How can models help data scientists?

Lessons learned from an undercover agent

Sébastien Mosser
Winter Modelling Meeting #2
20.01.2020, San Vigilio, Italy

UQÀM | Département d'informatique

ace

Context: a 🐈 among the 🦡s

A. Charpentier
Math Department
(actuarial science) 🦡

M.-J. Meurs
CS Department
(natural language processing) 🦡

Yours truly
(cat among the raccoons) 🐈

Meanwhile, during a (not so) fictive lunch ...

🤬 - To be honest, I think that **Software Engineering is bullshit**, and that **Software Engineering profs are crooks** (at best).

😲 - Yeah, obvi... Wait, what?? Why?

😡 - Well, **you'll come to my group**, do nothing, but ...

- state that **we're doing nothing the right way**.
- Then you'll **tell us what to do** in a very arrogant (🇫🇷) way.
  - If we **succeed**, you'll say it's **obviously thanks to you**.
  - If we **fail**, you'll say it's **because we didn't follow / understand your advices**
- I mean, you look like a nice guy, but really, **stay far from us**".

😭 - *[cries in French]*

1 The RELAI Project

2 Undercover Report

3 Burying the hatchet

CIHR IRSC
Canadian Institutes of Health Research / Instituts de recherche en santé du Canada

SSHRC CRSH
Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada

NSERC CRSNG

RELAI

Respectful & ExpLainable AI to support struggling people and mental health practitioners

New Frontiers in Research Fund
Fonds Nouvelles frontières en recherche

UQÀM | Département d'informatique
FACULTÉ DES SCIENCES
Université du Québec à Montréal

RELAI is about Mental Health Care

- **4,9M** 15_{yo}+ Canadians needed mental health care *(in 2015)*
  - **1,6M** felt that their needs were not met (or only partially)
  - **33%** of Ontario's students (aged ~ $[12\text{-}17]_{yo}$) reported the need to talk to someone about their mental health *(in 2017)*.

- ***Mental illness is a leading cause of disabilities in Canada***
- **Challenges** :
  - *Canada's population is ~38M (~1 California, $\frac{1}{2}$ Italy)*
  - *Canada is the 2nd largest inhabited territory in the World*

population density: 4 inhabitants/km2. 90% of the population lives less than 150km from the border

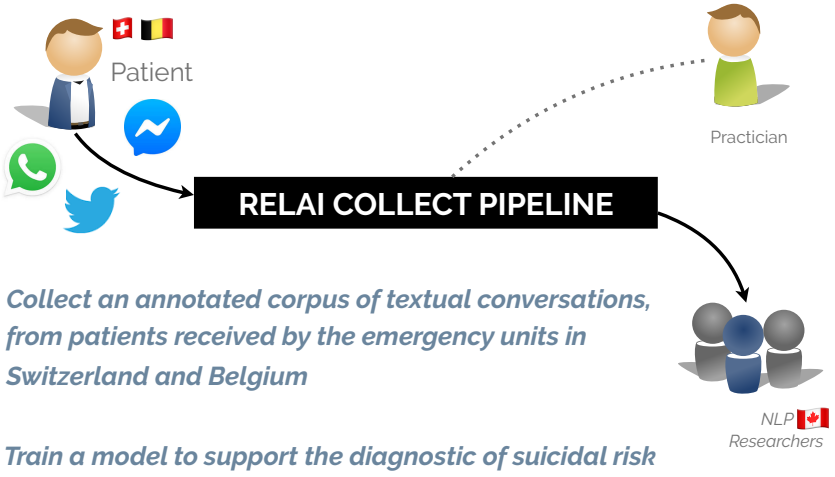
New Frontiers in Research: Interdisciplinary

- ***High Risk***, ***High Reward*** research fund
  - **Joint coordination** of the three federal funding agencies
    - *Natural Science & Engineering, Social Science, Health*
  - **Early career** fund (P.I. = Ass. Prof ≤ 5 years)
  - Created in 2018, Budget: 275M$ over five years

- Project must includes at least **researchers from two funds**
  - **Interdisciplinary** by construction

New Frontiers in Research Fund
Fonds Nouvelles frontières en recherche

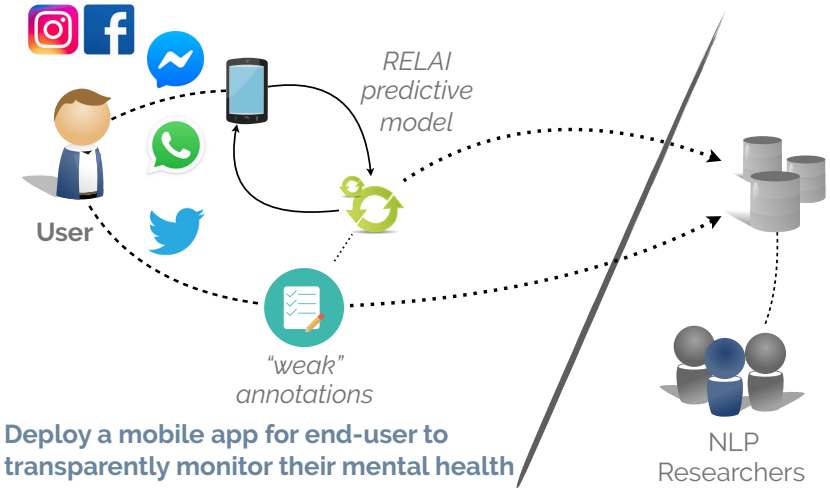
Vision: Frontline service for mental health [1/2]



Collect an annotated corpus of textual conversations, from patients received by the emergency units in Switzerland and Belgium

Train a model to support the diagnostic of suicidal risk

Vision: Frontline service for mental health [2/2]



Deploy a mobile app for end-user to transparently monitor their mental health

The RELAI team


M.-J. Meurs

- ***Principal Investigator*** : Marie-Jean Meurs
  - *Natural Language Processing, Artificial Intelligence*
- ***Co-Applicants & collaborators' expertises***:
  - ***Health*** : *Psychiatrists (3), Digital health*
  - ***Social sciences***: *Online user behaviour, Philosophy, Ethics, Technology appropriation (2), Lawyer.*
  - ***Engineering*** : *Privacy, Modelling*
- *The project aims to fund ~10 students (Postdocs, Ph.D.s & M.Sc.s)*

Undercover report

2

2019-...

Before talking, start by installing Eclipse.

- **Overheard** at WMM2020:
  - *It is usual to have 4 to 6 Eclipse instances on your Desktop*
  - **Can we agree that THIS IS NOT NORMAL!!!!**
    - ***Bran: "Eclipse is a dead-end. I am ~~Sirius~~ serious"***
- **Accidental complexity** of your approaches
  - *Not blaming Eclipse, it is a community problem*
  - ***You define YOUR tools***, *where integration is not a first priority*
- In the ***DevOps era***, lock-in with non-integrable tools is not an option

Strong state of practice

E.g., R, Python & de facto libraries. ***Nothing else (not bankable)***

It's MY job. Please do YOURS.

- **Graphical & LowCode approaches are out of questions**
- Also reject a lot of "SE for AI" current approaches
  - "*It is my job to optimize a model, please stop trying*".
  - "*You're rebranding your stuff as AI / Deep Learning, but often a simple decision tree would be better*"
  - *"You guys are 🤡 jumping into the 💰💰💰 wagon by doing 💩*"
- Assumption: **Why should I change? What can you bring to me?**
  - New tools are out of the question. The code is the truth.
  - ***"I can do my job without you guys. You're useless"***

Why should I write maintainable code?

"The new version speeds up the micro bench time, from 20s to 60ms."

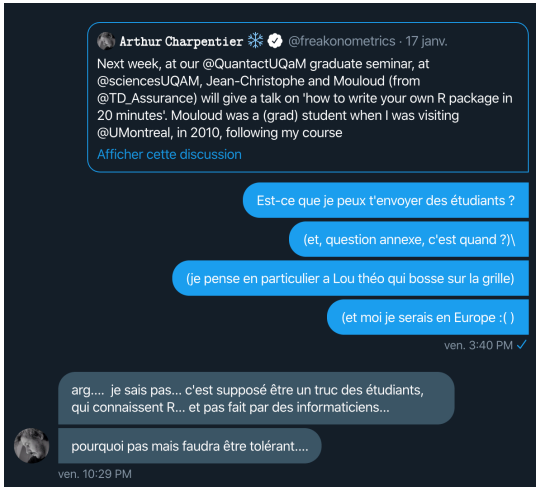
```
tags = [0,800,2*8*95,3*8*90,4*8*80,5*8*75]

def best(t) :
    t=tuple(sorted(t))
    m=min(t[0],t[2]-t[1])
    return tags[5]*(t[0]-m)+ tags[4]*(2*m+t[1]-t[0])+
           tags[3]*(t[2]-t[1]-m)+ tags[2]*(t[3]-t[2])+
           tags[1]*(t[4]-t[3])

def compute(l) :
    return best(tuple(l.count(i) for i in range(5)))/100
```

"This code will be thrown out in a month, after the experimental campaign"

Fear of judgment



(1)
Arthur announces a seminar about writing R package for mathematicians

(2)
Seb asks if he can send his student (***who works on deploying Arthur's R code on the federal grid***) to listen to the talk

(3)
Arthur freaks out.
(radio silence mode)

Tests / Tries / Demos / Campaigns

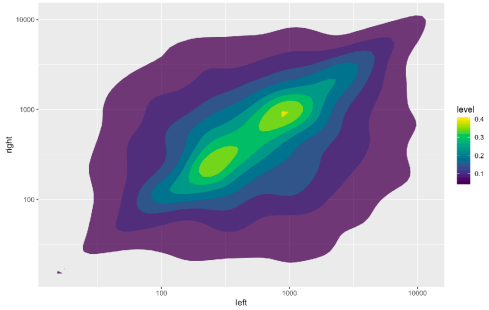
- **Surprisingly large culture of Unit Testing / CI**
- **Model tuning** relies a lot on "**trials-and-errors**" approaches
  - This is part of the job, and involve long running tasks
  - Tasks on Compute Canada can run **up to 26 days**
    - *RAC min. requirements: 50 core-years, 10 GPU-year, 10Tb*
- When the model / computation is ready
  - Building **nice demos** / visualization with the result
  - Participate to **competitions** *(e.g., eRisk)*

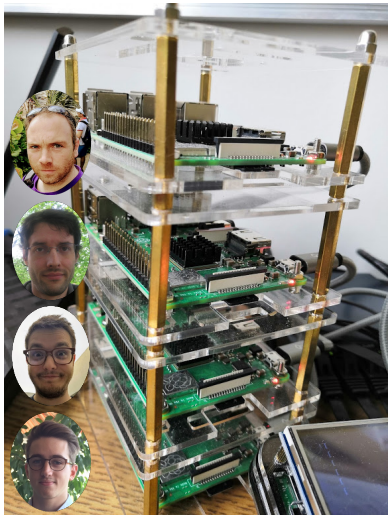
Without a RAC, you have only access to 20% of the resources (all comers partition)

Burying the hatchet

3

Preliminary step: invite 🦡s to 🐈's parties





🐈 **Quantitative analysis of Git commits (grid computing, large-scale)**

🦡 ***Needs: statistical methods, principal component analysis***

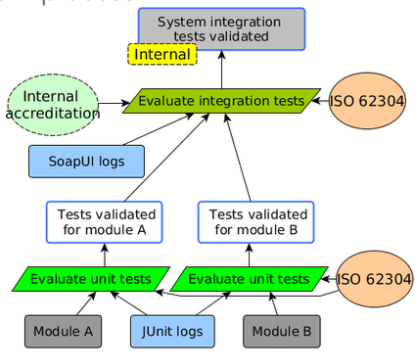
Requirements "models"

"trust in me"

- Switched to **facilitator / coach** role
  - *Who are your personas?*
  - *What are the epics we are targeting in the project?*
  - *Can we validate such epics with the practicians?*
- **Immediate insights** :
  - **Lightweight** approach (*i.e., brainstorming sessions*)
  - **Integrated** with GitLab / on-premise cloud service
  - **Measurable benefits** : misunderstanding / fuzzy areas

Convincing Ethic committees

- Need to obtain **three different ethics committees certifications**
  - (**Super**) Long & (**Super**) tedious process *(15 days meeting in 🇨🇭 right now).*
- Looks like a **classical "certification"** process
  - **Justification diagrams**!
- **Insights (ongoing)**:
  - Strengthen the DMP
  - Give confidence to defend it
  - **Even if not executable!**

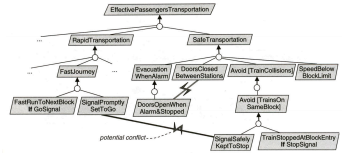


Patient's trust for data collection

- **Patients' trust is mandatory** in the healing process
- Patients need to give access to **VERY personal data**
  - *Conversations with "significant other", family, close friends*
  - *Social networks private messages*
- We all heard news of several "**data leaks**" in recent headlines
  - **Do you have any guarantees?**

- **Proposition : Use a "software factory" (SPL) approach**
  - *We generate your very own app, ensuring your level of consent*

User's application optimization



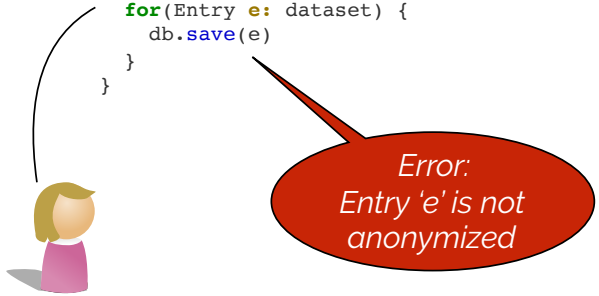
- User's give us access to private information in the mobile app.
  - But "***not all data sources are created equals***"
  - It depends of the ***suicidal risk*** of the user + ***cultural factors***
- ***Proposition :***
  - ***Drive the requirements gathering process with goal models***

Experiment deployment that scale

- **Using a computation grid / cluster is not that complicated**
  - *Pre-requisite: you need to know shell scripting* ✅
- **Using PROPERLY such equipment is complicated**
  - Need to have a ***proper understanding of system administration***: *networking, time estimation, shared disks, file system permissions, parallelizable tasks, monitoring …*
  - *Know how to **"play" with the scheduler** to stay under the radar*
- **Proposition**: *Speedup: from two years to two days.*
  - **Model the workflow** *(with bash 😱), automate as much as possible*

Pipeline Development

```
public void method store(Set<Entry> dataset) {
  Database db = RemoteStorage("...");
  for(Entry e: dataset) {
    db.save(e)
  }
}
```



*Enhance tooltips and error handlers in IDEs with **composable requirements** that are driven by external & evolving concerns*

Time to conclude !

Takeaway message

- **AI / Data science researchers**
  - have tons of very ***interesting modelling problems***
  - **do not rely on our tools**,
    - **do not want to use them**
  - **Do not understand the benefits** of working with us
    - *Think we will be add complexity to already complex problems.*
  - are **not convinced (yet) by the "SE for AI"** literature
- **Maybe we are doing it the wrong way 😇!**