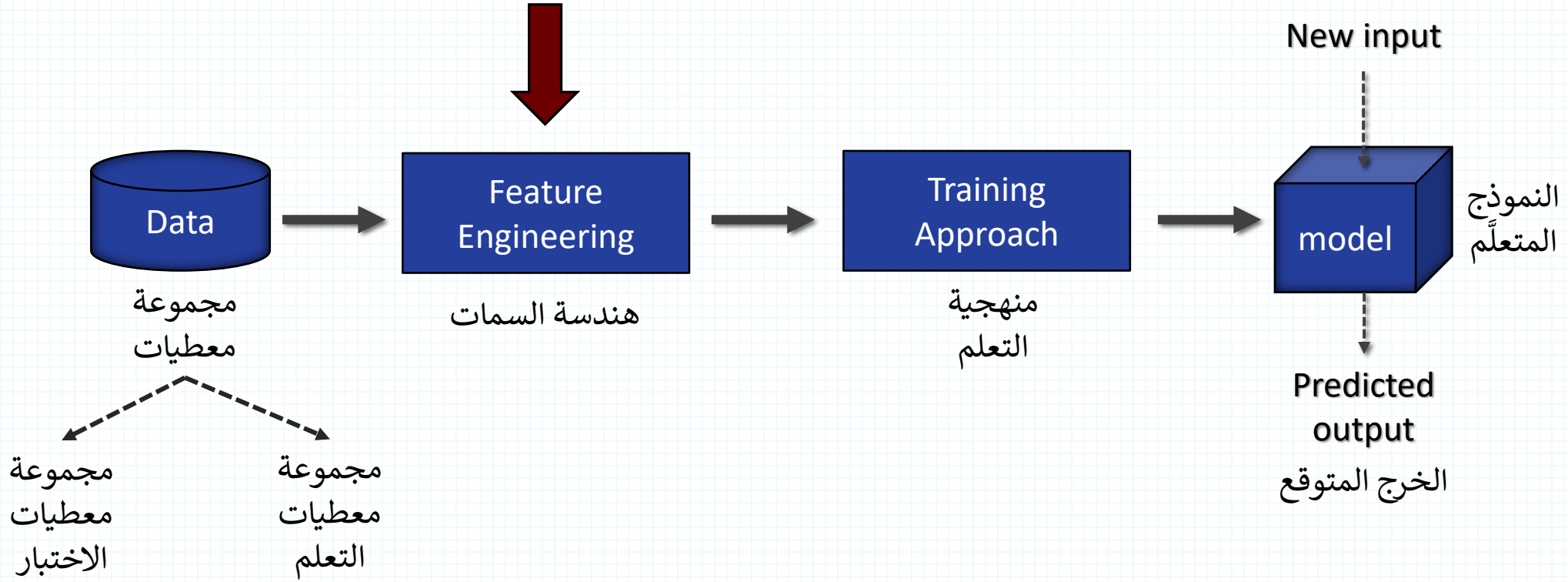
تعلم الآلة	كلية الهندسة المعلوماتية	المحاضرة الخامسة

هندسة السمات

Feature Engineering: Feature extraction & Feature Selection

د. رياض سنبل

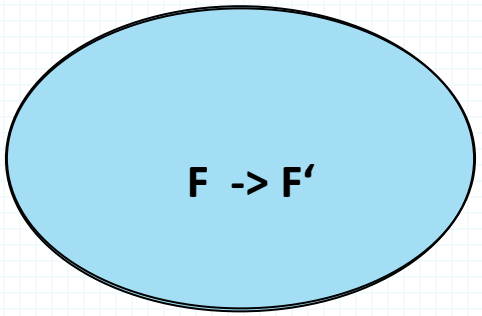
ML Pipeline


Data
مجموعة معطيات
مجموعة معطيات الاختبار
مجموعة معطيات التعلم
Feature Engineering
هندسة السمات
Training Approach
منهجية التعلم
New input
model
النموذج المتعلَّم
Predicted output
الخرج المتوقع

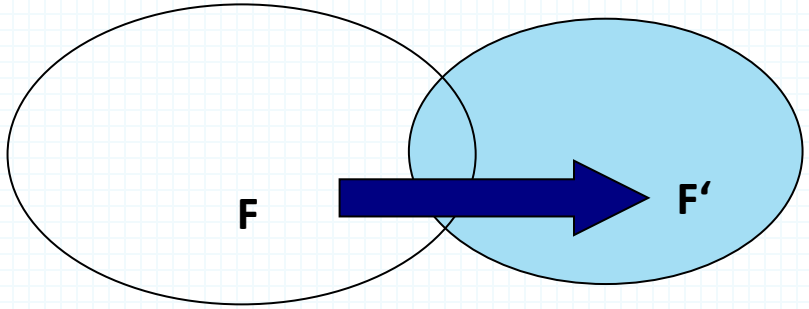
Feature (Preprocessing vs Selection vs Extraction)

- **Feature Preprocessing:** Clean, normalize, transform features the values of specific feature using a defined formula.
- **Feature extraction:** Creates new features (dimensions) defined as functions over all features
- **Feature selection:** Chooses subset of features

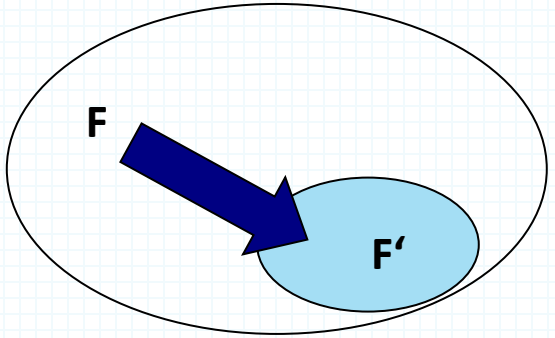
Feature Preprocessing



Feature extraction



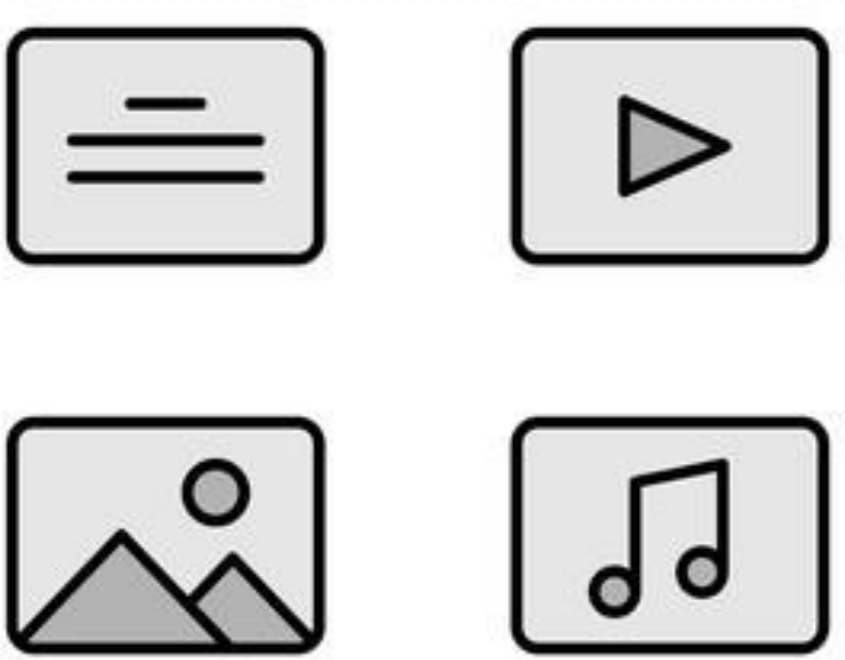
Feature selection



Feature Extraction

Feature Extraction

- Feature extraction is a process in which you take raw data, often in the form of complex and high-dimensional variables, and transform it into a reduced and more manageable set of features.



Feature Extraction: SMS Spam

- SMS Message (arbitrary text) -> 5 dimensional array of binary features
  - 1 if message is longer than 40 chars, 0 otherwise
  - 1 if message contains a digit, 0 otherwise
  - 1 if message contains word ‘call’, 0 otherwise
  - 1 if message contains word ‘to’, 0 otherwise
  - 1 if message contains word ‘your’, 0 otherwise

“SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info”

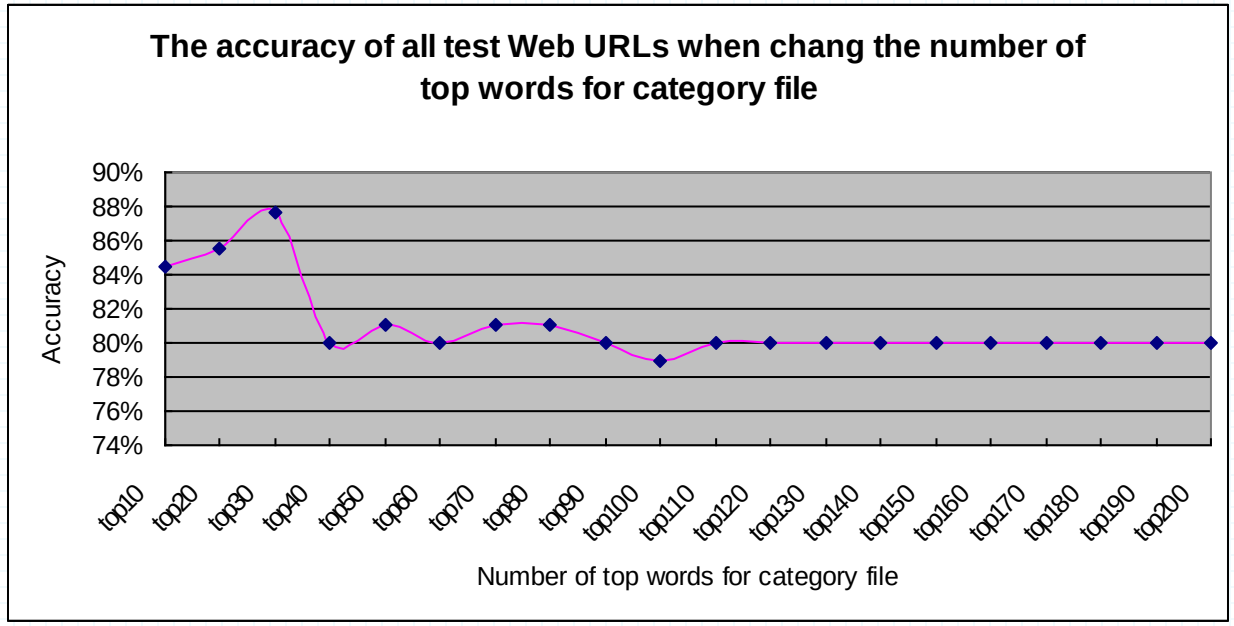
Long?	HasDigit?	ContainsWord(Call)	ContainsWord(to)	ContainsWord(your)

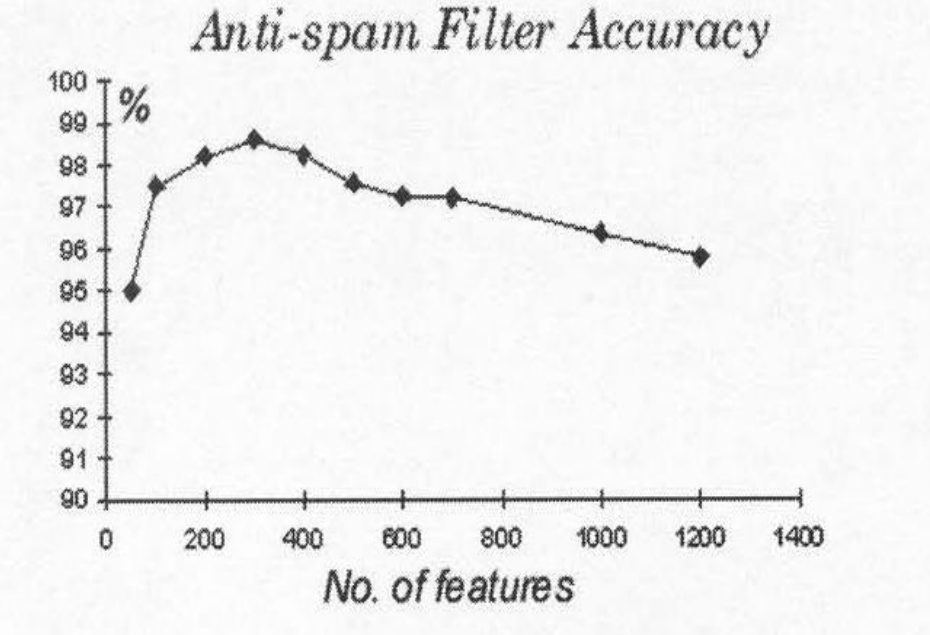
Feature Selection

Feature Selection: What

- You have a data has 100,000 fields (features)
  - Examples?
- You want to use it to build a classifier, so that you can predict something
  - What are the possible problems?
- you need to cut it down to 1,000 fields before you try machine learning. Which 1,000?
  - How to do that => Feature Selection

Feature Selection: Why





Why accuracy reduces

Why accuracy reduces

- **Noise:** The additional features typically **add noise**. Machine learning will pick up on **fake correlations**, that might be true in the training set, but not in the test set (**overfitting**).
  - Example: what will happen if you learn ID3 with too many noisy data?

- **Explosion:** For some ML methods, more features means **more parameters to learn** (more NN weights, more decision tree nodes, etc…) – the increased space of possibilities is **more difficult to search**.

Univariate feature selection

- Univariate feature selection works by selecting the best features based on univariate statistical tests. It can be seen as a preprocessing step to an estimator.
  - Example: In scikit, *SelectKBest* removes all but the h highest scoring features based on a scoring function.
- Methods are used to rank features by importance
  - Pearson correlation coefficient
  - F-score
  - Chi-square
  - Signal to noise ratio
  - And more such as mutual information,

```
>>> from sklearn.datasets import load_digits
>>> from sklearn.feature_selection import SelectKBest, chi2
>>> X, y = load_digits(return_X_y=True)
>>> X.shape
(1797, 64)
>>> X_new = SelectKBest(chi2, k=20).fit_transform(X, y)
>>> X_new.shape
(1797, 20)
```