



# Applied Social Data Science - Coding Camp

Trajche Panov, PhD

Teaching Fellow

September 2 - 6, 2024



# Introductions



My  
Background



My research

# Introductions



Who are we?



Background and  
current course



Coding familiarity?



► Expectations with  
the program/  
Coding camp week

# Schedule

► Monday: Introduction (Maxwell Theatre)

► Wednesday: More R: Good practices (Maxwell Theatre)

► Friday: How to report and share results (Maxwell Theatre)

Monday

Tuesday

Wednesday

Thursday


Friday

10:00 AM – 12:00 PM

► Tuesday: R basics (Maxwell Theatre)

► Thursday: Python basics (Maxwell Theatre)

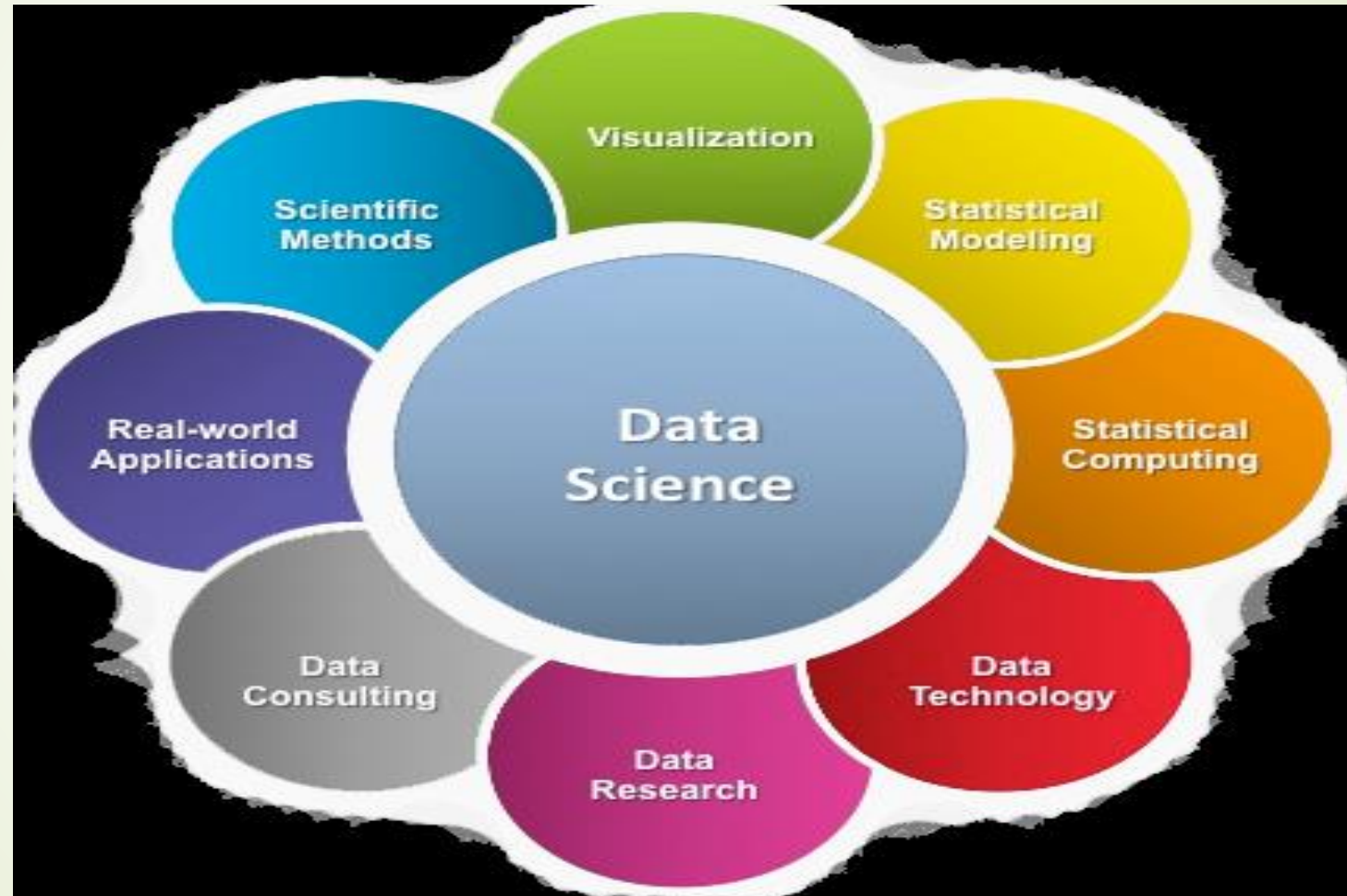
15min break?



# Today's class

- What is data science?
- Quantitative Programming Environments: R and Python
- Expectations

# What is Data Science?



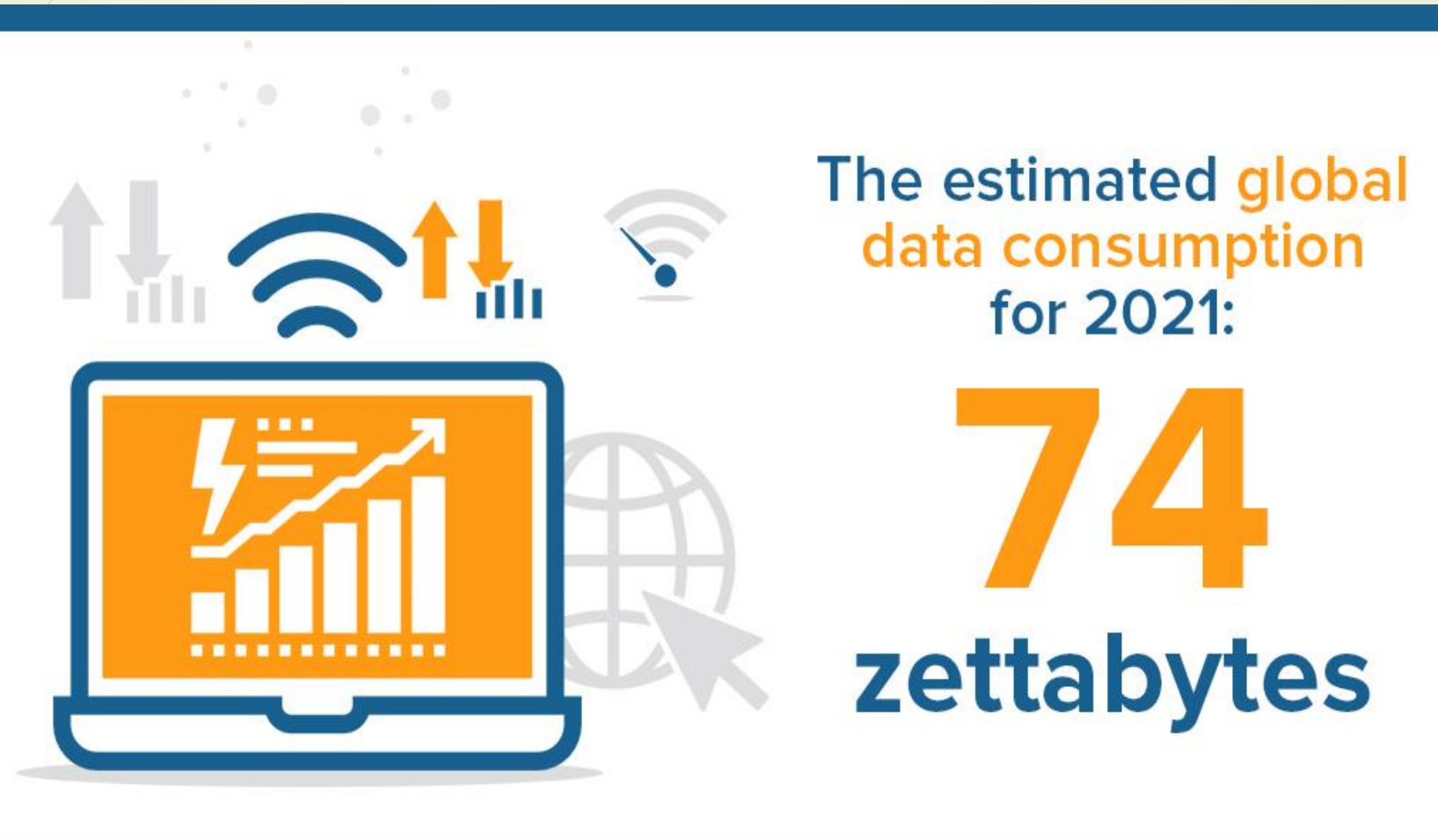


# What is Data Science?

- ▶ 'The science of learning from data' Donoho, 2017.
- ▶ statistics + computer science
- ▶ data mining, data analysis, knowledge discovery...
- ▶ Involves principles, processes, and methods for identifying and understanding phenomena via the automated or semi-automated analysis of data.
- ▶ Used for:
  - – Better decisions
  - – Predictive analysis
  - – Pattern discoveries, etc



# Big Data?!



Source: IDC & Statista, 2020



# Really big data

## A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion - fuelled by internet of things and the use of connected devices - are hard to comprehend, particularly when looked at in the context of one day

**500m**  
tweets are sent every day  
Twitter



**4PB**  
of data created by Facebook, including

350m photos  
100m hours of video watch time  
Facebook Research

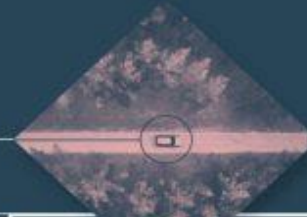
320bn  
emails to be sent each day by 2021

**294bn**  
billion emails are sent  
Facebook Group

306bn  
emails to be sent each day by 2020

**3.9bn**  
people use emails

**4TB**  
of data produced by a connected car  
Ford



### DEMYSTIFYING DATA UNITS

From the more familiar "KB" or "megabyte", larger units of measurement are more frequently being used to explain the masses of data.

Unit	Value	Size
b	0 or 1	1/8 of a byte
B	8 bits	1 byte
KB	1,000 bytes	1,000 bytes
MB	1,000 <sup>3</sup> bytes	1,000,000 bytes
GB	1,000 <sup>3</sup> bytes	1,000,000,000 bytes
TB	1,000 <sup>3</sup> bytes	1,000,000,000,000 bytes
PB	1,000 <sup>3</sup> bytes	1,000,000,000,000,000 bytes
EB	1,000 <sup>3</sup> bytes	1,000,000,000,000,000,000 bytes
ZB	1,000 <sup>3</sup> bytes	1,000,000,000,000,000,000,000 bytes
YB	1,000 <sup>3</sup> bytes	1,000,000,000,000,000,000,000,000 bytes

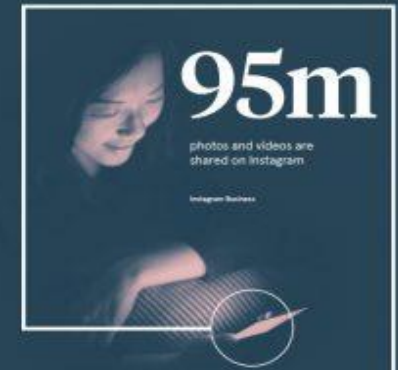
\*A lowercase "b" is used as an abbreviation for bits, while an upper case "B" represents bytes.

**65bn**  
messages sent over WhatsApp and two billion minutes of voice and video calls made  
Facebook



**463EB**  
of data will be created every day by 2025  
IAC

**95m**  
photos and videos are shared on Instagram  
Instagram Business



**28PB**  
to be generated from wearable devices by 2020  
Statista



### ACCUMULATED DIGITAL UNIVERSE OF DATA

4.4ZB  
2013  
PwC

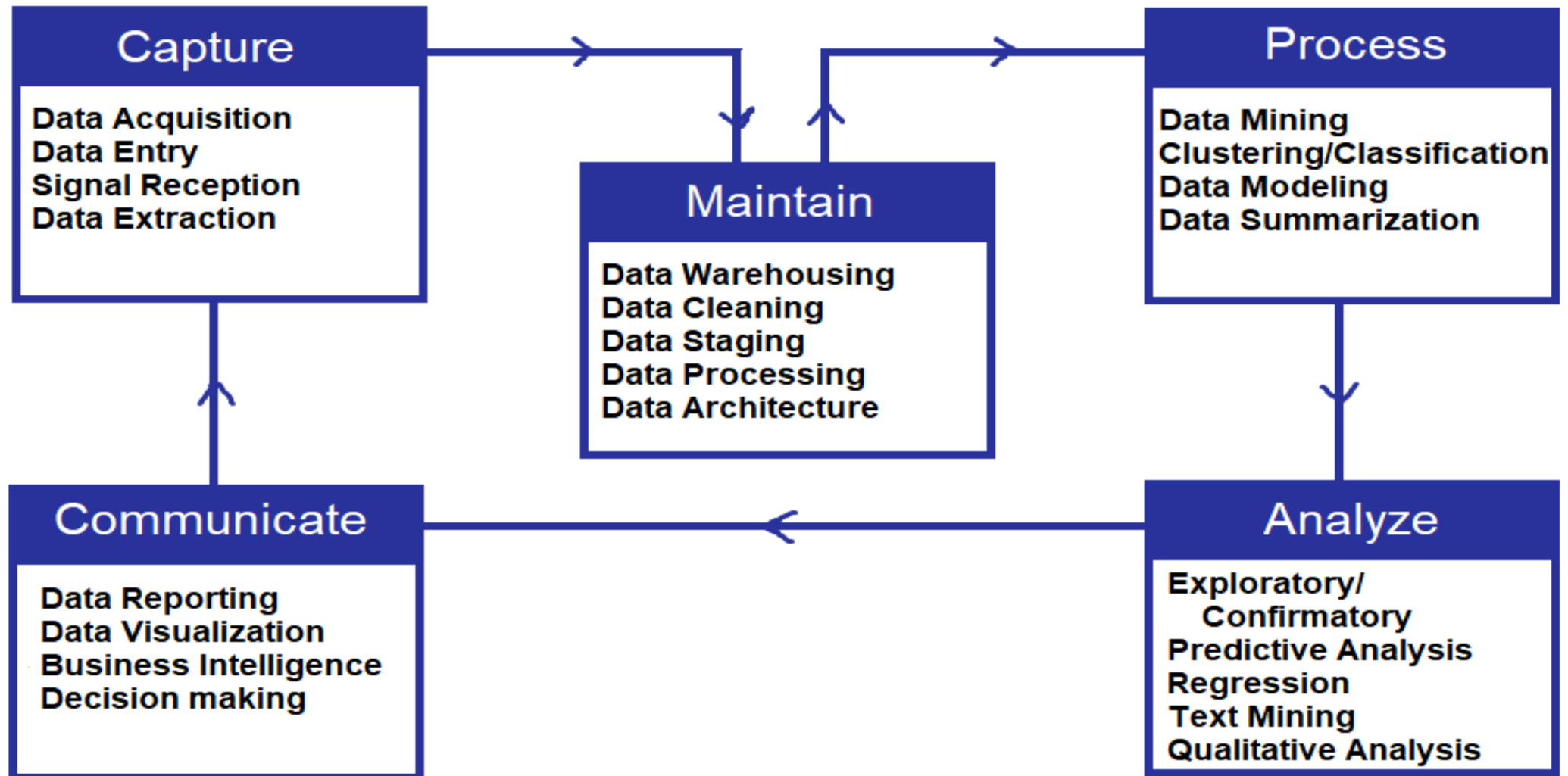
44ZB  
2020

Searches made a day  
5bn

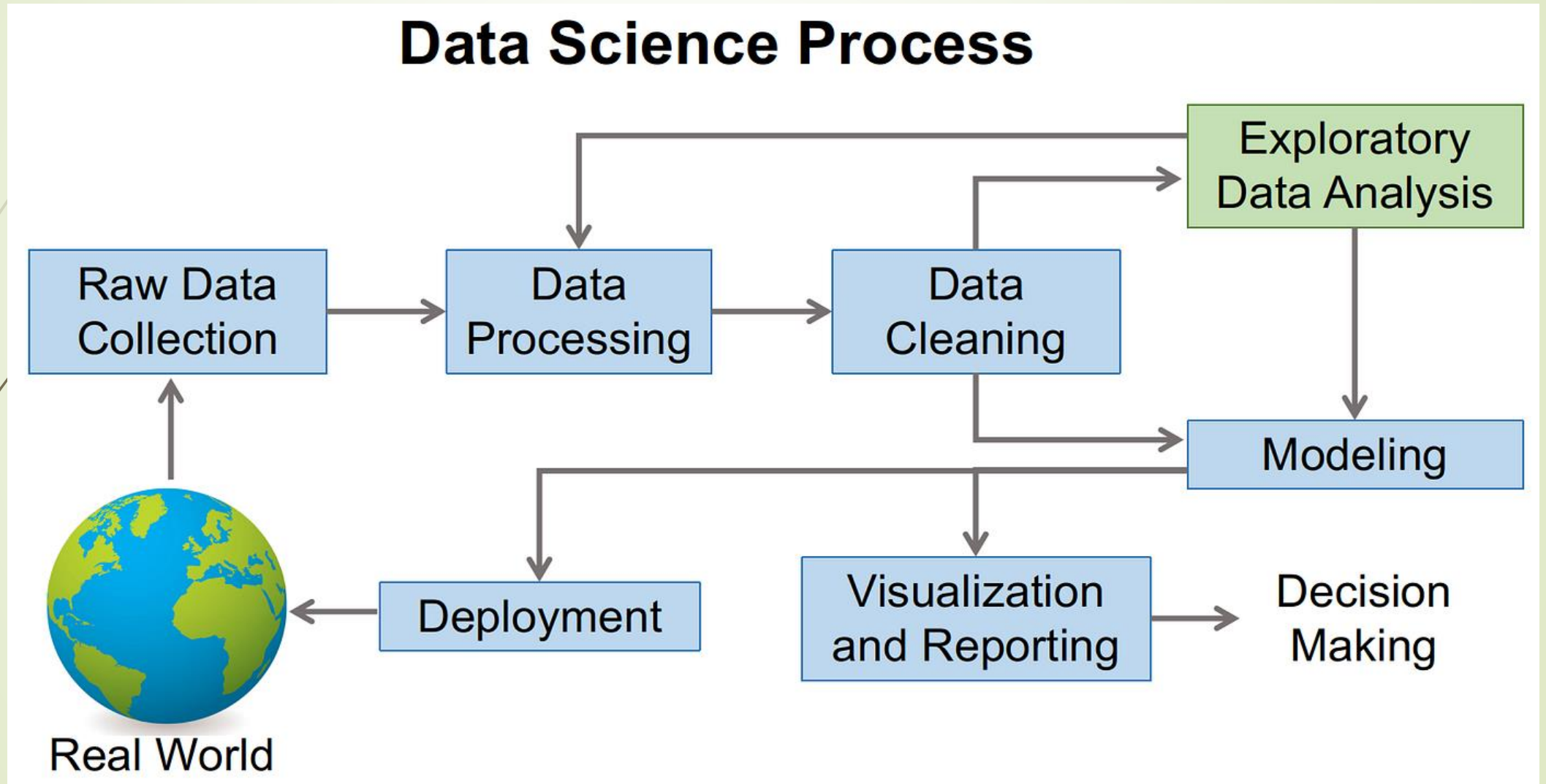
Searches made a day from Google  
3.5bn  
Smart Insights



# Data Science



# Data Science Process



# A brief history...

- ▶ 'All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.'
- ▶ The Future of Data Analysis, 1962.



Figure: John Tukey,  
1915-2000





# A brief history

- ▶ 'Four major influences act on data analysis today:
  - ▶ 1. The formal theories of statistics
  - ▶ 2. Accelerating developments in computers and display devices
  - ▶ 3. The challenge, in many fields, of more and ever larger bodies of data
  - ▶ 4. The emphasis on quantification in an ever wider variety of disciplines'
- ▶ (Tukey, 1962!)



# Timeline:

- ▶ 1960s - 1980s: advances in computer technology allow for new methods in processing and analysing data
- ▶ 1977: Exploratory Data Analysis, John Tukey
- ▶ 1990s: 'data mining' and 'knowledge discovery' emerge as terms for finding patterns in increasingly large datasets
- ▶ 1996: 'data science' included for first time in International Federation of Classification Societies (IFCS) conference title
- ▶ 2000s: analytics becomes increasingly important to businesses, 'big data' becomes a thing



# The job market

- 'I keep saying the sexy job in the next ten years will be statisticians [...] The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades.
- Hal Varian, Google Chief Economist, Jan. 2009



# Things to keep in mind

- ▶ Statistics: the mathematics associated with inference
- ▶ Data science: the practices associated with working with data
- ▶ Not everyone agrees with this distinction...
- ▶ Data science and statistics are essentially the same, but in practice they are coming to mean different things

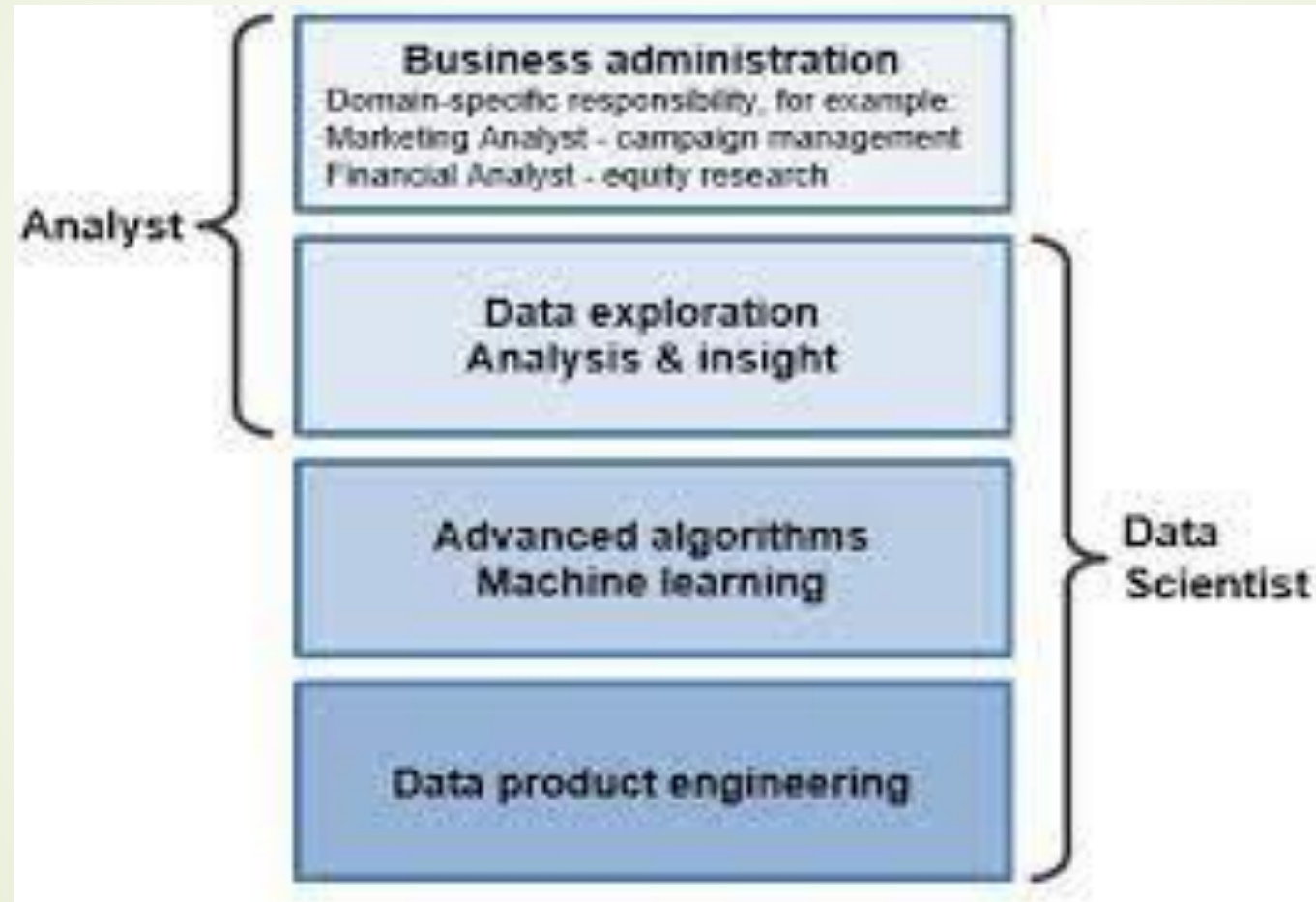


**Statistician**



**Data Scientist**

# Analyst vs. Data Scientist





# Data Science vs Data Analytics



## ► ► Data Science

Involves principles and methods for identifying and understanding phenomena via the automated or semi-automated analysis of data - predict future outcomes (broader field)

► Programming, statistics, machine learning and algorithms towards combining, preparing and examining large datasets

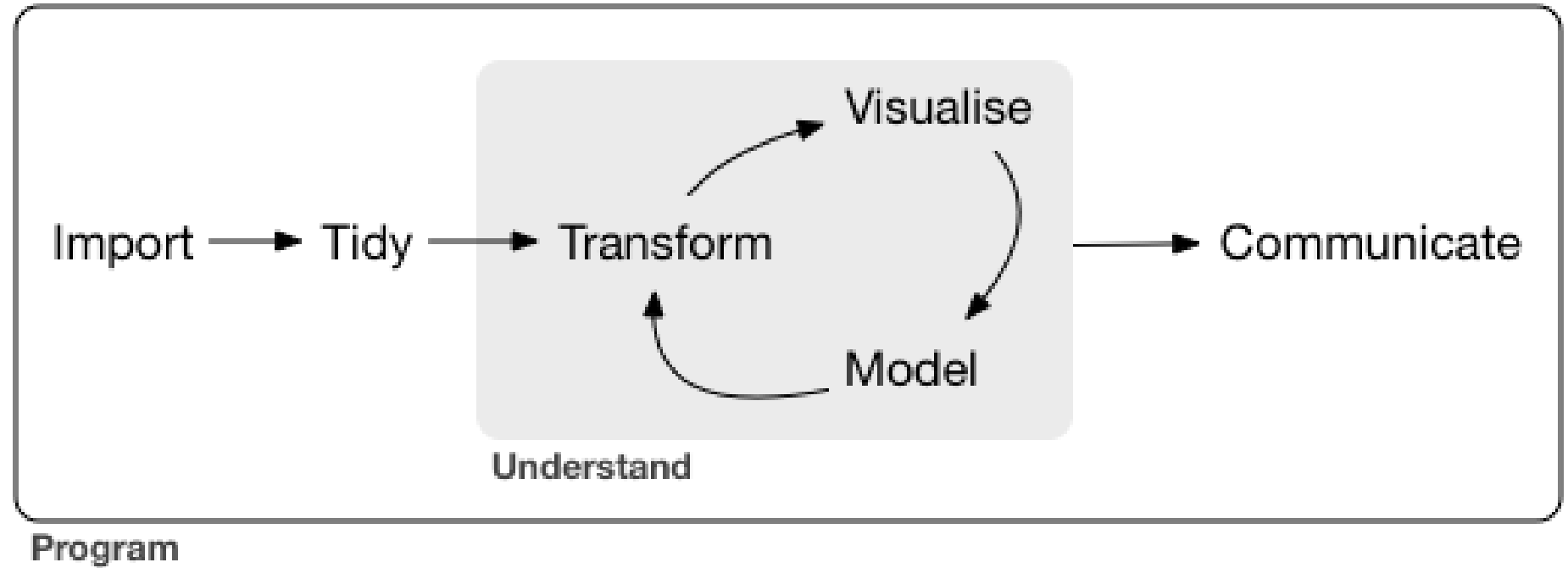
## ► ► Data Analytics

Analyses data to gain insights and inform decisions - past data for present decisions, specific questions.



# Scientific vs. Engineering mindsets

- ▶ the scientific mindset seeks to understand the underlying process (generative modeling)
- ▶ the engineering mindset looks to find the best prediction (predictive modeling)
- ▶ In the social sciences, we often want to understand what's inside the 'black box', but not all data science methods are designed for this.



Data science is a process

- Figure: Data science tools and workflow, c/o Hadley Wickham (R for Data Science)

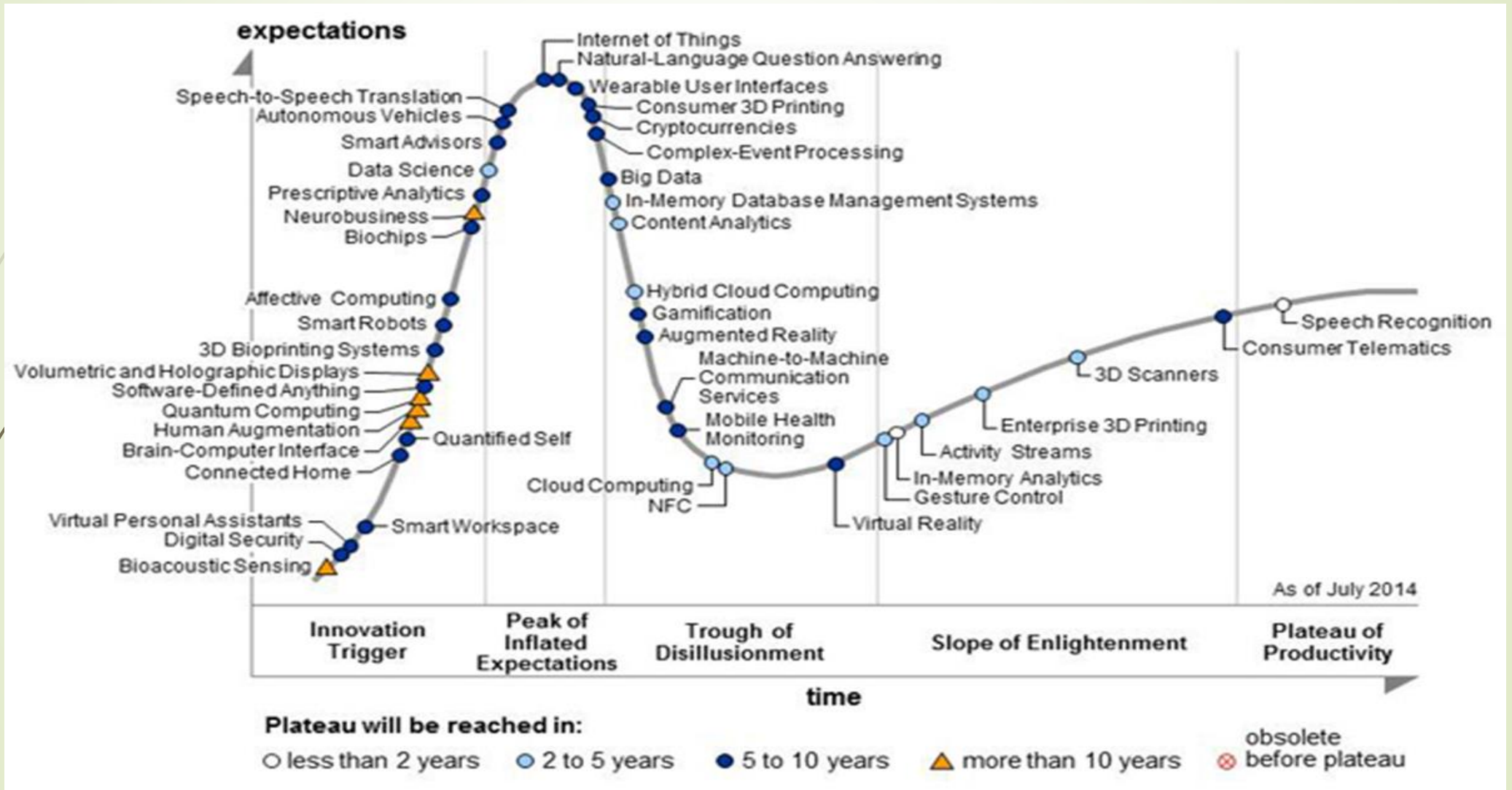


# Six Divisions

- The activities of 'Greater Data Science' are classified into six divisions (Donoho,2017):
- 1. Data Gathering, Preparation, and Exploration
- 2. Data Representation and Transformation
- 3. Computing with Data (several languages!)
- 4. Data Modeling (generative vs. predictive models)
- 5. Data Visualization and Presentation
- 6. Science about Data Science



# Gartner's 2014 Hype Cycle





# Data Scientists

The Sexiest Job of the 21st Century

- ▶ They find stories, extract knowledge.  
They are not reporters



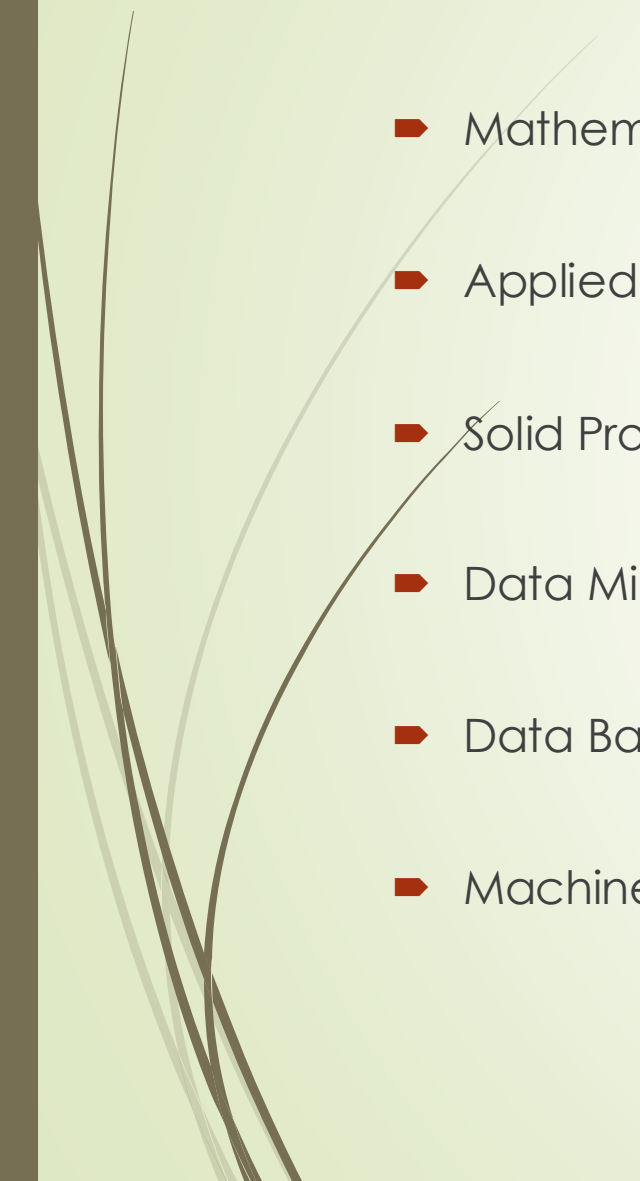


# DSperson

- Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions



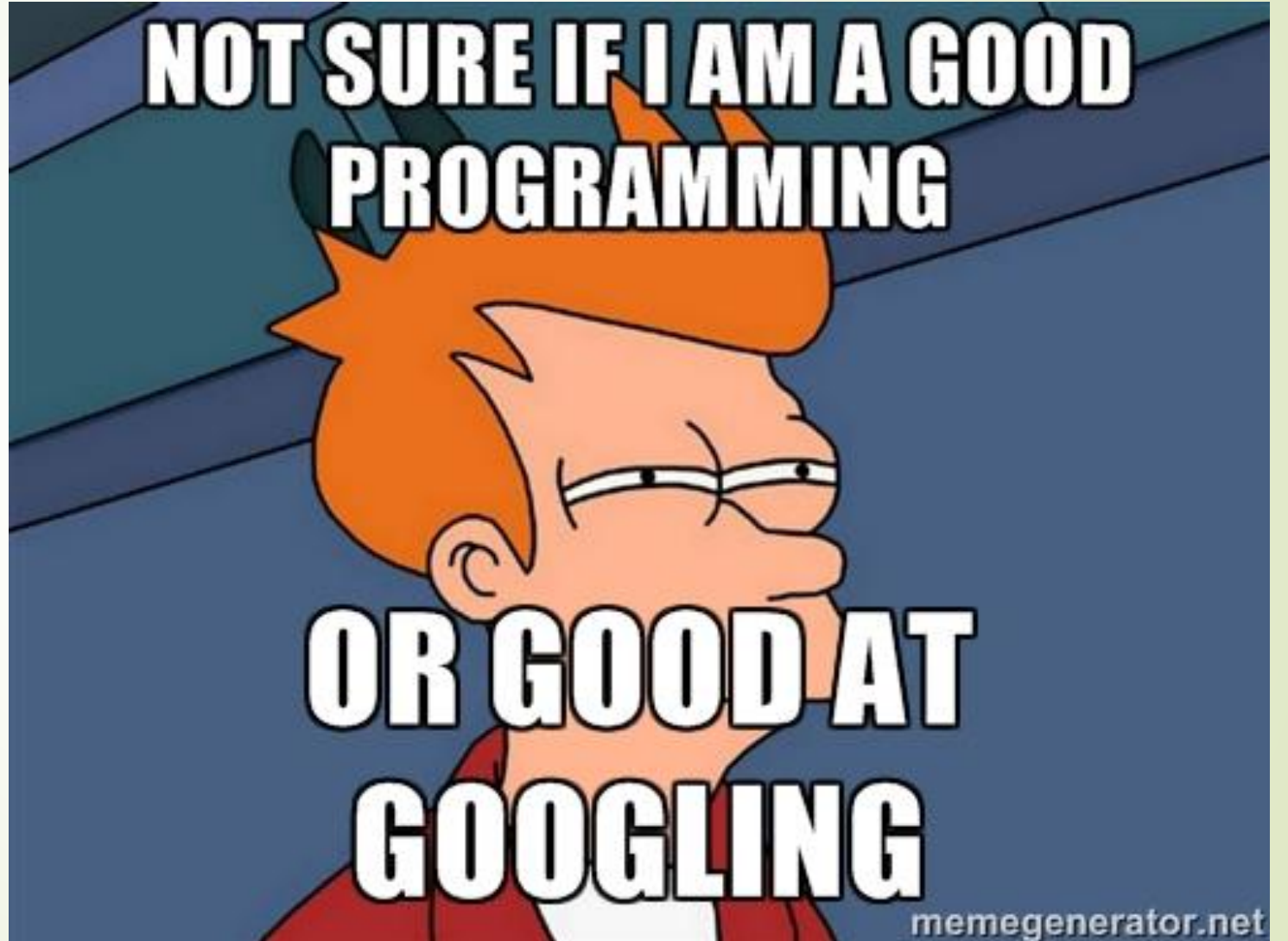
# Concentration in Data Science

- Mathematics and Applied Mathematics
  - Applied Statistics/Data Analysis
  - Solid Programming Skills (R, Python, Julia, SQL)
  - Data Mining
  - Data Base Storage and Management
  - Machine Learning and discovery
- 



# Expectations?!

- There is a lot to learn
- ▶ There is a steep learning curve
- ▶ Slow and steady wins the race
- ▶ We are here to help





# Useful Resources

- R for Data Science (2e): <https://r4ds.hadley.nz/>
- Python for Data Analysis (3e): <https://wesmckinney.com/book/>
- GitHub: <https://docs.github.com/en/get-started/quickstart/hello-world>
- Posit Primers: <https://posit.cloud/learn/primers>



Thank you for your  
attention!

[panovt@tcd.ie](mailto:panovt@tcd.ie)