DEPARTMENT OF
COMPUTER SCIENCE

THE NEW AGE OF DISCOVERY

What Is Computing Today?ʊ Deconstruction

Gregory D. Hager
Professor and Chair

JOHNS HOPKINS
UNIVERSITY
WHITING SCHOOL OF ENGINEERING

Questions

- How does ʟstuff workᵐ 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 ʟstuffᵐ changed with time?

2012 NITRD update by PCAST

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)

*Atkinson, R. D., & Stewart, L. A. (2013). Just the FACTS: The Economic Benefits of Information and Communications Technologies

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/σWSDM09-keynote.pdf

The Origins of PageRank

- Stanford WebBase project (1996 - 1999)
  http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/
  http://dbpubs.stanford.edu:8091/diglib/

- funded by NSF through DLI1
  http://www.dli2.nsf.gov/dlione/

"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 DLI1 website

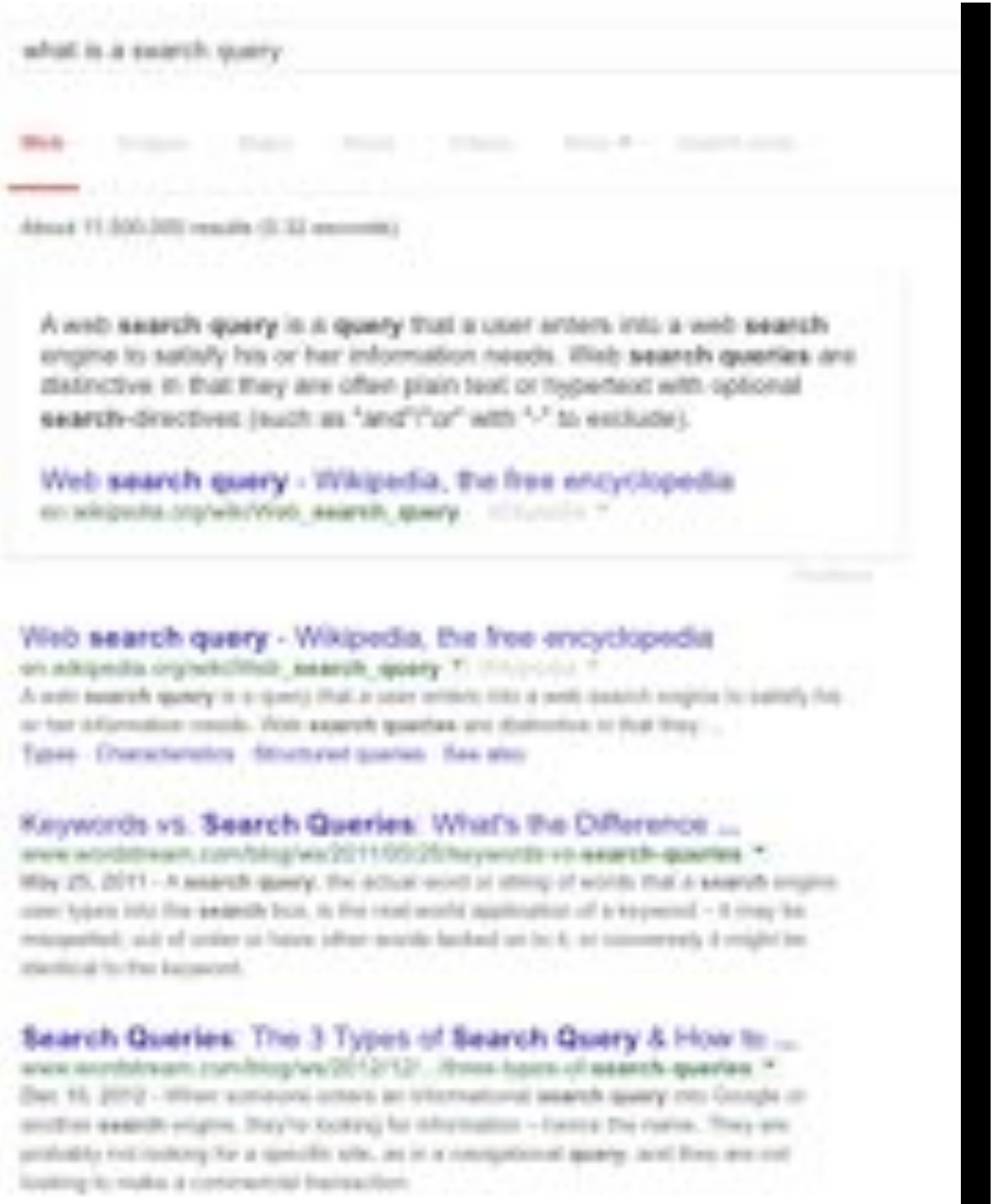
Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. (1999).
Brin, Sergey, and Lawrence Page. The anatomy of a large-scale hypertextual Web sea engine. *Computer networks and ISDN systems* 30, no. 1 (1998): 107-117.

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 ☺.

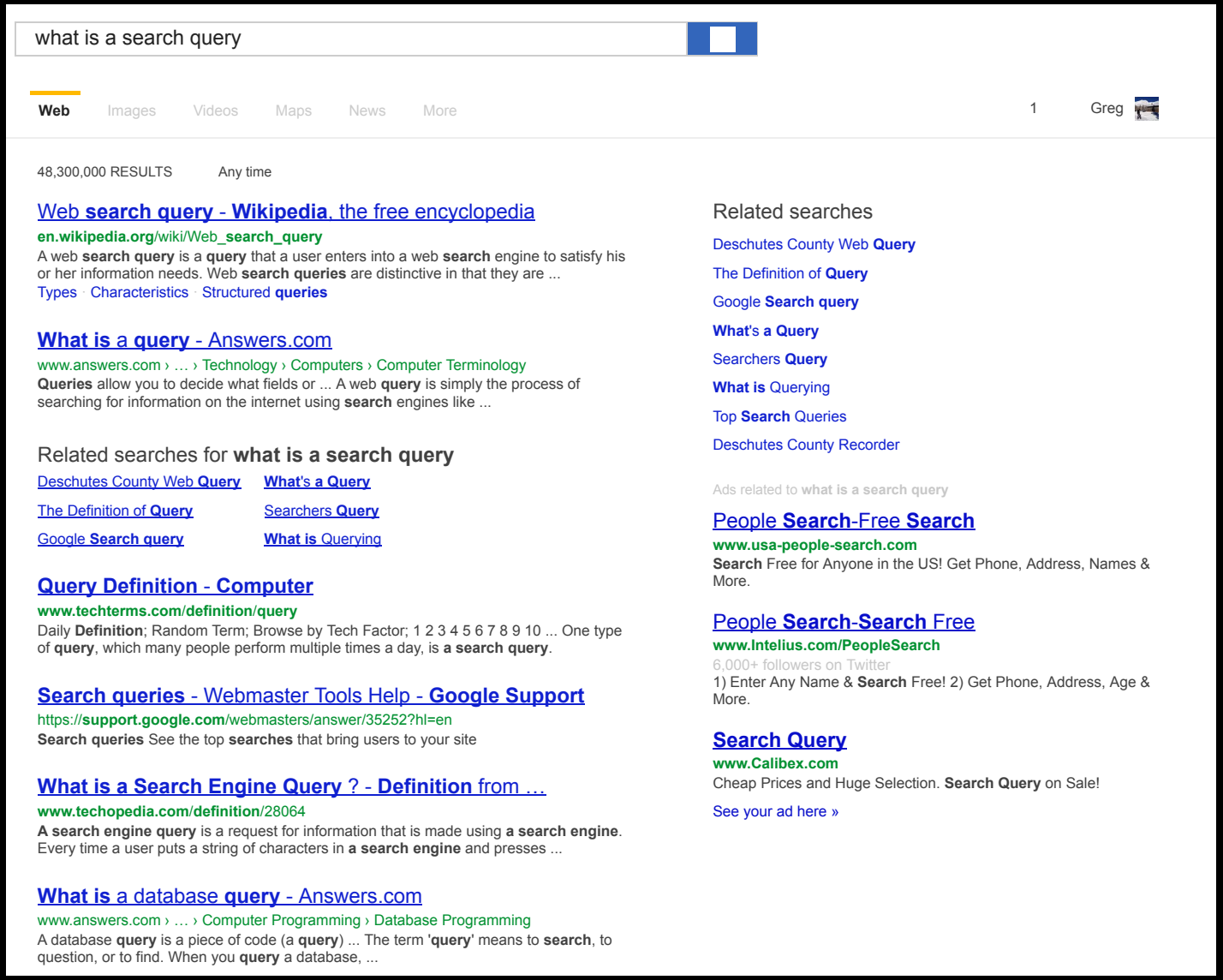
What Is ʊ a Search Query?

Google؛s answer



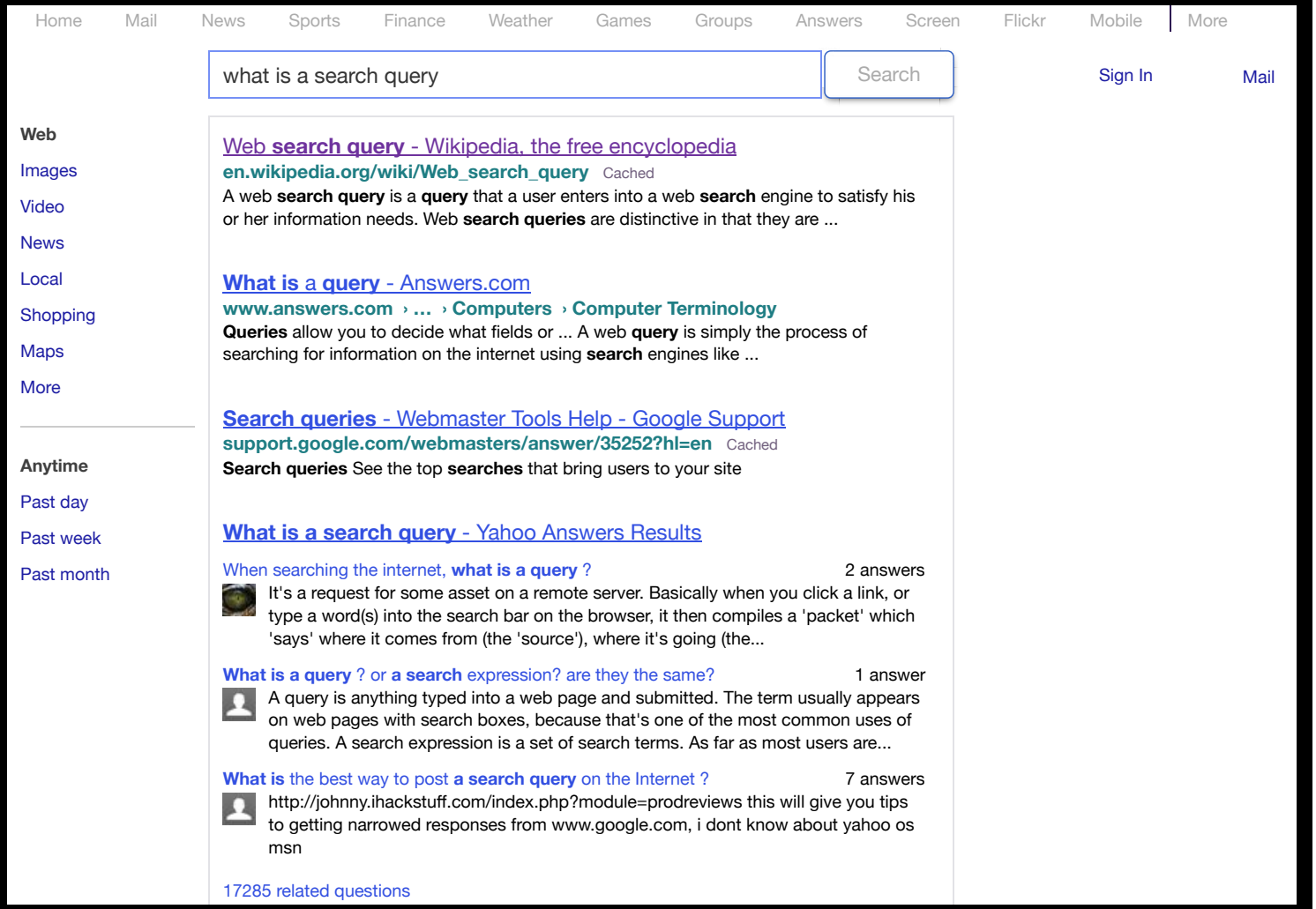
What Is a Search Query?

Bing's answer

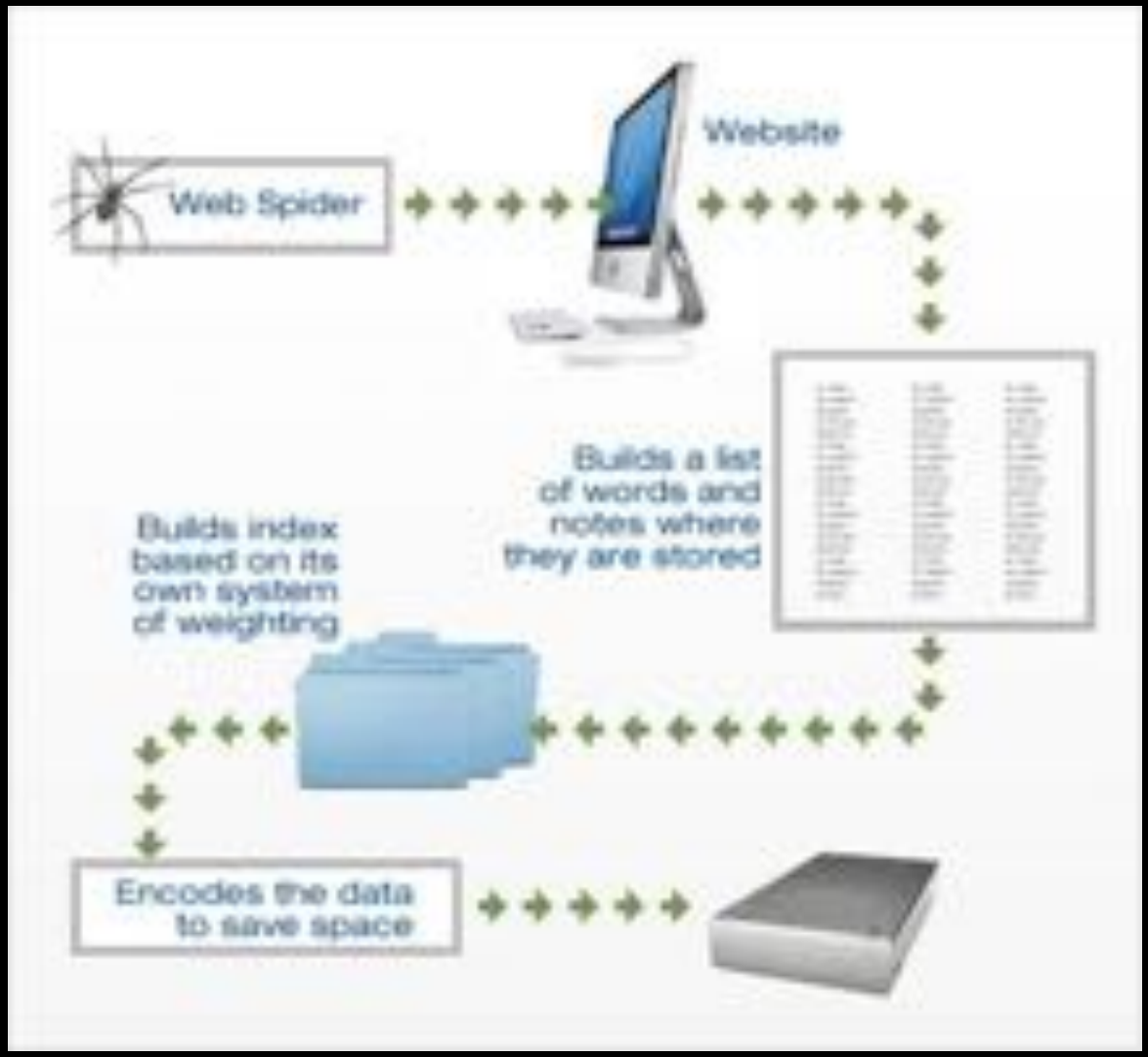


What Is a Search Query?

Yahoo's answer



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:

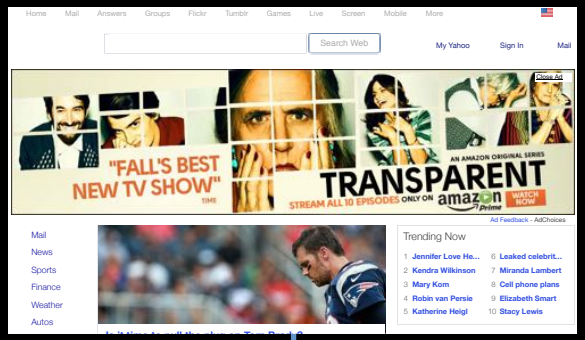
1. "The quick brown fox jumped over the lazy dog"
2. "Quick brown foxes leap over lazy dogs in summer"

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:

```
Term      Doc_1  Doc_2
-------------------------
Quick   |       |  X
The     |   X   |
brown   |   X   |  X
dog     |   X   |
dogs    |       |  X
fox     |   X   |
foxes   |       |  X
in      |       |  X
jumped  |   X   |
lazy    |   X   |  X
leap    |       |  X
over    |   X   |  X
quick   |   X   |
summer  |       |  X
the     |   X   |
------------------------
```


http://www.elasticsearch.org/guide/en/elasticsearch/guide/current/inverted-index.html

Browser to Computer to Internet to Search Engine to Cloud



Then We Need Horsepower t

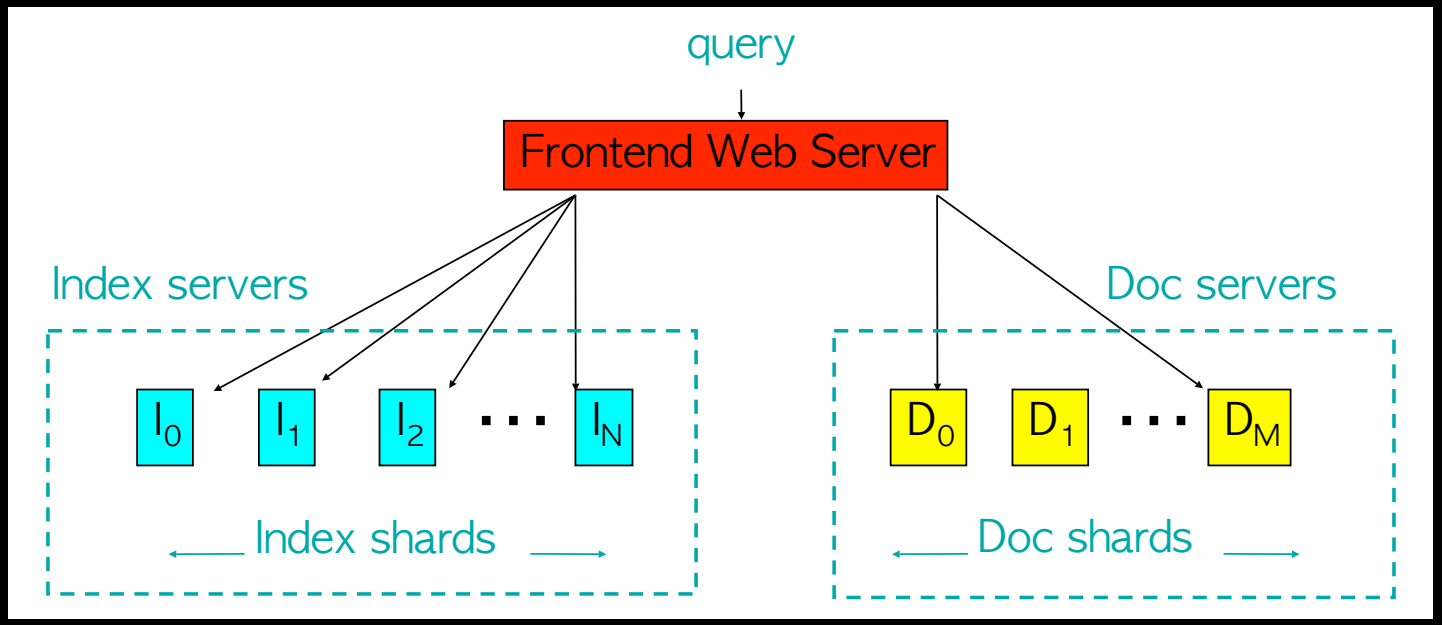


Image from Dean, Google 199

- 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

Some Questions

- Where are the bottlenecks?
- What is missing?

Adding Speed and Income

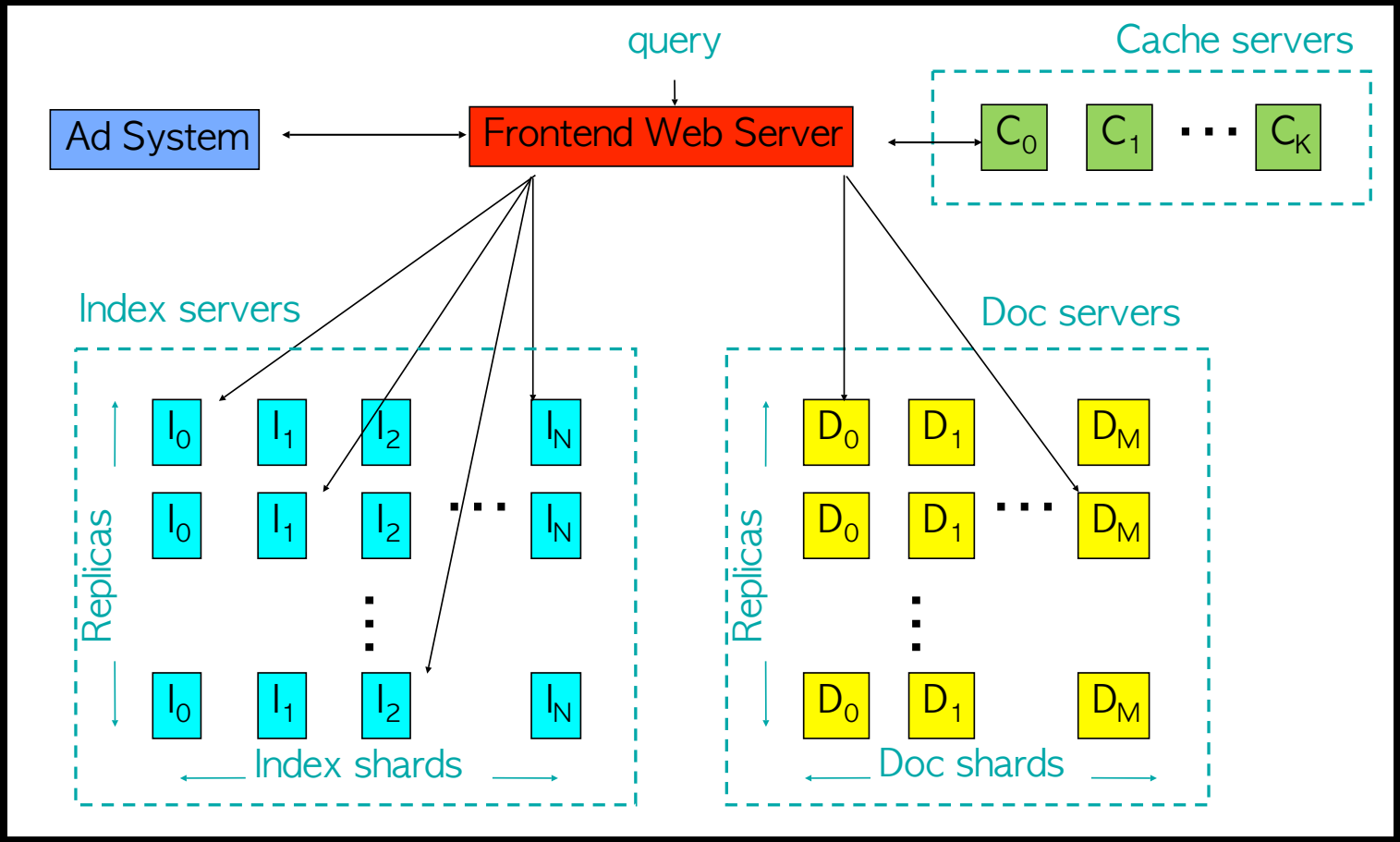
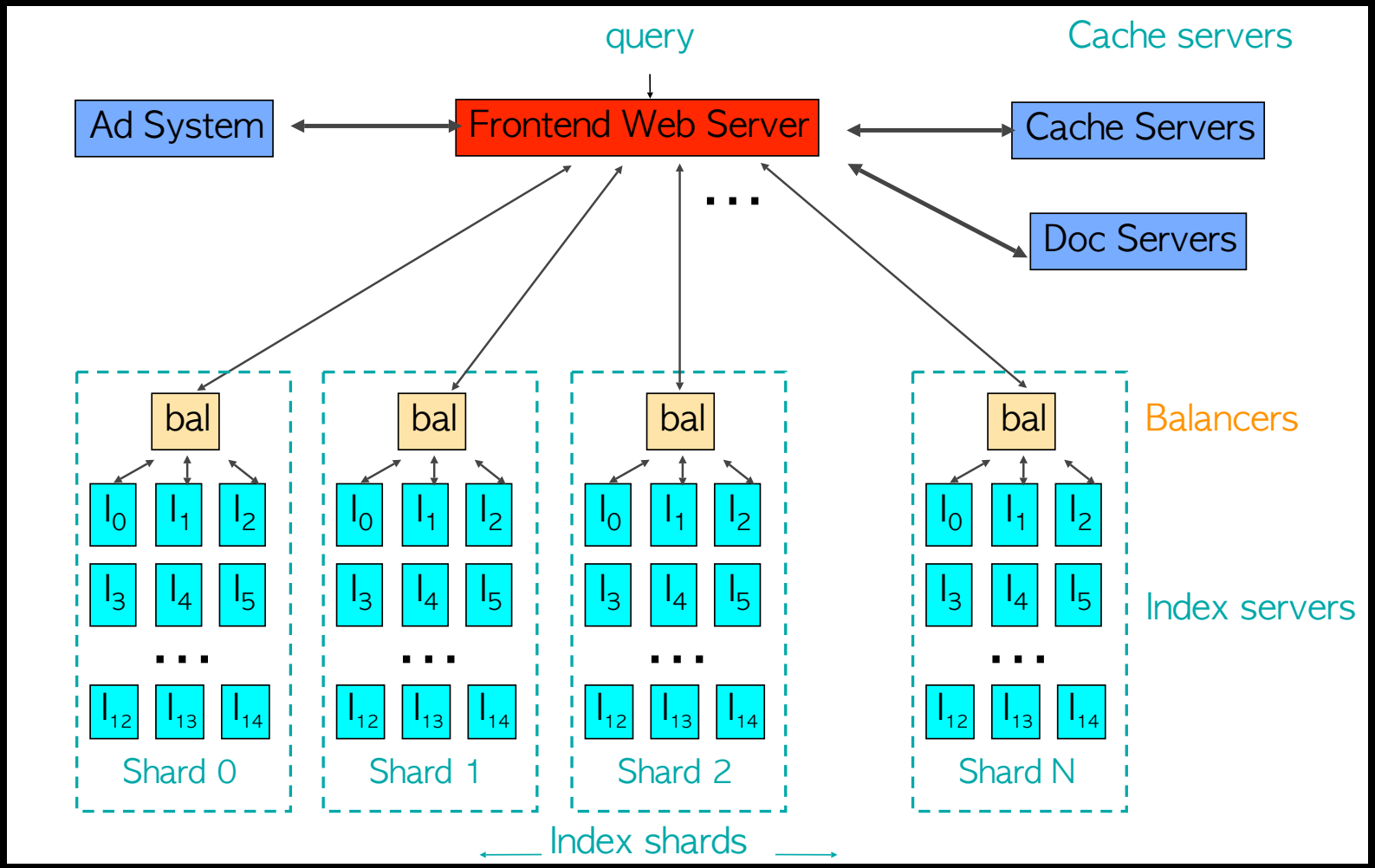


Image from Dean, Google 1999

What Makes an Effective t

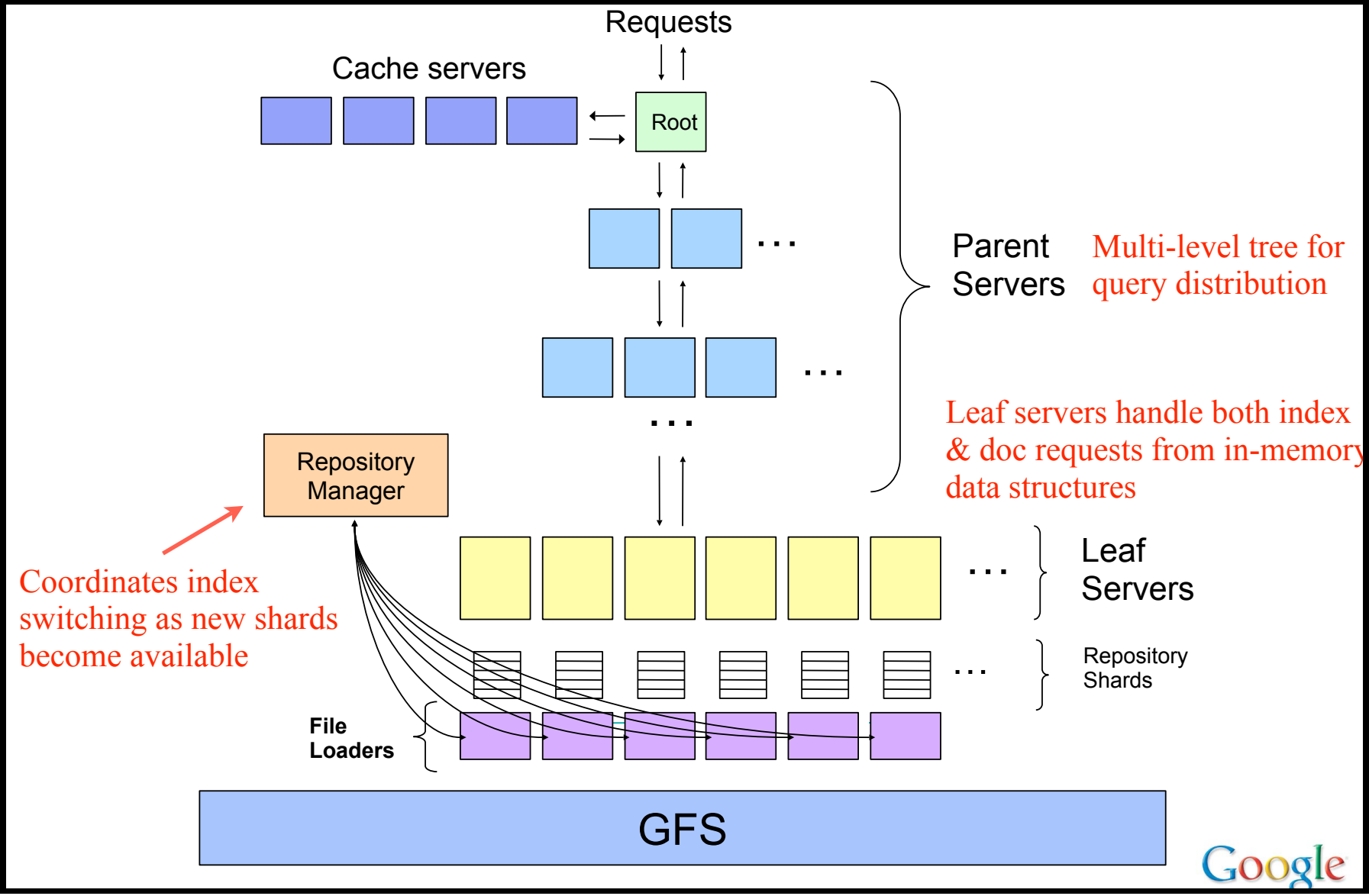
- Search?
- Ad?

More Scale?



Dean, Google 2001

A More Complete Picture



Some Interesting Plusses and Minuses

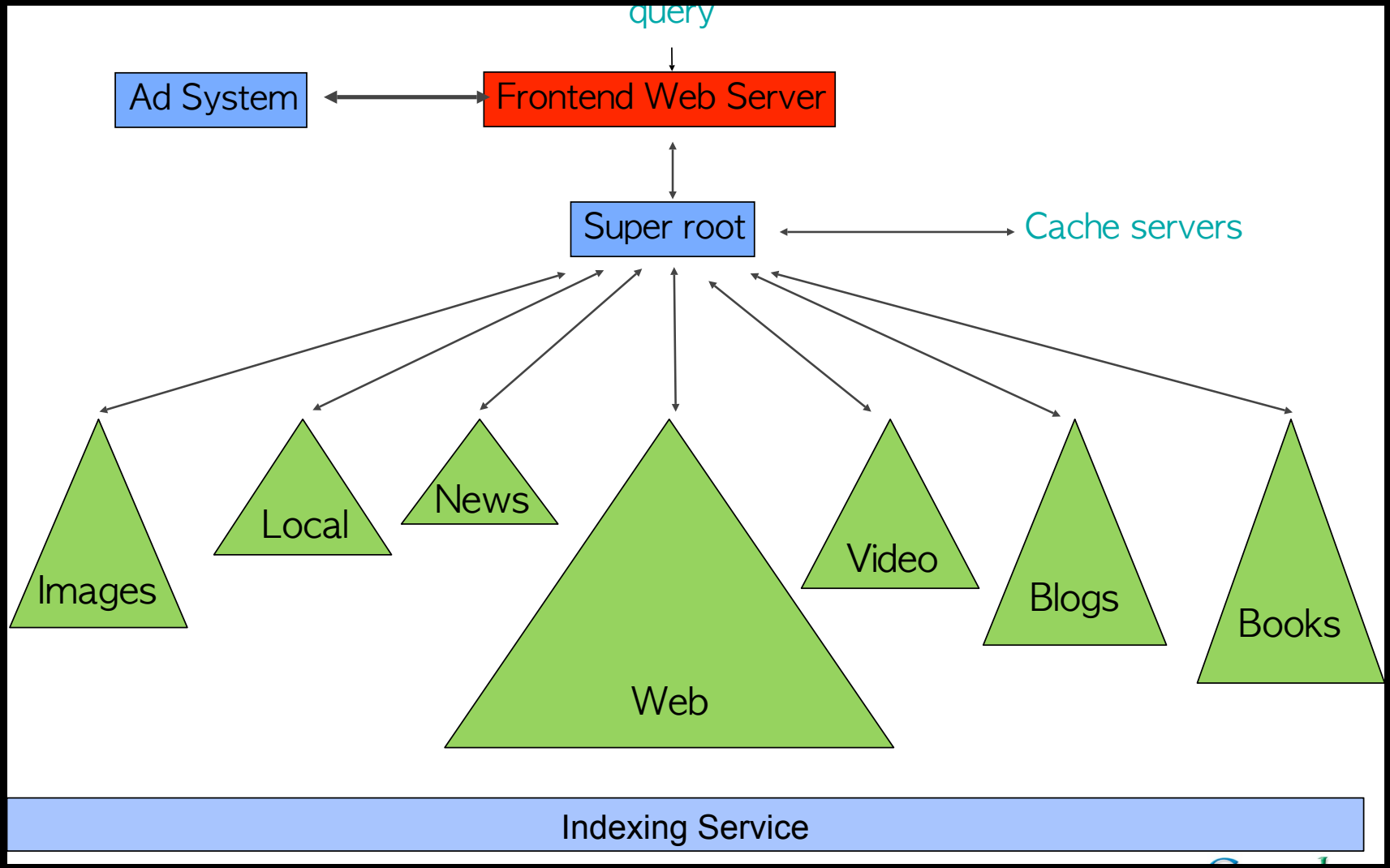
- Queries are now fast – particularly at tail
- Throughput is high

- 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



An Aside g Visual Search

- Key breakthrough due to Lowe (SIFT, 1999)ʊ
- 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

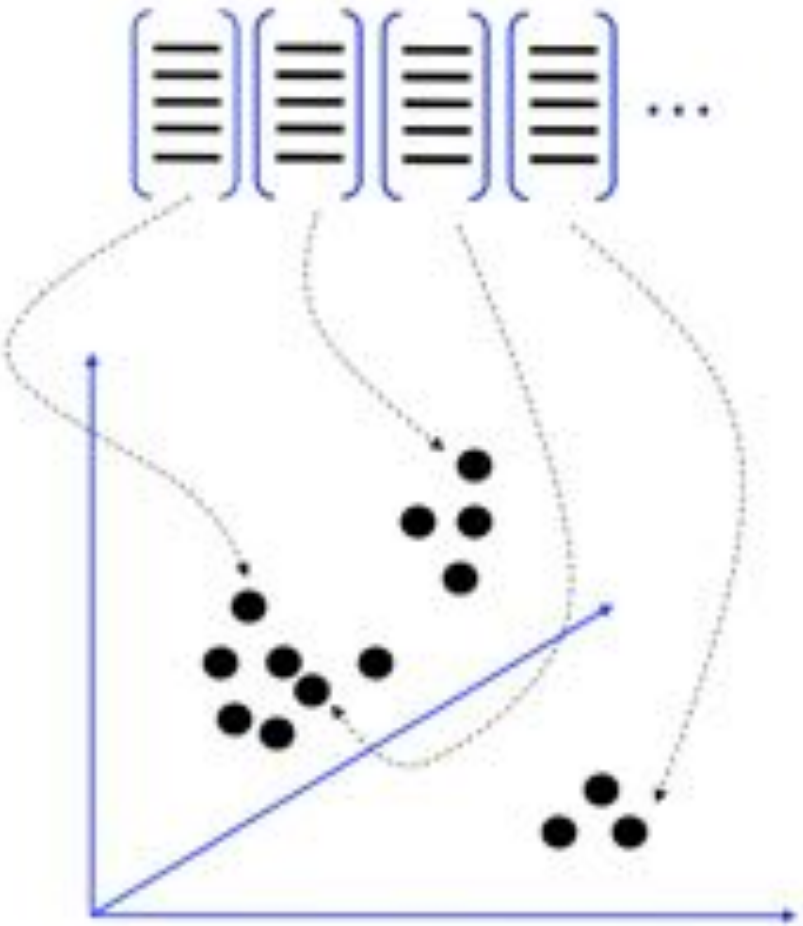
[Mikojaczyk and Schmid '02]

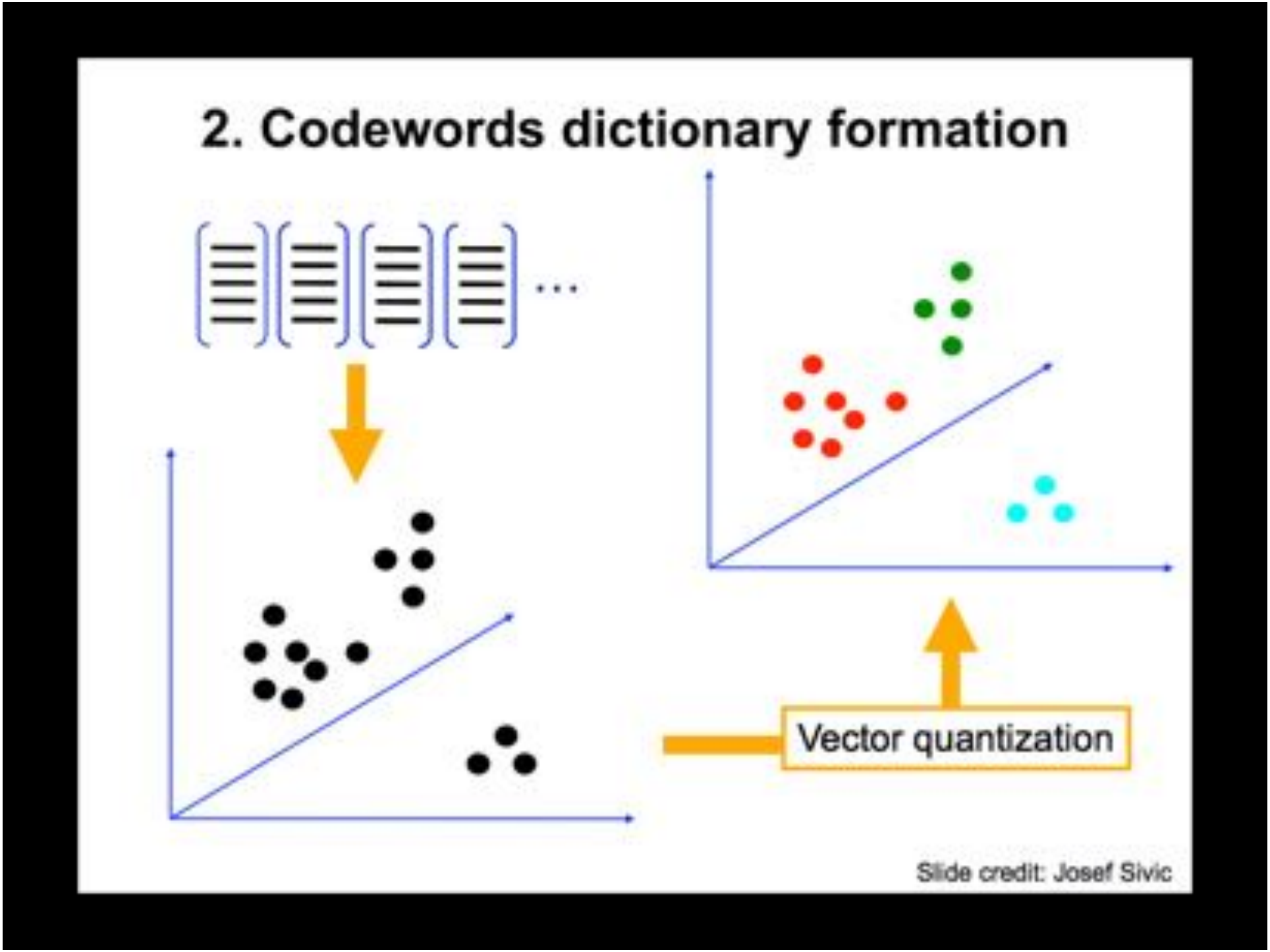
[Mata, Chum, Urban & Pajdla, '02]

[Sivic & Zisserman, '03]

Slide credit: Josef Sivic

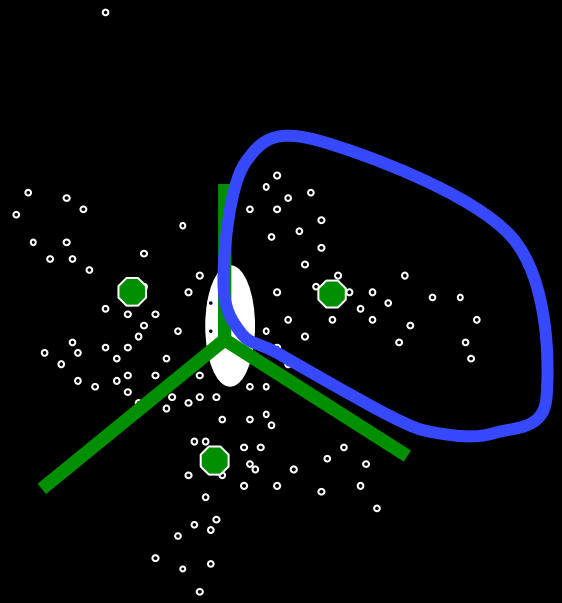
2. Codewords dictionary formation

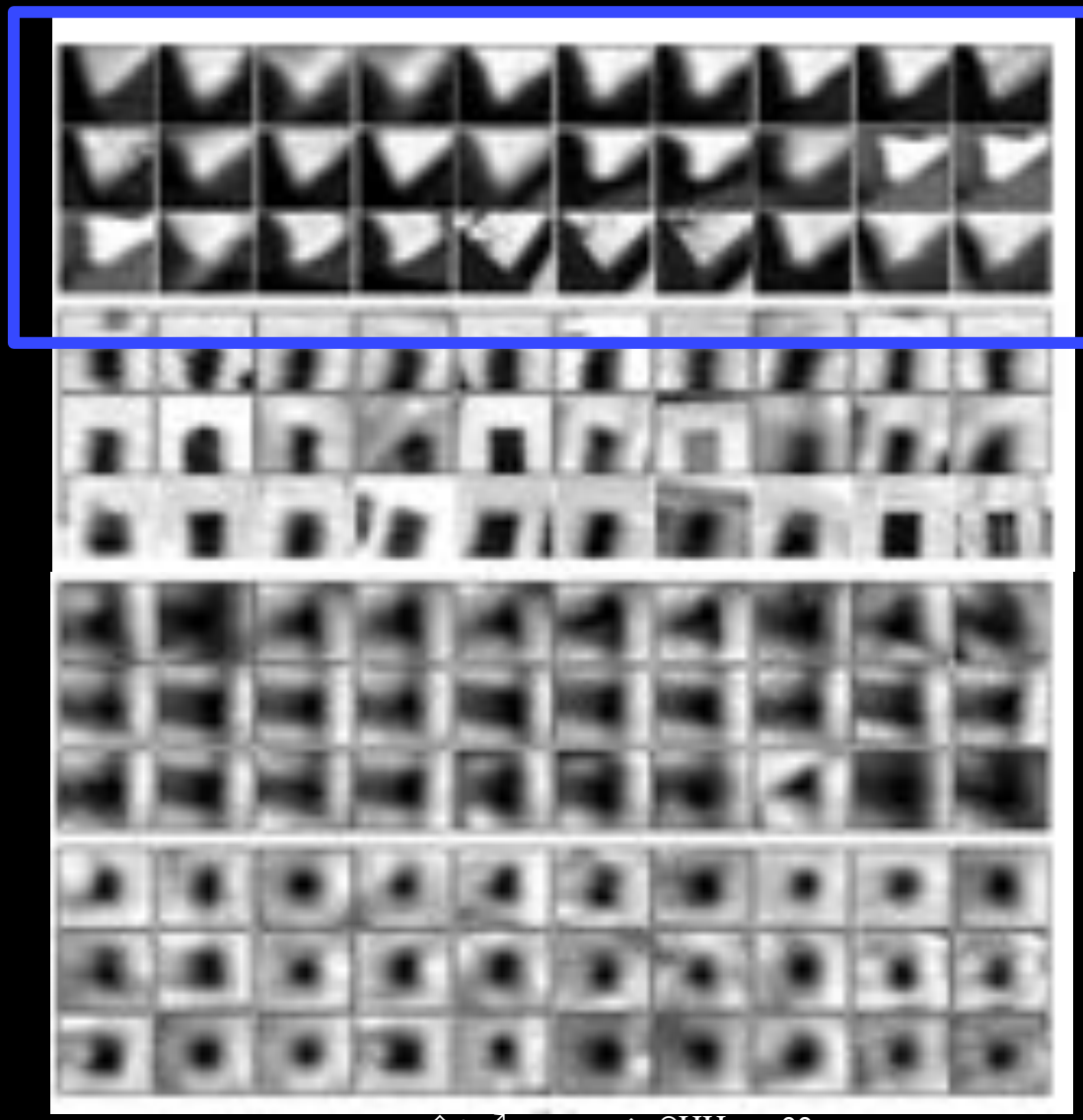



2. Codewords dictionary formation
Vector quantization
Slide credit: Josef Sivic

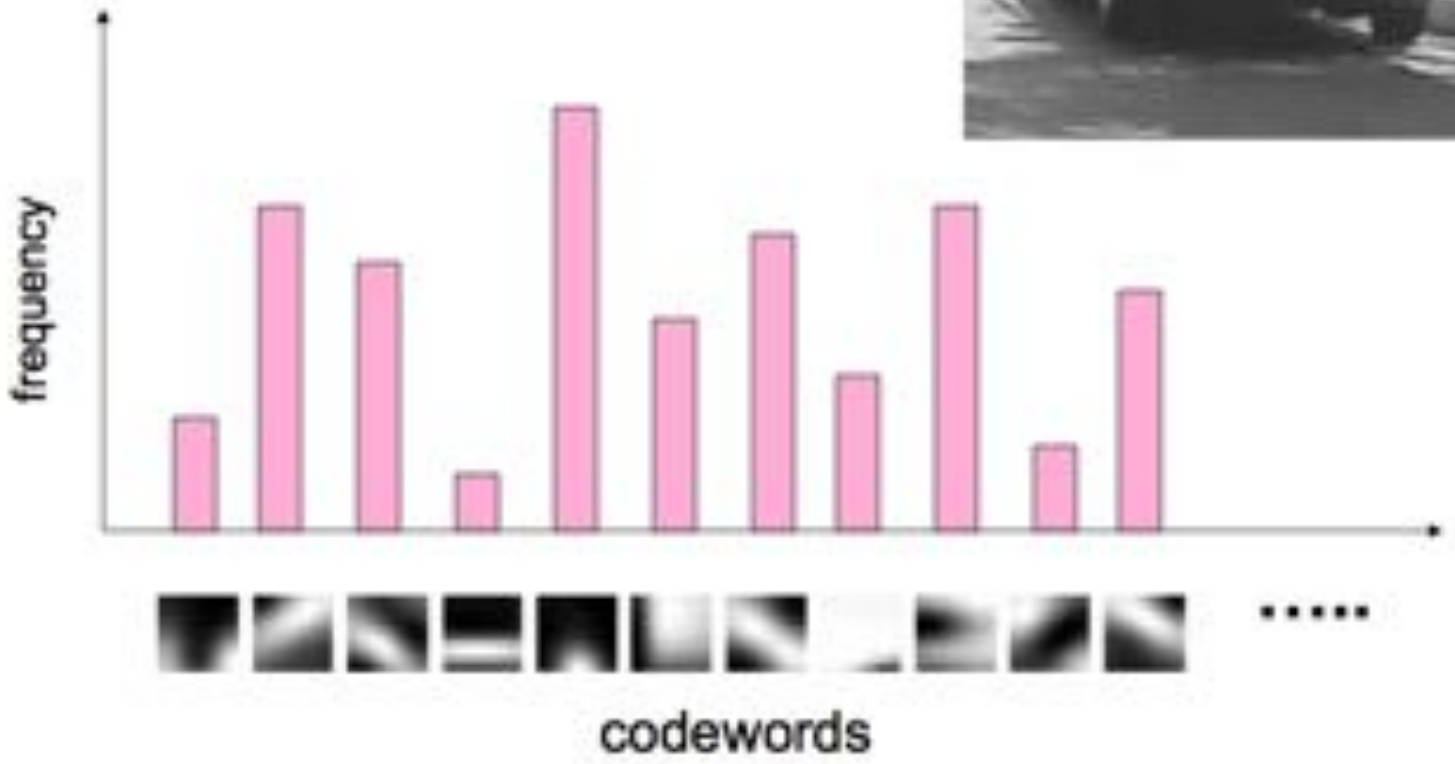
Visual Words

- Example: each group of patches belongs to the same visual word





3. Image representation



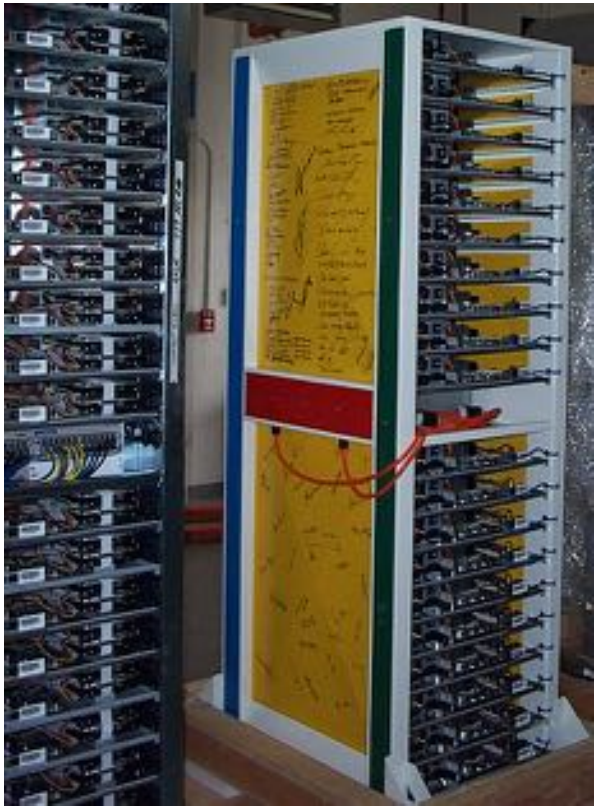
Object recognition results

- 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

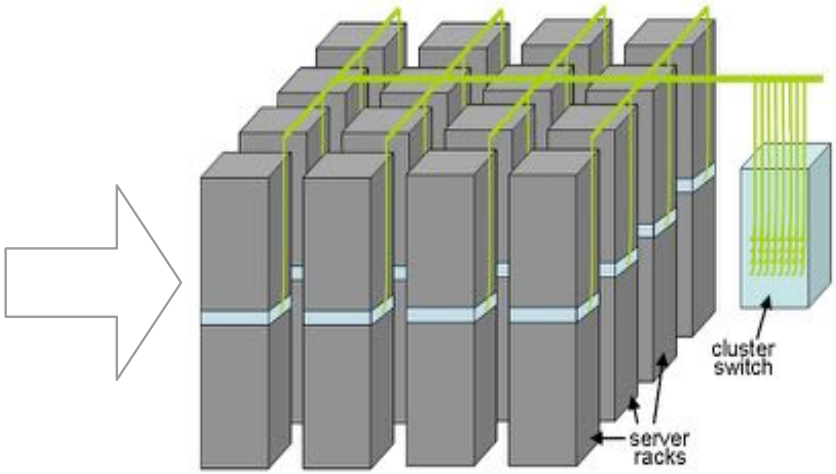
Slide credit: Kristen Grauman

Running in The Real World

Machines + Racks



Clusters



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

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?[ʊ]

[ʊ] 3.5 billion queries/day -> 40k/second -> **every 25 seconds**,

A Slight Digression g SAT and Program Verification

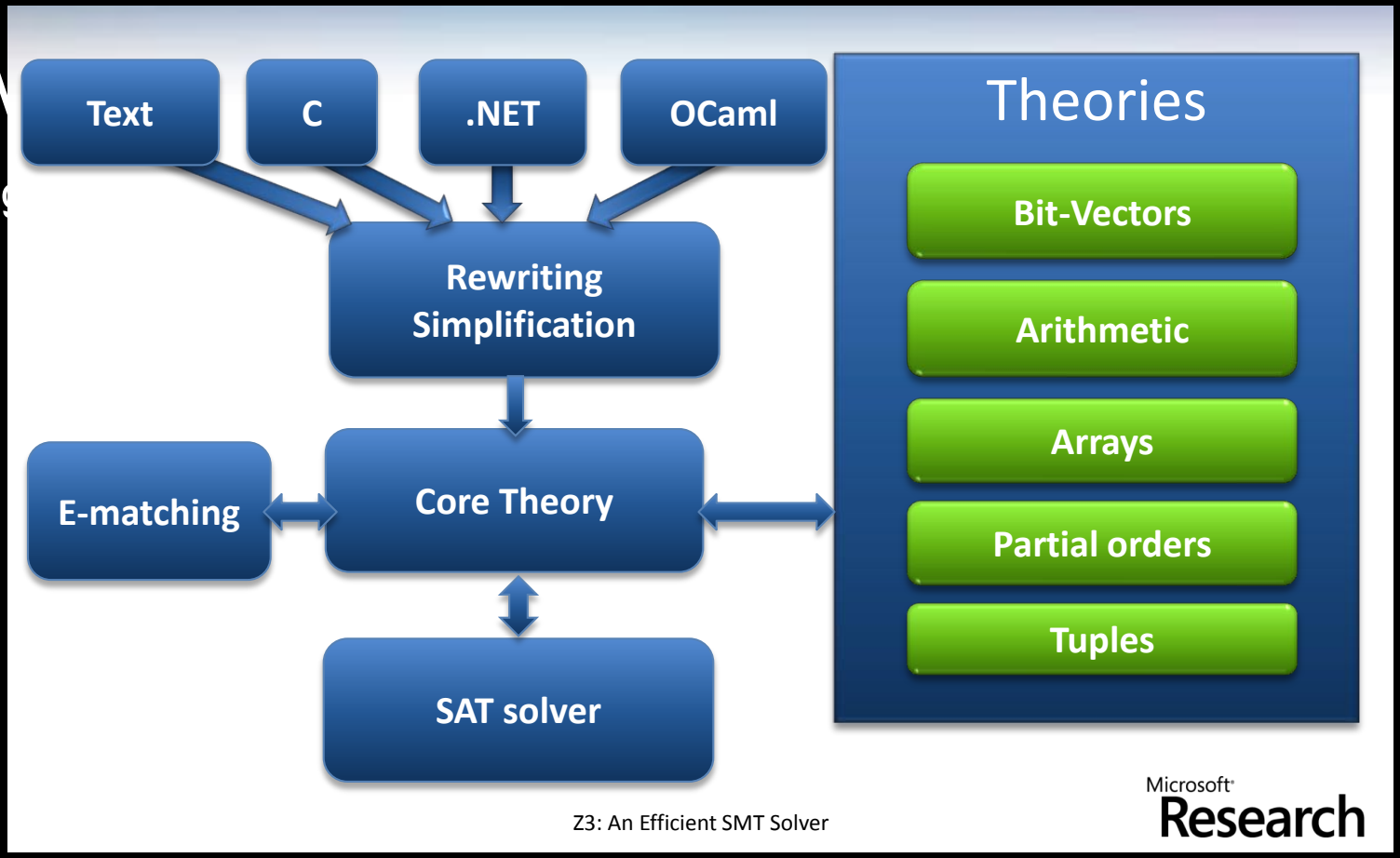
- Simple problem: is a boolean formula satisfiable?
  - A & B === yes A=1, B=1
  - A & ~A === no!
- Original NP complete problem (Cook 1971)
- So What?

A Slight Digression -- SAT

Courtesy Vijay Ganesh

Add Some Horsepower

- M
  - s



A large fraction of windows bugs now found by program verifications

http://research.microsoft.com/en-us/um/redmond/projects/z3/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**

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

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

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!

DEPARTMENT OF
COMPUTER SCIENCE

THE NEW AGE OF DISCOVERY

Where is Computing Going?

Gregory D. Hager
Professor and Chair

JOHNS HOPKINS
UNIVERSITY
WHITING SCHOOL OF ENGINEERING

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.

Fig. 1. The Navlab

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

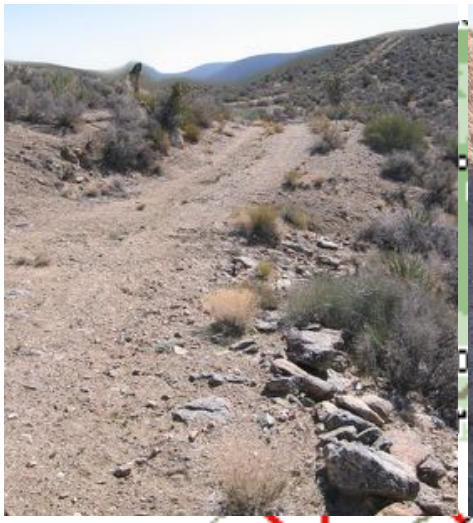
Automated Driving 1996

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**

(Pomerleau et al. CMU)

No Hands Across America

2797/2849 miles (98.2%)
automated driving



- 150 mile off-road robot race across the Mojave desert
- Natural and manmade hazards
- No driver, no remote control
- No dynamic passing
- Fastest vehicle wins the race (and 2 million dollar prize)

Courtesy Thrun

Straight Roads

Mountain Roads

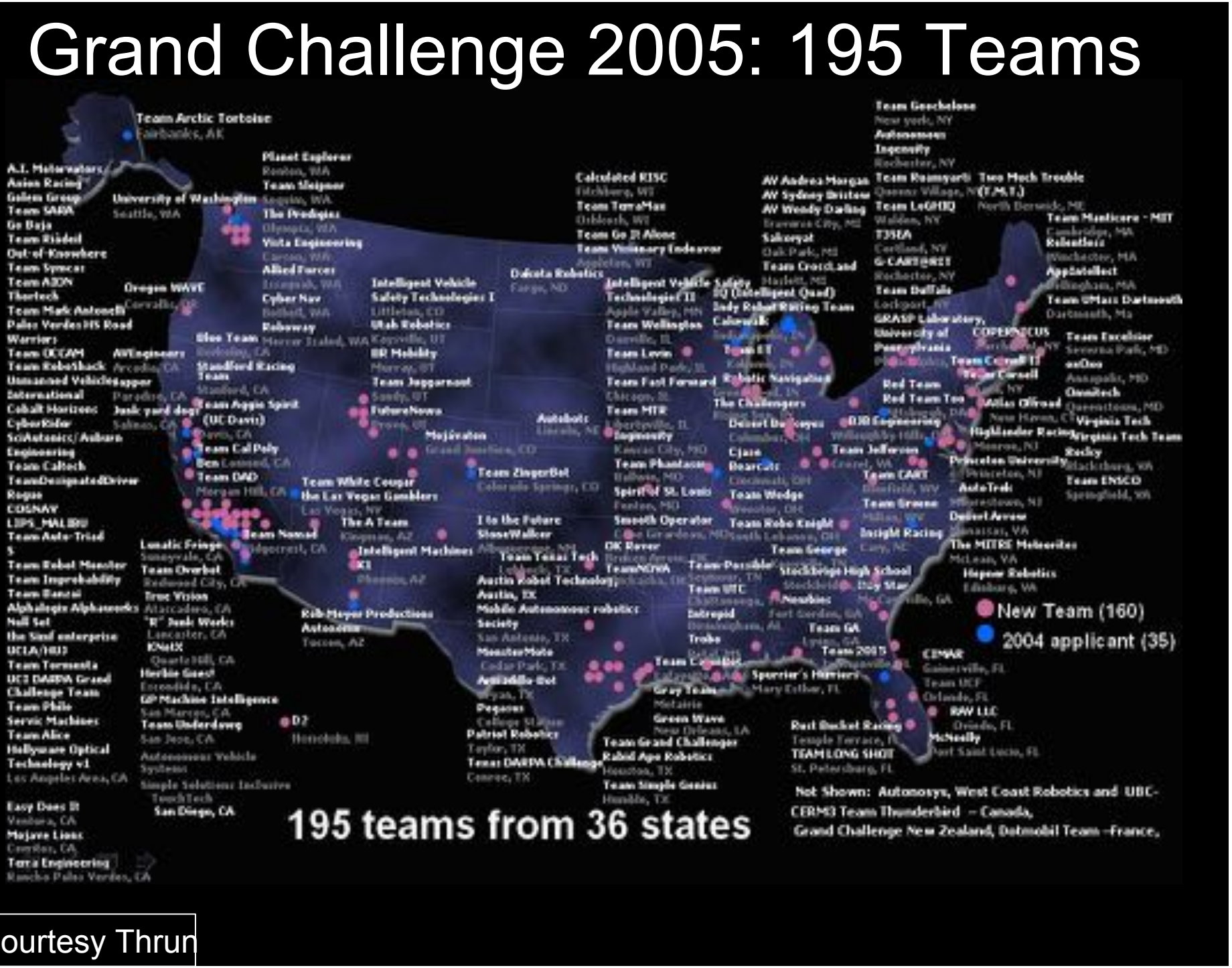
Cattle Guards

Open Lake

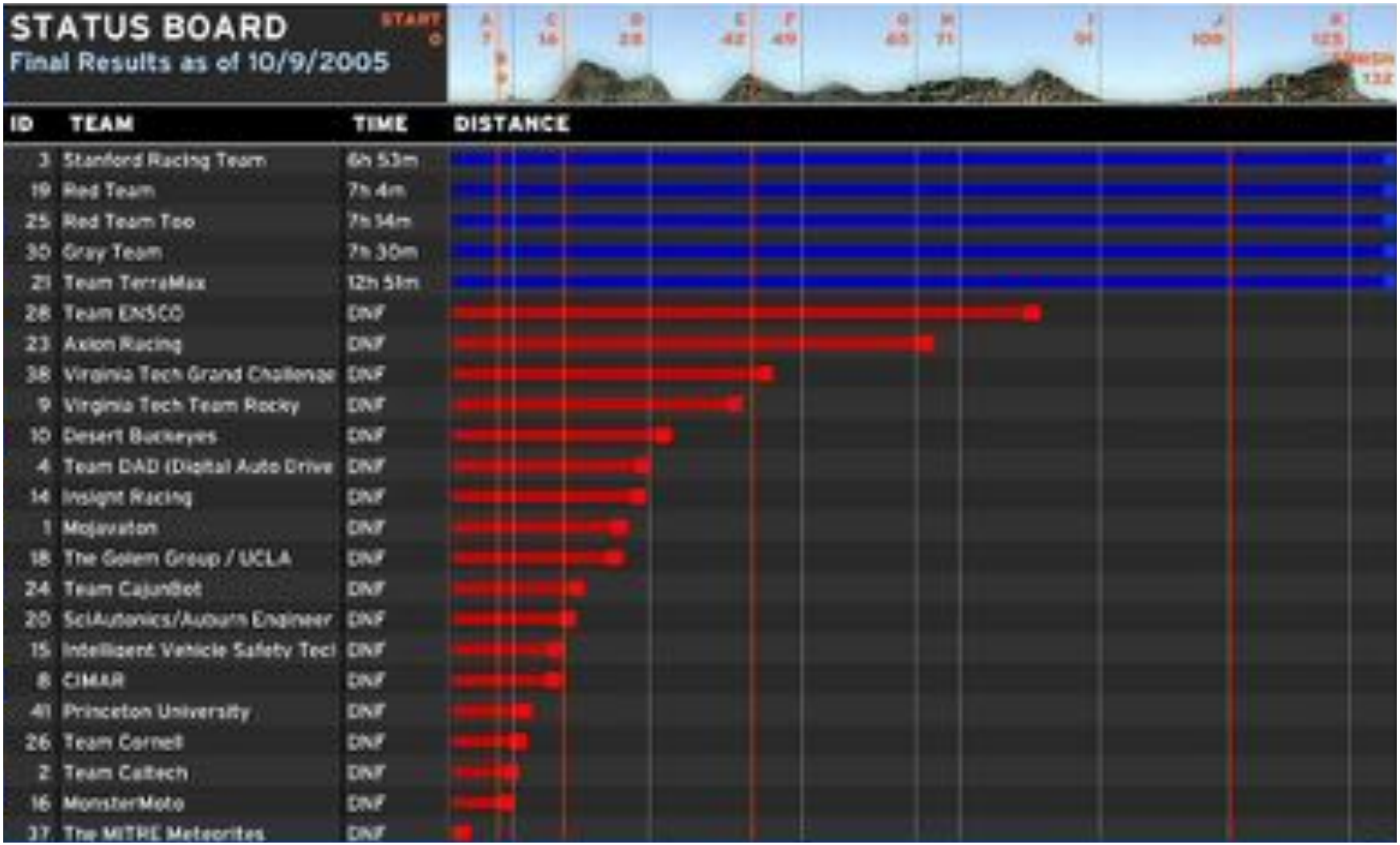
The 2004 DARPA Grand Challenge

- 150 mile course through Mojave Desert
- 15 entrants who made it to the final https://www.youtube.com/watch?v=FaBJ5sPPmcI
- After a few hours, all failed
- Sandstorm (CMU) made it 7.4 miles

Grand Challenge 2005: 195 Teams



Courtesy Thrun

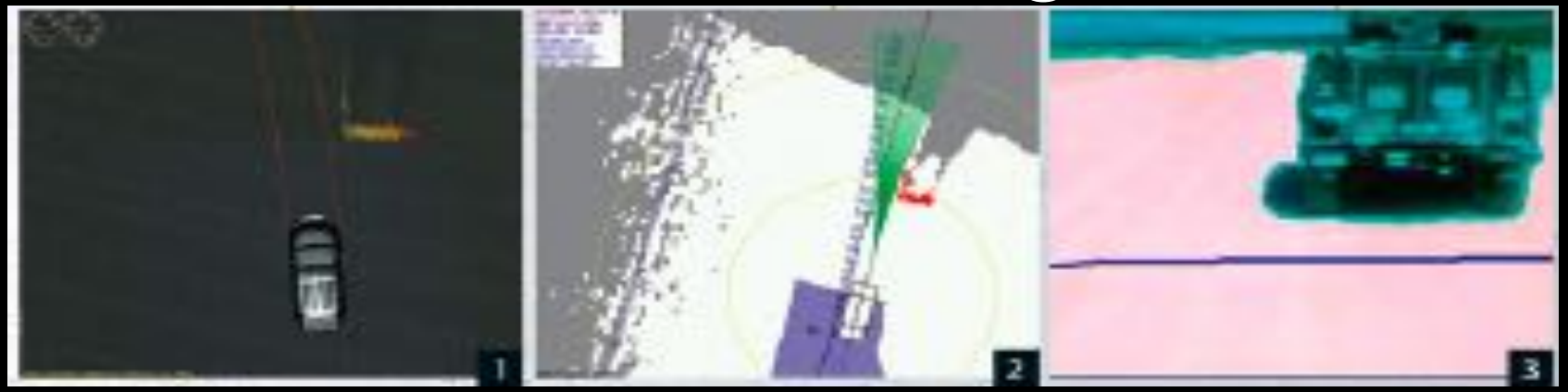

STATUS BOARD
Final Results as of 10/9/2005
START
ID
TEAM
TIME
DISTANCE
3 Stanford Racing Team 6h 53m
19 Red Team 7h 4m
25 Red Team Too 7h 14m
30 Gray Team 7h 30m
21 Team TerraMax 12h 51m
28 Team ENSCO DNF
23 Axion Racing DNF
38 Virginia Tech Grand Challenge DNF
9 Virginia Tech Team Rocky DNF
10 Desert Buckeyes DNF
4 Team DAD (Digital Auto Drive DNF
14 Insight Racing DNF
1 Mojavaton DNF
18 The Golem Group / UCLA DNF
24 Team CajunBot DNF
20 SciAutonics/Auburn Engineer DNF
15 Intelligent Vehicle Safety Tec DNF
8 CIMAR DNF
41 Princeton University DNF
26 Team Cornell DNF
2 Team Caltech DNF
16 MonsterMoto DNF
17 The MITRE Meteorites DNF

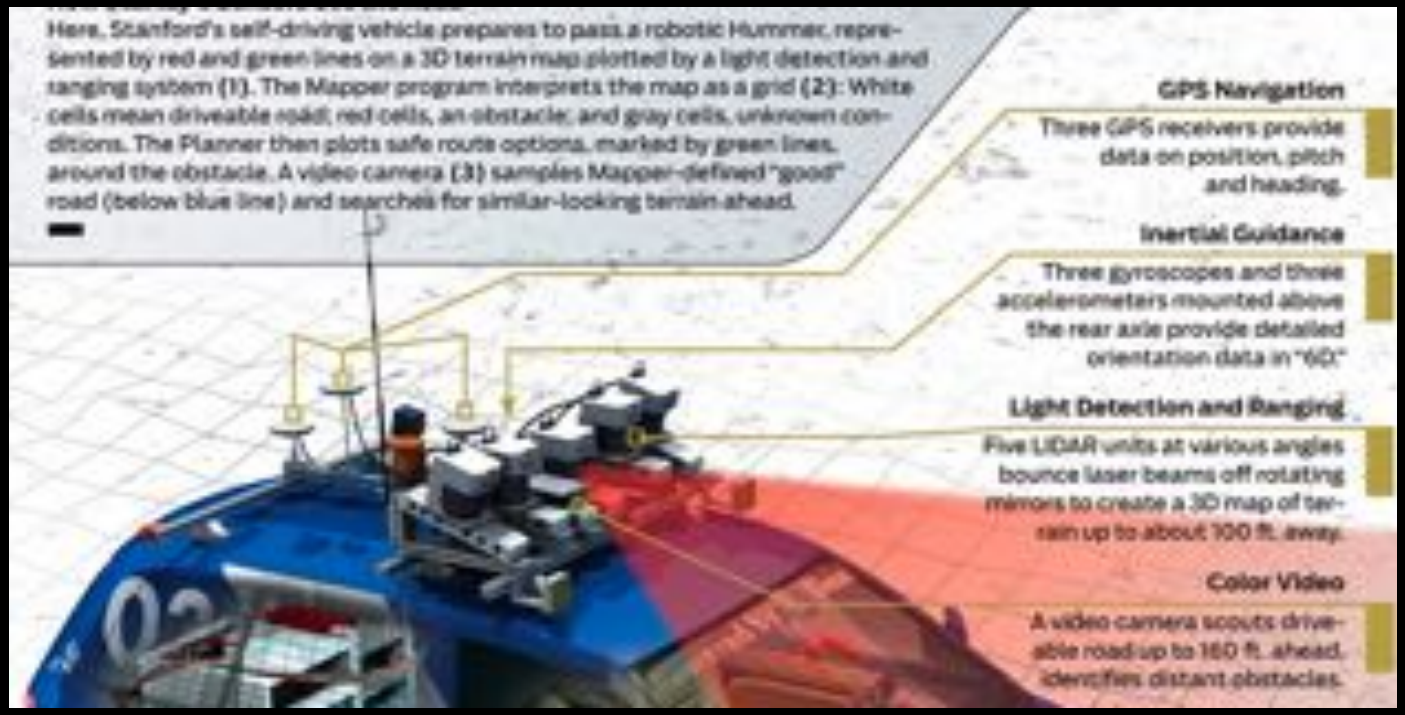

MDV


22


25

Automated Driving 2006





Stanford DARPA Grand Challenge Vehicle (Thrun)

(Popular Mechanics Jan '06)

SENSOR INTERFACE

RDDF database

Laser 1 interface

Laser 2 interface

Laser 3 interface

Laser 4 interface

Laser 5 interface

Camera interface

Radar interface

GPS position

GPS compass

IMU interface

Wheel velocity

Brake/steering

PERCEPTION

Road finder

Laser mapper

Vision mapper

Radar mapper

UKF Pose estimation

Surface assessment

PLANNING&CONTROL

Top level control

Path planner

Steering control

Throttle/b

Healt

USER INTERFACE

Touch screen UI

Wireless E-Stop

VEHICLE INTERFACE

Touareg interface

Process controller

GLOBAL SERVICES

Data logger

Communication requests

Communication channels

Inter-process communication (IPC) server

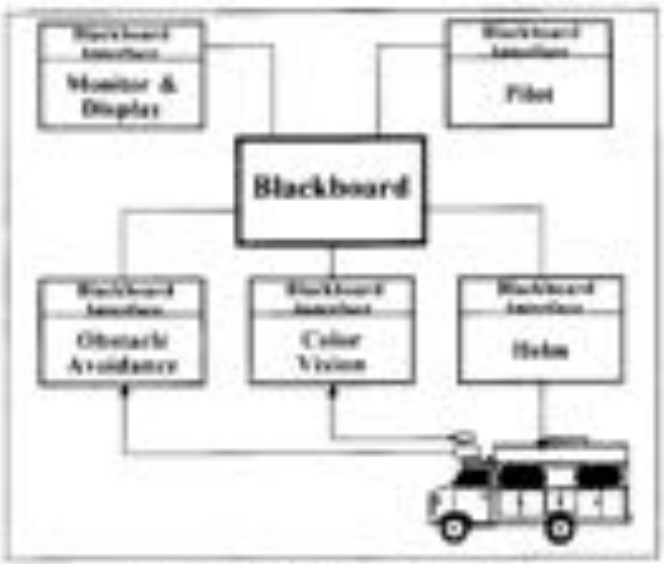
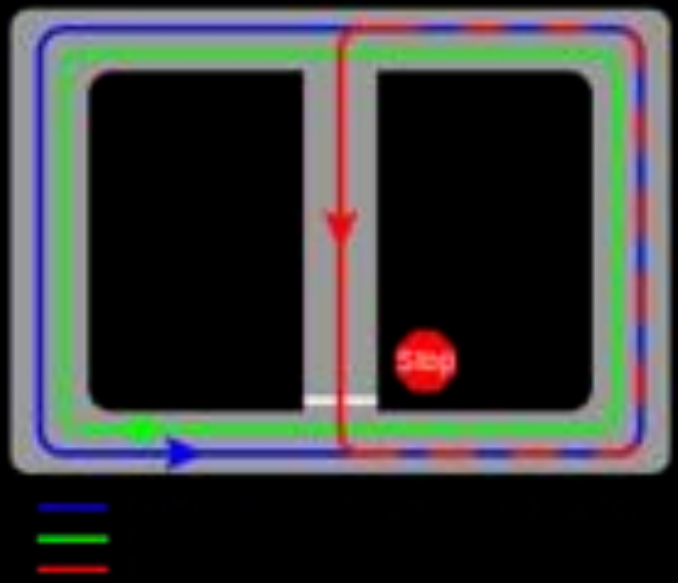


Fig. 20. Navlab software architecture

The Next Challenges

- Darpa Urban Challenge: https://www.youtube.com/watch?v=0wJAANgG-Vg



Sensors
Lasers, radars and cameras detect objects in all directions

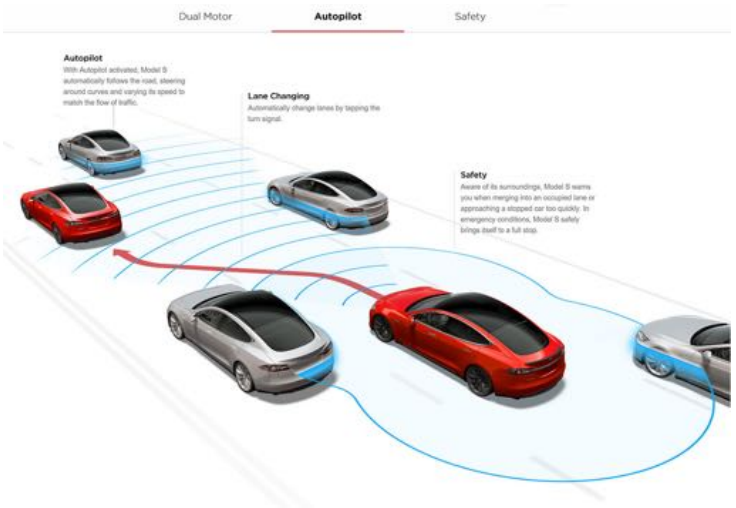
Interior
Designed for riding, not for driving

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

0 Comments | Print | Reprints | Respond | Recommend 0

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.

Stanford Racing Team

DARPA Robotics Challenge

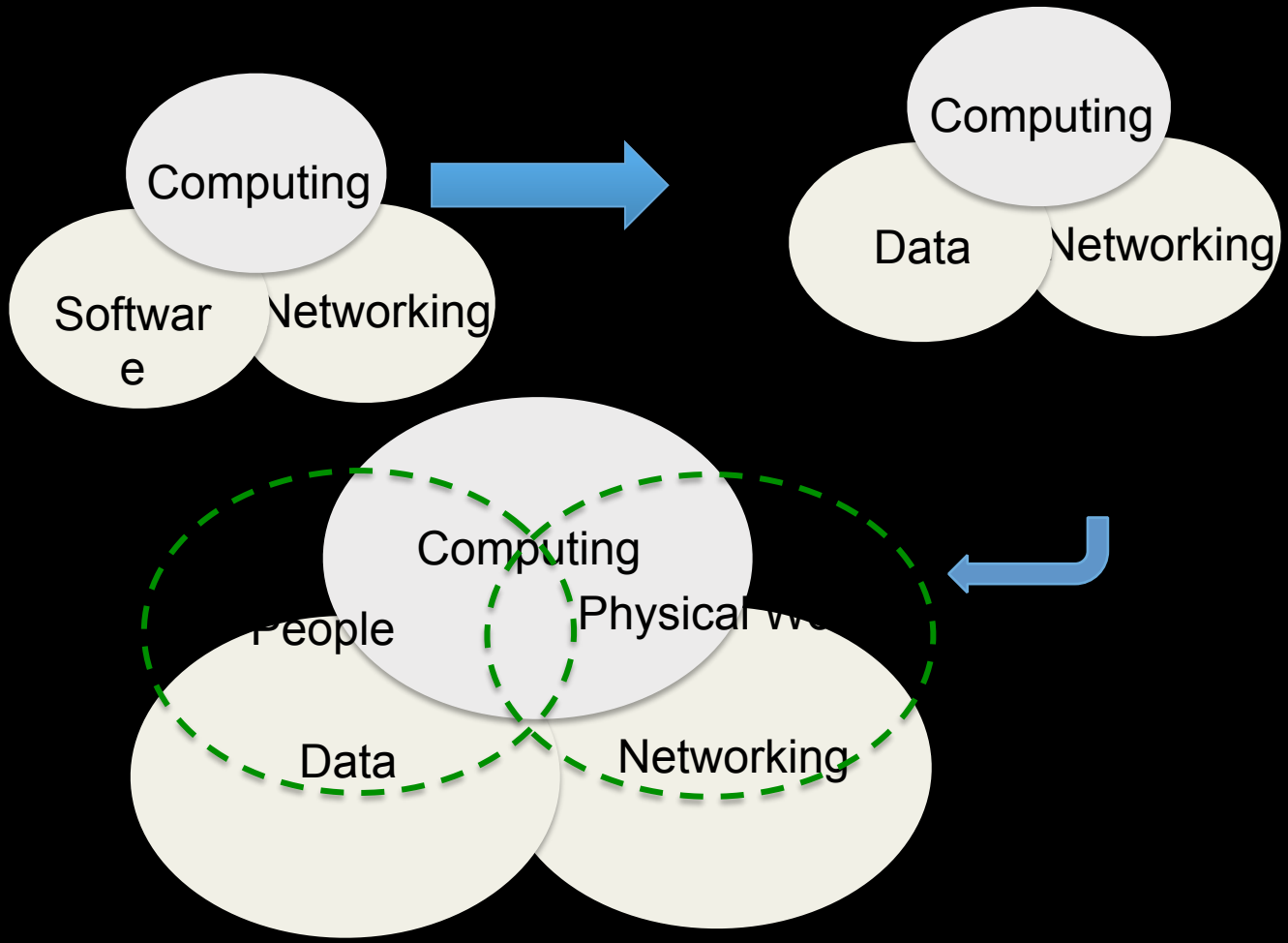
- 1. Drive a utility vehicle at the site.
- 2. Travel dismounted across rubble.
- 3. Remove debris blocking an entryway.
- 4. Open a door and enter a building.
- 5. Climb an industrial ladder and traverse an industrial walkway.
- 6. Use a tool to break through a concrete panel.
- 7. Locate and close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

DARPA Robotics Challenge

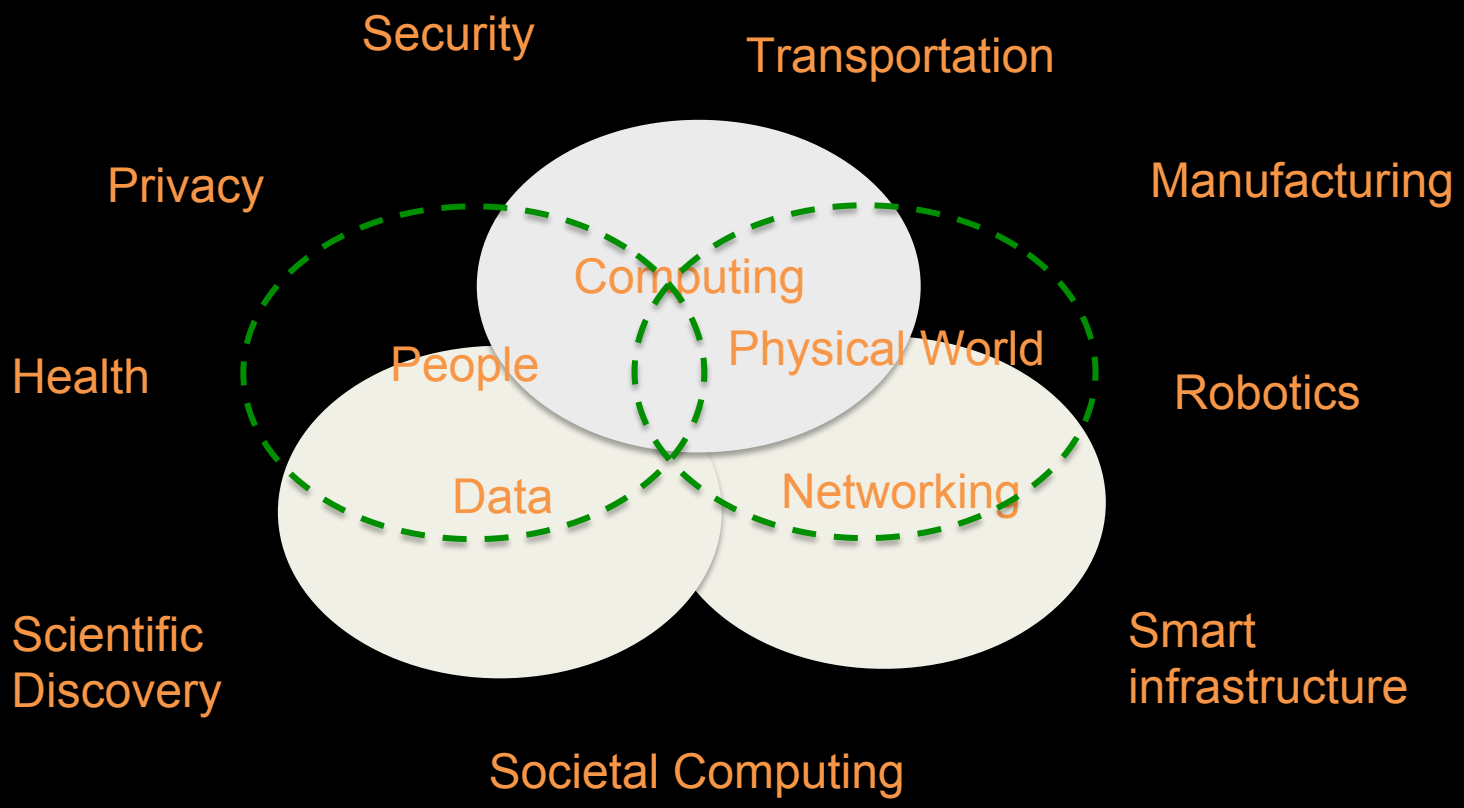
From Data to Information

Gregory D. Hager

Evolution of NIT 1991-2015



Evolution of NIT 1991-2015



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?

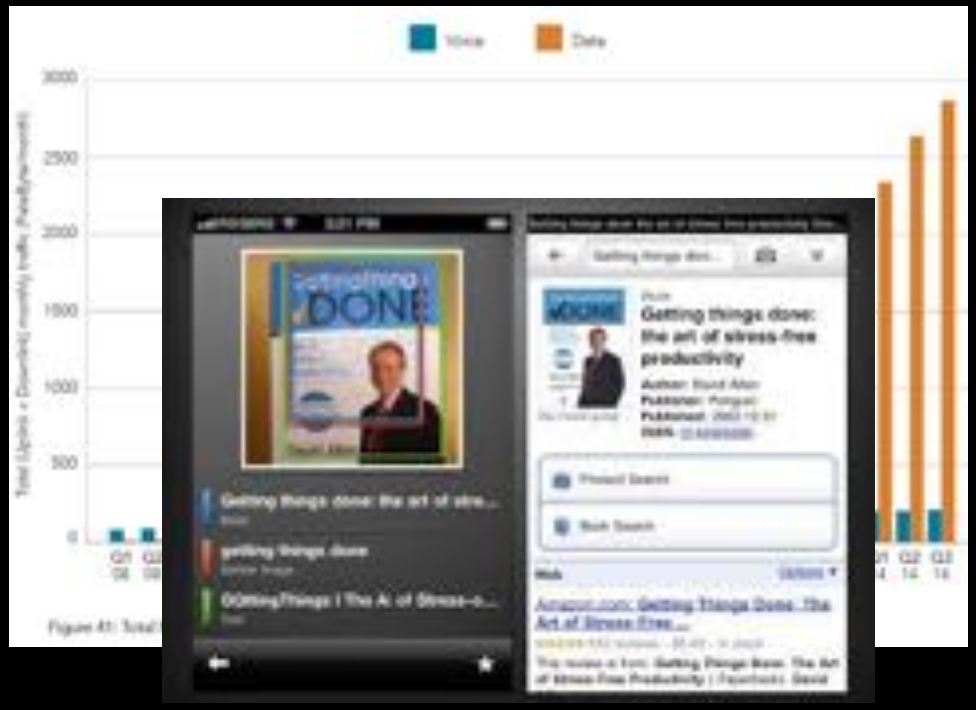
Geometric Computer Vision at Scale

Noah Snavely, Steven M. Seitz, and Richard Szeliski. Modeling the world from Internet photo collections. *International Journal of Computer Vision, 80(2):189-210, November 2008.*

Hamlyn 2014, Copyright GD Hager

The Rest of the World

Mobile Data Traffic by Quarter



Credit: akamai

IMAGENET

14,197,122 images, 21841 synsets indexed

Explore Download Challenges Publications CoolStuff About

Not logged in. Login | Signup

Copyright GD Hager 2015

Data at Scale Has its Benefits

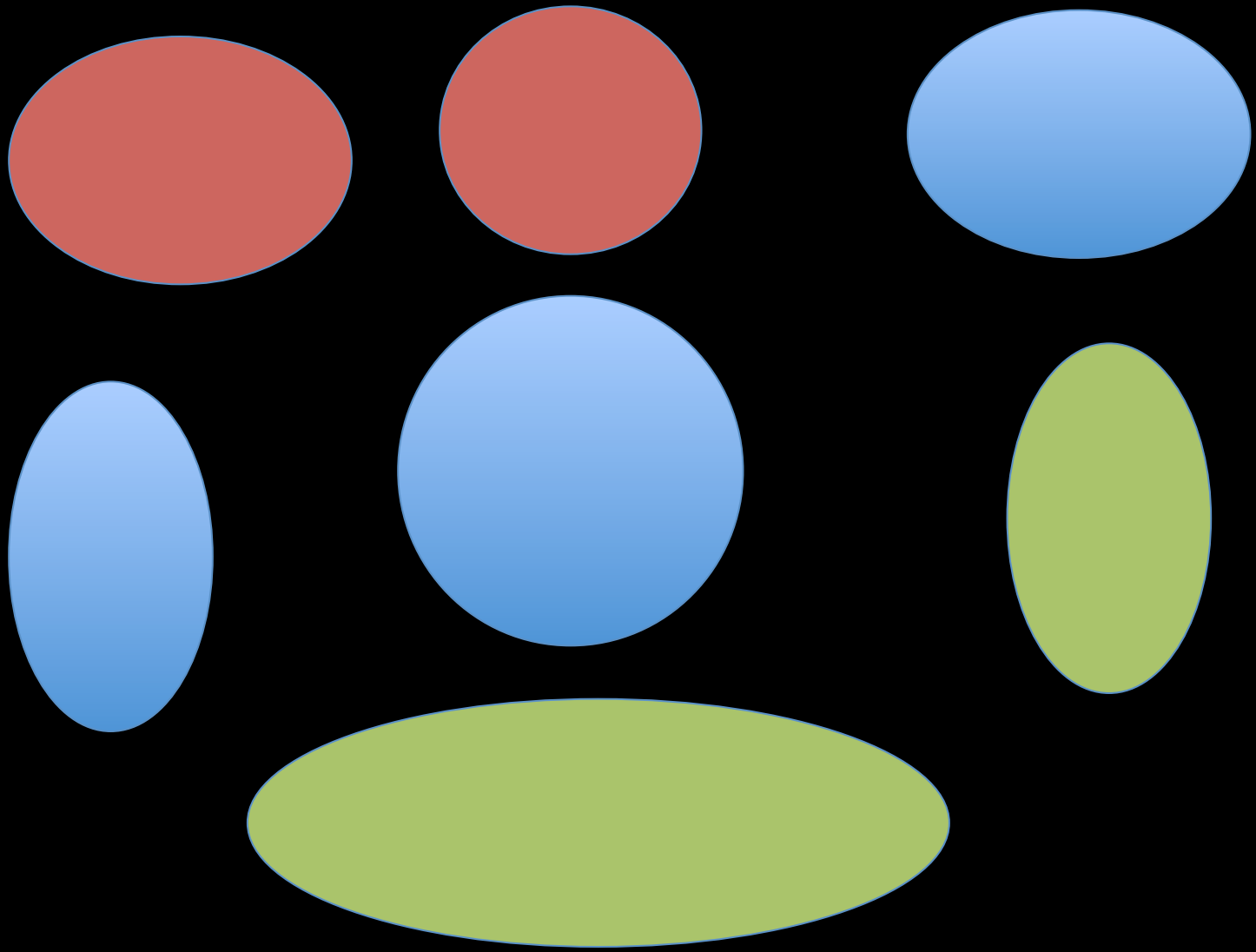
Copyright GD Hager 2015

How Do We Learn From Data?

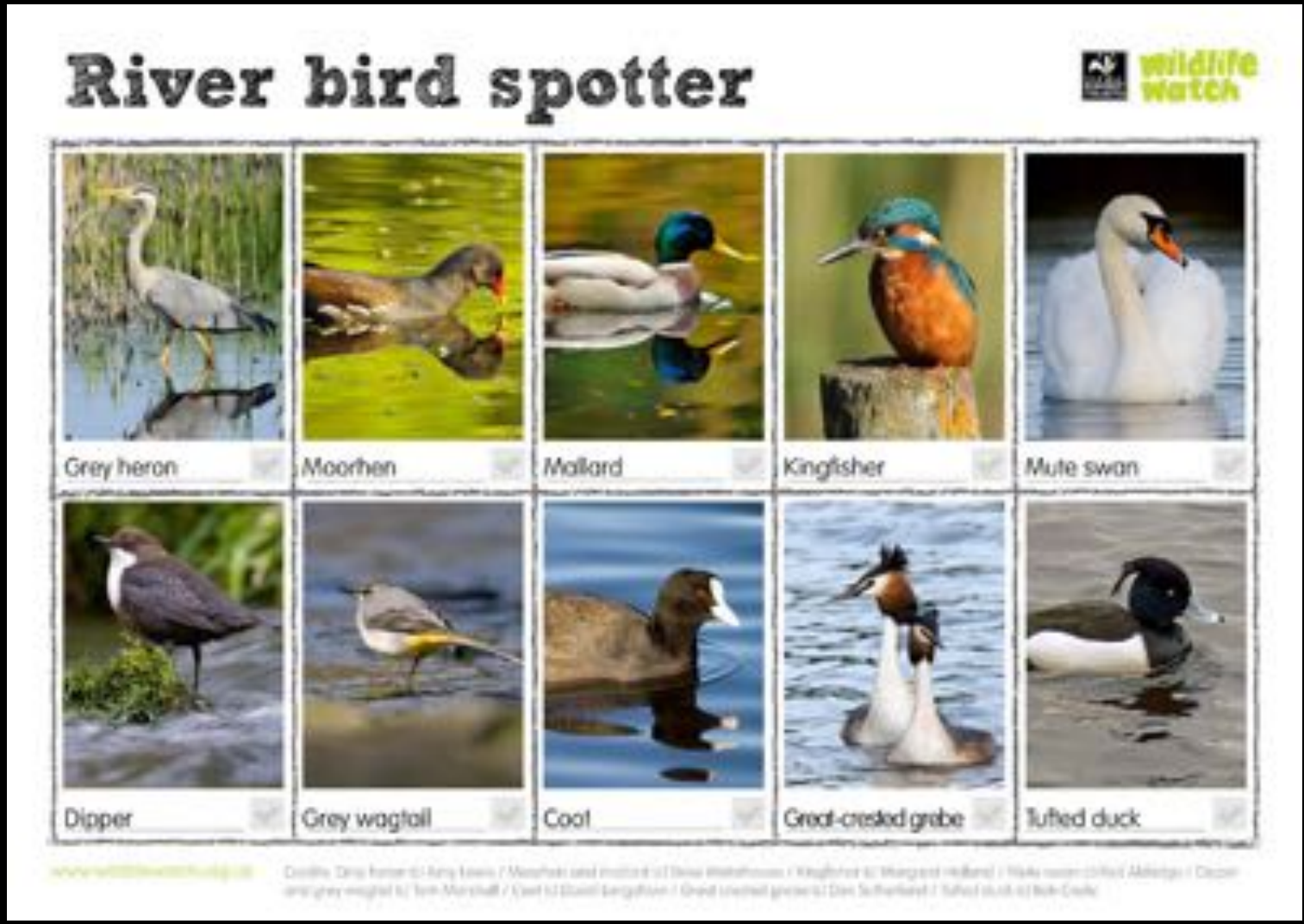
- Unsupervised Learning (data structuring)
- Supervised Learning
- Reinforcement Learning

Machine learning: Trends, perspectives, and prospects, MI Jordan, TM Mitchell. Science 349 (6245), 255-260

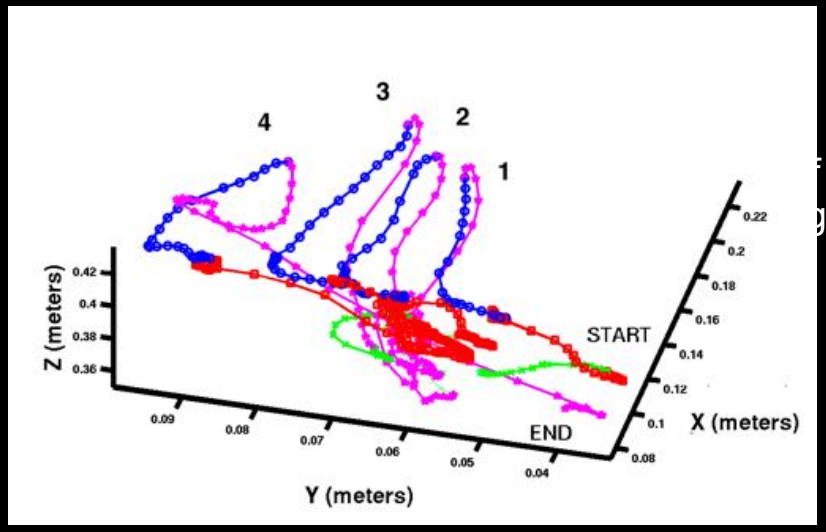
Unsupervised g An Example

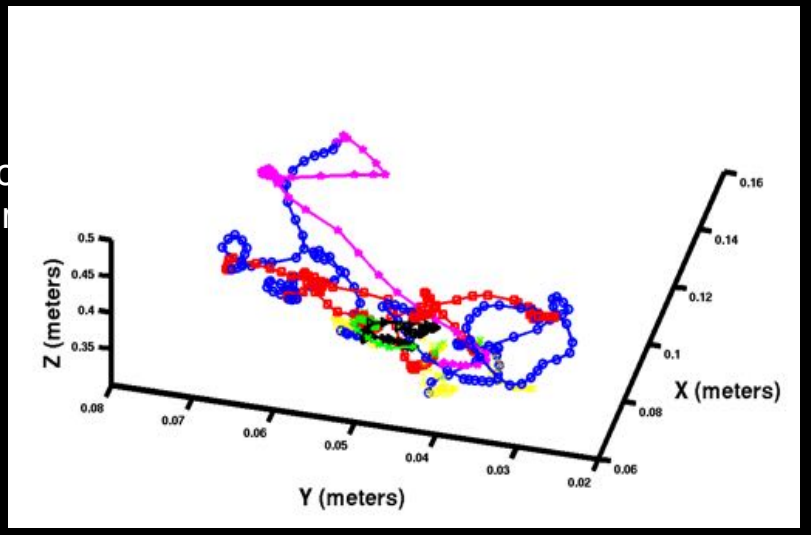


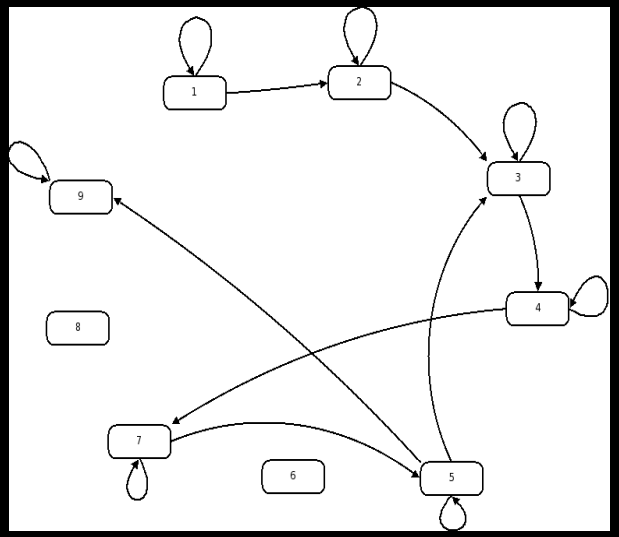
Supervised g An Example

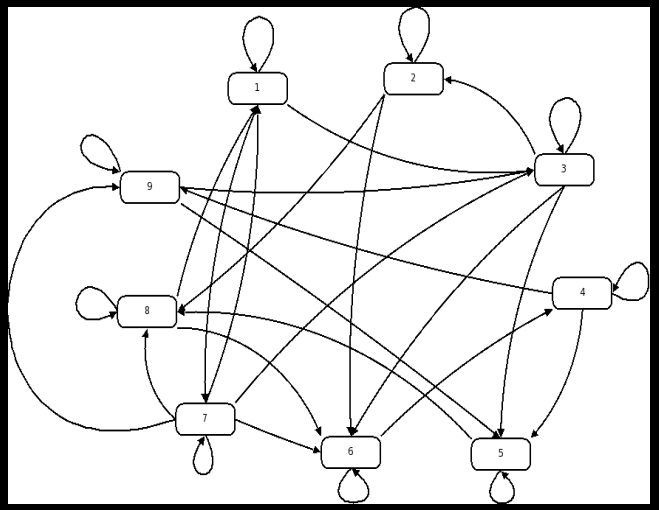


A Real Example


3
4
2
1
START
END
Z (meters)
Y (meters)
X (meters)


Z (meters)
Y (meters)
X (meters)


1
2
3
4
5
6
7
8
9

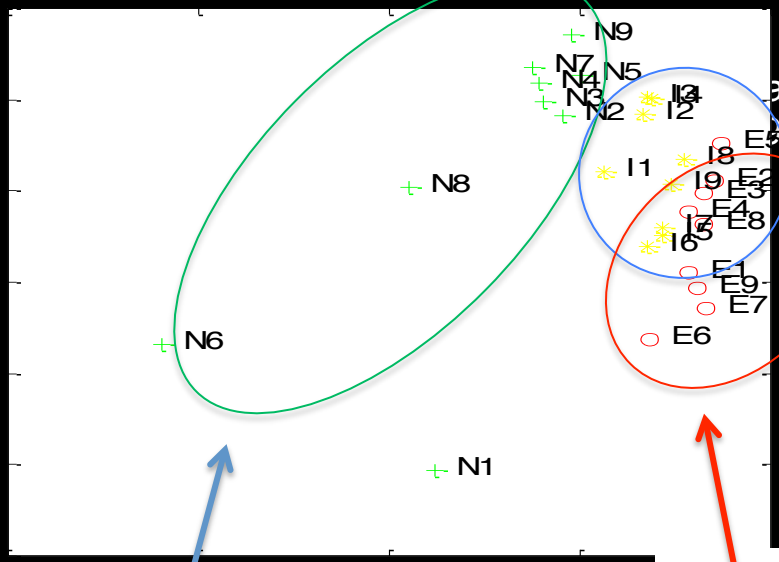

1
2
3
4
5
6
7
8
9

Task versus Subt

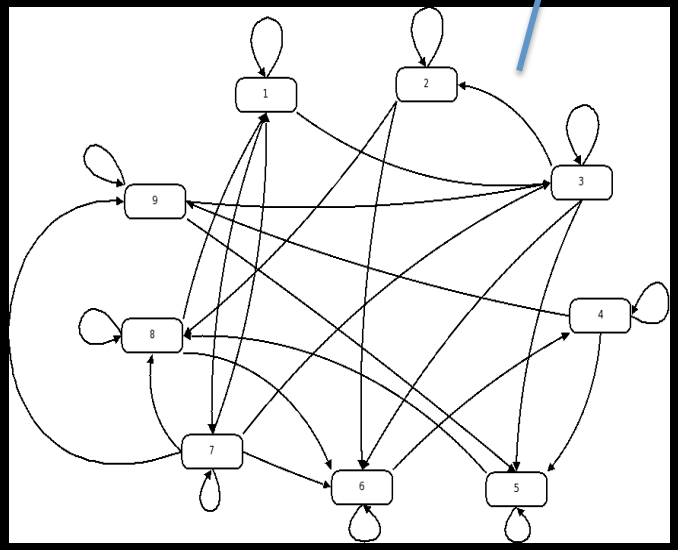
asive Surgery Carol

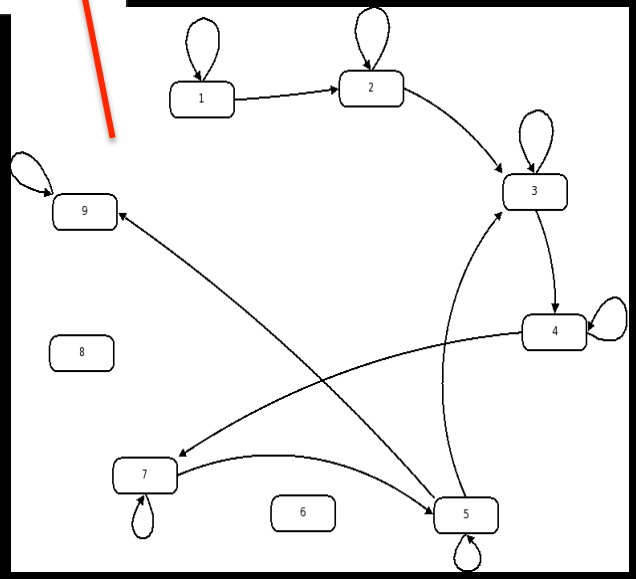
Novices

Intermediates

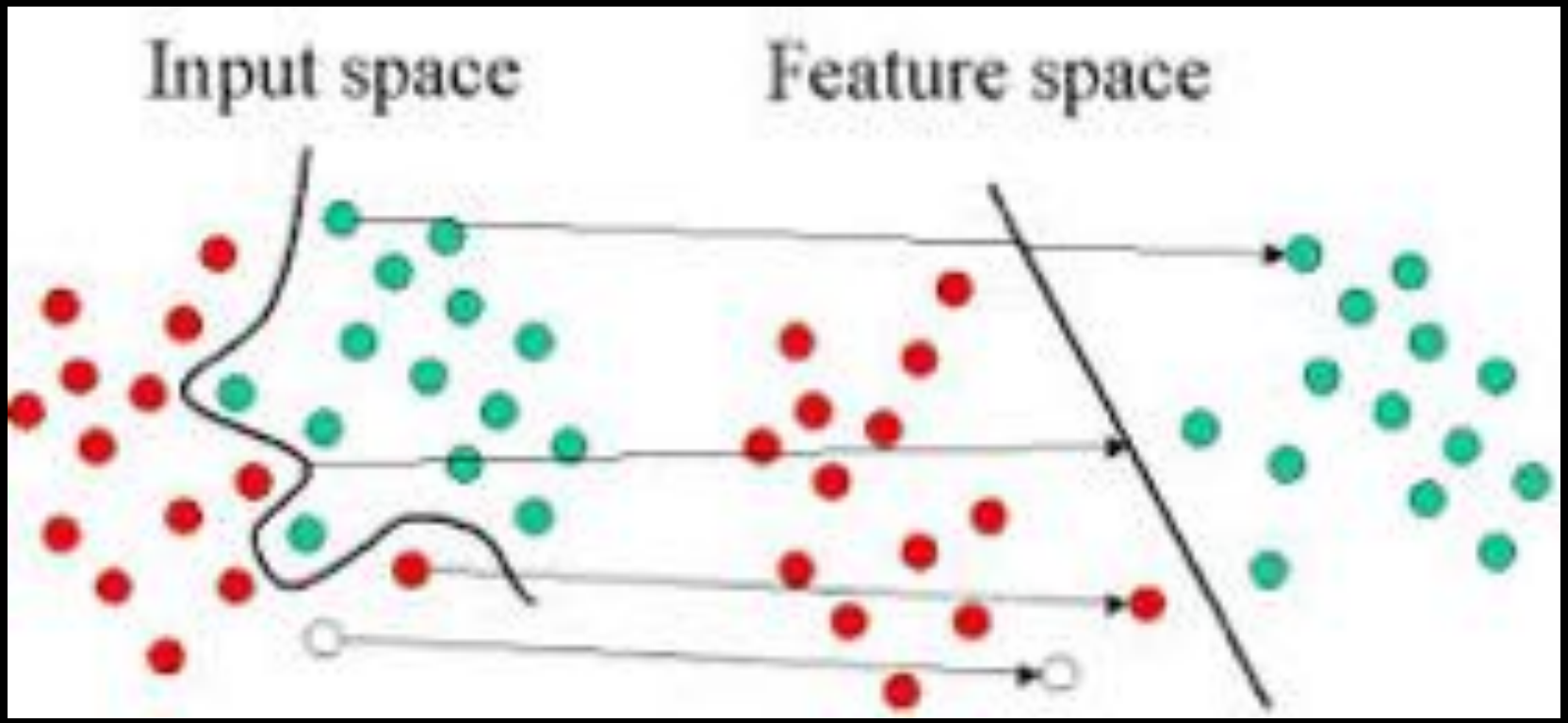


Experts

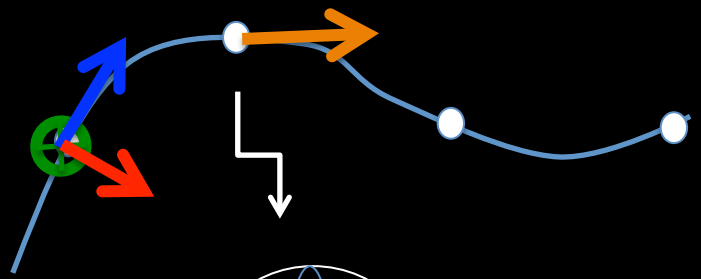


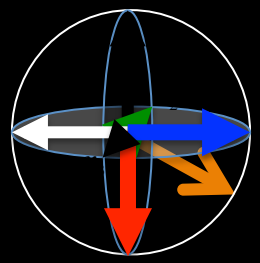


Supervised g An Example



String Motifs as a Representation



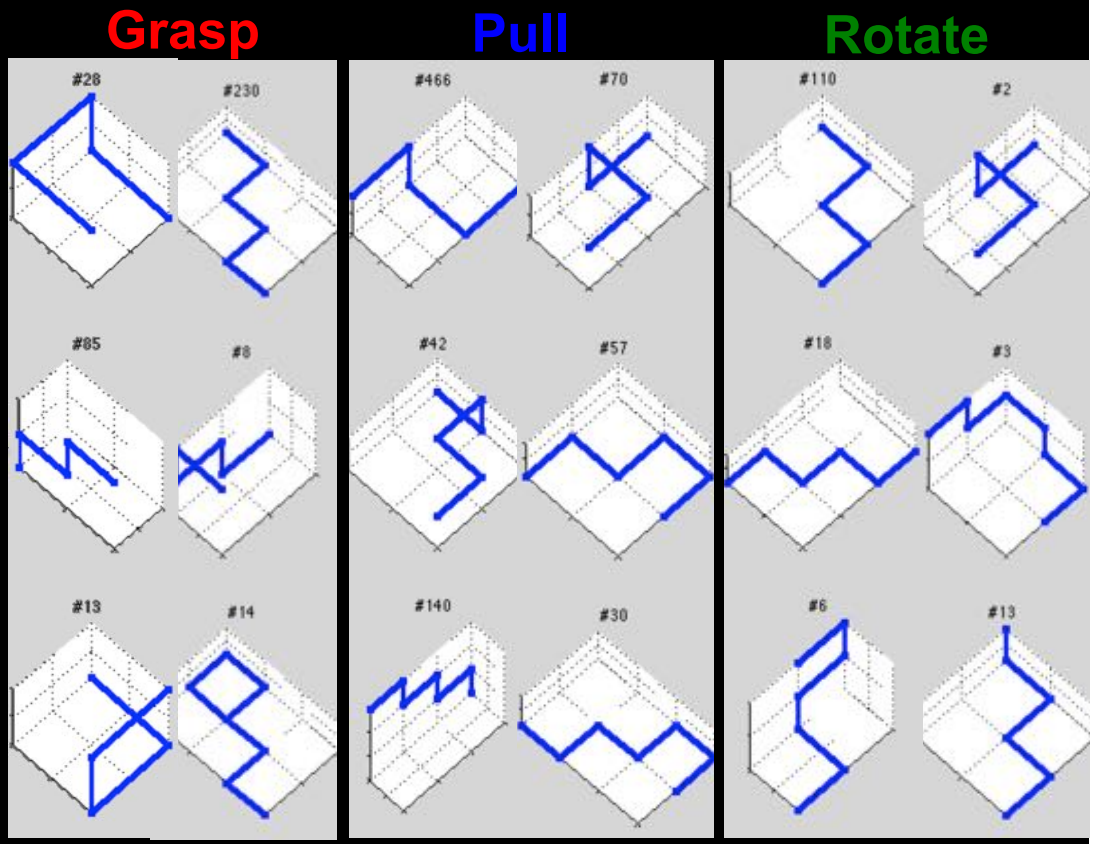


```
ID=1
gddfdfffffffeggefgeegefdeffffgeedfggfefefe

ID=2
cggeefcdffeffegcfgeeffgefgefgefggeddggcffff

ID=3
ecffdfgeeggcfgeeffggcffffegeeggfegggeccffff
```

Encode motion as a string



Build category-specific dictionaries

Ahmidi, et al. "String Motif-Based Description of Tool Motion for Detecting Skill and Gestures in Robotic Surgery." *MICCAI*, 2013.

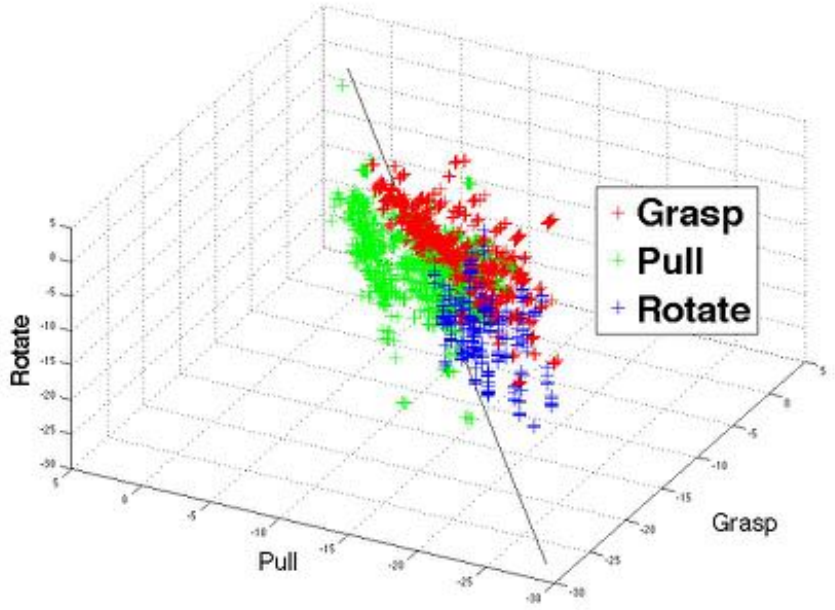
String Similarity

likelihood length location

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

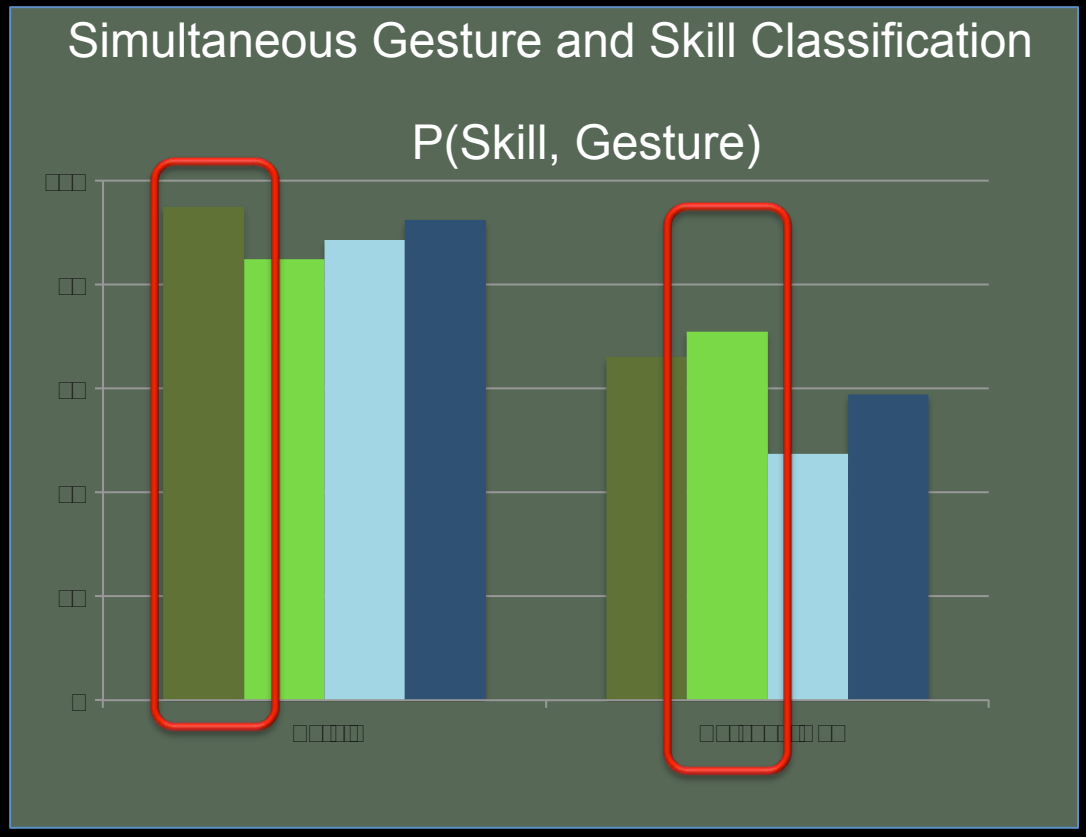
$$A_i = \frac{1}{1 + \sum_{J \in O} |Y - J|}$$

String	#	P	|C|	A	%
dfffgg	9	0.000103	6	0.001447	0.002275
fdfffg	3	0.000034	6	0.004464	0.002260
gffgee	3	0.000034	6	0.008264	0.001566
ddgggg	3	0.000034	6	0.003367	0.002578
fgeefe	2	0.000023	6	0.008850	0.001952
ffgeef	2	0.000023	6	0.009346	0.001890
ggffge	2	0.000023	6	0.012821	0.001534
dggggg	2	0.000023	6	0.004348	0.002752
ffddgg	2	0.000023	6	0.004831	0.002634
dggffg	1	0.000011	6	0.034483	0.001211
ffggff	1	0.000011	6	0.015385	0.002120
fffggf	1	0.000011	6	0.019608	0.001847
fedfdf	1	0.000011	6	0.012821	0.002325
fddggg	1	0.000011	6	0.011364	0.002461



Ahmidi, et al. "String Motif-Based Description of Tool Motion for Detecting Skill and Gestures in Robotic Surgery." *MICCAI*, 2013.

Results at the Gesture Level



94.98% **70.91%**

Chance = 11%

Deep Networks

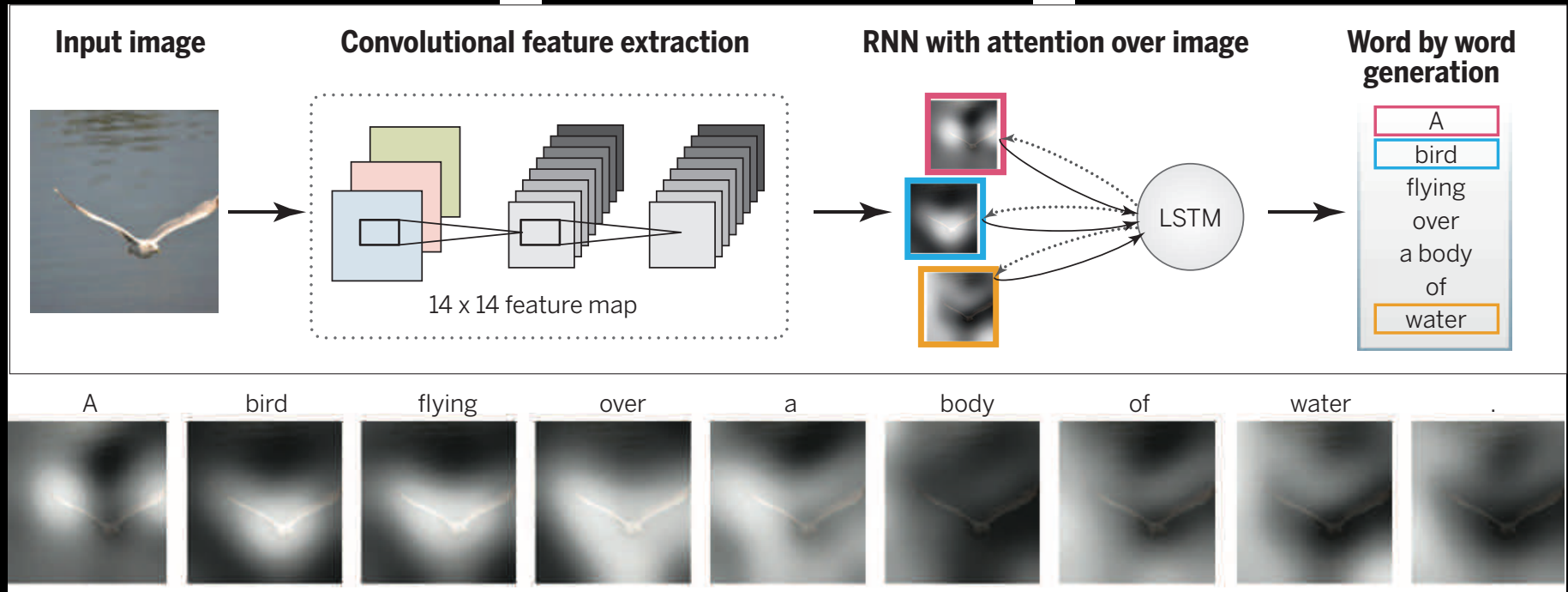
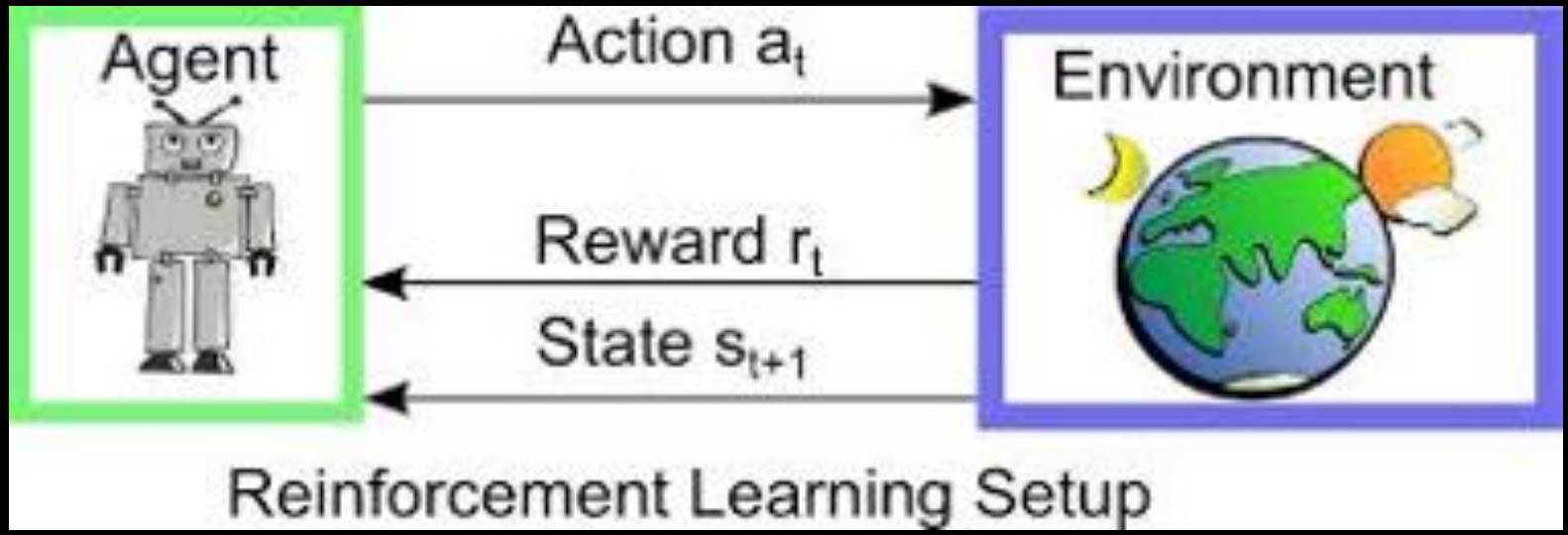


Fig. 2. Automatic generation of text captions for images with deep networks. A convolutional neural network is trained to interpret images, and its output is then used by a recurrent neural network trained to generate a text caption (top). The sequence at the bottom shows the word-by-word focus of the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

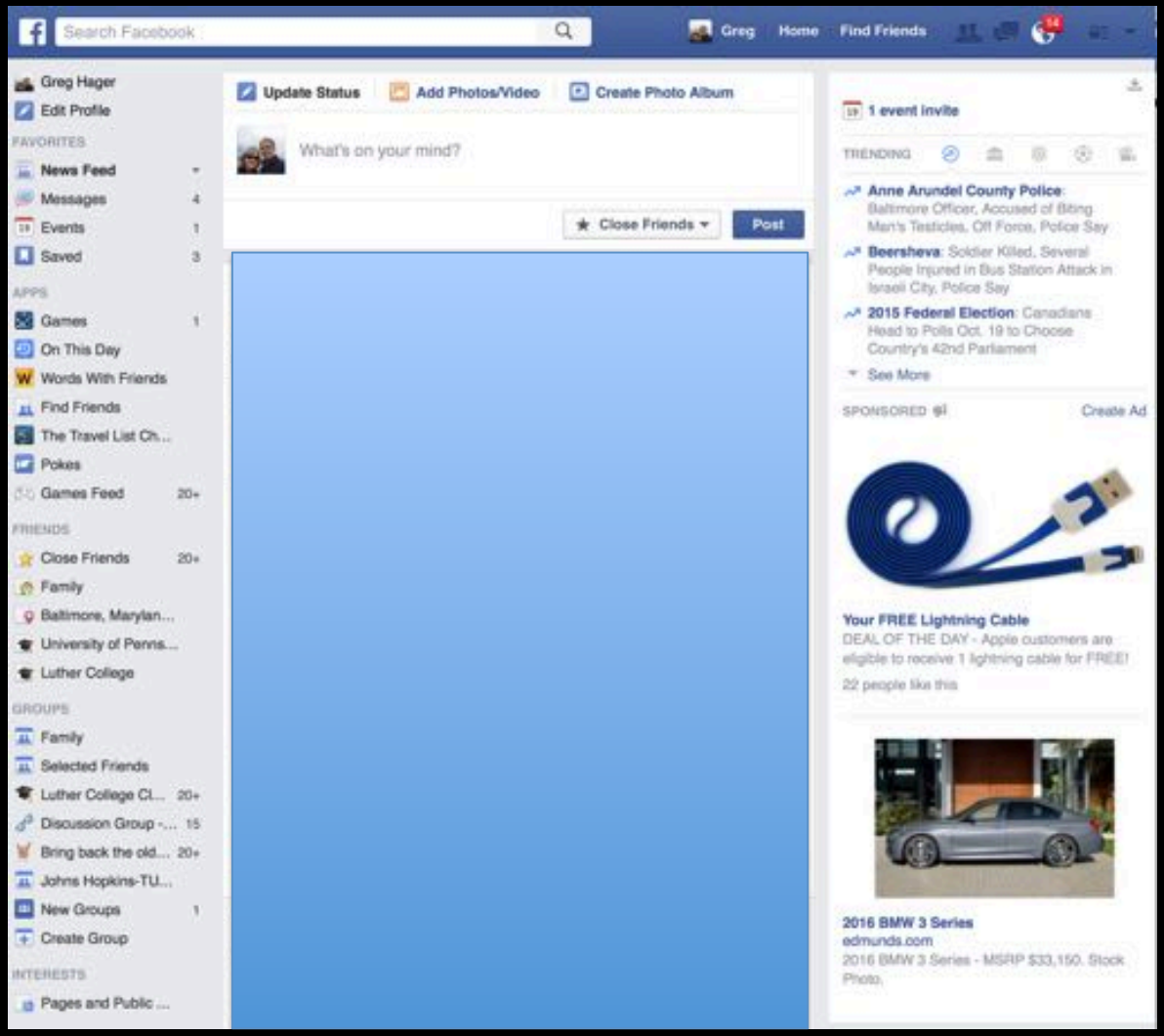
Machine learning: Trends, perspectives, and prospects, MI Jordan, TM Mitchell. Science 349 (6245), 255-260

Reinforcement Learning



Goal: Find a policy (choice of actions) that$^{\upsilon}$ **maximizes** reward

Reinforcement Learning




NETFLIX
AGENT.45
TAXI
bella ciao
KALIBER
hotel social
ZEN&ZACK
ZOMBIE COLONY
KENYA
OPIUM
Gift suggestion ...
Cross
(Audio CD) ~ Justice
(48)
$13.98 $12.99
Add to Your Cart
... based on Sarah's Facebook likes
Crystal Castles, The Polish Ambassador, Mr. Oizo, MGMT, Cut Copy
Do you have feedback? Let us know

Some of the Problems with Data

See 1 citation found by title matching your search:

Science. 2013 Jan 18;339(6117):321-4. doi: 10.1126/science.1229566.

Identifying personal genomes by surname inference.

Gymrek M[1], 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.

AI AND HEALTH

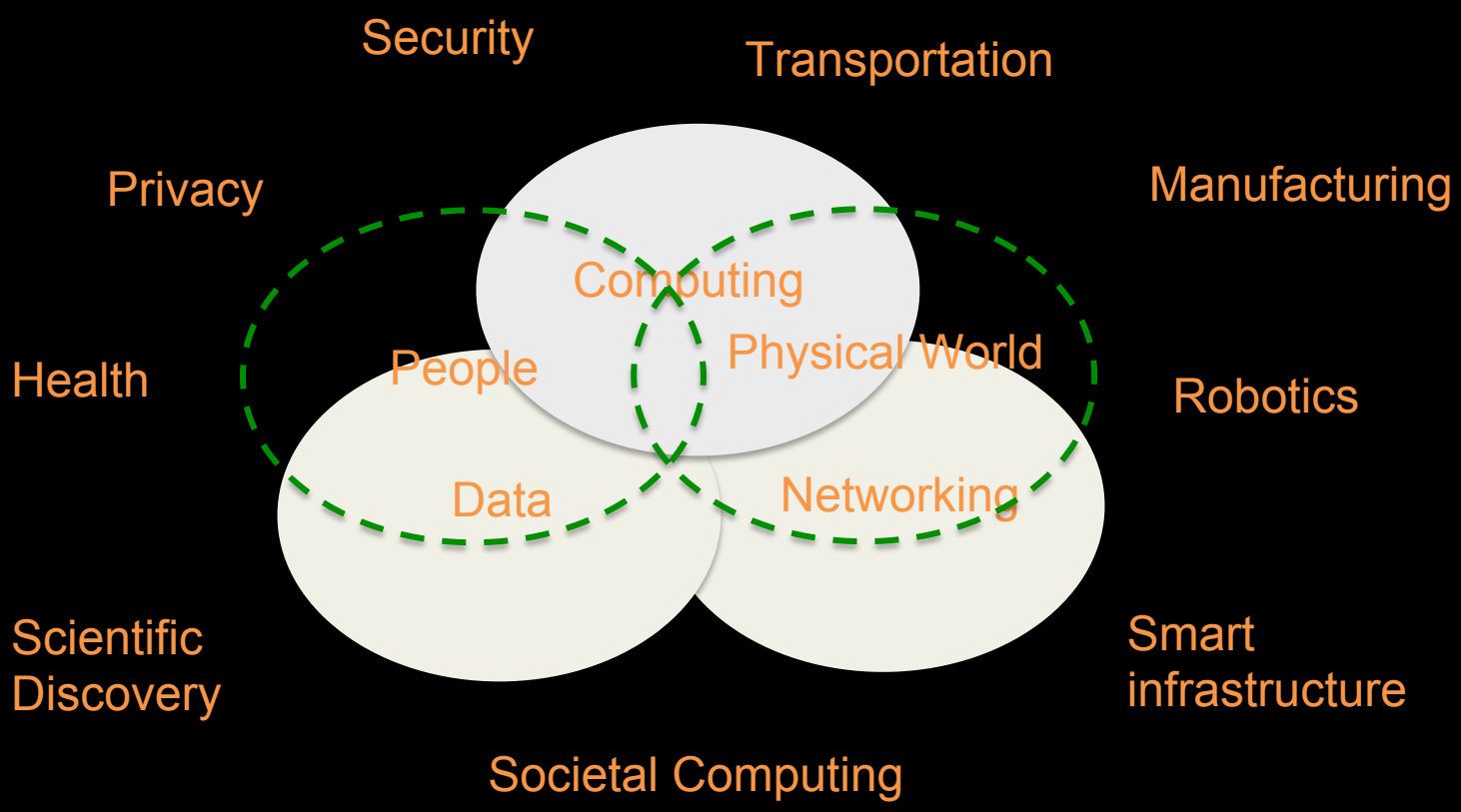
A $3 Trillion Challenge to Computational Scientists: Transforming Healthcare Delivery

IEEE Intelligent Systems,

Computing and Societal Needs

Gregory D. Hager

The Broader Impact of Computing

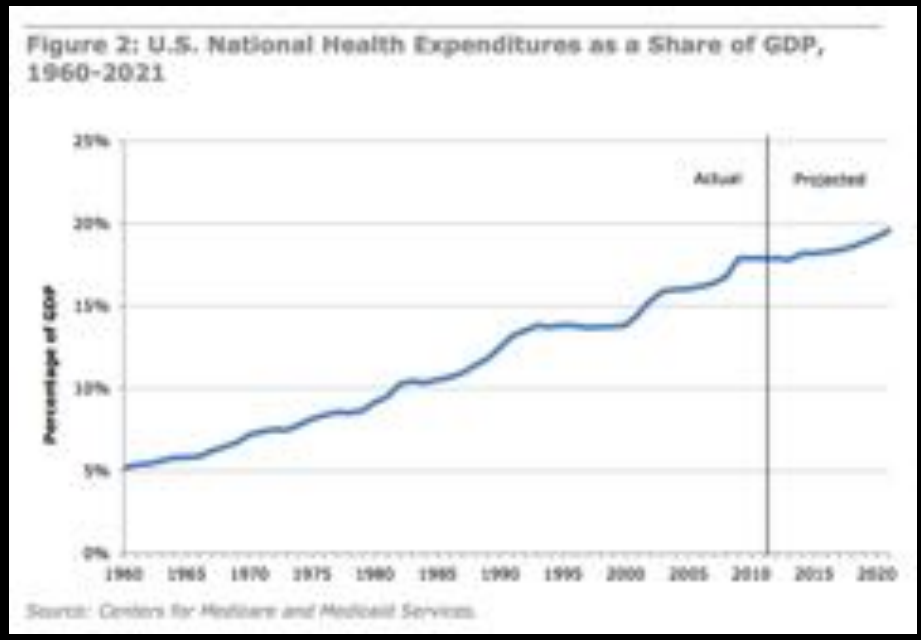


Medicine vs. Healthcare

- 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 o**rganization 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

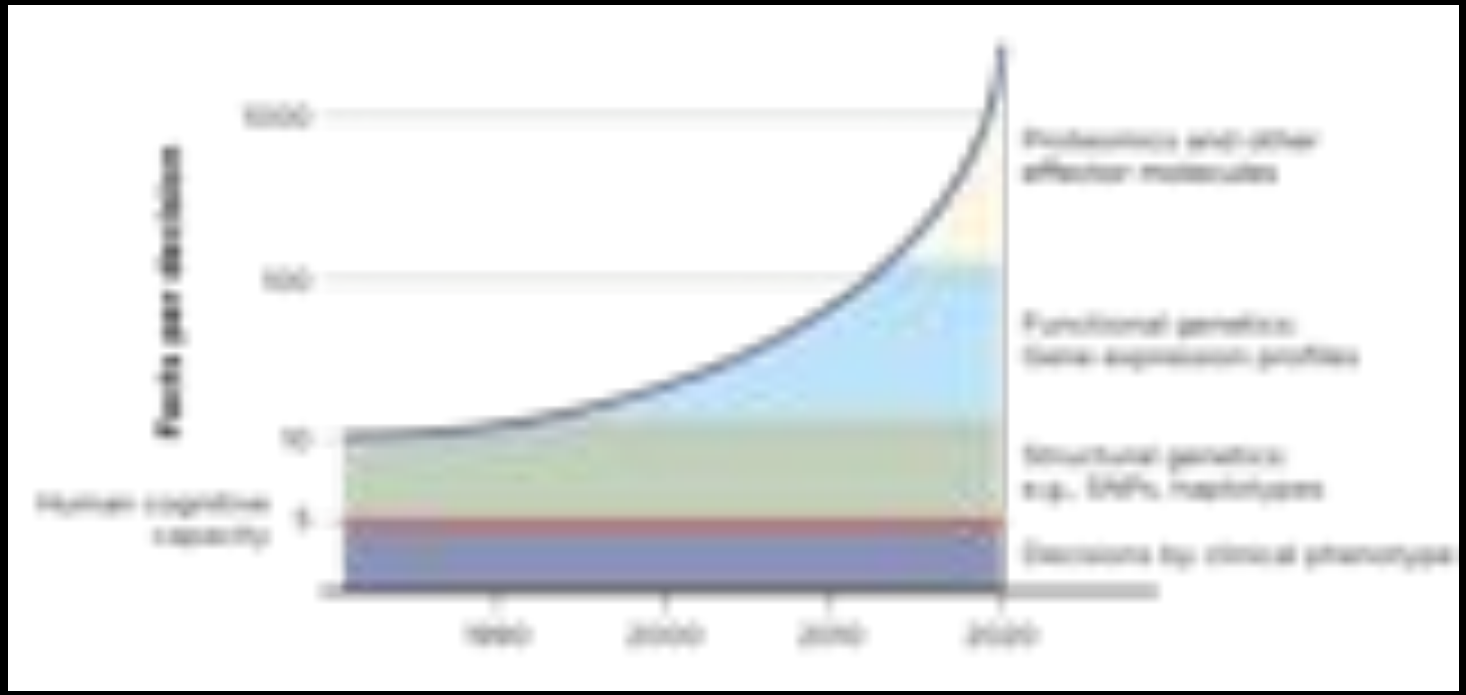
- Complexity

  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

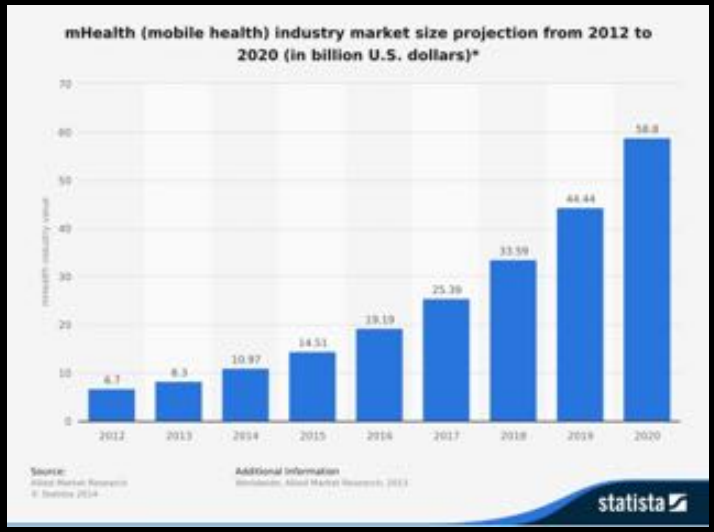
Health Care: Some Numbers

- 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) Reponses

- HITECH Act -- $15.5B to support EHRs
- Presidential Precision Health Initiative
- Big Data initiative
- Explosion of mobile health devices/apps in the consumer market



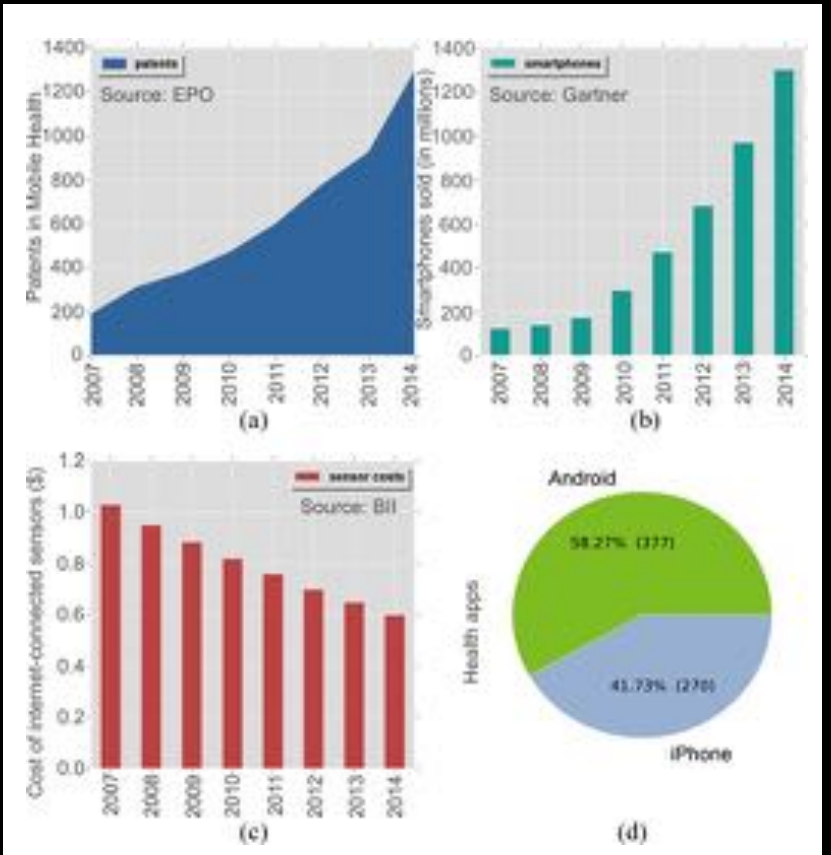


Fig. 6. a) Evolution of the number of patents published in the area of mobile health (source: European Patent Office); b) evolution of the number of smartphones sold per year in million units (source: Gartner); c) evolution of the cost of Internet-enabled sensors in dollars (source: Business Intelligence International); d) number of mobile health apps published in Google play and iTunes as of May 2015.

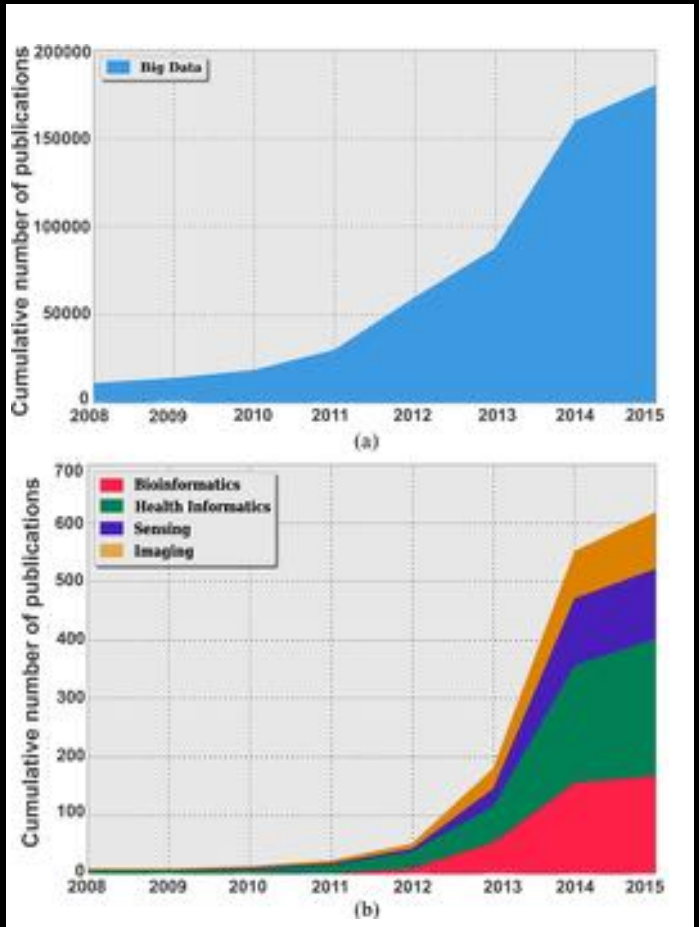


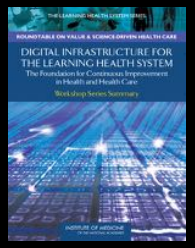
Fig. 1. (a) Cumulative number of publications referring to "big data" indexed by Google Scholar. (b) Cumulative number of publications per health research area referring to "big data," as indexed in IEEE Xplore, ACM Digital library, PubMed (National Library of Medicine, Bethesda, MD), Web of Science, and Scopus.

Big Data for Health, Andreu-Perez et al. Biomedical and Health Informatics, 2015

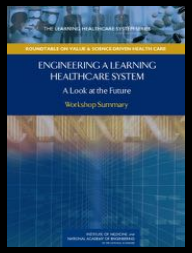
The Engineering Opportunity Space

The Learning Health System Series

Continuous improvement and innovation in health and health care



DIGITAL PLATFORM



SYSTEMS ENGINEERING

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

Executive Office of the President

President's Council of Advisors on Science and Technology

National Science Foundation
WHERE DISCOVERIES BEGIN

Directorate for Computer & Information Science & Engineering

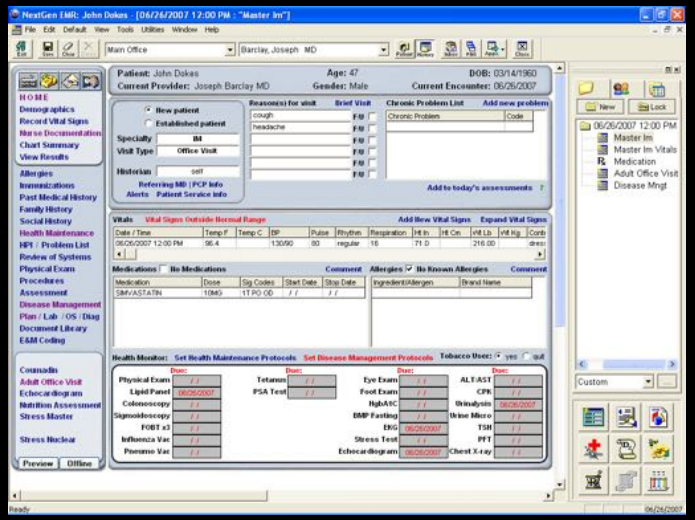
SMART HEALTH AND WELLBEING (SHW)

CONTACTS

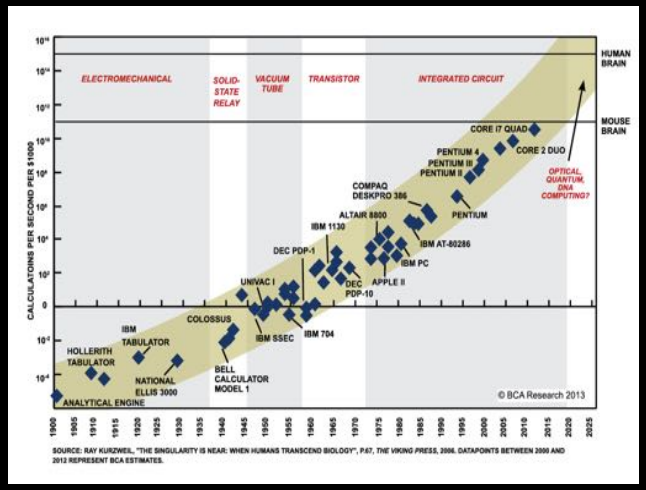
See program guidelines for contact information.

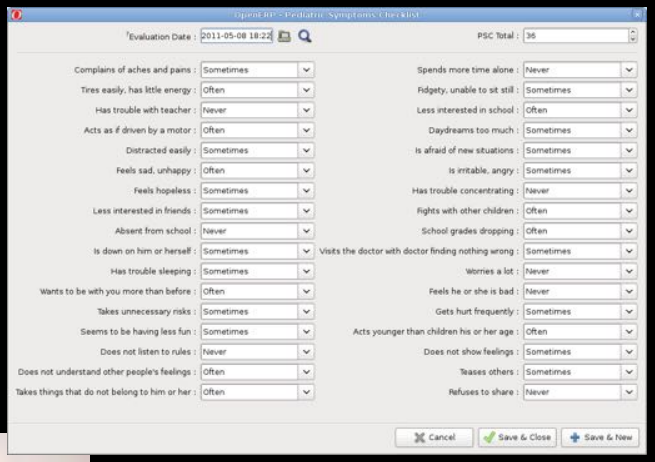
SYNOPSIS

What Blocks the Path to Innovation?




Geolocation

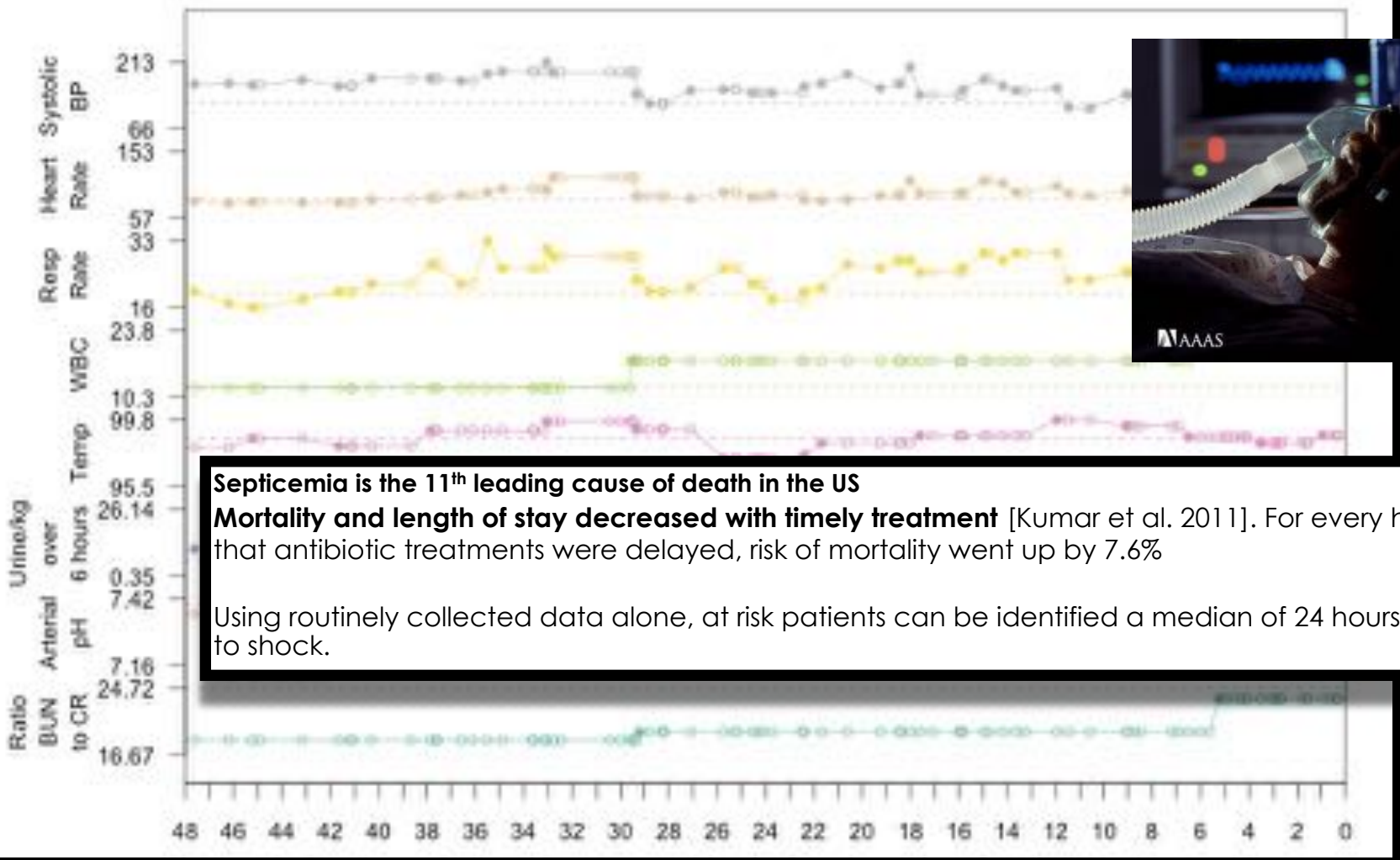

ELECTROMECHANICAL
SOLID-STATE RELAY
VACUUM TUBE
TRANSISTOR
INTEGRATED CIRCUIT
HUMAN BRAIN
MOUSE BRAIN
OPTICAL, QUANTUM, DNA COMPUTING?
CALCULATIONS PER SECOND PER $1000
ANALYTICAL ENGINE
HOLLERITH TABULATOR
IBM TABULATOR
NATIONAL ELLIS 3000
BELL CALCULATOR MODEL 1
COLOSSUS
IBM SSEC
IBM 704
UNIVAC I
DEC PDP-1
DEC PDP-10
IBM 1130
ALTAIR 8800
APPLE II
IBM PC
IBM AT-80286
COMPAQ DESKPRO 386
PENTIUM
PENTIUM II
PENTIUM III
PENTIUM 4
CORE 2 DUO
CORE I7 QUAD
© BCA Research 2013
SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, THE VIKING PRESS, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.


OpenEMR: Pediatric Symptoms Checklist
Evaluation Date : 2011-05-08 18:22
PSC Total : 36
Complains of aches and pains : Sometimes
Tires easily, has little energy : Often
Has trouble with teacher : Never
Acts as if driven by a motor : Often
Distracted easily : Sometimes
Feels sad, unhappy : Often
Feels hopeless : Sometimes
Less interested in friends : Sometimes
Absent from school : Never
Is down on him or herself : Sometimes
Has trouble sleeping : Sometimes
Wants to be with you more than before : Often
Takes unnecessary risks : Sometimes
Seems to be having less fun : Sometimes
Does not listen to rules : Never
Does not understand other people's feelings : Often
Takes things that do not belong to him or her : Often
Spends more time alone : Never
Fidgety, unable to sit still : Sometimes
Less interested in school : Often
Daydreams too much : Sometimes
Is afraid of new situations : Sometimes
Is irritable, angry : Sometimes
Has trouble concentrating : Never
Fights with other children : Often
School grades dropping : Often
Visits the doctor with doctor finding nothing wrong : Sometimes
Worries a lot : Never
Feels he or she is bad : Never
Gets hurt frequently : Sometimes
Acts younger than children his or her age : Often
Does not show feelings : Sometimes
Teases others : Sometimes
Refuses to share : Never
Cancel
Save & Close
Save & New

Three Examples

- Assessing OR performance -- Hager
- Data to diagnosis → Saria
- Data in the wild to detection → Ferry
- IT → manage behavior -- Mynatt

Early Warning System for Sepsis



Septicemia is the 11th leading cause of death in the US

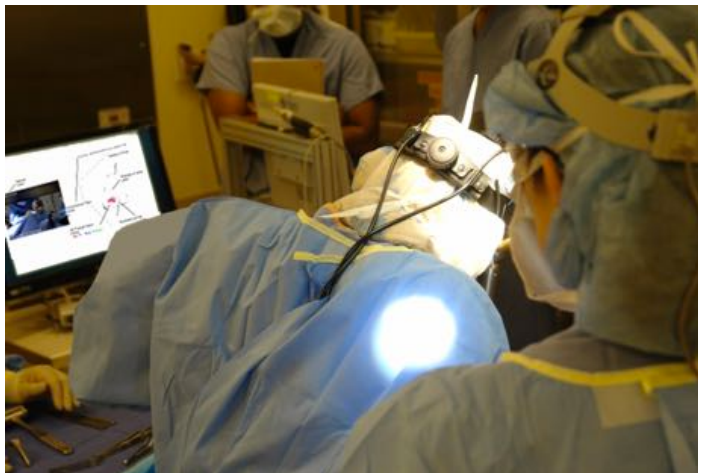
Mortality and length of stay decreased with timely treatment [Kumar et al. 2011]. For every hour that antibiotic treatments were delayed, risk of mortality went up by 7.6%

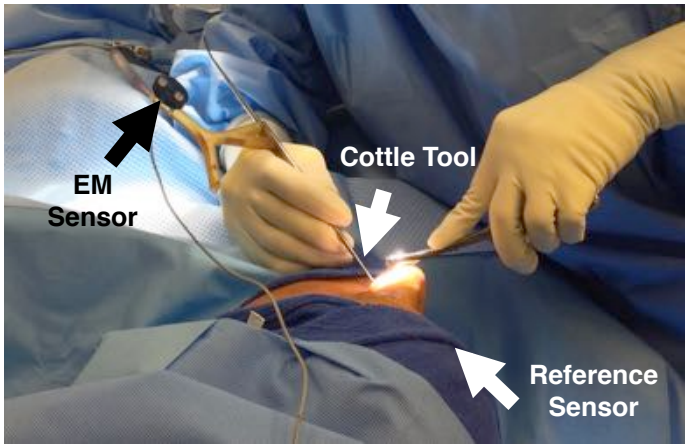
Using routinely collected data alone, at risk patients can be identified a median of 24 hours prior to shock.

K. Henry, D. Hager, P. Pronovost, S. Saria. A Targeted Real-time Early Warning Score (TREWScore) for Septic Shock. *Science Translational Medicine 2015. Vol. 7, Issue 299.*

K. Dyagilev, S. Saria. Learning (Predictive) Risk Scores in the Presence of Censoring due to Interventions. *Machine Learning, 2015 (in print).*

Measurement in the OR





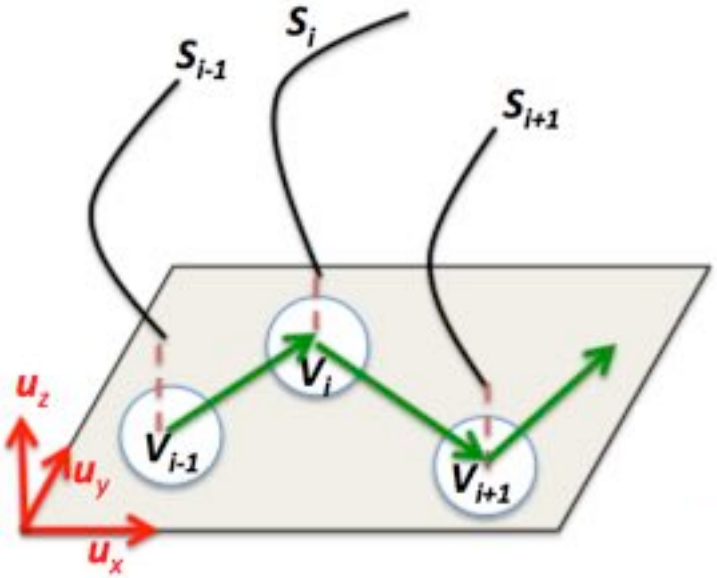
Data from **5 Johns Hopkins Hospital sites (2012-2014)**

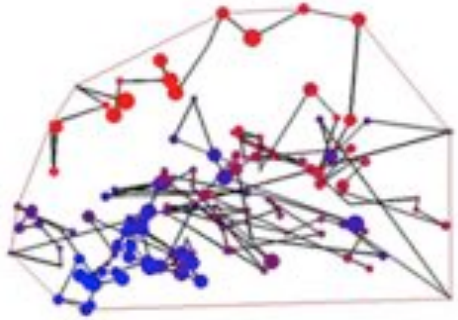
Native Septoplasty:

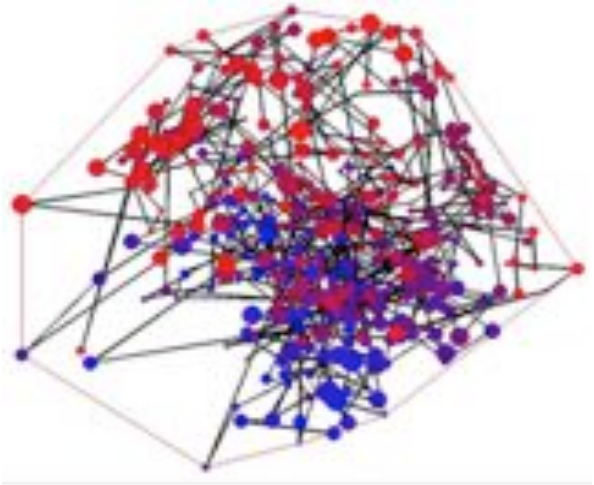
-- 5 Experts (60 trials)
-- 9 Novices (26 trials)
-- Multi-surgeons (14 trials)

Joint work with Masaru Ishii, MD, PhD and Lisa Ishii MD

From Data to Structure







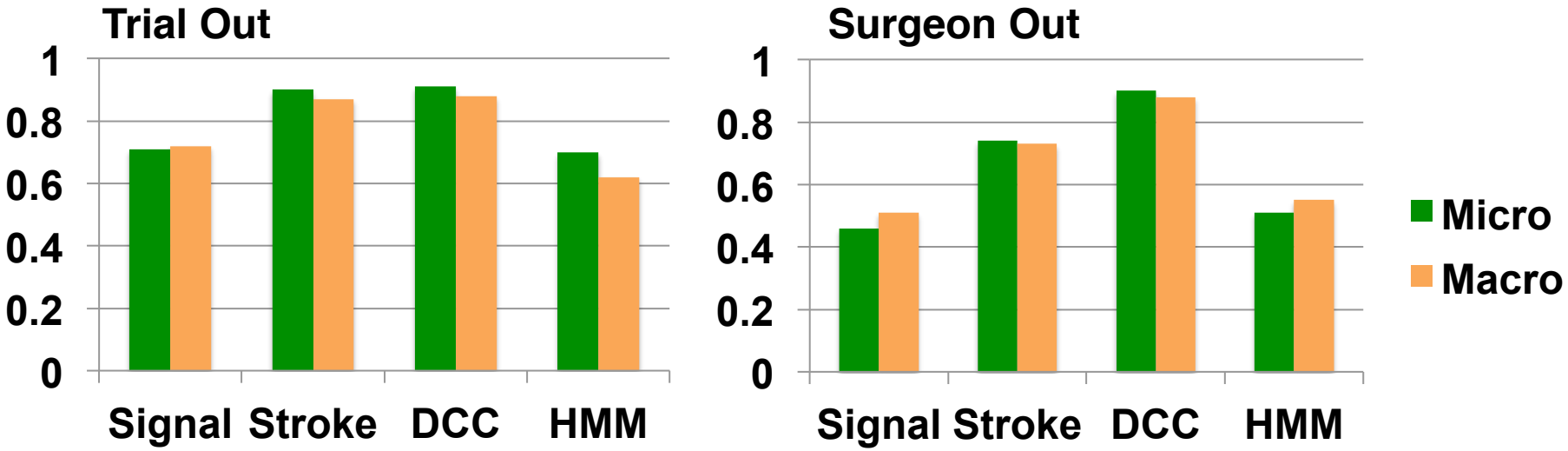
EXPERT **NOVICE**

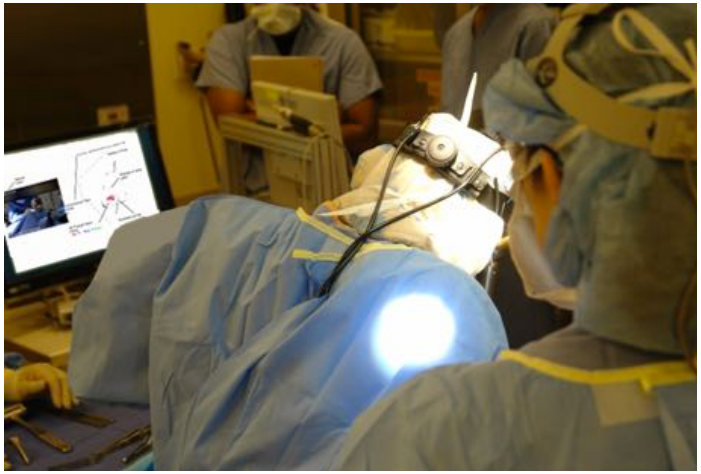
Start of Surgery

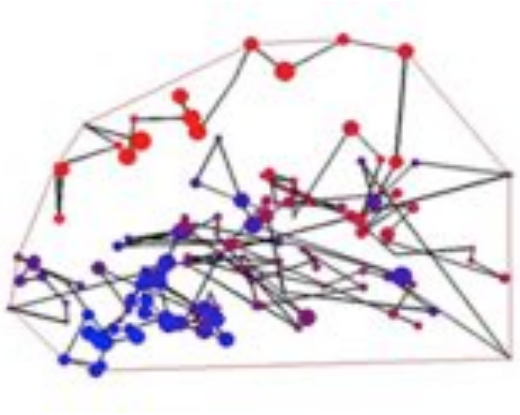
End of Surgery

Stroke Height

From Structure to Assessment

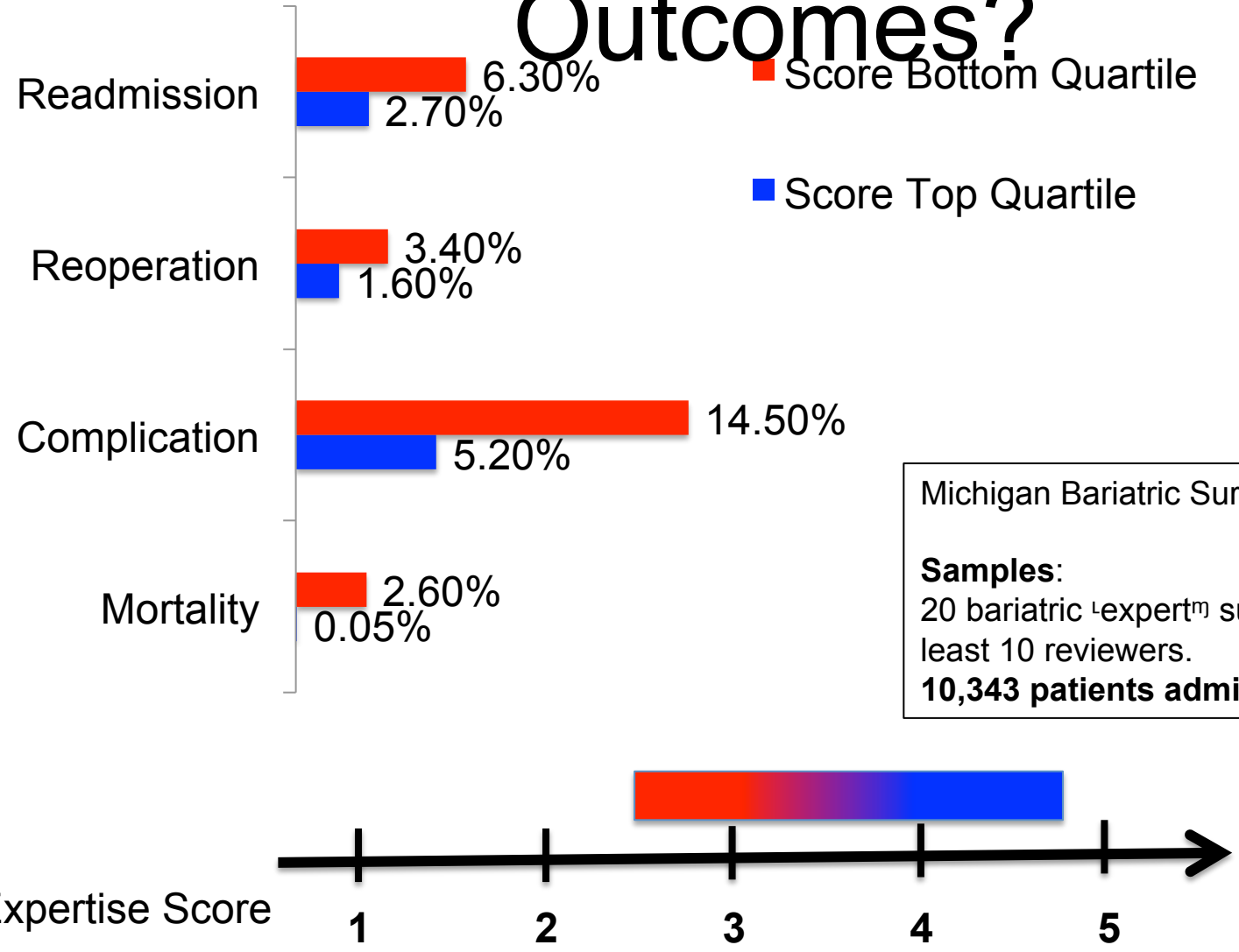






[DCC] Ahmidi N, et al.: String Motif-Based Description of Tool Motion for Detecting Skill and Gestures in Robotic Surgery. MICCAI (2013)
[HMM] Varadarajan B, et al.: Data-derived models for segmentation with application to surgical assessment and training. MICCAI (2009)

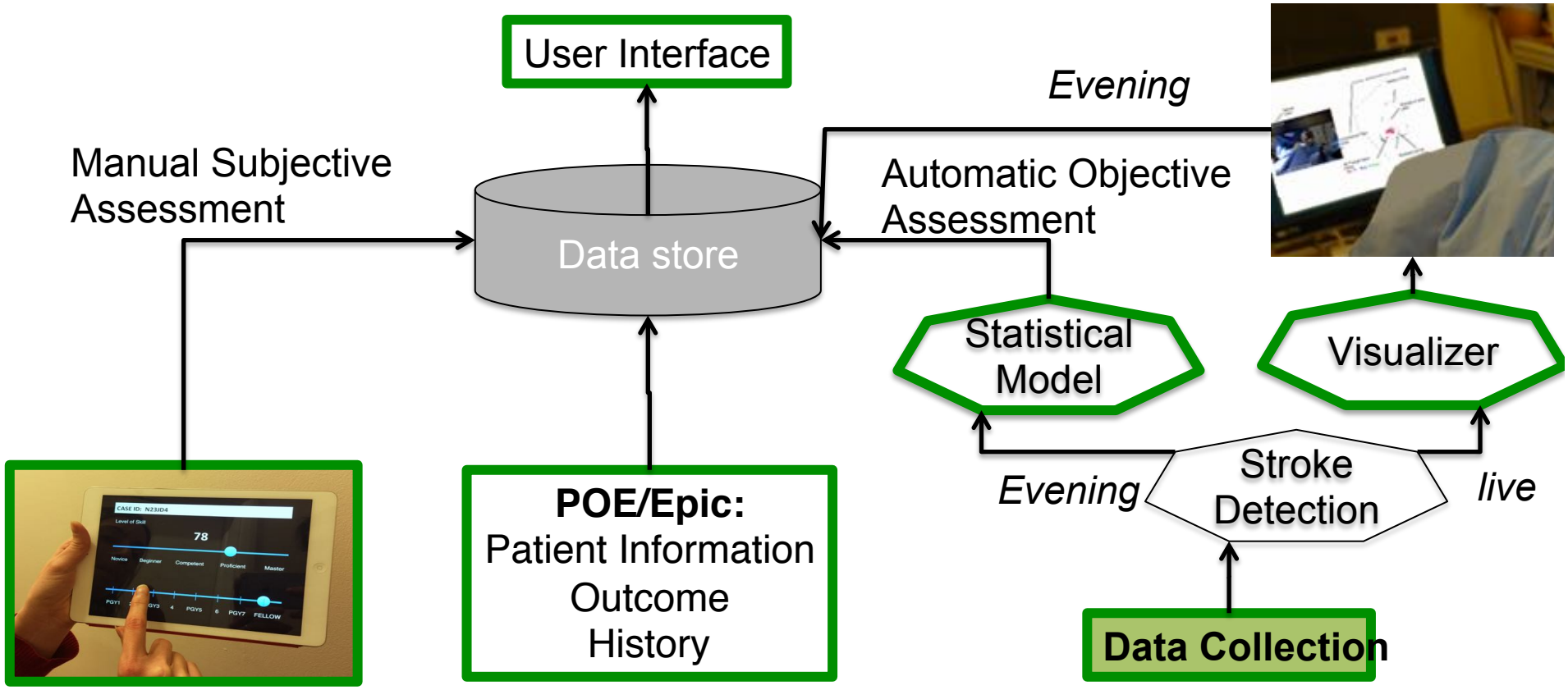
From Assessment to Outcomes?



Michigan Bariatric Surgery Collaborative

Samples:
20 bariatric ʟexpertᵐ surgeons ranked by at least 10 reviewers.
10,343 patients admitted 2006-2012

Birkmeyer J.D, et al. Surgical Skill and Complication Rates after Bariatric Surgery. NEJM, 2013.

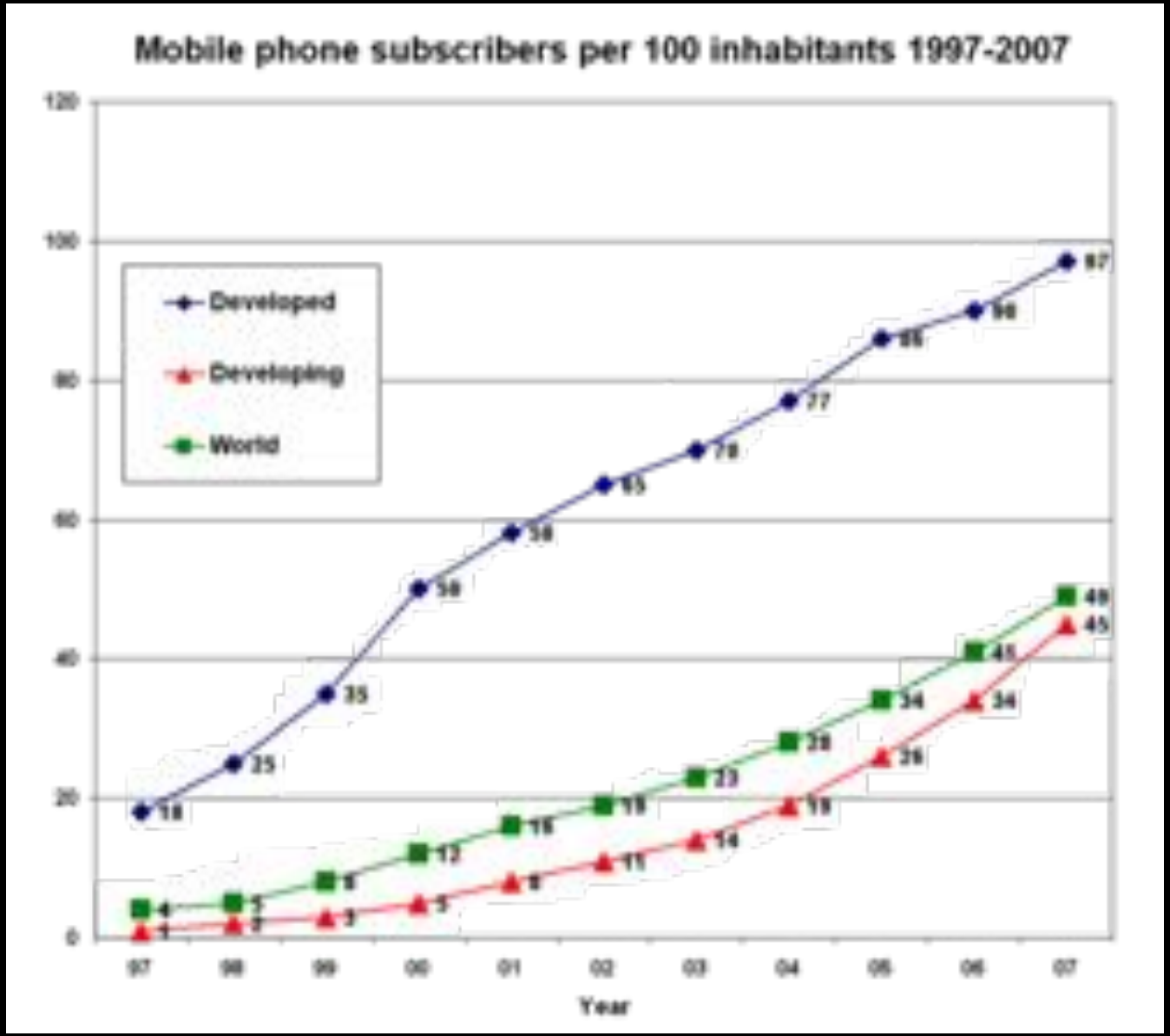


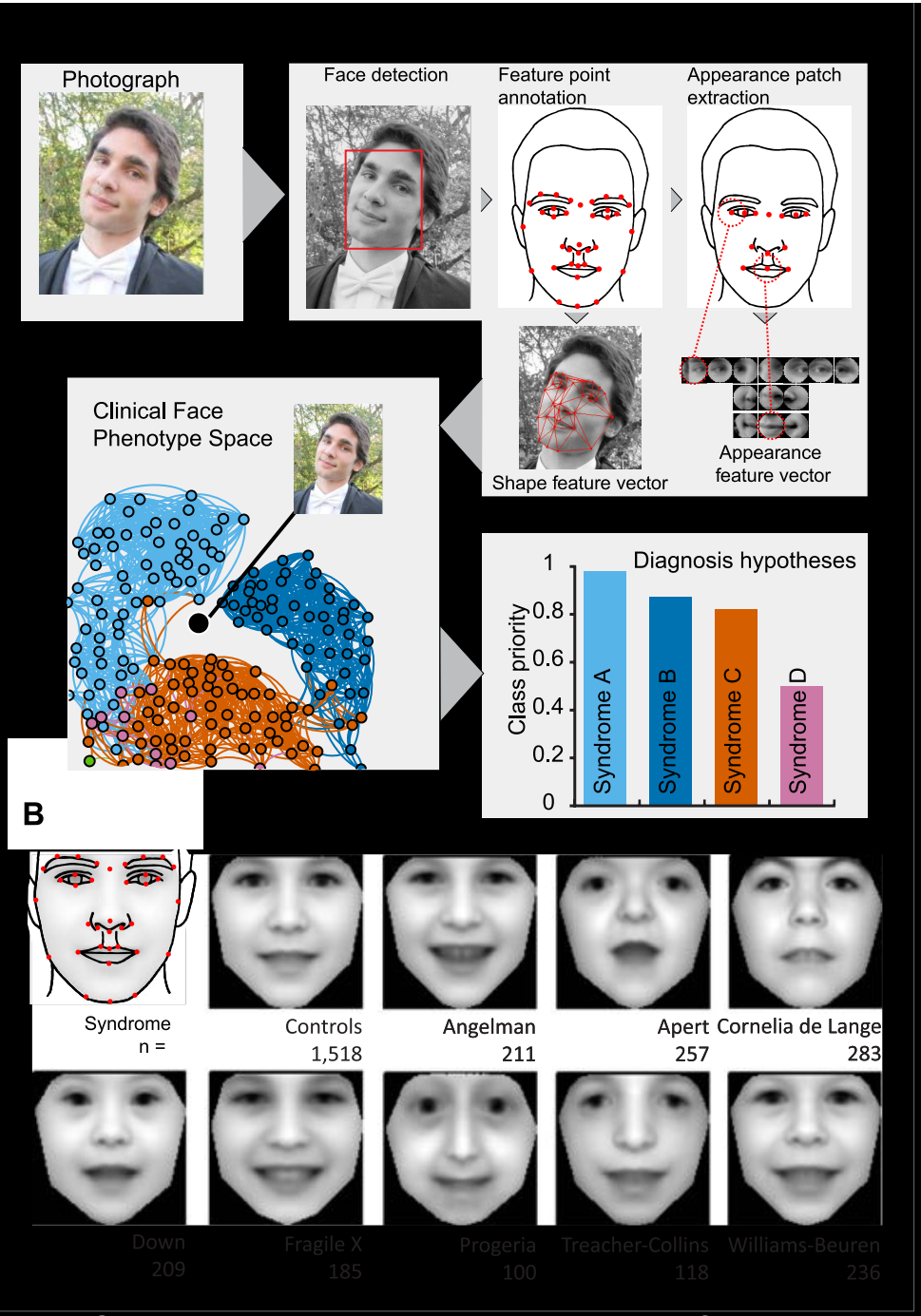
Trainee/Faculty Log

Case ID:
Traineeʝs ID:
Supervisorʝs ID:
Assessment of Supervisor: Notes, comments, scores
Automatic Assessment: Scores, Feedback

Outcome/case difficulty

Case ID:
Date and Location:
Patientʝs ID:
Patient Medical History:
Patient Outcome: Interval assessme


Mobile phone subscribers per 100 inhabitants 1997-2007
Developed
Developing
World
Year



Ferry, Quentin, et al. "Diagnostically relevant facial gestalt information from ordinary photos." *Elife* 3 (2014): e02020

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/

Georgia Tech

INSTITUTE for
PEOPLE and TECHNOLOGY

The convergence of people and technology

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.

Georgia Tech

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

Institute for People and Technology

TRANSFORMING HEALTHCARE

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

Rowan, Jim, and Elizabeth D. Mynatt. "Digital family portrait field trial: Support for aging in place." *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2005.

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

Institute for People and Technology

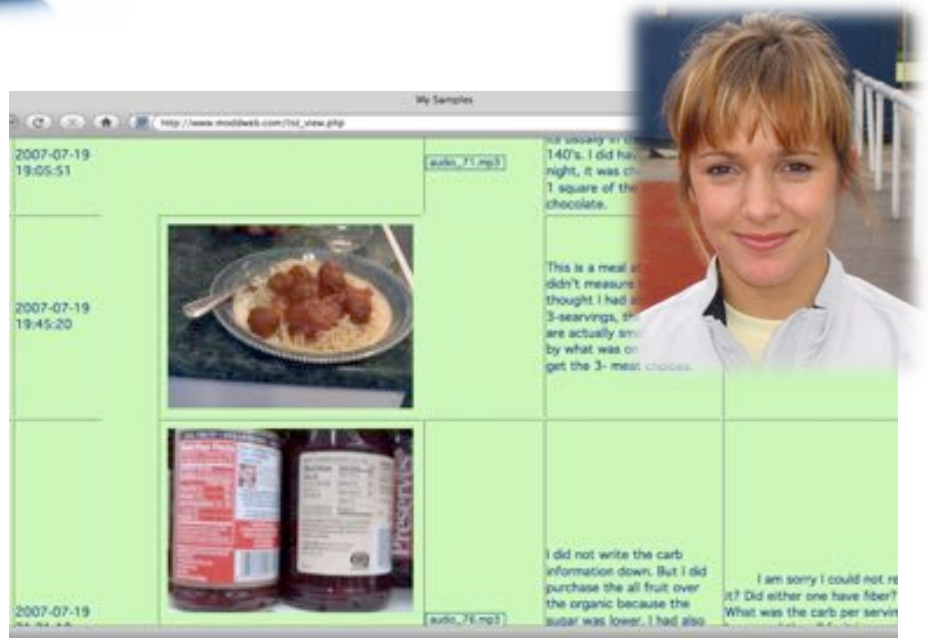
TRANSFORMING HEALTHCARE

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



Mamykina, Lena, et al. "MAHI: investigation of social scaffolding for reflective thinking in diabetes management." CHI-CONFERENCE-. Vol. 1. ACM INC, 2008.

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

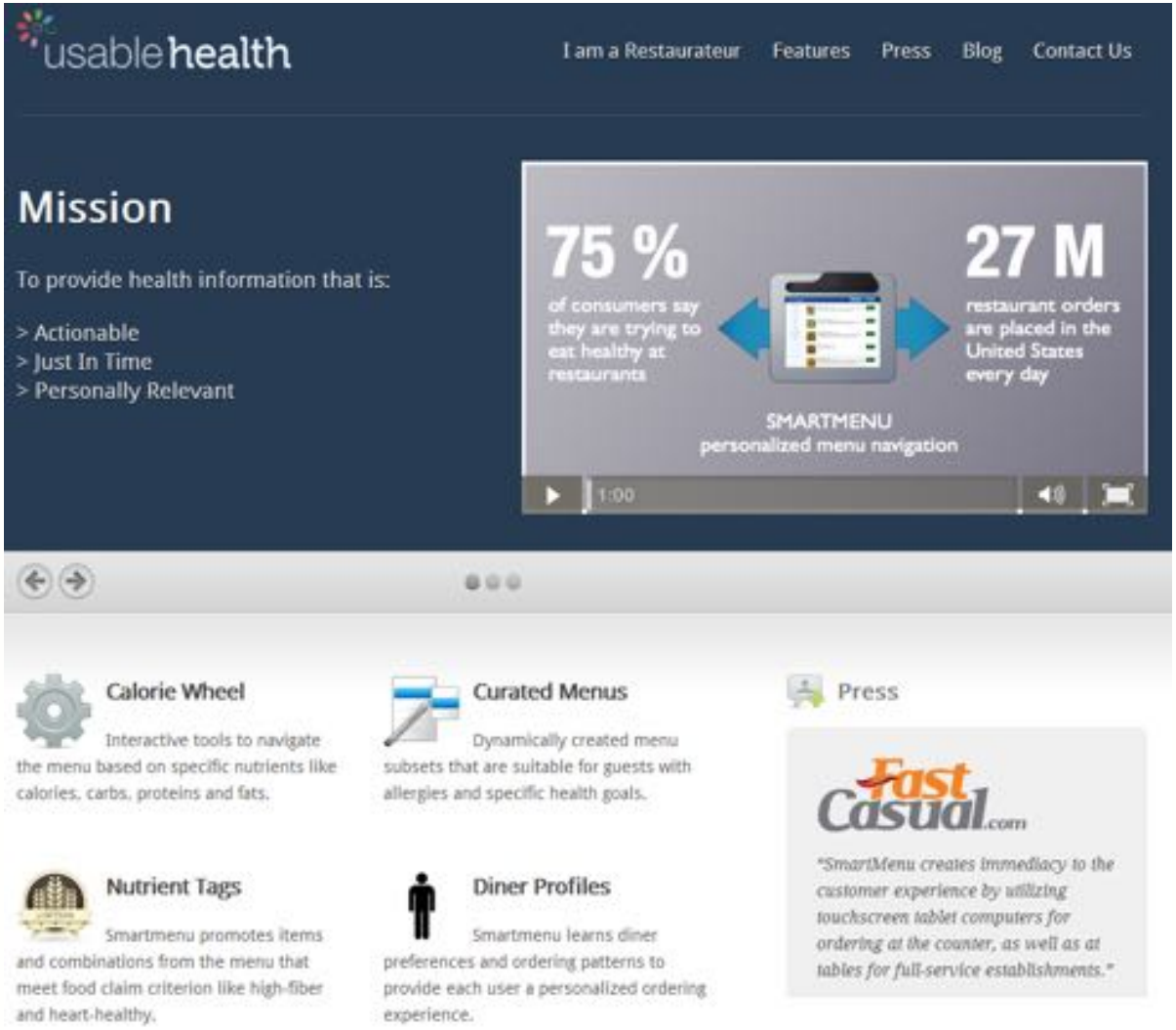
INSTITUTE for PEOPLE and TECHNOLOGY

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

Miller, A. D., Poole, E., Xu, Y., Eiriksdottir, E., Kestranek, D., Catrambone, R., & Mynatt, E. (2012). **The work of play: supporting a pervasive health behavior change intervention for us middle school students**. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12)*. Seattle, WA.

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*

Georgia Tech

INSTITUTE for PEOPLE and TECHNOLOGY

Adolescents & Physical Activity

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

- Identity
- Gaming
- Offline behavior

Miller, A.D., Pater, J.A., Mynatt, E.D. "Design Strategies for Youth-Focused Pervasive Social Health Games," 7th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2013),5-8 May 2013

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

Miller, A.D., Mynatt, E.D. "StepStream: A School-based Pervasive Social Fitness System for EverydayAdolescentHealth."Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2014).

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.

Jacobs, M., Clawson, J., Mynatt, E.D. 2014. "Cancer Navigation: Opportunities and Challenges for Facilitating the Breast Cancer Journey," Proceedings of the ACM 2014 conference on Computer Supported Cooperative Work (CSCW 2014),

Jacobs, M., Clawson, J., Mynatt, E.D. 2014. "MJC: A Preliminary Investigation of a Mobile Tool for Cancer Patients," in Proceedings of the 32nd International Conference on Human Factors in Computing Systems (CHI '14)

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
  - http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1072229&download=yes
- Netflix prize
- Facebook
- Gmail
  - 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/

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? ÄWhat do we (as a society) gain? What are some examples of the dangers?
- Generalize: Should information gathering and use be regulated? ÄAt what point? ÄHow would you govern it? Provide concrete examples.
- 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

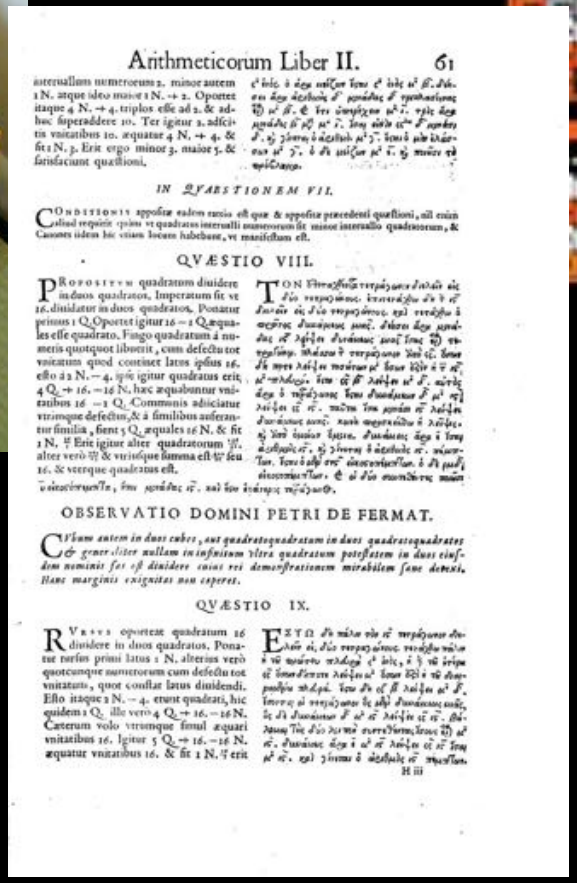
Simon 1957

- 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



Arithmeticorum Liber II. 61

IN QVÆSTIONEM VII.

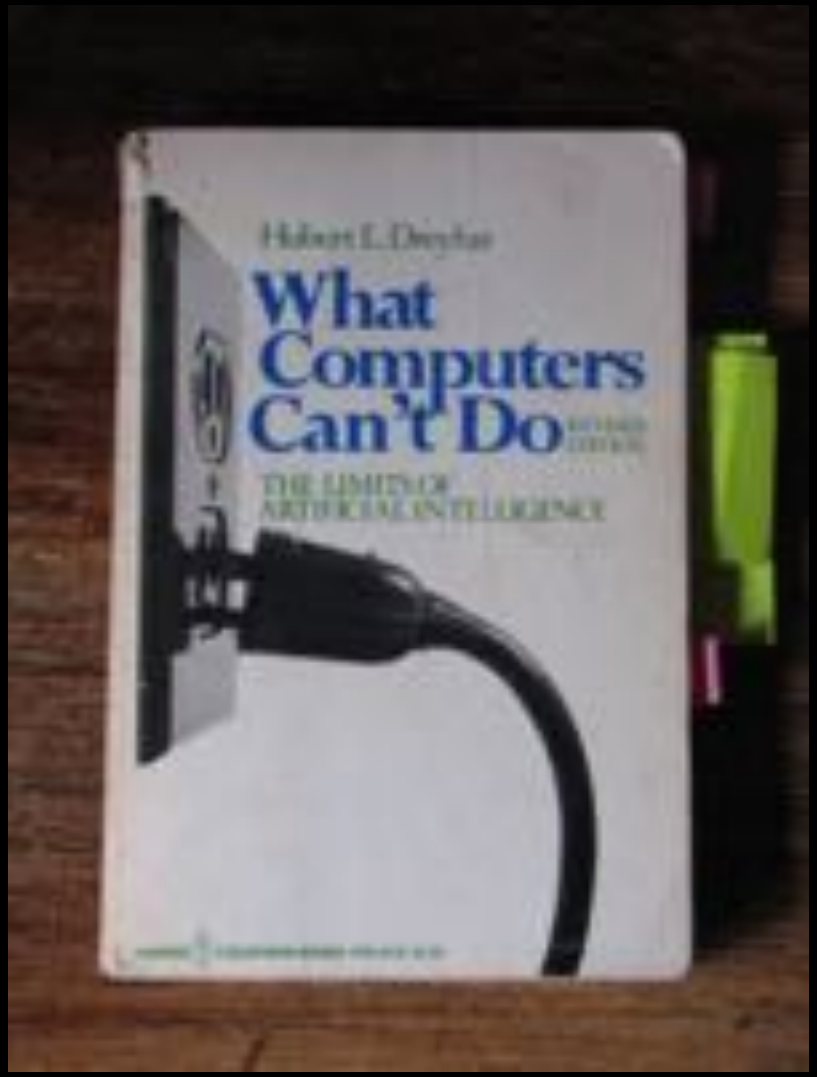
QVÆSTIO VIII.

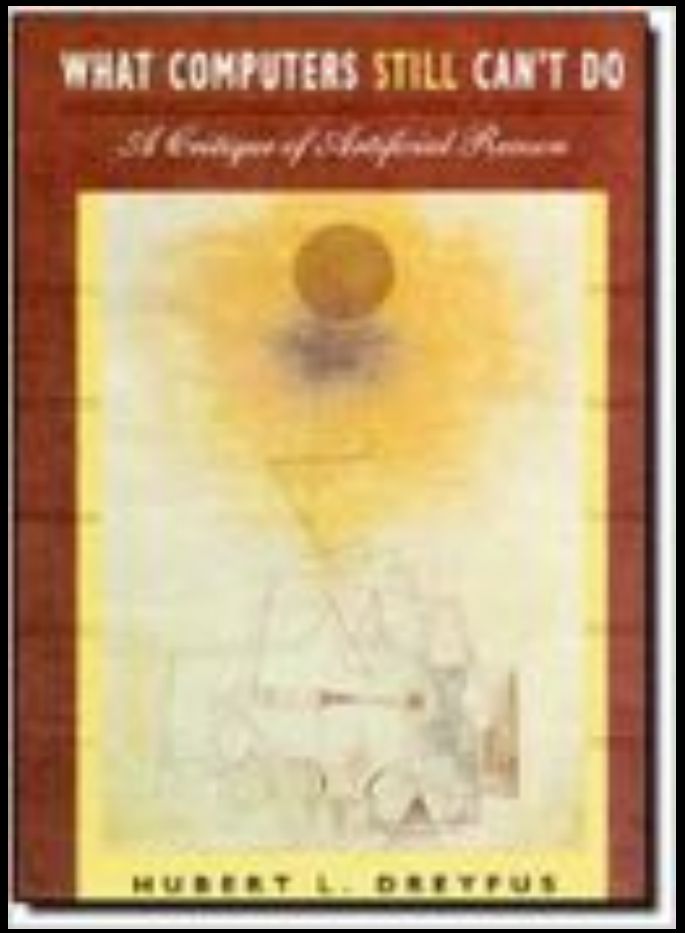
OBSERVATIO DOMINI PETRI DE FERMAT.

QVÆSTIO IX.

$A^n+B^n = C^n$ for Integers A, B, C and n>2?

Dreyfus 1972, 1992





AI Will Fail Because ŧ.

- **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

Some Questions

- Can a computer create music?
  - https://www.youtube.com/watch?v=PzrcoqpnZqA
- 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?
  - https://www.youtube.com/watch?v=PzrcoqpnZqA
- 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 ț Friona piled up the steals, swiping eight bags in all ț

http://www.narrativescience.com

Some Questions

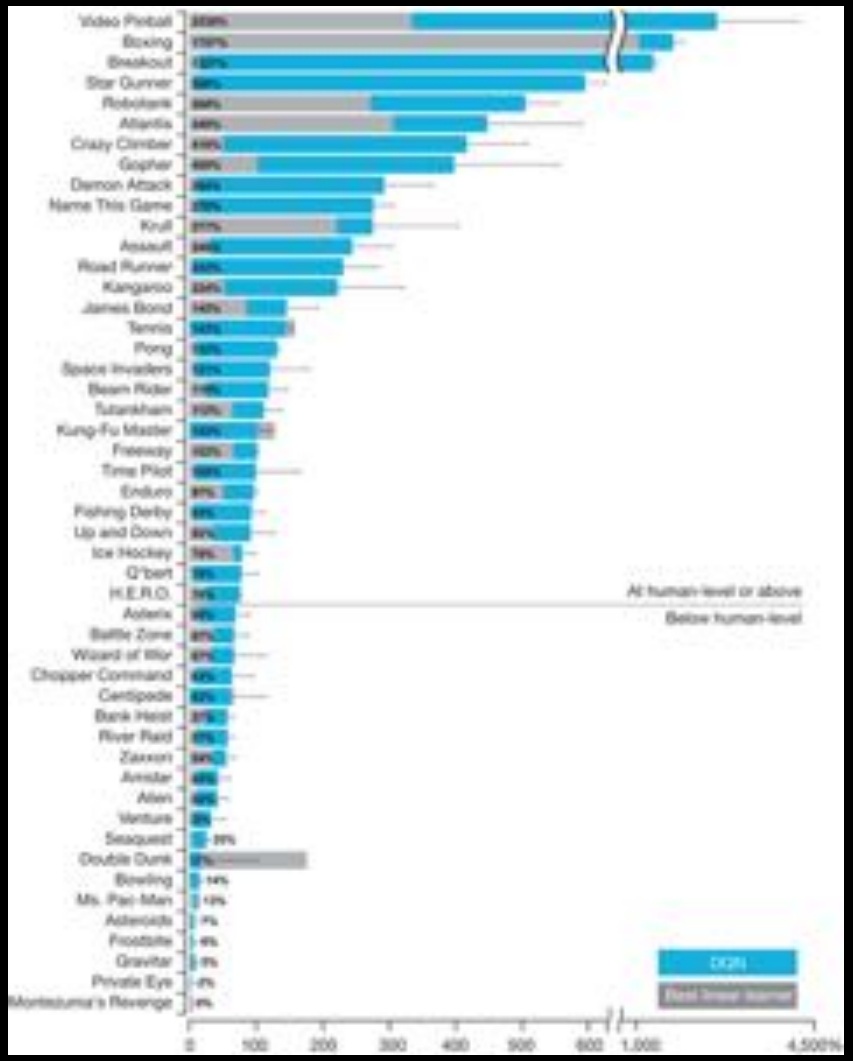
- Can a computer create music?
  - https://www.youtube.com/watch?v=PzrcoqpnZqA
- Can a computer write an article?
- Can a computer paint?

http://www.thepaintingfool.com

Is This Good? Bad? Inevitable?

- Elon Musk:
  - http://observer.com/2015/08/stephen-hawking-elon-musk-and-bill-gates-warn-about-artificial-intelligence/
- Paul Allen:
  - http://www.technologyreview.com/view/425733/paul-allen-the-singularity-isnt-near/

Two Perspectives



Mnih, Volodymyr, et al. "Human-level control through deep reinforcement learning." *Nature* 518.7540 (2015): 529-533.

Two Perspectives

What Is Our Future? ʊ How Will We Shape It?

Is Computing the ʊ future of thought ʊ and discourse?

Is it the beginning of the end of Computing asʊ we know it?

Is Computing creating a new ways to combineʊ and create?