Master Degree in Artificial Intelligence for Science and Technology

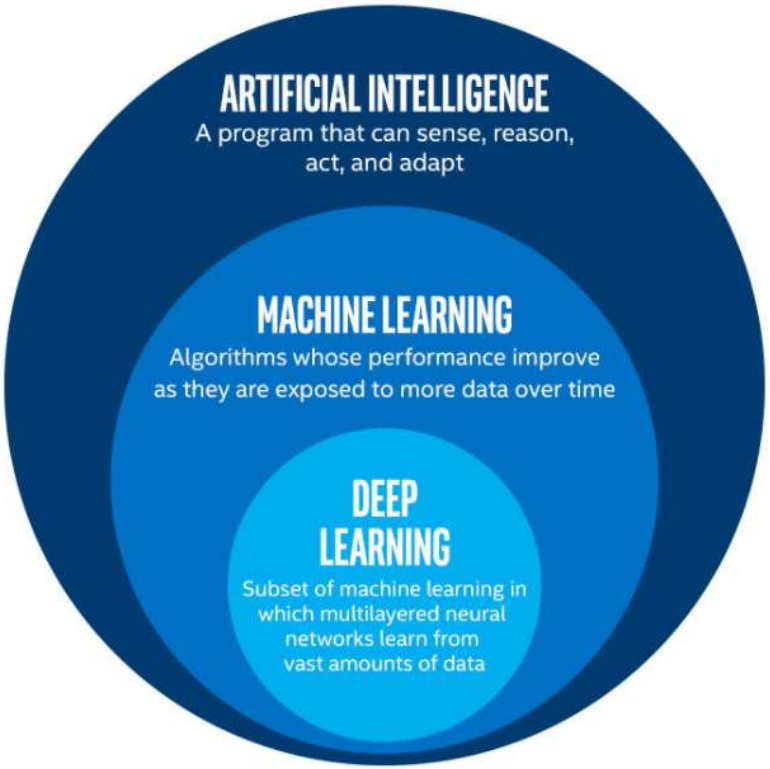
Unsupervised Learning

Fabio Stella
Department of Informatics, Systems and Communication
University of Milan-Bicocca
fabio.stella@unimib.it


Lecco
Bergamo
Saronno
E64
Monza
Rho
Milano
Brescia
Vigevano
E70
E62
Lodi
Crema
Pavia
Codogno
Cremona

Artificial Intelligence; intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans.

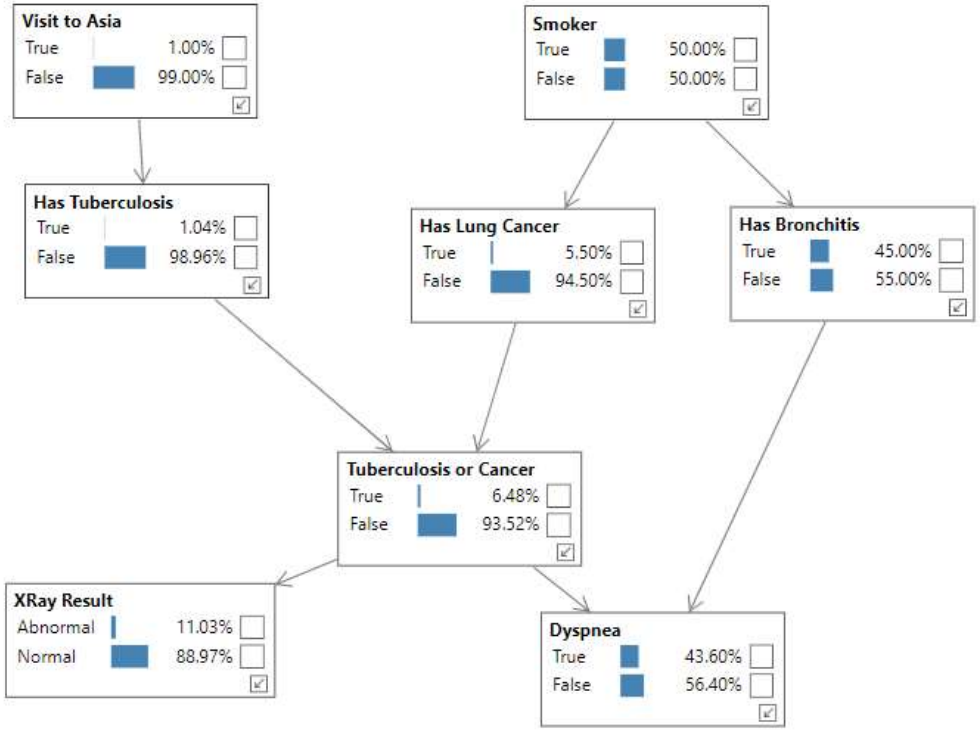
- Colloquially, the term Artificial Intelligence (**AI**) is used to describe machines/computers that mimic “cognitive” functions that humans associate with other human minds, such as "learning" and "problem solving".
- Two kinds of AI:
  - ✓ Weak
  - ✓ Strong



Artificial Intelligence

Bayesian Networks

- A type of statistical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG).
- Bayesian networks are ideal for taking an event that occurred and predicting the likelihood that any one of several possible known causes was the contributing factor.
- Structural Causal Models.

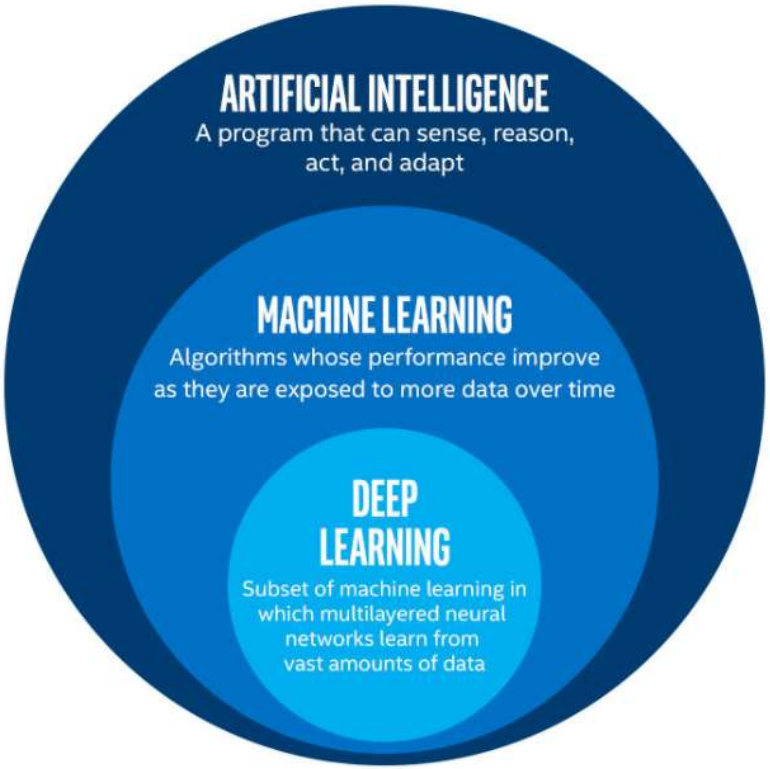


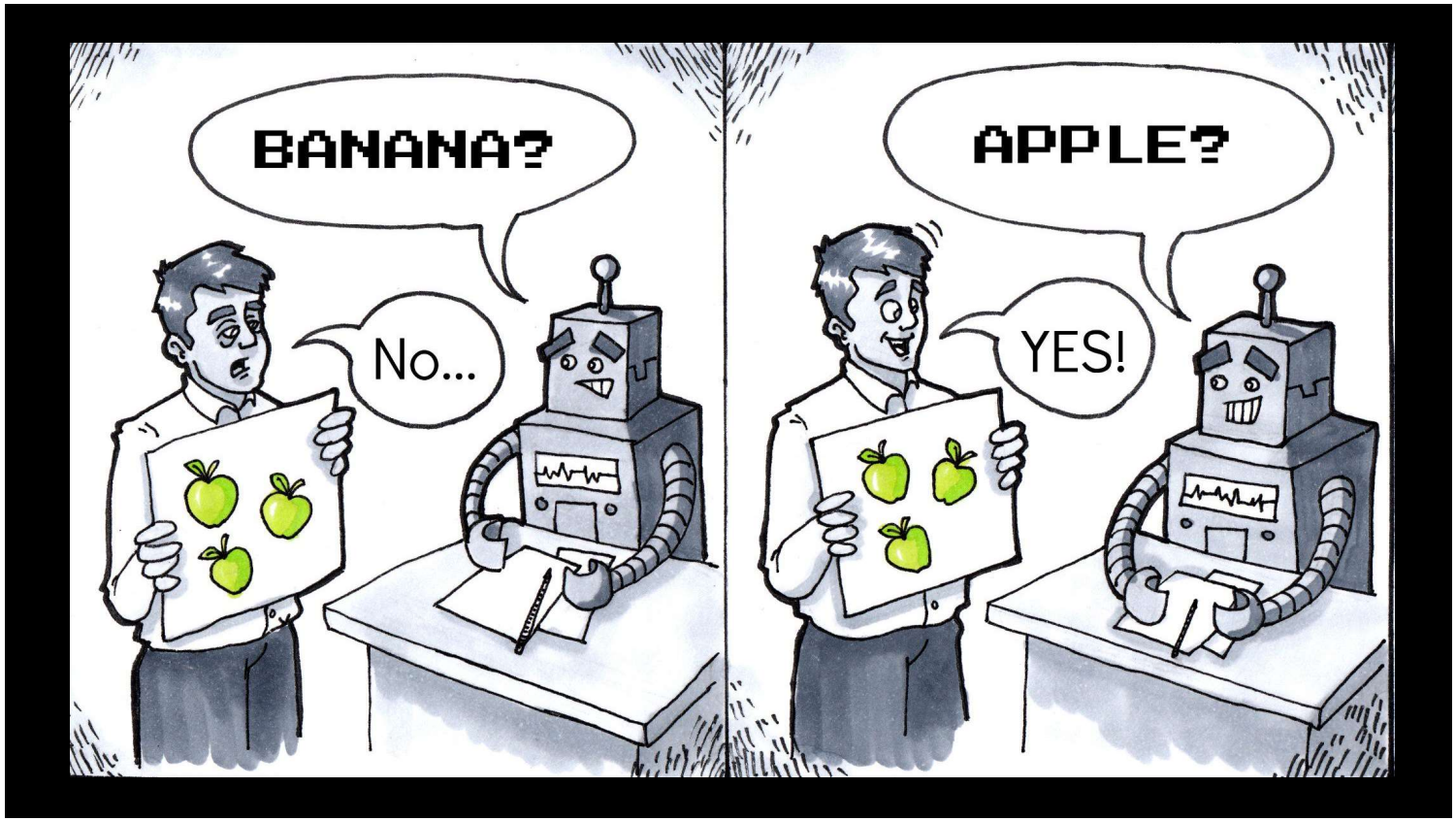
Machine Learning

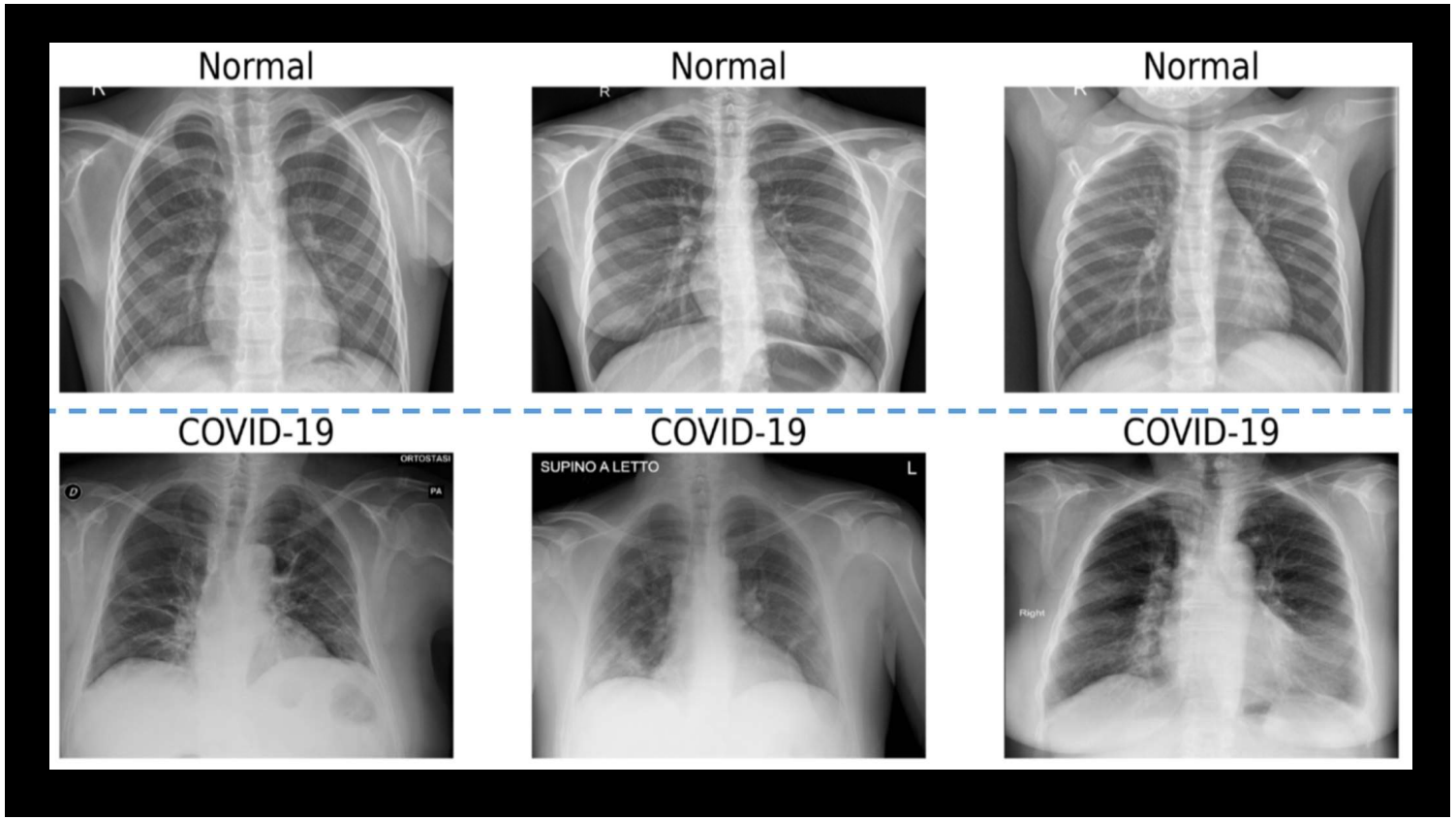
algorithms and statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions, relying on patterns and inference instead.

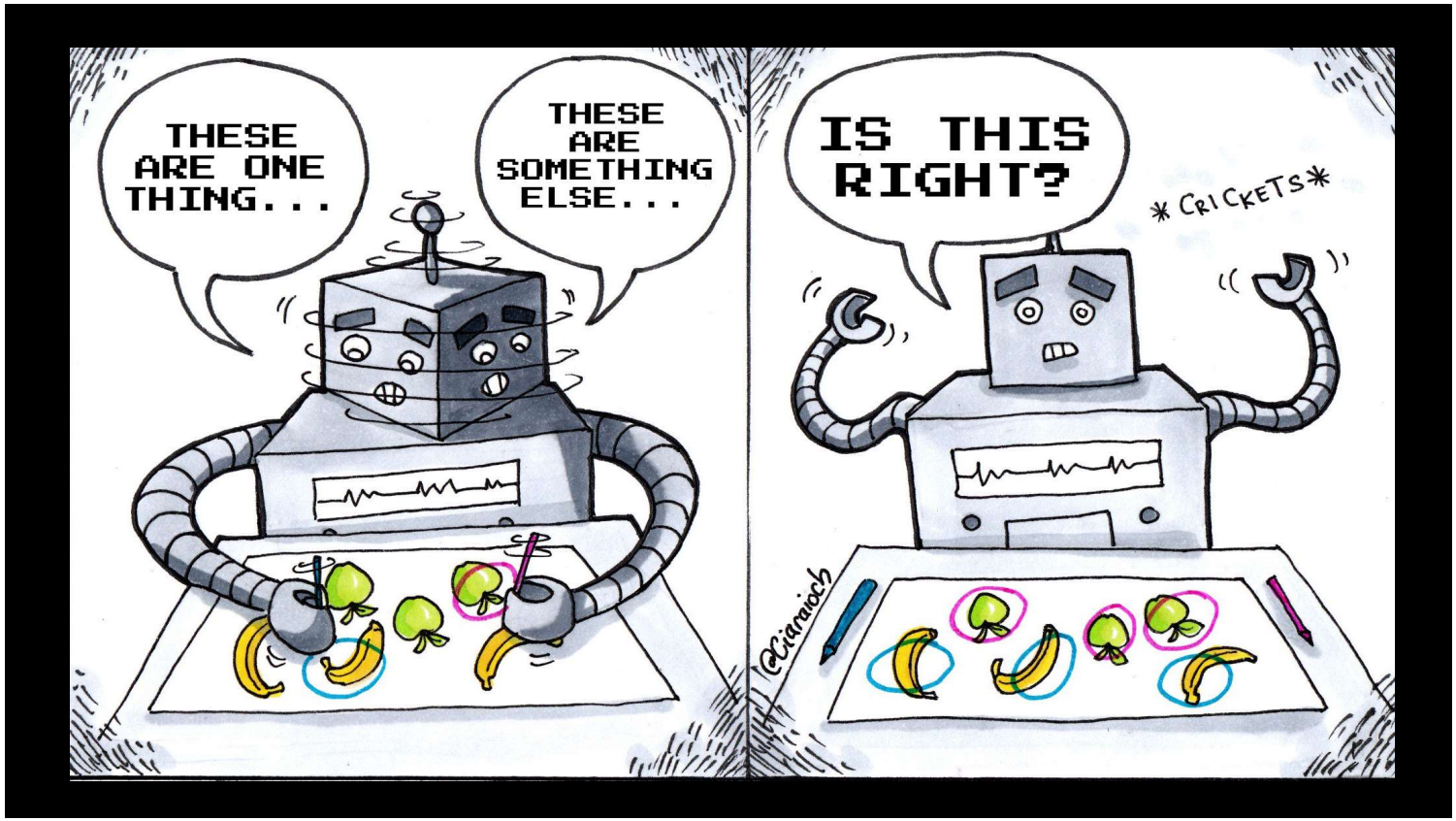
Three kinds of ML;

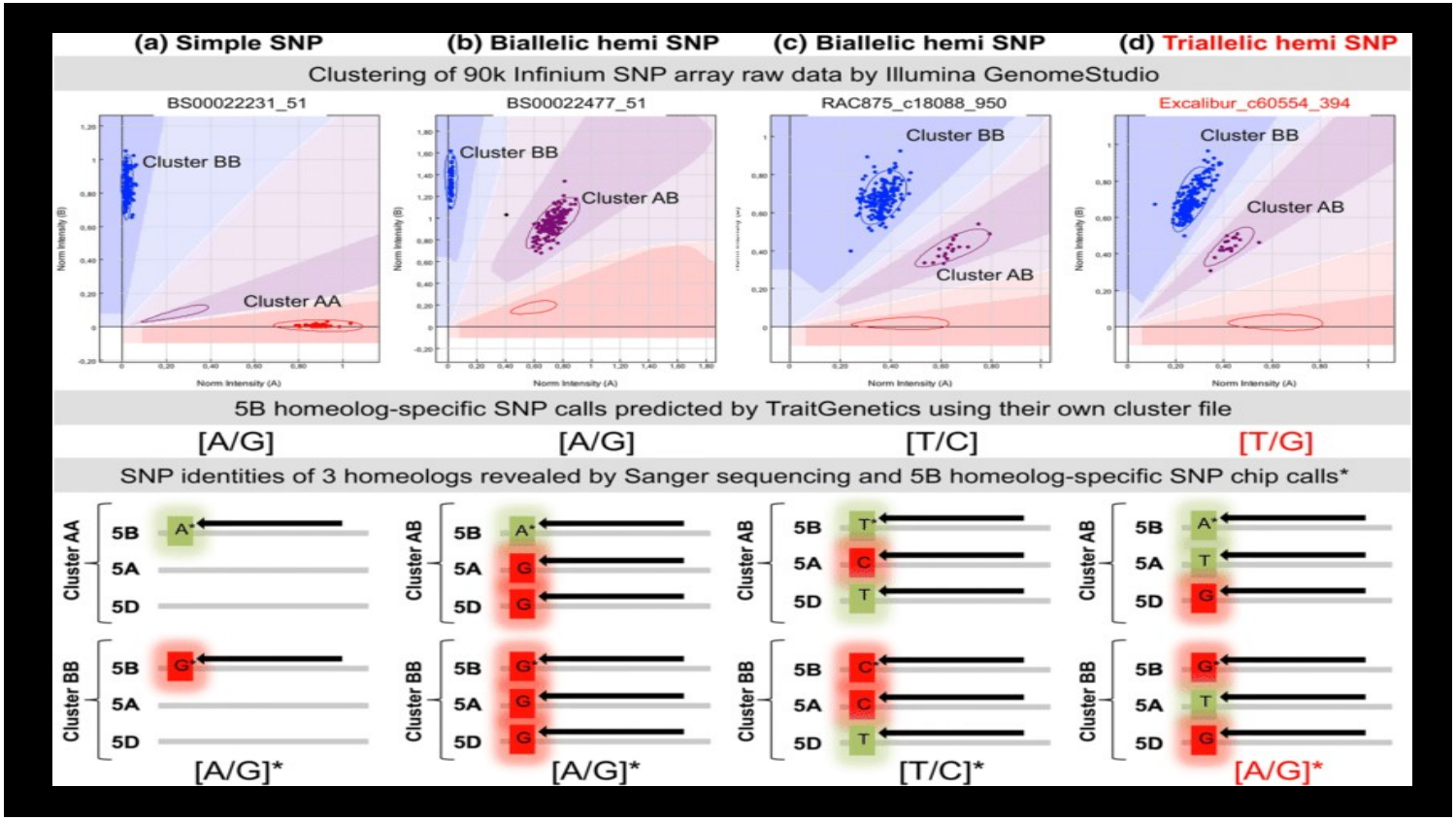
- Supervised
- Unsupervised
- Reinforcement Learning

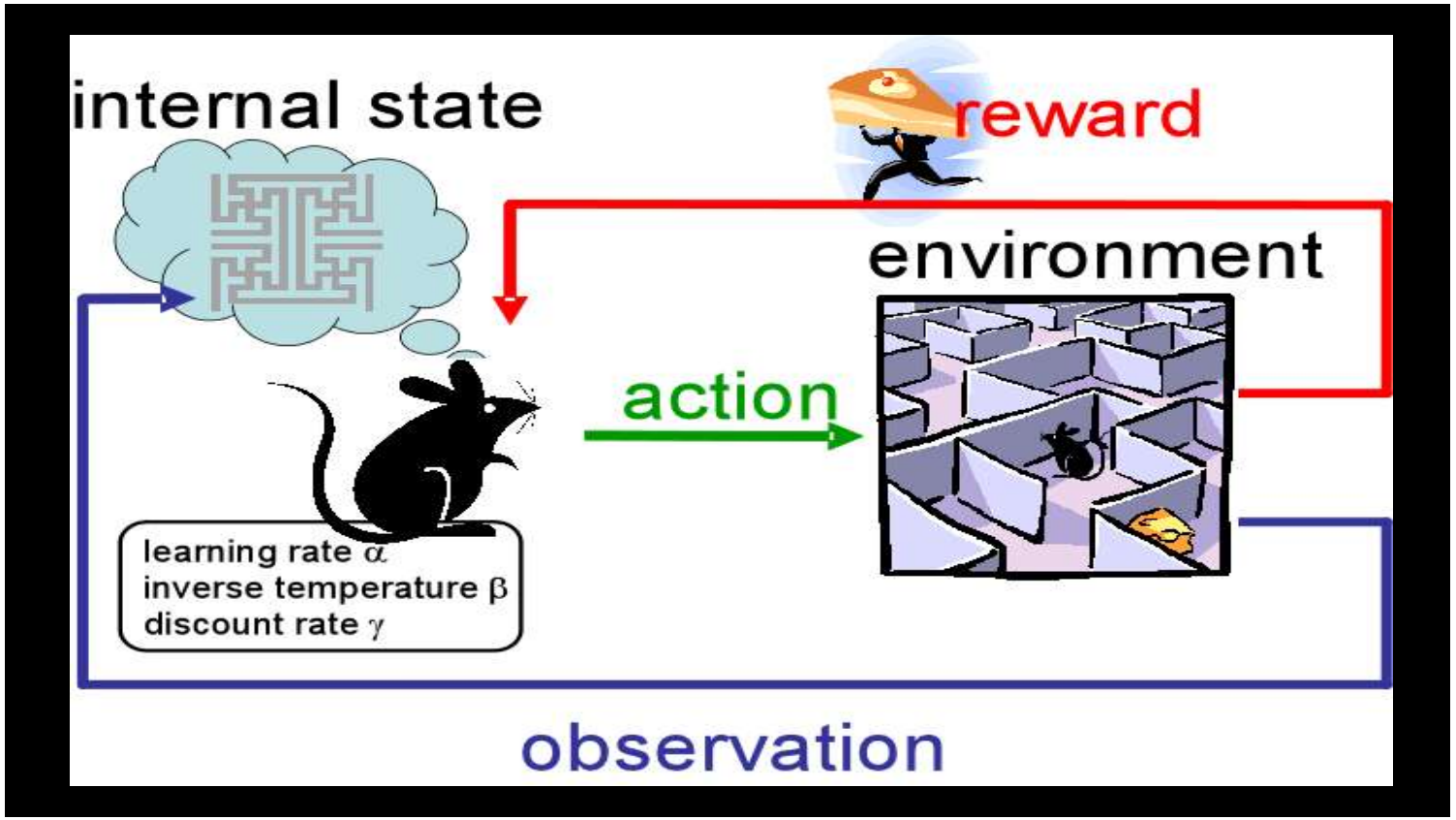



BANANA?
No...
APPLE?
YES!

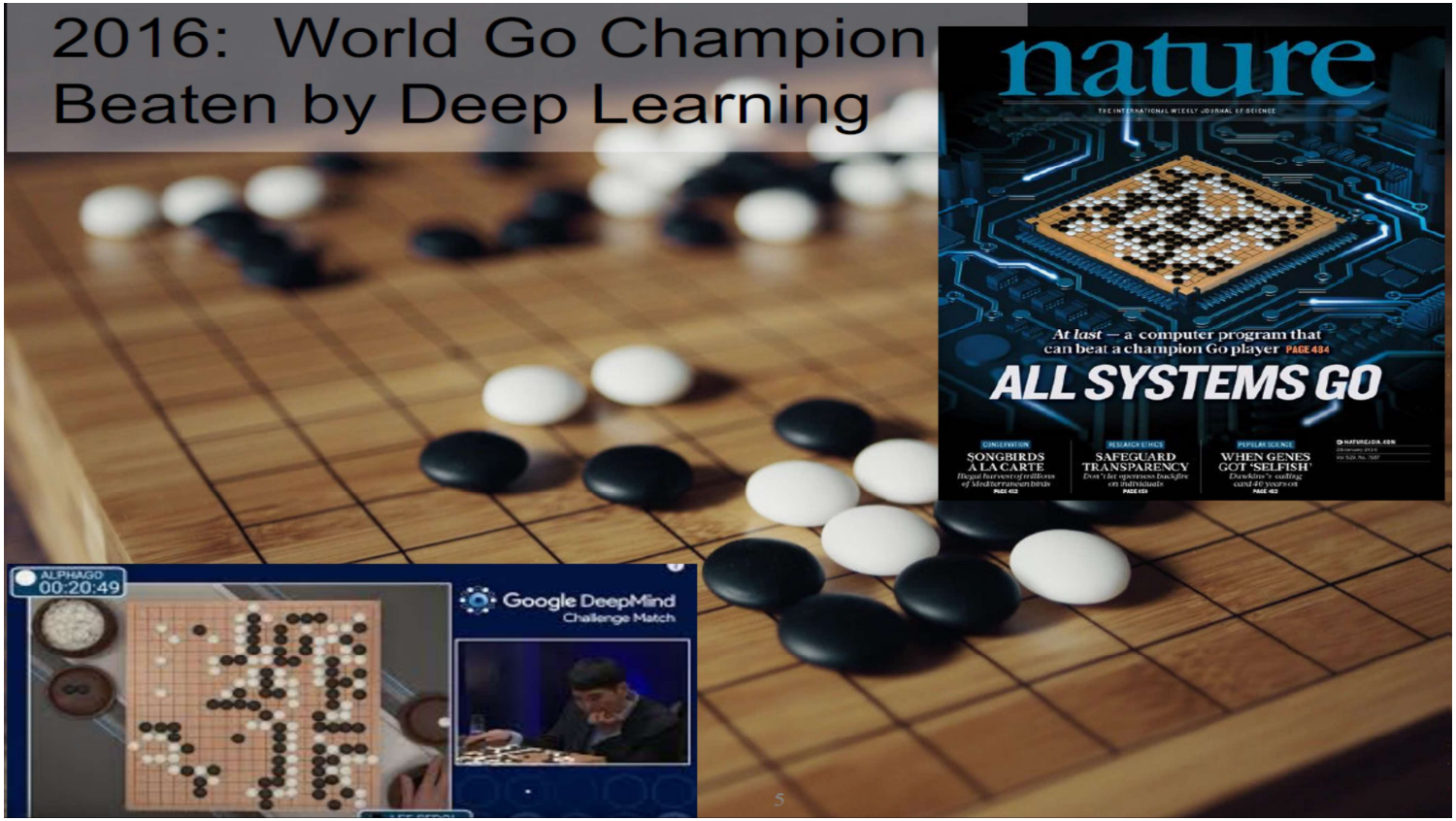

Normal
Normal
Normal
COVID-19
COVID-19
COVID-19
ORTOSTASI
PA
SUPINO A LETTO
L
Right


THESE ARE ONE THING...
THESE ARE SOMETHING ELSE...
IS THIS RIGHT?
CRICKETS
@Ciaraioch


(a) Simple SNP
(b) Biallelic hemi SNP
(c) Biallelic hemi SNP
(d) Triallelic hemi SNP
Clustering of 90k Infinium SNP array raw data by Illumina GenomeStudio
BS00022231_51
BS00022477_51
RAC875_c18088_950
Excalibur_c60554_394
Cluster BB
Cluster AB
Cluster AA
Norm Intensity (A)
Norm Intensity (B)
5B homeolog-specific SNP calls predicted by TraitGenetics using their own cluster file
[A/G]
[A/G]
[T/C]
[T/G]
SNP identities of 3 homeologs revealed by Sanger sequencing and 5B homeolog-specific SNP chip calls*
Cluster AA
Cluster AB
Cluster BB
5B
5A
5D
A*
G*
T*
C*
G
T
[A/G]*
[A/G]*
[T/C]*
[A/G]*


internal state
reward
environment
action
learning rate α
inverse temperature β
discount rate γ
observation

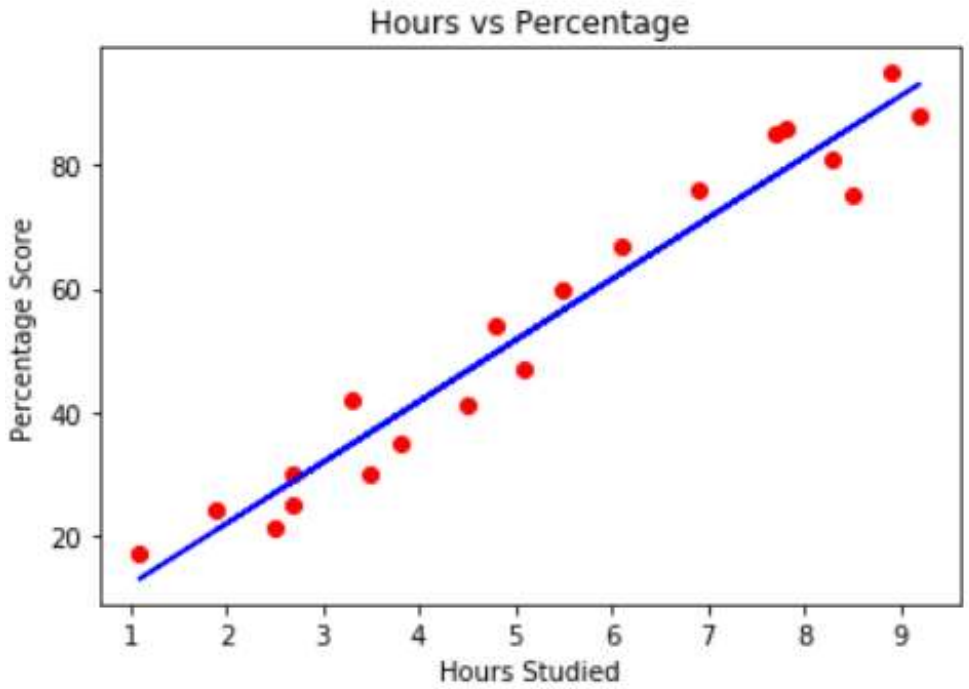
2016: World Go Champion Beaten by Deep Learning




(tensorflow) C:\Users\fra_v>

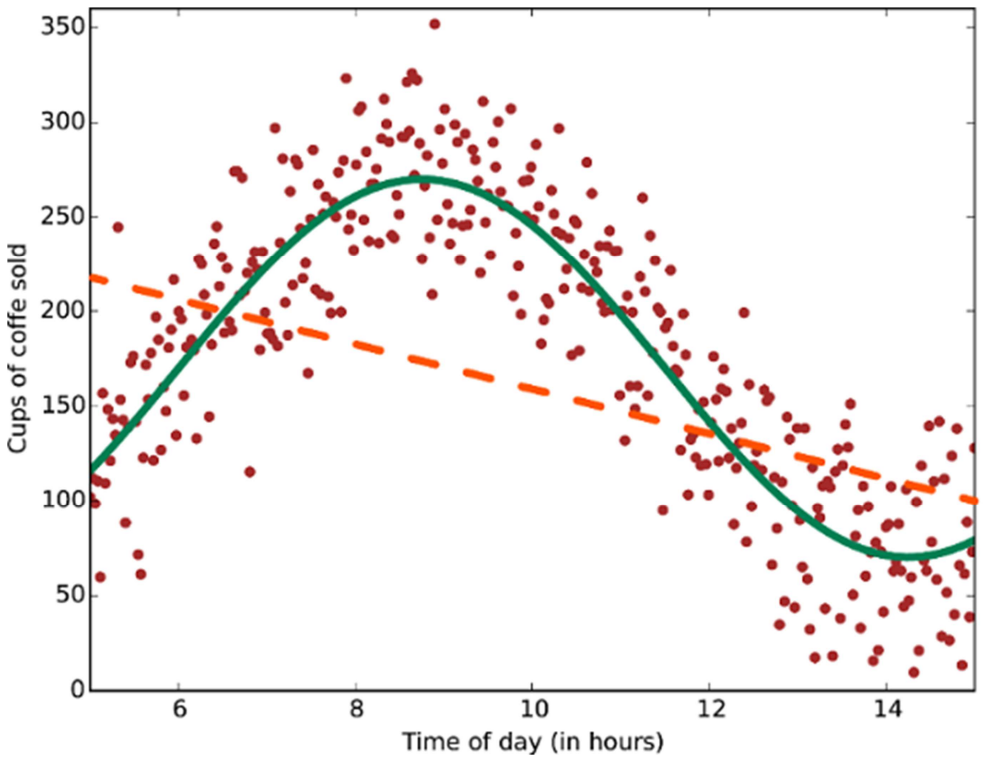
Machine Learning

- Curve fitting (correlations) - linear



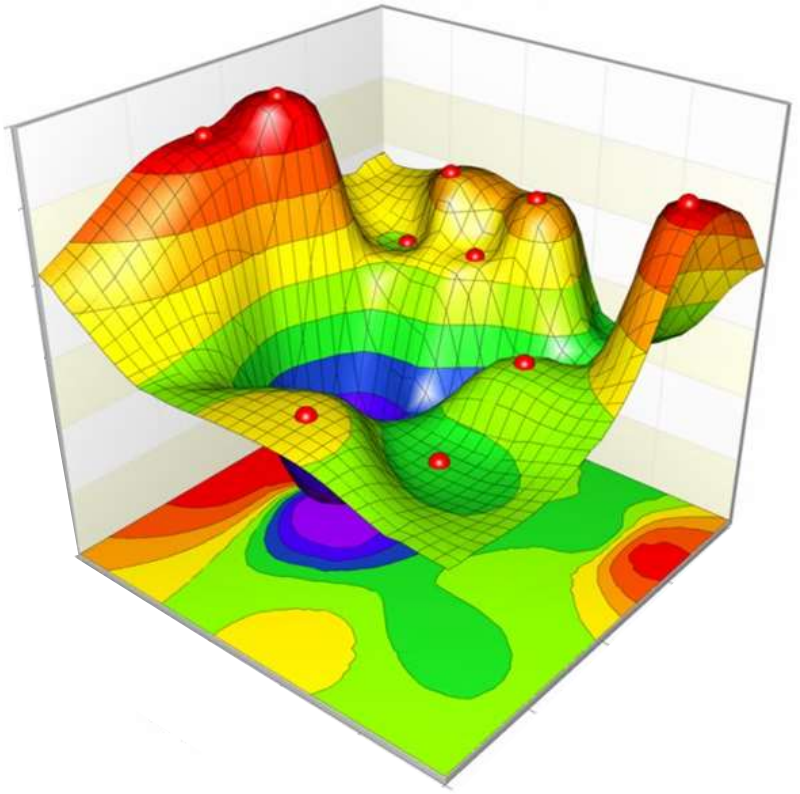
Machine Learning

- Curve fitting - nonlinear



Machine Learning

- Curve fitting - nonlinear



Machine Learning

- Deep Neural Networks

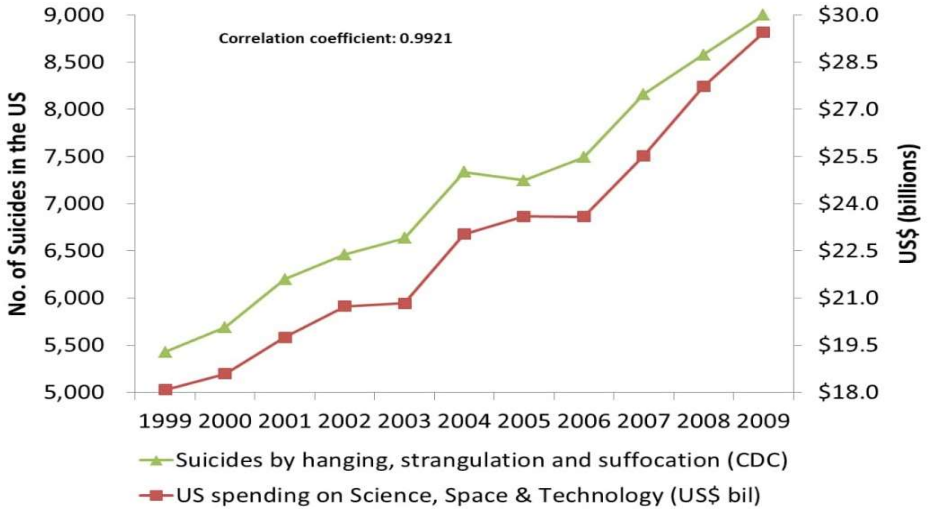
Highly dimensional, highly nonlinear

curve fitting

Machine Learning

- Spurious Correlations

Fitting can be highly misleading



Spurious Correlations

DOI:10.1145/3271625

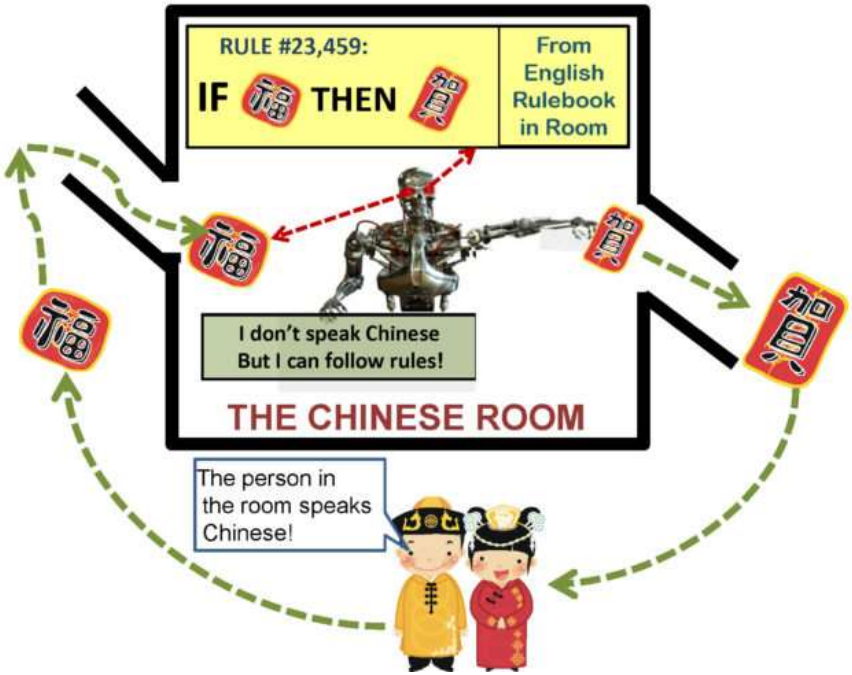
What just happened in artificial intelligence and how it is being misunderstood.

BY ADNAN DARWICHE

Human-Level Intelligence or Animal-Like Abilities?

"The vision systems of the eagle and the snake outperform everything that we can make in the laboratory, but snakes and eagles cannot build an eyeglass or a telescope or a microscope."

Fitting does not give us any understanding

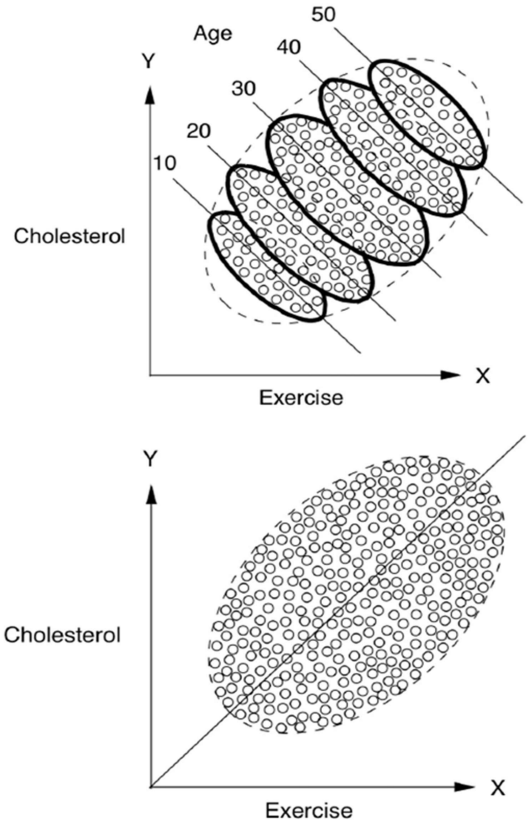


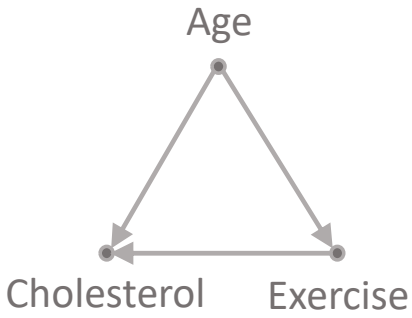
What can truly be achieved?

Does Exercise affects Cholesterol?

Simpson's Paradox

- No matter how much data you collect, the question can not be answered when using the data alone





DOI:10.1145/3271625

What just happened in artificial intelligence and how it is being misunderstood.

BY ADNAN DARWICHE

Human-Level Intelligence or Animal-Like Abilities?

"The vision systems of the eagle and the snake outperform everything that we can make in the laboratory, but snakes and eagles cannot build an eyeglass or a telescope or a microscope."

"The vision systems of the eagle and the snake outperform everything that we can make in the laboratory, but snakes and eagles cannot build an eyeglass or a telescope or a microscope."

— *Judea Pearl*

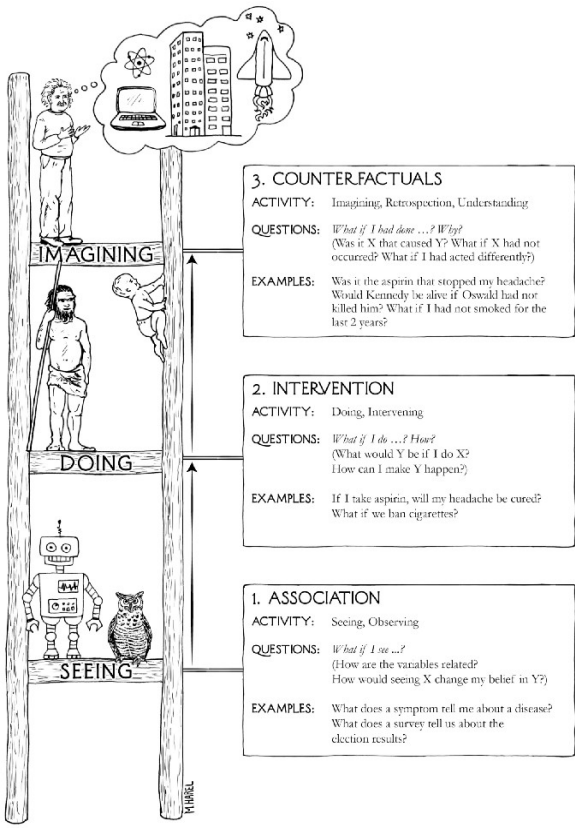
JUDEA PEARL

WINNER OF THE TURING AWARD

AND DANA MACKENZIE

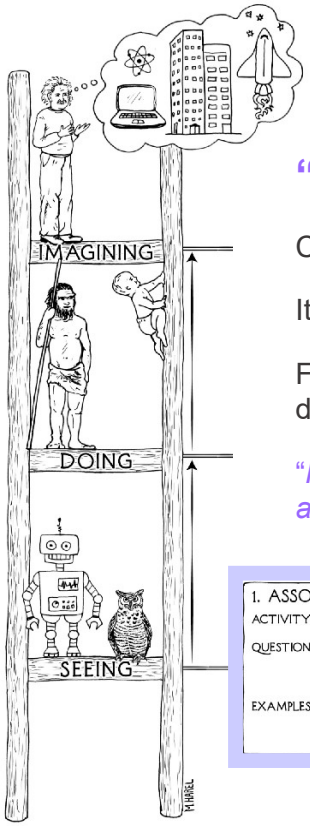
THE BOOK OF WHY

THE NEW SCIENCE OF CAUSE AND EFFECT



The Ladder of Causation

Seeing; we are looking for regularities in observations.



"What if I see …?"

Calls for predictions based on passive observations.

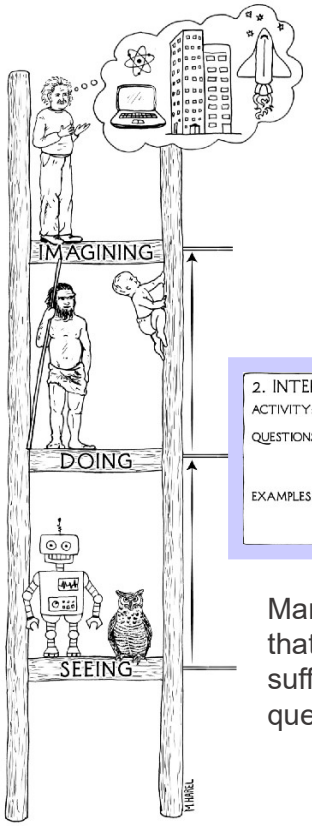
It is characterized by the question "*What if I see …?*"

For instance, imagine a marketing director at a department store who asks,

"*How likely is a customer who bought toothpaste to also buy dental floss?*"

1. ASSOCIATION	
ACTIVITY:	Seeing, Observing
QUESTIONS:	*What if I see ...?* (How are the variables related? How would seeing X change my belief in Y?)
EXAMPLES:	What does a symptom tell me about a disease? What does a survey tell us about the election results?

Intervention; ranks higher than association because it involves not just seeing but changing what is.



"What if do …?" & "How?"

We step up to the next level of causal queries when we begin to change the world. A typical question for this level is

"What will happen to our floss sales if we double the price of toothpaste?"

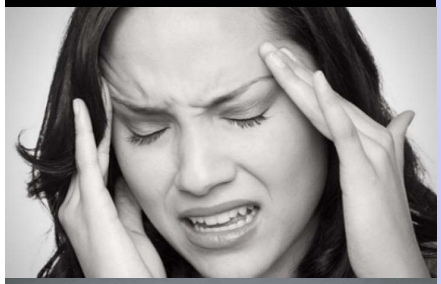
2. INTERVENTION	
ACTIVITY:	Doing, Intervening
QUESTIONS:	*What if I do …? How?* (What would Y be if I do X? How can I make Y happen?)
EXAMPLES:	If I take aspirin, will my headache be cured? What if we ban cigarettes?

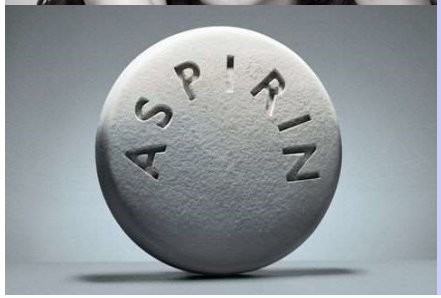
This already calls for a new kind of knowledge, absent from the data, which we find at rung two of the Ladder of Causation, **Intervention**.

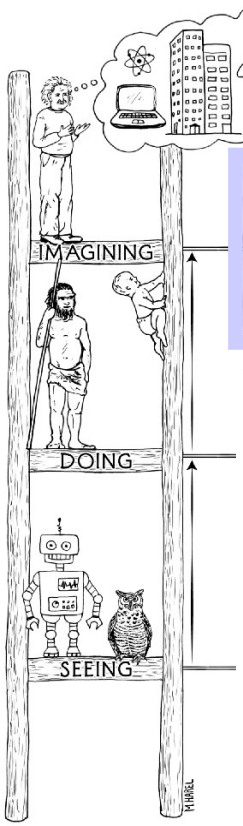
Many scientists have been quite traumatized to learn that none of the methods they learned in statistics is sufficient even to articulate, let alone answer, a simple question like

"What happens if we double the price?"

Counterfactuals; ranks higher than intervention because it involves **imagining**, **retrospection** and **understanding**.







"What if I had done …?" & "Why?"

3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?* (Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

We might wonder, My headache is gone now, but

- *Why?*
- *Was it the aspirin I took?*
- *The food I ate?*
- *The good news I heard?*

These queries take us to the top rung of the Ladder of Causation, the level of **Counterfactuals**, because to answer them we must go back in time, change history, and ask,

"What would have happened if I had not taken the aspirin?"

No experiment in the world can deny treatment to an already treated person and compare the two outcomes, so we must import a whole new kind of knowledge.

CONTENTS

- **DATA TYPES**; to list different types of data and to learn hw they must be used for unsupervised learning.
- **DATA PREPROCESSING**; to preprocess data in such a way it can be used by unsupervised learning tasks,
- **CLUSTERING LEARNING**; to form homogeneous groups of observations and/or attributes using a given proximity measure,
- **CLUSTERING VALIDATION**; to evaluate and compare different clustering solutions to select the one to deploy.
- **ANOMALY DETECTION**; to find anomalous observations, to discover outliers observations, under different theoretical settings.
- **BAYESIAN NETWORKS**; to learn probabilistic/causal structure from data and to make decisions under uncertainty.

ASSESSMENT – ONGOING EXAM (MAX 33)

- **Lab Reports** 11pts
- **Project** 12pts
- **Interview** 10pts
- Students are allowed/suggested to work in pairs during the lab activity
- Students are allowed/encouraged to work in pair to design, develop and document their project.

ASSESSMENT – FULL EXAM (MAX 30)

- **Project** 20pts
- **Interview** 10pts

PROJECT

- Students are allowed/suggested to work in pairs to design, develop and document their project
- A common project or a list of projects among which to choose will be given after the first half of the course (about mid of April)
- At the end of the project a technical report must be delivered (8 pages, a template will be given)
- The report will be graded analytically (e.g., methodological correctness, clarity of exposition, etc.). More details will be given together with the project.

INTERVIEW

- 20 minutes in total
- Of which 10/15 min for the presentation of the project
- The remaining 5/10 min for the interview on methodological aspects

LABORATORY

- Check slides from Giulia Cisotto PhD

LECTURE'S CALENDAR

Day	Time	Topic	where?
01-03-2023	14:30 - 16:30	Introduction	U1-04
06-03-2023	14:30 - 16:30	Types of Data	U3-09
15-03-2023	14:30 - 16:30	Proximity Measures	U1-04
20-03-2023	14:30 - 16:30	Introduction to Cluster Analysis	U3-09
22-03-2023	14:30 - 16:30	Cluster Analysis: K-means Clustering	U1-04
29-03-2023	14:30 - 16:30	Cluster Analysis: Hierarchical Clustering	U1-04
03-04-2023	14:30 - 16:30	Cluster Analysis: Density-based Clustering	U3-09
05-04-2023	14:30 - 16:30	Cluster Analysis: Clustering Validation	U1-04
12-04-2023	14:30 - 16:30	Introduction to Anomaly Detection	U1-04
19-04-2023	14:30 - 16:30	Anomaly Detection: Nearest-neighbor based	U1-04
26-04-2023	14:30 - 16:30	Anomaly Detection: Clustering Based, Statistical Approaches and Reconstruction Based	U1-04
03-05-2023	14:30 - 16:30	Anomaly Detection: Additional Algorithms	U1-04
10-05-2023	14:30 - 16:30	Introduction to Bayesian Networks	U1-04
17-05-2023	14:30 - 16:30	Bayesian Networks: Inference	U1-04
24-05-2023	14:30 - 16:30	Bayesian Networks: learning	U1-04
31-05-2023	14:30 - 16:30	Bayesian Networks and Healthcare	U1-04

ASKING QUESTIONS

- **Course issues**, i.e., lectures, lab, exams, general, …; write a post using the Moodle platform and discuss it with your peers and teachers at the same time
  - Lectures topics — Forum (Lecture)
  - Lab issues — Forum (Lab)
  - General issues — Forum (General)

- **Specific issues**, drop an email to fabio.stella@unimib.it and/or to giulia.cisotto@unimib.it

WE START BY ASKING YOU TO ANSWER FEW QUESTIONS