

# Capturing and Sharing Data for Rare Disease: Building Algorithms, Software, and a Community

---

**Michael Brudno**

Centre for Computational Medicine  
Hospital for Sick Children & University of Toronto

CS/Medicine, 27 February 2018

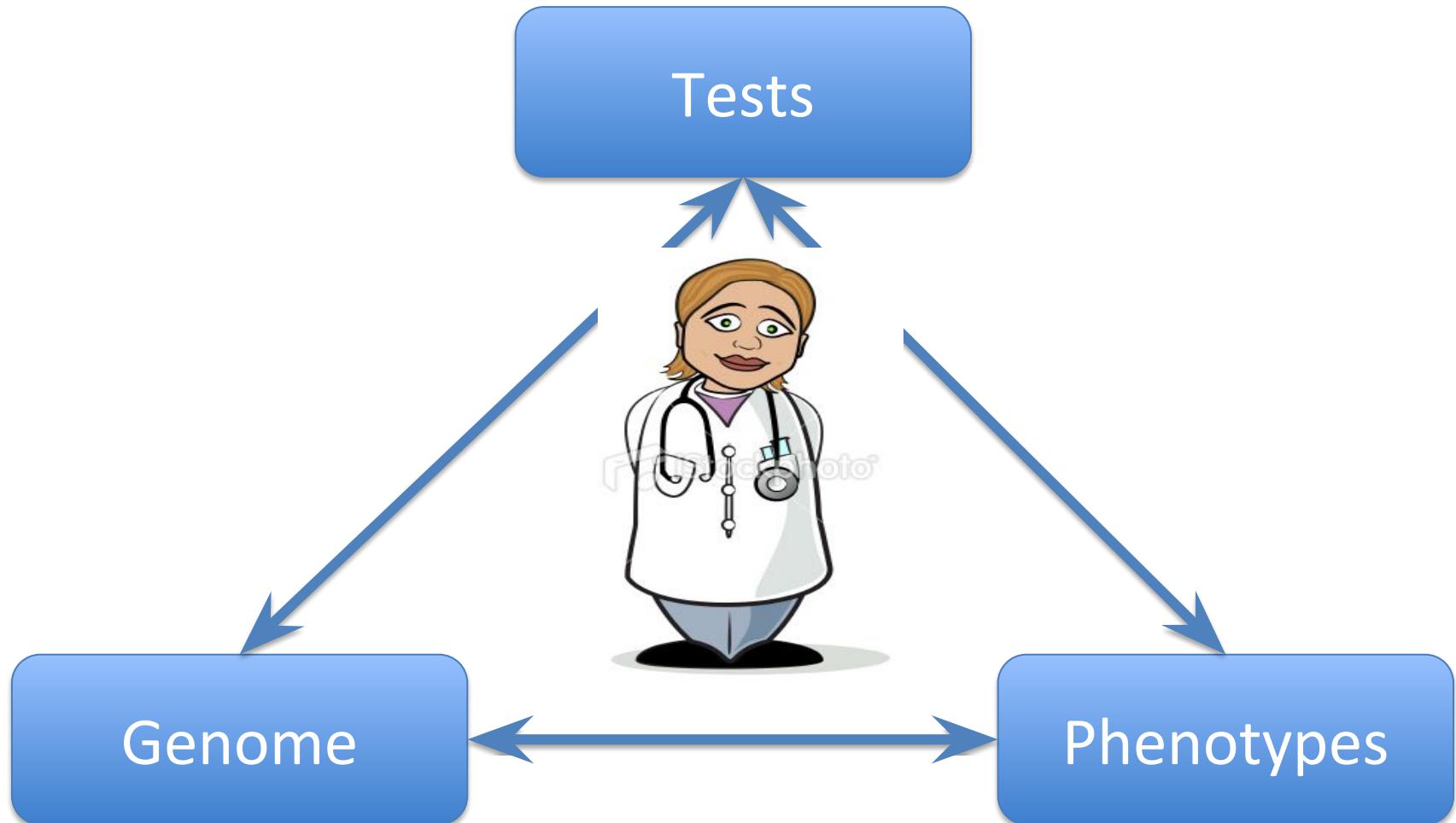


**SickKids®**

# This is a modern EHR



# What an EHR *should* be



- An EHR Should GUIDE the clinician around complex (genomic) data!

# Deep Phenotyping

- > Describe the features of an individual, rather than a disease
  - To enable a diagnosis (especially of a rare disease)
  - To distinguish between similar disorders
  - To enable genotype-phenotype correlations

# Previous State of Clinical Phenotyping

- Two Alternatives: free text or checkboxes

## Dysmorphic features

- df
- dysmorphic
- dysmorphic faces
- dysmorphic features

## Congenital malformation/anomaly:

- congenital anomaly
- congenital malformation
- congenital anamoly
- congenital anomly
- congenital anomaly
- congenital anomaly
- congenital anomalies
- cong. m.
- cong. Mal
- cong. malfor
- congenital malform
- congenital m.
- multiple congenital anomalies
- multiple congenital abnormalities
- multiple congenital abnormalities

## Phenotypic description (Clinical symptoms)

### Behavior, Cognition and Development

- Global development delay
- Fine motor delay       Gross motor delay
- Language delay
- Learning disability
- Mental retardation

### Examples of lists:

- \* Moderate dd. cong. malfor. behav. pro.
- \* Severe dd. mental retardation
- Attention deficit hyperactivity disorder
- Autism
- Pervasive developmental delay
- Psychiatric disorders (Specify below)
- Other: mental retard.short stature

### Neurological

- Hypotonia
- Seizures
- Ataxia
- Dystonia
- Chorea

### Cardiac

- ASD
- VSD
- AV canal defect
- Coarctation of aorta
- Tetralogy of fallot

Other: \_\_\_\_\_

### Craniofacial

- Craniosynostosis
- Cleft lip
- Microretrognathia
- Facial dysmorphism (Specify below)
- Other: \_\_\_\_\_

Cleft palate

Retrognathia

### Eye Defects

- Blindness
- Coloboma
- Epicanthus
- Eyelid abnormality (Specify bellow)
- Other: \_\_\_\_\_

# Problems with the status quo

- Phenotypic descriptions that are very evocative for humans, unreadable to a computer:
  - “first words at 5 years”
  - “has trouble spelling”
  - “recognizes only close relatives”
- Multiple terms have the same meaning: “generalized amyotrophy”, “generalized muscle atrophy”, “muscular atrophy, generalized”
- It is difficult to do computation with phenotypes!

# Football Analogy

- A word to you may mean something different than to me:



# Ontologies are terms with relationships

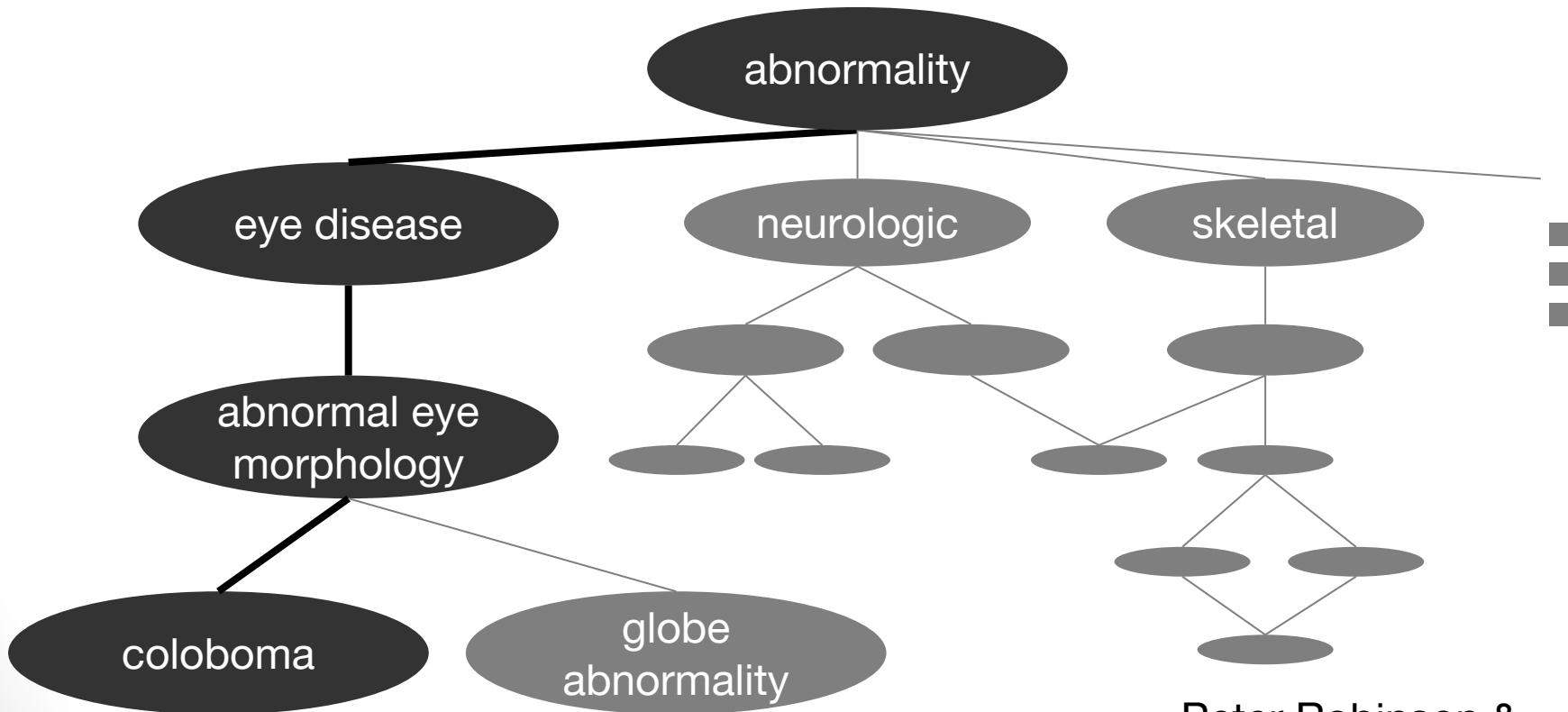
- Sports
  - Ball Sports
    - Football-related sports
      - North American Football
        - American Football
        - Canadian Football
      - Association Football (Soccer)
      - Rugby derivatives
        - Union
        - Aussie-rules
        - Gaelic
      - ...

# Next-generation phenotyping

Human Phenotype Ontology (HPO):

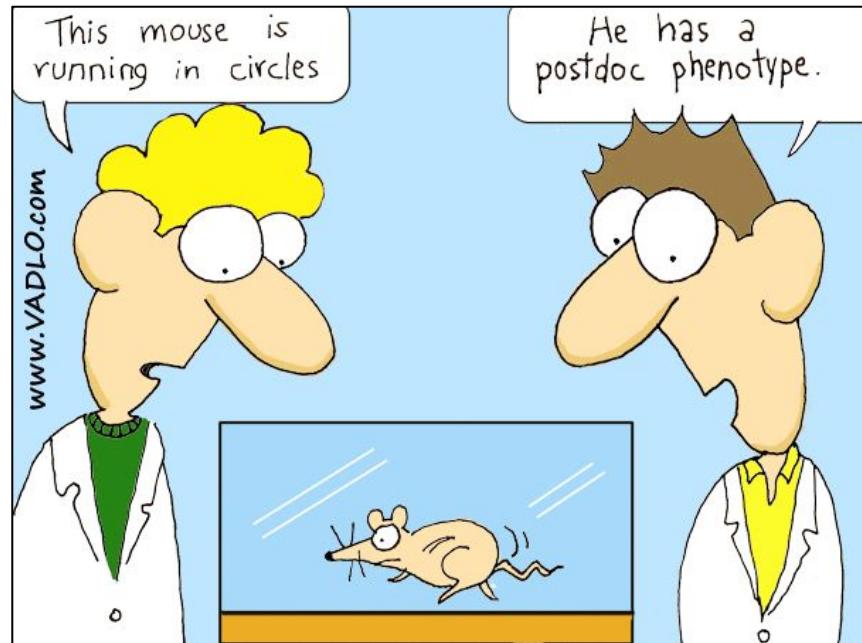
11,000+ terms

100,000+ links to 5,000+ OMIM/ORDO Disorders



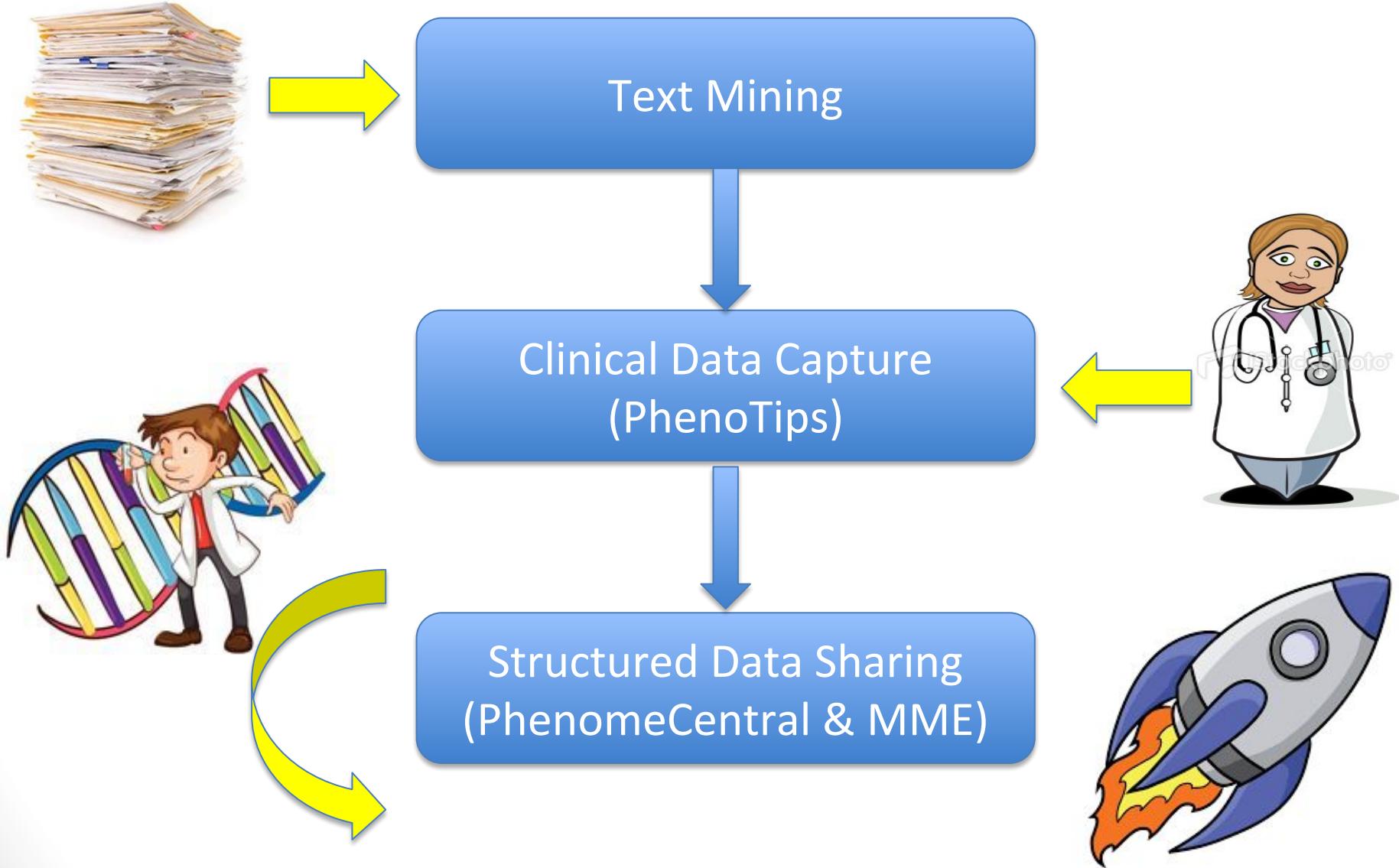
# “Model Organism” Ontologies

- [NA Y N] [HP:0000708] Behavioural/Psychiatric Abnormality
  - ▶ [NA Y N] [HP:0100851] **Abnormal emotion/affect behavior**
    - ▶ [NA Y N] [HP:0006919] Abnormal aggressive, impulsive or violent behavior
    - ▼ [NA Y N] [HP:0100852] Abnormal fear/anxiety-related behavior
      - [NA Y N] [HP:0000756] Agoraphobia
      - ▶ [NA Y N] [HP:0000739] Anxiety
      - [NA Y N] [HP:0000712] Emotional lability
      - [NA Y N] [HP:0001575] Mood changes
      - [NA Y N] [HP:0000720] Mood swings
    - [NA Y N] [HP:0012154] Anhedonia
    - [NA Y N] [HP:0000741] Apathy
  - ▶ [NA Y N] [HP:0000729] Autism spectrum disorder
  - [NA Y N] [HP:0100024] Conspicuously happy disposition
  - ▶ [NA Y N] [HP:0000716] Depression
  - [NA Y N] [HP:0010529] Echolalia
  - ▼ [NA Y N] [HP:0000719] Inappropriate behavior
    - [NA Y N] [HP:0000734] Disinhibition
    - ▶ [NA Y N] [HP:0000748] Inappropriate laughter
    - [NA Y N] [HP:0008768] Inappropriate sexual behavior
  - [NA Y N] [HP:0000732] Inflexible adherence to routines or rituals
  - [NA Y N] [HP:0000737] Irritability
  - [NA Y N] [HP:0000757] Lack of insight
  - [NA Y N] [HP:0000745] Lack of motivation
  - [NA Y N] [HP:0000721] Lack of spontaneous play
  - [NA Y N] [HP:0000744] Low frustration tolerance
  - [NA Y N] [HP:0002300] Mutism
  - [NA Y N] [HP:0010865] Oppositional defiant disorder
  - [NA Y N] [HP:0100025] Overfriendliness



- [NA Y N] [HP:0100025] Overfriendliness
- [NA Y N] [HP:0002193] Pseudobulbar behavioral symptoms
- ▶ [NA Y N] [HP:0000711] Restlessness
- [NA Y N] [HP:0000723] Restrictive behavior
- [NA Y N] [HP:0100962] Shyness

# Talk Overview



---

# Concept-recognition of Clinical Terms: Ontology-aware Deep Learning

---



Aryan Arbabi

# Introduction

- > Goal:
  - automatic annotation of unstructured text with phenotypes
- > Example:

There was marked variability in the severity and nature of manifestations with 2 having severe hand and foot involvement in addition to craniofacial changes compatible with a diagnosis of Freeman-Sheldon syndrome. Other apparently unrelated hereditary disorders in the family included ectrodactyly, biliary atresia, and Brachmann-de Lange syndrome.

# More details on the annotation example

There was marked variability in the severity and nature of manifestations with 2 having severe hand and foot involvement in addition to craniofacial changes compatible with a diagnosis of Freeman-Sheldon syndrome. Other apparently unrelated hereditary disorders in the family included ectrodactyly, biliary atresia, and Brachmann-de Lange syndrome.

HP:0002260  
Abnormal facial shape

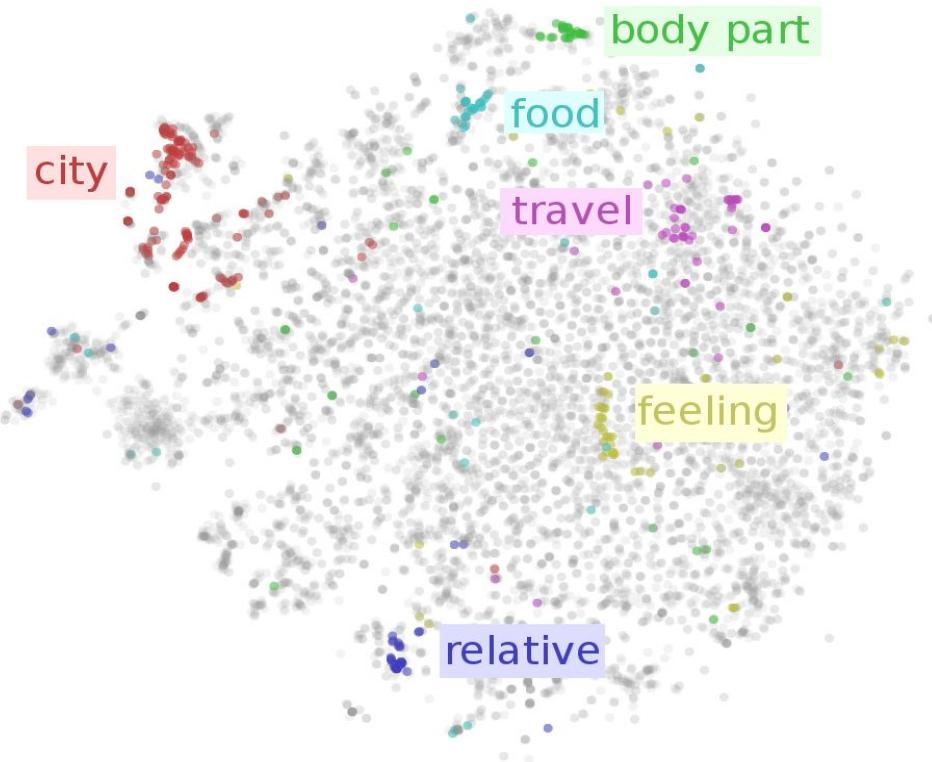
HP:0005912  
Biliary atresia

## ***Synonyms for HP:0002260***

Facial dysmorphism  
Malformation of face  
Unusual facial appearance  
Unusual facies  
Facial Dysmorphism  
Funny looking face  
Dysmorphic facial features  
Deformity of face  
Dysmorphic facies  
Distinctive facies  
Distortion of face

# Word embeddings

- > Each word is mapped to a vector in high-dimensional space
- > The vector representations are learned by training on large corpora
- > Words that occur in similar contexts tend to have closer vectors



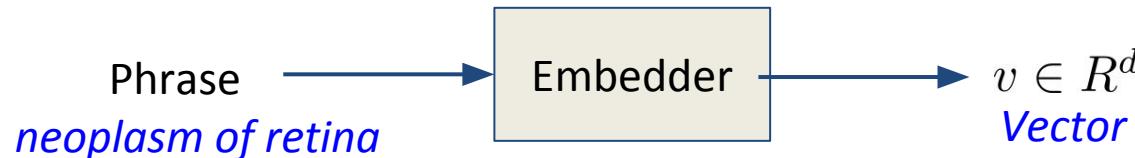
# Word embedding - biomedical text

- > Used GloVe (Pennington et al.) to learn word embeddings by training on publicly available articles in PubMedCentral (~ 1m documents)
- > Examples of 5 closest words:

eye	hospital	seizure	intron	short	BRCA
eyes	hospitals	seizures	intronic	long	BRCA2
ocular	department	epileptic	exon	longer	BRCA1
vision	medical	epilepsy	exons	shorter	MSH6
visual	outpatient	electrographi c	introns	end	MSH2
optic	admitted	epileptiform	untranslated	extended	HNPPCC

# Intuitions on method

- Learn a function that embeds phrases into vectors:



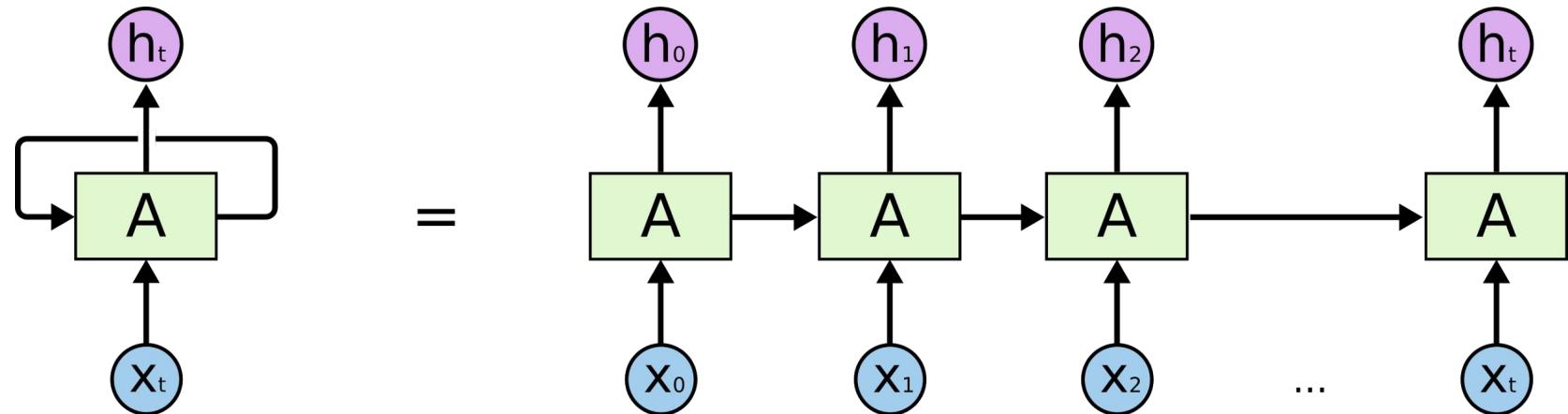
- Learn embeddings for all HPO concepts

HP:0000118	$v_0$
HP:0002715	$v_1$
...	...
HP:0000667	$v_n$

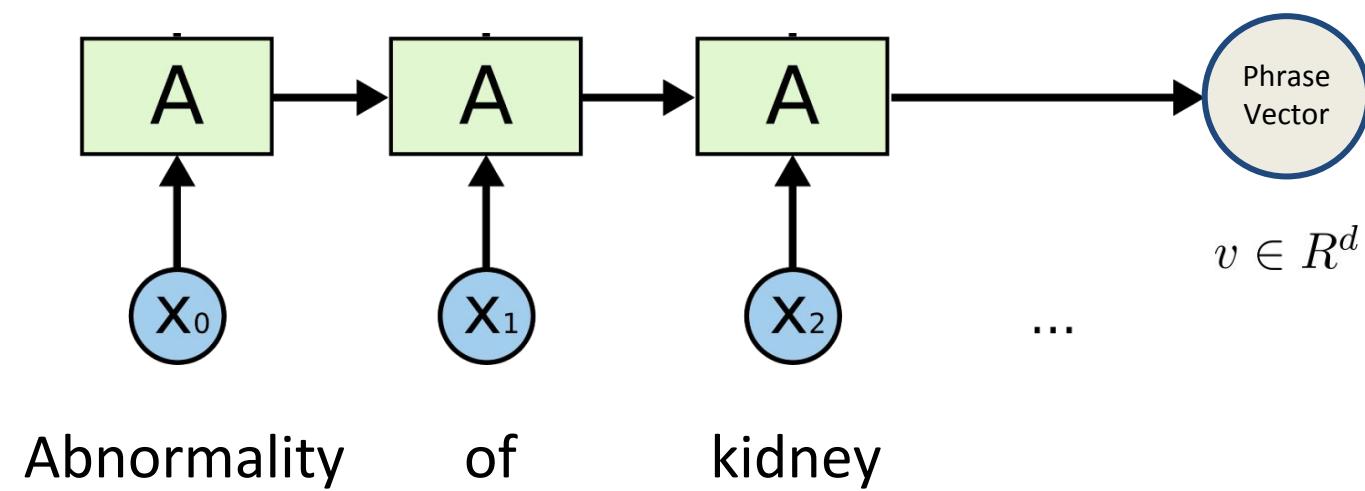
- Goal: Phrases “close” in HPO have close vectors

# How do we embed phrases?

- > Recurrent Neural Networks (RNN)



# Embedding phrases with RNN

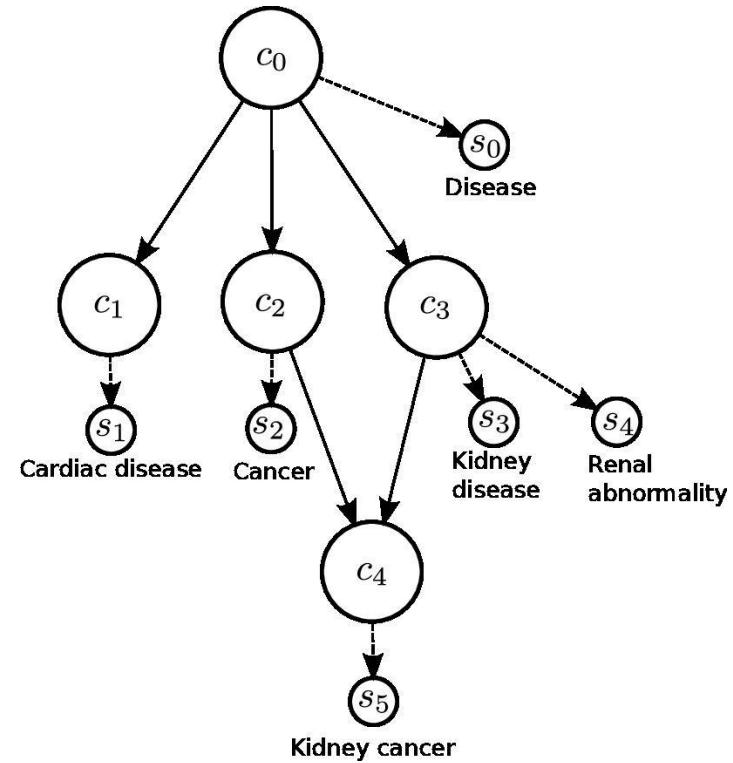


# How to use the ontology information?

- > Vendrov et al. (2015) introduced a method named Order-Embedding for modeling hierarchies
- >  $x$  and  $y$  are two vector representations with  $N$  dimensions

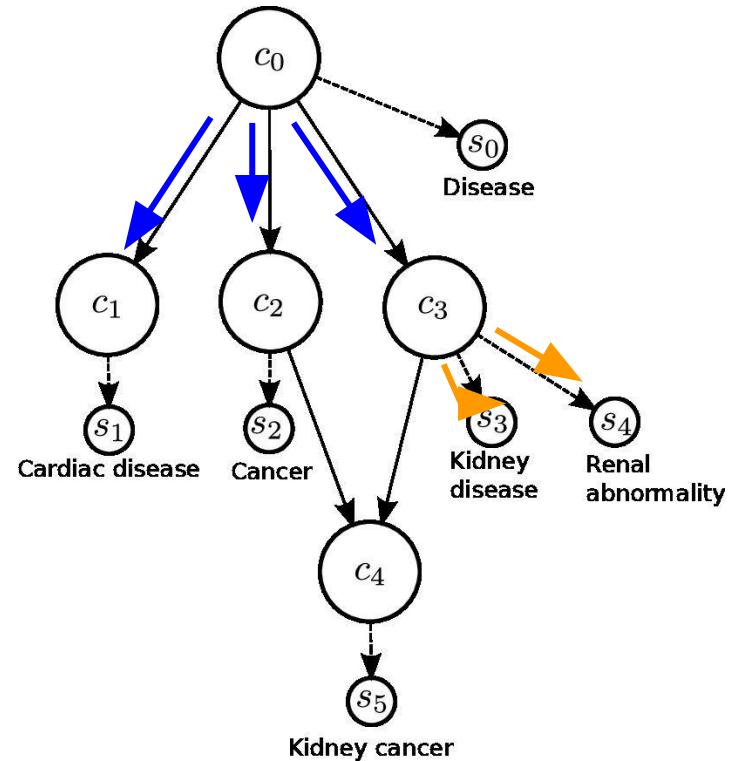
$$x \preceq y \text{ if and only if } \bigwedge_{i=1}^N x_i \geq y_i$$

- > A distance function is defined  $D(x,y)$ 
  - Measures how much the above property is violated



# How to use the ontology information?

- › During the training, a loss function is used that includes the following:
- › The order-embedding requirements are maintained between the concepts ( $c_i$ )
- › The phrase embedding for all synonyms of a concept are as close as possible to the concept's vector



# Experiments

- > We tested on the data-set provided for the Bio-Lark annotator (Cohen et al. 2010)
- > Data includes 229 paragraphs, manually annotated with HPO terms
- > There are 1933 total annotations
- > 955 unique HPO phrases

# Experiments

> We experimented both on phrase level and sentence level

> Phrase level:

Craniofacial changes → HP:0002260

> Sentence level:

There was marked variability in the severity and nature of manifestations with 2 having severe hand and foot involvement in addition to craniofacial changes compatible with a diagnosis of Freeman-Sheldon syndrome.

HP:0002260

# Results



(a) Hierarchy used for training



(b) Hierarchy NOT used for training

# Results

## Phrase level:

Ensemble experiments (train:500, test:455)	Top-1	Top-5	Top-10
PhenoTips	48.79	65.05	69.45
NCR	54.94	71.20	75.60
NCR+PhenoTips	55.60	71.64	77.36

## Sentence level:

Method	Micro (%)			Macro (%)			Extended (%)			Jaccard (%)
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	
BioLarK	78.5	60.5	68.3	71.3	60.7	65.6	79.0	68.2	73.2	74.2
cTAKES	72.2	55.6	62.8	64.4	56.1	60.0	<b>82.5</b>	61.9	70.7	71.0
NCR	<b>79.5</b>	<b>62.1</b>	<b>69.7</b>	<b>74.2</b>	<b>62.2</b>	<b>67.7</b>	81.6	69.0	<b>74.8</b>	<b>75.9</b>
NCR -H	69.9	60.5	64.9	64.4	<b>62.2</b>	63.3	71.3	<b>70.5</b>	70.9	69.6
NCR -N	78.7	61.6	69.1	73.0	61.6	66.8	79.8	68.4	73.7	74.5
NCR -HN	66.0	60.0	62.9	62.3	61.5	61.9	68.8	69.3	69.0	66.5

# Discussion

- > Word embeddings help with occurrences of word synonyms
- > Using the hierarchical structure of HPO to learn how words are contributing
- > Multiple synonyms for terms help better learning synonyms for words, and their importance
- > The method is not yet character-aware (typos or lexical differences are not caught)

# PhenoTips: Deep Phenotyping Platform for clinical and research use

---



Marta Girdea

# Mining Clinical Records

- > Billing codes are commonly misused, and for rare diseases lack specificity
- > Most valuable data in an EHR is unstructured (free text charts and notes)
- > Mining free text is messy & inaccurate
  - E.g. 73% accuracy at getting whether a patient has dementia in recent study
  - Difficult to use in patient care or in rare disease, where each phenotype is important

# Using ontologies patient-side

- Ontologies are large (HPO has > 11,000 terms) and difficult to use
- Re-mapping data to an ontology post-visit is time consuming and prone to error
- Best time to phenotype using ontologies is during the patient visit
- Goals of our work
  - Make deep phenotyping simple
  - Make it “faster than paper”
  - Enable broad data sharing for diagnosis

# Demo

---

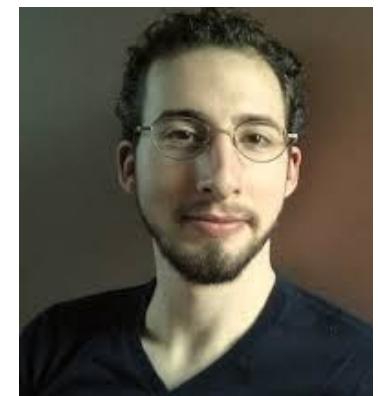
# Advantages of ontologies & PhenoTips

- Integration of data between studies, identification of unrelated patients
- Better and more thorough automated genome analysis and variant prioritization
- Phenotype patients in the exam room – more accurate, with less redundant data entry
- Training for next generation:
  - Diagnosis assistance
  - Identify previously seen similar patients
  - Decisions based on prior outcomes

---

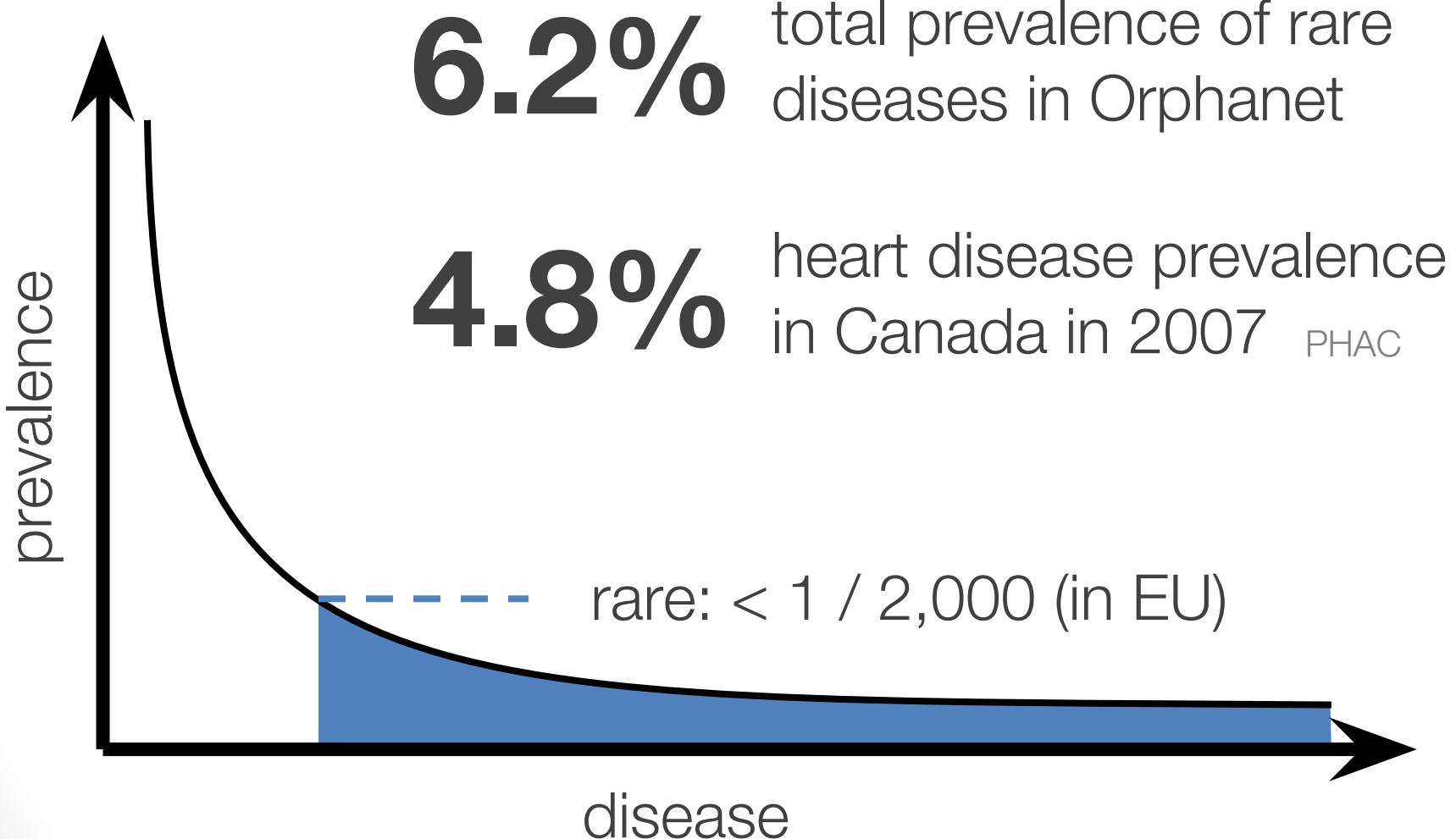
# PhenomeCentral & MatchMaker Exchange sharing phenotype and genotype data for rare genetic disorders

---



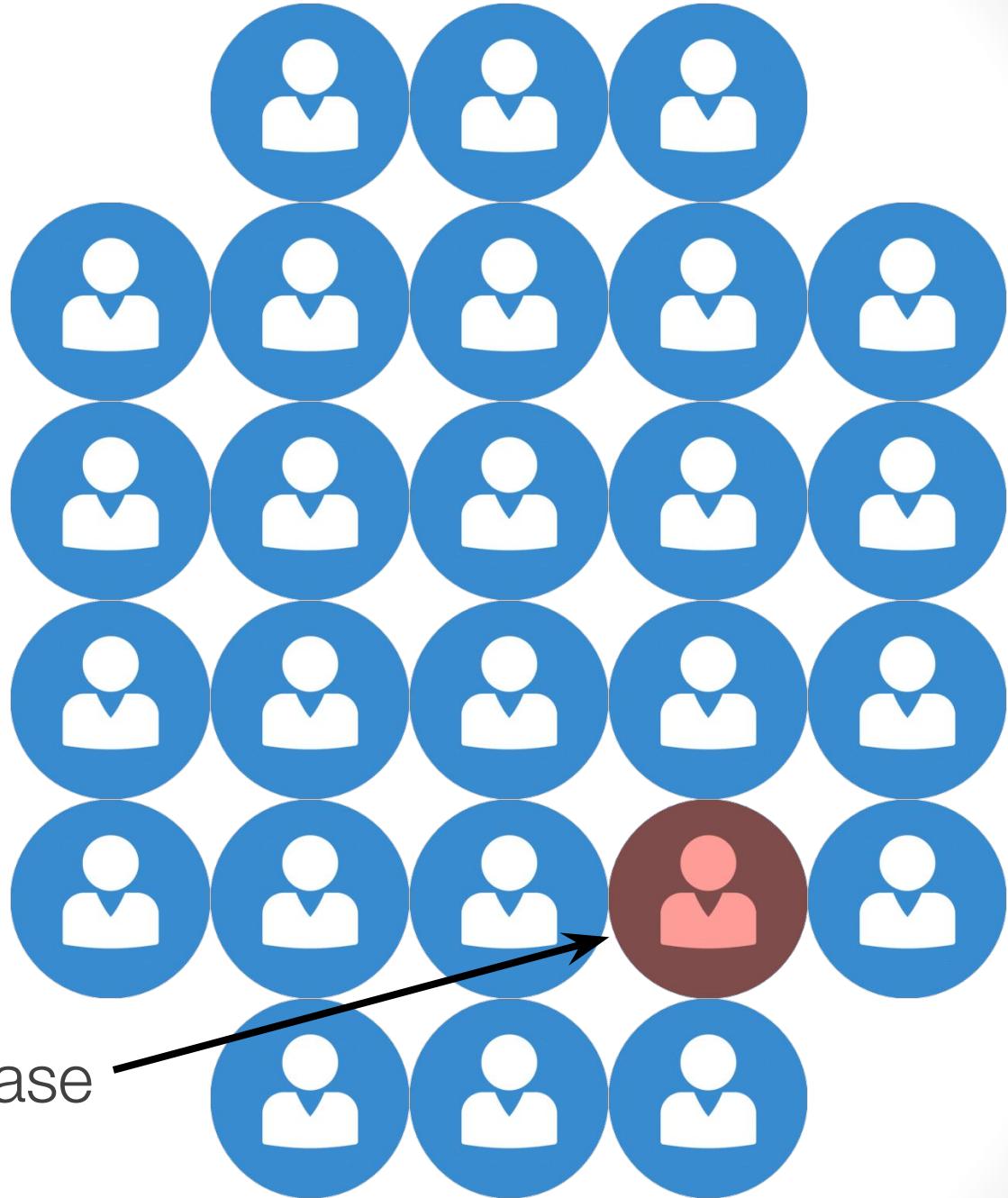
Orion Buske

# Rare diseases and the long tail





rare genetic disease

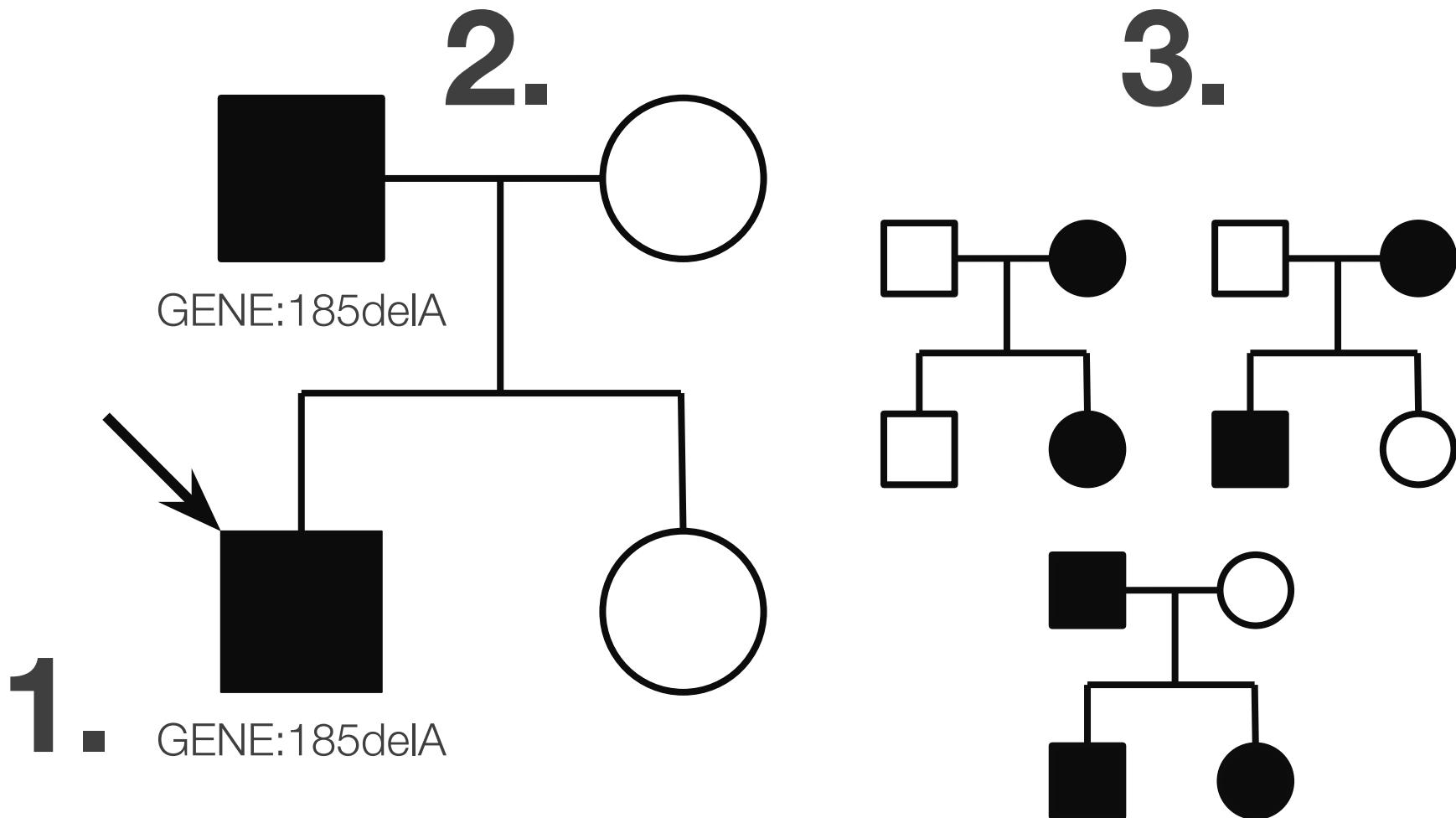




- > might not recognize known disease
- > insufficient sample size for novel gene



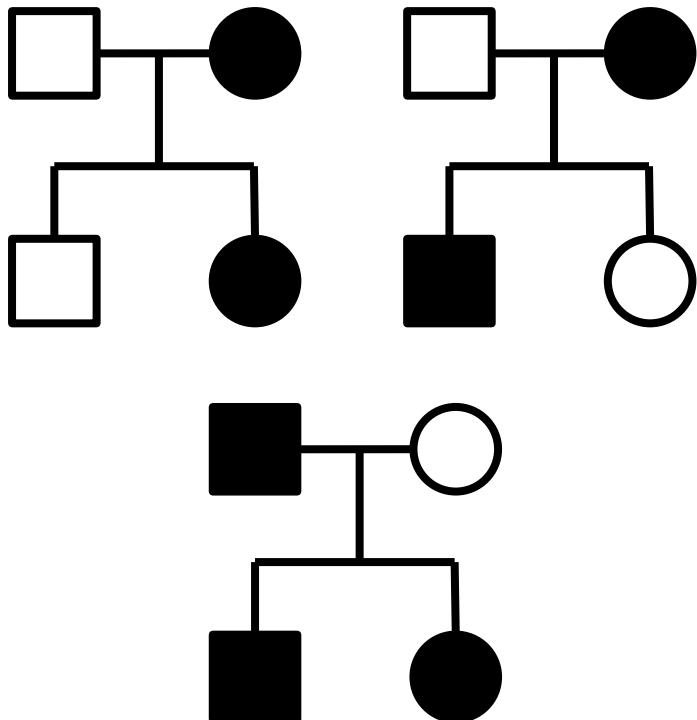
# Finding the cause of a genetic disease...

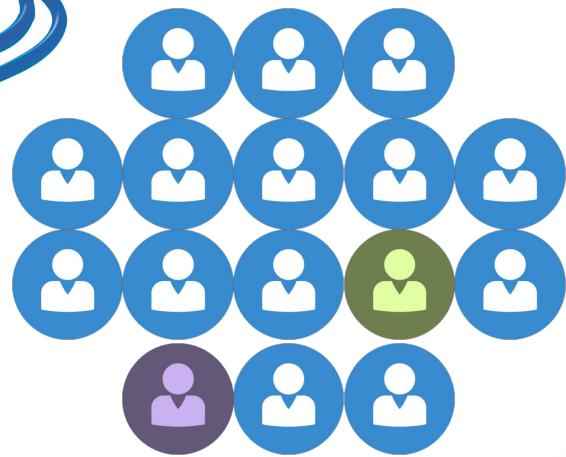
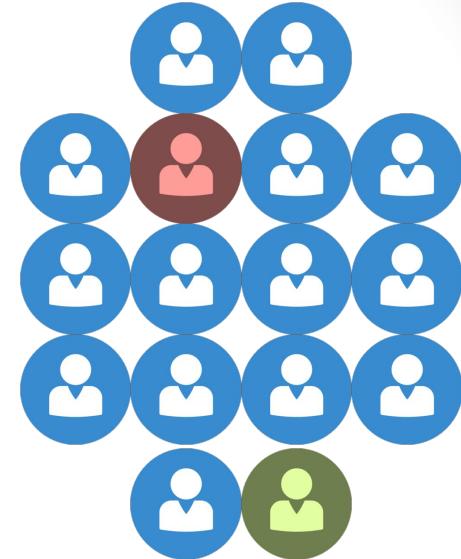
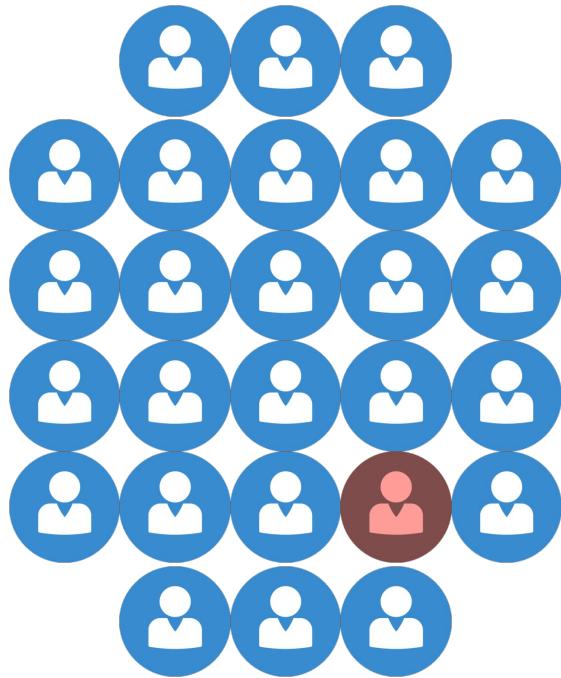


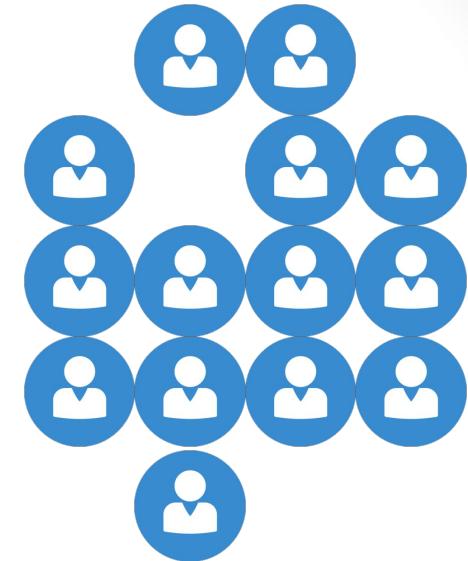
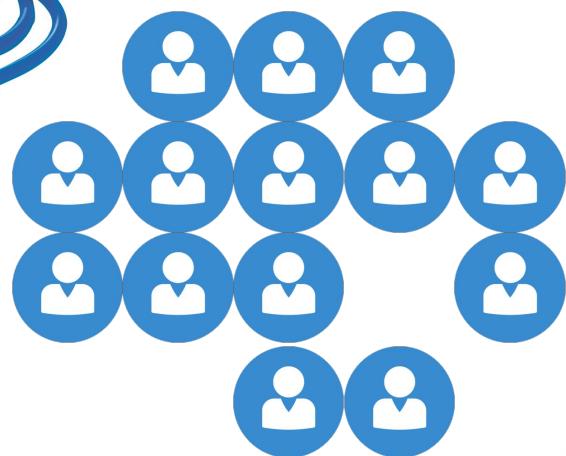
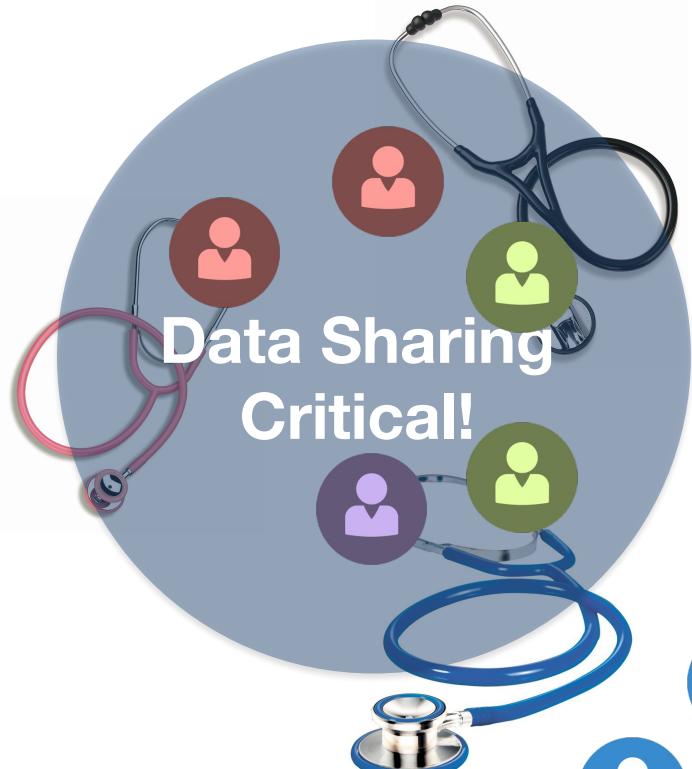
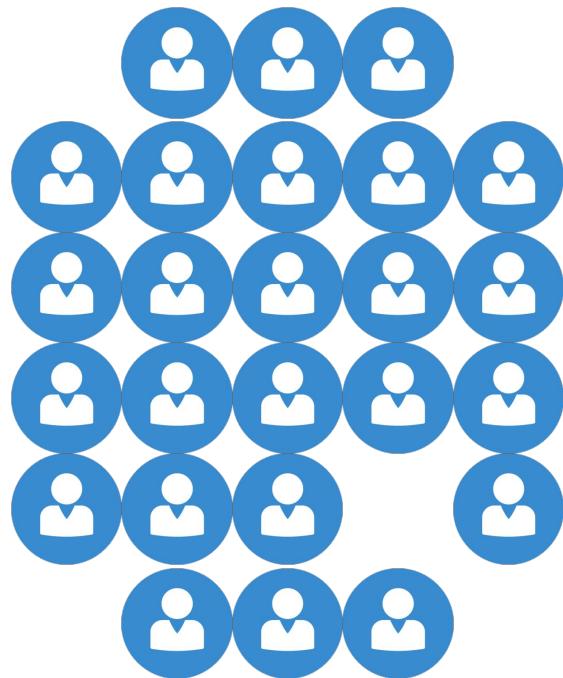
Finding additional families can be difficult...

**rare  
diseases  
are  
rare.**

**3.**

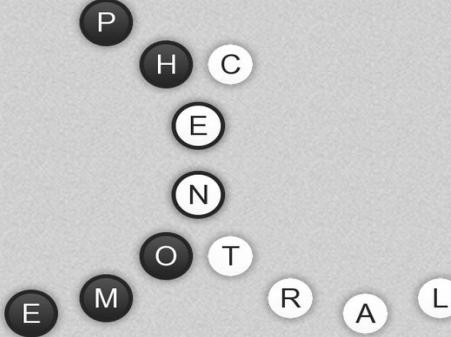






— to encourage data sharing we built a  
user-friendly, privacy-aware  
portal

— for discovering patients similar to yours



# PhenomeCentral

A hub for secure data sharing within the rare disorder community

[Sign up](#)

[Login](#)

[Read more...](#)

- > PhenomeCentral is a Matchmaker
  - Find out about other similar patients
  - Easily connect with other clinicians
- > Each Patient Record can be:
  - *Public* – Visible to all registered users
  - *Private* – Only visible to specified users/consortia
  - *Matchable* – Private visibility, but existence can be "discovered" by users who submit similar patients

# Step 1: submit your patient

The screenshot shows the PhenotypeXpress interface for submitting a patient record. At the top, there's a 'QUICK PHENOTYPE SEARCH:' bar with a search input field. Below it, under 'CURRENT SELECTION', 'BEHAVIOR, COGNITION AND DEVELOPMENT' is selected, showing 'Delayed gross motor development' as the current choice. The main content area is divided into sections: 'Ownership', 'Global visibility', 'Collaborators', and 'Medical report (optional)'. In the 'Ownership' section, the owner has full editing and managing capabilities. In 'Global visibility', options include 'hidden', 'private', 'matchable' (selected), and 'public'. The 'Collaborators' section shows 'Kym Boycott' as a collaborator. The 'Medical report (optional)' section is currently empty.

QUICK PHENOTYPE SEARCH:

Enter keywords and choose among suggested ontology terms

CURRENT SELECTION

BEHAVIOR, COGNITION AND DEVELOPMENT

Delayed gross motor development Delete Add details

Intellectual disability, moderate Delete Add details

Access rights for P0000190

**Ownership**

The owner has full editing and managing capabilities for a case.

This case is owned by you. You can:

- Keep current owner
- Transfer ownership to one of your groups New owner group
- Transfer ownership to another user New owner

**Global visibility**

Global visibility refers to how registered users can VIEW this case. Visibility levels do not change editing or managing permissions.

- hidden Hidden cases are only accessible to their owners, and don't contribute to aggregated statistics.
- private Private cases are only accessible to their owners, but they do contribute to aggregated statistics.
- matchable Users can discover the existence of matchable cases when they contribute a case that is similar to them, but cannot have access to the full case without explicit permissions from the owner.
- public All registered users can view public cases.

**Collaborators**

Collaborators can be given access to view, edit or share specific cases, regardless of the case's visibility setting. The type of advanced privileges can be established by the owner when selecting a collaborator:

Kym  
hide suggestions  
Kym Boycott  
KBoycott

Other (enter free text and choose among suggested ontology terms)

Medical report (optional):

None available + UPLOAD AND MANAGE

**CARDIAC**

Defect in the atrial septum Ventricular septal defect

- > Built-in PhenoTips interface
- > Export anonymized records from other instances
- > Add a VCF file and/or gene list
- > Set permission and add collaborators

# Step 2: see patients similar to yours

**F0000010**      Reported by **Marta Girdea (admin)** on 2013/09/29 18:10 · Last modified by **Marta Girdea** on 2013/09/30 14:00

This case is owned by Care4Rare, it is public and it is shared with 1 collaborator.

**Patient information**

Identifier: KB\_174\_FHS1-1  
Sex: Female

---

**Clinical symptoms and physical findings**

**PHENOTYPIC FEATURES BREAKDOWN**

CRANIOFACIAL	
Low hanging ear	■ ■ ■ ■ □ 75%
Thin upper lip	
Short philtrum	
Triangular face	
Wide nose	
Prominent nares	
Narrow nasal bridge	
Long nose	
NO Wide mouth	

EAR DEFECTS	
Low-set ears	■ ■ ■ ■ □ 66%
Recurrent otitis	

MUSCULOSKELETAL	
Broad fingers	■ ■ ■ ■ □ 62%
Brachydactyly	
Broad thumb	

GENITOURINARY	
Nephrocalcinosis	■ ■ ■ ■ □ 47%
Hydronephrosis	

BEHAVIOR, COGNITIVE	
Moderate expressive language delay	■ ■ ■ ■ □ 13%

**Diagnosis**

OMIM disorder:

**Similar cases**  
Showing 10 similar cases

Case ID	Diagnosis
F0000021	#136140 HUMAN INGAP-HARBOR SYNDROME, FLHS
F0000019	#136140 HUMAN INGAP-HARBOR SYNDROME, FLHS
F0000012	#136140 HUMAN INGAP-HARBOR SYNDROME, FLHS
F0000009	#136140 HUMAN INGAP-HARBOR SYNDROME, FLHS
F0000011	#136140 HUMAN INGAP-HARBOR SYNDROME, FLHS
F0000020	#136140 HUMAN INGAP-HARBOR SYNDROME, FLHS
F0000014	#136140 HUMAN INGAP-HARBOR SYNDROME, FLHS
F0000017	#136140 HUMAN INGAP-HARBOR SYNDROME, FLHS
F0000016	#136140 HUMAN INGAP-HARBOR SYNDROME, FLHS
F0000015	#136140 HUMAN INGAP-HARBOR SYNDROME, FLHS

**UNMATCHED**

The current patient (P0000081) presented with:

- Hypoplastic toenails
- Delayed CNS myelination
- Defect in the atrial septum
- Downslanted palpebral fissures
- Seizures
- Global developmental delay
- Malar flattening
- Stenosis of the external auditory canal
- Conductive hearing impairment
- Choanal atresia
- Cleft eyelid

The matched patient presented with:

- 

Matches found for 11 out of 17 features.

Show matches...

**GENE MATCHING BREAKDOWN**

EDN1	
Estimated relevance for the observed phenotype in the current patient (P0000081):	■ ■ ■ ■ □ 95%
Estimated relevance for the observed phenotype in the matched patient:	■ ■ ■ ■ □ 95%

VARIANT	ESTIMATED HARMFULNESS
chr6: 12290906 - 12290915	■ ■ ■ ■ □ 95%
GCCAAGGGAGC → - (FS_DELETION)	

**STAG2**

Estimated relevance for the observed phenotype in the current patient (P0000081): ■ ■ ■ ■ □ 77%

Estimated relevance for the observed phenotype in the matched patient: ■ ■ ■ ■ □ 77%

**ABC1**

Estimated relevance for the observed phenotype in the current patient (P0000081): ■ ■ ■ ■ □ 73%

Estimated relevance for the observed phenotype in the matched patient: ■ ■ ■ ■ □ 50%

**ANKRD20A4**

Estimated relevance for the observed phenotype in the current patient (P0000081): ■ ■ ■ ■ □ 34%

Estimated relevance for the observed phenotype in the matched patient: ■ ■ ■ ■ □ 34%

**ANKRD20A3**

Estimated relevance for the observed phenotype in the current patient (P0000081): ■ ■ ■ ■ □ 13%

Estimated relevance for the observed phenotype in the matched patient: ■ ■ ■ ■ □ 13%

**HIDE VARIANTS...**

**SHOW VARIANTS...**

**SHOW VARIANTS...**

**SHOW VARIANTS...**

**SHOW VARIANTS...**

Step 3: contact the other submitter

## PHENOTYPIC FEATURES BREAKDOWN

### PROXIMAL MUSCLE WEAKNESS

The current patient (P0000131) presented with:  
Proximal muscle weakness

■ ■ □ □ 59%

The matched patient (P0000371) presented with:  
Proximal muscle weakness

■ ■ □ □ 55%

### MYOPATHY

The current patient (P0000131) presented with:  
Myopathy

The matched patient (P0000371) presented with:  
EMG: myopathic abnormalities

■ ■ □ □ 55%

### THROMBOCYTOPENIA

The current patient (P0000131) presented with:  
Thrombocytopenia

The matched patient (P0000371) presented with:  
Thrombocytopenia

■ □ □ □ 27%

### ABNORMALITY OF THE MUSCULATURE

The current patient (P0000131) presented with:  
Exercise-induced muscle fatigue

The matched patient (P0000371) presented with:  
Frequent falls

■ □ □ □ 15%

### ABNORMALITY OF THE NERVOUS SYSTEM

The current patient (P0000131) presented with:  
Areflexia  
Delayed gross motor development  
Waddling gait

The matched patient (P0000371) presented with:  
Abnormality of peripheral nerve conduction

### UNMATCHED

The current patient (P0000131) presented with:  
Petechiae  
Bruising susceptibility  
Impaired platelet aggregation  
Limited hip extension  
Epistaxis  
Decreased hip abduction

The matched patient (P0000371) presented with:  
Hepatic steatosis  
Elevated serum creatine phosphokinase

## GENE MATCHING BREAKDOWN

### STIM1

Estimated relevance for the observed phenotype in the current patient (P0000131):  1%

HIDE VARIANTS...

Estimated relevance for the observed phenotype in the matched patient (P0000371):  1%

#### VARIANT

#### ESTIMATED HARMFULNESS

chr11: 4045175 - 4045175

■■■■ 80%

A → T  
(NONSYNONYMOUS)

#### VARIANT

#### ESTIMATED HARMFULNESS

chr11: 4045175 - 4045175

■■■■ 80%

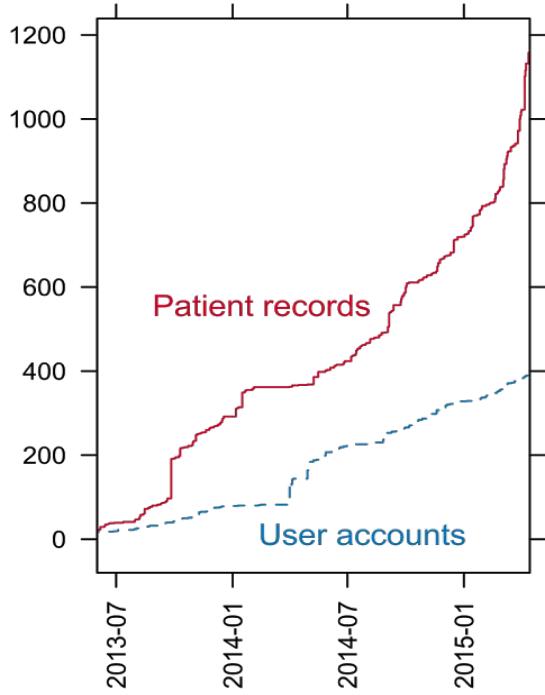
A → T  
(NONSYNONYMOUS)

includes data from:

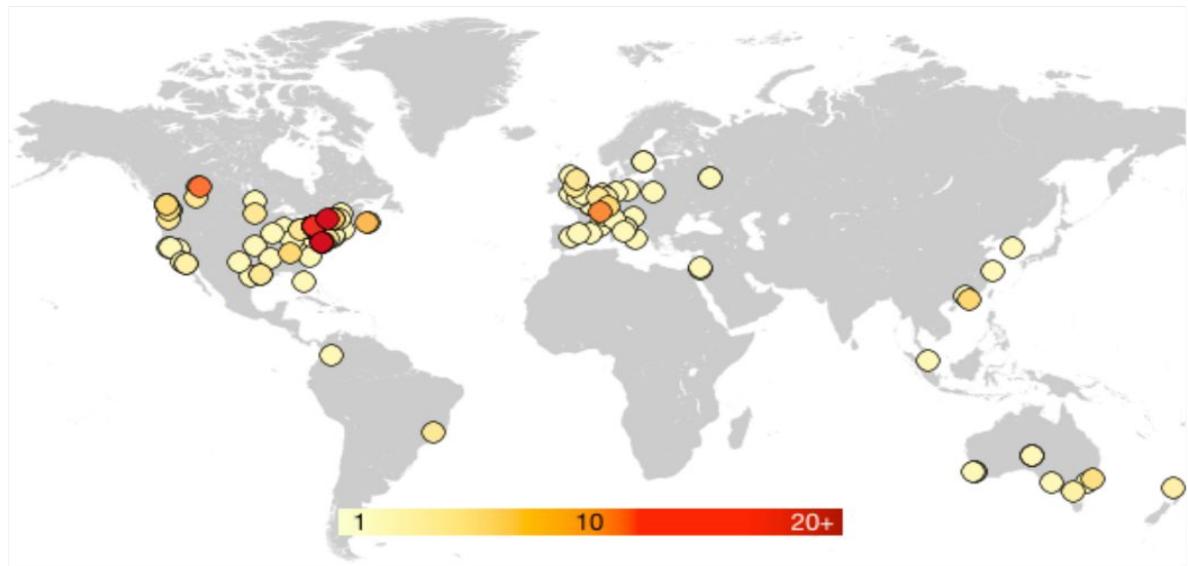


Two similar patients with STIM1 mutations matched despite inconsistent terminology

**A**



**B**



> >2000 cases

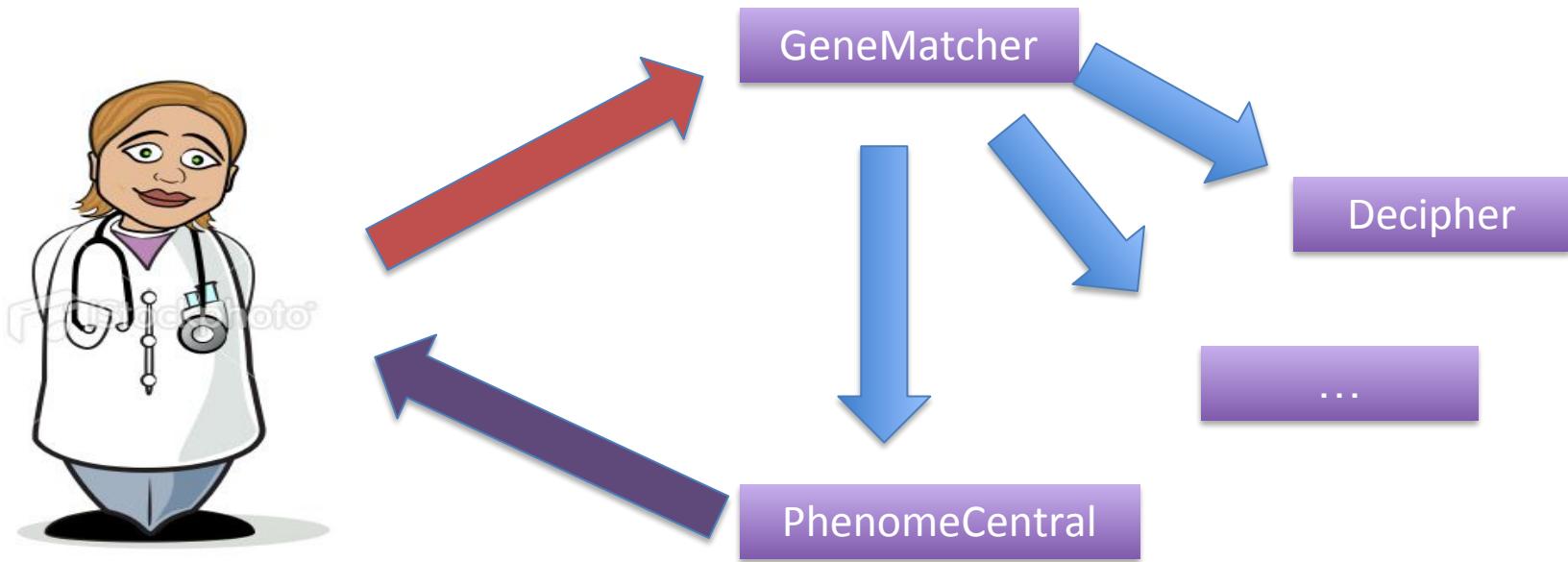
> >500 users

includes data from:

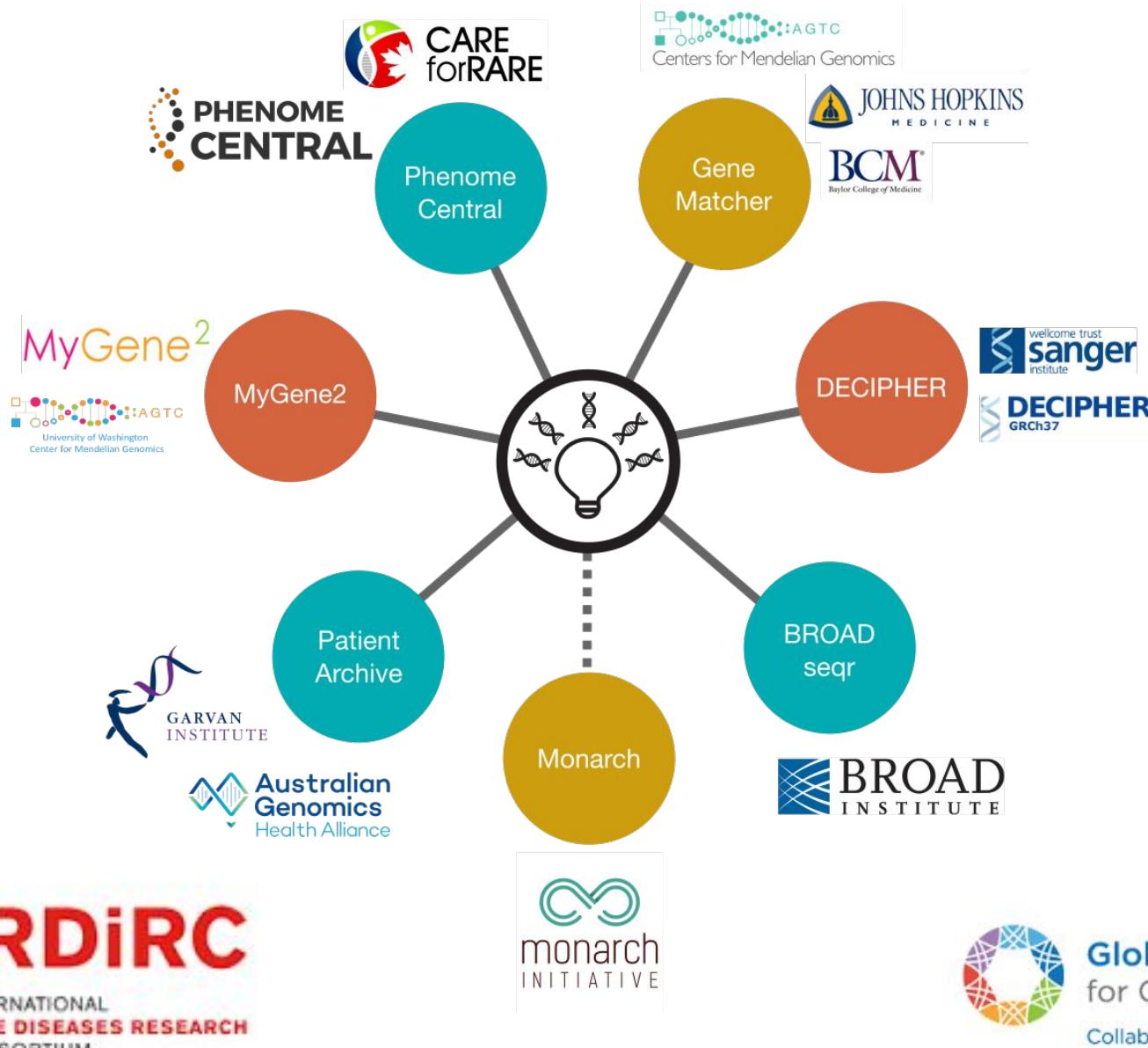


# So Many MatchMakers, So Little Time

- Rare disorder data is collected at many databases (GeneMatcher, Decipher, GEM.app, PhenomeCentral)
- Busy clinicians don't have time to use all of them



# MME Network



# MME Network (in PC)

## Remote server: Decipher Production Server

Anyone who proposes to publish material which uses data obtained from the DECIPHER database agrees to:

Acknowledge the DECIPHER Consortium; and

Contact the coordinator of the centre that entered the data on any individual who they wish to include in their report and offer appropriate agreed recognition of their contribution, which may include co-authorship if the magnitude of the contribution warrants it to at least one representative from the project/participating centre (possibly the member who submitted the patient data). This can be achieved by emailing a request to [decipher@sanger.ac.uk](mailto:decipher@sanger.ac.uk)

Access to bulk data may be obtained from DECIPHER and is subject to a Data Access Agreement, in which the user certifies that no attempt to identify individual patients will be undertaken. The same restrictions apply to the public data displayed on this website: no one is authorized to attempt to identify patients by any means.

The DECIPHER consortium provides these data in good faith as a research tool, but without verifying the accuracy, clinical validity or utility of the data. The DECIPHER consortium, makes no warranty, express or implied, nor assumes any legal liability or responsibility for any purpose for which the data are used.

[REFRESH](#)

Showing 5 similar cases

Remote Case ID	Potential diagnoses	Contact	Local relevance score	Remote relevance score	Details
263271	<i>Undiagnosed</i>	DECIPHER	9%	95%	<a href="#">SHOW PHENOTYPE AND GENOTYPE SIMILARITY...</a>

## PHENOTYPIC FEATURES BREAKDOWN

### ABNORMALITY OF THE PHILTRUM

The current patient (F0000010) presented with:  
Short philtrum



The matched patient (263271) presented with:  
Long philtrum



MED12

[SHOW VARIANTS...](#)

### ABNORMALITY OF THE THUMB

The current patient (F0000010) presented with:  
Broad thumb



The matched patient (263271) presented with:  
Short thumb

### LOW-SET EARS

The current patient (F0000010) presented with:



The matched patient (263271) presented with:

# PhenoTips.org PhenomeCentral.org

Development Team:

**Marta Gîrdea, Orion Buske, Sergiu Dumitriu, Felicia Collura, Veronika Koltunova, Jonathan Zung, Andriy Misyura, Anton Kats, Bailey Gallinger**

Consortia:

- CARE for RARE Canada  
**Kym Boycott, Taila Hartley, Sarah Sawyer, Chandree Beauleiu**
- NIH-UDP  
**David Adams, William Gahl, Neal Boerkoel, William Bone**
- Undiagnosed Diseases Network  
**Rachel Badovinac, Isaac Kohane, William Gahl**
- RD-Connect /Neuromics  
**Hanns Lochmuller, Rachel Thompson**

HPO, Exomiser, Monarch:

**Peter Robinson, Melissa Haendel, Damian Smedley, Sebastian Kohler, Nicole Washington**

Funding:

Genome Canada  
(CARE for RARE),  
CIHR, NSERC,  
NIH...)

# RDs are so cool... How can I help?

- › Phenotate – a platform to crowd-source RD knowledge from med students!
- › Get a “test” on rare diseases, where you annotate diseases with phenotypes
- › Your answers are used to supplement databases such as the HPO.



<http://phenotate.org>   [brudno@cs.toronto.edu](mailto:brudno@cs.toronto.edu)

# Want to join us?



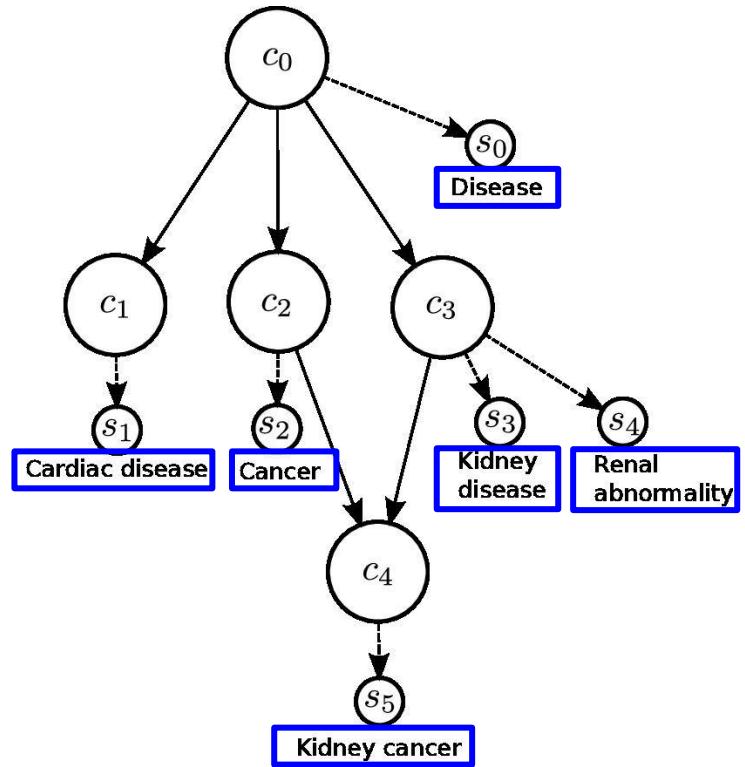
Looking for talented Grad Students, Postdocs and  
to work on computational analysis of clinical data  
[brudno@cs.toronto.edu](mailto:brudno@cs.toronto.edu)



**SickKids®**

# Training details

- > Stochastic Gradient is used for training
- > The following parameters are trained:
  - The RNN parameters (embedder)
  - The HPO concept embeddings
- > The training is repeated for 20 epochs
- > In each epoch data is divided in several batches
- > Training data includes all the HPO phrases (synonyms)



# How to recognize concepts in a sentence?

## > Method 1 - Window based approach

Embed every window of 1 to 7 words and find the closest HPO vector

*There was marked variability in the severity and nature of manifestations with 2 having severe hand and foot involvement in addition to craniofacial changes compatible with a diagnosis of Freeman-Sheldon syndrome.*

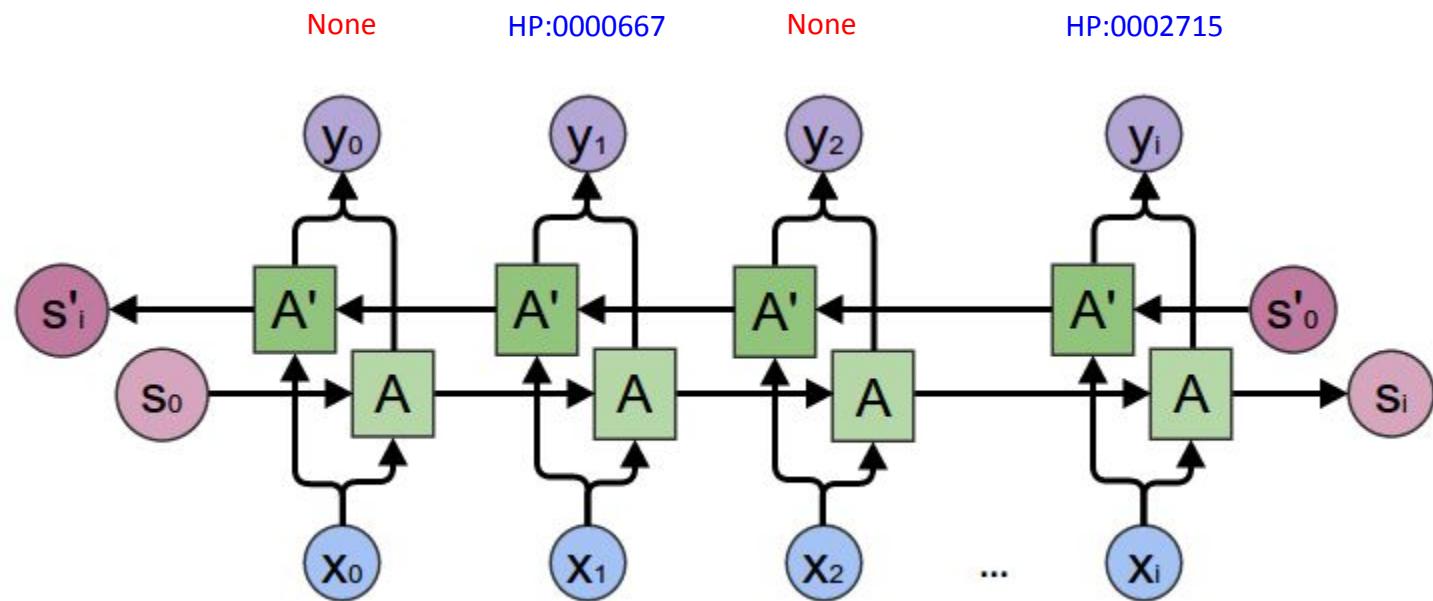
**Vector**

**Find the closest HPO vector**

HP:00001	$v_0$
18	$v_1$
HP:00027	
15	$v_n$
...	...
HP:00006	
67	

# Recurrent Neural Networks (RNN)

- > Method 2 - Use a bidirectional RNN on whole sentence  
For each HPO concept, check if exists any close RNN outputs



Another trick; add non-phenotype phrases

- › When the method is only trained on HPO, it thinks the world only includes phenotypes!
- › During the training, we can present some examples that are not phenotypes
- › It would be better if these examples are in similar context
- › We used a subset of terms in UBERON in some experiments: