SELVENTA™

Big Data Problem?
or Big Problem with Data?

William Hayes, PhD

SVP Platform Dev, Selventa

Who am I?

- ex-Aerospace Engineer
- Defected to Bioinformatics (PhD in Molecular Biology)
- bioPharma Bioinformatics and Text Analytics
- Ran Biogen Idec Library & Literature Informatics
- Currently Head of Platform Development for Selventa
  - a Systems Diagnostics Company
- Visionary for a new Book Catalog
  - eLCe.us
- Co-founded a Drug Regulatory Documents Data Service
  - drugregdocs.com

Big Data sheds light on bioPharma's 'Small Data' problems

"The best thing about the Big Data hype in pharma is how effectively it's shed light on all of the Small Data problems the industry is facing."

"Industry analysts noticed this quickly, redefining Big Data in terms of the three (or four) V's--not just ***volume*** but also ***variety***, ***velocity***, and ***variability***."

http://www.fiercebiotechit.com/story/big-data-sheds-light-pharmas-small-data-problems/2013-03-27

http://www.fiercebiotechit.com/story/big-data-sheds-light-pharmas-small-data-problems/2013-03-27

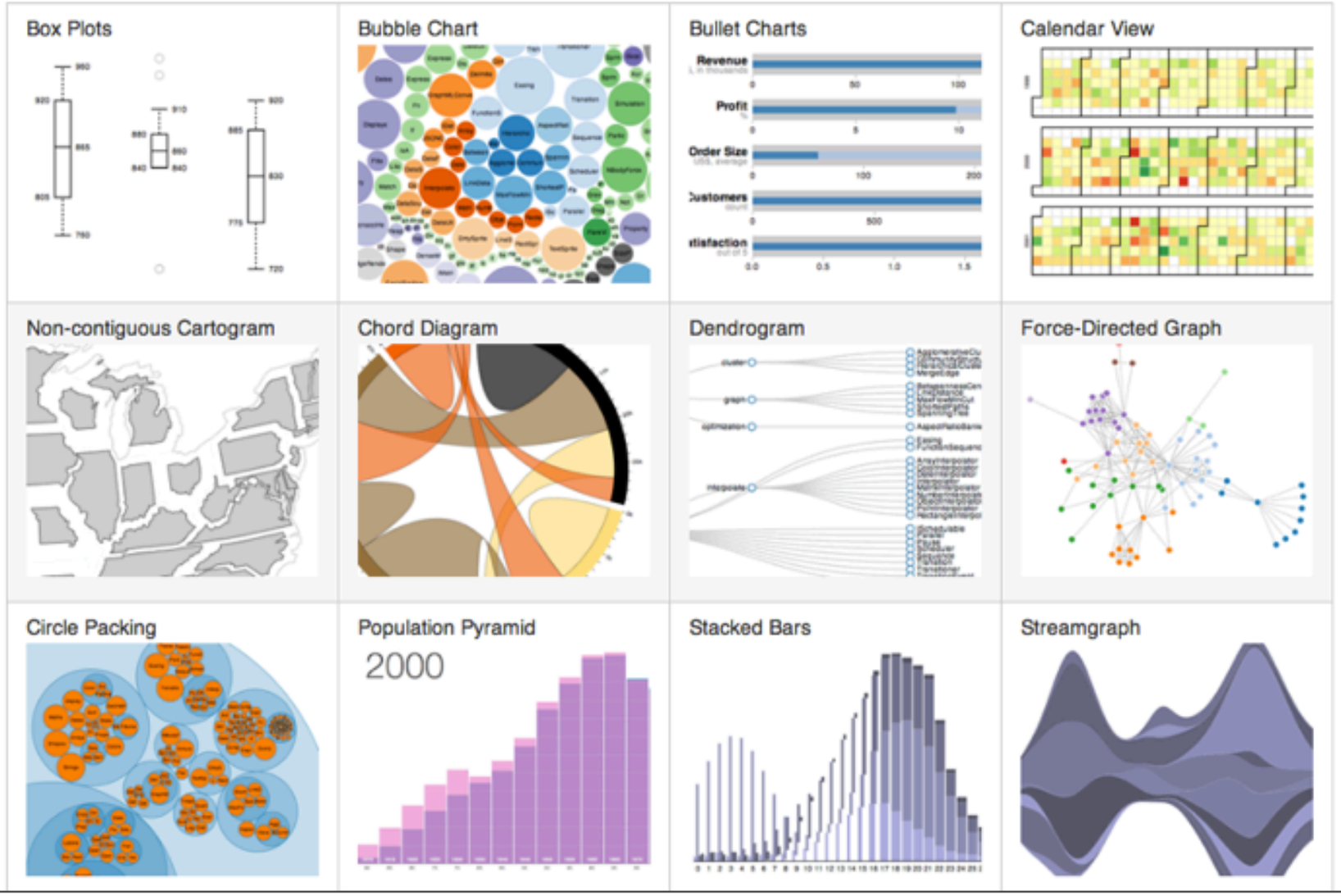
Trends

- Access to knowledge - no longer rare or hard to manage
  - Books and Journals were very hard to source and maintain
- Information Portfolio Management and Acquisition

- Information Integration - getting harder
- More Sophistication in Information Delivery
  - Flowing Data: http://flowingdata.com

D3js - Data-Driven Documents

Visual Index



Overall Information Service Workflow

- Collect/Acquire - success!
- Manage - success!
- Integrate - careful, danger ahead
- Filter/Analyze
- Transform
- Visualize
- Deliver/Re-use

Smart Data

- Coined by Lee Feigenbaum at Cambridge Semantics
  - Semantic University: http://www.cambridgesemantics.com/semantic-university

- Smart Data - is more:
  - Flexible
  - Understandable
  - Universal
  - Repurposable

http://www.cambridgesemantics.com/technology/smart-data-explained

Smart Data transforms the way data is discovered, integrated, searched, visualized, and analyzed. Unlike regular data, Smart Data is:

- *More flexible* — a smart data model accommodates new data as needed
- *More understandable* — data is represented using the same terminology and relationships that subject matter experts use to think about their domains
- *More universal* — data from very diverse sources with different (or no!) data model can be mapped onto a smart data model without losing any of the meaning of the original data
- *More repurposable* — smart data holds its meaning and so can be reused for many use cases, including ones not originally anticipated

Example

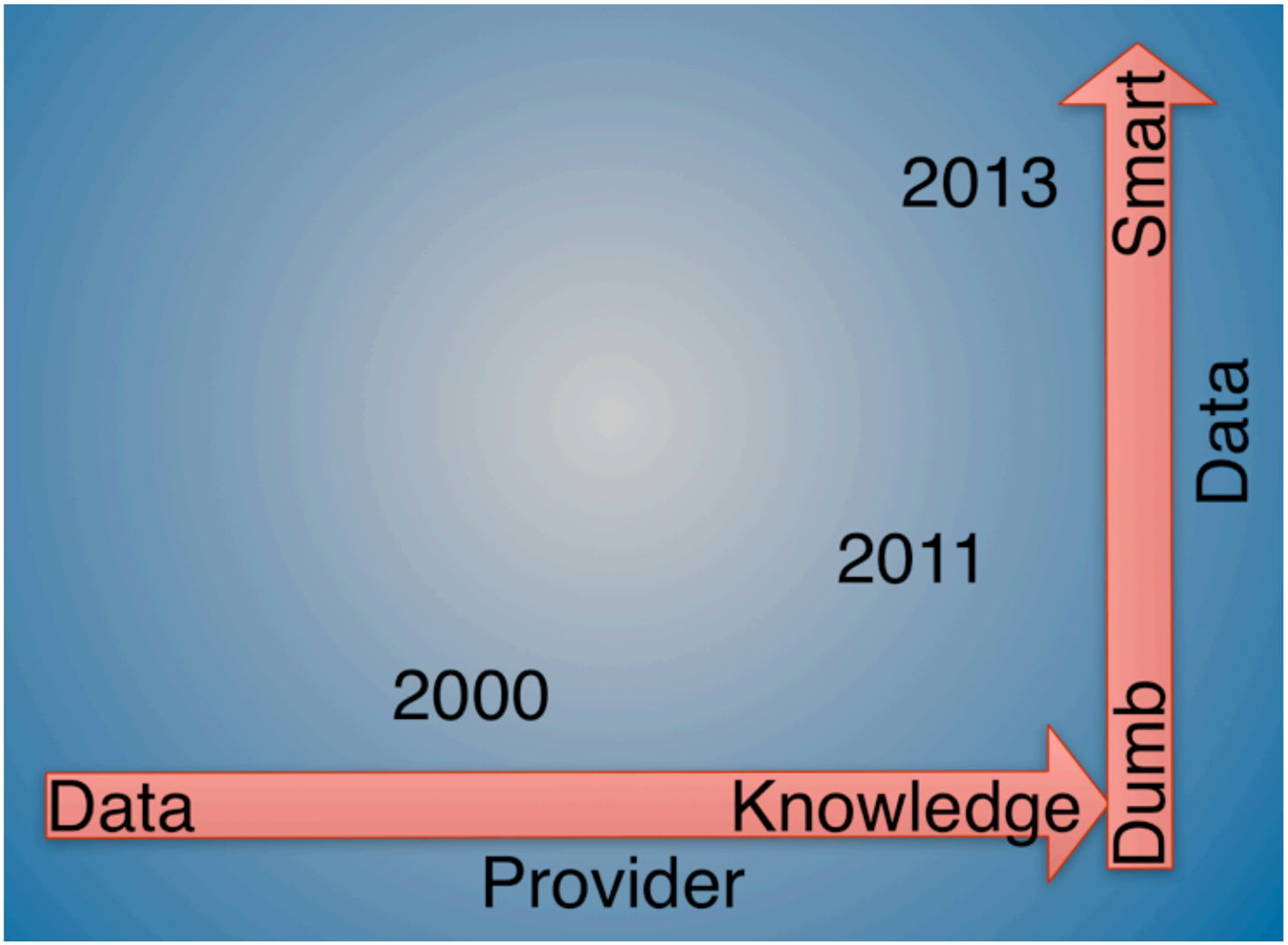
Example

EGFR

Example

HGNC:EGFR (I'm a human gene, that becomes the cell-surface receptor for Epidermal Growth Factor)

Where do you fit?



Grand Challenge I

Databases and Information Products

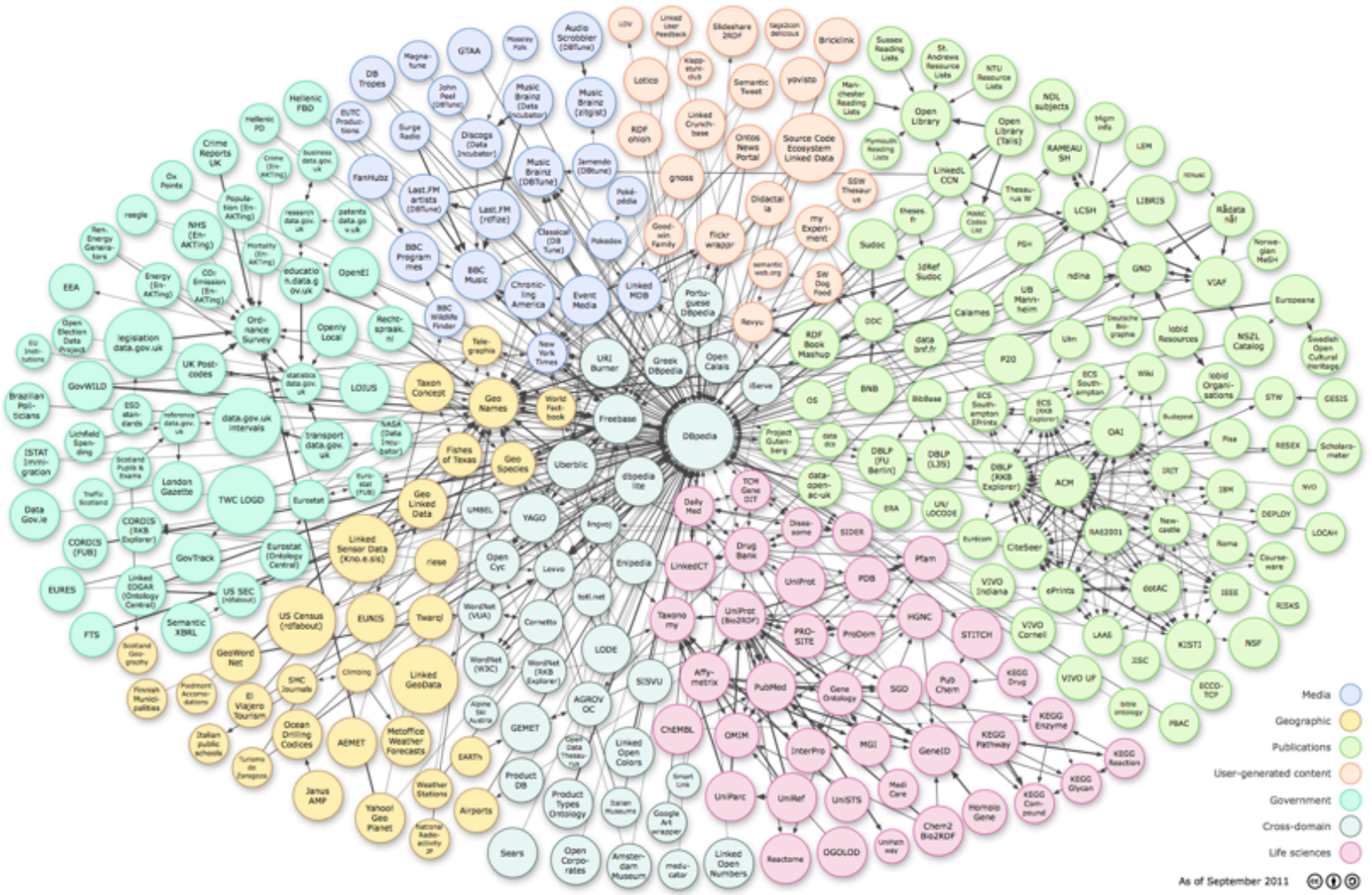
- Thomson Reuters
  - Financial = 165
  - Science = 102
- Elsevier = 65
- NAR Database Issue: 1380
- Metabase: ~1800

Semantic

Indexicus

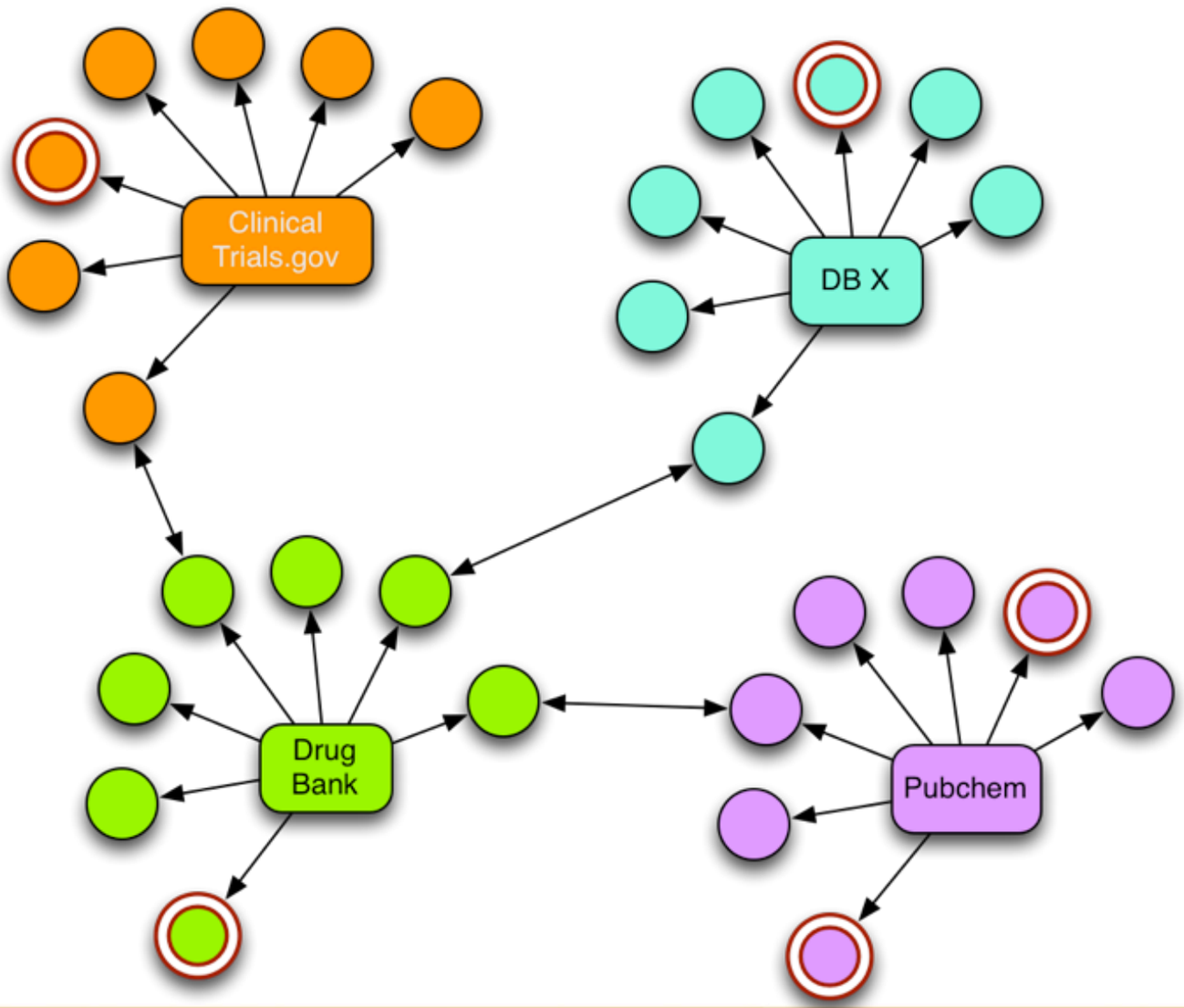
Databasicus

Linked Data Cloud



"Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"

Database of Databases



Grand Challenge II

"We are close to having
a $1,000 genome sequence,
but
this may be accompanied by
a $1,000,000 interpretation."

- Bruce Korf, Former President of the American College of Medical Genetics

This is evidence that we can't actually use the information in journal articles. That's a problem, because the domain here is very complicated, and it's very difficult (if not impossible) for scientists to synthesize lots of disparate information from a wide array of sources into something usable.

"We (Novartis) focus on a pathway mechanisms

...

Blockbusters are being re-defined as a drug working on the same mechanism across multiple diseases."

- Joseph Jimenez, CEO Novartis

http://money.cnn.com/video/magazines/fortune/2013/03/28/fortune-leadership-novartis-drugs.fortune

Re-using Knowledge

Major frustration while at Biogen Idec

OpenBEL

The Biologist's version of the Chemical Reaction Language

Repurposable and computable knowledge
from biological findings extracted from the literature

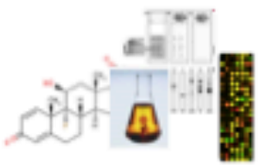
Capture and Re-use Biological Knowledge

Scientific Literature

"RNA expression of RBL2 is directly mediated via activation of the FOXO3 transcription factor"

```
tscript(p(HGNC:FOXO3)) => r(HGNC:RBL2)
```

J Biol Chem 2002 Nov 22 277(47) 45276-84



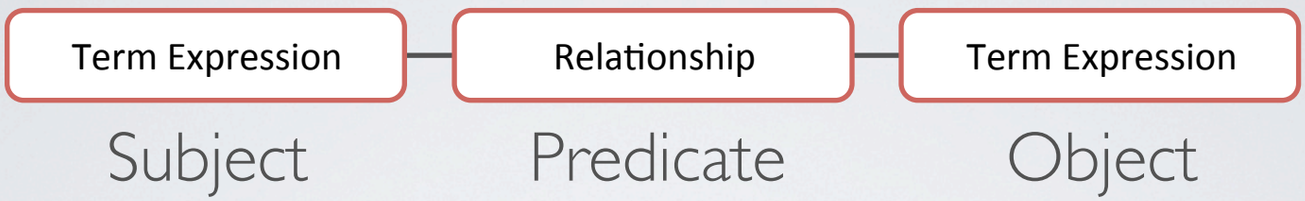
Original Research

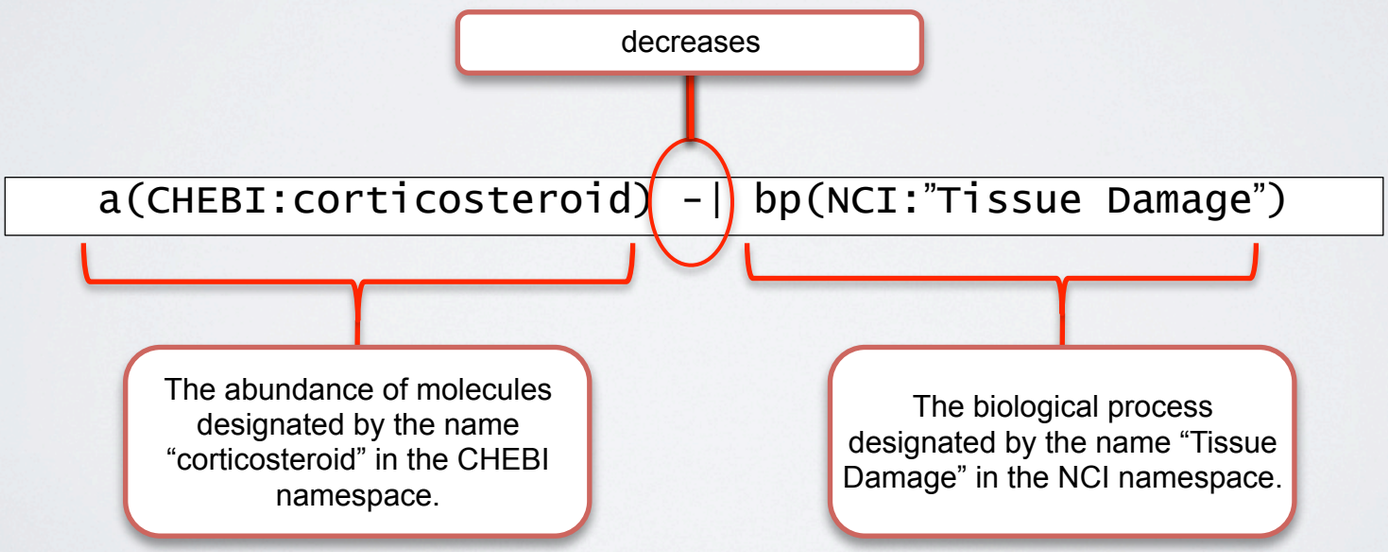
"LY294002 inhibits the activity of the PI3K alpha catalytic subunit"

```
a(PubChem:3973) -| kin(p(HGNC:PI3KCA))
```

XYZ Corp Document 12345

BEL Statements





BEL Statement Annotations Provide Context & Provenance

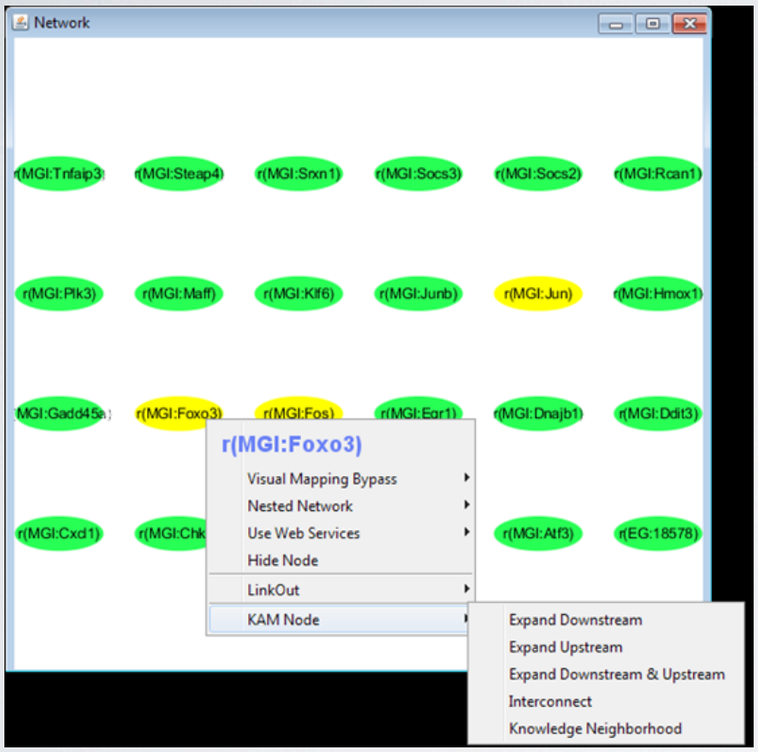
```
SET Citation = {“PubMed”, “J Mol Med (Berl). 2003 Mar;
81(3):168-74.“ , “12682725”}

SET Evidence =
       “high-dose steroid treatment decreases
       vascular inflammation and ischemic tissue damage
       after myocardial infarction and stroke through
       direct vascular effects involving the
       nontranscriptional activation of eNOS”


SET Tissue = “vascular system”


SET Disease = “Stroke”


a(CHEBI:corticosteroid) -| bp(NCI:”Tissue Damage”)
```

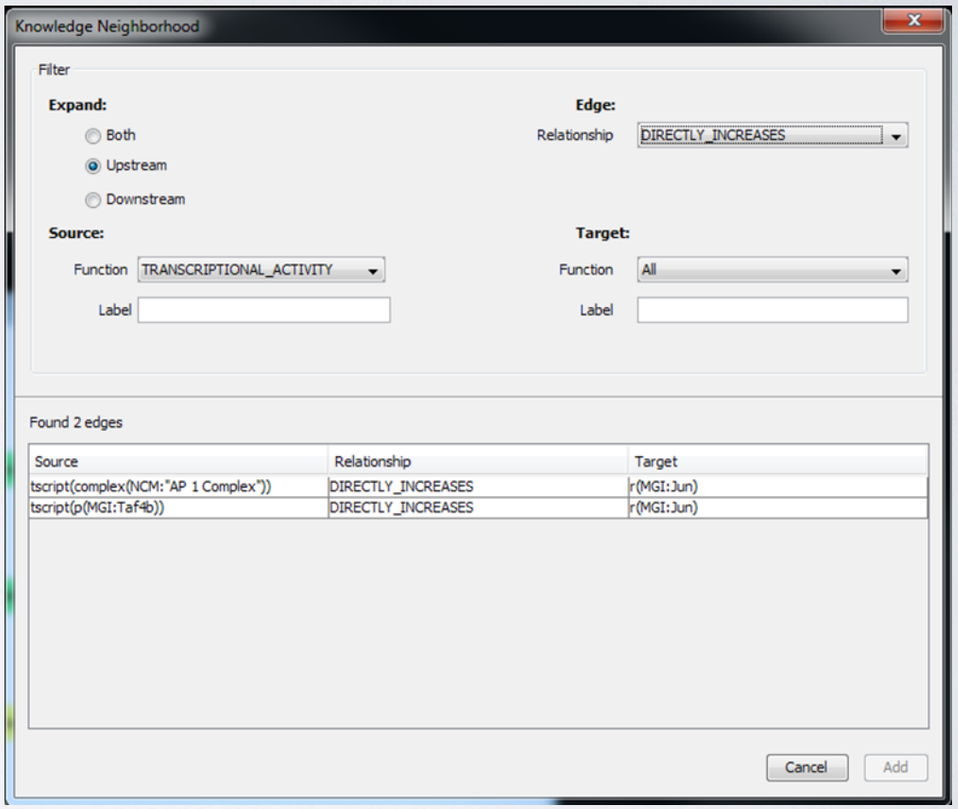
KAM Navigator
Visualize and explore networks in Cytoscape

KAM Navigator

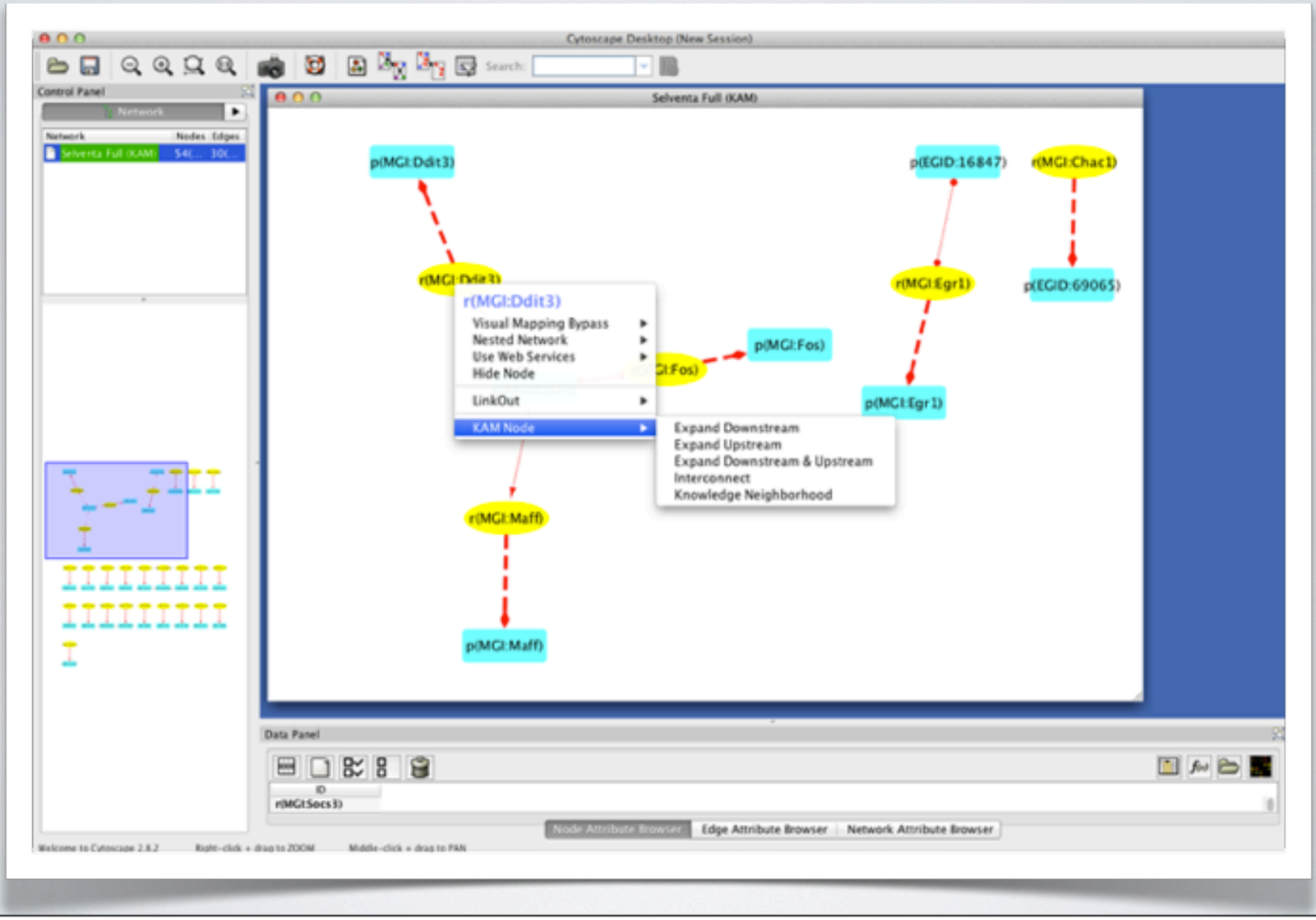
- Cytoscape plugin which enables the exploration of large BEL Knowledge Assembly Models (KAMs) and creation of sub-networks
  - Requires BEL Framework 2.0.0 or later, Cytoscape 2.8
  - BEL Framework Web server must be running and KAM(s) available in KAM store
- KAM nodes can be added to a Cytoscape network via:
  - Selecting by BEL Function and node label
  - From a list of namespace values
  - Neighbors of nodes already in the network

KAM Navigator – Adding Nodes and Edges





Cytoscape and KAM Navigator



BEL for bioPharma

- Library services – seeking, extracting and saving causal biology for use in R&D (disease biology, MoA, Toxicology, Clinical Trials, Drug Safety)
- Integrating databases (Ingenuity, AriadneGenomics, GeneGo, internal research requests) into single Knowledgebase
- Model Organism Translatability - e.g. how to check if mouse biology maps well with human biology for pre-clinical drug studies
- Disease vs Normal Biology Comparisons - comparing known biological differences between disease and normal states in an organism's tissue
- Assay Development - add tool compound and drug/target relations into BEL Network to aid development of biological assays

Starting Questions

- Does treating mouse model X with drug Y correlate to human biology?
- What is known biology for EGFR? Tissues expressed, interacting proteins, associated disease conditions, …
- What is found in the literature to associate kidney failure to steatosis (liver adverse event)?
- We need more detailed and highly contextualized information for Pathway X than what is in Ingenuity/GeneGo/etc
- How could using drug X with MoA Y affect tissue Z?

OpenBEL

http://openbel.org