

Ontologies to Unify Genomics and Phenomics Across Species

Anne E Thessen

annethessen@gmail.com

@diatomsRcool 



University of Colorado
Anschutz Medical Campus

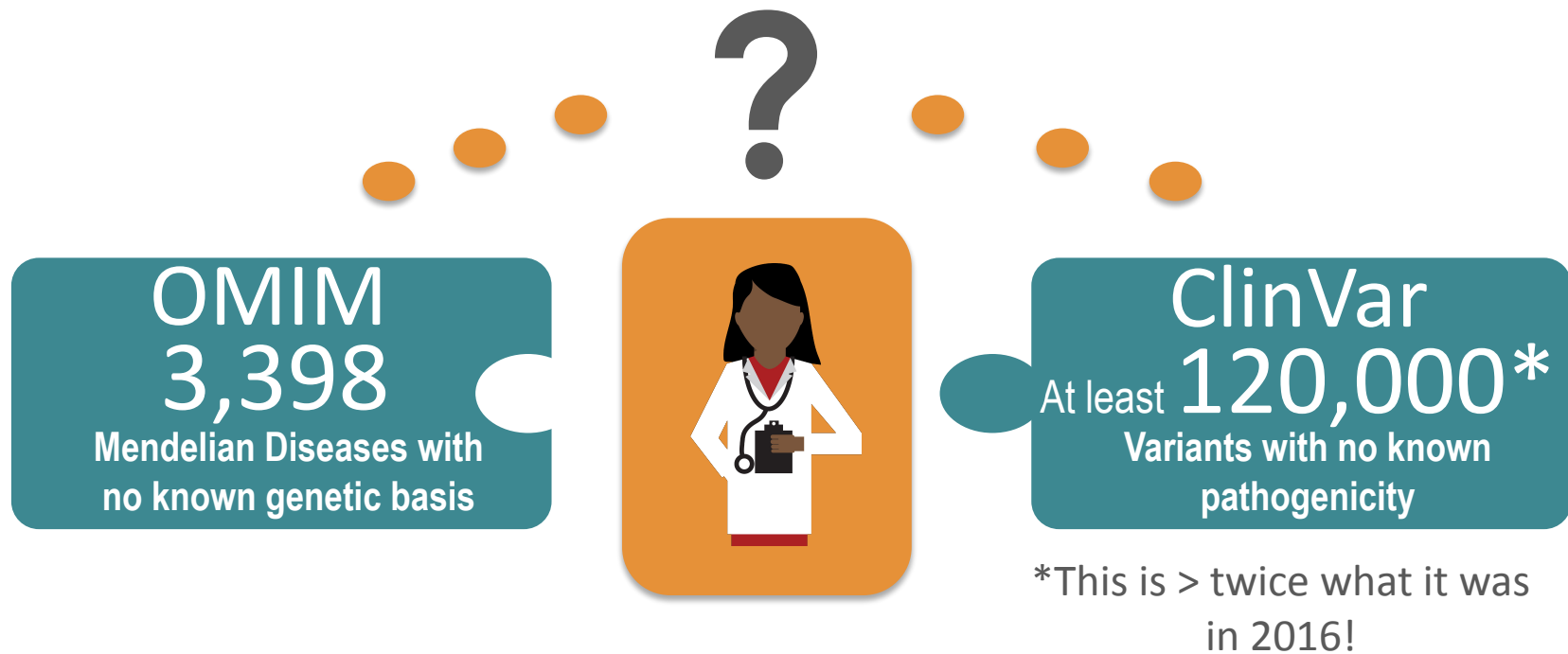
On behalf of:

*Melissa Haendel
Chris Mungall
Chris Chute
Damian Smedley
David Osumi-Sutherland
Peter Robinson
Tudor Groza
Sebastian Köhler
Ada Hamosh
Monica Muñoz-Torres
Jules Jacobsen
Kent Shefchek
Nomi Harris
Tim Putman
Anne Thessen
Harshad Hedge
Justin Reese
Kevin Schaper
Lauren Chan
Matt Brush
Nicolas Matentzoglou
Nicole Vasilevsky
Sabrina Toro
Seth Carbon
Sierra Moxon
Tudor Groza*

Funded by:
NIH Office of Director:
1R24OD011883;
NIH-UDP:
HHSN268201300036C,
HHSN268201400093P;
NCINCI/Leidos #15X143,
BD2K
U54HG007990-S2
(Haussler) & BD2K
PA-15-144-U01
(Kesselman)

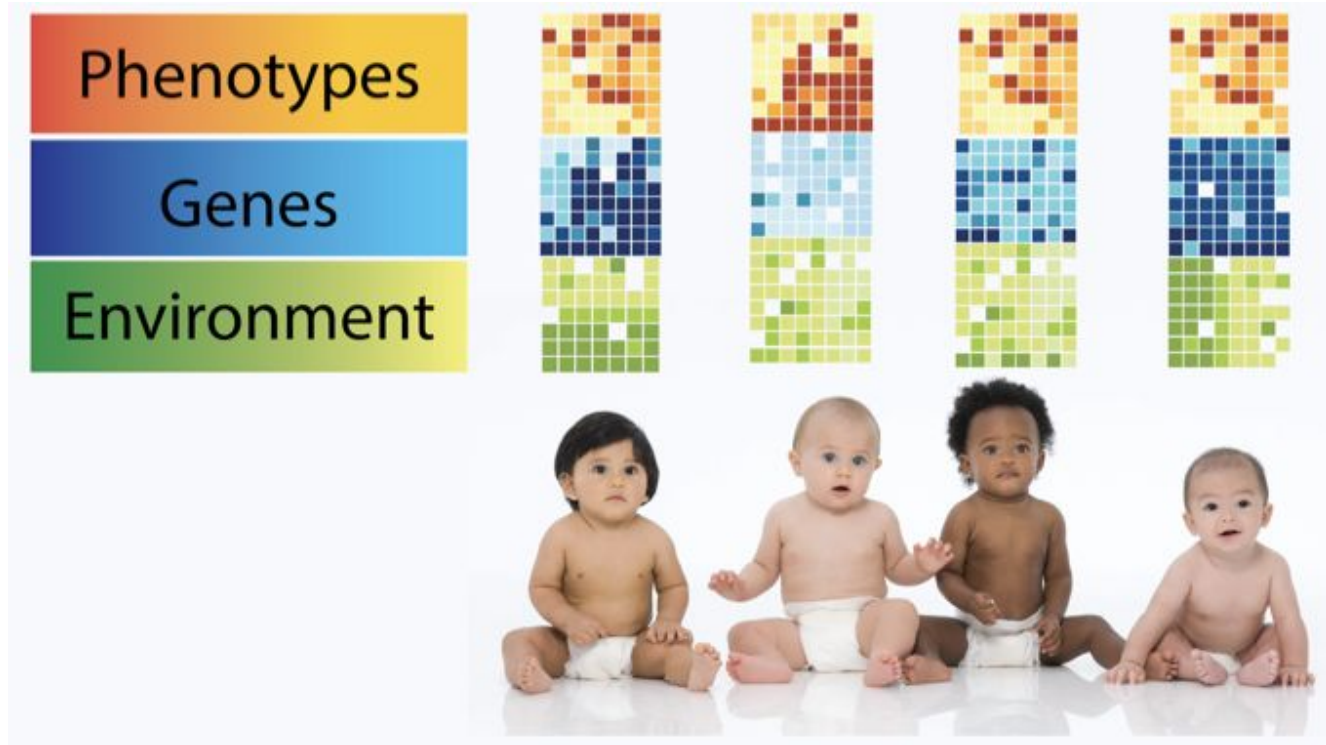
These slides are at bit.ly/NAScompanion

The genome is sequenced, but....



...we still don't know very much about what it does

The genome is sequenced, but....



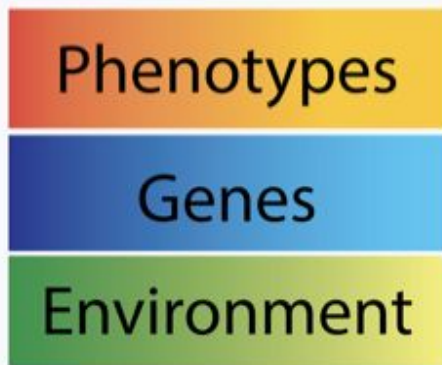
...we don't understand environmental effects

1a) Individual person
as biological subject

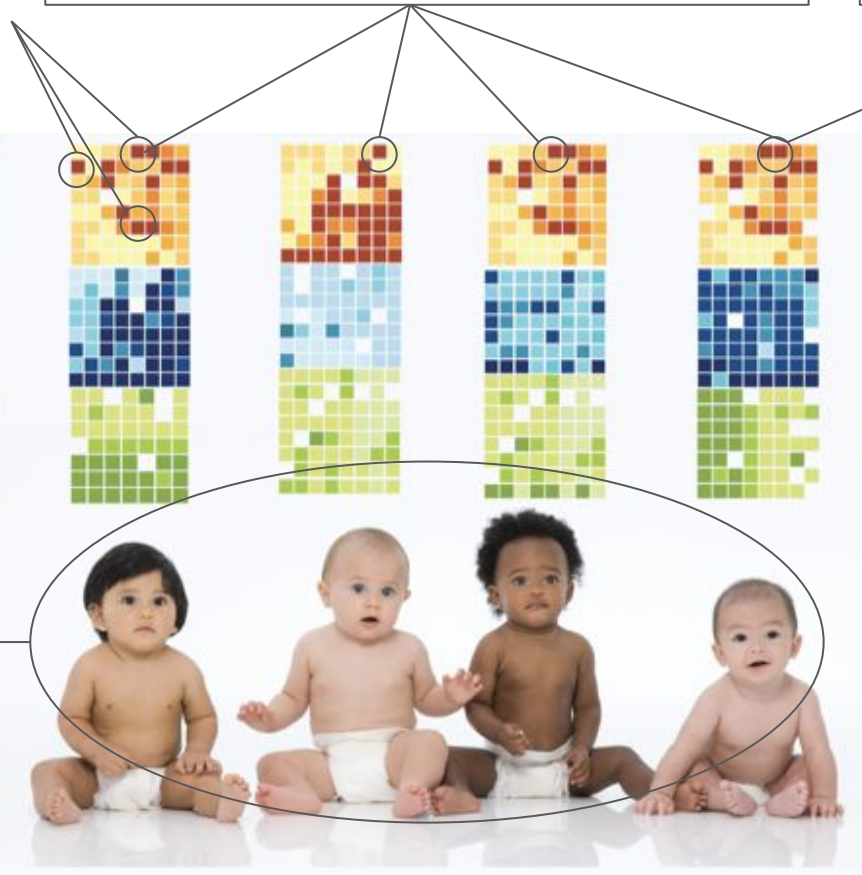
1b) Individual person
as patient

2) A disease as reference knowledge
(derived from many patient's presentation)

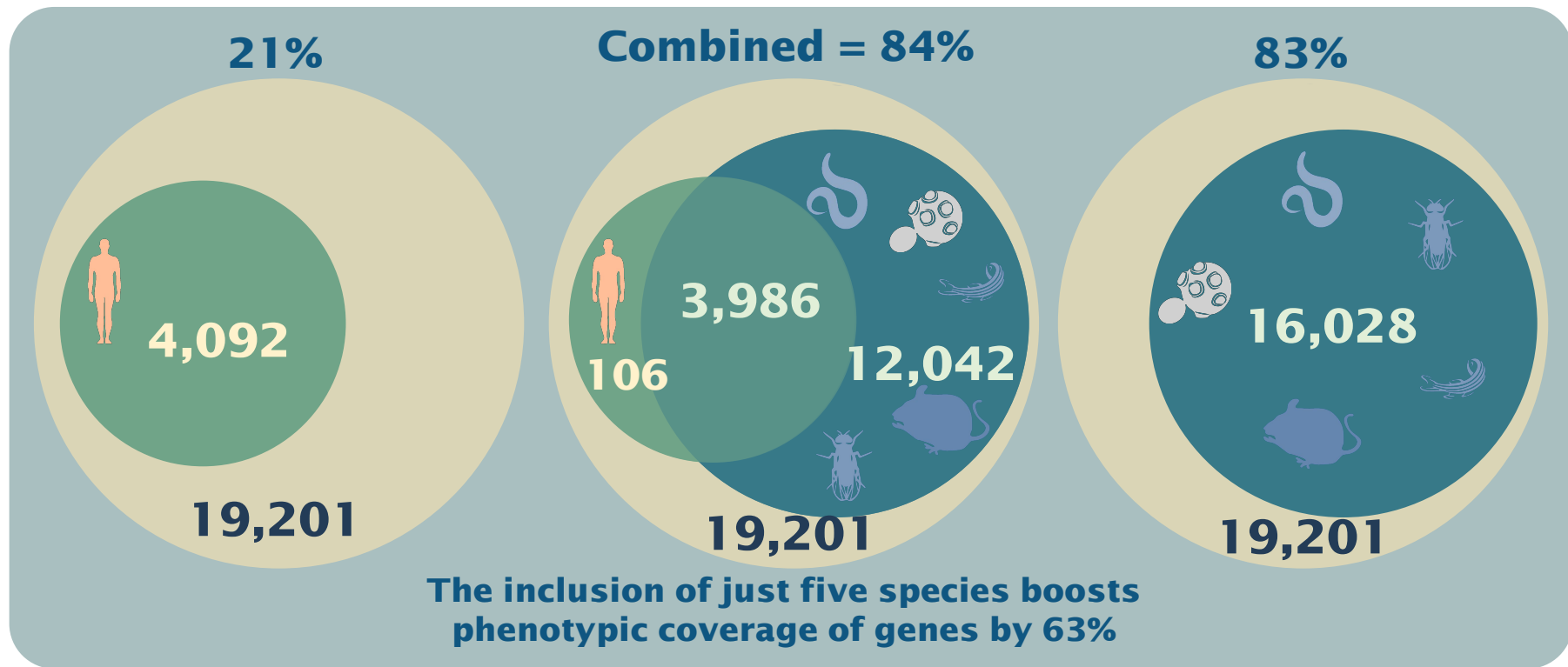
3) Cross species
comparison for diagnosis



4) Many individuals in
the EHR



More species, more coverage



Other species aren't just relevant, each has unique phenotypes



- The dog's retina has area centralis (analogous to the human macula) & fovea-like region, similar to humans; useful to study naturally occurring cone diseases



- Aged cats are natural models of Alzheimer's Disease: they form Abeta oligomers, neurofibrillary tangles, and have neuronal loss



- Naked Mole Rats don't get cancer



- Armadillos are a natural host of *M. leprae*, the mycobacterium that causes leprosy (only one besides humans)



- Tree shrews' glioblastomas are morphologically & genetically similar to humans (& more similar than mouse models)



- Great pond snails are models of inflammation-mediated memory dysfunction, and show evidence of spontaneous neural tissue regeneration after injury



- Silkworms are a model for uric acid metabolism. Decreases in plasma uric acid are correlated with clinical progression of Parkinson's Disease

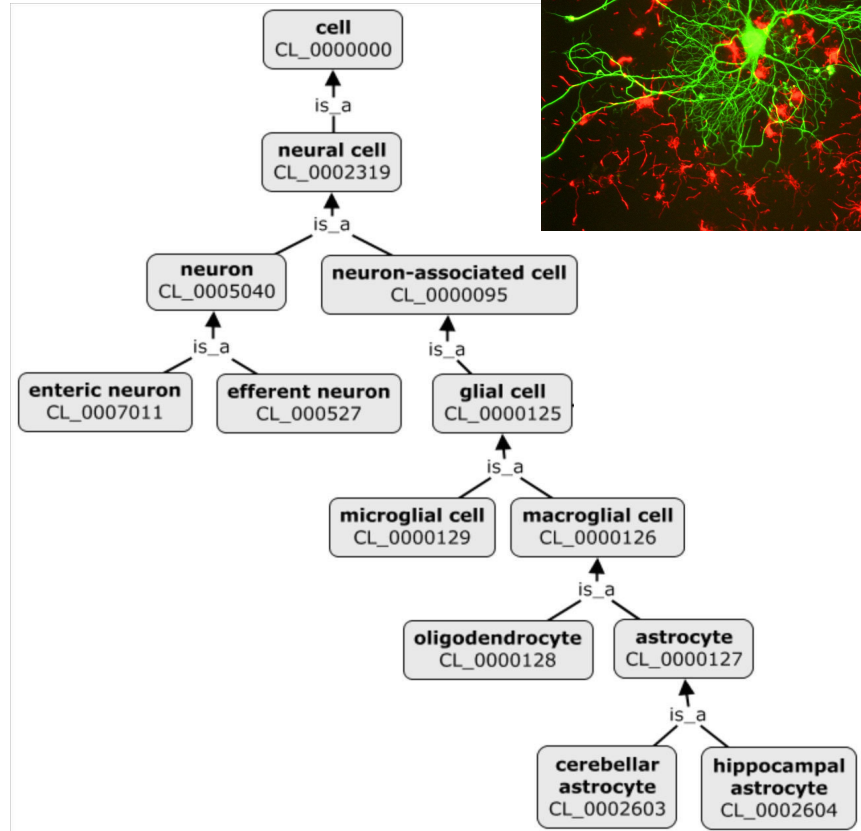
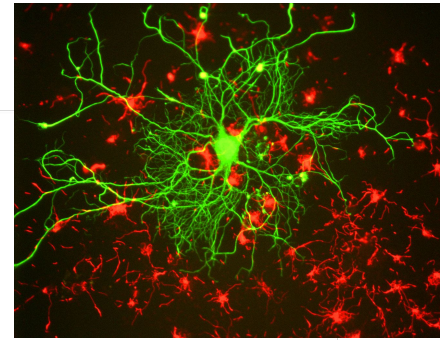
What is an ontology?

DEFINITION:

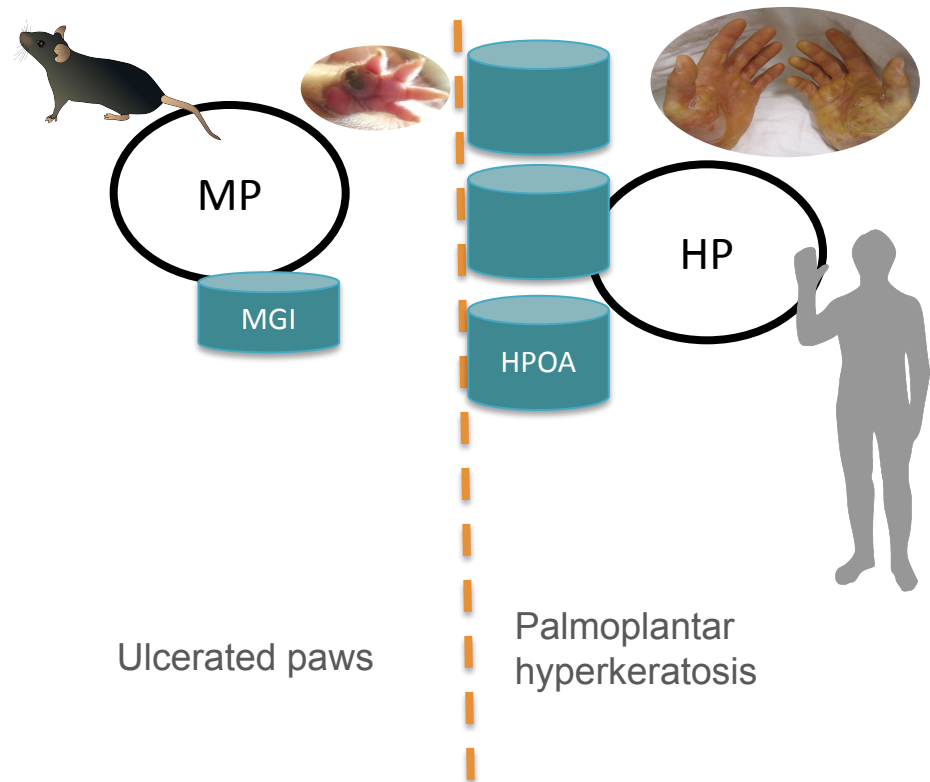
A formal, computational representation of knowledge in a particular domain.

KEY FEATURES:

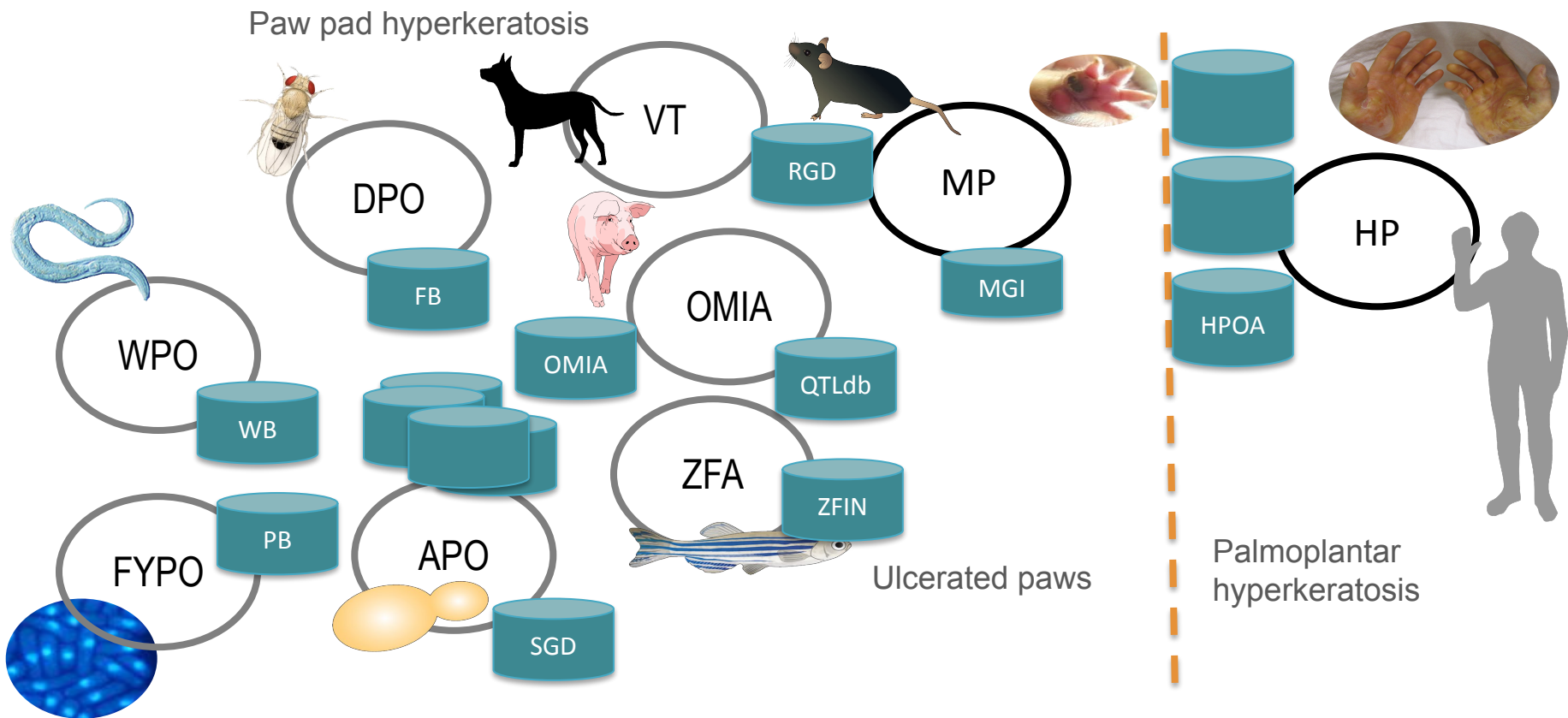
- Terms are defined
- Semantics - relationships between terms are defined, allowing logical inference and sophisticated data queries
- Terms are arranged in a hierarchy
- Expressed in a knowledge representation language such as RDFS, OBO, or OWL



Challenge: Each data source uses its own vocabulary



Challenge: Each data source uses its own vocabulary



Can we help machines understand phenotypes?

Human phenotype

“Palmoplantar
hyperkeratosis”



I have absolutely
no idea what
that means



Decomposition of complex concepts allows interoperability



**“Palmoplantar
hyperkeratosis”**

=



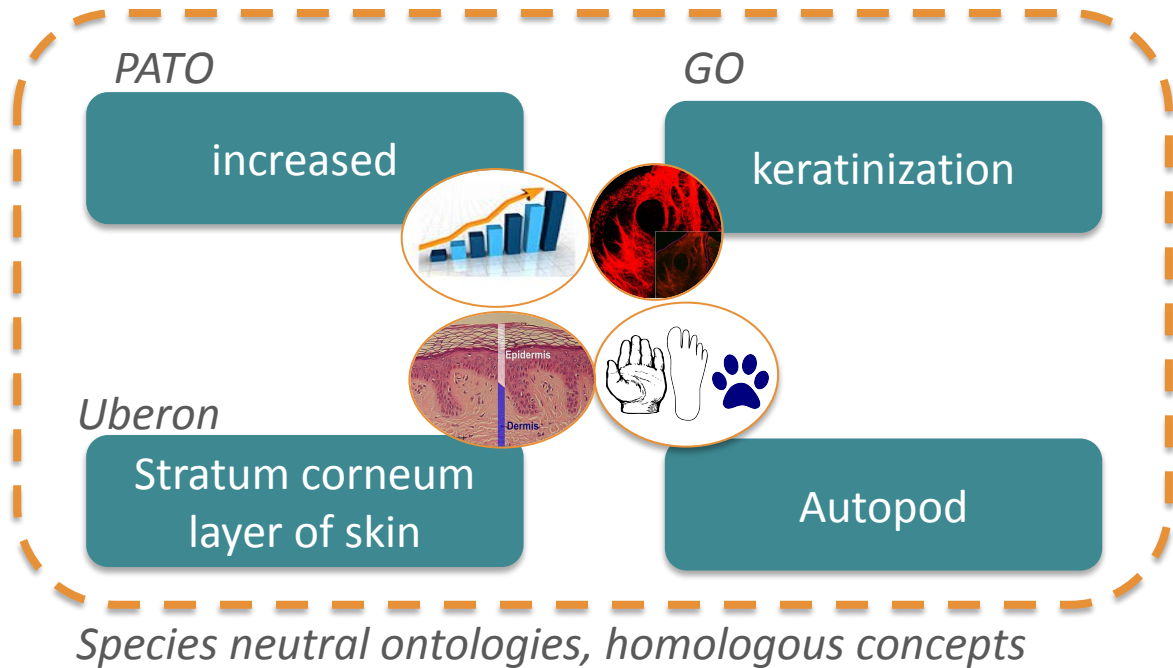
“Hairy dog feet”

=

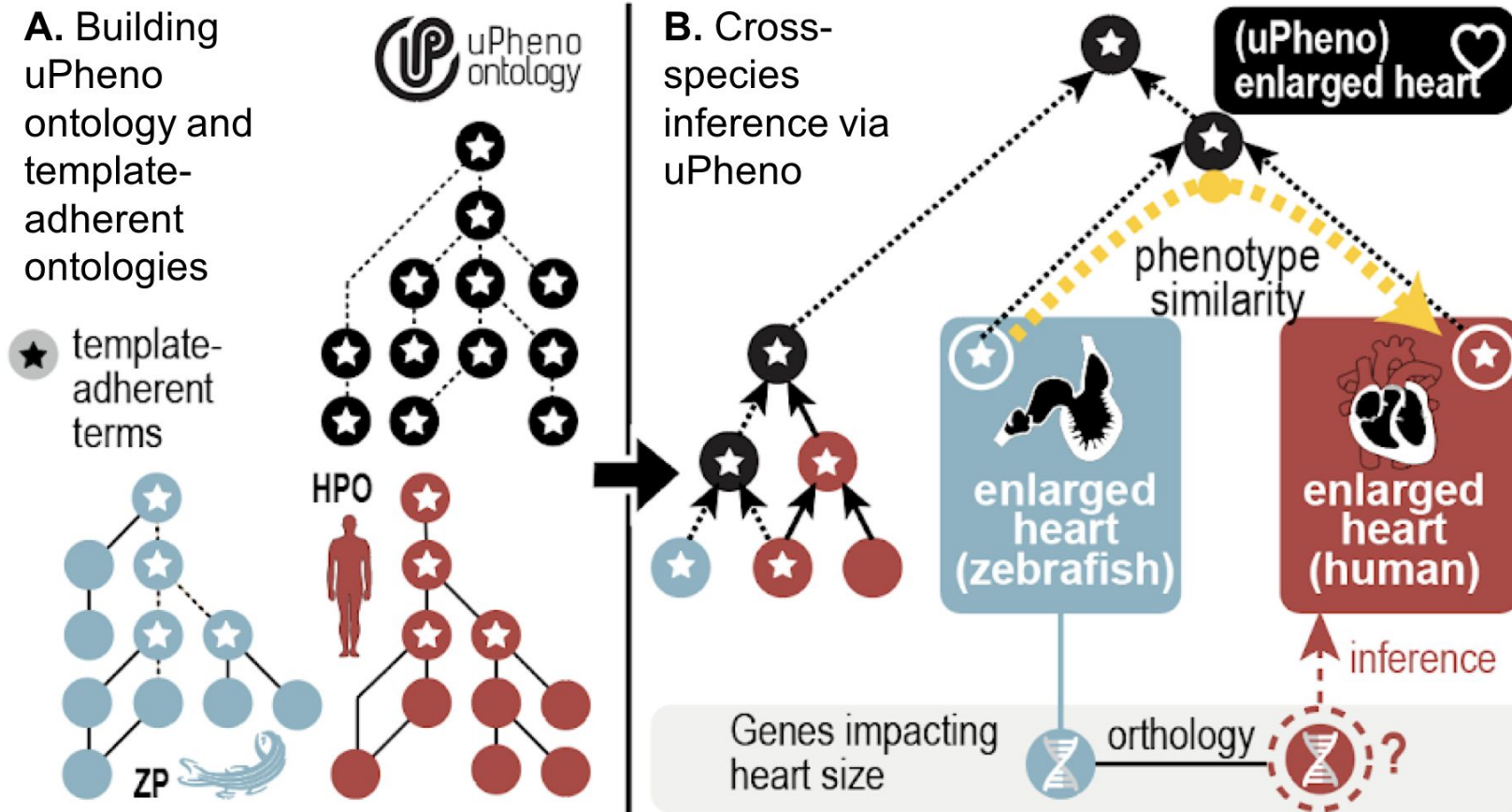


**“Ulcerated
paws”**

=

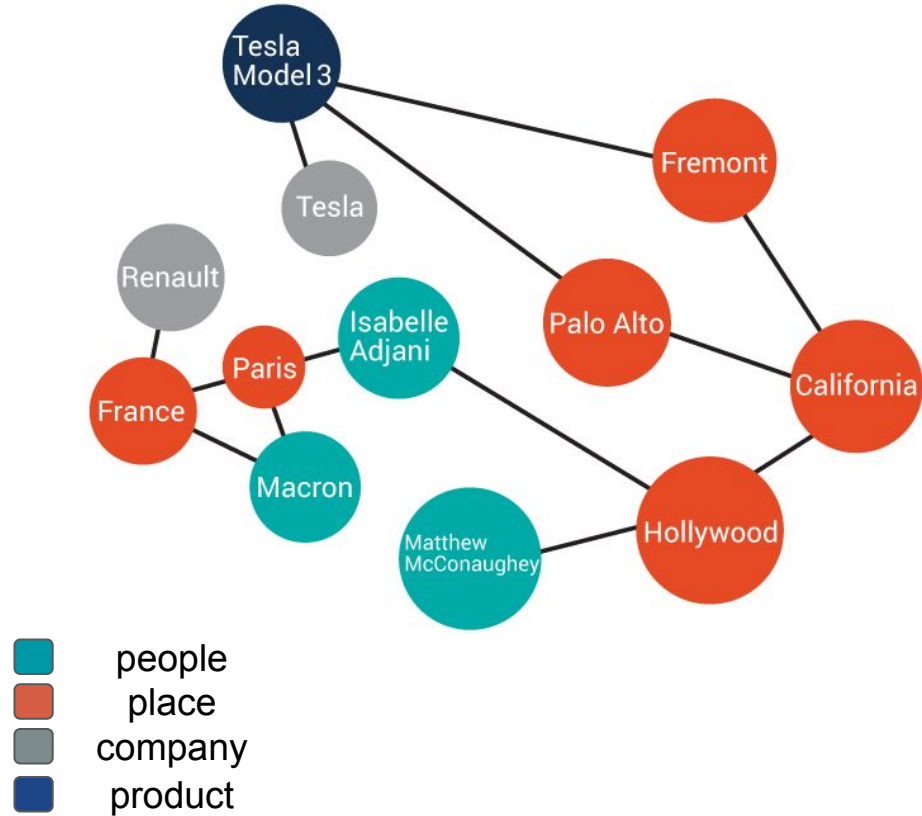


Why do we care about phenotype integration?



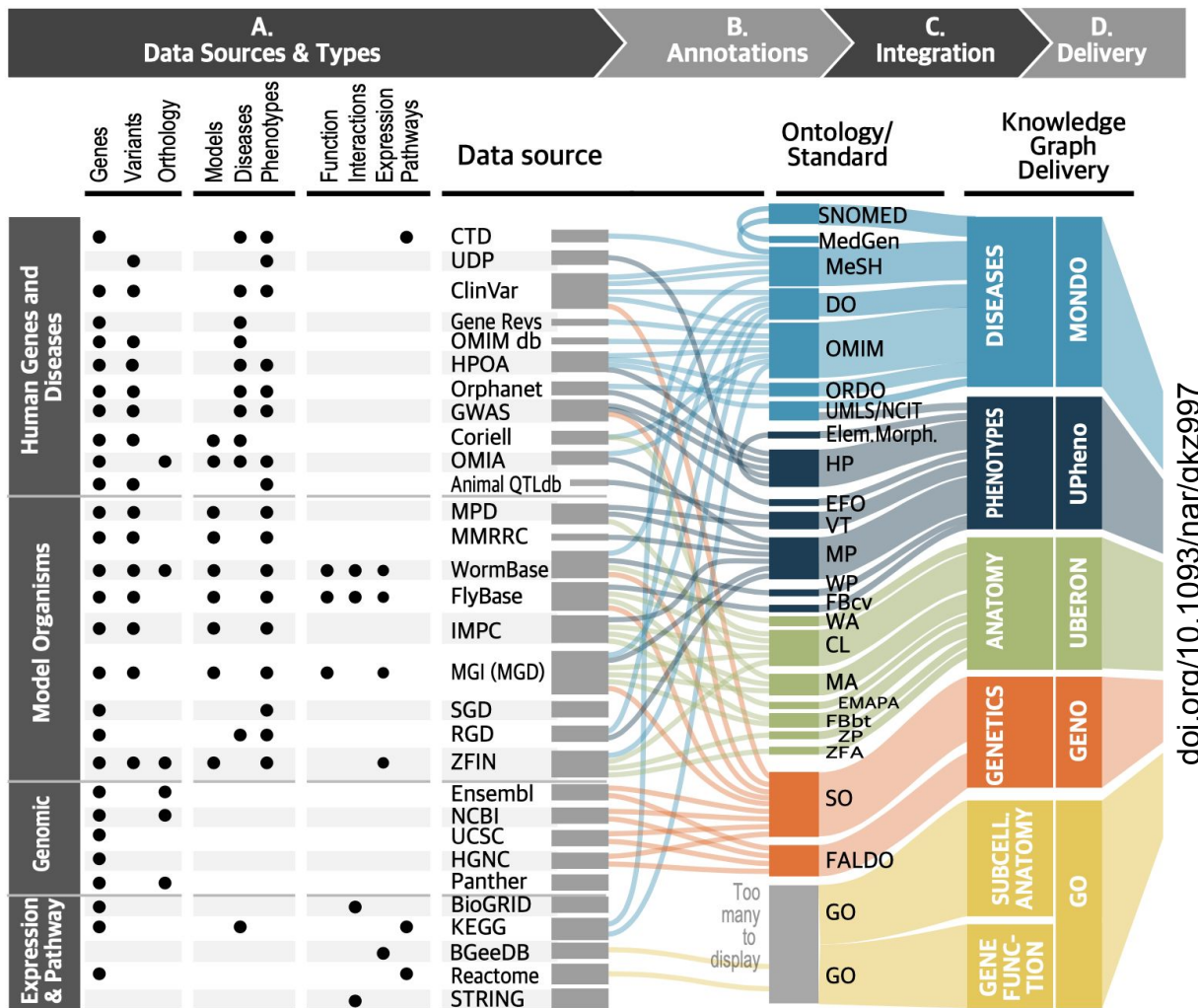
Ontology or Knowledge Graph?

- Two different ways to represent the same data
- Require slightly different modeling techniques
- Ontologies are community-driven schema that moves very slowly
- Knowledge graphs grow more quickly and can be more responsive
- More tools and services available for knowledge graphs



How did we do this?

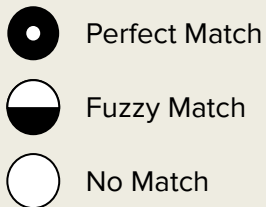
- monarchinitiative.org
- Integrator of cross species genotype-phenotype data
- OWL and DOS-DP
- Uses OBO Foundry ontologies
- obofoundry.org



Use Case: Rare disease diagnostics

- Fuzzy matching
- Collections of phenotypes and genes

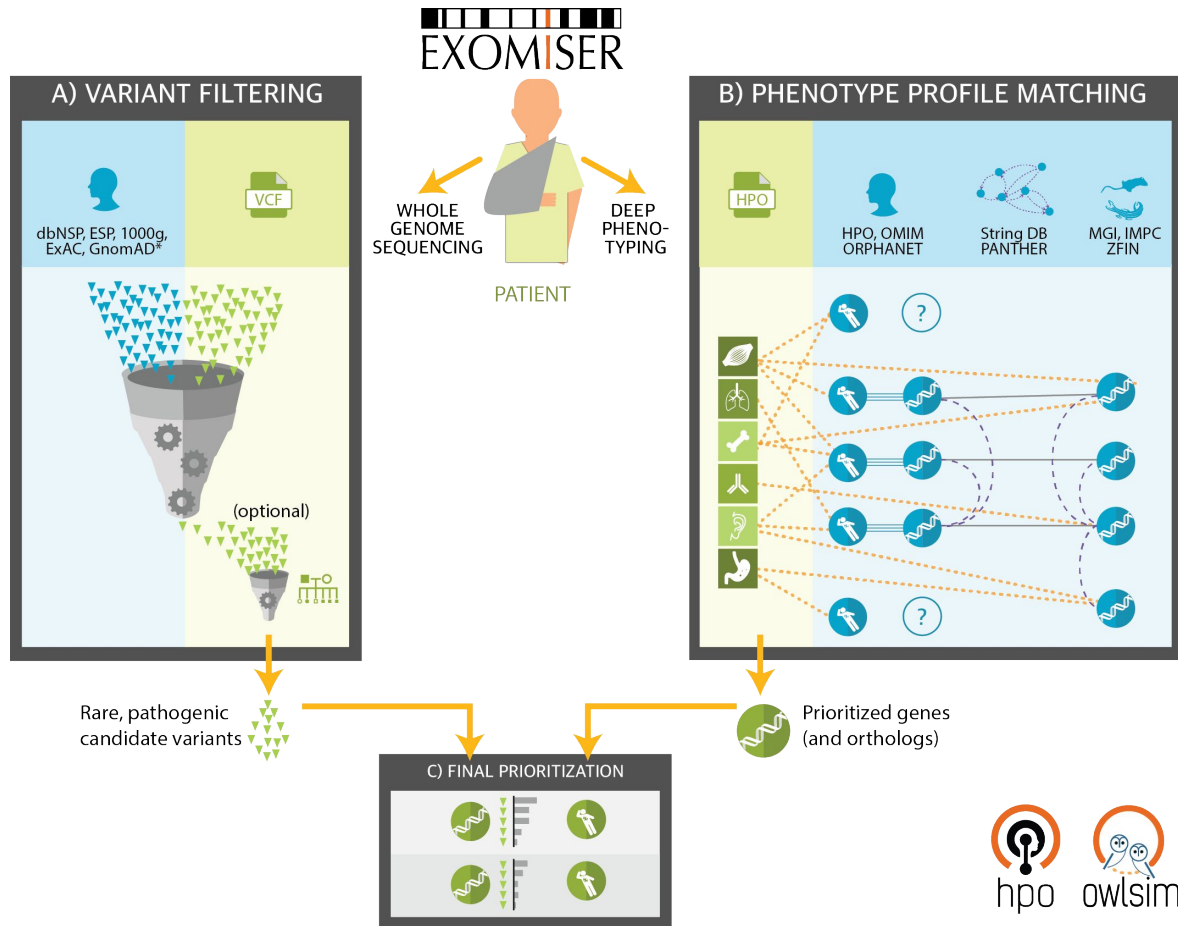
Legend



	Patient P1 profile (3 year old girl)	Wiedemann Steiner syndrome profile	Patient P8 profile (14 year old boy)
digits	None	Short toe HP:0001831	Long toe HP:0010511
	Cone-shaped epiphysis of the phalanges of the hand HP:0010230	Short middle phalanx of finger HP:0005819	None
development	Delayed speech and language development HP:0000750	Delayed speech and language development HP:0000750	Delayed speech and language development HP:0000750
	Global developmental delay HP:0001263	Intellectual disability HP:0001249	Global developmental delay HP:0001263
	Microcephaly HP:0000252	Microcephaly* HP:0000252	Macrocephaly HP:0000256
skeletal	Proportionate short stature HP:0003508	Short stature HP:0004322	None
	Thick upper lip vermillion HP:0000215	Thin upper lip vermillion HP:0000219	None
facial	Hypertelorism HP:0000316	Hypertelorism HP:0000316	None
	None	Blepharophimosis HP:0000581	Blepharophimosis HP:0000581
	None	Epicanthus HP:0000286	Epicanthus inversus HP:0000537

Use Case: Variant prioritization for clinical diagnostics

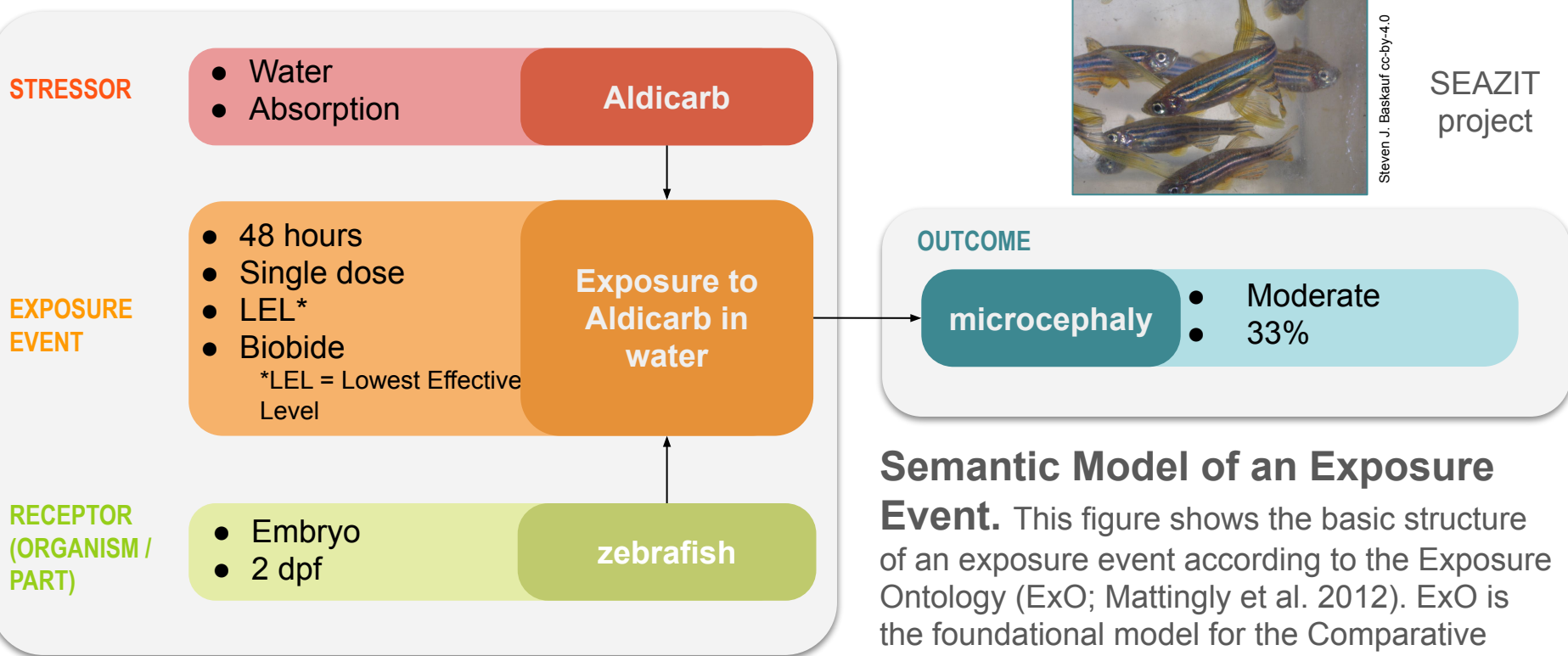
- Phenotype profile includes abnormally low bone mineral density
- Find all human orthologs of genes for which the presence of variants (alleles) is correlated with a phenotypic effect on bone mineral density.



Exposures: The missing piece



Exposure Ontology make environmental data computable



Semantic Model of an Exposure Event. This figure shows the basic structure of an exposure event according to the Exposure Ontology (ExO; Mattingly et al. 2012). ExO is the foundational model for the Comparative Toxicogenomics Database (CTD; Mattingly et al. 2006).

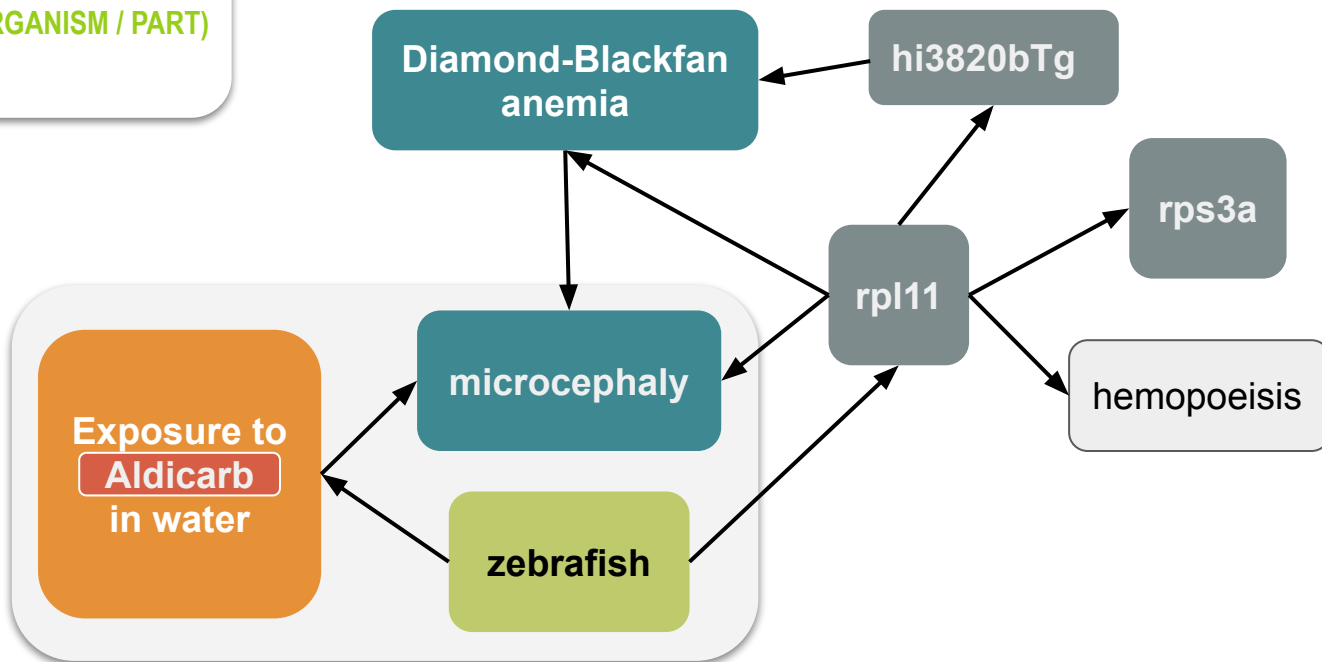
Integration with the Monarch Knowledge Graph

STRESSOR

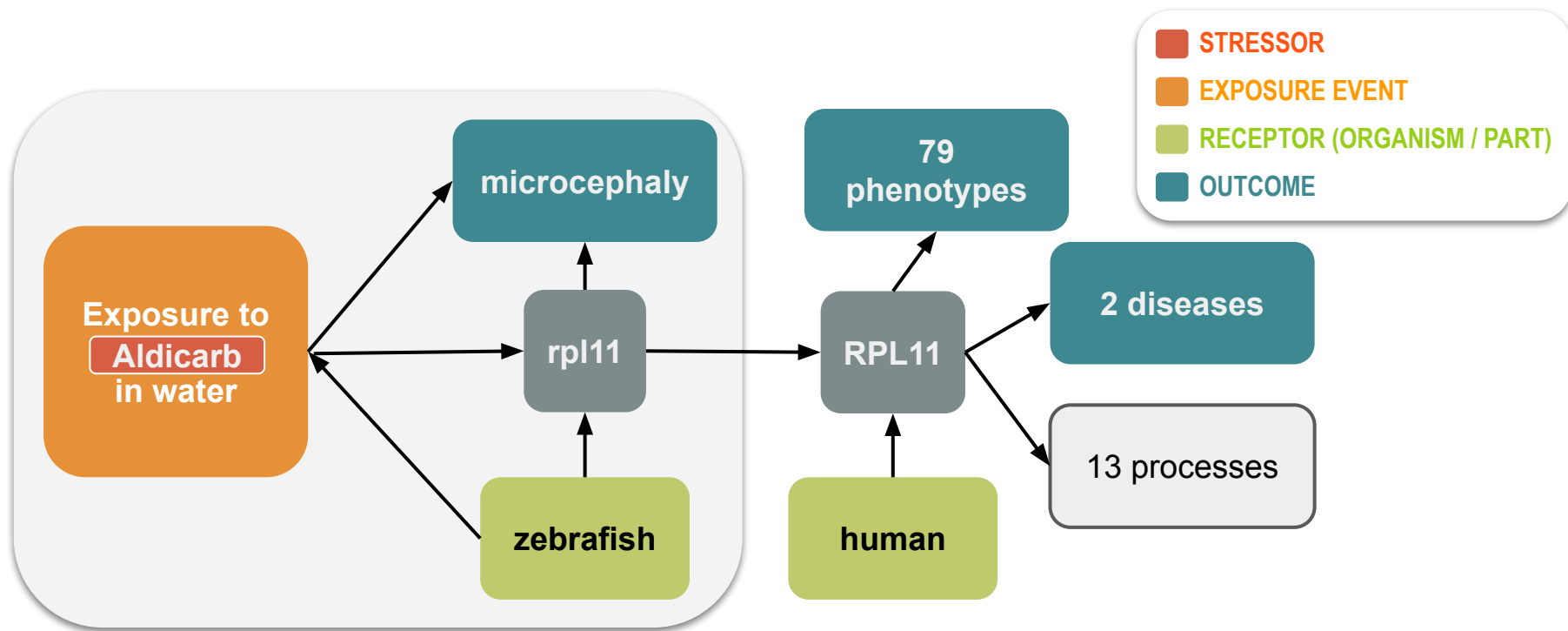
EXPOSURE EVENT

RECEPTOR (ORGANISM / PART)

OUTCOME

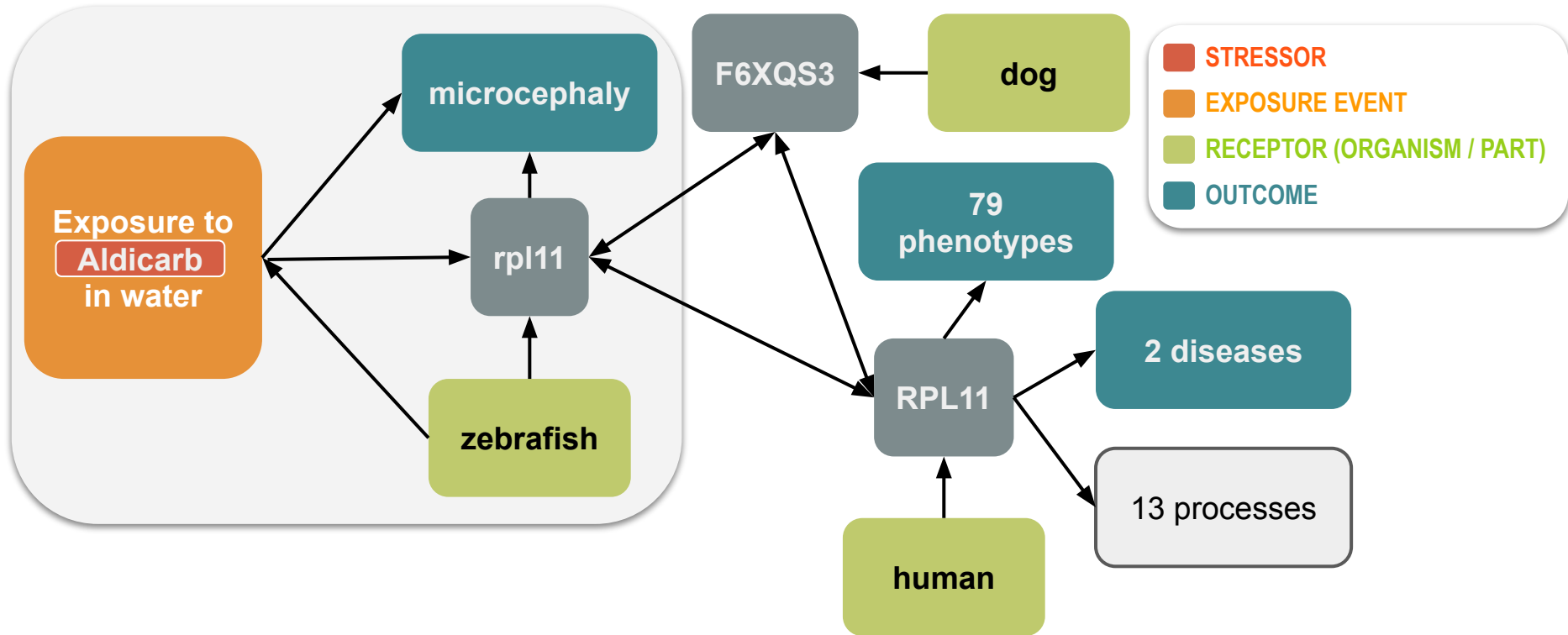


Integration with the Monarch Knowledge Graph



Integration with the Monarch Knowledge Graph

Not many connections for dog - need
standardized data sets

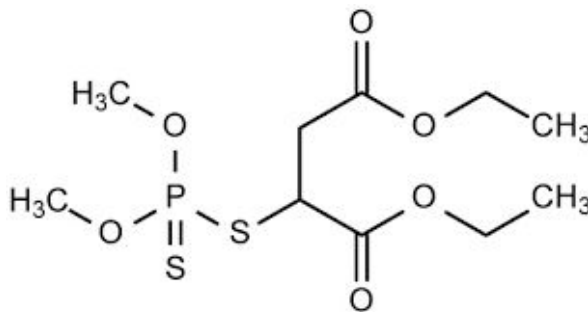


We have a machine-readable language for describing some exposures



CHEBI:6651

= CCOC(=O)CC(SP(=S)(OC)OC)C(=O)OCC

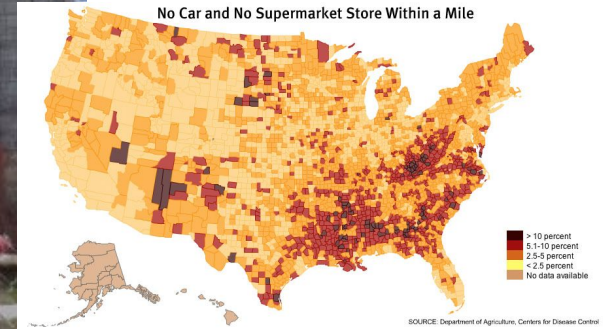


CheBI is a chemical ontology

But others are harder to define



Image: Zol87 CC by/nc



Summary

- Ontologies and knowledge graphs are an effective tool for integrating biological data across species and terminologies
- Integrated data sets can be used for inference and semantic similarity analysis
- Data standards are required (Dog Phenotype Ontology)
- Future work will include development of environmental exposure data models
- Companion animals share many human exposures can add power to studies

Want to help?

- annethessen@gmail.com
- Join the team (we and our partners are hiring trainees and staff)
 - 21772 - Instructor Senior Instructor - Scientific / Medical Writer
 - 23591 - Postdoctoral Fellow
 - 21903 - Consortium Program Assistant
 - Professional Research Assistant
 - Postdoctoral Fellow (Veterinary)
- Give us feedback on the Monarch tools and data. (T shirts and swag for user experience interviews) Inquire at info@tislab.org
- Promote responsible licensing of data reusabledata.org

The Monarch Initiative

www.monarchinitiative.org

annethessen@gmail.com



PDs: Melissa Haendel, Chris Mungall, Peter Robinson

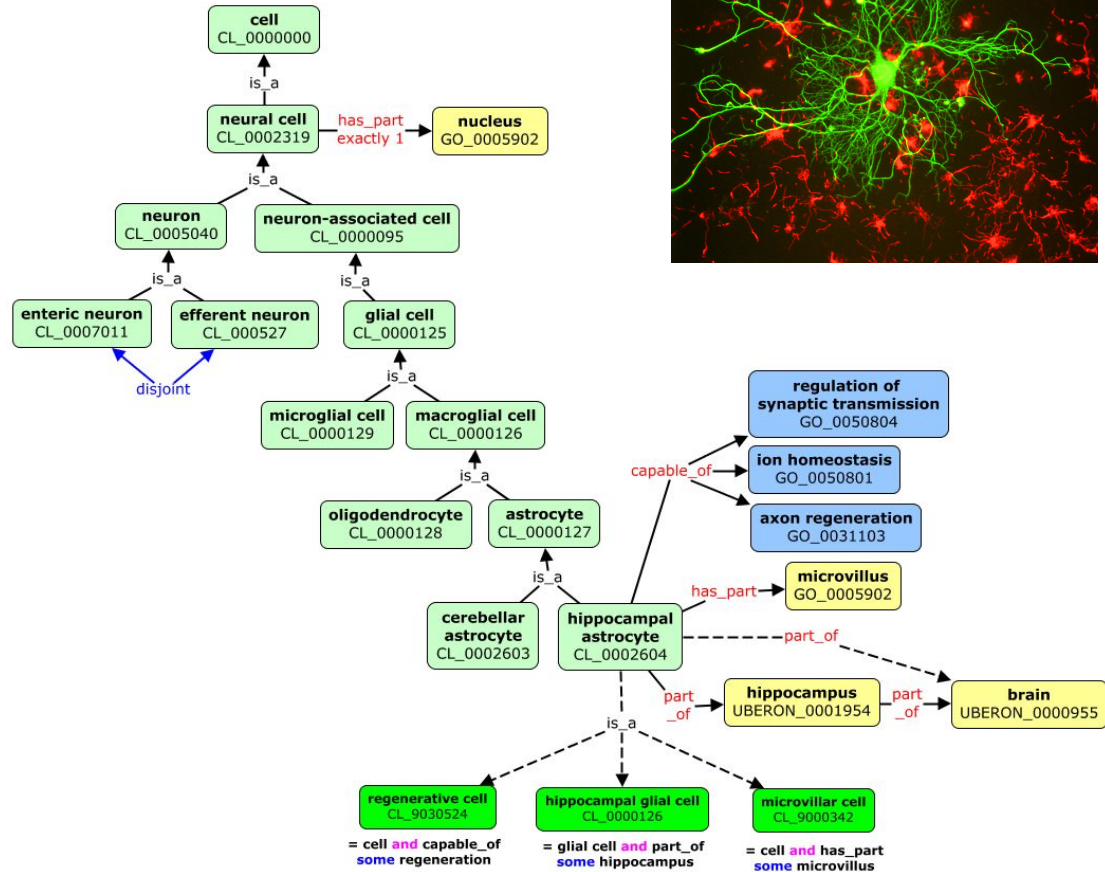
What is an ontology?

DEFINITION:

A formal, computational representation of knowledge in a particular domain.

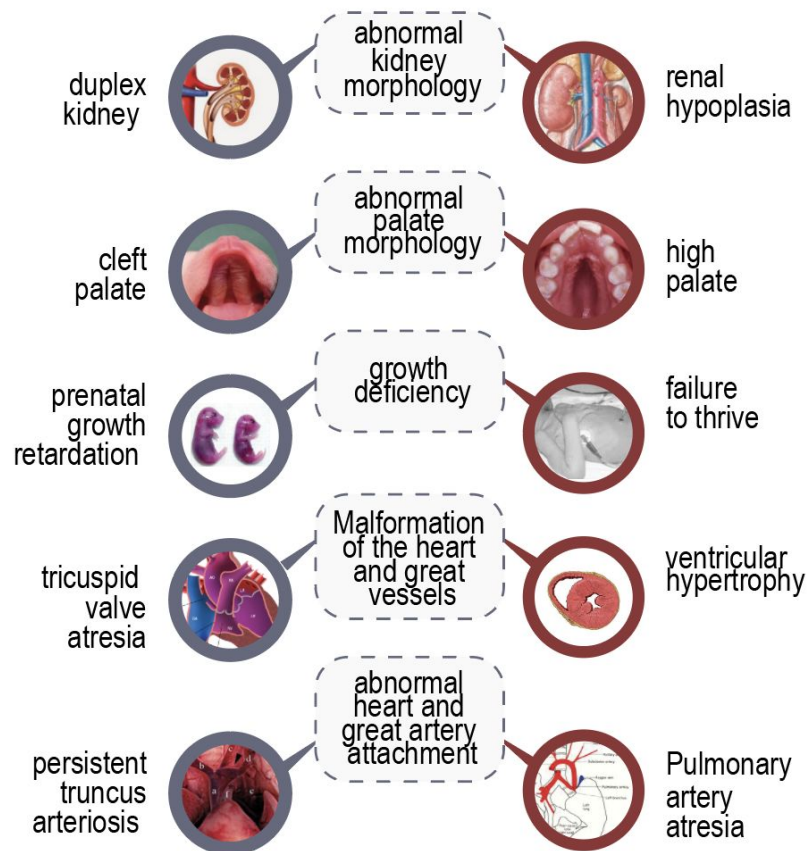
KEY FEATURES:

- Terms are defined
- Semantics - relationships between terms are defined, allowing logical inference and sophisticated data queries
- Terms are arranged in a hierarchy
- Expressed in a knowledge representation language such as RDFS, OBO, or OWL



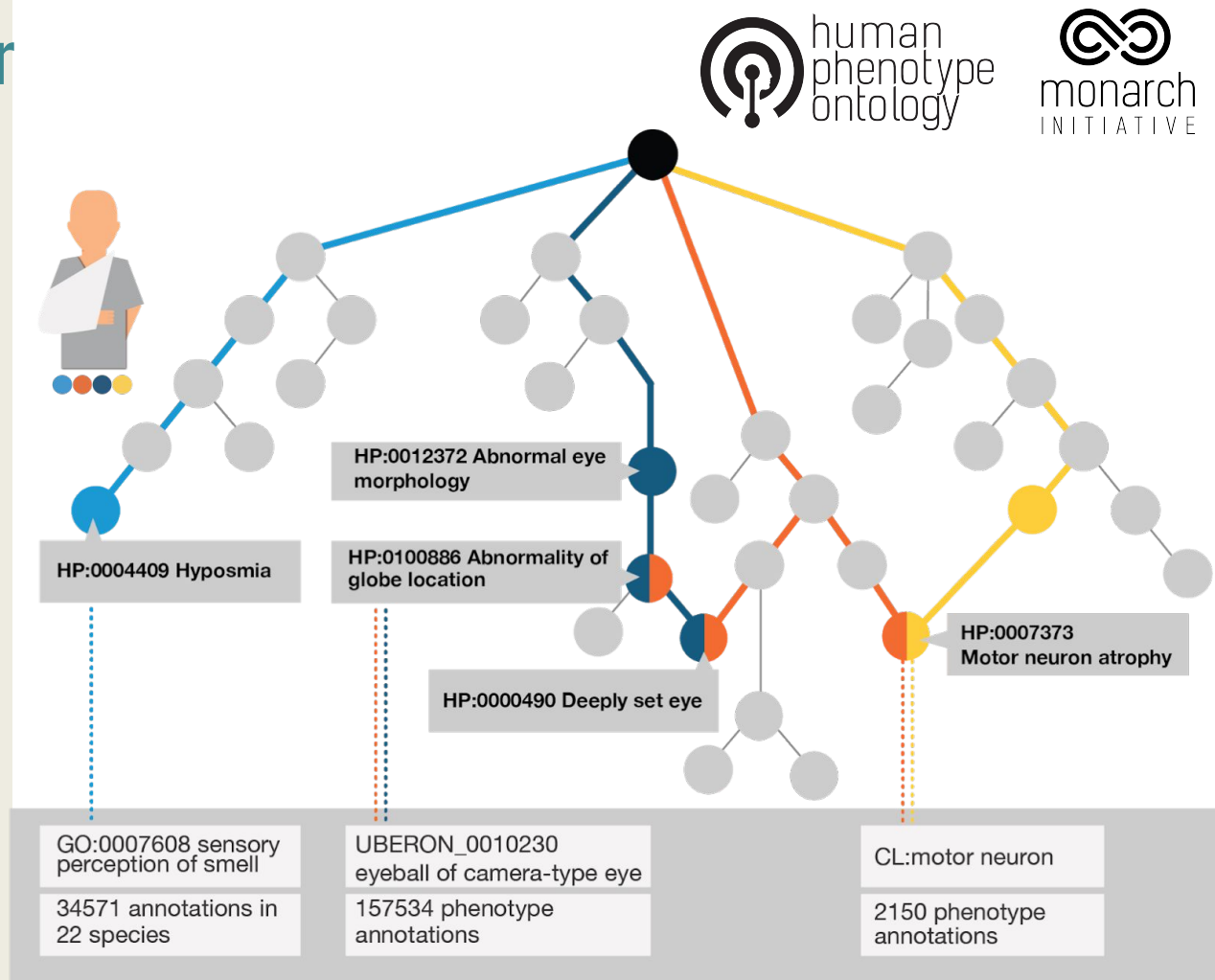
Ontologies as a tool for data integration

- Across species
- MAYBE REMOVE



Ontologies as a tool for data integration

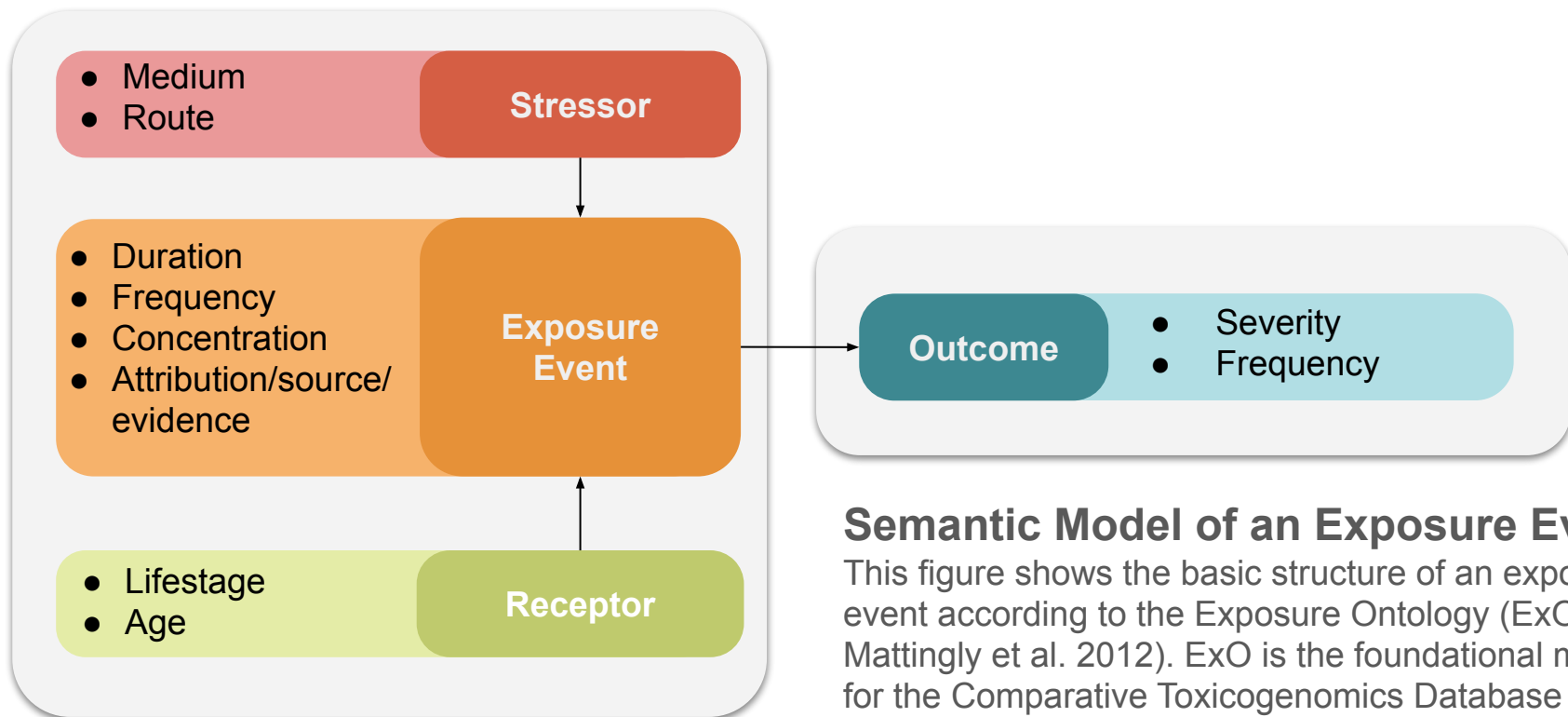
- Across species
- Granularity mismatch
- MAYBE REMOVE
BECAUSE WE HAVE 8



Exposures: The missing piece



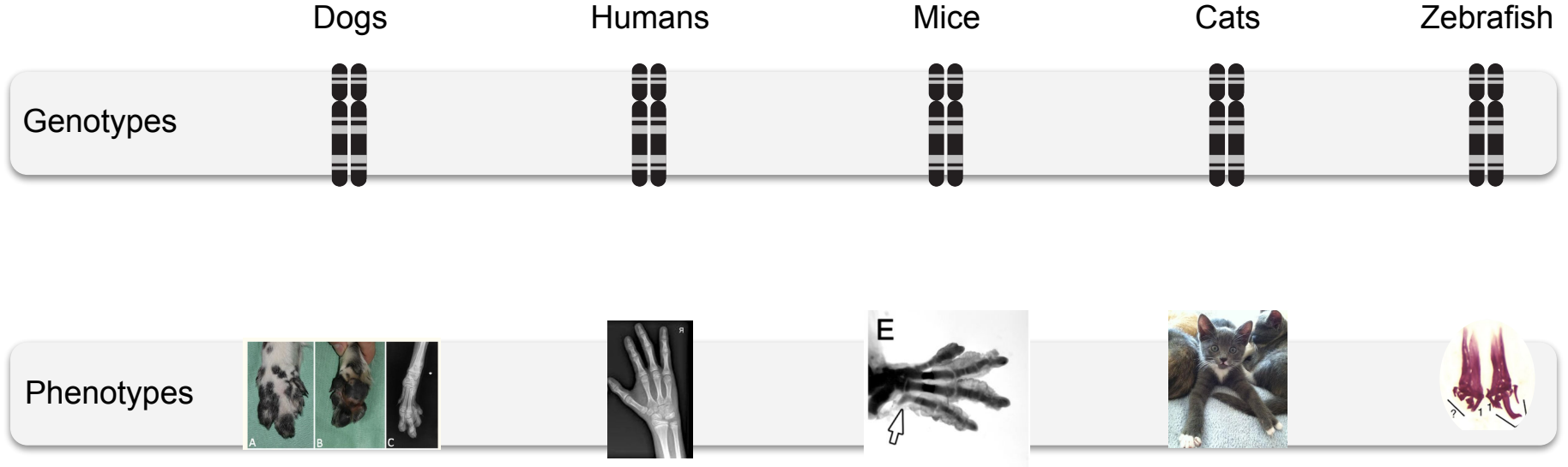
Environmental Exposures Modeling with ExO



Semantic Model of an Exposure Event.

This figure shows the basic structure of an exposure event according to the Exposure Ontology (ExO; Mattingly et al. 2012). ExO is the foundational model for the Comparative Toxicogenomics Database (CTD; Mattingly et al. 2006).

Why do we care about phenotype integration?

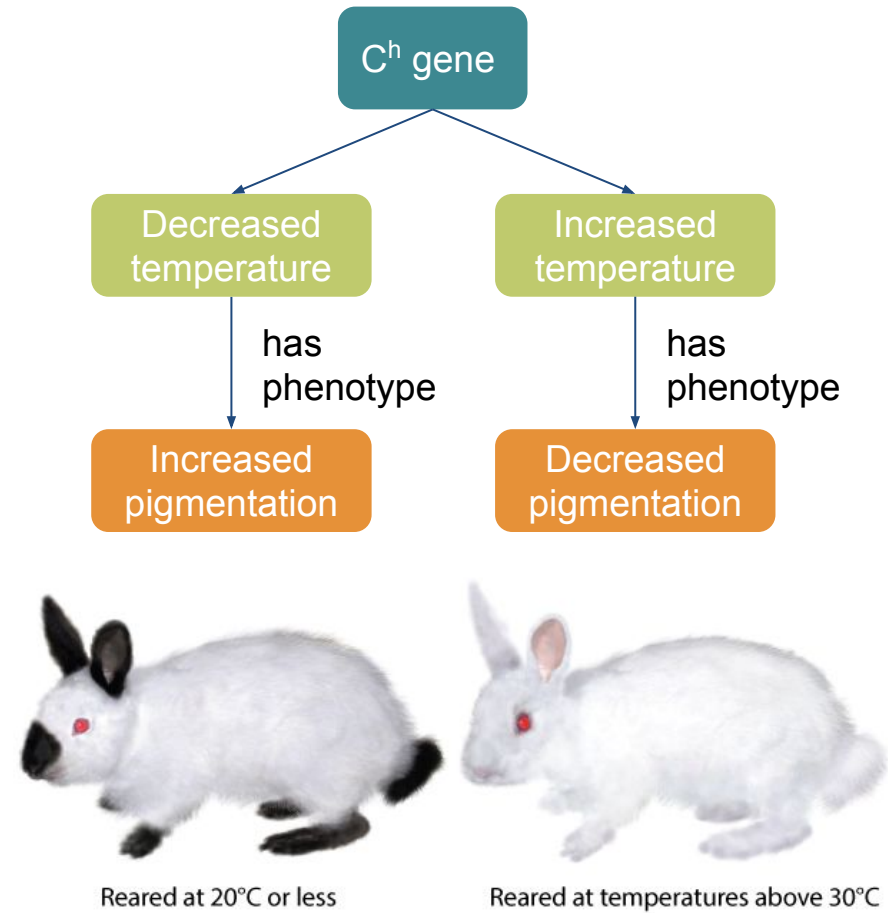


Modeling Genotype + Environment = Phenotype

Genotype

Environment

Phenotype



1. What are the strategies for standardizing, sharing, and aggregating health records and relevant metadata across species? What are the best practices for collection, storage, and analysis of biosamples to assess exposures (e.g., biorepository resources, DNA susceptibility, DNA methylation, microbiome, etc.)?
2. What are the obstacles (e.g., scientific, infrastructure, ethical, and financial) to using companion animals as biomonitors, and what are potential solutions?
3. What are the best practices for data integration of human and companion animal data? Have you identified best practices or lessons learned in efforts to link datasets?
4. Are there opportunities to promote data sharing and collaboration to advance research on the role of companion animals as sentinels for predicting environmental exposure effects? If so, what are they?
5. What incentives are needed to encourage data sharing and collaboration? How can disincentives for data sharing be overcome and the use of incentives become more standard? What infrastructure and resources are needed? Which groups need to be engaged?
6. What are the implications for expanded, systematic collection of this data? Could this help to identify environmental hazards and provide an early warning system for public health interventions?