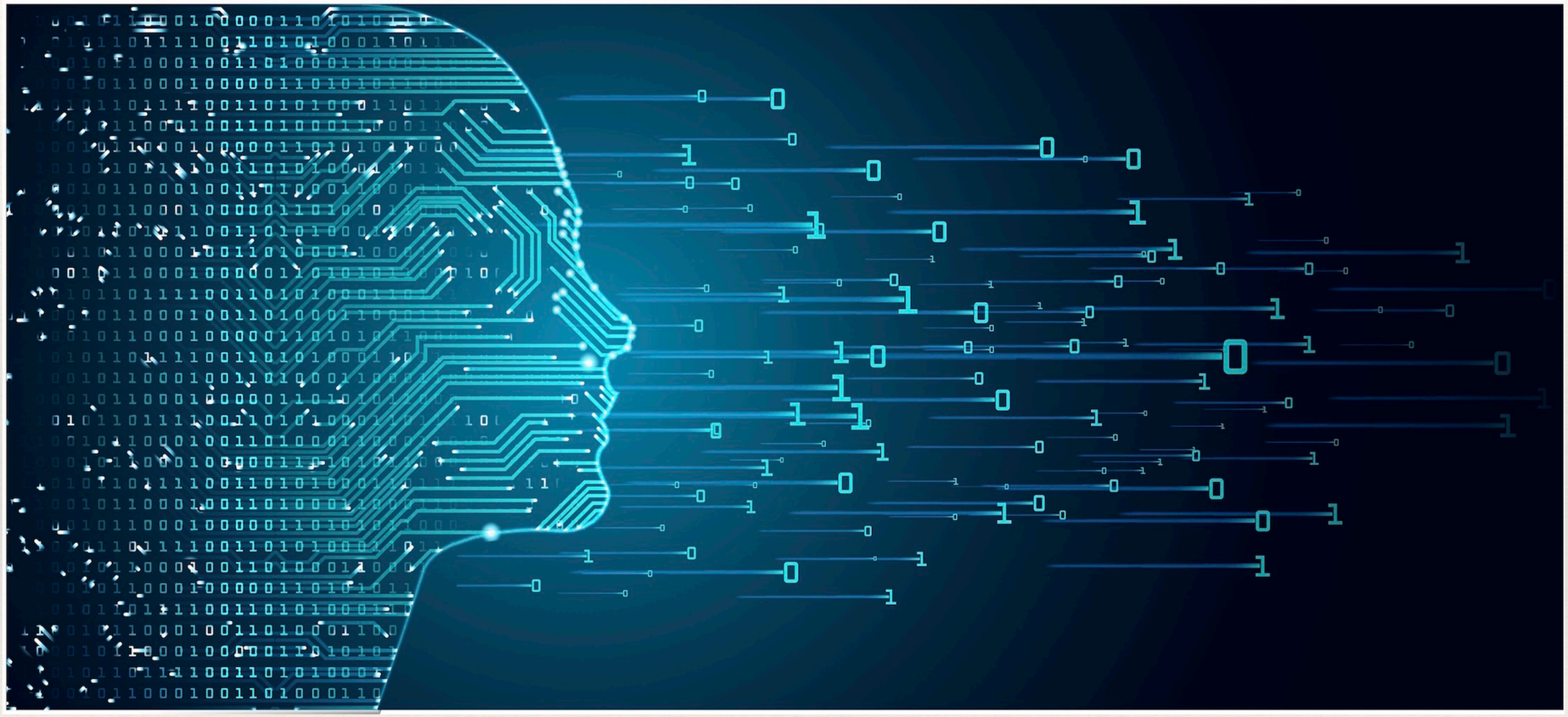
Lecture 1

Introduction

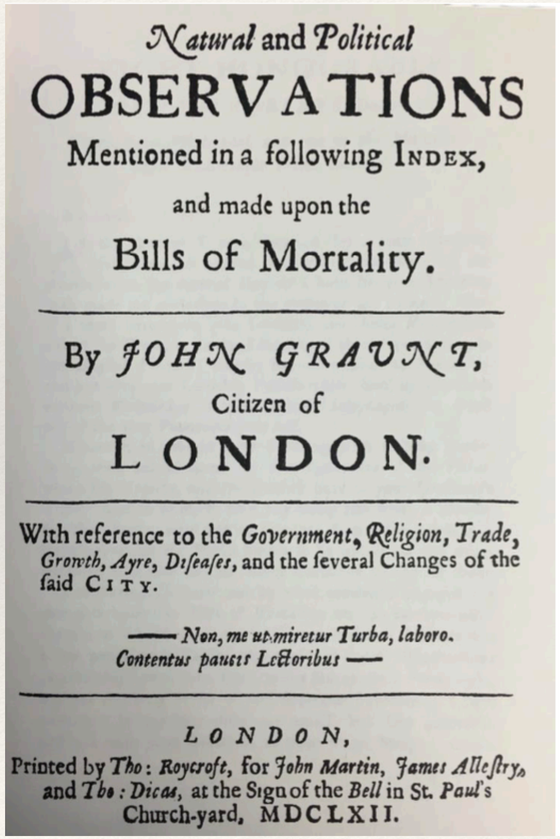
Dr. David Zmiaikou

Data Science



- Data scientist has been called "*the sexiest job of the 21st century*".
- Data science is a hot and growing field.

⏳ A bit of history…



Natural and *Political*

OBSERVATIONS

Mentioned in a following INDEX,

and made upon the

Bills of Mortality.

By JOHN GRAUNT,

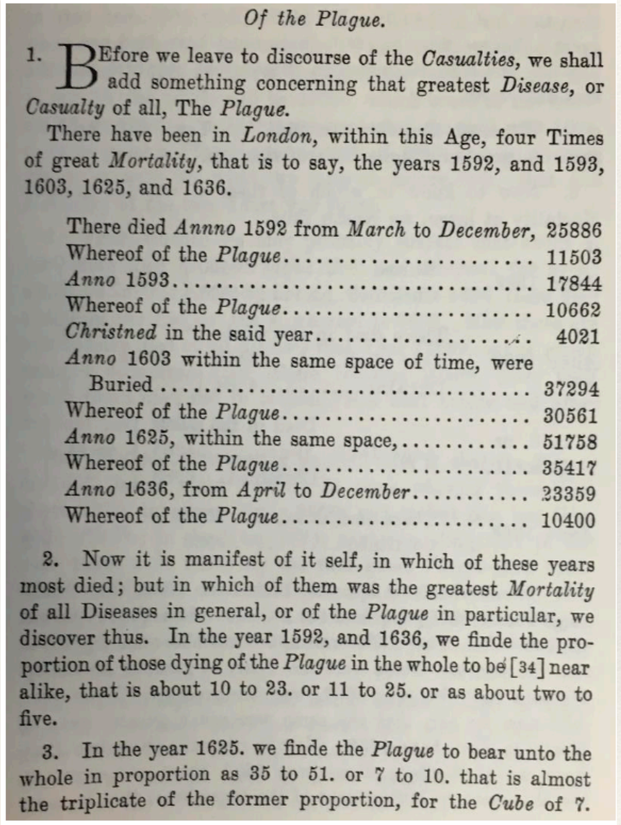
Citizen of

LONDON.

With reference to the *Government*, *Religion*, *Trade*, *Growth*, *Ayre*, *Diseases*, and the several Changes of the said CITY.

—— *Non, me ut miretur Turba, laboro.*
Contentus paucis Lectoribus ——

LONDON,
Printed by *Tho: Roycroft*, for *John Martin*, *James Allestry*, and *Tho: Dicas*, at the Sign of the *Bell* in St. *Paul's* Church-yard, MDCLXII.



Of the Plague.

1. BEfore we leave to discourse of the *Casualties*, we shall add something concerning that greatest *Disease*, or *Casualty* of all, The *Plague*.

There have been in *London*, within this Age, four Times of great *Mortality*, that is to say, the years 1592, and 1593, 1603, 1625, and 1636.

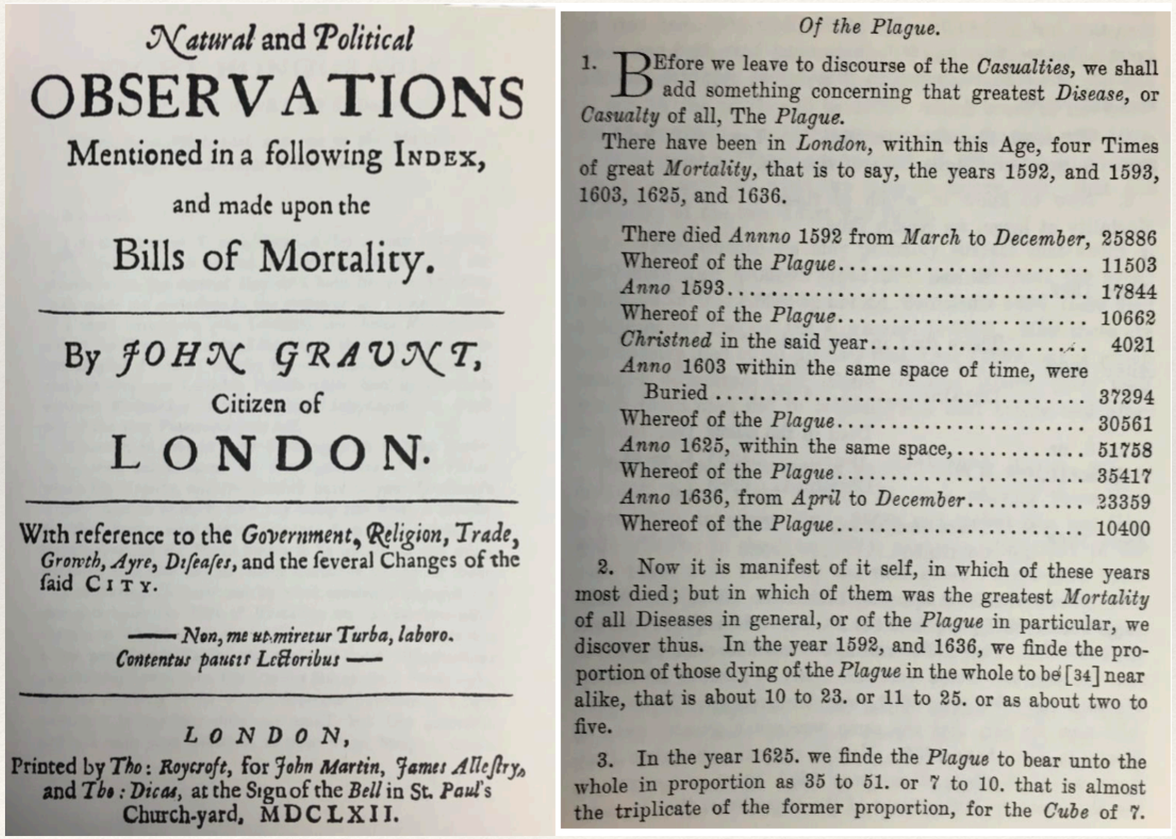
There died *Annno* 1592 from *March* to *December*,	25886
Whereof of the *Plague*.....	11503
Anno 1593.....	17844
Whereof of the *Plague*.....	10662
Christned in the said year.....	4021
Anno 1603 within the same space of time, were Buried	37294
Whereof of the *Plague*.....	30561
Anno 1625, within the same space,.....	51758
Whereof of the *Plague*.....	35417
Anno 1636, from *April* to *December*.....	23359
Whereof of the *Plague*.....	10400

2. Now it is manifest of it self, in which of these years most died; but in which of them was the greatest *Mortality* of all Diseases in general, or of the *Plague* in particular, we discover thus. In the year 1592, and 1636, we finde the proportion of those dying of the *Plague* in the whole to be [34] near alike, that is about 10 to 23. or 11 to 25. or as about two to five.

3. In the year 1625. we finde the *Plague* to bear unto the whole in proportion as 35 to 51. or 7 to 10. that is almost the triplicate of the former proportion, for the *Cube* of 7.

- The techniques for transforming data into information go back hundreds of years.
- A good starting point is 1592 with the publication of John Graunt's weekly "bills of mortality" in London.

⏳ A bit of history…



Natural and *Political*

OBSERVATIONS

Mentioned in a following INDEX,

and made upon the

Bills of Mortality.

By JOHN GRAUNT,

Citizen of

LONDON.

With reference to the *Government*, *Religion*, *Trade*, *Growth*, *Ayre*, *Diseases*, and the several Changes of the said CITY.

—— *Non, me ut miretur Turba, laboro.*
Contentus paucis Lectoribus ——

LONDON,

Printed by *Tho: Roycroft*, for *John Martin*, *James Allestry*, and *Tho: Dicas*, at the Sign of the *Bell* in St. *Paul's* Church-yard, MDCLXII.

Of the Plague.

1. BEfore we leave to discourse of the *Casualties*, we shall add something concerning that greatest *Disease*, or *Casualty* of all, The *Plague*.

There have been in *London*, within this Age, four Times of great *Mortality*, that is to say, the years 1592, and 1593, 1603, 1625, and 1636.

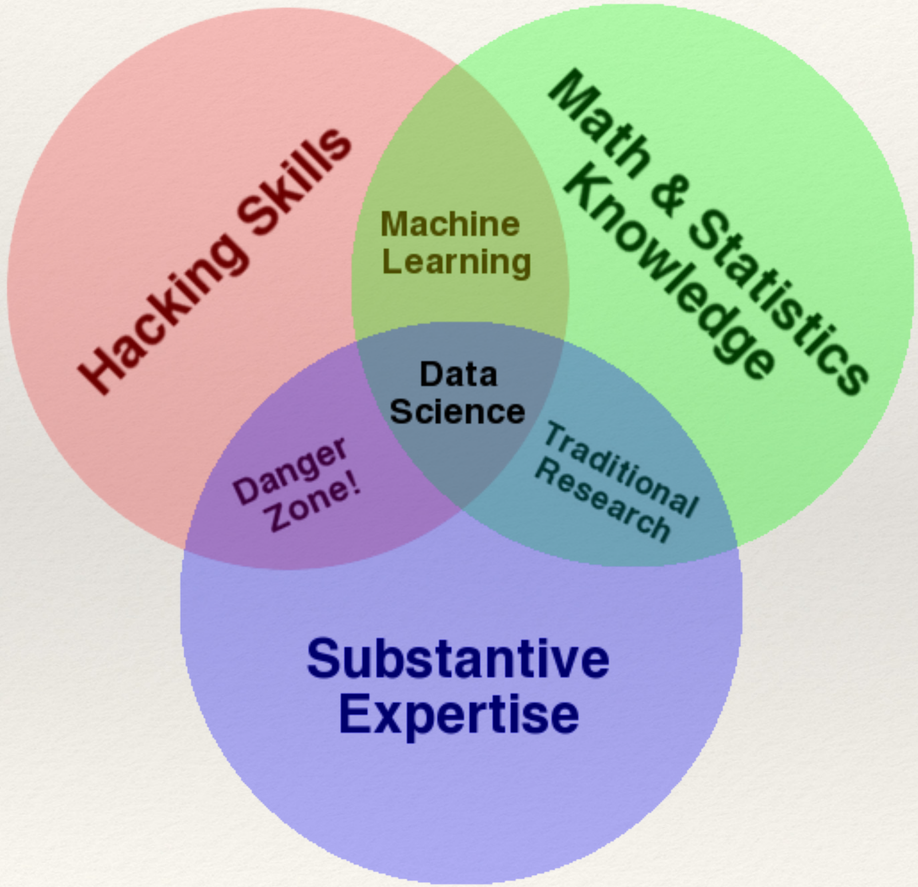
There died *Annno* 1592 from *March* to *December*,	25886
Whereof of the *Plague*.....	11503
Anno 1593.....	17844
Whereof of the *Plague*.....	10662
Christned in the said year.....	4021
Anno 1603 within the same space of time, were Buried	37294
Whereof of the *Plague*.....	30561
Anno 1625, within the same space,.....	51758
Whereof of the *Plague*.....	35417
Anno 1636, from *April* to *December*.....	23359
Whereof of the *Plague*.....	10400

2. Now it is manifest of it self, in which of these years most died; but in which of them was the greatest *Mortality* of all Diseases in general, or of the *Plague* in particular, we discover thus. In the year 1592, and 1636, we finde the proportion of those dying of the *Plague* in the whole to be [34] near alike, that is about 10 to 23. or 11 to 25. or as about two to five.

3. In the year 1625. we finde the *Plague* to bear unto the whole in proportion as 35 to 51. or 7 to 10. that is almost the triplicate of the former proportion, for the *Cube* of 7.

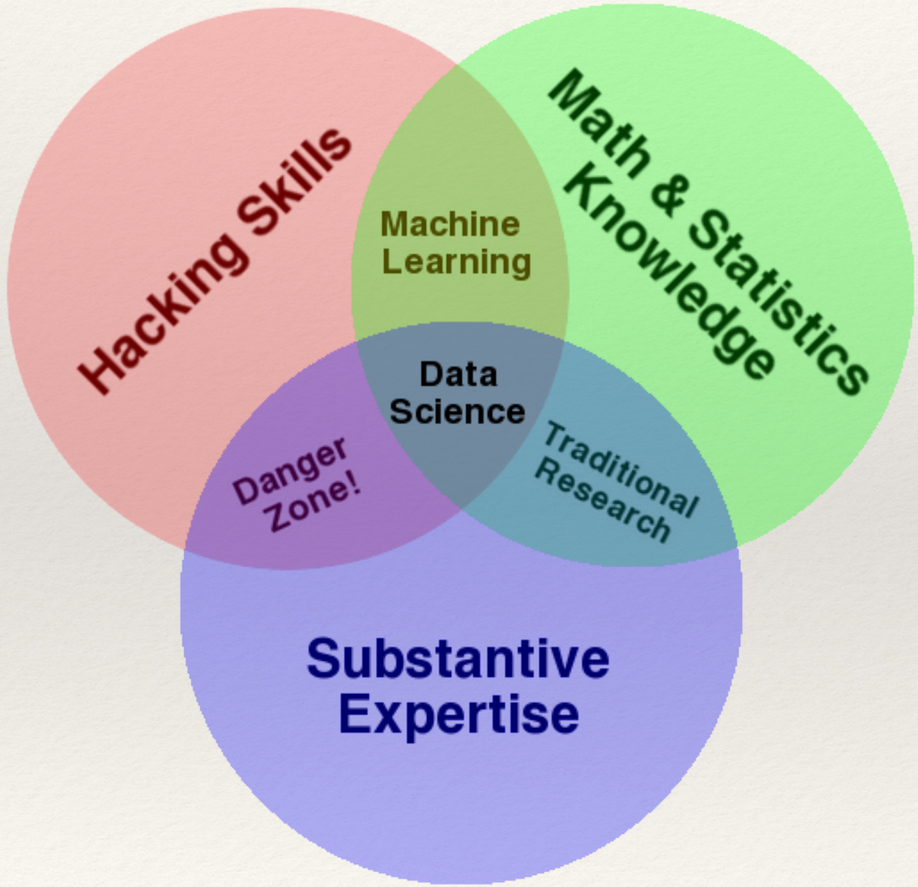
- Over the centuries, as data became larger, machines were introduced to speed up the tabulations.
- A major step: Herman Hollerith's development of punched cards and an electrical tabulating system for the United States Census of 1890.
- Also in the late 19th century, statistical methods began to develop rapidly.
  These methods have been tremendously important in interpreting data, but they were not intrinsically tied to mechanical data processing.
  Generations of students have learned to carry out statistical operations by hand on small sets of data.

$\stackrel{\text{def}}{=}$ Data Science



- Data science lies at the intersection of:
  - Hacking skills
  - Math and statistics knowledge
  - Substantive expertise

$\stackrel{\text{def}}{=}$ Data Science

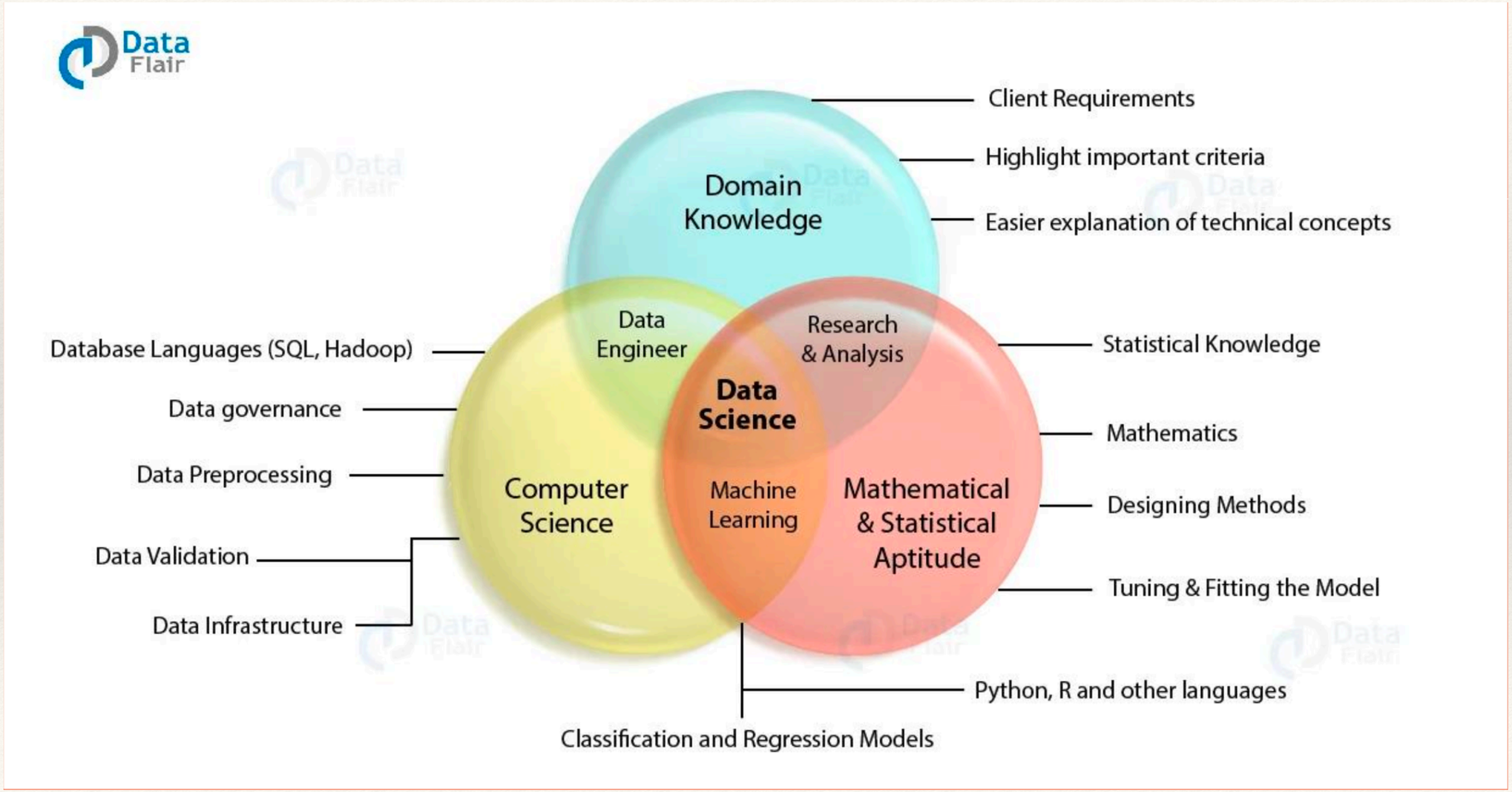


- Data science is about

  - extraction,
  - preparation,
  - visualisation,
  - maintenance of information.

It is a cross-disciplinary field which uses scientific methods and processes to draw insights from data.

$\stackrel{\text{def}}{=}$ Data Science



Example 1: A dating site

The dating site OkCupid asks its members to answer thousands of questions in order to find the most appropriate matches for them. But it also sometimes analyzes these results to figure out innocuous-sounding questions you can ask someone to find out how likely someone is to sleep with you on the first date.

Example 2: A social network

Facebook asks you to list your hometown and your current location, ostensibly to make it easier for your friends to find and connect with you. But it also analyzes these locations to identify global migration patterns and where the fanbases of different football teams live.

Example 3: A large retailer

As a large retailer, Target tracks your purchases and interactions, both online and in store. And it uses the data to predictively model which of its customers are pregnant, to better market baby-related purchases to them.

Example 4: Presidential campaigns

In 2012, the Obama campaign employed dozens of data scientists who data-mined and experimented their way to identifying voters who needed extra attention, choosing optimal donor-specific fundraising appeals and programs, and focusing get-out-the-vote efforts where they were most likely to be useful.

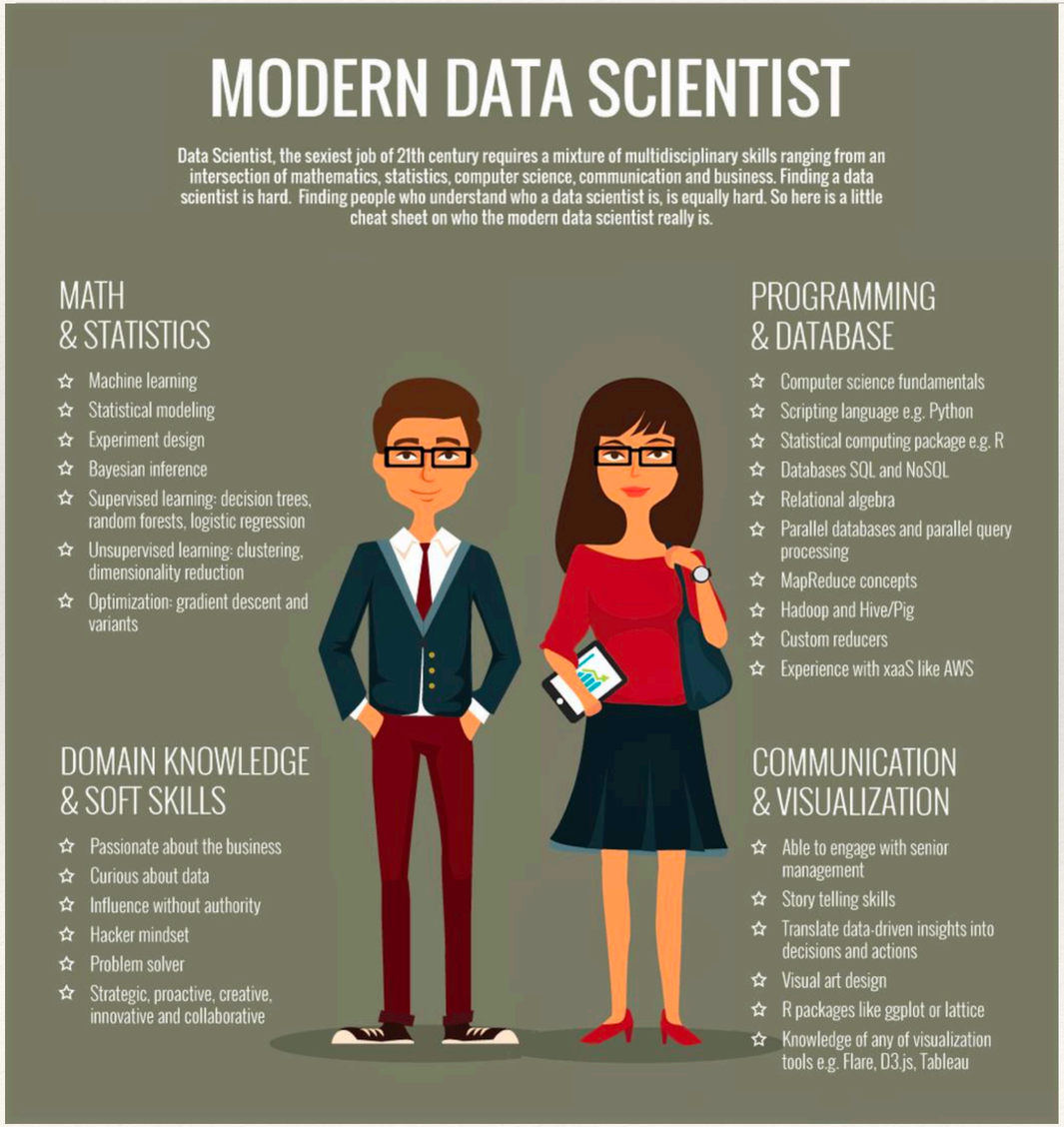
In 2016, the Trump campaign tested a staggering variety of online ads and analyzed the data to find what worked and what didn't.

Examples 5: Many good goals

Some data scientists also occasionally use their skills for good:

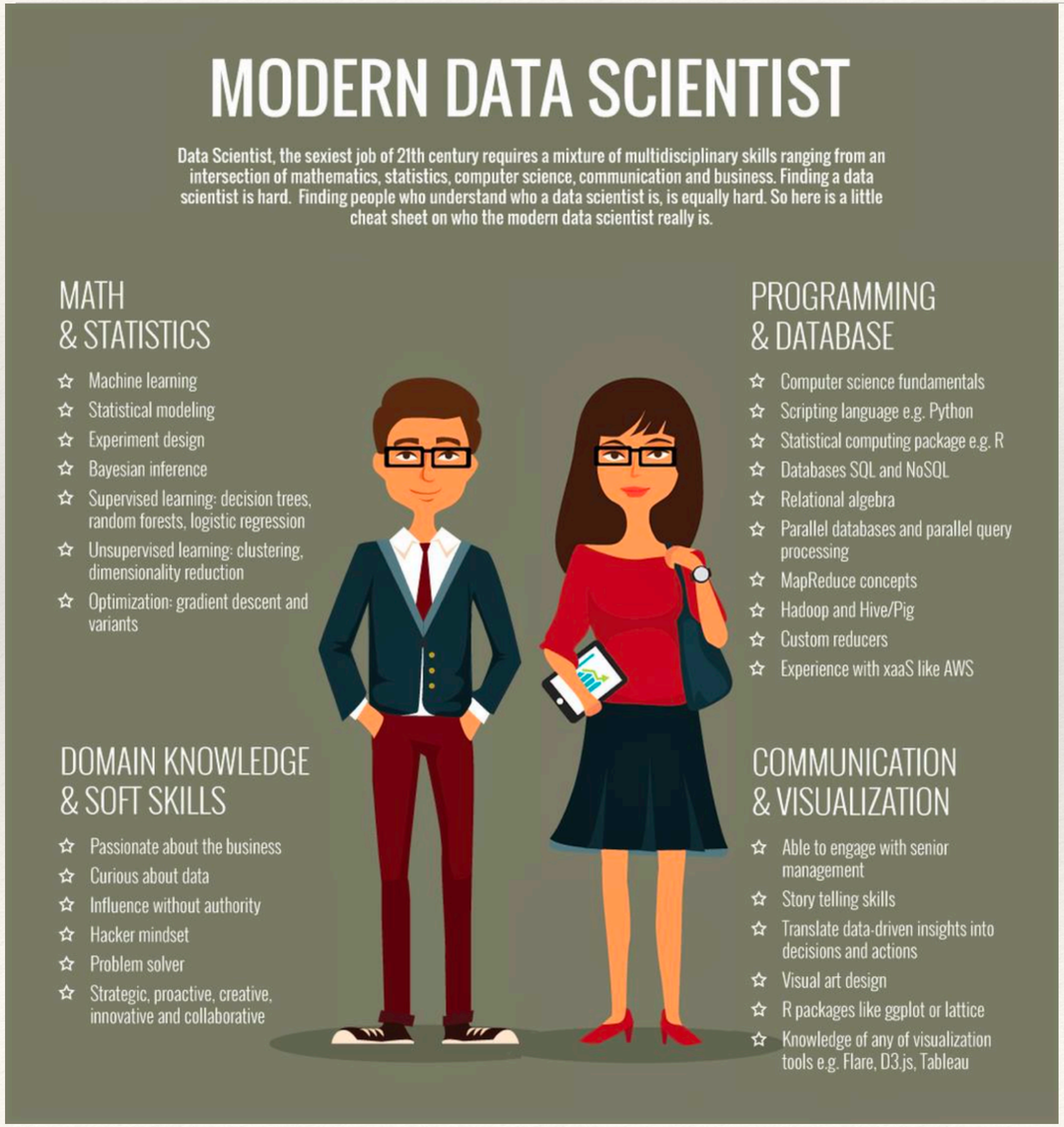
— using data to make government more effective,
— to help the homeless,
— to improve public health,
— and many more.

$\stackrel{\text{def}}{=}$ Data Scientist



- There's a joke that says a data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.
- We'll say that a **data scientist** is someone who extracts insights from messy data.

Data Scientist: Skills

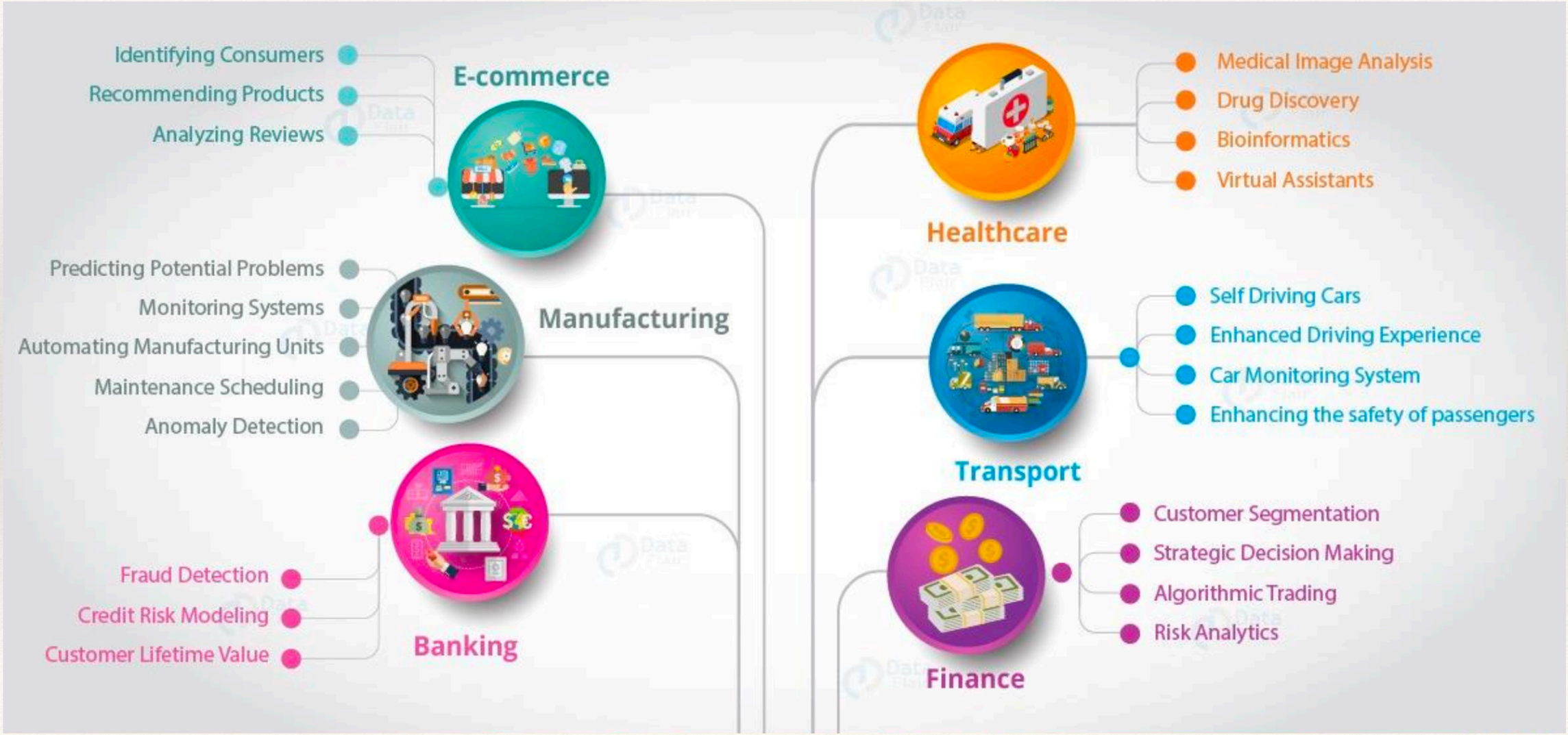


- Data scientists have

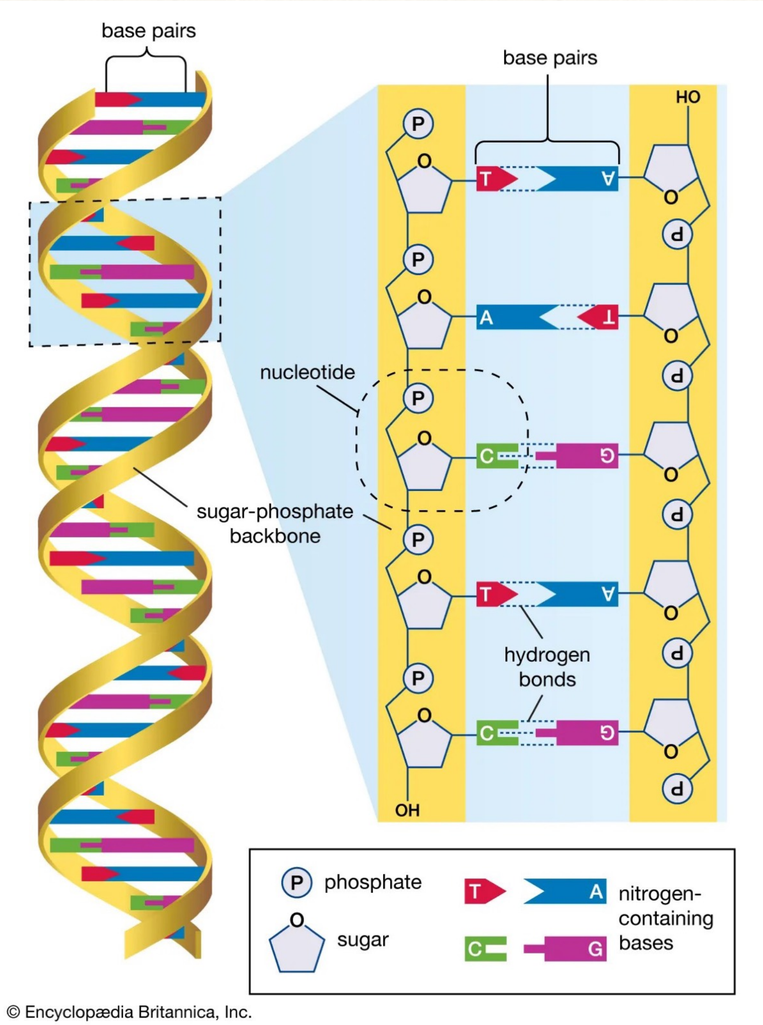
  — the technical skills to solve complex problems,

  — the curiosity to explore what problems need to be solved.

Data Science: Applications



Data Science: Applications



- Nowadays, it is common to have data sets that are so large they can be processed only by machine.
- The settings where such data arise are diverse:

  — the genome,
  — satellite observations of Earth,
  — entries by Web users,
  — sales transactions, etc.

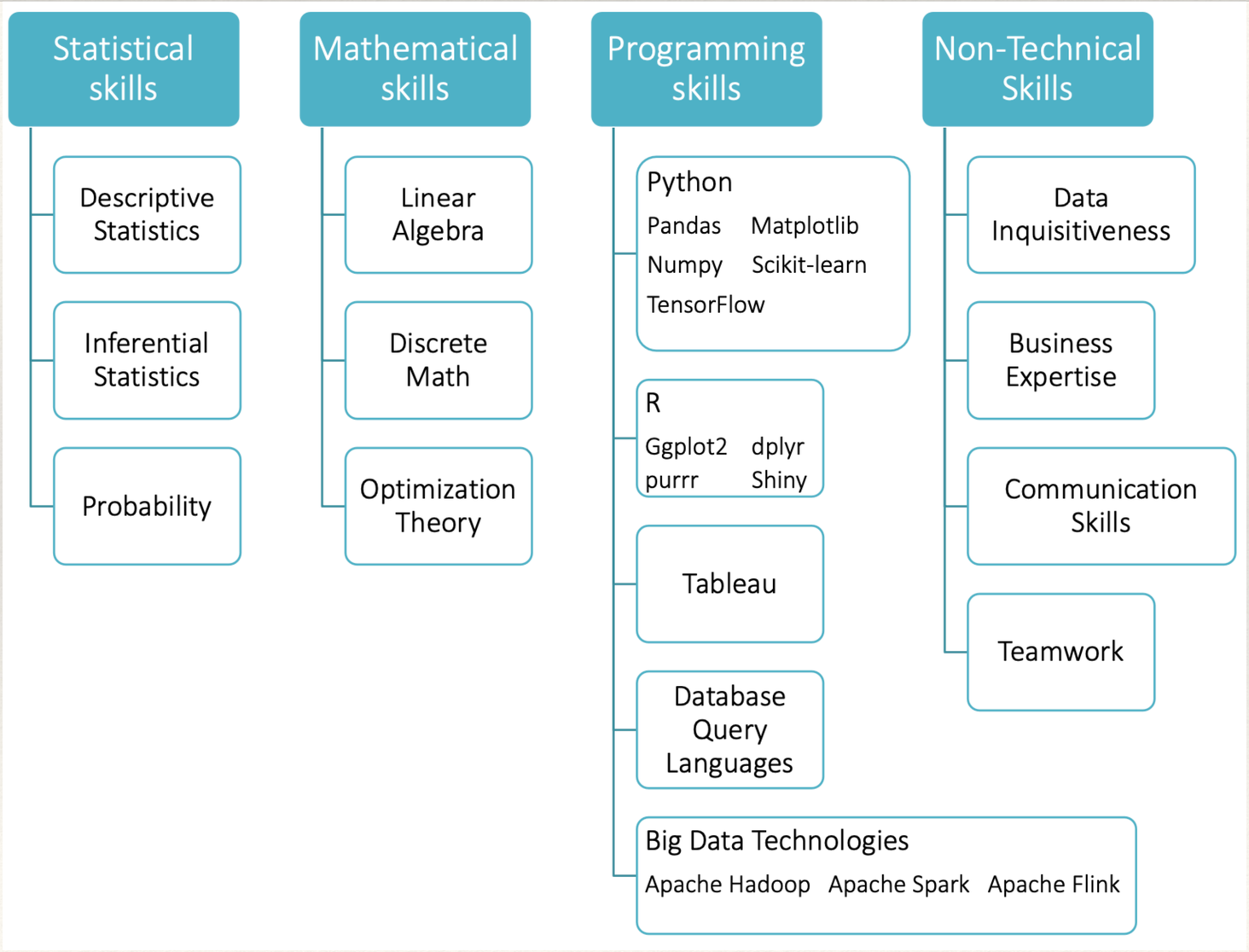
Data Science: Some Techniques

- There are new opportunities for finding and characterising patterns using techniques such as

  — data mining,
  — machine learning,
  — data visualisation, and so on.

  Such techniques require computer processing.

Data Scientist: Skills

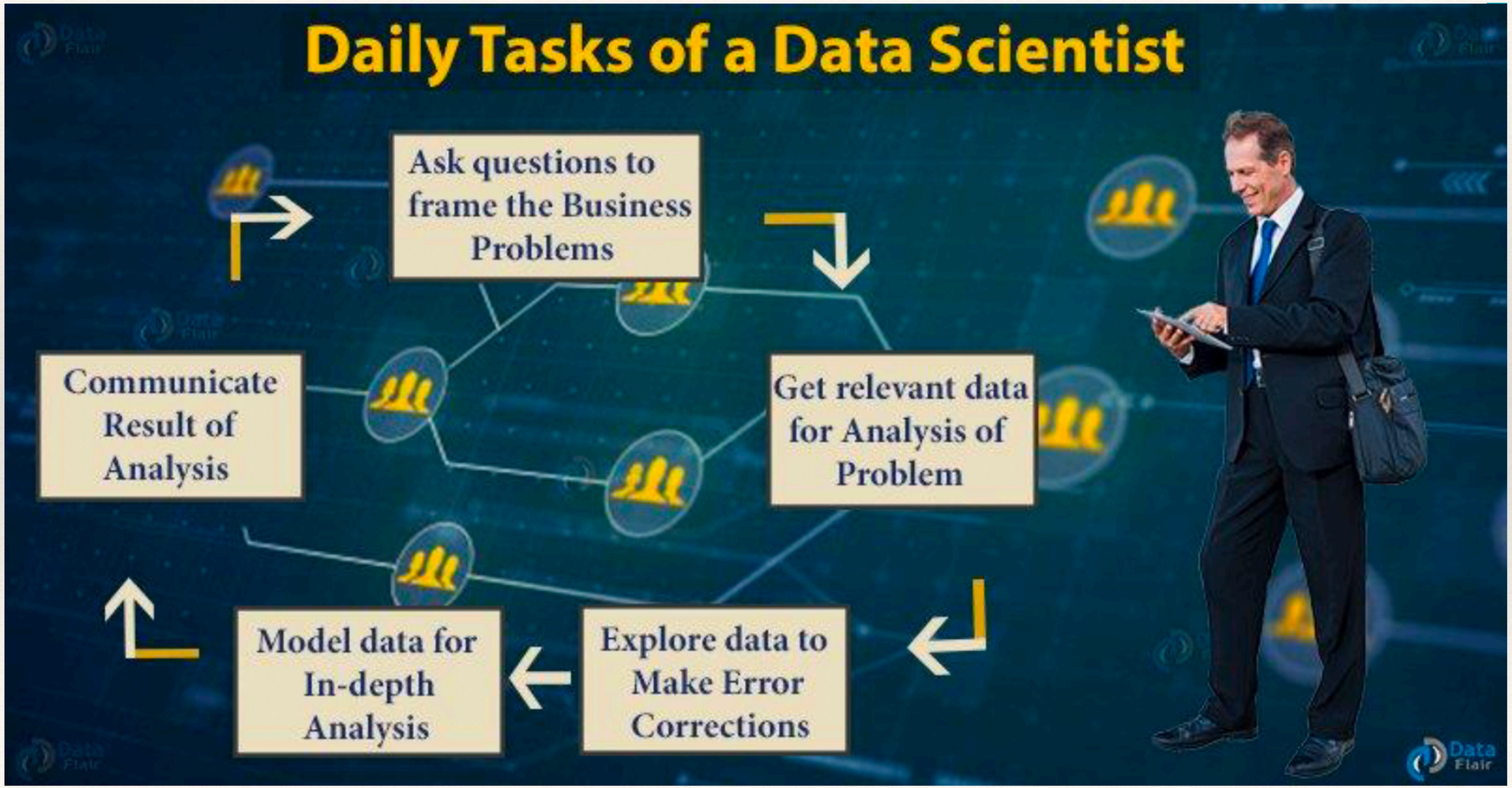


Why will we use Python?

- Python has several features that make it well suited for learning (and doing) data science:
  - It's free.
  - It's relatively simple to code in (and, in particular, to understand).
  - It has lots of useful data science–related libraries.

Data Science Process

Data Science Process



Step 1. Ask Questions to Frame the Business Problem

For analysis of the problem, you need to start by asking a lot of questions:

— Who are the target market and the customers?
— How do you approach the target market?
— How does the sales process look currently?
— What information do you have about the target market?
— How can we identify customers who are more likely to buy our product?

Step 2. Get Relevant Data for Analysis of the Problem

Before gathering the data, you should ask if the data required is already available with the company?

You find that most of the customer — related data is available in the company's Customer Relationship Management (CRM) software, managed by the sales team.

If you think the data available is not sufficient, then you must make arrangements to collect new data. You can even take feedback from your visitors and customers by displaying or distributing a feedback form. I agree, that is a lot of engineering work and requires time and effort.

The data you have collected is actually 'raw data' that contains errors and missing values. So before you analyze the data, you need to clean the data.

Step 3. Explore the Data to Make Error Corrections

Exploring the data is actually cleaning and organising it. More than 70-80% of the data scientist's time is spent on this process.

First, you need to make sure the data is clean and free from errors. This is the most important step in the process which requires patience and focus.

Then, you start answering these questions:
— Are there missing values in the data i.e. are there customers without their contact numbers?
— Are there any invalid values? If there are, how can you fix it?
— Are there multiple datasets? Is merging datasets a good choice? If yes, then how should you merge them?

Once you have uncovered missing and false values in your data, it is ready for analysis. Remember that getting the wrong insights from the data is worse than having no insight at all.

Step 4. Model the Data for In-depth Analysis

In this step, you analyse the data to get information from it. Analysing the data requires applying various algorithms that will draw out meaning from it:

- Build a model of the data to answer the question.
- Validate the model against the data collected.
- Usage of various visualisation tools to present data.
- Perform the necessary algorithms and statistical analysis.
- Compare results against other techniques and sources.

However, answering these questions will only give you hints and hypotheses.

Data modelling is a simple way to approximate data in a proper equation that the machine understands. You should be able to make predictions based on the model.

You might have to try several models in order to find the best fit.

Step 5. Communicate the Results of the Analysis

You need to effectively communicate the results of the problem specified previously:

- Graph or chart the information for presentation with tools: R, Python, Tableau, Excel.
- Use “storytelling” to fit the results.
- Answer the various follow-up questions.
- Present data in different formats — reports, websites.

Thank you!