Machine Learning and NLP: Advances and Applications
Day 1: Machine Learning Basics

1/22/2020

Yoshi Suhara

Course Overview

- Goal: Learning ML/NLP basics and how to apply the techniques to your own problems

- The course will cover
  - both theory (lecture) and practice (hands-on)

Course Overview

- Day 1: Machine Learning Basics
- Day 2: NLP Basics
- Day 3: Advanced Techniques and Applications

Course Overview

- Day 1: Machine Learning Basics
  - **Hands-on material 1**
- Day 2: NLP Basics
  - **Hands-on material 2**
- Day 3: Advanced Techniques and Applications
  - **Hands-on material 3**

Day 1 **Day 2** **Day 3**

Please Ask Questions!

Day 1

- What is ML?
- Supervised Learning
- ML Problem Formulation
- Basic ML Algorithm
- ML Evaluation
- ML Pipeline

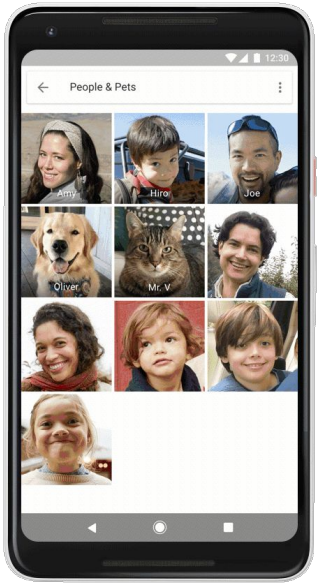
What is ML?

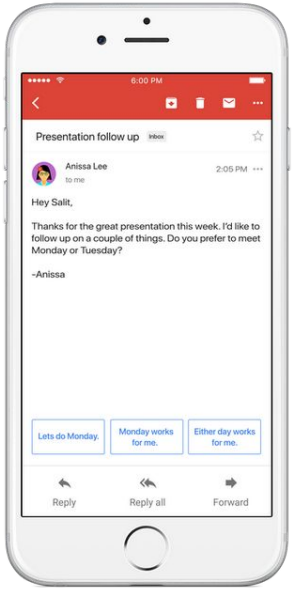
What is Machine Learning?

- A magic that makes your laptop/smartphone really smart
  - Can you name apps/functions that use machine learning?

What is Machine Learning?

- A magic that makes your laptop/smartphone really smart





What is Machine Learning?

"A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*." - Tom Mitchell (1997)

What is Machine Learning?

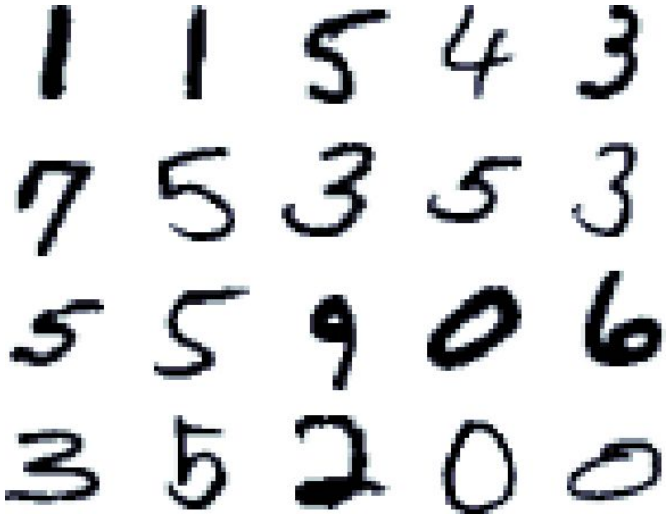
"A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*." - Tom Mitchell (1997)

- experience *E* = data
- performance measure *P* != Our goal

Machine Learning Categories

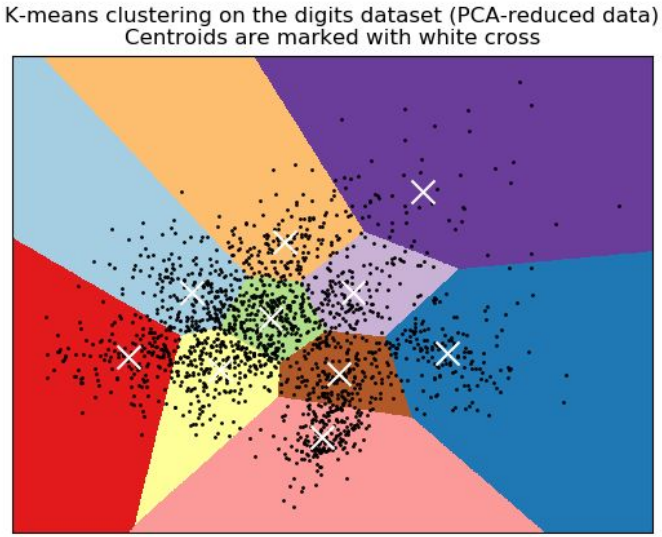
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

(1) Supervised Learning



(2) Unsupervised Learning

- Clustering or Representation Learning for Visualization or better supervised learning models



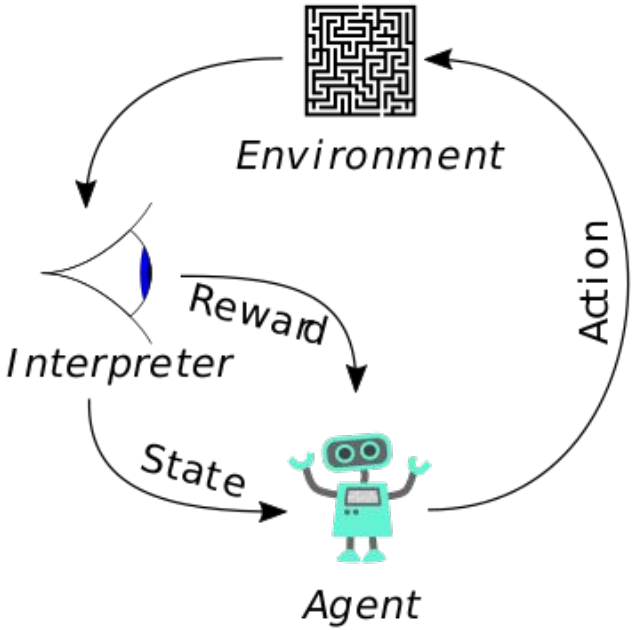
Clustering + PCA

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Topic Models

(3) Reinforcement Learning

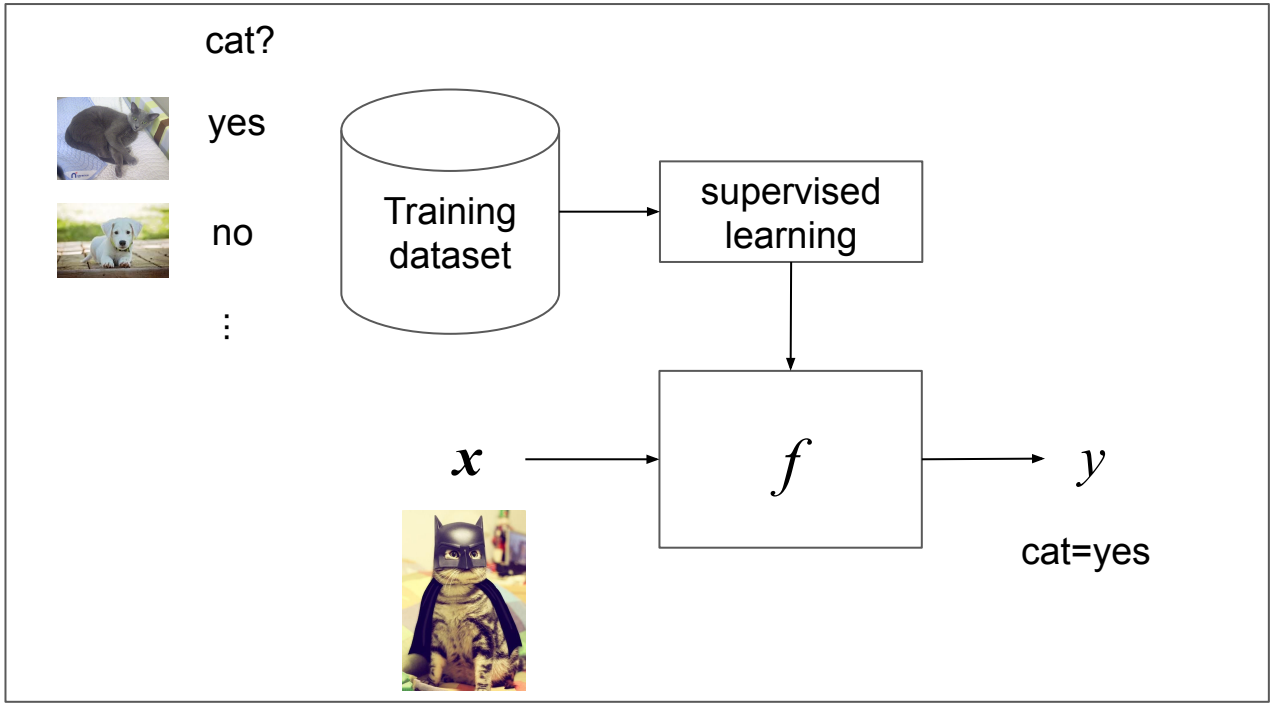


RL is not be covered by this course

Supervised Learning

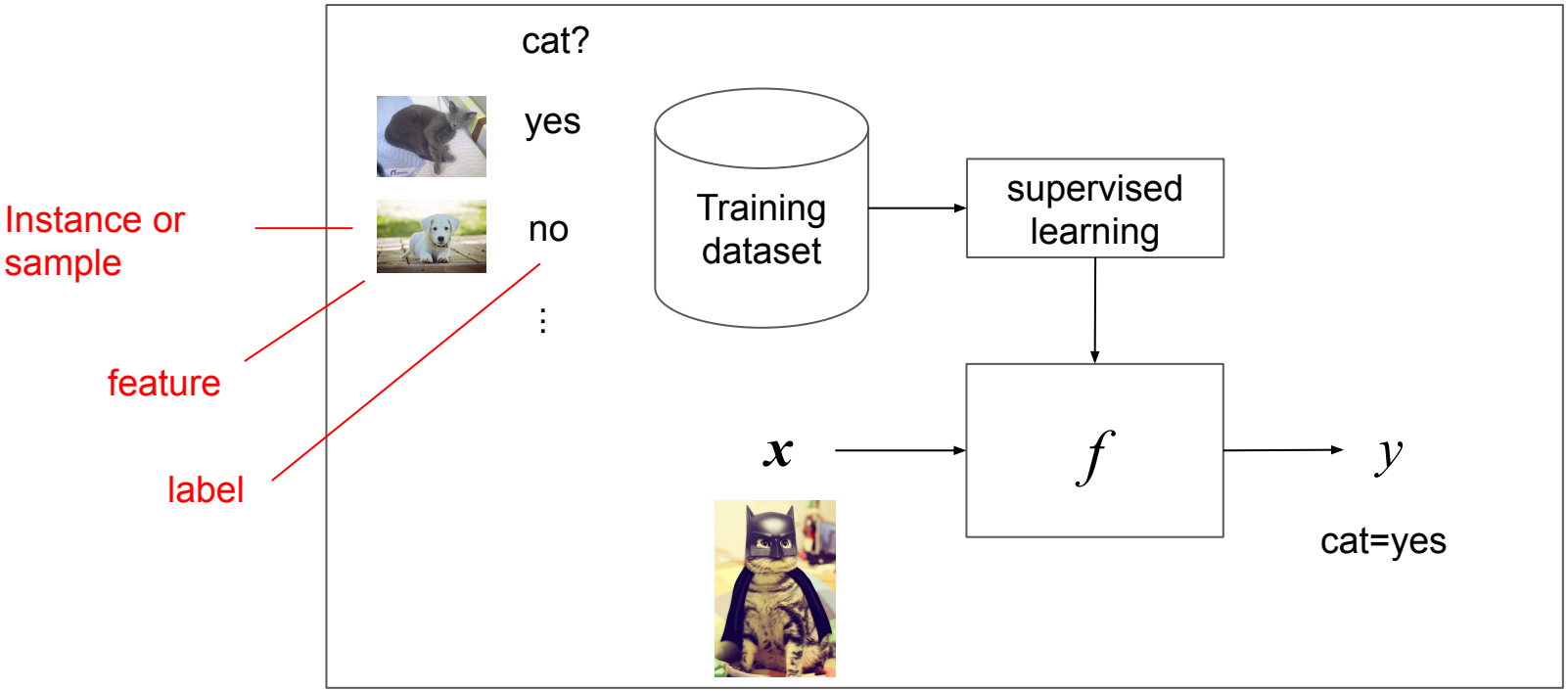
Supervised Learning

Supervised learning is a framework that builds a predictive model based on "labeled" training data



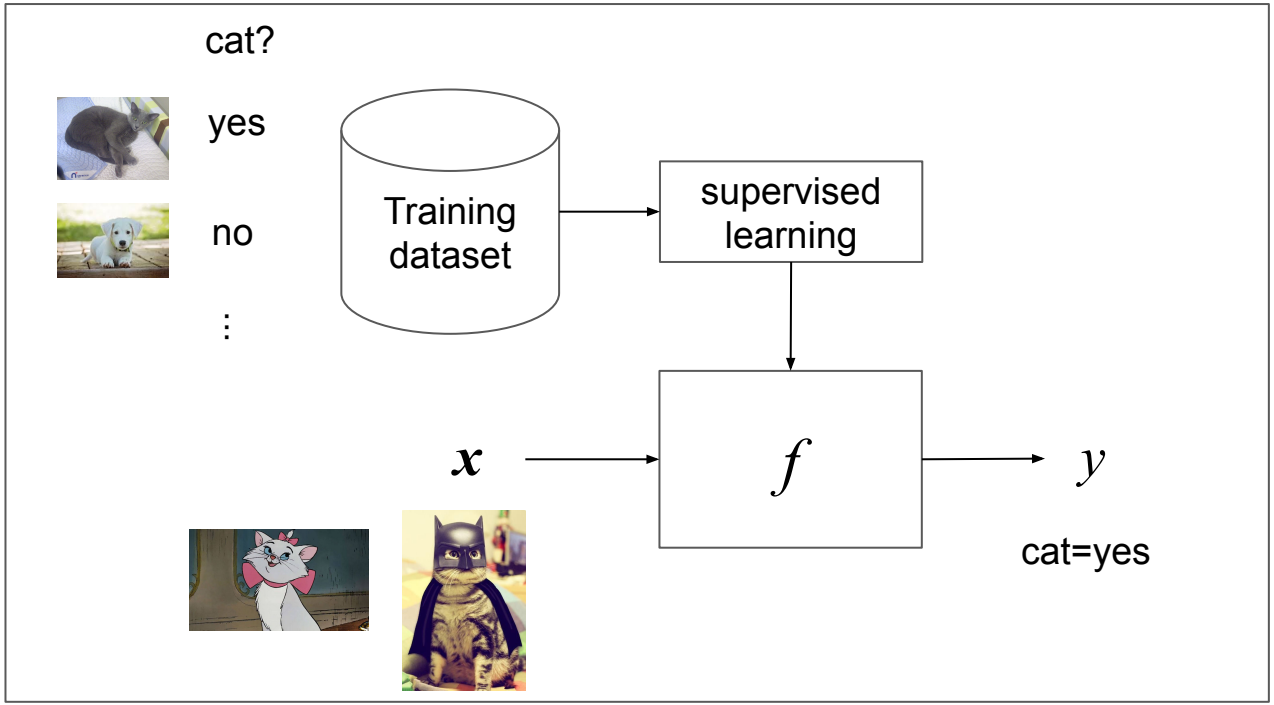
Supervised Learning

Supervised learning is a framework that builds a predictive model based on "labeled" training data



Note: Generalization (in-domain/out-domain)

- Machine Learning models are NOT good at predicting something they haven't seen before



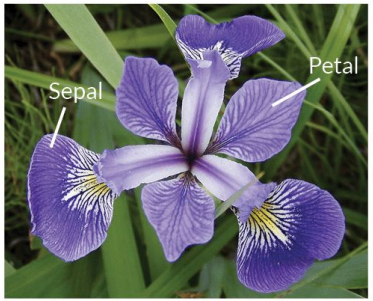
Supervised Learning = Learning a function

Supervised learning algorithm learns a function that maps a **feature vector** into a **target value**

$$f : \boldsymbol{x} \rightarrow y$$

$\{0, 1\}$	Binary classification
$\{0, 1, 2, \ldots, N\}$	Multi-class classification
$\mathbf{R}$	Regression

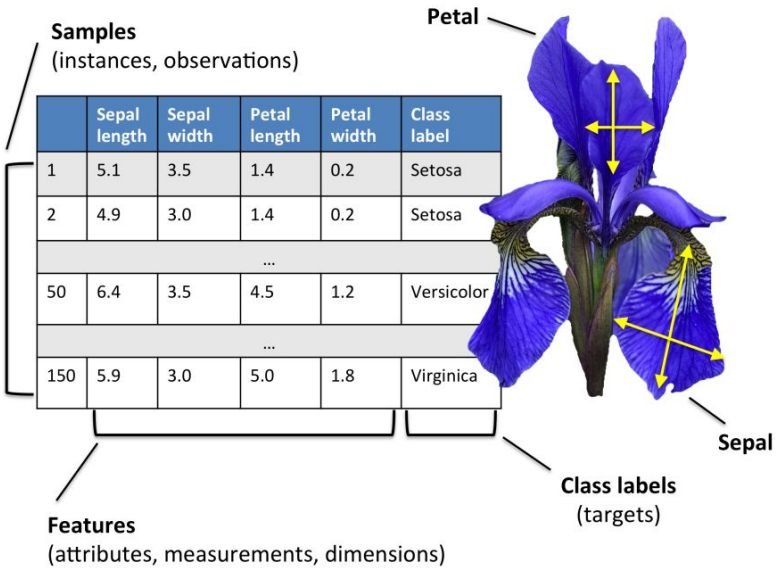
Example: Fisher's Iris Datasets



Iris Versicolor

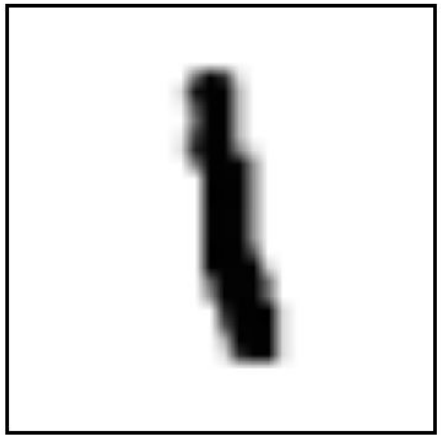
Iris Setosa

Iris Virginica



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Example: Handwritten Digit Recognition



$\approx$

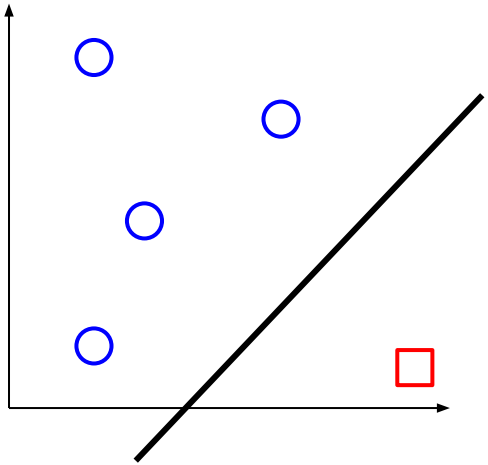
$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .6 & .8 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .7 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .7 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .5 & 1 & .4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & .4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & .4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & .7 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & .9 & 1 & .1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & .3 & 1 & .1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

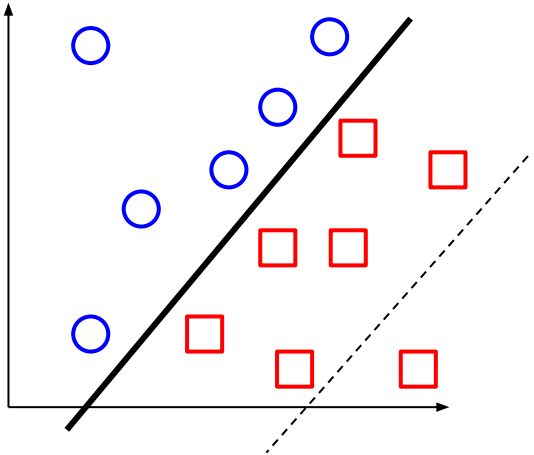
$\rightarrow$

x	y
	1

Positive/negative examples

- ML algorithms need both **positive/negative** examples to have a good model (especially, **borderline examples**)





Imbalanced dataset may need a special care (sampling or class-weights)

Don't worry: Training is 3-line Python Code!

```
>>> clf = LogisticRegression()
>>> clf.fit(X, y)
>>> clf.predict(X)
>>> ...
```

ML Problem Formulation

Different ML Problems

- Different target sets (often) need different classes of ML algorithms

$$f : \boldsymbol{x} \rightarrow y$$

$\{0, 1\}$	Binary classification
$\{0, 1, 2, \ldots, N\}$	Multi-class classification
$\mathbf{R}$	Regression

Quiz 1. Spam Detection

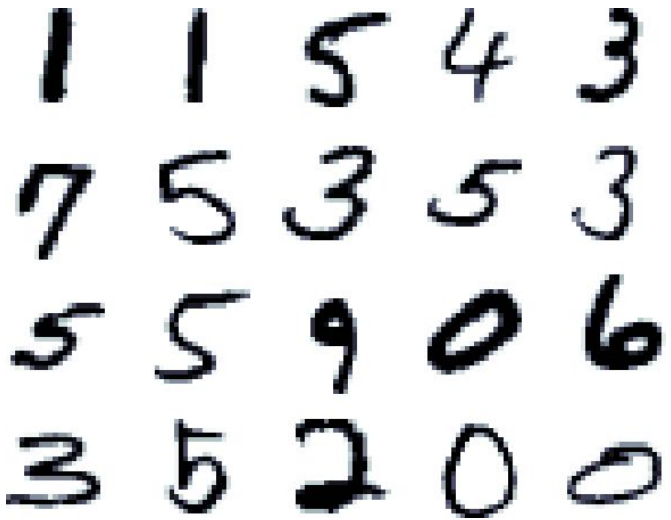
- Input:
- Target:
- Task:

Quiz 1. Spam Detection

- Input: Subject, Content, Sender etc.
- Target: Spam or Ham
- Task: Binary classification

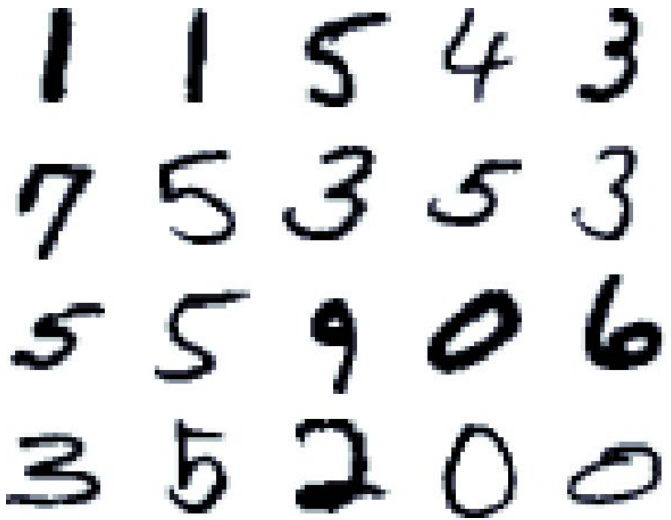
Quiz 2. Photo Classification

- Input:
- Target:
- Task:



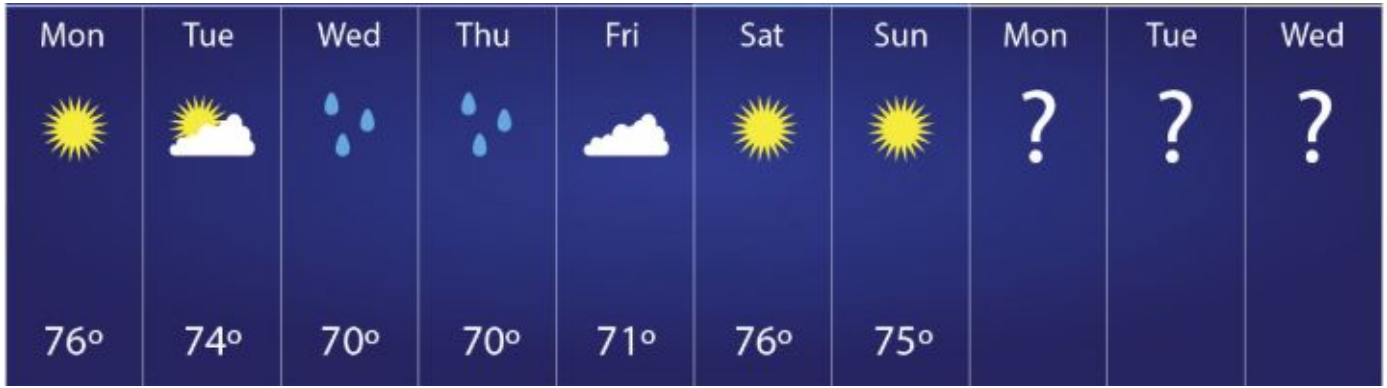
Quiz 2. Handwritten Digit Recognition

- Input: Grayscale pixel data
- Target: 10 digits (0-9)
- Task: Multi-class classification



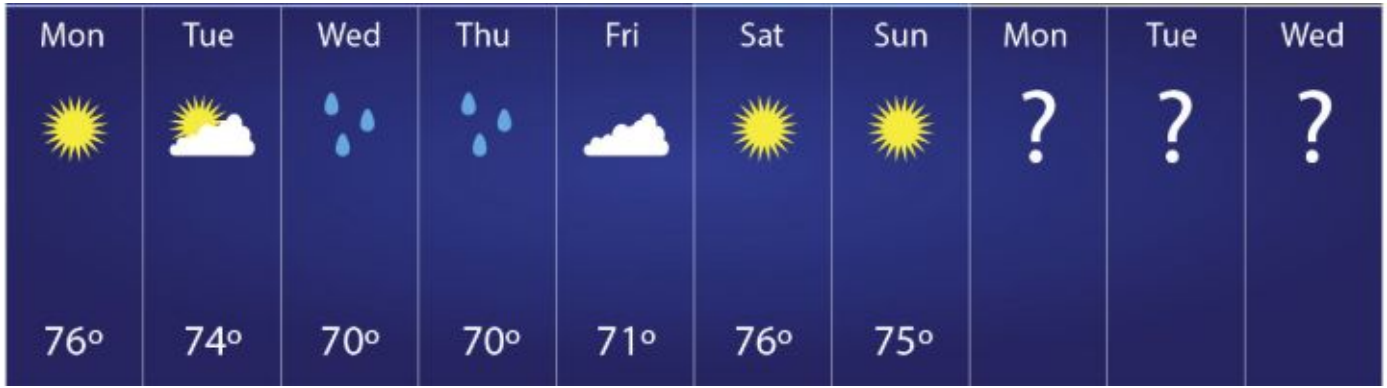
Quiz 3. Temperature Prediction

- Input:
- Target:
- Task:



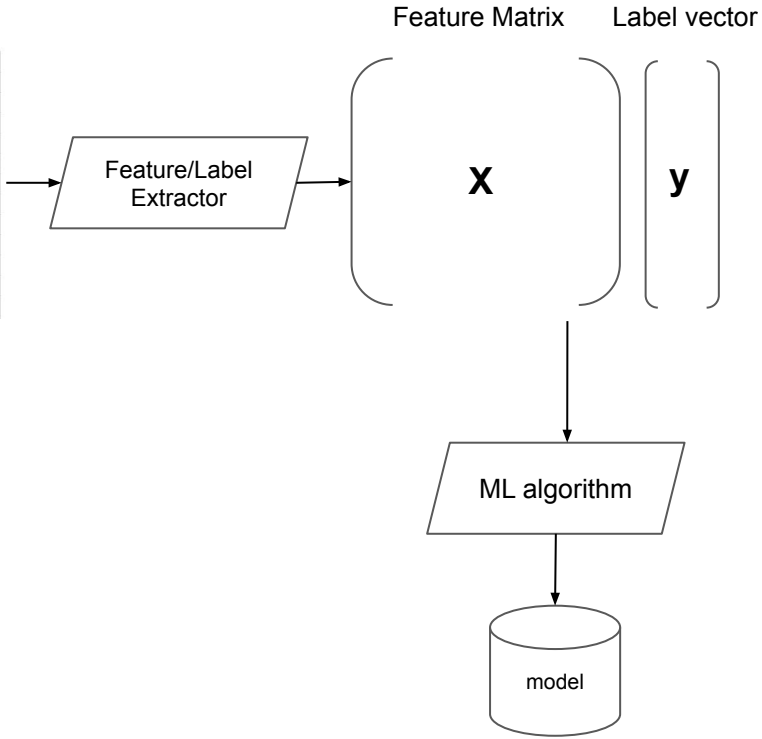
Quiz 3. Temperature Prediction

- Input: Temperature/Humidity/Weather in past days
- Target: Temperature (of the next day)
- Task: Regression



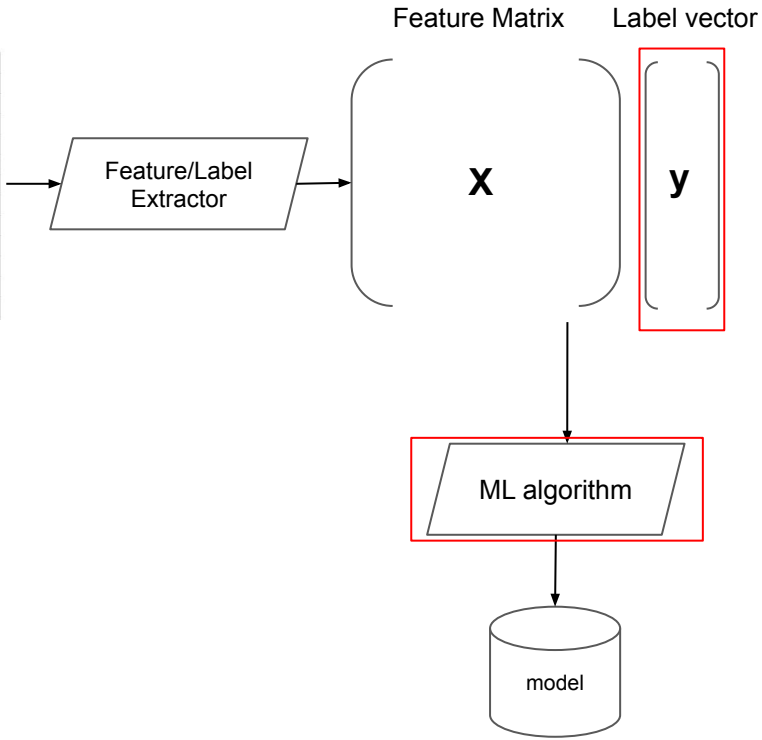
ML Workflow

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S



ML Workflow

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S



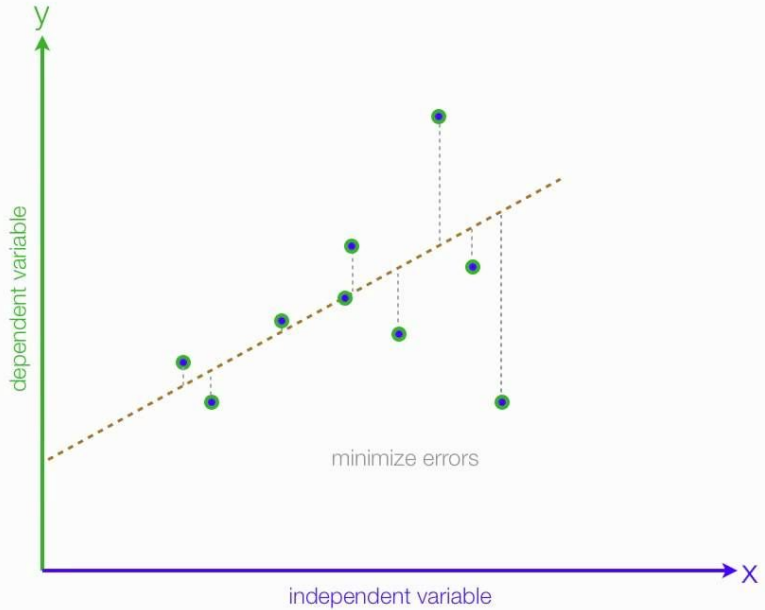
Key points

- 1) Feature matrix: **X**
- 2) Label vector: **y**
- 3) Task (e.g., classification, regression, etc.)

Basic ML Algorithms

Linear Regression (Ordinary Least Square)

- Linear model: $y = \boldsymbol{w}^T \boldsymbol{x}_i + b$
- OLS fits a linear model that minimizes the sum of squared errors

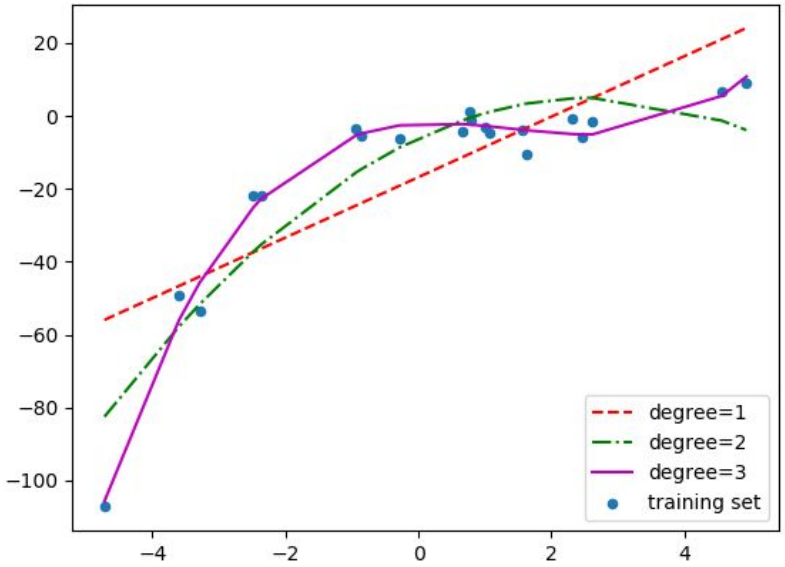


$$l(D;\ w) = \sum_{i=1}^{N} (y - \hat{y})^2$$

$$w = (X^T X)^{-1} X^T y$$

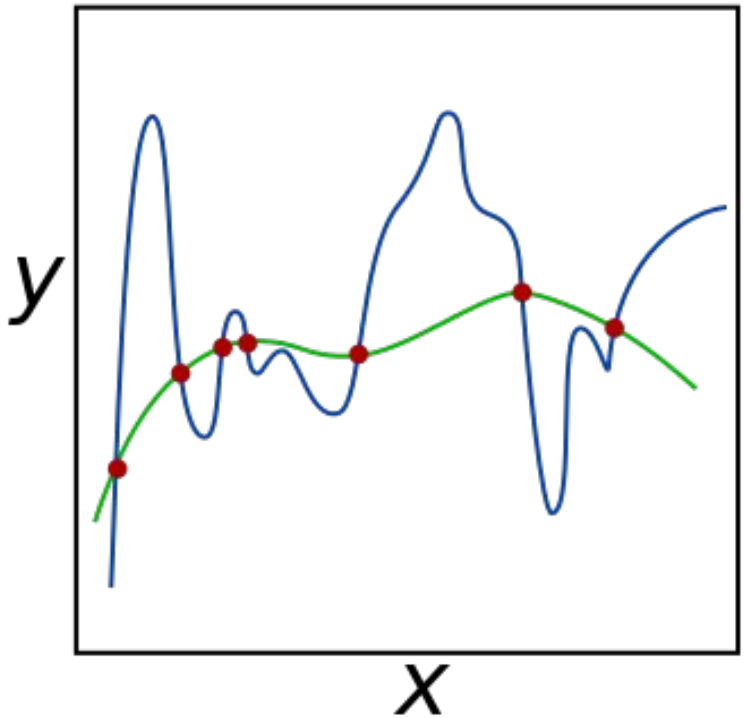
Polynomial Regression is Linear Regression

- Linear wrt "weight" parameters

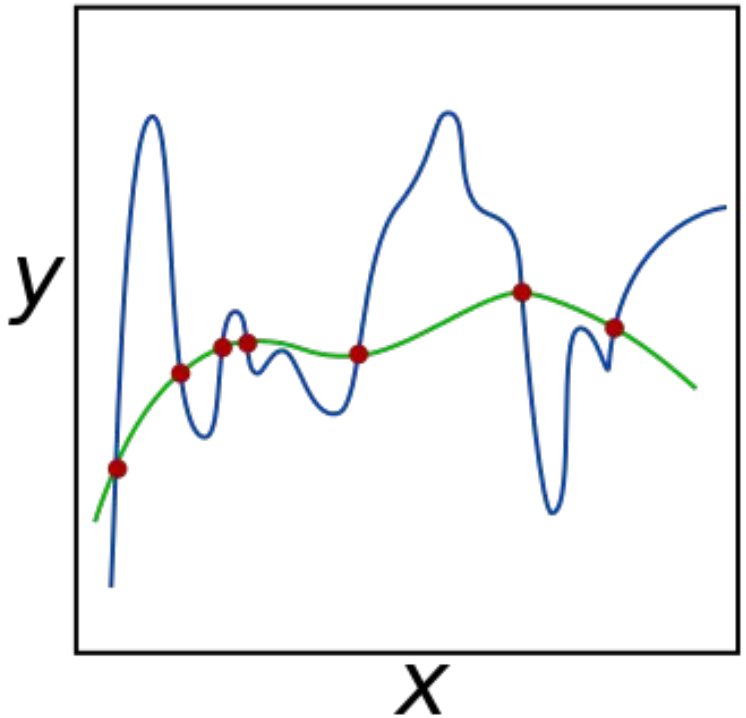


$$y = w_1 x + w_2 x^2 + b$$

Regularization: Giving penalty to parameters



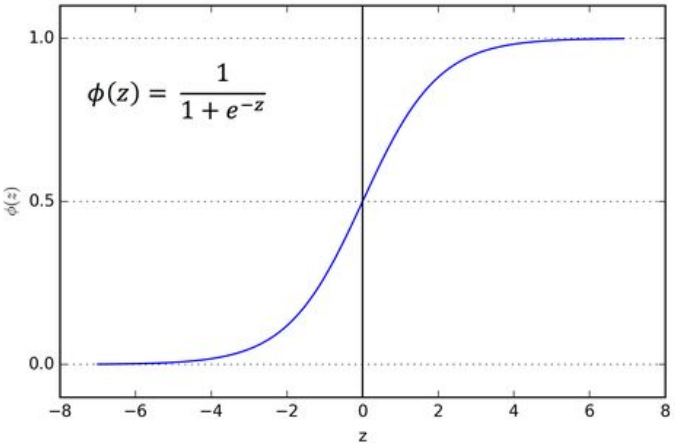
Regularization: Giving penalty to parameters



$$l(D;\ w) = \sum_{i=1}^{N}(y - \hat{y})^2 \quad +\lambda \sum_{j} w_j^2$$

Logistic Regression

- Linear model + logistic sigmoid function: $y = \sigma(\boldsymbol{w}^T \boldsymbol{x}_i + b)$



$$P(y = 1|x) = 1/\{1 + \exp(- w^T x)\}$$

$$L(D; w) = \prod_{i=1}^{N} P(y = 1|x_i)^{y_i} \cdot (1 - p(y = 0|x_i)^{(1-y_i)}$$

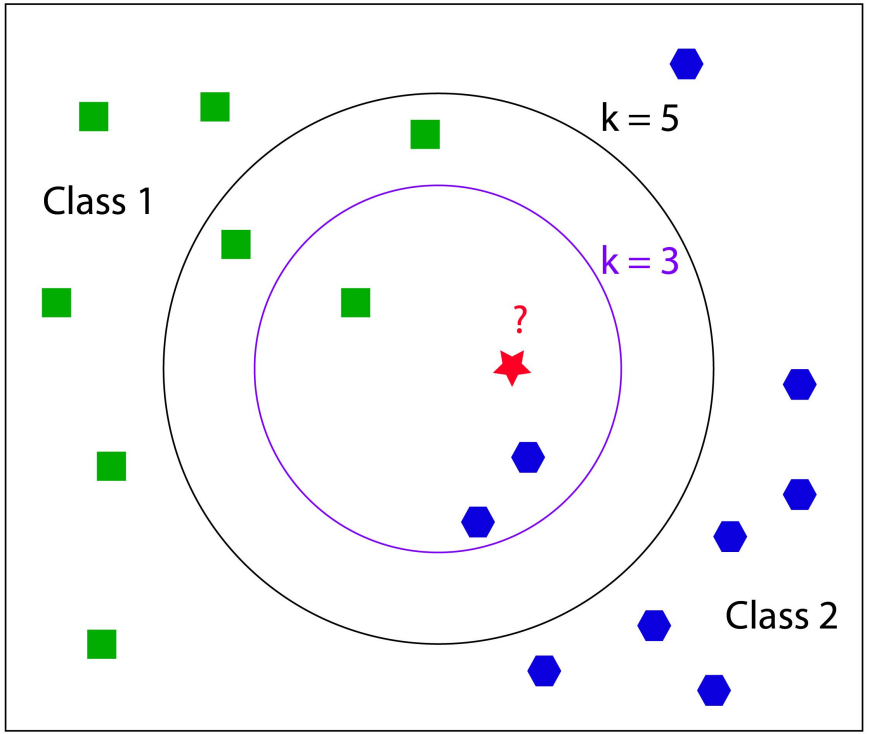
Easy extension to multi-class classification using softmax function

Model & Objective/Loss function

- 1) How to calculate a prediction (Model)
- 2) How to optimize the parameters (Objective function)

Other Basic ML Algorithms

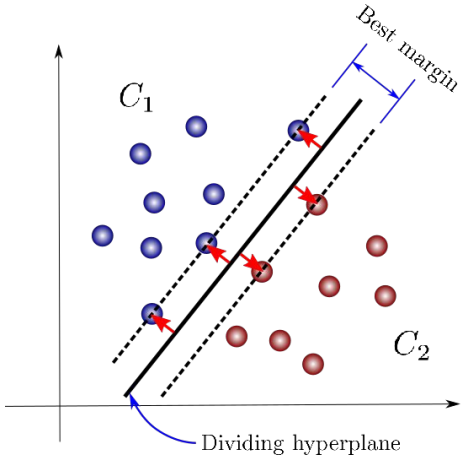
k-Nearest Neighbors (k-NN) algorithm

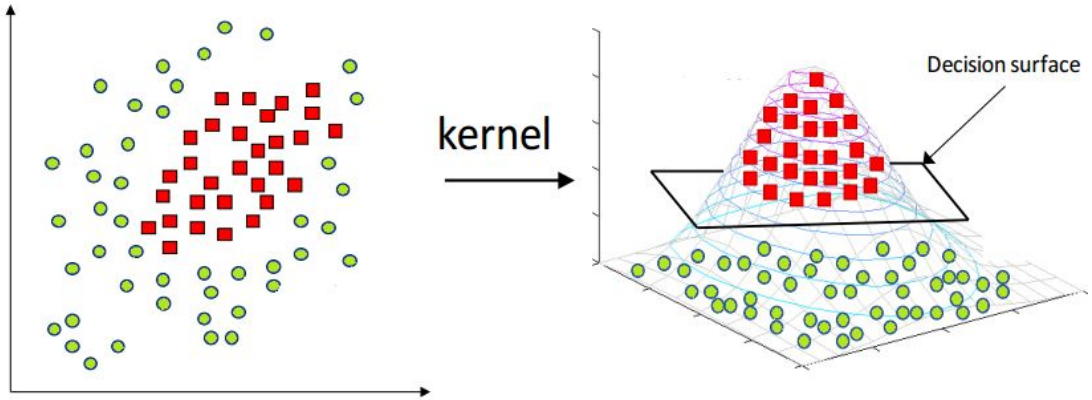


Sensitive to distance function

Support Vector Machine (SVM)

- Max-margin linear model + kernel trick
  - =~ As powerful as non-linear models
- Convex loss function → global optimum (cf. Multi-layer NN)
  - =~ As simple as linear models

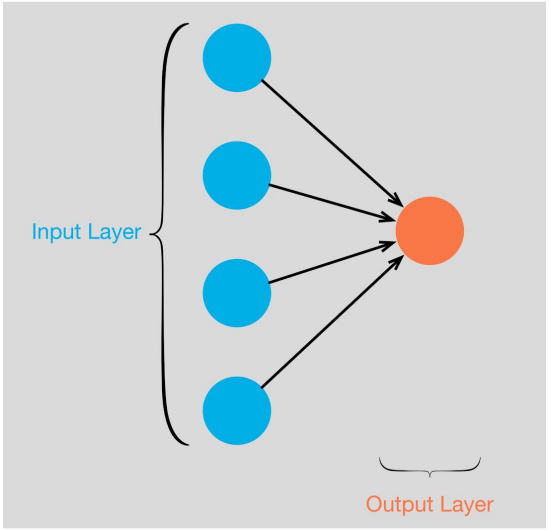


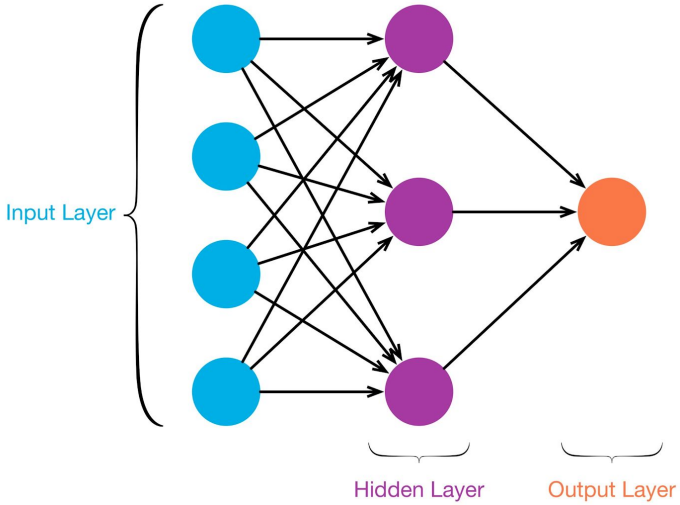


We will discuss Deep Learning in detail in Day 3

(Conventional) Neural Networks

- Node output = $\mathbf{a(w^T x + b)}$
- Training model parameters through backpropagation

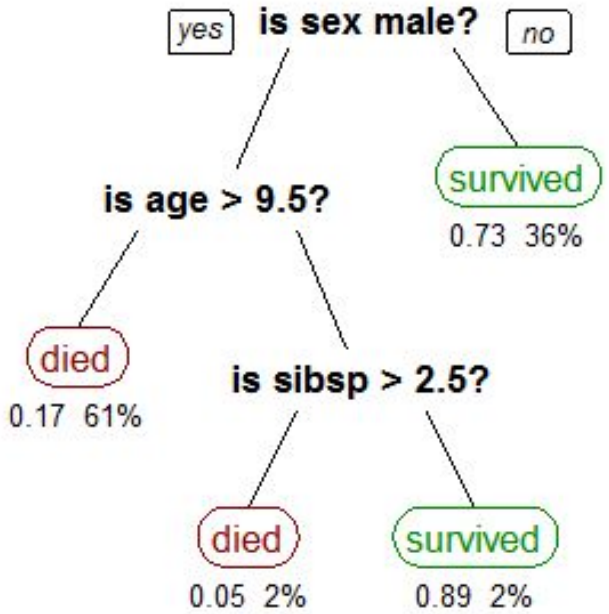




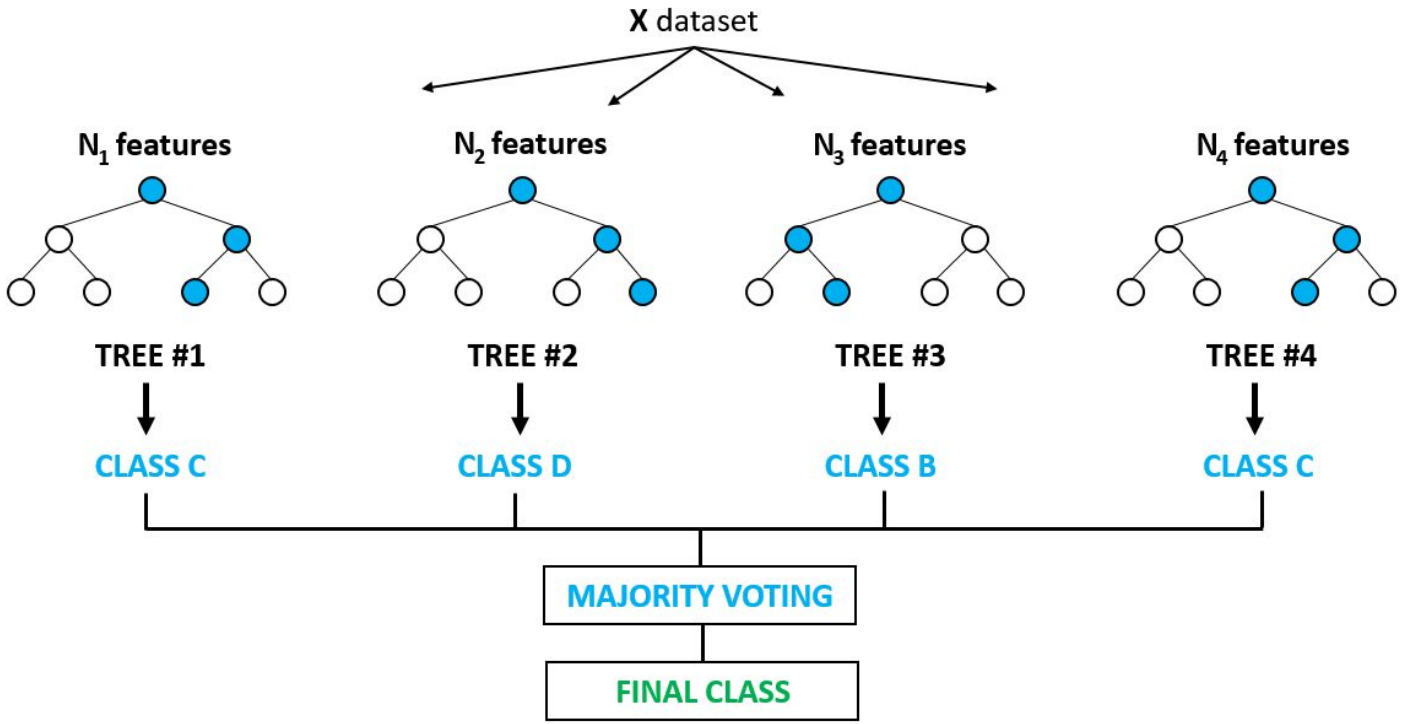
Gradient information

Decision Trees

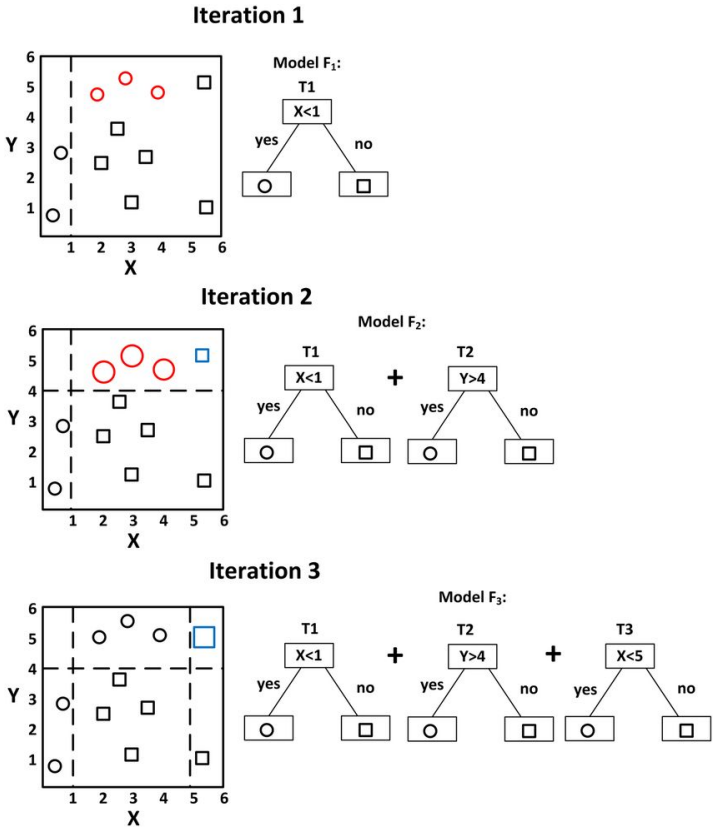
Titanic dataset example



Ensemble Models: Random Forest



Ensemble Models: Gradient Boosted Trees



What are differences? Should we care?

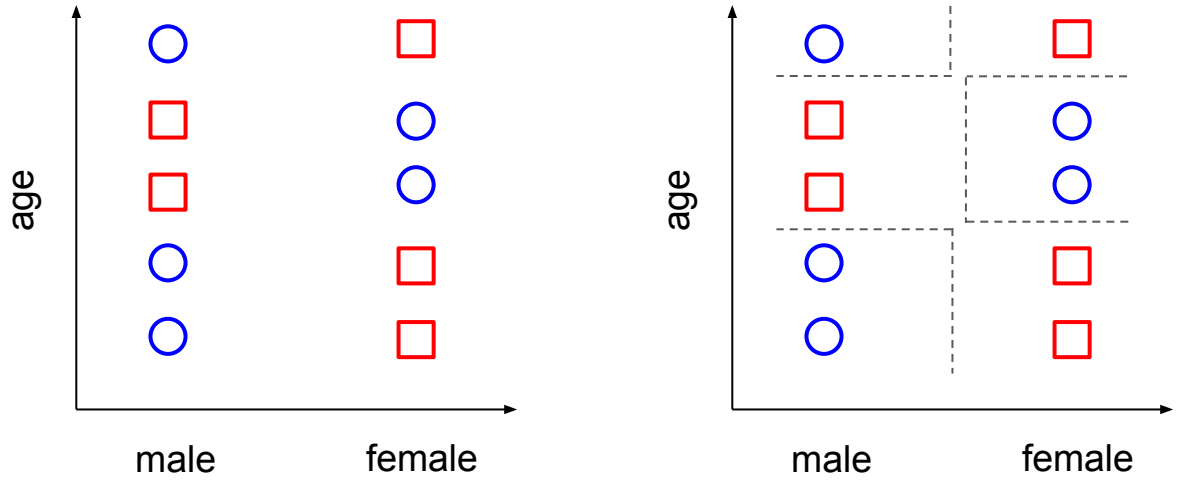
- Not Really
  - Easy to try out existing algorithms by changing just a few lines

- My rule of thumbs for classification
  - Logistic Regression
    - Will explain that LR is the simplest NN model in Day 3
  - Tree-based methods
    - I like Gradient Boosted Trees

Q. What are Tree-based methods good at? (compared to LR)

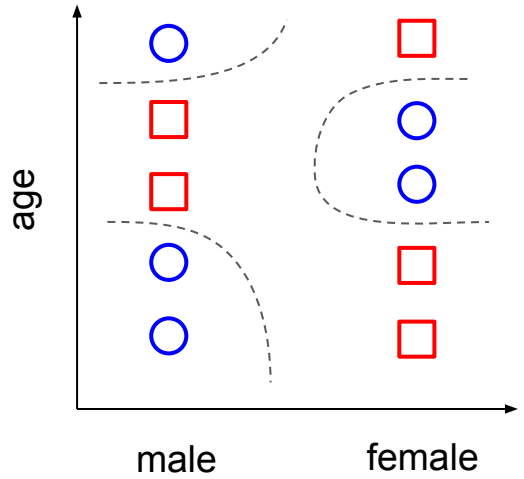
Q. What are Tree-based methods good at? (compared to LR)

- Region segmentation
  - Especially if features involve many categorical values

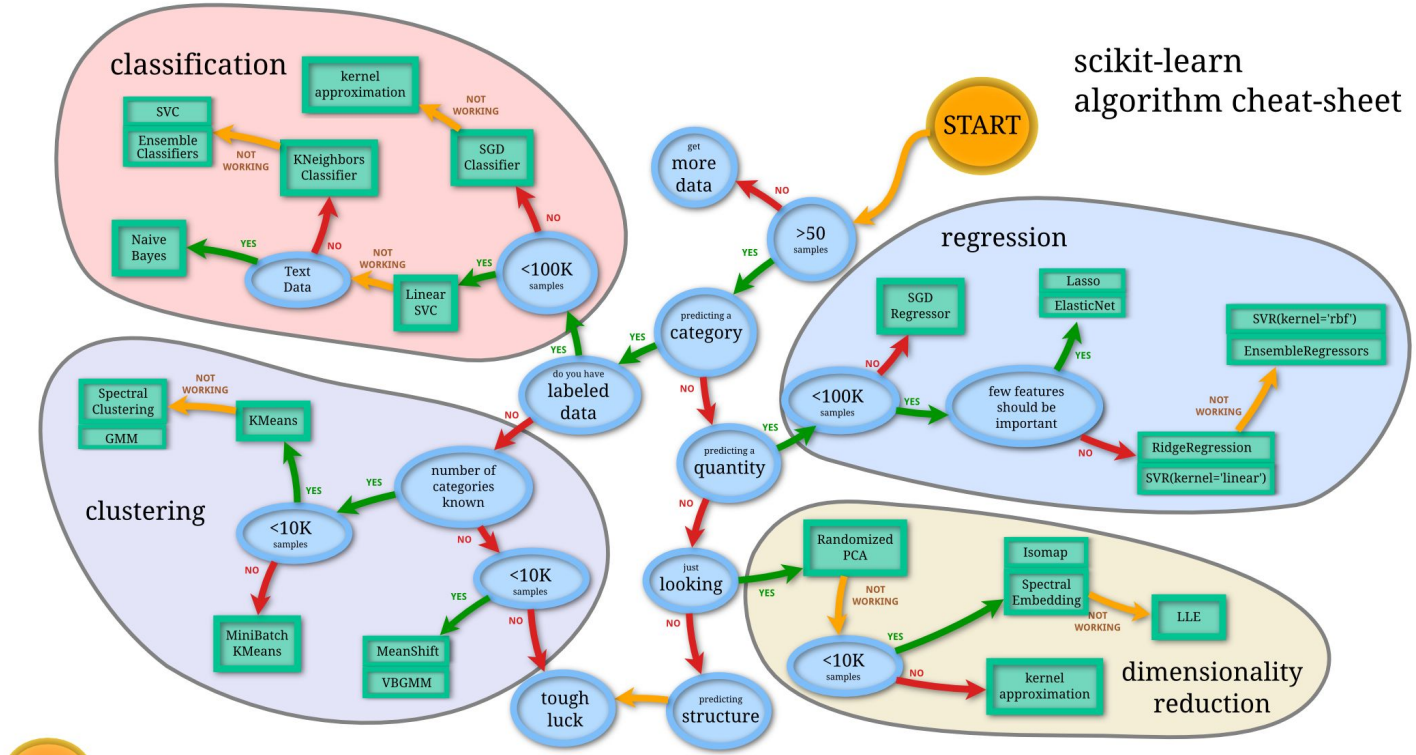


Q. What are Tree-based methods good at? (compared to LR)

- How about Deep Learning? Should it be able to learn such representations?
- → Yes. But, to me it's overkill. Why should we use many parameters to learn simple rules.

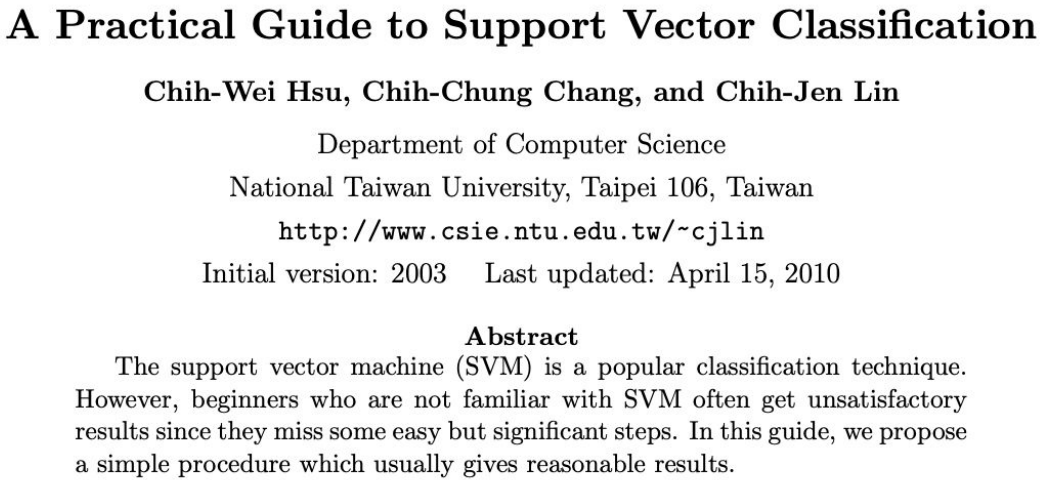


ML Cheat Sheet from Scikit-learn



Further Reading: A classic guide to ~~SVM~~ ML

- Feature scaling
- Two-stage hyper-parameter tuning
- Discussion on # Feature vs # instances



A Practical Guide to Support Vector Classification

Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin

Department of Computer Science

National Taiwan University, Taipei 106, Taiwan

http://www.csie.ntu.edu.tw/~cjlin

Initial version: 2003 Last updated: April 15, 2010

Abstract

The support vector machine (SVM) is a popular classification technique. However, beginners who are not familiar with SVM often get unsatisfactory results since they miss some easy but significant steps. In this guide, we propose a simple procedure which usually gives reasonable results.

Evaluation Metric

Accuracy

- # of correct prediction / # of total prediction

f

cat

non-cat

non-cat

= 2/3 = 66.6%

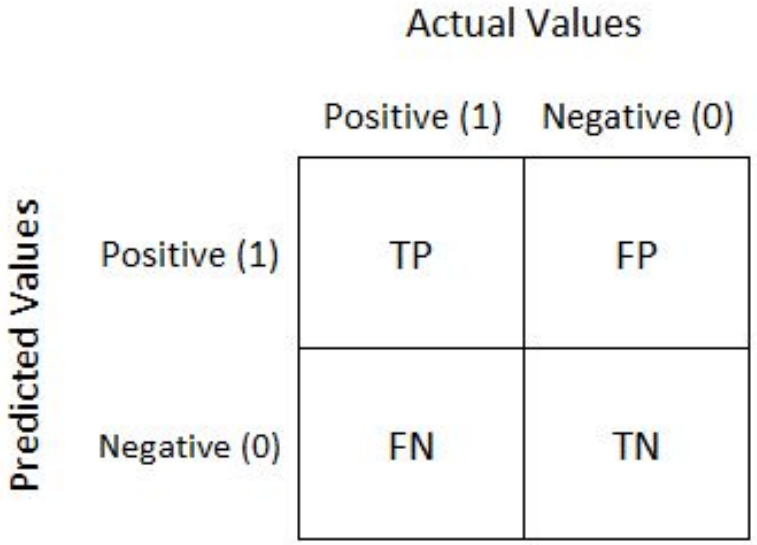
Quiz

- You got an accuracy value of 99% but something wrong seems to happen
- What do you think?

Random guessing can achieve 99%

- If the pos/neg ratio is 99:1

Precision, Recall, and F1 score



- Precision = TP / (TP + FP)
- Recall = TP (TP + FN)

- F1 = 2 P R / (P + R)
  - Harmonic Mean

Looking at Confusion Matrix is ALWAYS good idea

		True/Actual		
		Cat ()	Fish ()	Hen ()
Predicted	Cat ()	4	6	3
	Fish ()	1	2	0
	Hen ()	1	2	6

Evaluation Methods

Got 100% Accuracy! Yay!!

```
>>> clf = LogisticRegression()
>>> clf.fit(X, y)
>>> y_pred = clf.predict(X)
>>> accuracy_score(y, y_pred)
1.0
```

Can we publish a paper?

No. Unfortunately.

```
>>> clf = LogisticRegression()
>>> clf.fit(X, y)
>>> y_pred = clf.predict(X)
>>> accuracy_score(y, y_pred)
1.0
```

It's cheating

No. Unfortunately.

```
>>> clf = LogisticRegression()
>>> clf.fit(X, y)
>>> y_pred = clf.predict(X)
>>> accuracy_score(y, y_pred)
1.0
```

It's cheating

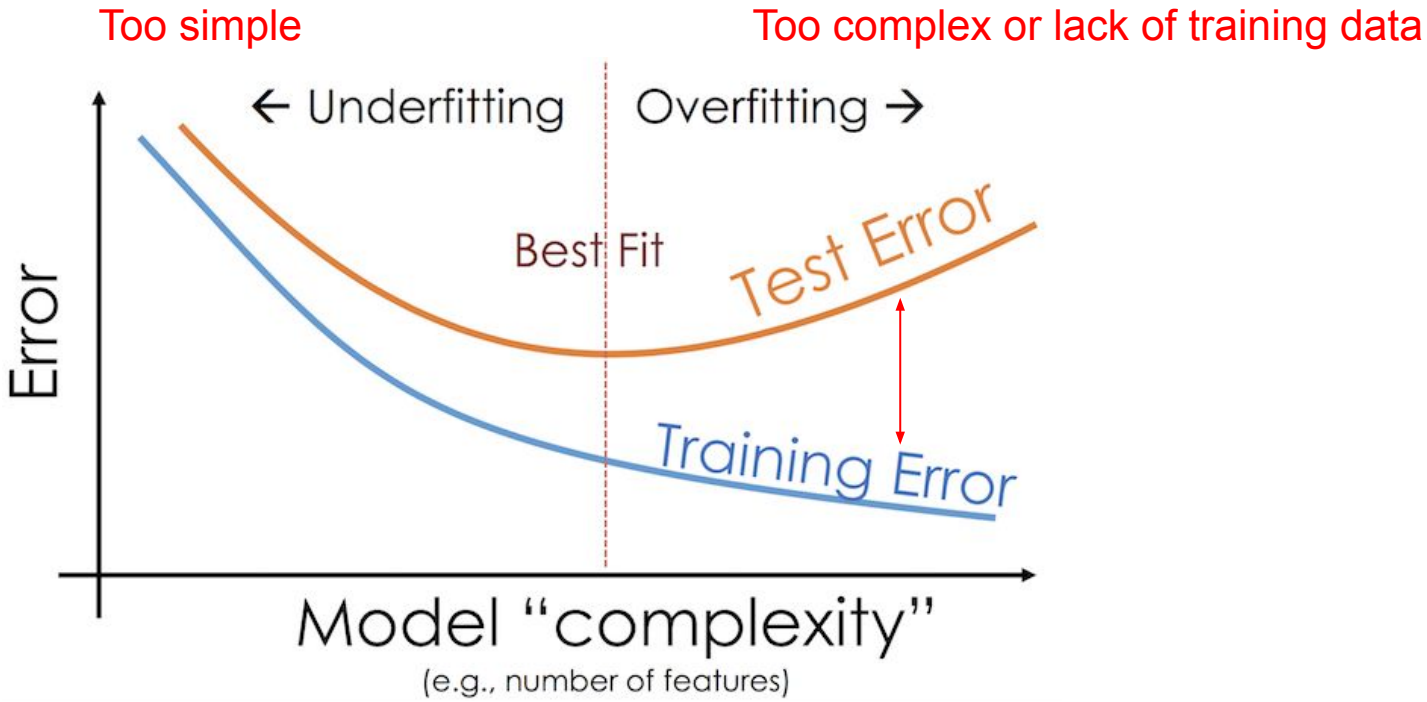
k-NN (k=1) model always achieves 100% accuracy on training data

We should create a holdout dataset for evaluation

- We can also explicitly create a validation set

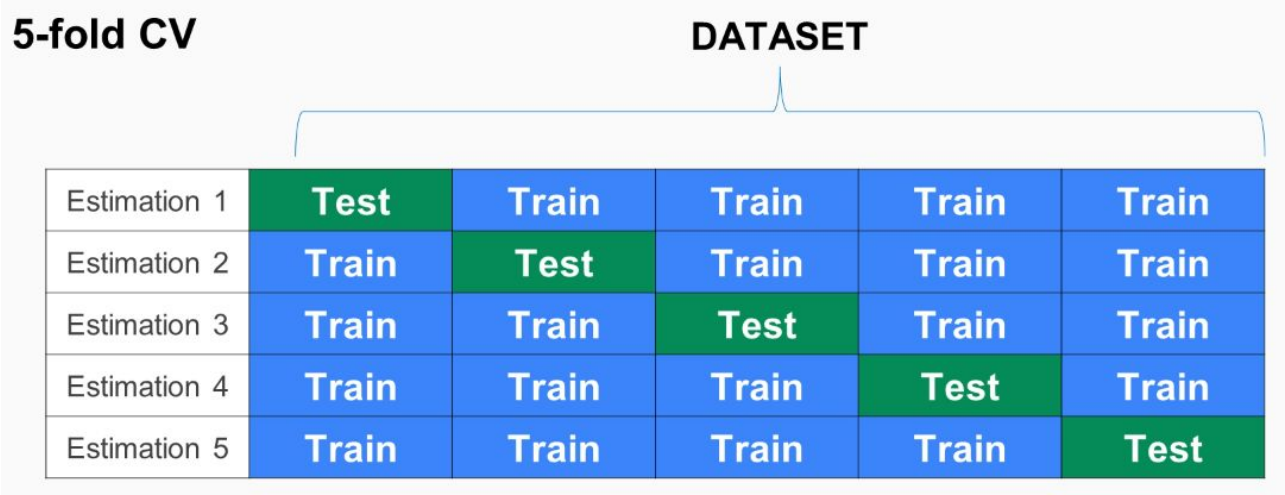
```
>>> from sklearn.model_selection import train_test_split
>>> X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                        test_size=0.33,
                                                        random_state=1)


>>> clf.fit(X_train, y_train)
>>> y_pred clf.predict(X_test)
>>> accuracy_score(y_test, y_pred)
0.83
```

Look at Training/Test Errors: Underfitting and Overfitting



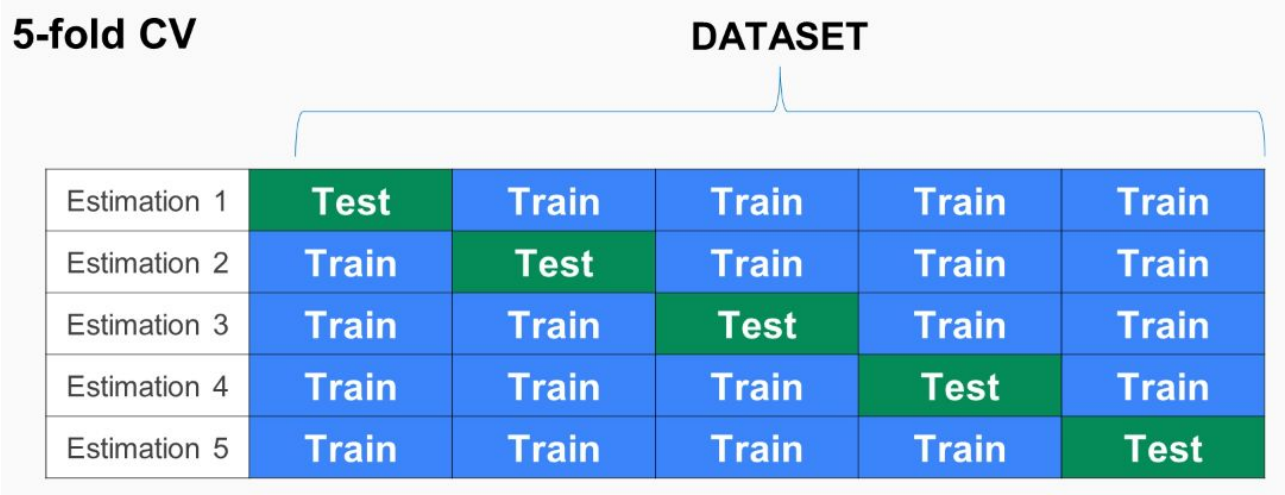
k-fold Cross-validation

- Iterate k different split for training and evaluation



k-fold Cross-validation

- Iterate k different split for training and evaluation



	Method A	Method B
	66.8	68.1
	65.4	68.3
	67.2	68.7
	70.1	68.5
	62.3	68.0
	66.4 (± 2.8)	68.3 (± 0.3)

Wait. What about Regression Analysis?

- In Regression Analysis, we evaluate a model based on goodness of fit, not generalization accuracy

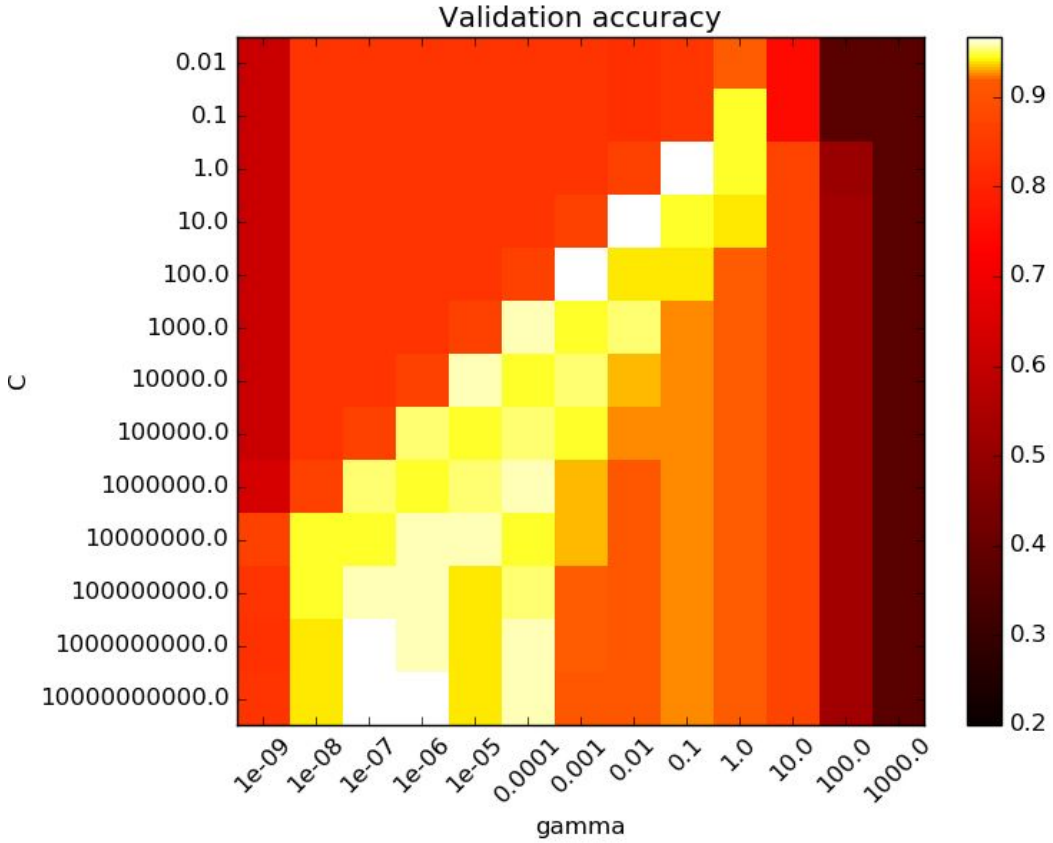
- This is a good borderline between machine learning and statistical models

Hyper-parameter Tuning

Hyper-parameter tuning

- Hyper-parameters are parameters that defines the training configuration and are NOT updated during the training process
  - For example,
    - Regularization coefficient C for Logistic Regression
    - # of trees, sub-sampling rate for Random Forest

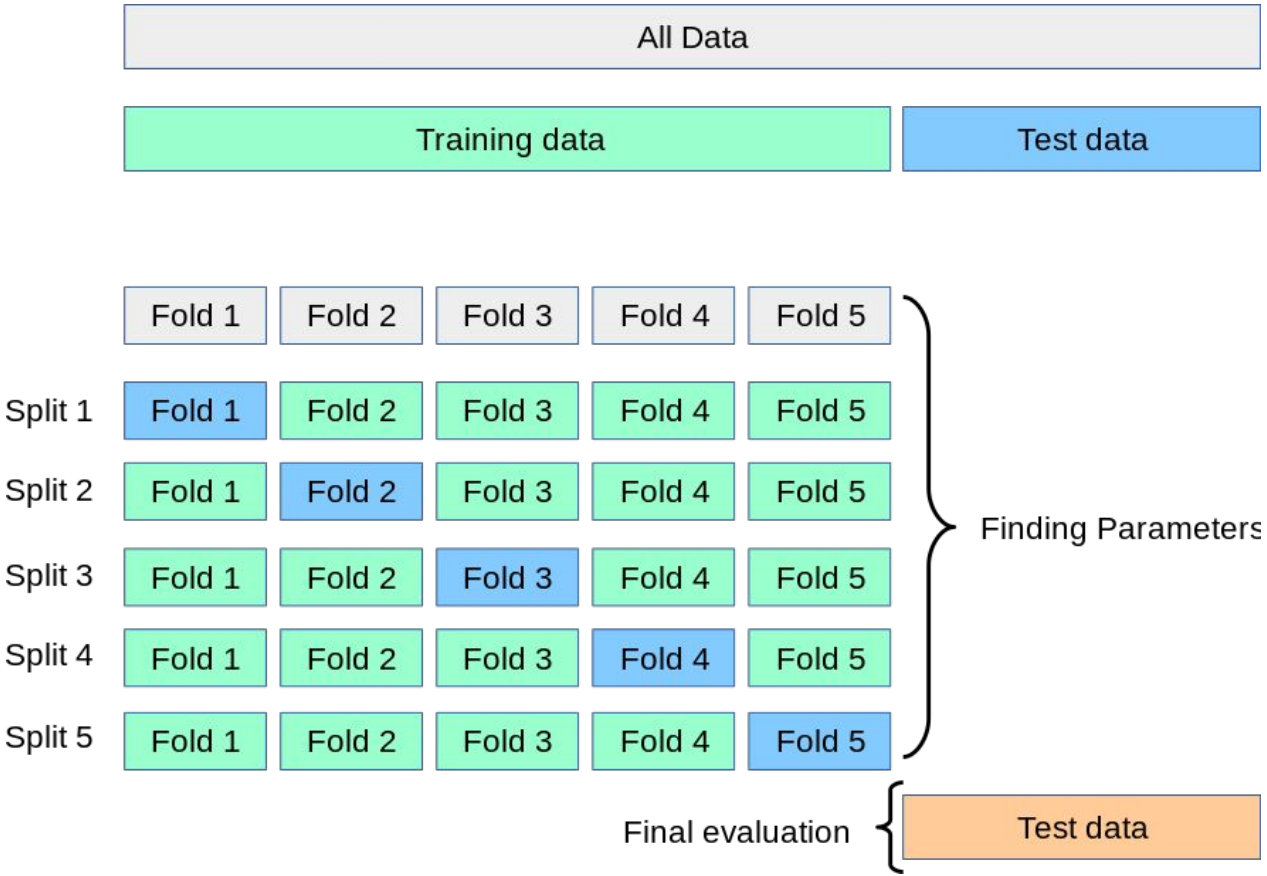
Grid Search


Validation accuracy
C
0.01
0.1
1.0
10.0
100.0
1000.0
10000.0
100000.0
1000000.0
10000000.0
100000000.0
1000000000.0
10000000000.0
gamma
1e-09
1e-08
1e-07
1e-06
1e-05
0.0001
0.001
0.01
0.1
1.0
10.0
100.0
1000.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

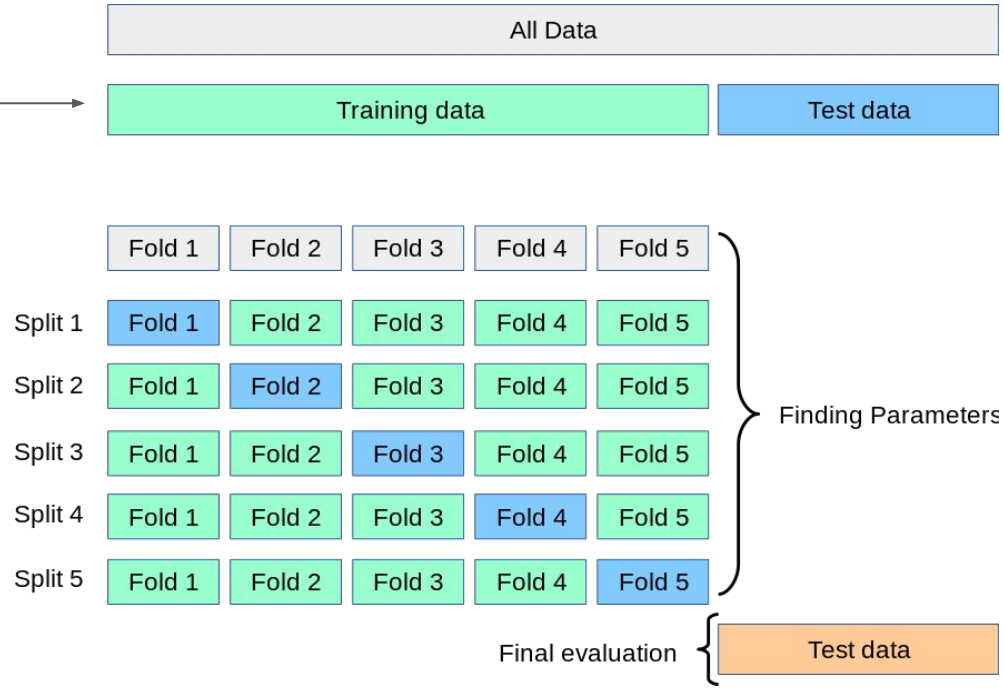
Grid Search: Scikit-learn Example

```
>>> from sklearn.model_selection import GridSearchCV
>>> clf = GridSearch(LogisticRegression(),
                     param_grid={"C": [0.001, 0.01, 0.1, 1.0]},
                     cv=5)
>>> clf.fit(X_train, y_train)
>>> y_pred = clf.predict(X_test)
>>> accuracy_score(y_test, y_pred)
0.87
```

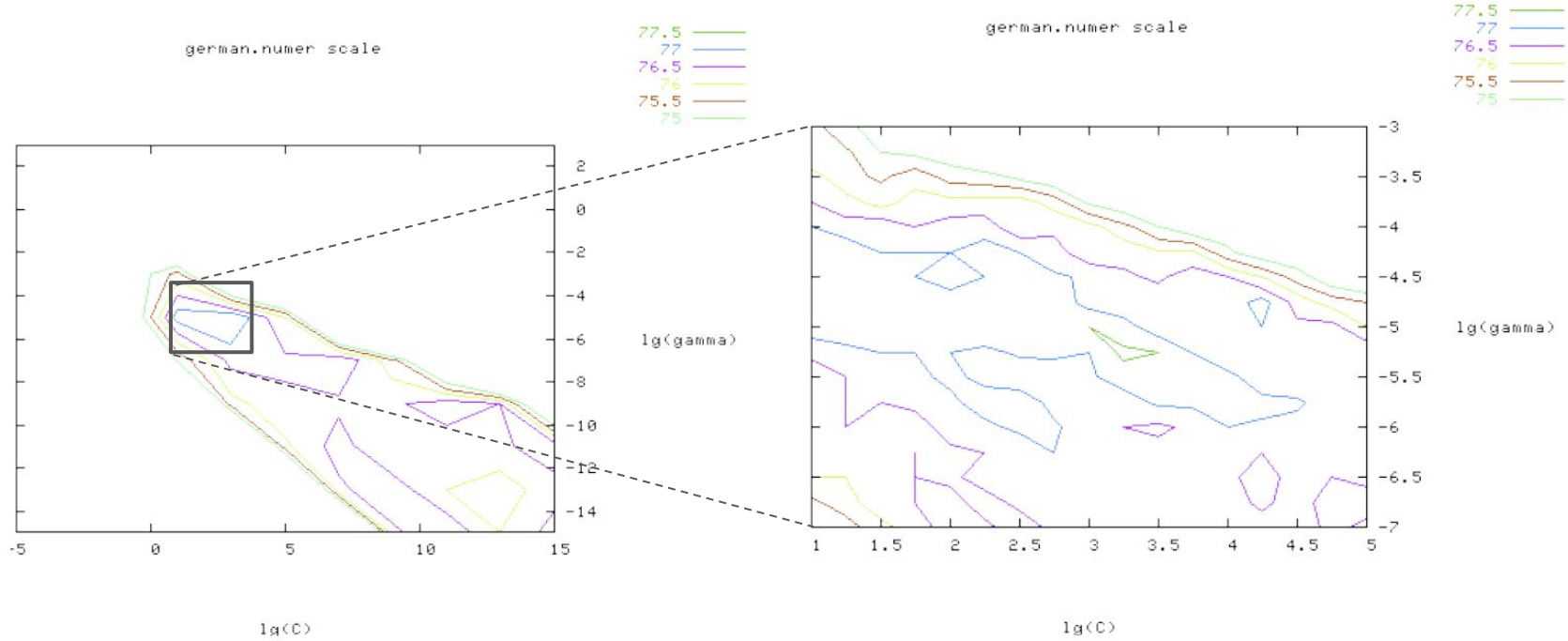
Cross-validation for Grid Search



cf. Cross-validation for Evaluation

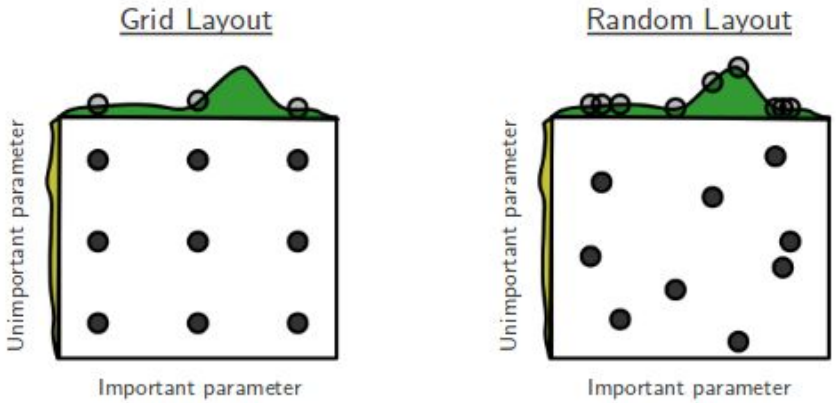


Loose Grid Search + Fine Grid Search



Grid Search vs Random Search

- Famous paper says Random Search >> Grid Search for NN models
- In my opinion, well-configured Grid Search works better than Random Search even for NNs



Prior knowledge vs (Fully) Data-driven

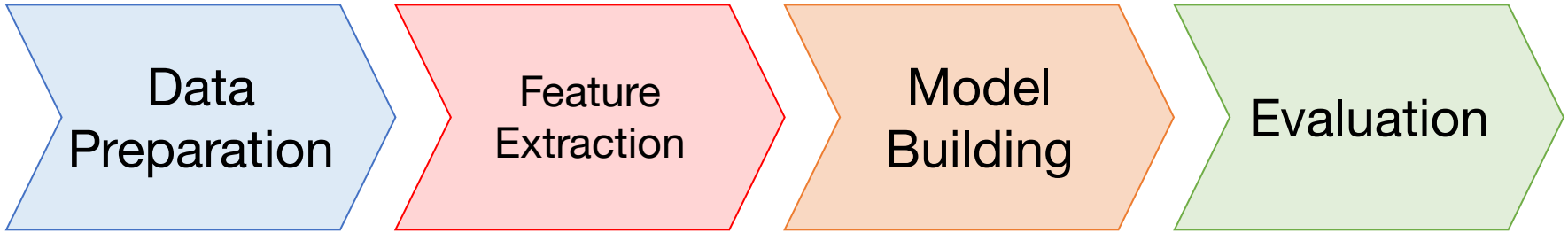
- How do we incorporate our prior knowledge into ML models?
- Many ways ...

Prior knowledge vs (Fully) Data-driven

- How do we incorporate our prior knowledge into ML models?
- Many ways ...
  - Data cleaning
  - Feature extraction
  - Algorithm selection
  - Hyper-parameter candidates for search
  - etc.

ML Pipeline

Machine Learning Pipeline



Data Preparation

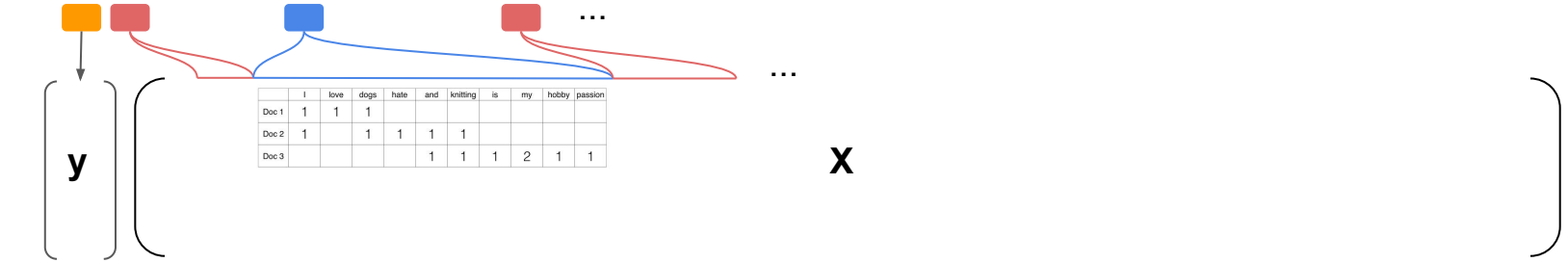
- Define target
- Data cleaning
  - Missing value imputation
  - Canonicalization

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S

Feature Extraction

- Use your domain knowledge!

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S

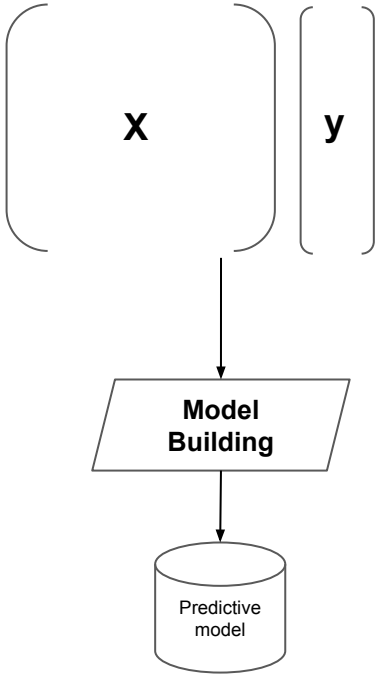


- Label extractor
- Category one-hot encoder
- Bag-of-words vectorizer
- ⋮

Model Building

- Training a model using feature matrix **X** and label vector **y**

```
>>> clf = LogisticRegression()
>>> clf.fit(X, y)
```

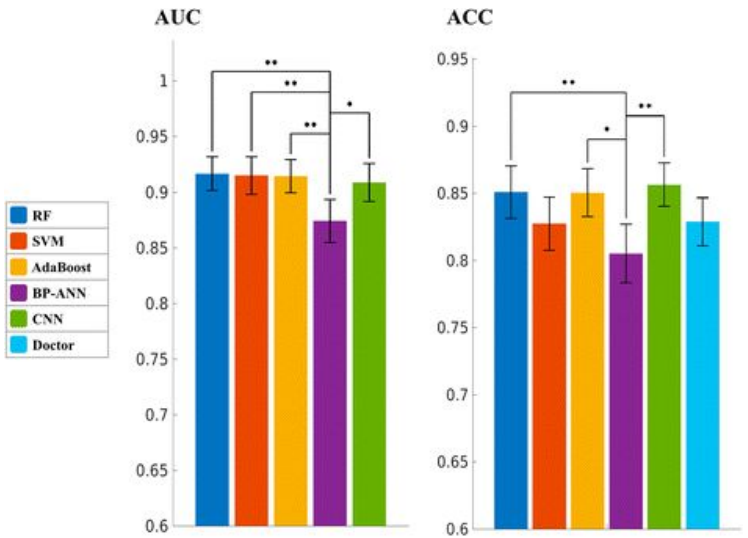


Evaluation

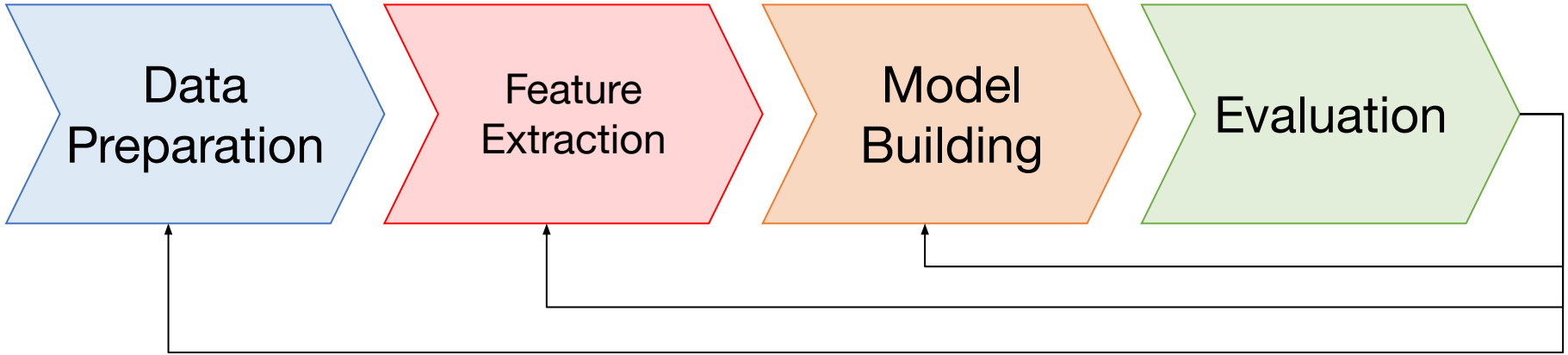
- Comparative evaluation and error analysis

5-fold CV — **DATASET**

Estimation 1	Test	Train	Train	Train	Train
Estimation 2	Train	Test	Train	Train	Train
Estimation 3	Train	Train	Test	Train	Train
Estimation 4	Train	Train	Train	Test	Train
Estimation 5	Train	Train	Train	Train	Test



ML Cycle


Data Preparation
Feature Extraction
Model Building
Evaluation

Break

Google Colab

HappyDB :)

Tech Policy

100,000 happy moments

What makes people happy? A huge database is making it possible to discern the answer at last.

by **Emerging Technology from the arXiv** Feb 5, 2018

HappyDB: A Corpus of 100,000 Crowdsourced Happy Moments

Akari Asai*, Sara Evensen†, Behzad Golshan‡, Alon Halevy‡, Vivian Li‡
Andrei Lopatenko‡, Daniela Stepanov‡, Yoshihiko Suhara‡, Wang-Chiew Tan‡, Yinzhan Xu†
*Univ. of Tokyo, †MIT, ‡Recruit Institute of Technology
akari-asai@g.ecc.u-tokyo.ac.jp, {sevensen, xyzhan}@mit.edu,
{behzad, alon, vivian, andrei, daniela, suharay, wangchiew}@recruit.ai

Abstract

The science of happiness is an area of positive psychology concerned with understanding what behaviors make people happy in a sustainable fashion. Recently, there has been interest in developing technologies that help incorporate the findings of the science of happiness into users' daily lives by steering them towards behaviors that increase happiness. With the goal of building technology that can understand how people express their happy moments in text, we crowd-sourced HappyDB, a corpus of 100,000 happy moments that we make publicly available. This paper describes HappyDB and its properties, and outlines several important NLP problems that can be studied with the help of the corpus. We also apply several state-of-the-art analysis techniques to analyze HappyDB. Our results demonstrate the need for deeper NLP techniques to be developed which makes HappyDB an exciting resource for follow-on research.

Keywords: science of happiness, positive psychology, happyDB corpus, crowdsourcing

What is HappyDB?

- HappyDB is a corpus of 100,000 crowd-sourced happy moments
- To advance the state of the art of understanding the causes of happiness that can be gleaned from text

Crowdsourcing

Question

- ${time} = 24 hours or 3 months

What made you happy today? Reflect on the past **${time}**, and recall three actual events that happened to you that made you happy. Write down your happy moment in a complete sentence.
(Write three such moments.)

Examples of Happy Moments

1. When I was on top of a hotel, looking at the city below me.

2. in the morning I received my college degree, receiving the title turn and behind me all my proud of my family was, for the goal that had just turned.

3. today was a school holiday for my son , woke up and played with him.

4. IT WAS VERY RELAXING TO COME HOME AFTER A LONG DAYS WORK.

5. The kitchen now gleams with new paint. Our annual renovation is over and all the colors we chose are set for at least a year. I love our new colors.

Happiness Category

Category	Definition	Examples
Achievement	With extra effort to achieve a better than expected result	Finish work. Complete marathon.
Affection	Meaningful interaction with family, loved ones and pets	Hug. Cuddle. Kiss.
Bonding	Meaningful interaction with friends and colleagues	Have meals w coworker. Meet with friends.
Enjoy the moment	Being aware or reflecting on present environment	Have a good time. Mesmerize.
Exercise	With intent to exercise or workout	Run. Bike. Do yoga. Lift weights.
Leisure	An activity done regularly in one's free time for pleasure	Play games. Watch movie. Bake cookies.
Nature	In the open air, in nature	Garden. Beach. Sunset. Weather