BIOS 7747: Machine Learning for Biomedical Applications

Course presentation - Introduction to machine learning

Antonio R. Porras (antonio.porras@cuanschutz.edu)

Department of Biostatistics and Informatics
Colorado School of Public Health
University of Colorado Anschutz Medical Campus

Course summary

- Credits: 3
- Target audience: MS or PHD students in Biostatistics, Bioengineering, or Computational Bioscience. But others are welcome!
- Prerequisites:
  - Biostatistical methods (e.g., BIOS 6611, BIOS 6612)
  - Linear algebra (e.g., MATH 3191)
  - Python programming (e.g., BIOS 6642)
- Classes: Mondays and Wednesdays – 9:00-10:20AM
- Location: Education 2 South L28-2306
- Office hours: Wednesdays, 1-2pm. Building 500, W4132

Course summary

❑ Materials

- Reading requirements: no book required.
- Programming environment: Python 3 (Visual Studio Code recommended as editor)
  - Note: non-native Python development environment problems will not be addressed
- Supporting materials:
  - Introduction to Machine Learning. Ethem Alpaydin. Third Edition. 2014. ISBN 0262028182.
  - Deep Learning. Ian Goodfellow and Yoshua Bengio and Aaron Courville. MIT Press. 2016. https://www.deeplearningbook.org/.
  - Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Springer, 2013. Corrected 8th printing, 2017. ISBN 1461471370.
  - Deep Learning with PyTorch: Build, train, and tune neural networks using Python tools 1st Edition. Eli Stevens, Luca Antiga, and Thomas Viehmann. Manning. 1617295264.

Course summary

- Evaluation

  - Assignments / homework: 20%
  - Paper presentations 10%
  - Flipped classrooms 15%
  - Classroom: 5%
  - Final exam: 50%

Course summary

Date	Format	Topic
08/28	Lecture	Course introduction
08/30	Practical	Practical warm-up class: Python setup and use of common libraries
09/04		Labor Day
09/06	Lecture	Guest lecture by Prof. Matthew DeCamp. Ethics and Biases in Machine Learning: More than the Data.
09/11	Lecture	Supervised machine learning: regression, regular gradient descent optimization, linear and non-linear regression.
09/13	Practical	Regression and optimization with Python: Statsmodels, Scikit-learn and Scipy.
09/18	Lecture	Feature exploration, visualization, pre-processing and normalization.
09/20	Practical	Feature exploration and pre-processing for a non-linear regression problem.
09/25	Lecture	Supervised machine learning: classification and logistic regression. Performance evaluation and cross-validation.
09/27	Practical	Cross-validation of logistic regression-based classification methods.
10/02	Flipped classroom	Supervised machine learning: K-nearest neighbors, decision trees and random forests. Strong vs. weak learners (boosting, bootstrap and bagging), Gradient boosting.
10/04	Practical	K-nearest neighbors, decision trees and random forests in Python.
10/09	Flipped classroom	Supervised machine learning: Lagrange multipliers and support vector machines. The kernel trick. Platt's algorithm. Support vector regression.
10/11	Practical	Support vector machines, class imbalance, Platt's algorithm, understanding and visualizing overfitting in Python.
10/16	Flipped classroom	Unsupervised learning: clustering, mixture models and other alternatives. Selecting the appropriate data for clustering. Performance evaluation.
10/18	Practical	Clustering and visualization in Python.
10/23	Lecture	The curse of dimensionality and dimensionality reduction. Unsupervised dimensionality reduction using principal component analysis. Principal component analysis-based modeling. Supervised dimensionality reduction using linear discriminant analysis.
10/25	Practical	Clustering and dimensionality reduction.

Course summary

10/30	Student presentations	Presentations of feature-based machine learning research papers.
11/01	Student presentations	Presentations of feature-based machine learning research papers.
11/06	Lecture	Introduction to neural networks. Feed-forward networks, activation and backpropagation. Examples of biomedical applications.
11/08	Practical	Introduction to Neural Networks with Pytorch and Tensorboard in Python.
11/13	Lecture	Neural network details and training
11/15	Practical	Introduction to Neural Networks with Pytorch and Tensorboard in Python.
11/20	Lecture	Working with time series and images: convolutional neural networks. Examples and application to biomedical data.
11/22		Thanksgiving Wednesday
11/27	Practical	Convolutional neural networks
11/29	Practical	Convolutional neural networks
12/04	Student presentations	Presentations of deep learning research papers.
12/06	Student presentations	Presentations of deep learning research papers.
12/11		Exams week
12/13		Exams week

Introduction to machine learning

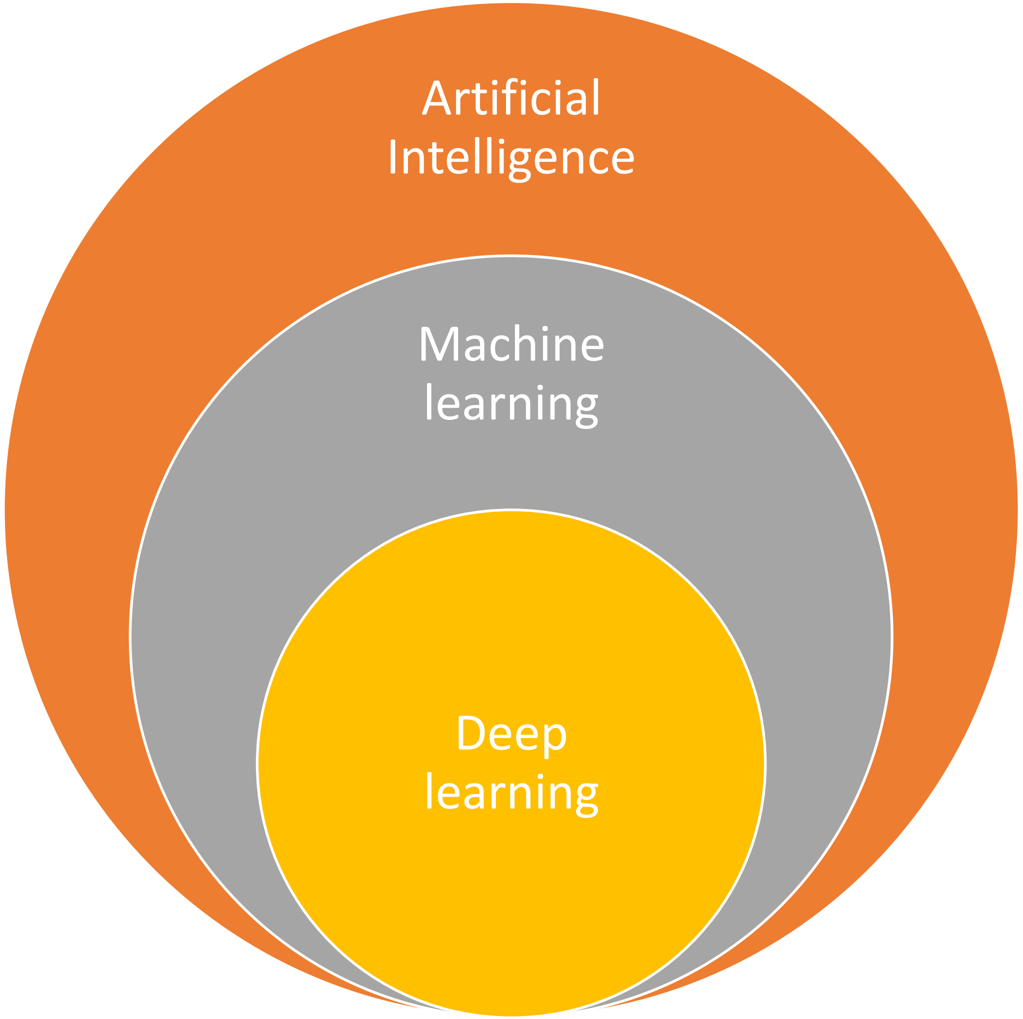
Intelligence: capability of inferring new information, retaining is as knowledge that can be applied within a context or environment

Human intelligence: capability of humans to reach correct conclusions about what is true and false, and to solve problems. It is marked by complex cognitive skills and high levels of motivation and self-awareness.

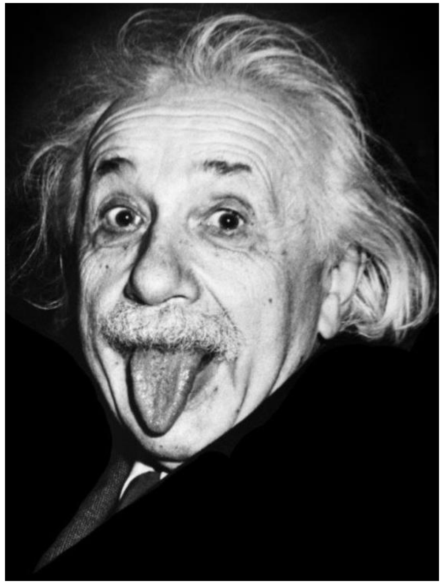
Artificial intelligence: Systems or machines that can mimic human intelligence to perform specific tasks that can iteratively improve themselves based on collected information.

Machine learning: Branch of artificial intelligence and computer science that focuses on developing algorithms that imitate that way humans learn

Deep learning: Branch of machine learning that uses neural networks to leverage large amounts of data



Introduction to machine learning



Human intelligence

- Fast learning
- Can learn millions of highly complex tasks
- Creativity and originality
- Conscious
- Self-aware
- Power-efficient

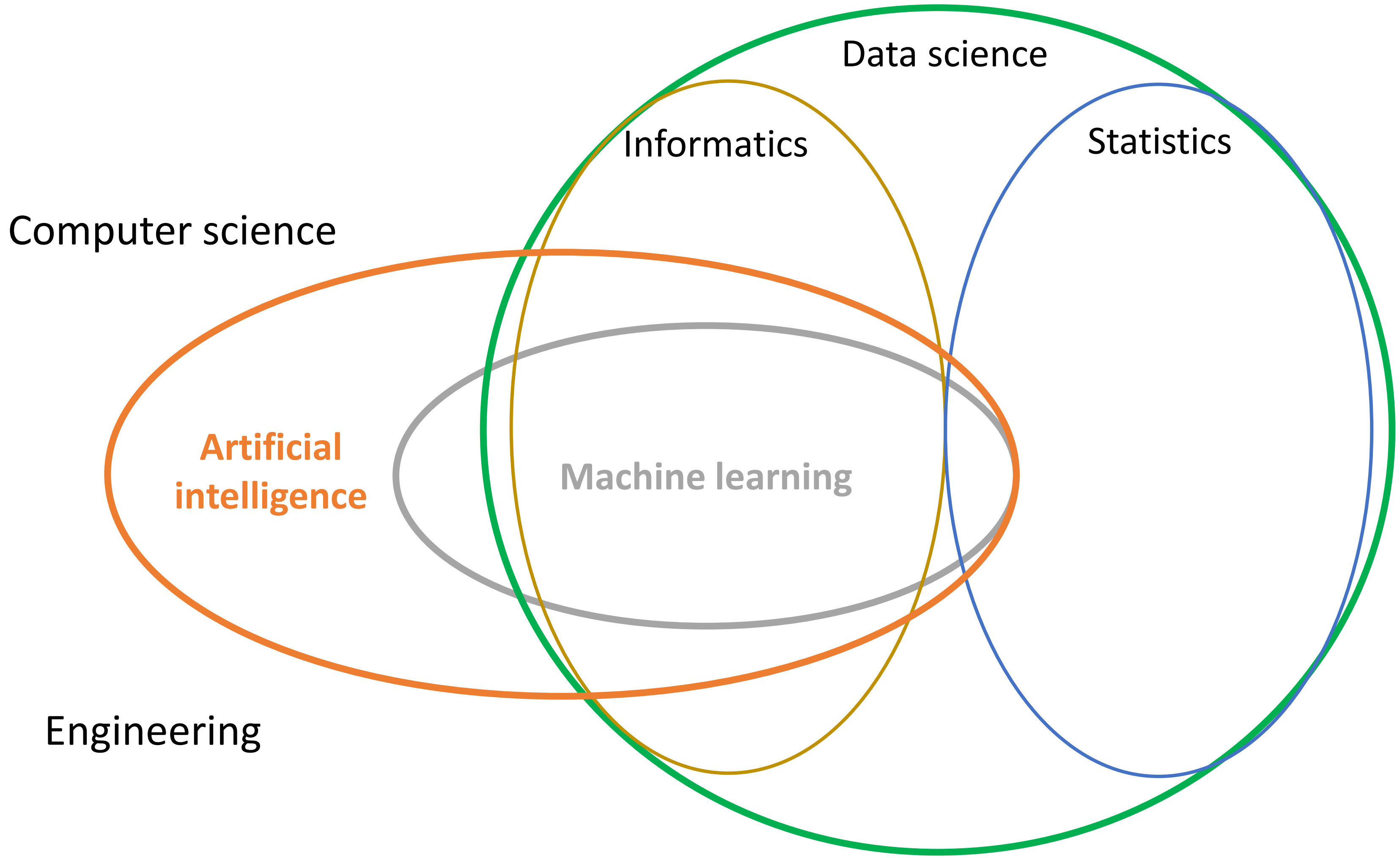
- Influenced by emotions
- Inexact
- Slow
- Forgetful

Artificial intelligence

- Slow learning process
- Can learn a limited amount of simple tasks
- Highly limited creativity
- Unconscious
- Not self-aware
- Power-inefficient

- Repeatable
- Exact
- Fast
- Persistent data

Introduction to machine learning



It's all math!

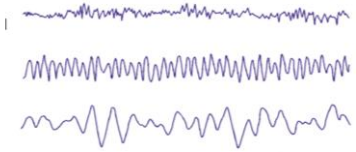
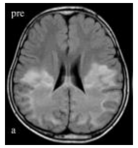
Introduction to machine learning for biomedical applications

An overview of the machine learning approach in biomedicine

1. Data collection

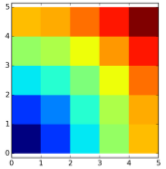




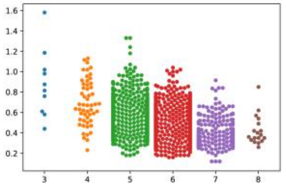
2. Data pre-processing

3. Data representation



10	20	30	40	50	60	70	80	90	100

4. Data wrangling (and more pre-processing) and exploratory analysis



5. Feature selection and/or feature space transformation

6. Model construction

7. Model evaluation

8. Deployment

Machine learning?

Machine learning?

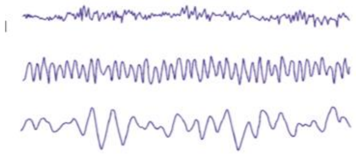
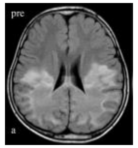
Introduction to machine learning for biomedical applications

An overview of the machine learning approach in biomedicine

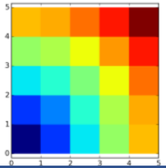
1. Data collection



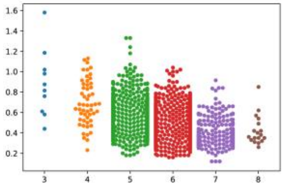


2. Data pre-processing
3. Data representation



10	20	30	40	50	60	70	80	90	100



4.
5.
6.

Deep learning

7. Model evaluation
8. Deployment

Machine learning?

Machine learning?

Introduction to machine learning for biomedical applications

Why using machine learning in biomedical research?

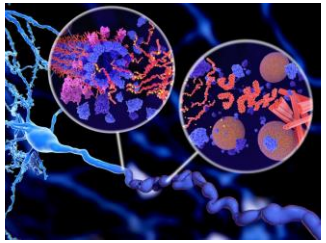
- Detailed analysis: machine learning methods can consider subtle quantitative variables that may be important to identify and/or predict the course of a disease or its treatment.

- Computational analysis: machine learning methods can consider large amounts of data and identify complex relationships between them to enable repetitive and reliable analysis.

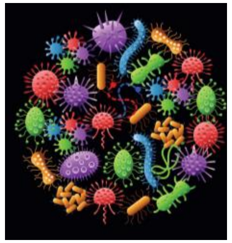
Most common types of data

Omics

Genomics



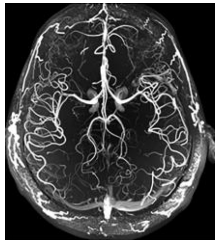
Proteomics



Microbiomics

Healthcare data

EHR



Imaging

Physiological signals

Representing biomedical information

- Molecular information
- Cell information
- Tissue information
- Patient information
- Population information

Representing biomedical information

- Molecular information

  - Genomics: studies DNA molecules with two strands containing the genetic information coded as sequences of adenine, thymine, guanine and cytosine.
    - Structural: sequencing and mapping.
    - Functional: gene expression and their function – transcription, translation and interactions

  - Transcriptomics: studies RNA transcripts produced by the genome and how they are affected by factors such as environment, drugs, etc.

  - Proteomics: studies the structure and function of the proteome (set of all proteins).

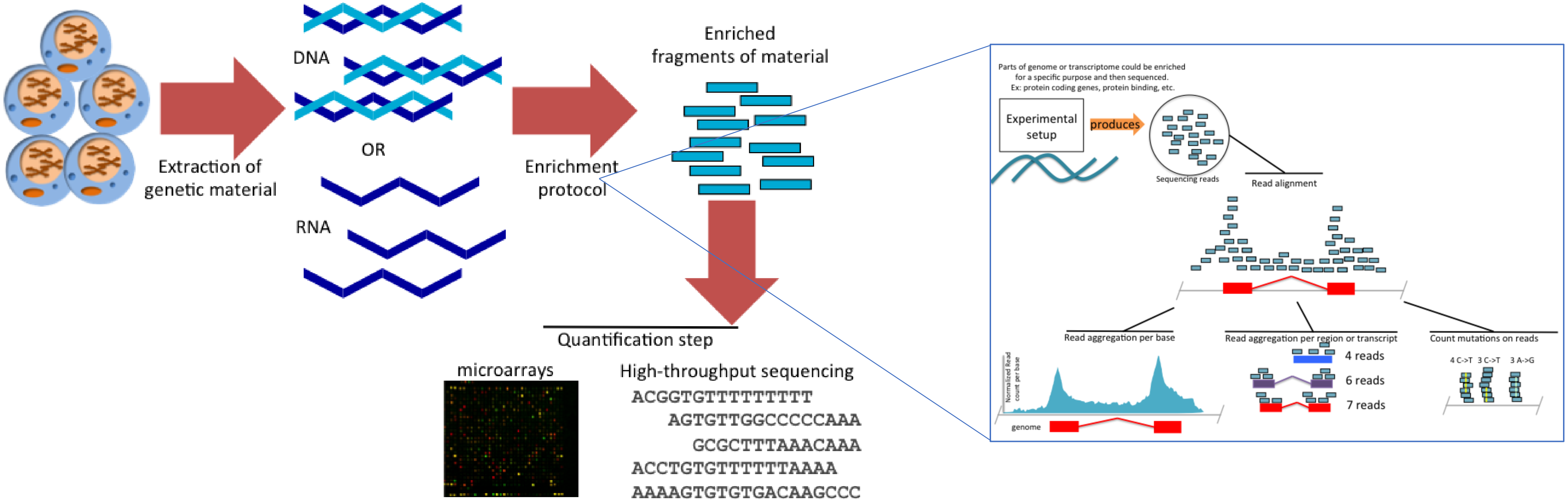
  - Epigenomics: studies the epigenetic modifications of genetic materials (e.g., DNA methylation, histone modification).

  - Others: lipidomics, glycomics, metabolomics…

Representing biomedical information

- Molecular information

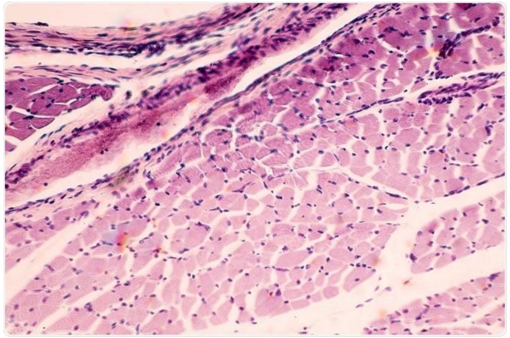
Which one is the data?
Every step depends on previous one



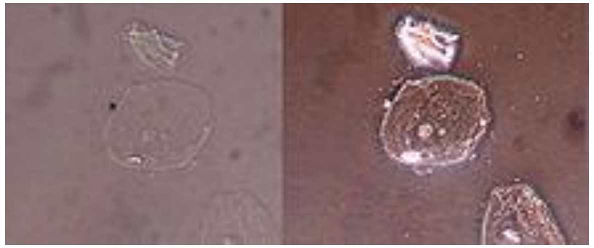
Representing biomedical information

- Cell and tissue information:
  - Highly driven by microscopy imaging

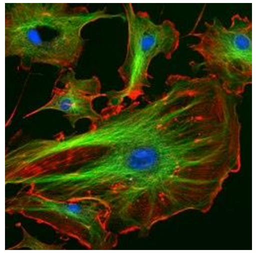
Optical microscopy



Phase contrast microscopy

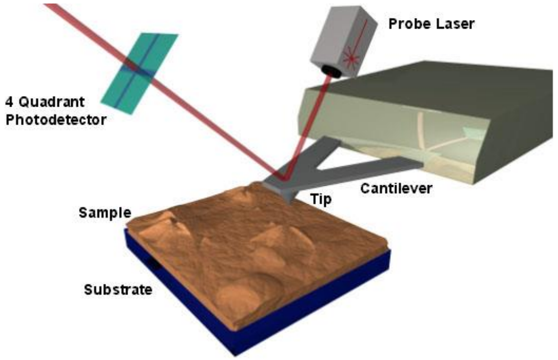


Fluorescent microscopy



Scanning electron microscopy

Atomic forces microscopy



Representing biomedical information

- Patient information:
  - Anatomical
    - Imaging: computed tomography, magnetic resonance, etc.
  - Functional:
    - Blood tests
    - Measured signals: electrocardiogram, electroencephalogram, electromyogram, etc.
  - Other
    - Electronic health records: demographic, symptoms and history data

Image or signal data is not the same than image- and signal-derived data

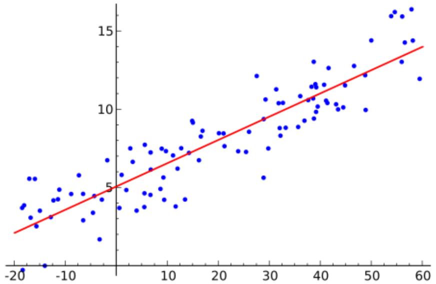
- Population data:
  - Any of all previous information from many individuals
  - Survey data

Main goals in machine learning

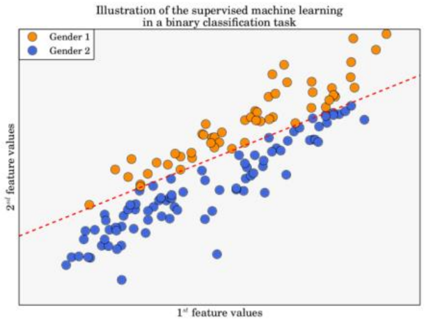
Supervised learning

Goal: make predictions

Data: labeled



Regression

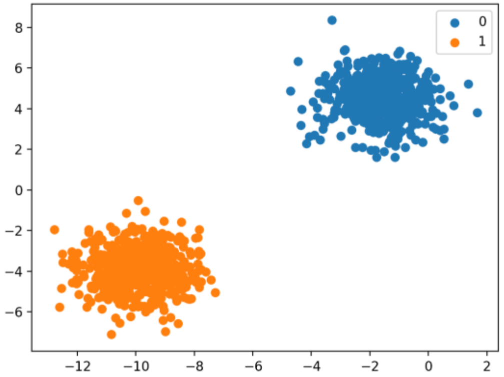


Classification

Unsupervised learning

Goal: find structure

Data: unlabeled



Clustering

Dimensionality reduction

Reinforcement learning

Goal: improve actions from feedback

Next class

- Have a Python 3 installation

- Jupyter could be handy in the first few weeks (at your own risk).

- Visual Studio Code is recommended.

- Install:
    - Pandas
    - Numpy
    - Scipy
    - Scikit-learn
    - Matplotlib
    - Statsmodels
    - Xlsxwriter

- “Play” with Numpy (data representation and matrix operations), Pandas (data I/O and representation) and Shelve (model and experiment persistent storage).