Big Clinical Data

V. Claveau

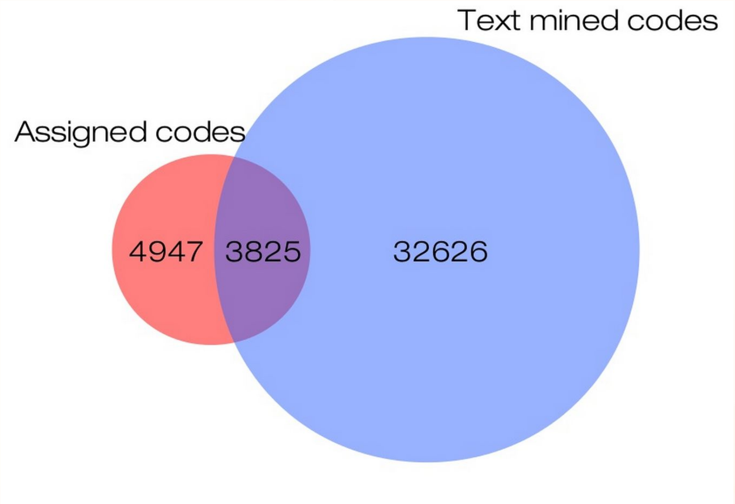
Objectives of the project

Leveraging the large unstructured data in the medical field

- Electronic Health Records
- Biomedical literature
- Research databases (eg. clinical trial protocols)
- Social Networks, specialized websites…

Focus on French

- but also Eng, Br

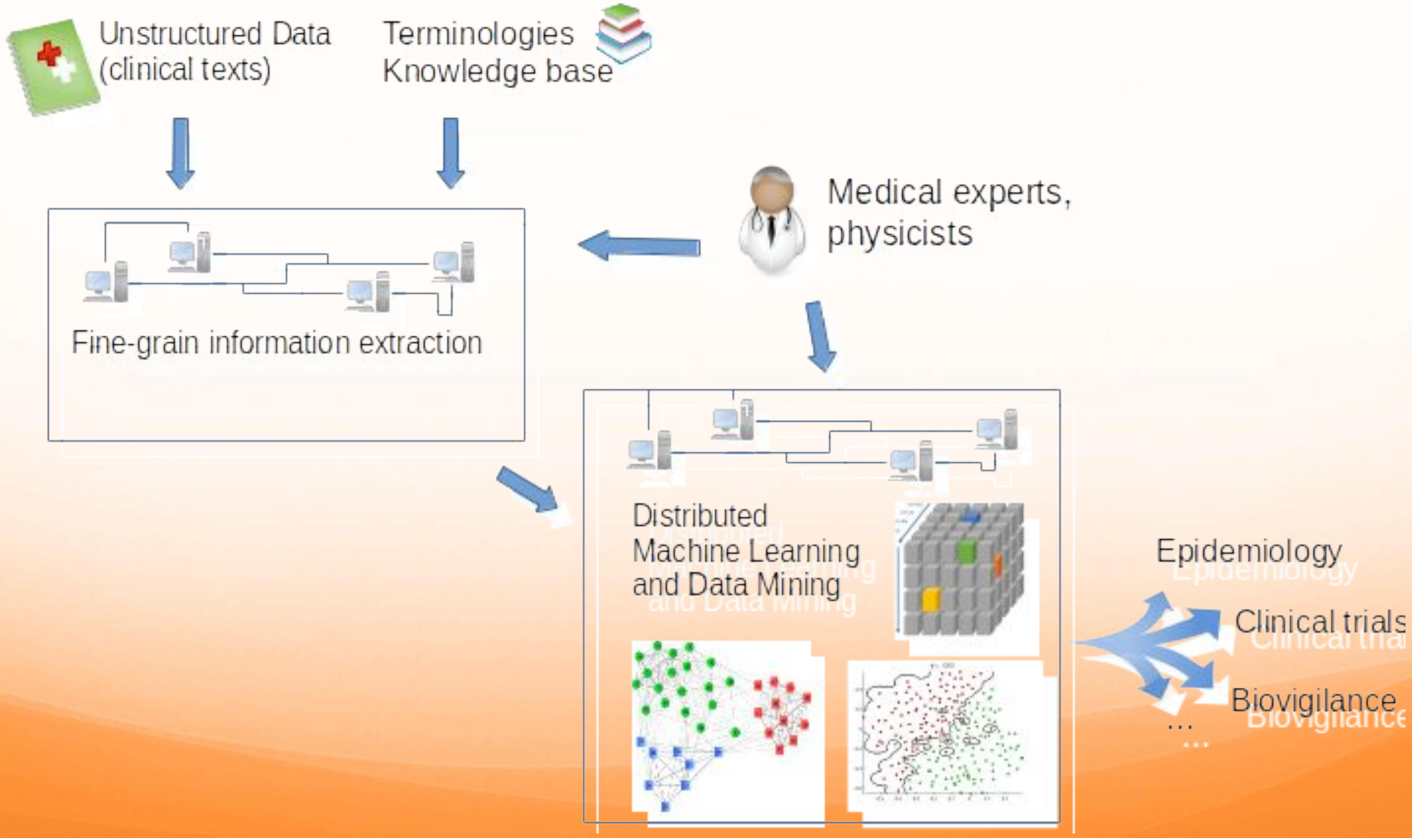


Codes of disease / medical activity (ICD10 ~ PMSI) found by text mining in patient records vs. codes manually assigned [Roque et al. 2011]

Objectives of the project

1 - Exploiting the clinical narratives in hospitals

- extract medically relevant information



Unstructured Data (clinical texts)

Terminologies Knowledge base

Medical experts, physicists

Fine-grain information extraction

Distributed Machine Learning and Data Mining

Epidemiology

Clinical trials

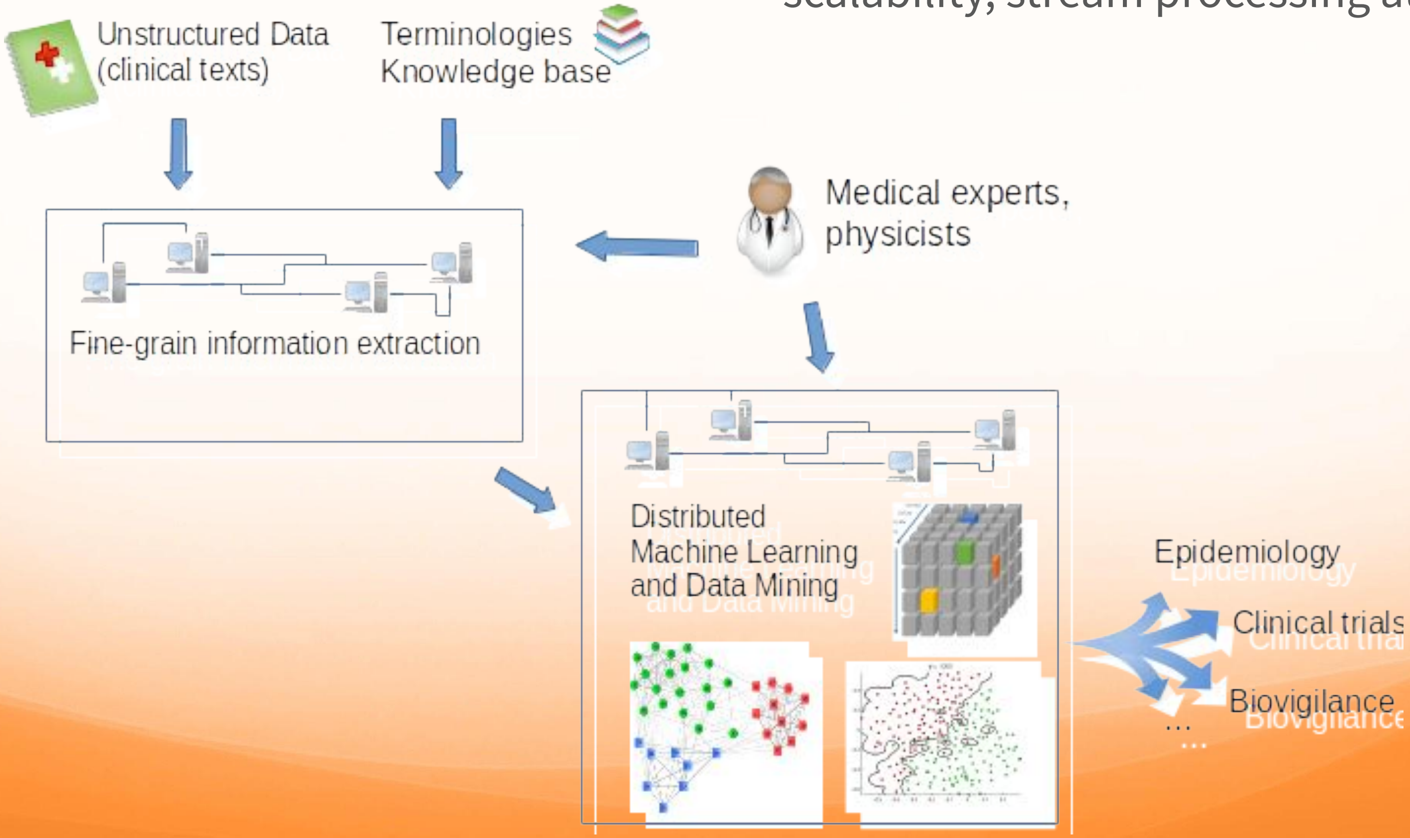
Biovigilance

…

Objectives of the project

2 - Addressing the distributed systems issues

- scalability, stream processing at runtime…



Unstructured Data (clinical texts)

Terminologies Knowledge base

Fine-grain information extraction

Medical experts, physicists

Distributed Machine Learning and Data Mining

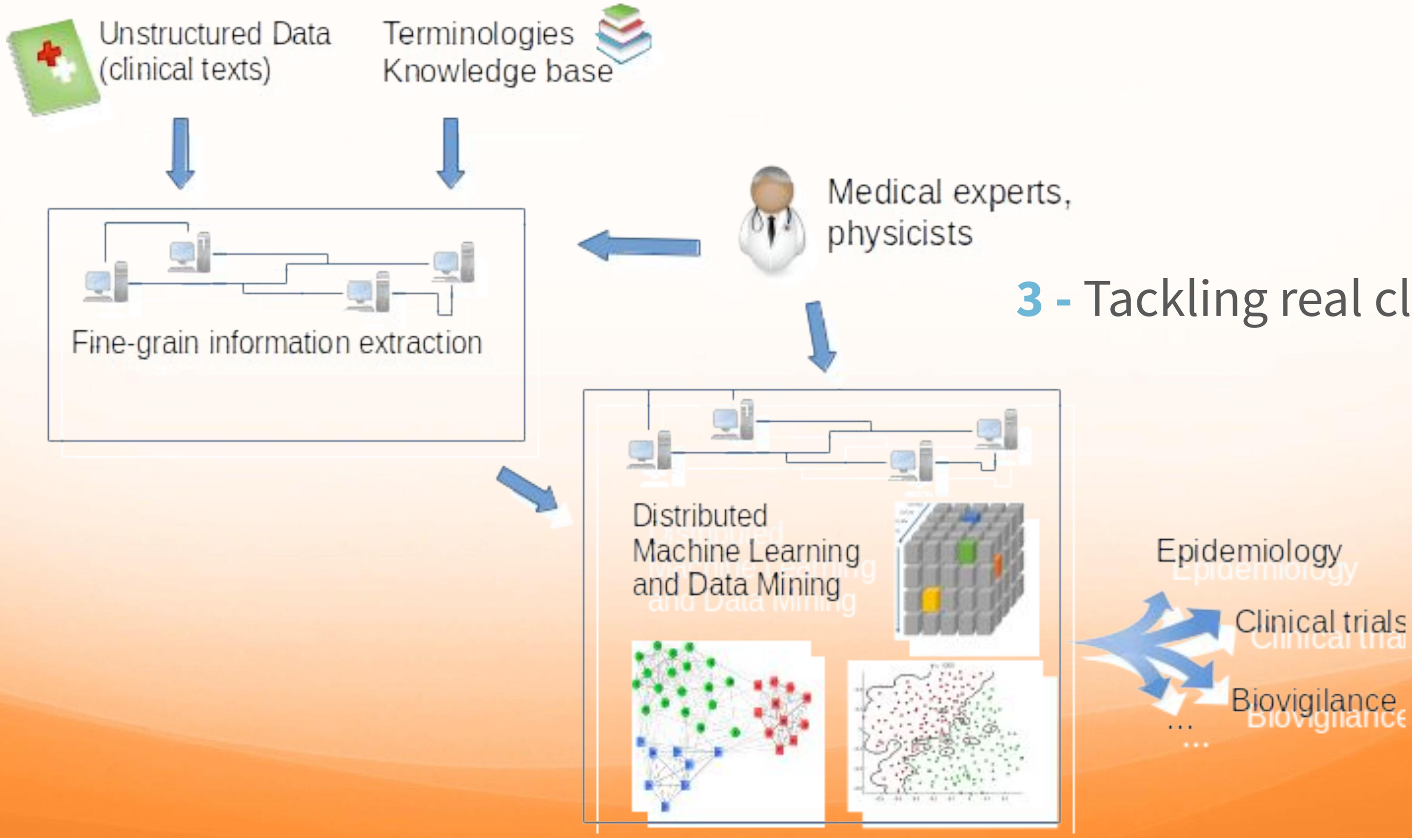
Epidemiology

Clinical trials

Biovigilance

…

Objectives of the project



3 - Tackling real clinical tasks

Summary of the project

Started: fall 2016,
End: Dec ~~2019~~ 2020

Funding

- 2 PhD scholarships
- 1 12-month Post-Doc

HBD (Health Big Data) SEPIA-LTSI/INSERM

- medical expertise, clinical data warehousing, with CHU Rennes

CIDRE, IRISA-Inria

- distributed algorithms, large-scale data stream processing

LinkMedia, IRISA-Inria

- Information retrieval, Natural Language Processing

External partner: CNRS - STL

- Biomedical Natural Language Processing

Main results

Collaboration

Health Informatics dept @ PUCPR, Curitiba, Brazil

- additional funding from CNRS/FAP
- same scientific objectives: nice synergy
- provide data + expertise in Br. Portuguese
- visits, student exchange + training
- several common publications

Caribbean Sea
Nicaragua
Venezuela
Guyana
Suriname
Colombia
RR
AP
Ecuador
AM
PA
MA
CE
RN
PB
PI
PE
AL
Brazil
AC
TO
SE
Peru
RO
BA
MT
GO
Bolivia
MG
ES
MS
Paraguay
RJ
PR
SC
Chile
RS
Uruguay
Argentina

Main scientific results

Neural approaches for negation and uncertainty detection

- pioneer work for French, state-of-the-art results for English
- available as a web-service

New top-*k* (most frequent) items detection algorithm

- one-pass, time-decaying, resource efficient and distributed
- especially suited for massive and highly dynamic data streams

Neural approaches for ICD-10 automatic coding

- on real massive patient dataset from Hospital of Rennes

Resources, software

Several new datasets Fr, Br and Eng

- several layers of medical and linguistic annotation
- can be shared!

Negation and uncertainty detection

- especially suited for clinical documents
- available as a web-service: https://allgo.inria.fr/app/negdetect

Part-of-Speech tagger for French

- especially suited for clinical documents
- available as a web-service: https://allgo.inria.fr/app/tagex

Main use-cases

Cohort selection for clinical trials

- n2c2 challenge (Eng), Br

Pharmacovigilance

- SMM4H challenge (Eng)
- self report of Adverse Drug Reaction

Epidemiology

- medical events (ICD-10 codes) in patient reports (Fr)

Indexing, research of expertise

- DeFT challenge (Fr)

Patient information

- DeFT challenge (Fr)

Animation

International workshop, Rennes 2016

- organized during the kickoff-meeting with NLP and medical informatics teams

Text-mining challenge DeFT, Toulouse 2019

- with T. Hamon and C. Grouin (LIMSI)
- Information extraction in French clinical texts
- 9 teams involved; 30 participants at the workshop; published proc

Training, teaching

- master 2 bio-informatics, Rennes 1: C. Moro, L. Oliveira N. Grabar, V. Claveau
- master 2 medical informatics, Rennes 1: NLP, N. Grabar, V. Claveau
- PhD student exchanges with Brazil (Fr→Br, Br→Fr)

Publications

4 journals, 1 WS proceedings, 9 int conferences, 7 Fr conferences

- Vincent Claveau and Ewa Kijak: **Direct vs. indirect evaluation of distributional thesauri**. *International Conference on Computational Linguistics, COLING* : 2016
- Vincent Claveau, Ewa Kijak. **Strategies to select examples for Active Learning with Conditional Random Fields**, *CICLing 2017 - 18th International Conference on Computational Linguistics and Intelligent Text Processing*, Apr 2017, Budapest, Hungary. pp.1-14.
- Clément Dalloux. **Détection de l'incertitude et de la négation : un état de l'art**, *RECITAL 2017 - 18ème Rencontre des Étudiants Chercheurs en Informatique en Traitement Automatique des Langues*, Jun 2017, Orléans, France. pp.1-14
- Clément Dalloux, Vincent Claveau, Natalia Grabar. **Détection de la négation : corpus français et apprentissage supervisé**, *SIIM 2017 - Symposium sur l'Ingénierie de l'Information Médicale*, Nov 2017, Toulouse, France. pp.1-8
- Vincent Claveau, Lucas Oliveira, Guillaume Bouzillé, Marc Cuggia, Claudia Cabral Moro *et al.* **Numerical eligibility criteria in clinical protocols: annotation, automatic detection and interpretation**, *AIME 2017 - 16th Conference in Artificial Intelligence in Medecine*, Jun 2017, Vienne, Austria. pp.203-208.
- Natalia Grabar, Vincent Claveau. **Critères numériques dans les essais cliniques : annotation, détection et normalisation**, *Actes de la conférence TALN 2017*, Jun 2017, Orléans, France
- Natalia Grabar, Vincent Claveau, Clément Dalloux. **CAS: French Corpus with Clinical Cases**, *LOUHI 2018 - The 9th International Workshop on Health Text Mining and Information Analysis*, Oct 2018, Bruxelles, Belgium. pp.1-7
- Clément Dalloux, Natalia Grabar, Vincent Claveau, Claudia Moro. **Portée de la négation : détection par apprentissage supervisé en français et portugais brésilien**, *TALN 2018 - 25e conférence sur le Traitement Automatique des Langues Naturelles*, May 2018, Rennes, France. 1, Actes de la conférence TALN 2018 - Traitement Automatique de la Langue Naturelle
- Anne-Lyse Minard, Christian Raymond, Vincent Claveau. **IRISA at SMM4H 2018: Neural Network and Bagging for Tweet Classification**, *SMM4H 2018 - Social Media Mining for Health Applications, Workshop of EMNLP*, Oct 2018, Brussels, Belgium. Pp.1-8

CominLabs

Publications

- Emmanuelle Anceaume, Yann Busnel, E. Shulte-Geers, Bruno Sericola, **Optimization results for a Generalized Coupon Collector Problem**, *Journal of Applied Probability*, 53(2): 622-629 (2016).
- Emmanuelle Anceaume, Yann Busnel, **Lightweight Metric Computation for Distributed Massive Data Streams**, *Transactions on Large-Scale Data and Knowledge-Centered Systems*. Volume 33. 2017.
- Emmanuelle Anceaume, Yann Busnel, Vasile Cazacu. **Finding Top-k Most Frequent Items in Distributed Streams in the Time-Sliding Window Model**, *DSN 2018 - 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, Jun 2018, Luxembourg,
- Emmanuelle Anceaume, Yann Busnel, Vasile Cazacu. **On the Fly Detection of the Top-k Items in the Distributed Sliding Window Model**, *NCA 2018 - 17th IEEE International Symposium on Network Computing and Applications*, IEEE, Nov 2018, Boston, United States.
- Emmanuelle Anceaume, Yann Busnel, Vasile Cazacu. **L'art d'extraire des éléments du top-k en temps réel sur des fenêtres glissantes réparties**, *ALGOTEL 2019 - 21èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications*, Jun 2019, Saint Laurent de la Cabrerisse, France
- Natalia Grabar, Cyril Grouin, Thierry Hamon, Vincent Claveau, **Corpus annoté de cas cliniques en français**, *conference TALN 2019*, Toulouse, France
- Natalia Grabar, Cyril Grouin, Thierry Hamon, Vincent Claveau, **Actes de l'atelier et compétition DeFT 2019**, Jul 2019, Toulouse, France
- Natalia Grabar, Cyril Grouin, Thierry Hamon, Vincent Claveau, **Recherche et extraction d'information dans des cas cliniques. Présentation de la campagne d'évaluation DEFT 2019**, *Proc. of the DeFT 2019 workshop*, Juillet 2019, Toulouse, France
- Clément Dalloux, Vincent Claveau, Natalia Grabar. **Détection de la négation : corpus français et apprentissage supervisé**, *journal TSI - Technique et science informatiques*, to appear 2019
- Cyril Grouin, Natalia Grabar, Vincent Claveau, Thierry Hamon, **Clinical Case Reports for NLP**, *workshop BioNLP 2019*, Nov 2019, Hong-Kong
- C. Dalloux, V. Claveau, N. Grabar, **Speculation and Negation detection in French biomedical corpora**, conference RANLP 2019
- N. Grabar, C. Dalloux, V. Claveau, **CAS: corpus of clinical cases in French**, Journal of Biomedical Semantics (JBMS), to appear 2020
-

Publications

Submitted, under review

- C. Dalloux, V. Claveau, N. Grabar, L Oliveira , C Moro, Y. Gumiel, D. Carvalho, **Cross-lingual and Cross-domain Detection of Negation and of its Scope**, *journal Natural Language Engineering, special issue on Negation*
- Y. Gumiel, L. Oliveira, V. Claveau, N. Grabar, E. Paraiso, C. Moro, D. Carvalho, **Temporal Relation Extraction in Clinical texts: A Systematic Review,** *journal ACM survey*
- C. Dalloux, G. Bouzillé V. Claveau, N. Grabar, M. Cuggia : **ICD-10 automatic coding from clinical big data : a neural network approach.**

To be submitted

- Y. Gumiel, L. Oliveira, D. Carvalho, L. Ronnau, A. Pucca da Silva, V. Claveau, N. Grabar, M. Cuggia, C. Dalloux, C. Moro, **Multilingual algorithm for eligibility criteria detection in English and Portuguese Language,** journal
- Cyril Grouin, Natalia Grabar, Vincent Claveau, Thierry Hamon (Editors). **Recent advances in clinical text-mining.** Journal RiDOWS special issue, 2020.

Future

Still on-going work

More use-cases on CHU Rennes datasets

- embed the developed algorithms in the data-warehouse
- address new use-cases

More annotated datasets

- corpus CAS - sharable clinical texts
- still growing + still adding reference annotation
- planned to be used for more text-mining challenges

More software

- completing the pipeline for French clinical
- adding web-services for all our information extraction modules

Continued collaboration

PUCPR, Curitiba, Brazil

- co-advised PhD
- planned exchange as invited professor, one month in 2019/2020
- several common publications to be submitted
- actively looking for funding!

Bonus

PhD students

Clément Dalloux

- Indexing and information extraction in clinical texts
- Started in Dec 2016
- PhD to be defended in February 2020

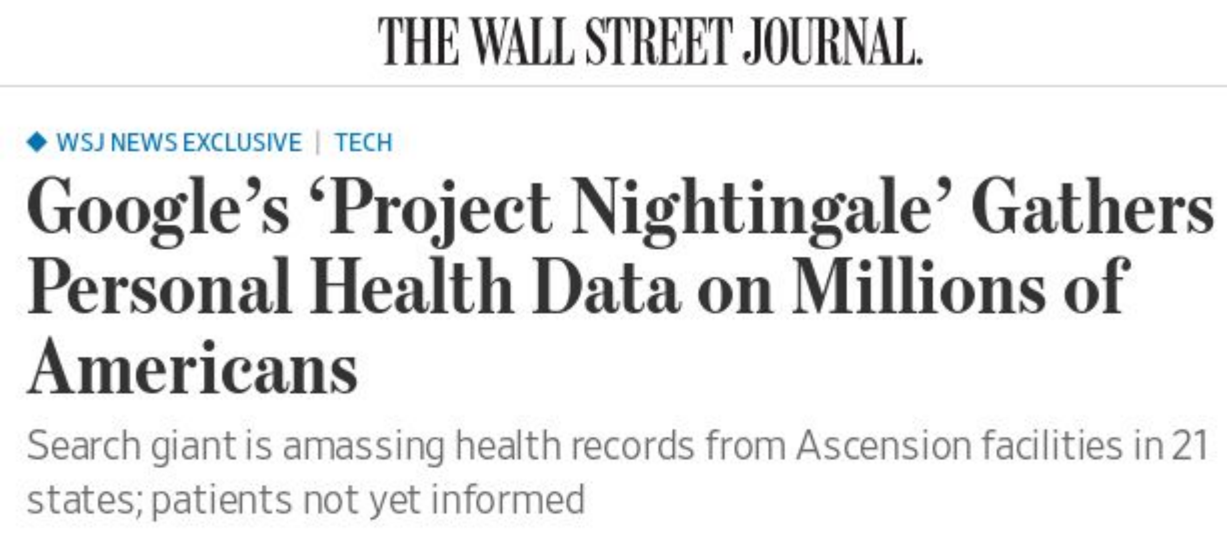
Vasile Cazacu

- Distributed computing for clinical data mining
- Started in Feb 2017
- PhD to be defended in February 2020

Research vs. hospital

A matter of security and sovereignty

- seldom access to French datasets slows down research
- at the same time:



THE WALL STREET JOURNAL.

◆ WSJ NEWS EXCLUSIVE | TECH

Google's 'Project Nightingale' Gathers Personal Health Data on Millions of Americans

Search giant is amassing health records from Ascension facilities in 21 states; patients not yet informed

Others

Datasets

- clinical cases (Fr), annotated with keywords, gender, age, outcome, disease, symptoms, entrance reasons, negation…
- clinical trials (En, Fr, Br), annotated with negation, numerical information
- medical concept embedding, any language
- eHOP Dataset : corpus of real patient data including discharge summary and ICD-10 coding

Software/API

- TagEx: Part-of-Speech tagger for French (medical and general)
- NegDetect: negation detection (cues and scope)
- Baseline systems from DeFT task 1, 2, 3

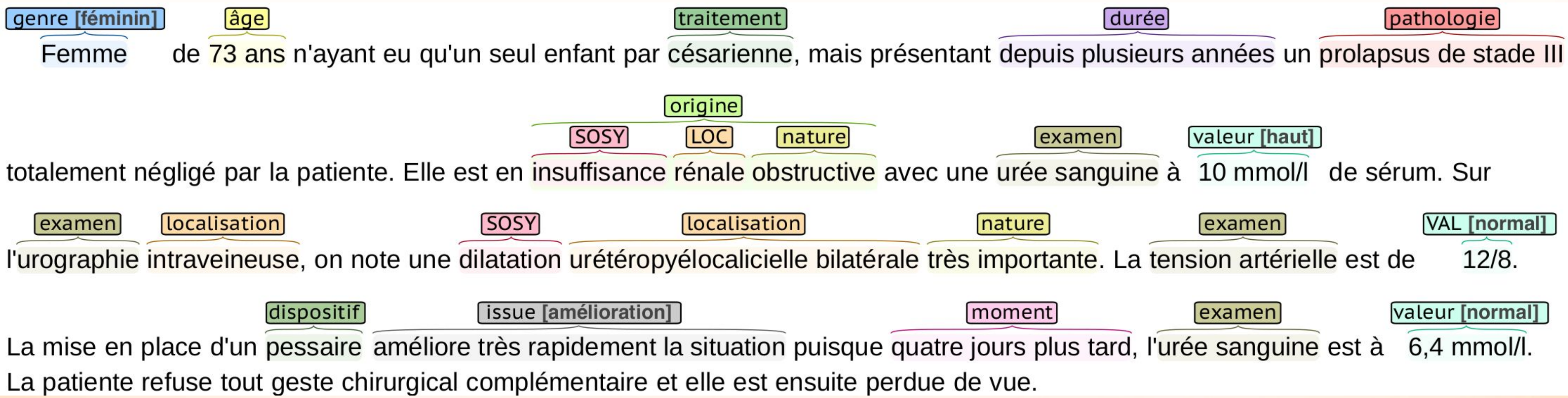
Annotated corpus

Clinical cases

- 4 300 cases, that is ~ 1 500 000 word occurrences
- different sources: scientific literature, didactic material, patient associations…
- different specialties : cardiology, urology, oncology, obstetrics, pneumology, pharmacology…

Annotated corpus

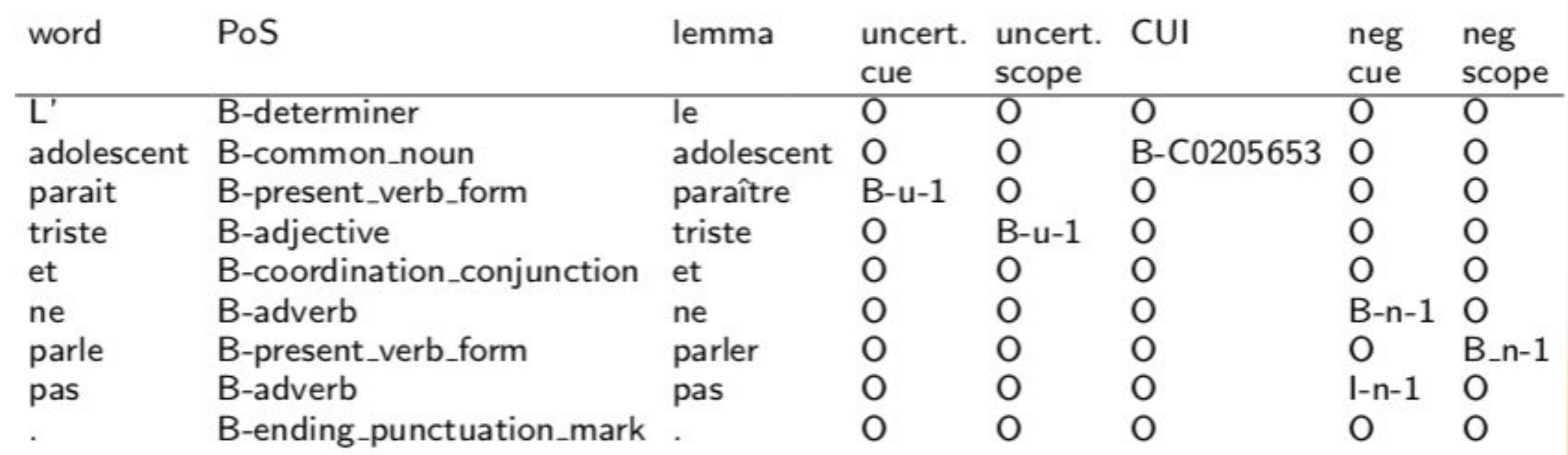
Clinical cases



Femme de 73 ans n'ayant eu qu'un seul enfant par césarienne, mais présentant depuis plusieurs années un prolapsus de stade III totalement négligé par la patiente. Elle est en insuffisance rénale obstructive avec une urée sanguine à 10 mmol/l de sérum. Sur l'urographie intraveineuse, on note une dilatation urétéropyélocalicielle bilatérale très importante. La tension artérielle est de 12/8. La mise en place d'un pessaire améliore très rapidement la situation puisque quatre jours plus tard, l'urée sanguine est à 6,4 mmol/l. La patiente refuse tout geste chirurgical complémentaire et elle est ensuite perdue de vue.

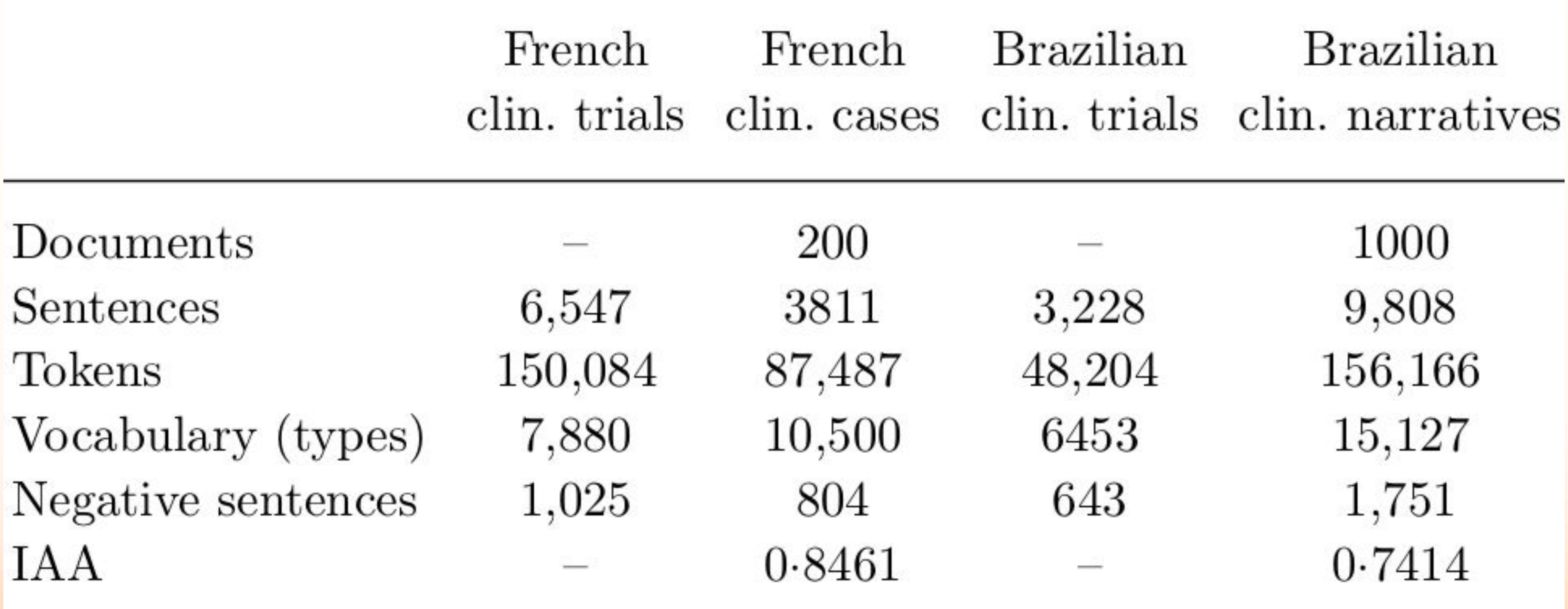
Annotated corpus

Semi-automatically annotated (with our NLP tools)



word	PoS	lemma	uncert. cue	uncert. scope	CUI	neg cue	neg scope
L'	B-determiner	le	O	O	O	O	O
adolescent	B-common_noun	adolescent	O	O	B-C0205653	O	O
parait	B-present_verb_form	paraître	B-u-1	O	O	O	O
triste	B-adjective	triste	O	B-u-1	O	O	O
et	B-coordination_conjunction	et	O	O	O	O	O
ne	B-adverb	ne	O	O	O	B-n-1	O
parle	B-present_verb_form	parler	O	O	O	O	B_n-1
pas	B-adverb	pas	O	O	O	I-n-1	O
.	B-ending_punctuation_mark	.	O	O	O	O	O

Annotated corpus



	French clin. trials	French clin. cases	Brazilian clin. trials	Brazilian clin. narratives
Documents	–	200	–	1000
Sentences	6,547	3811	3,228	9,808
Tokens	150,084	87,487	48,204	156,166
Vocabulary (types)	7,880	10,500	6453	15,127
Negative sentences	1,025	804	643	1,751
IAA	–	0·8461	–	0·7414

Social Media 4 Health

Pharmacovigilance on Twitter

1. automatic classification of tweets mentioning a drug name
2. automatic classification of tweets containing reports of first-person medication intake
3. automatic classification of tweets presenting self-reports of adverse drug reaction (ADR)
4. automatic classification of vaccine behavior mentions in tweets.

Neural networks (Bi-LSTM) with a bagging inspired re-sampling technique

N2C2: NLP for Clinical Data challenge

Cohort selection for clinical trials

- decide if a patient (document) satisfies each of the 13 inclusion criteria
  - treatments, medication, disorders, health problems, social characteristics, behaviour, lab results...
- 13 classes: for each patient and each criteria
- 300 sets of longitudinal patient records, annotated by medical professionals

Framework

- features: identification of medical concepts (CUIs)
- machine learning: Naïve Bayes, SVM, Logistic regression, Neural Networks
- rule based: regex, hand-encoded patterns for classes with few or no examples

DeFT 2019 - Clinical text-mining challenge & workshop

Tasks

- index clinical cases
  - identify keywords corresponding to a clinical case
- find expertise
  - pair a clinical case with a discussion
- extract information
  - age and gender of the patient
  - outcome (cured, improving, stable, worsening, death)
  - origine: part of text indicating why the patient is hospitalized

- Our approaches: neural approaches and information retrieval
- 9 participants (academics and industrial) / +30 participants at the workshop

Numerical information

On clinical trial protocols

Absolute neutrophil count 1,000 cells/l at time of enrollment .

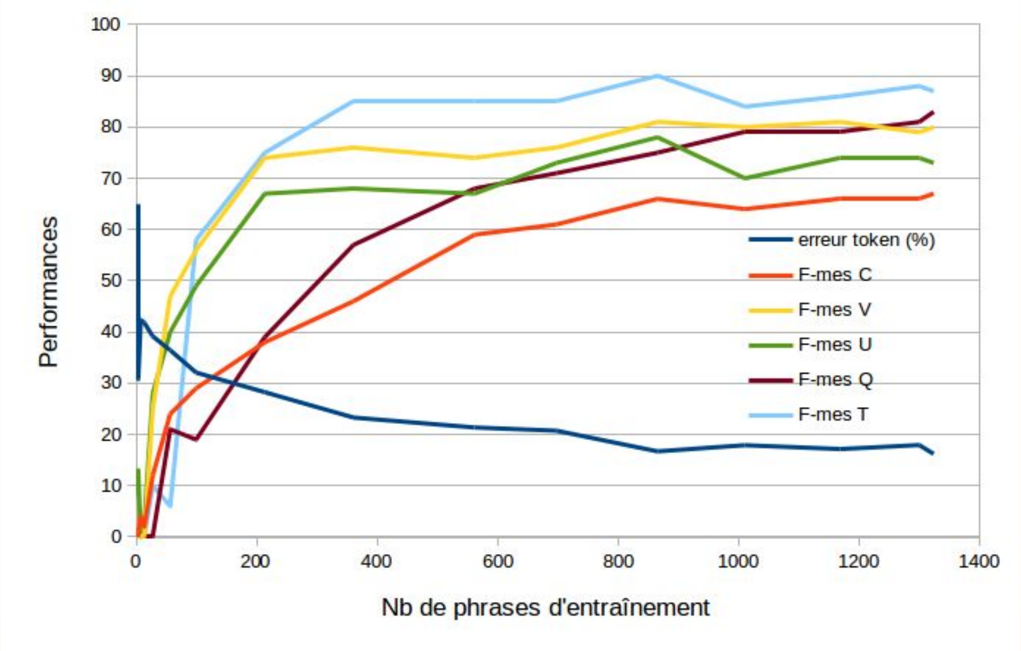
Exclude if T3 uptake is less than 19% ; T4 less than 2.9 (g/dL) ; free T4 index is less than 0.8 .

Elevated bilirubin within the past two years

🟧 measured concept
🟦 value
🟪 unit
🟦 quantifier
🟩 temporal validity

Numerical information

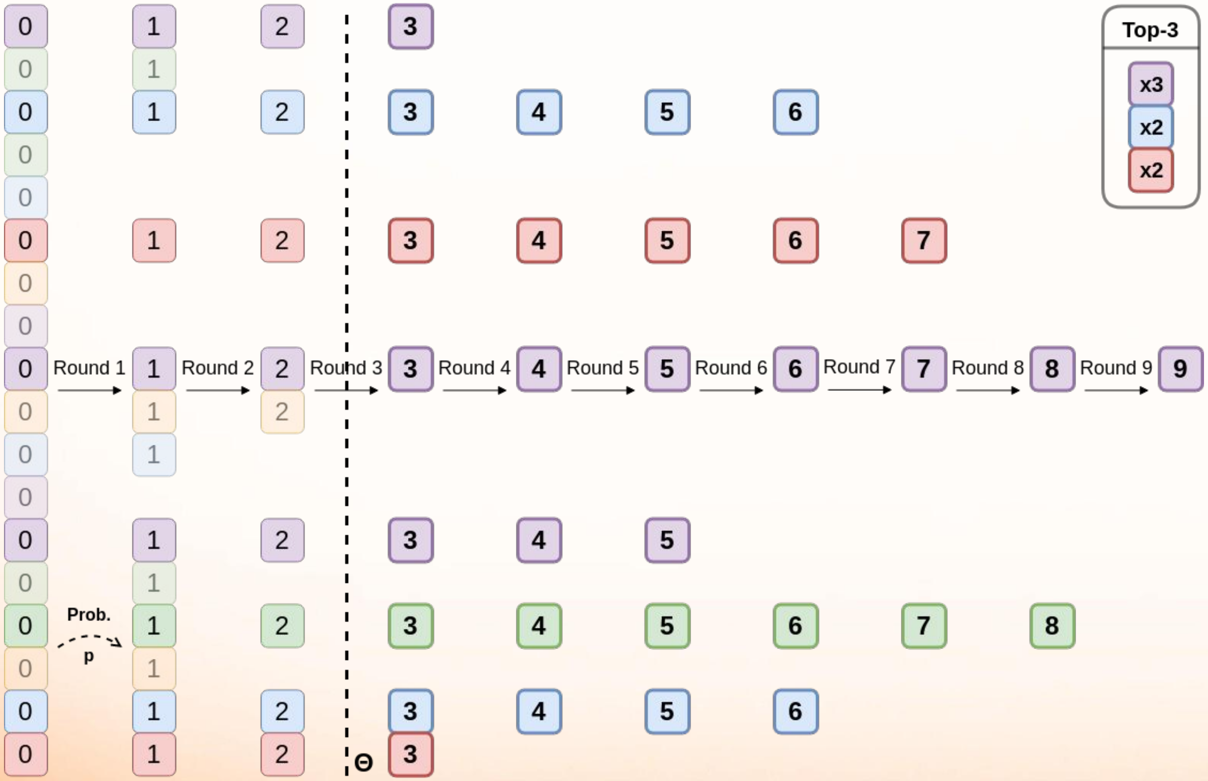
Results:

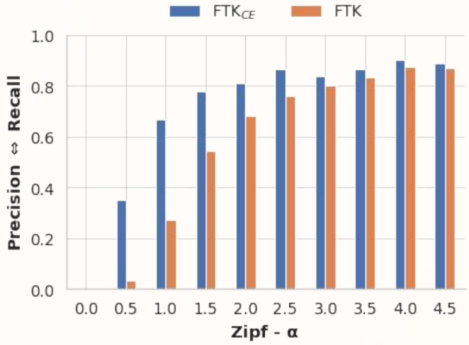


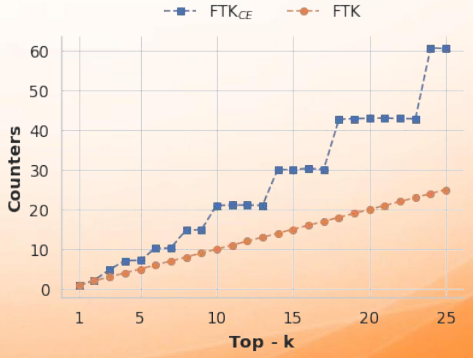
Normalization of units:

pack of cigarette per day → **cigarette per day** ; cigarette a day ; cigarette daily ; cigarette or equivalent per day ; cigarette/day ; pack per day ; cig/day ; pack/year ; pipe per day ; pack year ; cigarette per week

Top-k algorithm







Difficulties, weaknesses

Difficulties

- Recruitment
  - no recruitment for a 12-month post-doctoral position
  - PhD hired with a 3 month delay due to FSD clearance
- GDPR constraints for accessing to clinical patient data
  - The Clinical Data Center of CHU Rennes made available a massive dataset of real data and an infrastructure for distributed computing and deep learning on GPUs in mid 2018
  - does not allow reproducibility, difficult to supervise, limited computing power
  - partial solution: development of new corpora for French (clinical cases), work with English and Brazilian clinical texts but quite limitative for cost and clinical expertise reasons
- Need for more interaction
  - Developing efficient NLP tools require strong interdisciplinary skills which were hard to mobilize during the project. Further work on real data could have been made inside Clinical Data Center with more more NLP-skilled people.

instructions

- contributions scientifiques majeures obtenues

- éventuelles difficultés

- publications, logiciels et brevets

- devenir des doctorants, post-doctorants et ingénieurs CDD recrutés par le projet

- possibilités d'exploitation des résultats

Remarques

- séance publique ~15mn ; réunion privée 2h
- expliquer/exemplifier sigles : PMSI / ICD10...
- postdoc -> rendre ça plus positif ou ne pas en parler
- pb d'accès aux données -> rendre ça plus positif, expliquer comment on a rebondi
- améliorer le croisement des équipes (publis communes et aussi dans la façon de présenter le projet)
- suites: citer des initiatives possibles, Health Data Hub, MII en Allemagne
- Clément : insister sur les verrous scientifiques
- Vasile : faire le lien avec BigClin : veille sanitaire avec ICD-10, twitter pour la grippe, données *omics
- prolongation partielle pour le postdoc :