High Performance Artificial Intelligence

Claude TADONKI
MINES-PARISTECH / PSL

VI Escola Regional de Alto Desempenho do Rio de Janeiro (ERAD-RJ 2020)
November 30th – December 4th 2020

Universidade Federal Rural do Rio de Janeiro (UFRRJ) - BRAZIL

A Historical Case

Deep Blue versus Garry **Kasparov** (1997)

Deep Blue

- **Victory** of **Deep Blue** (IBM Supercomputer) over **Kasparov** (Human)
- Deep Blue had a database of the most important chess games of the 20th century
- Deep Blue was able to analyze 200 millions of moves per second
- Deep Blue was a **11.4 GFlops** machine, the current world fastest machine is **537 TFlops**

X 47×10^3

TOP500 List (November 2020)

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442,010.0	537,212.0	29,899
2	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
3	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
4	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
5	**Selene** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia NVIDIA Corporation United States	555,520	63,460.0	79,215.0	2,646

AlphaGo
Google DeepMind

- With **Fugaku**, we could imagine **10^4 billions moves per second** (scaling from **Deep Blue**)
- With current computing powers, **AI-based applications** are expected to be **highly efficient**

Real-Life Compromise

- Our actions should be **smarter** with **more time** to decision
- Our **time to decision** is always bounded and should be **shorter enough** to be useful

(Sport; Game; Work; Investments; Driving; ...)

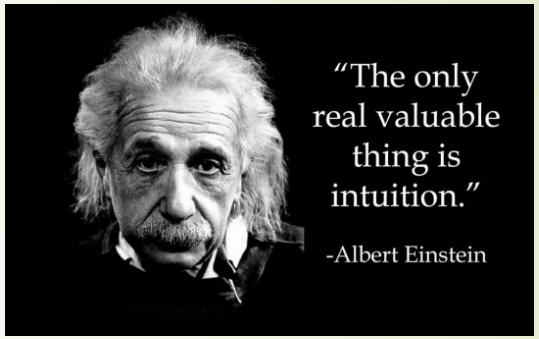
If you think too much before taking a step, then
you might spend the rest of your life with one leg in the air.

Chinese Proverb.

AI and Human Intelligence

- The fact that **AI can defeat or outperform a human** does **not mean "*smarter*"**
- **AI** is implemented through computers, thus it runs **deterministic algorithms**

HUMAN	AI
Brain (memory)	Data
Brain (connections)	Machine
Brain (memory + connections)	Algorithm
Intuition	-
Random	Pseudo-random
Emotion	-



- **AI** needs know-how (by design or through learning), while **Human** might invent (originality)
- With the increasingly powerful HPC support to **AI,** the **Turing Test** might become harder
- **AI** is made and driven by humans, could we thus imagine it going beyond ?

High Performance AI Applications

ECONOMY/FINANCE/BANKING

- Customer Service (build deep and personal relationships with customers)
- Security and Fraud Detection (detect fraudulent activities seen as abnormal behaviours)
- Mobile Banking (with AI-based only interaction, online banking can offer a round-the-clock service
- Algorithmic Trading and Risk Management (large-scale prediction and decision making)

(outcome, probability)
(outcome, probability)
⋮
(outcome, probability)

DECISION

- Chatbots and Other Bots (ubiquity while keeping close to human touch)

- 52% of financial services industry are investing in AI
- 72% of business decision makers believe that AI will be the business advantage of the future

According to research conducted by Autonomous Next
« **the aggregate potential cost savings for banks from AI applications is estimated at $447 billion by 2023** »
https://www.ciol.com/artificial-intelligence-every-bank-needs/

High Performance AI Applications

MONITORING

Sensing → Identification + Algorithm → Decision / Action

- Autonomous Driving (get the driving process managed by IA)

- Autonomous Surveillance (get the surveillance process managed by IA)

Autonomous Security Vehicle

Autonomous Surveillance Robot

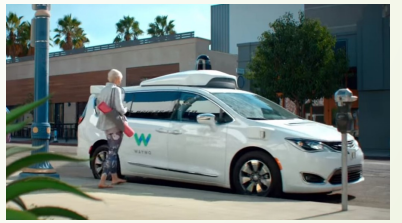
Autonomous Security Boat

- Autonomous Transportation (get the process process managed by IA)

Autonomous Aircraft (Embraer)

Autonomous Shuttle

Autonomous Bus



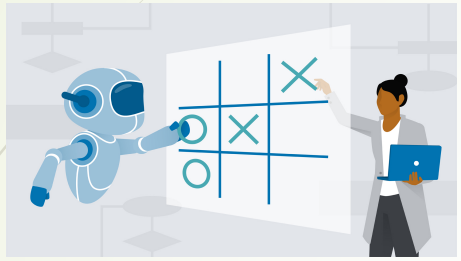
Autonomous Taxi

Autonomous Ambulance

High Performance AI Applications

GAMING

- **AI** is now pervasive in gaming (as a **full machine player** or as a **human player assistant**)

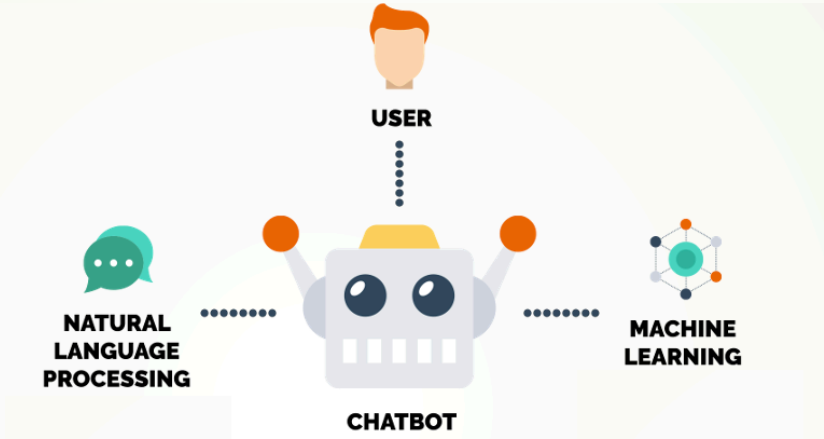



- Gaming is getting **more and more realistic** (video games are getting smarter and more creative)
- With high-precision design, **AI-based gaming** can even be used for **general purpose assistance** (disabled people, specialized education, patients daily assistance, …)
- AI-based games can **adapt from the player behaviour and records** (the interaction thus becomes incremental and more consistent)
- Coupled with virtual or augmented reality, AI-based games **close the gap** between **pure fiction** and **reality** (the gamer might feel that he is having a real-life experience)

High Performance AI Applications

INTERACTION & SERVICES

- Customer Service/support with AI Chatbots (expected to be **real-time** and **realistic**)



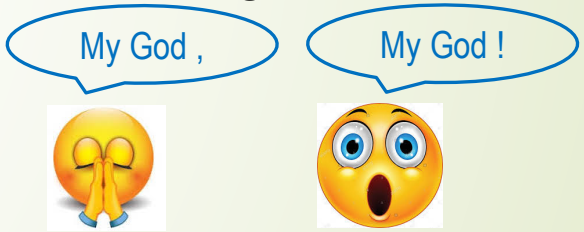

Domestic robot


Shopbot

- Emotional AI (**emotion** is an important user input that needs to be **identified** and **taken into account**)

This emotion can be sensed through

- Facial expression
- Voice intonation
- Language characteristic
- Specific behavior



- Customized Elements (answers/suggestions/adverts/…. The user feels understood and well guided)

AI challenges and the Need for HPC

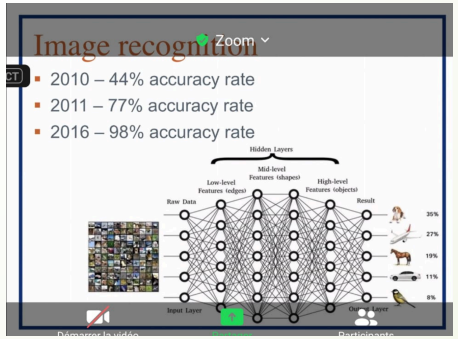
- The quality of **AI** approaches goes with complex algorithms
- Getting good **AI** results might require considering lot of data
- The conjunction of complex algorithms and lot of data ➡ **heavy computing workload**
- A **good AI** should be real-time
- **Large-scale machine learning** should be robust and efficient in order to scale AI

- ❖ Lot of (various) data
- ❖ Data-sensitive (even numerically)
- ❖ Complex evaluation procedure
- ❖ Repetitive learning process

Neurodegenerative diseases identified using HPC Artificial Intelligence
Mount Sinai Hospital – SC19 Award

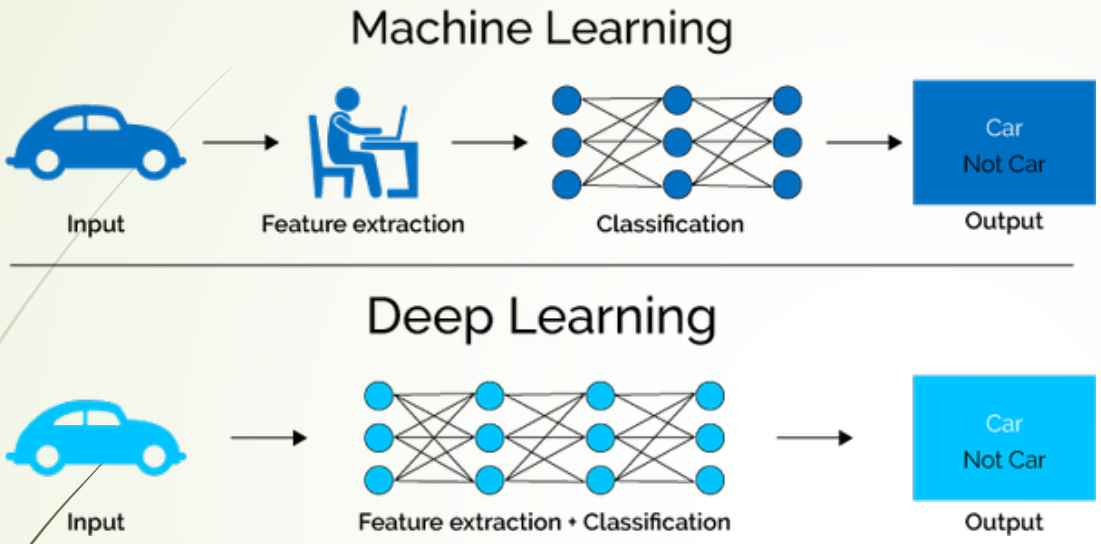
Fundamental HPC-AI Questions

- With the HPC advances, should we
  - consider **more powerful methods** (likely to yield better quality solution by design) or
  - **scale-up the scenario** of those already considered (more data, more training, …)
- For ordinary AI applications, how to deal with large-scale HPC infrastructures ?
  (for embedded solutions, remote computation might be the better way to go)
- Under the **influence of HPC**, should we pursue the **human brain target** ?
- How does the (new) practical horizon of AI looks like with HPC advances consideration ?
- What are the specificities of AI applications w.r.t scalable **HPC** ?
- What about the collateral damage of HPC issues on **high-performance AI ?**

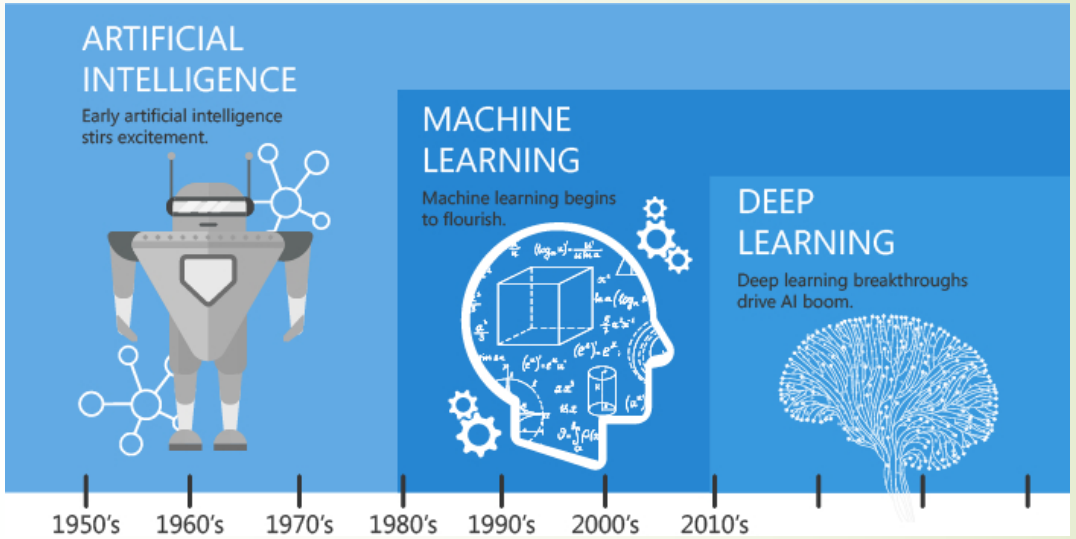


AI Methods and HPC

- The most popular AI approach is **Machine Learning**, which has led to **Deep Learning**



https://blog.dataiku.com/ai-vs.-machine-learning-vs.-deep-learning



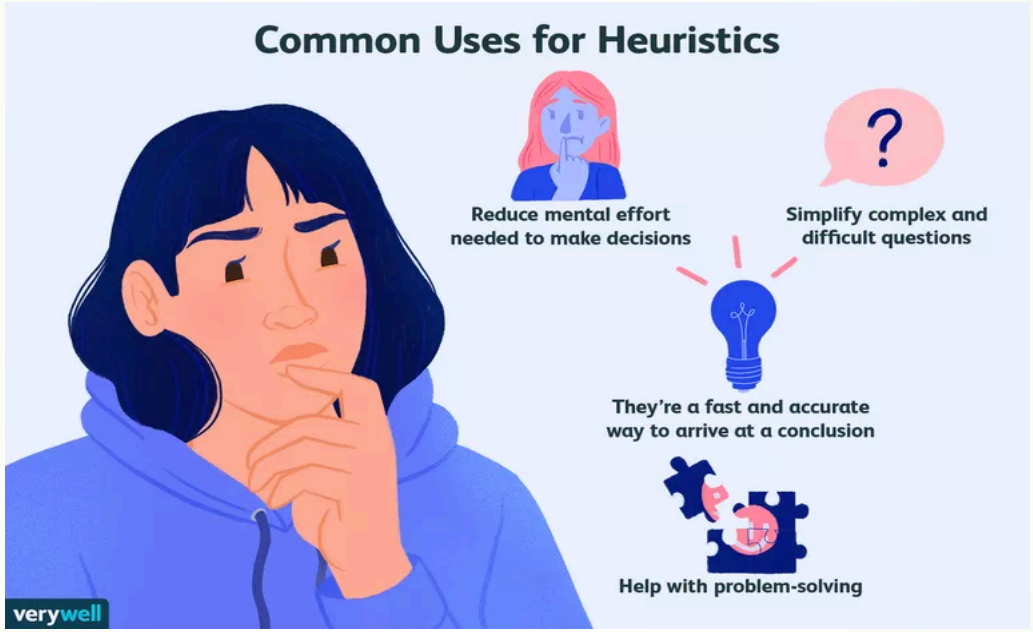
https://hackernoon.com/difference-between-artificial-intelligence-machine-learning-and-deep-learning-1pcv3zeg

- Since AI leads to a decision process, it uses (complex) **Operational Research algorithms**
- **HPC impact on AI** algorithms will mainly come from the **impact on OR advances**
- **HPC devices** that are tailored **for Deep Learning** are being considered
- **HPC libraries for AI** is a valuable software step
- **AI** will also impact **HPC** techniques (compilation, deployment, scheduling, …)

AI Methods and HPC

- **Heuristic**

  We are not always so dependent on having an **exact or optimal solution** and the **time to get one might be so long** that it won't be worth considering it anymore.



Verywell / Cindy Chung

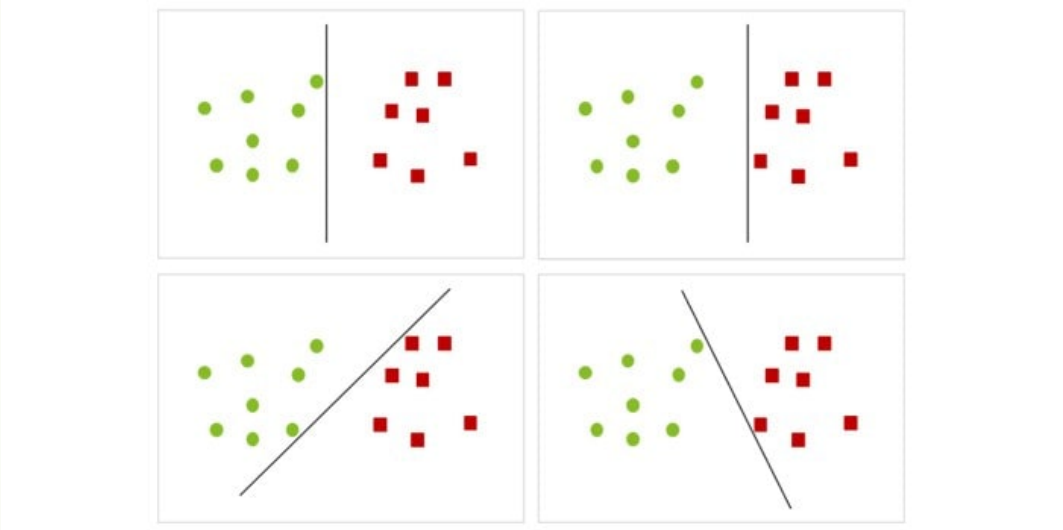
- We have to decide quickly (HPC itself might not be sufficient!)
- The path to the right decision my tolerate some deviation/simplification/omission
- Exact/optimal algorithms might be unscalable (thus inefficient with large-scale HPC)
- HPC implementations of heuristic algorithms need to be scalable enough

AI Methods and HPC

- **Support Vector Machines**

  Most of real-life decisions are based on a **data-oriented classification** that is expected to be **simple enough** so has to yield a **fast identification procedure**.

  Examples: A given email is a spam or not ? A given bank transaction is suspicious or not ?



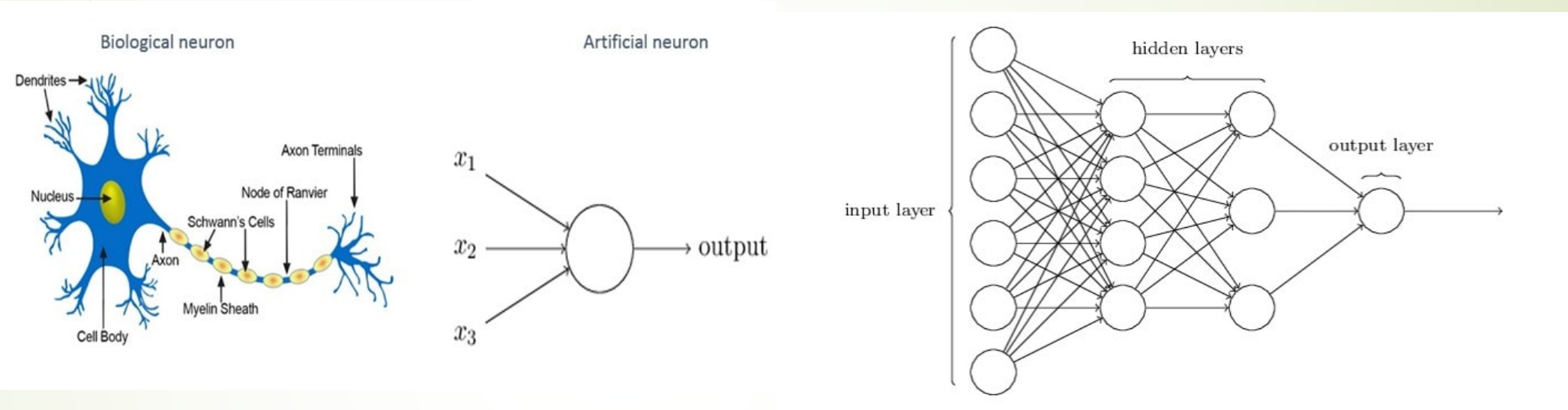
https://www2.deloitte.com/nl/nl/pages/data-analytics/articles/part-2-artificial-intelligence-techniques-explained.html

- Data intensive (might be highly multi-dimensional)
- Numerically sensitive (robust numerical method might be considered)
- A good quality separator might be more complex than desired (thus a HPC challenge)

AI Methods and HPC

Artificial Neural Networks

Artificial Neural Networks (ANN) is a **major paradigm used in AI**. ANN has a few neurons while human brain has hundred billions.



https://www2.deloitte.com/nl/nl/pages/data-analytics/articles/part-2-artificial-intelligence-techniques-explained.html

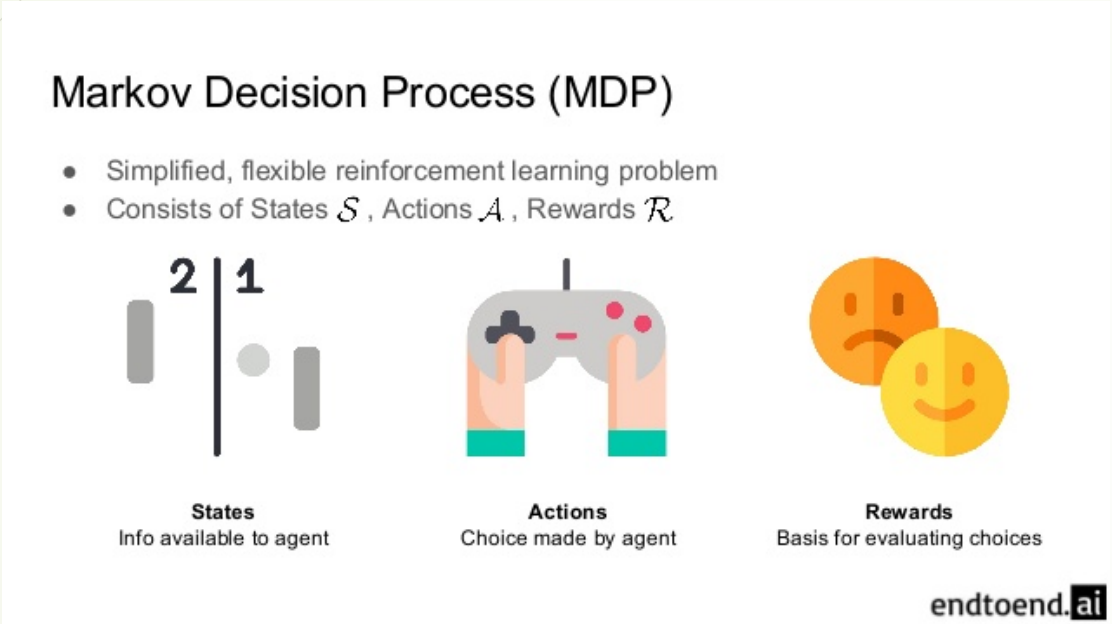
Typical applications: Image Recognition (CNN) and and Speech Recognition (RNN)

- Large-scale ANN faces the difficulty of maintain both efficiency and accuracy
- A large volume of data might come with redundancy
- Scalability is also challenging, especially with distributed memory parallelism (communications)

AI Methods and HPC

- **Markow Decision Process**

  Markov Decision Process (MPD) is **another paradigm used in AI**. MDP is appropriate for modelling a stepwise process under specific transition hypotheses.



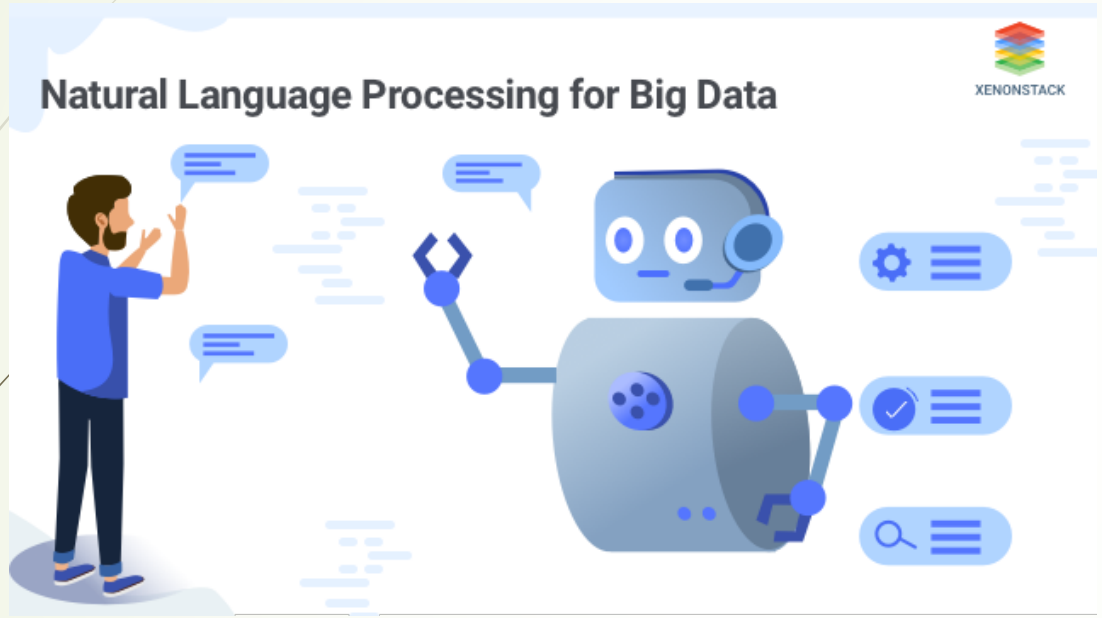
Typical applications: Path Monitoring, Inventory Management, Gaming

- MDP might be coupled with a ML algorithm (e.g. Obstacle Avoiding Robots)
- MDP has a strong linear (and Kronecker) algebra that is HPC challenging
- Numerical issues (iterative process) and scalability issues (multi-dimensional cases)

AI Methods and HPC

Natural Language Processing

Natural Language Processing (NLP) is an important topic in AI, covering techniques for Natural Language Understanding (NLU) and Natural Language Generation (NLG).



NLU
- ✓ Lexical Ambiguity
- ✓ Syntactic Ambiguity
- ✓ Semantic Ambiguity
- ✓ Anaphoric Ambiguity

NLG
- ✓ Text Planning
- ✓ Sentence Planning
- ✓ Realization

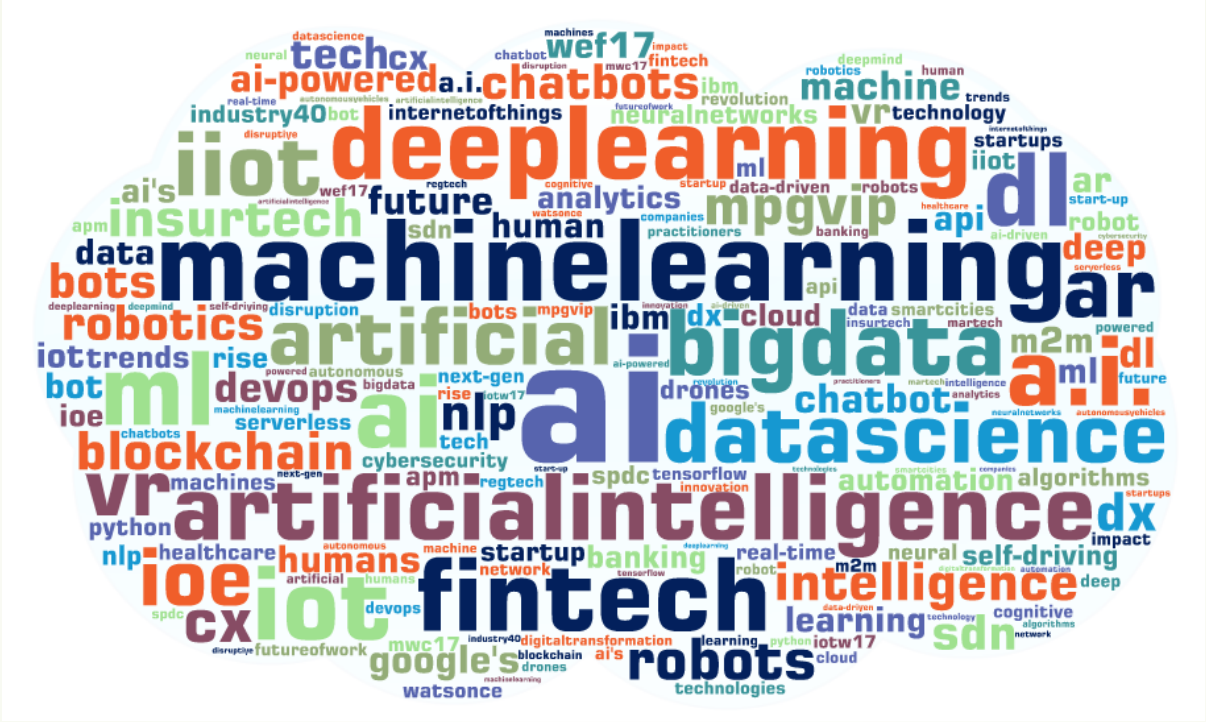
Typical applications: Chatbots, Log Analysis, Log Mining, Identification

- Ambiguity leads to a highly combinatorial process for NLP
- NLP can be coupled with ML and might involved a large volume of data
- Like any combinatorial algorithm, HPC efficiency and scalability are not trivial

Conclusion

- HPC advances tend to scale-up the expectations with AI
- Cutting-edge AI need to remain real-time, thus the strong need for HPC
- Connecting AI techniques might lead to heterogeneous HPC implementation
- AI-specialized HPC devices will be a central component for routine AI support
- As HPC is moving towards ambitious horizons, High Performance AI will follow similarly
- HPC ←→ AI will raise interesting fundamental/philosophical questions

END – QUESTIONS ?



High Performance Artificial Intelligence

VI Escola Regional de Alto Desempenho do Rio de Janeiro (ERAD-RJ 2020) – UFRRJ (Brazil)