



VIRTUAL ASTRONOMICAL OBSERVATORY

# Semantics in Biology & Data Mining

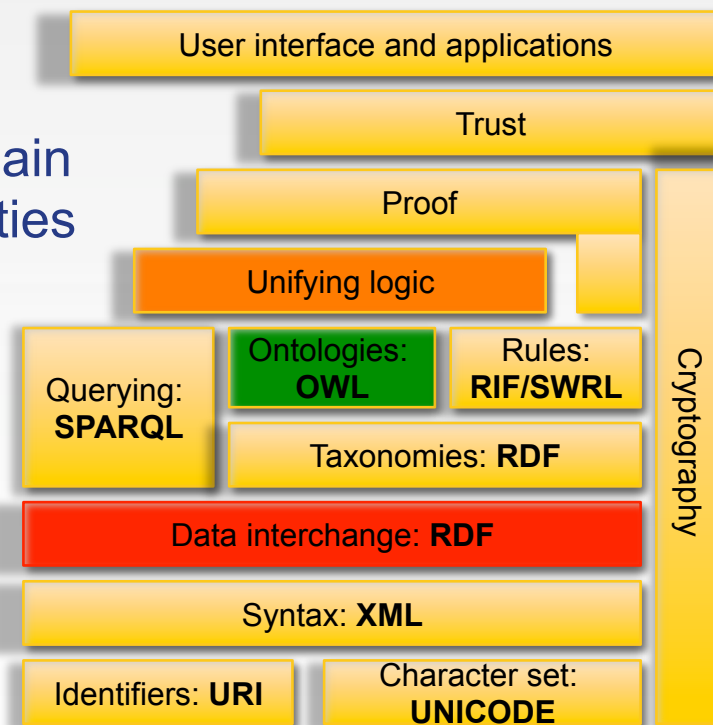
Matthew J. Graham, Caltech



The VAO is operated by the VAO, LLC.

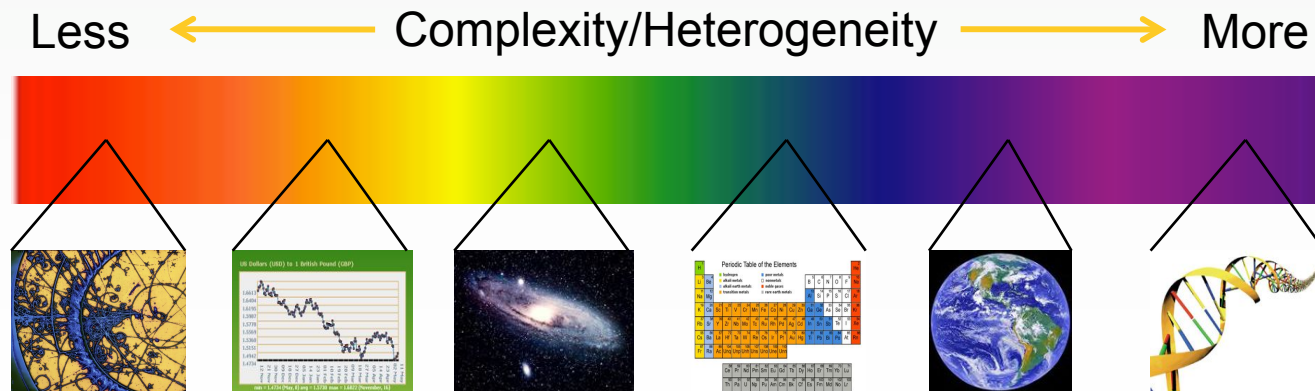
# What is a smart application?

- One built around technologies that *understand* data and *know* or can *infer* what to do with it
- What makes things smart?
  - **RDF**: all data can be represented as subject – predicate – object
  - **Ontology**: a conceptual model of domain knowledge in terms of classes, properties and relationships
  - **Description logic**: the backbone for inferencing and checking instances, relations, subsumption and concept consistency



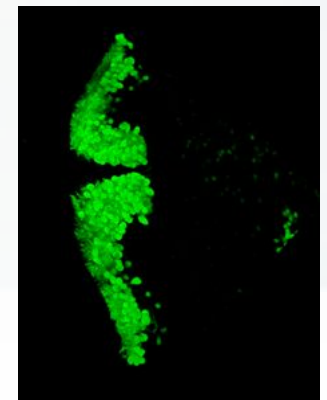
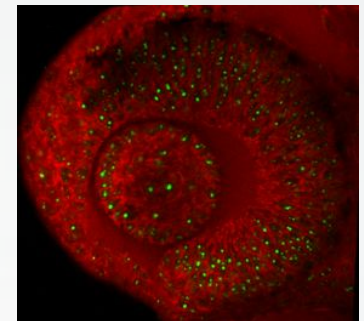
# Why is smartness prevalent in biology?

- X-informatics is the discipline of organizing, accessing, mining and analyzing information describing complex systems in (x = bio-, geo-, chemo-, astro-, econo-, ...)
- Bioinformatics was born in 1977 with the sequencing of the bacteriophage  $\Phi$ -X174
- Developments in genomic and information technology have produced a huge amount of complex and disparate *information*
- Smartness introduced via semantic technologies to address this



# The Zebrafish FlipTrap data repository

- A systems-based approach for analysis of gene function in developing vertebrate embryos in real time and space
- The FlipTrap screen is a gene trap that fuses the Citrine fluorescent protein to the trapped protein to generate a fully functional tagged version
- Expression patterns of the marked gene during development can then be imaged, etc.
- The data repository holds images, metadata, sequence data and annotations
- It makes extensive use of the Zebrafish anatomical ontology (2400 classes, 8 properties, 11038 entity annotation axioms) and the Gene ontology (30393 terms - 99.2% with definitions incl. 18939 biological process, 2735 cellular component and 8719 molecular function)





# Example: smart data entry

CALTECH  
ZEBRAFISH FLIPTR  
Center of Excellence in C

EGGS

Summary Image data Molecular data Functional data

Screen 1

Generation: F1

Day screened: Day 2

Stage: Gastrula (5.25-10.33 hrs)

Site of expression: hi

Choose one:

- hindbrain
- hindbrain commissure
- hindbrain interneuron

Add site Delete site

Subcellular expression:

- Cytoplasm
- Endoplasmic reticulum
- Extracellular matrix
- Golgi
- Membrane
- Mitochondria
- Not localized
- Nuclear envelope
- Nucleolus
- Nucleus

Additional comments:

No. of filesets to upload:

Add screen Delete screen

Only those anatomical structures defined by the ontology to be present at the selected "stage" are available for selection in the autocomplete drop-down.



# Example: smart querying

**ZEBRAFISH FLIPTRAP DATABASE**  
Center of Excellence in Genomic Science

SEARCH:

**BROWSE BY**

- Anatomical Location
- Stage
- Gene Name
- Allele
- Multiple Parameters

**DOCUMENTS**

- Contact Us

**RESOURCES**

- Additional Genomic Resources

**Multiple Parameter Search**

Expand one or more sections to enter search criteria and then click the Search button below.

Stage

<input type="checkbox"/> Zygote: 1-cell	<input type="checkbox"/> Gastrula	<input type="checkbox"/> Hatching
<input type="checkbox"/> Cleavage	<input type="checkbox"/> 50%-epiboly	<input type="checkbox"/> Long-pec
<input type="checkbox"/> 2-cell	<input type="checkbox"/> Germ-ring	<input type="checkbox"/> Pec-5n
<input type="checkbox"/> 4-cell	<input type="checkbox"/> Shield	<input type="checkbox"/> Larval
<input type="checkbox"/> 8-cell	<input type="checkbox"/> 75%-epiboly	<input type="checkbox"/> Protuding-mouth
<input type="checkbox"/> 16-cell	<input type="checkbox"/> 90%-epiboly	<input type="checkbox"/> Day 4
<input type="checkbox"/> 32-cell	<input type="checkbox"/> Bud	<input type="checkbox"/> Day 5
<input type="checkbox"/> 64-cell	<input type="checkbox"/> Segmentation	<input type="checkbox"/> Day 6
<input type="checkbox"/> Blastula	<input type="checkbox"/> 1-4 somites	<input type="checkbox"/> Days 7-13
<input type="checkbox"/> 128-cell	<input type="checkbox"/> 5-9 somites	<input type="checkbox"/> Days 14-20
<input type="checkbox"/> 256-cell	<input type="checkbox"/> 10-13 somites	<input type="checkbox"/> Days 21-29
<input type="checkbox"/> 512-cell	<input type="checkbox"/> 14-19 somites	<input type="checkbox"/> Juvenile
<input type="checkbox"/> 1k-cell	<input type="checkbox"/> 20-25 somites	<input type="checkbox"/> Days 30-44
<input type="checkbox"/> High	<input type="checkbox"/> 26+ somites	<input type="checkbox"/> Days 45-89
<input type="checkbox"/> Oblong	<input type="checkbox"/> Pharyngula	<input type="checkbox"/> Adult
<input type="checkbox"/> Sphere	<input type="checkbox"/> Prim-5	<input type="checkbox"/> Unknown
<input type="checkbox"/> Dome	<input type="checkbox"/> Prim-15	
<input type="checkbox"/> 30%-epiboly	<input type="checkbox"/> Prim-25	
	<input type="checkbox"/> High-pec	

Anatomical Location

<input type="checkbox"/> Zebrafish Anatomical Entity	<input type="checkbox"/> Anatomical Structure	<input type="checkbox"/> Portion Of Organism Substance
<input type="checkbox"/> Anatomical Line	<input type="checkbox"/> Acellular Anatomical Structure <input checked="" type="checkbox"/>	<input type="checkbox"/> Blood
<input type="checkbox"/> Groove <input checked="" type="checkbox"/>	<input type="checkbox"/> Anatomical Group <input checked="" type="checkbox"/>	<input type="checkbox"/> Cerebral Spinal Fluid
<input type="checkbox"/> Anatomical Space	<input type="checkbox"/> Cell <input checked="" type="checkbox"/>	<input type="checkbox"/> Dentine
<input type="checkbox"/> Bile Canaliculus	<input type="checkbox"/> Compound Organ <input checked="" type="checkbox"/>	<input type="checkbox"/> Enameloid
<input type="checkbox"/> Choroidal Fissure	<input type="checkbox"/> Embryonic Structure <input checked="" type="checkbox"/>	<input type="checkbox"/> Otolith <input checked="" type="checkbox"/>
<input type="checkbox"/> Coelom	<input type="checkbox"/> Extraembryonic Structure <input checked="" type="checkbox"/>	<input type="checkbox"/> Synovial Fluid
<input type="checkbox"/> Opercular Cavity	<input type="checkbox"/> Multi-tissue Structure <input checked="" type="checkbox"/>	<input type="checkbox"/> Unspecified
<input type="checkbox"/> Pericardial Cavity	<input type="checkbox"/> Organism Subdivision <input checked="" type="checkbox"/>	<input type="button" value="clear all"/> <input type="button" value="set all"/>
<input type="checkbox"/> Pleuropertoneal Cavity	<input type="checkbox"/> Portion Of Tissue <input checked="" type="checkbox"/>	
<input type="checkbox"/> Pupil	<input type="checkbox"/> Whole Organism <input checked="" type="checkbox"/>	

Terms and hierarchies generated dynamically from the ontology

Anatomical structures not present at the selected stage will be grayed out





# Example: smart results

Automatic links to literature about this gene

Full resolution images via appropriate viewer available at a click

Sequence information with exons highlighted and location of citrine marker indicated

**Gene Expression Report**

Allele	R122a
ZFIN	ZDB-GENE-980526-221
Ensembl	Ensembl
Gene Name	desm
Aliases	des   cb290   fb59a12   MGC109859   wu:fb59a12   cb290   des   fb59a12   wu:fb59a12   zgc:109859   desmin
Gene Description	desmin [Source:RefSeq peptide:Acc:NP_571036]
Gene Ontology	show
NCBI Nucleotide Hit	NM_130953 show details

Ensembl Gene Markers: 6987571

Comments: hide

mRNA sequence

```

CATTACACAGGGAC
ACCGCCACCTTTGG
GACTGACCTCCAGAT
TCCGAGGTGGCTCGG
AGGACTCCTCAACAC
AGGCGCCTTCCCGA
TTCCAGAGCTGTACGA
AGATCCAGAGGGACAA
AGCTGAAAACACCT
TCCAGGGTCTTACGA
AGCAGGTCAGGTGCA
ACGAGGCTATCCCTGC
TGAAACAGAAATACGA
GCCGATTTGACTCTCT
AGGCGGTGGTTATCA
TCCGGAGTACCAGGA
GAGAGGAGACAGGAT
ACCACACAGCAGGCA
GCCATGCCAGGTCCT
CAGAACTCCTGCATT
TGTGGATGACCATGG
CCTTCAGAAATGGCC

```

## Other smarts

- Suggestions
  - Search for related data products based on semantic similarity
- Environments
  - Virtual lab books linked to data and literature
  - Shared workflows with myExperiment.org
- Data mining
  - Incorporating domain knowledge into the discovery process

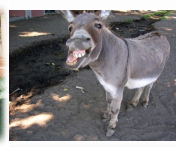




# Smart applications in astronomy?

- Linked data
    - ADS, CDS, NED
  - The Linnaean problem:
    - Linnaeus' original system in 1735 already had 6 levels of hierarchy: 3 kingdoms, 35 classes, orders, genera, species, subspecies. Is astronomical knowledge still too coarse-grained to warrant the depth of modelling that an ontology can provide?
    - Niche areas of taxonomic astronomy: solar system, exoplanets, supernovae?
- “The Eurybates family is a compact core inside the Menelaus clan, located in the L<sub>4</sub> swarm of Jupiter Trojans.”

– arXiv:1004.4180



# What use is semantics in KDD?

- Data mining is “the *semi-automatic* discovery of patterns, associations, changes, anomalies, and statistically significant structures and events in data”
- Such discoveries are **evaluated** (filtered) based on **relevance** (according to some metric of interestingness) and **content** (qualitative condition based on domain knowledge) constraints
- Traditionally the user assumes the responsibility of choosing which aspects of the domain knowledge are most important for the current task (hence *semi-automatic*)
- One of the ten challenging problems in data mining research is the incorporation of background or domain knowledge into the discovery process (Yang & Wu 2006)
- The main difficulty lies in representing and acquiring domain knowledge
- Ontologies are a viable construct for representing knowledge (OWL, SWRL, SPARQL/SQRWL)

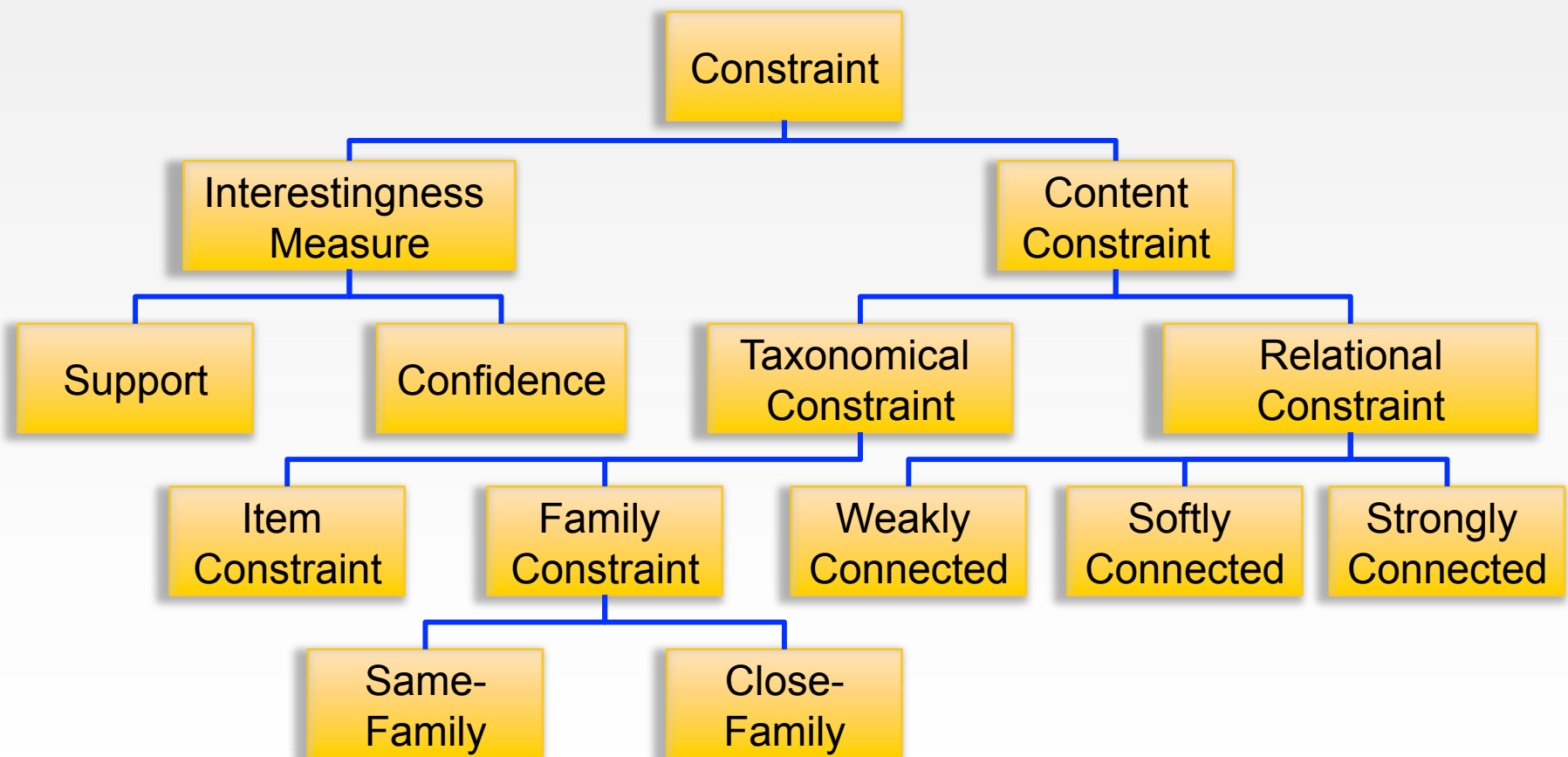


# Application ontologies

- Contains essential knowledge in order to drive data mining tasks
- Smart workflows
  - Recommender systems
  - Competitive intelligence tools
- OntoDM (<http://kt.ijs.si/panovp/OntoDM>):
  - dataset: data items
  - datatype: primitive, structured
  - data mining task: predictive modelling, pattern discovery, clustering, probability distribution estimation
  - generalization: predictive model, pattern, clustering, probability distribution
  - data mining algorithm: distance function, kernel function, refinement operator
  - function: aggregation function, prototype function, evaluation function, cost function
  - constraint: evaluation, language constraint
  - data mining scenario: query, inductive query

# Incorporating ontologies

- A simple way to incorporate an ontology into a data mining process is as a filter to prune those discoveries that do not meet the imposed constraint (derived from the ontology)



# Constraints

- A **constraint** is a predicate on the power set of the set of items  $I$ , that is, it is a function  $c: 2^I \rightarrow \{\text{true}, \text{false}\}$ . An itemset  $S$  is said to satisfy  $c$ , if and only if,  $c(S)$  is true.
- Interestingness metrics based on semantic similarity:
  - Edge counting: distance between ontology concepts
  - Information theoretic: information content of the lower common ancestor of two concepts
$$p_{ms}(c1, c2) = \min(\{p(c)\}) ; \text{sim}(c1, c2) = -\ln p_{ms}(c1, c2)$$
- Taxonomical based on family ties
  - {White dwarf, Massive} have same parent
  - {White dwarf/DA, Massive} have common ancestor and are at least  $n^{\text{th}}$  ( $n=1$ ) cousins to each other
- Relational based on relations between concepts
  - {Aperiodic, GRB, Massive} are weakly connected
  - No strongly connected itemsets



# Data mining with ontologies - I

- **Clustering:**

- Linkage-based:

- the similarity between two objects is measured based on the similarities between the objects linked with them

- Relational Fuzzy C-Means:

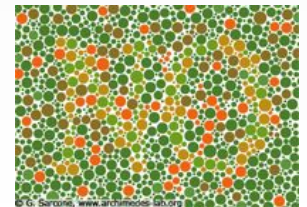
- processes  $n$  vectors in  $p$ -space as data input, and uses them, in conjunction with first order necessary conditions for minimizing the FCM objective functional, to obtain estimates for two sets of unknowns

- Correlation Cluster Validity

- Validate number of clusters by computing correlation between reconstruction matrix after fuzzy clustering and original dissimilarity matrix

- Ontological SOM

- Represent contribution of ontology term to description of associated node and replace distance metric with an ontology-based dissimilarity measure





# Data mining with ontologies - II

- **Detecting rare events via reasoning**

- Application of description-logic reasoning over an ontology to automate classification of instances into family and subfamily groups

- **Fuzziness**

- Markov Logic Networks – allows declarative domain knowledge to be expressed with real-valued weight indicating strength of statements



- **Association Rules**

- Discover strong rules between concepts/instances using different measures of interestingness

- **Network characterization**

- Establish functional relationships between instances and then predict functions and networks from these