Biomedical AI for Precision Health

Hoifung Poon

Microsoft Health Futures

Overview

AI for precision health

Biomedical NLP

Self-supervised learning

Domain-specific pretraining

Knowledge-rich self-supervision

Mission: Structure all medical data

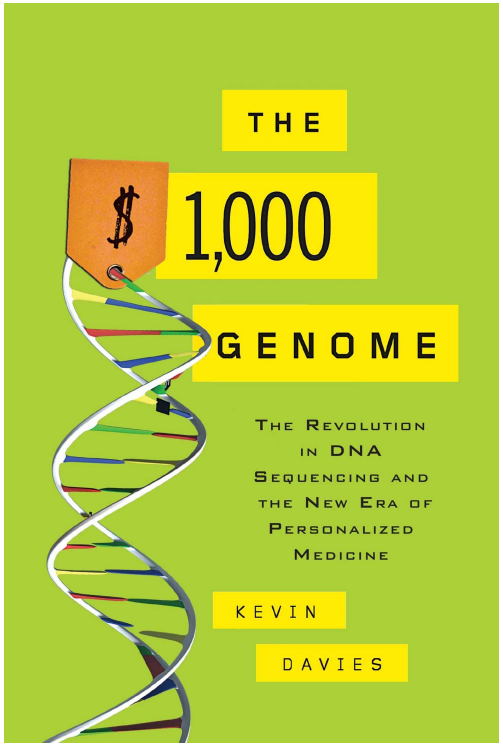
AI for Precision Health

Medicine Today Is Imprecise

Top 20 drugs
80% non-responders

Wasted
1/3 health spending
$1 Trillion / year

Disruption: Big Data

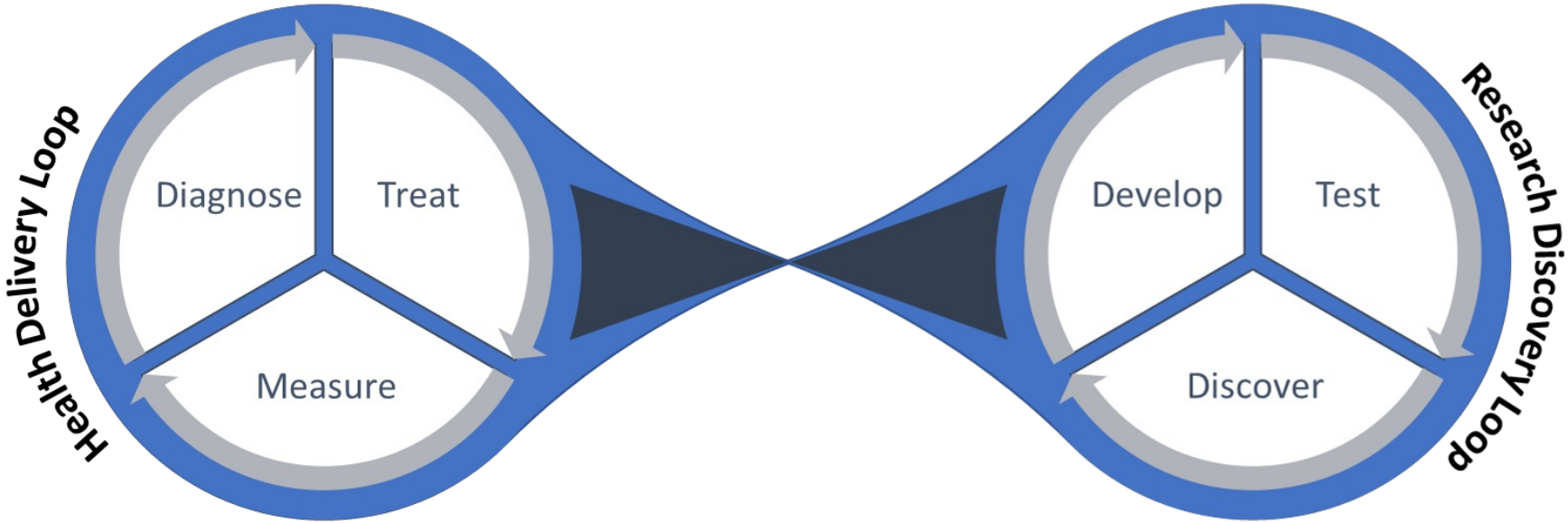


Accenture study: 93% of US doctors using EMRs

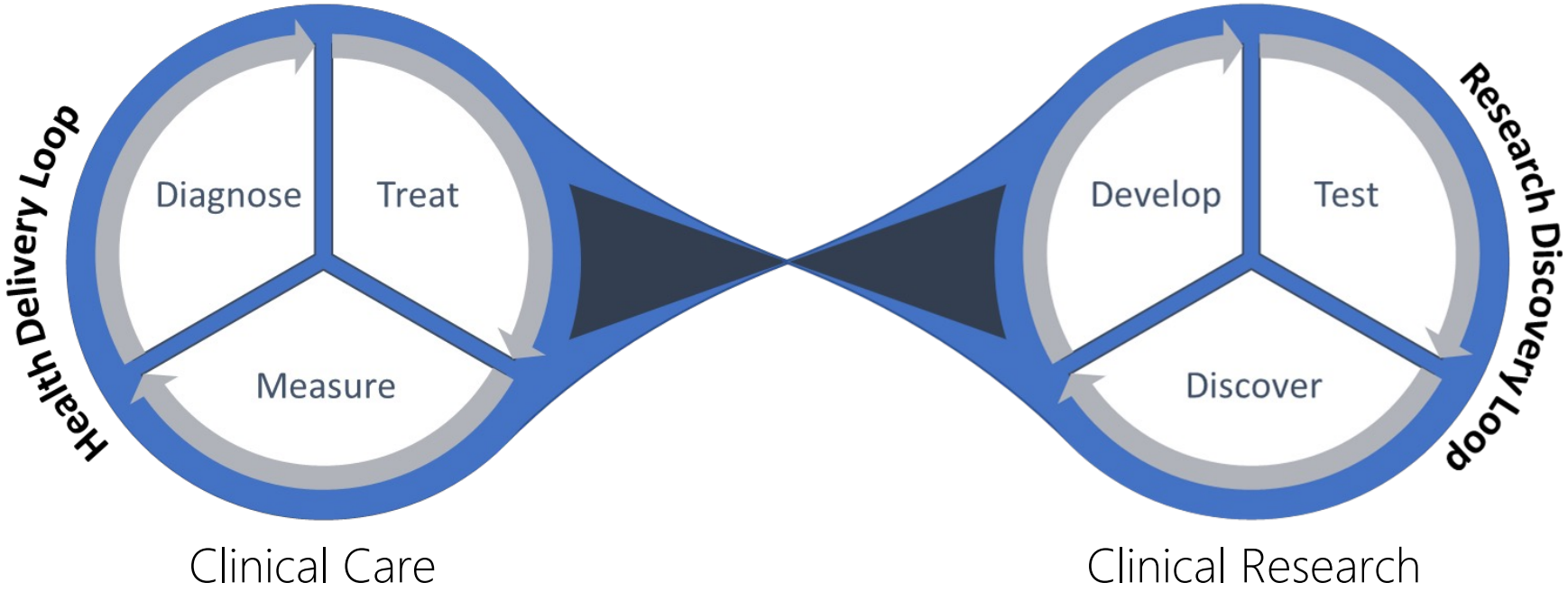
May 14, 2013 IHQRE informatics, IHQRE Journal Club EHR, EMR, Meaningful Use

2009 – 2013: 40% → 93%

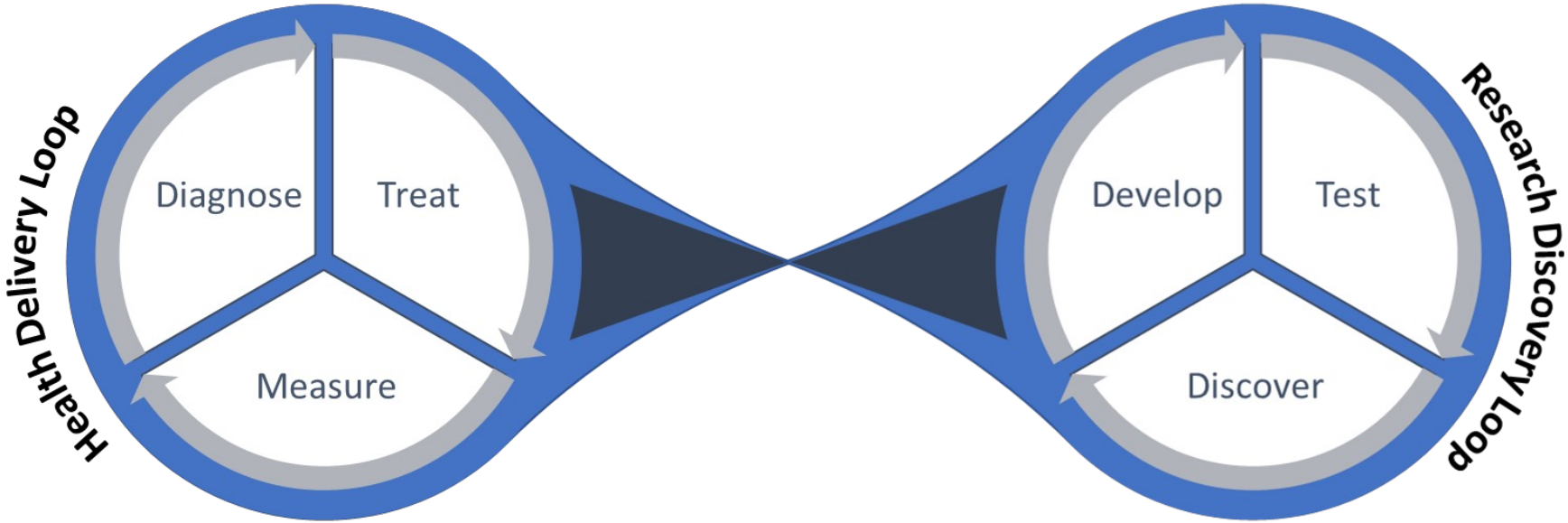
Towards Precision Health



India: two cancer centers per 10 millions
USA: 85% patients treated at community hospitals



New drug costs $2-10 billion and takes 10+ years
20% trials fail due to insufficient patients



Clinical Care

Clinical Research

https://imgur.com/rwUE5

Information Overload

EMR: 80% In Unstructured Text

Wolters Kluwer: Health Language Blog

Basic curation requires
2-3 expert hours per patient

1,23224,174680,2147-12-05,,,"Discharge summary","Report",,"","Admissi
on Date: [**2823-9-29**] Discharge Date: [**2823-10-1
7**]

Date of Birth: [**2768-10-11**] Sex: F

Service: SURGERY

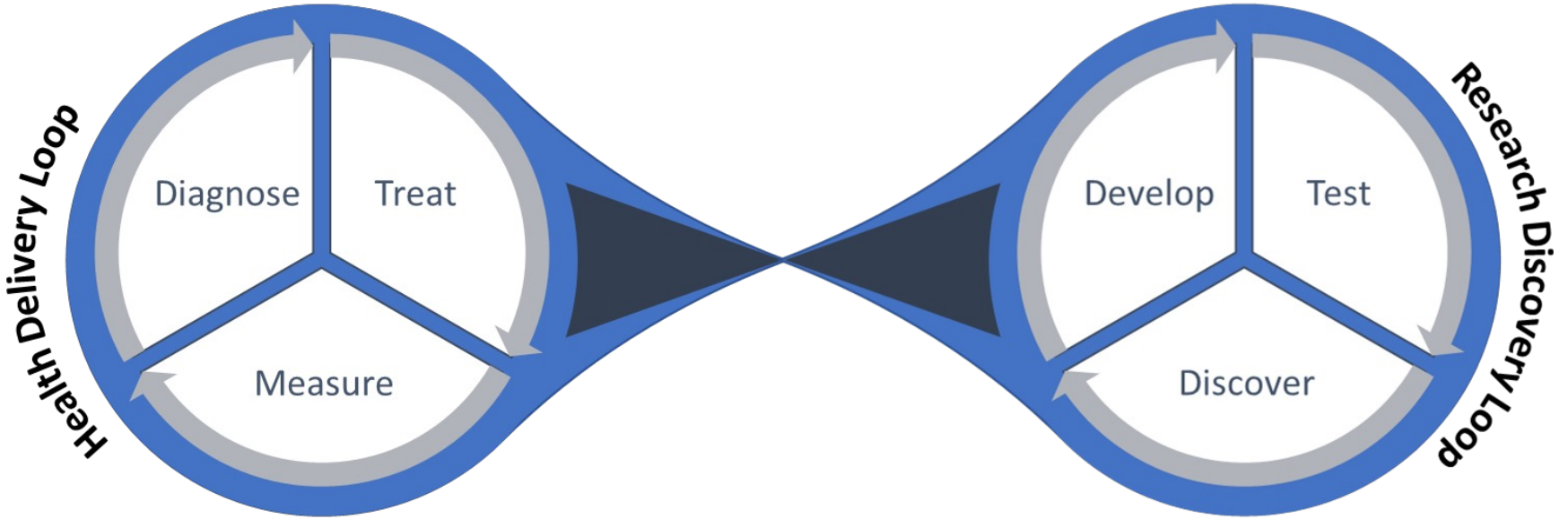
Allergies:
Patient recorded as having No Known Allergies to Drugs

Attending:[**First Name3 (LF) 1**]
Chief Complaint:
headache and neck stiffness

Major Surgical or Invasive Procedure:
central line placed, arterial line placed

History of Present Illness:
54 year old female with recent diagnosis of ulcerative colitis
on 6-mercaptopurine, prednisone 40-60 mg daily, who presents
with a new onset of headache and neck stiffness. The patient is
in distress, rigoring and has aphasia and only limited history
is obtained. She reports that she was awaken 1AM the morning of
[**2823-9-28**] with a headache which she describes as bandlike. She
states that headaches are unusual for her. She denies photo- or
phonophobia. She did have neck stiffness. On arrival to the ED
at 5:33PM, she was afebrile with a temp of 96.5, however she
later spiked with temp to 104.4 (rectal), HR 91, BP 112/54, RR
24, O2 sat 100 %. Head CT was done and relealved attenuation
within the subcortical white matter of the right medial frontal
lobe. LP was performed showing opening pressure 24 cm H2O WBC of
316, Protein 152, glucose 16. She was given Vancomycin 1 gm IV,
Ceftriaxone 2 gm IV, Acyclovir 800 mg IV, Ambesone 183 IV,
Ampicillin 2 gm IV q 4, Morphine 2-4 mg Q 4-6, Tylenol 1 gm ,
Decadron 10 mg IV. The patient was evaluated by Neuro in the
ED.

Biomedical AI for Precision Health



Democratize clinical care & scale clinical research

Translation, Dialog, Summarization

OCR, Imaging

Transcription

The deletion mutation on exon-19 of EGFR gene was present in 16 patients, while the L858E point mutation on exon-21 was noted in 10. All patients were treated with gefitinib and showed a partial response.

NLG

EGFR基因外大19的缺失突变在16例患者中存在，而 exon-21 的 L858E 点突变在10例中被注意到。所有患者均接受格菲替尼治疗,并表现出部分反应。

NLU

Machine Reading

KB

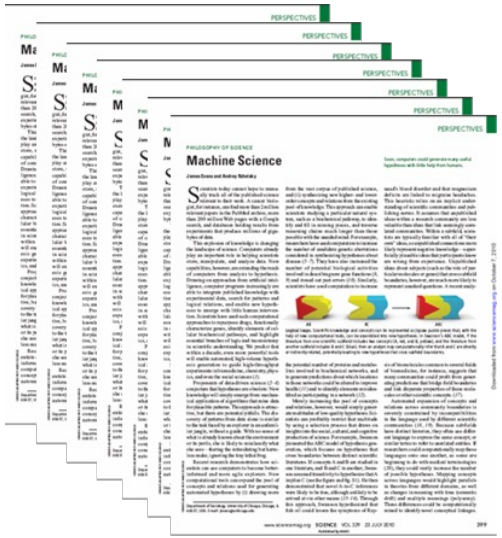
Machine Reading

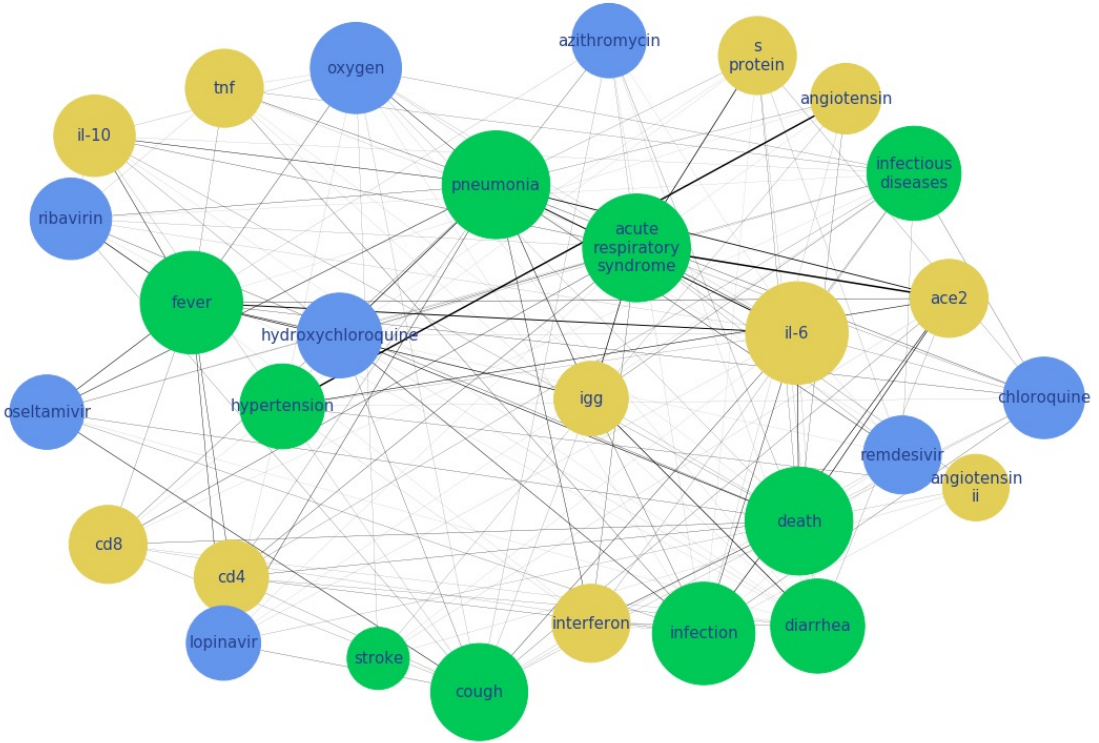
The deletion mutation on exon-19 of EGFR gene was present in 16 patients, while the L858E point mutation on exon-21 was noted in 10.

All patients were treated with gefitinib and showed a partial response.

TREAT(Gefitinib, EGFR, L858E)

Literature → Knowledge Graph





Case Study: Molecular Tumor Board

Problem: Hard to scale

U.S. 2021: 1.9 million new cases

902 cancer hospitals

Memorial Sloan Kettering

- Sequence: Tens of thousands
- Board can review: A few hundred

$200 rate X 10 experts X 3 hours X 1.9 m > $100 billion

OncoKB Team

OncoKB is developed and maintained by the Knowledge Systems group in the Marie Josée and Henry R. Kravis Center for Molecular Oncology at Memorial Sloan Kettering Cancer Center.

Design & Development

Debyani Chakravarty, PhD
Jianjiong Gao, PhD
Sarah Phillips, PhD
Hongxin Zhang, MSc
Ritika Kundra, MSc
Jiaojiao Wang, MSc
Ederlinda Paraiso, MPA
Julia Rudolph, MPA
David Solit, MD
Paul Sabbatini, MD
Nikolaus Schultz, PhD

Clinical Genomics Annotation Committee

Shrujal Baxi, MD, MPH
Margaret Callahan, MD, PhD
Sarat Chandarlapaty, MD, PhD
Alexandra Charen-Snyder, MD
Ping Chi, MD, PhD
Daniel Danila, MD
Mrinal Gounder, MD
James Harding, MD
Matthew Hellman, MD
Alan Ho, MD, PhD
Gopa Iyer, MD
Yelena Janjigian, MD
Thomas Kaley, MD
Maeve Lowery, MD
Antonio Omuro, MD
Paul Paik, MD
Michael Postow, MD
Dana Rathkopf, MD
Alexander Shoushtari, MD
Neerav Shukla, MD
Tiffany Traina, MD
Martin Voss, MD
Rona Yaeger, MD

Core Curators

Moriah Nissan, PhD
Lindsay Saunders, PhD
Tara Soumerai, MD
Fiona Brown, PhD
Tripti Shrestha Bhattarai, PhD
Kinisha Gala, BSc
Aphrothiti Hanrahan, PhD
Anton Henssen, MD
Phillip Jonsson, PhD
Iñigo Landa-Lopez, PhD
Eneda Toska, PhD

Quest Diagnostics

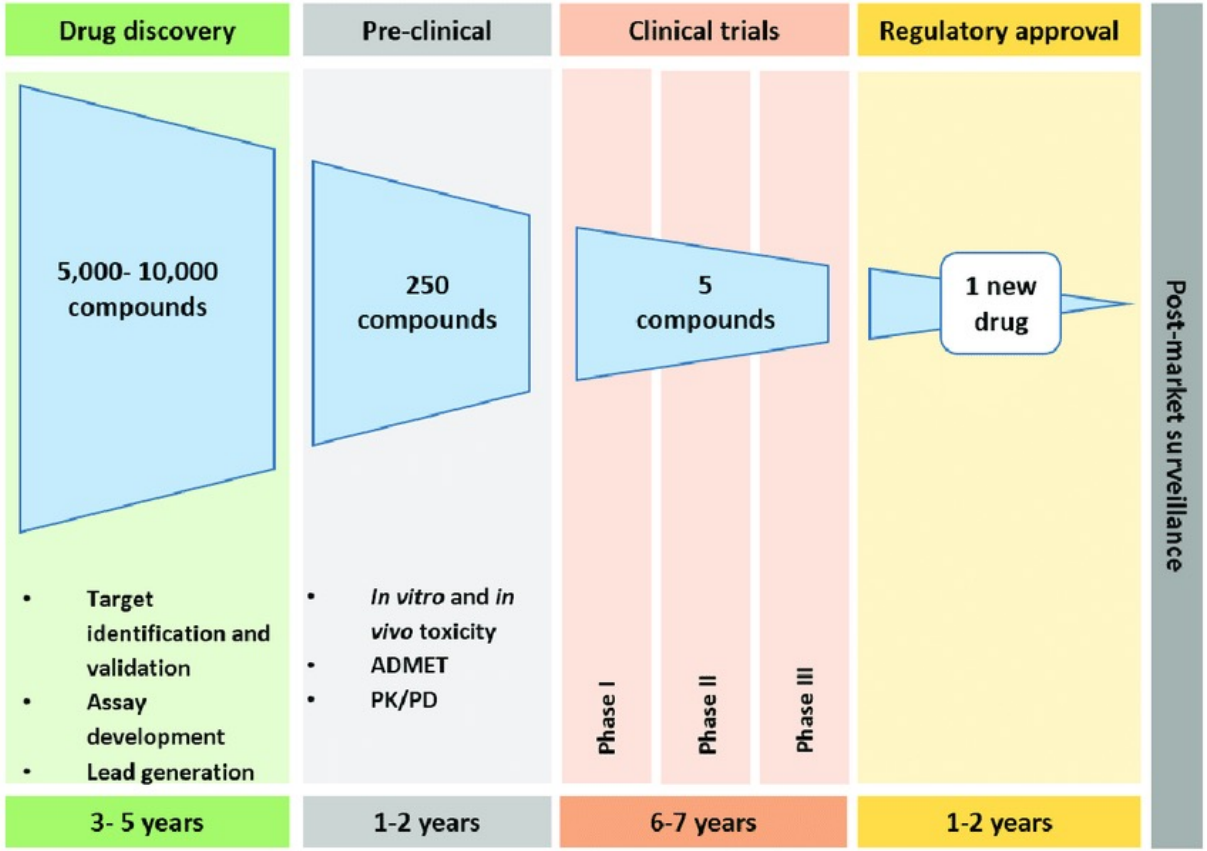
Feras M Abu Hantash, PhD
Andrew Grupe, PhD
Matthew Beer, BSc

Can we increase curation throughput by 10X?

EMR → Real-World Evidence (RWE)

Patient	Diagnosis	Treatment	Outcome
101	Lung Cancer	Gefitinib	remission
202	Leukemia	Imatinib	resistant
303	Lymphoma	Zaraparib	relapse
	.....		

Population-level high-definition patient journey



"Omics"-Informed Drug and Biomarker Discovery. Matthews et al. *Proteomes 2016*

Can we accelerate clinical research?

Drug Discovery	Clinical Trial	Post-Market
Target Identification	Eligibility	Adverse Event
Drug Repurposing	Synthetic Control	Comparative Effectiveness
	Virtual Trial	Off-Label Use
	Pragmatic Trial	

Real-World Evidence

Trillion-dollar opportunity:
Accelerate development; reduce cost; save lives

Case Study: Immunotherapy RWE

Keytruda: immunotherapy blockbuster ($14B, 2020)

FDA approved for many cancer indications

But only work for minority of patients. Why?

Case Study: Immunotherapy RWE

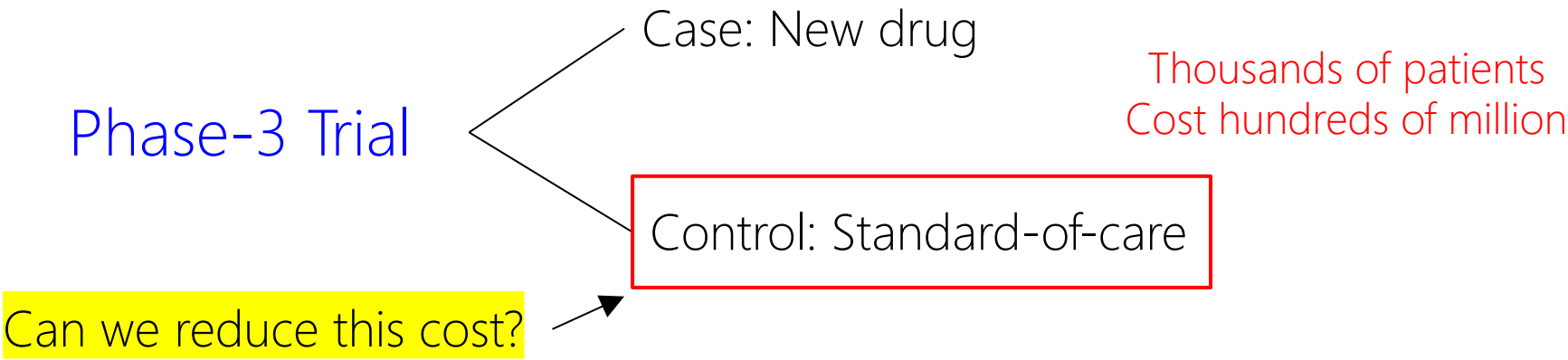
Given Keytruda cohort, find exceptional responder

Need to extract progression events

- "Patient's cancer in complete remission …"
- "… treatment was discontinued due to toxicity …"
- "… tumor continues to progress despite treatment …"

Case Study: Synthetic Control

Average cost of an FDA-approved drug		Annual number of FDA-approved drugs		Per Year
$2.5-10 billion	×	~50	=	$125-500 billion



Case Study: Synthetic Control

EMR: Standard of care ⇒ Virtual control arm

Case study: Flatiron

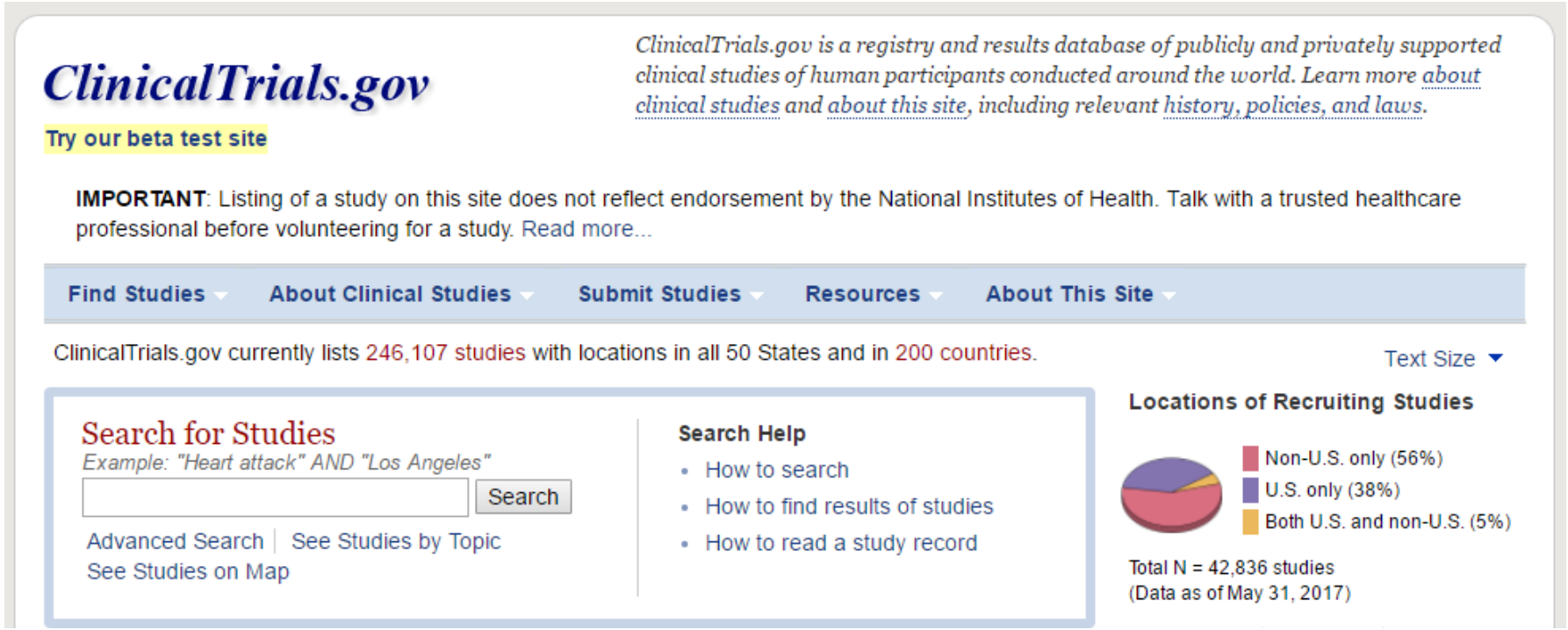
Hire over 1,000 abstractors

- Pfizer: Ibrance for male breast cancer
- Roche: Alectinib for ALK lung cancer

Roche to acquire Flatiron Health for $2.1 billion, with focus on real-world data

March 02, 2018 | Vol.44 No.09

Clinical Trial



Clinical Trial

Eligibility

Ages Eligible for Study: 18 Years and older (Adult, Senior)
Sexes Eligible for Study: Female
Accepts Healthy Volunteers: No

Criteria

Inclusion Criteria: A subject will be eligible for inclusion in this study only if all of the following criteria are met:

1. Female subjects, age ≥ 18 years at the time informed consent is signed
2. Pathologically confirmed adenocarcinoma of the breast
3. Pathologically confirmed as triple negative, source documented, defined as both of the following
    a. Estrogen Receptor (ER) and Progesterone Receptor (PgR) negative: < 1% of tumor cell nuclei are immunoreactive in the presence of evidence that the sample can express ER or PgR (positive intrinsic controls)
    b. Human Epidermal Growth Factor Receptor 2 (HER2) negative as per American Society of Clinical Oncology - College of American Pathologists (ASCO/CAP) guidelines i. Immunohistochemistry (IHC) 0 or 1 Fluorescence In Situ Hybridization (FISH) negative (or equivalent negative test). Subjects with IHC 2 must have a negative by Fluorescence In Situ Hybridization (FISH),, (or equivalent negative test).
4. Subjects with prior breast cancer history of different phenotypes (ie, ER/PgR/HER2 positive) must have pathologic confirmation of triple negative disease in at least one of the current sites of metastasis
5. Subjects must have received prior adjuvant or neoadjuvant anthracycline therapy; unless (a) anthracycline treatment was not indicated or was not the best treatment option for the subject in the opinion of the treating physician; and (b) anthracycline treatment remains not indicated or, in the opinion of the treating physician, is not the best treatment option for the subject's metastatic disease. a. Newly diagnosed subjects presenting with TNMBC are eligible for the study if anthracycline treatment is not indicated or is not the best treatment option for the subject in the opinion of the treating physician.
6. Subjects with measurable metastatic disease, defined by Response Evaluation Criteria in Solid Tumors 1.1 (RECIST 1.1) guidelines
7. Life expectancy ≥ 16 weeks from randomization
8. No prior cytotoxic chemotherapy for metastatic breast cancer. Prior immunotherapy and/or monoclonal antibody therapy are acceptable. Prior treatments must have been discontinued at least 30 days prior to start of study treatment and all related toxicities must have resolved to Grade 1 or less.
9. Prior neoadjuvant or adjuvant chemotherapy, if given, must have been completed at least 6 months before randomization with all related toxicities resolved, and documented evidence of disease progression per RECIST 1.1 guidelines is required. a. If prior neoadjuvant or adjuvant chemotherapy contained taxane, gemcitabine, or platinum agents, the treatment must have completed at least 12 months before randomization
10. Prior radiotherapy must have completed before randomization, with full recovery from acute radiation side effects. At least one measurable lesion must be completely outside the radiation portal or there must be unequivocal radiologic or clinical exam proof of progressive disease within the radiation portal, in accordance with RECIST 1.1 guidelines
11. At least 30 days from major surgery before randomization, with full recovery
12. Eastern Cooperative Oncology Group (ECOG) performance status 0-1
13. Subject has the following blood counts at screening:
    - Absolute Neutrophil Count (ANC) ≥ 1500/mm^2 ;
    - Platelets ≥ 100,000/mm^2 ;
    - Hemoglobin (Hgb) ≥ 9 g/dL

Clinical Trial Matching

Marty Tenenbaum

Late-stage melanoma (late 1990s)

Initial prognosis: 6 months

Saved by Phase III trial of Canvaxin

20% trials failed due to insufficient patients

Post-Market Surveillance

Adverse event

Comparative effectiveness

Drug repurposing

Pragmatic trial

Great opportunities, but can we deliver?

Biomedical AI: Recurring Themes

General domain vs biomed

Mirage of supervised learning

Real-world impact

General Domain vs Biomed

Trevor Noah's Will Smith Joke During Monologue at the 2022 Grammy Awards

Most people figured that 2022 Grammy Awards host Trevor Noah would make a reference to Will Smith's Oscars slap of Chris Rock during his opening monologue.

The comedian and host of The Daily Show made a number of light-hearted jokes while opening the 64th Annual Grammy Awards on Sunday night. He almost made it the whole way through without mentioning the infamous Oscars moment, but he made a subtly when he told the audience they're going to be "keeping people's names" out of their mouths.

Imaging Report

TECHNIQUE:
Spot magnification views in the craniocaudal and medial lateral projection was obtained at the level of the nipple and retroareolar area.

FINDINGS:
There are pleomorphic calcifications located in the left retroareolar region approximately 2 cm from the nipple. These extend over approximately 2 cm. These are not clearly appreciated on prior studies. These are suspicious. Biopsy will be necessary for further evaluation. These are at an anterior depth.

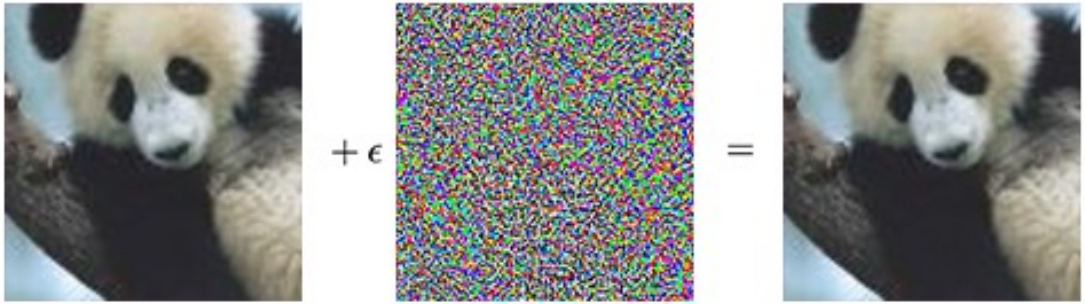
General Domain vs Biomed

Two cows are grazing in the field

IMPRESSION

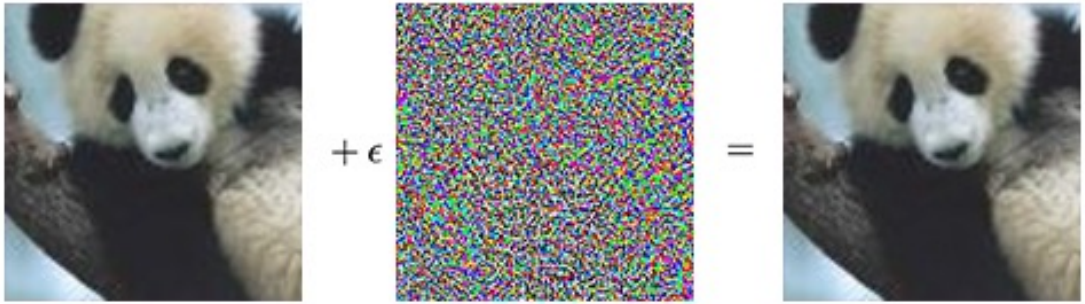
No significant change in right middle and low lobe pneumonia. Small increase in left pleural effusion. ……

Mirage of Supervised Learning



Overfit to spurious correlation

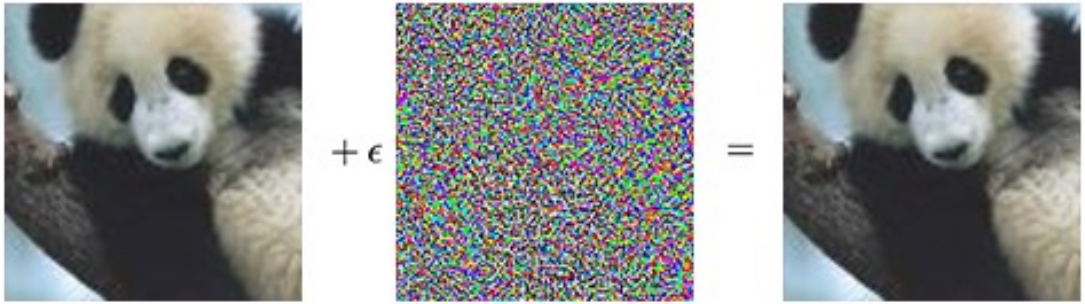
Mirage of Supervised Learning



ML Assumption 101
Training & test instances are drawn from same distribution

Overfit to spurious correlation

Mirage of Supervised Learning

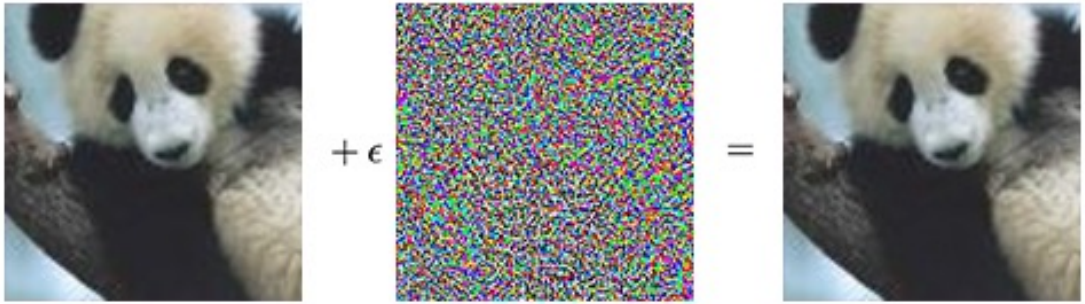


ML Assumption 101
Train & test instances drawn from same distribution

Usually, not quite

Overfit to spurious correlation

Mirage of Supervised Learning



~~ML Assumption 101~~
~~Training and test instances drawn from same distribution~~

Worse, may not have training data at all

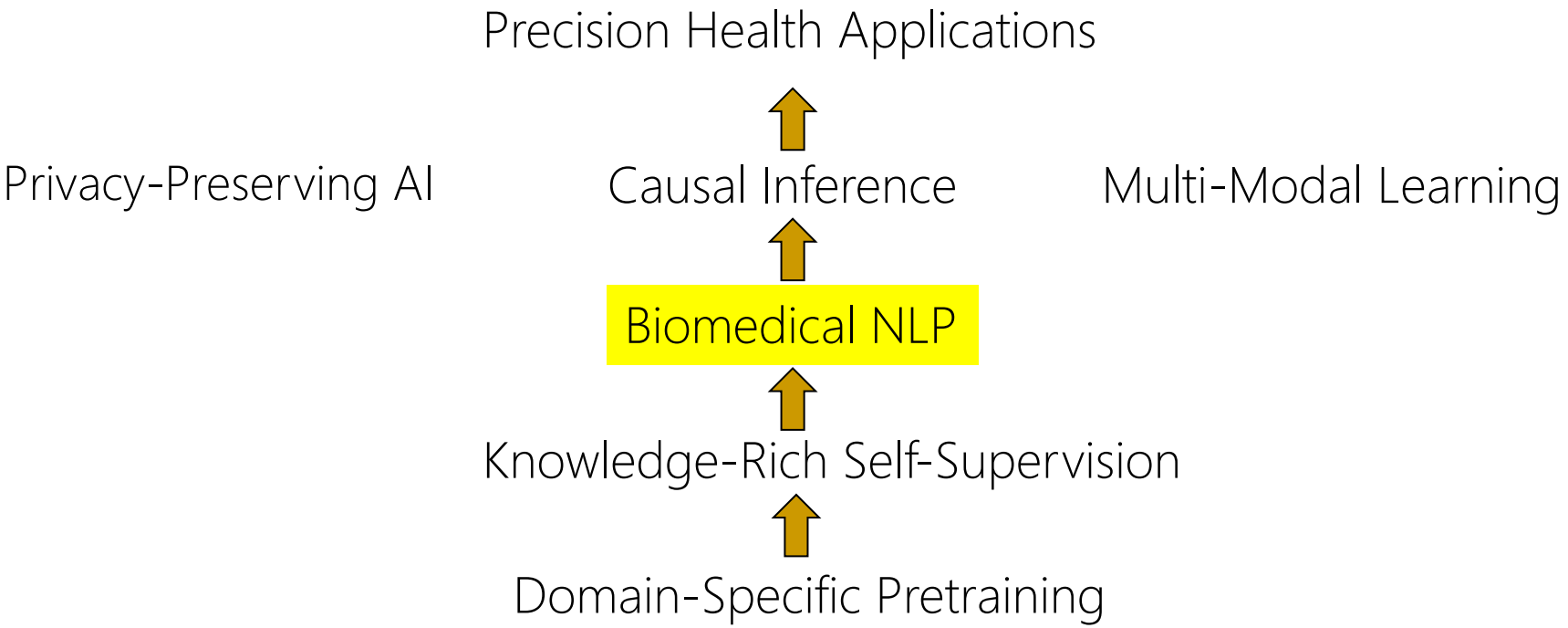
Real-World Impact

AI needs to fit into end-to-end scenarios

Deep partnership w. key stakeholders

Rigorous evaluation on real-world data

Biomed AI for Precision Health



Biomedical NLP

Machine Reading

Microsoft buys LinkedIn

Machine Reading

Microsoft buys LinkedIn

Machine Reading

Microsoft buys LinkedIn

COMPANY: Microsoft

COMPANY: LinkedIn

Named Entity Recognition (NER)

Machine Reading

Microsoft buys LinkedIn

Relation Extraction

ACQUIRE(Microsoft, LinkedIn)

Machine Reading

The deletion mutation on exon-19 of EGFR gene was present in 16 patients, while the L858E point mutation on exon-21 was noted in 10.

All patients were treated with gefitinib and showed a partial response.

Machine Reading

The deletion mutation on exon-19 of EGFR gene was present in 16 patients, while the L858E point mutation on exon-21 was noted in 10.

All patients were treated with gefitinib and showed a partial response.

DRUG: Gefitinib

GENE: EGFR

MUTATION: L858E

Named Entity Recognition (NER)

Machine Reading

The deletion mutation on exon-19 of EGFR gene was present in 16 patients, while the L858E point mutation on exon-21 was noted in 10.

All patients were treated with gefitinib and showed a partial response.

Relation Extraction

TREAT(Gefitinib, EGFR, L858E)

Machine Reading

DRUG: Talazoparib

GENE: KRAS

Mutation: A146T

To explore whether MEKi could re-sensitize PARPi resistant cells to effects of PARPi , we developed PARPi resistant cells by culturing highly PARPi sensitive cells (UWB1.289 and A27980CP , both RAS wild type , see Fig . 2) in the continued presence of BMN673 for 3 to 4 months , at which time drug resistant clones emerged . A2780CP PARPi resistant (A2780CP_R) and UWB1.289 PARPi resistant (UWB1.289_R) clones were highly resistant to BMN673 and cross resistant to olaparib (Fig . 3A-B) . RPPA analysis demonstrated that RAS / MAPK pathway activity (increased pMEK , pBAD , and pFOXO3a (inactive form)) was upregulated in PARPi resistant clones (Fig . 3C) . Moreover , resistant clones showed lower total FOXO3a and BIM , as expected from increased RAS / MAPK pathway activity . The decreased PAR and PARP1 expression in the resistant cells could also contribute to PARPi resistance , as PARP1 expression is associated with PARPi sensitivity (22) . Western blotting confirmed increased RAS / MEK pathway activity with concomitant decreases in FOXO3a and BIM in resistant cells (Fig . 3D) . Overall , the signaling changes in long-term PARPi resistant cells exhibited many similarities to adaptive responses to short-term PARPi treatment (see Fig . 1) . Despite increased RAS / MEK pathway activity , KRAS sequencing demonstrated that the resistant lines did not acquire classical activating KRAS mutations . However , deep NGS sequencing as well as Sanger sequencing of individual PARPi resistant clones from A2780CP_R demonstrated the presence of KRAS A146T , KRAS A59T and MAP2K1 A283T in 19 , 11 and 6 % of cells respectively but not in A2780CP parental cells . Importantly , prolonged culture of the lines without PARPi resulted in loss of the mutant KRAS and MAP2K1 clones . The KRAS A146T mutant has been demonstrated to be modestly activating (30) . The selection of KRAS mutations in a PARPi resistant line supports the concept that RAS mutations and RAS / MAPK pathway activation is a key mediator of PARP resistance . As expected by increases in RAS / MAPK activity in PARPi resistant cell lines and KRAS and MAPK1 mutations , A2780CP_R were markedly more sensitive and UWB1.289_R were modestly more sensitive to MEKi (Fig . 3E-F) . MEKi re-sensitized both PARPi resistant clones to PARPi (Fig . 3E-F) . Thus MEKi have the potential to re-sensitize PARPi resistant human tumors to PARPi .

RESISTANT(Talazoparib, KRAS, A146T)

Machine Reading

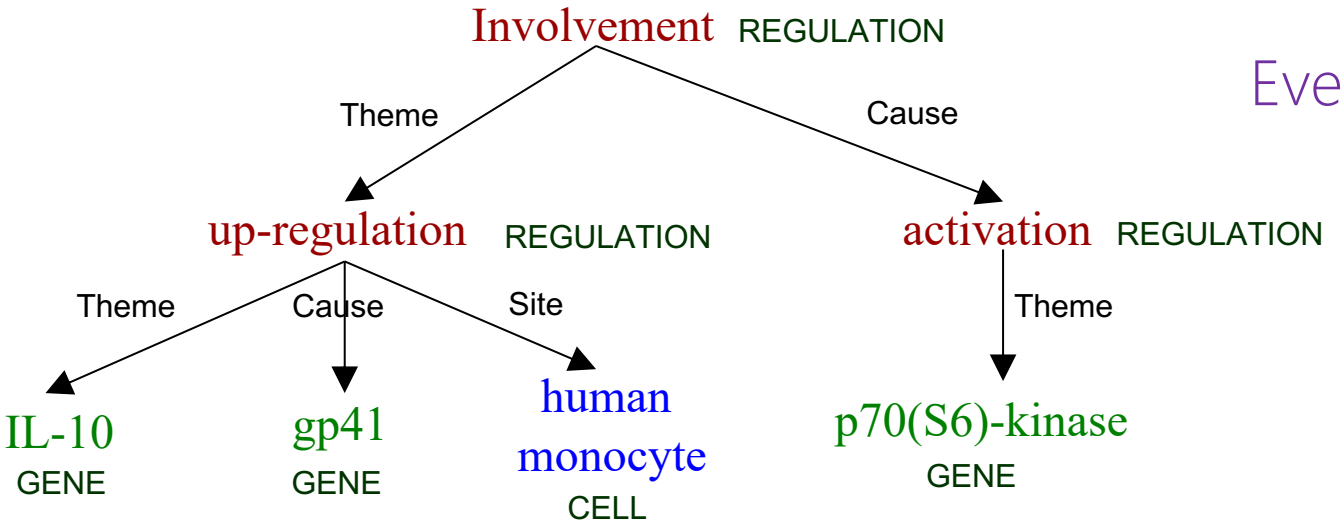
"We next expressed ALK F1174L, ALK F1174L/L1198P, ALK F1174L/G1123S, and ALK F1174L/G1123D in the original SH-SY5Y cell line."

(. . . 15 sentences and 2 figures . . .)

"The 2 mutations that were only found in the neuroblastoma resistance screen (G1123S/D) are located in the glycine-rich loop, which is known to be crucial for ATP and ligand binding and are the first mutations described that induce resistance to TAE684, but not to PF02341066."

Machine Reading

Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...

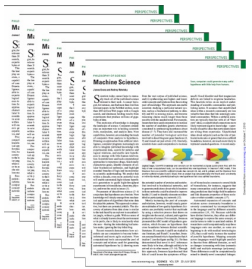


Event Extraction

NLP: The Big Picture

NLP is not a single thing, but spans a diverse range of tasks, domains, and complexities

NLP Pipeline



Sentence breaking
Tokenization
Morphology

POS Tagging
Syntactic Parsing

NER
Coreference
Entity Linking
Semantic Parsing
Discourse Parsing

Information Extraction
Question Answering
Machine Translation
.....

Component Tasks

Applications

Sentence breaking
Tokenization
Morphology

POS Tagging
Syntactic Parsing

NER
Coreference
Entity Linking
Semantic Parsing
Discourse Parsing

Information Extraction
Question Answering
Machine Translation
.....

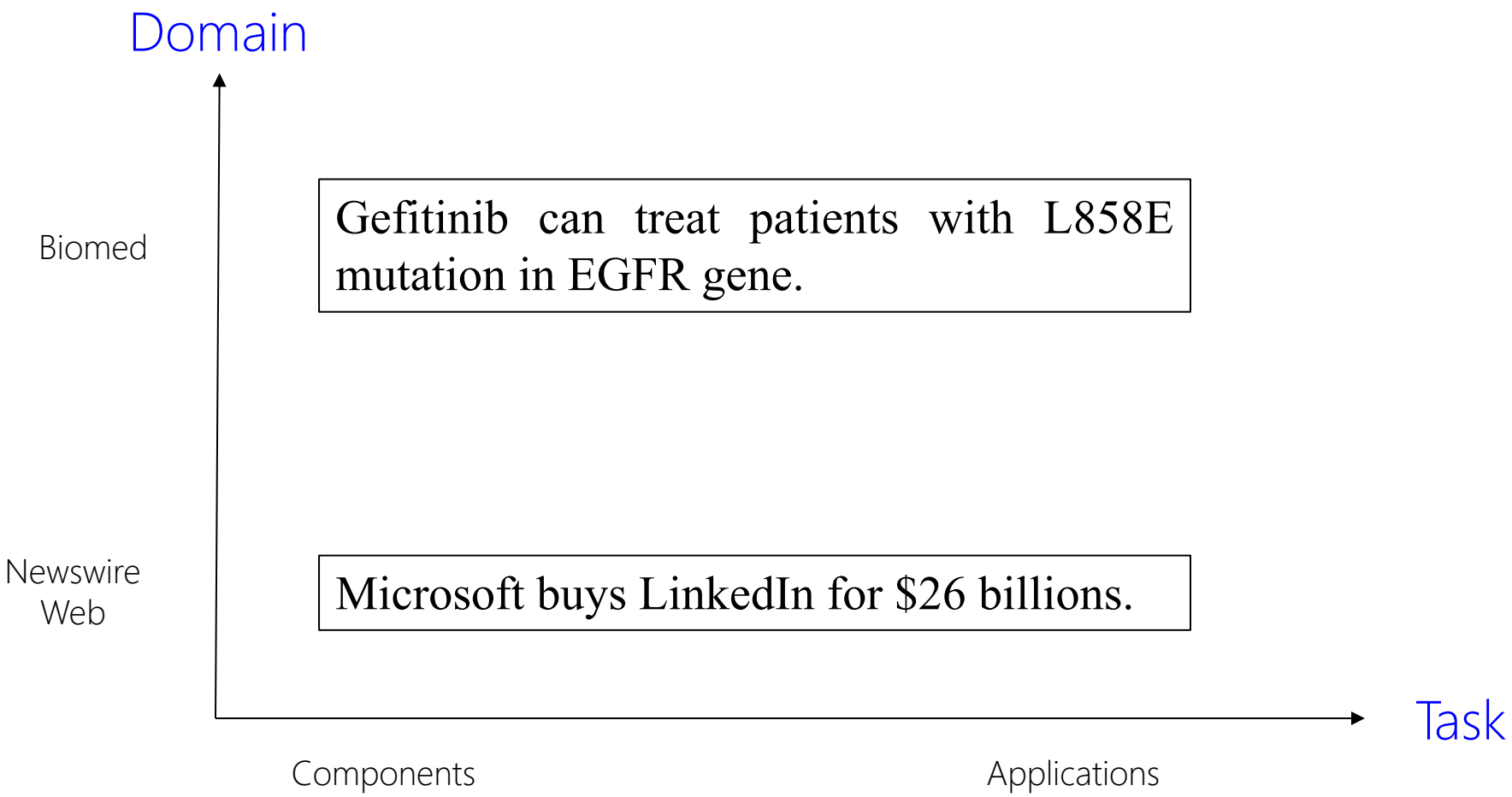
Applications

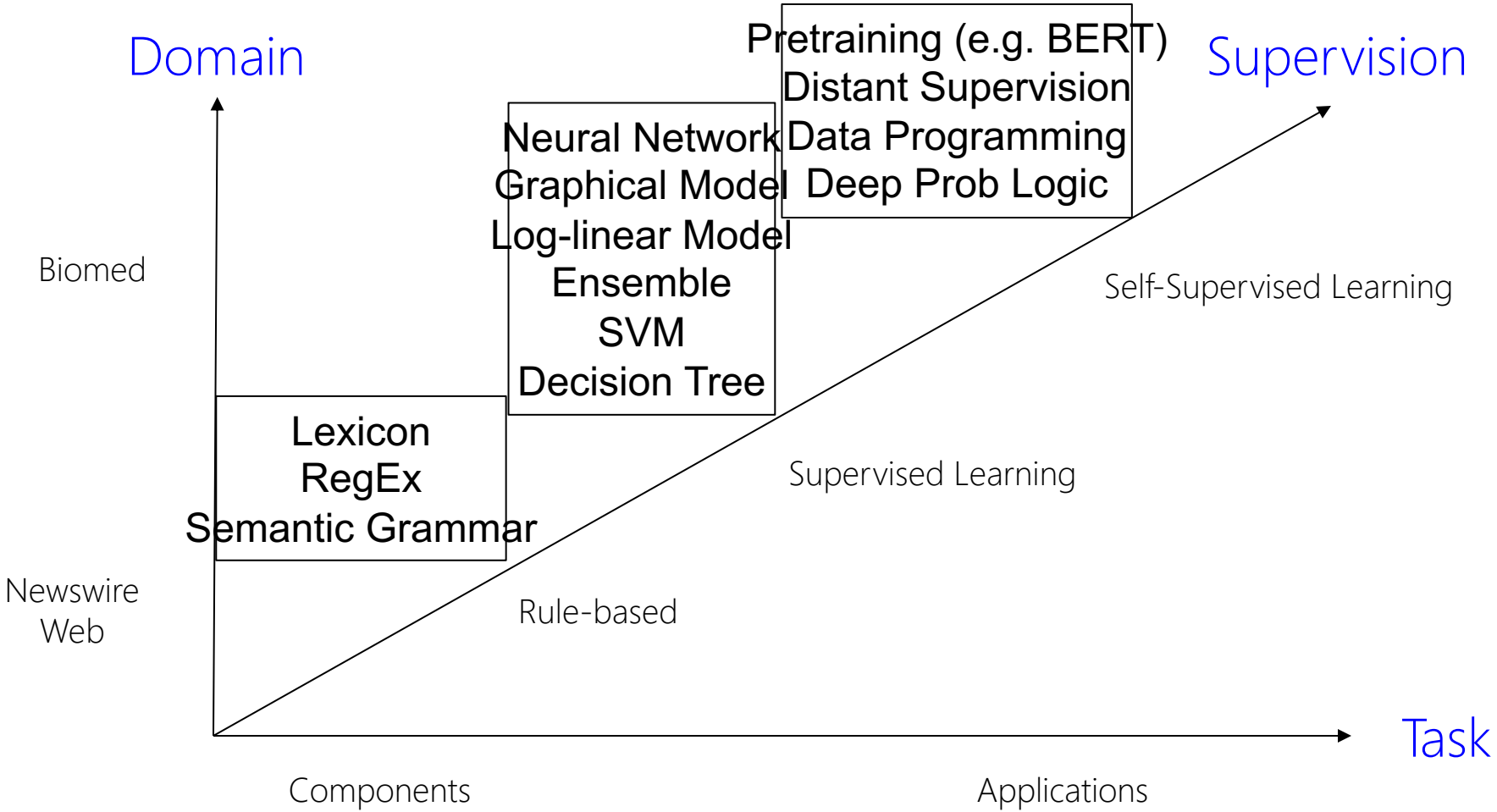
Component Tasks

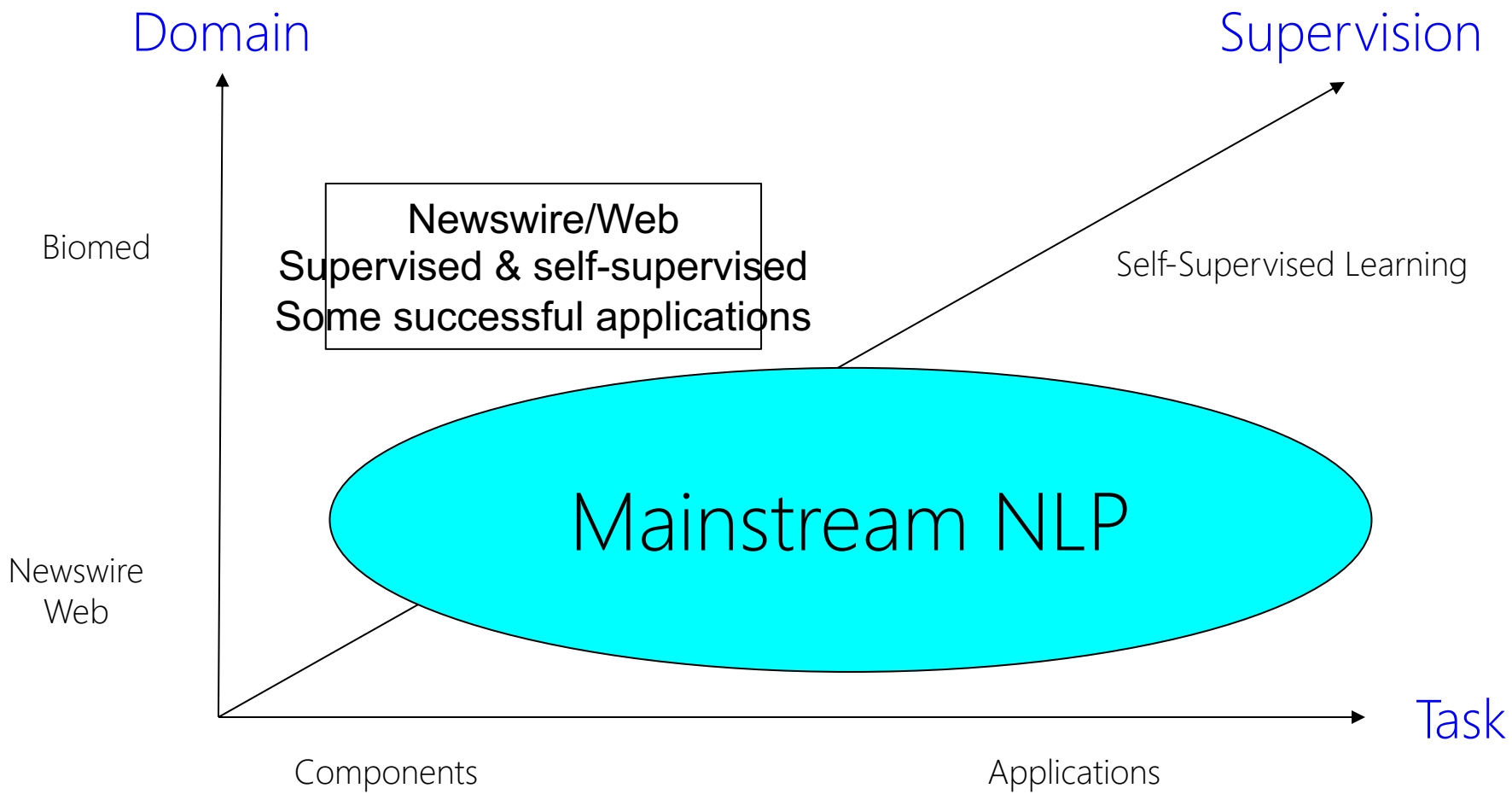
Components

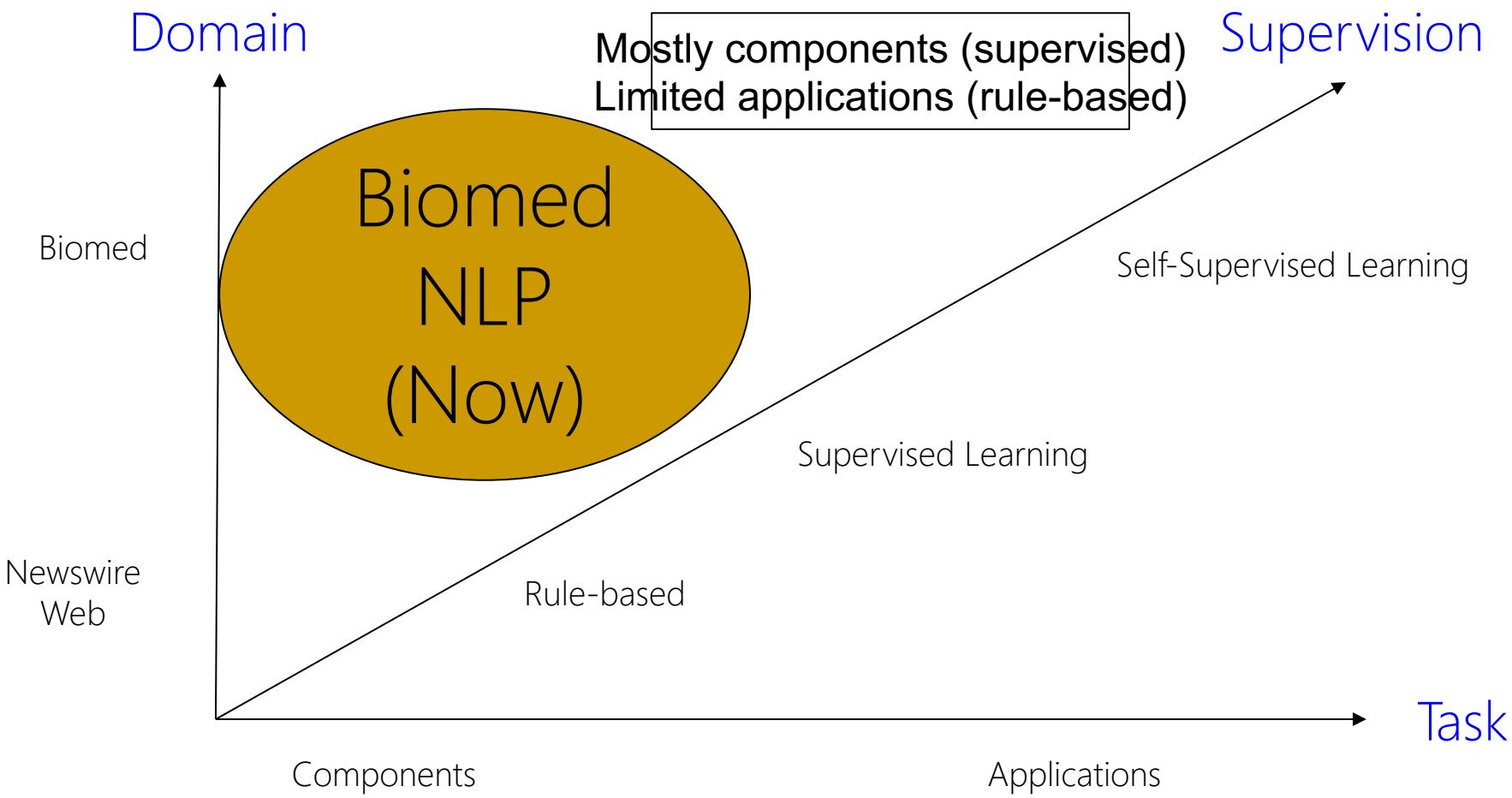
Applications

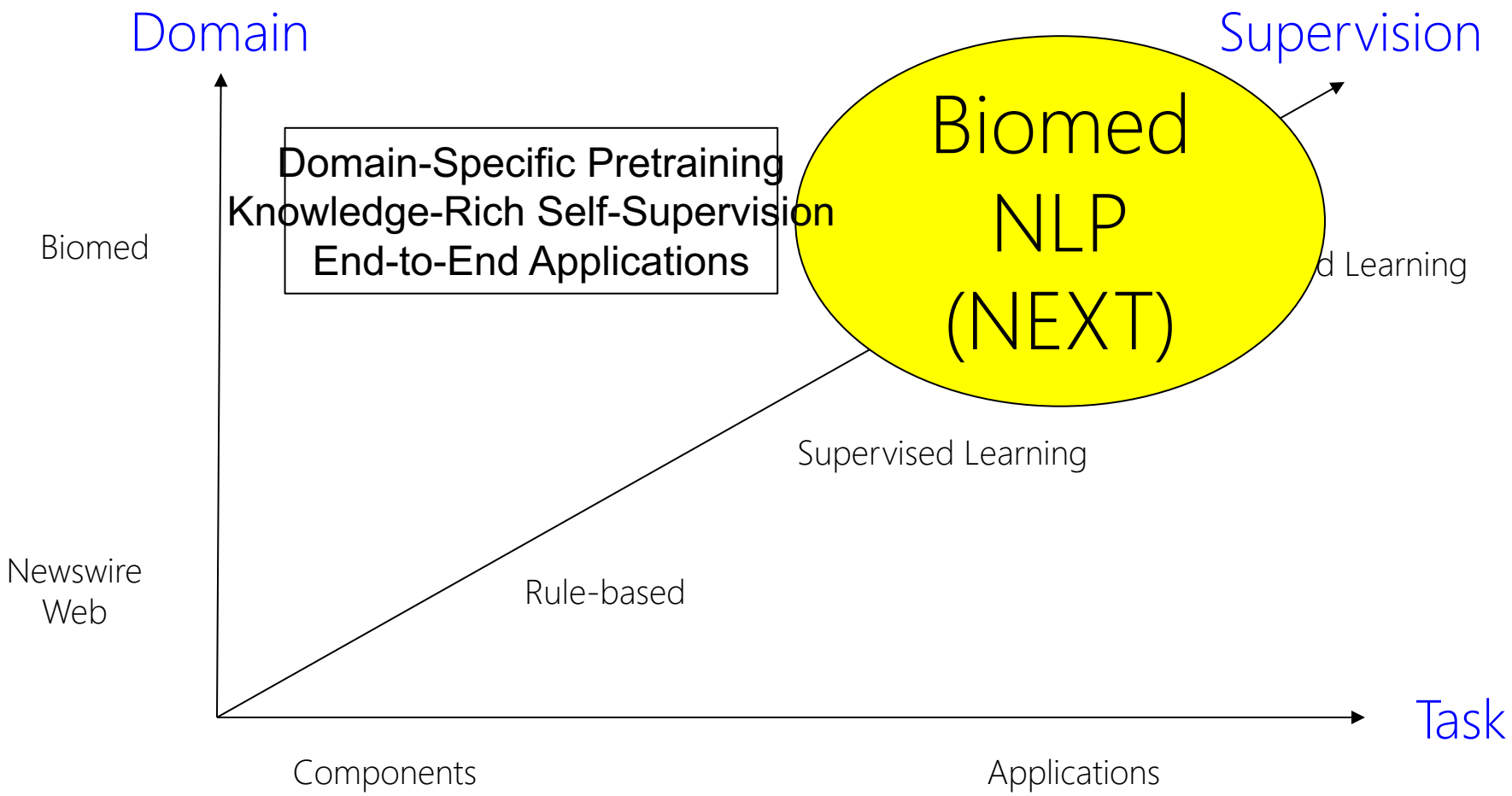
Task


Domain
Biomed
Gefitinib can treat patients with L858E mutation in EGFR gene.
Newswire
Web
Microsoft buys LinkedIn for $26 billions.
Task
Components
Applications


Domain
Supervision
Pretraining (e.g. BERT)
Distant Supervision
Data Programming
Deep Prob Logic
Neural Network
Graphical Model
Log-linear Model
Ensemble
SVM
Decision Tree
Lexicon
RegEx
Semantic Grammar
Biomed
Newswire
Web
Self-Supervised Learning
Supervised Learning
Rule-based
Task
Components
Applications


Domain
Supervision
Biomed
Newswire/Web
Supervised & self-supervised
Some successful applications
Self-Supervised Learning
Mainstream NLP
Newswire
Web
Task
Components
Applications


Domain
Supervision
Mostly components (supervised)
Limited applications (rule-based)
Biomed NLP (Now)
Biomed
Newswire
Web
Self-Supervised Learning
Supervised Learning
Rule-based
Task
Components
Applications


Domain
Supervision
Biomed
Newswire
Web
Domain-Specific Pretraining
Knowledge-Rich Self-Supervision
End-to-End Applications
Biomed
NLP
(NEXT)
d Learning
Supervised Learning
Rule-based
Task
Components
Applications

NLP Pipeline

Off-the-shelf biomed NLP tools

Supervised NER + rule-based Entity Linking

Sentence breaking
Tokenization
Morphology

POS Tagging
Syntactic Parsing

NER
Coreference
Entity Linking
Semantic Parsing
Discourse Parsing

Information Extraction
Question Answering
Machine Translation
.....

Applications

Component Tasks

A Brief History of NLP

Big Bang | GOFAI | Statistical Revolution | Deep Learning

Computer, AI, NLP
Turing Test, 1950
AI Birth (Dartmouth, Hanover NH), 1956
Chomsky ("Syntactic Structures"), 1957

Machine Translation
Cold war: Russian to English
Demo: IBM-Georgetown, 1954
Crash: ALPAC Report, 1966

Lesson: Pretty demo not enough
Need rigorous evaluation & benchmarks

1940-60 | 1970-80 | 1990-2010 | 2010-Present

Big Bang | GOFAI | Statistical Revolution | Deep Learning

Rule-base
Lexicon
RegEx
Semantic Grammar

Dialog, Question-Answering
Eliza, 1964
BASEBALL (Green et al.), 1961
SHRDLU (Winograd et al.), 1973
LUNAR (Woods et al.), 1978

Still used in most "clinical NLP" and "biomedical NLP" today

Negation Detection
Hedge Detection
Ontology-Based Entity Linking
.....

1940-60 | 1970-80 | 1990-2010 | 2010-Present

Big Bang | GOFAI | Statistical Revolution | Deep Learning

Statistical Machine Learning

Classification: Decision tree, Random Forest, Naïve Bayes, SVM, kernel methods, log-linear models, …

Structured Prediction: Dynamic Programming, HMM, CRF, probabilistic logic, …

Morphology, Syntactic Parsing, Named Entity Recognition (NER), Information Extraction, Question Answering, Machine Translation, …

Penn Treebank, 1990s

ACE, 2003

PropBank, 2005

Newswire / Web

Most on component tasks

1940-60 ..... 1970-80 | 1990-2010 | 2010-Present

Big Bang | GOFAI | **Statistical Revolution** | Deep Learning

Then: "NLP is all about feature engineering"

Sentence breaking
Tokenization
Morphology

POS Tagging
Syntactic Parsing

NER
Coreference
Entity Linking
Semantic Parsing
Discourse Parsing

Information Extraction
Question Answering
Machine Translation
.....

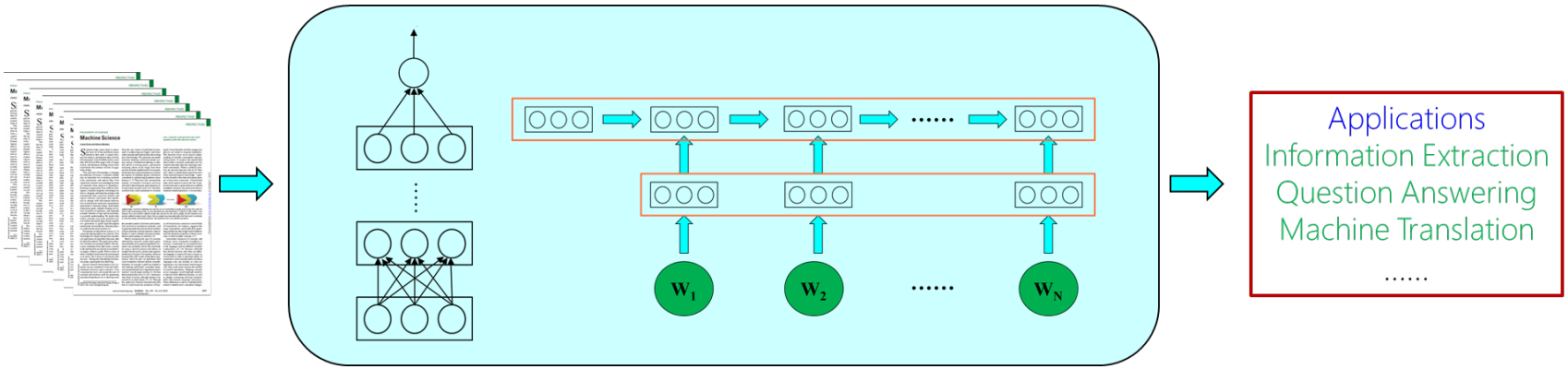
Component Tasks

Applications

1940-60 | 1970-80 | 1990-2010 | 2010-Present

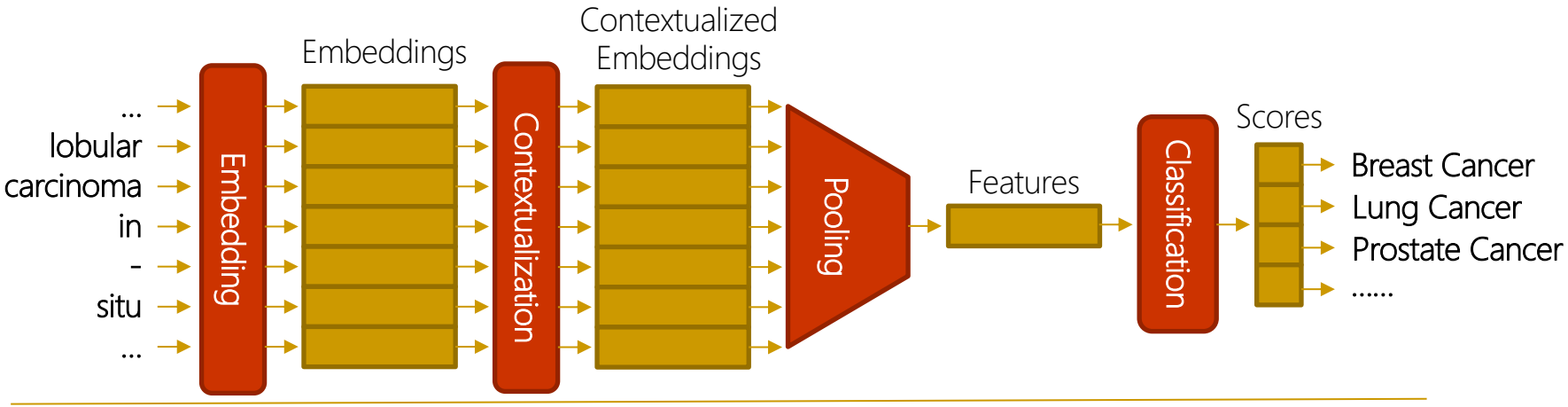
Big Bang | GOFAI | Statistical Revolution | Deep Learning

Now: End-to-end deep learning



1940-60 | 1970-80 | 1990-2010 | 2010-Present

End-to-End Deep Learning



GloVe
Word2Vec

CNN
RNN

Average
Max
Attention

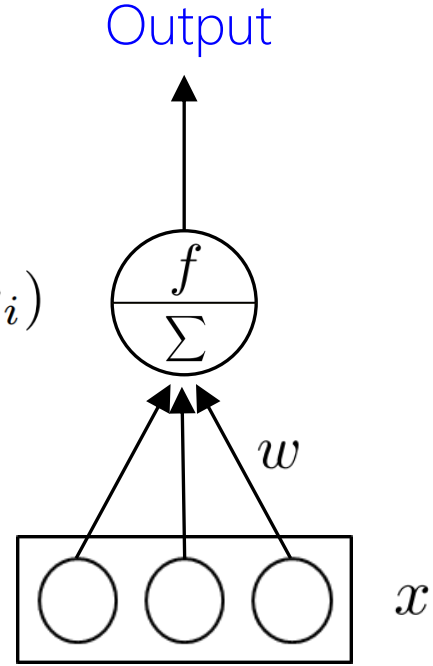
Linear
MLP
Logistic Regression
.....

ELMo, BERT, ELECTRA, ...

A Brief History of Deep Learning

Neural Unit

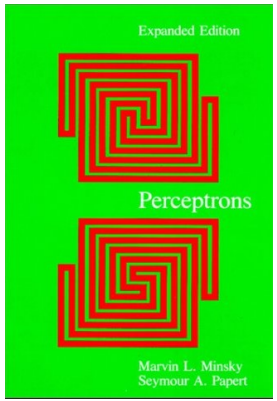
$$O = f(\sum_{i=1}^{n} w_i \cdot x_i)$$



Source: Arvin Calspan Advanced Technology Center; Hecht-Nielsen, R. Neurocomputing (Reading, Mass.: Addison-Wesley, 1990)

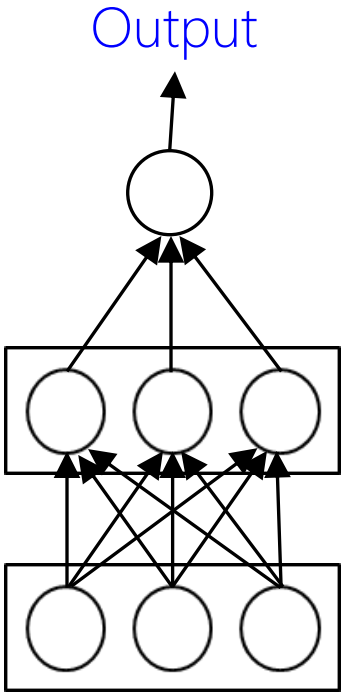
First Wave

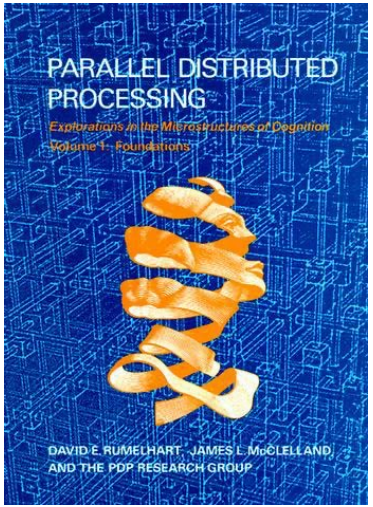
Perceptron
[Rosenblatt, 1957]



Can not represent complex functions such as XOR
[Minsky & Papert, 1969]

Neural Network



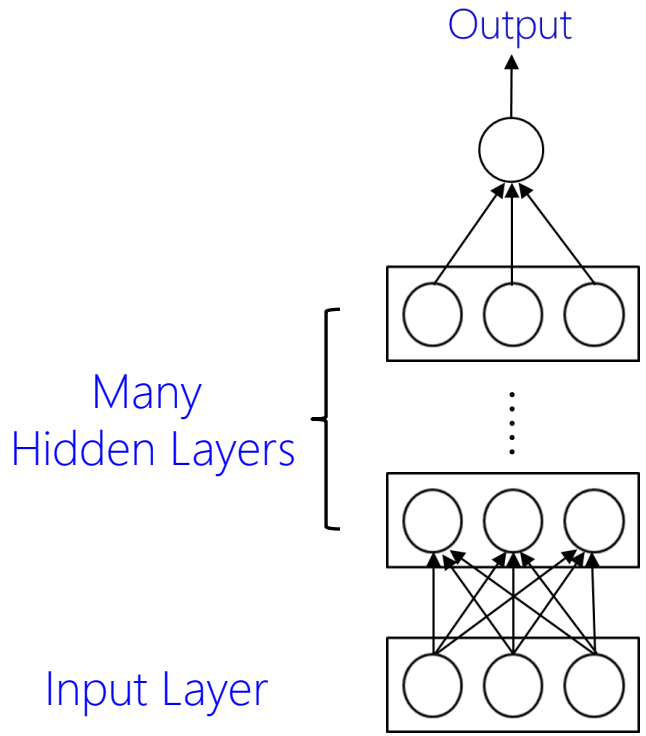


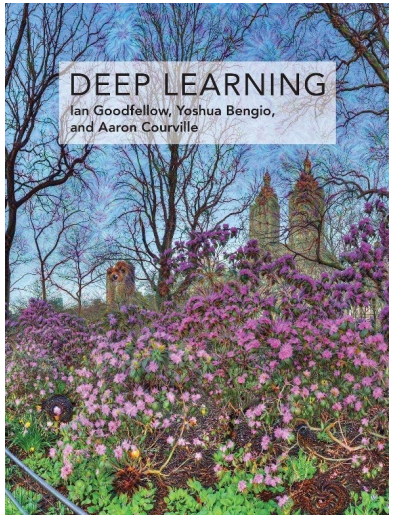
Second Wave

Backpropagation [Rummelhart, Hinton, Williams, 1986]

Gradient diffusion or explosion:
Can not learn more than a few layers

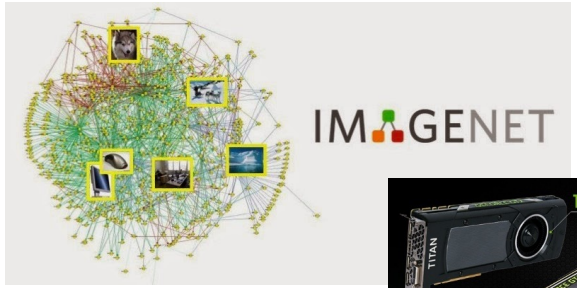
Deep Learning





Third Wave

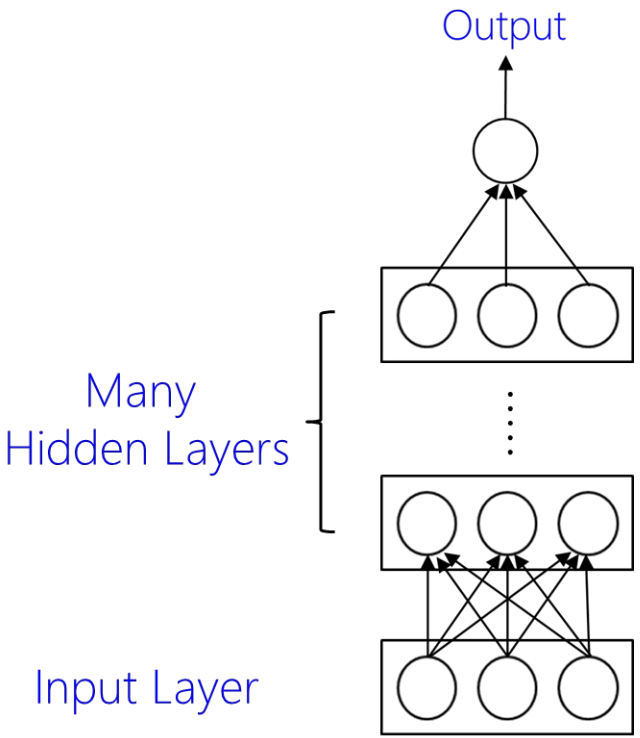
SGD, ReLU, dropout, …
[Hinton, LeCun, Bengio, Schmidhuber, Hochreiter, …]

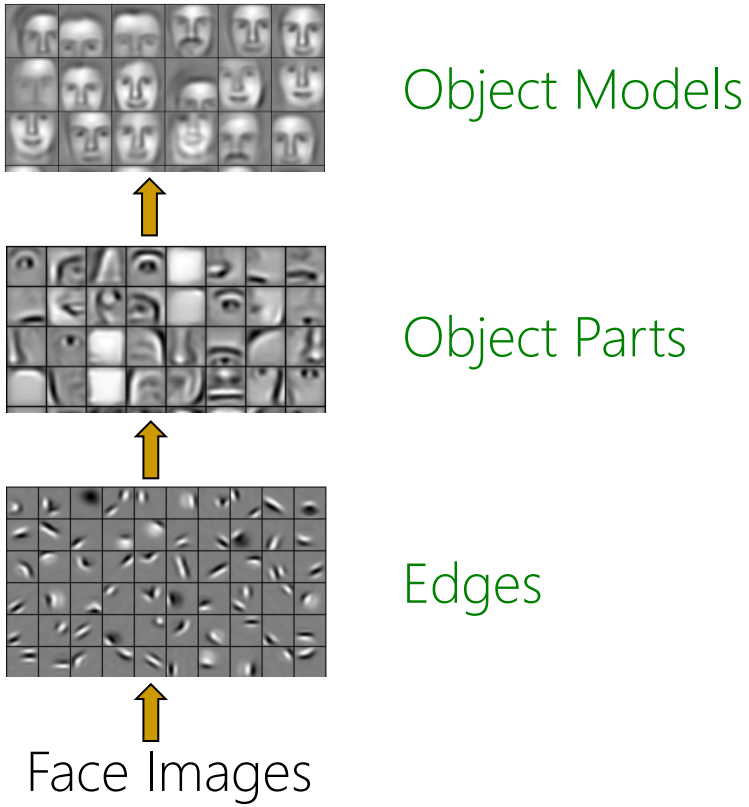


Big labeled data

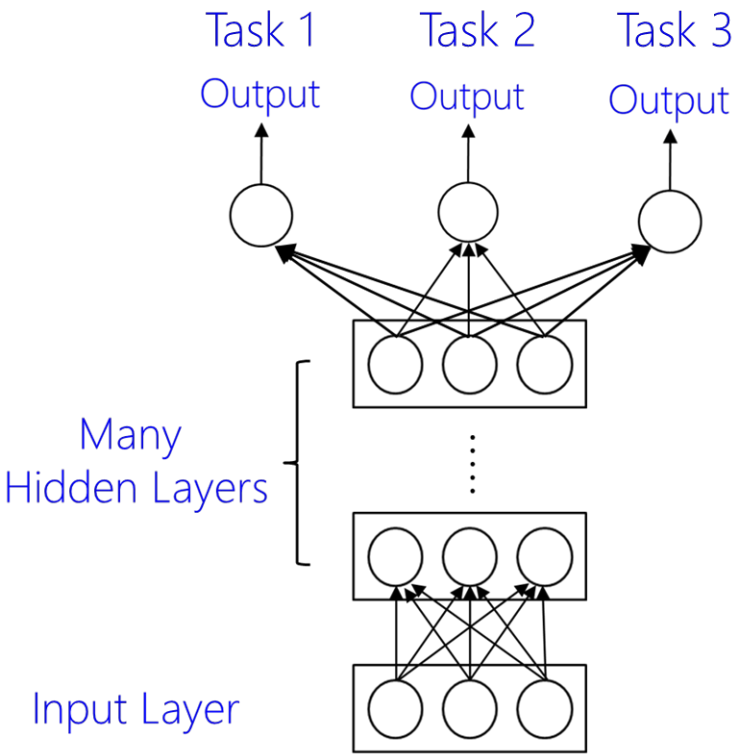
Fast computation

Promise: Representation Learning





Promise: Transfer Learning



E.g.: Chest X-rays v.s. ImageNet

Biomedical NLP: The Challenges

Challenge: Variations

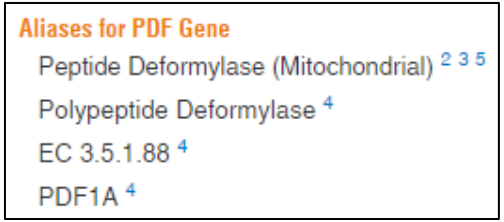
TP53 inhibits BCL2.
Tumor suppressor P53 down-regulates the activity of BCL-2 proteins.
BCL2 transcription is suppressed by P53 expression.
The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 …
……

negative regulation

532 inhibited, 252 inhibition, 218 inhibit, 207 blocked, 175 inhibits, 157 decreased, 156 reduced, 112 suppressed, 108 decrease, 86 inhibitor, 81 Inhibition, 68 inhibitors, 67 abolished, 66 suppress, 65 block, 63 prevented, 48 suppression, 47 blocks, 44 inhibiting, 42 loss, 39 impaired, 38 reduction, 32 down-regulated, 29 abrogated, 27 prevents, 27 attenuated, 26 repression, 26 decreases, 26 down-regulation, 25 diminished, 25 downregulated, 25 suppresses, 22 interfere, 21 absence, 21 repress ……

Challenge: Ambiguity

In eubacteria and eukaryotic organelles the product of this gene, peptide deformylase (PDF), removes the formyl group from the initiating methionine of nascent peptides. …… The discovery that a natural inhibitor of PDF, actinonin, acts as an antimicrobial agent in some bacteria has spurred intensive research into the design of bacterial-specific PDF inhibitors. …… In humans, PDF function may therefore be restricted to rapidly growing cells.



Challenge: Document-Level Complex Relation

"We next expressed ALK F1174L, ALK F1174L/L1198P, ALK F1174L/G1123S, and ALK F1174L/G1123D in the original SH-SY5Y cell line."

(. . . 15 sentences and 2 figures . . .)

"The 2 mutations that were only found in the neuroblastoma resistance screen (G1123S/D) are located in the glycine-rich loop, which is known to be crucial for ATP and ligand binding and are the first mutations described that induce resistance to TAE684, but not to PF02341066."

Challenge: Cross-Note Complex Extraction

Pathology:

LEFT BREAST, NEEDLE CORE BIOPSIES:
- Positive for high-grade ductal carcinoma in situ, pending

Imaging:

TECHNIQUE: Spot magnification views in the craniocaudal and medial lateral projection was obtained at the level of the nipple and retroareolar area.
FINDINGS: There are pleomorphic calcifications located in the left retroareolar region approximately 2 cm from the nipple. These extend over approximately 2 cm. These are not clearly appreciated on prior studies. These are suspicious. Biopsy will be necessary for further evaluation. These are at an anterior depth.

Tumor Site: ???
Histology: ???

Challenge: Annotation Bottleneck

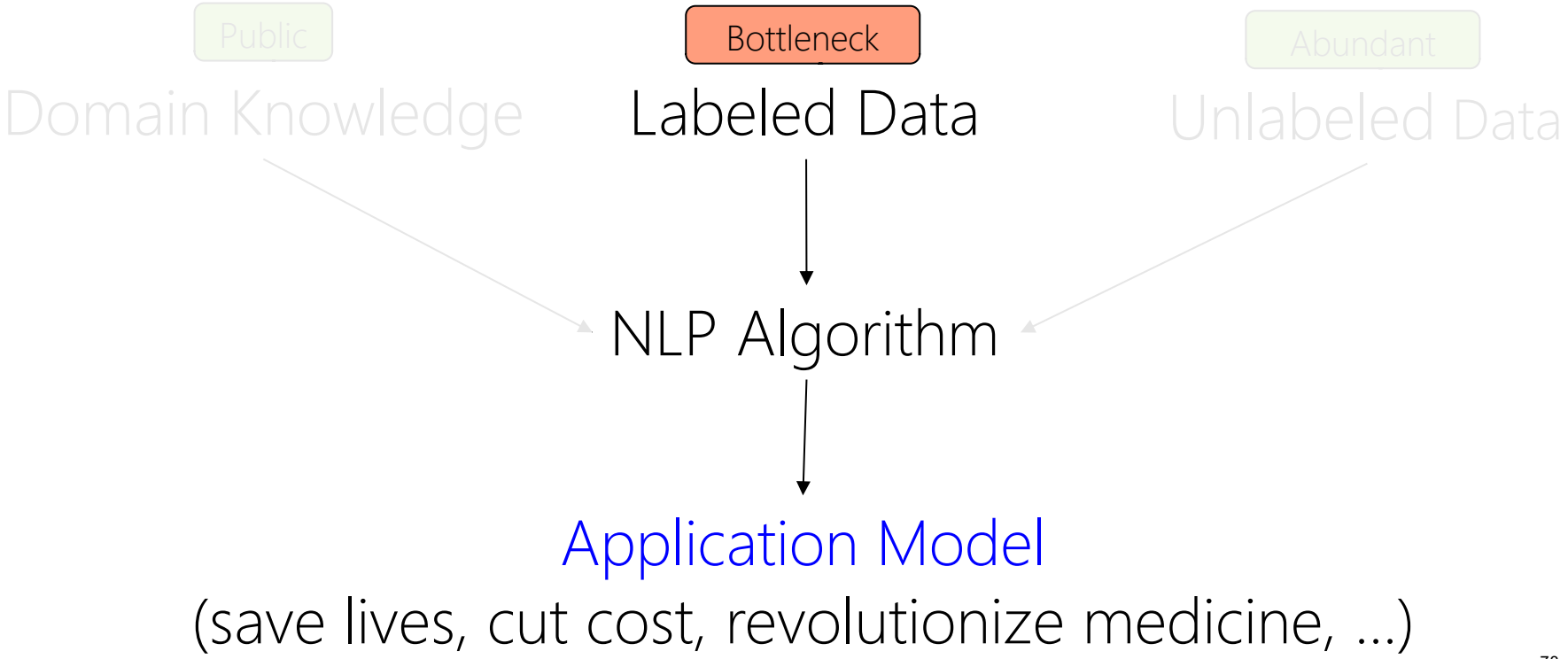
Deep learning requires many labeled examples

Hire experts to label: Not scalable

Crowdsource: Lack domain expertise

Self-Supervised Learning

Supervised Learning



What's the alternative?

Many Candidates ……

Distant supervision

Weak supervision

Incidental supervision

Semi-supervised learning

Indirect supervision

Unsupervised learning

Distant Supervision: A simple example [Mintz et al. 2009]

Distant Supervision

NCI Pathway KB

Regulation	Theme	Cause
Positive	A2M	FOXO1
Positive	ABCB1	TP53
Negative	BCL2	TP53
…	…	…

Distant Supervision

NCI Pathway KB

Regulation	Theme	Cause
Positive	A2M	FOXO1
Positive	ABCB1	TP53
Negative	BCL2	TP53
…	…	…

Distant Supervision

NCI Pathway KB

Regulation	Theme	Cause
Positive	A2M	FOXO1
Positive	ABCB1	TP53
Negative	BCL2	TP53
…	…	…

TP53 inhibits BCL2.
Tumor suppressor P53 down-regulates the activity of BCL-2 proteins.
BCL2 transcription is suppressed by P53 expression.
The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 …
……

Distant Supervision

NCI Pathway KB

Regulation	Theme	Cause
Positive	A2M	FOXO1
Positive	ABCB1	TP53
Negative	BCL2	TP53
…	…	…

TP53 inhibits BCL2.
Tumor suppressor P53 down-regulates the activity of BCL-2 proteins.
BCL2 transcription is suppressed by P53 expression.
The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 …
……

Distant Supervision

NCI Pathway KB

Regulation	Theme	Cause
Positive	A2M	FOXO1
Positive	ABCB1	TP53
Negative	BCL2	TP53
…	…	…

TP53 inhibits BCL2.
Tumor suppressor P53 down-regulates the activity of BCL-2 proteins.
BCL2 transcription is suppressed by P53 expression.
The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 …
……

Weak Supervision

Distant supervision ++

- Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations [Hoffman et al. 2011]
- Multi-instance learning

Data programming

- "noisy, limited, or imprecise sources are used to provide supervision signal"
- Stanford Chris Re & Gang (e.g., [Ratner et al. 2017])
- Labeling functions → Model instances I.I.D.

Incidental Supervision

Dan Roth [AAAI 2017]

- Defined by example: “Leveraging the international news cycle to learn transliteration models for named entities”
- Framed as distinct from distant supervision

Semi-Supervised Learning

Canonical setup: (small) labeled + (large) unlabeled

Graph-based

- Label propagation
- Similarity metric

Self-Training

- First: Syntactic parsing (PTB → Brown); e.g., McClosky [2007]
- Effectively a form of hard EM

Indirect Supervision

Opposite of “direct supervision” (i.e., supervised)

Precise, but excluding labels when available

Yann LeCun's cake

- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Unsupervised/Predictive Learning (cake)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

- **(Yes, I know, this picture is slightly offensive to RL folks. But I'll make it up)**

Unsupervised Learning

Commonly associated with clustering

Misnomer: Need supervision, just not direct

ML101: No-Free-Lunch Theorem

Any learning requires inductive bias (“supervision”)

Evolution of Yann LeCun's Cake

Unsupervised learning

Predictive learning

Self-supervised learning

How Much Information is the Machine Given during Learning?

- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

Yann LeCun's cake V2

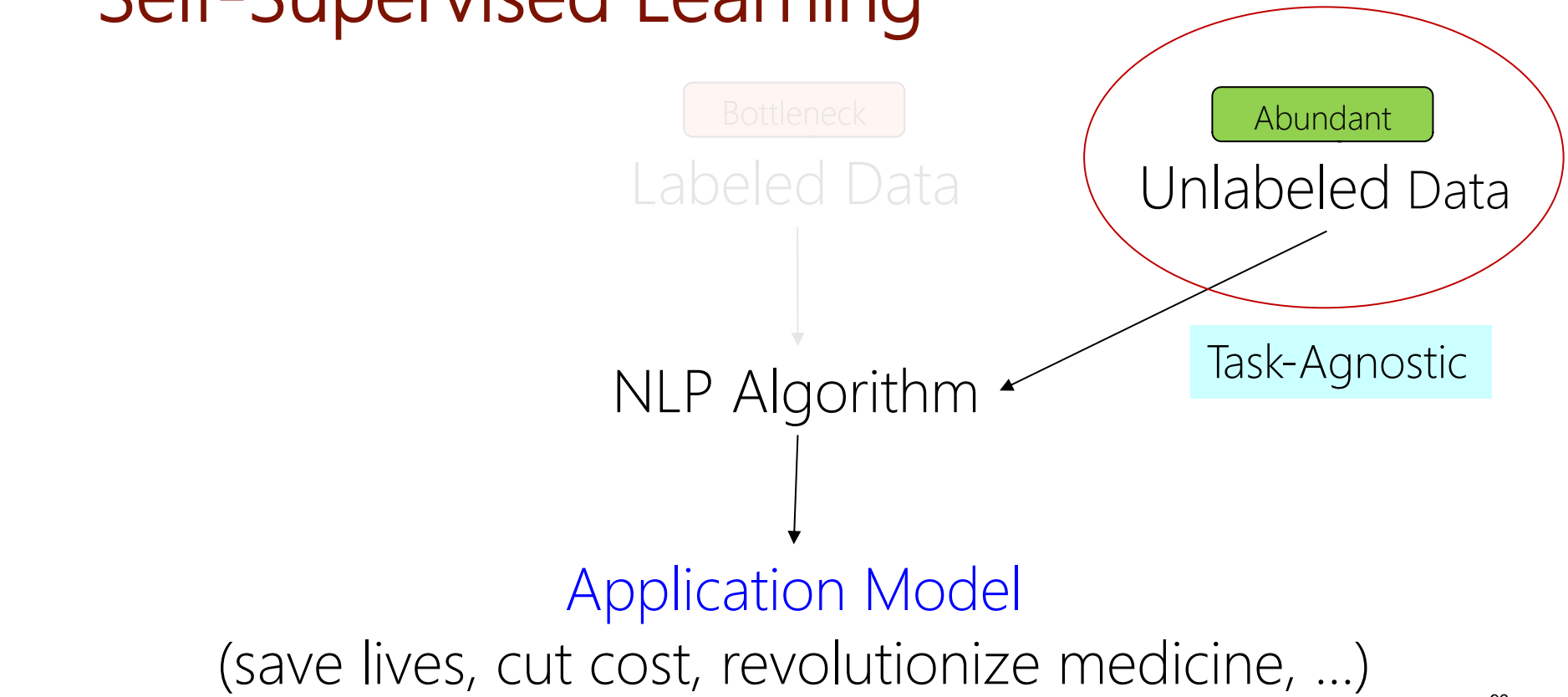
Self-Supervised Learning

Self-training = Early special form

Current: Pretraining (task-agnostic)

- NLP: BERT, RoBERTa, GPT3, ELECTRA, UniLM, …
- Vision: GAN, AutoEncoder, SEER
- VLP: Contrastive Learning; CLIP

Self-Supervised Learning



Self-Supervised Learning

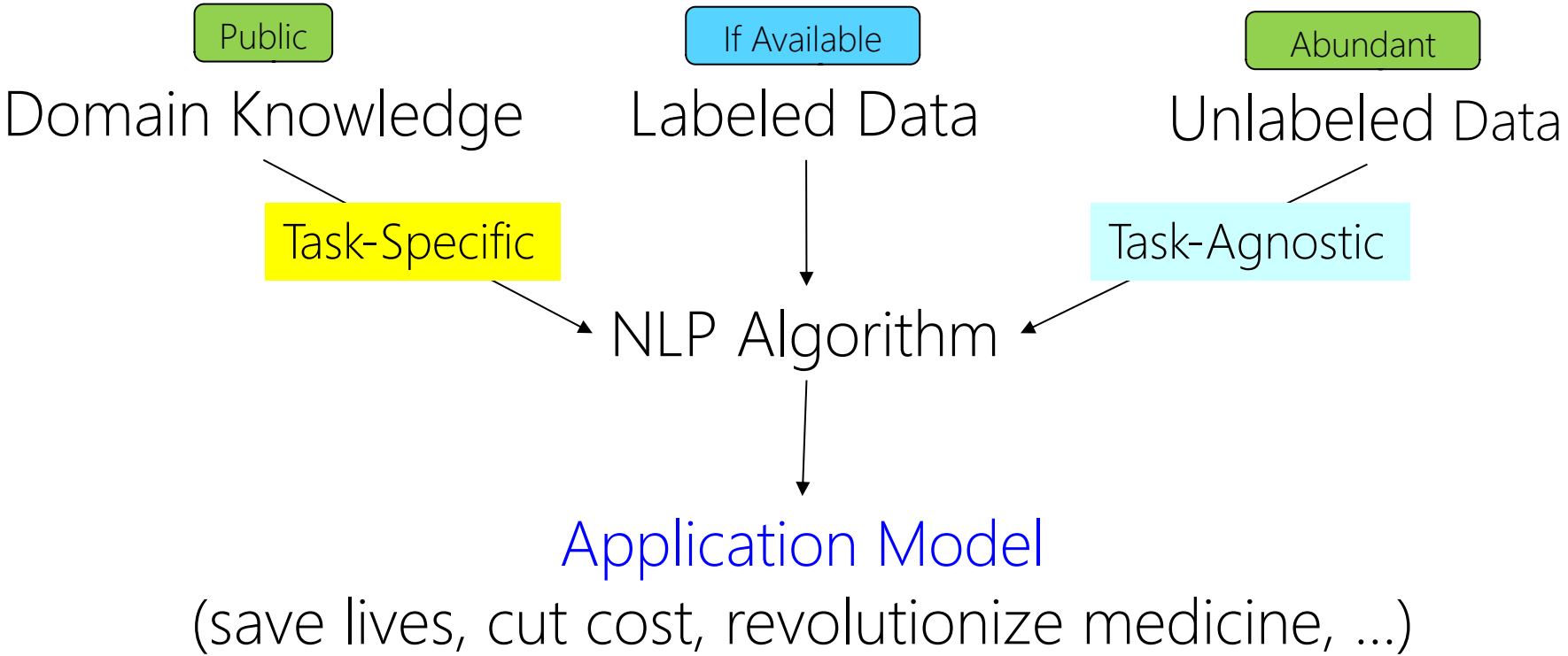
Self-training = Early special form

Current: Pretraining (task-agnostic)

New: Unified view of supervision

- Both task-agnostic and task-specific
- Domain knowledge, expert heuristics, …
- Self choice of supervision sources & composition

Self-Supervised Learning



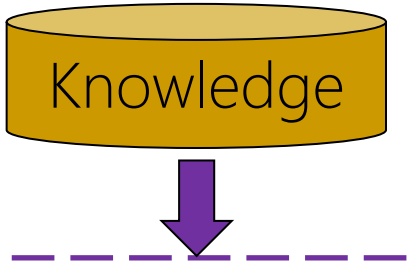
Self-Supervised Learning

Mission: Empower domain experts to efficiently develop task-specific machine-learning systems

Key idea: Admit & compose diverse forms of supervision

- User: ML experts → Domain experts
- Supervision: Direct → Self-supervised (scalable)
- Primitives: low-level (python) → high-level (domain-aware)
- Training: Static → Interactive

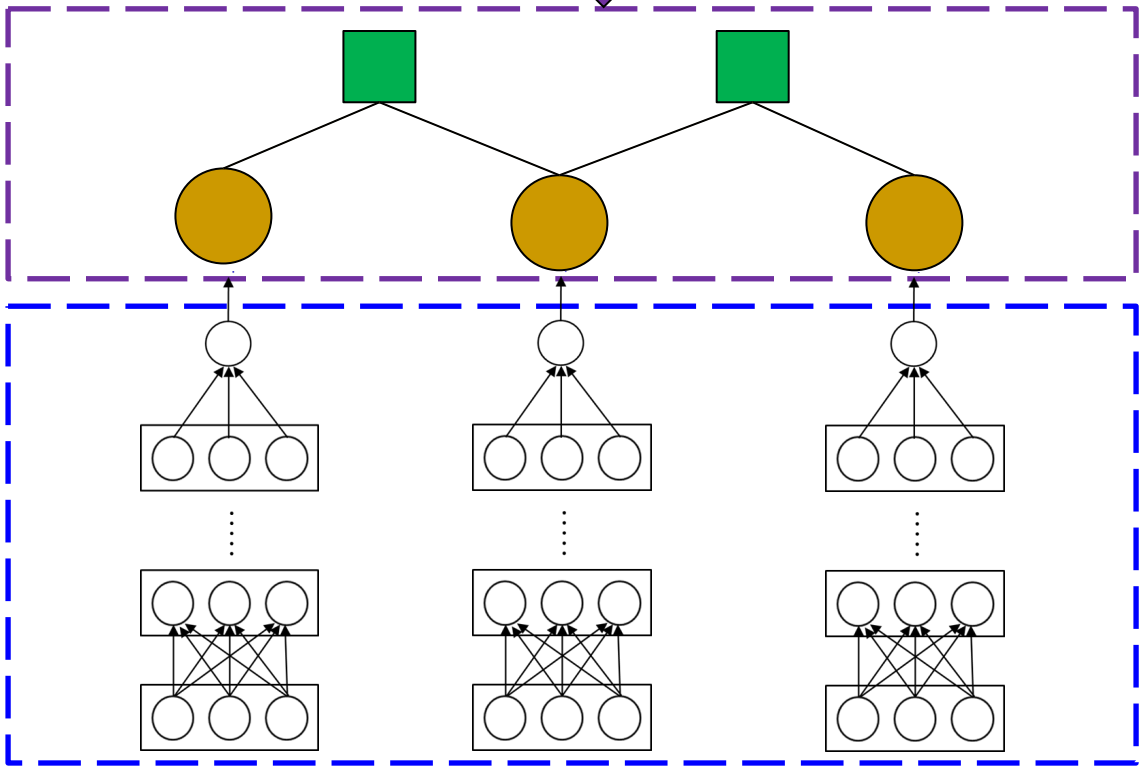
Wang & Poon. "Deep Probabilistic Logic: A Unifying Framework for Indirect Supervision", *EMNLP-18*.



Deep Probabilistic Logic

Virtual Evidence

Latent Variable



Probabilistic Logic

Deep Learning

Related Directions

AutoML

Active Learning

AutoML

Automate selection, composition, parametrization of ML models

E.g., hyperparameter tuning (Bayes Opt, NAS)

Complimentary to supervision sources

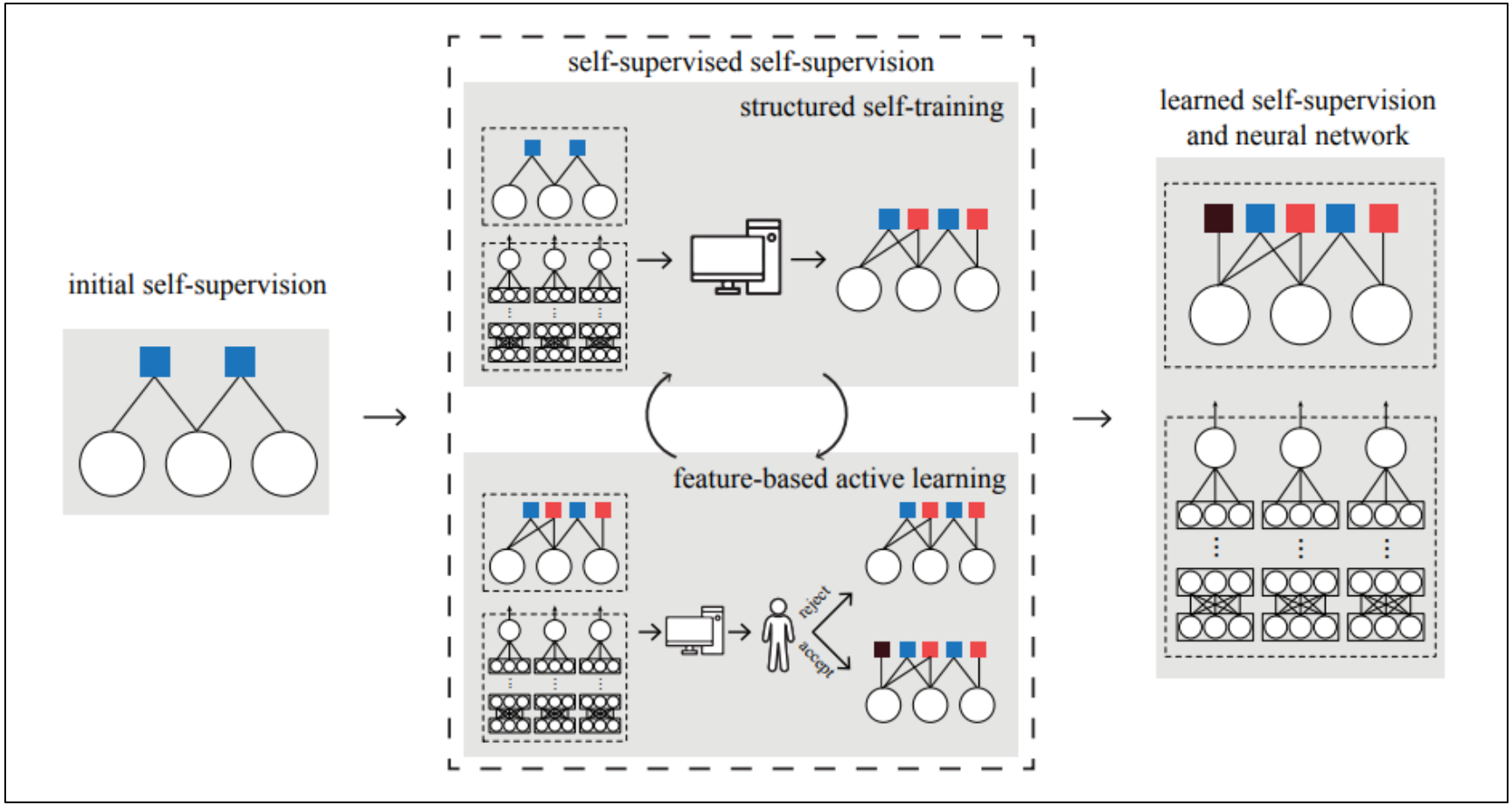
Active Learning

Still focus on acquiring instance-level labeling

E.g., select instance w. max info gain to label

New frontier: Active Self-Supervised Learning

Generate/validate new self-supervision



Lang & Poon. "Self-Supervised Self Supervision by Combining Deep Learning and Probabilistic Logic", *AAAI-21*.

Self-Supervised Learning

Public

Bottleneck

Abundant

Domain Kr

labeled Data

Free lunch is great

Knowledge is power

Less can be more

Application Model

(save lives, cut cost, revolutionize medicine, …)

Domain-Specific Pretraining

Neural Language Model Pretraining

The 2 mutations that were only found in the neuroblastoma resistance screen (G1123S/D) are located in the glycine-rich loop, which is known to be crucial for ATP and ligand binding and are the first mutations described that induce resistance to TAE684, but not to PF02341066

Unlabeled text

Neural Language Model Pretraining

The 2 mutations that were only found in the [MASK] resistance screen (G1123S/D) are [MASK] in the glycine-rich loop, which is known to be [MASK] for ATP and ligand [MASK] and are the first mutations described that induce resistance to TAE684, but not to [MASK]

Masked Language Model

Neural Pretraining: Key Ideas

Contextualized representation

Text completion

Self Attention

Word piece

Contextualized Representation

Early: Learning multiple word-sense vectors

ELMo: Bi-LSTM (slow)

BERT: Use self-attention (Transformer)

Text Completion

Probabilistic models

- ELMo / GPT2: Generative (~ HMM)
- BERT: Discriminative (~ Word2vec, GloVe)

Masking

- Past: Single word (~ Cloze)
- BERT: Random subset

Self-Attention

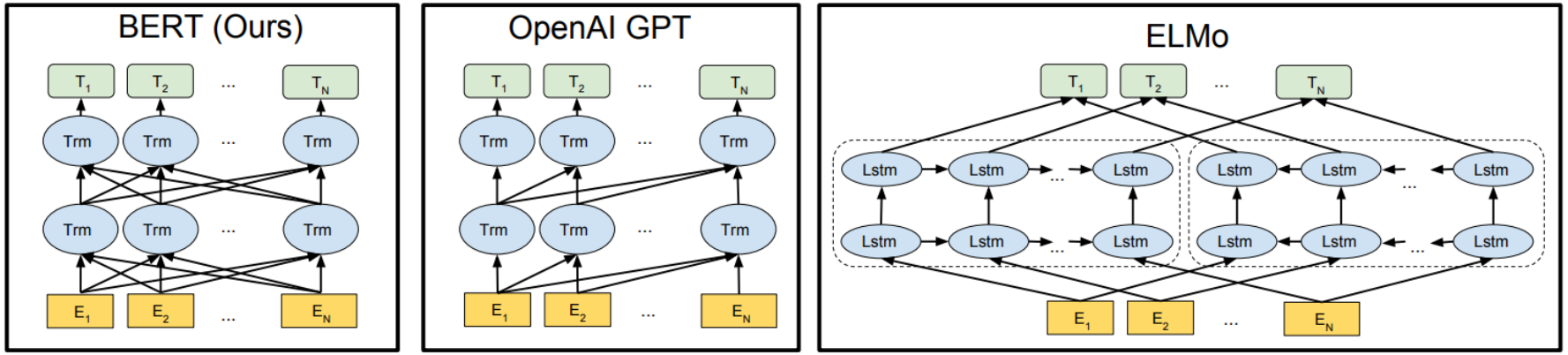


Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks. Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. *NAACL 2019*

Word Piece

Out-of-vocabulary (OOV)

Character n-gram vs word

Idea: Start with characters, add frequent n-grams

E.g., Byte Pair Encoding (BPE)

Note: Domain-specific, corpus-dependent

Neural Pretraining

General Domain

BERT: Base 110M, Large 340M; 12G
GPT2: 1.5B; 40G
T5: 11B; 750G
Turing: 17B / GPT3: 170B

GLUE
SQuAD
MARCO
……

Text | Language Model | Evaluation

Neural Pretraining

Biomed NLP Underexplored

Biomedical NLP

General Domain

Mixed-Domain

BioBERT
SciBERT
Clinical BERT

???

BERT: Base 110M, Large 340M; 12G
GPT2: 1.5B; 40G
T5: 11B; 750G
Turing: 17B / GPT3: 170B

GLUE
SQuAD
MARCO
……

Text

Language Model

Evaluation

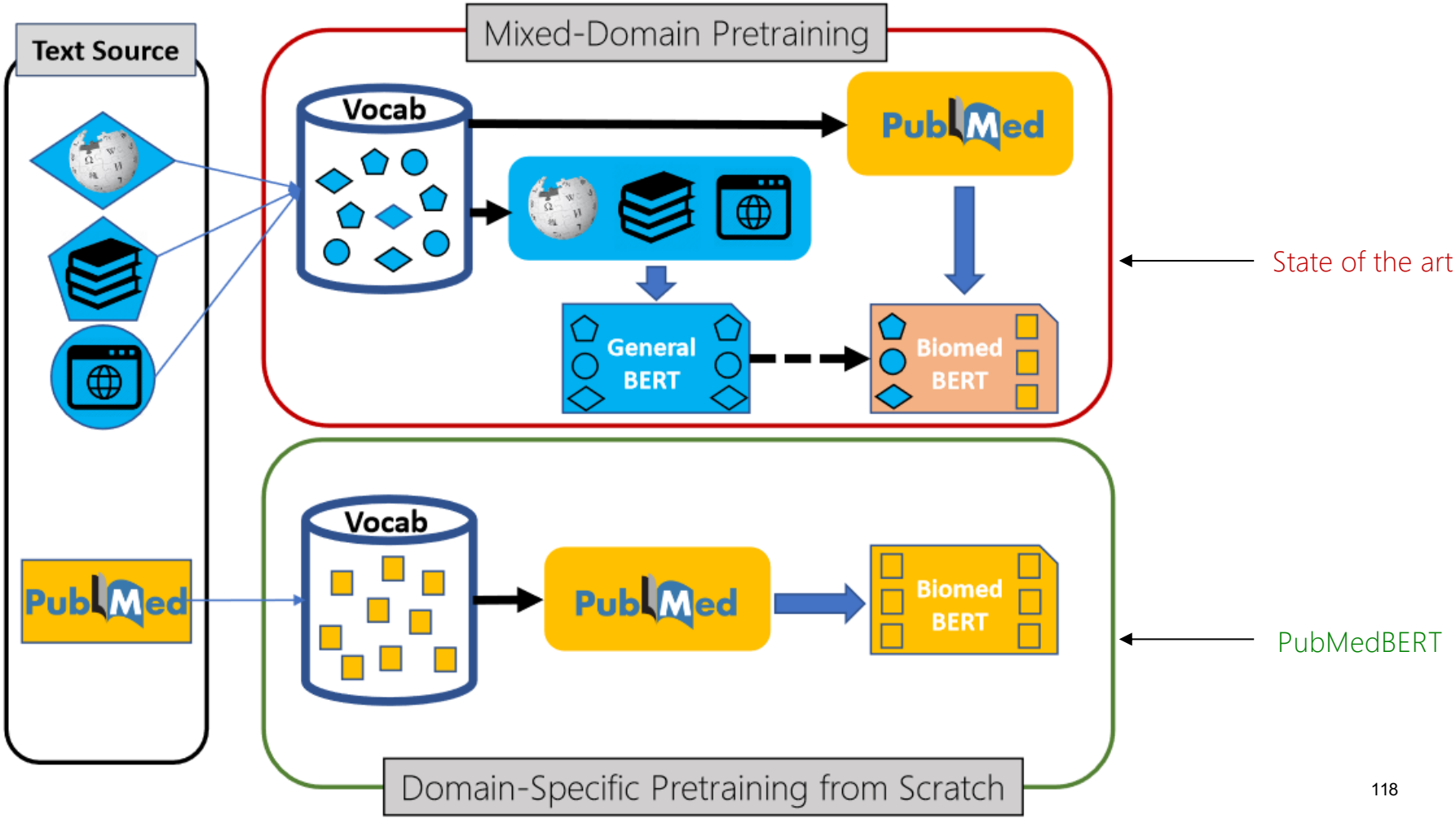
Is more data always better?

Biomed-Specific Pretraining

Biomedical Term	Category	BERT	SciBERT	PubMedBERT (Ours)
diabetes	disease	✓	✓	✓
leukemia	disease	✓	✓	✓
lithium	drug	✓	✓	✓
insulin	drug	✓	✓	✓
DNA	gene	✓	✓	✓
promoter	gene	✓	✓	✓
hypertension	disease		✓	✓
nephropathy	disease		✓	✓
lymphoma	disease		✓	✓
lidocaine	drug		✓	✓
oropharyngeal	organ			✓
cardiomyocyte	cell			✓
chloramphenicol	drug			✓
RecA	gene			✓
acetyltransferase	gene			✓
clonidine	drug			✓
naloxone	drug			✓

lymphoma → l, ##ym, ##ph, ##oma

acetyltransferase → ace, ##ty, ##lt, ##ran, ##sf, ##eras, ##e


Text Source
Mixed-Domain Pretraining
Vocab
PubMed
General BERT
Biomed BERT
State of the art
PubMed
Vocab
PubMed
Biomed BERT
PubMedBERT
Domain-Specific Pretraining from Scratch

PubMedBERT

Yu et al. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing", *ACM Special Issue 2021*.

BLURB: Biomed NLP Benchmark

	BERT		RoBERTa	BioBERT	SciBERT		ClinicalBERT	BlueBERT	PubMedBERT
	uncased	cased	cased	cased	uncased	cased	cased	cased	uncased
BC5-chem	89.25	89.99	89.43	92.85	92.49	92.51	90.80	91.19	**93.33**
BC5-disease	81.44	79.92	80.65	84.70	84.54	84.70	83.04	83.69	**85.62**
NCBI-disease	85.67	85.87	86.62	**89.13**	88.10	88.25	86.32	88.04	87.82
BC2GM	80.90	81.23	80.90	83.82	83.36	83.36	81.71	81.87	**84.52**
JNLPBA	77.69	77.51	77.86	78.55	78.68	78.51	78.07	77.71	**79.10**
EBM PICO	72.34	71.70	73.02	73.18	73.12	73.06	72.06	72.54	**73.38**
ChemProt	71.86	71.54	72.98	76.14	75.24	75.00	72.04	71.46	**77.24**
DDI	80.04	79.34	79.52	80.88	81.06	81.22	78.20	77.78	**82.36**
GAD	77.72	77.28	77.72	80.94	80.90	79.66	78.40	77.24	**82.34**
BIOSSES	82.68	81.40	81.25	89.52	86.25	87.15	91.23	85.38	**92.30**
HoC	80.20	80.12	79.66	81.54	80.66	81.16	80.74	80.48	**82.32**
PubMedQA	51.62	49.96	52.84	**60.24**	57.38	51.40	49.08	48.44	55.84
BioASQ	70.36	74.44	75.20	84.14	78.86	74.22	68.50	68.71	**87.56**
BLURB score	75.96	75.73	76.30	80.26	78.78	78.05	77.17	76.16	**81.07**

PubMedBERT

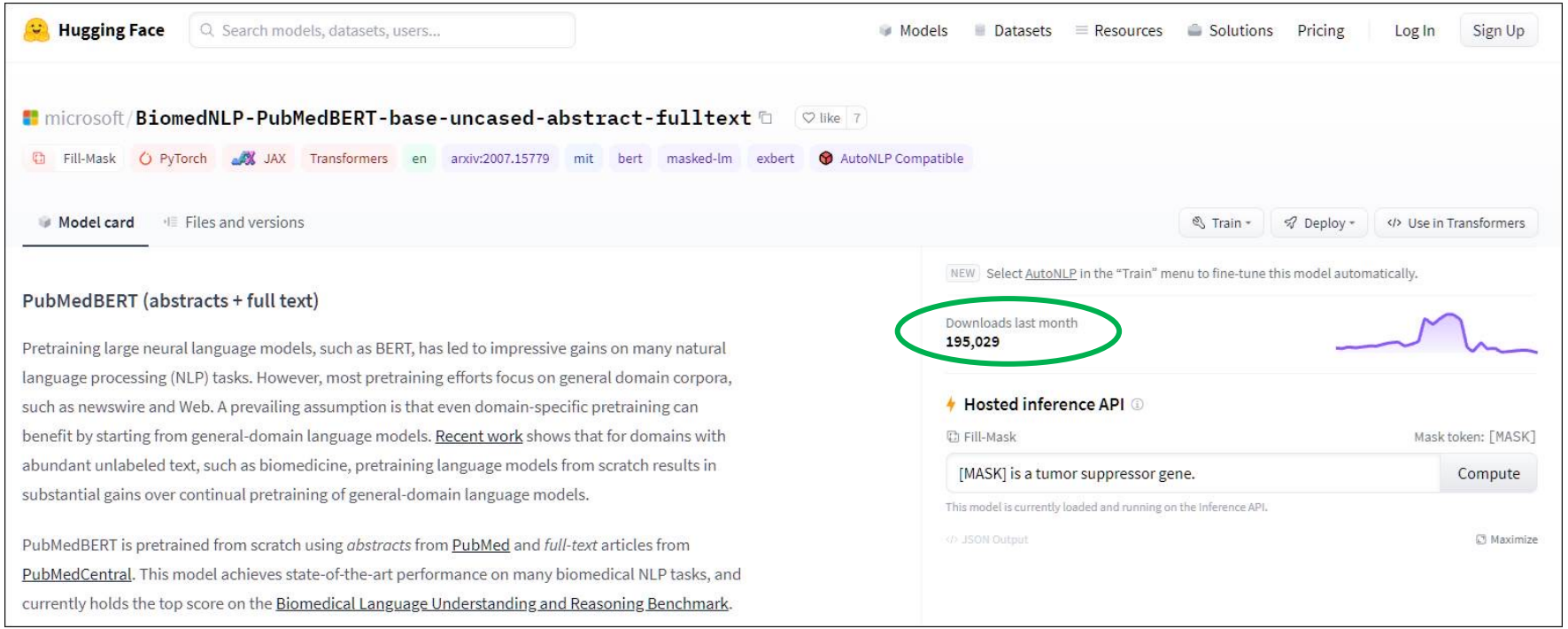
Wang et al. "Domain-Specific Pretraining for Vertical Search", *KDD-21*.

Model	NDCG@10	P@5
BERT	55.0 (±1.2)	63.4 (±2.3)
RoBERTa	53.5 (±1.6)	61.1 (±2.3)
UNILM	55.0 (±1.2)	62.0 (±1.8)
SciBERT	58.9 (±1.5)	67.7 (±2.2)
PubMedBERT	61.5 (±1.1)	69.5 (±1.8)
PubMedBERT-COVID	65.6 (±1.0)	73.2 (±1.1)

Biomedical Search

Domain-specific LMs attained top performance at TREC-COVID

Monthly Downloads: Hundreds of Thousands



OncoBERT: Oncology RWE

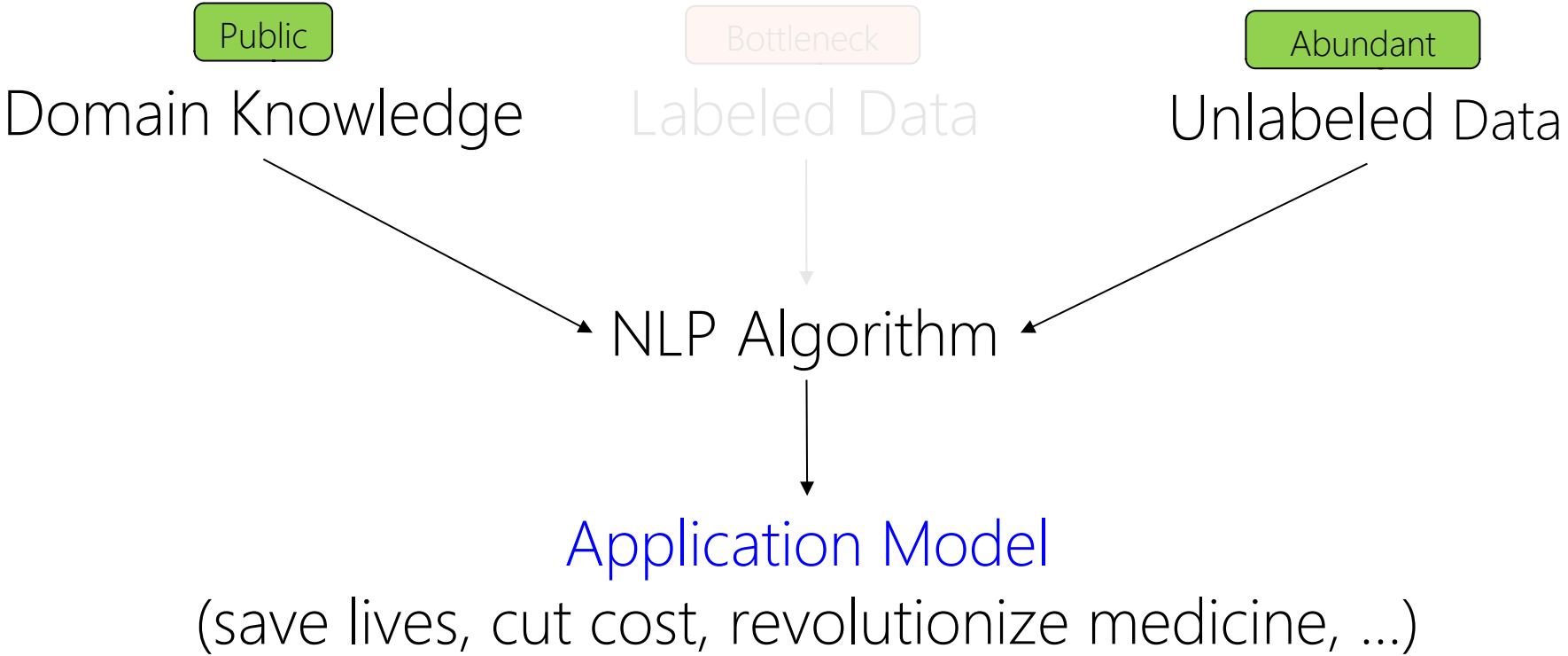
Providence

	Tumor Site	Histology	Clinical T	N	M	Pathological T	N	M
Ontology	19.4	19.2	-	-	-	-	-	-
BOW	62.8	76.6	70.4	96.6	98.4	72.1	90.7	**98.9**
OncoGloVe + CNN	72.0	84.4	74.2	96.5	98.6	83.9	93.1	98.5
OncoGloVe + HAN/GRU	74.0	85.9	76.2	97.1	98.7	86.4	94.2	98.5
BERT + HAN/GRU	75.1	86.2	77.0	96.6	98.4	86.4	94.4	98.2
PubMedBERT + HAN/GRU (ours)	76.7	87.2	79.3	97.2	98.7	87.2	95.2	98.6
OncoBERT + HAN/GRU (ours)	**77.1**	**87.6**	**81.4**	**97.5**	**99.0**	**87.6**	**95.5**	**98.9**

Preston, Wei, et al. "Towards Structuring Real-World Data at Scale: Deep Learning for Extracting Key Oncology Information from Clinical Text with Patient-Level Supervision", *in submission*.

Knowledge-Rich Self-Supervision

Self-Supervised Learning



Example: Distant Supervision

NCI Pathway KB

Regulation	Theme	Cause
Positive	A2M	FOXO1
Positive	ABCB1	TP53
Negative	BCL2	TP53
…	…	…

TP53 inhibits BCL2.

Tumor suppressor P53 down-regulates the activity of BCL-2 proteins.

BCL2 transcription is suppressed by P53 expression.

The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 …

……

Example: Joint Inference

In eubacteria and eukaryotic organelles the product of this gene, peptide deformylase (PDF), removes the formyl group from the initiating methionine of nascent peptides. …… The discovery that a natural inhibitor of PDF, actinonin, acts as an antimicrobial agent in some bacteria has spurred intensive research into the design of bacterial-specific PDF inhibitors. …… In humans, PDF function may therefore be restricted to rapidly growing cells.

Coreferent

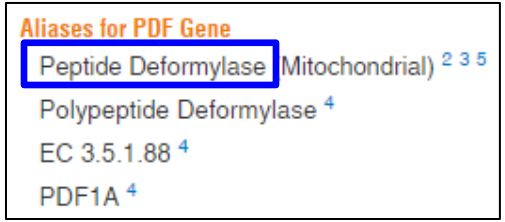
Example: Joint Inference

In eubacteria and eukaryotic organelles the product of this gene, peptide deformylase (PDF), removes the formyl group from the initiating methionine of nascent peptides. …… The discovery that a natural inhibitor of PDF, actinonin, acts as an antimicrobial agent in some bacteria has spurred intensive research into the design of bacterial-specific PDF inhibitors. …… In humans, PDF function may therefore be restricted to rapidly growing cells.

Apposition

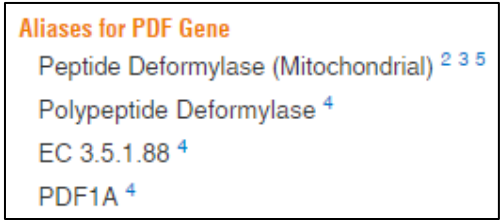
Example: Joint Inference

In eubacteria and eukaryotic organelles the product of this gene, peptide deformylase (PDF), removes the formyl group from the initiating methionine of nascent peptides. …… The discovery that a natural inhibitor of PDF, actinonin, acts as an antimicrobial agent in some bacteria has spurred intensive research into the design of bacterial-specific PDF inhibitors. …… In humans, PDF function may therefore be restricted to rapidly growing cells.



Example: Joint Inference

In eubacteria and eukaryotic organelles the product of this gene, peptide deformylase (PDF), removes the formyl group from the initiating methionine of nascent peptides. …… The discovery that a natural inhibitor of PDF, actinonin, acts as an antimicrobial agent in some bacteria has spurred intensive research into the design of bacterial-specific PDF inhibitors. …… In humans, PDF function may therefore be restricted to rapidly growing cells.



Case Study: Immunotherapy RWE

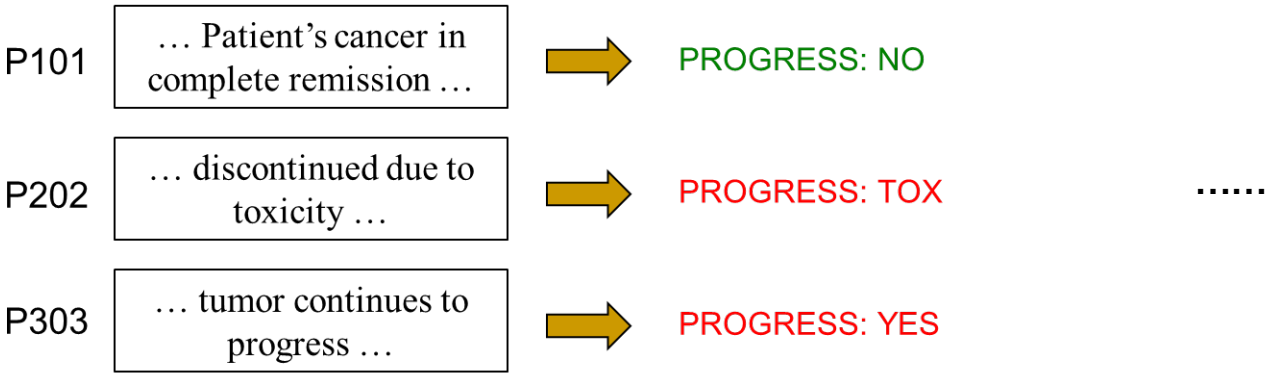
Given Keytruda cohort, find exceptional responder

Need to extract progression events

- "Patient's cancer in complete remission …"
- "… treatment was discontinued due to toxicity …"
- "… tumor continues to progress despite treatment …"

Case Study: Keytruda RWE

What's best way to leverage domain experts?



Supervised ML: Label many, many examples; need many, many hours …

Case Study: Keytruda RWE

What's best way to leverage domain experts?

Supervised ML: Label many, many examples; need many, many hours …

Example: Data Programming

What's best way to leverage domain experts?

If Patient's progress note contains words like "remission", "progression free", label as PROGRESS:NO.	Confidence: Medium
If Patient's medication table specifies discontinuation, label as PROGRESS:TOX.	Confidence: High
If Patient's radiology report contains words like "advance", "grow", label as PROGRESS:YES.	Confidence: Low

Self-supervised learning: Identify a few task-specific "rules"; easy to do

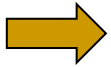
Example: Data Programming

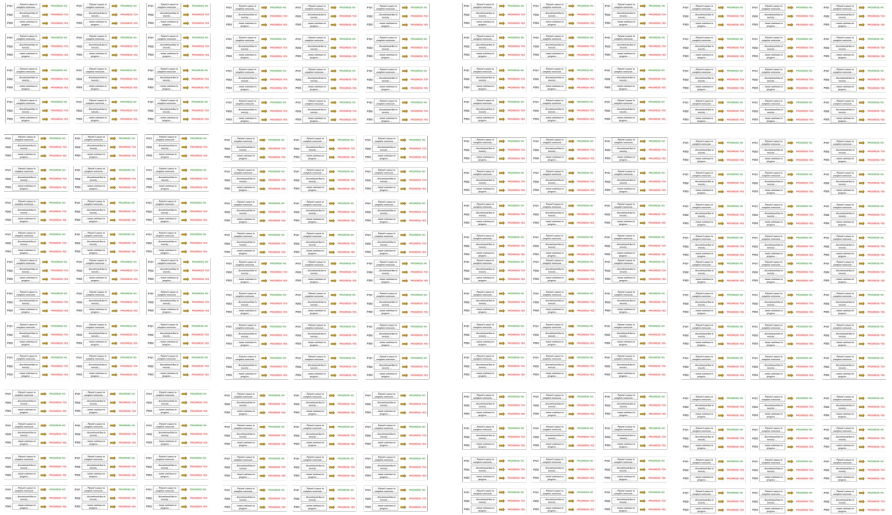
What's best way to leverage domain experts?

If Patient's progress note contains words like "remission", "progression free", label as PROGRESS:NO. — Confidence: Medium

If Patient's medication table specifies discontinuation, label as PROGRESS:TOX. — Confidence: High

If Patient's radiology report contains words like "advance", "grow", label as PROGRESS:YES. — Confidence: Low





...

+

Unlabeled EMRs

Self-supervised learning: Identify a few task-specific "rules"; easy to do

Deep Probabilistic Logic

Probabilistic Logic

Distant Supervision $f_{KB}(X_i, Y_i) = \mathbb{I}[\texttt{In-KB}(X_i, r) \wedge Y_i = r]$

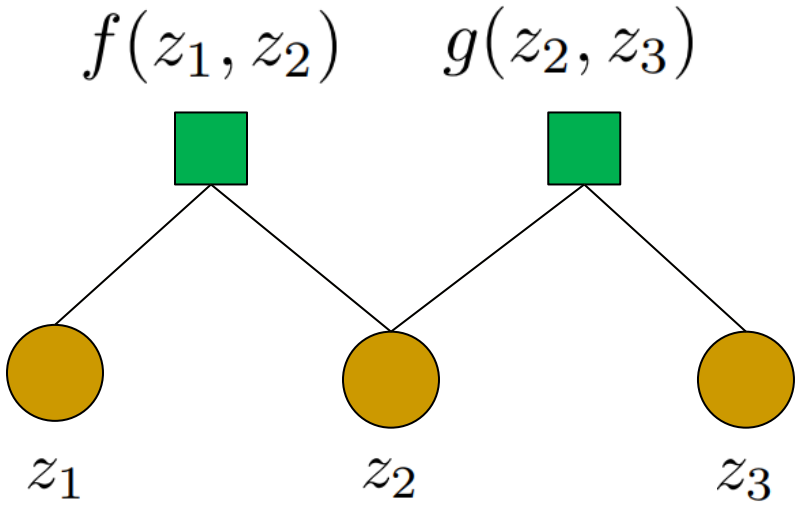
Data Programming $f_L(X_i, Y_i) = \mathbb{I}[L(X_i) = Y_i]$

Joint Inference $f_{\texttt{Joint}}(X_i, Y_i, X_j, Y_j) = \mathbb{I}[\texttt{Coref}(X_i, X_j) \wedge Y_i = Y_j]$

Probabilistic Logic

Factor

Variable



Probability $p(z_1, z_2, z_3) \propto \exp(w_f \cdot f(z_1, z_2) + w_g \cdot g(z_2, z_3))$

Challenge: Computation

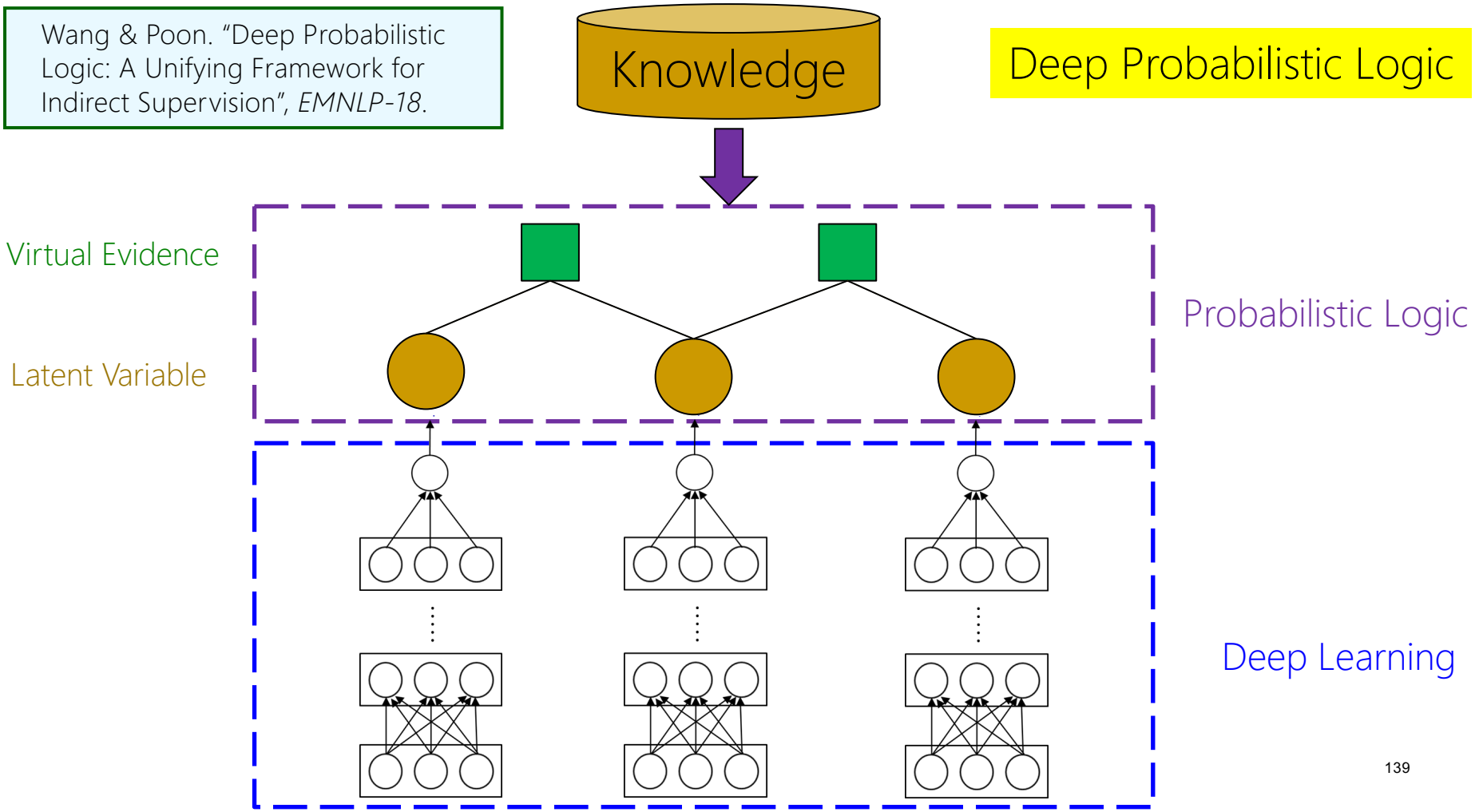
End-to-end modeling is generally intractable

Inference: Graphical model explodes in size

Learning: Require inference during iteration

Feature: Manually specified

Wang & Poon. "Deep Probabilistic Logic: A Unifying Framework for Indirect Supervision", *EMNLP-18*.



Virtual Evidence $P(v = 1|y = l)$

Bayesian-like prior (Pearl 1988, Reynolds & Bilmes 2005)

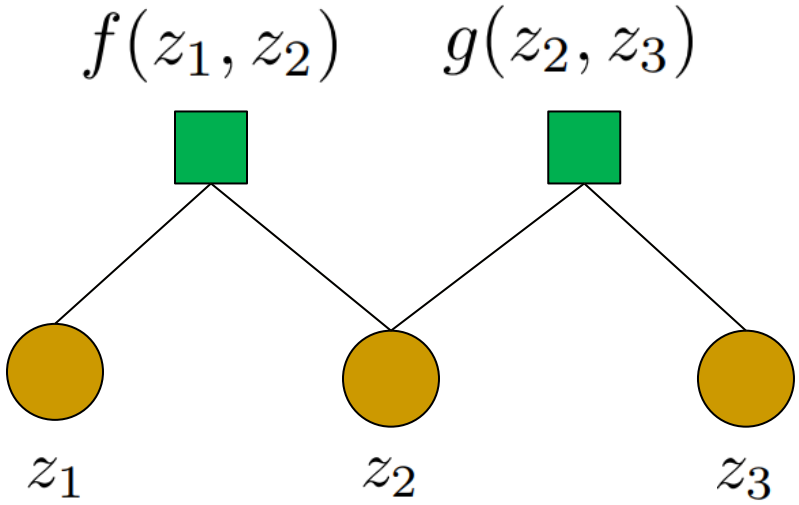
Generalize: variables → arbitrary factors

Learning maximizes conditional likelihood of virtual evidence

Probabilistic Logic

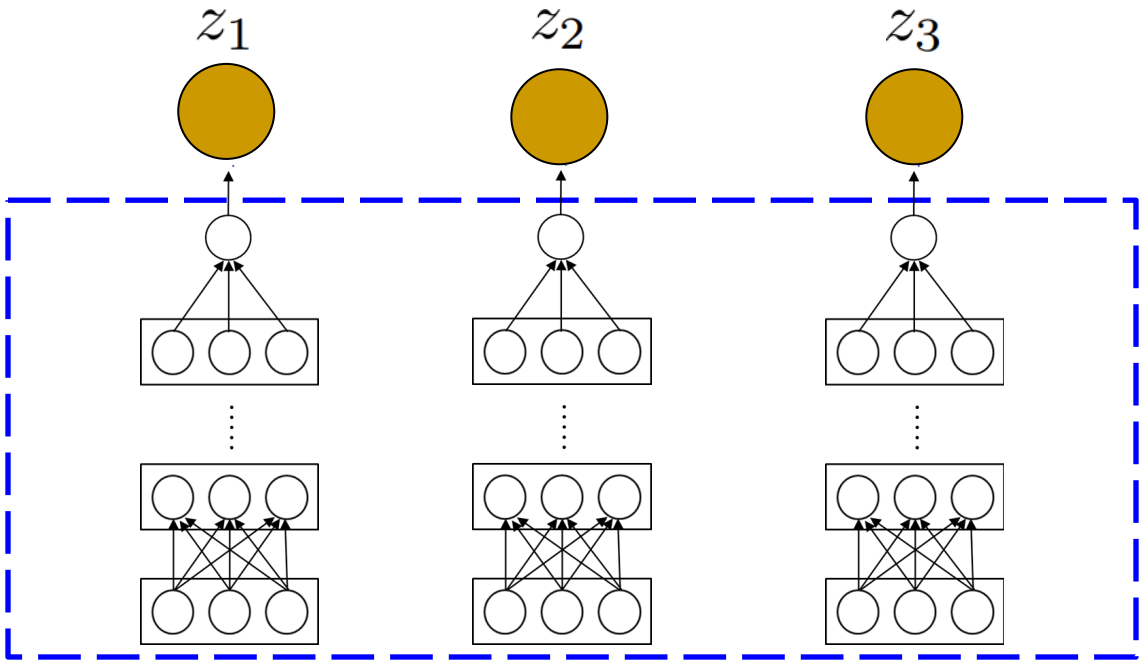
Factor

Variable



Probability $p(z_1, z_2, z_3) \propto \exp(w_f \cdot f(z_1, z_2) + w_g \cdot g(z_2, z_3))$

Deep Learning



Variational EM

Marginal $\sim p(z_1, z_2, z_3)$

Probabilistic Labels

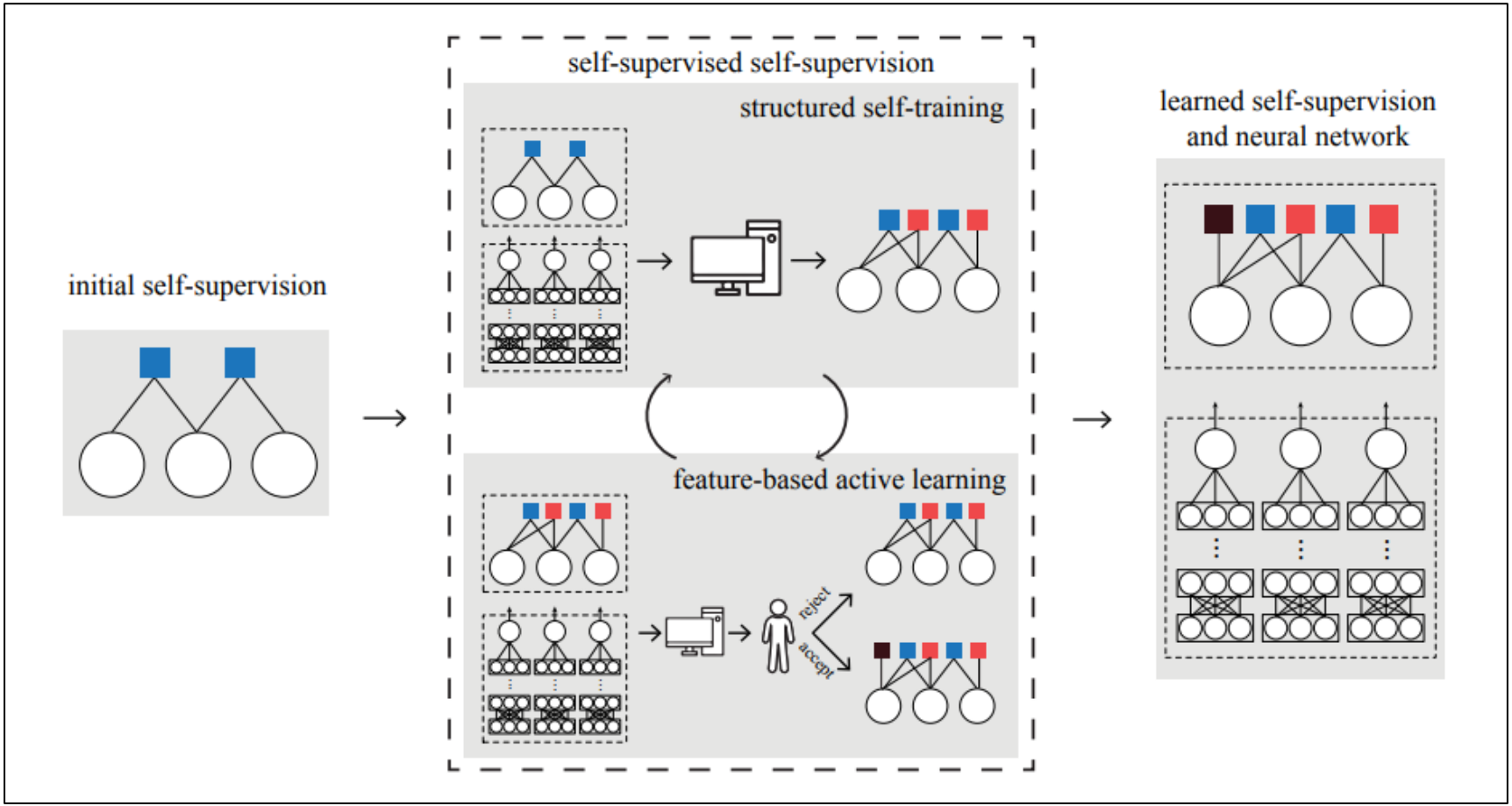
Related Work

Deep generative model [Kingma et al. 2014, Johnson et al. 2016]

Differentiable logic [Rocktaschel & Riedel 2017, Evans & Grefenstette 2018]

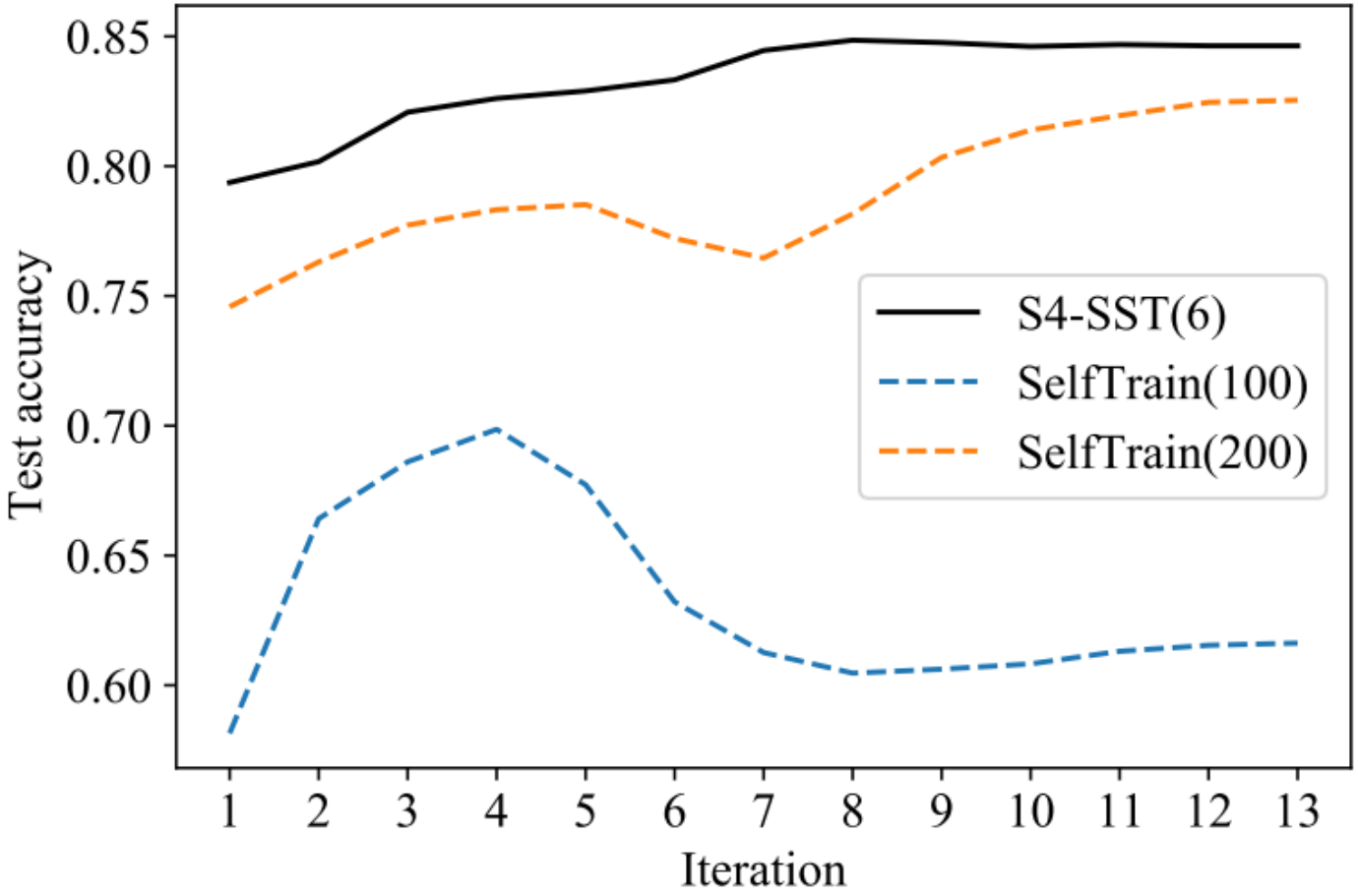
Knowledge compilation [Hu et al. 2016]

Self-Supervised Self-Supervision



Lang & Poon. "Self-Supervised Self Supervision by Combining Deep Learning and Probabilistic Logic", *AAAI-21*.

IMDb Test: Outperform self-training w. orders of magnitude more labeled data

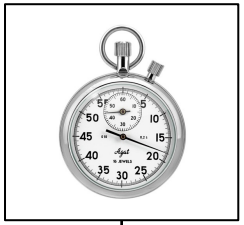


Substantially outperforms Snorkel

KRISSBERT: Contrastive Learning for Knowledge-Rich Self-Supervision

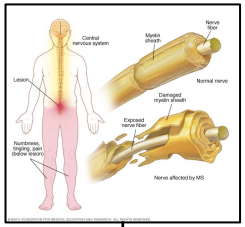
Entity Linking

Temporal concept



Millisecond

Disease

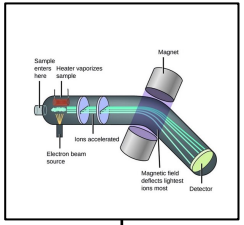


Multiple Sclerosis

Conductivity Unit

Millisiemens

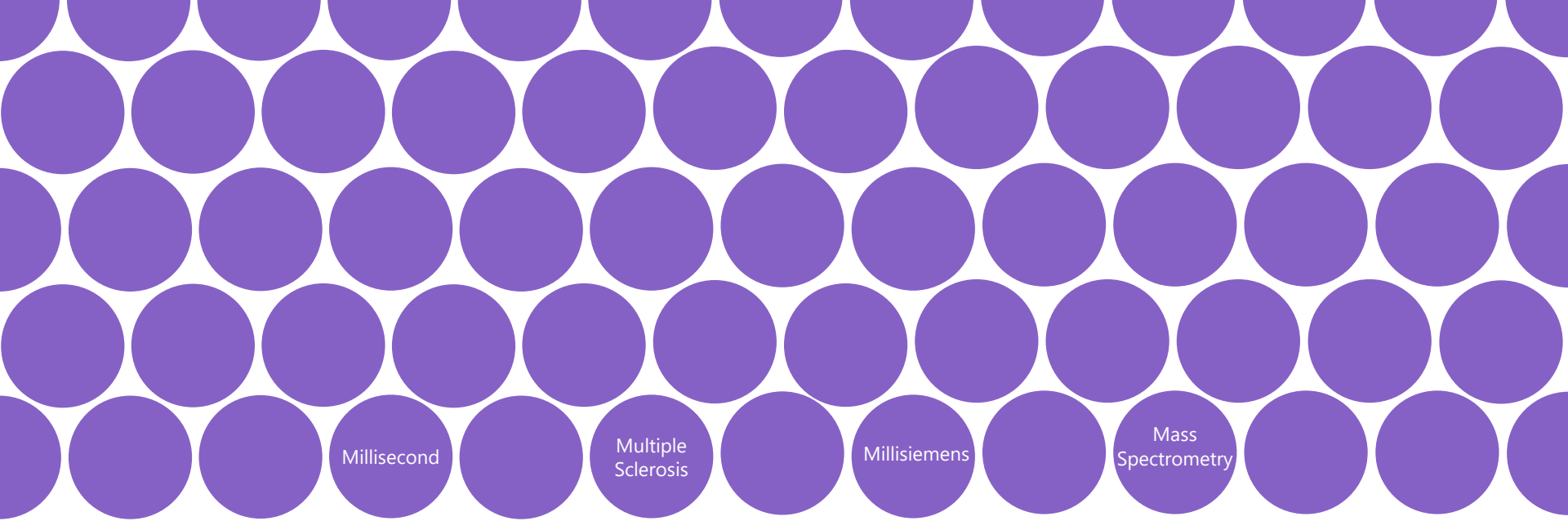
Analytical tool



Mass Spectrometry

Variation, Ambiguity,

… NMR spectroscopic data as well as CD and **MS** analysis. All isolates were tested for their …



Variation, Ambiguity, Gigantic Vocab, Lack of Annotations, Zero Shot

... NMR spectroscopic data as well as CD and **MS** analysis. All isolates were tested for their ...

Unified Medical Language System (UMLS)

Metathesaurus:

- A repo of biomedical vocabularies developed by the US National Library of Medicine.
- 14 million names for 4.5 million concepts from more than 200 biomedical vocabularies and 25 languages.

Semantic Network:

- Broad categories (semantic types) and their relationships (semantic relations).

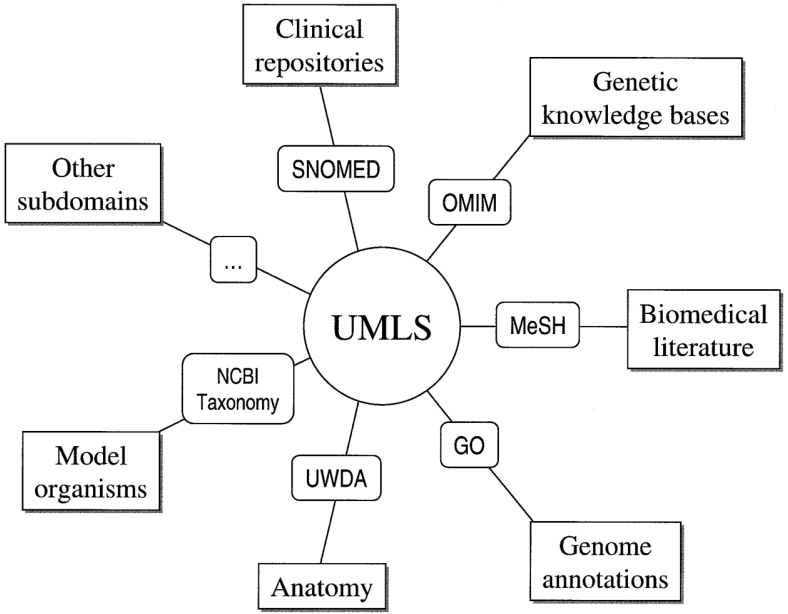
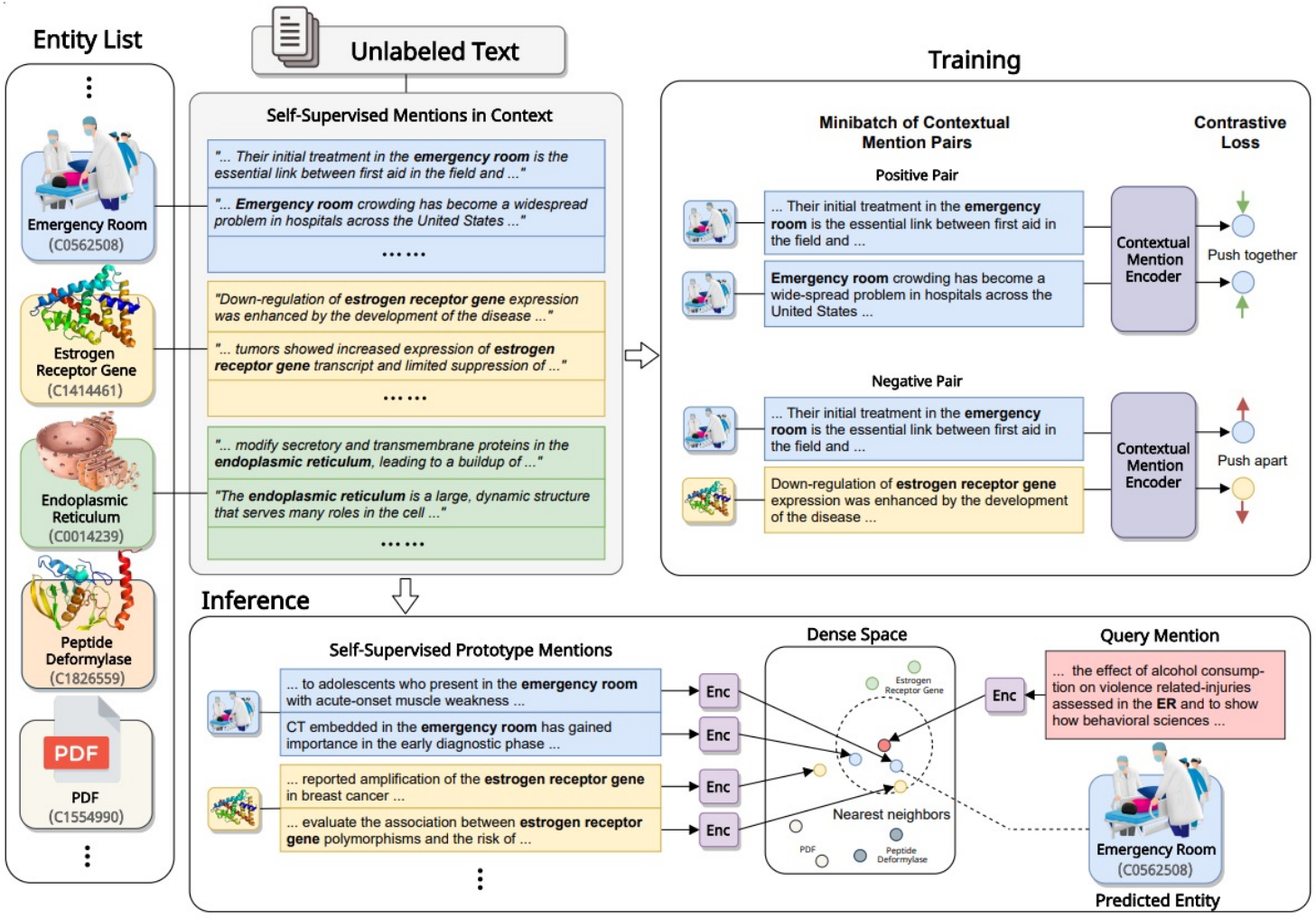


Figure 1. The various subdomains integrated in the UMLS.

Bodenreider (2004)



Knowledge-Rich Self-Supervised Entity Linking

Test Accuracy

MedMention

	NCBI	BC5CDR-d	BC5CDR-c	ShARe	N2C2	MM (full)	MM (st21pv)	Mean
QuickUMLS	39.7	47.5	34.9	42.1	29.8	12.1	20.0	32.3
BLINK	49.0	48.7	52.0	32.8	25.1	13.9	19.4	34.4
SapBERT†	63.0	83.6	96.2	80.4	59.7	37.6	44.2	66.4
KRISSBERT (self-supervised)	**83.2**±0.5	**85.5**±0.2	**96.5**±0.1	**84.0**±0.1	**67.8**±0.1	**61.4**±0.1	**63.5**±0.1	**77.4**
MedLinker	50.5	62.0	80.5	56.8	37.6	32.9	57.6	54.0
ScispaCy	66.8	64.0	85.3	66.6	54.6	53.1	52.9	63.3
KRISSBERT (supervised only)	76.9±0.9	85.5±0.7	93.8±0.3	53.9±0.4	29.2±1.2	60.7±0.3	63.7±0.4	66.2
KRISSBERT (lazy supervised)	**89.9**±0.1	**90.7**±0.1	**96.9**±0.1	**90.4**±0.1	**80.2**±0.1	**70.7**±0.1	**70.6**±0.1	**84.2**

Outperforms SapBERT and other prior SOTA by large margin

Prior SapBERT Scores Are Incorrect

SapBERT [Liu et al. 2021] reported wrong scores

Only predicts surface form, not canonical ID

"Correct" if matching one of gold surface forms

Completely ignore mention context

Impossible to resolve ambiguities

Same errors in recent biomed entity linking papers

Originated from BIOSYN [Sung et al. 2020]

	Mention As-is	SapBERT	KRISSBERT
NCBI	76.9	92.0	91.3
BC5CDR-d	83.4	93.8	92.8
BC5CDR-c	92.3	96.5	97.2
ShARe	74.5	85.6	87.3
N2C2	61.2	67.9	76.1
MM (full)	47.1	52.2	71.3
MM (st21pv)	48.3	53.8	72.2
Mean	69.1	77.4	84.0

Table 4: Accuracy comparison based on the evaluation metric used by Liu et al. (2021).

In representative datasets like MedMention, SapBERT not much better than string matching

	Ambiguous(%)	SapBERT	KRISSBERT
NCBI	43.2	57.1	64.5
BC5CDR-d	30.7	63.9	64.5
BC5CDR-c	11.5	76.4	76.5
ShARe	48.5	67.5	72.4
N2C2	67.5	50.7	58.2
MM (full)	67.8	24.8	48.9
MM (st21pv)	69.4	29.6	52.5

Table 5: Accuracy comparison on ambiguous cases.

In representative datasets like MedMention,
SapBERT's inability to resolve ambiguities is most apparent

Example Ambiguous Case

Mention: "... *Hence, we aimed to find drug targets using the 2DE / **MS** proteomics study of a dexamethasone - resistant cell line ...*"

SapBERT prediction: Master of Science (C1513009), Montserrat Island (C0026514), Mass Spectrometry (C0037813), ...

KRISSBERT prediction: Mass Spectrometry (C0037813)

KRISSBERT predicted prototype: "... ***mass spectrometry*** *is a widely used technique for enrichment and sequencing of phosphopeptides ...*"

Example Ambiguous Case

Example: "... *every patient followed up accordingly within ten days of* ***discharge*** ..."
SapBERT prediction: Discharge, Body Substance, Sample (C0600083), Body Fluid Discharge (C0012621), Patient Discharge (C0030685)
`KRISSBERT` **prediction**: Patient Discharge (C0030685)
`KRISSBERT` **predicted prototype**: *"Performance of the Hendrich Fall Risk Model II in Patients* ***Discharged*** *from Rehabilitation Wards ..."*

Example Ambiguous Case

Example: "*... we added separately, live cells and heat-killed* ***cells*** *of E. coli C600 ...*"
SapBERT prediction: Clone Cells (C0009013), Cell Count (C0007584), Cell Line, Tumor (C0085983), Cells (C0007634), ...
`KRISSBERT` **prediction**: Cells (C0007634)
`KRISSBERT` **predicted prototype**: *"... gram-positive rods such as C. liquefaciens activate T and A* ***cells*** *..."*

Comparison w. Supervised SOTA

	KRISSBERT (lazy supervised)	Supervised State of the Art
NCBI	89.9	89.1 (Ji et al., 2020)
BC5CDR	93.7	91.3 (Angell et al., 2021)
ShARe	90.4	91.1 (Ji et al., 2020)
N2C2	80.2	81.6 (Xu et al., 2020)
MM (full)	70.7	45.3† (Mohan and Li, 2019)
MM (st21pv)	70.6	74.1 (Angell et al., 2021)

Single model, no dataset-specific training or advanced techniques like joint inference

Knowledge-Rich Self-Supervised Entity Linking

Contrastive learning: Applicable to NER, EL, RE, QA …

Unlabeled Text + Entity List → SOTA Biomed EL

Universal entity linker for all 4 million UMLS entities

Outperforms prior best by over 20 points

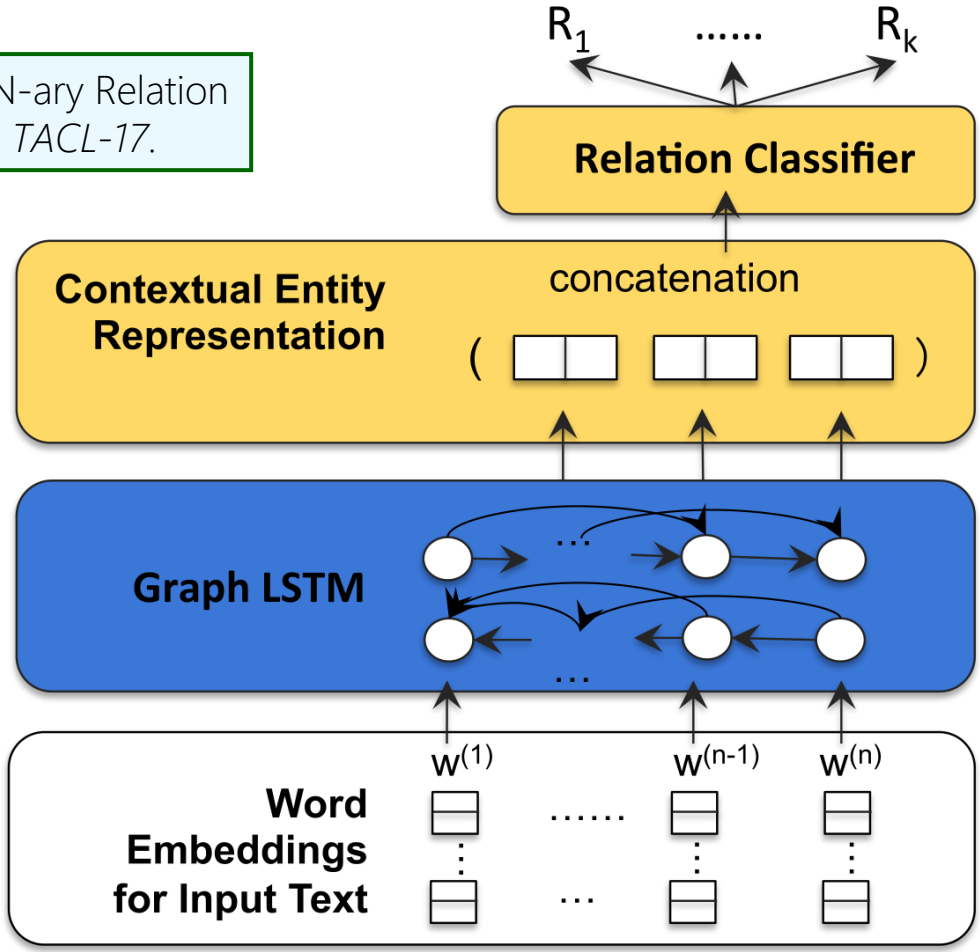
Zhang, et al. "Knowledge-Rich Self-Supervised Entity Linking", *in submission*.

KRISSBERT: http://aka.ms/krissbert

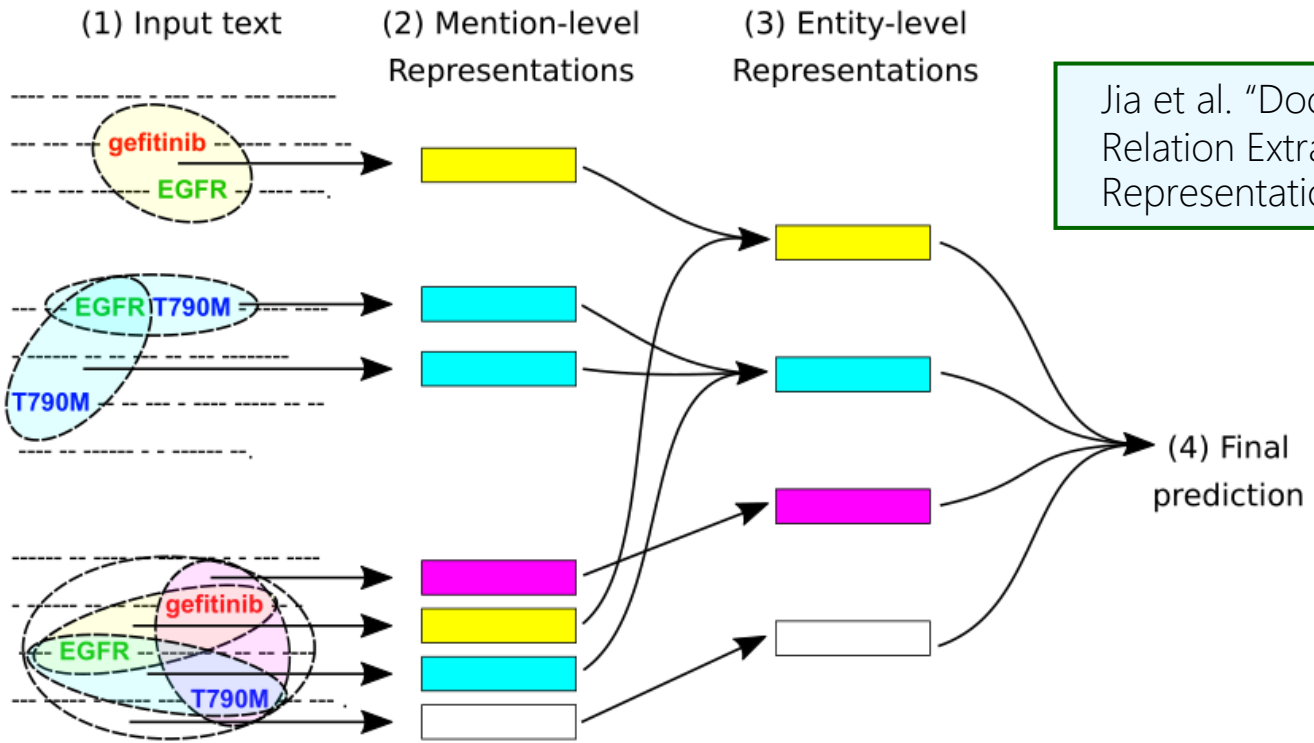
Novel Neural Architectures

Peng et al. "Cross-Sentence N-ary Relation Extraction with Graph LSTM", *TACL-17*.

Graph LSTM

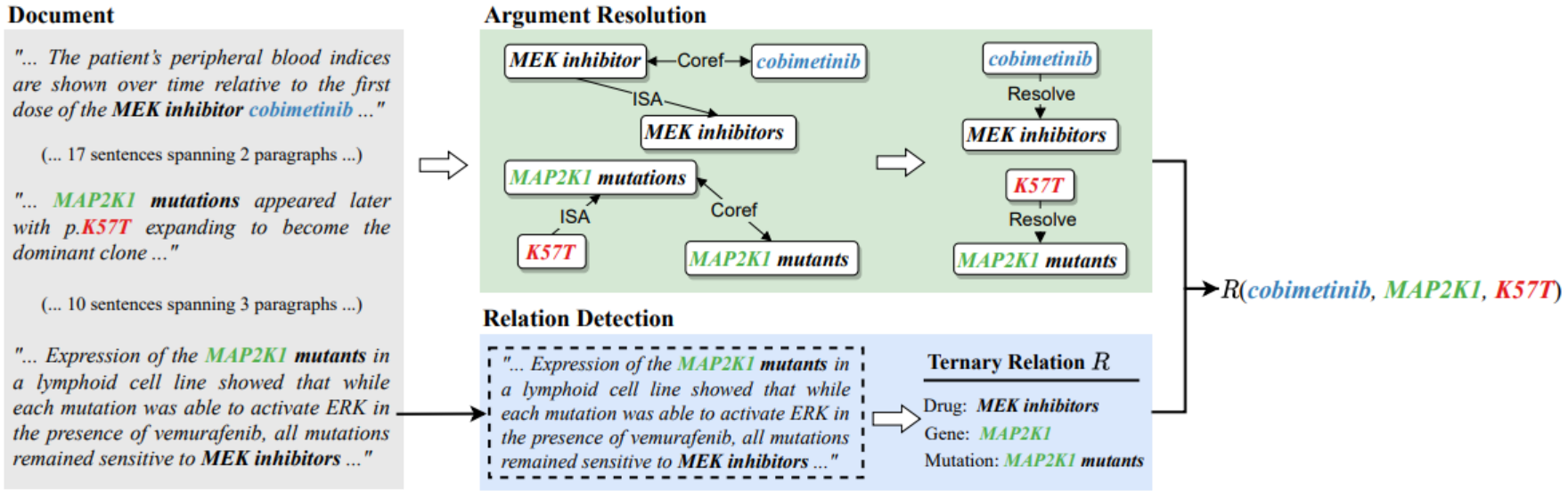


Multiscale Representation Learning



Jia et al. "Document-Level n-ary Relation Extraction with Multiscale Representation Learning", *NAACL-19*.

Modular Self-Supervision

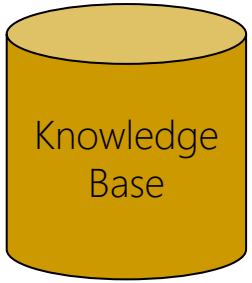


Zhang et al. "Modular Self-Supervision for Document-Level Relation Extraction", *EMNLP-21*.

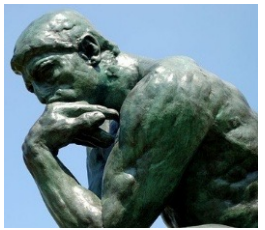
Mission: Structure All Medical Data

Project Hanover: AI for Precision Health

Machine Reading

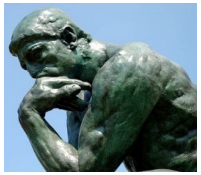


Decision Support



Structure all medical data

Assisted Curation



Do I trust this info?

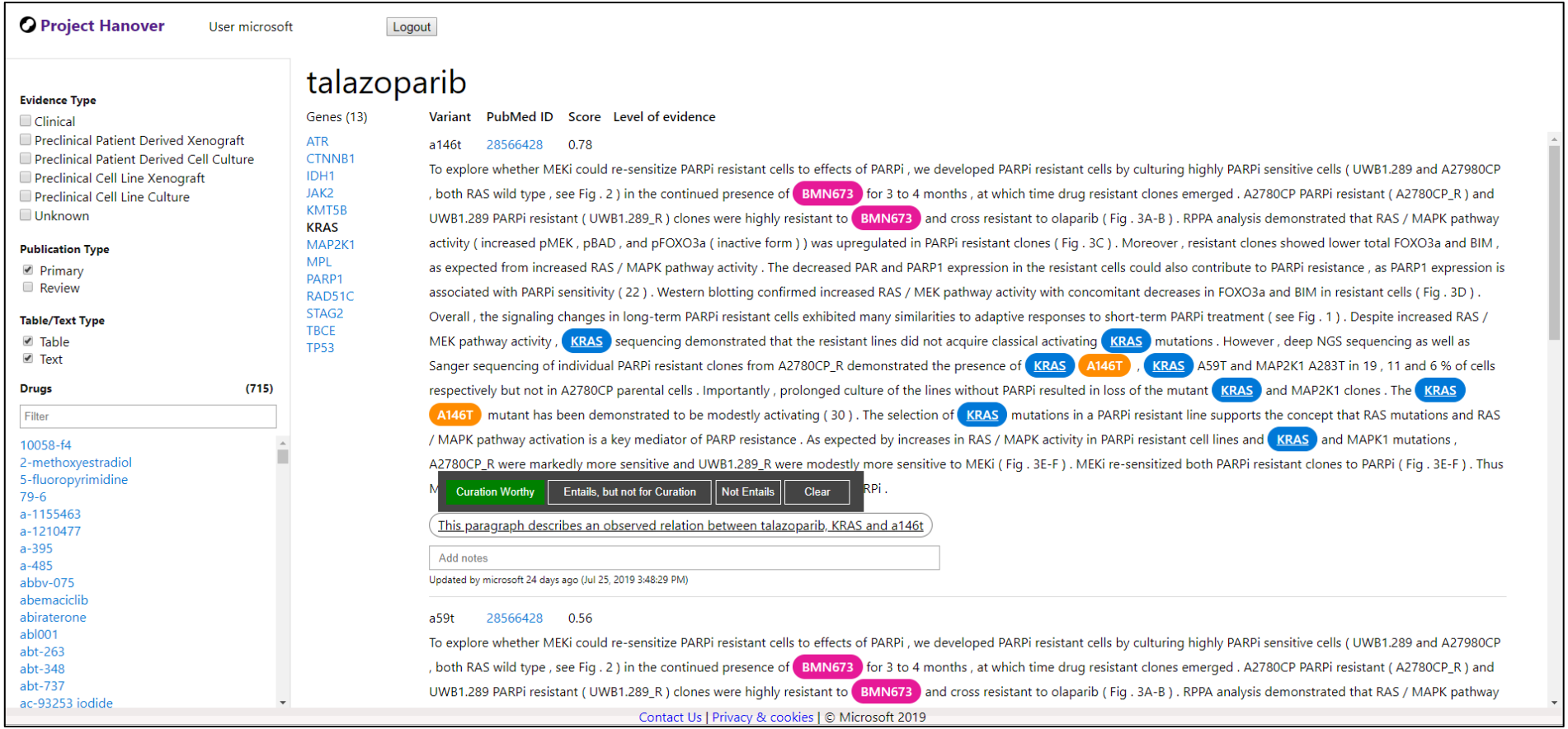
Does text state this info?

Curation

Machine Reading

Goal: Empower curators with super speed

Precision oncology: No labeled data required



Precision Ontology Knowledge Graph

"In our recent user study, manual curation identified 823 papers for one cancer therapy that had to be manually reviewed to extract 2 relevant patient responses. Hanover was able to narrow down the same therapy-related papers to 43, in seconds, and highlight 22 relevant patient responses for curation into CKB."

Thousands of expert-hours (for one drug) → 1 hour

The Jackson Laboratory

Wong et al. "Breaching the curation bottleneck with human-machine reading symbiosis", *in submission*.

Cancer RWE NLP

Providence

RWD curation: 1.2 million cancer patients

Order of magnitude more vs cancer registry (135K)

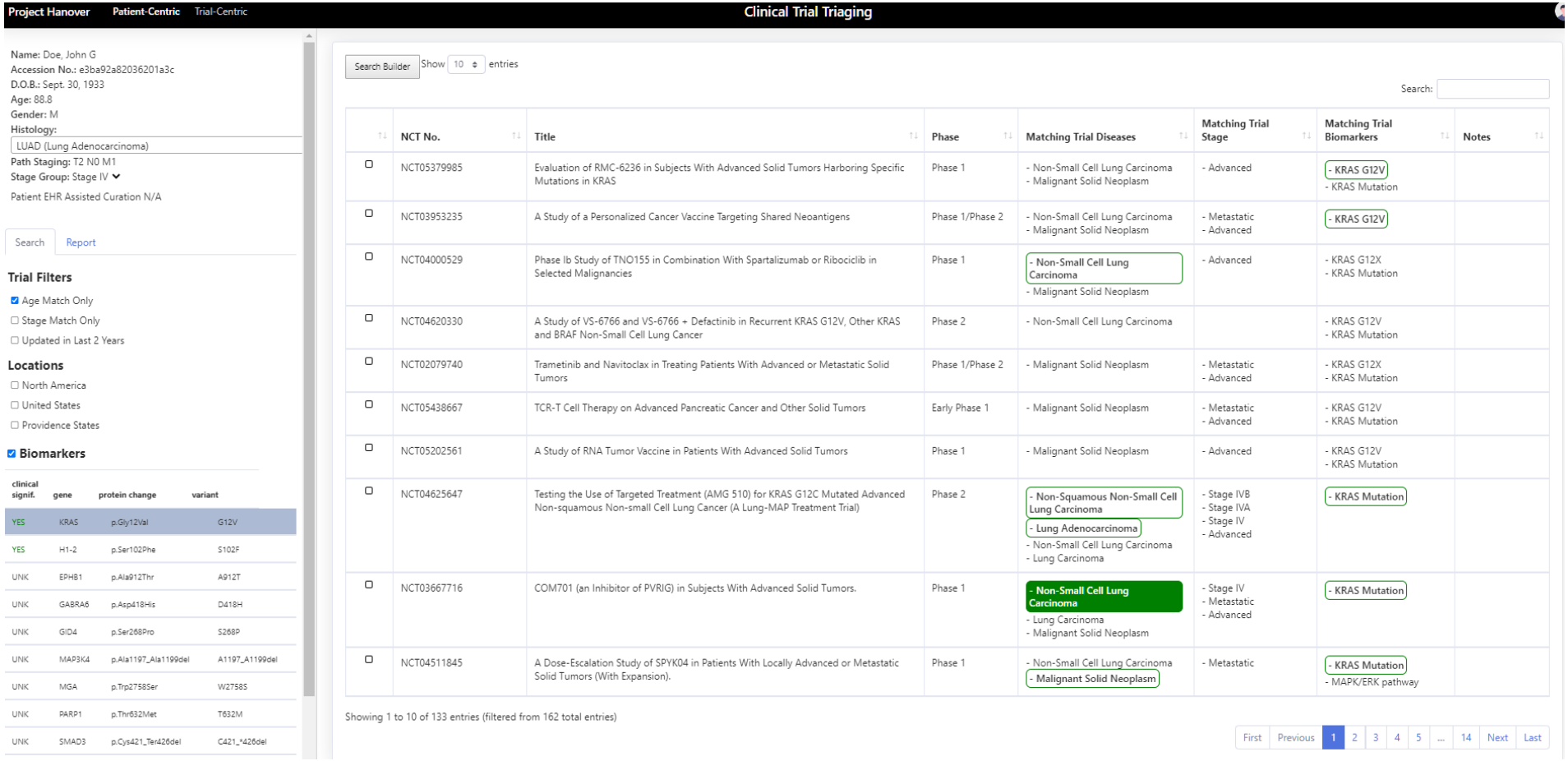
Daily: 100K patients, 790K notes

Precision oncology: PubMed knowledge graph

Clinical trial understanding: 400K trials (ct.gov)

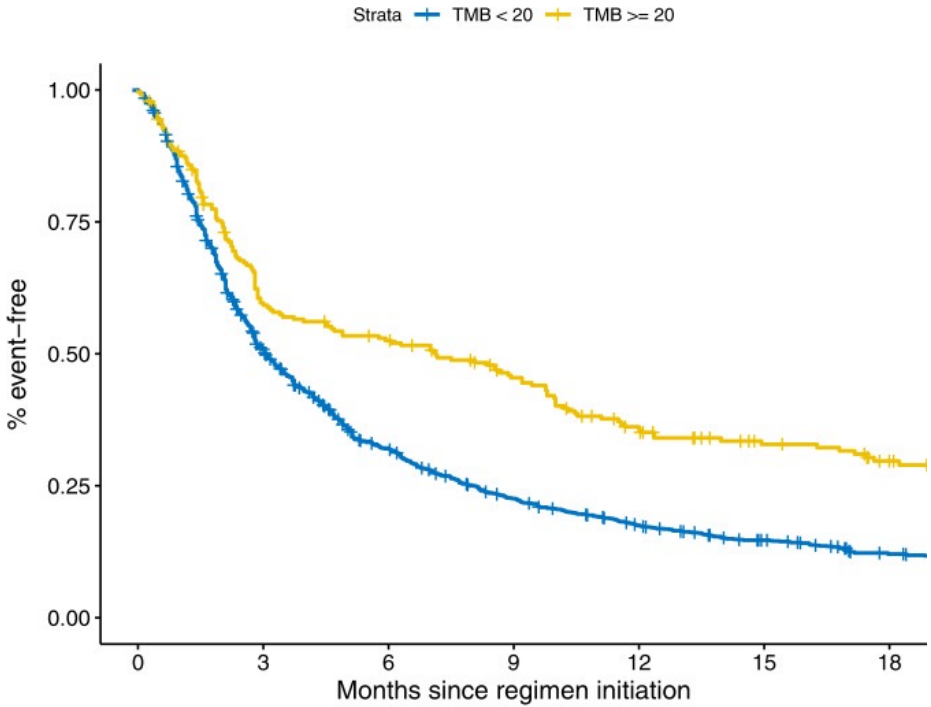
Power tumor board & clinical trial matching

Preston et al. "Towards Structuring Real-World Data at Scale: Deep Learning for Extracting Key Oncology Information from Clinical Text with Patient-Level Supervision", *in submission*.



Clinical Trial Matching (synthetic data, no PHI)

Cancer RWE NLP: Prior Work

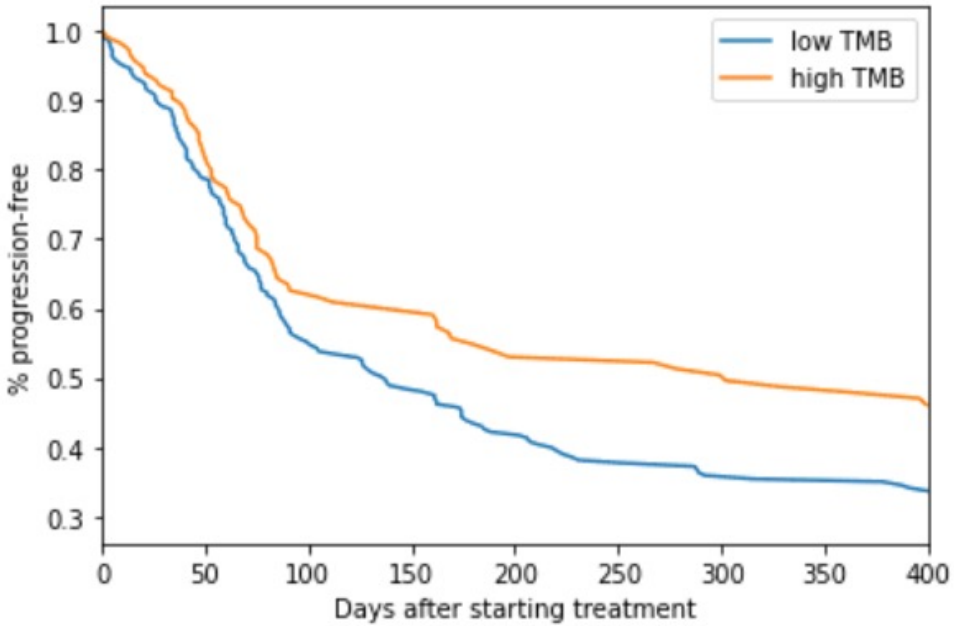


Immunotherapy response vs. tumor mutation burden (TMB)

Dana Farber experts manually annotated 60,000+ notes

Artificial intelligence-aided clinical annotation of a large multi-cancer genomic dataset. Kenneth et al. *Nature Communications 2021*.

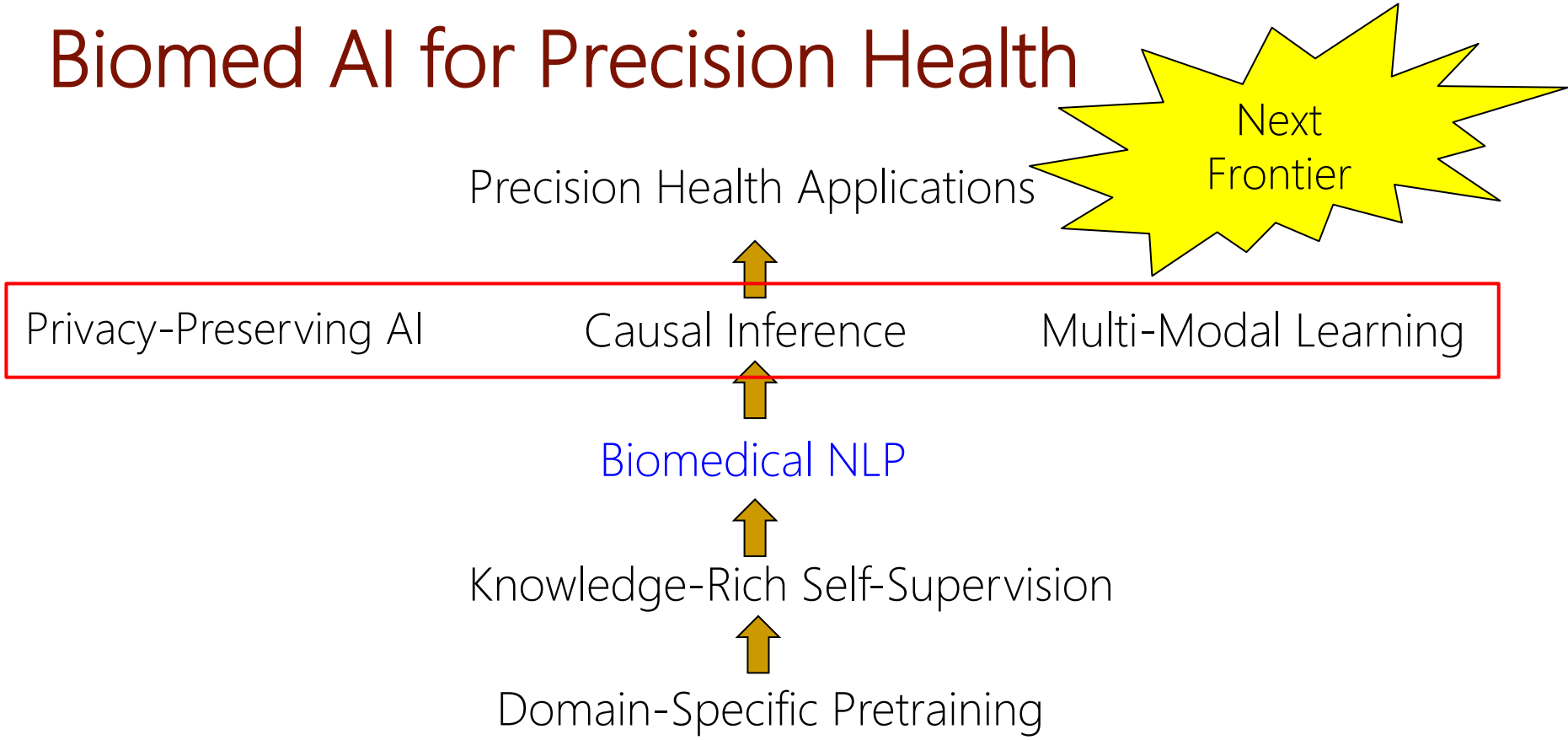
Cancer RWE NLP: Self-Supervision



OncoBERT
+
30 self-supervision rules
w.
Deep Probabilistic Logic

Immunotherapy response vs. tumor mutation burden (TMB)

Biomed AI for Precision Health



Real-World Evidence (RWE)

Pharma, Payor, Regulator

$Trillion Opportunity

Clinical trial recruitment
Synthetic control
Pragmatic trial
Post-market surveilance
Drug repurposing
Label expansion
.....

Insight Marketplace

RWE AI

ML for Observational Data

Privacy-Preserving AI | Causal RWE | Multi-Modal RWE

Domain-Specific Pretraining | Knowledge-Rich Self-Supervision

Data Lake Layer

RWE Evolution

Past: Claim data
Now: Clinical text
Next: Multi-modal EMR+omics+img

Provider, EHR Vendor

RWE AI: Modalities

Wang et al. "Classification of common human diseases derived from shared genetic and environmental determinants", Nature Genetics, 2017.

Structured data: ICD, CPT, …

Missing granularity & detail

RWE AI: Modalities

PubMedBERT, KRISSBERT, Deep Probabilistic Logic, Self-Supervised Self-Supervision, ……

Structured data: ICD, CPT, …

Biomedical text: Publications, clinical notes, …

Biomedical NLP

RWE AI: Modalities

Boecking*, Usuyama*, et al. Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. *ECCV 2022*.

Structured data: ICD, CPT, …

Biomedical text: Publications, clinical notes, …

Medical imaging: Radiology, pathology, …

Biomedical Vision-Language Processing

Vision-Language Processing (VLP)

Biomed vs general-domain (caption, VQA)

Research opportunities

- Structured: anatomy, conditions, composition
- Relational: temporal
- Scale: ImageNet 400 X 300 vs WSI 30K X 30K

Digital pathology: New frontier

Multi-Modal Pretraining

Boecking*, Usuyama*, et al. Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. *ECCV 2022*.

Image Encoder

Image Embedding

Projection

Projected Image Embedding

Contrastive Loss

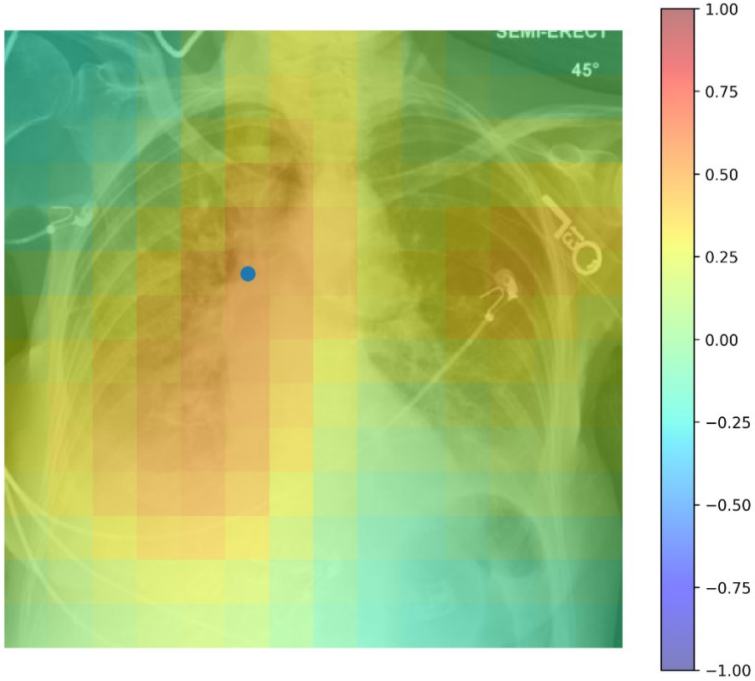
Text Encoder

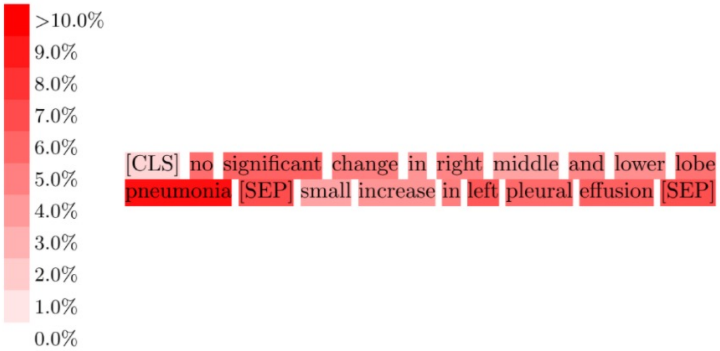
Text Embedding

Projection

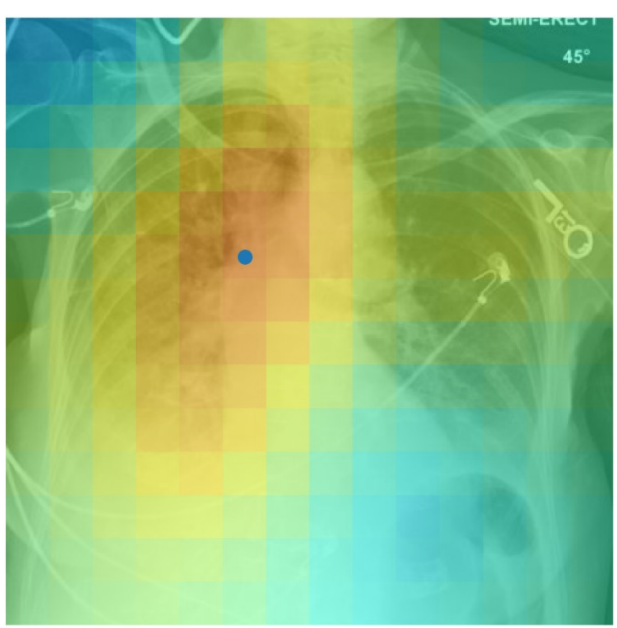
Projected Text Embedding

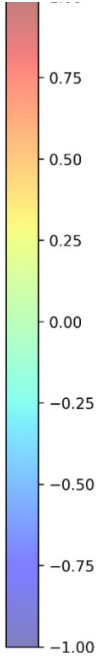
Multi-Modal Pretraining



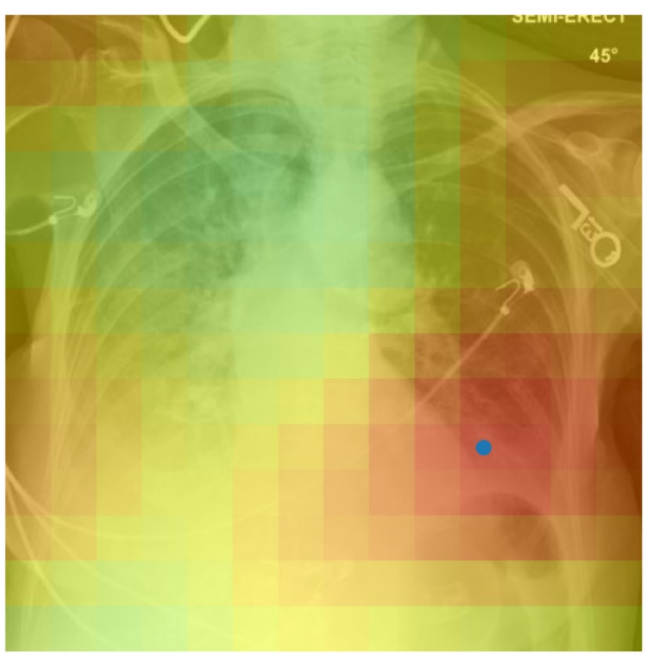


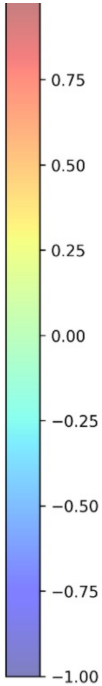
Multi-Modal Pretraining

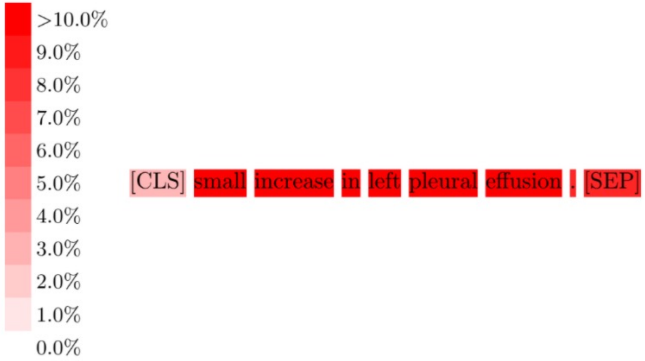




Multi-Modal Pretraining







RWE AI: Modalities

Structured data: ICD, CPT, …

Biomedical text: Publications, clinical notes, …

Medical imaging: Radiology, pathology, …

Emerging: Omics, IoT, …

Multi-Modal RWE

Multi-Modal RWE

Moonshot: Predict checkpoint inhibitor response

- Keytruda ($14B Merck 2020)
- Only minority of patients respond
- Simplistic companion diagnosis (TMB, PD-L1)

Multi-modal pretraining & fusion

EMR + Omics + Radiology / Pathology Imaging

Casual RWE: Counter-Factual Reasoning

Article | Published: 07 April 2021

Evaluating eligibility criteria of oncology trials using real-world data and AI

Ruishan Liu, Shemra Rizzo, Samuel Whipple, Navdeep Pal, Arturo Lopez Pineda, Michael Lu, Brandon Arnieri, Ying Lu, William Capra, Ryan Copping & James Zou

Nature **592**, 629–633 (2021) | Cite this article

Javier González

Near term: Optimize trial design

Moonshot: Optimize policy (adaptive trial)

RWE → Counter-factual playground

Privacy-Preserving Computing

Security: Access control, encryption, eyes-off

Reidentification risk: Differential privacy

Silos: Federated computing

Compete → Hesitant to Share

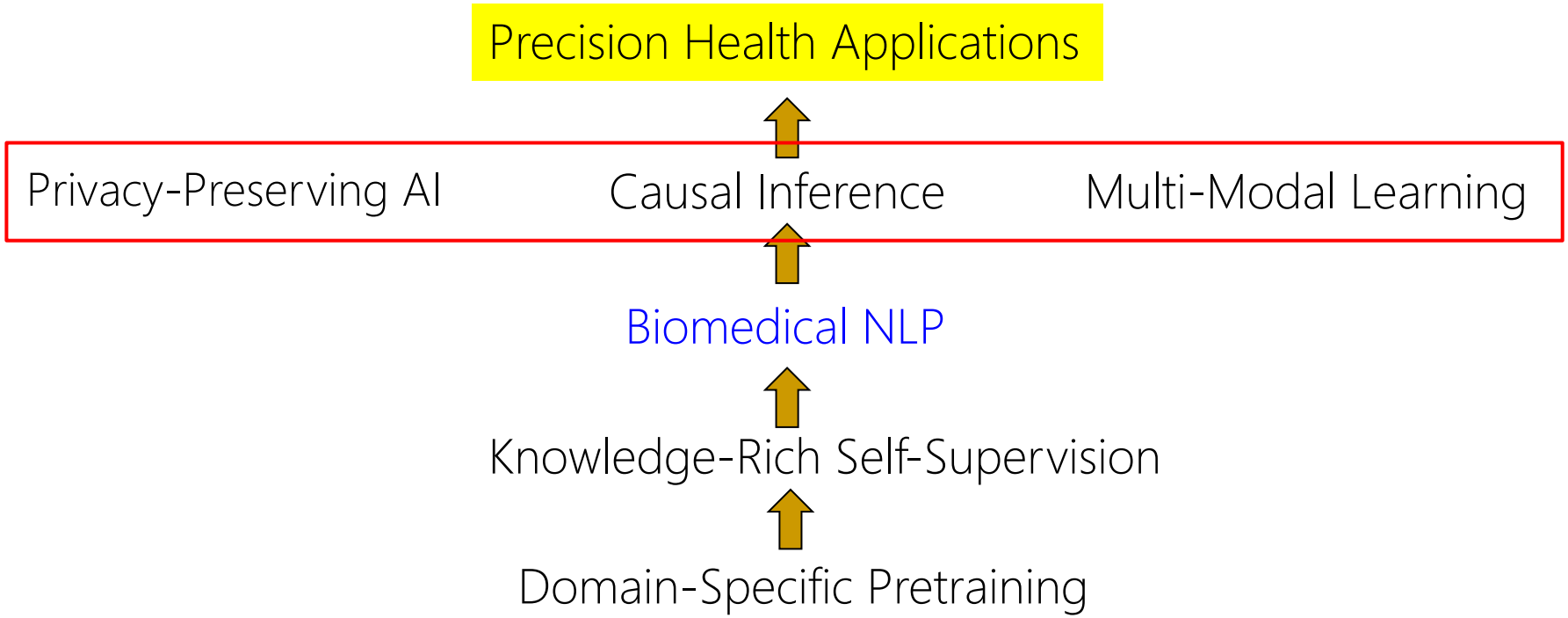
Common practice: De-id then sell

- Expensive; AI-complete for notes
- Prevent key use cases requiring identification (e.g., trial)
- Kick the can down the road (to Pharma / researchers)

Federated RWE AI → change status quo

Data valuation → viable business model

Biomed AI for Precision Health



Hanover Team

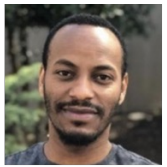

Cliff Wong


Tristan Naumann


Rajesh Rao

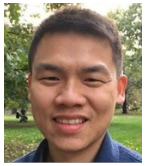

Naoto Usuyama

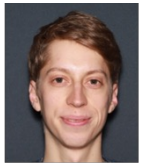

Sheng Zhang


Zelalem Gero


Javier González


Sam Preston


Mu Wei


Robert Tinn


Aiden Gu


Pratik Ghosh


Jass Bagga


Odeline Mateu-Silvernail

Collaborators

JAX: Susan Mockus, Sara Patterson

Fred Hutchinson: Christopher Li, Kathi Malone

Providence: Carlo Bifulco, Brian Piening

Knight Cancer Institute: Brian Druker, Jeff Tyner, Steve Kurtz

U. Chicago: Andrey Rzhetsky

MSR: Xiaodong Liu, Hao Cheng, Jianfeng Gao, Paul Bennett, Chenyan Xiong, Ozan Oktay, Javier Alvarez-Valle, Naveen Valluri

Interns: Maxim Grechkin, Ankur Parikh, Victoria Lin, Sheng Wang, Stephen Mayhew, Daniel Fried, Violet Peng, Hai Wang, Robin Jia, Matthew McDermott, Alexis Ross, Zelalem Gero, Sarthak Jain, Jenny Chen, Hunter Lang, Benedikt Boecking, Varsha Kishore, Jinfeng Xiao, Michelle Li, Wenxuan Zhou, Neha Hulkund, Risa Ueno