ULB, Oct/2023

Introduction to Machine Learning Methods

Bryan Zaldívar (IFIC, Valencia)

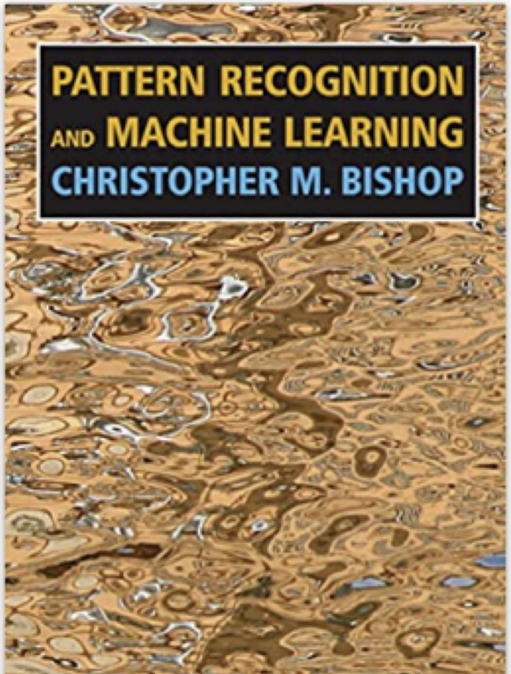
OUTLINE OF THE LECTURES

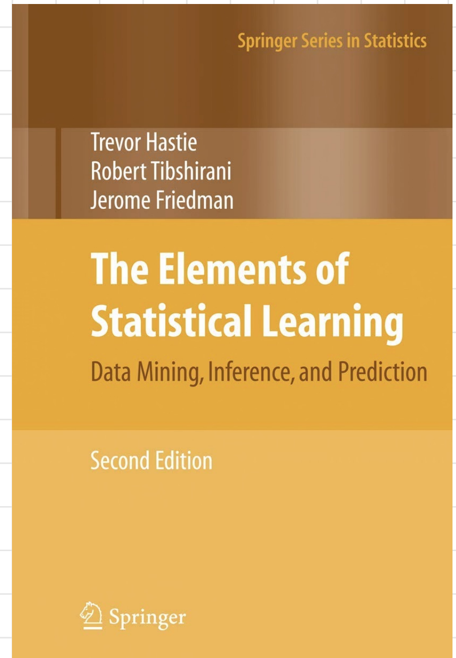
1. Overview of Machine Learning
2. Summary of statistics

} today

3. Regression & overfitting control
4. Bayesian learning
5. Classification methods
6. Neural networks
7. ... you decide

MAIN BIBLIOGRAPHY



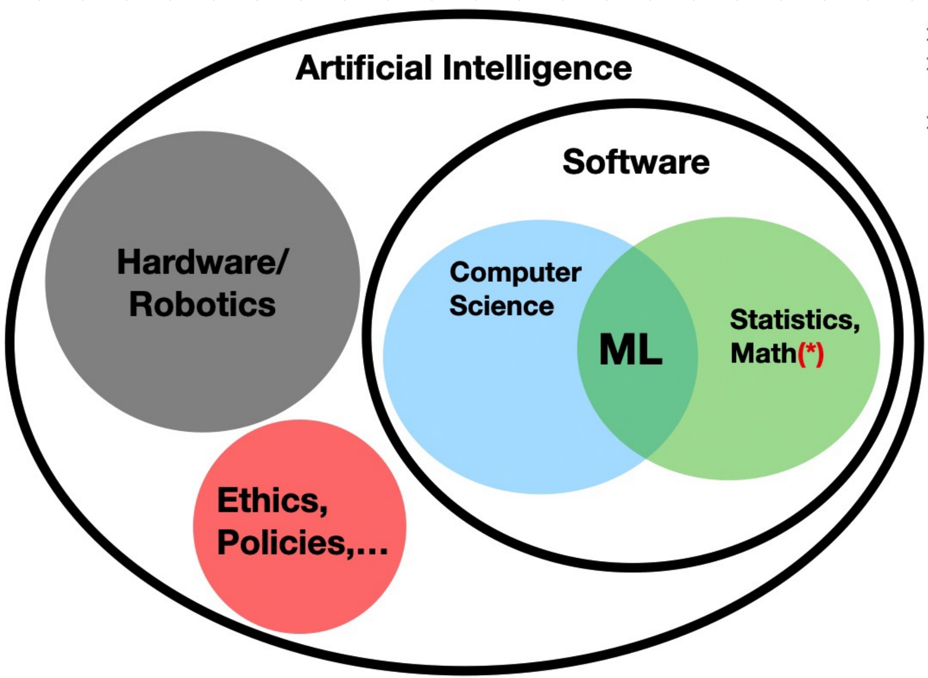


Lecture #1: Overview of ML

you are not supposed to understand these slides in detail...

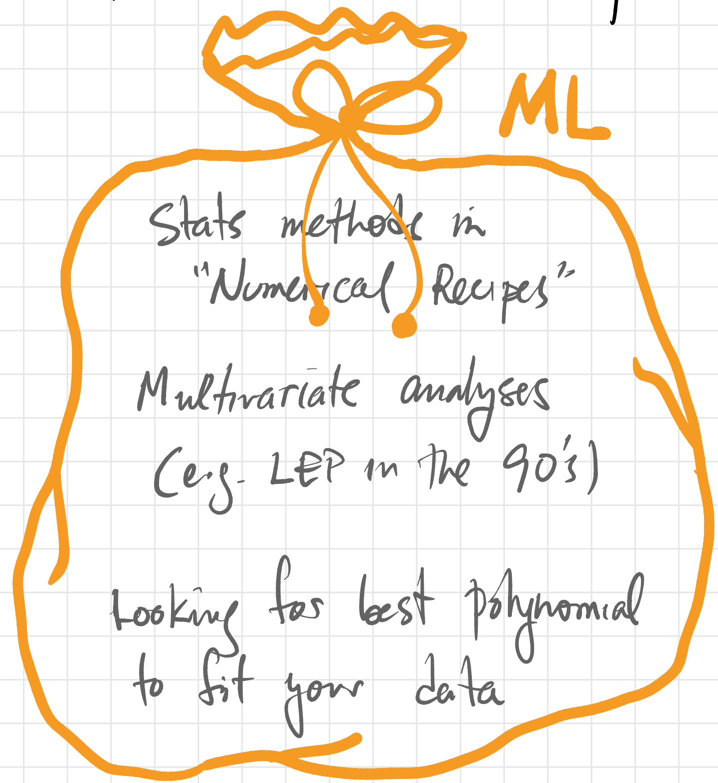
just watch them in a relaxed mood...

DEFINITION



(*) should be tailored to the domain expertise (e.g. physics) !

Very broadly... ML is about implementing stats on a computer



DEFINITION

"Machine Learning" coined by Arthur Samuel (IBM, 1959)

Quote by Tom M. Mitchell (prof. US, 1997):

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

DEFINITION

Historically, ML associated to:

- Games (chess, checkers, Go, ...)
- Image / sound / video recognition
- Natural language processing

Nowadays, ML synonym of:

- Data mining
- Big Data
- Deep Learning

SOME HISTORICAL MILESTONES

1950: Alan Turing's learning machine (based on primitive form of genetic algorithm)

1951: The 1st neural network is created (founded by Air Force Office)

1952: Arthur L. Samuel designed a computer program able to play checkers (IBM)

1967: Nearest Neighbor algorithm is created. The algorithm was used to map routes

1970: Back-propagation was invented

1985: NetTalk: a program that learns to pronounce written words in English

1997: IBM's Deep Blue beats Kasparov

2016: Google's "Alpha Go" beats an unhandicapped Go player

2020: Google's "Alpha Fold 2" is able to predict how proteins fold from amino acid sequence

2022: Open AI launches "ChatGPT 3", marking a breakthrough in Language Processing

WHAT DOES ML BRINGS TO PHYSICS ?

MOTIVATION

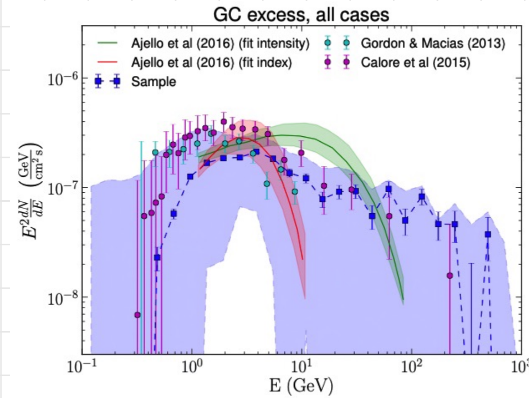
* Physical knowledge of background is limited
  (common problem in astrophysics)

MOTIVATION

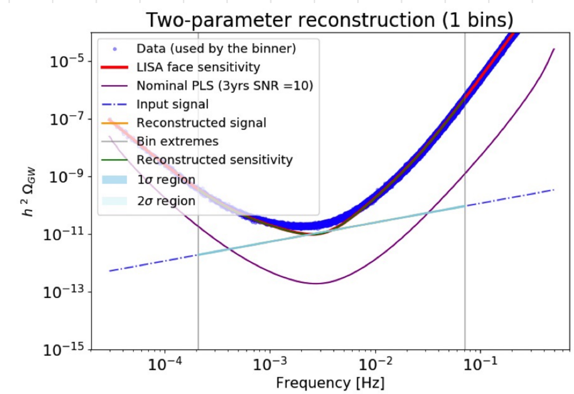
* Physical knowledge of background is limited
(common problem in astrophysics)

e.g. • γ-rays
• GW's

Fermi-LAT, 1704.03910



Caprini et al, 1906.09244

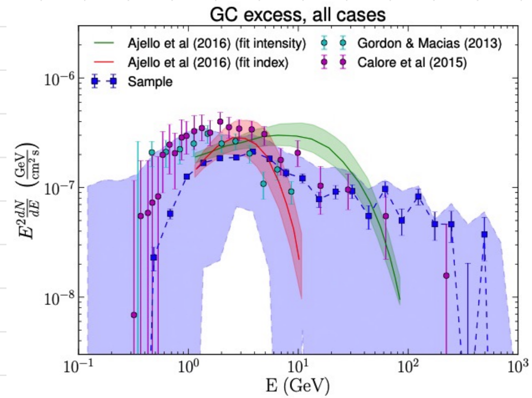


MOTIVATION

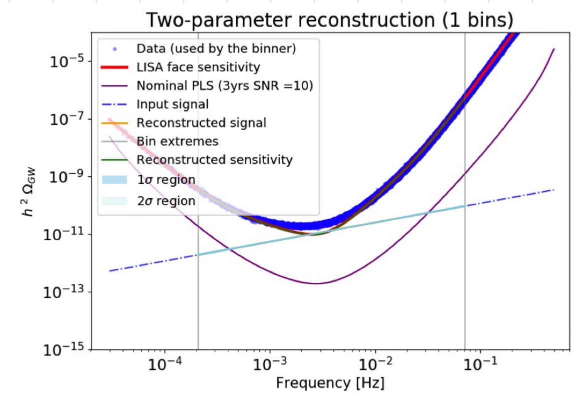
* Physical knowledge of background is limited
(common problem in astrophysics)

e.g. • γ-rays
• GW's

Fermi-LAT, 1704.03910



Caprini et al, 1906.09244



Signal + Background $\Longleftrightarrow$ Data

Signal ↳ interest (Physical model)

Background ↳ nuisance (data-driven model)

MOTIVATION

* Physics well known, but observables very complicated to compute

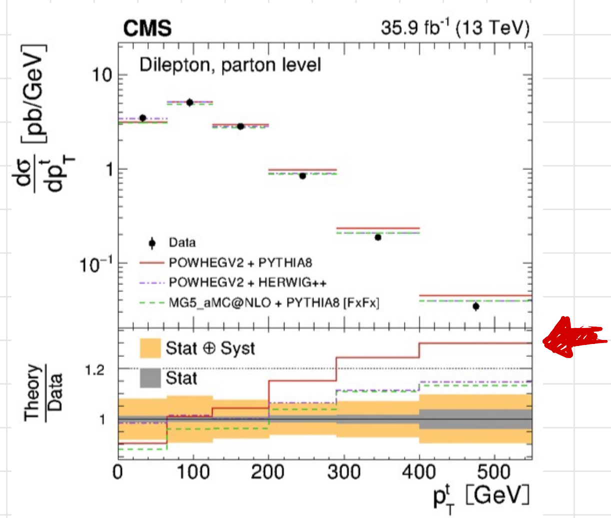
(complex topology, particle misidentification, etc)

MOTIVATION

* Physics well known, but observables very complicated to compute

(complex topology, particle misidentification, etc)

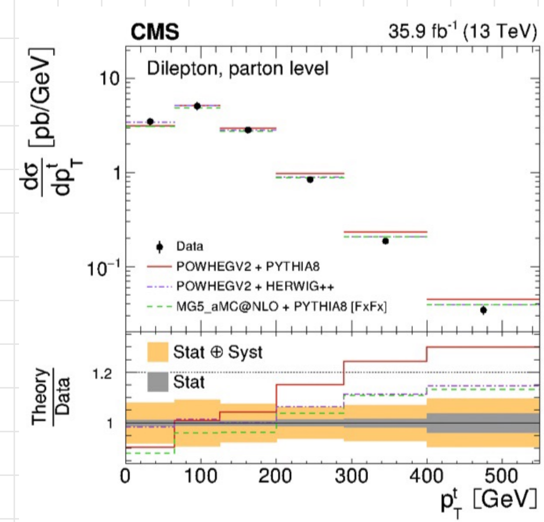
TOP-17-014-PAS



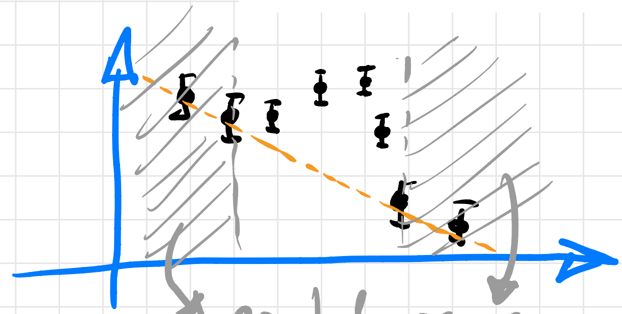
MOTIVATION

* Physics well known, but observables very complicated to compute (complex topology, particle misidentification, etc)

TOP-17-014-PAS



Data-driven methods to infer the background from inter/extrapolations



control regions

MOTIVATION

* Statistical bottleneck

MOTIVATION

* Statistical bottleneck

- More complex datasets

$\Downarrow$

More complex physical modeling

$\Downarrow$

More complex simulatory

MOTIVATION

* Statistical bottleneck

- More complex datasets

$\Downarrow$

More complex physical modeling

$\Downarrow$

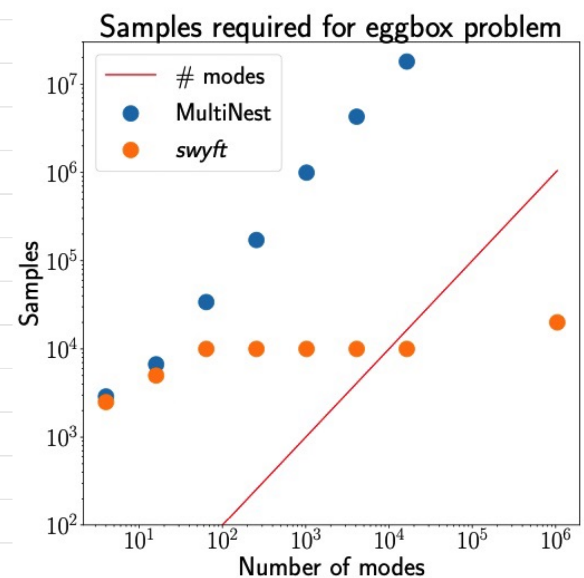
More complex simulators

$\Downarrow$

Better statistical treatment!

- better scaling
- more descriptive
- Higher statistical power

Miller et al, 2011.13951



MOTIVATION

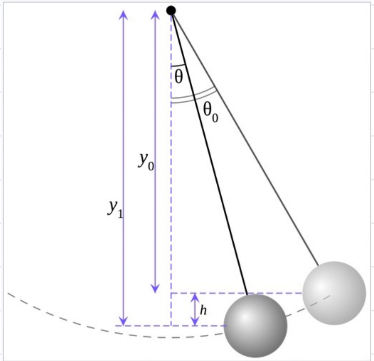
* Hints about the underlying physics

MOTIVATION

* Hints about the underlying physics

physical variables

- The intrinsic dimension of
  - Single pendulum : 2
  - Lava lamp : ?

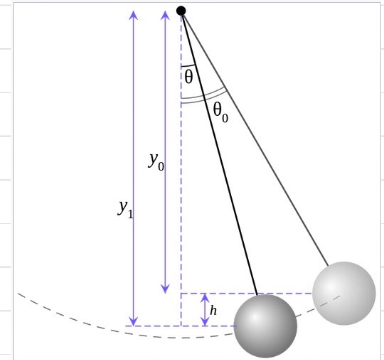


MOTIVATION

* Hints about the underlying physics

- The intrinsic dimension of

physical variables



  - Single pendulum : 2
  - Lava lamp : ?

ID = 7 – 8 ⟸ Statistical model (ML)

(see 2112.10755)

MAIN ML PARADIGMS

SUPERVISED LEARNING

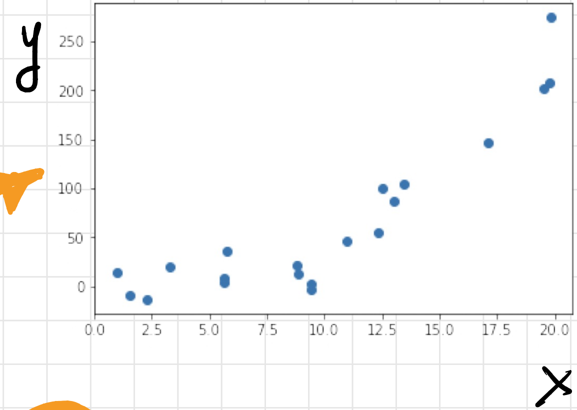
Data $\{X, \vec{y}\}$

x_1	x_2	x_3	y

⋮

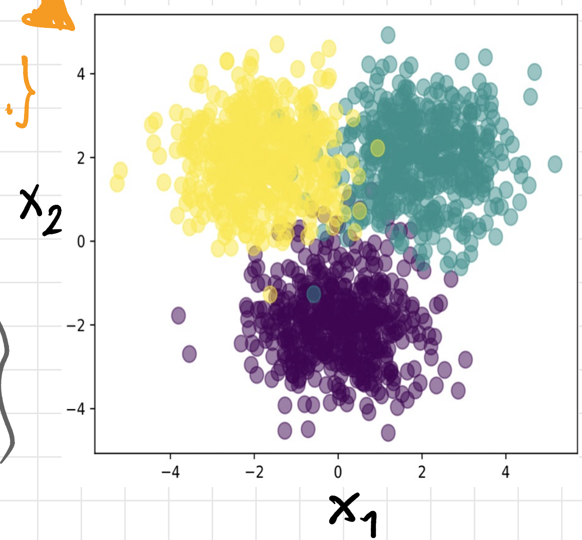
Regression

$y \in \mathbb{R}$



Classification

$y \in \{\ell_1, \ell_2, \ell_3, \ldots\}$



Tasks:

1 - Statistical inference of the data's probability distribution

2 - Predict the output for a new point

SUPERVISED LEARNING

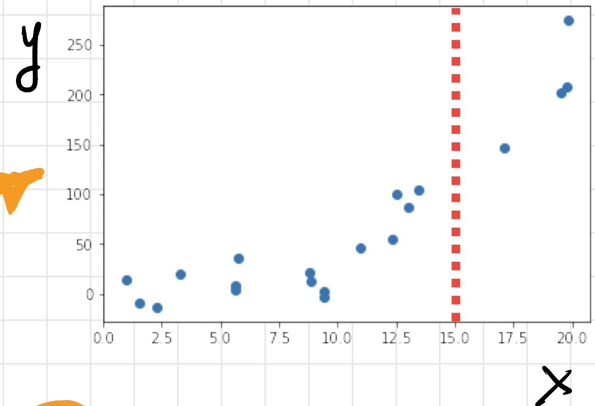
Data $\{X, \vec{y}\}$

x_1	x_2	x_3	y

$\vdots$

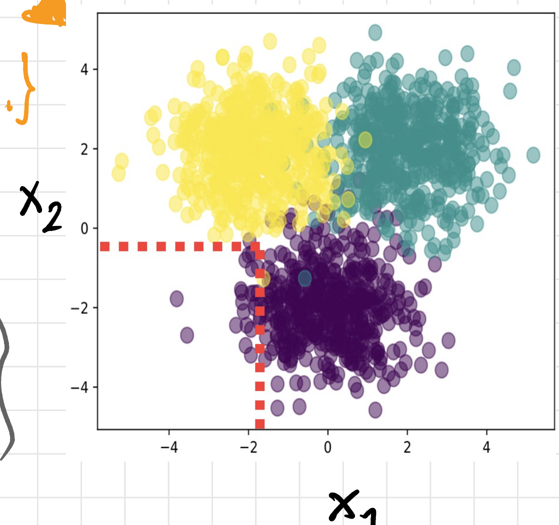
Regression

$y \in \mathbb{R}$



Classification

$y \in \{l_1, l_2, l_3, \dots\}$



Tasks:

1 - Statistical inference of the data's probability distribution

2 - Predict the output for a new point

SUPERVISED LEARNING

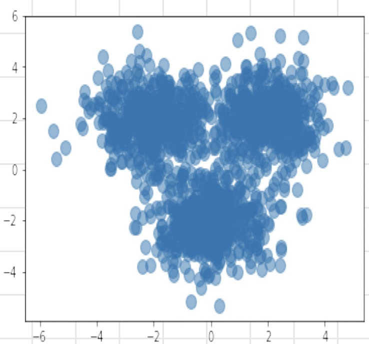
How is this different from fitting a function to some data?

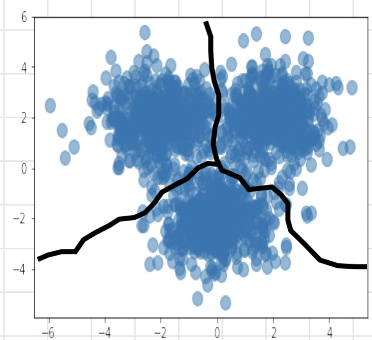
- Optimization algorithms, adapted to arbitrarily complex fitting functions
- Procedures to avoid overfitting (beyond e.g. $\chi^2/_{D.O.F}$)
- Choose the best function from a catalog
- Types of functions typically used (Neural networks, Decision Trees, kernels, ...)

UNSUPERVISED LEARNING

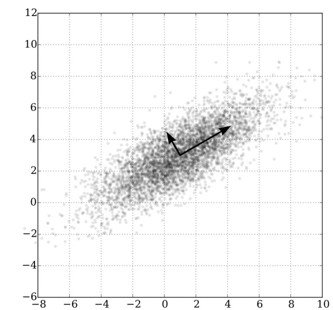
Data $\{X\}$ [no output / labels]

1) Clusterize the data

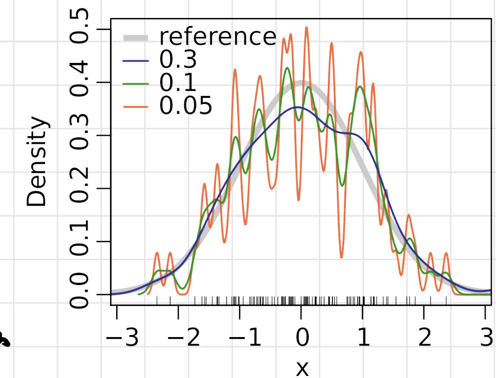




2) Dimensionality reduction



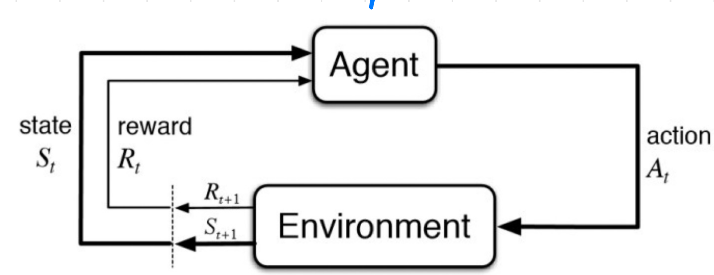
3) Probability density estimation



REINFORCEMENT LEARNING

(similar in spirit to the way humans learn from their environment)

computer program



"Environment"

Sistem in which the program operates

"Policy"

Method to map the program's state to an action

"Action"

Done by the program to move to a new state

"Reward"

Feedback from the environment

RL commonly used for learning to play games (chess, Go, ...)

Nowadays also used in science; eg. physics (quantum systems & computing) MCMC improvements, ...

DOI: 10.1103/PhysRevE.98.063303

BTW ChatGPT uses a form of RL!