Scalable survival analysis in a French hospital?

Jean Feydy

HeKA team, Inria Paris, Inserm, Université Paris-Cité

Methodological and Computational Advances in Survival Analysis

Tuesday, November 26, 2024 – Inria Paris

Today’s presentation

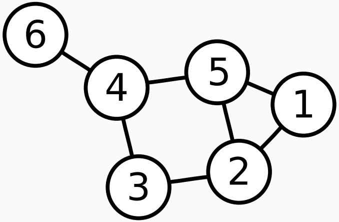
1. **Technological** context: the **GPU** revolution.

2. **Human** context: a triple intersection **CS + Stats + Medicine**.

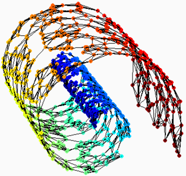
3. **Experience** feedback: three years of work in **pharmaco-vigilance**.

Technological context

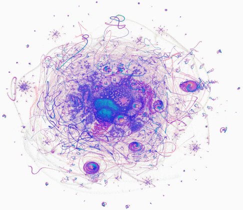
Towards a continuous description of large datasets [Pey11, Wil]



Simple **graph**.



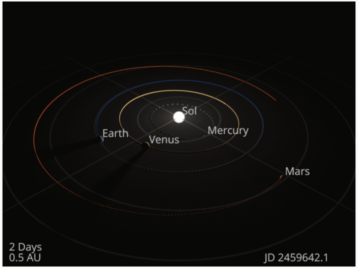
Underlying **surface**.

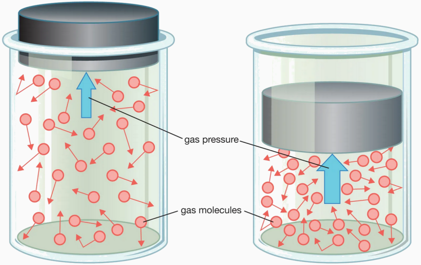


Visualization with UMAP.

The **language of continuous mathematics** has become mainstream in **data analysis** : gradient, density, manifold, test function…

A rich history in physics [Dat18, Bri, NWRC22]





The **solar system**.

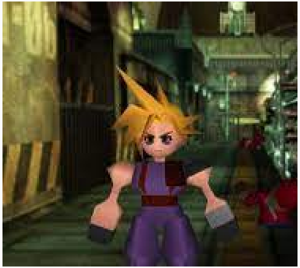
The **ideal gas** model.

Fluid simulation.

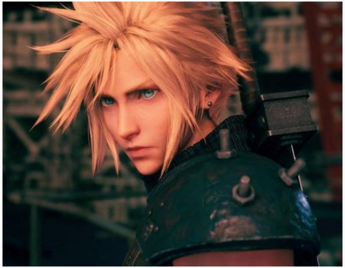
Research in **physics** $\Longleftrightarrow$ **High Performance super-Computers**

Access restricted to **institutional centres**.

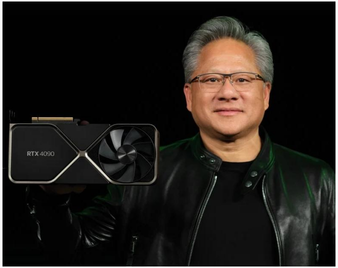
A recent history in computer graphics



FFVII on the PS1 – 1997.



FFVII on the PS4 – 2020.



Jensen Huang – 2022.

Research in **computer graphics** $\Longleftrightarrow$ **Graphics Processing Units**

Accessible to most research labs: revolutionary impact.

Modern hardware is the workhorse of the “AI revolution”

Statistics and machine learning have been studied for **decades**.

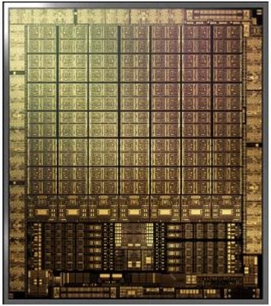
Breakthrough in 2010-15 : using **PlayStations** to do **science** became **easy**.

Research effort at all levels towards:

- Increasingly powerful **computers**.
- Increasingly convenient **software toolkits**.
- Increasingly relevant **models**.

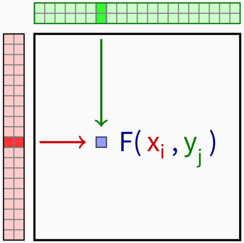
Spectacular results in a few applications

⟹ massive **investments**, industry + governments.

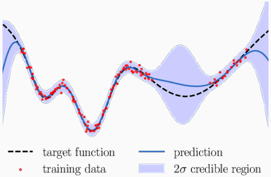


10,000 cores on a GPU.

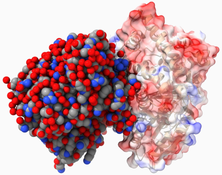
Example : the KeOps library www.kernel-operations.io



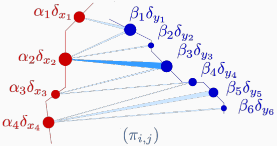
Symbolic matrices: distances, kernels, discrete transforms, point convolutions, attention...



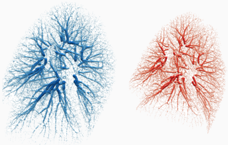
Gaussian processes.



Protein docking.



Optimal transport.



Lung registration.

Human context

HeKA: a translational research team

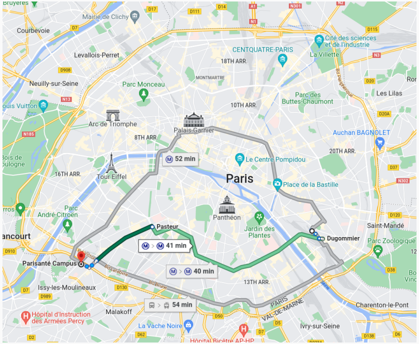
Hôpitaux

Inria

Inserm

Universités

PariSanté Campus: a fancy workplace for research on digital health



Mission: highlight public data warehouses

Our main sources:

- Patient records from the Pompidou and Necker Hospitals.
- National Rare Diseases Data Bank.
- Data from French **cartes vitales** – décret R 1461-12 du 29 juin 2021.

Strong support from Inria, Inserm, UPCité for
methodological research **and** **“simple” applications** to new data.

$\Longrightarrow$ A proven **“translational”** research model that is
well-established in the UK and growing at Inria.

Status differences

At **Inria** – plenty of **freedom**:

- Tenure around age 30.
- Full-time research, with remote work.
- Support for projects planned over 5 years.
- Dedicated **support** teams: skilled and supportive.

At the **hospital** – significant **pressure** :

- Tenure around age 40.
- **Care + Teaching + Research**, **day and night.**
- Fully embracing the *publish or perish* ethos with **SIGAPS** points.
- A **stable internet connection** is **not even** guaranteed!

Cultural differences

What is a **scientific truth**?

- In math: a formal **proof**.
- In computer science: well-**tested** software.
- In medicine: an **expert consensus**… and great responsibilities.

In medicine and biology, the **“prestige of the white gown”** is immense.

Analysts are often seen as **“pen pushers”** or **subordinates**.

Shifting **mindsets** is a challenge for our field.

Considering these factors is essential to succeed

Long-term strategy, enabled by Inria:

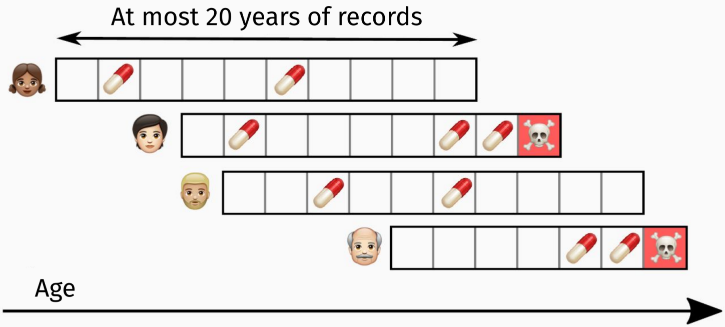
1. **Understand** the existing consensus,
   through lengthy discussions with doctors.

2. **Gradually introduce** modern tools (GPUs…),
   ensuring perfect **backward compatibility** with standard methods.

3. Leverage **experience** and **credibility** gained
   to introduce **new methods**, when **necessary**.

Experience in pharmaco-vigilance

Survival analysis: a classification problem on time series



Fundamental problem for:

- **Factories**: Which part will break next?
- **Businesses**: Which customer will stop returning first?
- **Doctors**: Which patient will develop cancer next?

Survival analysis: in practice

Standard Model: Cox Proportional Hazards (1972).
Time-dependent descriptors: Weighted Cumulated Exposures (WCE), …

Implementation: `survival` and `WCE` packages for R – 10M+ downloads.

Excellent packaging, but lacks GPU support:

- Acceptable for clinical trials (1k–10k patients).
- Prohibitively **slow** for large-scale studies, especially with time-dependent variables.

Epi-Phare with Anne-Sophie Jannot – 150k€ to scale up to nation-wide cohorts.
20 years of “receipts” from **cartes vitales**, for **70M+ French citizens**.

2021 – Implementation on GPUs

Striking similarity between survival and learning models:

- **Cox model** = **logistic regression** on a **graph** (1 node = 1 patient).
- Weighted Cumulative Exposures = **kernel** features.

I developed a **fast GPU solver** for these methods,
which Alexis Van Straaten packaged into an **R library**.

survivalGPU (for R and Python) produces **exactly** the same output
as the standard **survival** and **WCE** packages, but is **1,000 times faster**.

Two main outcomes:

- **Scalability**: work with **millions of patients** in just a few minutes.
- **Bootstrap and permutations**: repeat the same experiment 1,000 times
  to estimate confidence intervals.

2022 – Gaining access to carte vitale data (SNDS)

- **Inria** received authorization from the authorities via decree in June 2021.
- Inria funded **my one-week training** in May 2022 – thanks!

Problem: The `Ameli.fr` cloud is **not designed for innovative methods**.

Obsolete hardware configuration, with only three software options (c. 2005):

- Microsoft **Excel**.
- SAS (a **SQL** query system).
- **R** with a **frozen** collection of standard packages.

We need access to a modern machine (GPU + Python + R),
aligned with the **security** guidelines documented by the health insurance system.

2023 – Gaining access to a secure computer…

We got stuck on this point for **more than a year**:

- **Plan A:** Our request to use the Pompidou Hospital cloud was rejected.
- **Plan B:** We purchased a **modern workstation** (€5,000)…
  only to realize that it could not be approved for use.
- **Plan C:** Inria's IT department is working on a secure cloud solution…
  but cannot prioritize **specific procedures** for this data.

The situation around secure clouds is very **confusing**:

- **Old procedures** have been deprecated.
- **New platforms** like the Health Data Hub only support a few **pilot projects**.

$\Longrightarrow$ Fortunately, we met **Emmanuel Bacry** at PariSanté Campus!

2024 – Our first steps with the data

Funding for the project now comes from the Pr[ai]rie institute.
Emmanuel hired **Antoine Poirot-Bourdain** for a 2 years contract + 3 years of PhD thesis.

Getting access to drug reimbursement records over 5 years for 3,000,000 citizens:

- Hard to get past the usual regulatory boards without a specific, **narrow clinical question**.
- Fairly easy to **motivate the development of original methods** via the **Inria** access, thanks to Anne Combe and Michel Dojat.
- Then, the **Health Data Hub** could help us to get the data **on their secure platform**, thanks to Lise Vasteenkiste.

2025 – Finally some real medicine?

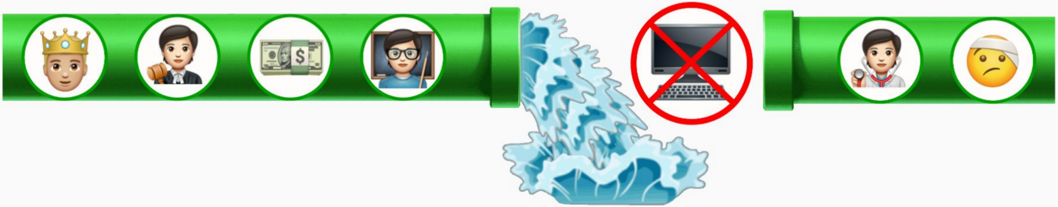
The HDH platform has **rough edges**… but **it works**!

Our first questions:

- What does real-life carte vitale data **look like**?
- How do we estimate **drug exposure** from reimbursement data?
- What are the adverse effects that are **easy to infer** from our data?
- Which drugs can we use as positive and negative **test cases**?
- Basic methods for adverse effect detection raise too many **false positives**. Do **modern methods** perform better?

We now work with a **healthy mix of mathematicians and pharmacists**.
I am also **tidying up survivalGPU** and adding support for Lasso, ElasticNet, etc.

We are getting there...



The **work accomplished** over the last decade is **substantial**.

However, the issue of digital **infrastructure** remains a **major blind spot**.

Information flow is poor:

- A proliferation of **incomprehensible** and **deprecated** procedures.
- **Leadership** is generally unaware of **on-the-ground bottlenecks**.
- We got lucky.

Conclusion

Hospitals are contrasting environments:

- **Futuristic** equipment in interventional radiology.
- “Our unit **cannot access its emails** today.”
- **Amazing doctors** working under absurd conditions.

As an **Inria researcher**:

- I have learnt a great deal through **collaborating with doctors and pharmacists**.
- I want to prioritize projects **that benefit all French citizens**.

I am **tired, but cautiously optimistic**: we’re finally ready to do some science.
Hopefully, the **new generation** will be able to build on this groundwork.

References

References i

Encyclopædia Britannica.

Ideal gas.

https://www.britannica.com/science/ideal-gas.

Datumizer.

Solar system orrery inner planets.

https://commons.wikimedia.org/wiki/File:Solar_system_orrery_inner_planets.gif, 2018.

CC BY-SA 4.0.

References ii

Mohammad Sina Nabizadeh, Stephanie Wang, Ravi Ramamoorthi, and Albert Chern.

Covector fluids.

ACM Transactions on Graphics (TOG), 41(4):113:1–113:15, 2022.

Gabriel Peyré.

The numerical tours of signal processing-advanced computational signal and image processing.

IEEE Computing in Science and Engineering, 13(4):94–97, 2011.

References iii

John Williamson.

What do numbers look like?

https://johnhw.github.io/umap_primes/index.md.html.