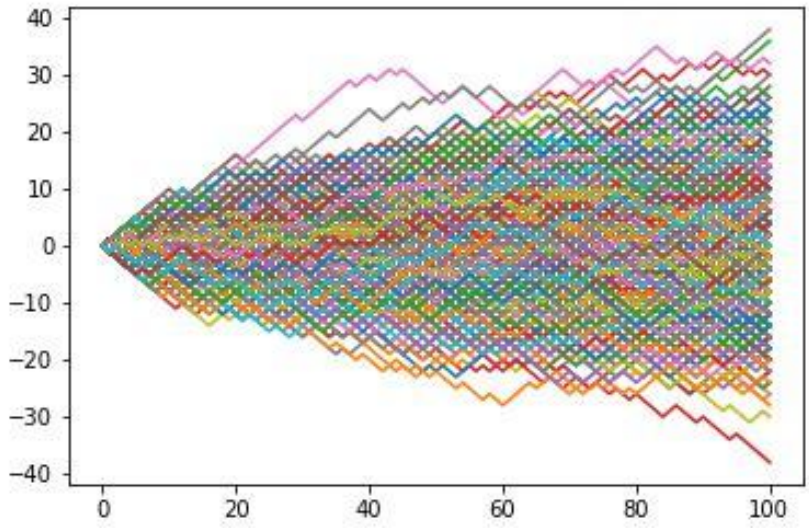
Un chemin pavé de probabilités, algorithmes et données : la théorie appliquée aux transports

Un quart de siècle pour un quart de plan

Dr. Arnaud de La Fortelle

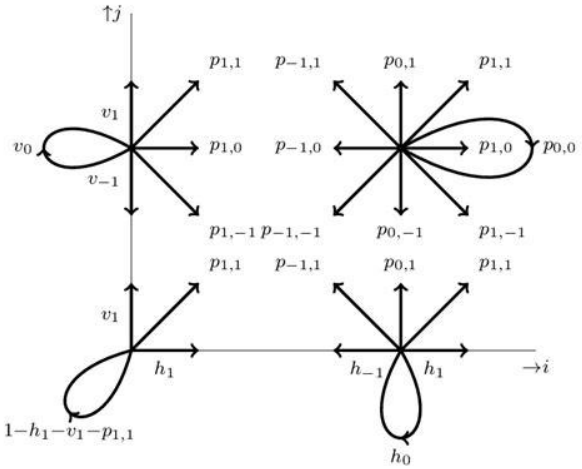
Large deviations for random walks

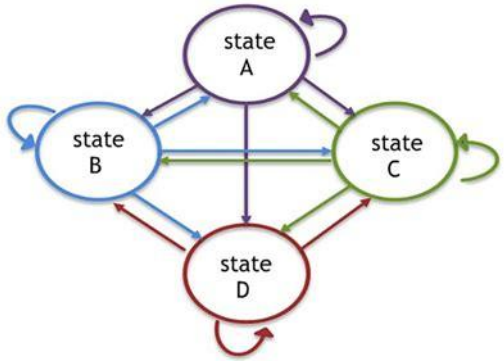
- Local dynamics (discrete or continuous time)
- Rate function = Entropy (or relative information)
- Proof relying on a twisted process
  - Related to harmonic functions, Martin boundary
- Several layers of theorems:
  - Empirical mean (of jumps)
  - Empirical measure of transitions
  - Empirical process distribution (sample path)



The quarter plane & higher dimensions

- Queueing networks
- Second vector field
  - boundary effects (from i.i.d. to Markov)
  - Non-homogeneous solutions
    - Difference between Large Deviations and Martin boundary
  - Optimization and space-time harmonic functions

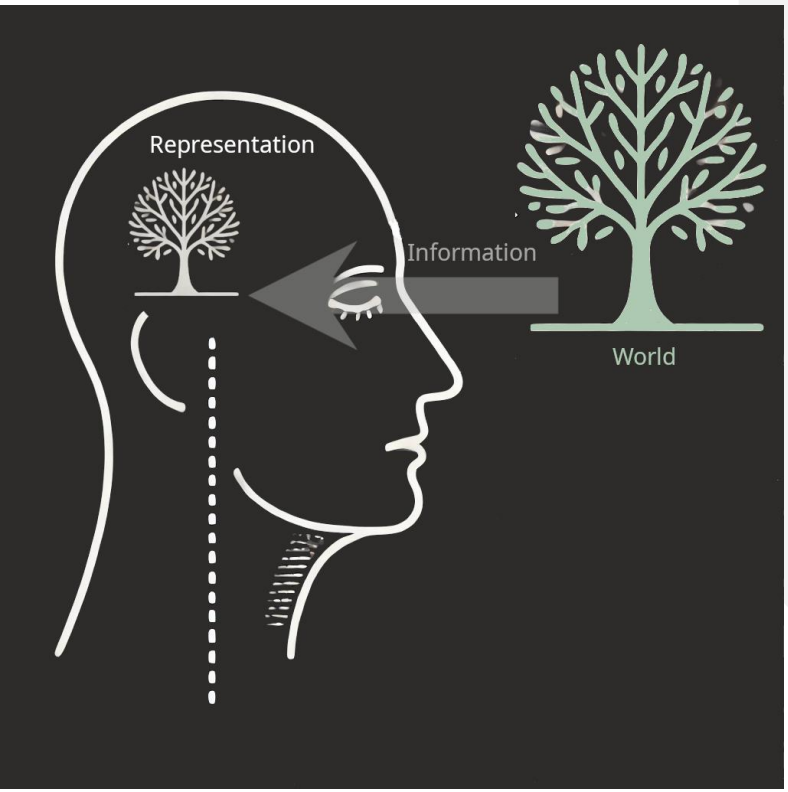




The lessons learned

- Local, homogeneous models
  - i.i.d. or Markovian jumps
  - Models and rules for events (e.g. simulations)
- Global properties, asymptotics
  - Direction of the random walk
  - Structures and measures (rare events)
- Building generative or harmonic functions
  - Strong links with the work of Flageolet
  - Mapping properties

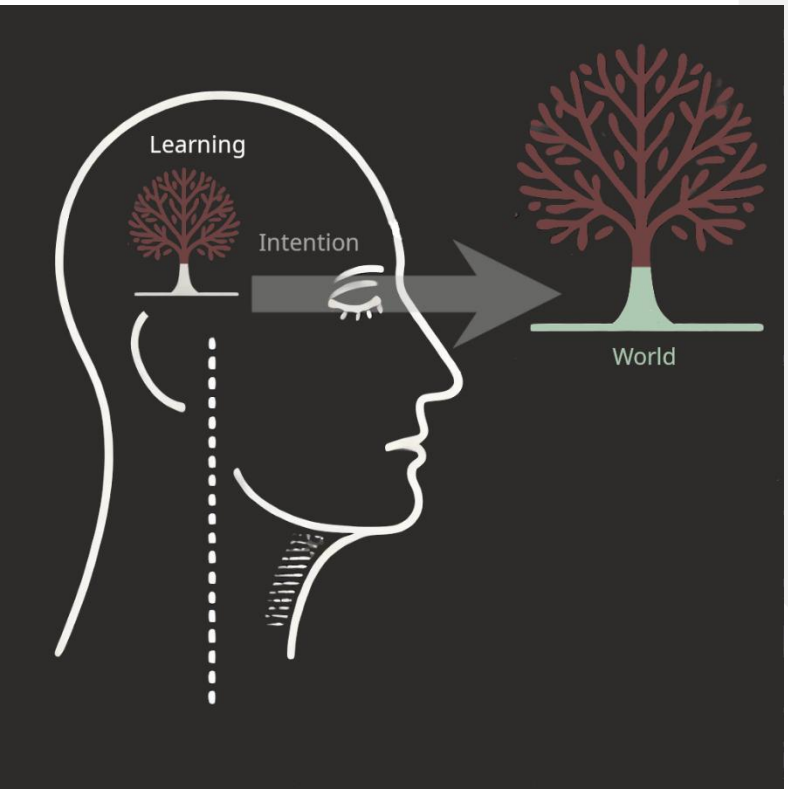
Intelligence & Data – Learning



Intelligence needs data

- Data as a way to access to reality
  - Data → information→ representation
  - Discovering order, laws
- Data opposed to models
  - Reasoning vs determinism

Intelligence & Data – Smart-Data



Data needs intelligence

- Curse of dimensionality
- Big Data limitation
  - Edge cases and the long tail
- Smart-Data
  - Active learning
  - Targeting the data

Data for AI - Expanding the learning base

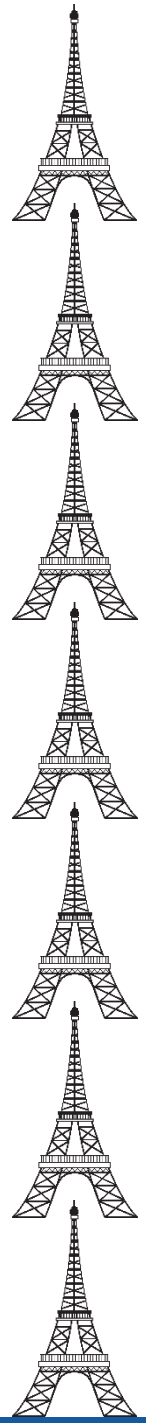
- Deep learning (Lecun)
  - MNIST dataset (1994): 60,000 32x32 images
  - LeNet Deep CNN (1998): 7 layers
- ImageNet: 15 million 300x300 images (~2014)
- GAN (Goodfellow, 2012)
- LLM: ChatGPT (2022)
  - 300 billion words = 500 GB for ChatGPT
  - Who has read 3 million books?
  - Bayesian inference is powerful, Bayes would be happy!
- Now Autonomous Driving?

Data from the real world

- 5 TB/h/veh.
  - Mainly because of cameras
- PB/day for a small fleet
- EB/year
  - 1 exabyte = 7 Eiffel towers of hard drives

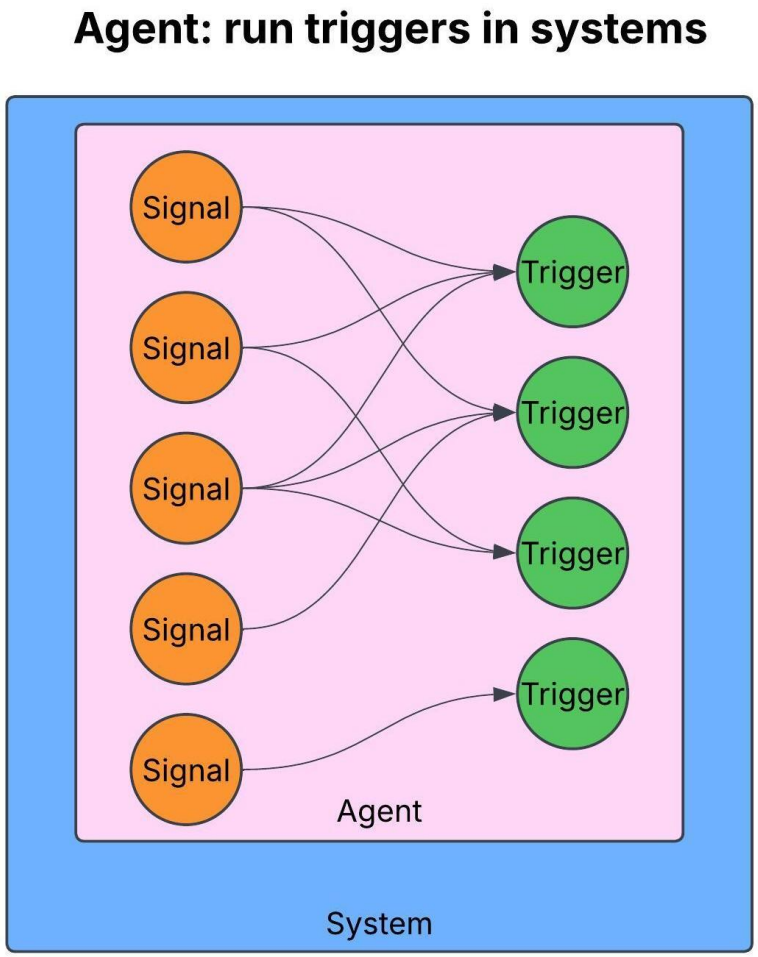
This is BIG DATA

- Not sustainable & not useful



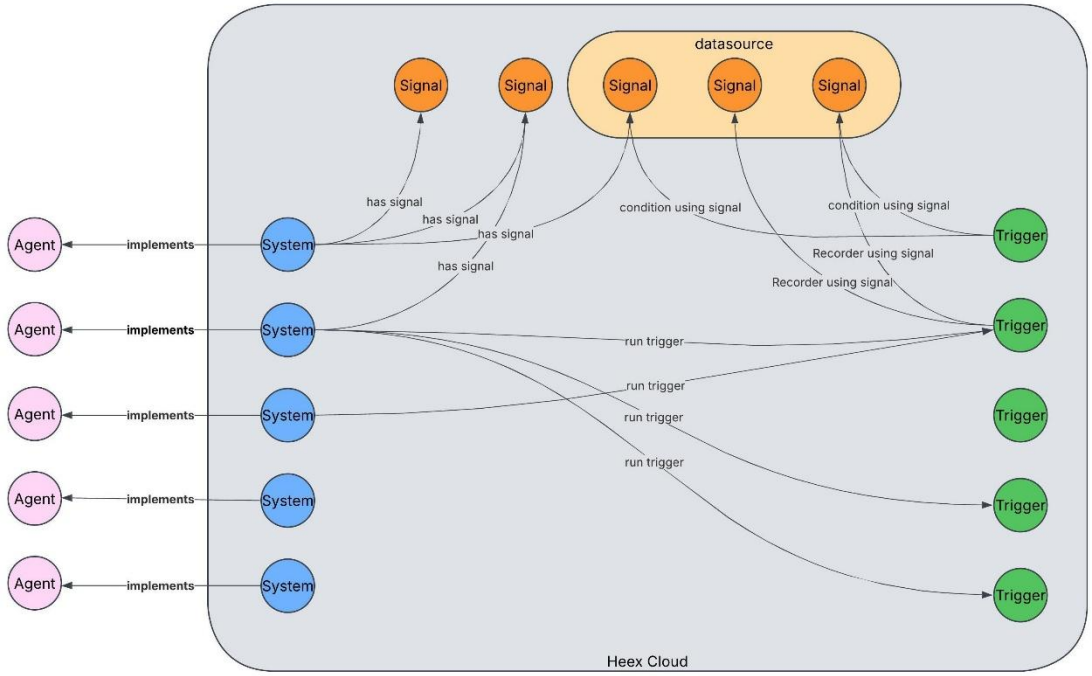
Smart-Data

- Event-driven data extraction
- What is the relevant data?
  - What do we need and want to learn?
  - Learning everything is not possible, nor useful
  - Target relevant events with Triggers



How do you define events?

- “Importance sampling”
  - Defines rules for Triggers
  - Optimization: failure, risk…



The challenges

- How to organize event definition (Triggers)?
  - Structures in the problem?
  - Measures of relevance?
  - Low-dimensional processes and sets?
- Can we learn from the signals?
  - Information about “hidden” states?
  - Are the signal equally informative?
- Is the data meaningful?
  - Underlying distribution of datasets?

The challenges

- How to organize event definition?
  - Structures in the problem?
  - Measures of relevance?
  - Low-dimensional processes and sets?
- Can we learn from the signals?
  - Information about “hidden” states?
  - Are the signal equally informative?
- Is the data meaningful?
  - Underlying distribution of datasets?

Conclusion

- Interesting events are rare
  - But there is no good definition
  - Structuring their definition is a direction
  - Semantics is probably needed here
- Real systems are difficult to model
  - Signals, failures, risks are very heterogeneous
  - AI and robotics challenge us
- There is a real impact
  - Data, Ai & robotics

Thank you!