Responsible NLP in the making: contributions from ethics and reproducibility

Aurélie Névéol

LISN (formerly, LIMSI) CNRS

HealTAC, June 18 2021

Acknowledgements

- In 2020, ACL adopted the ACM code of ethics
  - I was a member of the EMNLP 2020, NAACL 2021 and ACL 2021 ethics review committees
- Long standing interest in reproducibility in clinical NLP
  - Organisation of shared tasks
  - Survey and literature studies for a better understanding
- (lack of) accessibility of clinical corpus in French

Bender EM. Academic freedom, academic integrity, and ethical review in NLP. Medium blog post 2021

Bender EM, Fort K. NAACL Ethics Review Process Report-Back. ACL report 2021.

Ethical considerations

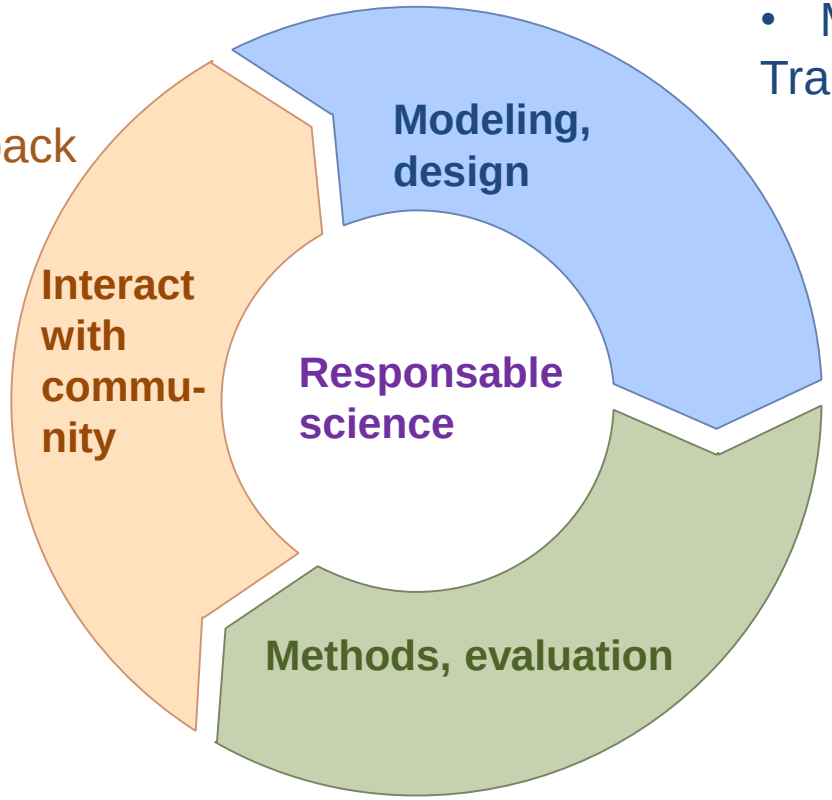
Responsible science:

gatekeeping vs. **benevolent guidance**

BioNLP research process

- Task definition
- Modeling in corpus: Train/test data

- Application
- Community feedback



- Design, apply NLP methods
- Evaluate using appropriate metrics

- Consider Ethics
- Ensure reproducibility
- Evaluate impact

Ethics is a
nuanced discussion of (at least)
three aspects of research problems
intended to help us strive for
better research

Is the problem meaningful and well designed?

Experiment:

Use a cloud hosted language model to test GPT3's ability to provide:

- Admin chat (e.g. appointment taking) with patient
- Mental health support
- Medical diagnosis
- …

Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. New England Journal of Medicine. 2018;378(11):981-983..

Human subjects

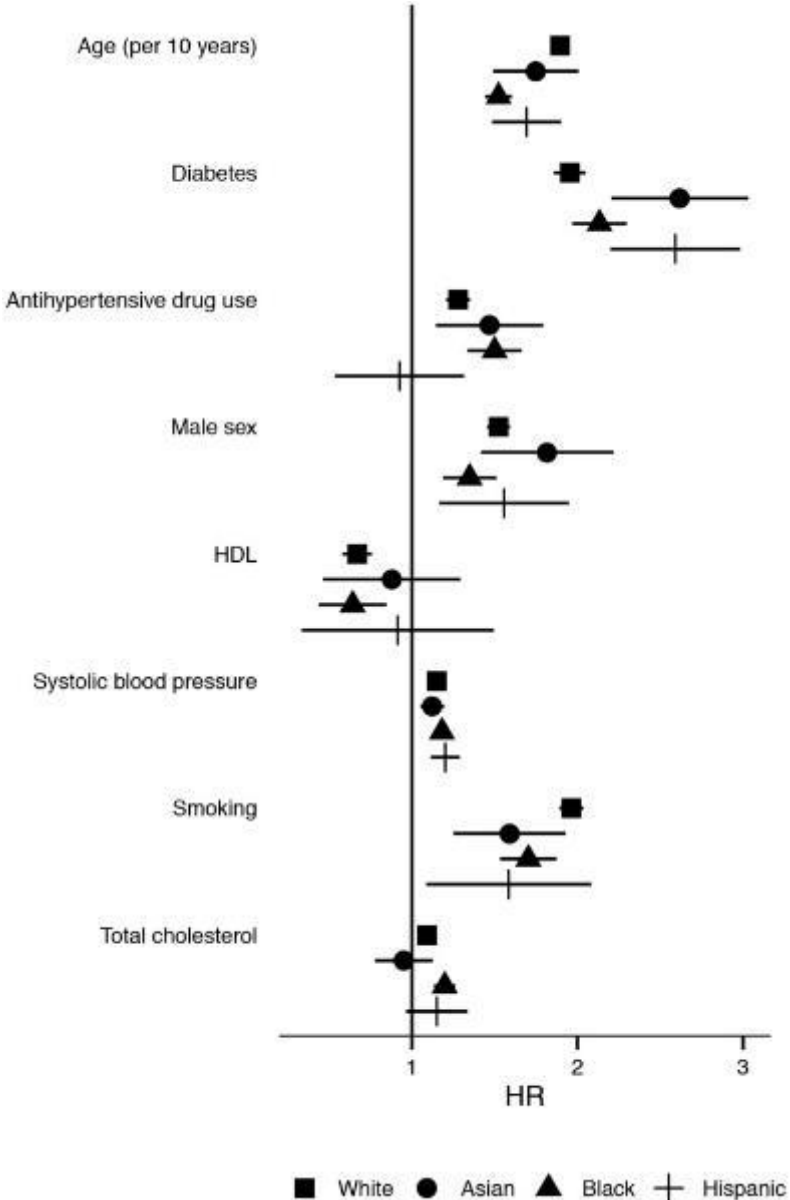
- When and how to interact with ethics committees, IRB?
- Language involves humans
  - Corpus sources → privacy, potential harm, including essentializing identity characteristics
  - Corpus processing → fair treatment of research participants
  - Corpus downstream users → direct and indirect

Larson B Gender as a Variable in Natural-Language Processing: Ethical Considerations. Proc. ACL Workshop on Ethics in Natural Language Processing 1-11. (2017).

Fort K, Adda G, Cohen KB. Amazon Mechanical Turk: Gold Mine or Coal Mine? Computational Linguistics, 37(2):413-420 (2011).

How are data and methods designed?

Bias: imbalanced data, diverging data processing, bias amplification



Gijsberts CM, Groenewegen KA, Hoefer IE, et al. Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. PLoS One. 2015; 10(7):e0132321

Lwowski B, Rios A. The risk of racial bias while tracking influenza-related content on social media using machine learning. J Am Med Inform Assoc. 2021 Mar 18;28(4):839-849. doi: 10.1093/jamia/ocaa326.

Datasets and corpus development

- Provenance and availability
- Terms of use, including confidentiality, copyrights
  - Some information is always sensitive (e.g. health, religion)
- Detailed description
  - Language, volume
  - Selection and collection method
  - Quality assessment, including biases

Bender E and Friedman B Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational Linguistics 6:587-604. (2018).

Impact of data on evaluation?

- Similarity between training and test corpus
  - 4 biomedical English benchmark datasets
  - Compare performance in redundant vs. non redundant scenarios
- Characterization of memorization vs. generalization
  - What is realistic in a real-life setting?

Elangovan A, He J, Verspoor K. Memorization vs. Generalization : Quantifying Data Leakage in NLP Performance Evaluation. Proc. EACL 2021: 1325-1335

What is the impact of deployment?

Dr Murphy (aka David Watkins)
@DrMurphy11

Is this another negligent #Triage from the @babylonhealth #GPatHand #AI #Chatbot App?

48yr old obese 30/day male smoker develops sudden onset central chest pain & sweating....

I say call 999, the Babylon App says see your GP...

Environmental impact

Consumption	CO_2e (lbs)
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

Underestimation not accounting for life cycle of computer equipment

Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL) 3645–3650. (2019).

Environmental impact

- Reporting of computational resource use
  - Carbon tracker https://github.com/lfwa/carbontracker
  - Green Algorithms http://www.green-algorithms.org/

- Benefit/risk analysis beyond leaderboard performance

Ethayarajh K and Jurafsky D. Utility is in the Eye of the User: A Critique of NLP Leaderboards. Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP) 4846-53. (2020).

Reproducibility

Challenges in Reproducibility

- Reports of a reproducibility crisis in many disciplines
  - Poll of 1,500 scientists (2016) 225 NLP researchers (2019)

Discipline	Failed to reproduce others' experiment	Failed to reproduce own experiment
Chemistry	90%	60%
Biology	80%	60%
Physics and engineering	70%	50%
Medicine	70%	60%
Earth and environment science	60%	40%
Other	60%	50%
Natural language processing	60%	30%

Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016 May 25;533(7604):452-4.

Mieskes M, Fort K, Névéol A, Grouin C, Cohen KB. NLP Community Perspectives on Replicability. Proc. RANLP. 2019:768–775.

Types of reproducibility and gains

	Hardware	Software/parameters	Data/method	Gain
Repeat	✓	✓	✓	determinism
Replicate	✗	✓	✓	robustness
Reproduce	✗	✗	✓	portability
Reuse	✗	✗	✗	generalizability

Cohen-Boulakia S, Belhajjame K, Collin O, et al.Scientific workflows for computational reproducibility in the life sciences: status, challenges and opportunities. Future Gen Comput Syst 2017; 75: 284–98.

Shared tasks foster reproducibility

- Primary goal is to provide a forum for direct comparison of approaches

- Availability of shared material
  - Specific definition of a “task”
  - Corpora and annotations, split into training, development and test sets
  - Evaluation metrics and scripts

A replicability study at CLEF eHealth 2016 set-up

- As part of the ICD10 coding task, participant could submit their system for replication
  - 4 analysts committed to replicate results in their usual work environement
  - 3 teams submitted systems

- Replication assessment
  - Scoring sheet documenting install/run/results
  - Timing

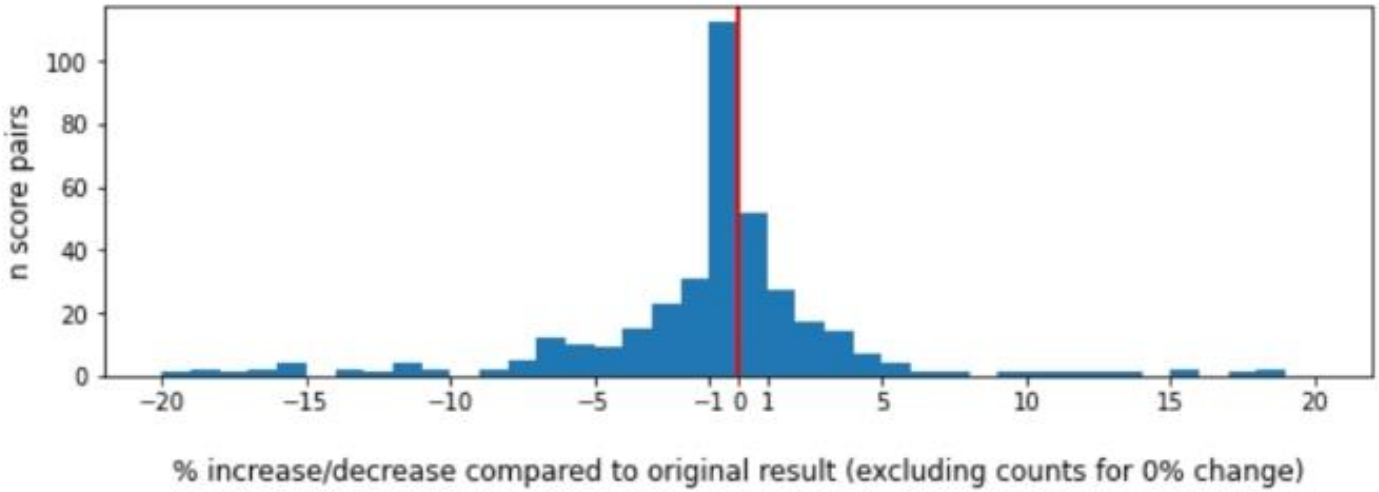
Névéol A, Cohen KB, Grouin C, Robert A. Replicability of Research in Biomedical Natural Language Processing: a pilot evaluation for a coding task. Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis, LOUHI. 2016:78-84

A replicability study at CLEF eHealth 2016 results

- Results were replicated…
  - No single analyst was able to replicate all
  - Time to replication varied greatly
- But replication is not trivial!
- Replication requires resources
  - For authors to produce quality systems, documentation
  - For others to understand and conduct

Névéol A, Cohen KB, Grouin C, Robert A. Replicability of Research in Biomedical Natural Language Processing: a pilot evaluation for a coding task. Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis, LOUHI. 2016:78-84

A broader look on NLP reproducibility



- Data and code are still elusive

Belz A, Agarwal S, Shimorina A, Reiter E. A Systematic Review of Reproducibility Research in Natural Language Processing. EACL 2021:381–393

Mieskes M. A quantitative study of data in the NLP community. Proc ACL Workshop on Ethics in NLP. 2017

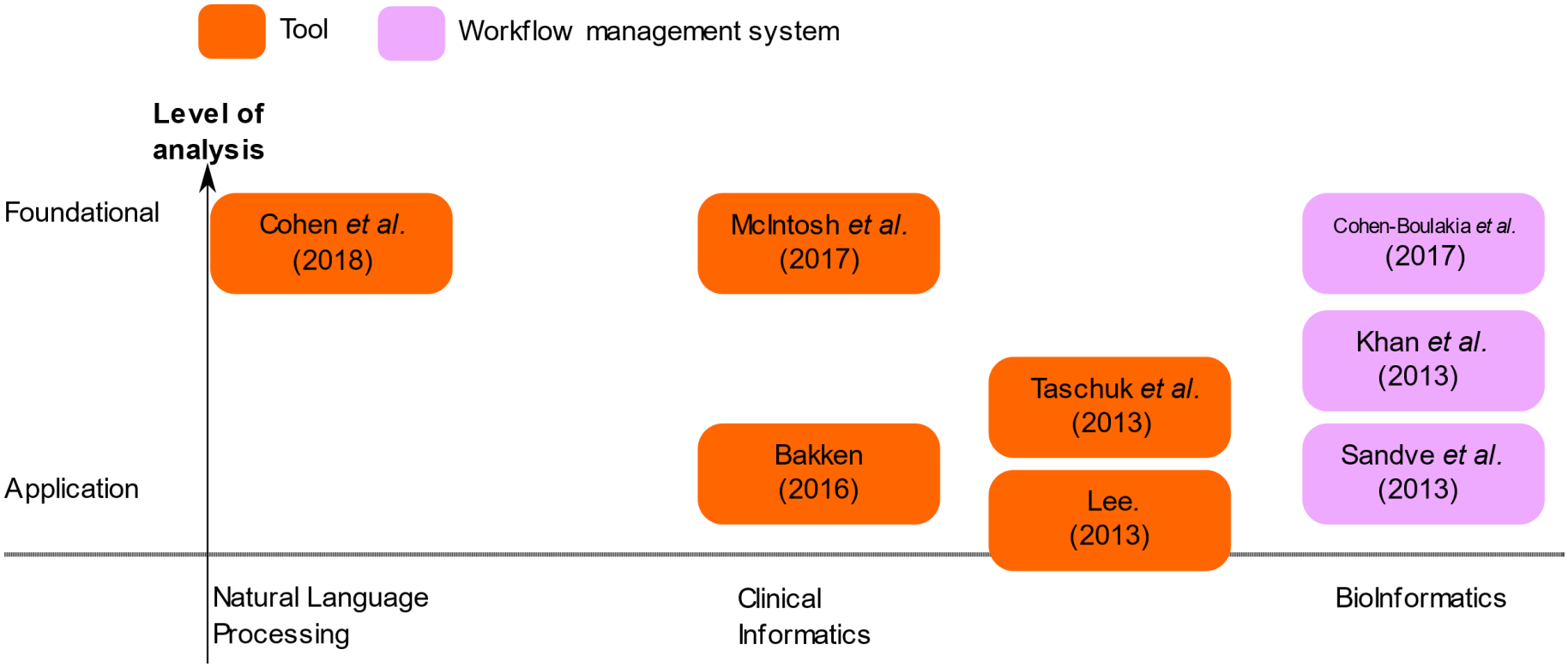
Towards actionable reproducibility

- From research to hospital operations
  - Need for standardization, traceability, automation
- Leveraging expertise and experience accross disciplines
  - Reproducibility criteria/desiderata expressed by the bioinformatics, medical informatics, NLP communities
- Characterize clinical NLP w. r. t. reproducibility
  - Analysis of 7 clinical NLP systems (for English)

Digan W, Névéol A, Neuraz A, Wack M, Baudoin B, Burgun A, Rance B. Can reproducibility be improved in clinical natural language processing? A study of 7 clinical NLP suites. J Am Med Inform Assoc. 2021. Mar 1;28(3):504-515.

Literature review to identify reproductibility criteria

- MEDLINE search and snowbowling

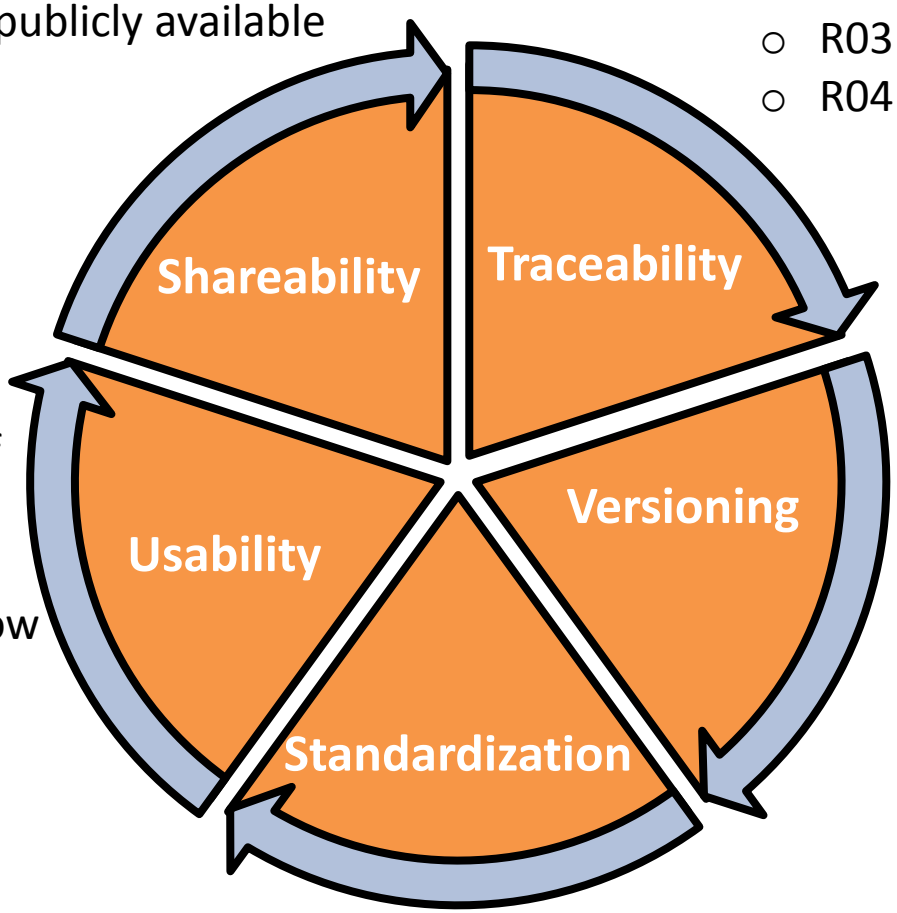


Digan W, Névéol A, Neuraz A, Wack M, Baudoin B, Burgun A, Rance B. Can reproducibility be improved in clinical natural language processing? A study of 7 clinical NLP suites. J Am Med Inform Assoc. 2021.

40 reproducibility criteria

- R39 Input data publicly available
- R40 Resources publicly available

- R01 Provenance Metadata
- R03 System Metadata
- R04 Record Parameters



- R28 Absence of manual steps
- R30 Ability to resume workflow

- R06 Pipeline versioning
- R07 Tool versioning
- R08 Resource versioning

Digan W, Névéol A, Neuraz A, Wack M, Baudoin B, Burgun A, Rance B. Can reproducibility be improved in clinical natural language processing? A study of 7 clinical NLP suites. J Am Med Inform Assoc. 2021.

Evaluation of clinical NLP systems

NLP systems

Relies on **UIMA or Gate**

- cTakes[1] 18/40
- CLAMP [2] 17/40
- GATE[3] 17/40

WMS Systems

Relies on **Galaxy**

- **LAPPGrid [4] 26/40**
- OpenMinTed [5] 22/40
- Textflows [6] 17 /40

NLP toolbox

python library

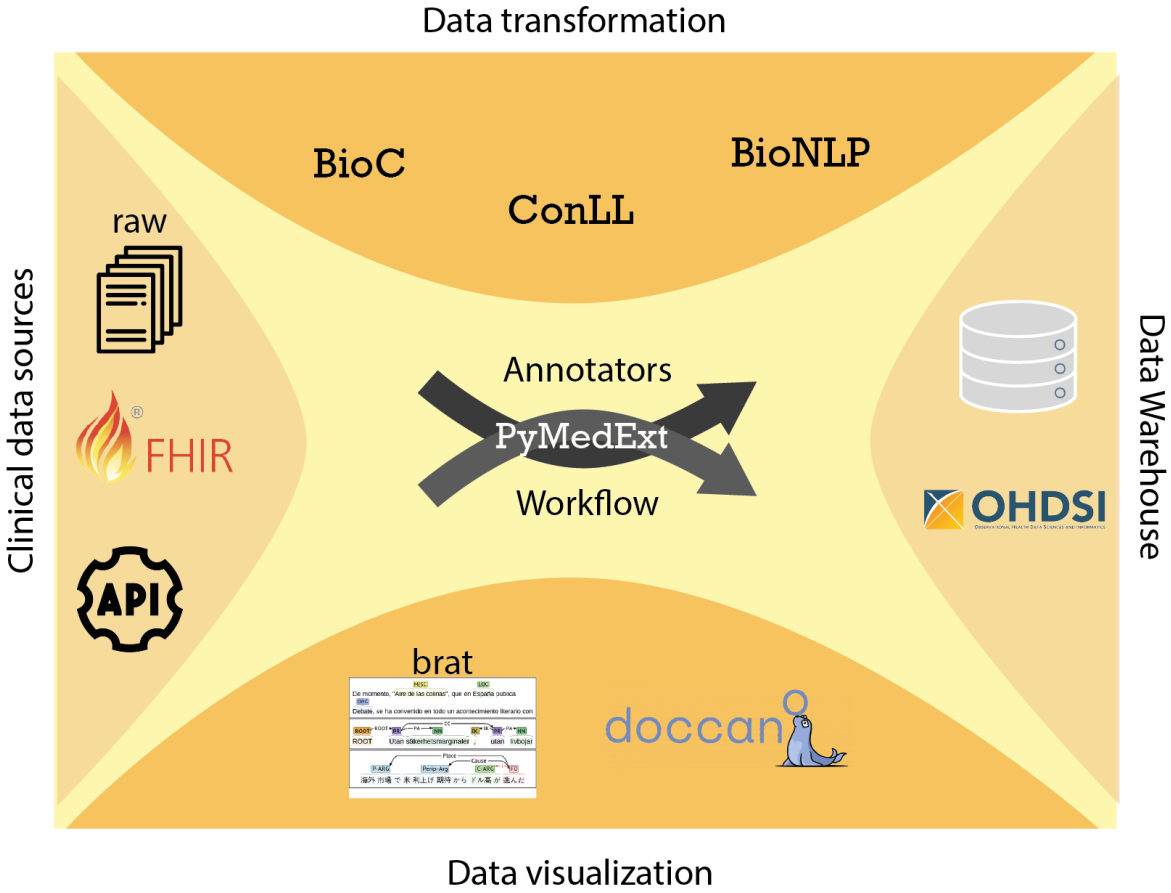
- ScispaCy [7] 17/40

- Reproducibility can be improved
  - Especially versioning, standardization and shareability
  - Experience from bioinformatics suggests modularity and workflows can help

Digan W, Névéol A, Neuraz A, Wack M, Baudoin B, Burgun A, Rance B. Can reproducibility be improved in clinical natural language processing? A study of 7 clinical NLP suites. J Am Med Inform Assoc. 2021.

Workflow management for French Clinical NLP?

Allows the implementation of a simple workflow integrating “annotators” and text format conversion



https://github.com/equipe22/pymedext_core

Shared clinical corpus
in languages other than English

Few corpora are available

- Death certificates
  - French, Hungarian, Italian [Névéol et al. 2018]
- Patient referals
  - Spanish [Báez et al. 2020]
- Creative solutions
  - Synthetic clinical narratives (Japanese [Aramaki et al. 2014], Norvegian [Rama et al. 2018])
  - Clinical case reports in French [Cardon et al. 2018] and Spanish [Miranda Escalada et al. 2020]

De-identification: a solved problem?

- Research stimulated by shared tasks
  - I2b2 2006, 2014, NGRID 2016
- Adressed in several languages
  - English, French, Swedish…
- Still different from anonymization

Synthetic corpus

- Successful attempts for English using neural models
  [Melamud and Shivade 2019, Ive et al. 2020]
  - Trained on MIMIC III
  - Keyphrases and ICD10 codes used as prompts
- Synthetic text visibly different from real corpus but beneficial as data augmentation for processing real data
- Is synthetic data anonymous?
  - Generative models can be tuned for privacy
  - Few rare n-grams retained from original corpus

Summary

- Ethics offers guidelines to question research
  - motives
  - methods and data (including evaluation)
  - impact
- Reproducibility is complex and hard to achieve
  - But we have to keep trying!
  - And document…
- We need to broaden the scope of shareable clinical corpus

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

cnrs

Thank you!

université PARIS-SACLAY