Introduction to Machine Learning
NPFL 054

http://ufal.mff.cuni.cz/course/npfl054

Barbora Hladká
hladka@ufal.mff.cuni.cz

Martin Holub
holub@ufal.mff.cuni.cz

Charles University in Prague,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Lecture 1 — Introduction to Machine Learning

Outline

- Motivation examples
- Supervised Machine Learning
- Searching for a good hypothesis
- Development cycle
- Brief overview of the course
- Examination requirements

Motivation example

Word-sense disambiguation (WSD)
Assign the correct sense of a word in a sentence.
Let's work with the word *line*:

- I've got Inspector Jackson on the **line** for you.
- Outside, a **line** of customers waited to get in.
- He quoted a few **lines** from Shakespeare.
- He didn't catch many fish, but it hardly mattered.
  With his **line** out, he sat for hours staring at the Atlantic.
- ...

Motivation example

Word-sense disambiguation
Assign the correct sense of a word in a sentence.
Let's work with the word *line* and its following senses:

- `CORD`
- `DIVISION`
- `FORMATION`
- `PHONE`
- `PRODUCT`
- `TEXT`

Motivation example — Word-sense disambiguation

?CORD ?DIVISION ?FORMATION ?PHONE ?PRODUCT ?TEXT

- I've got Inspector Jackson on the **line** for you. `PHONE`
- Outside, a **line** of customers waited to get in. `FORMATION`
- He quoted a few **lines** from Shakespeare. `TEXT`
- He didn't catch many fish, but it hardly mattered.
  With his **line** out, he sat for hours staring at the Atlantic. `CORD`
- The company has just launched a new **line** of small,
  low-priced computers. `PRODUCT`
- Draw a **line** that passes through the points P and Q. `DIVISION`
- This has been a very popular new **line**. `PRODUCT? FORMATION?`

Motivation example

Word-sense disambiguation

- What knowledge do you use to assign the senses?
- What are the keys for the correct decision?

Motivation example

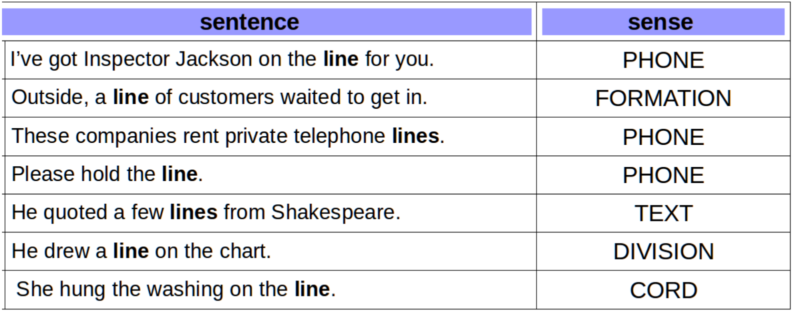
- We – human beings – do word sense disambiguation easily using the **context in the sentence** and our **knowledge of the world**.
- We want computers to master it as well.

Let's prepare examples and guide computers to learn from them.

That is Machine Learning!

Machine learning

Intuitively we need a large set of recognized **examples** to learn the essential knowledge necessary to recognize correct output values. Examples used for learning are called **training data**.



sentence	sense
I've got Inspector Jackson on the **line** for you.	PHONE
Outside, a **line** of customers waited to get in.	FORMATION
These companies rent private telephone **lines**.	PHONE
Please hold the **line**.	PHONE
He quoted a few **lines** from Shakespeare.	TEXT
He drew a **line** on the chart.	DIVISION
She hung the washing on the **line**.	CORD

What computers extract from examples

In the WSD task, both humans and computers need to know the **context of the target word** (“line”) to recognize correct senses.

Humans use their reason, intuition, and their real world knowledge.

Computers need to extract a limited set of useful **context clues** that are then used for automatic decision about the correct sense.

- Formally, the context clues are called **attributes or features** and should be exactly and explicitly defined.
- Then each object (e.g. a sentence) is characterized by a list of features, which is called **feature vector**.

Computer makes feature vectors from examples.

Intuitive feature extraction – examples

To choose an effective set of features we always need our intuition. Only then all experiments with data can start.

A few example hints:

class	**a feature to recognize the class – will be useful?**
`CORD`	immediately preceding word
`FORMATION`	immediately following word
`PHONE`	can be often recognized by characteristic verbs

“Examples” in ML – two meanings

1) Real examples – Each real object that is already recognized or that we want to recognize is an example.

2) Data instances – In ML, each real example is represented as a data instance. In this sense

$$\textbf{example} = \textbf{feature vector} + \textbf{output value}$$

Data instances

Sometimes we do not know the output value; in this case data instances are not different from feature vectors.

data instance = feature vector (+ output value, if it is known)

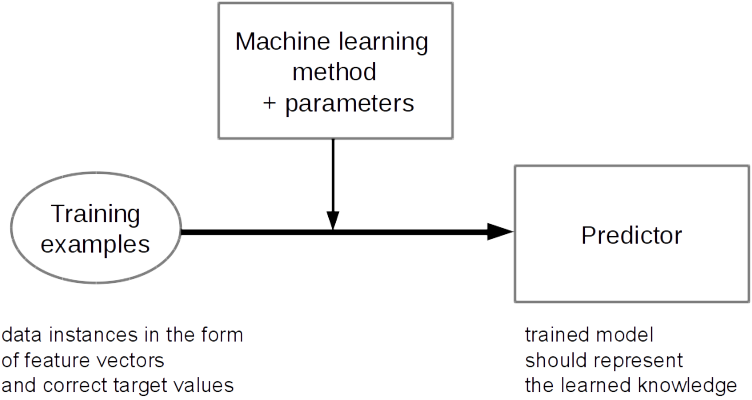
A data instance is either a feature vector or a complete example.

Formal definition of ML by Mitchell (1997)

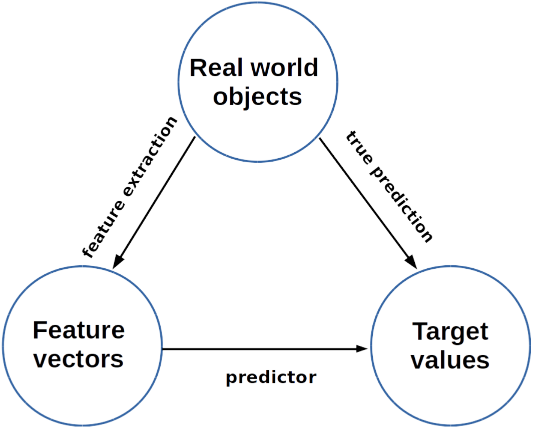
A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Supervised learning process

Supervised Machine Learning = computer learns "essential knowledge" extracted from a (large) set of examples


Machine learning method + parameters
Training examples
Predictor
data instances in the form of feature vectors and correct target values
trained model should represent the learned knowledge

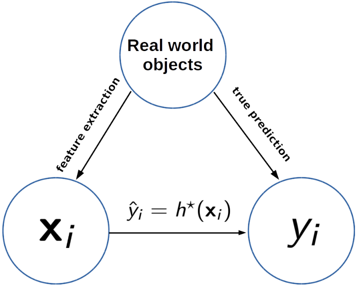
Machine learning as building a prediction function



- if target values are *continuous* numbers, we speak about **regression**
  = estimating or predicting a continuous response
- if target values are *discrete/categorical*, we speak about **classification**
  = identifying group membership

Prediction function and its relation to the data

Idealized model of supervised learning



- $\mathbf{x}_i$ are **feature vectors**, y_i are true **predictions**
- **prediction function** $h^\star$ is the "best" of all possible hypotheses h
- **learning process** is searching for $h^\star$, which means to search the **hypothesis space** and minimize a predefined **loss function**
- ideally, the learning process results in $h^\star$ so that predicted $\hat{y}_i = h^\star(\mathbf{x}_i)$ is equal to the true target values y_i

Loss function

A loss function $\mathrm{L}(\hat{y}, y)$ measures the cost of predicting $\hat{y}$ when the true value is y. Commonly used loss functions are

- squared loss $\mathrm{L}(\hat{y}, y) = (\hat{y} - y)^2$
  for regression
- zero-one loss $\mathrm{L}(\hat{y}, y) = \mathrm{I}(\hat{y} \neq y)$
  for classifiation; *indicator variable* I is 1 if $\hat{y} \neq y$, 0 otherwise

The goal of learning can be stated as producing a model with the smallest possible loss; i.e., a model that minimizes the average $\mathrm{L}(\hat{y}, y)$ over all examples.

Important notes

- Loss function is sometimes also known as "cost function".
- In a broader sense, loss function means the value that summarizes the loss over a sample of examples, e.g. $\sum \mathrm{L}(\hat{y}, y)$ or $\mathsf{E}\,[\mathrm{L}(\hat{y}, y)]$.
- A more general term is "objective function", which is sometimes used for the function that should be optimized (minimized or maximized); yes, typically the objective function is in fact the loss function computed over a sample of development test examples.

Training data vs. test data

- **Training data** = a set of examples
  – used for **learning process**
- **Test data** = another set of examples
  – used for **evaluation** of a trained model
- **Important**: the split of all available examples into the training and the test portions should be **random**!

Supervised ML task and data instances

Supervised machine learning necessarily requires learning examples

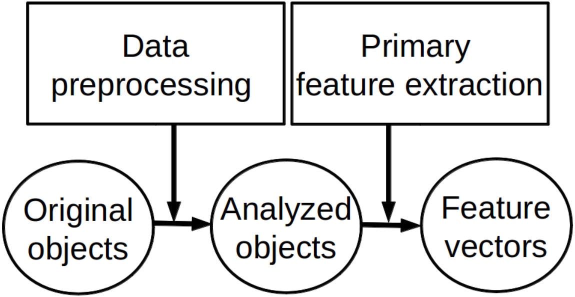
- **Features** are properties of examples that can be observed or measured
  – are numerical (discrete or continuous), or categorical (incl. binary)
- **Feature vector** is an ordered list of selected features
- **Data instance** = feature vector (+ target class, if it is known)
- **Training data** = a set of examples used for **learning process**
- **Test data** = another set of examples used for **evaluation**

Terminology – features and target values

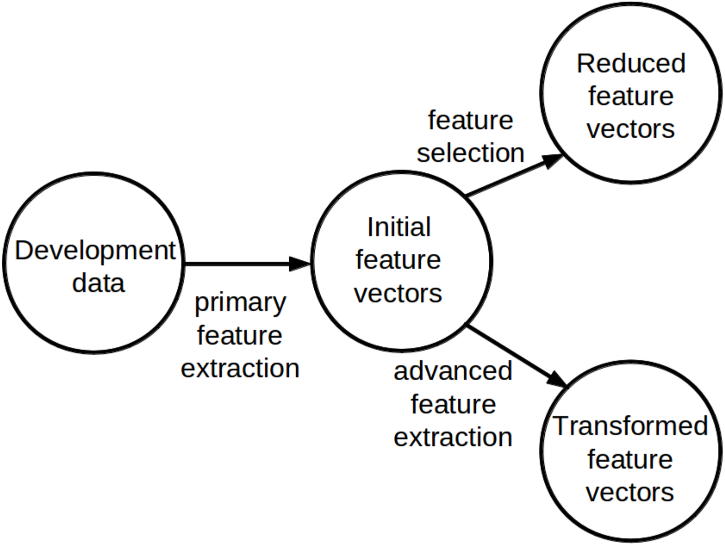
- **How different people call values that describe objects**

	observed (known) object characteristics	**values or categories to be predicted**
computer scientists	**features**	**(target) value or class**
mathematicians (statisticians)	attributes or predictors	response (value) or output value

Data preprocessing and feature extraction



Feature extraction and feature selection


Development data
primary feature extraction
Initial feature vectors
feature selection
Reduced feature vectors
advanced feature extraction
Transformed feature vectors

Sample error and generalization error

Sample error of a hypothesis h with respect to a data sample S of the size n is usually measured as follows

- for **regression**: **mean squared error** $\mathrm{MSE} = \dfrac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$
- for **classification**: **classification error** $= \dfrac{1}{n} \sum_{i=1}^{n} \mathrm{I}(\hat{y}_i \neq y_i)$

Generalization error (aka "true error" or "expected error") measures how well a hypothesis h generalizes beyond the used training data set, to unseen data with distribution $\mathcal{D}$. Usually it is defined as follows

- for **regression**: $\mathrm{error}_{\mathcal{D}}(h) = \mathsf{E}\,(\hat{y}_i - y_i)^2$
- for **classification**: $\mathrm{error}_{\mathcal{D}}(h) = \mathsf{Pr}\,(\hat{y}_i \neq y_i)$

Accuracy and error rate

To measure the performance of classification tasks we often use (sample) *accuracy* and (sample) *error rate*

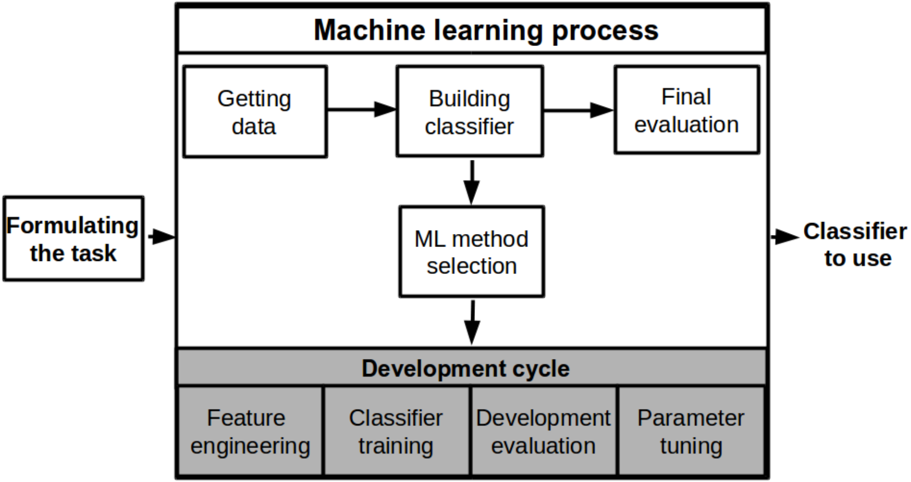
Sample accuracy is the number of correctly predicted examples divided by the number of all examples in the predicted set

Sample error rate is equal to **1 - accuracy**

Training error rate is the sample error rate measured on the training data set

Test error rate is the sample error rate measured on the test data set

Machine learning process — development cycle



Formulating the task

Machine learning process

Getting data

Building classifier

Final evaluation

ML method selection

Development cycle

Feature engineering

Classifier training

Development evaluation

Parameter tuning

Classifier to use

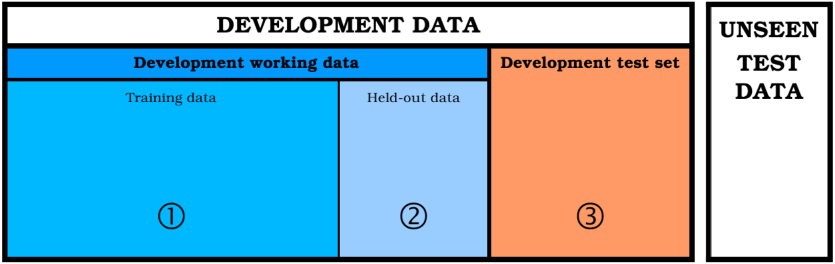
Terminological note on building predictors

The purpose of the learning process is search for the best prediction function parameters

learning parameters	hypothesis parameters
= parameters of the learning process	= parameters of the prediction function

- **Method** = approach/principle to learning. i.e. to building predictors
- **Model** = method + set of features + learning parameters
- **Predictor** = trained model, i.e. an output of the machine learning process, i.e. a particular method trained on a particular training data.
- **Prediction function** = predictor (used in mathematics). It's a function calculating a response value using "predictor variables".
- **Hypothesis** = prediction function – not necessarily the best one (used in theory of machine learning).

Development data and its division



DEVELOPMENT DATA

Development working data

Training data ①

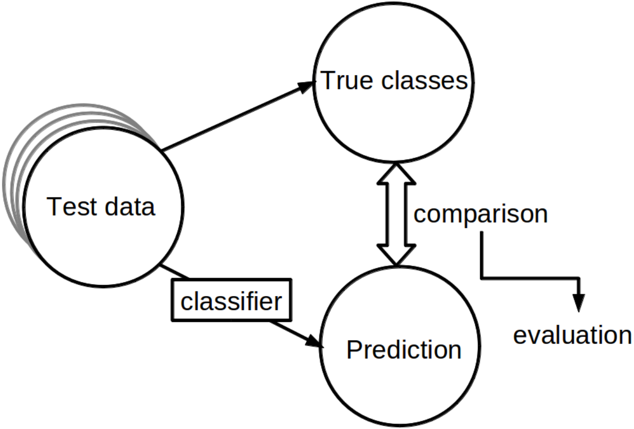
Held-out data ②

Development test set ③

UNSEEN TEST DATA

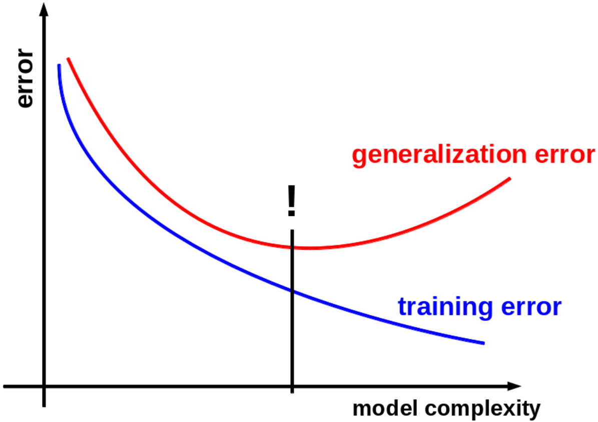
All subsets should be selected randomly in order to represent the characteristic distribution of both feature values and target values in the available set of examples.

Evaluation – basic scheme



Minimizing generalization error

Finding a model that minimizes generalization error
... is one of central goals of the machine learning process

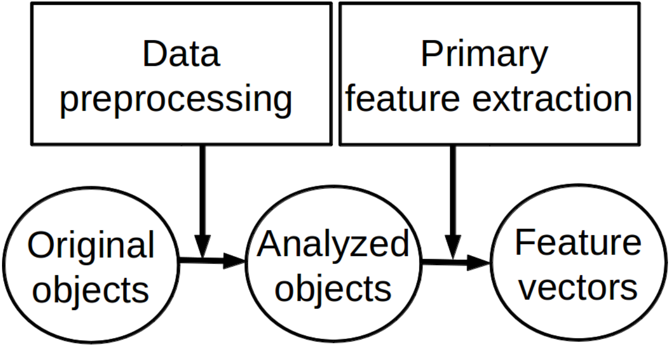


Formulating the task

1. **Task description**
   WSD: Assign the correct sense to the target word "line"
2. **Object specification**
   WSD: Sentences containing the target word
3. **Specification of desired output Y**
   WSD: $Y = \texttt{SENSE}$
   SENSE = {CORD,DIVISION,FORMATION,PHONE,PRODUCT,TEXT}

Getting data

Step 1: Getting feature vectors



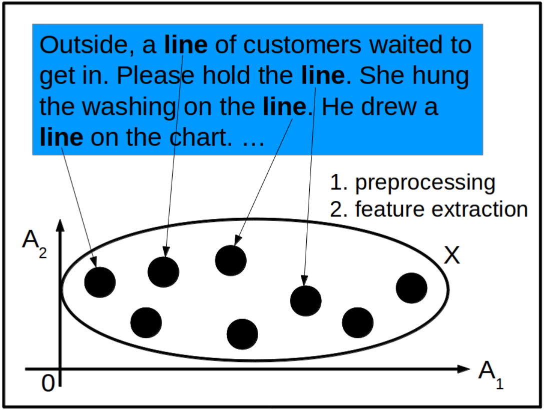
Getting data

Step 1: Getting feature vectors

- Features as variables A_1, ..., A_m
  - **numerical**
    – either discrete or continuous
  - **categorical**
    – any list of discrete values, non-numerical
  - **binary** (0/1, True/False, Yes/No)
    – can be viewed as a kind of categorical
- Feature values $x_1, ..., x_m$, $x_i \in A_i$
- Each object represented as feature vector $\mathbf{x} = \langle x_1, ..., x_m \rangle$
- Feature vectors are elements in an m-dimensional feature space
- Set of instances $X = \{\mathbf{x} : \mathbf{x} = \langle x_1, ..., x_m \rangle, x_i \in A_i\}$.

Getting data

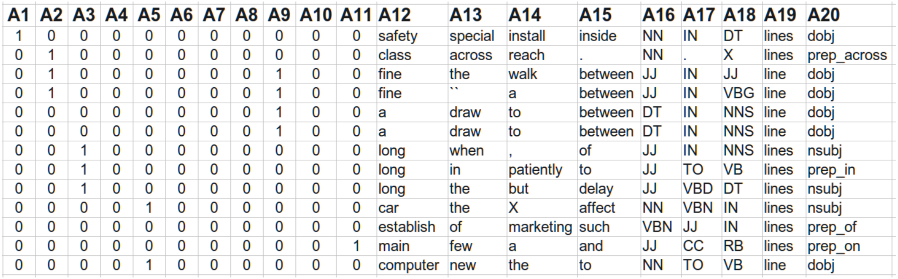
Step 1: Getting feature vectors – Example



Outside, a **line** of customers waited to get in. Please hold the **line**. She hung the washing on the **line**. He drew a **line** on the chart. …

1. preprocessing
2. feature extraction

Example feature vectors – the WSD task



A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20
1	0	0	0	0	0	0	0	0	0	0	safety	special	install	inside	NN	IN	DT	lines	dobj
0	1	0	0	0	0	0	0	0	0	0	class	across	reach	.	NN	.	X	lines	prep_across
0	1	0	0	0	0	0	0	1	0	0	fine	the	walk	between	JJ	IN	JJ	line	dobj
0	1	0	0	0	0	0	0	1	0	0	fine	``	a	between	JJ	IN	VBG	line	dobj
0	0	0	0	0	0	0	0	1	0	0	a	draw	to	between	DT	IN	NNS	line	dobj
0	0	0	0	0	0	0	0	1	0	0	a	draw	to	between	DT	IN	NNS	line	dobj
0	0	1	0	0	0	0	0	0	0	0	long	when	,	of	JJ	IN	NNS	lines	nsubj
0	0	1	0	0	0	0	0	0	0	0	long	in	patiently	to	JJ	TO	VB	lines	prep_in
0	0	1	0	0	0	0	0	0	0	0	long	the	but	delay	JJ	VBD	DT	lines	nsubj
0	0	0	0	1	0	0	0	0	0	0	car	the	X	affect	NN	VBN	IN	lines	nsubj
0	0	0	0	0	0	0	0	0	0	0	establish	of	marketing	such	VBN	JJ	IN	lines	prep_of
0	0	0	0	0	0	0	0	0	0	1	main	few	a	and	JJ	CC	RB	lines	prep_on
0	0	0	0	1	0	0	0	0	0	0	computer	new	the	to	NN	TO	VB	line	dobj

See the feature description `wsd.attributes.pdf` at
`https://ufal.mff.cuni.cz/course/npfl054/materials`

Getting data

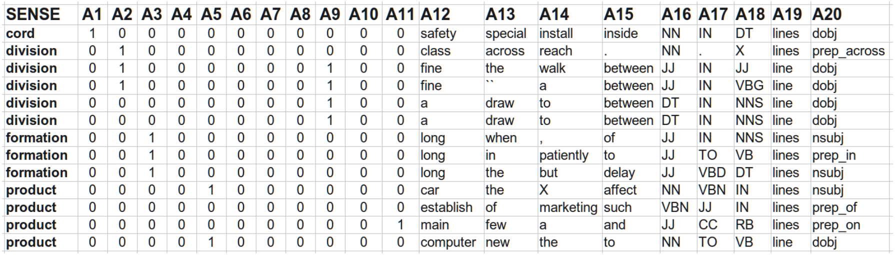
Step 2: Assigning true prediction

- Take *a number* of original objects and assign true prediction to each of them, e.g. **do manual annotation**.
- Take these objects and their true prediction, do preprocessing and feature extraction. It results in **Gold Standard Data** $Data = \{\langle \mathbf{x}, y \rangle : \mathbf{x} \in X, y \in Y\}$.

Getting data

Step 2: Assigning true prediction

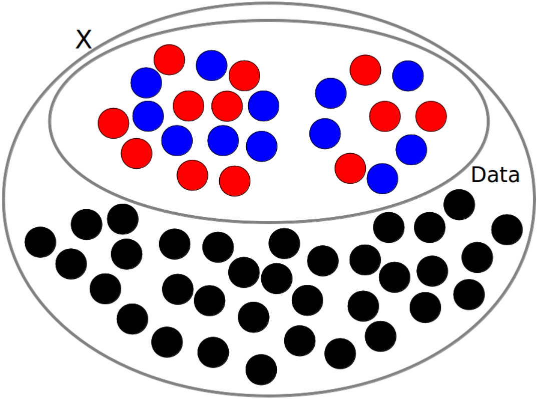
Example: $Y = \texttt{SENSE} = \{\texttt{CORD},\texttt{DIVISION},\texttt{FORMATION},\texttt{PHONE},\texttt{PRODUCT},\texttt{TEXT}\}$



SENSE	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20
cord	1	0	0	0	0	0	0	0	0	0	0	safety	special	install	inside	NN	IN	DT	lines	dobj
division	0	1	0	0	0	0	0	0	0	0	0	class	across	reach	.	NN	.	X	lines	prep_across
division	0	1	0	0	0	0	0	0	1	0	0	fine	the	walk	between	JJ	IN	JJ	line	dobj
division	0	1	0	0	0	0	0	0	1	0	0	fine	``	a	between	JJ	IN	VBG	line	dobj
division	0	0	0	0	0	0	0	0	1	0	0	a	draw	to	between	DT	IN	NNS	line	dobj
division	0	0	0	0	0	0	0	0	1	0	0	a	draw	to	between	DT	IN	NNS	line	dobj
formation	0	0	1	0	0	0	0	0	0	0	0	long	when	,	of	JJ	IN	NNS	lines	nsubj
formation	0	0	1	0	0	0	0	0	0	0	0	long	in	patiently	to	JJ	TO	VB	lines	prep_in
formation	0	0	1	0	0	0	0	0	0	0	0	long	the	but	delay	JJ	VBD	DT	lines	nsubj
product	0	0	0	0	1	0	0	0	0	0	0	car	the	X	affect	NN	VBN	IN	lines	nsubj
product	0	0	0	0	0	0	0	0	0	0	0	establish	of	marketing	such	VBN	JJ	IN	lines	prep_of
product	0	0	0	0	0	0	0	0	0	0	1	main	few	a	and	JJ	CC	RB	lines	prep_on
product	0	0	0	0	1	0	0	0	0	0	0	computer	new	the	to	NN	TO	VB	line	dobj

Getting data

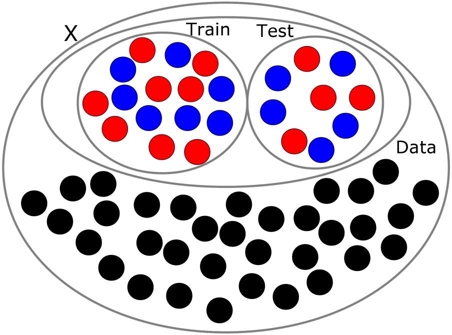
Step 2: Assigning true prediction
Example: $Y = \{red, blue\}$



Getting data

Step 3: Selecting training set *Train* and test set *Test*

- $Train \subseteq Data$, $Test \subseteq Data$
- $Train \cap Test = \emptyset$
- $Train \cup Test = Data$



Summary of Examination Requirements

You should be familiar with the key machine learning terms

- Machine learning process
- Development cycle
- Examples, feature vector, data instance, gold standard data, training data, test data
- Manual annotation (true prediction)
- Model, hypothesis, predictor
- Supervised learning, unsupervised learning
- Classification, regression
- Overfitting