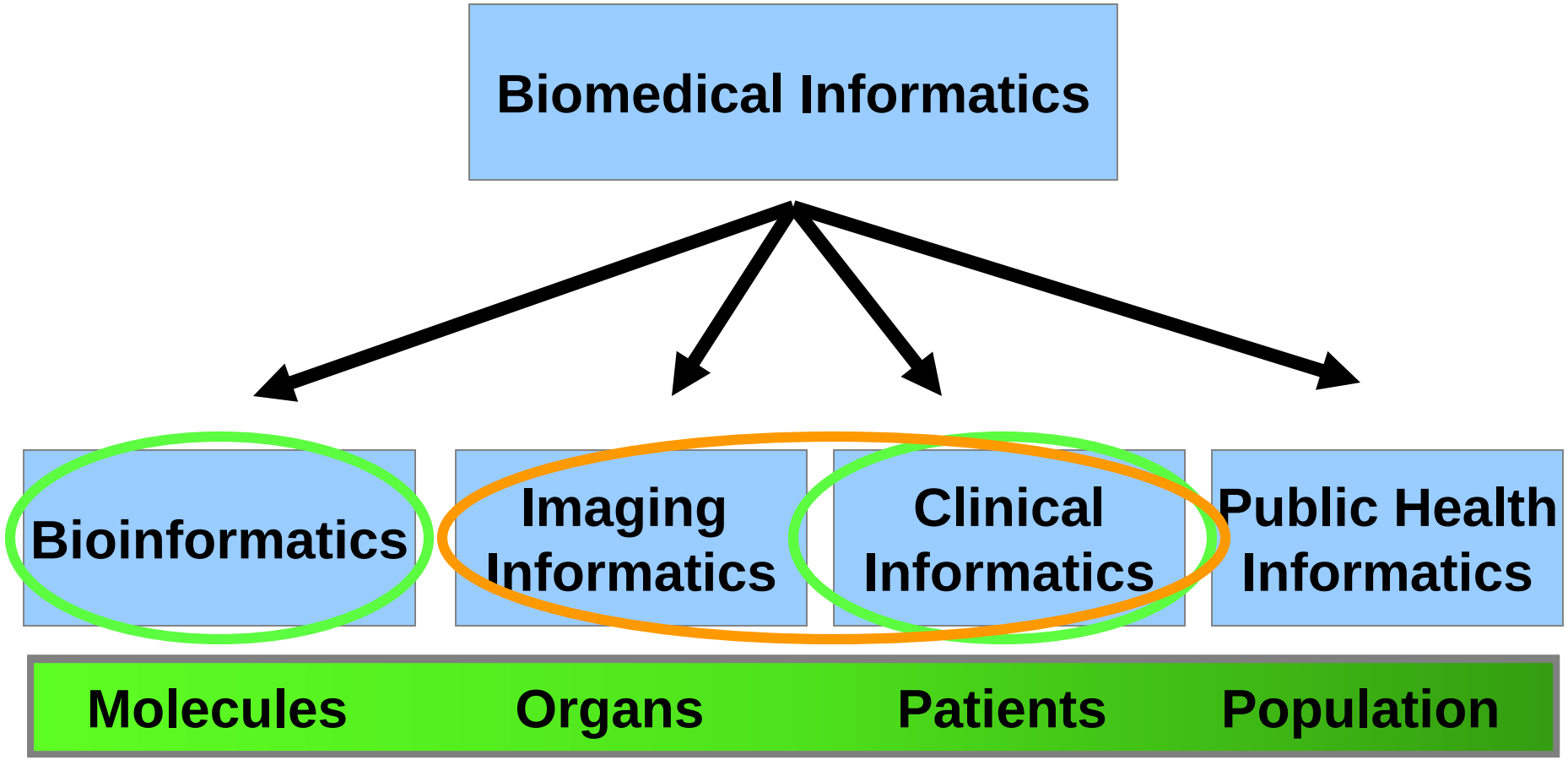
Informatics I: Data Standards

Jessie Tenenbaum, PhD, FACMI
jessie.tenenbaum@duke.edu
@jessiet1023

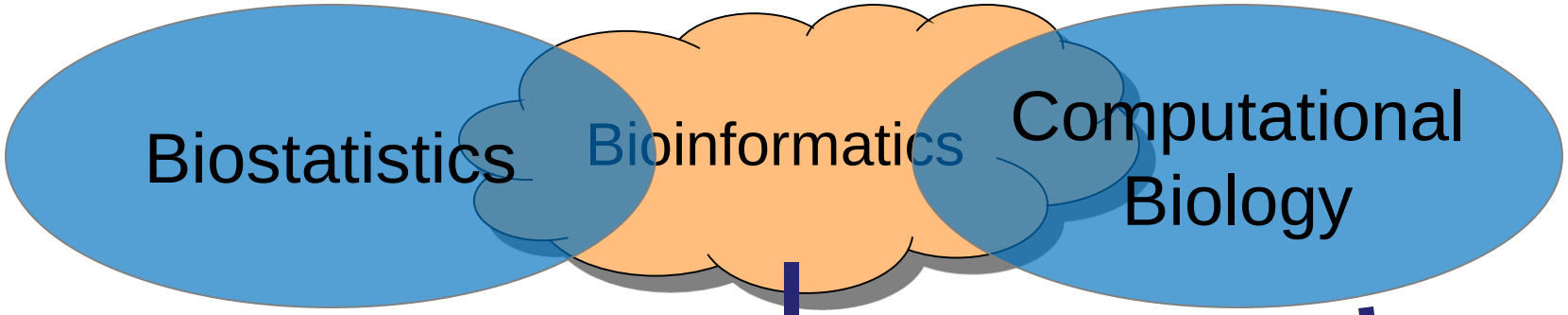
JDT Intro

- Division of Translational Biomedical Informatics in B&B Dept.
- Previous life as Program Manager at Microsoft
- PhD 2007 in Biomedical Informatics
- At Duke for 10 years, 2 as faculty
- Research
  - Data standards & research data warehousing
  - Infrastructure to support precision medicine, esp in mental health
    - EHR data mining
    - Clinical decision support

Biomedical Informatics
(as described by E. Shortliffe)



Bioinformatics in context



Interpretation and inference of numerical biological data

Databases, data management, knowledge representation

Algorithms, computationally intensive techniques- machine learning, signal detection, pattern recognition

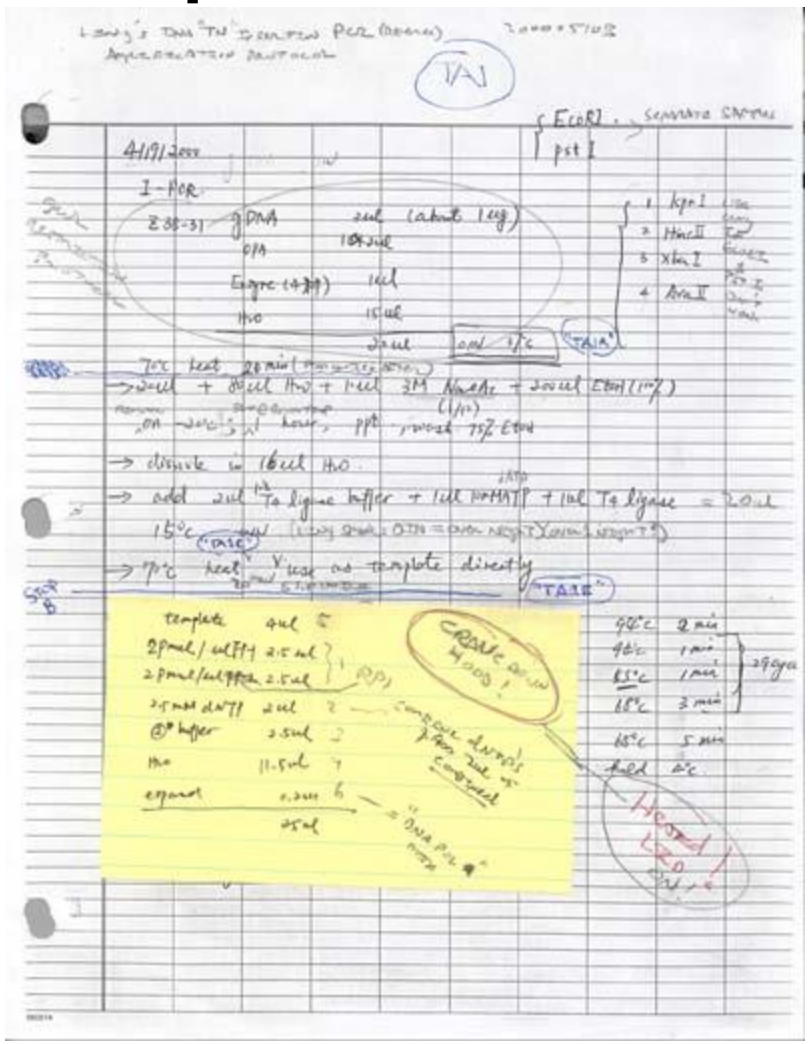
Large scale visualization, data retrieval

Data modeling and simulation

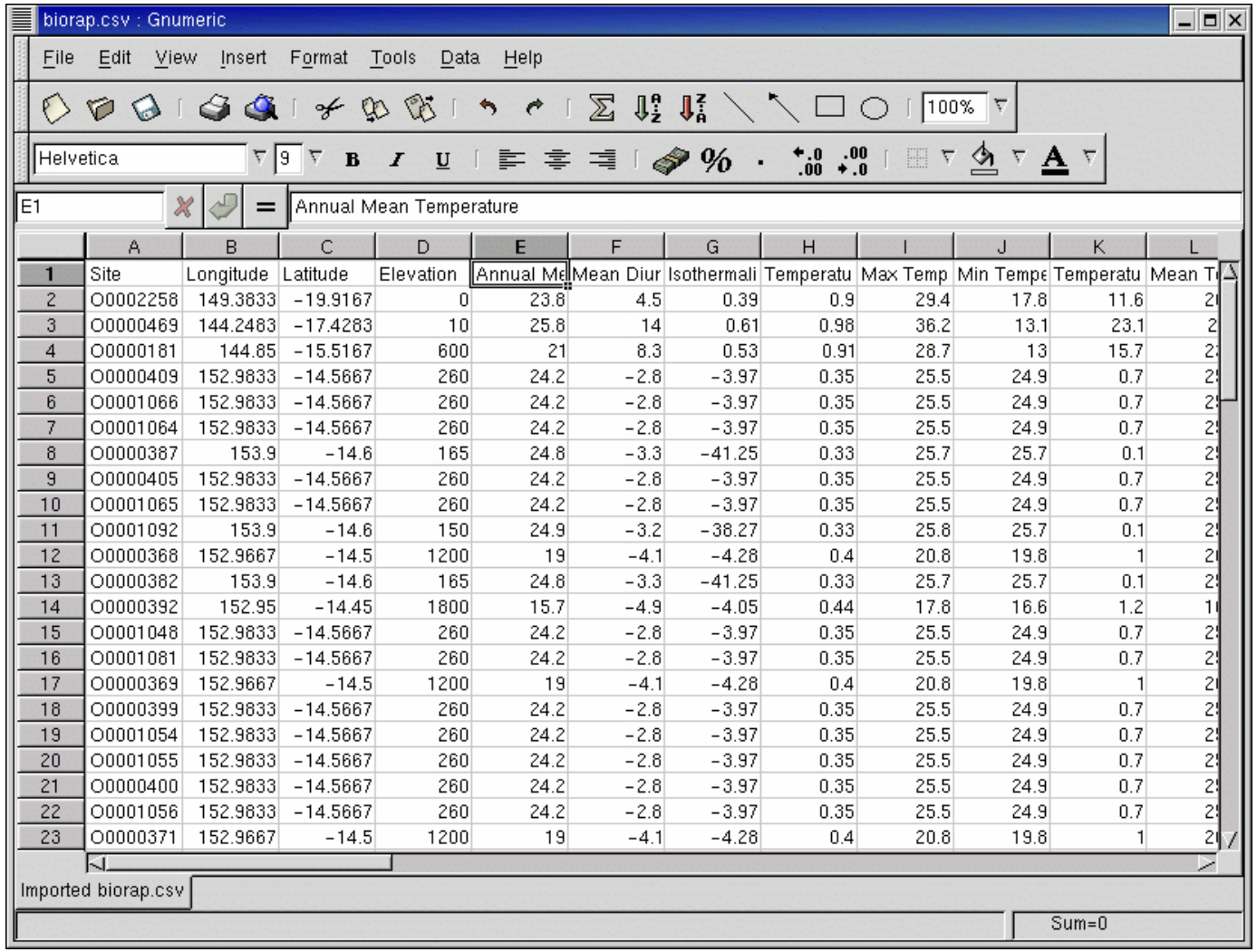
Other semantic axes as modifiers to "Informatics"

- Research ↔ Translational ↔ Clinical
  - Research- for the purpose of scientific discovery, not for use in clinical care
  - Translational- research related, but delivered (ultimately) to actual people
  - Clinical- as part of patient care
- Bio ↔ Clinical
  - Bio: molecular level
  - Clinical: patient/person level

At one end of the “data management” spectrum…

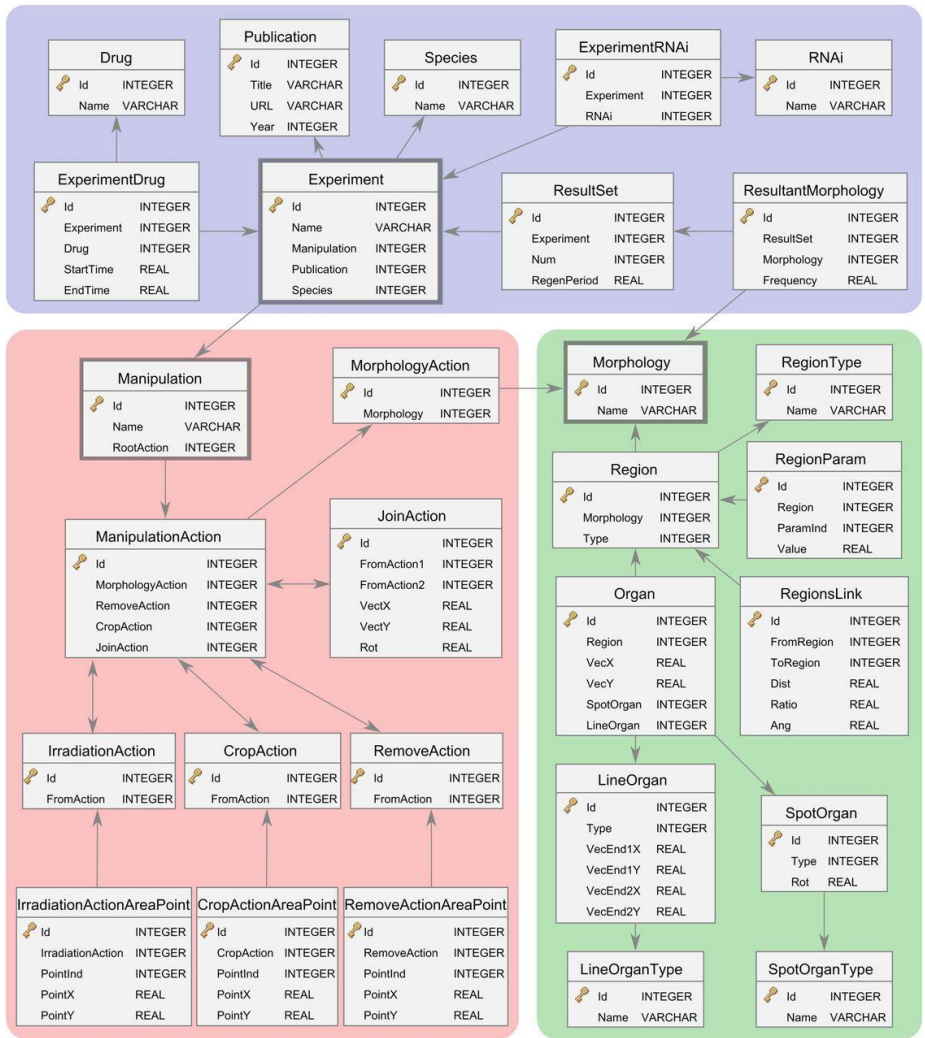


Somewhere in the middle…

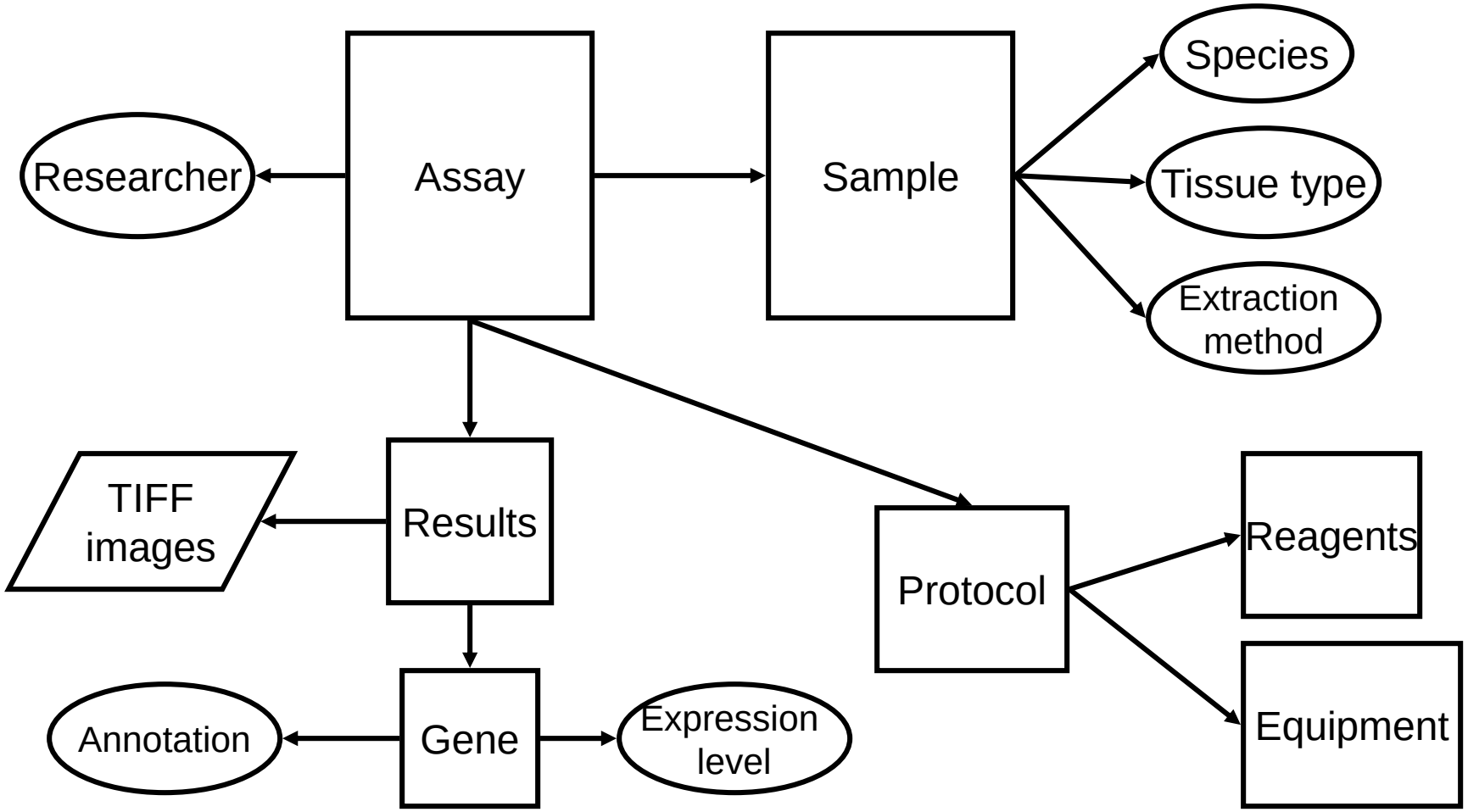


	A	B	C	D	E	F	G	H	I	J	K	L
1	Site	Longitude	Latitude	Elevation	Annual Me	Mean Diur	Isothermali	Temperatu	Max Temp	Min Tempe	Temperatu	Mean T
2	O0002258	149.3833	-19.9167	0	23.8	4.5	0.39	0.9	29.4	17.8	11.6	2
3	O0000469	144.2483	-17.4283	10	25.8	14	0.61	0.98	36.2	13.1	23.1	2
4	O0000181	144.85	-15.5167	600	21	8.3	0.53	0.91	28.7	13	15.7	2
5	O0000409	152.9833	-14.5667	260	24.2	-2.8	-3.97	0.35	25.5	24.9	0.7	2
6	O0001066	152.9833	-14.5667	260	24.2	-2.8	-3.97	0.35	25.5	24.9	0.7	2
7	O0001064	152.9833	-14.5667	260	24.2	-2.8	-3.97	0.35	25.5	24.9	0.7	2
8	O0000387	153.9	-14.6	165	24.8	-3.3	-41.25	0.33	25.7	25.7	0.1	2
9	O0000405	152.9833	-14.5667	260	24.2	-2.8	-3.97	0.35	25.5	24.9	0.7	2
10	O0001065	152.9833	-14.5667	260	24.2	-2.8	-3.97	0.35	25.5	24.9	0.7	2
11	O0001092	153.9	-14.6	150	24.9	-3.2	-38.27	0.33	25.8	25.7	0.1	2
12	O0000368	152.9667	-14.5	1200	19	-4.1	-4.28	0.4	20.8	19.8	1	2
13	O0000382	153.9	-14.6	165	24.8	-3.3	-41.25	0.33	25.7	25.7	0.1	2
14	O0000392	152.95	-14.45	1800	15.7	-4.9	-4.05	0.44	17.8	16.6	1.2	1
15	O0001048	152.9833	-14.5667	260	24.2	-2.8	-3.97	0.35	25.5	24.9	0.7	2
16	O0001081	152.9833	-14.5667	260	24.2	-2.8	-3.97	0.35	25.5	24.9	0.7	2
17	O0000369	152.9667	-14.5	1200	19	-4.1	-4.28	0.4	20.8	19.8	1	2
18	O0000399	152.9833	-14.5667	260	24.2	-2.8	-3.97	0.35	25.5	24.9	0.7	2
19	O0001054	152.9833	-14.5667	260	24.2	-2.8	-3.97	0.35	25.5	24.9	0.7	2
20	O0001055	152.9833	-14.5667	260	24.2	-2.8	-3.97	0.35	25.5	24.9	0.7	2
21	O0000400	152.9833	-14.5667	260	24.2	-2.8	-3.97	0.35	25.5	24.9	0.7	2
22	O0001056	152.9833	-14.5667	260	24.2	-2.8	-3.97	0.35	25.5	24.9	0.7	2
23	O0000371	152.9667	-14.5	1200	19	-4.1	-4.28	0.4	20.8	19.8	1	2

Relational database…



…Enables queriable knowledge base



Trend toward data sharing

Science 9 October 2009:
Vol. 326. no. 5950, pp. 234 - 236
DOI: 10.1126/science.1180598

POLICY FORUM

MEGASCIENCE: 'Omics Data Sharing

Dawn Field,[1,*,†,‡] Susanna-Assunta Sansone,[1,2,†] Amanda Collis,[3,†] Tim Booth,[1] Peter Dukes,[4] Susan K. Gregurick,[5] Karen Kennedy,[6] Patrik Kolar,[7] Eugene Kolker,[8] Mary Maxon,[9] Siân Millard,[10] Alexis-Michel Mugabushaka,[11] Nicola Perrin,[12] Jacques E. Romacle,[7] Karin Remington,[13] Philippe Rocca-Serra,[12] Chris

Science

AAAS

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: Th
Principles for
management

Mark D. Wilkinson *et al.*[#]

December 10, 2009

, and the good annotation
epends on, must become part
of daily research for researchers

nature biotechnology

First, design for data sharing

John Wilbanks & Stephen H Friend

To upend current barriers to sharing clinical data and insights, we need a framework that not only accounts for choices made by trial participants but also qualifies researchers wishing to access and analyze the data.

Byline: a GenomeWeb
Newsletter: GenomeWeb Daily News
GenomeWeb Daily News - December 10, 2009

NEW YORK (GenomeWeb News) – The White House plans to make data and information from federally-funded research available for public access and use as part of the Open Government Directive that President Barack Obama announced this week.

FAIR Principles

- **Findable-** appropriate metadata, indexed and persistent identifier
- **Accessible-** retrievable with standard open protocol, metadata available independent of data
- **Interoperable-** use standards for terminologies, knowledge representation
- **Re-usable-** provenance, licensing, follow community standards

Data Standards

What are they, and why should we use them?

Data Standards

- An agreement
- Provide specifications that permit sharing of data

Adapted from MN Zozus

Behavior colors:

Purple-great day!

Blue- a warning-no consequence

Yellow- note home; time out

Red-phone call home

Behavior colors:

Green-great day!

Yellow- a warning-no consequence

Orange- note home; time out

Red-phone call home

*** Blue-something happened today that was SUPER!!

3 Types of Data Standards

- Content standards
  - What data will be collected
- Semantic standards
  - How variables are named and defined, how they relate to each other
- Transmission standards
  - How data is transferred among machines

Adapted from MN Zozus

What about these other data standards?

- Common Data Elements (CDEs)
- Information models
- Domain analysis models
- Data models

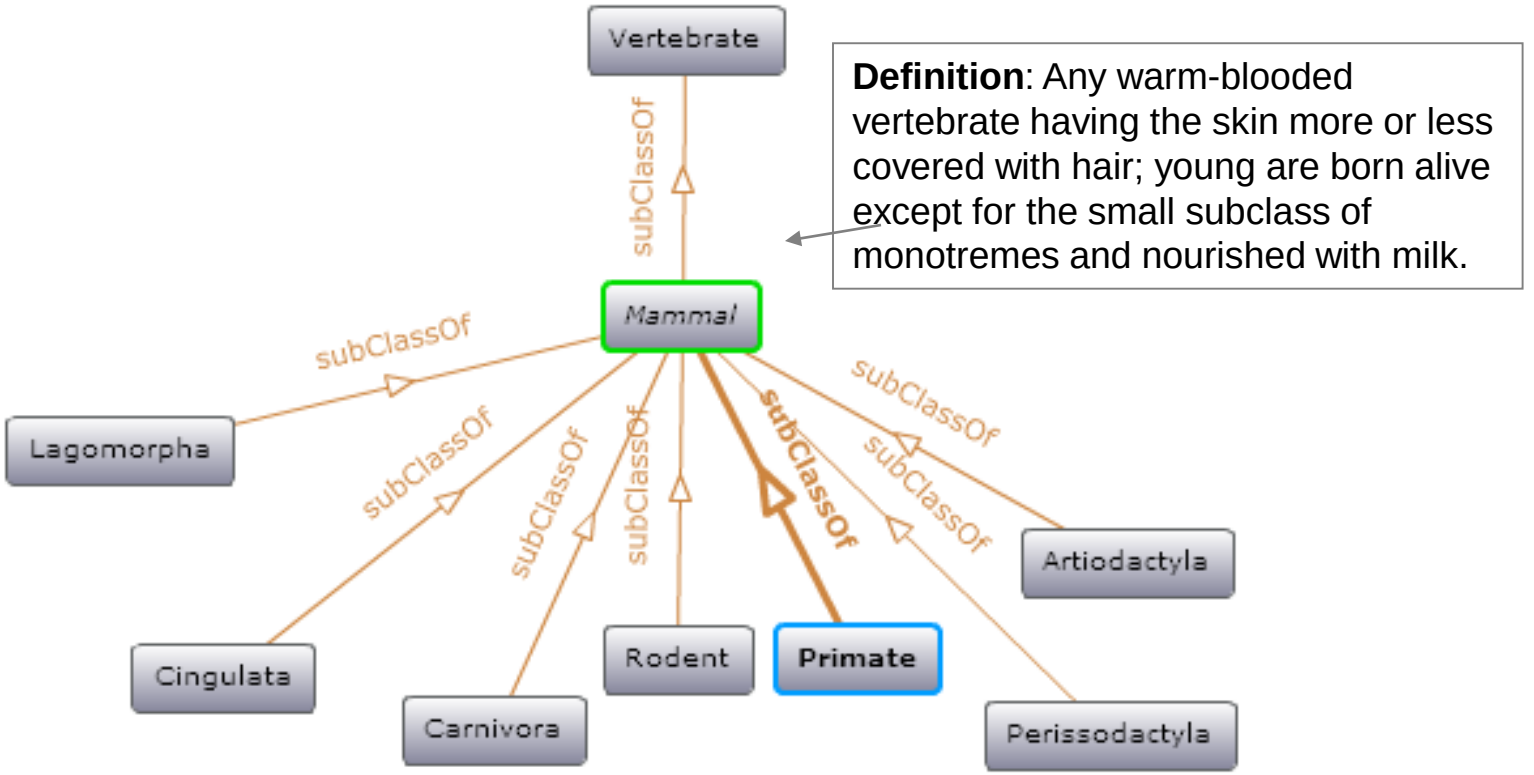
- All essentially some combination of semantics, content, format.

Content

- *Enumeration* of data elements
- “Minimum Information” lists
- E.g. MIAME- Minimum Information about a Microarray Experiment- 6 critical elements:
  1. Raw data
  2. Processed/normalized data
  3. Sample annotation
  4. Experimental design
  5. Array annotation
  6. Lab and data processing protocols

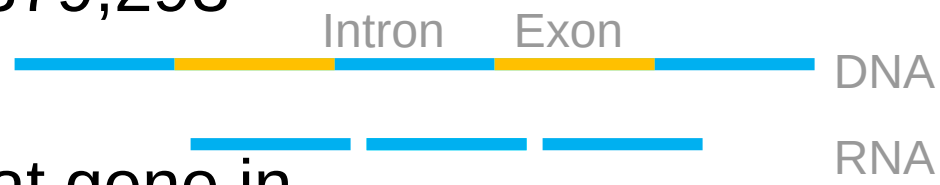
Semantics: The meaning

Ontology: a formal representation of a set of concepts within a domain and the relationships between those concepts.



Semantic illustration: “gene”

- Terminology:
  - Toll-like receptor 4
  - TLR4, tlr4, TLR-4
  - Toll-like receptor 4 in humans
  - ACTGATCAGGATCAGATCATCGCATTTACCCA…
  - Chromosome 9 positions 345,293 to 379,293
- Meaning:
  - All exons and introns that make up that gene in human
  - Only the protein coding parts of the DNA
  - The RNA that is measured in gene expression



Syntax

- Exchange format
- E.g. XML

```
<gene>
    <name=tlr4>
    <gbacc=NC_000070.5 >
    <loc=9q345,987>
</gene>
```

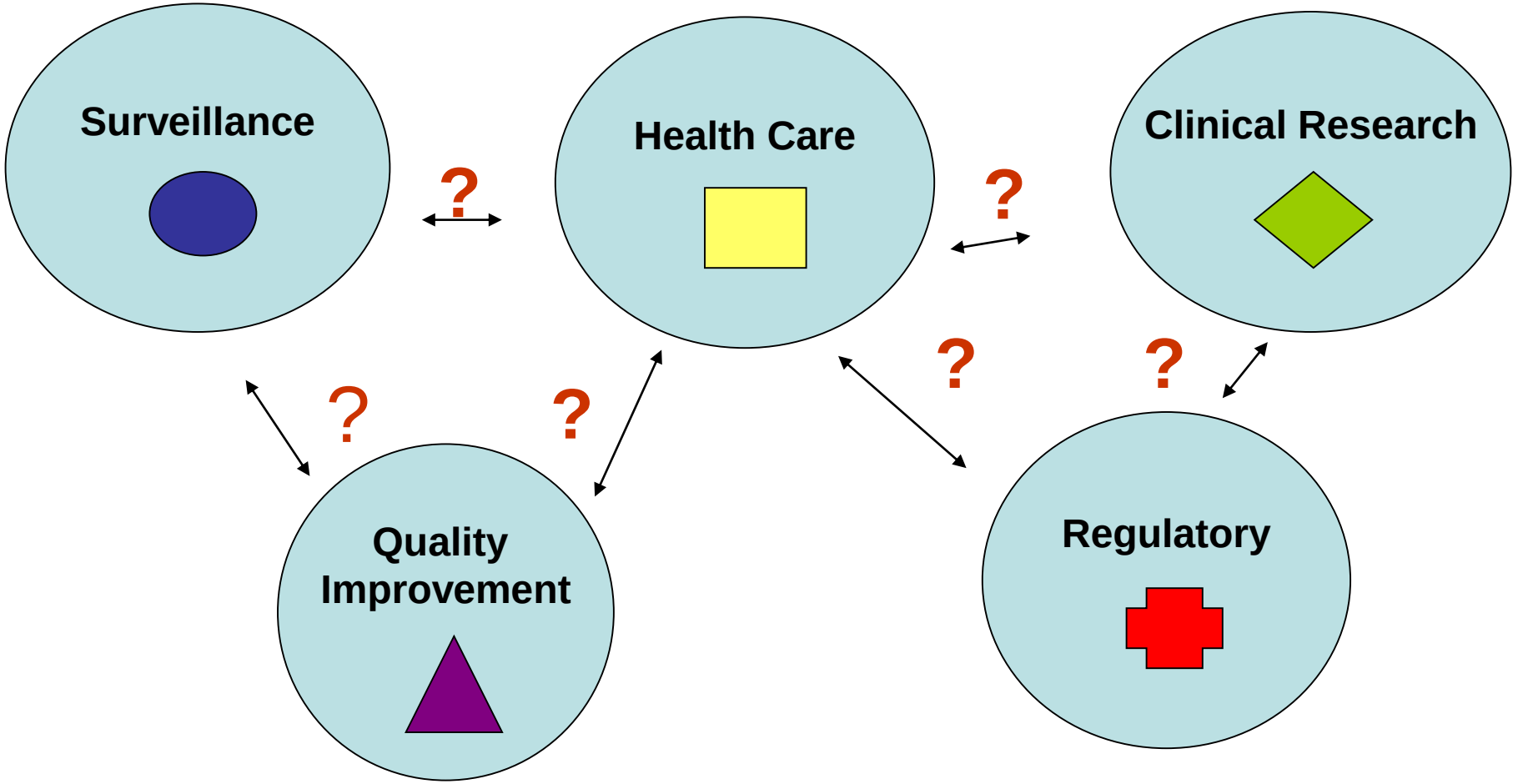
- E.g. tabular format
  - Rows are subjects, columns are lab values
  - Rows are genes, columns are subjects

Reasons to Use Standards

- Data more easily shared among researchers to answer broader questions
- Easier pooling of data from multiple sources for secondary analysis
- Data more easily combined across studies
- Lifetime & return on investment for data increased
  - Data can be used to answer other research questions
- Enhanced visibility and references
- NIH requirements

Adapted from MN Zozus

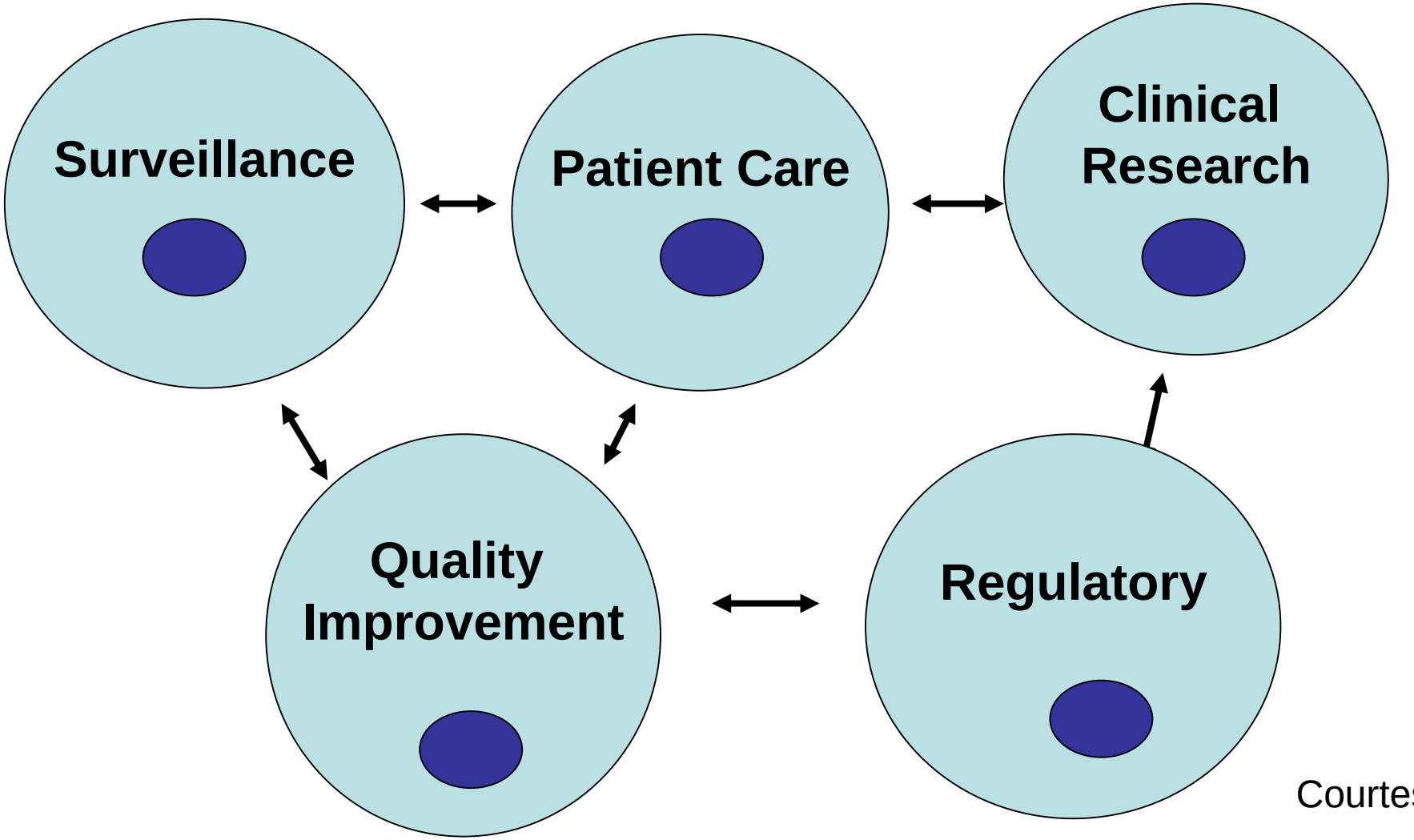
Data Exchange May Be Impossible



Carol Dukes Hamilton, MD, NIH Roadmap PI

Courtesy of MN Zozus

Data Exchange Possible



Courtesy of MN Zozus

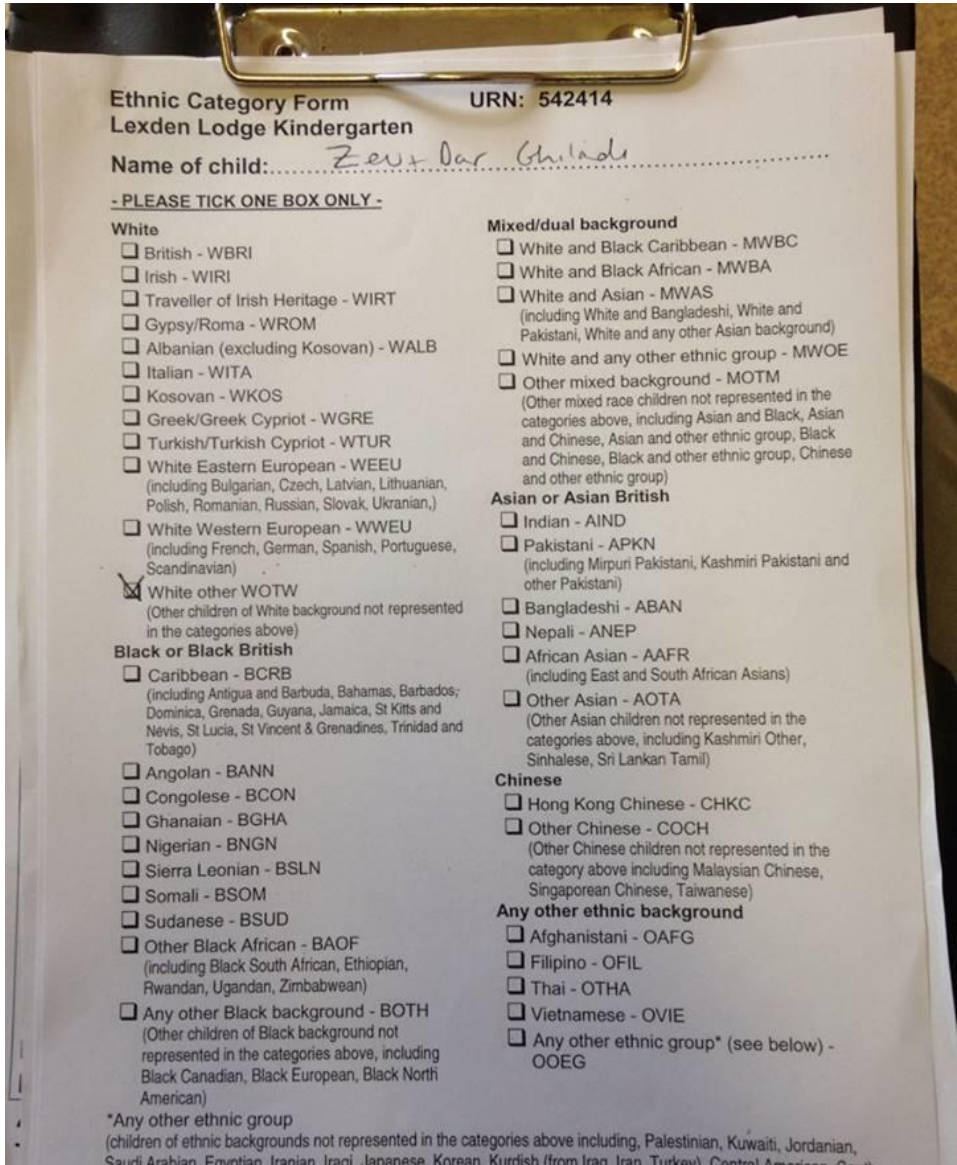
Carol Dukes-Hamilton, MD, NIH Roadmap PI

Example: Smoking status

- Do you smoke?
  - Yes/No
  - Never/Sometimes/Frequently
  - No, never have/No, quit/Yes
- How much do you smoke?
  - Never/Sometimes/Frequently
  - Occasionally/1 pack per day/More than 1 pack per day

Example: Race & Ethnicity

- NIH:
  - American Indian or Alaska Native
  - Asian
  - Black or African American
  - Hispanic or Latino
  - Native Hawaiian or Other Pacific Islander
  - White



Ethnic Category Form URN: 542414
Lexden Lodge Kindergarten
Name of child: Zeus Dar Ghilardi

- PLEASE TICK ONE BOX ONLY -

White
- ☐ British - WBRI
- ☐ Irish - WIRI
- ☐ Traveller of Irish Heritage - WIRT
- ☐ Gypsy/Roma - WROM
- ☐ Albanian (excluding Kosovan) - WALB
- ☐ Italian - WITA
- ☐ Kosovan - WKOS
- ☐ Greek/Greek Cypriot - WGRE
- ☐ Turkish/Turkish Cypriot - WTUR
- ☐ White Eastern European - WEEU (including Bulgarian, Czech, Latvian, Lithuanian, Polish, Romanian, Russian, Slovak, Ukranian,)
- ☐ White Western European - WWEU (including French, German, Spanish, Portuguese, Scandinavian)
- ☒ White other WOTW (Other children of White background not represented in the categories above)

Black or Black British
- ☐ Caribbean - BCRB (including Antigua and Barbuda, Bahamas, Barbados, Dominica, Grenada, Guyana, Jamaica, St Kitts and Nevis, St Lucia, St Vincent & Grenadines, Trinidad and Tobago)
- ☐ Angolan - BANN
- ☐ Congolese - BCON
- ☐ Ghanaian - BGHA
- ☐ Nigerian - BNGN
- ☐ Sierra Leonian - BSLN
- ☐ Somali - BSOM
- ☐ Sudanese - BSUD
- ☐ Other Black African - BAOF (including Black South African, Ethiopian, Rwandan, Ugandan, Zimbabwean)
- ☐ Any other Black background - BOTH (Other children of Black background not represented in the categories above, including Black Canadian, Black European, Black North American)

Mixed/dual background
- ☐ White and Black Caribbean - MWBC
- ☐ White and Black African - MWBA
- ☐ White and Asian - MWAS (including White and Bangladeshi, White and Pakistani, White and any other Asian background)
- ☐ White and any other ethnic group - MWOE
- ☐ Other mixed background - MOTM (Other mixed race children not represented in the categories above, including Asian and Black, Asian and Chinese, Asian and other ethnic group, Black and Chinese, Black and other ethnic group, Chinese and other ethnic group)

Asian or Asian British
- ☐ Indian - AIND
- ☐ Pakistani - APKN (including Mirpuri Pakistani, Kashmiri Pakistani and other Pakistani)
- ☐ Bangladeshi - ABAN
- ☐ Nepali - ANEP
- ☐ African Asian - AAFR (including East and South African Asians)
- ☐ Other Asian - AOTA (Other Asian children not represented in the categories above, including Kashmiri Other, Sinhalese, Sri Lankan Tamil)

Chinese
- ☐ Hong Kong Chinese - CHKC
- ☐ Other Chinese - COCH (Other Chinese children not represented in the category above including Malaysian Chinese, Singaporean Chinese, Taiwanese)

Any other ethnic background
- ☐ Afghanistani - OAFG
- ☐ Filipino - OFIL
- ☐ Thai - OTHA
- ☐ Vietnamese - OVIE
- ☐ Any other ethnic group* (see below) - OOEG

*Any other ethnic group
(children of ethnic backgrounds not represented in the categories above including, Palestinian, Kuwaiti, Jordanian, Saudi Arabian, Egyptian, Iranian, Iraqi, Japanese, Korean, Kurdish (from Iraq, Iran, Turkey), Central American,

Standards and Your Science

- Standards offer a way to share results and increase scientific productivity.
- Standards should NOT inhibit or alter the collection of data.

If the standard doesn’t fit the science,

Don’t use it.

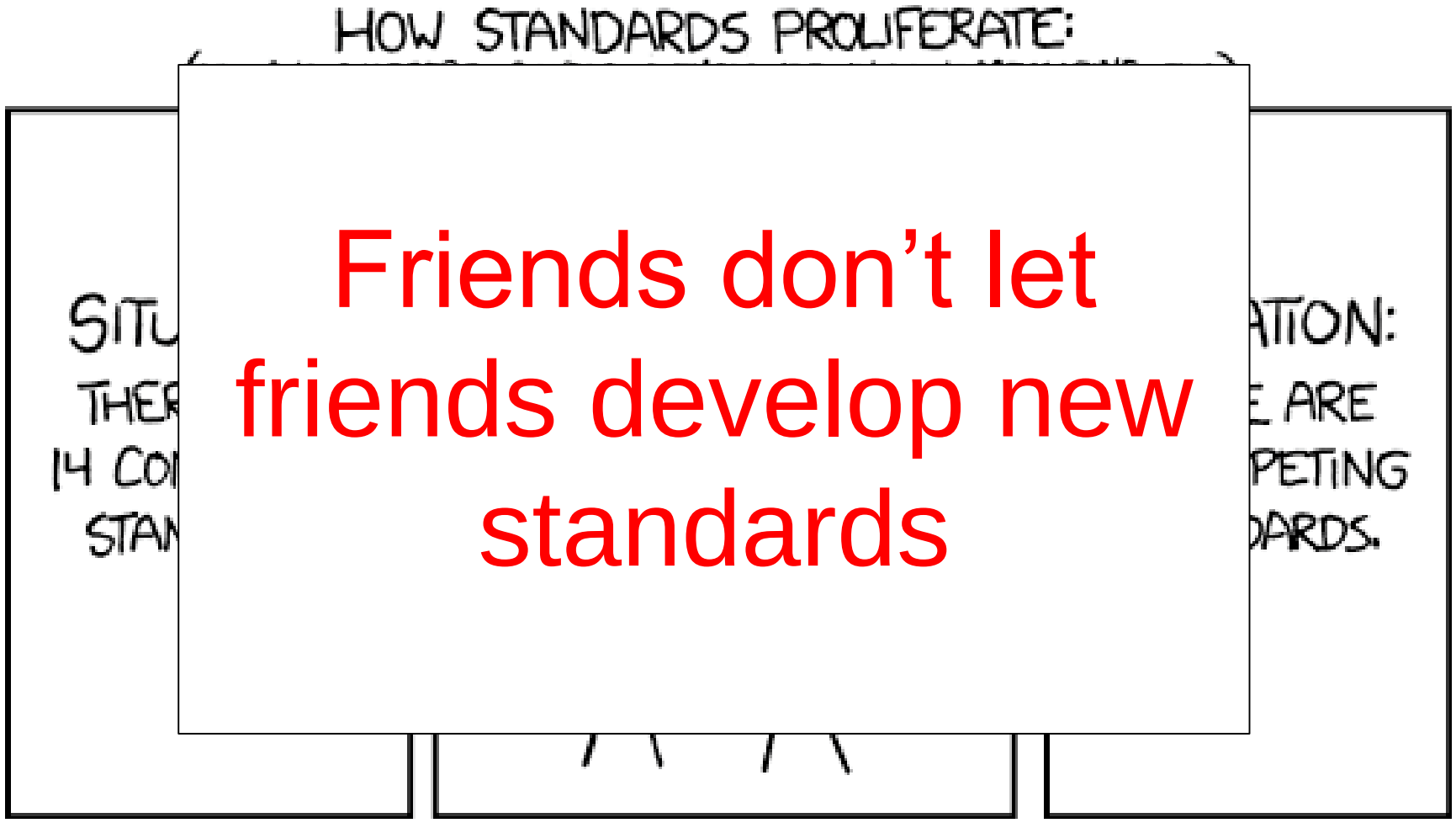
A few other notes about standards…

- Dynamic and ever-changing landscape
- Lots of politics at play
- Rarely any funding for their development
  - Developed by the volunteers who show up
  - Often as part of a specific project, for which funding then ends…

Implementing Standards: Challenges

- Time
  - Identifying applicable standards
  - Implementing use of standards in processes and workflow
- Expense
  - Implementation can increase study costs
- If standards do not exist, you will need to spend time and expense to develop them

Comic relief



Friends don’t let friends develop new standards

Relevant Data standards for DNA Sequences

- Content
  - MINSEQE (like MIAME)
- Terminology
  - Gene Ontology
  - Sequence Ontology
- Format
  - FASTA, FASTQ
  - SAM/BAM
  - VCF

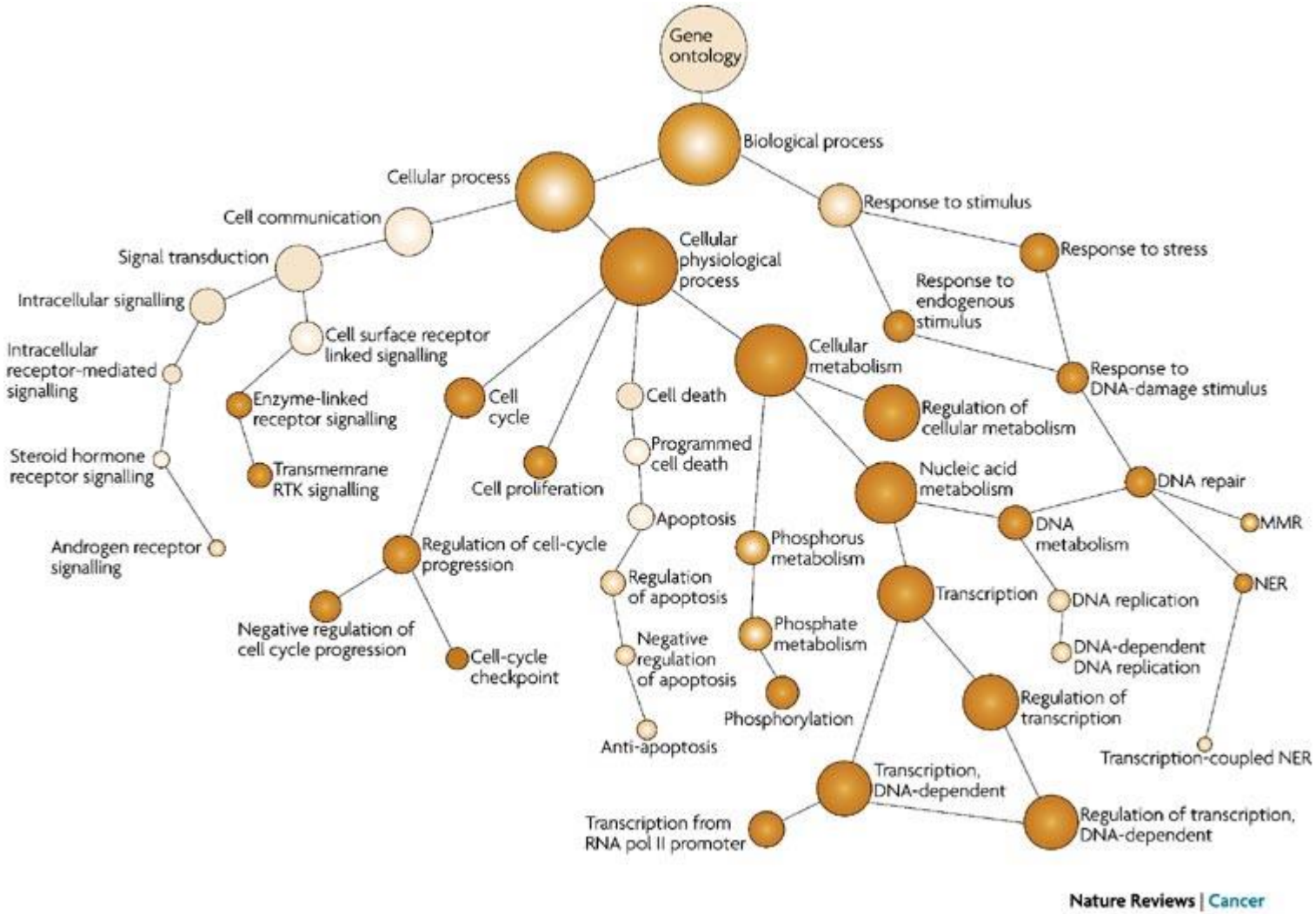
MINSEQE

1. Description of the biological system, samples, experimental variables
2. Sequence data for each assay- read sequences, base-level quality scores
3. ‘Final’ processed (or summary) data- the data on which the conclusions are based, and descriptions of the data format
4. General information about the experiment and sample-data relationships
5. Essential experimental and data processing protocols

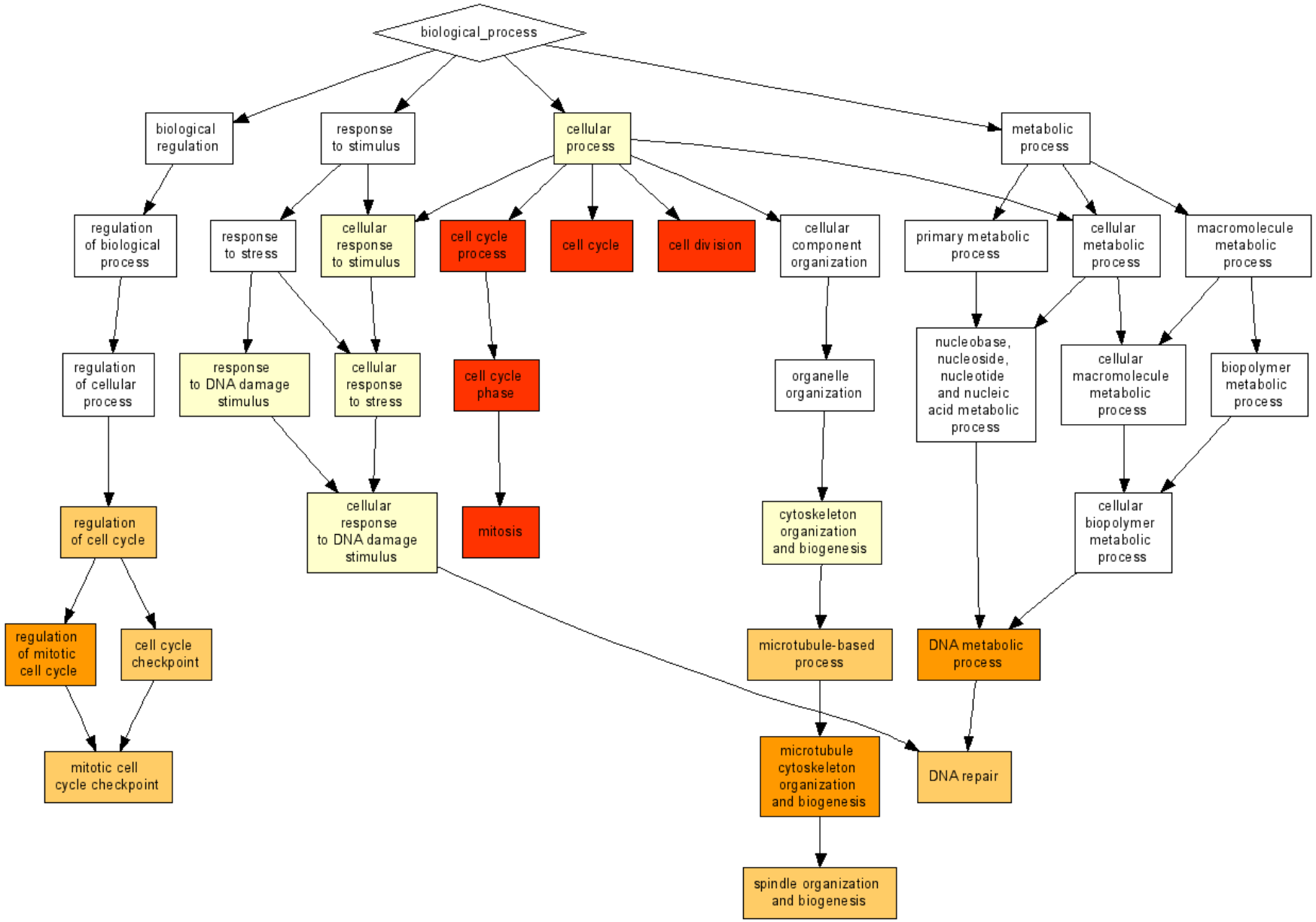
Gene Ontology

- Defines concepts/classes used to describe gene function, and relationships between these concepts.
  1. molecular function- molecular activities of gene products
  2. cellular component- where gene products are active
  3. biological process- pathways and larger processes made up of the activities of multiple gene products
- Enables “GO Enrichment” analysis
  - Given a set of genes that are up-regulated under a certain condition, what GO terms are over-represented among annotations of those genes?

Section of biological process branch of GO

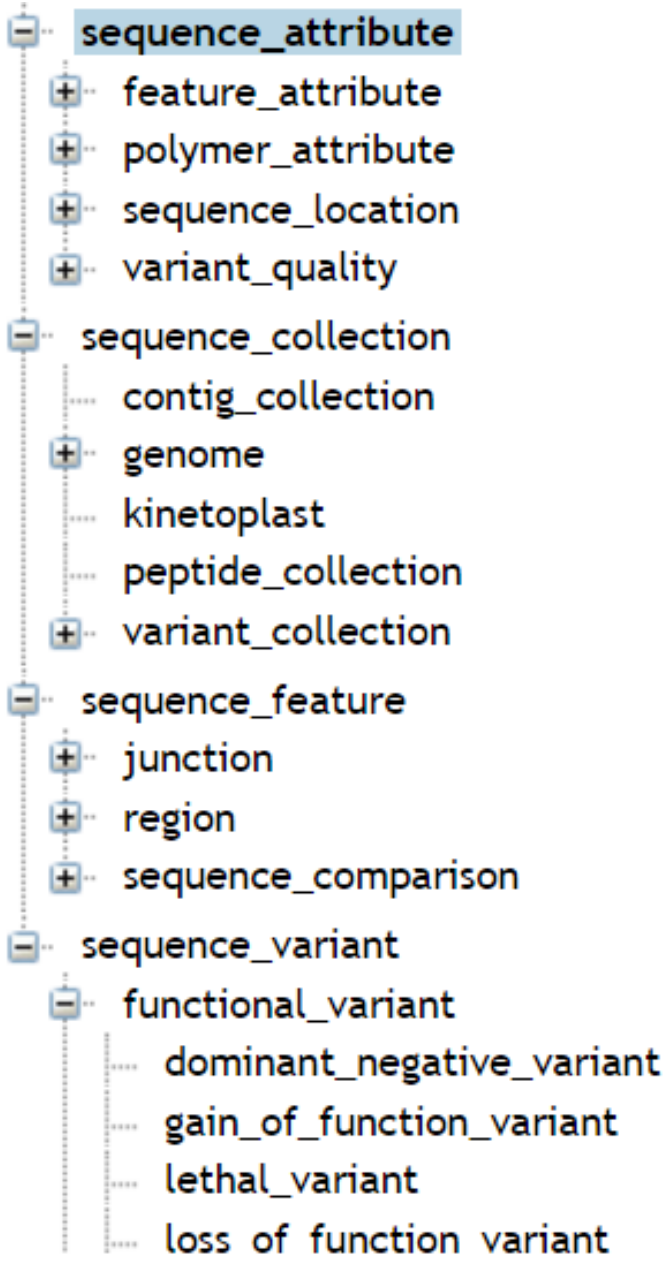


GO Enrichment

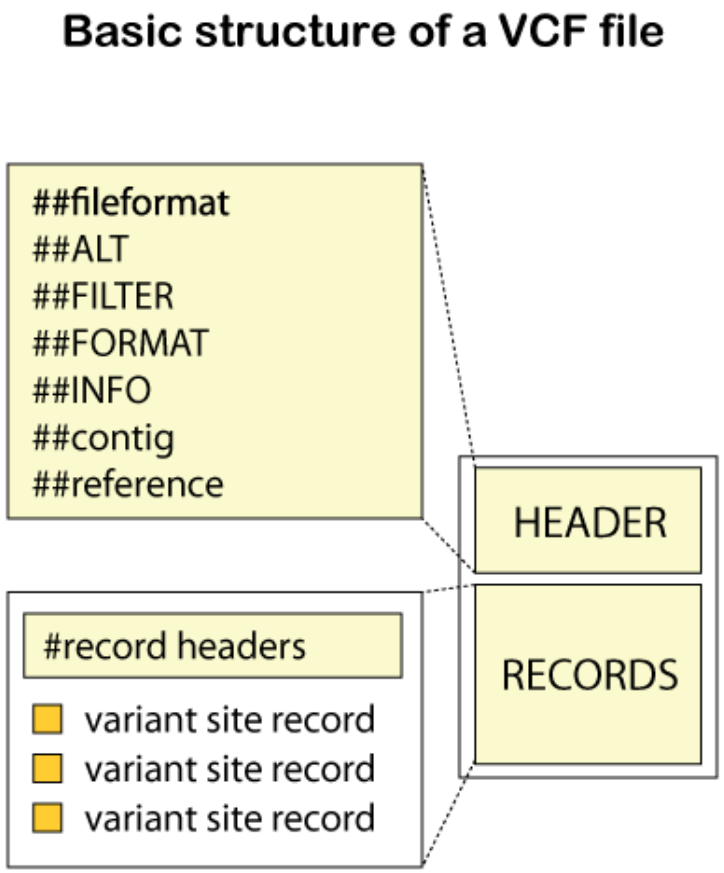


Sequence Ontology

- What it is
  - Set of terms and relationships used to describe the features and attributes used in biological sequence annotation
  - Part of OBO (Open Biomedical Ontologies) Library
  - View in BioPortal
- What it enables (in theory)
  - Query all sequence databases for e.g. all genes whose transcripts are edited, or trans-spliced, or are bound by a particular protein.



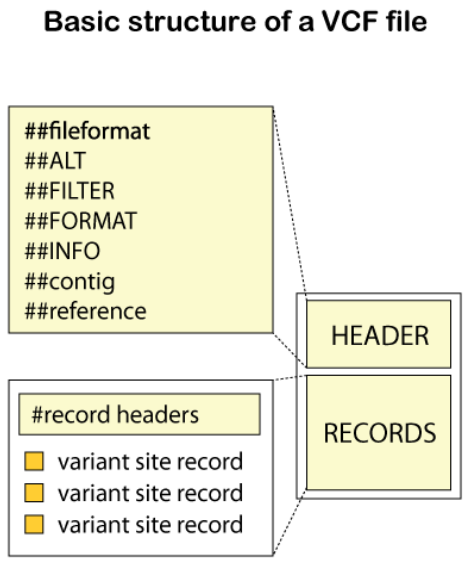
VCF



https://gatkforums.broadinstitute.org/gatk/discussion/1268/what-is-a-vcf-and-how-should-i-interpret-it

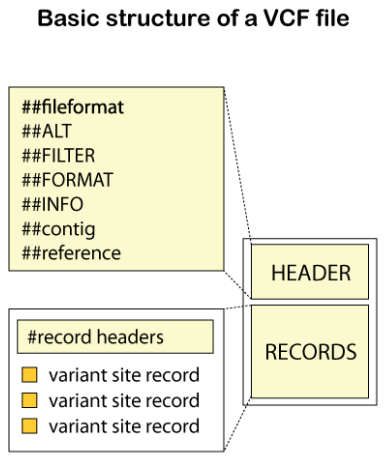
VCF- variant call format

- Text file format- human readable
- Metadata lines (## key = value)
- Header line- names 8 fixed, mandatory columns
  - #CHROM- Chromosome
  - POS- position
  - ID- unique identifiers where available
  - REF- reference base(s)
  - ALT- alternate base(s)
  - QUAL- quality score
  - FILTER- filter status- pass, or how many failed what filter
  - INFO- additional information- standard list or specified in metadata



VCF example

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```



Exercise: Identify standards in genomics

Standard	Type
MIAME	Reporting guideline
ISA-TAB	Exchange format
MAGE-TAB	Exchange format
MAGE-ML	Exchange format
SOFT	Exchange format
MIMiML	Exchange format
GO	Terminology artifact
EFO	Terminology artifact
OBI	Terminology artifact
MGED Ontology	Terminology artifact
MAGE-OM	Object model
FuGE	Object model
SEND	Exchange format
GEML	Exchange format
FUGO	Terminology artifact
MAML	Exchange format

Use cases

Level of Rigor	Use case example	Explanation
Low	Inter-lab collaboration	Data should meet minimal standards for structure and documentation to enable comprehension, but answers to questions are just an email/phone call/hallway away. At least until that person leaves the lab.
Medium	Publishing	Data should use standardized formats and annotation sufficient to enable both comprehension and reproducibility, with little or no interaction with the data owner.
High	Make available through public data repository	In addition to being comprehensible and reproducible, annotation should be structured in a way that enables querying for datasets that match specific criteria.

The Punchline[s]

1. Many different definitions for what constitutes a ‘data standard’.
2. No one standard is the ‘right’ standard- depends on particular needs.
3. Resources are needed to help researchers navigate the standards landscape

Standards Criteria

- The standard itself
- Adoption and user community
- Additional factors

The Standard Itself

- Specification documentation
- Ease of implementation (e.g. need for programmer support)
- Human and machine readability
- Formal structure
- Expressivity—breadth of information that can be represented
- Ease of use, e.g., minimal required fields, text-based interface familiarity to biologists.

Adoption and User Community

- Broad adoption and implementation
- Support supplied by the user community
- Use by community databases
- Software development that supports the standard
- Responsiveness to community requests
- Availability of examples of use
- Requirements of relevant authoritative bodies, e.g. funders, publishers, etc.

Additional Factors

- Integration/compatibility with other standards
- Extensibility and flexibility to cover new domains
- Conversion and mapping, when applicable
- Cost (e.g., open vs. licensing fee)

In Summary

- Standards are useful to enable exchange and reuse of data
- There are 3 main types of standards- content, semantic, format
- There is (alas) no single standard that is appropriate for all use cases
- The standards landscape is dynamic and changing
- There are resources that can help! (next lecture…)

Questions?

Jessie.Tenenbaum@duke.edu