

The Neuroscience Information Framework:

A Scalable Platform for Information Exploration and
Semantic Search Computing

Amarnath Gupta

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
- Who uses them?
- Who creates them?
- How can we find them?
- How can we make them better in the future?

- PDF files
- Desk drawers

Discovery and utilization of web-based resources for neuroscience

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

Literature

Database
Federation

NIF NAVIGATOR



LITERATURE →

PubMed (22361958)

NIF DATA FEDERATION →

DATA TYPE

Animals (136449)
Annotation (17450074)
Antibodies (2240902)
Atlas (267687512)
Biospecimen (253392)
Brain Activation Foci (56588)
Clinical Trials (274674)
Connectivity (50316)
Dataset (10397)
Disease (3459177)
Drugs (3412505)
Genes (246030)
Grants (2720110)
Images (923534)
MRI (574809)
Microarray (312869038)
Models (1414)
Multimedia (87247)
Pathways (789668)
People (377)
Phenotype (564493)
Plasmids (27649)
Registries (6371)
Software (1857)

NERVOUS SYSTEM LEVELS

Brain Regions (42777)
Cellular Level (59136)
Genes (62671895)
Molecular Level (693109)
Nervous System
Function (68872)

NIF REGISTRY (6078) →

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"

NIF

NEUROSCIENCE INFORMATION FRAMEWORK

Search for All Things Neuroscience

Search NIF

SEARCH TIPS | WHAT IS THIS? (example searches: cerebellum, genetic analysis software, gene:grm1)

Search Ne



Have Data?



Share your data with NIF

Data Sharing Plans | Lab Data Management | Large Data

Click to find out more

Data Sharing with NIF

Community News & Events

Twitter

How long does it take to get a mouse source brain.

June 4th, 2013 05:24 PM

Believe it or not, there is a wonderful community

Registry

▼ ABOUT

About NIF

People

Publications

Presentations

Brochures

Testimonials

Release Notes 5.1

FAQ

▼ NIF PRODUCTS

▼ NIF DATA SHARING

▼ NIF SYSTEM

▼ SOCIAL MEDIA



Registered with NIF?
Place this icon on your site.



NIH Blueprint

for Neuroscience Research

June 10, 2013

<http://neuinfo.org>

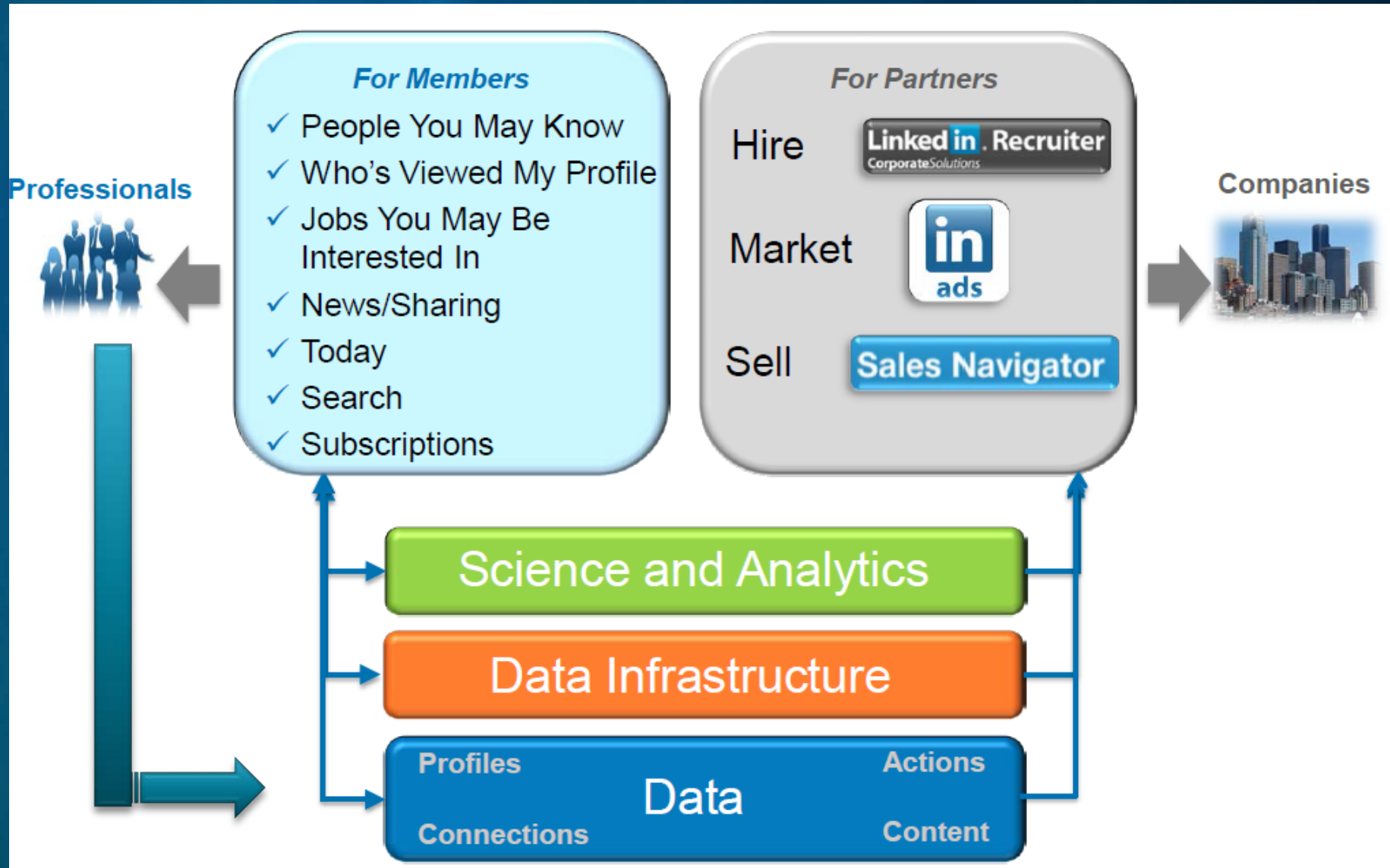
A Big, Fat Techno-Philosophical Question

- 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:



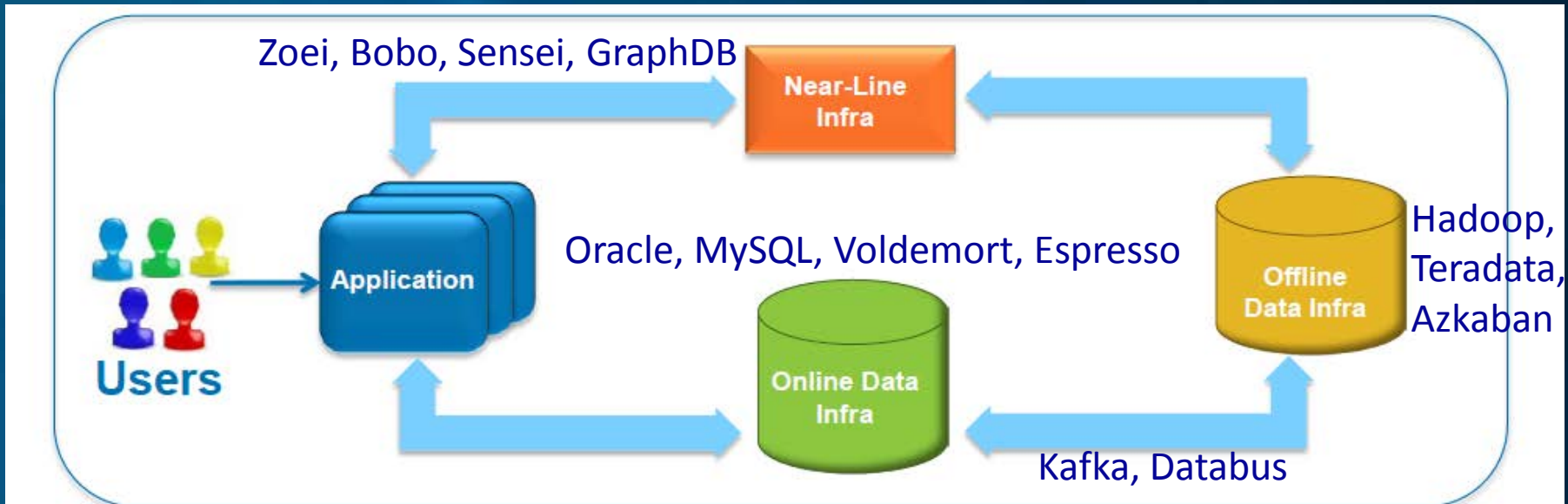
A Digression – Looking at LinkedIn

Credit: Bhaskar Ghosh, LinkedIn



A Digression – Looking at LinkedIn

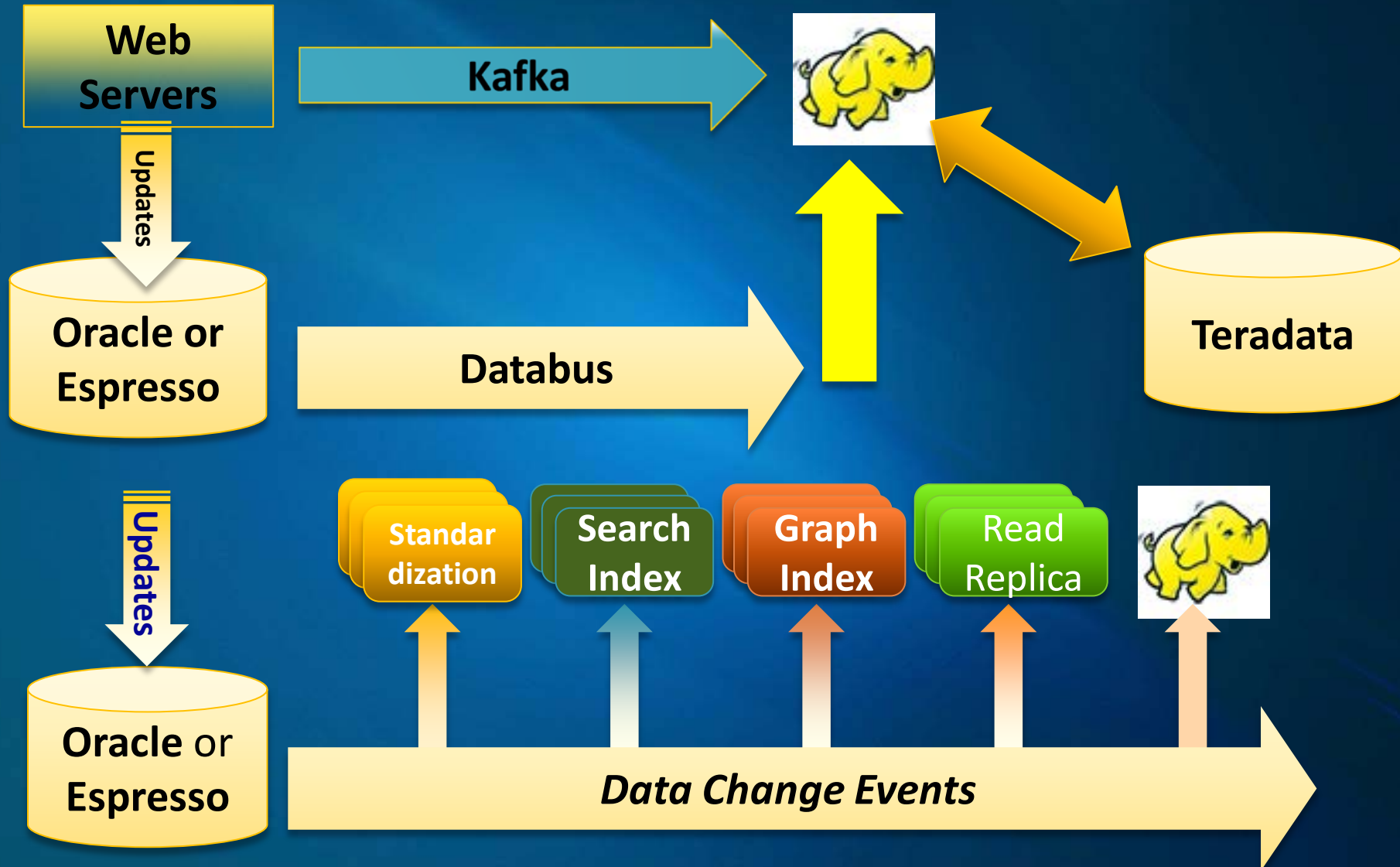
Credit: Bhaskar Ghosh, LinkedIn



Infrastructure	Latency & Freshness Requirements	Products
Online	Activity that should be reflected immediately	<ul style="list-style-type: none"> Member Profiles Company Profiles Connections Messages Endorsements Skills
Near-Line	Activity that should be reflected soon	<ul style="list-style-type: none"> Activity Streams Profile Standardization News Recommendations Search Messages
Offline	Activity that can be reflected later	<ul style="list-style-type: none"> 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

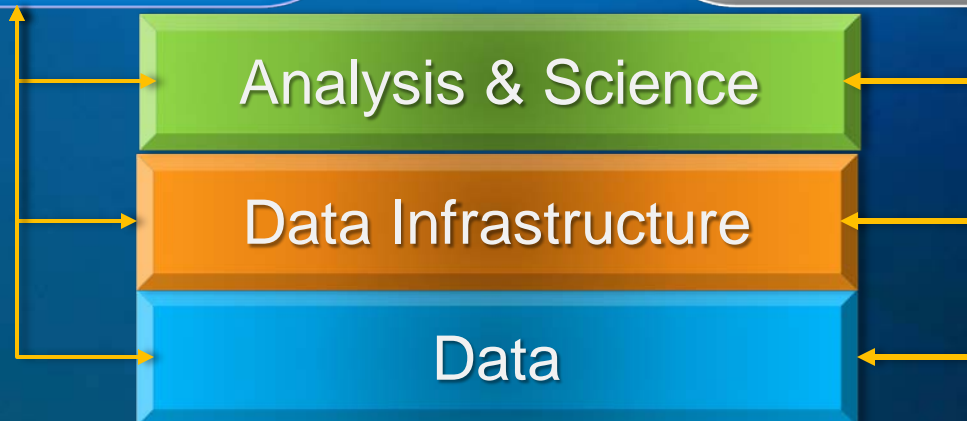


Researcher Activity

- ✓ 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

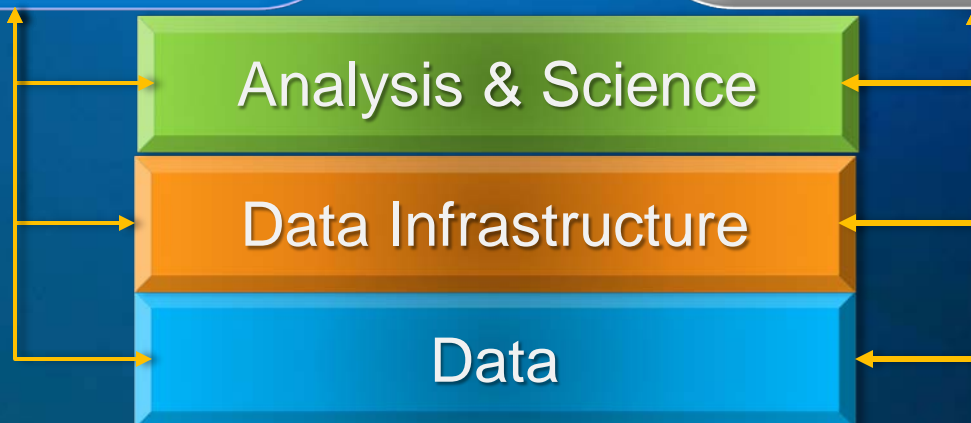
Where is data about X?

- ✓ 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

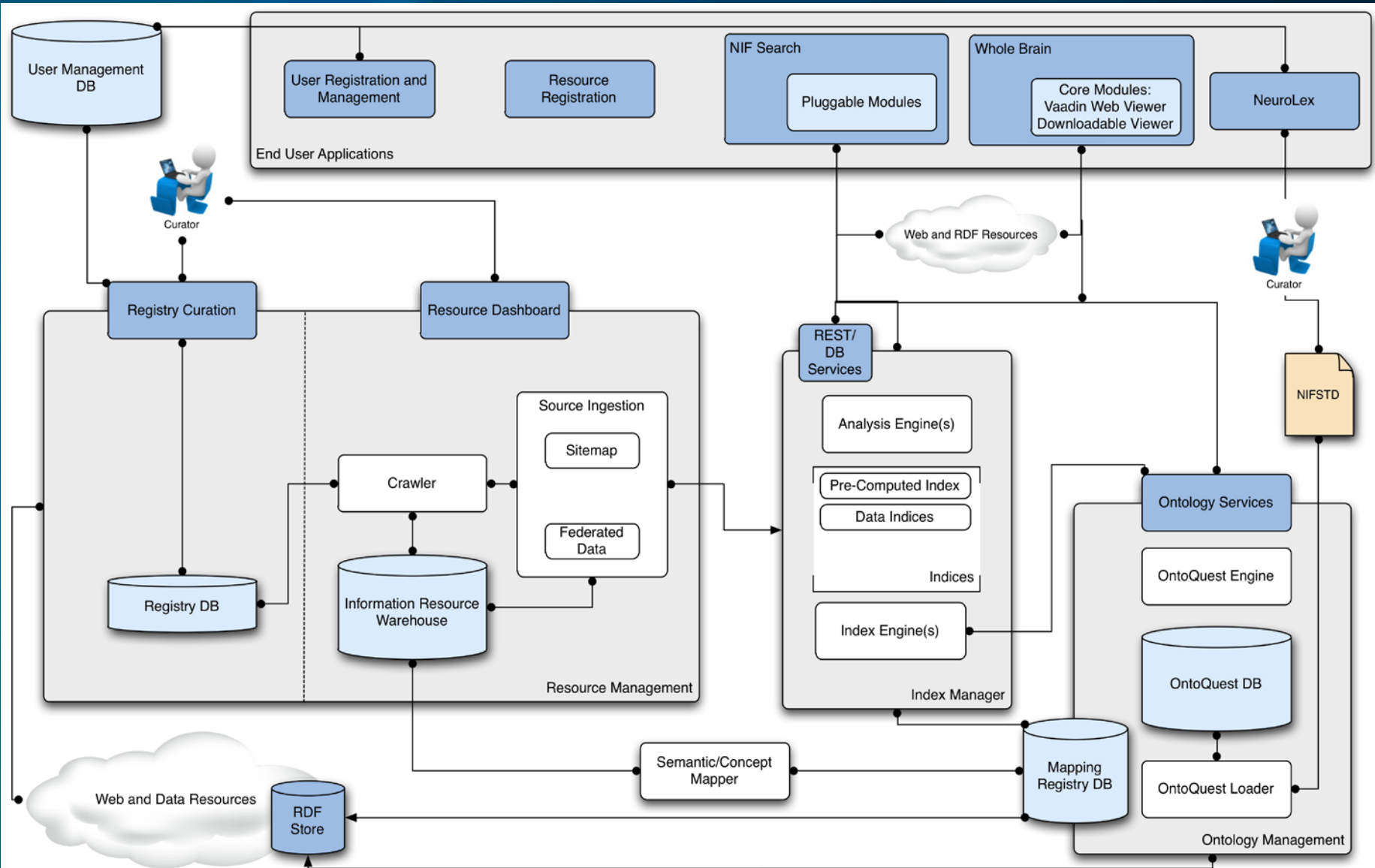
Resource Activity

- ✓ 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

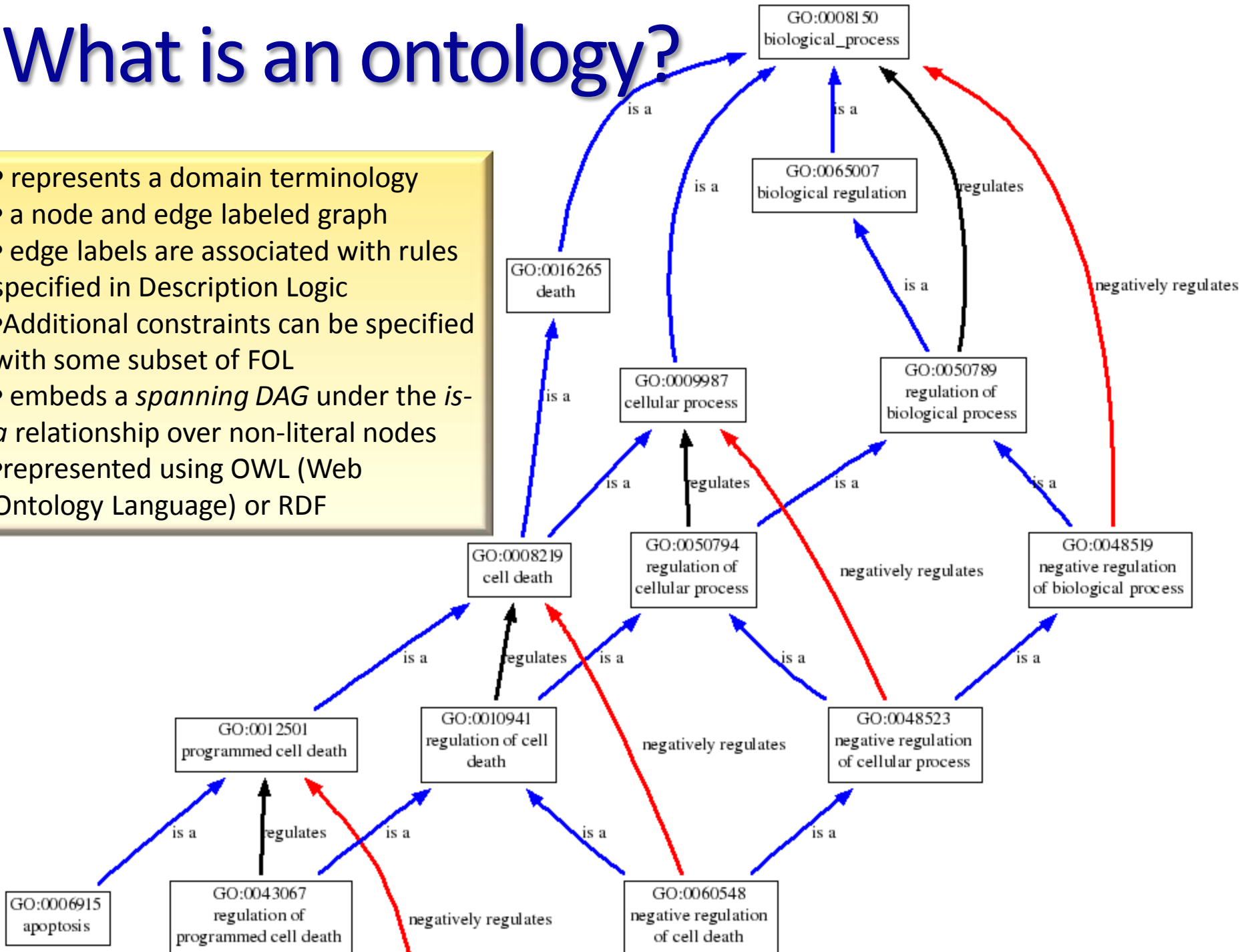
- Commercial solutions exist
 - They are not very scalable as number of schemas increase
- Many groups working on it
 - Data Tamer – an MIT-Intel partnership project on Big Data
 - Matches schema where possible
 - Crowd-sources ambiguous cases
 - Performs entity-consolidation by data clustering
- Our 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

```
--<success>
--<data>
--<classes>
--<class>
<id InternalId="357740-1">HP_0001049</id>
<name>HP_0001049</name>
--</class>
--</classes>
--</data>
--</success>

--<class>
<id InternalId="264158-1">ZFA_0001249</id>
<name>ZFA_0001249</name>
<label>exocrine pancreas</label>
<uri>http://purl.obolibrary.org/obo/ZFA_0001249</uri>
--<comments>
--<comment>
The exocrine pancreas is composed of acinar epithelial cells and ductal epithelium that manufacture the
</comment>
--</comments>
--<other_properties>
<property name="has_obo_namespace">zebrafish_anatomy</property>
<property name="database_cross_reference">ZFAN_ZDB-ANAT-050711-6</property>
<property name="id">ZFA_0001249</property>
<property name="OBO foundry unique label">exocrine pancreas (zebrafish)</property>
<property name="id">ZFA_0001249</property>
<property name="database_cross_reference">TAO_0001249</property>
--</other_properties>
--</class>
--<class>
<id InternalId="247995-1">UBERON_0000017</id>
<name>UBERON_0000017</name>
<label>exocrine pancreas</label>
<uri>http://purl.obolibrary.org/obo/UBERON_0000017</uri>
--<comments>
--<comment>
The exocrine pancreas produces and store zymogens of digestive enzymes, such as chymotrypsinogen a
epithelium that manufacture the proteolytic enzymes and bicarbonate required for digestion. [definition]
</comment>
--</comments>
--<other_properties>
<property name="database_cross_reference">CALOHA:TS-1241</property>
<property name="database_cross_reference">EV_0100093</property>
--</other_properties>
--</class>
```

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

```
-<relationship>
  -<subject InternalId="118064-3" id="SomeValuesFrom Restriction">
    something [exists]"develops from" "future cardiac ventricle"
  </subject>
  <property InternalId="4558-15" id="RO_0002202">RO_0002202</property>
  <object InternalId="590274-1" id="UBERON_0010226">future cardiac ventricle</object>
</relationship>
-<relationship>
  <subject InternalId="591402-1" id="ZFA_0009351">thromboplast</subject>
  <property InternalId="4558-15" id="RO_0002202">RO_0002202</property>
  <object InternalId="591954-1" id="ZFA_0009022">megakaryocyte erythroid progenitor cell</object>
</relationship>
-<relationship>
  <subject InternalId="585974-1" id="UBERON_0004764">intramembranous bone tissue</subject>
  <property InternalId="4558-15" id="RO_0002202">RO_0002202</property>
  <object InternalId="583746-1" id="UBERON_0003104">mesenchyme</object>
</relationship>
-<relationship>
  -<subject InternalId="114818-3" id="SomeValuesFrom Restriction">
    something [exists]"develops from" "early telencephalic vesicle"
  </subject>
  <property InternalId="4558-15" id="RO_0002202">RO_0002202</property>
  <object InternalId="585190-1" id="UBERON_0009676">early telencephalic vesicle</object>
</relationship>
-<relationship>
  <subject InternalId="584423-1" id="UBERON_0000080">mesonephros</subject>
  <property InternalId="4558-15" id="RO_0002202">RO_0002202</property>
  <object InternalId="588719-1" id="UBERON_0002120">pronephros</object>
</relationship>
```

“equivalenceClass”

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

```
<relationship>
  <subject InternalId="777193-1" id="ZFA_0005494">skeletal element</subject>
  <property InternalId="5360-15" id="equivalentClass">equivalentClass</property>
  <object InternalId="585975-1" id="UBERON_0004765">skeletal element</object>
</relationship>
```

Subclasses of “skeletal element” (incl. its equivalenceClasses) in ZFA

<http://nif-services-stage.neuinfo.org/ontoquest-lamhdi/rel/subclasses/term/skeletal%20element?level=3>

```
<relationship>
  <subject InternalId="778440-1" id="ZFA_0001635">intramembranous bone</subject><property InternalId="5389-15" id="subClassOf">subClassOf</property><object InternalId="777874-1" id="ZFA_0001514">bone element</object>
</relationship>
```

The Entity Recognition Problem

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

A first-order linear-chain CRF

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

where

- \mathbf{y} is output vector
- \mathbf{x} input data.
- $f_k(\mathbf{y}_{t-1}, \mathbf{y}_t, \mathbf{x}, t)$ output(IOB label)
- λ_k weight of t th maximum likelihood
- $Z(\mathbf{x})$ normalization
- 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 image displays two screenshots of the Named Entity Annotator Tool interface. The left screenshot shows the main text area with a highlighted sentence: "After rinsing in PBS for 5 min, the sections were incubated with blocking solution containing 5 % goat antiserum at room temperature for 1 h and then incubated with rabbit anti-lba". A legend below the text identifies named entities: source (green), organism (blue), target (red), antibody (purple), and clonality (orange). The right screenshot shows a detailed view of an extracted record for an antibody, listing its frame, vendor, source organism, target organism, clonality, catnum, and target.

Named Entity Annotator Tool

File Tools Help

Named Entities Extracted Records

antibody

Sentences

Next, the sections were incubated with 1:50 anti-D2 receptor mouse monoclonal IgG -LRB- D2DR, Santa Cruz Biotechnology -RRB- and 1:1000 anti-5HT1A receptor rabbit polyclonal antibodies -LRB- anti-ERK, anti-p-ERK, anti-JNK, anti-p-JNK, anti-p38, anti-p-p38, anti-β actin and anti-cleaved caspase 3 -RRB- were obtained from Cell Signaling. Anti-BrdU antibody was obtained from Sigma, anti-Tuj antibody was from Covance, anti-S100 antibody was from Dako, and anti-MBP was from Novus Biologicals. OPG and RAN Single cells were cultured in DMEM-high glucose and F12 -LRB- mixed 1:1 -RRB- supplemented with N2 and B27 -LRB- Invitrogen -RRB-, EGF -LRB- 20 ng/mL; Chemicon -RRB- For immunofluorescence double labeling -LRB- n = 3 -RRB-, brain sections -LRB- = 2.56 to - 3.30 caudal to the bregma, i.e. levels 1 and 2 -RRB- were first incubated in BS for 1 h. After rinsing in PBS for 5 min, the sections were incubated with blocking solution containing 5 % goat antiserum at room temperature for 1 h and then incubated with rabbit anti-lba antibody -LRB- 1:500, Invitrogen, Carlsbad, CA -RRB- After washing with PBS three times for 5 min, the sections were reacted with Alexa Fluor 488-conjugated anti-rabbit IgG antibody -LRB- 1:500, Invitrogen, Carlsbad, CA -RRB- The sections were incubated with primary antibodies -LRB- RANTES Antibody -LRB- AF478, 1:10 dilution, R&D SYSTEMS, Minneapolis, MN -RRB-, NeuN -LRB- MAB377, 1:50 dilution, Chemicon -RRB- The following primers -LRB- Sigma-Aldrich, Japan -RRB- were used: mouse RANTES -LRB- forward, 5' - ATATGGCTCGGACCACTC-3'; reverse, 5' - TGACAAAGACGACTG-3'. The primary antibodies used in this study were purchased from Cell Signaling Technology -LRB- CST, Boston, MA -RRB- -LRB- total Akt -LRB- # 4691, 1:1000 dilution -RRB-, p-ERK -LRB- # 4370, 1:1000 dilution -RRB-, p-JNK -LRB- # 4668, 1:1000 dilution -RRB-, p-p38 -LRB- # 4660, 1:1000 dilution -RRB-, p-p38 -LRB- # 4660, 1:1000 dilution -RRB-, p-actin -LRB- # 4957, 1:1000 dilution -RRB-, p-actin -LRB- # 4957, 1:1000 dilution -RRB-, p-actin -LRB- # 4957, 1:1000 dilution -RRB-, p-actin -LRB- # 4957, 1:1000 dilution -RRB-. Gene-specific primer pairs for amplification of TRPC1, TRPC3, TRPM2, HO-1 and GAPDH were designed using Primer Express v2.0 -LRB- Applied Biosystems, Foster City, CA -RRB-. For immunodetection of TRPC3 protein, nitrocellulose membranes -LRB- Perkin-Elmer, Woodbridge, ON -RRB- were blocked with 0.5 % egg white albumin -LRB- 1 h, room temperature -RRB-. For immunodetection of TRPM2, membranes were blocked with 5 % milk/1 % bovine serum albumin -LRB- 1 h, RT -RRB-, then incubated with 1:500 anti-TRPM2 antibody -LRB- 1:500, Invitrogen, Carlsbad, CA -RRB-. For β-actin, a blocking solution of 5 % milk was used, 1:2,000 anti-β-actin antibody -LRB- New England Biolabs, Pickering, ON -RRB- -LRB- overnight, 4 °C -RRB- followed by monoclonal antibodies: The anti-sac monoclonal antibodies were developed in our laboratory -LRB- Zippin et al., 2003 -RRB-. For EM, sections were incubated in colloidal gold -LRB- 1 nM -RRB- conjugated goat anti-mouse IgG -LRB- 1:50; Electron Microscopy Sciences, EMS -RRB- in 0.08 % BSA, 0.01 % BSA. The following cell-type specific markers were used: rabbit polyclonal anti-Olig2 -LRB- Cat #AB 9610, 1:400, Millipore, Temecula, CA -RRB-, rabbit polyclonal anti-doublecortin -LRB- Cat #AB 9610, 1:400, Millipore, Temecula, CA -RRB-. Sections were stained with rat anti-BrdU antibodies -LRB- Cat #MCA 2060T, 1:500, AbD Serotec, Raleigh, NC -RRB- or mouse monoclonal anti-PCNA -LRB- Cat #sc-25280, 1:500, Santa Cruz Biotechnology, Santa Cruz, CA -RRB-.

NER Annotation

After rinsing in PBS for 5 min, the sections were incubated with blocking solution containing 5 % goat antiserum at room temperature for 1 h and then incubated with rabbit anti-lba antibody -LRB- 1:500, Wako, Osaka -RRB- for microglia.

Legend - Named Entities for Sentences

source organism target antibody clonality

Overlapping Entities

After rinsing in PBS for 5 min, the sections were incubated with blocking solution containing 5 % goat antiserum at room temperature for 1 h and then incubated with rabbit anti-lba antibody -LRB- 1:500, Wako, Osaka -RRB- for microglia.

Named Entity Annotator Tool

File Tools Help

Named Entities Extracted Records

antibody Filter Sentence contains Search Full Set NE Stats

Sentences

Next, the sections were incubated with 1:50 anti-D2 receptor mouse monoclonal IgG -LRB- D2DR, Santa Cruz Biotechnology -RRB- and 1:1000 anti-5HT1A receptor rabbit polyclonal antibodies -LRB- anti-ERK, anti-p-ERK, anti-JNK, anti-p-JNK, anti-p38, anti-p-p38, anti-β actin and anti-cleaved caspase 3 -RRB- were obtained from Cell Signaling. Anti-BrdU antibody was obtained from Sigma, anti-Tuj antibody was from Covance, anti-S100 antibody was from Dako, and anti-MBP was from Novus Biologicals. OPG and RAN Single cells were cultured in DMEM-high glucose and F12 -LRB- mixed 1:1 -RRB- supplemented with N2 and B27 -LRB- Invitrogen -RRB-, EGF -LRB- 20 ng/mL; Chemicon -RRB- For immunofluorescence double labeling -LRB- n = 3 -RRB-, brain sections -LRB- = 2.56 to - 3.30 caudal to the bregma, i.e. levels 1 and 2 -RRB- were first incubated in BS for 1 h. After rinsing in PBS for 5 min, the sections were incubated with blocking solution containing 5 % goat antiserum at room temperature for 1 h and then incubated with rabbit anti-lba antibody -LRB- 1:500, Invitrogen, Carlsbad, CA -RRB- After washing with PBS three times for 5 min, the sections were reacted with Alexa Fluor 488-conjugated anti-rabbit IgG antibody -LRB- 1:500, Invitrogen, Carlsbad, CA -RRB- The sections were incubated with primary antibodies -LRB- RANTES Antibody -LRB- AF478, 1:10 dilution, R&D SYSTEMS, Minneapolis, MN -RRB-, NeuN -LRB- MAB377, 1:50 dilution, Chemicon -RRB- The following primers -LRB- Sigma-Aldrich, Japan -RRB- were used: mouse RANTES -LRB- forward, 5' - ATATGGCTCGGACCACTC-3'; reverse, 5' - TGACAAAGACGACTG-3'. The primary antibodies used in this study were purchased from Cell Signaling Technology -LRB- CST, Boston, MA -RRB- -LRB- total Akt -LRB- # 4691, 1:1000 dilution -RRB-, p-ERK -LRB- # 4370, 1:1000 dilution -RRB-, p-JNK -LRB- # 4668, 1:1000 dilution -RRB-, p-p38 -LRB- # 4660, 1:1000 dilution -RRB-, p-p38 -LRB- # 4660, 1:1000 dilution -RRB-, p-actin -LRB- # 4957, 1:1000 dilution -RRB-, p-actin -LRB- # 4957, 1:1000 dilution -RRB-, p-actin -LRB- # 4957, 1:1000 dilution -RRB-. Gene-specific primer pairs for amplification of TRPC1, TRPC3, TRPM2, HO-1 and GAPDH were designed using Primer Express v2.0 -LRB- Applied Biosystems, Foster City, CA -RRB-. For immunodetection of TRPC3 protein, nitrocellulose membranes -LRB- Perkin-Elmer, Woodbridge, ON -RRB- were blocked with 0.5 % egg white albumin -LRB- 1 h, room temperature -RRB-. For immunodetection of TRPM2, membranes were blocked with 5 % milk/1 % bovine serum albumin -LRB- 1 h, RT -RRB-, then incubated with 1:500 anti-TRPM2 antibody -LRB- 1:500, Invitrogen, Carlsbad, CA -RRB-. For β-actin, a blocking solution of 5 % milk was used, 1:2,000 anti-β-actin antibody -LRB- New England Biolabs, Pickering, ON -RRB- -LRB- overnight, 4 °C -RRB- followed by monoclonal antibodies: The anti-sac monoclonal antibodies were developed in our laboratory -LRB- Zippin et al., 2003 -RRB-. For EM, sections were incubated in colloidal gold -LRB- 1 nM -RRB- conjugated goat anti-mouse IgG -LRB- 1:50; Electron Microscopy Sciences, EMS -RRB- in 0.08 % BSA, 0.01 % BSA. The following cell-type specific markers were used: rabbit polyclonal anti-Olig2 -LRB- Cat #AB 9610, 1:400, Millipore, Temecula, CA -RRB-, rabbit polyclonal anti-doublecortin -LRB- Cat #AB 9610, 1:400, Millipore, Temecula, CA -RRB-. Sections were stained with rat anti-BrdU antibodies -LRB- Cat #MCA 2060T, 1:500, AbD Serotec, Raleigh, NC -RRB- or mouse monoclonal anti-PCNA -LRB- Cat #sc-25280, 1:500, Santa Cruz Biotechnology, Santa Cruz, CA -RRB-.

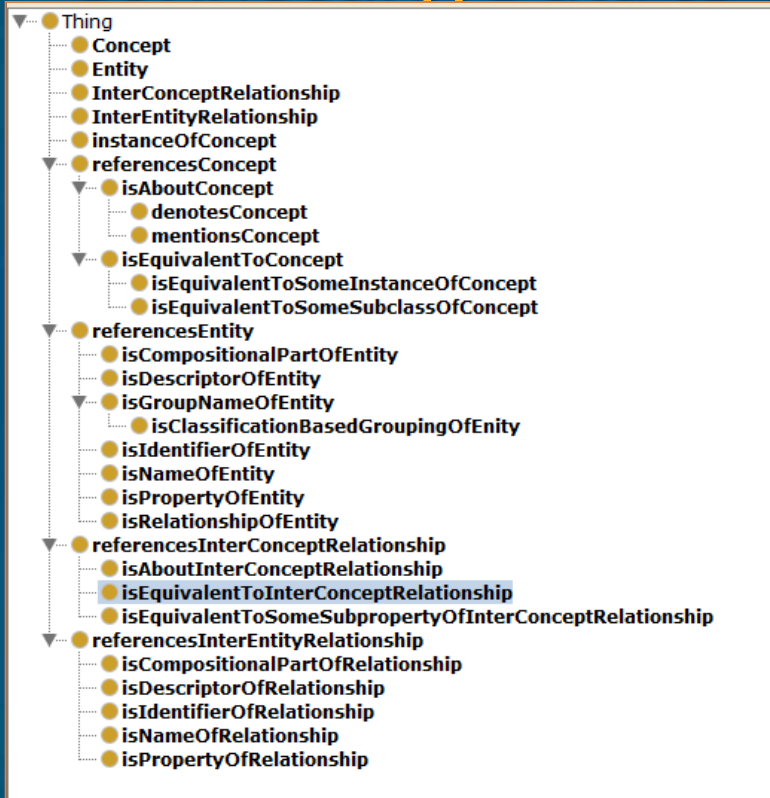
Antibody 1

-Antibody Frame

antibody	polyclonal antibody -LRB- 1:500, Wako, Osaka -RRB- for microglia
vendor	Wako
source_organism	rabbit
target_organism	
clonality	polyclonal
catnum	
target	anti-lba 1

/Users/bozyurt/elsevier10_idx_ie.xml

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

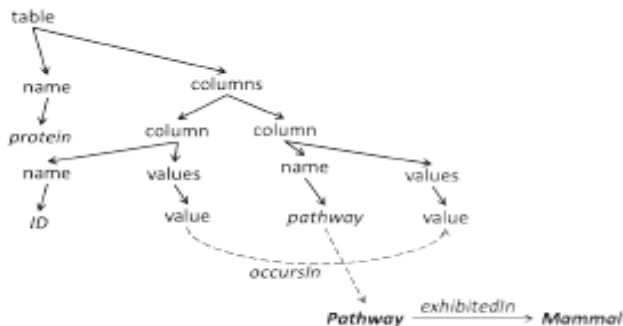


Table: protein(ID, symbol, name, pathway)

Mapping:

pathway mapsTo ontoID(Pathway).

(exhibitedIn(ontoID(Pathway), ontoID(mammal))

AND occursIn(value(ID), value(pathway)))

The Concept Mapping Tool

Concept Map Tool

Browse Add Import

Basic Information
View Definition
Mappings
Display

View Definition (SQL)

```
select distinct
replace(b.e_uid,':','') as e_uid,
replace(replace (b.resource_name, ' ',
resource_name,
b.abbrev,
b.availability,
b.comment,
replace(replace (b.definition, '<BR>',
b.curationstatus,
b.definingcitation as url,
lower(replace (replace (replace (replace
ce (replace (replace (replace (replace (rep
', ' from
b.has role),' ','),'0',''),'9',''),'8
,'3',''),'2',''),'1',''),' (birnlex_'
(nlx_res ','),'nlx_inv ',''),'',''))
replace(b.is_part_of, 'Resource:', '')
b.keyword,
b.created_date as date_created,
b.last_update_date as date_updated,
case when b.pmid is null then '' else
nif_pmid_display,

```

Vocabulary View

Check View & Update Columns

Columns

Name	Facet Name	Data type
------	------------	-----------

Modify Column Information

Basic Information

Name:

Facet Name:

Data Type:

Weight:

Delimiter:

Indexable?: Yes No

Facet?: Yes No

Is Key?: Yes No

Exportable?: Yes No

Description

| A⁺ A⁻ >>

Add Semantic Mapping

Column Mappings

Add New Delete Modify

Concept Id	Concept Name	Ontology	Rule	Last Upda...	Update Time
birnlex_2	Organism	NIF		anita	Thu Feb 21 16...

Save Cancel

2013 Summer Institute: Dis

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?source=1&columnName=transgenic_line

Now group and rank merge

in parallel if needed

Get top 10

```
<results method="getColumnValueFrequencies">
  <value freq="209"/>
  <value freq="4">Etv1-CreERT2</value>
  <value freq="4">Grik4-Cre</value>
  <value freq="3">Syt6-Cre</value>
  <value freq="3">Rbp4-Cre</value>
  <value freq="3">Nr5a1-Cre</value>
  <value freq="3">Gal-Cre</value>
  <value freq="2">Slc6a4-CreERT2</value>
  <value freq="2">Pmch-Cre</value>
  <value freq="2">Lepr-IRES-Cre</value>
  <value freq="2">Ntsr1-Cre</value>
  <value freq="1">Pomc-Cre (ST)</value>
  <value freq="1">Oxt-IRES-Cre</value>
  <value freq="1">Erbb4-2A-CreERT2</value>
  <value freq="1">Scnn1a-Tg3-Cre</value>
</results>
```

```
<results method="getOntoTermColumnMappings">
  <result source="nif-0000-00096" table="nif-0000-00096-1" column="Gene / Protein Name"/>
  <result source="nlx_37081" table="nlx_37081-1" column="gene_symbol"/>
  <result source="nif-0000-34000" table="nif-0000-34000-1" column="allele_name"/>
  <result source="nif-0000-07730" table="nif-0000-07730-1" column="AB_TARGET"/>
  <result source="nlx_22354" table="nlx_22354-1" column="GENE_SYMBOL"/>
  <result source="nif-0000-03213" table="nif-0000-03213-1" column="odor_ligands"/>
  <result source="nif-0000-23326" table="nif-0000-23326-1" column="allele_type"/>
  <result source="nif-0000-02683" table="nif-0000-02683-1" column="chemicalname"/>
  <result source="nif-0000-08127" table="nif-0000-08127-1" column="gene_symbol"/>
  <result source="nlx_152726" table="nlx_152726-1" column="model_receptors"/>
  <result source="nlx_98194" table="nlx_98194-1" column="ensembl_gene_symbol"/>
  <result source="nlx_149225" table="nlx_149225-1" column="reverse_primer"/>
  <result source="nlx_149225" table="nlx_149225-1" column="encoding_protein_product"/>
  <result source="nlx_152726" table="nlx_152726-1" column="model_neurotransmitters"/>
  <result source="nlx_149225" table="nlx_149225-1" column="ma_polymerase"/>
  <result source="nlx_23971" table="nlx_23971-1" column="ensembl_gene_symbol"/>
  <result source="nif-0000-02683-2" table="nif-0000-02683-2" column="chemicalid"/>
  <result source="nif-0000-34000-1" table="nif-0000-34000-1" column="allele_symbol3"/>
  <result source="nif-0000-00517" table="nif-0000-00517-1" column="geneset_name"/>
  <result source="nif-0000-20925-1" table="nif-0000-20925-1" column="symbol"/>
  <result source="nif-0000-20925-1" table="nif-0000-20925-1" column="cas"/>
  <result source="nif-0000-34000-1" table="nif-0000-34000-1" column="marker_symbol"/>
  <result source="nlx_146253" table="nlx_146253-1" column="transgenic_line"/>
  <result source="nif-0000-20925-1" table="nif-0000-20925-1" column="chebi"/>
  <result source="nlx_149225" table="nlx_149225-1" column="gene"/>
```

Search: The Keyword Query Interface

NIF Home | myNIF | Neurolex | Search | Recommend a Resource Login | Register | Tutorial | Help

NIF search for (e.g., cerebellum, "pulvinar nucleus")

Hippocampus gene.* ☆ 🔍

Data () Search Options Semantic Expansion [-]

Literature () ammon's horn, ammon horn, hippocampus proper, cornu ammonis

Registry () WormBase: WormPhenotypes

Funding () WormBase provides anatomical and genetic information of C. elegans and related research nematodes.

Web (←) Information on WormBase 399,688 Results

NIF Home | myNIF | Neurolex | Search | Recommend a Resource Login | Register | Tutorial | Help

NIF search for (e.g., cerebellum, "pulvinar nucleus")

Hippocampus gene.* ☆ 🔍

AutDB:AnimalModels Source Options Semantic Expansion [+]

AutDB provides information on animal models used in autism research.

[More Information on AutDB](#)

Displaying results 1 - 20 out of 2271 total results.

Hide search filters

Animal Model [+]

Gene Symbol [-]

- mecp2 (264)
- fmr1 (140)
- disc1 (122)
- ube3a (89)
- gabbr3 (78)
- nf1 (78)
- foxp2 (69)
- oxtr (67)
- shank3 (65)
- reln (52)

Animal Model	Gene Symbol	Gene Name	Aliases	Phenotype Profile	Experimental Paradigm	Reference For Phenotype	Reference For Model
UBE3A_1_KO_HM	Ube3a	ubiquitin protein ligase E3A	Hpv6a; KIAA4216; mKIAA4216; 4732496B02; 5830462N02Rik; A130086L21Rik; Ube3a	Neurophysiology: Description: Decreased chloride inhibitory currents in cornu ammonis 1 (CA1) pyramidal ne ...[more]	Intracellulr recordings after hippocampal cannulation and drug infusion	PMID:19430469, 20211139, 20696245, 21974935, 21974935, 22381732, 9808466, 19430469, 21624971, 9808466, 9808466, 9808466	PMID:11543639, 16575182, 16754645, 18846633, 19404257, 19430469, 21624971, 8988171
NRP2_4_KO_HT	Nrp2	neuropilin 2	RP23-149A5.1, 1110048P06Rik, Np-2, Np2, Npn-2, Npn2	Neurophysiology: Description: Increased excitation as inferred by population spike amplitude in field CA1 ...[more]	I/O analysis	PMID:10707970, 10707970, 18657176, 17427189, 17443771, 18657176, 18657177	PMID:10707971, 17259176, 17329436, 17427189, 17443771, 9288754
NTNG1_1_KO_HM	Ntng1	Netrin G1	laminet 1; laminet-1; netrin-G1	Molecular profile: Description: Decreased Ngl-1 immunoreactivity in hippocampus, layer I of the parietal c ...[more]	Immunohistochemical analysis	PMID:17785411	PMID:14595443, 15870826, 16980967, 17785411, 17973922, 18384956
FMR1_1_KO_HM	Fmr1	fragile X mental	FMRP; Fmr-1; Fmr1	Molecular profile:		PMID:16055059	PMID:11773805

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 specific fields of a document	The query title:gaba searches for “gaba” only in the title of the document
(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?

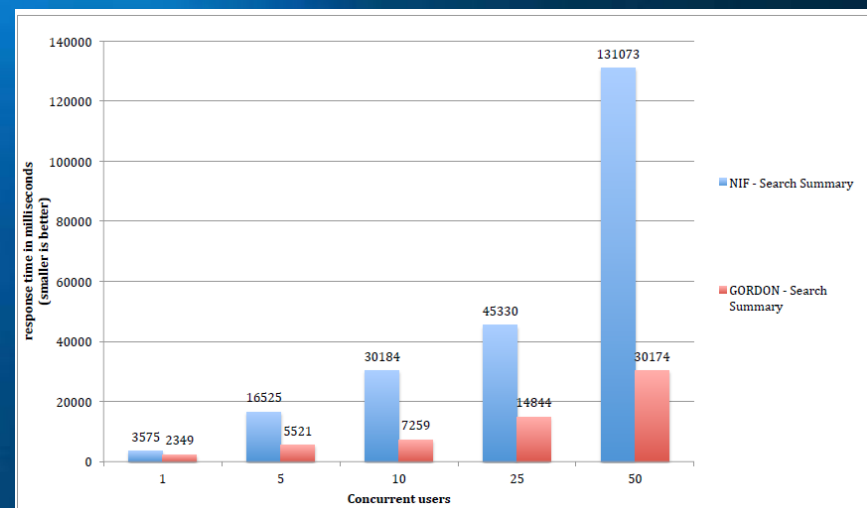
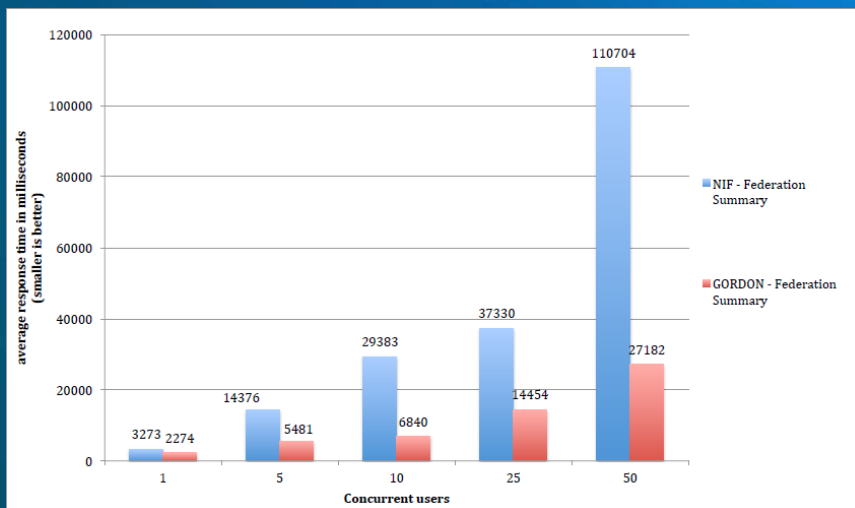
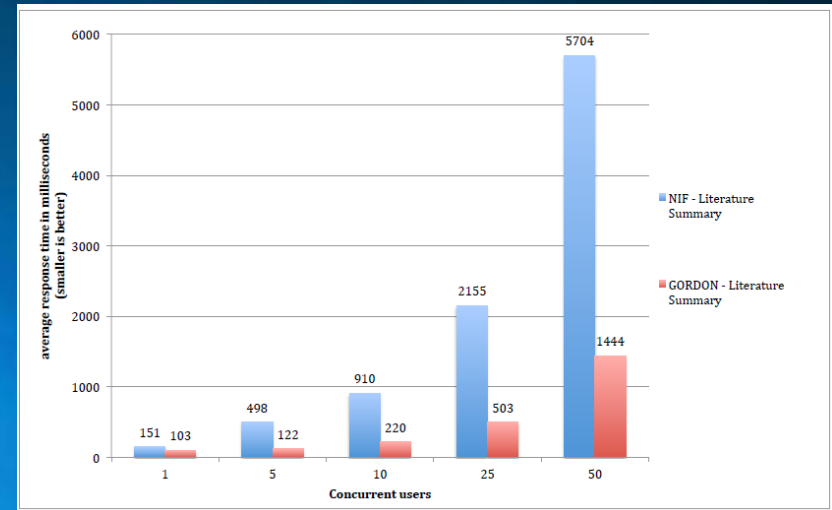
1	gene	projNum	projTitle
2	bptf	1ZIABC005263-28	Eukaryotic Chromatin Structure and Gene Regulation
3	bptf	1ZIABC005263-29	Eukaryotic Chromatin Structure and Gene Regulation
4	bptf	1ZIABC005263-31	Eukaryotic Chromatin Structure and Gene Regulation
5	bptf	1ZIABC005263-30	Eukaryotic Chromatin Structure 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