Mastering Applied Data Science

Day 2: Data Preparation and Exploration with Visualization, External Data Acquisition

Agenda - Day 2

THE DATA SCIENCE BOOTCAMP

Introduction to Data Science	• Matplotlib, Pandas visualization • Seaborn • Tableau • Supermarket Data visualization
Data Preparation and Exploration	• Feature Engineering • Missing Value Imputation • Creating Dummy Variables
External Data Acquisition	• Web Scraping • API Pulls

What is Data Visualization?

THE DATA SCIENCE BOOTCAMP

Data visualization is the representation of data or information in a graph, chart, or other visual format.

It communicates relationships of the data with images.

Data Visualization

THE DATA SCIENCE BOOTCAMP

- Words don't always paint the clearest picture. Raw data doesn't always tell the most compelling story.
- The human mind is very receptive to **visual information**. That's why data visualization is a powerful tool for communication.
- Visualization is important step in data exploration and data communication
- Gain insight into data that will guide analysis approaches
- Great technique to help explain patterns, trends, and correlations

Why Data Visualization?

THE DATA SCIENCE BOOTCAMP

A visual summary of information makes it easier to identify patterns, trends and outliers than looking through thousands of rows on a spreadsheet.

It's the way the human brain works.

Since the purpose of data analysis is to gain insights, data is much more valuable when it is visualized.

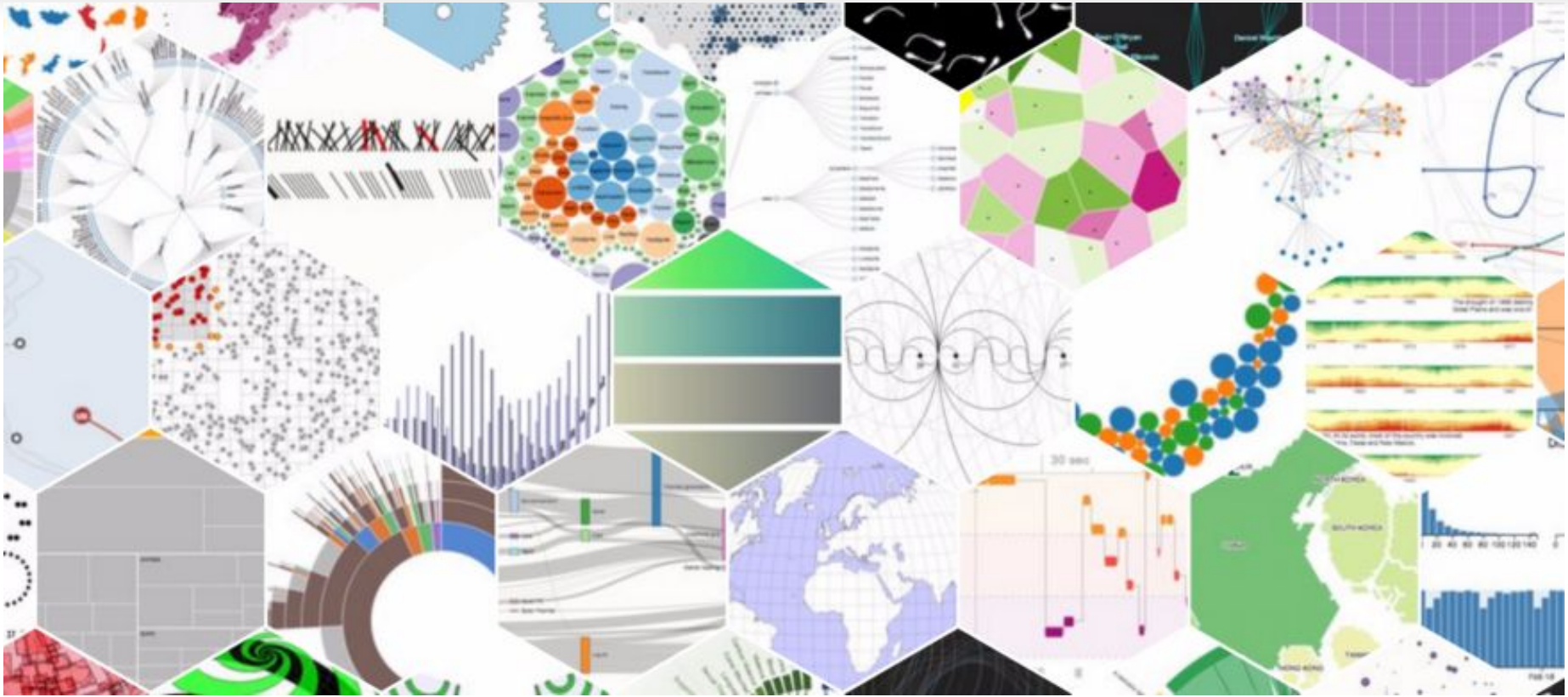
“A picture is worth a thousand words”

THE DATA SCIENCE BOOTCAMP

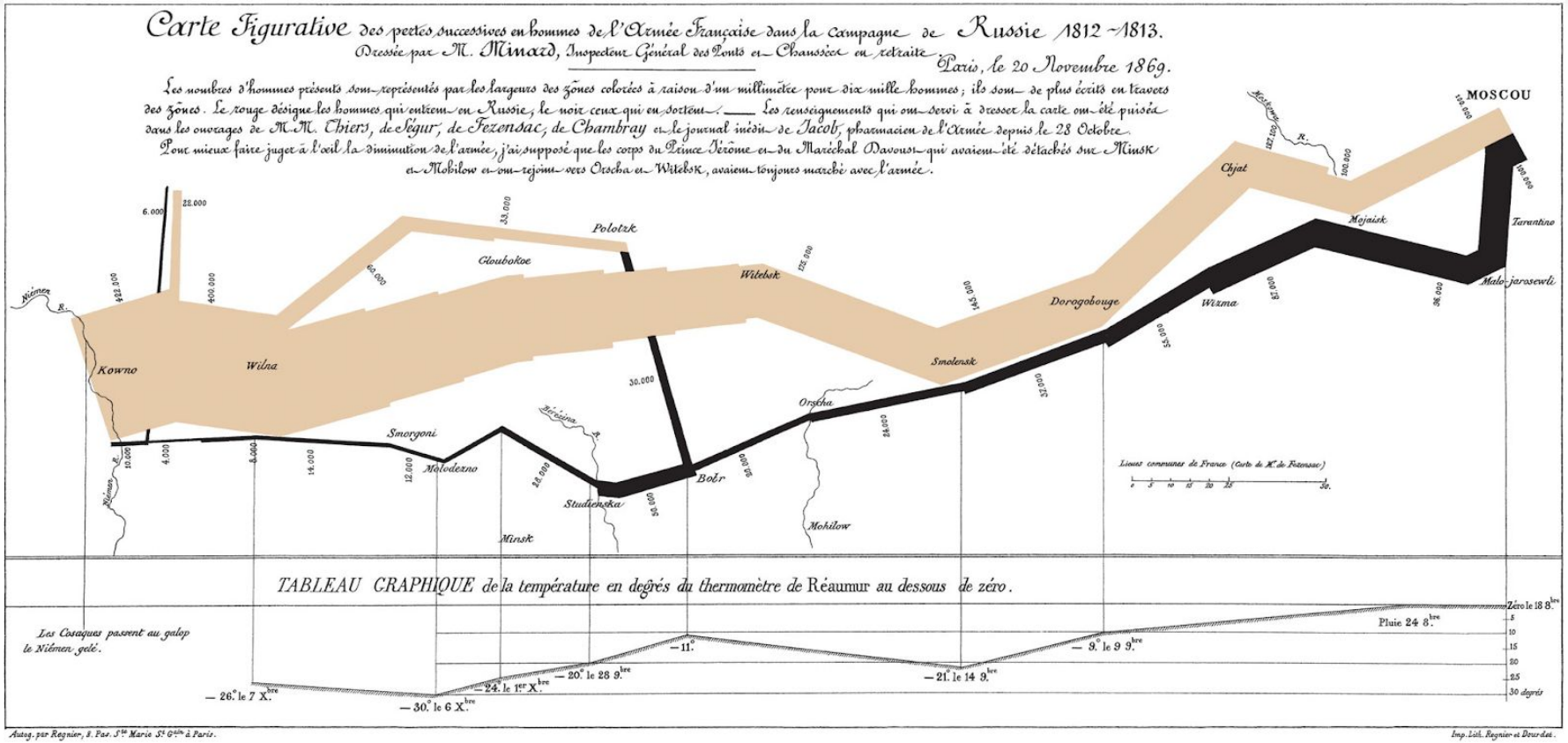
Our eyes are drawn to colors and patterns. We can quickly identify red from blue, square from circle. Our culture is visual, including everything from art and advertisements to TV and movies

Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message.

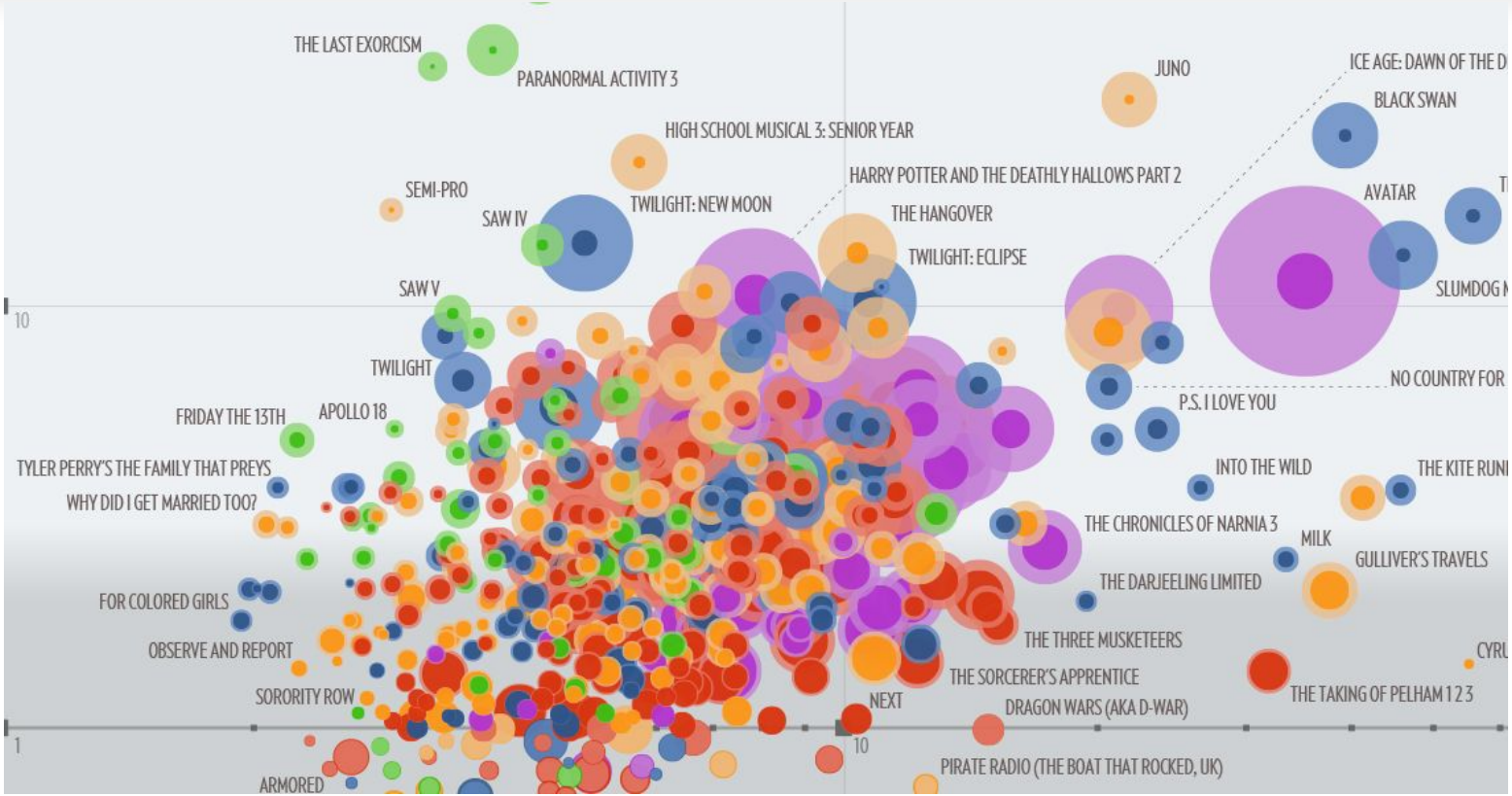
Data Visualization



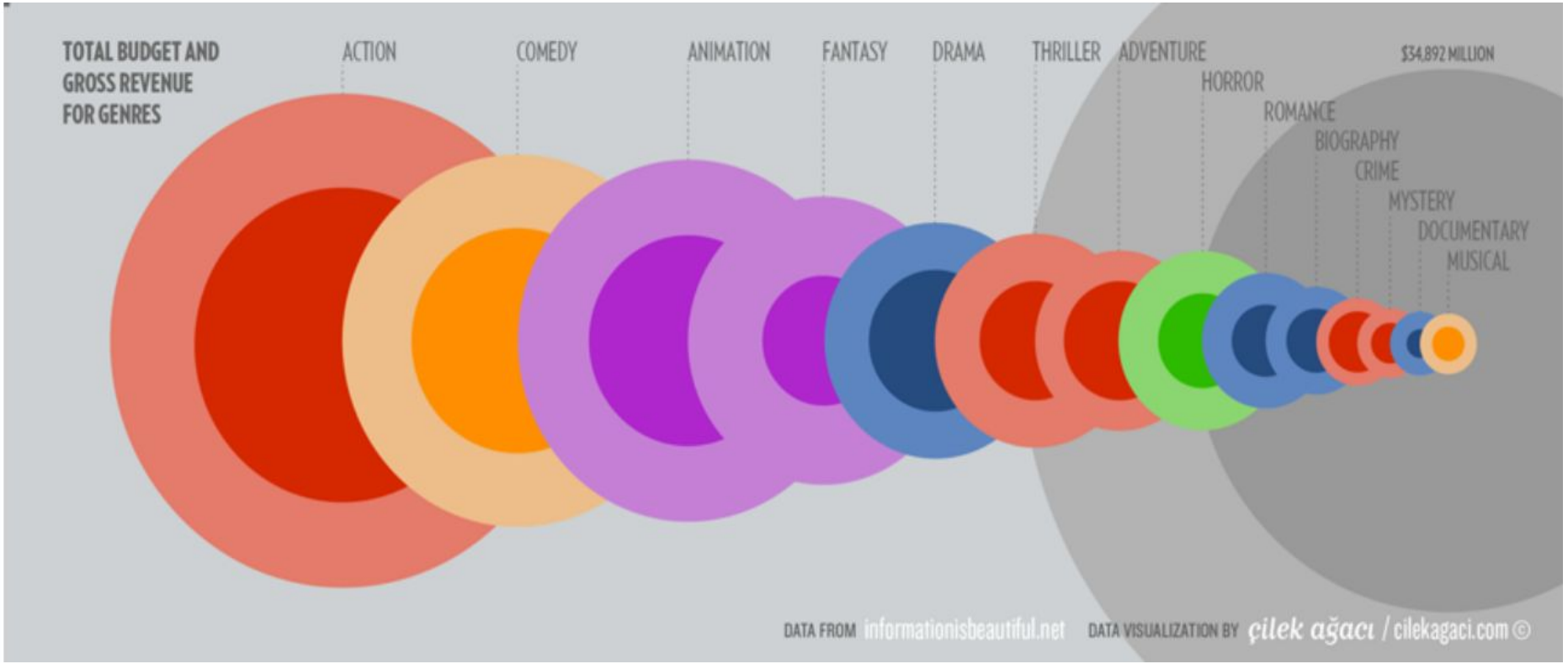
Data Visualization



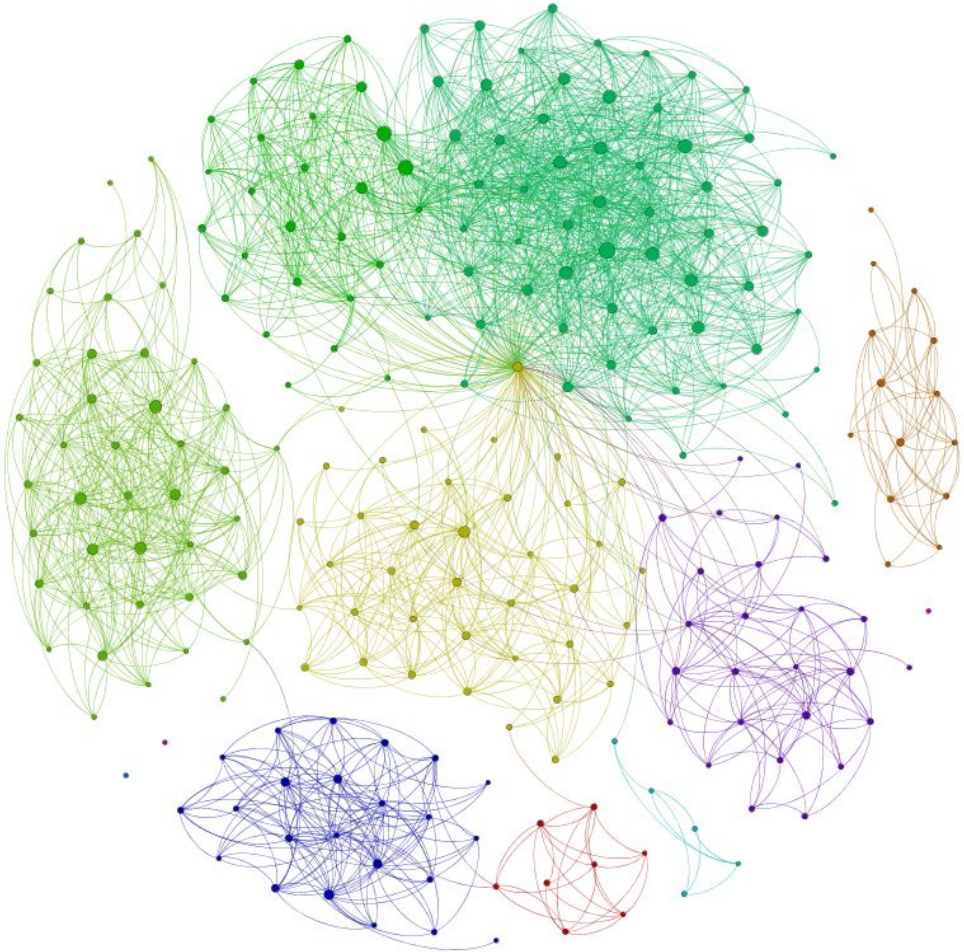
Hollywood Economics

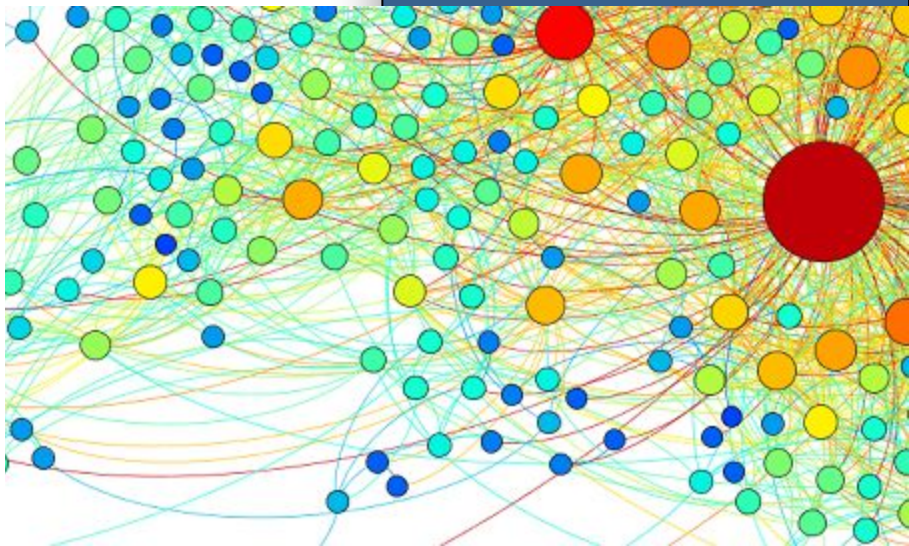


Hollywood Economics

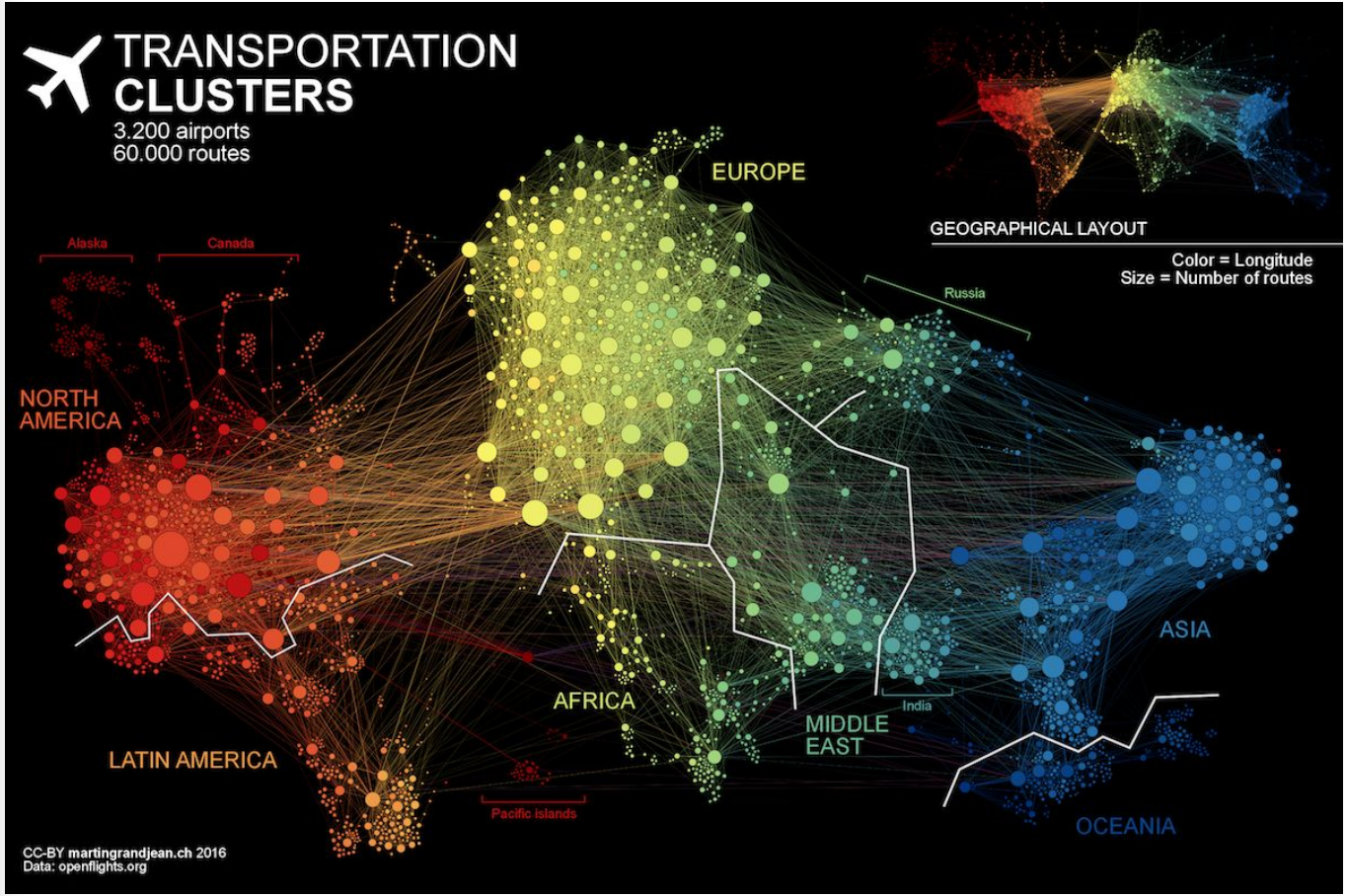


Graph Theory



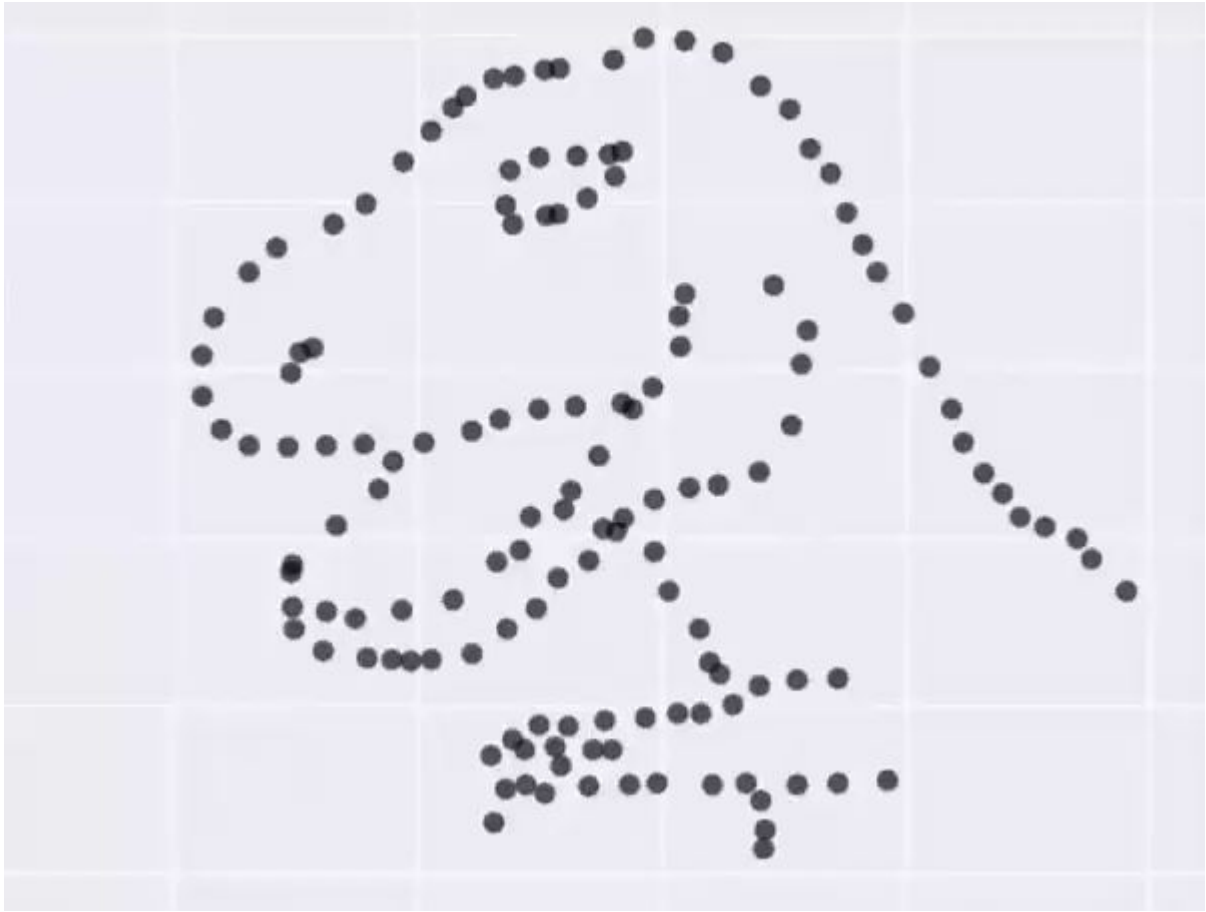


Air Traffic Control



Never trust summary statistics alone; always visualize your data

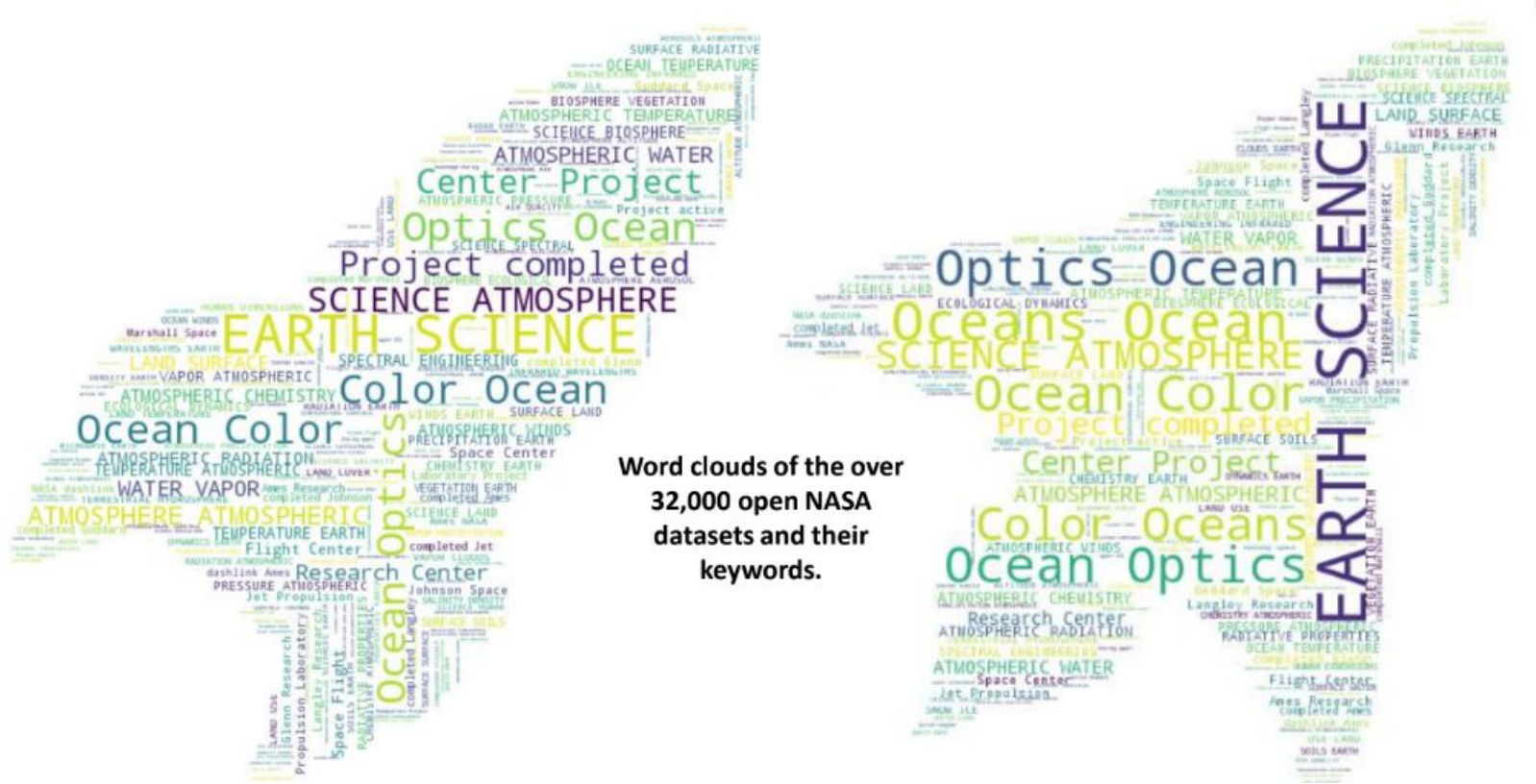
Case in point - all datasets below have the same mean, standard deviation, and correlation (both x/y) to 2 decimal places, while being drastically different.



2021 This Is What Happens In An Internet Minute

What happens in an internet minute?

Word Cloud



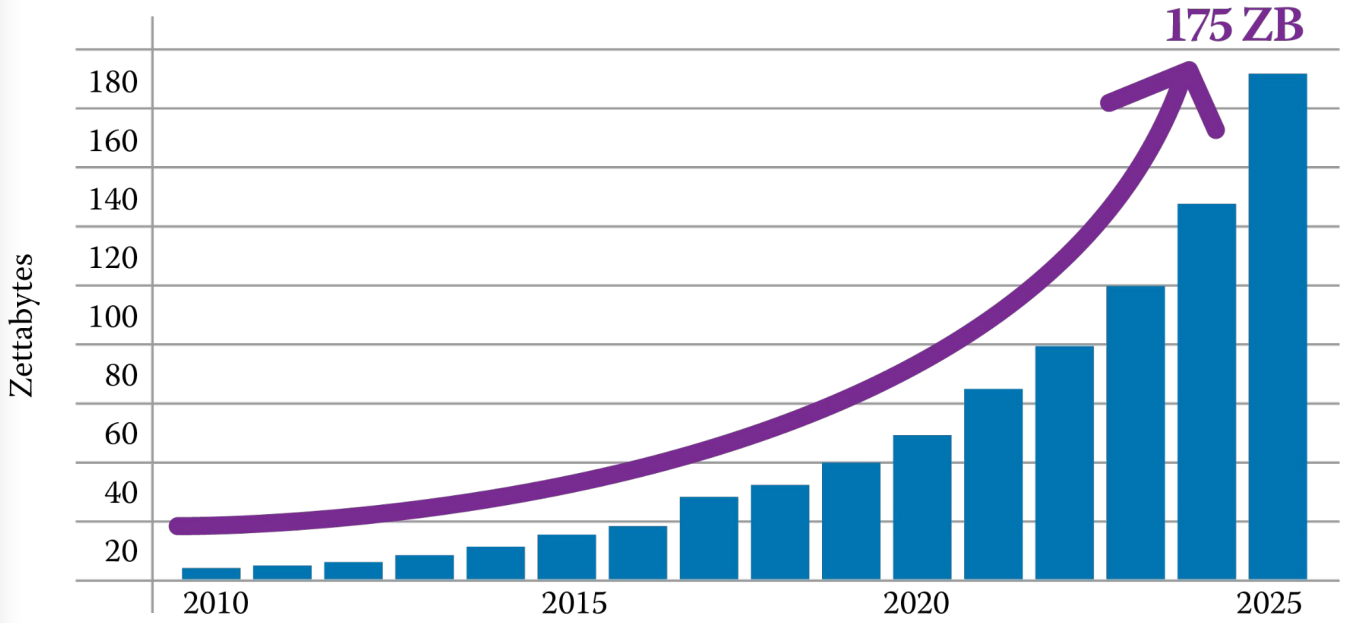
Word clouds of the over 32,000 open NASA datasets and their keywords.

Word Cloud

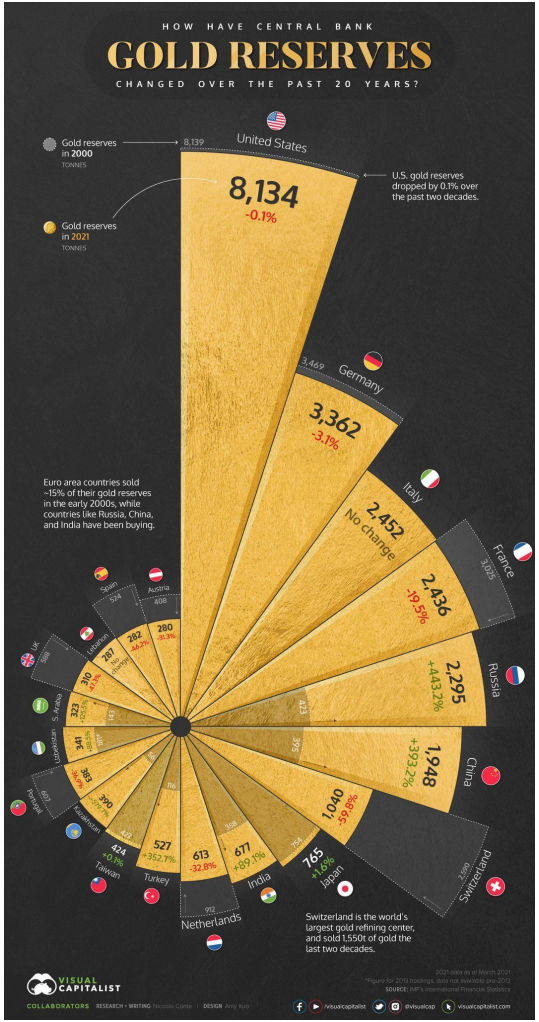

Alice
little
Queen
know
went
King
one
thing
time
see
head
began
Hatter
Gryphon
Mock Turtle
now
way
first
much
back
Duchess
say
voice
quite
looked
large
think
door
Dormouse
Caterpillar
well
came
never
looking
round
thought
tone
got
go
Mouse
Cat
last
tell
March Hare
come
day
nothing
hand
found
look
word
Oh


BIG
DATA
INFORMATION
STORAGE
LARGE
PETABYTES
SEARCH
ANALYSIS
RESEARCH
MANAGEMENT
TECHNOLOGIES
SOFTWARE
VOLUME
SETS
MILLION
DEVELOPMENT
TECHNOLOGY
SCIENCE
FUTURE
NEW
INTERNET
LEARNING
BASED
PROCESS
SAN
NAS
GOVERNMENT
SOCIAL
SYSTEMS
DATABASES
PROCESSING
SIZE
ANALYTICS
EXABYTES
FRAMEWORK
BETWEEN
BUSINESS
APPLICATIONS
AMOUNT
REAL
HIGH
EVERY
WORLD
BILLION
PARALLEL
INITIATIVE
ECONOMIC
TERABYTES
TIME
USED
NETWORKS
HYPOTHESIS
INCLUDE
PAST
COMPLEX
MASSIVE
FACTOR
ONLINE
ARCHITECTURE

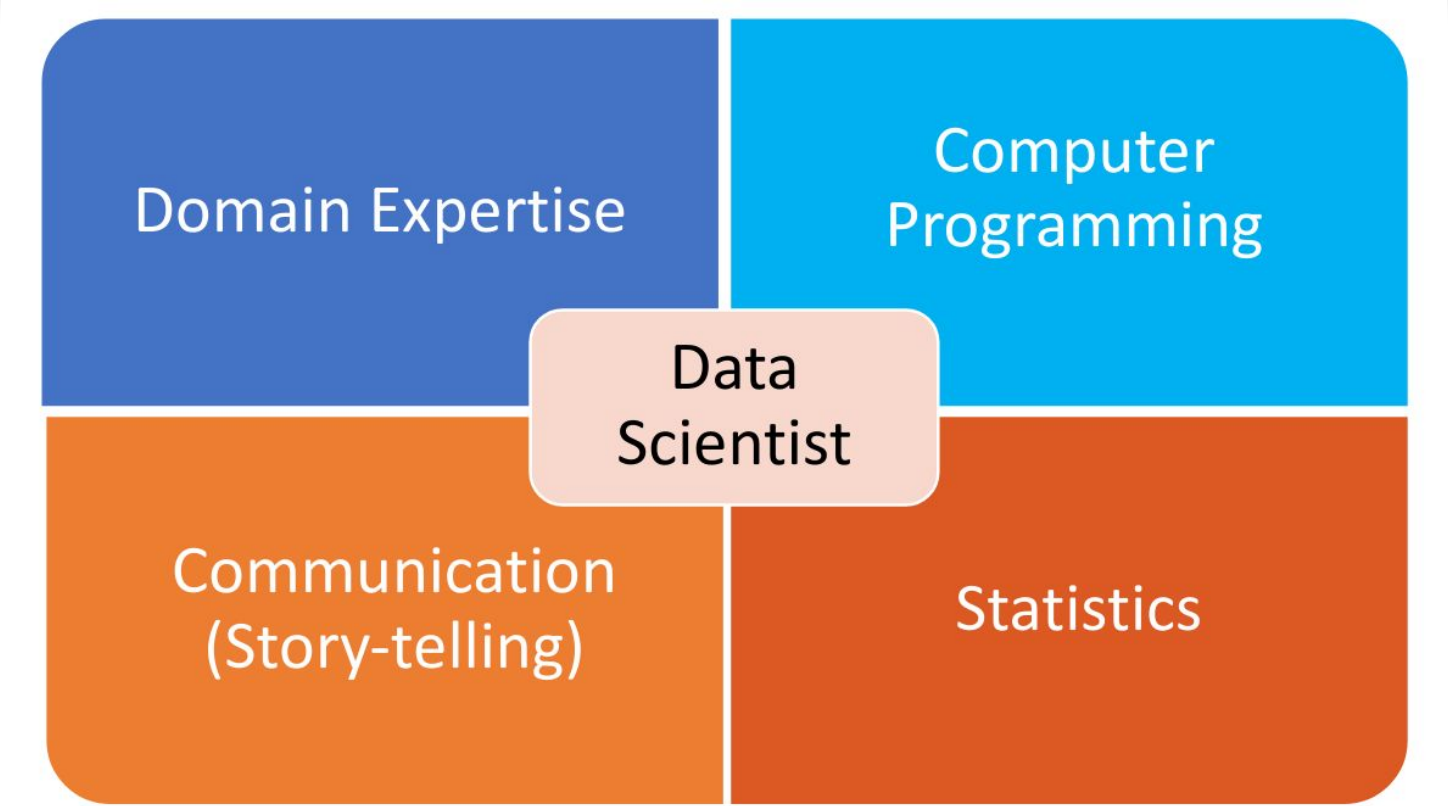
Massive Data Growth



Gold Reserves



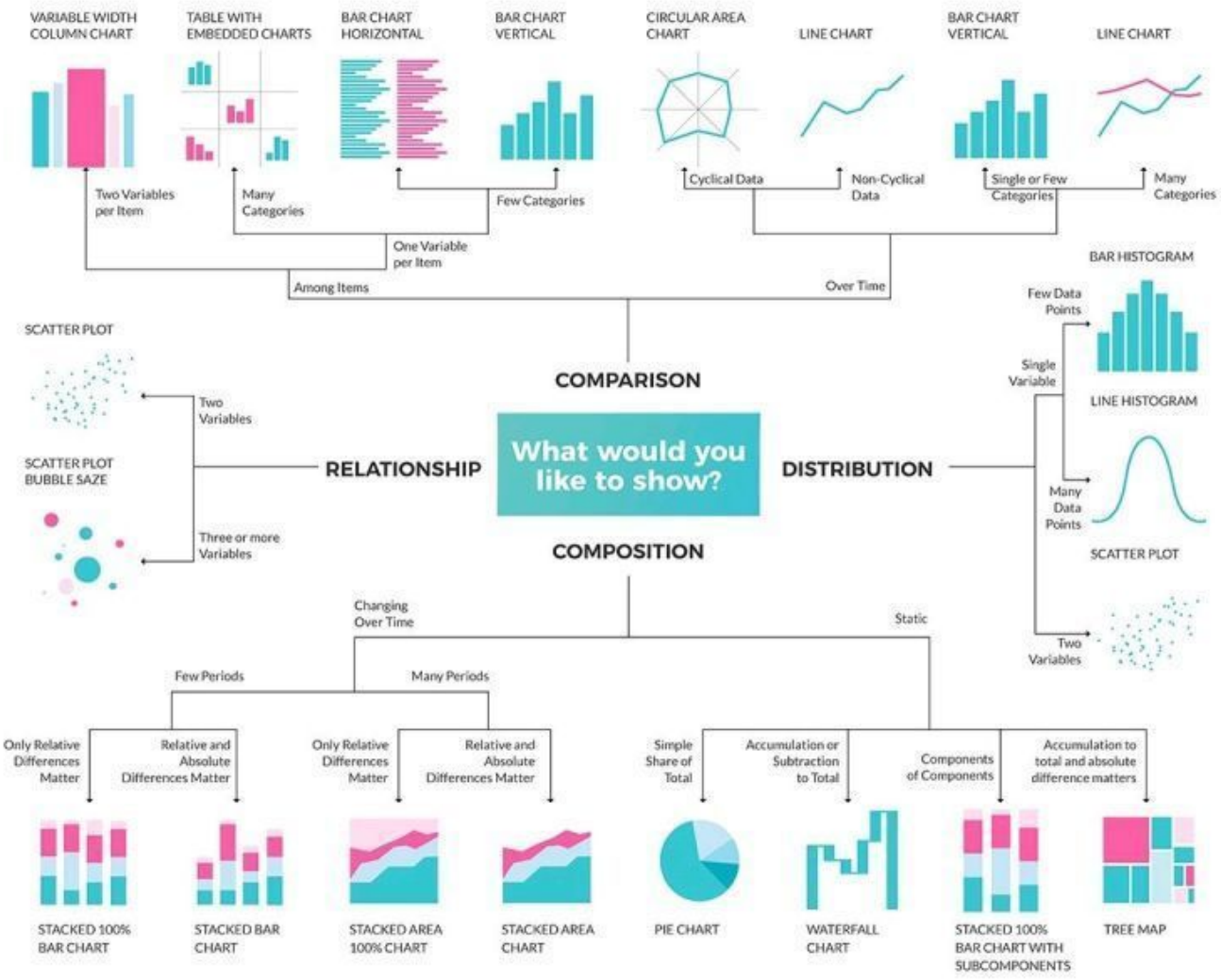
Data Science is Multidisciplinary



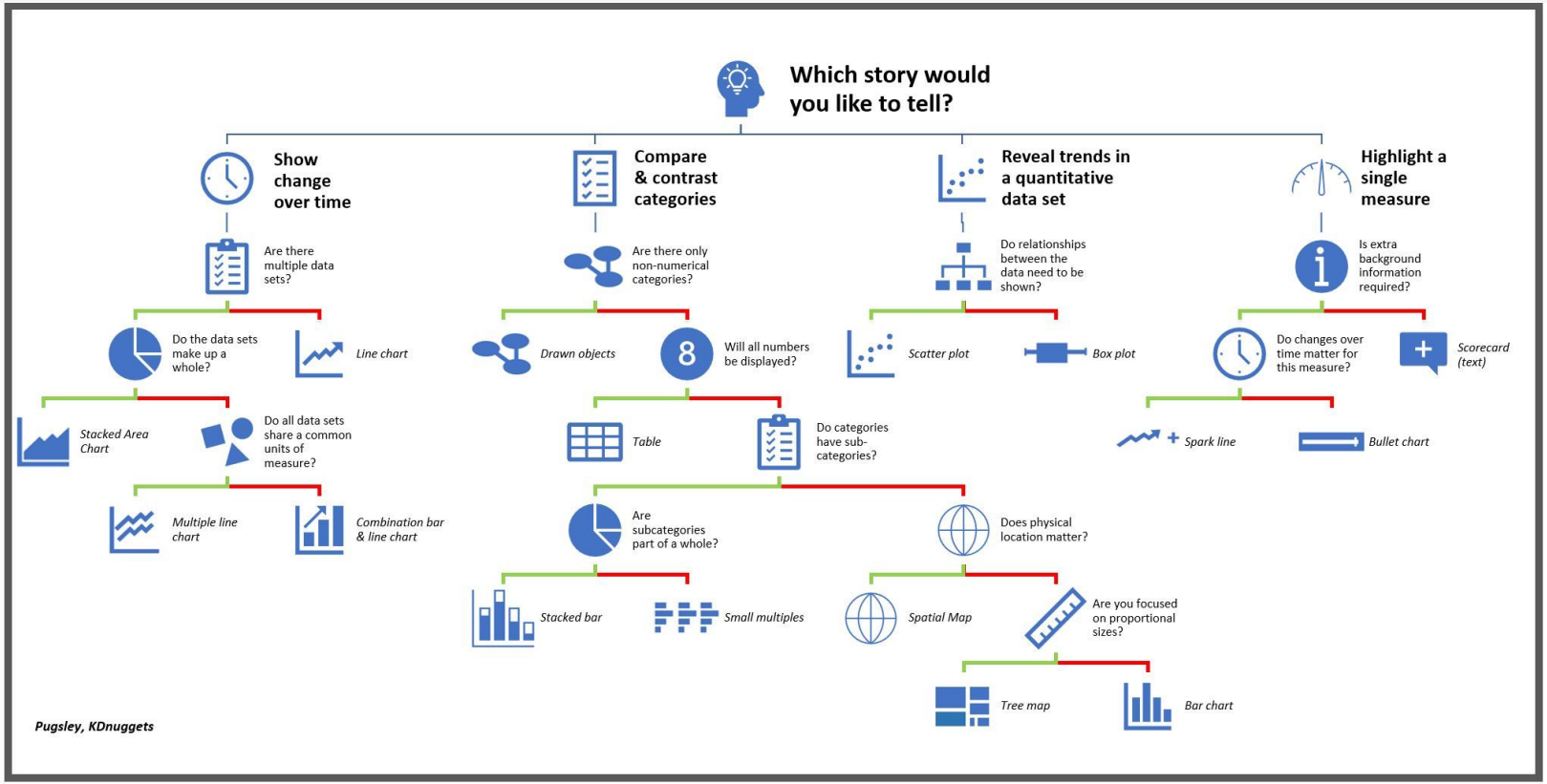
Data Visualization

Data Visualization

https://datavizcatalogue.com/



Data Visualization



Visualization

- Most popular graphing package in Python
- Visualization is important step in exploratory analysis and communicating findings

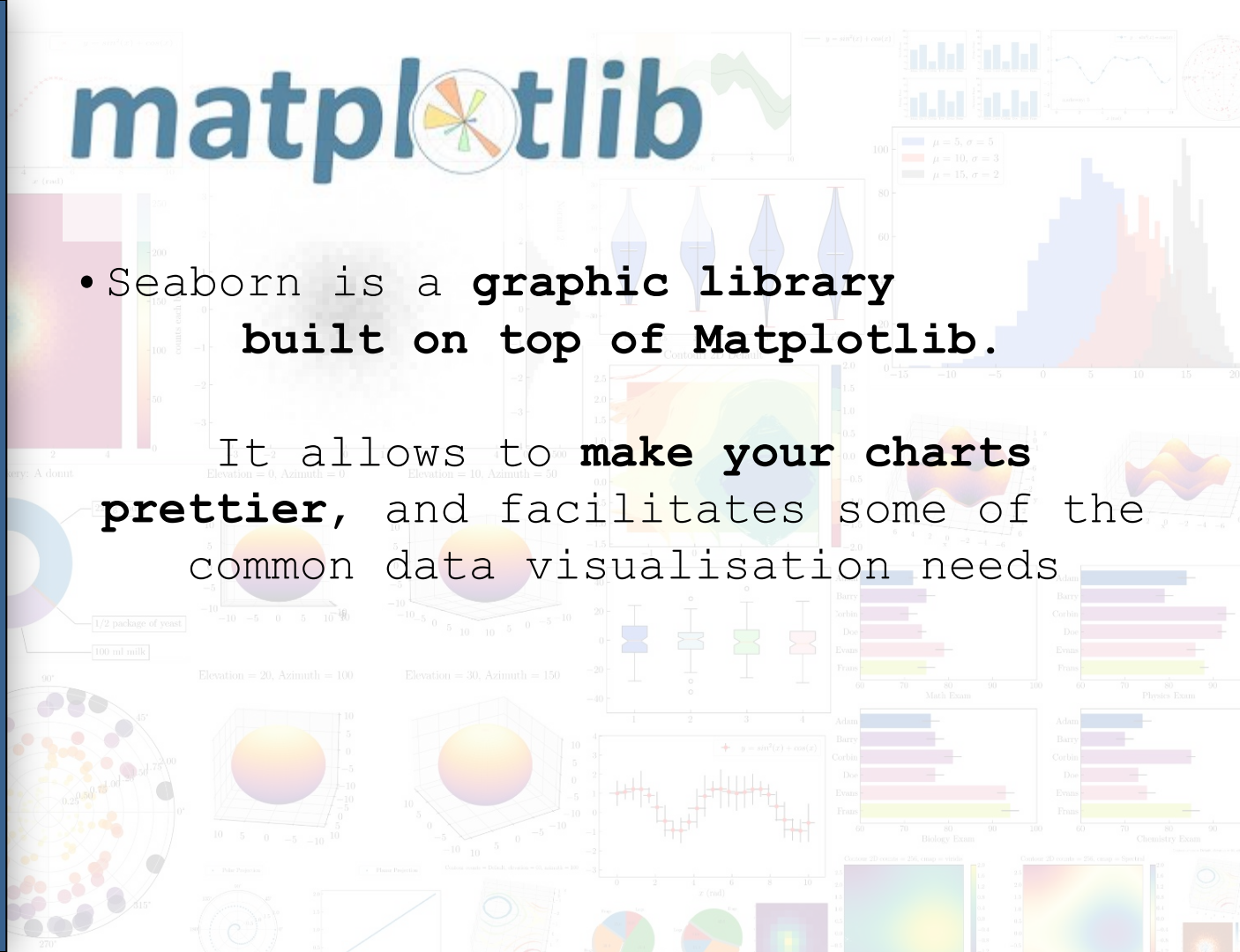
- We'll create five types of graphs here:
  - Line Plot
  - Scatter Plot
  - Histogram
  - Box plot

Pandas Visualization

• Pandas is an open source high-performance, easy-to-use library providing data structures, such as dataframes, and data analysis tools like the visualization tools.
• Pandas Visualization makes it really easy to create plots out of a pandas dataframe and series. It also has a higher level API than Matplotlib and therefore we need less code for the same results.
• Here let's look at five common used graphs:
    • Scatter plot
    • Line chart
    • Histogram
    • Bar chart
    • Box plot

Pandas Visualization

Visualization



- Seaborn is a **graphic library built on top of Matplotlib.**

It allows to **make your charts prettier,** and facilitates some of the common data visualisation needs

Any Questions?

Zafer Acar
Senior Data Scientist
zafer.acar@cern.ch
The Data Science Bootcamp