

A complex network diagram with nodes of various sizes and colors (red, dark blue, grey) connected by lines. The diagram is overlaid on a red horizontal band.

Basic data collection and pre-processing

5ARE0: DATA ANALYSIS & LEARNING METHODS (2025 – 2026)

Uzay Kaymak, Jheronimus Academy of Data Science, u.kaymak@tue.nl

Master: Artificial Intelligence & Engineering Systems

Recap

- **Data has multiple facets**
- **Distinction between primary data collection and secondary data collection**
 - In engineering systems, primary data collection has focus
 - Data science deals also with secondary data collection
- **Capturing context information is very relevant**

Data collection: process of gathering data for use in (business) decision-making, strategic planning, research and other purposes

Outline

- **Categories of data**
- **Signal conditioning**
- **Data storage**
- **Experiment design**
 - System excitation
 - Sampling
- **Data conditioning**
- **CRISP-DM methodology**

Categories of data

- **Attribute – value pairs**
 - Matrix/table representation
- **Unstructured data**
 - text
- **Sequence data**
 - Time series
 - Event logs
- **Graph data**
 - Non-linear data represented as nodes and vertices

Attribute – value pairs (fields in a record)

Diabetes Patients Medication Status
Source: Centers for Disease Control and Prevention (CDC) <https://www.cdc.gov/>

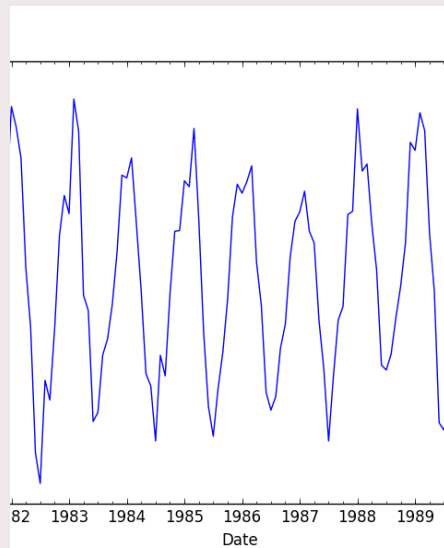
<i>Year</i>	<i>Pills Only</i>	<i>Insulin Only</i>	<i>Insulin and Pills</i>	<i>Any Medication</i>	<i>No Medication</i>
2000	6.3	2.3	1.4	10.0	1.8
2001	7.0	2.2	1.6	10.8	1.9
2002	7.6	2.1	1.7	11.4	2.0
2003	8.1	2.1	1.8	12.0	2.1
2004	8.6	2.2	2.0	12.9	2.2
2005	9.2	2.2	2.2	13.6	2.5
2006	9.8	2.2	2.2	14.2	2.6
2007	10.1	2.2	2.4	14.8	2.8
2008	10.8	2.3	2.7	15.7	3.0
2009	11.3	2.6	2.9	16.8	3.2
2010	11.8	2.8	2.9	17.5	3.2
2011	11.8	2.9	2.9	17.7	3.1

Unstructured data

Diabetes Patients Medication Status
Source: Centers for Disease Control and Prevention (CDC)
<https://www.cdc.gov/>




“ ... From **1997** to 2011, the number of adults aged 18 years or older with diagnosed diabetes who reported taking diabetes medication increased for those taking either insulin, pills, or both. The number of adults with diagnosed diabetes who did not report taking diabetes medication also increased during the period. For those taking insulin only, trends showed little or no change until **2007** and increased afterwards. ...”

Gene sequence



Last 30 Days

Is accessed logs | Show older logs

count	User	Organization	Action
account	User	Organization Name	Documents
in	Jenn Kaine	Happy Frog - Jackie Tilds	Document create
in	Jenn Kaine	Happy Frog - Jackie Tilds	Document create
in	Jenn Kaine	Happy Frog - Jackie Tilds	Document create
in	Jenn Kaine	Happy Frog - Jackie Tilds	Document create
in	Jenn Kaine	Happy Frog - Jackie Tilds	Document create
in	Jenn Kaine	Happy Frog - Jackie Tilds	Document create
in	Jenn Kaine	Happy Frog - Jackie Tilds	Document create
in	Jenn Kaine	Happy Frog - Jackie Tilds	Document create
in	 Jenn Kaine	jenn	Document create
in	 Jenn Kaine	jenn	Document create
in	 Jenn Kaine	jenn	Document create

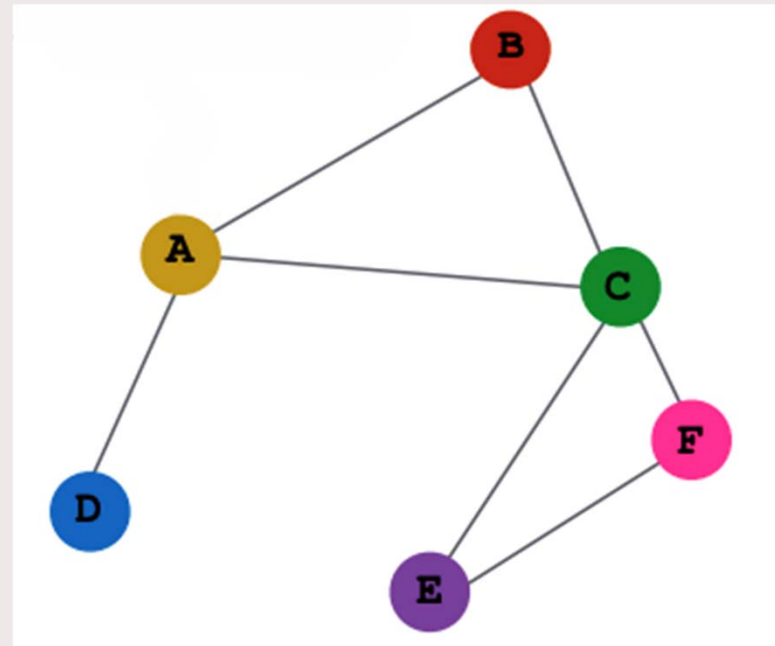


Graph data

Information represented as:

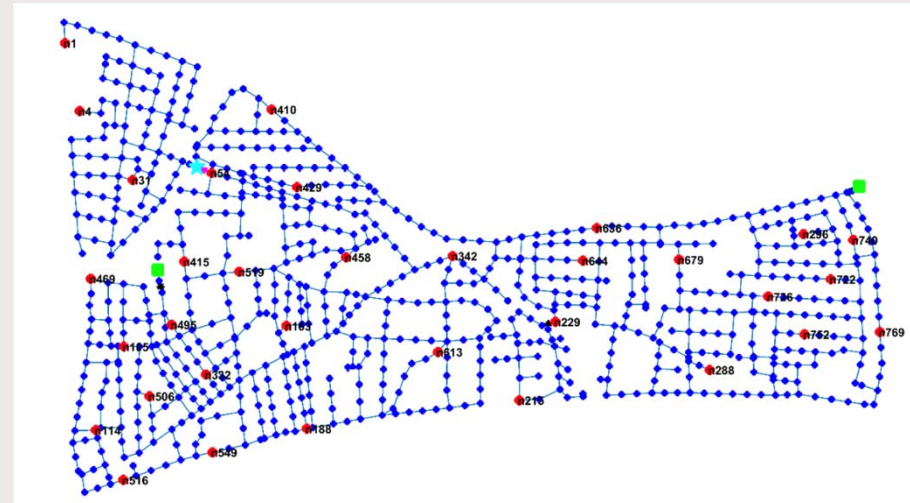
- nodes,
- vertices,
- attributes on nodes and/or vertices.

Captures structural elements better



Many problems have a natural graph representation: e.g. A Water Distribution Network

- A town with 10,000 population
- 2 pumps, 1 tank, 900 pipes and 782 junctions
- Each node is a service connection that supplies several households.



Vrachimis, Stelios G., et al. "BattLeDIM: Battle of the leakage detection and isolation methods." Proc., 2nd Int. CCWI/WDSA Joint Conf. 2020.

Data sources

- **Sensors**
- **Surveying and monitoring**
 - Audio
 - Video
- **Surveys and questionnaires**
- **Interviews**
- **Etc.**



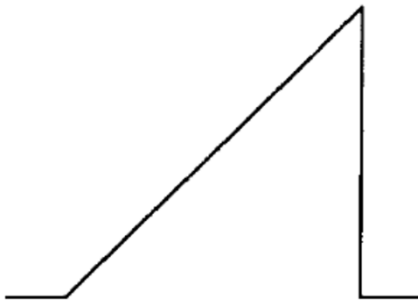
Examples of common data collection methods

- automated data collection from business applications, websites and mobile apps
- sensors that collect operational data from equipment, vehicles, etc.
- data from external data sources (*e.g.* data streams)
- tracking online channels (*e.g.* social media, discussion forums, twitter)
- surveys, questionnaires and forms (online, in person, by phone, etc.)
- focus groups and one-on-one interviews
- direct observation of participants in a research study

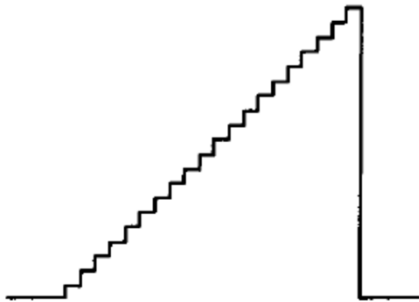
Observing the analogue world – sensors



www.techbriefs.com



(a) Analog Waveform

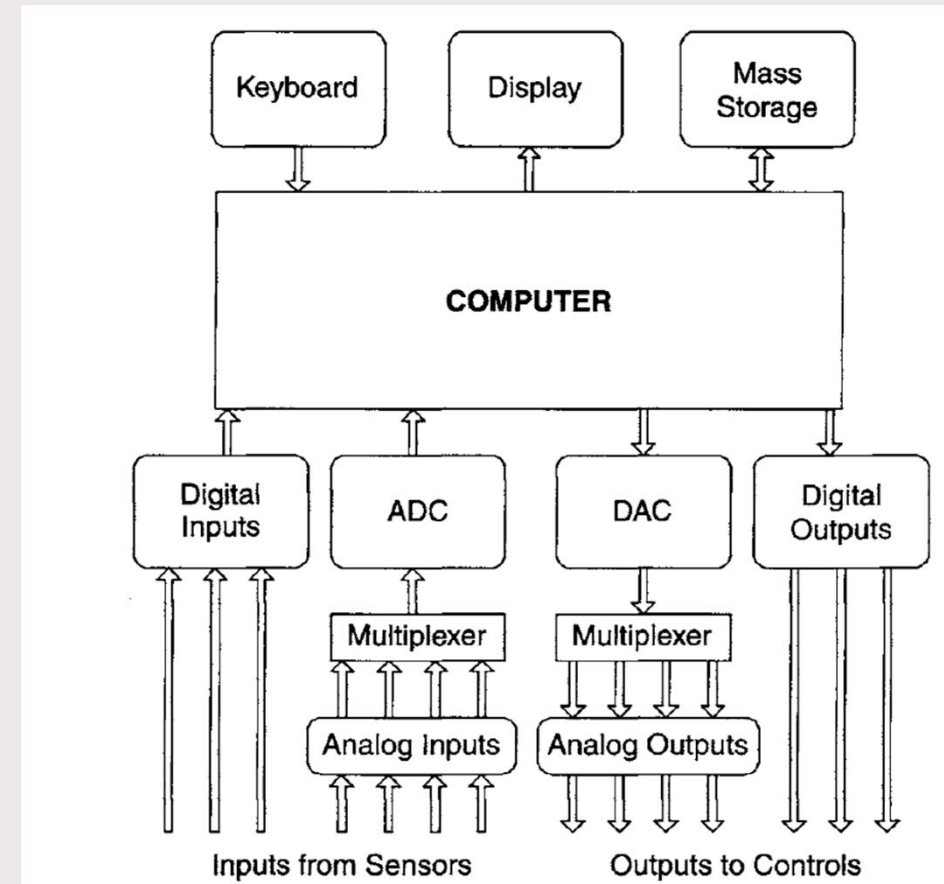


(b) Digitized Waveform

Data acquisition system (simplified)

ADC: analogue – digital converter

DAC: digital – analogue converter



Transducers

Convert a physical quantity into another

Properties

- **Sensitivity**
- **Stability**
- **Noise**
- **Dynamic range**
- **Linearity**

Examples

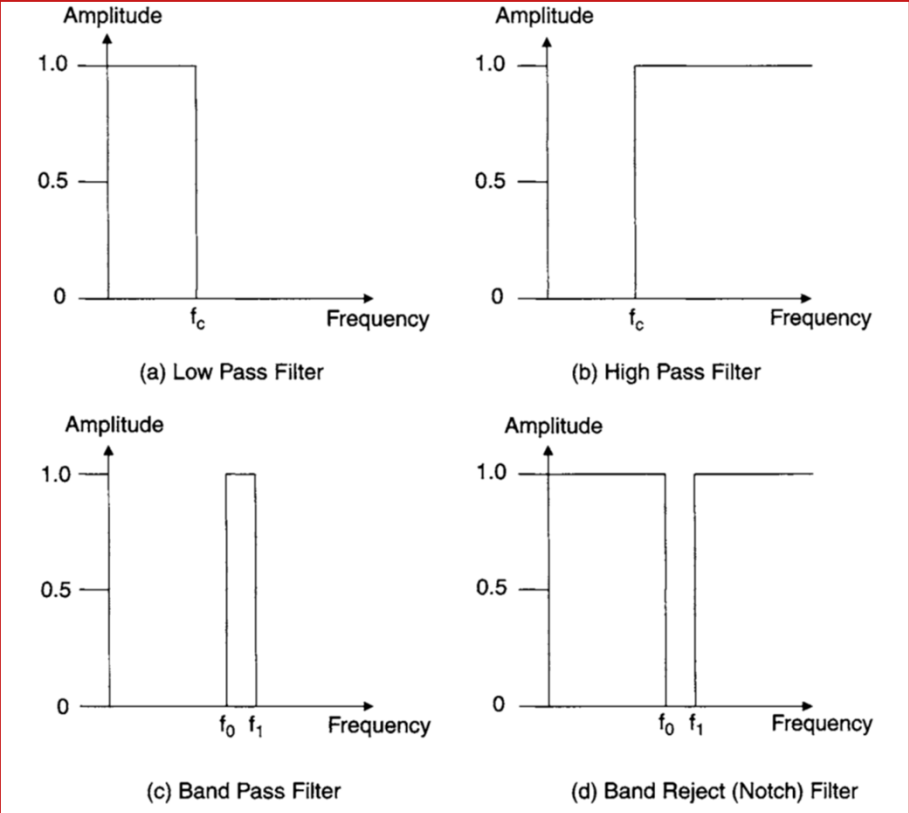
- **Temperature sensors**
- **Force and pressure transducers**
- **Magnetic field sensors**
- **Ionizing radiation sensors**
- **Displacement (position) sensors**
- **Fiber optic sensors**
- **MEMS (micro electromechanical systems)**

Signal conditioning

**Modifying the (analogue) signal
before being processed
(*e.g.* digitizing)**

- **Amplification**
- **Attenuation**
- **Input coupling**
- **Excitation**
- **Linearization**
- **Filtering**
 - Remove noise
 - Isolate relevant/interesting signal
 - Anti-aliasing
 - Smoothing
- **Surge protection**
- **(Electrical) isolation**
- **Etc.**

Common filter types

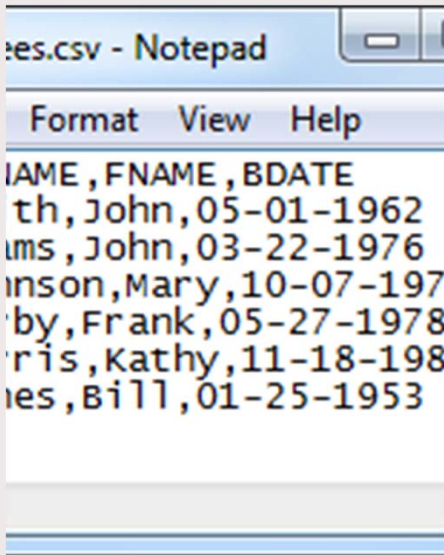


- **Direct writing to hardware (embedded systems, low-level OS access)**
- **Text files**
 - CSV (comma separated value) file (flat file)
 - Markup file (*e.g.* XML, JSON)
- **Spreadsheets**
- **Relational databases**
- **NoSQL databases (for graph data)**



Text files

CSV file



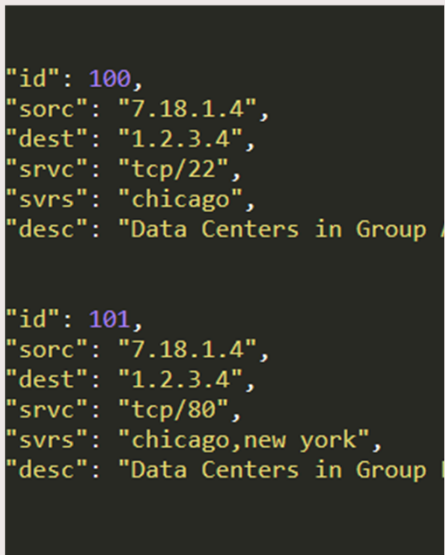
```
es.csv - Notepad
Format View Help
NAME,FNAME,BDATE
th,John,05-01-1962
ms,John,03-22-1976
nson,Mary,10-07-197
by,Frank,05-27-1978
ris,Kathy,11-18-198
es,Bill,01-25-1953
```

XML



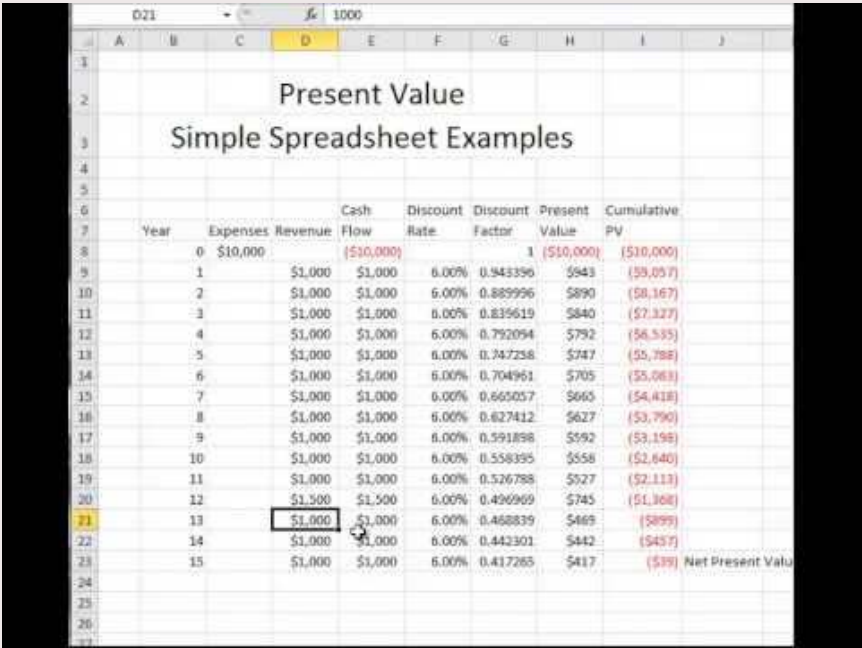
```
xml version="1.0" encoding="UTF-8"
employeeData>
  <employee id="34594">
    <firstName>Heather</firstName>
    <lastName>Banks</lastName>
    <hireDate>1/19/1998</hireDate>
    <deptCode>BB001</deptCode>
    <salary>72000</salary>
  </employee>
  <employee id="34593">
    <firstName>Tina</firstName>
    <lastName>Young</lastName>
    <hireDate>4/1/2010</hireDate>
    <deptCode>BB001</deptCode>
    <salary>65000</salary>
  </employee>
employeeData>
```

JSON



```
{
  "id": 100,
  "src": "7.18.1.4",
  "dest": "1.2.3.4",
  "src": "tcp/22",
  "svrs": "chicago",
  "desc": "Data Centers in Group 1"
},
{
  "id": 101,
  "src": "7.18.1.4",
  "dest": "1.2.3.4",
  "src": "tcp/80",
  "svrs": "chicago,new york",
  "desc": "Data Centers in Group 2"
}
```

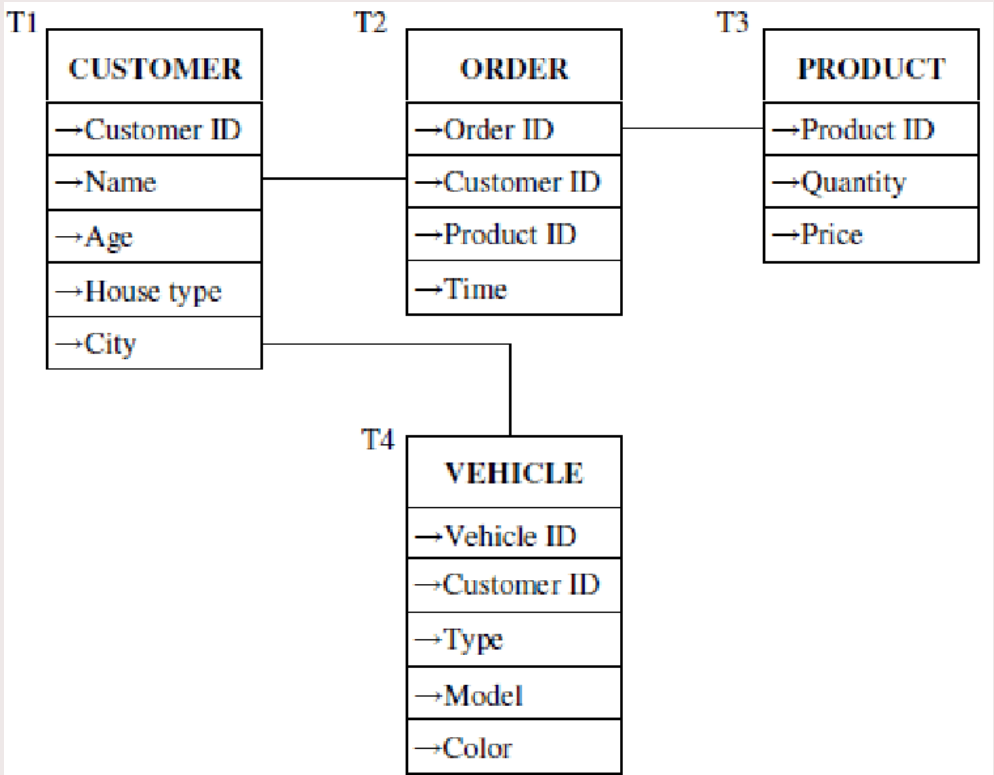
Spreadsheets



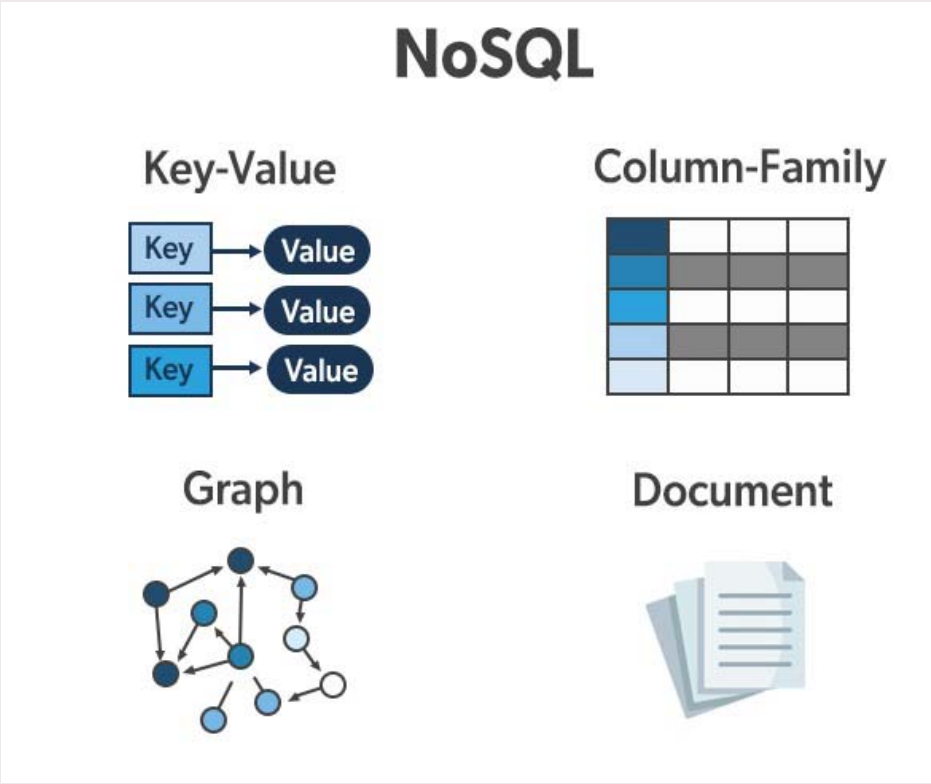
The screenshot shows a spreadsheet with the following data:

Year	Expenses	Revenue	Cash Flow	Discount Rate	Discount Factor	Present Value	Cumulative PV
0	\$10,000		(\$10,000)		1	(\$10,000)	(\$10,000)
1		\$1,000	\$1,000	6.00%	0.943396	\$943	(\$9,057)
2		\$1,000	\$1,000	6.00%	0.889996	\$890	(\$8,167)
3		\$1,000	\$1,000	6.00%	0.839619	\$840	(\$7,327)
4		\$1,000	\$1,000	6.00%	0.792094	\$792	(\$6,535)
5		\$1,000	\$1,000	6.00%	0.747258	\$747	(\$5,788)
6		\$1,000	\$1,000	6.00%	0.704961	\$705	(\$5,083)
7		\$1,000	\$1,000	6.00%	0.665057	\$665	(\$4,418)
8		\$1,000	\$1,000	6.00%	0.627412	\$627	(\$3,790)
9		\$1,000	\$1,000	6.00%	0.591898	\$592	(\$3,198)
10		\$1,000	\$1,000	6.00%	0.558395	\$558	(\$2,640)
11		\$1,000	\$1,000	6.00%	0.526788	\$527	(\$2,113)
12		\$1,500	\$1,500	6.00%	0.496969	\$745	(\$1,368)
13		\$1,000	\$1,000	6.00%	0.468839	\$469	(\$899)
14		\$1,000	\$1,000	6.00%	0.442301	\$442	(\$457)
15		\$1,000	\$1,000	6.00%	0.417285	\$417	(\$39)

Relational database



NOSQL database

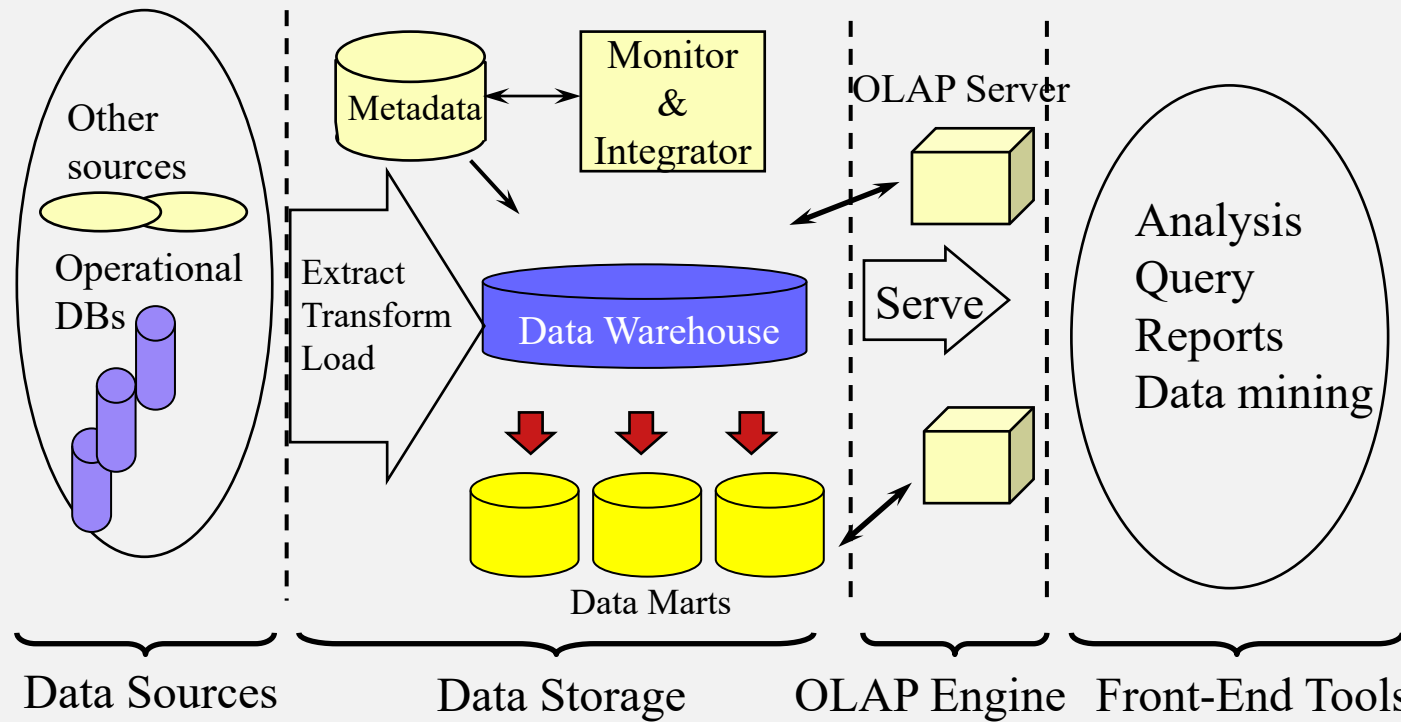


Raw data

- **Store as much as possible raw data (actual “measurements”)**
- **Raw data is often atomic or transactional**
- **For analysis, derive the required features from raw data (feature extraction)**
- **Consider: system to analyze characteristics of web site visitors by studying the time they spend on various pages and the order they visit the pages**
 - Time spent on URL 1
 - Next URL to visit

Question: Which data to store?

Data warehouses and data marts



Source: Jiawei Han



Experiment design (which data and how to collect?)

Cross-sectional data (static)

- Population sampling
- Active learning

Whenever possible, primary data collection, where there is control over collected data should be preferred.

Longitudinal data (dynamic)

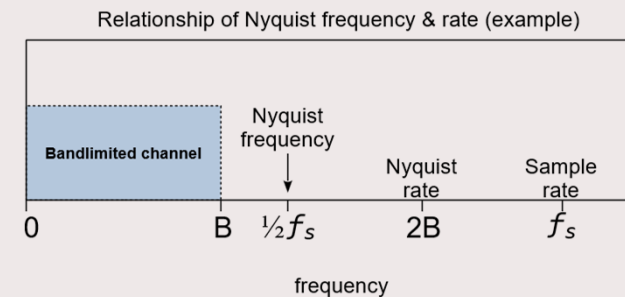
- System excitation
- Sampling frequency

System excitation

- **System should see sufficient relevant examples**
- **Rich in types/regions of input**
- **Linear systems (open loop):**
 - Pseudo-random binary sequence (all frequencies present in signal)
- **More complex for nonlinear systems**

Sampling frequency

- **If the system is bandwidth limited, use the Nyquist rate to reconstruct the signal**
Nyquist rate = 2 * bandwidth



Data conditioning

Modifying the (digitized) data for analysis, easy storage and/or further processing

- **Decomposition (*e.g.* frequency analysis)**
- **Aggregation**
- **Smoothing**
- **Interpolation**
- **Normalization**
- **Synchronizing**

Decomposition

Represent the signal as a superposition of constituent parts

- **Fourier transform**
- **Z-transform**
- **Wavelets**

Aggregation

Replace multiple values by a single value

- **Mean**
- **Median**
- **Minimum**
- **Maximum**
- **Etc.**

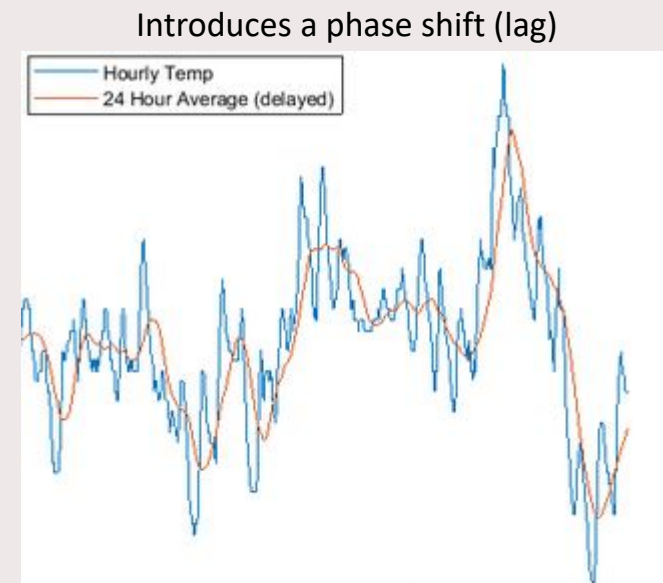
Example: daily stock prices



Smoothing

Moving average filters

- p-period simple moving average (mean of last p samples)
- p-period weighted moving average (weighted mean of last p-samples)
- A common smoothing filter is the exponentially weighted moving average $y'(t) = a y(t) + (1 - a) y'(t-1)$; $0 < a < 1$



Interpolation

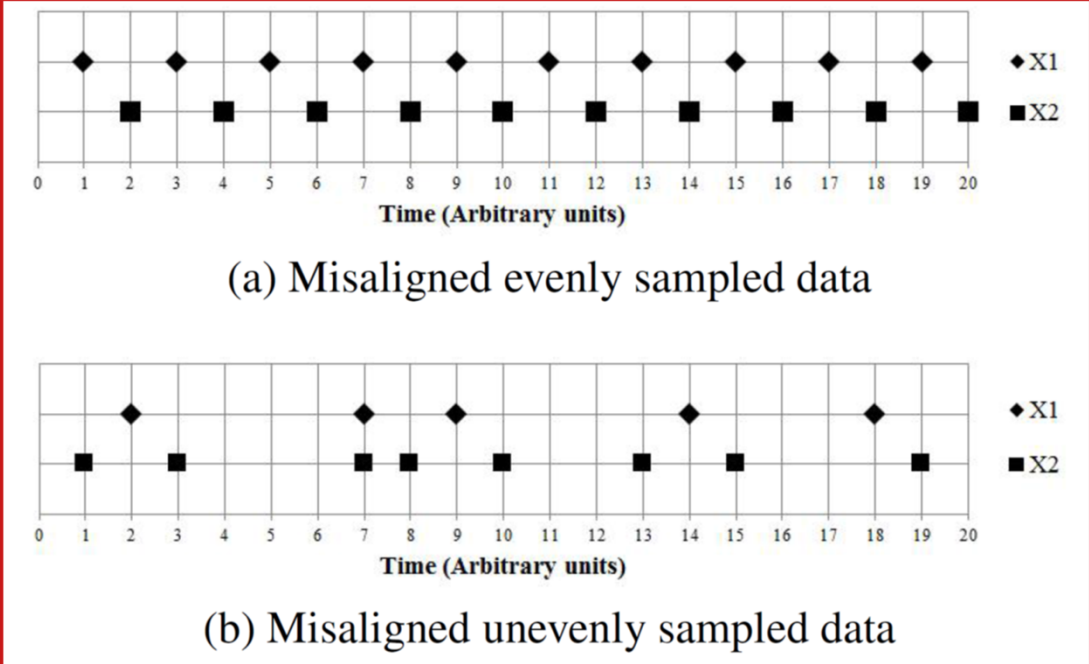
Linear table lookup

Polynomial interpolation

Splines

Etc...

Data synchronization



Missing values

MCAR: Missing Completely at Random

No systematic differences between missing and non-missing data

MAR: Missing at Random

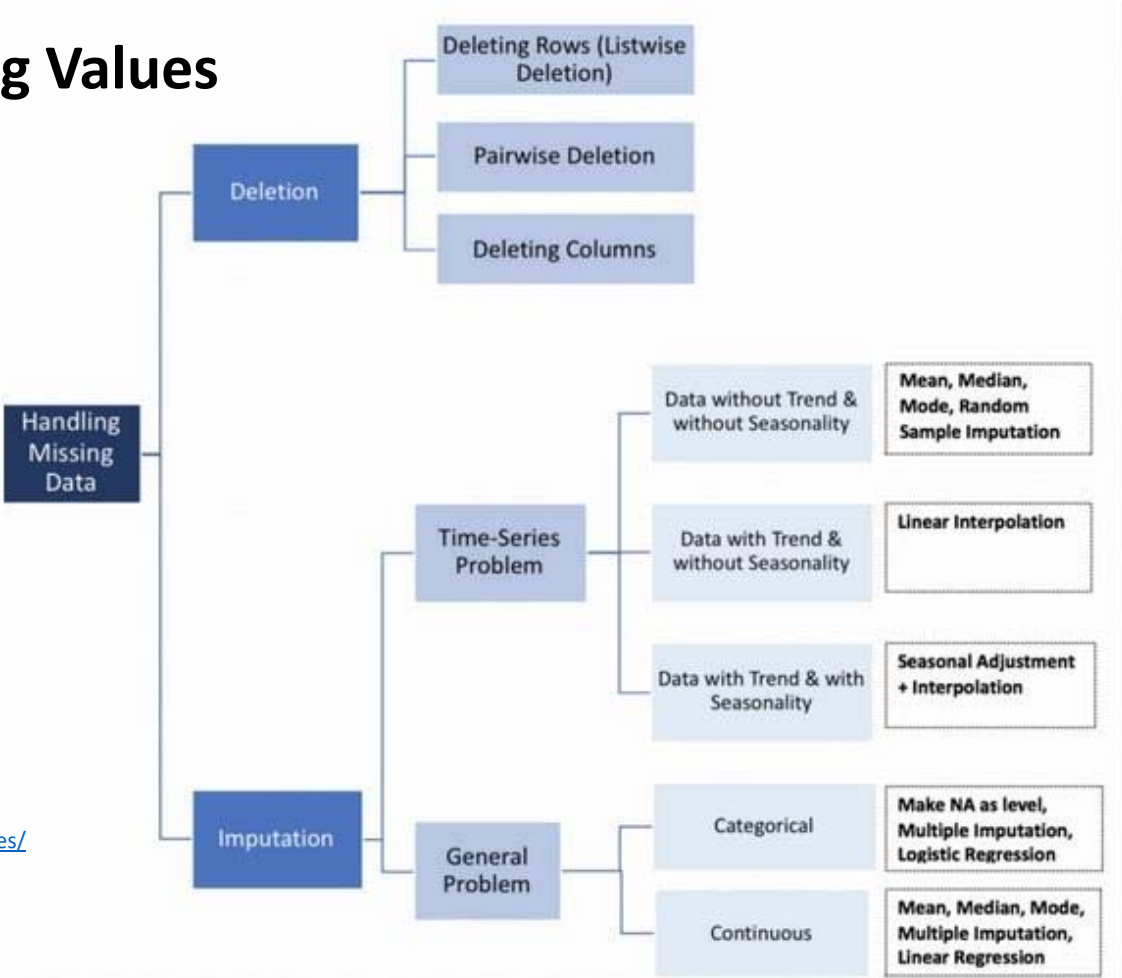
Differences between missing and non-missing data, which can be explained entirely by non-missing features

MNAR: Missing Not at Random

Missing with bias (e.g. value of variable explains missing data)

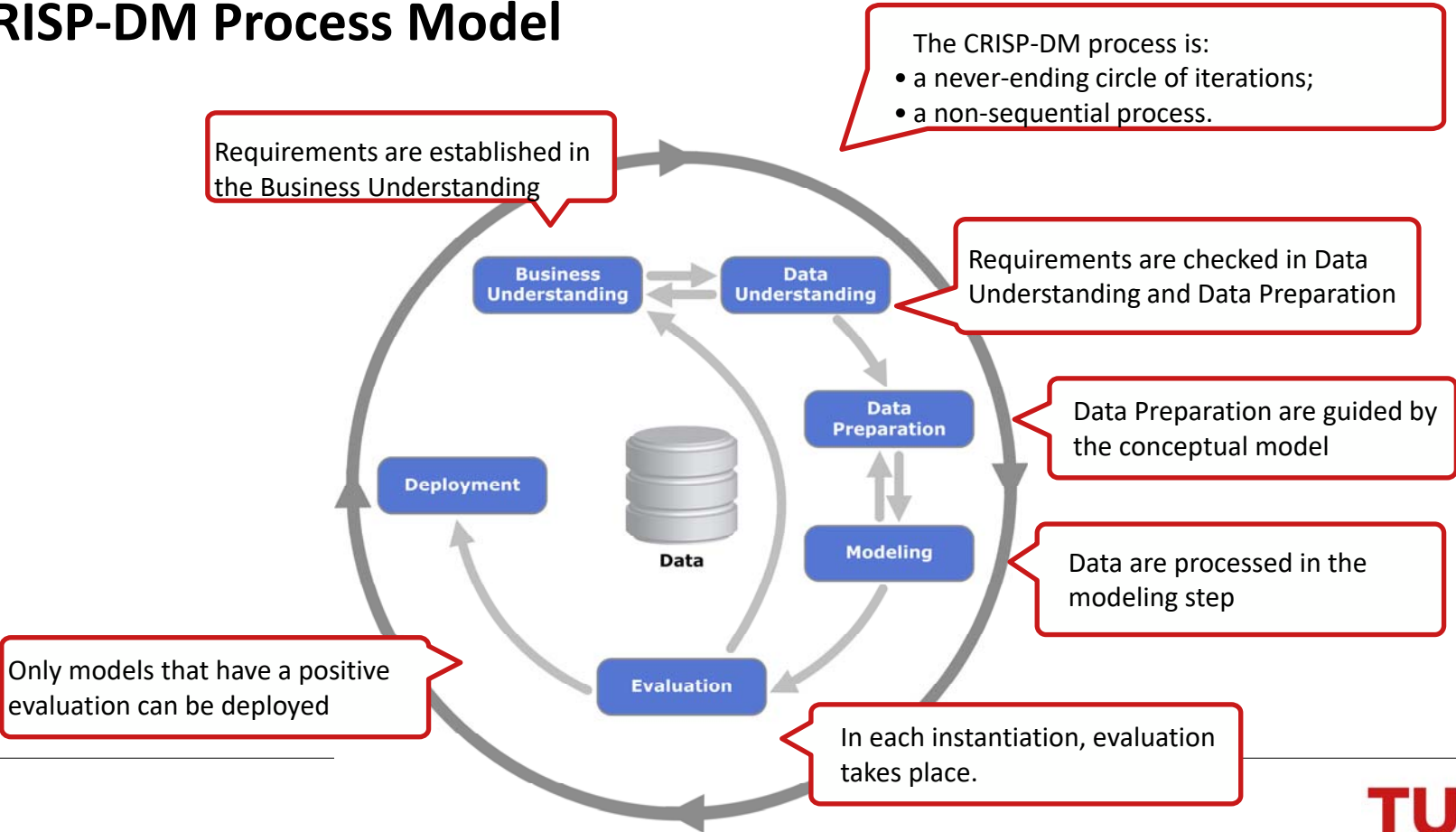


Handling Missing Values



<https://hrngok.github.io/posts/missing%20values/>

CRISP-DM Process Model





Tremor assessment

- **Tremor: rhythmic and involuntary movements of any body part, resulting from a neurological disorder**
- **Most common neurodegenerative disease with tremor:**
 - Parkinson's disease — *1 million in US, 75% tremor*
 - Essential tremor — *10 million in US*

Tremor symptoms

	Parkinson’s disease	Essential Tremor
Tremor types	Resting	Postural, kinetic
Frequency	4-6 Hz	7-12 Hz
Presence in hands	More than 70%	95%

- Essential tremor is not life threatening, but disabling
- Treatment: Medication, or in severe cases Deep Brain Stimulation

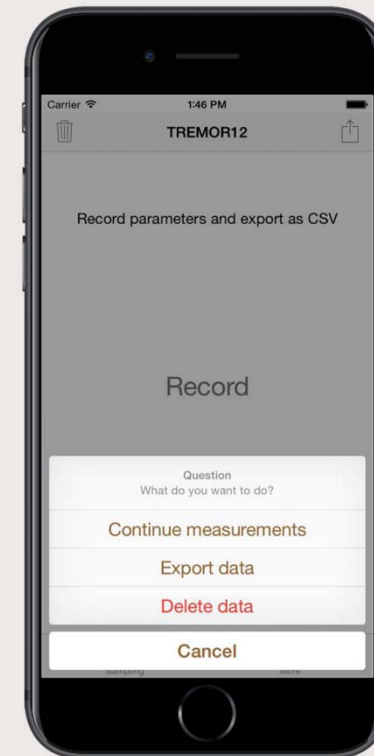
Tremor assessment

- **Rating scales – *most used, biased***
 - QQuality of life in ESsential Tremor (QUEST)
 - Filled in by patient
 - *e.g.* “I have lost interest in my hobbies because of tremor”
 - Essential Tremor Rating Scale (ETRS)
 - Filled in by clinician
 - *e.g.* “Rest tremor severity of the head”
- **Computerized analyses – *objective***



TREMOR12

- **Open-source smartphone application**
- **Created at Maastricht University Medical Center, The Netherlands**
- **Sensors:**
 - Accelerometer: Acceleration (in g)
 - Gyroscope: Rotation speed (in radians/s)
- **Sampling rate: 100Hz**



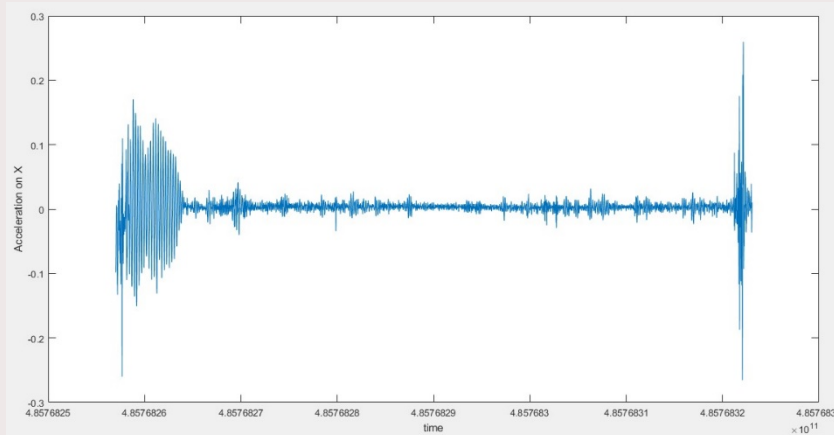
Data Collection

- **Five tests, both right and left wrist**
 - Rest: 1 test, 1 minute
 - Postural: 2 test – 1 minute
 - Kinetic: Glass & finger-nose test, each repeated 3 times
- 20 ET patients, ETRS and QUEST score known

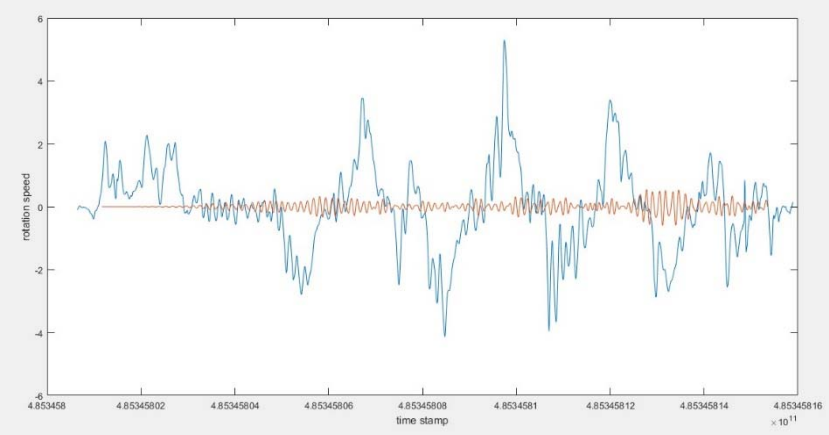


Noise filtering

- Noise from tapping start and stop button
- First and last 50 samples omitted



- Voluntary movement
- Equiripple Finite Impulse Response (FIR) filter 7 to 12 Hz



Feature extraction and selection (more on this later)

- **Extracted features for each test:**
 - Signal strength
 - Root-mean-square
 - Signal period
 - Dominant magnitude
 - Welch method
 - Daubechies 8 wavelet method
 - Dominant frequency
 - Welch method
 - Daubechies 8 wavelet method
- Sequential forward feature selection

Recap

- Four main categories (types) of data
- Data pre-processing prepares data for further analysis
- A lot of data is collected from sensors (signals from transducers)
- Signal conditioning for sensor data
- Data conditioning is an important step in pre-processing
- Handling missing data
- CRISP-DM

