e-Science

Elmer V. Bernstam, MD
Professor
Biomedical Informatics and Internal Medicine
UT-Houston

Acknowledgements

- Todd Johnson (UTH → UKy)
- Jack Smith (Dean at UTH SBMI)
- CTSA informatics community
- Luciano Floridi
  - philosophy of information

- Portions adapted from:
  - Bernstam EV, Smith JW and Johnson TR. What is biomedical informatics? J Biomed Inform. 2010 Feb;43(1):104-10. Epub 2009 Aug 13.
  - Bernstam EV, Hersh WR, Johnson SB, *et al.* Synergies and Distinctions between Computational Disciplines in Biomedical Research: Perspective from the Clinical and Translational Science Award Programs. Acad Med, 2009 Jul;84(7):964-70.
  - Bernstam EV, Hersh WR, Sim I , *et al.* Unintended consequences of health information technology: a need for biomedical informatics. J Biomed Inform. 2010 Oct;43(5):828-30. Epub 2009 Jun 7.
  - Bernstam EV, Johnson TR. Why health information technology doesn't work. Bridge, 39:4.

Biomedical (informatics) perspective

- My background
- Examples drawn from this domain
- However, most should be generally applicable

What is e-Science?

- **E-Science** (or **eScience**) is computationally intensive science that is carried out in highly distributed network environments, or science that uses immense data sets that require grid computing; the term sometimes includes technologies that enable distributed collaboration, such as the Access Grid.
  - Wikipedia
    - http://en.wikipedia.org/wiki/E-Science, Accessed 1/24/2012

I prefer…

- Science that requires computation
  - Not science (of) computation = computer science
- Why?
  - Don't have to have networks.
  - Don't have to have large datasets.
  - Doesn't have to be distributed or require a grid.
  - But must require computation…

Data, information and knowledge

- Often used, many definitions
- Philosophy of computing:
  - Data: observations about the world.
    - Example: 35
  - Information: data + meaning
    - Example: Body Mass Index (BMI) = 35
  - Knowledge: justified true belief
    - Example: Persons with a BMI $\geq$ 30 are at greater risk of diabetes mellitus.

Caveats

- It is difficult for humans to discuss data
  - How can we talk about anything without considering its' meaning?
  - Humans are meaning (information) processors
- I will try to separate opinion from fact
  - Can be difficult sometimes, so be wary

Why is this important?

- Our (current) technology deals with data
  - IT is a misnomer
- People deal with information and knowledge
- Semantic gap
  - Difference between data and information
  - $1 vs. pneumonia

Computerization

- Fields where the difference between data and information is small → computerized
  - E.g., banking, $1
- Fields where the difference between data information is large → not computerized
  - E.g., clinical medicine, pneumonia

Bank Account Model

- Mapping to floating point representation, plus procedures for + and - are sufficient
  - Note: ignoring interest, etc.
- The representation of a number admits a simple procedure to compute + and –
  - procedure for 150+10 is the same as 2000+562
- The symbolic representation of a number plus simple procedures are sufficient to model bank accounts
- We can ignore most of the economic concepts

Consider biomedical concepts

- Most concepts represented using words, such as “Hypertension”
- Consider the operations:
  - Is it a disease?
  - What are its symptoms?
  - What kind of disease is it?
  - What systems does it affect?
- Nothing in the representation admits an easy procedure for answering these questions
  - Consider ICD-9-CM codes: 401 (Hypertension)
  - Leibniz Classification System: Each attribute is a prime, concepts are products of primes
    - » Deciduous: 3, Plant: 5
    - » Deciduous Plant: 3*5 = 15
    - » If Vine is 105, we know it is a Deciduous plant, because it is divisible by 15

Big data vs. small data

- Different challenges
- Big data – usually associated with e-Science
  - Challenge is volume of data (scale)
    - Climate
    - Genomics/proteomics (-omics)
    - Physical sciences
  - Semantic gap is generally small

Big data vs. small data

- Small data
  - Challenge is making sense of the data
    - Clinical informatics
- Clearly there are big data approaches to small data problems
  - E.g., automated translation – statistical vs. semantic approaches
    - Alon Halevy , Peter Norvig , Fernando Pereira, The Unreasonable Effectiveness of Data, IEEE Intelligent Systems, v.24 n.2, p.8-12, March 2009 [doi>10.1109/MIS.2009.36]

Big data

Big data

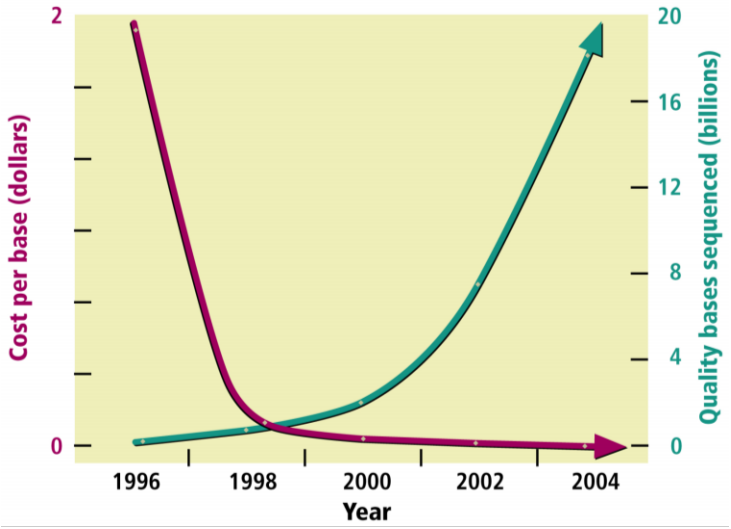
- Large data sets (>>GB)
- Often single elements are simple
  - E.g., bases in DNA (A, T, G, C)
- Difficulty comes from volume of data
  - How do you store the output of genetic sequencing machines
- Traditionally (but not necessarily) associated with biological data
  - Imaging, some public health databases, etc.

Example

- Solexa Genome Analyzer II (GAII) by Illumina
- Single sequencing run
  - 115,200 TIFF files (images)
  - Each TIFF file ~ 8MB
  - ~1TB/run

- How many runs will a typical storage system handle?
  - Current cost for research data storage at UTH:
    - $3,000 - 9,000/TB (January 2010)

Richter BG, Sexton DP, 2009 Managing and Analyzing Next-Generation Sequence Data. PLoS Comput Biol 5(6): e1000369. doi:10.1371/journal.pcbi.1000369



http://www.ornl.gov/sci/techresources/Human_Genome/project/whydoe.shtml

"Big data" problems

- Funding research infrastructure (i.e., hardware, software that enable science)
- Typically funded from
  - Institutional funds
    - Philanthropy, clinical income, research income
  - Research funds
    - Grants have direct costs + indirect costs
      - Direct costs = funding to do your research
      - Indirect costs = funding research infrastructure
      - Is IT "research" or "research infrastructure"

"Big data" problems

- "Typical" NIH individual research grant
  - ~$1M direct costs / 4 years
  - 50% indirect costs (i.e., $500k/4 years)
- $5k/TB → 300TB using all funds (no actual research)
- Bottom line: research data storage needs are now a significant (financial) problem with no clear solution.

Why does it cost so much?

- Data are stored multiple times
- People are expensive – cost is for installation, maintenance, backup, monitoring, etc.
- Backup
  - Periodic
  - Off site

Possible solution

- Consumer-level hardware
  - ~$50-$100/TB and dropping rapidly
  - But… much less reliable, no built-in backup, need for off site backup, etc.
- Approach
  - Buy lots of consumer-level hardware, deal with the failures, replace as needed
- Using consumer-level hardware promising, but whether it is cost-effective is yet an open question
  - Google seems to think so…
    - Ghemawat S, Gobioff H and Leunk ST. The Google file system. Presented at 19th ACM Symposium on Operating Systems Principles, Lake George, NY, October 2003.
      - http://labs.google.com/papers/gfs-sosp2003.pdf

Small data

Small data

- Generally (but not necessarily) associated with clinical datasets
- May be large, e.g., GB but not as large as "big data" datasets
- Usually text or numerical data
- Semantic gap is large

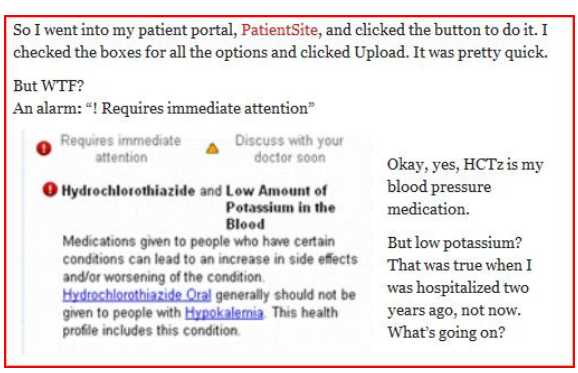
"Small data" problems

- "Big data" = computer-bound
- "Small data" = meaning (?human)-bound
- Challenge is to make sense of the data
  - Data → information
  - Can also be said of genetic data – e.g., functional genomics
- Examples:
  - Concept extraction from clinical text
  - Standards/vocabularies
  - Ontology maintenance and reconciliation
  - Data access/privacy

"Small data" problems

- Staff knowledge and costs are the rate-limiting resources
- Computers are important but usually not the rate-limiting components

E-patient Dave


So I went into my patient portal, PatientSite, and clicked the button to do it. I checked the boxes for all the options and clicked Upload. It was pretty quick.

But WTF?
An alarm: "! Requires immediate attention"

Requires immediate attention | Discuss with your doctor soon

Hydrochlorothiazide and Low Amount of Potassium in the Blood

Medications given to people who have certain conditions can lead to an increase in side effects and/or worsening of the condition. Hydrochlorothiazide Oral generally should not be given to people with Hypokalemia. This health profile includes this condition.

Okay, yes, HCTz is my blood pressure medication.

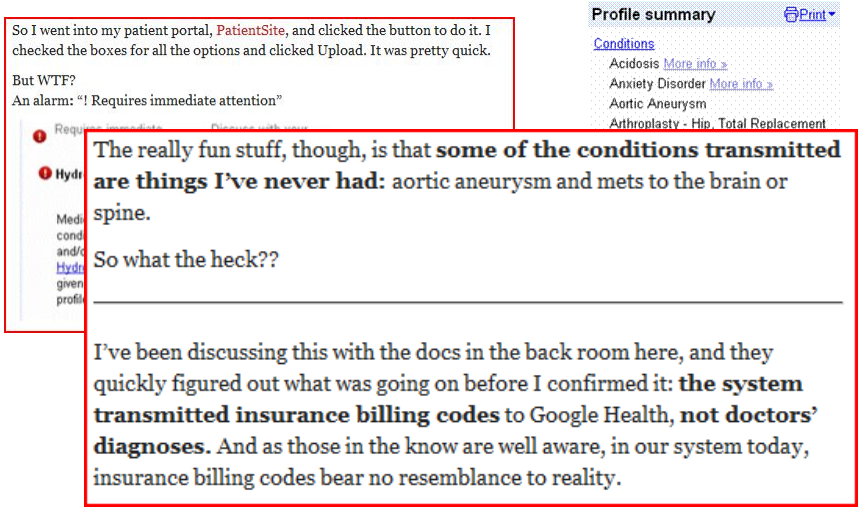
But low potassium? That was true when I was hospitalized two years ago, not now. What's going on?


Profile summary Print

Conditions
- Acidosis More info »
- Anxiety Disorder More info »
- Aortic Aneurysm
- Arthroplasty - Hip, Total Replacement
- Bone Disease
- CANCER
- Cancer Metastasis to Bone
- Cardiac Impairment
- CHEST MASS
- Chronic Lung Disease
- Depressed Mood More info »
- DEPRESSION More info »
- Diarrhea
- Elevated Blood Pressure More info »
- Hair Follicle Inflammation with Abscess in Sweat Gland Areas
- HEALTH MAINTENANCE
- HYDRADENITIS
- HYPERTENSION More info »
- Inflammation of the Large Intestine More info »
- Intestinal Parasitic Infection

http://e-patients.net/archives/2009/04/imagine-if-someone-had-been-managing-your-data-and-then-you-looked.html, accessed 7/3/2009

E-patient Dave



So I went into my patient portal, PatientSite, and clicked the button to do it. I checked the boxes for all the options and clicked Upload. It was pretty quick.

But WTF?
An alarm: "! Requires immediate attention"

Profile summary Print

Conditions
Acidosis More info »
Anxiety Disorder More info »
Aortic Aneurysm
Arthroplasty - Hip, Total Replacement

The really fun stuff, though, is that **some of the conditions transmitted are things I've never had:** aortic aneurysm and mets to the brain or spine.

So what the heck??

I've been discussing this with the docs in the back room here, and they quickly figured out what was going on before I confirmed it: **the system transmitted insurance billing codes** to Google Health, **not doctors' diagnoses.** And as those in the know are well aware, in our system today, insurance billing codes bear no resemblance to reality.

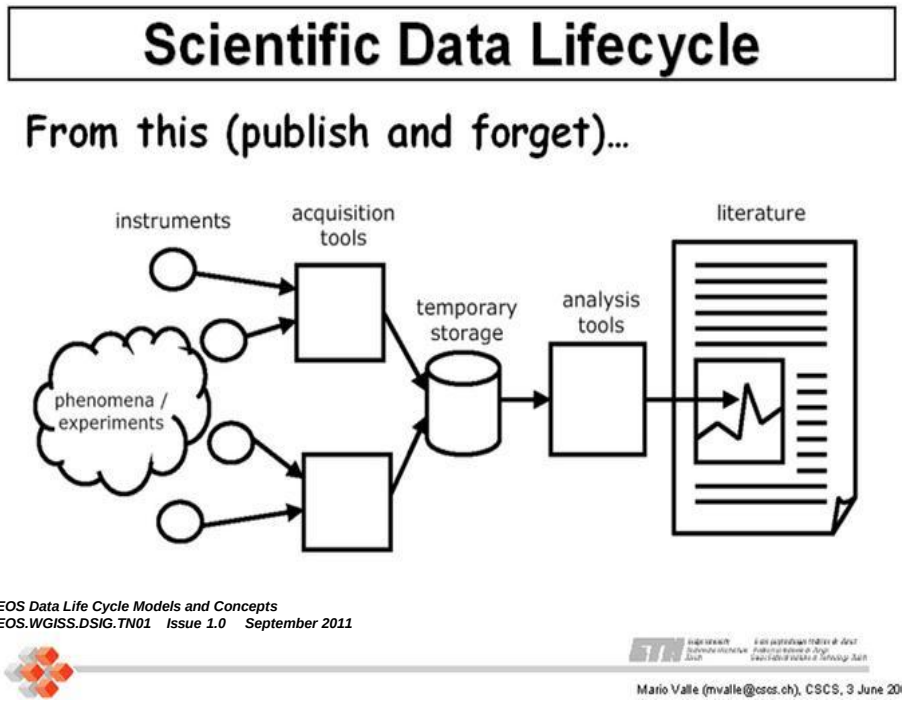
http://e-patients.net/archives/2009/04/imagine-if-someone-had-been-managing-your-data-and-then-you-looked.html, accessed 7/3/2009

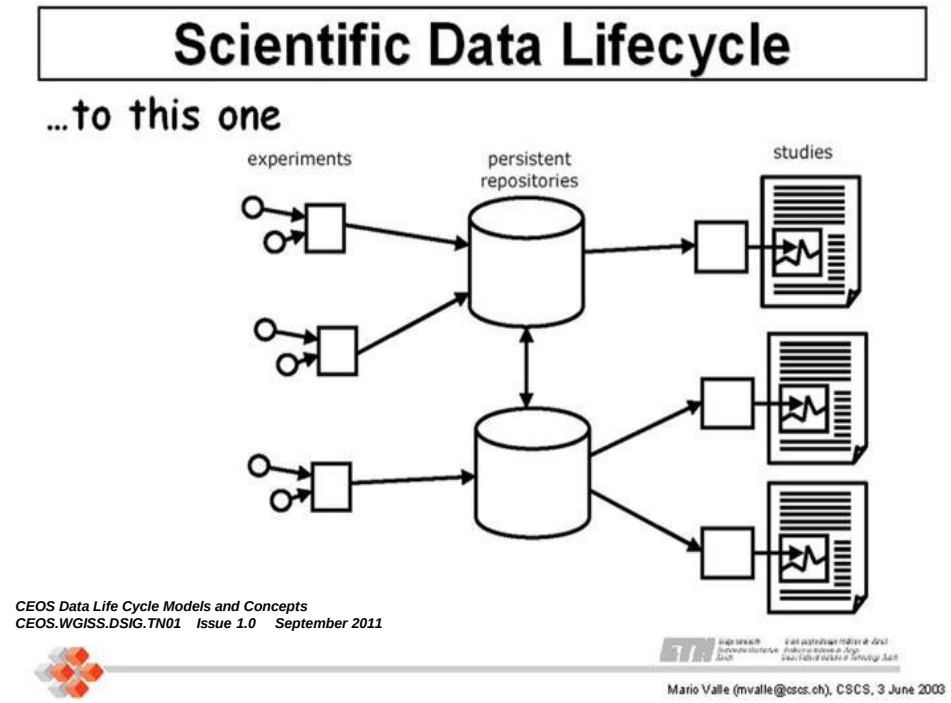
"Small data" problems: data access

- Suppose that I have a large clinical data set…
  - E.g., I am a provider (e.g., hospital, clinic)
- Why should I give you access to my data?
- What am I risking?
- What am I gaining?

"Small data" problems: data access

- HIPAA and privacy laws
  - Very open to interpretation
- Cost/benefit of privacy
  - No consensus or even rational conversation
    - E.g., What does the lack of a unique patient identifier cost?
- Sometimes privacy is used as an excuse to avoid sharing data
- Note that the issue is not: "can I," but "may I"
  - Faster computer doesn't help

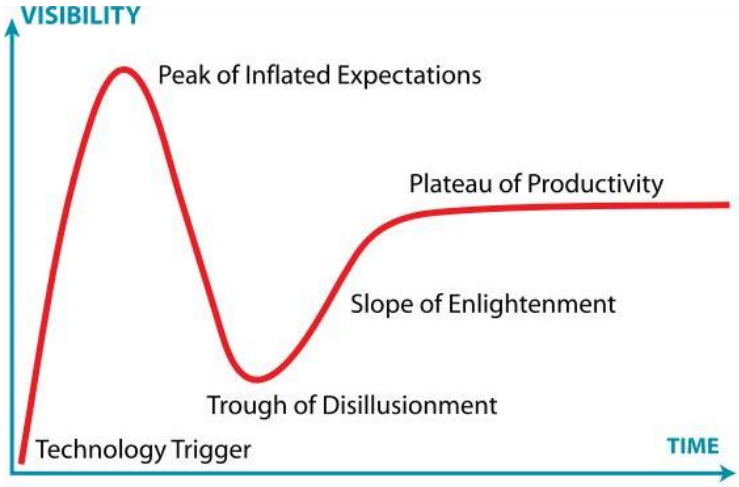




Research data life cycle issues

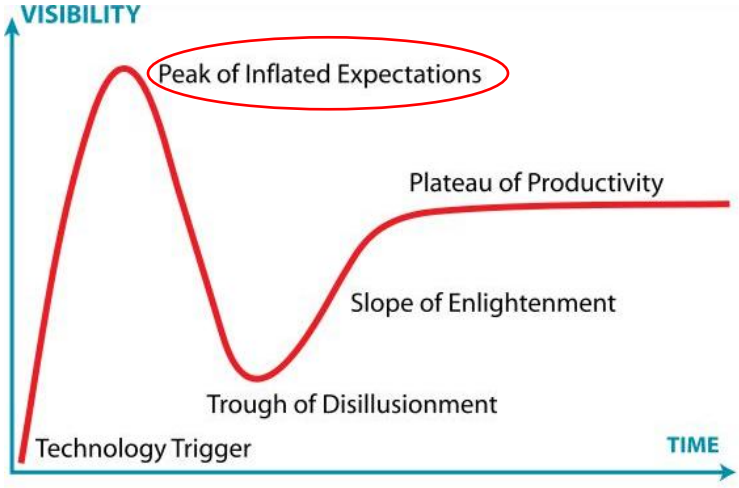
- How do you document a data set?
  - At UTH we allow only intermediated access to our clinical data
  - Calculator argument repeat?
- Meaning is central
  - Must understand data provenance (interpreted broadly) to re-use data

Gartner Hype Cycle



http://en.wikipedia.org/wiki/Hype_cycle

Gartner Hype Cycle



http://en.wikipedia.org/wiki/Hype_cycle

"AI winters" (plural) – cycles of boom (enthusiasm) and bust (disappointment) associated with technology

Your **continued donations** keep Wikipedia running!

Log in / create account

article | discussion | edit this page | history

WIKIPEDIA
The Free Encyclopedia

navigation
- Main page
- Contents
- Featured content
- Current events
- Random article

search
Go Search

interaction
- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

AI winter

From Wikipedia, the free encyclopedia
(Redirected from AI Winter)

In the history of artificial intelligence, an **AI winter** is a period of reduced funding and interest in artificial intelligence research.[1] The process of hype, disappointment and funding cuts are common in many emerging technologies (consider the railway mania or the dot-com bubble), but the problem has been particularly acute for AI. The pattern has occurred many times:

- 1966: the failure of machine translation,
- 1970: the abandonment of connectionism,
- 1971–75: DARPA's frustration with the Speech Understanding Research program at Carnegie Mellon University,
- 1973: the large decrease in AI research in the United Kingdom in response to the Lighthill Report,
- 1973–74: DARPA's cutbacks to academic AI research in general,
- 1987: the collapse of the Lisp machine market,
- 1988: the cancellation of new spending on AI by the Strategic Computing Initiative,
- 1993: expert systems slowly reaching the bottom,
- 1990s: the quiet disappearance of the fifth-generation computer project's original goals,
- and the generally bad reputation AI has had since.

The worst times for AI were 1974–80 and 1987–93. Sometimes one or the other of these periods (or some part of them) is referred to as "the" AI winter.[2]

http://en.wikipedia.org/wiki/AI_Winter, accessed 7/3/2009

Promising research directions

- Emphasis on cognitive science
  - How does this technology improve human performance?
- Comparative effectiveness research
  - Just like the eye doctor: Better 1? Better 2?
  - Emphasized recently by federal government
- Natural language processing
  - Clinical knowledge is in free text, not billing data
    - ePatient Dave, mammogram = breast cancer diagnosis
- Outcome-based informatics research
  - Does system improve outcome?

Summary

- To realize promise of e-Science requires unprecedented collaboration
- Different disciplines
  - philosophy, computer science, psychology, biomedicine...
  - Different cultures, very different values
- Challenge and opportunity
  - Fundamentally different perspectives on old stubborn problems