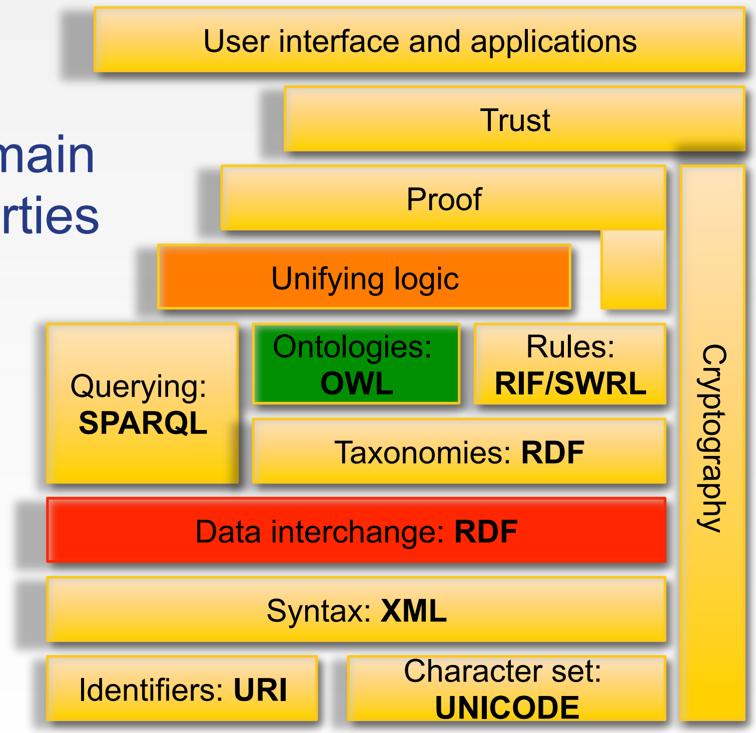
VIRTUAL ASTRONOMICAL OBSERVATORY

Semantics in Biology & Data Mining

Matthew J. Graham, Caltech

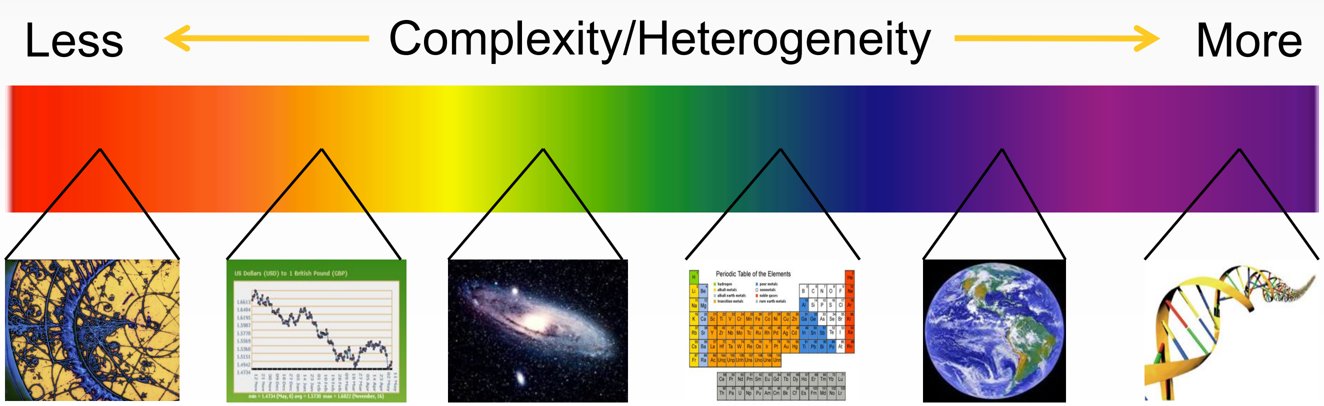
What is a smart application?

- 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



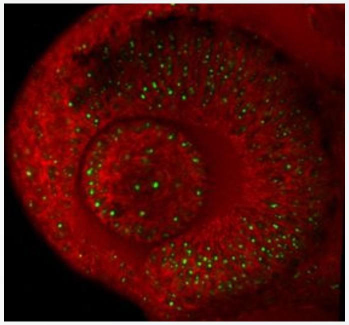
Why is smartness prevalent in biology?

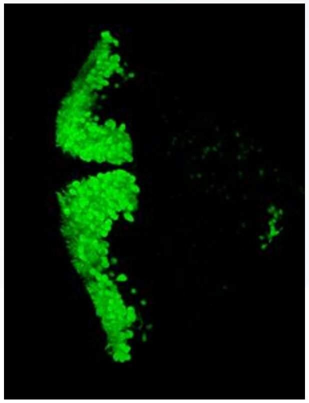
- X-informatics is the discipline of organizing, accessing, mining and analyzing information describing complex systems in (x = bio-, geo-, chemo-, astro-, econo-, …)
- Bioinformatics was born in 1977 with the sequencing of the bacteriophage Φ-X174
- Developments in genomic and information technology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



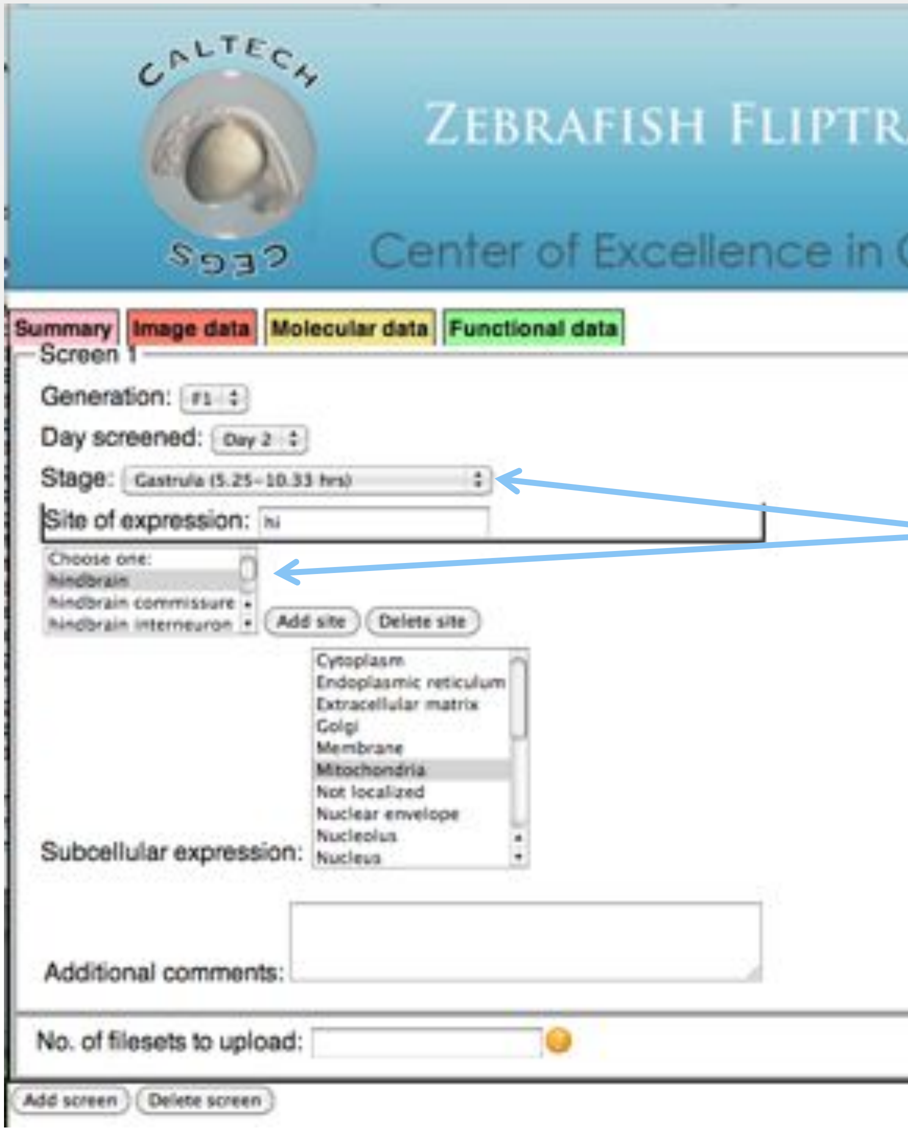
The Zebrafish FlipTrap data repository

- 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



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

Terms and hierarchies generated dynamically from the ontology

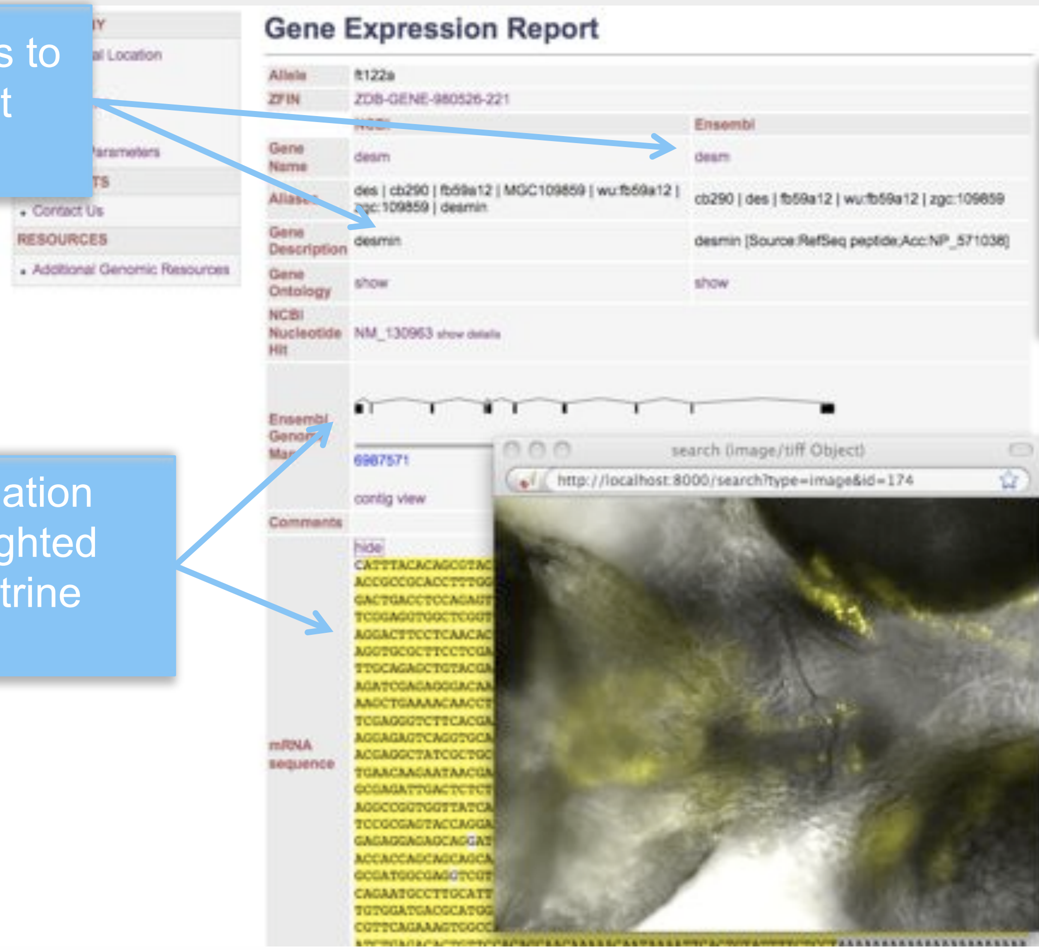
Anatomical structures not present at the selected stage will be grayed out

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



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 taxonometric astronomy: solar system, exoplanets, supernovae?

  "The Eurybates family is a compact core inside the Menelaus clan, located in the L_4 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*)

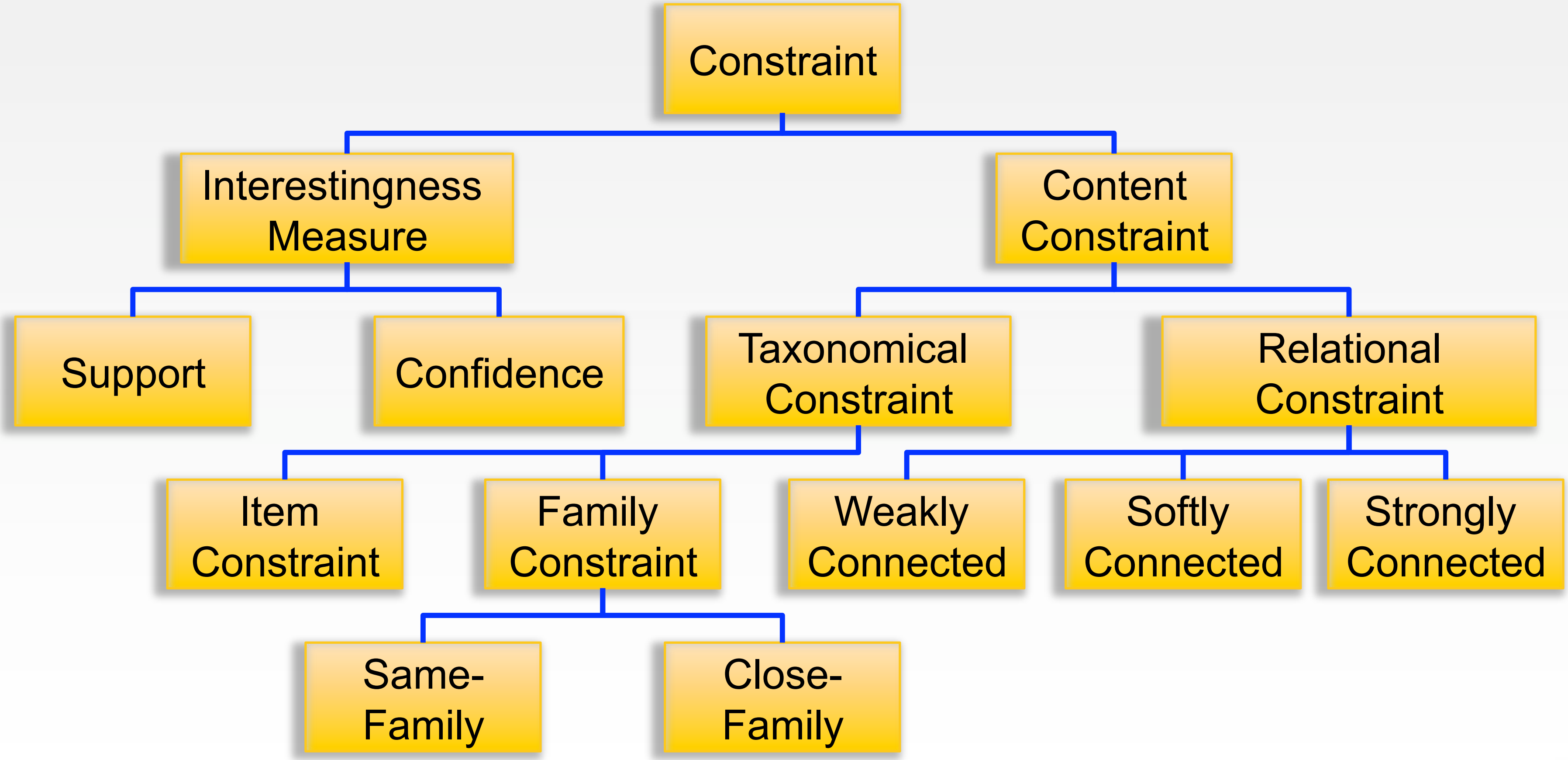
- One of the ten challenging problems in data mining research is the incorporation of background or domain knowledge into the discovery process (Yang & Wu 2006)
- The main difficulty lies in representing and acquiring domain knowledge
- Ontologies are a viable construct for representing knowledge (OWL, 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)



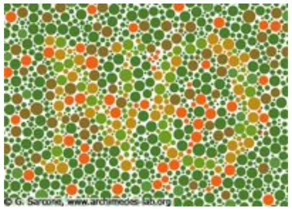
Constraints

- A **constraint** is a predicate on the power set of the set of items I, that is, it is a function *c*: $2^{\mathcal{I}}$ -> {true, 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

    $p_{ms}(c1,c2) = \min(\{p(c)\}$; $sim(c1, c2) = -\ln p_{ms}(c1, c2)$
- Taxonomical based on family ties
  - {White dwarf, Massive} have same parent
  - {White dwarf/DA, Massive} have common ancestor and are at least n^{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