Accelerating Science
A Grand Challenge for AI?

CCC Task Force on Convergence of Data and Computing
Vasant Honavar, Mark Hill, Kathy Yelick
Presentation to DARPA Defense Science Office
March 16, 2017

CCC
Computing Community Consortium
Catalyst

Computing Community Consortium

The **mission** of Computing Research Association's Computing Community Consortium (CCC) is to **catalyze** the computing research community and **enable** the pursuit of innovative, high-impact research.

Promote Audacious Thinking:

- Community Initiated Visioning Workshops
- Blue Sky Ideas tracks at conferences

Inform Science Policy

- Outputs of visioning activities
- Task Forces – e.g., Artificial Intelligence, Data and Computing, Health, Internet of Things, Privacy

Engage the Community:

- CCC Blog - http://cccblog.org/
- Computing Research in Action Videos
- Research “Highlight of the Week”

Promote Leadership and Service:

- Computing Innovation Fellows Project
- Leadership in Science Policy Institute

Accelerating Science: A Grand Challenge for AI?

- Discussion based in part on:
  - Accelerating Science: A Computing Research Agenda. A Computing Community Consortium White Paper, Vasant Honavar, Mark Hill, and Katherine Yelick, 2016 http://cra.org/crn/2016/03/3501/
  - AAAI Fall Symposium on Accelerating Science: A Grand Challenge for AI
- Other related events:
  - NSF Workshop on Discovery Informatics, February 2012
  - AAAI Fall Symposium on Discovery Informatics, November 2012
  - CMUSV Symposium on Cognitive Systems and Discovery Informatics, 2013
  - AAAI Fall Symposium on Discovery Informatics, November 2013
  - AAAI Workshop on Discovery Informatics, July 2014
  - ACM SIGKDD Workshop on Discovery Informatics, August 2014
  - PSB Workshop on Discovery Informatics, January 2015

CCC
Computing Community Consortium
Catalyst

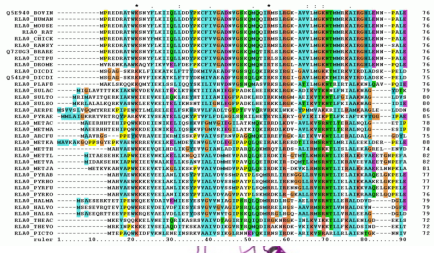
Accelerating Science: A Grand Challenge for AI?

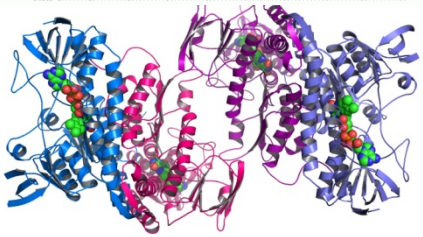
- All science is either stamp collecting or “physics”
- Big data = spectacular stamp collections!
- Big data ≠ Demise of the scientific method
- Accelerating science presents a grand challenge for AI:
  - Analysis and synthesis of computational abstractions of both
    - Universes of scientific discourse
    - Scientific artifacts and scientific process
  - Cognitive tools that augment and extend human intellect
  - Collaborative human-machine infrastructure for science
- Accelerating science calls for
  - Foundational advances within and across virtually all subfields of Artificial Intelligence
  - Concomitant advances in collaborative data and computing infrastructure

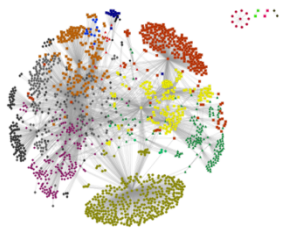
CCC
Computing Community Consortium
Catalyst

Big Data: Challenges and Opportunities

Omics





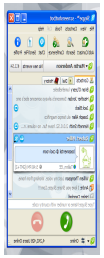


Digital Media

VIDEO

BLOGS

MOBILE



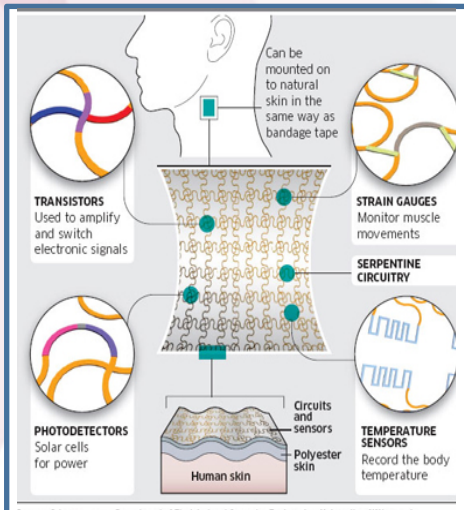
VOIP

EMAIL

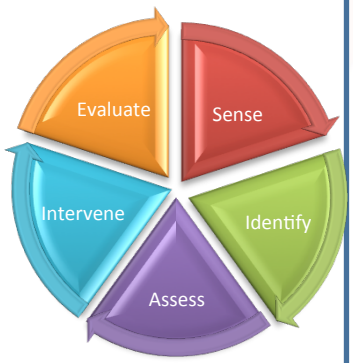
IM

Human Sensors

Source: Keith Marzullo



Health Care



sortium

Big Data

Opportunities offered by big data are real

- Understanding the structure and dynamics of complex systems – cells, brains, individuals, organizations, societies
- Improving population health
- Anticipating and responding to crises
- Personalizing teaching and learning
- Defending critical infrastructure and services
- Making better decisions, e.g., public policy
- Making cities and communities smarter
- Improving food, energy, and water security
- ….

CCC
Computing Community Consortium
Catalyst

Big data = the end of the scientific method?

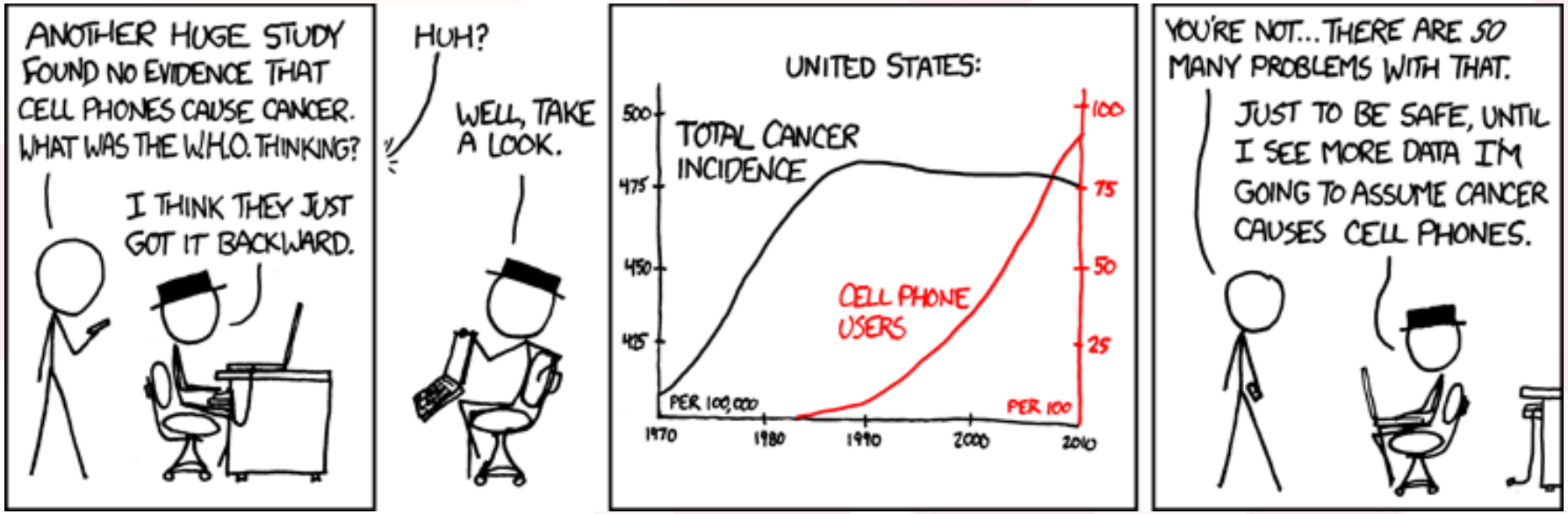
CHRIS ANDERSON, The end of theory: The data deluge makes the scientific method obsolete, *Wired Magazine* 16.07 (June 23, 2008). http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

- "Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot."
- "Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all."
- Most machine learning and data mining algorithms are essentially sophisticated ways of finding correlations from data

CCC
Computing Community Consortium
Catalyst

Perils of fishing for wisdom in oceans of data

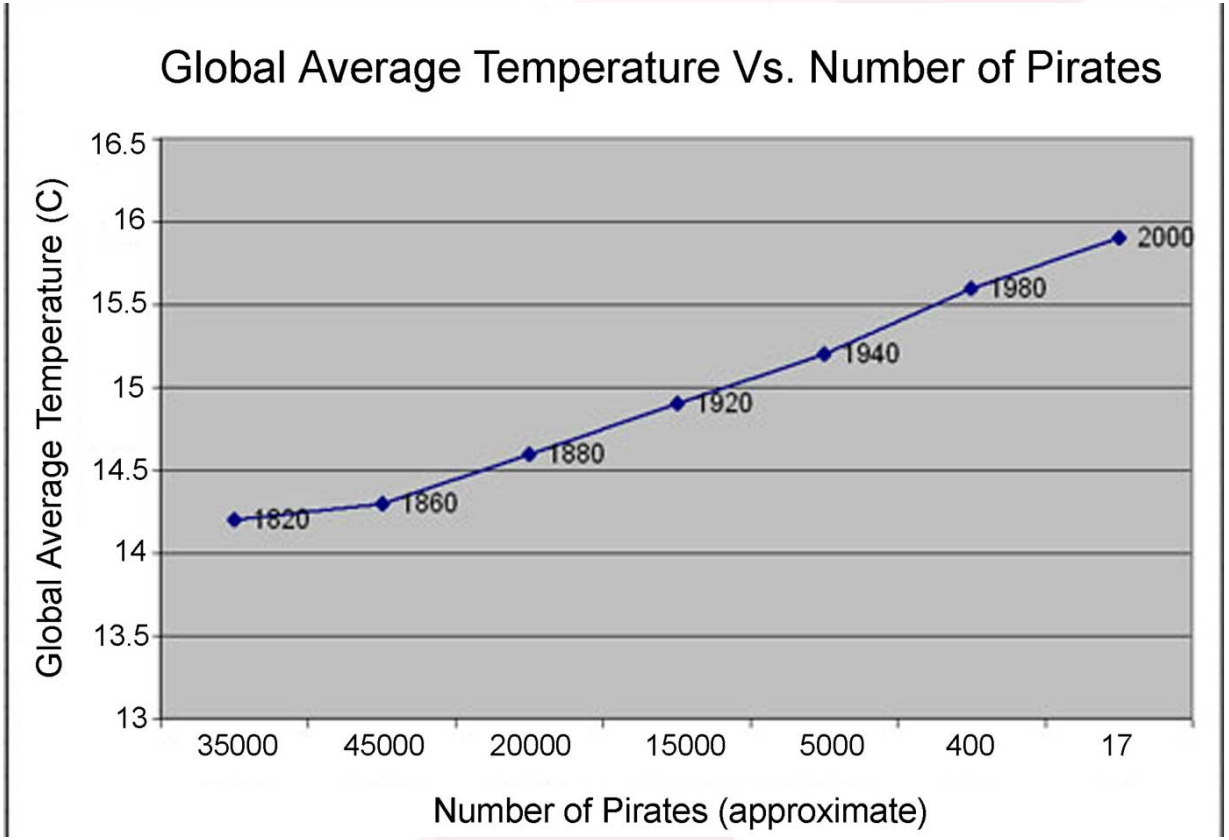
Does cancer cause cell phone use?



CCC
Computing Community Consortium
Catalyst

Perils of fishing for wisdom in oceans of data

Fight global warming! Become a pirate!

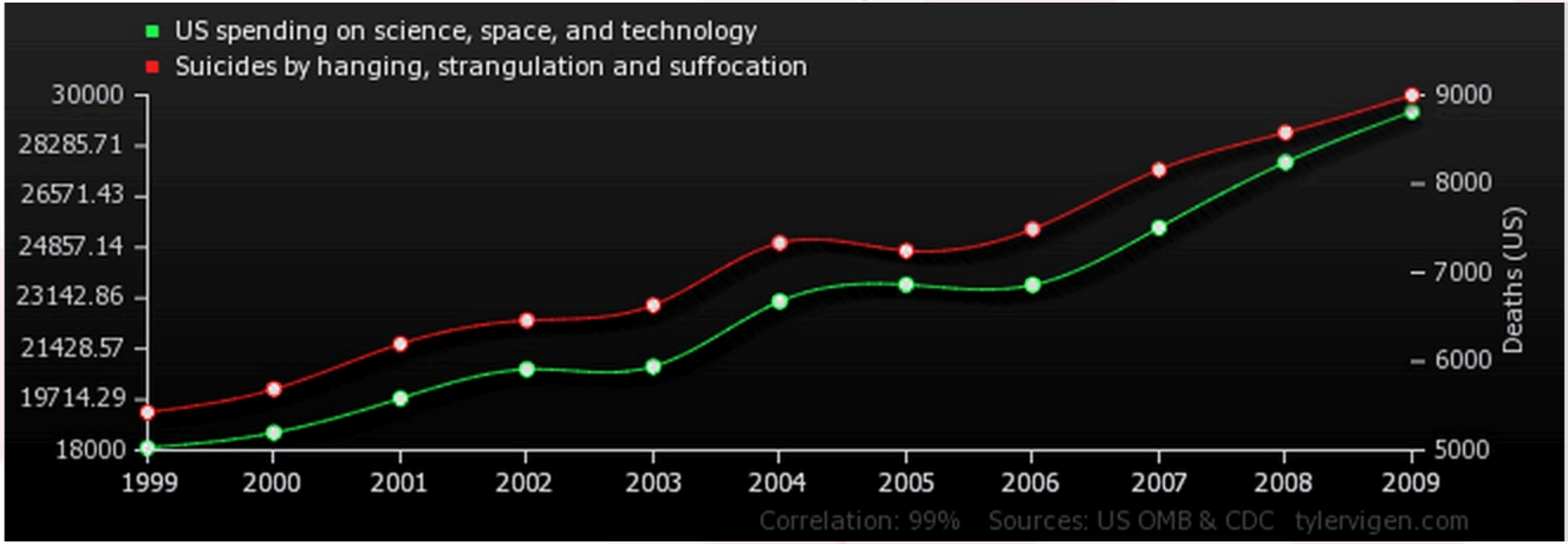


- Big data ≠ End of theory!
- Correlation ≠ Causation!

CCC
Computing Community Consortium
Catalyst

Perils of fishing for wisdom in oceans of data

Eliminate science funding to save lives!



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology *Millions of todays dollars (US OMB)*	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

Correlation: 0.992082

CCC
Computing Community Consortium
Catalyst

Perils of fishing for wisdom in oceans of data

- Sally Clark's 1st son died in 1996, due to SIDS
- Her 2nd son died in 1999, also as a result of SIDS
- Prosecutors charged Sally Clark for murder on the grounds that both deaths were too unlikely to be due to SIDS
- Rationale:
  - One in 8,543 infant deaths is due to SIDS.
  - So chance of 2 deaths = $1/ (8{,}543^2)$ (or 1 in 73 million)
- – What's wrong with this?

CCC
Computing Community Consortium
Catalyst

Data acquisition no longer the rate limiting step in science

- "We are close to having a $1,000 genome sequence, but this may be accompanied by a $1,000,000 interpretation[1]"
- >1300 NAR gene databases
- 1M new biomedical journal articles published per year (2700/day)

[1]Bruce Korf, Former President, American College of Medical Genetics

Slide courtesy Larry Hunter

CCC
Computing Community Consortium
Catalyst

Many aspects of data management and analytics no longer the rate limiting steps in science

Most of the recent advances and efforts are focused on:

- Data Management
  - Organizing
  - Indexing
  - Integrating
  - Storing
  - Querying
- Data Analytics
  - Machine learning
  - Scaling up
  - High dimensionality
  - Heterogeneity

CCC
Computing Community Consortium
Catalyst

Big Data = the end of the scientific method? A lesson from Physics

Transformation of physics from a descriptive science (pre Newton) into a predictive science (post Newton)

- Tycho Brahe gathered 20 years of extremely accurate astronomical measurements: positions of the stars and planets: big data

- Johannes Kepler, working for Tycho Brahe, fit the data in every way imaginable to discover laws of planetary motion: big data analytics

- Isaac Newton's invention of calculus provided the language to express, analyze, and communicate the unified laws of motion: knowledge representation for physics
- Big data did not make obsolete the scientific method then, and it does not do so now!

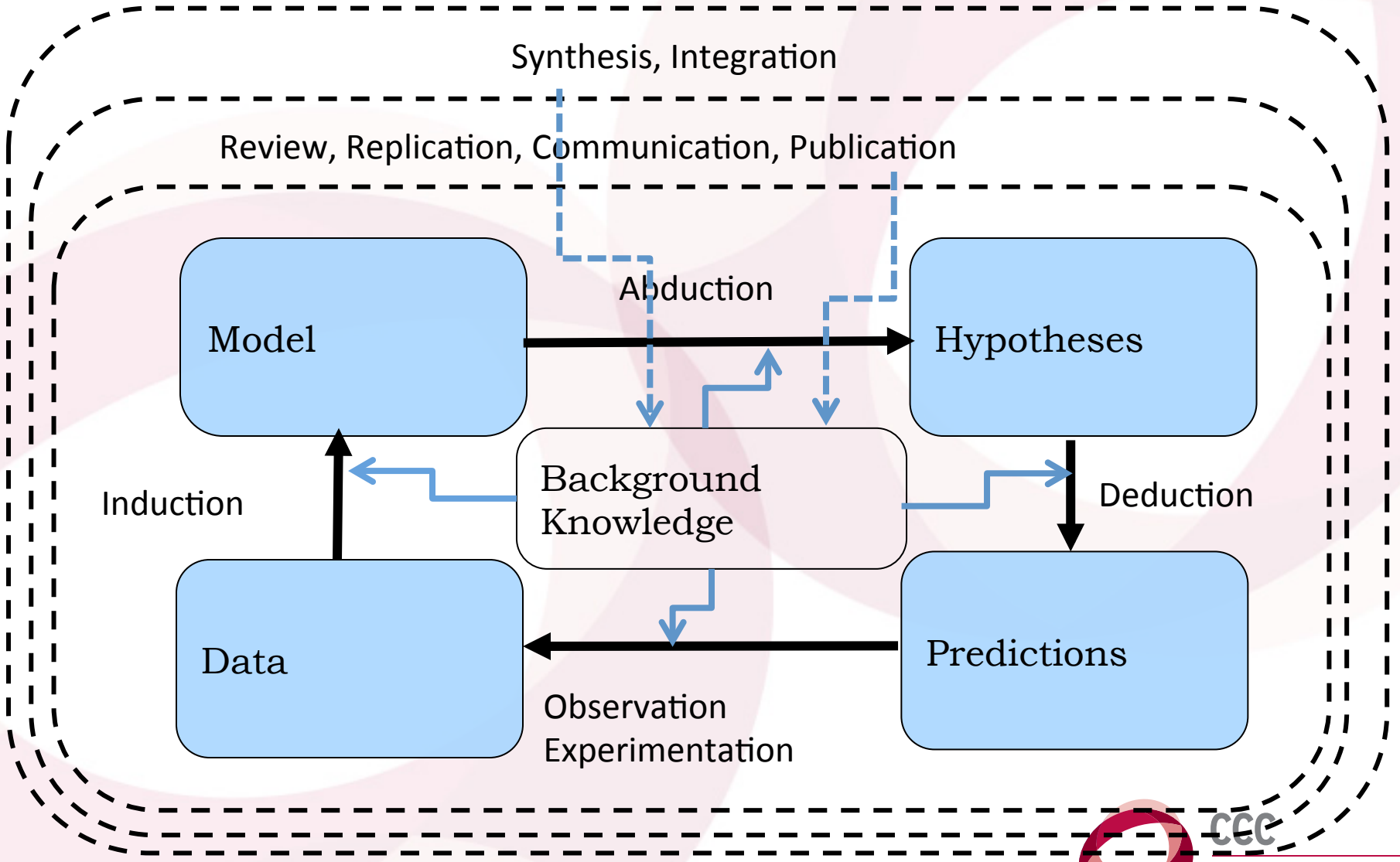
Big Data ≠ The end of the scientific method!

- Automation of “Big data” acquisition, management and analytics accelerates
  - Brahe’s part of the scientific endeavor (data acquisition, management)
  - And thanks to advances in machine learning increasingly, Kepler’s part (data analytics and model building)
  - But for the most part, leaves untouched, the other aspects of science, which become the rate limiting steps in science

➢ Accelerating science in the era of big data requires accelerating the rate-limiting steps of the scientific method!

CCC
Computing Community Consortium
Catalyst

A cartoon of the scientific method



How close are we to fully automating science?

http://www.wired.com/wiredscience/2011/04/robot-scientist-language/

How close are we to fully automating science?

Science 3 April 2009: Vol. 324 no. 5923 pp. 85-89

The Automation of Science

Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, Andrew Sparkes, Kenneth E. Whelan and Amanda Clare

ABSTRACT

The basis of science is the hypothetico-deductive method and the recording of experiments in sufficient detail to enable reproducibility. We report the development of Robot Scientist "Adam," which advances the automation of both. Adam has autonomously generated functional genomics hypotheses about the yeast Saccharomyces cerevisiae and experimentally tested these hypotheses by using laboratory automation. We have confirmed Adam's conclusions through manual experiments. …

How close are we to fully automating science?

- Not very, except perhaps in very carefully constrained settings
- Collaborative human-machine systems might offer a more realistic approach to accelerating science
- Accelerating science presents a grand challenge for AI:
  - Analysis and synthesis of computational abstractions of both
    - Universes of scientific discourse
    - Scientific artifacts and the scientific process
  - Cognitive tools that augment and extend human intellect
  - Collaborative human-machine infrastructure for science

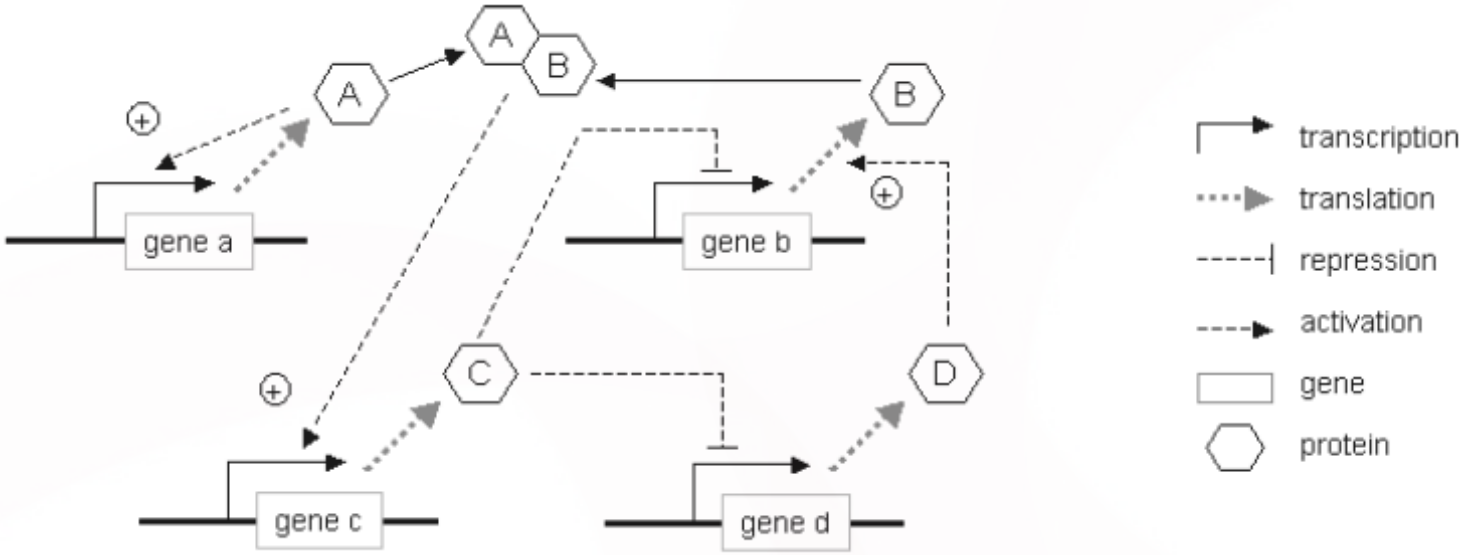
CCC
Computing Community Consortium
Catalyst

Computational abstractions of the universes of scientific discourse

- Church-Turing Thesis: Anything that can be described can be described by a computer program
- In any domain of scientific discourse, we need **computational abstractions** that describe objects, their properties, inter-relationships

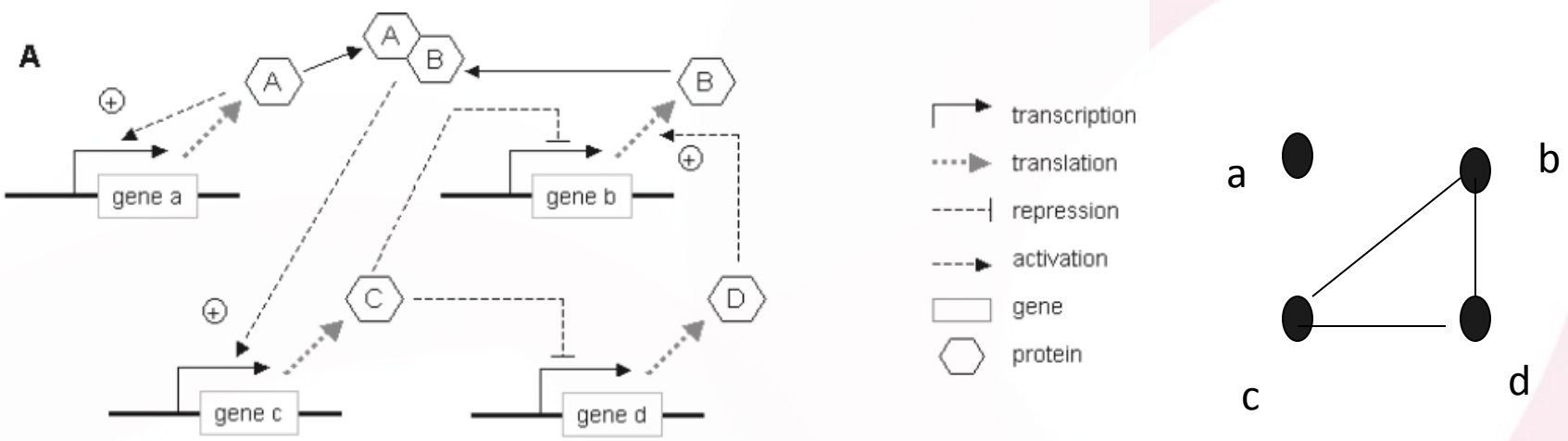
CCC
Computing Community Consortium
Catalyst

Example: Computational abstractions of bio-molecular networks



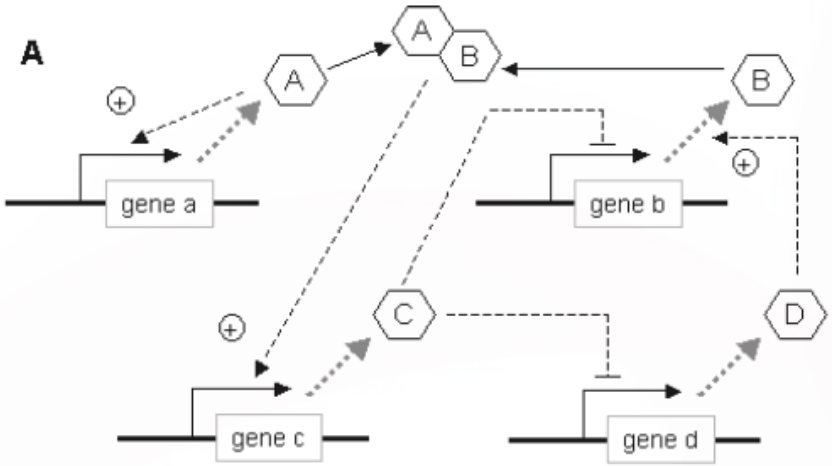
- Genes a, b, c, d code for proteins A, B, C, D
- Proteins A and B form a hetero-dimer that activates the expression of gene c
- Protein C inhibits the expression of (and co-regulates) genes b and d
- Protein D is necessary for the transcription of protein B

Example: Undirected Graphs as abstractions of bio-molecular interaction networks

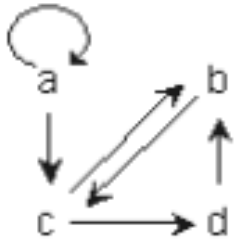


- Protein-protein interaction networks
  - Nodes represent proteins
  - Edges represent interactions e.g., protein *a* binds to protein *b*
  - Topological analysis reveals functional roles
  - Connected components suggest complexes or pathways
  - Comparative analyses (across species, tissues, etc.) reveal shared sub-networks

Directed labeled graphs as abstractions of bio-molecular interaction networks



Directed graphs

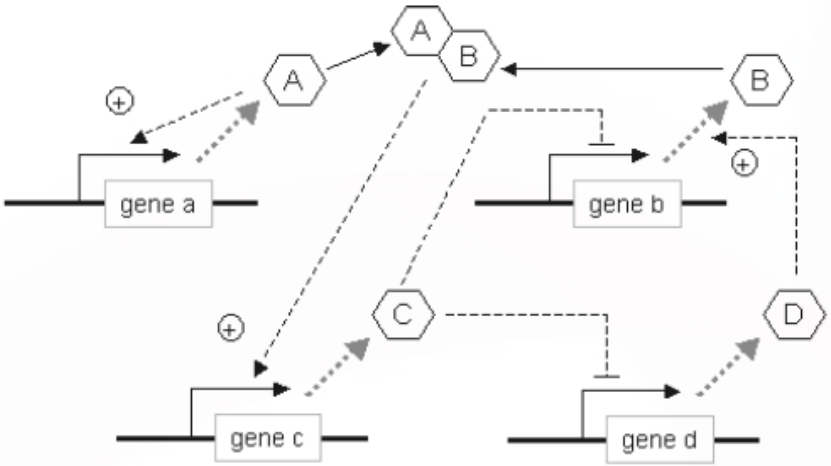


$V = \{a,b,c,d\}$

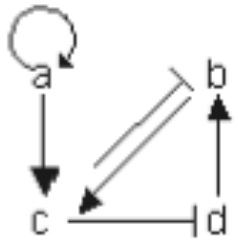
$E = \{(a,a,+),(a,c,+),(b,c,+),(c,b,-),(c,d,-),(d,b,+)\}$

- Nodes correspond to genes
- Edges correspond to regulatory interactions
- Edge labels can be used to denote the types of interactions, or lists of regulators and their influence on the specific edge, e.g., KEGG pathways
- Can help uncover sequences of regulatory events, cycles (feedback regulation), redundancy…

Boolean networks as abstractions of bio-molecular interaction networks



Boolean network



$a(t+1) = a(t)$

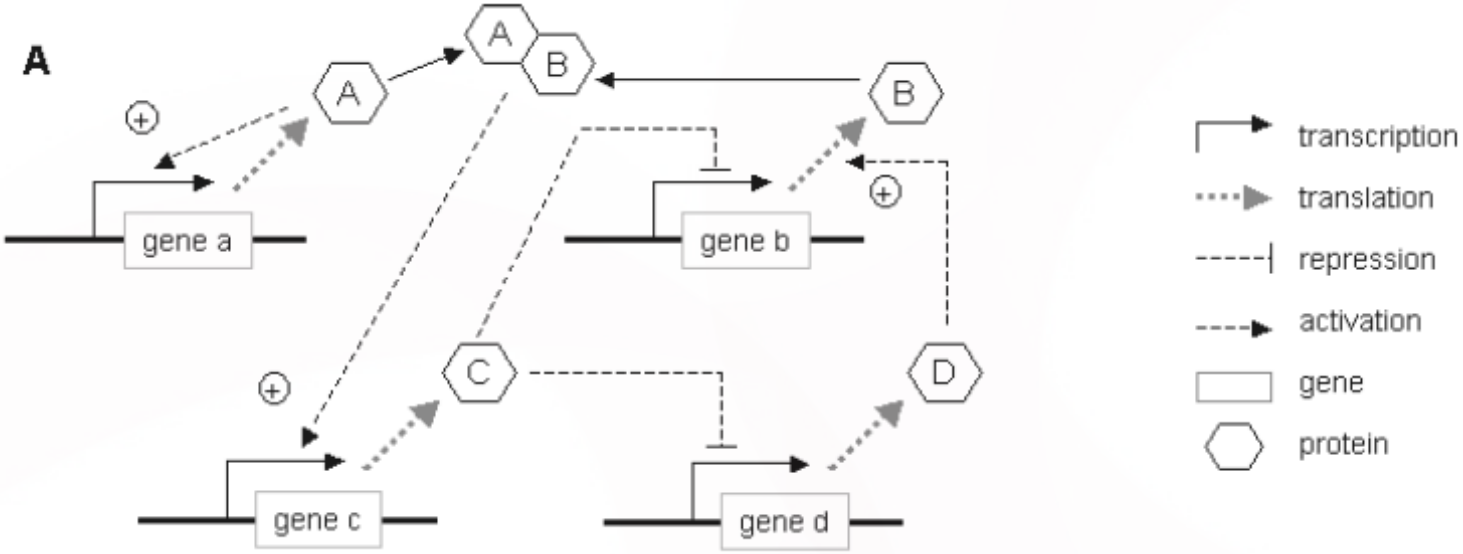
$b(t+1) = (\text{not } c(t)) \text{ and } d(t)$

$c(t+1) = a(t) \text{ and } b(t)$

$d(t+1) = \text{not } c(t)$

- Genes are modeled by binary variables - on, off (1, 0)
- States of genes are updated in discrete time steps
- State of a gene at time *t* +1 is a Boolean function of the states at time *t* of the genes that influence it
- An *N* gene Boolean network can in principle be in one of 2^N states
- Can help determine if the network can get from one state to another, the effect of gene knockout, etc.

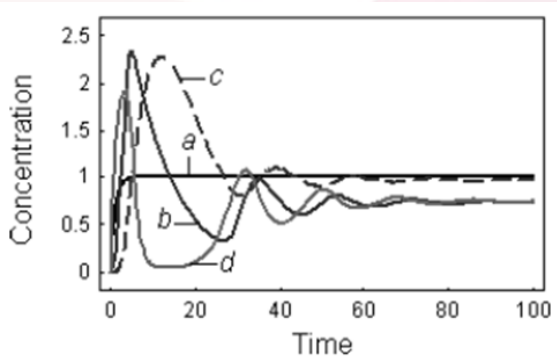
Differential equations as abstractions of bio-molecular interaction networks



Considering only the $mRNA$ abundances a, b, c, d

$$\frac{da}{dt} = f_a(a) \qquad \frac{db}{dt} = f_b(b,c,d)$$

$$\frac{dc}{dt} = f_c(a,b,c) \qquad \frac{dd}{dt} = f_d(c,d)$$



Differential equations can provide detailed information about kinetics

Accelerating science requires computational abstractions of the universes of scientific discourse

- Church-Turing Thesis: Anything that can be described can be described by a computer program
- In any domain of scientific discourse, we need computational abstractions that describe objects, their properties, inter-relationships, in domains of scientific discourse

CCC
Computing Community Consortium
Catalyst

How close are we to fully automating science?

- Not very, except perhaps in very carefully constrained settings
- Collaborative human-machine systems might offer a more realistic approach to accelerating science
- Accelerating science presents a grand challenge for AI:
  - Analysis and synthesis of computational abstractions of both
    - Universes of scientific discourse
    - Scientific artifacts and the scientific processes
  - Cognitive tools that augment and extend human intellect
  - Collaborative human-machine infrastructure for science

CCC
Computing Community Consortium
Catalyst

Computational abstractions of scientific artifacts and scientific processes

Examples of scientific artifacts

- Experimental protocols
- Data, metadata, provenance
- Assumptions
- Conjectures
- Hypotheses
- Analysis tools
- Findings
- Arguments
- Models
- Explanations
- Theories
- Workflows

Examples of scientific processes

- Designing, prioritizing, planning, executing, documenting, replicating experiments
- Acquiring and organizing data
- Building, evaluating, linking models
- Generating and ranking conjectures
- Generating and testing hypotheses
- Testing, refining, comparing theories
- Producing and ranking explanations
- Sharing data and other artifacts

CCC
Computing Community Consortium
Catalyst

Computational abstractions of model construction

- Models can be built from knowledge (using inference, e.g., abstraction, specialization), observations (machine learning), experiments (e.g., causal inference)
- The use of off the shelf machine learning methods (SVM, DNN, Bayesian Networks, Stochastic grammars, etc.) introduces a language gap between model builders and model users
- Need principled and generalizable approaches to
  - Construct and refine models that are
    - Accurate yet comprehensible
    - Explanatory
    - Communicable
    - Consistent with accepted background knowledge (e.g., laws of physics)
    - Lead to testable hypotheses
    - Models that span multiple levels of abstraction and scale
  - Assessing models with respect to not only predictive accuracy but also
    - Explanatory power
    - Coherence with models at higher and lower levels of abstraction
    - Simplicity …

How close are we to fully automating science?

- Not very, except perhaps in very carefully constrained settings
- Collaborative human-machine systems might offer a more realistic approach to accelerating science
- Accelerating science presents a grand challenge for AI:
  - Analysis and synthesis of computational abstractions of both
    - Universes of scientific discourse
    - Scientific artifacts and the scientific process
  - Cognitive tools that augment and extend human intellect
  - Collaborative human-machine infrastructure for science

CCC
Computing Community Consortium
Catalyst

Cognitive tools for scientists that augment and extend the human intellect

- Acquiring, organizing and maintaining background “knowledge”
- Assessing and finding gaps in scientific knowledge
- Formulating and prioritizing questions
- Formulating, prioritizing, planning, and documenting studies
- Designing, prioritizing, planning, executing, monitoring experiments
- Drawing inferences, constructing explanations and hypotheses
- Synthesizing findings from disparate observational and experimental studies
- Building accurate, communicable, testable models from knowledge, observations and experiments
- Linking data, models, scientific arguments, hypotheses, experiments
- Linking and reasoning with models at different levels of abstraction or across different facets
- Sharing data, models, hypotheses, and other scientific artifacts
- Integrating results into the larger body of knowledge

CCC
Computing Community Consortium
Catalyst

A representative cognitive tool for scientists

A scientist's associate that

- Learns what you and others in your field and related fields are are working on
- Finds and reads relevant literature
- Locates and ingests available knowledge and data
- Offers assistance
  - Here are some data that contradict your hypothesis
  - Here are arguments for and against your hypothesis
  - Here is some data from lab X that explains your finding
  - Here is why you should prefer model A to model B

CCC
Computing Community Consortium
Catalyst

Collaborative human-machine infrastructure for science

- Distributed collaboratories that support:
  - Sharable and communicable representations of scientific artifacts
  - Data and computational resources
- Organizational structures and processes for collaboration
  - Assembling teams
  - Prioritizing, assigning and scheduling tasks
  - Decomposing tasks, combining results
  - Incentivizing and engaging participants
  - Organizing citizen science

CCC
Computing Community Consortium
Catalyst

Accelerating science: A grand challenge for AI?

- Accelerating science calls for synergistic advances across multiple areas of AI (and computing)
  - Knowledge representation and inference
    - How to represent and reason about computational abstractions of scientific domains, scientific artifacts, and scientific processes?
  - Planning and robotics
    - How to design, plan, execute, monitor, experiments?
  - Machine learning and causal inference
    - How to build accurate, comprehensible, predictive, explanatory or causal models from knowledge, observations, and experiments?
  - Computer supported collaborative work
    - How to optimally organize and incentivize scientific collaborative teams?
  - Data and computational infrastructure for science
    - How to share and reuse scientific artifacts at scale?

CCC
Computing Community Consortium
Catalyst

Discussion

CCC
Computing Community Consortium
Catalyst