What is ML?

Machine Learning with R

Basel R Bootcamp

May 2019

What do you think?

No Googling :)

What is machine learning?

Machine learning is...

...a **field of artificial intelligence**...

...that uses **statistical techniques**...

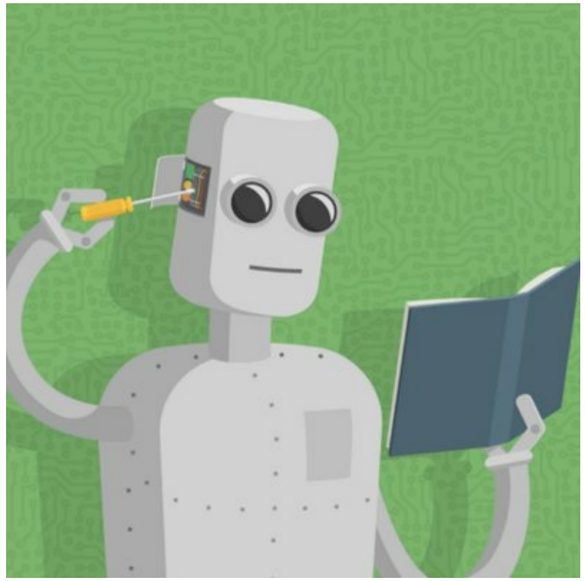
...to allow computer systems to **"learn"**,...

...i.e., to progressively **improve performance** on a specific task...

...from small or large amounts of **data**,...

....**without being explicitly programmed**....

....with the goal to **discover structure** or improve decision making and predictions.



from medium.com

ML's origin

Easy to confuse

AI is **intelligence demonstrated by machines**, in contrast to the natural intelligence displayed by humans and animals.

Statistics is a **branch of mathematics** dealing with data collection, organization, analysis, interpretation and presentation.

Big Data deals with data sets that are **too large or complex** to be dealt with by traditional data-processing application software.

Data Science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to **extract knowledge and insights** from structured and unstructured data.

Data-driven decisions

Predicting Heart Attacks

You are an intake nurse at an emergency room.

A patient comes in complaining of chest pain and thinks he is having a heart attack

How do you decide whether or not the patient is really having a heart attack?

from medium.com

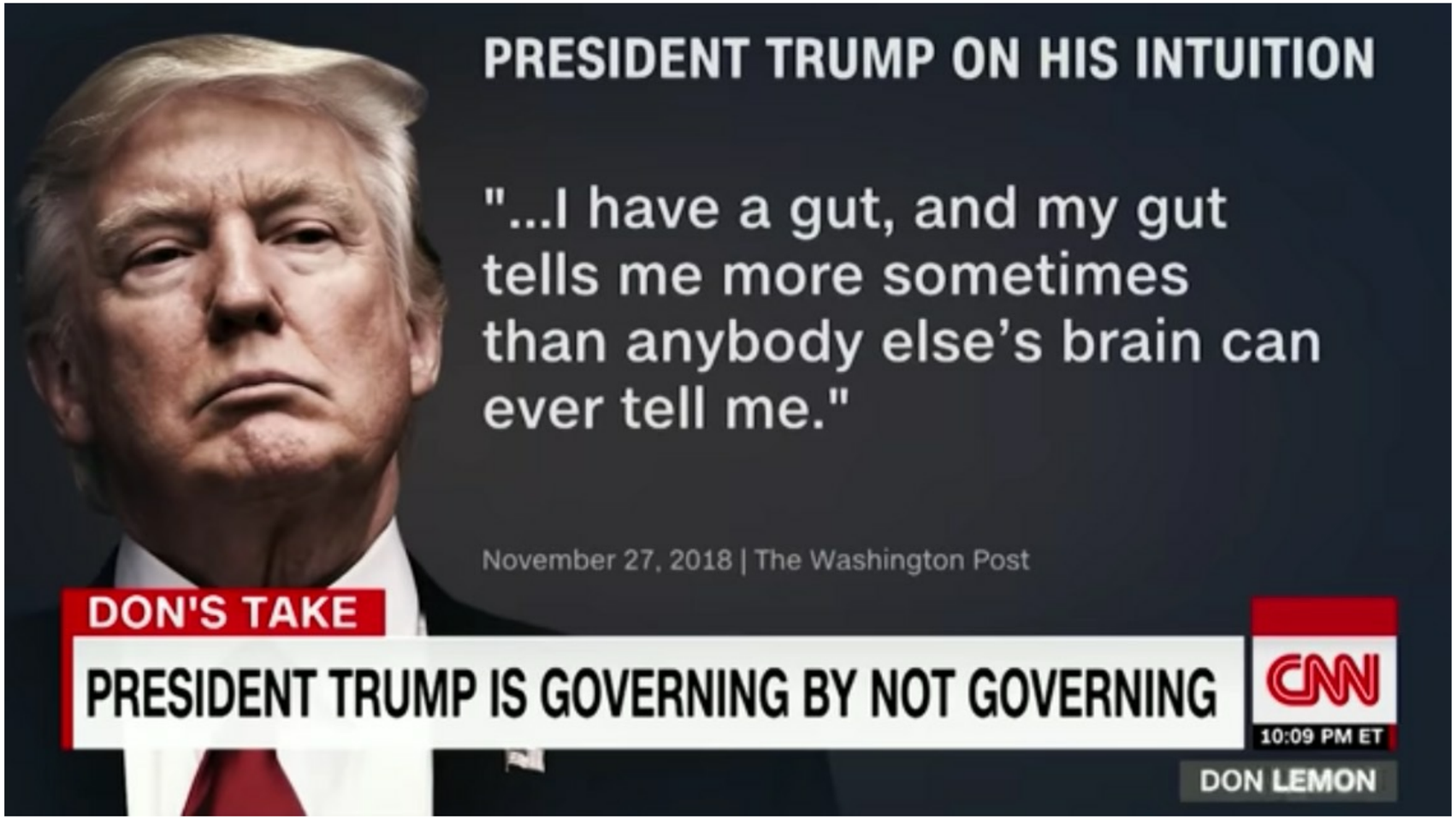
Predicting Sales

You are an analyst at a retail corporation.

The executive team is considering whether or not to open a new retail location in Basel.

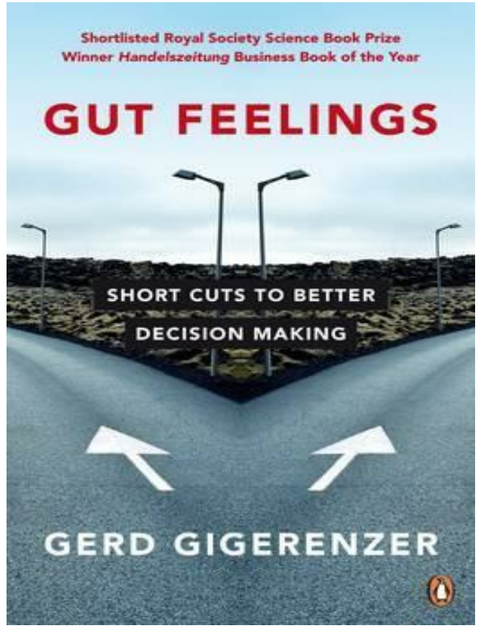
How can you predict what the sales of the new store would be?

from thirdmanrecords.com



from cnn.com

Can we trust our intuition?



from amazon.uk

from medium.com

from thirdmanrecords.com

Problems with intuition

Intuition...

What problems arise from trusting one's intuition?

from medium.com

from thirdmanrecords.com

Problems with intuition

Intuition...

...might not tell you anything about **how the prediction** has been made.

...could be based on **reasons other than accuracy**, e.g., self protection.

...impossible to know if **critical information is being ignored**.

...is **difficult to reproduce** and rarely permits rigorous evaluation.

...can always be **defended in hindsight**.

from medium.com

from thirdmanrecords.com

Machine learning is data-driven

Data-driven, ML-based heart rate prediction

Based on **data** from past patients **at this hospital**, a **regression model**, using the patient's **age, cholesterol level, and ecg**, **predicts** the probability that this patient is having a heart attack is only **45%**.

	diagnosis	age	sex	cp	trestbps	chol	fbs	restecg
1	FALSE	63	1	ta	145	233	1	hypertrophy
2	TRUE	67	1	a	160	286	0	hypertrophy
3	TRUE	67	1	a	120	229	0	hypertrophy
4	FALSE	37	1	np	130	250	0	normal
5	FALSE	41	0	aa	130	204	0	hypertrophy
6	FALSE	56	1	aa	120	236	0	normal
7	TRUE	62	0	a	140	268	0	hypertrophy
8	FALSE	57	0	a	120	354	0	normal
9	TRUE	63	1	a	130	254	0	hypertrophy
10	TRUE	53	1	a	140	203	1	hypertrophy
11	FALSE	57	1	a	140	192	0	normal

from medium.com

from thirdmanrecords.com

Benefits of ML

Machine learning *algorithms*....

What are benefits of machine learning?

from medium.com

from thirdmanrecords.com

Benefits of ML

Machine learning *algorithms*....

...can integrate all available **data**.

...make **explicit, reproducible, and quantitative** predictions of variables of interest.

...can tell you **which variables are important** and which are not.

...can give you **probability estimates**, and estimated errors, rather than single decisions or point estimates.

...can reveal **novel insights** about your data.

...can be **automated**.

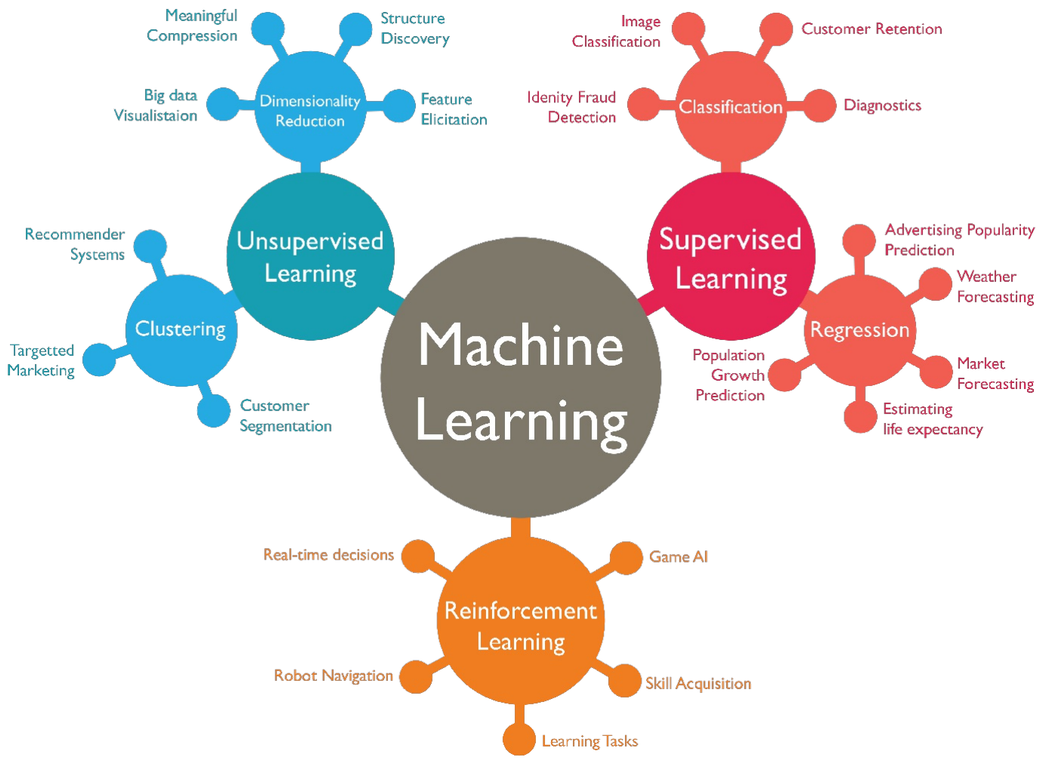
from medium.com

from thirdmanrecords.com

Types of machine learning tasks

There are many types of machine learning tasks, each of which call for different models.

We will focus on supervised machine learning.



from amazonaws.com

Data terminology

Term	Definition	Example
Case	A specific **observation** of data.	A patient, a site, etc.
Feature	An measurable **property** of cases. Also called predictors.	Age, temperature, country, etc.
Criterion	The **feature** that you want to **predict**.	Heart attack, sales, etc.
Data	Typically **rectangular** representation of cases (rows) and features (columns).	`.csv`, `.xls`, `.sav`, etc.

Criterion *Features*

Cases

diagnosis	age	sex	cp	trestbps	chol	fbs
FALSE	63	1	ta	145	233	1
TRUE	67	1	a	160	286	0
TRUE	67	1	a	120	229	0
FALSE	37	1	np	130	250	0
FALSE	41	0	aa	130	204	0
FALSE	56	1	aa	120	236	0
TRUE	62	0	a	140	268	0
FALSE	57	0	a	120	354	0
TRUE	63	1	a	130	254	0
TRUE	53	1	a	140	203	1
FALSE	57	1	a	140	192	0

Supervised learning

The **dominant type** of machine learning.

Supervised learning uses **labeled data** to learn **a model** that relates the criterion to the features.

<u>Verbal model</u>

```
if cp (chest pain) is not a
(asymptomatic) and age is larger than
60 then high probability of hearth
attack, otherwise low probability.
```

Criterion

diagnosis
FALSE
TRUE
TRUE
FALSE
FALSE
FALSE
TRUE
FALSE
TRUE
TRUE
FALSE

Model/ Algorithm

Some function that relates the criterion to the features.

criterion = *f*(features)

e.g., regression

Features

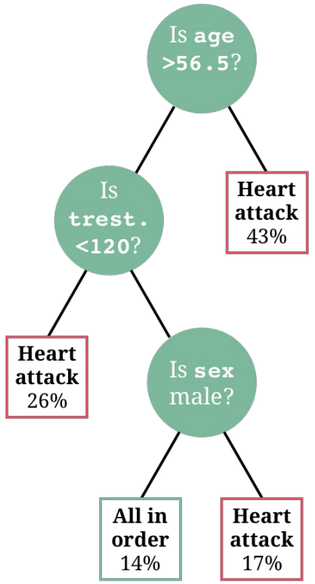
age	sex	cp
63	1	ta
67	1	a
67	1	a
37	1	np
41	0	aa
56	1	aa
62	0	a
57	0	a
63	1	a
53	1	a
57	1	a

3 key (supervised) models

Regression

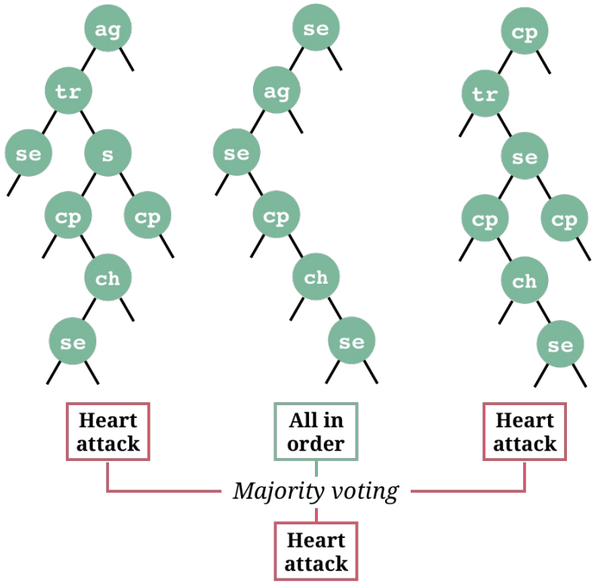
$$P\binom{\text{Heart}}{\text{attack}} = \beta_1 * \texttt{sex} + \beta_2 * \texttt{age} + \beta_3 * \texttt{tre.} + \beta_4 * \texttt{chol} + \beta_5 * \texttt{cp}$$

Decision tree



Random forest

(simplified)



2 types of supervised problems

There are two types of supervised learning problems that can often be approached using the same model.

Regression

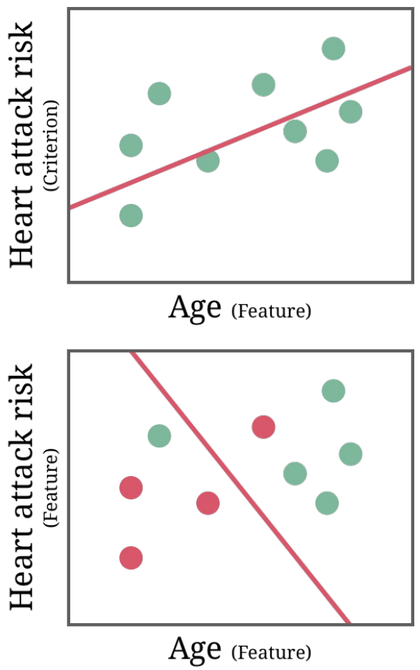
Regression problems involve the **prediction of a quantitative feature**.

E.g., predicting the cholesterol level as a function of age.

Classification

Classification problems involve the **prediction of a categorical feature**.

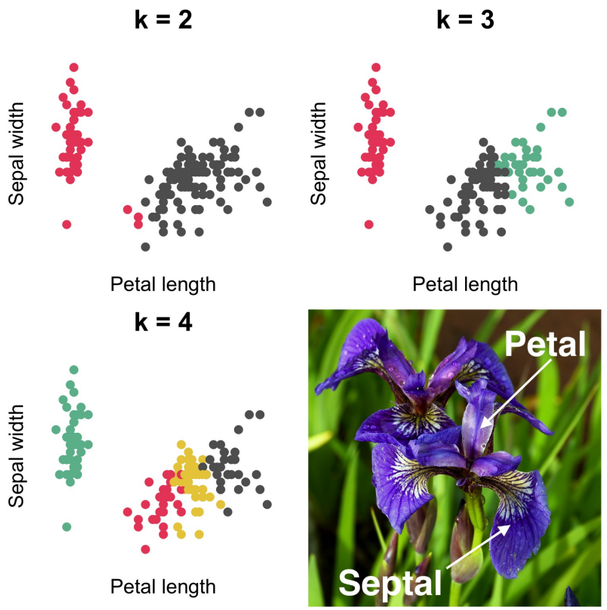
E.g., predicting the type of chest pain as a function of age.



Unsupervised learning

Analyzes the relationships among cases (**clustering**) or among features (**dimensionality reduction**) to **discover structures** such as groups or meta-features.

Approach	Description	Example
Clustering	Analyze distances between cases to identify **clusters of homogeneous cases**.	Types of customers or patients.
Dimension-ality reduction	Analyze correlations between features to identify **higher order features**.	Dimensions of personality or user experience.



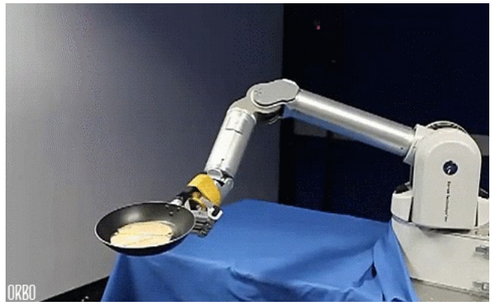
Reinforcement learning

Learns iteratively from minimal supervision provided by **performance feedback**.

RL is closely **related to psychological theories of learning**.

Examples

Application	Description
Model fitting	Iteratively **change model parameters** to improve prediction.
Robot movements	Iteratively **change movement** patterns to increase pancake-catch probability.
Games	Iteratively **change controller input** patterns to improve Mario Kart racing time.



from giphy.com

from nvidia.com

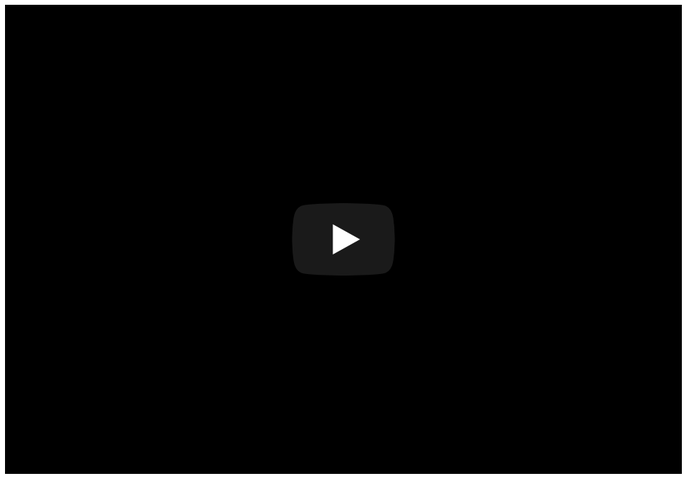
Reinforcement learning

Learns iteratively from minimal supervision provided by **performance feedback**.

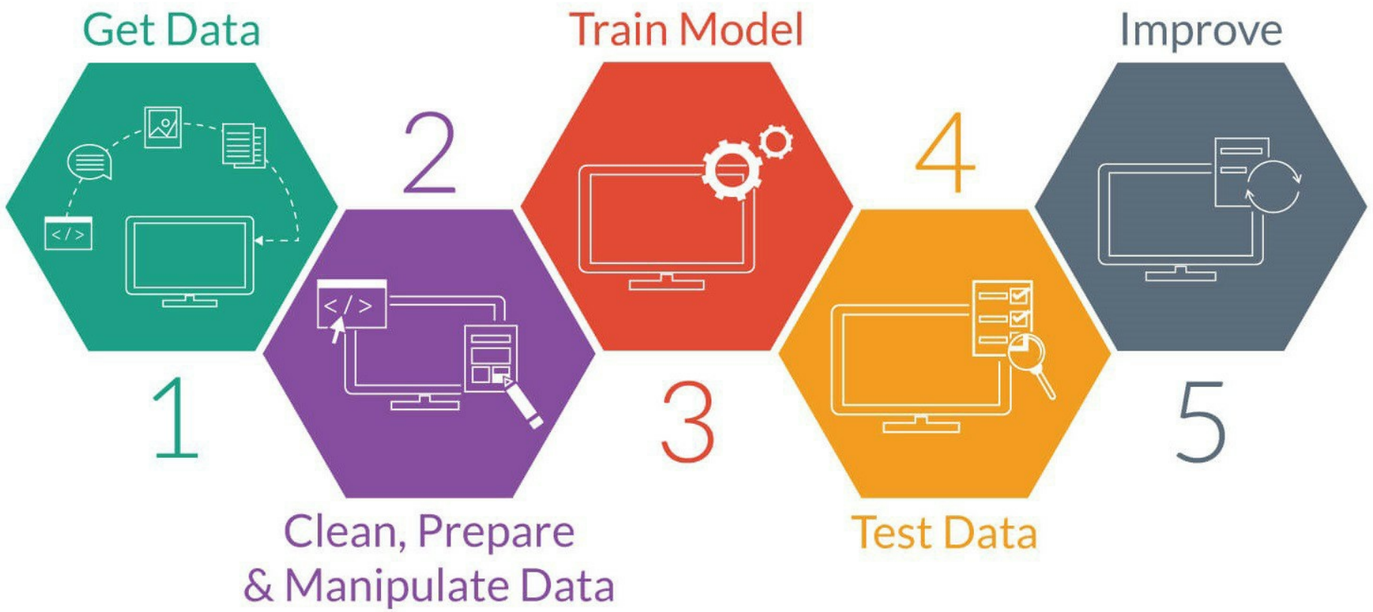
RL is closely **related to psychological theories of learning**.

Examples

Application	Description
Model fitting	Iteratively **change model parameters** to improve prediction.
Robot movements	Iteratively **change movement** patterns to increase pancake-catch probability.
Games	Iteratively **change controller input** patterns to improve Mario Kart racing time.



Machine learning is more than algorithms



from houseofbots.com

Schedule