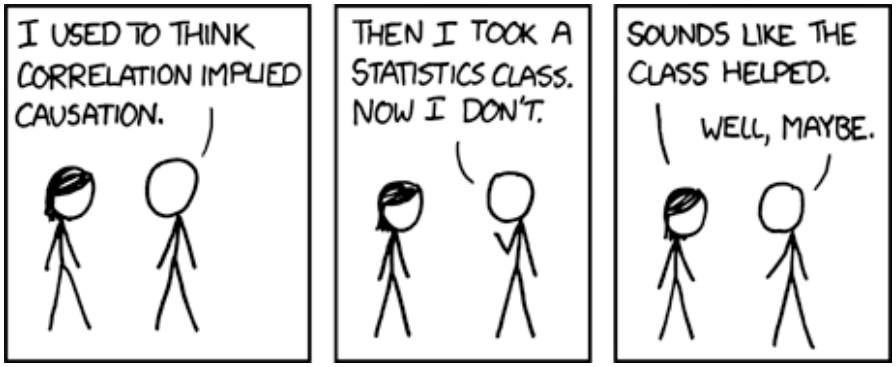
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Accelerating Discovery in the 21st Century

MARK GREAVES

PACIFIC NORTHWEST NATIONAL LABORATORY

The Changing Face of Science



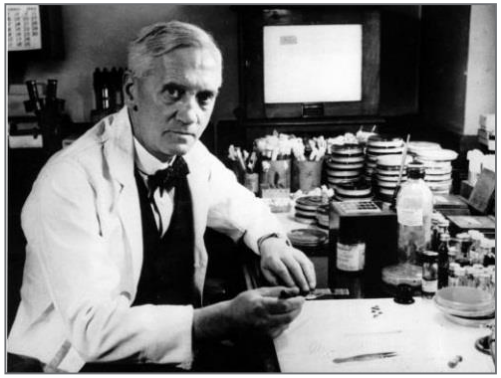
- **Most of scientific history:** observe, build a model and decide on a hypothesis about a phenomena, select an experimental method, derive data requirements, experiment, compare hypothesis against the data (often from a common repository), conclude a causal relationship.

- **Emerging 21st-century science**: start with the dataset(s), find correlations and suggestive patterns, propose a hypothesis, attempt to build model that explains why the correlation isn't spurious, test if you can…
  - Research in online social networks, computational linguistics, political science, etc., is largely driven by data availability
  - "Hard sciences" like microbiology, pharmacology, and materials science

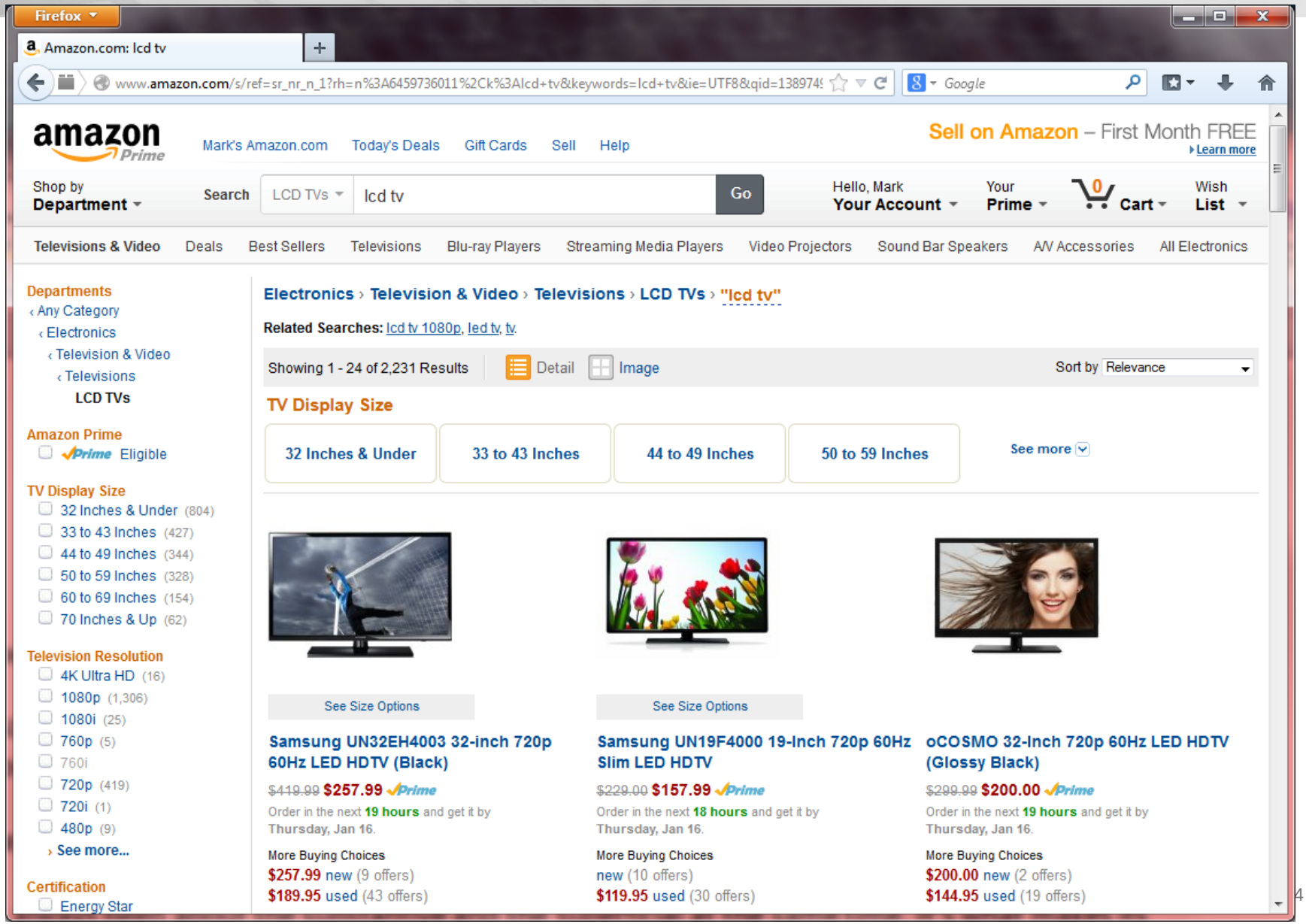
Synthesis is the new Analysis

Why is Synthesis Popular?

- **Innovation and Invention**
  - Invention is creating a new idea
  - Innovation is the application of the idea into a product
    - Ex: Penicillin: Fleming vs. Florey & Chain
- **20th-century Innovation**
  - Pattern: hierarchically-organized teams
  - Medium scale: Bell Labs, Xerox PARC
  - Large scale: Manhattan Project
- **Innovation via Massive Search**
  - A new innovation pattern
  - Couples with the increasing power of sensors, computers, instruments…
  - Success in ML systems
  - Ex: High-throughput drug discovery, Amazon page design



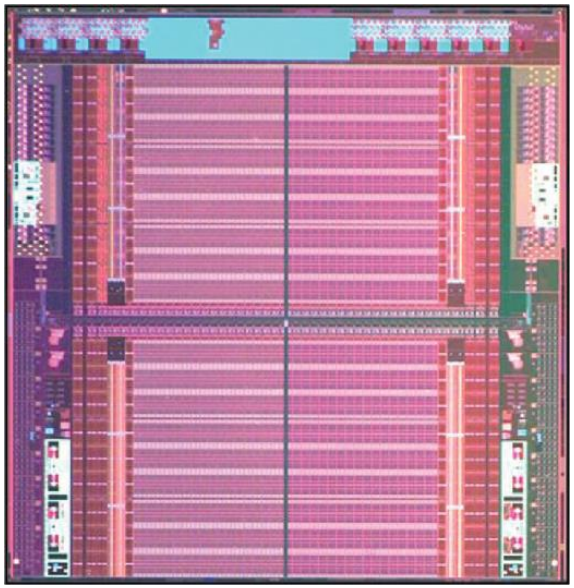
Amazon Web Page Design



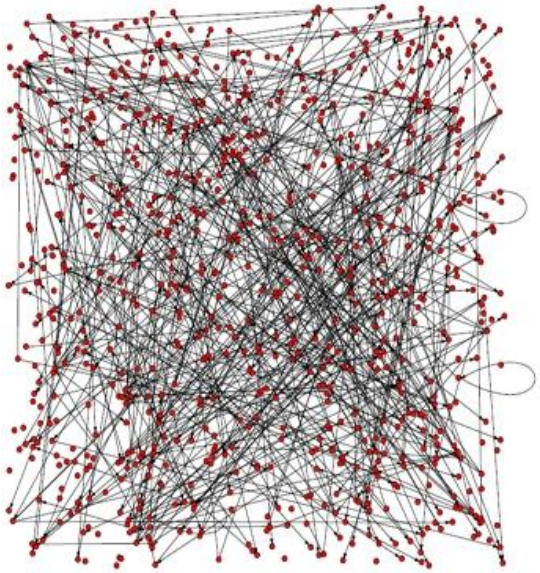
Solvable Problems by Scaling Search

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

364 Mbit SRAM



Arabidopsis Gene Regulation Network



Boeing 787 Full-Scale Simulation

- **Simultaneous real-time personalization of user experience and recommendations for billions of people**
- **Real-time language translation**
- **Machine learning for commerce (e.g., Target)**

Downside: Experimental Replicability

"An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment, [the complete data] and the complete set of instructions which generated the figures."

-- Buckheit and Donoho, "Wavelab and Reproducible Research" DOI 10.1.1.53.6201

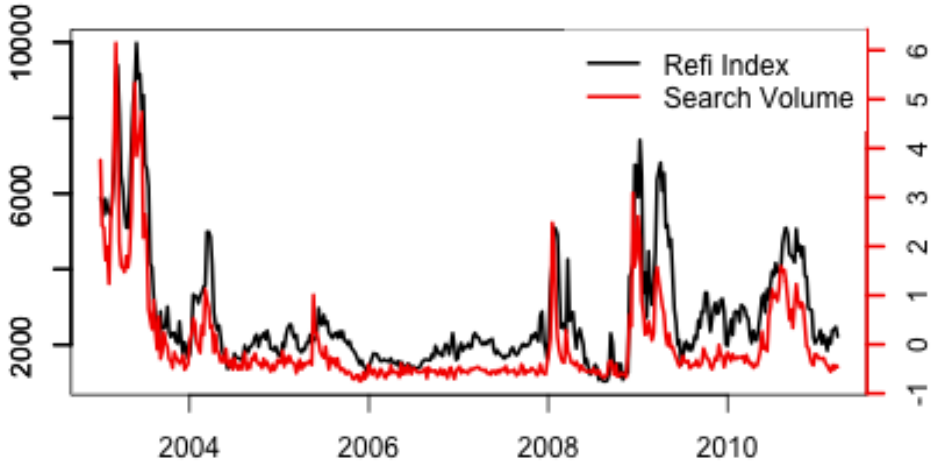
- **What does this include?**
  - Datasets
  - Data Collections
  - Algorithms
  - Configurations
  - Tools and Apps
  - Workflows / Scripts
  - Code Libraries
  - Services
  - Systems Software
  - Infrastructure
  - Compilers / Interpreters
  - Hardware
- **47/53 "Landmark" publications could not be replicated (Begley and Ellis, Nature 483, 2012).**

Difficulties at the Foundation

- **Scientists are now habitually collecting their own data at extremely high spatial and semantic resolution**
  - New generation of instruments and DIY analysis tools
  - Much higher probability of semantic divergence
  - Much higher amount of integration effort: scientists and engineers spend more than 60% of their time preparing data ("Computational modeling algorithms and Cyberinfrastructure" (NASA A.40, Dec 2011))
- **Data science should be as much a part of scientific training as foreign language training once was**
  - PNNL experience
- **Poor fit with the "standard" Kuhn model of scientific progress**
  - Data is much more temporally, spatially, and thematically individuated; leading to a "methodology per scientist" problem
  - Known replication issues and little open data means there is less trust
  - Variability makes it more difficult to identify conflicts between theories

Parallels to Intelligence Analysis

Search volume of refinancing calculator



Mohebbi, et. al. "Google Correlate" 2001

- **Analysts have enormous amounts of data available to them**
  - Data mining and correlation tools substitute for causal explanation,
  - Little open data and very little replicability
- **Forecast: Repeatable quantitative analytic argument is getting more difficult**
  - What counts as progress or as paradigm shift?
- **Forecast: Both face the Complexity Brake**
  - Scaling and correlation analysis is not sufficient to understand "wicked" systems

A Common Challenge

- **Scale is becoming the dominant meta-feature of science practice**
  - Psychological competition for grants and publications
  - Democratic supercomputing: You don't have to own a supercomputer to have access to supercomputing capability
    - Amazon added more capacity to its cloud offering in each day of 2013 than in the entirety of 2008
- **Correlation analysis has been fruitful at finding non-obvious connections**
  - Human inability to find complex patterns in high-dimensional spaces
  - Lots of problems have a satisficing correlation solution. Causality is overrated
- **But… causality is still the gold standard for explanation and theory**

What can analysis learn from progress in *discovery informatics*?

Familiar to this Audience…

Lesson 1: Don't Make the Applications Smarter; Make the Data Smarter

Why Should We Do This?

- **(We believe that) Smart data yields higher returns in productivity than better applications**
  - Avoid New Cuyama
  - Realization during the DAML Project
    - In 2001, DAML was about Reading the Web
    - In 2005, DAML was about Publishing Your Data
  - Higher efficiencies in scientific workflows, particularly in life sciences
- **Much of the work of the semantics technology community is here**
- **Success is hard**
  - Building good class hierarchies and ontological rules is often annoying to the SMEs, and exceptions are constant
    - Open World Assumption allows for this, but blocks many desirable inferences
    - Global inconsistency with local consistency
  - Sometimes we have to mostly solve the problem before creating the right representations with the right distinctions
    - Clear thinking is its own reward
  - DARPA SIMPLEX

Maybe Not So Familiar to this Audience?

Lesson 2: Rebalance Human Creative Effort

- **What is the role of semantics in directly aiding causal discovery?**
- **Semantic tech gives us a way to describe set-based regularities in the world**
  - Many of these regularities are not fully describable using sets and truth functions
  - The level of formalism we can do is incompatible with the uncertainty and vagueness in the data
- **Classical AI for science attempted to build expert systems <u>to aid human discovery and hypothesis forumulation</u>**
  - The DENDRAL project (1965… Feigenbaum, Buchanan, Lederberg, Djerassi)
    - Formulate hypotheses about unknown compounds and chemical structures from mass spectroscopy data
    - Heuristic Dendral and META-Dendral
  - Automated Mathematician (1977, Doug Lenat)
    - Generate, modify, and combine LISP fragments that correspond to mathematical concepts
    - Lots of heuristic rules
    - Found the Goldbach conjecture and the unique prime factorization theorem
  - BACON, DALTON, and several more.

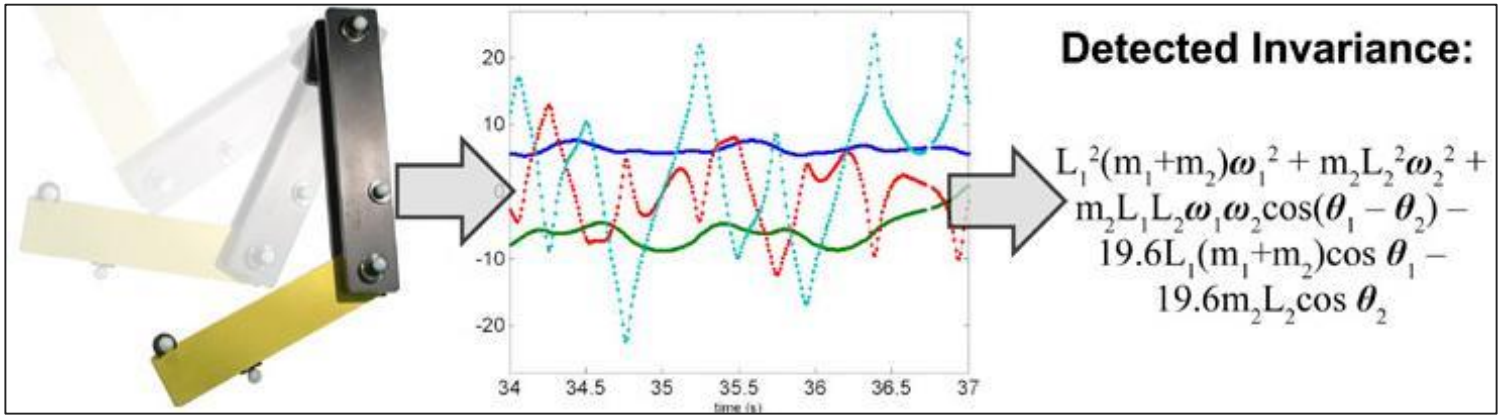
Discovery Informatics

- **How can we rebalance the effort of humans and machines?**
- **How can we use AI and semantics techniques to go beyond data description and to support human discovery?**

- **"AI Winter" killed most AI in science**

- **Examples of the resurgence of classic AI discovery themes:**
  - Eureqa
  - EU Large Knowledge Collider (LarKC) project
  - Netflix Quantum Theory
  - DARPA Big Mechanism Program
  - IBM Watson and Cognitive Computing
  - PNNL Analysis in Motion

Eureqa



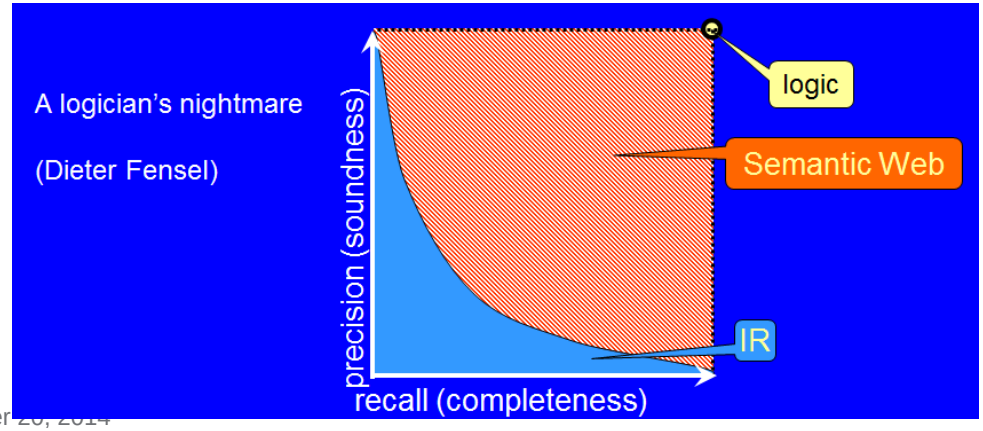
- **Eureqa examines data from an experiment, and produces equations that explain what happened**
  - Genetic algorithm search through sets of governing equations, using symbolic constraints
  - Searches through a large space of possible models
  - Multiple hypotheses possible

Eureqa

Large Knowledge Collider (LarKC) Project

- **Deductive inference with a given set of axioms at the Web scale is practically impossible**
  - Too many triples to process;
  - Too much processing power/time is needed
  - Data snapshots too difficult

- **LarKC aimed at contributing to a scalable Semantic Web reasoning platform via three techniques:**
  - Giving up on completeness
  - Combining heuristic search and logic reasoning into a new process
  - Parallelization



Netflix and the Limits of Data Analytics

VS.

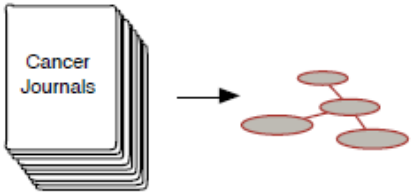
Slapstick Bounty-Hunter Movies
Set in Australia/NZ
About Reunited Lovers

- **Netflix largely ignored the winning Netflix Challenge algorithms**
  - Transitioned only Matrix Factorization and Restricted Boltzmann Machines
- **The Human Element**
  - Human appreciation of an object comes because of the contexts and associations aroused the user, and the cognitive embeddings
  - This is as true of a Netflix movie suggestion as it is of an analytic conclusion
- **"Netflix Quantum Theory" goes beyond the 5 star system**
  - Provide the human meaning that simple correlation analysis does not
  - Embed movies in human socio-linguistic structures

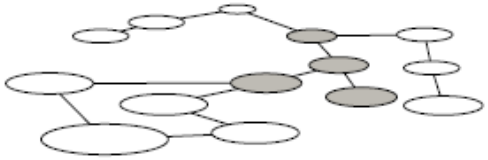
http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/

DARPA Big Mechanism

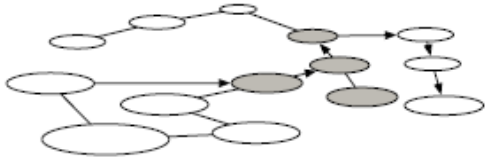
Technology Development Tasks



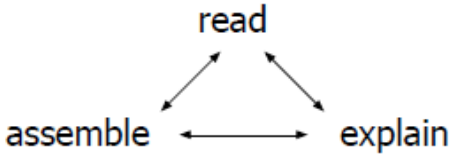
Read papers in cancer biology and extract causal fragments of signaling pathways, represented at all relevant semantic levels.



Assemble causal fragments into more complete pathways; discover and resolve inconsistencies.

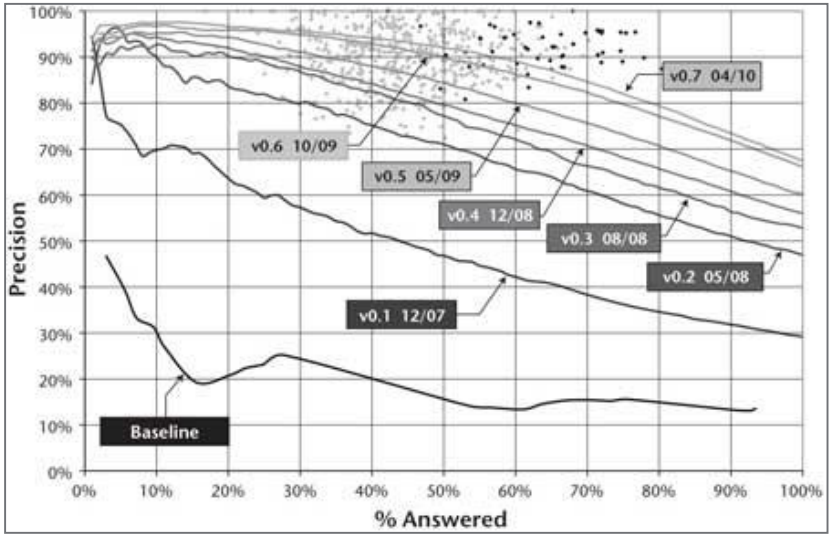


Explain phenomena in signaling pathways. Answer questions, including "reaching down to data," when it is available.



Integrate reading, assembly and explanation in a non-pipeline architecture that provides flexible control.

Watson for *Jeopardy*: Key Features



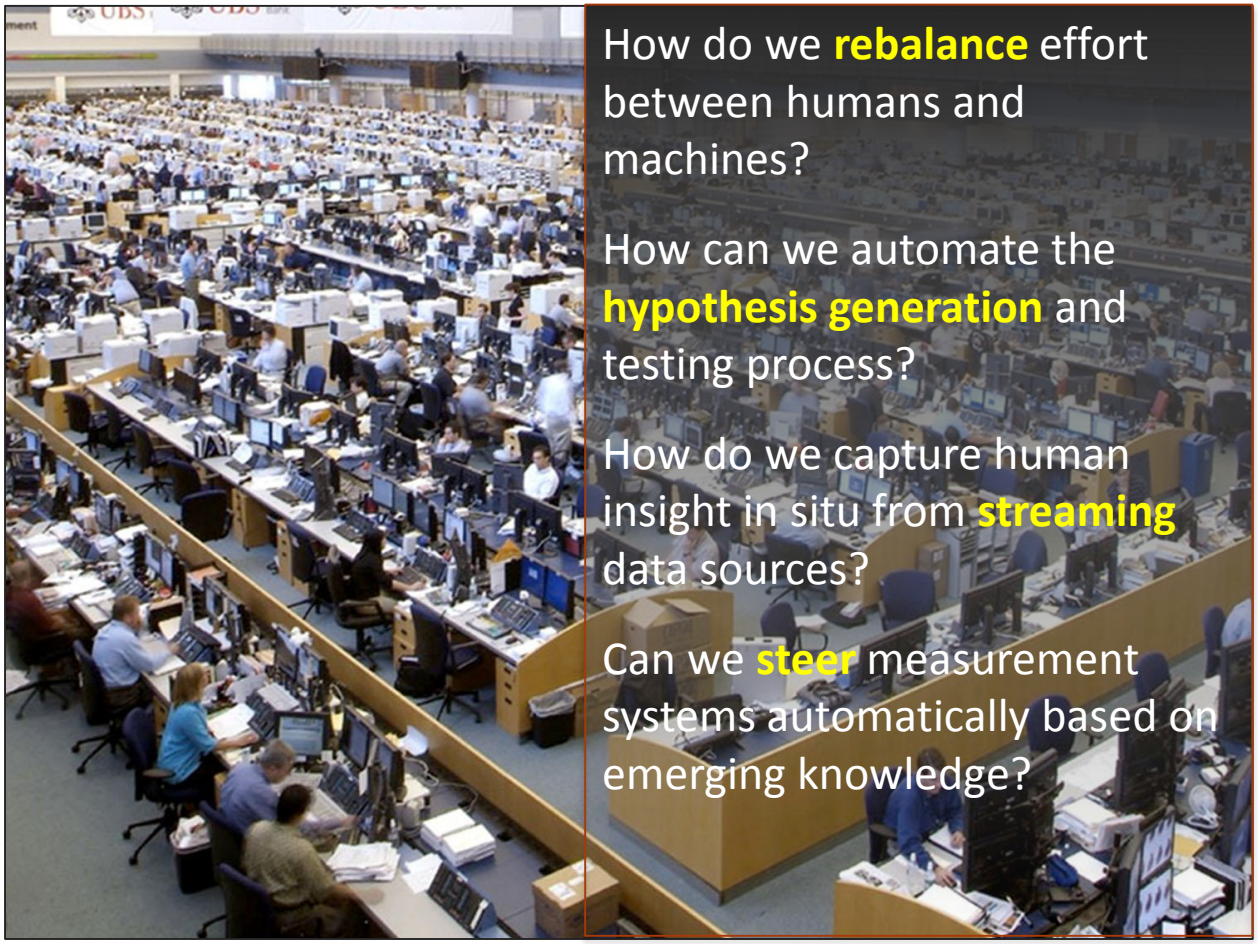
- **IBM Jeopardy Power7 cluster**
  - 2880 POWER7 cores at 3.5 GHz
  - 16 Terabytes of memory
  - 80 Teraflops, #94 on Top500
  - ~$3 million
  - Run DeepQA in <3 sec
- ***IBM Journal of R&D*, May 2012**
- ***Jeopardy's* central graph**
  - Metric: be in the winner's cloud
  - Multiple DeepQA systems at different levels of performance
  - Constant testing
    - ~40K official *Jeopardy* QA pairs
    - New QA pairs easy to create
    - Decomposable metric
    - Factoid answers

Lessons From Watson

- **Recognition that *Jeopardy* could be modeled**
  - An empirically-grounded model of 100s candidate Q-A pair types
  - A learned model of the ability of each solver to accurately answer a question type
  - A complete model of the Jeopardy rules, objectives, and buzzer management
  - An large but incomplete model of needed domain knowledge
  - Needed knowledge is static and mostly available

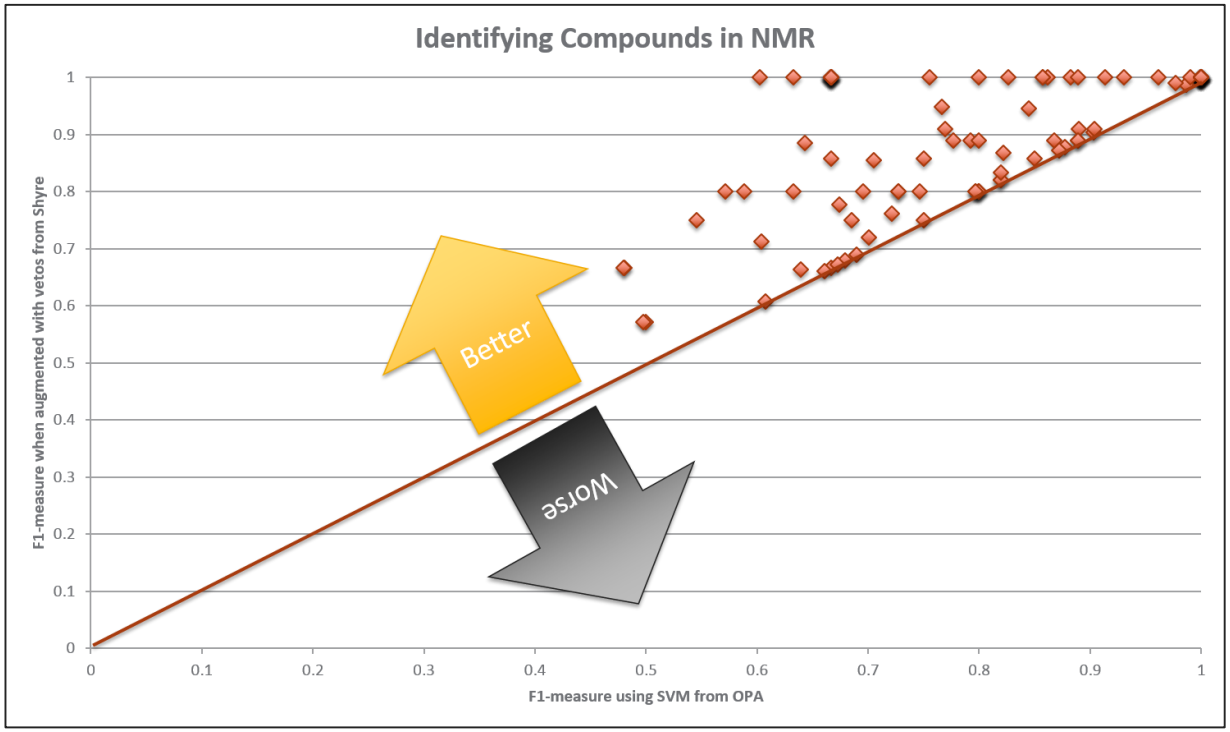
- **Some Key Watson Innovations (with Jim Hendler)**
  - Validation of a AI parallel "feed-forward" architecture
  - No reduction to a common statistical model or single logical KR – integration of lots of small things in a very large memory
    - Embrace Data Heterogeneity: Language-based interlingua vs. fixed database schema or pre-built formal ontology
  - AI as a large collection of small processes that are orchestrated with a context, vs. a small collection of very general processes
    - Question-Answering Architecture: Build a haystack, then find the needle vs. "1-5 carefully designed algorithms to rule them all"

Analysis in Motion

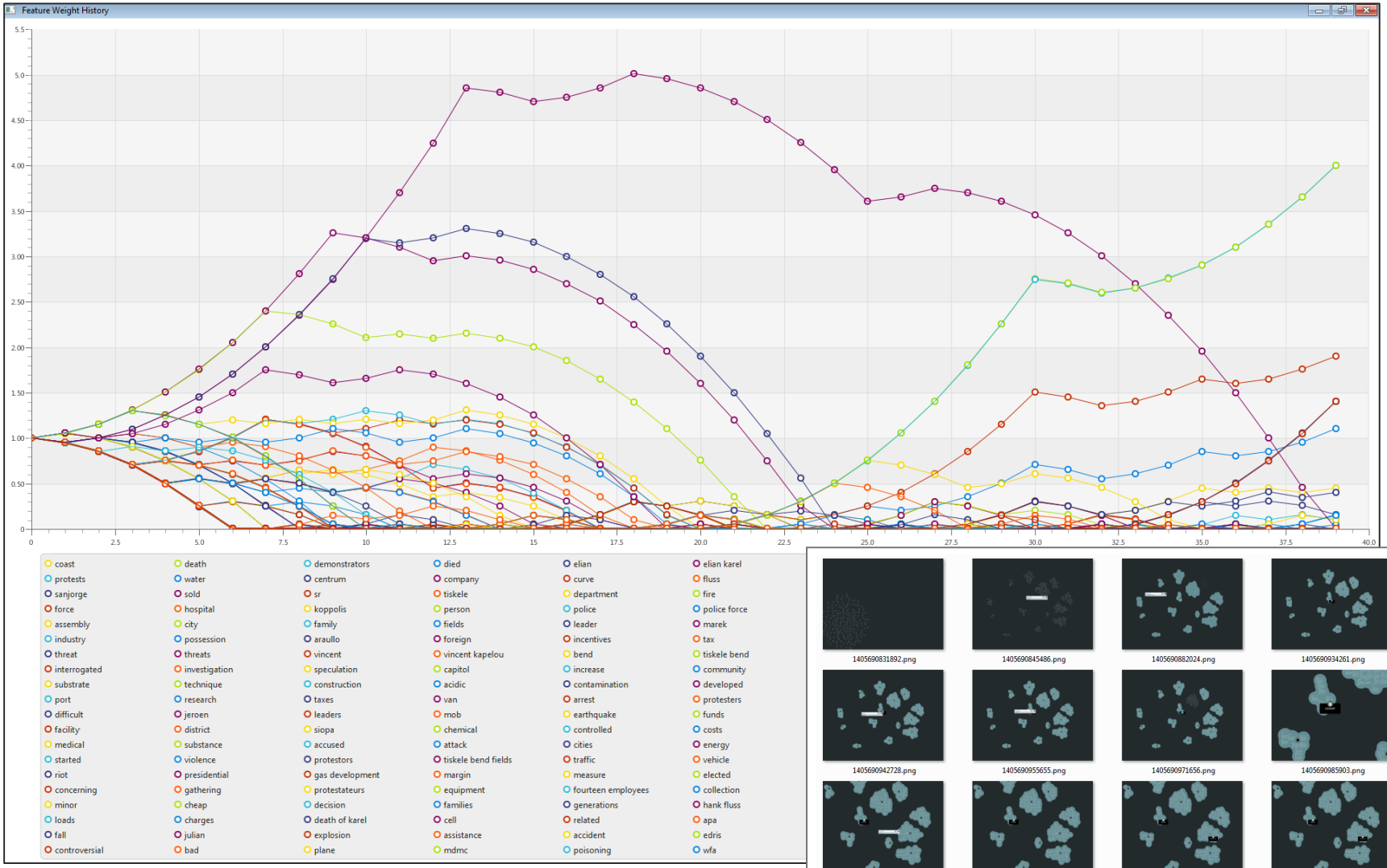


AIM Hypothesis Generation and Testing

- **Initial outcomes**
  - A prototype streaming symbolic reasoner with eviction policies, achieving high recall without large cache
  - Initial incremental learning techniques for streams (ANN, SVM, BN…), linked to deductive inference (new hybrid reasoning model)



AIM: Human-Machine Feedback



Conclusions

- **Semantic technology provides its advertised benefits**
  - Help users to publish and share data
  - Allow app development to be more efficient
- **AI and semantics have been combined in a new generation of discovery informatics tools that we should bring to analysts**
  - Combine search and synthesis; the human provides analysis
    - Model search like Eureqa
    - Scalable reasoning like LarKC
    - Human embeddings like Netflix
    - Causality like Big Mechanism
    - Language-based integration like Watson
    - Human-machine feedback (maybe with streams) like AIM
  - Rebalance human and machine effort
- **What is a new architecture for analysis?**

Thank You