Breakout Session 6:

Scaling Up Literature Annotations with Cloud Computing in PubTator 3.0

Dr. Zhiyong Lu (Moderator),
Deputy Director for Literature Search, NCBI; Senior Investigator, NIH/NLM

Scaling up literature annotations with cloud computing in PubTator 3.0

Zhiyong Lu, PhD FACMI FIAHSI

NCBI, NLM, NIH

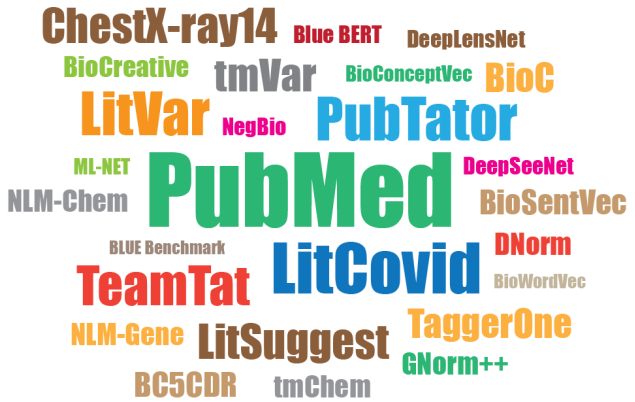
2024 NIH/ODSS Cloud Program PI Meeting

January 18, 2024

Our research at NLM IRP

- Research Areas
  - Machine Learning; Deep Learning
  - Natural Language Processing (NLP)
  - Medical Image Analysis

- Text & Image data
  - Biomedical Literature
  - Clinical notes, EMRs
  - CT, CXR & retinal images



- Application areas:
  - Literature Retrieval (e.g., PubMed Search; LitCovid)
  - Information Extraction/Curation (e.g., LitVar, PubTator, LitSuggest)
  - AI in Healthcare (e.g., machine diagnosis and prognosis)

Gene-disease-variant relations for genomic medicine

J Alzheimers Dis. 2012;32(2) Disease A Gene 2 Variant

Highly pathogenic Alzheimer's disease presenilin 1 P117R mutation causes a specific increase in p53 and p21 protein levels and cell cycle dysregulation in human lymphocytes.

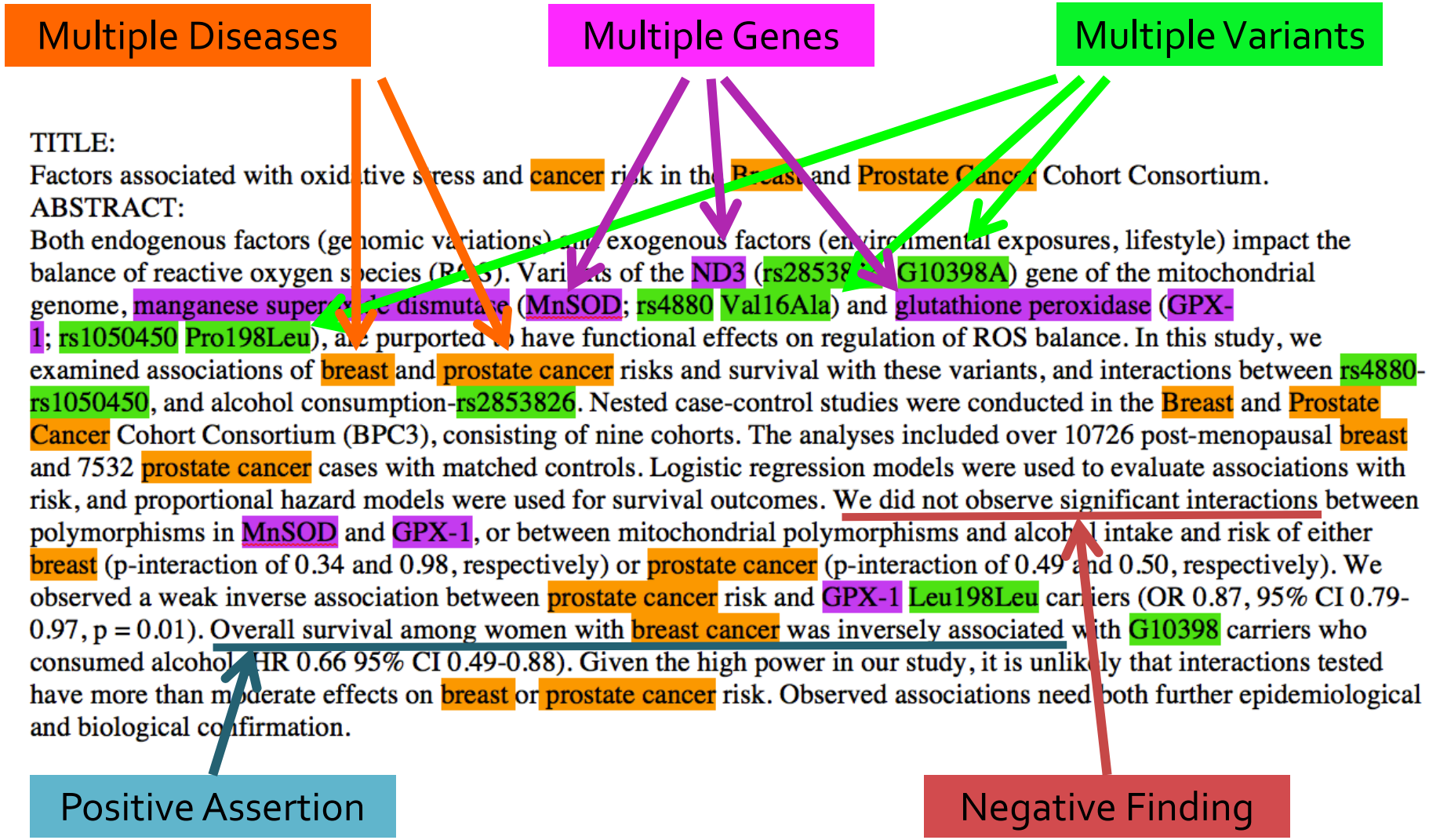
Bialopiotrowicz E[1], Szybinska A, Kuzniewska B, Buizza L, Uberti D, Kuznicki J, Wojda U.

Rank	Gene or Protein ID	Gene SYM	WTAA	MTAA	POS	Disease	PMIDs
1	Q13131	PRKAA1	Q	R	16	Breast cancer	16959974
2	P31749	AKT1	E	K	17	Breast cancer	17611497 18954143 19713527 21793738
3	P10275	AR	H	Y	874	Prostate cancer	17591767
	.....	....	....	....	....	....	.....

Automatic entity recognition is non-trivial

- Term variation
  - c.77A>C; c.77A->C; A77C;
- Synonymy
  - TP53; tumor protein p53; p53; BCC7; LFS1
- Ambiguity
  - Activated protein C (APC)
  - Antigen-presenting cell (APC)
  - Amino acid-polyamine-organocation (APC)
  - Anaphase-promoting complex (APC)
  - Argon plasma coagulation (APC)

Challenges in relation extraction



Our Entity Taggers

Disease

DNorm (2013) – 80.9%

Mutation

tmVar (2013) – 91.4%

tmVar 3.0 (2022) – 94.3%

Gene/Protein

GNormPlus – 86.7% (2015)
GNorm2 – 89.4% (2023)

Species

SR4GN (2012) – 85.4%

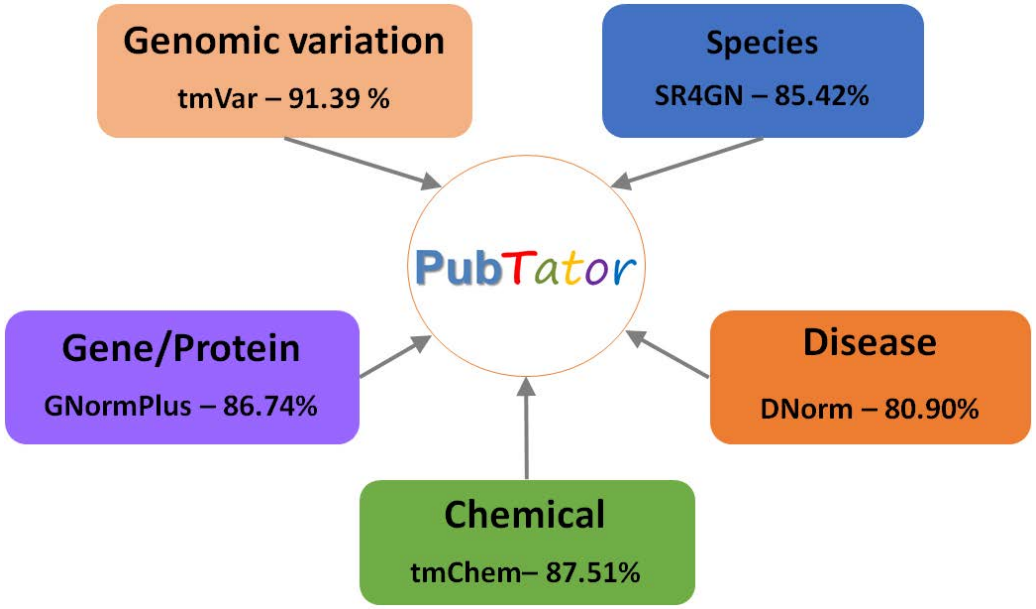
Chemical/Drug

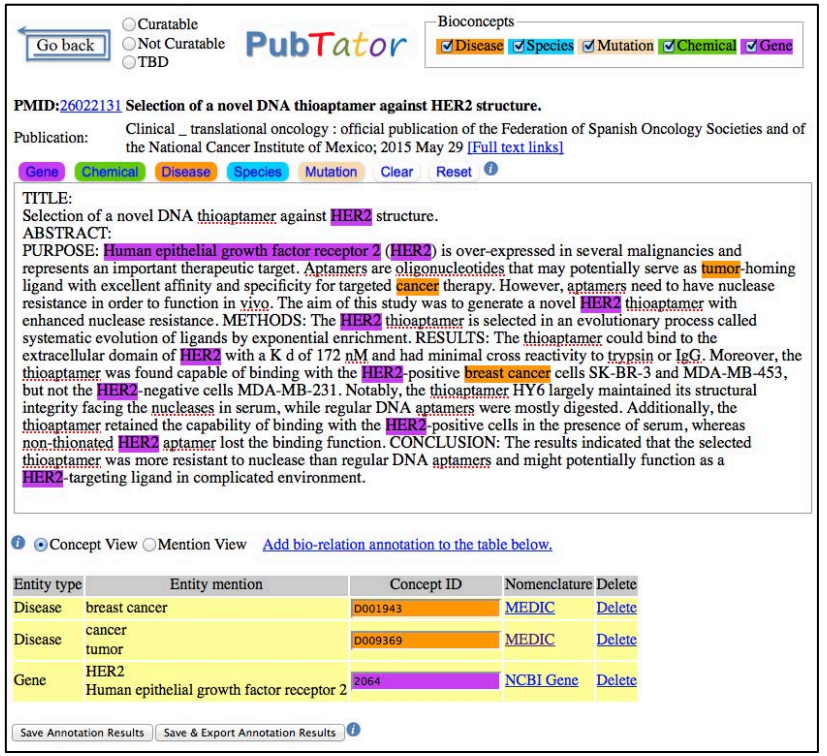
tmChem (2015) – 88.3%

All-in-one Tagger
AIONER (2023)

- State-of-the-art Performance
  - CHEMPROT-Elsevier Challenge Award, 2017
  - GNormPlus: 1st in BioCreative V CHEMDNER-GPRO Task
  - tmChem: 1st in BioCreative V CHEMDNER-CEMP Task
  - DNorm: 1st in ShARe/CLEF Disease Normalization Task
  - tmChem: 1st in BioCreative IV CHEMDNER-CEM Task
- Hybrid approach (mostly machine learning)
- Freely available and open source
- Scalable and interoperable
- Large user base

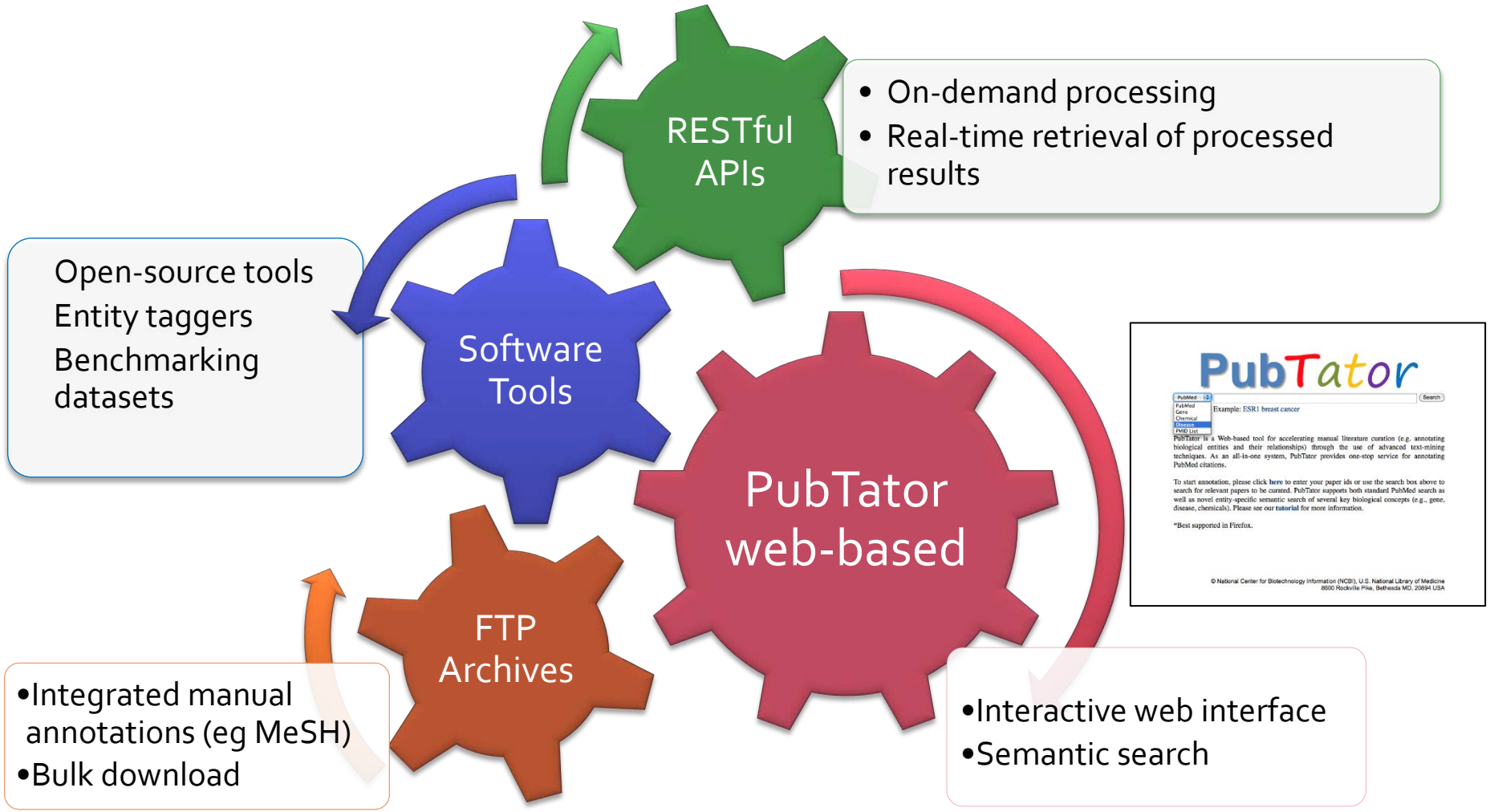
PubTator (2012 -): integrating text-mined results at PubMed scale



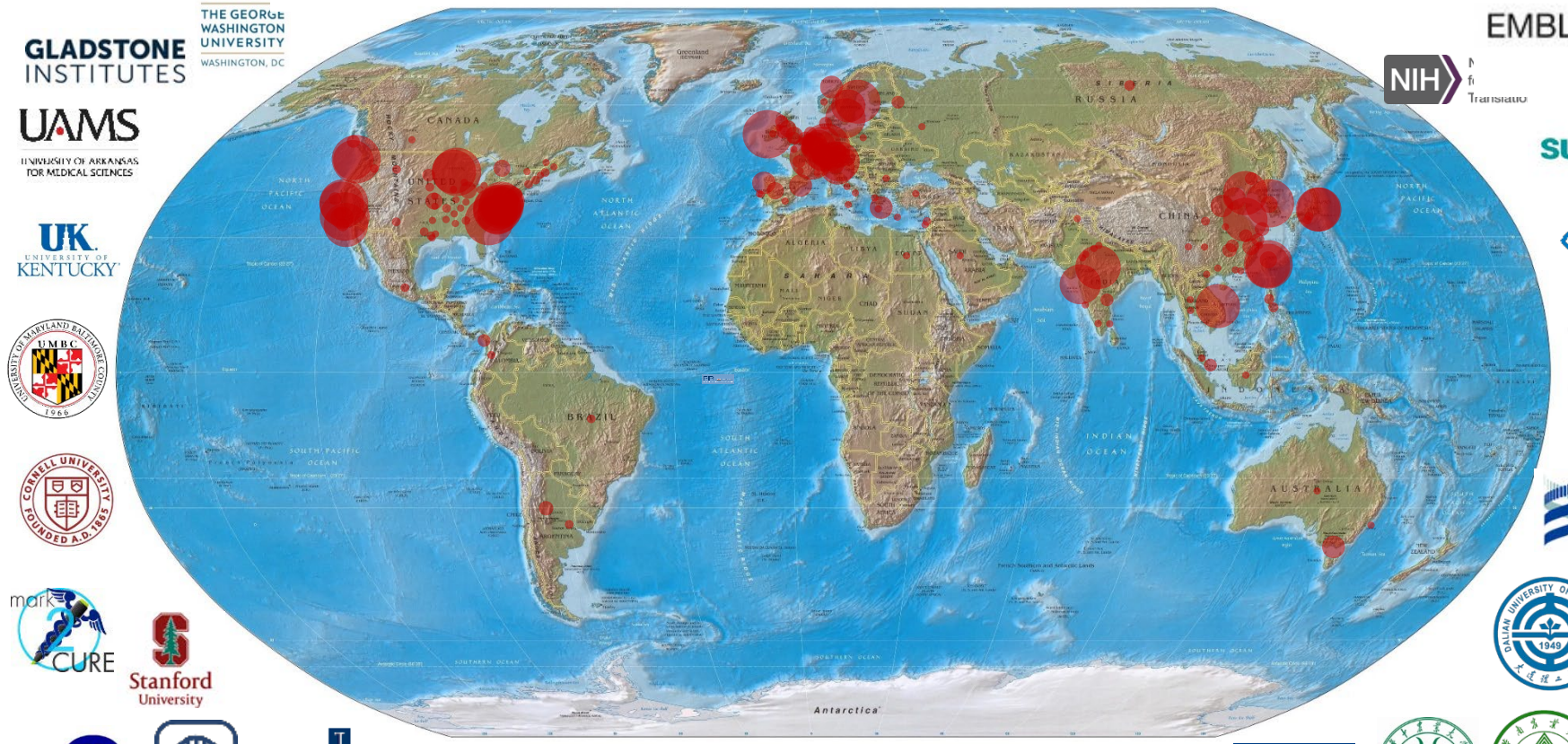


Wei, Kao, & Lu: PubTator: a Web-based text-mining tool for assisting biocuration, Nucleic Acids Research, 2013

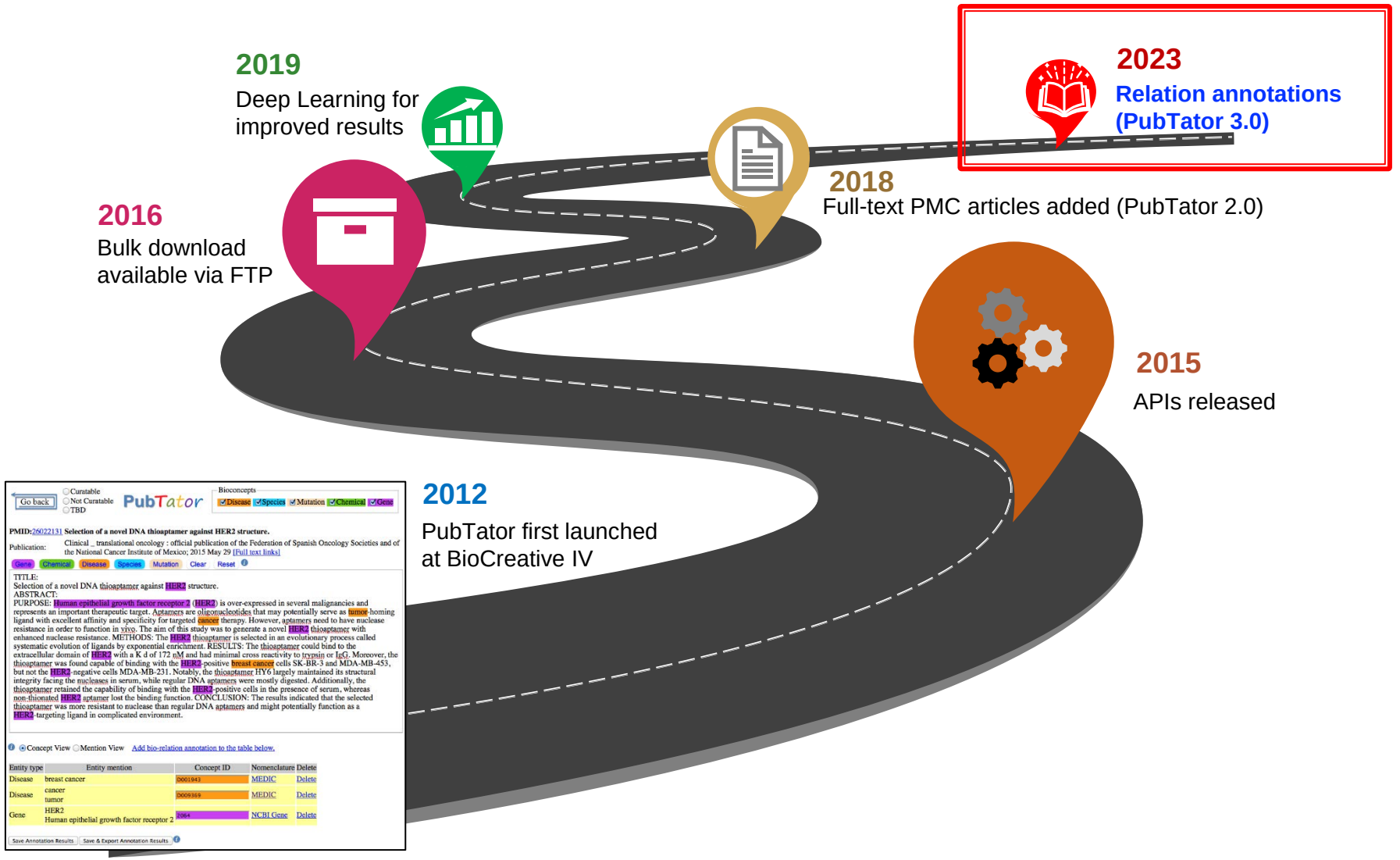
Data Sharing and Reuse



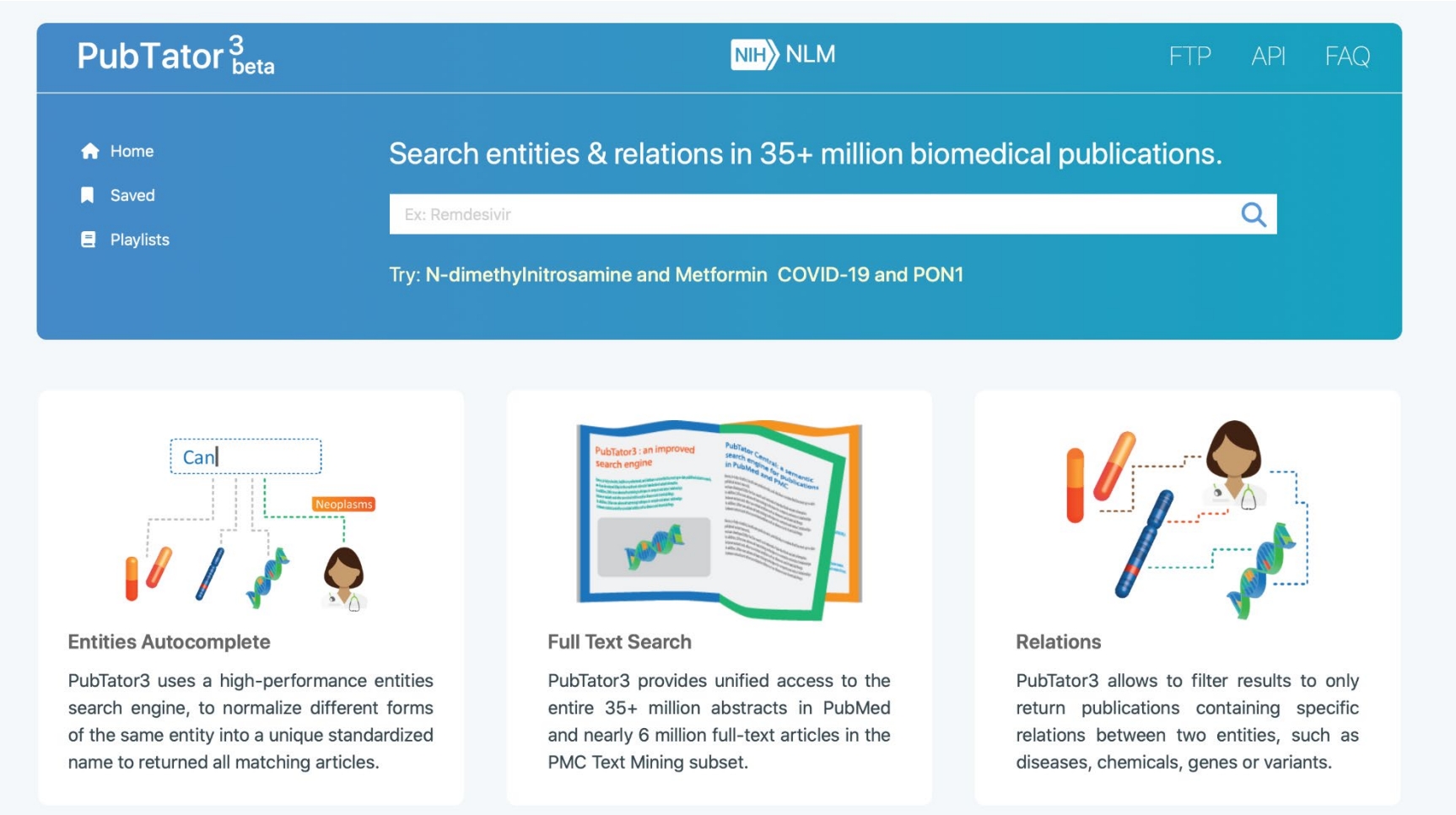
PubTator API usage: over 1 billion requests since April 2015



PubTator milestones (2012 -)



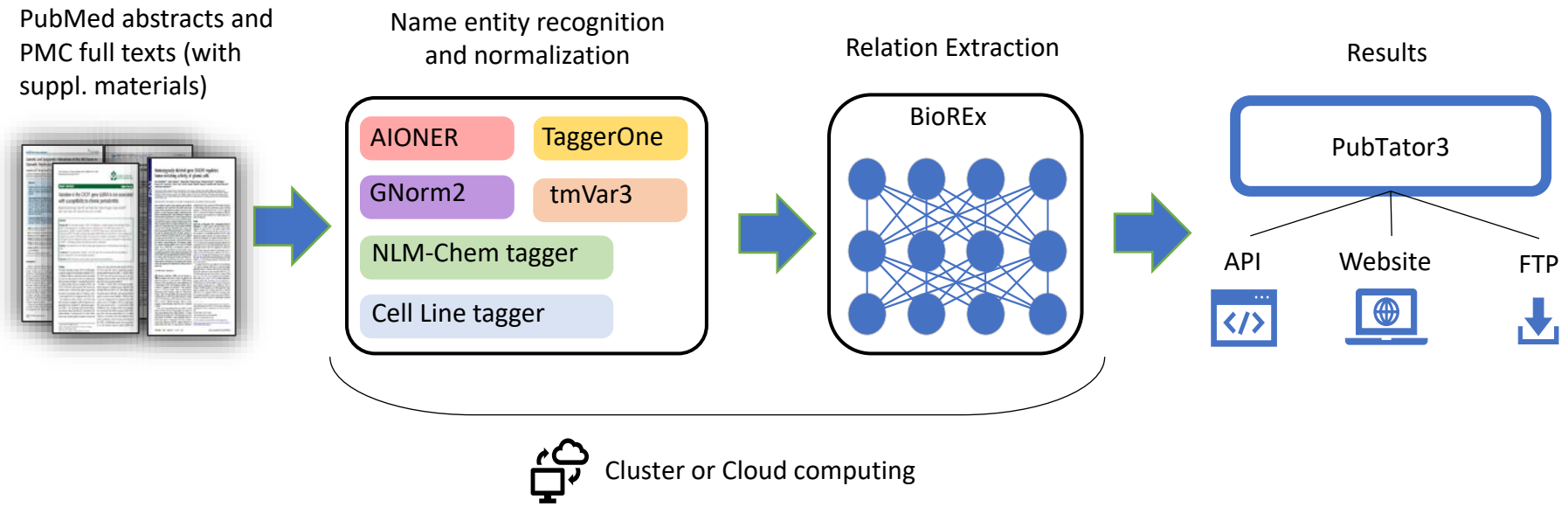
PubTator 3.0 beta is now available



https://www.ncbi.nlm.nih.gov/research/pubtator3/

Supported by NIH ODSS Exploratory Cloud Program

Computational tasks need GPU access



- IE methods enhancement
- (Re-)process the entire PubMed/PMC datasets
- Annotate new articles each day
  - 5k-100K PubMed abstracts / 1-5K full texts (daily)

Available computing resources

CPU/GPU	#	Resource
Nvidia A100	0-1	Biowulf
Nvidia V100	0-1	Biowulf
Nvidia Tesla K80	0-5	Biowulf
CPU	0-400	Biowulf
CPU	0-300	NCBI computer cluster

CPU/GPU	#	Resource
Nvidia A100	2	Google cloud
CPU	24	Google cloud

- Most of our IE tools are developed using deep learning, which rely on GPUs.
- The availability of NCBI/Biowulf resources is unstable and often occupied by other users.

Comparison

Tasks	NCBI Cluster/Biowulf	Google Cloud
Re-process the entire PubMed/PMC	100-600 CPUs + 1 K80 GPU	2 A100 +24 CPUs
	5+ months	1-2 months
Method improvement	GPU (not guaranteed)	2 A100
	Days to Weeks	Hours to Days
Regularly process new articles	100-300 CPUs	1 A100 + 24 CPUs
	1-5 days	Daily

Acknowledgments

Team members: Alexis Allot, Don Comeau, Qingyu Chen, Rezarta Dogan, Amr Elsawy, Aadit Kapoor, Won Kim, Qiao Jin, Robert Leaman, **Po-Ting Lai**, Ling Luo, Shubo Tian, **Chih-Hsuan Wei**, John Wilbur, Natalie Xie, Yifan Yang, Lana Yeganova, Qingqing Zhu

CC: Ron Summers, Le Lu, Xiaosong Wang, Ke Yan, etc.

NCATS: Tyler Beck, Christine Colvis, Noel Southall

NCI: Daniela Seminara, Travis Hyams, Brionna Hair

NHGRI: Vence Bonham, Larry Brody, Julia Byeon

NIAID: Morgan Similuk, Daniel Veltri, Sandhya Xirasagar, Rajarshi Ghosh, Andrew Oler

PubMed: Kathi Canese, Grisha Starchenko, etc.

dbSNP: Lon Phan, Juliana Feltz, Rama Maiti, Tim Hefferon

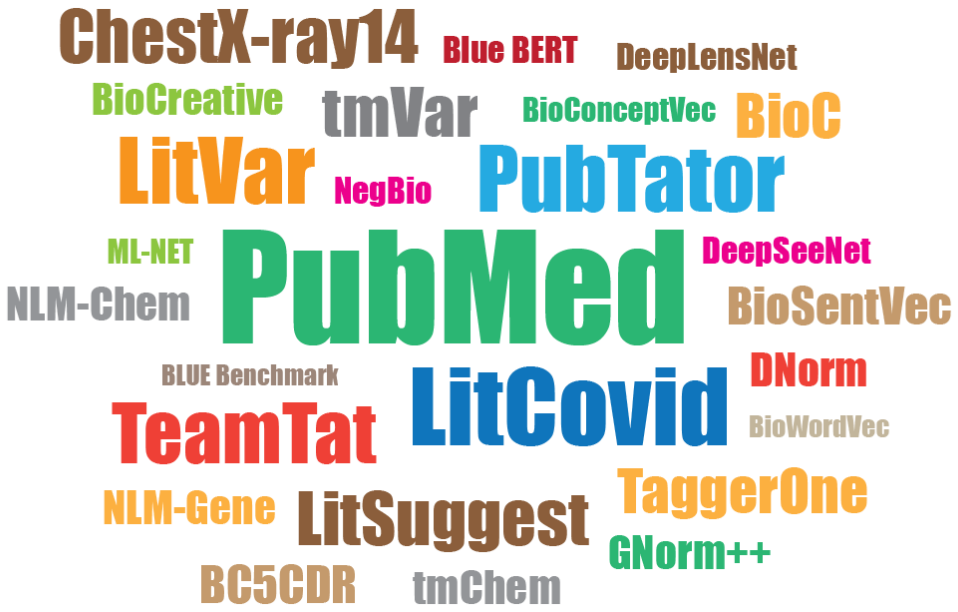
ClinVar: Melissa Landrum

UniProt: Cecilia Arighi, Alex Bateman, Alan Bridge, Livia Famiglietti, Michele Magrane, Sylvain Poux, Cathy Wu, Ioannis Xenarios, etc.

GWAS Catalog: Aoife McMahon, Jackie MacArthur, Fiona Cunningham, Helen Parkinson

And many others at NLM/NIH and beyond!!

Data Sharing and Open Science



zhiyong.lu@nih.gov