

# Citizen Science for Biocuration

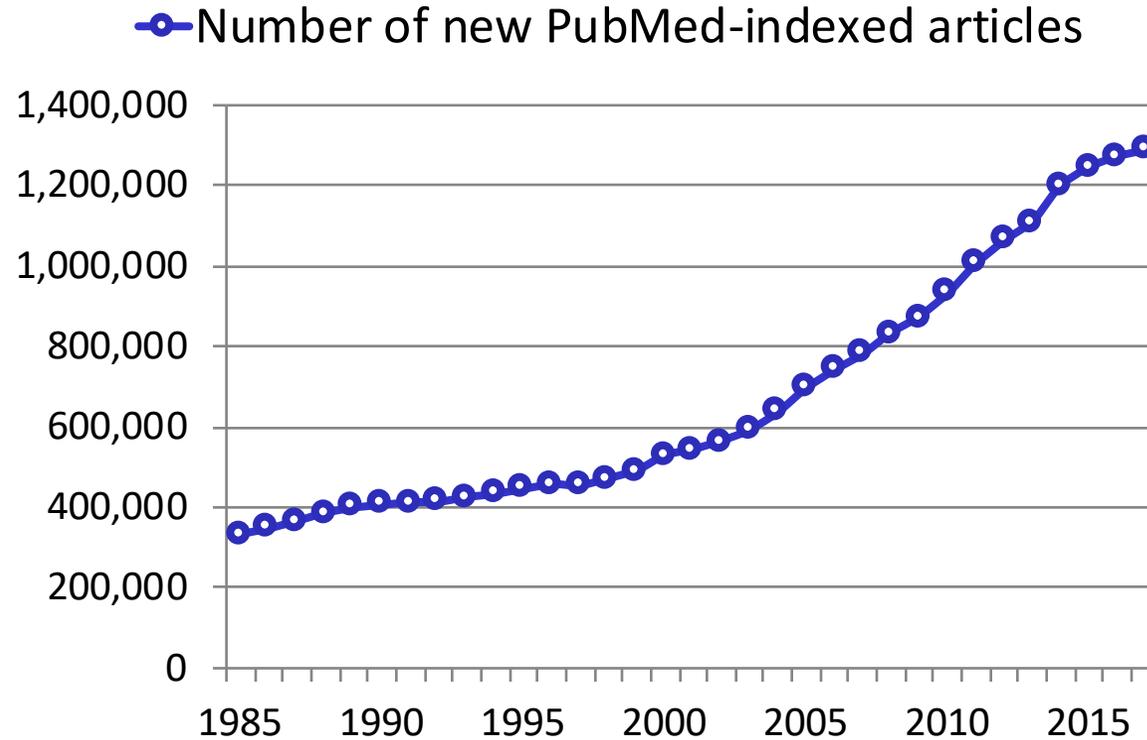


Andrew Su, Ph.D.  
@andrewsu   
<http://sulab.org>

May 31, 2019  
Curating the Clinical Genome  
Slides: <http://bit.ly/citsci-ccg19>



# The biomedical literature is massive...



**Over 29 million articles  
cumulative total**

... but it is very difficult to query and compute on

# Information extraction from biomedical text

## Named entity recognition (NER)

**DISEASES**

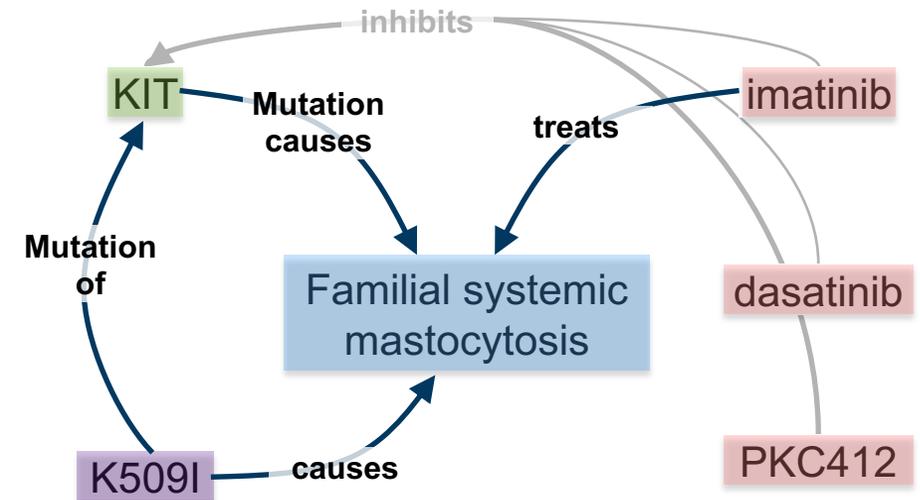
**GENES**

**VARIANTS**

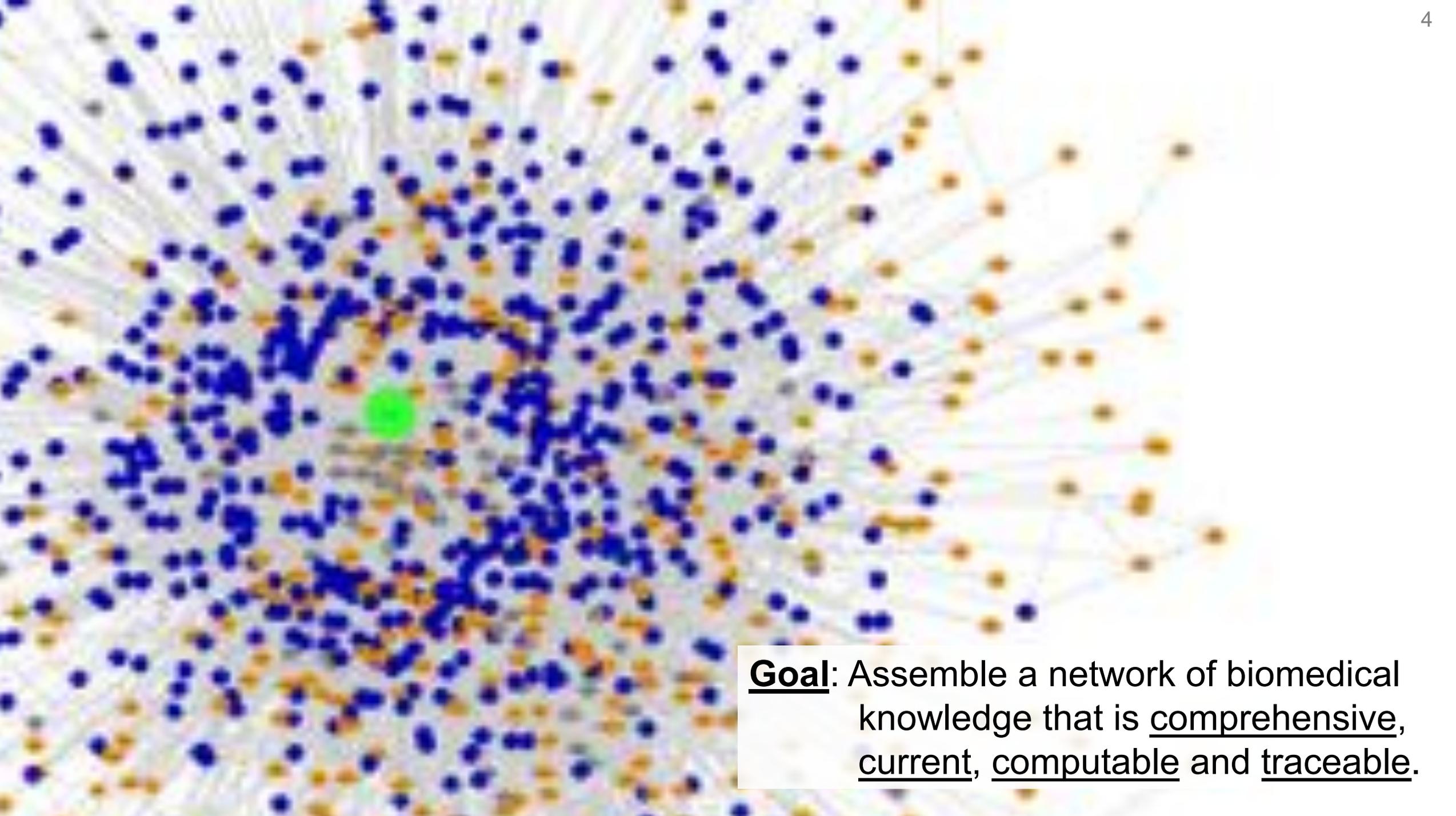
**DRUGS**

... We report a case of **familial systemic mastocytosis** with the rare **KIT K509I** germ line mutation. In vitro treatment with **imatinib**, **dasatinib** and **PKC412** reduced cell viability of primary mast cells harboring **KIT K509I** mutation. Both patients with **familial systemic mastocytosis** had remarkable hematological and skin improvement after three months of **imatinib** treatment.

## Relationship extraction



Subject	Predicate	Object
KIT	Mutation causes	Familial systemic mastocytosis
K509I	Mutation of	KIT
K509I	Causes	Familial systemic mastocytosis
Imatinib	Treats	Familial systemic mastocytosis
Imatinib	Inhibits	KIT
Dasatinib	Inhibits	KIT
PKC412	Inhibits	KIT



**Goal**: Assemble a network of biomedical knowledge that is comprehensive, current, computable and traceable.



## Gene Wiki Project

Building an open knowledge graph for all biomedical information within Wikidata

More info:

<http://bit.ly/wikidata-gene-wiki>



## MyVariant.info

Building high-performance web services for annotations of human genetic variants

More info:

<http://myvariant.info>  
<http://biothings.io>



Chunlei Wu

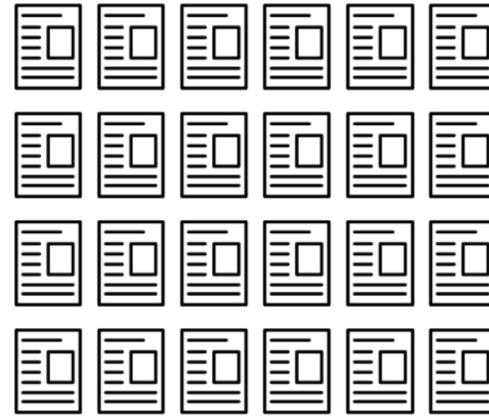
**Question:** Can a group of non-scientists collectively perform named entity recognition (NER) in biomedical texts?

# Experts versus crowd for concept identification



\$\$\$

**F = 0.78**



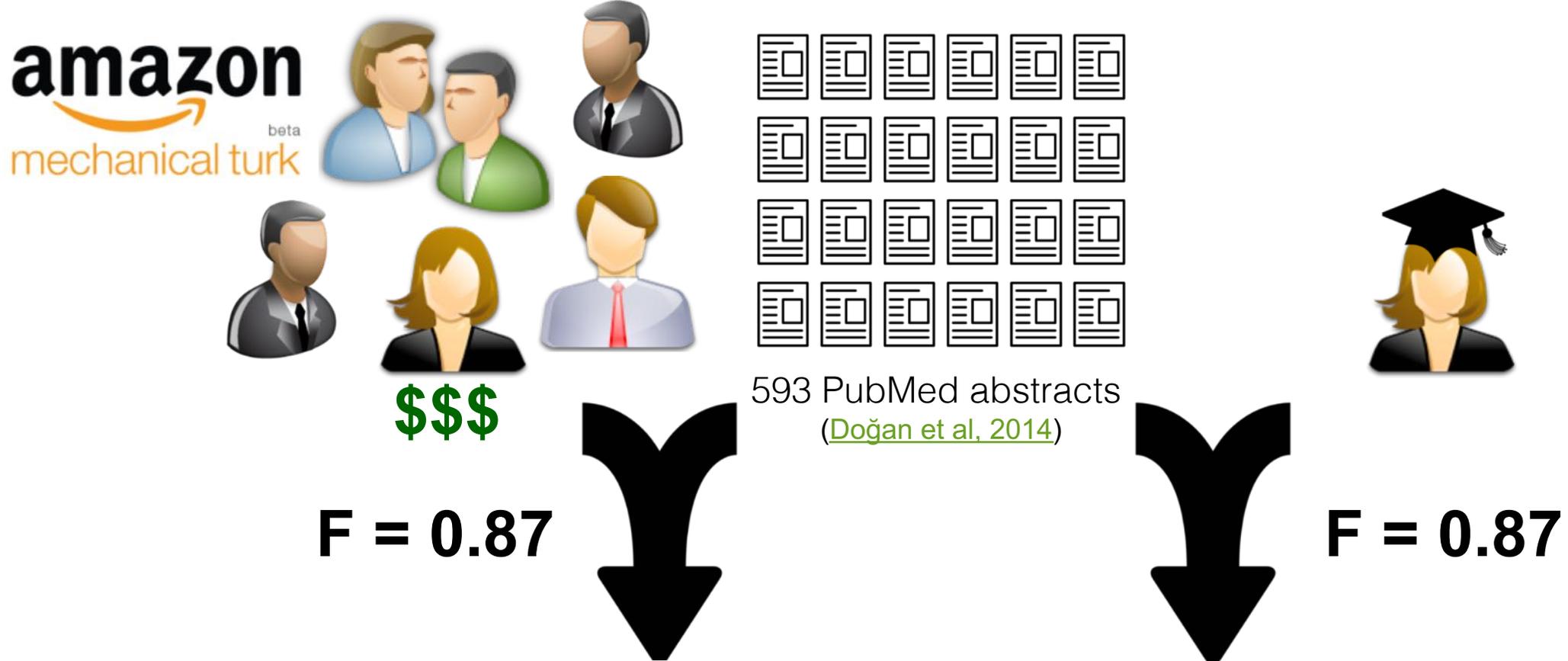
593 PubMed abstracts  
([Doğan et al, 2014](#))



**F = 0.87**

6,900 mentions of  
"disease concepts"

# Experts versus crowd for concept identification



- 9 days
- 145 workers
- Cost: \$630.96

6,900 mentions of  
“disease concepts”



Push research *forward*  
and discover cures *faster*.

Let's help scientists organize biomedical  
knowledge and uncover hidden links.

If you can READ, you can HELP

**START NOW**  
(BETA-EXPERIMENT)

*Learn More*



*News*

- 2015.02.01 - [News Article] Medical Researchers seek Public's Help
- 2015.01.30 - [Blog Post] Help us reach our goal so we can help you
- 2015.01.27 - [Blog Post] How the Rare disease community can help Mark2Cure help rare disease research

*Experiment Progress*



The current experiment  
is 20% complete

Help complete it!

**Start Now** or **Login**

Recruit your friends to  
Help complete it!

**Share on Twitter**

**Share on facebook**

*Join our mailing list*

**Join now**

## Paid crowdsourcing



\$\$\$

- **F = 0.87**
- 9 days
- 145 workers
- Cost: \$630.96

## Citizen Science



“Help science, please”

- **F = 0.84**
- 28 days
- 212 workers
- Cost: \$0 (\*\*\*)

# Does Citizen Science scale?

Number of annotation  
events per year needed

1,000,000 articles \* 10 AE / article

10,275 AE \* 365 days

212 annotators\* 28 days

=

**15,828  
volunteers  
needed**



Number of annotation  
events per year  
per volunteer done

AE = Annotation events







Φύτνης Καφάτος



# Bertrand was the first case of NGLY1, but he is not alone.

NGLY1 Researchers are racing to find clues in biomedical literature and need your help to uncover hidden links. If you can read, you can help.

About NGLY1

Get Started

▶ Watch Video

787,094 annotations have been submitted so far, but we're not done! Your help is still needed ... [Learn More](#) ▶

## Current MISSIONS.

 Stress Response and glcnac Total Docs: 73	 Stress Response and glcnac Total Docs: 75	 Galactosemia and Oxidative Stress Total Docs: 36	 Total Docs: 75	 Total Docs: 48	 GlcNAc and lipocalin Total Docs: 68	 Oxidative Stress Total Docs: 444
---	--	---	---	---	---	--

<http://mark2cure.org>



**Question:** Can crowdsourcing and citizen science be used for relationship extraction in biomedical texts?

Gene-Disease Relationships: Try it out! Remember, don't use outside information for **the relationship between TTR and Familial amyloidotic polyneuropathy**, but you're welcome to look up anything else (Eg- TTR, crest syndrome, multiple myeloma, etc.)

Use the menu in the box below to help relate the two concepts:

Select a Relationship below...

✕
TTR

relates to
has no relation to
cannot be determined

(may) exacerbate(s)
(may) treat(s)
(may) increase(s) risk of
(may) cause(s)
(may) prevent(s)
other relation, or relation unclear

✕

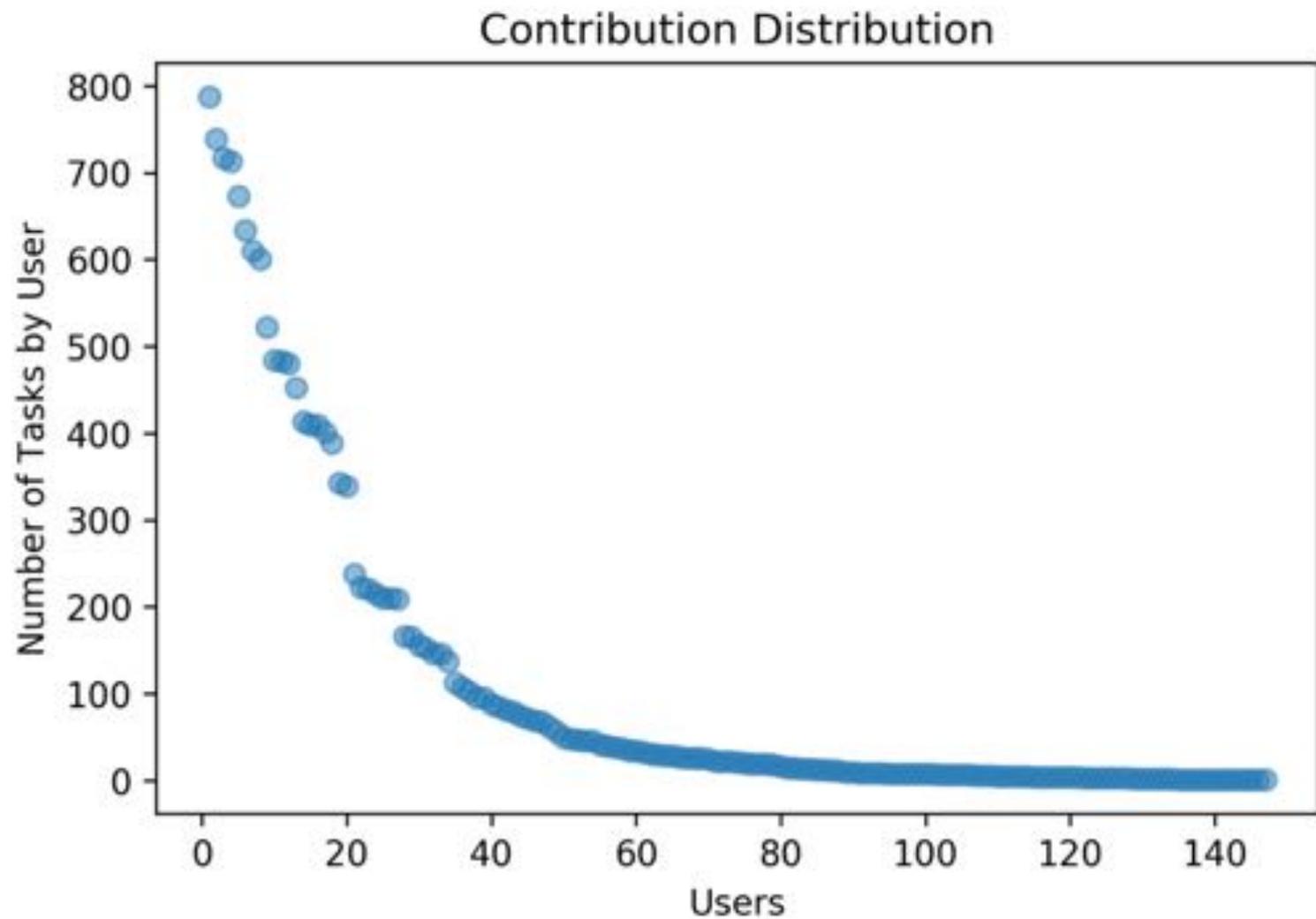
Familial  
amyloidotic  
polyneuropathy

Together with a case presenting in a patient with multiple myeloma, we describe 2 unique presentations including 1 associated with CREST syndrome in a patient with a previous history of breast carcinoma and another, also associated with cancer, with transthyretin deposits in a woman with a **TTR** gene mutation and a family history of **familial amyloidotic polyneuropathy**.

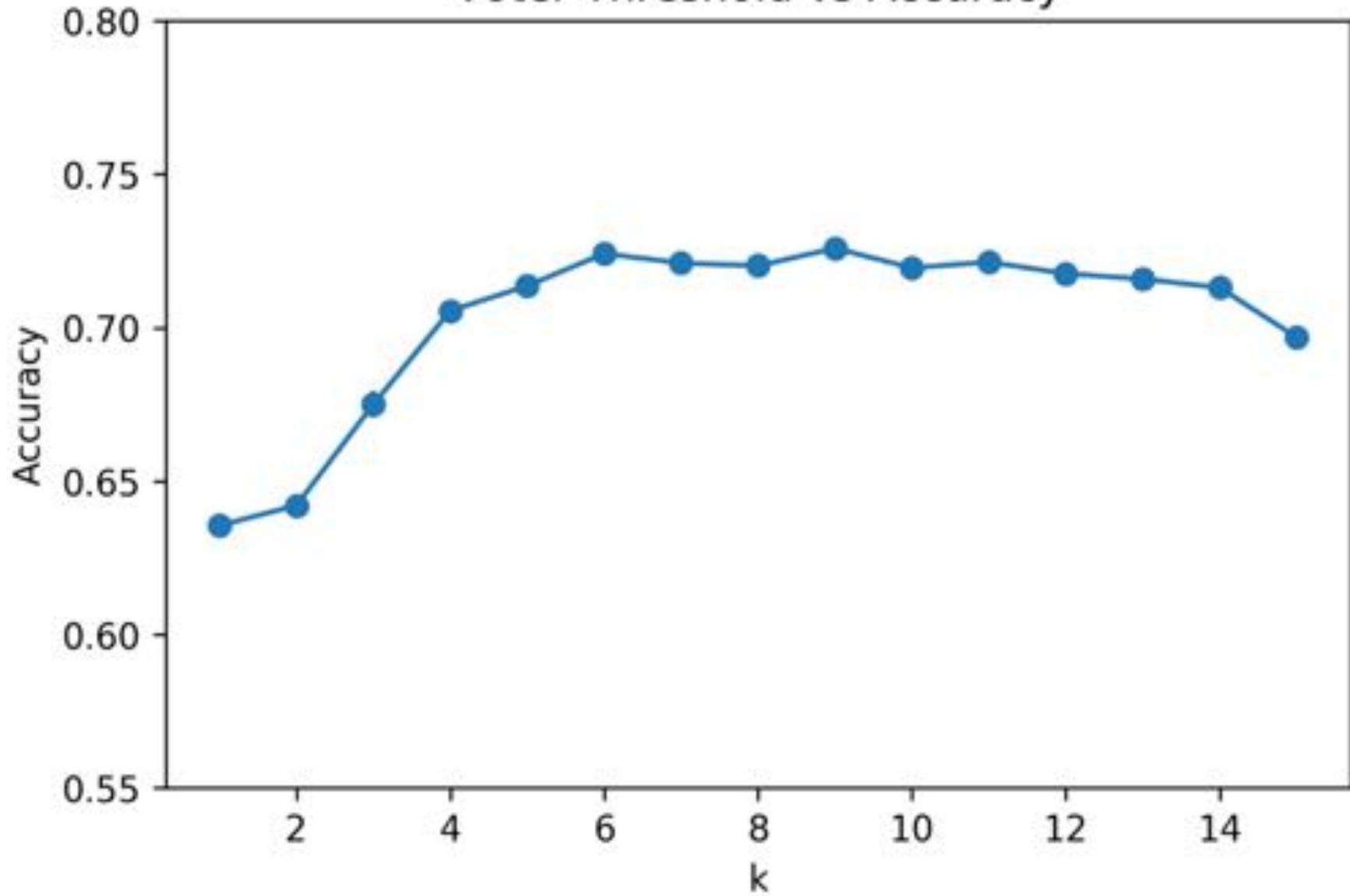
147 volunteers

15,739 annotations

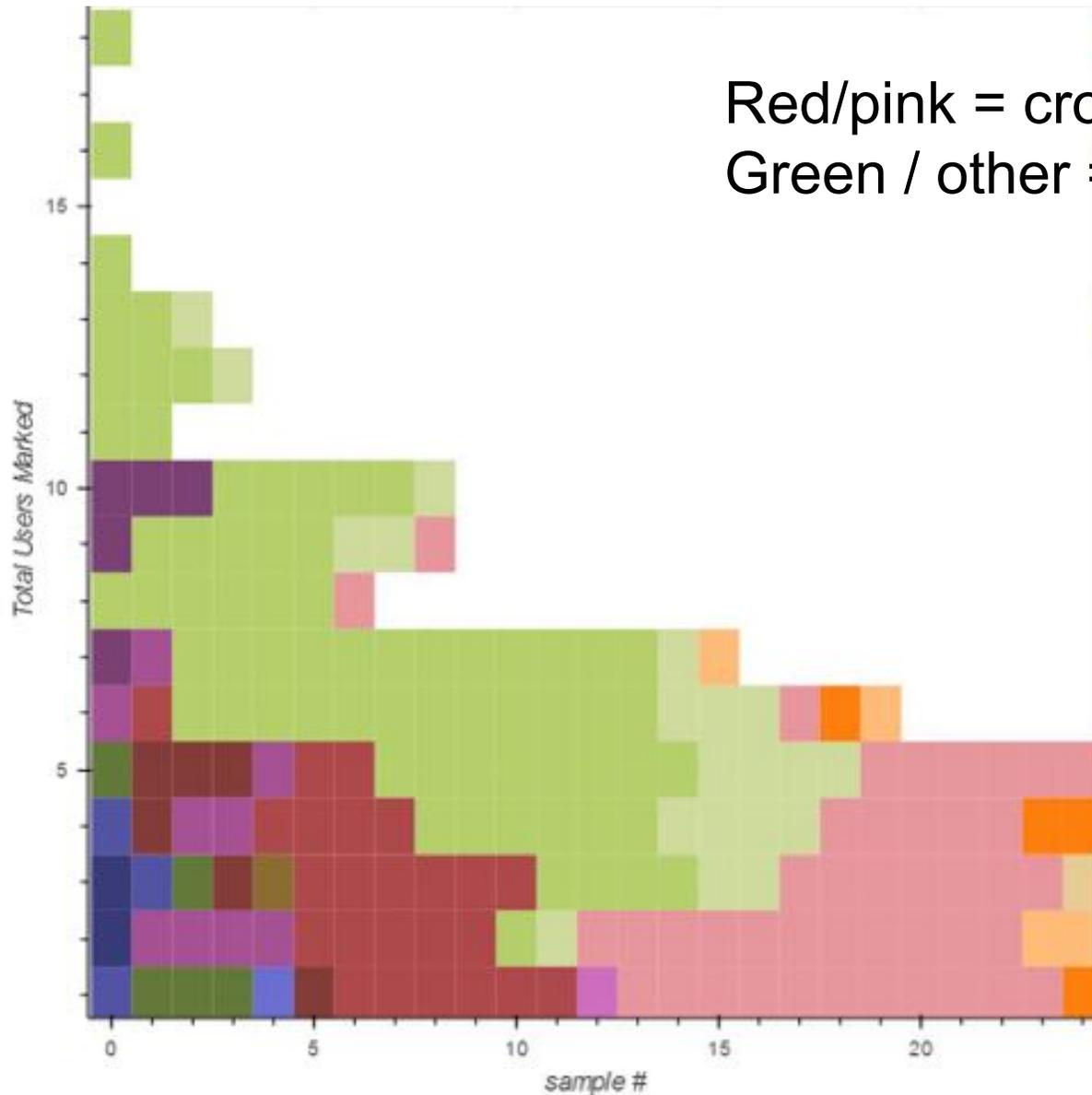
1009 concept pairs



Voter Threshold vs Accuracy



Red/pink = crowd disagreement with expert   
 Green / other = crowd agreement with expert 



-  relation with misdiagnosis
-  drug is altered in disease
-  drug failed
-  partial relationship
-  has available specific relationship
-  drug test for gene linked to disease
-  concept wrong
-  relation investigation in text
-  concepts vaguely related
-  drug is diagnostic reagent
-  gene resistance association
-  gene resistance is disease
-  nonspecific relation in text
-  gene resistance associated with disease that is often misdi
-  chem indirect relation to disease
-  drug treats overarching disease
-  disease confers resistance to drug
-  chemical is reagent for diagnostic test for different disease
-  chemical substrate for gene associated with disease
-  #!!!!
-  gene is marker

A few  
thoughts...





Citizen Scientists bring enthusiasm and energy

With training, they can be skilled members of a team...



Mark2Cure is a hammer...



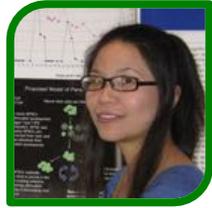
... in search of nails



Jennifer Fouquier



Max Nanis



Ginger Tsueng



Ben Good

# Crowd volunteers and partners worldwide



WIKIPEDIA  
The Free Encyclopedia



WIKIDATA



Estenik family



Leftwich family



SBP  
Hudson Freeze



UAB  
Matt Might & Might family



# Why do I Mark2Cure?

---

In memory of my daughter who had Cystic Fibrosis

Studied biology in college and I really miss it!

My 4 year old daughter Phoebe is living with and battling rare disease.

I have Ehlers Danlos Syndrome. I hope to help people learn about this painful and debilitating disorder, so that others like me can receive more effective medical care.

I am retired, have a doctorate in medical humanities, and have two children with Gaucher disease. I am just looking for some way to put my education to use.

Give back

I Mark2Cure in memory of my son Mike who had type 1 diabetes.

Take part in something that helps humanity.