



Hands-on Natural Language Processing for Clinical Informatics in Python

Hannah Eyre, Alec Chapman, Kelly Peterson, Patrick Alba, and
Scott DuVall

University of Utah and US Department of Veterans Affairs





"We are time's subjects, and time bids be gone."

- Shakespeare, *King Henry IV, Part II*

Link to training materials
and additional information:





Outline

- Why we developed medspaCy
- MedspaCy demo
- The changing NLP landscape
- Ideas (and experimental work) for incorporating new NLP methods and medspaCy
 - Few-shot learning for emerging events
 - Maintaining legacy rule-based NLP
 - Causal analysis using rule-based NLP



Patrick Alba



Olga Patterson



Mengke Hu



Alec Chapman



Hannah Eyre



Kelly Peterson



Scott Duvall



John Stanley



Annie Bowles



Qiwei Gan



Jianlin Shi



U.S. Department of Veterans Affairs

Epidemiology

Clinical Informatics

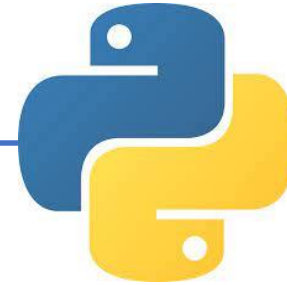
Data Science



Timeline of Clinical NLP Software



MedLEE



2020



SymText

MetaMap

LEO



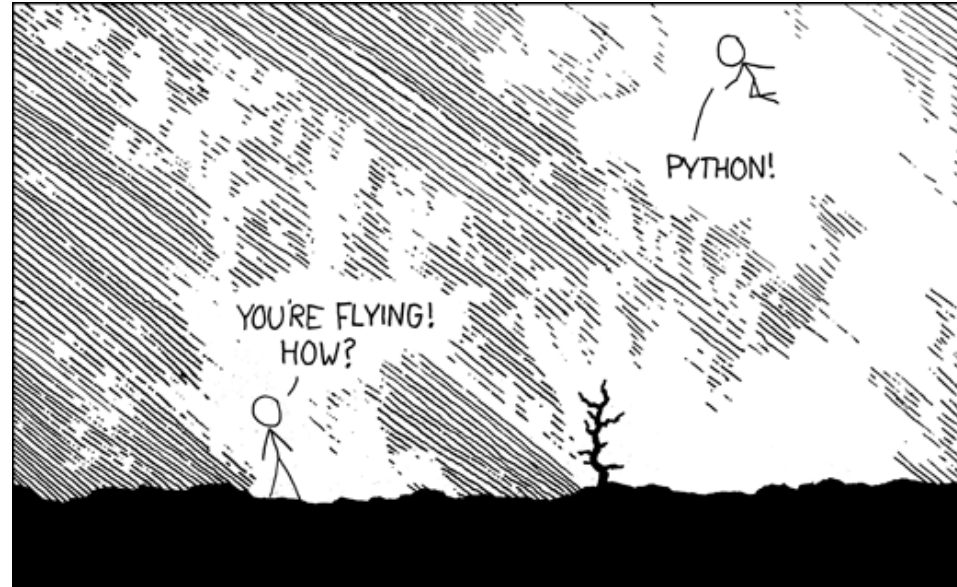
cTAKES

Unstructured
Information
Management
Architecture
An Apache Project



The magic of Python

- More accessible, easy to learn
- Making programming more accessible...





... to clinicians



Brian Bucher,
surgery

Rashmee Shah,
cardiology



Lori Gawron,
OB/GYN

Andrew Gawron,
gastroenterology





... and new programmers

Alec Chapman,
c. 2014

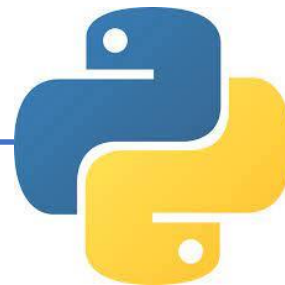


Brian Chapman teaching Python to
Alec Chapman, c. 2016



MEDINFO23

8 - 12 JULY 2023 | SYDNEY, AUSTRALIA



2020



TensorFlow

spaCy



What happened in 2020?

- Needed rapid, easy-to-implement clinical NLP software
- Used by VA/DoD researchers but be available to community at large





Introducing medspaCy

- Built on spaCy's extensible framework
- Designed to be interoperable with any other spaCy component
- Enabled integration of modern Data Science and NLP packages into traditional Clinical NLP workflows





Hands-on tutorial

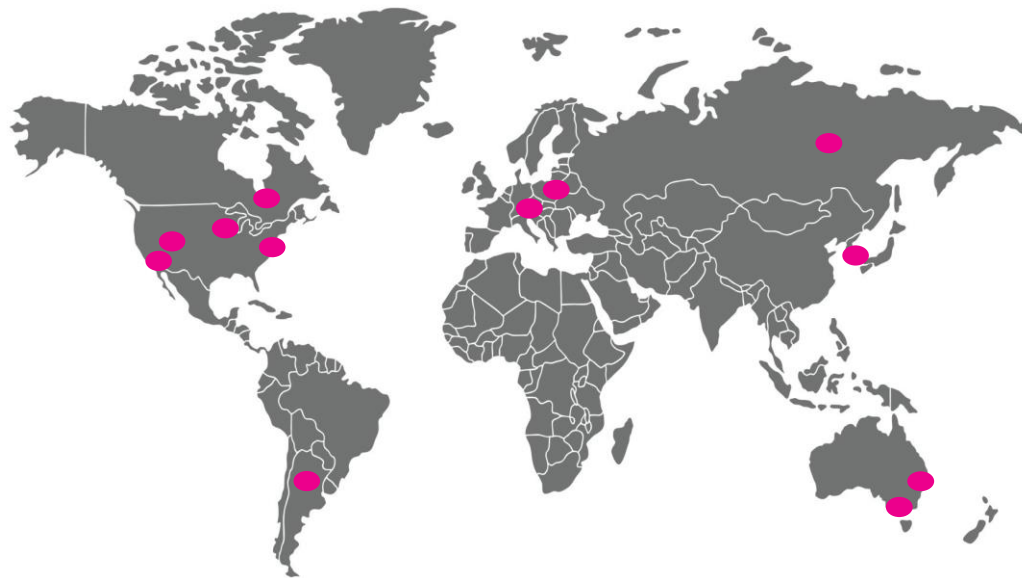
Follow along on
Google Colab or
your local machine



https://github.com/medspacy/medspaCy_MedInfo_2023



Impact



Covid-19

- Deployed in [VA biosurveillance system](#)
- Used in VA [operations](#) and [research studies](#)
- Parsing signs and symptoms of Covid-19 patients
- Surveillance system [replicated](#) in Toronto, Canada

A Natural Language Processing System for National COVID-19 Surveillance in the US Department of Veterans Affairs

Alec B Chapman^{1,2}, Kelly S Peterson^{1,2}, Augie Turano³, Tamara L Box⁴,
Katherine S Wallace⁵, Makoto Jones^{1,2}

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Comparative Effectiveness of BNT162b2 and mRNA-1273 Vaccines in U.S. Veterans

Barbra A. Dickerman, Ph.D., Hanna Gerlovin, Ph.D.,
Arin L. Madenci, M.D., Ph.D., Katherine E. Kurgansky, M.P.H.,
Brian R. Ferolito, M.Sc., Michael J. Figueroa Muñiz, B.Sc.,
David R. Gagnon, M.D., Ph.D., M.P.H., J. Michael Gaziano, M.D., M.P.H.,
Kelly Cho, Ph.D., Juan P. Casas, M.D., Ph.D., and
Miguel A. Hernán, M.D., Dr.P.H.

RESEARCH ARTICLE

Using Primary Care Clinical Text Data and Natural Language Processing to Identify Indicators of COVID-19 in Toronto, Canada

Christopher Meaney^{1*}, Rahim Moineddin¹, Sumeet Kalia¹, Babak Allarzadeh¹,
Michelle Greiver^{1,2}

¹ Department of Family and Community Medicine, Faculty of Medicine, University of Toronto, Toronto, Canada, ² North York Family Health Team, North York General Hospital, Toronto, Canada



Social Determinants of Health

- Evaluation of [VA homelessness programs](#)
- [MIMIC-SBDH](#): Public dataset for SDoH
- [Hanna Pethani](#) (University of Sydney): extracting SDoH from dental notes to study impact on dental health

MIMIC-SBDH: A Dataset for Social and Behavioral Determinants of Health

Hiba Ahsan

Emmie Ohnuki

Avijit Mitra

College of Information and Computer Sciences, University of Massachusetts, Amherst, MA, USA

Hong Yu

College of Information and Computer Sciences, University of Massachusetts, Amherst, MA, USA

Department of Computer Science, University of Massachusetts, Lowell, MA, USA

Center for Healthcare Organization & Implementation Research, Bedford, MA, USA

Department of Medicine, University of Massachusetts Medical School, Worcester, MA, USA

HAHSAN@UMASS.EDU

EOHNUKI@UMASS.EDU

AVIJITMITRA@UMASS.EDU

College of Information and Computer Sciences, University of Massachusetts, Amherst, MA, USA

HONG.YU@UMASSMED.EDU

Original Research

ReHouSED: A novel measurement of Veteran housing stability using natural language processing

Alec B Chapman^{a,b,*}, Audrey Jones^{a,b}, A Taylor Kelley^{a,c}, Barbara Jones^{a,d}, Lori Gawron^{a,e}, Ann Elizabeth Montgomery^{f,g,h}, Thomas Byrne^{h,i,j}, Ying Suo^{a,b}, James Cook^{a,b}, Warren Pettey^{a,b}, Kelly Peterson^{a,b,k}, Makoto Jones^{a,b}, Richard Nelson^{a,b,h}





...and more!

- Georgina Kennedy (Maridulu Budyari Gumal):
Extracting ECOG performance status
- [BioNLP](#)
- More examples and references in our [MedInfo repository](#)

75yo M, **ECOG 0** **ECOG_STATUS** SOCIAL HISTORY: << SOCIAL >>

Lives alone usually, (synthetic text)

Widowed, 2 daughters

Current **CURRENT** **ECOG 1** **ECOG_STATUS**, ET 200m, independent self-care

ECOG 2-3 **ECOG_STATUS**, last **ECOG 0-1** **ECOG_STATUS** over a year **BGO** **HISTORICAL** OE << EXAM >> SaO2 97%RA, PR 80 SR

hyperinflated lung fields, no creps or wheeze

68yroM **ECOG 1** **ECOG_STATUS** L OPX T2N2M0 (IVA) referred for ongoing management. PMH: << HISTORY >> Cardiac atherosclerosis on angiogram, hypercholesterolemia SHx: << SOCIAL >> Taxi driver, lives with wife, ex-smoker (40pyh) quit 12yros ago O/E: << EXAM >> Appears comfortable when resting. **ECOG 0** **ECOG_STATUS**, palpable L submandibular LN, no other palpable locoregional nodes, minimal decreased sensation.

Impression: << EXAM >> no fevers, some lethargy since hospital admission, **ECOG2** **ECOG_STATUS** with **premorbid** **HISTORICAL** **ECOG 0-1** **ECOG_STATUS**

**UMASS_BioNLP at MEDIQA-Chat 2023:
Can LLMs generate high-quality synthetic note-oriented doctor-patient conversations?**

Junda Wang * Zonghai Yao * Avijit Mitra Samuel Osebe Zhichao Yang Hong Yu

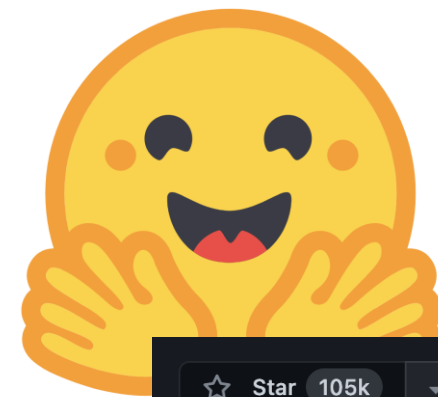
CICS, University of Massachusetts, Amherst, MA, USA

jundawang@umass.edu zonghaiyao@umass.edu

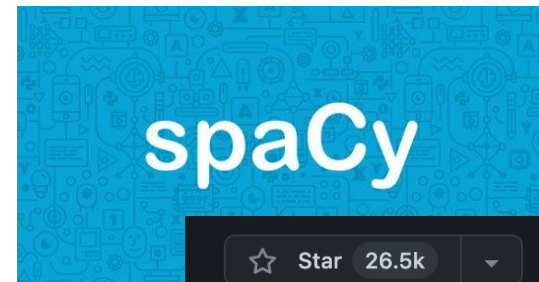


Changes in NLP Today

- New model architectures (Transformers) substantially improve performance on all NLP tasks
 - BERT, GPT-*, Megatron, etc.
- Libraries like Huggingface Transformers adopted
 - Huggingface Transformers more popular than spaCy since Oct. 2020



☆ Star 105k




☆ Star 26.5k



Changes in NLP Today

- Rapidly increasing capabilities AND size
 - Training capacity both more accessible than ever and extreme capacity is siloed
- Reframing old tasks for new transformers
- Deep-learning using pre-trained transformers is often a feasible first option



16GB nVIDIA Tesla V100 PCI Express 3.0 x16 GPU 699-2G500-0200-310
Brand New · NVIDIA

\$2,363.26
or Best Offer
Free shipping
Free returns

Top Rated Plus

NVIDIA DEVELOPER Home Blog Forums Docs Downloads Training

Technical Blog Search blog Filter

Conversational AI / NLP

Scaling Language Model Training to a Trillion Parameters Using Megatron

Apr 12, 2021 +3 Like

By Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand and Bryan Catanzaro



Who still cares?

- Compute is still scarce for many projects and organizations
 - Compute at state-of-the-art scale is extremely expensive
- Privacy concerns for cloud compute platforms and model APIs
- Data source for pre-trained transformers is often unknown



Who still cares?

- Transformers are best performing, but "best performing" is not always the metric for success
- Difference between .95 PPV and .85 PPV may not matter
- Annotated data is required for deep learning, not for rules
- Rules are explainable and debuggable



Future Directions and medspaCy's Role



Few-Shot Learning: Work in Progress

- Work-in-progress developing new capabilities for biosurveillance
- Many potential “early signs” for a public health event
 - One of these is animal exposure (i.e., zoonotic transmission)
- Difficult to annotate and develop NLP for emerging events



Poultry losses for the year from bird flu near record; the threat to humans isn't zero

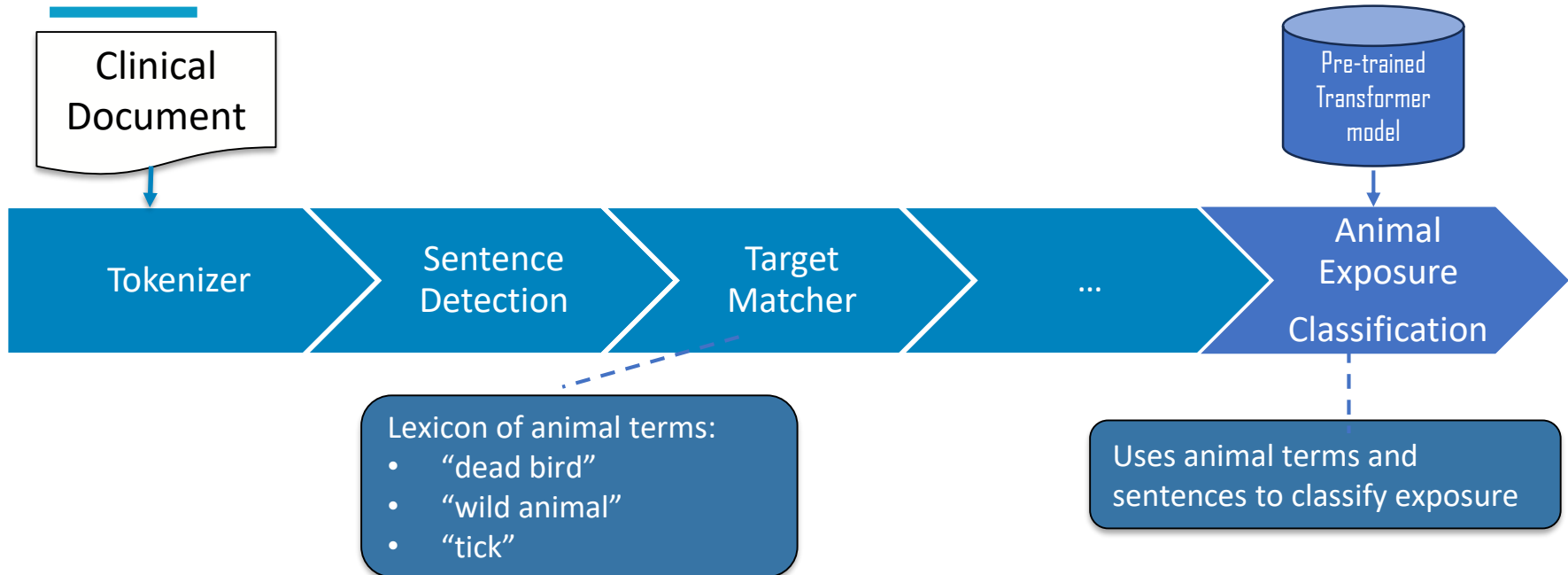


Few-Shot Learning: Work in Progress

- Simply identifying animal keywords not sufficient for true exposure:
 - “Pt lives in Antelope Valley”
 - “Complains of chest pain that feels like a kick from a mule.”
 - “Her son has an appetite like a horse”
 - “Pt smokes 1 pack a day; primarily camels.”
 - “Precaution: During travel, avoid contact with wild animals (i.e., avian influenza)”
 - “Pt denies knowledge of any tick bites”
 - “Pt manages anxiety by watching wild animals from their window”



Few-Shot Learning: Work in Progress





Few Shot Learning: Work in Progress

- Started with rules-based approach, not very accurate
- Annotated a few instances for Few-Shot Learning
- Leverages pre-built Transformer models
 - “SetFit” uses SentenceTransformers
- Very encouraging results so far

Efficient Few-Shot Learning Without Prompts

Lewis Tunstall¹, Nils Reimers², Unso Eun Seo Jo¹, Luke Bates³,
Daniel Korat⁴, Moshe Wasserblat⁴, Oren Pereg⁴

¹Hugging Face ²cohere.ai

³Ubiquitous Knowledge Processing Lab, Technical University of Darmstadt

⁴Emergent AI Lab, Intel Labs

¹firstname@huggingface.com ²info@nils-reimers.de

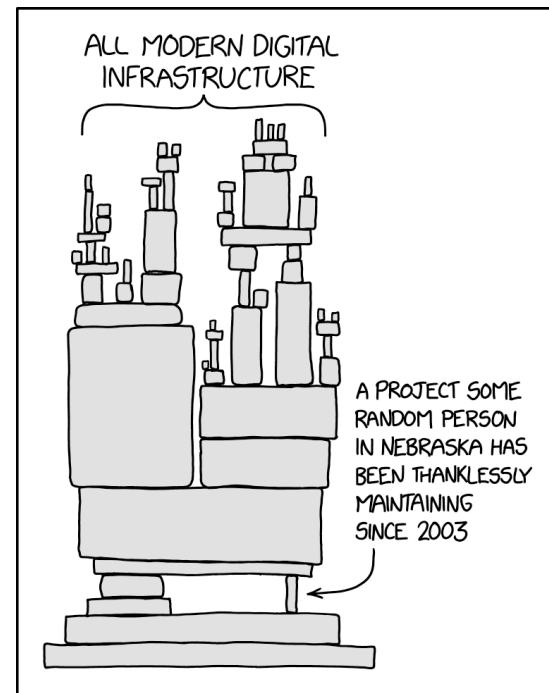
³bates@ukp.informatik.tu-darmstadt.de

⁴firstname.lastname@intel.com



Modernizing Legacy Rule-based NLP

- Operational datasets need regular updates
- Nobody wants to maintain legacy software
 - *"Hi all, is anybody still using UIMA-AS?"*
- UIMA developer to UIMA user listserv, 2022
- Rule-based NLP incorporates institutional knowledge that is not always available





Modernizing Legacy Rule-based NLP

- Sampled output from EchoExtractor
 - System is 6+ years old
 - Lots of VA research projects using output
 - Rules are complex, difficult to edit
- Extracting measures from echocardiograms
 - Frequency of labels vary widely from common (LVEF) to rare (mitral valve stenosis)

M-MODE MEASUREMENTS :

LV DIASTOLE : 50 (40–55mm)

LV SYSTOLE : 71 (25–30mm)

LT ATRIUM : 44 (25–35mm)

SOFTWARE

Open Access

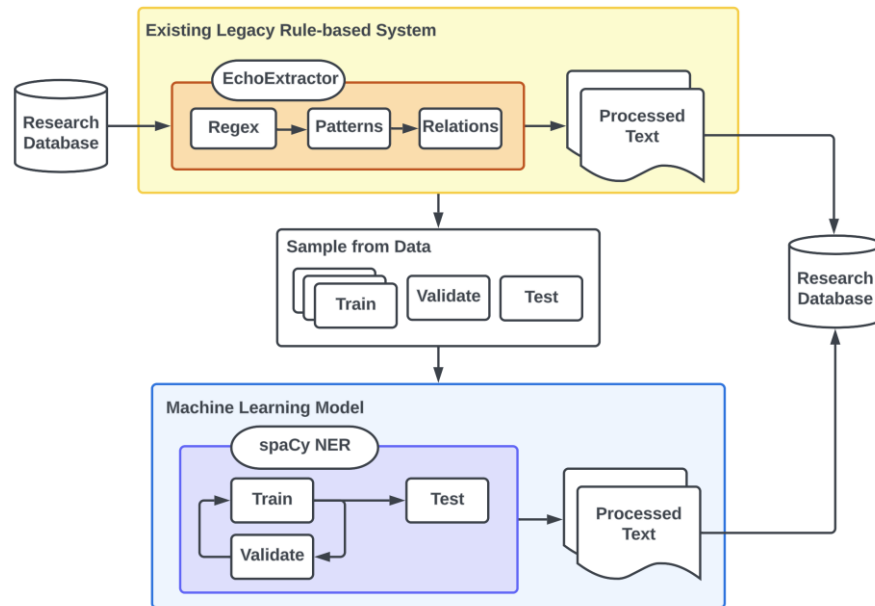
Unlocking echocardiogram measurements for heart disease research through natural language processing

Olga V. Patterson^{1,2*}, Matthew S. Freiberg^{3,4}, Melissa Skanderson⁵, Samah J. Fodeh⁶, Cynthia A. Brandt^{5,6} and Scott L. DuVal^{1,2}



Modernizing Legacy Rule-based NLP

- Used existing rule-based NLP output as training data
- Trained spaCy convolutional neural network (CNN)





Modernizing Legacy Rule-based NLP

- Frequent classes achieved $F1 > .90$
- Rare classes were inconsistent
 - Some were performant but others were never predicted
- Overall performance not there yet, but shows promise

Documents	PPV (Precision)	Sensitivity (Recall)	F1
1,000	88.2	56.9	69.2
10,000	92.9	59.3	72.4
100,000	94.4	69.2	79.8
670,000 (all)	95.8	94.7	95.2



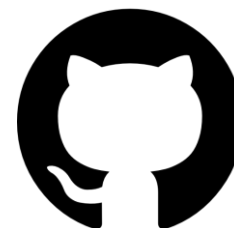
Modernizing Legacy Rule-based NLP

- Manual evaluation is needed
- Upcoming experiments using VA-trained BERT models
- Scaling inference speed of models is hard



Conclusion

- Exciting time to be in NLP
- Still room for traditional methods in clinical NLP
 - Alone and in combination with new methods
- Lots of open questions, problems to be solved





Thank You!

- Contact us:

- Hannah Eyre, hannah.eyre@utah.edu
- Alec Chapman, alec.chapman@hsc.utah.edu

- Collaboration

- We would love to help you use medspaCy in your work
- Open-source can always use contributors
- We are looking for multi-lingual rules!

