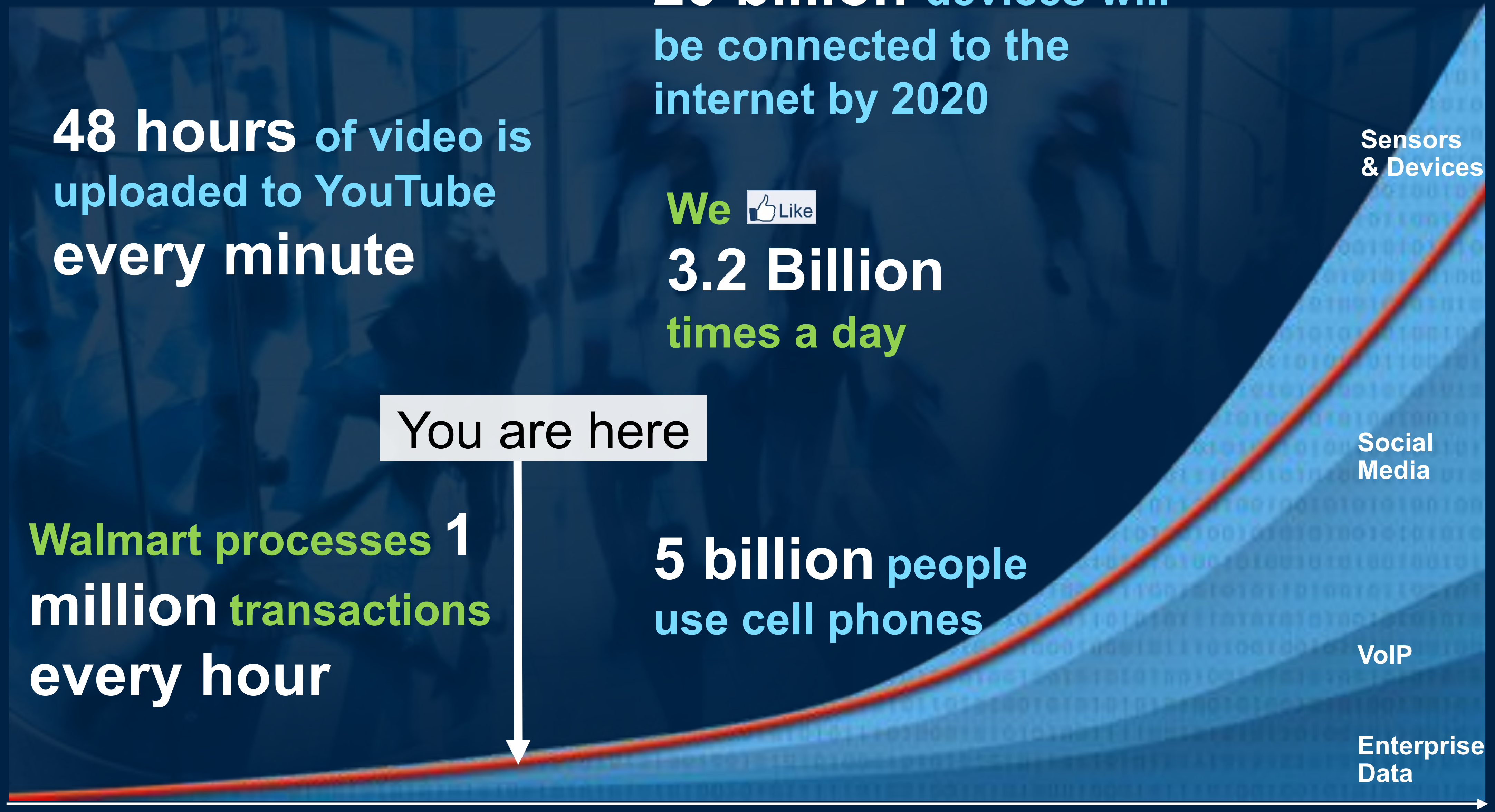
From Data to Insight to Change: Technologies and Opportunities

Laura Haas
IBM Fellow
Director, Accelerated Discovery Laboratory

Data is the fuel …



48 hours of video is uploaded to YouTube **every minute**

20 billion devices will be connected to the internet by 2020

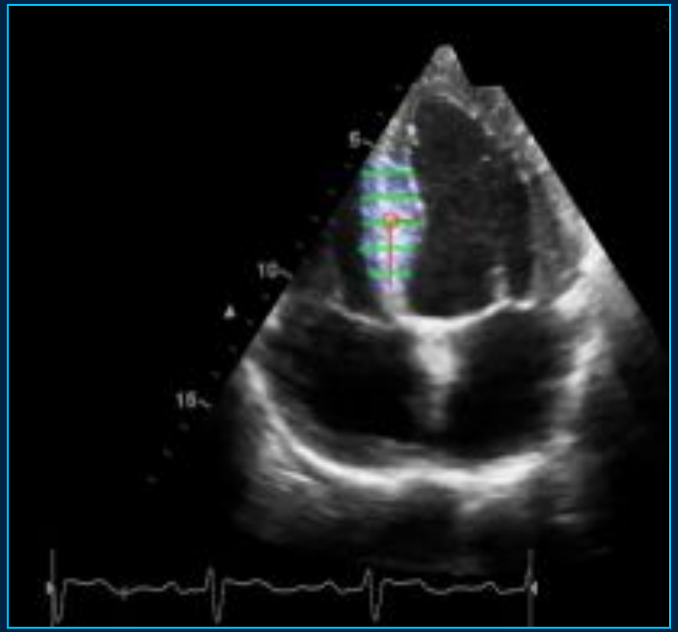
We Like **3.2 Billion** times a day

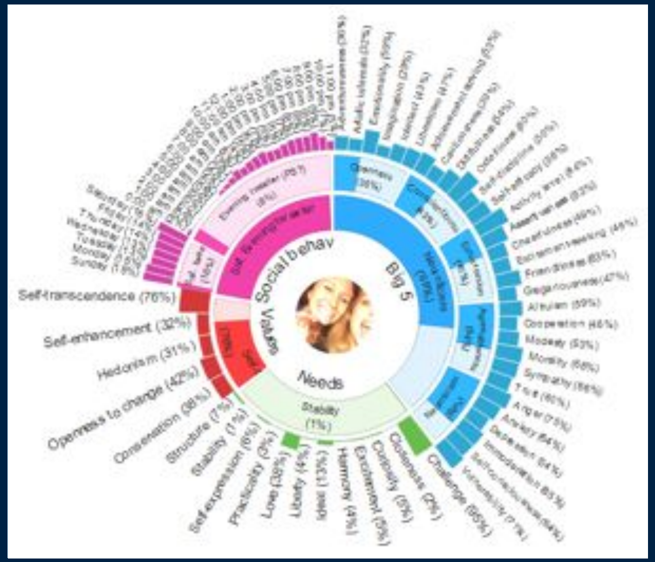
Walmart processes **1 million** transactions **every hour**

5 billion people use cell phones

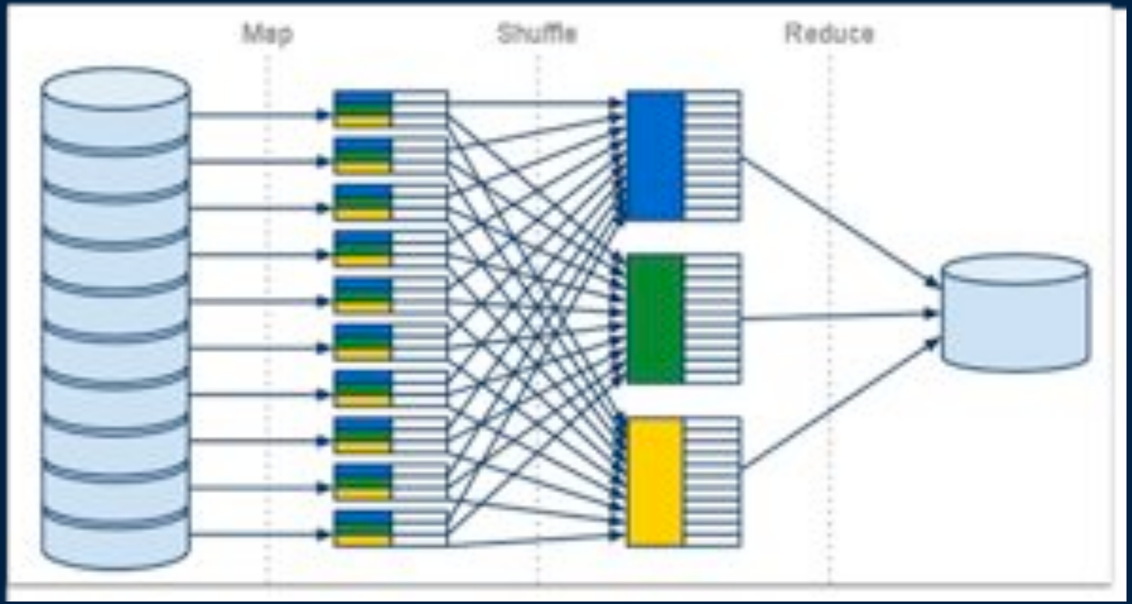
A zetabyte is **100 million** copies of the **Library of Congress**

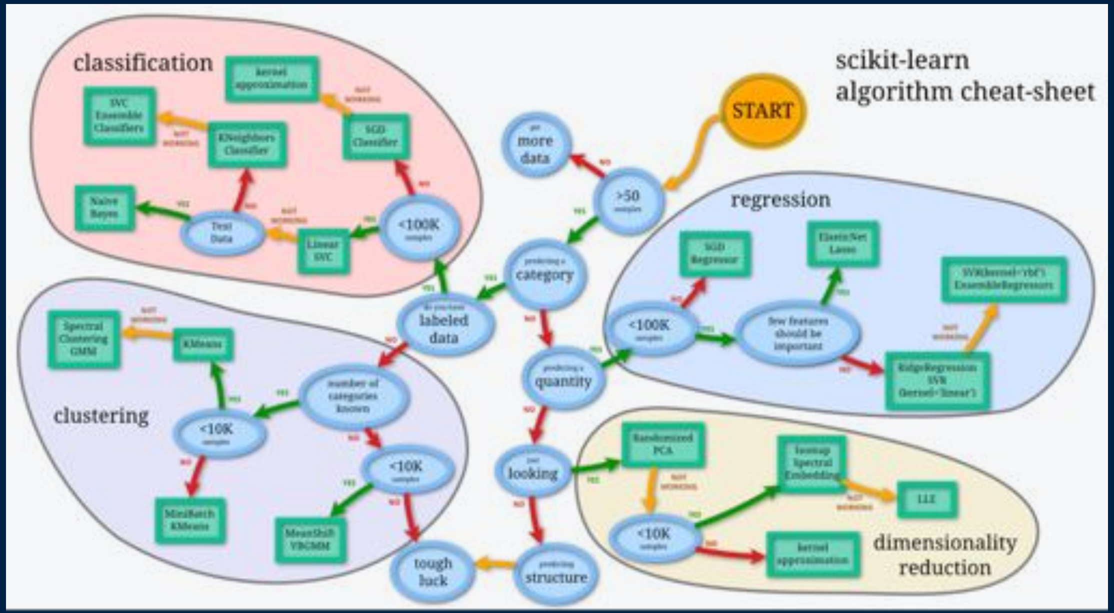
… behind social, science, government and business systems

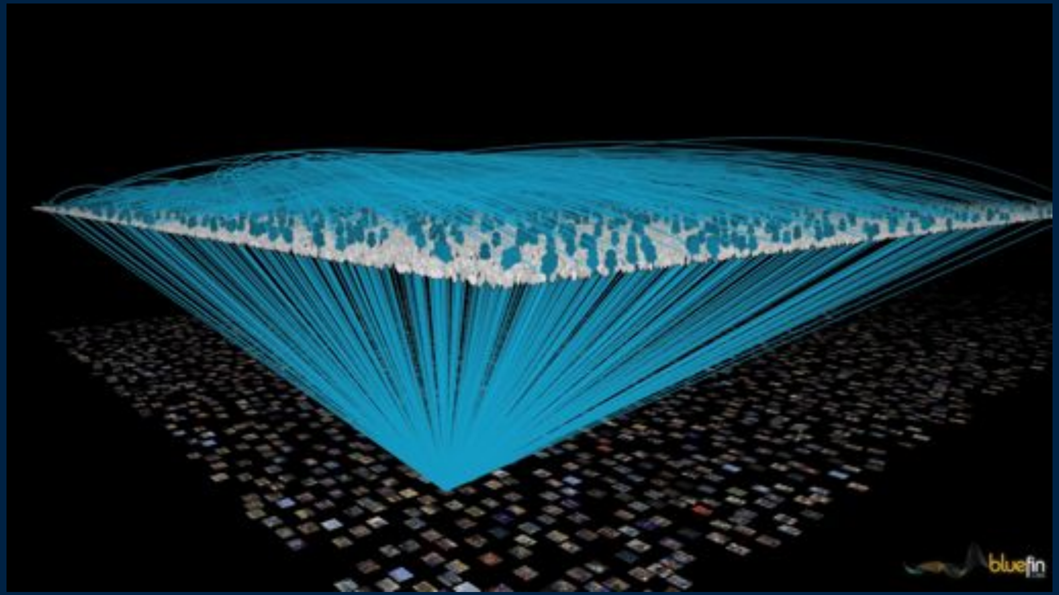


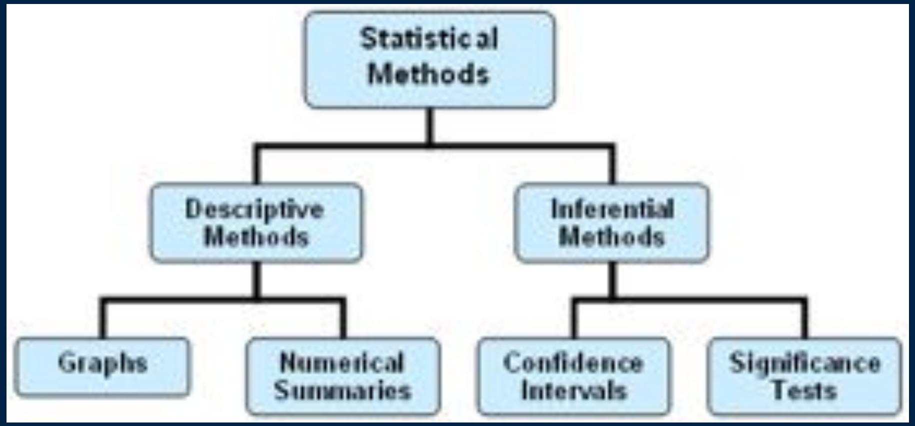


These advances are powered by a broad range of technologies









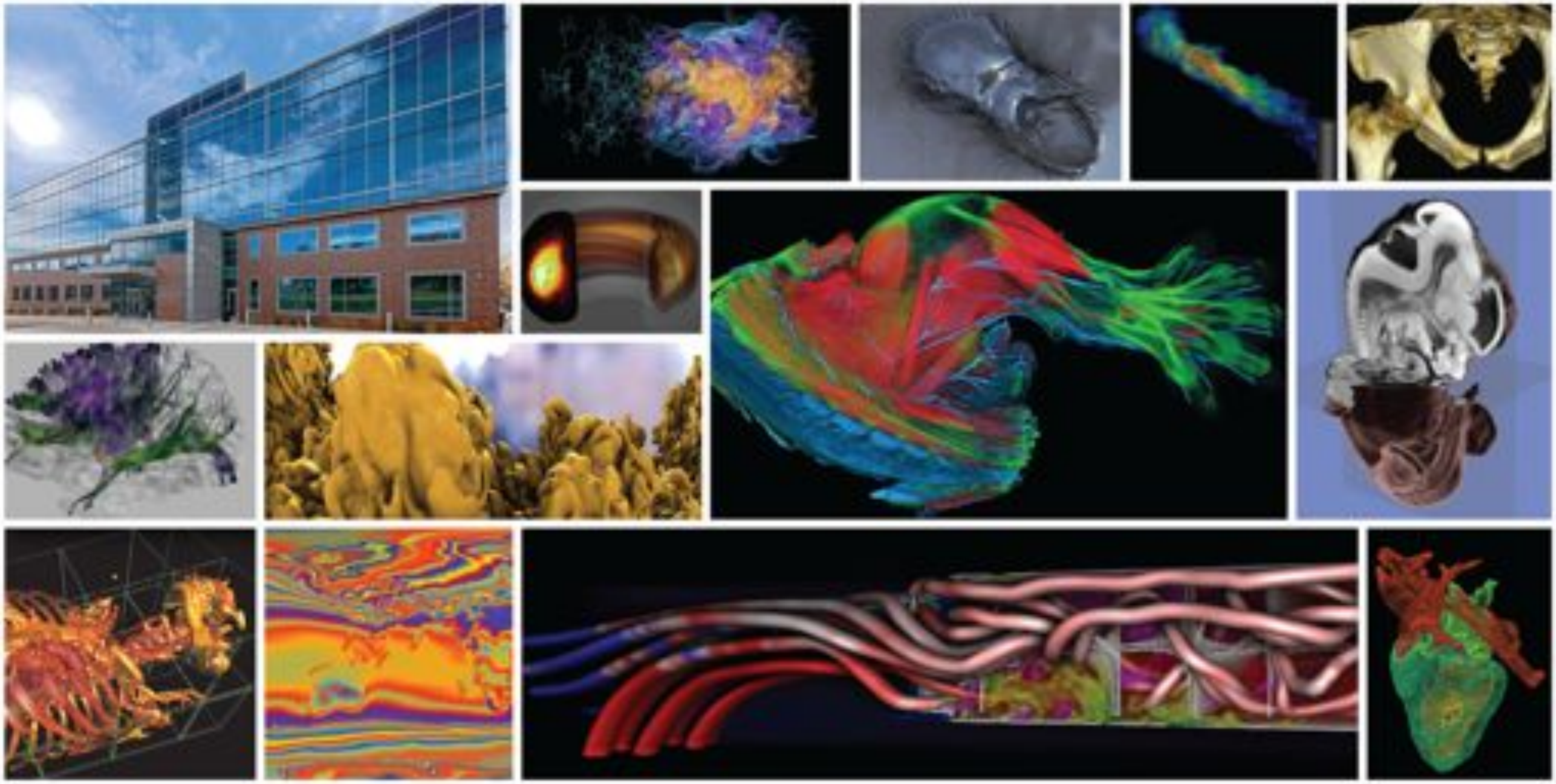
Session Plan

- Eric Horvitz, Microsoft Research
- Chris Johnson, University of Utah
- Brandon Johnson, Goldman Sachs
- Discussion

Discussion points

- What are the most exciting technology developments in “data science”?
- What is the use case that most excites you?
- Is there something “new” here?
- What should computer science’s role be?
- What do we need to do to prepare students for this brave new data-centric world?

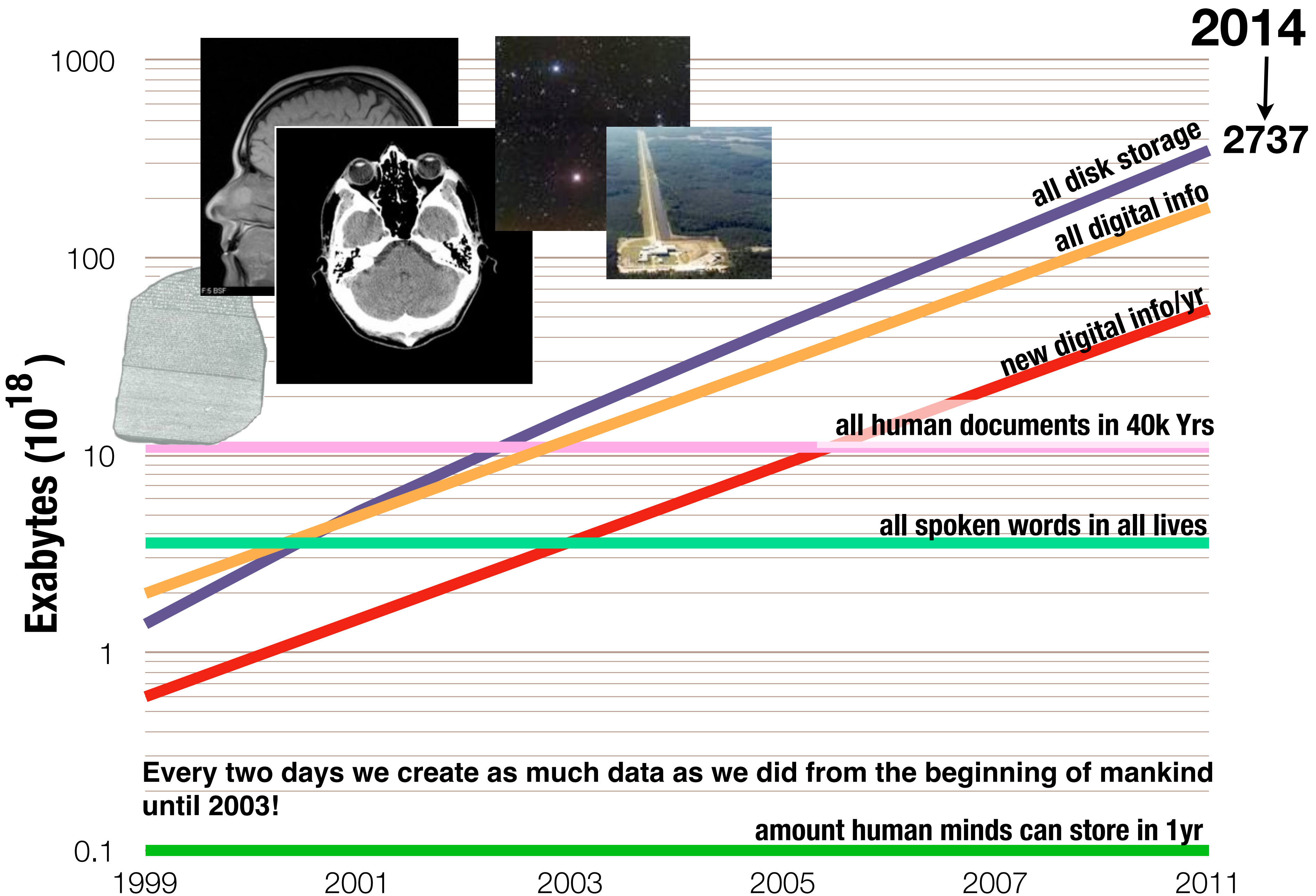
Data to Insight to Change



Chris Johnson

Scientific Computing and Imaging Institute

University of Utah



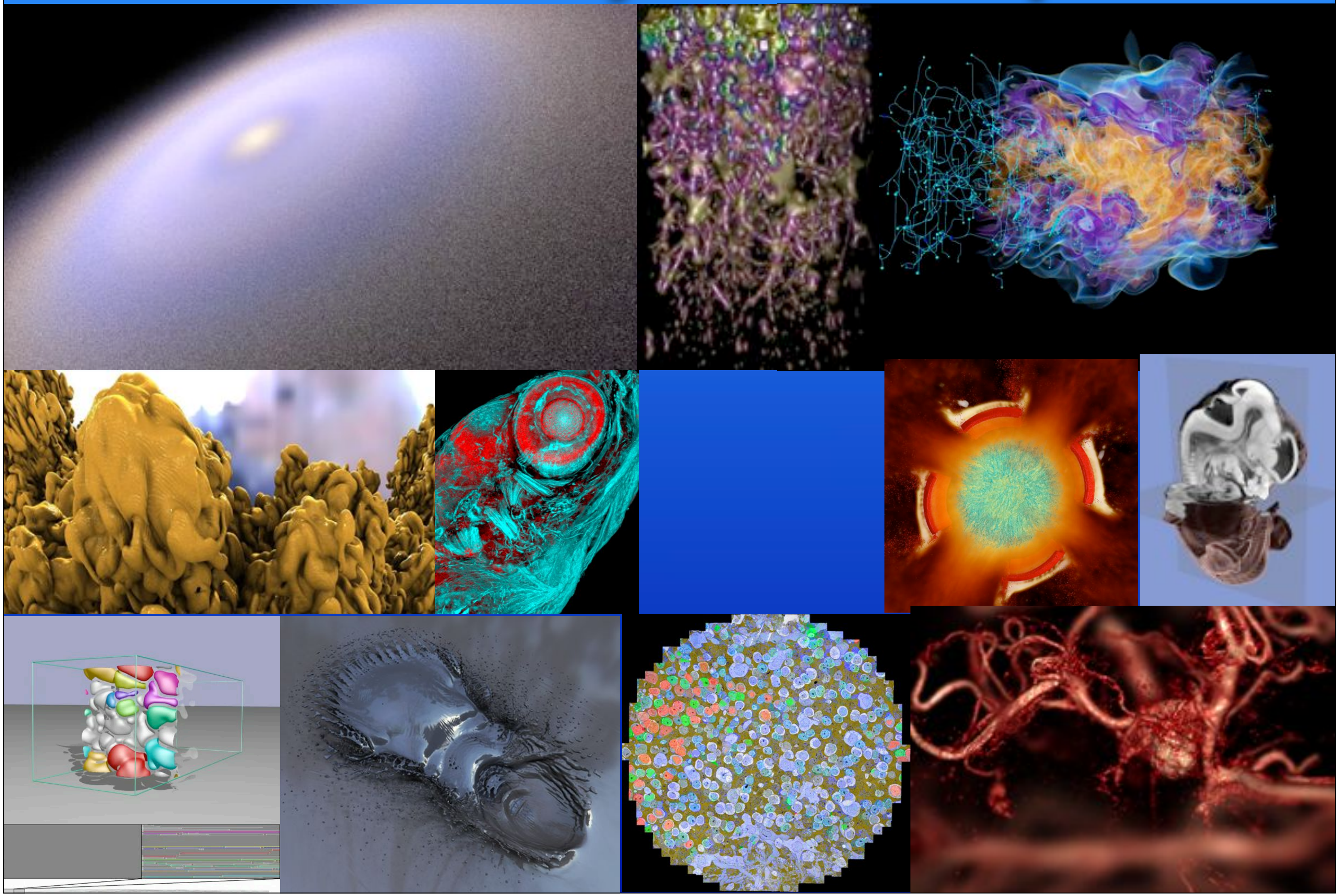
Sources: Lesk, Berkeley SIMS, Landauer, EMC, TechCrunch, Smart Planet

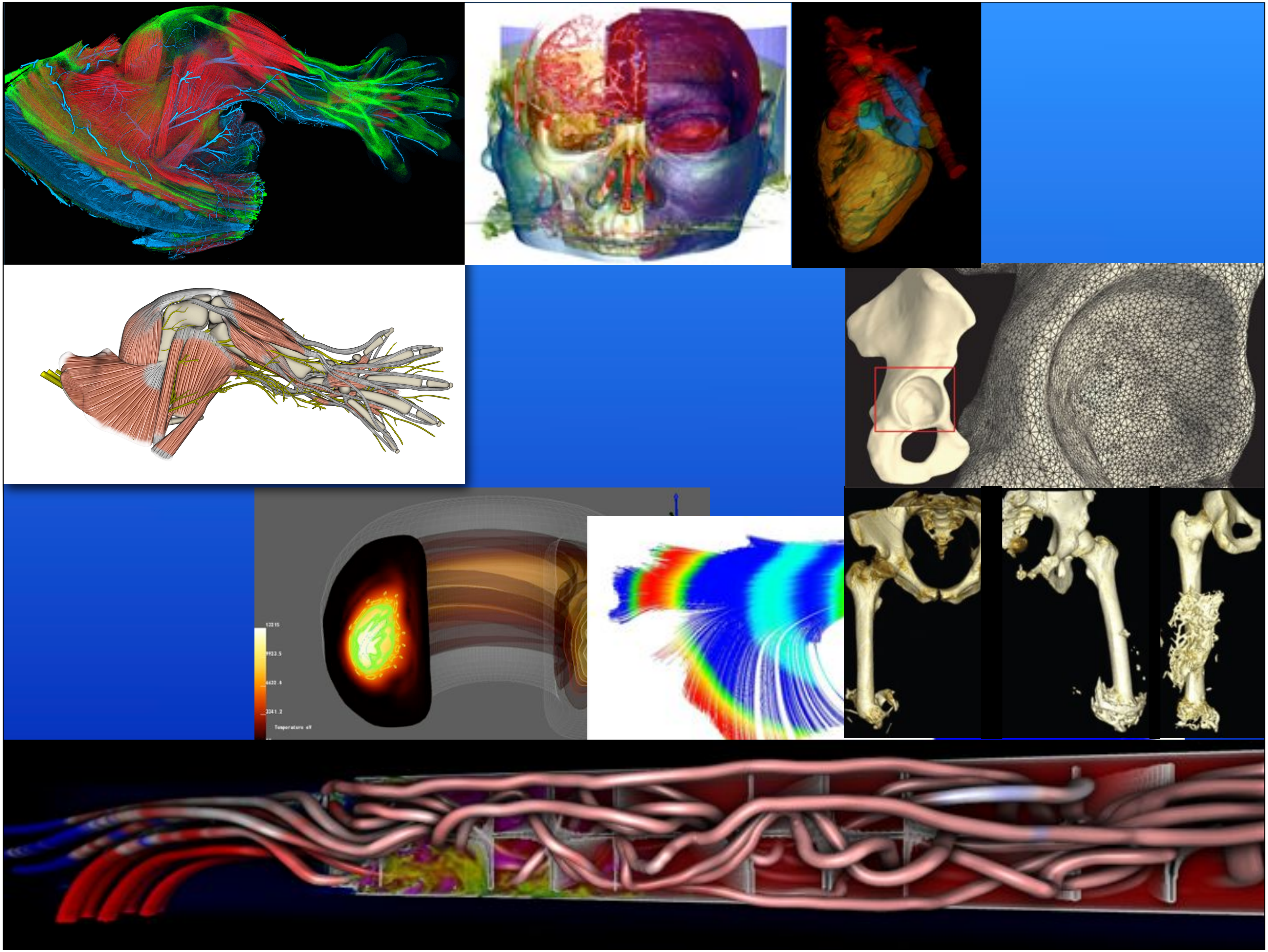
Big Data

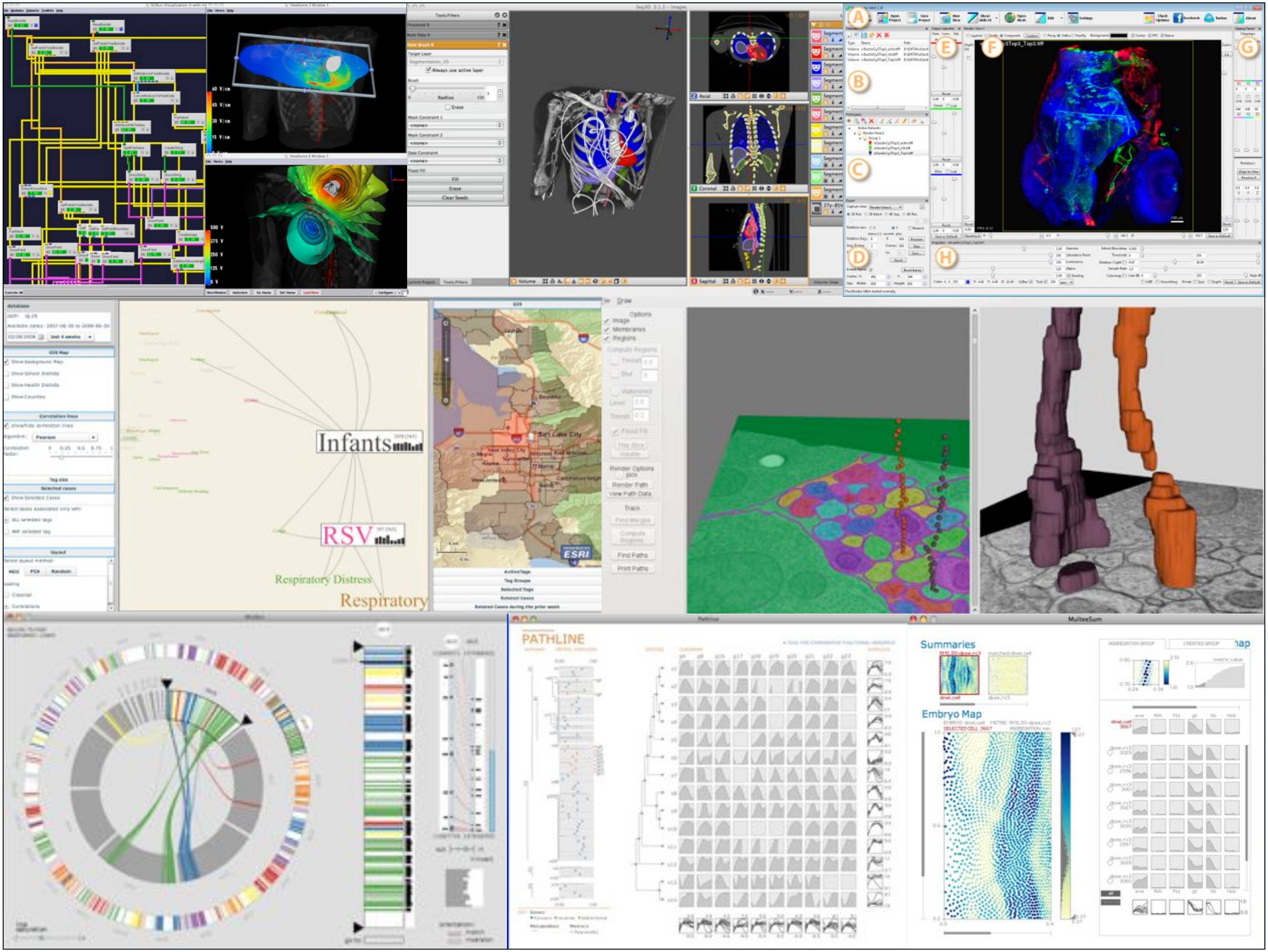
Big data is like teenage sex:
everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

Dan Ariely

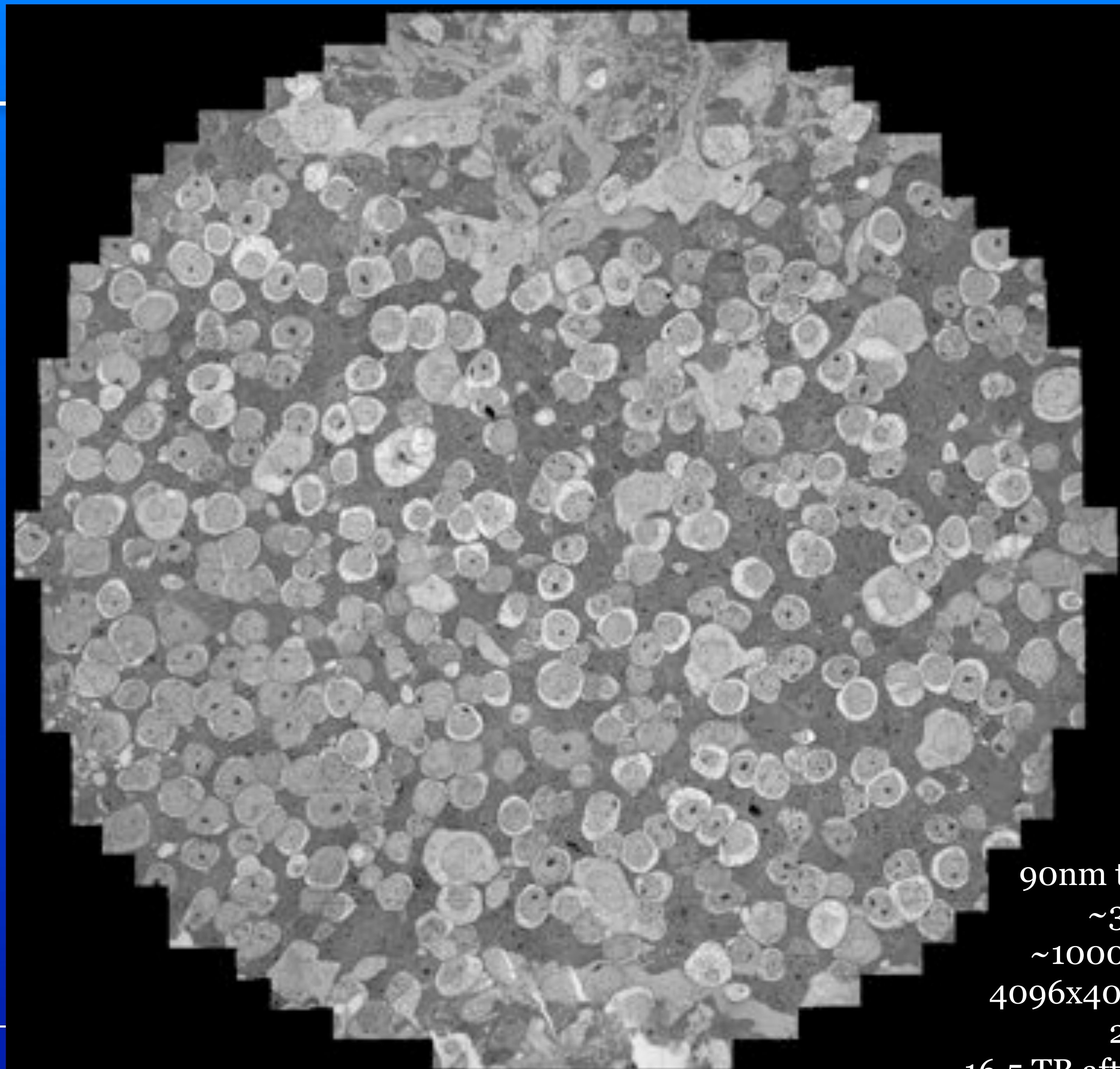
New Visual Analysis Techniques




13215
9923.5
6632.4
3341.2
Temperature eV



A
B
C
D
E
F
G
H
Axial
Coronal
Sagittal
Volume
Infants
RSV
Respiratory Distress
Respiratory
Salt Lake City
ESRI
PATHLINE
MulteeSum
Summaries
Embryo Map



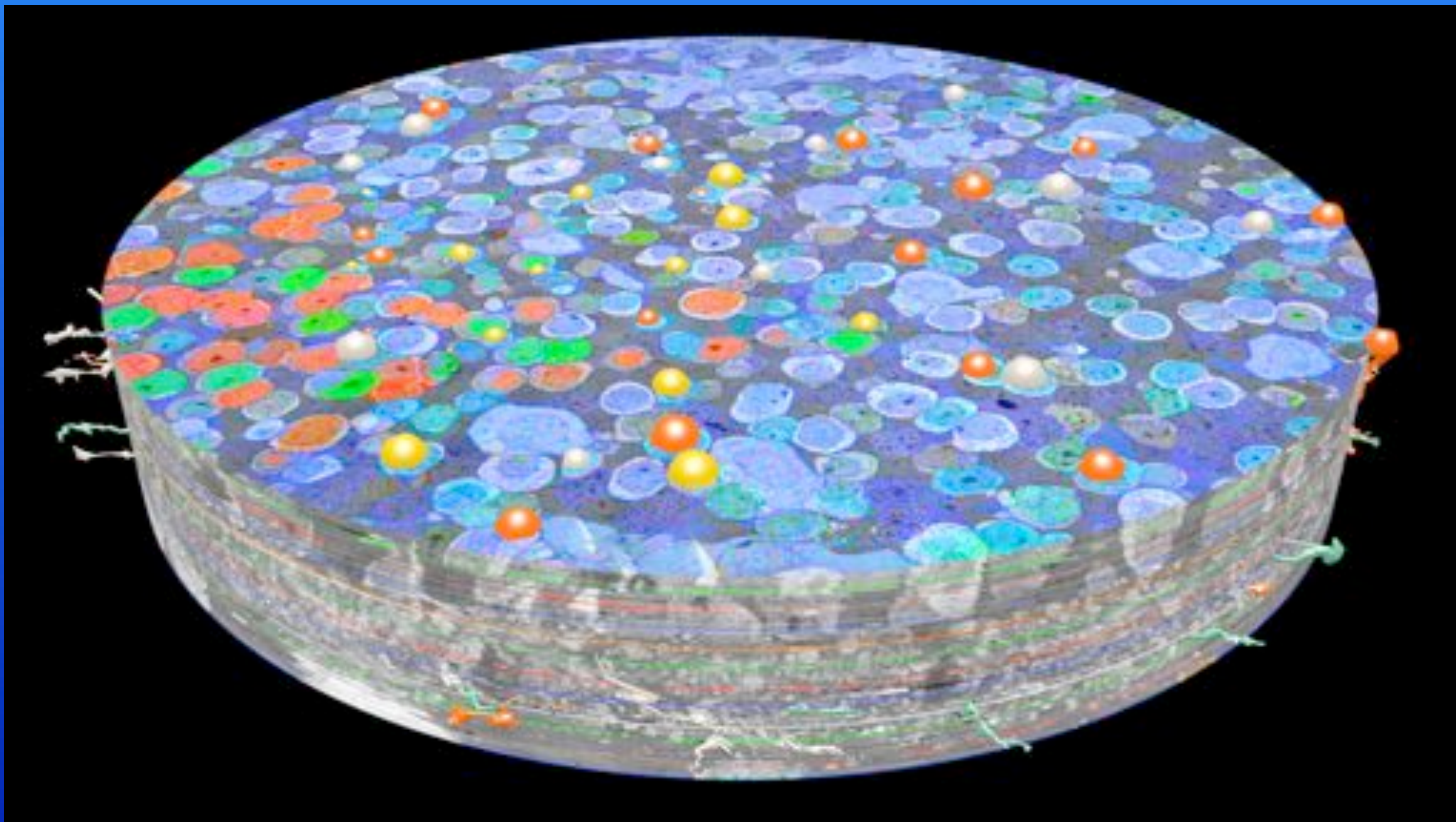
341 Sections
90nm thick sections
~32GB/Section
~1000 tiles/section
4096x4096 pixels/tile
2.18 nm/Pixel
16.5 TB after processing

Antony van Leeuwenhoek (1632-1723)

. . . my work, which I've done for a long time, was not pursued in order to gain the praise I now enjoy, but chiefly from a craving after knowledge, which I notice resides in me more than in most other men. And therewithal, whenever I found out anything remarkable, I have thought it my duty to put down my discovery on paper, so that all ingenious people might be informed thereof.

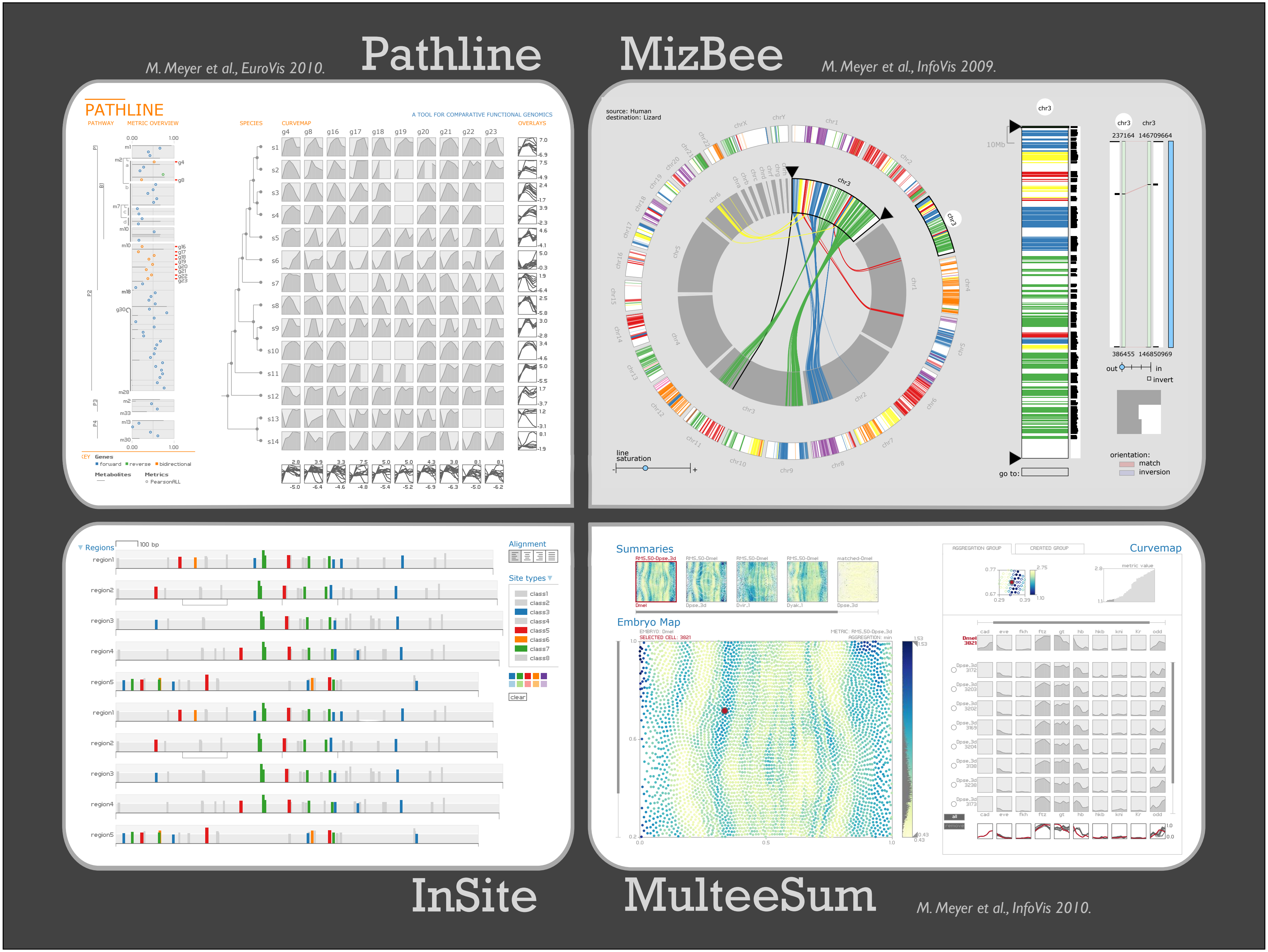
Antony van Leeuwenhoek. Letter of June 12, 1716

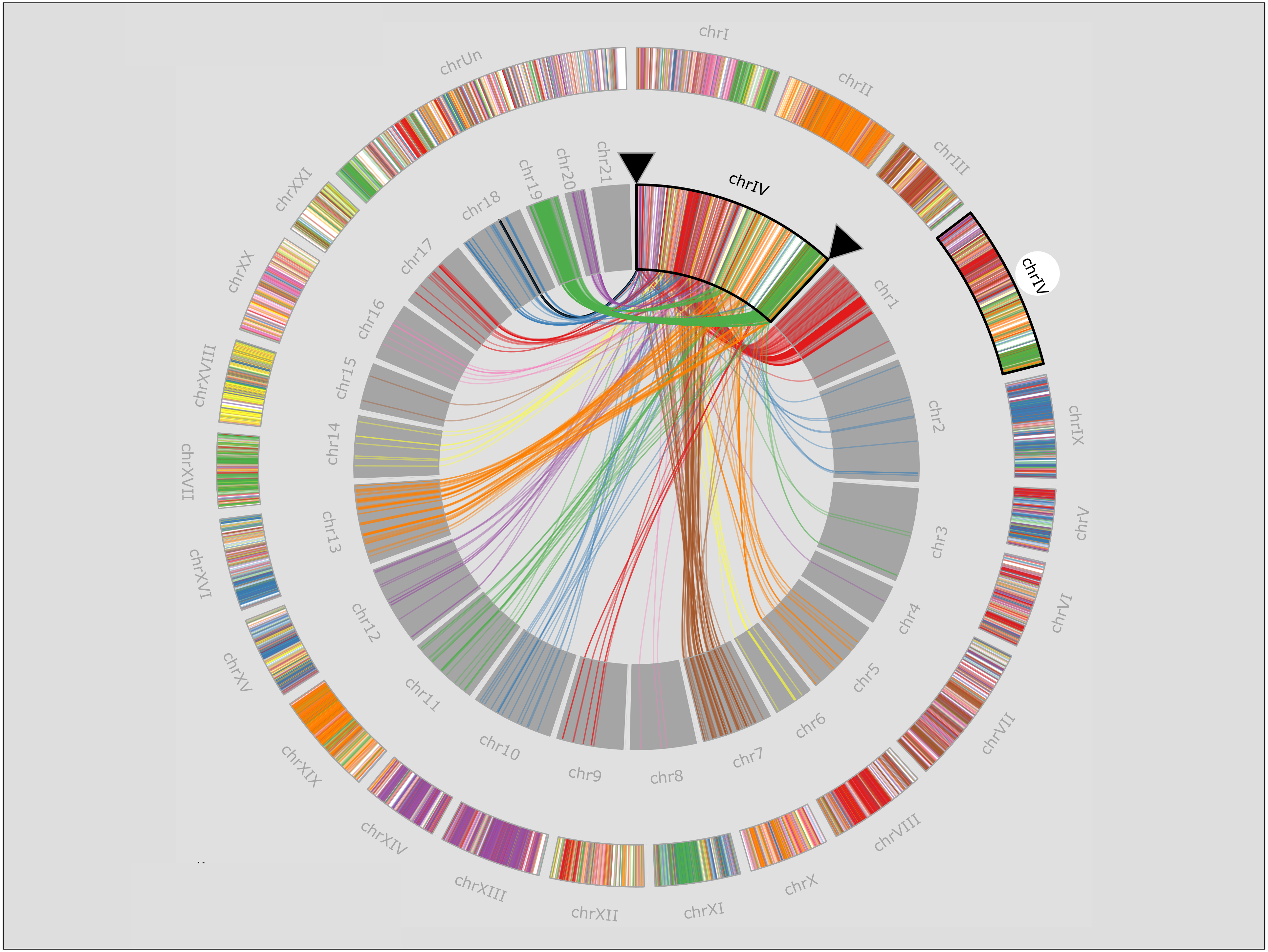
Connectome



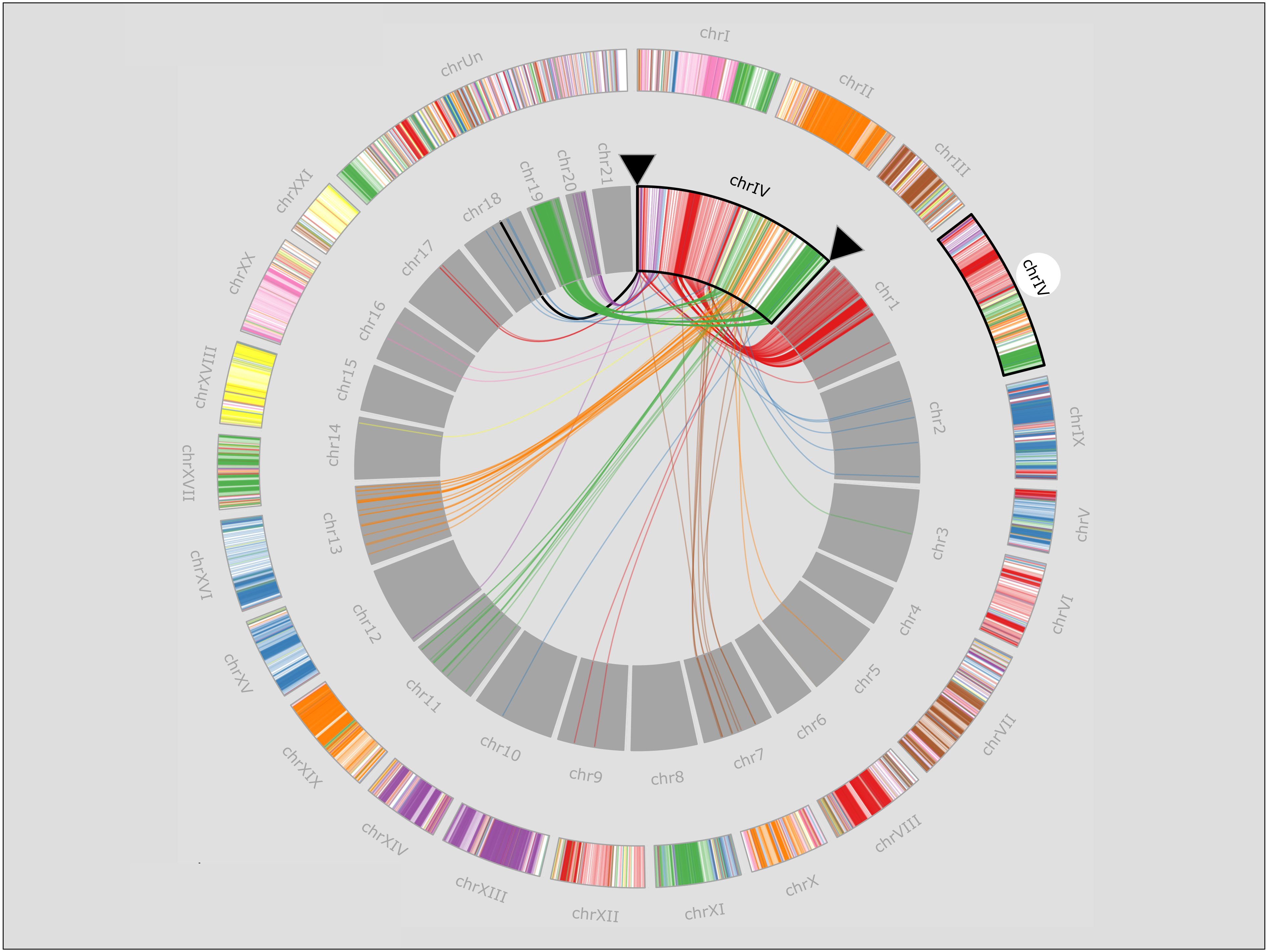
PROBLEM-DRIVEN VISUALIZATION RESEARCH *for biological data*

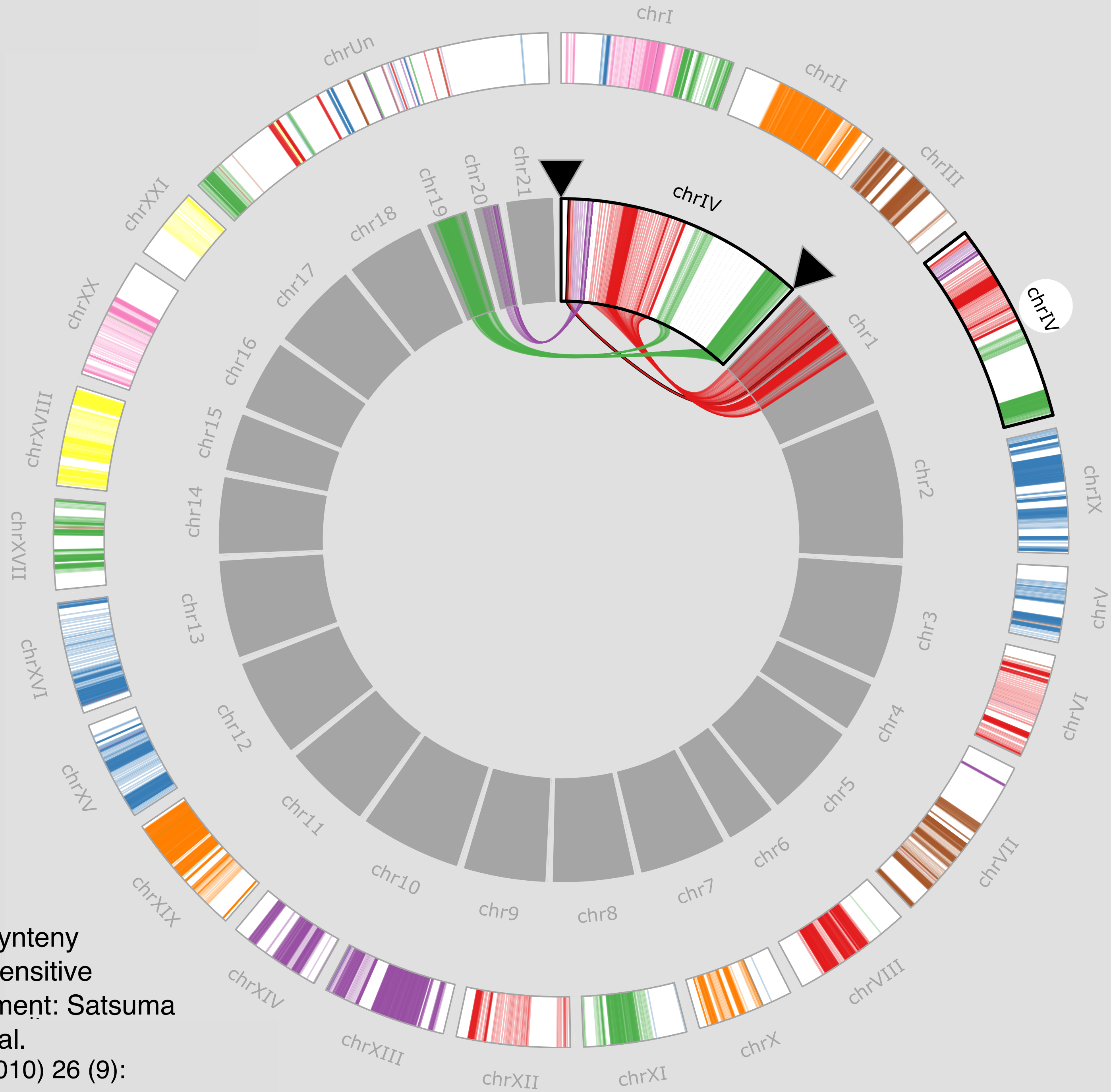
- target specific biological problems
- close collaboration with biologists
- rapid, iterative prototyping
- focus on genomic and molecular data


M. Meyer et al., EuroVis 2010.
Pathline
MizBee
M. Meyer et al., InfoVis 2009.
PATHLINE
A TOOL FOR COMPARATIVE FUNCTIONAL GENOMICS
source: Human
destination: Lizard
Regions
Alignment
Site types
Summaries
Embryo Map
Curvemap
InSite
MulteeSum
M. Meyer et al., InfoVis 2010.



chrI
chrII
chrIII
chrIV
chrIX
chrV
chrVI
chrVII
chrVIII
chrX
chrXI
chrXII
chrXIII
chrXIV
chrXIX
chrXV
chrXVI
chrXVII
chrXVIII
chrXX
chrXXI
chrUn
chrIV
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21

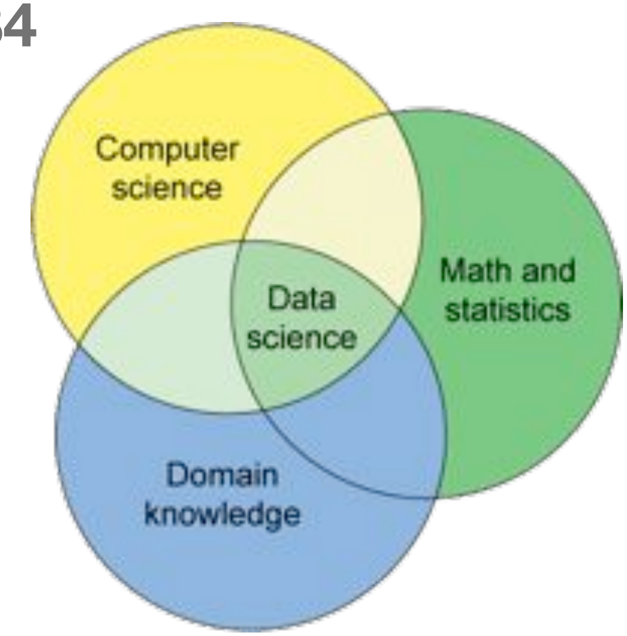

chrI
chrII
chrIII
chrIV
chrIX
chrV
chrVI
chrVII
chrVIII
chrX
chrXI
chrXII
chrXIII
chrXIV
chrXIX
chrXV
chrXVI
chrXVII
chrXVIII
chrXX
chrXXI
chrUn
chrIV
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21



Genome-wide synteny through highly sensitive sequence alignment: Satsuma
M. Grabherr, et al.
Bioinformatics (2010) 26 (9): 1145-1151.

Data Science Programs

- http://analytics.ncsu.edu/?page_id=4184
- 19 MS programs in Data Analytics
- 8 MS programs in Data Science
- 28 MS programs in Business Analytics
- Several additional “tracks” or “concentration” programs



Big Data Curriculum

Analytics Electives:

- **Data Mining** (required)
- **Machine Learning** (required)
- **Visualization** (required)
- **Artificial Intelligence**.
  Decision making under uncertainty.
- **Natural Language Processing**.
  Understanding textual data and language.
- **Probabilistic Modeling**.
  Advanced statistical techniques and tools (using R).
- **Image Processing**.
  Analysis and learning on image data.

Big Data Curriculum

Algorithmics Electives:

- **Advanced Algorithms** (required)
- **Models of Computation for Big Data**.
  How algorithmic bottlenecks change as data becomes very large; Relation to modern big data systems (e.g. MapReduce).
- **Computational Geometry**.
  Geometric interpretation of big data analysis and computation.
- **Computational Topology**.
  Topological data analysis and algorithms.

Big Data Curriculum

Management Electives:

- **Database Systems** (required)
- **Parallel Programming for Many-Core Architectures**.
  Parallel Computing and High Performance Computing.
  Scalable programming on GPUs, many-cores, and HPC clusters.
- **Advanced Computer Networks**.
  Large-scale network protocols, architectures, and applications.
- **Network Security**.
  Message integrity, access control, authentication, confidentiality.

Piloting in Adobe (Lehi)

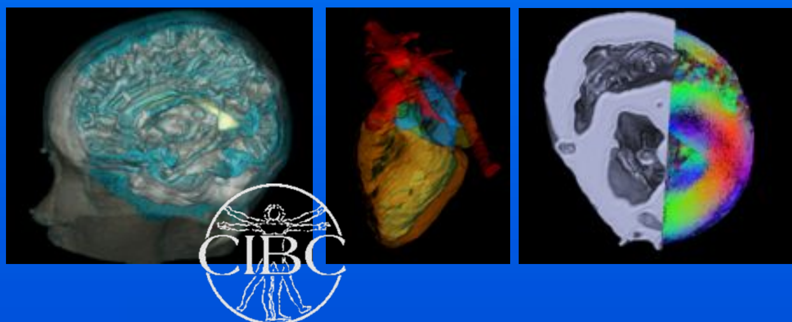
- Starting Fall 2014.
- Live 2-way streaming. Interaction across video.
  Fall 2014: Visualization: T-Th 9:10 - 10:30am
  Fall 2014: Adv. Algorithms: T-Th 10:45 - 12:05am
  (plan for early evening, e.g. Data Mining M-W 5:15-6:35pm)
- Potential for Instructor on site in future.
  Adobe lecture room open to others.

The SCI Institute

Productivity Machines

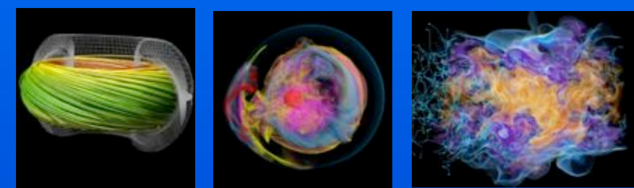
Acknowledgments

NIH/NIGMS Center for Integrative Biomedical Computing



SDAV

Scalable Data Management, Analysis and Visualization



UTAH Center for Computational Earth Sciences

Utah Center for Neuroimage Analysis

NIH NAMIC

Center for Extreme Data Management, Analysis, and Visualization

CEDMAV

KAUST

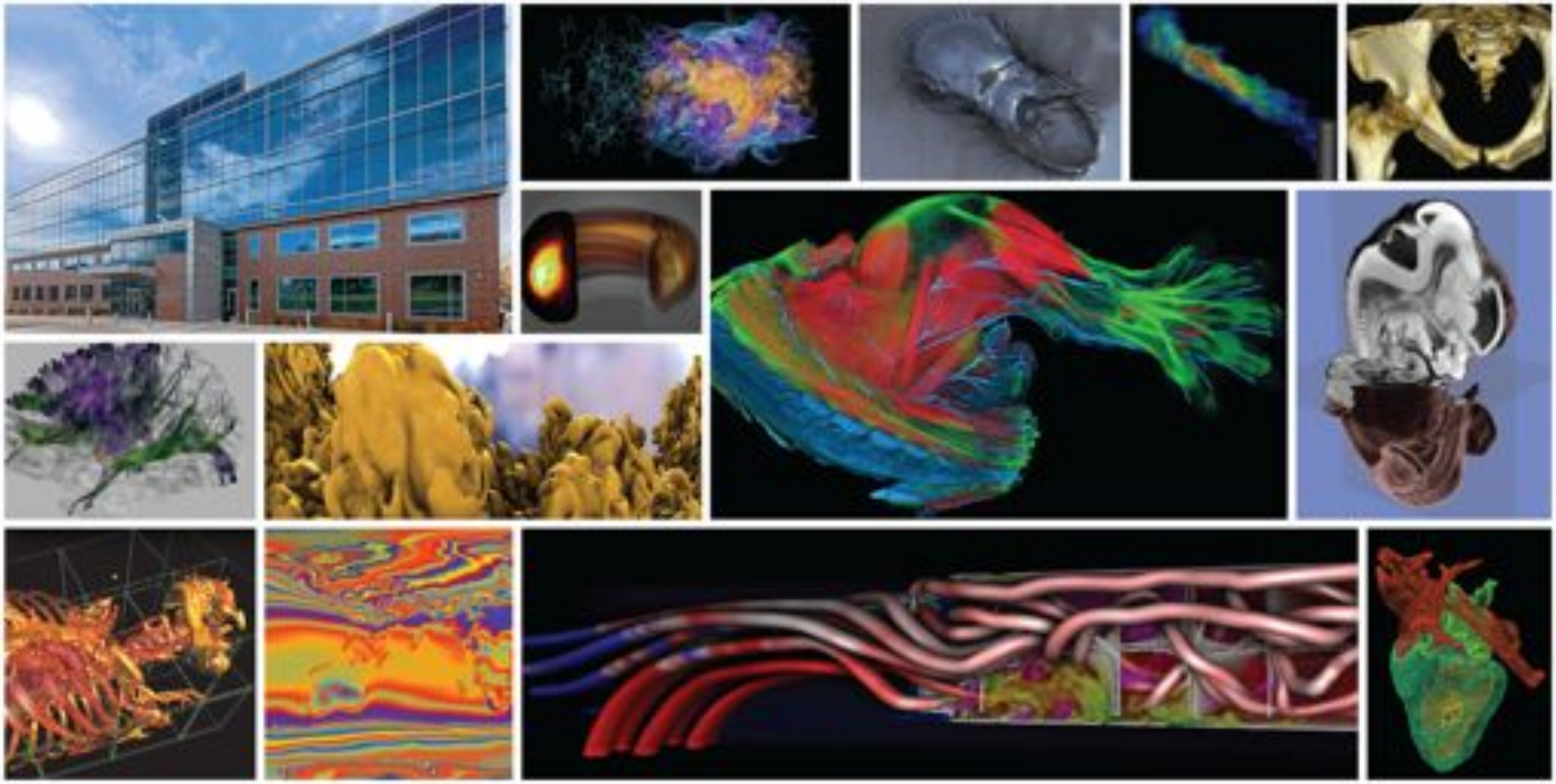
IAMCS
Institute for Applied Mathematics and Computational Science

More Information

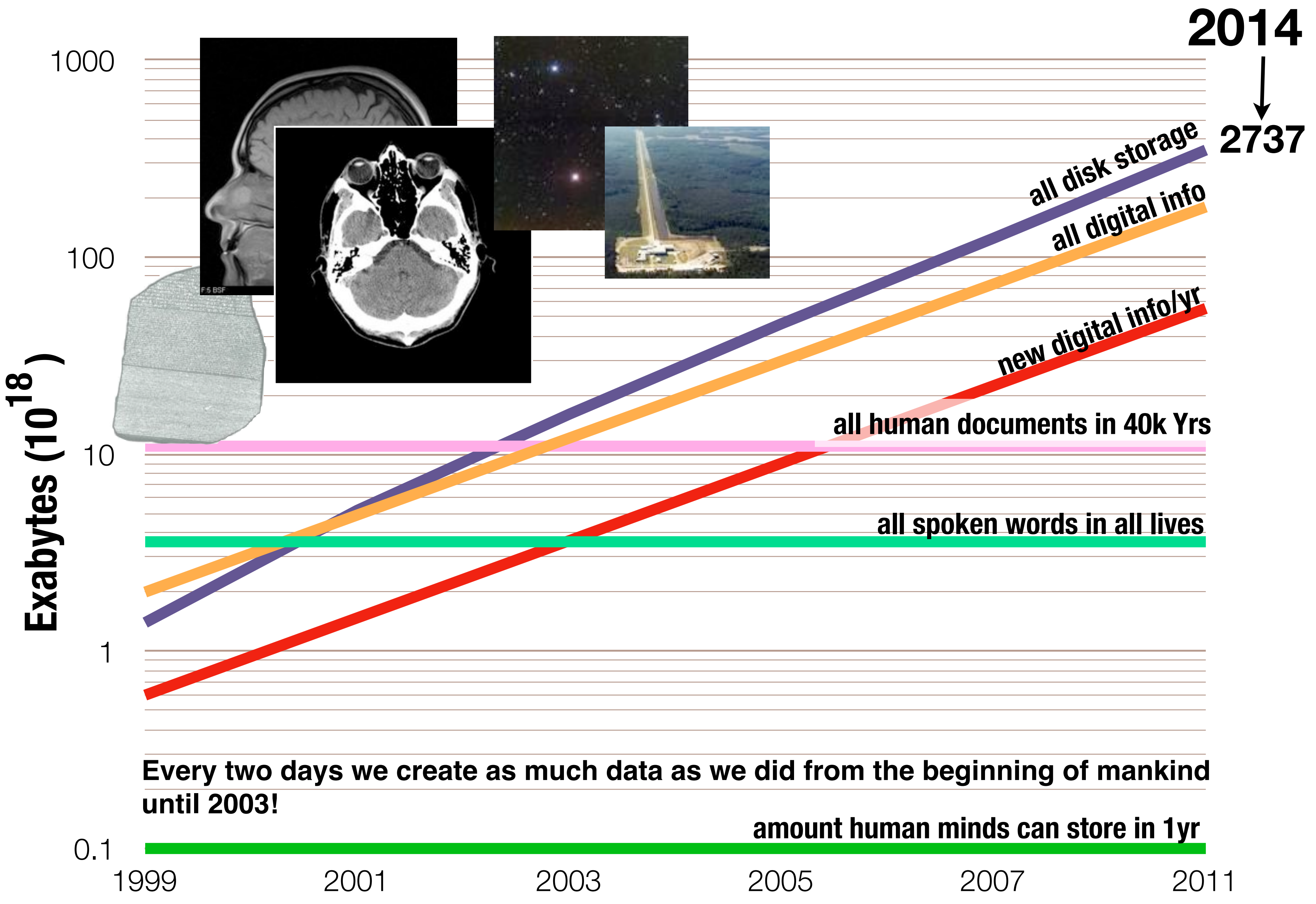
www.sci.utah.edu

crj@sci.utah.edu

Ecosystem Challenges Around Data Use



Chris Johnson
Scientific Computing and Imaging Institute
University of Utah



Sources: Lesk, Berkeley SIMS, Landauer, EMC, TechCrunch, Smart Planet

Big Data

Big data is like teenage sex:
everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

Dan Ariely

Panelists

Leonid Zhukov - Director of Data Science, Ancestry.com

Vance Checketts - VP and GM of EMC

Edison Ting, Solutions Architect, Pivotal

GS Big Data Platform

Data Philosophy

1. Instrument everything
2. Put all data in one place
3. Data first, questions later
4. Store first, structure later
5. Let everyone party on the data (with controls)
6. Keep raw data forever
7. Produce tools to support the whole research cycle
8. Modular and composable infrastructure

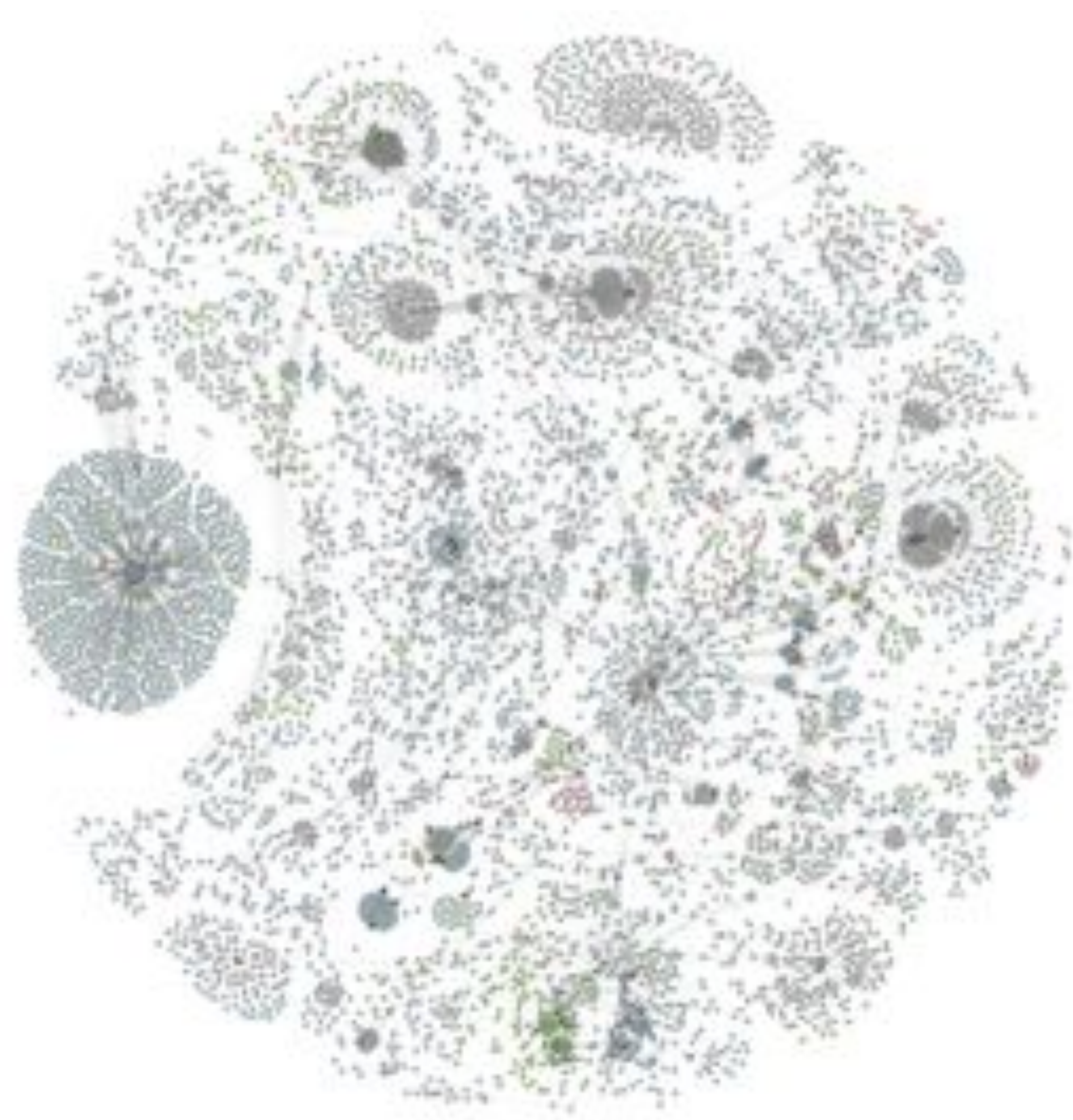
Distributed Systems; Distributed Data

Our 'Big' small data problem

- Highly functionally aligned systems
  - Excellent Data Segregation
  - Local Data Autonomy
  - Local Governance & retention
  - Locally negotiated data evolution

- Extensive ~~ab~~use of data movement technologies
  - 'Shared Data' (reference data) is broadly disseminated, but mostly from central locations
  - 'Event' data (Transactions) flow across systems and persisted at each stage
  - We rarely used centralized shared services like the reference data farm

- Our Data is an '*Asset'* and should be treated as such



Big Data Platform Goals

Create a 'GS Data Lake' to allow for many datasets to coexist and be available which is external to any specific GS application.

Creating a data registry to store the dataset metadata and allow for datasets to be discovered and used.

Create a facility to properly entitle access to the datasets (that code is typically custom logic embedded directly within each application)

Create the facilities to ingest the data and provide resiliency.

Build an integrated software stack using multiple data management software products to provide the full suite of function required for the Data Lake.

The platform will be composed of multiple products integrated and made consistent by GS developed infrastructure.

HDFS/Hive for deep petabyte scale online archive

MPP Column-store ParAccel for high performance aggregation/pivoting

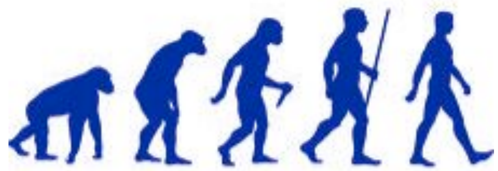
Graph database for non-relational queries and semantic search

Text search function for unstructured or semi-structured datasets

Metadata registry and entitlement model

The evolution of scale-out data management platforms

Perhaps the single biggest factor in enabling Big Data is the rapid innovation occurring for the "scale-out" of data management platforms.



 DBMS runs on single host only (traditional RDBMS)

 DBMS runs on multiple hosts, single copy of data, statically partitioned (DB2 DPF)

 DBMS runs on multiple hosts, two copies of data, statically partitioned (ParAccel)

 DBMS runs on multiple hosts, many copies of data, statically partitioned (MongoDB)

 DBMS runs on multiple hosts, many copies of data, dynamic partioning (H-Base)

Big Data Platform Desired Properties

Scalable

- No fixed upper bound to ultimate dataset size.
- Storage and CPU capacity must be able to be increased in an incremental and linear fashion.
- Platform technology stack should already be in use at larger scale than GS use-case.

Affordable

- Technology hardware stack should be based on a scale-out of commodity components.
- Technology software stack should be based on open source projects.
- Platform should be designed to run on GS Dynamic Compute nodes
- Vendor lock-in for any unique portion of the platform should be avoided when possible.
- Operating cost of platform should be kept to a minimum via low touch infra-structure that self manages.

Trusted

- Entire be resilient to individual component failure not requiring any manual intervention
- Must be easy to both self heal from failure and to scale-out additional capacity
- Must have facilities to allow for authentication, security and access entitlement

Big Data Platform Ongoing Research (1)

Entitlement for Big Data platform datasets

Two different entitlement problems to be solved for:

- How to model the entitlement rules on who should be able to see what data.
- How to implement those rules within the platform.

Products such as sqrrl and Accumulo are being looked at to provide the fine grained access control. Alternatively the rules could be implemented in a GS access layer software

Big Graph

- Graph databases can be powerful, allowing for queries that are difficult to express in SQL.
- Graph databases do not easily lend themselves to data shard'ing and scale-out.
- YarcData Urika product is being looked at for high performance Big Graph solution.
- Aurelius Titan graph database also being tested.

Text Search Data Store / Semantic Search Data Store

- Entitling data stored in an unstructured or semi-structured manner poses new challenges
- Elasticsearch product is used in several different applications within GS
- Attivio product is also in use at GS

Big Data Platform Ongoing Research (2)

Big Data platform - data movement

- Information loaded to the Big Data platform should be considered immutable.
- Data fed into the Big Data platform will need to be stored identically on multiple clusters.
- Gigabus will be instrumental in creating serialized streams of data across the Big Data platform
- Any new data created on the Big Data platform will need to be streamed back into the platform

Big Data platform – data retention

- Traditional concepts such as a ‘database backup’ or ‘transaction log’ need to be completely rethought for the Big Data platform.
- Forcing all data through a product with data retention such as Gigabus should be enforced.
- All products that feed data unto the Big Data platform should have a method of replaying datasets unto the platform on request.

Big Data platform – workload management

- All things being equal a fewer number of clusters is preferable to a greater number of clusters.
- YARN and other technologies are being tested to understand workload management functions
- Hadoop data federation technologies are being tested to bridge multiple products.

Big Data Product Status Q2 2014

GS Big Data Catalog

- Runs on standard Dynamic Compute nodes.
- Utilizes CKAN open source metadata repository application and UI.
- Metadata is externalized in Google DSPL format.
- Entire registry stack can be extended for GS specific requirements.

Hortonworks Hadoop 2.06

- Runs on Dynamic Compute large storage nodes.
- Major engineering effort underway to have full Kerberos integration.
- Standard monitoring to Fabric with 24x7 support team.
- HDFS file based technologies such as M-R, PIG and Hive currently used in production.
- H-Base key value database currently used in production.
- Site resiliency and data retention will not be provided via the Hadoop stack

ParAccel Relational DBMS

- Runs on Dynamic Compute large storage nodes.
- Mature high performance MPP columnar RDBMS.
- Cluster is inelastic and does not keep multiple copies of data.

Appendix

Using R: Statistical Data Analysis

You may have heard of….

Predictive Analytics, Data Mining, Data Analysis, Statistical Analysis, Data Visualization, Business Intelligence, Big Data

Who uses it? (who doesn't?)

Google, Facebook, Double-Click, LinkedIn

Credit Card Companies, Insurance Companies, Finance (of course)

Anywhere you want to extract value from your data

Development Patterns

Data Visualization

Graphing statistical summaries of data to gain insights

Modeling and Prediction

Model the system using statistical models, then use those models to check new data

Data visualization used to understand how the model performs

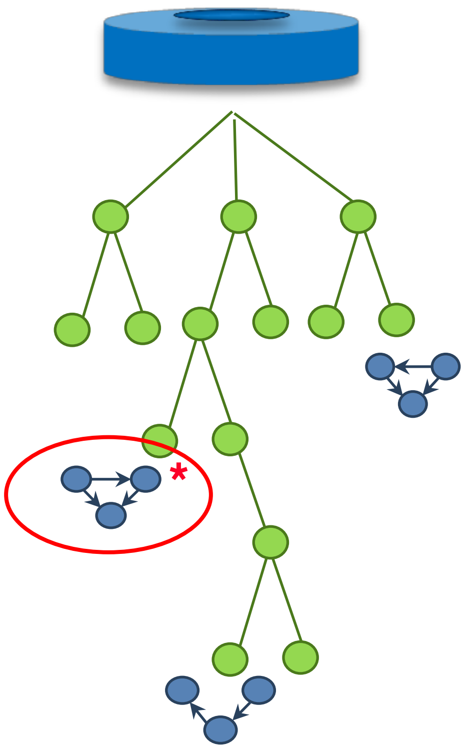
R is a great way for programmers to do statistics

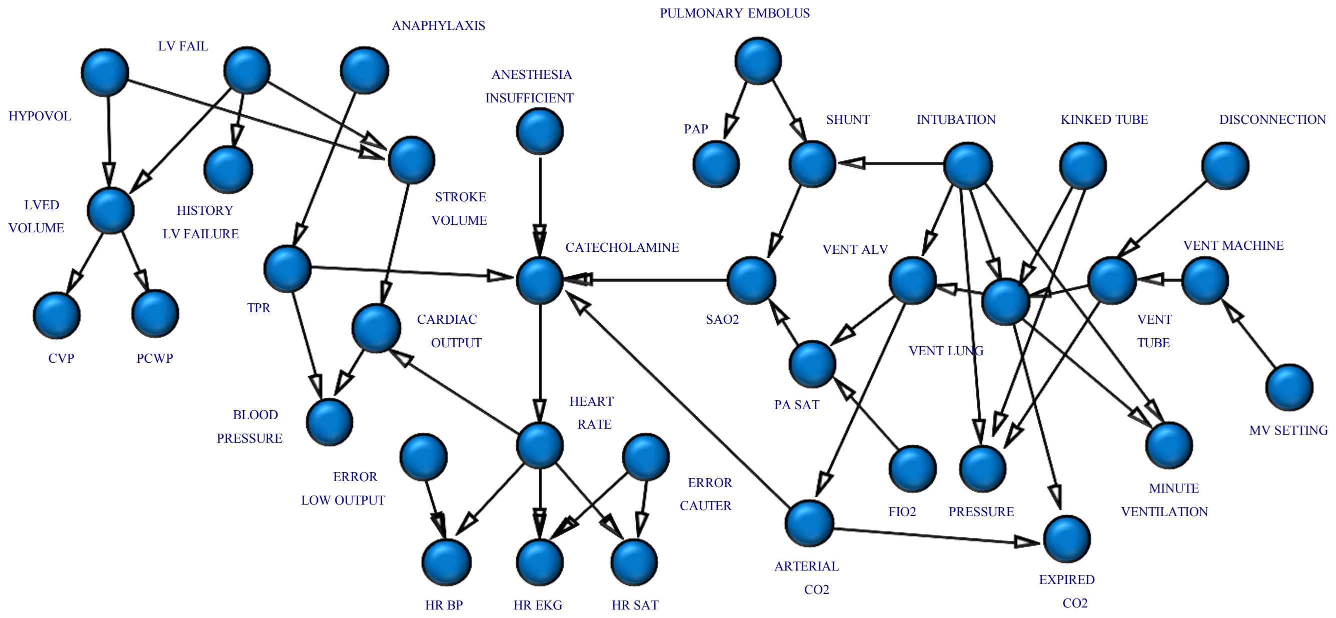
From Data to Insight to Change: Technologies and Opportunities

Eric Horvitz
Microsoft Research

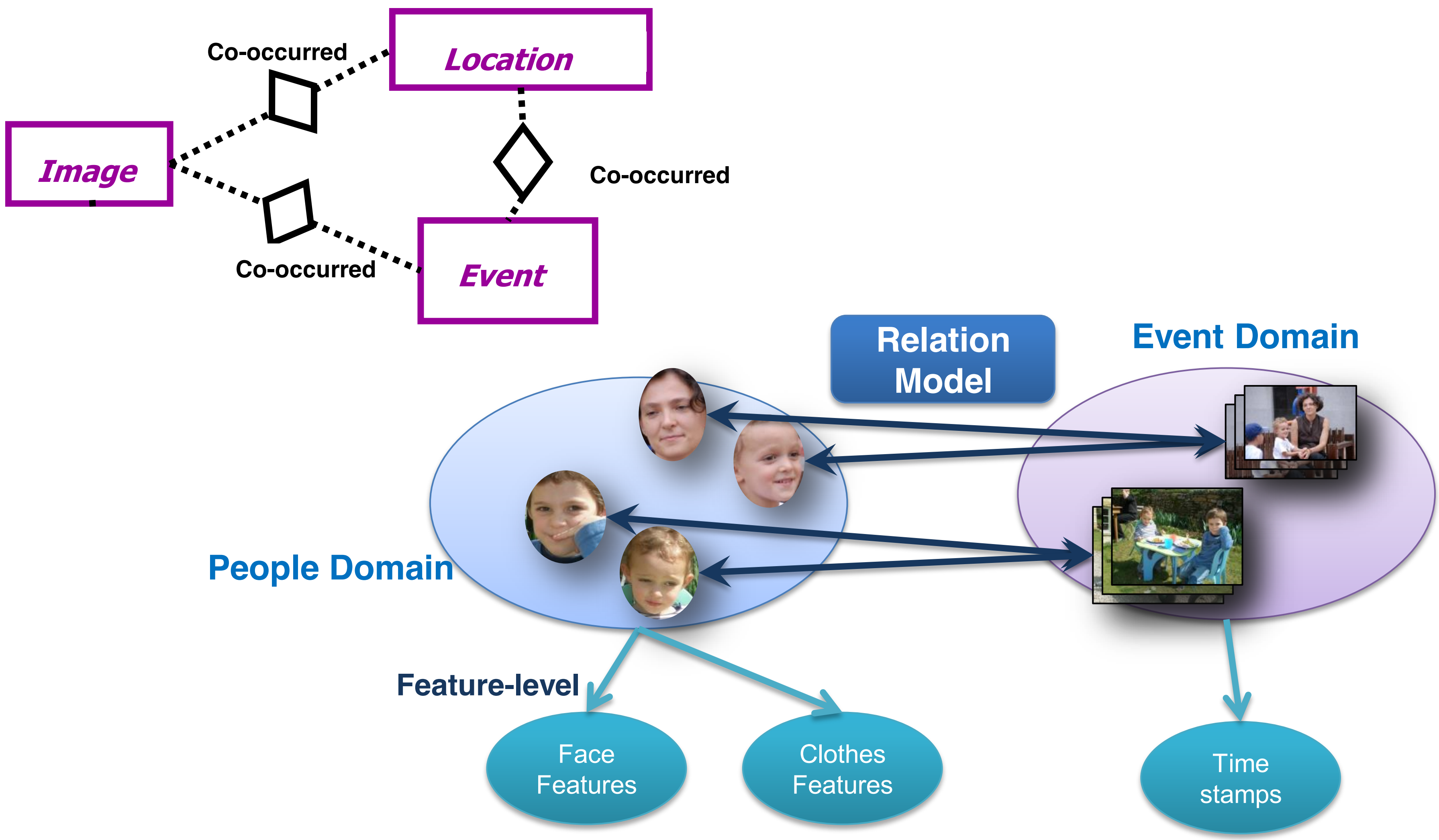
http://research.microsoft.com/~horvitz

Advances in learning and inference from data

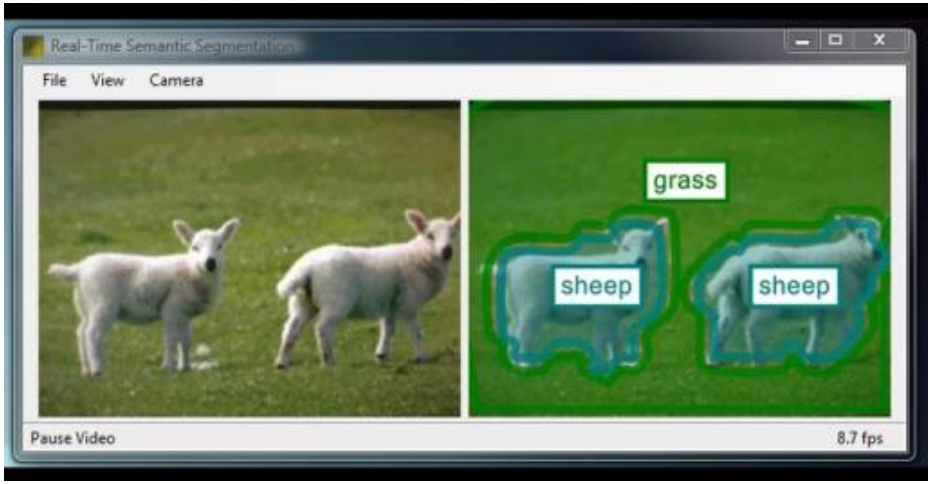




Rise of Rich Representations

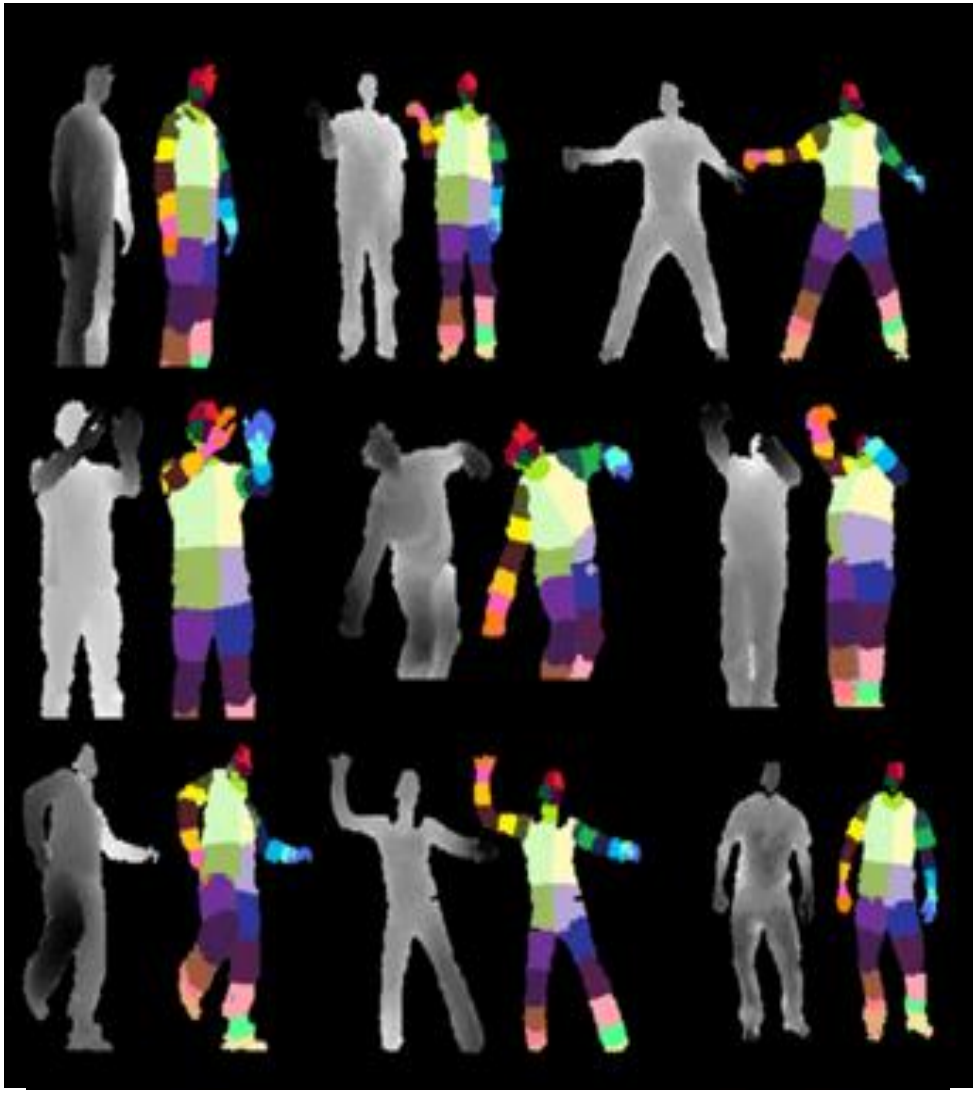


Rise of Rich Representations



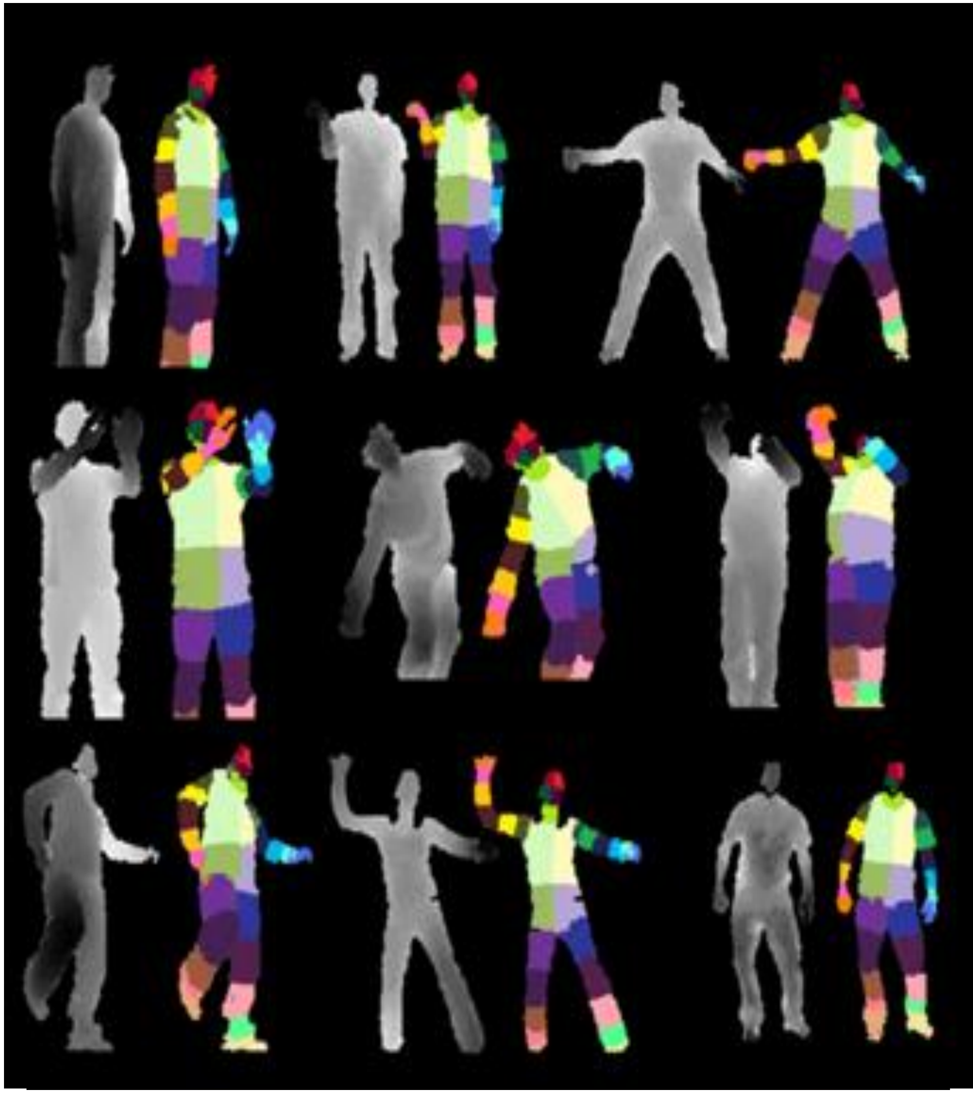
J. Shotton, J. Winn, C. Rother, A. Criminisi

Rise of Rich Representations



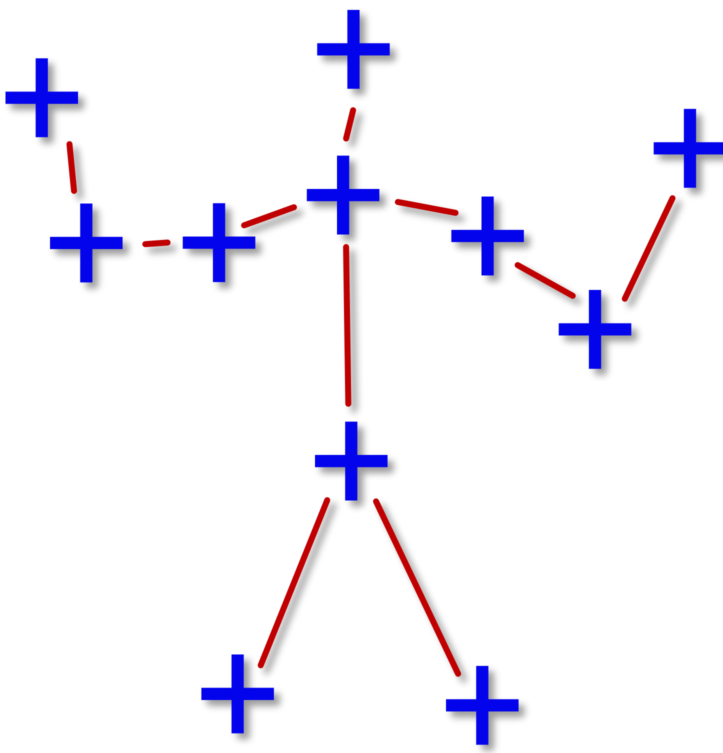
J. Shotton, J. Winn, C. Rother, A. Criminisi

Rise of Rich Representations



J. Shotton, J. Winn, C. Rother, A. Criminisi

Rise of Rich Representations



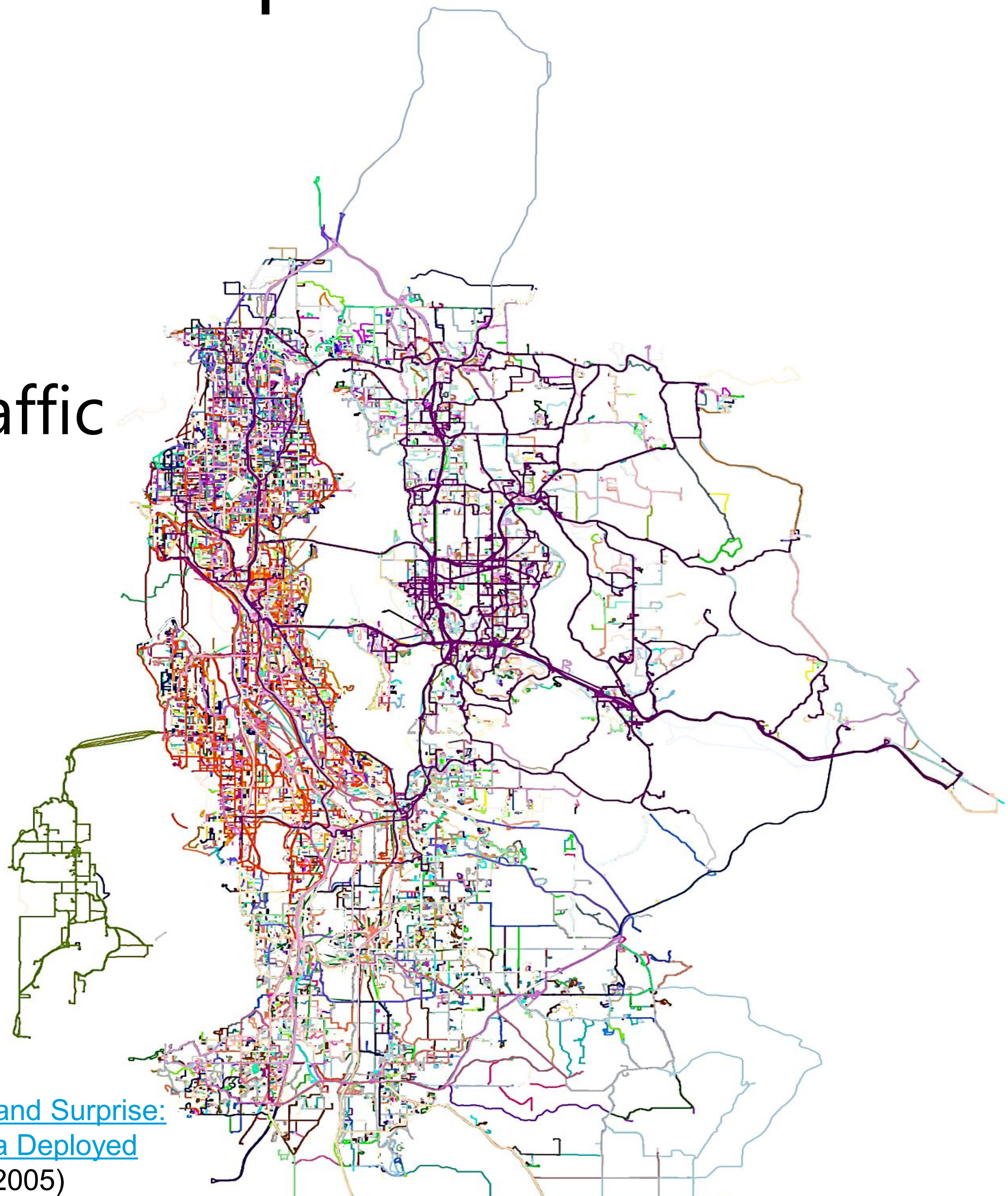
J. Shotton, J. Winn, C. Rother, A. Criminisi

Rise of Rich Representations

KINECT
for XBOX 360.

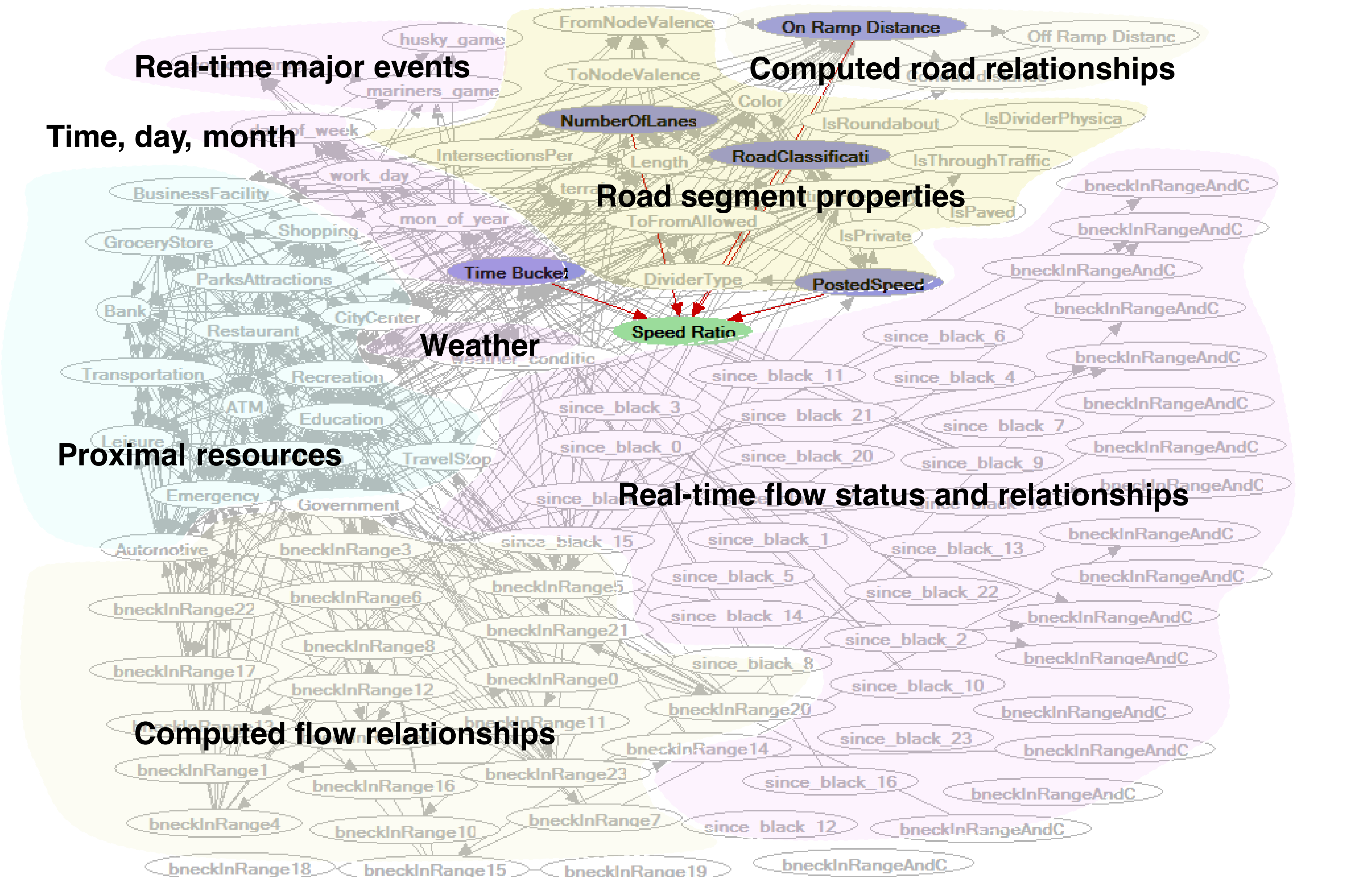
Rise of Rich Representations

Seattle traffic

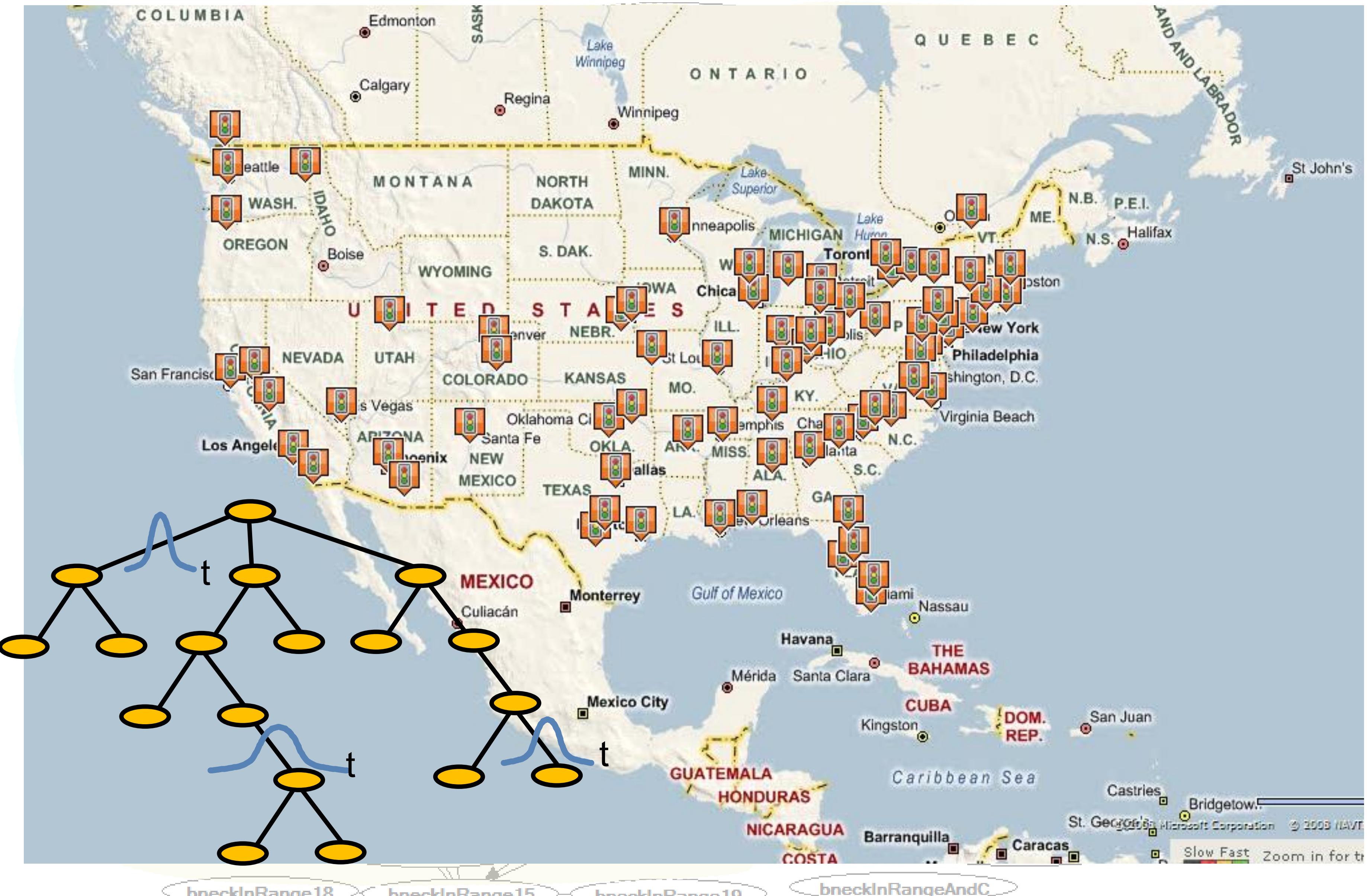


H. et al. [Prediction, Expectation, and Surprise: Methods, Designs, and Study of a Deployed Traffic Forecasting Service](#) (UAI 2005)

Rise of Rich Representations



Rise of Rich Representations



Rise of Rich Representations

The New York Times

Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

Search Tech News & 8,000+ Products [Go]

Browse Products -- Select a Product Category -- [Go]

Microsoft Introduces Tool for Avoiding Traffic Jams

By JOHN MARKOFF
Published: April 10, 2008

SAN FRANCISCO — Microsoft on Thursday plans to introduce a Web-based service for driving directions that incorporates complex software models to help users avoid traffic jams.

Related

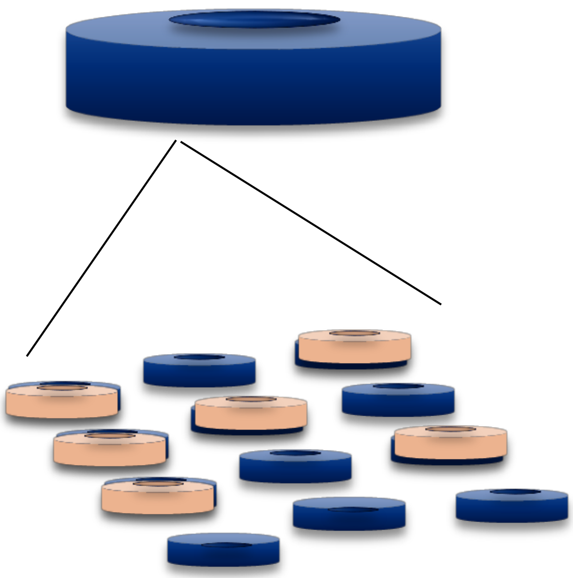
Times Topics: Microsoft Corporation

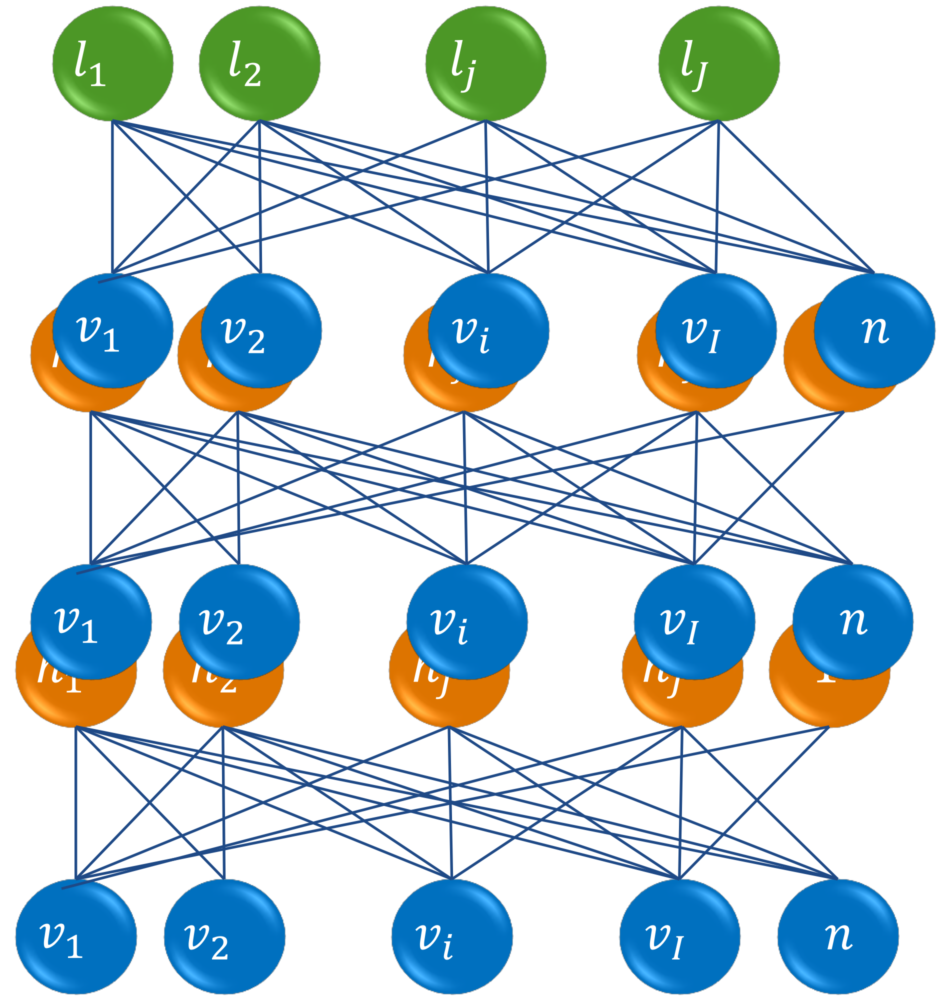
The new service's software technology, called Clearflow, was developed over the last five years by a group of artificial-intelligence researchers at the company's Microsoft Research laboratories. It is an ambitious attempt to apply machine-learning techniques to the problem of traffic congestion. The system is intended to reflect the complex traffic interactions that occur as

Microsoft now considers surface street traffic as well as freeway speeds in its routing.

Data and Power of Familiar Methods

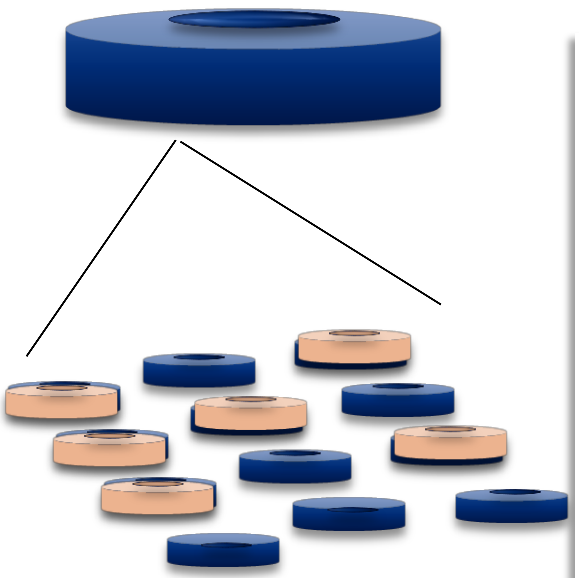
Pursuit of speech, vision with stacked representations

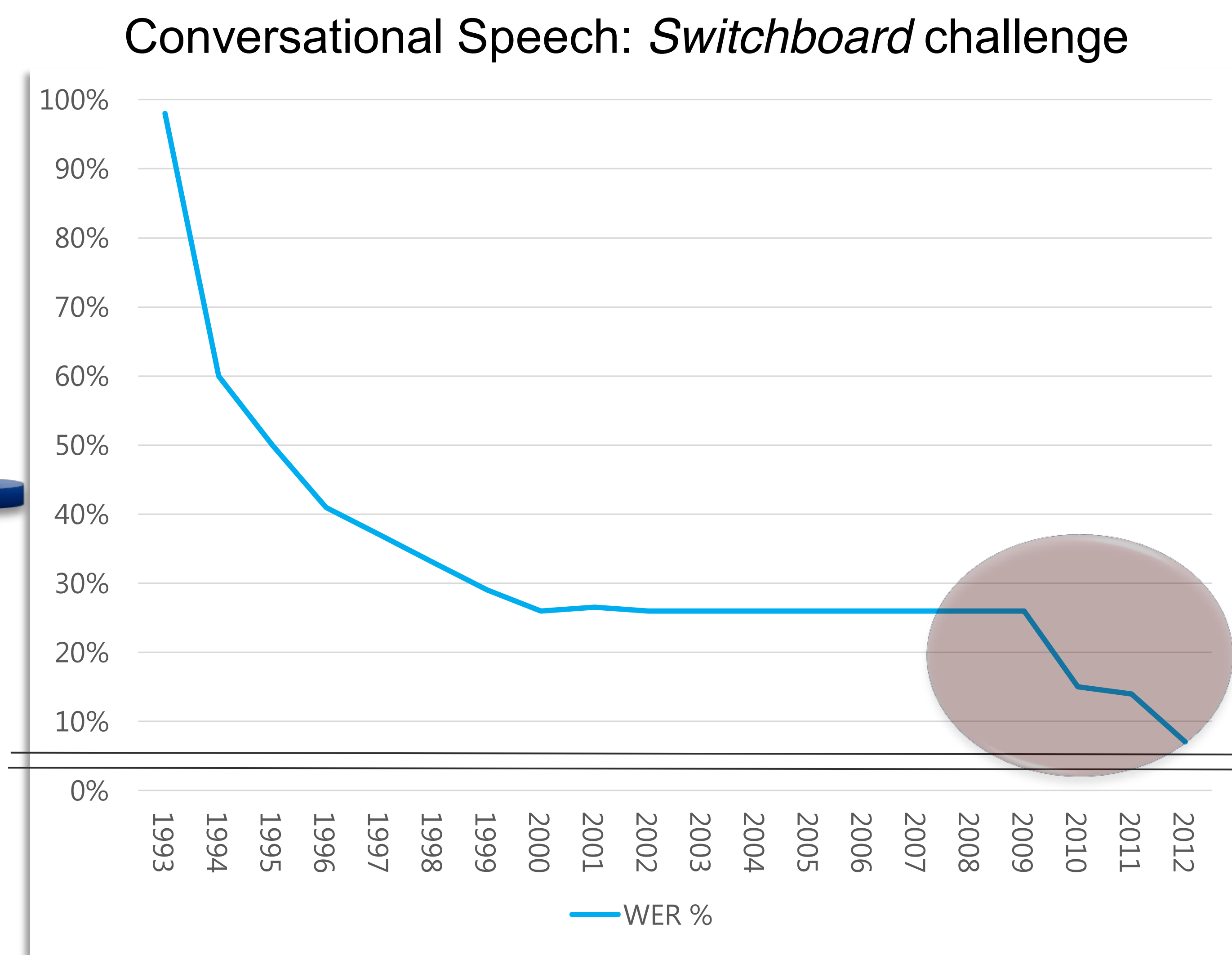




Data and Power of Familiar Methods

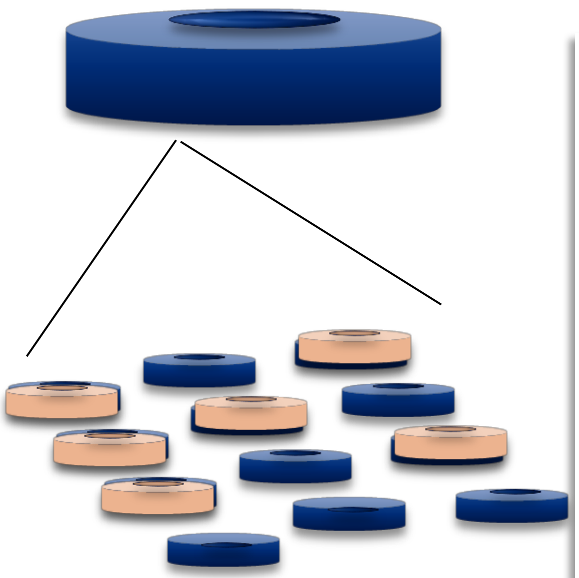
Pursuit of speech, vision with stacked representations



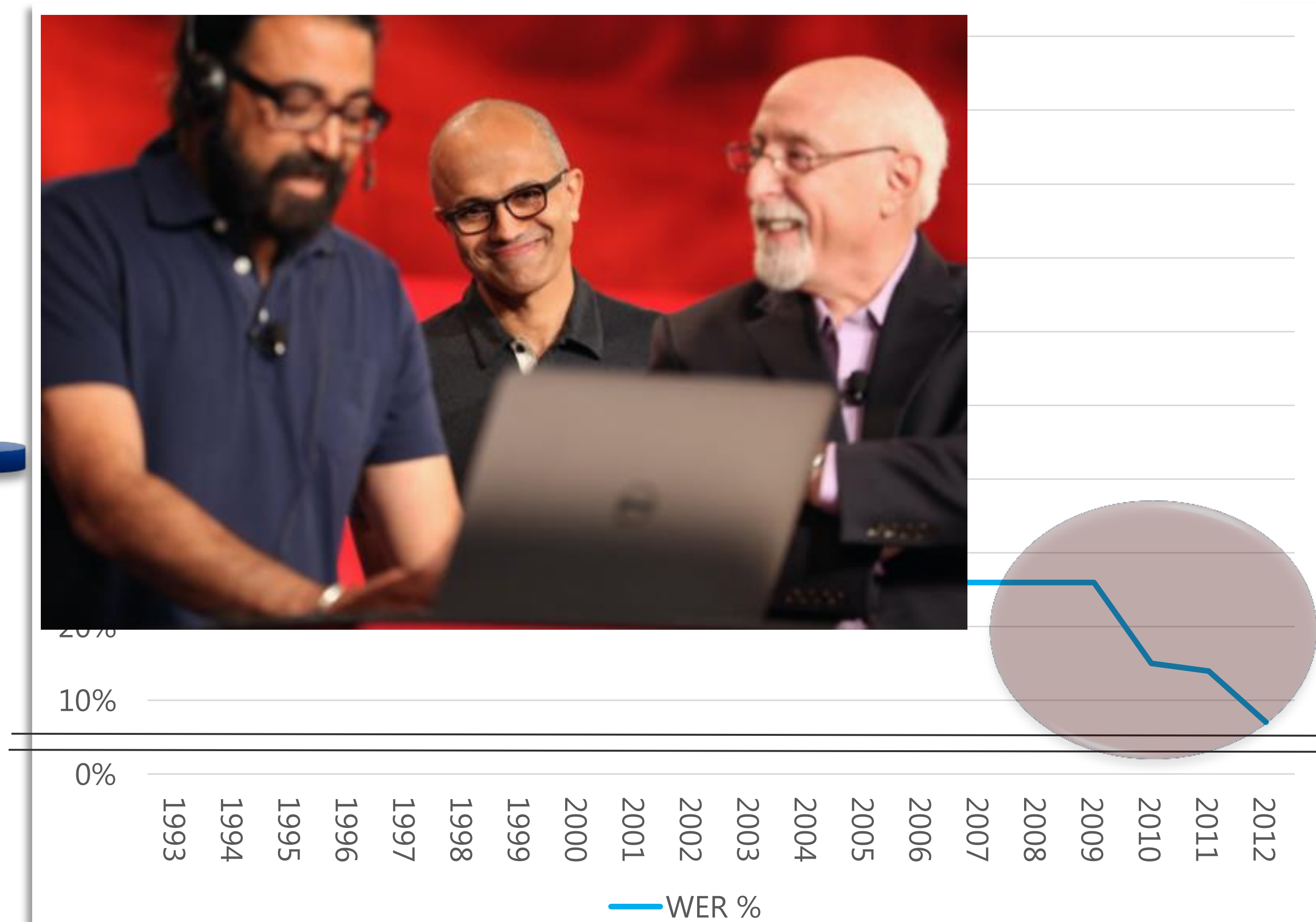


Data and Power of Familiar Methods

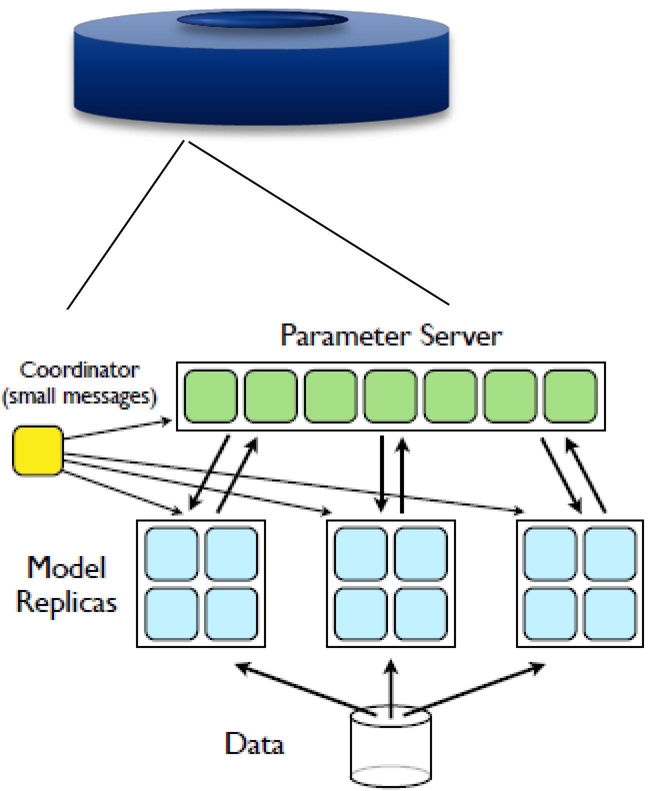
Pursuit of speech, vision with stacked representations

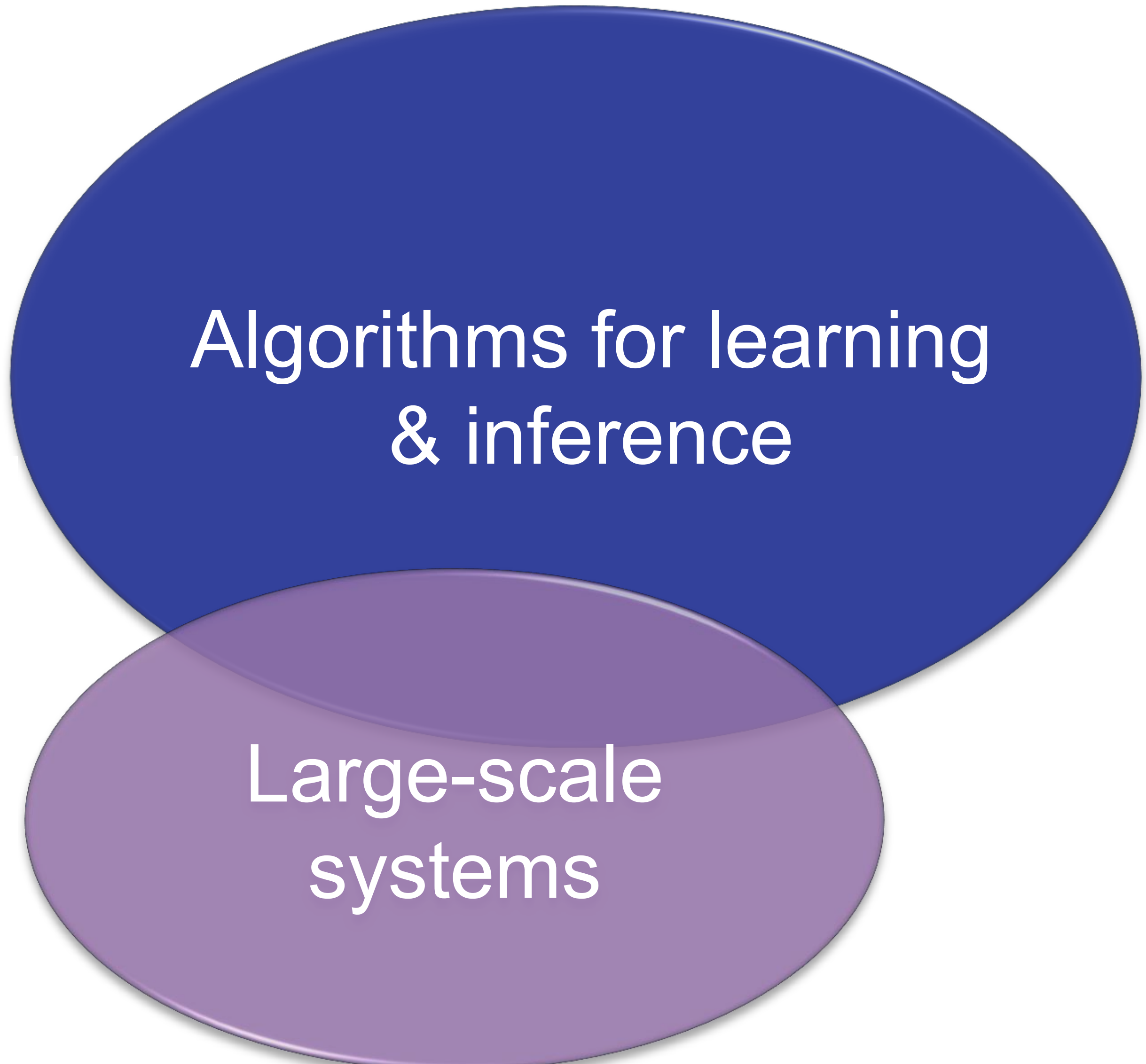


Conversational Speech: *Switchboard* challenge



Direction: Data, Learning, and Systems





Beauty and the Bottleneck

Hekaton: Database service

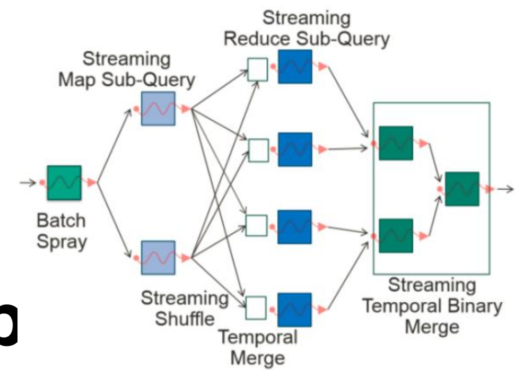
In-memory, manycore, latch-free:

30x speed-up

Trill: Streaming analytics

Column-oriented batches, P3 sort:

2-4 orders of magnitude speed-up



Catapult: Data center search perf.

Speed-ups via FPGA

40x speed-up

Machine Learning PREVIEW

Powerful cloud-based predictive analytics

- Designed for new and experienced users
- Proven algorithms from MS Research, Xbox and Bing
- First class support for the open source language R
- Seamless connection to HDInsight for big data solutions
- Deploy models to production in minutes
- Pay only for what you use. No hardware or software to buy.

Get started now

Machine Learning pricing details ▸ Machine Learning tutorials ▸

What Our Early Adopters Are Saying

SALES 1-800-867-1389 | MY ACCOUNT | PORTAL | Search

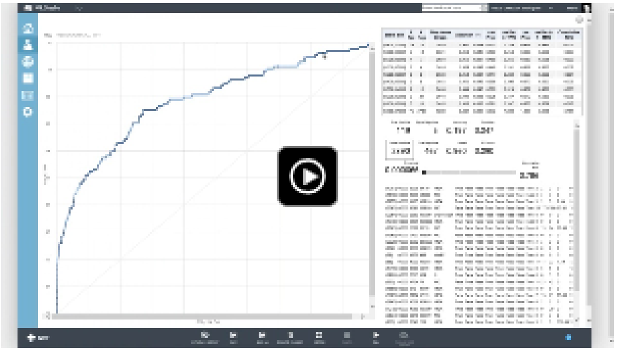
Microsoft Azure

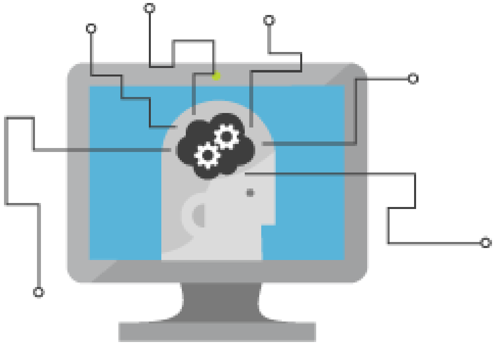
Features Pricing Documentation Downloads Gallery Community Support

FREE TRIAL

The power of machine learning

Machine learning–mining historical data with computer systems to predict future trends or behavior–touches more and more lives every day. Search engines, online recommendations, ad targeting, virtual assistants, demand forecasting, fraud detection, spam filters–machine learning powers all these modern services. But these uses barely scratch the surface of what's possible.





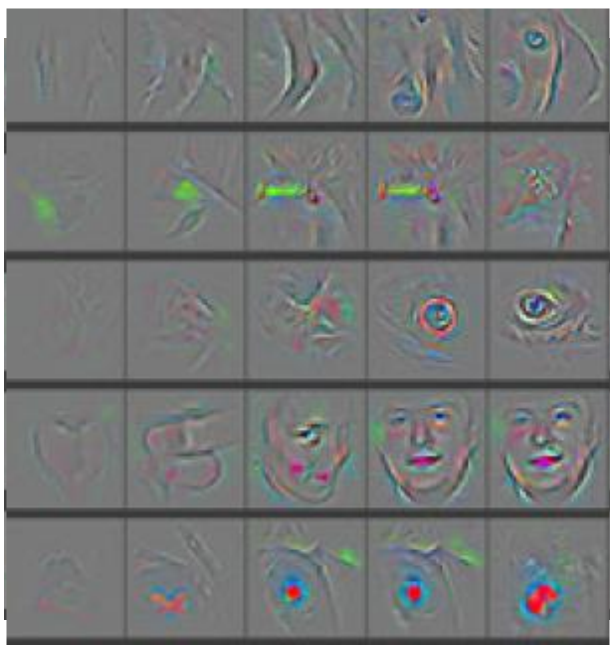
Meet Azure Machine Learning

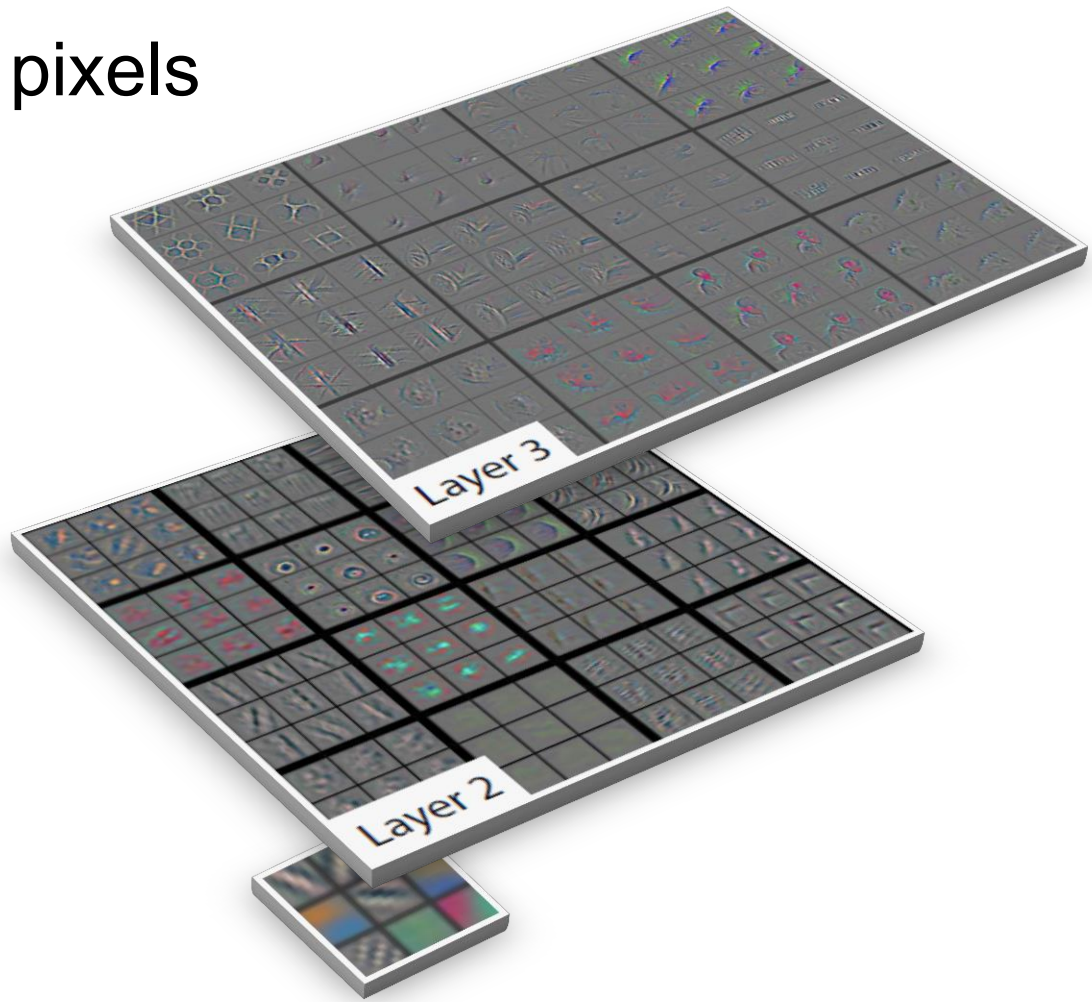
The problem? Machine learning traditionally requires complex software, high-end computers, and seasoned data scientists who understand it all. For many startups and even large enterprises, it's simply too hard and expensive. Enter Azure Machine Learning, a fully-managed cloud service for predictive analytics. By leveraging the cloud, Azure Machine Learning makes machine learning more accessible to a much broader audience. Predicting future outcomes is now attainable.

Direction: Insights via Visualization

Power of building visualization pipeline (Zeiler et al., 2011)

DNNs: Map features to pixels

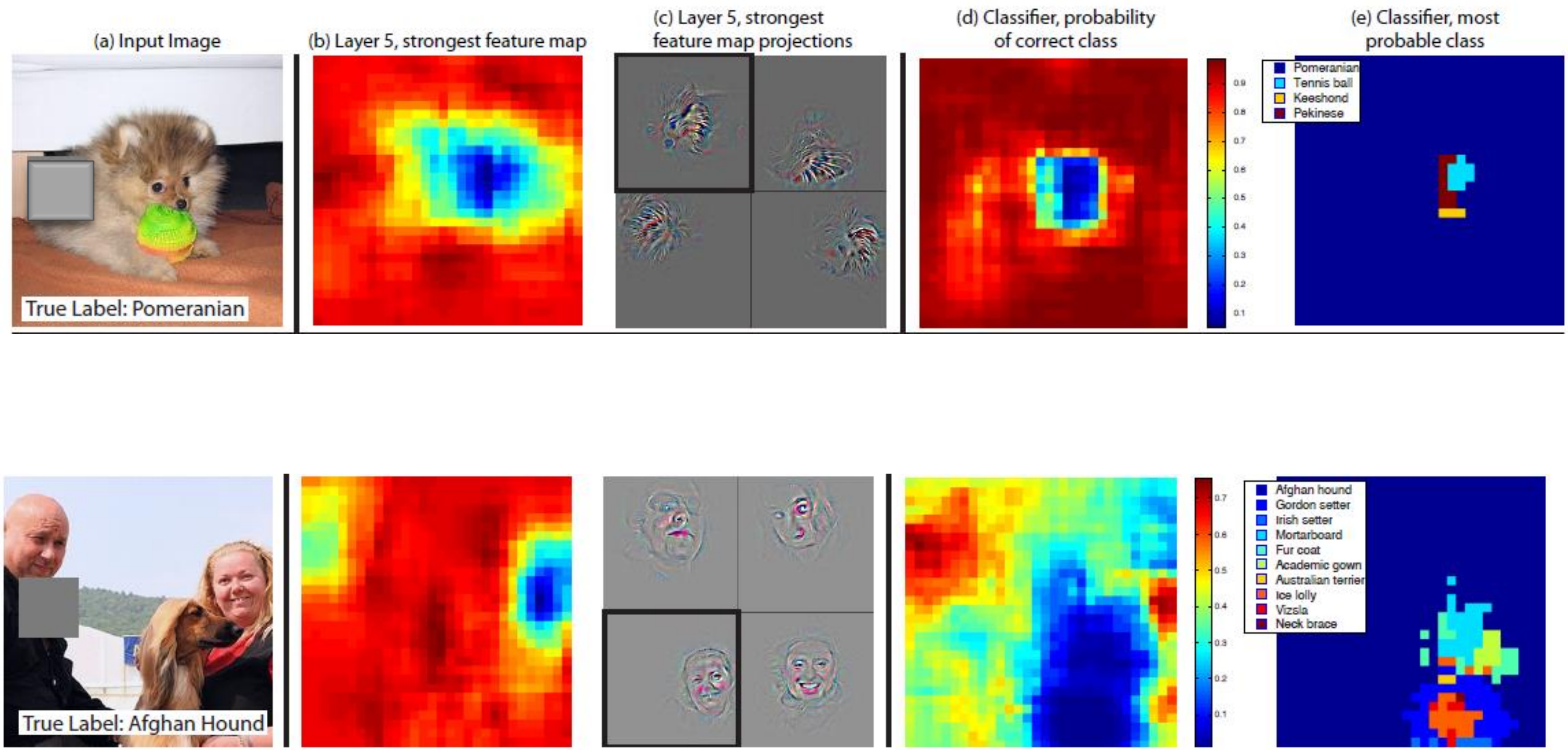




M. Zeiler, R. Fergus Visualizing and Understanding Convolutional Networks, Arxiv (2013)

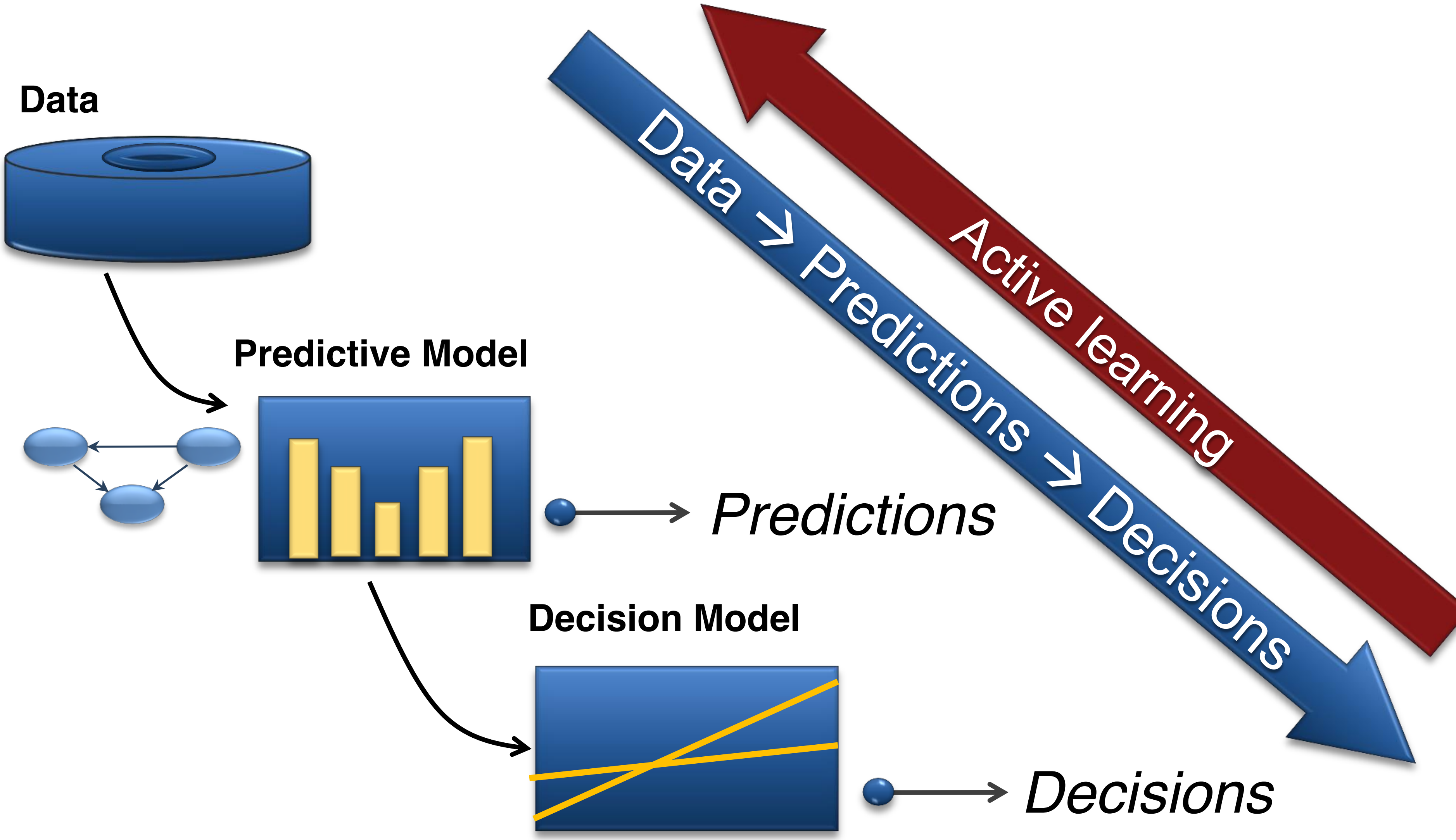
Direction: Insights via Visualization

Invariances and Sensitivities



M. Zeiler, R. Fergus Visualizing and Understanding Convolutional Networks, Arxiv (2013)

Direction: Predictions to Decisions



Direction: Predictions to Decisions

Readmissions Manager for Microsoft Amalga

Reducing Hospital Readmissions is an Impending Priority

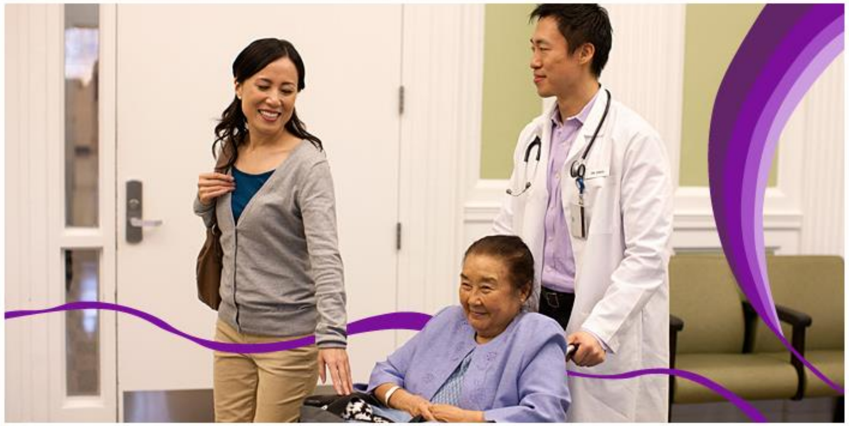
Overview

One in five Medicare inpatients is readmitted within 30 days. The Centers for Medicare and Medicaid Services (CMS) considers 40%-75% of these readmissions to be preventable.

In October 2012, CMS will begin to track readmission and impose financial penalties on hospitals with higher–than–expected readmission rates for certain conditions. Other payers will certainly follow.

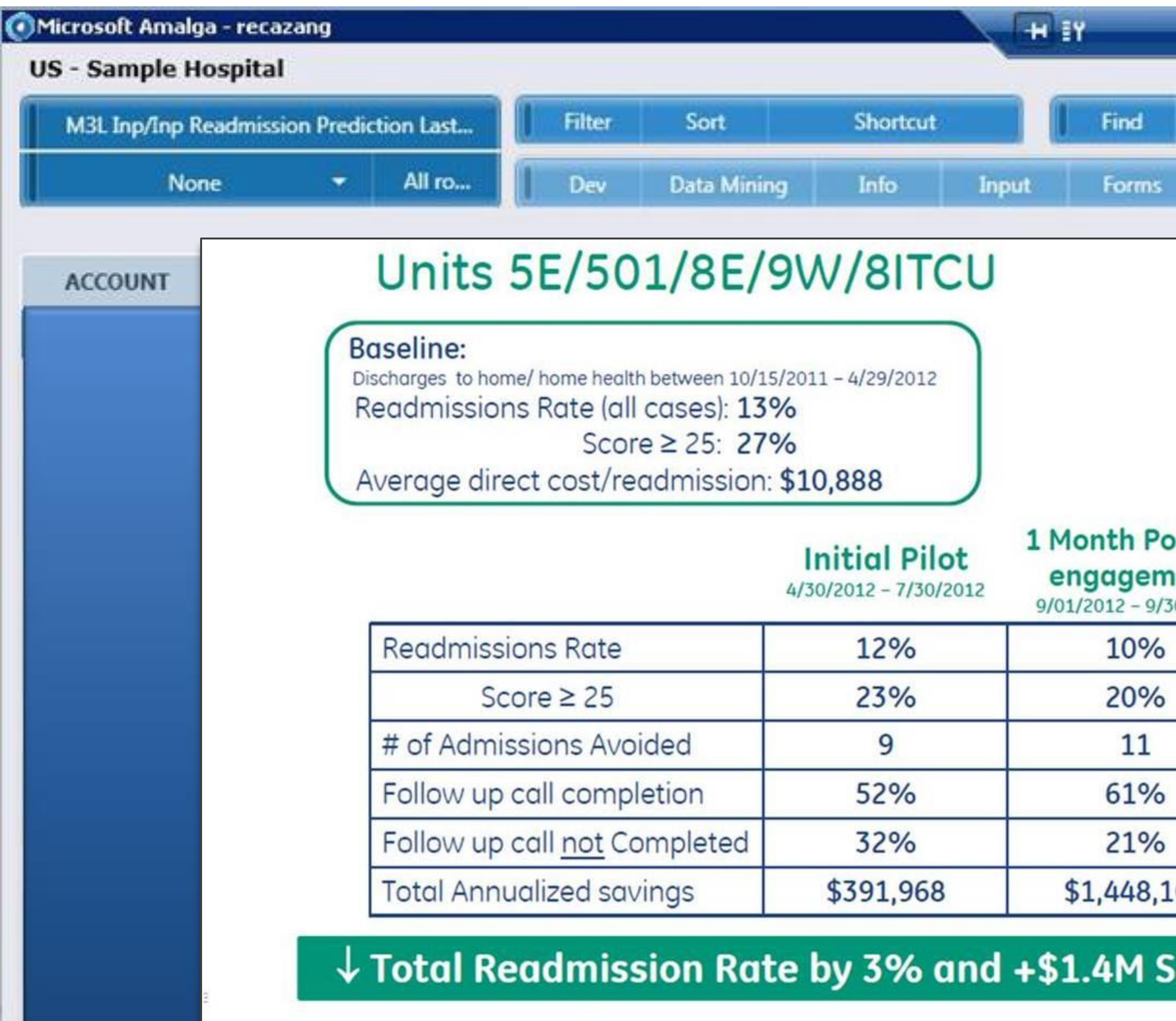
It is clear that hospital admissions and readmissions are becoming a critical parameter for tracking care delivery from both a financial and quality perspective.

Readmissions Manager for Microsoft Amalga is an innovative solution to help organizations address this very important business need.



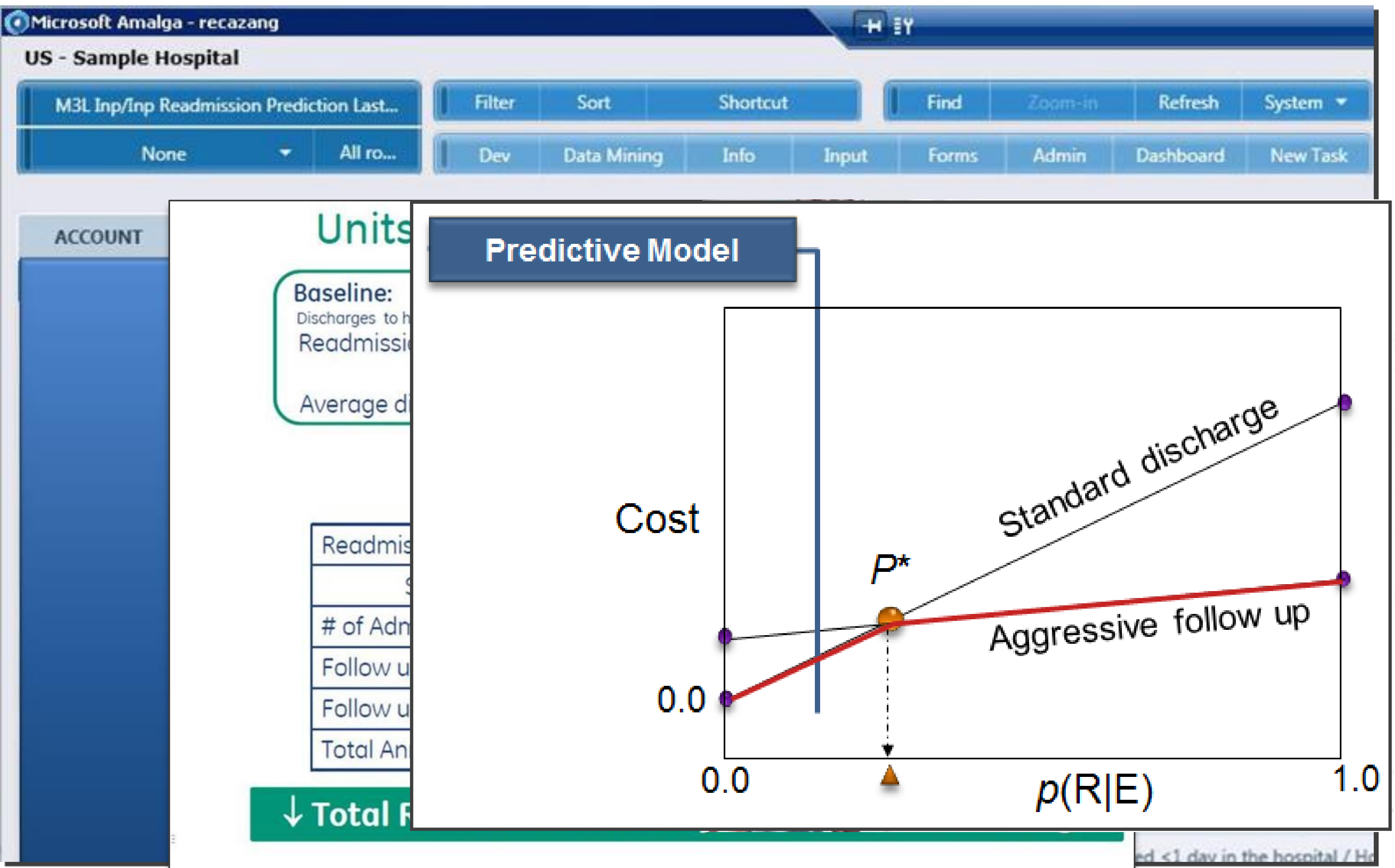
Readmissions Manager Targets Avoidable Hospital Readmissions

Direction: Predictions to Decisions

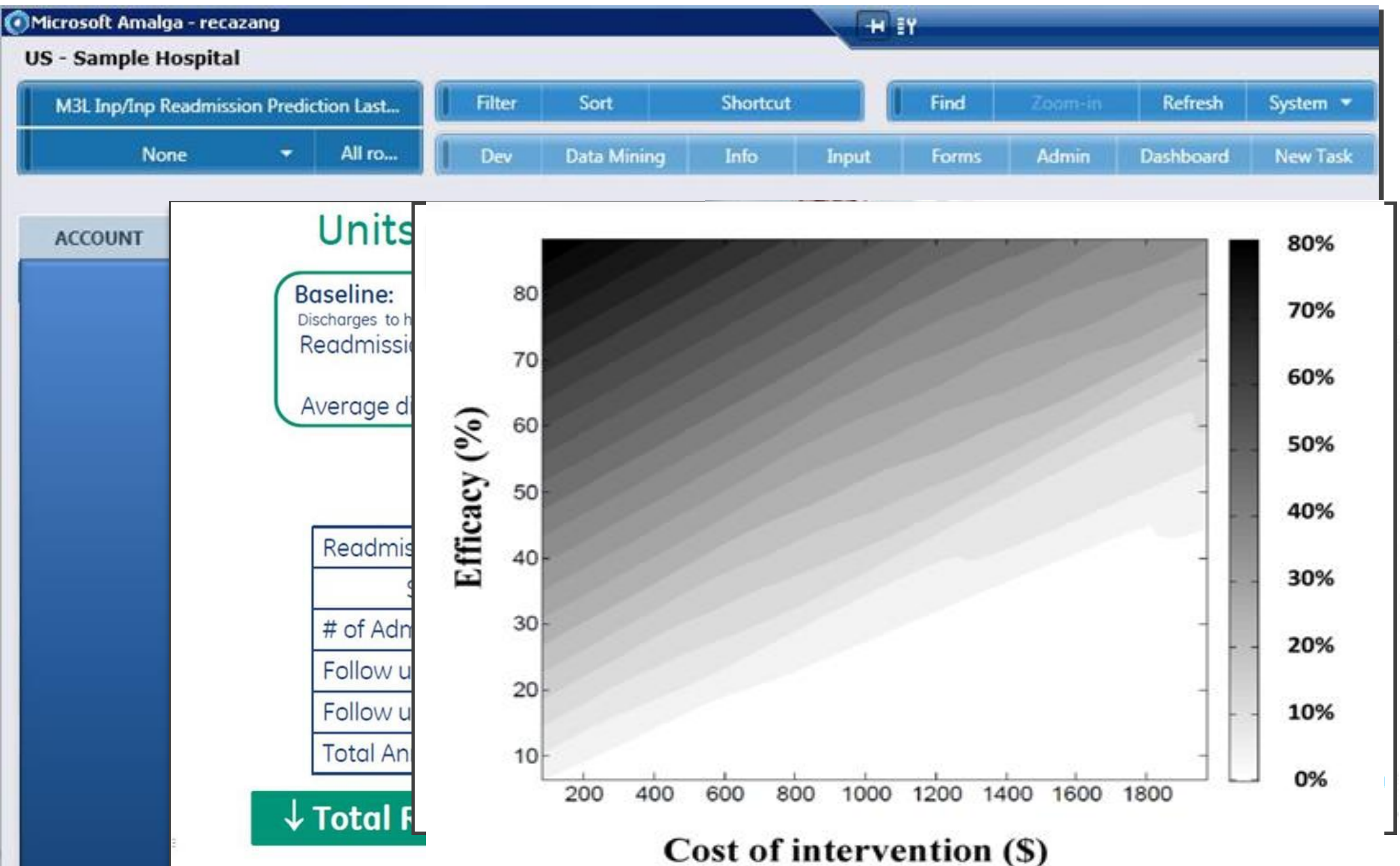


	Initial Pilot 4/30/2012 – 7/30/2012	1 Month Pos engageme 9/01/2012 – 9/30
Readmissions Rate	12%	10%
Score ≥ 25	23%	20%
# of Admissions Avoided	9	11
Follow up call completion	52%	61%
Follow up call not Completed	32%	21%
Total Annualized savings	$391,968	$1,448,10

Direction: Predictions to Decisions



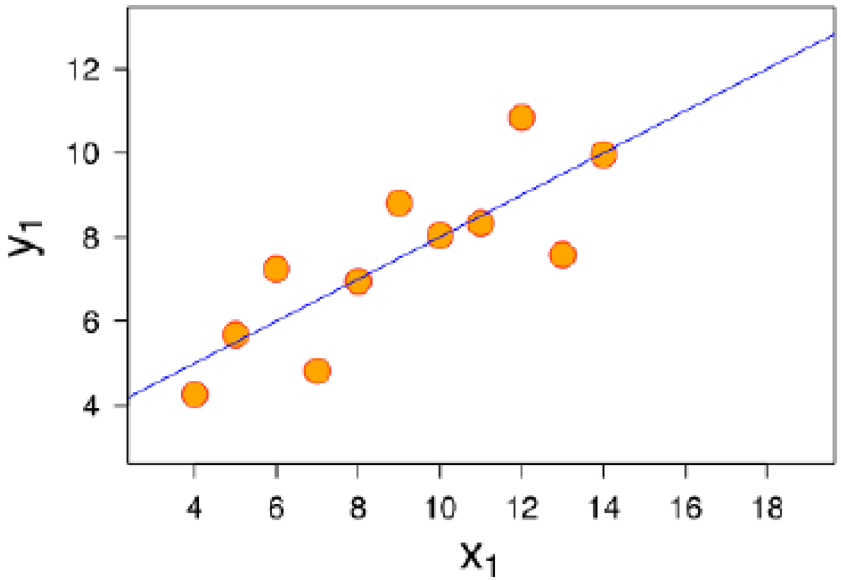
Direction: Predictions to Decisions



Direction: Interpretability & Explanation

DTTM	AGE	SEX	PROB_NUM_%	FACTORS_PRO_READMISSION	
	62	F	37.9	Num past 6m visits = 6 to 10 / Patient had dx = Disorders of fluid, electrolyte, an	stayed 3-6 days ir
	74	M	32.72	stayed <1 day in the hospital / Patient had dx = Disorders of fluid, electrolyte, and	Num past 3m visit
	48	M	30.83	Patient had dx = Chronic renal failure / 44 < Age < 60	Gap since first HF
	68	M	29.05	Patient had dx = Disorders of fluid, electrolyte, and acid-base balance / Patient ha	Num past 3m visit
	44	M	28.54		Gap since first HF
	61	M	27.36	Patient had dx = Acute renal failure / Patient had dx = Chronic renal failure	Num past 3m visit
	70	M	18.05	Patient had dx = Other personal history presenting hazards to health / Patient ha	Was NOT admitte
	68	M	16.57	stayed <1 day in the hospital	Was NOT admitte
	80	M	16.18	Patient had dx = Disorders of fluid, electrolyte, and acid-base balance / Patient ha	Num past 3m visit
	79	M	15.52		Num past 3m visit
	22	F	14.53	stayed <1 day in the hospital / Ave gap of past yr visits = between 15 and 30 days	Was NOT admitte
	25	F	14.42	stayed <1 day in the hospital / Patient had dx = Other personal history presenting	Was NOT admitte
	24	M	14.39	stayed <1 day in the hospital	Was NOT admitte
	53	F	13.59	stayed <1 day in the hospital / 44 < Age < 60	Was NOT admitte

Interpretability–Power Tradeoff



$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

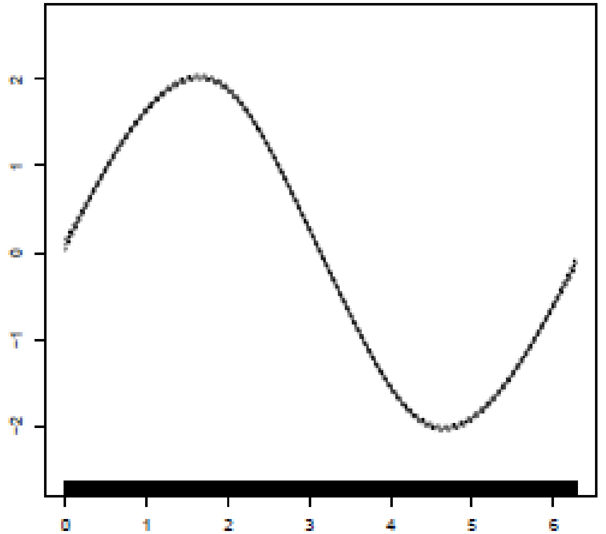
$$y = f_1(x_1) + \ldots + f_n(x_n)$$

$$y = f(x_1, \ldots, x_n)$$

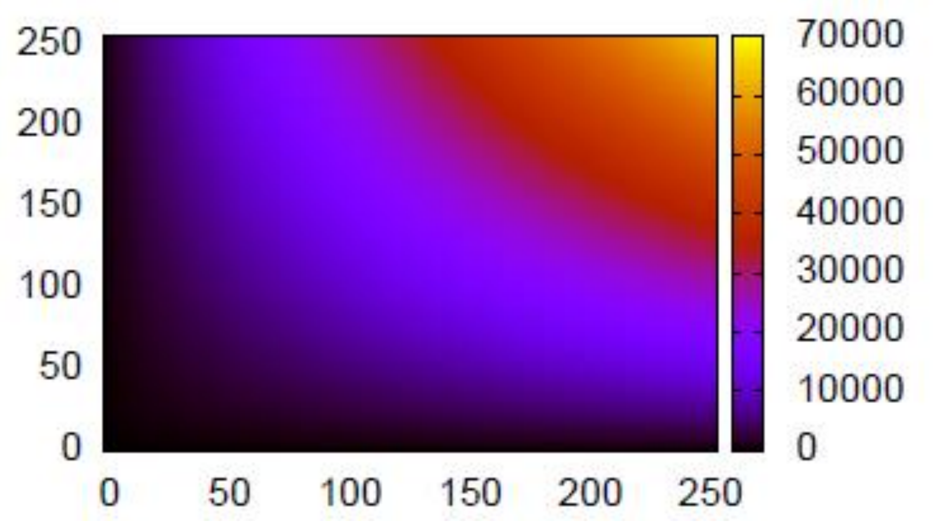
Capturing Key Interactions

Efficient means to identify pairwise interactions

$$y = \sum_i f_i(x_i) + \sum_{ij} f_{ij}(x_i, x_j)$$



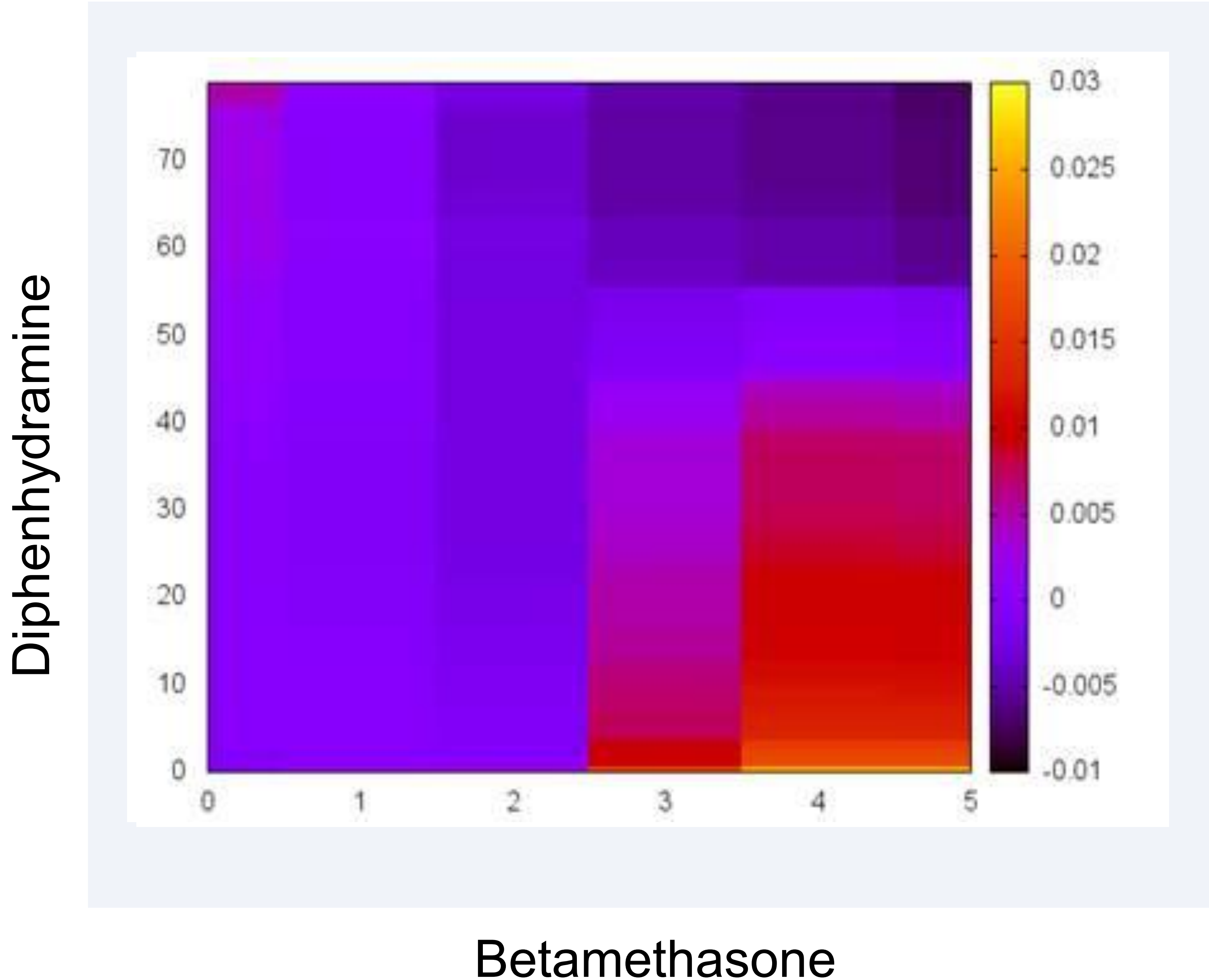
$f_i(x_i)$



$f_{ij}(x_i, x_j)$

Y. Lou, R. Caruana, J. Gehrke, and G. Hooker. Accurate Intelligible Models with Pairwise Interactions. In KDD, 2013.

Insights about Interactions

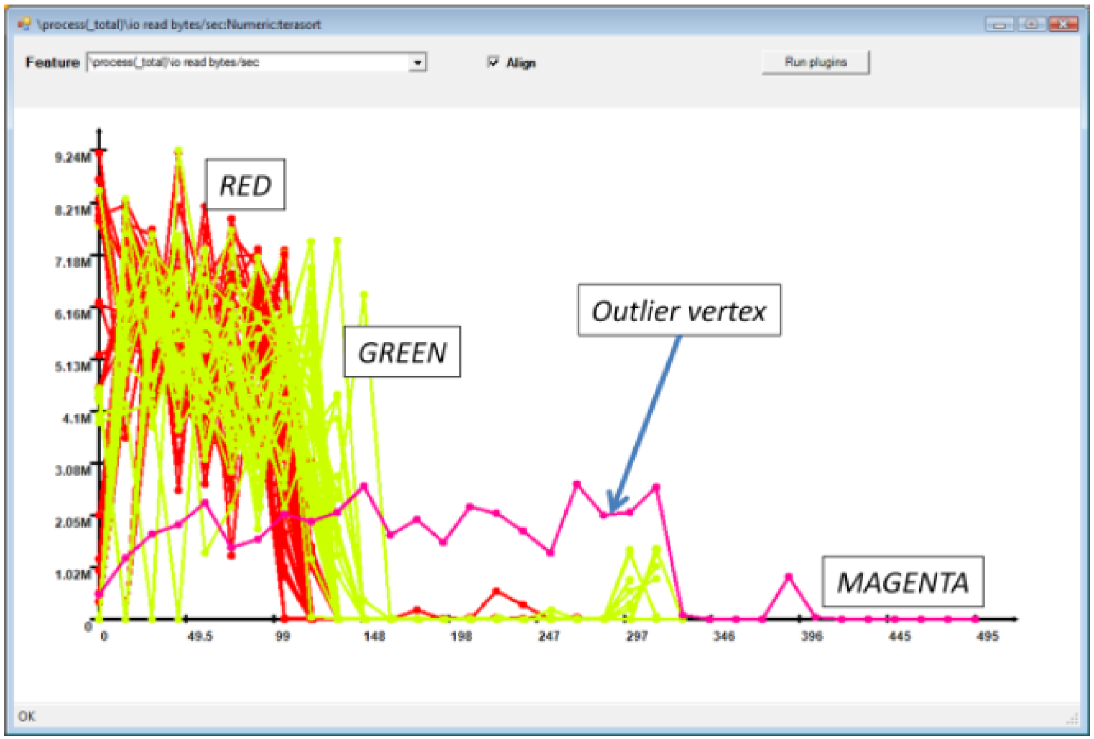


Direction: Identifying Causality

Predicting C. Difficile

- diabetes = TRUE
- history of C. Diffi = TRUE
- hospital service = gsg (general surgery)
- meds= acetylcysteine (n-acetylcys)
- meds = lidocaine hcl
- meds = clindamycin phosphate
- platelet count = C (thrombocytosis)
- unit = 2g
- albumin = L (hypoalbuminemia)
- admission source = transfer
- attending MD= XXXXXX
- unit = 2d
- CO2 = L (hypocapnea)
- city = XXXXXX
- employer name = Not Employed
- monocyte percent = H
- 70<=age<80
- wbc = H (white blood cell count)
- admission procedure = catheterizatio
- admission complaint =gastrointestinal
- last visit meds = fentanyl citrate
- meds = hydromorphone hcl

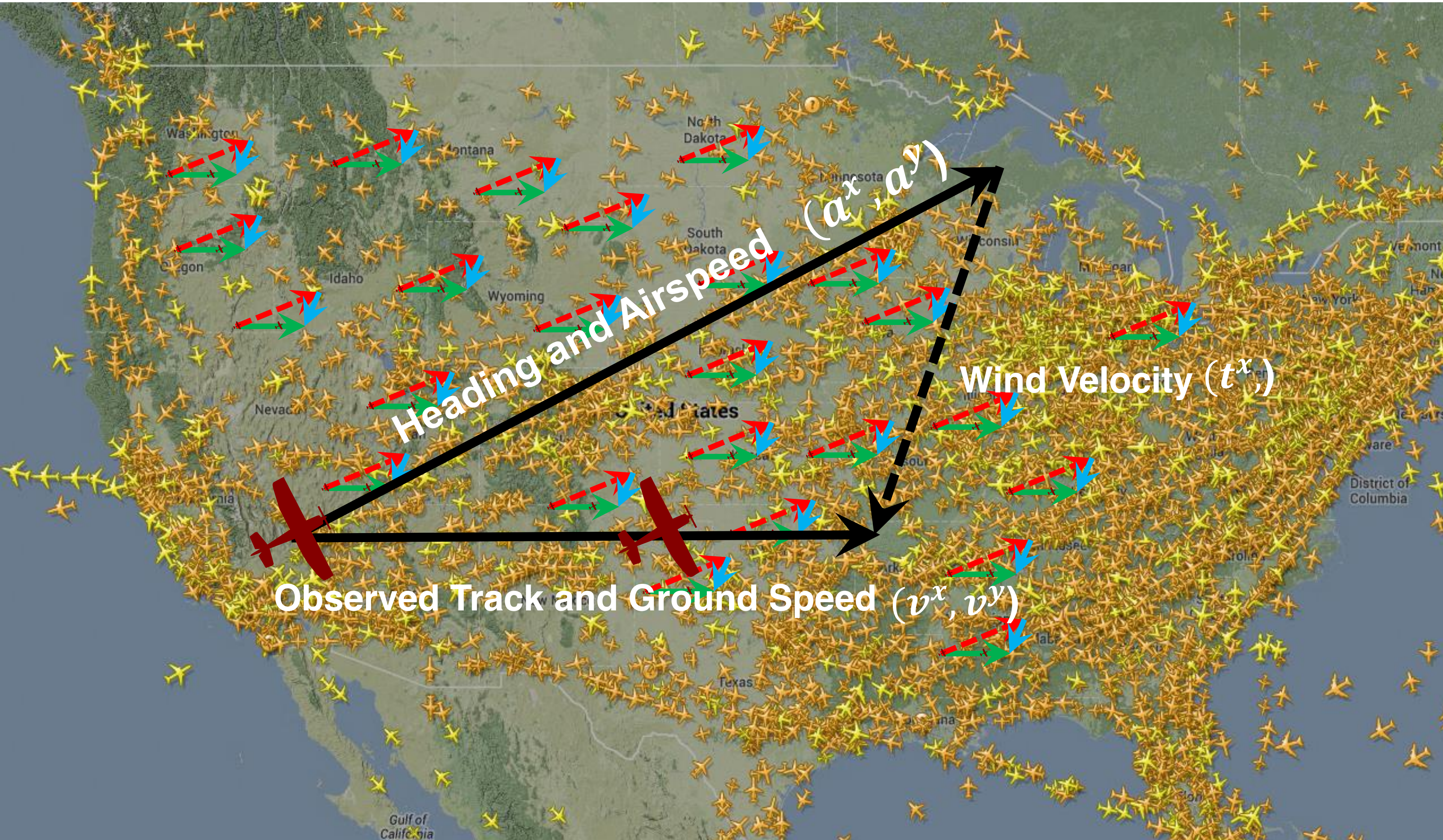
Root source of datacenter slowdown



J. Wiens, J. Guttag, E. Horvitz. [Patient Risk Stratification for Hospital-Associated C. Diff as a Time Series Classification Task] (NIPS 2012)

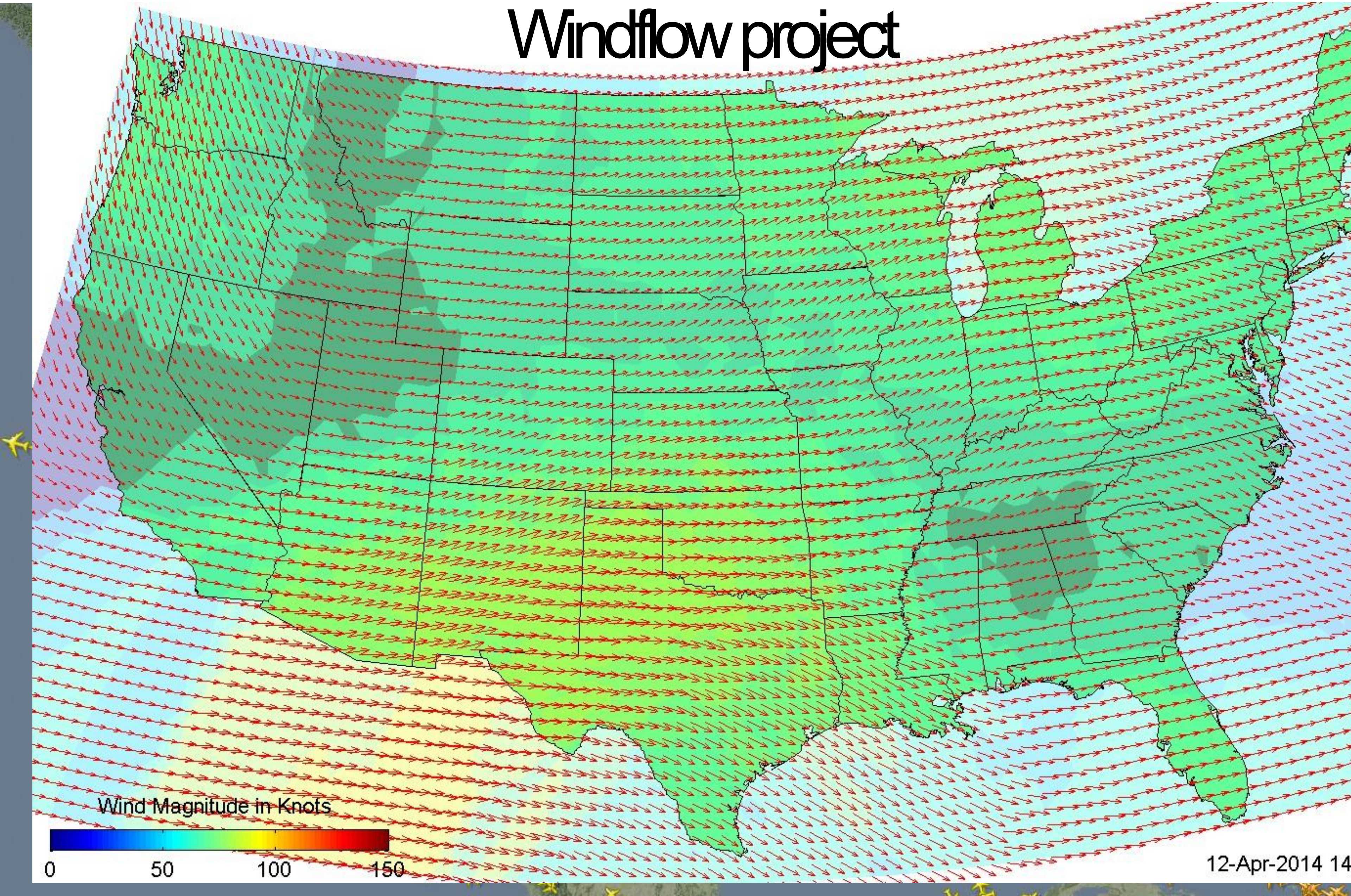
Direction: Selective Sensing

Airplanes Aloft as a Sensor Network for Wind Forecasting (IPSN 2014)
Access live inferences about continental windflows.



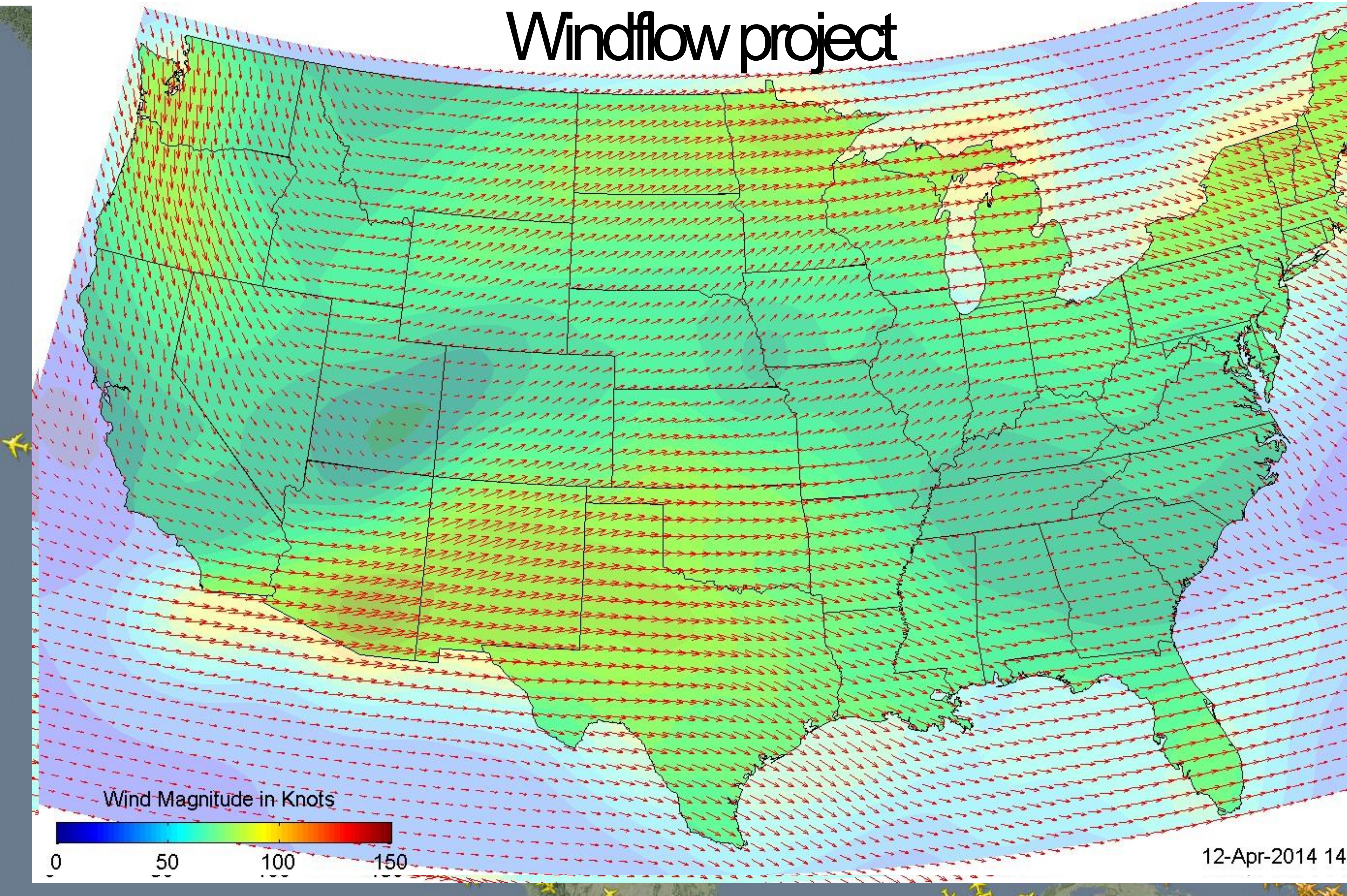
Direction: Selective Sensing

Windflow project



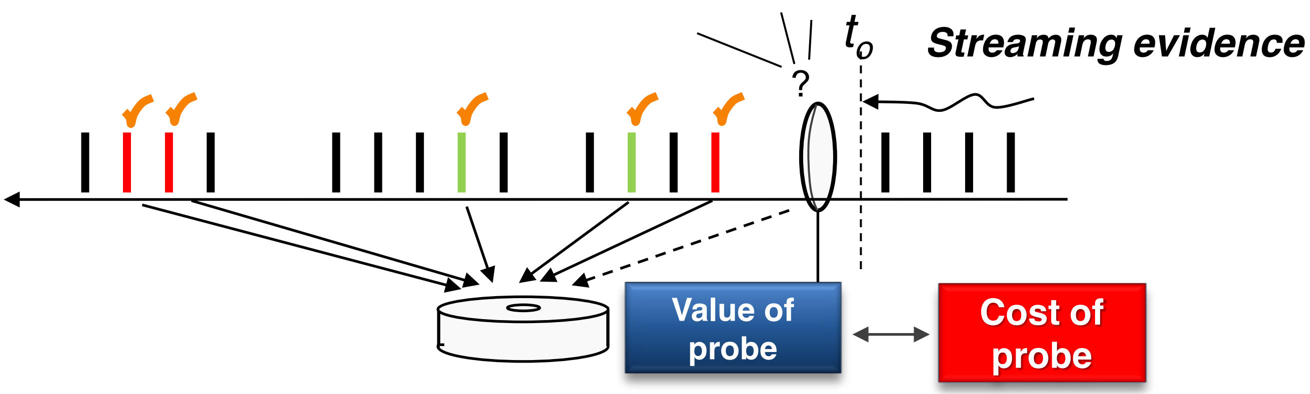
Direction: Selective Sensing

Windflow project



Direction: Active Learning

When do I need more data?
Value and cost of acquisition?



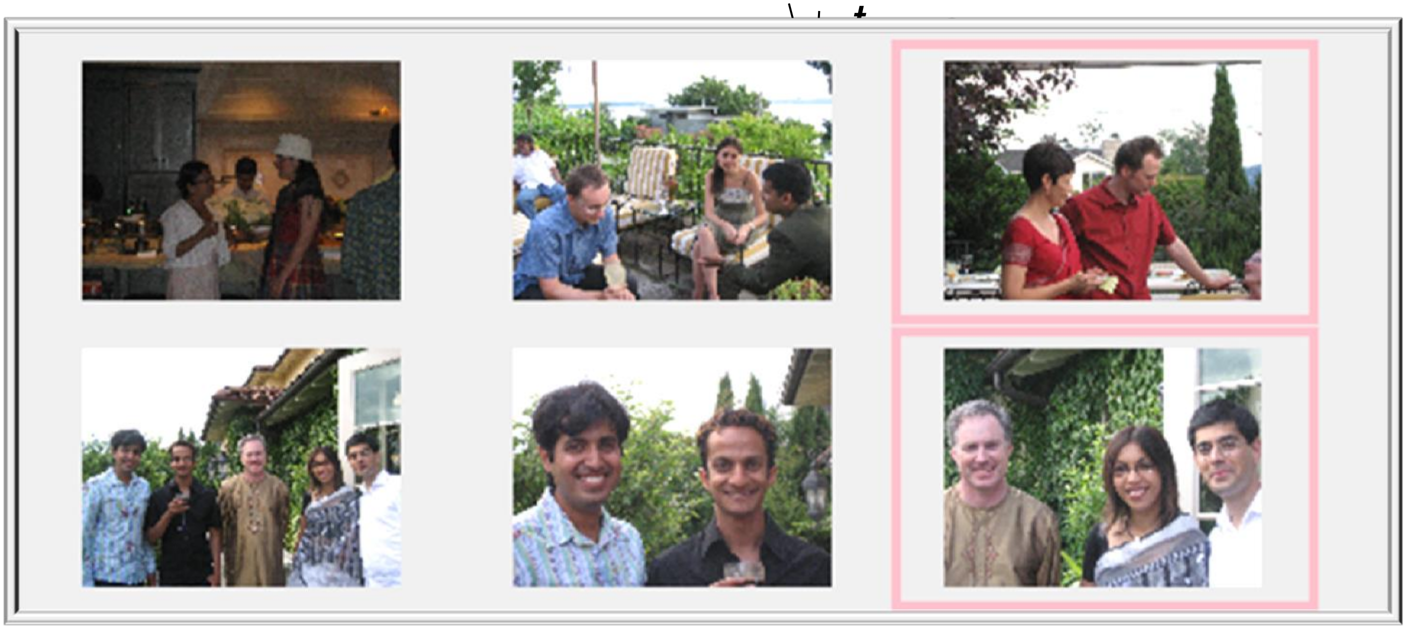
Kapoor & H. *Principles of Lifelong Learning for Predictive User Modeling* (UM 2007)

Kapoor & H. *On Discarding, Caching, and Recalling Samples in Active Learning* (UAI 2007)

Kapoor & H. *Breaking Boundaries: Active Information Acquisition Across Learning and Diagnosis* (NIPS 09)

Direction: Active Learning

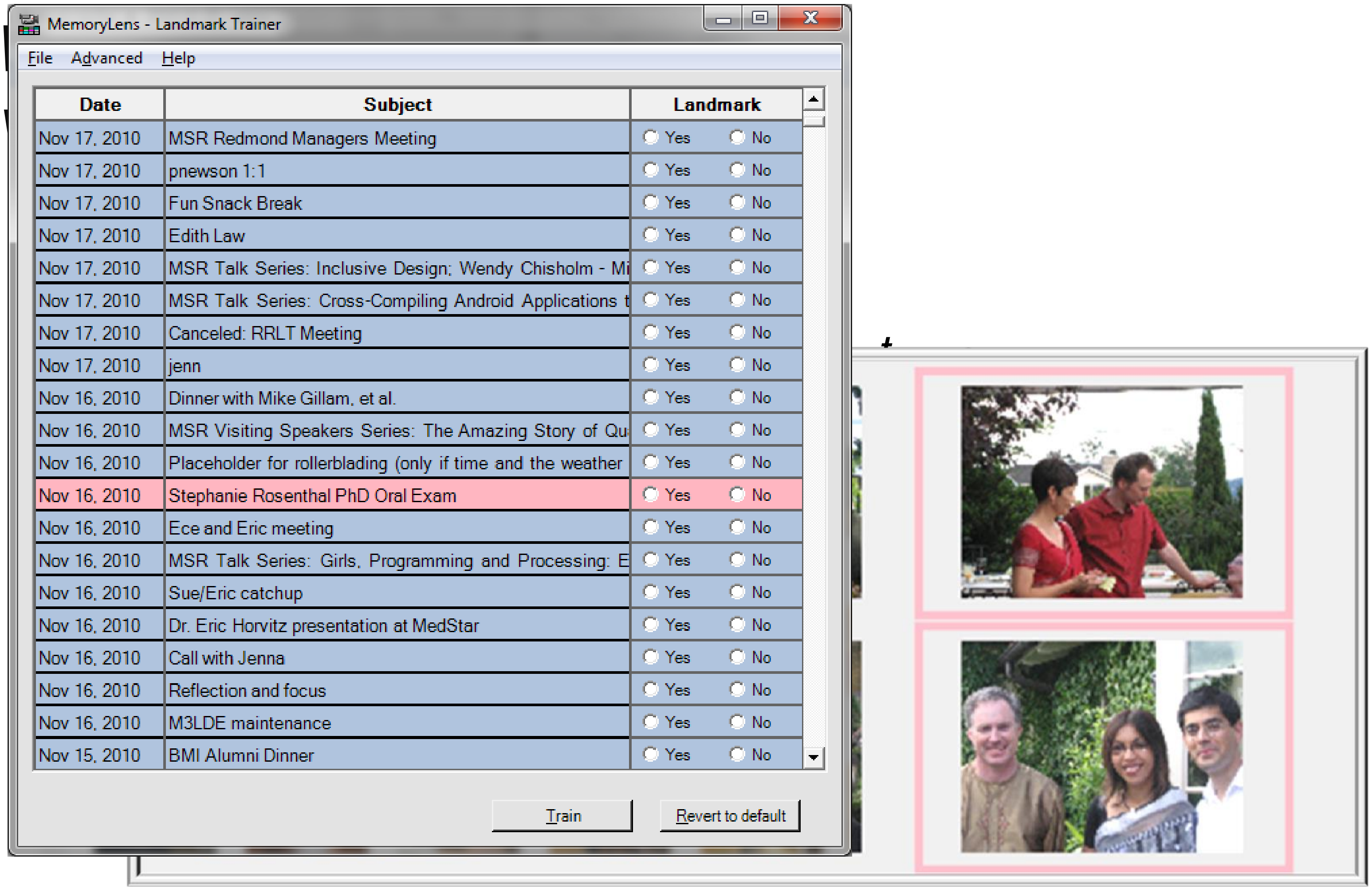
When do I need more data?
Value and cost of acquisition?



H., et al. *Learning Predictive Models of Memory Landmarks* (CogSci 2004)

Kapoor, et al. Selective Supervision: Guiding Supervised Learning with Decision-Theoretic Active Learning (IJCAI 2007)

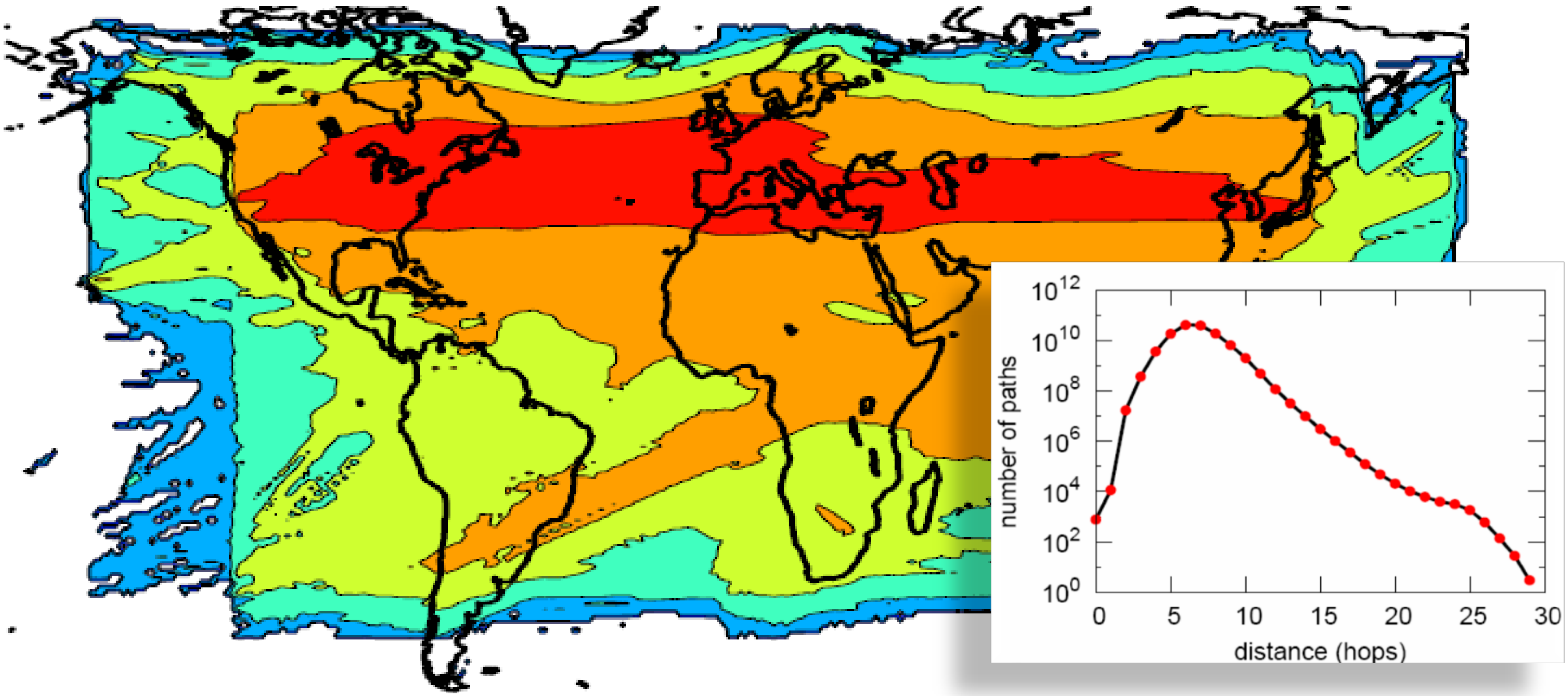
Direction: Active Learning



H., et al. [Learning Predictive Models of Memory Landmarks](#) (CogSci 2004)

Kapoor, et al. [Selective Supervision: Guiding Supervised Learning with Decision-Theoretic Active Learning](#) (IJCAI 2007)

Challenge: Sharing Industry Data



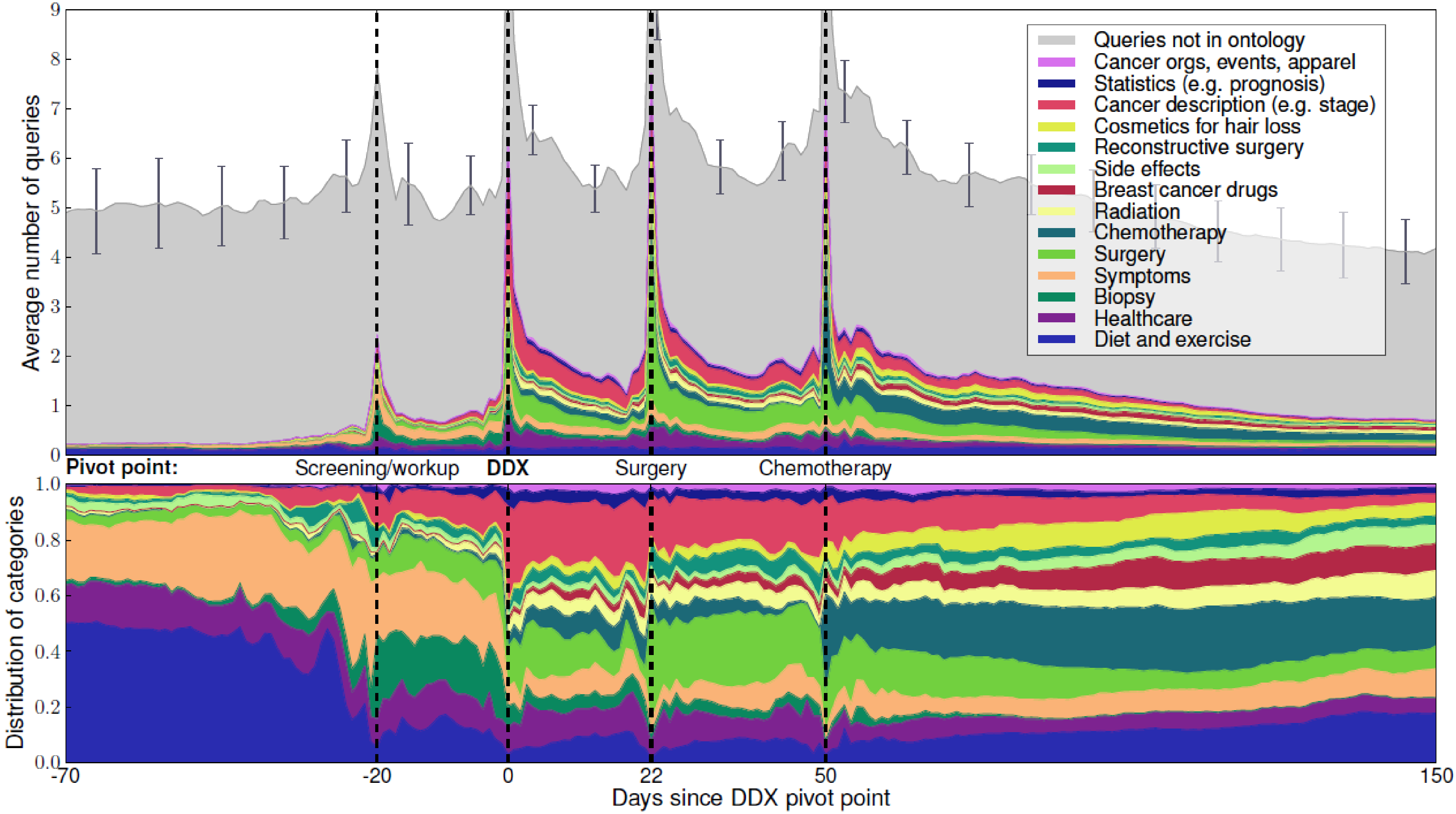
Messenger communication graph

30 billion conversations (30 days)

255 billion messages exchanged, 1.3 billion edges

J. Leskovec, H. [Planetary-Scale Views on a Large Instant-Messaging Network](#) (WWW 2008).

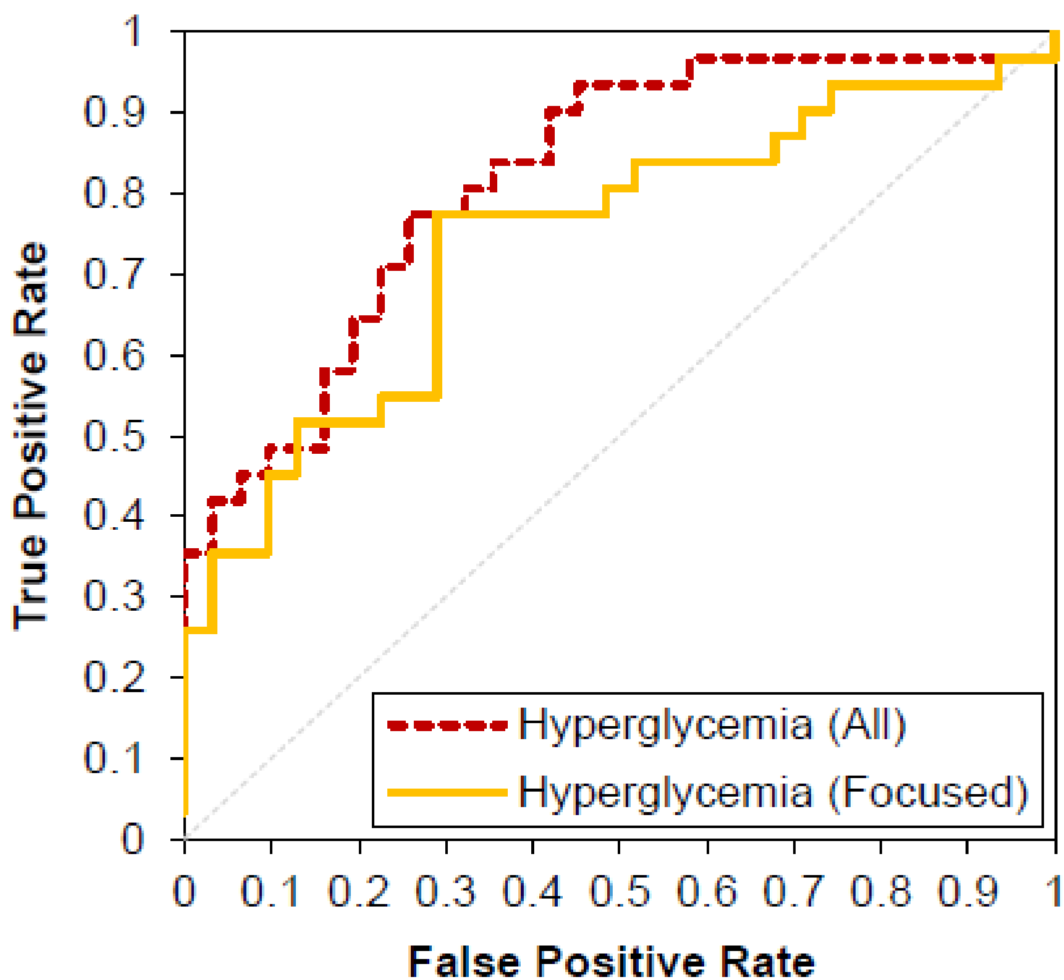
Challenge: Sharing Industry Data



M. Paul, R. White, H.

Challenge: Sharing Industry Data

Web-Scale Pharmacovigilance



Label	*Drug 1*	*Drug 2*
TP	dobutamine	hydrocortisone
TP	dobutamine	triamcinolone
TP	dobutamine	prednisolone
TP	betamethasone	dobutamine
TP	glipizide	phenytoin
TP	dobutamine	methylprednisolone
TP	prednisolone	salmeterol
TP	salmeterol	triamcinolone
TP	betamethasone	terbutaline
TP	dexamethasone	dobutamine
TP	budesonide	salmeterol
TN	hydrochlorothiazide	tazobactam
TN	clindamycin	montelukast
TN	lamotrigine	nystatin
TN	methylprednisolone	rosuvastatin
TP	budesonide	formoterol
TN	loratadine	nystatin
TN	hydroxychloroquine	prochlorperazine
TN	labetalol	sertraline
TN	ciprofloxacin	vecuronium

/	2.438, 3.094	< 0.0001
l	2.189, 2.767	< 0.0001

/2012 4:00:00 pm
5 pm

RFPs: Search Logs for Research

Workshop on Web Search Click Data, held in conjunction with WSDM 2009

February 9, 2009
Barcelona, Spain

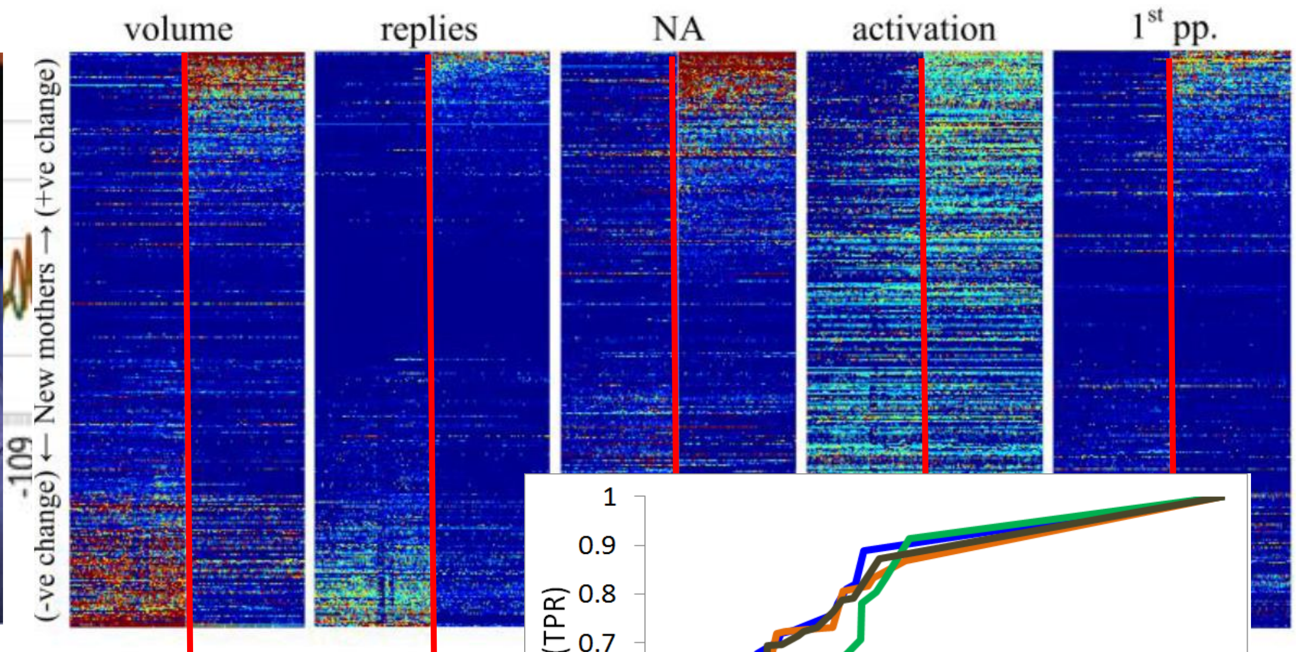
Organizers

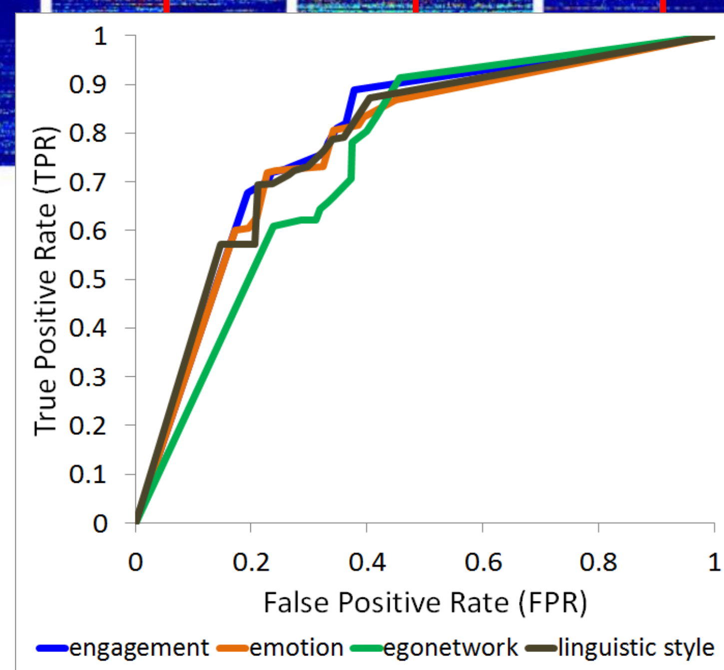
- Nick Craswell, Microsoft
- Rosie Jones, Yahoo! Labs
- Georges Dupret, Yahoo! Labs
- Evelyne Viegas, Microsoft

Workshop Program [Full proceedings at ACM.org, and video of talks at VIDEOLECTURES.net.]

9:00-9:05	Welcome and Introductions
9:05-10:00	Invited speaker: Alissa Cooper **A Policy Perspective on Query Log Privacy-Enhancing Techniques**
10:00	**Survey and evaluation of query intent detection methods** David J. Brenes, Daniel Gayo Avello and Kilian Pérez-González
10:30-11:00	Coffee Break
11:00	**Analysis of Long Queries in a Large Scale Search Log** Michael Bendersky and Bruce Croft
11:30	**Search Shortcuts Using Click-Through Data** Ranieri Baraglia, Fidel Cacheda, Victor Carneiro, Vreixo Formoso, Raffaele Perego and Fabrizio Silvestri
12:00	**Query Suggestions Using Query-Flow Graphs** Paolo Boldi, Francesco Bonchi, Carlos Castillo, Debora Donato and Sebastiano Vigna
12:30	**Intentional Query Suggestion: Making User Goals More Explicit During Search** Markus Strohmaier, Mark Kröll and Christian Körner

Direction: Privacy, Ethics, and Behavioral Data





Predicting before birth:

Who will suffer postpartum depression?

Predicting Postpartum Changes in Emotion and Behavior via Social Media (CHI 2013).

Microsoft Research Ethics Advisory Board

Researchers engage in structured, critical discussions with educated peers. Unproblematic designs approved via an expedited process, while red flags provoke a full review.

Direction: Datasets & challenge problems

Microsoft COCO
Common Objects in Context

Tsung-Yi Lin (Cornell),
Michael Maire (CalTech),
James Hayes (Brown),
Deva Ramanan (UCI),
Serge Belongie (CornellTech),
Pietro Perona (Caltech),
Piotr Dollar (MSR),
Larry Zitnick (MSR)

Rolled out at CVPR this coming week.

COCO: Common Objects in Context

COCO: images with objects in natural context

ImageNet: iconic images

<u>Commonsense</u>: Children (age 4-8) asked to name all objects seen in indoor & outdoor environments

→ 90 object types recognizable by 4 yr. old

.

Piotr Dollar, Larry Zitnick, et al. ECCV 2014, Zurich, Sept. 2014.

ImageNet: Iconic object images



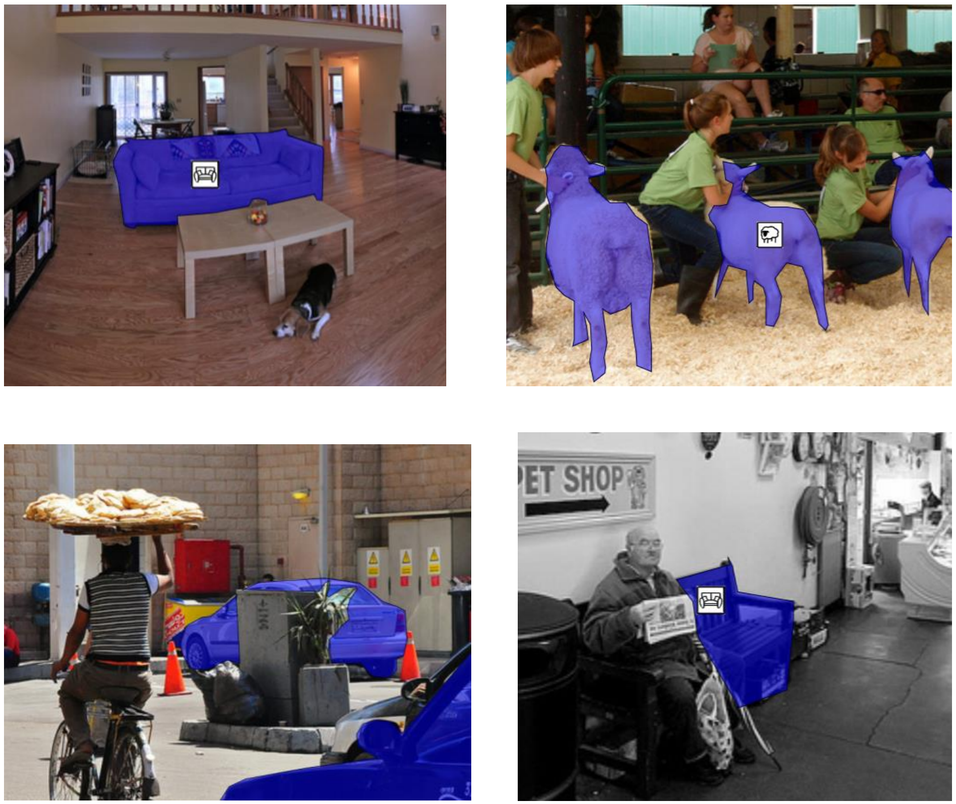

Iconic scenes



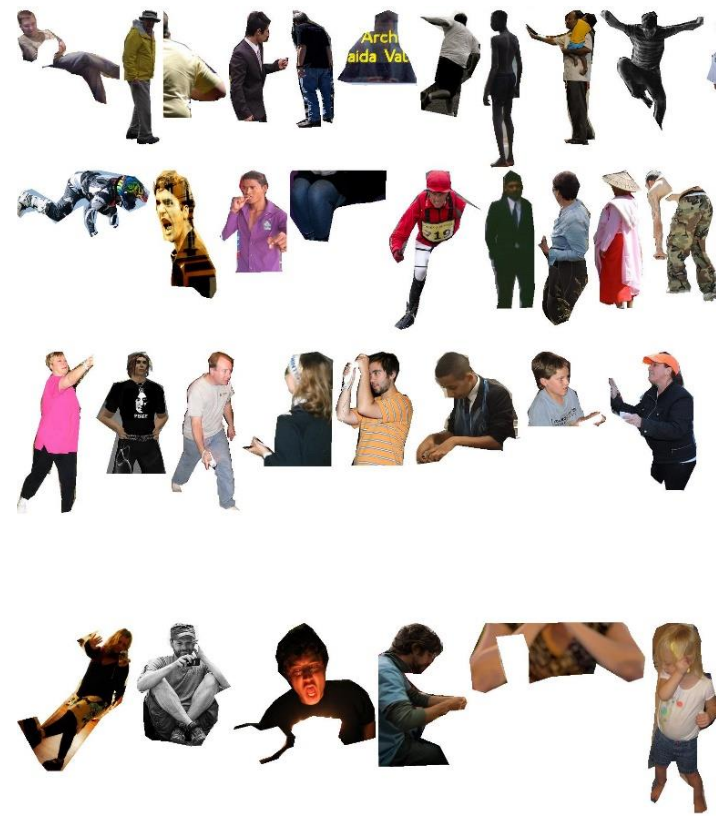

Non-iconic scenes



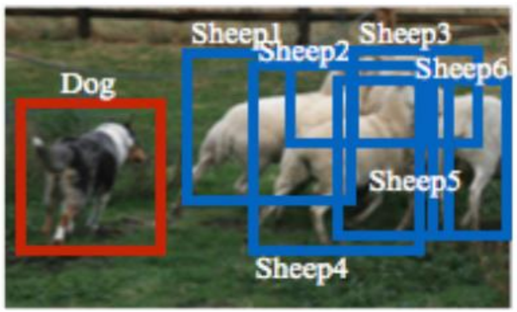

Contextual information



People in context

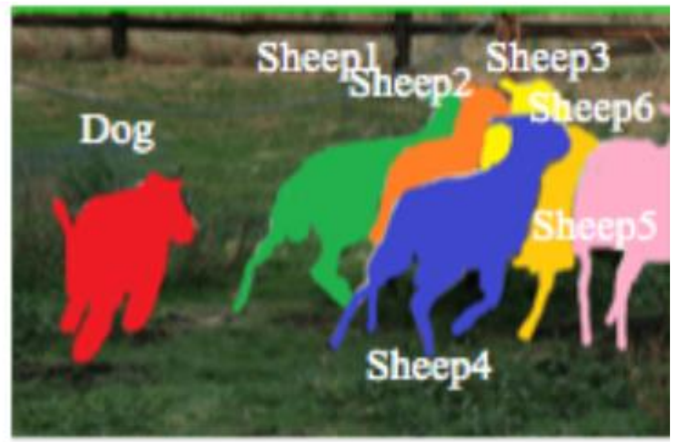


Object classification



Object detection

Semantic segmentation



90 categories
10,000 instances / category