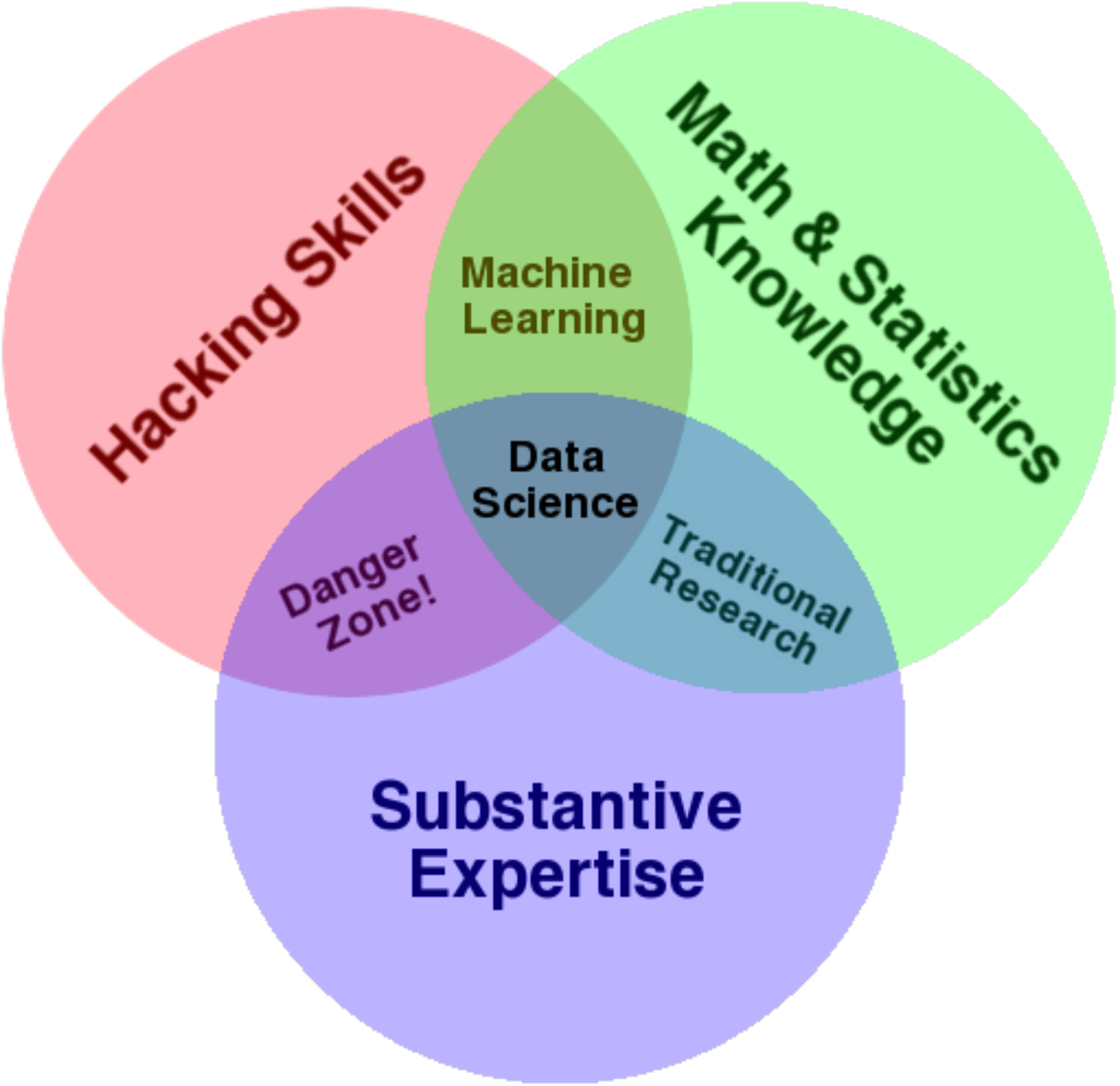
The era of Data Science

What is Data Science?

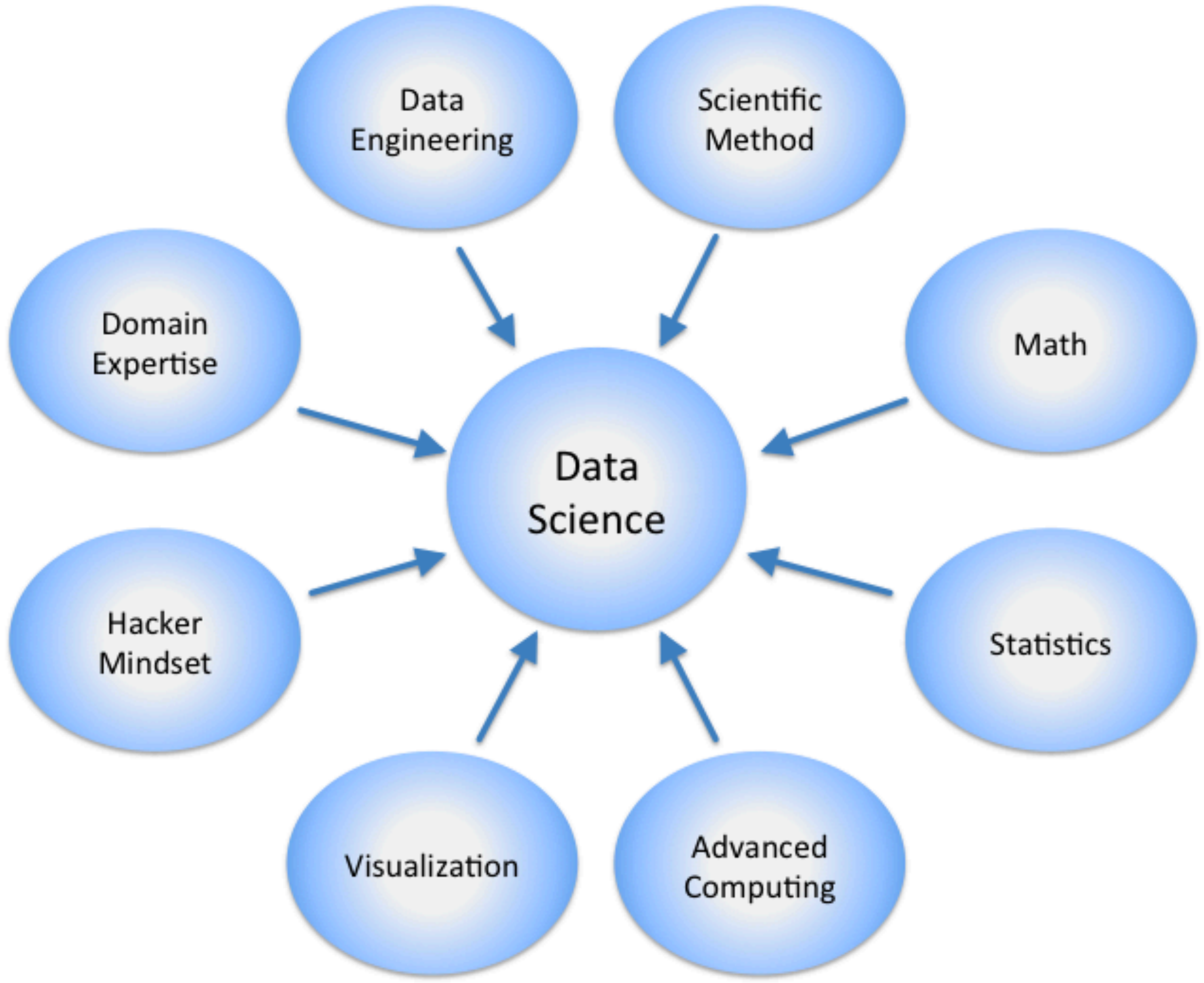
What is Data Science?

- “Data science [...] is perhaps the best label we have for the ***cross-disciplinary set of skills*** that are becoming increasingly important in many applications across industry and academia.”

-- Jake VanderPlas



Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram



Source: https://upload.wikimedia.org/wikipedia/commons/4/44/DataScienceDisciplines.png

The goal of data science is to improve decision making by basing decisions on insights extracted from large data sets. As a field of activity, data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting nonobvious and useful patterns from large data sets. It is closely related to the fields of data mining and machine learning, but it is broader in scope. Today, data science drives decision making in nearly all parts of modern societies.

Source: Kelleher and Tierney, "Data Science" (MIT Press, 2018)

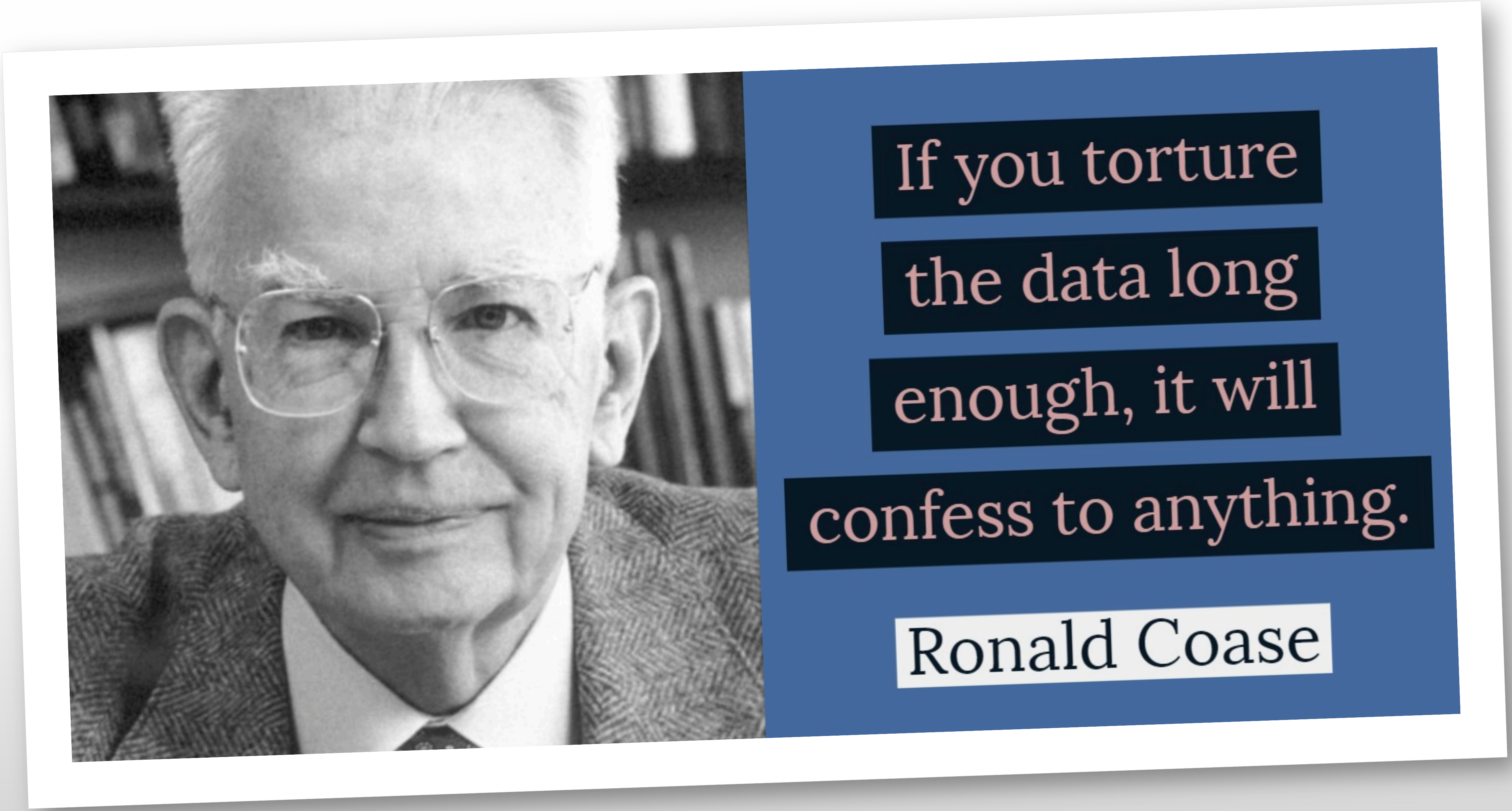
What is a Data Scientist?

What is a data scientist?

Searching the web for more information about the emerging term "data science," we encounter the following definitions from the Data Science Association's "Professional Code of Conduct"[6]

"Data Scientist" means a professional who uses scientific methods to liberate and create meaning from raw data.

Source: David Donoho (2017) 50 Years of Data Science, Journal of Computational and Graphical Statistics, 26:4, 745-766, DOI: 10.1080/10618600.2017.1384734



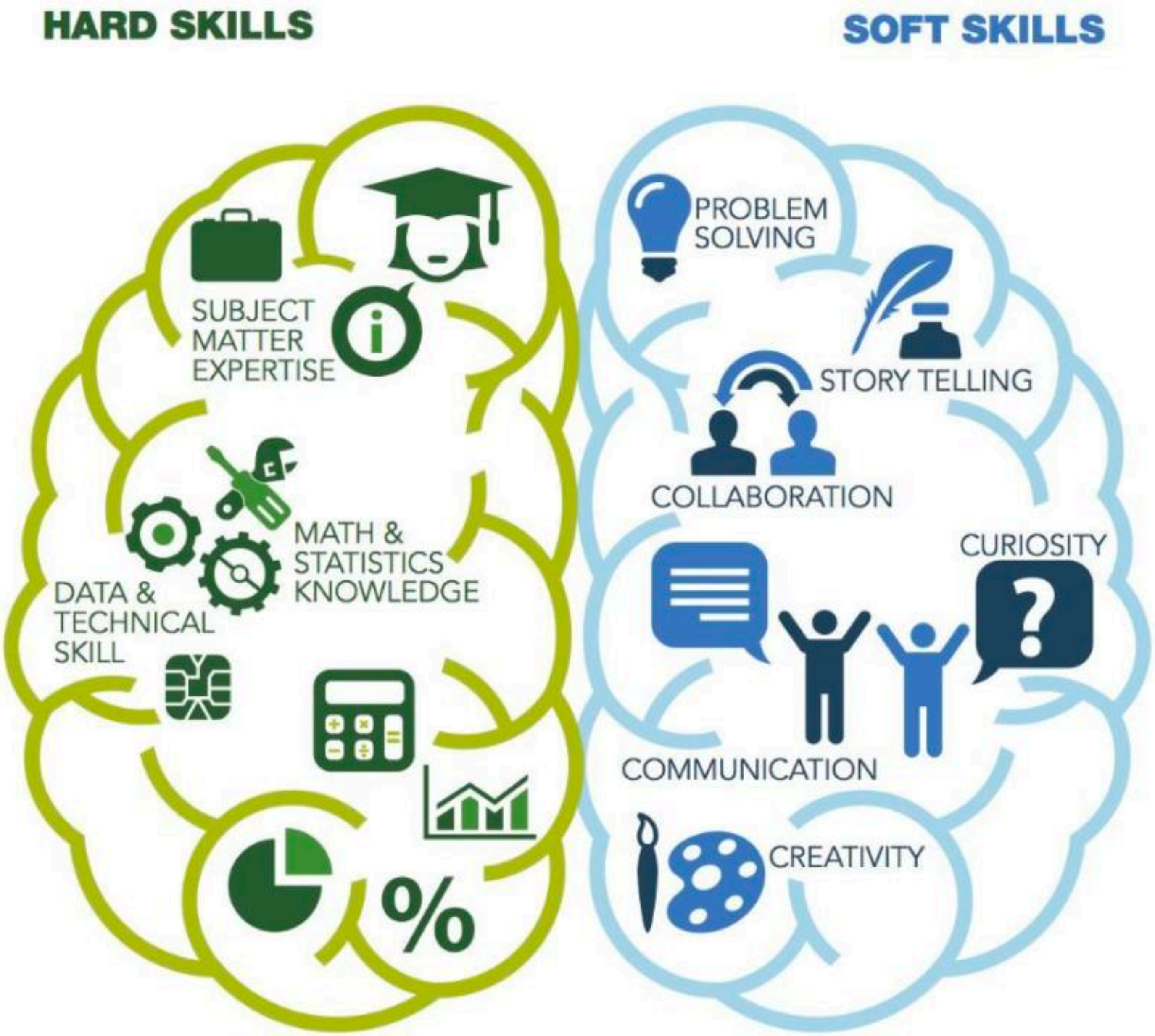
Source: https://www.reddit.com/r/QuotesPorn/comments/b76ujr/if_you_torture_the_data_long_enough_it_will/

What is a data scientist?

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

Source: David Donoho (2017) 50 Years of Data Science, Journal of Computational and Graphical Statistics, 26:4, 745-766, DOI: 10.1080/10618600.2017.1384734

Top 10 skills for Data Scientists



Source: https://www.linkedin.com/feed/update/urn:li:activity:6609150977335455744/

Data scientists use data to...

- Test hypotheses
- Model processes
- Predict outcomes
- Detect anomalies
- Train machine learning models
- Explain the world
- Extract information (and eventually knowledge)

What is driving Data Science?

What is driving Data Science?

- Enormous availability of (raw) data
- Open source tools
- Easy access to code and datasets
- Numerous use cases / applications
- Faster / ubiquitous computing platforms
- Lower barriers to enter

Why did you sign up for this tutorial?

Data Science is a hot field!

LinkedIn Workforce Report | United States | August 2018

Published on Aug 10, 2018

Skills Gaps | Demand for data scientists is off the charts

In 2015, there was a national surplus of people with data science skills. An employer in **Dallas** or **Atlanta** who wanted to hire data scientists had plenty of options; aside from in a few tech or finance-heavy cities like **San Francisco**, **New York City** and **Boston**, there weren't many local shortages.

But today, 3 years later, the picture has changed markedly: data science skills shortages are present in almost every large U.S. city. Nationally, we have a shortage of 151,717 people with data science skills, with particularly acute shortages in **New York City** (34,032 people), the **San Francisco Bay Area** (31,798 people), and **Los Angeles** (12,251 people). As more industries rely on big data to make decisions, data science has become increasingly important across all industries, not just tech and finance. In that sense, it's a good proxy for how today's cutting-edge skills like AI & machine learning may spread to other industries and geographies in the future.

Source: https://economicgraph.linkedin.com/resources/linkedin-workforce-report-august-2018

Linked in The Learning Blog

Paul Petrone May 9, 2019

The Skills New Grads Are Learning the Most

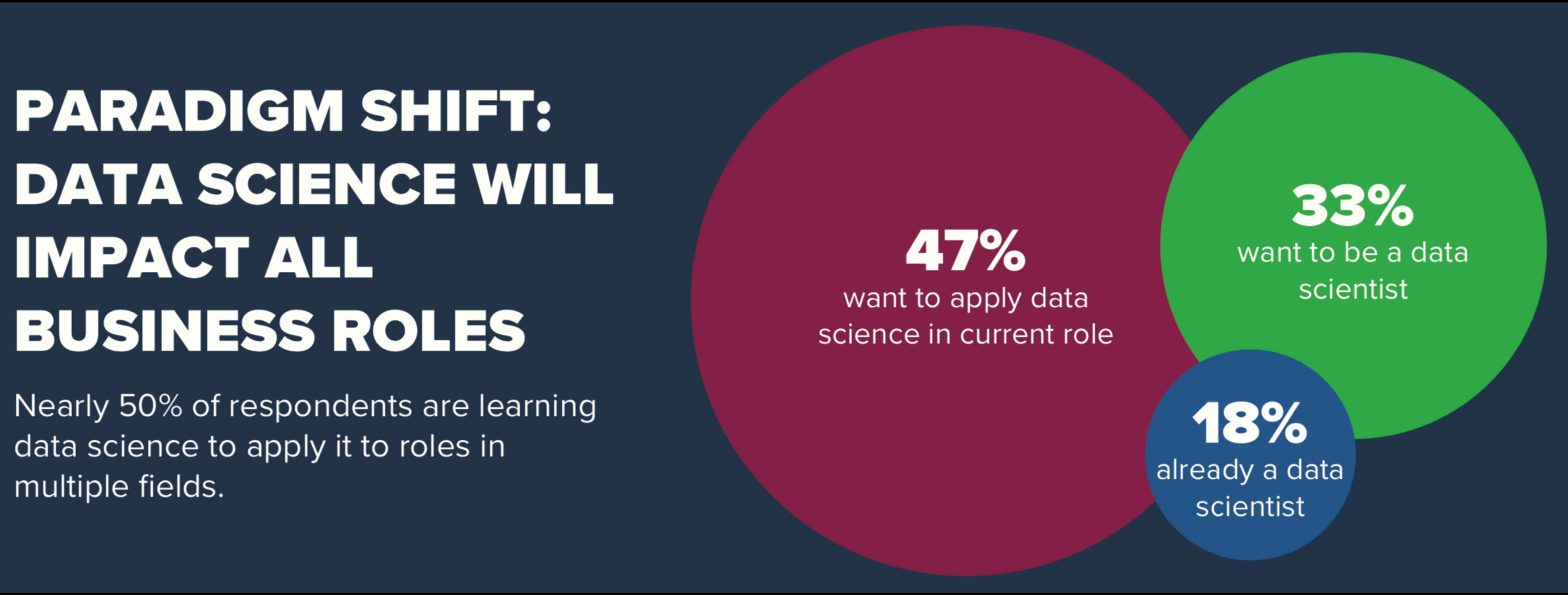
- Here are the five skills recent college graduates are disproportionately learning on LinkedIn Learning, compared to other professionals:
  - **Data Visualization**
  - **Data Modeling**
  - **Python**
  - **Web Analytics**
  - **Databases**
- This paints a clear picture – **all five skills directly relate to analyzing and storytelling with data**. And these skills are only becoming more important, as organizations become more data driven.

Source: https://learning.linkedin.com/blog/top-skills/the-skills-new-grads-are-learning-the-most

2019 STATE OF DATA SCIENCE

This spring, we surveyed nearly 5,000 members of the Anaconda community to understand current trends in data science. Here's what we found.

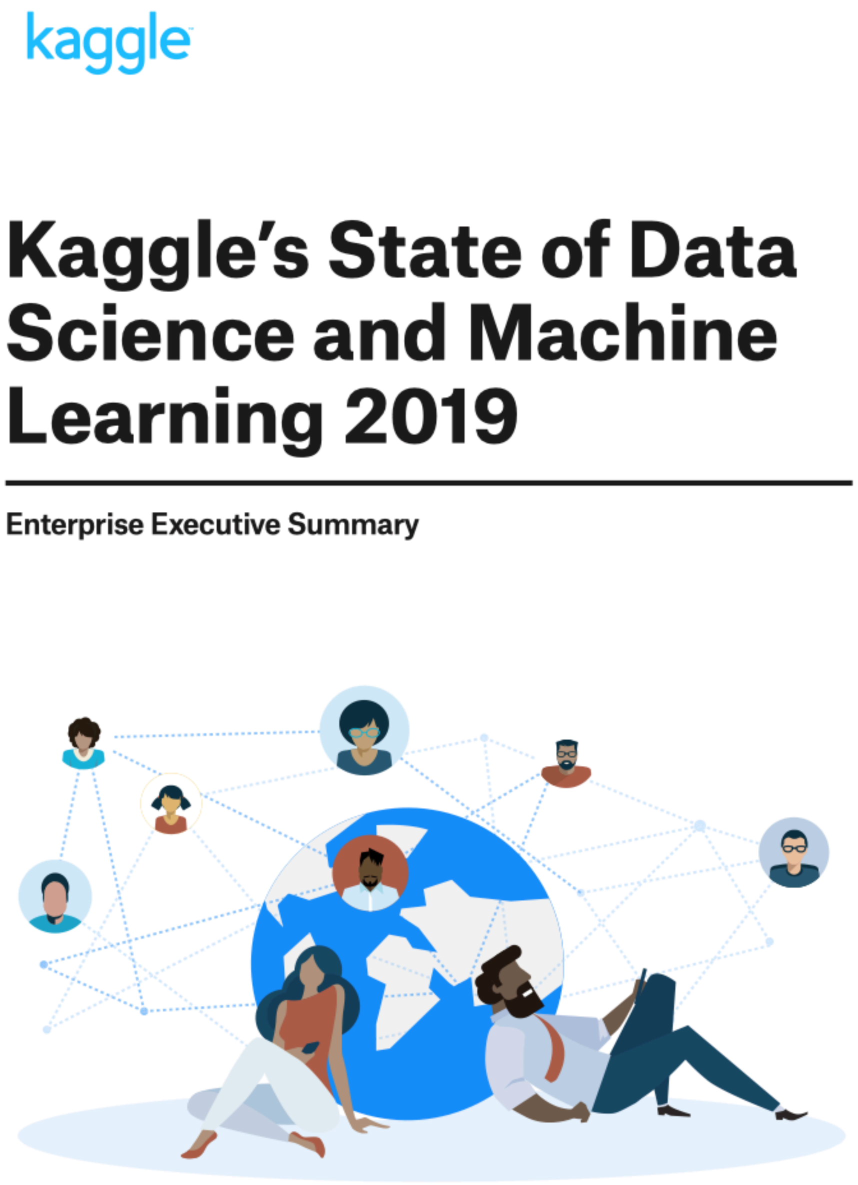
Source: https://know.anaconda.com/rs/387-XNW-688/images/2019-SoDS-Infographic-Anaconda.pdf



Source: https://know.anaconda.com/rs/387-XNW-688/images/2019-SoDS-Infographic-Anaconda.pdf

Highlights from

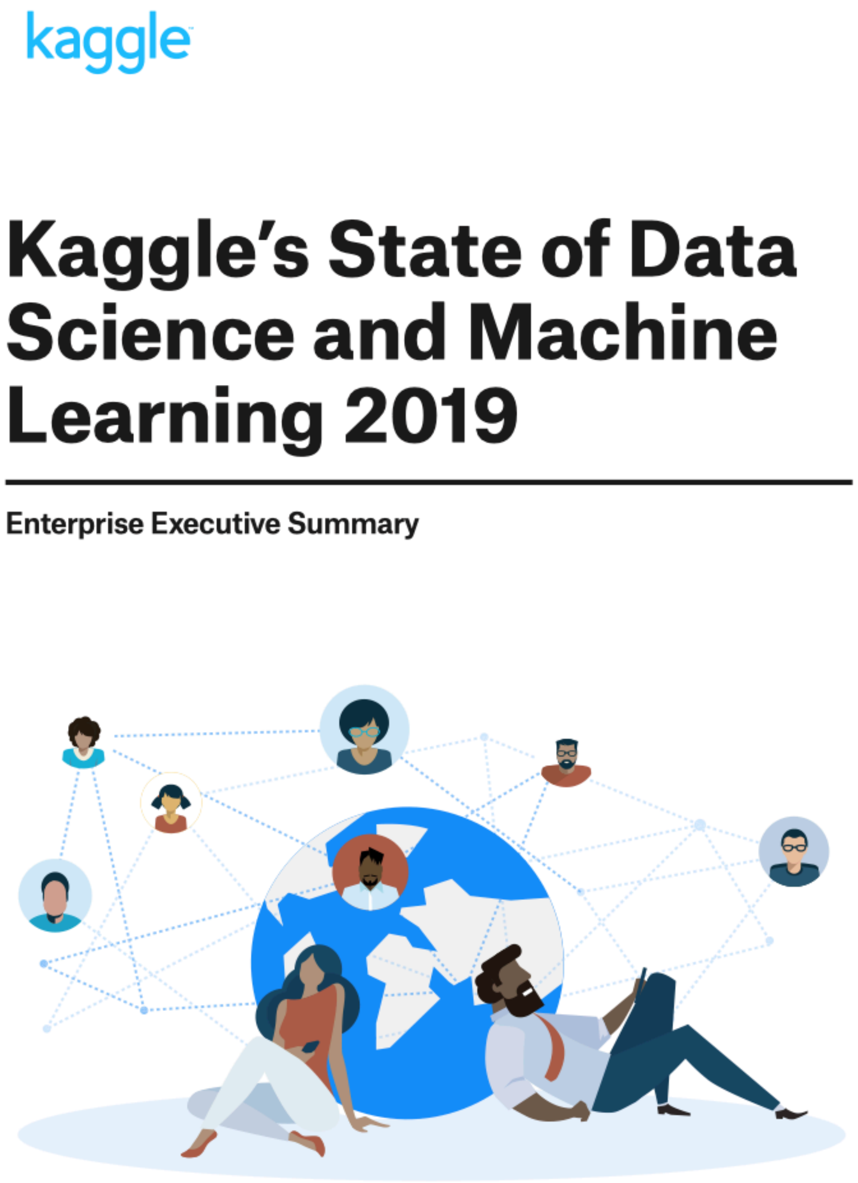
"Kaggle's State of Data Science and Machine Learning 2019"



Source: https://www.kaggle.com/kaggle-survey-2019

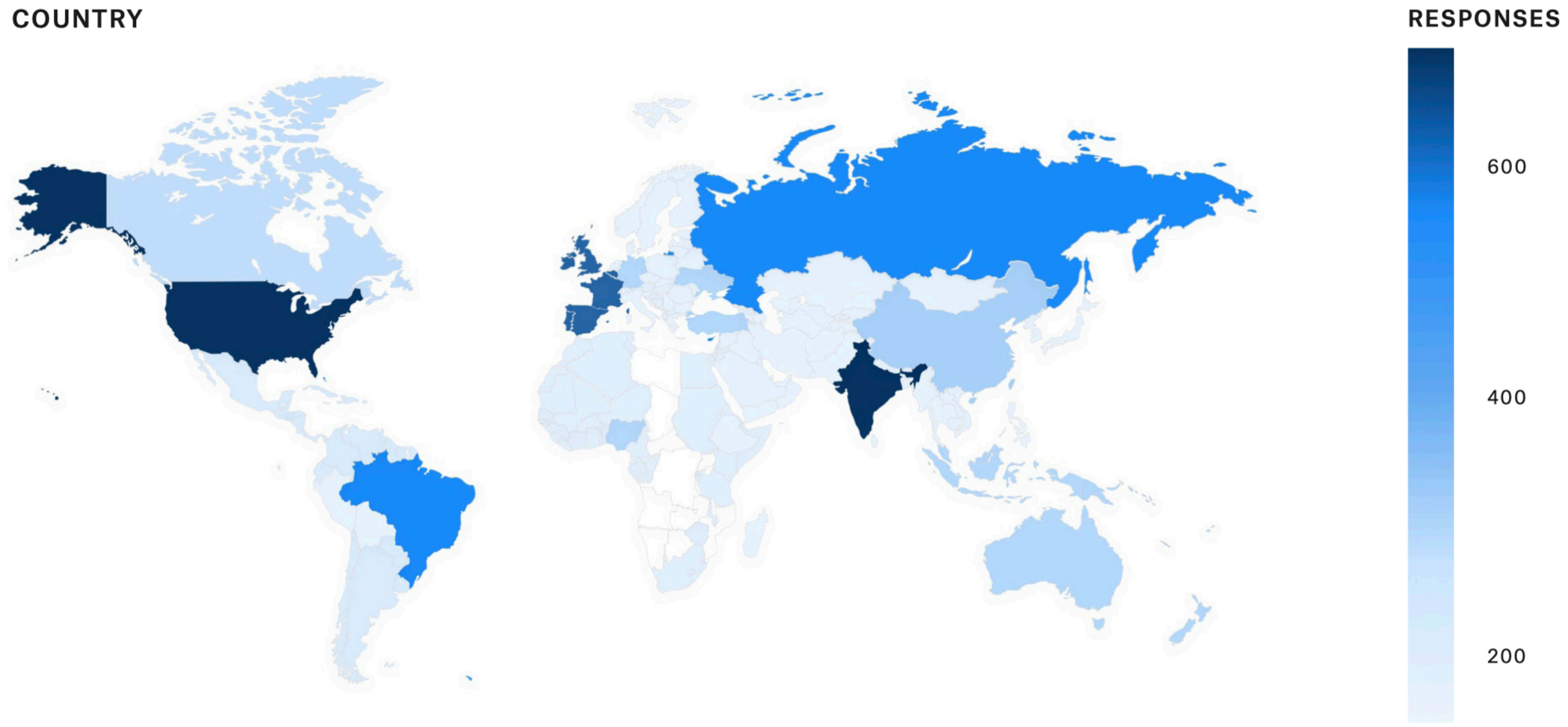
Respondents

- **19,717 Kaggle members worldwide**
- Selected charts and results are culled from professional data scientists (covering 21% of respondents)
- Kaggle has published the complete dataset of responses for the community to review:
  - https://www.kaggle.com/kaggle-survey-2019



Source: https://www.kaggle.com/kaggle-survey-2019

Geographical distribution of respondents



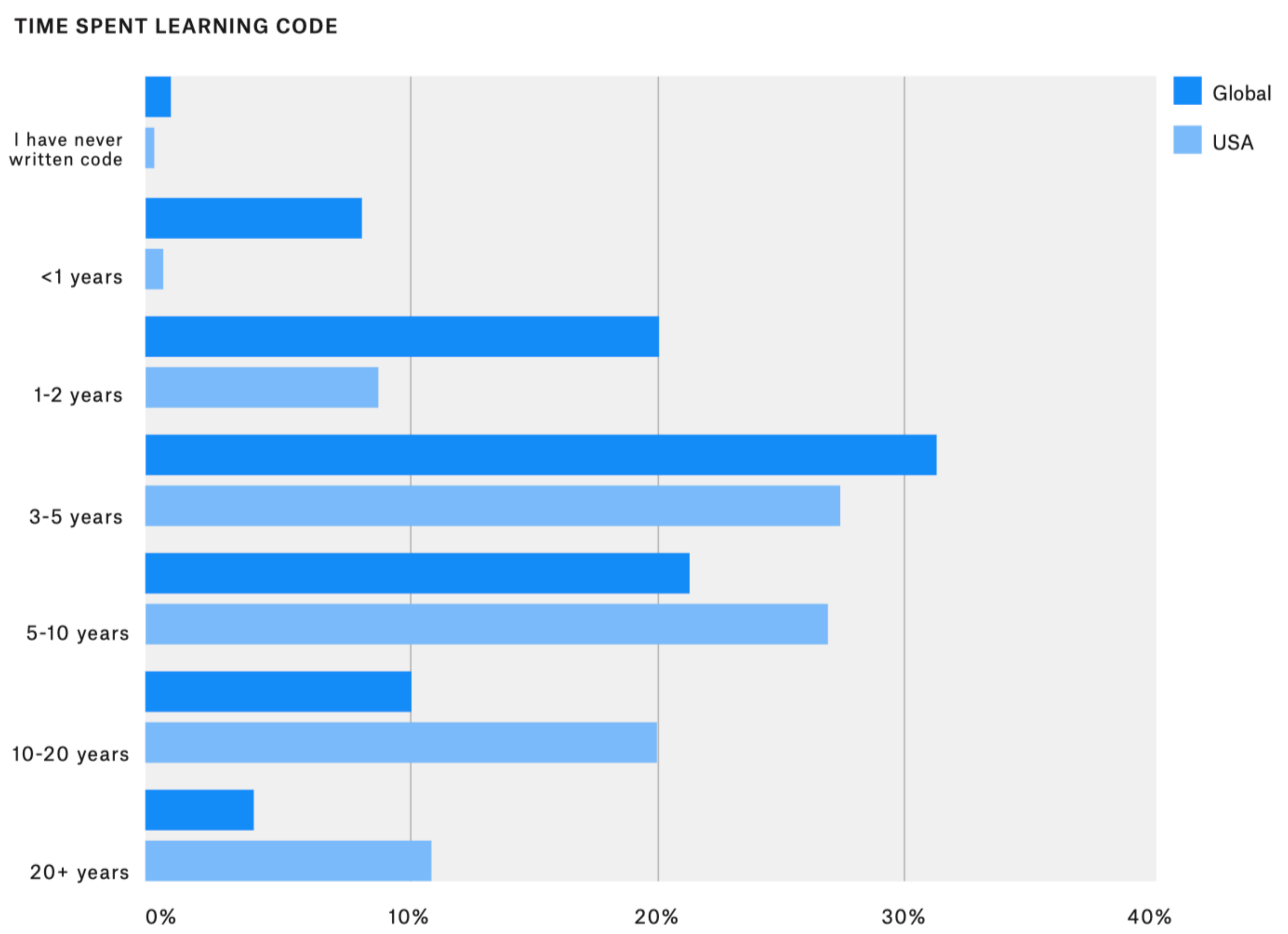
Source: https://www.kaggle.com/kaggle-survey-2019

Key results

- Data science is mostly male, an imbalance that has remained unchanged from previous years.
- Over half of data scientists are less than thirty years old.
- Unsurprisingly, data scientists are highly educated, with well over half obtaining advanced degrees.
- More than half of respondents have fewer than five years of coding experience and even less experience with machine learning.
- Salaries for data scientists in the United States far exceed other countries.
- Local development environments are the most common way data scientists perform their work.
- Nearly one in four professional data scientists have still not adopted cloud computing.
- Simple methods, such as linear regressions and decision trees, dominate despite the power of more complex techniques.

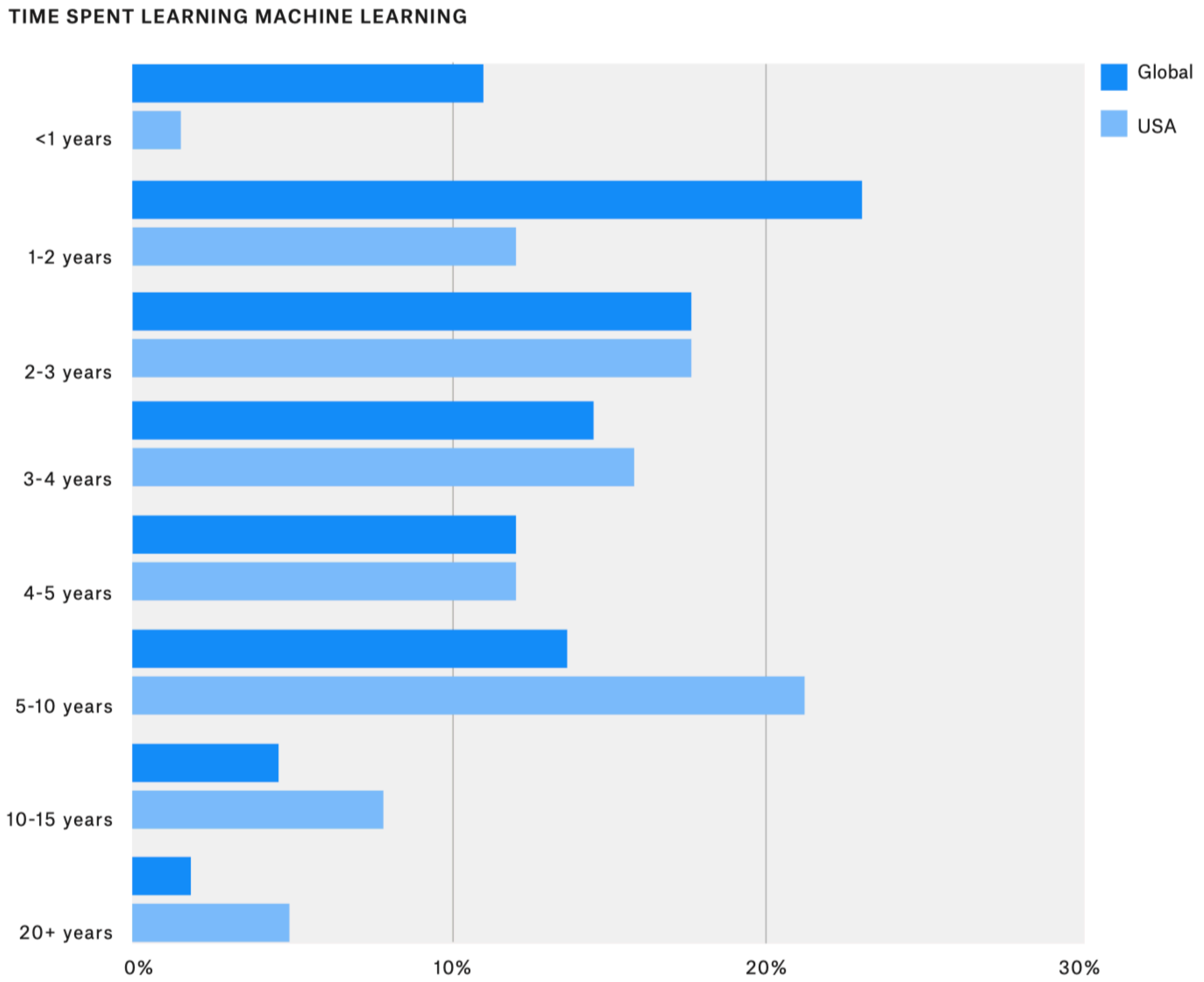
Source: https://www.kaggle.com/kaggle-survey-2019

Data Science & Machine Learning Experience



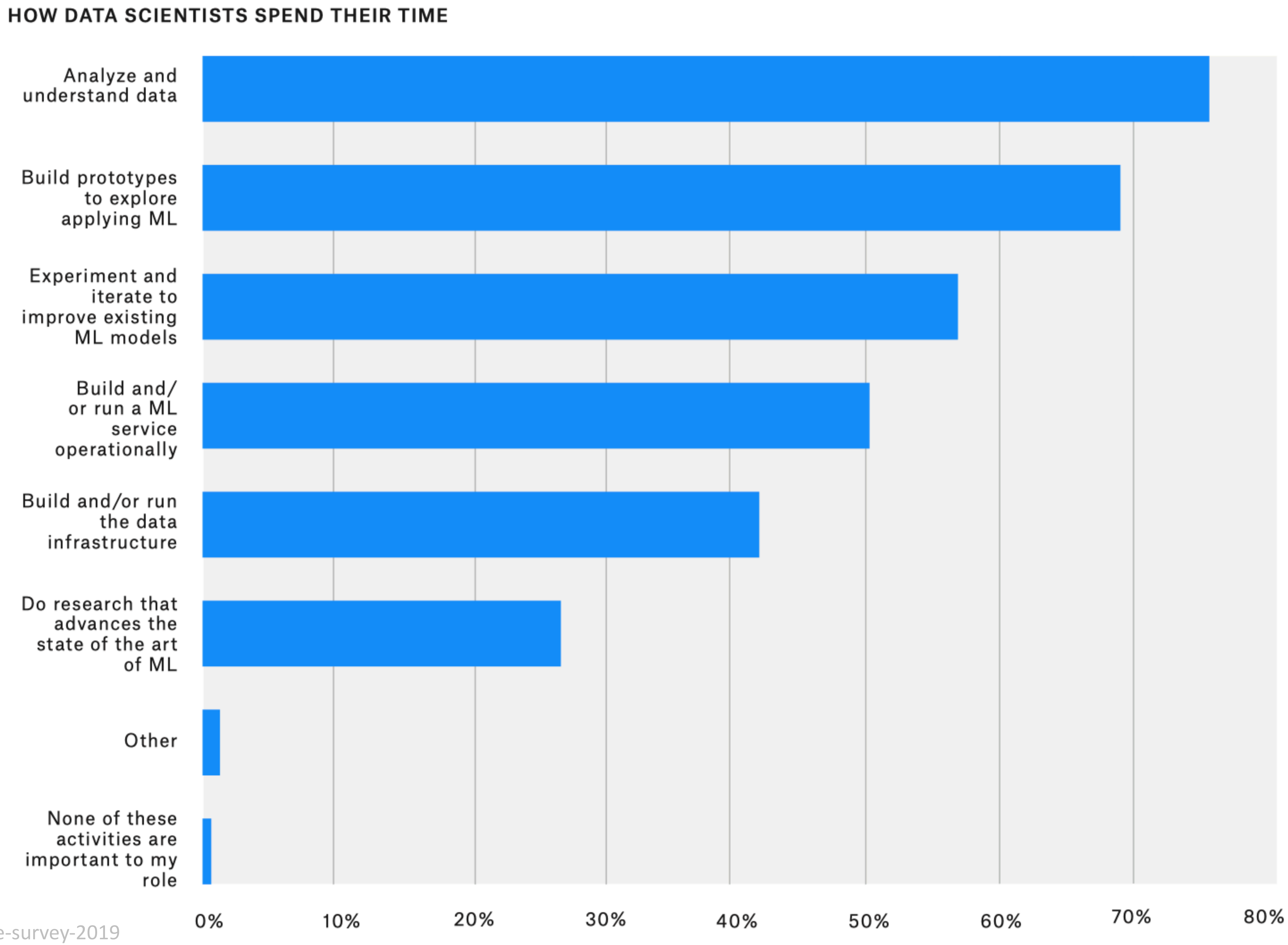
Source: https://www.kaggle.com/kaggle-survey-2019

Data Science & Machine Learning Experience



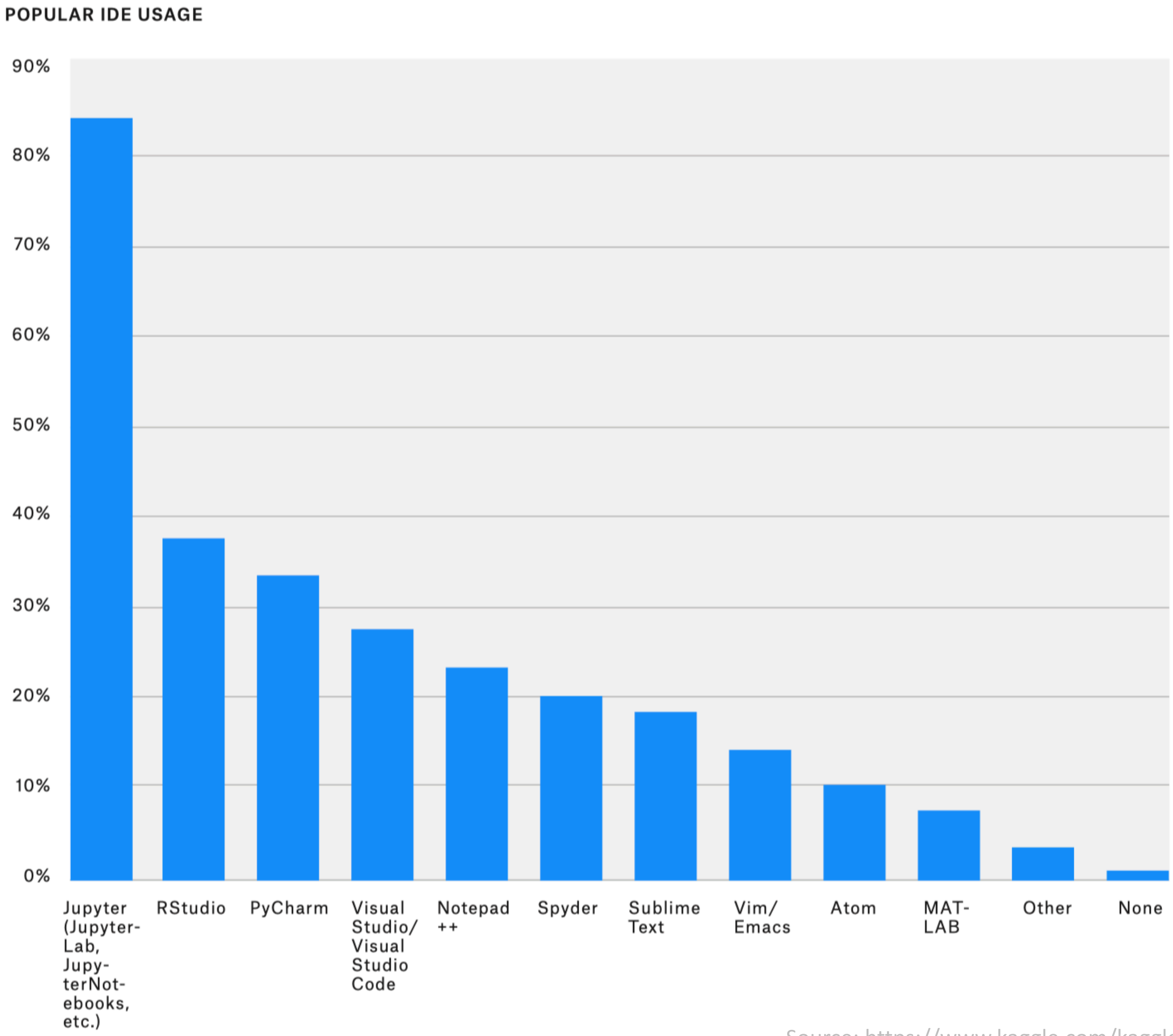
Source: https://www.kaggle.com/kaggle-survey-2019

Time



Source: https://www.kaggle.com/kaggle-survey-2019

IDEs



Source: https://www.kaggle.com/kaggle-survey-2019

What will we cover in this tutorial?

- How to use contemporary tools to develop a "data science workflow/pipeline"
  - Python
  - Jupyter notebooks
  - NumPy
  - Pandas
  - Matplotlib / seaborn
  - scikit-learn
- Selected math/statistics topics
- Selected Machine Learning algorithms
- Critical thinking, perspective, broad view of data science problems
- Learning resources

What will this tutorial not cover?

- Python programming
- Advanced Machine Learning algorithms
- Neural Networks
- Deep Learning
- “Big Data” tools, frameworks, etc.
- …

Let’s get started!

Part 2:

Data Science concepts and terminology

General remarks

- The goal of data science is to <u>improve</u> decision making by basing decisions on <u>insights</u> extracted from large data sets.
- Data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting <u>nonobvious</u> and <u>useful</u> patterns from large data sets.
- Many of the elements of data science have been developed in <u>related fields</u> such as *machine learning* and *data mining*.
- In general, data science becomes <u>useful</u> when we have a large number of data examples and when the patterns are too complex for humans to discover and extract manually.

Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)

If a human expert can easily create a pattern in his or her own mind, it is generally not worth the time and effort of using data science to "discover" it.

Source: Kelleher and Tierney, "Data Science" (MIT Press, 2018)

Important concepts

- A <u>datum</u> or a piece of information is an abstraction of a real-world entity (person, object, or event).
- The terms <u>variable</u>, <u>feature</u>, and <u>attribute</u> are often used interchangeably to denote an individual abstraction.
- Each <u>entity</u> is typically described by a number of attributes.
  - For example, a book might have the following attributes: author, title, topic, genre, publisher, price, date published, word count, number of chapters, number of pages, edition, ISBN, etc.
- A <u>dataset</u> consists of the data relating to a collection of entities, with each entity described in terms of a set of attributes.
  - In its most basic form, a data set is organized in an n-by-m data matrix, sometimes called the *analytics record*, where <u>n</u> is the number of entities (rows) and <u>m</u> is the number of attributes (columns).

Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)

Example: analytics record for a data set of classic books

ID	Title	Author	Year	Cover	Edition	Price
1	*Emma*	Austen	1815	Paperback	20th	$5.75
2	*Dracula*	Stoker	1897	Hardback	15th	$12.00
3	*Ivanhoe*	Scott	1820	Hardback	8th	$25.00
4	*Kidnapped*	Stevenson	1886	Paperback	11th	$5.00

- Each row in the table describes one book.
- The terms instance, example, entity, object, case, individual, and record are used in data science literature to refer to a row.
- So a dataset contains a set of instances, and each instance is described by a set of attributes.

Source: Kelleher and Tierney, "Data Science" (MIT Press, 2018)

Types of attributes

- **Numeric**: describe measurable quantities that are represented using integer or real values
  - **Interval scale**
    - measured on a scale with a fixed but arbitrary interval and arbitrary origin (e.g., temperature in C or F, time, date)
  - **Ratio scale**
    - Possess a true-zero origin (e.g., temperature in K, exam grades, height, weight)

Source: Kelleher and Tierney, "Data Science" (MIT Press, 2018)

Types of attributes

- **Nominal (Categorical)**: take values from a finite set.
  - Examples: marital status [single, married, divorced] and beer type [ale, pale ale, porter, stout, etc.].

  - A **binary attribute** is a special case of a nominal attribute where the set of possible values is restricted to just two values.
    - Example: the binary attribute “spam,” which describes whether an email is spam (true) or not spam (false)

Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)

Types of attributes

- **Ordinal**: similar to nominal attributes, with the difference that *it is possible to apply a rank order* over the categories of ordinal attributes.
  - Example: an attribute describing the response to a survey question might take values from the domain "strongly dislike, dislike, neutral, like, and strongly like."

Source: Kelleher and Tierney, "Data Science" (MIT Press, 2018)

			Interval-scale	Categorical	Ordinal	Ratio-scale
ID	**Title**	**Author**	**Year**	**Cover**	**Edition**	**Price**
1	*Emma*	Austen	1815	Paperback	20th	$5.75
2	*Dracula*	Stoker	1897	Hardback	15th	$12.00
3	*Ivanhoe*	Scott	1820	Hardback	8th	$25.00
4	*Kidnapped*	Stevenson	1886	Paperback	11th	$5.00

Source: Kelleher and Tierney, "Data Science" (MIT Press, 2018)

The data type of an attribute (numeric, ordinal, nominal) affects the methods we can use to analyze and understand the data.

Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)

Perspectives on data: structured or not?

- **Structured data**:
  - can be represented as a **table**, and every instance in the table has the same structure (i.e., set of attributes).
  - can be easily stored, organized, searched, reordered, and merged with other structured data.

- **It is relatively easy to apply data science to structured data** because, by definition, it is already in a format that is suitable for integration into an analytics record.

Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)

Perspectives on data: structured or not?

- **Unstructured data**:
  - each instance in the data set may have its own internal structure, and this structure is not necessarily the same in every instance.
  - ***much more common than structured data***.
- Examples of unstructured data:
  - collections of human text (emails, tweets, text messages, posts, novels, etc.)
  - collections of sound, image, music, video, and multimedia files.
- The variation in the structure between the different elements means that it is difficult to analyze unstructured data in its raw form.
  - We can often extract structured data from unstructured data using AI techniques (e.g., NLP and ML), digital signal processing, and computer vision.
    - *Expensive and time-consuming*

Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)

Two forms of raw data

1. **Captured data** are collected through a direct measurement or observation that is designed to gather the data.
   - Example: responses from a survey whose primary purpose is to gather specific data on a particular topic of interest.

2. **Exhaust data** are a by-product of a process whose primary purpose is something other than data capture (including *metadata*)
   - Example: byproducts of interactions with different primary purposes, such as social media technologies
     - Goal: to enable users to connect with other people.
     - However, for every image shared, blog posted, tweet retweeted, or post liked, a range of exhaust data is generated: who shared, who viewed, what device was used, what time of day, which device was used, how many people viewed/liked/retweeted, etc.

Source: Kelleher and Tierney, "Data Science" (MIT Press, 2018)

Derived attributes

It is frequently the case that the real value of a data science project is the identification of one or more important derived attributes that provide insight into a problem.

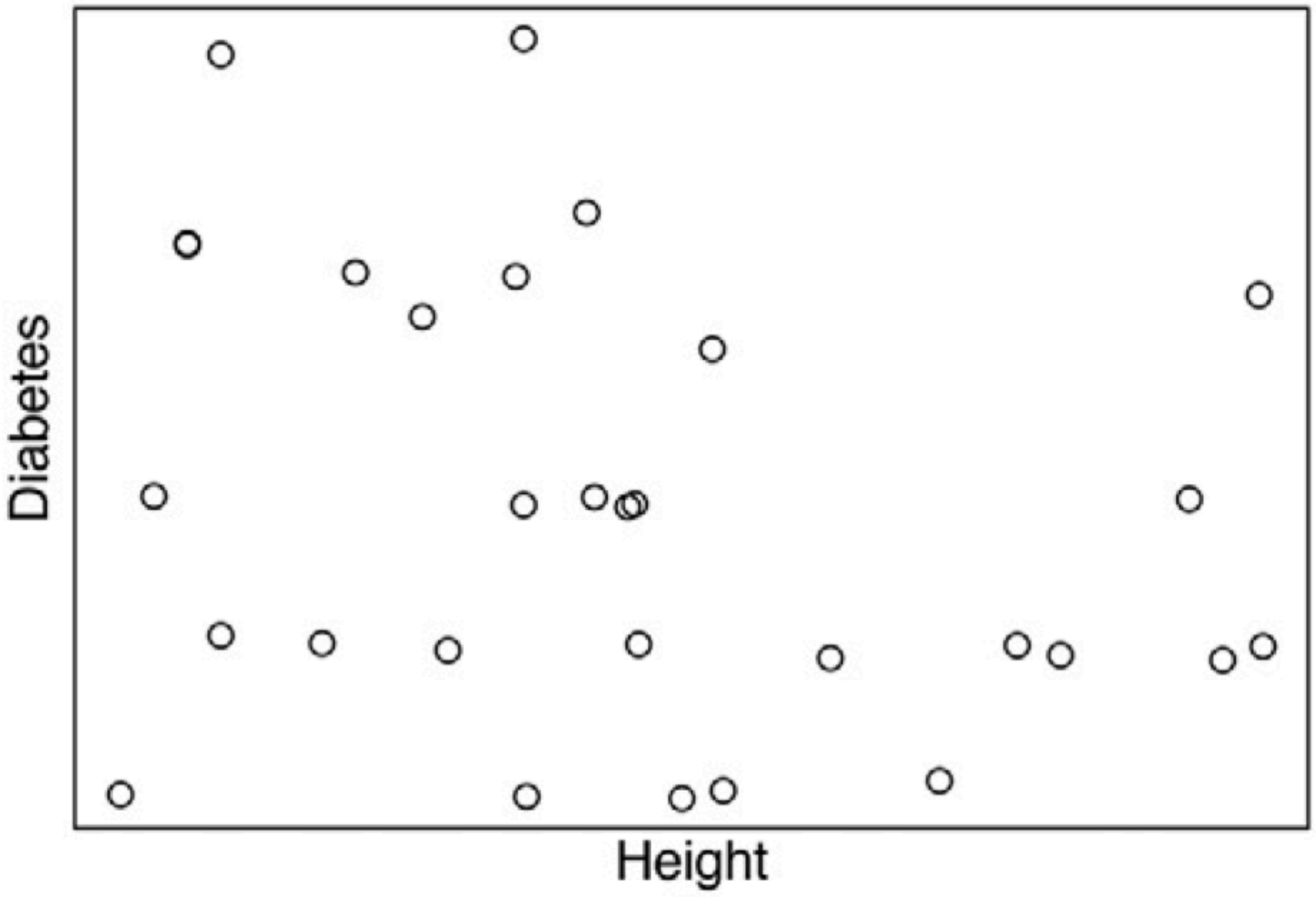
Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)

Derived attributes

- Imagine we are trying to study the **causes of Type 2 diabetes in white American adult males**
- We are interested in identifying if any of the attributes have a strong correlation with the target attribute describing a person's likelihood of developing diabetes.
- We could begin by examining the raw attributes of individuals, such as their height and weight, but the results would not be encouraging.

Source: Kelleher and Tierney, "Data Science" (MIT Press, 2018)

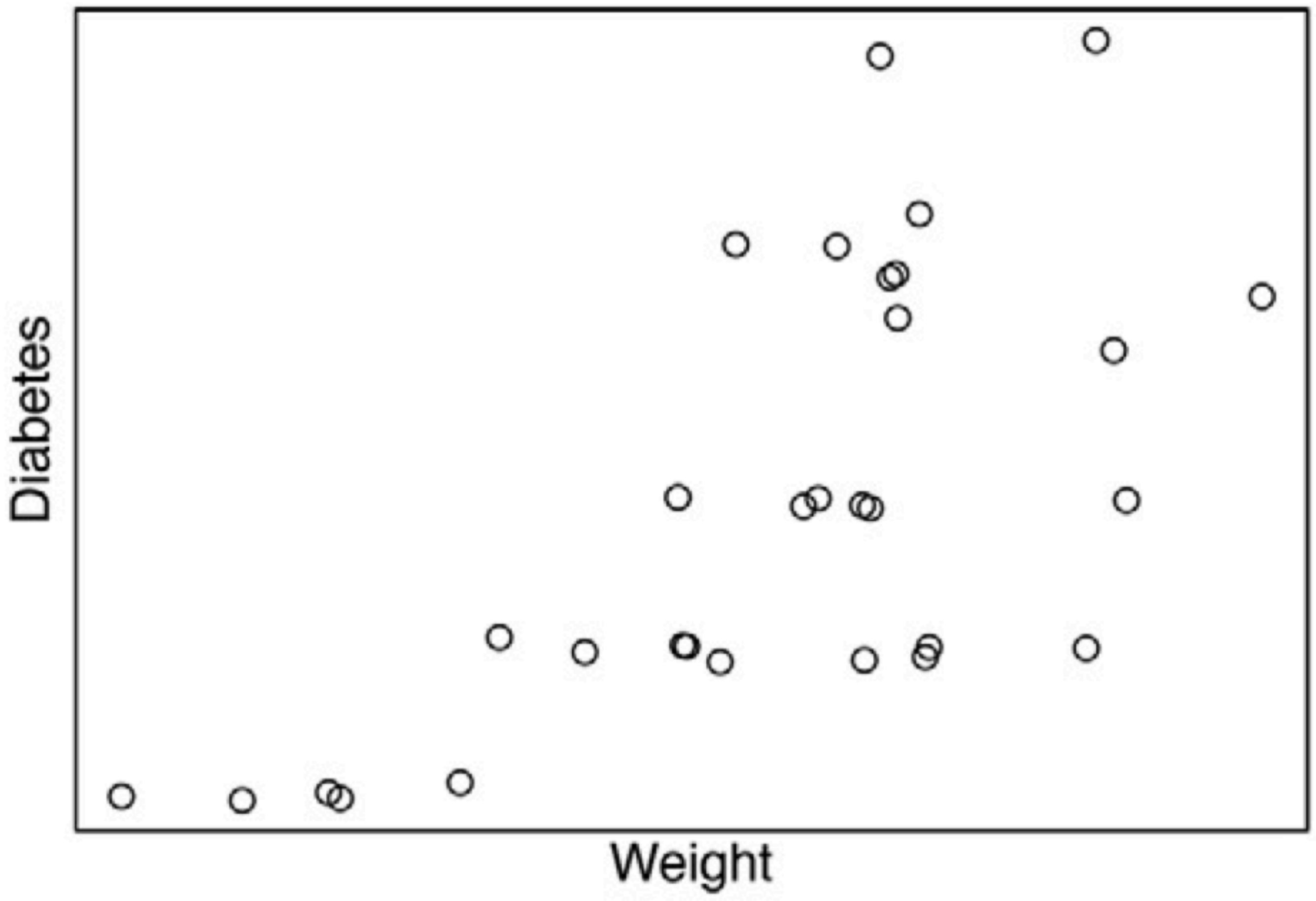
Example: Type 2 diabetes



Pearson coefficient: $r = -0.277$

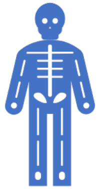
Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)

Example: Type 2 diabetes



Pearson coefficient: r = 0.655

Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)



Example: Type 2 diabetes

- After studying the problem for some time we might end up designing a more informative derived attribute such as the **Body Mass Index (BMI)**.

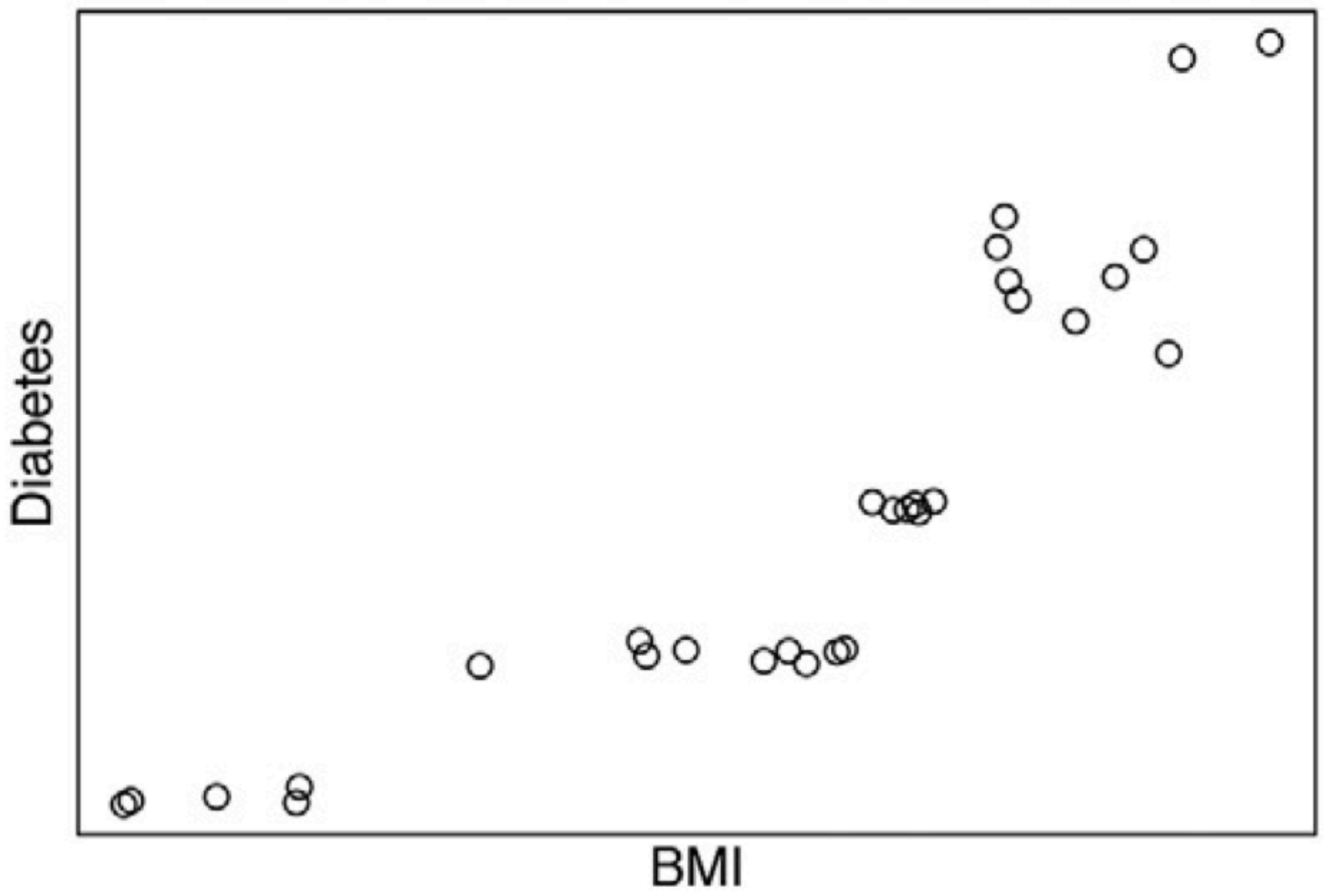
  - BMI is the ratio of weight (in kilograms) divided by height (in meters) squared.
    - Invented in the 19^{th} century by a Belgian mathematician, Adolphe Quetelet

  - The ratio of weight and height is used because BMI is designed to have a similar value for people who are in the same category (*underweight*, *normal weight*, *overweight*, or *obese*) irrespective of their height.

  - We know that weight and height are positively correlated (generally, the taller someone is, the heavier he is), so by dividing weight by height, we control for the effect of height on weight.

Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)

Example: Type 2 diabetes



Pearson coefficient: r = 0.877

Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)

The key to success is getting the right data and finding the right attributes.

Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)

Feature engineering

- The process of using domain knowledge of the data to create features that make machine learning algorithms work.
- Feature engineering is fundamental to the application of machine learning, and is both difficult and expensive.
  - The need for manual feature engineering can be obviated by automated feature learning.

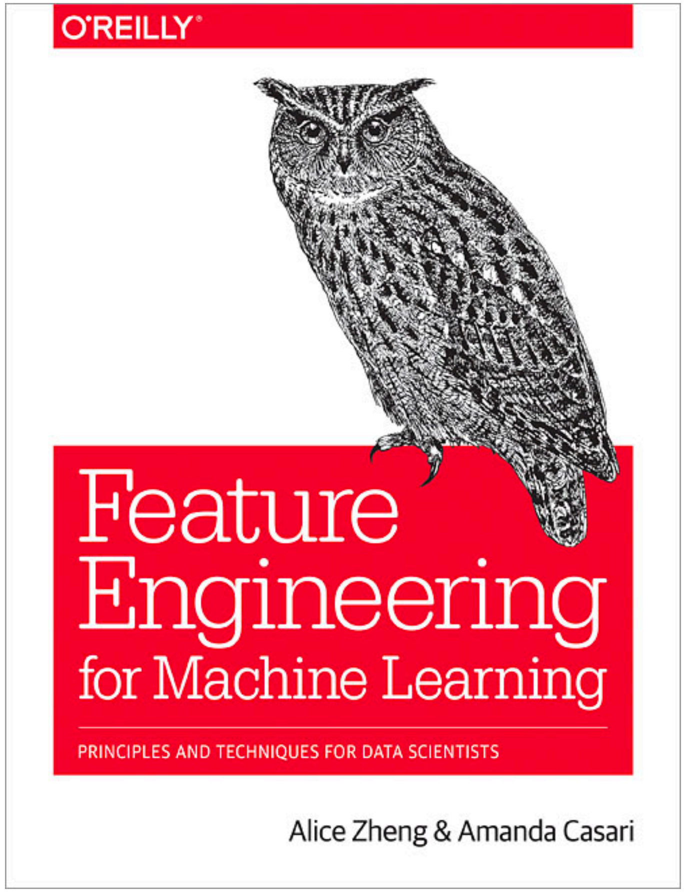
Coming up with features is difficult, time-consuming, requires expert knowledge.

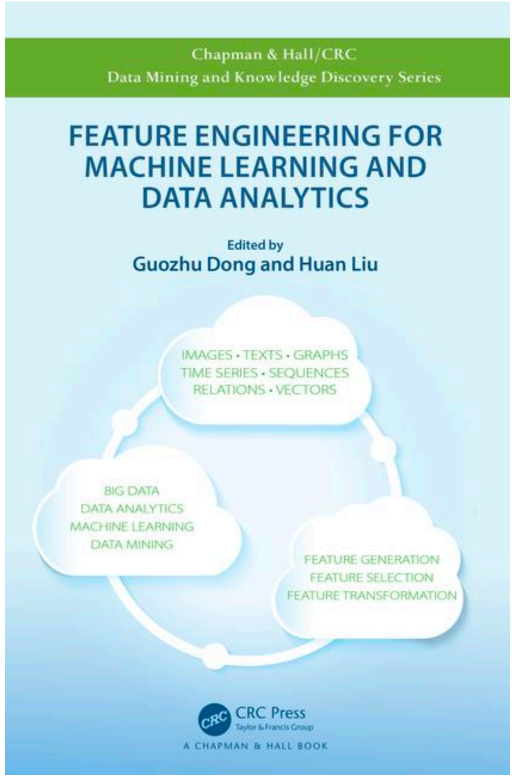
"Applied machine learning" is basically feature engineering.

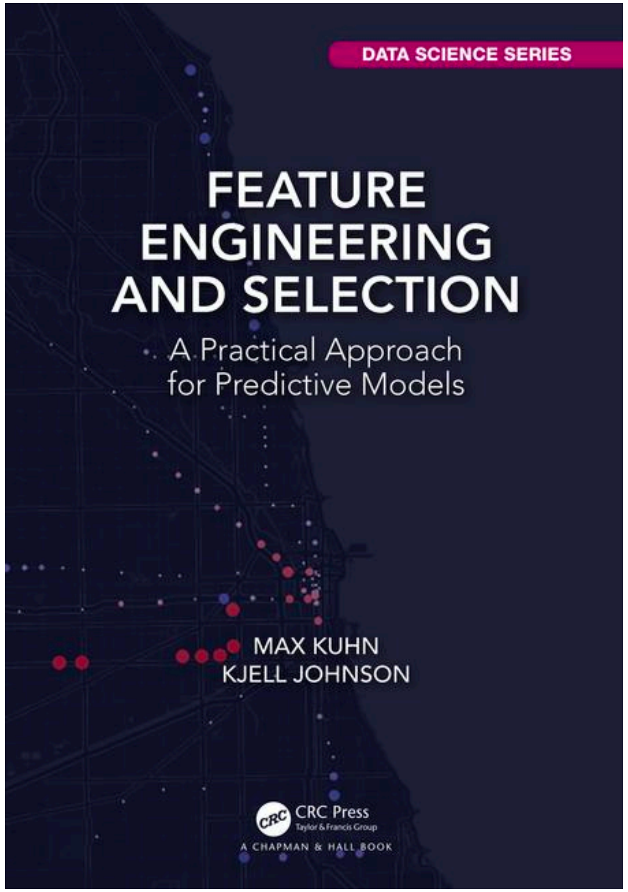
— Andrew Ng

Source: Wikipedia

Feature engineering







https://bookdown.org/max/FES/

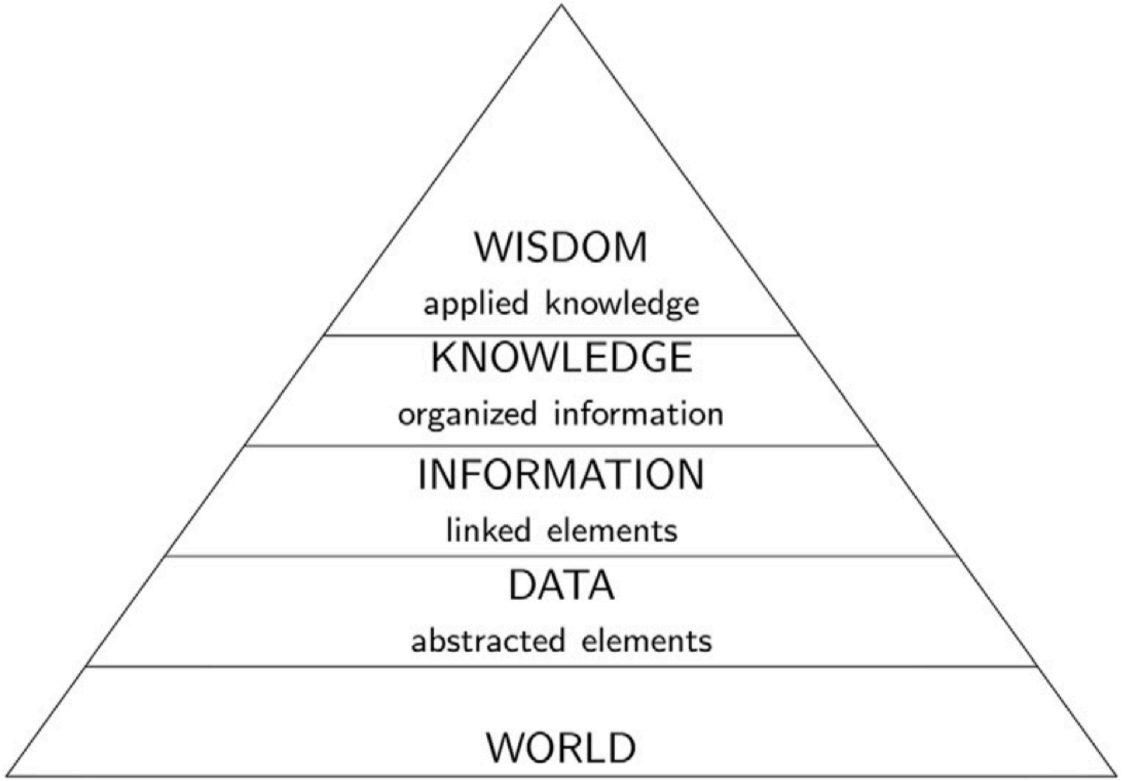
From data to insight

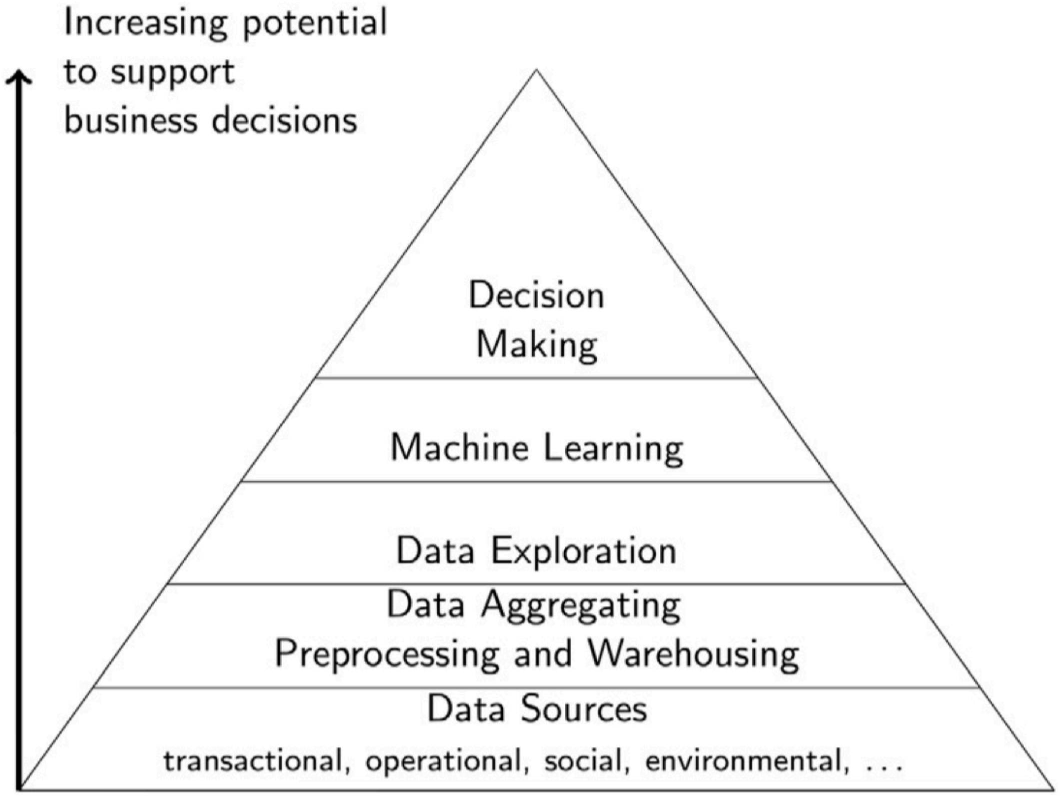
- Data are generated through a process of abstraction, so **any data are the result of human decisions and choices**.
  - For every abstraction, somebody (or some set of people) will have made choices with regard to what to abstract from and what categories or measurements to use in the abstracted representation.
  - The implication is that **data are never an objective description of the world**. They are instead always partial and biased.

- The data we use for data science are not a perfect representation of the real-world entities and processes we are trying to understand, but if we are careful in how we design and gather the data that we use, then the results of our analysis will provide useful insights into our real-world problems.

Source: Kelleher and Tierney, "Data Science" (MIT Press, 2018)

The DIKW and Data Science pyramids

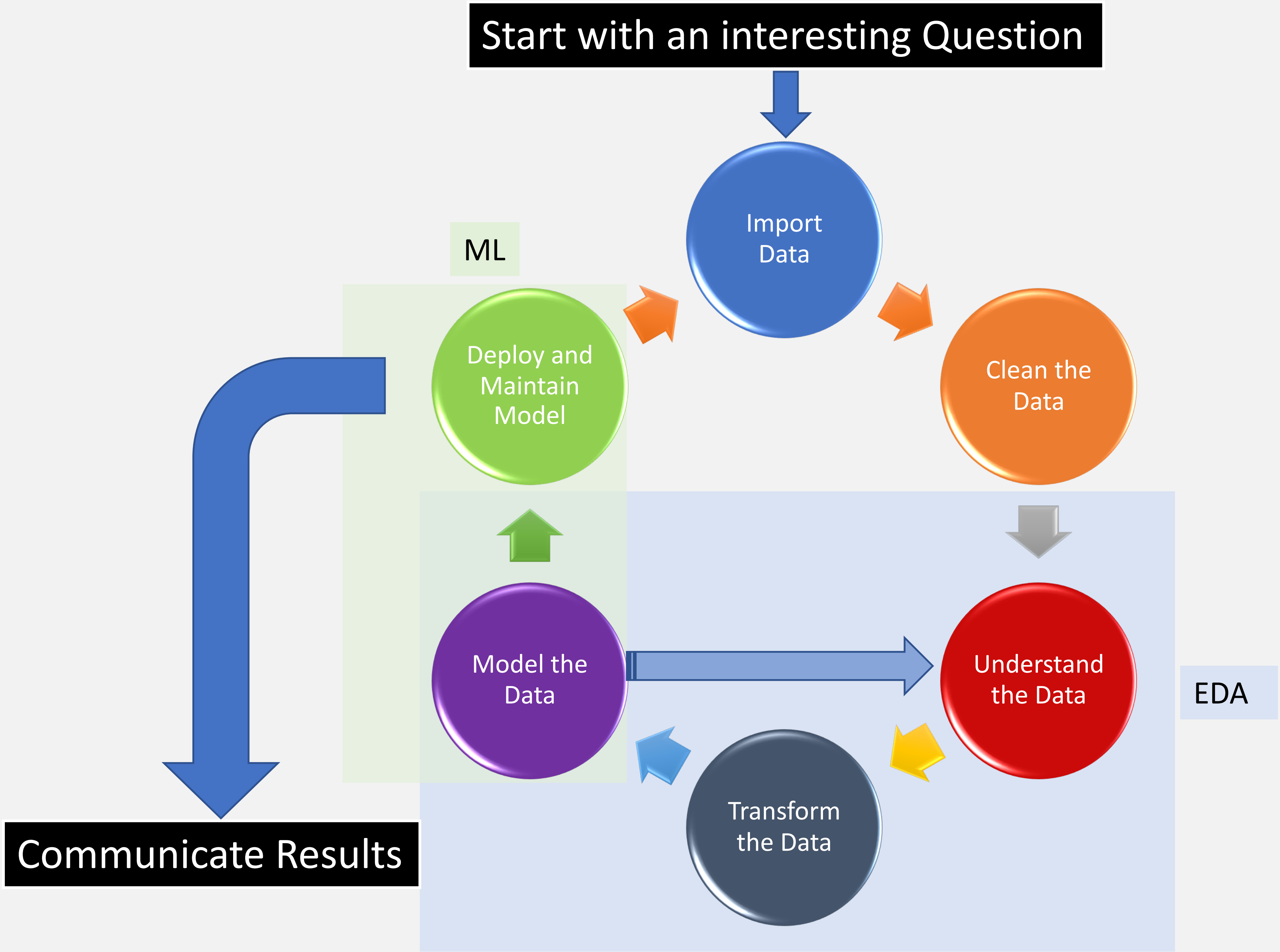




Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)

Part 3:

The Data Science workflow / ecosystem

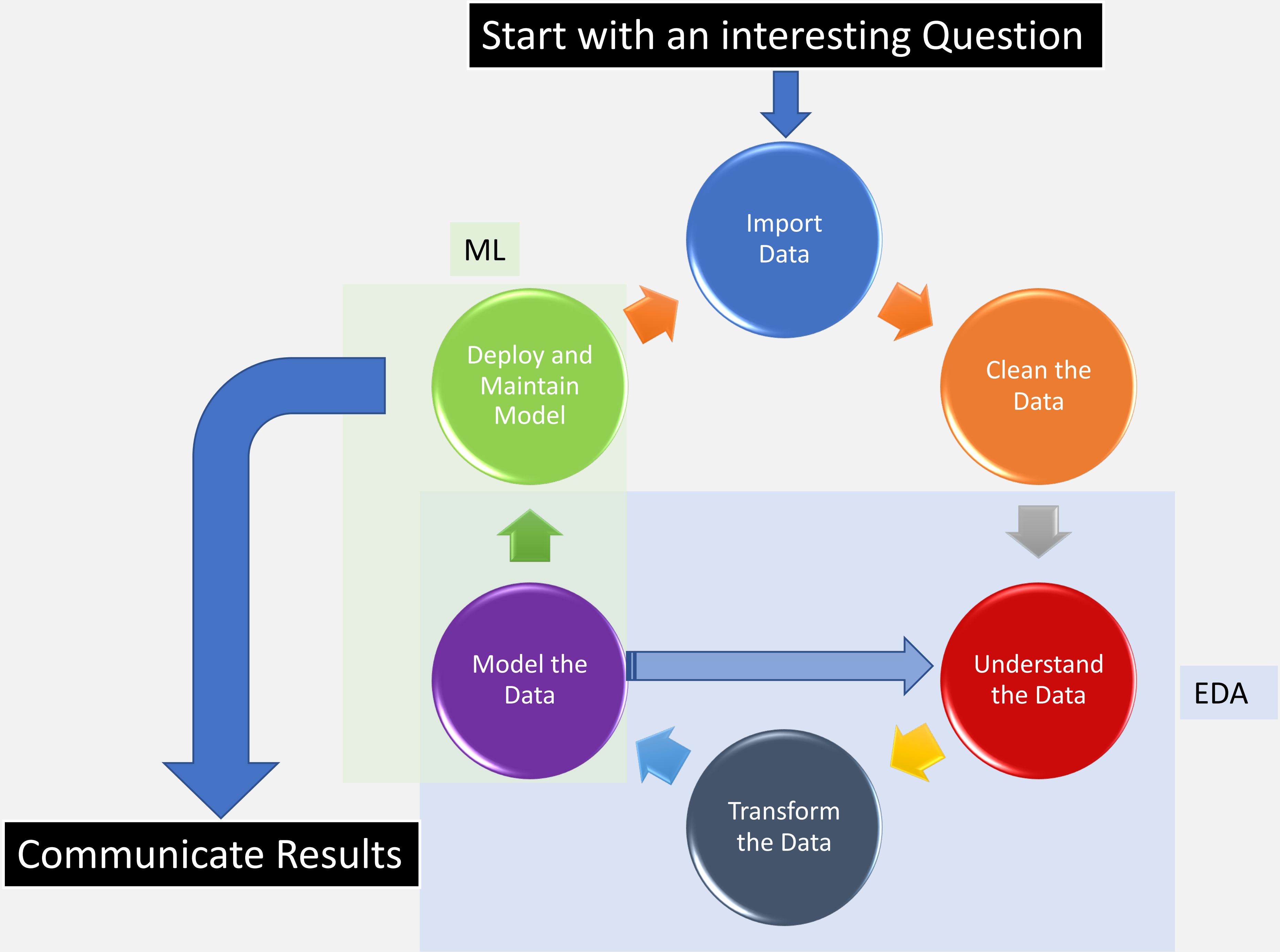

Start with an interesting Question
Import Data
Clean the Data
Understand the Data
Transform the Data
Model the Data
Deploy and Maintain Model
ML
EDA
Communicate Results

Key takeaways

- Data preparation / wrangling is the most time-consuming task
- Spending time doing meaningful EDA pays off
- The cycles never end…
  - … but you must stop them at some point
- Don’t lose sight of the original question
- Keep the stakeholders in mind…
  - … and adjust how you communicate your findings
- Deploying and maintaining the model is a major undertaking

Part 4:

Exploratory Data Analysis (EDA)


Start with an interesting Question
Import Data
Clean the Data
Understand the Data
Transform the Data
Model the Data
Deploy and Maintain Model
ML
EDA
Communicate Results

Key questions

- What are your **goals**?
- Which **techniques** should you use?
- Which **tools** can you use?

Key questions (and answers)

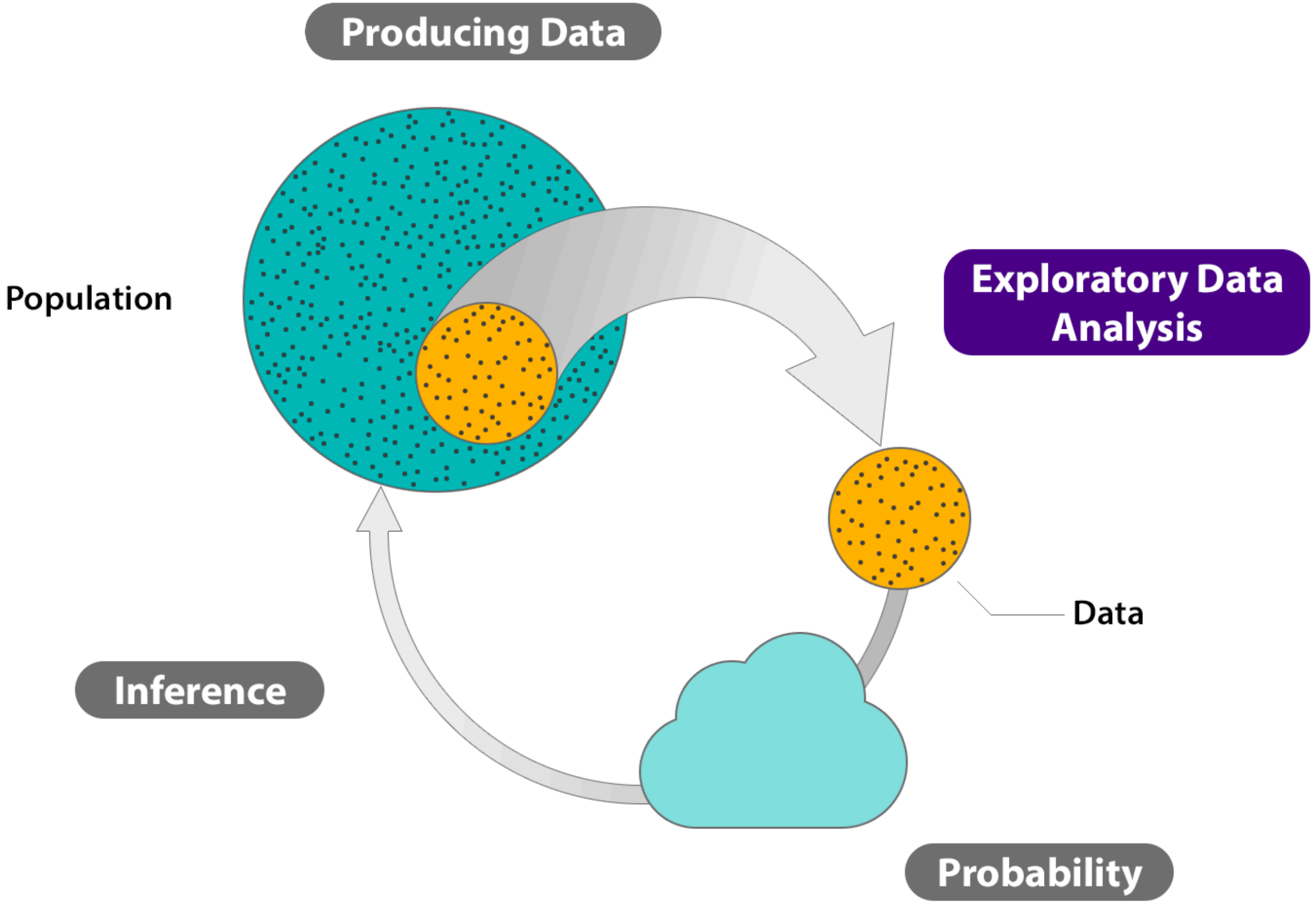
- What are your **goals**?
  - To understand and trust my data.
- Which **techniques** should you use?
  - Summary statistics and visualization.
- Which **tools** can you use?
  - Numpy, Pandas, Matplotlib (the “Python Data Science stack”)

Start with the right question

> Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.
>
> —John Tukey

Source: https://www.goodreads.com/quotes/7284366-far-better-an-approximate-answer-to-the-right-question-which

Exploratory Data Analysis (EDA) in context



Source: https://timewithai.com/2019/07/20/the-big-picture-inference/

Summary statistics and visualization tools

- Summary statistics can only go so far as providing a general feel for the distribution of the data
- Visualization helps tremendously!

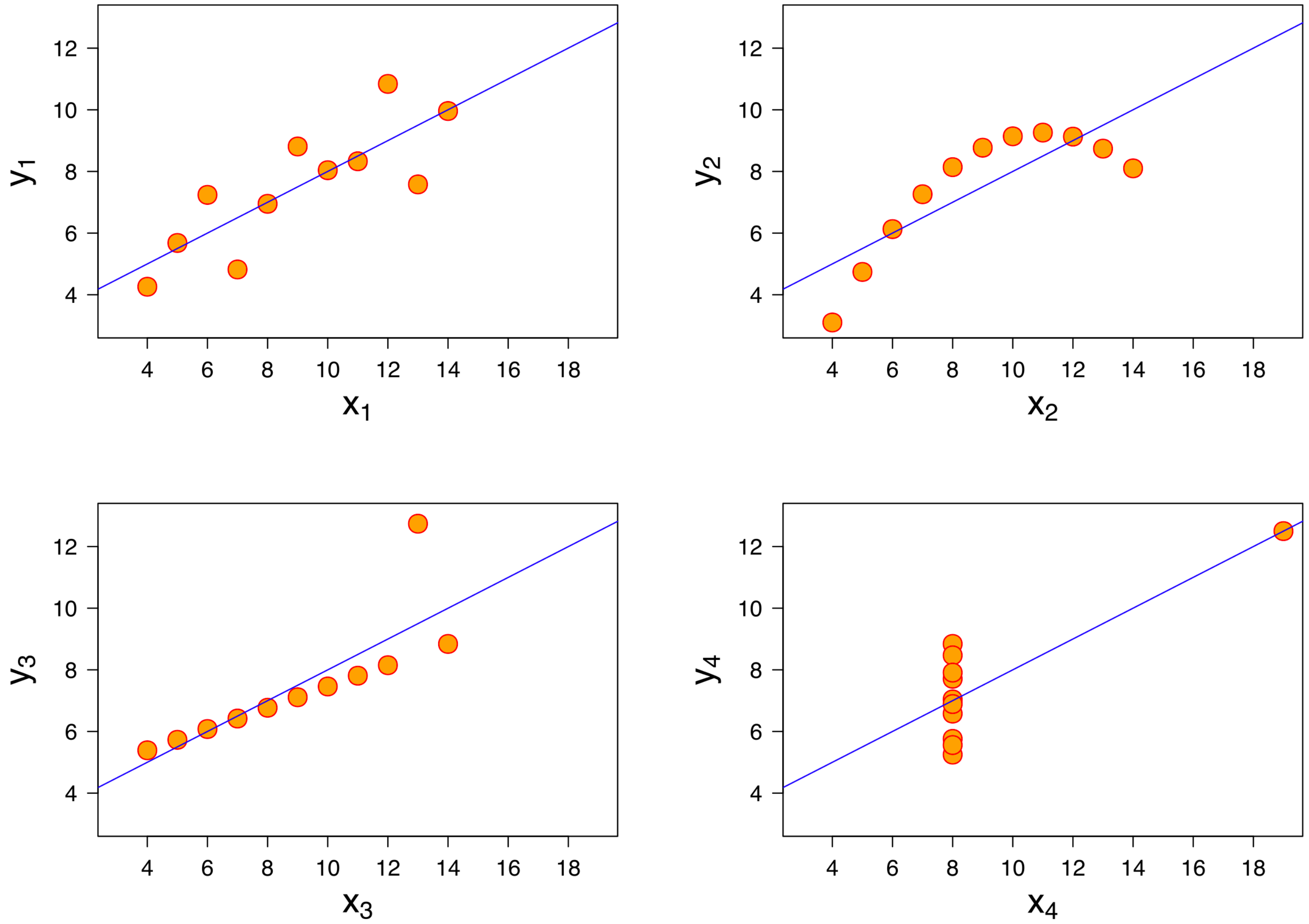
"The world cannot be understood without numbers.

And it cannot be understood with numbers alone."

Source: Hans Rosling et al., Factfulness: Ten Reasons We're Wrong about the World—And Why Things Are Better Than You Think (New York: Flatiron Books, 2018).

Anscombe's quartet

Constructed in 1973 by statistician **Francis Anscombe** to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.



Source: Wikipedia

EDA: techniques and recommendations

- **Single variable explorations**: start by examining one variable at a time, finding out what the variables mean, looking at distributions of the values, and choosing appropriate summary statistics.

- **Pair-wise explorations**: to identify possible relationships between variables, look at tables and scatter plots, and compute correlations and linear fits.

- **Multivariate analysis**: if there are apparent relationships between variables, use multiple regression to add control variables and investigate more complex relationships.

Source: Allen B. Downey, "Think Stats" 2nd ed. – O'Reilly

EDA: techniques and recommendations

- **Estimation and hypothesis testing**: When reporting statistical results, it is important to answer three questions:
  - How big is the effect?
  - How much variability should we expect if we run the same measurement again?
  - Is it possible that the apparent effect is due to chance?

- **Visualization**: During exploration, visualization is an important tool for finding possible relationships and effects.
  - Then, if an apparent effect holds up to scrutiny, visualization is an effective way to communicate results.

Source: Allen B. Downey, "Think Stats" 2nd ed. – O'Reilly

python™

IP[y]:
IPython

jupyter

CO

NumPy

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

matplotlib

Seaborn

scikit learn

ANACONDA®

pip

Jupyter notebooks (examples)

tinyurl.com/icmla2019

Hands on!

Christian Garbin

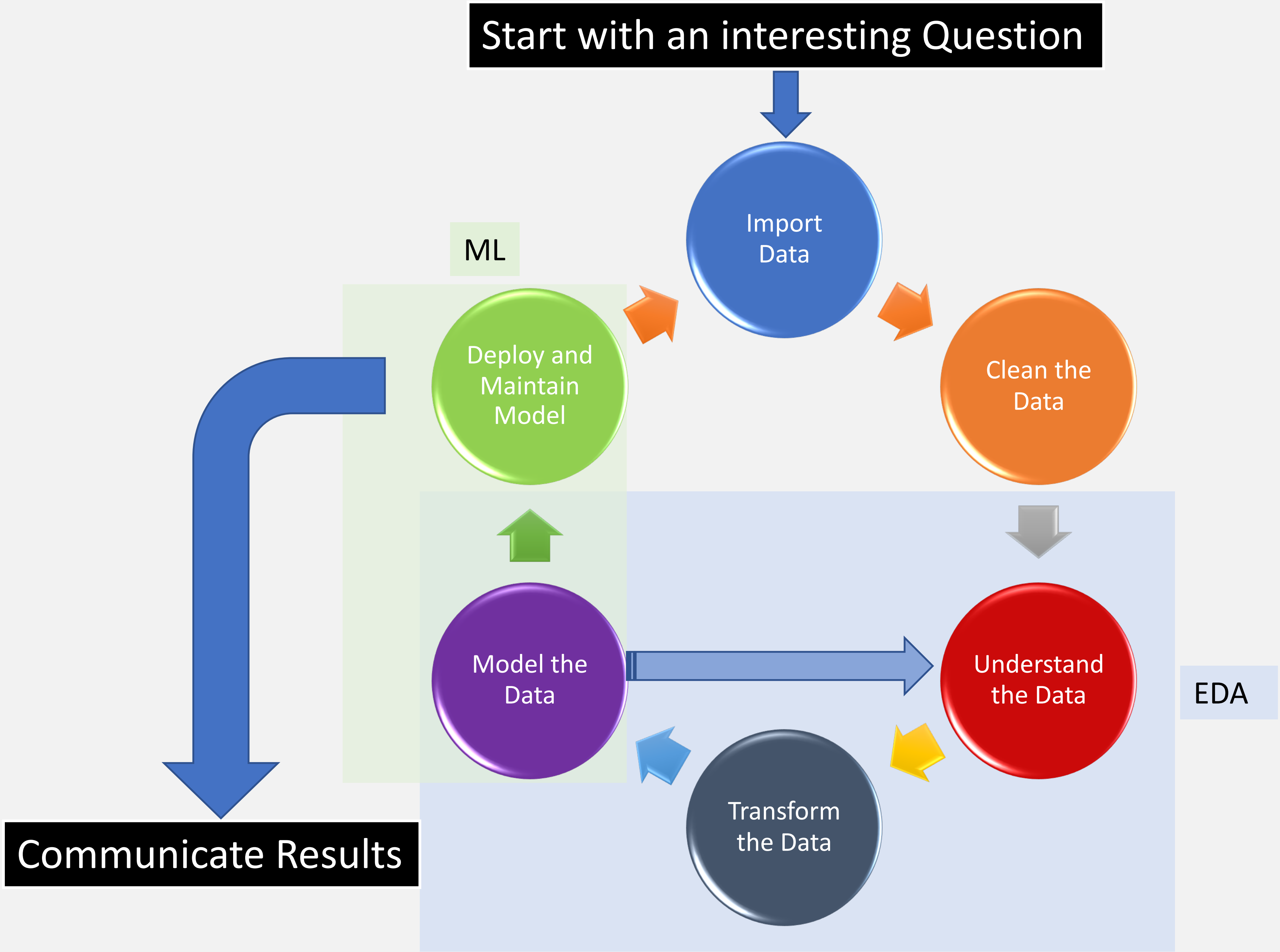
Senior Architect and
Distinguished Expert at
Unify Inc., an Atos company
(Boca Raton, FL)

Example 1: Exploratory Data Analysis

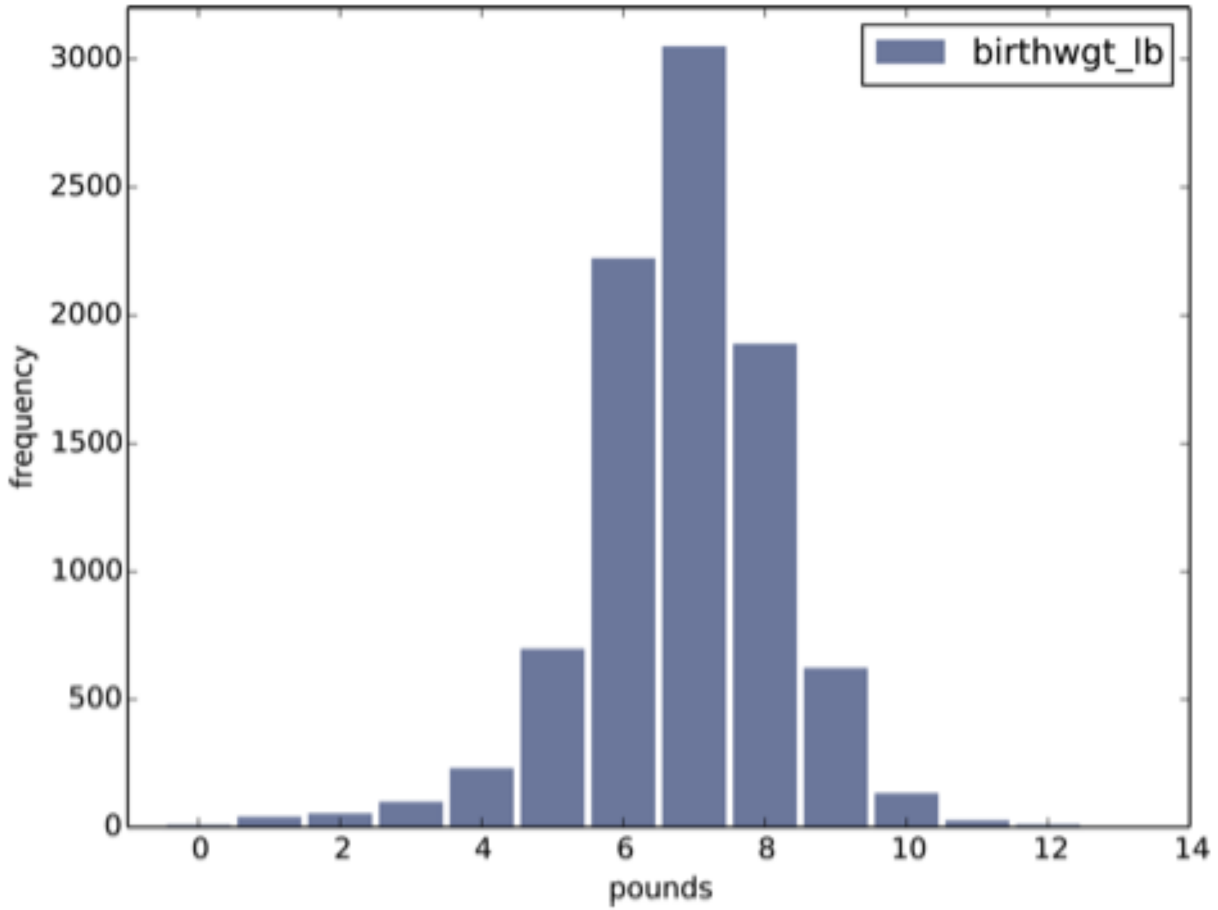
tinyurl.com/icmla2019

Part 5:

Statistics and Data Science


Start with an interesting Question
Import Data
Clean the Data
Understand the Data
Transform the Data
Model the Data
Deploy and Maintain Model
ML
EDA
Communicate Results

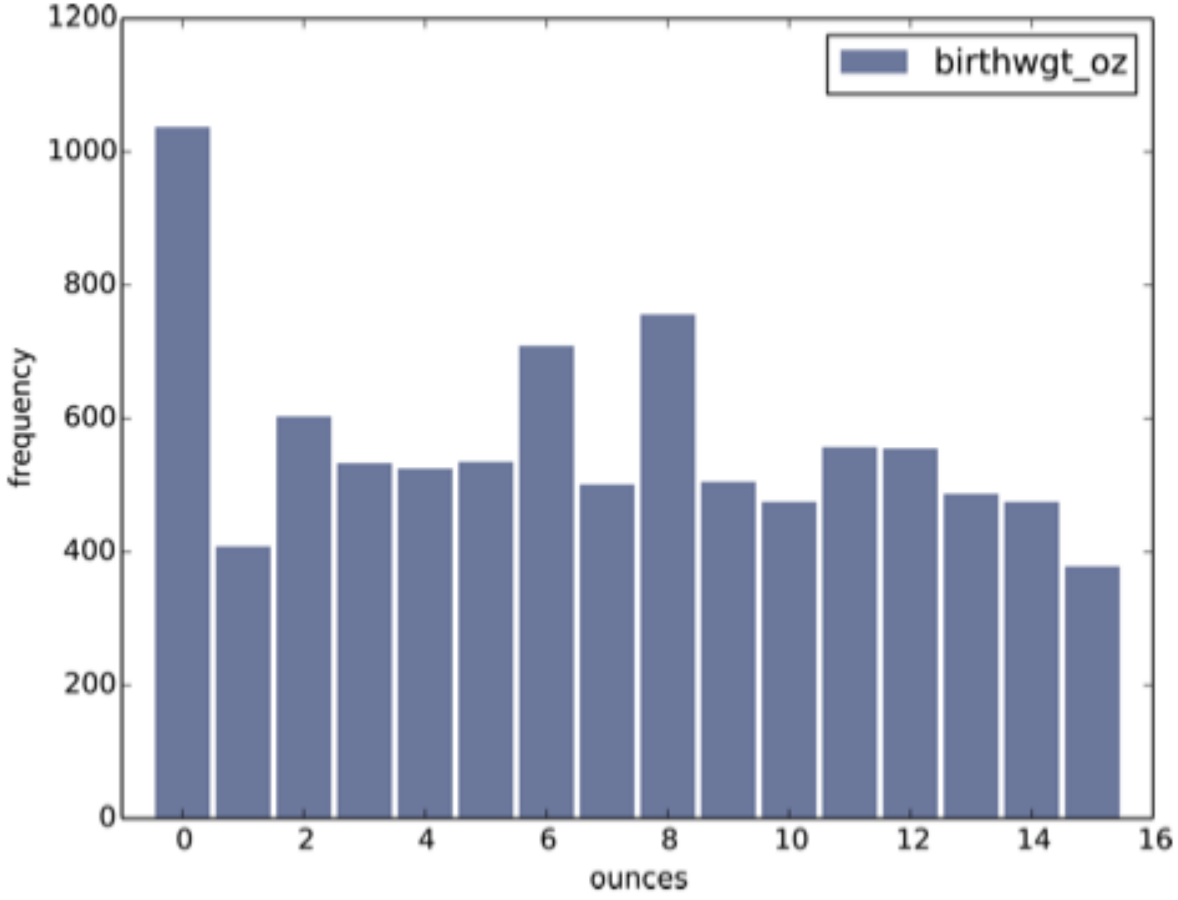
Histograms



National Survey of Family Growth (NSFG) Cycle 6 (Jan 2002 - Mar 2003)
U.S. Centers for Disease Control and Prevention (CDC)

Source: Allen B. Downey, "Think Stats" 2nd ed. – O'Reilly

Histograms

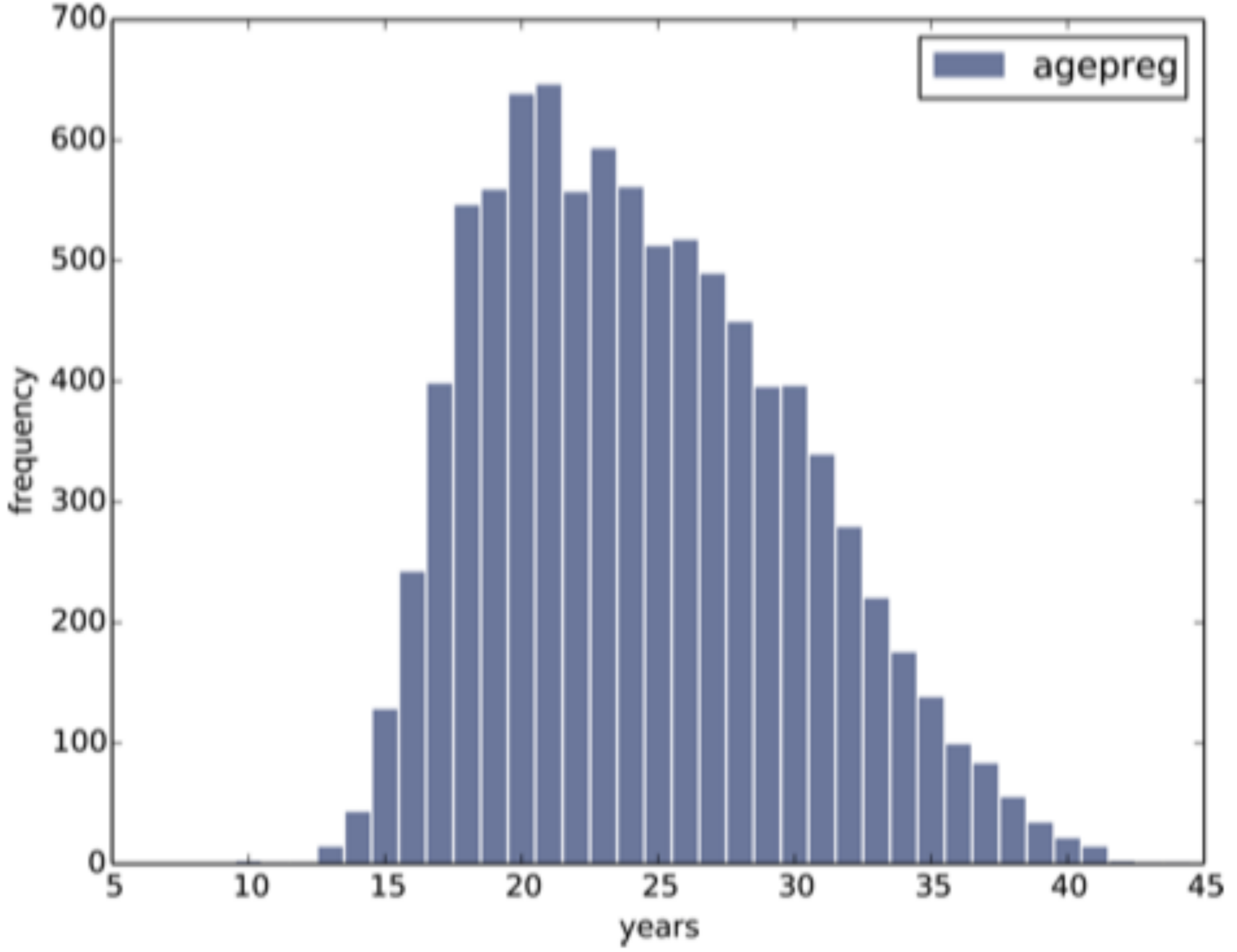


National Survey of Family Growth (NSFG) Cycle 6 (Jan 2002 - Mar 2003)
U.S. Centers for Disease Control and Prevention (CDC)

Source: Allen B. Downey, “Think Stats” 2nd ed. – O’Reilly

Histograms

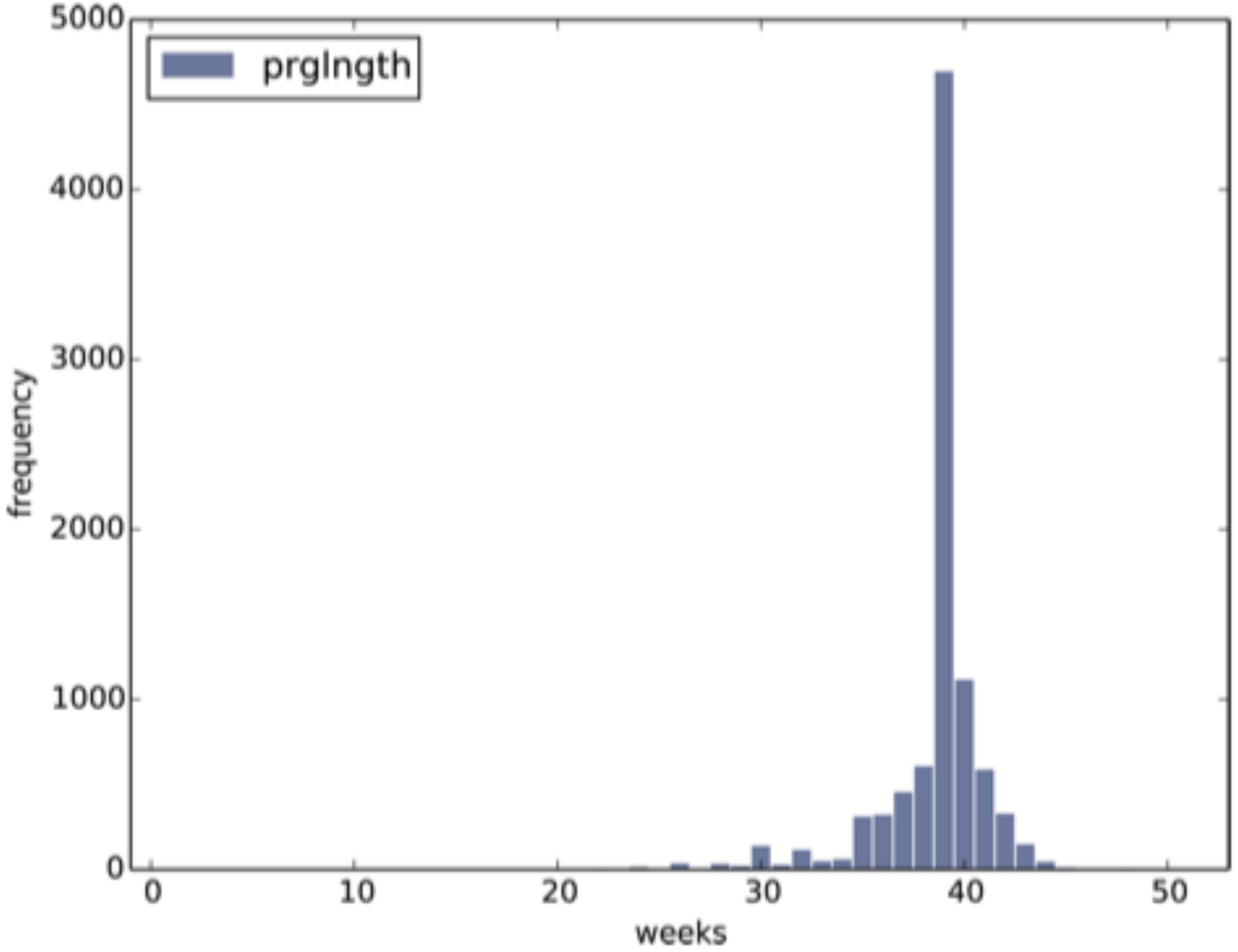
National Survey of Family Growth (NSFG) Cycle 6 (Jan 2002 - Mar 2003)
U.S. Centers for Disease Control and Prevention (CDC)



Source: Allen B. Downey, "Think Stats" 2nd ed. – O'Reilly

Histograms

National Survey of Family Growth (NSFG) Cycle 6 (Jan 2002 - Mar 2003)
U.S. Centers for Disease Control and Prevention (CDC)



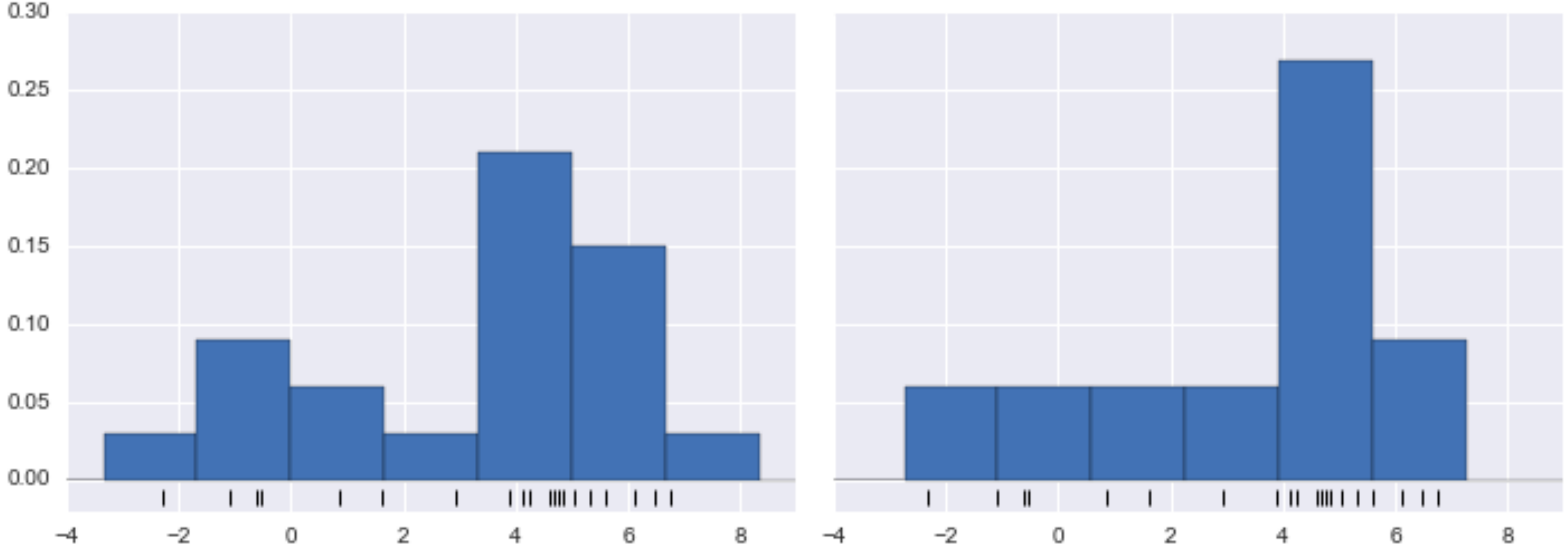
Source: Allen B. Downey, “Think Stats” 2nd ed. – O’Reilly

Histograms help provide insights

- **central tendency**
  - Do the values tend to cluster around a particular point?
- **modes**
  - Is there more than one cluster?
- **spread**
  - How much variability is there in the values?
- **tails**
  - How quickly do the probabilities drop off as we move away from the modes?
- **outliers**
  - Are there extreme values far from the modes?

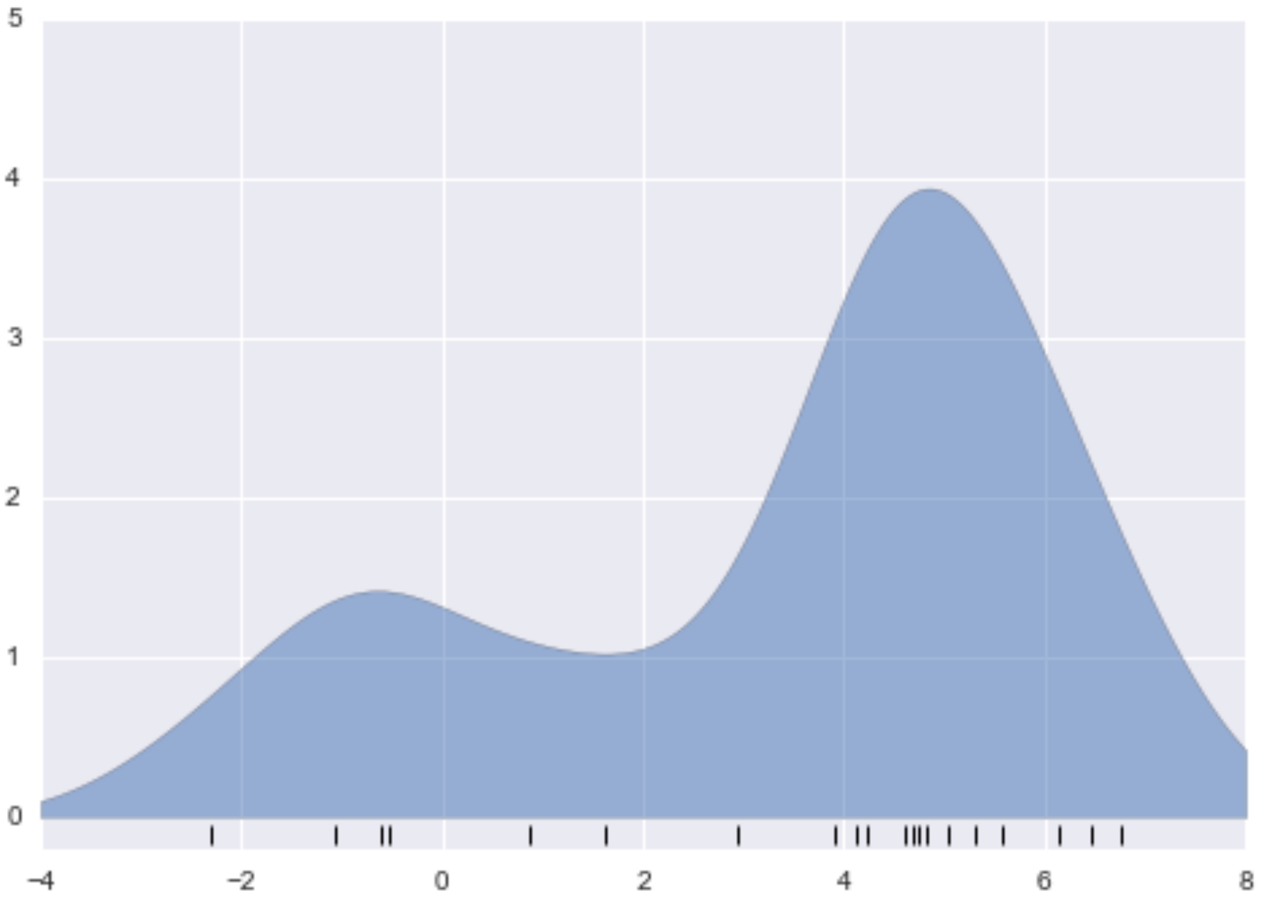
Source: Allen B. Downey, “Think Stats” 2nd ed. – O’Reilly

The binning problem



Source: https://jakevdp.github.io/PythonDataScienceHandbook/05.13-kernel-density-estimation.html

The binning problem: a solution (KDE: Kernel Density Estimation)



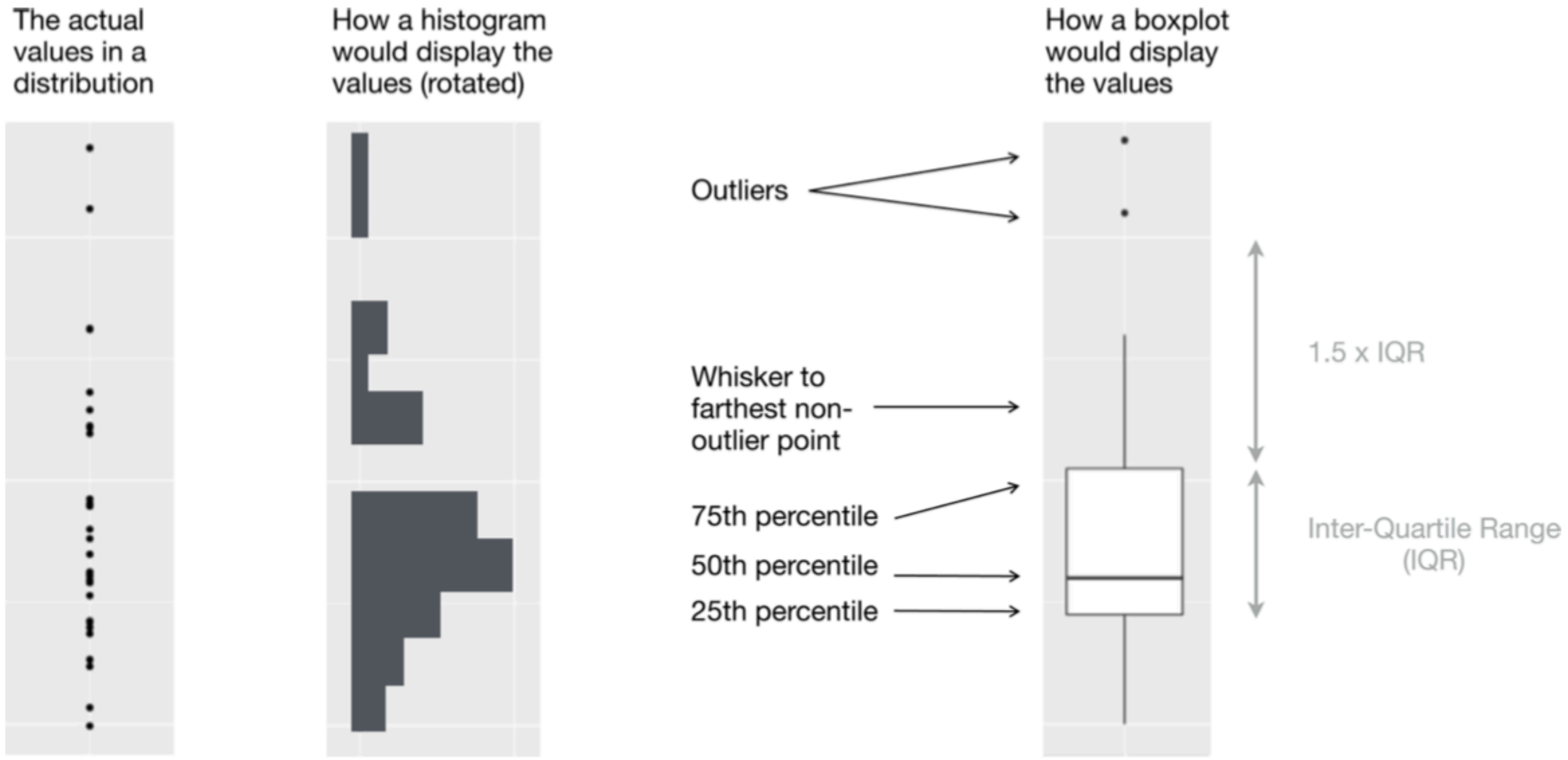
Source: https://jakevdp.github.io/PythonDataScienceHandbook/05.13-kernel-density-estimation.html

Percentile-based statistics

- Median = 50th percentile
  - A measure of central tendency of the distribution

- Interquartile Range (IQR) = the difference between the 75th and 25th percentiles
  - A measure of the spread of the distribution

Box(-and-whisker) plots



Source: "R for Data Science" – Wickham & Grolemund (O'Reilly)

Hands on!

Christian Garbin

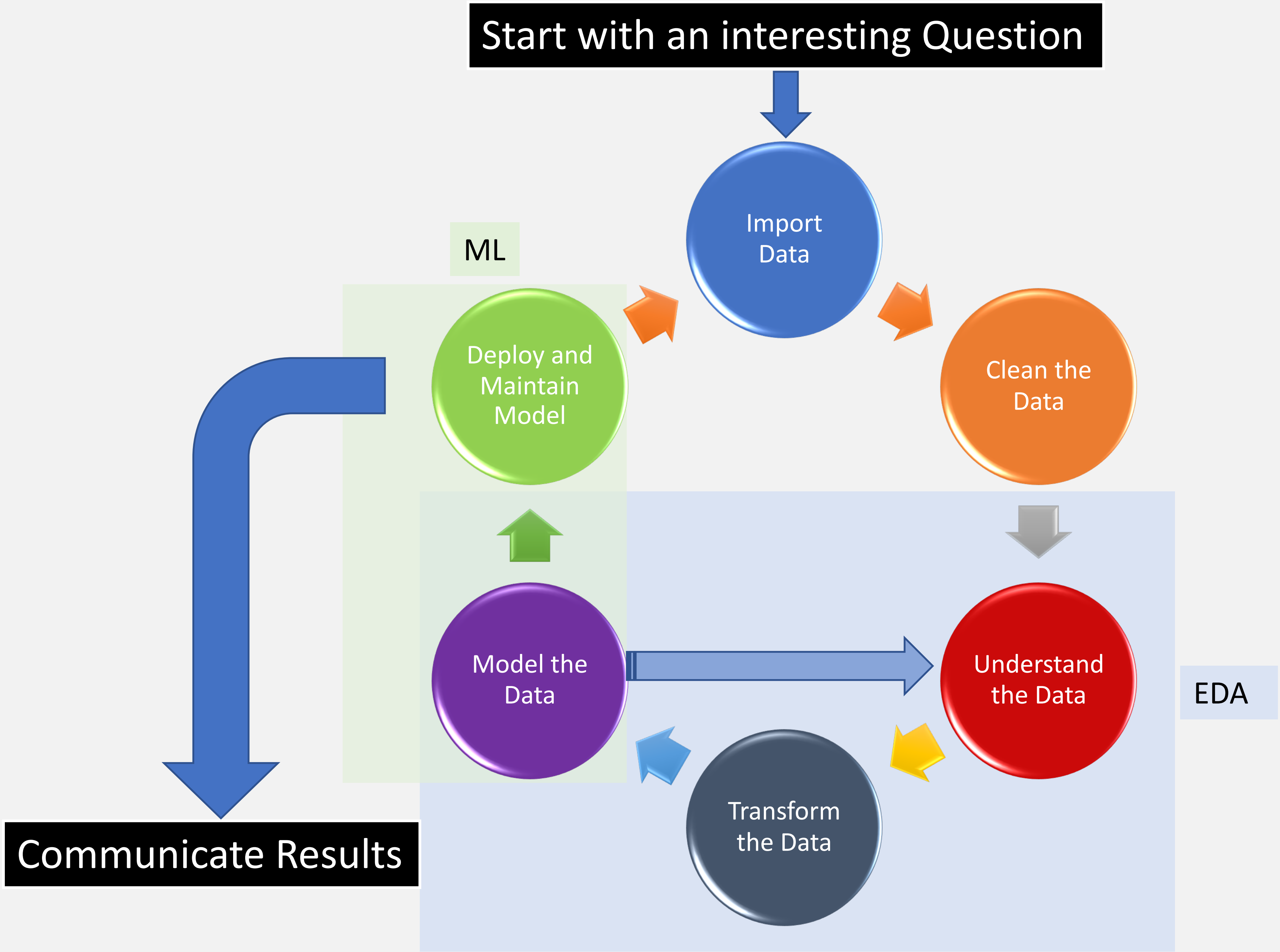
Senior Architect and
Distinguished Expert at
Unify Inc., an Atos company
(Boca Raton, FL)

Example 2: Statistics and Data Science

tinyurl.com/icmla2019

Part 6:

Using data to answer questions


Start with an interesting Question
Import Data
Clean the Data
Understand the Data
Transform the Data
Model the Data
Deploy and Maintain Model
ML
EDA
Communicate Results

How can we use data to…

… answer questions?

… confirm suspicions?

… dismiss misconceptions?

… test hypotheses?

Two main paths

Informal

- Slice-and-dice
- EDA
- Visual observations
- Simple calculations and comparisons

Formal

- Hypothesis testing
- Statistical significance

Correlation and covariance

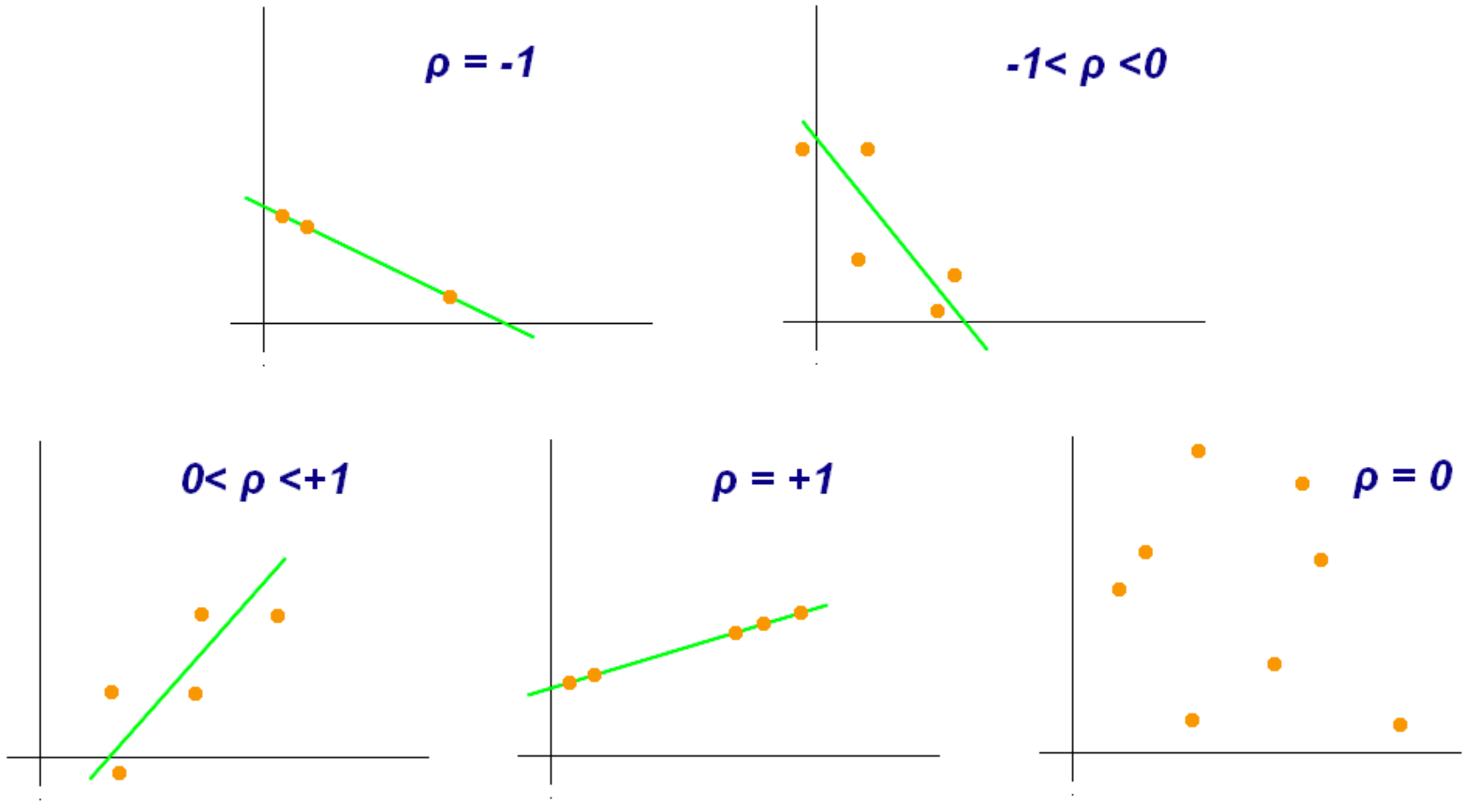
Correlation is a statistic intended to quantify the strength of the relationship between two variables.

Covariance is a measure of the tendency of two variables to vary together.

Source: Allen B. Downey, "Think Stats" 2nd ed. – O'Reilly

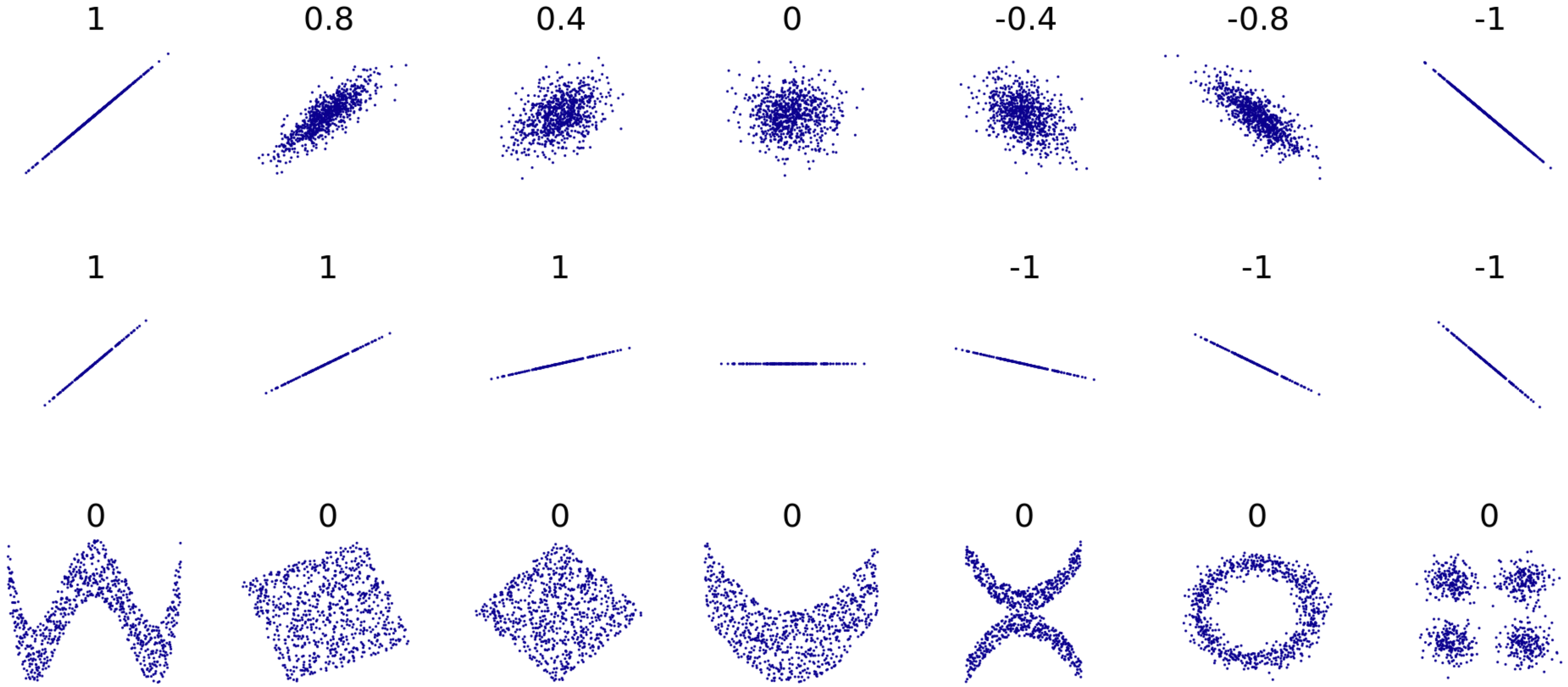
Pearson's correlation

- Pearson's correlation is always between -1 and +1 (including both).
  - If ρ is positive, we say that the correlation is positive, which means that when one variable is high, the other tends to be high.
  - If ρ is negative, the correlation is negative, so when one variable is high, the other is low.



Source: Wikipedia

Pearson’s correlation



Source: Wikipedia

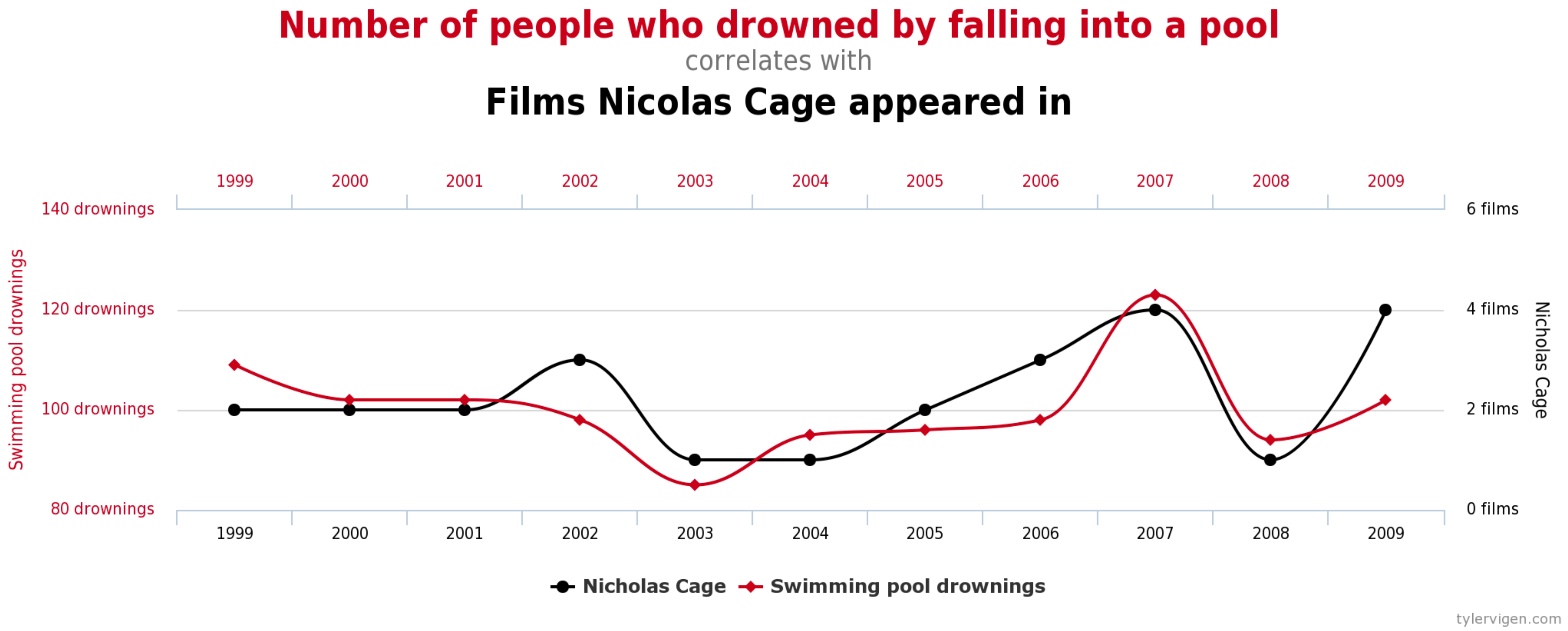
Correlation and causation

- If variables A and B are correlated, there are three possible explanations:
  - A causes B,
  - B causes A, or
  - some other set of factors causes both A and B.

- These explanations are called "causal relationships".

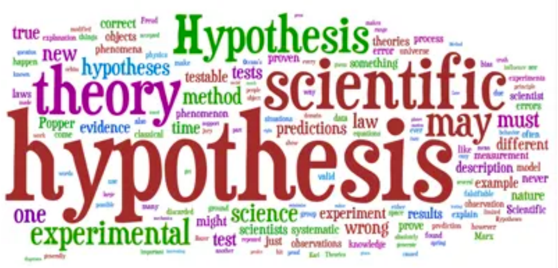
Source: Allen B. Downey, "Think Stats" 2nd ed. – O'Reilly

Correlation and causation



Source: https://www.tylervigen.com/spurious-correlations

Classical hypothesis testing



- The goal of classical hypothesis testing is to answer the question:

  “Given a sample and an apparent effect, what is the probability of seeing such an effect by chance?”

- The logic of the process is similar to a proof by contradiction.
  - To prove a mathematical statement, A, you assume temporarily that A is false.
  - If that assumption leads to a contradiction, you conclude that A must actually be true.

Source: Allen B. Downey, “Think Stats” 2^{nd} ed. – O’Reilly

Classical hypothesis testing: steps

1. Quantify the size of the apparent effect by choosing a **test statistic**.
2. Define a **null hypothesis**, which is a model of the system based on the assumption that the apparent effect is not real.
3. Compute a **p-value**, which is the probability of seeing the apparent effect if the null hypothesis is true.
4. **Interpret the result**.

   If the p-value is low, the effect is said to be **statistically significant**, which means that it is unlikely to have occurred by chance.

   In that case we infer that the effect is more likely to appear in the larger population.

Source: Allen B. Downey, "Think Stats" 2nd ed. – O'Reilly

Hands on!

Christian Garbin

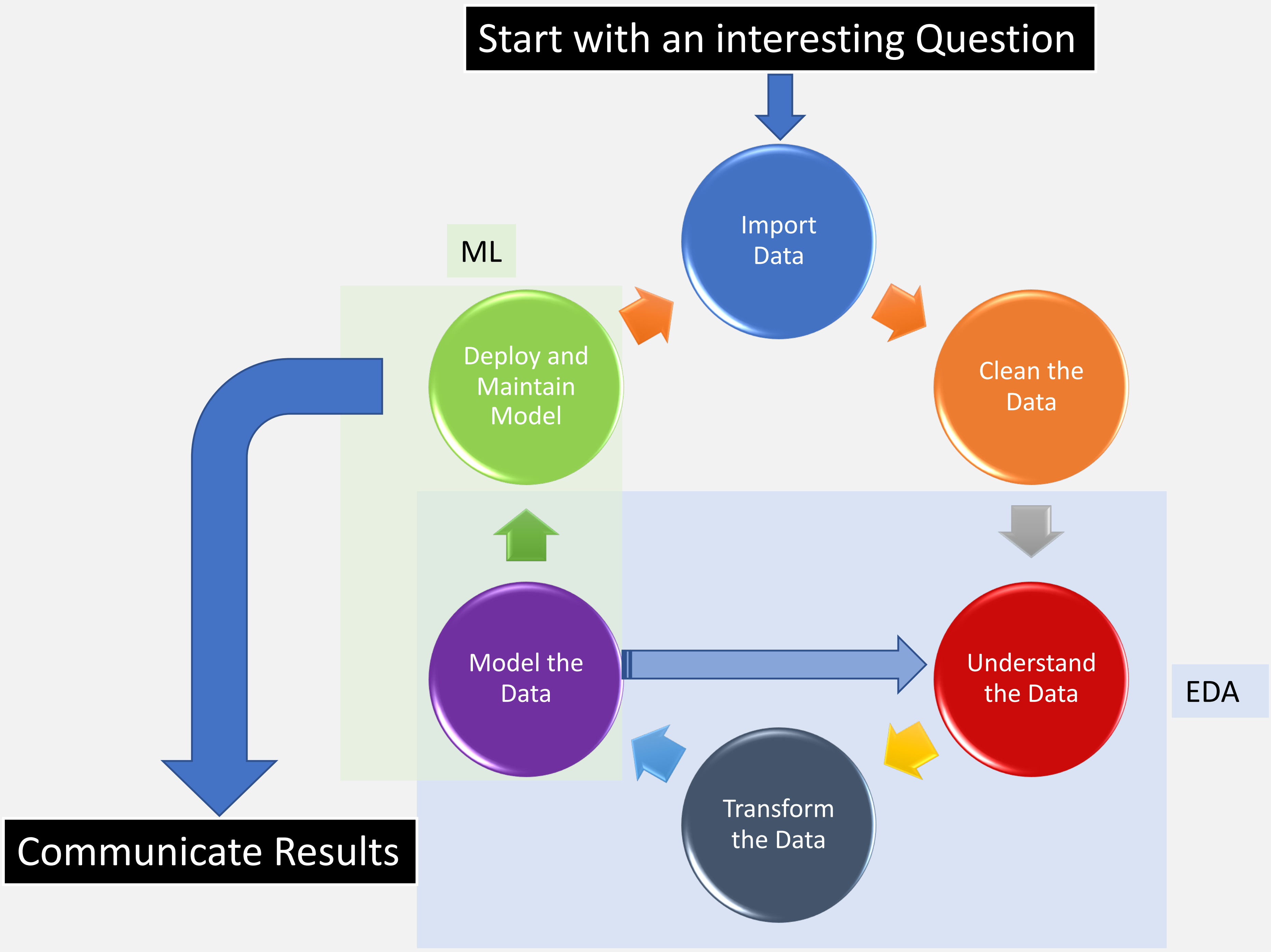
Senior Architect and Distinguished Expert at Unify Inc., an Atos company (Boca Raton, FL)

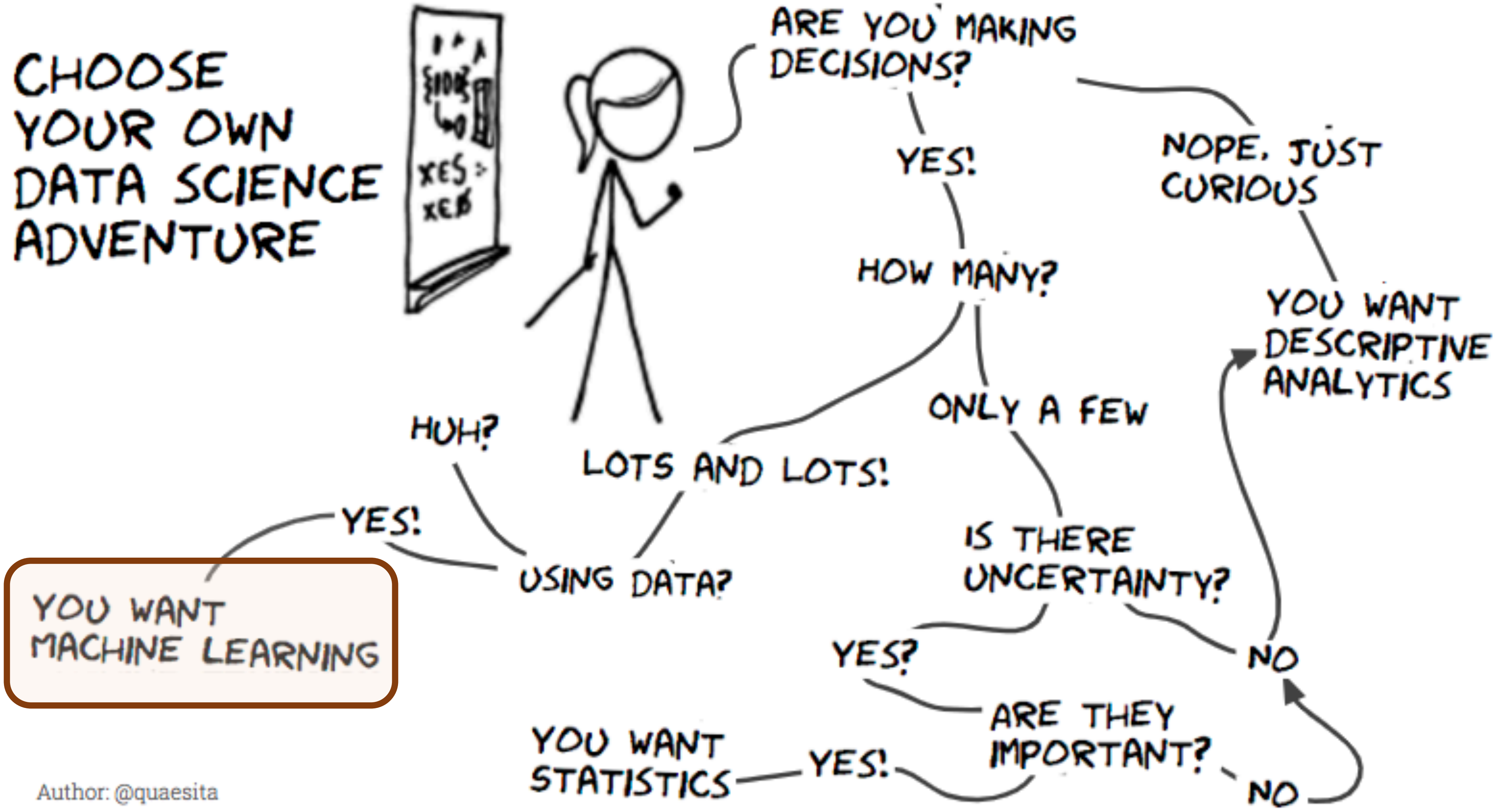
Example 3: Using data to answer questions

tinyurl.com/icmla2019

Part 7:

Machine Learning


Start with an interesting Question
Import Data
Clean the Data
Understand the Data
Transform the Data
Model the Data
Deploy and Maintain Model
ML
EDA
Communicate Results



Source: https://hackernoon.com/what-on-earth-is-data-science-eb1237d8cb37

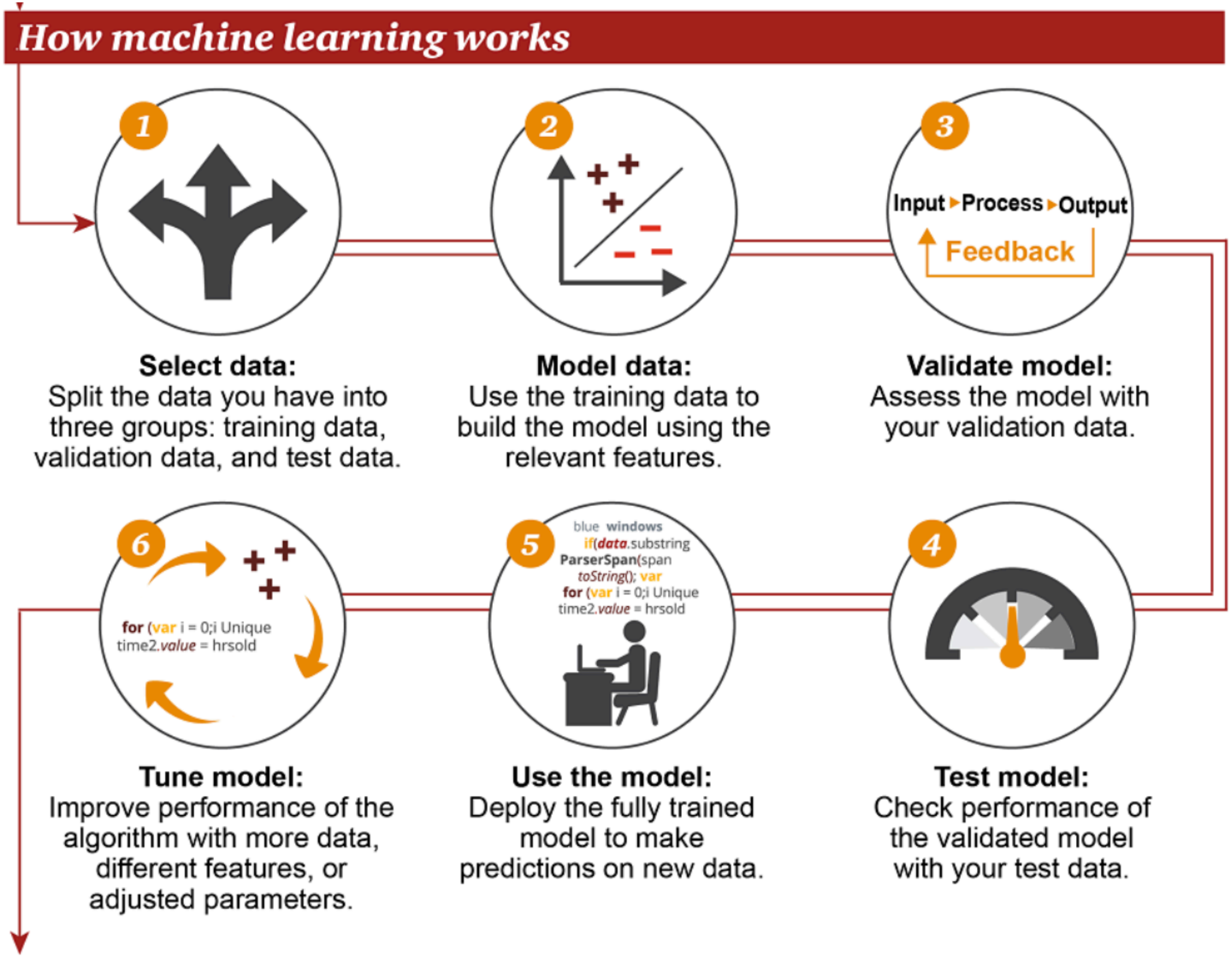
What is Machine Learning?

Machine learning teaches computers to do what comes naturally to humans and animals: learn from experience.

Machine learning algorithms use computational methods to “learn” information directly from data without relying on a predetermined equation as a model.

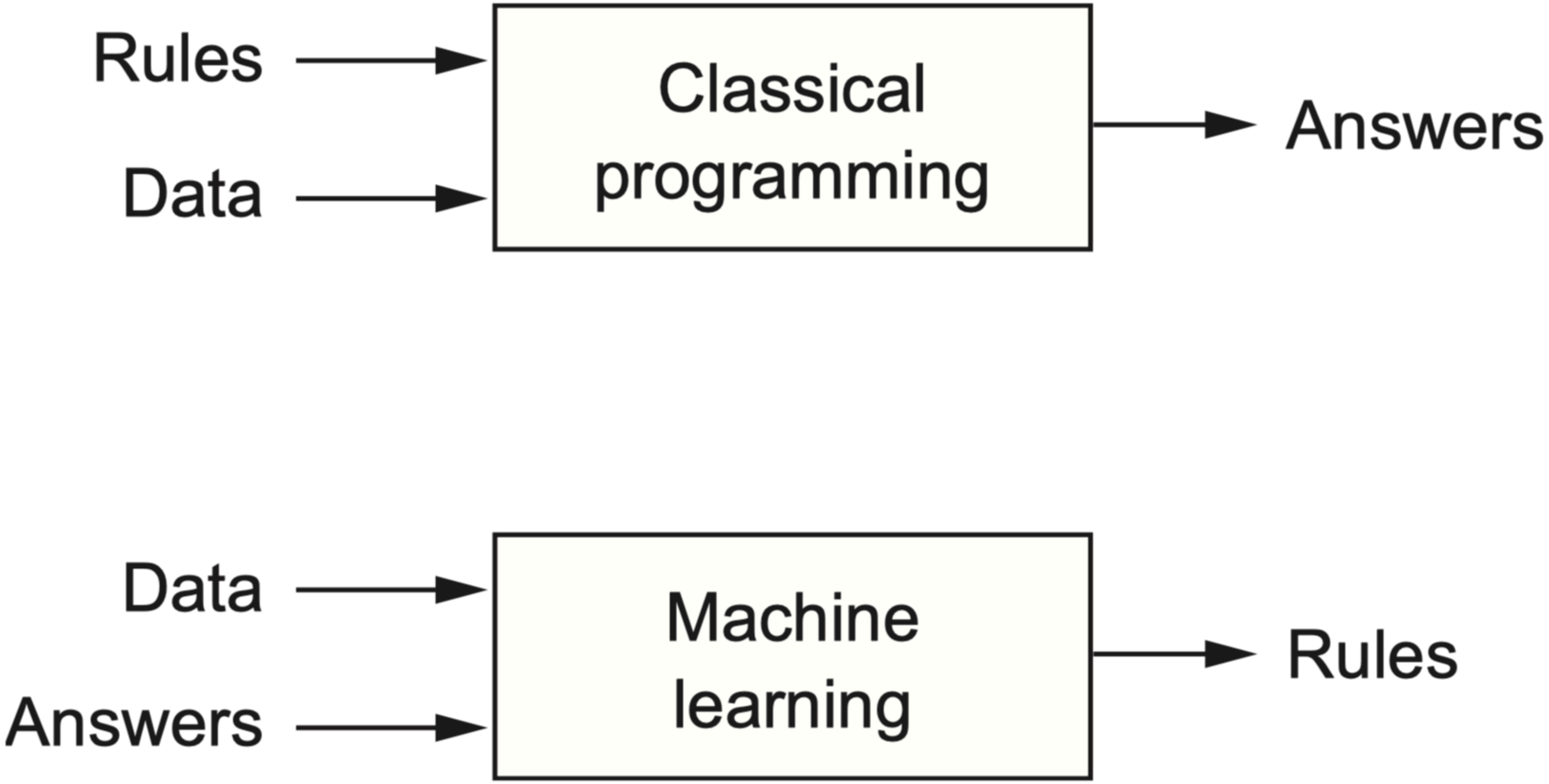
The algorithms adaptively improve their performance as the number of samples available for learning increases.

Source: (Mathworks 2016)



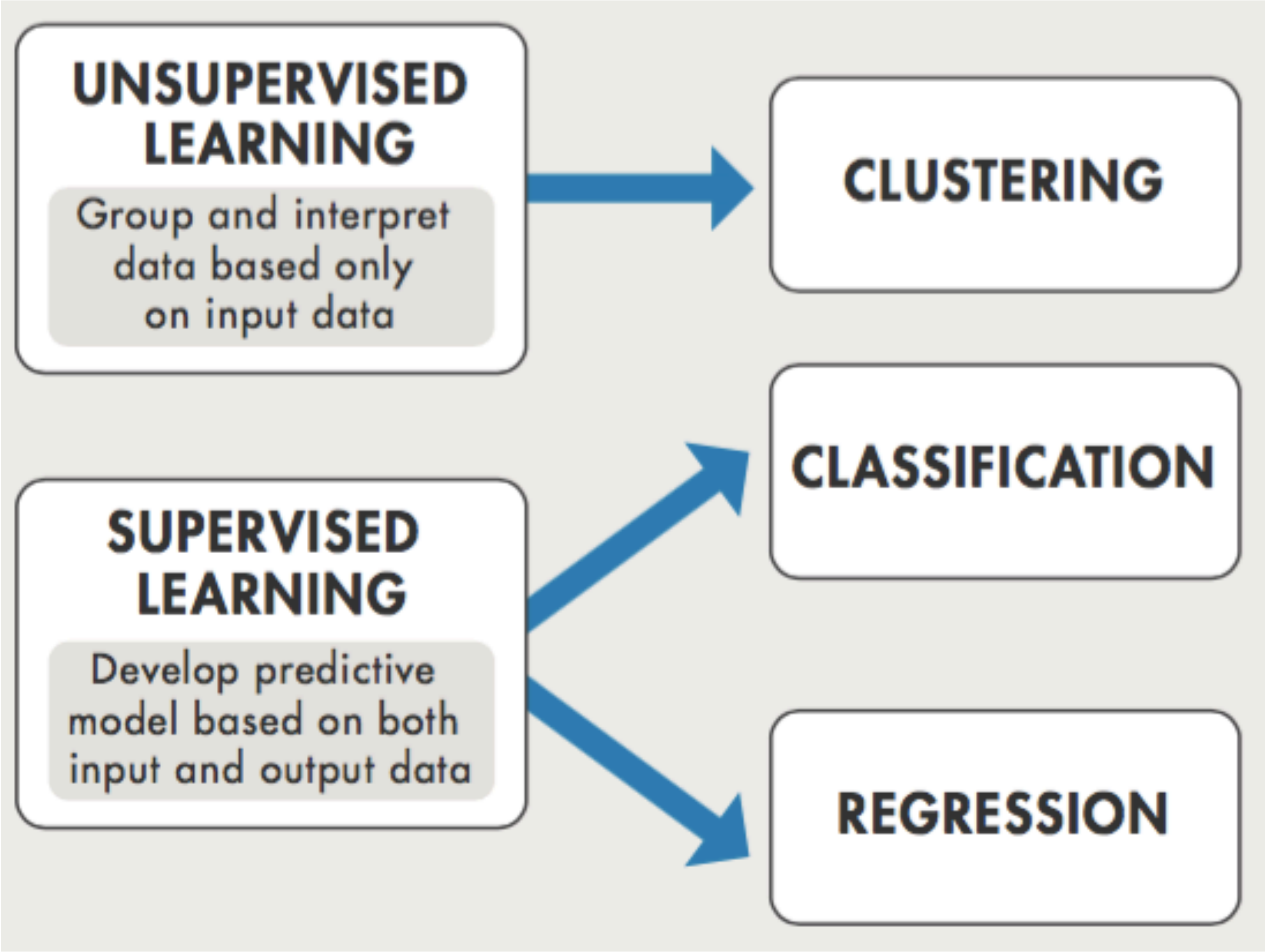
Source: http://usblogs.pwc.com/emerging-technology/machine-learning-methods-infographic/

Machine Learning: a new programming paradigm



Source: F. Chollet, “Deep Learning with Python” - Manning

Machine Learning Techniques



Source: (Mathworks 2016)

Types of learning

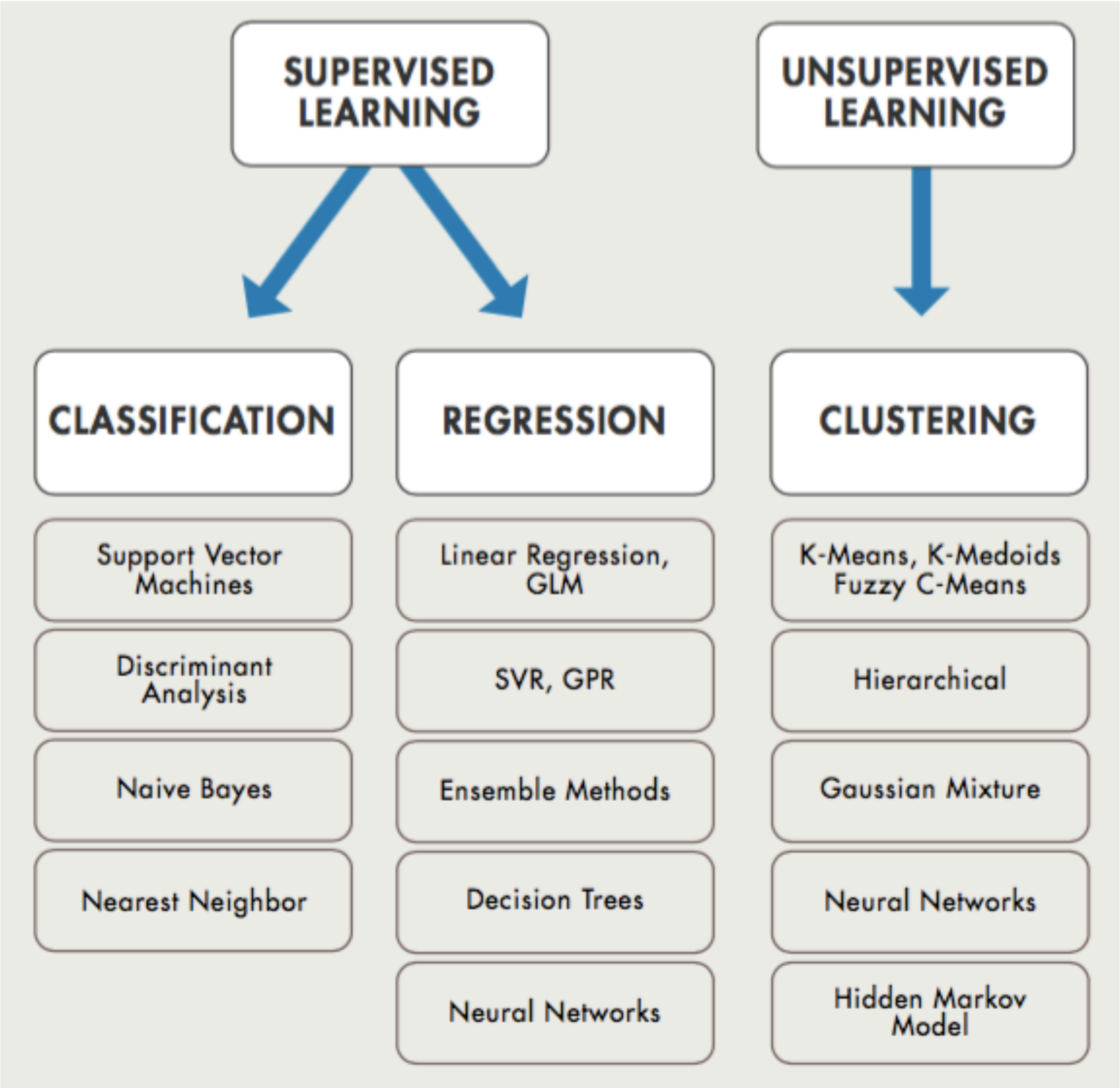
- In **unsupervised learning** the agent learns patterns in the input even though *no explicit feedback is supplied*.
- In **supervised learning** the agent *observes some example input–output pairs and learns a function that maps from input to output*.
- In **reinforcement learning** the agent learns from a series of *reinforcements—rewards or punishments*.

Source: (Russell & Norvig 2011)

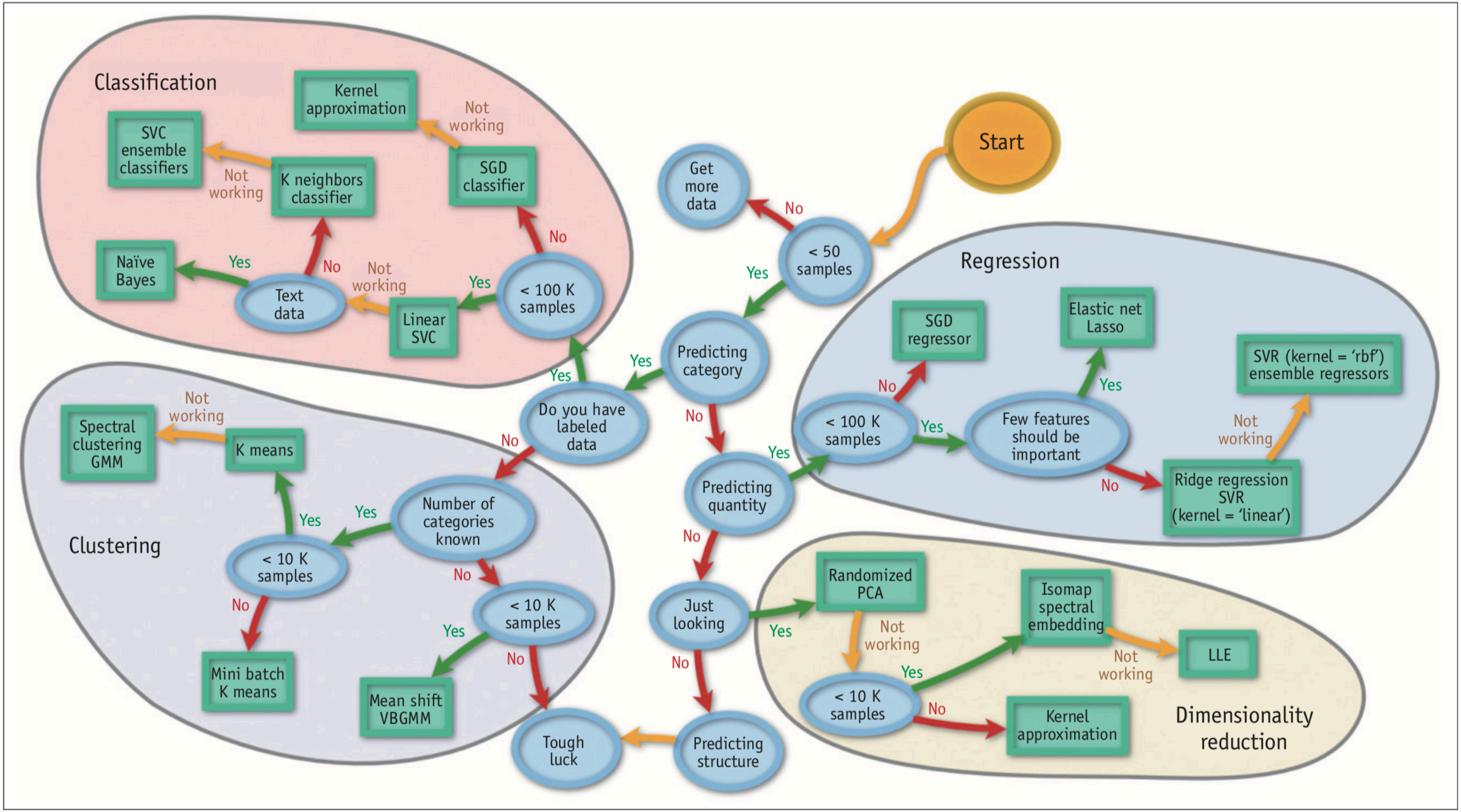
Which ML algorithm to use?

- A potentially overwhelming task!
  - There are dozens of supervised and unsupervised machine learning algorithms, and each takes a different approach to learning.
- There is no best method or one size fits all.
  - Finding the right algorithm is partly just trial and error—even highly experienced data scientists can’t tell whether an algorithm will work without trying it out.
- But algorithm selection also depends on the size and type of data you’re working with, the insights you want to get from the data, and how those insights will be used.

Which ML algorithm to use?



Source: (Mathworks 2016)



Source: scikit-learn

The real challenge in using ML is to find the algorithm whose learning bias is the best match for a particular data set.

A look at *Machine learning methods*

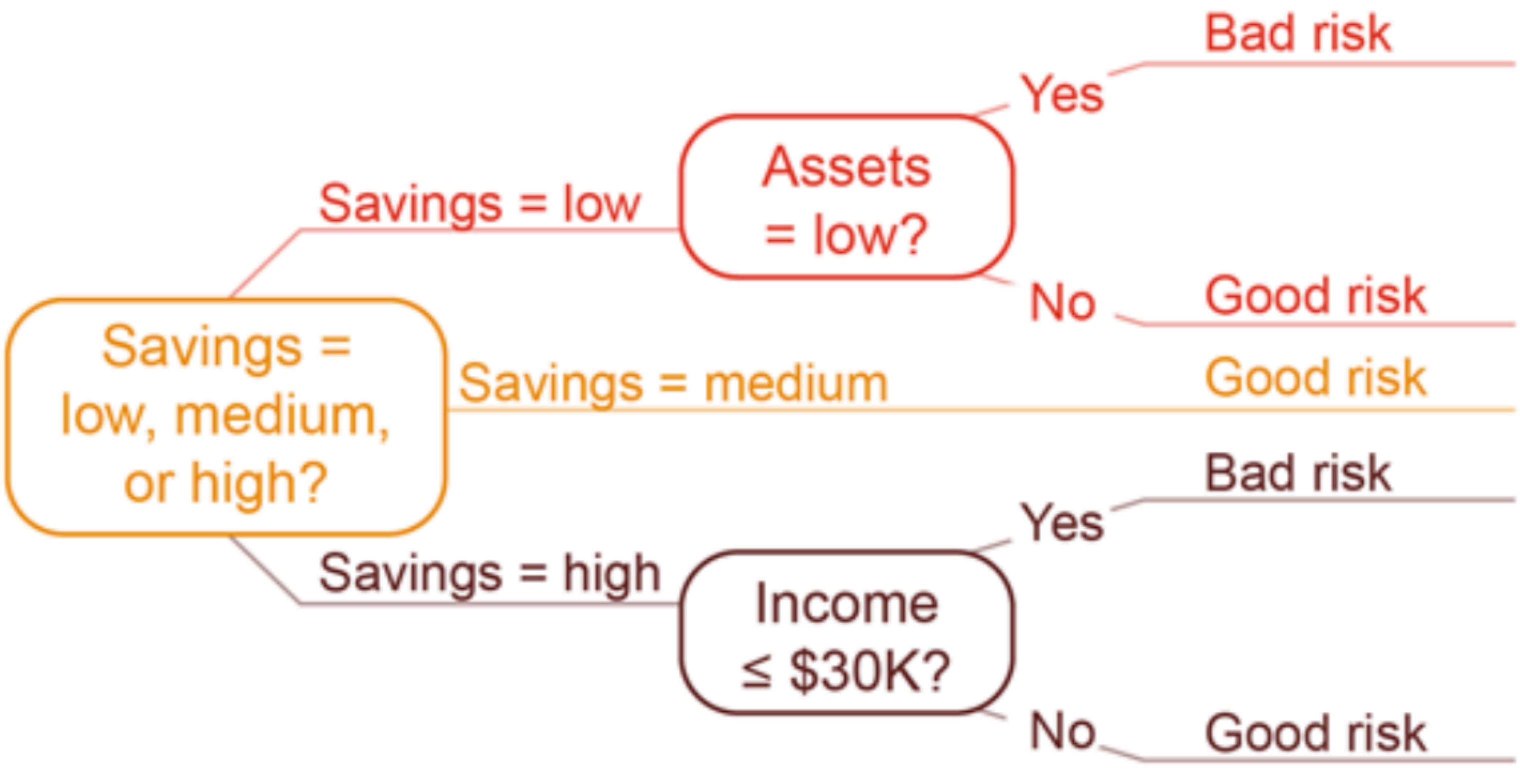
Introduction

Which machine learning algorithm should you use? A lot depends on the characteristics and the amount of the available data, as well as your training goals, in each particular use case. Avoid using the most complicated algorithms unless the end justifies more expensive means and resources. Here are some of the more common algorithms ranked by ease of use.

Source: http://usblogs.pwc.com/emerging-technology/machine-learning-methods-infographic/

Decision trees

Decision tree analysis typically uses a hierarchy of variables or decision nodes that, when answered step by step, can classify a given customer as creditworthy or not, for example.



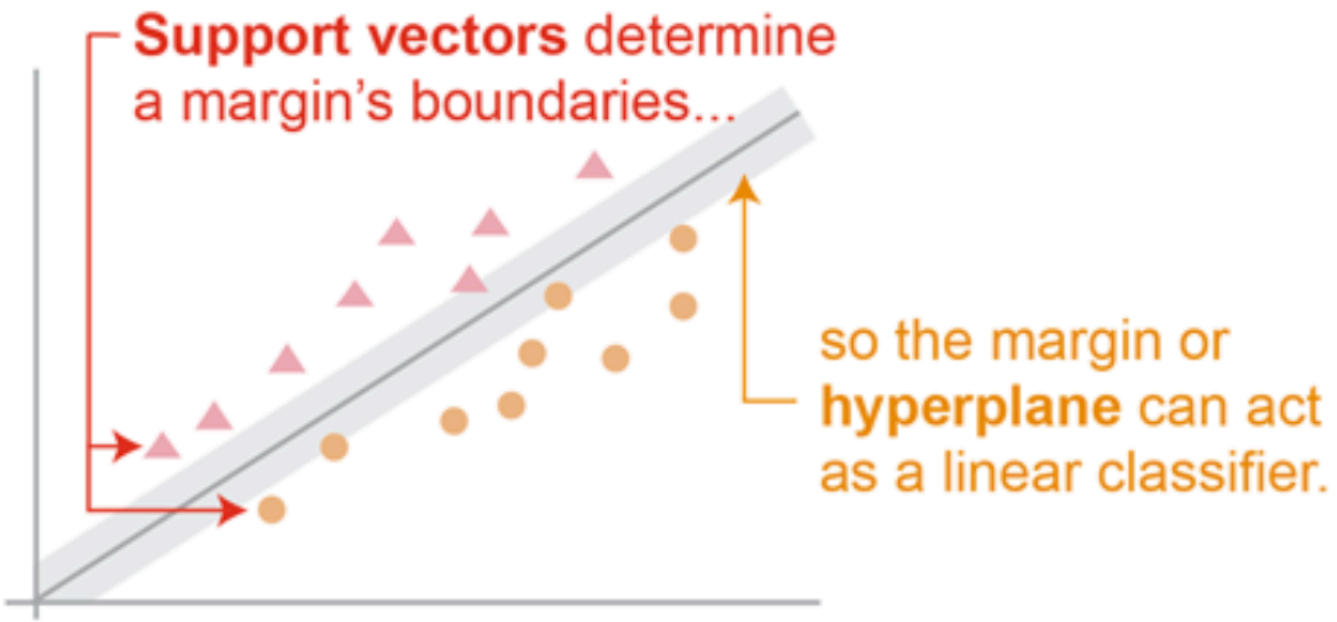
Advantages	Use cases
Decision trees are useful when evaluating lists of distinct features, qualities, or characteristics of people, places, or things.	Rule-based credit risk assessment, horse race performance prediction

Source: Daniel T. Larose and Chantal D. Larose, *Data Mining and Predictive Analytics*, 2nd Edition, John Wiley & Sons, 2015

Source: http://usblogs.pwc.com/emerging-technology/machine-learning-methods-infographic/

Support vector machines

Support vector machines classify groups of data with the help of hyperplanes.



Advantages	Use cases
Support vector machines are good for the binary classification of X versus other variables and are useful whether or not the relationship between variables is linear.	News categorization, handwriting recognition

Source: Matthew Kelly, *Computer Science: Source*, 2010

Source: http://usblogs.pwc.com/emerging-technology/machine-learning-methods-infographic/

Regression

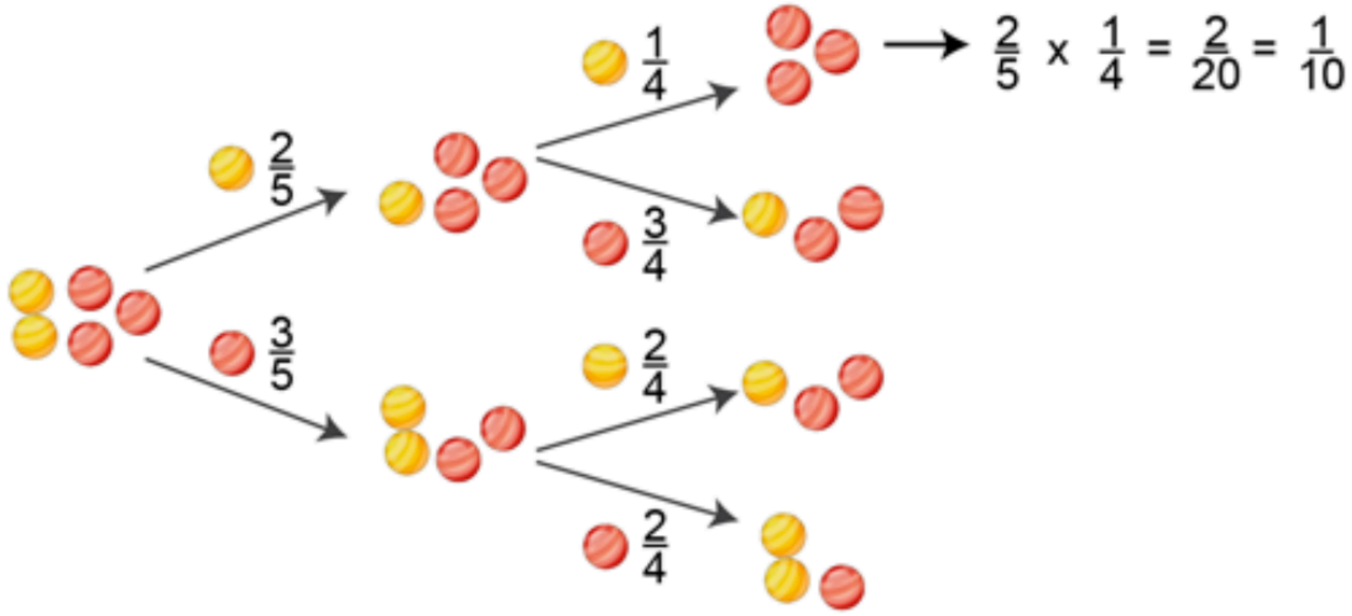
Regression maps the behavior of a dependent variable relative to one or more dependent variables. In this example, logistic regression separates spam from non-spam text.

Advantages	Use cases
Regression is useful for identifying continuous (not necessarily distinct) relationships between variables.	Traffic flow analysis, email filtering

Source: http://usblogs.pwc.com/emerging-technology/machine-learning-methods-infographic/

Naive Bayes classification

Naive Bayes classifiers compute probabilities, given tree branches of possible conditions. Each individual feature is “naive” or conditionally independent of, and therefore does not influence, the others. For example, what’s the probability you would draw two yellow marbles in a row, given a jar of five yellow and red marbles total? The probability, following the topmost branch of two yellow in a row, is one in ten. Naive Bayes classifiers compute the combined, conditional probabilities of multiple attributes.



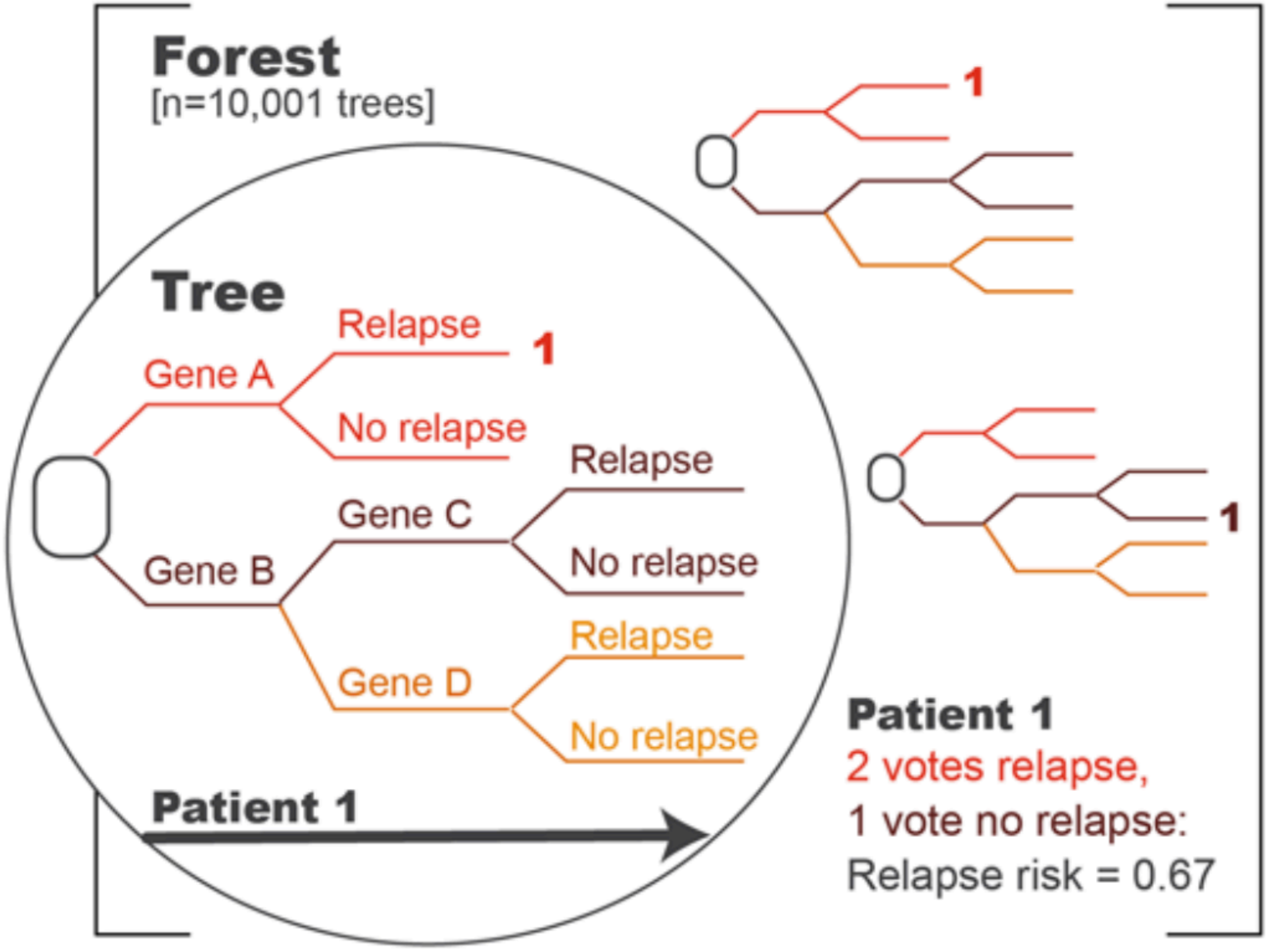
Advantages	Use cases
Naive Bayes methods allow the quick classification of relevant items in small data sets that have distinct features.	Sentiment analysis, consumer segmentation

Source: Rod Pierce, et al., *MathIsFun*, 2014

Source: http://usblogs.pwc.com/emerging-technology/machine-learning-methods-infographic/

Random forest

Random forest algorithms improve the accuracy of decision trees by using multiple trees with randomly selected subsets of data. This example reviews the expression levels of various genes associated with breast cancer relapse and computes a relapse risk.



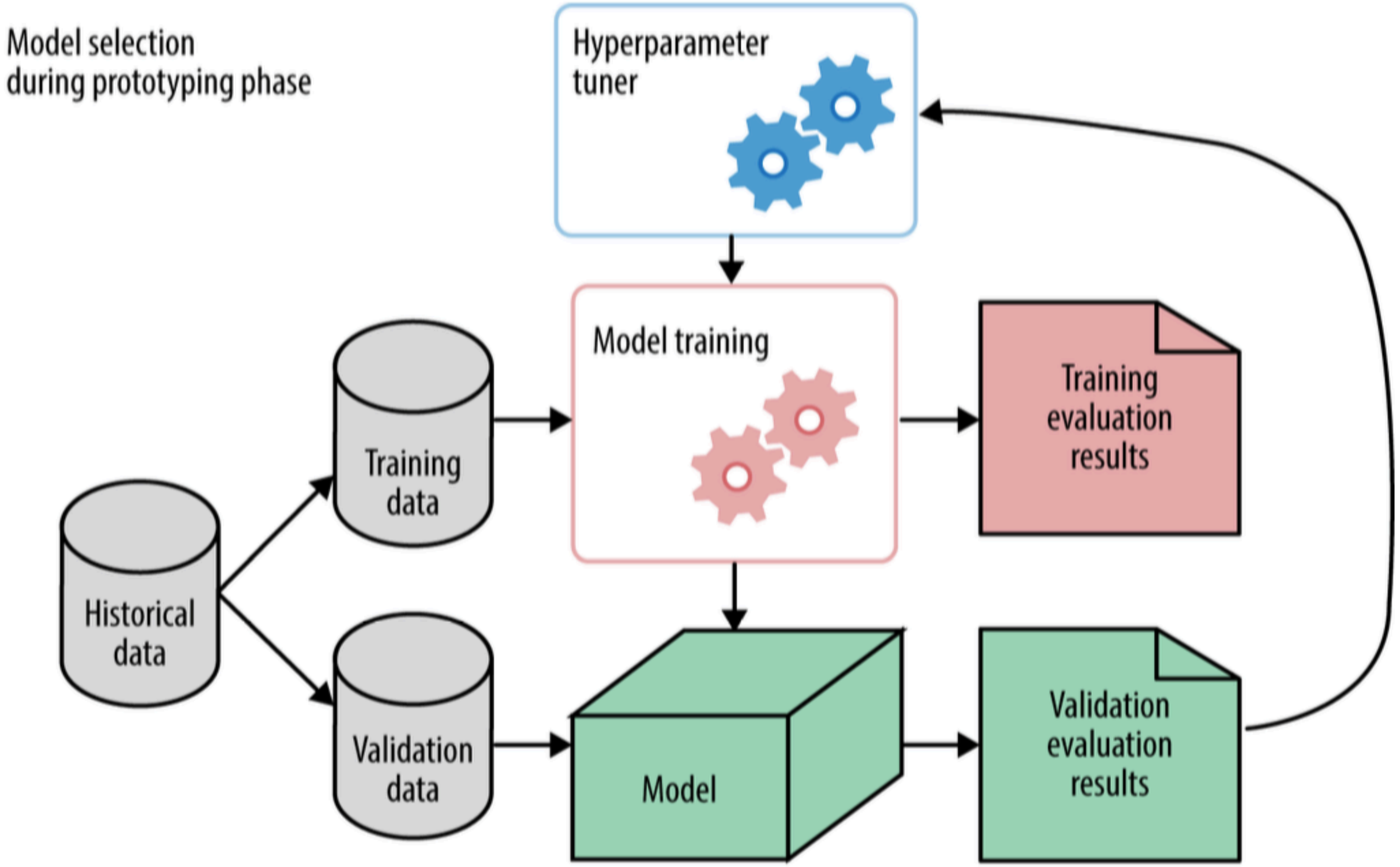
Advantages	Use cases
Random forest methods prove useful with large data sets and items that have numerous and sometimes irrelevant features.	Customer churn analysis, risk assessment

Source: Nicolas Spies, Washington University, 2015

Source: http://usblogs.pwc.com/emerging-technology/machine-learning-methods-infographic/

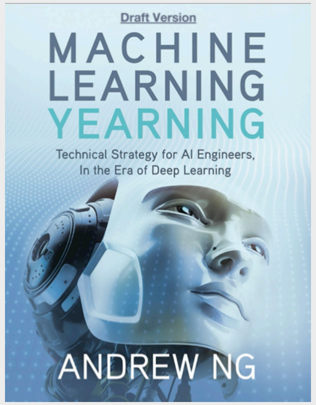
ML workflow

The prototyping phase of building a ML model



Source: (Zheng 2015)

Practical hints and best practices



- **Setting up development and test sets**
  - Choose dev and test sets from a distribution that reflects what data you expect to get in the future and want to do well on.
    - This may not be the same as your training data's distribution.
  - Choose dev and test sets from the same distribution, if possible.
  - The old heuristic of a 70%/30% train/test split does not apply for problems where you have a lot of data; the dev and test sets can be much less than 30% of the data.
  - Your dev set should be large enough to detect meaningful changes in the accuracy of your algorithm, but not necessarily much larger.
  - Your test set should be big enough to give you a confident estimate of the final performance of your system.

Source: https://www.deeplearning.ai/machine-learning-yearning/

Source: A. Ng, "Machine Learning Yearning"

The golden rule for evaluating models is that models should never be tested on the same data they were trained on.

Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)

Practical hints and best practices

- **Beware of peeking!**
  - Do not use the test set to make any decisions regarding the algorithm, including whether to roll back to the previous week’s system.
    - If you do so, you will start to **overfit** to the test set, and can no longer count on it to give a completely unbiased estimate of your system’s performance.

Source: A. Ng, “Machine Learning Yearning”

Practical hints and best practices

- Consider having a **single-number evaluation metric** (such as accuracy)
  - It allows you to sort all your models according to their performance on this metric, and quickly decide what is working best.
  - It speeds up your ability to make a decision when you are selecting among a large number of classifiers.
  - It gives a clear preference ranking among all of them, and therefore a clear direction for progress.

Source: A. Ng, "Machine Learning Yearning"

Evaluation criteria

- Confusion matrix

	P' (Predicted)	N' (Predicted)
P (Actual)	True Positive	False Negative
N (Actual)	False Positive	True Negative

We measure these answers by counting the number of:

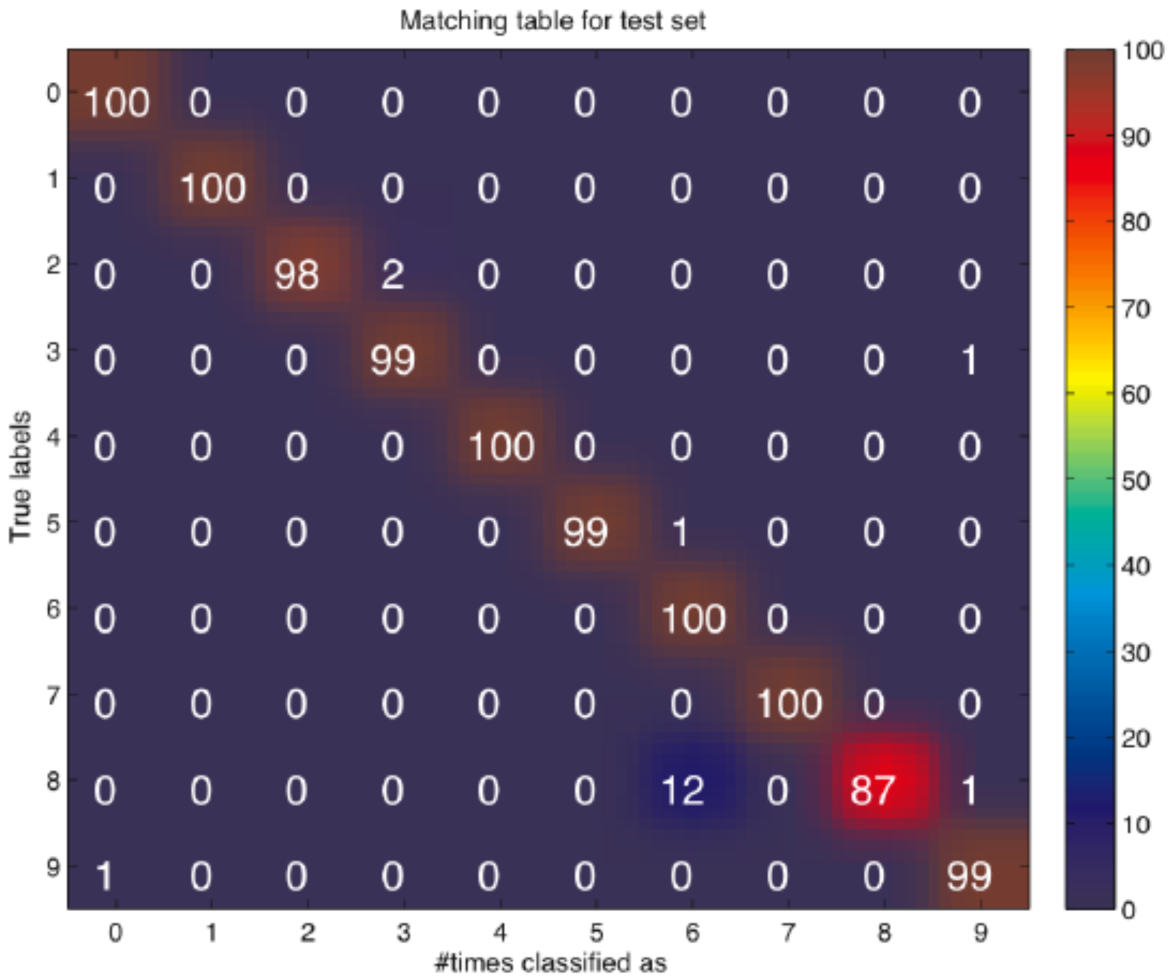
true positives
— positive prediction
— label was positive
false positives
— positive prediction
— label was negative
true negatives
— negative prediction
— label was negative
false negatives
— negative prediction
— label was positive

Source: (Gibson&Patterson 2016)

Evaluation criteria

- Confusion matrix: example

10-digit classifier (OCR)



Source: (Marques 2011)

Evaluation criteria

- Sensitivity, specificity, and accuracy
  - **Sensitivity** quantifies how well the model avoids false negatives.

    Sensitivity = TP / (TP + FN)

  - **Specificity** quantifies how well the model avoids false positives.

    Specificity = TN / (TN + FP)

  - **Accuracy** is the degree of closeness of measurements of a quantity to that quantity's true value.

    Accuracy = (TP + TN) / (TP + FP + FN + TN)

Source: (Gibson&Patterson 2016)

Evaluation criteria

- Precision, recall, and F1
  - **Precision** (also known as *the positive prediction value*) is the degree to which repeated measurements under the same conditions give us the same results.

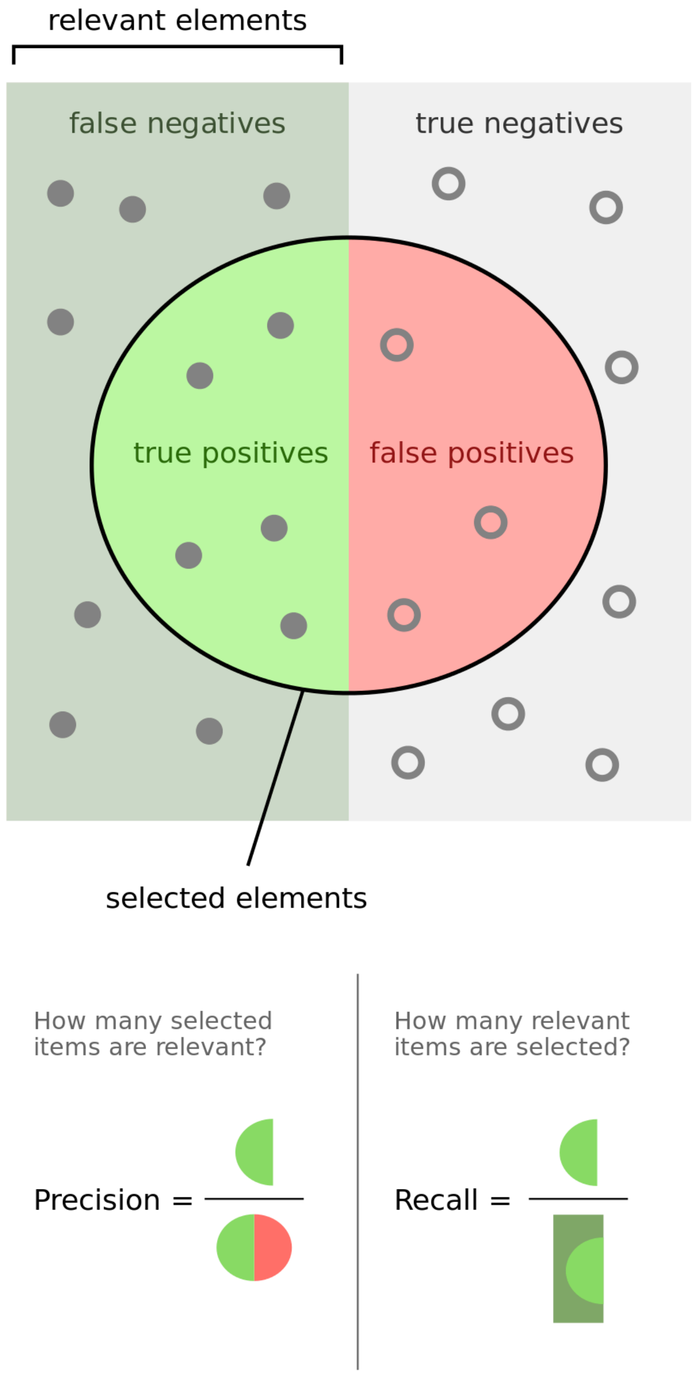
$$\text{Precision} = TP / (TP + FP)$$

  - **Recall** is the same as **sensitivity**

$$\text{Recall} = TP / (TP + FN)$$

  - The **F1 score** is the harmonic mean of both the precision and recall measures into a single score

$$F1 = 2TP / (2TP + FP + FN)$$



Source: (Gibson&Patterson 2016)

Practical hints and best practices

- Q: **Which model is best?**
- A: Classifier A

Classifier	Precision	Recall
A	95%	90%
B	98%	85%

Classifier	Precision	Recall	F1 score
A	95%	90%	**92.4%**
B	98%	85%	**91.0%**

Source: A. Ng, "Machine Learning Yearning"

Practical hints and best practices

- Q: **Which model is best?**

- A: It depends...
  - If running time < 100 ms = "satisficing metric"
  - Then Classifier B is best according to the "optimizing metric" (accuracy)

Classifier	Accuracy	Running time
A	90%	80ms
B	92%	95ms
C	95%	1,500ms

Source: A. Ng, "Machine Learning Yearning"

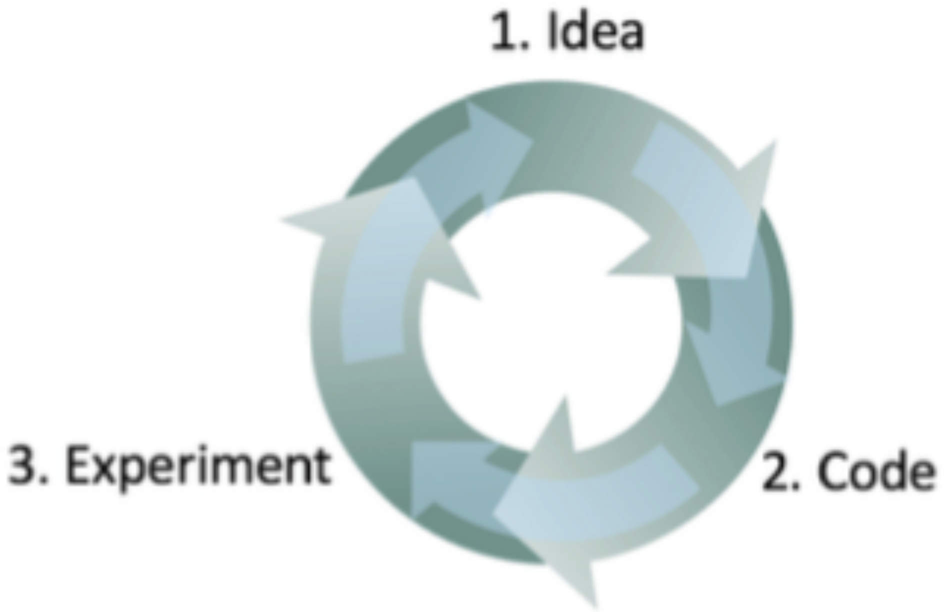
Practical hints and best practices

- If you are trading off N different criteria:
  - set N-1 of the criteria as “satisficing” metrics, i.e., you simply require that they meet a certain value.
  - then define the final one as the “optimizing” metric.

Practical hints and best practices

Building a machine learning system is an iterative process:

1. Start off with some **idea** on how to build the system.
2. Implement the idea in **code**.
3. Carry out an **experiment** which tells how well the idea worked.
   - Based on these learnings, go back to generate more ideas, and keep on iterating.

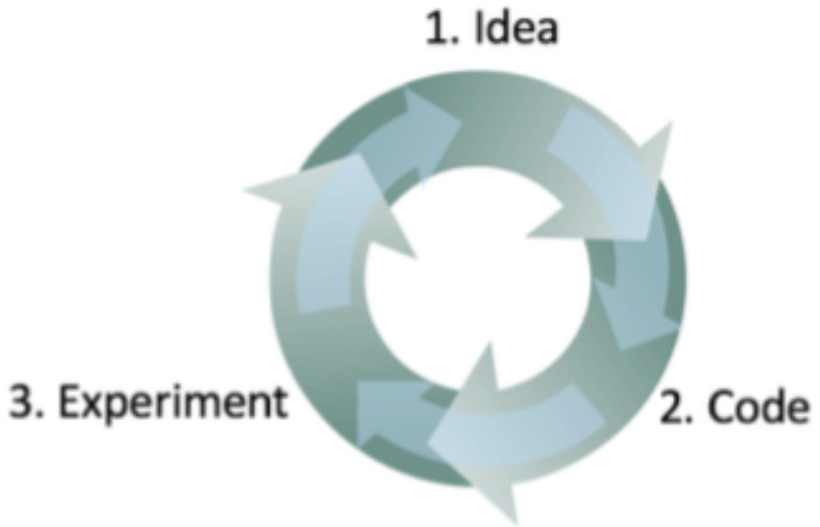


Source: A. Ng, “Machine Learning Yearning”

Practical hints and best practices

Having a dev set and metric speeds up iterations

- The faster you can go round this loop, the faster you will make progress.
- This is why having dev/test sets and a metric are important:
  - Each time you try an idea, measuring your idea's performance on the dev set lets you quickly decide if you're heading in the right direction.



Source: A. Ng, "Machine Learning Yearning"

Practical hints and best practices

- **If ever your dev set and metric are no longer pointing your team in the right direction, quickly change them**:
  - (i) If you had overfit the dev set, get more dev set data.
  - (ii) If the actual distribution you care about is different from the dev/test set distribution, get new dev/test set data.
  - (iii) If your metric is no longer measuring what is most important to you, change the metric.

Source: A. Ng, “Machine Learning Yearning”

Practical hints and best practices

- **Invest time into Error Analysis**
  - “Error Analysis” refers to the process of (manually) examining dev set examples that your algorithm misclassified, so as to understand the underlying causes of the errors.
  - This can both help you prioritize projects and inspire new directions.
  - However, it does not result in a rigid mathematical formula that tells you what should be the highest priority task.

Source: A. Ng, “Machine Learning Yearning”

Practical hints and best practices

Error Analysis example

- Your cat detector solution has problems:
    1. Occasionally *dogs* are being recognized as cats.
    2. Sometimes "*great cats*" (lions, panthers, etc.) are recognized as house cats (pets).
    3. The system's performance on *blurry* images should be improved.
- Which one would you tackle first?

Image	Dog	Great cat	Blurry	Comments
1	✔			Usual pitbull color
2			✔	
3		✔	✔	Lion; picture taken at zoo on rainy day
4		✔		Panther behind tree
...	...	...	...	...
% of total	8%	43%	61%	

Source: A. Ng, "Machine Learning Yearning"

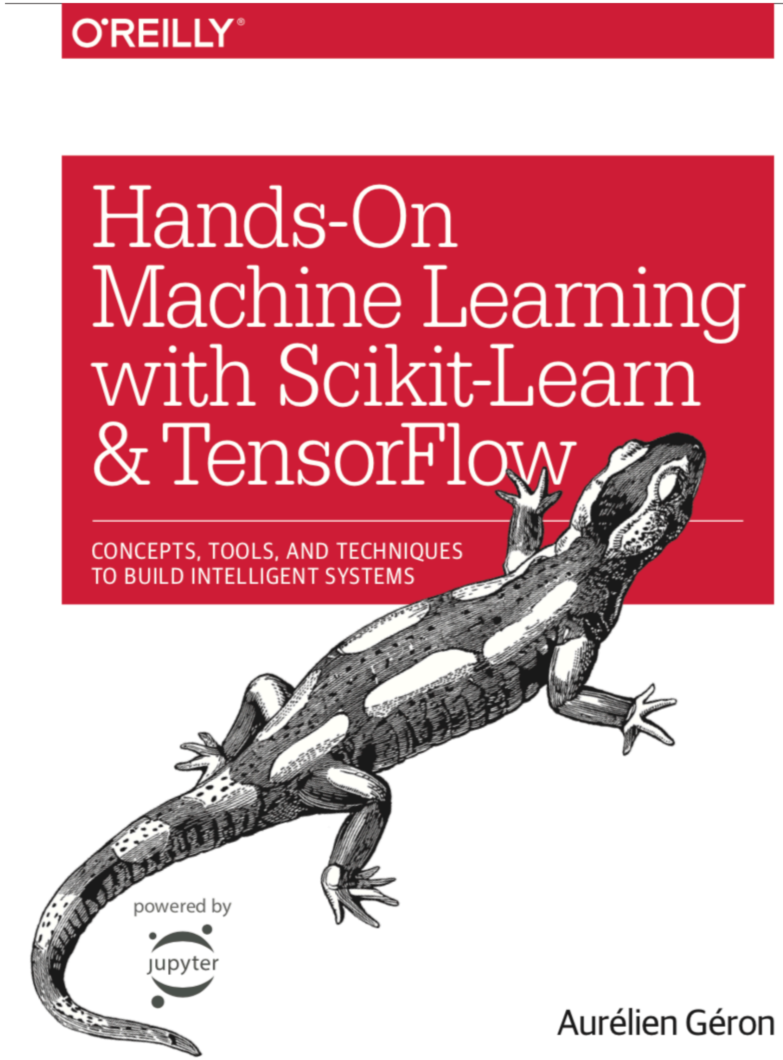
Every successful data science project begins by clearly defining the problem that the project will help solve.

Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)

A recipe for practitioners

End-to-end Machine Learning

1. Look at the big picture.
2. Get the data.
3. Discover and visualize the data to gain insights.
4. Prepare the data for Machine Learning algorithms.
5. Select a model and train it.
6. Fine-tune your model.
7. Present your solution.
8. Launch, monitor, and maintain your system.



Source: (Géron 2017)

Hands on!

Christian Garbin

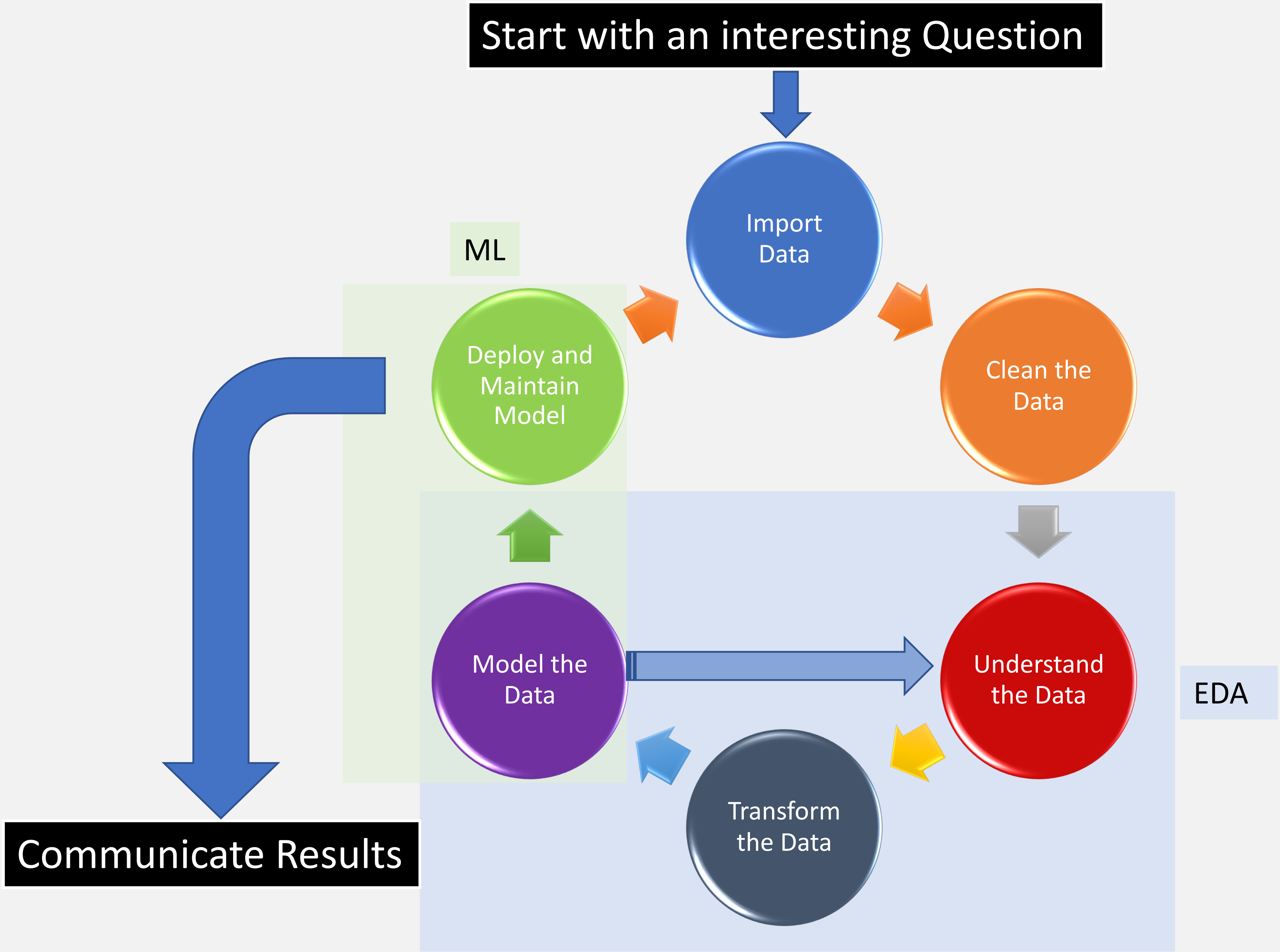
Senior Architect and
Distinguished Expert at
Unify Inc., an Atos company
(Boca Raton, FL)

Example 4:
Machine Learning and Data Science

tinyurl.com/icmla2019

Part 8:

Data Science beyond the code

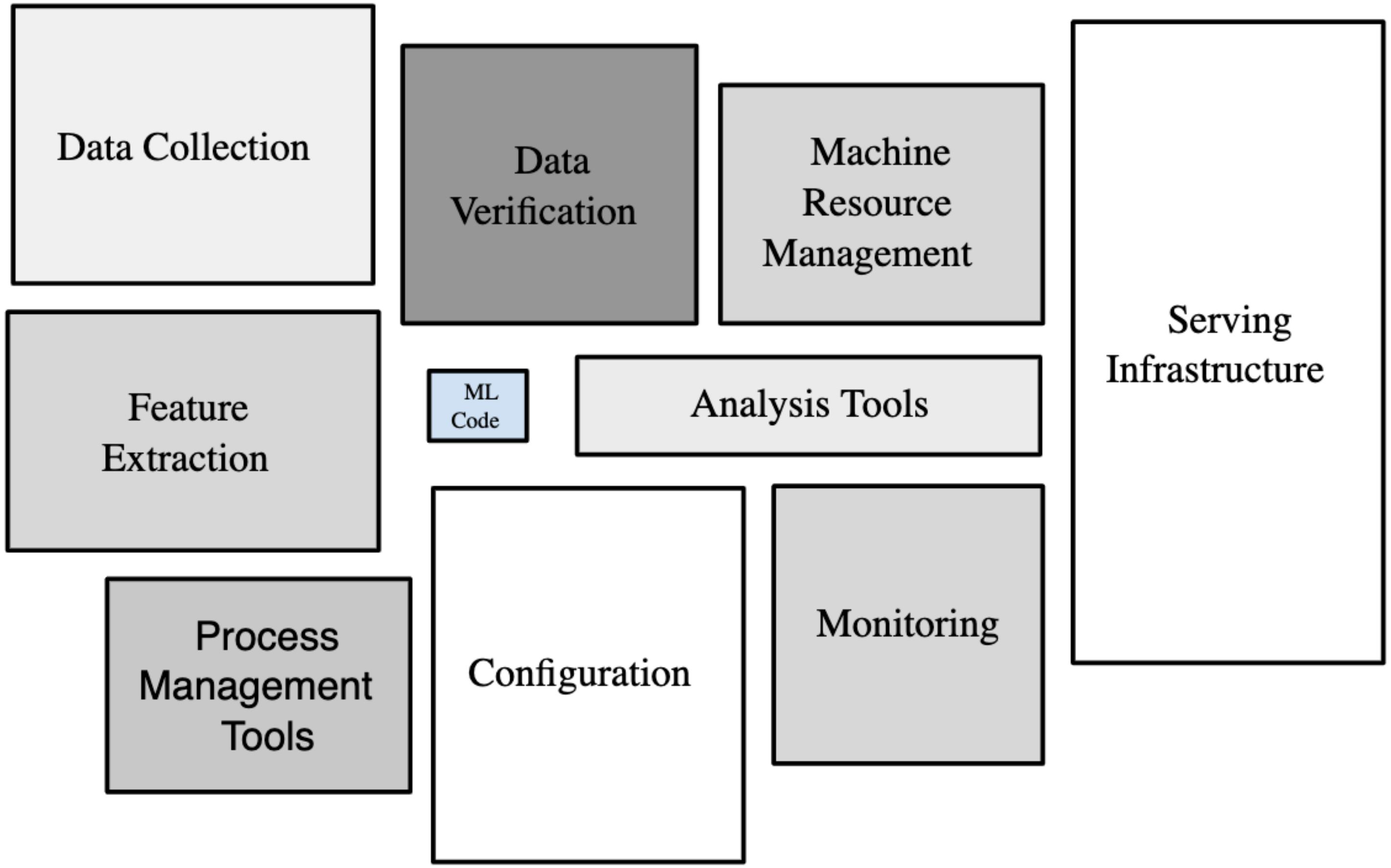

Start with an interesting Question
Import Data
Clean the Data
Understand the Data
Transform the Data
Model the Data
Deploy and Maintain Model
ML
EDA
Communicate Results

Key takeaways

- Data scientists’ concerns reach far beyond the scope of our “main diagram” for this tutorial
- “There’s a huge difference between building a Jupyter notebook model in the lab and deploying a production system that generates business value.” (Andrew Ng)
- The growing use of AI, machine learning, deep learning, and big data analytics leads to many social, legal, and ethical challenges and implications
- We need effective strategies to prepare for an AI-heavy data-driven lifestyle

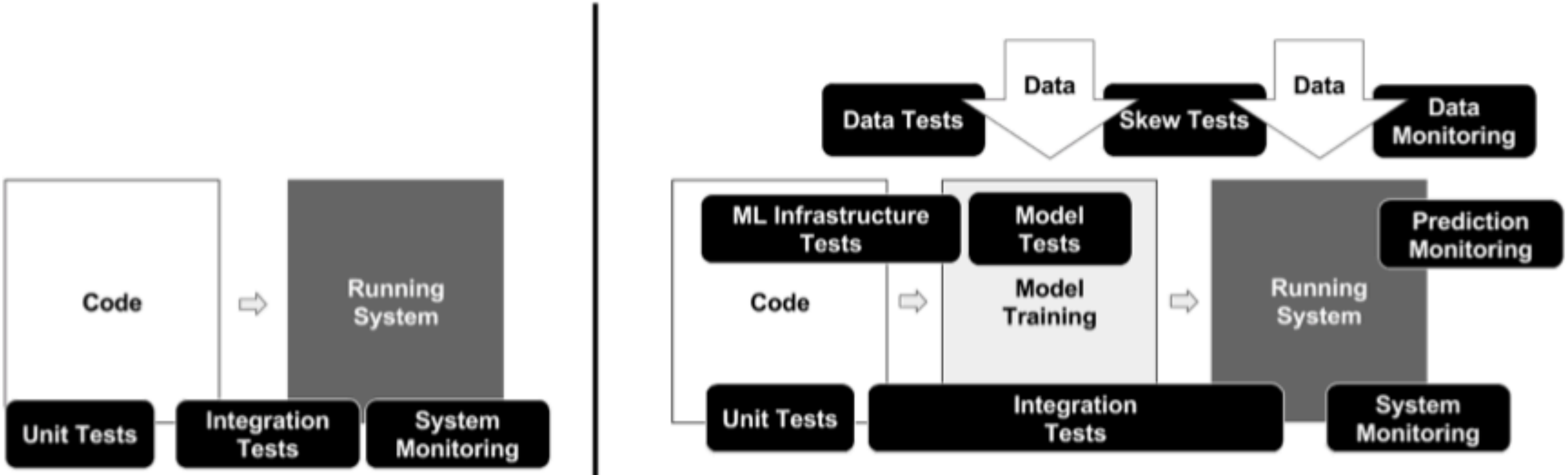
Real-world production ML system

"The ML code is at the heart of a real-world ML production system, but that box often represents only <u>5% or less of the overall code</u> of that total ML production system."



Source: https://developers.google.com/machine-learning/crash-course/production-ml-systems

ML Systems Require Extensive Testing and Monitoring



Traditional System Testing and Monitoring

ML-Based System Testing and Monitoring

Source: https://research.google/pubs/pub46555/

ALGORITHMIA

2020 state of enterprise machine learning

- Based on a survey of nearly 750 people including machine learning practitioners, managers overseeing machine learning projects, and executives at large tech corporations.
- More than 2/3 of the subgroup that was asked about budgets reported increased spending on AI between 2018 and 2019
- Nonetheless, 43% of respondents cited difficulty scaling machine learning projects to their company's needs, up 13% from last year's survey.
- Half of respondents said their company takes between a week and three months to deploy a model. 18% said it takes from three months to a year.

Source: deeplearning.ai "The Batch" Dec 18, 2019

Legal challenges of AI / Data Science

- New regulations
  - General Data Protection Regulation (GDPR) (EU)
    - Adopted April 2016, enforced as of May 2018
  - California Consumer Privacy Act (2020)

- Data privacy

- The right to explanation

- Discrimination and bias: age, gender, and racial bias
  - Where is the bias coming from?

Source: AI Accountability Essential Training (B. Poulson)

Ethical implications of AI / Data Science

- Profiling and discrimination
- Examples
  - Diagnosing diseases with high risks and costs
  - Parole applications
  - Predicting the risk of defaulting on a loan

Personalization can result in preferential treatment for some and marginalization of others.

Unless used very carefully, data science can actually perpetuate and reinforce prejudice.

Source: AI Accountability Essential Training (B. Poulson)

Source: Kelleher and Tierney, “Data Science” (MIT Press, 2018)

Confronting the challenges
Strategies for developers

- **Preparing your project**
  - What are you trying to maximize?
  - What data is available and legally usable?
  - Is AI really necessary?
- **Dealing with bias**
  - Include diverse training data
  - Give special focus to small groups and edge cases
  - Know what's happening in any packages you use
- **Protecting your work**
  - Adversarial examples, exploitation
  - Build as much transparency as possible

Source: AI Accountability Essential Training (B. Poulson)

Confronting the challenges

Strategies for executives

- **Look at the ROI**
  - AI works better for some projects than others
  - Base ROI calculations on reasonable projections
  - Don't get caught in the hype cycle
- **Beware of data restrictions**
- **Simpler is better**
  - Narrow AI has been more successful than general AI
- **Stay in communication with your team**
  - Legal department
  - CIO and data security team
  - PR managers

Source: AI Accountability Essential Training (B. Poulson)

Confronting the challenges

Strategies for consumers

- Know your rights
- Fight for your rights
- Use your rights
- Choose ethical services / companies

Source: AI Accountability Essential Training (B. Poulson)

Confronting the challenges

Strategies for all of us

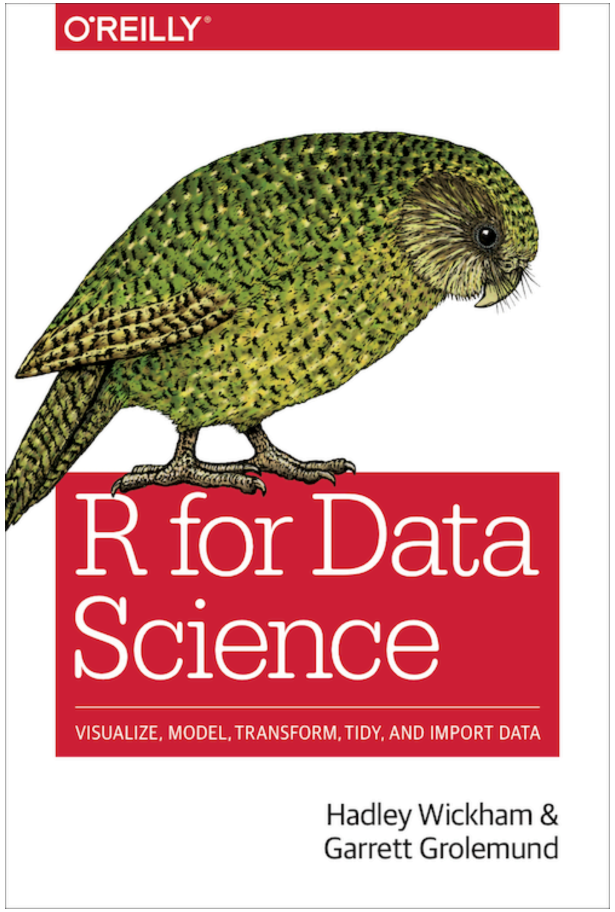
- Stay *relevant*

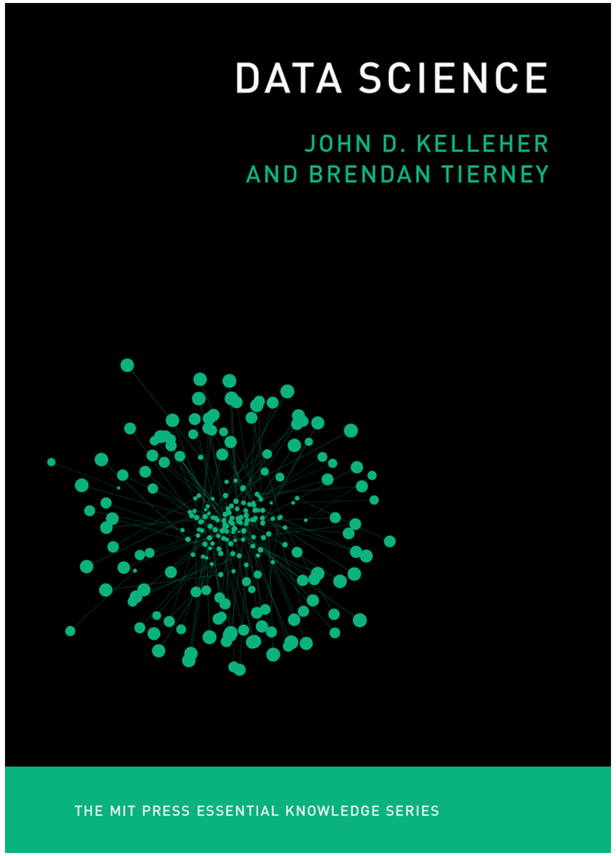
- Adopt a *lifelong learning* attitude

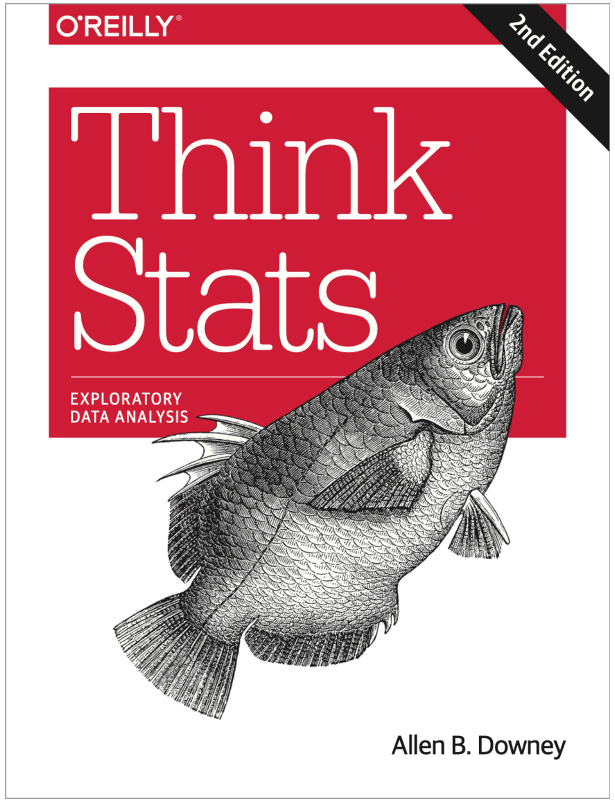
Part 9:

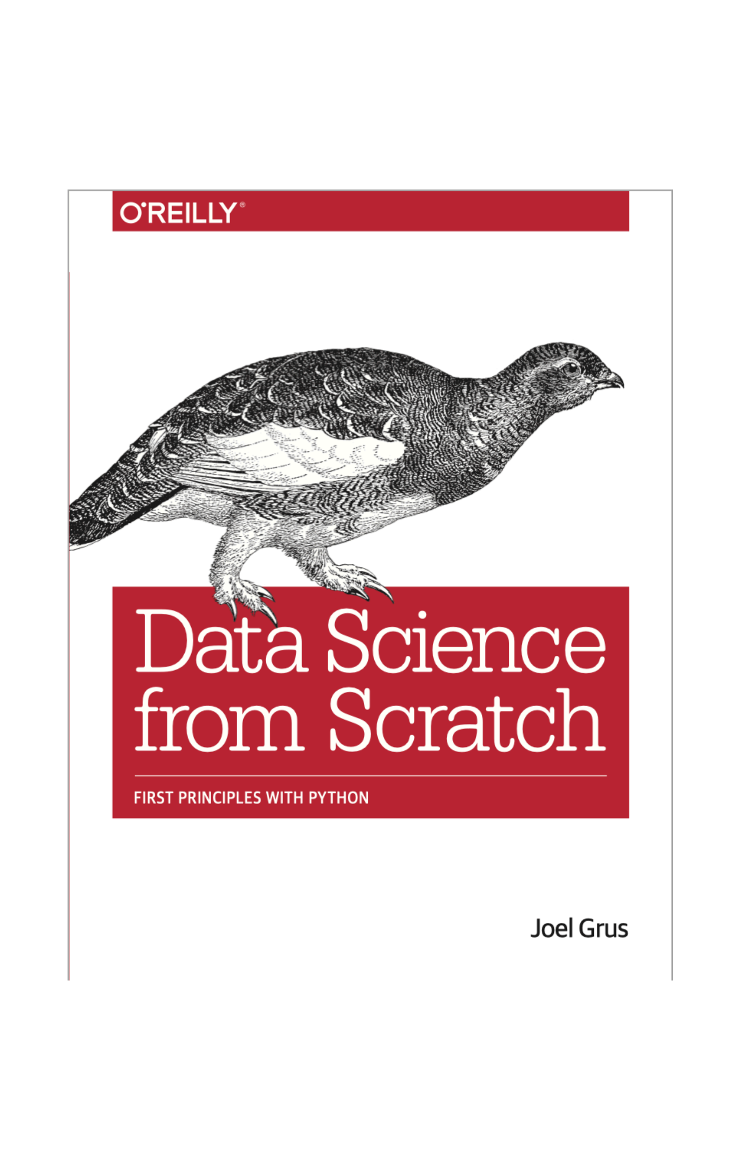
Recommended books and resources

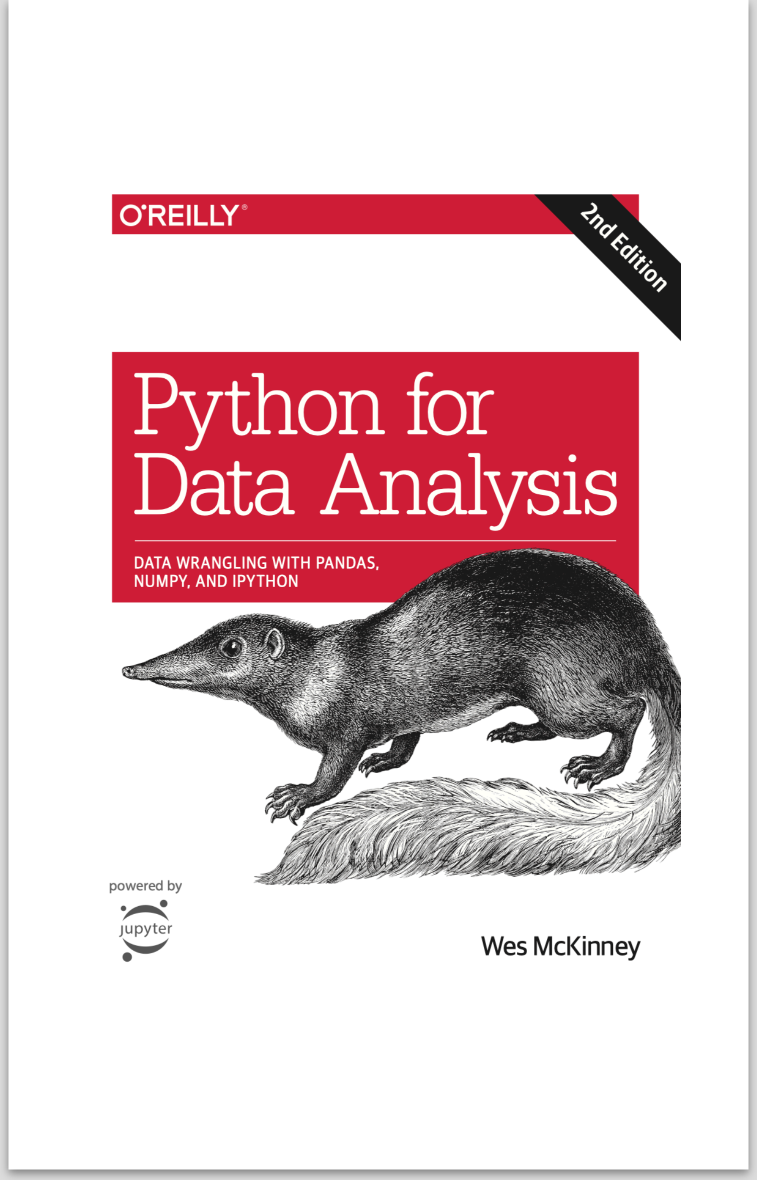
BOOKS

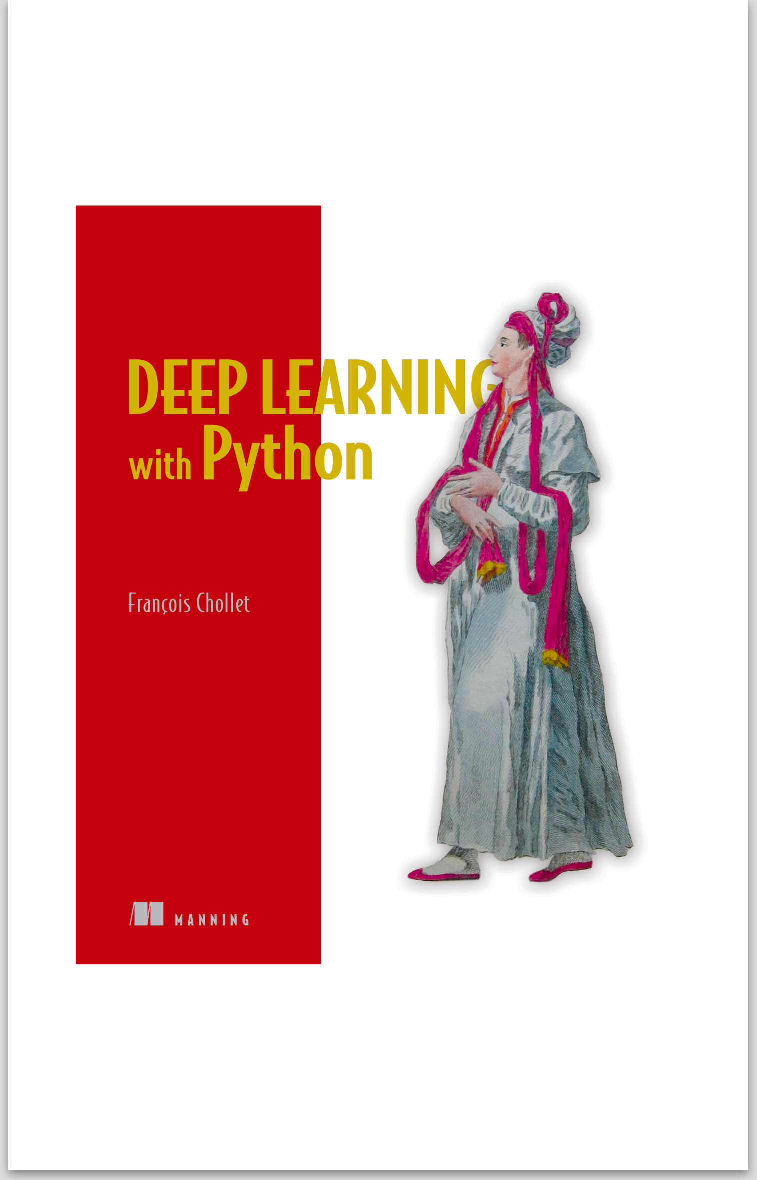

O'REILLY
R for Data Science
VISUALIZE, MODEL, TRANSFORM, TIDY, AND IMPORT DATA
Hadley Wickham &
Garrett Grolemund

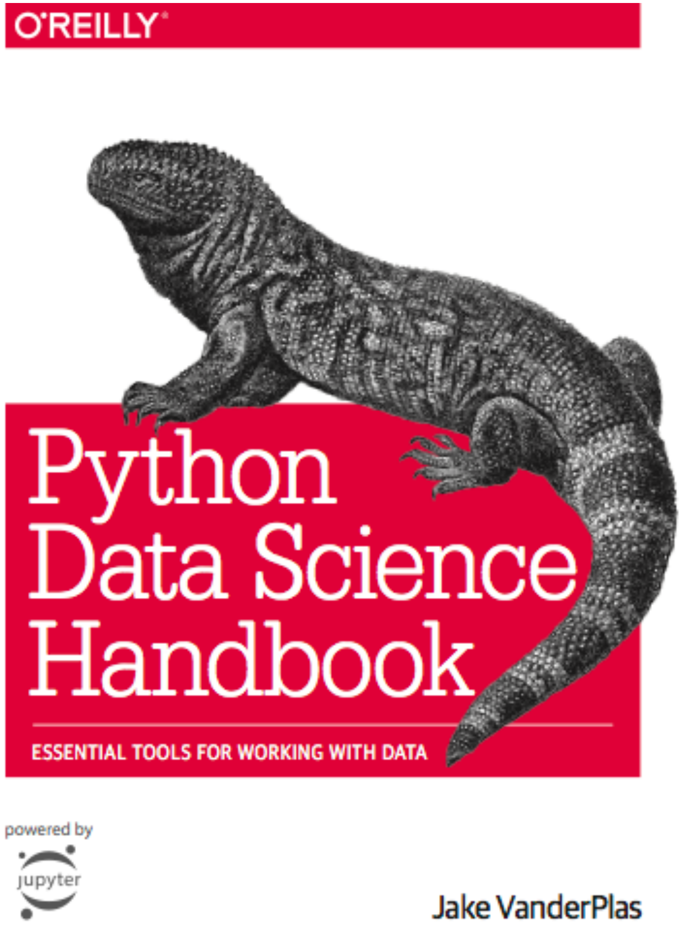

DATA SCIENCE
JOHN D. KELLEHER
AND BRENDAN TIERNEY
THE MIT PRESS ESSENTIAL KNOWLEDGE SERIES

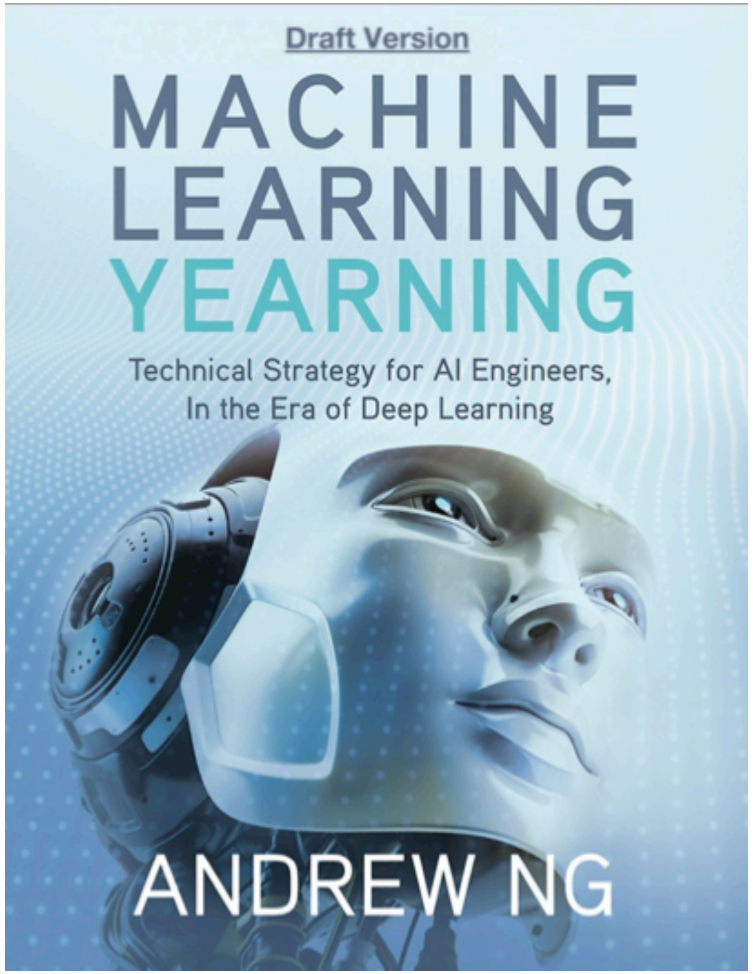

O'REILLY
2nd Edition
Think Stats
EXPLORATORY
DATA ANALYSIS
Allen B. Downey

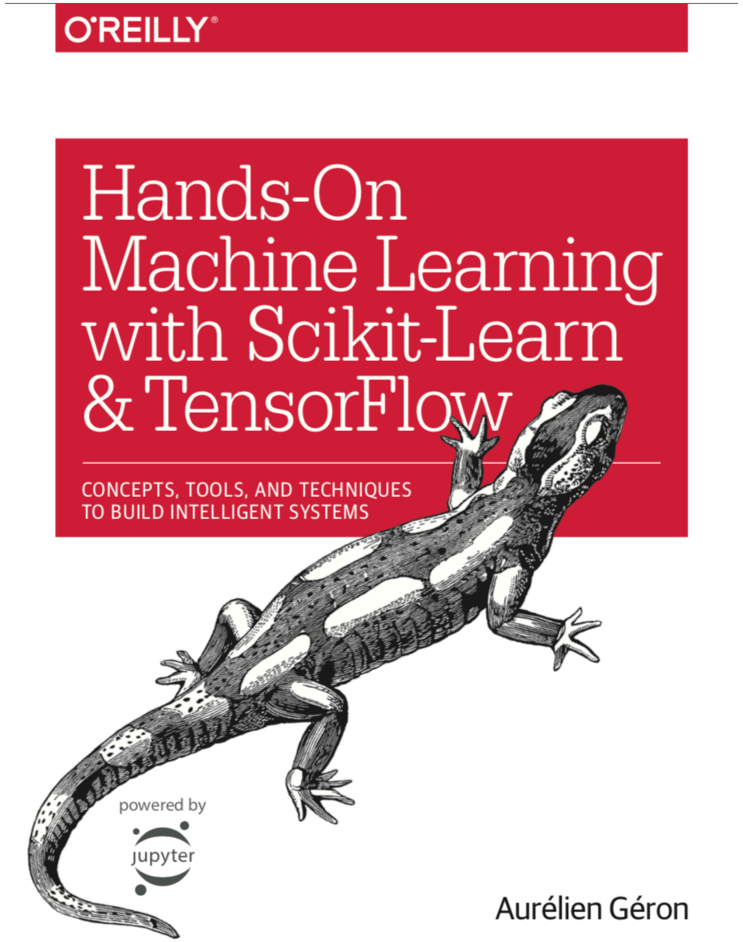

O'REILLY
Data Science
from Scratch
FIRST PRINCIPLES WITH PYTHON
Joel Grus

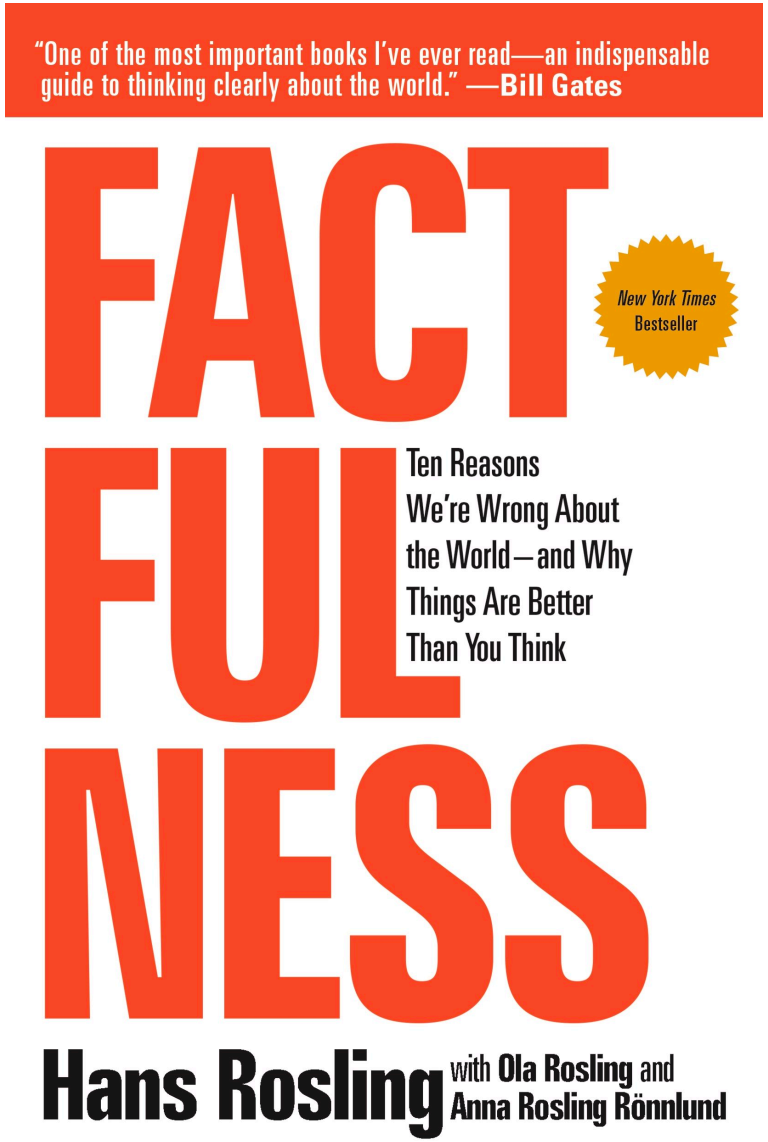

O'REILLY
2nd Edition
Python for
Data Analysis
DATA WRANGLING WITH PANDAS,
NUMPY, AND IPYTHON
powered by
jupyter
Wes McKinney

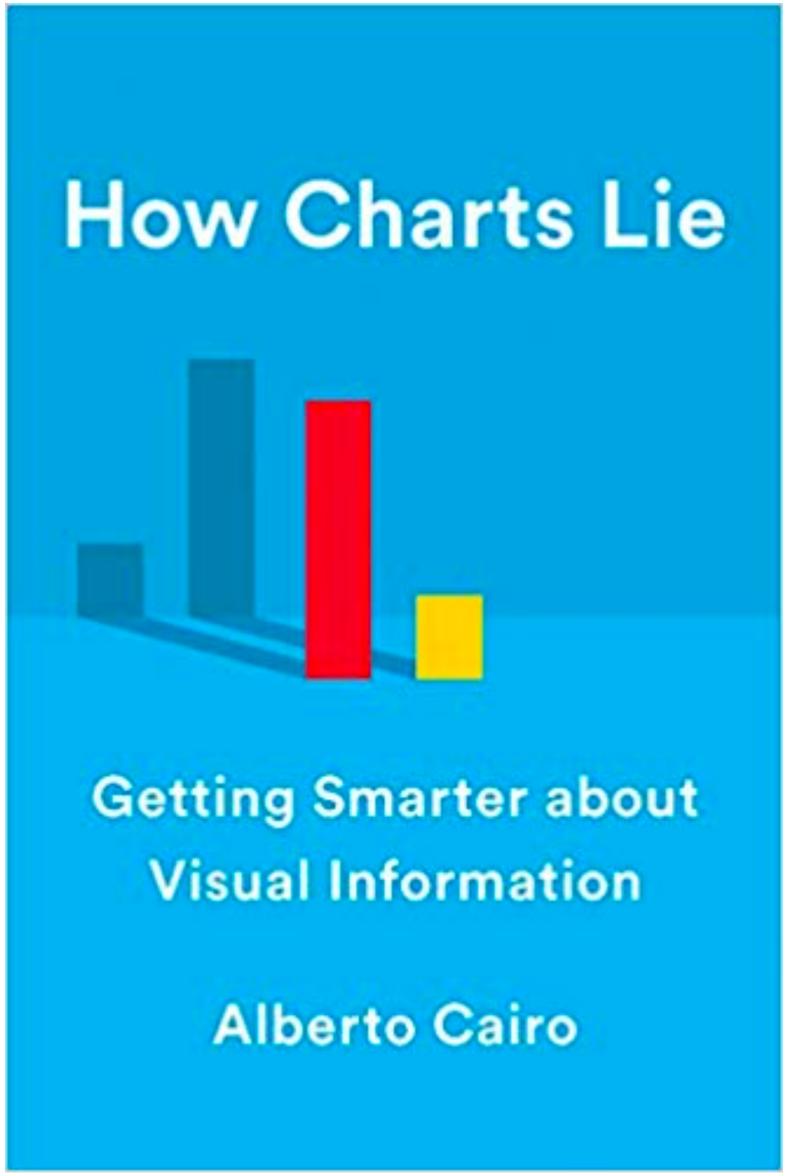

DEEP LEARNING
with Python
François Chollet
MANNING


O'REILLY
Python Data Science Handbook
ESSENTIAL TOOLS FOR WORKING WITH DATA
powered by
jupyter
Jake VanderPlas


Draft Version
MACHINE LEARNING YEARNING
Technical Strategy for AI Engineers, In the Era of Deep Learning
ANDREW NG

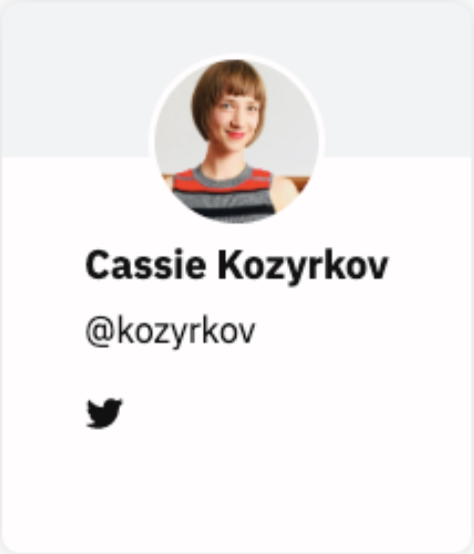

O'REILLY
Hands-On Machine Learning with Scikit-Learn & TensorFlow
CONCEPTS, TOOLS, AND TECHNIQUES TO BUILD INTELLIGENT SYSTEMS
powered by
jupyter
Aurélien Géron


"One of the most important books I've ever read—an indispensable guide to thinking clearly about the world." —Bill Gates
FACT
FUL
NESS
New York Times Bestseller
Ten Reasons We're Wrong About the World—and Why Things Are Better Than You Think
Hans Rosling with Ola Rosling and Anna Rosling Rönnlund


How Charts Lie
Getting Smarter about Visual Information
Alberto Cairo

OTHER RESOURCES

kaggle™



DataCamp

M

Towards Data Science

Sharing concepts, ideas, and codes

Oge Marques, Ph.D.
Professor

FAU
FLORIDA ATLANTIC UNIVERSITY

Computer & Electrical Engineering and Computer Science (CEECS)
777 Glades Road
Boca Raton, FL 33431-0991
tel: 561.297.3857
fax: 561.297.2800
email: omarques@fau.edu
Skype: ProfessorOge

ProfessorOgeMarques

@ProfessorOge