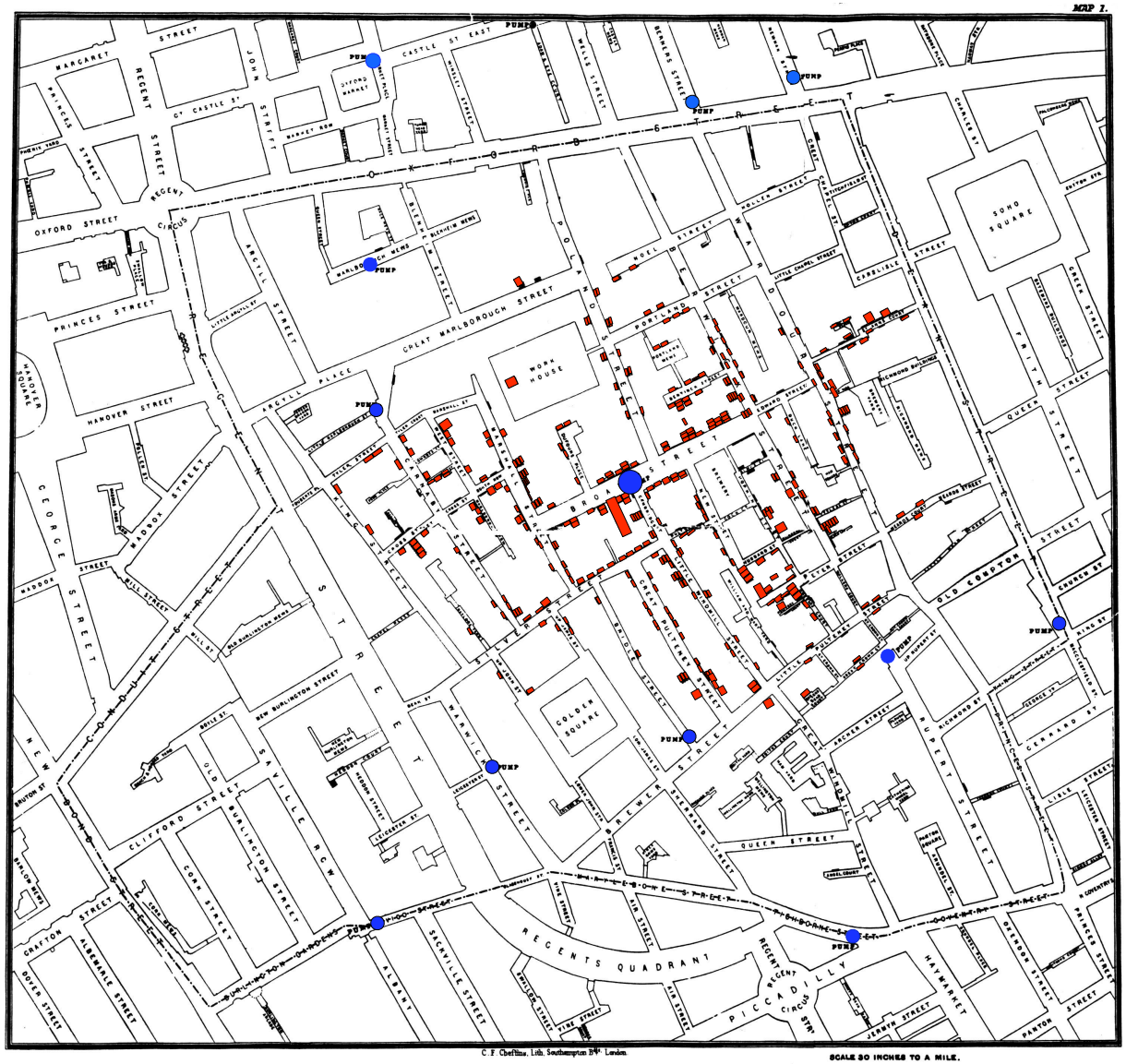
Big Mechanism

Paul Cohen

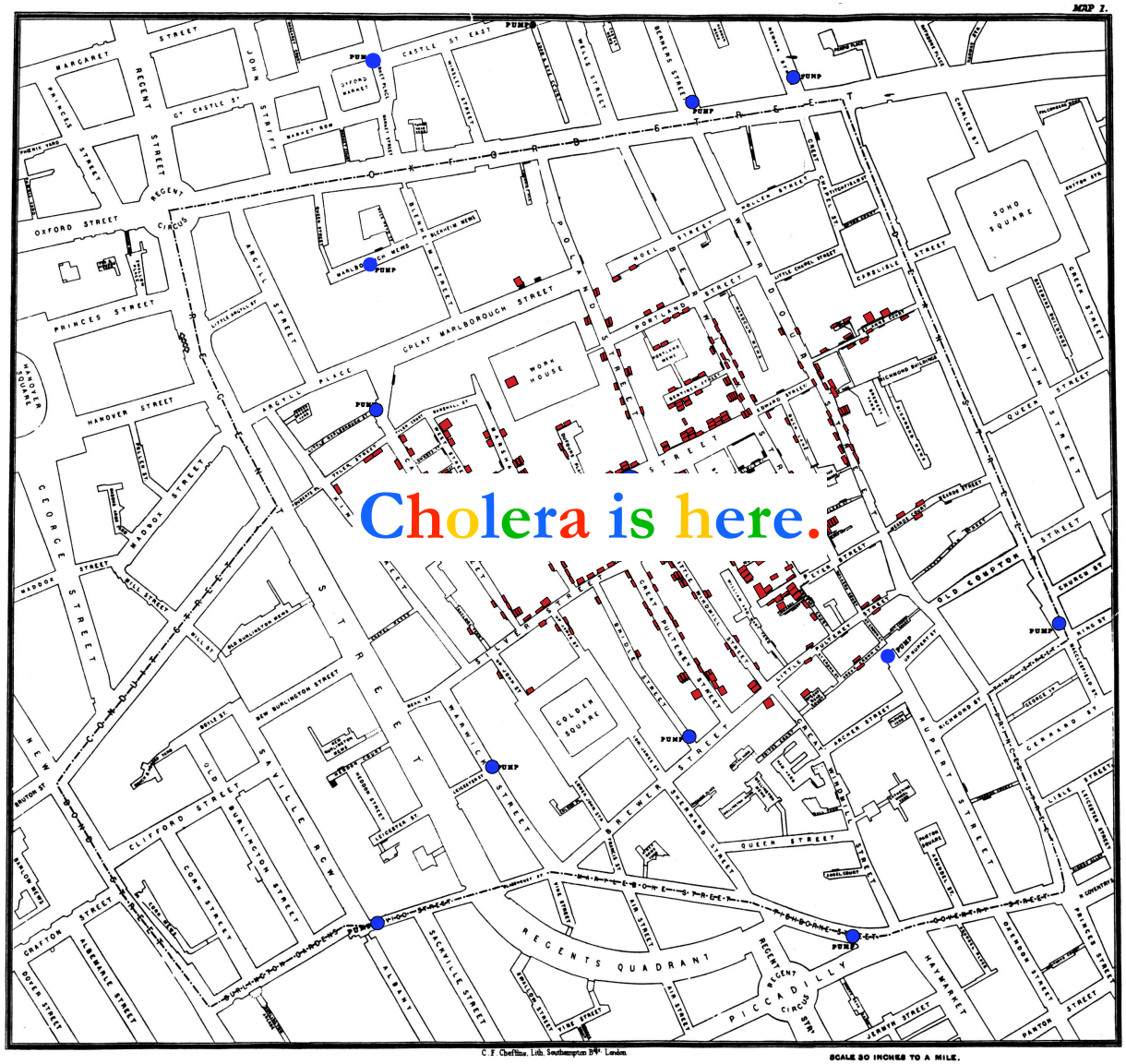
2014/01/31

Big Data, 1854



Approved for Public Release, Distribution Unlimited

Big Data, 2013

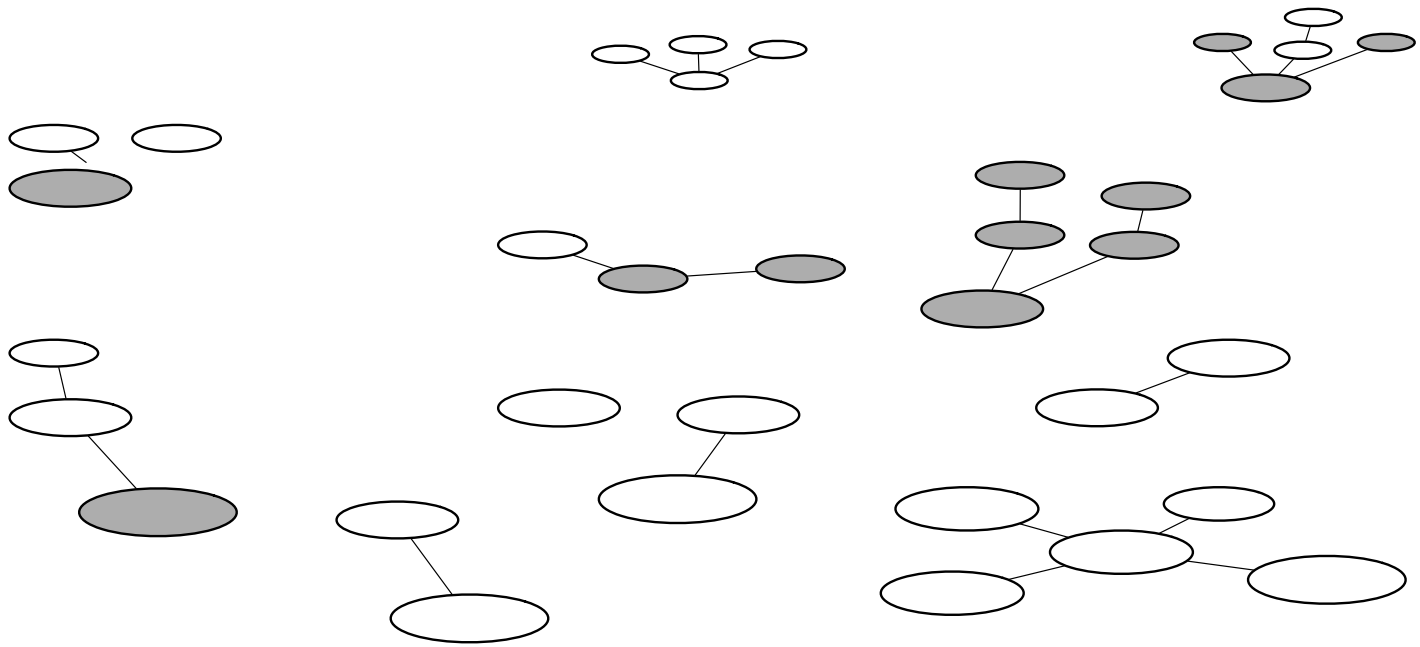


Approved for Public Release, Distribution Unlimited

The Problem

We need to understand the mechanisms of big, complicated systems.

But our knowledge of these mechanisms is increasingly fragmented, voluminous and inconsistent.



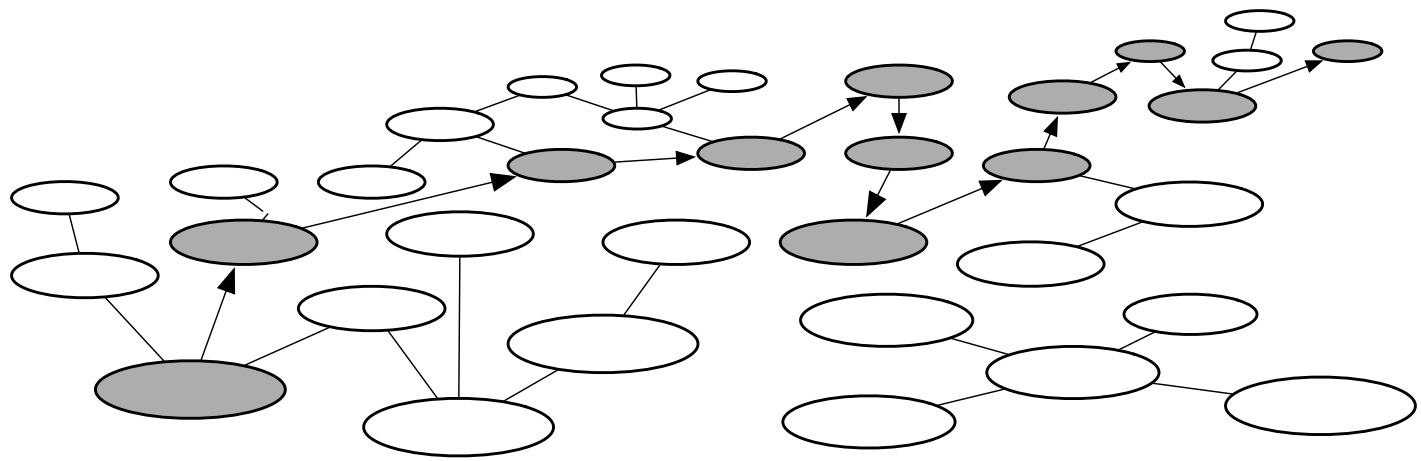
Approved for Public Release, Distribution Unlimited

The Solution: Read, Assemble, Explain

We need to understand the mechanisms of big, complicated systems.

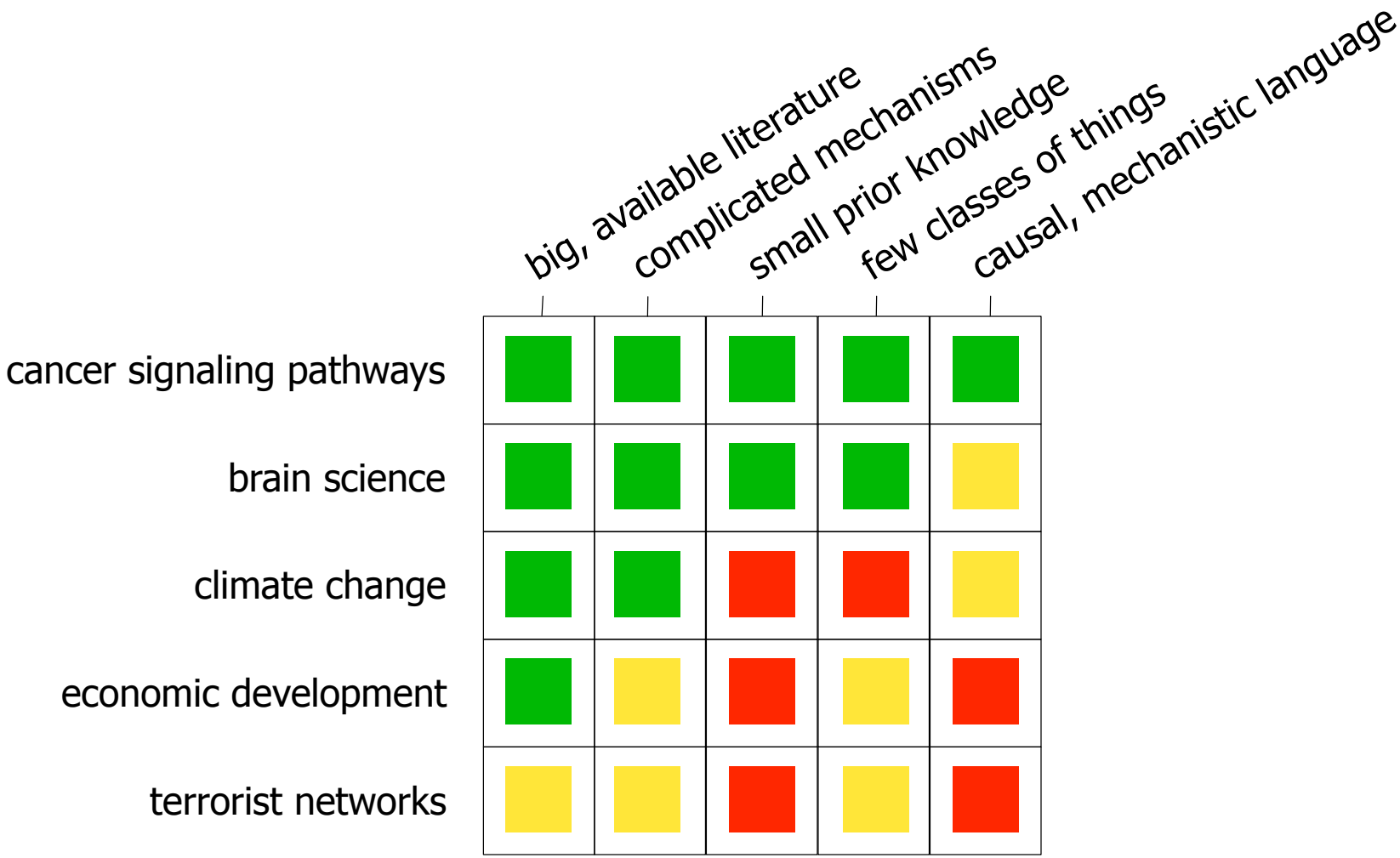
But our knowledge of these mechanisms is increasingly fragmented, voluminous and inconsistent.

The Solution: Make computers **read** documents and data, and **assemble** the fragments they contain into Big Mechanisms that **explain** causes and effects within systems.



Approved for Public Release, Distribution Unlimited

Read, Assemble, Explain for Which Systems?



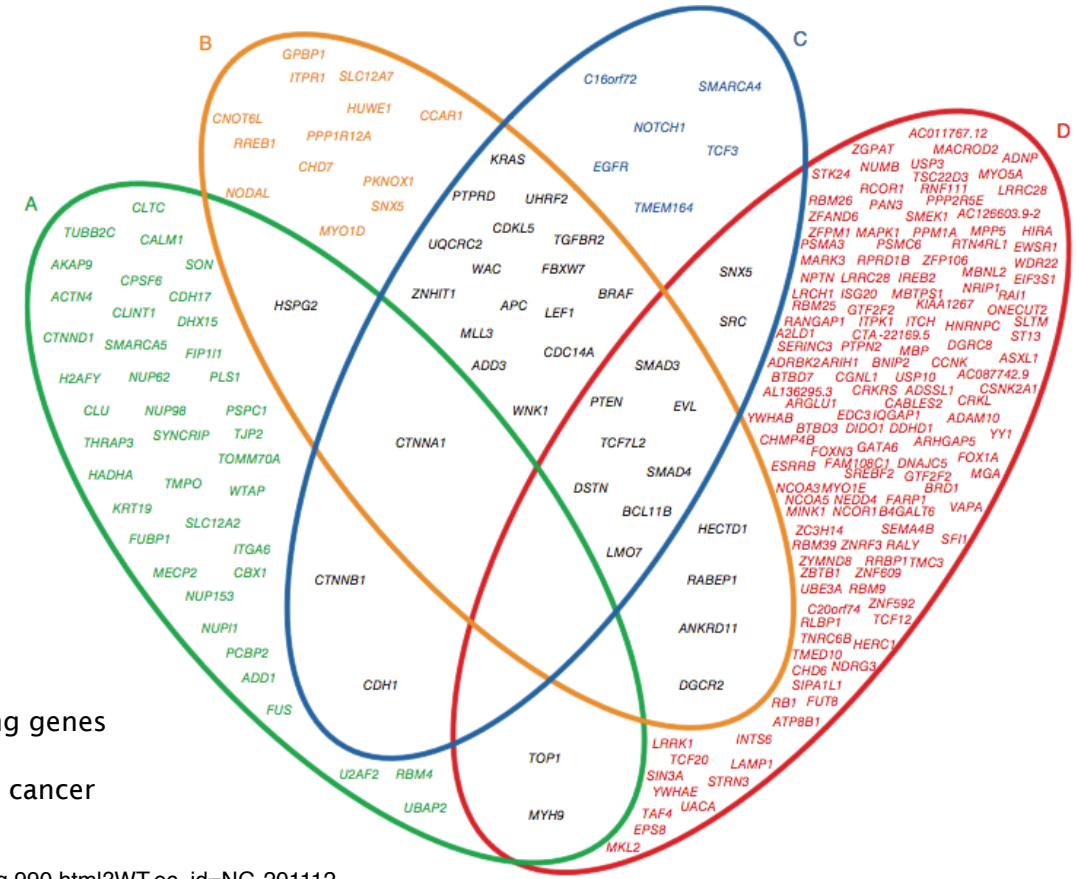
Approved for Public Release, Distribution Unlimited

Complicated Systems

Cancer biology is a complicated system.

Different authors estimate that on average 7 – 15 somatic mutations are required for a normal human cell to undergo malignant transformation.

The proteins and genes in the figure are *all* implicated in colorectal cancer.



1. Force mutagenesis
2. Find common insertion sites (CIS)

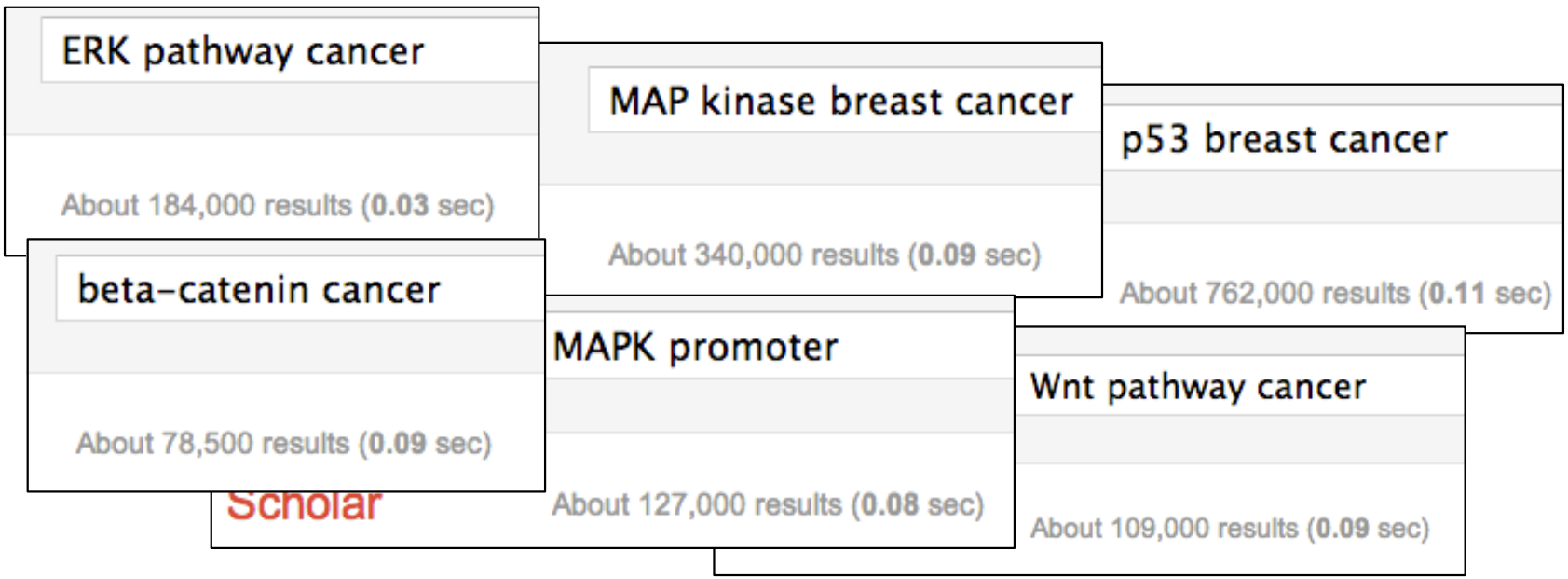
A. Proteins coded by CISs
B. Human orthologs of CIS-containing genes
C and D. Previously reported human orthologs associated with colorectal cancer

Source: Nature Genetics 2011
http://www.nature.com/ng/journal/v43/n12/full/ng.990.html?WT.ec_id=NG-201112

Approved for Public Release, Distribution Unlimited

Knowledge is Fragmented and Voluminous

Cancer biology knowledge is fragmented and voluminous:



PubMed contains 23,000,000 abstracts and grows at roughly 500,000 abstracts per year (source: PubMed)

Approved for Public Release, Distribution Unlimited

Knowledge is Inconsistent

Cancer biology knowledge is inconsistent:

"We observed remarkably poor agreement (consistently less than 10%) among different databases regarding pathway components."

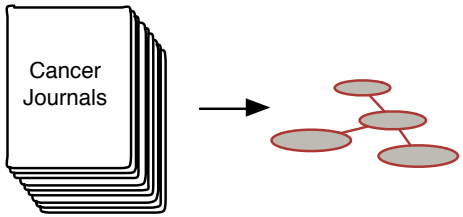
"What constitutes a 'canonical' pathway is database specific. This inconsistency ... may reflect underlying biology, in that signal transduction events are often context-dependent, or it may reflect the absence of a controlled vocabulary."

"This raises a significant problem for mechanistic modeling, since ... it is not clear which genes/proteins to include for modeling or experimental measurement."

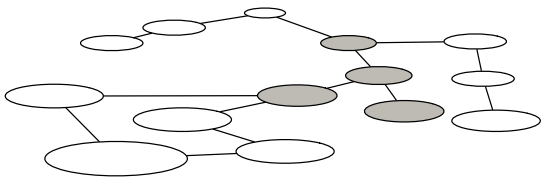
– Kirouac et al. BMC Systems Biology 2012, 6:29
http://www.biomedcentral.com/1752-0509/6/29

Approved for Public Release, Distribution Unlimited

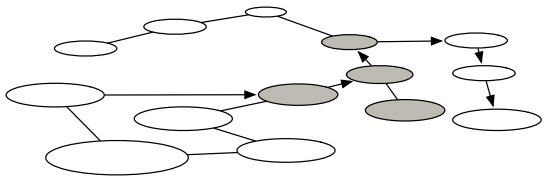
Technology Development Tasks



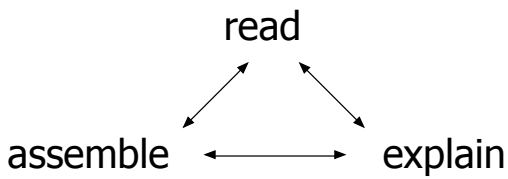
Read papers in cancer biology and extract causal fragments of signaling pathways, represented at all relevant semantic levels.



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.

Approved for Public Release, Distribution Unlimited

The Reading Task: Technology Challenges

Current reading technology extracts semantically shallow assertions; this program intends to "go deep" to extract causality, kinetics, abstraction.

Involvement	"β-catenin is a critical component of Wnt-mediated transcriptional activation"
Causal/Promotion	"ARF6 activation promotes the intracellular accumulation of β-catenin."
Kinetic	"L-cells treated with the GSK3β inhibitor LiCl (50 mM) or the proteasome/calpain inhibitor MG132 (25 μM) showed a marked increase in β-catenin fluorescence within 30 – 60 min"
Modular	"...via a mechanism that involves the endocytosis of growth factor receptors and robust activation of extracellular signal-regulated kinase."

Shallower

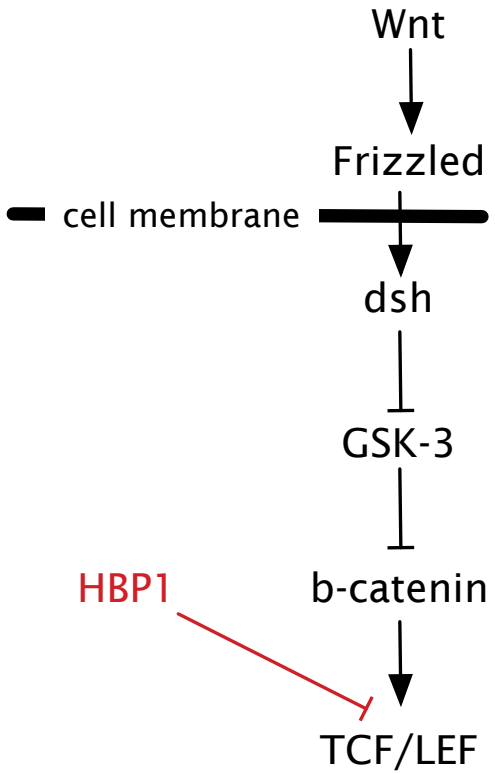
Deeper

Approved for Public Release, Distribution Unlimited

The Reading Task: The Language

HBP1 is a repressor of the cyclin D1 gene and inhibits the Wnt signaling pathway. The inhibition of Wnt signaling and growth requires a common domain of HBP1. The apparent mechanism is an inhibition of TCF/LEF DNA binding through physical interaction with HBP1.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC125566/#



Everything except HBP1 inhibition was known before this 2001 article was published.

Approved for Public Release, Distribution Unlimited

The Reading Task: Previous Results

Extracting Regulatory Gene Expression Networks from PubMed

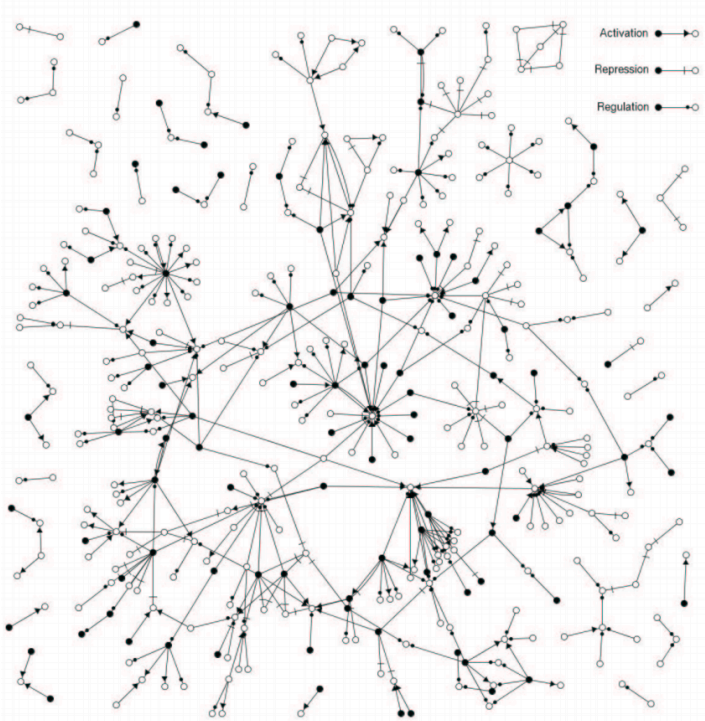
Saric, Jensen, Ouzounova, Rojas & Bork EMBL, Heidelberg, Germany

Extracted a regulatory network of 441 pairwise relations from 58,664 PubMed abstracts about Brewers' Yeast.

Semantic accuracy of 83 – 90% for different roles (e.g., nx_prom, contain)

Extremely small problem!

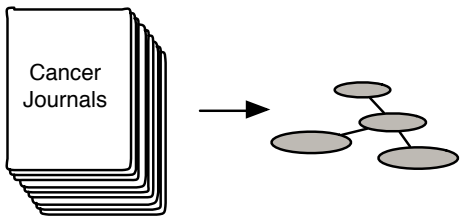
Semantically shallow (no kinetics, modules, etc.); "... we decided against trying to extract how the regulation take place."



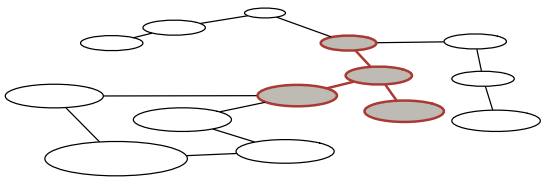
[nx_prom the ATR1 promoter region]
[$contain$ contains]
[nx_uas_pt
  [$dt-a$ a] [bs binding site] [for for]
  [$nx\_activator$ the GCN4 activator protein]]

Approved for Public Release, Distribution Unlimited

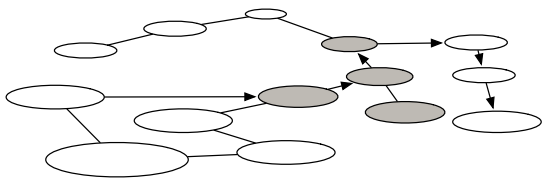
Technology Development Tasks



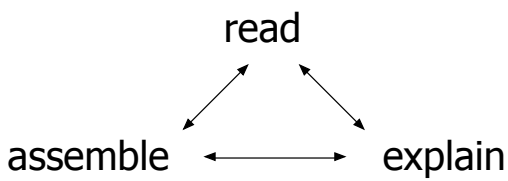
Read papers in cancer biology and extract causal fragments of signaling pathways, represented at all relevant semantic levels.



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.

Approved for Public Release, Distribution Unlimited

The Assembly Task – Technology Challenges

Which entities "match up"? Are fragments semantically consistent? In which formal language is Big Mechanism to be represented?

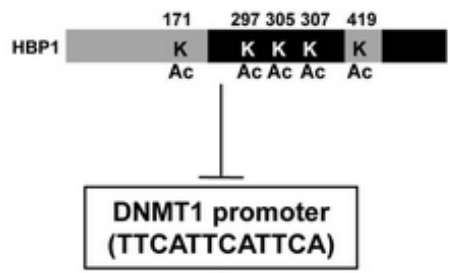
Fragment 1

HBP1 is a **repressor** of the **cyclin D1 gene** and inhibits the Wnt signaling pathway.

HBP1 ⊣ Cyclin D1

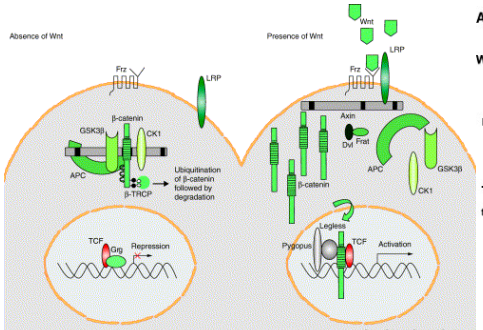
Fragment 2

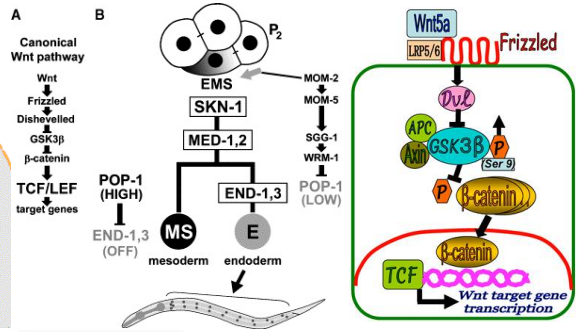
HBP1 represses the **DNMT1 promoter** through sequence-specific binding (of the type TTCATTCATTCA) and the activity of **HBP1** itself is regulated through acetylation at any of 5 sites in **the protein.**

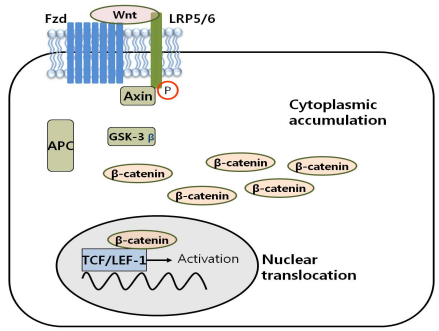


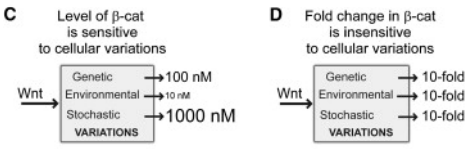
Approved for Public Release, Distribution Unlimited

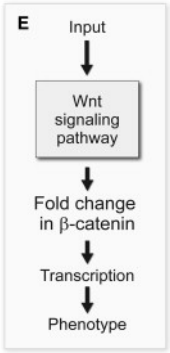
The Assembly Task – Ad Hoc Representation

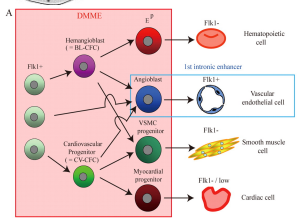


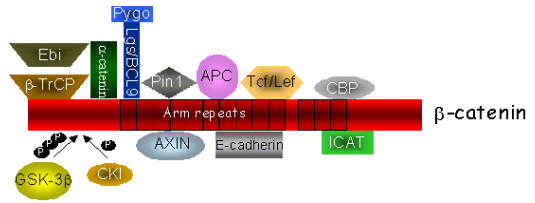


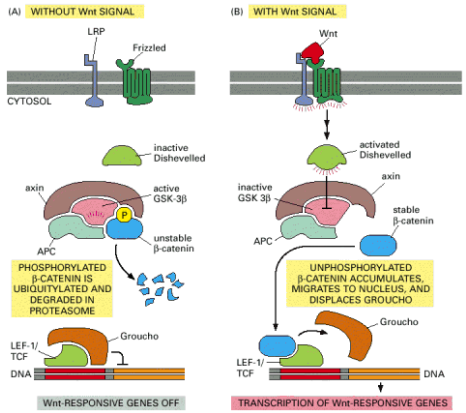


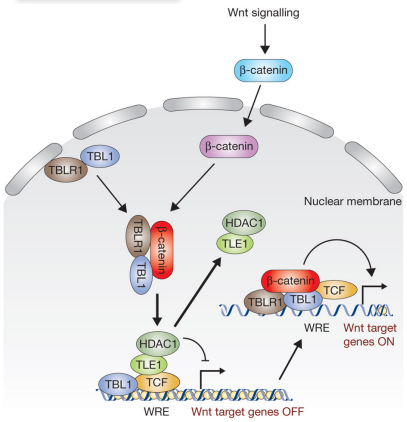


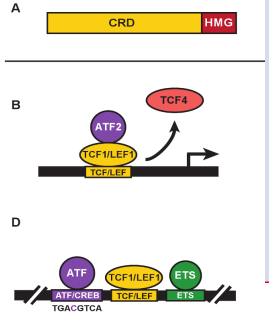


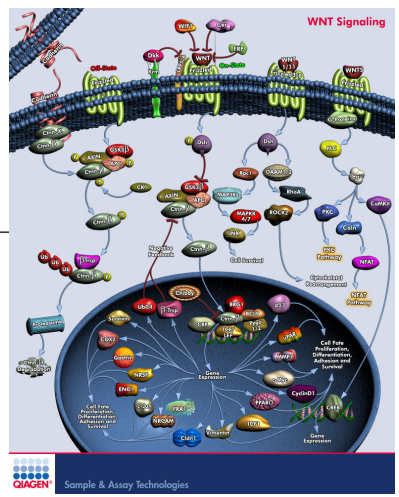










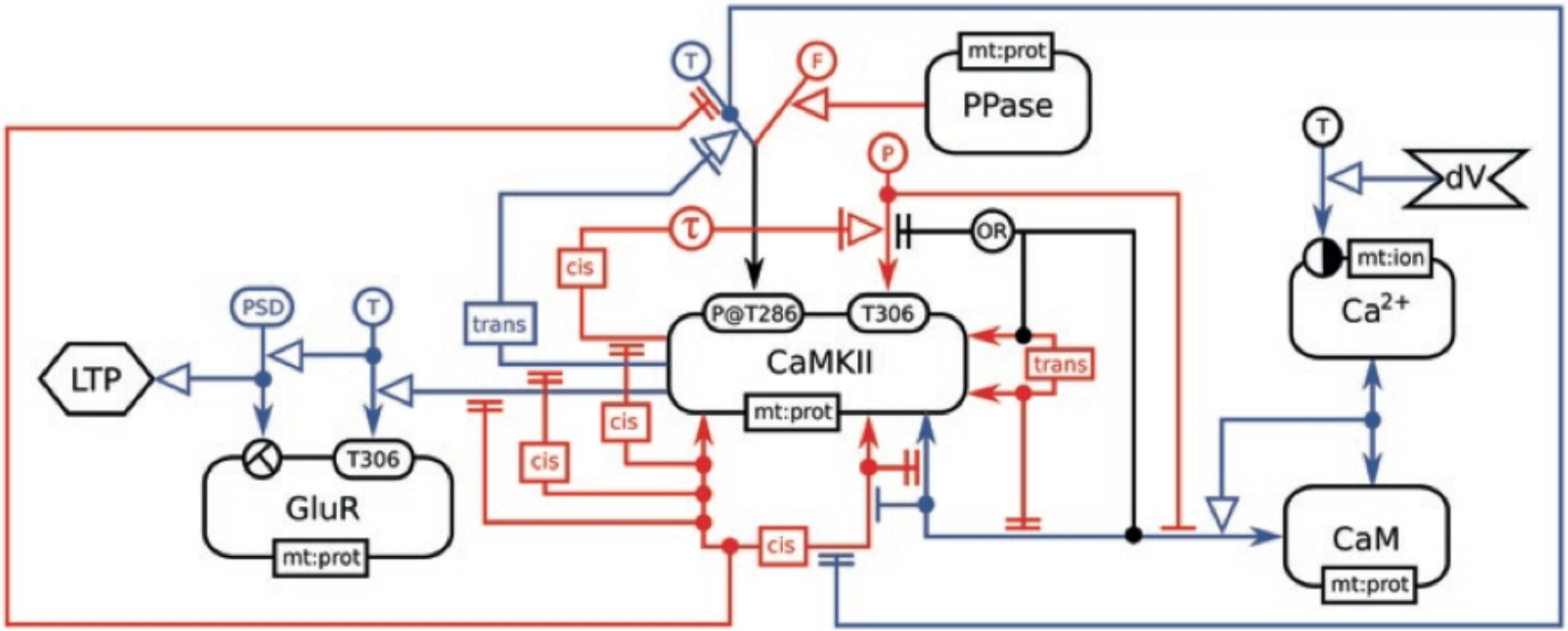


Source:
Paul Cohen

Approved for Public Release, Distribution Unlimited

The Assembly Task – Previous Results

Toward a standard cell biology modeling language:

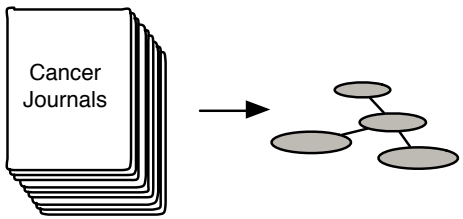


Manual encoding of process of long-term potentiation at synapses

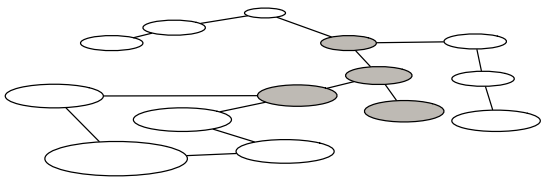
Source: The Systems Biology Graphical Notation. Nature Biotechnology volume 27 number 8 2009
http://www.nature.com/nbt/journal/v27/n8/full/nbt.1558.html

Approved for Public Release, Distribution Unlimited

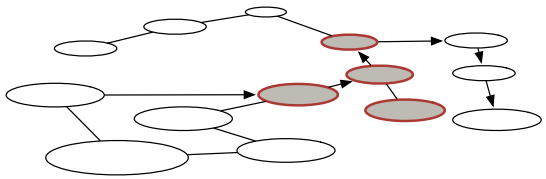
Technology Development Tasks



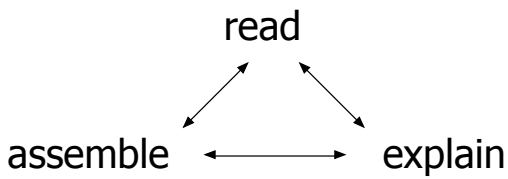
Read papers in cancer biology and extract causal fragments of signaling pathways, represented at all relevant semantic levels.



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.

Approved for Public Release, Distribution Unlimited

The Explanation Task – Technology Challenges

Diverse reasoning methods – probabilistic, deductive, abductive, kinetic simulation, qualitative simulation – and data mining methods, all contribute to explanations; which are best, when, and what's missing?

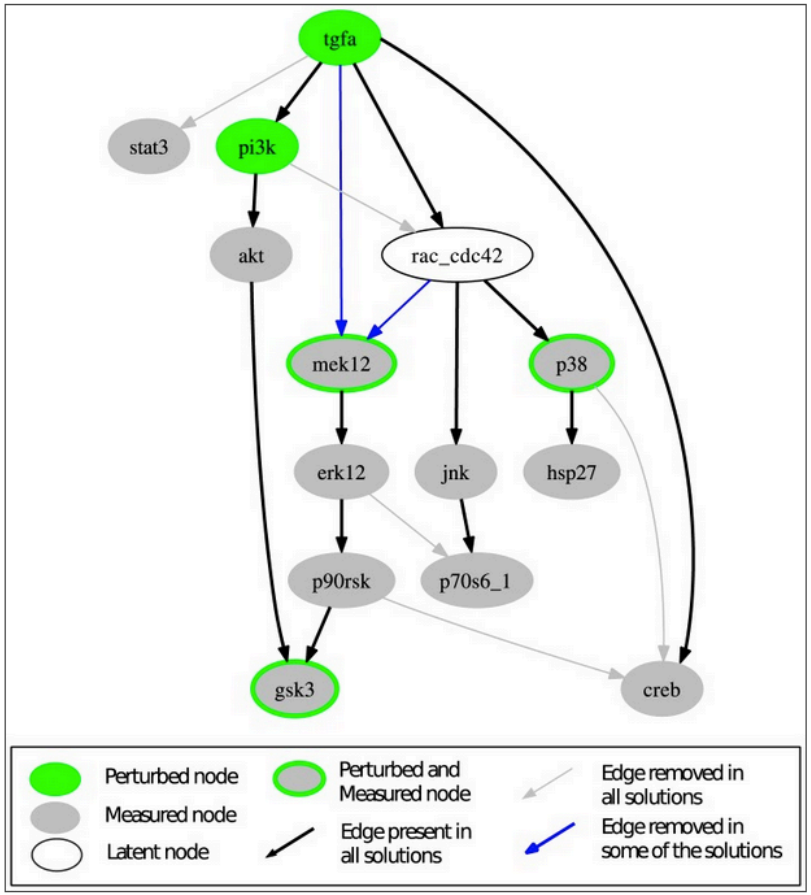
Examples:

- Can a contradiction in a causal model be resolved by a kinetic model?
- What are likely consequences of enabling or inhibiting a protein?

- Use data to evaluate pathway models
- Use data to create or modify pathway models

Approved for Public Release, Distribution Unlimited

The Explanation Task – Previous Results



Source:
http://www.ncbi.nlm.nih.gov/pubmed/24039561

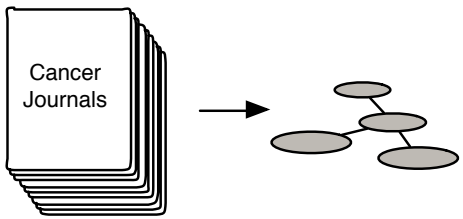
Detecting and Removing Inconsistencies between Experimental Data and Signaling Network Topologies Using Integer Linear Programming on Interaction Graphs.

Melas, Samaga, Alexopoulos, and Klamt.

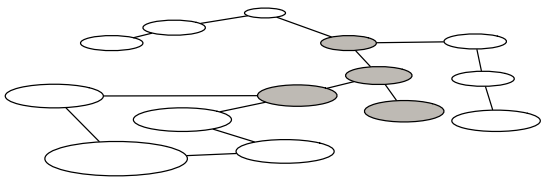
[We] predict the possible qualitative changes (up, down, no effect) of the activation levels of the nodes for a given stimulus. We detect and remove inconsistencies between measurements and predicted behavior ... [We] detect interactions ... and provide suggestions for new interactions that, if included, would significantly improve the goodness of fit.

Approved for Public Release, Distribution Unlimited

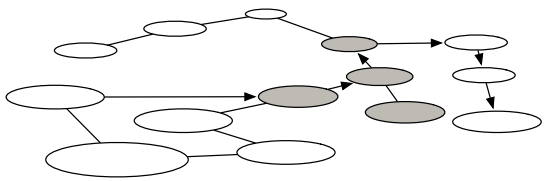
Technology Development Tasks



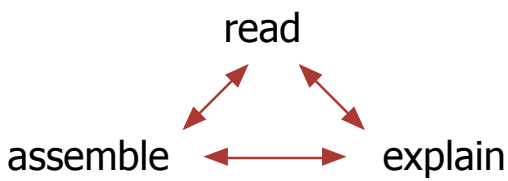
Read papers in cancer biology and extract causal fragments of signaling pathways, represented at all relevant semantic levels.



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.

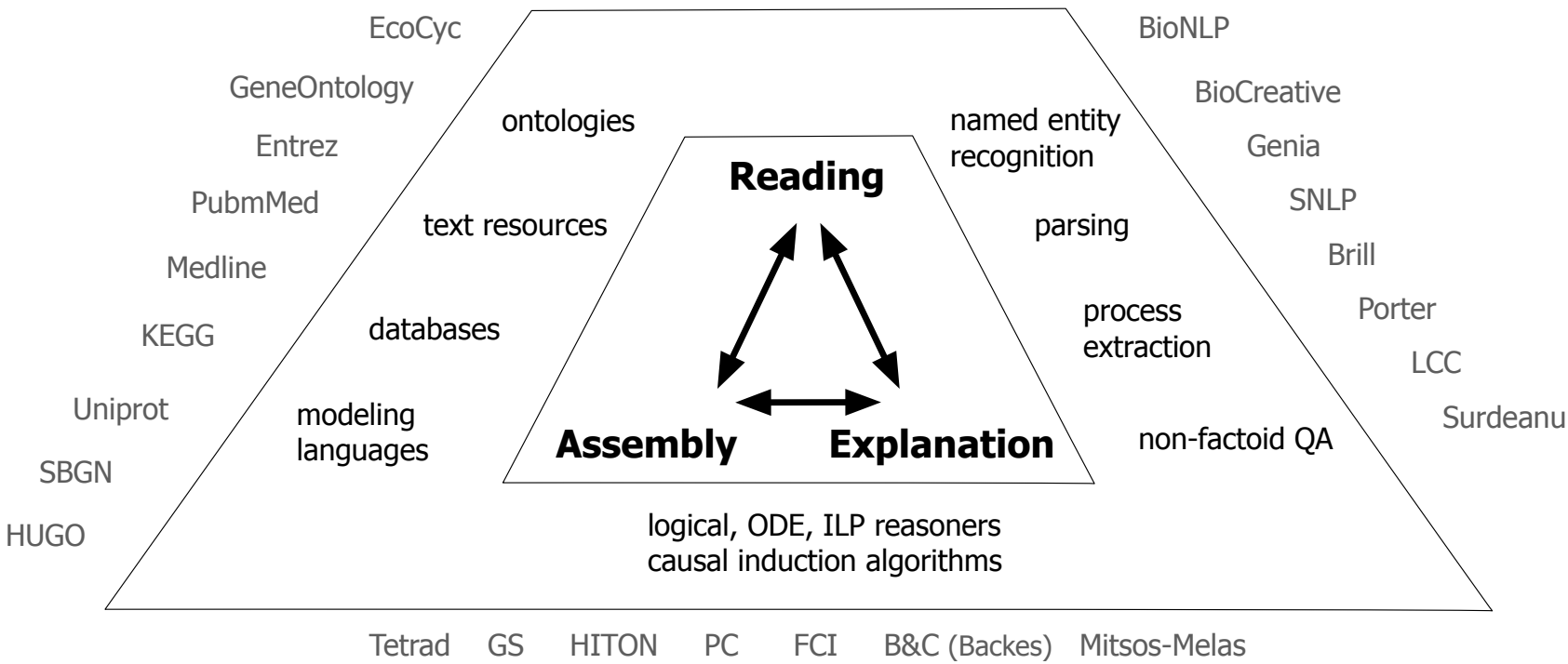


Integrate reading, assembly and explanation in a non-pipeline architecture that provides flexible control.

Approved for Public Release, Distribution Unlimited

The Integration Task – Technology Challenges

Reading, Assembly and Explanation shouldn't be pipelined but should use each other opportunistically. Need flexible, non-pipelined control, plus significant software integration.



Approved for Public Release, Distribution Unlimited

Conclusion

It's a big problem

Distinct communities to coordinate

statistical NLP

systems biology

knowledge-based NLP

data mining

ontology, databases

representation and reasoning

mathematical biology

Many domains

Potentially a new way to do science

BIG MECHANISM

Approved for Public Release, Distribution Unlimited