The Neuroscience Information Framework:

A Scalable Platform for Information Exploration and Semantic Search Computing

Amarnath Gupta

SDSC

UCSD

NIF

NEUROSCIENCE INFORMATION FRAMEWORK

- NIF is an initiative of the NIH Blueprint consortium of institutes
  - What types of resources (data, tools, materials, services) are available to the neuroscience community?
  - How many are there?
  - What domains do they cover? What domains do they not cover?
  - Where are they?
    - Web sites
    - Databases
    - Literature
    - Supplementary material
    - **PDF files**
    - **Desk drawers**
  - Who uses them?
  - Who creates them?
  - How can we find them?
  - How can we make them better in the future?

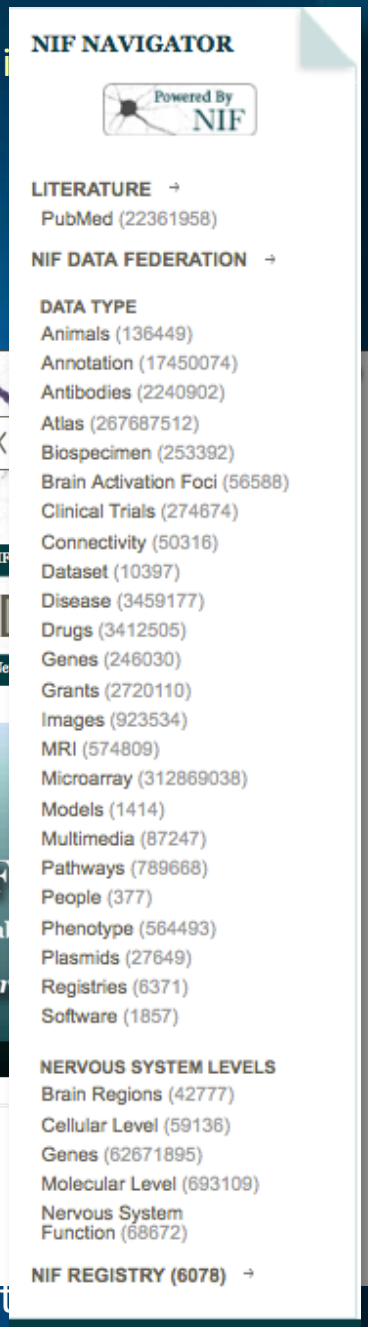
Discovery and utilization of web-based resources for neuroscience

UCSD, Yale, Cal Tech, George Mason, Washington Uni

Literature

Database Federation

Registry



A portal for finding and using neuroscience resources

A consistent framework for describing resources

Provides simultaneous search of multiple types of information, organized by category

Supported by an expansive ontology for neuroscience

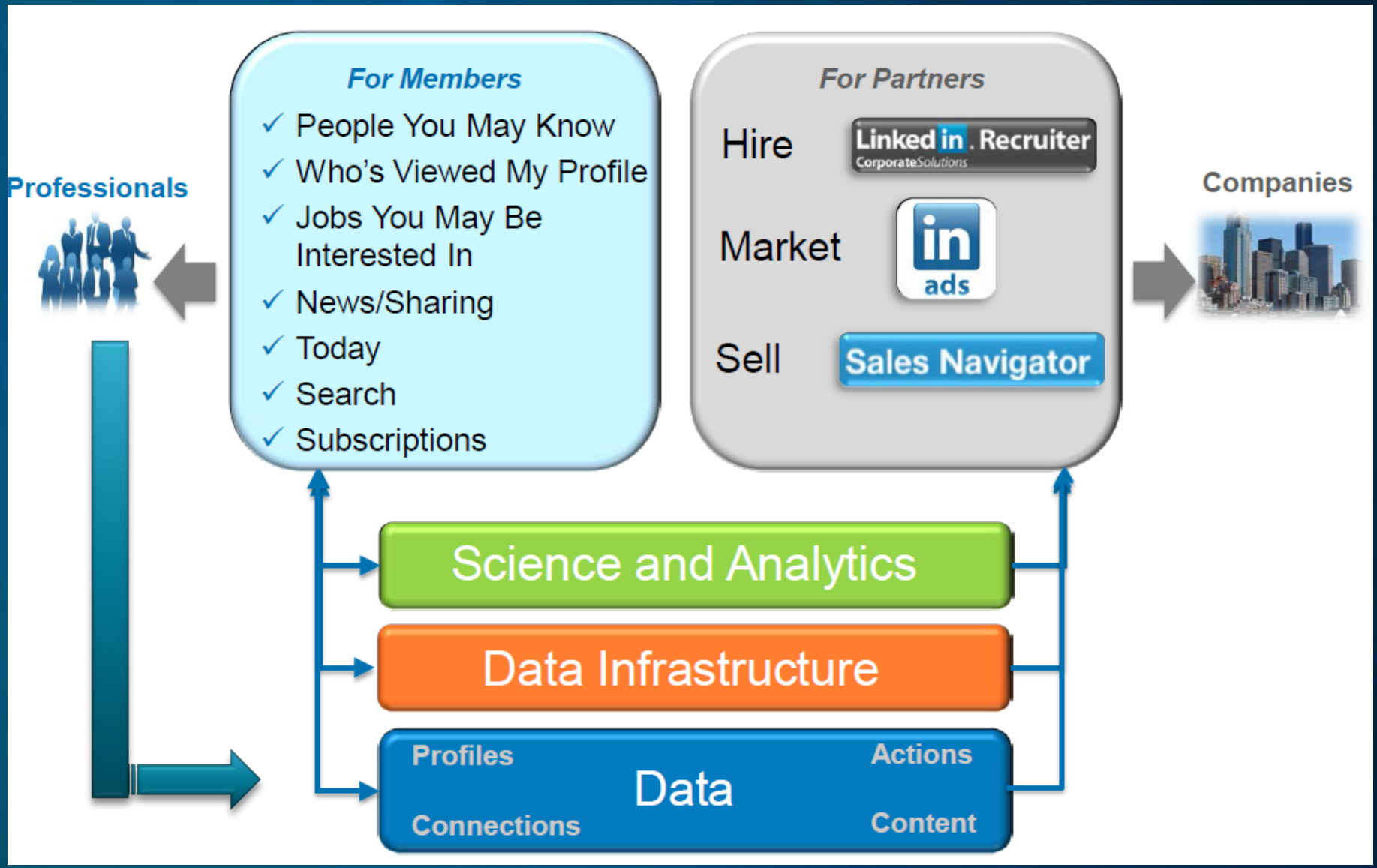
Utilizes advanced technologies to search the “hidden web”

A Big, Fat Techno-Philosophical Question

- *How do we create an information infrastructure that is able to connect a person or a community with the resources they need to accomplish their task at hand?*
- Resource
  - Anything that is tangible and accessible
    - a product, a person, an institution, a piece of data, a connection …
- Information Infrastructure
  - Enables the entire life cycle of information from acquisition to (potential)archival
  - Allows people to find, access, understand and work with information
- A domain-specific example:

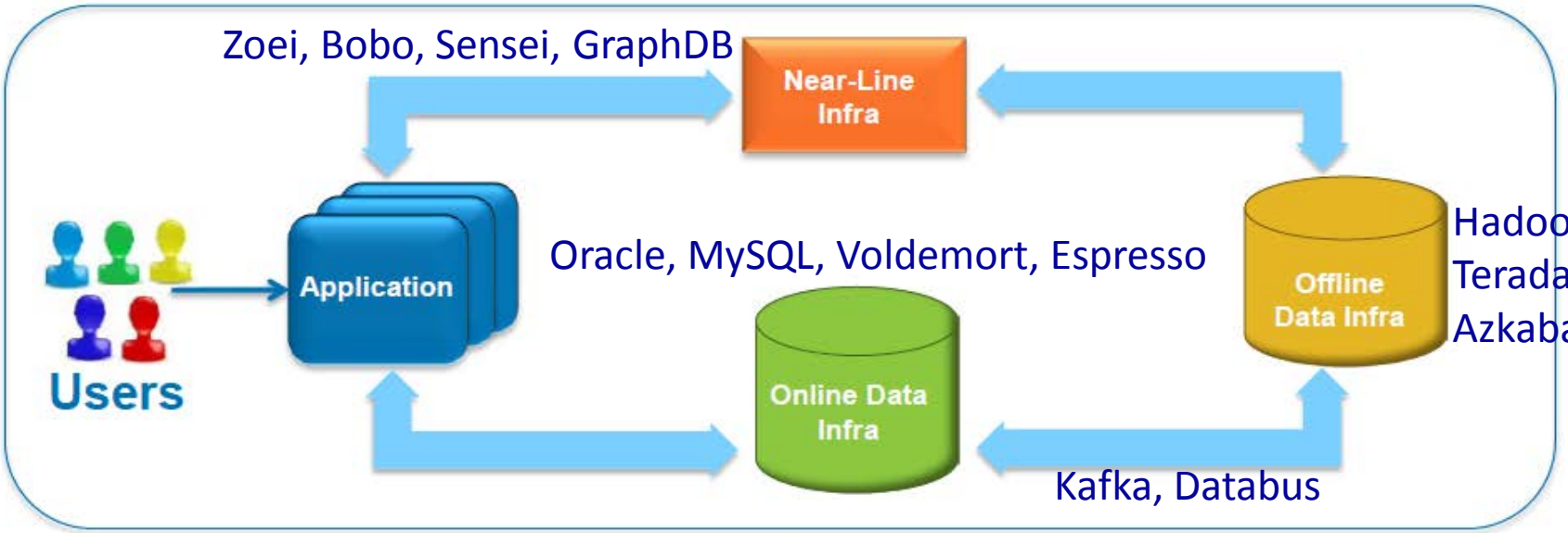
A Digression – Looking at LinkedIn

Credit: Bhaskar Ghosh, LinkedIn



A Digression – Looking at LinkedIn

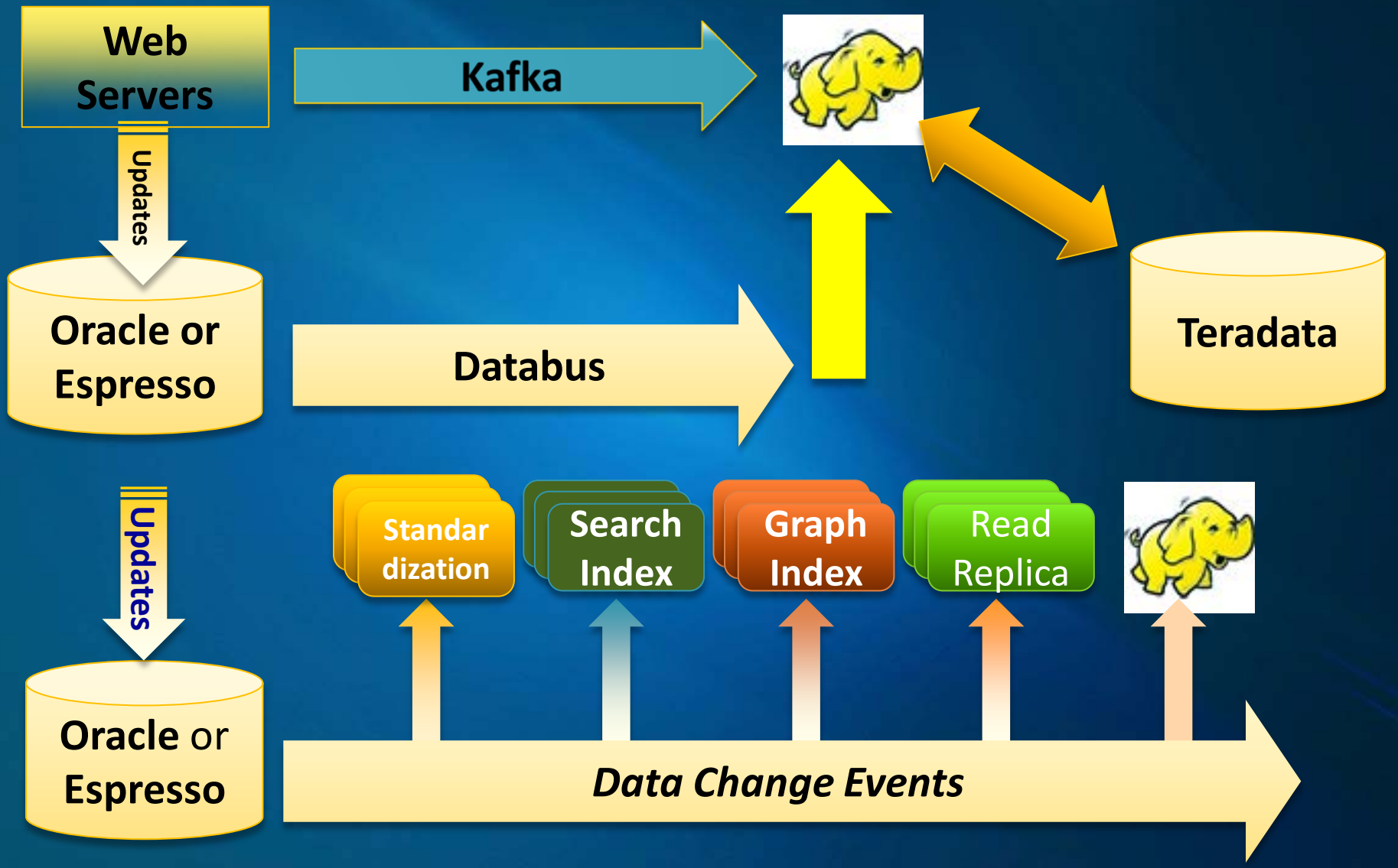
Credit: Bhaskar Ghosh, LinkedIn



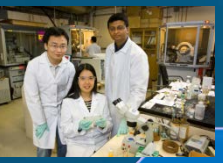
Infrastructure	Latency & Freshness Requirements	Products	
Online	Activity that should be reflected **immediately**	• Member Profiles • Company Profiles • Connections	• Messages • Endorsements • Skills
Near-Line	Activity that should be reflected **soon**	• Activity Streams • Profile Standardization • News	• Recommendations • Search • Messages
Offline	Activity that can be reflected **later**	• People You May Know • Connection Strength • News	• Recommendations • Next best idea…

A Digression – Looking at LinkedIn

Credit: Hien Luu, Sid Anand, LinkedIn



Resource Finding in Biomedical Sciences

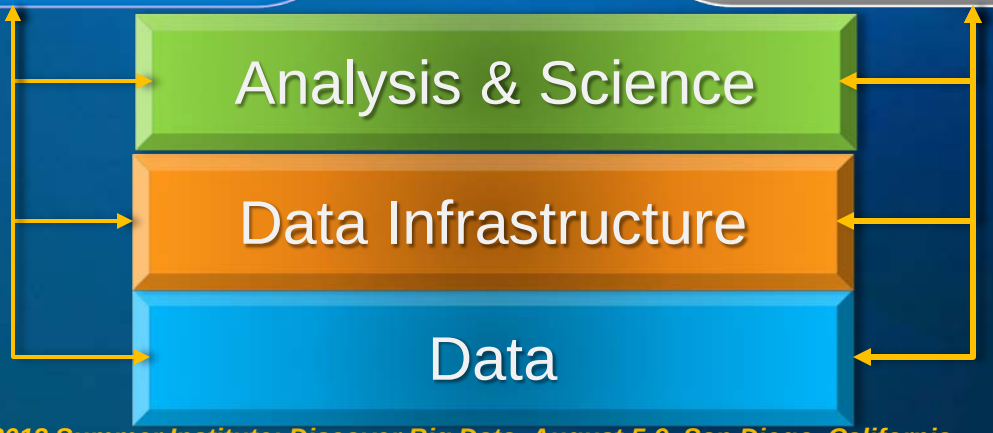


Researcher Activity

- ✓ Where is data about X?
- ✓ How does Y relate to Z?
- ✓ Accumulate and Analyze
- ✓ Compare X and Y
- ✓ Subscribe to topic T
- ✓ Recommend Resource
- ✓ Funding reports
- ✓ Search and Explore
- ✓ News

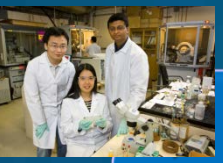
Resource Activity

- ✓ Resource Promotion
- ✓ Utilization Search
- ✓ Cross-Utilization
- ✓ Experiential Services



And yet, biomedical resource finding is a hard problem

Resource Finding in Biomedical Sciences

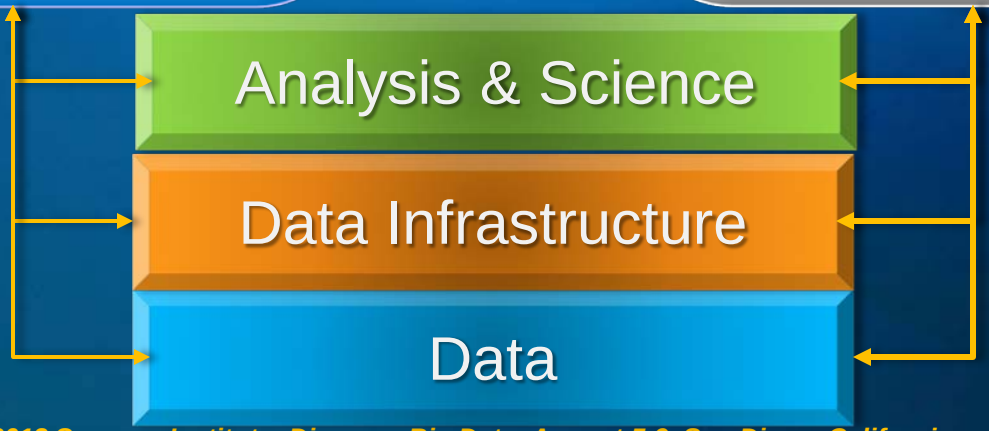


Researcher Activity

Where is data about X?
- ✓ How does Y relate to Z?
- ✓ Accumulate and Analyze
- ✓ Compare X and Y
- ✓ Subscribe to topic T
- ✓ Recommend Resource
- ✓ Funding reports
- ✓ Search and Explore
- ✓ News

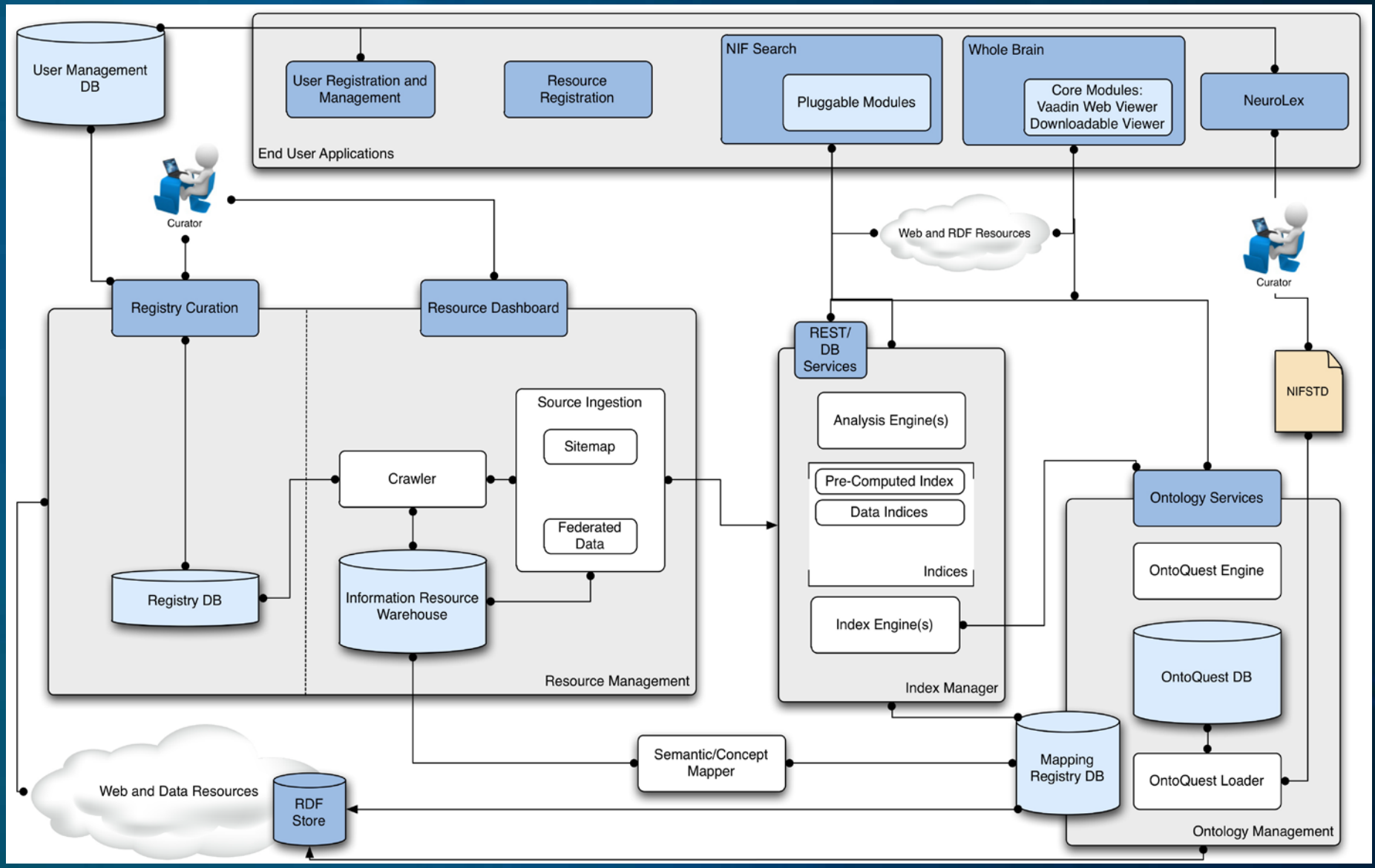
Resource Activity

- ✓ Resource Promotion
- ✓ Utilization Search
- ✓ Cross-Utilization
- ✓ Experiential Services



The problem starts here ?

NIF Architecture



The "Resource Identification" Problem

- *Given an infinite number of web accessible resources, which are relevant for Neuroscience?*
- Easy Case
  - A resource assigned by a trusted source
- Reasonably Easy Case
  - A resource recommended by a potentially trustable source
- Not-so-easy Case
  - "Mine" for resources from literature/crawler
  - Auto-filter by semantic classification
  - Fully validate by curatorial staff/community

Semantic Classification

- Is this a potential neuroscience resource?
  - Two-pronged classification problem
    - If it belongs to the class, a reasonable portion of the document term vector will align with a neuroscience vocabulary
      - Necessary but not sufficient
    - The “spread” of the document term vector with respect to a reasonable domain ontology will be limited
  - Pragmatic problems
    - What additional (recognizable) descriptors does the resource have?
    - Is the resource “current”?
    - Are there “other ways” of getting to the content of the site?

Ingestion and Transformation Provenance

- DISCO – NIF's Ingestion Manager and Data Tracker
  - "Relationalizes" incoming data when needed
  - Feeds the curators' dashboard for ingestion, update and index management
  - Executes automatic updates per schedule
  - Keeps track of chains of derived views defined by curators
  - Maintains annotations on data and its views
  - Propagates data updates to all derived views through curator notification

The "Information Variety" Problem

- The data come
  - From too many disparate sources
    - 6000+ neuroscience resources
  - In too many different formats and models
    - Relational, XML, RDF, Text, domain-specific, …
  - Having all too diverse semantics
    - "GRM1": a string, a gene, a chromosomal region, a list of interesting SNPs in mice?
- There is a massive *data integration problem* because only integration of data will lead to insight
  - *What possible drugs might be repurposed for human inclusion body myopathy (HIBM)?*
  - Data about/from the following to be integrated
    - Organisms, diseases, cross-organism anatomy, phenotypes, genes, proteins, interactions, pathways, genomic variations, pharmaceutical compounds, assays and publications

Hybrid Integration Strategy

- Data integration using schema mappings for similar resources
- Semantics-based integration
  - Using ontologies as the unifying structure
  - Using vocabularies as the connecting substrate
- Using linked-data graph where applicable
  - Link inference
  - Link prediction

  } When possible
- Using machine learning
  - Term association using active learning with conditional random fields

Integration using Schema Mapping

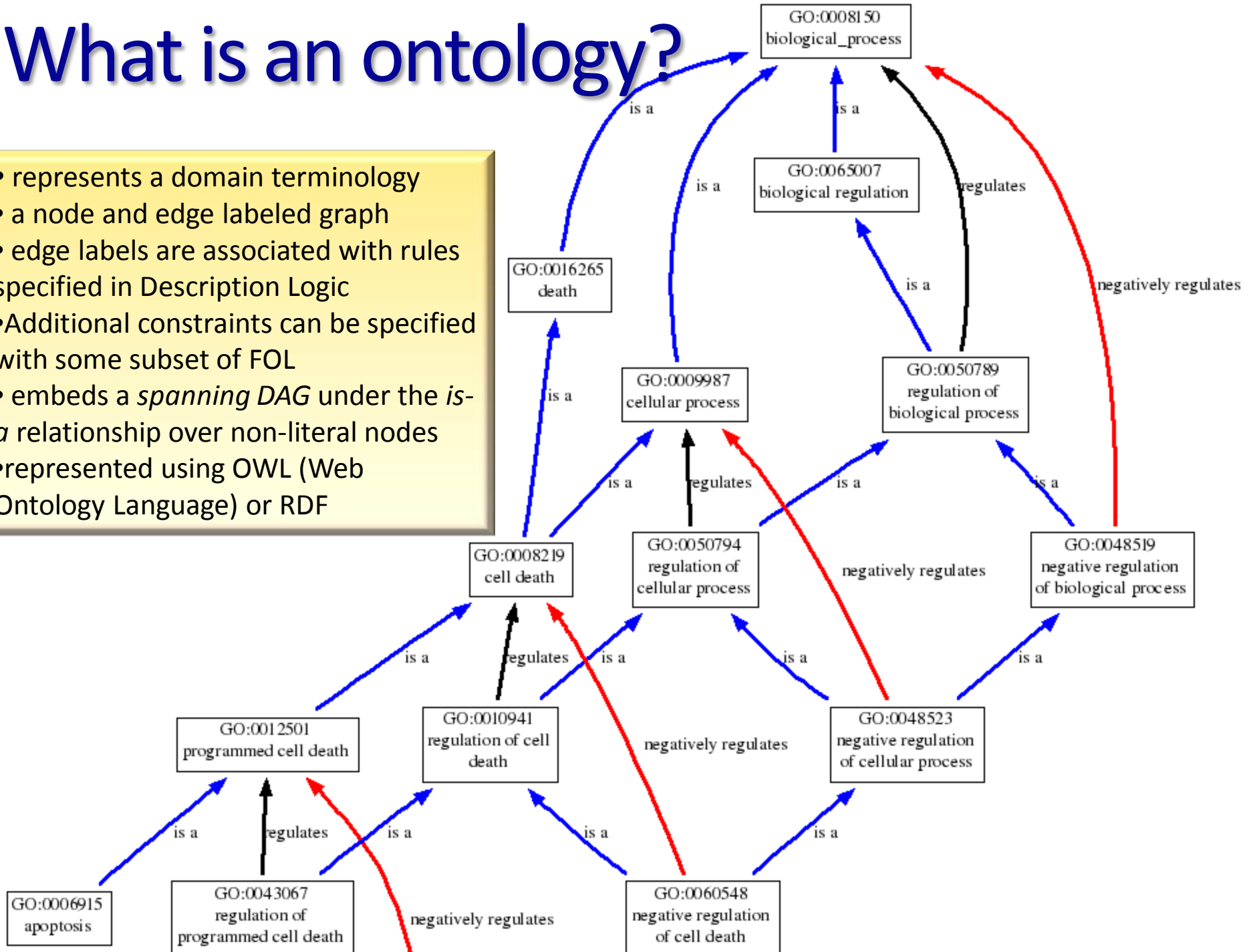
- Commercial solutions exist
  - They are not very scalable as number of schemas increase
- Many groups working on it
  - Data Tamer – an MIT-Intel partnership project on Big Data
    - Matches schema where possible
    - Crowd-sources ambiguous cases
    - Performs entity-consolidation by data clustering
- Our approach
  - Curators use small scale schema mappings using ontology and Google Refine

Know Thy Landscape – Semantic Scouting

- Gather the language of the domain
  - Terms, their variations and term properties
  - “Verbs”, relationships, their linguistic variations and their properties
  - Constraints that hold in the domain
- Sources
  - Ontologies
  - Folksonomies
    - image tags, text annotations, …
  - Literature
    - figure and table captions
  - Data (structured or semi-structured)

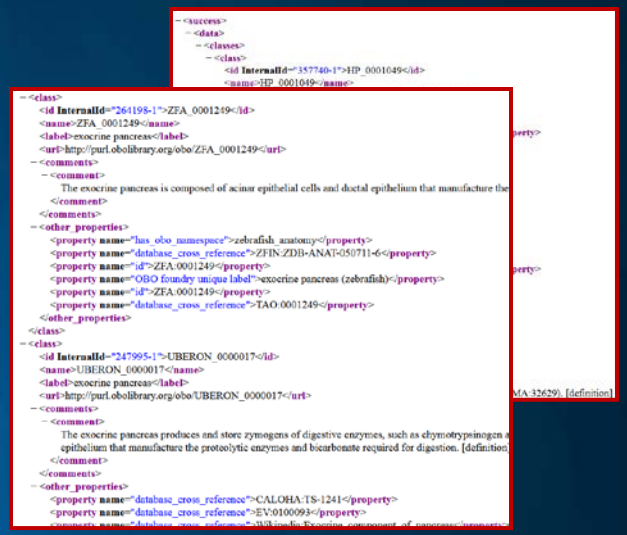
What is an ontology?

- represents a domain terminology
- a node and edge labeled graph
- edge labels are associated with rules specified in Description Logic
- Additional constraints can be specified with some subset of FOL
- embeds a *spanning DAG* under the *is-a* relationship over non-literal nodes
- represented using OWL (Web Ontology Language) or RDF



Ontology Processing with OntoQuest

- A graph query engine for ontological graphs
  - Accepts OWL, RDFS, RDF
  - Ingestors for special cases can be developed
    - MESH XML DTD
  - Has a service API

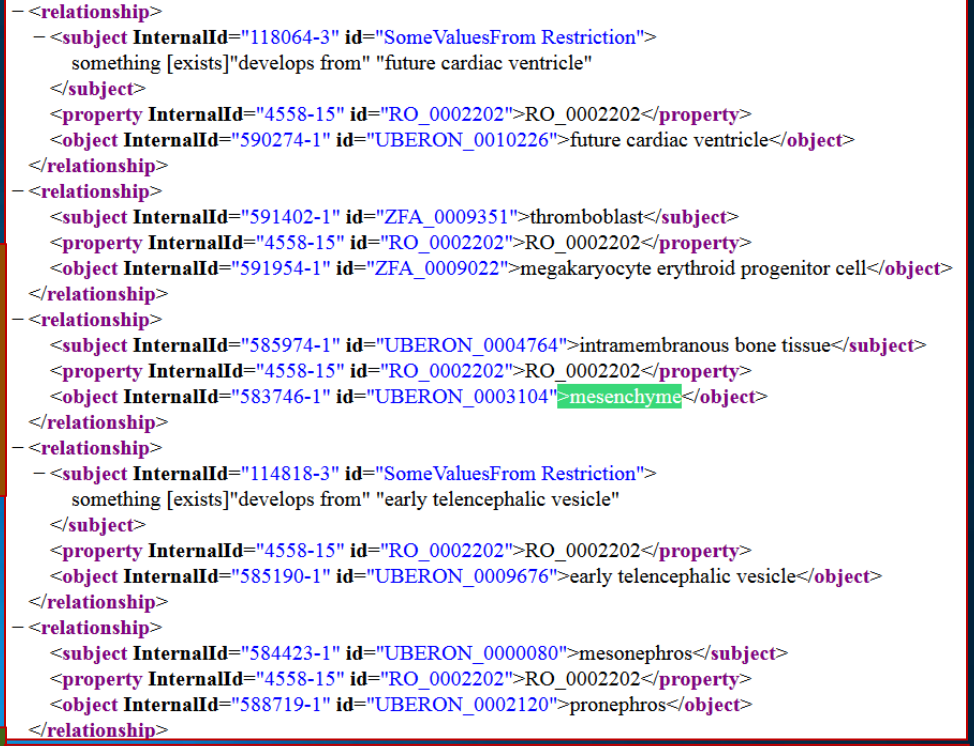


1. Get all human phenotypes
http://nif-services-stage.neuinfo.org/ontoquest-lamhdi/concepts/search/HP_
2. Find superclasses of "exocrine pancreas"
First: http://nif-services-stage.neuinfo.org/ontoquest-lamhdi/concepts/term/exocrine+pancreas
Then, from the result of the above: http://nif-services-stage.neuinfo.org/ontoquest-lamhdi/rel/superclasses/UBERON_0000017
3. Find all direct properties of UBERON_0000017
http://nif-services-stage.neuinfo.org/ontoquest-lamhdi/rel/children/UBERON_0000017
COMBINED WITH
http://nif-services-stage.neuinfo.org/ontoquest-lamhdi/rel/parents/UBERON_0000017

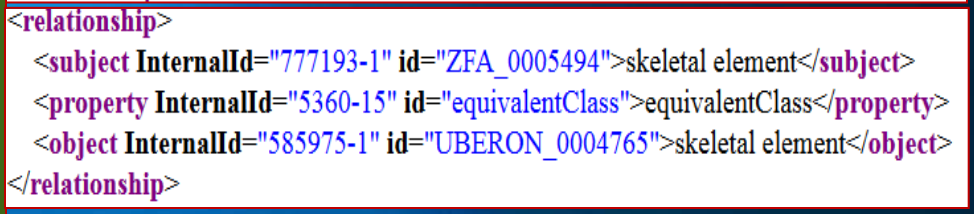
Computing with OntoQuest

- Which skeletal structures in the zebrafish develop from the mesenchyme?
  - Return $x where
    - ($x (subclassOf)* ‘skeletal element’) and
    - ($x develops_from* ‘ZFA:mesenchyme’)
    - Return $y where
      - ($y subclassOf* $x)
  - Query Rewriting
  - Return $x where
    - ($x develops_from* `mesenchyme’) and
    - ($x has_ontology $o) ($x equivalent_to $z) ($z has_ontology ‘ZFA’) and
    - ($x (subclassOf)* ‘skeletal element’)

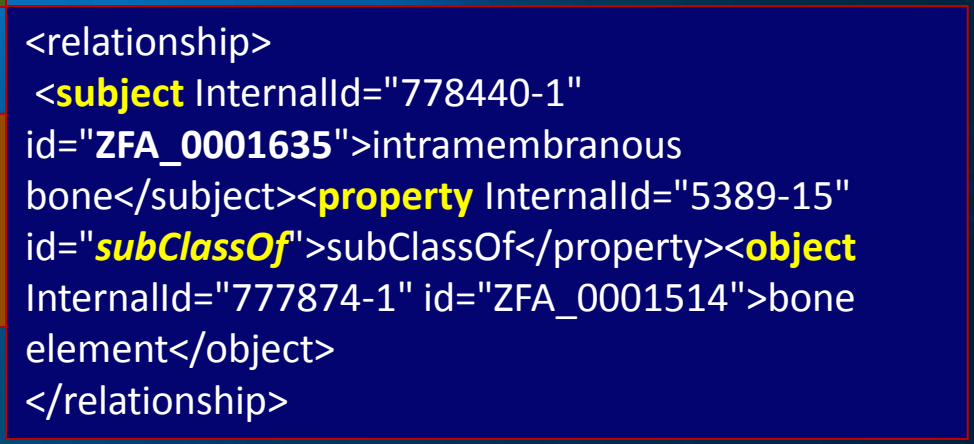
“develops_from”
http://nif-services-stage.neuinfo.org/ontoquest-lamhdi/rel/edge-relation/id/RO_0002202



“equivalenceClass”
http://nif-services-stage.neuinfo.org/ontoquest-lamhdi/rel/children/term/skeletal%20element?level=1



Subclasses of “skeletal element” (incl. its equivalenceClasses) in ZFA
http://nif-services-stage.neuinfo.org/ontoquest-lamhdi/rel/subclasses/term/skeletal%20element?level=3



The Entity Recognition Problem

- Tagging of Antibody Records using a machine learning technique with Conditional Random Fields

A first-order linear-chain CRF

$$p_\Lambda(\mathbf{y}|\mathbf{x}) = \frac{1}{\mathbf{Z}(\mathbf{x})} \exp\left(\sum_{\mathbf{t}=1}^{\mathbf{T}} \sum_{\mathbf{k}=1}^{\mathbf{K}} \lambda_\mathbf{k} \mathbf{f_k}(\mathbf{y_{t-1}}, \mathbf{y_t}, \mathbf{x}, \mathbf{t})\right)$$

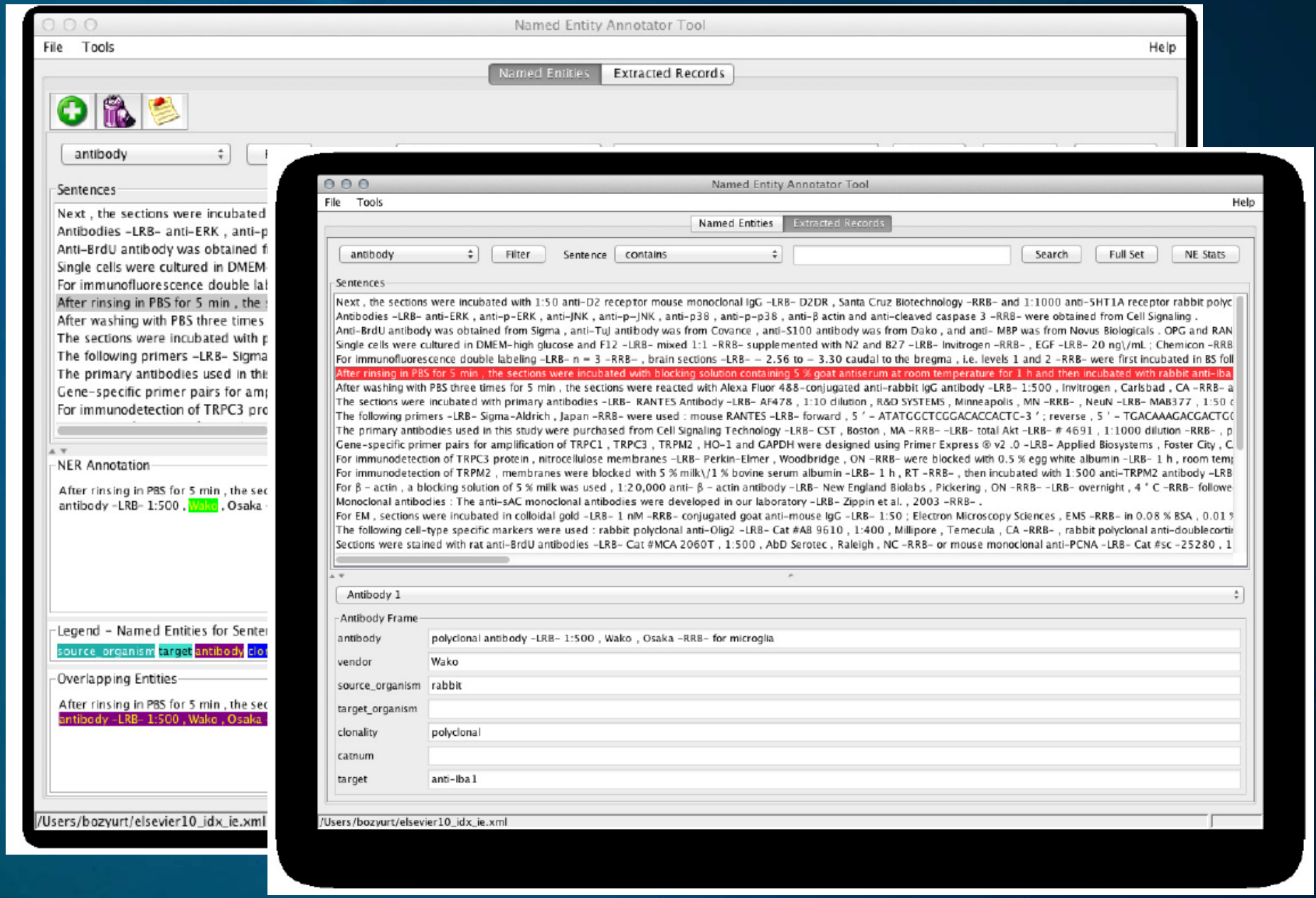
where

- **y** is output ve
- **x** input data.
- $f_k(\mathbf{y_{t-1}}, \mathbf{y_t}, \mathbf{x},$ output(IOB la
- λ_k weight of t maximum like
- $Z(\mathbf{x})$ normaliz
- t index to the

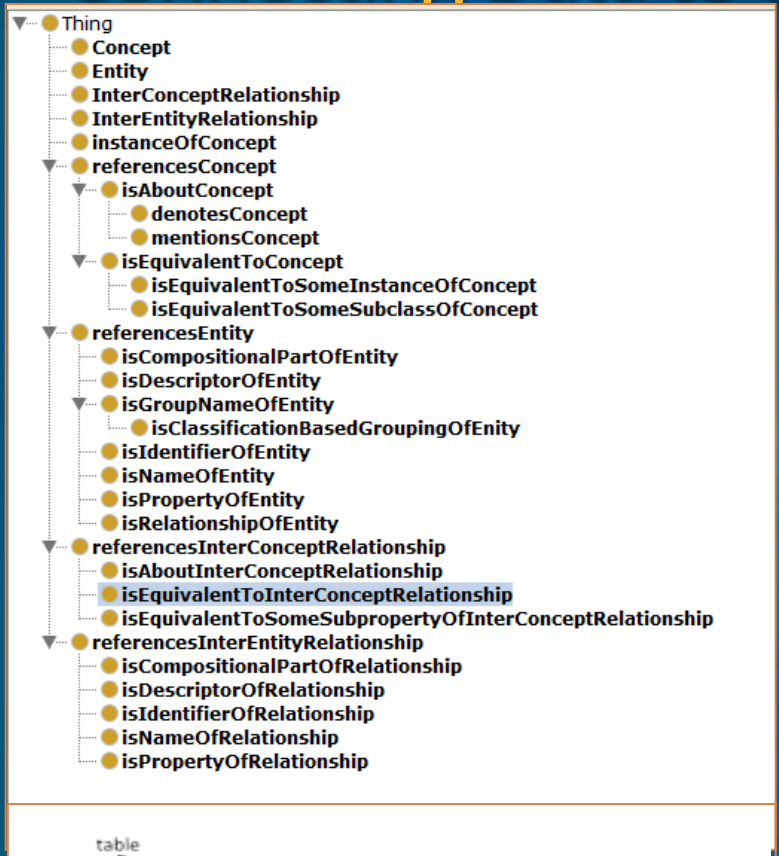
```
Expression|O:{NN,NPC} of|O:{IN}
glutamate|B-PROTEIN:{NN,NPC}
carboxypeptidase|I-PROTEIN:{VBP} II|I-PROTEIN:{NNP,NPC}
in|O:{IN} human|O:{JJ} brain|O:{NN,NPC} .|O:{.}
```

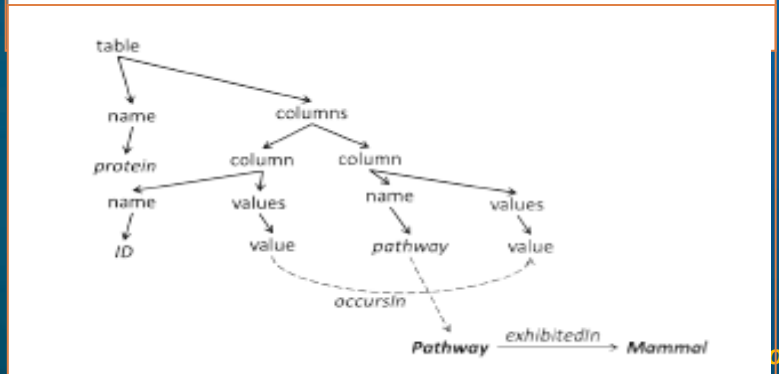
- An example sentence labeled via extended IOB format. The prefix `B` denotes the beginning on a named entity.
- The rest of terms in the NE are denoted by the prefix `I` and terms not belonging to any NE by `O`.
- IOB format is extended to include additional information, namely, part-of-speech tags and noun phrase membership.

The Learning Tool and Minimal Models



The Mapping Problem –
Connecting Data to the Ontology

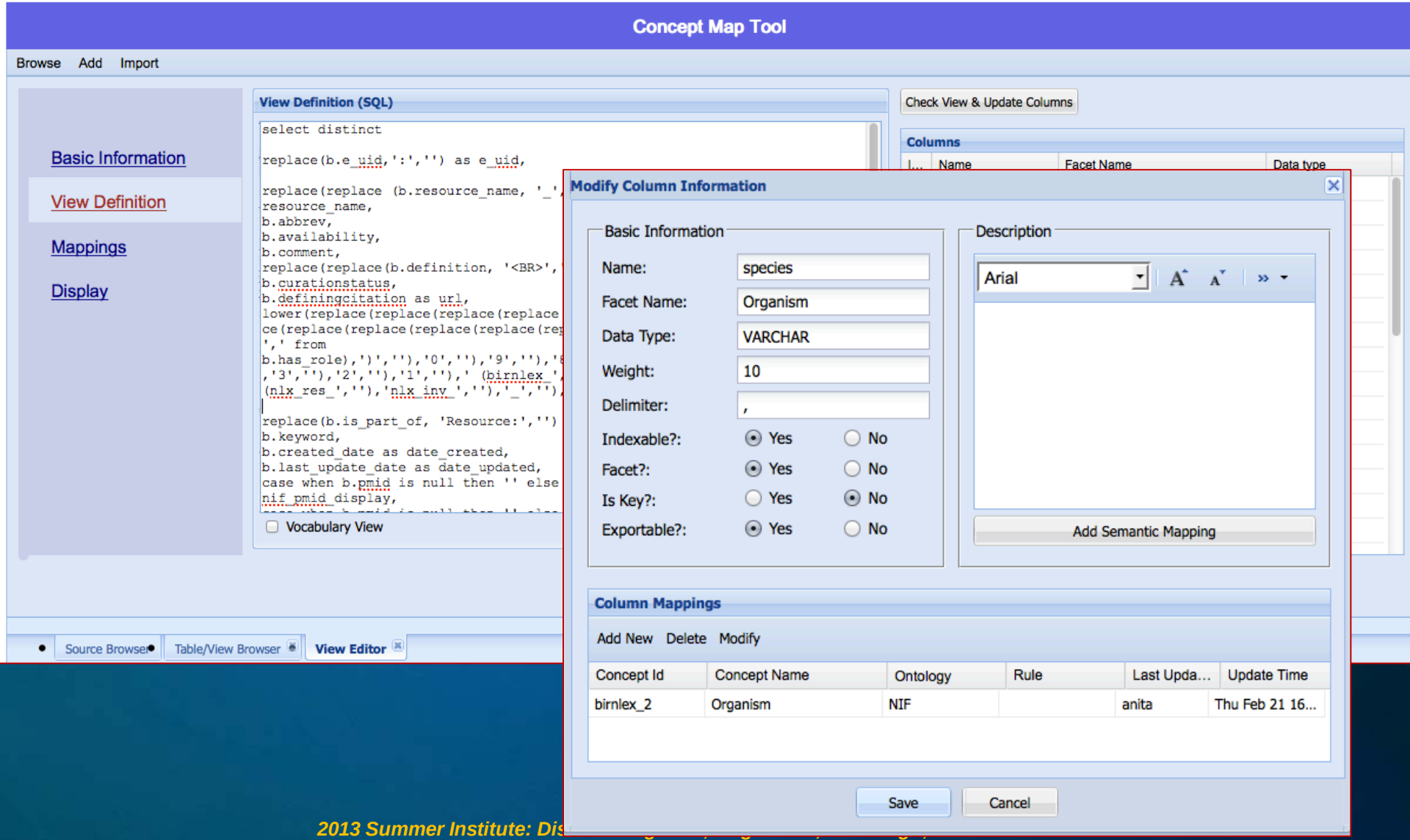




- A high-level statement
  - The ontology O is a graph of concepts and inter-concept relationships
  - The data is a semistructured object S with groupings
  - A mapping structure is a graph of mapping edges from S to O + intra-source map edges + intra-ontology map edges

Table: protein(ID, symbol, name, pathway)
Mapping:
pathway mapsTo ontoID(Pathway).
(exhibitedIn(ontoID(Pathway), ontoID(mammal))
AND occursIn(value(ID), value(pathway)))

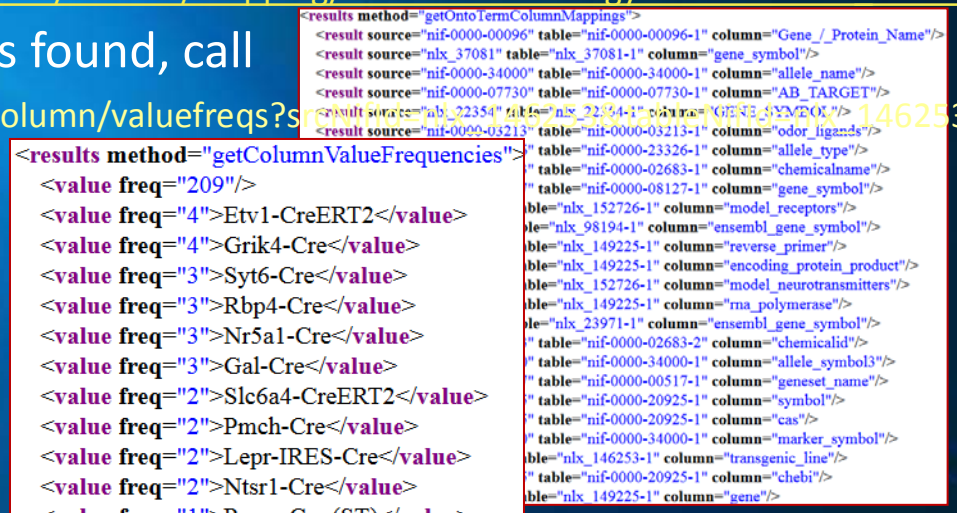
The Concept Mapping Tool



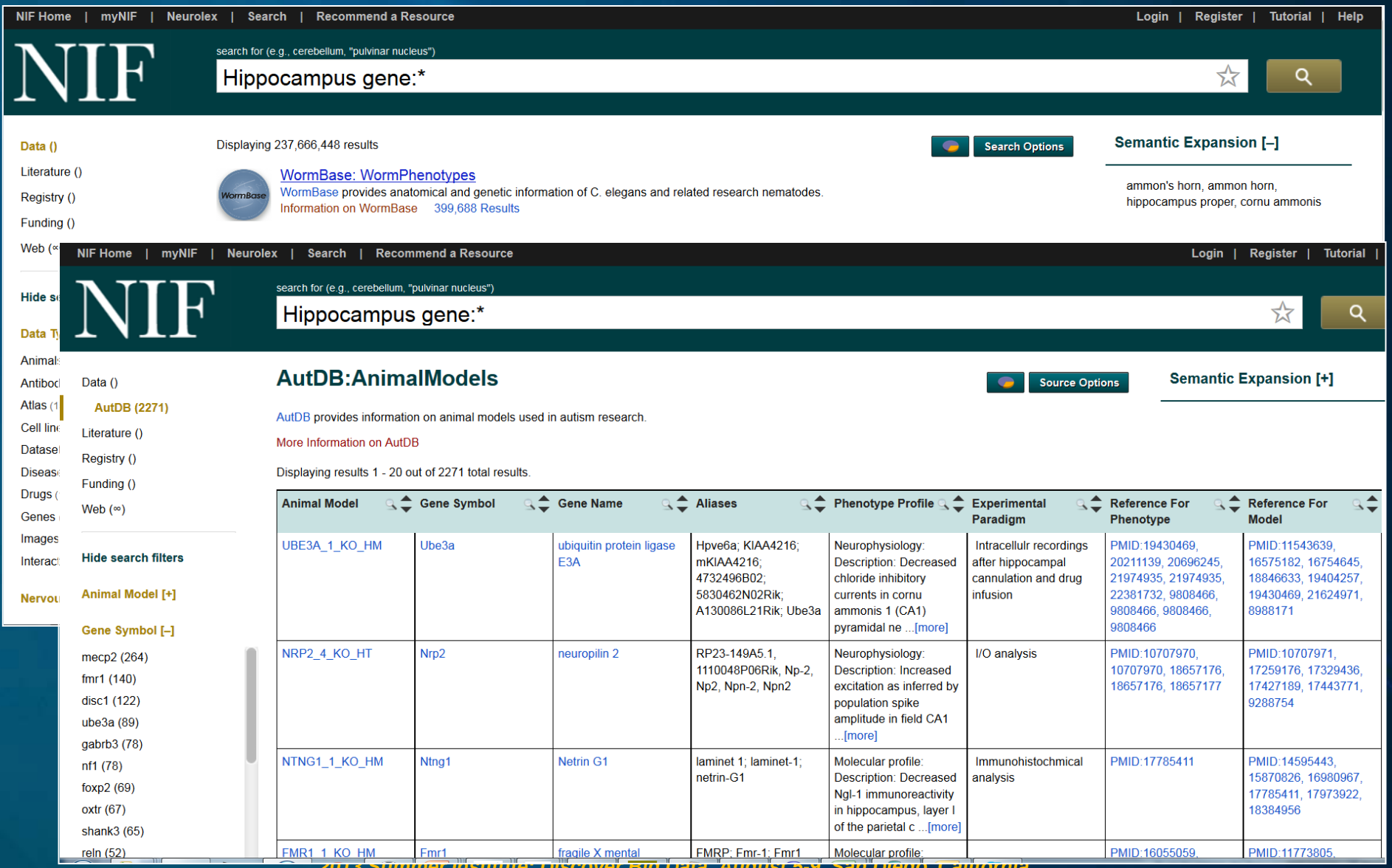
Querying the Mapping Structure

- What are the top 10 genes you have most information about?
  - Which sources have information about "genes"?
    - Find the ontology term for gene → CHEBI_23367
    - Call http://cm.neuinfo.org:8080/cm_services/column/mapping/ontoterm?ontologyTermId=CHEBI_23367
    - For each table and column thus found, call

      http://cm.neuinfo.org:8080/cm_services/column/valuefreqs?s...1&columnName=transgenic_line
    - Now group and rank merge
      - in parallel if needed
    - Get top 10



Search: The Keyword Query Interface



SKEYQL – extending a Keyword QL

- Lucene Query

+

Feature	Description	Example
Boolean search	in addition to search with explicit AND, OR, NOT, a query can specify terms to be included (with a +) and terms to be excluded	The query `+(neuron protein) -gene` searches for documents `neuron` and `protein` and not with `gene`
Fielded search	A search can be issued against stpecific fields of a document	The query `title:gaba` searches for "gaba" only in the title of the doxument
(Extended) Dismax search	User queries are phrases without Boolean connectives; where matches are performed across multiple fields of a document, and in the case of multiple matches of a term in different fields, the max score is used. The query also allows parameters like the minimal number of terms that must be matched, and "query slop factor" that determines the importance of the proximity between query terms.	The query `international knockout mouse` with minimal match = 2 and query slop = 1 will penalize the matching text "the international conference on genetic mouse design was a knockout success".
phrase weights	A query phrase can be given additional multiplicative weight	

- **FIND:**(image video) hippocampus
- *anatomy:*hippocampus *component:*"plasma membrane"
- *anatomy::*organism:human
- *anatomy:*hippocampus[::organism:human]
- **RELATED:**(Tenascin rabbit) **RELATED**::measuredBy:("cell signaling" cytometry)

Semantic Rewriting of SKEYQL Queries

- **FIND:**(image video) hippocampus
  - Find the ontological class of the query term “hippocampus”
    - Ans: anatomical_entity
  - Does “hippocampus” have a non-empty has_part tree underneath?
    - Every node in the ontology keeps an approximate statistics of descendant counts across various edge labels
  - Rewrite query to:
    - FIND: (image video) has_part*(synonyms(hippocampus))
  - Issue:
    - A cell is a part of any brain region.
      - Should the expansion include the cells of hippocampus?
    - No, because “cell” is a different module of the ontology whose top-level is “cell”.
    - Partonomic expansion stays within module of the ontology
  - Final rewrite:
    - FIND: (image video) anatomy::(has_part*(synonyms(hippocampus)))

Computing with NIF Services

The PDSP K_i database is a unique resource in the public domain which provides information on the abilities of drugs to interact with an expanding number of molecular targets. The K_i database serves as a data warehouse for published and internally-derived K_i, or affinity, values for a large number of drugs and drug candidates at an expanding number of G-protein coupled receptors, ion channels, transporters and enzymes.

Which marijuana related genes are of interest to NCI? Let's just use PDSP as an example.

Which marijuana related genes are of interest to NCI? Let's just use PDSP as an example.

Computing with NIF Services

```
public static final String THC_QUERY = "cannabis thc marijuana";

public void demonstrateFederation() throws Exception {
  final String kiDatabaseId = "nif-0000-01866-1";
  FederationQuery query = FederationQuery.builder(kiDatabaseId, THC_QUERY).get();
  for (Facets facets: searcher.getFacets(query, 10, 0, 1)) {
    // Find all receptor (gene) facets
    if (!facets.getCategory().equals("Receptor")) {
      continue;
    }
    for (Facet facet: facets.getFacets()) {
      // Get grants related to these genes from NCI
      query = FederationQuery.builder("nif-0000-10319-1", "\"" + facet.getFacet() +"\"")
          .facet("Funding Institute", "national cancer institute")
          .exportType(ExportType.data)
          .rows(1000).get();
      TableData data = searcher.getTableData(query);

      for (FederationModelData model: data.getResult()) {
        System.out.println(facet.getFacet() + "," + model.get("project_number") + "," + model.get("project_title"));
      }
    }
  }
}
```

Which marijuana related genes are of interest to NCI?

cannabinoid cb2	3R01CA142115-04S1	Cannabinoid CB2 Agonists for Treatment of Breast Cancer-Induced Bone Pain
cannabinoid cb2	1R01CA142115-01A1	Cannabinoid CB2 Agonists for Treatment of Breast Cancer-Induced Bone Pain
sigma	1R01CA163764-01	SIGMA-2/PEPTIDOMIMETIC CONJUGATES TARGET APOPTOSIS IN PANCREATIC CANCER
sigma	1ZIABC008714-35	Bacterial Functions Involved in Cell Growth Control
sigma	1ZIABC010632-09	Transcription Regulation in E. coli and H. pylori
sigma	1ZIABC010632-08	Transcription Regulation in E. coli and H. pylori
sigma	1ZIABC010632-07	Transcription Regulation in E. coli and H. pylori
sigma	1ZIABC008714-32	Bacterial Functions Involved in Cell Growth Control
sigma	1ZIABC008714-34	Bacterial Functions Involved in Cell Growth Control
sigma	1ZIABC008714-33	Bacterial Functions Involved in Cell Growth Control
sigma	1R21CA173887-01A1	Nanomicellar Formulation for Synergistic Targeting of Prostate Cancer
sigma	1F32CA171543-01	Synthesis of Vinblastine Analogues with Improved Physiochemical Properties
sigma	1ZIABC010378-13	Macromolecular Crystallography Research with Synchrotron Radiation
sigma	1ZIABC011203-04	Proteolysis and Regulation of Bacterial Cell Growth Control
sigma	1ZIABC011203-03	Proteolysis and Regulation of Bacterial Cell Growth Control
sigma	1ZIABC010845-04	p53-induced Regulation of Transcription in the Chromatin Context
sigma	1ZIABC011203-02	Proteolysis and Regulation of Bacterial Cell Growth Control
sigma	1ZICBC010517-08	Large Databases of Small Molecules - Drug Development Tool and Public Resource
sigma	2R56CA107510-06	The role of p53 and 14-3-3 in genomic instability
sigma	1ZIABC011203-01	Proteolysis and Regulation of Bacterial Cell Growth Control

What about ALL Databases?

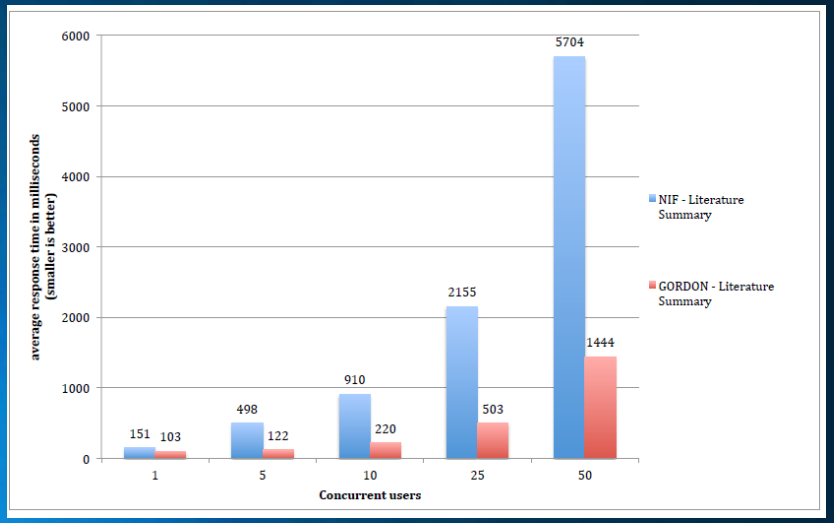
	gene	projNum	projTitle
2	bptf	1ZIABC005263-28	Eukaryotic Chromatin Structure and Gene Regulation
3	bptf	1ZIABC005263-29	Eukaryotic Chromatin Structure and Gene Regulation
4	bptf	1ZIABC005263-31	Eukaryotic Chromatin Structure and Gene Regulation
5	bptf	1ZIABC005263-30	Eukaryotic Chromatin Structure and Gene Regulation
6	itk	1ZIABC011267-03	Preclinical drug development in pancreatic cancer
7	itk	1R41CA167907-01	Calibrated Methods for Quantitative PET/CT Imaging
8	itk	1ZIABC009281-26	Receptor Mediated T and B Cell Activation
9	itk	1ZIABC010304-14	Biochemical Basis of T Cell Activation
10	itk	1ZIABC009281-25	Receptor Mediated T and B Cell Activation
11	itk	1ZIABC010304-13	Biochemical Basis of T Cell Activation
12	itk	1ZIABC010944-04	Control of the immune response for cancer vaccine development
13	itk	1ZIABC009281-24	Receptor Mediated T and B Cell Activation
14	itk	1ZIABC010304-12	Biochemical Basis of T Cell Activation
15	itk	1ZIABC010944-03	Control of the immune response for cancer vaccine development
16	itk	1ZIABC010304-11	Biochemical Basis of T Cell Activation
17	itk	1ZIABC010944-02	Control of the immune response for cancer vaccine development
18	itk	3R01CA112663-10S1	T-bet and Tumor Immunity
19	crkl	1ZIASC006892-23	Molecular Biology of Pediatric Tumors
20	crkl	1ZIASC006892-24	Molecular Biology of Pediatric Tumors
21	crkl	1ZIASC006892-22	Molecular Biology of Pediatric Tumors
22	crkl	1ZIASC006892-21	Molecular Biology of Pediatric Tumors
23	tetrahydrocannabinol (thc)	3R01CA111196-04S1	MODULATION OF ONCOGENIC AGENTS BY MARIJUANA
24	brca1	2R01CA089239-13A1	Analysis of BRCA1 function in DNA Repair
25	brca1	7R01CA111436-04	Regulation of BRCA1 Function by Protein Phosphatase 1
26	brca1	1R01CA137023-01A1	The Role of BRCA1/BARD1 in Basal-like Breast Cancer
27	brca1	4R01CA129440-03	ROLE OF BRCA1/AKT1 PATHWAY IN THE TUMORIGENESIS
28	brca1	1R01CA129440-01A2	ROLE OF BRCA1/AKT1 PATHWAY IN THE TUMORIGENESIS
29	brca1	7R01CA089239-11	Analysis of BRCA1 Function in DNA Repair
30	brca1	1R01CA174904-01	Roles of Chromatin Modification in BRCA1 Dependent DNA Repair
31	brca1	1ZIABC010847-06	Gene-specific Mechanisms of BRCA1 transcriptional Control

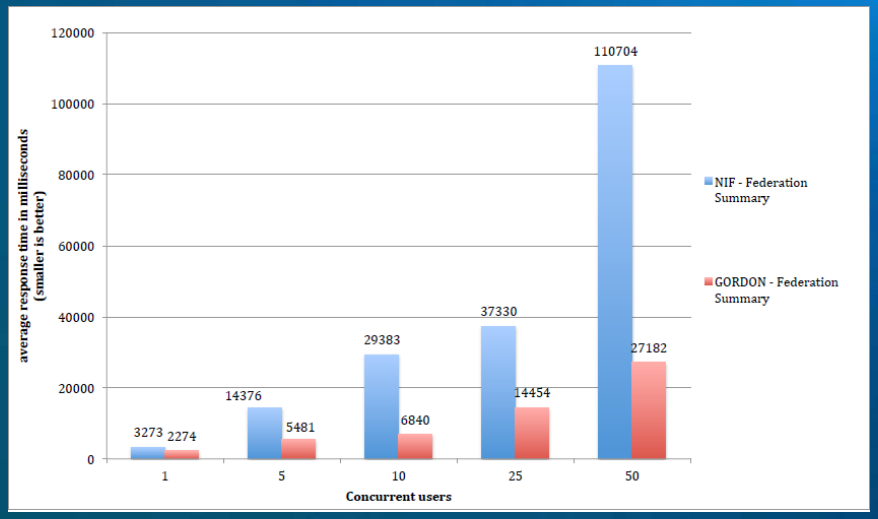
Where is SDSC in all of this?

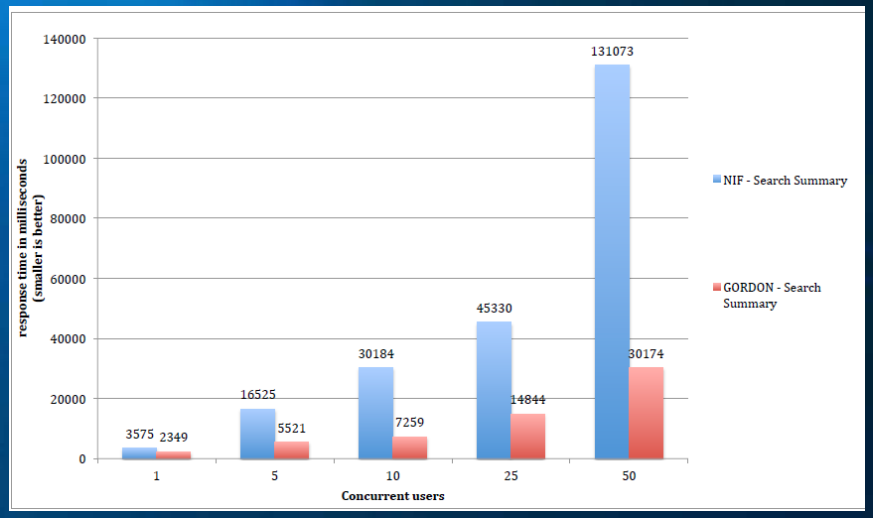
Running on Gordon (with no tweaking)

Configuration

- 1 server
  - 98 GB RAM
  - 24 core Intel Xeon CPU @2.8GHz
- RAID 5 SSD
- 2 Solr instances serving the
- Federation and literature cores.







Conclusion

- The Neuroscience Information Framework is not really dependent on Neuroscience
- Applying it to
  - Diabetes and kidney diseases
  - Model organisms
  - Earthcube for geo-science data
  - (Hopefully) social science data for economists
- We need
  - More scalability
  - Improved complex query handling
  - A distribution framework