EMBEDDED VISION DESIGN 3

DATA

JEROEN VEEN

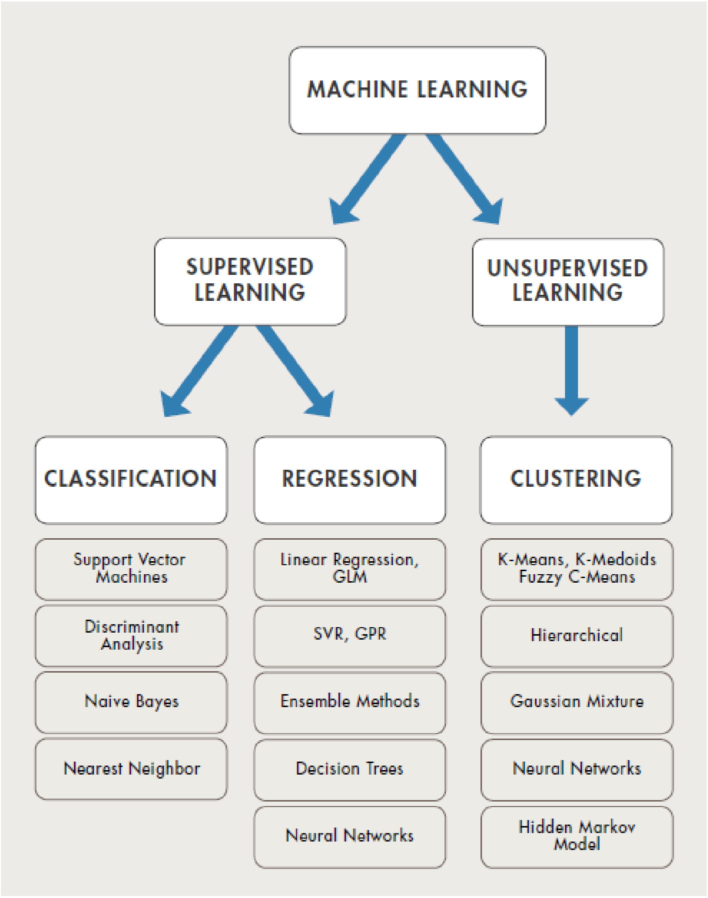
HAN_UNIVERSITY OF APPLIED SCIENCES

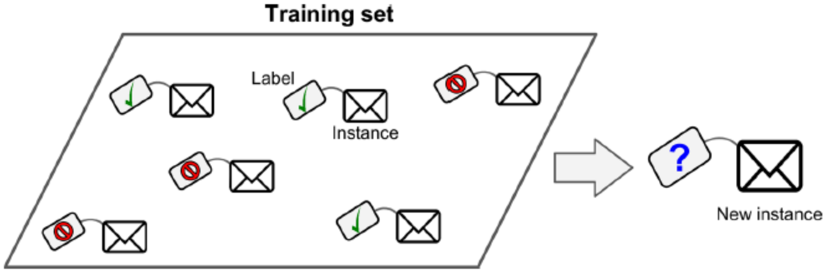
CONTENTS

- Thinking about data
- Splitting your data
- Feature engineering
- Exploring your data
- Data preparation

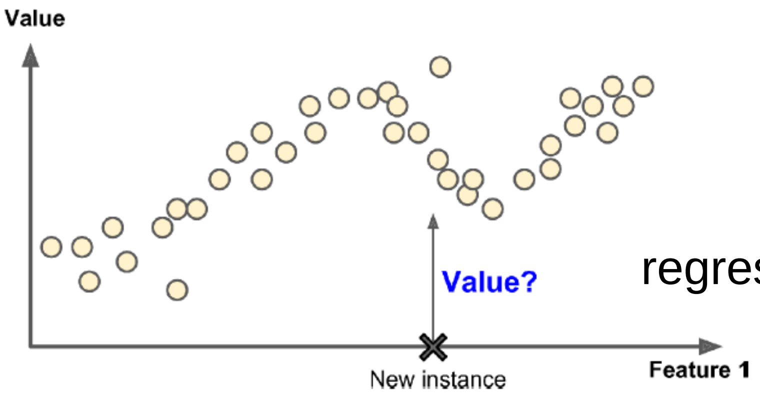
"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts." — Sir Arthur Conan Doyle, *Sherlock Holmes*

RECAP: MACHINE LEARNING APPROACHES





classification

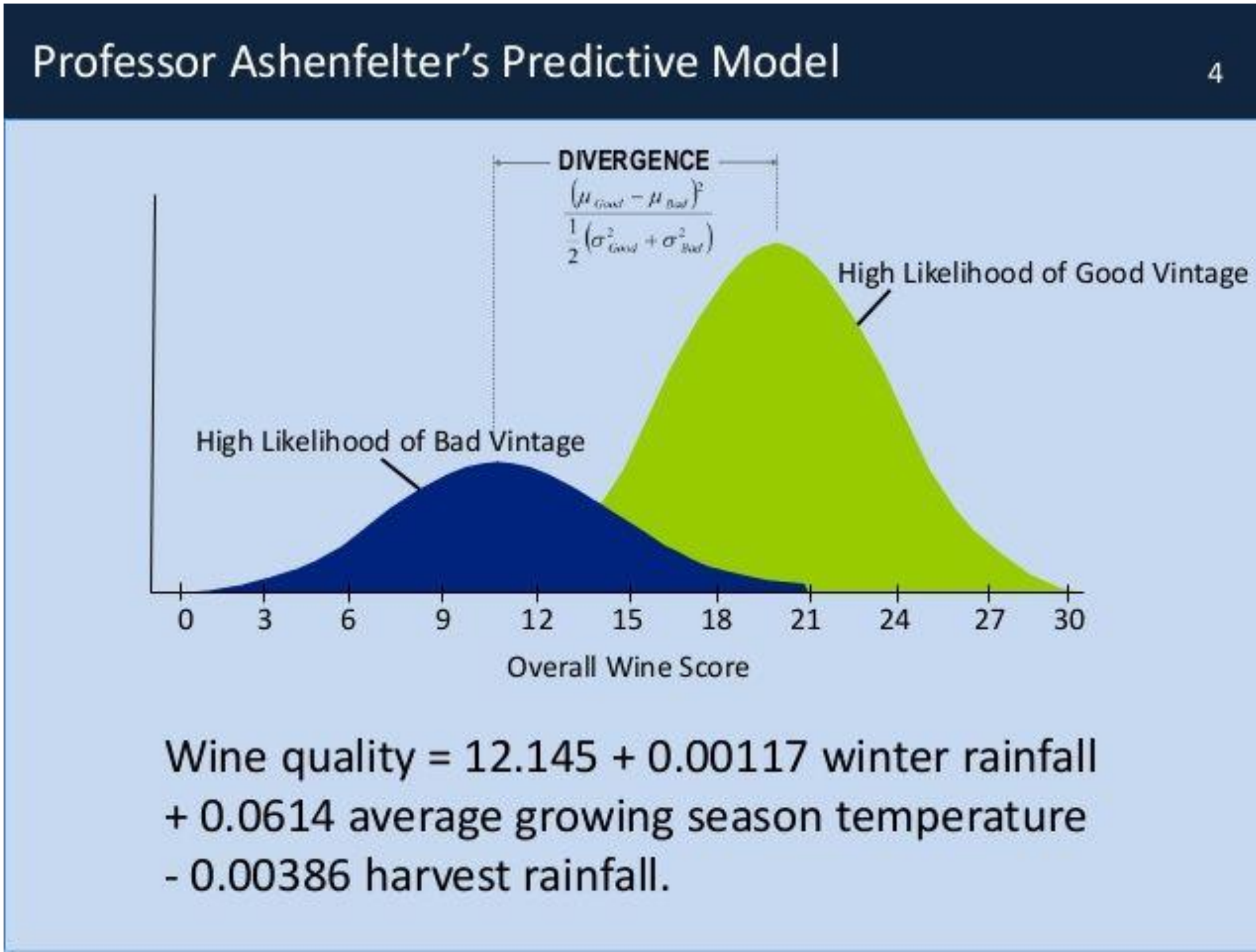


regression

Source: Géron, ISBN: 9781492032632

HAN_UNIVERSITY OF APPLIED SCIENCES

REGRESSION EXAMPLE



WORKING WITH REAL DATA

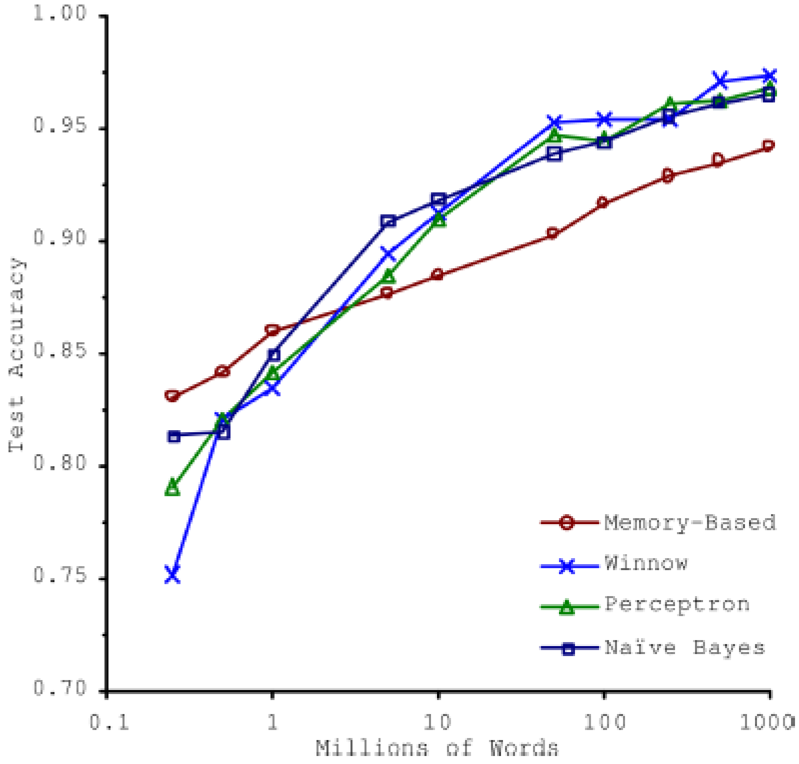
- Numerical information
- Values of quantitative variables
- Collected through measurement
- Usable for processing

Source: https://en.wikipedia.org/wiki/Data#/media/File:Data_types_-_en.svg

Ultimate goal of data processing: Turn information into insight!

UNREASONABLE EFFECTIVENESS OF DATA

- Data matters more than algorithms!



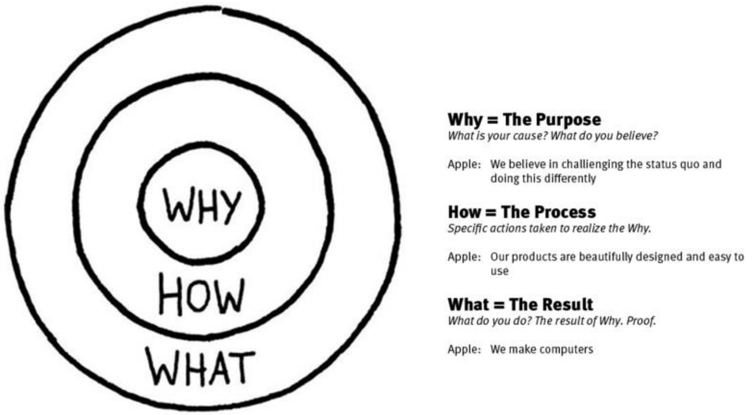
Source: Peter Norvig et al 2009

DEFINE YOUR OBJECTIVE

- What do you want to achieve?
  - > Define a SMART objective

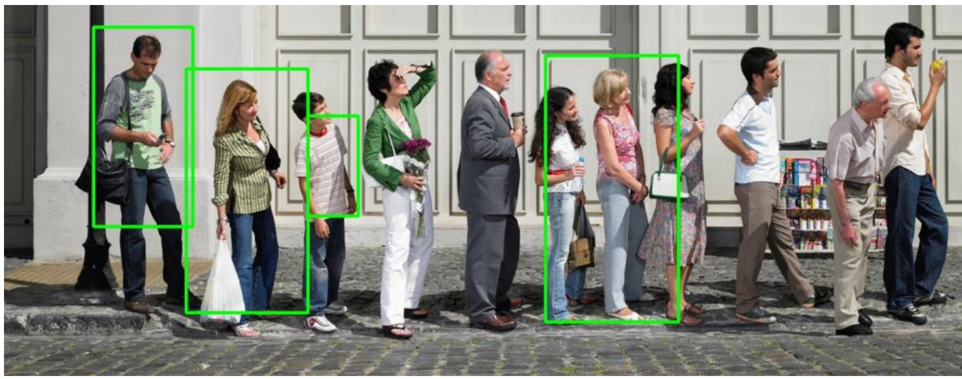
- What classes apply?
- What data is available?
- What attributes are present?
- What data should be collected?

- What features matter?



Source: Simon Sinek

EXAMPLE: PEOPLE DETECTION



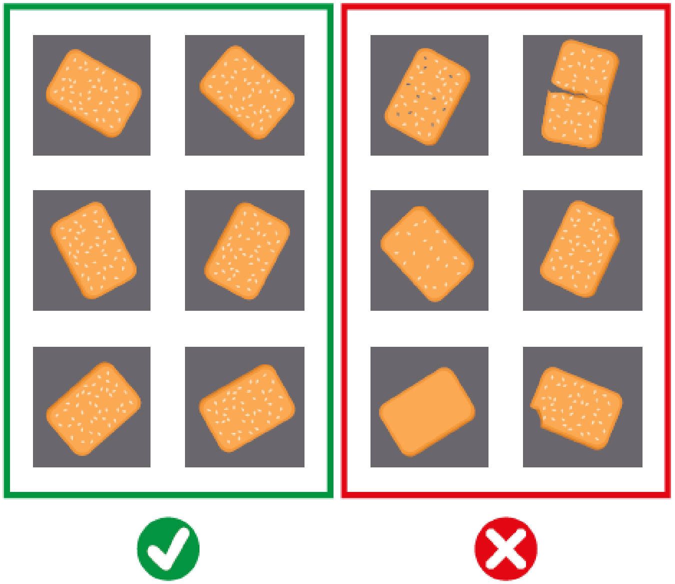
Source: https://thenextweb.com/wp-content/blogs.dir/1/files/2013/02/queue.jpg

Source:
https://www.researchgate.net/profile/Hayley_Hung2/publication/24979

GETTING LABELED DATASETS

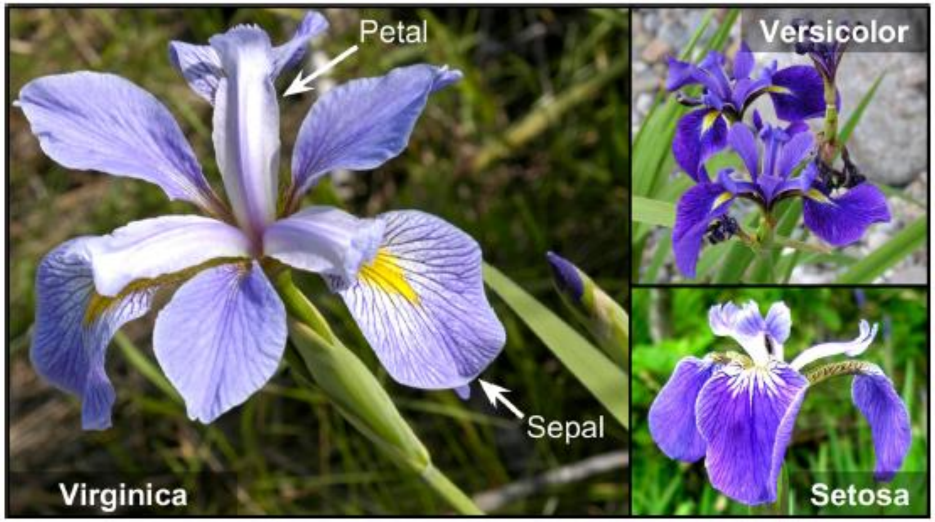
- Data acquisition
  - Field campaign
  - Controlled test set-ups
  - Scraping

- Data labelling
  - Domain experts
  - Hire data services
  - Control the experiments



Source: Basler, Artificial Intelligence in Image Processing

EXAMPLE: PUBLIC DATASETS



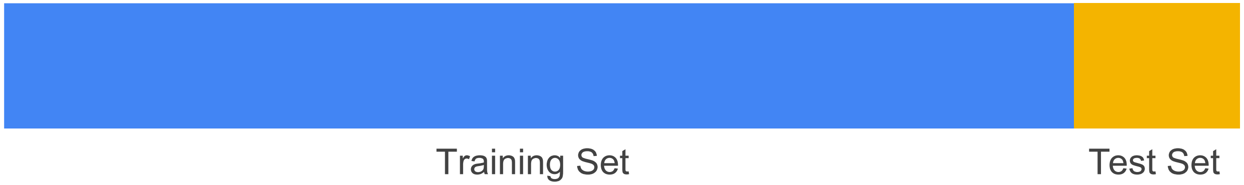
Source: Iris flower dataset

Source: MNIST database

See e.g. Scikit learn, Kaggle, Quandl, Google, Amazon

TRAINING AND TEST SETS: SPLITTING DATA

- **training set**—a subset to train a model.
- **test set**—a subset to test the trained model.
- You could imagine slicing the single data set as follows:

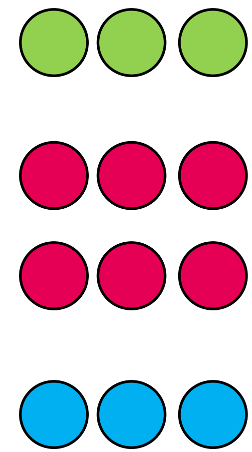


- Make sure that your test set meets the following two conditions:
  - Is large enough to yield statistically meaningful results.
  - Is representative of the data set as a whole. In other words, don't pick a test set with different characteristics than the training set.

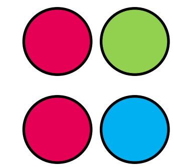
https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data

STRATIFIED SAMPLING

- Make sure the subsets set properly reflect the population



population



sample

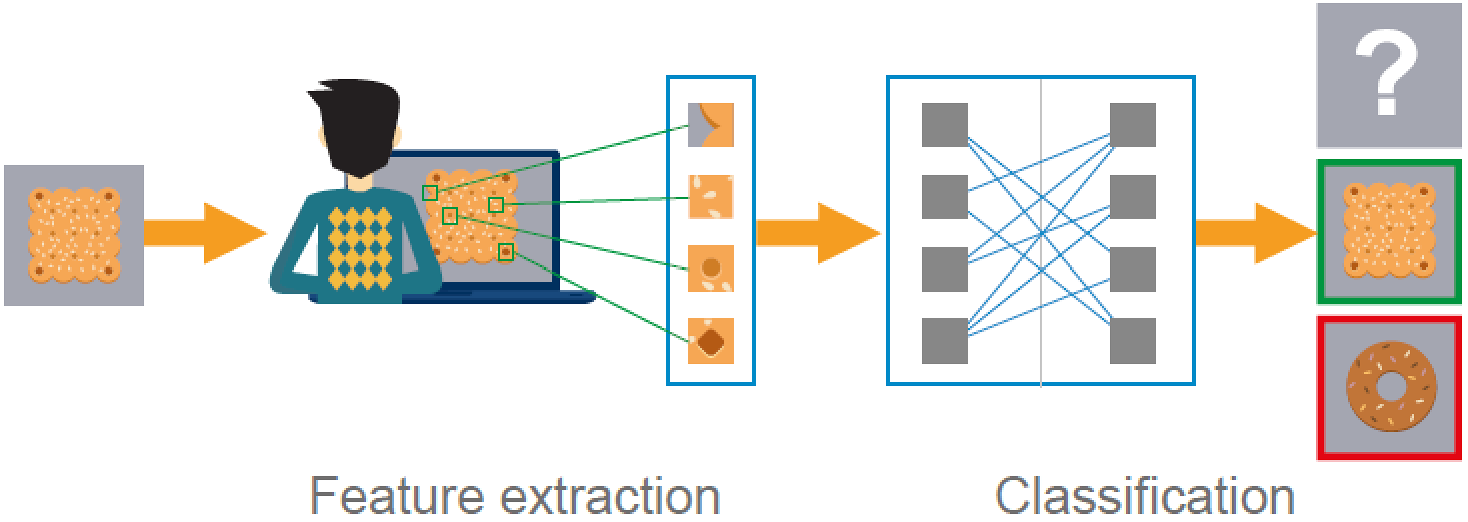
Never train on test data.
If you are seeing surprisingly good results on your evaluation metrics, it might be a sign that you are accidentally training on the test set. For example, high accuracy might indicate that test data has leaked into the training set.

https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data

https://developers.google.com/machine-learning/crash-course/validation/another-partition

FEATURE ENGINEERING

- Turn data into **feature vectors**
- Abstraction of an image



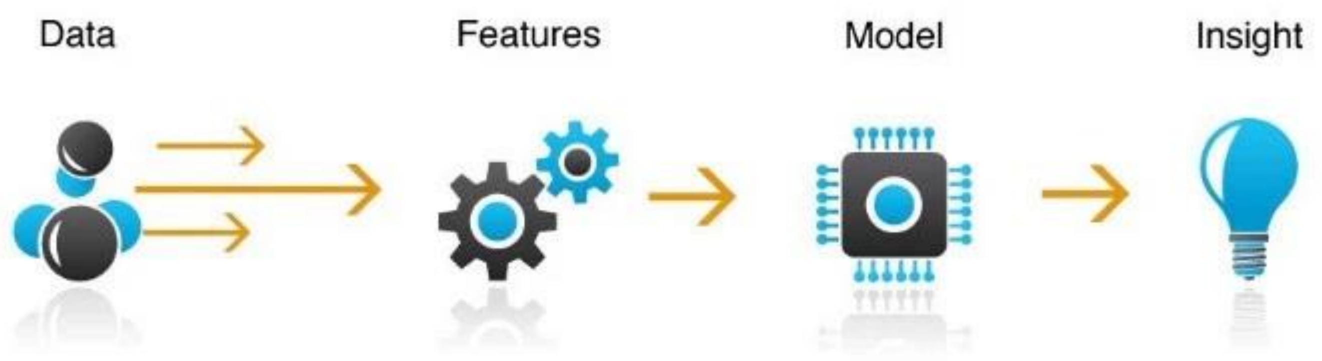
Source: Basler, Artificial Intelligence in Image Processing

WHAT MAKES A GOOD FEATURE?

https://www.youtube.com/watch?v=N9fDIAflCMY&feature=youtu.be

FEATURE ENGINEERING

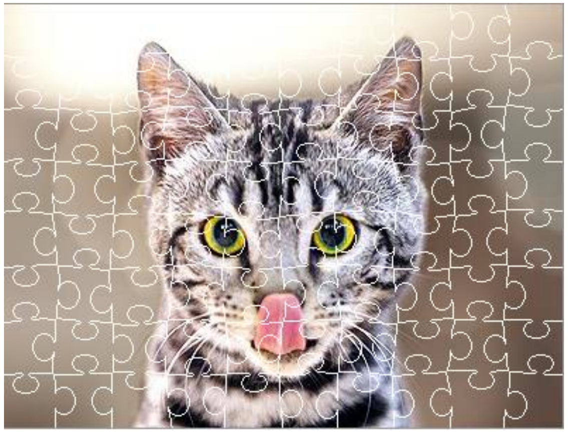
- Select features
- Decompose features (e.g. area -> length, width)
- Extract features (e.g. aggregate, combinations)
- Creating new features by gathering new data
- Add promising transformations of features (e.g., log(*x*), sqrt(*x*), x^2, etc.).



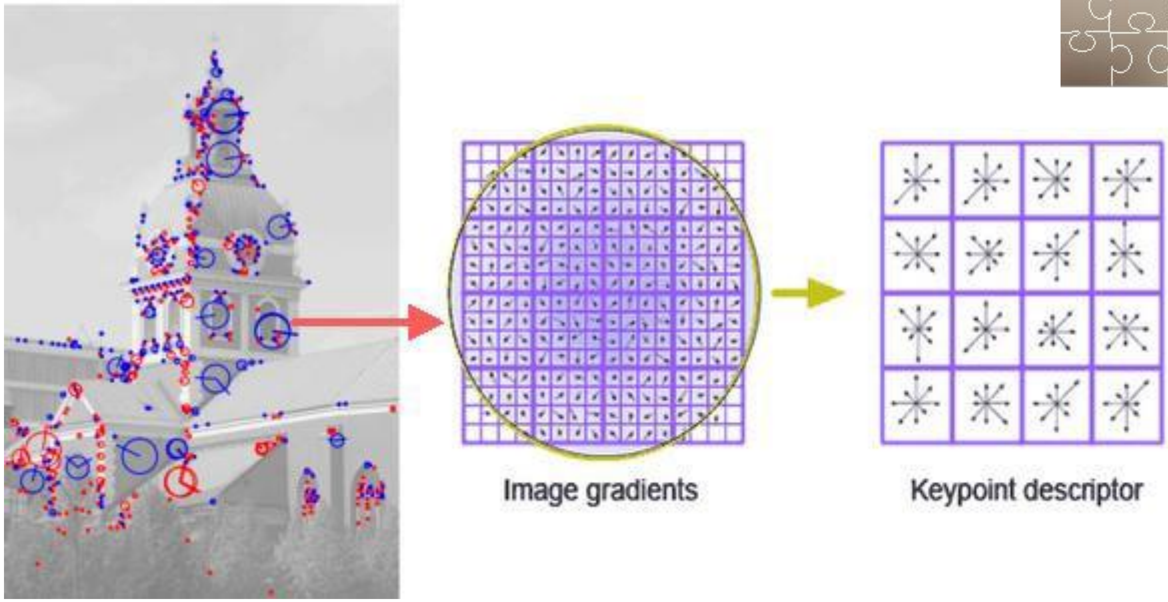
Source: VentureBeat

IMAGE FEATURE ENGINEERING

- Keypoints
- Extract descriptors
- Rotational and scaling invariance



Source: https://whyevolutionistrue.files.wordpress.com/2013/01/screen-shot-2013-01-30-at-8-36-49-am.png



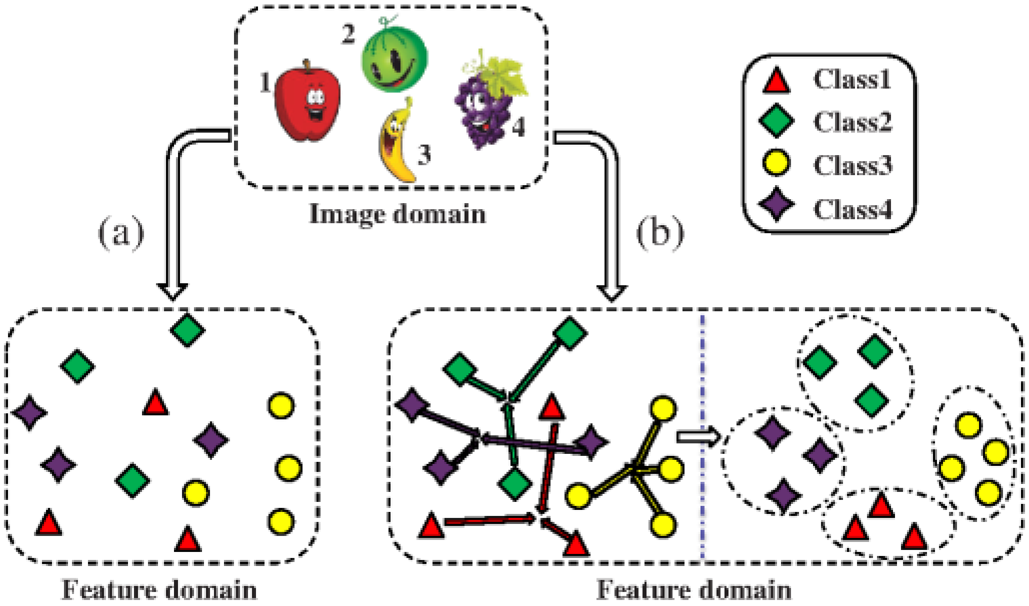
Source: https://miro.medium.com/max/1186/1*K68boX7fmtsYmyG2LlcmhQ.jpeg

KEYPOINT DETECTOR METHODS

- FAST: simple, and prone to error?
- SIFT: computationally expensive, but highly expressive.
- SURF: faster and more robust
- Star: optimized for measuring camera self-motion
- BRIEF: extracting feature descriptions
- BRISK
- ORB
- FREAK
- ….

QUALITIES OF GOOD FEATURES

- Informative
- Discriminating
- Independent
- Nearly unique

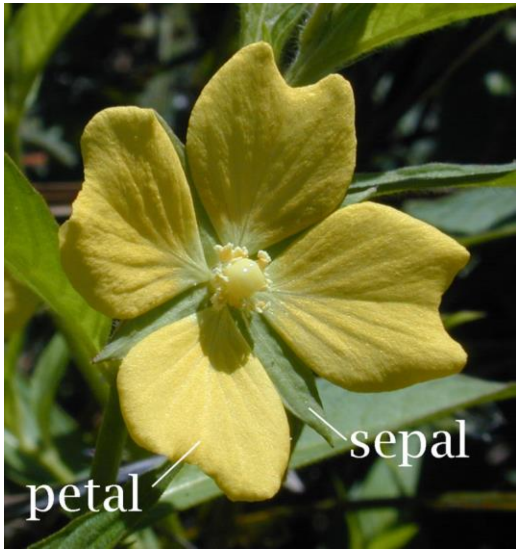


Source: https://www.spiedigitallibrary.org/ContentImages/Journals/JEIME5/26/1/013023

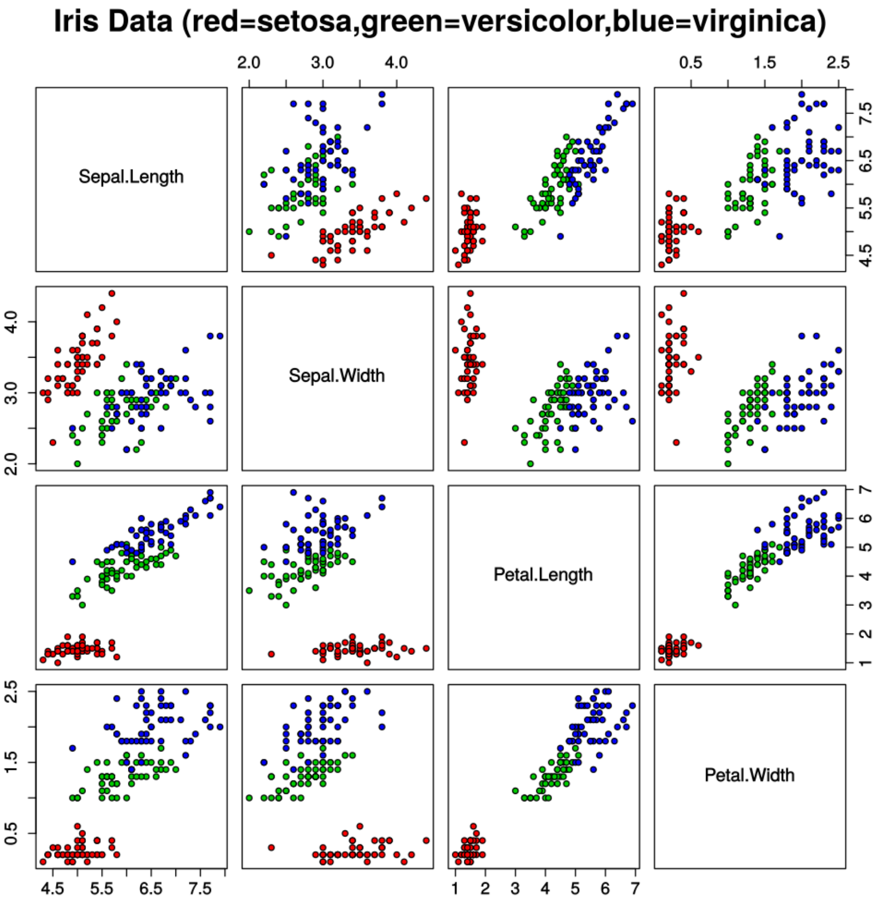
- NB later on feature scaling may be required

EXAMPLE: *IRIS* FLOWER DATA SET

- Sepal and petal width and length



Source: https://en.wikipedia.org/wiki/Sepal

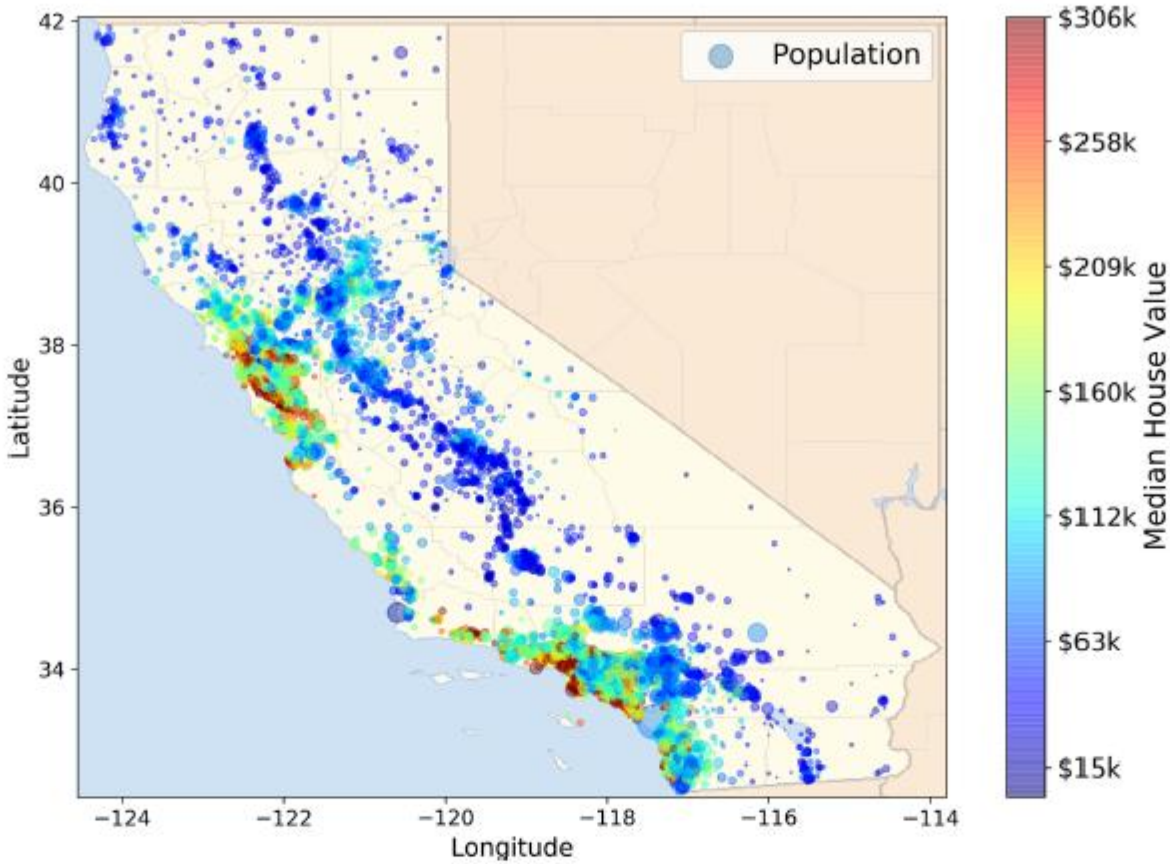


Source: https://en.wikipedia.org/wiki/Iris_flower_data_set

EXPLORE THE DATA

- Get insights from a domain expert
- Set aside a subset of the data for exploration
- Study each attribute and its characteristics
  - categorical, int/float, bounded/unbounded, text, structured,
  - Noisiness and type of noise (stochastic, outliers, rounding errors)
- Visualize the data
- Study the correlations between attributes
- Think about how you would solve the problem manually

EXAMPLE: CALIFORNIA HOUSING PRICES

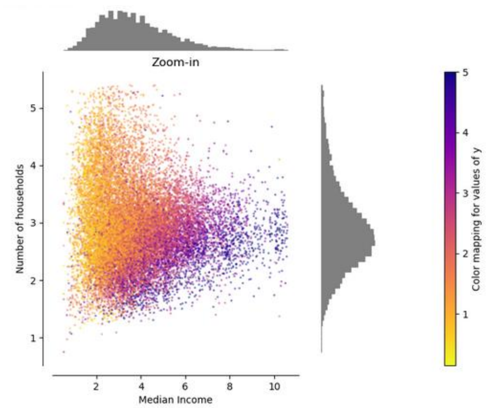


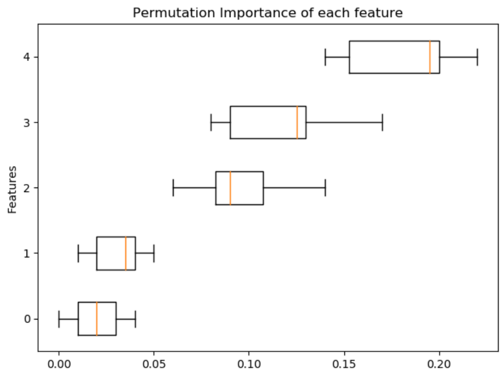
Source: Géron, ISBN: 9781492032632

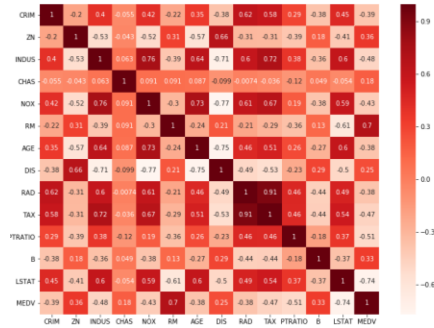
HAN_UNIVERSITY OF APPLIED SCIENCES

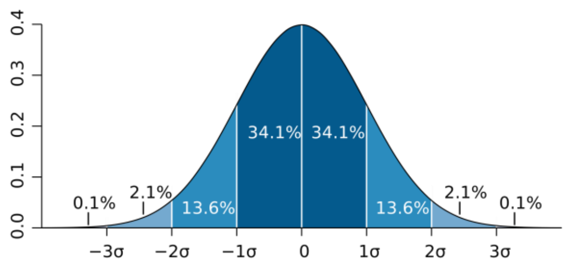
TOOLS FOR EXPLORATORY DATA ANALYSIS

- Univariate analysis
- Histogram
- Scatterplot
- Boxplot
- Correlation heatmap



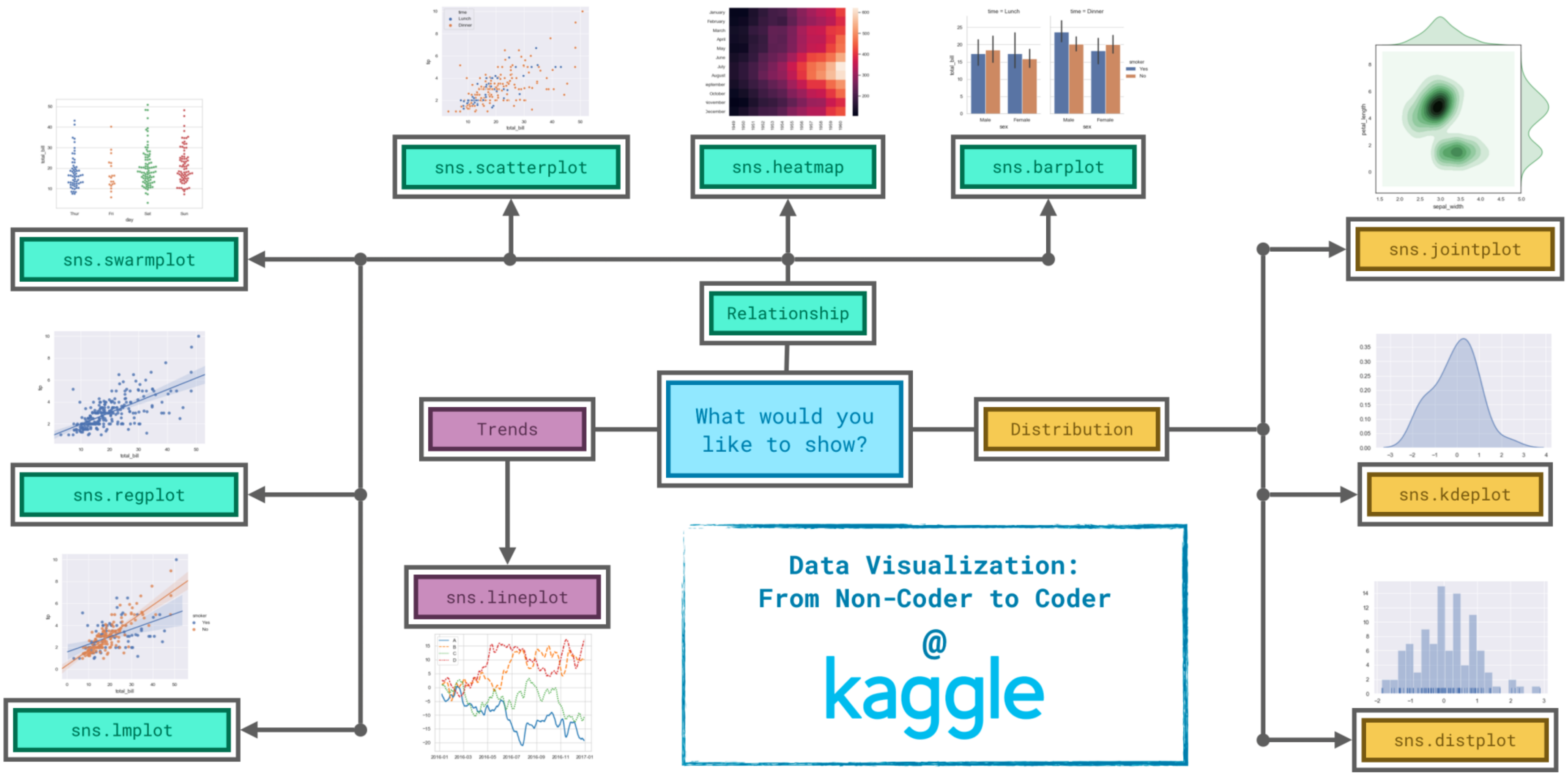






Sources:
https://en.wikipedia.org/wiki/Exploratory_data_analysis
https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html
https://en.wikipedia.org/wiki/Box_plot

HAN_UNIVERSITY OF APPLIED SCIENCES



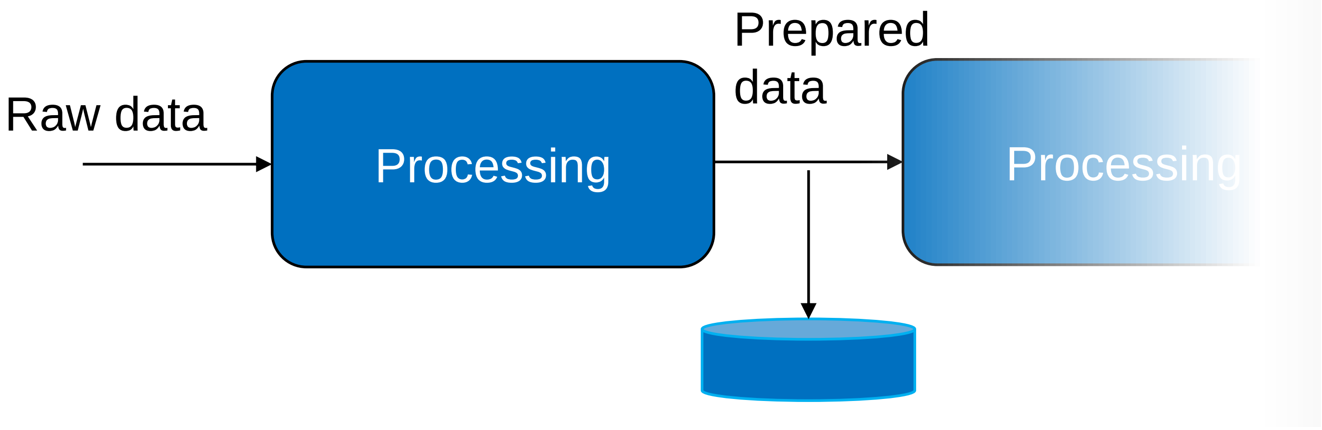
HAN_UNIVERSITY OF APPLIED SCIENCES

DATA QUALITY ISSUES

- Insufficient data. ML needs massive amounts of training data.
- Messy data. Data that contains a large amount of conflicting or misleading information.
- Dirty data. Data that contains missing values, categorical and character features with many levels, and inconsistent and erroneous values.
- Sparse data. Data that contains very few actual values and is instead composed of mostly zeros or missing values.
- Inadequate data. Data that is either unbalanced, incomplete or biased.

PIPELINES

- Sequence of data processing components
- First step is preparing the data



PREPARING DATA

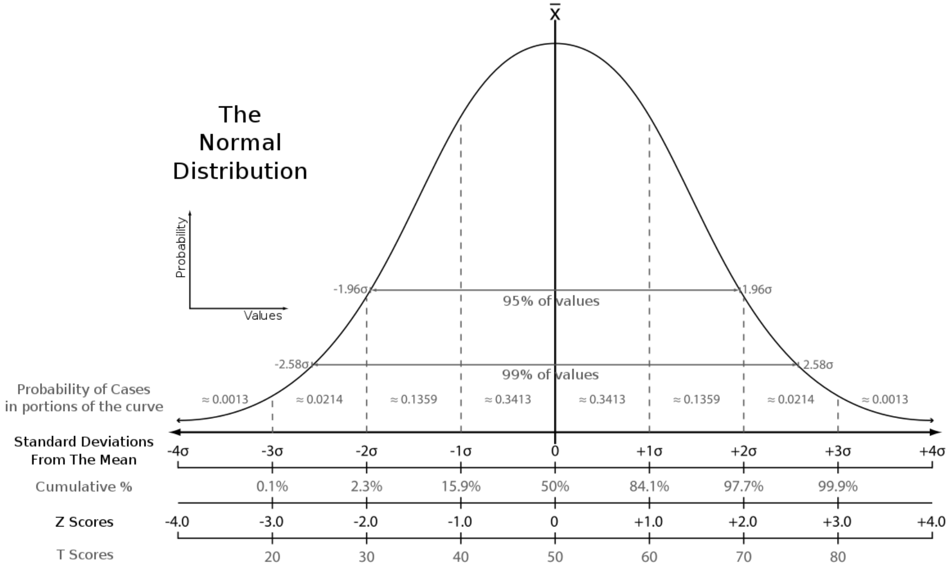
- Data cleaning:
  - Fix or remove outliers (optional)
  - Fill in missing values (e.g., with zero, mean, median…) or drop their rows (or columns).

- Feature computation:
  - Selection
  - Transformation

- Feature scaling:
  - Standardize or normalize features.

EXAMPLE: OUTLIER DETECTION

- Assume feature values are normally distributed
- Compute Z-score of value
- Detect if z-score is above threshold
- Typically used in low dimensional feature space



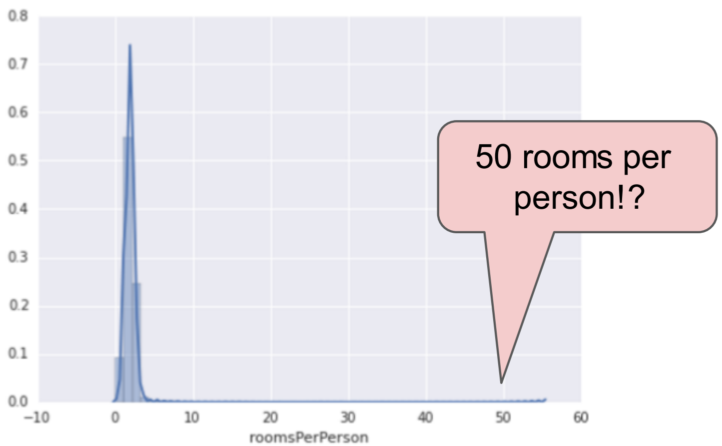
Source: https://en.wikipedia.org/wiki/Standard_score

CLEANING DATA

- Scrubbing
  - Detect omitted values or duplicated examples and remove
  - Detecting bad feature values or labels can be far trickier
  - Outlier detection
  - Limited or sparse features / attributes

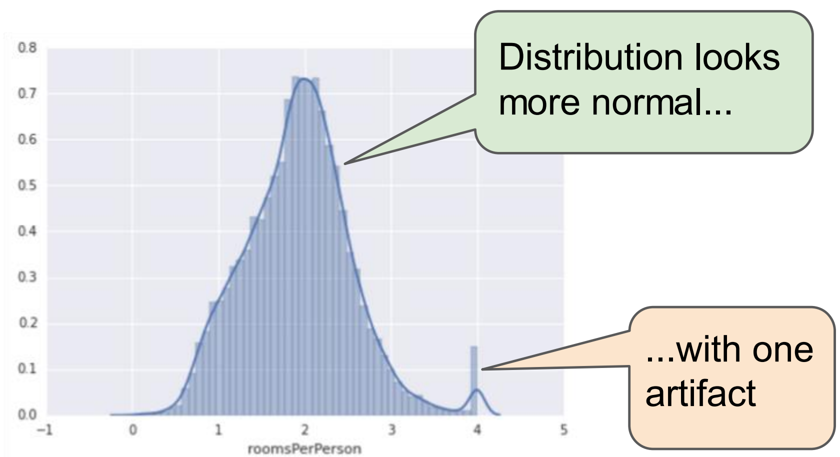
- Scaling
  - Avoid algorithm bias to features having a wider range
  - Help algorithms converge more quickly
  - Handling extreme outliers, e.g. log scaling, clipping

EXAMPLE: CLIPPING



`roomsPerPerson = totalRooms / population`

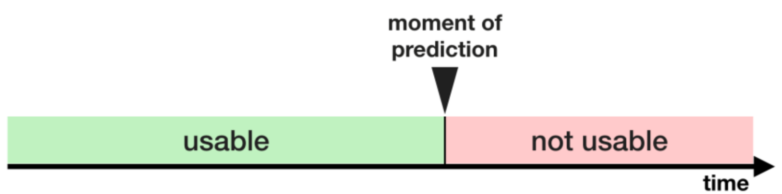
https://developers.google.com/machine-learning/crash-course/representation/cleaning-data



`roomsPerPerson = min(totalRooms / population, 4)`

PITFALLS

- **Insufficient data**
- **Sampling bias:** your dataset is not representative of the cases you want to generalize to
- **Unbalanced data:** your dataset does not represent classes equally (skewed, nonresponse)
- **Non-stationary data**: distribution changes within the data set
- **Over/underfitting:** optimizing for the wrong thing by considering too many or too few features
- **Train-test contamination:** you fail to distinguish training data from validation data.
- **Target leakage:** your training data includes data that will not be available at the time you make predictions



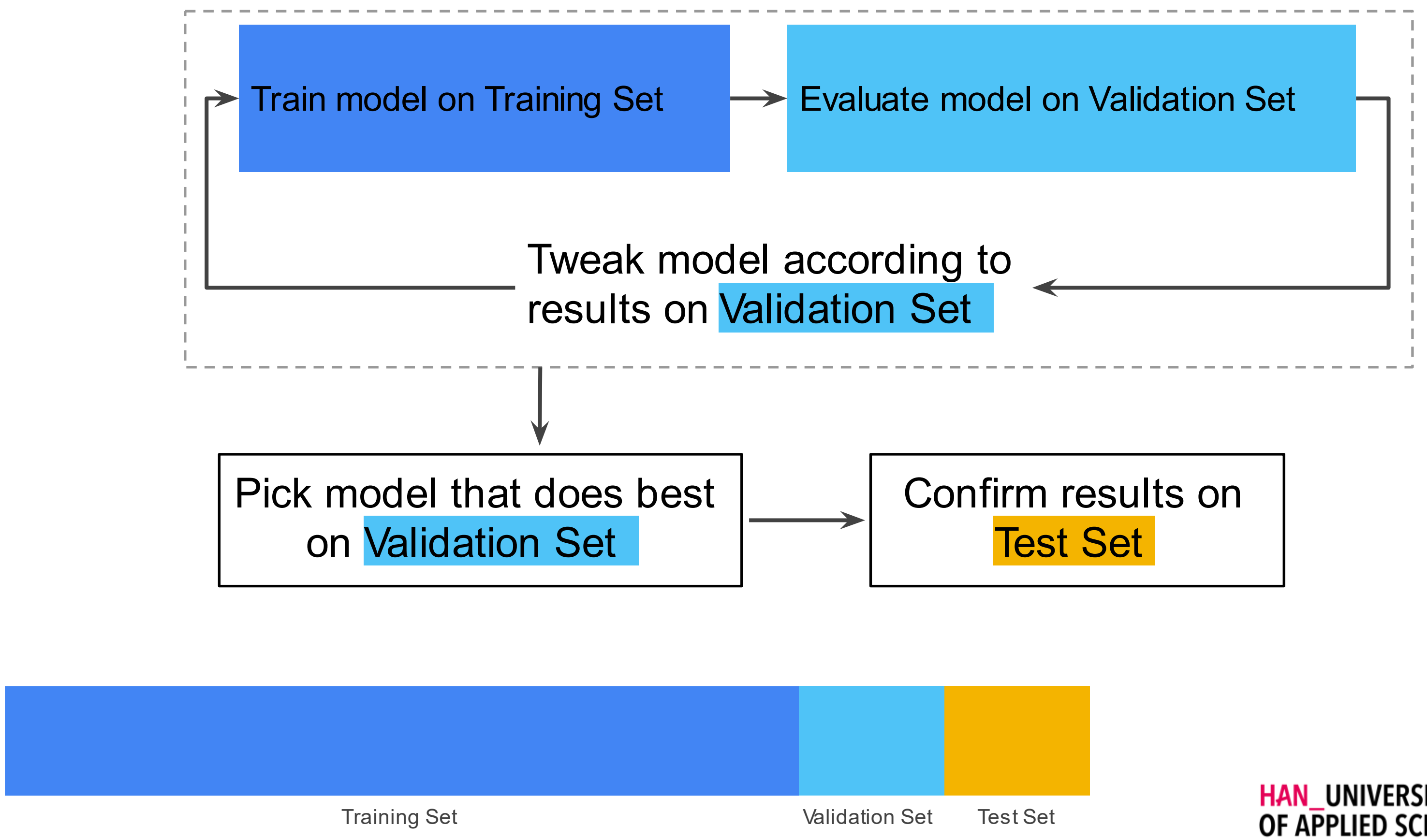
HAN_UNIVERSITY
OF APPLIED SCIENCES

BIASES

- Selection bias: tendency to implicitly filter data based on some arbitrary criteria and then try to make sense out of it without realizing or acknowledging that we’re working with incomplete data
- Availability bias: tendency to work with data that’s easier to obtain rather than looking for data that is harder to gather but is more informative.
- False causality: tendency to assume that correlation implies causation
- Sunk cost fallacy: tendency to make decisions based on how much is already invested

INCLUDE VALIDATION

- Validation Set: Another Partition



https://developers.google.com/machine-learning/crash-course/validation/another-partition

NEXT WEEK

- Theory:
- Quiz
- Supervised machine learning: Support vector machines, Decision trees

- Hands-on:
- Exploratory data analysis
- -> to maximally benefit start collecting data