Limitations of Data Mining in Healthcare.

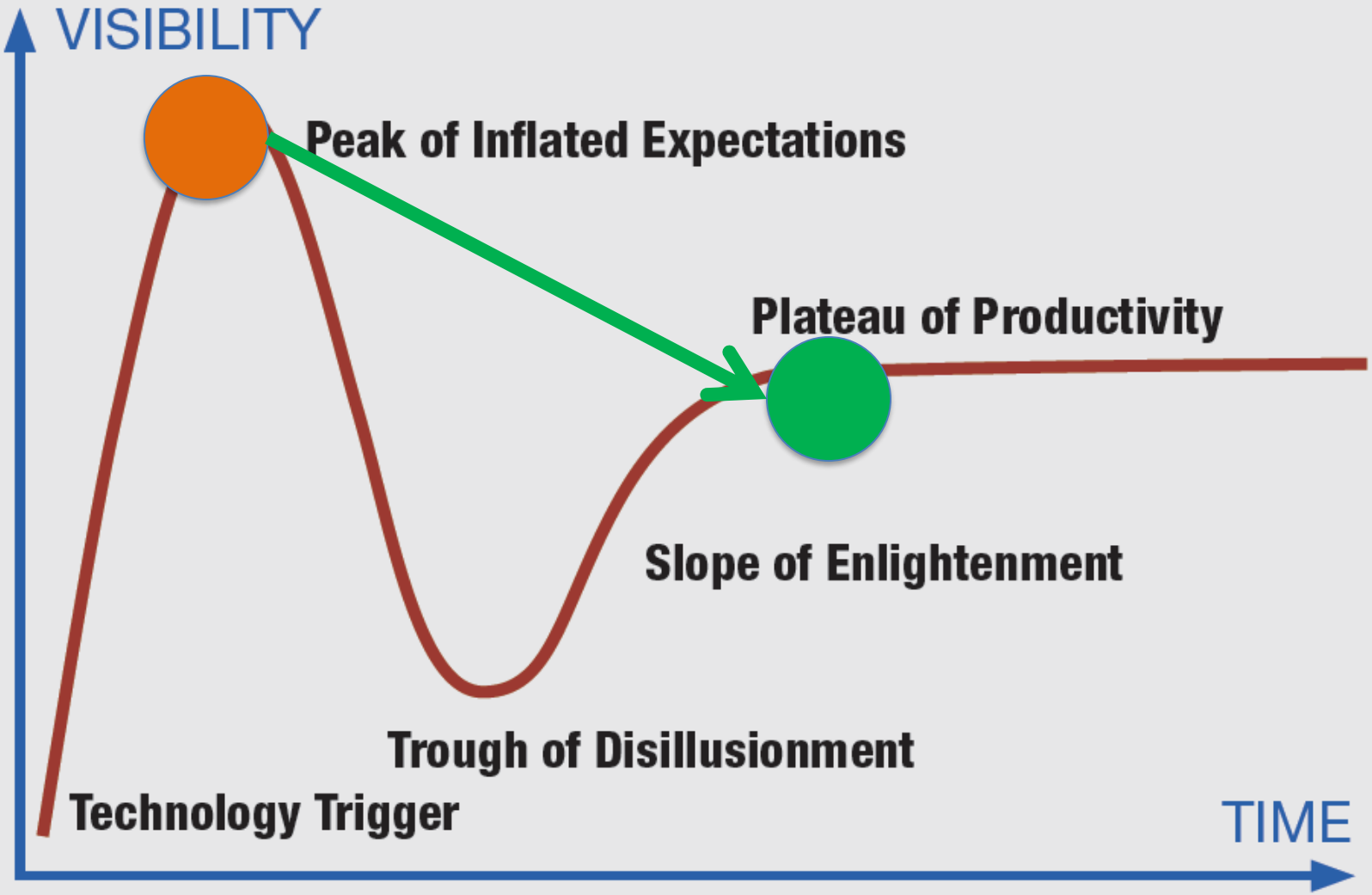
Vitaly Herasevich, MD, PhD, FCCM
Associate Professor of Anesthesiology and Medicine,
Department of Anesthesiology and Perioperative Medicine, Division of Critical Care

Multidisciplinary Epidemiology and Translational Research in Intensive Care (METRIC)

vitaly@mayo.edu

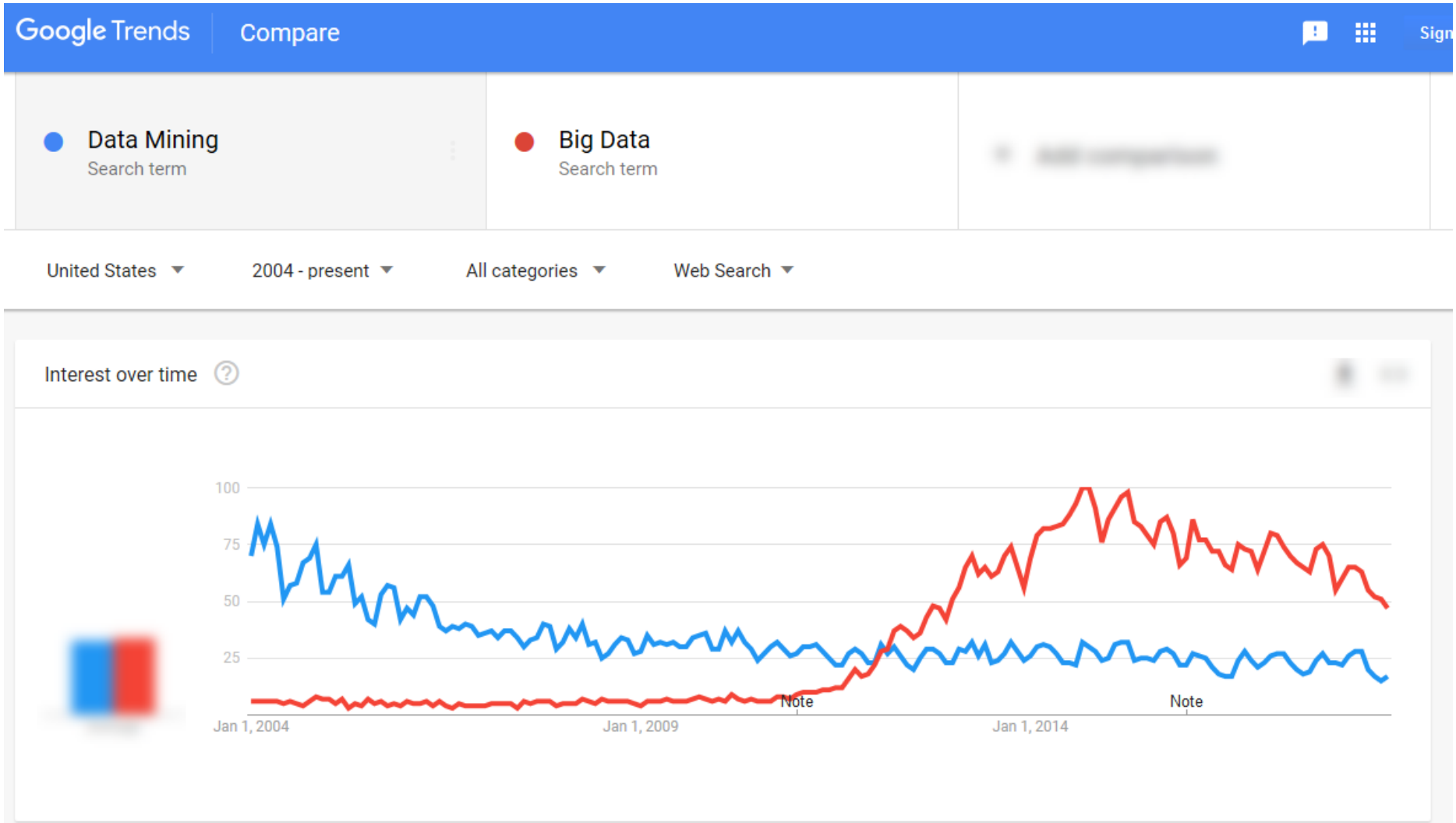
Outline

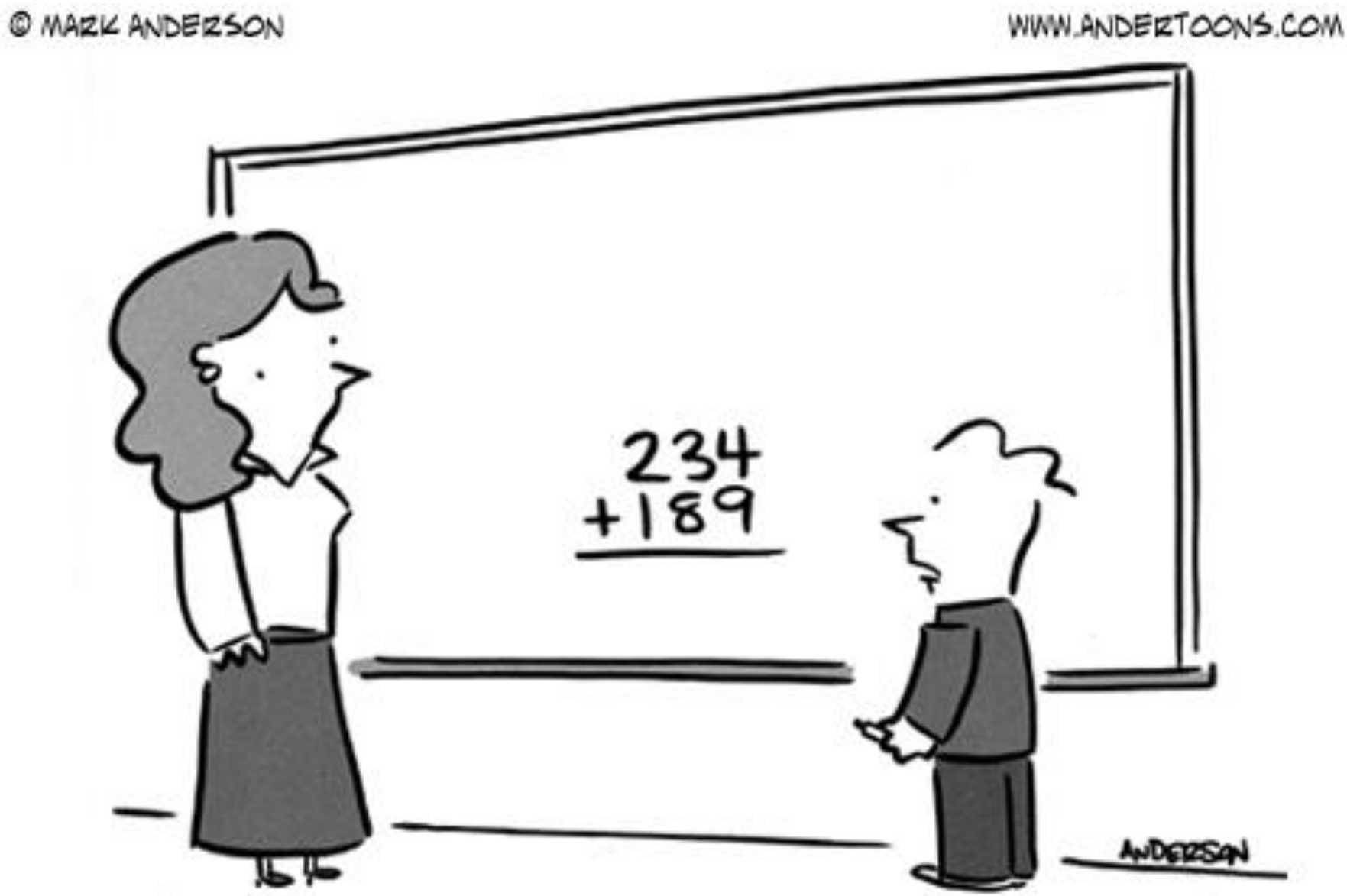
1. What is Data Mining (Big Data/Artificial Intelligence)
2. What we can and cannot do with it in clinical medicine.
3. Importance of Health Information Evaluation



Above is the Gartner curve that former MaineHealth CIO Barry Blumenfeld, MD, referred to in a 2012 interview with *Healthcare IT News*. The curve shows the various stages people who go implement or begin using new technology traditionally experience.

14 years of Google searches

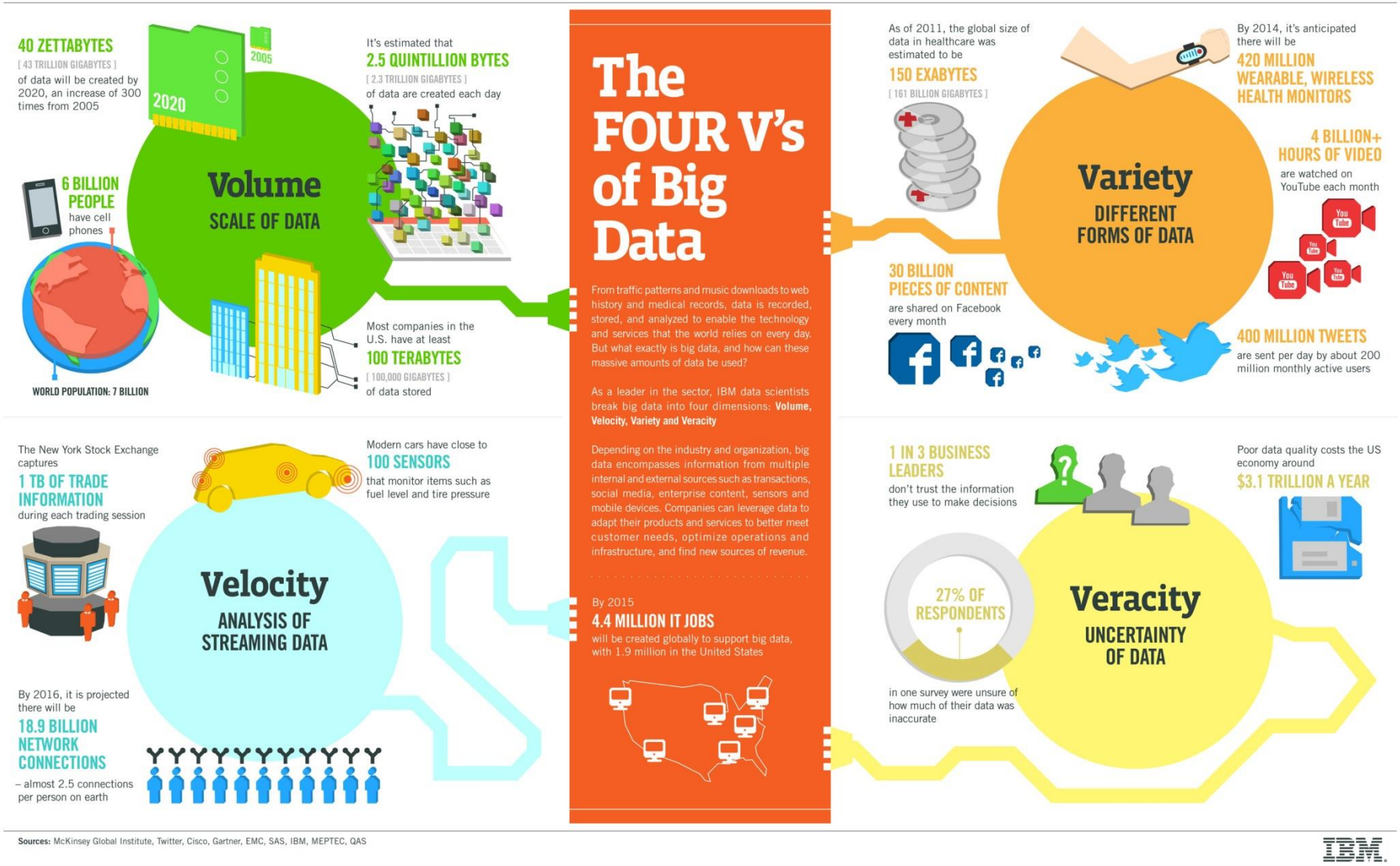




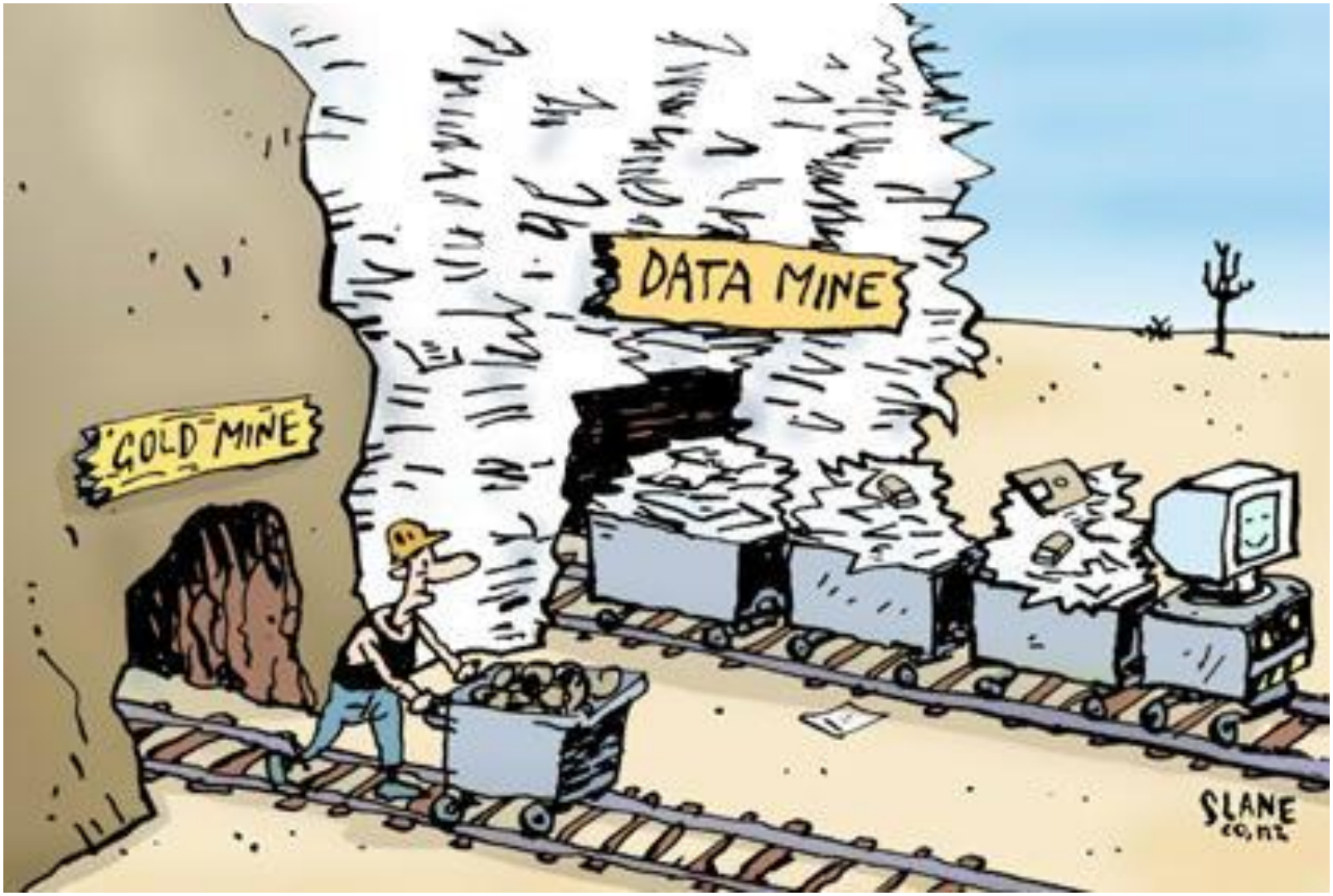
"Does this count as big data?"

Big data

- Large volumes of high velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information.



Big data encompasses such characteristics as **volume**, **variety**, **velocity** and, with respect specifically to healthcare, **veracity**.


GOLD MINE
DATA MINE
SLANE
co.nz

Data mining

Is a process to turn **raw data** into **useful information**. Data mining is the **process** of finding **anomalies**, **patterns** and **correlations** within **large data** sets to **predict** outcomes.

Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, reduce risks and more.

History

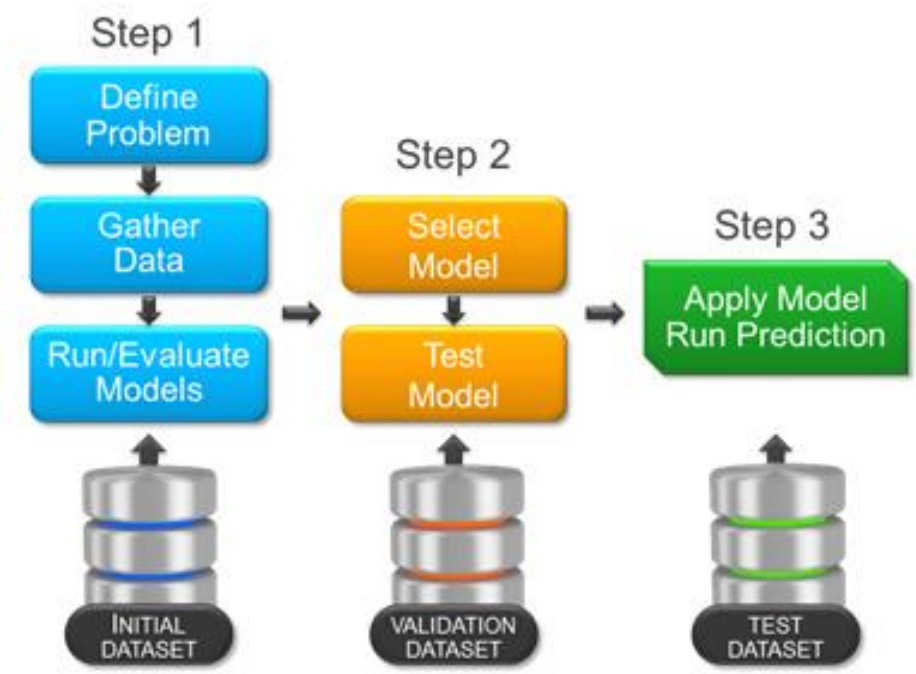
Sperry UNIVAC 1108 at NYU's UHMC

- Magnetic memory - 512K bytes.
- FASTRAND magnetic drum - 90 MB.
- Running at 1 MHz

Served the entire engineering school AND ran a real-time transaction system for the NYU Medical Center.

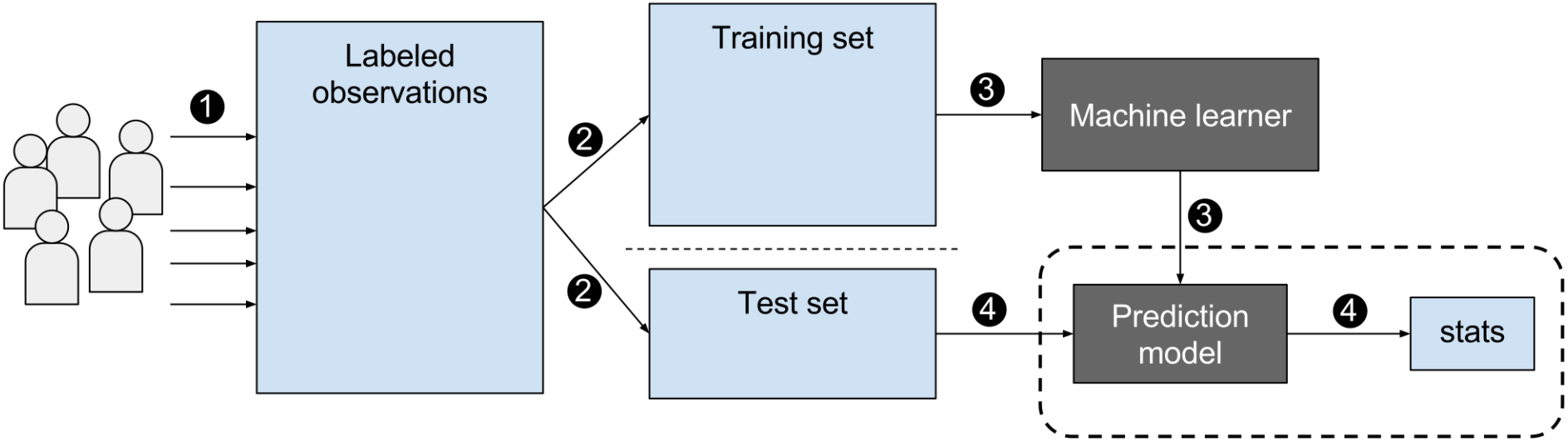
The process of digging through data to discover hidden connections and predict future trends <u>has a long history</u>. Sometimes referred to as "knowledge discovery in databases," the term <u>"data mining" wasn't coined until the 1990s</u>.

Data Mining



- Decision trees
- Random forests
- Support vector machines
- Nearest-neighbor
- K-means clustering
- Bayesian networks
- Multivariate regression
- **Neural networks**

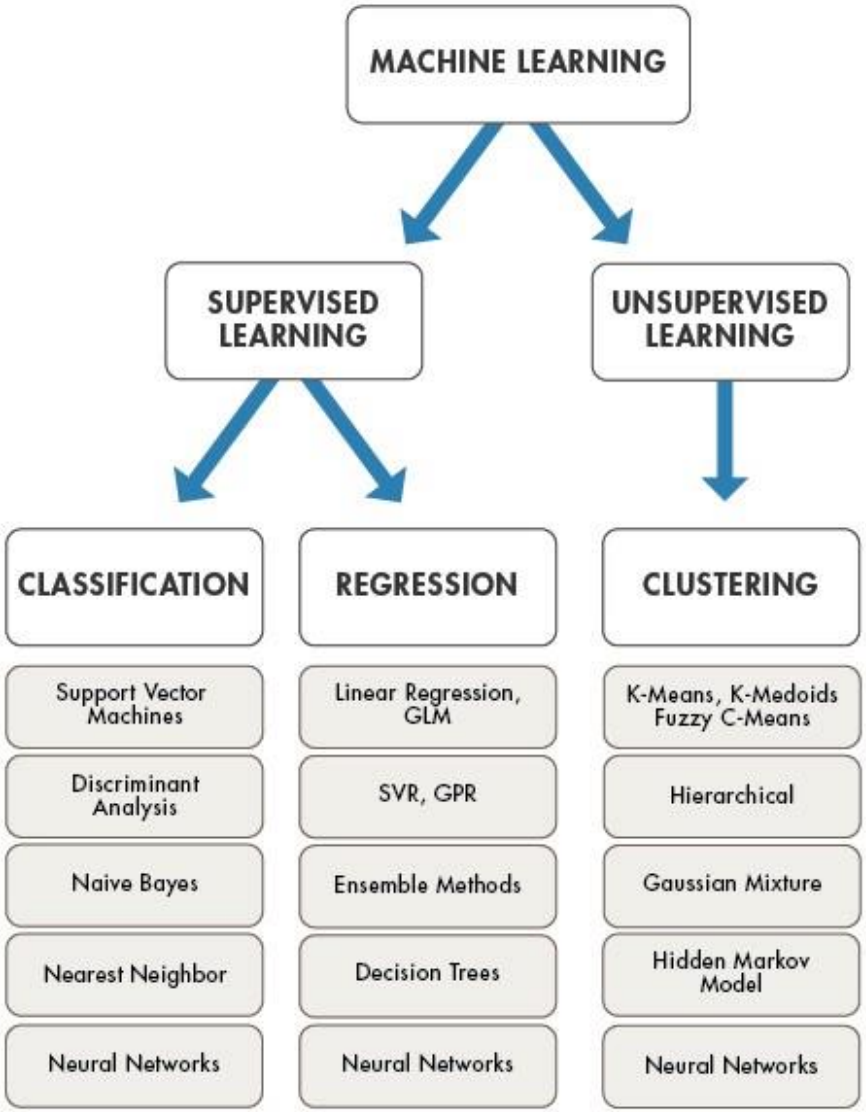
Machine learning



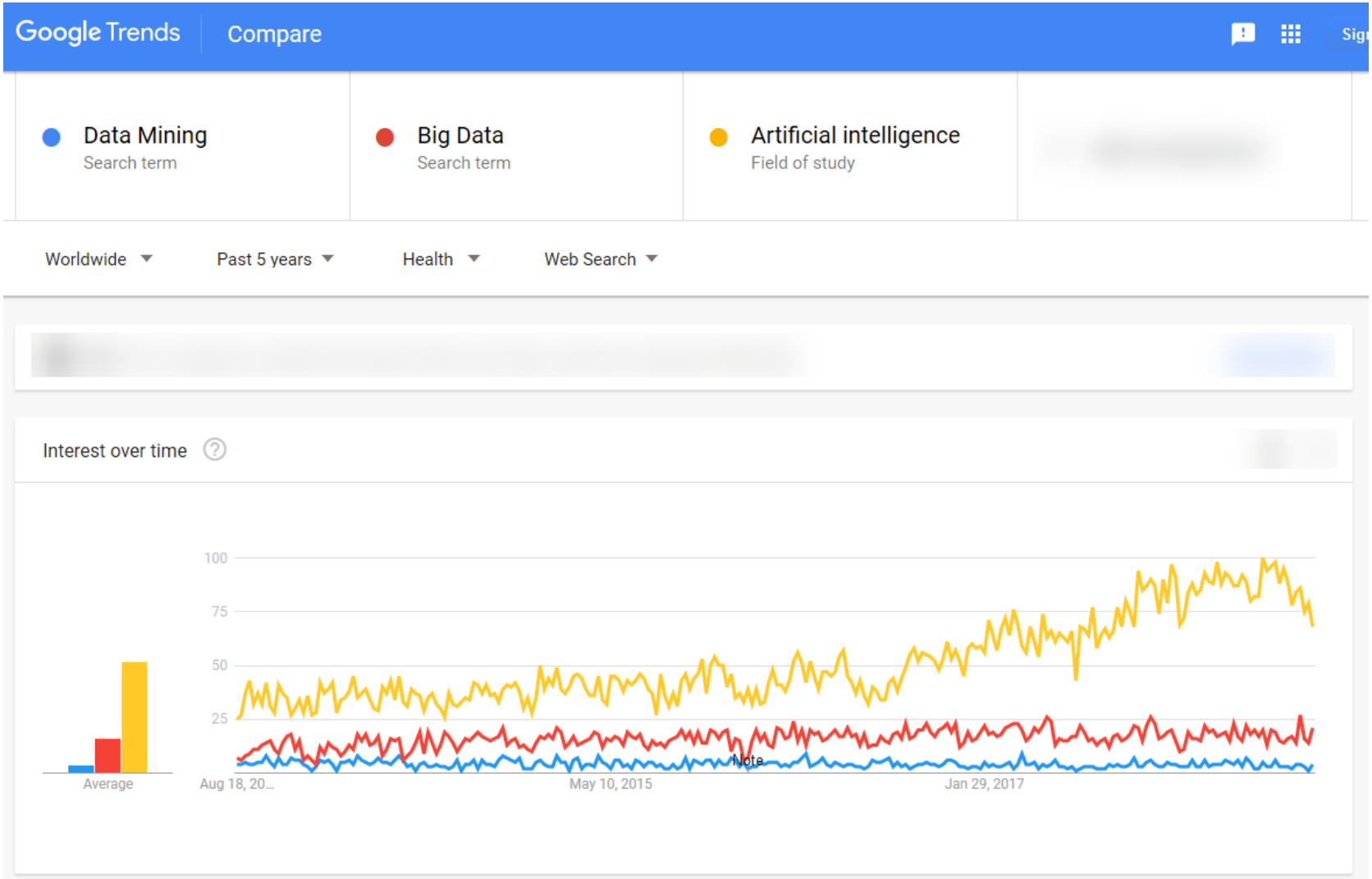
Is a method of data analysis that automates analytical model building.

Machine learning vs. data mining

- Machine learning and data mining use the same methods and overlap significantly
- **Machine learning** focuses on **prediction**,
- **Data mining** focuses on the **discovery** of unknown properties in the data.



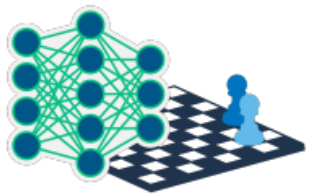
Last 5 years of Google search



Artificial Intelligence (AI)

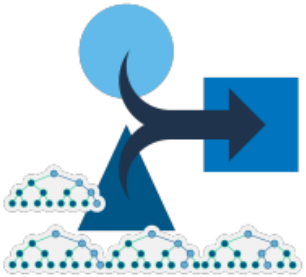
- Makes it possible for machines to learn from (human) experience, adjust to new inputs and perform human-like tasks.
- Most AI examples today – from chess-playing computers to self-driving cars – rely heavily on **deep learning**.
- Computers can be trained to accomplish specific tasks by processing large amounts of data and recognizing patterns in the data.

Not a novel



1950s–1970s

Neural Networks



1980s–2010s

Machine Learning

Present Day

Deep Learning

- 1956 - The term **AI was coined** in.
- 1960s - US DoD began training computers to **mimic basic human reasoning**.
- 1970s - DARPA completed **street mapping** project.
- 2003 - DARPA produced **intelligent personal assistant** - long before Siri, Alexa.

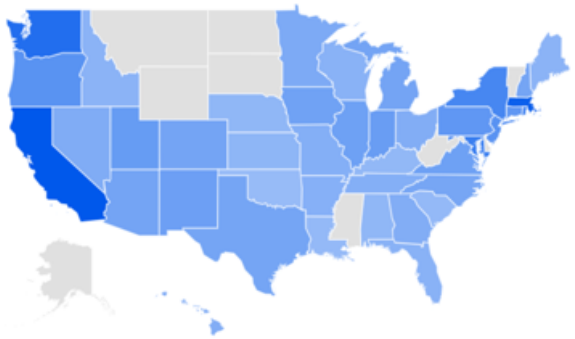
AI has become more popular today - increased data volumes, advanced algorithms, and improvements in computing power and storage.

Google Trends | Explore

Sign

deep learning
Search term

+ Compare



1	California	100
2	Massachusetts	93
3	Washington	81
4	Maryland	67
5	New York	60

Deep learning

- **Deep learning** is a type of **machine learning** that trains a computer to perform human-like tasks.
- Instead of organizing data to run through predefined equations, deep learning sets up basic parameters about the data and **trains the computer to learn on its own** by recognizing patterns using many layers of processing.
- Deep learning is one of the foundations of **artificial intelligence (AI),** and the current interest in deep learning is due in part to the buzz surrounding AI.

Deep learning

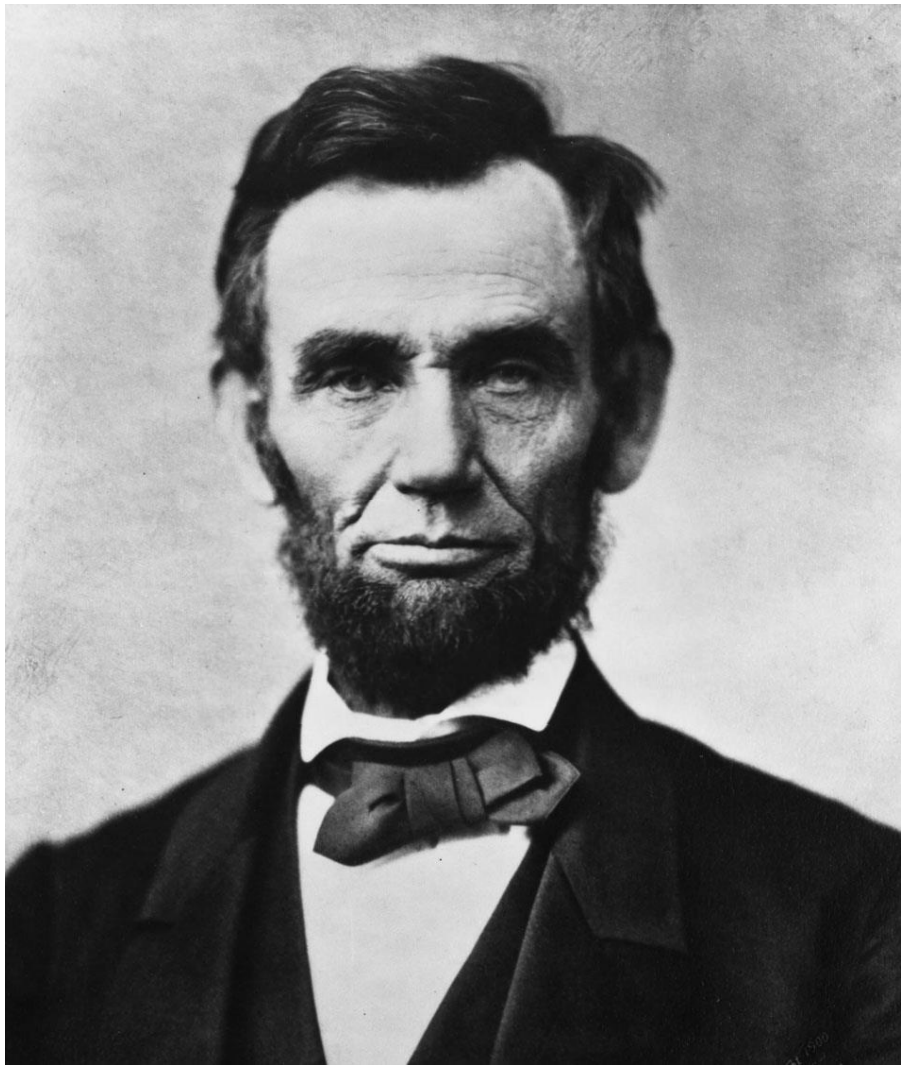
- Speech Recognition
- Natural Language Processing
- Image (Video) Recognition
- Recommendation Systems

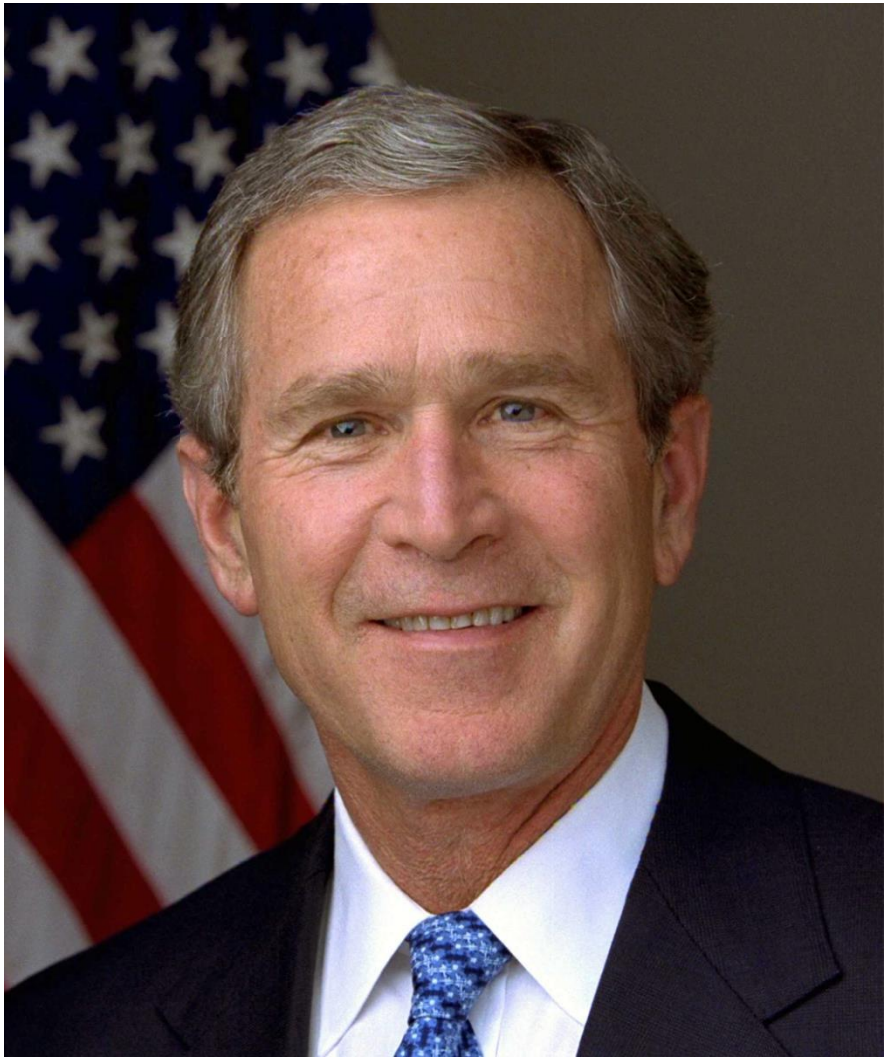
Pattern recognition

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

http://www.mrc-cbu.cam.ac.uk/~mattd/Cmabrigde/
http://www.stevesachs.com/jumbler.cgi

~~Ambient~~ Human intelligence





"Will AI replace me? The short answer is **NO**."

What are the limitation of AI?

The principle limitation of AI is that it learns from the data.

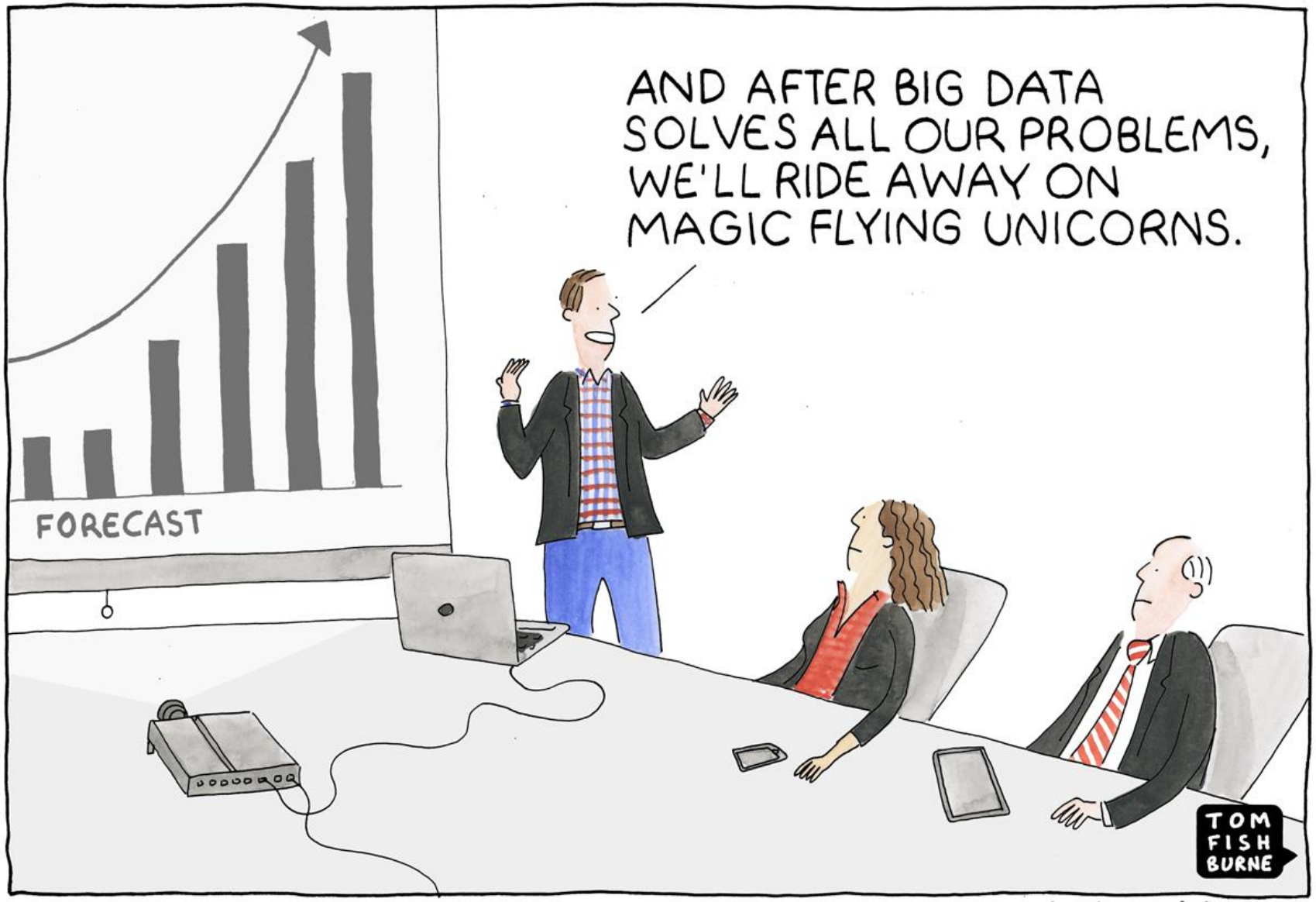
- There is no other way in which knowledge can be incorporated.
- That means any inaccuracies in the data will be reflected in the results.
- And any additional layers of prediction or analysis have to be added separately.

What are the limitation of AI?

Today's AI systems are trained to do a clearly defined task.

- The system that plays poker cannot play solitaire or chess.
- The system that detects fraud cannot drive a car or give you legal advice.
- In fact, an AI system that detects health care fraud cannot accurately detect tax fraud or warranty claims fraud.

The imagined AI technologies that you see in movies and TV are still science fiction.



© marketoonist.com

2011

Paging Dr. Watson: AI Jeopardy! Soon To Be Physician's Assistant

By **Jeremy Ford** - Mar 09, 2011 5,032

Is there an AI doctor in the house?

Will you ever be treated by Dr. Watson? Not Sherlock Holmes's right-hand man, but the AI Jeopardy! champion who's poised to be a sidekick for future physicians. IBM and Nuance Healthcare have teamed up with Columbia University and the University of Maryland to build a medical Watson that's fine-tuned to address the queries of doctors. The goal is

USA TODAY | News Subscribe | Mobile Google USA TODA

Home | News | Travel | Money | Spor

Tech: Blogs | Products | Gaming | Science & Space

IBM's Watson delving into medicine

By Jim Fitzgerald, Associated Press Posted 5/21/2011 4:28:22 PM | 3

YORKTOWN, N.Y. — Some guy in his pajamas, home sick with bronch complaining online about it, could soon be contributing to a digital colle information designed to help speed diagnoses and treatments.

2013

THE UNIVERSITY OF TEXAS
MD Anderson
~~Cancer~~ Center

myMDAnderson | Locations |

PATIENTS & FAMILY | PREVENTION & SCREENING | DONORS & VOLUNTEERS | FOR PHYSICIANS | RESEARCH

MD Anderson Taps IBM Watson to Power "Moon Shots" Mission

MD Anderson News Release 10/18/13

The University of Texas MD Anderson Cancer Center and IBM today announced that MD Anderson is using the IBM Watson cognitive computing system for its mission to eradicate cancer. Following a year-long collaboration, IBM and

2016

Special Review of
Procurement Procedures Related to the
M.D. Anderson Cancer Center Oncology Expert Advisor Project

Through August 31, 2016, approximately $62.1 million has been paid to external firms for planning, project management, and development of OEA. More than half of the funding used towards the system came from restricted gifts donated or pledged specifically for this purpose. This total reflects payments to external entities only; it does not include internal resources such as staff time, technology infrastructure, or administrative support. OEA has not been updated to integrate with MD Anderson's new electronic medical records system, and is not in clinical use.

2018

THE WALL STREET JOURNAL.

U.S. Edition ▾ | August 15, 2018 | Today's Paper | Video

Home World U.S. Politics Economy **Business** Tech Markets Opinion Life & Arts Real Estate WSJ. Magazine Search

BUSINESS

IBM Has a Watson Dilemma

Big Blue promised its AI platform would be a big step forward in treating cancer. But after pouring billions into the project, the diagnosis is gloomy.

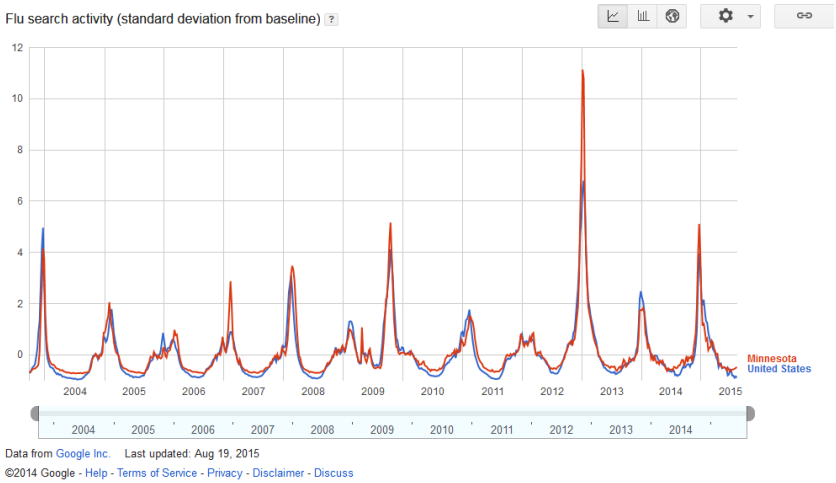
By Daniela Hernandez and Ted Greenwald

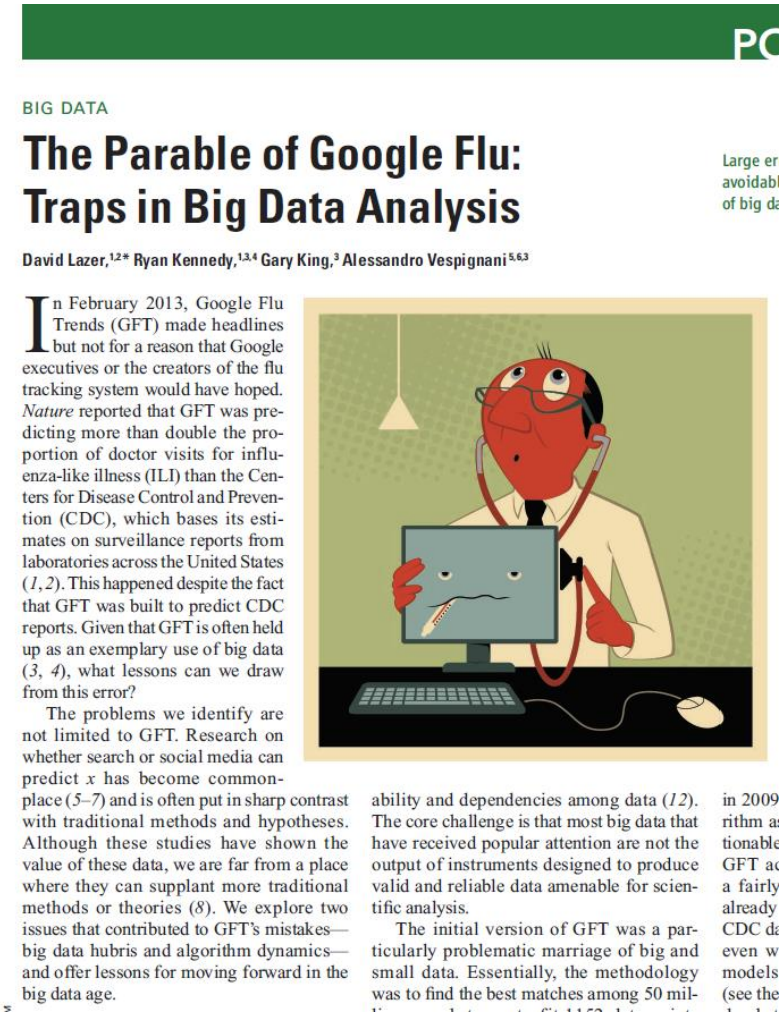
Aug. 11, 2018 12:19 a.m. ET

Can Watson cure cancer?

Tracking Disease Outbreaks

- One of the earliest examples was Google Flu Trends, which began offering real-time data to the public in 2008. Based on people's Internet searches for flu-related terms, this tool monitored flu outbreaks worldwide.





BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large er
avoidabl
of big da

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can predict *x* has become commonplace (*5*–*7*) and is often put in sharp contrast with traditional methods and hypotheses. Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories (*8*). We explore two issues that contributed to GFT's mistakes—big data hubris and algorithm dynamics—and offer lessons for moving forward in the big data age.

ability and dependencies among data (*12*). The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.

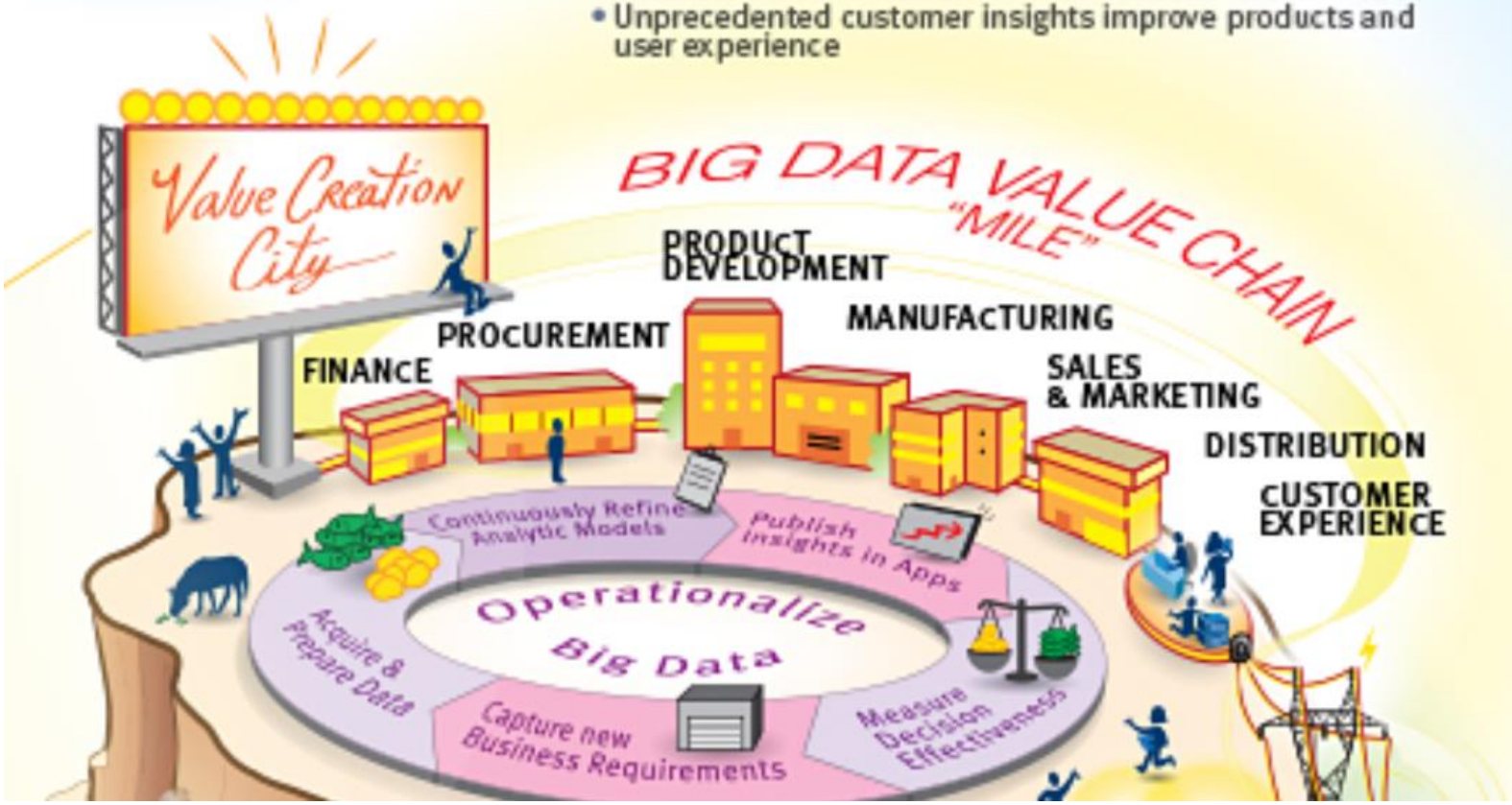
The initial version of GFT was a particularly problematic marriage of big and small data. Essentially, the methodology was to find the best matches among 50 mil-

in 2009
rithm a
tionable
GFT ac
a fairly
already
CDC da
even w
models
(see the

EMC²

FUTURE STATE

- Big Data analytics continuously drive business innovations
- Real-time insights optimize operations and boost profits
- Unprecedented customer insights improve products and user experience

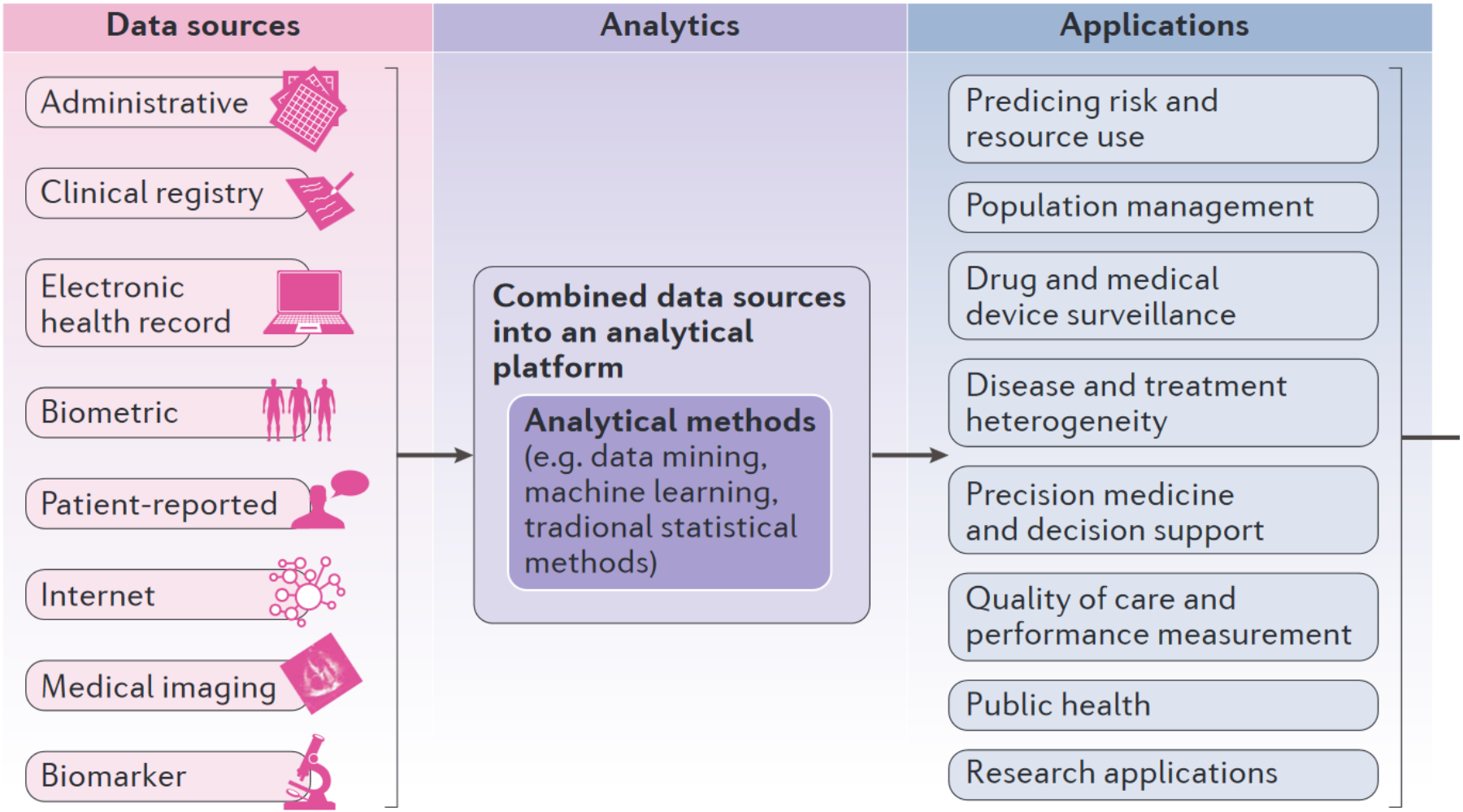


Big Data in Healthcare

In Medicine…

1. Use of big data to drive better health delivery
2. Application of “big data” to improve health care performance
3. Accelerate new discoveries

Overview of big data applications



*Rumsfeld, J. S., et al. (2016). "Big data analytics to improve cardiovascular care: promise and challenges." Nature Reviews Cardiology **13: 350.***

BUT…. Big data/Data mining is NOT magic:

- Data mining will not automatically discover solutions without guidance, will not sit inside of your database and send you an email when some interesting pattern is discovered.
- Data mining may find interesting patterns, but it does not tell you the value of such patterns.

http://callingbullshit.org/

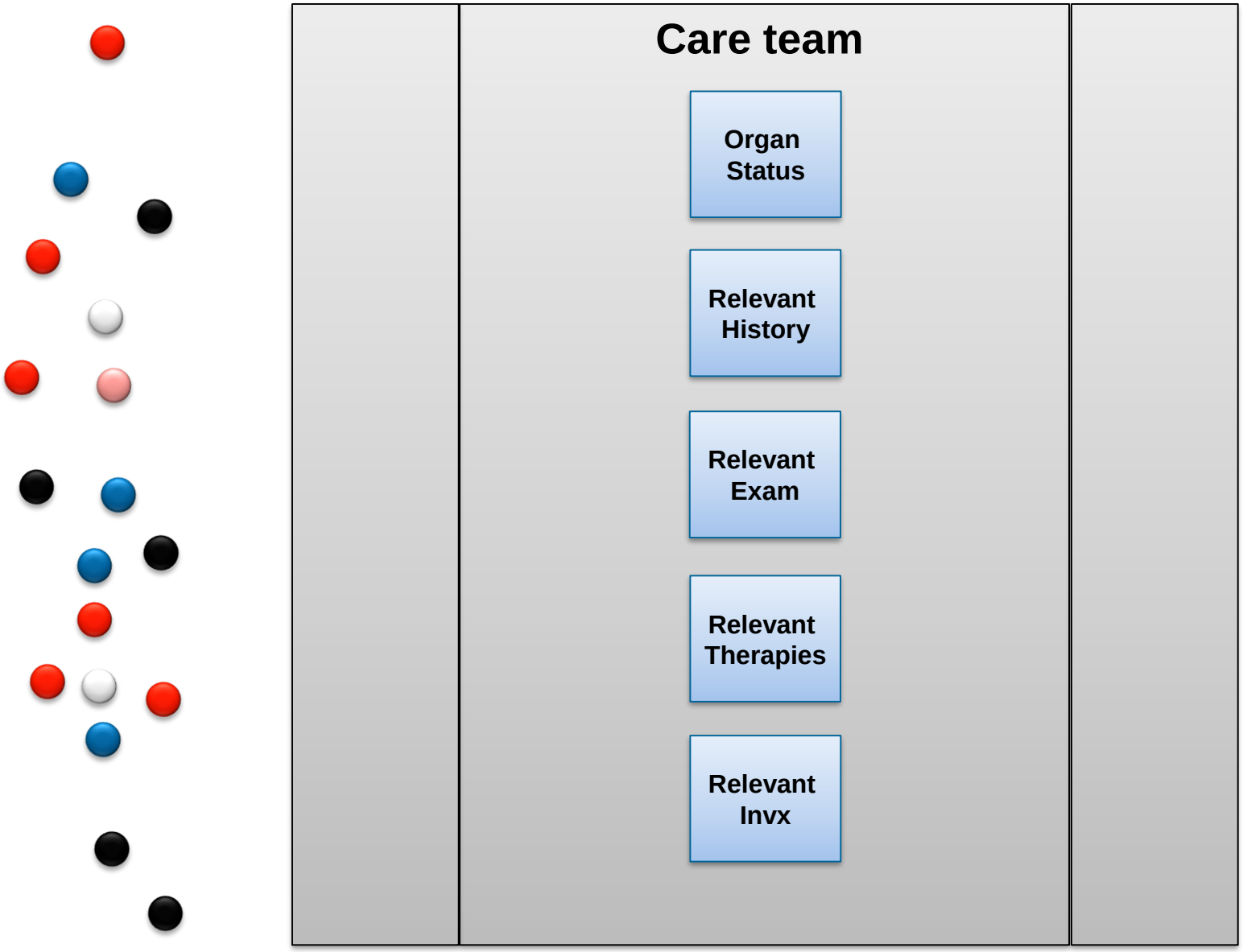
Calling Bullshit in the Age of Big Data

Logistics

Course: INFO 198 / BIOL 106B. University of Washington
To be offered: Spring Quarter 2017
Credit: 1 credit, C/NC
Time and Location: Wednesday 3:30-4:20 MGH 389
Enrollment: 160 students
Instructors: Carl T. Bergstrom and Jevin West
Synopsis: Our world is saturated with bullshit. Learn to detect and defuse it.

The course will be offered as a 1-credit seminar this spring through the Information School at the University of Washington. We aim to expand it to a 3 or 4 credit course for 2017-2018. For those who cannot attend in person, we aim to videotape the lectures this spring and make video clips freely available on the web.

Information Flow in the ICU



Data

Information Packets

Clinical Decision

Processing clinical data – the 30000 ft. view

Tools and methods for data entry

- Sensors
- Laboratory
- Imaging
- Natural context
- CPOE
- etc

Control, managing and storing of the data

- Databases
- Warehouses
- Data marts

Standardization and transferring of the data

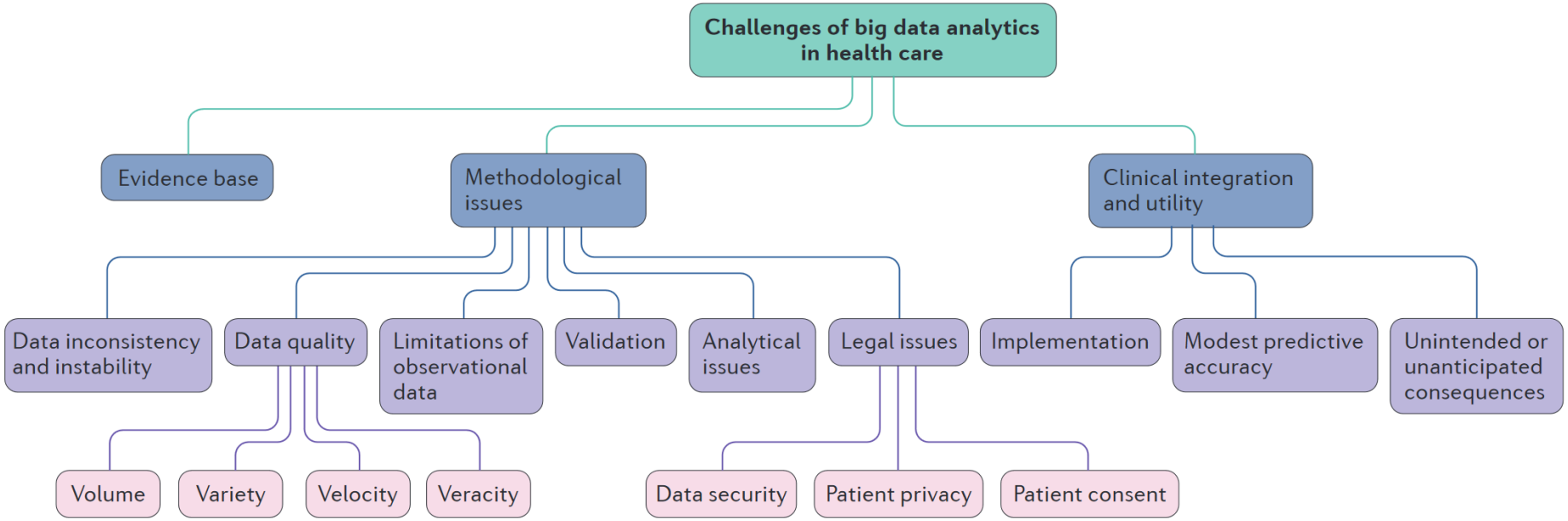
- HL7
- DICOM
- ICD
- SNOMED CT
- CPT
- etc

Clinical use of information

- EMR viewers
- Displays
- CDSS

Herasevich V, Litell J, Pickering B. Electronic medical records and mHealth anytime, anywhere. Biomed Instrum Technol. 2012 Fall;Suppl:45-8. PMID: 23039776.

Challenges in healthcare big data

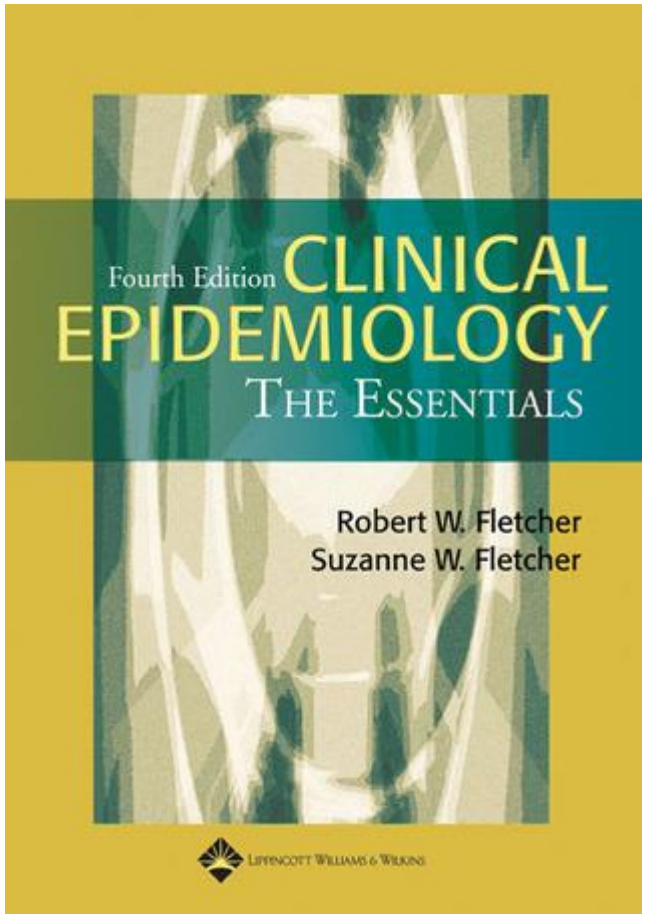


*Rumsfeld, J. S., et al. (2016). "Big data analytics to improve cardiovascular care: promise and challenges." Nature Reviews Cardiology **13: 350.***

MAYO CLINIC

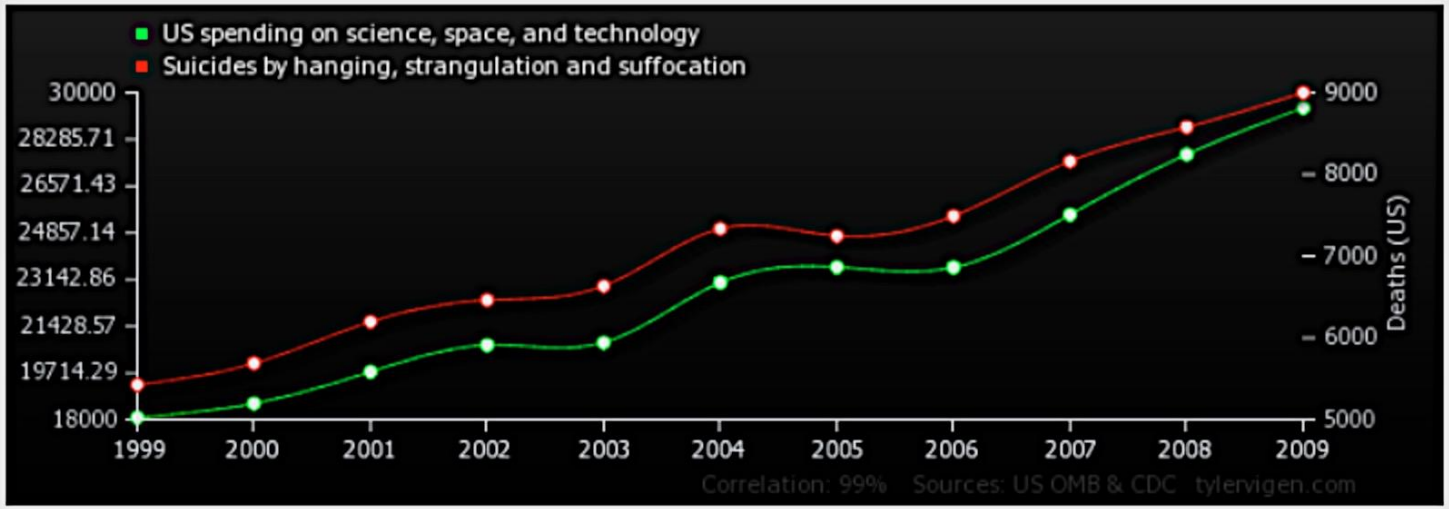
Problem: Big Data Hubris

- "Big data hubris" is the often implicit assumption that big data are a **substitute** for, rather than a **supplement** to, traditional data collection and analysis.



Problem: Data mining does not infer causality

US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
US spending on science, space, and technology *Millions of todays dollars (US OMB)*	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578

Correlation: 0.992082

http://www.tylervigen.com/view_correlation?id=1597

Problem: EMR data has pre-test probability

- EMR data has characteristics that decrease the practicality of most predictive models.
- It is Pretest Probability which is the probability of a patient having a target disorder before a diagnostic test result is known.
- Data is present in the EMR when clinicians cause it to be there as they suspect a specific health problem. For example, a diagnostic troponin test is ordered because a physician suspects myocardial infarction.

Problem: data quality

Additional complexity added by missing data or **delayed data** in the EMR.

Virtual Special Issue on Improving and Maintaining Health Data Quality: Guest Editorial

HIMJ

Data quality: "Garbage in – garbage out"

Health Information Management Journal
1–3

DOI: 10.1177/1833358318774357
journals.sagepub.com/home/himj
SAGE

Monique F Kilkenny, *BAppSc(MRA), GradDipEpid/Biostats, MPH, PhD*[1,2],
Kerin M Robinson, *BHA, BAppSc(MRA), MHP, PhD, CHIM*[3]

Introduction

"Garbage in – garbage out" is a colloquial recognition of poor quality data entry leading to unreliable data output.

Trinh et al. (2017), in their comparison of two routinely collected datasets (*Incident Information Management System and the health information exchange in hospitals in*

Advantages to healthcare

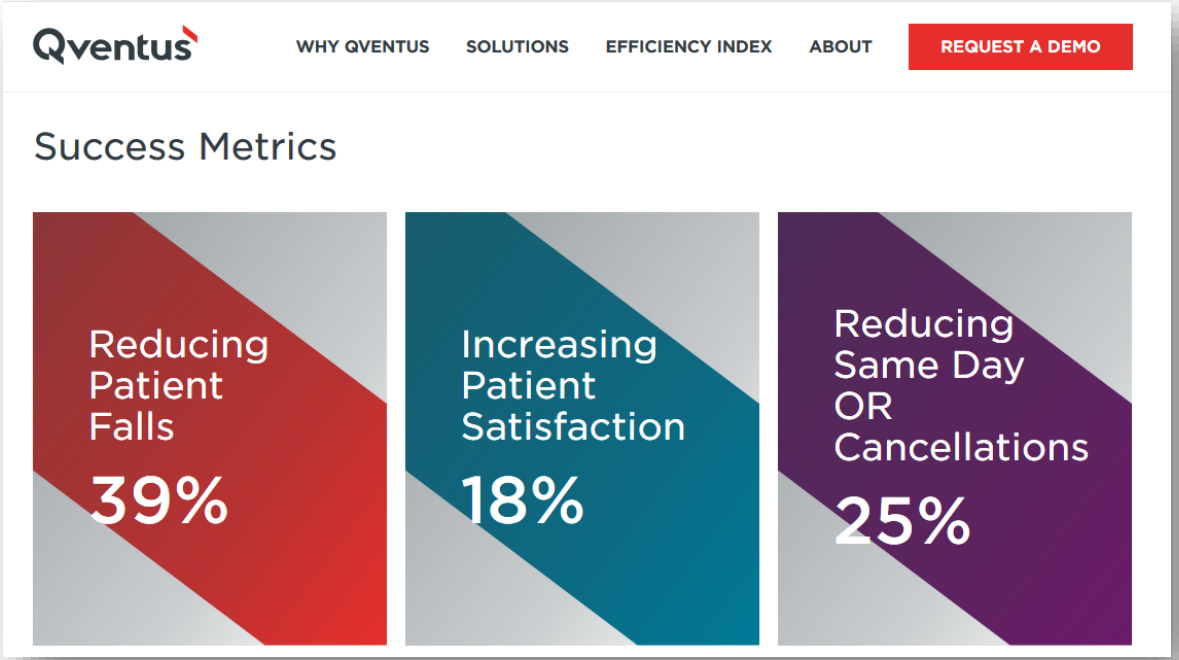
- Performance Evaluation
- Financial Planning
- Patient Satisfaction
- Healthcare Management
- Quality Scores and Outcome Analysis
- Labor Utilization

Advantages to healthcare

1. **Clinical operations:** Comparative effectiveness research.
2. **Research & development:** 1) predictive modeling 2) improve clinical trial design and patient recruitment.
3. **Public health:** analyzing disease patterns and tracking disease outbreaks.
4. **Evidence-based medicine:** Combine and analyze a variety of structured and unstructured data.
5. **Genomic analytics.**
6. **Pre-adjudication fraud analysis:** to reduce fraud, waste and abuse.
7. **Device/remote monitoring:** safety monitoring and adverse event prediction;
8. **Patient profile analytics:** to identify individuals who would benefit from proactive care or lifestyle changes.

HIT evaluation

Start with question – not technology



1. **What is the setting?**
2. **What is the sample size?**
3. **What is the comparison group?**
4. **How biases controlled?**
5. **How statistical analysis was done?**

Clever marketing?

Reported effect

Publication bias

Funding bias

Measurement bias

Observer bias

Recall bias

Selection bias

Confounding

Random error

Real causal effect

HIT Stakeholders

Technology

Patients and Families

- Is it safe?
- Is it help me?

Clinicians

- Is it fast?
- Is it accurate?
- Is it user-friendly?

IT and Security

- Does it work?
- Will they use it?
- Is it secure?

Administrators/Purchasers

- What is the cost/benefit?
- Is it reliable?

Herasevich V, Pickering BW Health Information Technology Evaluation Handbook: From Meaningful Use to Meaningful Outcome, 2017, 208 pages, CRC Press, ISBN 978-1498766470

April 1995

College of Healthcare Information Management Executives

April 1995
$4.95

healthcare INFORMATICS

with a leader.
Phone 1-800-4-PENKEY (toll-free) to learn what the future offers you. It's your call.

information needs for tomorrow.

systems integration performance without vendor involvement — look to LDE™. It puts you in control.

q&a

Building the Electronic Information Warehouse

dy 1994 report published by Kidder, Peabody & Co. Inc., New York, claimed, "Cost is the most serious issu ng care today. Providers, medical institutions and third-party payors are aggressively planning and executin e-rein in prices and cut costs. At the same time, purchasers of healthcare, such as large employers, the gov nt dividuals are demanding more value for these increased costs."

Using Artificial Intelligence to Predict Myocardial Infarction

1995

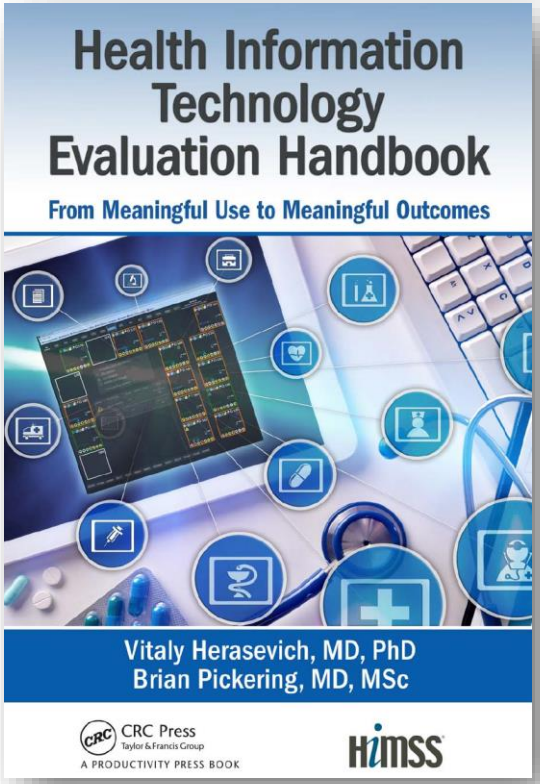
shifting gears

Reengineering Information Systems with the Customer in Mind

The challenge:
Control costs while improving the quality and accessibility of care

Last points

- Know your data
- Use high quality data
- Understand limitation of mining approach
- Use clinical reasoning



ISBN-10: 1498766471

Thank You!

Google → "Clinical informatics Mayo"

vitaly@mayo.edu