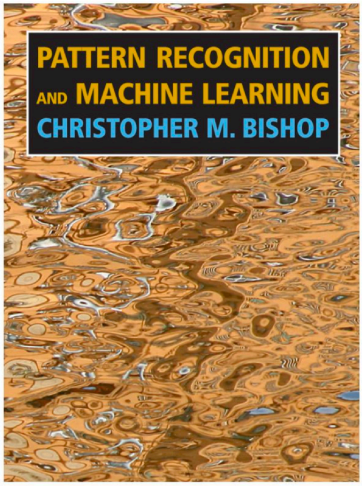
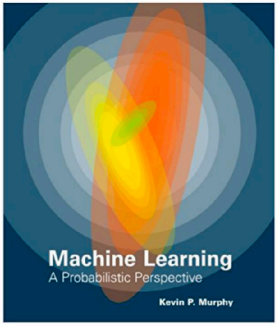
# Introduction to Machine(Deep) Learning

Dr. Alejandro Veloz

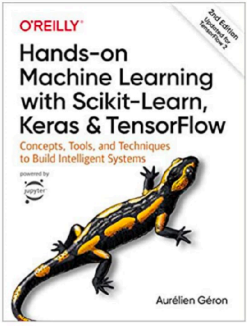
inf395.github.io




Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, 2006.




Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.




Géron, *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow*, O'Reilly, 2nd Edition, 2019.

# What is Machine Learning

**Arthur Samuel (1959):** Field of study that gives computers the ability to learn without being explicitly programmed.

**Shapire:** Machine learning studies how to automatically learn to make predictions based on past observations.

The field of machine learning is concerned with building and understanding systems that can automatically extract information from empirical data to improve their performance.

As a scientific discipline, machine learning is an interdisciplinary (and relatively young) field that focuses both on the theoretical foundations of systems that learn, reason, and act, as well as on the practical applications of these systems.

# Multiple disciplines

- **Statistics:** Inference from data, probabilistic models, learning theory, …
- **Mathematics:** Optimization theory, numerical methods, tools for theory, …
- **Engineering:** Signal processing, system identification, robotics, control, information theory, data-mining, …
- **Computer science:** Artificial intelligence, computer vision, information retrieval, data-structures, implementations, …
- **Economics:** decision theory, operations research, econometrics, …
- **Psychology/Cognitive science:** Computational linguistics, learning, reinforcement learning, movement control, …
- **Physics:** Energy minimization principles, entropy, capacity, …
- **Computational Neuroscience:** Neural networks, principles of neural information processing, …

Frequently information flowing back in from application domains, e.g. tools for bioinformatics getting used in other domains, ...

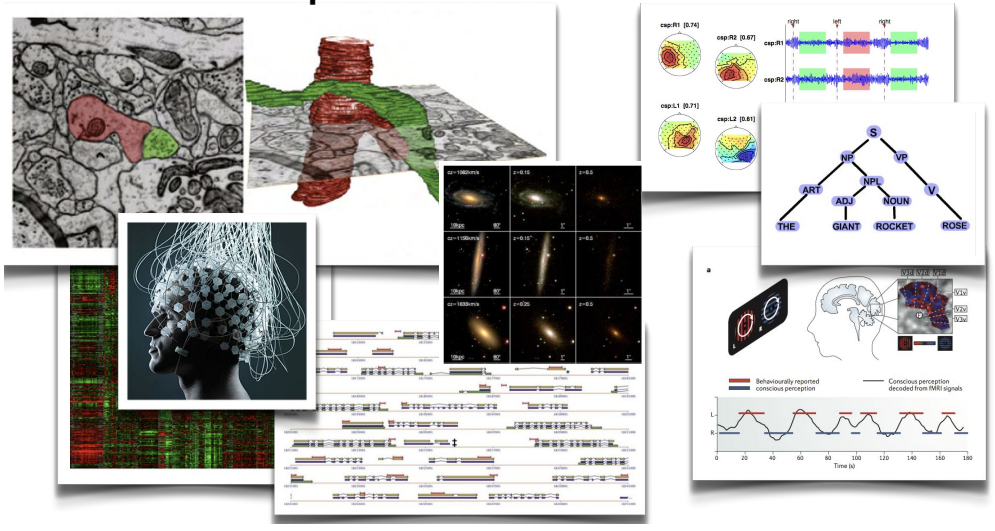
# Machine Learning / Statistical Learning

- We would like to design an algorithm that help us to solve different **prediction problems**.
- The algorithm is designed based on a mathematical model or function and a dataset.
- Extract knowledge from data.

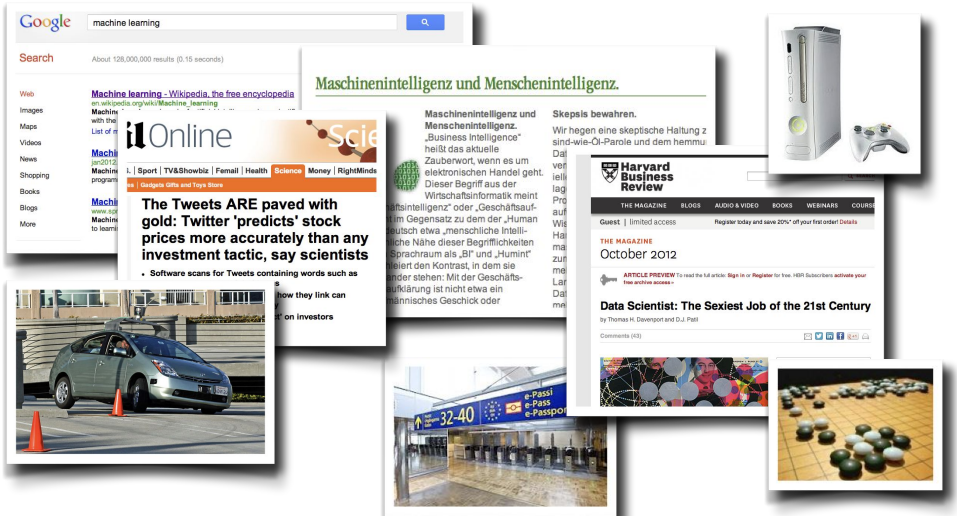
# Main challenges of machine learning

- Insufficient quantity of training data.
- Nonrepresentative training data.
- Poor-quality data.
- Irrelevant features.
- Overfitting the training data.
- Underfitting the training data.

# Varias disciplinas científicas



# … y en la vida cotidiana

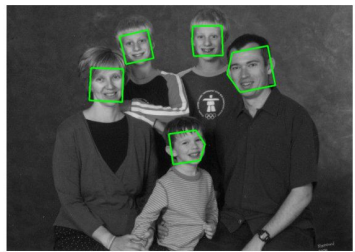


# Examples of ML problems

Handwritten digit recognition

# Examples of ML problems

Face detection and face recognition



From Murphy (2012).

# Examples of ML problems

Predicting the age of a person looking at a particular YouTube video.

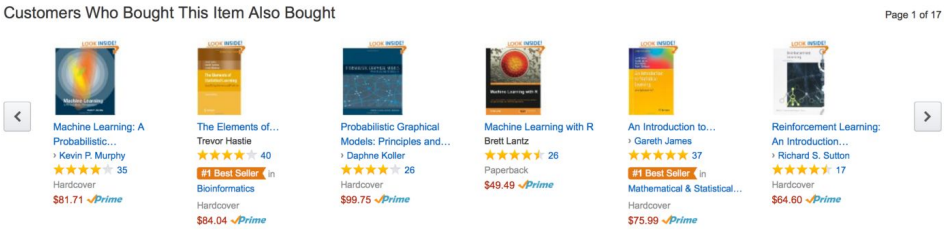
# Examples of ML problems

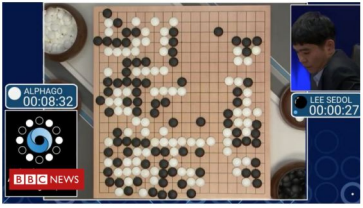
Stock market

# Examples of ML problems

Clustering: segmenting customers in e-commerce

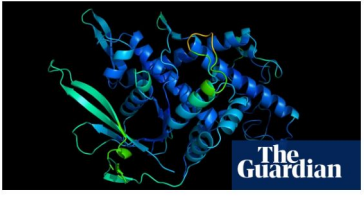
# Recommendation systems






AlphaGo

Autonomous driving




AlphaFold

# Six characteristics of ML problems and solutions

- **Problem class:** What is the nature of the training data and what kinds of queries will be made at testing time?
- **Assumptions:** What do we know about the source of the data or the form of the solution?
- **Evaluation criteria:** What is the goal of the prediction or estimation system? How will the answers to individual queries be evaluated? How will the overall performance of the system be measured?
- **Model type:** Will an intermediate model of the world be made? What aspects of the data will be modeled in different variables/parameters? How will the model be used to make predictions?
- **Model class:** What particular class of models will be used? What criterion will we use to pick a particular model from the model class?
- **Algorithm:** What computational process will be used to fit the model to the data and/or to make predictions?

# Problem class

- Supervised learning:
  - Variable $y$ is discrete: **classification.**
  - Variable $y$ is continuous: **regression.**
- Unsupervised learning:
  - Find similar groups: **clustering.**
  - Find a probability function for $\mathbf{x}$: **density estimation.**
  - Find a lower dimensionality representation for $\mathbf{x}$: **dimensionality reduction and visualization.**
- Other types of learning: reinforcement learning, semi-supervised learning, active learning, multi-task learning.

Problem classes vary according to what kind of data provided and what kind of conclusions to draw from it.

# Supervised learning

Supervised learning is a system where inputs are paired with **known outputs**. The task is to learn the mapping from inputs to outputs.

- **Classification:** Output is from a *small, finite set*.
  - **Binary:** Only two classes.
  - **Multi-class:** More than two classes.
- **Regression:** Output is from a *large, ordered set or continuous set* (real numbers).

**Training set:** a set of $N$ instances and their labels/targets:

$$\mathscr{D}_{\text{train}} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$$

where $\mathbf{x}_i$ is typically a $d$-dimensional vector (input), and $y_i$ is the predicted real-valued output (target).

**Generalization:** ability to correctly predict the label/target $y_{N+1}$ of a new instance $\mathbf{x}_{N+1}$.

# Unsupervised learning

- Given samples

$$\mathscr{D}_{\text{train}} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}, \qquad \mathbf{x}_i \in \mathbb{R}^D$$

  the problem is to **re-represent them** as points in a $d$-dimensional space, where $d < D$.

- Unsupervised learning does not rely on learning a mapping from inputs to outputs using pairs of input and output examples.

- Instead, it involves working with a dataset without explicit labels, with the goal of discovering patterns or underlying structure within the data.

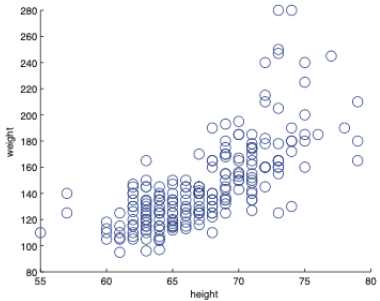
# Unsupervised learning - clustering

Given samples $\mathscr{D}_{\text{train}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_i \in \mathbb{R}^d$, the goal is to find a partitioning (or "clustering'') of the samples that group similar samples.

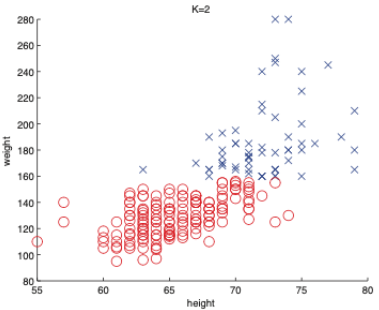
Objectives vary based on sample similarity definitions, aiming to minimize average intra-cluster distance and maximize average inter-cluster distance.

Other methods perform a *soft* clustering, in which samples may be assigned 0.9 membership in one cluster and 0.1 in another.

# Unsupervised learning

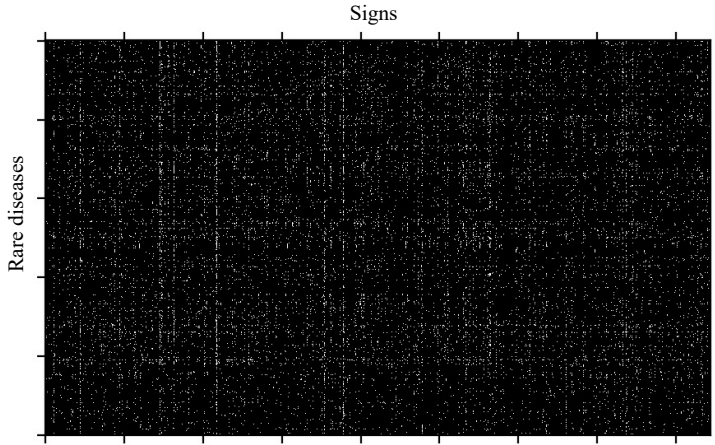



(a)

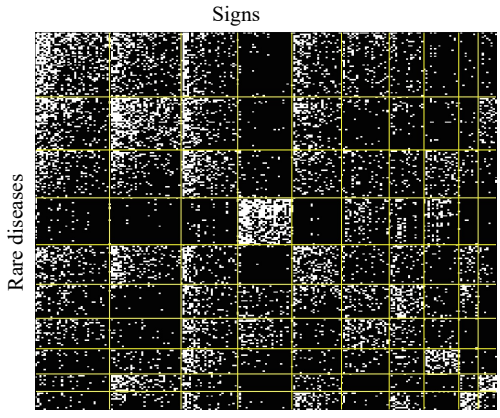


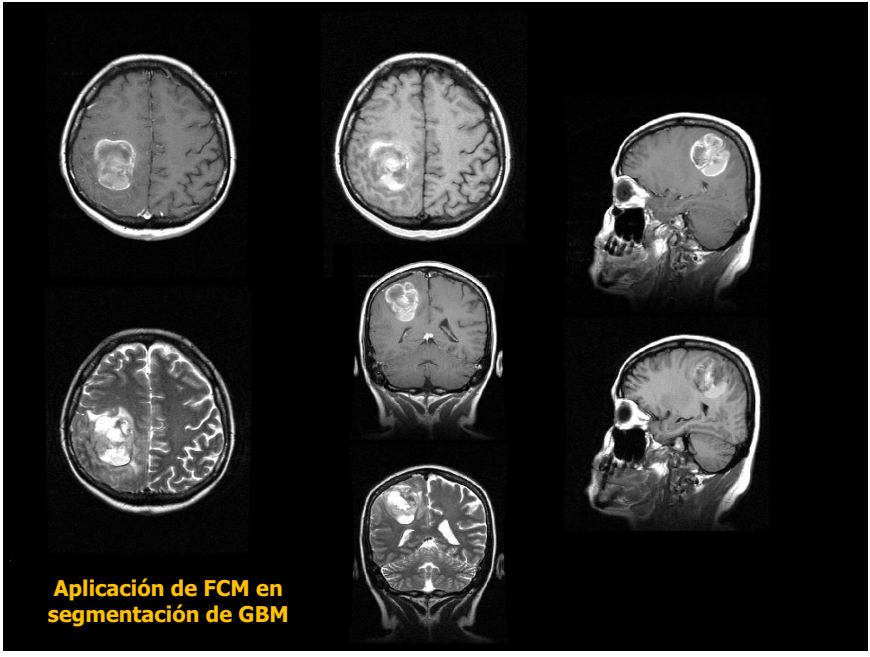

(b)

# Clustering rare diseases example



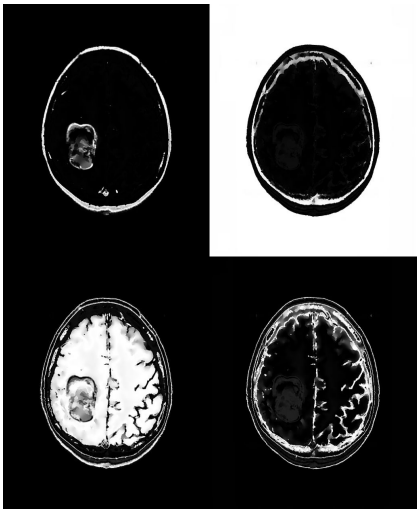
# Clustering rare diseases example




Aplicación de FCM en
segmentación de GBM

# Resultado para cuatro clases visto como imágenes (U)

- Clases: materia blanca, materia gris, líquido cefalorraquídeo – fondo, tumor-grasa,
  C = 4
- Exponente de difusividad,
  m = 2
- Criterio de similitud,
  ε = 10-5
- Nº de iteraciones máximas,
  n = 100



# Unsupervised learning - density estimation

Given samples

$$\mathscr{D}_{\text{train}} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}, \qquad \mathbf{x}_i \in \mathbb{R}^d$$

drawn i.i.d. from some distribution $\Pr(X)$, the goal is to **predict the probability** $\Pr(X = \mathbf{x}_{N+1})$ of an element drawn from the same distribution.

Density estimation sometimes also plays a role as a “subroutine” in the overall learning method for supervised learning.

ICD+

# #188400

# DIGEORGE SYNDROME; DGS

*Alternative titles; symbols*
CHROMOSOME 22q11.2 DELETION SYNDROME
HYPOPLASIA OF THYMUS AND PARATHYROIDS
THIRD AND FOURTH PHARYNGEAL POUCH SYNDROME

Other entities represented in this entry:

## DIGEORGE SYNDROME CHROMOSOME REGION, INCLUDED; DGCR, INCLUDED

TAKAO VCF SYNDROME, INCLUDED
CATCH22, INCLUDED

**Phenotype-Gene Relationships**

| Location | Phenotype | Phenotype MIM number | Inheritance (in progress) | Phenotype mapping key | Gene/Locus | Gene/Locus MIM number |
|---|---|---|---|---|---|---|
| 22q11.21 | DiGeorge syndrome | 188400 | AD | 3 | TBX1 | 602054 |

Clinical Synopsis

### TEXT

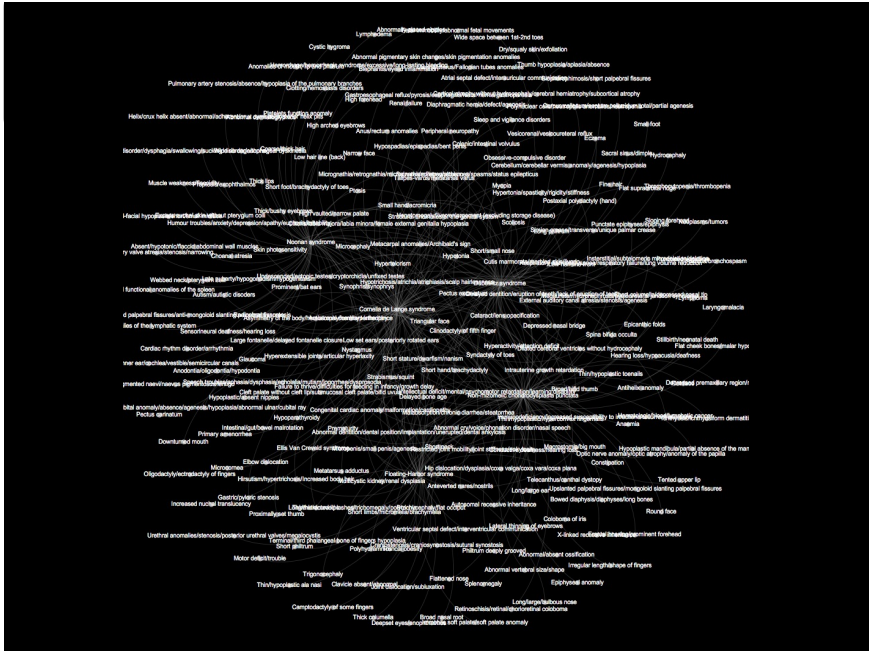
A number sign (#) is used with this entry because DiGeorge syndrome is caused by a 1.5- to 3.0-Mb hemizygous deletion of chromosome 22q11.2. Haploinsufficiency of the TBX1 gene (602054) in particular is responsible for most of the physical malformations. There is evidence that point mutations in the TBX1 gene can also cause the disorder.

### Description

DiGeorge syndrome (DGS) comprises hypocalcemia arising from parathyroid hypoplasia, thymic hypoplasia, and outflow tract defects of the heart. Disturbance of cervical neural crest migration into the derivatives of the pharyngeal arches and pouches can account for the phenotype. Most cases result from a deletion of chromosome 22q11.2 (the DiGeorge syndrome chromosome region, or DGCR). Several genes are lost including the putative transcription factor TUPLE1 which is expressed in the appropriate distribution. This deletion may present with a variety of phenotypes: Shprintzen, or velocardiofacial, syndrome (VCFS; 192430); conotruncal anomaly face (or Takao syndrome); and isolated outflow tract defects of the heart including tetralogy of Fallot, truncus arteriosus, and interrupted aortic arch. A collective acronym CATCH22 has been proposed for these differing presentations. A small number of cases of DGS have defects in other chromosomes, notably 10p13 (see 601362). In the mouse, a transgenic Hox A3 (Hox 1.5) knockout produces a phenotype similar to DGS as do the teratogens retinoic acid and alcohol.

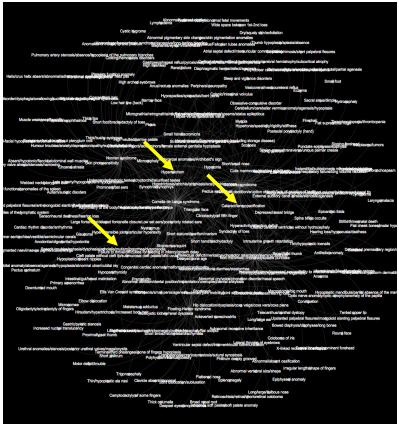
### Nomenclature

DiGeorge syndrome overlaps clinically with the disorder described by the Japanese as 'conotruncal anomaly face syndrome' (Kinouchi et al., 1976; Takao et al., 1980; Shimizu et al., 1984), where the cardiovascular presentation is the focus of attention. The term conotruncal anomaly face syndrome is cumbersome and has the disadvantage of using embryologic assumptions as a title. It would be appropriate to use Takao syndrome for those cases with a preponderant cardiac presentation in contrast to the low T cell and hypocalcemic presentation in infancy of DiGeorge syndrome and the craniofacial and palatal abnormalities typical of Shprintzen syndrome. These 3 phenotypes may be seen in the same family and most cases of all 3 categories have been shown to have a 22q11 deletion. This led Wilson et al. (1993) to propose the acronym CATCH22 (Cardiac Abnormality/abnormal facies, T cell deficit due to thymic hypoplasia, Cleft palate, Hypocalcemia due to hypoparathyroidism resulting from 22q11 deletion) as a collective acronym for those with the common genetic etiology. Shprintzen (1994) objected to 'lumping' velocardiofacial syndrome with the DiGeorge anomaly, arguing that there is 'no valid evidence to suggest that velocardiofacial syndrome is etiologically heterogeneous...[whereas] the DiGeorge anomaly is known to be so.' Hall (1993) cited data of Driscoll et al. (1993) indicating that velocardiofacial syndrome is etiologically heterogeneous. She stated that '...68% of Shprintzen syndrome patients...have been recognised to have deletions of 22q11.' Shprintzen (1994) refuted her statement, maintaining that it could accurately be stated that deletion was found in 68% of patients sent to the Driscoll laboratory with a diagnosis of velocardiofacial syndrome made by other clinicians. Shprintzen (1994) said that in his sample, 100% had deletion.

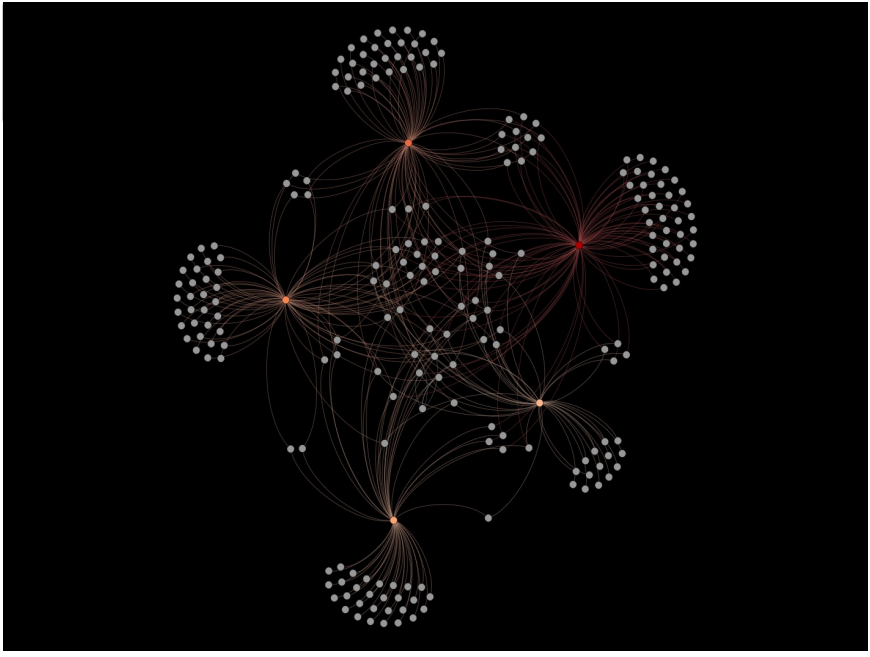

Abnormal fetal movements
Lymphedema
Wide space between 1st-2nd toes
Cystic hygroma
Dry/sparse scalp vessels
Abnormal pigmentary skin changes
Thumb hypoplasia/aplasia/absence
Pulmonary artery stenosis/absence/hypoplasia of the pulmonary branches
Clotting/hematologic disorders
Atrial septal defect/muscular communication
Short palpebral fissures
Sleep and vigilance disorders
Small foot
High arched eyebrows
Arca/rectus anomalies
Peripheral neuropathy
Vesicoureteral reflux
Eczema
Hypospadias/epispadias/bent penis
Intestinal volvulus
Low hair line (back)
Nevus flammeus
Obsessive-compulsive disorder
Sacral sinus/dimple
Hydrocephaly
Cerebellum/cerebellar vermis anomaly/agenesis/hypoplasia
Muscle weakness/hypotonia
Ptosis
Short foot/brachydactyly of toes
Myalgia
Hypertonia/spasticity/rigidity/stiffness
Finger clinodactyly
Thrombocytopenia/thrombopenia
Postaxial polydactyly (hand)
Facial hypoesthesia/hypalgesia
Small hands/micromelia
Scoliosis
Punctate epiphyseal stippling
Short stature
Humour troubles/anxiety/depression/apathy
Microcephaly
Metacarpal anomalies/Archibald's sign
Short/small nose
Absent/hypotonic/flaccid abdominal wall muscles
Hypoplasia
Cutis marmorata
Hypertelorism
Webbed neck/pterygium colli
Synophrys/synophria
Pectus excavatum
Abnormal dentition/eruption of teeth
Lack of vitamin K-dependent coagulation factors
External auditory canal atresia/stenosis/agenesis
Laryngomalacia
Cornelia de Lange syndrome
Triangular face
Cataract/lens opacification
Depressed nasal bridge
Epicanthic folds
Sensorineural deafness/hearing loss
Clinodactyly of fifth finger
Spina bifida occulta
Large fontanelle/delayed fontanelle closure
Low set ears/posteriorly rotated ears
Hyperactivity
Dilatation of cerebral ventricles without hydrocephaly
Flat cheek bones
Cardiac rhythm disorder/arrhythmia
Nystagmus
Hypermetensible joints/articular hypertaxity
Syndactyly of toes
Hearing loss/hypacusis/deafness
Short stature/dwarfism/nanism
Hand/clasp-knife deformity
Strabismus/squint
Short hand/brachydactyly
Intrauterine growth retardation
Thin/hypoplastic toenails
Detached premature
Cleft palate without cleft lip
Abnormal cleft palate/cleft uvula
Hypoplastic/absent nipples
Delayed bone age
Congenital cardiac anomaly
Pectus carinatum
Hypoparathyroidism
Abnormal cry/voice/phonation
Abnormal/delayed speech
Primary amenorrhea
Intestinal gut/bowel malrotation
Abnormal orbital/orbital position/implantation/orientation
Anaemia
Downturned mouth
Ella Van Creveld syndrome
Short/thick/abnormal nail
Micrognathia
Big mouth
Hypoplastic mandible/partial absence of the man
Elbow dislocation
Microdontia
Metatarsus adductus
Hip dislocation/dysplasia/coxa valga/coxa vara
Constipation
Oligodactyly/ectrodactyly of fingers
Hirsutism/hypertrichosis/increased body hair
Floating-Harbor syndrome
Telecanthus/canthal dystocy
Tented upper lip
Gastrocolytic stenosis
Antevorted apres/incertis
Long face
Increased nuclear translucency
Bowed diaphysis/diaphyses/long bones
Round face
Proximal/great thumb
Short limb-hand type dysplasia
Coloboma of iris
Urethral anomalies/stenosis/posterior urethral valves/megacystis
Ventricular septal defect/interventricular communication
X-linked recessive inheritance
Prominent forehead
Terminal/third phalanges/bone of fingers hypoplasia
Short philtrum
Craniosynostosis/sutural synostosis
Polythelia/supernumerary nipples
Philtrum deeply grooved
Motor delay/disturbance
Abnormal/absent ossification
Irregular length/shape of fingers
Trigonocephaly
Abnormal vertebral size/shape
Flattened nose
Thin/hypoplastic ala nasi
Clavicle absent/hypoplastic/abnormal
Epiphyseal anomaly
Long/large/bulbous nose
Camptodactyly of some fingers
Retinoschisis/retinal/chorioretinal coloboma
Thick calcaneus
Broad apical mid
Deepset eyes
Prominent soft palate
Cleft palate anomaly

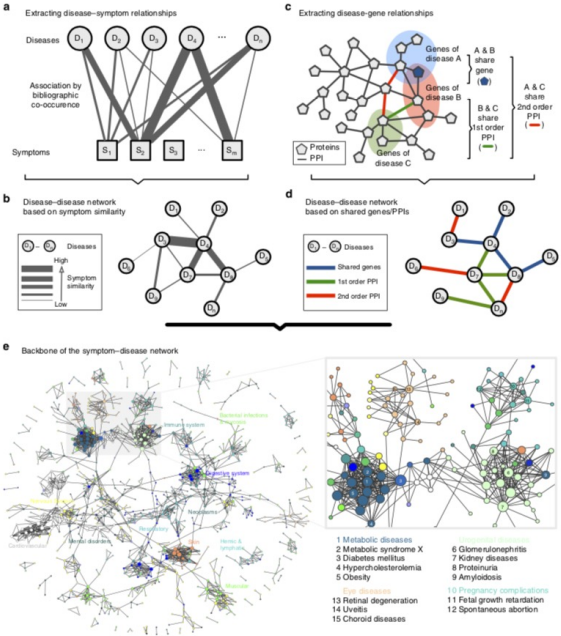
Abnormal fetal movements
Lymphedema
Wide space between 1st-2nd toes
Cystic hygroma
Dry/waxy skin exfoliation
Abnormal pigmentary skin changes/skin pigmentation anomalies
Thumb hypoplasia/aplasia/absence
Pulmonary artery stenosis/absence/hypoplasia of the pulmonary branches
Bleeding/hemostasis disorders
Atrial septal defect/muscular communication
Short palpebral fissures
Renal failure
Diaphragmatic hernia/defect/eventration
Sleep and vigilance disorders
Small foot
Arcus/rectus anomalies
Peripheral neuropathy
Vesicoureteral reflux
Eczema
Hypocapidies/epispadias/bent penis
Intestinal volvulus
Low hair line (back)
Narrow face
Obsessive-compulsive disorder
Sacral sinus/dimple
Hydrocephaly
Cerebellum/cerebellar vermis anomaly/agenesis/hypoplasia
Short foot/brachydactyly of toes
Ptosis
Myalgia
Hypertonia/spasticity/rigidity/stiffness
Finger/toe clinodactyly
Thrombocytopenia/thrombopenia
Postaxial polydactyly (hand)
Small hands/micromicria
Scoliosis
Punctate epiphyseal stippling
Humour troubles/anxiety/depression/apathy
Absent/hypotonic/flaccid/abdominal wall muscles
Short/small nose
Hypoplasia
Webbed neck/pterygium colli
Hypertelorism
Cutis marmorata
Synophrys/hypertrichosis
Pectus carinatum
Dentition/eruption of teeth
Cornelia de Lange syndrome
External auditory canal atresia/stenosis/agenesis
Laryngomalacia
Triangular face
Cataract/lens opacification
Depressed nasal bridge
Epicanthic folds
Sensorineural deafness/hearing loss
Clinodactyly of fifth finger
Spina bifida occulta
Large fontanelle
Low set ears/posteriorly rotated ears
Stillbirth/neonatal death
Nystagmus
Hypertrophic cardiomyopathy
Flat chest bones
Cardiac rhythm disorder/arrhythmia
Hypertrophy of left ventricle without hydrocephaly
Syndactyly of toes
Hearing loss/hypacusis/deafness
Short stature/dwarfism/nanism
Short hand/brachydactyly
Intrauterine growth retardation
Thin/hypoplastic toenails
Strabismus/squint
Detached premature/tiny region/
Cleft palate without cleft lip
Hypoplastic/absent nipples
Antithelia anomaly
Congenital cardiac anomaly
Abnormal cry/voice/phonation disorder
Nasal speech
Anaemia
Primary amenorrhea
Hypoparathyroidism
Intestinal gut/bowel malrotation
Abnormal oriental position/implantation/orientation of teeth/dentition
Downturned mouth
Ella Van Creveld syndrome
Micrognathia
Cystic nares anomaly
Hypoplastic mandible/partial absence of the man
Elbow dislocation
Constipation
Microcornea
Metatarsus adductus
Hip dislocation/dysplasia/coxa valga/coxa varus
Oligodactyly/ectrodactyly of fingers
Hirsutism/hypertrichosis/increased body hair
Floating-Harbor syndrome
Telecanthus/canthal dystocy
Tented upper lip
Antevorted apex/incertitis
Long face
Underdeveloped palpebral fissures/malgoidal slanting palpebral fissures
Gastric/pyloric stenosis
Bowed diaphysis/diaphyses/long bones
Increased nuclear translucency
Autosomal recessive inheritance
Round face
Proximal/great thumb
Coloboma of iris
Urethral anomalies/stenosis/posterior urethral valves/megacystis
Ventricular septal defect/interventricular communication
X-linked recessive inheritance
Prominent forehead
Terminal/third phalanges/bone of fingers hypoplasia
Short philtrum
Polythelia/supernumerary nipples
Philtrum deeply grooved
Abnormal/absent ossification
Motor delay/disturbance
Irregular length/shape of fingers
Trigonocephaly
Abnormal vertebral size/shape
Flattened nose
Thin/hypoplastic ala nasi
Clavicle absent/dysplasia/subluxation
Splenomegaly
Epiphyseal anomaly
Long/large/bulbous nose
Camptodactyly of some fingers
Retinoschisis/retinal/chorioretinal coloboma
Thick calcaneus
Broad nasal root
Deepset eyes
Hypoplastic soft palate/cleft palate anomaly

# The network perspective



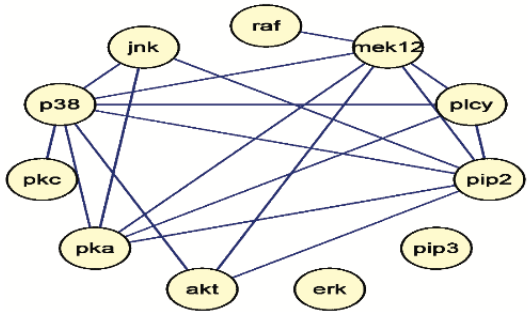
- What are the most probable outcomes?
- What are the most probable signs?
- What are other possible diseases?






Zhou et al.,
Human symptoms-disease network.
Nature Commun.
2014

# Unsupervised learning - causal discovery



# Unsupervised learning - dimensionality reduction

Given samples

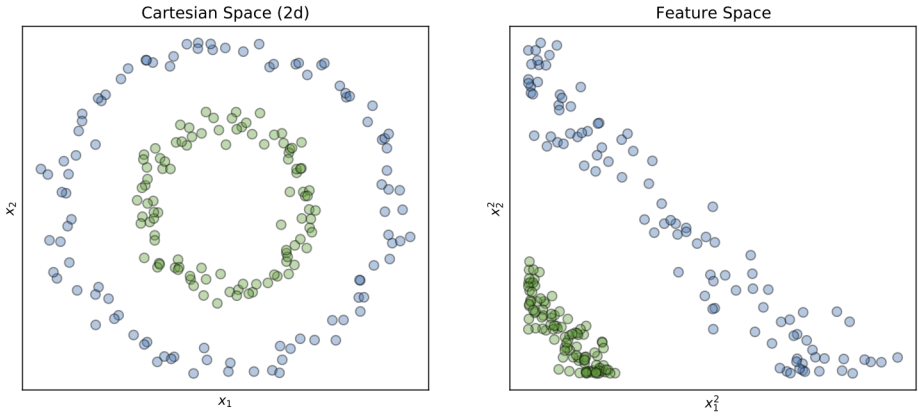
$$\mathscr{D}_{\text{train}} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}, \qquad \mathbf{x}_i \in \mathbb{R}^D$$

the problem is to **re-represent them** as points in a $d$-dimensional space, where $d < D$.

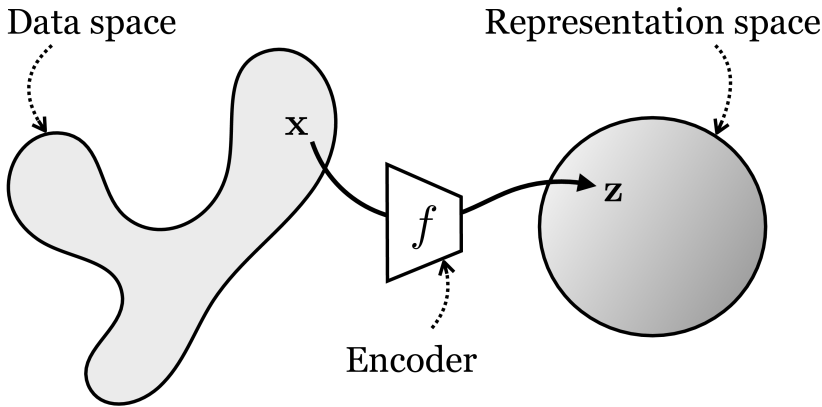
The goal is typically to **retain information** in the data set that will, e.g., allow elements of one class to be distinguished from another.

Dimensionality reduction is a standard technique that is particularly useful to visualize or understand high-dimensional data.

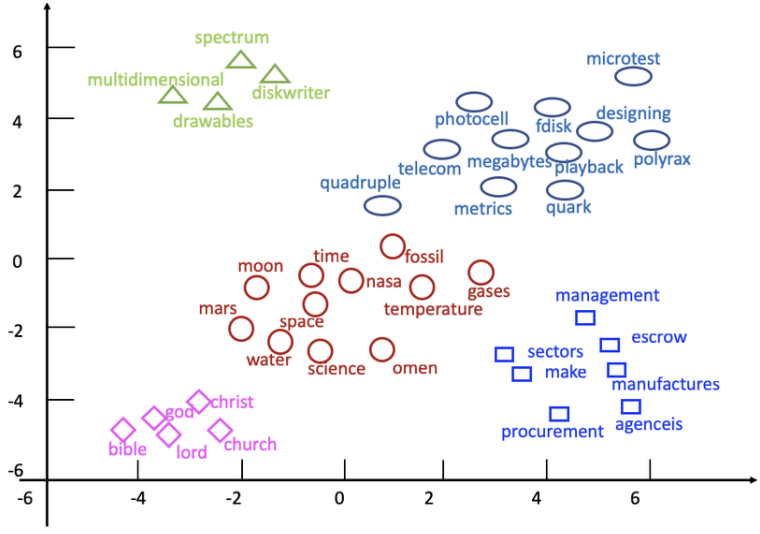
# Unsupervised learning



# Unsupervised learning



# Unsupervised learning

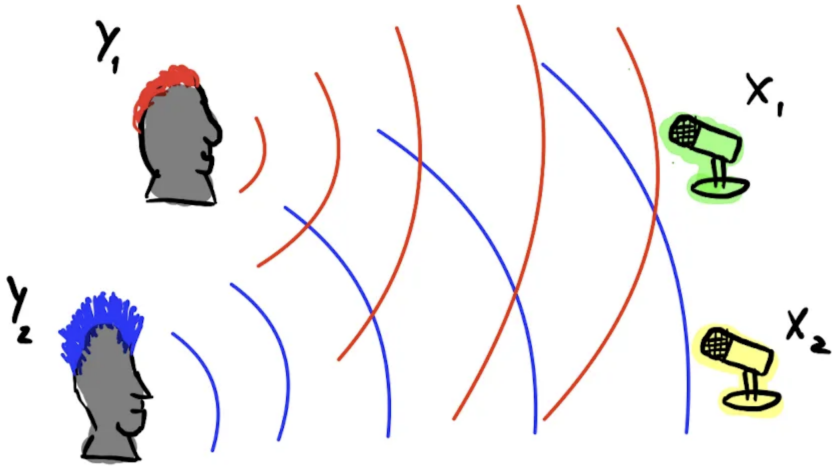


# Unsupervised learning

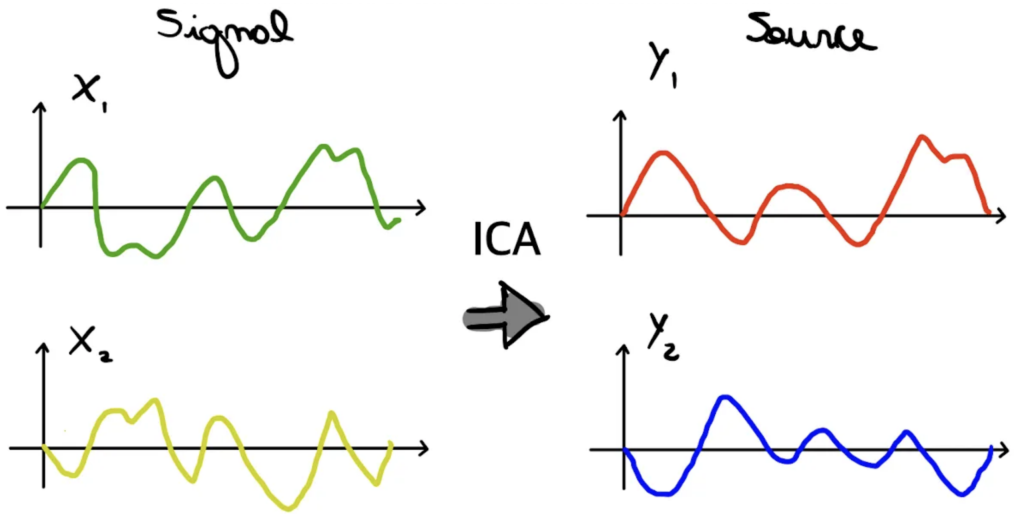
(a)

(b)

# Independent component analysis (ICA)



# Independent component analysis (ICA)



# Independent component analysis (ICA)

The ICA assumes that observations and sources are related through:

$$
\begin{aligned}
\mathbf{y}_1 &= a_{11}\mathbf{x}_1 + a_{12}\mathbf{x}_2 + ... + a_{1n}\mathbf{x}_n \\
\mathbf{y}_2 &= a_{21}\mathbf{x}_1 + a_{22}\mathbf{x}_2 + ... + a_{2n}\mathbf{x}_n \\
&\vdots \\
\mathbf{y}_p &= a_{p1}\mathbf{x}_1 + a_{p2}\mathbf{x}_2 + ... + a_{pn}\mathbf{x}_n
\end{aligned}
$$

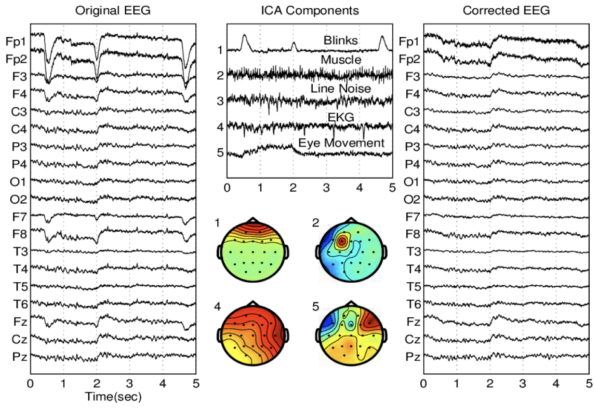
In matrix form:

$$\mathbf{X} = \mathbf{A}^{-1}\mathbf{Y}$$

where:

- $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]^T$ are the independent source signals
- $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_p]^T$ are the observed mixed signals
- $\mathbf{A}^{-1}$ is the mixing matrix (inverse of the unmixing matrix)

# ICA - EEG


Original EEG
ICA Components
Corrected EEG
Fp1
Fp2
F3
F4
C3
C4
P3
P4
O1
O2
F7
F8
T3
T4
T5
T6
Fz
Cz
Pz
0 1 2 3 4 5
Time(sec)
1
2
3
4
5
Blinks
Muscle
Line Noise
EKG
Eye Movement

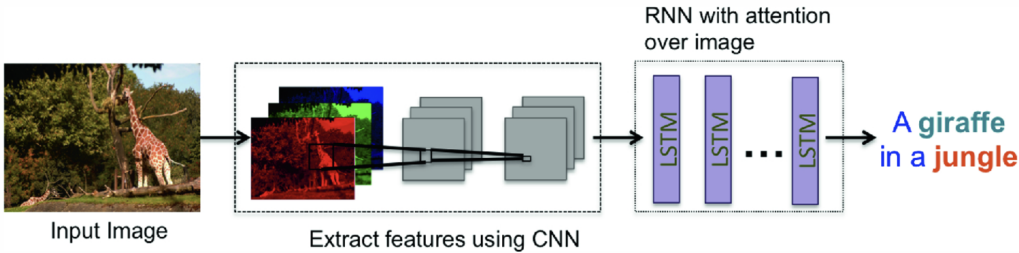
# Sequence learning

The goal is to learn a mapping from *input sequences* $x_0, \ldots, x_N$ to *output sequences* $y_1, \ldots, y_M$.

The mapping is typically represented as a *state machine*, with one function $f_s$ used to compute the next hidden internal state given the input, and another function $f_o$ used to compute the output given the current hidden state.

It is supervised in the sense that we are told what output sequence to generate for which input sequence, but the internal functions have to be learned by some method other than direct supervision, because we don't know what the hidden state sequence is.

# Image to text



# Reinforcement learning

- The goal is to learn a mapping from input values (typically assumed to be states of an agent or system, e.g. the velocity of a moving car) to output values (typically control actions, e.g. if to accelerate or hit the brake).
- We need to learn the mapping without a direct supervision signal to specify which output values are best for a particular input.
  - the learning problem is framed as an agent interacting with an environment.

# Reinforcement learning

- The interaction setting is the following:
  - The agent observes the current state $s_t$.
  - Select an action $a_t$.
  - It receives a reward, $r_t$, which typically depends on $s_t$ and possibly $a_t$.
  - The environment transitions probabilistically to a new state, $s_{t+1}$, with a distribution that depends only on $s_t$ and $a_t$.
  - The agent observes the current state, $s_{t+1}$.
  - …

# Reinforcement learning

- The goal is to find a policy $\pi$, mapping $s$ to $a$, (that is, states to actions) such that some long-term sum or average of rewards $r$ is maximized.
- This setting is very different from either supervised learning or unsupervised learning, because the agent's action choices affect both its reward and its ability to observe the environment.
- It requires careful consideration of the long-term effects of actions as well as all the other issues related to supervised learning.

# Other settings

- In **semi-supervised** learning, we have a supervised-learning training set, but there may be an additional set of $\mathbf{x}_i$ values with no known $y_i$. These values can still be used to improve learning performance (if they are drawn from $\Pr(X)$ that is the marginal of $\Pr(X, Y)$ that governs the rest of the data set).
- In **active** learning, it is assumed to be expensive to acquire a label $y_i$ (imagine asking a human to read an x-ray image), so the learning algorithm can sequentially ask for particular inputs $\mathbf{x}_i$ to be labeled and must carefully select queries to learn as effectively as possible while minimizing the cost of labeling.
- In **transfer** learning (also called *meta-learning*), there are multiple tasks, with data drawn from different, but related, distributions. The goal is for the experience with previous tasks to apply to learning a current task in a way that requires a decrease in experience with the new task.

# Assumptions

The kinds of assumptions that we can make about the data source or the solution include:

- The data are independent and identically distributed (i.i.d.).
- The data are generated by a Markov chain (i.e., the outputs depend only on the current *state*, without additional *memory*).
- The process generating the data might be adversarial.
- The “true’ ’ model that generates the data can be perfectly described by one of a particular set of hypotheses.

# Evaluation criteria

Once we have specified a problem class, we need to say what makes an output or the answer to a query good, given the training data.

We specify evaluation criteria at two levels: how an individual prediction is scored, and how the overall behavior of the prediction or estimation system is scored.

The quality of predictions from a learned model is often expressed in terms of a *loss function*.

A loss function $\mathscr{L}(g, a)$ tells you how much you will be penalized for making a guess $g$ when the answer is actually $a$.

# Evaluation criteria

There are many possible loss functions:

- **0-1 loss** applies to predictions drawn from finite domains.

$$\mathscr{L}(g,a) = \begin{cases} 0 & \text{if } g = a \\ 1 & \text{otherwise} \end{cases}$$

- **Squared loss**

$$\mathscr{L}(g,a) = (g-a)^2$$

- **Absolute loss**

$$\mathscr{L}(g,a) = |g-a|$$

- **Asymmetric loss**

$$\mathscr{L}(g,a) = \begin{cases} 1 & \text{if } g = 1 \text{ and } a = 0 \\ 10 & \text{if } g = 0 \text{ and } a = 1 \\ 0 & \text{otherwise} \end{cases}$$

# Evaluation criteria

Any given prediction rule will usually be evaluated based on multiple predictions and the loss of each one.

At this level, we might be interested in:

- Minimizing expected loss over all the predictions (also known as *risk*)
- Minimizing maximum loss: the loss of the worst prediction
- Minimizing or bounding regret: how much worse this predictor performs than the best one drawn from some class
- Characterizing asymptotic behavior: how well the predictor will perform in the limit of infinite training data
- Finding algorithms that are probably approximately correct: they probably generate a hypothesis that is right most of the time.

# Model type

- Non-parametric models
- Parametric models

# Model type / Non-parametric models

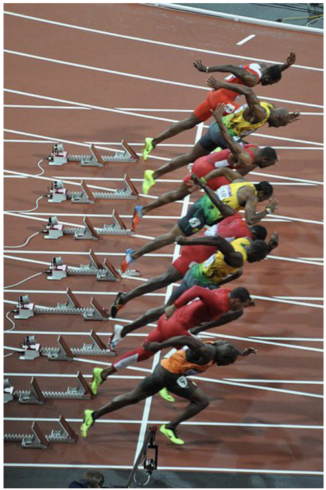
- In some cases, we can answer queries directly from the training data without building a model or learning parameters.
- For example, in regression or classification, we might generate an answer to a new query by averaging answers to similar queries, as in the *nearest neighbor* method.

# Model type / Parametric models

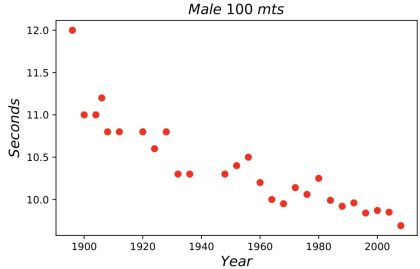
This two-step process is more typical:

1. “Fit’ ’ a model (with some a-prior chosen parameterization) to the training data.
2. Use the model directly to make predictions.

# Example: olympic 100m data



Dataset:




Model:

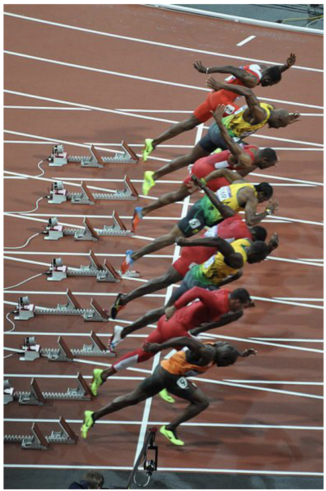
- Linear model $y = f(x, \mathbf{w})$, where $y$ is the time in seconds and $x$ the year of the competition.
- The linear model is given by:

$$y = w_1 x + w_0$$

where $w_0$ is the intercept and $w_1$ is the slope.

- We use $\mathbf{w}$ to refer both to $w_0$ and $w_1$.

# Example: olympic 100m data



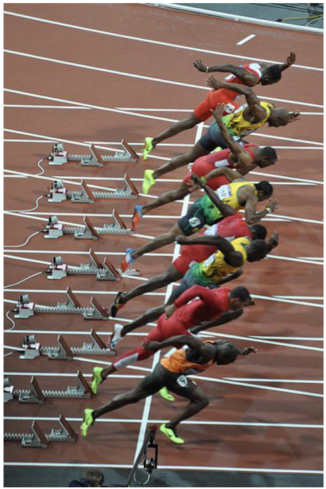
Evaluation criteria (loss function):

- We use an objective function to estimate the parameters $w_0$ and $w_1$ that best fit the data.
- In this example, we use a a least squares objective function:

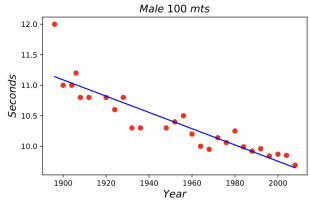
$$E(w_0, w_1) = \sum_{\forall i} (y_i - f(x_i))^2$$
$$= \sum_{\forall i} [y_i - (w_1 x_i + w_0)]^2 .$$

- By minimizing the error with respect to $\mathbf{w}$, we get the solution $w_0 = 36.4$ and $w_1 = -1.34 \times 10^{-2}$.

# Example: olympic 100m data



Data and model:




Predictions:

- What does the model predict for 2012?
- If we say $x = 2012$, then

$$\begin{aligned} y = f(x, \mathbf{w}) &= f(x = 2012, \mathbf{w}) \\ &= w_1 x + w_0 \\ &= (-1.34 \times 10^{-2}) \times 2012 + 36.4 = 9.59. \end{aligned}$$

(The actual value was 9.63)

# Model type / Parametric models

- The model will be some hypothesis or prediction rule $y = h(x; \Theta)$ for some functional form $h$.
- The term *hypothesis* has its roots in statistical learning and the scientific method, where models or hypotheses about the world are tested against real data and refined with more evidence or observations.
- Note that the parameters themselves are only part of the assumptions we are making about the world.
- The model itself is a hypothesis that will be refined with more evidence.
- The idea is that $\Theta$ is a set of one or more parameter values that will be determined by fitting the model to the training data and then remaining fixed during testing.

# Model type / Parametric models

- Given a new $\mathbf{x}_{N+1}$, we would then make the prediction $h(\mathbf{x}_{N+1}; \Theta)$.
- The fitting process is posed as an optimization problem:
  - Find parameter values $\Theta$ that minimize a criterion involving $\Theta$ and the data.
- If the true underlying data distribution $\Pr(X, Y)$ was known:
  - The optimal strategy would be to predict values that minimize the expected loss (also known as test error).

# Model type / Parametric models

- In practice, since Pr$(X, Y)$ is unknown:
  - Instead, we minimize the training error by finding $\Theta$ that minimizes the average loss in the training data.
- The typical criterion to minimize is:

$$\mathscr{E}(h; \Theta) = \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}(h(\mathbf{x}_i; \Theta), y_i) \quad ,$$

  where $\mathscr{L}(g, a)$ measures the penalty for predicting $g$ when the true value is $a$.
- Minimizing training error alone is often not sufficient:
  - This can lead to overfitting, where the model fits the current data well but does not generalize to new values of $x$.

# Model class and parameter fitting

A model *class* $\mathscr{M}$ is a set of possible models, typically parameterized by a vector of parameters $\Theta$.

# Model class and parameter fitting

- **Model Assumptions:**
  - We need to decide on the form of the model.
  - For regression, we might use a linear prediction rule:

$$h(\mathbf{x}; \theta, \theta_0) = \theta^\top \mathbf{x} + \theta_0$$

    where the parameter vector $\Theta = (\theta, \theta_0)$.

- **Model Classes:**
  - "Parametric'' models: Restricted to model classes with a fixed, finite number of parameters.
  - "Non-parametric'' models: Models that do not make this restriction.

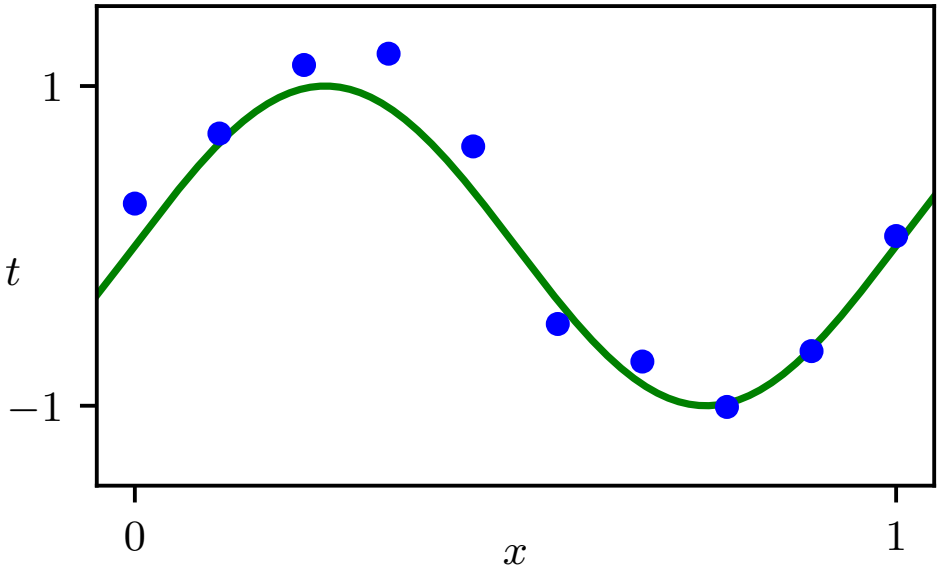
# Model class and parameter fitting

- **Model Selection:**
  - The model class can sometimes be specified directly by the practitioner.
  - Alternatively, several model classes may be tried, and the best is chosen based on some objective function.
  - This process is called *model selection*:
    - Model selection: pick a model class $\mathscr{M}$ from a (usually finite) set of possible model classes.
    - Model fitting: pick the specific model in $\mathscr{M}$ by specifying its (usually continuous) parameters $\Theta$.

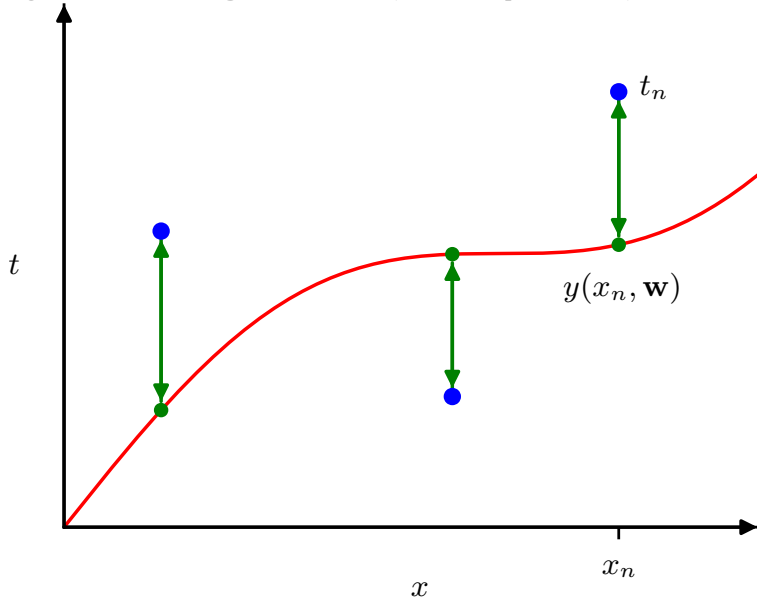
# Algorithm

- **Algorithmic Problem:**
  - After defining a class of models and a scoring method, the challenge is to design a computational procedure to find a good model in the class.
  - For example, finding the parameter vector that minimizes training error may use familiar optimization algorithms, such as least squares, when $h$ is being fitted to the data $x$.
- **Optimization Approaches:**
  - Sometimes, generic optimization software can be used to solve the parameter estimation problem.
  - In many cases, algorithms are specifically designed for machine learning problems or particular hypothesis/model classes.
- **Beyond Direct Optimization:**
  - Some algorithms, such as the perceptron for linear classifiers, do not explicitly optimize a particular criterion, yet are historically significant and effective for certain tasks.

## Example polynomial regression (Bishop 2006)



# Example polynomial regression (Bishop 2006)



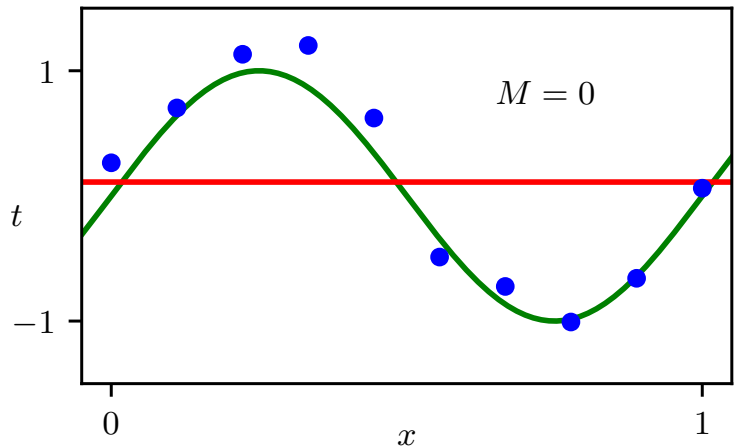
# Example polynomial regression (Bishop 2006)

$$h(x, w) = w_0 + w_1 x + w_2 x^2 + ... + w_M x^M$$

$$E(w) = \sum_{n=1}^{N} \left( h\left(x_n, w\right) - t_n \right)^2$$

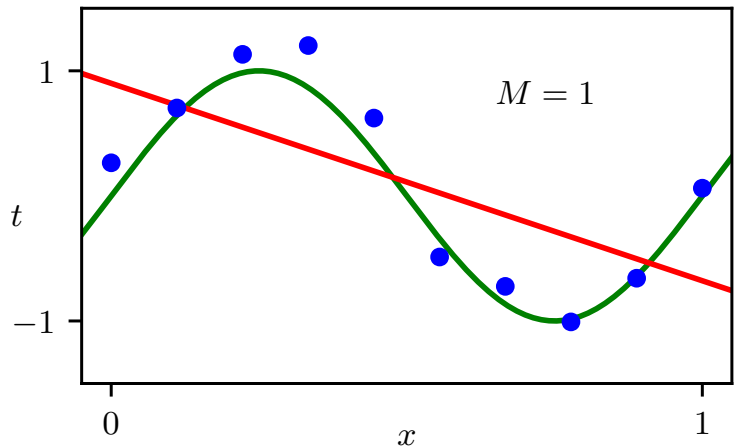
# Example polynomial regression (Bishop 2006)

$$h(x, w) = w_0 + w_1 x + w_2 x^2 + ... + w_M x^M$$



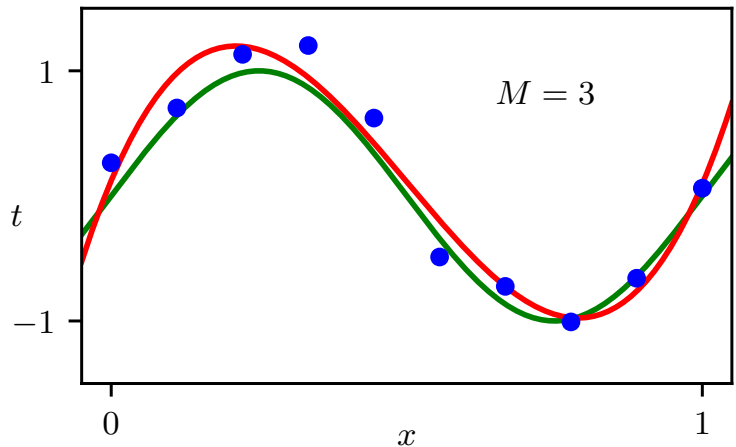
# Example polynomial regression (Bishop 2006)

$$h(x, w) = w_0 + w_1 x + w_2 x^2 + ... + w_M x^M$$



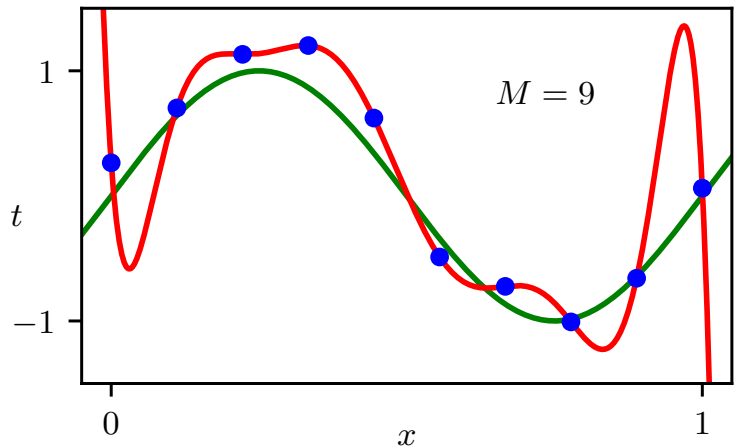
# Example polynomial regression (Bishop 2006)

$$h(x, w) = w_0 + w_1 x + w_2 x^2 + ... + w_M x^M$$

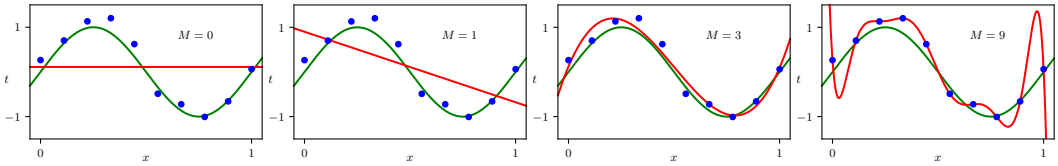


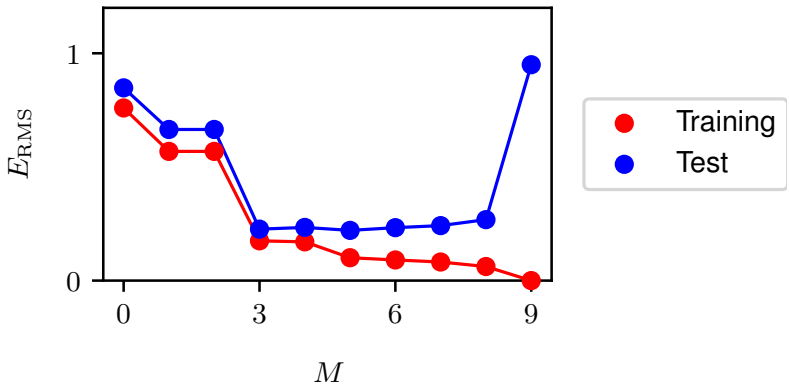
# Example polynomial regression (Bishop 2006)

$$h(x, w) = w_0 + w_1 x + w_2 x^2 + ... + w_M x^M$$

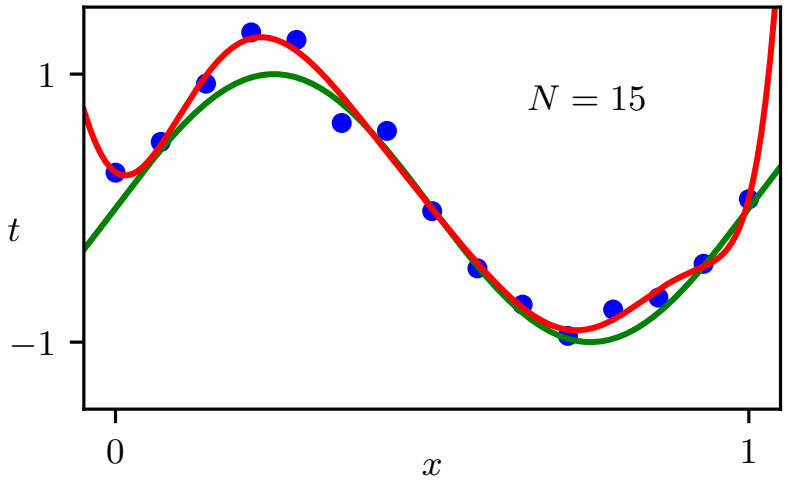


# Example polynomial regression (Bishop 2006)

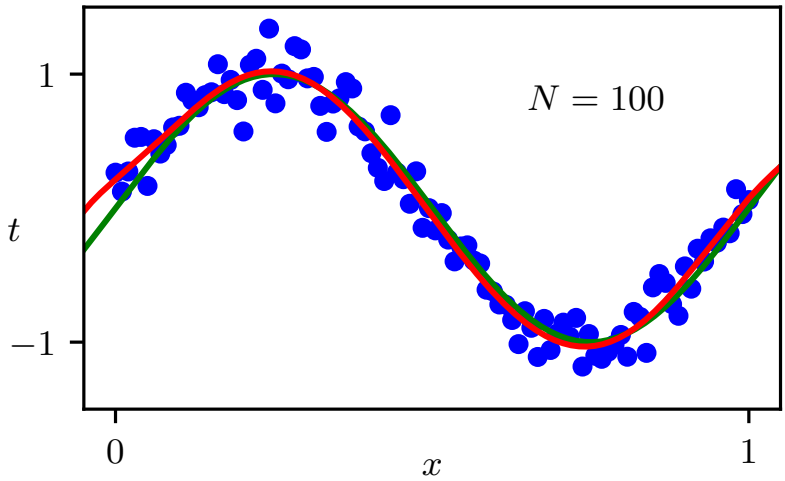


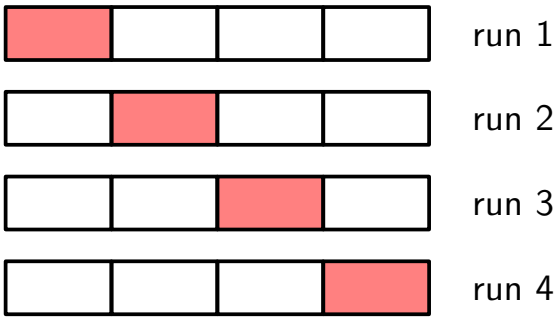
# Example polynomial regression (Bishop 2006)



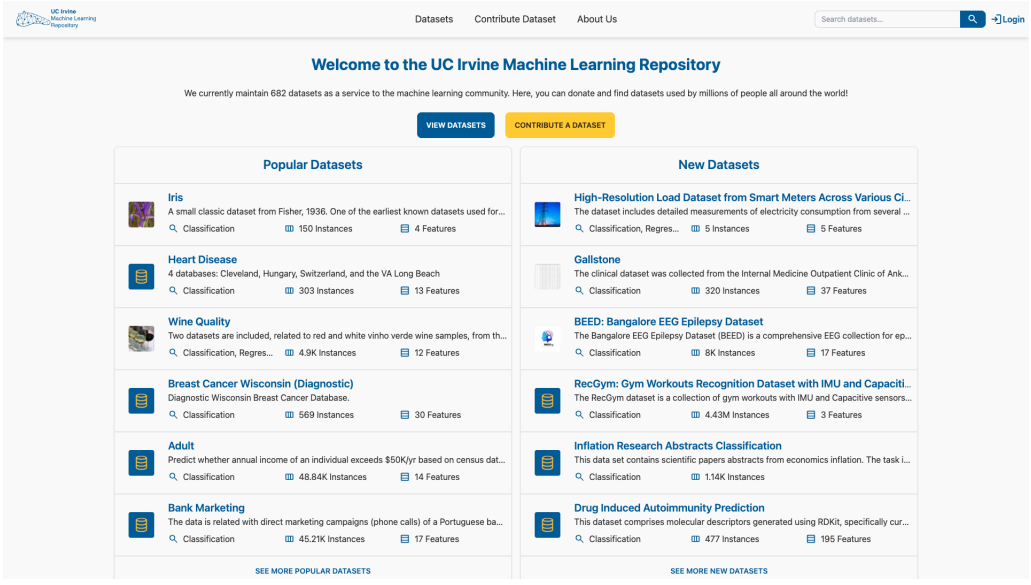
# Example polynomial regression (Bishop 2006)



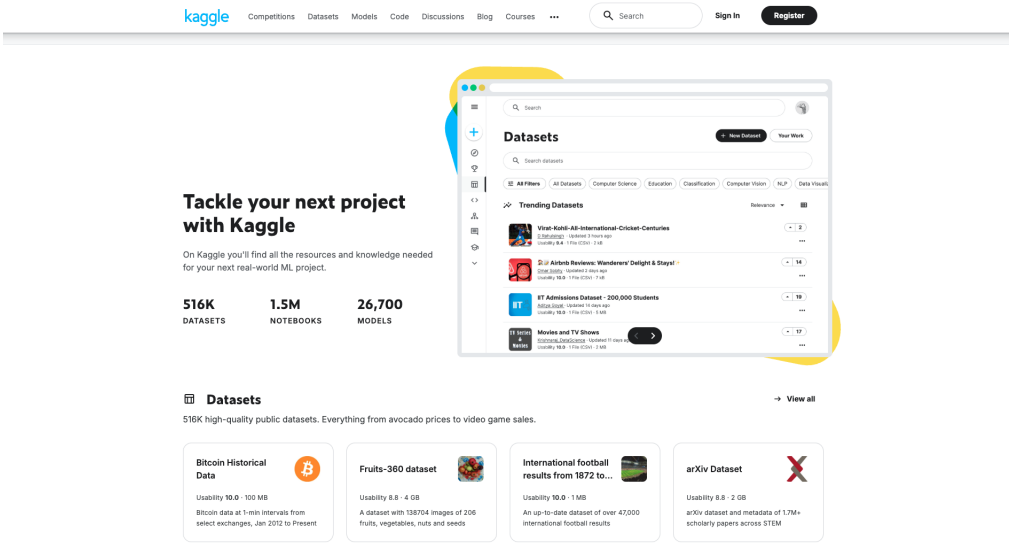
# Cross-validation



# Data sources



# Data sources



# Sources of health data

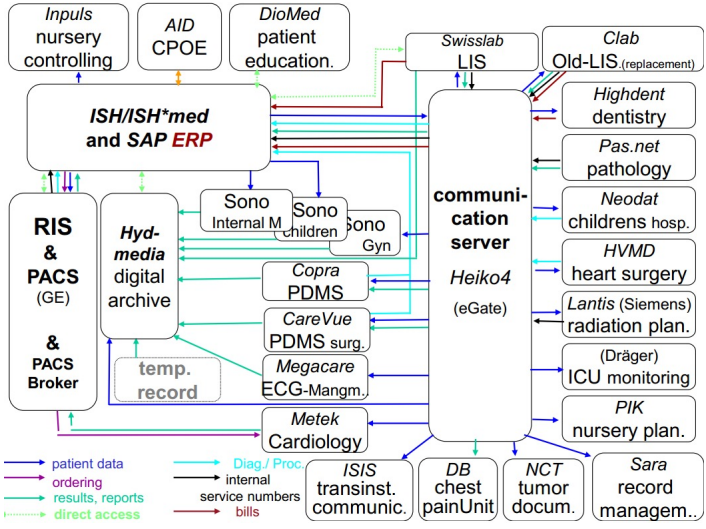
**Institution-based sources**

Hospitals, health centers, community-based institutions/service delivery mechanisms

**Research-related sources**

Publications repositories, diseases databases, specific data repositories.

# HIS application systems



# Sources of health data

- Electronic medical records (private).
- Publications databases, e.g. Pubmed.
- Public repositories: OMIM, ORPHANET, and many others.
- Images/signals (MRI, CT, RX)

- Data types:
  - Images
  - Text

# OMIM®

## Online Mendelian Inheritance in Man®

### An Online Catalog of Human Genes and Genetic Disorders

Updated May 25, 2018

Search OMIM for clinical features, phenotypes, genes, and more...

**Advanced Search :** OMIM, Clinical Synopses, Gene Map

**Need help? :** Example Searches, OMIM Search Help

**Mirror site :** mirror.omim.org

OMIM is supported by a grant from NHGRI, licensing fees, and generous contributions from people like you.

ICD+

# #188400

# DIGEORGE SYNDROME; DGS

*Alternative titles; symbols*
CHROMOSOME 22q11.2 DELETION SYNDROME
HYPOPLASIA OF THYMUS AND PARATHYROIDS
THIRD AND FOURTH PHARYNGEAL POUCH SYNDROME

Other entities represented in this entry:

## DIGEORGE SYNDROME CHROMOSOME REGION, INCLUDED; DGCR, INCLUDED

TAKAO VCF SYNDROME, INCLUDED
CATCH22, INCLUDED

**Phenotype-Gene Relationships**

| Location | Phenotype | Phenotype MIM number | Inheritance (in progress) | Phenotype mapping key | Gene/Locus | Gene/Locus MIM number |
|---|---|---|---|---|---|---|
| 22q11.21 | DiGeorge syndrome | 188400 | AD | 3 | TBX1 | 602054 |

Clinical Synopsis

### TEXT

A number sign (#) is used with this entry because DiGeorge syndrome is caused by a 1.5- to 3.0-Mb hemizygous deletion of chromosome 22q11.2. Haploinsufficiency of the TBX1 gene (602054) in particular is responsible for most of the physical malformations. There is evidence that point mutations in the TBX1 gene can also cause the disorder.

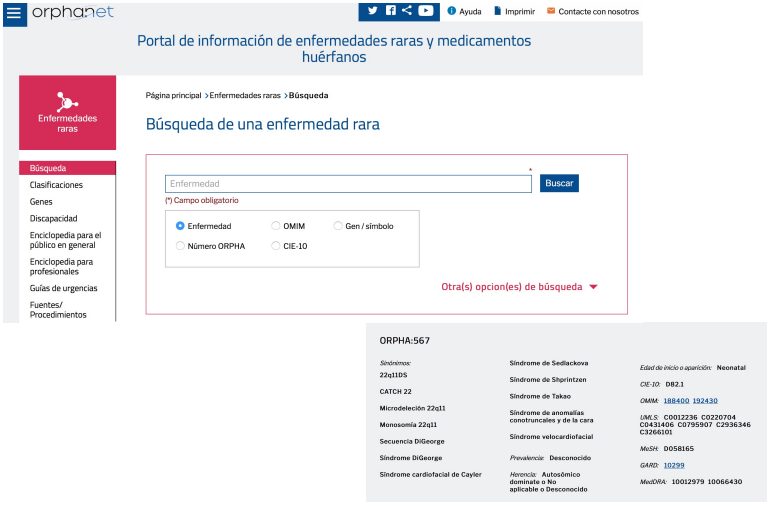
### Description

DiGeorge syndrome (DGS) comprises hypocalcemia arising from parathyroid hypoplasia, thymic hypoplasia, and outflow tract defects of the heart. Disturbance of cervical neural crest migration into the derivatives of the pharyngeal arches and pouches can account for the phenotype. Most cases result from a deletion of chromosome 22q11.2 (the DiGeorge syndrome chromosome region, or DGCR). Several genes are lost including the putative transcription factor TUPLE1 which is expressed in the appropriate distribution. This deletion may present with a variety of phenotypes: Shprintzen, or velocardiofacial, syndrome (VCFS; 192430); conotruncal anomaly face (or Takao syndrome); and isolated outflow tract defects of the heart including tetralogy of Fallot, truncus arteriosus, and interrupted aortic arch. A collective acronym CATCH22 has been proposed for these differing presentations. A small number of cases of DGS have defects in other chromosomes, notably 10p13 (see 601362). In the mouse, a transgenic Hox A3 (Hox 1.5) knockout produces a phenotype similar to DGS as do the teratogens retinoic acid and alcohol.

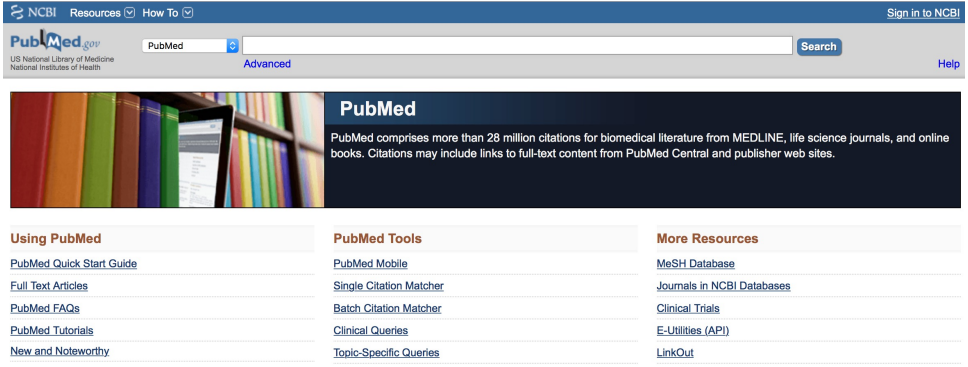
### Nomenclature

DiGeorge syndrome overlaps clinically with the disorder described by the Japanese as 'conotruncal anomaly face syndrome' (Kinouchi et al., 1976; Takao et al., 1980; Shimizu et al., 1984), where the cardiovascular presentation is the focus of attention. The term conotruncal anomaly face syndrome is cumbersome and has the disadvantage of using embryologic assumptions as a title. It would be appropriate to use Takao syndrome for those cases with a preponderant cardiac presentation in contrast to the low T cell and hypocalcemic presentation in infancy of DiGeorge syndrome and the craniofacial and palatal abnormalities typical of Shprintzen syndrome. These 3 phenotypes may be seen in the same family and most cases of all 3 categories have been shown to have a 22q11 deletion. This led Wilson et al. (1993) to propose the acronym CATCH22 (Cardiac Abnormality/abnormal facies, T cell deficit due to thymic hypoplasia, Cleft palate, Hypocalcemia due to hypoparathyroidism resulting from 22q11 deletion) as a collective acronym for those with the common genetic etiology. Shprintzen (1994) objected to 'lumping' velocardiofacial syndrome with the DiGeorge anomaly, arguing that there is 'no valid evidence to suggest that velocardiofacial syndrome is etiologically heterogeneous...[whereas] the DiGeorge anomaly is known to be so.' Hall (1993) cited data of Driscoll et al. (1993) indicating that velocardiofacial syndrome is etiologically heterogeneous. She stated that '...68% of Shprintzen syndrome patients...have been recognised to have deletions of 22q11.' Shprintzen (1994) refuted her statement, maintaining that it could accurately be stated that deletion was found in 68% of patients sent to the Driscoll laboratory with a diagnosis of velocardiofacial syndrome made by other clinicians. Shprintzen (1994) said that in his sample, 100% had deletion.

# ORPHANET


orphanet
Ayuda
Imprimir
Contacte con nosotros
Portal de información de enfermedades raras y medicamentos huérfanos
Enfermedades raras
Página principal > Enfermedades raras > Búsqueda
Búsqueda de una enfermedad rara
Búsqueda
Clasificaciones
Genes
Discapacidad
Enciclopedia para el público en general
Enciclopedia para profesionales
Guías de urgencias
Fuentes/ Procedimientos
Enfermedad
Buscar
(*) Campo obligatorio
Enfermedad
OMIM
Gen / símbolo
Número ORPHA
CIE-10
Otra(s) opcion(es) de búsqueda
ORPHA:567
Síndromes:
22q11DS
CATCH 22
Microdeleción 22q11
Monosomía 22q11
Secuencia DiGeorge
Síndrome DiGeorge
Síndrome cardiofacial de Cayler
Síndrome de Sedlackova
Síndrome de Shprintzen
Síndrome de Takao
Síndrome de anomalías conotruncales y de la cara
Síndrome velocardiofacial
Prevalencia: Desconocido
Herencia: Autosómico dominate o No aplicable o Desconocido
Edad de inicio o aparición: Neonatal
CIE-10: D82.1
OMIM: 188400 192430
UMLS: C0012236 C0220704 C0431406 C0795907 C2936346 C3266101
MeSH: D058165
GARD: 10299
MedDRA: 10012979 10066430

# Pubmed


NCBI
Resources
How To
Sign in to NCBI
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
Search
Advanced
Help
PubMed
PubMed comprises more than 28 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.
Using PubMed
PubMed Quick Start Guide
Full Text Articles
PubMed FAQs
PubMed Tutorials
New and Noteworthy
PubMed Tools
PubMed Mobile
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Topic-Specific Queries
More Resources
MeSH Database
Journals in NCBI Databases
Clinical Trials
E-Utilities (API)
LinkOut

# Medical Data for Machine Learning

https://github.com/beamandrew/medical-data/

This is a curated list of medical data for machine learning.
This list is provided for informational purposes only, please make sure you respect any and all usage restrictions for any of the data listed here.

## 1. Medical Imaging Data

**The National Library of Medicine presents MedPix®**
Database of 53,000 medical images from 13,000 patients with annotations. **Requires registration**.
Information: https://medpix.nlm.nih.gov/home

---

**ABIDE: The Autism Brain Imaging Data Exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism.**
Function MRI images for 539 individuals suffering from ASD and 573 typical controls. These 1112 datasets are composed of structural and resting state functional MRI data along with an extensive array of phenotypic information. **Requires registration**.
Paper: http://www.ncbi.nlm.nih.gov/pubmed/23774715
Information: http://fcon_1000.projects.nitrc.org/indi/abide/
Preprocessed version: http://preprocessed-connectomes-project.org/abide/

---

**Alzheimer's Disease Neuroimaging Initiative (ADNI)**
MRI database on Alzheimer's patients and healthy controls. Also has clinical, genomic, and biomaker data. **Requires registration**.
Paper: http://www.neurology.org/content/74/3/201.short
Access: http://adni.loni.usc.edu/data-samples/access-data/

# Medpix

MedPix®  Search

CME  Cases  Topics  Login

By Diagnosis
Contribute Case
Download Cases

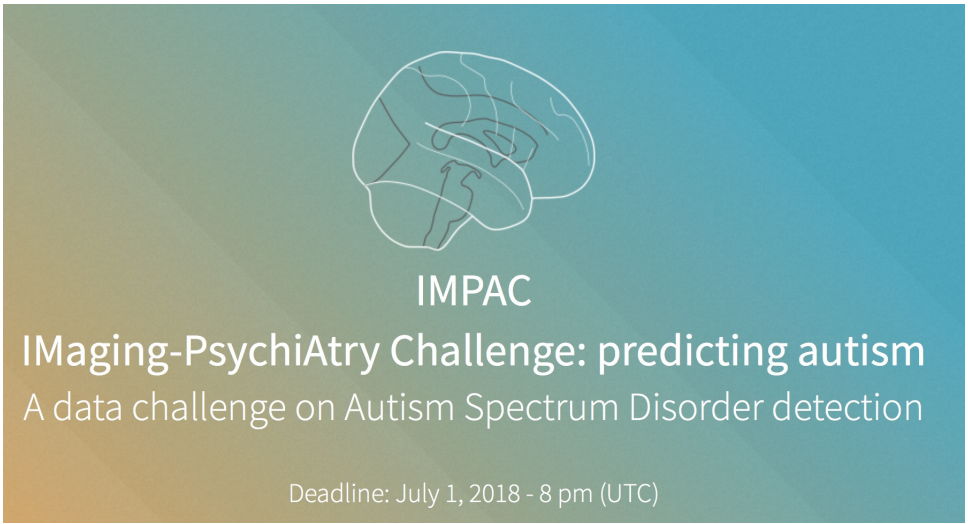
## The National Library of Medicine presen

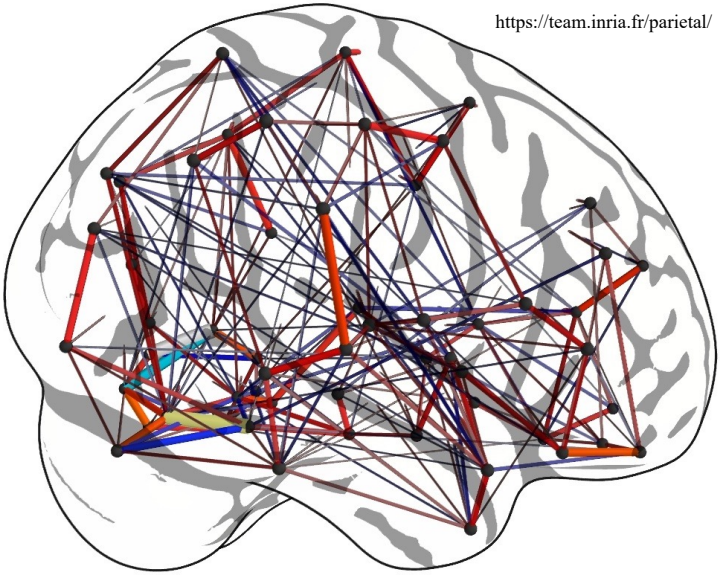
Case Of The Week( COW )

MedPix® is a free open-access online database of medical images, teaching cases, and clinical topics, integrating images and textual metadata including over 12,000 patient case scenarios, 9,000 topics, and nearly 59,000 images. Our primary target audience includes physicians and nurses, allied health professionals, medical students, nursing students and others interested in medical knowledge.

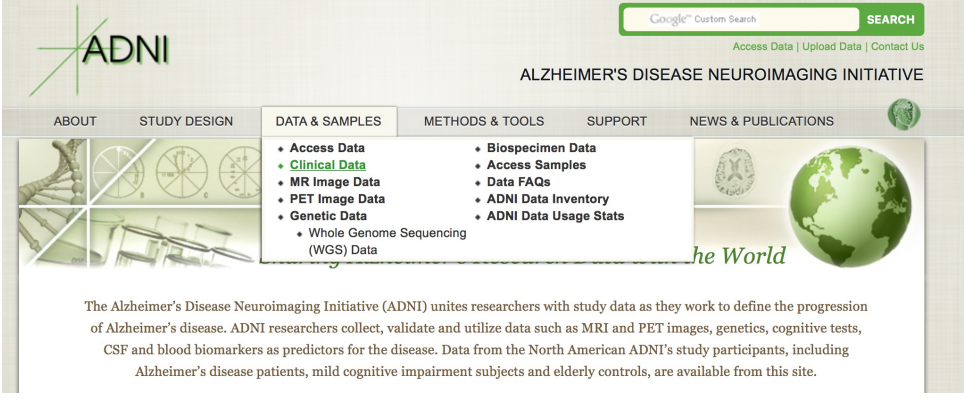
The content material is organized by disease location (organ system); pathology category; patient profiles; and, by image classification and image captions. The collection is searchable by patient symptoms and signs, diagnosis, organ system, image modality and image description, keywords, contributing authors, and many other search options.

# Autism


IMPAC
IMaging-PsychiAtry Challenge: predicting autism
A data challenge on Autism Spectrum Disorder detection
Deadline: July 1, 2018 - 8 pm (UTC)


https://team.inria.fr/parietal/

# ADNI


ADNI
Google™ Custom Search
SEARCH
Access Data | Upload Data | Contact Us
ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE
ABOUT
STUDY DESIGN
DATA & SAMPLES
METHODS & TOOLS
SUPPORT
NEWS & PUBLICATIONS
Access Data
Clinical Data
MR Image Data
PET Image Data
Genetic Data
Whole Genome Sequencing (WGS) Data
Biospecimen Data
Access Samples
Data FAQs
ADNI Data Inventory
ADNI Data Usage Stats
he World
The Alzheimer's Disease Neuroimaging Initiative (ADNI) unites researchers with study data as they work to define the progression of Alzheimer's disease. ADNI researchers collect, validate and utilize data such as MRI and PET images, genetics, cognitive tests, CSF and blood biomarkers as predictors for the disease. Data from the North American ADNI's study participants, including Alzheimer's disease patients, mild cognitive impairment subjects and elderly controls, are available from this site.

**Digital Retinal Images for Vessel Extraction (DRIVE)**
The DRIVE database is for comparative studies on segmentation of blood vessels in retinal images. It consists of 40 photographs out of which 7 showing signs of mild early diabetic retinopathy.
Paper: http://www.isi.uu.nl/Research/Publications/publicationview/id=855.html
Access: http://www.isi.uu.nl/Research/Databases/DRIVE/download.php

---

**AMRG Cardiac Atlas** The AMRG Cardiac MRI Atlas is a complete labelled MRI image set of a normal patient's heart acquired with the Auckland MRI Research Group 's Siemens Avanto scanner. The atlas aims to provide university and school students, MR technologists, clinicians...

**Congenital Heart Disease (CHD) Atlas** The Congenital Heart Disease (CHD) Atlas represents MRI data sets, physiologic clinical data and computer models from adults and children with various congenital heart defects. The data have been acquired from several clinical centers including Rady...

**DETERMINE** Defibrillators to Reduce Risk by Magnetic Resonance Imaging Evaluation, is a prospective, multicenter, randomized clinical trials in patients with coronary artery diseases and mild-to-moderate left ventricular dysfunction. The primary objective...

**MESA** Multi-Ethnic Study of Atherosclerosis, is a large-scale cardiovascular population study (>6,500 participants) conducted in six centres in the USA. It aims to investigate the manifestation of subclinical to clinical cardiovascular disease before...

# OMNIMEDICALSEARCH.COM

**Directories:** Medical Associations | Patient Forums | Medical Journals | Medical Images

## Medical Image Databases & Libraries

**General Category**

e-Anatomy.org - Interactive Atlas of Anatomy - e-anatomy is an anatomy e-learning web site. More than 1500 slices from normal CT and MR exams were selected in order to cover the entire sectional anatomy of human body. Images were labeled using Terminologia Anatomica. A user-friendly interface allows to cine through multi-slice image series combined with interactive textual information, 3D models and anatomy drawings.

Medical Pictures and Definitions
Welcome to the largest database of medical pictures and definitions on the Internet. There are many sites sites that provide medical information but very few that provide medical pictures. As far as we know we are the only one that provides a medical picture database with basic information about each term pictured. Editor's Note: Nice website with free access & no pesky registration to 1200+ health and medical related images with definitions.

Nucleus Medical Art - Medical Illustrations, Medical Art. Includes 3D animations. "Nucleus Medical Art, Inc. is a leading creator and distributor of medical illustrations, medical animations, and interactive multimedia for publishing, legal, healthcare, entertainment, pharmaceutical, medical device, academia and other markets, both in the U.S. and abroad. Editors Note: Great website.

Medical Image Databases on the Internet (UTHSCSA Library) A directory of links to websites with topic specific medical related images.

Surgery Videos - A National Library of Medicine MedlinePlus collection of links to 100s and 100s of different surgical procedures. You must have RealPlayer media player on your computer to view these videos which are free of charge.

The ADAM Medical Encyclopedia with Illustrations. Perhaps one of the best illustrated medical works on the internet today, the ADAM Medical Encyclopedia includes over 4,000 articles about diseases, tests, symptoms, injuries, and surgeries. It also contains an extensive library of medical photographs and illustrations to back up those 4,000 articles. These illustrations and articles are free to the public.

Hardin MD - Medical and Disease Pictures, is a Free and established resource that has been offered by the University of Iowa for quite some time. The home page is in directory style where users will have to drill down to find the images they are looking for, many of which go offsite. Nevertheless, Hardin MD is an excellent gateway to 1,000s of detailed medical photos and illustrations.

Health Education Assets Library (HEAL)

## 2. Challenges/Contest Data

**Visual Concept Extraction Challenge in Radiology** Manually annotated radiological data of several anatomical structures (e.g. kidney, lung, bladder, etc.) from several different imaging modalities (e.g. CT and MR). They also provide a cloud computing instance that anyone can use to develop and evaluate models against benchmarks.

Access: http://www.visceral.eu/

---

**Grand Challenges in Biomedical Image Analysis**

A collection of biomedical imaging challenges in order to *facilitate better comparisons between new and existing solutions*, by standardizing evaluation criteria. You can create your own challenge as well. As of this writing, there are 92 challenges that provide downloadable data sets.

Access: http://www.grand-challenge.org/

---

# VISCERAL Retrieval2 dataset

The Retrieval2 dataset consists 2311 volumes originated from various modalities (CT, MRT1, MRT2). These scans have been acquired during the daily clinical routine work from three different data providers. For a subset of these volumes we provide from the volume's radiological report extracted anatomy-pathology terms in the form of csv files. The following table gives an overview of the dataset in which a participant should perform the retrieval task.

| Modality | Body Region | Volumes | Available A-P term files |
|---|---|---|---|
| CT | Abdomen | 336 | 213 |
| CT | Thorax | 971 | 699 |
| CT | Thorax + Abdomen | 86 | 86 |
| CT | Unknown | 211 | 211 |
| CT | Whole body | 410 | 410 |
| MRT1 | Abdomen | 167 | 114 |
| MRT1 | Unknown | 24 | 24 |
| MRT2 | Abdomen | 68 | 18 |
| MRT2 | Unkwnown | 38 | 38 |
| **TOTAL** | | **2311** | **1813** |

The anatomy-pathology term files list pathological terms that occur in the report of a volume together with its anatomy. Both entities are described textually and additionally with their corresponding Radlex ID (RID). Radlex is a unified language of radiology terms that can be used for standardized indexing and retrieval of radiology information resources. Each term file lists both, occurring and explicitly in the report negated pathologies.

**Kaggle diabetic retinopathy**

High-resolution retinal images that are annotated on a 0–4 severity scale by clinicians, for the detection of diabetic retinopathy. This data set is part of a completed Kaggle competition, which is generally a great source for publicly available data sets.

Access: https://www.kaggle.com/c/diabetic-retinopathy-detection

---

**Cervical Cancer Screening**

In this kaggle competition, you will develop algorithms to correctly classify cervix types based on cervical images. These different types of cervix in our data set are all considered normal (not cancerous), but since the transformation zones aren't always visible, some of the patients require further testing while some don't.

Access: https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening/data

---

**Multiple sclerosis lesion segmentation**

challenge 2008. A collection of brain MRI scans to detect MS lesions.

Access: http://www.ia.unc.edu/MSseg/

---

**Multimodal Brain Tumor Segmentation Challenge**

Large data set of brain tumor magnetic resonance scans. They've been extending this data set and challenge each year since 2012.

**Coding4Cancer**

A new initiative by the Foundation for the National Institutes of Health and Sage Bionetworks to host a series of challenges to improve cancer screening. The first is for digital mammography readings. The second is for lung cancer detection. The challenges are not yet launched.

Access: http://coding4cancer.org/

---

**EEG Challenge Datasets on Kaggle**

- Melbourne University AES/MathWorks/NIH Seizure Prediction - Predict seizures in long-term human intracranial EEG recordings

Access: https://www.kaggle.com/c/melbourne-university-seizure-prediction

- American Epilepsy Society Seizure Prediction Challenge - Predict seizures in intracranial EEG recordings

Access: https://www.kaggle.com/c/seizure-prediction

- UPenn and Mayo Clinic's Seizure Detection Challenge - Detect seizures in intracranial EEG recordings

Access: https://www.kaggle.com/c/seizure-detection

- Grasp-and-Lift EEG Detection - Identify hand motions from EEG recordings

Access: https://www.kaggle.com/c/grasp-and-lift-eeg-detection

## 3. Data derived from Electronic Health Records (EHRs)

**Building the graph of medicine from millions of clinical narratives**
Co-occurence statistics for medical terms extracted from 14 million clinical notes and 260,000 patients.
Paper: http://www.nature.com/articles/sdata201432
Data: http://datadryad.org/resource/doi:10.5061/dryad.jp917

---

**Learning Low-Dimensional Representations of Medical Concept**
Low-dimensional embeddings of medical concepts constructed using claims data. Note that this paper utilizes data from *Building the graph of medicine from millions of clinical narratives*
Paper: http://cs.nyu.edu/~dsontag/papers/ChoiChiuSontag_AMIA_CRI16.pdf
Data: https://github.com/clinicalml/embeddings

---

**MIMIC-III, a freely accessible critical care database**
Anonymized critical care EHR database on 38,597 patients and 53,423 ICU admissions. **Requires registration**.
Paper: http://www.nature.com/articles/sdata201635
Data: http://physionet.org/physiobank/database/mimic3cdb/

---

**Clinical Concept Embeddings Learned from Massive Sources of Medical Data**
Embeddings for 108,477 medical concepts learned from 60 million patients, 1.7 million journal articles, and clinical notes of 20 million patients
Paper: https://arxiv.org/abs/1804.01486
Embeddings: https://figshare.com/s/00d69861786cd0156d81
Interactive tool: http://cui2vec.dbmi.hms.harvard.edu

PARTNERS HEALTHCARE

FOUNDED BY BRIGHAM AND WOMEN'S HOSPITAL AND MASSACHUSETTS GENERAL HOSPITAL

Partners HealthCare Biobank Disease Challenge



## Welcome to the Partners HealthCare Biobank Disease Challenge

Partners HealthCare is hosting a "Biobank Disease Challenge," an articial intelligence and machine learning data analytics competition and it is open to researchers across the United States.

The goal of this competition is to enable major translational data science players to leverage the Partners HealthCare Biobank in order to develop better phenotypic algorithms for clinical and basic research.

Registration will be open on April 23 and close on June 29, 2018 to 50 teams. Prizes will be awarded for the best machine learning models and visualizations.

### Tentative Challenge Schedule

- April 23, 2018 // Registration Opens
- June 29, 2018 // Registration Closes
- July 2 – August 19, 2018 // Practice on Platform
- September 12 – October 10, 2018 // Competition

### The following prize awards are as follows:

- First place: $15,000
- Second place: $10,000
- Third place: $5,000

Part 1: Develop Computed Phenotypes for 5 Diseases

In this challenge we are asking participants to develop computed phenotype machine learning algorithms to identify 5 disease states (to be released at the beginning of the challenge) of all patients enrolled in the Partners Biobank. Algorithms must be developed against available EHR data and limited training data. These algorithms will be assessed against gold standard labels from clinician assessment of a patient's full clinical chart. All submissions should include an algorithm score for each patient and disease state. Scores will be ranked for each disease and an area under the ROC curve (AUC) will be calculated based on the validation dataset. The final score will be partially based on the average of the 5 disease AUCs.

## 5. UCI Datasets

**Liver Disorders Data Set**
Data on 345 patients with and without liver disease. Features are 5 blood biomarkers thought to be involved with liver disease.
Data: https://archive.ics.uci.edu/ml/datasets/Liver+Disorders

**Thyroid Disease Data Set**
Data: https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease

**Breast Cancer Data Set**
Data: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

**Heart Disease Data Set**
Data: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

**Lymphography Data Set**
Data: https://archive.ics.uci.edu/ml/datasets/Lymphography

## 6. Biomedical Literature

**PMC Open Access Subset**
Collection of all the full-text, open access articles in Pubmed central.
Information: http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
Archived files: http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/#Data_Mining

**PubMed 200k RCT**

Collection of pubmed abstracts from randomized control trials (RCTs). Annotations for each sentence in the abstract are available.

Paper: https://arxiv.org/abs/1710.06071

Data: https://github.com/Franck-Dernoncourt/pubmed-rct

## 6. TREC Precision Medicine / Clinical Decision Support Track

Text REtrieval Conference (TREC) is running a track on Precision Medicine / Clinical Decision Support from 2014.

**2014 Clinical Decision Support Track**
Focus: Retrieval of biomedical articles relevant for answering generic clinical questions about medical records.
Information and Data: http://www.trec-cds.org/2014.html

**2015 Clinical Decision Support Track**
Focus: Retrieval of biomedical articles relevant for answering generic clinical questions about medical records.
Information and Data: http://www.trec-cds.org/2015.html

**2016 Clinical Decision Support Track**
Focus: Retrieval of biomedical articles relevant for answering generic clinical questions about medical records. Actual electronic health record (EHR) patient records are be used instead of synthetic cases.
Information and Data: http://www.trec-cds.org/2016.html

**2017 Clinical Decision Support Track**
Focus: Retrieve useful precision medicine-related information to clinicians treating cancer patients.
Information and Data: http://www.trec-cds.org/2017.html

## Documents

There are two target document collections for the Precision Medicine track: scientific abstracts and clinical trials. Both XML and TXT versions are available for both sets. Note that the XML is the official collection, as it has the complete information for each abstract/trial. The TXT versions are provided for ease of processing, but no guarantees are made that all information is contained within these files.

### Obtaining the Collection

**Scientific Abstracts**: A January 2017 snapshot of PubMed abstracts is used for the scientific abstracts. Additionally, abstracts obtained from AACR and ASCO proceedings are included in this category (these are more targeted toward cancer therapy, and likely to include precision medicine studies not in PubMed). These are only available as TXT files, and the file name (without extension) should be used as the ID in the submission files.

1. medline_xml.part1.tar.gz [4.9 GB]
2. medline_xml.part2.tar.gz [4.9 GB]
3. medline_xml.part3.tar.gz [4.9 GB]
4. medline_xml.part4.tar.gz [4.9 GB]
5. medline_xml.part5.tar.gz [1.2 GB]
6. medline_txt.tar.gz [8.8 GB]
7. extra_abstracts.tar.gz [58 MB]

**Clinical Trials**: An April 2017 snapshot of ClinicalTrials.gov is used for the clinical trial descriptions.

1. clinicaltrials_xml.tar.gz [700 MB]
2. clinicaltrials_txt.tar.gz [275 MB]

## Topics

The topics for the track consist of synthetic patient cases created by MD Anderson precision oncologists. The topics consist of the disease, genetic variants, demographic, and potentially other information about the patients. For example:

| | **Patient1** | **Patient2** |
|---|---|---|
| **Disease:** | Acute lymphoblastic leukemia | thyroid cancer |
| **Variant:** | ABL1, PTPN11 | RET, BRAF |
| **Demographic:** | 12-year-old male | 63-year-old female |
| **Other:** | No relevant factors | Ecog grade of 2 |

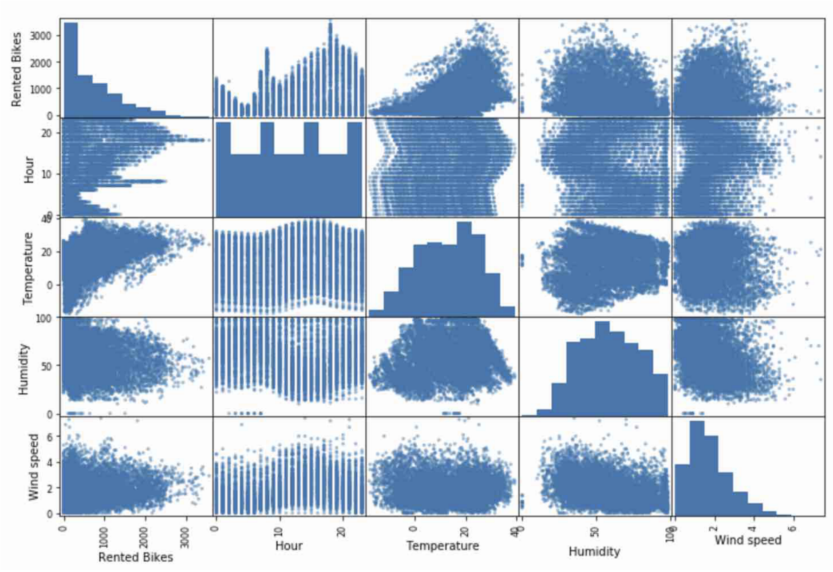
# Machine Learning project ckecklist

1. Frame the problem and look at the big picture.
2. Get the data.
3. Explore the data to get insights.
4. Prepare the data to better expose the underlying data patterns.
5. Explore many different models and shortlist the best ones.
6. Fine-tune your models and combine them into a solution.
7. Present your solution.

# Dataset: bike rentals

- The feature vector $\mathbf{x}$ includes the following features: hour, temperature, humidity, wind speed, visibility, Dew point temperature, solar radiation, rainfall, snowfall, seasons, holiday, functioning day.
- The variables hour, temperature, humidity, wind speed, visibility, Dew point temperature, solar radiation, rainfall, snowfall can be considered as continuous.
- The variables seasons, holiday, and functioning day are categorical variables.
- The output variable $y$ is the number of bikes rented.

# Explore the data



# Study correlations between attributes

The correlation coefficient between two RVs $X$ and $Y$ is given by:

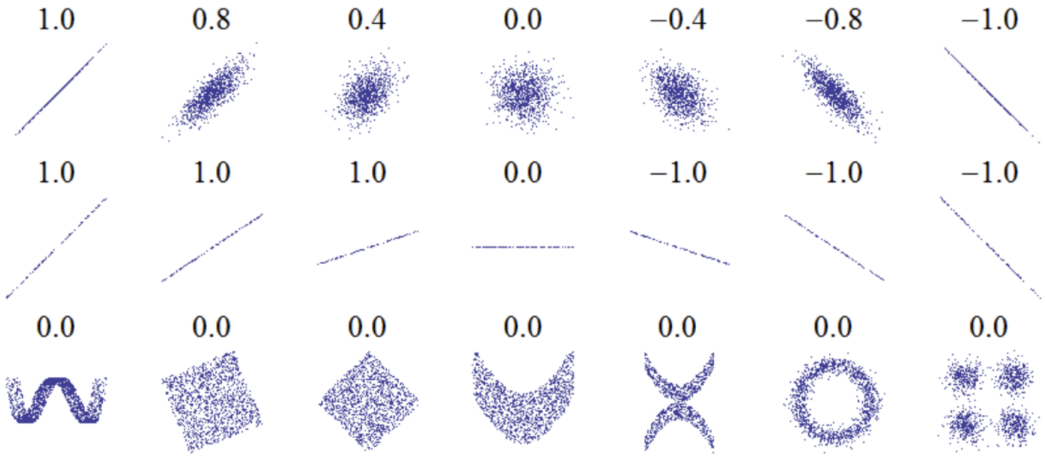
$$\rho_{X,Y} = \frac{E\{(X-\mu_X)(Y-\mu_Y)\}}{\sigma_X \sigma_Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y},$$

where $-1 < \rho_{X,Y} < 1$ and $\sigma_{X,Y}$ is known as the covariance between $X$ and $Y$.

```
In [18]: corr_matrix["Rented Bikes"].sort_values(ascending=False)

Out[18]: Rented Bikes             1.000000
         Temperature              0.538558
         Hour                     0.410257
         Dew point temperature    0.379788
         Solar Radiation          0.261837
         Visibility               0.199280
         Wind speed               0.121108
         Rainfall                -0.123074
         Snowfall                -0.141804
         Humidity                -0.199780
         Name: Rented Bikes, dtype: float64
```

# Correlation and dependence



# Data preparation - cleaning

Remove the outliers in your data (optional).

Handle the missing values:

- Filling them in using the mean, the median, or any other value.
- Drop the feature if most of the instances have a missing value.
- Drop the instance if you have several instances with missing values.
- You can add an additional feature indicating whether the instance has a missing feature or not and then use a value of 0 for the missing feature.

# Data preparation - cleaning

- Most ML methods require features that are numbers rather than categories (usually appearing as text).
- Also, if the feature is categorical, it is useful to use a different representation.
- In the previous example of bike rentals, there were three categorical features:
  - Season that can take four categories.
  - Holiday and functioning day, each taking two categories.

# Data preparation - cleaning

- For example, the feature season takes values autumn, winter, spring and summer.
- The way to handle this feature is to use a representation known as one-hot encoding to obtain a higher-dimensional binary representation for each value.

$$\begin{aligned}\text{autumn} &= [1,0,0,0]\\ \text{winter} &= [0,1,0,0]\\ \text{spring} &= [0,0,1,0]\\ \text{summer} &= [0,0,0,1]\end{aligned}$$

- The values of the feature season do not have a natural order, therefore one should not map these values to numbers like 1 for autumn, 2 for winter, 3 for spring and 4 for summer.
- The ML method will try to find regularities within these ordered values even though they do not exist.

# Data preparation - Feature selection/engineering

- You have the option to remove features that are uninformative.
- You have the option to discretize a continuous feature (e.g. binning).
- You can also create new features from the ones available.
- For example, instead of using the feature $x$, you can use $\log(x)$, $\sqrt{x}$, $x^2$, etc.

# Feature scaling

- Several ML methods do not perform well when the input features have very different scales.
- In the rental bike example, the variable humidity is in the range 0 to 100, whereas the wind speed is in the range 0 to 8.
- Two ways to get all features to have the same scale are normalization (or min-max scaling) and standardization (or z-score normalization).

# Feature scaling

- In normalization, we map the range of values that a feature takes to the range $[-1, 1]$ or $[0, 1]$.
- The normalization formula is given by:

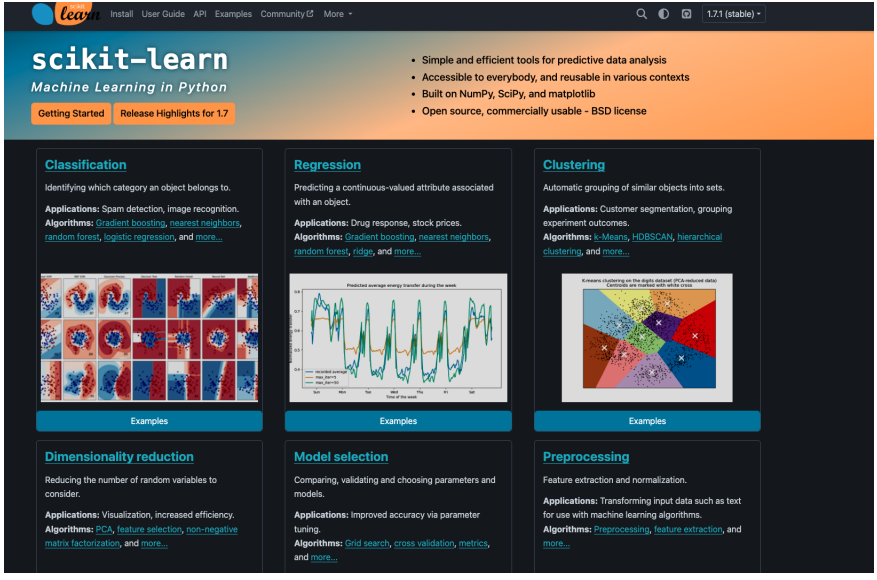
$$\bar{x}_j = \frac{x_j - \min x_j}{\max x_j - \min x_j},$$

  where $\min x_j$ and $\max x_j$ are the minimum and maximum values of the feature in the training set.
- In standardization, the features are scaled so that they have mean zero and standard deviation equal to one,

$$\hat{x}_j = \frac{x_j - \mu_j}{\sigma_j},$$

  where $\mu_j$ and $\sigma_j$ are the mean and standard deviation of the feature $x_j$.

# scikit-learn



# What is Deep Learning

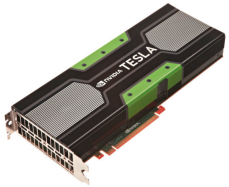
Good old Neural Networks, with more layers/modules.

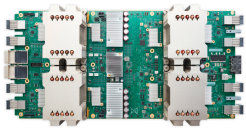
Non-linear, hierarchical, abstract representations of data.

Flexible models with any input/output type and size.

# Why Deep Learning Now?

- Better algorithms & understanding
- Computing power (GPUs, TPUs, …)
- Data with labels
- Open source tools and models



PYTORCH

Microsoft CNTK

Caffe2

dmlc mxnet

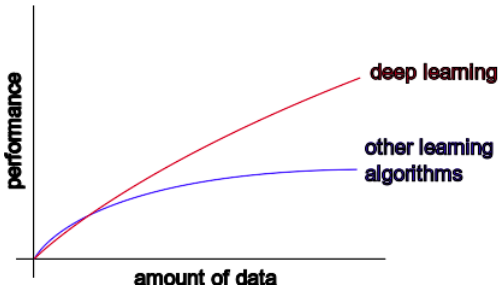
gensim

spaCy

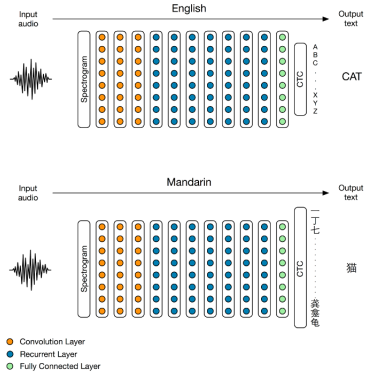
theano

# Why Deep Learning Now?

- Better algorithms & understanding
- Computing power (GPUs, TPUs, …)
- Data with labels
- Open source tools and models

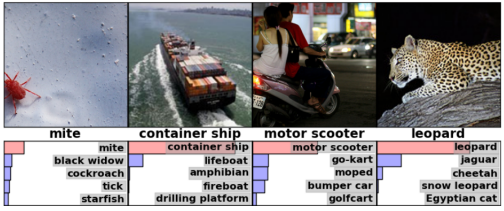


# DL Today: Speech-to-Text



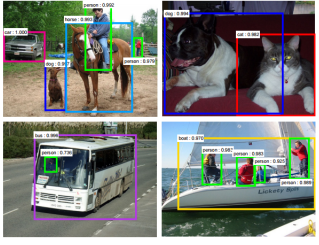

[Baidu 2014]

# DL Today: Vision




[Krizhevsky 2012]

[Ciresan et al. 2013]

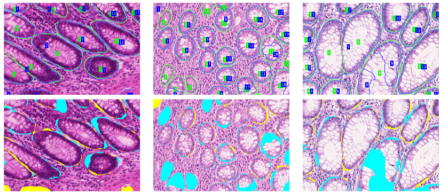


[Faster R-CNN - Ren 2015]

[NVIDIA dev blog]

# DL Today: Vision

[Stanford 2017]




[Nvidia Dev Blog 2017]

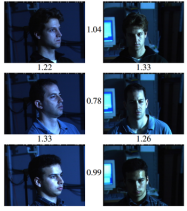
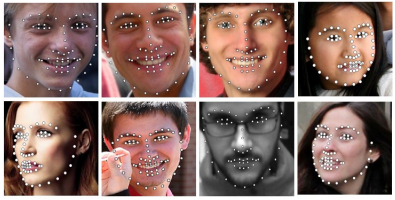



Figure 1. **Illumination and Pose invariance.**

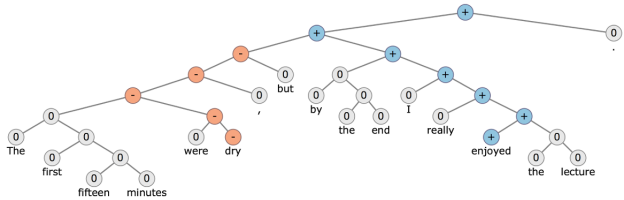
[FaceNet - Google 2015]



[Facial landmark detection CUHK 2014]

# DL Today: NLP

[Google Translate System - 2016]



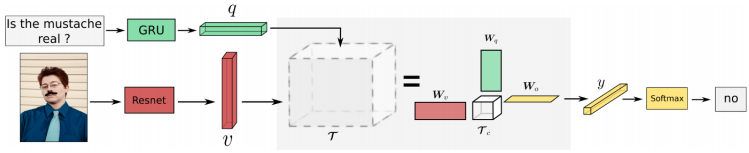

[Socher 2015]

# DL Today: NLP

[Google Inbox Smart Reply]

[Amazon Echo / Alexa]

# DL Today: Vision + NLP




[VQA - Mutan 2017]

[Karpathy 2015]

# DL Today: Image translation

[DeepDream 2015]

[Gatys 2015]

[Ledig 2016]

# DL Today: Generative models

Sampled celebrities [Nvidia 2017]

# DL Today: Generative models

StackGAN v2 [Zhang 2017]

# Language / Image models

Open-AI GPT-3, or DALL-E: https://openai.com/blog/dall-e/

TEXT PROMPT

an armchair in the shape of an avocado [...]

AI-GENERATED IMAGES

View more or edit prompt ↓

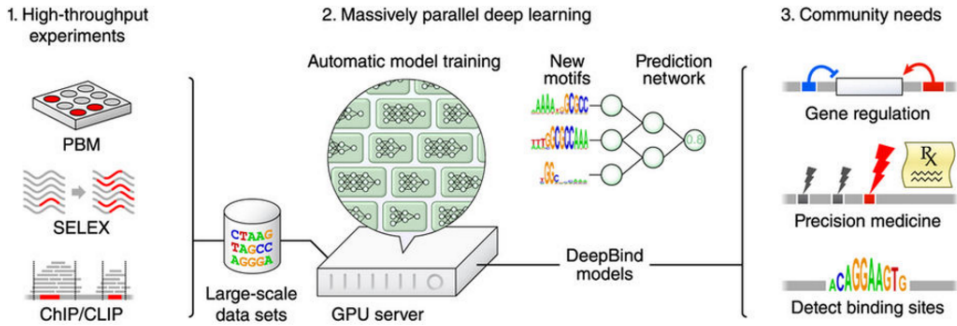
TEXT PROMPT

a store front that has the word 'openai' written on it [...]

AI-GENERATED IMAGES

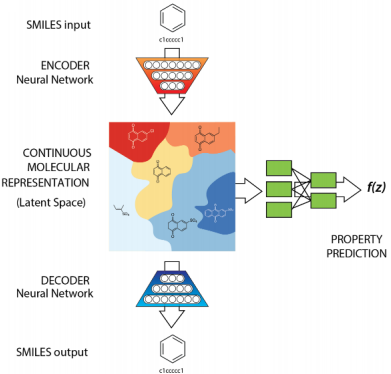
View more or edit prompt ↓

# DL in Science: Genomics




[Deep Genomics 2017]

# DL in Science: Chemistry, Physics




[Gómez-Bombarelli 2016]

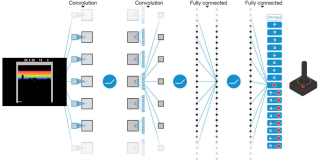
[Tompson 2016]

# DL for AI in games

w 20/30/0, b 8/40/2

1...e5 g3 d5 cxd5 ♘f6 ♗g2 ♘xd5 ♘f3

[Deepmind AlphaGo / Zero 2017]




[Atari Games - DeepMind 2016]

[Starcraft 2 for AI research]

AlphaGo/Zero: Monte Carlo Tree Search, Deep Reinforcement Learning, self-play

# Frameworks and Computation Graphs

PYTORCH

Microsoft CNTK

Caffe2

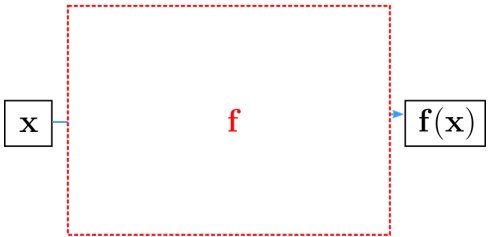
dmlc mxnet

gensim

spaCy

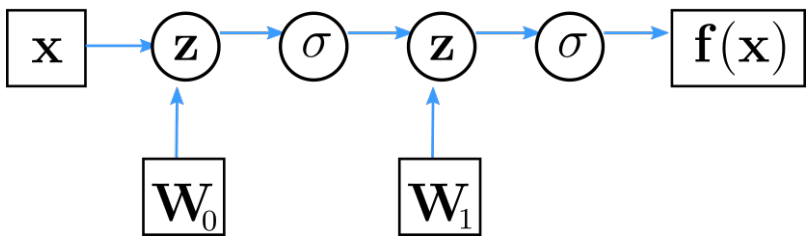
theano

# Computation Graph



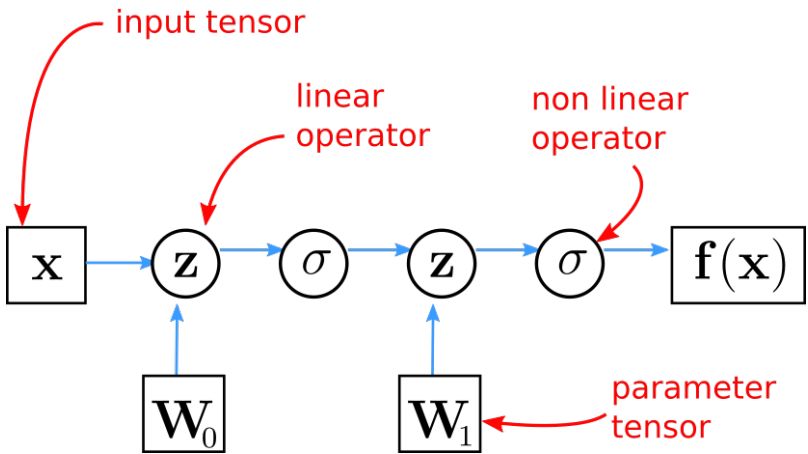

Neural network = parametrized, non-linear function

# Computation Graph



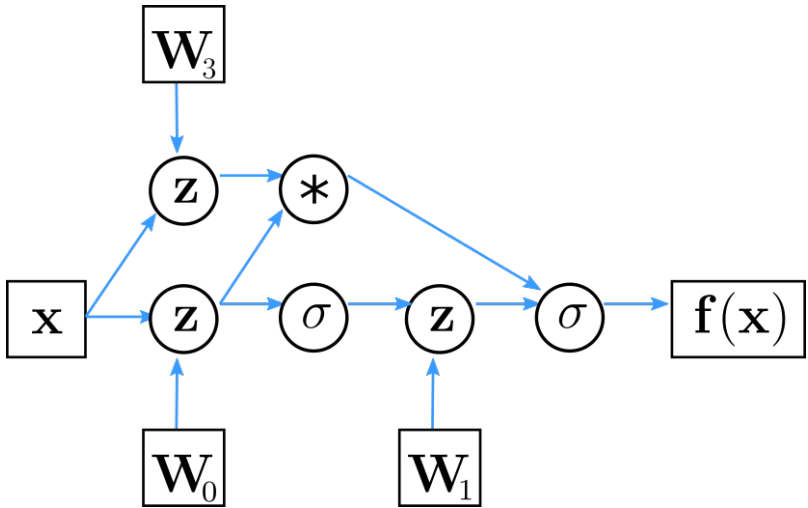

Computation graph: Directed graph of functions, depending on parameters (neuron weights)

# Computation Graph



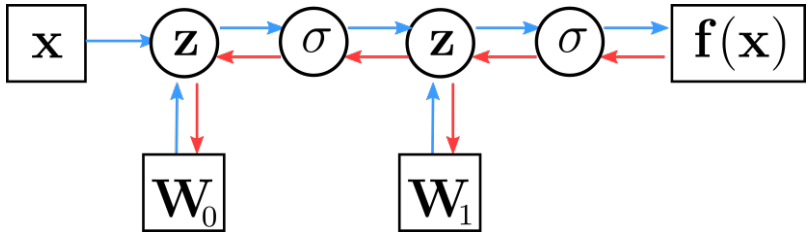

Combination of linear (parametrized) and non-linear functions

# Computation Graph




Not only sequential application of functions

# Computation Graph



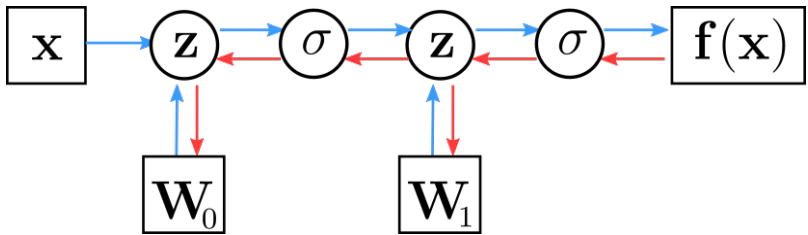

Automatic computation of gradients: all modules are **differentiable**!

Theano (now Aesara), **Tensorflow 1**, etc. build a static computation graph via static declarations.

**Tensorflow 2**, **PyTorch**, **JAX**, etc. rely on dynamic differentiable modules: “define-by-run”.

Vector computation on **CPU** and accelerators (**GPU** and **TPU**).

# Computation Graph



# Simple keras implementation

```
model = Sequential()
model.add(Dense(H, input_dim=N))  # defines W0
model.add(Activation("tanh"))
model.add(Dense(K))               # defines W1
model.add(Activation("softmax"))
```