The Co$t of Curation

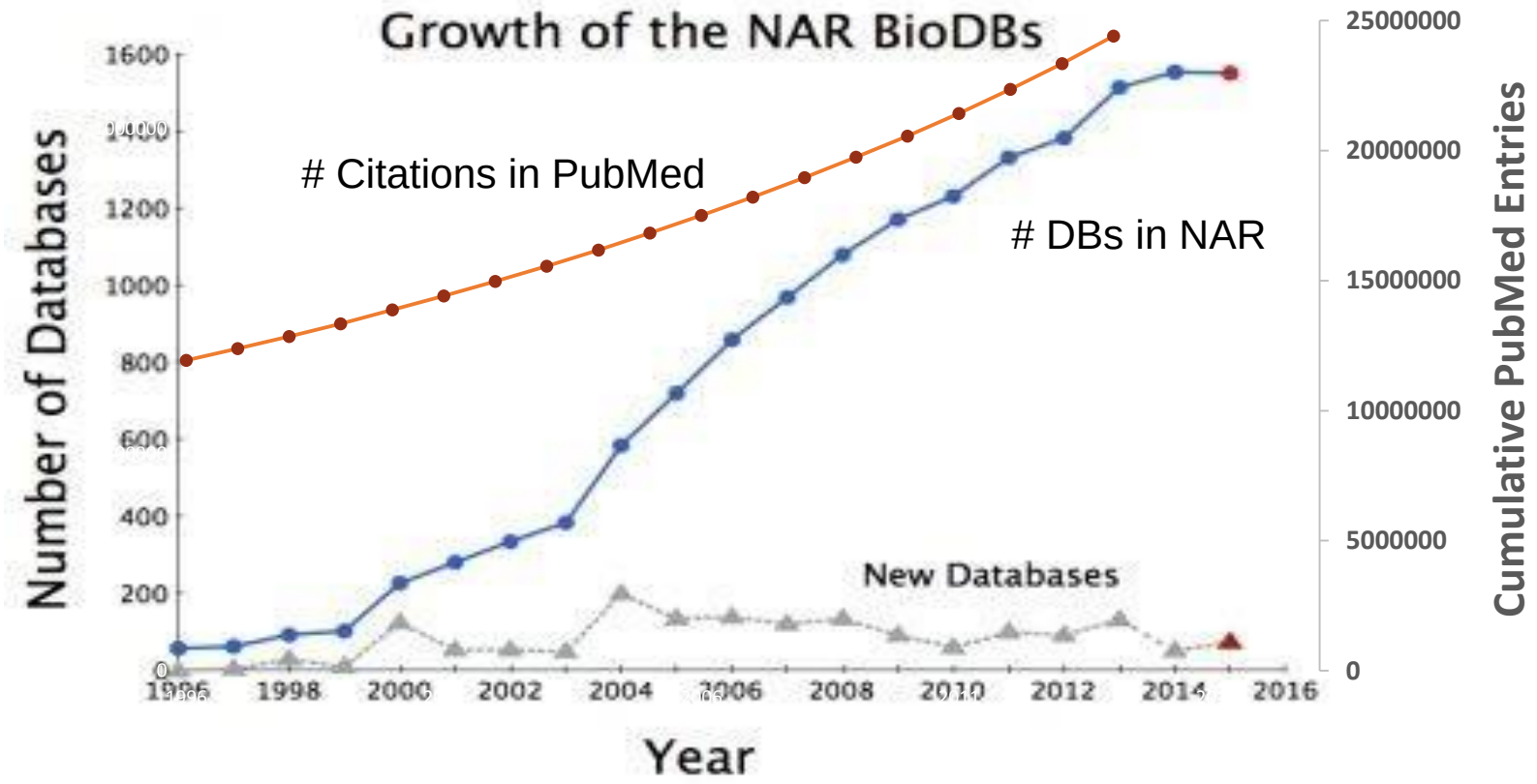
Lynette Hirschman
The MITRE Corporation
Swiss Institute of Bioinformatics Workshop
June 4-5, 2015

MITRE

Outline

- **Cost of Curation**
- **Hybrid Curation**
  - An approach to sustainable quality curation
- **Lessons Learned**

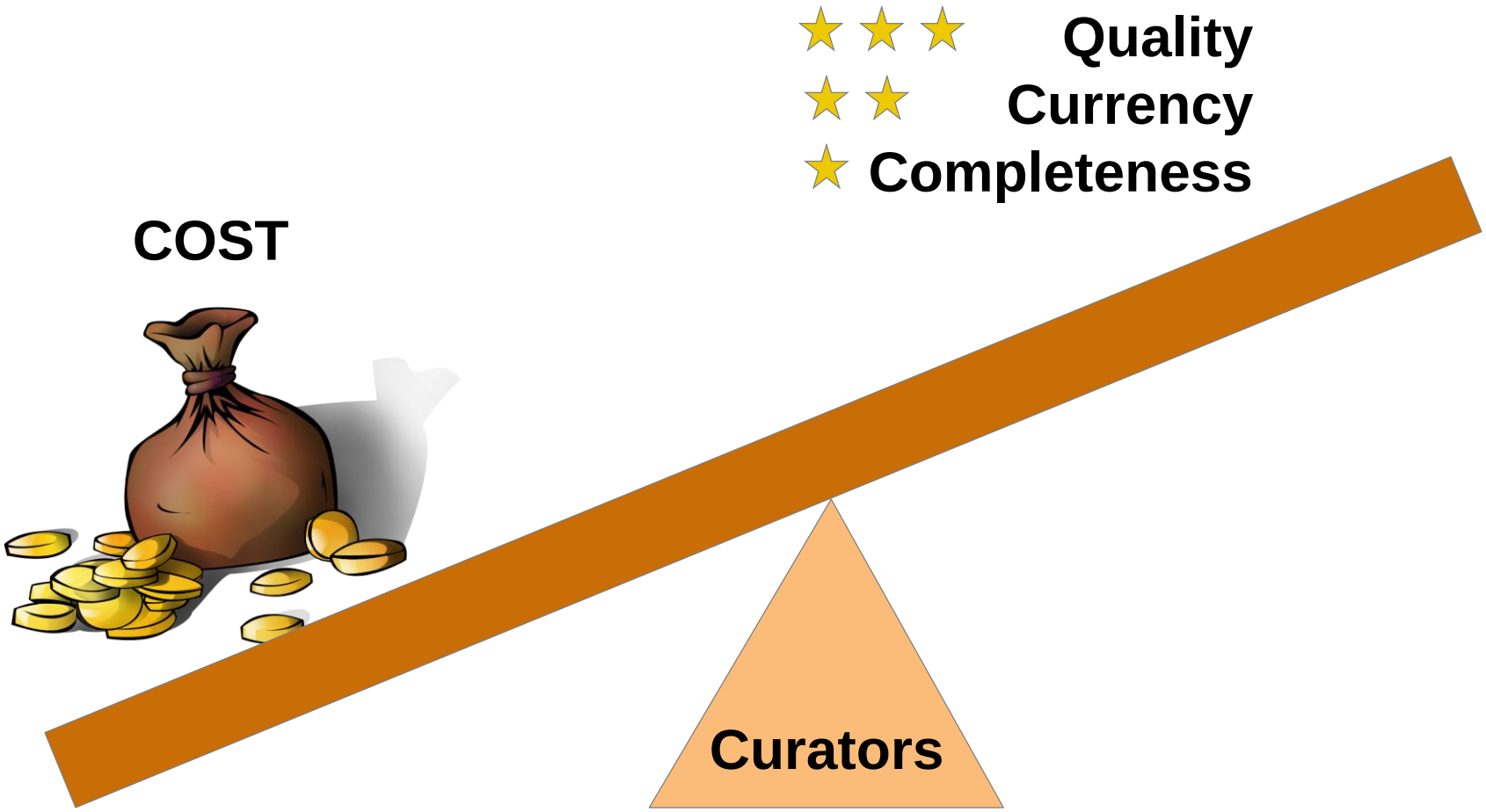
More Publications, More Databases…



Growth of NAR BioDBs from: @finchtalk: Bio Databases 2015
http://scienceblogs.com/digitalbio/2015/01/30/bio-databases-2015/

MITRE

A Balancing Act for Curators: the Four C's



MITRE

How to Keep Up?

- **The consumer perspective**
  - How to achieve maximum utility for a sustainable cost
- **The curator perspective**
  - How to keep up with the data tsunami on a fixed budget
- **The challenge**
  - Optimizing the cost-benefit trade-offs

How to quantify utility or benefit?

How to measure cost?

What are the trade-offs?

Measuring the 4 C's

- **Cost**
  - ✓ Steady state cost (for some projects)
  - **?** Maintenance (little information)
- **Quality**
  - **X** Correctness against a gold standard – no gold std!
  - **$$** Consistency among curators (inter-curator agreement)
- **Currency (Throughput)**
  - ✓ Rate of curation (some data)
- **Completeness**
  - ✓ What is covered – often related to throughput

Case Study #1
Linguistic Annotation: Propbanking[1]

- **Task: Linguistic annotation**
  - Annotation of propositional structure, e.g., subject-verb-object relations, to train & evaluate automated natural language processing systems
- **Annotator cost ($25/hr)**
  - 2.5 predicates per sentence; 2+ hours for 60 instances (24 sentences)
  - Double annotation, adjudication, pre & post processing
- **Throughput:**
  - 5K sentences in 14-16 weeks @ 30 hrs/wk
- **Overall costs:**
  - $13,200 for 5K sentences + $7K for set up
- **Estimated steady state:**
  - Cost: ~$2.60/sentence; ~$1.00 per annotation
  - Quality: Interannotator agreement high: Kappa[2] > 0.90
  - Throughput: 1600 sentences/month or ~160 short docs/month

[1]Data supplied by Martha Palmer, U Colorado

[2]Kappa coefficient measures inter-rater agreement taking chance agreement into account

Case Study #2: The Comparative Toxicogenomics Database[1]

- **Task: Curation of full text articles for deposit in a database**
  - Entities: genes, diseases, chemicals
  - Interactions: gene/disease, gene/chemical, chemical/disease
- **Throughput**
  - 6K articles/year (plus special projects) for the curation team
- **Curation statistics**
  - 93% time spent on "curatable" papers; 7% on rejecting papers
  - Ave. time/curatable paper: 20 min/paper for ~30 interactions
- **Estimated steady state**
  - Cost: $10/paper (@ $30/hr); ~ $0.33/interaction
  - Quality: Average precision = 0.91; average recall = 0.71[2]
  - Throughput: 3 papers/hr per curator

[1]Data from Carolyn Mattingly, A.P. Davis, Comparative Toxicogenomics Database

[2]Wiegers TC, Davis AP, Cohen KB, et al. Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (CTD). BMC Bioinformatics. 2009;10:326

Case Study #3: Medical Records[1]

- **Task: Data for the i2b2[2] Challenge Evaluation**
  - Medical concepts & negation/uncertainty, relations, coreference
  - 10-12 annotators for review and adjudication
  - Plus overhead of ~15%-20% for supervision
  - Adjudication and additional machine-assisted layers of review to generate the final reference standard
- **2010 i2b2 Challenge: $40K**
  - 150K annotations, 800 documents (patient notes)
- **2011 i2b2 Challenge: $38K**
  - 80K annotations, 800 documents (patient notes)
- **Estimated steady state**
  - Cost: ~ $50/patient note; $0.25-0.50/annotation
  - Throughput: ~100 notes/month
  - Quality: [double annotation plus adjudication]

[1]Data from Brett South, Salt Lake City VA and U of Utah

[2]Informatics for Integrating Biology and the Bedside

MITRE

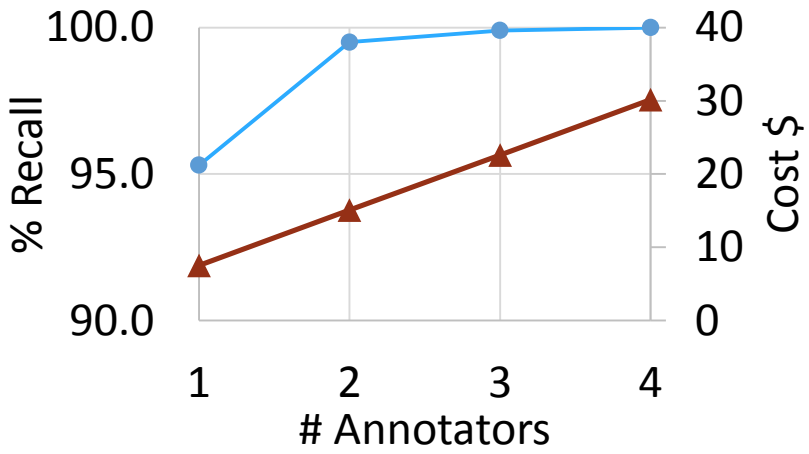
Case Study #4: De-identification of Clinical Notes[1]

- **Task: De-identification of free text in clinical notes of electronic health records**
  - Removal of Personal Health Information: 18 classes of information per US HIPAA[2] regulations, including patient name, address, social security number, phone number, etc.
  - Annotators identify and redact all types of PHI in a patient note
- **Corpus**
  - 100 clinical records were de-identified by 4 annotators,
  - 1093 PHI instances total (~10 instances per note)
- **Estimated steady state**
  - Cost: $7.50/patient note/annotator; $0.70/annotation
  - Throughput: NA
  - Quality: 95% recall w single annotator; 99.5% w 2 annotators

[1]From: Is the Juice Worth the Squeeze? Costs and Benefits of Multiple Human Annotators for Clinical Text De-identification, David S. Carrell, David J. Cronkite, Bradley A. Malin, John S. Aberdeen, Lynette Hirschman, submitted for publication

[2]US Health Insurance Portability and Accountability Act

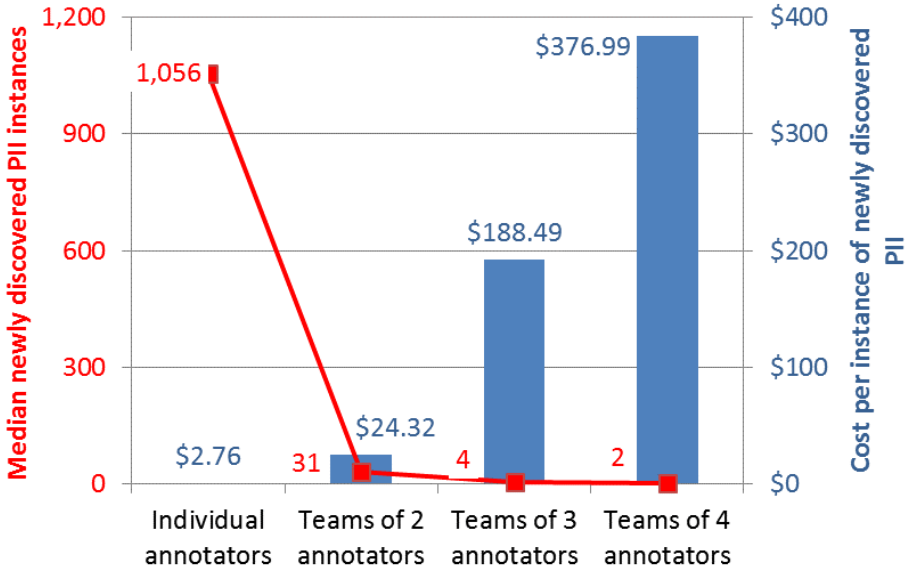
De-Identifying Patient Records: Recall vs. Cost



Recall (blue) and cost (red) with increasing # of annotators

Left (red): Median number of PII* instances discovered by increasing #s of annotators

Right (blue) Cost per additional PII* instance discovered in 2014 dollars.



*Personally Identifying Information

MITRE

Annotation/Curation Costs - Summary

Corpus	Type	Cost/unit	Unit	Cost/ Annot	# Units Annot	# Annot	Elapsed Time	Quality
PropBank	Linguistic propositions	$3	sentence	**$0.98**	5,000	13,000	14-16 wks	> .9 Kappa
2010 i2b2 Challenge	Medical concepts	$50	patient note	**$0.27**	800	150,000	48 wks	
2011 i2b2 Challenge	Medical concepts	$48	patient note	**$0.48**	800	80,000	48 wks	
De-identification	Personal Health Identifiers	$7.50 (1x) $15 (2x)	patient note	**$0.69**	100	1,093		Recall 95% (1 X) 99.5 (2x)
Comparative Toxicogenomics	Biomedical Literature	$11	journal article	**$0.42**	2,400	60,000	7 wks	Prec 0.91 Rec 0.71

- **Cost per annotation ranges from $0.25 to $1.00**
- **Cost per "document" depends on annotation density and document length**

Linguistic annotation  Medical annotation  Biological curation 

MITRE

How Can We Make Curation Cheaper?

- **Option 1: Automated Tools (Text Mining)**
  - Text mining & information extraction tools can assist
  - But how to control for quality?
    - Are text mining tools accurate enough?
    - Is automated curation plus human review cost-effective?
- **Option 2: Crowdsourcing**
  - Crowdsourcing has been effective for linguistic annotation
    - But linguistic annotation may not require specialized expertise
  - Crowdsourcing from citizen scientists can be effective
    - Quality control is an issue
- **Option 3: Combine these for Hybrid Curation**

Crowdsourcing for Curation: A Hybrid Approach

Hypothesis

We can use automated preprocesses to accelerate human annotation

- ✓ **Automated entity extraction has reasonable accuracy**
- ✓ **Automated relation extraction is still hard**
- ✓ **Crowdsourcing enables new models of annotation**

Proposed annotation workflow

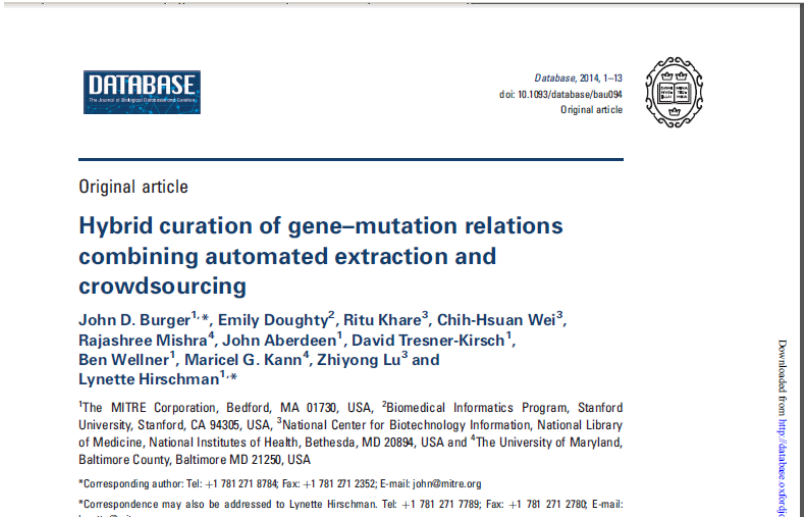
1. **Automatically** extract elements (e.g., genes, mutations)
2. **Automatically** prepare candidate relations for human annotation
3. **Collect and process human judgments on these relations**

Research Context: Unlocking Information in Free Text

- **Critical medical observations are locked in narrative (free text) –**
  - For patient records
  - For findings reported in the biomedical literature
- **This inhibits**
  - Secondary use of clinical information
  - Combining patient data with findings from biomedical research
- **Unlocking this information will**
  - Support personalized medicine
  - Enable discovery of correlations between patient genotype (genetic variation) and phenotype (e.g., disease, drug response)

MITRE

Based on Two Publications in *Database: The Journal of Biological Databases and Curation*



DATABASE

Database, 2014, 1–13
doi: 10.1093/database/bau094
Original article

Original article

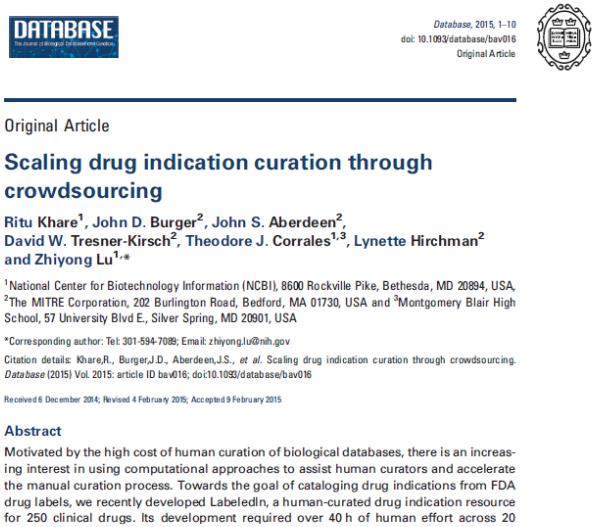
Hybrid curation of gene–mutation relations combining automated extraction and crowdsourcing

John D. Burger[1,*], Emily Doughty[2], Ritu Khare[3], Chih-Hsuan Wei[3], Rajashree Mishra[4], John Aberdeen[1], David Tresner-Kirsch[1], Ben Wellner[1], Maricel G. Kann[4], Zhiyong Lu[3] and Lynette Hirschman[1,*]

[1]The MITRE Corporation, Bedford, MA 01730, USA, [2]Biomedical Informatics Program, Stanford University, Stanford, CA 94305, USA, [3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA and [4]The University of Maryland, Baltimore County, Baltimore MD 21250, USA

*Corresponding author: Tel: +1 781 271 8784; Fax: +1 781 271 2352; E-mail: john@mitre.org

*Correspondence may also be addressed to Lynette Hirschman. Tel: +1 781 271 7789; Fax: +1 781 271 2780; E-mail:



DATABASE

Database, 2015, 1–10
doi: 10.1093/database/bav016
Original Article

Original Article

Scaling drug indication curation through crowdsourcing

Ritu Khare[1], John D. Burger[2], John S. Aberdeen[2], David W. Tresner-Kirsch[2], Theodore J. Corrales[1,3], Lynette Hirchman[2] and Zhiyong Lu[1,*]

[1]National Center for Biotechnology Information (NCBI), 8600 Rockville Pike, Bethesda, MD 20894, USA, [2]The MITRE Corporation, 202 Burlington Road, Bedford, MA 01730, USA and [3]Montgomery Blair High School, 57 University Blvd E., Silver Spring, MD 20901, USA

*Corresponding author: Tel: 301-594-7089; Email: zhiyong.lu@nih.gov

Citation details: Khare,R., Burger,J.D., Aberdeen,J.S., *et al.* Scaling drug indication curation through crowdsourcing. *Database* (2015) Vol. 2015: article ID bav016; doi:10.1093/database/bav016

Received 6 December 2014; Revised 4 February 2015; Accepted 9 February 2015

Abstract

Motivated by the high cost of human curation of biological databases, there is an increasing interest in using computational approaches to assist human curators and accelerate the manual curation process. Towards the goal of cataloging drug indications from FDA drug labels, we recently developed LabeledIn, a human-curated drug indication resource for 250 clinical drugs. Its development required over 40 h of human effort across 20

Organization	Collaborators
MITRE	John Aberdeen, John Burger, Lynette Hirschman, David Tresner-Kirsch, Ben Wellner
National Center for Biotechnology Information (NCBI)	Ritu Khare, Zhiyong Lu, Chih-Hsuan Wei
University Maryland Baltimore County	Rajashree Mishra, Maricel Kann
Stanford University	Emily Doughty

MITRE

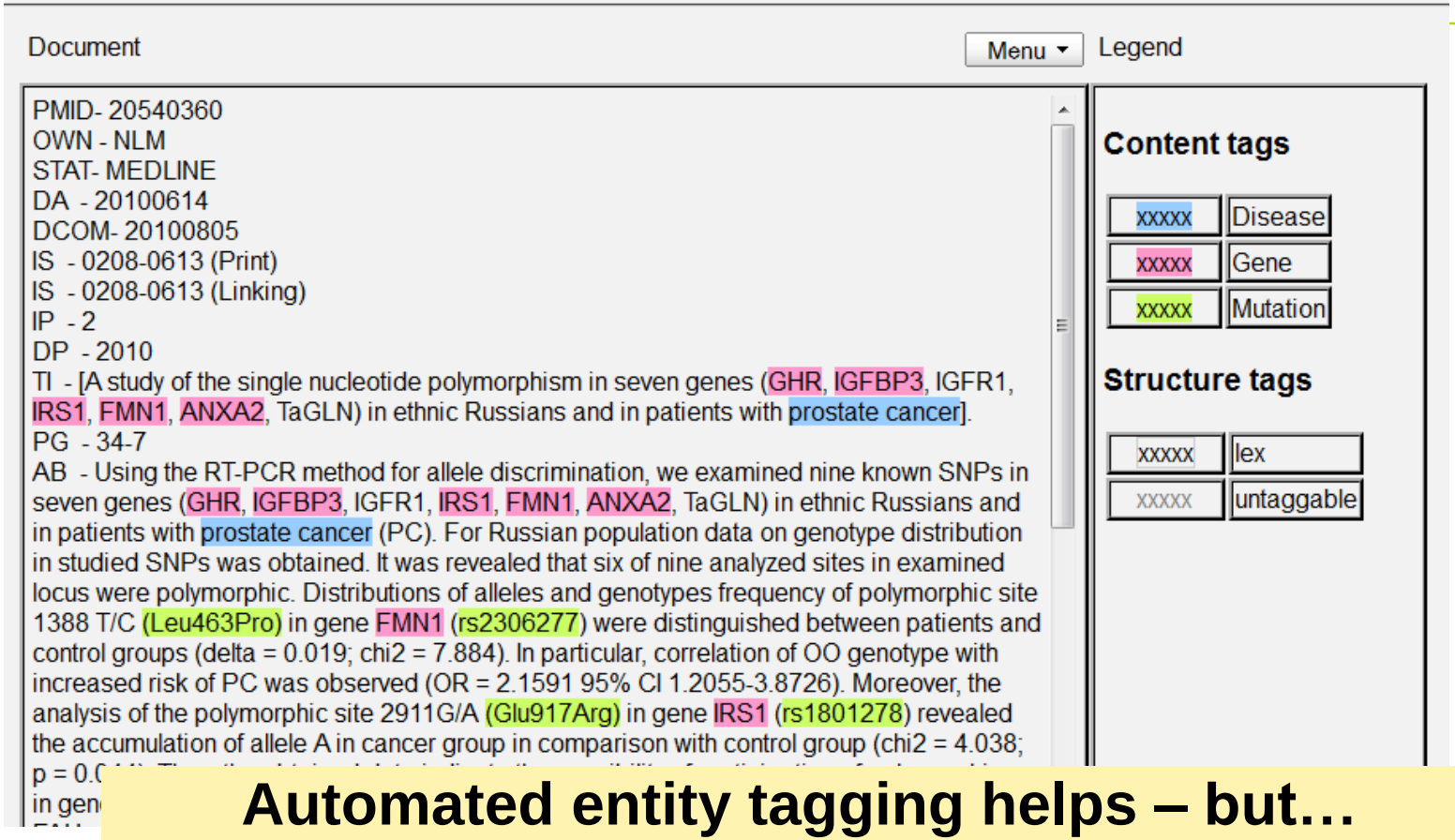
Application 1: Genetic Mutations Associated with Disease

- **What gene-mutation-disease relation(s) are in this abstract (PMID 20540360)?**

TI - [A study of the single nucleotide polymorphism in seven genes (GHR, IGFBP3, IGFR1, IRS1, FMN1, ANXA2, TaGLN) in ethnic Russians and in patients with prostate cancer].
PG - 34-7
AB - Using the RT-PCR method for allele discrimination, we examined nine known SNPs in seven genes (GHR, IGFBP3, IGFR1, IRS1, FMN1, ANXA2, TaGLN) in ethnic Russians and in patients with prostate cancer (PC). For Russian population data on genotype distribution in studied SNPs was obtained. It was revealed that six of nine analyzed sites in examined locus were polymorphic. Distributions of alleles and genotypes frequency of polymorphic site 1388 T/C (Leu463Pro) in gene FMN1 (rs2306277) were distinguished between patients and control groups (delta = 0.019; chi2 = 7.884). In particular, correlation of OO genotype with increased risk of PC was observed (OR = 2.1591 95% CI 1.2055-3.8726). Moreover, the analysis of the polymorphic site 2911G/A (Glu917Arg) in gene IRS1 (rs1801278) revealed the accumulation of allele A in cancer group in comparison with control group (chi2 = 4.038; p = 0.044). Thus, the obtained data indicate the possibility of participation of polymorphism in genes FMN1 and IRS1 in formation of predisposition to PC.

It's hard!

Highlighting Genes, Mutations, Diseases



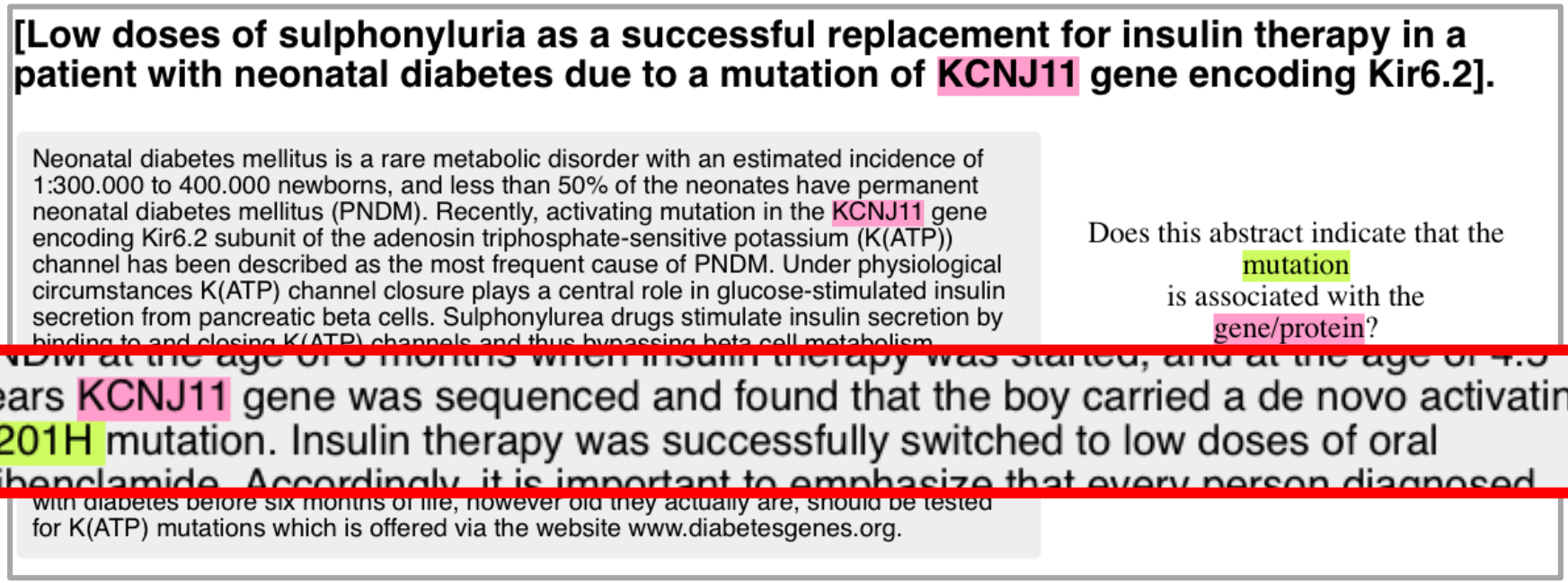
Automated entity tagging helps – but…

20 possible Gene-Mutation-Disease relations

How to find the valid ones?

MITRE

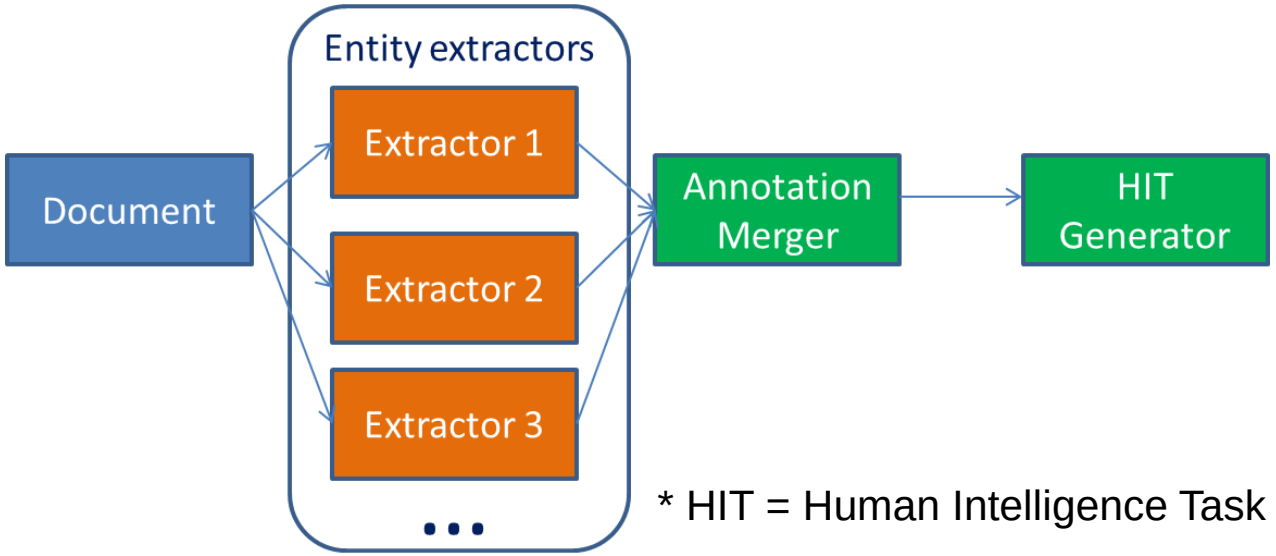
Crowdsourcing: Relation Annotation in Context



- **Split task into simple judgments:**
  - E.g., is this mutation associated with this gene?
- **Extract and highlight entities in context**
- **Frame task as simple yes/no questions**

MITRE

Automatic Extraction Workflow: Gene-Mutation Relations

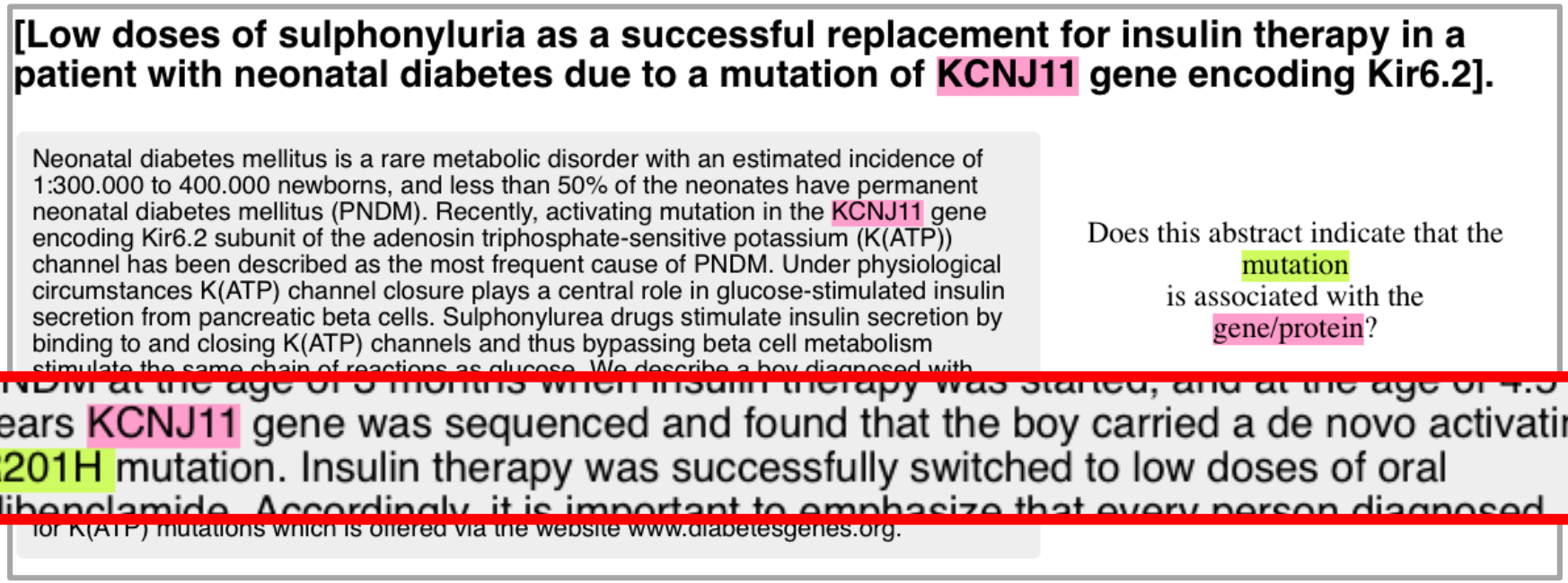


* HIT = Human Intelligence Task

- **Extractor 1: Genes**
  - GenNorm gene tagger from NCBI
- **Extractor 2: Mutations**
  - Extractor of MUtations (EMU) from UMBC[1]
- **Extracted entities are merged and highlighted in text**
- **One gene/one mutation are highlighted in an abstract, presented for judgment to Amazon's Mechanical Turk**

[1]University of Maryland, Baltimore County

HIT Generation (Human Intelligence Task)

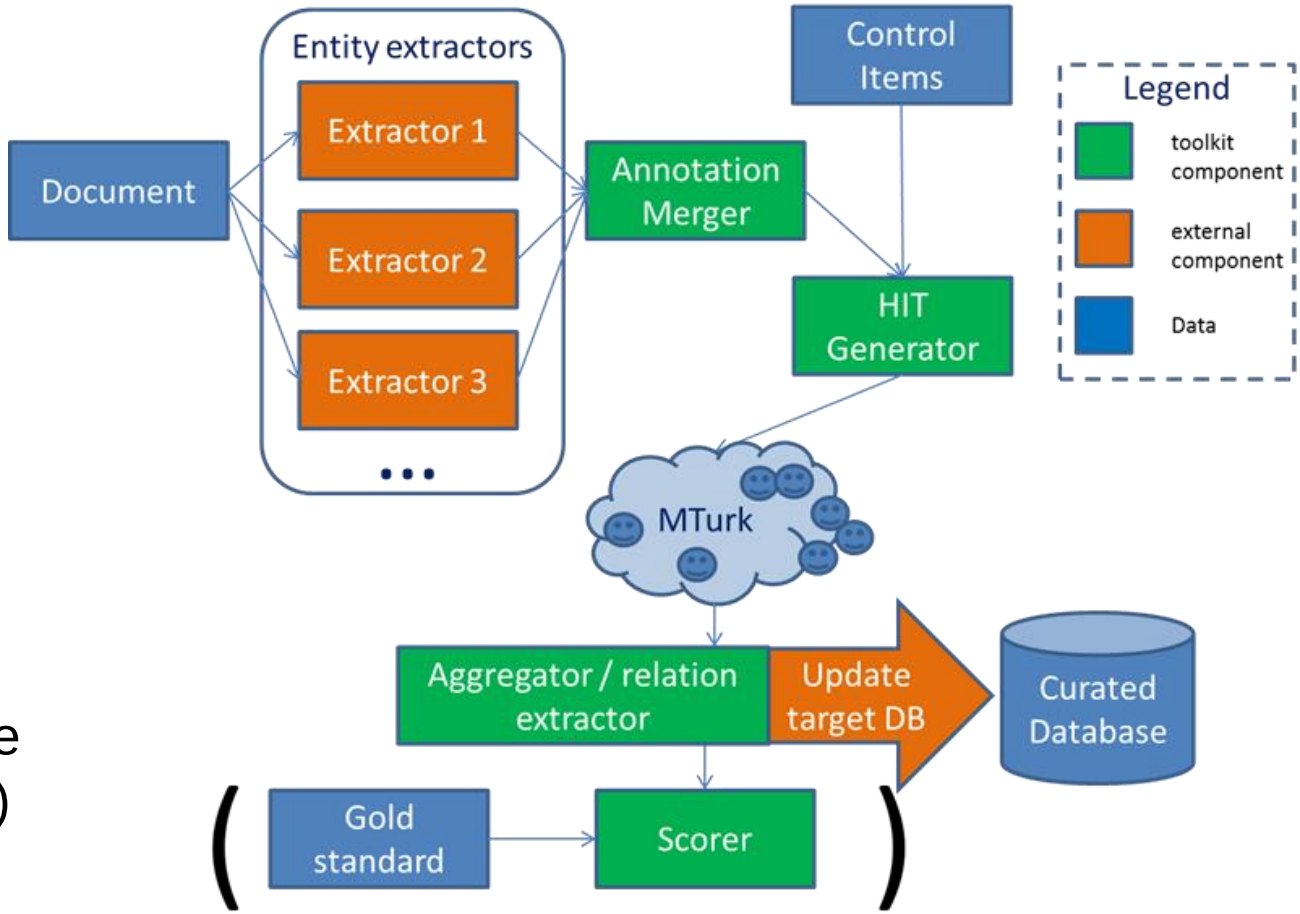


- **Select 1 gene, 1 mutation for display**
- **If context is local, judgment can be fast**
- **Pay a few cents/judgment**

MITRE

Hybrid Curation: Full Workflow

- Present candidate relations as HITs
- Inject control items (with known answers) to weight Turker responses
- Aggregate Turker judgments
- Validation: Score for accuracy
- Production: Deposit in database (with expert review)



Experiment Data on Gene-Mutation Relations

	Expt 1	Expt 2
# Turkers	23	24
# Abstracts	250	275
# Gene-mutation pairs (GOLD)	578	444
# Gene-mutation candidates posted (including control items)	1733	1354
Elapsed time	1.5 days	11 days*
Total cost @ 7¢ per judgment/5x Turking	$670	$521
Cost per abstract (5x Turking)	$2.68	$1.90

* Data released over the holidays

Results

- **Combining results from multiple Turkers improved accuracy**
- **Used Naïve Bayes to weight Turker efficacy based on performance on control items**
- **Precision/Recall for Expt 2:**
  - Precision* = 71.9
  - Recall** = 78.8

Accuracy		
	Expt 1	Expt 2
Individual responses	**75.5%**	**75.9**
Naïve Bayes	**84.5**	**85.3**

*Precision = # of correct answers returned/total answers returned

**Recall = # of correct answers returned/total # correct answers possible

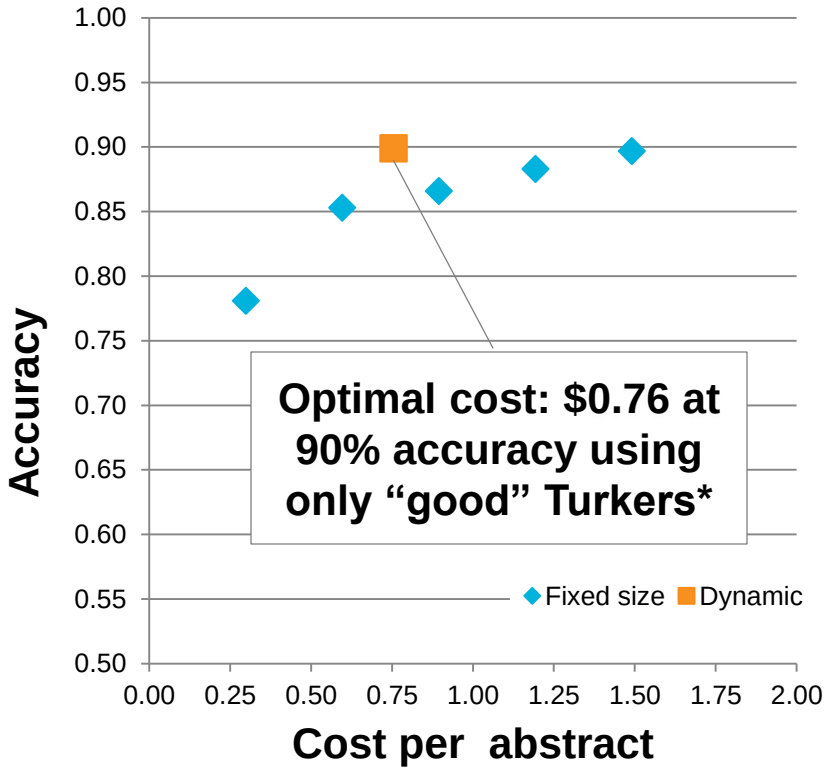
Turker Ablation Simulations

- **Requested five Turkers for each item**
  - Interested in performance with fewer Turkers
  - Straightforward linear reductions in cost
- **Simulate by ignoring last *n* responses for each item**
  - Aggregate with Naïve Bayes on the remaining
- **Note that earlier Turkers were better Turkers**
  - Due to notifications sent to best Turkers from Experiment 1

Turkers	5	4	3	2	1
Concept Accuracy	0.865	0.857	0.844	0.832	0.765
Recall (Turker)	0.937	0.951	0.948	0.926	0.869
Precision	0.771	0.752	0.734	0.725	0.652

Do We Need 5 Turkers? Simulation Experiments

- **Initially requested five Turkers for each item**
  - Interested in performance with fewer Turkers
- **Linear reduction:**
  - Simulate by ignoring last *n* responses for each item
  - Aggregate with Naïve Bayes on the remaining
- **Dynamic reduction:**
  - "Fire" any Turker who gets less than 50% on control items



*Turkers scoring at least 50% on control items

MITRE

Application 2: Drug Inserts*

- **FDA receives 2000+ new inserts per month**
  - In a specific XML format
- **Mostly prescription and over-the-counter drugs**
  - Also homeopathics, animal, some ingredients and devices, etc.
- **NCBI is interested in drug-disease relationships**
  - Indications vs. contraindication, risk factors, etc.
  - With plans to expand to other annotations

*Scaling Drug Indication Curation through Crowdsourcing, Ritu Khare, John D. Burger, John S. Aberdeen, David W. Tresner-Kirsch, Theodore J. Corrales, Lynette Hirschman, Zhiyong Lu; accepted for publication in Database (2015)

The Problem:
Text Mining Drug Indications from Drug Descriptions

SIDER-2, Neveol and Lu 2010, Wei et al. 2013, Fung et al. 2013

Drug	Indication Excerpts in DailyMed	
d1	Dutasteride capsules are indicated for the treatment of symptomatic benign prostatic hyperplasia. Dutasteride is not approved for the prevention of prostate cancer	contraindication
d2	Ranitidine is indicated in the treatment of GERD. Concomitant antacids should be given for pain relief to patients with GERD.	other drug's
d3	In patients with coronary heart disease, but with multiple risk factors for coronary heart disease such as retinopathy, albuminuria, smoking, or hypertension.	risk factors

Due to the challenges above, automatic methods are useful but not sufficient (65-80% accuracy) in creating ground truth.

MITRE

Setting Up the Experiment

- **NCBI created a gold standard for a subset of labels**
  - 534 most frequent searches on DailyMed website
  - Tagged yes or no by subject matter experts
- **Used same Hybrid Curation pipeline as in the gene-mutation experiments**
  - Automatically tagged drugs and diseases and other medical conditions
  - Turkers asked whether the **highlighted disease or condition** is an indication for the **highlighted drug**
- **Used a seven-way categorization to gather more data**
  - Single “yes” category
  - “No” category split according to reason for “no” judgment
  - “Uncertain” category

Annotate the Indications for a Prescription Medi

Show instructions

Please read the drug label below and indicate whether the highlighted disease or condition is an indication for the highlighted medication. Make sure you have read the instructions carefully. You *must* make a selection for every HIT—submissions with incomplete items run the risk of being be rejected. Thank you for your efforts!

Lidocaine

Lidocaine
Ointment USP, 5% (Spearmint flavored)
INDICATIONS AND USAGE

Lidocaine Ointment USP, 5% is indicated for production of anesthesia of accessible mucous membranes of the oropharynx.

It is also useful as an anesthetic lubricant for intubation and for the temporary relief of pain associated with minor burns, including sunburn, abrasions of the skin, and insect bites.

Does the text above state that lidocaine is used in the treatment, prevention, management, or relief of insect bites?

- 1. Yes
- 2. No - Characteristic or risk factor of the indicated disease
- 3. No - Side effect of the highlighted drug
- 4. No - Contraindication of the highlighted drug
- 5. No - Otherwise unrelated
- 6. No - Not a disease mention
- 7. Uncertain

Comments? We welcome your feedback.

Drug Indication Turking Run

- **74 Turkers did the task**
  - 5-fold judging per item
  - 6¢ per judgment
  - Median duration 12 seconds
- **706 drug labels**
  - 3004 items plus 20% control items from gold standard
- **18,775 judgments altogether**
  - Elapsed time 8 hours
  - $1,239 total
  - $1.75 per label
- **Results good!**
  - Lumping all “no” results together: Precision = 96%; Recall = 89%

Average control performance	
Accuracy (fine)	82.2%
Accuracy (coarse)	92.4
Precision	96.2
Recall	89.1

The 4 C's for Hybrid Curation Workflow

For curation, the workflow must be low cost, timely, and accurate

- **Currency/Throughput:** keep up with data flow
  - ✓ 10K papers/year on gene-mutation-disease (250 papers/wk)
  - ✓ 700 drug indication labels returned w 5x annotation in 8 hrs
- **Quality:** must be reliable (comparable to expert curation)
  - Expert curation: 90% precision @ recall > 70%
  - ✗ Hybrid gene-mutation curation: 82% precision @ 71% recall
  - ✓ Hybrid drug indication curation: 96% precision @ 90% recall
- **Completeness:**
  - ? Loss of information due to automatic extraction failure
- **Cost:** target is <$1 per abstract per relation type
  - ✓ Curation cost at $0.76/abstract for gene-mutation relations
  - ✓ Drug indication curation at $2/label for 5x annotation

Cost of Curation with Hybrid Curation

Corpus	Type	Cost/ unit	Unit	Cost/ Annot	# Units Annot	# Annot	Elapsed Time	Quality
PropBank	Linguistic propositions	$3	sentence	**$0.98**	5,000	13,000	14-16 wks	> .9 Kappa
2010 i2b2 Challenge	Medical concepts	$50	patient note	**$0.27**	800	150,000	48 wks	
2011 i2b2 Challenge	Medical concepts	$48	patient note	**$0.48**	800	80,000	48 wks	
De-identification	Personal Health Identifiers	$7.50(1x) $15 (2x)	patient note	**$0.69**	100	1,093		Recall 95% (1 X) 99.5 (2x)
Comparative Toxicogenomics	Biomedical Literature	$11	journal article	**$0.42**	2,400	60,000	7 wks	Prec 0.91 Rec 0.71
Gene-mutation relations	Biomedical Literature	$2	MEDLINE abstract	**$0.38**	225	1300	11 days	Prec 0.72 Rec 0.79
Drug Indications	Drug Inserts	$1.75	drug insert	**$0.41**	700	3000	8 hrs	Prec 0.96 Rec 0.89

Linguistic annotation
Medical annotation
Biological curation
Hybrid curation

MITRE

Hybrid Curation: Some Initial Lessons Learned

- **We can recruit Turkers with sufficient domain expertise**
  - Multiple Turkers with >90% accuracy on control items
- **Crowdsourcing can provide low latency, high throughput turnaround**
  - 1st expt: 36 hr turnaround for 250 abstracts
  - 2nd expt: 11 days for 275 abstracts – over Christmas holidays
  - 3rd expt (Drug-indications): 8 hours for ~700 drug labels
- **Aggregated results (5-fold Turking) gives reasonable accuracy**
  - Gene mutation relations: 85-90% accuracy
    - ~20% loss of recall (due to automated pre-process)
  - Drug indications: 96% precision at 89% recall
- **Cost**
  - For 5x judgments, cost is $2-3 per abstract or drug insert
  - Dynamic selection of Turkers can reduce cost, improve accuracy

Hybrid Curation: Are We There Yet?

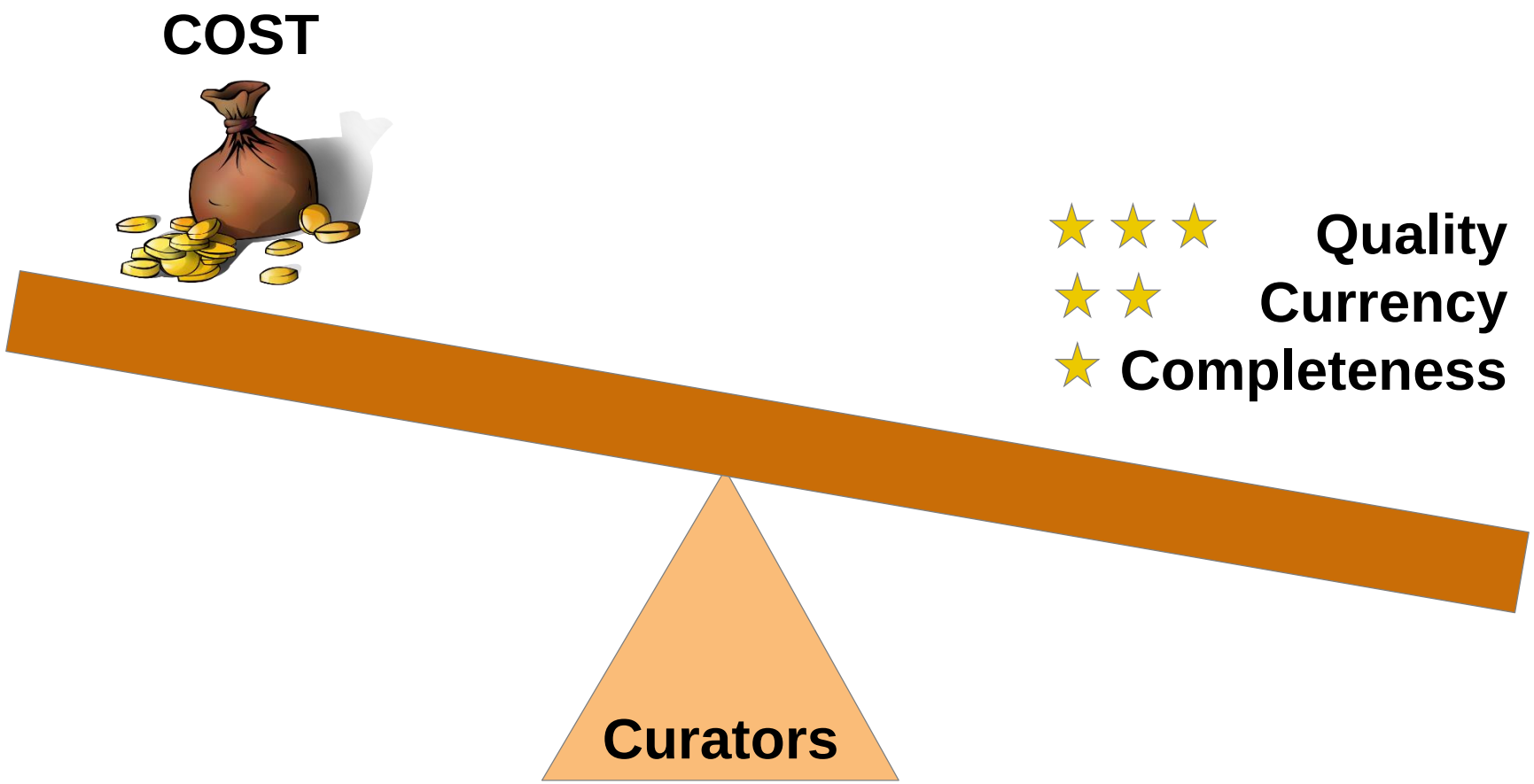
- **There is significant cost to setting up the experiment**
  - We plan to make the framework open source
  - But it still is taking ~2-3 staff months to prepare the task
- **Quality depends on the task**
  - Accuracy not comparable to expert curation for gene-mutation relations (yet)
  - However, drug indication labeling is fast and accurate
- **Method can produce good quality training corpora**
  - Which can lead to better automated extraction systems
- **Can more complex relations be captured?**
  - So far, only simple (binary) relations captured for short pieces of text

The Possibilities of Hybrid Curation

- **Next steps**
  - Improve the entity extractors for better recall
  - Add expert curator review step to improve quality
- **Explore additional strategies**
  - Dynamic Turking
  - Use an automated extractor as an extra curator
- **"Assembly-line" curation: curating in multiple easy steps**
  - Can tasks be chained to capture more complexity?
    - E.g., capturing gene-mutation-disease relations by first doing gene-mutation relations, then doing disease-mutation relations
  - Would expert curators want to use this approach for curation?
    - Or does it "dumb down" the decisions too much?

Lots of possibilities to explore!

How Can We Change the Balance?



Strategies to Reduce Cost

- **Reduce "completeness"**
  - Two-tiered curation a la UniProtKB
- **Use automated tools instead of manual curation**
  - Tools getting better (cf. BioCreative results), but…
  - Tools still don't do well on extracting complex relations
- **Get free curation**
  - Author and community-based curation (cf. FlyBase, SOL Genomics Network)
  - Crowdsourced curation, e.g., GeneWiki
- **Combine automated tools with expert curation**
  - Interactive tools – cf. BioCreative Interactive Track
  - Hybrid curation

Some Parting Questions

- **How much is curation worth – and to whom?**
  - What is a good "business model" for curation?
  - What is the value of a curated resource?
- **Can we couple publishing with curation?**
  - Tried in BioCreative II.5 with Elsevier and Federation of European Biochemical Societies (FEBS) Letters
    - Authors, automated systems, curators added value
    - Automated systems were better at certain aspects
    - This suggests a hybrid approach may be promising

Cost of curating an article (~1 staff hr)
is only a small fraction of the cost of
the experiment and writing the article (~1 staff yr)

Acknowledgements

- **Collaborators on hybrid curation experiments:**

Organization	Collaborators
MITRE	John Aberdeen, John Burger, Lynette Hirschman, David Tresner-Kirsch, Ben Wellner
NCBI	Ritu Khare, Zhiyong Lu, Chih-Hsuan Wei
University Maryland Baltimore County	Rajashree Mishra, Maricel Kann
Stanford University	Emily Doughty

- **Collaborators on De-identification**
  - David S. Carrell, David J. Cronkite (Group Health Cooperative, Seattle)
  - Bradley A. Malin (Vanderbilt); John Aberdeen (MITRE)
- **Curators/annotators who provided data**
  - Martha Palmer, University of Colorado
  - Brett South, Salt Lake City VA
  - Carolyn Mattingly, AP Davis, Tom Wiegers, Comparative Toxicogenomic Database

Back Up

Gene and Mutation Extraction

Element	Gold std	Extracted	Correct	Precision	Recall
Genes	246	582	222	0.381	0.902
Mutations	452	497	395	0.795	0.874
Gene–mutation	444	1078	374	0.347	0.842

Some Dimensions of Cost

- **Quality requirements:**
  - Precision/recall, reproducibility
- **Curators/annotators:**
  - Recruiting, training, number of staff, expertise
- **Value of information:**
  - Cost of errors (false positives, false negatives)
- **Scalability: speed, throughput**
  - What tools to insert where?
- **Maintenance of pipeline/workflow**

These dimensions are interrelated and conditioned by the use case:

More data vs. cheaper data vs. better data
Greater curator expertise vs. cheaper curation
One-off annotation effort vs. steady-state maintenance

Decreasing the Cost of Curation

- **What can be automated?**
  - Triage?
    - May not be the bottleneck, but curators are very interested
  - Indexing?
    - Automated tools can help with linkage to standard nomenclatures, external resources (BioCreative II.5)
  - Extraction?
    - Least reliable, but curators interested in interactive tools for full text
- **Role of the curator(s)**
  - Redundancy?
    - Value of reduced error rate vs. cost?
  - Expertise?
    - Experts are more expensive than non-experts; community curation?
  - Recruiting and training of curators?
    - Amortize training costs by retaining curators – vs. minimal training (e.g., crowdsourcing)

Currency and Throughput Requirements

- **BioCuration for literature curation**
  - Goal: keep up to date with new publications
  - Secondary goal: curate backlog of older publications
- **Linguistic annotation for machine learning systems**
  - Training data: Machine-learning based systems need annotated training data
    - Depending on variability: need 100's to 1000's of exemplars of each annotated data type
    - Quantity is more important than quality: single annotation may be good enough; crowd sourcing may be good enough
  - Evaluation data: needs to be high quality

Research Hypotheses

- **We can create a workflow for cost-effective curation of biomedical data to identify clinically relevant relations**
  - Using automated extraction of biological entities
  - Combined with human curation of relations
- **We can use crowdsourcing to obtain reliable (aggregate) judgments**
  - Even when the task appears to require domain-expertise

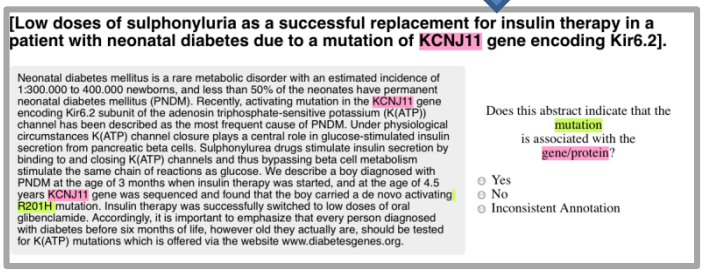
This hybrid approach can help to break the Curation Bottleneck

Evaluating Performance in Different Dimensions

	Displayed Surface Text (Turker View)	Concepts (Database View)
Entities	1. Gene spans and mutation spans in text	2. EntrezGene IDs; Mutation triple
Relations	3. Judgment on entity spans in context	4. Tuple of ⟨gene ID, mutation triple⟩

Gene, mutation recall, precision

Turker view


[Low doses of sulphonyluria as a successful replacement for insulin therapy in a patient with neonatal diabetes due to a mutation of KCNJ11 gene encoding Kir6.2].

Neonatal diabetes mellitus is a rare metabolic disorder with an estimated incidence of 1:300.000 to 400.000 newborns, and less than 50% of the neonates have permanent neonatal diabetes mellitus (PNDM). Recently, activating mutation in the KCNJ11 gene encoding Kir6.2 subunit of the adenosin triphosphate-sensitive potassium (K(ATP)) channel has been described as the most frequent cause of PNDM. Under physiological circumstances K(ATP) channel closure plays a central role in glucose-stimulated insulin secretion from pancreatic beta cells. Sulphonylurea drugs stimulate insulin secretion by binding to and closing K(ATP) channels and thus bypassing beta cell metabolism stimulate the same chain of reactions as glucose. We describe a boy diagnosed with PNDM at the age of 3 months when insulin therapy was started, and at the age of 4.5 years KCNJ11 gene was sequenced and found that the boy carried a de novo activating R201H mutation. Insulin therapy was successfully switched to low doses of oral glibenclamide. Accordingly, it is important to emphasize that every person diagnosed with diabetes before six months of life, however old they actually are, should be tested for K(ATP) mutations which is offered via the website www.diabetesgenes.org.

Does this abstract indicate that the mutation is associated with the gene/protein?

- Yes
- No
- Inconsistent Annotation

Eval against Gold Std view

pmid	Curation Code	Disease	wtaa	mtaa	pos	genes	geneid	type
1302001	8	AUTISTIC DISORDER	S	P	413	ASL	435	PROTEIN
1382850	8	PROSTATIC NEOPLASMS	Q	L	61	H-RAS	3265	PROTEIN
1565474	7	BREAST NEOPLASMS	A	G		P53	7157	DNA
1631125	3	PROSTATIC NEOPLASMS	V	M	730	AR	367	PROTEIN
1631125	3	PROSTATIC NEOPLASMS	G	A		AR	367	DNA

MITRE

Summary of Expt 2 Results By Quadrant

	Displayed Surface Text	Concepts
Entities: Genes; Mutations	1. NA (No Gold Std)	2. Genes P 38% R 90% Mutations P 80% R 87% Gene-Mut P 35% R 84%
Relations: Gene-Mutation	3. Accuracy: 91% Precision: 84% Turker Recall: 95%	4. Accuracy: 85% / 90%* Precision: 72% / 82% Turker Recall: 94% / 94% End-to-End Recall: 79% / 71%

- **Quad1: No gold standard**
- **Quad 2: Concept level entities: gene ID, mutation triple**
  - Recall of 84% for showing correct gene-mutation pairs to Turkers
- **Quad 3: (Estimated) Turker results: 91% accuracy**
- **Quad 4: Concept relations: Precision 82% @ 71% recall****

*Col 1: All HITS; **Col 2: Only HITs with local position for mutations

MITRE