Computing on the shoulders of giants:

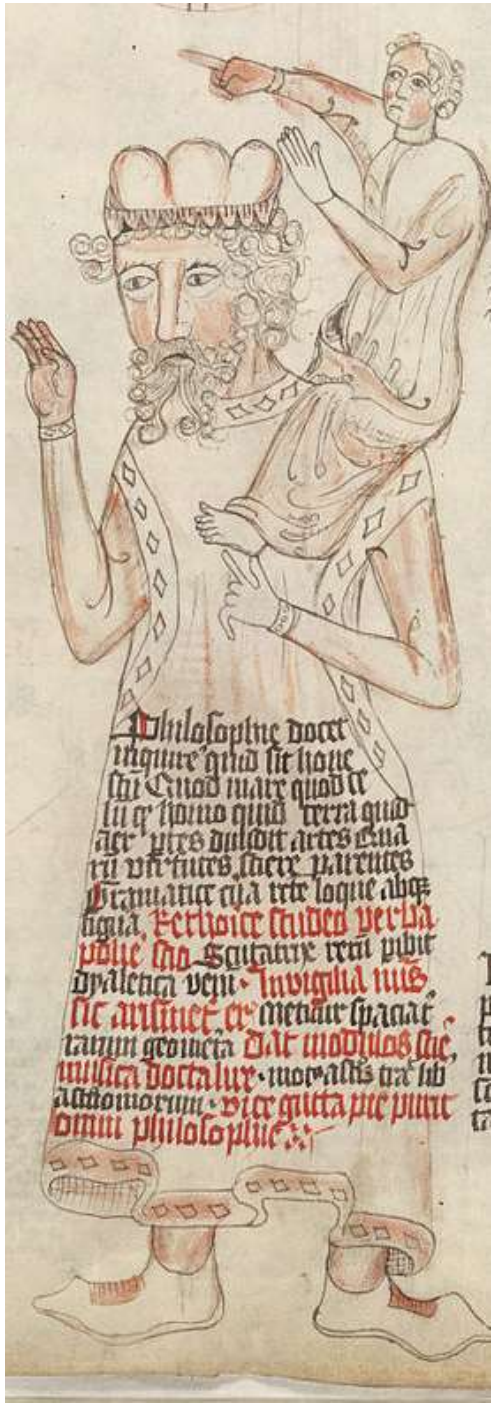
how existing knowledge is represented and applied in bioinformatics

Benjamin Good

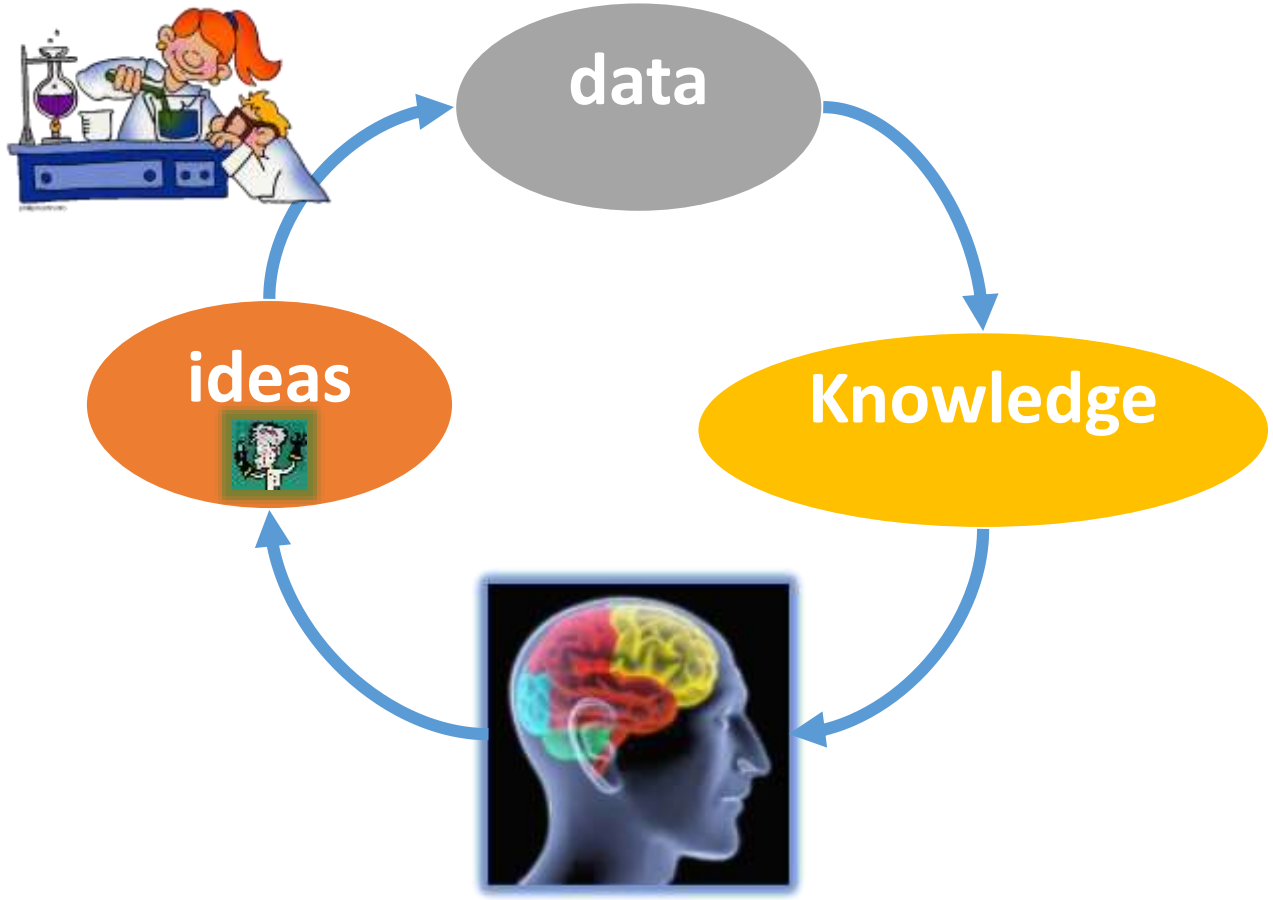
bgood@scripps.edu

Assistant Professor of the Department of Molecular and Experimental Medicine

Specialty: artificial intelligence, crowdsourcing



The more you can ‘know’ the better a scientist you can become

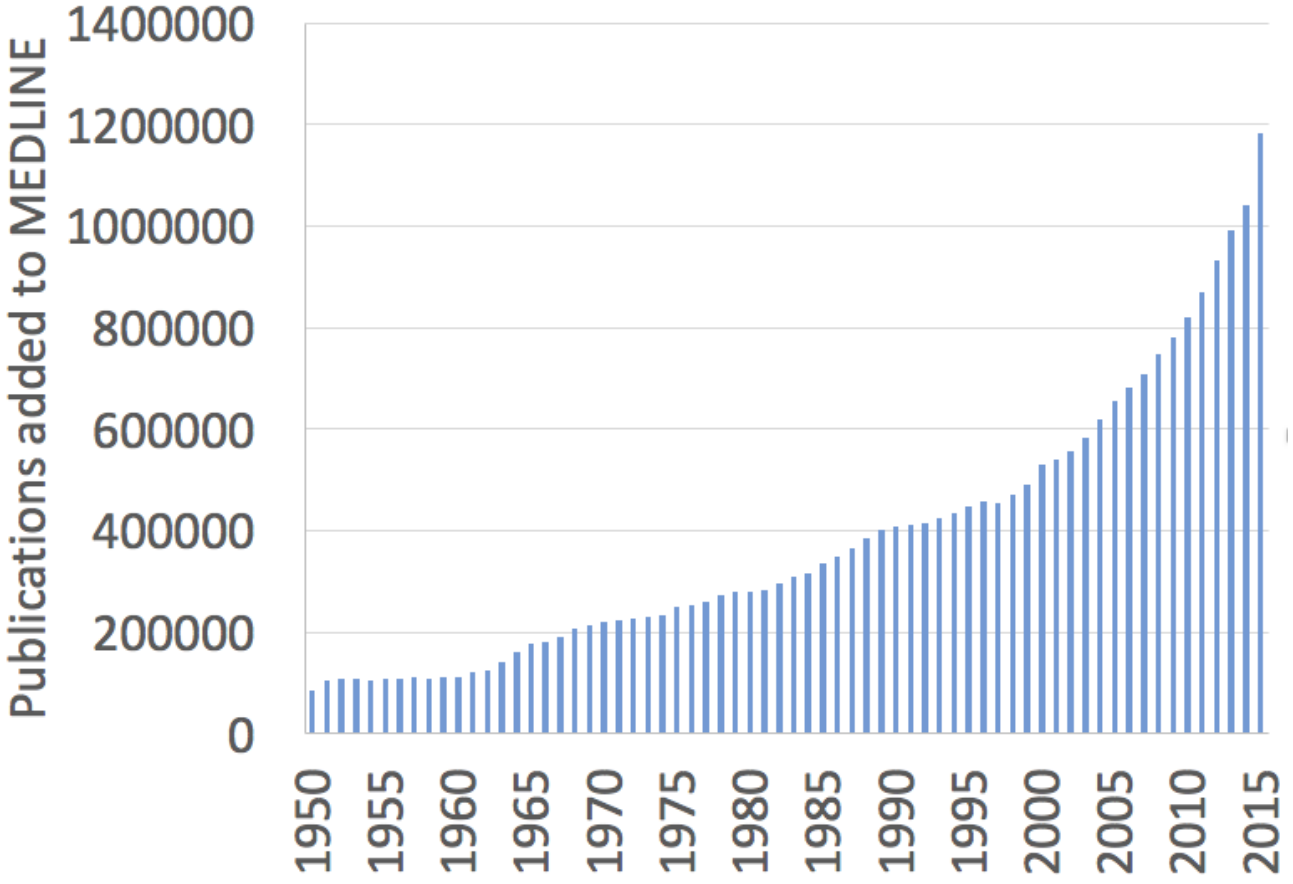


(Derived from Alex Pico, WikiPathways)

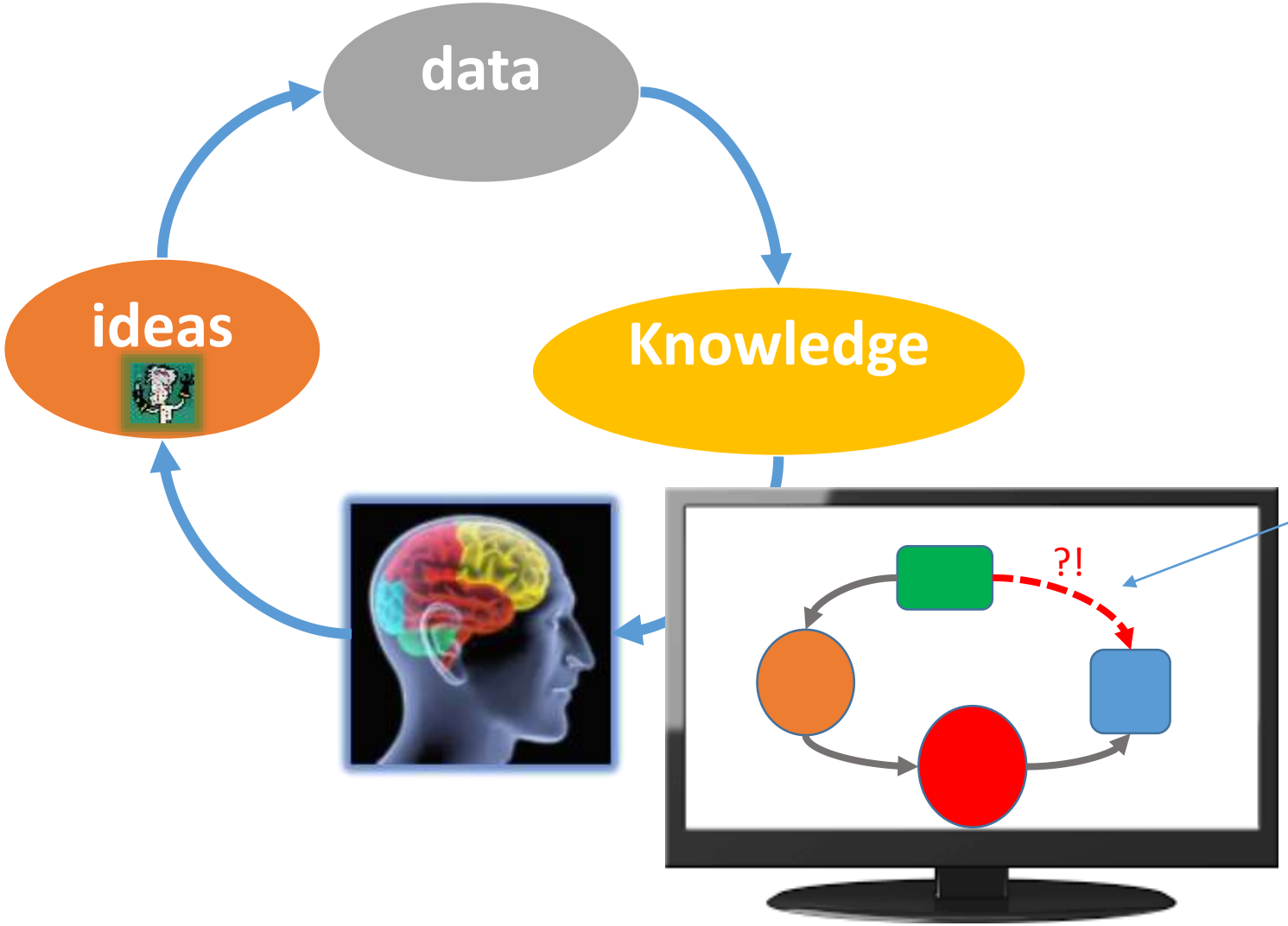
Too much to know

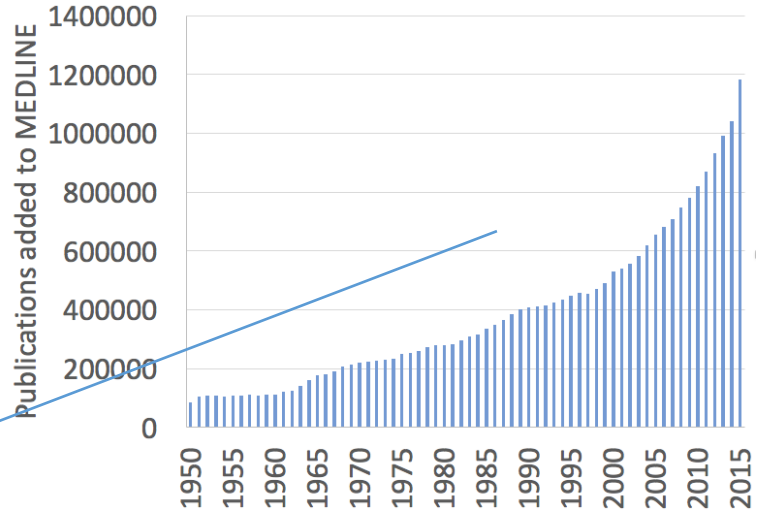
Knowledge

- PubMed lists > 1 million articles published each year (more than 2 per minute)
- **Your capacity to read and comprehend is limiting**



Knowledge representation





Goals for representing knowledge (outline)

- Make things (articles, genes, antibodies, etc.) easier to find
- Answer questions
- Generate hypotheses

Controlled vocabularies (MeSH)
Ontologies (Gene Ontology)

knowledge graphs on the Web: the SPARQL query language

knowledge plus computation = inference, the ABC model

Part 1:
Medical Subject Headings (MeSH)

Finding what to read:
controlled vocabularies for indexing PubMed

- What happens when you search PubMed?

PubMed | fainting | Search

Create RSS Create alert Advanced Help

Summary 20 per page Sort by Most Recent Send to:

Filters: Manage Filters

Search results

Items: 1 to 20 of 18088

<< First < Prev Page 1 of 905 Next > Last >>

-induced complete heart block and

2016-214474.

ng in patients with **syncope**].

a A, Broullón FJ, Álvarez-García N,

j.medcli.2016.03.041. [Epub ahead of

Results by year

PMC Images search for *fainting*

Titles with your search terms

Find related data

Search details

```
"syncope"[MeSH Terms] OR "syncope"
[All Fields] OR "fainting"[All
Fields]
```

Search See more...

PubMed | "fainting"

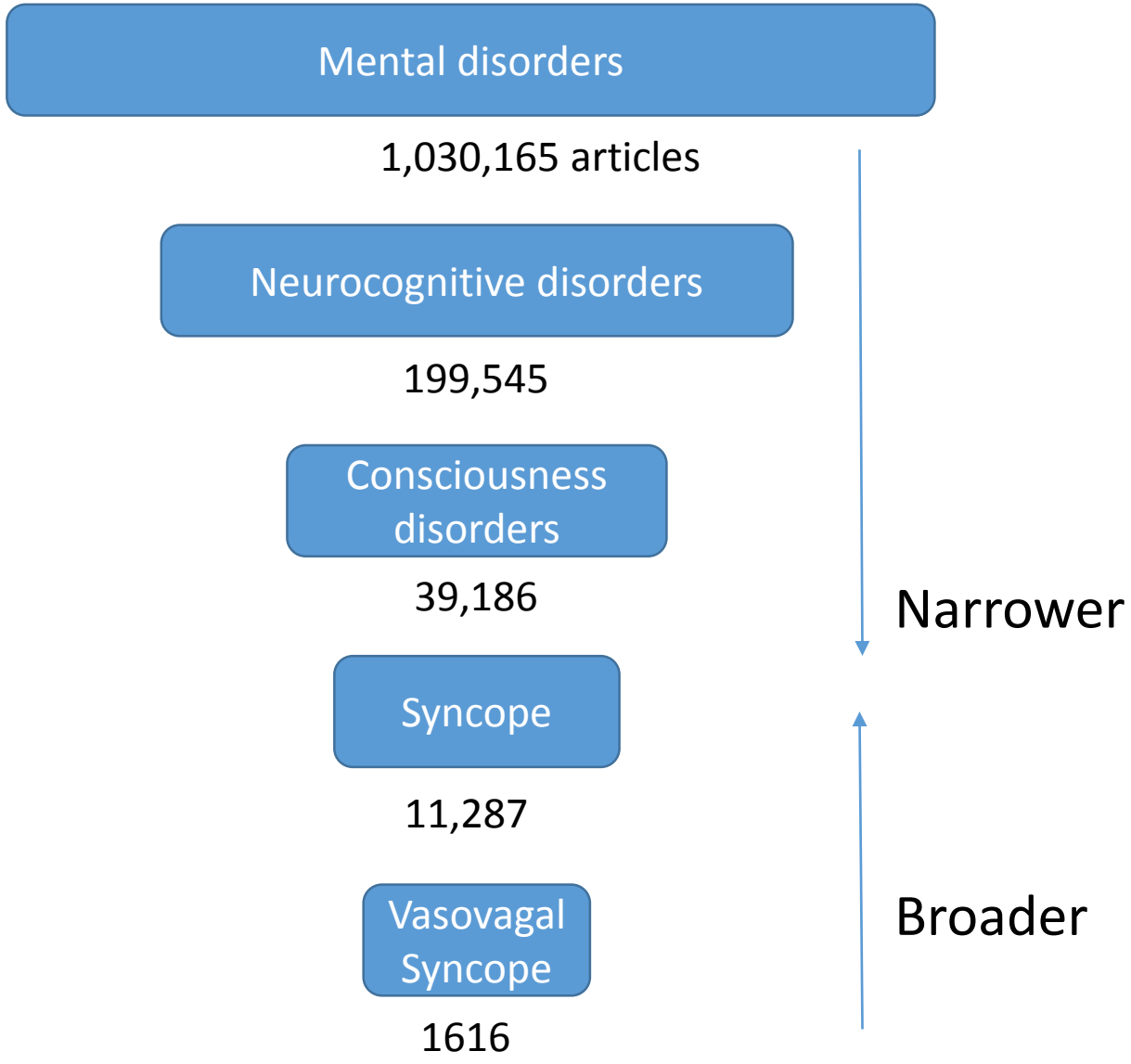
Create RSS Create alert

Summary 20 per page Sort by Most R

Search results

Items: 1 to 20 of 976

MeSH controlled vocabulary (AKA ‘thesaurus’)



- **Descriptor Unique ID**: D013575
- **Definition**: A transient loss of consciousness and postural tone caused by diminished blood flow to the brain…
- **Entry Terms**: Syncopes, Fainting, Syncopal Vertigo, Presyncope, Drop Attack, Carotid Sinus Syncope,…
- **Relations to other terms**

MeSH: medical subject headings

- >27,000 descriptors
- >87,000 entry terms
- 16 hierarchical trees
- Constantly being revised

1. + Anatomy [A]
2. − Organisms [B]
    - Eukaryota [B01] +
    - Archaea [B02] +
    - Bacteria [B03] +
    - Viruses [B04] +
    - Organism Forms [B05] +
3. + Diseases [C]
4. + Chemicals and Drugs [D]
5. + Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. + Psychiatry and Psychology [F]
7. + Phenomena and Processes [G]
8. + Disciplines and Occupations [H]
9. + Anthropology, Education, Sociology and Social Phenomena [I]
10. + Technology, Industry, Agriculture [J]
11. + Humanities [K]
12. + Information Science [L]
13. + Named Groups [M]
14. + Health Care [N]
15. + Publication Characteristics [V]
16. + Geographicals [Z]

Demo and play time

- View and explore the MeSH trees:
  - https://www.nlm.nih.gov/mesh/2016/mesh_browser/MeSHtree.html
- Use MeSH to query PubMed
  - Go to: http://www.ncbi.nlm.nih.gov/mesh
  - Search for the term 'fainting'
  - click 'Add to search builder'
  - click search PubMed
  - click back, search for other things..

Query demos

- **Query expansion**
  - Hand Bones [Mesh]
  - Hand Bones [Mesh:NoExp]
- **Boolean operators**
  - cardiac hypertrophy and use rodents besides mice and rats in their experiments
    - ("Cardiomegaly"[Mesh])
    - AND "Rodentia"[Mesh]
    - NOT "Mice"[Mesh] NOT "Rats"[Mesh]
- **Article type filter**
  - Review papers about cardiac hypertrophy
  - Cardiomegaly [MeSH] AND Review[ptyp]
  - Try with http://www.ncbi.nlm.nih.gov/pubmed/advanced

Questions about MeSH ?

- Good 3 minute tutorial video on practical use: http://www.youtube.com/watch?v=uyF8uQY9wys

Part 2: Ontology

“Ontology”

- The word comes from philosophy:
  - “the branch of metaphysics dealing with the nature of being”
- In practice they are:
  - **A set of concepts, definitions and inter-relationships.**
  - (The dividing line between “controlled vocabulary”, “thesaurus”, “ontology” is hazy and not terribly important for practical purposes.)
- We have hundreds of ontologies in biology, e.g. see:
  - http://www.obofoundry.org (100+)
  - http://bioportal.bioontology.org (500+)

The Gene Ontology

Started in 1999

As a collaboration between 3 Model Organism Databases

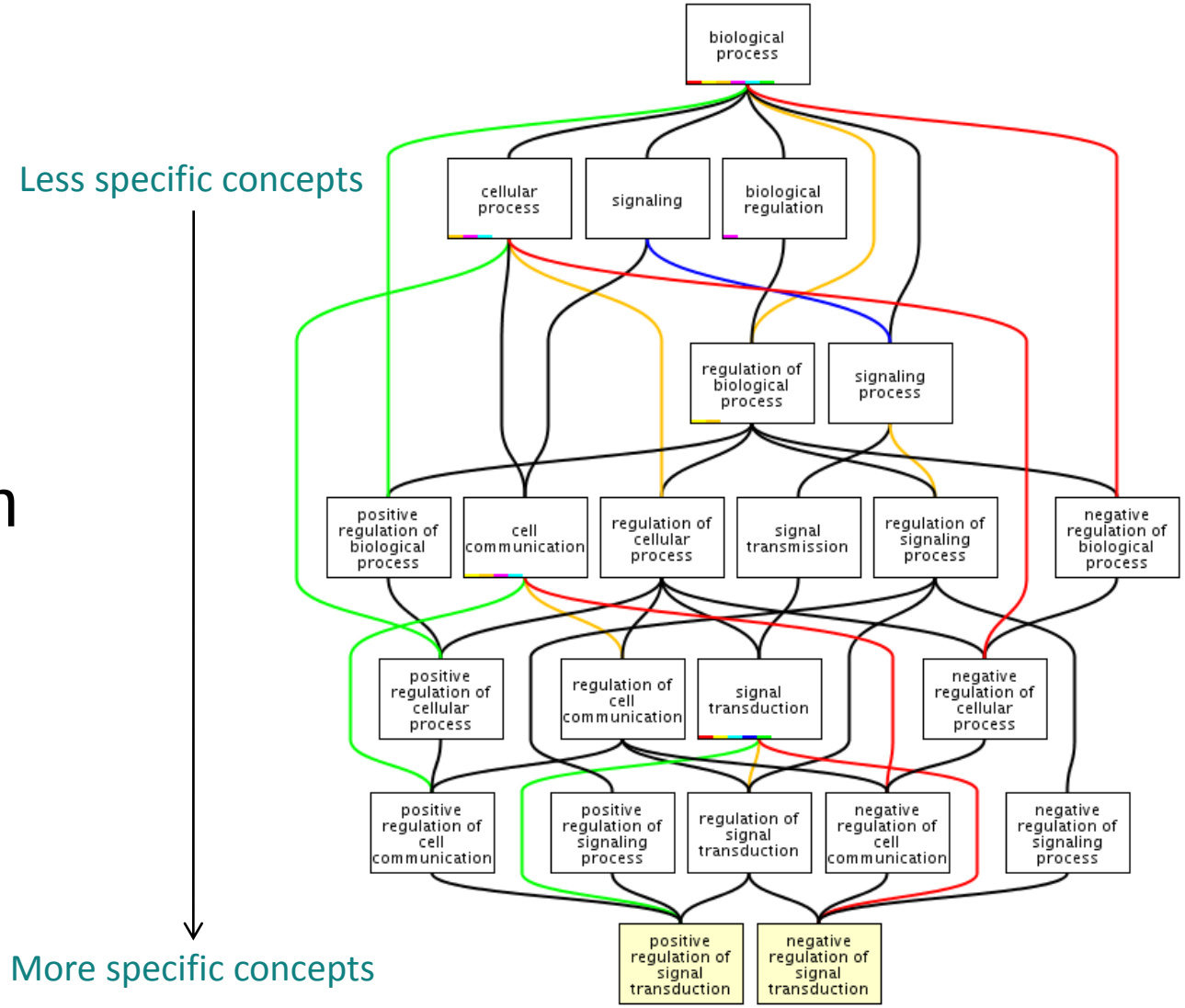
Slide credit: Mélanie Courtot, Ph.D.

Ashburner et al., Nat Genet. 2000 May;25(1):25-9.

The Gene Ontology

- A way to capture biological knowledge for individual gene products in a computable form

- A set of concepts and their relationships to each other arranged as a hierarchy



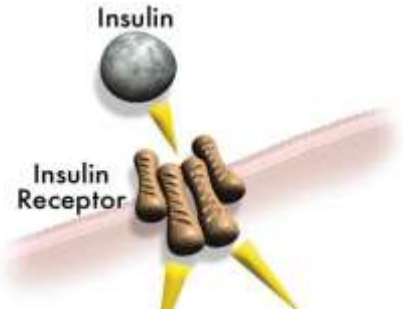
http://www.ebi.ac.uk/QuickGO

Slide credit: Mélanie Courtot, Ph.D.

The GO branches

1. Molecular Function

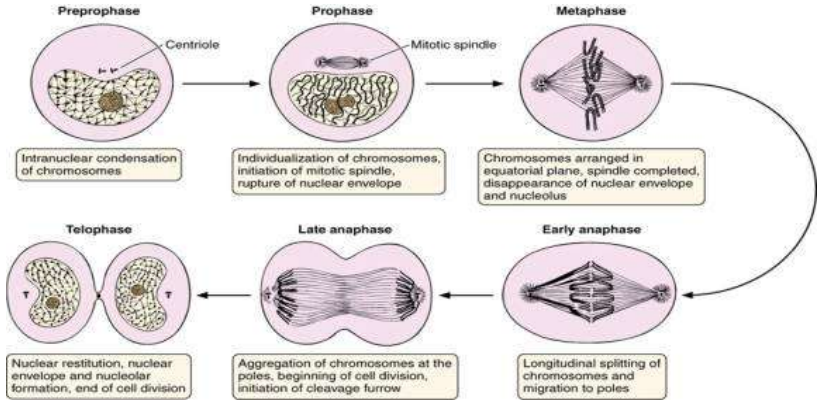
An elemental activity or task or job



- protein kinase activity
- insulin receptor activity

2. Biological Process

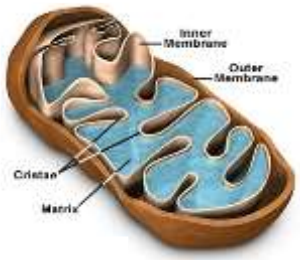
A commonly recognized series of events



- cell division

3. Cellular Component

Where a gene product is located

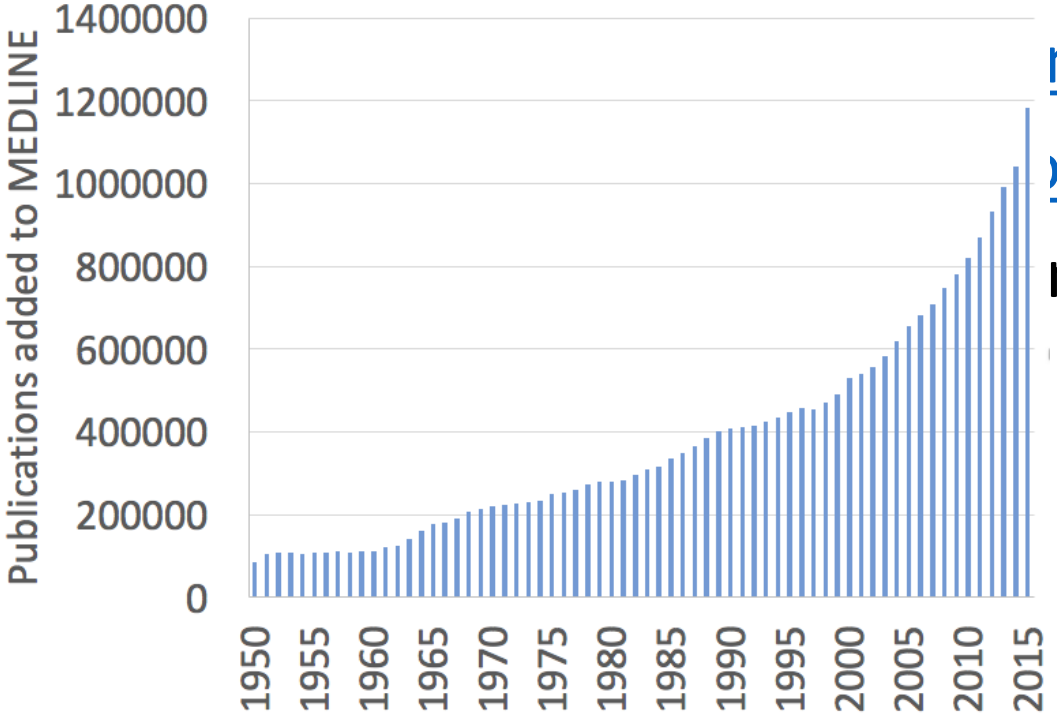


- mitochondrion
- mitochondrial matrix
- mitochondrial inner membrane

Slide credit: Mélanie Courtot, Ph.D.

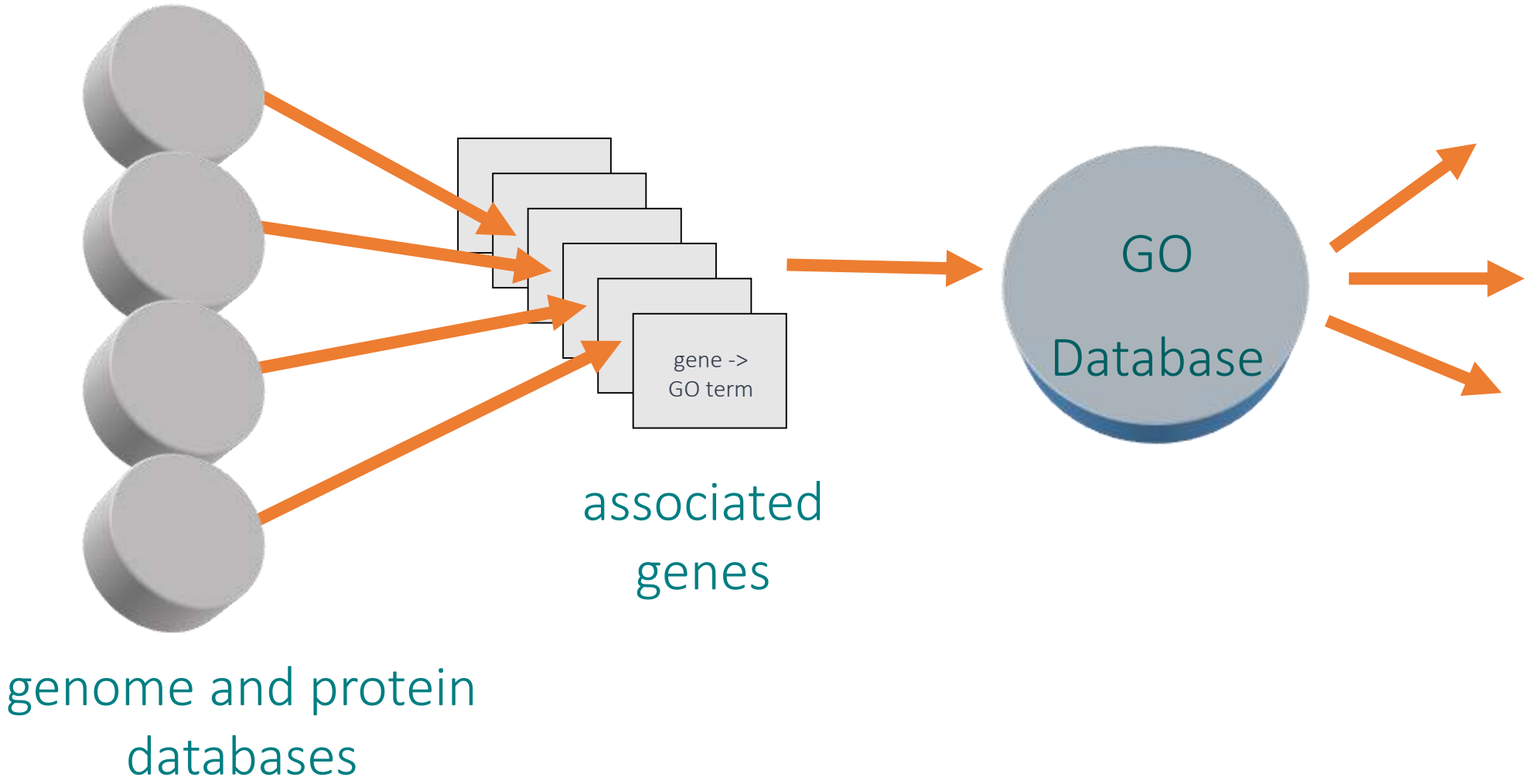
Building the GO
(now covering more than 40,000 terms)

- GO editorial team based at the European Bioinformatics Institute
- Submission via ntology/
- Submissions vi rg
- In principal, an ntology, but the GO editors make tl



Slide credit: Mélanie Courtot, Ph.D.

Using the GO to describe gene products



Slide credit: Mélanie Courtot, Ph.D.

Contributors

PomBase

dictyBase

MGI

UniProt

EnsemblGenomes

GO Annotating cardiovascular genes

EcoliHub

WormBase

go

FlyBase

interPro Protein sequence analysis & classification

tair

SGD

GRAMENE

REACTOME

Berkeley Lab

AgBase

ZFIN

J. Craig Venter Institute

University of Maryland School of Medicine Institute for Genome Sciences

RGD

Slide credit: Mélanie Courtot, Ph.D.

A GO annotation is …

…a statement that a gene product;

Accession	Name	GO ID	GO term name	Reference	Evidence code
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA

Slide credit: Mélanie Courtot, Ph.D.

A GO annotation is …

…a statement that a gene product;

1. has a particular **molecular function**
   or is involved in a particular **biological process**
   or is located within a certain **cellular component**

Accession	*Name*	*GO ID*	*GO term name*	*Reference*	*Evidence code*
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA

Slide credit: Mélanie Courtot, Ph.D.

A GO annotation is …

…a statement that a gene product;

1. has a particular **molecular function**
   or is involved in a particular **biological process**
   or is located within a certain **cellular component**
2. as described in a particular reference

Accession	*Name*	*GO ID*	*GO term name*	*Reference*	*Evidence code*
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA

Slide credit: Mélanie Courtot, Ph.D.

A GO annotation is …

…a statement that a gene product;

1. has a particular **molecular function**
   or is involved in a particular **biological process**
   or is located within a certain **cellular component**
2. as described in a particular reference
3. according to a particular method

Accession	*Name*	*GO ID*	*GO term name*	*Reference*	*Evidence code*
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA

Slide credit: Mélanie Courtot, Ph.D.

Kinds of evidence for GO annotations by curators

Experimental data

Inferred from Experiment (EXP)
Inferred from Direct Assay (IDA)
Inferred from Physical Interaction (IPI)
Inferred from Mutant Phenotype (IMP)
Inferred from Genetic Interaction (IGI)
Inferred from Expression Pattern (IEP)

Computational analysis

Inferred from Sequence or structural Similarity (ISS)
Inferred from Sequence Orthology (ISO)
Inferred from Sequence Alignment (ISA)
Inferred from Sequence Model (ISM)
Inferred from Genomic Context (IGC)
Inferred from Biological aspect of Ancestor (IBA)
Inferred from Biological aspect of Descendant (IBD)
Inferred from Key Residues (IKR)
Inferred from Rapid Divergence(IRD)
Inferred from Reviewed Computational Analysis (RCA)

Author statements/ curator inference

Traceable Author Statement (TAS)
Non-traceable Author Statement (NAS)
Inferred by Curator (IC)
No biological Data available (ND) evidence code

Slide credit: Mélanie Courtot, Ph.D.

http://geneontology.org/page/guide-go-evidence-codes

Inferred from Electronic Annotation (IEA)

The one evidence code used for completely automated annotation

http://geneontology.org/page/guide-go-evidence-codes

Manual annotations

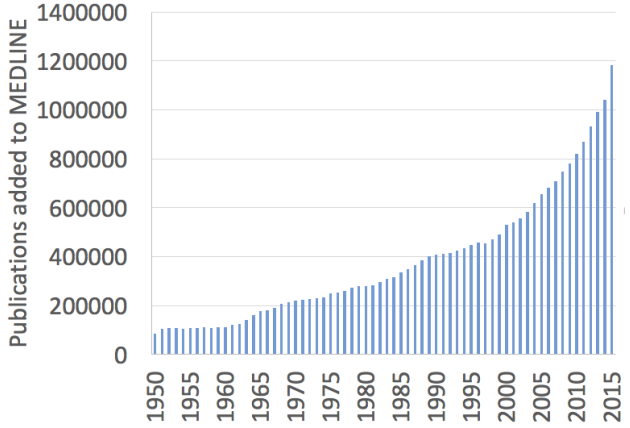
- Time-consuming process producing lower numbers of annotations (~2,800 taxons covered)
- More specific GO terms
- Manual annotation is essential for creating predictions

Aleksandra Shypitsyna

Elena Speretta

Alex Holmes

Tony Sawford



Slide credit: Mélanie Courtot, Ph.D.

Electronic Annotations (IEA)

- Quick way of producing large numbers of annotations
- Annotations use less-specific GO terms
- Only source of annotation for ~438,000 non-model organism species

orthology

interpro2go

taxon constraints

Slide credit: Mélanie Courtot, Ph.D.

A public resource of data and tools

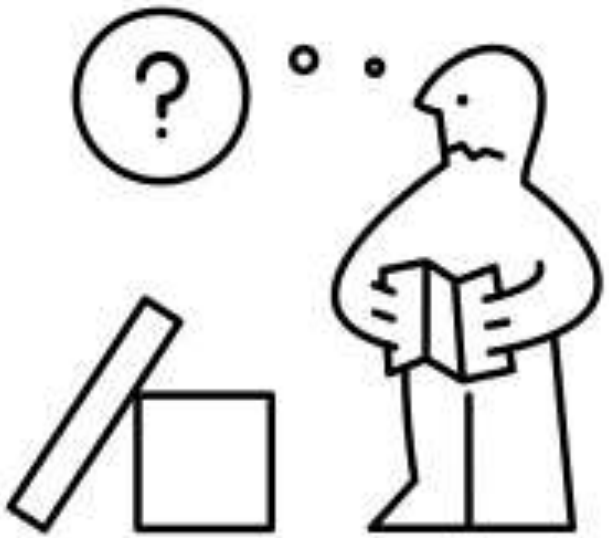
Number of annotations in UniProt-GOA database (March 2016)

Electronic annotations	269,207,317
Manual annotations*	2,752,604

* Includes manual annotations integrated from external model organism and specialist groups

https://www.ebi.ac.uk/QuickGO/
http://www.ebi.ac.uk/GOA

Slide credit: Mélanie Courtot, Ph.D.



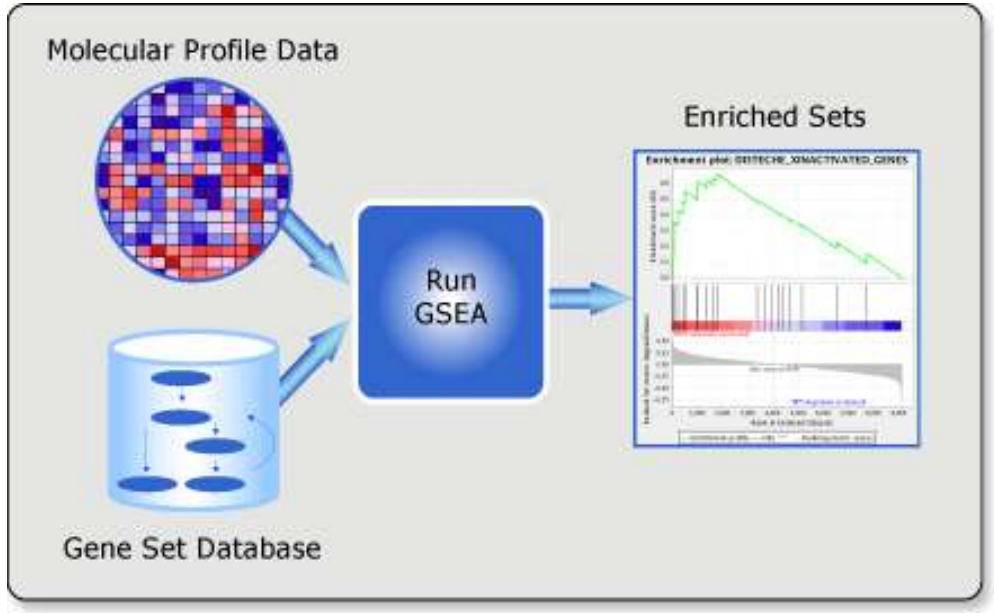
Slide credit: Mélanie Courtot, Ph.D. & IKEA

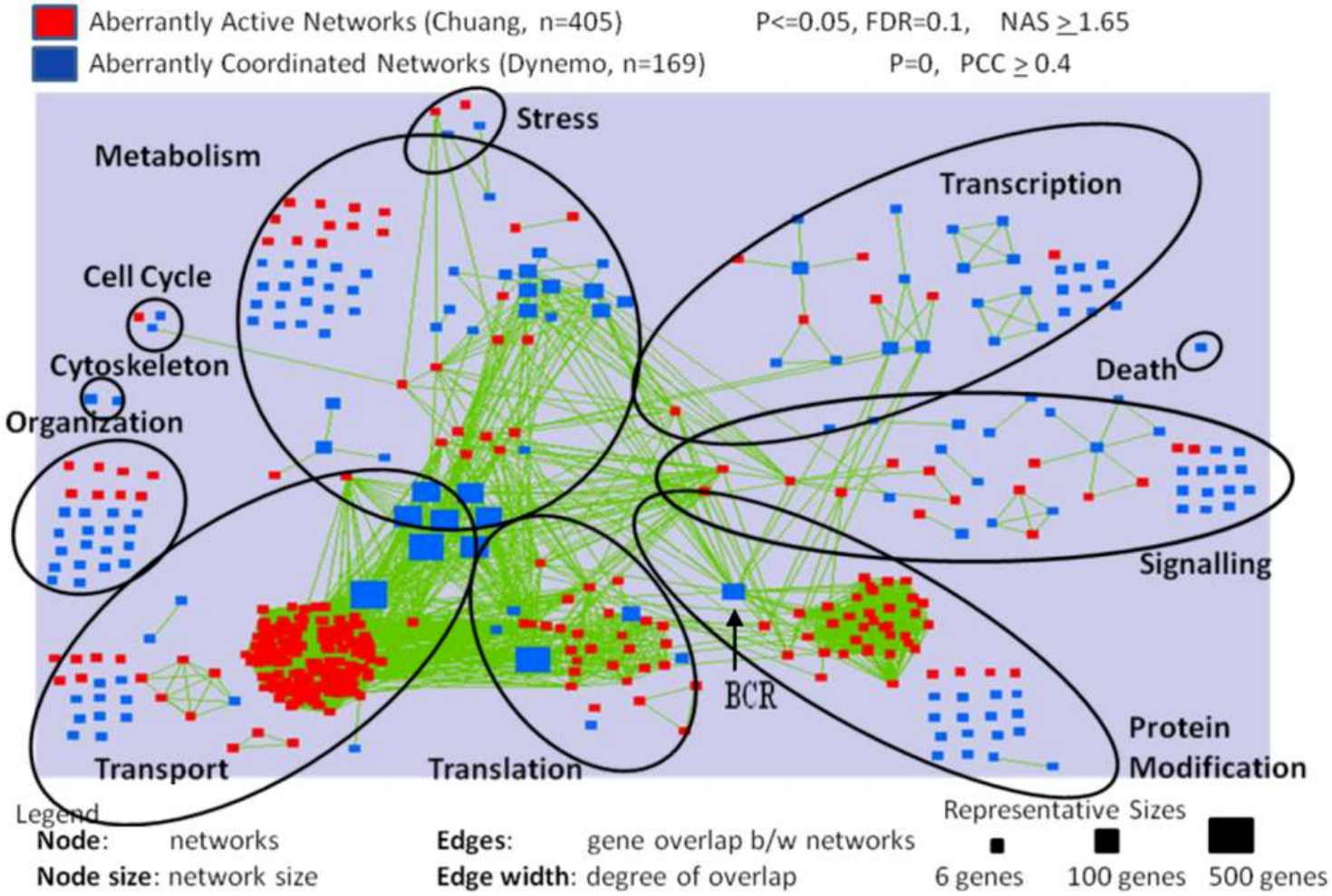
Using AMIGO2: http://amigo.geneontology.org

- Find the Gene Ontology term for Nucleus
- Find its child term Pronucleus
- Find a C. Elegans gene associated with this term and find the PubMed id of the reference supporting the annotation
- Repeat for a human gene, what is the evidence for the annotation?

http://geneontology.org/page/guide-go-evidence-codes

Gene Set Enrichment Analysis (previously covered)



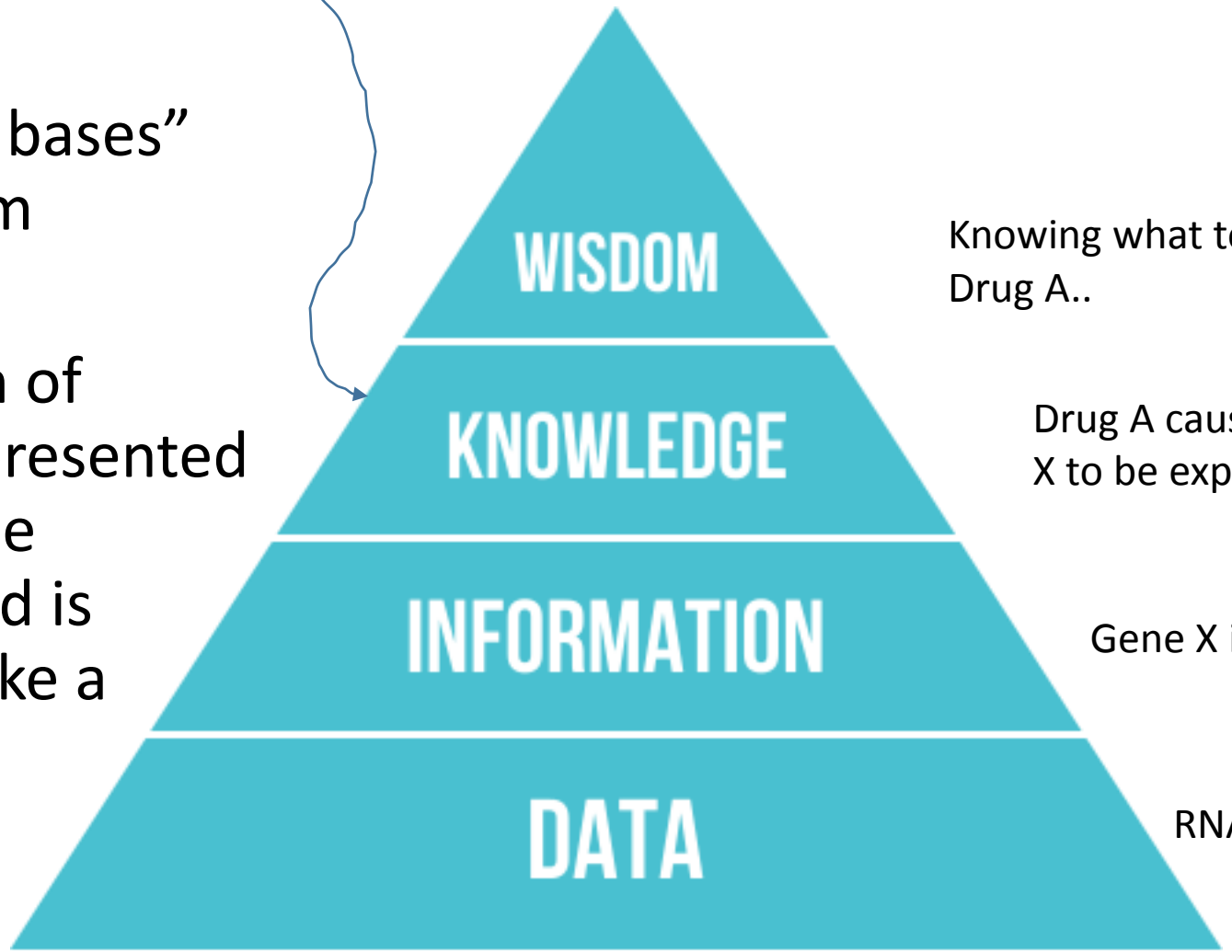


Questions about GO or other ontologies?

Part 3: Knowledge graphs

Knowledge Graphs

- Also called “knowledge bases” to distinguish them from databases.
- An integrated collection of assertions or claims represented in something that can be visualized as a graph and is technically very much like a database.



Knowing what to do with Drug A..

Drug A caused Gene X to be expressed

Gene X is expressed

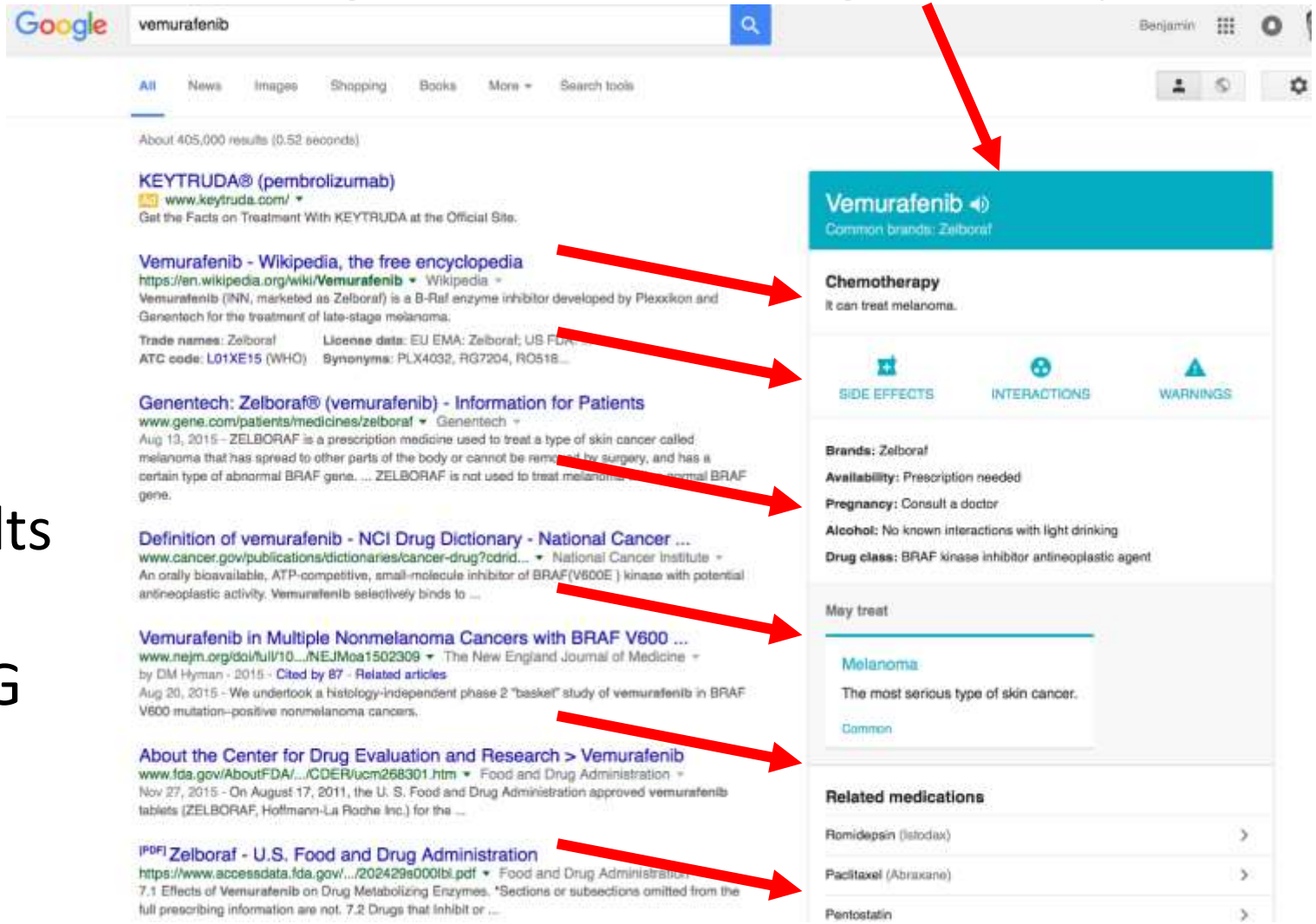
RNASeq reads

Example knowledge graphs

- Wikidata: The structured equivalent of Wikipedia
  - http://wikidata.org
- UniProt Knowledge Base: Manually curated Protein knowledge base
  - http://www.uniprot.org/uniprot/
- Microsoft Knowledge Graph (“Satori”)
- Google Knowledge Graph

Example: “Google Knowledge Graph” (GKG)

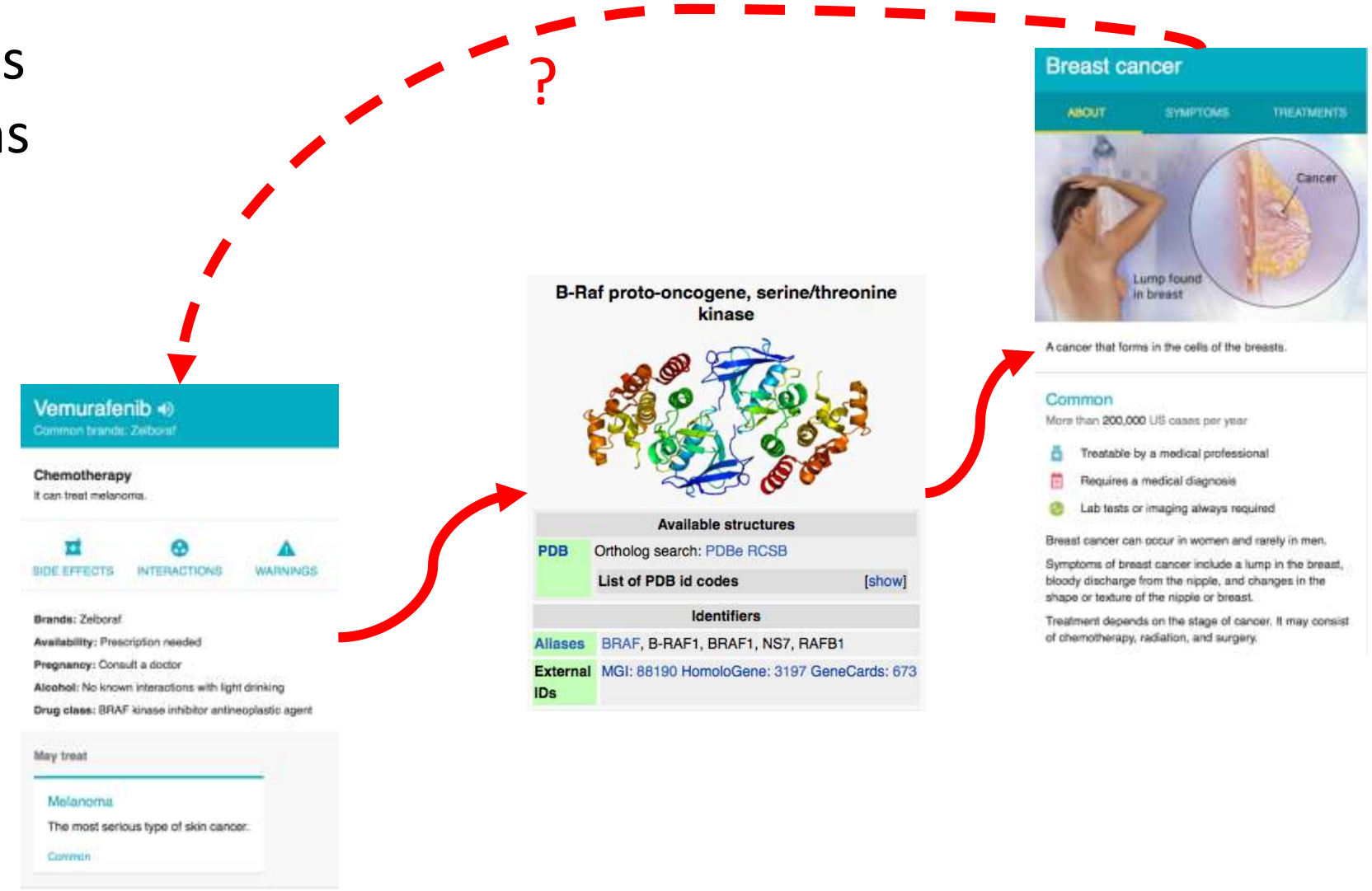
Vemurafenib
405,000 results
1 infobox
1 node in GKG



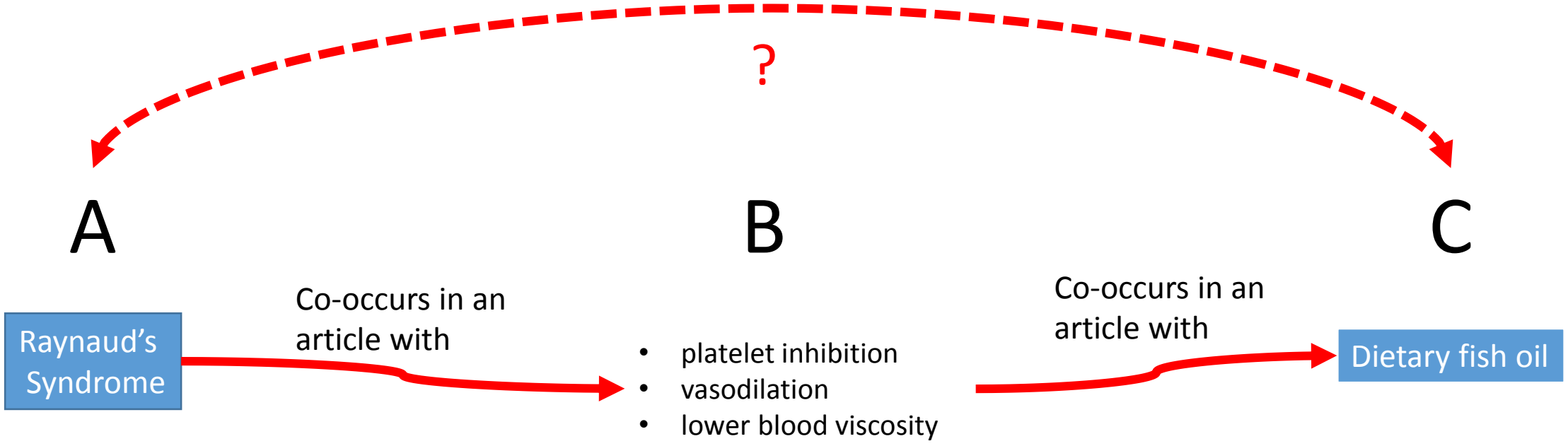
https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html

Why Knowledge Graphs?

- Answer explicit questions
- Uncover implicit relations



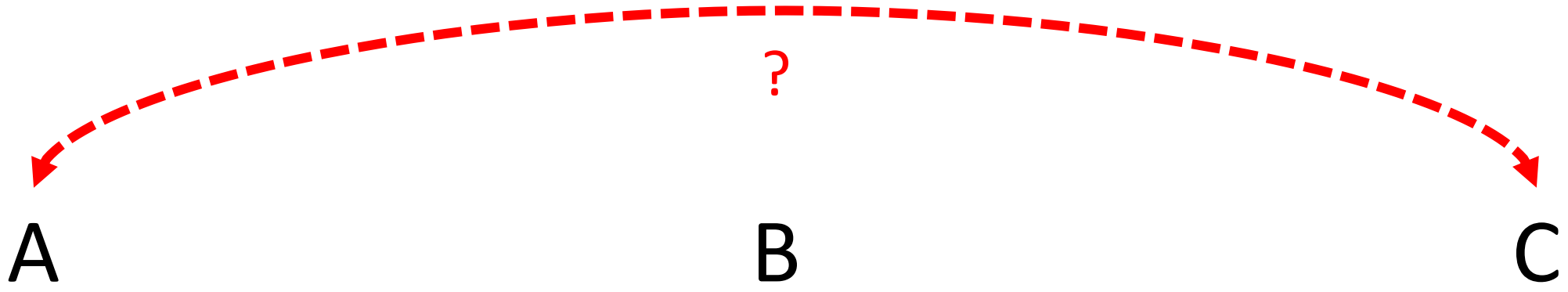
Implicit relations for hypothesis generation
ABC model



Swanson (1986) Fish oil, Raynaud's syndrome and undiscovered public knowledge
http://muse.jhu.edu/article/403510/pdf

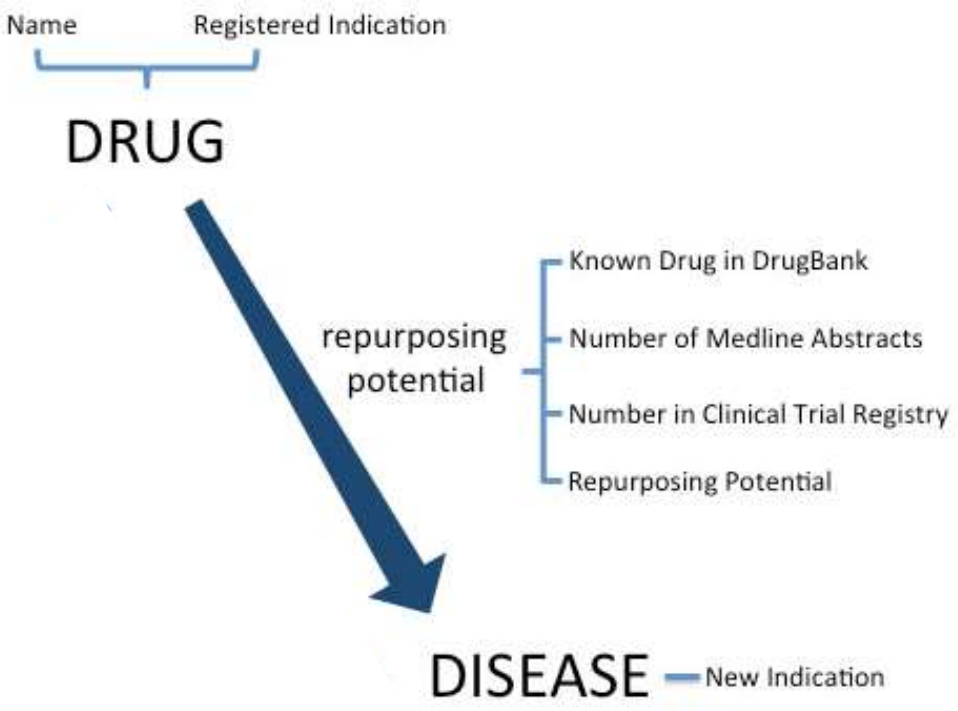
Open Discovery and Closed Discovery

- Open, you don’t know what C or B is (e.g. disease -> ?drug)
- Closed, you know what C is and are looking for B (e.g. disease – why? – drug)



Example question: drug repurposing

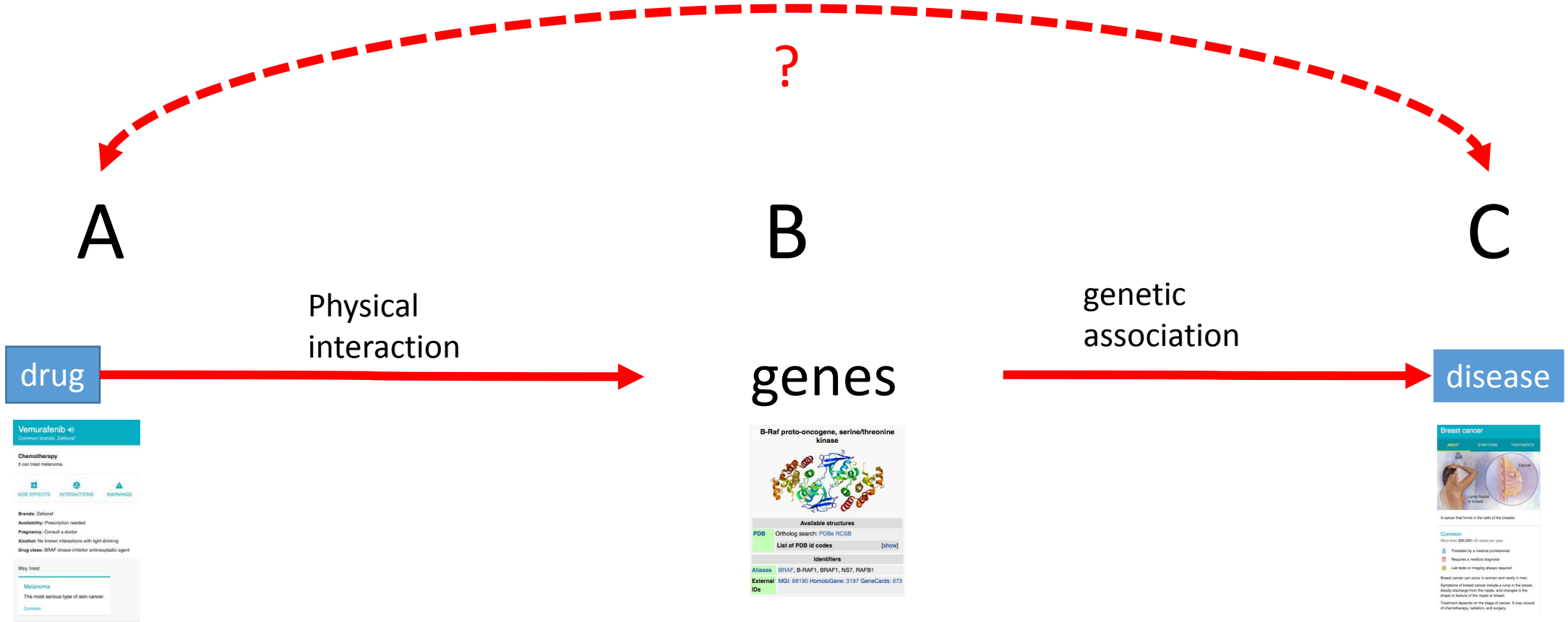
- For a given drug, what diseases might it be used to treat?



'RE:fine drugs': an interactive dashboard to access drug repurposing opportunities.

http://www.ncbi.nlm.nih.gov/pubmed/27189611

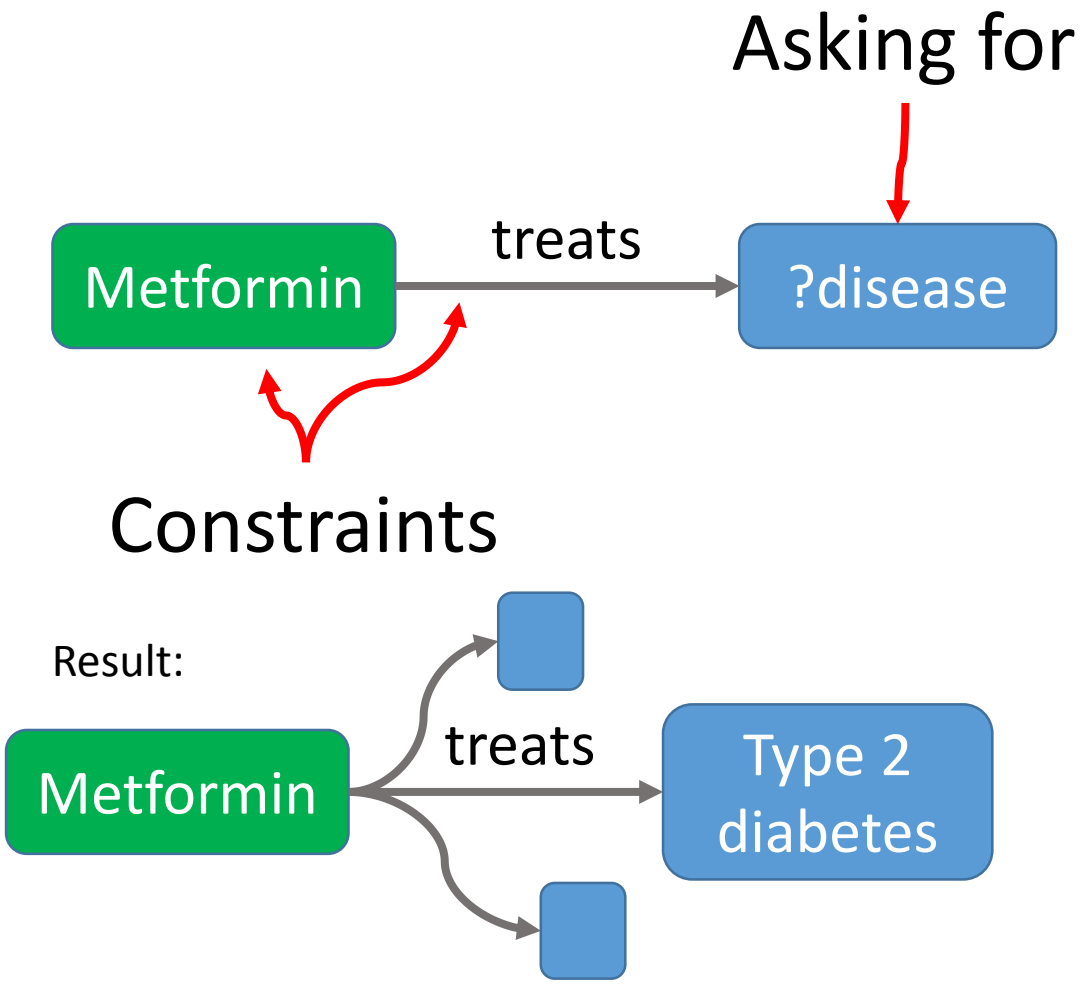
Implicit relations for hypothesis generation
ABC model for drug repurposing



Questions on ABC model ?

Querying a knowledge graph with SPARQL

- "SPARQL protocol and RDF query language"
- RDF: Resource Description Framework (common standard for storing knowledge graphs)
- **A SPARQL query = a partially completed graph**
  - **?'s show what you are looking for**
  - **rest constrains the search**



http://www.w3.org/TR/rdf-sparql-query/

Metformin's unique id: Q19484

Treats property id: P2175

```
#what disease does the drug treat ?
SELECT ?disease WHERE {
  wd:Q19484 wdt:P2175 ?disease .
}
```

=

disease
wd:Q3025883
wd:Q6717002

https://query.wikidata.org/

Metformin's unique id: Q19484
Treats property id: P2175

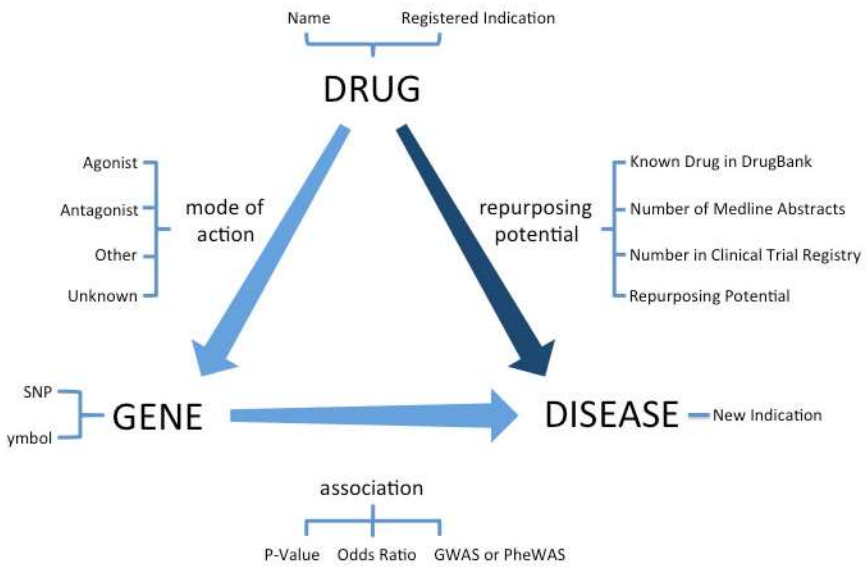
```
#what disease does the drug treat ?
SELECT ?disease ?diseaseLabel WHERE {
  wd:Q19484 wdt:P2175 ?disease .
  #add the labels
   SERVICE wikibase:label {
      bd:serviceParam wikibase:language "en" .
   }
}
```

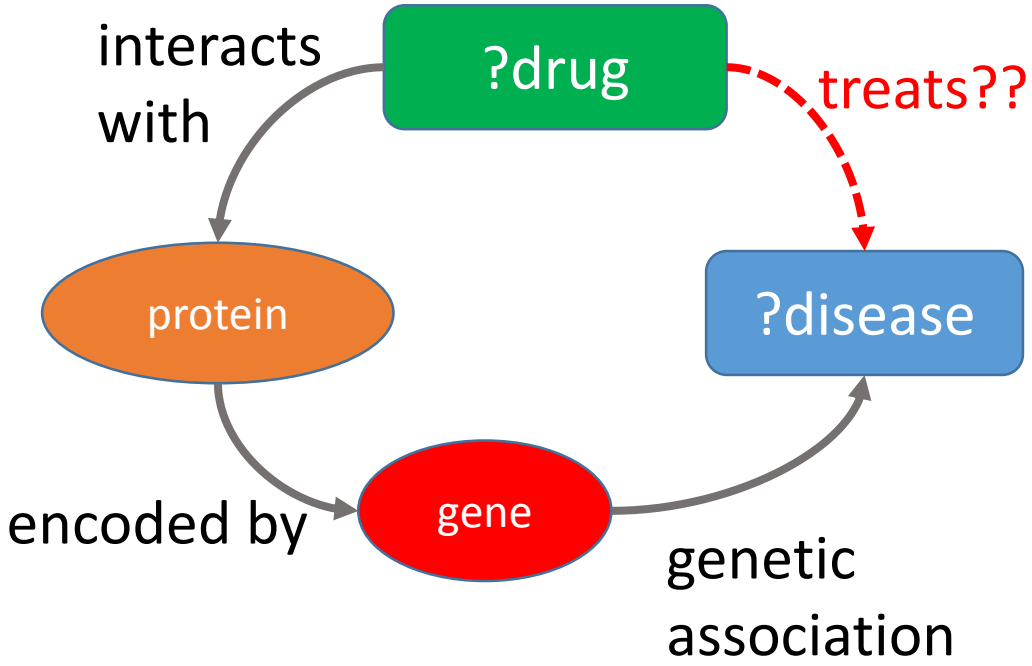
disease	diseaseLabel
wd:Q3025883	diabetes mellitus type 2
wd:Q6717002	MODY 2

http://tinyurl.com/gwd6pep

Example question: drug repurposing

"For a given drug, what diseases might it be used to treat?"

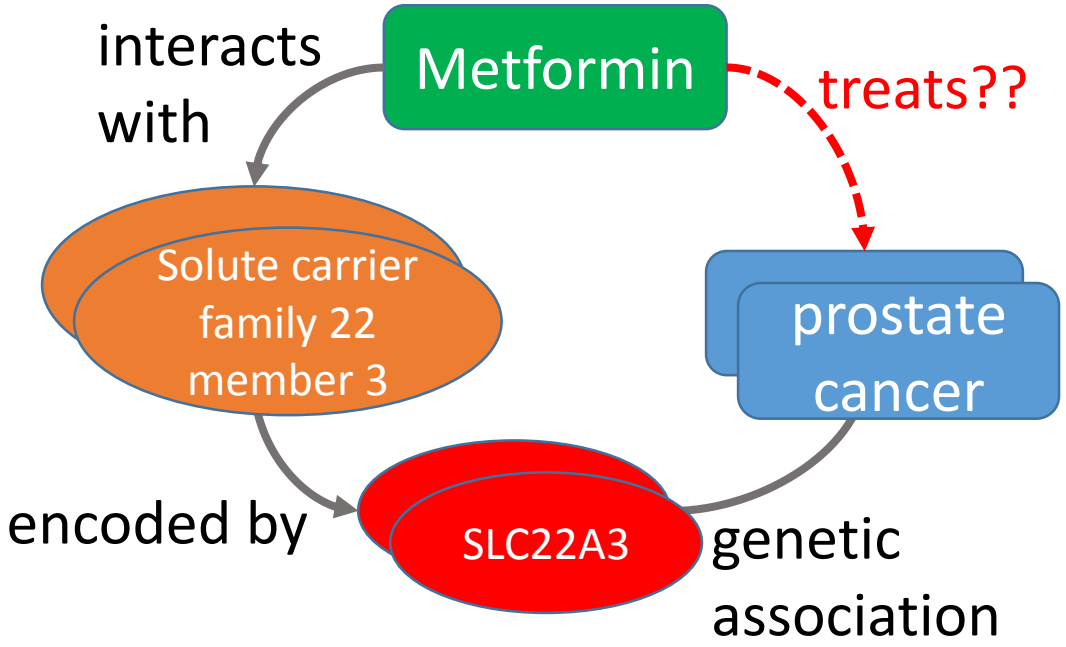




Example question: repurposing Metformin

```
SELECT ?gene ?geneLabel ?disease ?diseaseLabel WHERE {
  wd:Q19484 wdt:P129 ?gene_product .   # Metformin interacts with a gene_product
  ?gene_product wdt:P702 ?gene .  # gene_product is encoded by a gene
  ?gene wdt:P2293 ?disease .    # gene is genetically associated with a disease
  # add labels
   SERVICE wikibase:label {
      bd:serviceParam wikibase:language "en" .
   }
}
```

geneLabel	diseaseLabel
solute carrier family 22 (organic cation transporter), member 3	hepatitis C
solute carrier family 22 (organic cation transporter), member 3	prostate cancer
solute carrier family 22 (organic cation transporter), member 3	colorectal cancer
solute carrier family 22 (organic cation transporter), member 2	nephropathy



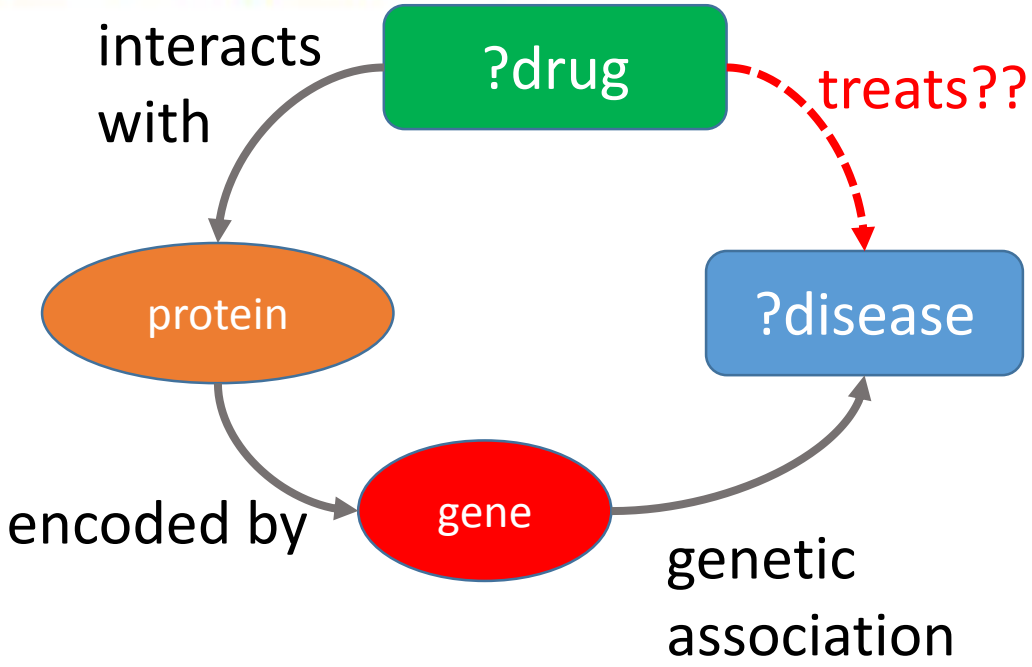
http://tinyurl.com/zem3oxz

Aside

- “Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality”
- https://jamia.oxfordjournals.org/content/22/1/179

Example question: repurposing all drugs

```
SELECT ?drug ?drugLabel ?gene ?geneLabel ?disease ?diseaseLabel WHERE {
  ?drug wdt:P129 ?gene_product .   # drug interacts with a gene_product
  ?gene_product wdt:P702 ?gene .  # gene_product is encoded by a gene
  ?gene wdt:P2293 ?disease .    # gene is genetically associated with a disease
  # add labels
   SERVICE wikibase:label {
      bd:serviceParam wikibase:language "en" .
   }
}
limit 1000
```



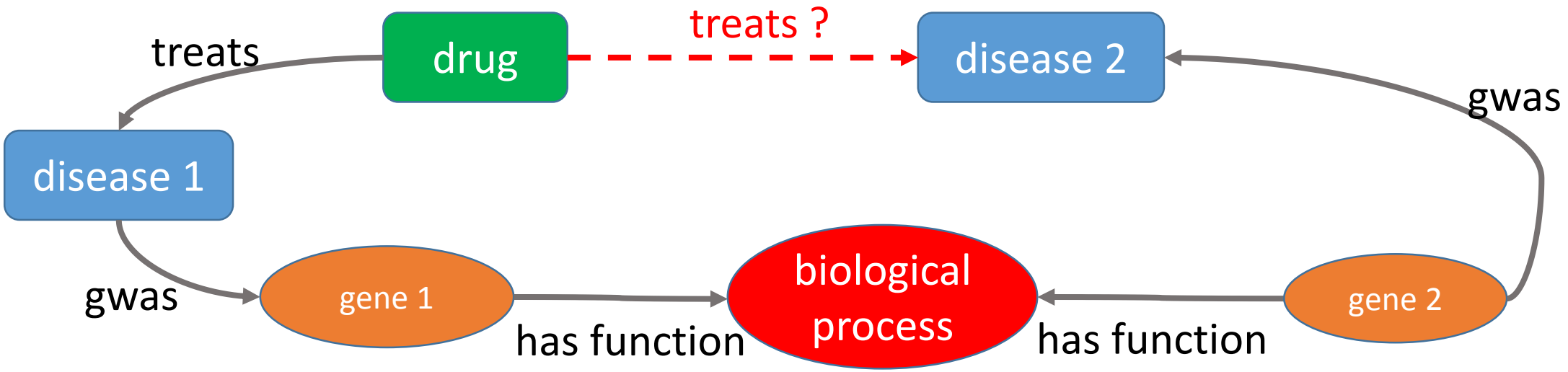
http://tinyurl.com/hwm9388

Adding constraints

- Find drugs that may treat disease
  - according to the drug->gene->disease model
  - constrained to focus on cancers
    - **?disease wdt:P279* wd:Q12078 .**
  - limited to genes related to cell proliferation
    - **?gene_product wdt:P682 ?biological_process**
    - **?biological_process wdt:P279* wd:Q14818032**
- http://tinyurl.com/j222k6g

Other patterns?

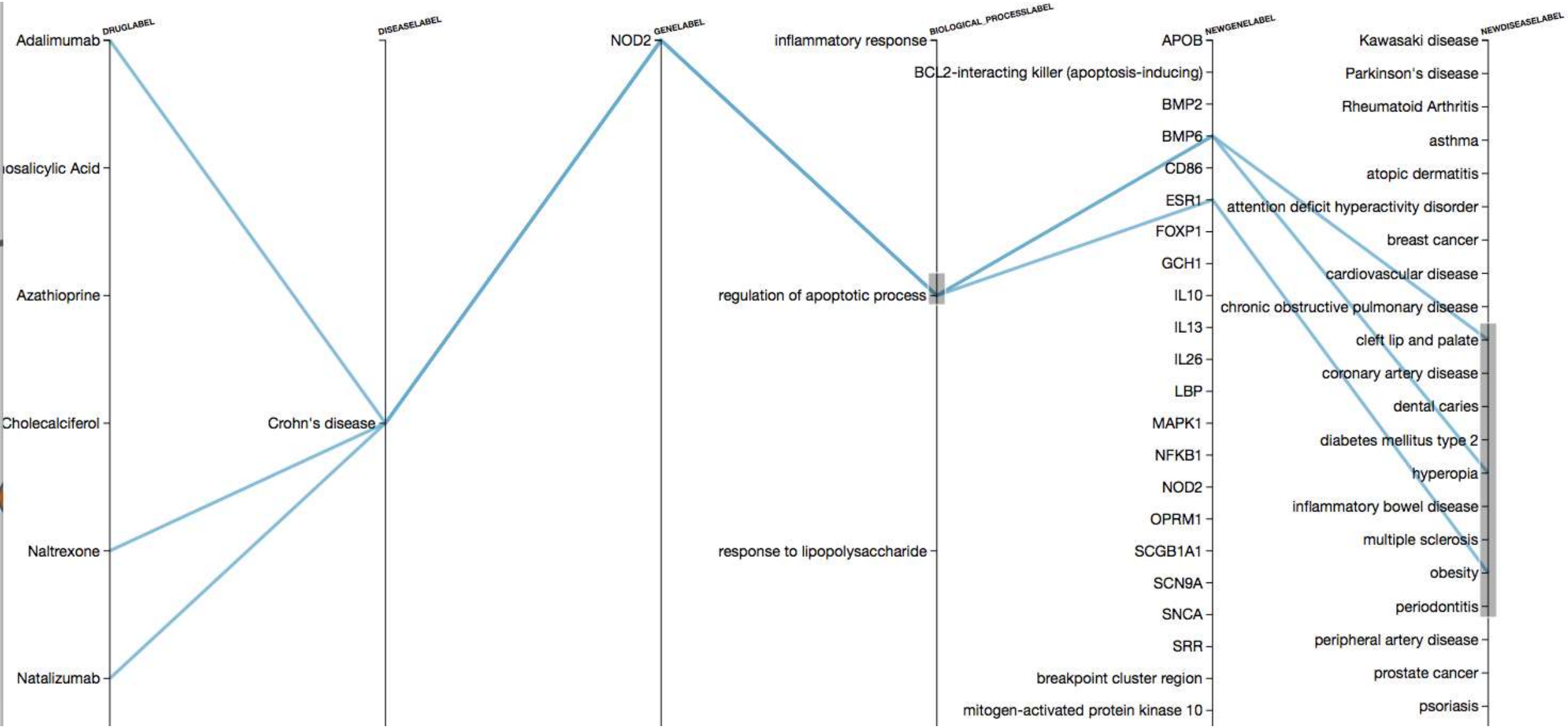
Is there a connecting path in the knowledge graph?
Is it meaningful?



http://tinyurl.com/gpfr9kj

Beta result viewer,
http://jonaskress.github.io/

http://tinyurl.com/jmoczaq



SPARQL endpoints of interest

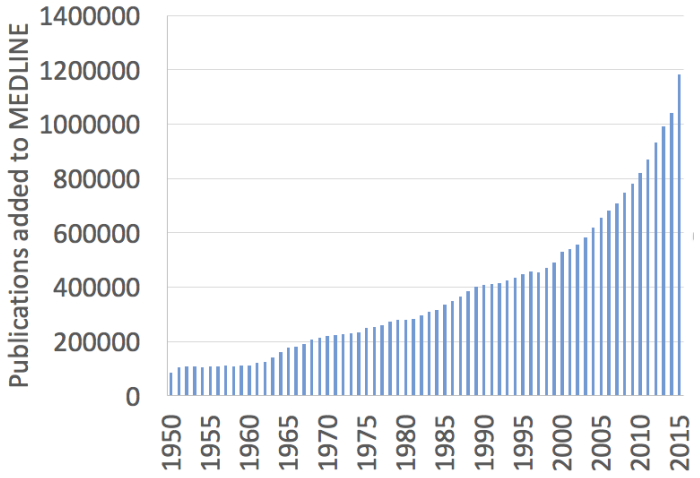
- Wikidata http://query.wikidata.org
- UniProt http://sparql.uniprot.org
- MeSH https://id.nlm.nih.gov/mesh/query
- EBI https://www.ebi.ac.uk/rdf/documentation/sparql-endpoints
- Bio2RDF https://github.com/bio2rdf/bio2rdf-scripts/wiki/Query-repository

2 problems with knowledge graphs

Not enough knowledge in the graph

text and data mining

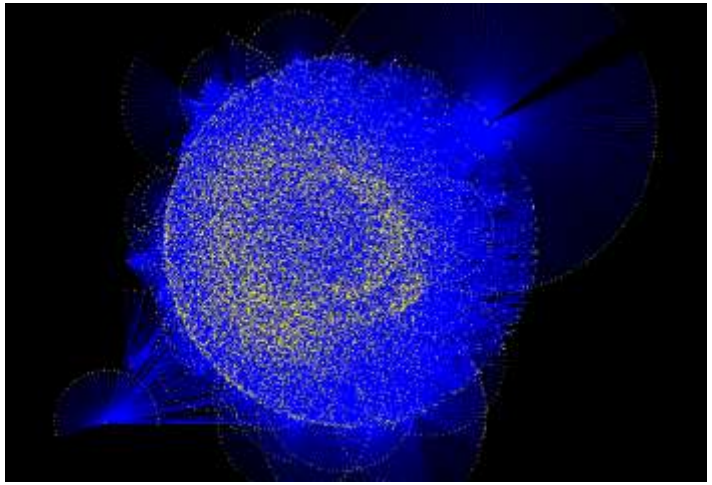
crowdsourcing ?



Too much knowledge in the graph

sorting algorithms

visualizations



http://i9606.blogspot.com/2010/05/gene-wiki-hairball-1.html

Plan for Thursday / Homework

- Implement and apply an ABC Model style hypothesis generating program
- Assignment: write the program, explain its logic, explain how you used it to generate a hypothesis, explain the hypothesis
- A Jupyter notebook with Python code will be provided to get you started
- If you do not want to program, there will be another option using online tools.

Suggested Reading

- Ontology
  - Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support
    - http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2592252/
  - Gene Ontology: tool for the unification of biology
    - http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037419/
- Knowledge-based hypothesis generation
  - Fish oil, Raynaud's syndrome and undiscovered public knowledge
    - http://muse.jhu.edu/article/403510/pdf
  - Knowledge discovery by automated identification and ranking of implicit relationships
    - http://bioinformatics.oxfordjournals.org/content/20/3/389.full.pdf
- Text mining
  - Literature mining for the biologist: from information retrieval to biological discovery
    - http://www.nature.com/nrg/journal/v7/n2/full/nrg1768.html