Artificial Intelligence in Medicine

Peter Szolovits
6.034
December 6, 2019

Massachusetts
Institute of
Technology

WHO Constitution defines “health”

“a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity”

- Physical
- Mental
- Social
    - very hard to measure

What are the Goals of Medicine?

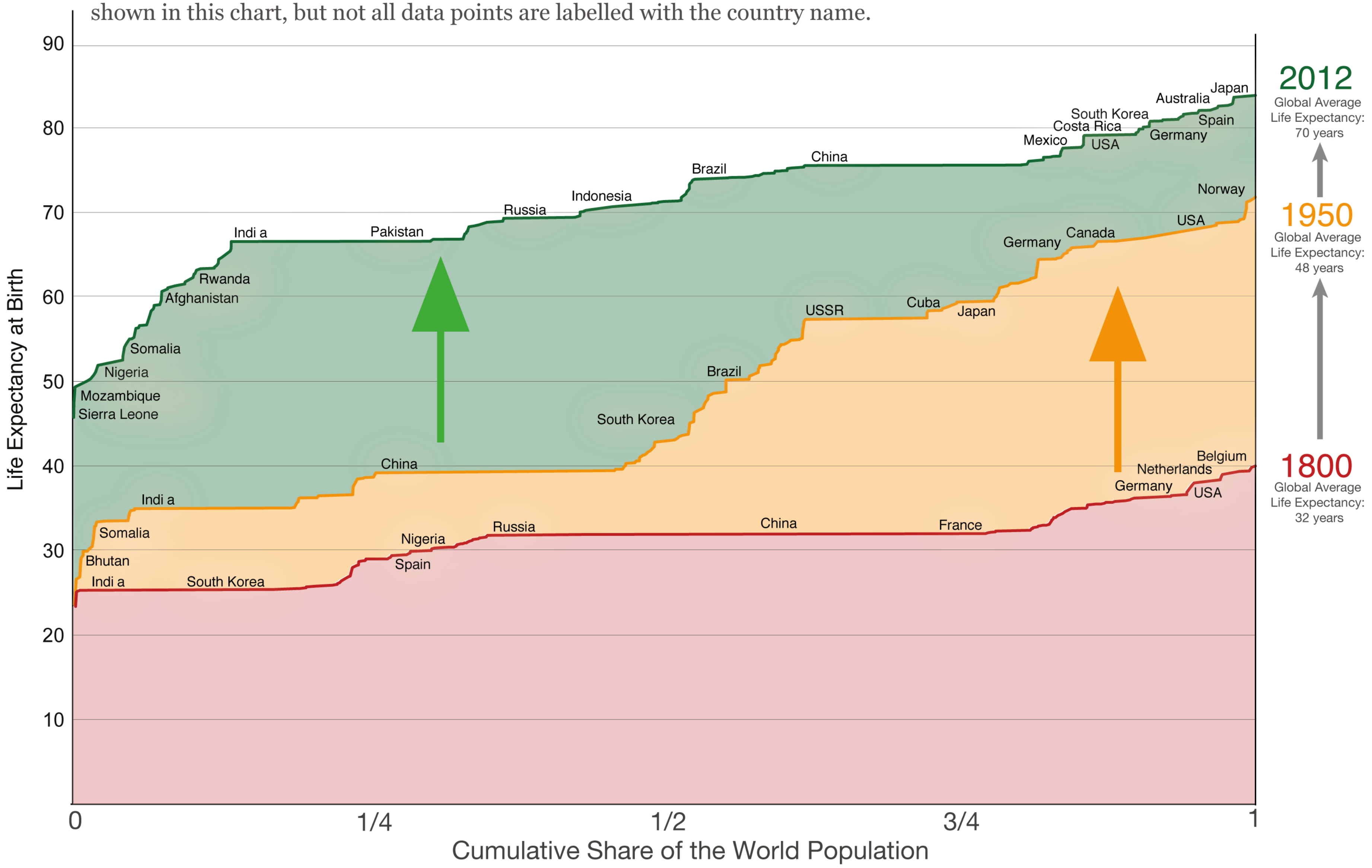
- Reduce mortality (death)
- Reduce morbidity (illness)
- Reduce disability

- Improve population health

Our World in Data

Life Expectancy of the World Population in 1800, 1950 and 2012

Countries are ordered along the x-axis ascending by the life expectancy of the population. Data for almost all countries is shown in this chart, but not all data points are labelled with the country name.



Data source: The data on life expectancy by country and population by country are taken from Gapminder.org.

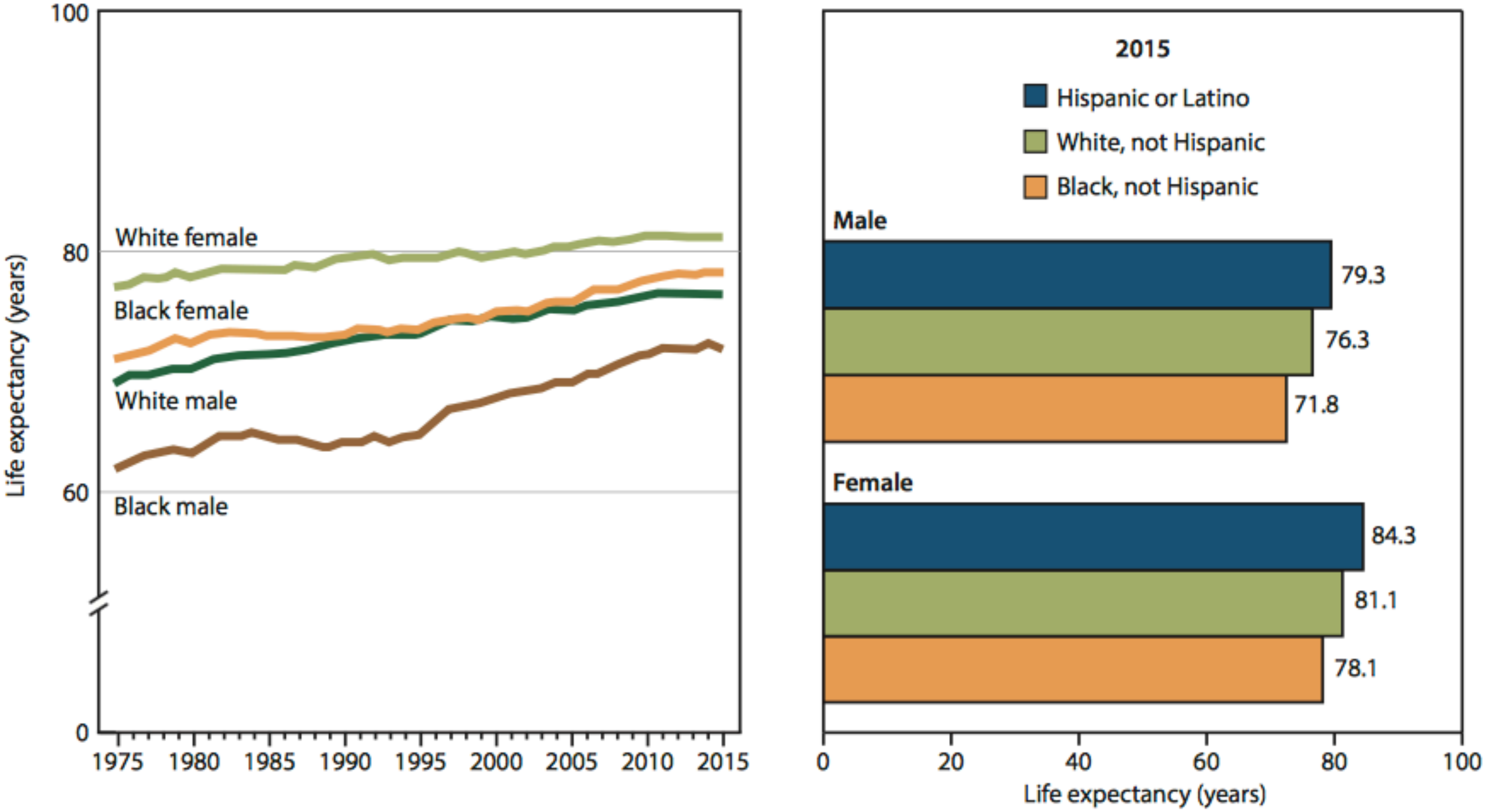
The interactive data visualisation is available at OurWorldinData.org. There you find the raw data and more visualisations on this topic.

Longevity at birth
(CIA World Fact Book, 2001, 2018)

Country	*Male*		*Female*	
	2018	*2001*	*2018*	*2001*
Rwanda	62.6	38.35	66.5	39.65
South Africa	62.7	47.64	65.6	48.56
Kenya	63.1	46.57	66.1	48.44
Cambodia	62.7	54.62	67.9	59.12
Russia	65.6	62.12	77.3	72.83
Brazil	70.7	58.96	78.0	67.73
Turkey	72.9	68.89	77.7	73.71
Albania	76.0	69.01	81.6	74.87
Israel	80.8	76.69	84.7	80.84
USA	77.8	74.37	82.3	80.05
France	78.9	75.01	85.3	83.01
Japan	82.2	77.62	89.0	84.15

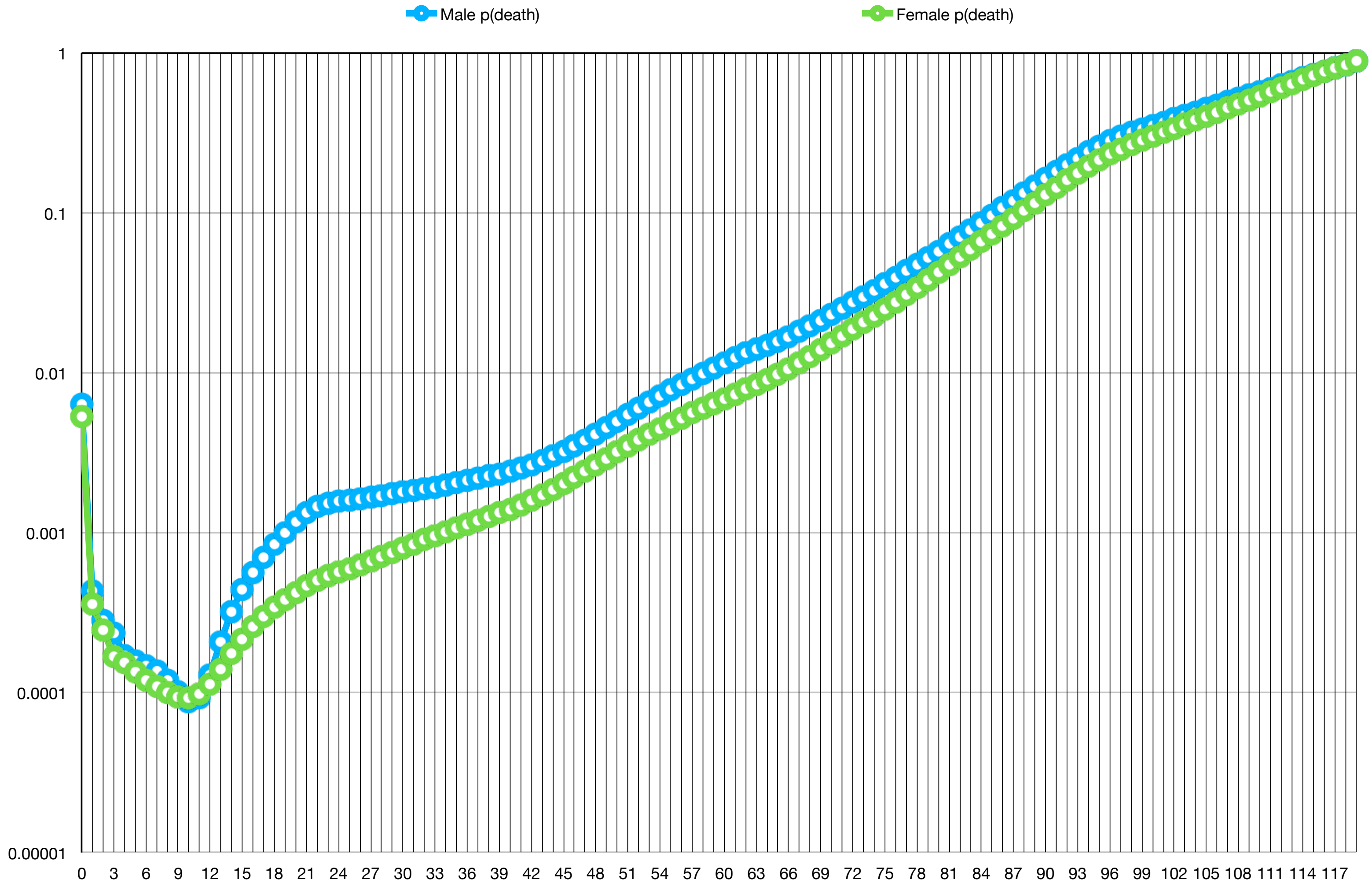
Ethnic Differences

Figure 6. Life expectancy at birth, by sex, race and Hispanic origin: United States, 1975–2015

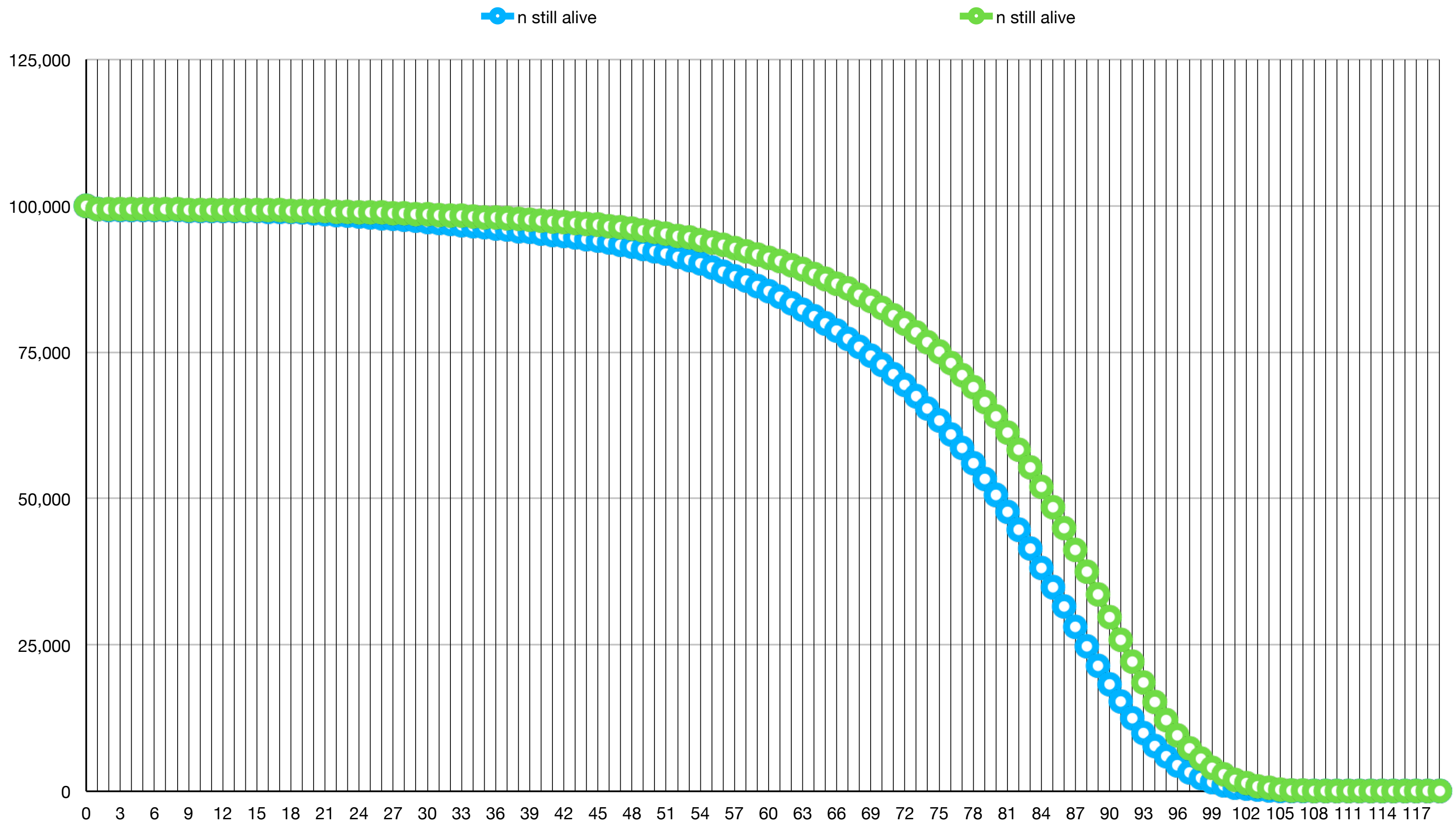


https://www.cdc.gov/nchs/data/hus/hus16.pdf#019

US Death Rates by Age (2016)



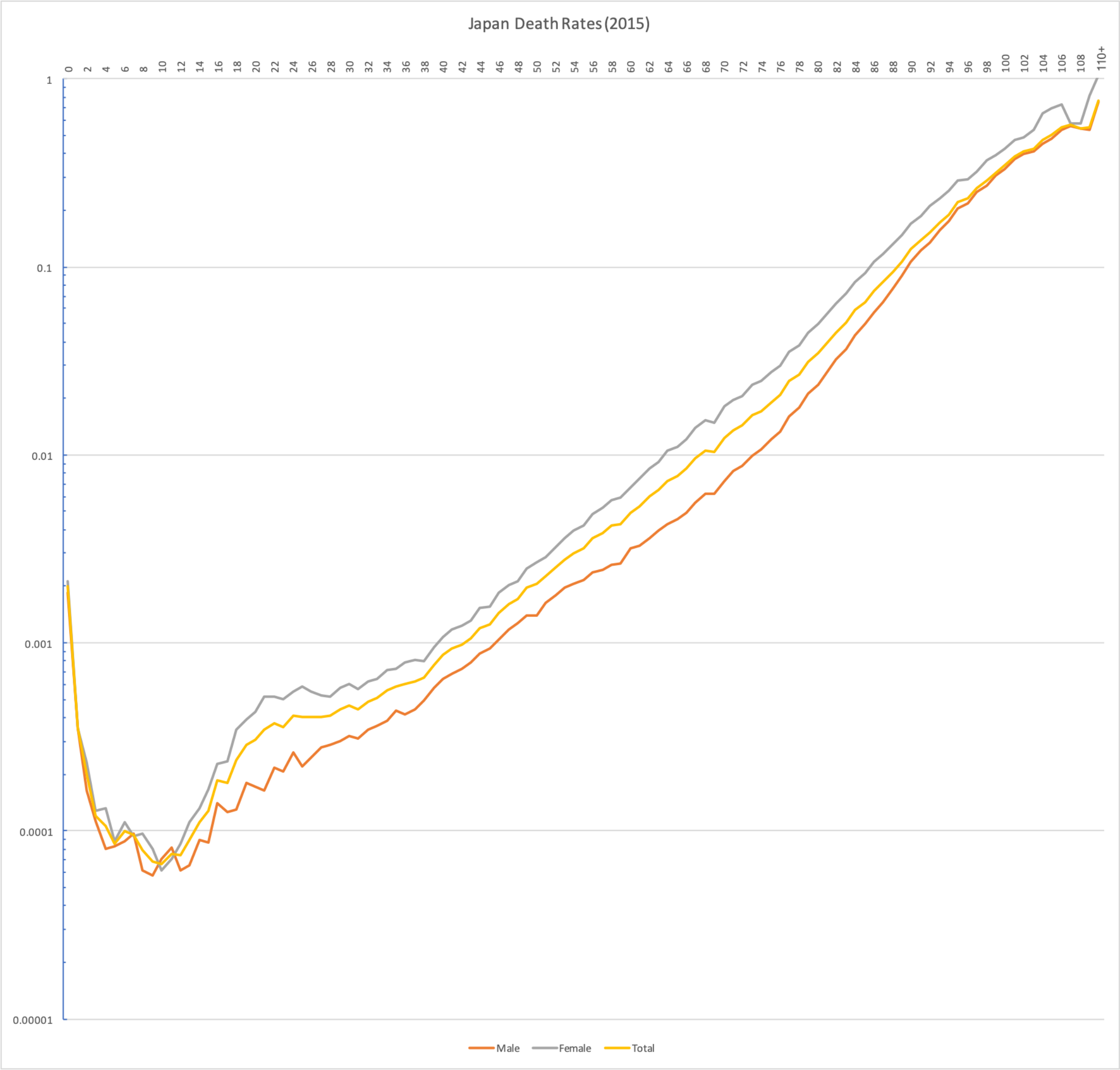
US Cohort Survival (2016)



Distribution of Death Rates by Age

- Life table deaths by year (Japan, 2015)

http://www.ipss.go.jp/p-toukei/JMD/00/STATS/Mx_1x1.txt



Causes of death
(USA, 2014)

Cause	**Deaths/100K**	**%**
Heart disease	192.7	23.4
Cancer	185.6	22.5
Chronic lower respiratory disease	46.1	5.6
Accidents	42.7	5.2
Stroke	41.7	5.1
Alzheimer's disease	29.3	3.6
Diabetes	24.0	2.9
Influenza and pneumonia	17.3	2.1
Kidney disease	15.1	1.8
Suicide	13.4	1.6
OTHER	*215.8*	*26.2*

https://www.medicalnewstoday.com/articles/282929.php

Morbidity: Top 10 Chronic Conditions
Persons aged ≥ 65

Condition	Both	Male	Female
Arthritis	49.6	40.7	55.7
Hypertension	39.0	33.0	43.2
Hearing impairment	30.0	35.2	26.3
Heart disease	25.7	26.9	24.9
Orthostatic impairment	16.8	15.7	17.8
Cataracts	15.5	11.3	18.4
Chronic sinusitis	15.2	13.7	16.2
Visual impairment	10.1	12.0	8.8
Genitourinary	9.9	11.3	8.9
Diabetes	8.9	7.8	9.7

U.S. Nat’l Ctr Health Stat, *Vital and Health Statistics*, 1985 (1982 data)

Quality of life

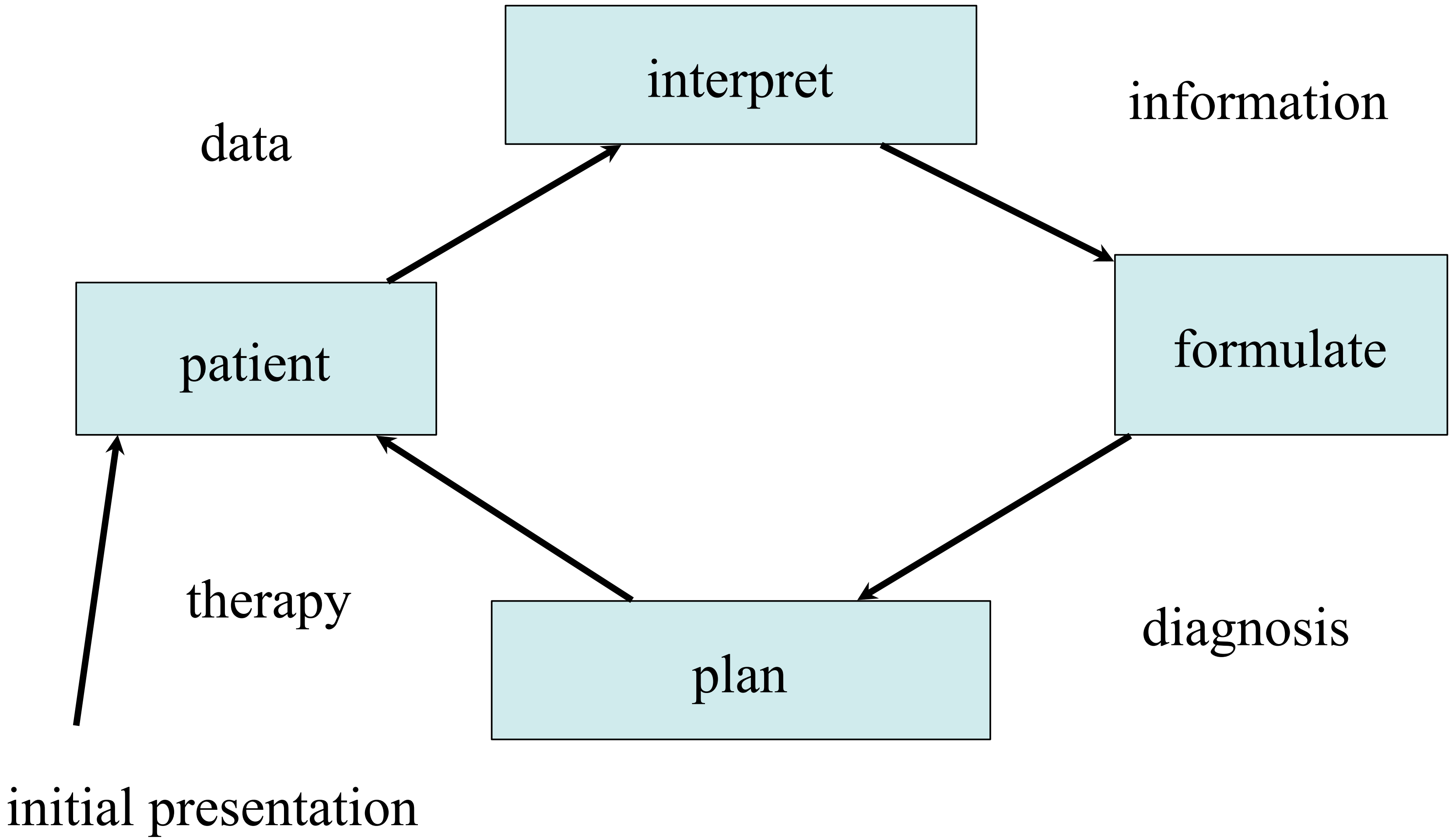
Value of a total life depends on

- Length, T (assume now is N)
- Quality (q) over time
- Discounts (g) for future or past
  - depends very much on what the value is to be used for
  - what is an appropriate discount factor?

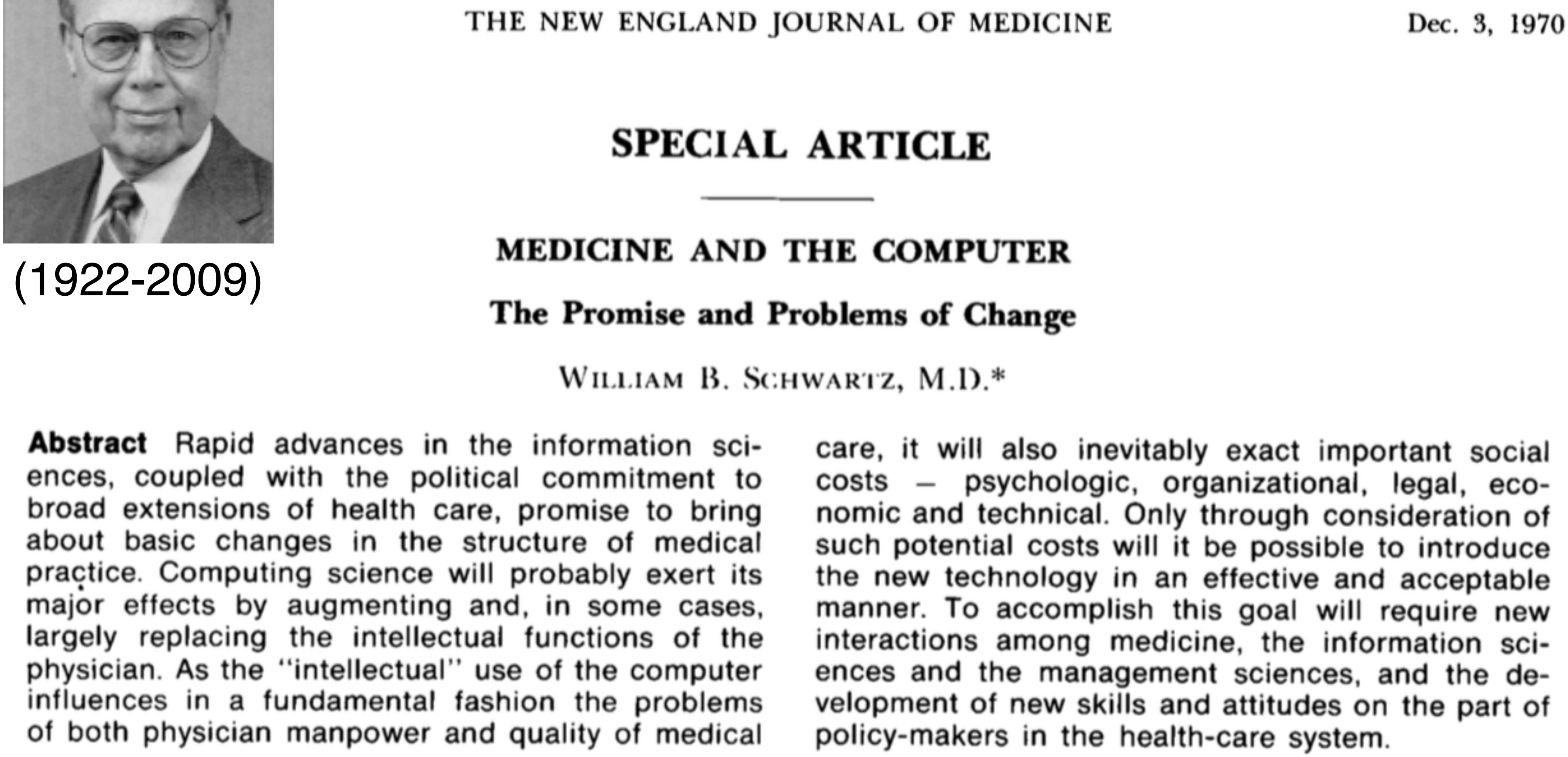
$$v_N = \int_{t=0}^{T} q(t)g(t-N)dt$$

The Medical Cycle

Diagnosis
Prognosis
Therapy Selection



(1922-2009)



THE NEW ENGLAND JOURNAL OF MEDICINE Dec. 3, 1970

SPECIAL ARTICLE

MEDICINE AND THE COMPUTER

The Promise and Problems of Change

WILLIAM B. SCHWARTZ, M.D.*

Abstract Rapid advances in the information sciences, coupled with the political commitment to broad extensions of health care, promise to bring about basic changes in the structure of medical practice. Computing science will probably exert its major effects by augmenting and, in some cases, largely replacing the intellectual functions of the physician. As the "intellectual" use of the computer influences in a fundamental fashion the problems of both physician manpower and quality of medical care, it will also inevitably exact important social costs – psychologic, organizational, legal, economic and technical. Only through consideration of such potential costs will it be possible to introduce the new technology in an effective and acceptable manner. To accomplish this goal will require new interactions among medicine, the information sciences and the management sciences, and the development of new skills and attitudes on the part of policy-makers in the health-care system.

Bill’s 1970 Predictions (emphasis added)

- “the ***computer as an intellectual tool*** can reshape the present system of health care, fundamentally alter the role of the physician, … — in short, the possibility that the health-care system by the year 2000 will be basically different from what it is today”
- “exploitation of the computer as an ***‘intellectual,’ ‘deductive’ instrument*** — a consultant that is built into the very structure of the medical-care system”
- “difficult challenge of maintaining a high level of physician competence in the face of a continued ***expansion of medical knowledge*** that tends to widen progressively the gap between what a doctor should know and what he can retain and utilize”
- “the physician and the computer will engage in ***frequent dialogue***, the computer continuously taking note of history, physical findings, laboratory data, and the like, alerting the physician to the ***most probable diagnoses and suggesting the appropriate, safest course of action***”
- “help free the physician to ***concentrate on the tasks that are uniquely human*** such as the application of bedside skills, the management of the emotional aspects of disease, and the exercise of good judgment in the nonquantifiable areas of clinical care”

Bill’s 1970 Predictions; more (emphasis added)

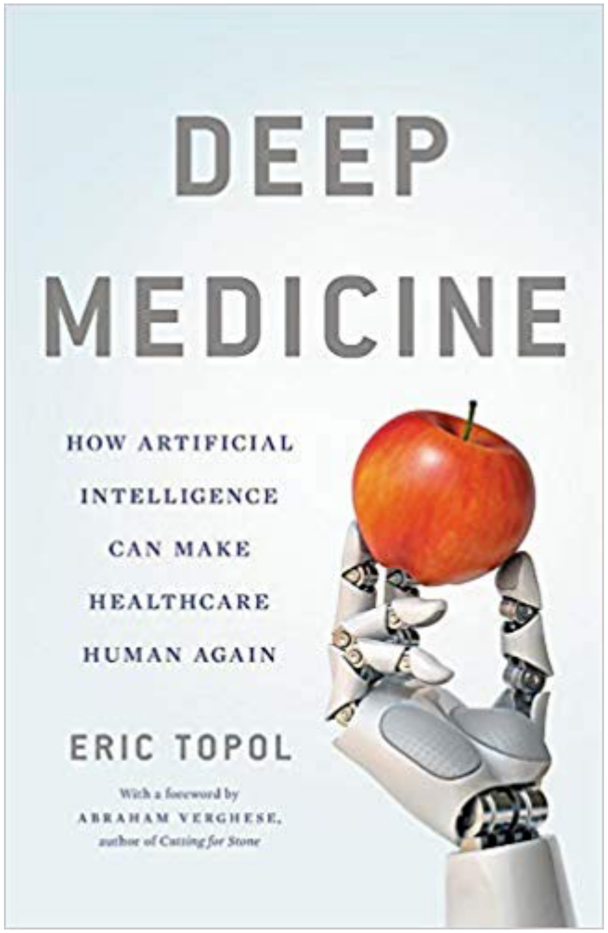
- “familiar projections envision the computer performing a wide variety of functions such as the scheduling of hospital admissions, the keeping of medical records and the operation of laboratory and pharmacy … in the area of ***“housekeeping”*** activities”
- “the increasing ***shortage of physician manpower*** and … geographic maldistribution” “Computer-supported ***“health-care specialists”***, aided by a variety of automated devices for history taking, blood analysis and other procedures, and trained to perform a careful physical examination, might take over a large segment of the responsibility for the delivery of primary medical care.”
- “it is conceivable that the computer could also ***take over a variety of specialized functions that are now performed by highly skilled physicians***. It is entirely possible, for example, that the administration of anesthesia — a function now uniquely human — could be largely or fully automated if new monitoring technics were combined with the capacity of the computer instantaneously to analyze and respond to large volumes of physiologic data”

Similar Projections Today for Tomorrow’s Medicine

- “Deep neural networks [are] algorithms that permit software to train itself to perform tasks…”
- Enabled by
  - Enormous collections of data
  - Enormous computational power
- Greatest successes in image interpretation:
  - Chest x-rays, retinopathy, dermatology, …
  - Growing capability in predictions, making sense of narratives
- US FDA approval process for AI tools
  - Handful of applications have been approved
- **Dream**: Virtual medical assistant
  - “AI can help achieve the gift of time with patients,”

  ⇒ Resurgence of empathy-based care

  medical care is about human care



DEEP MEDICINE

HOW ARTIFICIAL INTELLIGENCE CAN MAKE HEALTHCARE HUMAN AGAIN

ERIC TOPOL

With a foreword by
ABRAHAM VERGHESE,
author of Cutting for Stone

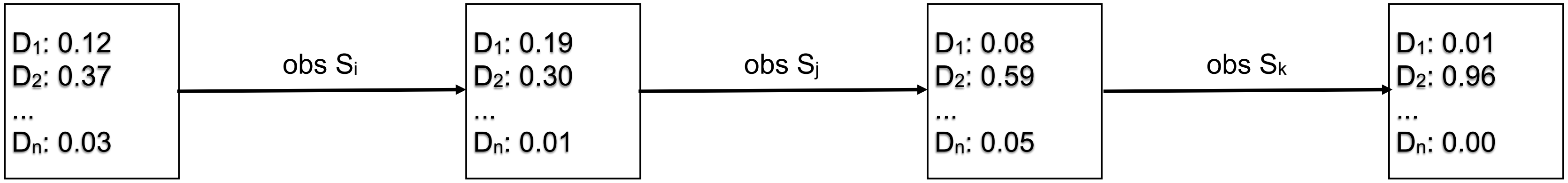
2019

Evolution of AI Approaches

- 1950s to 1960s — Simple Probabilistic Methods
  - Single-disease conditionally independent symptom models; e.g., appendicitis
  - Sequential Bayesian probability updates

Diagnostic Reasoning with Naive Bayes

- Exploit assumption of conditional independence among symptoms $P(s_1, s_2, \ldots, s_k \mid d_j) = P(s_1 \mid d_j) P(s_2 \mid d_j) \ldots P(s_k \mid d_j)$
- Sequence of observations of symptoms, s_i, each revise the distribution via Bayes' Rule
- $P^j(d_i \mid s_1, \ldots, s_j) = P^{j-1}(d_i) P(s_j \mid d_i) / P^{j-1}(s_j) = P^{j-1}(d_i) P(s_j \mid d_i) / \sum_{i=0}^{n} P^{j-1}(d_i) P(s_j \mid d_i)$



D$_1$: 0.12
D$_2$: 0.37
…
D$_n$: 0.03

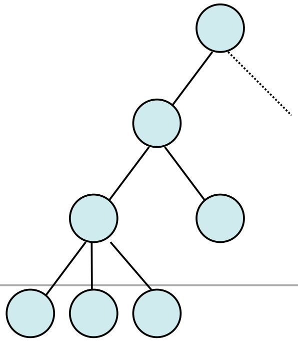
D$_1$: 0.19
D$_2$: 0.30
…
D$_n$: 0.01

D$_1$: 0.08
D$_2$: 0.59
…
D$_n$: 0.05

D$_1$: 0.01
D$_2$: 0.96
…
D$_n$: 0.00

Evolution of AI Approaches

- 1950s to 1960s — Simple Probabilistic Methods
- 1970s to 1980s — Symbolic Methods
  - Knowledge-based systems (rules, prototypes, logics)
  - E.g., Mycin (rule-based), multi-disease diagnosis (hypothesis space search)



Mycin—Rule-based Systems

- Task: Diagnosis and prescription for bacterial infections of the blood (and later meningitis)
- Method: Collection of modular rules
- To find out a fact
  - If there are rules that can conclude it, try them
  - Ask the user
- To “run” a rule
  - Try to find out if the facts in the premises are true
  - If they all are, then assert the conclusion(s), with a suitable certainty
- Backward chaining from goal to given facts
- Certainty factors

RULE037

IF the organism

> **1) stains grampos**
> **2) has coccus shape**
> **3) grows in chains**

THEN

> **There is suggestive evidence (.7) that the identity of the organism is streptococcus.**

Explore Mycin’s Use of Knowledge

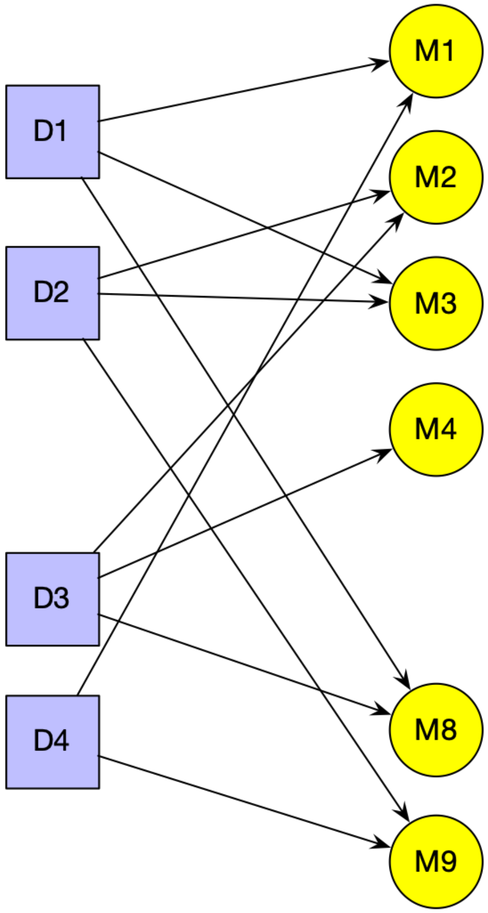
**** Did you use RULE 163 to find out anything about ORGANISM-1?**
RULE163 was tried in the context of ORGANISM-1, but it failed because it is not true that the patient has had a genito-urinary tract manipulative procedure (clause 3).

**** Why didn't you consider streptococcus as a possibility?**
The following rule could have been used to determine that the identity of ORGANISM-1 was streptococcus: RULE033
But clause 2 (“the morphology of the organism is coccus”) was already known to be false for ORGANISM-1, so the rule was never tried.
Davis, *et al.*, *Artificial Intelligence* **8**: 15-45 (1977)

Bipartite Graph Model

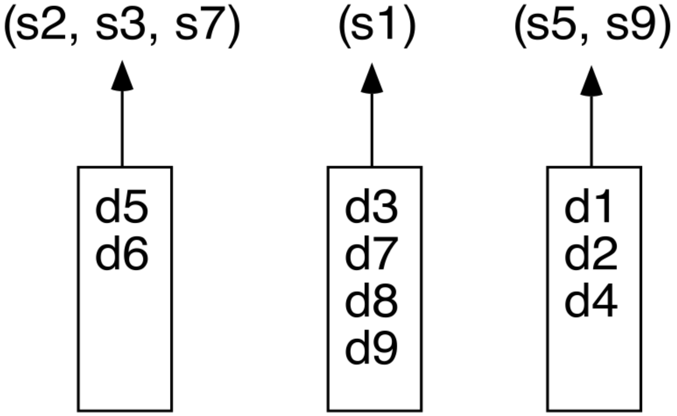
- Multiple diseases
- Diseases are independent
- Manifestations (signs, symptoms, lab results, etc.) depend only on which diseases are present
- Thus, they are conditionally independent

- Computationally intractable
- Various "greedy" search methods



Symptom Clustering for Multi-Disorder Diagnosis

- Assume a bipartite graph representation of diseases/symptoms
- Given a set of symptoms, how to proceed?
- If we could “guess” an appropriate clustering of the symptoms so that each cluster has a single cause …

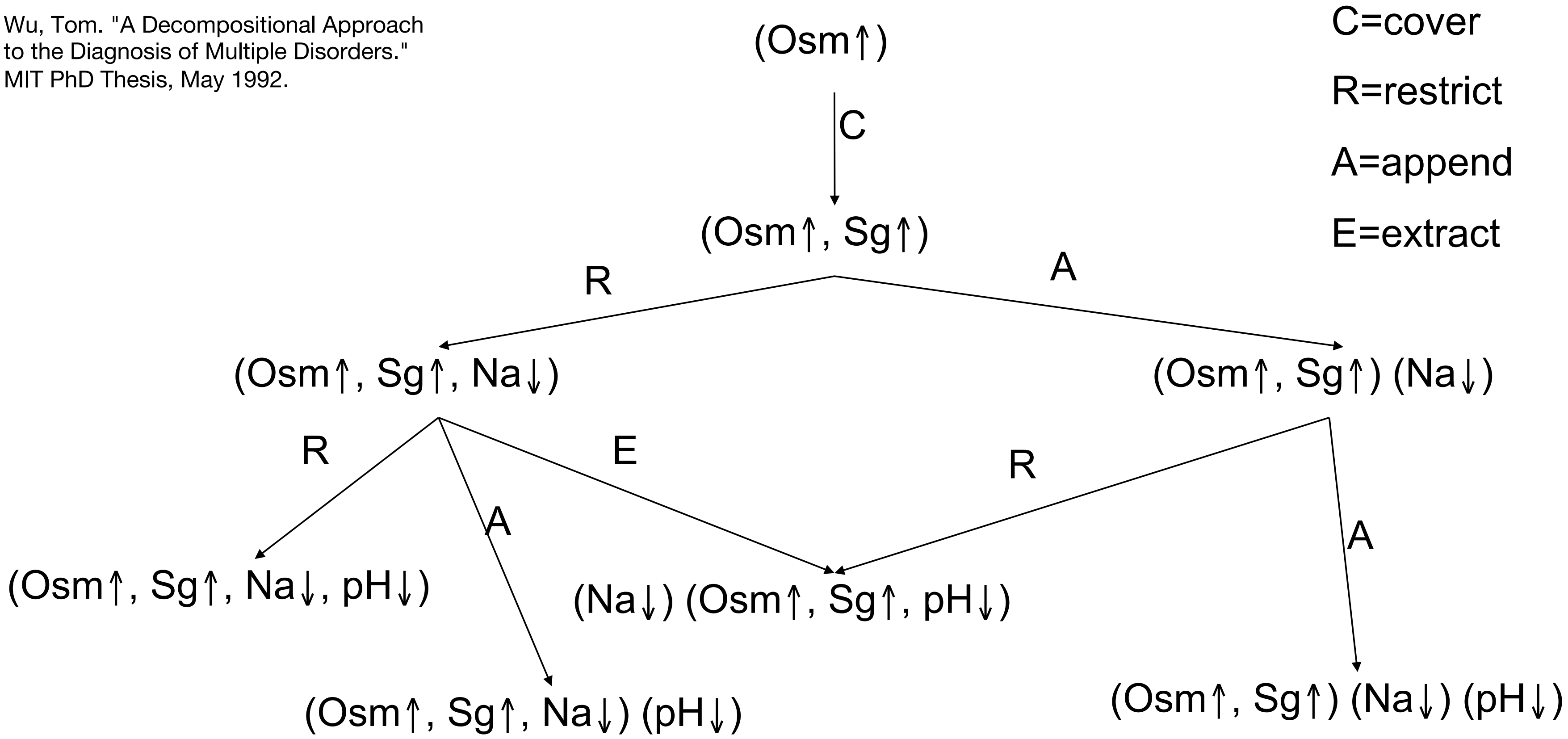


- … then the solution is (d5, d6) x (d3, d7, d8, d9) x (d1, d2, d4)

Wu, T. D. (1990). Efficient Diagnosis of Multiple Disorders Based on a Symptom Clustering Approach. Proc. National Conference on Artificial Intelligence, 357–364.

Search Through an Evolving Hypothesis Space

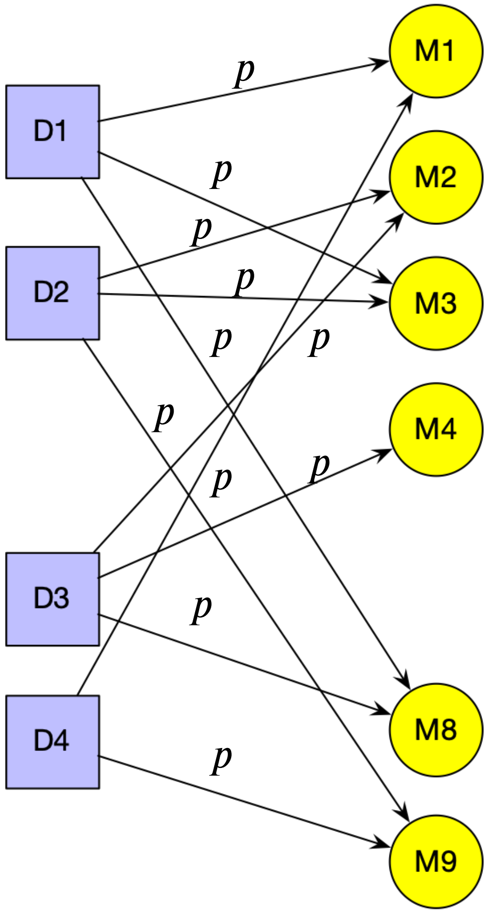
Wu, Tom. "A Decompositional Approach
to the Diagnosis of Multiple Disorders."
MIT PhD Thesis, May 1992.



	HTN	AGN	IgA	PRA	HRS	RV	CHF	Aldo	Peri	DKA	AN	HKN	CRF	RTA
Osm↑	X	X	X	X	X	X								
Sg↑	X	X	X	X	X	X	X							
Na↓				X	X		X	X	X					
pH↓		X		X						X	X	X	X	X

Evolution of AI Approaches

- 1950s to 1960s — Simple Probabilistic Methods
- 1950s to 1980s — Symbolic Methods
- Since 1980s — Probabilistic Methods
  - From Naïve Bayes to Bayesian Networks
    - Add probabilities to Bipartite Network
    - More complex, “deeper” networks
      - include dependencies among diseases, syndromes, signs & symptoms, treatments, risk factors, etc.
  - Learning Networks from Data



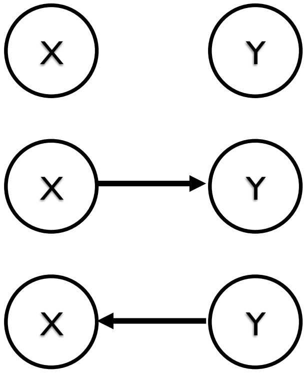
Learning Bayes Networks

- Learn structure G and parameters θ
  - structure captures (in)dependence among variables
  - parameters specify conditional probability of each node given its parents
- $P(G|D)= P(D|G)P(G)/P(D) \propto P(D|G)P(G)$
- $P(D|G)=\int P(D|\theta_G, G)P(\theta_G|G)P(G)d\theta_G$
- Search over all possible graphs and parameters
  - Maximize $P(D|G)$
  - Make practical:
    - Limit max number of parents
    - Constrain to partial order of nodes
  - Usual optimization methods; e.g., hill-climbing on $P(D|G)$

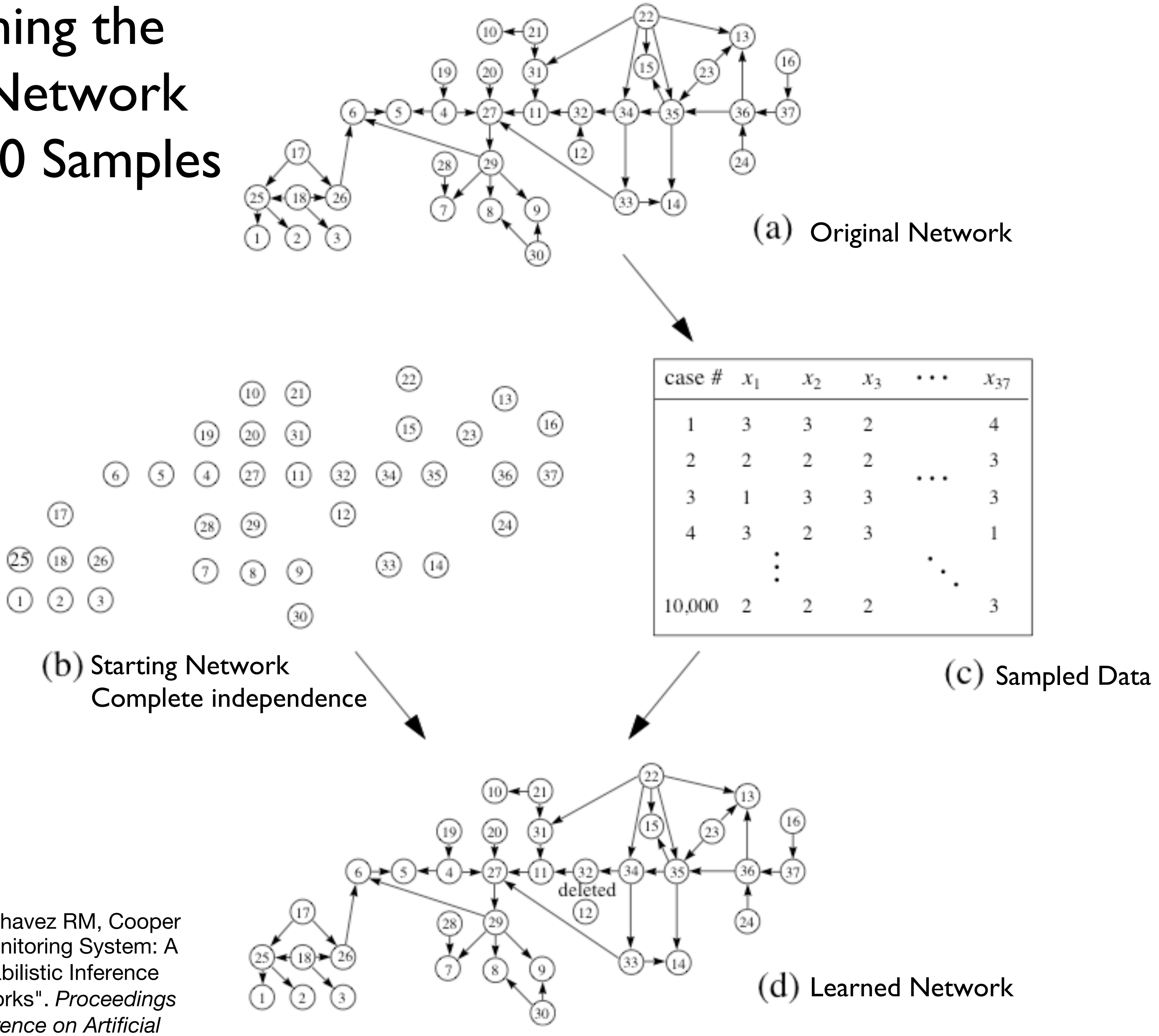
marginal likelihood

complexity penalty

parameter priors



Re-Learning the ALARM Network from 10,000 Samples



(a) Original Network

(b) Starting Network
Complete independence

(c) Sampled Data

case #	x_1	x_2	x_3	$\cdots$	x_{37}
1	3	3	2		4
2	2	2	2	$\cdots$	3
3	1	3	3		3
4	3	2	3		1
	$\vdots$			$\ddots$	
10,000	2	2	2		3

(d) Learned Network

Beinlich I, Suermondt HJ, Chavez RM, Cooper GF (1989). "The ALARM Monitoring System: A Case Study with Two Probabilistic Inference Techniques for Belief Networks". *Proceedings of the 2nd European Conference on Artificial Intelligence in Medicine*, 247–256.

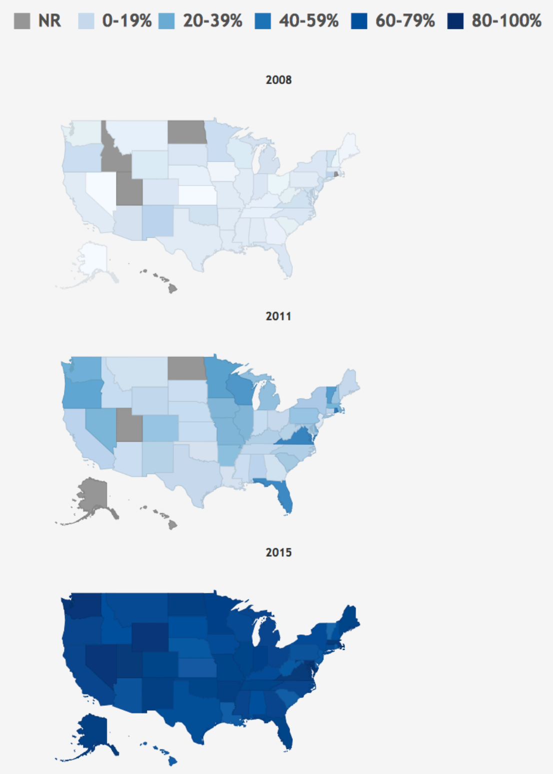
Evolution of AI Approaches

- 1950s to 1960s — Simple Probabilistic Methods
- 1950s to 1980s — Symbolic Methods
- Since 1980s — Sophisticated Probabilistic Methods
- Since 1990s — Big Data
  - Vast amounts of data are being collected “in the wild"
  - Even simple methods work well with enough data
  - Can observational data substitute for trial data?

The Advent of Clinical Data: ~1995-2015

- US National Adoption of Hospital EHR’s

Figure 4: Percent of non-federal acute care hospitals with adoption of at least a Basic EHR system at the State-Level for years 2008, 2011, and 2015



EXPERT OPINION

Contact Editor: **Brian Brannon,** bbrannon@computer.org

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, ***Google***

Eugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"[1] examines why so much of physics can be neatly explained with simple mathematical formulas

such as $f = ma$ or $e = mc^2$. Meanwhile, sciences that involve human beings rather than elementary particles have proven more resistant to elegant math-

behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

Learning from Text at Web Scale

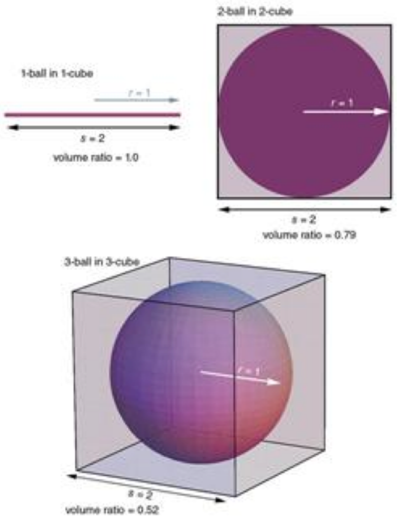
The biggest successes in natural-language-related machine learning have been statistical speech recognition and statistical machine translation. The reason for these successes is not that these tasks are

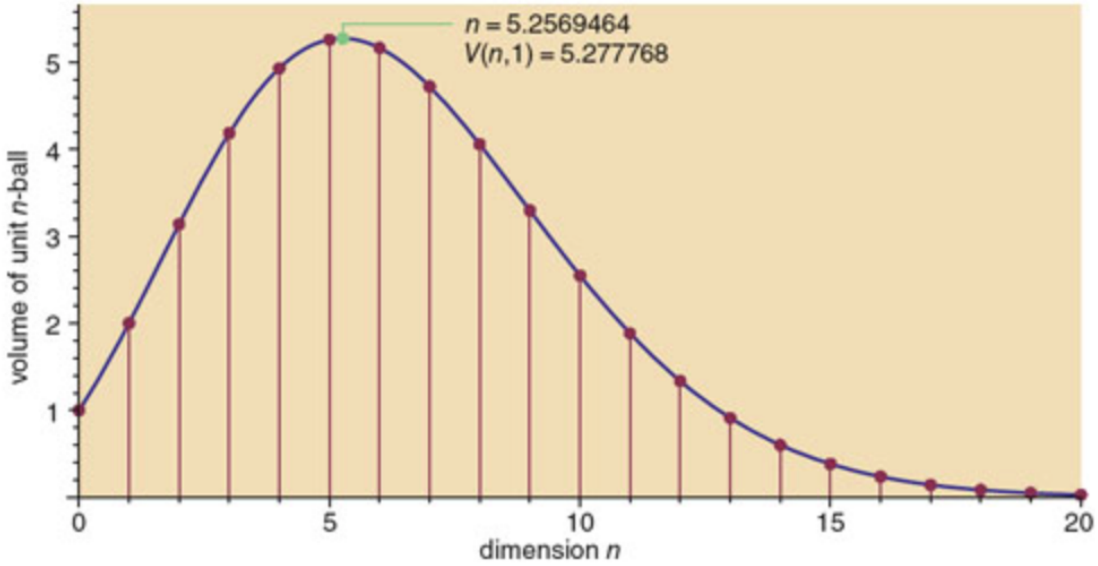
Peter Norvig
IEEE Intelligent Systems, 2009

... a large training set of the input-output behavior that we seek to automate is available to us in the wild.

Google’s Lessons

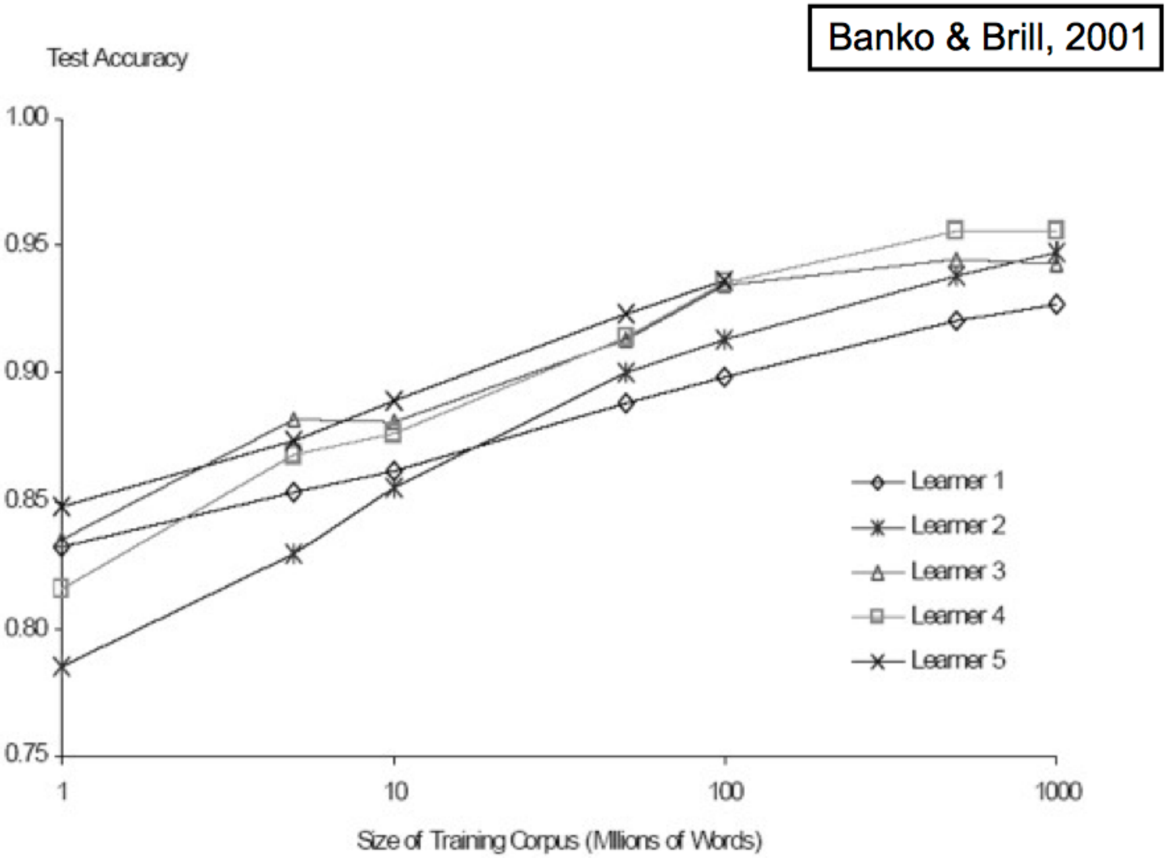
- Much of human knowledge is not like physics!
- “... invariably, simple models and a lot of data trump more elaborate models based on less data”
- “... simple n-gram models or linear classifiers based on millions of specific features perform better than elaborate models that try to discover general rules”
- “... all the experimental evidence from the last decade suggests that throwing away rare events is almost always a bad idea, because much Web data consists of individually rare but collectively frequent events”





Brian Hayes, http://www.americanscientist.org/issues/pub/an-adventure-in-the-nth-dimension

More Data vs. Better Algorithms (Word Sense Disambiguation Task)



Banko & Brill, 2001

Figure 2. Learning Curves for Confusable Disambiguation

http://www.stanford.edu/group/mmds/slides2010/Norvig.pdf

What is Evidence-Based Medicine?
RCTs, Meta-Analysis, Systematic Reviews

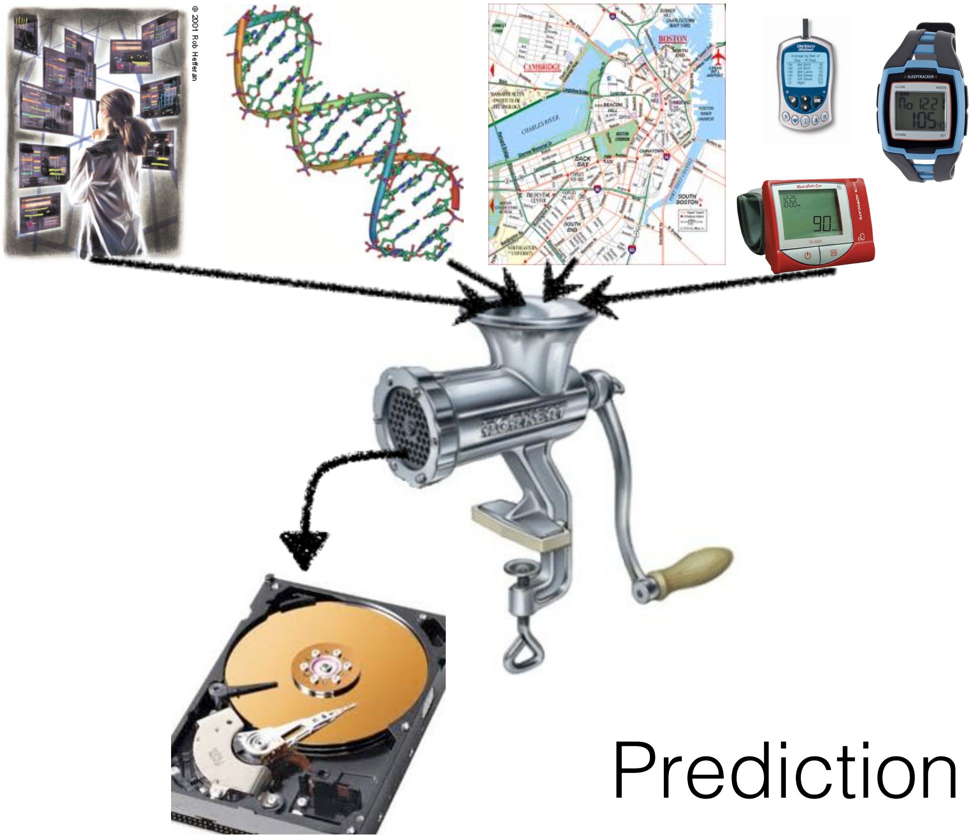
- Randomized Controlled Clinical Trials
  - E.g., is drug A more effective than drug B for condition X?
  - Narrow selection of patient cases and controls
  - Careful collection of systematically organized data
  - Statistical analysis of outcomes

  => Statistically significant conclusions
- But:
  - **Heterogeneity:** Most cases to which RCT results are applied do not fit trial criteria
  - **Short Follow-Up:** Trials run for limited times, but use is longer
  - **Small Samples:** Some effects are rare but devastating
- **Instead**: consider every patient's experience as a source of knowledge by which to improve health care
  - "The Learning Health Care System": "one in which progress in science, informatics, and care culture align to generate new knowledge as an ongoing, natural by-product of the care experience, and seamlessly refine and deliver best practices for continuous improvement in health and health care" —IOM (now National Academy of Medicine)

“The Learning Health Care System”

- “one in which progress in science, informatics, and care culture align to generate new knowledge as an ongoing, natural by-product of the care experience, and seamlessly refine and deliver best practices for continuous improvement in health and health care” —IOM (now National Academy of Medicine)
- Needs not currently met:
  - Comprehensive collation of all clinical, social, demographic, behavioral, ... data that are now captured in the health care system
  - Routine capture of novel data sources:
    - genomes, gene expression, etc.
    - environmental factors (e.g., metagenomics)
    - physiological response to life situations
      - (related to fitness and wellness)
  - Technical infrastructure
    - Storage and analysis of truly “big data”
  - Incentives and demonstrations of utility

Use All Possible Data

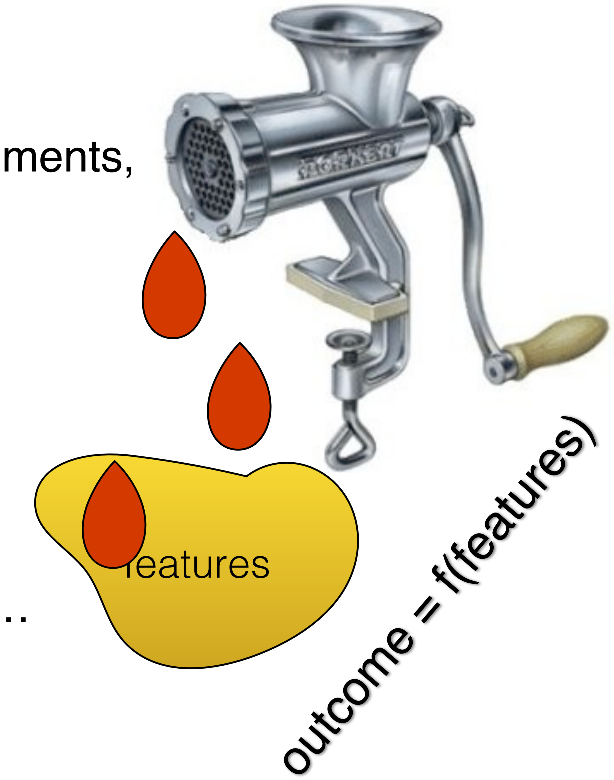


© 2001 Rob Hefferan

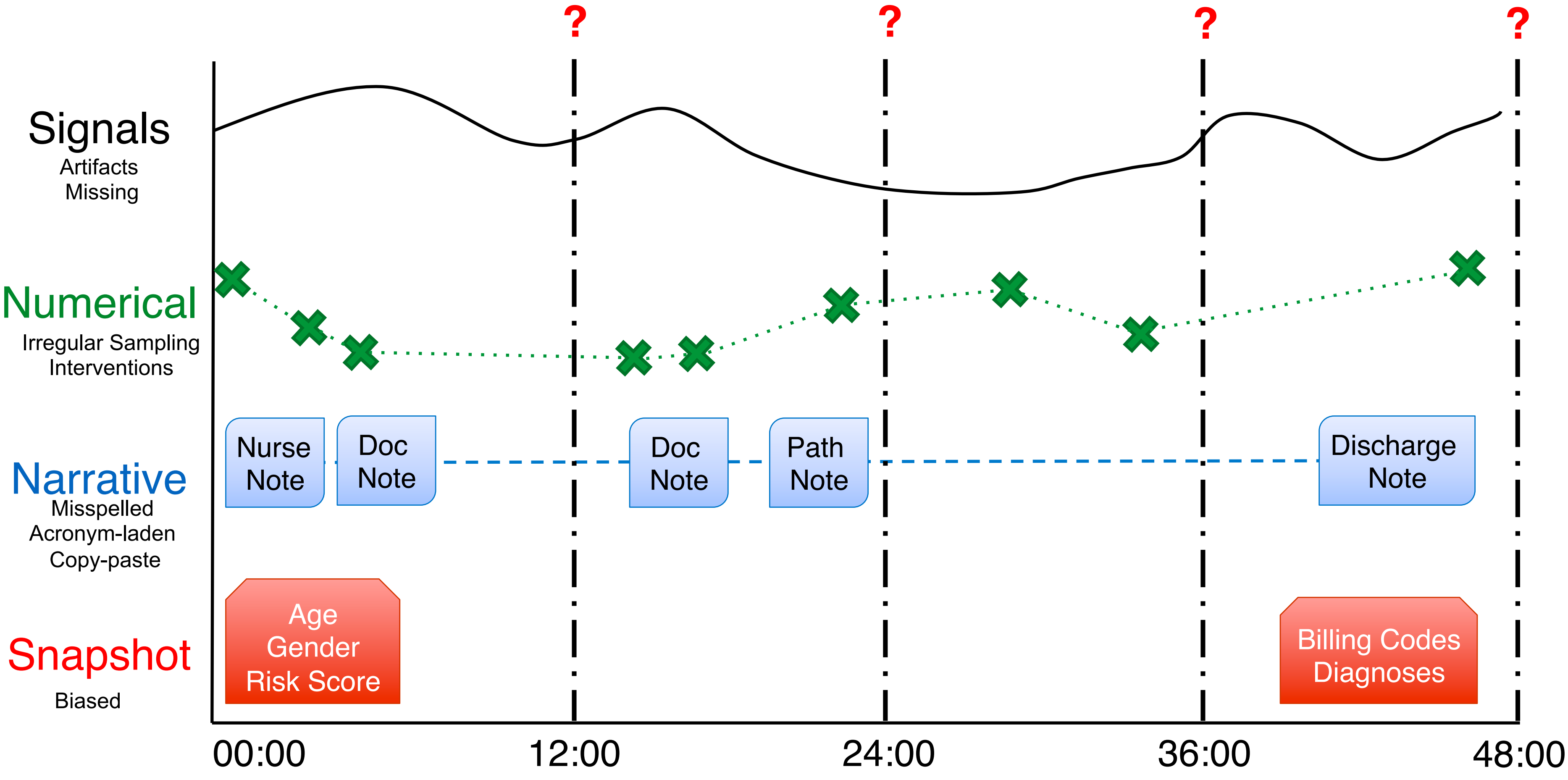
Prediction = f(inputs)

Heterogeneous Sources of Clinical Data

- **Tabular**, standardized data
  - Billing codes: encounter, most important conditions, severity, interventions, …
  - Demographics
  - Laboratory measurements
  - Medication prescriptions
  - Discrete measurements from monitors, wearable instruments,
- Continuously recorded **signals**
  - Bedside monitors, wearables, …
- **Narrative** reports
  - Doctors' and nurses' notes
  - Specialist reports: pathology, radiology, …
  - Patient self-observations, blogs, email exchanges, …
- **Questionnaires**: smoking, drugs, exercise habits, travel, …
- **Imaging**: x-rays, CT scans, MRI, PET, ultrasound, …
- **Environmental** conditions: pollution, epidemics, …
- **Genetics:** gene expression, SNP, CNV, exome, genome, epigenetics, metagenetics, …



Real Clinical Data is “Messy”



Slide from Dr. Maryzeh Ghassemi

What if we had all these data?
⇒ *Decision Support*

- For patients
  - Manage ongoing care of chronic conditions, titrate medications, change behaviors, notice actionable conditions, …
- For providers
  - Double-check data interpretation, improve diagnostic acumen, optimize therapeutic decisions, predict outcomes, help to follow up, …
- For institutions and public health
  - Focus on critical needs, rationalize priorities, organize public education, …
- For researchers
  - Provide insights into the science and practice of medicine
- Key: ability to ***predict future events*** in individuals or population
- How?
  - prediction = f(features)

Using MIMIC (ICU) Data to Build Predictive Models

http://dspace.mit.edu/handle/1721.1/46690

- Mortality
  - Comparison to SAPS II
  - Daily Acuity Scores
  - Real-time Acuity Scores (real-time risk assessment)
- Other clinical events
  - pressor weaning, intra-aortic balloon pump weaning
  - onset of septic shock, acute kidney injury
- Data set (MIMIC 2, earlier snapshot, today ~5-6x)
  - 10,066 patients: 7,048 development, 3,018 validation
  - selected cases with adequate data
  - excluded neurological and trauma cases
  - **only tabular data**
- Derived variables can summarize essential contributions of dynamic variation
  - integrals, slopes, ranges, frequencies, etc.
  - Transformed variables: inverse, abs, square, square root, log-abs, abs deviation from mean, log abs deviation, ...

https://upload.wikimedia.org/wikipedia/commons/a/a7/Respiratory_therapist.jpg

Example of Mortality Models

Based only on numeric data: labs, monitors, bedside measurements, predict 30-day mortality.

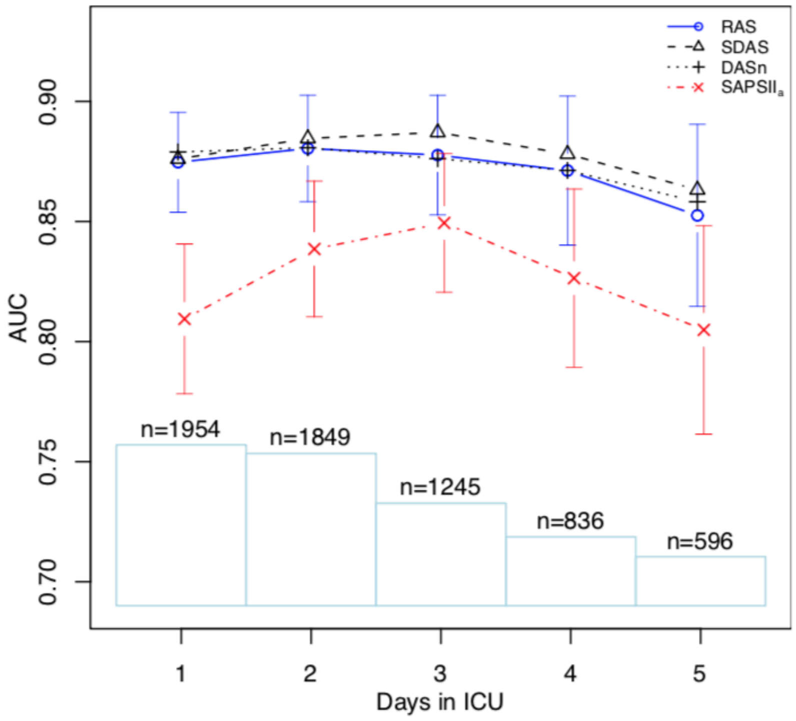


Figure 5-25: AUC versus day, first 5 ICU days (validation data). The 95% confidence intervals are shown for the RAS and SAPSII$_a$ performances.

Hug, C. W., & Szolovits, P. (2009). ICU acuity: real-time models versus daily models. *AMIA Annual Symposium Proceedings / AMIA Symposium AMIA Symposium*, *2009*, 260–264.

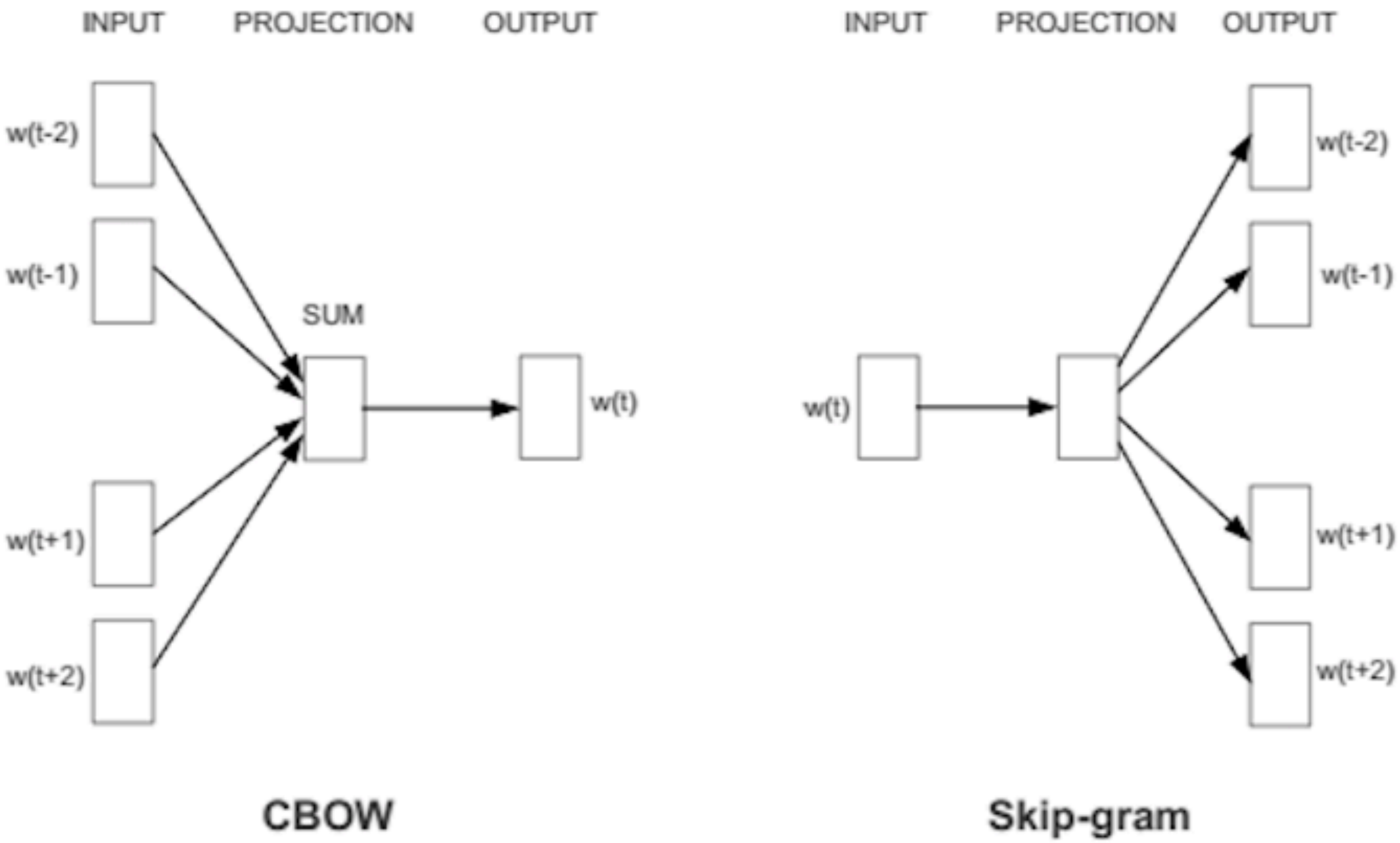
Mortality, Therapeutic Opportunities and Risks (MIMIC ICU Numeric Data)

Prediction	AUC
30-day mortality	0.89
Vasopressor weaning + Survival	0.83
Weaning from Intra-Aortic Balloon Pump	0.82
Onset of Septic Shock	0.84
Acute kidney injury	0.74

Evolution of AI Approaches

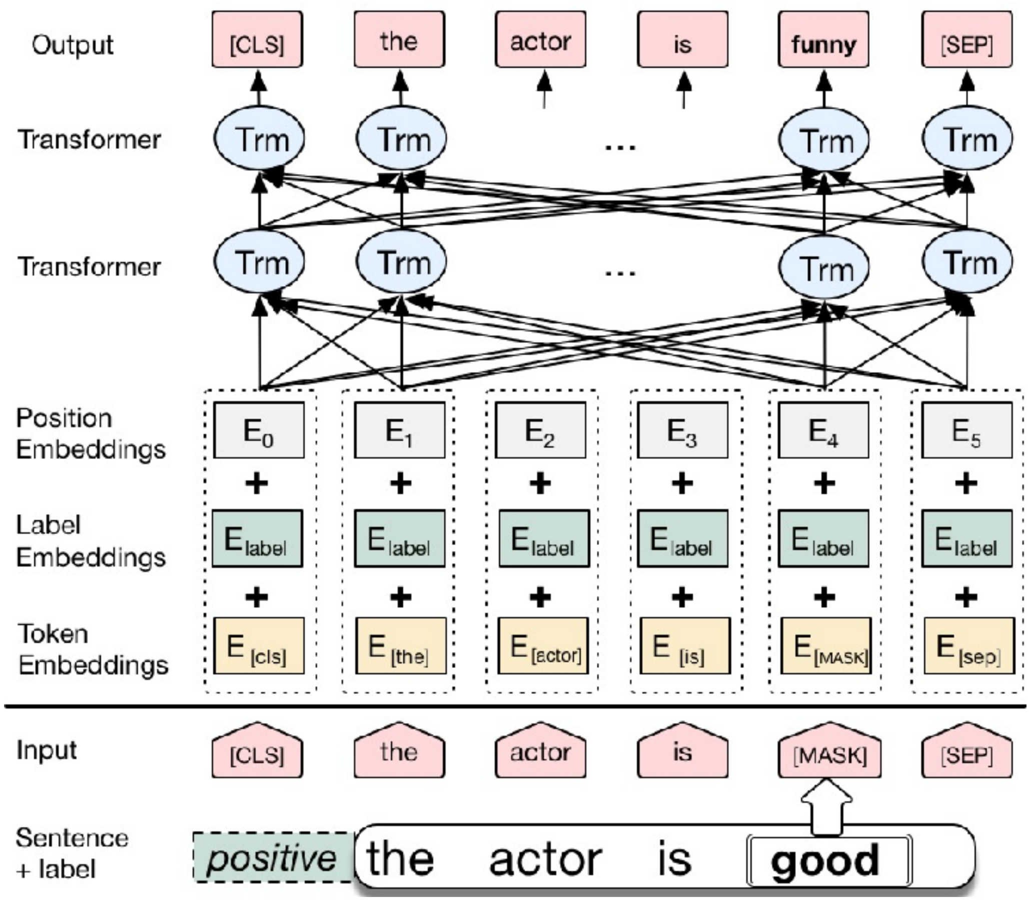
- 1950s to 1980s — Symbolic Methods
- Since 1980s — Probabilistic Methods
- Since 1990s — Big Data
- In 2010s — Artificial Neural Networks
  - Finding vector space representations or all data
  - From discrete to continuous optimization
  - Architectures for applying ANN

Turning Words into Vectors



Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013, October 16). Distributed Representations of Words and Phrases and their Compositionality. arXiv.

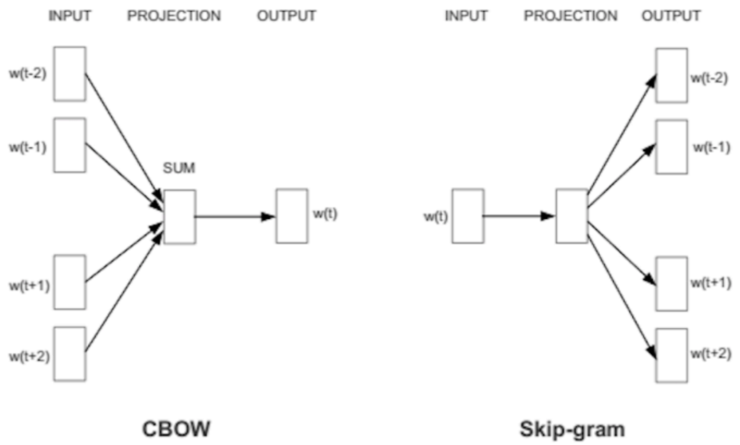
Turning Words into Vectors



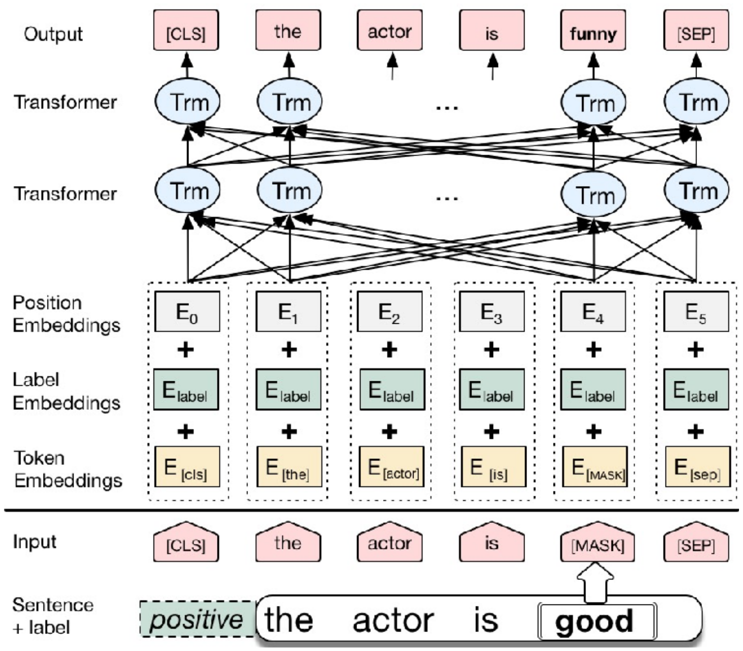
https://www.semanticscholar.org/paper/Conditional-BERT-Contextual-Augmentation-Wu-Lv/188024469a2443f262b3cbb5c5d4a96851949d68

How to Turn Discrete Data into Vector Spaces

- Word2Vec



- Contextual Embeddings
    - BERT, ELMO, GPT-2, …

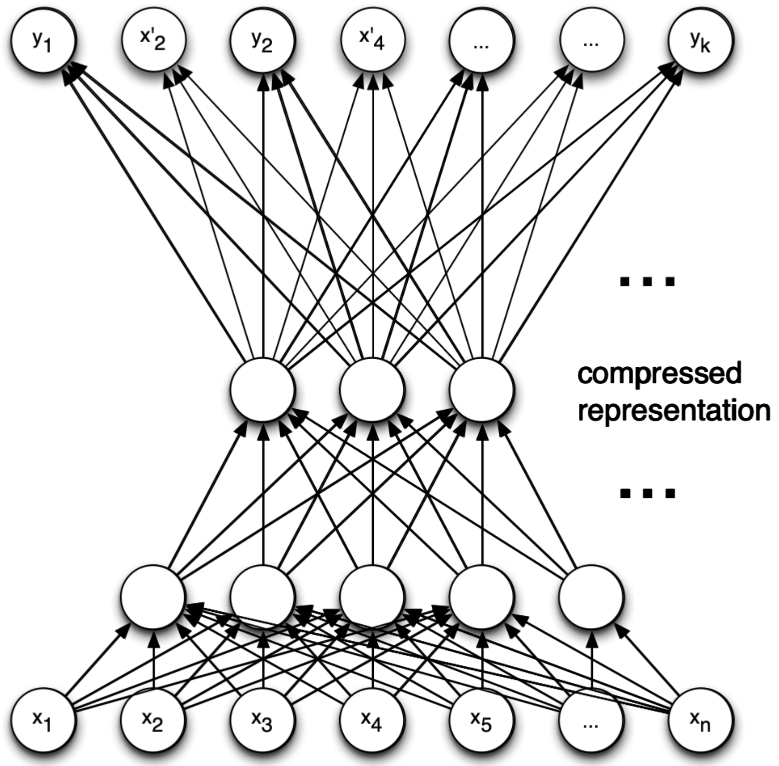


https://www.semanticscholar.org/paper/Conditional-BERT-Contextual-Augmentation-Wu-Lv/188024469a2443f262b3cbb5c5d4a96851949d68

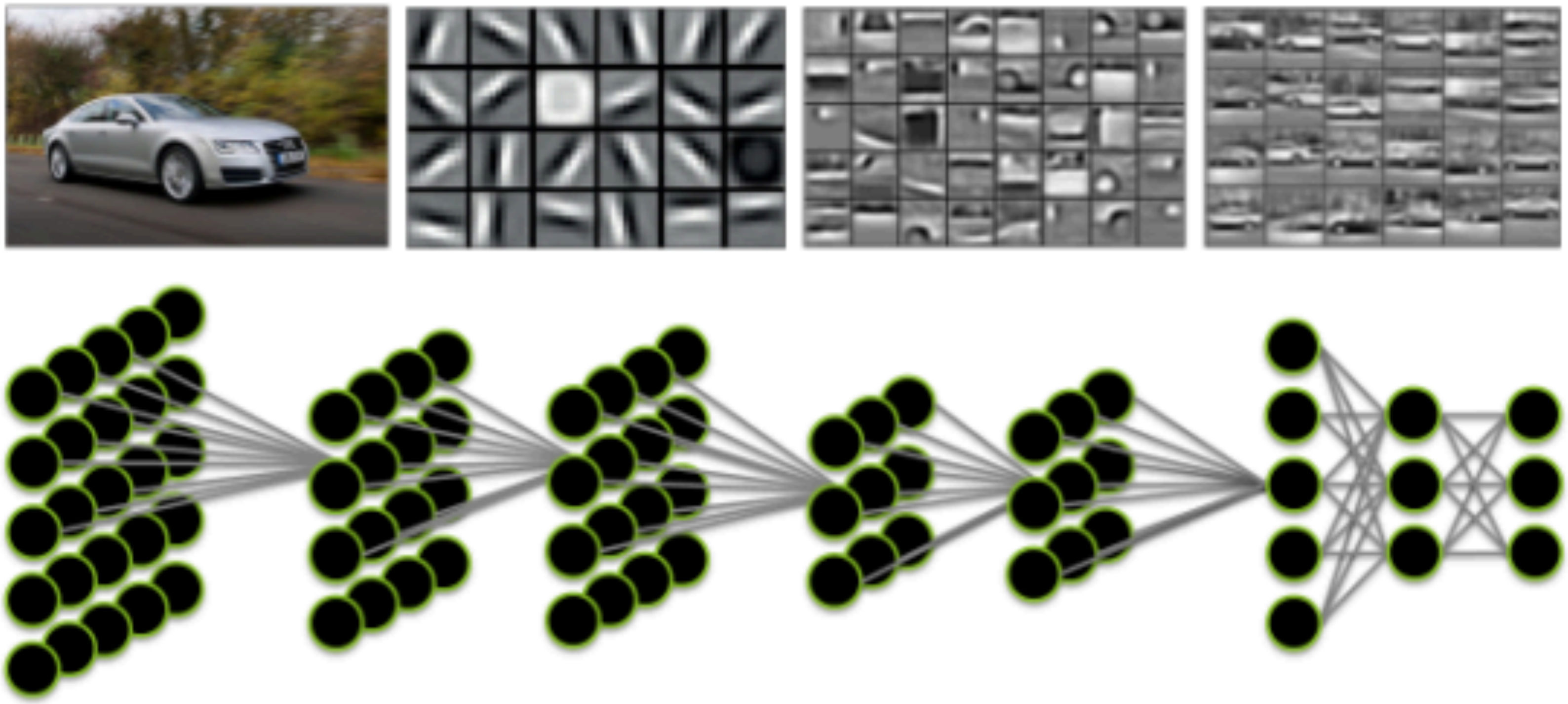
(Deep) Neural Networks

- Every node computes a logistic regression (or some other non-linear) function of its inputs
- Number of nodes in each layer may vary
- Number of layers is another hyper-parameter
- Dropout may omit some fraction of links
- Training by back propagation
  - change weights in proportion to error signal

- **Unsupervised**: Train to optimize unsupervised compression
- **Supervised**: Train to objective function on gold standard data



HOW A DEEP NEURAL NETWORK SEES

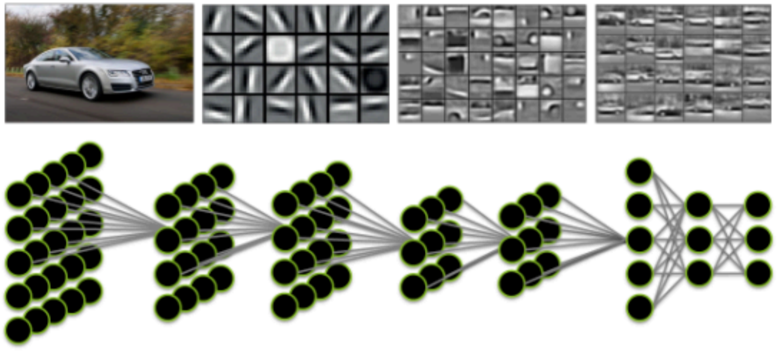


https://blogs.nvidia.com/blog/2018/09/05/whats-the-difference-between-a-cnn-and-an-rnn/

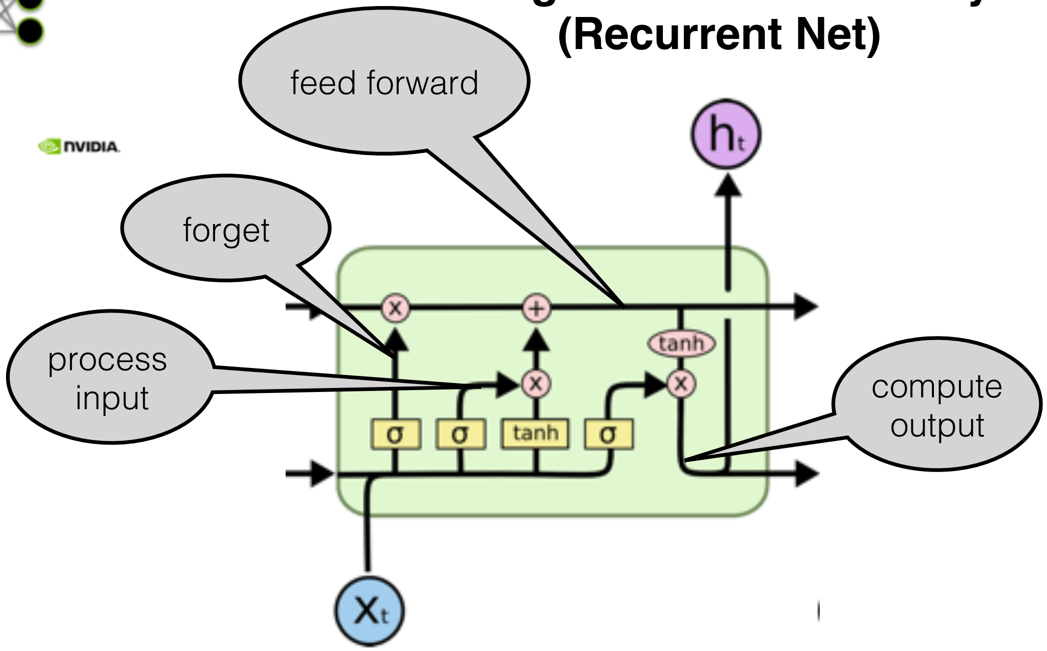
Convolutional and Recurrent Networks

HOW A DEEP NEURAL NETWORK SEES

https://blogs.nvidia.com/blog/2018/09/05/whats-the-difference-between-a-cnn-and-an-rnn/



Long Short Term Memory (Recurrent Net)



feed forward

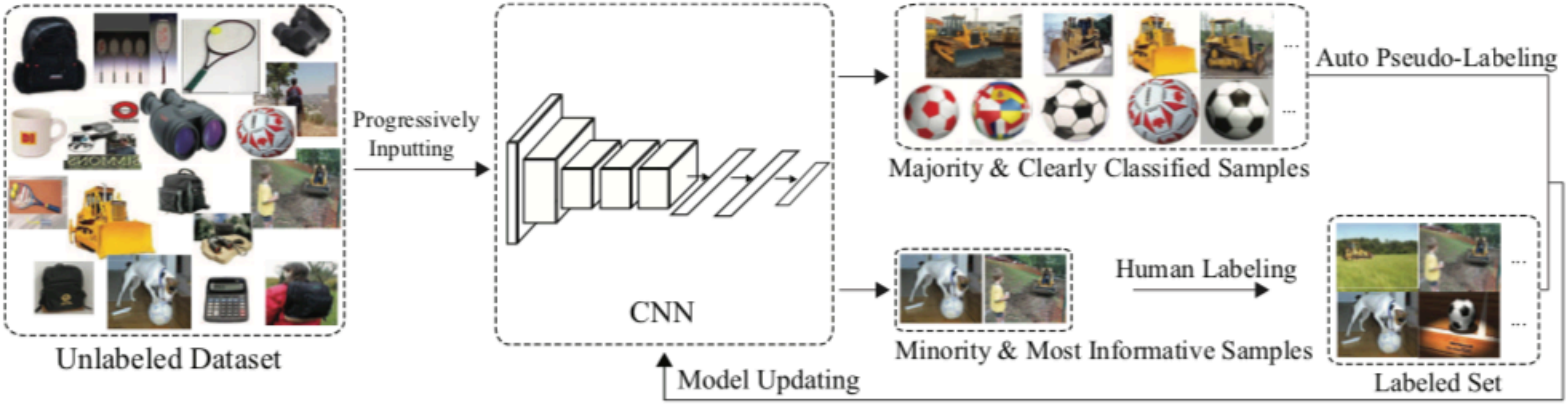
forget

process input

compute output

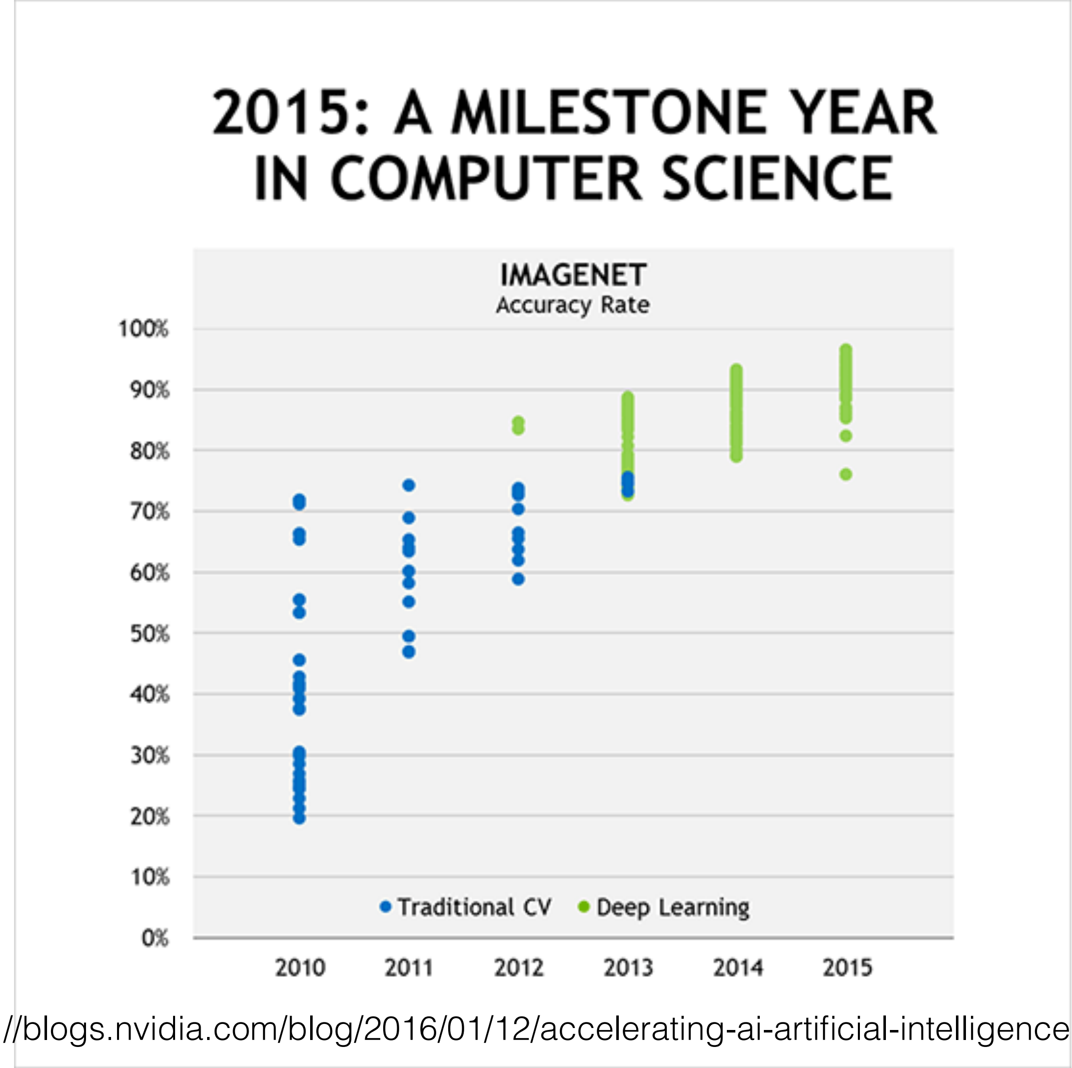
https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Deep Neural Networks for Image Classification



Wang, K., Zhang, D., Li, Y., Zhang, R. and Lin, L. Cost-Effective Active Learning for Deep Image Classification. IEEE Transactions on Circuits and Systems for Video Technology, 2017. https://arxiv.org/pdf/1701.03551.pdf

“Deep Learning Performs Miracles”



https://blogs.nvidia.com/blog/2016/01/12/accelerating-ai-artificial-intelligence-gpus/

Deep Learning Progress

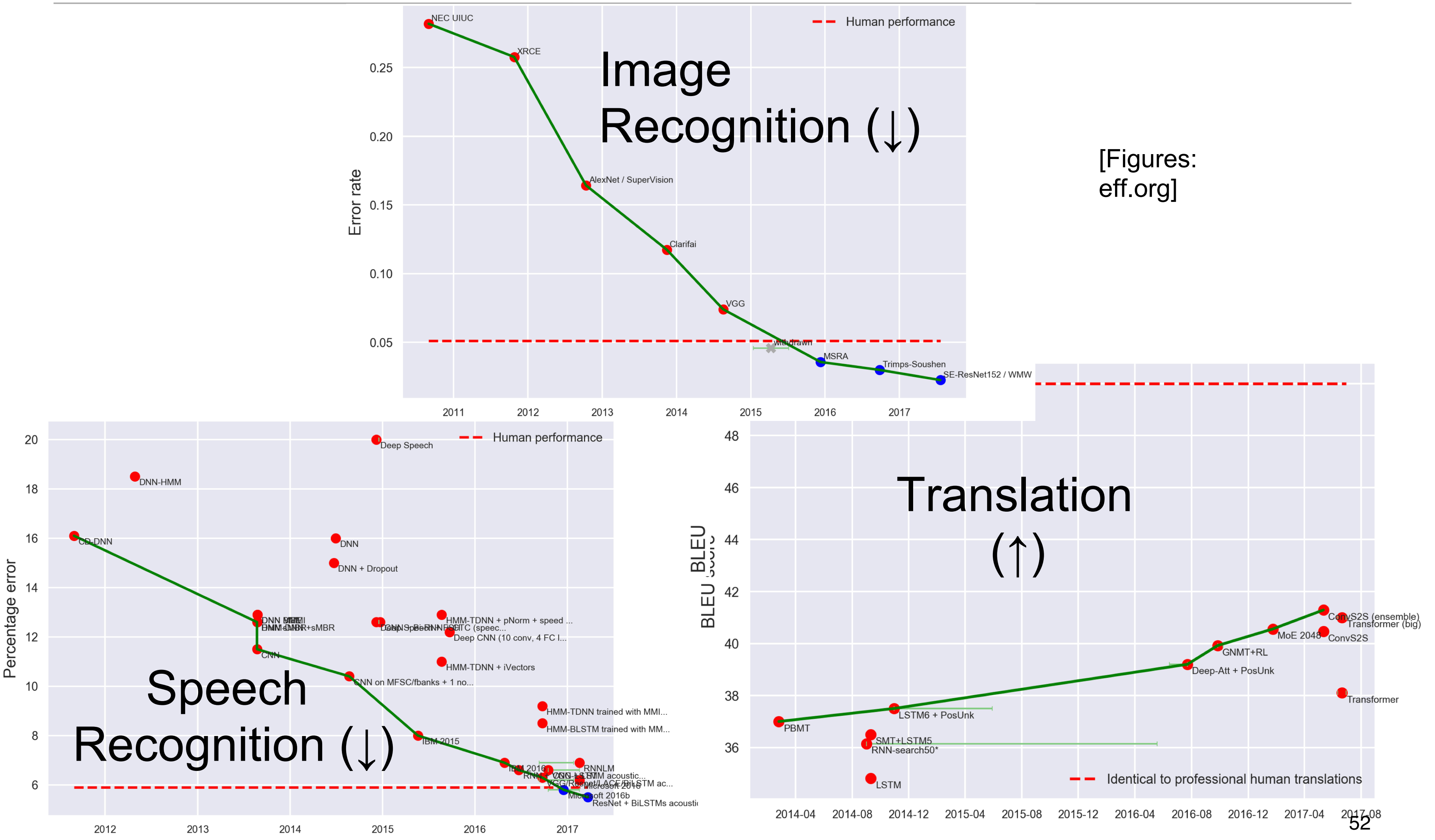


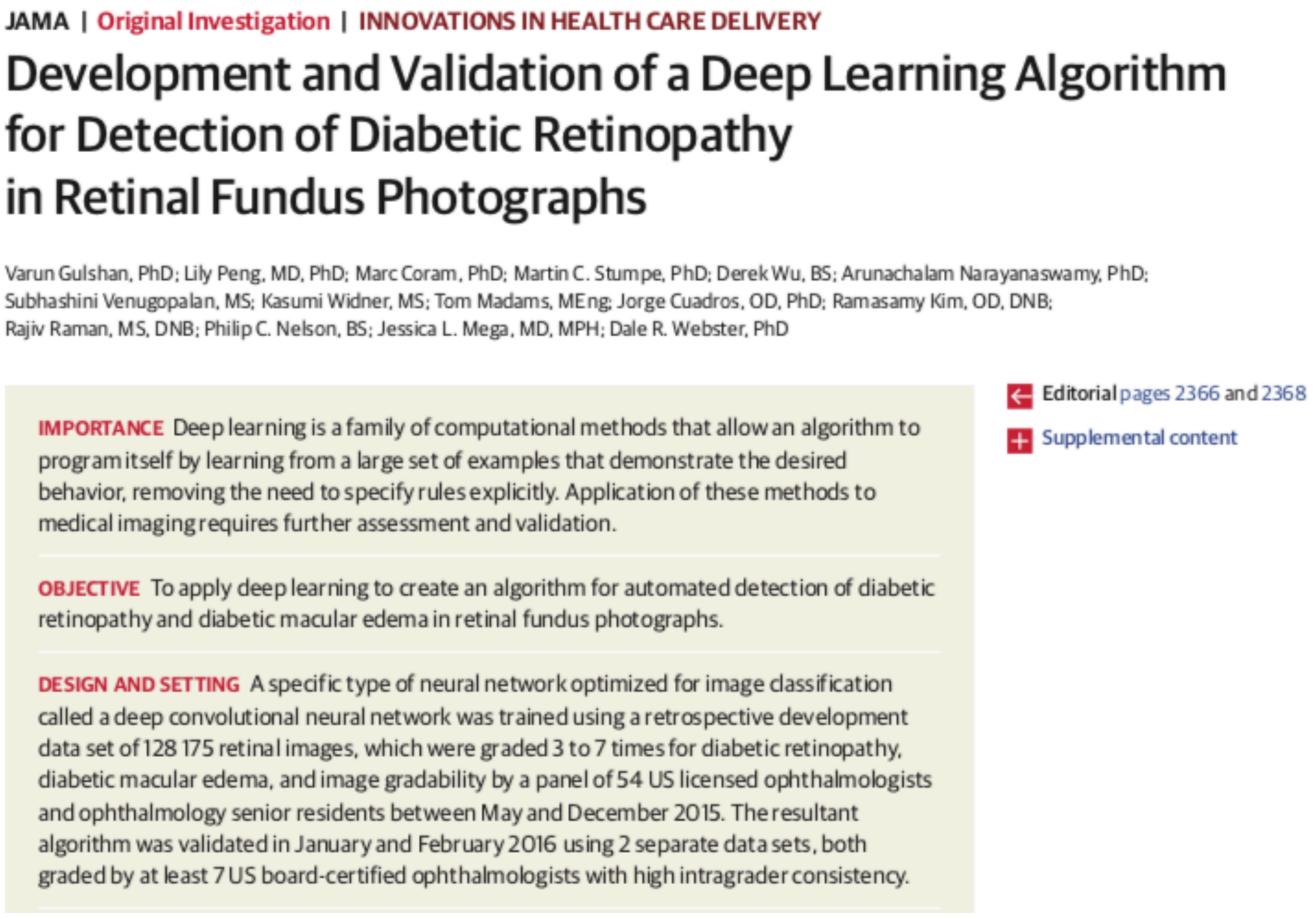
Image Recognition (↓)

Speech Recognition (↓)

Translation (↑)

[Figures: eff.org]

Image Analysis for Diabetic Retinopathy



JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD

Editorial pages 2366 and 2368

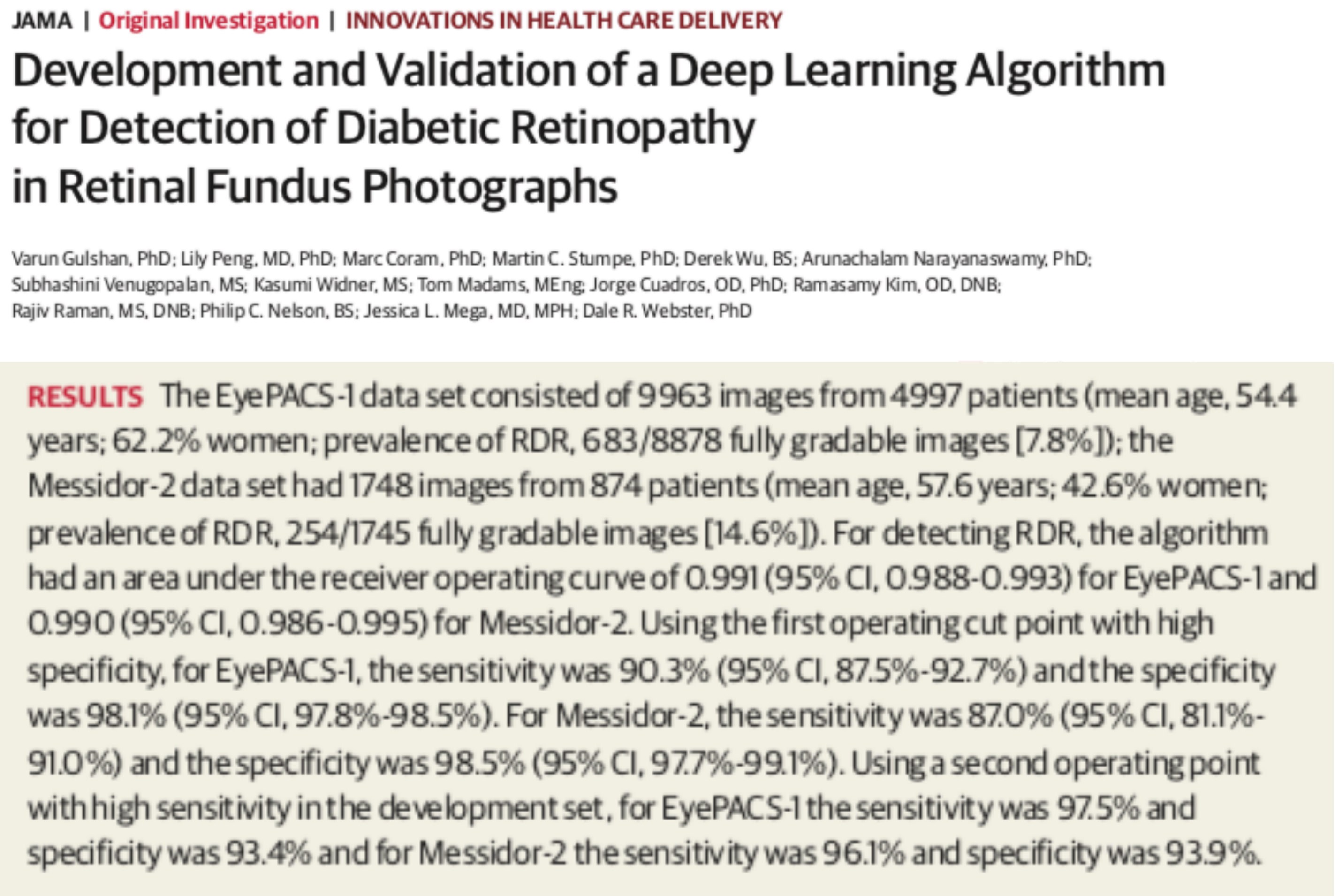
Supplemental content

IMPORTANCE Deep learning is a family of computational methods that allow an algorithm to program itself by learning from a large set of examples that demonstrate the desired behavior, removing the need to specify rules explicitly. Application of these methods to medical imaging requires further assessment and validation.

OBJECTIVE To apply deep learning to create an algorithm for automated detection of diabetic retinopathy and diabetic macular edema in retinal fundus photographs.

DESIGN AND SETTING A specific type of neural network optimized for image classification called a deep convolutional neural network was trained using a retrospective development data set of 128 175 retinal images, which were graded 3 to 7 times for diabetic retinopathy, diabetic macular edema, and image gradability by a panel of 54 US licensed ophthalmologists and ophthalmology senior residents between May and December 2015. The resultant algorithm was validated in January and February 2016 using 2 separate data sets, both graded by at least 7 US board-certified ophthalmologists with high intragrader consistency.

Image Analysis for Diabetic Retinopathy



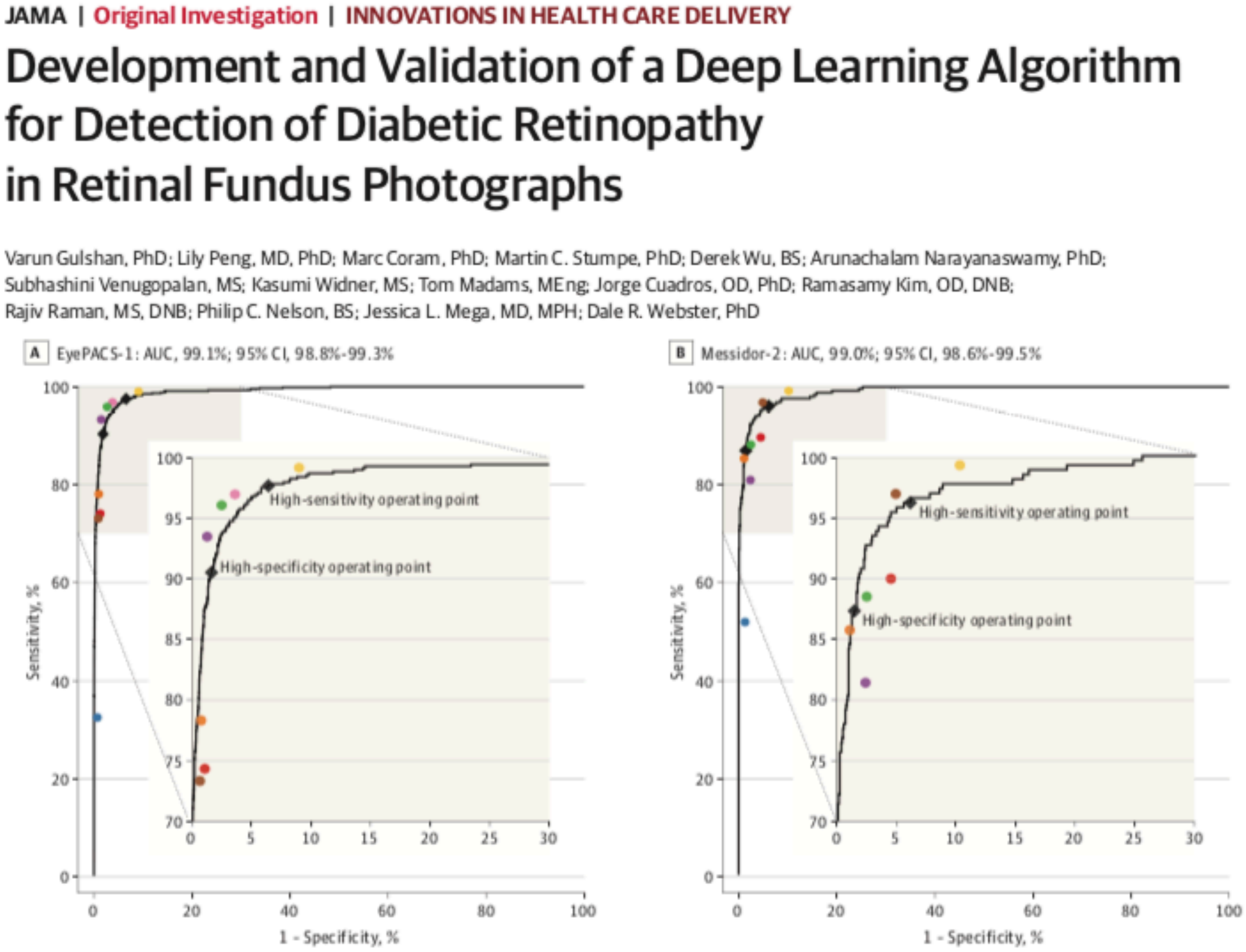
JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD

RESULTS The EyePACS-1 data set consisted of 9963 images from 4997 patients (mean age, 54.4 years; 62.2% women; prevalence of RDR, 683/8878 fully gradable images [7.8%]); the Messidor-2 data set had 1748 images from 874 patients (mean age, 57.6 years; 42.6% women; prevalence of RDR, 254/1745 fully gradable images [14.6%]). For detecting RDR, the algorithm had an area under the receiver operating curve of 0.991 (95% CI, 0.988-0.993) for EyePACS-1 and 0.990 (95% CI, 0.986-0.995) for Messidor-2. Using the first operating cut point with high specificity, for EyePACS-1, the sensitivity was 90.3% (95% CI, 87.5%-92.7%) and the specificity was 98.1% (95% CI, 97.8%-98.5%). For Messidor-2, the sensitivity was 87.0% (95% CI, 81.1%-91.0%) and the specificity was 98.5% (95% CI, 97.7%-99.1%). Using a second operating point with high sensitivity in the development set, for EyePACS-1 the sensitivity was 97.5% and specificity was 93.4% and for Messidor-2 the sensitivity was 96.1% and specificity was 93.9%.

Image Analysis for Diabetic Retinopathy



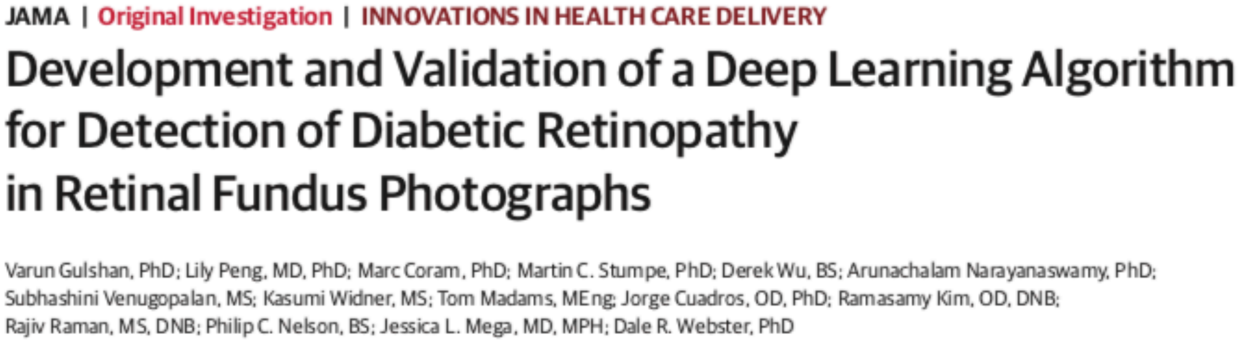
JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD

A
EyePACS-1: AUC, 99.1%; 95% CI, 98.8%-99.3%
Sensitivity, %
1 - Specificity, %
0
20
40
60
80
100
70
75
80
85
90
95
5
10
15
25
30
High-sensitivity operating point
High-specificity operating point
B
Messidor-2: AUC, 99.0%; 95% CI, 98.6%-99.5%
Sensitivity, %
1 - Specificity, %
High-sensitivity operating point
High-specificity operating point

Image Analysis for Diabetic Retinopathy



JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

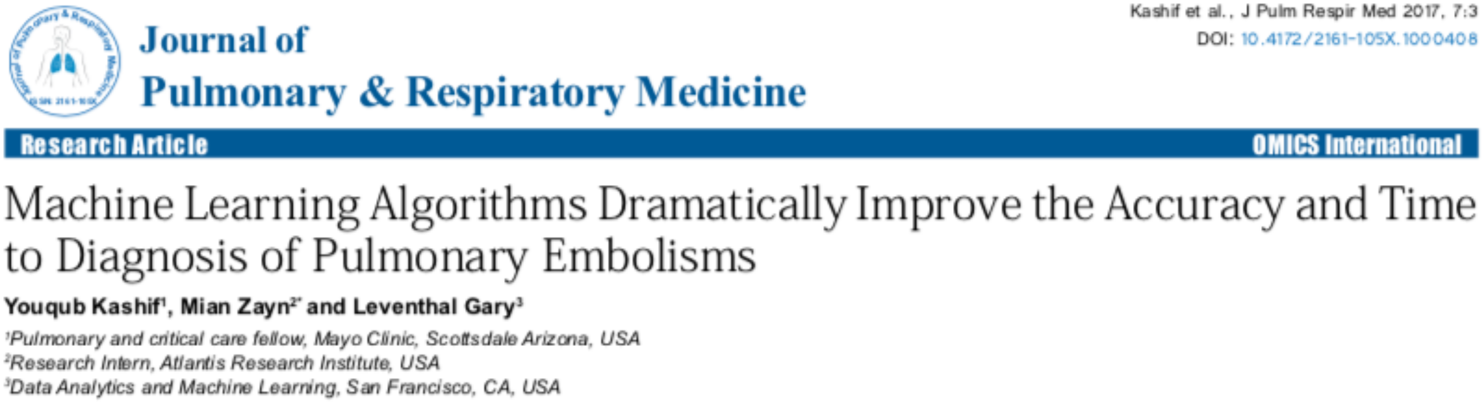
Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD

2016

	AUC	Favoring Specificity		Favoring Sensitivity	
		Sensitivity	Specificity	Sensitivity	Specificity
EyePACS-1 9963 images 4997 patients	0.991	90.3	**98.1**	**97.5**	93.4
Messidor-2 1748 images 874 patients	0.990	87.0	**98.5**	**96.1**	93.9

Image Analysis for Pulmonary Emboli



Kashif et al., J Pulm Respir Med 2017, 7:3
DOI: 10.4172/2161-105X.1000408

Journal of
Pulmonary & Respiratory Medicine

Research Article | OMICS International

Machine Learning Algorithms Dramatically Improve the Accuracy and Time to Diagnosis of Pulmonary Embolisms

Youqub Kashif[1], Mian Zayn[2*] and Leventhal Gary[3]

[1]Pulmonary and critical care fellow, Mayo Clinic, Scottsdale Arizona, USA
[2]Research Intern, Atlantis Research Institute, USA
[3]Data Analytics and Machine Learning, San Francisco, CA, USA

First pass analysis demonstrated 20 (600 case cohort) false positive results. … Zero false negative results were reported. 580 results were in concordance with the official report.

Second and third pass results demonstrated … only 3 false positive results related to incomplete/non-diagnostic contrast bolus.

Image Analysis for Pulmonary Emboli

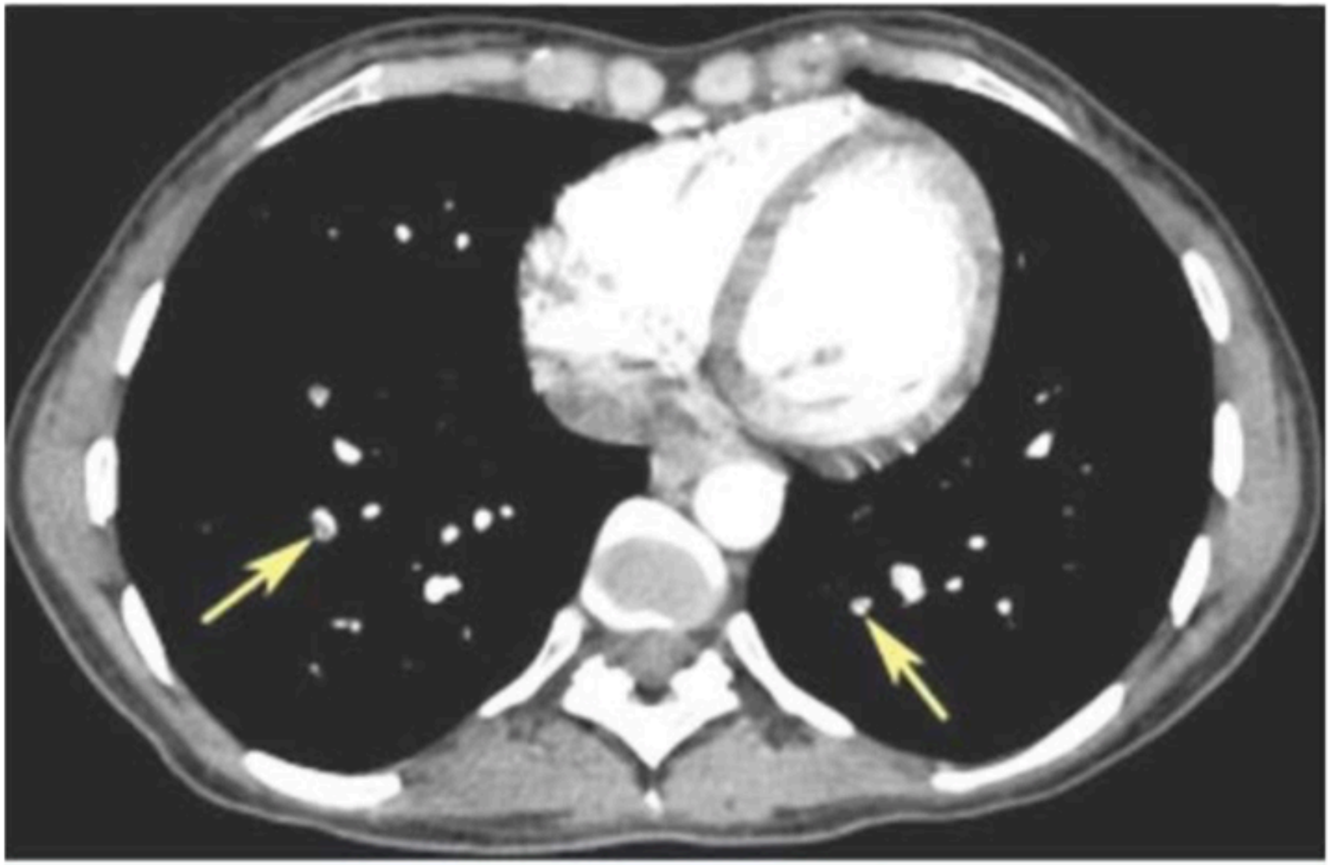
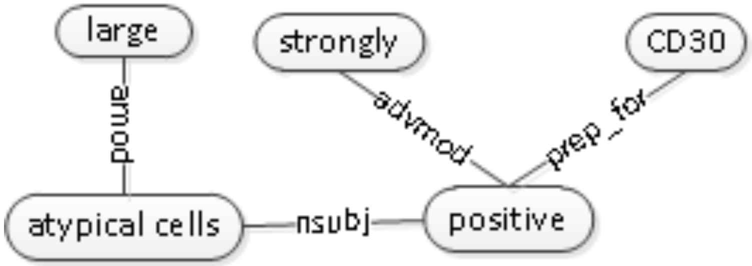


Figure 2: Bilateral pulmonary emboli in third and fourth order branch points of the pulmonary arteries. Small emboli distally located can be a diagnostic challenge.

Where do Features Come From? “Natural Features”

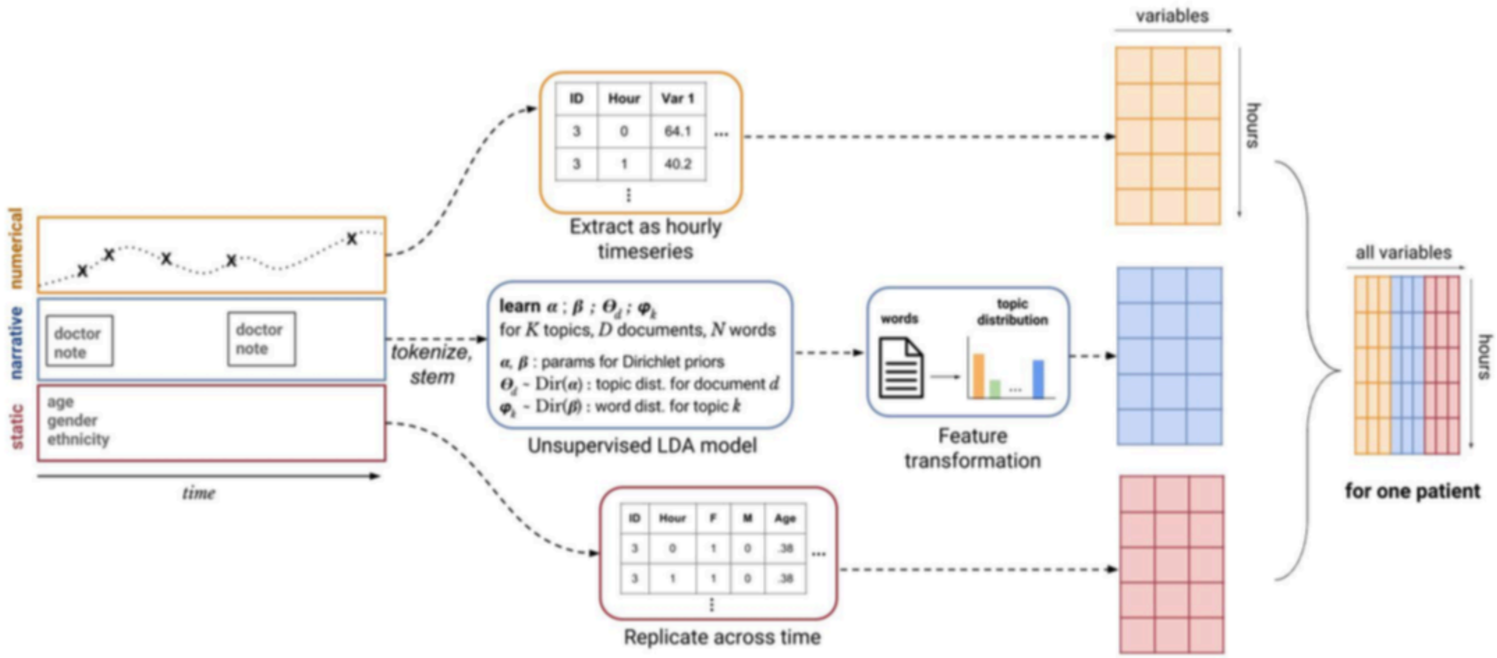
- Coded data
  - lab values
  - demographics (age, gender, ethnicity, socio-economic status, marital status)
  - medications
  - symptoms
  - diagnoses
  - procedures
  - insurance status
  - billing codes
  - death records, hospitalizations, ED visits
- Information extracted from narrative notes
  - <all of the above>
  - physical exam, presenting complaint, vital signs, nursing observations, specialists’ reports, …
  - richer structure: drug given for condition, side-effect of treatment, …



Where do Features Come From?
Abstractions

- Temporal trends
  - average, range, standard deviation, mean-crossings, linear trends, each over various lengths of time
- Clustering-based
  - commonly co-occurring values & trends, patient groupings, linguistic structures
    - unsupervised or supervised
  - bi- or tri-clustering, matrix or tensor factorization
    - mutually constrain clustering among different dimensions
  - topic models
    - bag-of-words
    - bag of more primitive features

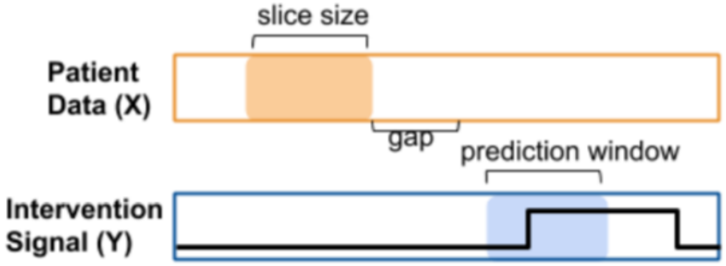
Neural Network Models Applicable to All Data e.g., predicting clinical interventions


numerical
narrative
static
doctor note
doctor note
age
gender
ethnicity
time
tokenize, stem
Extract as hourly timeseries
learn α ; β ; Θ_d ; φ_k
for K topics, D documents, N words
α, β : params for Dirichlet priors
Θ_d ~ Dir(α) : topic dist. for document d
φ_k ~ Dir(β) : word dist. for topic k
Unsupervised LDA model
words
topic distribution
Feature transformation
Replicate across time
variables
hours
all variables
hours
for one patient

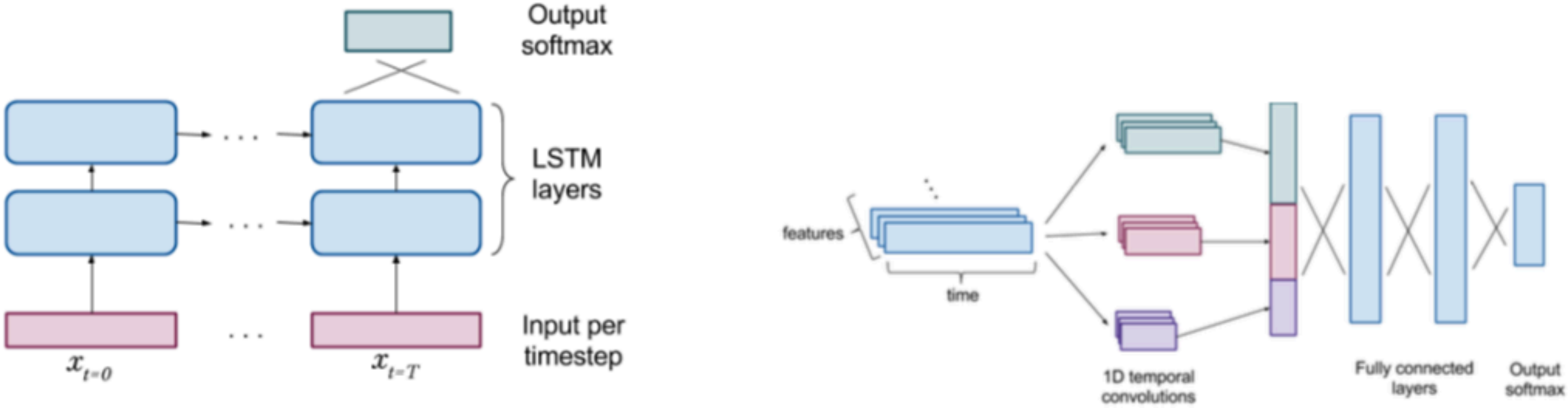
ID	Hour	Var 1
3	0	64.1
3	1	40.2

ID	Hour	F	M	Age
3	0	1	0	.38
3	1	1	0	.38

Figure 1: Data preprocessing and feature extraction with numerical measurements and lab values, clinical notes and static demographics.



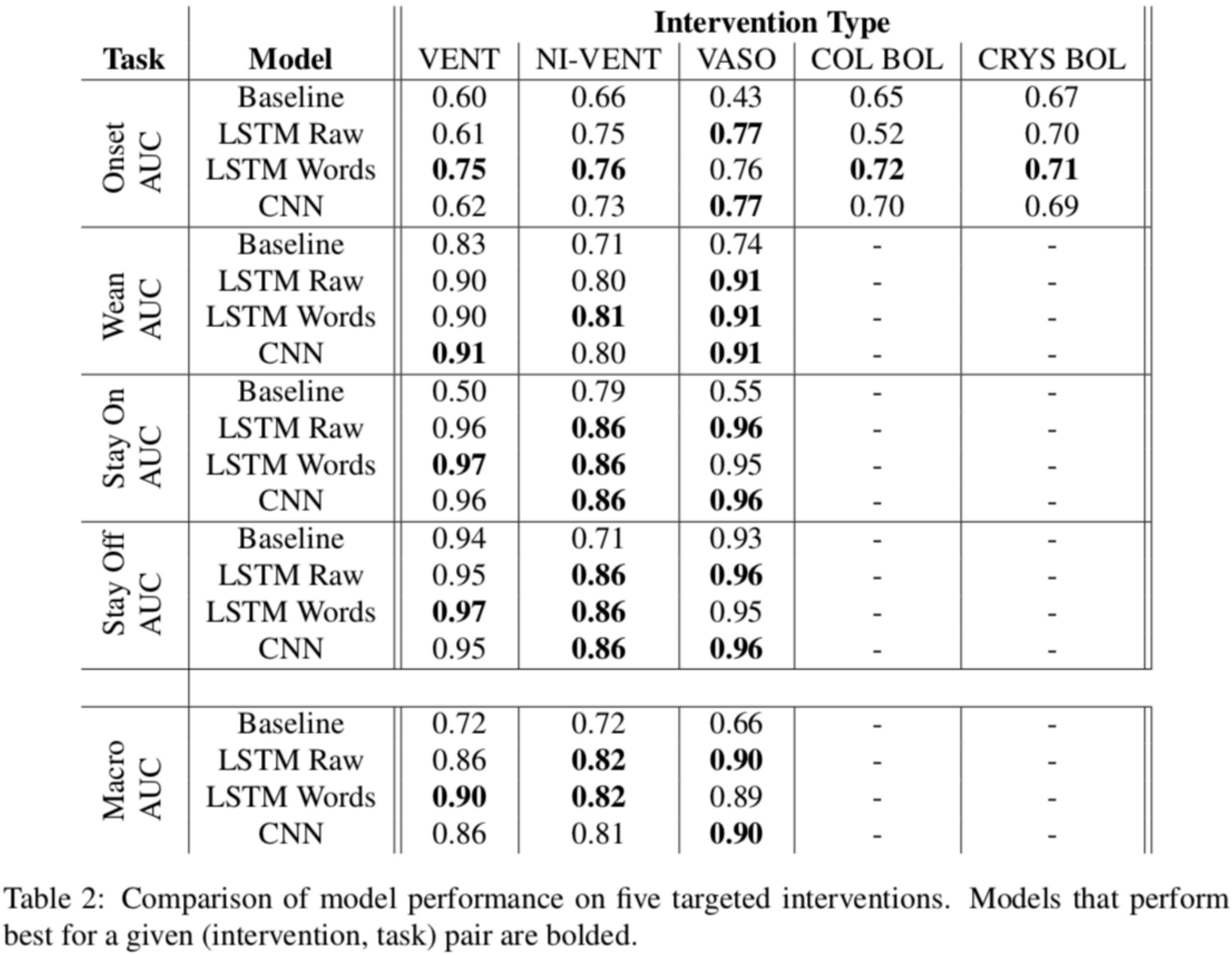
[1] Suresh, H., Hunt, N., Johnson, A., Celi, L. A., Szolovits, P., and Ghassemi, M. **Clinical intervention prediction and understanding with deep neural networks**. In *Proceedings of the 2nd Machine Learning for Healthcare Conference* (Boston, Massachusetts, 18–19 Aug 2017), F. Doshi-Velez, J. Fackler, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, Eds., vol. 68 of *Proceedings of Machine Learning Research*, PMLR, pp. 322–337.



(a) The LSTM consists of two hidden layers with 512 nodes each. We sequentially feed in each hour's data. At the end of the example window, we use the final hidden state to predict the output.

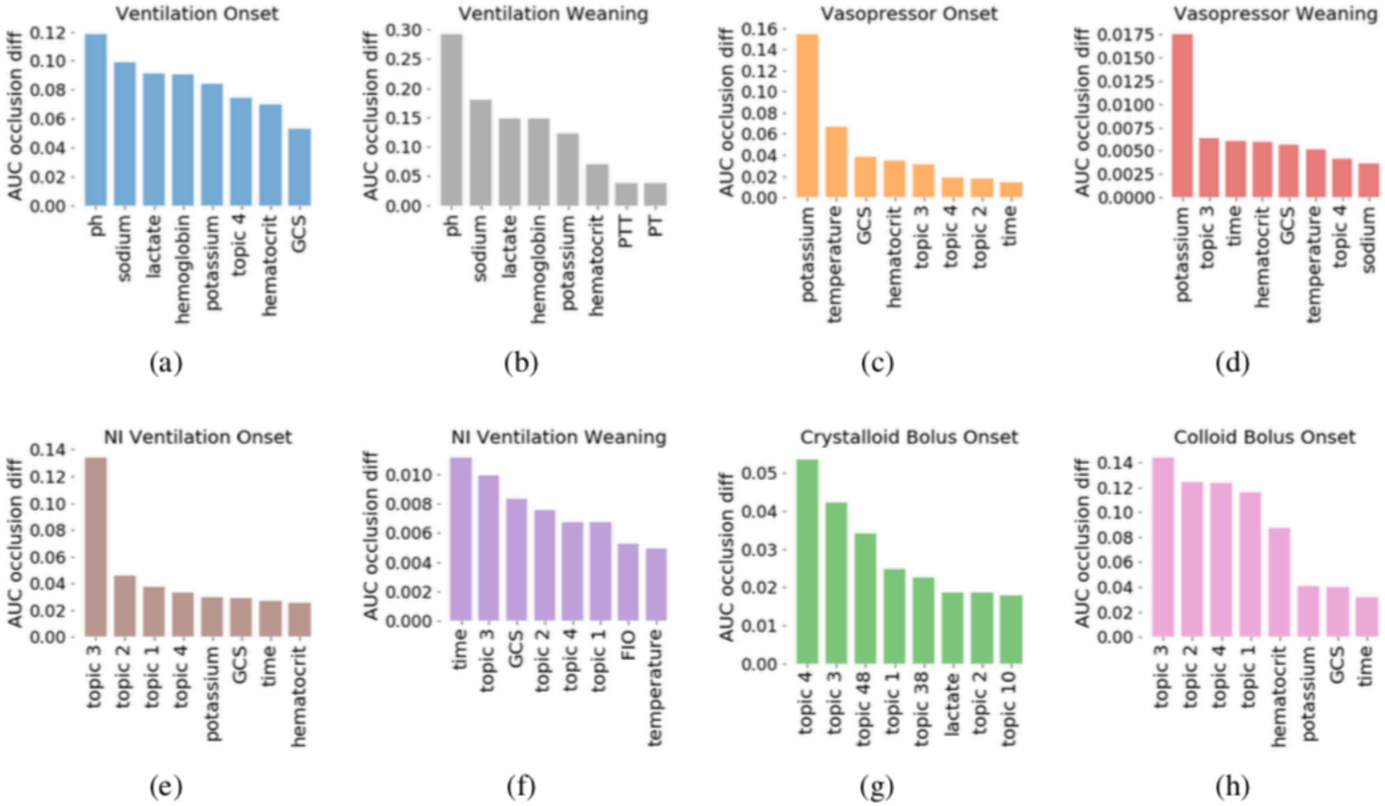
(b) The CNN architecture performs temporal convolutions at 3 different granularities (3, 4, and 5 hours), max-pools and combines the outputs, and runs this through 2 fully connected layers to arrive at the prediction.

Figure 3: Schematics of LSTM and CNN model architectures.



Task	Model	Intervention Type VENT	NI-VENT	VASO	COL BOL	CRYS BOL
Onset AUC	Baseline	0.60	0.66	0.43	0.65	0.67
	LSTM Raw	0.61	0.75	**0.77**	0.52	0.70
	LSTM Words	**0.75**	**0.76**	0.76	**0.72**	**0.71**
	CNN	0.62	0.73	**0.77**	0.70	0.69
Wean AUC	Baseline	0.83	0.71	0.74	-	-
	LSTM Raw	0.90	0.80	**0.91**	-	-
	LSTM Words	0.90	**0.81**	**0.91**	-	-
	CNN	**0.91**	0.80	**0.91**	-	-
Stay On AUC	Baseline	0.50	0.79	0.55	-	-
	LSTM Raw	0.96	**0.86**	**0.96**	-	-
	LSTM Words	**0.97**	**0.86**	0.95	-	-
	CNN	0.96	**0.86**	**0.96**	-	-
Stay Off AUC	Baseline	0.94	0.71	0.93	-	-
	LSTM Raw	0.95	**0.86**	**0.96**	-	-
	LSTM Words	**0.97**	**0.86**	0.95	-	-
	CNN	0.95	**0.86**	**0.96**	-	-
Macro AUC	Baseline	0.72	0.72	0.66	-	-
	LSTM Raw	0.86	**0.82**	**0.90**	-	-
	LSTM Words	**0.90**	**0.82**	0.89	-	-
	CNN	0.86	0.81	**0.90**	-	-

Table 2: Comparison of model performance on five targeted interventions. Models that perform best for a given (intervention, task) pair are bolded.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4: We are able to make interpretable predictions using the LSTM and occluding specific features. The top eight features that cause a decrease in prediction AUC for each intervention task. In general, physiological data were more important for the more invasive interventions — mechanical ventilation (4a, 4b) and vasopressors (4c, 4d) — while clinical note topics were more important for less invasive tasks — non-invasive ventilation (4e, 4f) and fluid boluses (4g, 4h). Note that all weaning tasks except for ventilation have significantly less AUC variance.

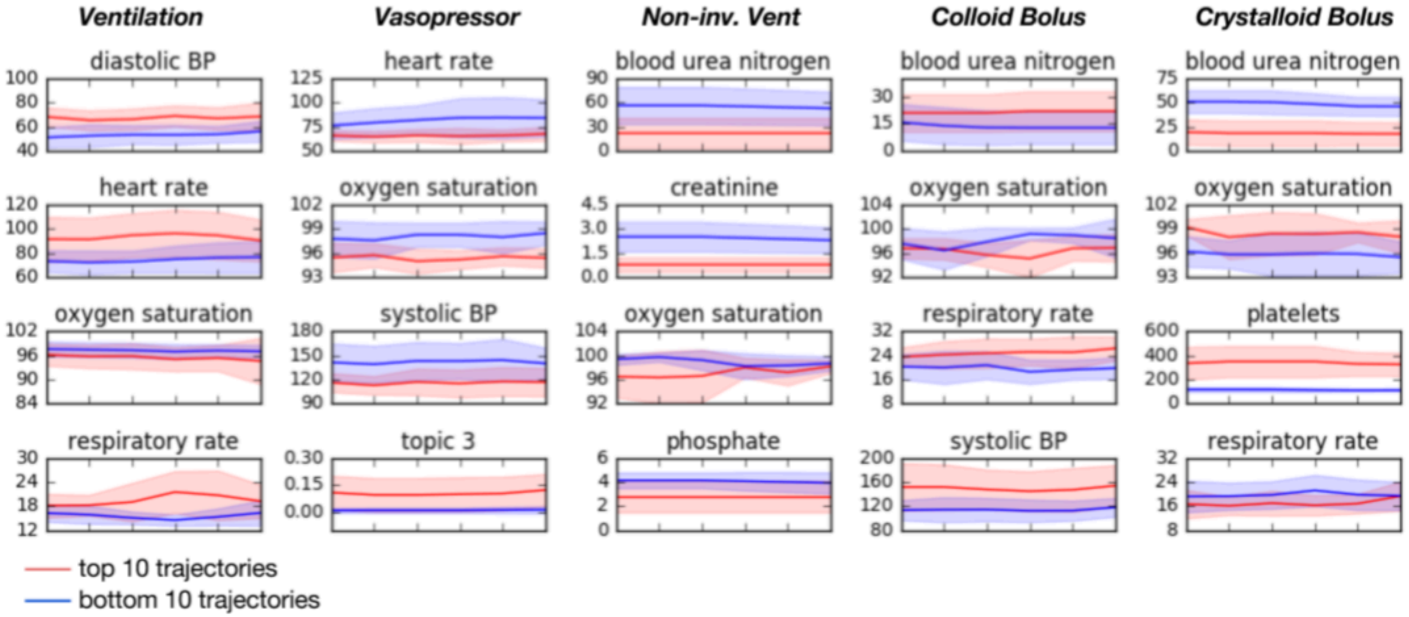


Figure 5: Trajectories of the 10 maximally and minimally activating examples for onset of each of the interventions.

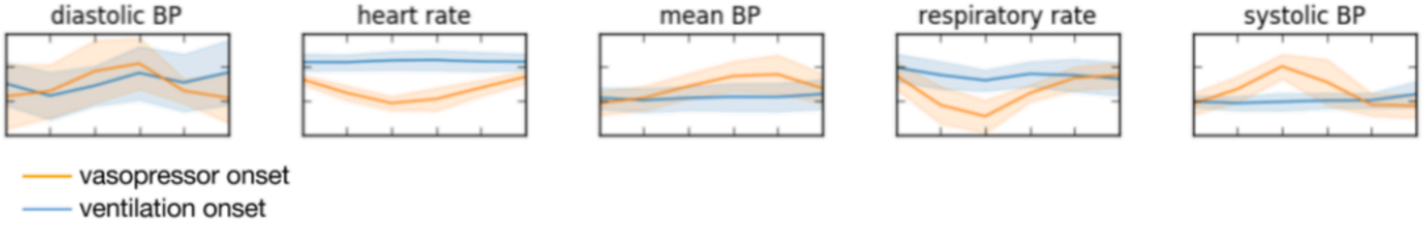


Figure 6: Trajectories generated by adjusting inputs to maximally activate a specific output node of the CNN.

ANN Model for De-identification of Clinical Narratives

- Label-sequence optimization layer

$$s(y_{1:n}) = \sum_{i=1}^{n} \mathrm{a}_i[y_i] + \sum_{i=2}^{n} T[y_{i-1}, y_i]$$

- Label prediction layer

- Character-enhanced token-embedding layer

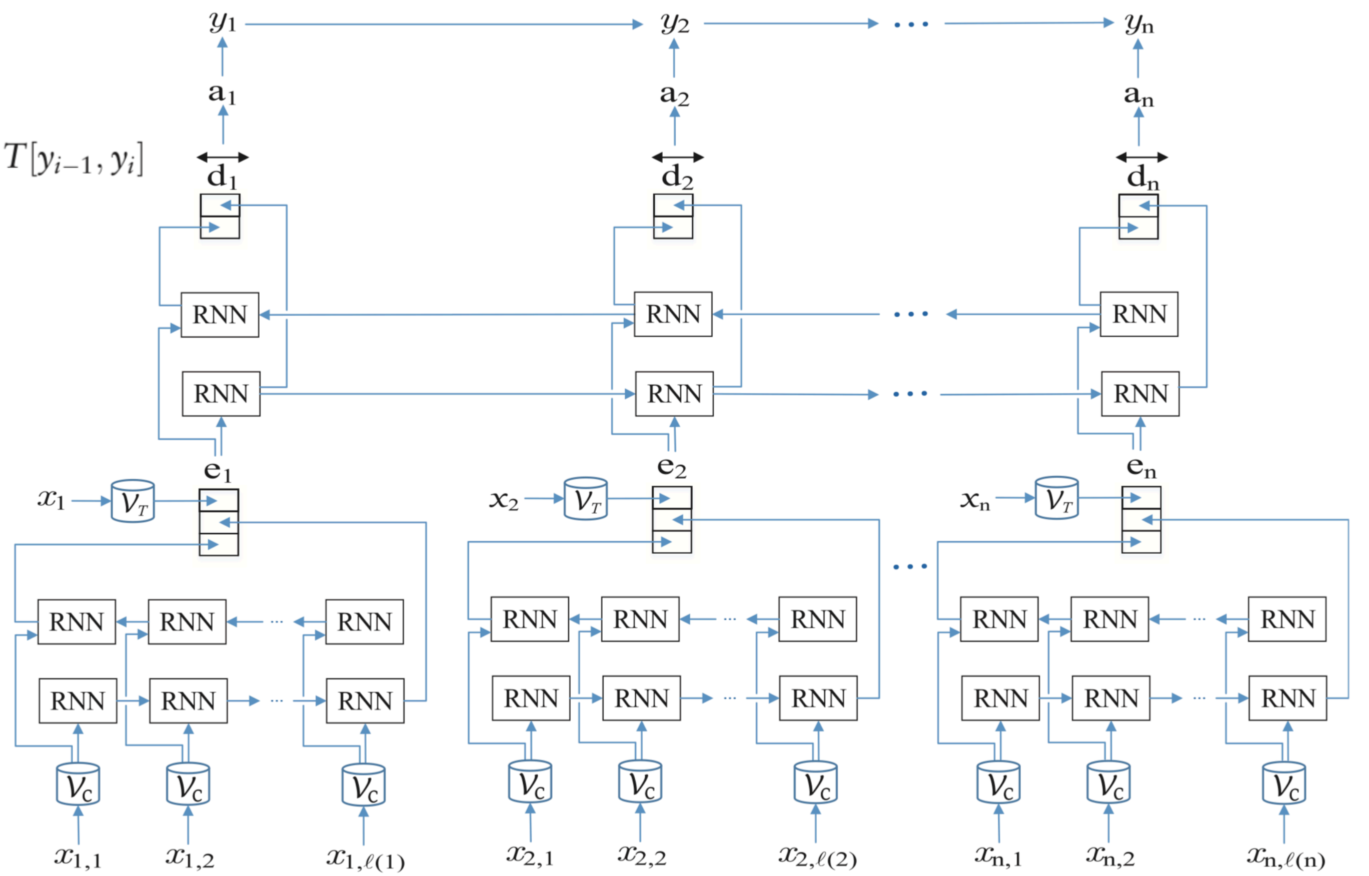
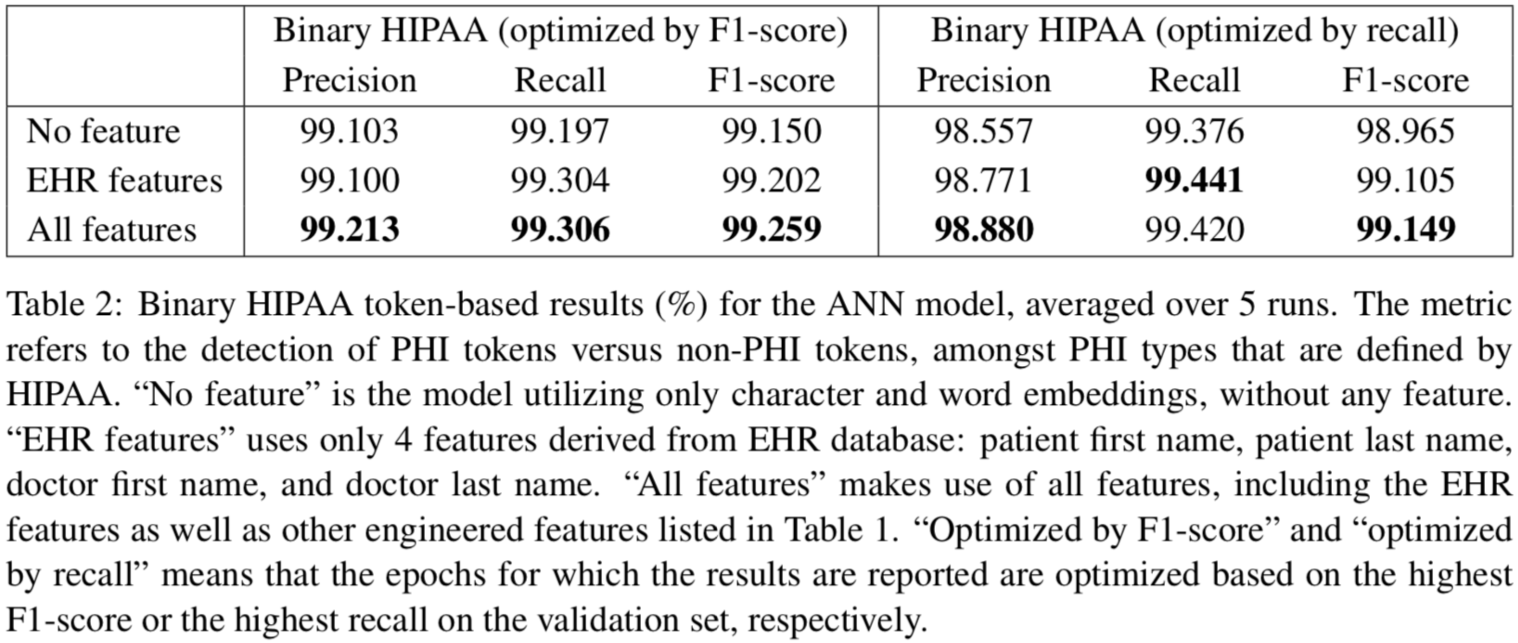


Figure 1. Architecture of the artificial neural network (ANN) model. (RNN, recurrent neural network.) The type of RNN used in this model is long short-term memory (LSTM). n is the number of tokens, and x_i is the i^{th} token. $\mathcal{V}_T$ is the mapping from tokens to token embeddings. $\ell(i)$ is the number of characters and $x_{i,j}$ is the j^{th} character in the i^{th} token. $\mathcal{V}_C$ is the mapping from characters to character embeddings. e_i is the character-enhanced token embeddings of the i^{th} token. $\overleftrightarrow{\mathrm{d}}_i$ is the output of the LSTM of the label prediction layer, a_i is the probability vector over labels, y_i is the predicted label of the i^{th} token.

Dernoncourt, F., Lee, J. Y., Uzuner, Ö., & Szolovits, P. (2016). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, ocw156. http://doi.org/10.1093/jamia/ocw156



	Binary HIPAA (optimized by F1-score)			Binary HIPAA (optimized by recall)		
	Precision	Recall	F1-score	Precision	Recall	F1-score
No feature	99.103	99.197	99.150	98.557	99.376	98.965
EHR features	99.100	99.304	99.202	98.771	**99.441**	99.105
All features	**99.213**	**99.306**	**99.259**	**98.880**	99.420	**99.149**

Table 2: Binary HIPAA token-based results (%) for the ANN model, averaged over 5 runs. The metric refers to the detection of PHI tokens versus non-PHI tokens, amongst PHI types that are defined by HIPAA. "No feature" is the model utilizing only character and word embeddings, without any feature. "EHR features" uses only 4 features derived from EHR database: patient first name, patient last name, doctor first name, and doctor last name. "All features" makes use of all features, including the EHR features as well as other engineered features listed in Table 1. "Optimized by F1-score" and "optimized by recall" means that the epochs for which the results are reported are optimized based on the highest F1-score or the highest recall on the validation set, respectively.

Opportunities

- What do you want to predict? How can that improve health care?
- Create very large scale research data sets
  - Toward personalized but evidence-based care
  - E.g., *All of Us* (https://allofus.nih.gov), PCORI, UK Biobank, …
- Engage vast army of interested students, researchers in machine learning
- Crowdsource improvements to data curation
- Improve on many fronts
  - scientific understanding of disease
  - instrumentation and observation
  - process of health care (“industrial engineering”)

Where will medical progress arise?

- Understanding mechanisms of disease
- Instrumentation
- Clinical knowledge and Data analysis ★
  - Predictive Modeling
  - Natural Language Processing

Current Projects

- Predictive Modeling
  - Progression of inflammatory bowel disease and similar auto-immune diseases
  - Complications of pregnancy and delivery
  - Most appropriate treatments at different stages of disease/care
  - Undiagnosed diseases
- Image Analysis (typically coupled with clinical data)
  - Generating (draft) radiology report from chest x-rays
  - Improving radiation therapy planning
  - Need for surgery in aneurysm patients
  - Time course of development of liver fibrosis, cirrhosis
- Natural Language Processing
  - Question answering (for providers, patients) from clinical records
  - Understanding information-seeking behavior of breast cancer patients
  - Translation of professional to lay language
  - Entity and relation (including temporal) extraction
  - Using NLP in predictions of clinical course, readmission, …
- Laboratory data
  - Imputation of unmeasured analytes
  - Effect of perturbagens on gene expression in psychiatric disorders