Business Intelligence

Data Literacy

tdi@ek.dk

Agenda

BI Intro Recap

Data Literacy

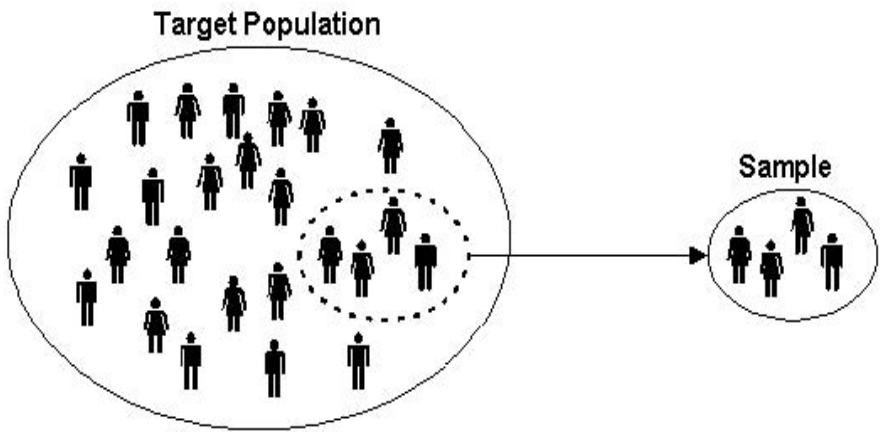
- Introduction to Terminology
- Overview of Data Categories

- Data Ingestion and data Wrangling
- Data Quality
- Problem Solving

Data Literacy

The ability to explore, understand, and communicate with data

Data Definitions



- Object – units of interest
- Attributes –features of the objects that can be measured – in quantity or in quality
  - in programming, the attributes are described by variables, for which the measures are values
- Data is a recorded set of values of qualitative or quantitative variables about one or more objects
- The set of objects of interest is called population
- The set of available objects is called sample
- Various attributes of the population create multi-dimensional data

BI Data Categories

Image from Wikipedia

- Quantitative = numeric measures
  - Numeric values measured in even intervals - age, distance, temperature
  - Measurement needs a scale
    - Discrete - 1, 2, 3, …
    - Continuous – 1.1, 1.11, 1.111, …
- Qualitative = labels, symbolic names
  - Ordinal data
    - there is a logical rank-order relationship between the range of values – such as {never, sometimes, mostly, always}; {good, better, best}
  - Nominal data
    - categories cannot be ranked - Asia, Europe, America, Australia; male, female

Examples of Quantitative vs Qualitative

Quantitative - measures	Qualitative / Ordinal	Qualitative / Nominal
Weight (10 kg. 35 kg, 100 kg)	Gold Silver Bronze	North America Europe Asia
Cost ($500, $2.5 M, $4 B)	Excellent Good Poor	Alice Bob Chris
Discount (0%, 2.5%, 5%)	January February March	Wine Beer Water
We need a scale to measure them	Ordered in some way	Equivalent in meaning

Measurement Scales

Numeric Measurement Scales

Interval

- no true zero
- the measurement can represent values above and below zero

Ratio

- there is true zero
- the measurement starts counting from zero

Scale					
Nominal	Numbers Assigned to Runners				Finish
Ordinal	Rank Order of Winners	 Third place	 Second place	 First place	Finish
Interval	Performance Rating on a 0 to 10 Scale	8.2	9.1	9.6	
Ratio	Time to Finish, in	15.2	14.1	13.4	

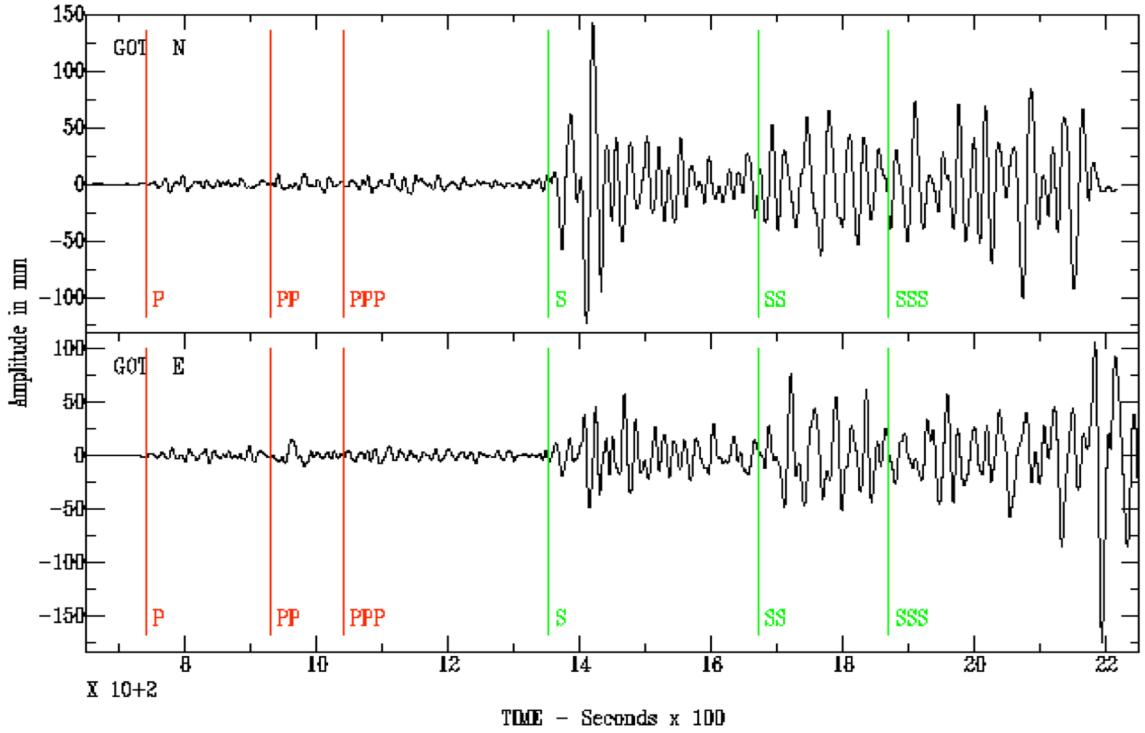
cphbusiness

Operations and Transformations with Data Categories

category	operations with a value	transformations to a value	explanation of the transformation
Nominal	(=, ≠)	All kind of permutations	If all CPR numbers get replaced, their meaning wouldn't change
Ordinal	(<, >)	By a function *new_value = f(old_value)*	{good, better, best} can be transformed to {1, 2, 3} or to{ 0.5, 1, 10}
Interval Quantities	(+, -)	*new_value =a * old_value + b, where a and b are constants*	Fahrenheit -> Celsius replacement of scales

Question

- How would you represent each of these candies as data for data analysis?
- The earthquake seismogram?



Discrete or Continuous?

- mail sender
- speed
- phone number
- shoe size

Type Conversion

How can you convert (change) the type of

- speed
- temperature

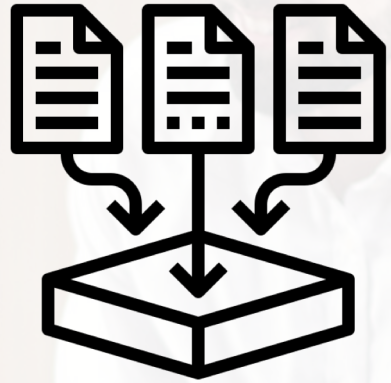
From discret to continuous?

From continuous to discret?

Minds-On

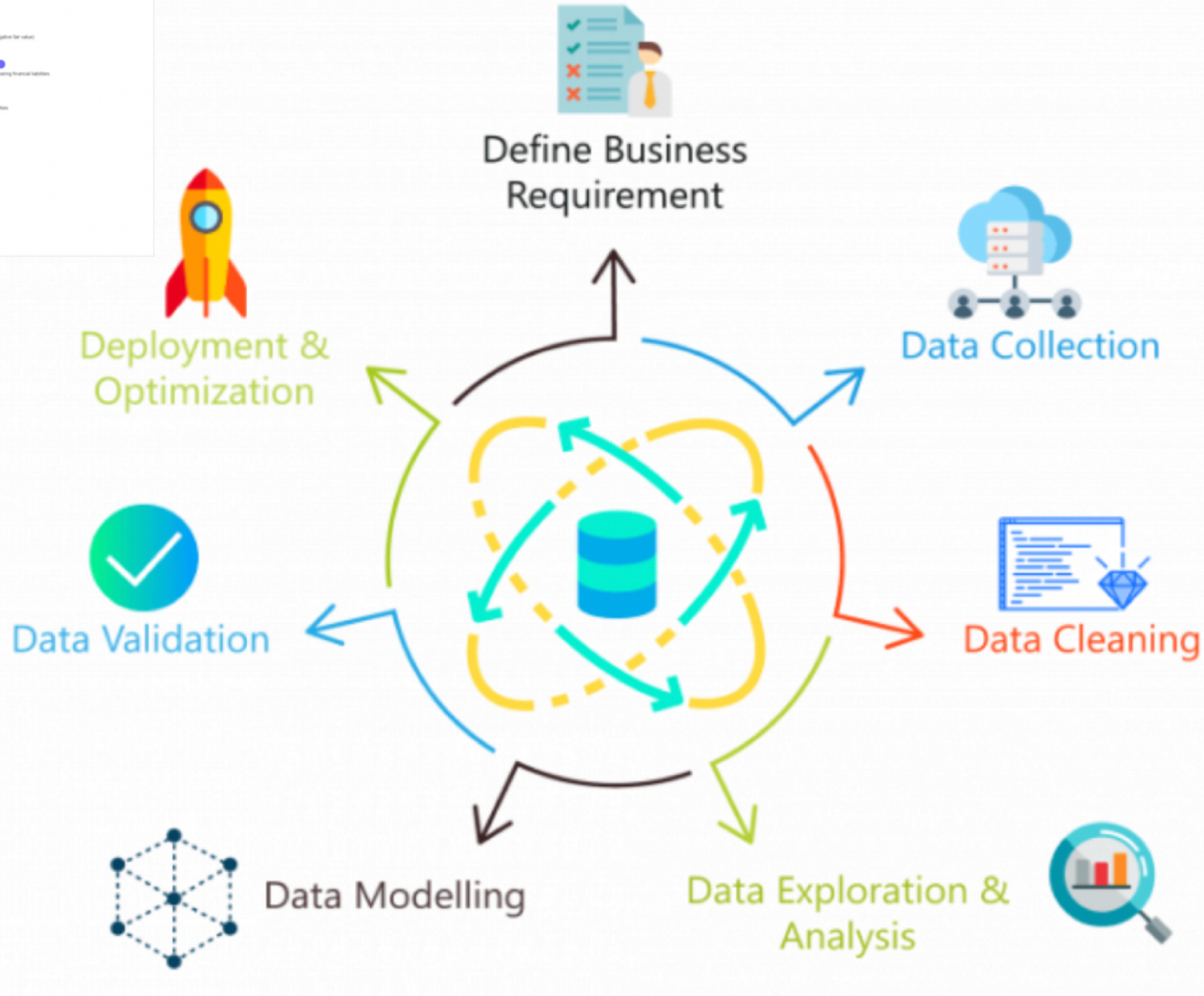
BI Data Categories

A-B Exercise

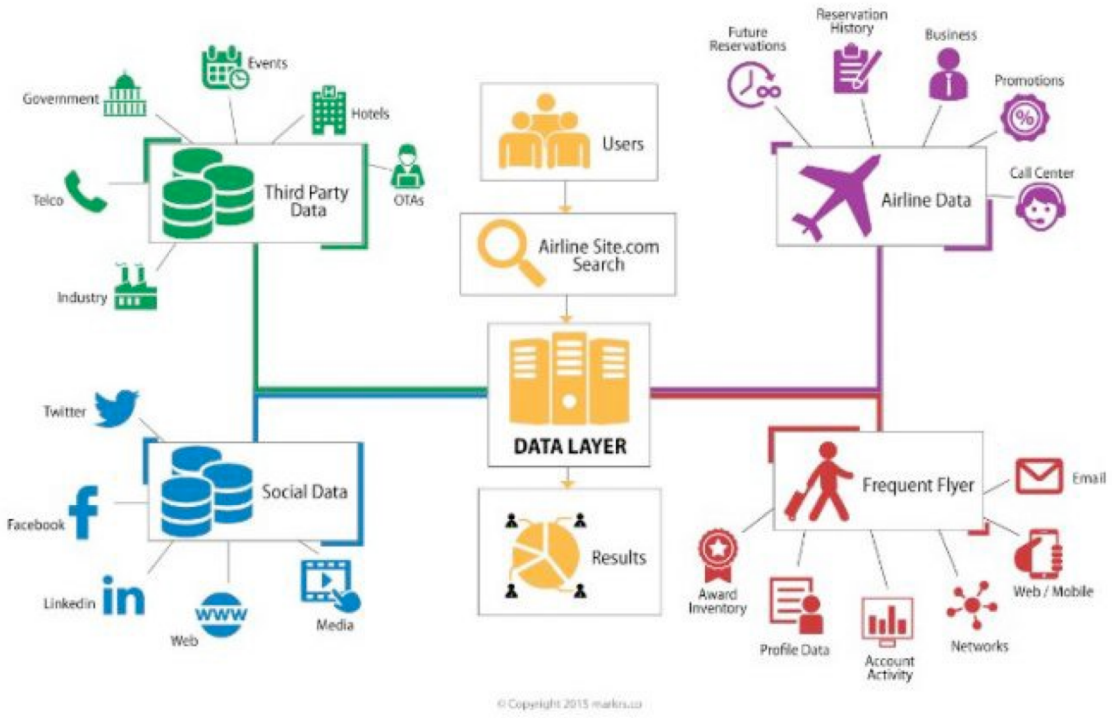


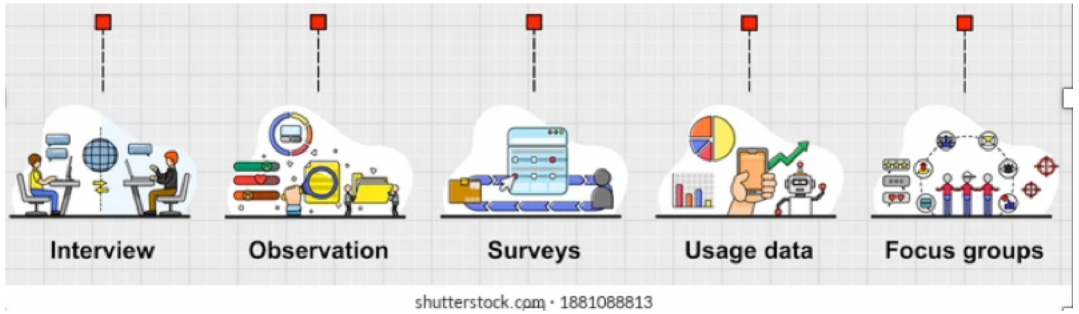
Data Ingestion and Wrangling

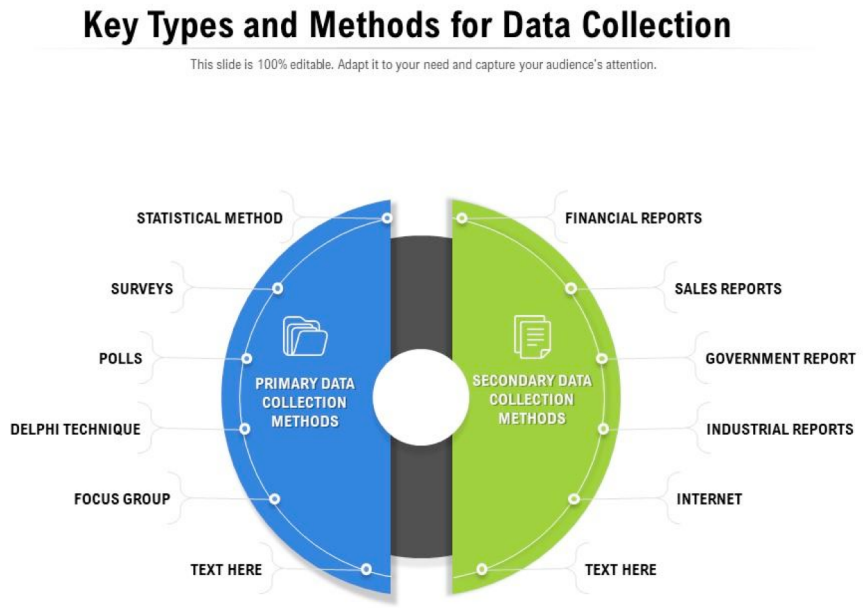
Data Collection and Pre-processing

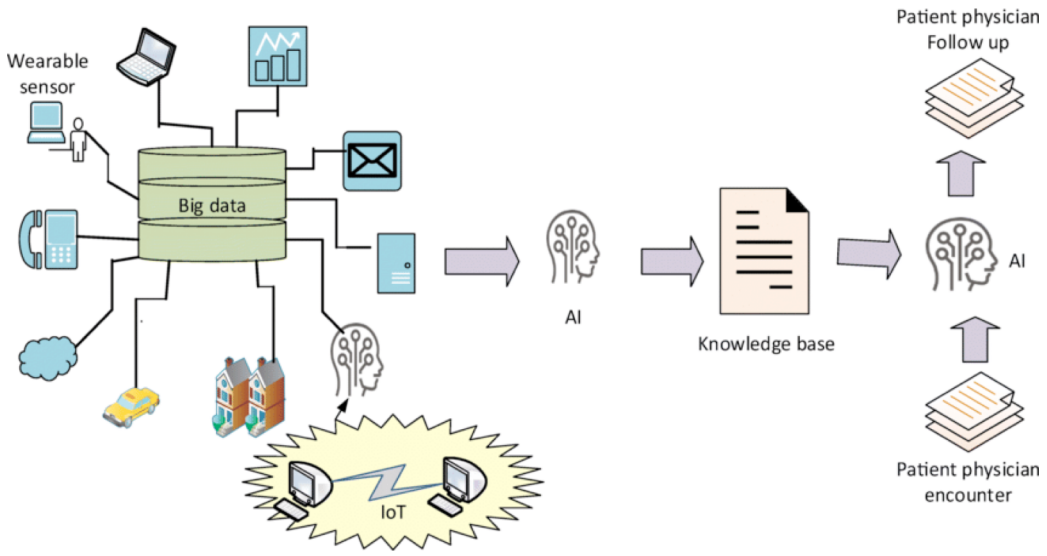

Define Business Requirement
Data Collection
Data Cleaning
Data Exploration & Analysis
Data Modelling
Data Validation
Deployment & Optimization

Data Comes From Everywhere



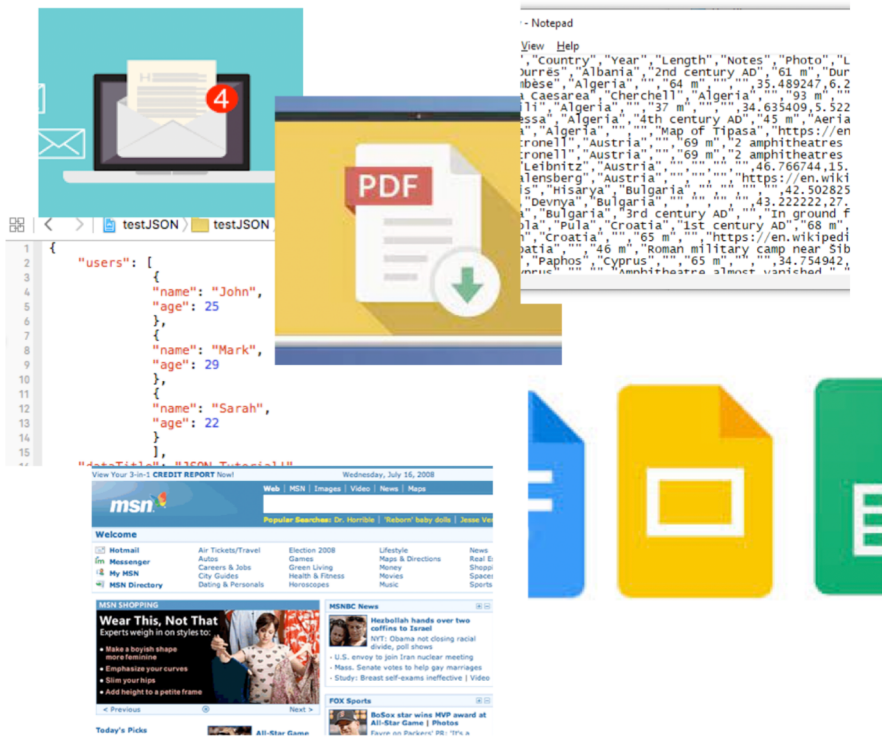






cphbusiness

Data Ingestion and Data Wrangling



Data Ingestion is the process of obtaining, importing, and processing data for later use or storage in a database.

This can be achieved manually, or automatically using a combination of software and hardware tools designed specifically for this task.

IBM

Data Wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.

Wikipedia

cphbusiness

Ingestion and Wrangling Operations

Data Cleaning

Finding and removing incorrect and inaccurate records from a set or a data source

- examples for garbage data: duplicate values, dummy values, missing data, and contradictory data
- examples for a cause: corruption in the technical systems – collection, transmission, storage

Data cleaning includes activities like

- removing typographical errors
- validating and correcting values
- harmonizing and standardizing data

Data Transformation

Converting and mapping data from one format to another format

- first, data is extracted from a data source in its raw format
- then, it is parsed into a predefined data structure or processed by an algorithm
- finally, it is stored in a storage unit for future use

Different tools for data wrangling are available

Example
https://www.varonis.com/blog/free-data-wrangling-tools/

Data Engineering

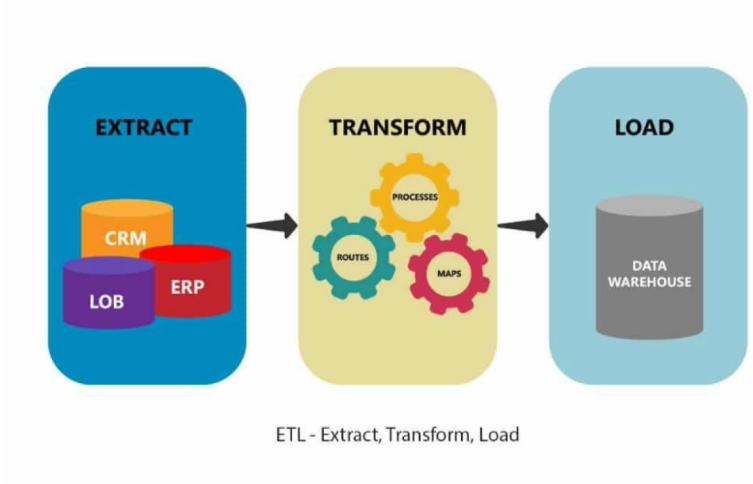
Re-engineering the initially available attributes

Can include:

- ignoring insignificant attributes
- calculation and generation of new attributes
- aggregating or merging attributes
- splitting attributes
- replacing all available attributes with a more appropriate set

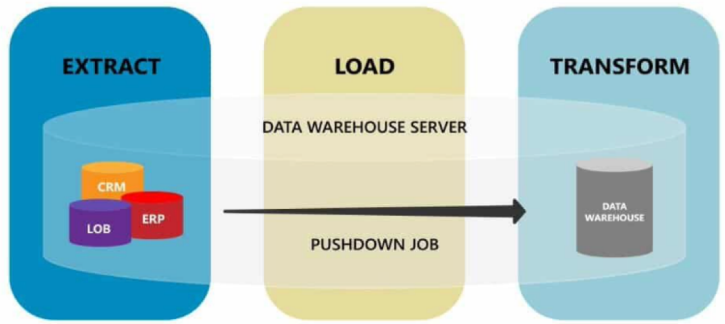
ETL and ELT

ETL- Extract, Transform, Load



ETL - Extract, Transform, Load

ELT- Extract, Load, Transform



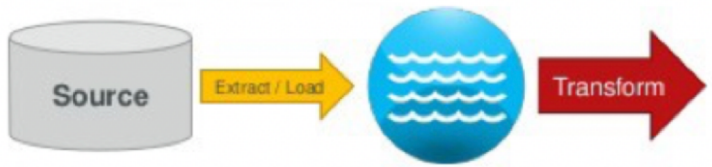
Pushdown Optimization Mode

https://www.astera.com/type/blog/etl-vs-elt-whats-the-difference/

Data Warehouses use a traditional ETL Process

Data is transformed when it enters the data warehouse

Data Lakes make use of the ELT Process

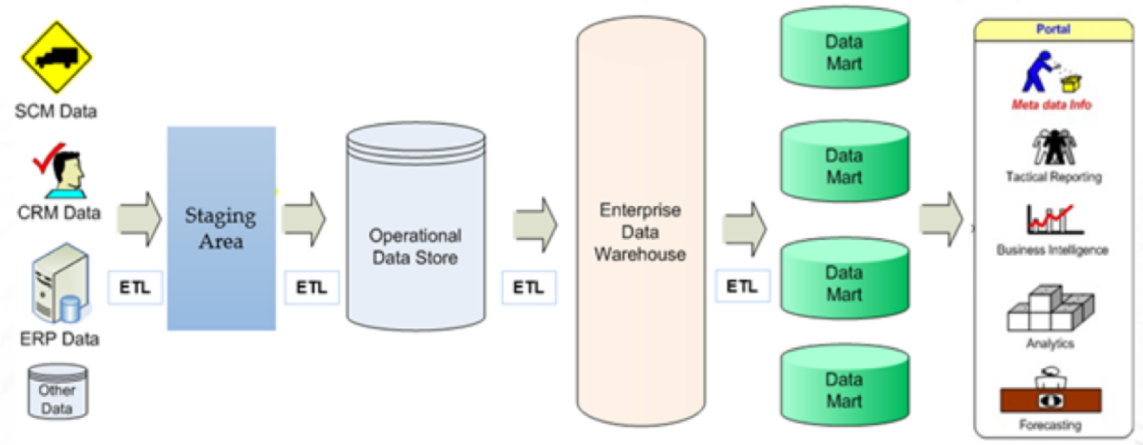


Data is transformed when it is retrieved from the data lake

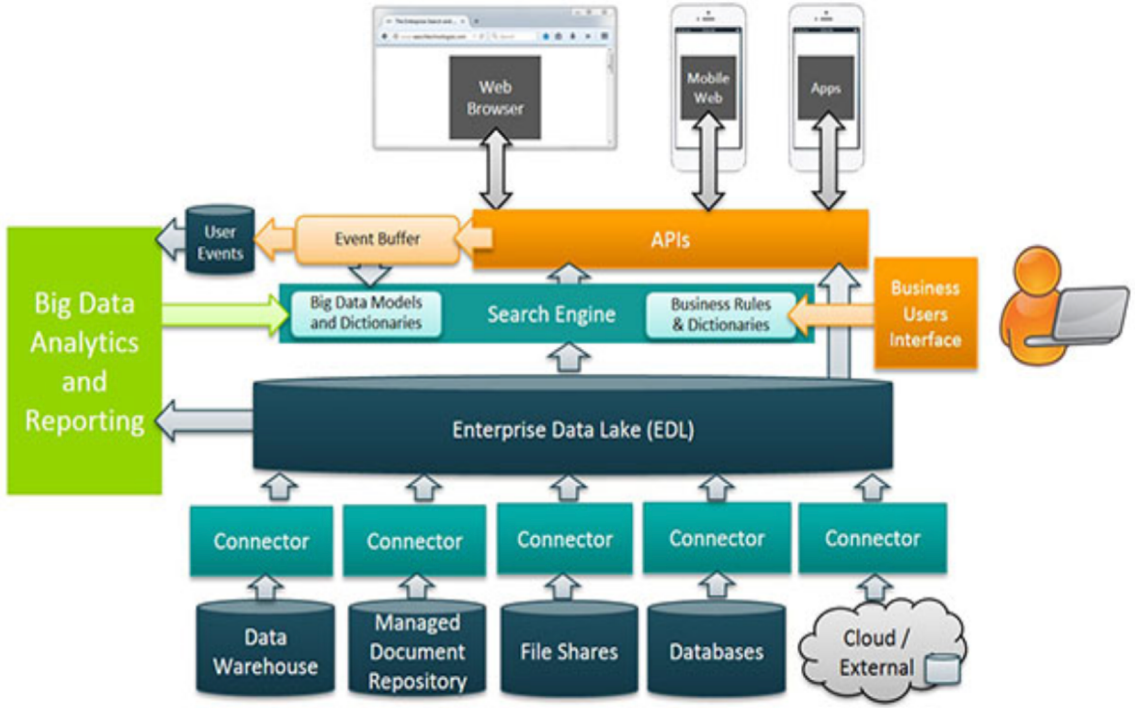
cphbusiness

Data Warehouse vs Data Lake

Data Warehouse + Data Marts

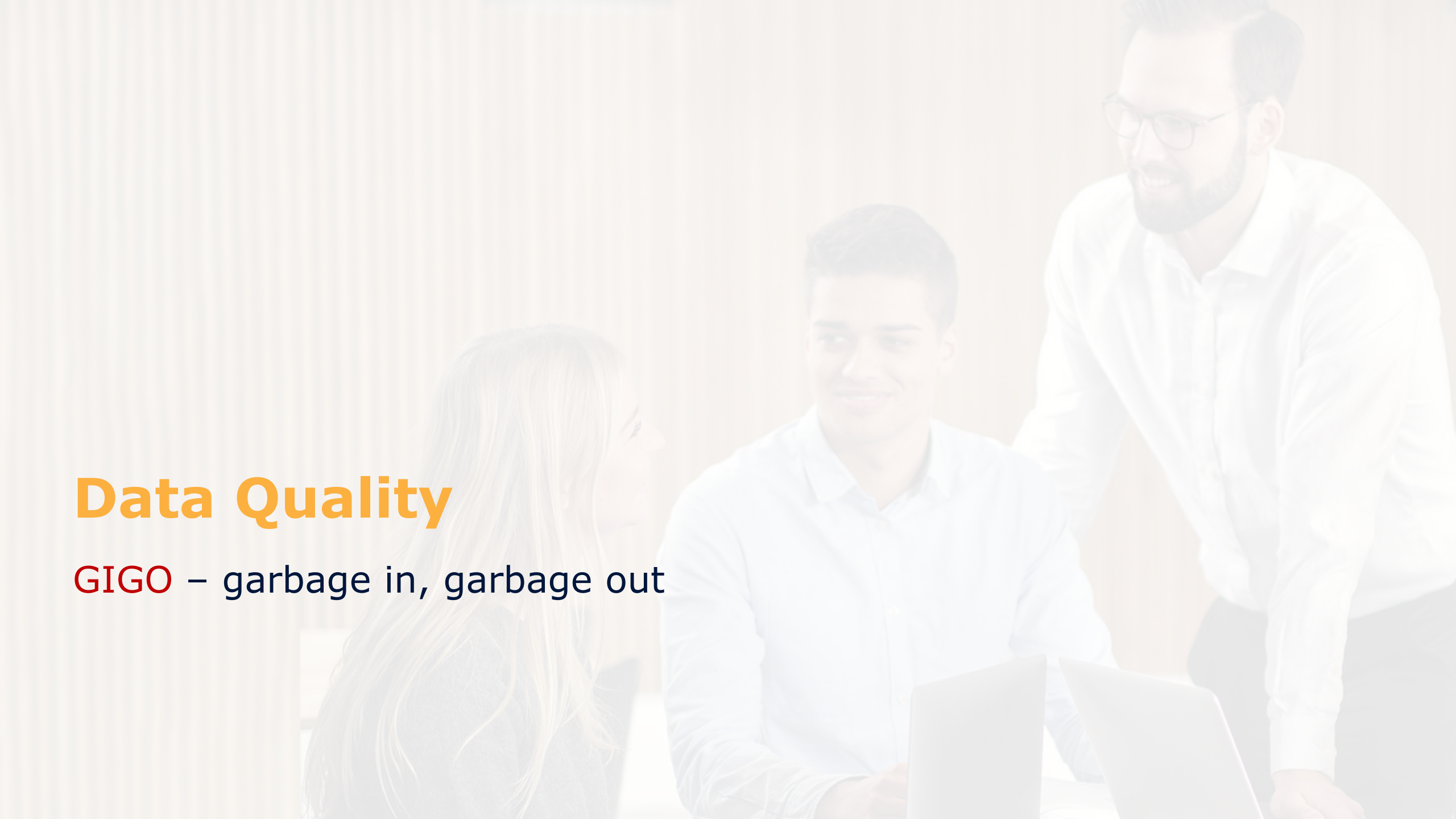


Data Lake



ETL vs ELT

Comparison Parameters	ETL	ELT
Ease of adoption to the tool	ETL is a well-developed process used for over 20 years, and ETL experts are easily available.	ELT is a new technology, so it can be difficult to find experts and develop an ELT pipeline
Data size	ETL is better suited for dealing with smaller data sets that require complex transformations.	ELT is better suited when dealing with massive amounts of structured and unstructured data.
Order of the process	Data transformations happens after extraction in the staging area. After transformation, the data is loaded into the destination system.	Data is extracted, loaded into the target system, and then transformed.
Transformation process	The staging area is located on the ETL solution's server.	The staging area is located on the source or target database.
Load time	ETL load times are longer than ELT because it's a multi-stage process: (1) data loads into the staging area, (2) transformations take place, (3) data loads into the data warehouse.	Data loading happens faster because there's no waiting for transformations and the data only loads one time into the target data system.



Data Quality

GIGO – garbage in, garbage out

How to Prepare Good Data?

1. Obtain meaningful data
   - correctly measured
   - usefully labelled
2. Acquire sufficient data
   - not possible to tell in advance
   - decided after testing
3. Shape it
4. Clean it
5. Wrangle it

Data Quality Factors

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Quantitative Features
  - the number of the observations
  - the number of observed attributes
- Qualitative Features
  - number of missing values
  - distorted distributions, outliers
  - redundancy/duplication of information
  - anomalous examples
- Can you think of more factors of good data?

What Is Good Data Quality?

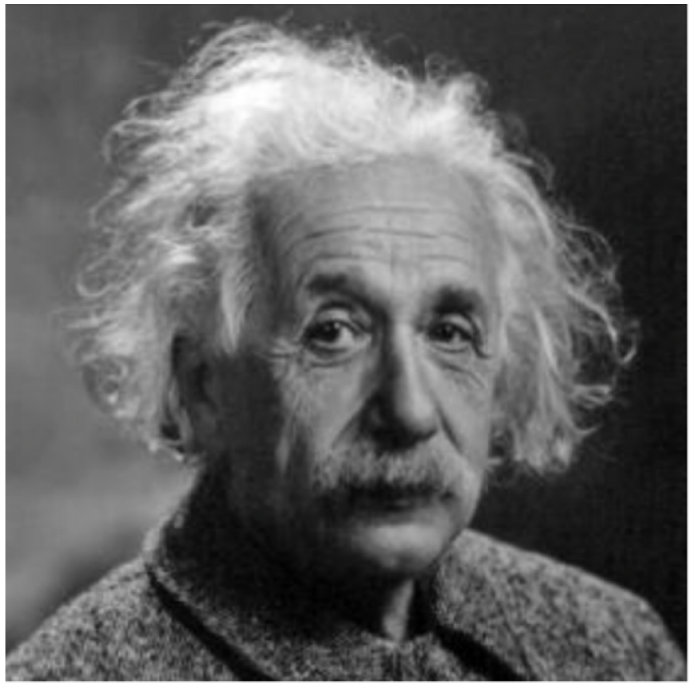
- Volume – larger amount better chance to find answers
- History – long-time collection contain patterns
- Variety – more quantitative and qualitative variables reveal better findings
- Consistency – fitting the old and new data
- Clarity – meaningful labels - better understanding
- Level of detail - lowest
- Segmentation – grouping by categories, as possible
- Transparency
  - clear origin
  - clear processing
- Clean data – repairing noise and missing data
- Good structure – optimal for data analysis operations

What Is Bad Data Quality?

- Quantitative criteria
  - too low number of observations
  - too high number of features
- Qualitative Criteria
  - missing values
  - distorted distributions, outliers
  - redundancy of information, duplication
  - anomalous examples
  - noise
  - invalid formats
  - wrong data structures
  - non-consistently written label

Keep Attention on Important Features

Keep the complexity under control



"Make it as simple, as possible, but not simpler "

(Albert Einstein)

Keep Attention on Important Features

Seek generalization

- don’t take a small population sample

Avoid *survivor bias*

- logical error of concentrating on the people or things that past some selection process, and overlooking those that did not, typically because of their lack of visibility
- during WW2 US Air Force enforced hit spots on airplanes, mathematician Abraham Wald pointed the opposite

Some Cleaning Techniques
Restoring Missing Data

In ndarray Python places special value, printed as NaN (Not a Number)

DataFrame can replace NaN or drop rows with Nan

Python Examples

```
# replaces NaN with zero
numpy.nan_to_num(dataset)

# drops rows with NaN
dataset = dataset.dropna()

# filters not null rows only
dataset[dataset.notnull()]

# replace NaN with the average of the rest
df['Sale'].fillna(int(df['Sale'].mean()), inplace=True)


# ffill() replaces NaN with the previous valid value,
and bfill() replaces it with the next valid value
```

- Restore missing data carefully
  - identify missing data
  - drop a variable, if 90% of instances are missing
  - otherwise, fill in approximate data
- Randomly missing data is replaced by
  - zero
  - mean or median
  - interpolation
  - place value out of the normal range, e.g. negative

Some Cleaning Techniques
Resolving Outliers

Due to

- sensor malfunctioning
- wrong data entry
- freak events

May damage the model

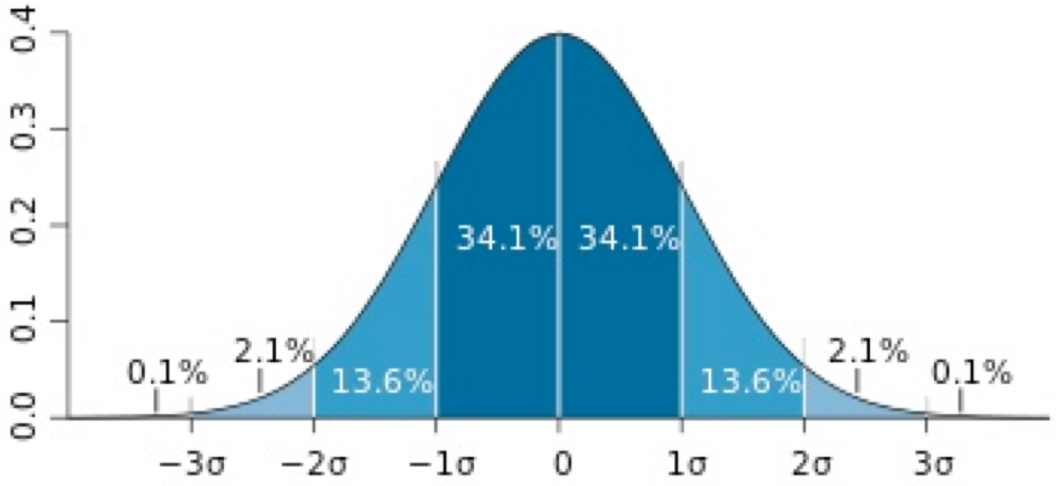
How to recognize them?

if mean is much different than median

When to filter them

x > mean - 2 * sd

x < mean + 2 * sd



cphbusiness

Some Wrangling Techniques
Data Framing

- Meaningful level of details
- Matching levels of details between sources
- Tall and narrow instead of short and wide

  - Each variable is in one column, with a column header
  - Each different observation of that variable is on a different row

Calculation Country Name	Calculation Country Code	Data 1961	Data 1962	Data 1963	Data 1964	Data 1965	Data 1966
Aruba	ABW	55,435	56,226	56,697	57,029	57,360	57,712
Andorra	AND	14,376	15,376	16,410	17,470	18,551	19,646
Afghanistan	AFG	8,953,544	9,141,783	9,339,507	9,547,131	9,765,015	9,990,125
Angola	AGO	5,056,688	5,150,076	5,245,015	5,339,893	5,433,841	5,526,653
Albania	ALB	1,659,800	1,711,319	1,762,621	1,814,135	1,864,791	1,914,573
United Arab Emirates	ARE	97,727	108,774	121,574	134,411	146,341	156,890

Calculation Country Name	Calculation Country Code	Pivot Year	Pivot Population
Aruba	ABW	1961	55,435
Andorra	AND	1961	14,376
Afghanistan	AFG	1961	8,953,544
Angola	AGO	1961	5,056,688
Albania	ALB	1961	1,659,800
United Arab Emirates	ARE	1961	97,727
Argentina	ARG	1961	20,959,241
Armenia	ARM	1961	1,934,239
American Samoa	ASM	1961	20,478
Antigua and Barbuda	ATG	1961	55,403
Australia	AUS	1961	10,483,000

Some Wrangling Techniques
Data Frames Restructuring

Pivoting – converting columns into rows

Employee	2/5/2020	2/6/2020	2/7/2020	2/8/2020	2/9/2020
Christine	10	10	10	10	10
Tristan	10				
Lily	10				10
Jamal	10		10		

Splitting - separating a column that contains multiple pieces of information into multiple columns, one for each piece of information

Airline
American Airlines: AA
Delta Airlines: DL
JetBlue Airways: B6
United Airlines: UA

Airline Name	Airline ID
American Airlines	AA
Delta Airlines	DL
JetBlue Airways	B6
United Airlines	UA

Employee	Date	Parking Fee
Christine	2/5/2020	10
Christine	2/6/2020	10
Christine	2/7/2020	10
Christine	2/8/2020	10
Christine	2/9/2020	10
Tristan	2/5/2020	10
Lily	2/5/2020	10
Lily	2/9/2020	10
Jamal	2/5/2020	10
Jamal	2/7/2020	10

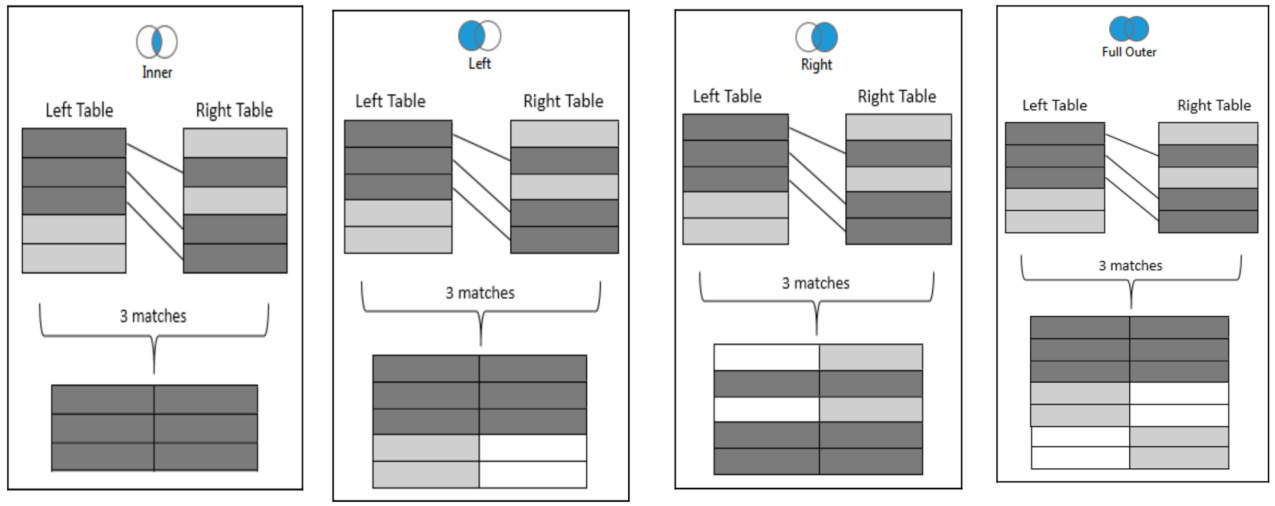
Some Wrangling Techniques

Data Aggregation

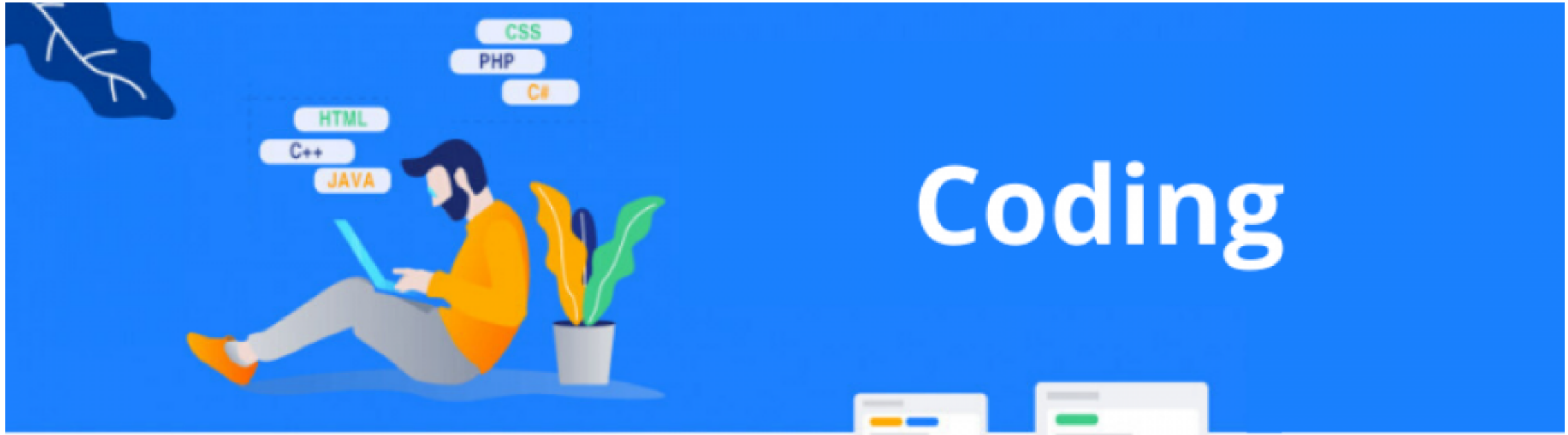
- Similar to SQL DB
- Two types
  - Joins - link sources on row-by-row scale

    aggregate data from one source with matching data from another source
  - Blends - links sources on aggregate level

    independent processing of data from two sources and consequent aggregation of the results



cphbusiness



Programming