Transforming Healthcare Using Machine Learning

John Guttag
Dugald C. Jackson Professor
Professor MIT EECS

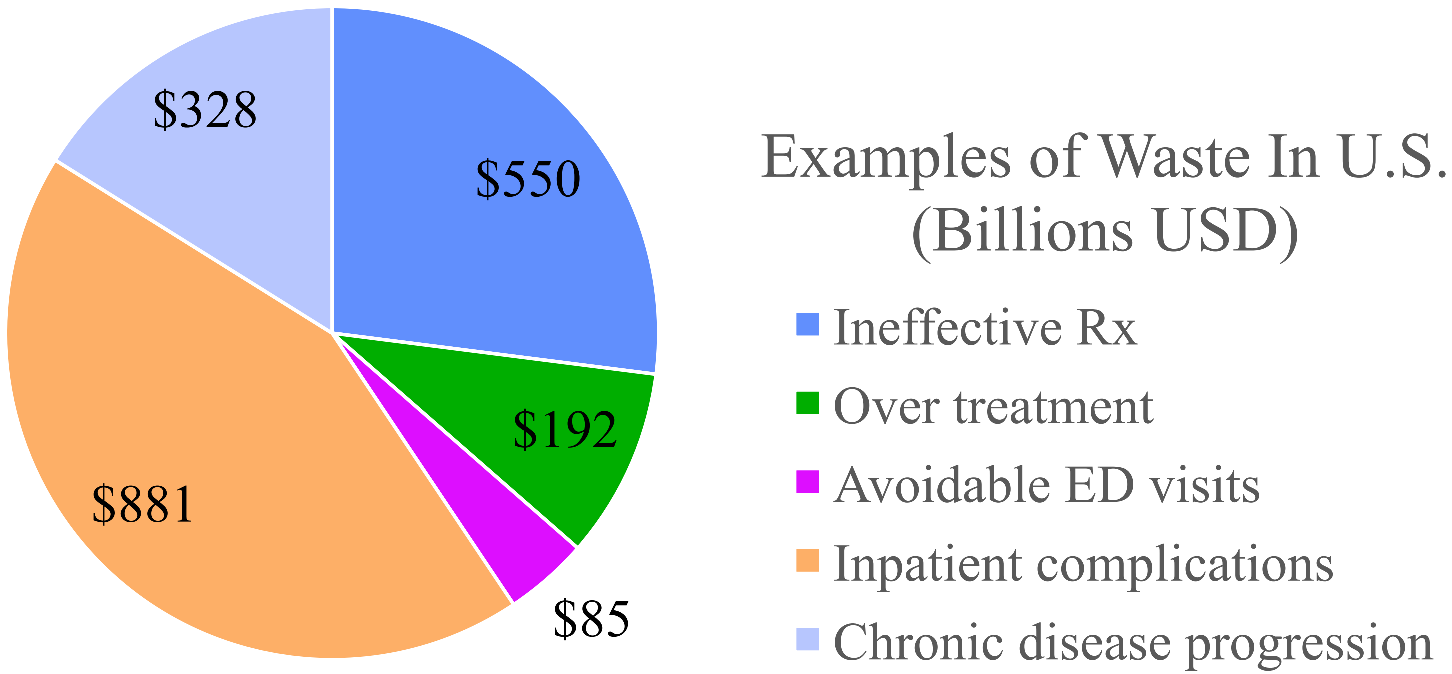
CSAIL

Conflict of Interest Disclosure

I am Chief Scientific Officer at Health[at]Scale Technologies, Inc. and have a financial interest in the company.

HEALTH [at] SCALE
TECHNOLOGIES

Why Healthcare Needs to be Transformed



Opportunity to improve care provided to people.
And save billions!

Machine Learning Can Help

By matching

Each patient

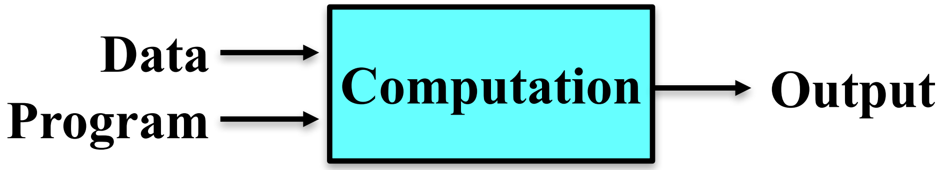
to the
Right treatment

by the
Right provider

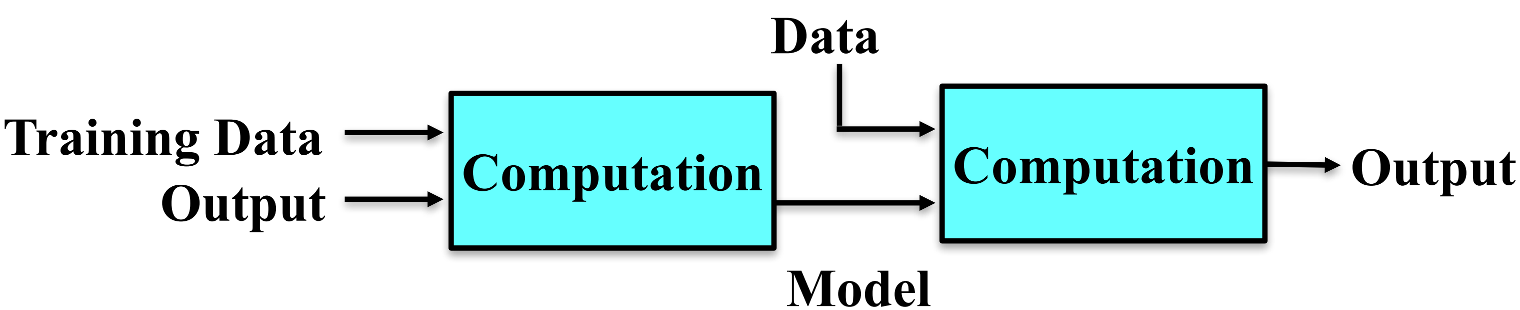
at the
Right time

What Is Machine Learning (ML)

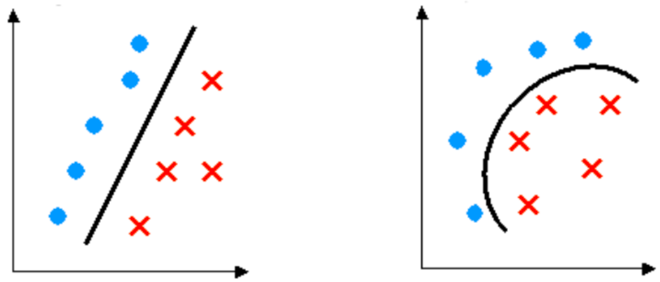
Traditional Programming



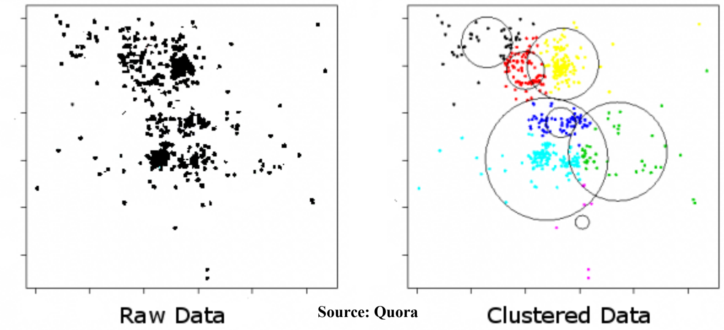
(Supervised) Machine Learning



Some Different Kinds of Machine Learning



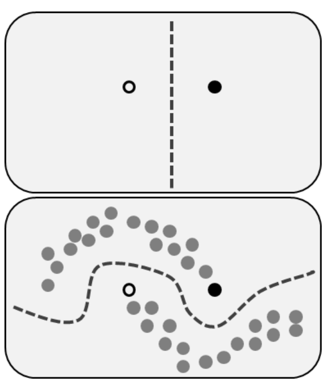
Source: Utah CS



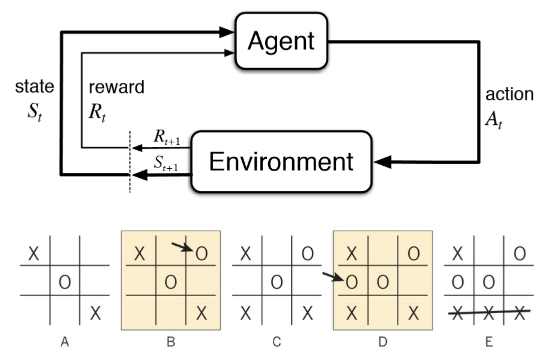
Source: Quora

Supervised Learning

Unsupervised Learning



Source: Wikipedia

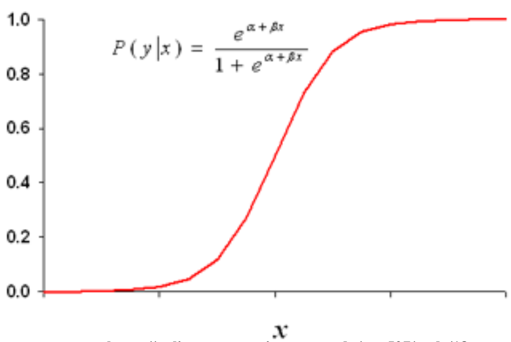


Source: Stanford CS and Nature

Semi-Supervised Learning

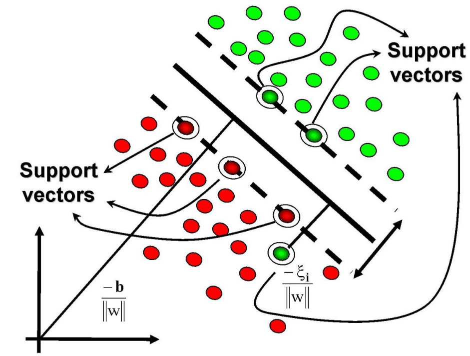
Reinforcement Learning

Some Classes of Machine Learning Algorithms



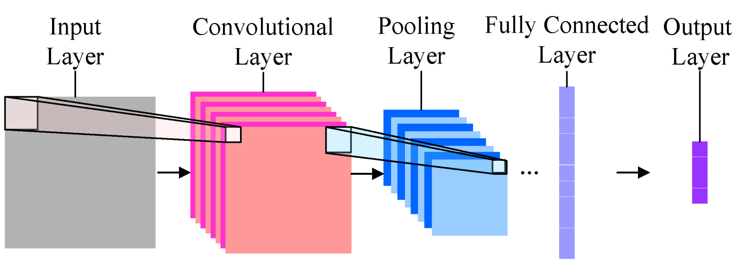
https://onlinecourses.science.psu.edu/stat507/node/18

Logistic Regression



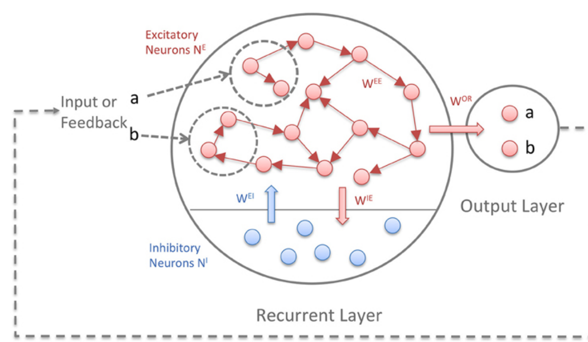
Source: https://medium.com/@haydar_ai/learning-data-science-day-11-support-vector-machine-8ef06da91bfc

Support Vector Machines



Source: http://www.mdpi.com

Convolutional Neural Networks



Source: www.frontiersin.org/articles/10.3389/fncom.2015.00036/full

Recurrent Neural Networks

CSAIL

How Might ML Be Useful in Healthcare

Better decisions about care

E.g., should patient *A* receive an ICD? When and how can we intervene to avoid an ED visit for patient *B*?

Improvements at the level of a patient

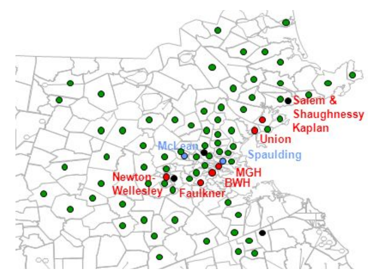
Benchmarking and improving institutions and providers

E.g., is hospital *C* over- or under-performing relative to peer institutions?

Improvements at the level of a hospital

Optimizing resource utilization

E.g., should patient *D* have a procedure done at hospital *E* or F?



Improvements at the level of a network

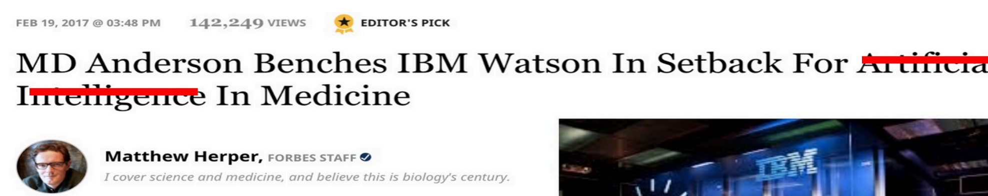
Impact of ML on Healthcare

Lots of slick marketing from industry

Lots of publications from academia

But, over all, disappointingly little change in

- Delivery of care
- Business models

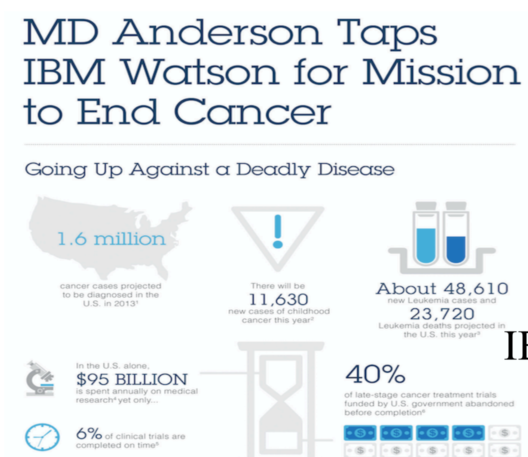

FEB 19, 2017 @ 03:48 PM 142,249 VIEWS EDITOR'S PICK

MD Anderson Benches IBM Watson In Setback For ~~Artificial Intelligence~~ In Medicine

Matthew Herper, FORBES STAFF

I cover science and medicine, and believe this is biology's century.

Forbes 2017



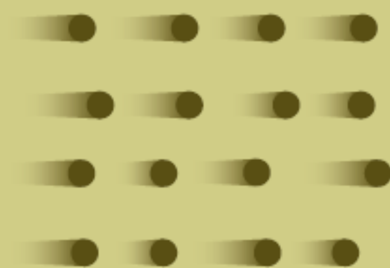
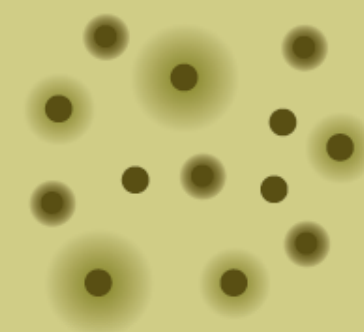
IBM 2013

Why?

Because it's hard

Because ML for healthcare is different

The Typical "Big Data" Problem

Too Large	Too Fast	Too Complex	Too Uncertain
Volume	**Velocity**	**Variety**	**Veracity**
			
Data at rest	**Data in motion**	**Data in many forms**	**Data in doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text and multimedia	Uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception and model approximations

Source: IBM

Not That Large

500 Petabytes

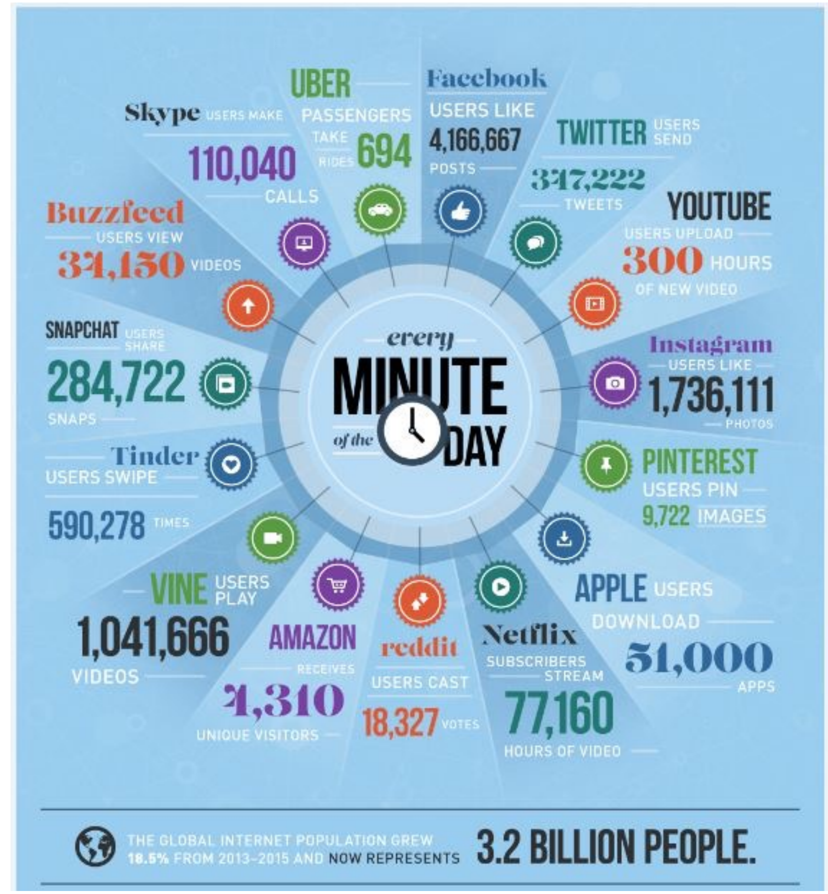
(1 petabyte = 1 million gigabytes)

Total amount of all the healthcare data existing in the world in 2012[1]

15 Exabytes

(1 petabyte = 1 billion gigabytes)

Total amount of data managed by a large web company alone in 2014[2]



Source: BGR

[1] IDC Report; [2] Cirrus Insight

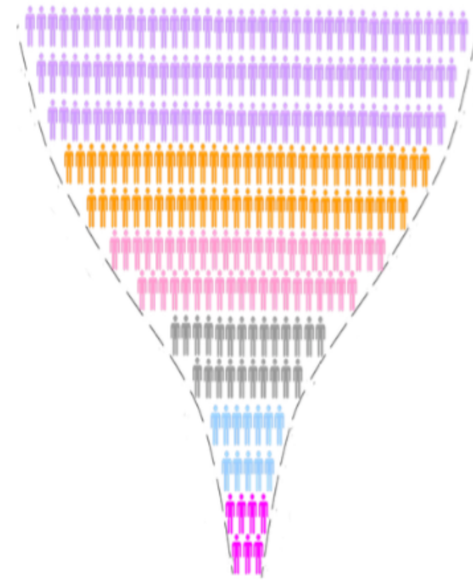
Never Enough Obviously Relevant Data

Filter by patient history

Example
Filter 1: How many patients have a craniotomy?
Filter 2: How many of these have intentional hypothermia?
Filter 3: How many of these have a post operative infection?
Filter 4: How many of these received an antibiotic?
...

Trying to find groups of patients with similar demographic, H&P findings, labs etc. can make big data small really quickly

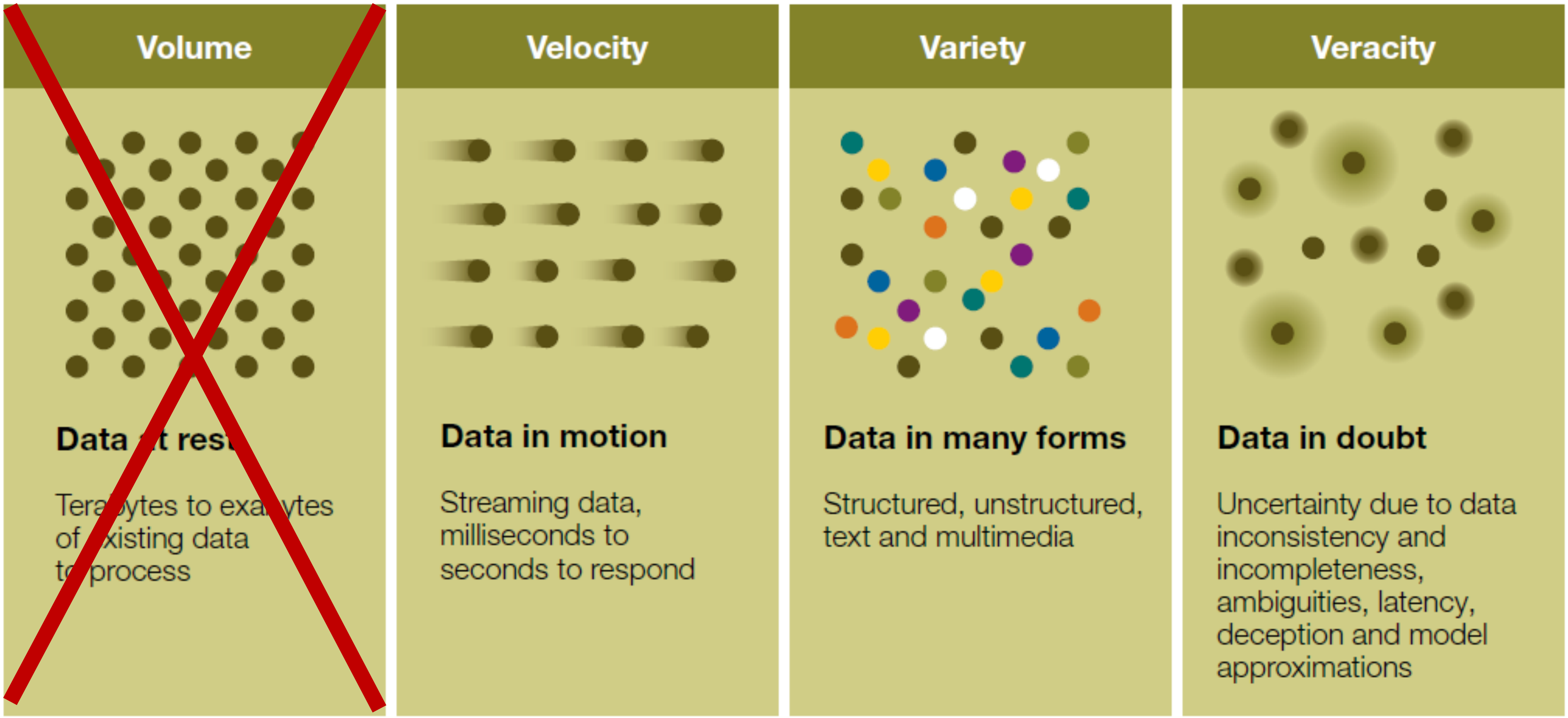
Source: Health at Scale Corporation



Similar story for institutions

Not the Typical "Big Data" Problem

Too Large | Too Fast | Too Complex | Too Uncertain



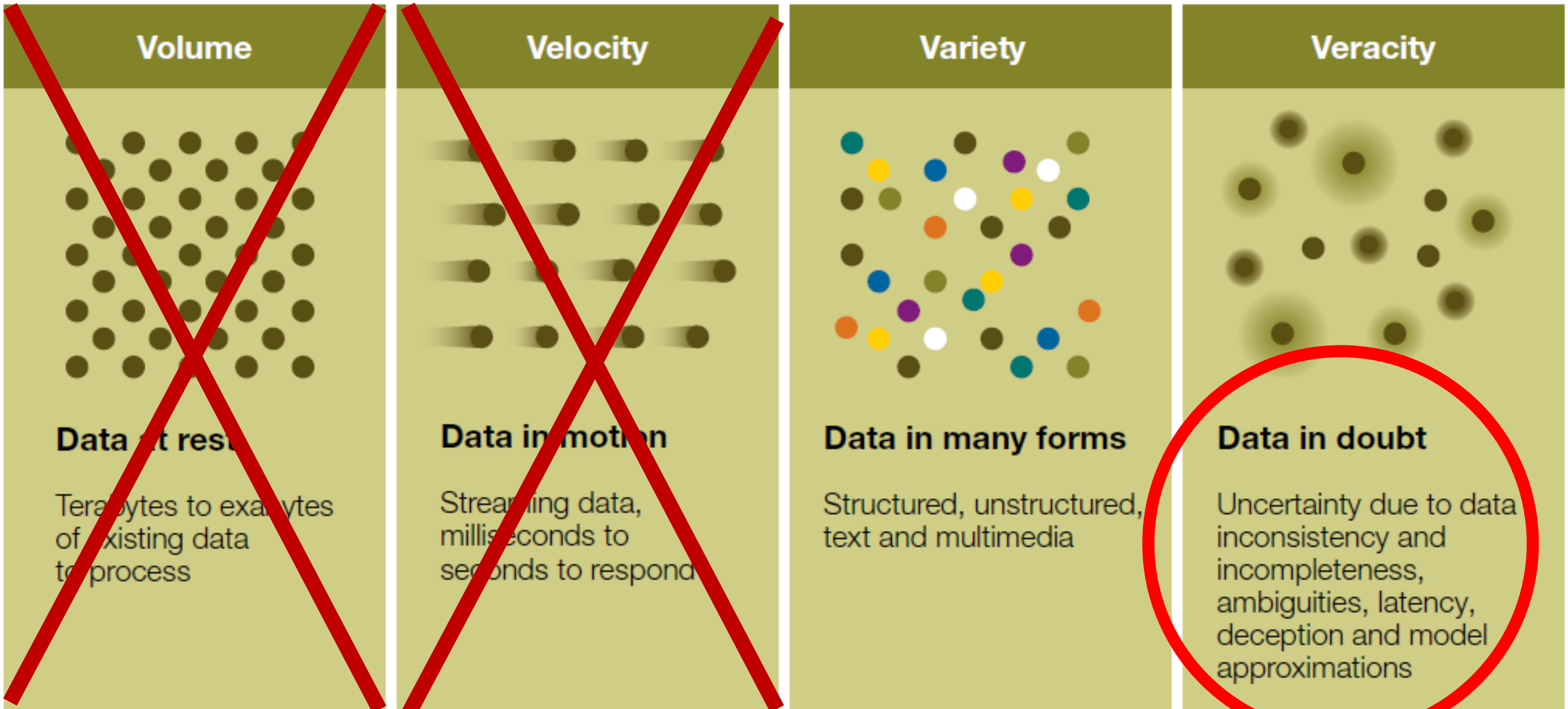
Source: IBM

Not That Fast

- Getting data about current patient can be urgent
- But getting data about large numbers of patients is rarely urgent
- For almost all medical decisions “real time” is minutes, not micro seconds

Not the Typical "Big Data" Problem

Too Large | Too Fast | Too Complex | Too Uncertain



Source: IBM

But Not Impossible

Unprecedented amount of relevant data

- Medical records
- Clinical trial data
- Billing data
- Ambulatory data

Economic pressure for reduced cost and better outcomes

- Payers
- Consumers

Improved technology

- Hardware
- ML methods
  - General purpose
  - Specialized to healthcare

CSAIL

Some Examples

Better outcomes and reduced cost for post-acute care

Billing data

CMS
CENTERS FOR MEDICARE & MEDICAID SERVICES

Reducing prevalence of nosocomial infections (HAIs)

Full electronic health records

Precision Steerage

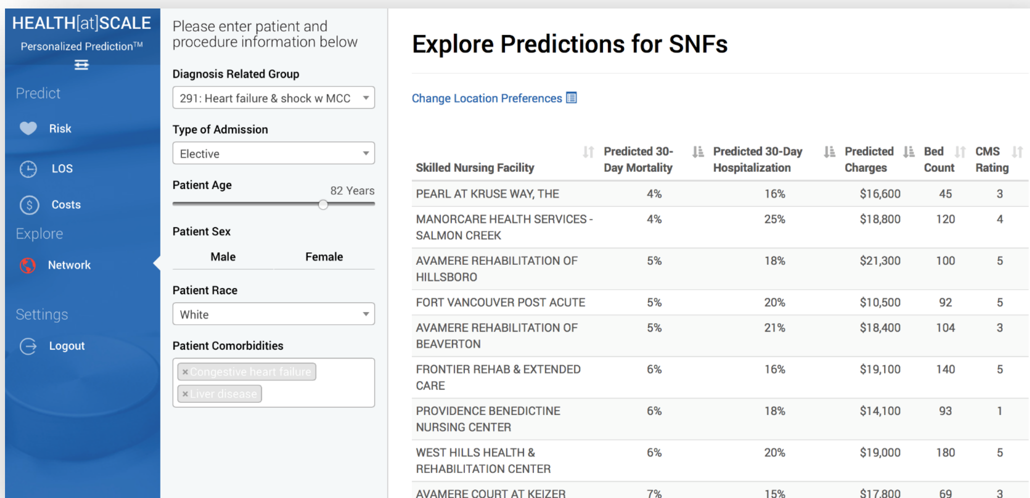
Skilled nursing facilities (SNFs) increasingly important

25% of patients readmitted to hospital within 30 days

2/3 of these preventable

Choosing skilled nursing facilities (SNFs) that are optimal for *individuals*

Built a predictive model using years of Medicare data



Source: Health[at]Scale

Precision Steerage

Predicted 30-Day Mortality	Predicted 30-Day Hospitalization	Predicted Charges	Bed Count	CMS Rating
4%	16%	$16,600	45	3
4%	25%	$18,800	120	4
5%	18%	$21,300	100	5
5%	20%	$10,500	92	5
5%	21%	$18,400	104	3
6%	16%	$19,100	140	5
6%	18%	$14,100	93	1
6%	20%	$19,000	180	5
7%	15%	$17,800	69	3

Impact of Choosing the Right SNF

		Rate of 30-Day Mortality and Hospital Readmission (N=3)	Rate of 30-Day Mortality and Hospital Readmission (N=5)
National	Patients Admitted to Top-N SNFs (Number of Patients in Group)	22.4% (604,428)	22.8% (713,549)
	Patients Admitted to Non-Top-N SNFs (Number of Patients in Group)	28.6% (393,034)	29.9% (283,913)
	Comparison P-Value	<0.001	<0.001

Improvements in mortality and readmission at the population level after factoring in competition for resources, patient preferences and capacities of SNFs

National		Volume Change (5%)	Volume Change (10%)	Volume Change (25%)
	SNFs within 5 Miles	-6%	-7%	-9%
	SNFs within 10 Miles	-8%	-10%	-12%
	SNFs within 25 Miles	-11%	-13%	-15%

Results on Medicare Fee-For-Service Population

Healthcare Associated Infections in the U.S.

Source: General Electric

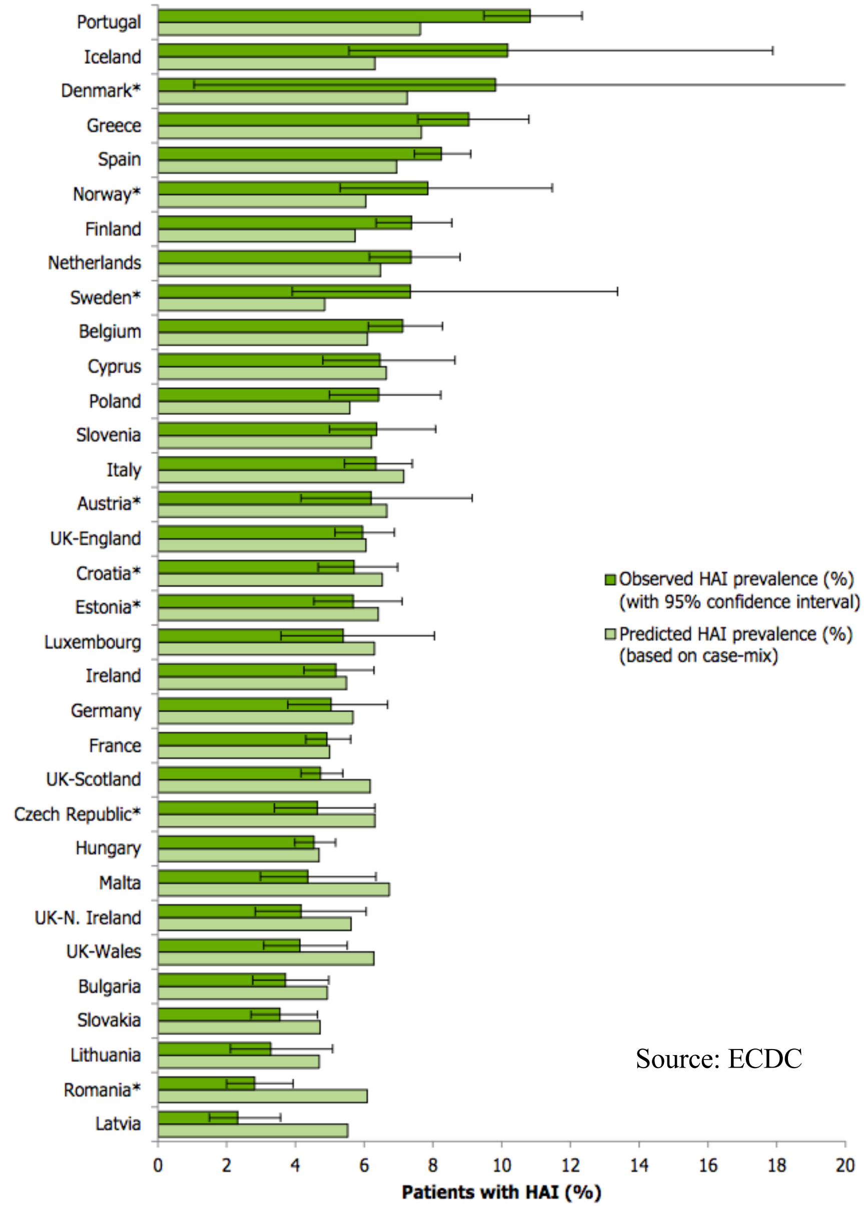
Not Just in U.S.

U.S.: 4.8%

EU: 7.1%

Low and middle income countries: 15.5%

Source: WHO

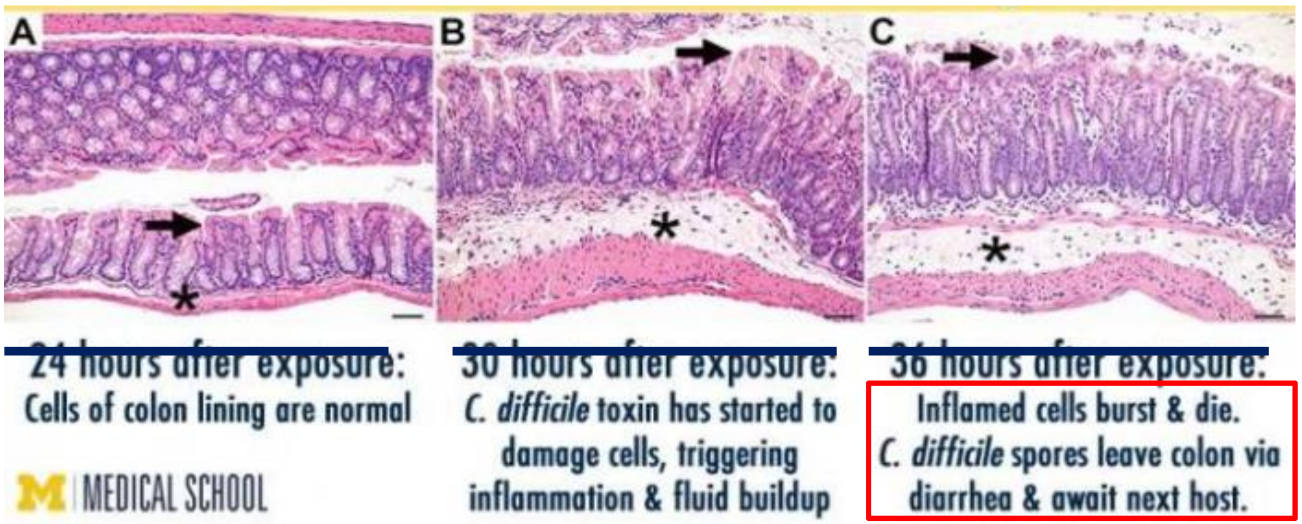


Source: ECDC

- **C. diff was established as the major cause of antibiotic-associated diarrhea in 1978**
  - Early association with clindamycin→ but since then many other antibiotics have been implicated including cephalosporins, fluoroquinolones
  - Early 2000s, NAP1/B1/027 strain emerged– more virulent, increased toxin production
- **500,000 infections each year in U.S.; 30,000 deaths**
  - 66% healthcare associated
- **About 25% will recur**

C. diff Infection (CDI)

C. diff an anaerobic, spore-forming, gram-positive bacillus



Risk Factors

Some are well established

- Antibiotic exposure
- Healthcare exposure
- Prior CDI
- Proton pump inhibitors
- Advanced age

Our focus

- Discovery of other factors that increase susceptibility
- Discovery of sources of infections
  - Colonization
  - Environmental exposure
  - Transmission paths

Transmitted by Environment

- Highly transmissible by fecal-oral route
- Patients can serve as a reservoir for environmental contamination

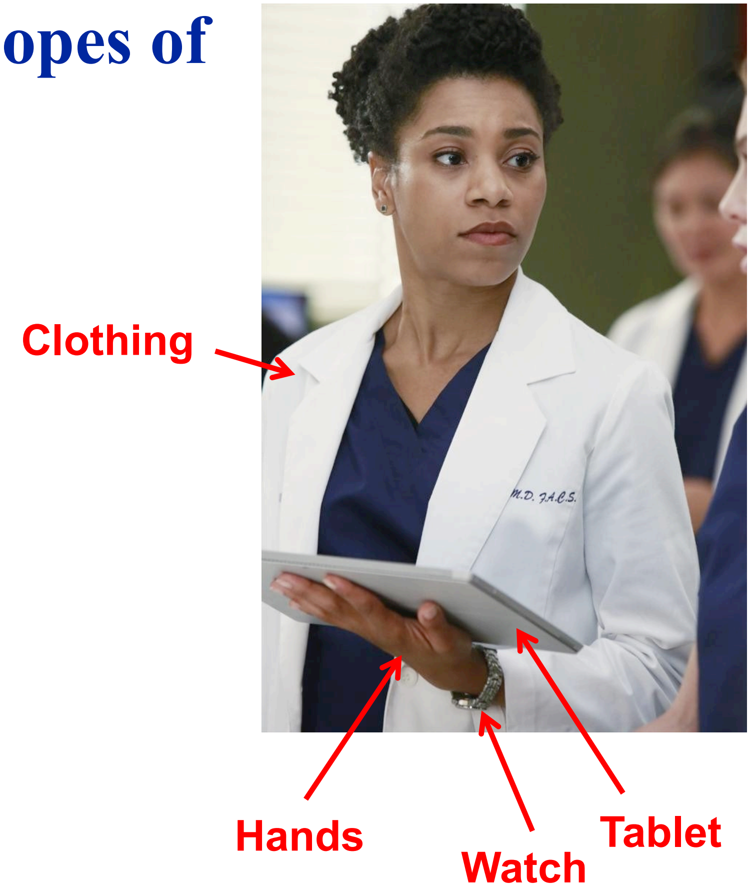
The room looks clean, but …

Source: Wikipedia

Transmitted by People

Has been cultured from hospital rooms, items in the room, and the hands, clothing, stethoscopes of healthcare workers

She looks clean, but …



Role of Asymptomatically Colonized

An open question

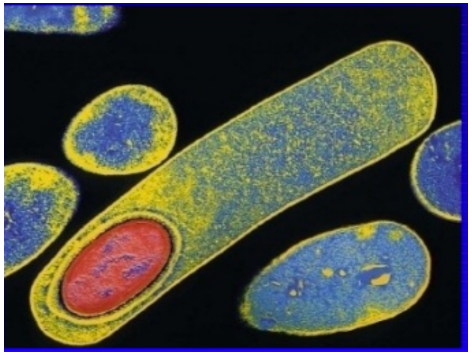
Estimates in literature all over the place

"Extremely rare" to "50% of all cases"

Colonized patients do shed spores

Spores last a long time in environment

Recent studies suggest role is important



One Way to Think About Things

Hospital system of (mostly) mobile "devices"

- Medical equipment and furniture
- Patients
- Caregivers

Properties of devices can be changed by coming into contact with other devices

- Patient becomes infected, x-ray table acquires spores, …

Learn

- Properties of individual devices
- How individual and classes of devices influence each other

Specific Questions

Who is at highest risk for CDI *and for being colonized*?

What is the contribution of *asymptomatic carriage* to CDI?

What are the most important *routes of transmission* over space *and* time?

Risk Prediction for CDI

Traditionally, takes the form of evaluating existing hypotheses, i.e., regression model incorporating antibiotics, PPIs, comorbidities, etc.

- This approach pre-specifies the variables that matter
- Heavy emphasis on susceptibility, **exposure largely ignored**
- Generally these models predict risk at time of admission to the hospital, however, ***we know that risk evolves over time***

Our Approach

- Leverages all available information to identify factors that confer risk
- Fine-grained inference of exposure
- Allows for evolution of risk (and relative importance of risk factors) over time

It’s Not About A Model

One size does not fit all!

It's Not About A Model

Models need to be institution specific

What is most important at MGH may or may not be most important elsewhere

Developed a method for building institution-specific models

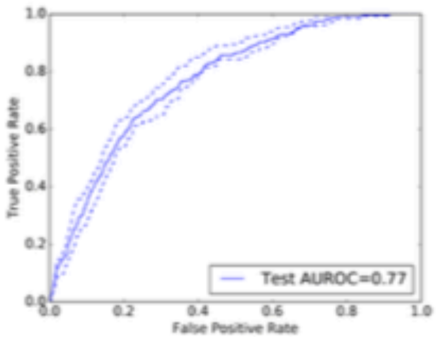
a generalizable approach **rather than** **a generalizable model**

Tested it by building separate models for two institutions, MGH and Univ. of Michigan Hospital

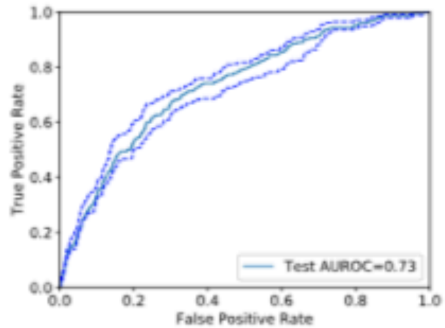
Results

AUROC 0.733-0.77

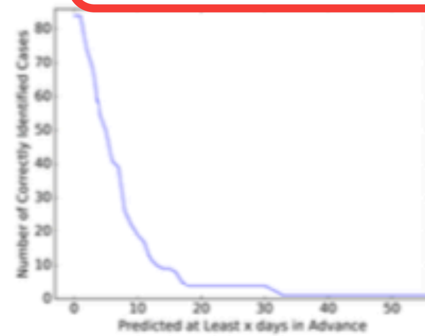
Median number of days in advance of diagnosis CDI: 5



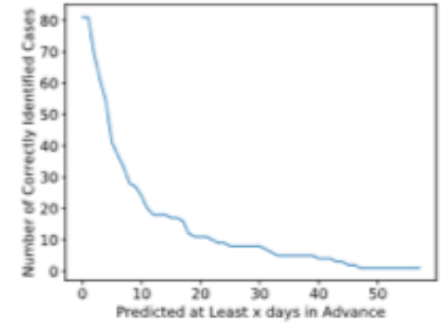
MM



MGH



MM



MGH

Risk Factor Rank	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
1	Admission to Medicine (0.39)	Chlorhexidine (0.36)	Chlorhexidine (0.32)	Chlorhexidine (0.30)	Allopurinol (0.26)	Allopurinol (0.26)
2	Any history of CDI (0.34)	Vancomycin (0.29)	Allopurinol (0.28)	Cefepime (0.27)	Furosemide (0.27)	Chlorhexidine (0.21)
3	History of CDI within prior year (0.34)	Admission to medicine service (0.29)	Vancomycin (0.28)	Allopurinol (0.26)	Omeprazole (0.25)	Omeprazole (0.20)
4	Chlorhexidine (0.32)	Cefepime (0.26)	Admission to Medicine (0.27)	Vancomycin (0.24)	Chlorhexidine (0.24)	Furosemide (0.20)
5	Allopurinol (0.30)	Omeprazole (0.26)	Cefepime (0.26)	Omeprazole (0.24)	Vancomycin (0.22)	History of CDI within prior year (0.19)

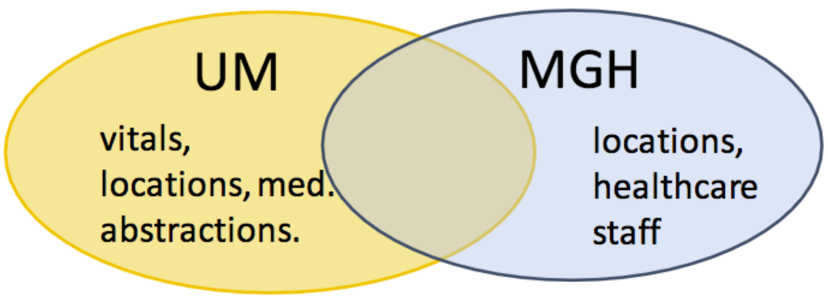
Some Institutional Differences

Demographic	UM	MGH
Female	54%	49%
Median age	56	62
CDI	1.1%	0.83%
CDI in past year	2.4%	1.55

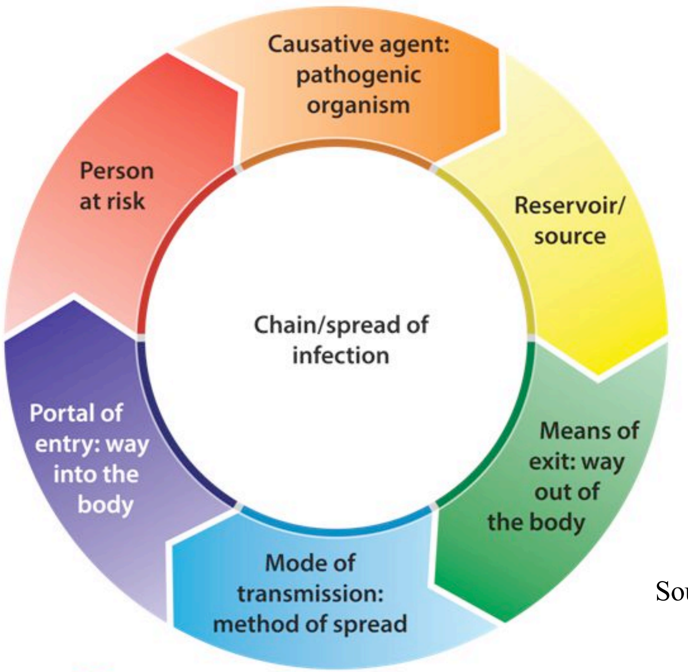
Features

UM d=4,836

MGH d=1,837



Current model is mostly about susceptibility→ what about exposure?

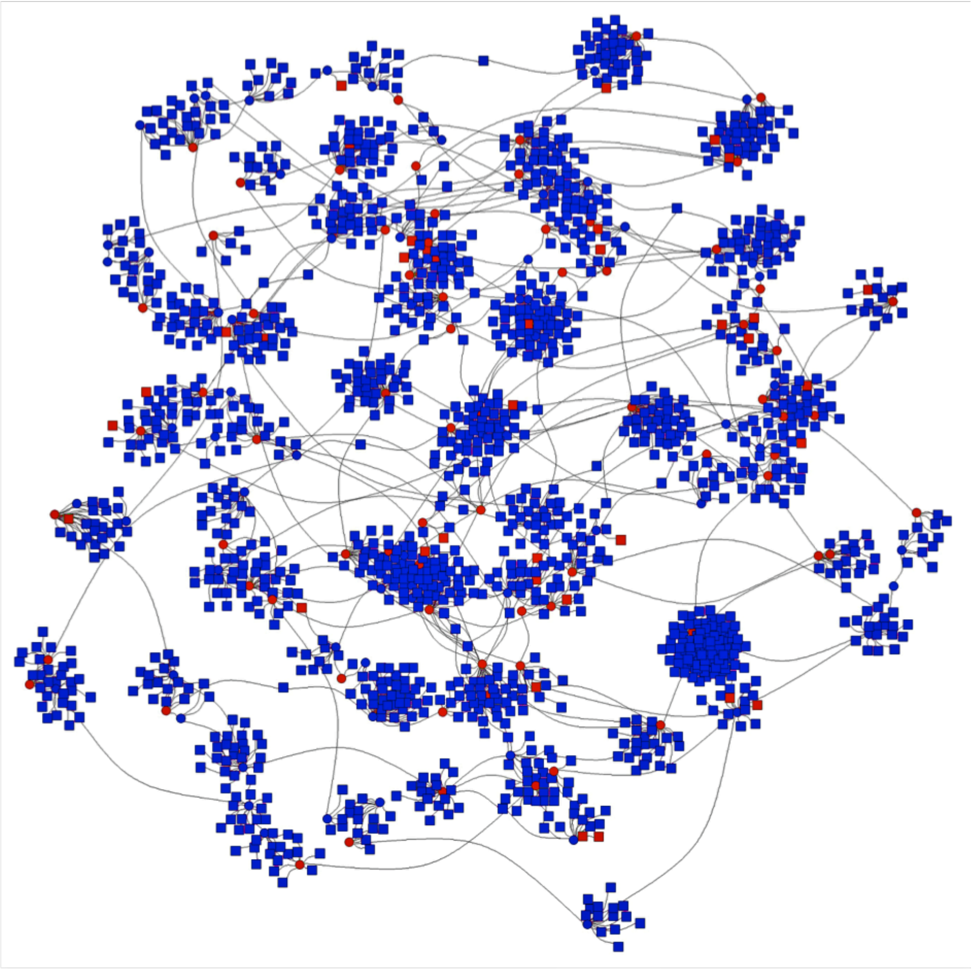


Source: Merryl Dawson

In order to understand exposure, we need to investigate the network and paths

MGH

MGH



Problem: Estimating Influence of Neighbors

Identify latent influencers based on

- Intrinsic characteristics of node
- Characteristics and labels of neighbors
  - Infected or not

Maximum likelihood estimation in presence of latent variables

Multiple definitions of "neighbor"

- Shared spaces over time (proxy for furniture)
- Shared care givers

Assumptions Underlying Work

Not infected does not imply not contagious

Colonized individuals shed spores

Two factors contributing to infection state:

Susceptibility: captured through observed individual characteristics

Exposure: captured through contact with **unobserved latent spreaders**

Network structure is observable

An Over-simplified View

Predict the spreader state (z) of individuals based upon their own characteristics

Predict who will become infected (y) based on their characteristics and the spreader states of their neighbors

$$p(\mathbf{z}_i, \theta_i, \eta_i | \mathcal{D}, \mathbf{u}, \mathbf{w}) = \frac{p(\mathbf{z}_i | \mathbf{u}, X_{n(i)}) p(\theta_i | \mathbf{z}_i) p(\eta_i | \theta_i) p(y_i | \mathbf{x}_i, \eta_i, \mathbf{w})}{\int_\theta \sum_{\mathbf{z}} \sum_\eta p(\mathbf{z}_i | \mathbf{u}, X_{n(i)}) p(\theta_i | \mathbf{z}_i) p(\eta_i | \theta_i) p(y_i | \mathbf{x}_i, \eta_i, \mathbf{w})}$$

Some Early Qualitative Results

Fine grained analysis of neighbor relation improves predictive power

- Shared rooms
  - Concurrent occupancy
  - Sequential occupancy
- Shared nurses

Strong *hypotheses* about *specific* sources of infection

- Not just prevalence in a ward, but which patients/care providers

Wrapping Up

Computer science **is poised to revolutionize healthcare**

Not new therapies

A better job of utilizing existing therapies

Requires using multiple technologies

Machine learning

Sensing

Signal processing

Computer vision

Etc.

Requires a transition path

Accounting for economic factors

Collaboration with practitioners

Collaboration with industry