Bringing Phenotype to the Genotype: Semantics for Maximizing Disease Discovery

Melissa Haendel, PhD

March 31, 2020

These slides: bit.ly/semantics-space

@ontowonka
@MonarchInit

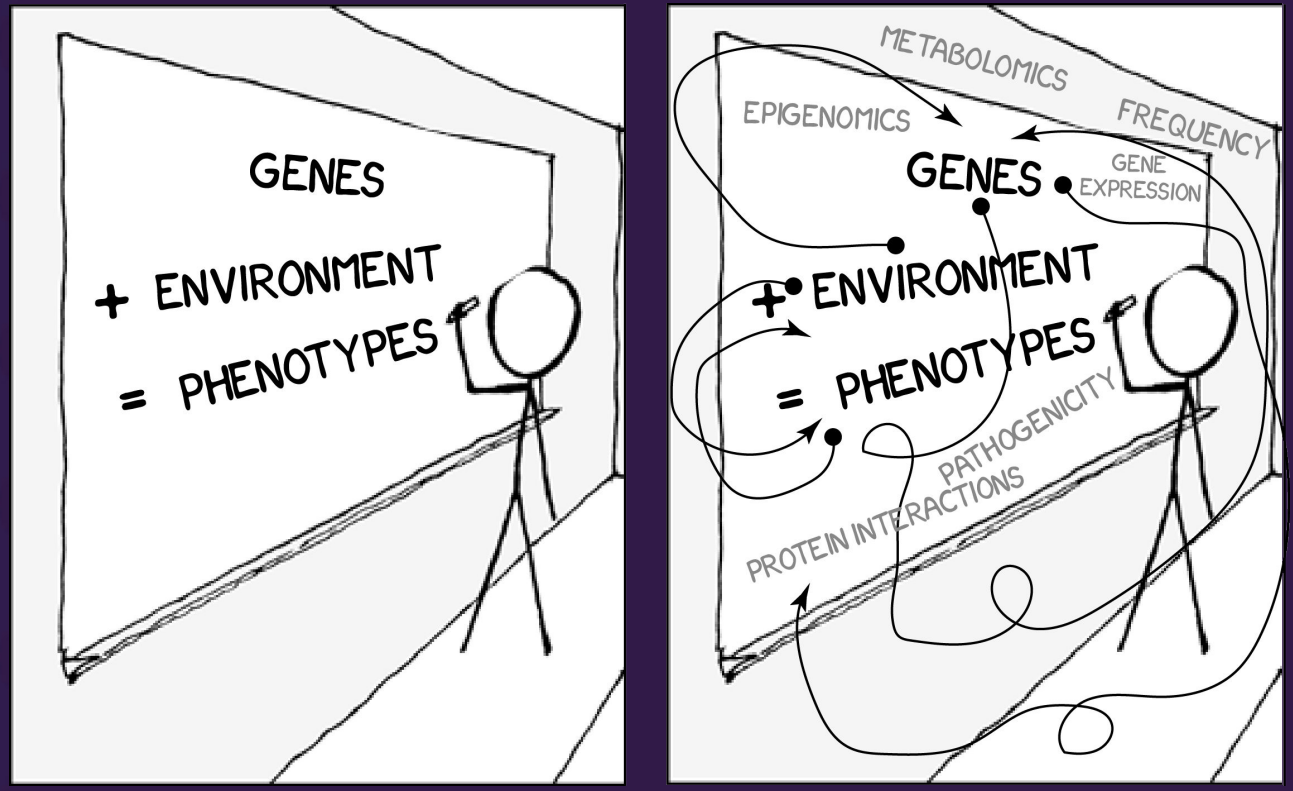
Jessica's story:
Evidence-based diagnosis requires evidence

Jessica (aged 4) has a rare condition which causes epilepsy, affects her movement and developmental delay. Standard genetics tests negative.

To solve her case requires the ability to compare Jessica to a multitude of other available data, both from humans and from other animals.

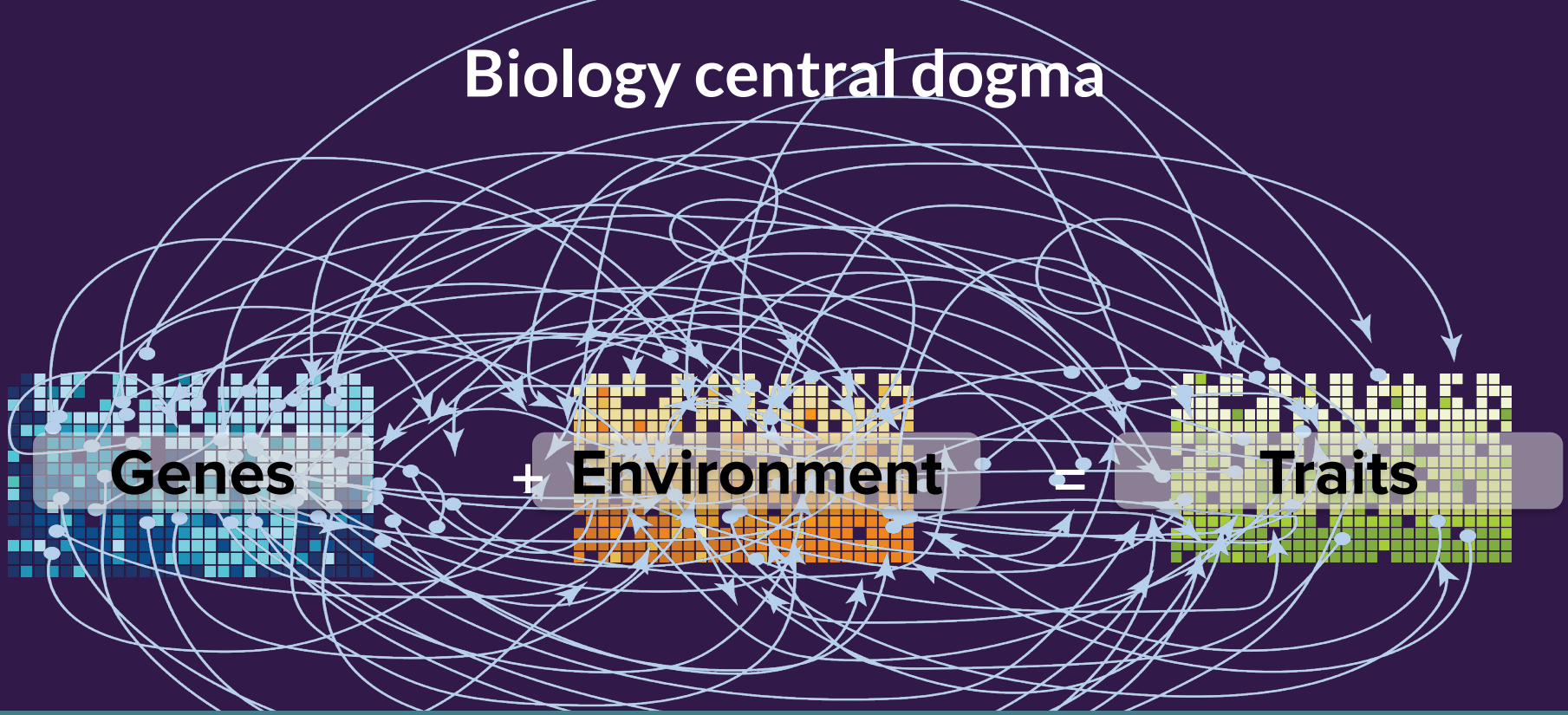
Genomics england

Biology central dogma



ADAPTED FROM xkcd.com/295

Biology central dogma



Standards for encoding and exchanging data must be up to these challenges.

It is not just the bits...

G-P or D (disease)
causes
contributes to
is risk factor for
protects against
correlates with
is marker for
modulates
involved in
increases susceptibility to

G-G (kind of)
regulates
negatively regulates (inhibits)
positively regulates (activates)
directly regulates
interacts with
co-localizes with
co-expressed with

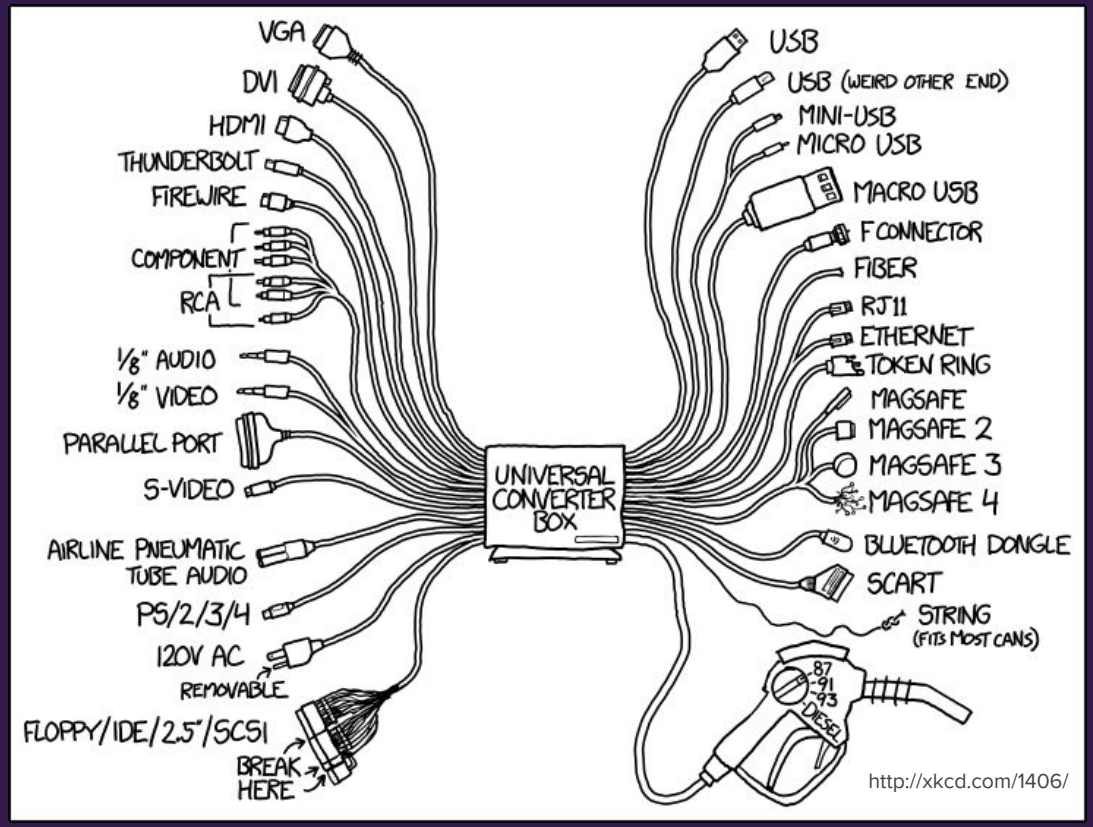
P/D - P/D
correlates with
hallmark of (P->D)

E-P
contributes to (E->P)
influences (E->P)
exacerbates (E->P)
manifest in (P->E)

G-E (kind of)
expressed in
expressed during

The relationships too must be captured

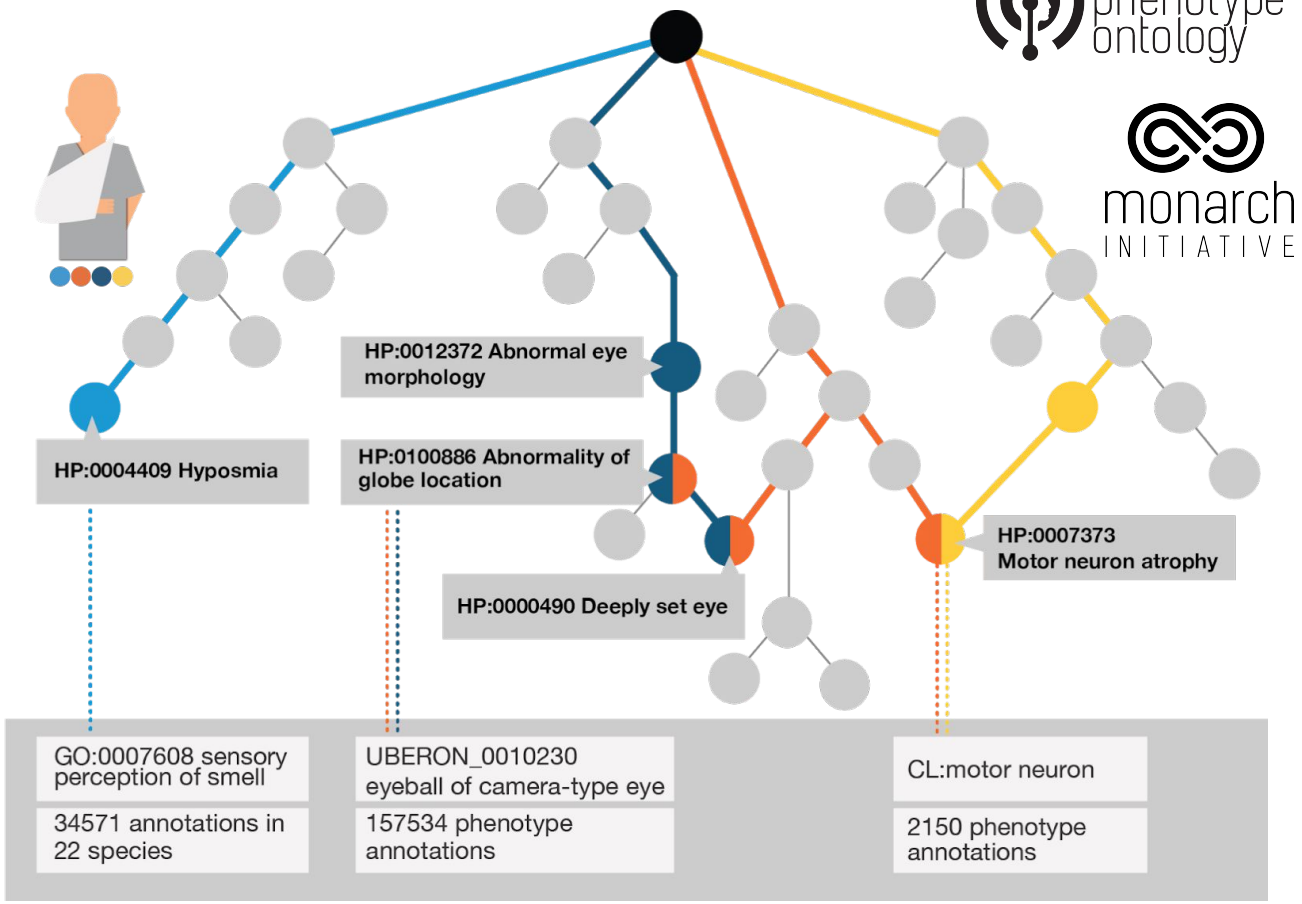
Semantics are the ultimate universal converter



Human Phenotype Ontology (HPO)

A standardized, machine readable vocabulary of phenotypic abnormalities encountered in human disease.

- Over 14,000 phenotype terms
- It is used to create computational models of disease



hpo.jax.org

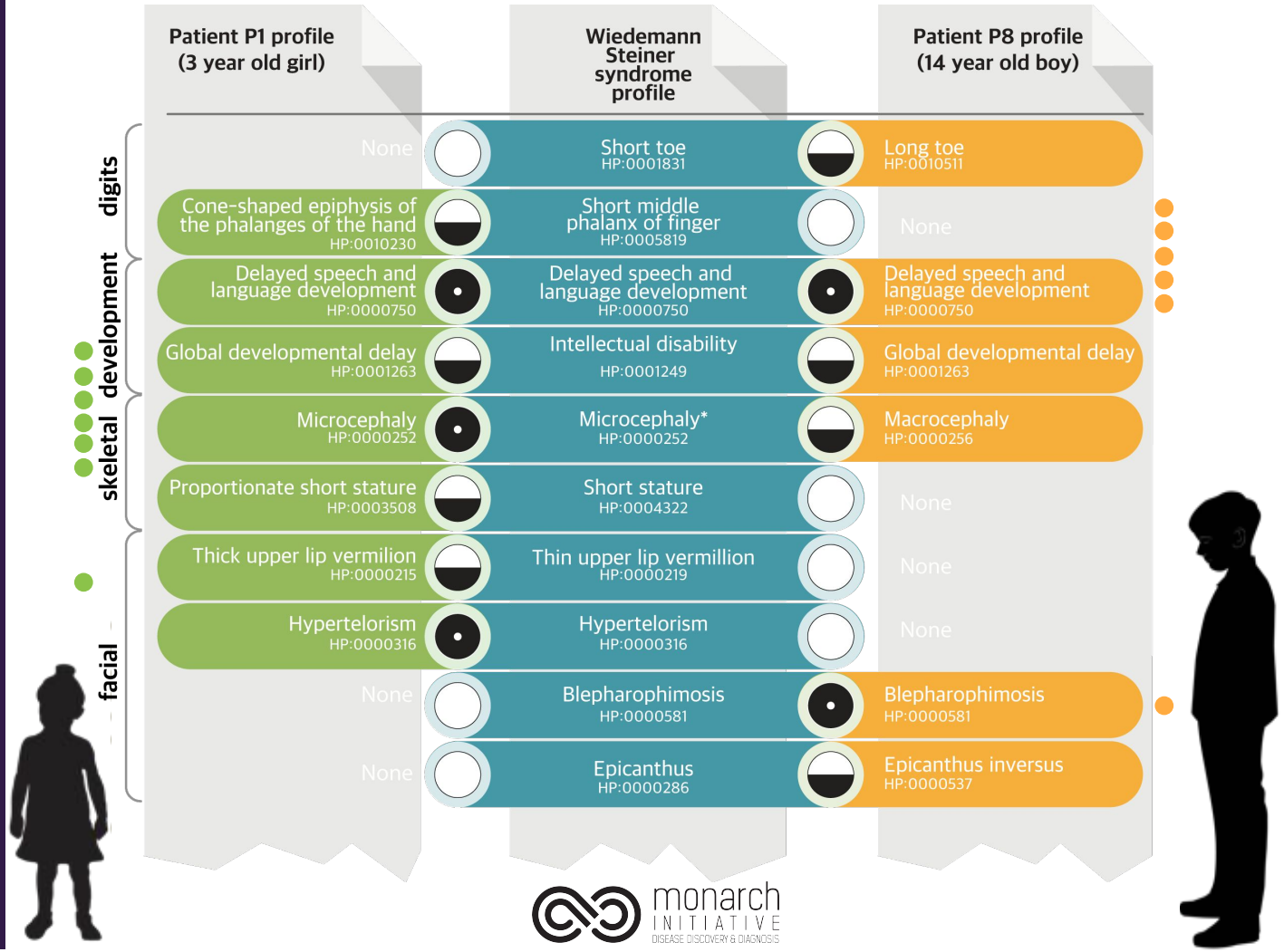
Fuzzy Phenotype Matching

Not same variant, but same disease and gene, KMT2A.

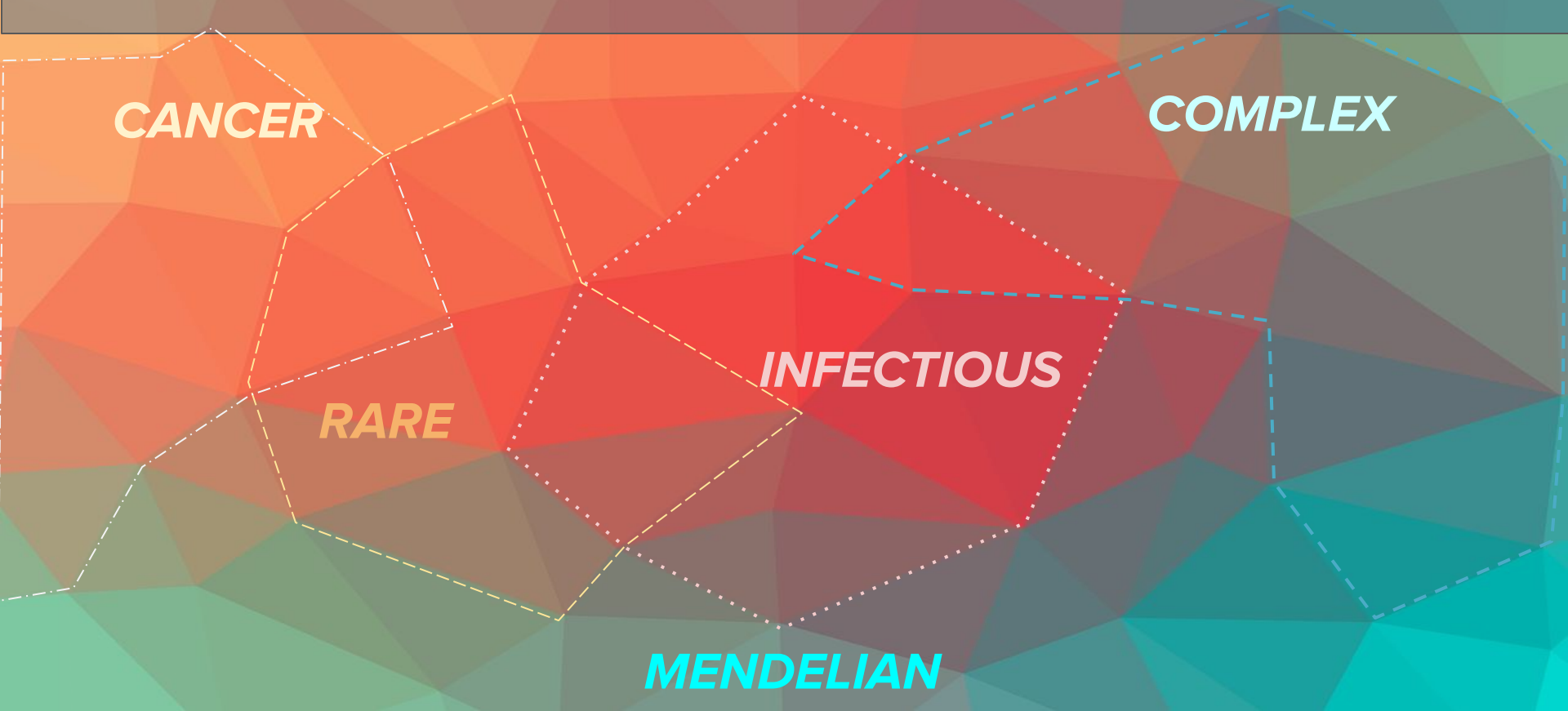
DOI: 10.1126/scitranslmed.3009262

Legend

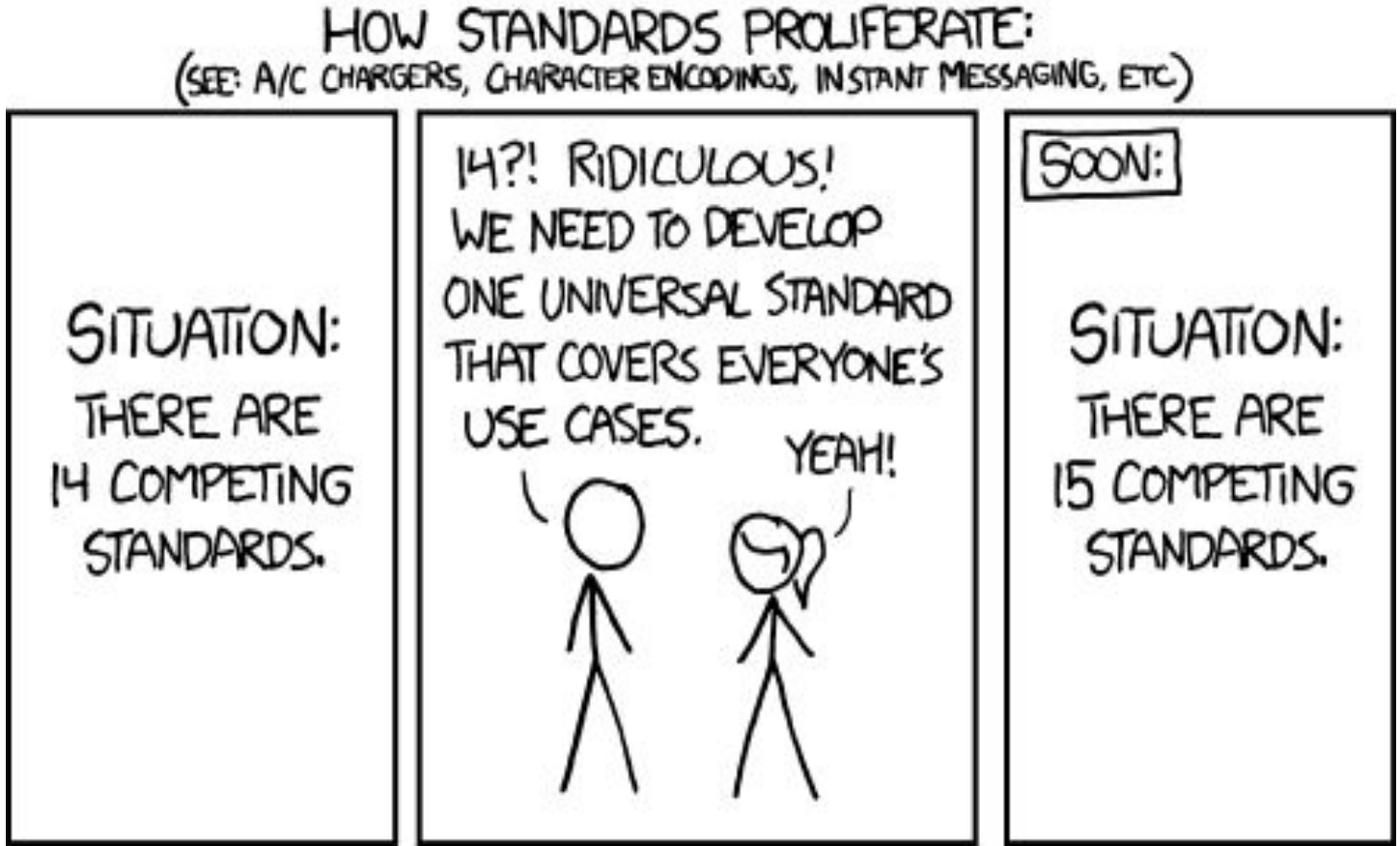
- Perfect Match
- Fuzzy Match
- No Match



What is the most clinically useful way to define and group diseases?



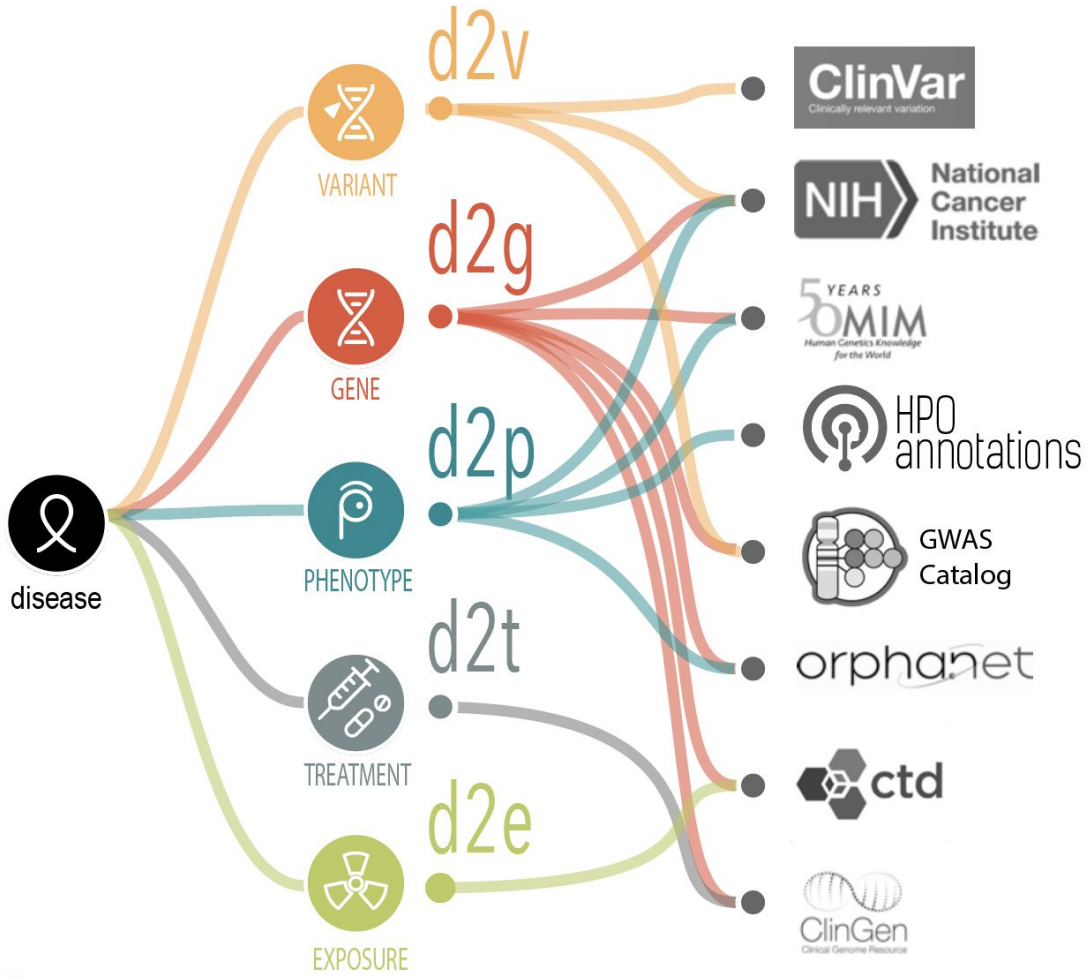
Standards proliferation: how do you know you need a new one?



xkcd.com/927

Standards proliferation: how do you know you need a new one?

Different communities define different aspects of diseases differently



The NEW ENGLAND JOURNAL of MEDICINE

Classification, Ontology, and Precision Medicine

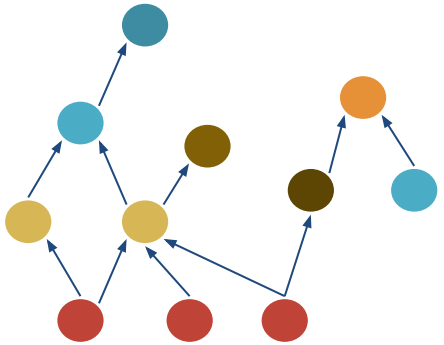
Evidence-based merging of equivalent disease concepts

kBOOM

Bayesian
OWL
Ontology
Merging

Logical +
Probabilistic
Inference

iterative curator-assisted equivalence inference



evaluate

mondo

curate

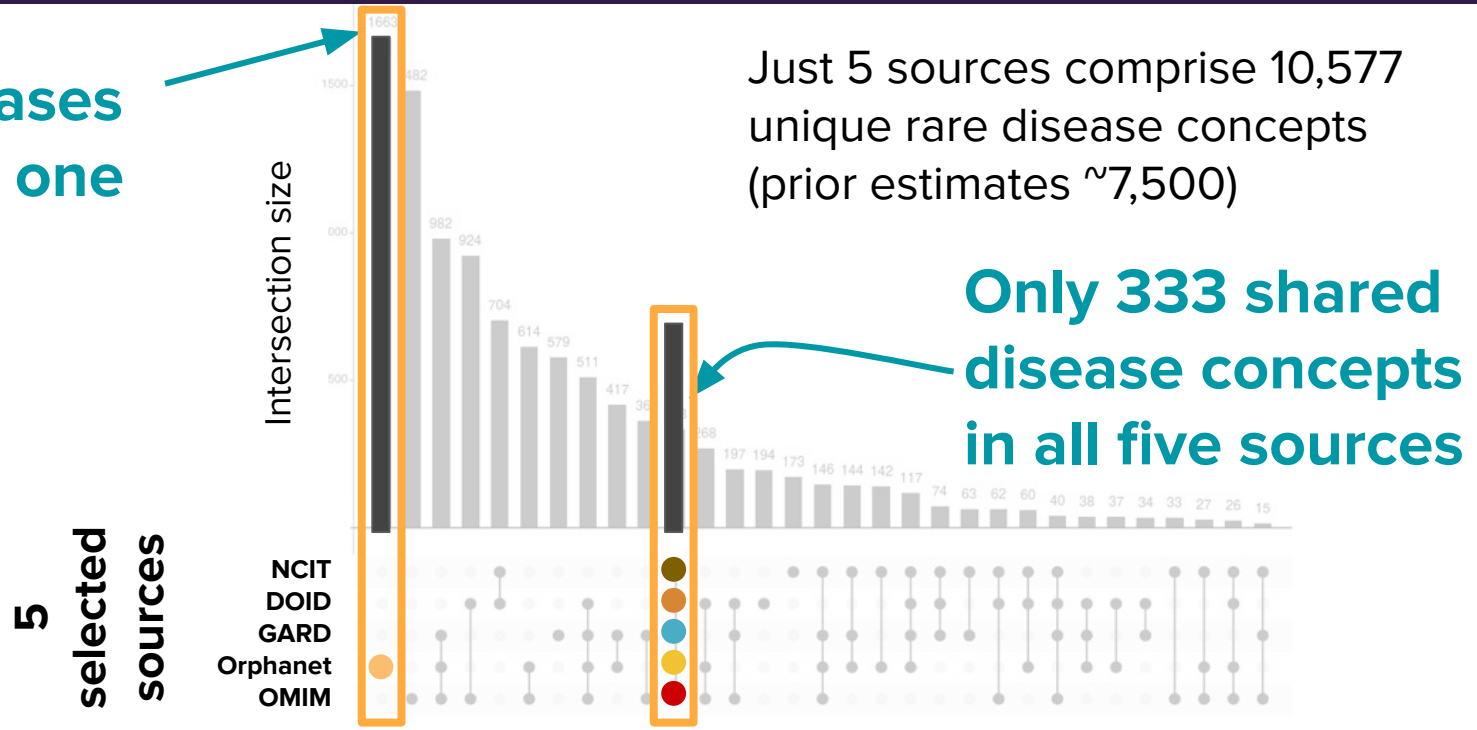
feedback

Curated Equivalence Relations

doi.org/10.1101/048843

If rare diseases are not counted, rare disease patients will not count

Many diseases are in only one source



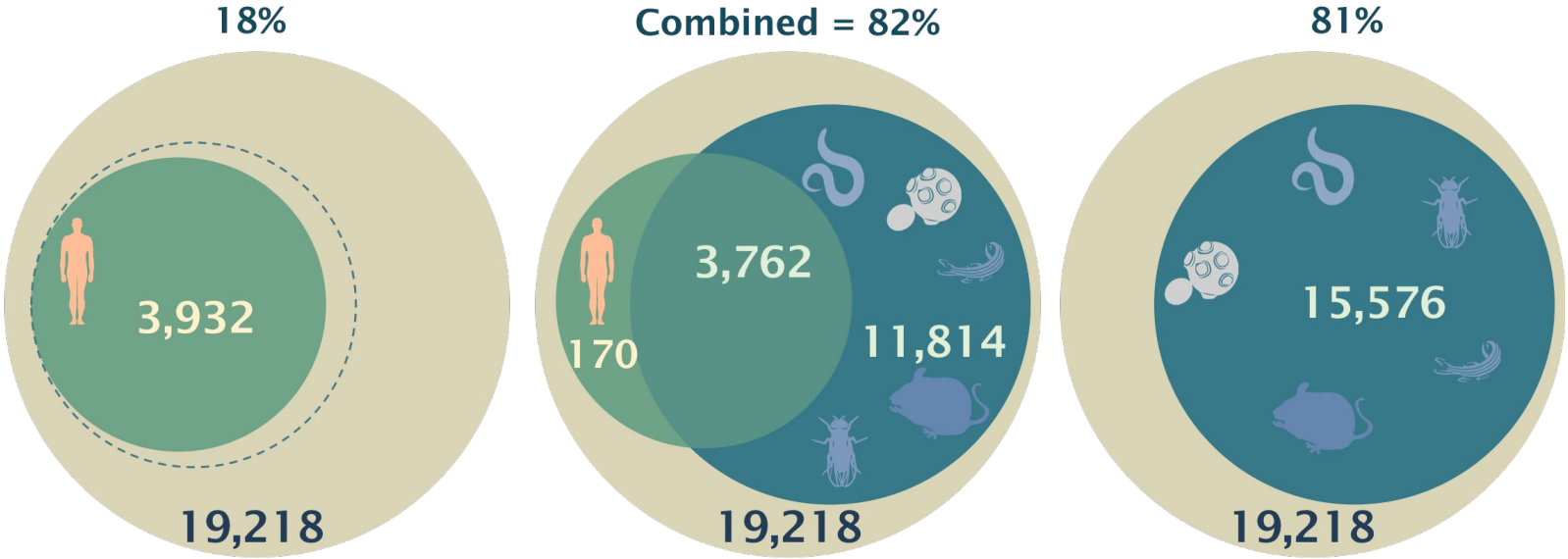
Just 5 sources comprise 10,577 unique rare disease concepts (prior estimates ~7,500)

Only 333 shared disease concepts in all five sources

Nature Reviews Drug Discovery (bit.ly/nature-rare-diseases)

Why model organisms matter to patients

More species = more coverage



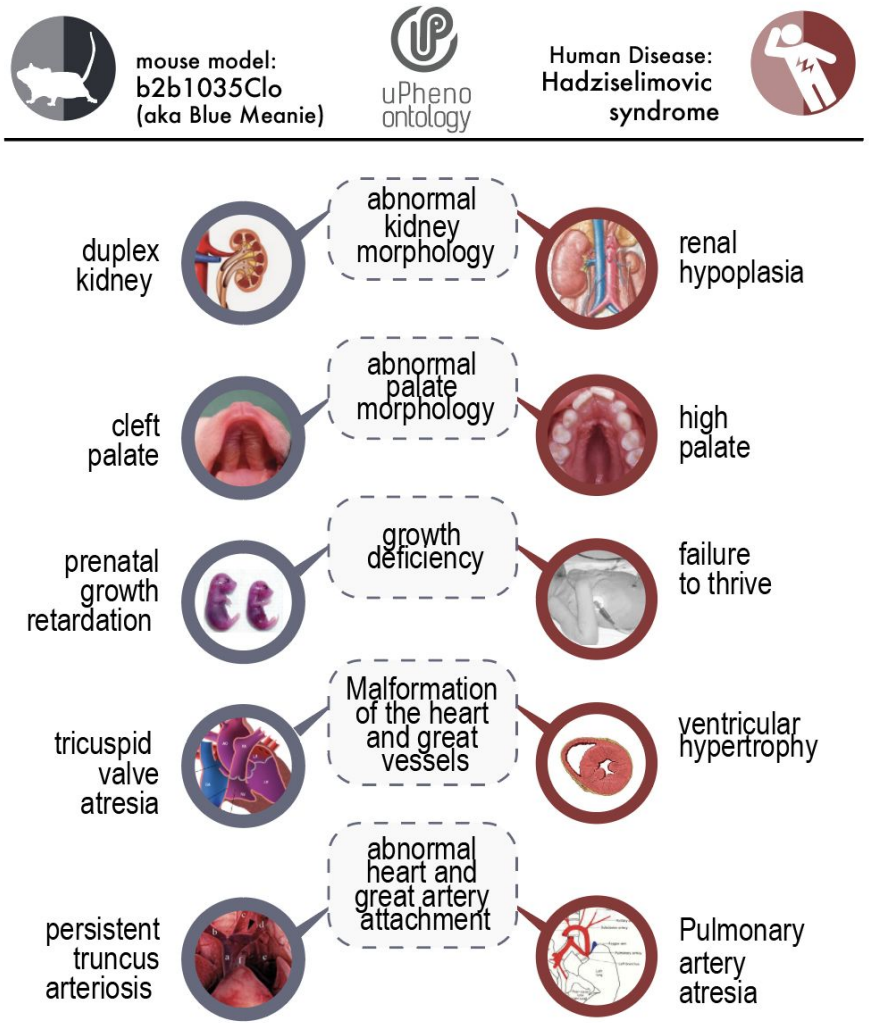
The inclusion of just five species boosts phenotypic coverage of genes by 64%

Latest data from monarchinitiative.org - 12-July-2019
See also Mungall et al. *Nucleic Acids Research* bit.ly/monarch-nar-2016

Fuzzy matching across species improves diagnostics

Genomics england

upheno ontology

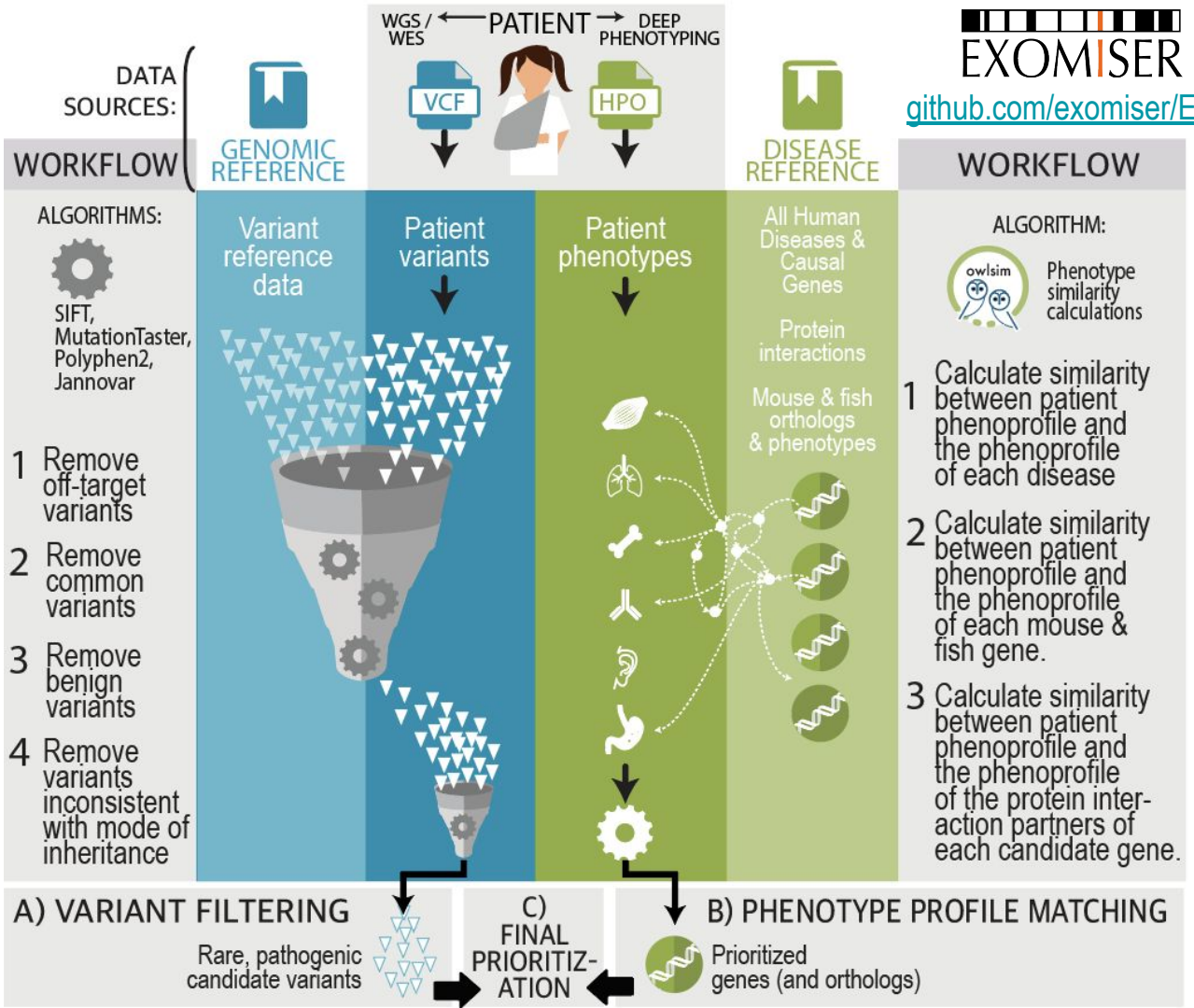


Combining genomic and phenomic data improves variant prioritization for diagnosis

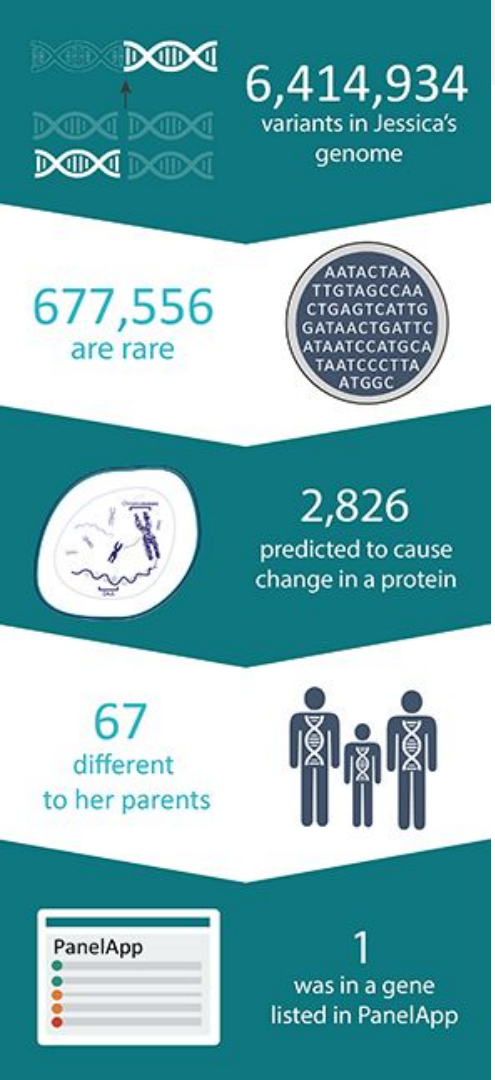
doi: 10.1038/gim.2015.137

EXOMISER

github.com/exomiser/Exomiser



Jessica

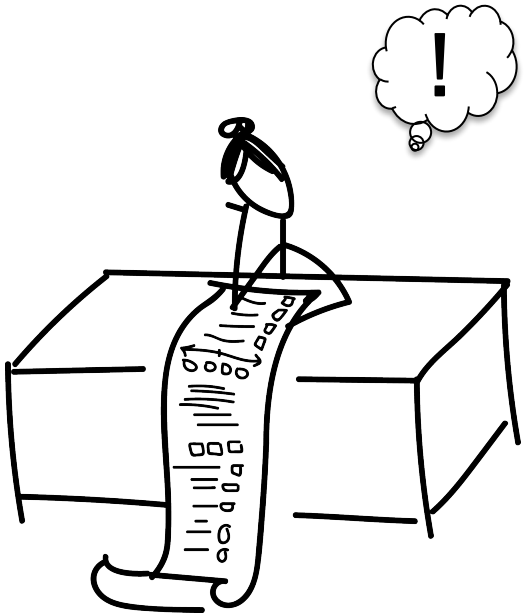


- Jessica (age 4) has a rare condition which causes epilepsy, affects her movement and developmental delay. Standard genetics tests negative.
- De novo deletion in *SLC2A1* identified as the cause of her Glut 1 deficiency syndrome
- Exomiser ranked this variant first
- Now being successfully treated with a ketogenic, low-carb diet
- Low risk for future pregnancies

Genomics england

Exomiser ranking 94% in top 3 candidates using human + model organism data

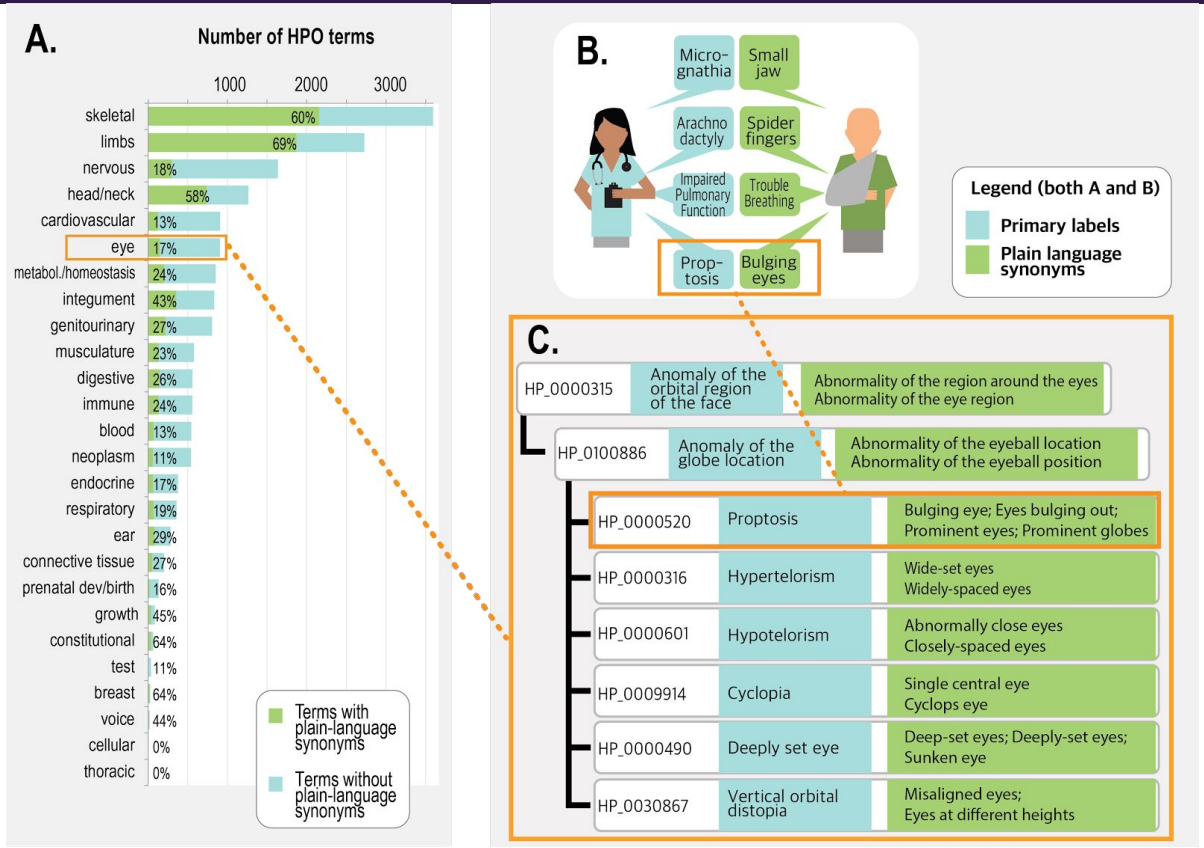
Phenotyping is not free … or easy; So how much is enough?



- The more phenotype data we have, the better able we are in answering that question
- We can help inform users whether their phenotyping is sufficient for analysis, and which new phenotypes to examine
- We need to improve case-level information sharing to better understand the heterogeneity of presentation and progression


Patients can and should be empowered to help

Plain language synonyms for patients to use



4887 of 13823 HPO terms have lay synonyms

National Covid-19 data and analytics platform

National Center for Advancing Translational Sciences

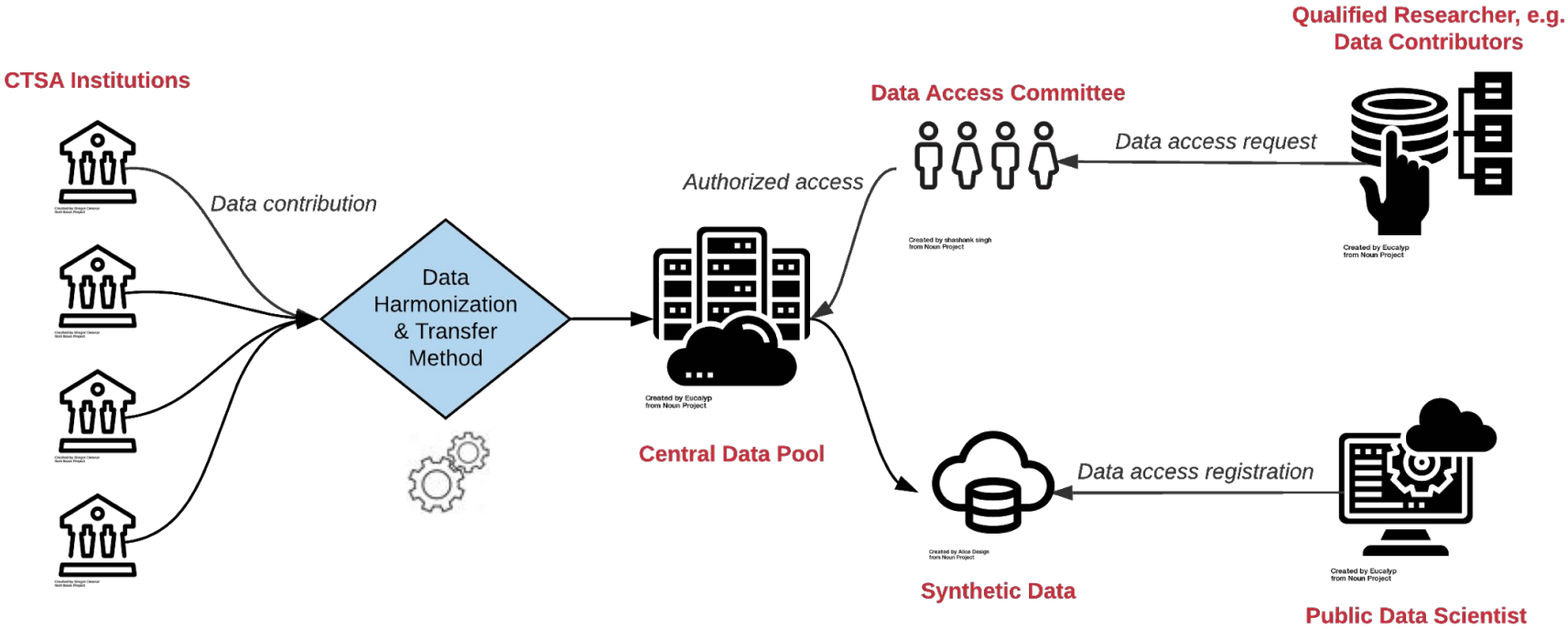
Shared, harmonized Covid data

Centralized model advantages

- Large dataset
- Consistency
- Improved machine learning applications & analytics over patient-level data
- Shared compute infrastructure and application deployment
- Purpose-driven curation/data modeling for covid-19

National Covid-19 data and analytics platform

National Center for Advancing Translational Sciences

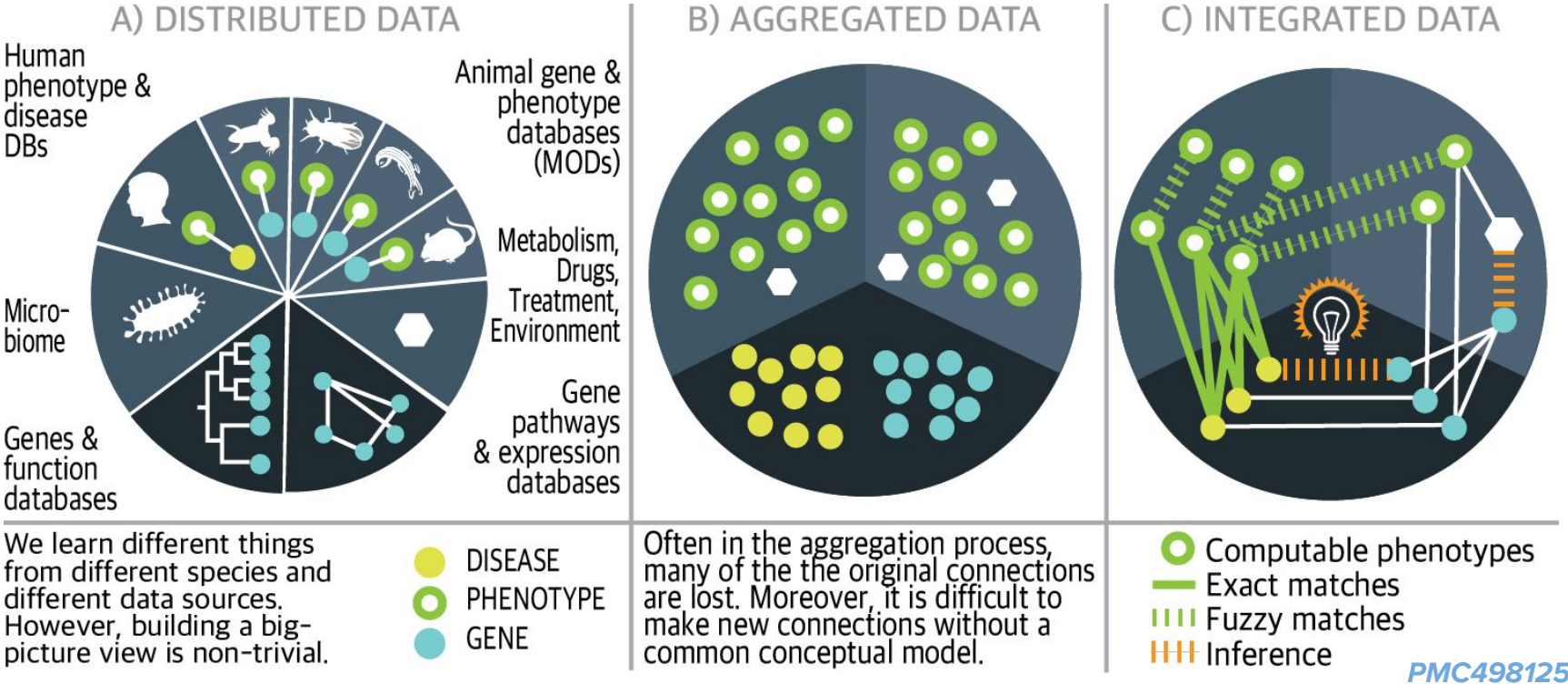


Info will be posted at covid.cd2h.org

How we can all help improve disease diagnosis and care

- We need deeper phenotyping of a patient's condition - beyond billing
- Ontologies can support discovery by aiding data integration and analytics across domains
- Use non-human data to support diagnostics, drug discovery, and mechanism discovery
- Use semantics to combine genomics with phenomics (and other data)!
- Richer and more widespread sharing of patient-level data is needed to fully understand disease heterogeneity
- **There is an urgent need to improve data collection and flow between all stakeholders** across full lifecycle of disease care -- from patients and families themselves through clinicians and hospitalists => this is true in all fields !!

Think about downstream data reuse



PMC4981258
NIH: R24 OD011883

Thank you

Chris Mungall
Peter Robinson
Ken Gersing
Chris Chute
Julie McMurry
Monica Munoz-Torres
Matt Brush
Jules Jacobsen
Nico Matentzoglu
Nicole Vasilevsky
Kent Shefchek
Tom Conlin
Nomi Harris
Marcin Joachimiak
Seth Carbon
Justin Reese
Deepak Unni
Sebastian Koeller
Tudor Groza

Monarch - Linking diseases to model organism resources 5 R24 OD011883
Biomedical Data Translator: OT3 TR002019-01S2
Center for Data To Health (CD2H) U24TR002306