Day 2: Introduction to Data Science

Samuel Schlenker

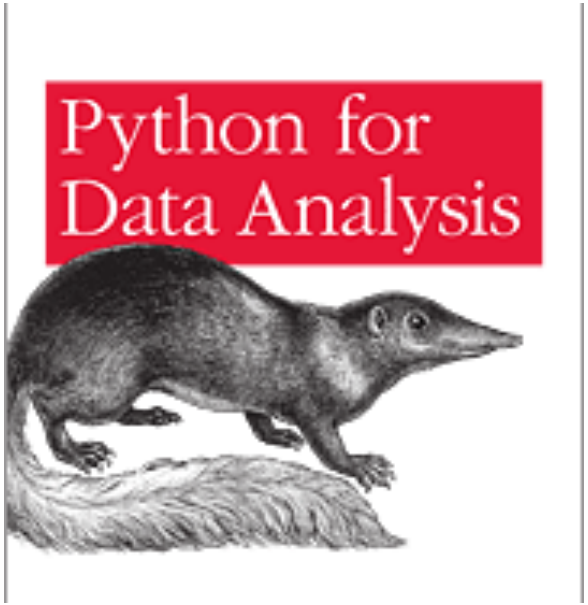
04.11.2025, WWI 2025F

Students should ...

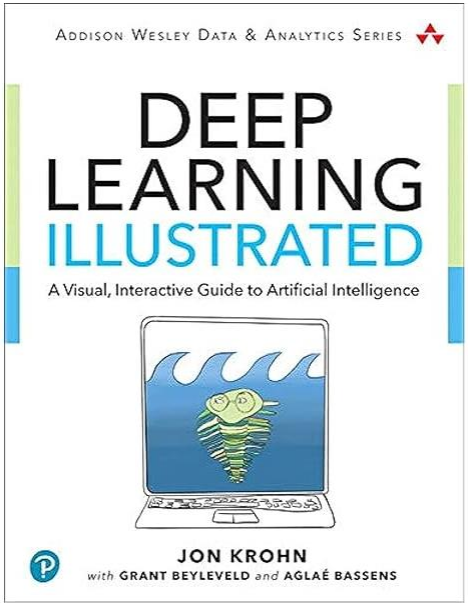
- Understand the fundamental definition and scope of data science as an interdisciplinary field
- Identify the three pillars of data science: domain expertise, statistics/mathematics, and computer science
- Recognize real-world applications of data science across industries (quality control, predictive maintenance, fraud detection, autonomous driving)
- Distinguish between AI, Machine Learning, Deep Learning, and Generative AI
- Understand the characteristics of Big Data (Volume, Velocity, Variety, Veracity, Value)
- Identify different data sources (open-source, private, commercial) and their accessibility
- Apply principles of effective data visualization and storytelling
- Recognize poor or misleading visualizations and understand common pitfalls
- Understand different chart types and their appropriate use cases (line charts, bar charts, scatterplots, heatmaps, etc.)
- Recognize the difference between correlation and causation
- Identify opportunities and risks associated with data science applications

Recommended Reading

Data Science from Scratch, by Joel Grus. O'Reilly

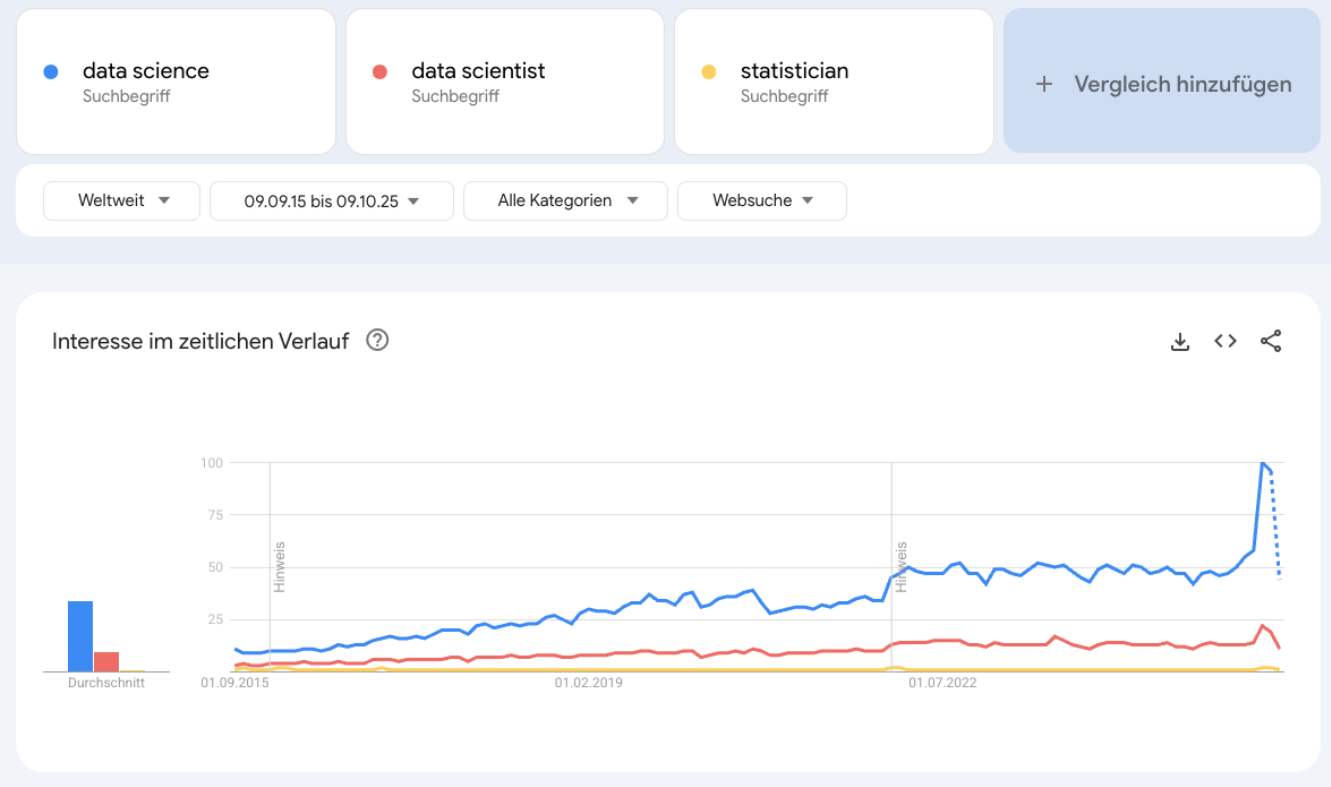


Python for Data Analysis, 2nd ed by Wes McKinney. O'Reilly

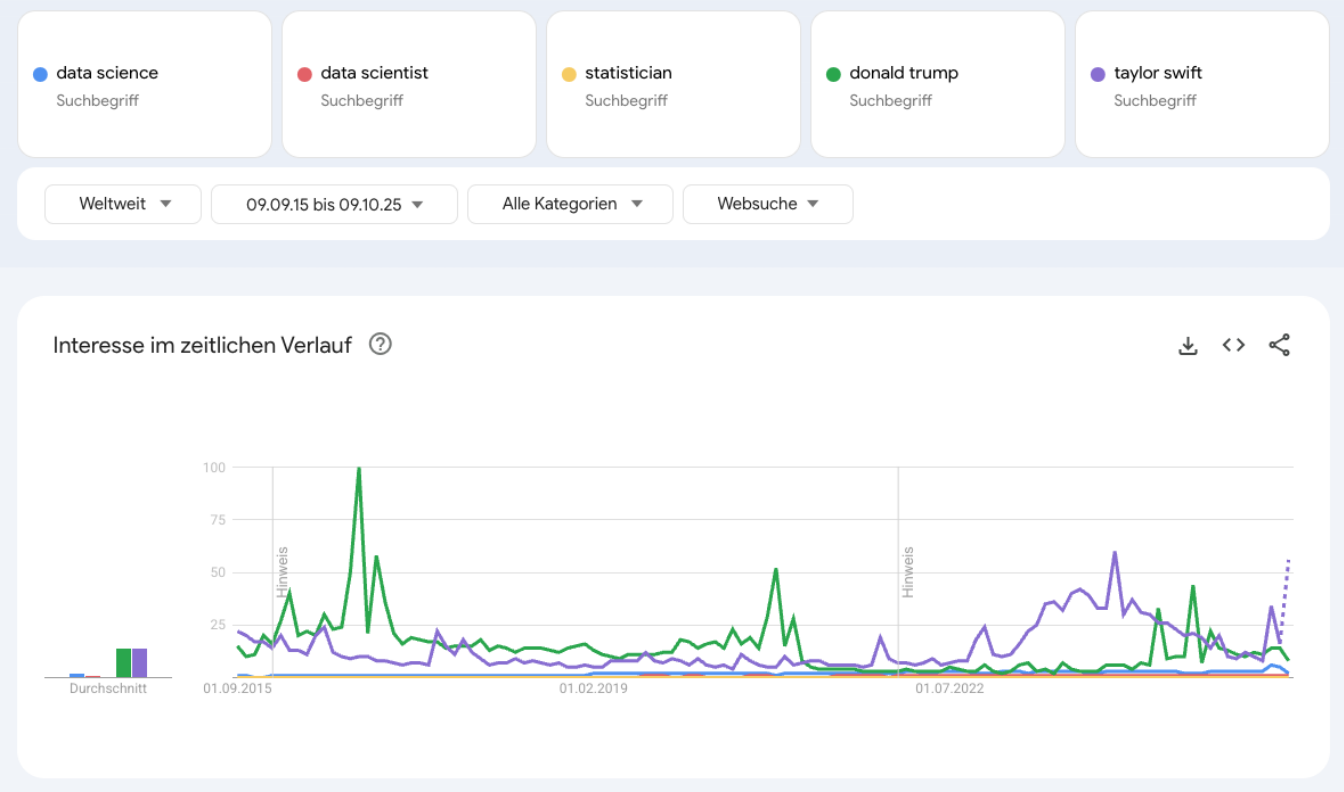


https://www.deeplearningillustrated.com/

Google Trends

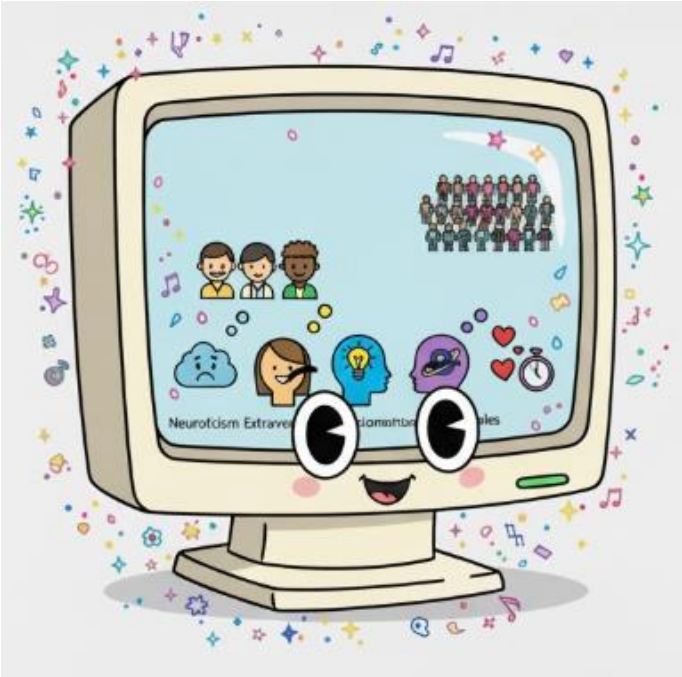


In comparison ...



Facebook knows us better than friends and family

- 2015: Study with 86,220 volunteers
- Collaboration between the University of Cambridge and Stanford University
- Questionnaire with 100 items on the Five Factor Model (FFM) of personality psychology / "Big Five"
- Neuroticism, extraversion, openness to experience, conscientiousness, and agreeableness
- Computer algorithm (linear regression) vs. assessment by individuals
  - From 10 likes: Computer is better than coworkers
  - From 70 likes: Computer is better than friends
  - From 150 likes: Computer is better than family
  - From 300 likes: Computer is better than spouse
- The average Facebook user shares 227 likes

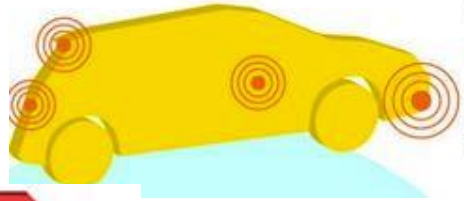


Self-test: https://applymagicsauce.com/demo

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]



Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

40 ZETTABYTES

[43 TRILLION GIGABYTES]

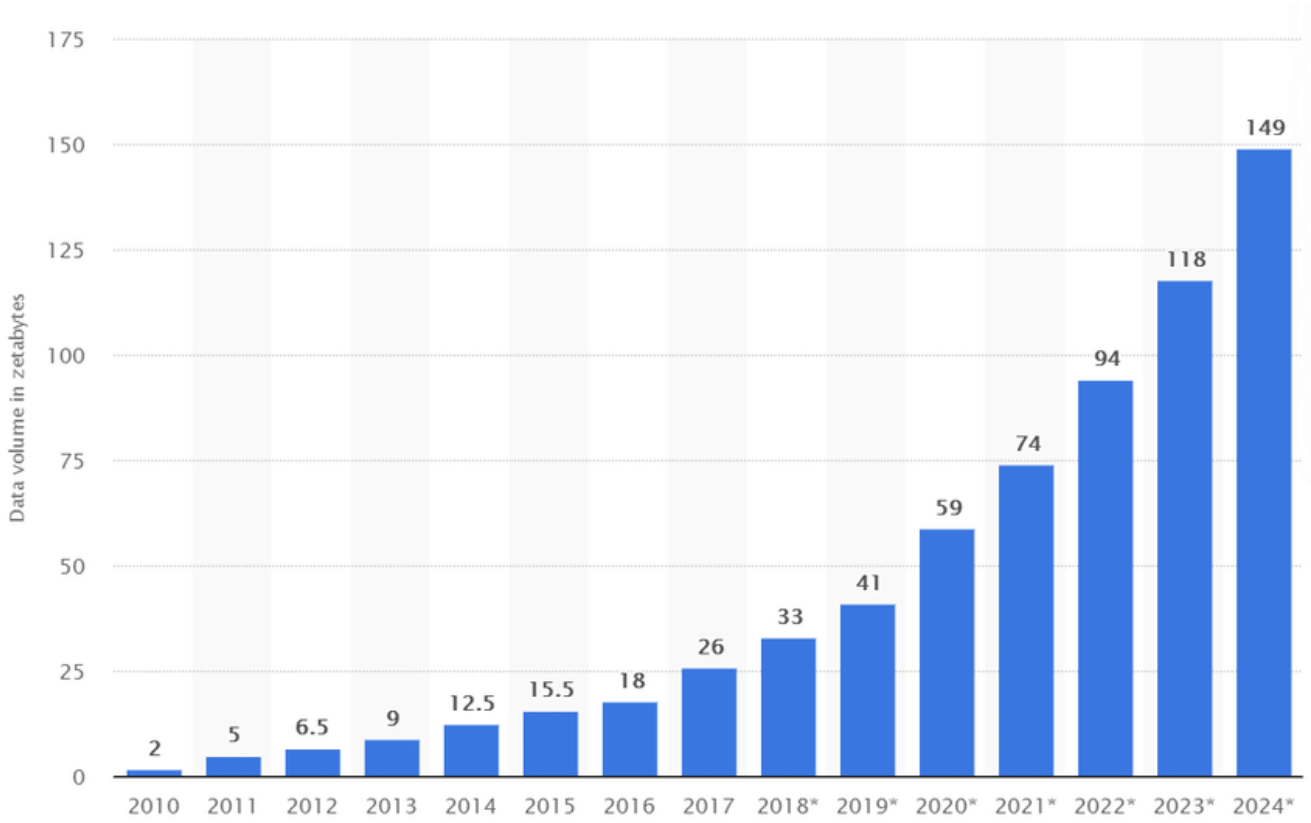
of data will be created by 2020, an increase of 300 times from 2005

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

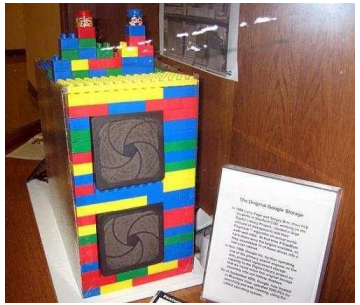
Pure Amount of data



Distributed Computing

This amount of data requires more than one computer.

Google – 1998:



Distributed Computing

This amount of data requires more than one computer.

Google – 2014:


INTRODUCTION TO DATA SCIENCE BY W&S
YOUR TURN!
EXERCISE / QUESTION

Data Science

Needs massive data processing

Distributed computing

Why is this the case?

Data Is Driving Everything

- Modern data acquisition is inexpensive!
    - Smartphones, embedded systems, inexpensive sensors,
    - Medical devices, simulators, …
- Data storage is inexpensive!
- Parallel (compute cluster) computation is inexpensive
    - The Cloud, clusters of computers, GPUs, tensor processors, …
- Science only has explanatory and predictive models in a few (mostly physical sciences-related) domains
- … So: can we use algorithms + data to understand phenomena? Build or augment models? Build detectors? Make diagnoses?

Data Is Driving Everything

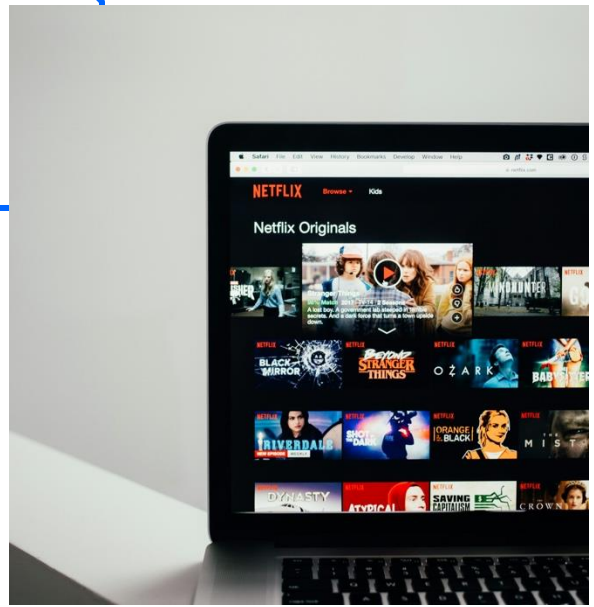
"Big data"
"Data science"
"Data lakes"
"Visual analytics"

"Deep learning"
"Statistical analysis"
"Biomedical informatics"
"Business analytics"

Lots of trends in pursuit of the same goals!
Discovery, models, decision-making, …

Also, new issues -
"Ethical algorithms"
"Reproducibility"

Data Science



Reccomandation engines

Autonomous driving

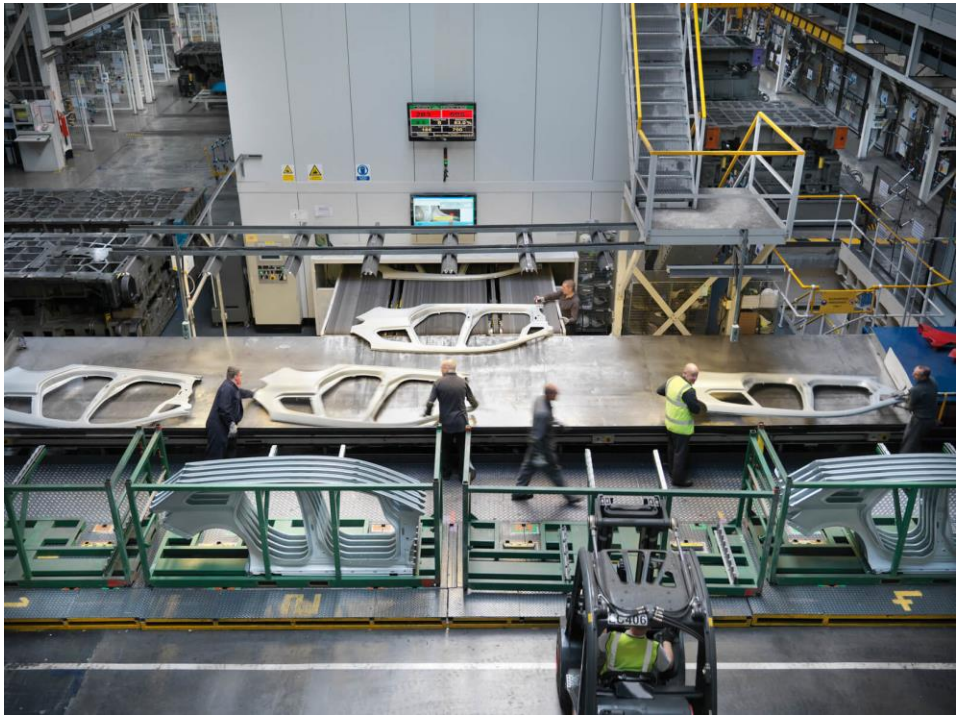
Business intelligence

Applications of data science (Use cases)

Quality control:

- Quantity: Is there the right number of salami slices on the frozen pizza?
- Correct product: Are the right components installed in the server?
- Quality: Are there scratches in the paintwork on the car door? Is the weld seam flawless?
- Results in: Fewer complaints, Cost reduction, Higher customer satisfaction

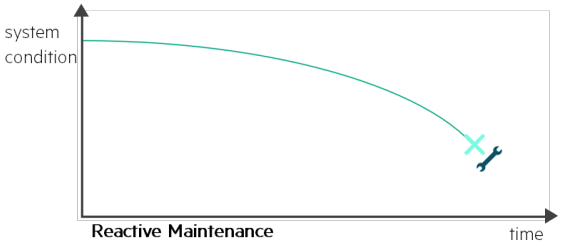
Applications of data science (Use cases)

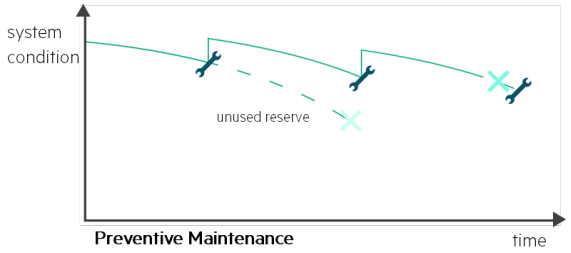


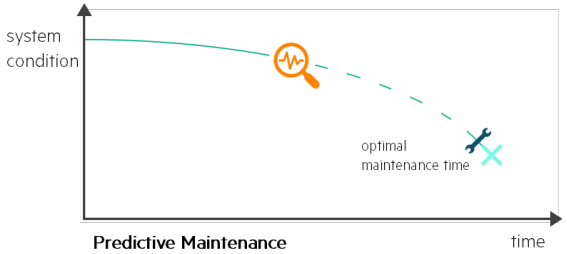
Predictive Quality: Predicting the quality of the product before the production process is complete.

- Early correction of the product, if possible, or discontinuation of production.
- Reduction of waste.
- Energy, water, and material savings by avoiding finishing steps for poor-quality products.
- Cost reduction.

Applications of data science (Use cases)







Predictive Maintenance:

Predicting the failure of a production machine

- Avoiding unplanned production downtime
- Replacing/repairing the component that is about to fail at the best possible time
- Using the production machine or individual parts for as long as possible

Applications of data science (Use cases)

Fraud Detection:

Detection of credit card fraud, for example, through anomaly detection

- Early account blocking
- Loss minimization

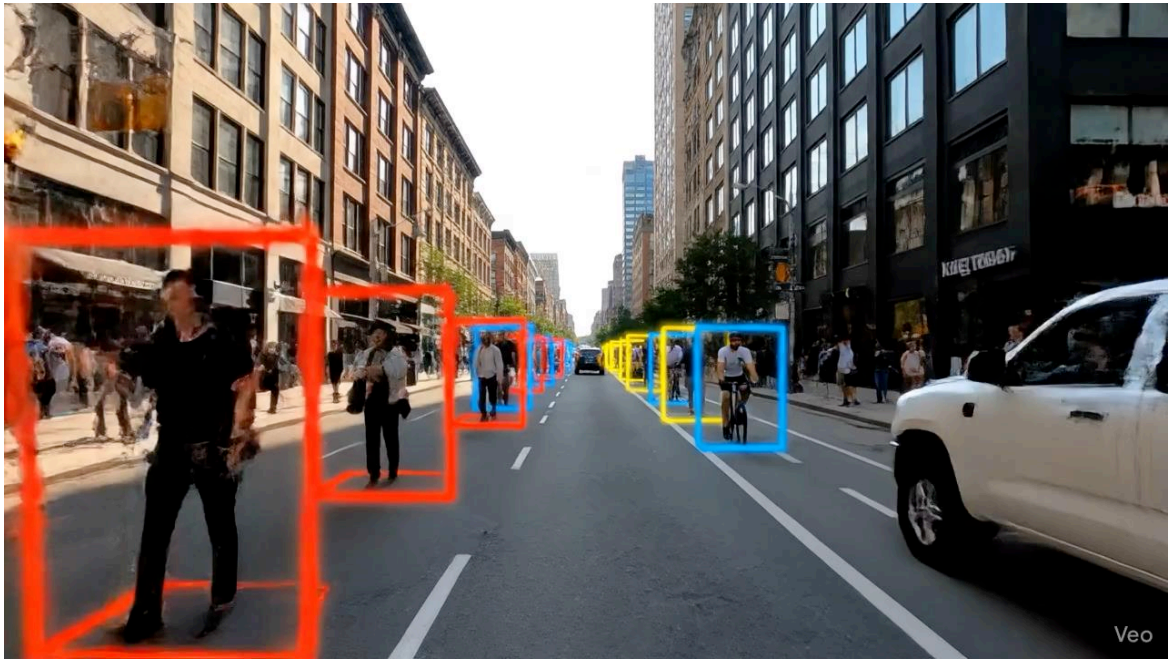
Applications of data science (Use cases)

Personalized recommendations:

For example, shopping cart analysis

- Individualized shopping experience
- Higher customer loyalty
- Higher shopping cart values
- Increased sales

Applications of data science (Use cases)



Autonomous driving:

Detection of pedestrians, their direction of travel, and derivation of the appropriate action

- Traffic sign recognition
- Enables innovation
- Minimization of human error
- Reduction in the number of accidents

Applications by industry

Financial Services

- Fraud Detection
- Risk Assessment
- Algorithmic Trading
- Sentiment Analysis

Healthcare

- Gene Research
- Drug Discovery
- Patient Assistant
- Medical Records

Public Sector

- Urban Planning & Design
- Predictive Policing
- Citizen Engagement
- Document Summarization

Manufacturing

- QC / Defect Detection
- Supply Chain Optimization
- Predictive Maintenance
- Generative Design

Telecommunications

- Customer Support
- Network Optimization
- Predictive Maintenance
- Fraud Detection

Financial Services & Insurance

Public Sector / Defense

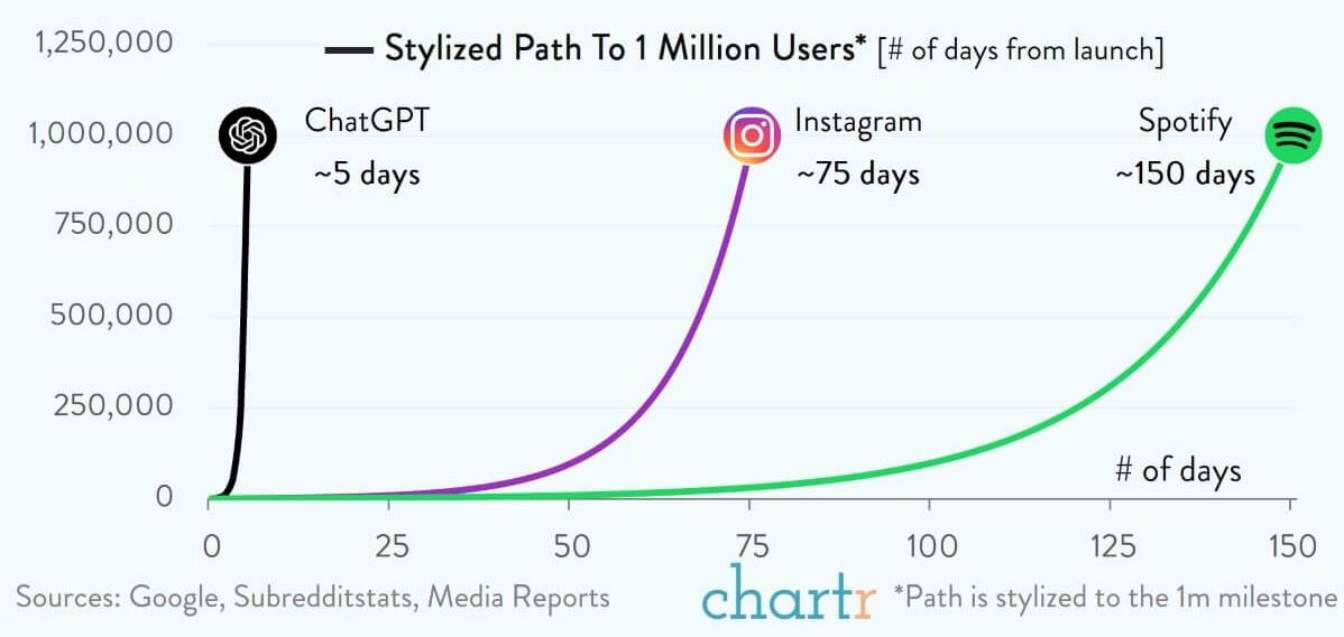
Healthcare / Life Sciences

Autonomous Driving

Manufacturing

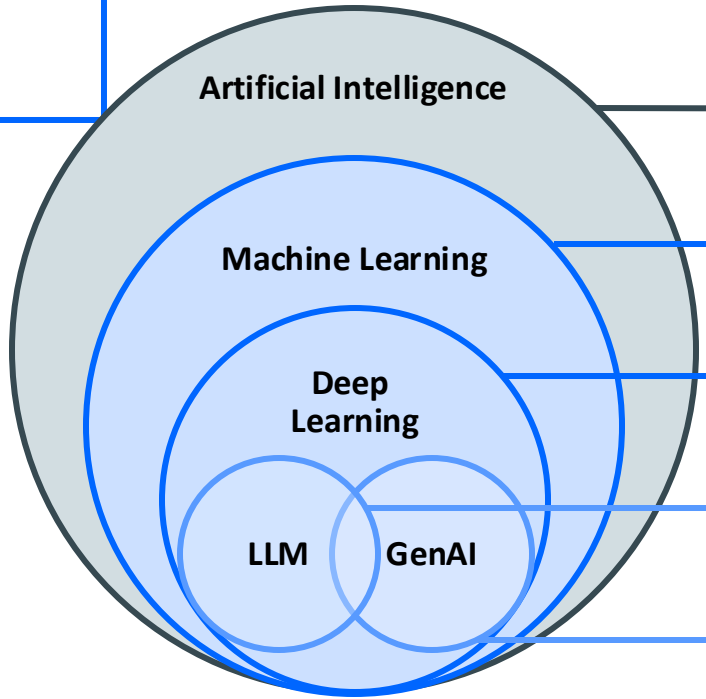
AI goes mainstream with ChatGPT

1M users in 5 days



But how does AI fit into the picture?

Navigating Artificial Intelligence: Understanding its fundamental building blocks



Artificial Intelligence (AI)
Any technology that enables machines to solve tasks in a way like humans do

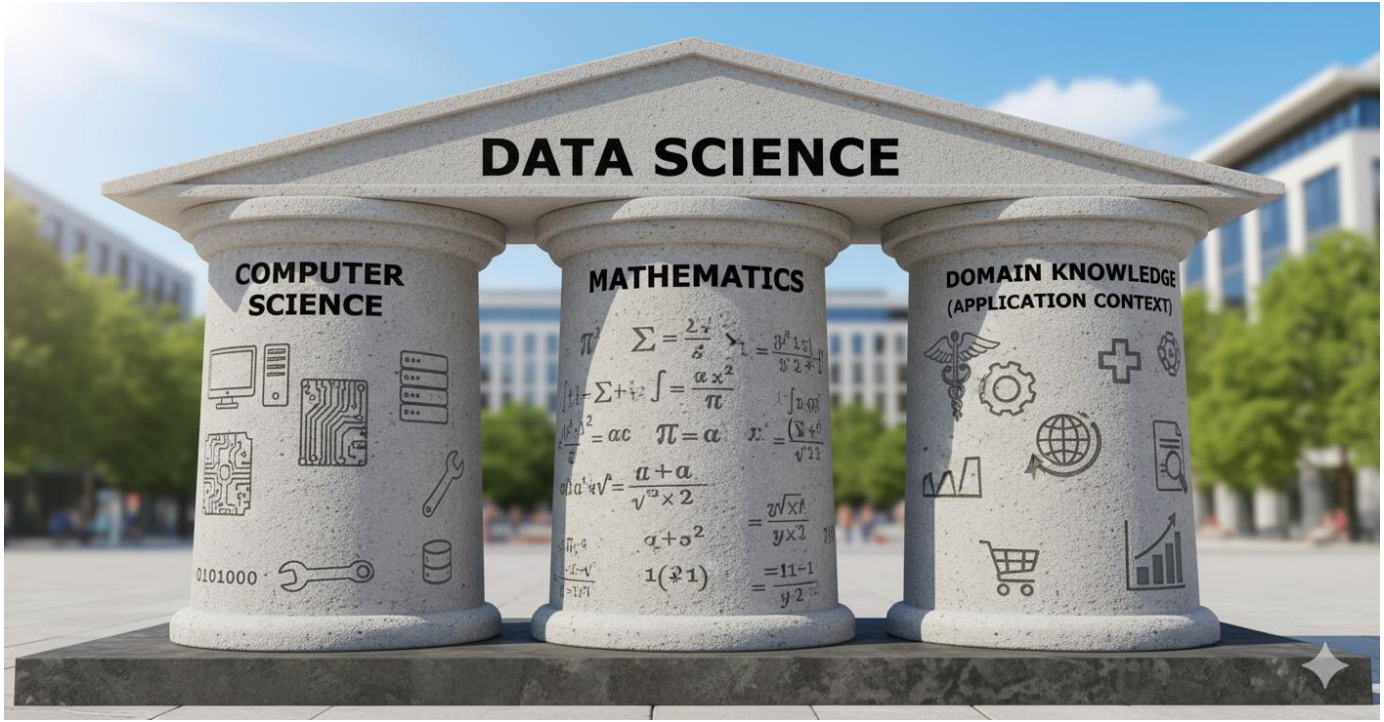
Machine Learning (ML)
Algorithms that allow computers to learn from examples without being explicitly programmed (supervised & unsupervised)

Deep Learning (DL)
Using deep artificial neural networks as models, inspired by the structure and function of the human brain

Large Language Models (LLM)
Models trained on massive datasets to understand and generate human-like text across diverse subjects

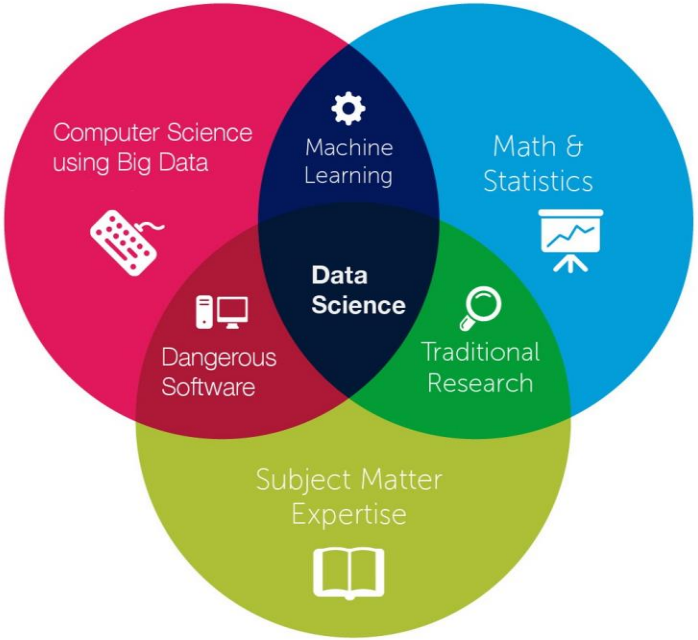
Generative Artificial Intelligence (GenAI)
Refers to technologies that utilize machine learning models to generate human-like text, images, or other content

The three pillars of data science



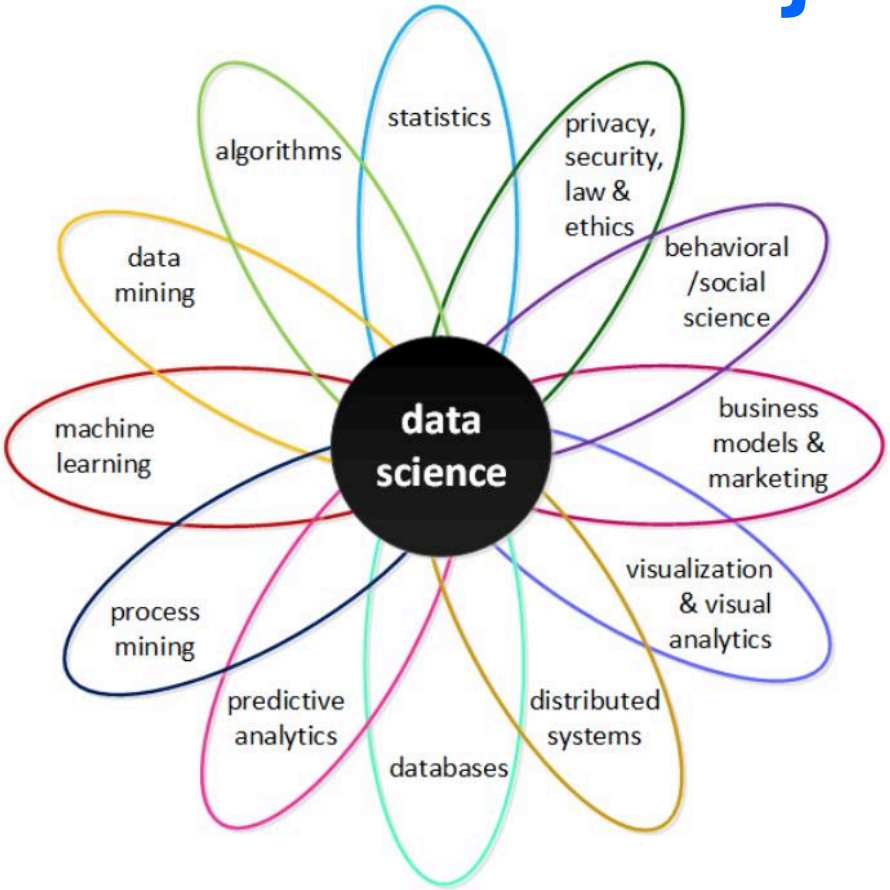
Why so many disciplines?

- Enormous amounts of data → Big Data
    - New algorithms. Must be scalable and efficient
    - New data management technologies and concepts
- High dimensionality of data
    - Data can have thousands of dimensions/attributes
    - High complexity and variability of data
- New, complex fields of application
- Data science is not (only) out-of-the-box, but domain-specific



https://miro.medium.com/max/1100/1*aXJWLmf-CYqVTrNiE2pdCw.png

So, what is a data scientist job?



https://sda.tech/wp-content/uploads/file-manager/Bonn-SDA-WvdA-2017.pdf

And more concrete ... ?

Data Steward

- Monitoring data quality and integrity
- Responsible for the technical accuracy of data

Tasks:

- Data owner
- Data governance
- Master of data

Data Engineer

- Data supply
- Develop, implement, test, and maintain architecture for data storage

Tasks:

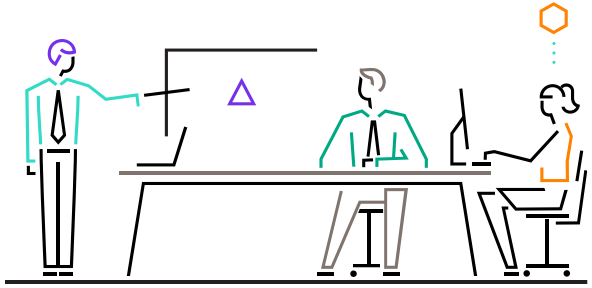
- Data Quality Improvement
- Data Transformation

Data Scientist

- Provides answers to analytical questions using data

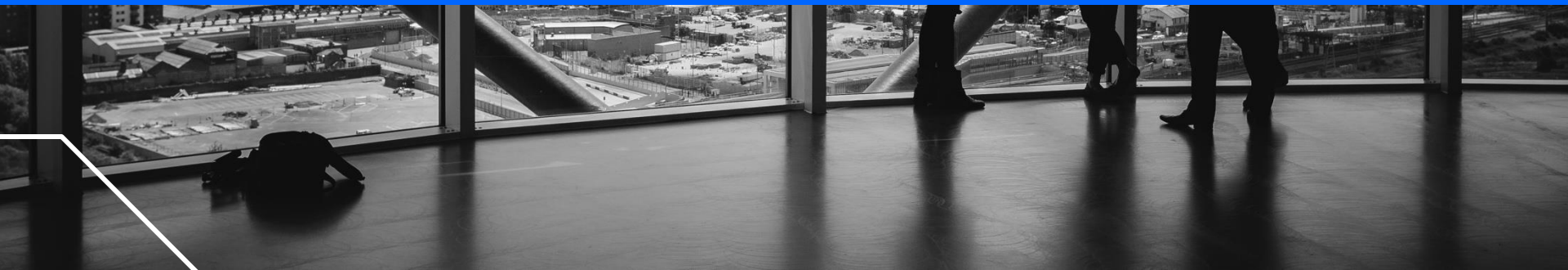
Tasks:

- Data Exploration & Visual Analytics
- Model Deployment & Scoring
- Big Data Handling & Manipulation

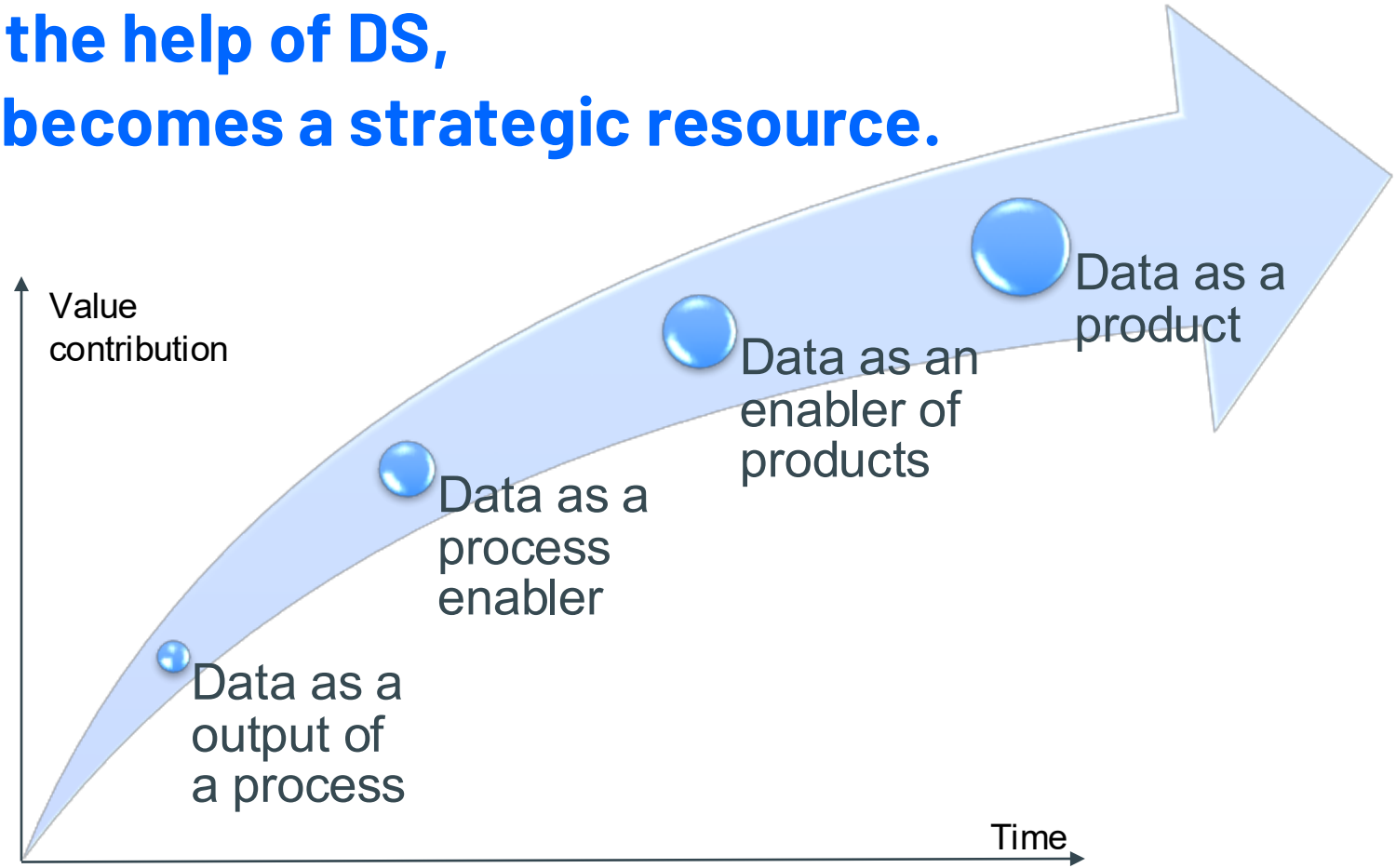


https://www.bertelsmann-university.com/fileadmin/Programs/AI_Curiculum/1_FAQ__%C3%9Cbersicht_Data_Curriculum_DE_final.pdf

Big Data



With the help of DS, data becomes a strategic resource.



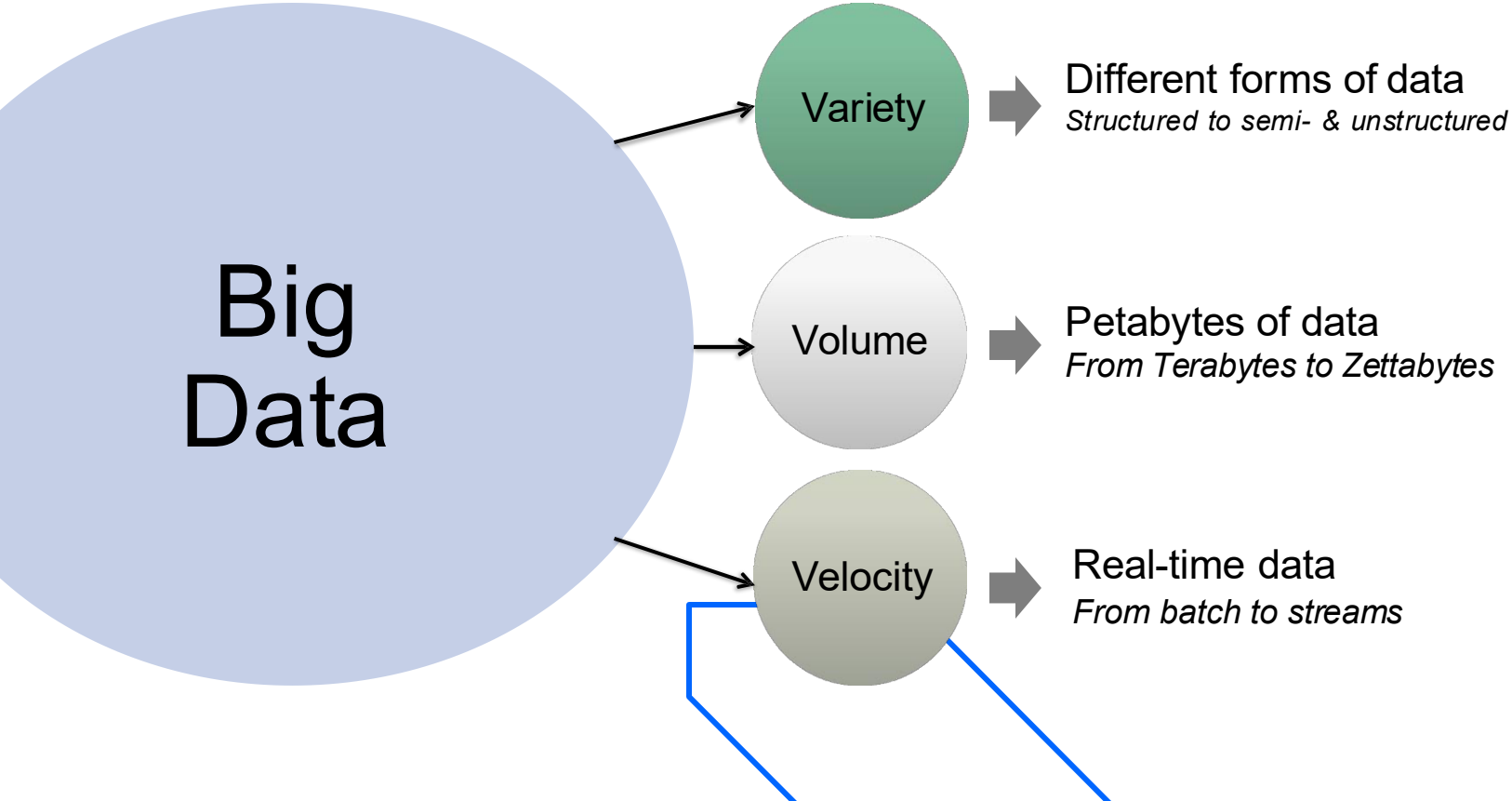
Otto, B.: Industrial Data Space im Überblick, Fraunhofer Institut für Materialfluss und Logistik, Dortmund, 30.10.2015, URL: http://de.slideshare.net/borisotto/berblick- zum-industrial-data-space

Big Data

Refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.

Big does not stand primarily for size, but as an analogy for “overwhelming”

Big can mean “high variety”, “high volume” or “high velocity”


Big Data
Variety
Different forms of data
Structured to semi- & unstructured
Volume
Petabytes of data
From Terabytes to Zettabytes
Velocity
Real-time data
From batch to streams

More Vs

Turning data into knowledge

Variation in meaning in different contexts

Uncertainty of the data

- Not easy to measure
- Depend on context and intended use

Where is the data coming from?

Possible data sources

- Questionnaires, interviews, ...
- Observations & measurements (e.g. in production)
- Analysis of content (documents, web scraping, social media...)

Raw data: Data received or collected

- no variables have been manipulated
- no data removed from the record
- no summary / aggregation

Not prepared
(pre-processed)

Where is the data coming from?

Private data

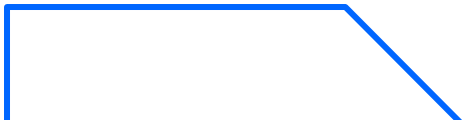
Created by customers
Created during business process execution

Commercial data

Cloud Marketplaces (e.g., AWS Data)
Qlik DataMarket, Statista

Open-source data

Data that is publicly available (check for limits on usage)



Open-Source Data Sources examples

Kaggle
World Health Organization
Our World in Data
Census Bureau (U.S.)
National Oceanic and Atmospheric Administration (U.S.)
UC Irvine Machine Learning Repository
Harvard Dataverse
AWS, Facebook, Google, Microsoft, ...

CAN YOU GET THE DATA OUT OF SILOS?

Value ≠ Quality

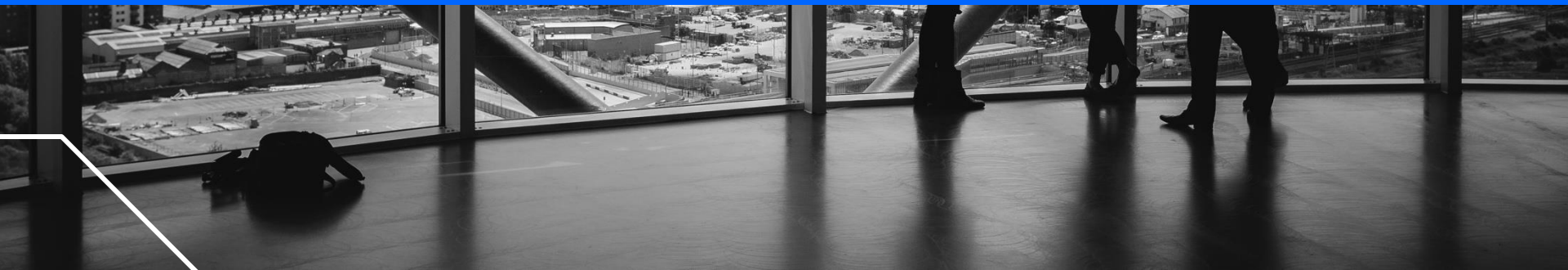
While value and quality of big data may be correlated, they are conceptually different. For example, one can have high quality data about the names of all the countries in North America, but this list of names may not have much perceived value. In contrast, even relatively incomplete data about the shopping habits of people can be quite valuable to online advertisers.

Xin Luna Dong, Google, 2015

Raw data (including high quality data) itself does not hold any value, unless it is processed in analytical tasks from which humans or downstream applications can derive insight.

Gerhard Weikum, Max Planck Institute for Informatics, 2015

Data Storytelling and Visualization



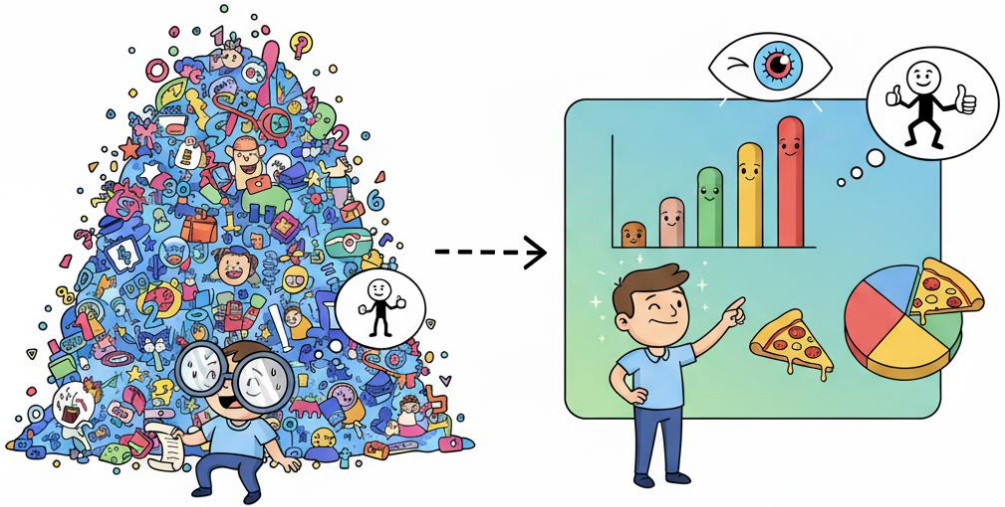
Data story telling

https://twitter.com/wcastillophd/status/1463600788140396550

Why do we visualize?

- Most datasets are far too large to be examined in their raw format.
- Visual analysis uses our **pre-attentive perception** - visual cues that humans process automatically and unconsciously.
- We can perceive and interpret these types of characteristics quickly and without any special effort.

- Example: Using the length of bars to represent sales volumes is an effective choice to indicate differences in sales between categories.



Possibilities of Visual Presentation

Length

- Very good for quantitative variables
- Poorly suited for qualitative variables

Possibilities of Visual Presentation

Width

- Limited to quantitative variables
- Poorly suited for qualitative variablest

Possibilities of Visual Presentation

Orientation

- Limited to quantitative variables
- Poorly suited for qualitative variables

Possibilities of Visual Presentation

Size

- Limited to quantitative variables
- Poorly suited for qualitative variables

Possibilities of Visual Presentation

Form

- Poor for quantitative variables
- Very suitable for qualitative variables

Possibilities of Visual Presentation

Position

- Very good for quantitative variables
- Poorly suited for qualitative variables

Possibilities of Visual Presentation

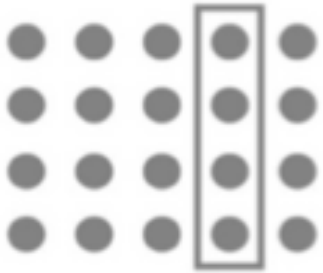
Grouping

- Limited to quantitative variables
- Very suitable for qualitative variables

Possibilities of Visual Presentation

Containment

- Poor for quantitative variables
- Limited suitable for qualitative variables



Possibilities of Visual Presentation

Hue

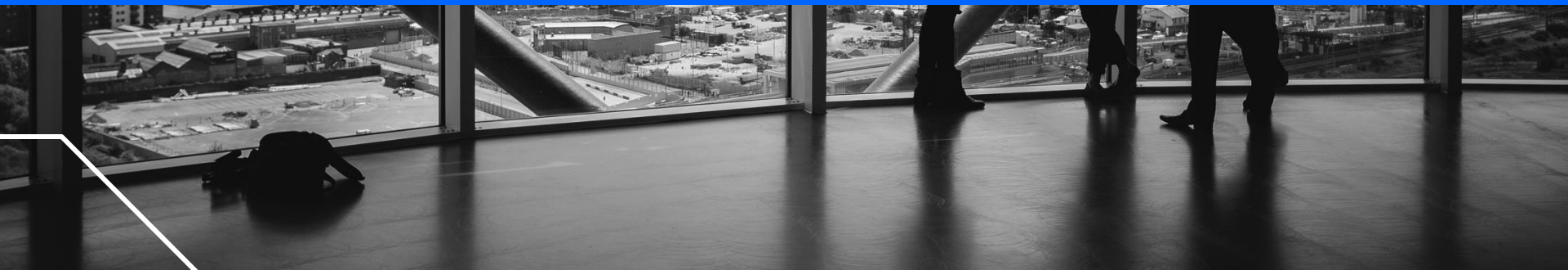
- Poor for quantitative variables
- Very suitable for qualitative variables

Possibilities of Visual Presentation

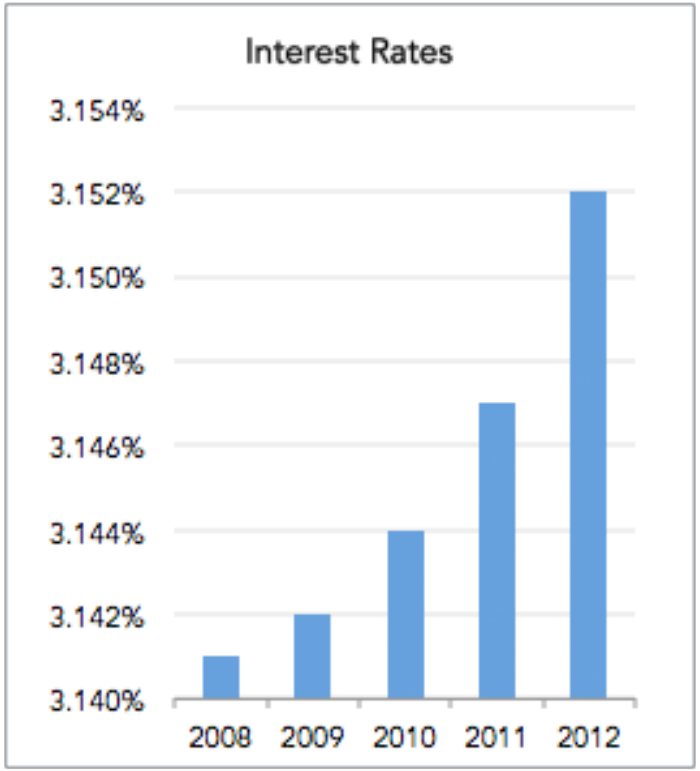
Colour intensity

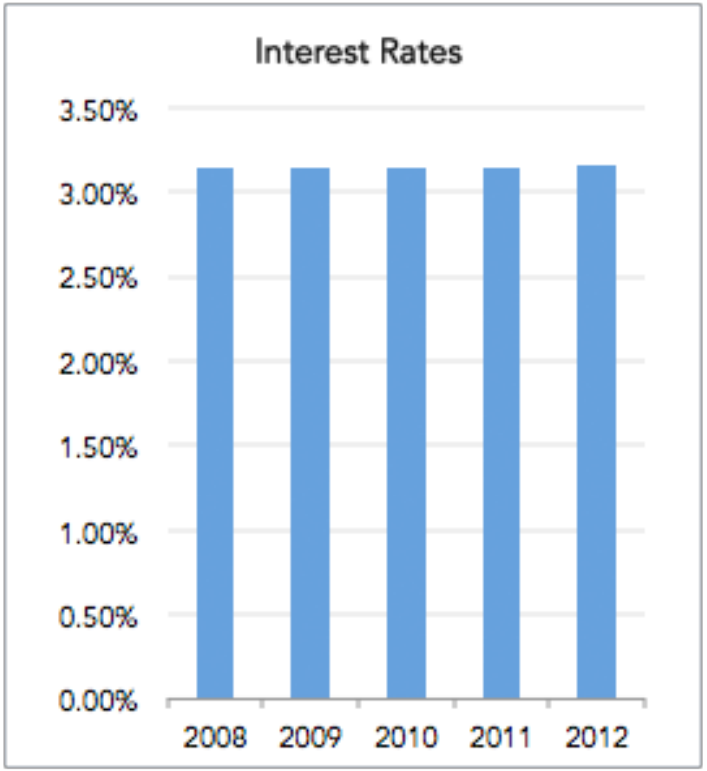
- Limited to quantitative variables
- Poorly suited for qualitative variables

Poor or misleading visualizations



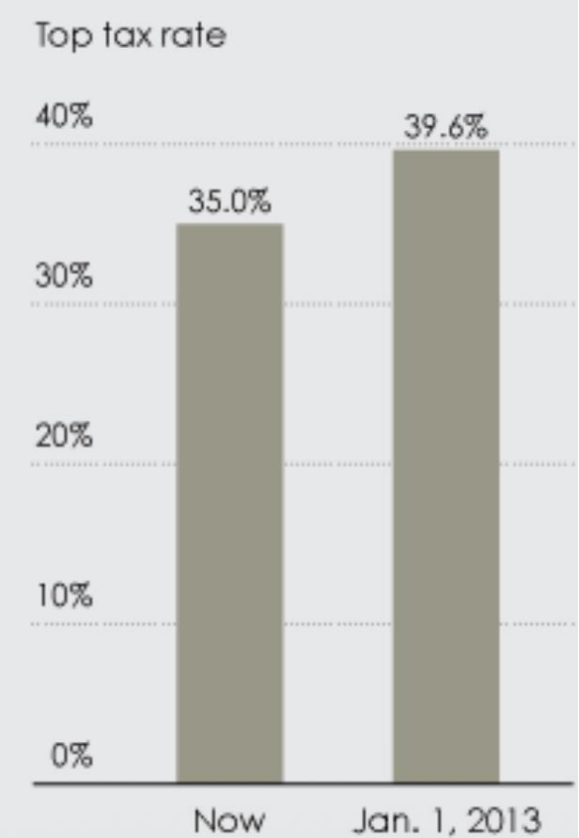
Poor or misleading visualizations

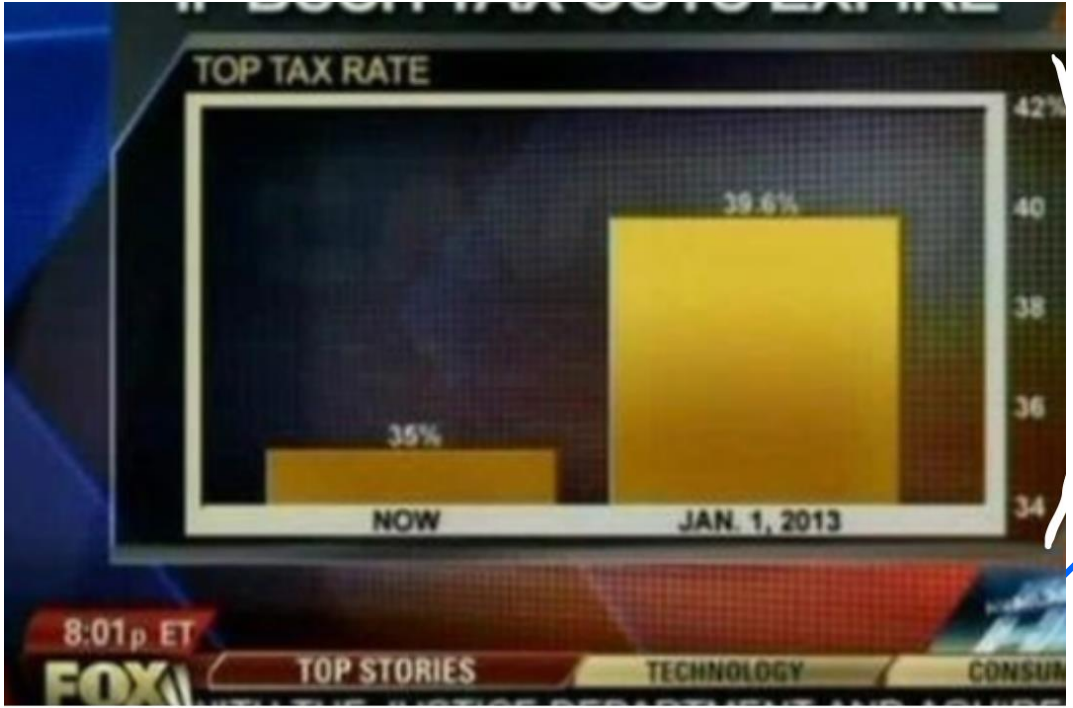




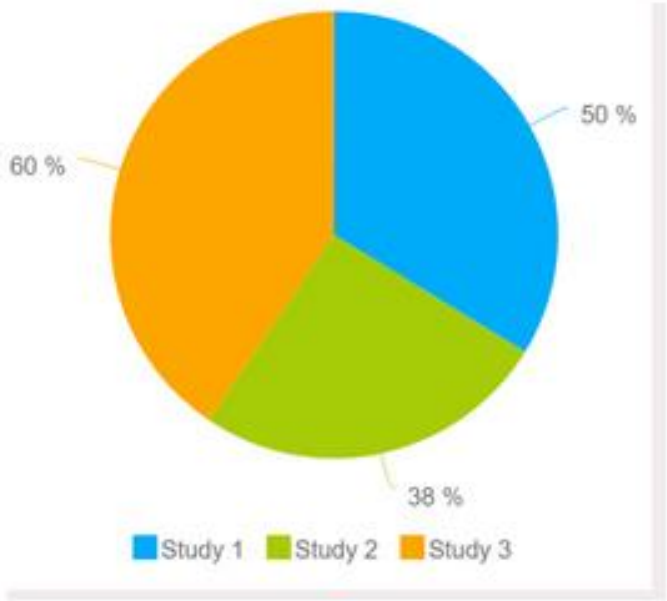
https://www.datapine.com/blog/misleading-data-visualization-examples/
https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations

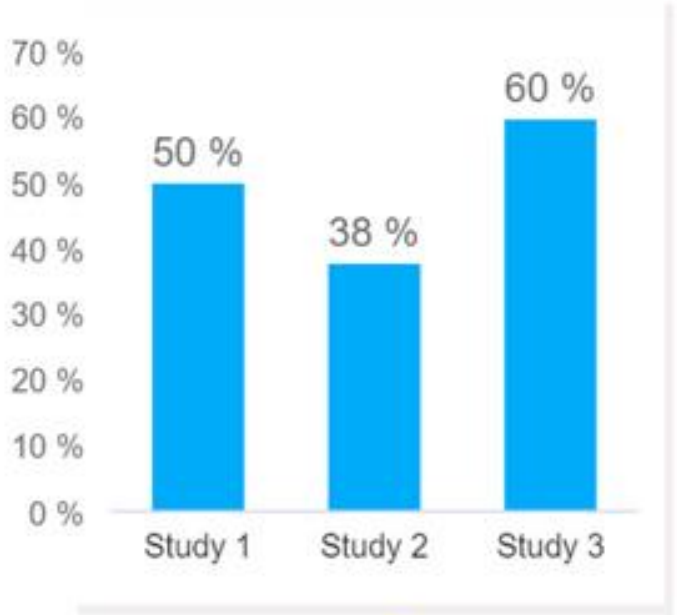
If Bush's tax rate reduction expires ...





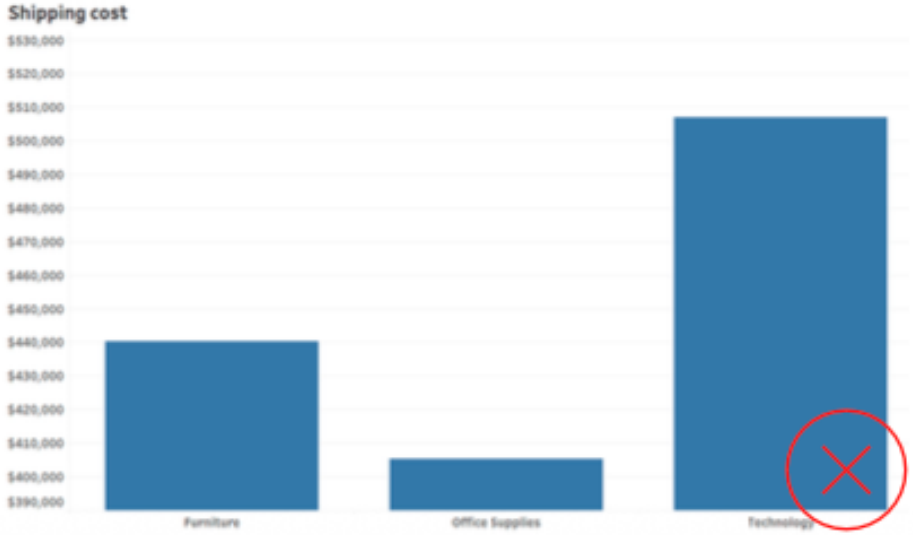
Poor or misleading visualizations

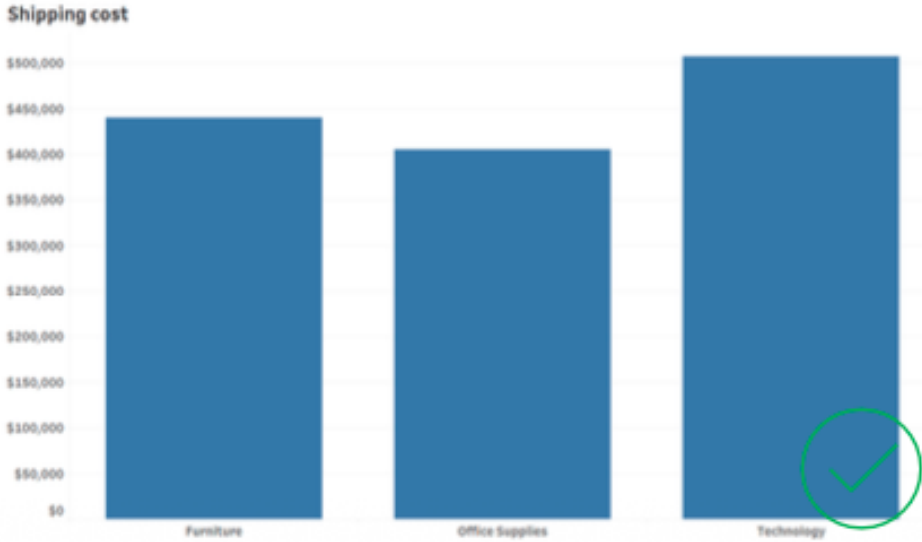




https://www.datapine.com/blog/misleading-data-visualization-examples/
https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations

Poor or misleading visualizations

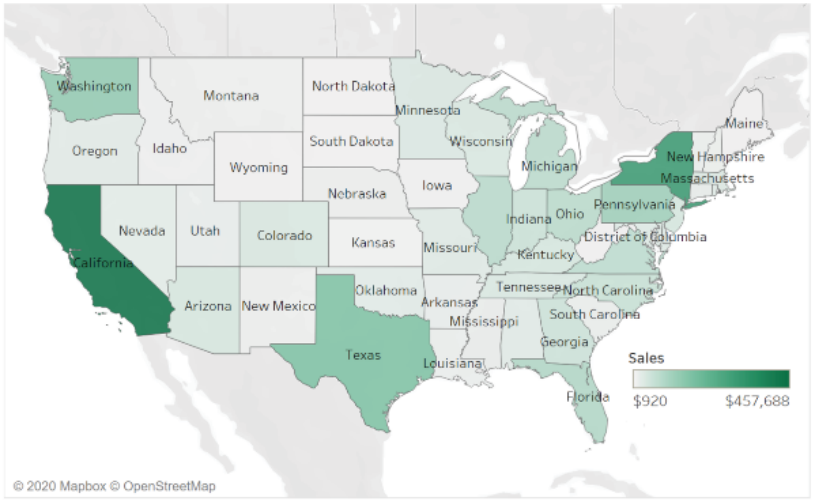




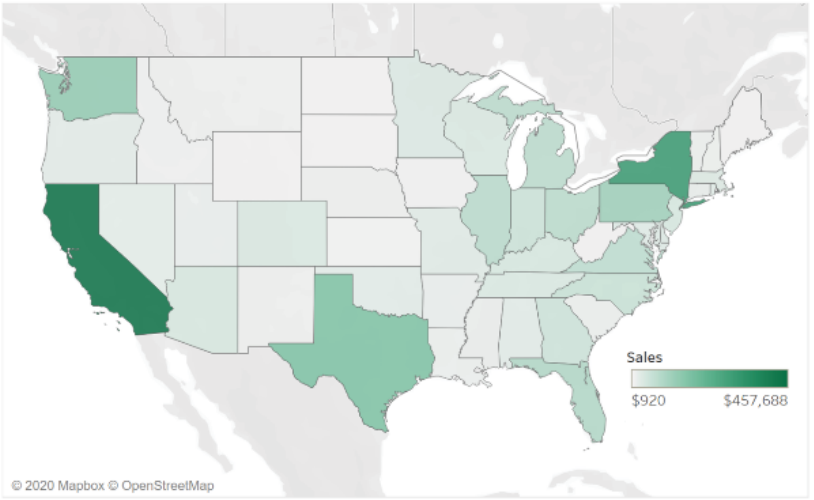
https://www.datapine.com/blog/misleading-data-visualization-examples/
https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations

Poor or misleading visualizations

total sales map



total sales map



ineffective

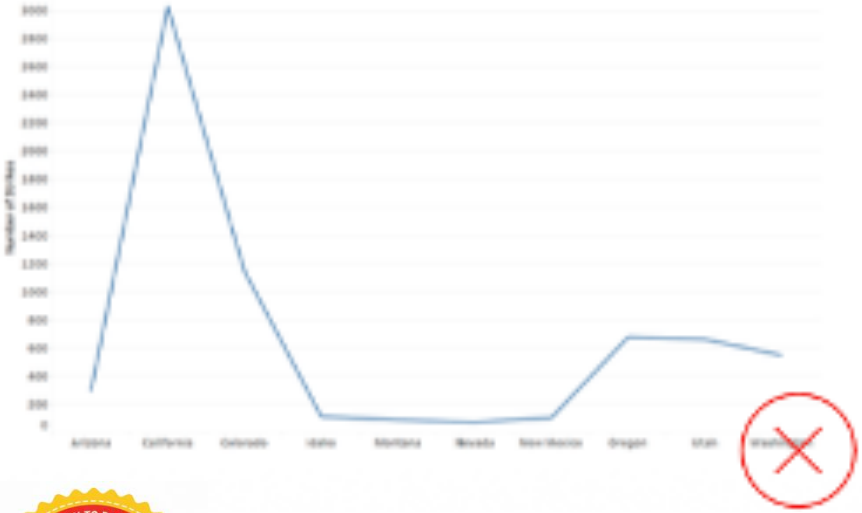
effective

https://www.datapine.com/blog/misleading-data-visualization-examples/
https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations

Poor or misleading visualizations

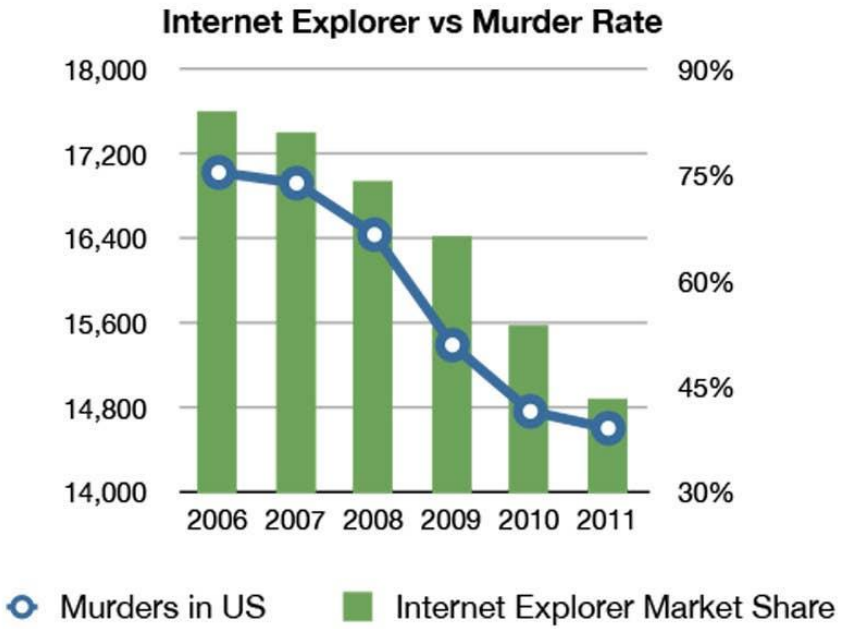
https://www.datapine.com/blog/misleading-data-visualization-examples/
https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations

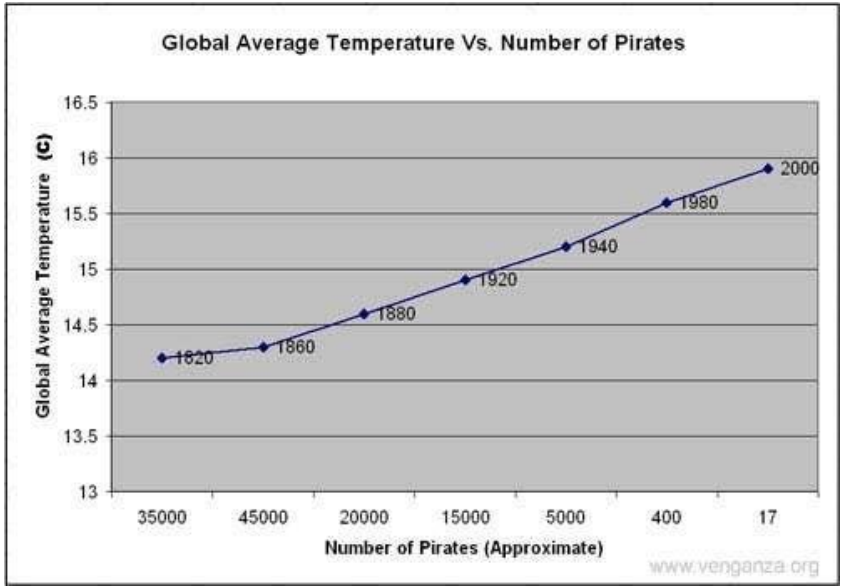
Poor or misleading visualizations



https://www.datapine.com/blog/misleading-data-visualization-examples/
https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations

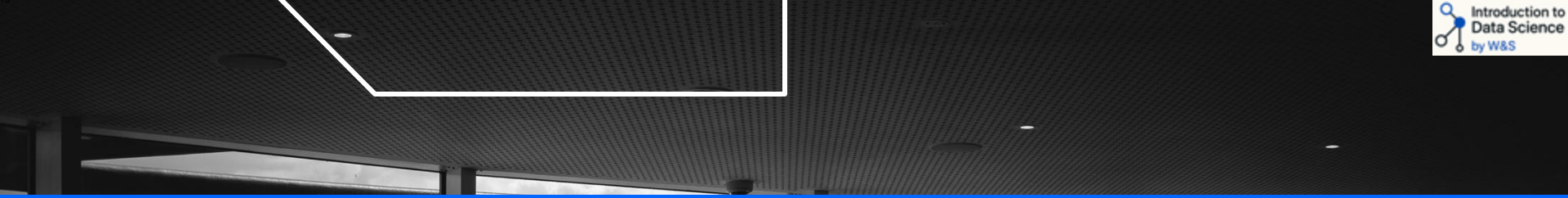
Correlation = causality?





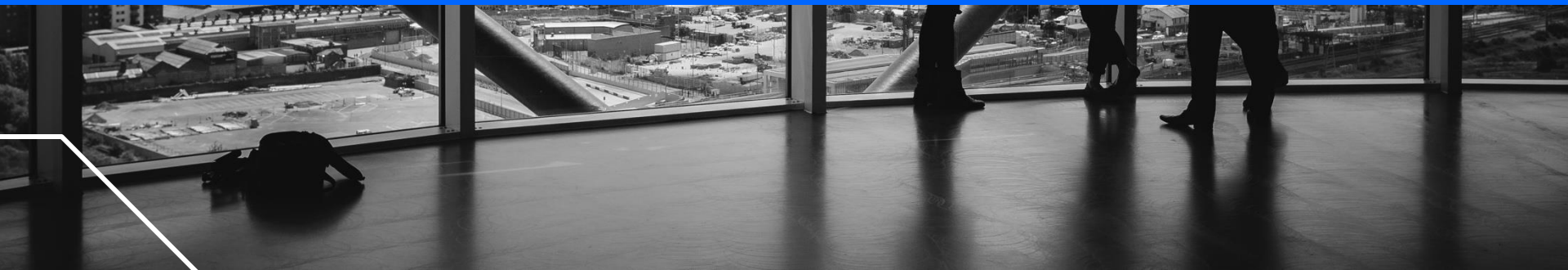
https://www.buzzfeednews.com/article/kjh2110/the-10-most-bizarre-correlations

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



Introduction to Data Science by W&S

CHART TYPES



Line Chart

- Displays the change over time for a measure
- What kind of questions does this chart answer?
  - How has this variable changed over the past period?
  - When did this variable change?
  - How quickly did this variable change?
  - What are the trends? Can future trends be derived?
- Example: Stock market prices of the last five years

Bar Chart (horizontal | vertical)

- Comparison of data of different categories (dimensions)
- What kind of questions does this chart answer?
  - Which of these categories shows the highest/lowest value?
  - Are there any extraordinary categories?
  - What is the gap (deviation) between the lowest and highest values of different categories?

- Example: Sales per department

Bullet Diagram (horizontal | vertical)

- Modification of a bar chart that shows the performance of a primary measure of achievement of key figures.
  - What kind of questions does this chart answer?
  - As with the bar chart
  - Additional: Comparison per bar against a key figure

- Example: What is the actual turnover compared to the expected turnover?

Histogram

- Representation of the distribution of values
- What kind of questions does this chart answer?
  - Are events grouped around a certain probability?
  - Which group shows the highest values?
  - Which area covers the most observations?
- Example: Students' performance in an exam

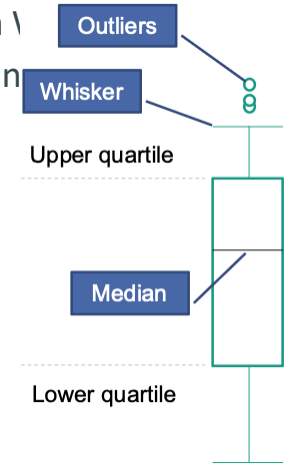
Boxplot

- Representation of the distribution within categories (dimensions)
- What kind of questions does this chart answer?
  - In which range are the values of most of the data in a category located?
  - Are there outliers in the data?
  - What is the median of values in a category?

- Example: Distribution of discounts in different product groups

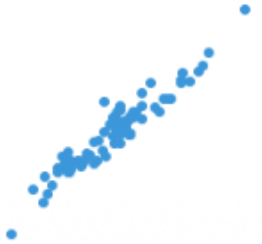
Boxplot

- Inside the box is the middle 50 percent of the data
- Whiskers (antennas) describe boundaries outside of which we speak of outliers
    - No uniform definition
    - Definition according to John \ of the whiskers to a maximum of 1.5 times the interquartile distan



Scatterplot

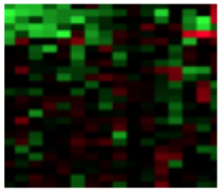
- Displays relationships between two numeric variables
- What kind of questions does this chart answer?
  - Are there any patterns when looking at the data points?
  - Are there correlation relationships between the variables?
  - Are there exits in the data?



- Example: Relationship between daily calorie intake and a person's body weight.

Heat Map

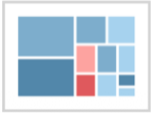
- Comparisons between two variables
- What kind of questions does this chart answer?
  - Is there a relationship between two variables?
  - Are certain areas particularly prominent?
  - Do two variables correlate?



- Example: Which nations won medals at the Olympics (divided into gold, silver,
- bronze)?

Tile Chart (Tree Map)

- Displays the proportion of an overall distribution.
- What kind of questions does this chart answer?
  - How much does this value contribute to the total?
  - How does the distribution of a variable change over time?



- Example: What proportion of total sales do an item's sales by sales region
- provide?
- Alternative chart types: pie charts, area charts, stacked bar charts

Pie Chart

- Displays the proportion of an overall distribution.
- What kind of questions does this chart answer?
  - How much does this value contribute to the total?
  - How does the distribution of a variable change over time?

- Example: What proportion of the votes did a party receive in the last elections?
- Alternative chart types: tree map, area charts, stacked bar charts

Map Charts

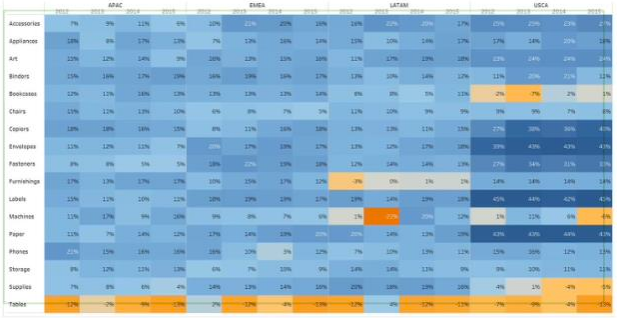
- Map charts represent positions and geographic patterns in the data.
- Variants: filled cards, point distribution cards, symbol cards, density cards...

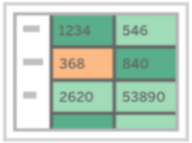
- What types of questions can this diagram answer?
  - Which place has the largest/smallest values?
  - Which region has the largest/smallest values?
  - How do you represent deviations geographically?

- Example: How is COVID-19 spreading worldwide?

Highlight Table

- Data table with color coding
- What kind of questions does this chart answer?
  - Shows all the details of the data points
  - Colors can be assigned to specific dimensions and categories or highlight key figures (quantiles, max/min values, ...)
  - Apply color markers from diagrams and thus display their details.





Tipps and Guidelines

Order your data

When displaying the value of several entities, ordering them makes the graph much more insightful.

To cut or not to cut?

Cutting the Y-axis is one of the most controversial practice in data viz. See why.

The spaghetti chart

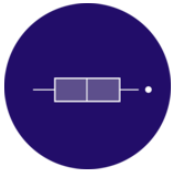
A line graph with too many lines becomes unreadable: it is called a spaghetti graph.

Pie chart

The human eye is bad at reading angles. See how to replace the most criticized chart ever.

Play with histogram bin size

Always try different bin sizes when you build a histogram, it can lead to different insights.



Do boxplots hide information?

Boxplots are a great way to summarize a distribution but hide the sample size and their distribution.

The problem with error bars

Barplots with error bars must be used with great care. See why and how to replace them.

Too many distributions.

If you need to compare the distributions of many variables, don't clutter your graphic.

https://www.data-to-viz.com/caveats.html

Presenting results

Promises

- Enhanced Analyses and Predictions: Leveraging large volumes of data can enable deeper insights and more accurate forecasts than ever before.
- Uncertainty at the Micro Level, Accuracy at the Macro Level: Big Data allows for precise identification of patterns and trends at the macro level, despite uncertainty at the individual level.
- Quality is Crucial: Despite the vast amounts of data, ensuring data quality is essential for drawing valid conclusions.

Risks

- Incorrect Modeling and Conclusions: The complexity and volume of Big Data can lead to errors in data modeling, resulting in false or misleading outcomes.
- Privacy and Individual Rights: Processing large datasets poses risks to privacy and can conflict with individual rights.

Recommendations for Mitigating Risks

- Rigorous Data Verification: Implement strict data quality verification processes to ensure the integrity of analyses.
- Transparent Modeling: Promote transparency in your modeling processes to build trust and avoid misinterpretations.
- Preserving Privacy: Adhere to privacy regulations and practices to protect individuals' rights and foster trust in Big Data initiatives.
- Reason about results: be able to explain; why have predictions been made?
- Reason about input: data set; metrics

Use of Diagrams

- Usually there are data sets with many variables (characteristics) or
- we want to determine the utility value of other variables

- Application of diagrams
  - Often the same / different diagrams are used one after the other / combined
  - Decision on the type of visualization based on
    - variables,
    - dimensions in our data, and
    - the question asked


HANDS-ON
TIME
INTRODUCTION TO DATA SCIENCE
BY W&S
YOUR TURN!
EXERCISE / QUESTION


TITANIC DATASET
TITANIC DATASET
INTRODUCTION TO DATA SCIENCE
BY W&S
YOUR TURN!
EXERCISE / QUESTION

Pandas Basics

https://media.datacamp.com/legacy/image/upload/v1676302204/Marketing/Blog/Pandas_Cheat_Sheet.pdf

Plotly Basics

datacamp plotly

Data Visualization with Plotly Express in Python

Learn Plotly online at www.DataCamp.com

What is plotly?

Plotly Express is a high-level data visualization package that allows you to create interactive plots with very little code. It is built on top of Plotly Graph Objects, which provides a lower-level interface for developing custom visualizations.

Interactive controls in Plotly

Plotly plots have interactive controls shown in the top-right of the plot. The controls allow you to do the following:

- **Download plot as a png:** Save your interactive plot as a static PNG.
- **Zoom:** Zoom in on a region of interest in the plot.
- **Pan:** Move around in the plot.
- **Box Select:** Select a rectangular region of the plot to be highlighted.
- **Lasso Select:** Draw a region of the plot to be highlighted.
- **Autoscale:** Zoom to a "best" scale.
- **Reset axes:** Return the plot to its original state.
- **Toggle Spike Lines:** Show or hide lines to the axes whenever you hover over data.
- **Show closest data on hover:** Show details for the nearest data point to the cursor.
- **Compare data on hover:** Show the nearest data point to the x-coordinate of the cursor.

Plotly Express code pattern

The code pattern for creating plots is to call the plotting function, passing a data frame as the first argument. The `x` argument is a string naming the column to be used on the x-axis. The `y` argument can either be a string or a list of strings naming column(s) to be used on the y-axis.

```
px.plotting_fn(dataframe, # Dataframe being visualized
               x=["column-for-x-axis"], # Accepts a string or a list of strings
               y=["columns-for-y-axis"], # Accepts a string or a list of strings
               title="Overall plot title", # Accepts a string
               xaxis_title="X-axis title", # Accepts a string
               yaxis_title="Y-axis title", # Accepts a string
               width=width_in_pixels, # Accepts an integer
               height=height_in_pixels) # Accepts an integer
```

Common plot types

Import plotly

```
# import plotly express as px
import plotly.express as px
```

Scatter plots

```
# Create a scatterplot on a DataFrame named clinical_data
px.scatter(clinical_data, x="experiment_1", y="experiment_2")
```

Set the `size` argument to the name of a numeric column to control the size of the points and create a bubble plot.

Line plots

```
# Create a lineplot on a DataFramed named stock_data
px.line(stock_data, x="date", y=["FB", "AMZN"])
```

Set the `line_dash` argument to the name of a categorical column to have dashes or dots for different lines.

Bar plots

```
# Create a barplot on a DataFramed named commodity_data
px.bar(commodity_data, x="nation", y=["gold", "silver", "bronze"],
       color_discrete_map={"gold": "yellow",
                           "silver": "grey",
                           "bronze": "brown"})
```

Swap the `x` and `y` arguments to draw horizontal bars.

Histograms

```
# Create a histogram on a DataFramed named bill_data
px.histogram(bill_data, x="total_bill")
```

Set the `nbins` argument to control the number of bins shown in the histogram.

Heatmaps

```
# Create a heatmap on a DataFramed named iris_data
px.imshow(iris_data.corr(numeric_only=True),
          zmin=-1, zmax=1, color_continuous_scale='rdbu')
```

Set the `text_auto` argument to `True` to display text values for each cell.

Customizing plots in plotly

The code pattern for customizing a plot is to save the figure object returned from the plotting function, call its `.update_traces()` method, then call its `.show()` method to display it.

```
# Create a plot with plotly (can be of any type)
fig = px.some_plotting_function()
# Customize and show it with .update_traces() and .show()
fig.update_traces()
fig.show()
```

Customizing markers in Plotly

When working with visualizations like scatter plots, lineplots, and more, you can customize markers according to certain properties. These include:

- `size`: set the marker size
- `color`: set the marker color
- `opacity`: set the marker transparency
- `line`: set the width and color of a border
- `symbol`: set the shape of the marker

```
# In this example, we're updating a scatter plot named fig_sct
fig_sct.update_traces(marker={"size": 24,
                              "color": "magenta",
                              "opacity": 0.5,
                              "line": {"width": 2, "color": "cyan"},
                              "symbol": "square"})
fig_sct.show()
```

Customizing lines in Plotly

When working with visualizations that contain lines, you can customize them according to certain properties. These include:

- `color`: set the line color
- `dash`: set the dash style ("solid", "dot", "dash", "longdash", "dashdot", "longdashdot")
- `shape`: set how values are connected ("linear", "spline", "hv", "vh", "hvh","vhv")
- `width`: set the line width

```
# In this example, we're updating a scatter plot named fig_ln
fig_ln.update_traces(patch={"line": {"dash": "dot",
                                     "shape": "spline",
                                     "width": 4}})
fig_ln.show()
```

Customizing bars in Plotly

When working with barplots and histograms, you can update the bars themselves according to the following properties:

- `size`: set the marker size
- `color`: set the marker color
- `opacity`: set the marker transparency
- `line`: set the width and color of a border
- `symbol`: set the shape of the marker

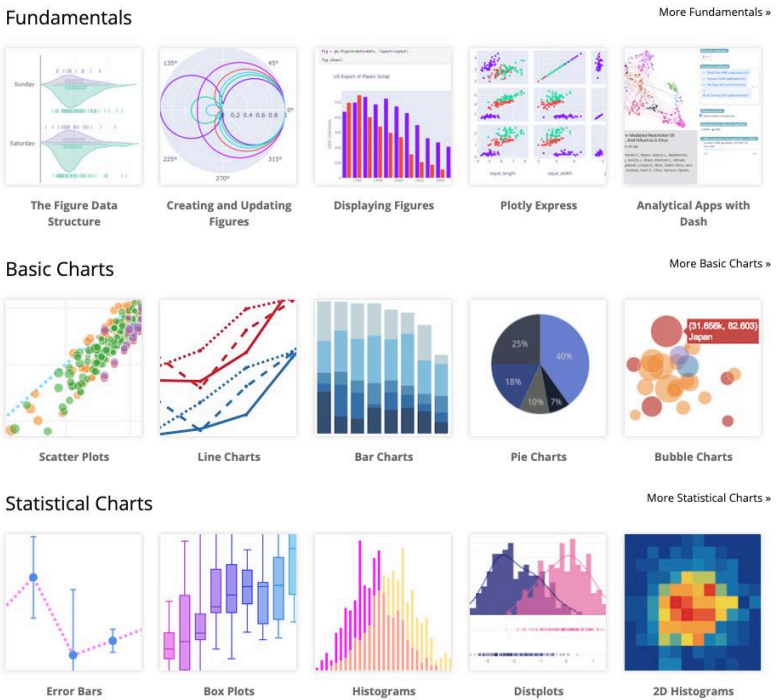
```
# In this example, we're updating a scatter plot named fig_bar
fig_bar.update_traces(marker={"color": "magenta",
                              "opacity": 0.5,
                              "line": {"width": 2, "color": "cyan"}})
fig_bar.show()

# In this example, we're updating a histogram named fig_hst
fig_hst.update_traces(marker={"color": "magenta",
                              "opacity": 0.5,
                              "line": {"width": 2, "color": "cyan"}})
fig_hst.show()
```

Learn Data Skills Online at www.DataCamp.com

datacamp

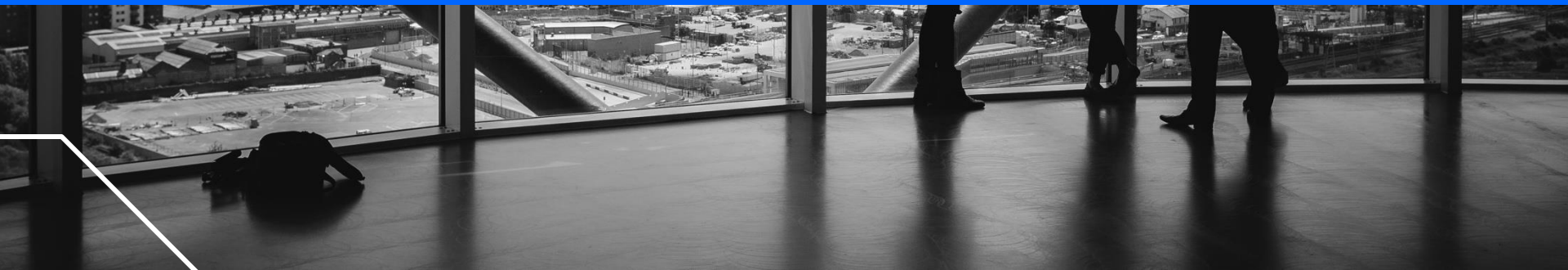
https://media.datacamp.com/legacy/image/upload/v1668605954/Marketing/Blog/Plotly_Cheat_Sheet.pdf

Plotly Documentation



https://plotly.com/python/

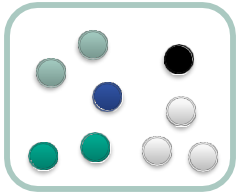
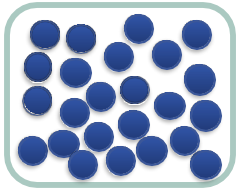
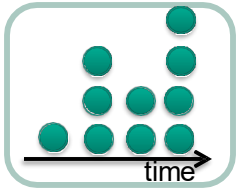
Challenges



Whats hard about data science

- Getting the data (usually)
- Overcoming assumptions
- Communication
    - With domain experts
    - Expectation management for client
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototype to Production transition
- Data pipeline complexity (team boundary)

Data Challenges (1)

	Variety	Volume	Velocity
			
Challenge	Handling multiplicity of types, sources and formats	Dealing with large volumes of data	Streams, sensors, near real-time data
Impacted Tasks	Data integration	Storage, processing & analytics	Processing & analytics
Solution	Semantic technologies are a good fit	Distributed storage & parallel processing	Real-time technologies

Data Challenges (2)

- **Data veracity:** coping with uncertainty, imprecision, missing values, misstatements or untruths.

- **Data quality:** determining the quality of datasets and relevance to particular issues. Depends on the use case:
  - How broad/complete is the data?
  - How fine is the sample resolution? How timely are the readings?
  - Does the data contain any “noise” (errors)? Is it representative?

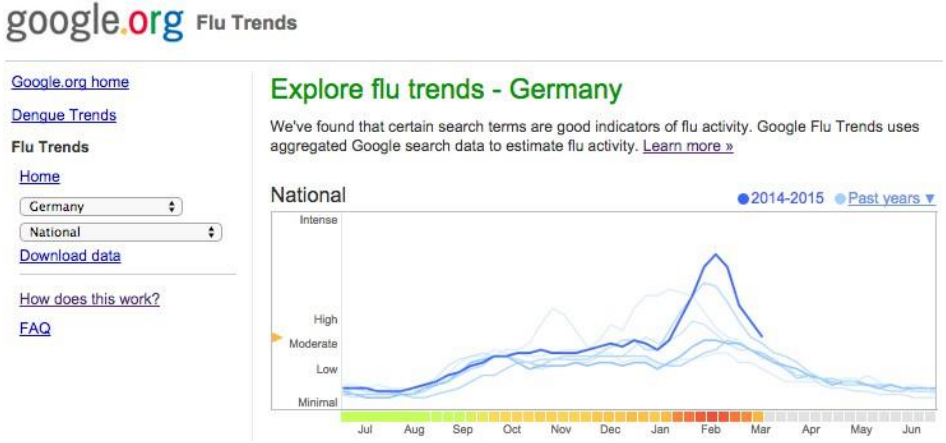
Data Challenges (3)

- **Data discovery:** finding relevant data from enormous amount of data available on the Web.

- **Data dogmatism:** analysis of Big Data can offer remarkable insights. However, data analysis should not entirely replace domain expert knowledge, but act as a tool to support /confirm facts.
    - E.g., Google Flu Trends

Data Challenges (3)

Example: Data dogmatism



Source: https://www.google.org/flutrends/de/#DE

"By counting how often we see these **[flu-related topics] search queries**, we can estimate **how much flu is circulating** in different countries and regions around the world."

Data Challenges (3)

Example: Data dogmatism

Comparison of Google Flu Trends model against medical reports.

SCIENCE BIG DATA

Google's Flu Project Shows the Failings of Big Data

Bryan Walsh @bryanrwalsh | March 13, 2014

A new study shows that using big data to predict the future isn't as easy as it looks—and that raises questions about how Internet companies gather and use information

Source: http://time.com/23782/google-flu-trends-big-data-problems/

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer[1,2,*], Ryan Kennedy[1,3,4], Gary King[3], Alessandro Vespignani[5,6,3]

+ Author Affiliations

*Corresponding author. E-mail: d.lazer@neu.edu.

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

Read the Full Text

Source: http://www.sciencemag.org/content/343/6176/1203

Process Challenges (1)

Big-Data Scientists *"Janitor Work"*

50% - of data scientists' time
80% - spent in "Data wrangling"

Data acquisition:
Data availability
Data permissions

Aligning/integrating data from different sources:
Syntactic challenge: Data in different formats
Semantic challenges: Resolving when two objects are the same, describing relationships between data points, resolving inconsistencies

Transforming, cleaning and organizing the data into a form suitable for analysis

Source: "For Big-Data Scientists, 'Janitor Work' is Key Hurdle to Insights" in The New York Times. 2014.

Process Challenges (2)

Modeling data:

Mathematically: mathematical models to describe the data Statistical methods or Machine Learning

Simulations

Knowledge representation: ontologies and rules

Understanding the output:

Interpreting the results

Visualizing

Sharing the results

Data privacy, security, and Governance:

- Ensuring that data is used correctly (abiding by its intended uses and relevant laws).
- Tracking how the data is used, transformed, derived, …
- Managing data lifecycle.

*"Many data warehouses contain sensitive data such as personal data. There are **legal** and **ethical concerns** with accessing such data. So the data must be secured and access controlled as well as logged for audits".*

Michael Blaha, Modelsoft Consulting Corporation, 2012

Source: http://www.odbms.org/blog/2012/03/data-modeling-for-analytical-data-warehouses-interview-with-michael-blaha/

Data science – two sides of the same coin

Opportunities	Risks
• Discover potential • Develop new services • Informed decisions based on forecasts • Service and product improvement	• Surveillance • Manipulation • Uselessness • Data protection

Bis Donnerstag

samuel.schlenker@lehre.dhbw-stuttgart.de