What Is Computing Today?
Deconstruction

Gregory D. Hager
Professor and Chair

JOHNS HOPKINS UNIVERSITY
WHITING SCHOOL OF ENGINEERING
Questions

- How does stuff work and how did it come to be?

- What are the basic research areas of CS that impacted it?

- What are commercial needs drove it?

- How has that stuff changed with time?
Some Background Information

- IT is around 1T$* of US economy (itself 18T$ GDP)
  - Apple Inc. (Nasdaq: AAPL), (560B/30B)
  - Exxon Mobil Corporation (NYSE: XOM),
  - Google Inc (Nasdaq: GOOG), (358B /12B)
  - Microsoft Corporation (Nasdaq: MSFT), (344B/20B)
  - Berkshire Hathaway Inc. (NYSE: BRK.B),
  - Wal-Mart Stores, Inc. (NYSE: WMT),
  - Johnson & Johnson (NYSE: JNJ),
  - General Electric Company (NYSE: GE),
  - Chevron Corporation (NYSE: CVX)
  - Wells Fargo & Co (NYSE: WFC)

Deconstructing a Search Query

Challenges in Building Large-Scale Information Retrieval Systems

Jeff Dean
Google Fellow
jeff@google.com

Credits to material used from
static.googleusercontent.com/media/research.google.com/en/us/people/jeff/
WSDM09-keynote.pdf
The Origins of PageRank

- Stanford WebBase project (1996 - 1999)
  http://dbpubs.stanford.edu:8091/diglib/

- funded by NSF through DLI
  http://www.dli2.nsf.gov/dlige/

“The Initiative's focus is to dramatically advance the means to collect, store, and organize information in digital forms, and make it available for searching, retrieval, and processing via communication networks -- all in user-friendly ways.” quote from the DLI website

Some Other Research Ideas

- Cache (M. Wilkes, 1965, Cambridge)
- The internet (Cerf, Kahn, 1969, ARPA)
- The Web and HTML (T. Berners-Lee, 1989, CERN)
- PageRank (Brin, Page, Motwani, Winograd, Stanford, 1997)
- SIFT Image Features (Lowe, UBC, 1999)
- Hadoop (Cutting, Cafarella, Yahoo/UW, 2005)
- Deep Learning (Hinton+others, Toronto+others, ??)

† GPUs.
What Is a Search Query?

Google's answer
What Is a Search Query?

A web search query is a request that a user enters into a search engine to satisfy his or her information needs. Web search queries are distinctive in that they are...
What Is a Search Query?

A web search query is a query that a user enters into a web search engine to satisfy his or her information needs. Web search queries are distinctive in that they are...

What is a query - Answers.com

Queries allow you to decide what fields or ... A web query is simply the process of searching for information on the internet using search engines like ...

What is a search query - Wikipedia, the free encyclopedia

A web search query is a query that a user enters into a web search engine to satisfy his or her information needs. Web search queries are distinctive in that they are ...

What is a search query - Yahoo Answers Results

When searching the internet, what is a query? 2 answers

It's a request for some asset on a remote server. Basically when you click a link, or type a word(s) into the search bar on the browser, it then compiles a 'packet' which 'says' where it comes from (the 'source'), where it’s going (the...
It All Starts With a Spider

http://programming4.us/website/15366.aspx
Inverted Index

For example, let’s say we have two documents, each with a `content` field containing:

1. “The quick brown fox jumped over the lazy dog”
2. “Quick brown foxes leap over lazy dogs in summer”

To create an inverted index, we first split the `content` field of each document into separate words (which we call `terms` or `tokens`), create a sorted list of all the unique terms, then list in which document each term appears. The result looks something like this:

<table>
<thead>
<tr>
<th>Term</th>
<th>Doc_1</th>
<th>Doc_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quick</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>The</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>brown</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>dog</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>dogs</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>fox</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>foxes</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>in</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>jumped</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>lazy</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>leap</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>over</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>quick</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>summer</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>the</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
Then We Need Horsepower

- Given search terms e.g. dog, cat
- Return pair <docid, score>
- Score is the secret sauce for ranking docs
- Doc servers return pre-formatted snippets plus doc address

Image from Dean, Google 1997
Some Questions

- Where are the bottlenecks?
- What is missing?
Adding Speed and Income

Image from Dean, Google 1999
What Makes an Effective Ad?

- Search?

- Ad?
More Scale?

Ad System

Frontend Web Server

query

Cache servers

Cache Servers

Doc Servers

bal

bal

bal

bal

Index servers

Shard 0

I_0

I_1

I_2

I_3

I_4

I_5

I_{12}

I_{13}

I_{14}

Shard 1

I_0

I_1

I_2

I_3

I_4

I_5

I_{12}

I_{13}

I_{14}

Shard 2

I_0

I_1

I_2

I_3

I_4

I_5

I_{12}

I_{13}

I_{14}

Shard N

I_0

I_1

I_2

I_3

I_4

I_5

I_{12}

I_{13}

I_{14}

Balancers

Dean, Google 2001
A More Complete Picture

Query Serving System, 2004 edition

Repository Manager

Cache servers

Requests

Parent Servers

Leaf Servers

File Loaders

Repository Shards

File Loaders

GFS

Multi-level tree for query distribution

Leaf servers handle both index & doc requests from in-memory data structures

Coordinates index switching as new shards become available
Some Interesting Plusses and Minuses

- Queries are now fast particularly at tail
- Throughput is high

- Now depending on a very large # of machines one key machine down and that query fails or takes a long time
- What if a query kills a machine you can mow down the entire cluster
Canary Requests

- Send a request to one machine and see if it dies
- If not, go ahead
- If it does, try a couple more; if they die, give up
Google 2007 Architecture

Query

Ad System

Frontend Web Server

Super root

Indexing Service

Cache servers

Images

Local

News

Web

Video

Blogs

Books
An Aside: Visual Search

- Key breakthrough due to Lowe (SIFT, 1999)

- Second key technology: use of weak labels from the Web

- Third key technology: learning technologies that can be applied at scale
1. Feature detection and representation

- Compute SIFT descriptor
  [Lowe '99]

- Normalize patch

- Detect patches
  [Mikolajczyk and Schmid '02]
  [Mata, Chum, Urban & Pajdla, '02]
  [Sivic & Zisserman, '03]

Slide credit: Josef Sivic
2. Codewords dictionary formation
2. Codewords dictionary formation

Vector quantization
Example: each group of patches belongs to the same visual word.
3. Image representation
Object recognition results

- Caltech objects database 101 object classes
- Features:
  - SIFT detector
  - PCA-SIFT descriptor, $d=10$
- 30 training images / class
- **43% recognition rate**
  (1% chance performance)
- 0.002 seconds per match
Running in The Real World

Machines + Racks

- In-house rack design
- PC-class motherboards
- Low-end storage & networking hardware
- Linux
- + in-house software

Clusters

From Dean
Running in the Real World

Typical first year for a new cluster:

~1 network rewiring (rolling ~5% of machines down over 2-day span)
~20 rack failures (40-80 machines instantly disappear, 1-6 hours to get back)
~5 racks go wonky (40-80 machines see 50% packetloss)
~8 network maintenances (4 might cause ~30-minute random connectivity losses)
~12 router reloads (takes out DNS and external vips for a couple minutes)
~3 router failures (have to immediately pull traffic for an hour)
~dozens of minor 30-second blips for dns
~1000 individual machine failures
~thousands of hard drive failures
slow disks, bad memory, misconfigured machines, flaky machines, etc.

Dean
Time for a Reality Check

- Suppose you have a bug that is exercised once in a million queries
- How often will that bug be exercised in a day at Google?
  - 3.5 billion queries/day -> 40k/second -> every 25 seconds,
A Slight Digression $^g$ SAT and Program Verification

$^g$ Simple problem: is a boolean formula satisfiable?

$^g$ A & B  === yes A=1, B=1
$^g$ A & $\neg$A === no!

$^g$ Original NP complete problem (Cook 1971)

$^g$ So What?
A Slight Digression -- SAT

SAT/SMT Solver Research Story
A 1000x Improvement

- Solver-based programming languages
- Compiler optimization using solvers
- Solver-based debugging
- Solver-based static analysis
- Formal verification
- Distributed computing
- Bio & Computation

- Console Testing
- Program Analysis
- Equivalence Checking

1,000,000 Constraints
100,000 Constraints
10,000 Constraints
1,000 Constraints


Courtesy Vijay Ganesh
Add Some Horsepower

A large fraction of windows bugs now found by program verifications

Z3_System.pdf
Time for a Reality Check

- Suppose you have a bug that is exercised once in a million queries
- How often will that bug be exercised in a day at Google?
  - 3.5 billion queries/day -> 40k/second -> **every 25 seconds**, 
- Even more complicated because queries are highly parallel!
A Paradigm Is Born

- Every problem has to deal with all of the possible hardware and software exceptions. Lots of work!

- BUT, many of the underlying computations are embarrassingly parallel.
  - Send a query to server (e.g. "do you have this term")
  - Aggregate the results

- The idea of Map-Reduce

11/8/15
CS@JHU M&Ms 2014, GD Hager
Map Reduce

- User writes two main functions
  - Map -> the work each worker has to do e.g. find docs for an index term
  - Reduce -> the work to combine the results e.g. find the top n queries based on ranking

- System handles
  - Distribution, load balancing, communication, checkpointing

- Apache Hadoop a common (open source) system for Map-Reduce
### Computing at Scale

<table>
<thead>
<tr>
<th></th>
<th>Aug, ‘04</th>
<th>Mar, ‘06</th>
<th>Sep, ’07</th>
<th>May, ’10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of jobs</td>
<td>29K</td>
<td>171K</td>
<td>2,217K</td>
<td>4,474K</td>
</tr>
<tr>
<td>Average completion time (secs)</td>
<td>634</td>
<td>874</td>
<td>395</td>
<td>748</td>
</tr>
<tr>
<td>Machine years used</td>
<td>217</td>
<td>2,002</td>
<td>11,081</td>
<td>39,121</td>
</tr>
<tr>
<td>Input data read (TB)</td>
<td>3,288</td>
<td>52,254</td>
<td>403,152</td>
<td>946,460</td>
</tr>
<tr>
<td>Intermediate data (TB)</td>
<td>758</td>
<td>6,743</td>
<td>34,774</td>
<td>132,960</td>
</tr>
<tr>
<td>Output data written (TB)</td>
<td>193</td>
<td>2,970</td>
<td>14,018</td>
<td>45,720</td>
</tr>
<tr>
<td>Average worker machines</td>
<td>157</td>
<td>268</td>
<td>394</td>
<td>368</td>
</tr>
</tbody>
</table>

Dean: Map Reduce Statistics
Some Lessons

- Reality is a harsh taskmaster many of the best ideas are forged from real problems.

- It’s usually not a single idea borrow from the best!

- It’s hard to trace the impact of ideas to fruition at best we can do an anecdotal approximation; don’t be fooled by an overly simplistic view!

- There are few truly failed ideas, just failed applications thereof persevere!
Where is Computing Going?

Gregory D. Hager
Professor and Chair
What Do Computers Do Well

- Office work: accounting, wordprocessing

- Simulation: science, gaming

- Automation: manufacturing, embedded systems
What is Hard?
Driving: A Case Study
Automated Driving 1986

Navlab 1 (CMU)
- 5 racks of computing equipment
- Warp supercomputer
- Vision, laser scanner

By late 80's software systems could drive at 20 m.p.h.

1986: Dickmanns demonstrates 60 mph driving using simple vision-based control approach
Automated Driving 1996

Navlab 5: 1990 Pontiac Trans Sport
- Portable Advanced Navigation Support
- Sparc LX with a color video digitizer
- Differential GPS, fiber optic rate gyro
- Position estimation, vehicle control, and safety monitoring.
- Powered from cigarette lighter

No Hands Across America
2797/2849 miles (98.2%)
automated driving

(Pomerleau et al. CMU)
The 2004 DARPA Grand Challenge

- 150 mile course through Mojave Desert
- 15 entrants who made it to the final
  [https://www.youtube.com/watch?v=FaBJ5sPPmcl](https://www.youtube.com/watch?v=FaBJ5sPPmcl)
- After a few hours, all failed
- Sandstorm (CMU) made it 7.4 miles
Grand Challenge 2005: 195 Teams

195 teams from 36 states

Courtesy Thrun
### Status Board

Final Results as of 10/9/2005

<table>
<thead>
<tr>
<th>ID</th>
<th>Team</th>
<th>Time</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Stanford Racing Team</td>
<td>6h 53m</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Red Team</td>
<td>7h 4m</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Red Team Too</td>
<td>7h 14m</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Gray Team</td>
<td>7h 30m</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Team TerraMax</td>
<td>12h 51m</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Team ENSCO</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Axion Racing</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>Virginia Tech Grand Challenge</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Virginia Tech Team Rocky</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Desert Buckeyes</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Team DAD (Digital Auto Drive)</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Insight Racing</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>MojaveBot</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>The Golem Group / UCLA</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Team CajunBot</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>SciAutonics/Auburn Engineer</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Intelligent Vehicle Safety Tec</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>CIMAR</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>Princeton University</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Team Carnel</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Team Caltech</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Monster/Moto</td>
<td>DNF</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>The MITRE Meteorites</td>
<td>DNF</td>
<td></td>
</tr>
</tbody>
</table>
The Winners

■ Stanley (Stanford) 6:54
■ Sandstorm (CMU) 7:05
■ Highlander (CMU) 7:14
Automated Driving 2006

Stanford DARPA Grand Challenge Vehicle (Thrun)

(Popular Mechanics Jan ‘06)
Stanford Racing Team

Touareg interface

Laser mapper

Wireless E-Stop

Top level control

Laser 2 interface

Laser 3 interface

Laser 4 interface

Laser 5 interface

Camera interface

Radar interface

UKF Pose estimation

Wheel velocity

GPS position

GPS compass

IMU interface

Surface assessment

Laser 1 interface

Laser 2 interface

Laser 3 interface

Laser 4 interface

Laser 5 interface

Camera interface

Radar interface

UKF Pose estimation

Wheel velocity

GPS position

GPS compass

IMU interface

Surface assessment

Laser 1 interface

Laser 2 interface

Laser 3 interface

Laser 4 interface

Laser 5 interface

Camera interface

Radar interface

UKF Pose estimation

Wheel velocity

GPS position

GPS compass

IMU interface

Surface assessment

Laser 1 interface

Laser 2 interface

Laser 3 interface

Laser 4 interface

Laser 5 interface

Camera interface

Radar interface

UKF Pose estimation

Wheel velocity

GPS position

GPS compass

IMU interface

Surface assessment
The Next Challenges

- Darpa Urban Challenge: https://www.youtube.com/watch?v=0wJAANgG-Vg
Sensors
Lasers, radars and cameras detect objects in all directions

Interior
Designed for riding, not for driving

Electric batteries
To power the vehicle

Rounded shape
Maximizes sensor field of view

Computer
Designed specifically for self-driving

Back-up systems
For steering, braking, computing and more
Toyota invests $50 million in artificial intelligence for smarter cars

Gabe Nelson
Automotive News
September 4, 2015 - 1:00 pm ET -- UPDATED: 9/4/15 1:16 pm ET --
adds comments

PALO ALTO, Calif. -- Seeking to make cars better at avoiding crashes, Toyota Motor Corp. will spend $50 million over five years to set up joint research centers at the Massachusetts Institute of Technology and Stanford University, the Japanese automaker said today during an event near Stanford’s campus here.
DARPA Robotics Challenge

- 1. Drive a utility vehicle at the site.
- 2. Travel dismounted across rubble.
- 3. Remove debris blocking an entryway.
- 4. Open a door and enter a building.
- 5. Climb an industrial ladder and traverse an industrial walkway.
- 6. Use a tool to break through a concrete panel.
- 7. Locate and close a valve near a leaking pipe.
- 8. Connect a fire hose to a standpipe and turn on a valve.
DARPA Robotics Challenge
From Data to Information

Gregory D. Hager
Evolution of NIT 1991-2015

- Health
- Smart infrastructure
- Robotics
- Scientific Discovery
- Transportation
- Security
- Networking
- Data
- Computing
- Smart manufacturing
- Physical World
- People
- Societal Computing
- Privacy
- Manufacturing
- Robotics
- Physical World
- Data
- Computing
- Societal Computing
- Privacy
- People
NIT R&D is essential to many National Priorities

**Cybersecurity** - research on cybersecurity by design, defense against attack, systems resilience, implementation support, better and faster attack attribution methods.

**Health** - research on treatment and outcomes, disease and wellness, mobile and biometric technologies for monitoring and care, actionable decision support, regulatory compliance.
What Drives the Value of the Web?

AWE-INSPIRING VISTAS

Wine Tasting in Yosemite!
Join us for Vintners' Holidays this fall at The Ahwahnee in Yosemite Valley.

PLAN YOUR GETAWAY
The Rest of the World

Mobile Data Traffic by Quarter

Credit: akamai
Data at Scale Has its Benefits
How Do We Learn From Data?

- Unsupervised Learning (data structuring)

- Supervised Learning

- Reinforcement Learning

Unsupervised An Example
Supervised An Example
A Real Example

![Image of faces with varying lighting conditions]
Task versus Subtask Surgical Skill Evaluation of Robotic Minimally Invasive Surgery

Carol E. Reiley and Gregory D. Hager, Proc. MICCAI 2009

Distances to Expertise

Experts

Intermediates

Novices
Supervised An Example
String Similarity

$$Score(T_y, D) = \frac{1}{d} \sum_{C_i \in D \cap T_y} \left( \log(P_i) + w_1 \cdot \log(|C_i|) + w_2 \cdot \log(A_i) \right)$$

$$A_i = \frac{1}{1 + \sum_{J \in O} |Y - J|}$$

| String | # | P   | |C| | A   | %     |
|-------|---|-----|---|---|-----|------|
| dftfgg| 9 | 0.000103 | 6 | 0.001447 | 0.002275 |
| fdfggg| 3 | 0.000034 | 6 | 0.004464 | 0.002260 |
| gffge | 3 | 0.000034 | 6 | 0.008264 | 0.001566 |
| ddggg | 3 | 0.000034 | 6 | 0.003367 | 0.002578 |
| fgeef | 2 | 0.000023 | 6 | 0.008850 | 0.001952 |
| ffgge | 2 | 0.000023 | 6 | 0.009304 | 0.001890 |
| gffge | 2 | 0.000023 | 6 | 0.012821 | 0.001534 |
| dgggg | 2 | 0.000023 | 6 | 0.004348 | 0.002752 |
| fddgg | 2 | 0.000023 | 6 | 0.004831 | 0.002634 |
| dgggg | 1 | 0.000011 | 6 | 0.034483 | 0.001211 |
| ffggg | 1 | 0.000011 | 6 | 0.015385 | 0.002120 |
| fffgg | 1 | 0.000011 | 6 | 0.019608 | 0.001847 |
| fedff | 1 | 0.000011 | 6 | 0.012821 | 0.002325 |
| fddgg | 1 | 0.000011 | 6 | 0.011364 | 0.002461 |

Results at the Gesture Level

Simultaneous Gesture and Skill Classification

P(Skill, Gesture)

94.98%  70.91%

Chance = 11%
Deep Networks

Fig. 2. Automatic generation of text captions for images with deep networks. A convolutional neural network is trained to interpret images, and its output is then used by a recurrent neural network trained to generate a text caption (top). The sequence at the bottom shows the word-by-word focus of the network on different parts of input image while it generates the caption word-by-word. [Adapted with permission from (30)]

Reinforcement Learning

Goal: Find a policy (choice of actions) that maximizes reward
Reinforcement Learning
Some of the Problems with Data

See 1 citation found by title matching your search:


Identifying personal genomes by surname inference.
Gymrek M¹, McGuire AL, Golan D, Halperin E, Erlich Y.

Author information

Abstract
Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.
Robust De-anonymization of Large Datasets
(How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin
February 5, 2008

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary’s background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world’s largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber’s record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.
A $3 Trillion Challenge to Computational Scientists: Transforming Healthcare Delivery

IEEE Intelligent Systems,
Computing and Societal Needs

Gregory D. Hager
The Broader Impact of Computing

- Health
- Smart infrastructure
- Robotics
- Scientific Discovery
- Transportation
- Security
- Privacy
- Manufacturing
- Networking
- Data
- Physical World
- People
- Societal Computing
- Smart infrastructure
Medical vs. Healthcare

- Medicine
  - the science or practice of the diagnosis, treatment, and prevention of disease.

- Healthcare
  - the maintenance and improvement of physical and mental health, especially through the provision of medical services.

- Healthcare system
  - the organization of people, institutions, and resources that deliver health care services to meet the health needs of target populations.
Health Care: Some Numbers

- **Costs**
  - **Absolute expenditures** $2.6 trillion 18% GDP (2012)
  - **Relative expenditures** – 76% increase health costs in past 10 years,
  - **Potential efficiency gains** $750 billion (2009) more than 25% of the total

From *Best Care At Lower Costs: The Path to Continuously Learning Health Care in America* Institute of Medicine, 2012
Health Care: Some Numbers

- Complexity
  
  **More conditions** — e.g. 79 year old patient with 19 meds per day
  **More clinicians** — e.g. 200 other doctors treating patients of PC Doctor
  **More choices** — e.g. hundreds of diagnostic factors; dozens of treatments
  **More activities** — e.g. ICU clinicians with 180 activities per person, per day

From "Best Care At Lower Costs: The Path to Continuously Learning Health Care in America" Institute of Medicine, 2012
Health Care: Some Numbers

- **Quality**
  - **Patient harm** – 1/5-1/3 of hospital patients preventably harmed during stay
  - **Recommended care** – Only about half of recommended care actually delivered.
  - **Outcome shortfalls** – If care quality matched highest statewide performance, there would have been 75,000 fewer deaths nationally.

From *Best Care At Lower Costs: The Path to Continuously Learning Health Care in America*© Institute of Medicine, 2012
Some (Digital) Responses

- HITECH Act -- $15.5B to support EHRs
- Presidential Precision Health Initiative
- Big Data initiative
- Explosion of mobile health devices/apps in the consumer market
Big Data for Health, Andreu-Perez et al. Biomedical and Health Informatics, 2015
The Engineering Opportunity Space

The Learning Health System Series
Continuous improvement and innovation in health and health care

REPORT TO THE PRESIDENT
BETTER HEALTH CARE AND LOWER COSTS:
ACCELERATING IMPROVEMENT THROUGH
SYSTEMS ENGINEERING

Directorate for Computer & Information Science & Engineering
SMART HEALTH AND WELLBEING (SHW)

CONTACTS
See program guidelines for contact information.

SYNOPSIS
What Blocks the Path to Innovation?
Geolocation

Finding your location, around you!
Three Examples

- Assessing OR performance -- Hager
- Data to diagnosis \( g \) Saria\( ^\circ \)
- Data in the wild to detection \( g \) Ferry
- IT \( g \) manage behavior -- Mynatt
Septicemia is the 11th leading cause of death in the US. Mortality and length of stay decreased with timely treatment [Kumar et al. 2011]. For every hour that antibiotic treatments were delayed, risk of mortality went up by 7.6%.

Using routinely collected data alone, at risk patients can be identified a median of 24 hours prior to shock.


Measurement in the OR

Data from 5 Johns Hopkins Hospital sites (2012-2014)

Native Septoplasty:
-- 5 Experts (60 trials)
-- 9 Novices (26 trials)
-- Multi-surgeons (14 trials)

Joint work with Masaru Ishii, MD, PhD and Lisa Ishii MD
From Data to Structure

- **Expert**
  - Start of Surgery: Blue Circle
  - End of Surgery: Red Circle
  - Stroke Height: Blue Arrow

- **Novice**
From Structure to Assessment


From Assessment to Outcomes?

<table>
<thead>
<tr>
<th>Category</th>
<th>Score Bottom Quartile</th>
<th>Score Top Quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readmission</td>
<td>6.30%</td>
<td>2.70%</td>
</tr>
<tr>
<td>Reoperation</td>
<td>3.40%</td>
<td>1.60%</td>
</tr>
<tr>
<td>Complication</td>
<td>5.20%</td>
<td>14.50%</td>
</tr>
<tr>
<td>Mortality</td>
<td>2.60%</td>
<td>0.05%</td>
</tr>
</tbody>
</table>

Michigan Bariatric Surgery Collaborative

**Samples:**
20 bariatric experts surgeons ranked by at least 10 reviewers.

**10,343 patients admitted 2006-2012**

POE/Epic: Patient Information Outcome History

Manual Subjective Assessment

Data store

Automatic Objective Assessment

Statistical Model

Visualizer

Stroke Detection

Evening

Data Collection

User Interface

Evening

live

Trainee/Faculty Log

Case ID:
Trainee’s ID:
Supervisor’s ID:
Assessment of Supervisor: Notes, comments, scores
Automatic Assessment: Scores, Feedback

Outcome/case difficulty

Case ID:
Date and Location:
Patient’s ID:
Patient Medical History:
Patient Outcome: Interval assessment

Figure 1. Overview of the computational approach and average faces of syndromes. (A) A photo is automatically analyzed to detect faces and feature points are placed using computer vision algorithms. Facial feature annotation points delineate the supra-orbital ridge (8 points), the eyes (mid points of the eyelids and eye canthi, 8 points), nose (nasion, tip, ala, subnasale and outer nares, 7 points), mouth (vermilion border lateral and vertical midpoints, 6 points) and the jaw (zygomatic mandibular border, gonion, mental protrubance and chin midpoint, 7 points). Shape and Appearance feature vectors are then extracted based on feature points and these determine the photo’s location in Clinical Face Phenotype Space (further details on feature points in Figure 1—figure supplement 1). This location is then analyzed in the context of existing points in Clinical Face Phenotype Space to extract phenotype similarities and diagnosis hypotheses (further details on Clinical Face Phenotype Space with simulation examples in Figure 1—figure supplement 2). (B) Average faces of syndromes in the database constructed using AAM models ('Materials and methods') and number of individuals which each average face represents. See online version of this manuscript for animated morphing images that show facial features differing between controls and syndromes (Figure 2).

DOI: 10.7554/eLife.02020.003

Figure 1. Continued on next page
ResearchKit for Developers

ResearchKit is an open source framework introduced by Apple that enables your iOS app to become a powerful tool for medical research. Easily create visual consent flows, real-time dynamic active tasks, and surveys using a variety of customizable modules that you can build upon and share with the community. And since ResearchKit works seamlessly with HealthKit, researchers can access even more relevant data for their studies — like daily step counts, calorie use, and heart rate.

https://www.apple.com/researchkit/
How Human-Centered Computing Research Can Help Transform Healthcare

Elizabeth D. Mynatt, Professor and Executive Director
Understanding Everyday Health

Health is personal, social, and negotiated.

Georgia Tech’s Aware Home is a unique laboratory resource for investigating home health solutions. Faculty from many disciplines (engineering, computing, design, psychology, digital media) work on interdisciplinary projects.
Human-Centered Computing and Healthcare

**Theoretical Base**
- Locus of control
- Social cognitive theory
- Identify presentation
- Health Belief Model
- Trans. Model of Change
- Social comparison theory
- Social support theory
- Sensemaking

**HCI Design Process**
- Ethnographic inquiry and informants
- Participatory design
- Field evaluation

**Interventions to Improve**
- Awareness (self and by caregivers)
- Problem Solving (“be a detective”)
- Self-efficacy and Internal LOC
- Social support and learning
- In the moment decisions & exploring new behaviors
- Healthcare Facilitation

**Technical Innovation**
- “Is today a normal day?”
- Sensing and activity recognition
- Modeling the healthcare journey

**Health Outcomes**
- Disease management
- Behavior change
- Independence
- Scale of healthcare delivery
Digital Family Portrait: Designing for Peace of Mind

Caregiver awareness
Motion sensing
Visualize 28 days of activity plus daily detail
Was today a normal day?
Interpret a social connection


NSF# 0121661- ITR/SY: The Aware Home: Sustaining the Quality of Life for an Aging Population
Tools for Diabetes Management

Mobile and web tools that empower patients to learn diabetes mgmt skills.

Patients can easily record and compare data from daily life activities

Learn to be a detective


NSF 0915934 - HCC: SMALL:Technologies for Nutrition and Diabetes Management
Nutrition and Decision Making

Just in time decision support for nutrition and diet daily choices

Nudge interfaces

Try out new behaviors

NSF 0915934 - HCC: SMALL: Technologies for Nutrition and Diabetes Management

NSF 1158766- I-Corps: SmartMenu
Adolescents & Physical Activity

Understand how social computing interventions can facilitate offline rituals & habits.

Funded by the Humana Foundation


Xu, Y., Poole, E. S., Miller, A. D., Eiriksdottir, E., Kestranek, D., Catrambone, R., & Mynatt, E. D. (2012). This is not a one-horse race: understanding player types in multiplayer pervasive health games for youth. Proceedings of the ACM 2012 conference on
Adolescents & Physical Activity

Understand how social computing interventions can facilitate offline rituals & habits.

Identity
Gaming
Offline behavior


NSF 1116801- SHB: Small: Social Tools for Everyday Adolescent Health
Adolescents & Physical Activity

Observational Learning (in contrast to competition)

Reaching the bottom tier

Changing media and gaming conventions

School as social context


NSF 1116801- SHB: Small: Social Tools for Everyday Adolescent Health
Healthcare Facilitators & Healthcare Journeys

Empowering patients (at home)
Allows patients to work through information in a private, self-paced setting.

Supporting the cancer journey
Tailor information delivery and interaction to phases of care and recovery.


Funded by GA DCH and ONC: Rome Challenge Project, Consumer Mediated Health Information Exchange #12036G-ARRA
Privacy

Gregory D. Hager
Security vs. Privacy

- Security: the practice of defending resources (computers, infrastructure, or data) from unauthorized access, use, disclosure, disruption, modification, recording or destruction.

- Privacy: The ability to anticipate and control the acquisition and use of personal data or information by a third party.
Privacy vs. Security Examples

- Sony DRM incident
- Netflix prize
- Facebook
- Gmail
  - [http://www.theguardian.com/technology/2013/aug/14/google-gmail-users-privacy-email-lawsuit](http://www.theguardian.com/technology/2013/aug/14/google-gmail-users-privacy-email-lawsuit)
- Interactive Barbie
  - [http://www.theregister.co.uk/2015/02/19/hello_barbie/](http://www.theregister.co.uk/2015/02/19/hello_barbie/)
Target Example

- Data collected from their own stores
- An observation that certain buying habits were generally predictive of pregnancy
- Push ads to target women who were likely pregnant based on buying behavior
Your Assignment

• The specific case: Should what Target did be allowed? What do we (as a society) gain? What are some examples of the dangers?


• Analyze: What are the technological problems that your solution introduces?

• Write an email summarizing the conclusions from your group and email it to us post-class. joanne@cs.jhu.edu, hager@cs.jhu.edu
Some Questions

- What are reasonable expectations for privacy with respect to online data and activity?

- How would one ensure such expectations are met?

- Are there mitigating circumstances supporting violation of privacy?

- How can one portray the value proposition of sharing vs. not?
What Computers Can’t Do (Yet?)
The Popular View of Computers
predicted that by 1967:

- A computer would be world champion in chess
- A computer would discover and prove an important new mathematical theorem
- Most theories in psychology will take the form of computer programs.
And so far †

- A computer would be world champion in chess
  - On May 11, 1997, the machine won the second six-game match against world champion Garry Kasparov, two to one, with three draws. ○

- A computer would discover and prove an important new mathematical theorem
  - The four color theorem was proven in 1976 by Kenneth Appel and Wolfgang Haken. It was the first major theorem to be proved using a computer. ○

- Most theories in psychology will take the form of computer programs.
  - Hmmm, still waiting †
Some Seemingly Simple Problems

\[ A^n + B^n = C^n \] for Integers A, B, C and \( n > 2 \)?
Dreyfus 1972, 1992
AI Will Fail Because ✧.

ências

The biological assumption
- The brain processes information in discrete operations by way of some biological equivalent of on/off switches.

The psychological assumption
- The mind can be viewed as a device operating on bits of information according to formal rules.

The epistemological assumption
- All knowledge can be formalized.

The ontological assumption
- The world consists of independent facts that can be represented by independent symbols.
And Now??

Can a computer create music?

https://www.youtube.com/watch?v=PzrcoqpnZqA
And Now??

- Can a computer create music?
  - [https://www.youtube.com/watch?v=PzrcoqpnZqA](https://www.youtube.com/watch?v=PzrcoqpnZqA)
Some Questions

瑁 Can a computer create music?

�� https://www.youtube.com/watch?v=PzrcoqpnZqA

瑁 Can a computer write an article?

This is the article generated by the LA Times algorithm: A shallow magnitude 4.7 earthquake was reported Monday morning five miles from Westwood, California, according to the U.S. Geological Survey. The temblor occurred at 6:25 a.m. Pacific time at a depth of 5.0 miles. According to the USGS, the epicenter was six miles from Beverly Hills, California, seven miles from Universal City, California, seven miles from Santa Monica, California and 348 miles from Sacramento, California. In the past ten days, there have been no earthquakes magnitude 3.0 and greater centered nearby. This information comes from the USGS Earthquake Notification Service and this post was created by an algorithm written by the author.
Some Questions

- Can a computer create music?  
  [video](https://www.youtube.com/watch?v=PzrcqpnZqA)

- Can a computer write an article?
  Friona fell 10-8 to Boys Ranch in five innings on Monday at Friona despite racking up seven hits and eight runs. Friona was led by a flawless day at the dish by Hunter Sundre, who went 2-2 against Boys Ranch pitching. Sundre singled in the third inning and tripled in the fourth inning. Friona piled up the steals, swiping eight bags in all.

[website](http://www.narrativescience.com)
Some Questions

- Can a computer create music?
  - [Link](https://www.youtube.com/watch?v=PzrcogpnZqA)

- Can a computer write an article?

- Can a computer paint?

[Image: http://www.thepaintingfool.com]
Is This Good? Bad? Inevitable?

- **Elon Musk:**

- **Paul Allen:**
Two Perspectives

Two Perspectives
What Is Our Future?  ᵇ
How Will We Shape It?

Is Computing the  ᵇ future of thought  ᵇ and discourse?

Is it the beginning of the end of Computing as  ᵇ we know it?

Is Computing creating a new ways to combine  ᵇ and create?

USC 2014, GD Hager