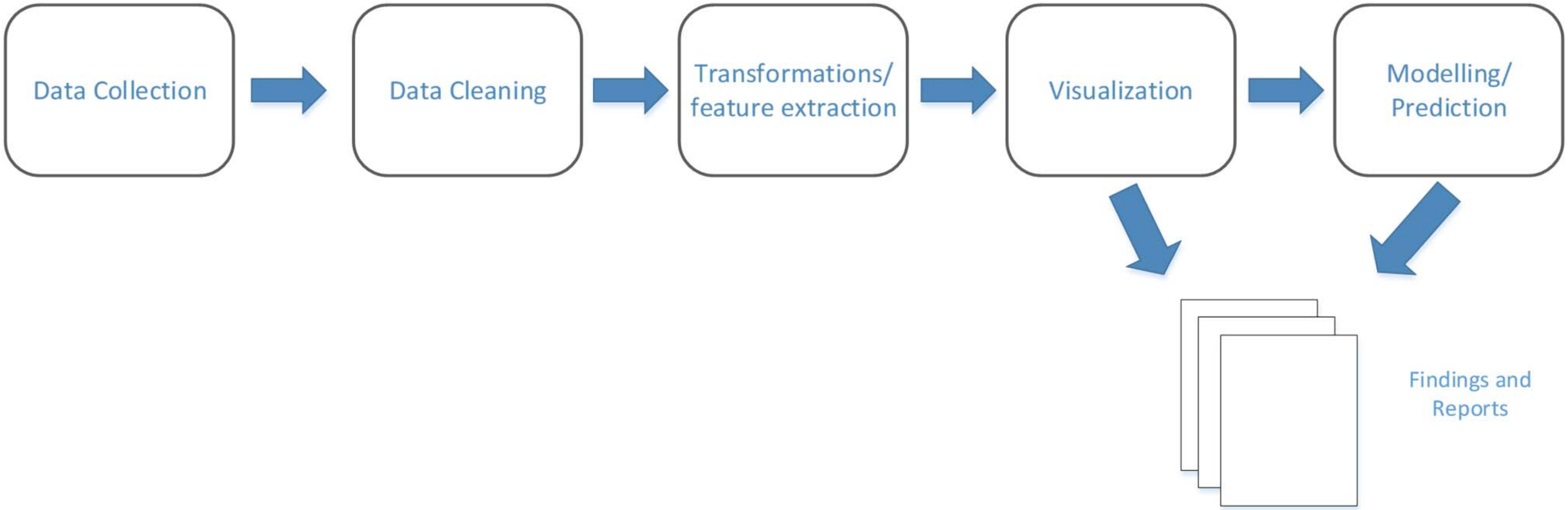
ENGINEERING ANALYTICS & MACHINE LEARNING

Data Cleaning

Data Analytic Process



What is raw data? (Recap)

- No software had been run on the data
- No manipulation ad been done any numbers in the data
- No data or number had been remove
- The data are not summarize in any way

❑ The strange binary file generated by the measurement machine

❑ The JSON data you got from scrapping the Twitter API or Facebook API

❑ The hand-entered numbers collected from paper survey forms

❑ Random like numbers generated by sensor network

Data Cleaning

Happy families are all alike; every unhappy family is unhappy in its own way ---- Leo Tolstoy

- Huge amount of effort is spent cleaning data to get it ready for analysis
- It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data.
- Data preparation is not just the first step but must be repeated over the course of analysis as new problems come to light or new data is collected
- Data cleaning is sometime painfully manual

Data Cleaning

This is an important stage and is necessary because the data that is collected is often “dirty” due to:

1. Missing data
2. Unacceptable formats
3. Erroneous values
4. Data could be embedded inside text or other information and needs to be extracted
5. Collected data in an unacceptable format
6. Remove or compensate for outliers that skews the data unrealistically

Missing Data In Pandas

1. None
   - Python objects
   - If we perform any aggregation function such as sum() or min() across an arrary with a None value will lead to an error
2. NaN : Not a Number
   - It is a special floating-point value recognized by all systems t hat use the standard IEEE floating-point representations
   - Regardless of the operation, the result of arithmetic with NaN will be another NaN

Data Transformation / Extraction

Data collected and read into analysis tools may need to be preprocessed because

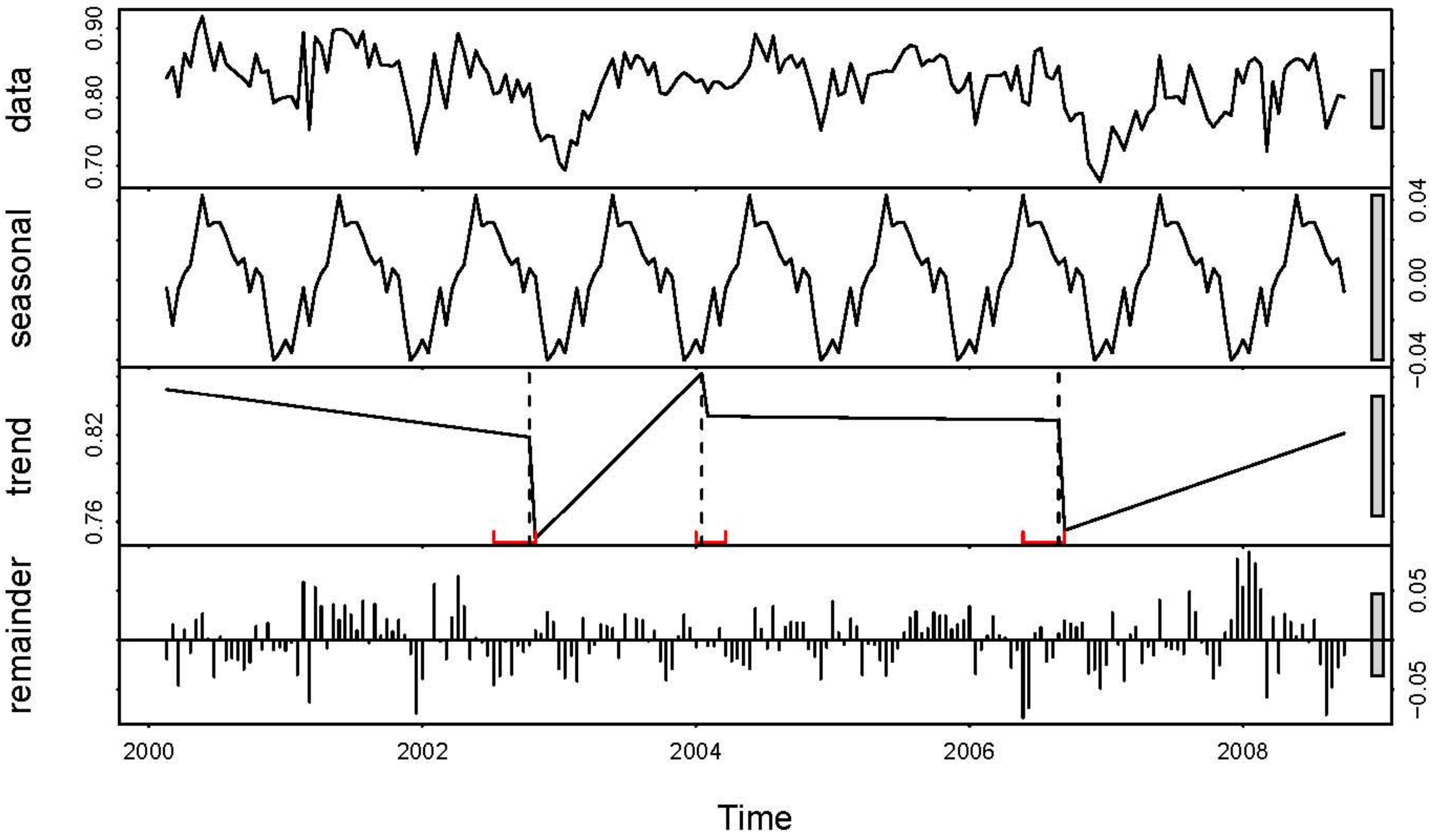
- The data types may be unsuitable for processing. For example, a floating point number may be wrongly formatted as strings. If mathematical computations are required. These “floating point strings” need to be reformatted as floating point numbers.

- The additional variables need to be calculated from the existing data to facilitate meaningful analysis. For example, the data sample below contains measurements of flower sepal and petal dimensions :

Date Time

- Engineering data such as those from sensors, measurement equipment or machines usually come with date time
- We usually refer to these type of data as time series data
- The data change according to time instead of some event
- The relationship between the data point and time such as cyclic effort or time-pattern event are pretty important
- The identification of such events would assist in analysis and also building of model for prediction purposes.

Data Transformation



AY2018/2019 APR SEMESTER

School of Engineering
TEMASEK POLYTECHNIC

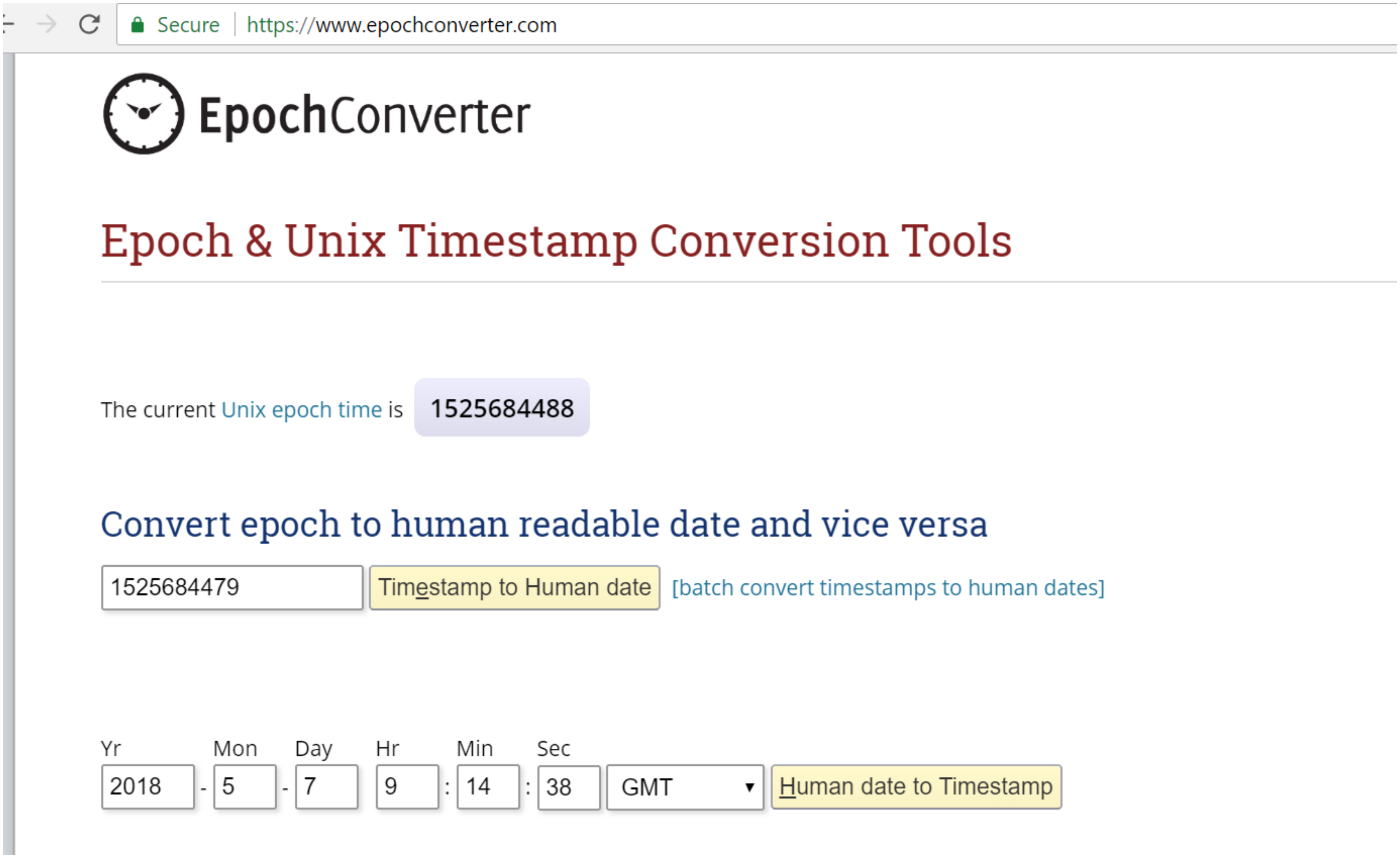
Unix Epoch

- The Unix epoch (or Unix time or POSIX time or Unix timestamp) is the number of seconds that have elapsed since January 1, 1970 (midnight UTC/GMT), not counting leap seconds (in ISO 8601: 1970-01-01T00:00:00Z).
- Literally speaking the epoch is Unix time 0 (midnight 1/1/1970), but 'epoch' is often used as a synonym for 'Unix time'.
- Most epoch time stamp is in seconds, milliseconds or microseconds.

Unix Epoch

Human readable time	Seconds
1 hour	3600 seconds
1 day	86400 seconds
1 week	604800 seconds
1 month (30.44 days)	2629743 seconds
1 year (365.24 days)	31556926 seconds

Online Epoch Converter



Secure | https://www.epochconverter.com

EpochConverter

Epoch & Unix Timestamp Conversion Tools

The current Unix epoch time is 1525684488

Convert epoch to human readable date and vice versa

1525684479 | Timestamp to Human date | [batch convert timestamps to human dates]

Yr		Mon		Day	Hr		Min		Sec		
2018	-	5	-	7	9	:	14	:	38	GMT	Human date to Timestamp

https://www.epochconverter.com/