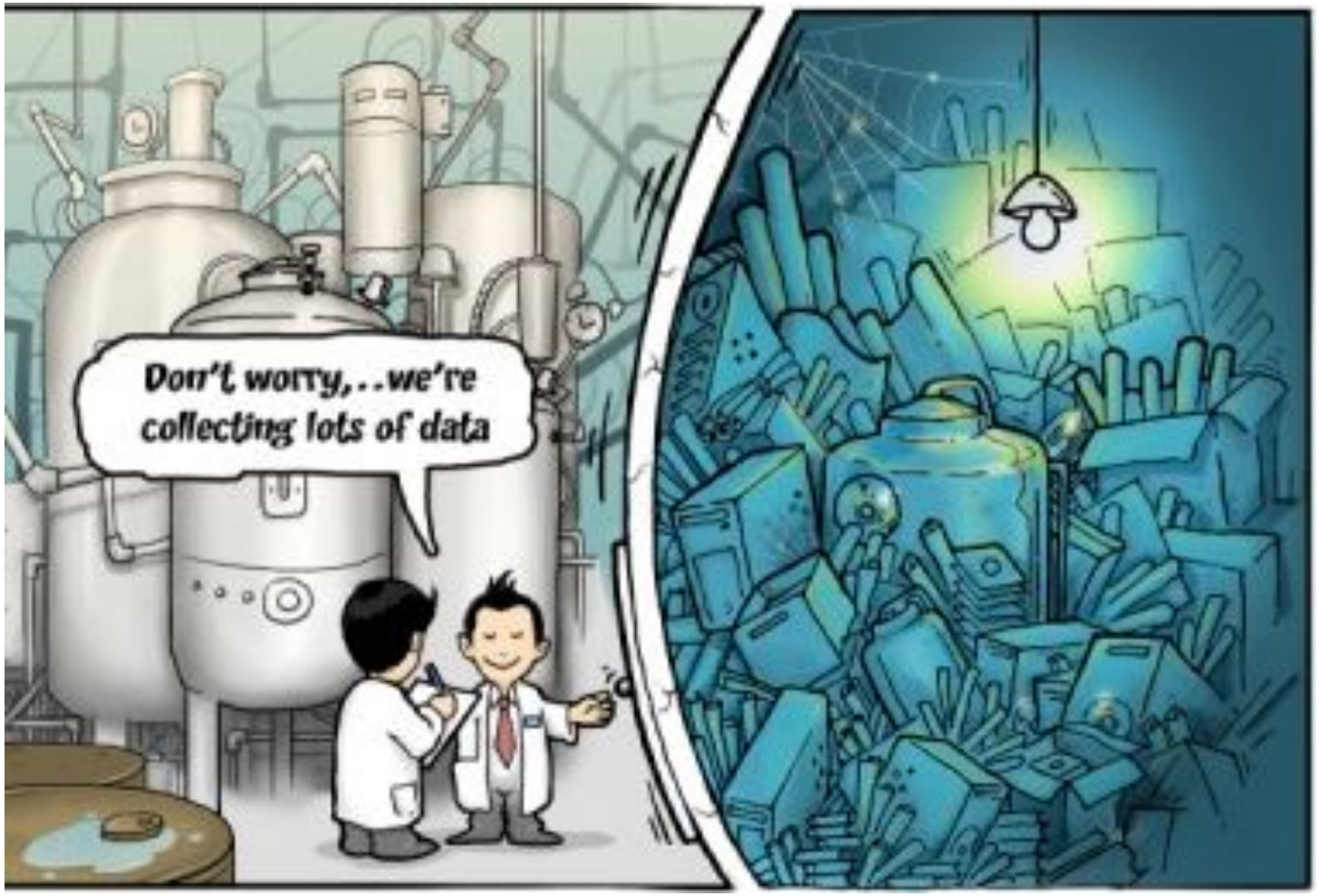
Capturing and Sharing Data for Rare Disease: Building Algorithms, Software, and a Community

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

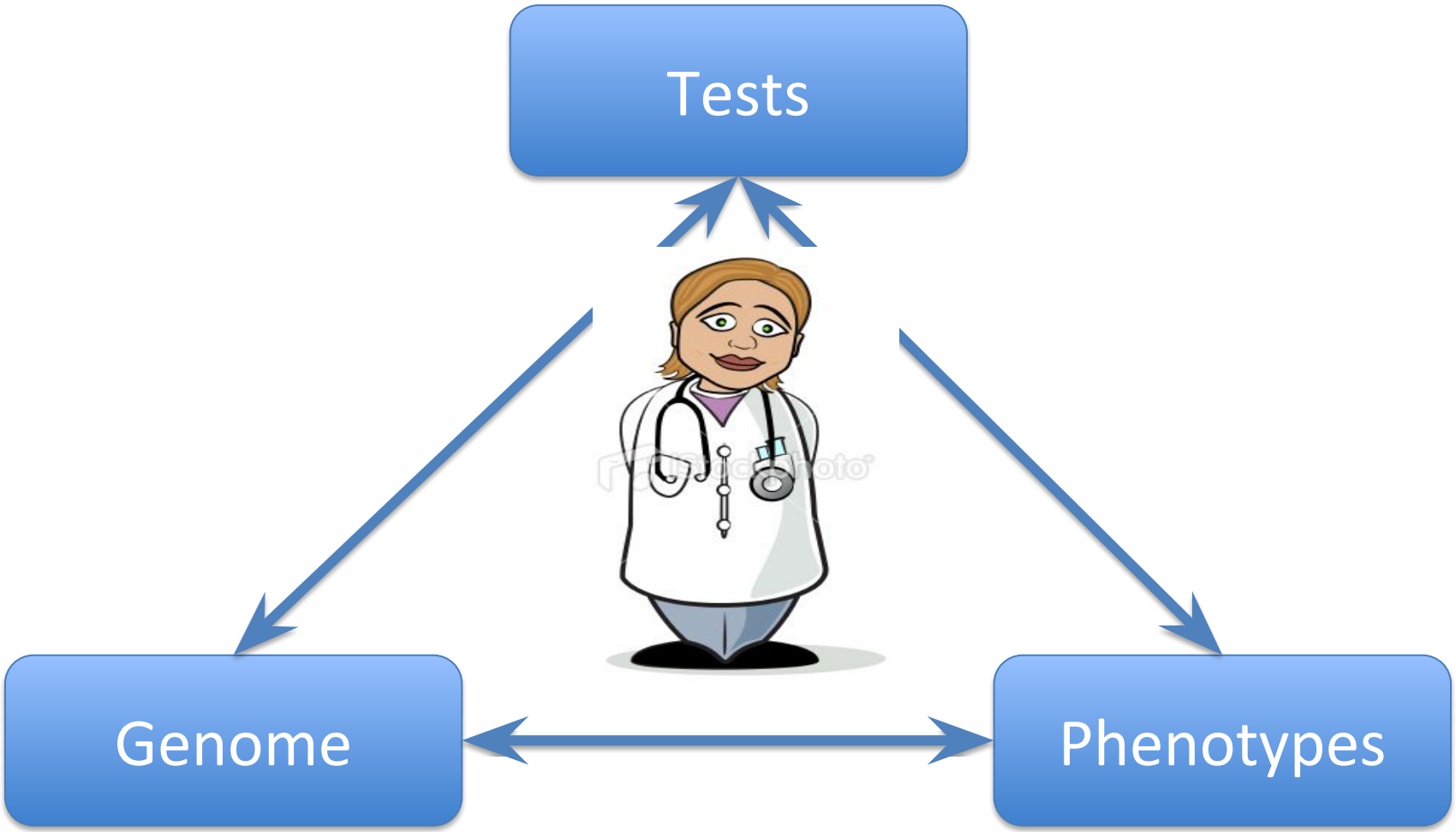
CS/Medicine, 27 February 2018

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
- congential anomaly
- congentital anomaly
- cong. m.
- cong. Mal
- cong. malfor
- congenital malform
- congenital m.
- multiple congenital anomalies
- multiple congenital abormalities
- 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
  - ☐ Mild
  - ☐ Moderate
  - ☐ Severe
- ☐ Attention deficit hyperactivity disorder
- ☐ Autism
- ☐ Pervasive developmental delay
- ☐ Psychiatric disorders (Specify below)
- ☐ Other: _____

Neurological
- ☐ Hypotonia
- ☐ Seizures
- ☐ Ataxia
- ☐ Dystonia
- ☐ Chorea

Cardiac
- ☐ ASD
- ☐ VSD
- ☐ AV canal defect
- ☐ Coarctation of aorta
- ☐ Tetralogy of fallot
- ☐ Other: _____

Craniofacial
- ☐ Craniosynostosis
- ☐ Cleft lip ☐ Cleft palate
- ☐ Microretrognathia ☐ Retrognathia
- ☐ Facial dysmorphism (Specify below)
- ☐ Other: _____

Eye Defects
- ☐ Blindness
- ☐ Coloboma
- ☐ Epicanthus
- ☐ Eyelid abnormality (Specify bellow)
- ☐ Other: _____

Examples of lists:
* dd. cong. malfor. behav. pro.
* dd. mental retardation
* df< delayed puberty
* df&lt
* dd of mr
* mental retar.short stature

Problems with the status quo

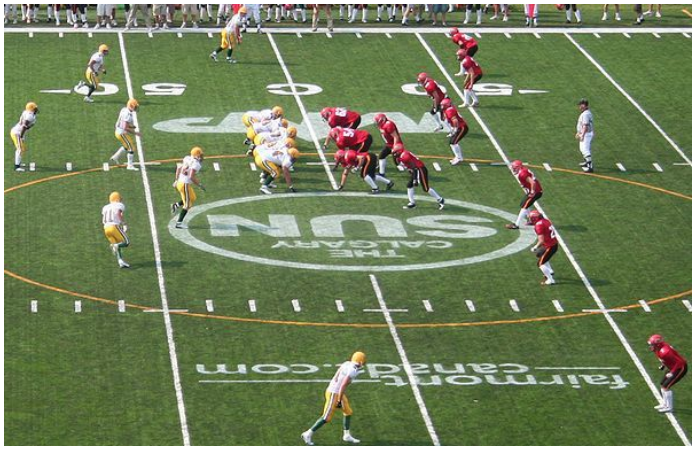
- Phenotypic descriptions that are very evocative for humans, unreadable to a computer:
  - “first words at 5 years”
  - “has trouble spelling”
  - “recognizes only close relatives”

- 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:



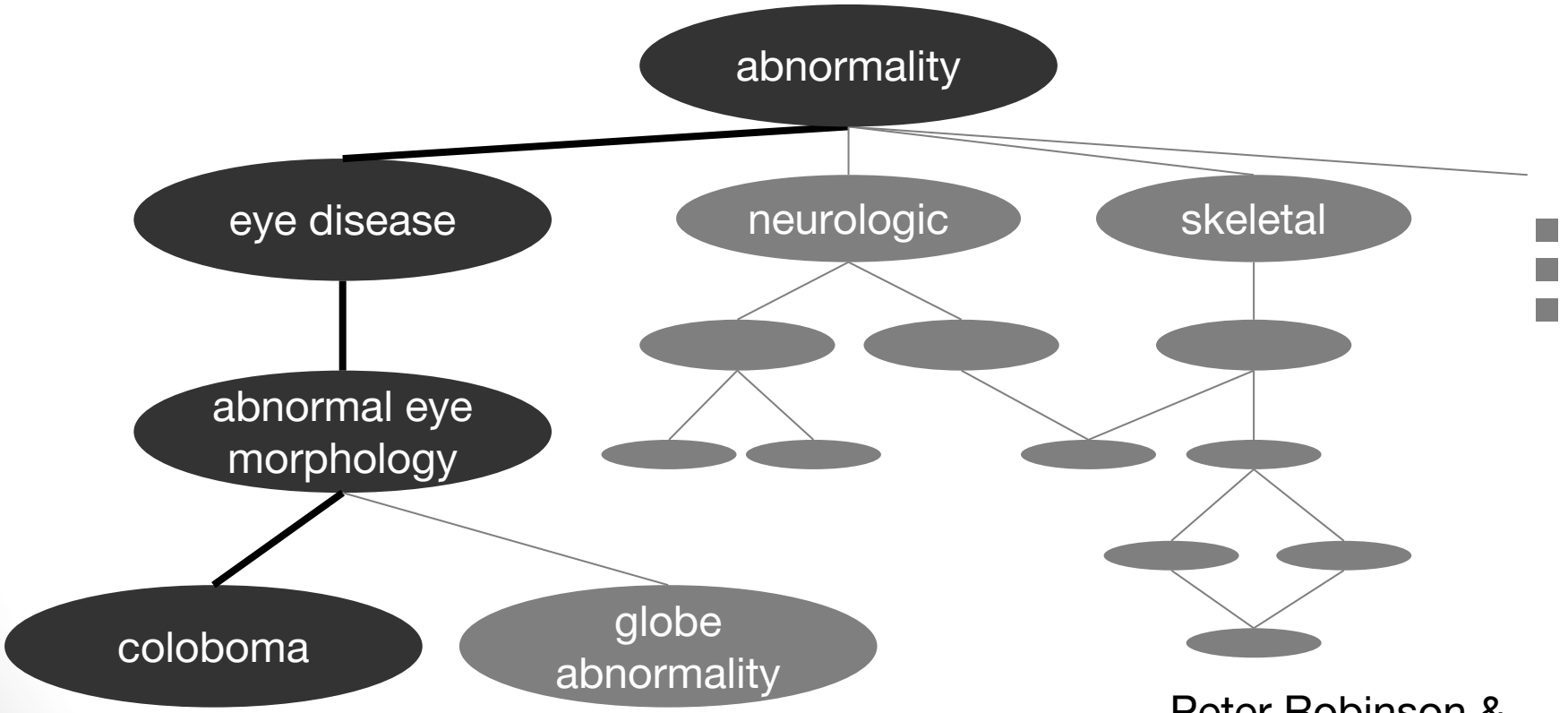
Ontologies are terms with relationships

- 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):

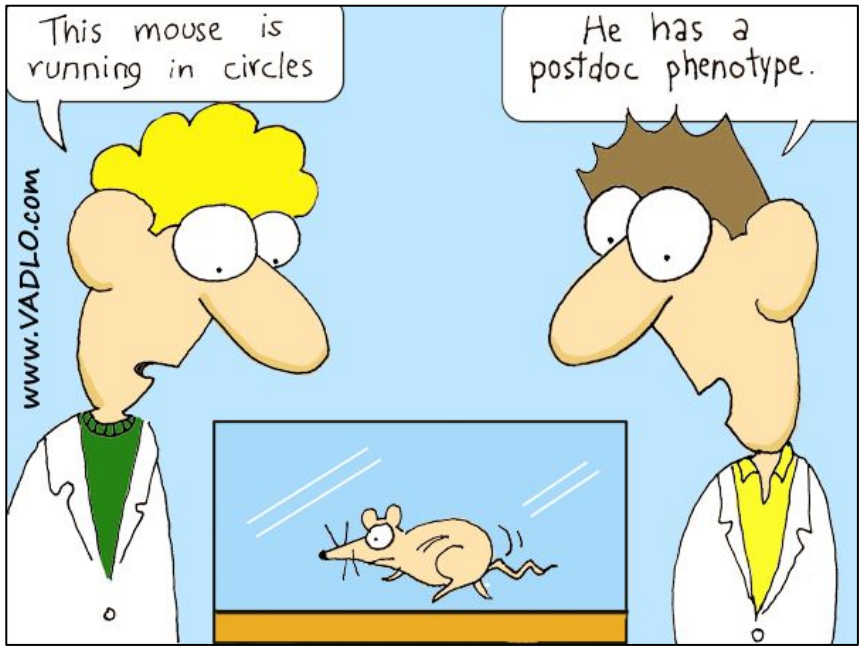
- 11,000+ terms
- 100,000+ links to 5,000+ OMIM/ORDO Disorders



Peter Robinson &
Monarch Consortium

“Model Organism” Ontologies

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



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

Talk Overview

Text Mining

Clinical Data Capture
(PhenoTips)

Structured Data Sharing
(PhenomeCentral & MME)

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

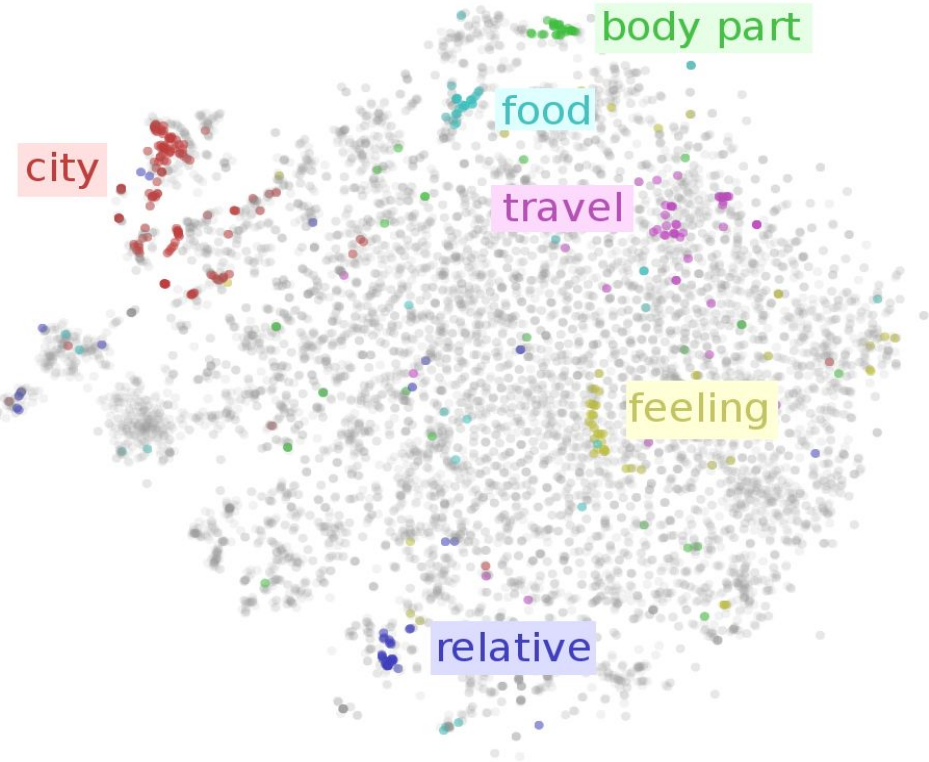
Dysmorphic facies

Distinctive facies

Distortion of face

Word embeddings

- Each word is mappe a vector in high-dimensional sp
- The vector representations are learned by training o large corpora
- Words that occur in similar contexts tend to have closer vectors



Word embedding - biomedical text

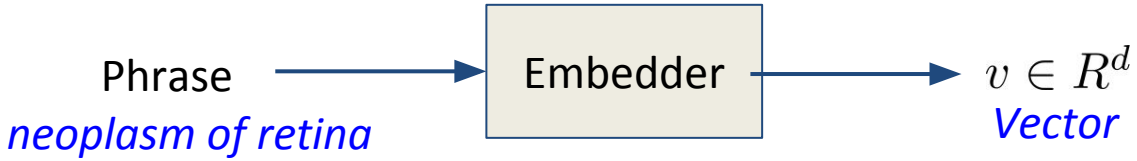
- Used GloVe (Pennington et al.) to learn word embeddings by training on publicly available articles in PubMedCentral (~ 1m documents)

- 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	electrographic	introns	end	MSH2
optic	admitted	epileptiform	untranslated	extended	HNPCC

Intuitions on method

- Learn a function that embeds phrases into vectors:



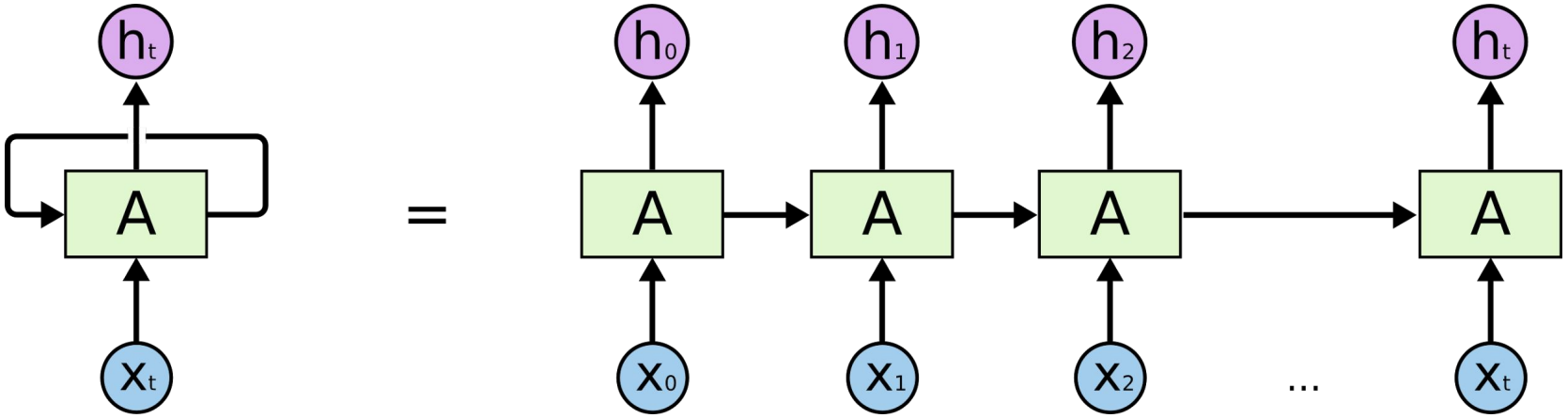
- Learn embeddings for all HPO concepts

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

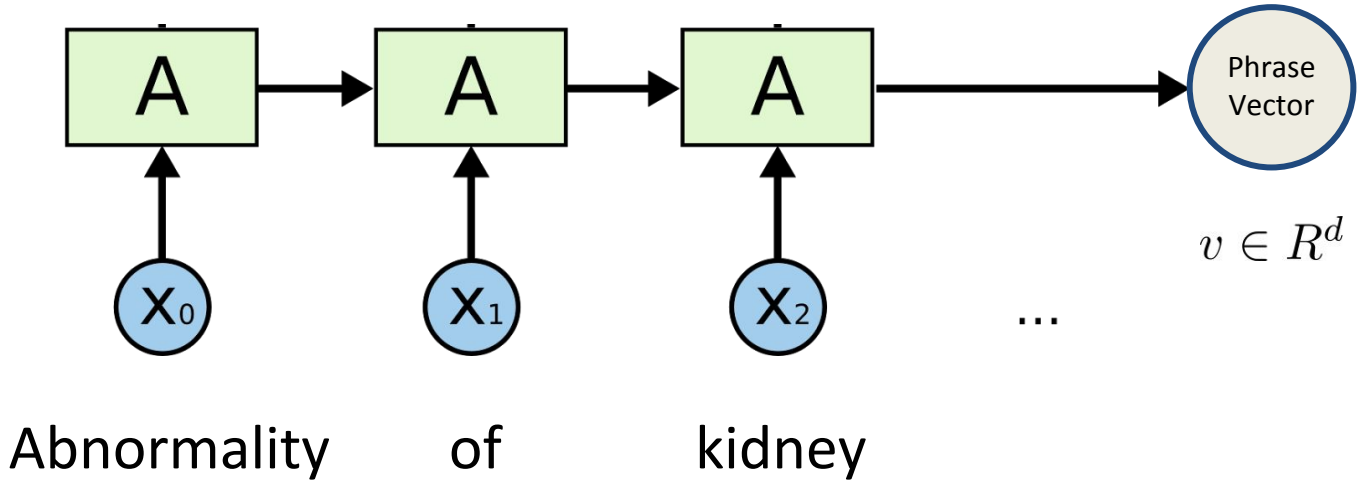
- Goal: **Phrases** "close" in HPO have **close vectors**

How do we embed phrases?

> Recurrent Neural Networks (RNN)



Embedding phrases with RNN

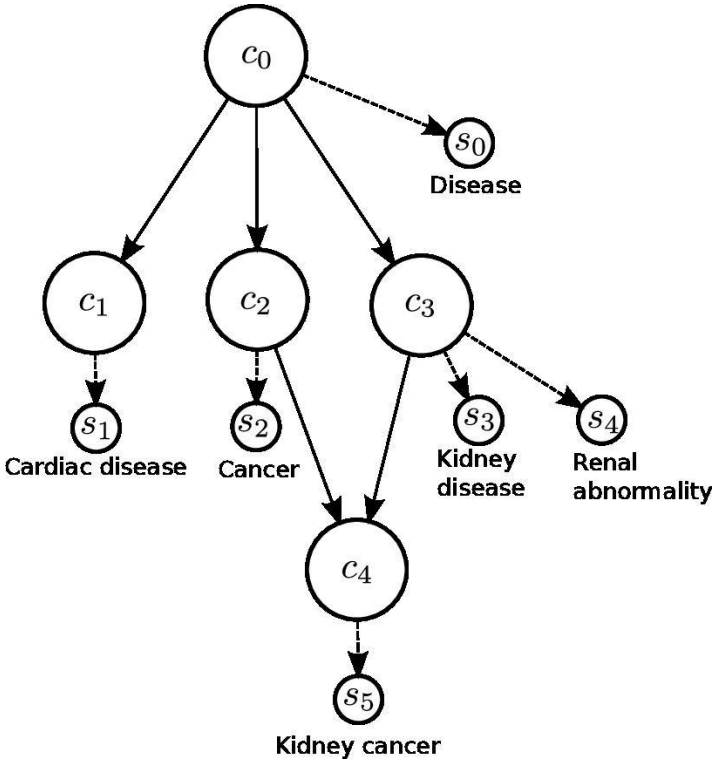


How to use the ontology information?

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

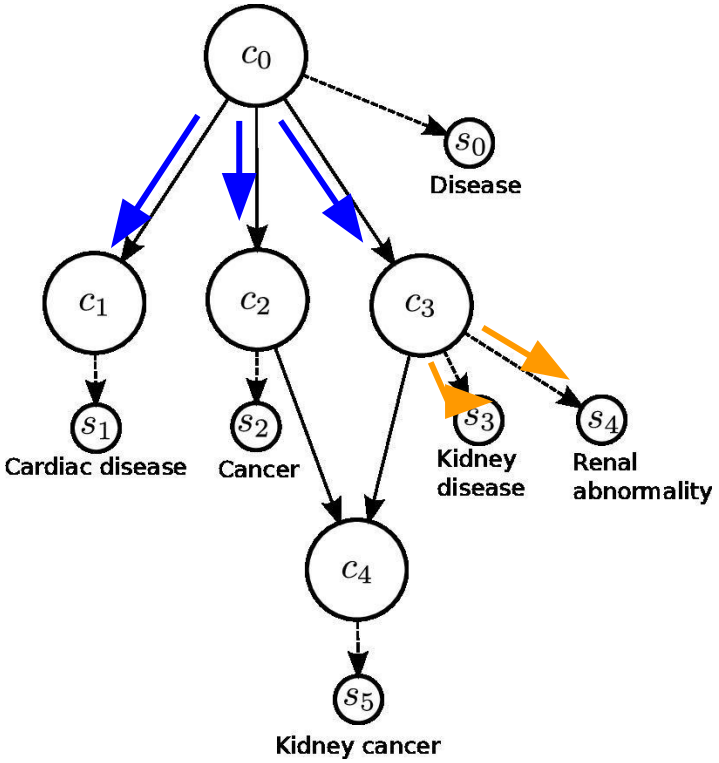
$$x \preceq y \text{ if and only if } \bigwedge_{i=1}^{N} x_i \geq y_i$$

- 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:

	Micro (%)			Macro (%)			Extended (%)			
Method	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Jaccard (%)
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	**82.5**	61.9	70.7	71.0
NCR	**79.5**	**62.1**	**69.7**	**74.2**	**62.2**	**67.7**	81.6	69.0	**74.8**	**75.9**
NCR -H	69.9	60.5	64.9	64.4	**62.2**	63.3	71.3	**70.5**	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.org

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

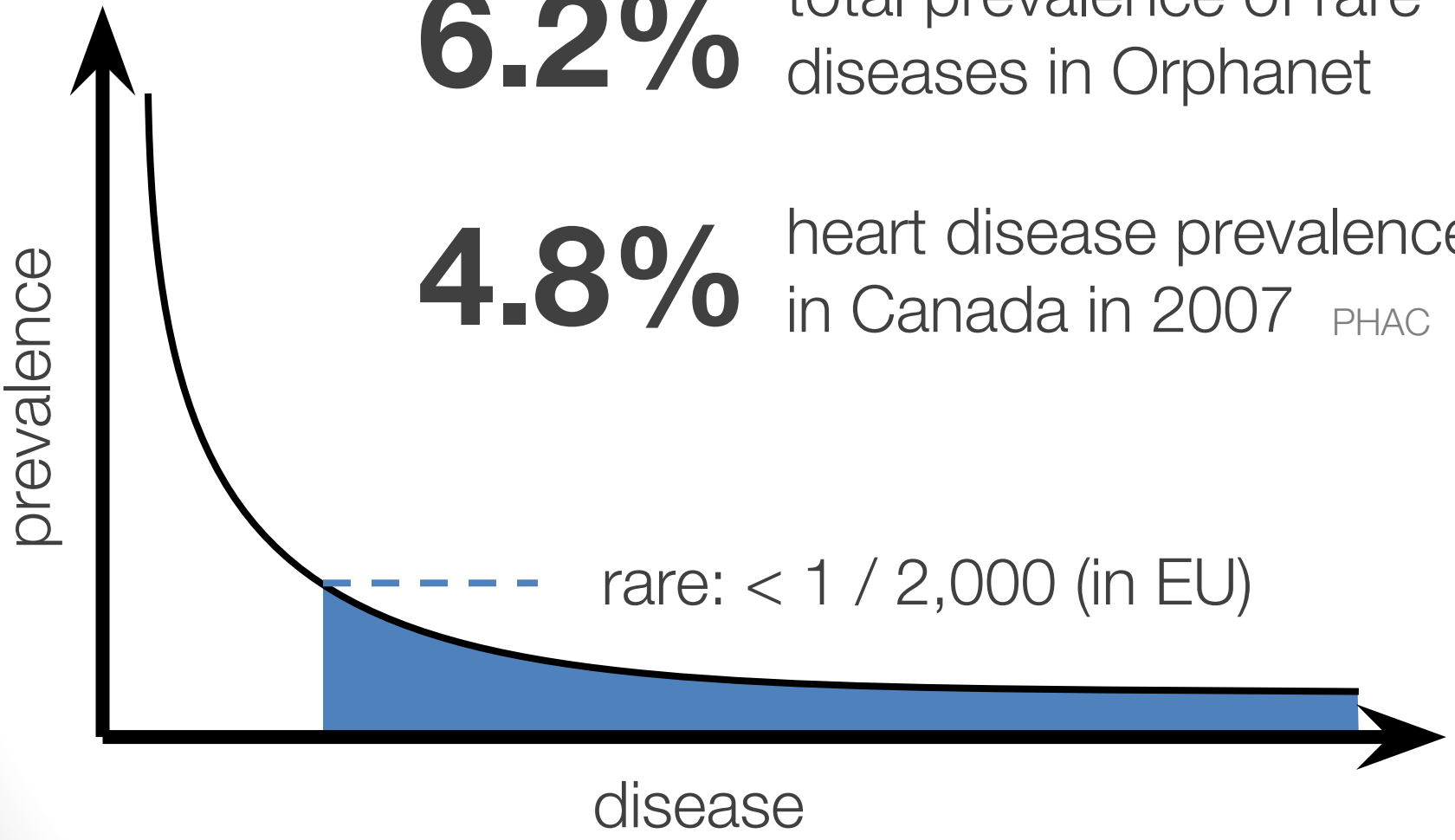
for rare genetic disorders

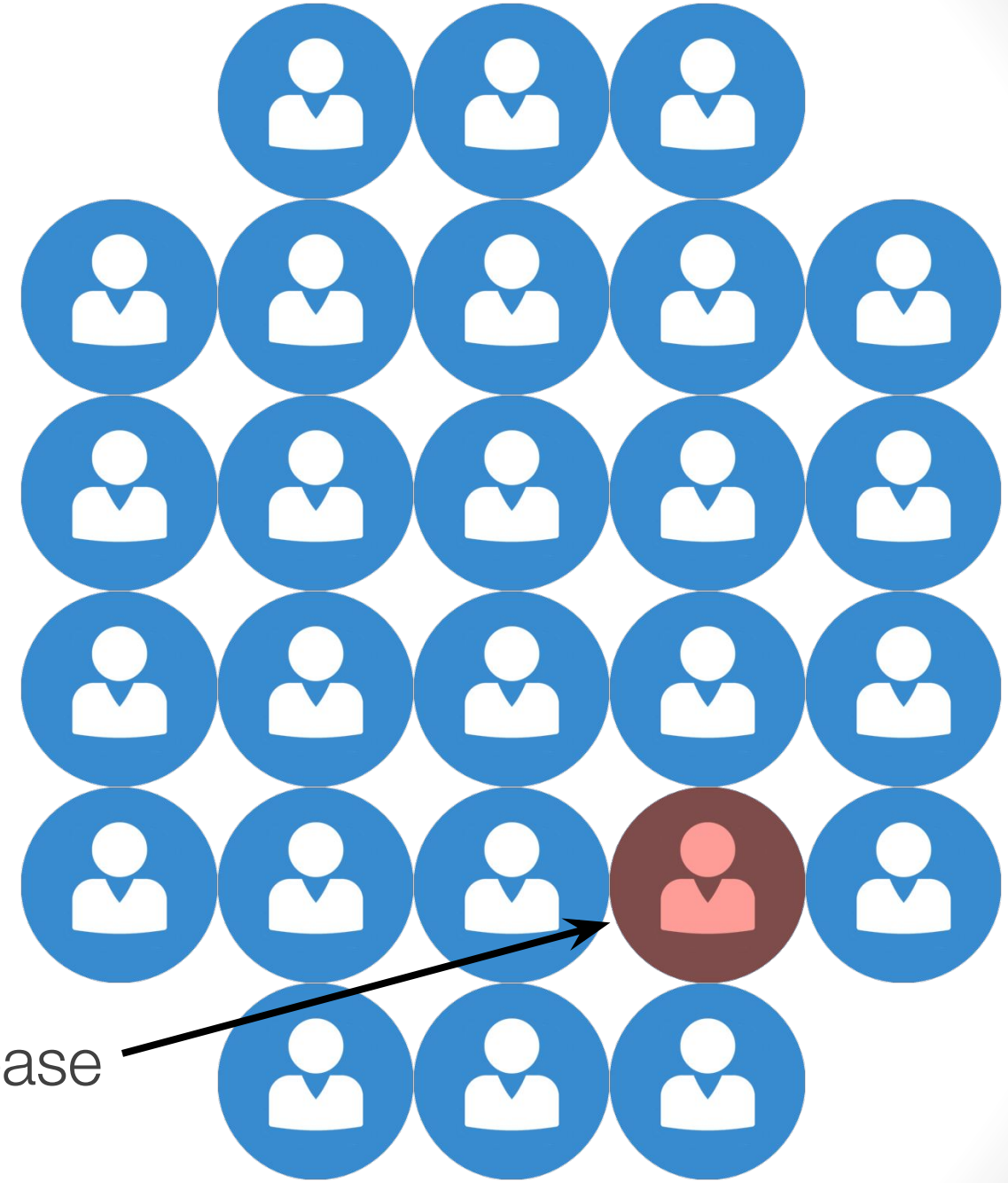
Orion Buske

Rare diseases and the long tail

6.2% total prevalence of rare diseases in Orphanet

4.8% heart disease prevalence in Canada in 2007 PHAC

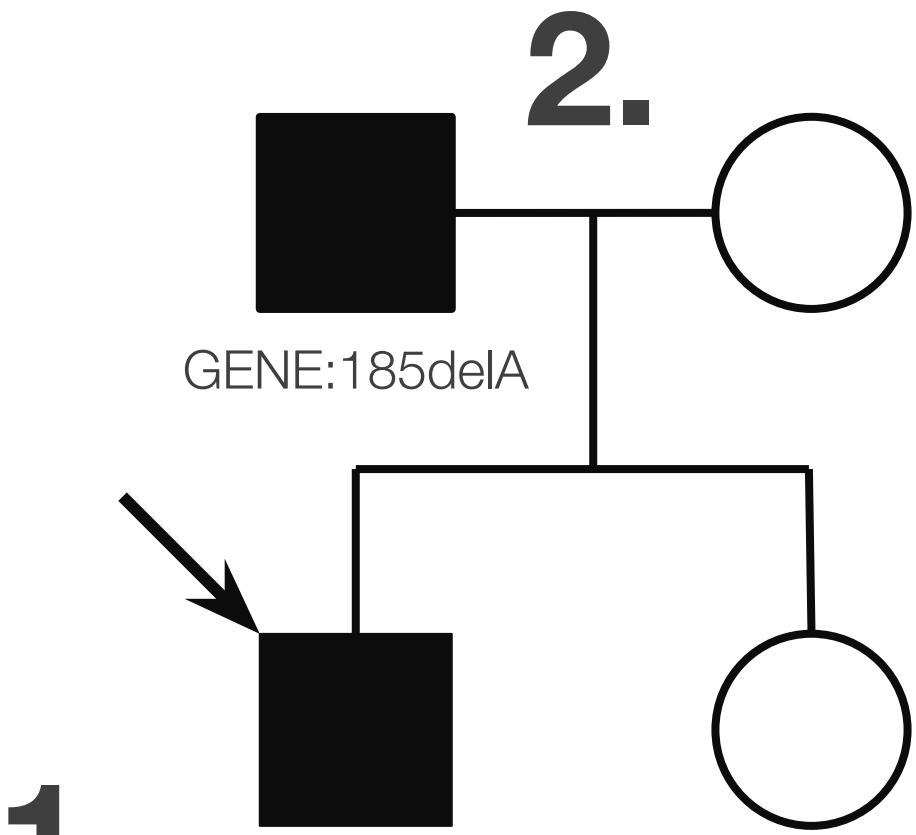


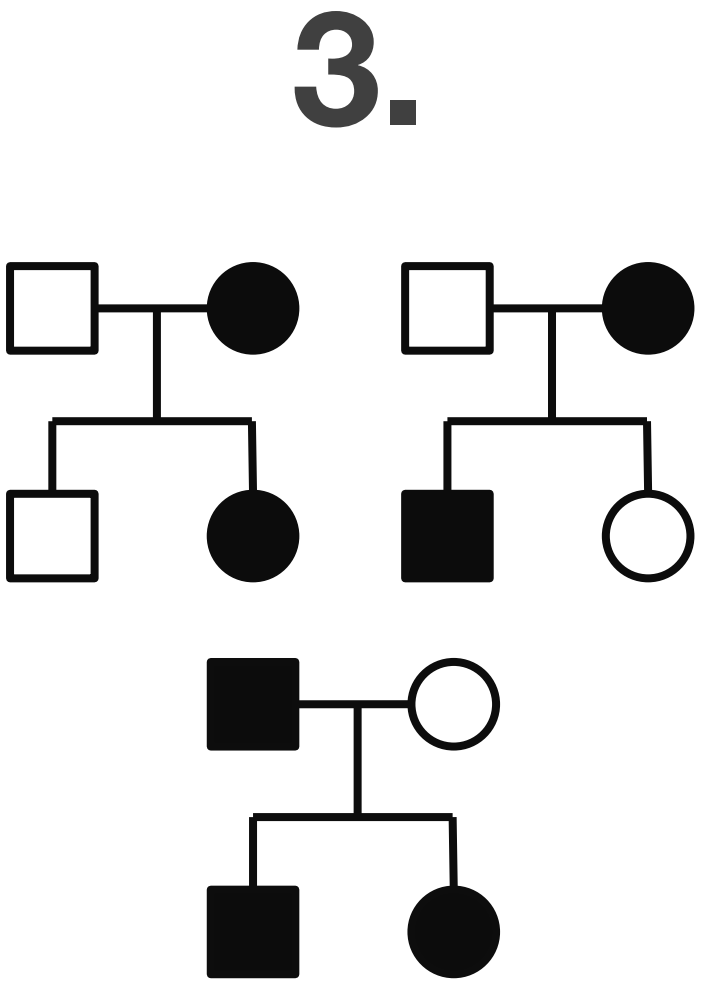


rare genetic disease

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

Finding the cause of a genetic disease...

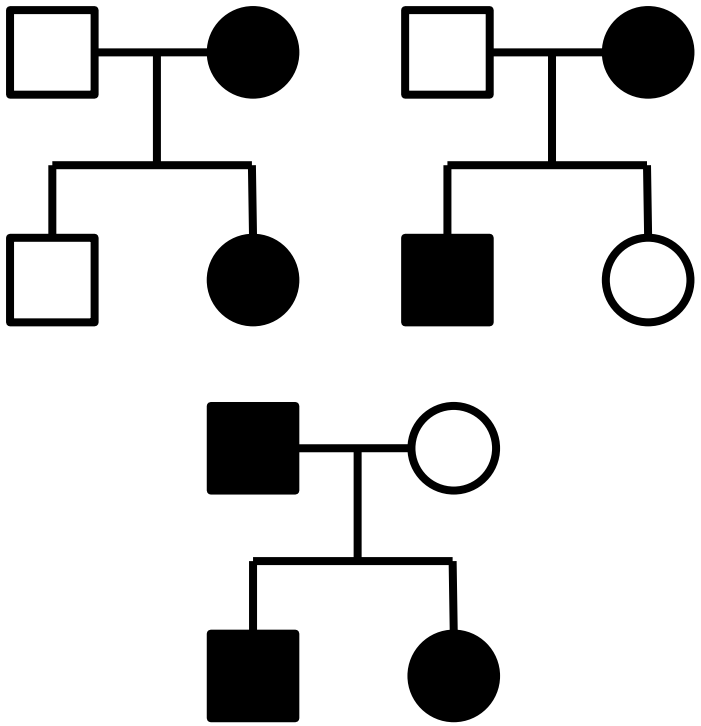


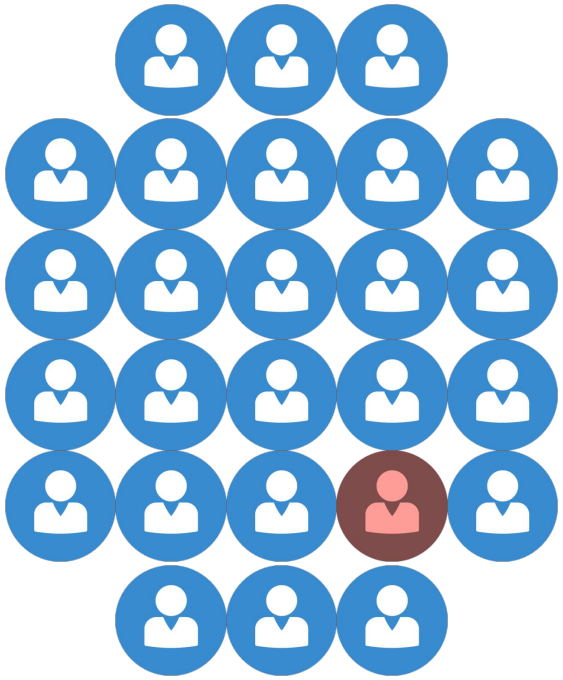


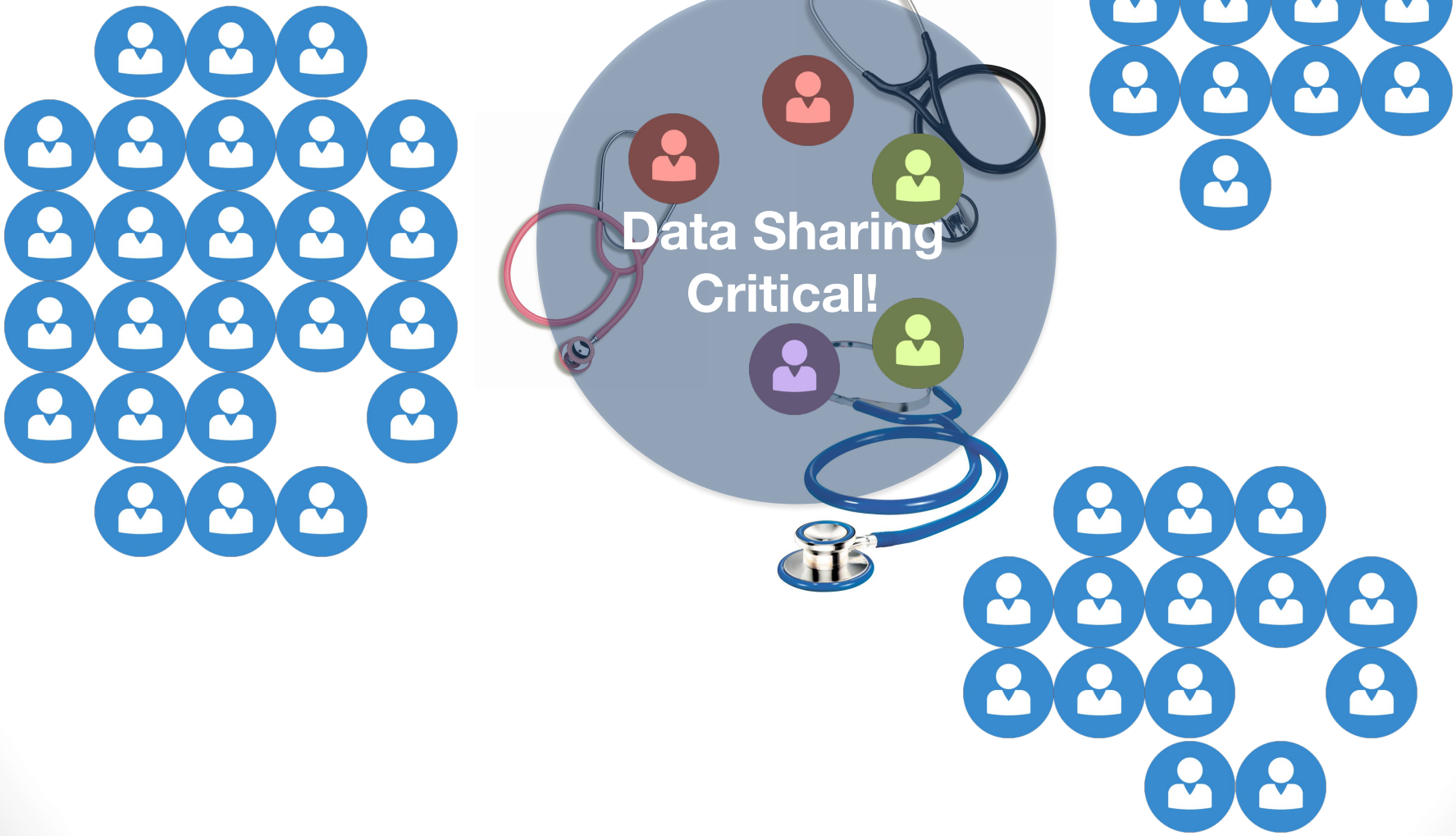
Finding additional families can be difficult...

3.

rare diseases are rare.



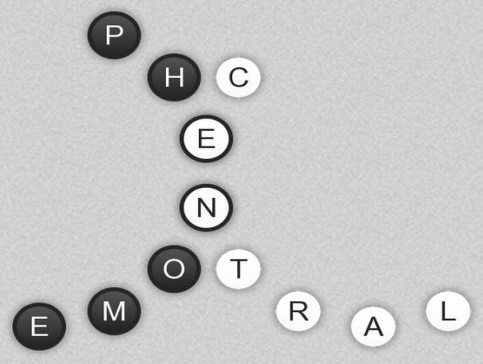



Data Sharing
Critical!

to encourage data sharing we built a

user-friendly, privacy-aware portal

for discovering patients similar to yours



PhenomeCentral

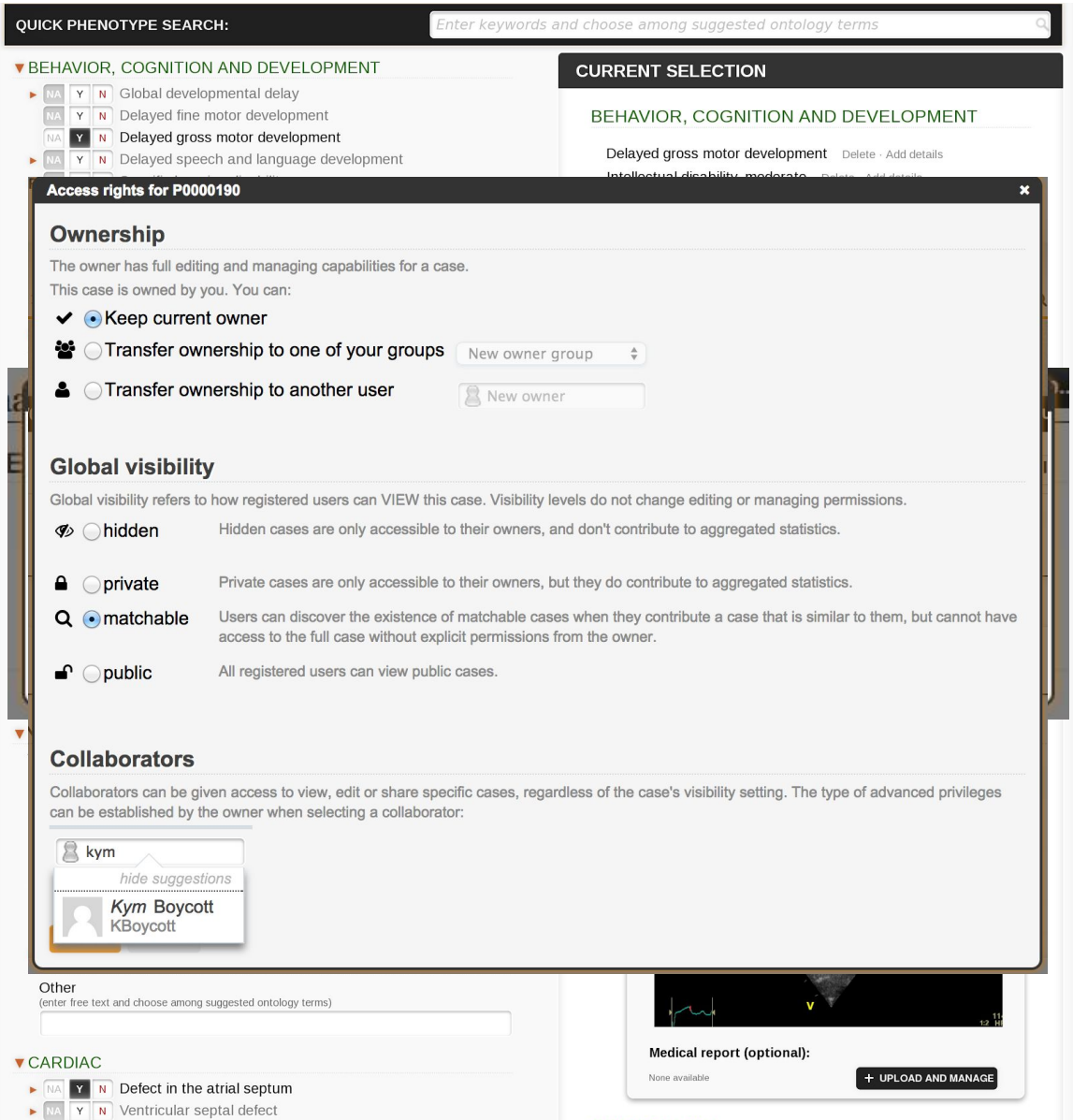
A hub for secure data sharing within the rare disorder community

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

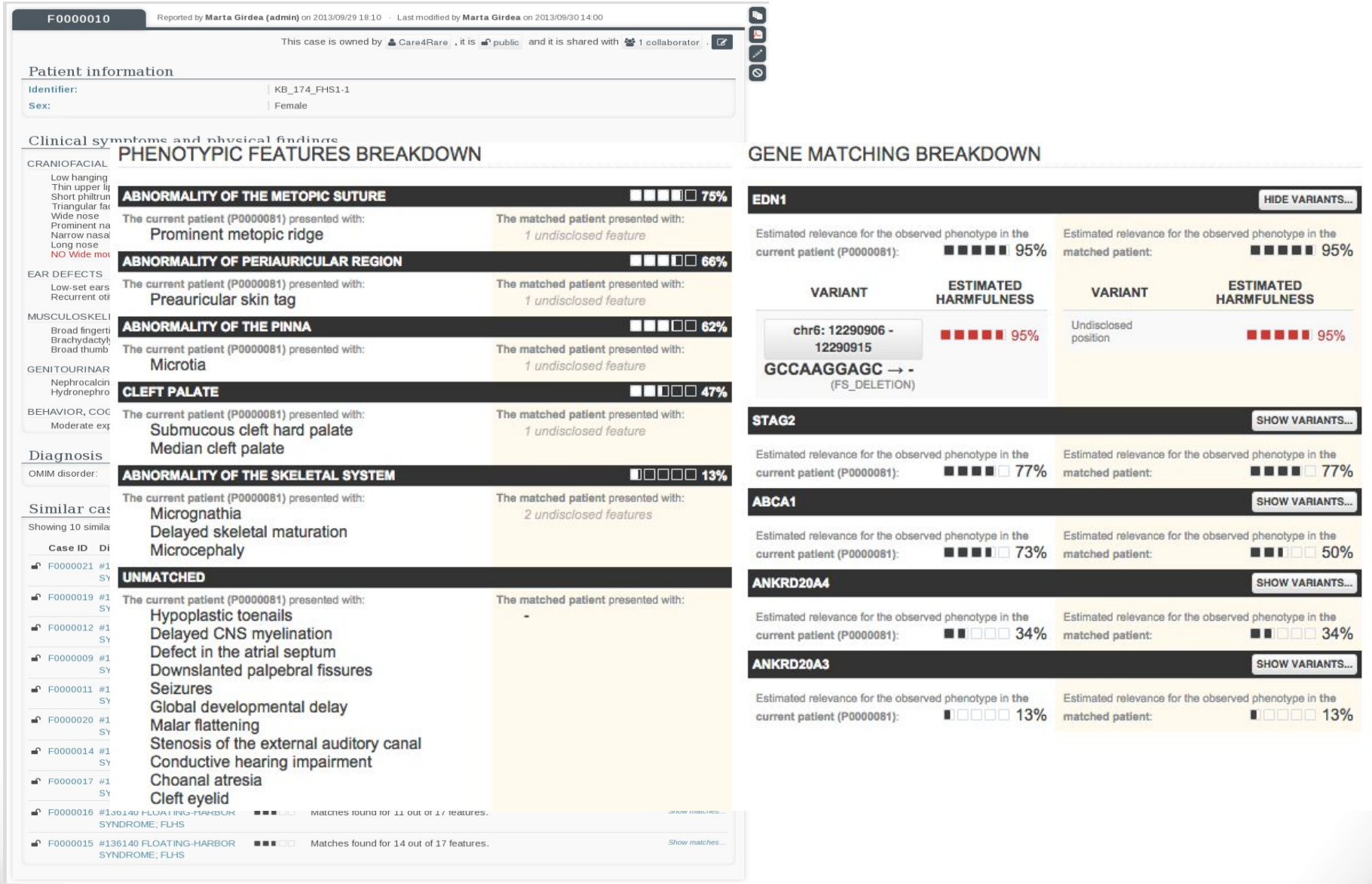
phenomecentral.org

Step 1: submit your patient

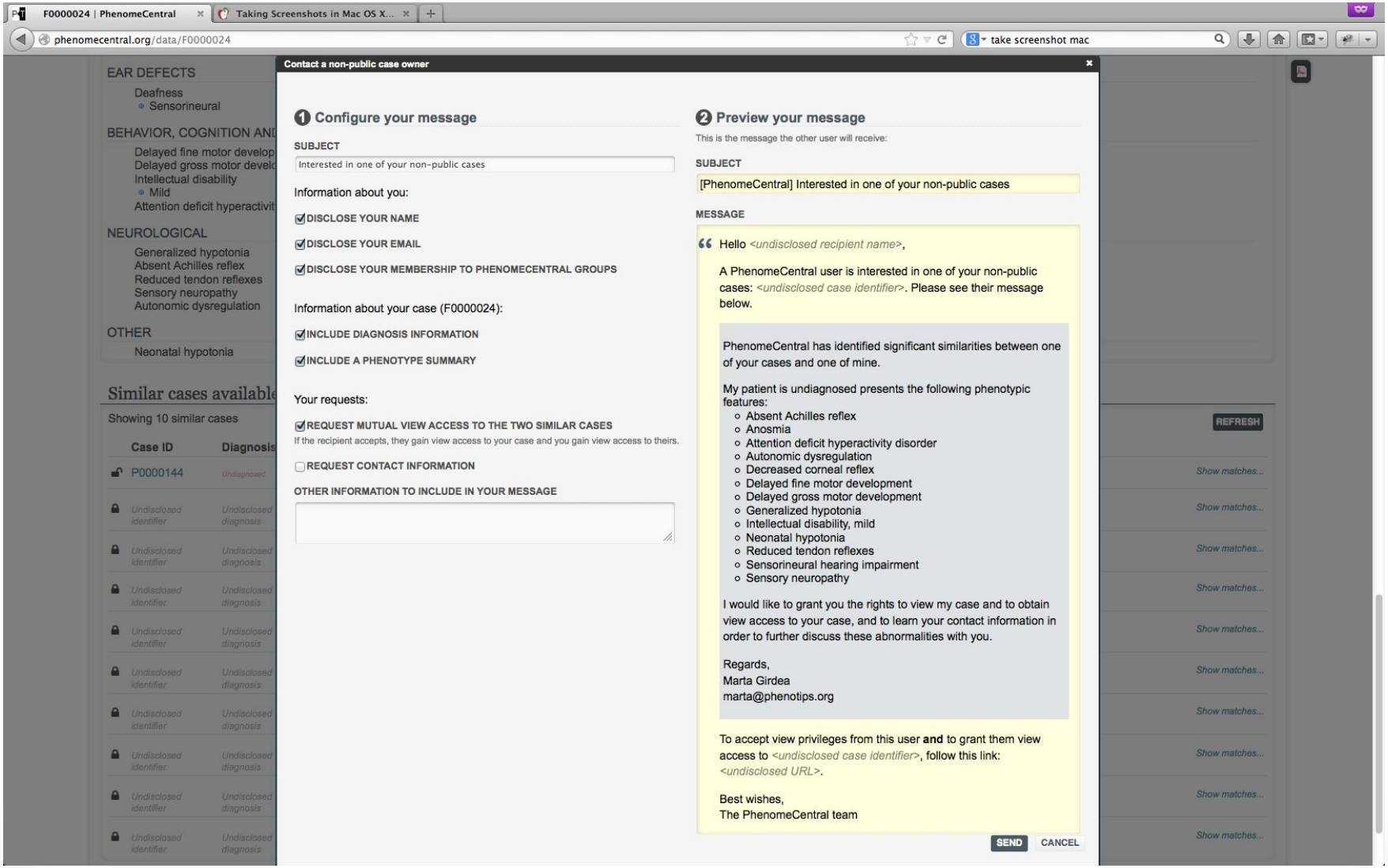


- Built-in PhenoTips interface
- Export anonymized records from other instances
- Add a VCF file and/or gene list
- Set permission and add collaborators

Step 2: see patients similar to yours

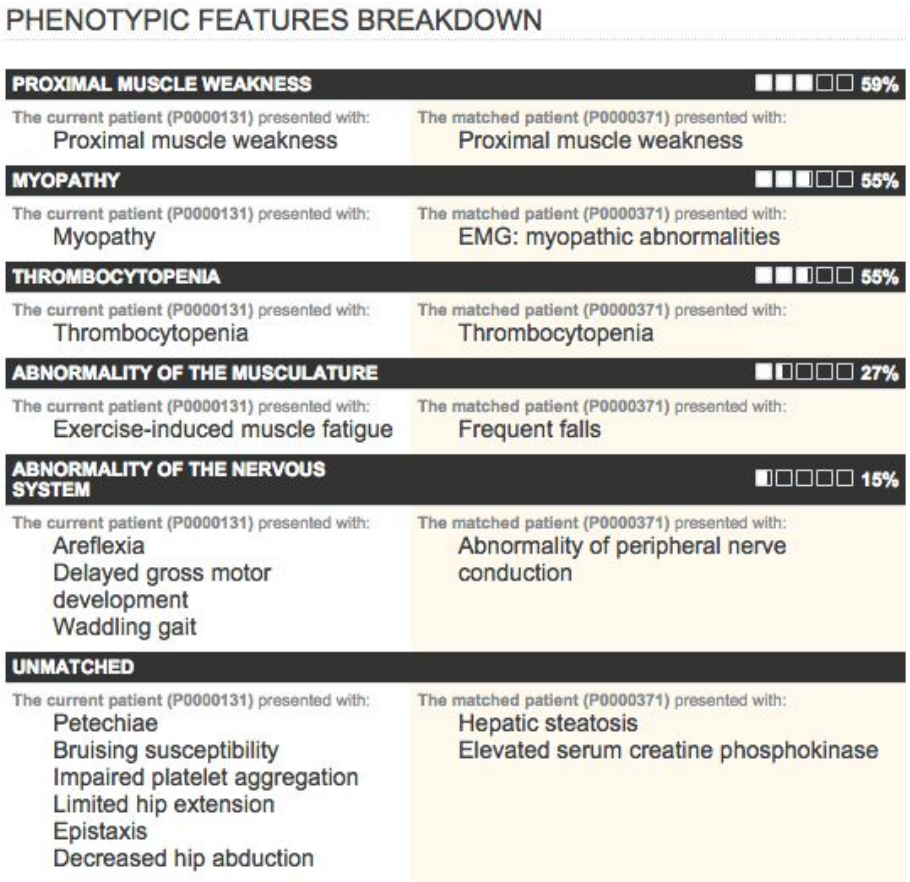


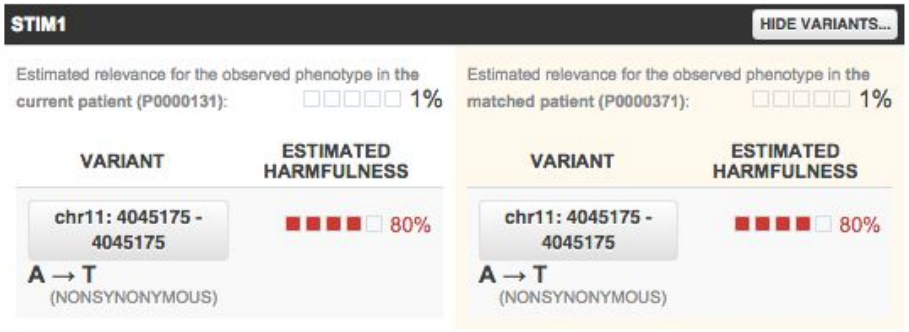
Step 3: contact the other submitter



PhenomeCentral

phenomecentral.org





Two similar patients with STIM1 mutations matched despite inconsistent terminology

includes data from:

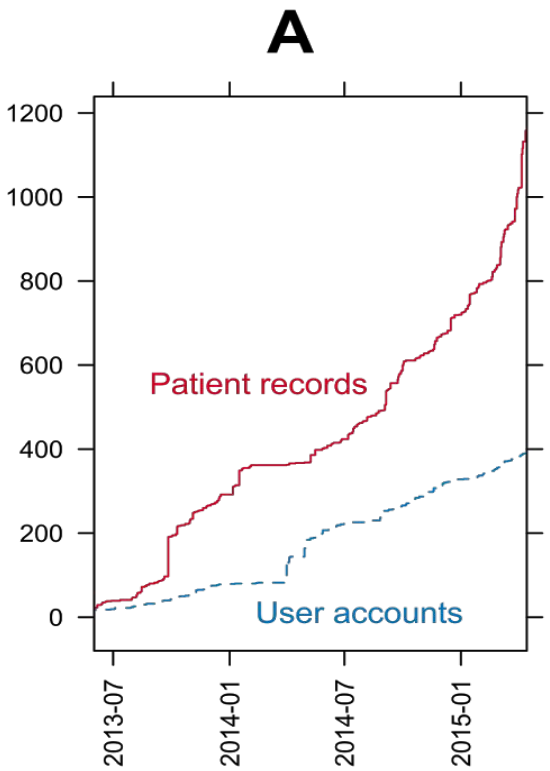
CARE forRARE

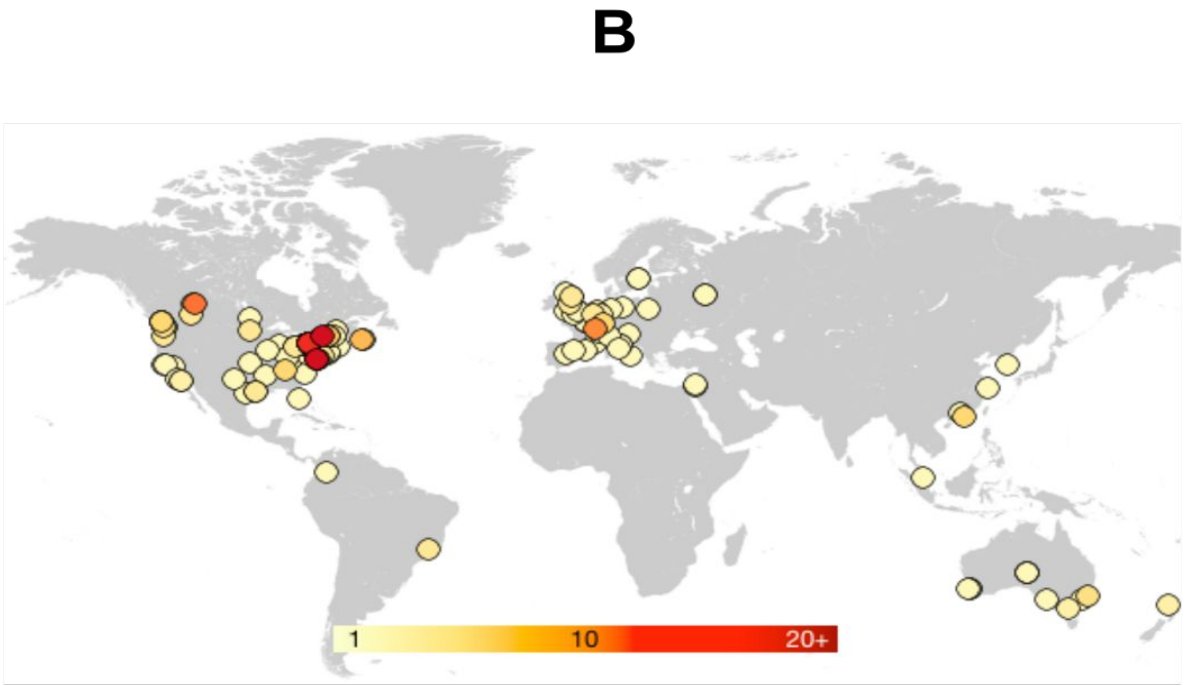
RD Connect

NIH Undiagnosed Diseases Program

PhenomeCentral

phenomecentral.org





- **>2000** cases
- **>500** users

includes data from:

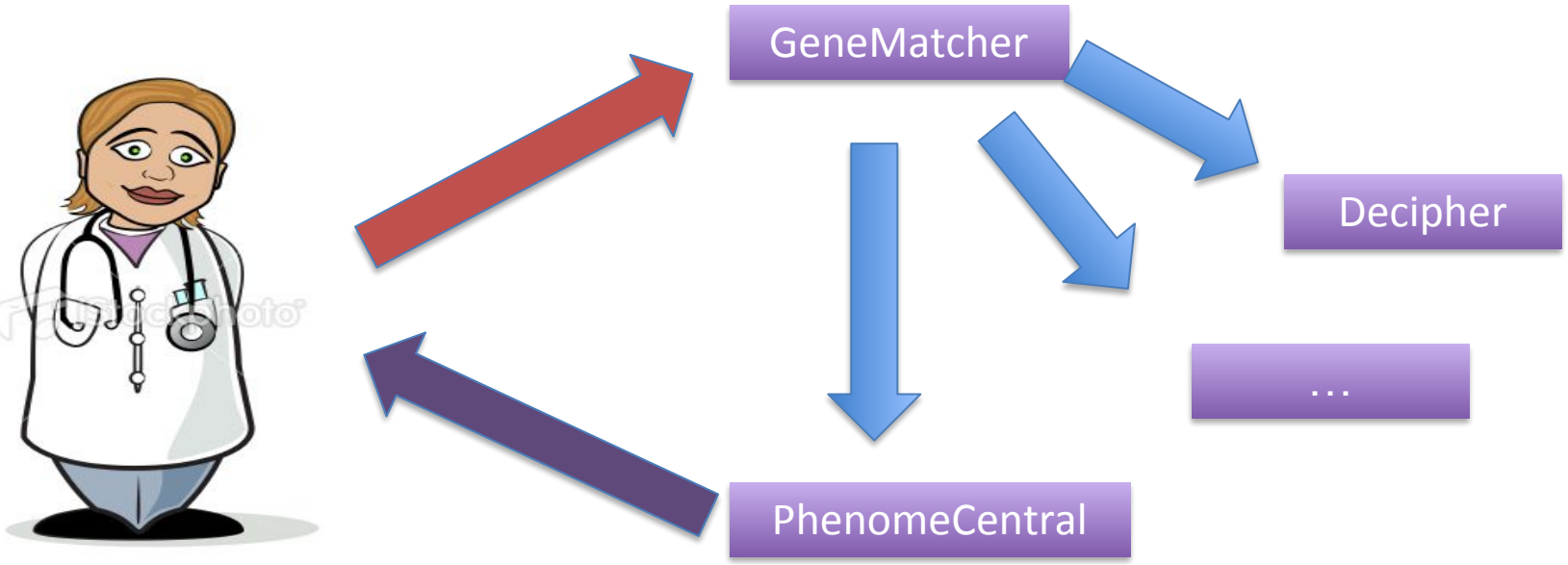
CARE forRARE

RD Connect

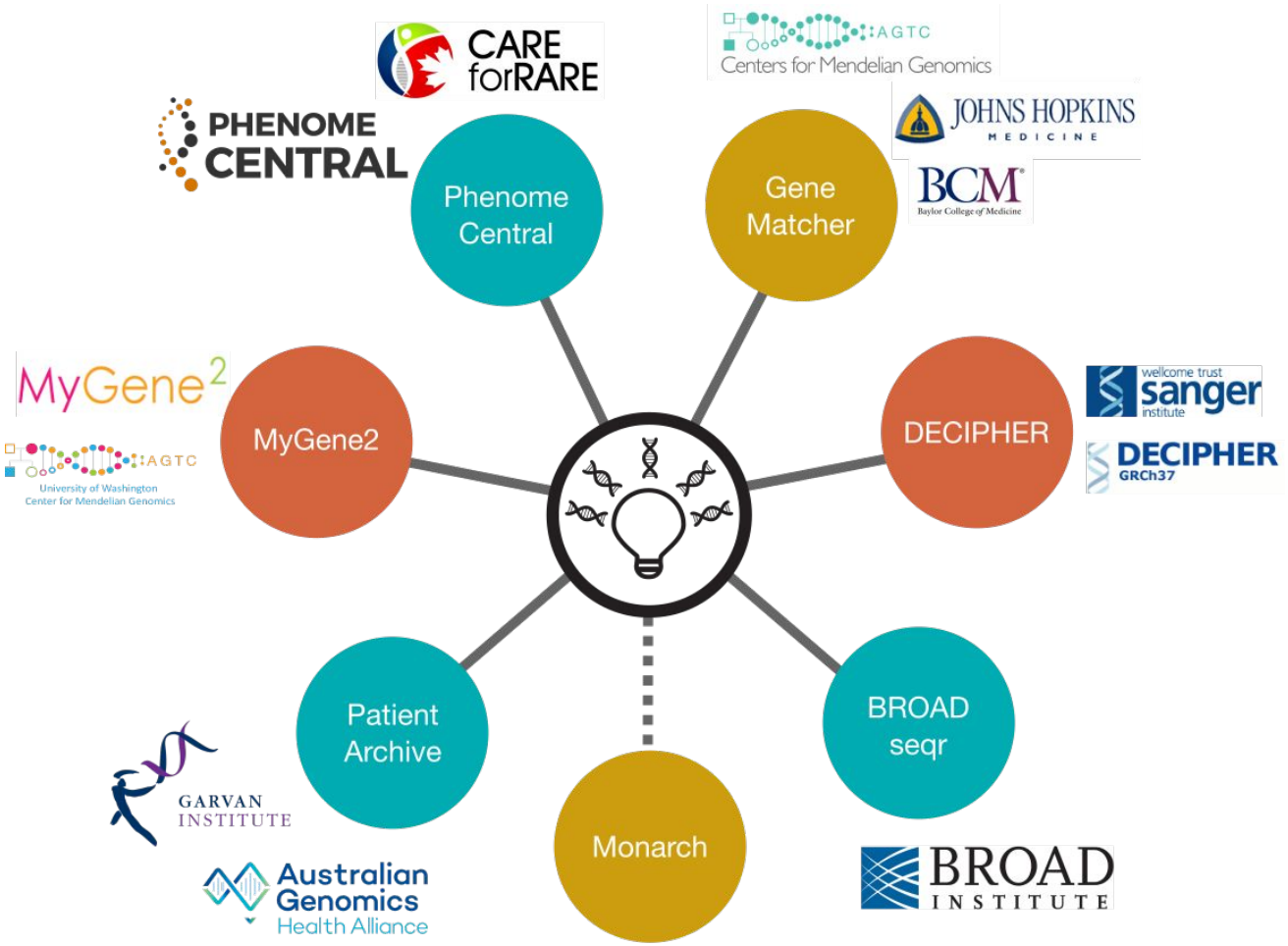
NIH Undiagnosed Diseases Program

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



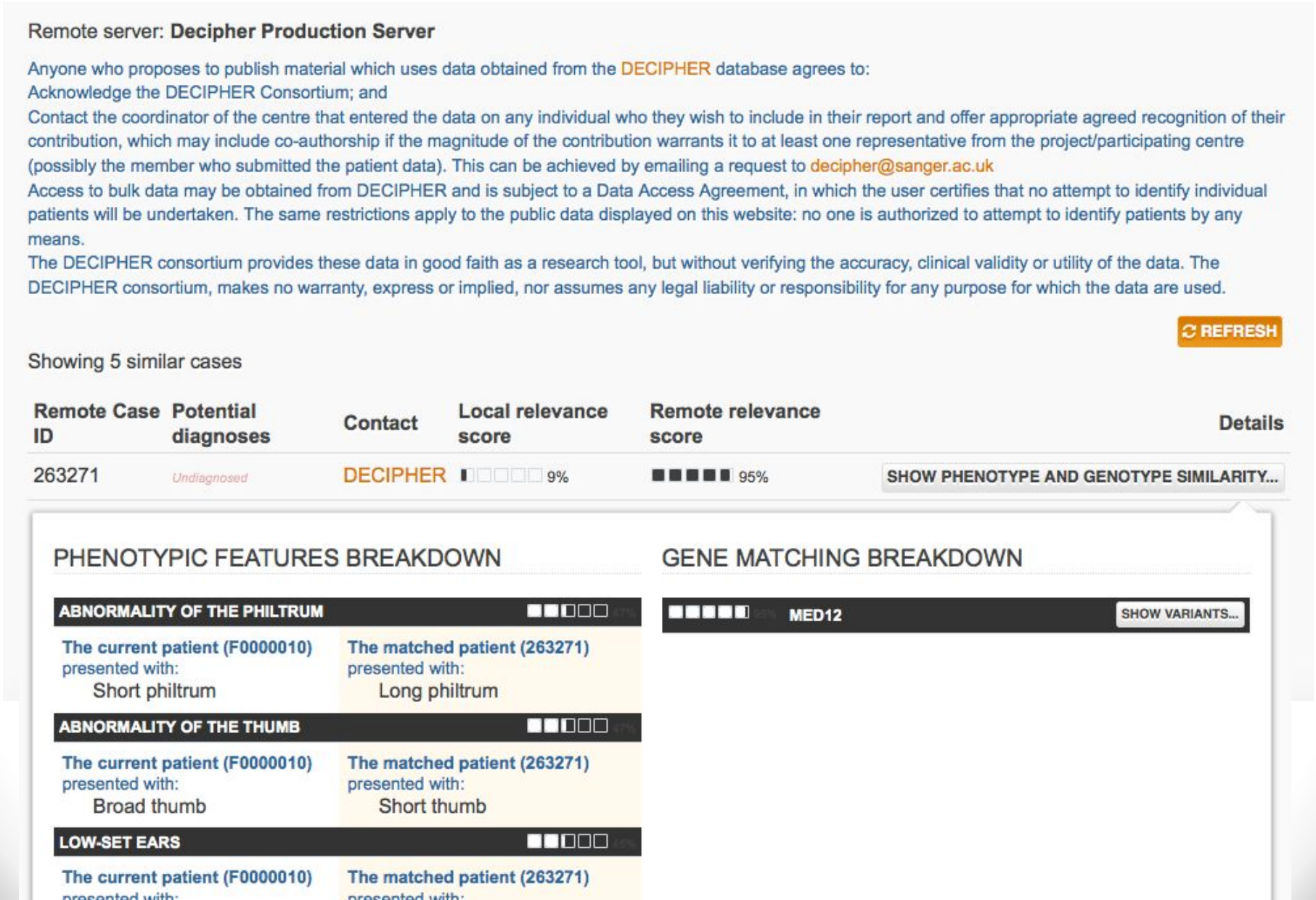
MME Network



IRDiRC
INTERNATIONAL RARE DISEASES RESEARCH CONSORTIUM

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

MME Network (in PC)



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

Want to join us?

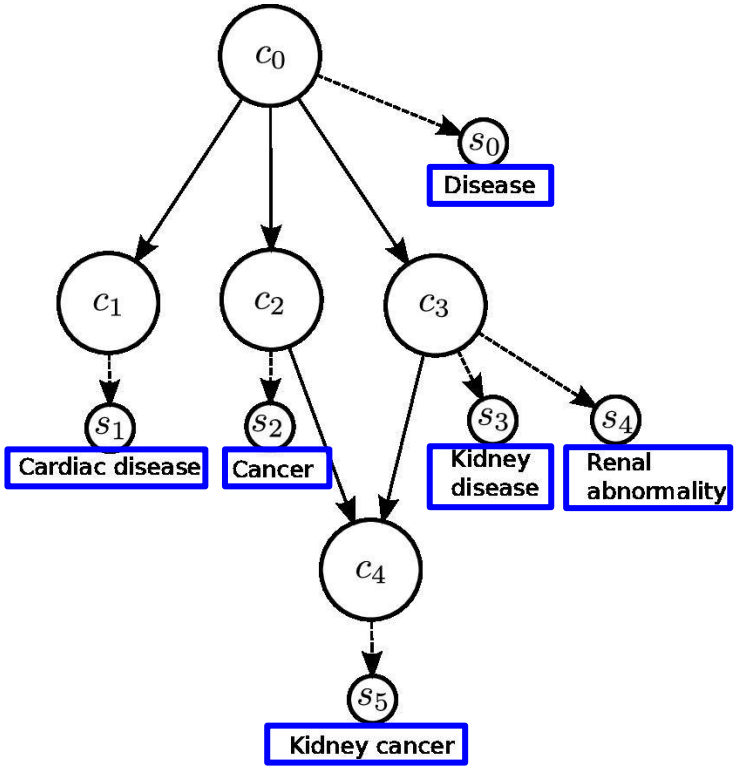
king for talented Grad Students, Postdocs and
to work on computational analysis of clinical data

brudno@cs.toronto.edu

SickKids®

Training details

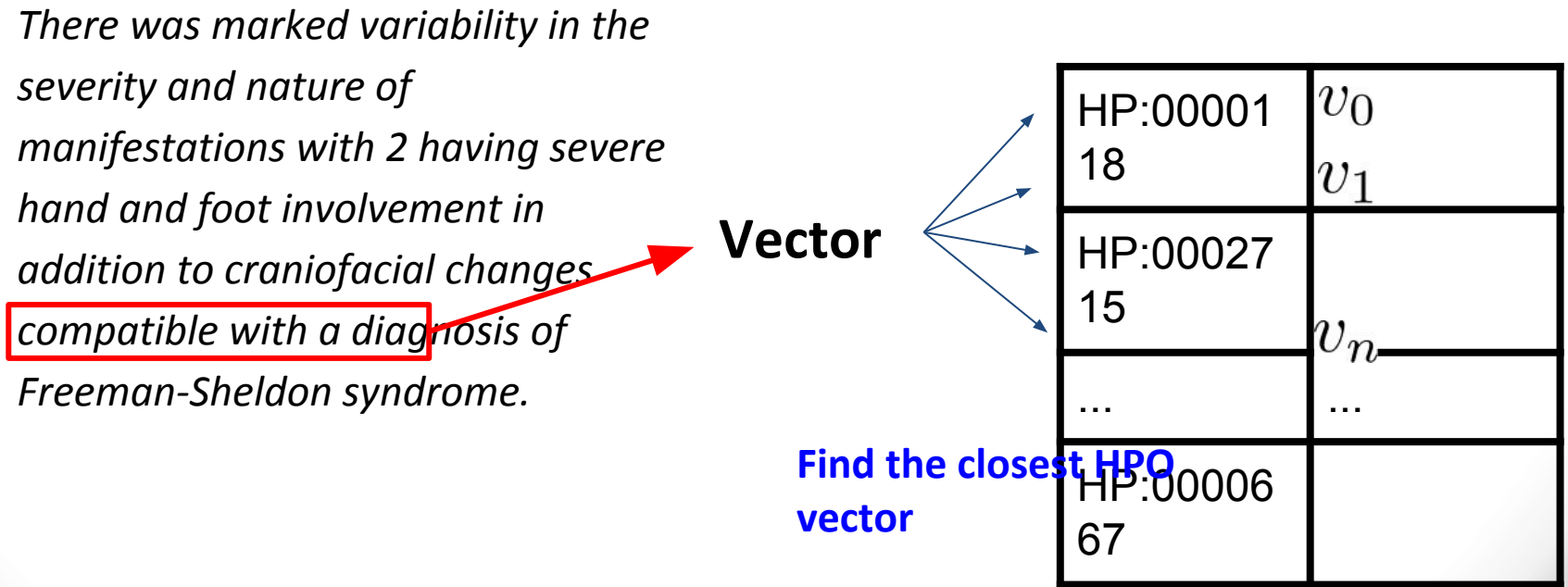
- Stochastic Gradient is used for training
- The following parameters are trained:
  - The RNN parameters (embedder)
  - The HPO concept embeddings
- The training is repeated for 20 epochs
- In each epoch data is divided in several batches
- Training data includes all the HPO phrases (synonyms)



How to recognize concepts in a sentence?

- Method 1 - Window based approach

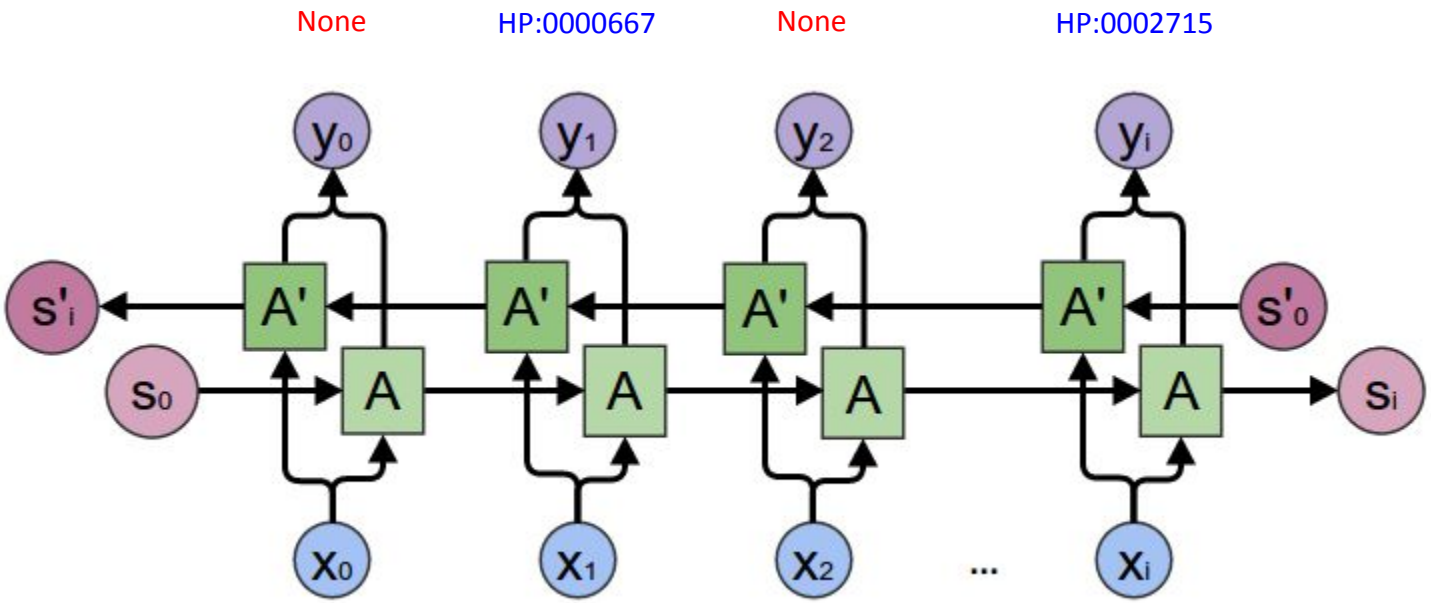
  Embed every window of 1 to 7 words and find the closest HPO vector



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: