Technologies, methods and challenges to effective public data sharing and aggregation

Mark D Wilkinson
Medical Genetics, UBC
PI Bioinformatics
Heart + Lung Institute at St. Paul's Hospital

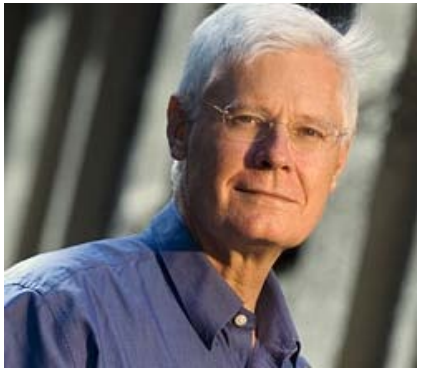
markw@illuminae.com
http://wilkinsonlab.ca

Thanks in advance

(very few of these ideas are my own!)

Paul Gordon – Sun Center of Excellence, U Calgary

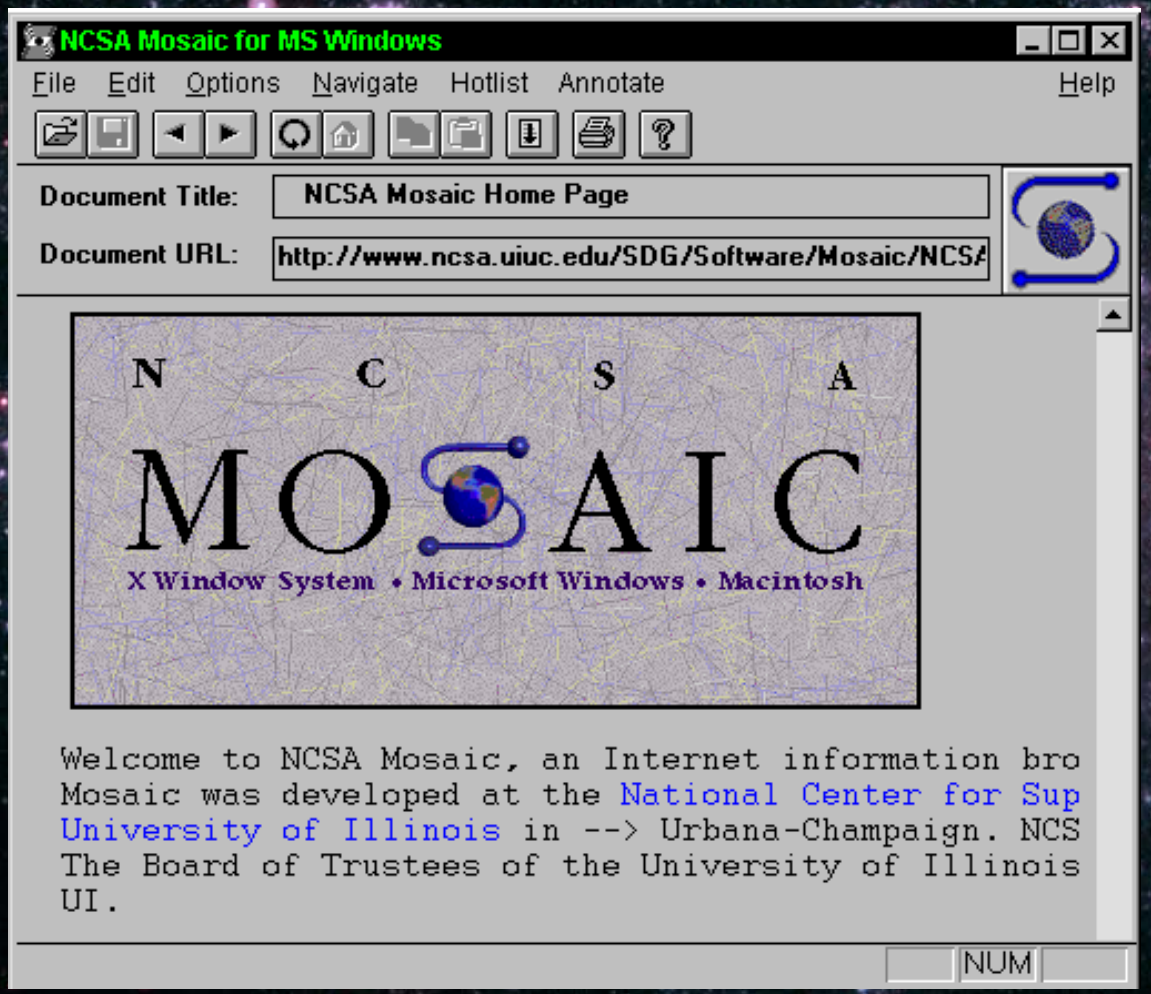
Carole Goble – University of Manchester

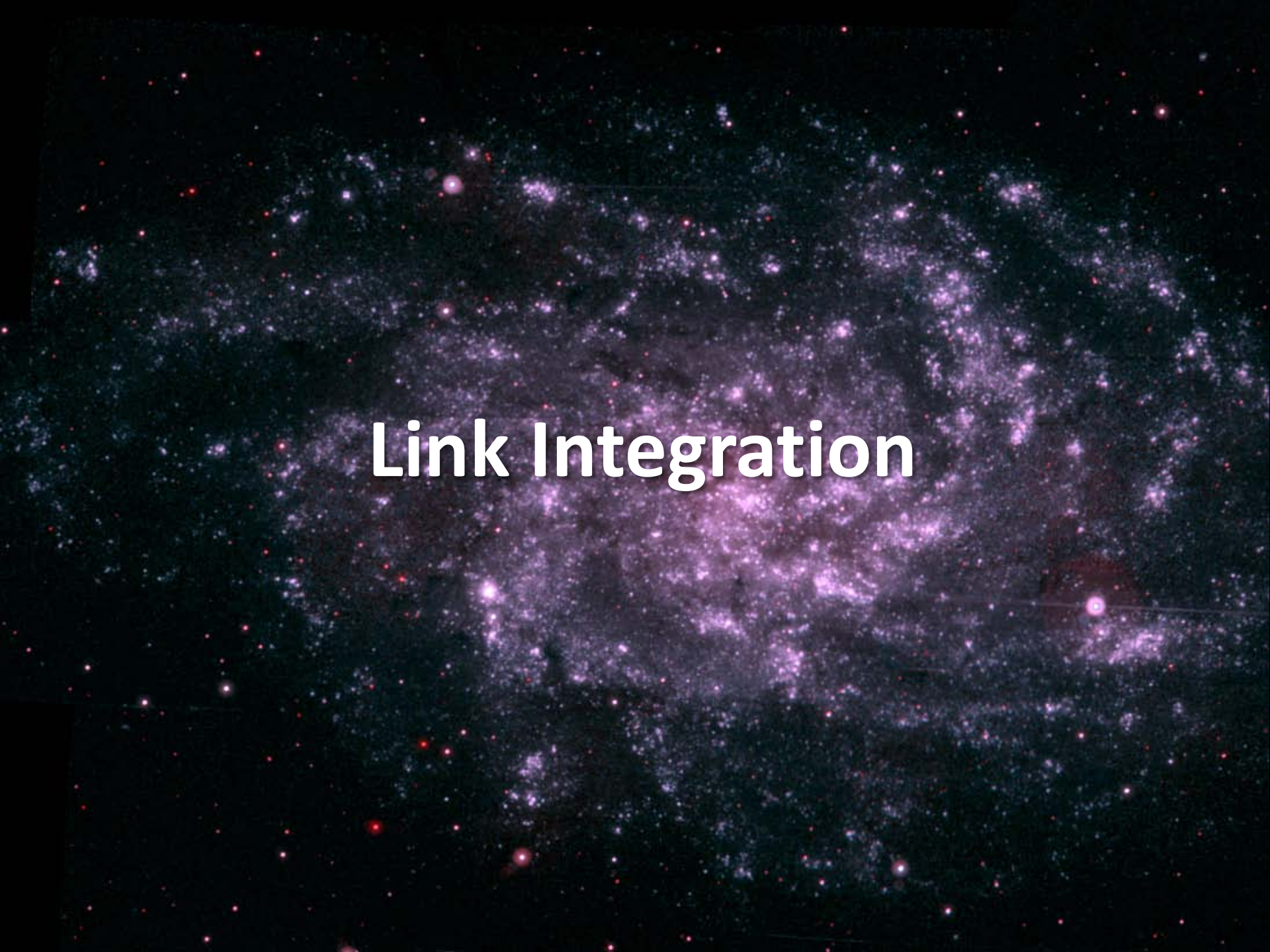


Charles Petrie – Stanford University

We’ve come a long way!!

In the beginning...




Link Integration

Show All entries | Show/hide columns | Filter:

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
BRCA1-001	ENST00000357654	7094	ENSP00000350283	1863	Protein coding	CCDS11453
BRCA1-003	ENST00000497488	779	ENSP00000418986	177	Protein coding	-
BRCA1-004	ENST00000477152	1980	ENSP00000419988	622	Protein coding	-
BRCA1-005	ENST00000471181	5939	ENSP00000418960	1885	Protein coding	-
BRCA1-006	ENST00000493795	5732	ENSP00000418775	1816	Protein coding	-
BRCA1-007	ENST00000468300	3273	ENSP00000417148	699	Protein coding	-
BRCA1-008	ENST00000493919	1948	ENSP00000418819	572	Protein coding	-
BRCA1-009	ENST00000478531	1972	ENSP00000420412	623	Protein coding	-
BRCA1-011	ENST00000470026	2108	ENSP00000419274	649	Protein coding	-
BRCA1-013	ENST00000494123	1612	ENSP00000419103	473	Protein coding	-
BRCA1-014	ENST00000491747	2376	ENSP00000420705	758	Protein coding	-
BRCA1-015	ENST00000484087	1495	ENSP00000419481	498	Protein coding	-
BRCA1-016	ENST00000489037	455	ENSP00000420781	98	Protein coding	-
BRCA1-017	ENST00000476777	769	ENSP00000417554	222	Protein coding	-
BRCA1-018	ENST00000473961	958	ENSP00000420201	319	Protein coding	-
BRCA1-019	ENST00000487825	800	ENSP00000418212	266	Protein coding	-
BRCA1-022	ENST00000461574	726	ENSP00000417241	242	Protein coding	-
BRCA1-201	ENST00000309486	7173	ENSP00000310938	1567	Protein coding	CCDS11459
BRCA1-202	ENST00000346315	6456	ENSP00000246907	1624	Protein coding	CCDS11455
BRCA1-203	ENST00000351666	3624	ENSP00000338007	680	Protein coding	CCDS11454
BRCA1-204	ENST00000352993	3747	ENSP00000312236	721	Protein coding	-
BRCA1-205	ENST00000353540	7050	ENSP00000013772	1822	Protein coding	-
BRCA1-206	ENST00000354071	6378	ENSP00000326002	1598	Protein coding	CCDS11456
BRCA1-207	ENST00000393680	7236	ENSP00000377285	1354	Protein coding	-
BRCA1-208	ENST00000393683	6484	ENSP00000377288	1567	Protein coding	CCDS11459
BRCA1-209	ENST00000393691	7370	ENSP00000377294	1863	Protein coding	CCDS11453
BRCA1-210	ENST00000412061	7167	ENSP00000397145	1863	Protein coding	CCDS11453
BRCA1-002	ENST00000492859	1584	ENSP00000420253	59	Nonsense mediated decay	-
BRCA1-010	ENST00000461221	5693	ENSP00000418548	63	Nonsense mediated decay	-
BRCA1-020	ENST00000461798	582	ENSP00000417988	63	Nonsense mediated decay	-
BRCA1-012	ENST00000467274	4497	No protein product	-	Retained intron	-
BRCA1-021	ENST00000472490	561	No protein product	-	Retained intron	-

Integration of What?

Specially-formatted Files

```
FT   CDS             73..1212
FT                   /db_xref="GDB:135679"
FT                   /db_xref="GOA:P27361"
FT                   /db_xref="HGNC:6877"
FT                   /db_xref="UniProtKB/Swiss-Prot:P27361"
FT                   /gene="ERK1"
FT                   /product="protein serine/threonine kinase"
FT                   /protein_id="CAA42744.1"
FT                   /translation="MAAAAAQGGGGGEPRRTEGVGPGVPGEVEMVKGQPFDVGPRYTQL
FT                   QYIGEGAYGMVSSAYDHVRKTRVAIKKISPFEHQTYCQRTLREIQILLRFRHENVIGIR
FT                   DILRASTLEAMRDVYIVQDLMETDLYKLLKSQQLSNDHICYFLYQILRGLKYIHSANVL
FT                   HRDLKPSNLLSNTTCDLKICDFGLARIADPEHDHTGFLTEYVATRWYRAPEIMLNSKGY
FT                   TKSIDIWSVGCILAEMLSNRPIFPGKHYLDQLNHILGILGSPSQEDLNCIINMKARNYL
FT                   QSLPSKTKVAWAKLFPKSDSKALDLLDRMLTFNPNKRITVEEALAHPYLEQYYDPTDEP
FT                   VAEEPFTFAMELDDLPKERLKELIFQETARFQPGVLEAP"
FT   misc_binding    1354..1379
FT                   /Affymetrix="probe:HG_U95Av2:1000_at;399:559;"
FT   misc_binding    1366..1391
FT                   /Affymetrix="probe:HG_U95Av2:1000_at;544:185;"
```

EMBL Format

Specially-formatted Files

```
misc_feature    447283..484667
                /note="Putative prophage; Putative lysogenic prophage
                region. Best Hits not completed. Putative attL/R; PHAGE01"
gene            complement(447387..448511)
                /locus_tag="BA0427"
                /db_xref="GeneID:1087434"
                /db_xref="prophinder:34501"
                /db_xref="sherw:116"
                /db_xref="phage_finder:16305"
CDS             complement(447387..448511)
                /locus_tag="BA0427"
                /note="identified by match to PFAM protein family HMM
                PF00589"
                /codon_start=1
                /transl_table=11
                /product="prophage LambdaBa04, site-specific recombinase,
                phage integrase family"
                /protein_id="NP_842969.1"
                /db_xref="GI:30260592"
                /db_xref="GeneID:1087434"
                /db_xref="aclame:protein:vir:6181"
                /translation="MKGYFRKRGEKWSFTIDIGKDPITGKRKQKTASGFKTKKEAERA
                CNELIHQFNTGSLVDDKNFTLSEYLQEWLENTAKQRVRETTFTNYKRAINSRIIPVLG
                SHKLKDLKPLHGQRFVKSLIDEGLSPAYIEYIFIVLKGSLEDAVRWELLFKNPFQHVE
                IPRPRKVVNSTWSIEETKKFLNRTKFENVIYYHLFLLALNTGMRRGEILGLKWKNFDL
                NEGKISVTETLIYDENGFRFTEPKTHGSKRLISIDQNLCKEFKSYKAKQNEFKLLFGQ
                SYEDNDLVFAKETGQPILPRTMTTTFNQFIKKADVPQIRFHDLRHTHATILLKLGINP
                KIVSERLGHSSIKTTLDTYSHVTIDMQESAVLKLSEALKS"
                /db_hit="ID=aclame:protein:vir:6181#Eval=6e-77#bits=281"
```

GenBankFormat

Specially-formatted Files

```
>AB000263 |acc=AB000263|descr=Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

FASTA Format

Specially-formatted Files

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
AB000263  Length: 368  Check: 4514  ..
       1  acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg
      61  ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg
     121  caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc
     181  aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag
     241  gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
     301  agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca
     361  gacctgaa
```

GCG Format

At least 20 different formats for representing DNA sequences...

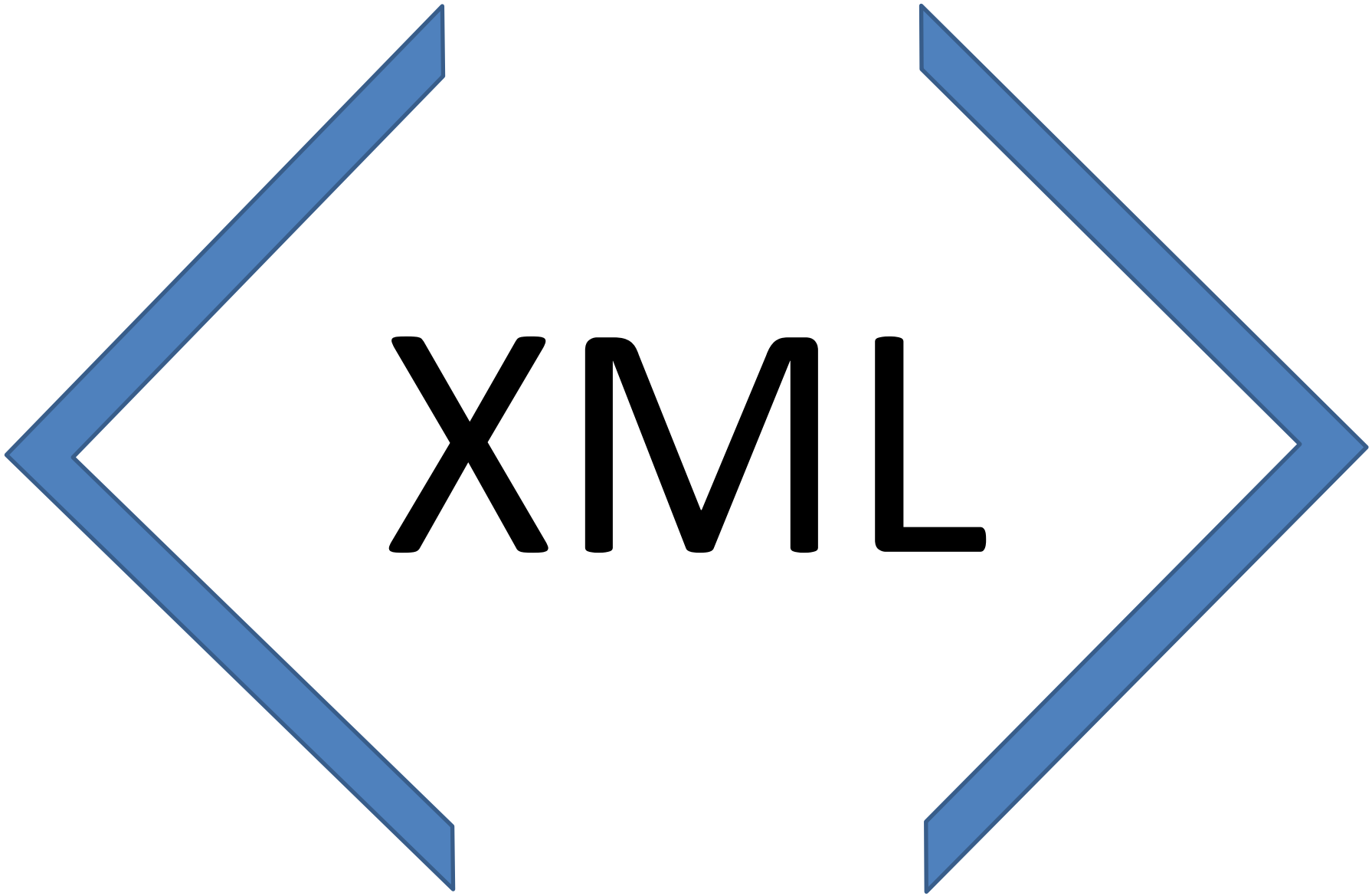
Lord et al. 2004

Many formats contained a wide variety of related, but different, information

DNA sequence
Sequence features
Translation
Date/time/method
Publication cross-references
...
...

Each file format required it’s own parser...

...and that problem wasn’t limited to DNA...


XML

What did XML do for us?

"...advent of XML meant that we didn't have to write our own parsers anymore..."

Individual data elements in a file
can be automatically located and extracted

Predictable way to represent data

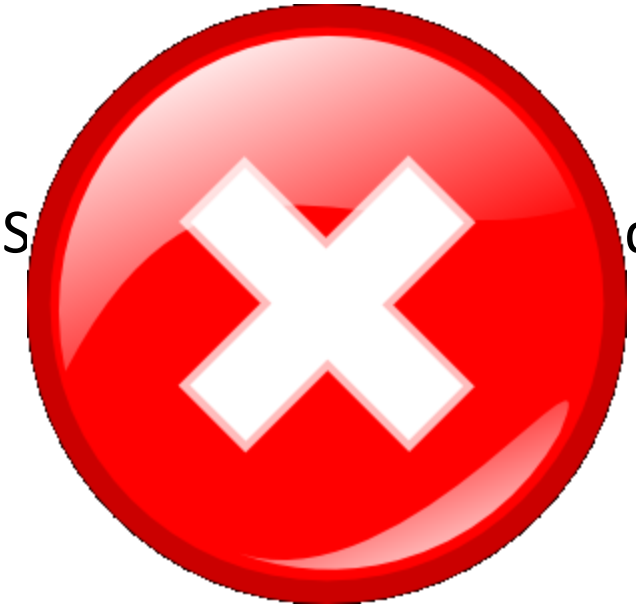
Makes it easier for machines to encode/extract

```
      <value>
        BRCA1
      </value>
    </qualifier>
    <qualifier name="product">
      <value>
        breast and ovarian cancer susceptibility
      </value>
    </qualifier>
    <qualifier name="note">
      <value>
        influences susceptibility to breast and ovarian cancer
      </value>
    </qualifier>
    <qualifier name="protein_id">
      <value>
        AAA73985.1
      </value>
    </qualifier>
    <qualifier name="translation">
      <value>
        MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKGPSQCPLCKNDITKRSLQESTRFSQLVEELLKIICAFQLDTGLEYANS
      </value>
    </qualifier>
    <location type="single" complement="false">
      <locationElement type="range" accession="U14680" version="1" complement="false">
        <basePosition type="simple">
          120
        </basePosition>
        <basePosition type="simple">
          5711
        </basePosition>
      </locationElement>
    </location>
  </feature>
  <feature name="exon">
    <qualifier name="gene">
      <value>
        BRCA1
      </value>
    </qualifier>
    <qualifier name="number">
      <value>
        3
      </value>
    </qualifier>
```

EMBL Record for BRCA1
In XML

```
      </GBQualifier_name>
      <GBQualifier_value>
        breast and ovarian cancer susceptibility
      </GBQualifier_value>
    </GBQualifier>
    <GBQualifier>
      <GBQualifier_name>
        protein_id
      </GBQualifier_name>
      <GBQualifier_value>
        AAA73985.1
      </GBQualifier_value>
    </GBQualifier>
    <GBQualifier>
      <GBQualifier_name>
        db_xref
      </GBQualifier_name>
      <GBQualifier_value>
        GI:555932
      </GBQualifier_value>
    </GBQualifier>
    <GBQualifier>
      <GBQualifier_name>
        translation
      </GBQualifier_name>
      <GBQualifier_value>
        MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKGPSQCPLCKNDITKRSLQESTRFSQLVEELLKIICAFQLDTGLEYANSYNFAKKEN
      </GBQualifier_value>
    </GBQualifier>
  </GBFeature_quals>
</GBFeature>
<GBFeature>
  <GBFeature_key>
    exon
  </GBFeature_key>
  <GBFeature_location>
    200..253
  </GBFeature_location>
  <GBFeature_intervals>
    <GBInterval>
      <GBInterval_from>
        200
      </GBInterval_from>
      <GBInterval_to>
        253
      </GBInterval_to>
      <GBInterval_accession>
```

GenBank Record for BRCA1
In XML

S  d

```
<qualifier name="translation">
  <value>
    MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIF
  </value>
</qualifier>
```

EMBL Record for BRCA1
In XML

```
<GBQualifier>
  <GBQualifier_name>
    translation
  </GBQualifier_name>
  <GBQualifier_value>
    MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIF
  </GBQualifier_value>
</GBQualifier>
```

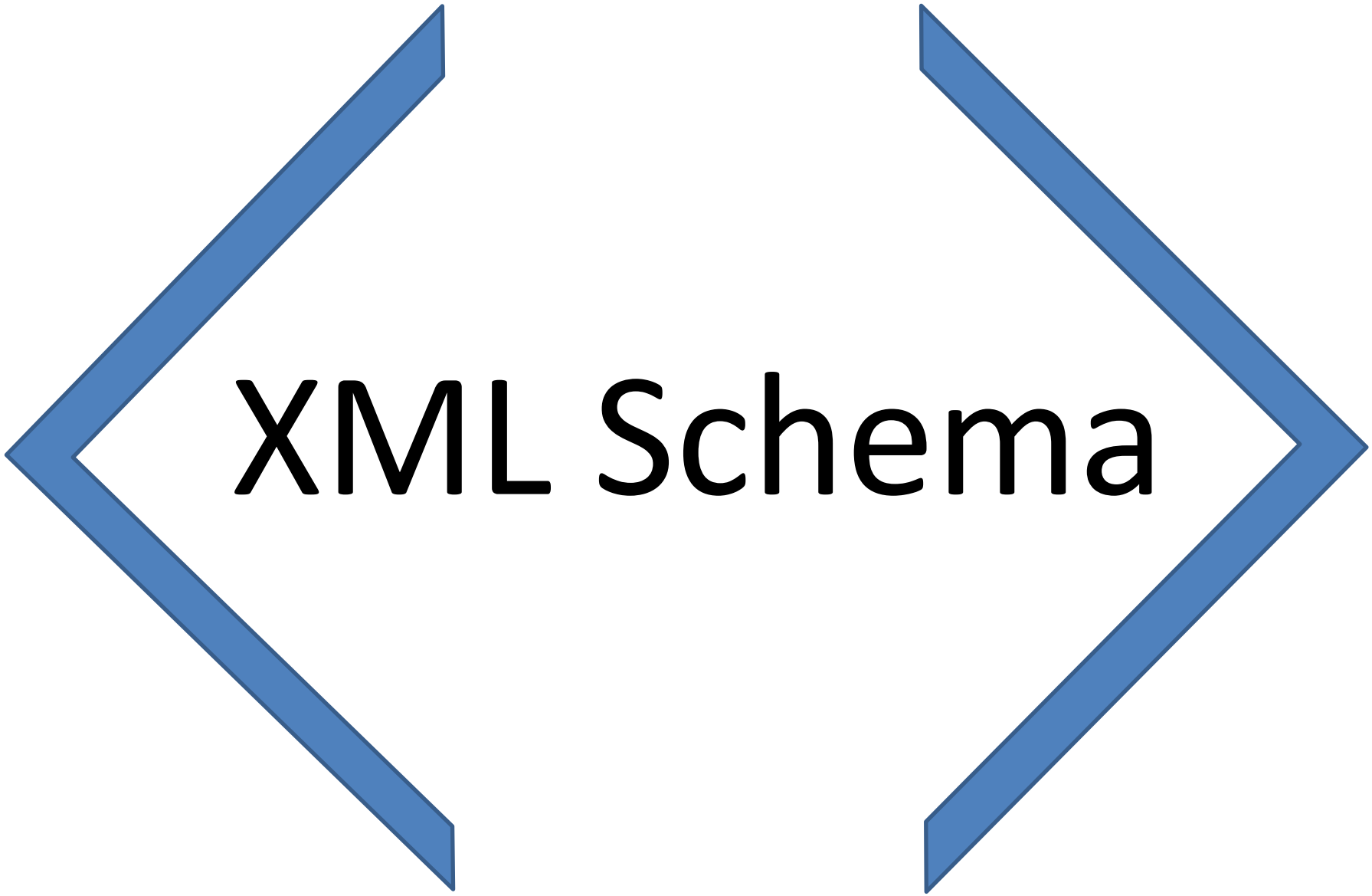
GenBank Record for BRCA1
In XML

S  d

...because it isn’t (just) a parsing problem...

Various resources have various data models

So... Let’s find a way to describe the data models!


XML Schema

```
<qualifier name="translation">
  <value>
    MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIF
  </value>
</qualifier>
```

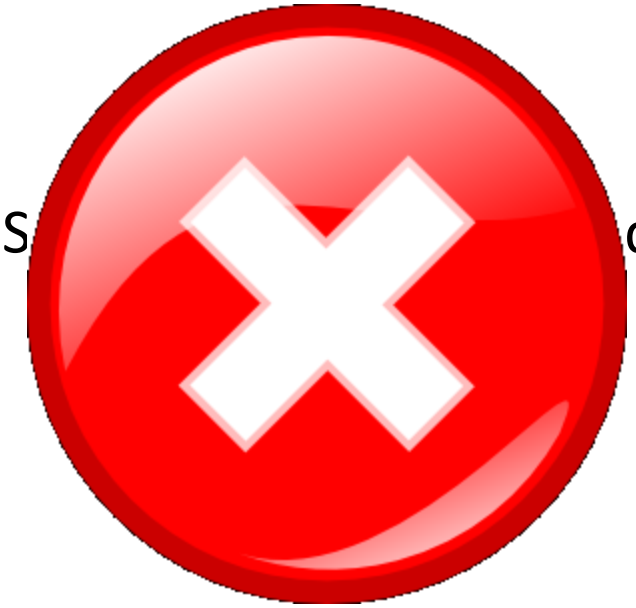
XML Schema

*There will be an element called "**qualifier**"*
*It will have an attribute called "**name**"*
*The content of that attribute will be **text***
*There will be a child element called "**value**"*
*The content of that child element will be **free-text***

```
<GBQualifier>
  <GBQualifier_name>
    translation
  </GBQualifier_name>
  <GBQualifier_value>
    MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIF
  </GBQualifier_value>
</GBQualifier>
```

XML Schema

*There will be an element called "**GBQualifier**"*
*There will be a child element called "**GBQualifier_name**"*
*The content of that child element will be **free-text***
*There will be a child element called "**GBQualifier_value**"*
*The content of that child element will be **free-text***

S  d

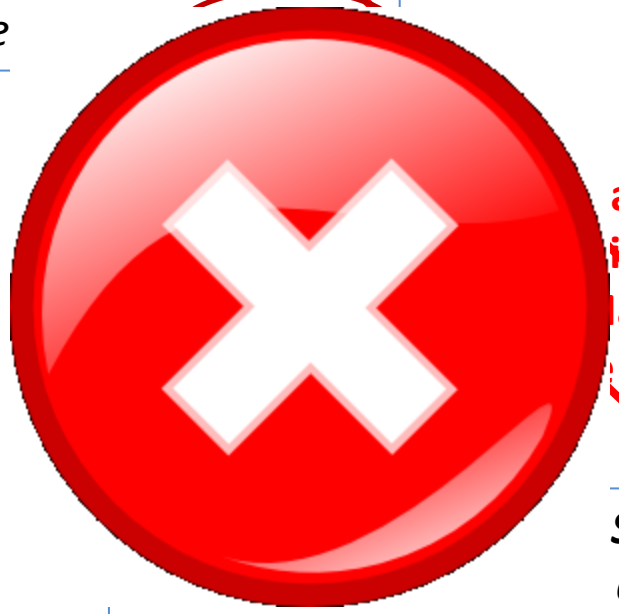
What did XML Schema do for us?

"...XML Schema (among other things) allowed us to ~automate the creation of (in-memory) Structures which could hold the given XML-formatted data..."

Does not solve the integration or aggregation problem

XML Schema

*There will be an element called "**qualifier**"*
*It will have an attribute called "**name**"*
*The content of that attribute will be **text***
*There will be a child element called "**value**"*
The content of that child eleme



aning" of each
it, we resort to
lapping"
e the data

Schema

*called "**GBQualifier**"*
*There will be a child element called "**GBQualifier_name**"*
*The content of that child element will be **free-text***
*There will be a child element called "**GBQualifier_value**"*
*The content of that child element will be **free-text***

Nevertheless...

Web Services

“Service Oriented Architectures”

WSDL

(and many other 4-letter words)

Web Services & SOA's

Allow you to expose software

(e.g. a database, analytical tool, or service)

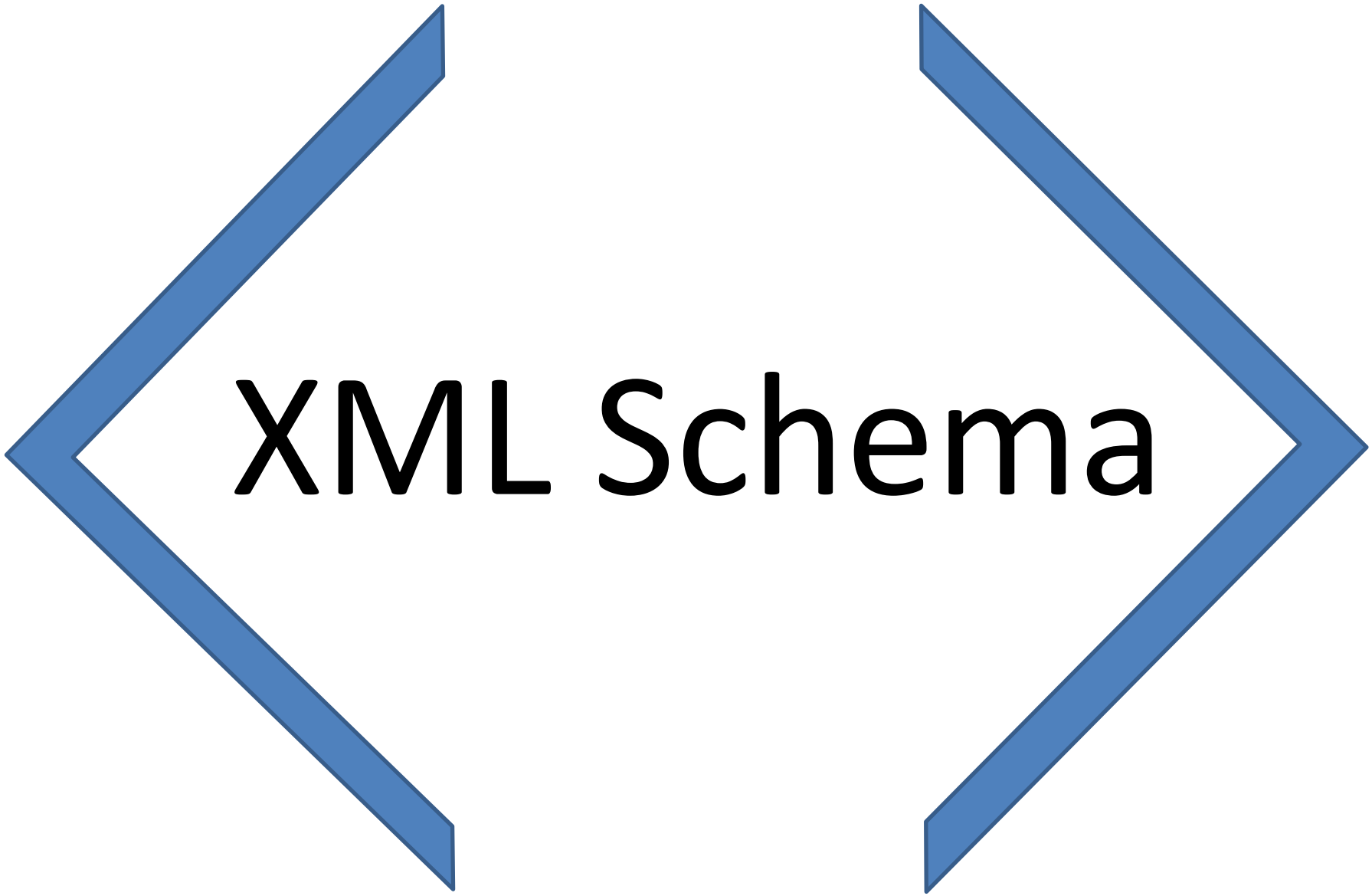
on the Web

so that others can use it

(in their own analytical pipelines)

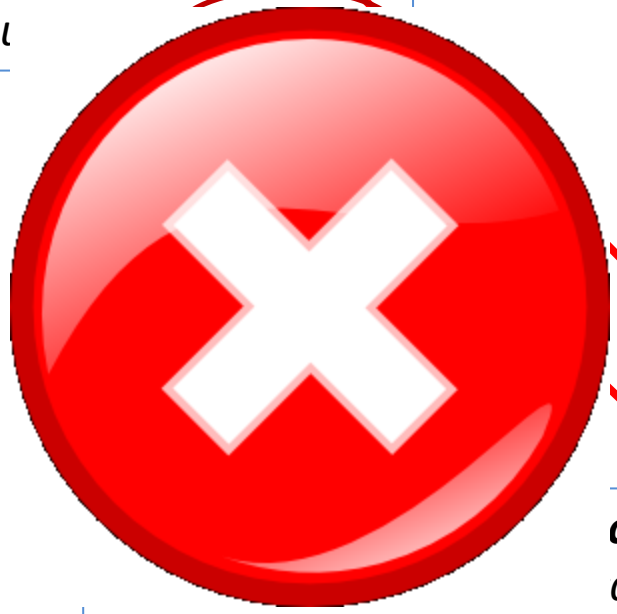
Excellent!!

But...


XML Schema

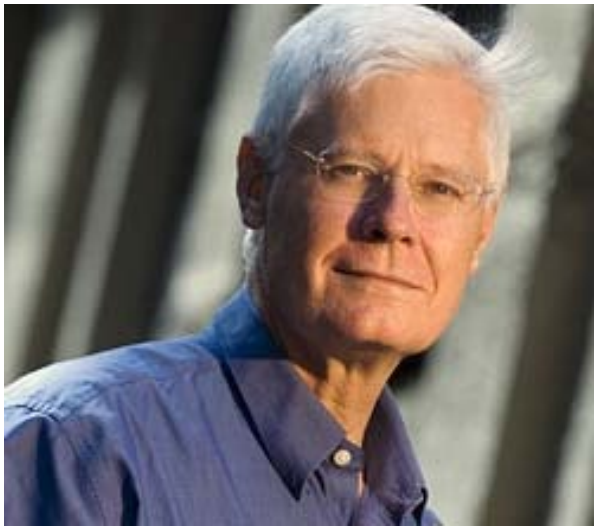
XML Schema

*There will be an element called "**qualifier**"*
*It will have an attribute called "**name**"*
*The content of that attribute will be **text***
*There will be a child attribute called "**value**"*
The content of that child attribu



chema

*alled "**GBQualifier**"*
*There will be a child attribute called "**GBQualifier_name**"*
*The content of that child attribute will be **free-text***
*There will be a child attribute called "**GBQualifier_value**"*
*The content of that child attribute will be **free-text***



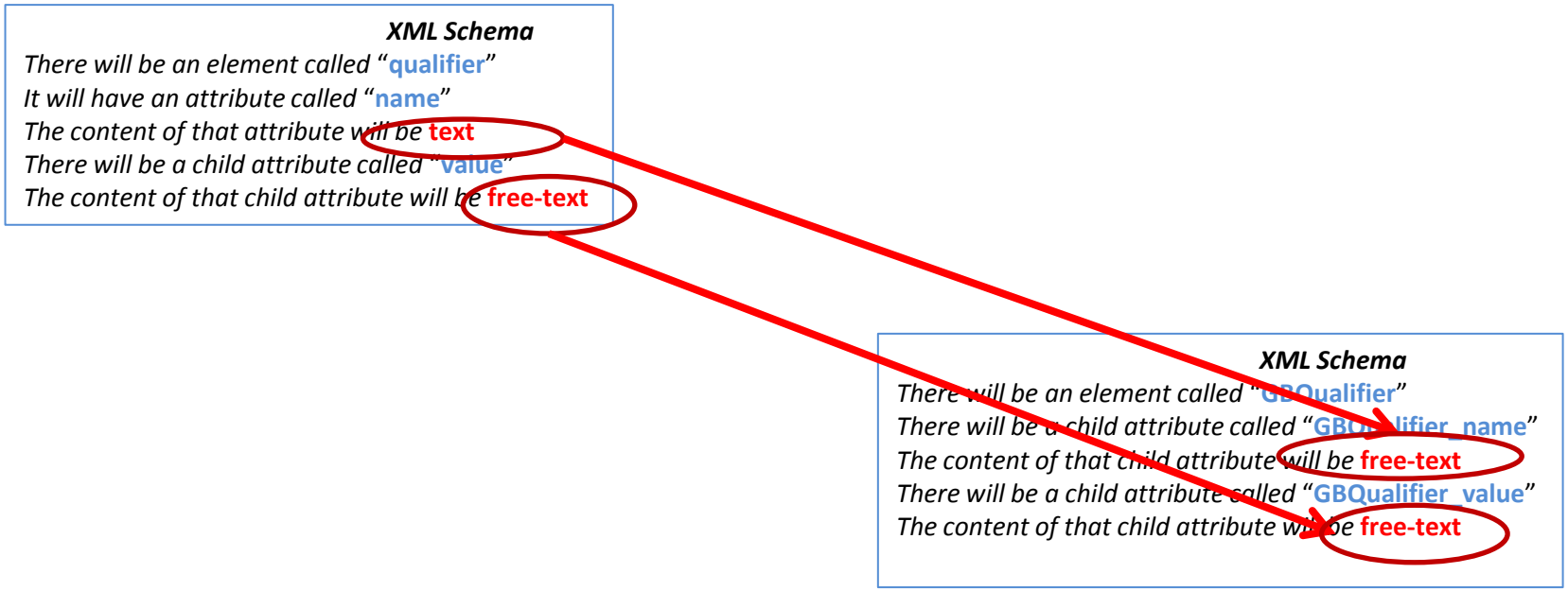
"The phrase 'practical Web Services' is not intrinsically an oxymoron, but [I] argue that there are few in existence."

Why?

Because this problem is ***so disruptive*** that ***there is little point*** in building "public" Web Services...

They are simply too difficult to integrate with other "public" Web Services.

-- adapted from Petrie, SWSIP 2009



...and that’s pretty much
where the world is right now...


But there is hope!

“Linked Data” movement

Resource Description Framework
“RDF”

Two new technologies & communities

The “Semantic Web” movement

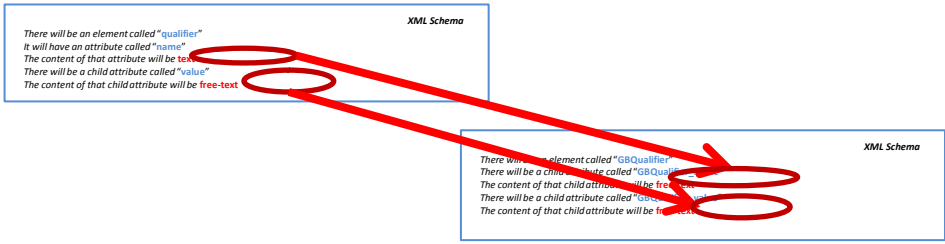
Web Ontology Language
“OWL” (+ RDF)

What does RDF do for us?

"...RDF replaces XML Schema, because RDF says that ***there is only one data model***..."

What does OWL do for us?

"...the semantics are ***no longer implicit*** in the data model..."



So what?

The Semantic Web

Gives us the opportunity to re-think
how we build our health data infrastructures

The Semantic Web

isn’t

“yet another layer of technology”

The Semantic Web

changes

the way we write software

The Semantic Web

What to do & How to do it

is no longer encoded in your software

The Semantic Web

What to do & How to do it

is part of the data

The Semantic Web

What to do & How to do it
is part of a shared,
expert understanding

The Semantic Web

What to do & How to do it

IS* PERSONAL!

*** can be...**

One piece of software

Any question... Any answer

Let me demonstrate what I mean

SADI

Find. Integrate. Analyze.

Semantic **A**utomated **D**iscovery and **I**ntegration

http://sadiframework.org

Microsoft Research

HEART & STROKE FOUNDATION OF BC & YUKON

Finding answers. For life.

Founding partner

CARDIO
SHARE

Data + Knowledge
for Cardiac Researchers

Semantic **H**ealth **A**nd **R**esearch **E**nvironment

(a Semantic Web question answerer...)

Example #1

Show me the latest Blood Urea Nitrogen and Creatinine levels of patients **who appear to be rejecting their transplants**

```
SELECT ?patient ?bun ?creat
FROM <http://sadiframework.org/ontologies/patients.rdf>
WHERE {
        ?patient rdf:type patient:LikelyRejecter .
        ?patient l:latestBUN ?bun .
        ?patient l:latestCreatinine ?creat .
}
```

Likely Rejecter:

A patient who has creatinine levels that are increasing over time

- - Wilkinson “MD”

Likely Rejecter:

…but there is no “likely rejecter” column or table in our database…

Likely Rejecter:

Our database contains various blood chemistry measurements at various time-points

SHARE determines

by itself

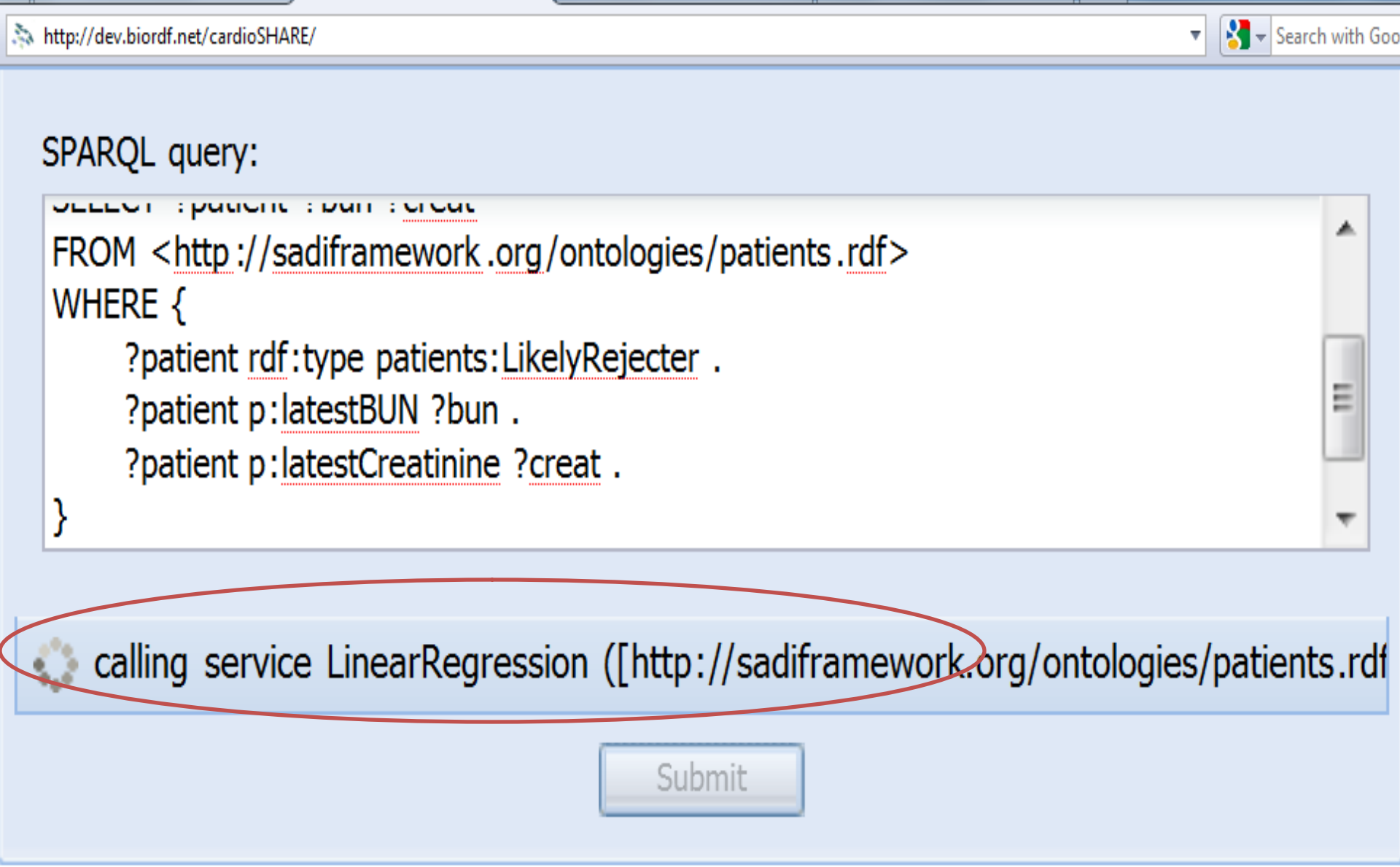
the **need** to do a
Linear Regression analysis over
Creatinine blood chemistry measurements

SHARE determines

by itself

how and where that analysis can be done

and **does it**



The SHARE system utilizes Semantics (via SADI) to discover and access analytical services on the Web that do linear regression analysis

SPARQL query:

```
WHERE {
      ?patient rdf:type patients:LikelyRejecter .
      ?patient p:latestBUN ?bun .
      ?patient p:latestCreatinine ?creat .
}
```

View results as RDF. There were warnings executing the query. Click for details.

Submit

VOILA!

Query results

bun	creat	patient
5.861790	1.215768	http://biordf.net/moby/Dumm...
17.673603	1.000161	http://biordf.net/moby/Dumm...
7.997613	1.146408	http://biordf.net/moby/Dumm...
2.977437	0.953866	http://biordf.net/moby/Dumm...
10.995189	1.247073	http://biordf.net/moby/Dumm...
1.168096	1.185007	http://biordf.net/moby/Dumm...
7.570712	0.986164	http://biordf.net/moby/Dumm...

Neither SADI nor SHARE

know anything about

blood chemistry, or mathematics

Example #2

From a (contrived) integrated dataset,
retrieve the blood pressure measurements

```
SELECT ?output ?unit ?value
FROM <http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl>
WHERE {

    ?output rdf:type sbp:BloodPressure .
    ?output local:hasCanonicalAttribute ?pr .
    ?pr sio:SIO_000221 ?unit .
    ?pr sio:SIO_000300 ?value .

}
```

This should be extremely straightforward...

...except for one problem...

```
<owl:NamedIndividual rdf:about="http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl#pressureinstance1">
    <rdf:type rdf:resource="&galen;SystolicBloodPressure"/>
    <resource:SIO_000300>0.137</resource:SIO_000300>
    <resource:SIO_000221 rdf:resource="&ucum;unit/pressure/meter-of-mercury-column"/>
  </owl:NamedIndividual>


  <owl:NamedIndividual rdf:about="http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl#pressureinstance2">
    <rdf:type rdf:resource="&galen;SystolicBloodPressure"/>
    <resource:SIO_000300>12.45</resource:SIO_000300>
    <resource:SIO_000221 rdf:resource="http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl#centi-meter-of-mercury-column"/>
  </owl:NamedIndividual>


  <owl:NamedIndividual rdf:about="http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl#pressureinstance3">
    <rdf:type rdf:resource="&galen;SystolicBloodPressure"/>
    <resource:SIO_000300>5.3</resource:SIO_000300>
    <resource:SIO_000221 rdf:resource="&ucum;unit/pressure/inch-of-mercury-column"/>
  </owl:NamedIndividual>
```

<owl:NamedIndividual rdf:about="http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl#pressureinstance1">
<rdf:type rdf:resource="&galen;SystolicBloodPressure"/>
<resource:SIO_000300>**0.137**</resource:SIO_000300>
<resource:SIO_000221 rdf:resource="&ucum;unit/pressure/**meter-of-mercury**-column"/>
</owl:NamedIndividual>

<owl:NamedIndividual rdf:about="http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl#pressureinstance2">
<rdf:type rdf:resource="&galen;SystolicBloodPressure"/>
<resource:SIO_000300>**12.45**</resource:SIO_000300>
<resource:SIO_000221 rdf:resource="&ucum;unit/framingham/sbpfeb.owl#**centi-meter-of-mercury**-colum
</owl:NamedIndividual>

<owl:NamedIndividual rdf:about="http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl#pressureinstance3">
<rdf:type rdf:resource="&galen;SystolicBloodPressure"/>
<resource:SIO_000300>**5.3**</resource:SIO_000300>
<resource:SIO_000221 rdf:resource="&ucum;unit/pressure/**inch-of-mercury**-column"/>
</owl:NamedIndividual>

Example #2

From a (contrived) integrated dataset,
retrieve the blood pressure measurements

```
SELECT ?output ?unit ?value
FROM <http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl>
WHERE {

    ?output rdf:type sbp:BloodPressure .
    ?output local:hasCanonicalAttribute ?pr .
    ?pr sio:SIO_000221 ?unit .
    ?pr sio:SIO_000300 ?value .

}
```

My semantic definition of "Blood Pressure" includes the units that I want...

This is enough to trigger SHARE to automatically discover an online unit-conversion service...

SPARQL query:

```
?output rdf:type sbp:BloodPressure .
?output sbp:hasCanonicalAttribute ?pr .
?pr sio:SIO_000221 ?unit .
?pr sio:SIO_000300 ?value .

}
```

View results as RDF. There were warnings executing the query. Click for details.

Submit

Query results

output	unit	value
http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl#instance3	http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl#'milli-meter-of-mercury-column	134.61999999999998
http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl#instance2	http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl#'milli-meter-of-mercury-column	124.5
http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl#instance1	http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl#'milli-meter-of-mercury-column	137.0

unit	value
http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl#'milli-meter-of-mercury-column	134.61
http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl#'milli-meter-of-mercury-column	124.5
http://es-01.chibi.ubc.ca/~soroush/framingham/sbpfeb.owl#'milli-meter-of-mercury-column	137.0

Neither SADI nor SHARE

know anything about

units or unit conversions

Many of the challenges
to data aggregation and sharing
now have solutions that work!

What, in my opinion, is the greatest remaining challenge?

To a biologist...

...“data mining” means “this data is mine!”

The challenge to us all

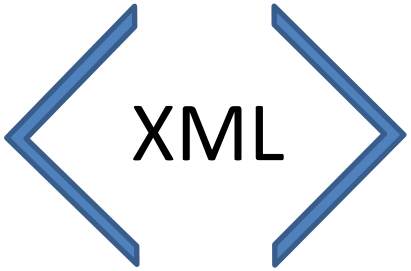
Move from Data Mine-ing

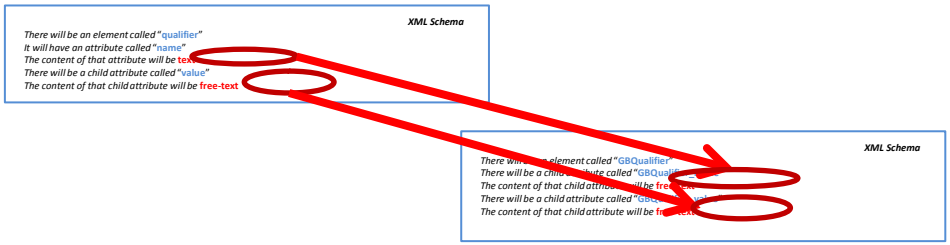
To Data Ours-ing

-- Len Silverston, 2007


```
FT   CDS             73..1212
FT                   /db_xref="GDB:135679"
FT                   /db_xref="GOA:P27361"
FT                   /db_xref="HGNC:6877"
FT                   /db_xref="UniProtKB/Swiss-Prot:P27361"
FT                   /gene="ERK1"
FT                   /product="protein serine/threonine kinase"
FT                   /protein_id="CAA42744.1"
FT                   /translation="MAAAAAQGGGGGEPRRTEGVGPGVPGEVEMVKGQPFDVGPRYTQL
FT                   QYIGEGAYGMVSSAYDHVRKTRVAIKKISPFEHQTYCQRTLREIQILLRFRHENVIGIR
FT                   DILRASTLEAMRDVYIVQDLMETDLYKLLKSQQLSNDHICYFLYQILRGLKYIHSANVL
FT                   HRDLKPSNLLSNTTCDLKICDFGLARIADPEHDHTGFLTEYVATRWYRAPEIMLNSKGY
FT                   TKSIDIWSVGCILAEMLSNRPIFPGKHYLDQLNHILGILGSPSQEDLNCIINMKARNYL
FT                   QSLPSKTKVAWAKLFPKSDSKALDLLDRMLTFNPNKRITVEEALAHPYLEQYYDPTDEP
FT                   VAEEPFTFAMELDDLPKERLKELIFQETARFQPGVLEAP"
FT   misc_binding    1354..1379
FT                   /Affymetrix="probe:HG_U95Av2:1000_at;399:559;"
FT   misc_binding    1366..1391
FT                   /Affymetrix="probe:HG_U95Av2:1000_at;544:185;"
```

We've come a long way!!





TEAM:

Luke McCarthy
Benjamin Vandervalk
Soroush Samadian

Microsoft *Research*