Introduction to Data Mining

Data Mining for Business and Governance*
29/8/2017

*Formerly known as Social Data Mining

Instructors

Lectures: Willem Huijbers

Topic: Neuroimaging

Video Lectures/Practicals: Chris Emmery

Topic: Text Mining

Overview Data Mining

- Practical Notes
- What is Data Mining?
- What makes prediction possible?
- Data Mining as Applied Machine Learning
  - Supervised Learning
  - Unsupervised Learning

Practical Notes

- Lectures: Attendance is expected, slides are not self explanatory, but will be posted online
- Additionally video lectures on blackboard: up to 1 content video and 1 instructional video
- Reading Materials: on blackboard
- Weekly mandatory assignment: **pass/fail**
- **Mid-term** (Sep. 26th)
- Office hours:
  - times will be listed on GitHub.
  - for practical issues only!

More Practical Notes

- Check Github and Subscribe to the course forum on BlackBoard
- Ask any question regarding course content and organization
- Try to answer fellow students' questions
- Chris and me will be monitoring the forum
- Email only for personal issues

https://github.com/tcsai/data-mining

Course Schedule

`v24.06.2017` (subject to change – always check the latest version!)

#	Date	Lectures (Theory - Willem)	Date	Video Lectures (Applications - Chris)	Video Practicals & Notebooks
1	29-08	Introduction to Data Mining	31-08	Introduction to Data Science	Introduction to jupyter, pandas, and scikit-learn
2	05-09	Regression	07-09	Representing Data: Vectors, Types, Databases	Handling & Interpreting Data, Feature Construction
3	12-09	Classification	14-09	Working with Textual Data	Visualizing Data, Distributions, and Models
4	19-09	Algorithm Fitting & Tuning	21-09	Mining Huge Datasets	Preprocessing, Developing and Evaluating Pipelines
5	26-09	**Midterm**	28-09	Discuss Midterm Results + Open Q&A	Online / Out-of-Core Learning and MNIST Challenge
6	03-10	Data Reduction & Decomposition	05-10	Social Media and Multi-modal Data	?
7	10-10	Time Series Analysis	12-10	Applications of Deep Learning	?
8	17-10	Clustering and Graphs	19-10	Explaining Models, Ethics, Privacy	Unsupervised Learning: Intuitions and Metrics

Course on
Data Mining vs Machine Learning

- Beginner: no prerequisites
- Broader, building intuition
- Less technical detail
- Practicals with Python notebooks and R

- Programming in Python
- More focused
- More technical detail
- Practicals with Python

Overlap in course content

Recommendation: Don't try to follow both Data Mining and Machine Learning at the same time

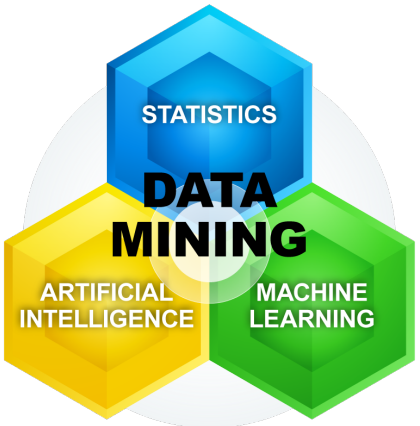
What is Data Mining?

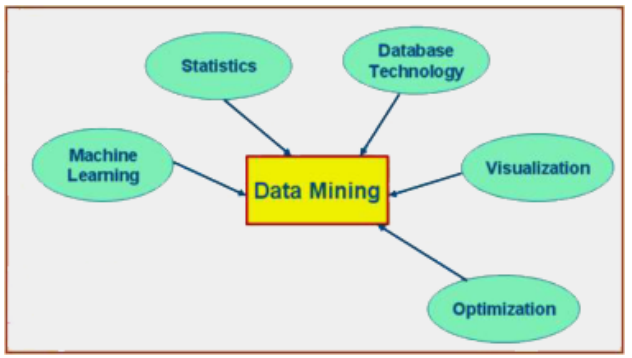
"Data mining is the **computational** process of discovering patterns in **large data sets** involving methods at the intersection of **artificial intelligence**, **machine learning**, **statistics**, and **database** systems."

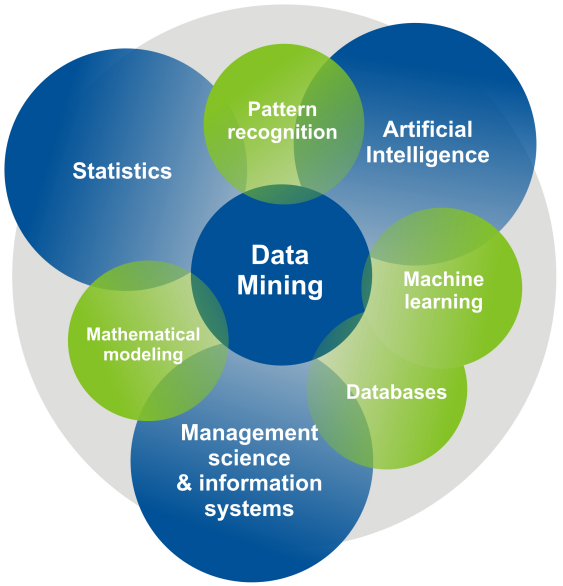
https://en.wikipedia.org/wiki/Data_mining

What is Data Mining?

"Data mining is the **computational** process of discovering patterns in **large data sets** involving methods at the intersection of **artificial intelligence**, **machine learning**, **statistics**, and **database** systems."







https://en.wikipedia.org/wiki/Data_mining

Related Scientific Disciplines:

- **Statistics**: branch of mathematics focused on data
- **Machine Learning**: branch of Computer Science studying learning from data
- **Artificial Intelligence**: Interdisciplinary field aiming to develop intelligent machines

Key aspects

- Computation vs Large data sets

  trade-off between processing time and memory

- Computation enables analysis of large data sets:

  computers as a tool and with growing data

- Data Mining often implies data discovery from data bases

  from unstructured data to structured knowledge

What is large amounts or big data?

Volume

- Too big for manual analysis
- Too big to fit in RAM
- Too big to store on disk

Variety

- Range of values: variance
- Outliers, confounders and noise
- Interactions, data is codependent

Velocity

- Data changes quickly: require results before data changes
- Streaming data (no storage)

Introduction into Data Mining

- Data Mining is as Applied Machine Learning: it requires skill and as with most skills you get better with practice and experience.

Application of data mining

Companies: Business Intelligence

amazon

Booking.com

Science: Knowledge Discovery

Market analysis and management

- Target marketing, customer relation management
- Risk analysis and management
- Forecasting, customer retention, quality control, competitive analysis
- Fraud detection and management

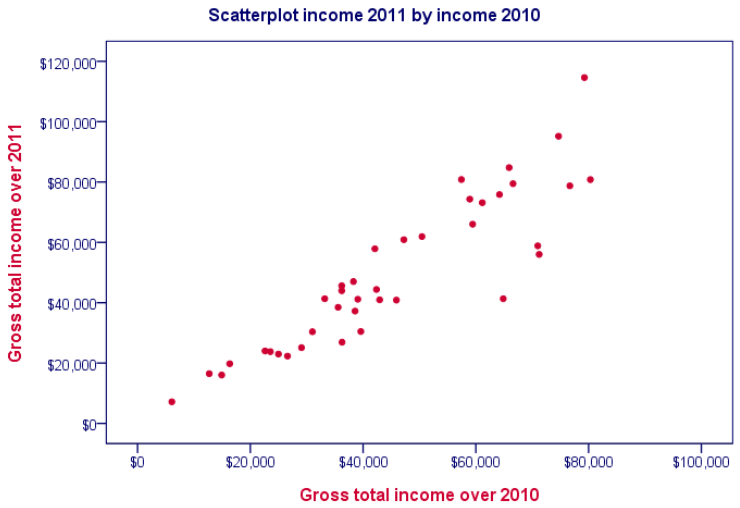
Scientific discovery in large data

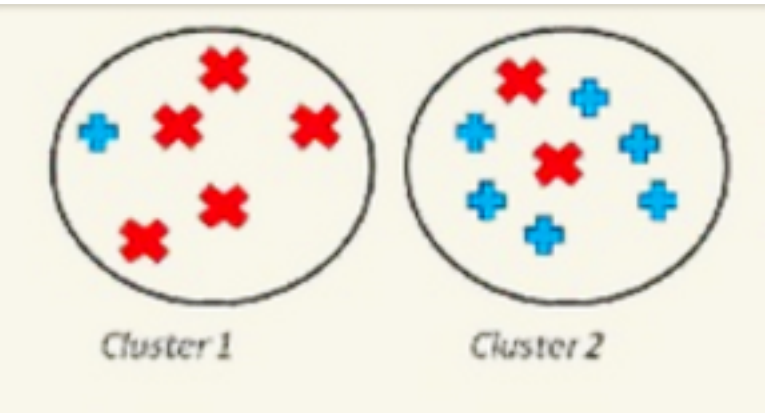
- DNA: sequence data
- SETI program, time series
- Electronic Health Records (Electronisch Patienten Dossier)

Text Mining (natural language processing): going from unstructured text → structured knowledge

What makes prediction possible?

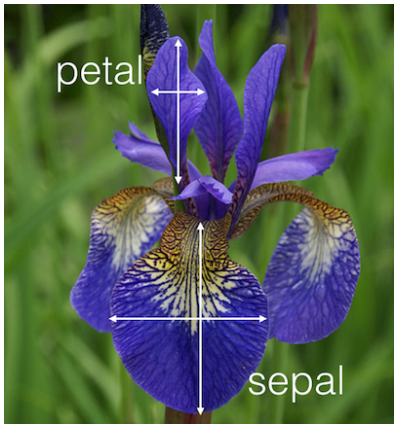
- Associations between features/target
- Numerical: correlation coefficient
- Categorical: mutual information Value of x_1 contains information about value of x_2



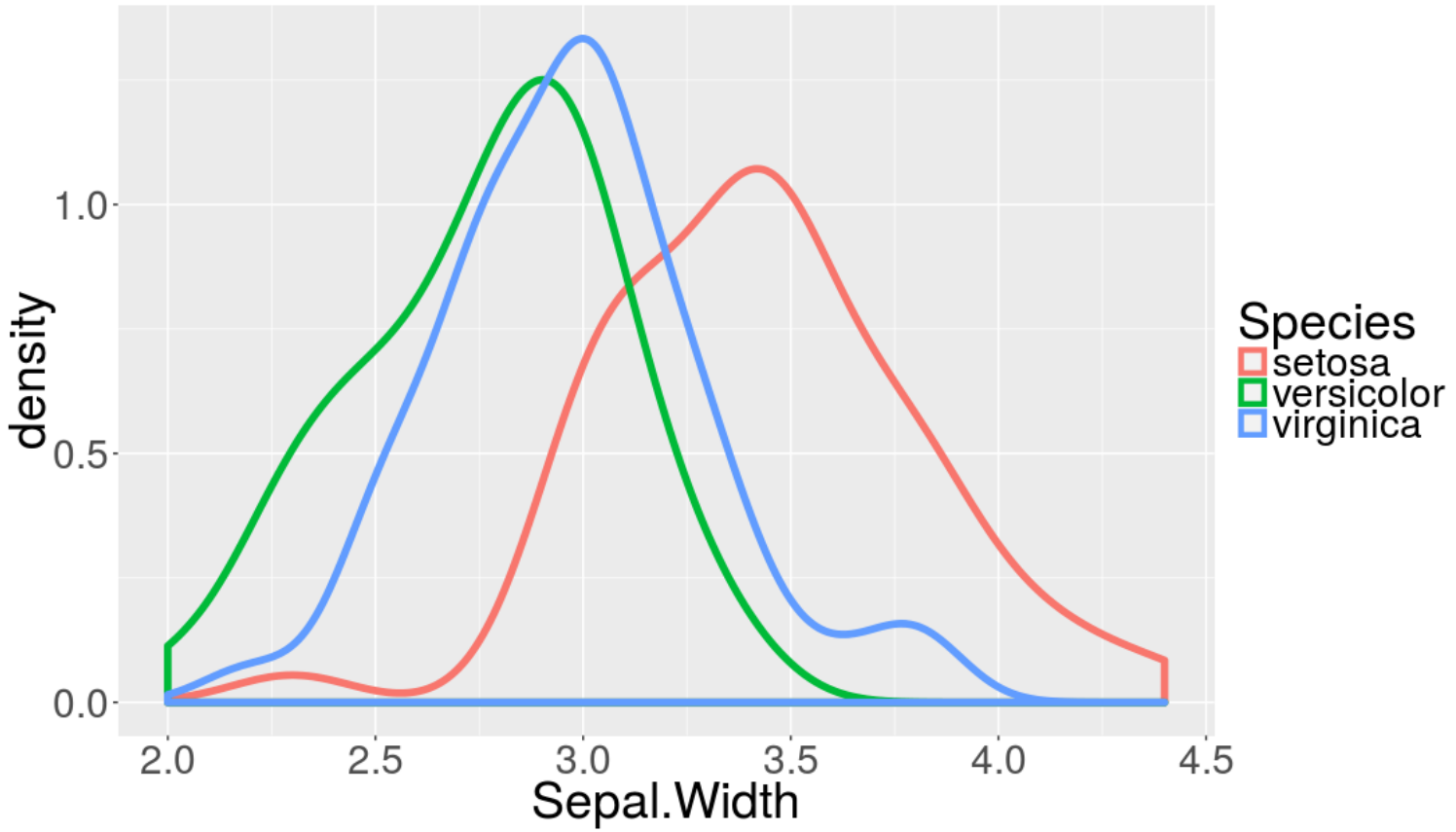


An example: the Iris dataset

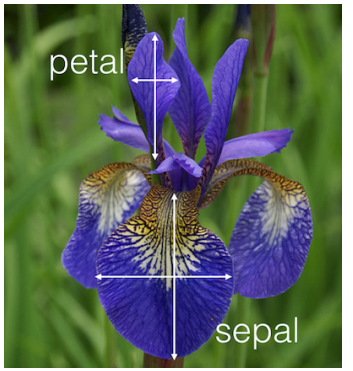
Nr	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
94	6.0	2.9	4.5	1.5	versicolor
95	5.4	3.0	4.5	1.5	versicolor
96	6.7	3.1	4.7	1.5	versicolor
97	6.0	2.2	5.0	1.5	virginica
98	6.3	2.8	5.1	1.5	virginica
99	6.3	3.3	4.7	1.6	versicolor

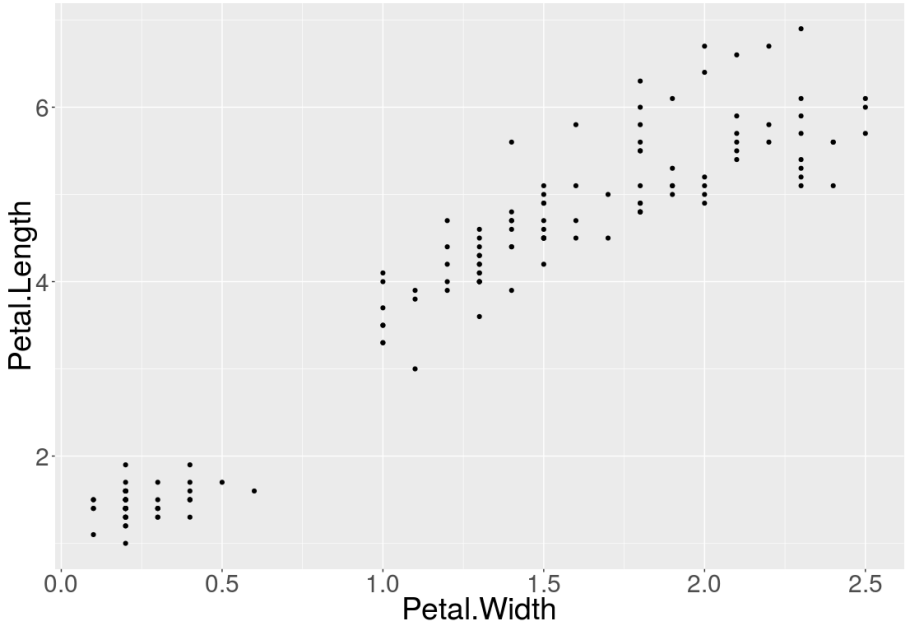


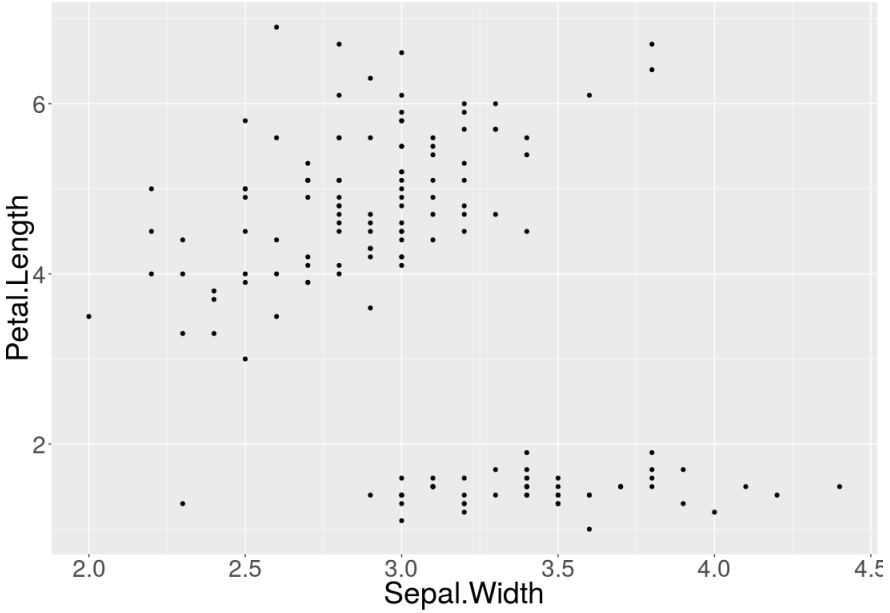
Iris: Sepal Width vs Species



Iris: Petal Length







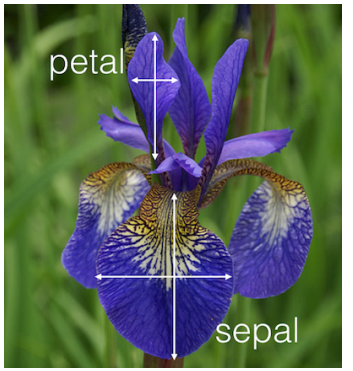
Pearson's correlation coefficient

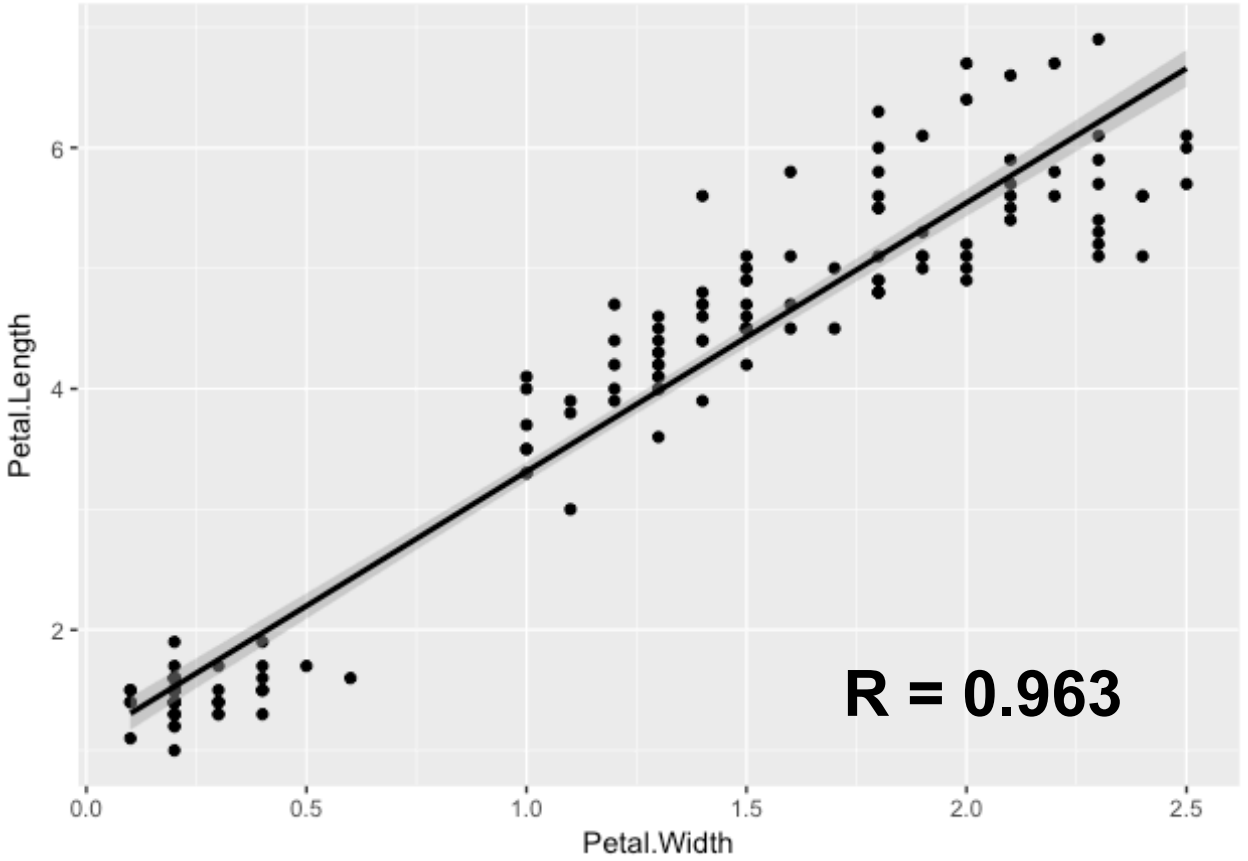
- Numerator: **covariance**. To what extent the features change together.

- Denominator: **product of standard deviations.** Makes correlations independent of units.

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

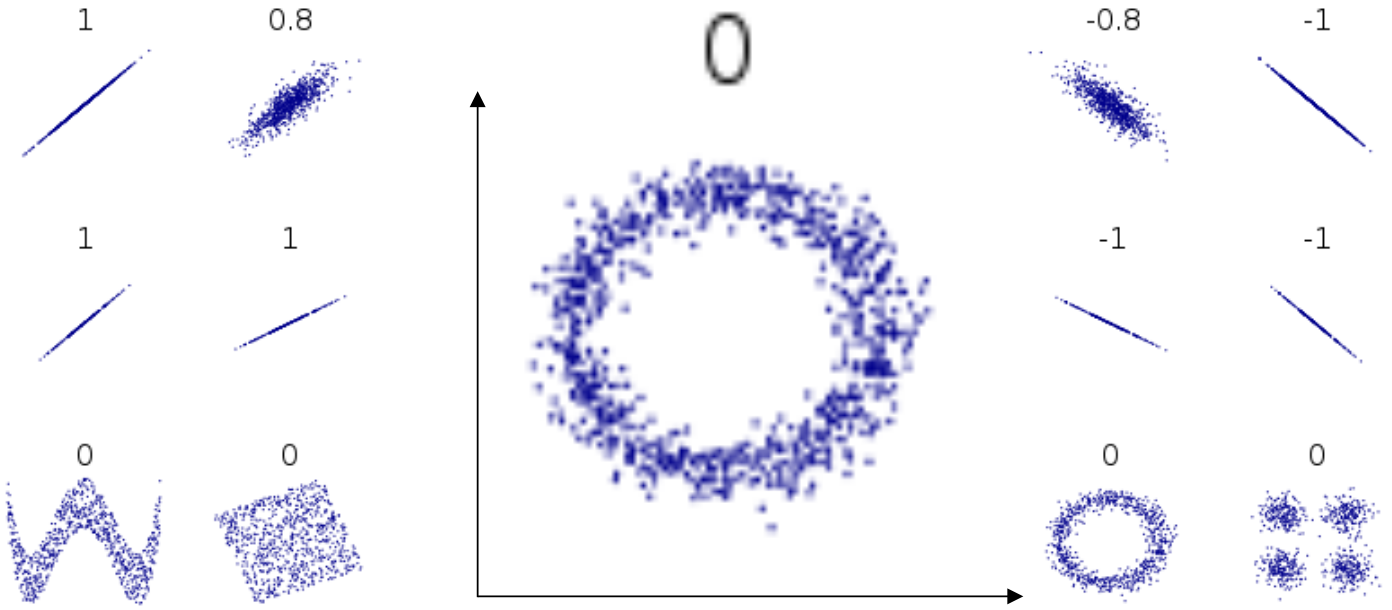
Pearson's coefficient of Petal Length by Petal Width





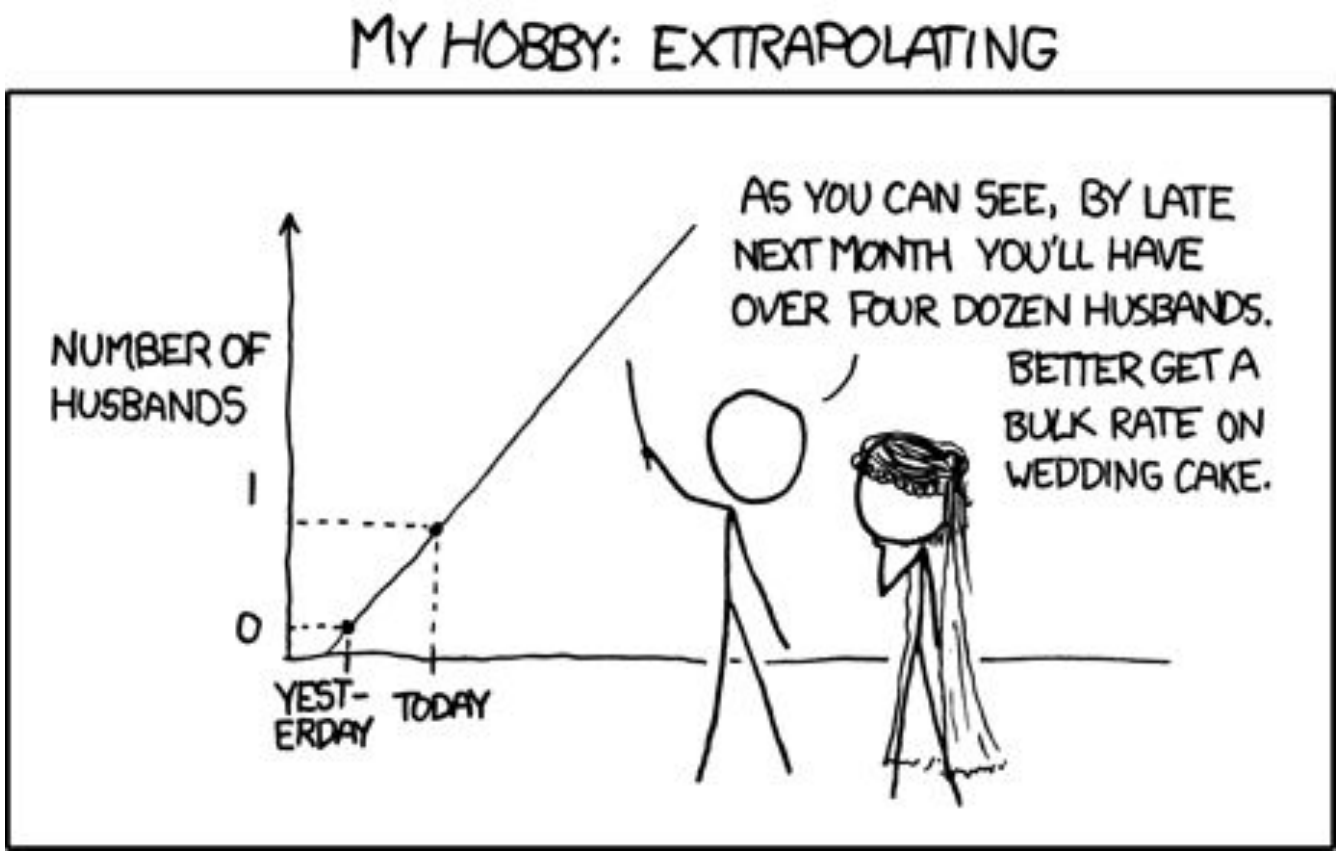
Caveats

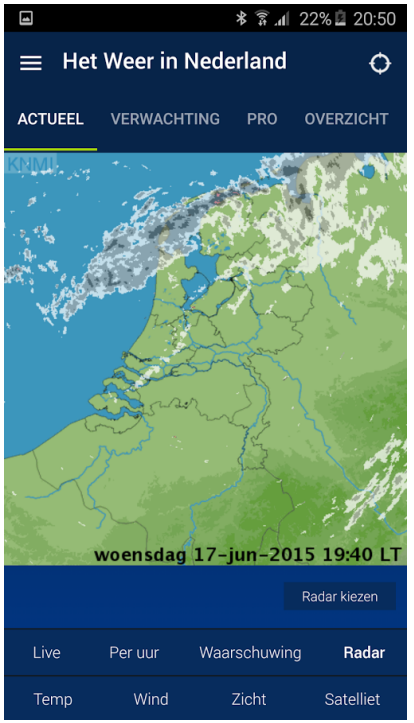
- Pearson's *r* only measures **linear** dependency
- Other types of dependency can also be used for prediction!
- **Correlation** does not imply **causation** but it may still enable prediction



What makes prediction possible?

- Fitting data is easy, but predictions are hard





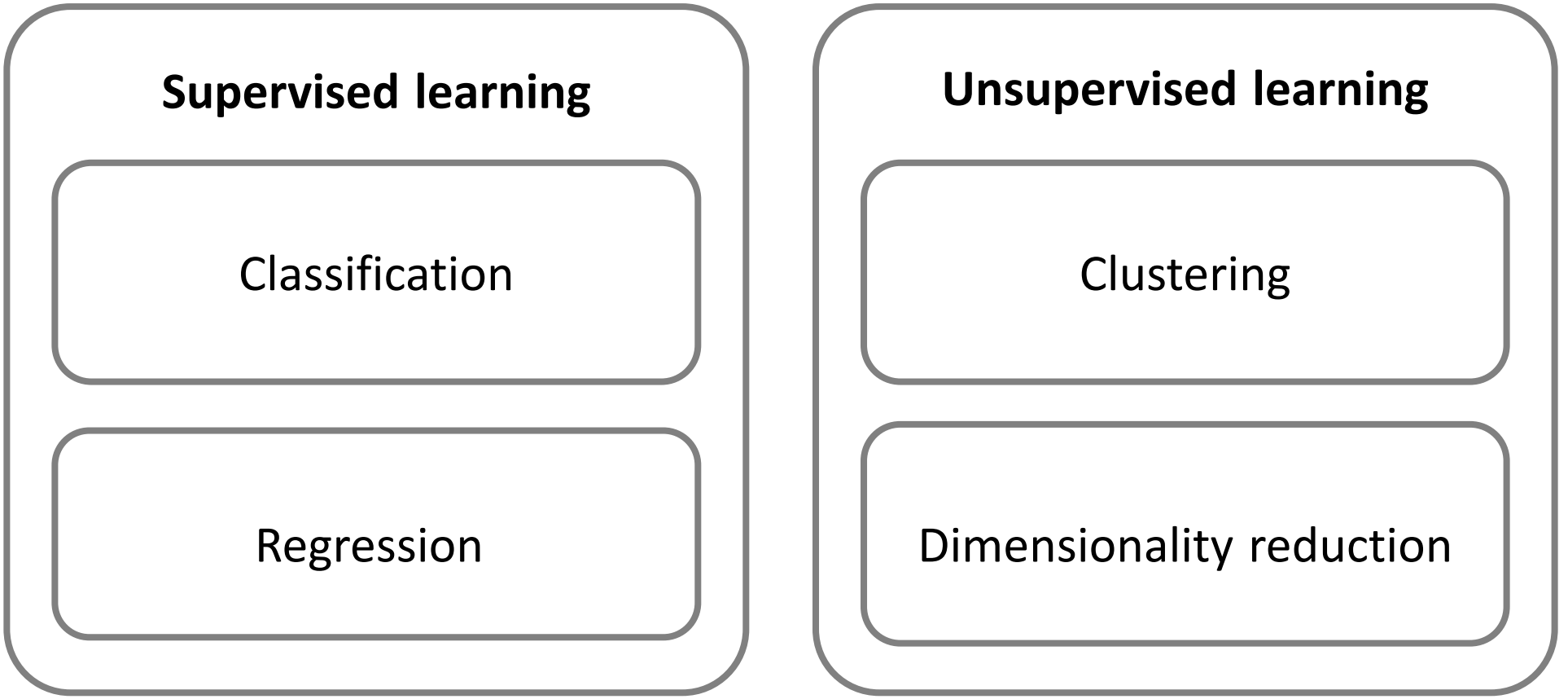
Course Schedule

`v24.06.2017` (subject to change – always check the latest version!)

#	Date	Lecture I (Theory - Willem)	Date	Lecture 2 (Applications - Chris)	Video Practicals
1	00-00	Introduction to Data Mining	00-00	Introduction to Data Science	Introduction to Jupyter, Pandas, and Scikit-learn
2	00-00	Regression	00-00	Representing Data: Vectors, Types, Databases	Handling & Interpreting Data, Feature Construction
3	00-00	Classification	00-00	Working with Textual Data	Visualizing Data, Distributions, and Models
4	00-00	Algorithm Fitting & Tuning	00-00	Mining Huge Datasets	Preprocessing, Developing and Evaluating Pipelines
5	00-00	**Midterm**	00-00	Discuss Midterm Results + Open Q&A	Online / Out-of-Core Learning and MNIST Challenge
6	00-00	Data Reduction & Decomposition	00-00	Social Media and Multi-modal Data	Building a System for Music Recommendation
7	00-00	Time Series Analysis	00-00	Applications of Deep Learning	Imaging the Brain as a Sequence
8	00-00	Clustering and Graphs	00-00	Explaining Models, Ethics, Privacy	Unsupervised Learning: Intuitions and Metrics

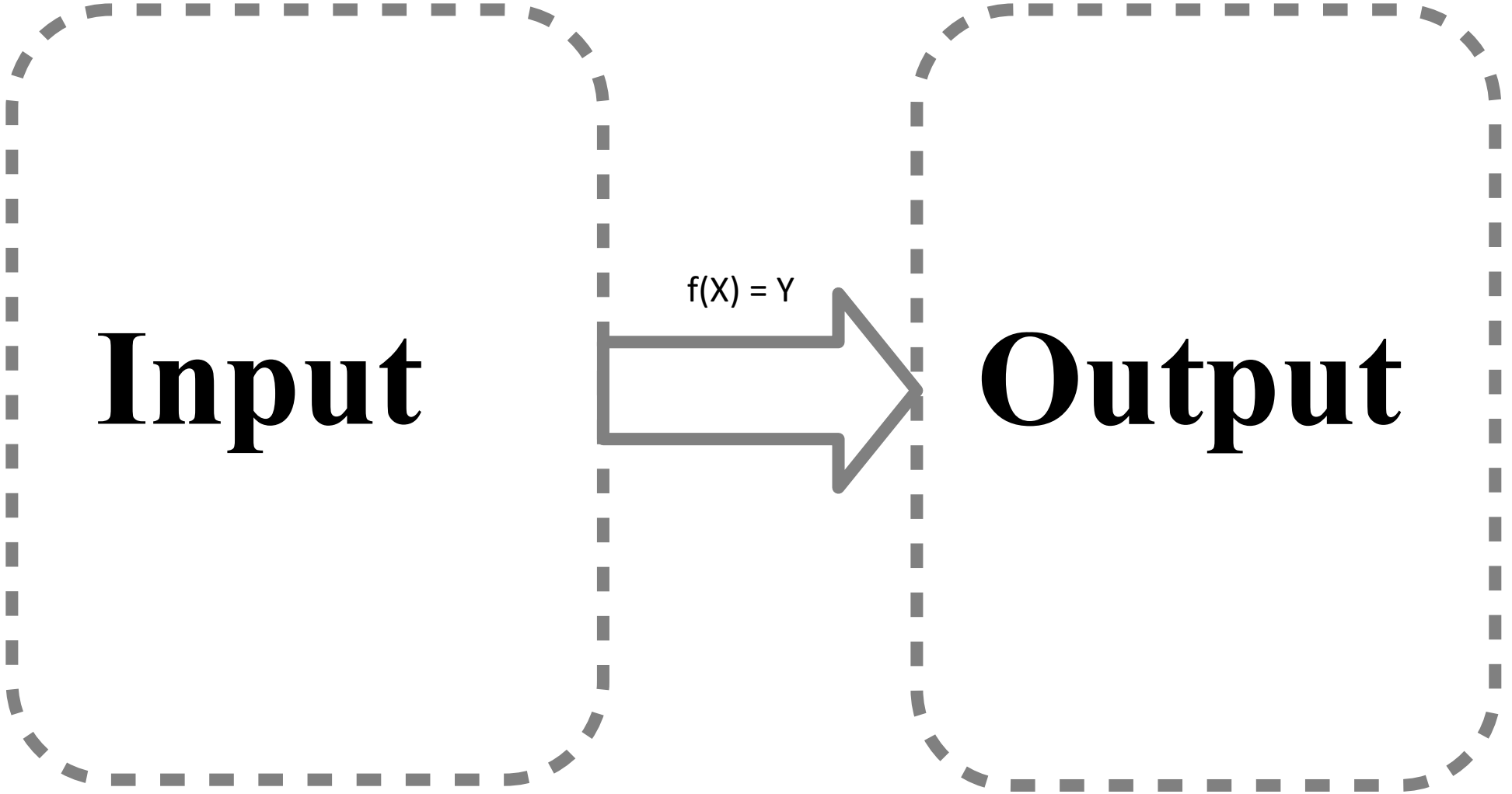
What is machine learning?

"An program is said to learn from experience (E) on task (T) and a performance measure (P), if its performance at tasks in T as measured by P improves with E"

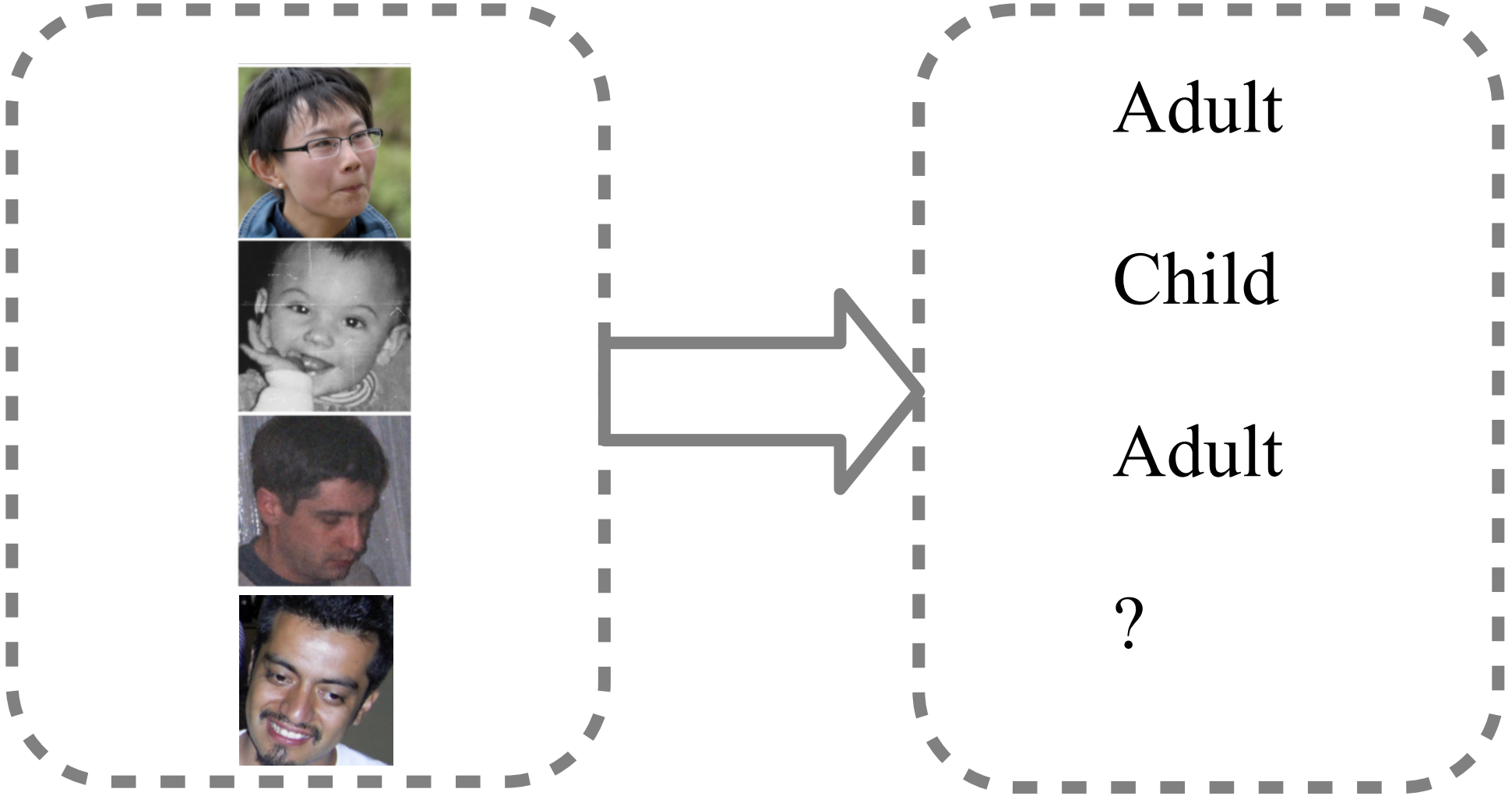


Other form of ML are Semi-supervised learning, Reinforcement Learning…

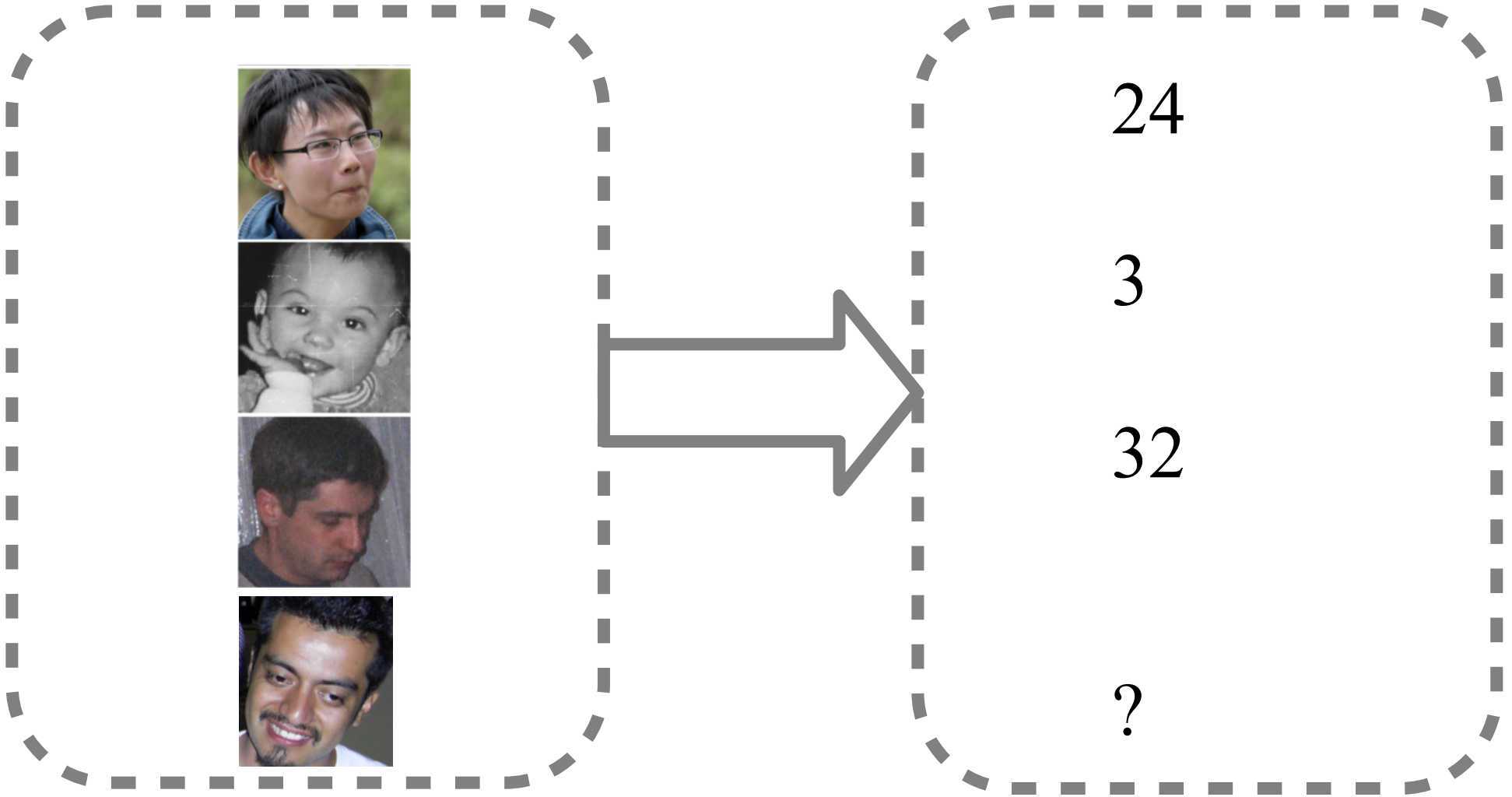
Supervised learning



Classification Task



Regression Task



What is machine learning?

"An algorithm is said to learn from experience (E) on task (T) and a performance measure (P), if its performance at tasks in T as measured by P improves with E"

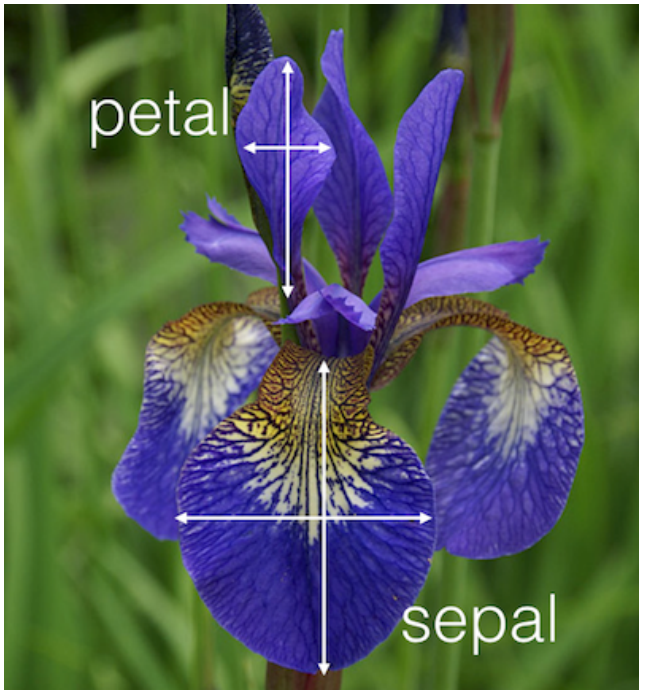
Supervised learning

- Classification
- Regression

Unsupervised learning

- Clustering and Graphs
- Dimensionality reduction

Semi-supervised learning, Reinforcement Learning…



$$[123, 189, 5, 123, 232, \ldots]$$

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
5.1	3.5	1.4	0.2

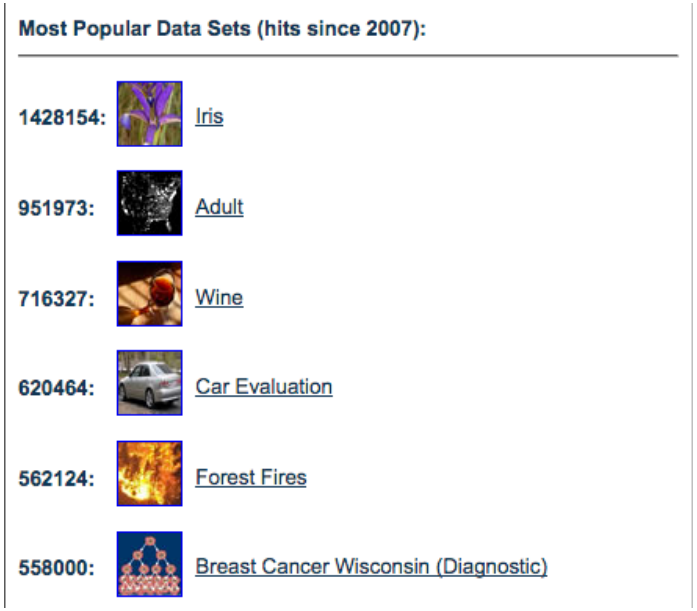
Supervised learning Workflow

1. Collect data
2. Label examples
3. Choose representation
4. Train model(s)
5. Evaluate

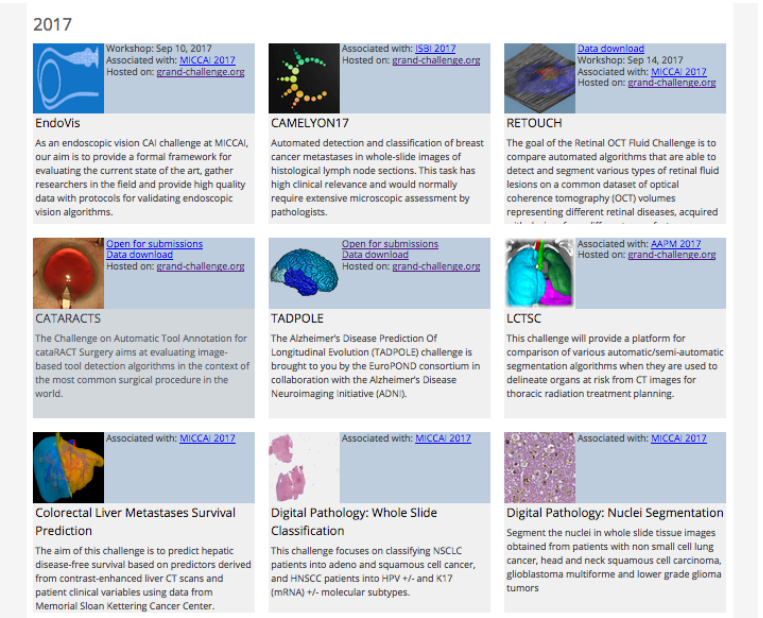
1. Collect data

- How do you select your sample?
- Reliability of measurement
- Privacy and other regulations

Open Access Repositories



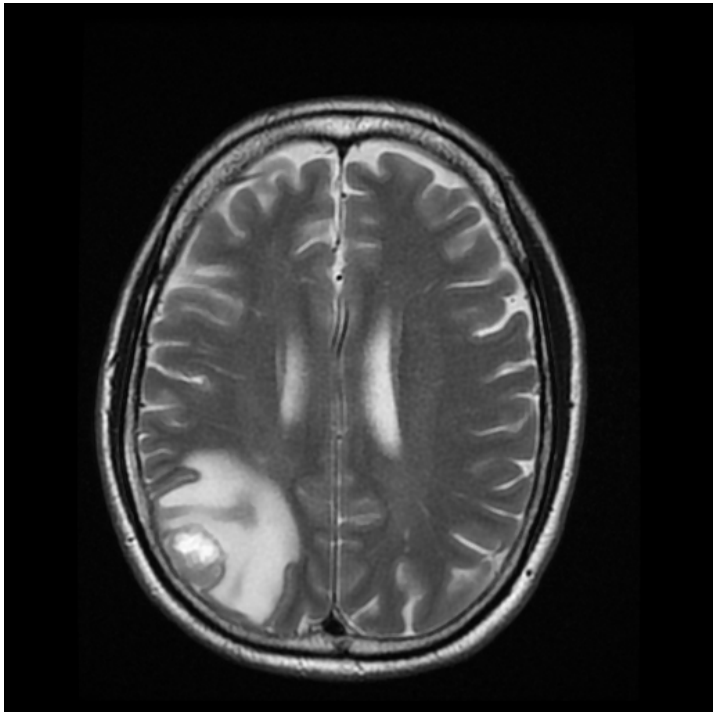
ULC Repository



Grand Challenges .org

2. Label examples

- Annotation guidelines
- Measure inter-annotator agreement
- Crowdsourcing

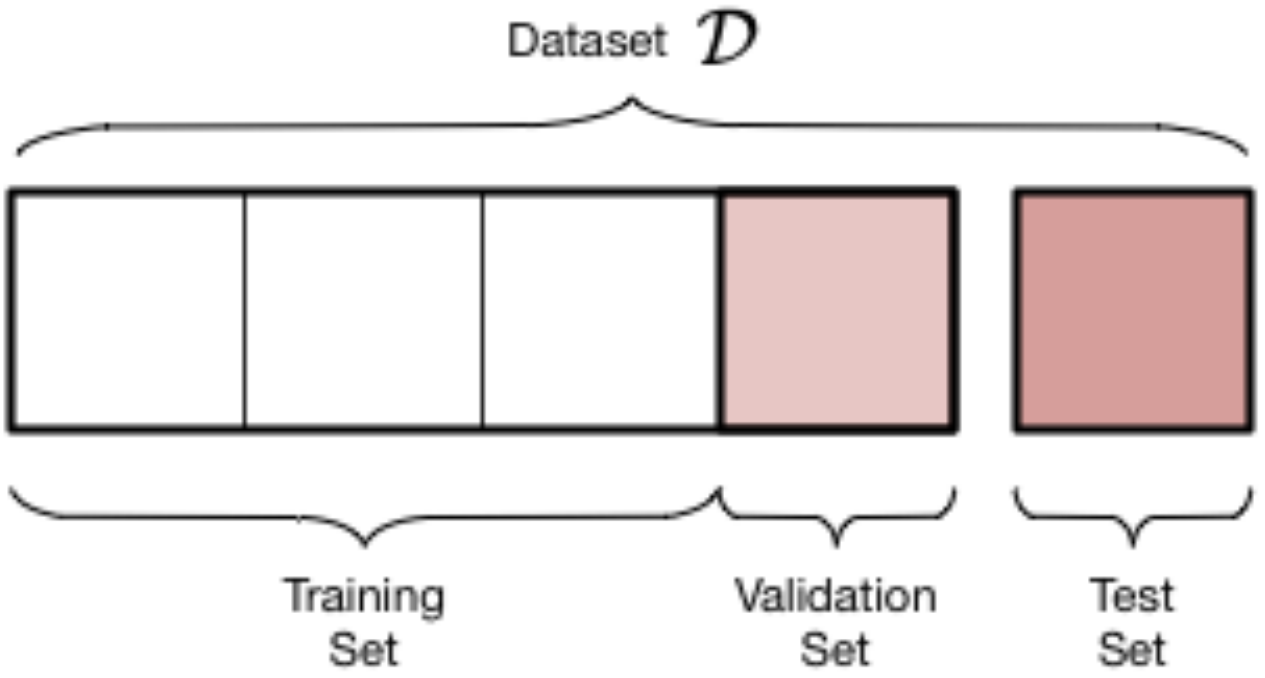


3. Representation

- Features: attributes describing examples
  - Numerical
  - Categorical
- Possibly convert to feature **vector**
  - A vector is a fixed-size list of numbers
  - Some learning algorithms require examples represented as vectors

4. Train

- Keep some examples for final evaluation: test set
- Use the rest for:
  - Learning: training set
  - Tuning: validation set



Parameter or model Tuning

- Learning algorithms typically have settings (aka **hyperparameters**)

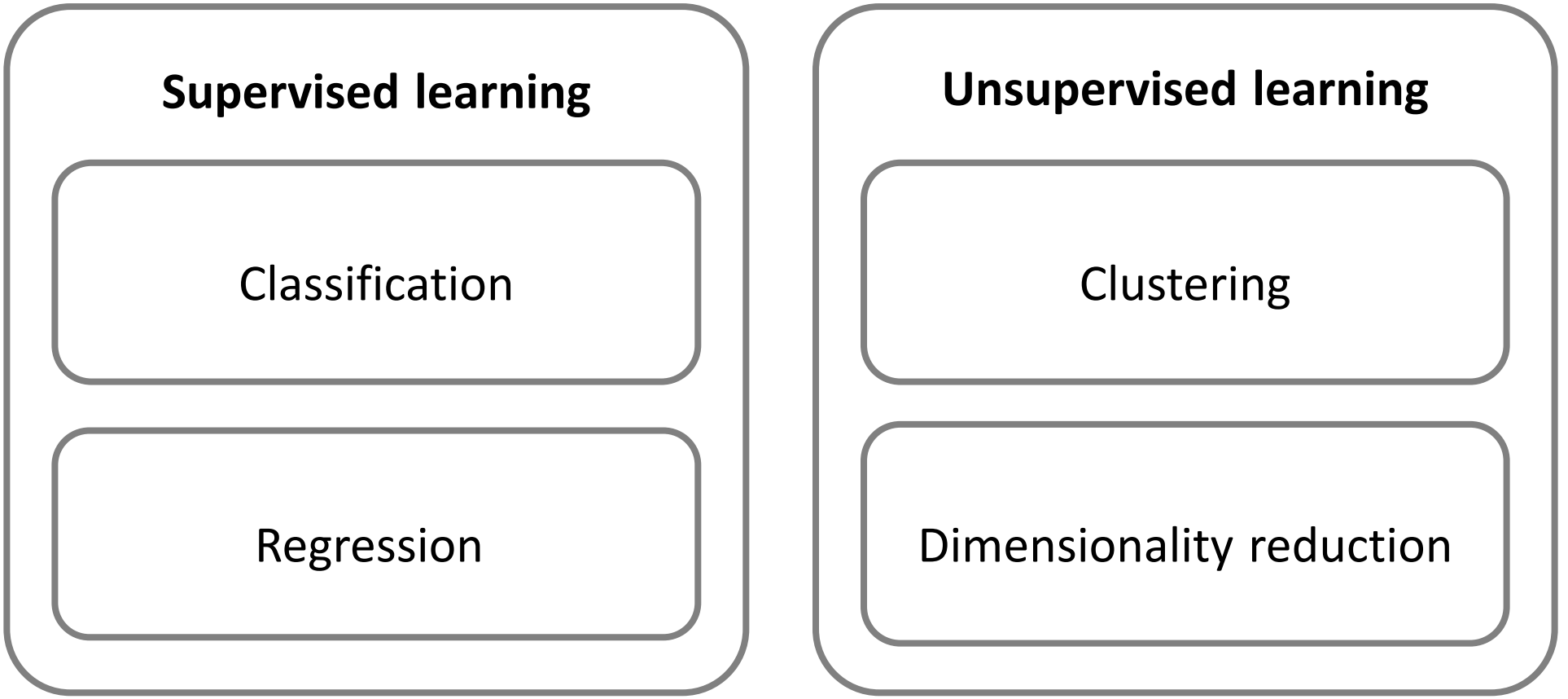
- For each value of hyperparameters:
  - Apply algorithm to training set to learn
  - Check performance on validation set
  - Find/Choose best-performing setting

5. Evaluate

- Check performance of tuned model on test set
- Goal: estimate how well your model will do in the **real world**.
- Keep evaluation realistic.

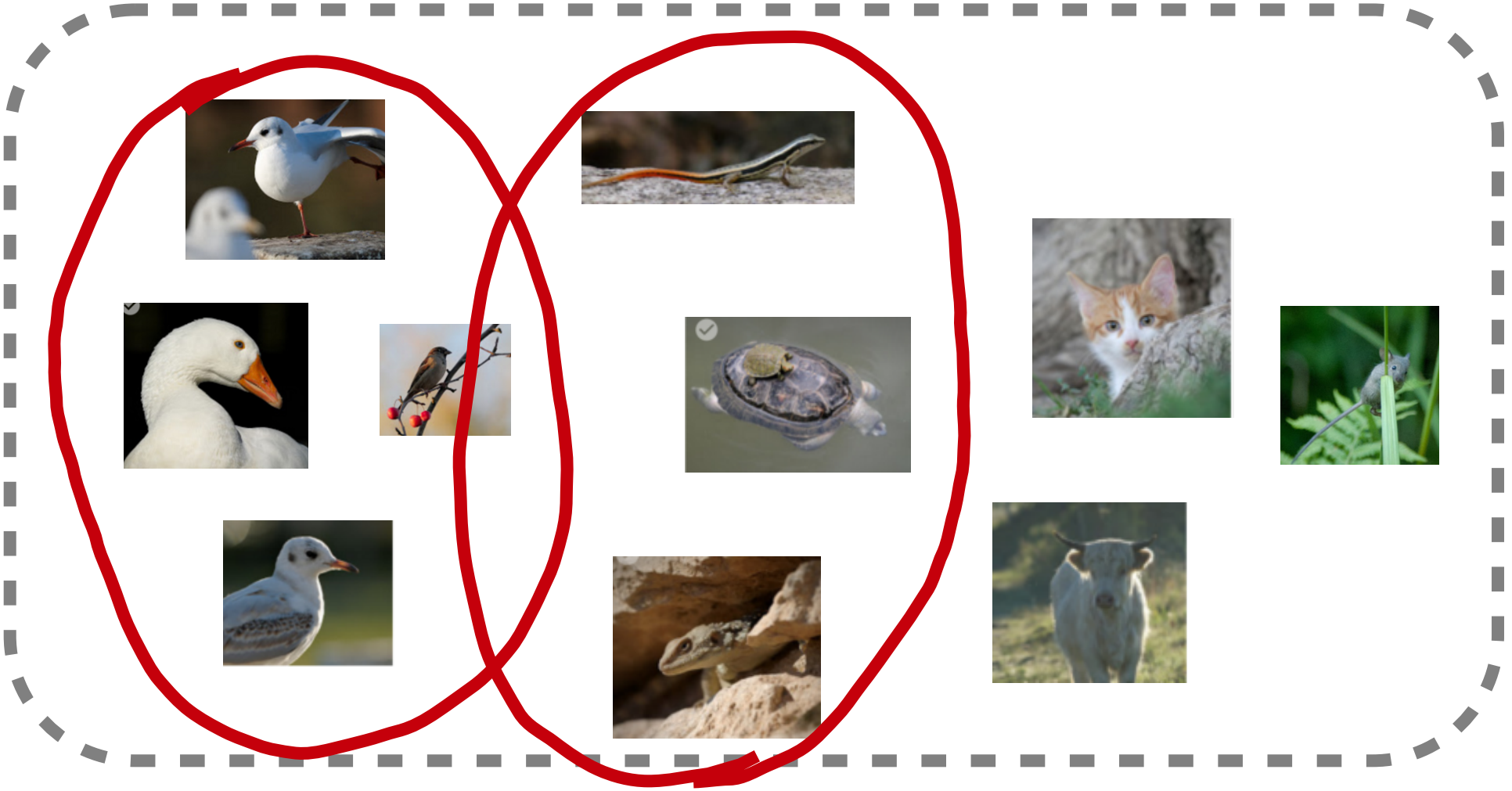
What is machine learning?

"An program is said to learn from experience (E) on task (T) and a performance measure (P), if its performance at tasks in T as measured by P improves with E"

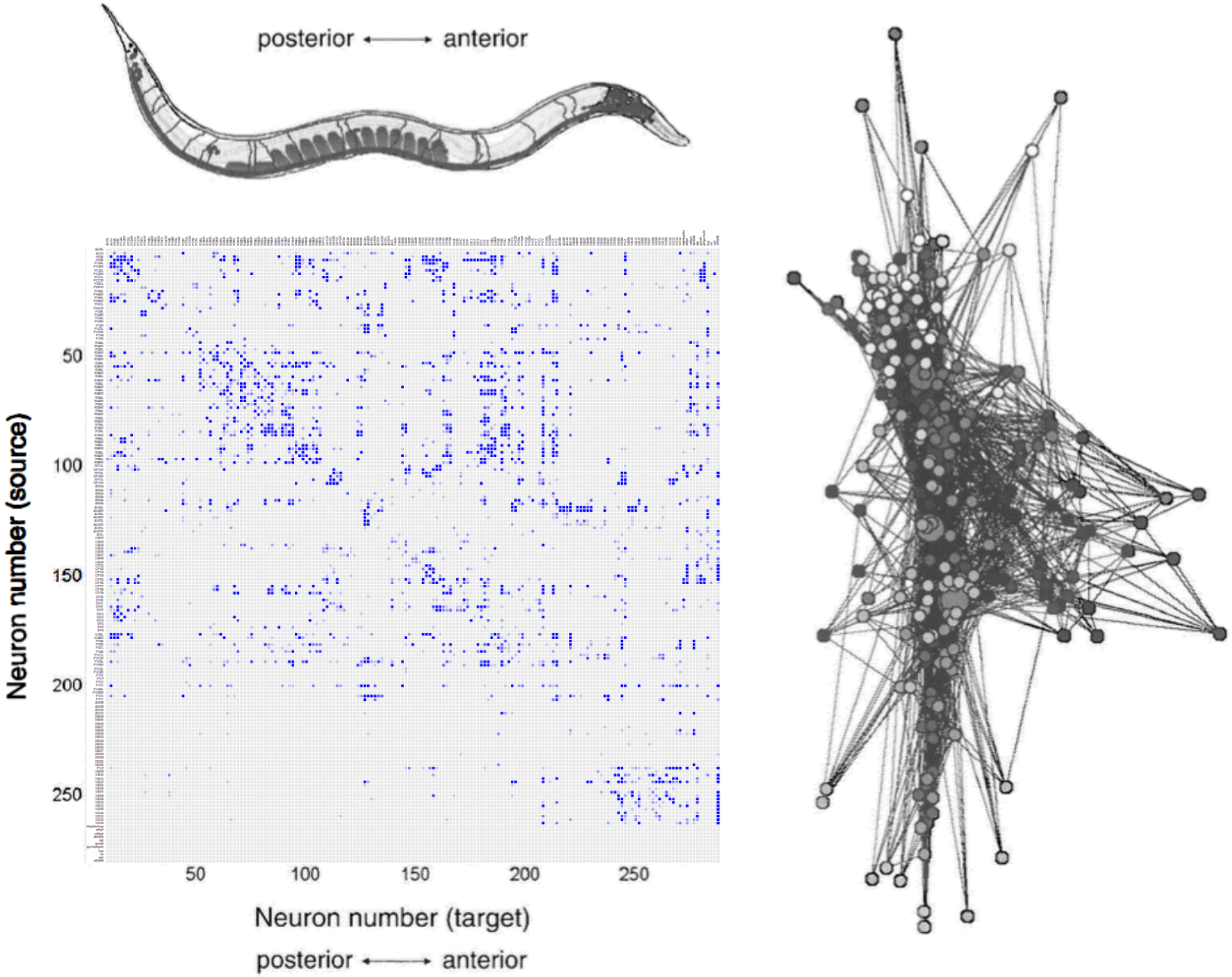


Other form of ML are Semi-supervised learning, Reinforcement Learning…

Clustering



Clustering and graph analysis



What is machine learning?

"An algorithm is said to learn from experience (E) on task (T) and a performance measure (P), if its performance at tasks in T as measured by P improves with E"

Supervised learning

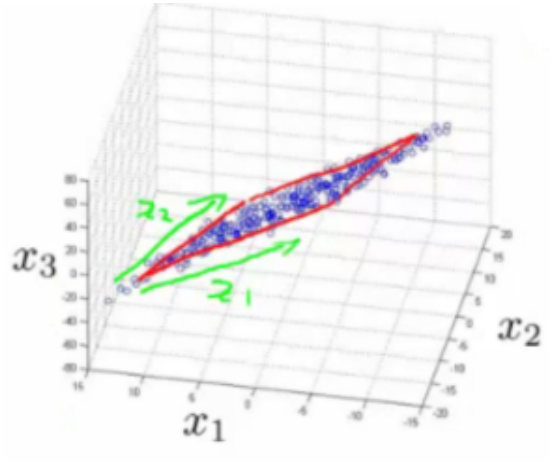
- Classification
- Regression

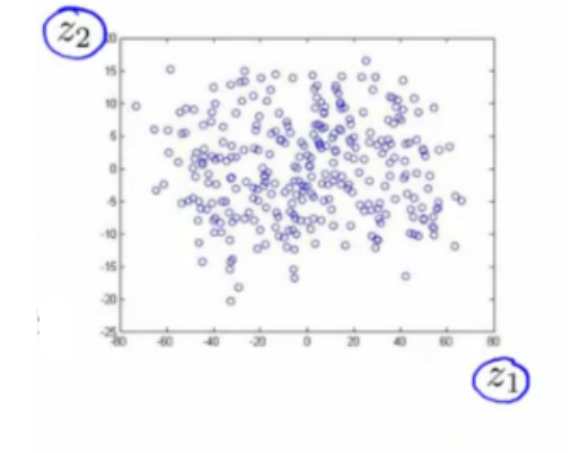
Unsupervised learning

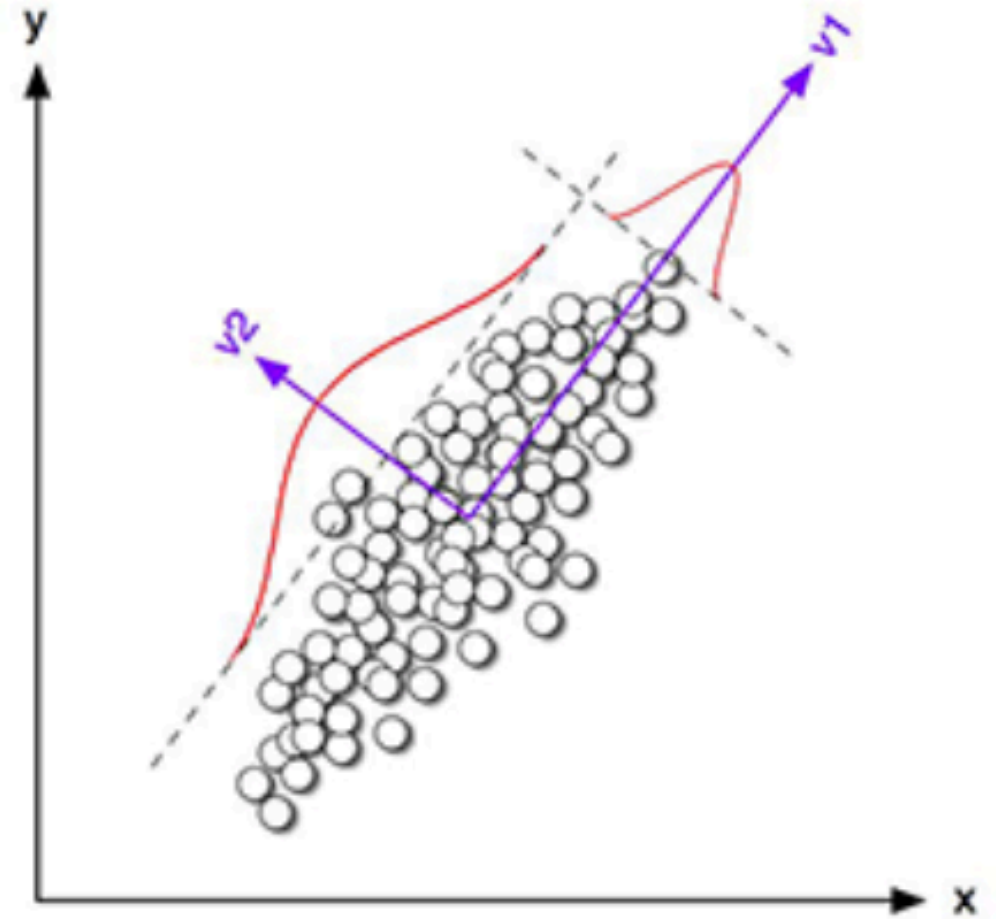
- Clustering
- Dimensionality reduction

Semi-supervised learning, Reinforcement Learning…

Dimensionality reduction: principle component analysis



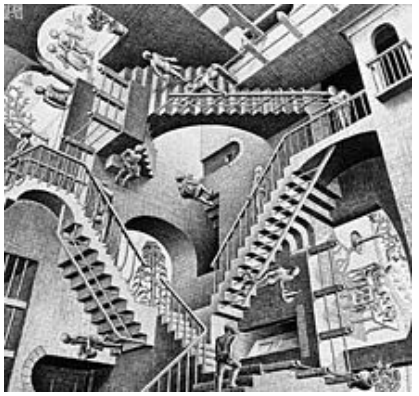




Dimensionality reduction

Feature selection: reduce the large amount of data

- Reduce Complexity and easier interpretation
- Reduce demand on resources (computation/memory)
- Reduce the “curse of dimensionality”
- Reduce chance of over-fitting



Feature extraction: often domain specific

- Image Processing: edge detection
- From pixels to reduced set of features
- Often part of preprocessing, but might harbor the hard problems

Course Schedule

`v24.06.2017` (subject to change – always check the latest version!)

#	Date	Lecture I (Theory - Willem)	Date	Lecture 2 (Applications - Chris)	Video Practicals
1	00-00	Introduction to Data Mining	00-00	Introduction to Data Science	Introduction to Jupyter, Pandas, and Scikit-learn
2	00-00	Regression	00-00	Representing Data: Vectors, Types, Databases	Handling & Interpreting Data, Feature Construction
3	00-00	Classification	00-00	Working with Textual Data	Visualizing Data, Distributions, and Models
4	00-00	Algorithm Fitting & Tuning	00-00	Mining Huge Datasets	Preprocessing, Developing and Evaluating Pipelines
5	00-00	**Midterm**	00-00	Discuss Midterm Results + Open Q&A	Online / Out-of-Core Learning and MNIST Challenge
6	00-00	Data Reduction & Decomposition	00-00	Social Media and Multi-modal Data	Building a System for Music Recommendation
7	00-00	Time Series Analysis	00-00	Applications of Deep Learning	Imaging the Brain as a Sequence
8	00-00	Clustering and Graphs	00-00	Explaining Models, Ethics, Privacy	Unsupervised Learning: Intuitions and Metrics

https://github.com/tcsai/data-mining

Dimensionality reduction

Handedness Questionnaire

Most people are either right-handed or left-handed. However, there are different "degrees" of handedness. Some people use one hand for jobs that require skill and the other hand for jobs that involve reaching. Other people use the same hand for these different jobs. Use this "Handedness Questionnaire" to measure the strength of handedness. Place a mark in a box for each question that describes you best.

	LEFT Hand	RIGHT Hand	EITHER Hand
1. Which hand do you use to write?			
2. Which hand do you use to draw?			
3. Which hand do you use to throw a ball?			
4. Which hand do you hold a tennis racket?			
5. With which hand do you hold a toothbrush?			
6. Which hand holds a knife when you cut things?			
7. Which hand holds a hammer when you nail things?			
8. Which hand holds a match when you light it?			
9. Which hand holds an eraser when you erase things?			
10. Which hand removes the top card when you deal from a deck?			
11. Which hand holds the thread when you thread a needle?			
12. Which hand holds a fly swatter?			
TOTAL			

How to Determine your Score

1. Count the number of LEFT, RIGHT and EITHER responses.
2. Multiply the number of RIGHT responses by 3. This number = R
3. Multiply the number of EITHER responses by 2. This number = E
4. Add R + E + (number of LEFT responses). This sum is your score.

Here is a table to help:
Number of RIGHT responses x 3 = ____
Number of EITHER responses x 2 = ____
Number of LEFT responses = ____
TOTAL = _____

Score Handedness
33 to 36 = Strongly Right-Handed
29 to 32 = Moderately Right-Handed
25 to 28 = Weakly Right-Handed
24 = Ambidextrous
20 to 23 = Weakly Left-Handed
16 to 19 = Moderately Left-Handed
12 to 15 = Strongly Left-Handed