# An Introduction to Machine Learning

Fabio González, PhD

MindLAB Research Group - Universidad Nacional de Colombia

Introducción a los Sistemas Inteligentes

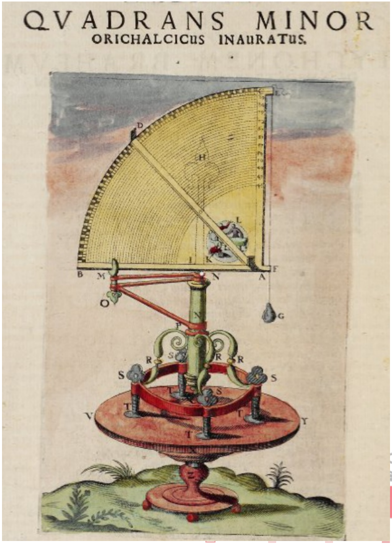
# Outline

1. Introduction
2. Machine learning
   - What's machine learning
   - History
   - Supervised learning
   - Non-supervised learning
3. The machine learning process
   - Model learning
   - Model evaluation
   - Feature extraction
   - Model application

# Observation and analysis

# Tycho Brahe

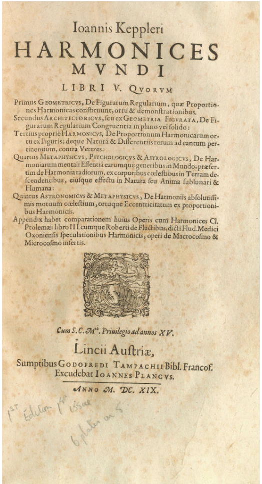

TYCHONIS BRAHE
ASTRONOMIÆ
INSTAURATÆ
MECHANICA
WANDESBURGI
ANNO


QVADRANS MINOR
ORICHALCICUS INAURATUS.

# Tycho Brahe

| | Date, Old Style | | | Longitude | | | | | | Latitude | | | Mean Longitude | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Year | Day | Month | H | M | D | M | S | Sign | D | M | | S | D | M | S |
| I | 1580 | 18 | November | 1 | 31 | 6 | 28 | 35 | Gemeni | 1 | 40 | N. | 1 | 25 | 49 | 31 |
| II | 1582 | 28 | December | 3 | 58 | 16 | 55 | 30 | Cancer | 4 | 6 | N. | 3 | 9 | 24 | 55 |
| III | 1585 | 30 | January | 19 | 14 | 21 | 36 | 10 | Leo | 4 | 32 | N. | 4 | 20 | 8 | 9 |
| IV | 1587 | 6 | March | 7 | 23 | 25 | 43 | 0 | Virgo | 3 | 41 | N. | 6 | 0 | 47 | 40 |
| V | 1589 | 14 | April | 6 | 23 | 4 | 23 | 0 | Scorpio | 1 | 12 | N. | 7 | 14 | 18 | 26 |
| VI | 1591 | 8 | Jume | 7 | 43 | 26 | 43 | 0 | Sagitt. | 4 | 0 | S. | 9 | 5 | 43 | 55 |
| VII | 1593 | 25 | August | 17 | 27 | 12 | 16 | 0 | Pisces | 6 | 2 | S. | 11 | 9 | 49 | 31 |
| VIII | 1595 | 31 | October | 0 | 39 | 17 | 31 | 40 | Taurus | 0 | 8 | N. | 1 | 9 | 55 | 4 |
| IX | 1597 | 13 | December | 15 | 44 | 2 | 28 | 0 | Cancer | 3 | 33 | N. | 2 | 23 | 11 | 56 |
| X | 1600 | 18 | January | 14 | 2 | 8 | 38 | 0 | Leo | 4 | 30 | N. | 4 | 4 | 35 | 50 |
| XI | 1602 | 20 | Febuary | 14 | 13 | 12 | 27 | 0 | Virgo | 4 | 10 | N. | 5 | 14 | 59 | 37 |
| XII | 1604 | 28 | March | 16 | 23 | 18 | 37 | 10 | Libra | 2 | 26 | N. | 6 | 27 | 0 | 12 |

# Johannes Kepler



Ioannis Keppleri

HARMONICES MVNDI

LIBRI V. QVORVM

Primus GEOMETRICVS, De Figurarum Regularium, quæ Proportiones Harmonicas conſtituunt, ortu & demonſtrationibus.

Secundus ARCHITECTONICVS, ſeu ex GEOMETRIA FIGVRATA, De Figurarum Regularium Congruentia in plano vel ſolido:

Tertius propriè HARMONICVS, De Proportionum Harmonicarum ortu ex Figuris; deque Naturâ & Differentiis rerum ad cantum pertinentium, contra Veteres:

Quartus METAPHYSICVS, PSYCHOLOGICVS & ASTROLOGICVS, De Harmoniarum mentali Eſſentiâ earumque generibus in Mundo; præſertim de Harmonia radiorum, ex corporibus cœleſtibus in Terram deſcendentibus, eiuſque effectu in Natura ſeu Anima ſublunari & Humana:

Quintus ASTRONOMICVS & METAPHYSICVS, De Harmoniis abſolutiſſimis motuum cœleſtium, ortuque Eccentricitatum ex proportionibus Harmonicis.

Appendix habet comparationem huius Operis cum Harmonices Cl. Ptolemæi libro III. cumque Roberti de Fluctibus, dicti Flud. Medici Oxonienſis ſpeculationibus Harmonicis, operi de Macrocoſmo & Microcoſmo inſertis.

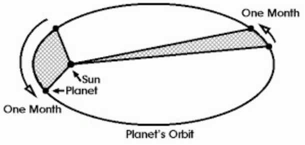
Cum S.C. M^tis. Priuilegio ad annos XV.

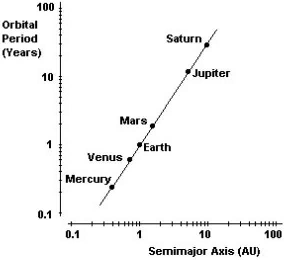
Lincii Auſtriæ,

Sumptibus GODOFREDI TAMPACHII Bibl. Francof.
Excudebat IOANNES PLANCVS.

ANNO M. DC. XIX.

1st Edition 1st issue



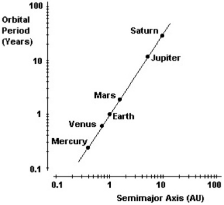



# Data and models

Data

| | Date, Old Style | | | Longitude | | | | | | Latitude | | | Mean Longitude | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Year | Day | Month | H | M | D | M | S | Sign | D | M | | S | D | M | S |
| I | 1580 | 18 | November | 1 | 31 | 6 | 28 | 35 | Gemeni | 1 | 40 | N. | 1 | 25 | 49 | 31 |
| II | 1582 | 28 | December | 3 | 58 | 16 | 55 | 30 | Cancer | 4 | 6 | N. | 3 | 9 | 24 | 55 |
| III | 1585 | 30 | January | 19 | 14 | 21 | 36 | 10 | Leo | 4 | 32 | N. | 4 | 20 | 8 | 9 |
| IV | 1587 | 6 | March | 7 | 23 | 25 | 43 | 0 | Virgo | 3 | 41 | N. | 6 | 0 | 47 | 40 |
| V | 1589 | 14 | April | 6 | 23 | 4 | 23 | 0 | Scorpio | 1 | 12 | N. | 7 | 14 | 18 | 26 |
| VI | 1591 | 8 | Jume | 7 | 43 | 26 | 43 | 0 | Sagitt. | 4 | 0 | S. | 9 | 5 | 43 | 55 |
| VII | 1593 | 25 | August | 17 | 27 | 12 | 16 | 0 | Pisces | 6 | 2 | S. | 11 | 9 | 49 | 31 |
| VIII | 1595 | 31 | October | 0 | 39 | 17 | 31 | 40 | Taurus | 0 | 8 | N. | 1 | 9 | 55 | 4 |
| IX | 1597 | 13 | December | 15 | 44 | 2 | 28 | 0 | Cancer | 3 | 33 | N. | 2 | 23 | 11 | 56 |
| X | 1600 | 18 | January | 14 | 2 | 8 | 38 | 0 | Leo | 4 | 30 | N. | 4 | 4 | 35 | 50 |
| XI | 1602 | 20 | Febuary | 14 | 13 | 12 | 27 | 0 | Virgo | 4 | 10 | N. | 5 | 14 | 59 | 37 |
| XII | 1604 | 28 | March | 16 | 23 | 18 | 37 | 10 | Libra | 2 | 26 | N. | 6 | 27 | 0 | 12 |

Model



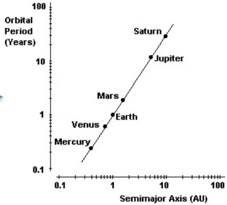
# Machine Learning

Data

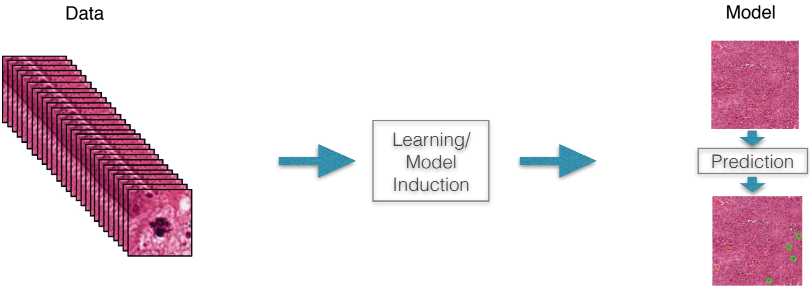
| | Date, Old Style | | | Longitude | | | | | | Latitude | | | Mean Longitude | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Year | Day | Month | H | M | D | M | S | Sign | D | M | | S | D | M | S |
| I | 1580 | 18 | November | 1 | 31 | 6 | 28 | 35 | Gemeni | 1 | 40 | N. | 1 | 25 | 49 | 31 |
| II | 1582 | 28 | December | 3 | 58 | 16 | 55 | 30 | Cancer | 4 | 6 | N. | 3 | 9 | 24 | 55 |
| III | 1585 | 30 | January | 19 | 14 | 21 | 36 | 10 | Leo | 4 | 32 | N. | 4 | 20 | 8 | 9 |
| IV | 1587 | 6 | March | 7 | 23 | 25 | 43 | 0 | Virgo | 3 | 41 | N. | 6 | 0 | 47 | 40 |
| V | 1589 | 14 | April | 6 | 23 | 4 | 23 | 0 | Scorpio | 1 | 12 | N. | 7 | 14 | 18 | 26 |
| VI | 1591 | 8 | Jume | 7 | 43 | 26 | 43 | 0 | Sagitt. | 4 | 0 | S. | 9 | 5 | 43 | 55 |
| VII | 1593 | 25 | August | 17 | 27 | 12 | 16 | 0 | Pisces | 6 | 2 | S. | 11 | 9 | 49 | 31 |
| VIII | 1595 | 31 | October | 0 | 39 | 17 | 31 | 40 | Taurus | 0 | 8 | N. | 1 | 9 | 55 | 4 |
| IX | 1597 | 13 | December | 15 | 44 | 2 | 28 | 0 | Cancer | 3 | 33 | N. | 2 | 23 | 11 | 56 |
| X | 1600 | 18 | January | 14 | 2 | 8 | 38 | 0 | Leo | 4 | 30 | N. | 4 | 4 | 35 | 50 |
| XI | 1602 | 20 | Febuary | 14 | 13 | 12 | 27 | 0 | Virgo | 4 | 10 | N. | 5 | 14 | 59 | 37 |
| XII | 1604 | 28 | March | 16 | 23 | 18 | 37 | 10 | Libra | 2 | 26 | N. | 6 | 27 | 0 | 12 |

Learning/
Model
Induction

Model



# Machine Learning with Images

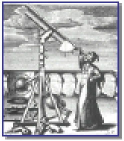


# The fourth paradigm

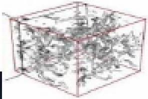
## Emergence of a Fourth Research Paradigm

1. Thousand years ago – **Experimental Science**
   – Description of natural phenomena
2. Last few hundred years – **Theoretical Science**
   – Newton's Laws, Maxwell's Equations…
3. Last few decades – **Computational Science**
   – Simulation of complex phenomena
4. Today – **Data-Intensive Science**
   – Scientists overwhelmed with data sets from many different sources
   - Data captured by instruments
   - Data generated by simulations
   - Data generated by sensor networks

➤ eScience is the set of tools and technologies to support data federation and collaboration
- For analysis and data mining
- For data visualization and exploration
- For scholarly communication and dissemination



$$\left(\frac{\dot a}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



(With thanks to Jim Gray)

# Machine Learning

- Construction and study of systems that can learn from data
- Main problem: to find patterns, relationships, regularities among data, which allow to build descriptive and predictive models.
- Related fields:
    - Statistics
    - Pattern recognition and computer vision
    - Data mining and knowledge discovery
    - Data analytics

# Machine Learning

- Construction and study of systems that can learn from data
- Main problem: to find patterns, relationships, regularities among data, which allow to build descriptive and predictive models.
- Related fields:
    - Statistics
    - Pattern recognition and computer vision
    - Data mining and knowledge discovery
    - Data analytics

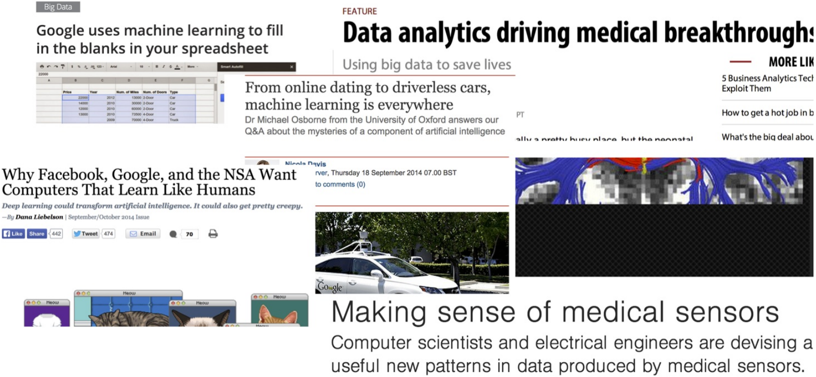
# Machine Learning

- Construction and study of systems that can learn from data
- Main problem: to find patterns, relationships, regularities among data, which allow to build descriptive and predictive models.
- Related fields:
    - Statistics
    - Pattern recognition and computer vision
    - Data mining and knowledge discovery
    - Data analytics

# Brief history

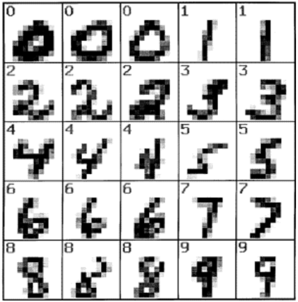
- Fisher's linear discriminant (Fisher, 1936)
- Artificial neuron model (MCCulloch and Pitts, 1943)
- Perceptron (Rosenblatt, 1957) (Minsky&Papert, 1969)
- Probably approximately correct learning (Valiant, 1984)
- Multilayer perceptron and back propagation (Rumelhart et al., 1986)
- Decision trees (Quinlan, 1987)
- Bayesian networks (Pearl, 1988)
- Support vector machines (Cortes&Vapnik, 1995)
- Efficient MLP learning, deep learning (Hinton et al., 2007)

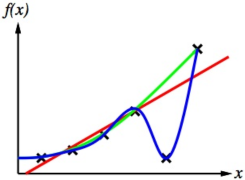
# Machine Learning in the news



# Supervised learning

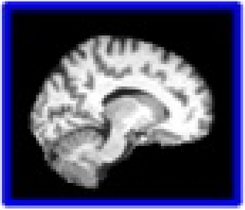
- **Fundamental problem**: to find a function that relates a set of inputs with a set of outputs
- Typical problems:
  - Classification
  - Regression

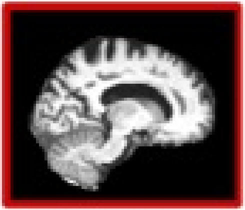


# Supervised learning

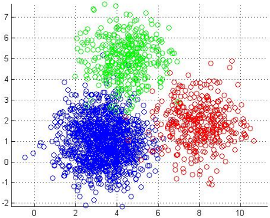
- **Fundamental problem**: to find a function that relates a set of inputs with a set of outputs
- Typical problems:
  - Classification
  - Regression





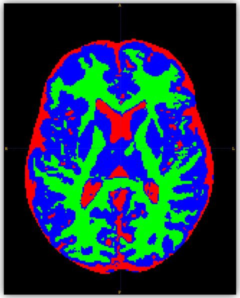
# Non-supervised learning

- There are not labels for the training samples
- **Fundamental problem**: to find the subjacent structure of a training data set
- Typical problems: clustering, segmentation, dimensionality reduction, latent topic analysis
- Some samples may have labels, in that case it is called semi-supervised learning

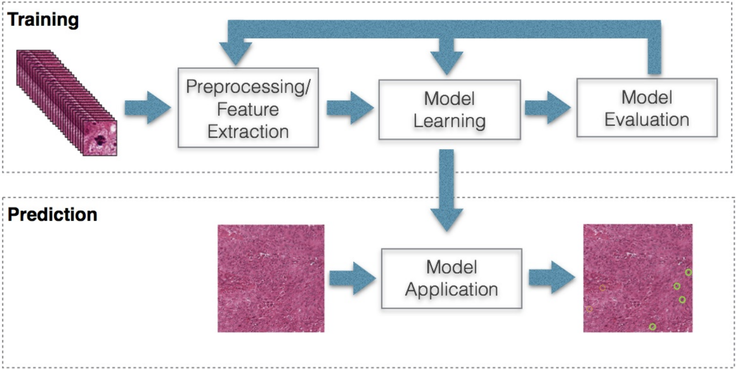


# Non-supervised learning

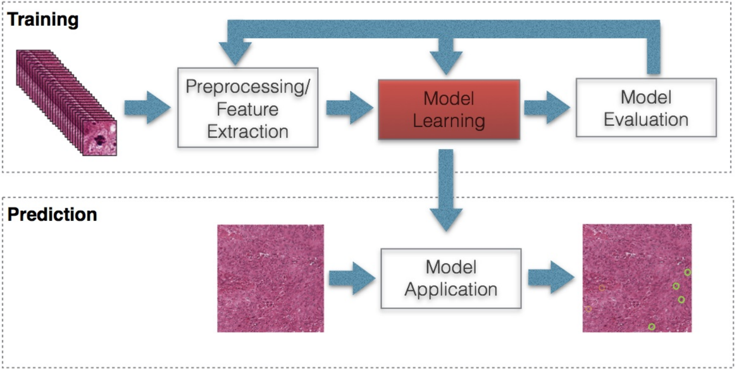
- There are not labels for the training samples
- **Fundamental problem**: to find the subjacent structure of a training data set
- Typical problems: clustering, segmentation, dimensionality reduction, latent topic analysis
- Some samples may have labels, in that case it is called semi-supervised learning



# The machine Learning process



# Model learning



# Model induction from data

- Learning is an *ill-posed* problem (more than one possible solution for the same particular problem, solutions are sensitive to small changes on the problem)
- It is necessary to make additional assumptions about the kind of pattern that we want to learn
- **Hypothesis space**: set of valid patterns that can be learnt by the learning algorithm
- Occam's razor: "All things being equal, the simplest solution tends to be the best one."

# Model induction from data

- Learning is an *ill-posed* problem (more than one possible solution for the same particular problem, solutions are sensitive to small changes on the problem)
- It is necessary to make additional assumptions about the kind of pattern that we want to learn
- **Hypothesis space**: set of valid patterns that can be learnt by the learning algorithm
- Occam's razor: "All things being equal, the simplest solution tends to be the best one."

# Model induction from data

- Learning is an *ill-posed* problem (more than one possible solution for the same particular problem, solutions are sensitive to small changes on the problem)
- It is necessary to make additional assumptions about the kind of pattern that we want to learn
- **Hypothesis space**: set of valid patterns that can be learnt by the learning algorithm
- Occam's razor: "All things being equal, the simplest solution tends to be the best one."

# Model induction from data

- Learning is an *ill-posed* problem (more than one possible solution for the same particular problem, solutions are sensitive to small changes on the problem)
- It is necessary to make additional assumptions about the kind of pattern that we want to learn
- **Hypothesis space**: set of valid patterns that can be learnt by the learning algorithm
- Occam's razor: "All things being equal, the simplest solution tends to be the best one."

# Approaches to learning

- Probabilistic:
  - Generative models: model $P(Y,X)$
  - Discriminative models: model $P(Y|X)$
- Geometrical:
  - Manifold learning: model the geometry of the space where the data lives
  - Max margin learning: model the separation between the classes
- Optimization:
  - Energy/loss/risk minimization

# Approaches to learning

- Probabilistic:
  - Generative models: model $P(Y, X)$
  - Discriminative models: model $P(Y|X)$
- Geometrical:
  - Manifold learning: model the geometry of the space where the data lives
  - Max margin learning: model the separation between the classes
- Optimization:
  - Energy/loss/risk minimization

# Approaches to learning

- Probabilistic:
  - Generative models: model $P(Y, X)$
  - Discriminative models: model $P(Y|X)$
- Geometrical:
  - Manifold learning: model the geometry of the space where the data lives
  - Max margin learning: model the separation between the classes
- Optimization:
  - Energy/loss/risk minimization

# Learning as optimization

- General optimization problem:

$$\min_{f \in H} L(f, D),$$

with $H$:hypothesis space, $D$:training data, $L$:loss/error

- Example, logistic regression:
  - Hypothesis space:

$$y(x) = P(C_+|x) = \sigma(w^T x)$$

  - Cross-entropy error:

$$E(w) = -\ln p(\mathbf{t}|w) = -\sum_{n=1}^{\ell} [t_n \ln y_n + (1 - t_n) \ln(1 - y_n)]$$

# Learning as optimization

- General optimization problem:

$$\min_{f \in H} L(f, D),$$

with $H$:hypothesis space, $D$:training data, $L$:loss/error

- Example, logistic regression:
  - Hypothesis space:

$$y(x) = P(C_+|x) = \sigma(w^T x)$$

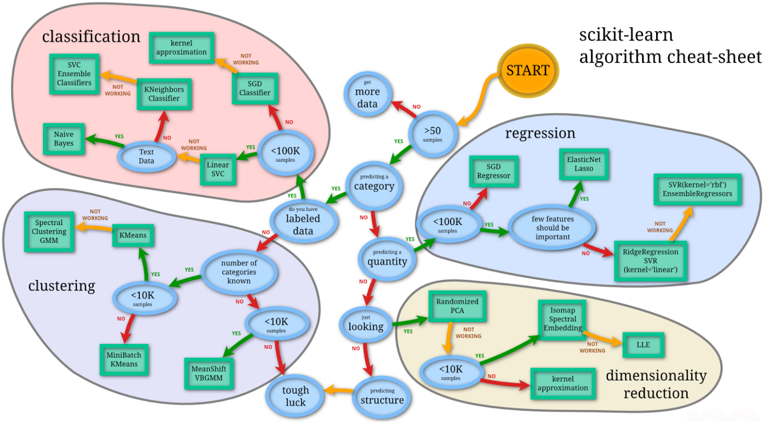
  - Cross-entropy error:

$$E(w) = -\ln p(\mathbf{t}|w) = -\sum_{n=1}^{\ell} [t_n \ln y_n + (1 - t_n)\ln(1 - y_n)]$$

# Methods

- Supervised generative:
  - Naïve Bayes
  - Graphical models
  - Markov random fields
  - Hidden markov models
- Supervised discriminative:
  - Logistic regression
  - Ridge regression
  - Conditional random fields
- Supervised geometrical
  - Max margin classification (SVM)
  - $k$-nearest neighbors
- Non-supervised generative:
  - Latent semantic analysis
  - Latent Dirichlet allocation
  - Gaussian mixtures
- non-supervised geometrical:
  - $k$-means
  - PCA
  - Manifold learning
- Other
  - Neural networks (deep learning)
  - Decision tress
  - Association rules

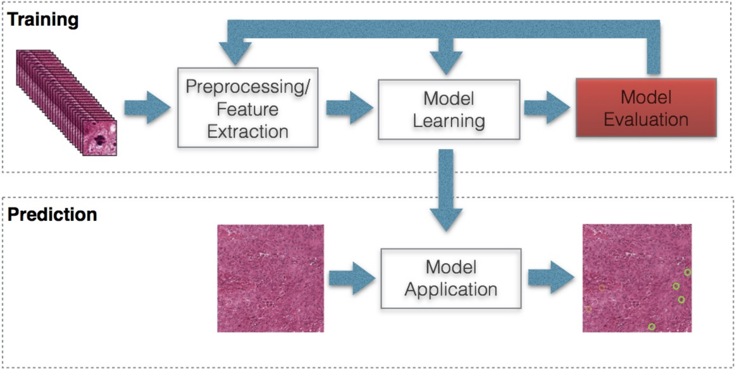
# Methods


classification
scikit-learn
algorithm cheat-sheet
START
kernel approximation
SVC Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
Naive Bayes
Text Data
Linear SVC
<100K samples
get more data
>50 samples
predicting a category
do you have labeled data
regression
SGD Regressor
ElasticNet Lasso
SVR(kernel='rbf') EnsembleRegressors
few features should be important
RidgeRegression SVR (kernel='linear')
predicting a quantity
clustering
Spectral Clustering GMM
KMeans
number of categories known
<10K samples
MiniBatch KMeans
MeanShift VBGMM
just looking
Randomized PCA
Isomap Spectral Embedding
LLE
kernel approximation
dimensionality reduction
tough luck
predicting structure
YES
NO
NOT WORKING

# Strategies

- Optimization (non-linear, convex, etc)
- Stochastic gradient descent
- Kernel methods
- Maximum likelihood estimation
- Maximum a posteriori estimation
- Bayesian estimation (variational learning, Gaussian processes)
- Expectation maximization
- Maximum entropy models
- Sampling (Markov Chain Monte Carlo, particle filtering)

# Evaluation


Training
Preprocessing/
Feature
Extraction
Model
Learning
Model
Evaluation
Prediction
Model
Application

# Training error vs generalization error

- Training error:

$$\sum_{i=1}^{\ell} L(f_w, S_i)$$

- Generalization error:

$$E[(L(f_w, S)]$$

# Training error vs generalization error

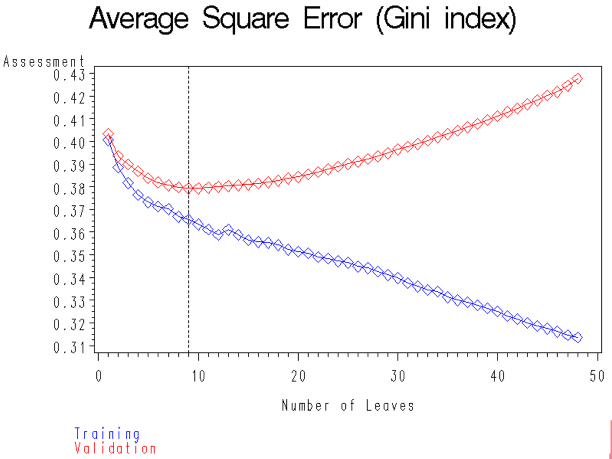
- Training error:

$$\sum_{i=1}^{\ell} L(f_w, S_i)$$

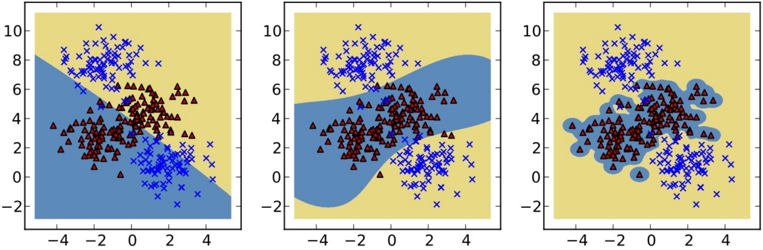
- Generalization error:

$$E[(L(f_w, S)]$$

# Cross validation



# Overfitting and underfitting

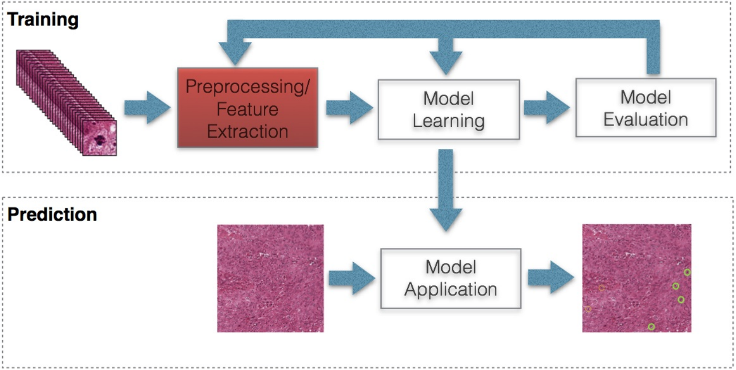


# Regularization

$$\min_{w} -C \overbrace{\sum_{n=1}^{\ell} [t_n \ln y_n + (1-t_n)\ln(1-y_n)]}^{\text{original objective function}} + \overbrace{\|w\|^2}^{\text{regularizer}}$$

- Controls the complexity of a learned model
- Usually, the regularization term corresponds to a norm of the parameter vector ($L_1$ or $L_2$ the most common)
- In some cases, it is equivalent to the inclusion of a prior and finding a MAP solution.

# Feature extraction



# Features

- Features represent our prior knowledge of the problem
- Depend on the type of data
- Specialized features for practically any kind of data (images, video, sound, speech, text, web pages, etc)
- Medical imaging:
  - Standard computer vision features (color, shape, texture, edges, local-global, etc)
  - Specialized features tailored to the problem at hand
- New trend: learning features from data

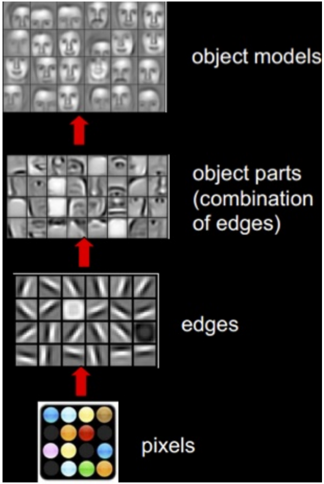
# Features

- Features represent our prior knowledge of the problem
- Depend on the type of data
- Specialized features for practically any kind of data (images, video, sound, speech, text, web pages, etc)
- Medical imaging:
  - Standard computer vision features (color, shape, texture, edges, local-global, etc)
  - Specialized features tailored to the problem at hand
- New trend: learning features from data

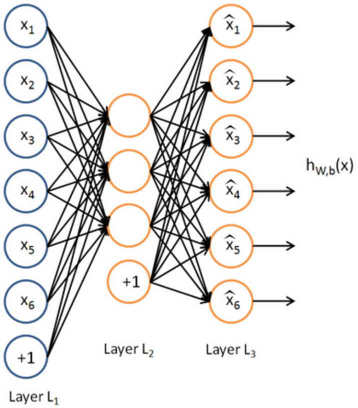
# Features

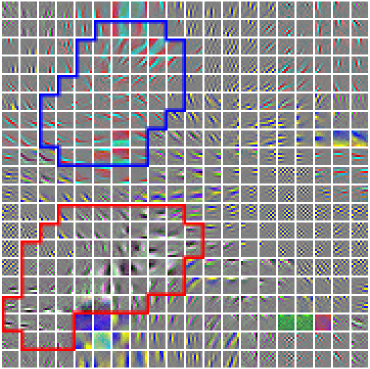
- Features represent our prior knowledge of the problem
- Depend on the type of data
- Specialized features for practically any kind of data (images, video, sound, speech, text, web pages, etc)
- Medical imaging:
  - Standard computer vision features (color, shape, texture, edges, local-global, etc)
  - Specialized features tailored to the problem at hand
- New trend: learning features from data

# Feature learning



# Unsupervised feature learning



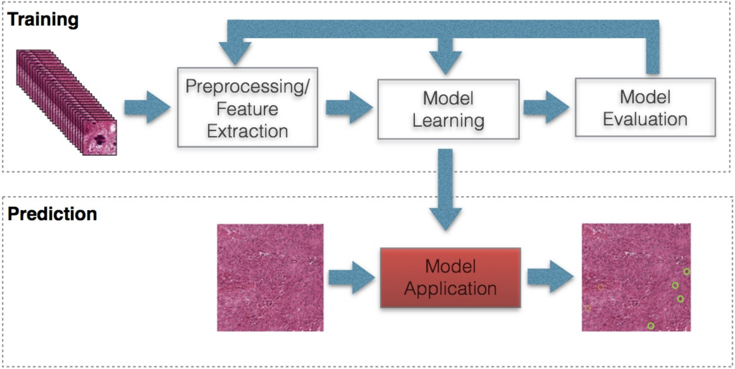



# AMIDA-MICCAI 2013 Challenge

| Team name | Precision | Recall | $F_1$-Score |
|---|---|---|---|
| IDSIA | 0.610 | 0.612 | 0.611 |
| DTU | 0.427 | 0.555 | 0.483 |
| SURREY | 0.357 | 0.332 | 0.344 |
| ISIK | 0.306 | 0.351 | 0.327 |
| PANASONIC | 0.336 | 0.310 | 0.322 |
| CCIPD/MINDLAB | 0.353 | 0.291 | 0.319 |
| WARWICK | 0.171 | 0.552 | 0.261 |
| POLYTECH/UCLAN | 0.186 | 0.263 | 0.218 |
| MINES | 0.139 | 0.490 | 0.217 |
| SHEFFIELD/SURREY | 0.119 | 0.107 | 0.113 |
| SEOUL | 0.032 | 0.630 | 0.061 |
| NTUST | 0.011 | 0.685 | 0.022 |
| UNI-JENA | 0.007 | 0.077 | 0.013 |
| NIH | 0.002 | 0.049 | 0.003 |

→ UFL using deep-leranin
→ Donut like histograms—SVM
→ Shape, color, texture—SVM

# Model application



# High-throughput data analytics

- Large scale machine learning (big-data):
  - Large number of samples
  - Large samples (whole-slide images, 4D high-resolution volumes)
- Scalable learning algorithms (on-line learning)
- Distributed computing architectures (Hadoop, Spark)
- GPGPU computing and multicore architectures

# High-throughput data analytics

- Large scale machine learning (big-data):
  - Large number of samples
  - Large samples (whole-slide images, 4D high-resolution volumes)
- Scalable learning algorithms (on-line learning)
- Distributed computing architectures (Hadoop, Spark)
- GPGPU computing and multicore architectures

# High-throughput data analytics

- Large scale machine learning (big-data):
  - Large number of samples
  - Large samples (whole-slide images, 4D high-resolution volumes)
- Scalable learning algorithms (on-line learning)
- Distributed computing architectures (Hadoop, Spark)
- GPGPU computing and multicore architectures

# High-throughput data analytics

- Large scale machine learning (big-data):
  - Large number of samples
  - Large samples (whole-slide images, 4D high-resolution volumes)
- Scalable learning algorithms (on-line learning)
- Distributed computing architectures (Hadoop, Spark)
- GPGPU computing and multicore architectures

# Questions?

fagonzalezo@unal.edu.co

http://www.mindlaboratory.org