Social Machines and Social Data

Peter Buneman
University of Edinburgh

Thanks to: Tony Harmar, Sarah Cohen Boulakia, Susan Davidson, Jamie Davies, Wenfei Fan, James Frew, Andreas Rauber, Joanna Sharman and Gianmaria Silvello

Social Machine???

“A social machine is an environment comprising humans and technology interacting and producing outputs or action which would not be possible without both parties present.”

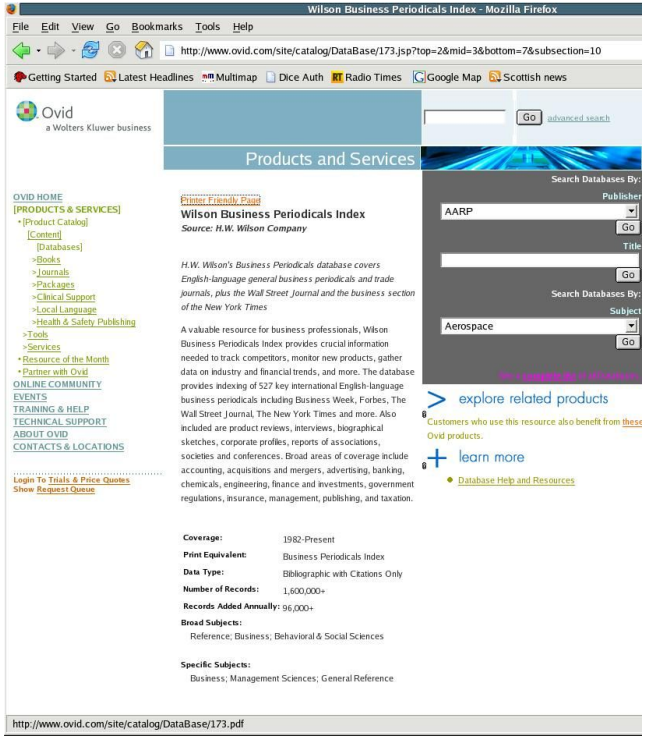
Examples:

Citizen science projects (Galaxy Zoo, SETI@home, QMC@home, butterfly counts, bird counts….). Certain forms of “crowdsourcing”

Social Media (Facebook, Twitter, Linkedin, Tumblr, ….) Newsgroups

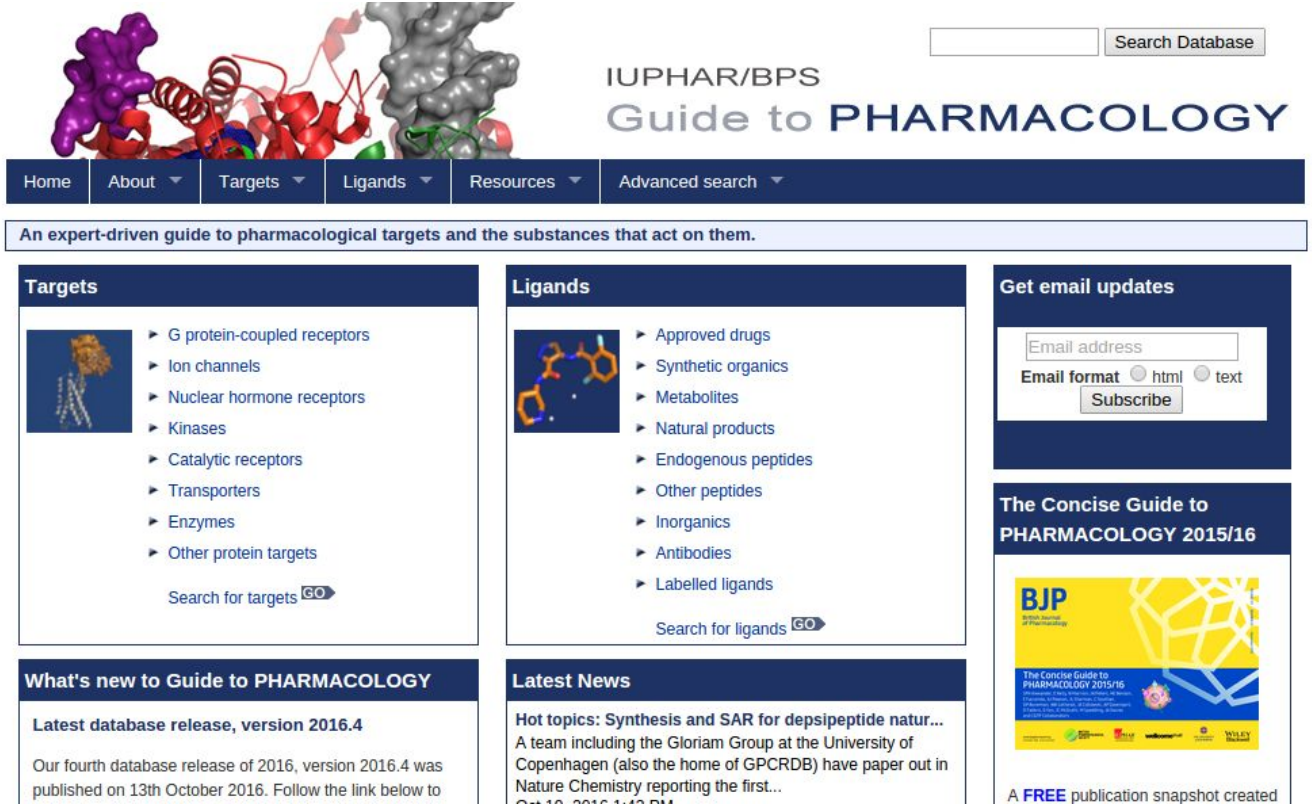
And curated databases (expert-sourcing)?

Curated databases?



- A curated database is one that is maintained with a lot of human effort
- Curare: Latin “to care for”
- Typically replacing reference works, encyclopedias, gazetteers, etc

GtoPdb: The leading curated database on pharmacological receptors (drugs)



Drilling down we find some text….

G protein-coupled receptors

Contents

Overview

« Hide

G protein-coupled receptors (GPCRs) are the largest class of membrane proteins in the human genome. The term "7TM receptor" is commonly used interchangeably with "GPCR", although there are some receptors with seven transmembrane domains that do not signal through G proteins. GPCRs share a common architecture, each consisting of a single polypeptide with an extracellular N-terminus, an intracellular C-terminus and seven hydrophobic transmembrane domains (TM1-TM7) linked by three extracellular loops (ECL1-ECL3) and three intracellular loops (ICL1-ICL3). About 800 GPCRs have been identified in man, of which about half have sensory functions, mediating olfaction (~400), taste (33), light perception (10) and pheromone signalling (5) [6]. The remaining ~350 non-sensory GPCRs mediate intersignalling by ligands that range in size from small molecules to peptide to large proteins; they are the targets for the majority of drugs in clinical usage [8,10], although only a minority of these receptors are exploited therapeutically. The first classification scheme to be proposed for GPCRs [4] divided them, on the basic of sequence homology, into six classes. These classes and their prototype members were as follows: **Class A** (rhodopsin-like), **Class B** (secretin receptor family), **Class C** (metabotropic glutamate), **Class D** (fungal mating pheromone receptors), **Class E** (cyclic AMP receptors) and **Class F** (frizzled/smoothened). Of these, classes D and E are not found in vertebrates. An alternative classification scheme "GRAFS" [11] divides vertebrate GPCRs into five classes, overlapping with the A-F nomenclature, *viz*:

Glutamate family (class C), which includes metabotropic glutamate receptors, a calcium-sensing receptor and $GABA_B$ receptors, as well as three taste type 1 receptors [class C list] and a family of pheromone receptors (V2 receptors) that are abundant in rodents but absent in man [6].

Rhodopsin family (class A), which includes receptors for a wide variety of small molecules, neurotransmitters, peptides and hormones, together with olfactory receptors, visual pigments, taste type 2 receptors and five pheromone receptors (V1 receptors). [Class A list]

Adhesion family GPCRs are phylogenetically realted to class B receptors, from which they differ by possessing large extracellular N-termini that are autoproteolytically cleaved from their 7TM domains at a conserved "GPCR proteolysis site" (GPS) which lies within a much larger (~320 residue) "GPCR autoproteolysis-inducing" (GAIN) domain, an evolutionary ancient mofif also found in polycystic kidney disease 1 (PKD1)-like proteins, which has been suggested to be both required and sufficient for autoproteolysis [9]. [Adhesion family list].

And then some “data”

Natural/Endogenous Ligands

adrenomedullin 2/intermedin {Sp: Human} , adrenomedullin 2/intermedin {Sp: Mouse} , adrenomedullin 2/intermedin {Sp: Rat}

amylin {Sp: Human} , amylin {Sp: Mouse, Rat}

calcitonin {Sp: Human} , calcitonin {Sp: Mouse, Rat}

α-CGRP {Sp: Human}

β-CGRP {Sp: Human} , β-CGRP {Sp: Mouse}

α-CGRP {Sp: Mouse, Rat}

β-CGRP {Sp: Rat}

Comments: Amylin, α-CGRP, and β-CGRP are the most potent endogenous agonists

Rank order of potency (Human)

calcitonin (salmon) ≥ amylin (*IAPP*, P10997) ≥ α-CGRP (*CALCA*, P06881) > adrenomedullin 2/intermedin (*ADM2*, Q7Z4H4) ≥ calcitonin (*CALCA*, P01258) > adrenomedullin (*ADM*, P35318)

Download all structure-activity data for this target as a CSV file Go

Agonists

Key to terms and symbols Click column headers to sort

Ligand	Sp.	Action	Affinity	Units	Reference
calcitonin (salmon)	Rn	Full agonist	9.0	pK_i	3
calcitonin {Sp: Human}	Hs	Full agonist	8.9 – 11.3	pEC_{50}	1,7,13
amylin {Sp: Mouse, Rat}	Hs	Full agonist	9.0 – 10.7	pEC_{50}	1,5,7,10
α-CGRP {Sp: Human}	Hs	Full agonist	8.7 – 10.8	pEC_{50}	7,10-11,13,20
pramlintide	Hs	Agonist	9.4 – 9.4	pEC_{50}	5
amylin {Sp: Human}	Hs	Full agonist	9.0 – 9.7	pEC_{50}	5
β-CGRP {Sp: Human}	Hs	Full agonist	9.2	pEC_{50}	7
Tyr^0α-CGRP (human)	Hs	Full agonist	7.6 – 9.5	pEC_{50}	7,10
[Cys(Et)2,7]α-CGRP (human)	Hs	Full agonist	7.8 – 8.4	pEC_{50}	7,10
adrenomedullin 2/intermedin {Sp: Human}	Hs	Full agonist	8.0	pEC_{50}	8
adrenomedullin {Sp: Human}	Hs	Full agonist	6.5 – 8.4	pEC_{50}	7,11

View species-specific agonist tables

Agonist Comments

The AMY_1 receptor is a heterodimeric complex of the calcitonin receptor and RAMP1 [15]. The variability in potency values reported is likely to reflect cell background such as the presence of other endogenous RAMPs and the calcitonin receptor-like receptor [18]. It is difficult to ascertain the contribution of such factors to the reported values.

View species-specific agonist tables

Agonist Comments

The AMY_1 receptor is a heterodimeric complex of the calcitonin receptor and RAMP1 [15]. The variability in potency values reported is likely to reflect cell background such as the presence of other endogenous RAMPs and the calcitonin receptor-like receptor [18]. It is difficult to ascertain the contribution of such factors to the reported values. Human amylin is rarely used because of its propensity to aggregate.

Antagonists

Key to terms and symbols **View all chemical structures** Click column headers to sort

Ligand	Sp.	Action	Affinity	Units	Reference
α-CGRP-(8-37) (human)	Hs	Antagonist	6.6	pK_B	7
AC187	Hs	Antagonist	8.0	pK_i	7
CT-(8-32) (salmon)	Hs	Antagonist	7.8	pK_i	7

Primary Transduction Mechanisms

Transducer	Effector/Response
G_s family	Adenylate cyclase stimulation

References: 3,7,10-11,13,15

Tissue Distribution

Lung > fundus (stomach) > spleen, brainstem, hypothalamus > liver, cortex, cerebellum.

Note: At present there is virtually no information on the co-localisation of CT with RAMP1. This data is based on the binding of [^{125}I]-amylin and so is an aggregate for AMY_1, AMY_2 and AMY_3 receptors.

Species: Rat
Technique: Radioligand binding.
References: 2

Functional Assays

Measurement of cAMP levels in COS-7 cells transfected with AMY_1 receptors (CT receptor plus RAMP1).

Species: Human
Tissue: COS-7 cells.
Response measured: cAMP accumulation.
References: 3,7

Physiological Functions

Amylin inhibits [^{14}C]glycogen accumulation in isolated skeletal muscle.

Species: Rat
Tissue: *Ex vivo*
References: 4,12

Curated databases are social machines

GtoPdb represents contributions and collaboration by over 1000 scientists worldwide. It is “expert-sourced”

Nearly every traditional reference work is now a curated database

Over 1000 curated databases in molecular biology alone.

Database topics from curated databases

* Data integration/transformation

* Data formats (pre and post XML)

* Data provenance

* Annotation

  Ontologies

* Data Citation

As well as all the other expected database topics

Annotation

Studied sporadically by DB community over 15 years [Bhagwat, Deepavali, et al. VLDB, 2004.]

Major question: propagation of annotation through queries (Provenance semirings [Tannen et al])

Increasing demand for practical annotation systems:

- Open up (e.g. GtoPDB) for general annotation
- Construct databases that consist of annotation (e.g. UNIPROT)

What is annotation? How is it different from any other data?

Annotation is the Communications Infrastructure of Social Machines

- Social machines mediate/assist human communication
  - Without this they would not be “social”
- The way we communicate using social machines differs from conventional communication (speech, letters, books, email, broadcast media etc.)
- Social machines provide some kind of framework to which we attach data
- The process of attaching data to that framework is *annotation*
- Examples ...

Facebook, Twitter, etc

Underlying structure: a massive graph with $O(10^9)$ nodes and $O(10^{11})$ edges representing social relationships (friend, follower etc)

Communication: adding data (messages, images, …) to that graph.

Other examples

Galaxy zoo: Underlying framework: (objects in) the celestial coordinate system

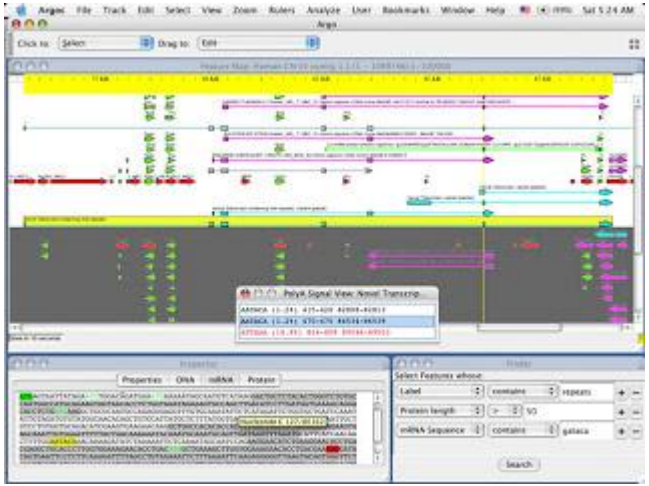
Citizen science: often some terrestrial coordinates (lat/long, postcodes,...)

Oxford English Dictionary: (Pre-computer) was largely crowdsourced. Annotation of English words.

GtoPdb: “We want to open up our database for external annotation”

Human Genome project

Scientists started to communicate through quasi-linear coordinate system of the human gene.



Tools were developed (Distributed Annotation Server) to allow scientists to communicate through a variety of GUIs

Curated databases

UNIPROT. The curators have a clear idea of "annotation" – value added by scientists

```
ID   143B_HUMAN     STANDARD;      PRT;   245 AA.
AC   P31946;
DT   01-JUL-1993 (REL. 26, CREATED)
DT   01-FEB-1996 (REL. 33, LAST SEQUENCE UPDATE)
DT   01-OCT-1996 (REL. 34, LAST ANNOTATION UPDATE)
DE   14-3-3 PROTEIN BETA/ALPHA (PROTEIN KINASE C INHIBITOR PROTEIN-1)
DE   (KCIP-1) (PROTEIN 1054).
GN   YWHAB.
OS   HOMO SAPIENS (HUMAN).
OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC   EUTHERIA; PRIMATES.
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=KERATINOCYTES;
RX   MEDLINE; 93294871.
RA   LEFFERS H., MADSEN P., RASMUSSEN H.H., HONORE B., ANDERSEN A.H.,
RA   WALBUM E., VANDEKERCKHOVE J., CELIS J.E.;
RL   J. MOL. BIOL. 231:982-998(1993).
. . .
```

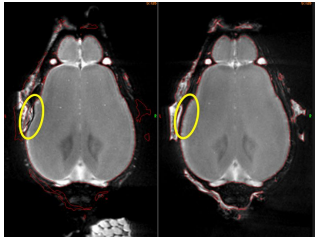
```
. . .
CC   -!- FUNCTION: ACTIVATES TYROSINE AND TRYPTOPHAN HYDROXYLASES IN THE
CC       PRESENCE OF CA(2+)/CALMODULIN-DEPENDENT PROTEIN KINASE II, AND
CC       STRONGLY ACTIVATES PROTEIN KINASE C. IS PROBABLY A MULTIFUNCTIONAL
CC       REGULATOR OF THE CELL SIGNALING PROCESSES MEDIATED BY BOTH
CC       KINASES.
CC   -!- SUBUNIT: HOMODIMER.
CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC.
CC   -!- TISSUE SPECIFICITY: 14-3-3 PROTEINS ARE LOCALIZED IN NEURONS, AND
CC       ARE AXONALLY TRANSPORTED TO THE NERVE TERMINALS. THEY MAY BE ALSO
CC       PRESENT, AT LOWER LEVELS, IN VARIOUS OTHER EUKARYOTIC TISSUES.
CC   -!- PTM: ISOFORM ALPHA DIFFERS FROM ISOFORM BETA IN BEING
CC       PHOSPHORYLATED (BY SIMILARITY).
CC   -!- ALTERNATIVE PRODUCTS: TWO FORMS ARE PRODUCED BY ALTERNATIVE
CC       INITIATION (BY SIMILARITY).
CC   -!- SIMILARITY: BELONGS TO THE 14-3-3 FAMILY OF PROTEINS.
DR   EMBL; X57346; G23114; -.
DR   MIM; 601289; -.
DR   PROSITE; PS00796; 1433_1; 1.
DR   PROSITE; PS00797; 1433_2; 1.
KW   BRAIN; NEURONE; PHOSPHORYLATION; ACETYLATION; MULTIGENE FAMILY;
KW   ALTERNATIVE INITIATION.
FT   INIT_MET     0     0      BY SIMILARITY.
FT   INIT_MET     2     2      IN SHORT FORM (BY SIMILARITY).
FT   MOD_RES      1     1      ACETYLATION (BY SIMILARITY).
FT   MOD_RES      2     2      ACETYLATION (IN SHORT FORM)
FT                             (BY SIMILARITY).
FT   MOD_RES    185   185      PHOSPHORYLATION (BY SIMILARITY).
SQ   SEQUENCE   245 AA;  27951 MW;  CE0EADFE CRC32;
     TMDKSELVQK AKLAEQAERY DDMAAAMKAV TEQGHELSNE ERNLLSVAYK NVVGARRSSW
     RVISSIEQKT ERNEKKQQMG KEYREKIEAE LQDICNDVLE LLDKYLIPNA TQPESKVFYL
     KMKGDYFRYL SEVASGDNKQ TTVSNSQQAY QEAFEISKKE MQPTHPIRLG LALNFSVFYY
     EILNSPEKAC SLAKTAFDEA IAELDTLNEE SYKDSTLIMQ LLRDNLTLWT SENQGDEGDA
        GEGEN
//
```

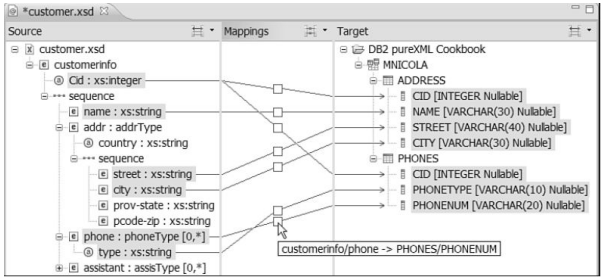
Mechanical Turk is not “Social”

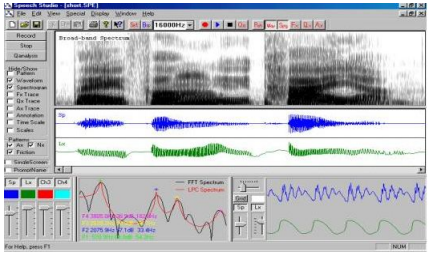
Does not really support human communication

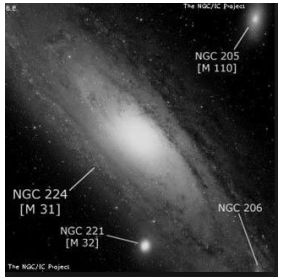
No clearly defined framework/coordinate system

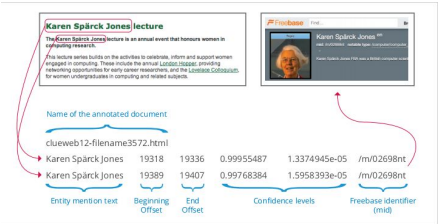
If people pumping computers for information is not a social machine why should computers pumping people be considered “social”?

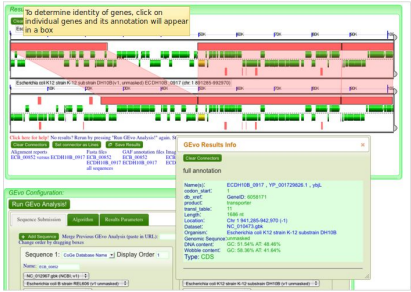


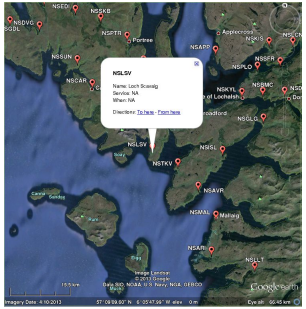

*customer.xsd
Source
Mappings
Target
customer.xsd
customerinfo
Cid : xs:integer
sequence
name : xs:string
addr : addrType
country : xs:string
sequence
street : xs:string
city : xs:string
prov-state : xs:string
pcode-zip : xs:string
phone : phoneType [0,*]
type : xs:string
assistant : assisType [0,*]
DB2 pureXML Cookbook
MNICOLA
ADDRESS
CID [INTEGER Nullable]
NAME [VARCHAR(30) Nullable]
STREET [VARCHAR(40) Nullable]
CITY [VARCHAR(30) Nullable]
PHONES
CID [INTEGER Nullable]
PHONETYPE [VARCHAR(10) Nullable]
PHONENUM [VARCHAR(20) Nullable]
customerinfo/phone -> PHONES/PHONENUM

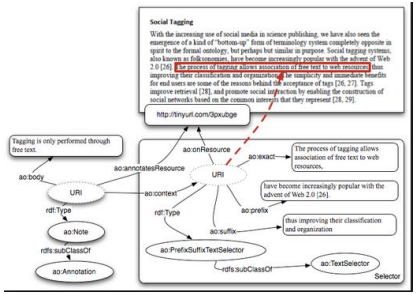


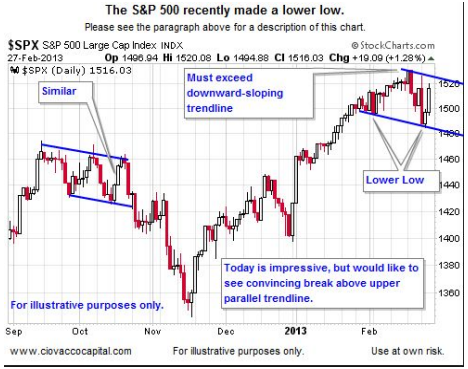

The NGC/IC Project
NGC 205
[M 110]
NGC 224
[M 31]
NGC 206
NGC 221
[M 32]
The NGC/IC Project


Karen Spärck Jones lecture
Freebase
Name of the annotated document
clueweb12-filename3572.html
Karen Spärck Jones 19318 19336 0.99955487 1.3374945e-05 /m/02698nt
Karen Spärck Jones 19389 19407 0.99768384 1.5958393e-05 /m/02698nt
Entity mention text
Beginning Offset
End Offset
Confidence levels
Freebase identifier (mid)


To determine identity of genes, click on individual genes and its annotation will appear in a box
GEvo Results Info
GEvo Configuration:
Run GEvo Analysis!


NSLSV
Name: Loch Scaraig
Service: NA
When: NA
Directions: To here - From here
Google earth


Social Tagging
http://tinyurl.com/3pxubge
Tagging is only performed through free text.
ao:body
URI
ao:annotatesResource
ao:onResource
URI
ao:exact
The process of tagging allows association of free text to web resources,
ao:context
have become increasingly popular with the advent of Web 2.0 [26].
rdf:Type
ao:prefix
ao:Note
ao:suffix
thus improving their classification and organization
rdfs:subClassOf
ao:PrefixSuffixTextSelector
ao:Annotation
rdfs:subClassOf
ao:TextSelector
Selector


The S&P 500 recently made a lower low.
Please see the paragraph above for a description of this chart.
$SPX S&P 500 Large Cap Index INDX
©StockCharts.com
27-Feb-2013 Op 1496.94 Hi 1520.08 Lo 1494.88 Cl 1516.03 Chg +19.09 (+1.28%)
$SPX (Daily) 1516.03
Similar
Must exceed downward-sloping trendline
Lower Low
Today is impressive, but would like to see convincing break above upper parallel trendline.
For illustrative purposes only.
Sep
Oct
Nov
Dec
2013
Feb
www.ciovaccocapital.com
For illustrative purposes only.
Use at own risk.

Annotation of databases

Here the “coordinate system” or “framework” is a database (database = any evolving structured collection of data: relational, XML, JSON, RDF)

So annotation is the *attachment of data to existing data*

- How do we specify that attachment?
- How is annotation different from adding data?
- What happens to the annotation if the underlying database changes?
- How does the annotation propagate through a query?
- Do annotations have structure, or are they “opaque”?

Does annotation have structure?

Annotating with comments

Bill is underpaid

Bill likes Mary

Mary likes champagne

Emps:

Id	Name	Sal	Dept
123456	Joe	40k	Sales
123321	Bill	20k	Research
654321	Mary	50k	Research

Depts:

Dept	Manager	Budget
Research	Mary	500k
Sales	Jane	800k

```
SELECT Name, Manager
FROM Emps, Depts
WHERE Emps.Dept = Depts.Dept
AND Id = 123321
```

Name	Manager
Bill	Mary

Bill is underpaid

Bill likes Mary

Mary likes champagne

We probably want the *union* of the comments on the input

Annotating with beliefs: the people who *believe* a tuple to be true

Emps:

Id	Name	Sal	Dept
123456	Joe	40k	Sales
123321	Bill	20k	Research
654321	Mary	50k	Research

{Jean, Sue, Tim}

Depts:

Dept	Manager	Budget
Research	Mary	500k
Sales	Jane	800k

{Sue, Tim, Bob}

```
SELECT Name, Manager
FROM Emps, Depts
WHERE Emps.Dept = Depts.Dept
AND Id = 123321
```

Name	Manager
Bill	Mary

{Sue, Tim}

We want the *intersection* of the believers of the input tuple

Annotating with beliefs for another query:

Emps:

Id	Name	Sal	Dept
123456	Joe	40k	Sales
123321	Bill	20k	Research
654321	Mary	50k	Research

{Jean, Sue, Tim}

Depts:

Dept	Manager	Budget
Research	Mary	500k
Sales	Jane	800k

{Sue, Tim, Bob}

```
SELECT Name
FROM Emps
UNION
SELECT Manager
FROM Dept
```

Name
Joe
Bill
Mary
Jane

{Jean, Sue, Tim, Bob}

For UNION queries we want the *union* of the believers of the input tuples

Provenance/Annotation Semirings (Tannen atelier: PODS ’07, ‘08 & '11)

R:

a	b	c	p
d	b	e	r
f	b	e	s

V:

a c	$p + (p \cdot p)$
a e	$p \cdot r$
d c	$r \cdot p$
d e	$r + (r \cdot r) + (r \cdot s)$
f e	$s + (s \cdot s) + (s \cdot r)$

$V(X, Z) :- R(X, \_, Z)$

$V(X, Z) :- R(X, Y, \_), R(\_, Y, Z)$

Tuples are created by :

“joining” other tuples (join): $p \cdot r$

“merging” other tuples (project and union): $p + r$

Both the “$\cdot$” and “+” are commutative and associative,

“$\cdot$” distributes over “+”: $p \cdot (r + s) = (p \cdot r) + (p \cdot s)$

Provenance semirings describe how (tuple) annotations combine and propagate through queries.

They provide an elegant generalization of things we have been studying: bag semantics, c-tables, probabilistic data, why-provenance …

We also need them later in the talk

Annotation is the attachment of data to existing data

But *how* is the annotation data attached? To what *part* of the database

- [Bhagwat, et al. VLDB, 2004.] – values in a table
- [Tannen atelier] – tuples
- [Geerts *et al. Mondrian,* ICDE 2006] – “rectangular” subtables (select/project queries)
- [Buneman *et al,* TODS 2008] – values, tuples, tables,... in a nested relational model.

But *how* is the annotation data attached? To what *part* of the database.
In general we’d like to attach an annotation to a *view*

And an annotation propagates through a query if the view can be computed from the query!!!

This turned out to be nice but too general. (But we’ll use the idea later)

Some annotations that the GtoPdb pharmacologists want (translated into terms we can understand)

Id	Name	Shoesize	Waistline	...
1234	Joe	6	38	...
9876	Jane	7	28	...

Annotation

What is being annotated, and when is the annotation valid?

Example 1. *Annotation* = “Joe’s shoesize is bigger than 6”

How do we identify the tuple?

SELECT … FROM R WHERE Name = “Joe”
SELECT … FROM R WHERE Id = 1234
SELECT … FROM R WHERE Id = 1234 AND Name = “Joe” AND Shoesize = 6 AND Waistline =38 AND...

What part of the tuple is being annotated?

SELECT Shoesize FROM R WHERE … ? Not really what we want.

When is it valid?

SELECT … FROM R WHERE … AND Shoesize ≤ 6

Id	Name	Shoesize	Waistline	...
1234	Joe	6	38	...
9876	Jane	7	28	...

Annotation

- There is no reason to expect that we can express everything in SQL, but remember that SQL is the *only* access method for RDBs, so it's going to figure.
- Any method of specifying *what* is being annotated is probably going to specify a *set* but the annotations apply to members of that set.

Example 2. *Annotation* = "6 looks like a US or UK shoe size"
How do we identify the tuple?

```
SELECT … FROM R WHERE Shoesize = 6
```

Example 3. *Annotation* = "Shoesizes are generally greater than the square root of the Waistline"
How do we identify the tuple?

```
SELECT … FROM R WHERE Shoesize*Shoesize <= Waistline
```

Nothing remarkable about this, but the annotation could be on *both* Shoesize and Waistline

Example 4. *Annotation* = "The average shoesize is 6.5"
Although about a set, it might be appropriate to attach it to an individual tuple.

So what do we learn from shoe sizes?

We need a way of specifying what parts of a tuple are being annotated.

We need to specify conditions under which the “part” receives an annotation and what happens if the database changes.

We didn’t ask where we physically store the annotation. It would be nice if we could put it in the DB itself, but an RDB schema makes this difficult. We need to treat things like column names as values.

The last remark suggests that we might profitably look at schema-less data models (JSON, RDF…)

A possible semistructured model: nested terms

Believes(John, Likes(Lucy, Cheese))

Comment(James, Likes(Lucy, Cheese)), “but not smelly cheese”)

Underlying data is in black, annotation is in blue, and annotation is indicated by *nesting*. Attachment is always to a term.

Annotations on annotations are easy

These examples indicate that we can (and should) have several “kinds” of annotation, but for the time being we’ll use just one kind, Annot, e.g.
Annot(Likes(Lucy, Cheese), “so does Jane”)

Using an RDF-like representation

Id	Name	Shoesize	Waistline	...
1234	Joe	6	38	...
9876	Jane	7	28	...

Annotation

{ Name(1234, Joe), Shoesize(1234, 6), Waistline(1234, 38)
Name(9876, Jane, Shoesize(9876, 7), Waistline(9876, 28) }

Annot(Shoesize(1234, 6), “6 is too low”) ⟵ Shoesize(1234, 6)

or maybe

Annot(Shoesize(1234, x), Too-low(x)) ⟵ Shoesize(1234, x) $\wedge$ $x \leq 6$

Annotations are specified by rules

So why not?

- Nobody uses a nested term model
- What we have “invented” is (syntactically) Prolog. It may be highly constrained, but we could still have infinite recursion, e.g.,

  Believes(x, Believes(x,y)) ←Believes(x,y).
- [B. Kostylev, Vansummeren ICDT 2014] Annotations are Relative. Database is large graph of nested terms.

However, in RDF it is now becoming common to treat the graph “name” (the 4th column) as an identifier for a single triple.

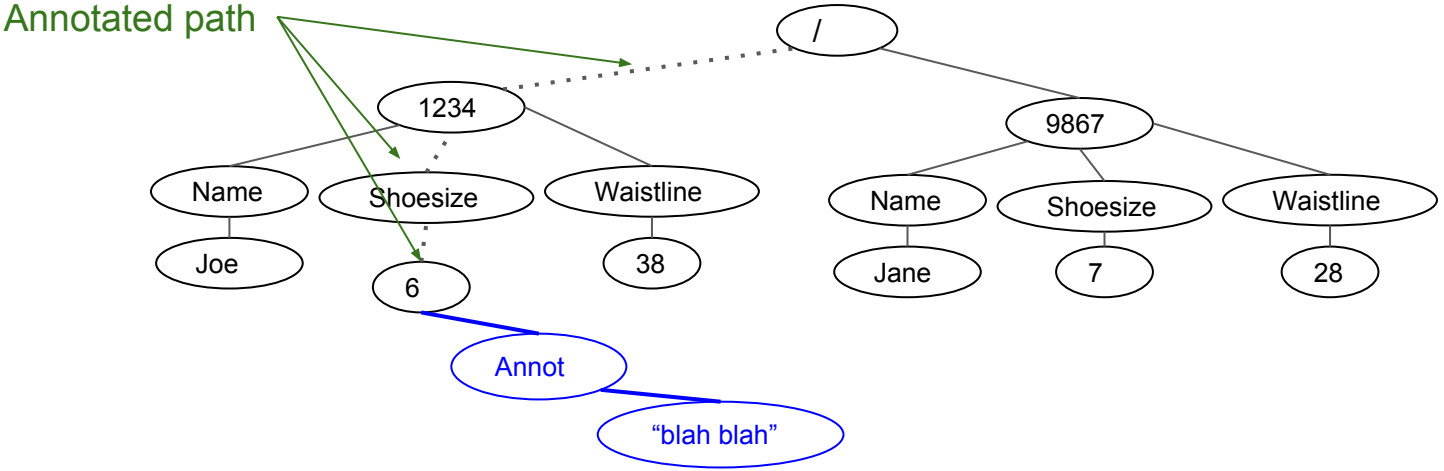
This is almost equivalent to a nested term model

Another approach: annotate hierarchies

{ 1234: {Name: Joe, Shoesize: 6, Waistline: 38},
  9876: {Name: Jane, Shoesize: 7, Waistline: 28}}

Id	Name	Shoesize	Waistline	...
1234	Joe	6	38	...
9876	Jane	7	28	...

Annotation



So what does an annotation rule for JSON Look like?

It has to specify a path (or set of paths) to be annotated. XPath does this so maybe something like

This represents the simplest form of annotation: clicking on something and adding text

/R/1234/Shoesize/6 :+ {Comment: “Too low”}

/R/*[Name/Joe]/Shoesize/6 :+ {Comment: “Too low for Joe”}

/R/y[Name/Joe]/Shoesize/x, x ≤ 6:+ {Comment: “Too low for Joe”}

/R/y[Name/Joe]/Shoesize/x, x ≤ 30 :+ {Comment: {Not-European: x}}

The first two are (more or less) standard XPath on the left with JSON on the right. We have added variables and conditions to the last two.

Constituents of a hierarchical annotation language

/R/y[Name/Joe]/Shoesize/x, x ≤ 30 :+ {Comment: {Not-European: x}}

XPath⁻ with variables

Optional condition

JSON

The only interesting question is how do we interpret XPath⁻ with variables.

Idea: Think of a JSON tree as a set of *paths* – a prefix-closed set of sequences of labels and values.

Given a JSON tree ($T \subseteq \mathcal{L}^*$), the meaning of an XPath⁻ expression E is an assignment of a *set of substitutions* (of variables in E to labels) to paths in T. If E contains no variables then we have an ordinary XPath expression which assigns

- {} --The empty set, if the node is *not* in the result of E
- {{}} – The set containing the empty substitution, if the node is in the result of E

Syntax of XPath$^-$: l ranges over labels in $\mathcal{L}$; v over variables.

$$q ::= . \mid l \mid v \mid q/q \mid q//q \mid q[q]$$

If S_1 and S_2 are substitutions, which agree on their common variables, their join $\bowtie$ is the substitution which maps all their variables to the appropriate label. Extend the join to sets in the obvious way:

$$S_1 \bowtie S_2 \quad = \quad \{s \mid s = s_1 \bowtie s_2 \text{ for } s_1 \in S_1, s_2 \in S_2\}$$

The other operation we need on substitution sets is union

We can now write down the evaluation rules [[Q]]$T(p)$ which give the set of substitutions produced by the query Q on the path p in the JSON tree T

$$\begin{array}{rcl}
[\![.]\!]_T(\emptyset) & = & \{\{\}\} \\
[\![a]\!]_T(a) & = & \{\{\}\} \\
[\![x]\!]_T(a) & = & \{\{x : a\}\} \\
[\![Q/Q']\!]_T(p) & = & \bigcup\{[\![Q]\!]_T(p_1) \bowtie [\![Q']\!]_{T|p_1}(p_2) \mid p = p_1p_2\} \\
[\![Q//Q']\!]_T(p) & = & \bigcup\{[\![Q]\!]_T(p_1) \bowtie [\![Q']\!]_{T|p_1r}(p_2) \mid p = p_1rp_2\} \\
[\![Q[Q']]\!]_T(p) & = & [\![Q]\!]_T(p) \bowtie \bigcup\{[\![Q']\!]_{T|p}(r) \mid r \in T \mid p\}
\end{array}$$

Nice properties

- Evaluation rules “well-defined”
- PTIME data complexity
- Efficient in practice (very efficient without //)
- Each substitution set binds all the (relevant) variables (no disjunction)
- Efficient (time and space) *incremental* & *external* evaluation (under investigation)
- XPath^{-} allows us to express both the “attachment point(s)” and the conditions, and
- seems to express what the GtoPDB pharmacologists want.

Some of these properties depend on the model being JSON (nested dictionary/ deterministic) not XML.

[Hidders *et al.* PODS 2017] “logical foundations” of JSON querying. Similar set up to ours, but includes *path variables.*

Conclusions on annotation

Fundamental observation is that annotations are *rules.*

- Maybe very simple rules (e.g. the thing being annotated has to exist), but still rules
- This view may also support annotation privacy etc.

Annotation requires some kind of semistructured/schema-less data model.

People who build social machines/curated databases would benefit greatly from generic annotation tools. Annotation propagation (~ provenance) is critical.

Data citation

GtoPdb is a reference work, created by a thousand or more academics around the world who contribute material to it.
But it's also a database. You can:

- See it in HTML pages
- Run SQL on it
- Run SPARQL on the RDF representation

Question posed by Tony Harmar 10 years ago:
How do I get people to cite GtoPdb?

The academics should get the same credit that they get for any other publication

Increasing demand for data citation

Large number of organizations: Datacite DataONE, GEOSS, D-Lib Alliance, DCC, COPDES, Force-11, AGU, ESIP, DCMI, CODATA, ICSTI, IASSIST, ICSU

Force 11: “Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.”

DataCIte: “We believe that you should cite data in just the same way that you can cite other sources of information, such as articles and books.”

Amsterdam Manifesto: “Data should be considered citable products of research.”

Oxford University (on behalf of EPSRC) “Describe your data ... to enable other researchers to … cite them”

What is a (conventional) citation?

A collection of “snippets” of information: authors, title, date, etc. and some kind of access mechanism (DOI, URL, ISBN, shelf number etc.) Something like this [2]

Not exactly provenance

Self contained, immutable (to within some choice of format)

Needed for a variety of reasons: kudos, currency, authority, recognition, access…

[2] Blondel, V. D., Gajardo, A., Heymans, M., Senellart, P., & Van Dooren, P. (2004). A measure of similarity between graph vertices: Applications to synonym extraction and web searching. *SIAM review*, *46*(4), 647-666.

So what's the problem?

Citations vary with what part of of the database is being cited.

There is a huge (maybe infinite) number of "parts" of a database, the "part" being defined by some database query

Web	URI/CGI
RDB	SQL
XML	XPath/XQuery
RDF	SPARQL
File system	set of paths

We cannot expect to put a citation for each "part" into DBLP. We are going to have to generate citations on the fly. And we can't expect the authors to do it.

It gets worse

Start of a 700 line SQL component of some OLAP API

```
SELECT /*+ NOPARALLEL bypass_recursive_check */
SP_ALIAS_190,
((CASE SP_ALIAS_191
WHEN 1
THEN 'PROVIDER::ALL_PROV::'
WHEN 0
THEN 'PROVIDER::PROV::'
ELSE NULL END) || SP_ALIAS_190) ALIAS_3553,
SP_ALIAS_194,
SP_ALIAS_191,
SP_ALIAS_192,
SP_ALIAS_193,
SP_ALIAS_205,
D4_AGE_GROUP_ET,
((CASE D4_AGE_GROUP_GID
WHEN 1
THEN 'AGE_GROUP::ALL_AGE_GRP::'
WHEN 0
```

Start of Datacite 400 line XML schema specification for citation

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- Revision history
    2010-08-26   Complete revision according to new common specification by the metadata work group after review. AJH, DTIC
                2010-11-17 Revised to current state of kernel review, FZ, TIB
                2011-01-17 Complete revsion after community review. FZ, TIB
                2011-03-17 Release of v2.1: added a namespace; mandatory properties got minLength; changes in the definitions of relationTypes
                IsDocumentedBy/Documents and isCompiledBy/Compiles; changes type of property "Date" from xs:date to xs:string. FZ, TIB
                2011-06-27 v2.2: namespace: kernel-2.2, additions to controlled lists "resourceType", "contributorType", "relatedIdentifierType", and "descriptionType". Removal of intermediate include-files.
    2013-05 v3.0: namespace: kernel-3.0; delete LastMetadataUpdate & MetadateVersionNumber; additions to controlled lists "contributorType", "dateType", "descriptionType", "relationType", "relatedIdentifierType" & "resourceType"; deletion of "StartDate" & "EndDate" from list "dateType" and "Film" from "resourceType";  allow arbitrary order of elements; allow optional wrapper elements to be empty; include xml:lang attribute for title, subject & description; include attribute schemeURI for nameIdentifier of creator, contributor & subject; added new attributes "relatedMetadataScheme", "schemeURI" & "schemeType" to relatedIdentifier; included new property "geoLocation"
                2014-08-20 v3.1: additions to controlled lists "relationType", contributorType" and "relatedIdentifierType"; introduction of new child element "affiliation" to "creator" and "contributor"-->
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
xmlns="http://datacite.org/schema/kernel-3" targetNamespace="http://datacite.org/schema/kernel-3"
elementFormDefault="qualified" xml:lang="EN">
                <xs:import namespace="http://www.w3.org/XML/1998/namespace"
schemaLocation="http://www.w3.org/2009/01/xml.xsd"/>
                <xs:include schemaLocation="include/datacite-titleType-v3.xsd"/>
                <xs:include schemaLocation="include/datacite-contributorType-v3.1.xsd"/>
                <xs:include schemaLocation="include/datacite-dateType-v3.xsd"/>
                <xs:include schemaLocation="include/datacite-resourceType-v3.xsd"/>
                <xs:include schemaLocation="include/datacite-relationType-v3.1.xsd"/>
                <xs:include schemaLocation="include/datacite-relatedIdentifierType-v3.1.xsd"/>
                <xs:include schemaLocation="include/datacite-descriptionType-v3.xsd"/>
                <xs:element name="resource">
```

Another principle/recommendation

Unless we couple the process of generating a citation with the act of extracting the data, the advocacy of data citation is pointless.

The main problem

Given a database D and a query Q, generate an appropriate citation.

NB. The citation depends on *both* Q and D

The database problem

Looks hard because any analysis of a query is likely to be hard, if not undecidable, but there's hope.

Key idea: *It is common for authors/publishers to formulate citations for some "parts" of the database.* These are views $V_1 \ldots V_n$. So given a query Q, can it be factored through a view? That is, is there a Q_i and V_i such that

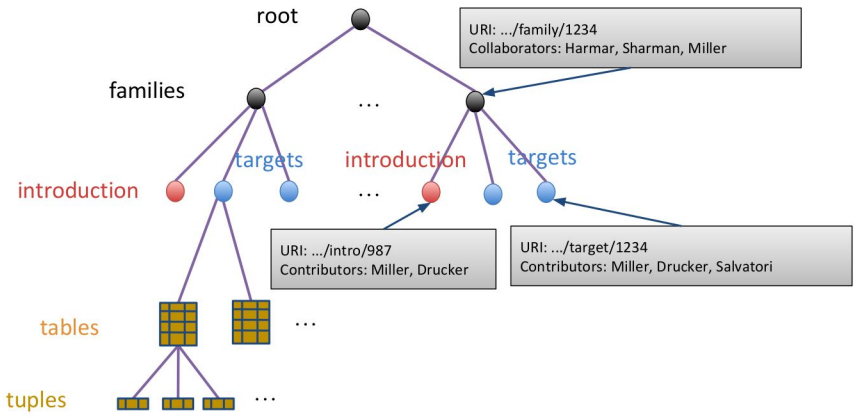
$$\forall D \in S.\ Q(D) = Q_i(V_i(D))$$

If so, the citation for V_i is a possible citation for Q.

This is a well-known database problem that comes from optimization. In fact our problem is a bit more subtle because the citation also depends on D, and we have to introduce the notion of a *parameterized* view. But the known machinery can be adapted. Can also be formulated for SPARQL & XQUERY

Hierarchical data (files, XPath, some URLs)

A simple pattern-matching language
for generating citations in a hierarchy



{ DB: IUPHAR, Version: $v, Family: $$f, Contributors: $a,
URI: ”www.iuphar.org”, DOI: 10.3.14159}
←
/Root[VersionNumber: $v]/Family[FamilyName: $$f]
/Introduction[Contributor-list: $a]

{ DB: IUPHAR, Version: 26, Family: ”Calcitonin”, Contributors: [”Debbie Hay”, ”David R. Poyner”], URI: ”www.iuphar.org”, DOI: 10.3.14159}

Type:	Nonsense mutation
Species:	Human
Description:	Rare variant identified in attention-deficit hyperactivity disorder (ADHD) patient, premature STOP codon with impaired cell surface expression and cAMP inhibition
Amino acid change:	Y170X
Nucleotide accession:	NM_005958
Protein accession:	NP_005949
References:	15

Type:	Missense mutation
Species:	Human
Description:	Common variant identified in control population with reduced ERK1/2 activation
Amino acid change:	A266V
Nucleotide accession:	NM_005958
Protein accession:	NP_005949
References:	16

General Comments

The molecular pharmacology of ovine melatonin receptors has been shown to be different to human recombinant melatonin receptors [49].

Available Assays

DiscoveRx	OPEN ECN PathHunter® eXpress MTNR1A CHO-K1 β-Arrestin GPCR Assay *(Cat no. 93-0510E2CP0M)* PathHunter® CHO-K1 MTNR1A β-Arrestin Cell Line *(Cat no. 93-0951C2)*	more info

References

Show »

How to cite this page

Philippe Delagrange, Margarita L. Dubocovich, James Olcese.
Melatonin receptors: MT_1 receptor. Last modified on 29/06/2015. Accessed on 21/09/2015. IUPHAR/BPS Guide to PHARMACOLOGY,
http://www.guidetopharmacology.org/GRAC/ObjectDisplayForward?objectId=287.

But views may have order and citations may have structure

Views can be ordered. $V_i \leq V_j$ if $\exists F. \forall D \in S.\ V_i(D) = F(V_j(D))$

This is the hierarchical ordering in GtoPdb, and the rule is always to choose the “least” or “finest” citation. (Cite the paper not the journal)

What happens if a citation requires the conjunction or disjunction of views?

- “The <u>calcitonin receptors</u> show greater blahblah that the <u>melatonin receptors”</u> (conjunction needed)
- This phenomenon is seen both in <u>calcitonin receptors</u> and <u>melatonin receptors</u> (disjunction needed)

Sounds like semiring provenance. Could citations form a semiring?

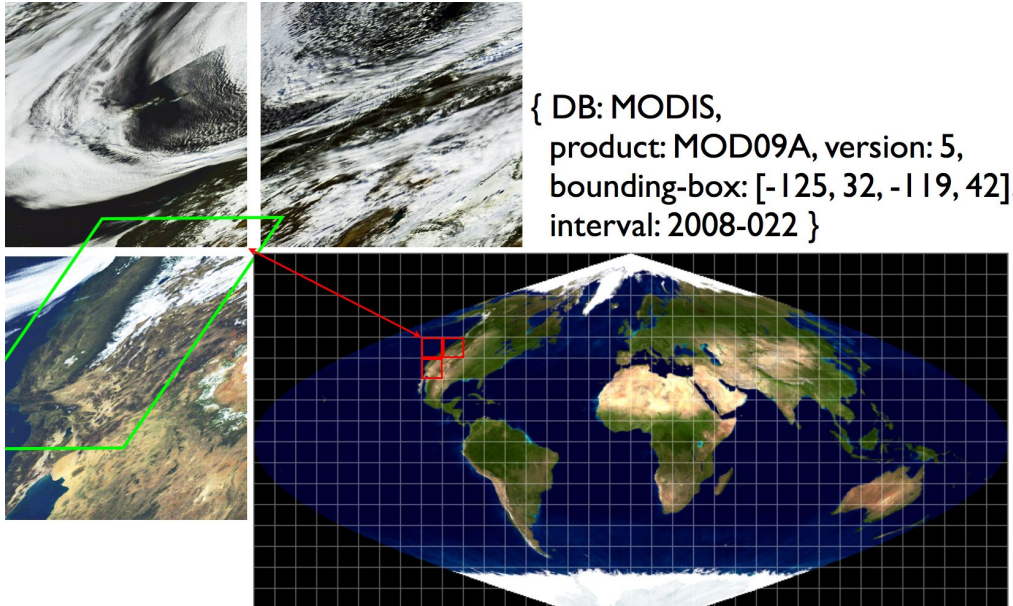
Yes they can …

(MODIS is a huge database of terrestrial satelite images)

{ DB : ”MODIS”, product : $$p, version: $v, bounding-box : [$$minlong, $$minlat, $$maxlong, $$maxlat], interval: [$$mint, $$maxt]}

←

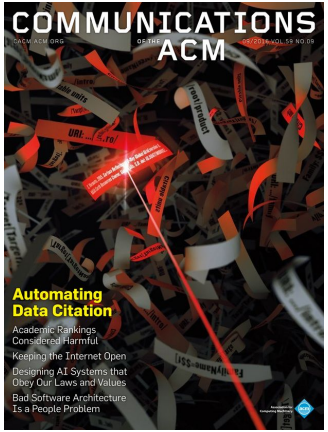
/root/product[ProdName=$p]/file[Lat ≥ $$minlat and Lat < $$maxlat and Lon ≥ $$minlon and Lon < $$maxlon and Time ≥ $$mint and Time < $$maxt]



Developing these ideas

[Davidson *et al* CIDR 2017] propose alternative semirings for citation that involve dictionaries and sets.

[Alawini et al JCDL2017] Use this to generate citations for the eagle-i database.

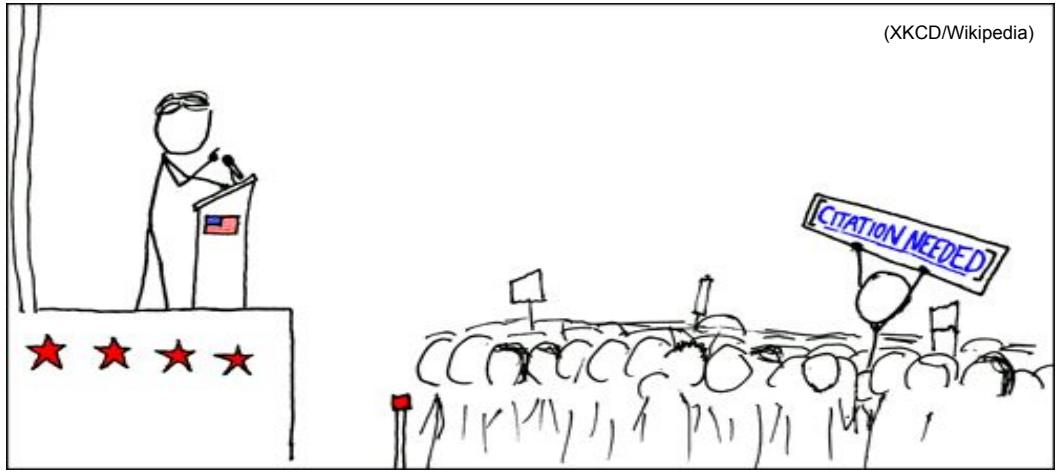


Bibliometrists and others are considering radically new forms of citation and publication

- the 10,000 author paper and the 10,000 citation paper
- transitive citations (some kind of PageRank)
- citation ontologies (why do we cite something)

Can we do the same or more for databases?

More generally, could we use ideas of provenance/citation into other social machines (Facebook, Twitter,...)?



“The technical community has the opportunity to produce tools that can be used by Internauts everywhere to separate quality information from dross, but the application of those tools falls to individual users willing to exercise critical thinking to get at the facts. Will liberty survive the Digital Age? Yes, I think it can, but only if we make it so.”
Vinton Cerf Can Liberty Survive the Digital Age? CACM May 2017

Thank you. Questions:

BL Cotton Nero A. X

Cotton Otho A. XII

Ann. Phys., Lpz 18 639-641

Nature, 171,737-738

Peter Buneman

```
wget -qO - http://mirror.hmc.edu/ctan/FILES.byname | grep ".bst$" \
| sed 's/.*\/\(.*\)/\1/' | sort -u | wc -l
```

Executed on 18 November 2011

Aad, G. *et al*. (ATLAS Collaboration, CMS Collaboration) *Phys. Rev. Lett.* **114**, 191803 (2015).