



**SARDIGNA CHIRCAS
SARDEGNA RICERCHE**

open:campus

Artificial Intelligence for developers

8 weekend per diventare Machine Learning Specialist



Natural Language Processing

Maurizio Atzori

Università degli Studi di Cagliari

atzori@unica.it

February 9-10 , 2024



Intelligent Data Exploitation

Dipartimento di Matematica ed Informatica

https://web.unica.it/unica/en/intelligent_data.page

Faculty:

Maurizio Atzori - atzori@unica.it

Cecilia Di Ruberto

Giovanni Puglisi

Barbara Pes

Andrea Loddo

Manuela Sanguinetti

+ *Research Fellows / PhD / students*

Topics of the Lab:

- **Artificial Intelligence and Natural Language Processing**
- Knowledge Graphs and Semantic Web
- Biomedical Image Analysis, Precision Agriculture, Image Retrieval
- Computer Vision, Multimedia Forensics
- Data Mining and Machine Learning
- High-dimensional Data Analysis and Feature Selection

Featured on:



Google
Faculty Research Awards

Wikimedia

Research Newsletter

NewScientist



COMMUNICATIONS
OF THE
ACM

ARTIFICIAL INTELLIGENCE AND KNOWLEDGE ENGINEERING (IEEE AIKE 2019)

June 3-5, 2019 - Cagliari, Sardinia, Italy

CALL FOR PAPERS

KEYNOTE SPEAKERS

IEEE Artificial Intelligence & Knowledge Engineering 2019 (IEEE AIKE 2019)

<https://aike2019.unica.it/>

3-5 Giugno 2019 all'Hotel Regina Margherita

SEBD 2024 Villasimius

CALLS

ATTENDING

PROGRAM



Georg Gottlob

FRS, Professor at University of Calabria, Italy



Themis Palpanas

Distinguished Professor at University Paris Cite, France

SEBD 2024 32nd SYMPOSIUM ON ADVANCED DATABASE SYSTEMS

June 23-26, 2024 - Villasimius, Sardinia, Italy

CONFERENCE PROGRAM

SUBMIT NOW!

KEYNOTE SPEAKERS

Sistemi Evoluti per Basi di Dati (SEBD 2024)

<https://sebd2024.unica.it/>

June 23-26 at Tanka Resort Villasimius

Outline of the course

- **Intro on AI, ML and NLP**
- **Text Processing**
- **Words and Corpora**
- **Lexical similarity**
- **Language Modeling**
- **Text Classification**
- **Semantic similarity**
- **Knowledge Graphs**
- **Intro to Large Language Models**

Teaching Material: Books

➡ **Speech and Language Processing by Dan Jurafsky and James H. Martin**

- **Free** online version (draft Feb 2024): <https://web.stanford.edu/~jurafsky/slp3/>
- Introduction to Natural Language Processing, The MIT Press 2019 by Jacob Eisentein (Georgia Tech, now Google AI)
 - **Free** online version (2018): <https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf>
- Goldberg: A Primer on Neural Network Models for Natural Language Processing
 - **Free** at <http://u.cs.biu.ac.il/~yogo/nnlp.pdf>
- Artificial Intelligence: Foundations of Computational Agents, second edition, Cambridge University Press 2017 by David L. Poole and Alan K. Mackworth (University of British Columbia, Vancouver)
 - **Free** online: <https://artint.info/2e/html/ArtInt2e.html>

Teaching Material: other resources

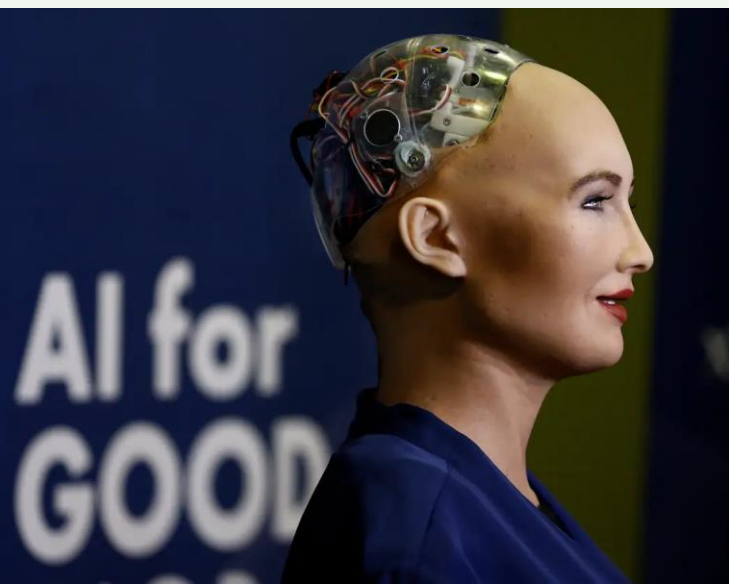
- Free Python libraries:
 - Spacy <https://spacy.io/>
 - NLTK <http://www.nltk.org/book/>
 - Sklearn (Scikit-learn), Numpy, Scipy, Pandas
 - PyTorch <https://pytorch.org/>
 - Textblob
 - Transformers (huggingface)



What are they?

Intro on AI, ML and NLP

1



Top Chess players Feb 2024

Sort: Rating ▾

1 2 3 4 5 > Last



GM Magnus Carlsen

2830 | #1

Norway

To many, GM Magnus Carlsen is the best to ever play the game, although GMs Garry Kasparov and Bobby Fischer remain in the conversation. At any rate, the clear and remarkable point is that before turning 30 years old, Carlsen had already earned a spot at the top, and he has continued to dominate into...



GM Fabiano Caruana

2804 | #2

United States

The prodigy who broke the record held by GM Hikaru Nakamura for America's youngest chess grandmaster, Fabiano Caruana, has climbed the mountain of chess and reached the summit. This grandmaster has had a ranking as high as number-two in the world and has won numerous tournaments in his career. In 2018...



GM Hikaru Nakamura

2788 | #3

United States

Hikaru Nakamura was born December 9, 1987 in Hirakata, Japan. His family moved to the United States when he was just two years old, and the Stars and Stripes are the only national banner he has known as a chess player. Nakamura has been one of the world's top players for well over a decade...



GM Anish Giri

2762 | #4

Netherlands

GM Anish Giri is a four-time Dutch champion and a world-class player. He has been ranked as high as number three in the world and reached his peak rating of 2802 on February 15, 2015. Giri has represented the Netherlands at the Olympiads since 2010. He won the 2012 Reggio Emilia tournament and the 2017...



GM Ding Liren

2762 | #4

China

Ding Liren is the reigning FIDE World Chess Champion after defeating GM Ian Nepomniachtchi in the 2023 World Championship. Like most world champions, a combination of dominance and close calls have defined Ding's career leading up to winning the title. Ding won his first Chinese Chess Championship at...



GM Alireza Firouzja

2760 | #6

France

Alireza Firouzja is an Iranian-born grandmaster who now plays for France. He is a two-time world championship candidate and two-time Iranian champion. In late 2019 and early 2020, Firouzja electrified the chess world with his second place finish in the World Rapid Championship (one point behind World...



GM Ian Nepomniachtchi

2758 | #7

Russia

GM Ian Nepomniachtchi ("Nepo" for short) is a Russian super grandmaster who has

Remove Ads

Top Players

- Top Chess Players
- Live Chess Ratings
- Games Database

Q Search Top Players

Recent Players

- Maja Zielinska
- Federico Silva
- M P Yuridu Akesh Samarasinghe
- Maria Del Carmen Palacios Garcia
- Jesus Palacian Campodarye

Remove Ads

Top Chess players Feb 2024

Sort: Rating ▾

< 1 2 3 4 5 > Last



GM Magnus Carlsen

2830 | #1

Norway

To many, GM Magnus Carlsen is the best to ever play the game, although GMs Garry Kasparov and Bobby Fischer remain in the conversation. At any rate, the clear and remarkable point is that before turning 30 years old, Carlsen had already earned a spot at the top, and he has continued to dominate into...



GM Fabiano Caruana

2804 | #2

United States

The prodigy who broke the record held by GM Hikaru Nakamura for America's youngest chess grandmaster, Fabiano Caruana, has climbed the mountain of chess and reached the summit. This grandmaster has had a ranking as high as number-two in the world and has won numerous tournaments in his career. In 2018...



GM Hikaru Nakamura

2788 | #3

United States

Hikaru Nakamura was born December 9, 1987 in Hirakata, Japan. His family moved to the United States when he was just two years old, and the Stars and Stripes are the only national banner he has known as a chess player. Nakamura has been one of the world's top players for well over a decade...



GM Anish Giri

2762 | #4

Netherlands

GM Anish Giri is a four-time Dutch champion and a world-class player. He has been ranked as high as number three in the world and reached his peak rating of 2802 on February 15, 2015. Giri has represented the Netherlands at the Olympiads since 2010. He won the 2012 Reggio Emilia tournament and the 2017...



GM Ding Liren

2762 | #4

China

Ding Liren is the reigning FIDE World Chess Champion after defeating GM Ian Nepomniachtchi in the 2023 World Championship. Like most world champions, a combination of dominance and close calls have defined Ding's career leading up to winning the title. Ding won his first Chinese Chess Championship at...



GM Alireza Firouzja

2760 | #6

France

Alireza Firouzja is an Iranian-born grandmaster who now plays for France. He is a two-time world championship candidate and two-time Iranian champion. In late 2019 and early 2020, Firouzja electrified the chess world with his second place finish in the World Rapid Championship (one point behind World...



GM Ian Nepomniachtchi

2758 | #7

Russia

GM Ian Nepomniachtchi ("Nepo" for short) is a Russian super grandmaster who has

Remove Ads

Top Players

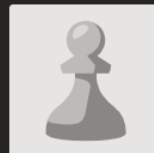
- Top Chess Players
- Live Chess Ratings
- Games Database

Search Top Players

Recent Players

- Maja Zielinska
- Federico Silva
- M P Yurid Alesh Samarasinghe
- Maria Del Carmen Palacios Garcia
- Jesus Palacian Campodarye

Remove Ads



IM Ilja Schneider

Germany

2460 | #993

GM Milos Jirovsky

Czech Republic

2459 | #1003


GM Ilya Khmelniker

Israel

2459 | #1003

Sort: Rating ▾


1 2 3 4 5 > Last



GM Magnus Carlsen 2830 | #1

Norway


To many, GM Magnus Carlsen is the best to ever play the game, although GMs Garry Kasparov and Bobby Fischer remain in the conversation. At any rate, the clear and remarkable point is that before turning 30 years old, Carlsen had already earned a spot at the top, and he has continued to dominate into...



GM Fabiano Caruana 2804 | #2

United States


The prodigy who broke the record held by GM Hikaru Nakamura for America's youngest chess grandmaster, Fabiano Caruana, has climbed the mountain of chess and reached the summit. This grandmaster has had a ranking as high as number-two in the world and has won numerous tournaments in his career. In 2018...



GM Hikaru Nakamura 2788 | #3

United States


Hikaru Nakamura was born December 9, 1987 in Hirakata, Japan. His family moved to the United States when he was just two years old, and the Stars and Stripes are the only national banner he has known as a chess player. Nakamura has been one of the world's top players for well over a decade...



GM Anish Giri 2762 | #4

Netherlands


GM Anish Giri is a four-time Dutch champion and a world-class player. He has been ranked as high as number three in the world and reached his peak rating of 2802 on February 15, 2015. Giri has represented the Netherlands at the Olympiads since 2010. He won the 2012 Reggio Emilia tournament and the 2017...



GM Ding Liren 2762 | #4

China


Ding Liren is the reigning FIDE World Chess Champion after defeating GM Ian Nepomniachtchi in the 2023 World Championship. Like most world champions, a combination of dominance and close calls have defined Ding's career leading up to winning the title. Ding won his first Chinese Chess Championship at...



GM Alireza Firouzja 2760 | #6

France

Alireza Firouzja is an Iranian-born grandmaster who now plays for France. He is a two-time world championship candidate and two-time Iranian champion. In late 2019 and early 2020, Firouzja electrified the chess world with his second place finish in the World Rapid Championship (one point behind World...



GM Ian Nepomniachtchi 2758 | #7

Russia

GM Ian Nepomniachtchi ("Nepo" for short) is a Russian super grandmaster who has

Remove Ads

Top Players

- Top Chess Players
- Live Chess Ratings
- Games Database

Search Top Players

Recent Players

- Maja Zielinska
- Federico Silva
- M P Yuridit Akesh Samarasinghe
- Maria Del Carmen Palacios Garcia
- Jesus Palacian Campodarve

Remove Ads

As of June 2023, Stockfish is the highest-rated engine according to the computer chess rating list (CCRL), with a rating of approximately 3530



Artificial Intelligence: some definitions

It is disputed

- The goal of artificial intelligence is to build software and robots with the same range of abilities as humans (Russell and Norvig, 2009)
- AI is whatever hasn't been done yet
- Intelligence is the computational part of the ability to achieve goals in some world
- If it can pass some tests (e.g. Turing Test)

AI: the beginning

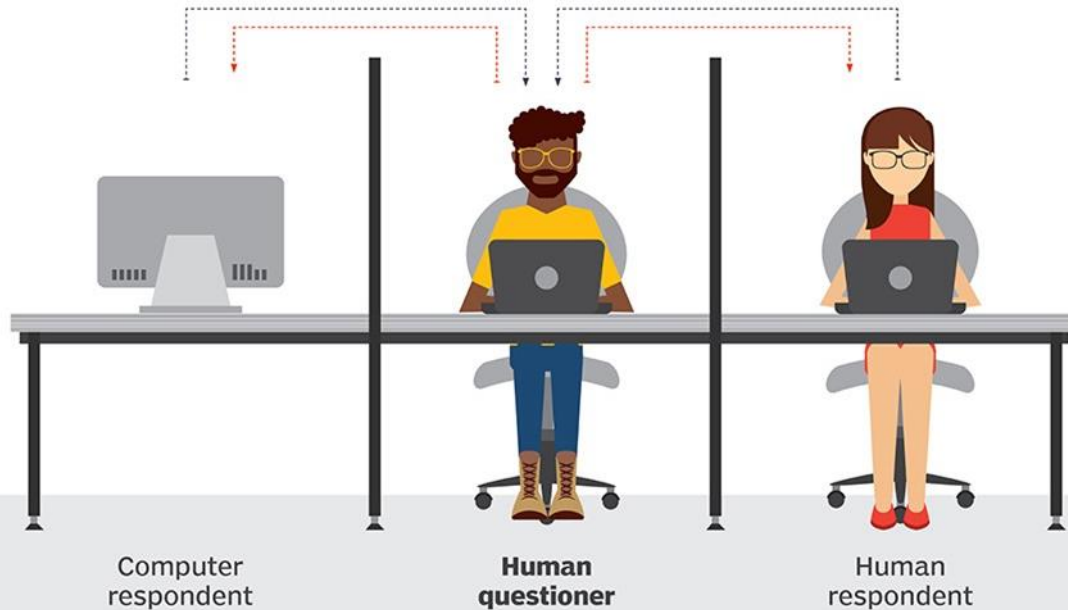
After WWII (1939-1945), a number of people independently started to work on intelligent machines. The English mathematician **Alan Turing** may have been the first. He gave a lecture on it in 1947. He also may have been the first to decide that AI was best researched by programming computers rather than by building machines. By the late 1950s, there were many researchers on AI, and most of them were basing their work on programming computers.

<https://www.kurzweilai.net/what-is-artificial-intelligence>

Turing Test (1950)

During the Turing test, the human questioner asks a series of questions to both respondents. After the specified time, the questioner tries to decide which terminal is operated by the human respondent and which terminal is operated by the computer.

■ QUESTION TO RESPONDENTS ■ ANSWERS TO QUESTIONER



Other tests for intelligence

- The **Marcus Test**- In which a program which can 'watch' a television show is tested by being asked meaningful questions about the show's content.
- **The Lovelace Test 2.0**- Which is a test made to detect AI through examining its ability to create artwork (<https://arxiv.org/abs/1410.6142>),
 - "Until a machine can originate an idea that it wasn't designed to, it can't be considered intelligent in the same way humans are." Ada Lovelace
 - e.g. "tell a story in which a boy falls in love with a girl, aliens abduct the boy, and the girl saves the world with the help of a talking cat."
- **Winograd Schema Challenge**- a test that asks multiple-choice questions in a specific format

Ada Lovelace

first computer programmer



Winograd schemas and Reasoning

The trophy doesn't fit into the brown suitcase because **it** is too [small/large].

Q: What is "it" referring to?

Solving this example requires spatial reasoning; other schemas require **reasoning** about actions and their effects, emotions and intentions, and social conventions.

Considerations on Tests for intelligence (1/2)

- Turing Test (as well as others) is a restrictive definition of AI, since it does not admit other **goal-oriented behavior** such as insect behavior as intelligence.

This is why there is now a distinction on different levels of AI:

Artificial General Intelligence (AGI, aka strong AI, full AI) is the intelligence of a machine that can understand or learn *any intellectual task* that a human being can (currently very far from solving this problem)

Artificial Intelligence or Applied AI (aka weak AI, narrow AI) the use of software to study or accomplish specific problem solving or reasoning tasks. It does not attempt to perform the full range of human cognitive abilities (e.g. Siri, self-driving cars, machine translation, ...)

Considerations on Tests for intelligence (2/2)

- These tests on intelligence all rely on a very characteristic ability of human beings: **natural language understanding**

Natural language understanding not only falls into AI topics, but it is also considered **AI-complete**, that is, solving it would solve any other AI task. In other words, it is the most difficult problem in AI

Eliza (1966)



"Psychotherapist" by Joseph Weizembaum

Nice online demo at <https://www.masswerk.at/eliza/>

<http://psych.fullerton.edu/mbirnbaum/psych101/eliza.htm>

Natural Language Processing (NLP)

Natural Language Processing is the set of methods for making human language accessible to computers... or, "whatever you do with NL (text/voice)"

Some examples:

- speech recognition,
- natural language understanding (NLU), and
- natural language generation (NLG).

ELIZA is an example of NLP application which mimic NLU

NLP: applications

Real examples of NLP applications:

- **automatic machine translation** is ubiquitous on the web/media
- **text classification** keeps our email inboxes free of spam
- **search engines** have moved beyond string matching, IR, and network analysis to a high degree of linguistic sophistication (with **knowledge graphs**)
- **chatbots/dialog systems** provide an increasingly common and effective way to get and share information
- **voice assistants** such as Alexa help to accomplish daily tasks at home

Circa 29.800.000 risultati (0,66 secondi)

NLP tasks

Da fonti sul Web



Named-entity recognition ▾



Sentiment analysis ▾



Speech recognition ▾



Stemming ▾



Keyword extraction ▾



Coreference resolution ▾



Customer service autom... ▾



Similarity ▾



Summarization ▾



Part-of-speech tagging ▾



Chatbots ▾



Text generation ▾



Language identification ▾



Search results ▾



Email filtering ▾



Dependency parsing ▾



Machine translation



Text classification



Question answering



Lexical analysis



Spam detection



Translation



Identifying stopwords



Sentence segmentation

Mostra meno ^

NLP-progress

Repository to track the progress in Natural Language Processing (NLP), including the datasets and the current state-of-the-art for the most common NLP tasks.

Tracking Progress in Natural Language Processing

Table of contents

English

- Automatic speech recognition
- CCG
- Common sense
- Constituency parsing
- Coreference resolution
- Data-to-Text Generation
- Dependency parsing
- Dialogue
- Domain adaptation
- Entity linking
- Grammatical error correction
- Information extraction
- Intent Detection and Slot Filling
- Language modeling
- Lexical normalization
- Machine translation
- Missing elements
- Multi-task learning
- Multi-modal
- Named entity recognition
- Natural language inference
- Part-of-speech tagging
- Paraphrase Generation
- Question answering
- Relation prediction
- Relationship extraction
- Semantic textual similarity
- Semantic parsing
- Semantic role labeling
- Sentiment analysis
- Shallow syntax
- Simplification
- Stance detection
- Summarization
- Taxonomy learning
- Temporal processing
- Text classification
- Word sense disambiguation

NLP: related areas

These diverse applications are based on a common set of ideas, drawing on algorithms, linguistics, logic, statistics.

Computational Linguistics: the object of the study is the language itself, and it's done by means of computational methods (e.g., as in computational astronomy)

Artificial Intelligence: the object of the study of reproducing intelligence of any form, by means of several techniques such as reasoning over knowledge bases

Machine Learning (ML): the object of the study is automatic learning, usually from examples. Currently most NLP approaches are based on ML, either supervised or unsupervised

Machine Learning: a brief introduction

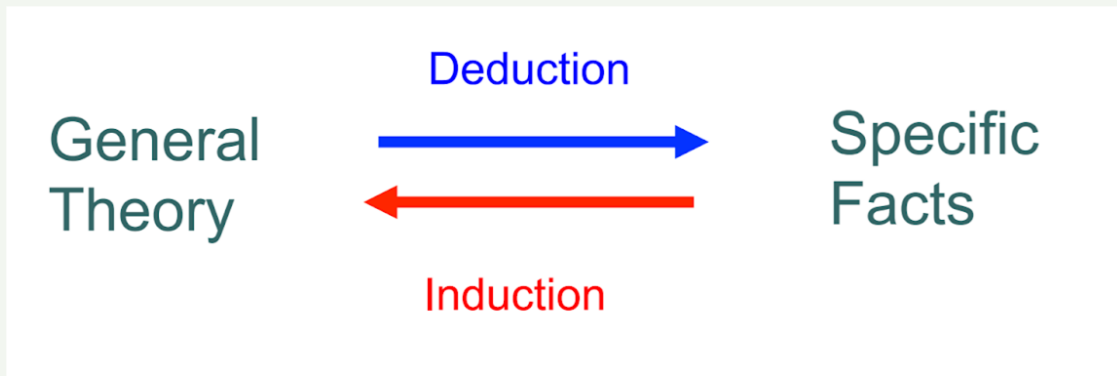
What is Machine Learning?

- Programs that get better with experience given a task and some performance Measure... some examples:
 - Learning to classify news articles
 - Learning to recognize spoken words
 - Learning to play board games
 - Learning to navigate a virtual world
 - Usually involves some sort of inductive reasoning step
- Usually involves some sort of **inductive reasoning** step

Next slides from <https://homepage.cs.uri.edu/faculty/hamel/courses/2015/spring2015/csc481/lecture-notes/> based on work by D. Poole and A. Mackworth (see their book)

Reasoning

- Deductive reasoning (rule based reasoning)
 - From the general to the specific (All Cretans are liars, Epimenides is a cretan)
- Inductive reasoning
 - From the specific to the general (Epimenides is a liar, Kresilas is a liar, Nearchus is a liar)



Arrows represent inference. **Inference** is the act or process of drawing a conclusion based solely on what one already knows.

Note: very different from Mathematical Induction!

Inductive Reasoning: example

- Facts: every time you see a swan you notice that the swan is white.
- Inductive step: you infer that all swans are white.

Observed Swans
are white.



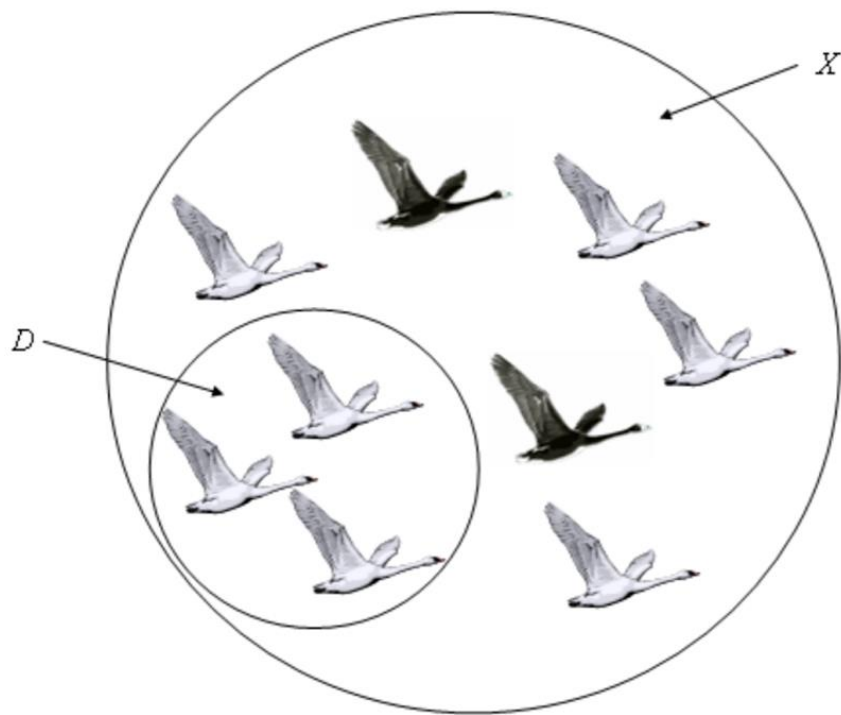
Induction

All Swans
are white.



Reasoning

- Deduction is “truth preserving”
 - If the rules employed in the deductive reasoning process are sound, then, what holds in the theory will hold for the deduced facts
 - There exist also probabilistic deductive reasoning frameworks
- Induction is NOT “truth preserving”
 - It is more of a statistical argument
 - The more swans you see that are white, the more probable it is that all swans are white. But this does not exclude the existence of black swans



$D \equiv$ observations

$X \equiv$ universe of all swans

Different styles of ML: supervision

- Supervised Learning

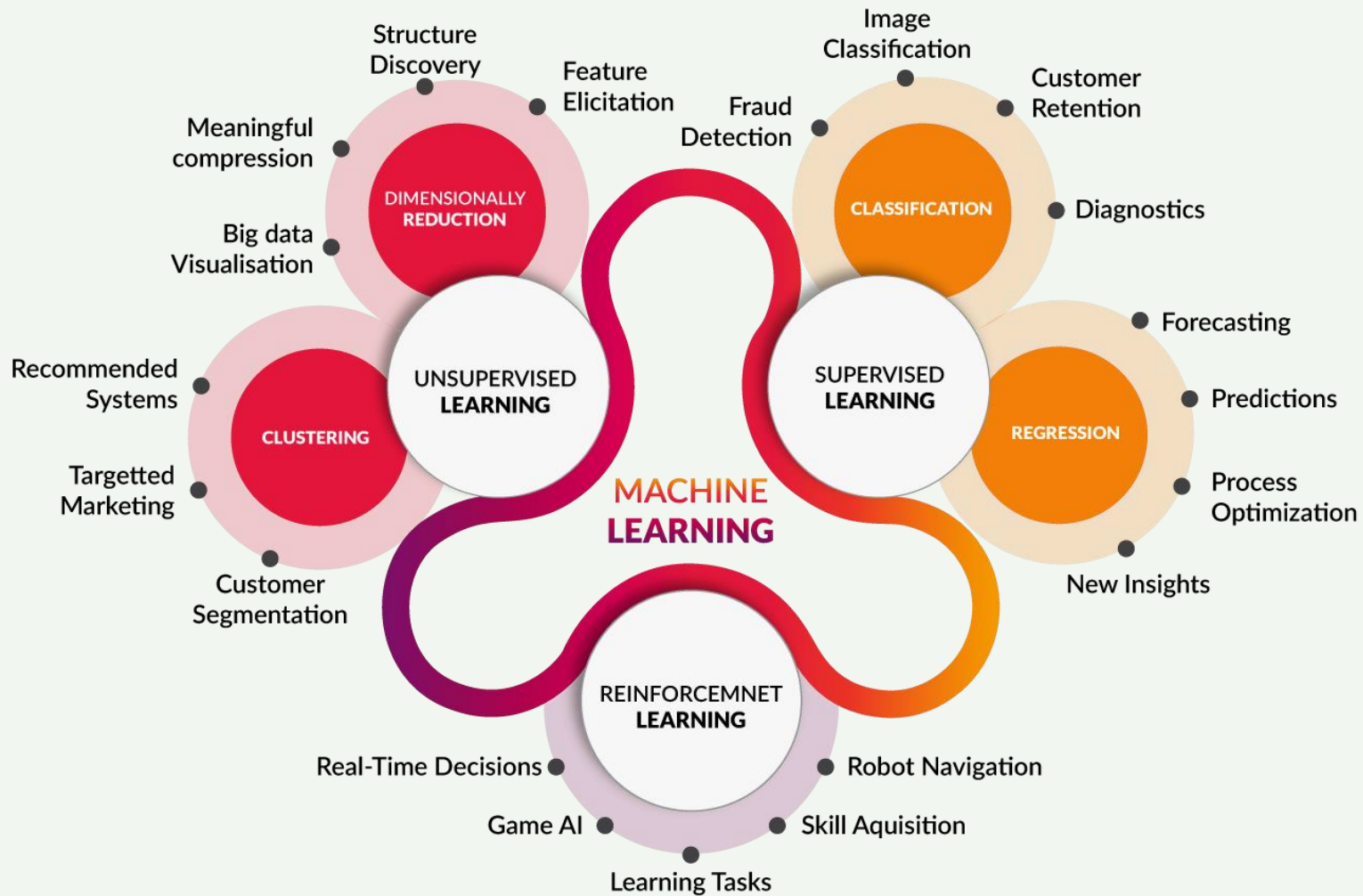
- The learning needs explicit examples of the concept to be learned (e.g. white swans...)

- Unsupervised Learning

- The learner discovers autonomously any structure in the domain that might represent an interesting concept

- Reinforcement Learning

- The learner gets feedbacks from his own past behaviour and learn from it. E.g. it plays against itself, improving time over time



Different styles of ML: knowledge representation

- Symbolic Learners (transparent models), e.g.:
 - If-then-else rules
 - Decision trees
 - Association rules
- Sub-Symbolic Learners (non-transparent models), e.g.:
 - Neural Networks
 - Clustering (Self-Organizing Maps, k-Means)
 - Support Vector Machines



What are they?

Intro on AI, ML and NLP



Performance Evaluation

Intro on AI, ML and NLP

Evaluation

Let's consider just binary text classification tasks

Imagine you're the CEO of Delicious Pie
Company

You want to know what people are saying about
your pies

So you build a "Delicious Pie" tweet detector

- Positive class: tweets about Delicious Pie Co
- Negative class: all other tweets

The 2-by-2 confusion matrix

gold standard labels

		gold positive	gold negative	
<i>system output labels</i>	system positive	true positive	false positive	precision = $\frac{tp}{tp+fp}$
	system negative	false negative	true negative	
		recall = $\frac{tp}{tp+fn}$		accuracy = $\frac{tp+tn}{tp+fp+tn+fn}$

Evaluation: Accuracy

Why don't we use **accuracy** as our metric?

Imagine we saw 1 million tweets

- 100 of them talked about Delicious Pie Co.
- 999,900 talked about something else

We could build a dumb classifier that just labels every tweet "not about pie"

- It would get 99.99% accuracy!!! Wow!!!!
- But useless! Doesn't return the comments we are looking for!
- That's why we use **precision** and **recall** instead

Evaluation: Precision

% of items the system detected (i.e., items the system labeled as positive) that are in fact positive (according to the human gold labels)

$$\textbf{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

Evaluation: Recall

% of items actually present in the input that were correctly identified by the system.

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Why Precision and recall

Our dumb pie-classifier

- Just label nothing as "about pie"

Accuracy=99.99%

but

Recall = 0

- (it doesn't get any of the 100 Pie tweets)

Precision and recall, unlike accuracy, emphasize true positives:

- finding the things that we are supposed to be looking for.

A combined measure: F

F measure: a single number that combines P and R:

$$F_{\beta} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

We almost always use balanced F_1 (i.e., $\beta = 1$)

$$F_1 = \frac{2PR}{P + R}$$

Development Test Sets ("Devsets") and Cross-validation

Training set

Development Test Set

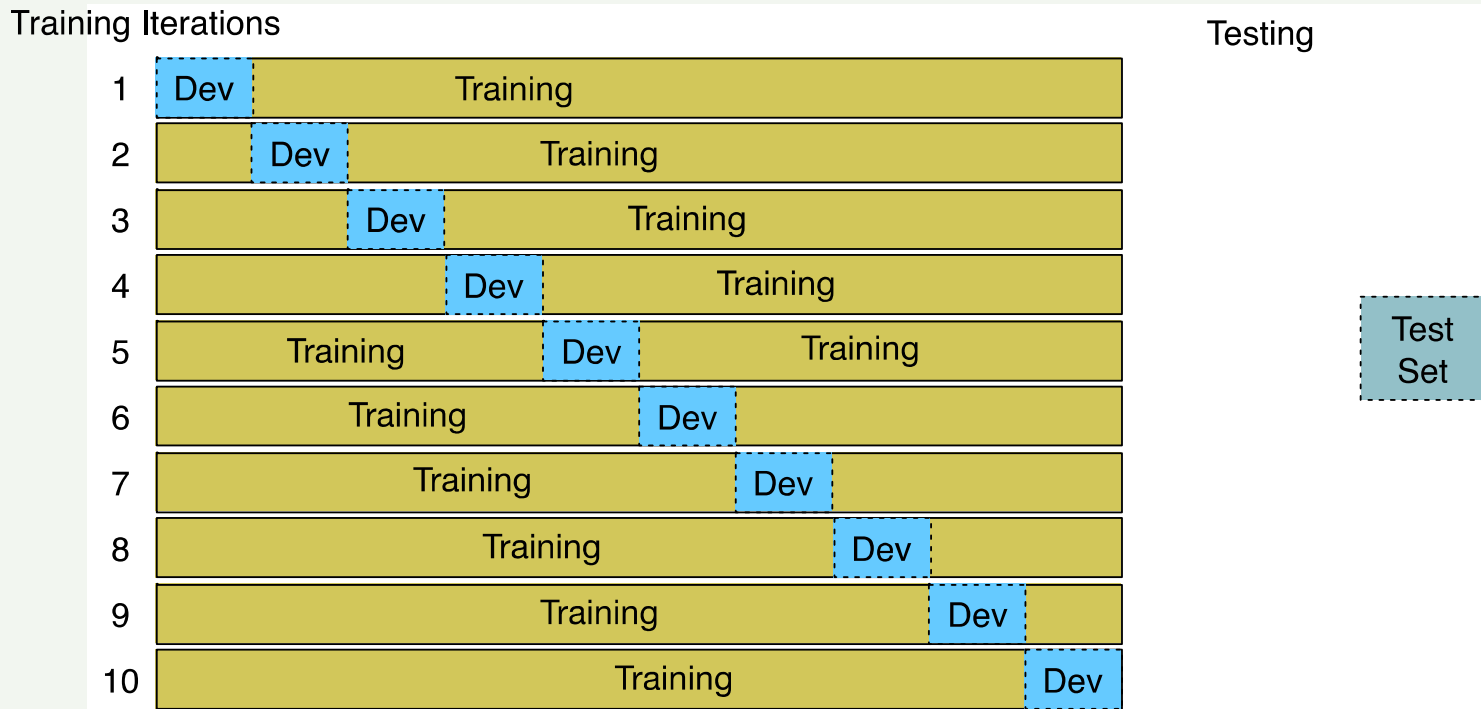
Test Set

Train on training set, tune on devset, report on testset

- This avoids overfitting ('tuning to the test set')
- More conservative estimate of performance
- But paradox: want as much data as possible for training, and as much for dev; how to split?

Cross-validation: multiple splits

Pool results over splits, Compute pooled dev performance



Exercise

Precision: percentage of correct results among those found

Recall: percentage of correct results among all that are known to be correct

On Google Colaboratory



Performance Evaluation

Intro on AI, ML and NLP