

Using Natural Language Processing to Uncover Signals of Mental State

(accompanied by a brief rant about data)

Philip Resnik
University of Maryland
resnik@umd.edu

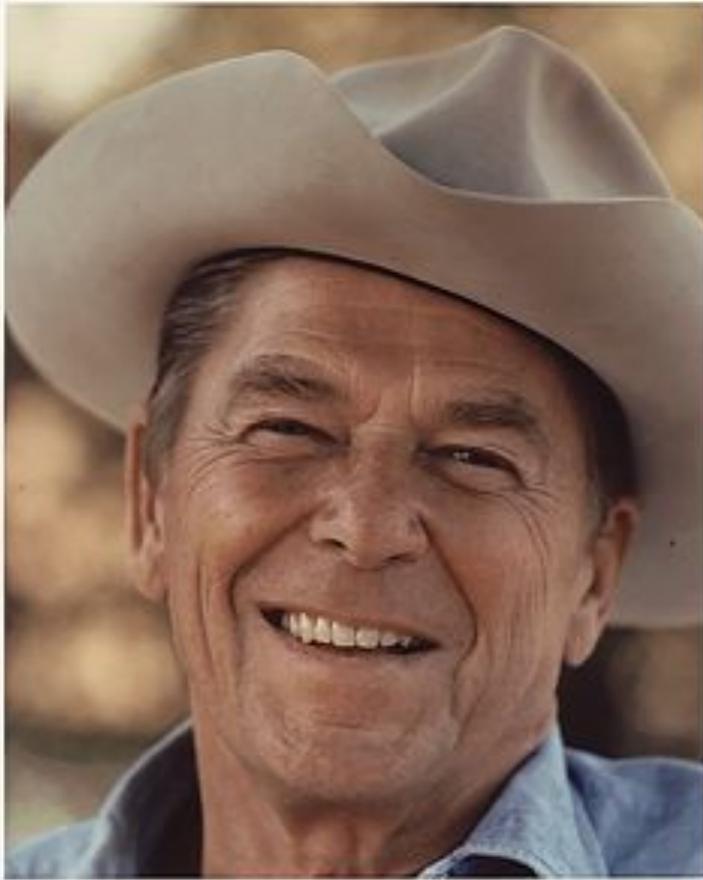
NLP State of Science Conference
Veteran's Administration
September 9, 2015



“Magic is a rich and largely untapped source of insight into perception and awareness. Insofar as the understanding of behaviour and perception goes, there are specific cases in which the magician's intuitive knowledge is superior to that of the neuroscientist.”

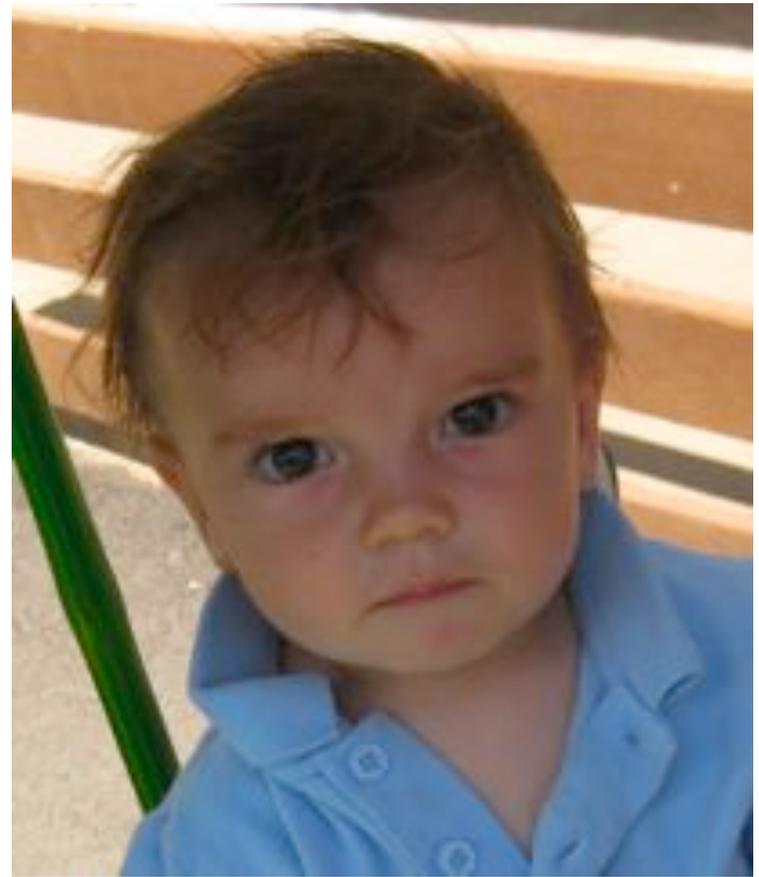
“Mistakes were made.”

Ronald Reagan, January 27, 1987



“My toy broke.”

Jay Resnik, frequently





Kentucky clerk walks free

By Emma Margolin



<p>HUMANITARIAN CRISIS Migrants overwhelm European nations Watch now</p> 	<p>WOMEN'S RIGHTS Is Clinton's message breaking through? What's working</p>	<p>NOW ON SHIFT Sports Matters: Legalized sports gambling Watch now</p> 
---	--	--

DAVIS OUT OF JAIL: Ky. clerk who denied gay marriage licenses is freed



KENTUCKY CLERK KIM DAVIS, who refused to issue marriage licenses to gay couples, was released from jail Tuesday and greeted by a crowd of cheering supporters, including GOP presidential candidate Mike Huckabee (left).

- **READ THE OFFICIAL ORDER:** U.S. District Court order releasing Kim Davis from custody
- **VIDEO:** Kim Davis' lawyer: 'She has no regrets'
- **VIDEO:** Davis joined by attorney, Huckabee at jail release
- **VIDEO:** Davis' attorney says last week's licenses are void



CONCERNS CONFIRMED Intel review backs claim Clinton emails 'top secret'

- **MAKING IMPROVEMENTS:** Kerry taps State Dept 'transparency' czar to oversee records
- **MEDIA BUZZ:** Why Hillary won't apologize for email fiasco
- **VIDEO:** Clinton says no email apology



VICTORY FOR COAL Colo. mine avoids closure over group's green gambit

- **VIDEO:** EPA on the hot seat after mine spill disaster
- **VIDEO:** Clinton doubles down on Obama's green energy agenda



MOTORCYCLE BAR BLAZE 'World's largest biker bar' burns to the ground in SD

- **DEADLY BLAZE:** Fire at popular nightclub in Cambodia's capital kills 5 women



"walked free"

About 339,000 results (0.69 seconds)

Accused of sex acts with children, he walked free. Here's why:
www.news-leader.com/...walked-free/7172156... Springfield News-Leader - 4 days ago - The case against John C. Heisler relied on witness testimony that wasn't always clear.

walked free - French translation - bab.la English-French ...
en.bab.la | bab.la Dictionary > English-French >
 Translation for 'walked free' in the free French dictionary. More French translations for: to free, free, walked.

The Cannibal that Walked Free (TV Movie 2007) - IMDb
www.imdb.com/title/tt1145515/ Internet Movie Database >
 The Cannibal that Walked Free (also known as Cannibal Superstar) is a British documentary produced by Visual Voodoo for Channel Five which explores ...

Florida's youngest convicted murderer walked free Tuesday ...
www.orlandoweekly.com/.../floridas-youngest-convicte... Orlando Weekly - Jul 28, 2015 - Almost 16 years after 12-year-old Curtis Jones and his 13-year-old sister were sentenced to prison for shooting and killing their father's ...

DAVIS OUT OF JAIL: Ky. clerk who denied gay marriage licenses is freed



KENTUCKY CLERK KIM DAVIS, who refused to issue marriage licenses to gay couples, was released from jail Tuesday

"refused to" -gay

About 91,900,000 results (0.44 seconds)

Refuse - definition of refuse by The Free Dictionary
www.thefreedictionary.com/refuse >
 Refuse implies determination and often brusqueness: "The commander ... refused to discuss questions of right" (George Bancroft). "I'll make him an offer he can't ...
 Refuse - Refuse collector - Refuse heap - Refusenik

refuse - Dictionary Definition : Vocabulary.com
www.vocabulary.com/dictionary/refuse >
 France has conducted airstrikes against Islamic State's forces in Iraq, but has previously refused to hit the radical Islamist group in its Syrian home territory.

Apple refused to wiretap an iMessage account for the ...
www.theverge.com/.../apple-wiretap-imessage-justice-departm... The Verge - 2 hours ago - For years, Tim Cook has been telling users that iMessage's encryption makes it impossible to wiretap — and now, the Justice Department ...

12 Homeowners Who Refused To Be Forced Out Of Their ...
www.littlebudha.com/12-resilient-underdogs-refused-forced-homes/ >
 Everyone loves a good underdog story, be that the American "Miracle" Olympic hockey team versus the old Soviet Union's, David versus Goliath, or the 99% ...

Multi-level modeling

[The] press may not be successful much of the time in telling people what to think, but it is stunningly successful in telling its readers *what to think about*. The world will look different to different people depending on the map that is drawn for them by writers, editors, and publishers of the paper they read."

Cohen, B.C. (1963). *The press and foreign policy*. Princeton. (emph. added)

What framing does is to "select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described."

Entman, R.M. (1993). "Framing: Toward clarification of a fractured paradigm". *Journal of Communication* 43 (4): 51–58.

Latent Dirichlet Allocation (LDA), Blei et al. (2003)

For every document, pick the mixture of topics that will be used.

For every word position, pick a topic for that word.

Prior probabilities for topic-word distributions

Prior probabilities for document-topic distributions

Generate a word associated with that topic.

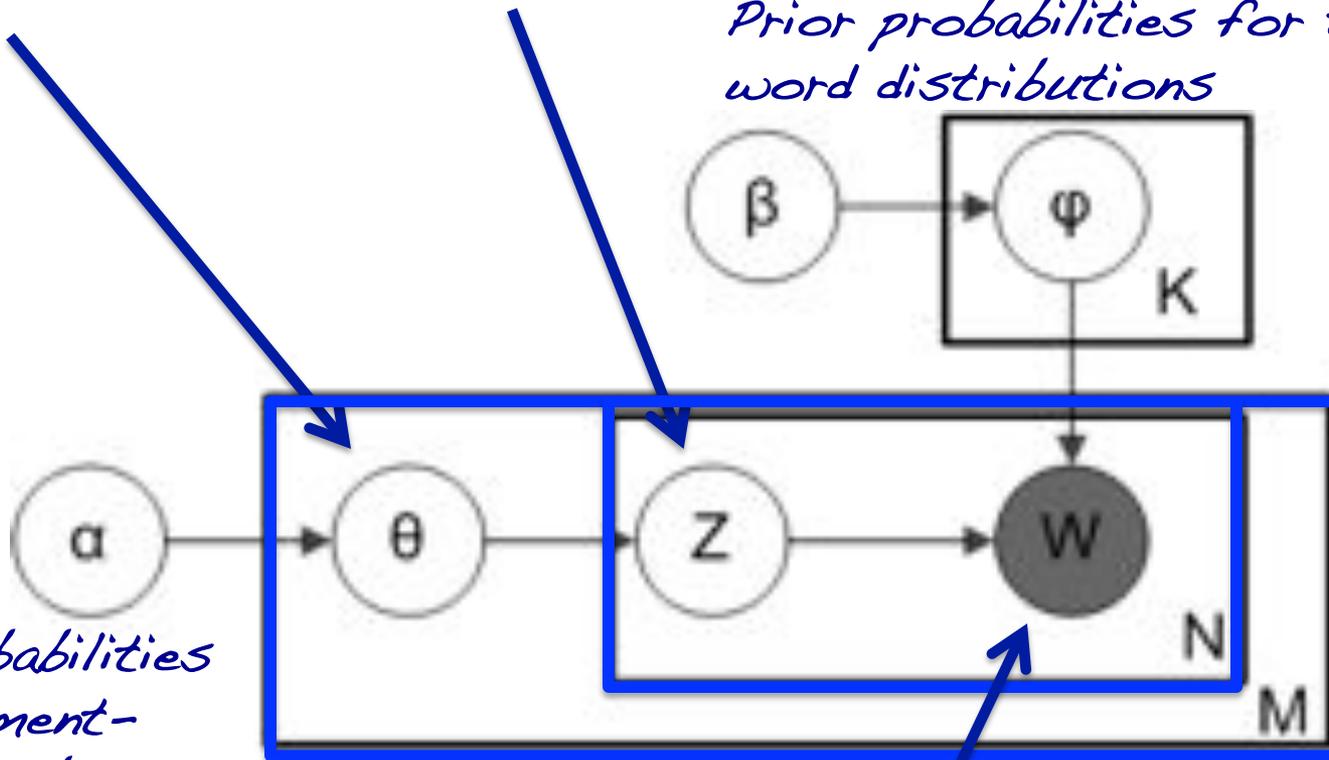
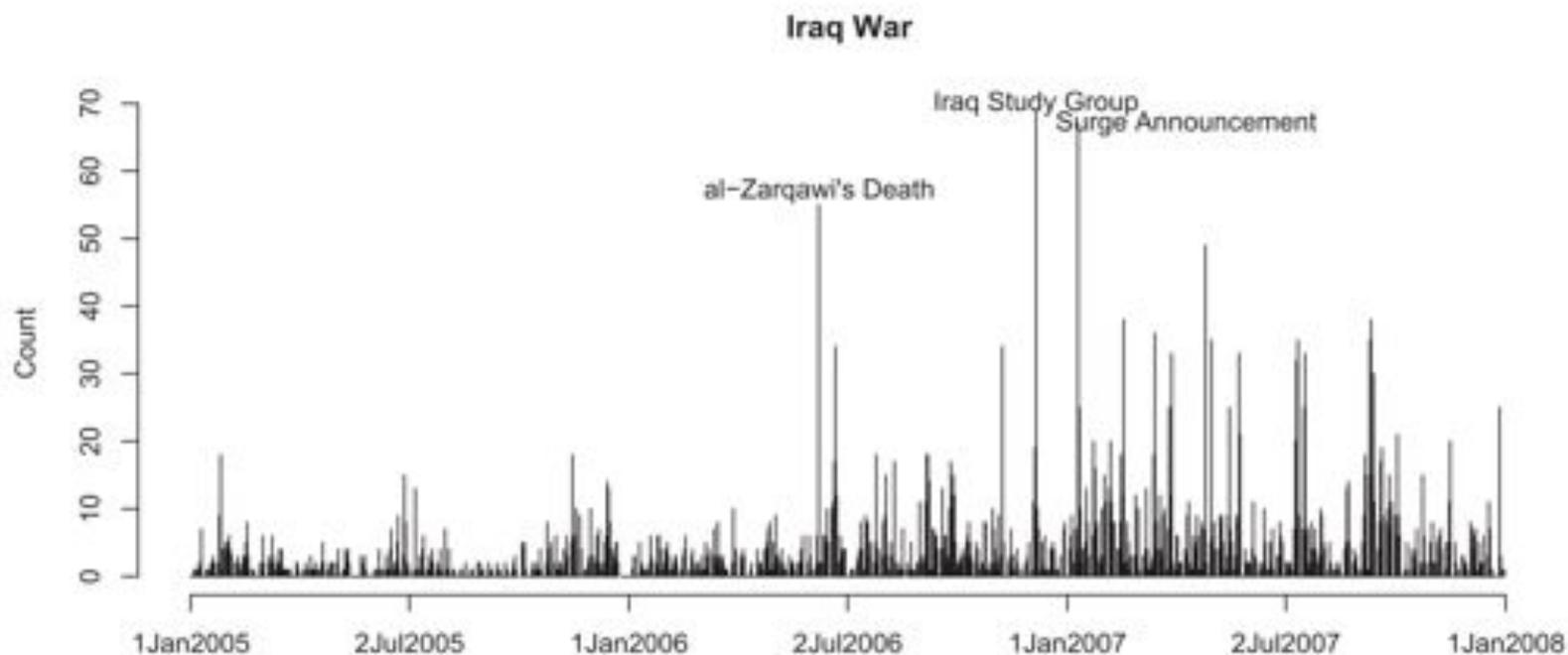


Table 2 Ten most discussed topics

<i>Label</i>	<i>Identifying stems</i>	<i>% Press releases</i>
Appropriations/grants	fund,project,000,million,water,transport,develop,improv,airport,citi	8.6
Honorary	honor,servic,school,serv,american,veteran,academi,famili,student,world	8.2
Iraq war	iraq,troop,war,iraqi,american,militari,polit,secur,support,countri	6.6
Health grants	health,program,educ,children,school,fund,student,care,servic,000	6.3
Homeland security	secur,homeland,port,border,depart,fund,guard,air,servic,transport	5.3
Judicial nominations	court,vote,justic,american,judg,case,hous,congress,constitut,protect	4.8
Hurricanes/disasters	disast,assist,hurrican,fema,flood,damag,fund,katrina,storm,declar	4.5
Taxes	tax,american,budget,social,secur,wage,famili,worker,increas,benefit	4.4
Defense projects	million,defens,fund,air,militari,base,facil,guard,armi,project	4.2
Health policy	health,care,drug,medicar,senior,prescript,plan,medic,program,cost	3.8



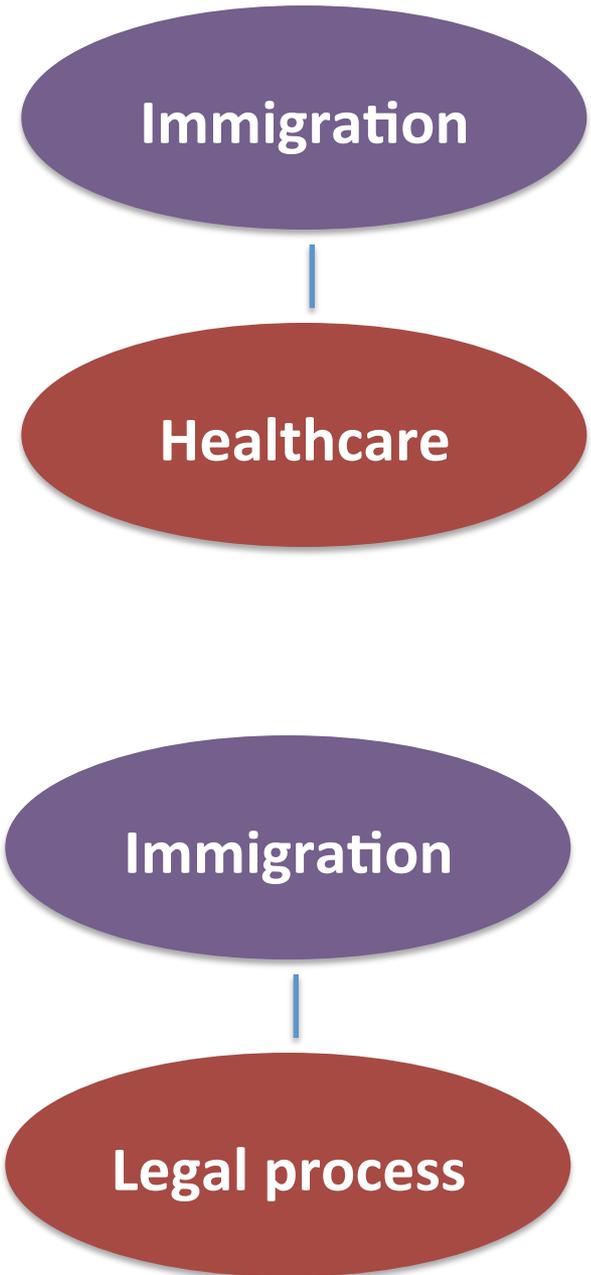


Swine flu surfaces at Texas-Mexico border among illegals...
Washington Times - 10 hours ago
The first case of swine flu has been found among the scores of **illegal children** who have been crossing into America at the Texas-Mexico ...

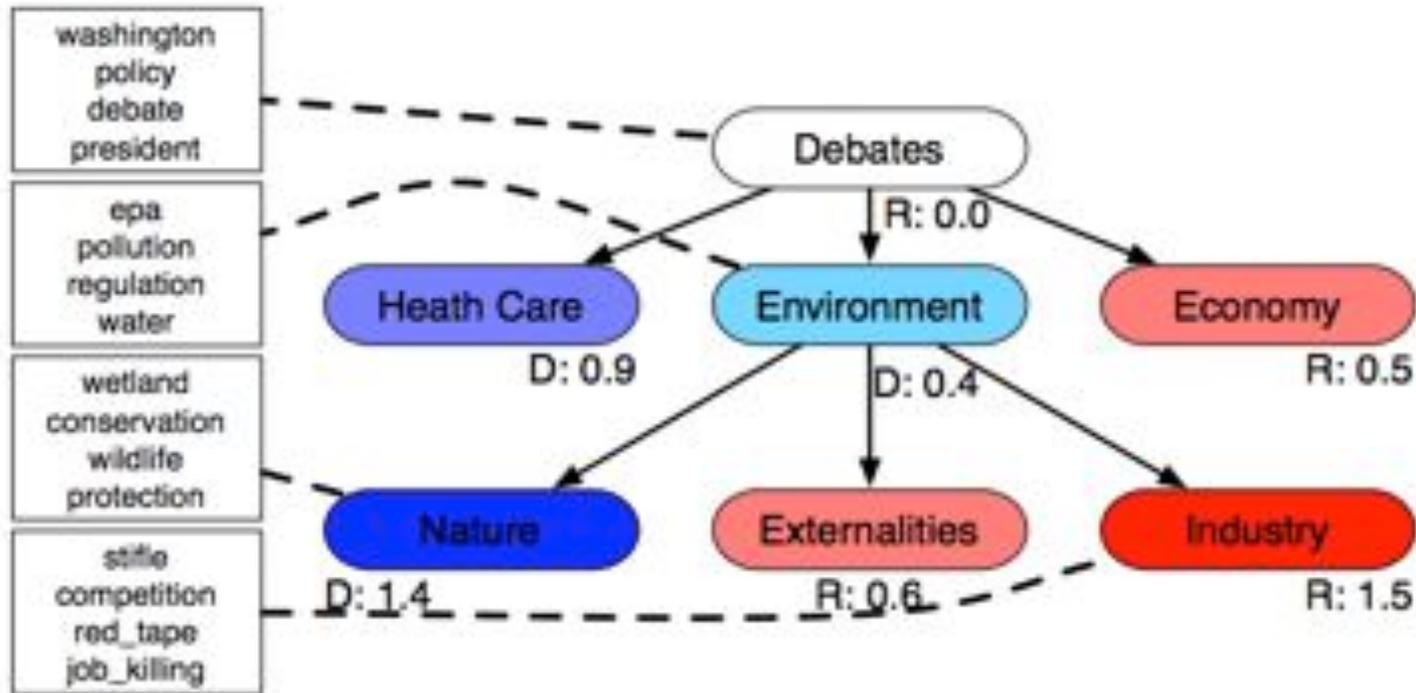
The swine flu finding only fuels fears from law enforcement along the border who say the **illegal immigrants are not being properly screened for diseases and contagious sicknesses** before moving along to other facilities for holding across the nation.

“Some of the children who have come to this country may not have a valid legal basis to remain, but some will. Yet, **it is virtually impossible for a child to assert a valid claim under immigration law in the absence of legal representation.** ... It is a fantasy to believe that unrepresented children have a fair shot in an immigration proceeding”

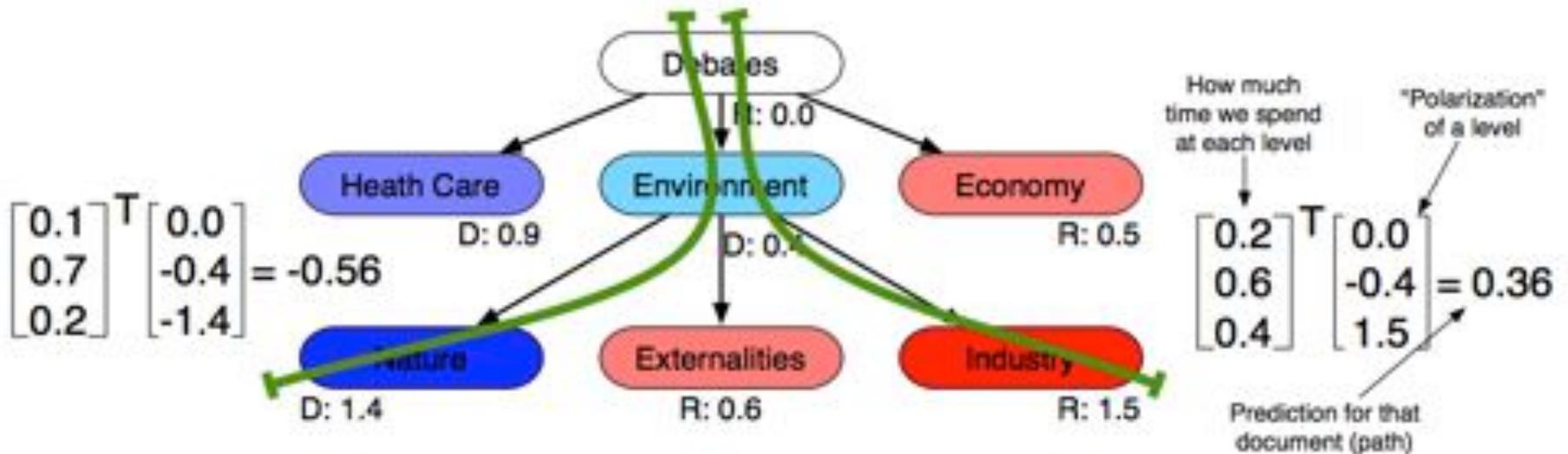
--Rep. Hakeem Jeffries, D-N.Y.



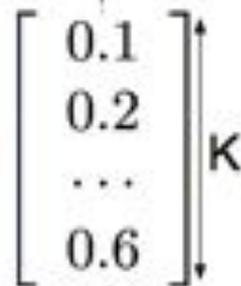
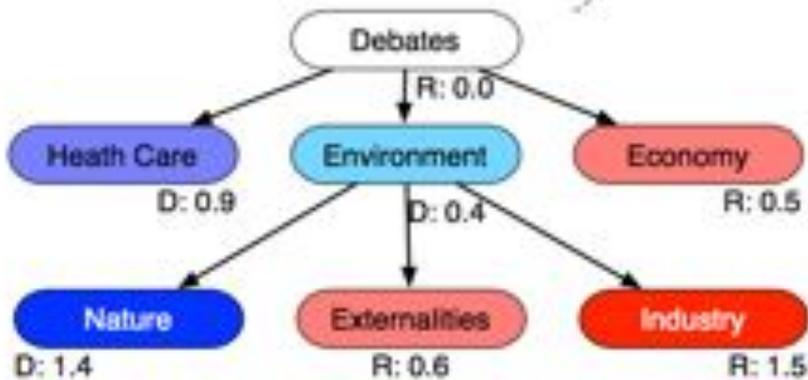
Supervised Hierarchical LDA (SHLDA)



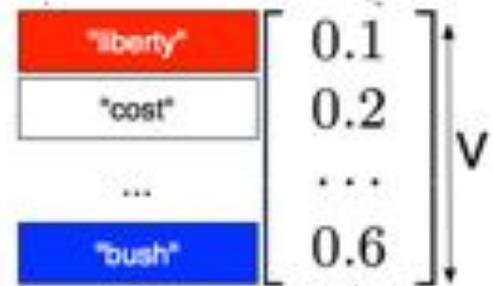
Supervised Hierarchical LDA (SHLDA)



$$y_d \sim \mathcal{N}(\eta^T \bar{z}_d + \tau^T \bar{w}_d, \rho)$$

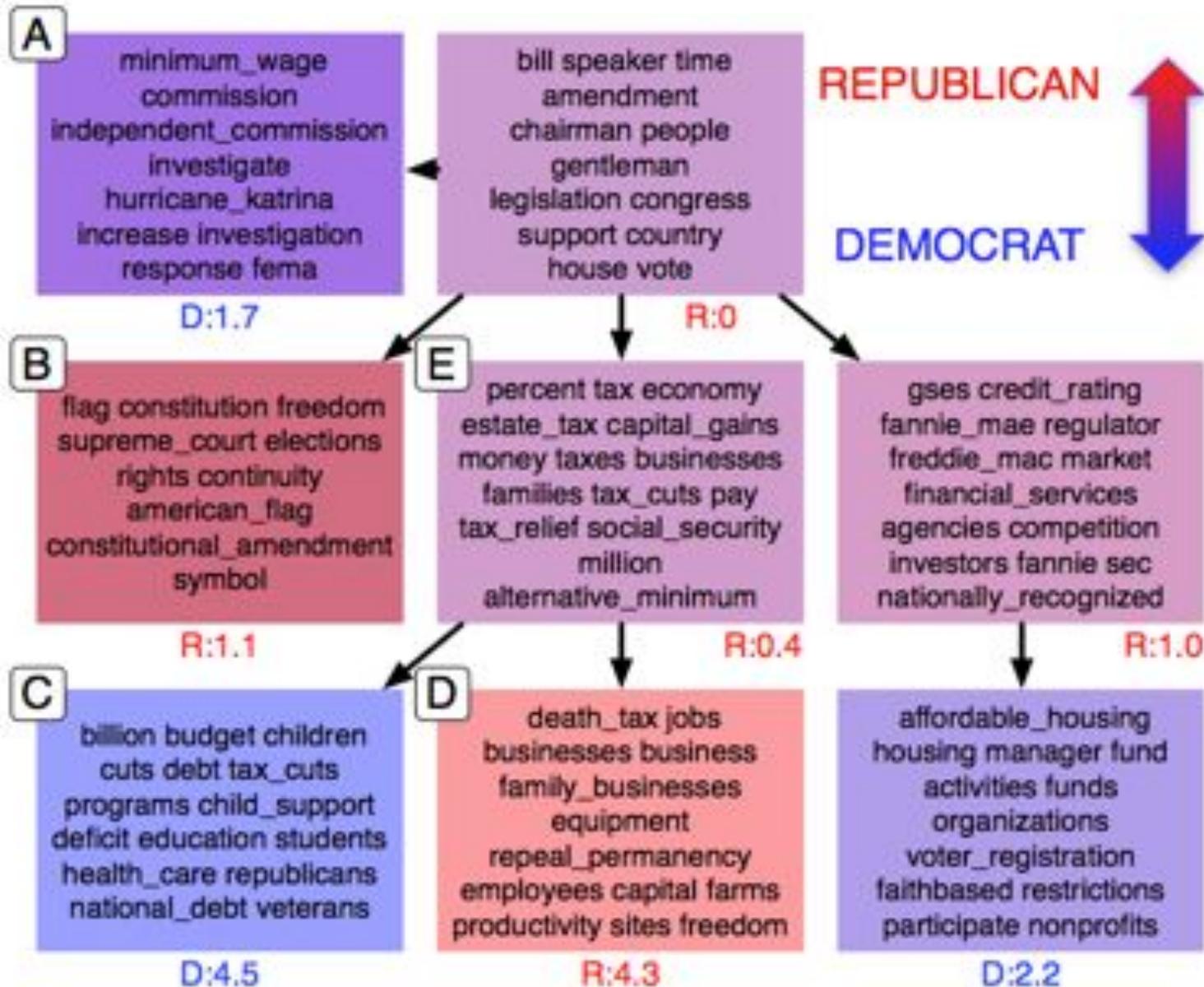


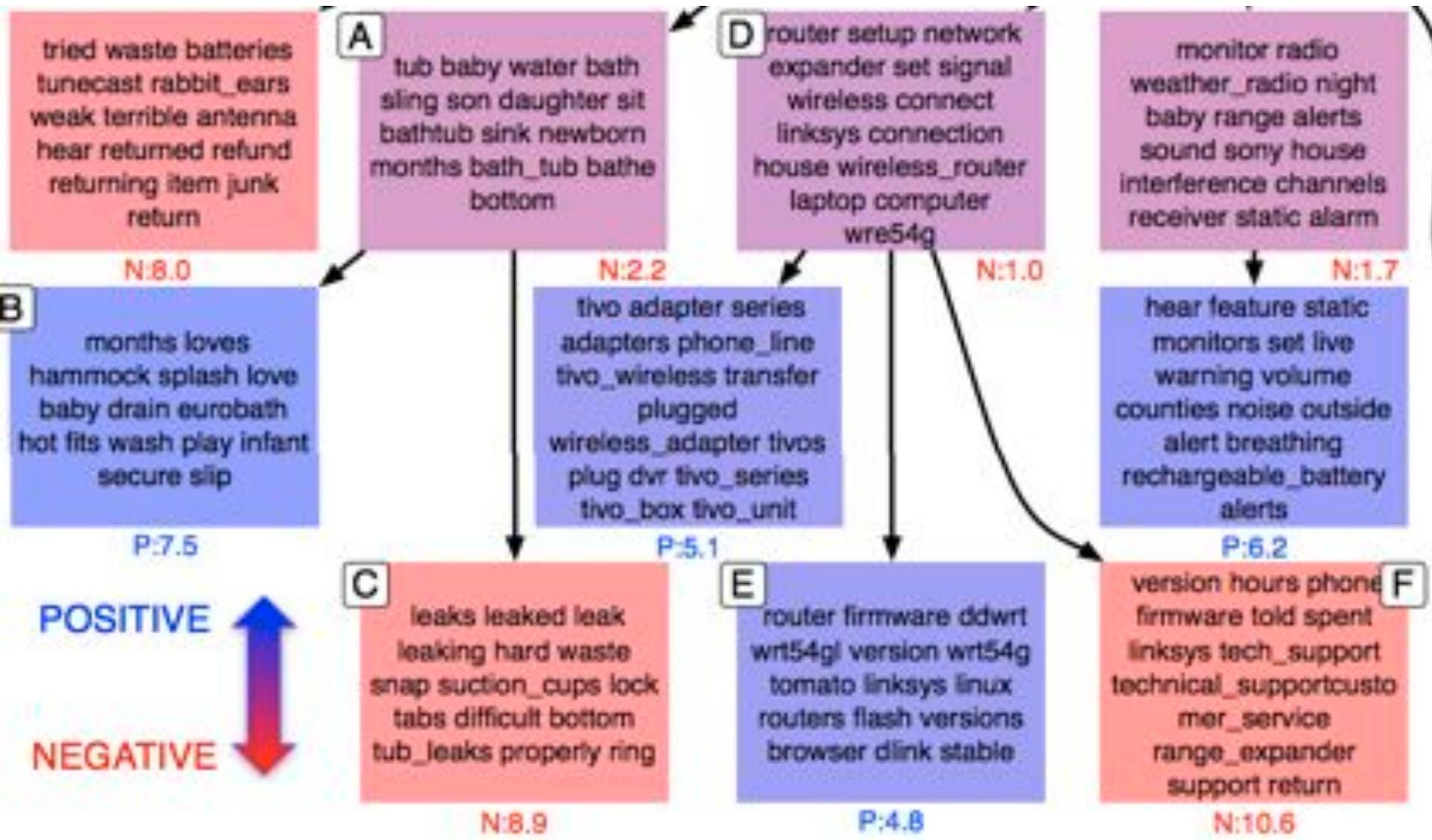
Document empirical distribution over nodes



Document normalized word vector

Supervised Hierarchical LDA (SHLDA)





Models	Floor Debates				Amazon Reviews		Movie Reviews	
	House-Senate		Senate-House		PCC ↑	MSE ↓	PCC ↑	MSE ↓
	PCC ↑	MSE ↓	PCC ↑	MSE ↓				
SVM-LDA ₁₀	0.173	0.861	0.08	1.247	0.157	1.241	0.327	0.970
SVM-LDA ₃₀	0.172	0.840	0.155	1.183	0.277	1.091	0.365	0.938
SVM-LDA ₅₀	0.169	0.832	0.215	1.135	0.245	1.130	0.395	0.906
SVM-VOC	0.336	1.549	0.131	1.467	0.373	0.972	0.584	0.681
SVM-LDA-VOC	0.256	0.784	0.246	1.101	0.371	0.965	0.585	0.678
MLR-LDA ₁₀	0.163	0.735	0.068	1.151	0.143	1.034	0.328	0.957
MLR-LDA ₃₀	0.160	0.737	0.162	1.125	0.258	1.065	0.367	0.936
MLR-LDA ₅₀	0.150	0.741	0.248	1.081	0.234	1.114	0.389	0.914
MLR-VOC	0.322	0.889	0.191	1.124	0.408	0.869	0.568	0.721
MLR-LDA-VOC	0.319	0.873	0.194	1.120	0.410	0.860	0.581	0.702
SLDA ₁₀	0.154	0.729	0.090	1.145	0.270	1.113	0.383	0.953
SLDA ₃₀	0.174	0.793	0.128	1.188	0.357	1.146	0.433	0.852
SLDA ₅₀	0.254	0.897	0.245	1.184	0.241	1.939	0.503	0.772
SHLDA	0.356	0.753	0.303	1.076	0.413	0.891	0.597	0.673

For another related model see Nguyen et al., Tea Party in the House: A Hierarchical Ideal Point Topic Model and Its Application to Republican Legislators in the 112th Congress. Association for Computational Linguistics, Beijing, July 2015.

(Rant coming, not there yet.)

What framing does is to "select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described."



Entman, R.M. (1993). "Framing: Toward clarification of a fractured paradigm". *Journal of Communication* 43 (4): 51–58.

What framing does is to "select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described."



Entman, R.M. (1993). "Framing: Toward clarification of a fractured paradigm". *Journal of Communication* 43 (4): 51–58.

“It is not pleasant to experience **decay**, to find yourself exposed to the ravages of an almost **daily** rain, and to know that you are **turning into** something feeble, that **more and more of you** will blow off with the first strong wind, making you **less and less**.

— Andrew Solomon, *The Noonday Demon: An Atlas of Depression*

“[Despair], owing to some evil trick played upon the sick brain by the inhabiting psyche, comes to resemble the diabolical **discomfort** of being **imprisoned** in a fiercely **overheated** room. And because no breeze stirs this cauldron, because there is no escape from this **smothering confinement**, it is entirely natural that the victim begins to think ceaselessly of oblivion.”

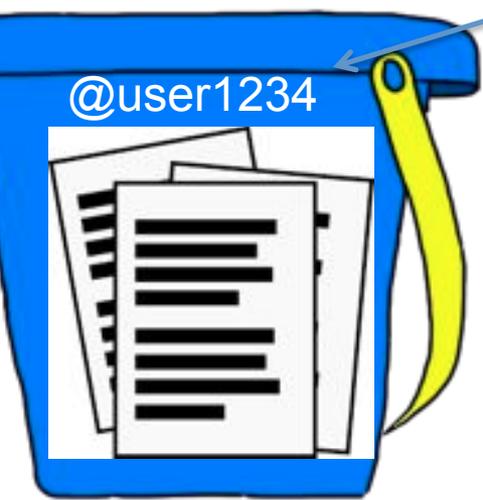
— Dr. Kay Redfield Jamison, *Night Falls Fast: Understanding Suicide*

“Ptds is like being **nailed to the cross** over and over and over until each body surface, organ and area has had its countless repetitive turns at **assaults**: has been **split and punctured** wide open and then left to **fry alive** in tossed blame, and another’s abandoned shame”

— PTSD discussion board

Notes	Valence	Regression value	Top 20 words
social engagement	p	-1.593	game play football team watch win sport ticket texas season practice run basketball lose so
social engagement	p	-1.122	music song listen play band sing hear sound guitar change remind cool rock concert voice
social engagement	p	-0.89	party night girl time fun sorority meet school house tonight lot rush drink excite fraternity
social engagement	p	-0.694	god die church happen day death lose doe bring care pray live plan close christian control
high emotional valence	e	-0.507	hope doe time bad wait glad nice happy worry guess lot fun forget bet easy finally suck fin
somatic complaints	n	-0.205	cold hot hair itch air light foot nose walk sit bear eye rain nice sound smell freeze weather
poor ego control; immature	n	0.177	yeah wow minute haha type funny suck hmm guess blah bore gosh ugh stupid bad lol hey
relationship issues	n	0.234	call talk miss phone hope mom mad love stop tonight glad dad weird stupid matt email any
homesick; emotional distress	n	0.34	home miss friend school family leave weekend mom college feel parent austin stay visit lot
social engagement	p	0.51	friend people meet lot hang roommate join college nice fun club organization stay social to
negative affect*	n	0.663	suck damn stupid hate hell drink shit fuck doe crap smoke piss bad kid drug freak screw cr
high emotional valence	e	0.683	life change live person future dream realize mind situation learn goal grow time past enjoy
sleep disturbance*	n	0.719	sleep night tire wake morning bed day hour late class asleep fall stay nap tomorrow leave n
high emotional valence	e	0.726	love life happy person heart cry sad day feel world hard scar perfect feeling smile care stro
memories	n	0.782	weird talk doe dog crazy time sad stuff funny haven happen bad remember day hate lot sca
somatic complaints*	n	0.805	hurt type head stop eye hand start tire feel time finger arm neck move chair stomach bother
anxiety*	n	1.111	feel worry stress study time hard lot relax nervous test focus school anxious concentrate pr
emotional discomfort	n	1.591	feel time reason depress moment bad change comfortable wrong lonely feeling idea lose ga
homesick; emotional distress*	n	2.307	hate doe sick feel bad hurt wrong care happen mess horrible stupid mad leave worse anymo

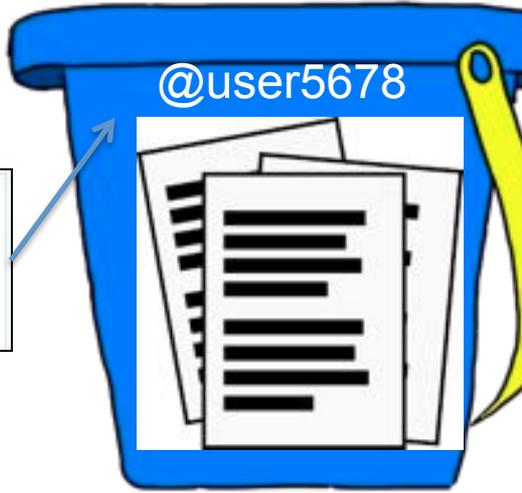
SLDA topics from undergraduate stream-of-consciousness essays identified by a clinician as most relevant for assessing depression. Supervision (regression) is based on Z-scored Big-5 scores for emotional instability (neuroticism).



“Depressed”

John Doe @johndoe Aug 25
Ethical dilemma, woman I’m dating doesnt know I’m diagnosed with depression and struggling with anxiety

Philip Resnik @psresnik · Jul 17
Interesting (though still small-N) study on patterns of cell phone use as a potential source of signal for depression jmir.org/2015/7/e175/



“Control”

(Rant coming up very, very soon.)

Mental health in social media

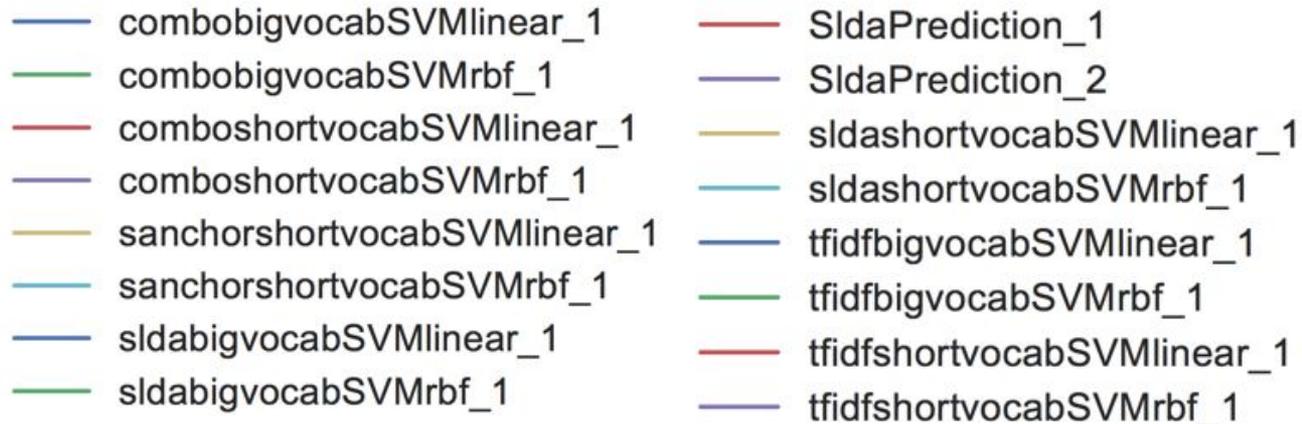
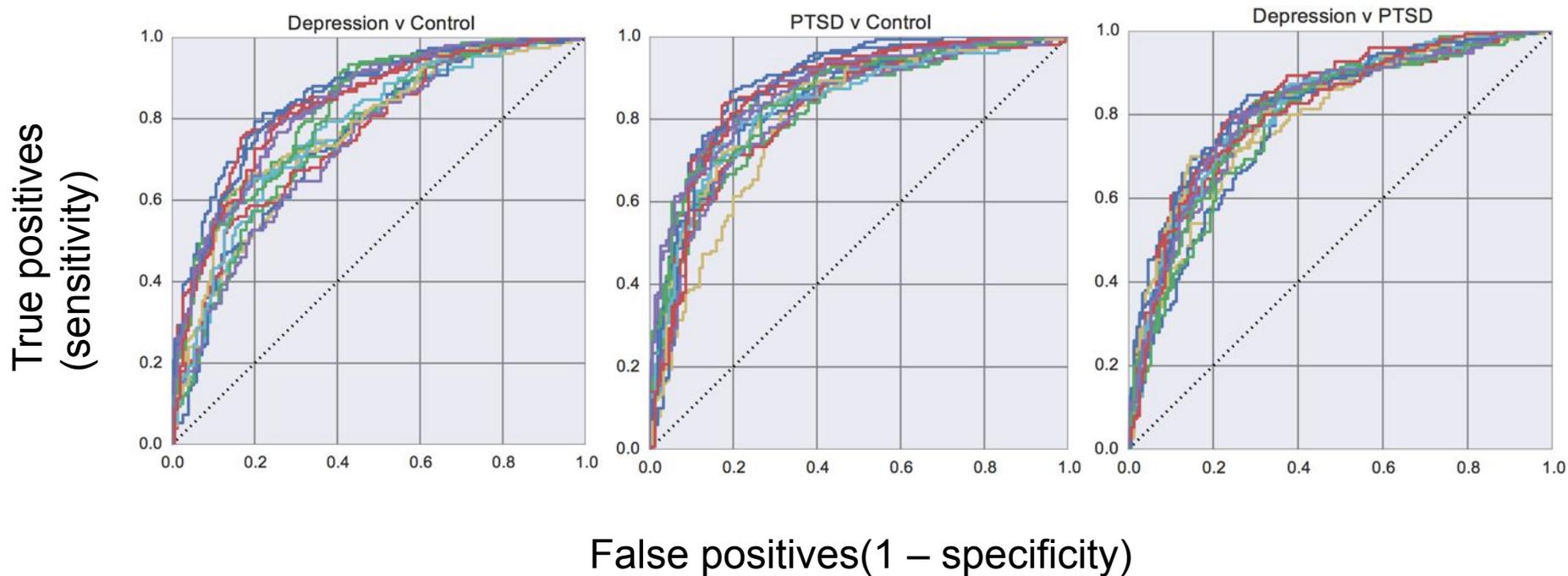
Regression value	Top 20 words
4.319	omg cry love gonna demi cute guy feel perfect meet idk tweet omfg pls god wanna song literally bye ily
4.318	people woman doe person human kid word read child understand happen world joke remember real reason write stop change wrong
4.29	fuck shit bitch smoke hate people drink gonna sex damn fuckin dick suck wtf weed life hell feel piss stupid
3.536	feel eat die fat cut hate lose people line cry stop body care cross friend sick hurt life scar start
3.394	home watch week time wait day bed hour cat tomorrow feel call morning friend hope leave buy sleep night ago
3.093	girl guy boy people friend cute mom wear hot hate school life wanna date picture talk boyfriend kiss literally pretty
2.78	week post baby inbox month hey day ago start pregnant feel time pain girl private boy bad doe period child
...	...
0.148	lmao lol talk girl lmfao text love tho baby miss bae phone wanna mad shit fuck call damn bitch oomf
0.062	hair buy nail love dress wear red color blue cute beautiful fall pink black eye flower shoe beauty pretty spring
-0.042	photo post facebook photoset share tumblr skinny picture time update tag pic life timeline day story repost month video challenge
-0.043	happy birthday love day hope guy reason babe miss start nice stop time life night literally bad alive world song
-0.066	EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI EMOJI girl EMOJI EMOJI EMOJI EMOJI love EMOJI EMOJI people EMOJI wanna
-0.083	game video iphone apple app play add youtube ipad google phone note update internet review free super galaxy playlist pro
...	...
-1.432	nigga shit bitch hoe bout gotta real wanna ima tho aint smh damn lil wit tryna money call cuz female
-1.594	sleep wanna feel gonna hate bed tire wake love day miss baby time people text hungry annoy hair bad nap
-1.837	guy love pic miss hey lav wait die plz hope smile tweet watch true wat soo fan sweet cont day
-1.959	school class tomorrow day college teacher homework study start test hate hour home math sleep people sit friday senior grade
-2.348	lol lmao money yea smh damn dat gotta yal ppl kid time dont dude remember jayz baby lil hell woman
-2.742	night tonight tomorrow time miss wait party weekend summer ready home drink hour week saturday excite friend gonna fun leave

Most extreme and neutral sLDA topics from Twitter dataset containing authors with self-reported depression (positive) and controls (negative).

Mental health in social media

-0.641	post facebook pretty comment doe bad tag yesterday move cool scan lady stuff nice l
-0.314	cool guy super dude stuff pretty nice play doe yep apparently sweet terrible pick
9.814	ugh cute picture shit tattoo guy suck makeup omg gonna stupid pierce nude free
...	...
1.165	night sleep late bed hour tomorrow morning wake bus start nap class alexis relax fall
-2.404	morning tomorrow excite wait friday night weekend saturday monday till sunda
1.561	sleep wake bed feel hour asleep nap fall awake mood finally lay wait sleepy hav

Selected hierarchical topics derived from Twitter training data using supervised nested latent Dirichlet allocation (SNLDA)



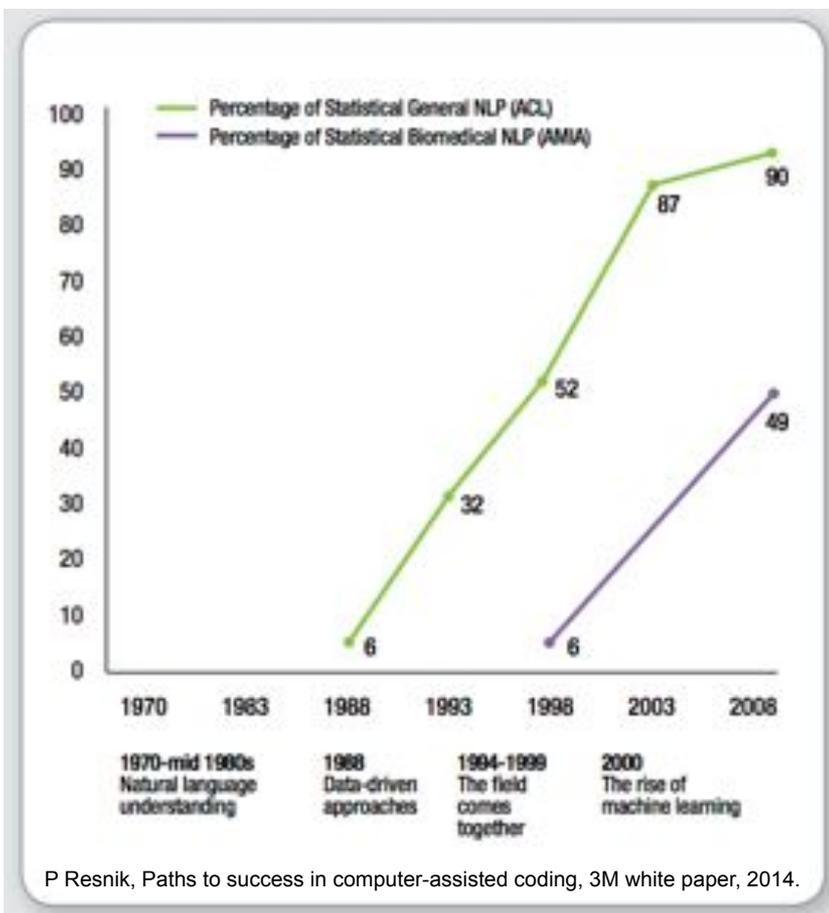
Resnik et al., Beyond LDA: Exploring Supervised Topic Modeling for Depression-Related Language in Twitter. NAACL Workshop on Computational Linguistics and Clinical Psychology, Denver, CO, June 2015.

Take-aways

- **Bayesian topic models** provide a way to uncover latent structure in text content
- **Hierarchical models** can capture not only topics, but how those topics are *framed* – an indication of underlying mental state
- **Integrating supervision** makes it possible to predict response variables of interest

(here is the rant I promised)

- The mantra of NLP progress is that *it's all about the data*.
- The clinical NLP data shortage is **fatally desperate**.



- **We absolutely must have *large scale, representatively variable* clinical data if we are to make rapid progress.**
- **RESEARCHERS CANNOT SOLVE THIS PROBLEM!!!**

Can you?

Collaborators and thanks

- Jordan Boyd-Graber
- Viet-An Nguyen
- Deborah Cai
- Amber Boydstun
- Rebecca Glazier
- Matt Pietryka
- Tim Jurka
- Stephan Greene
- Noah Smith
- Justin Gross
- Kris Miler
- Rebecca Resnik
- William Armstrong
- Leonardo Claudino
- Thang Nguyen
- Glenn Coppersmith
- Meg Mitchell
- Kristy Hollingshead
- Mark Dredze
- Jamie Pennebaker
- IARPA SCIL program
- NSF SOCS program

Thanks!



Linguistic Structured Sparsity in Text Categorization

Dani Yogatama and Noah A. Smith

Language Technologies Institute

Carnegie Mellon University

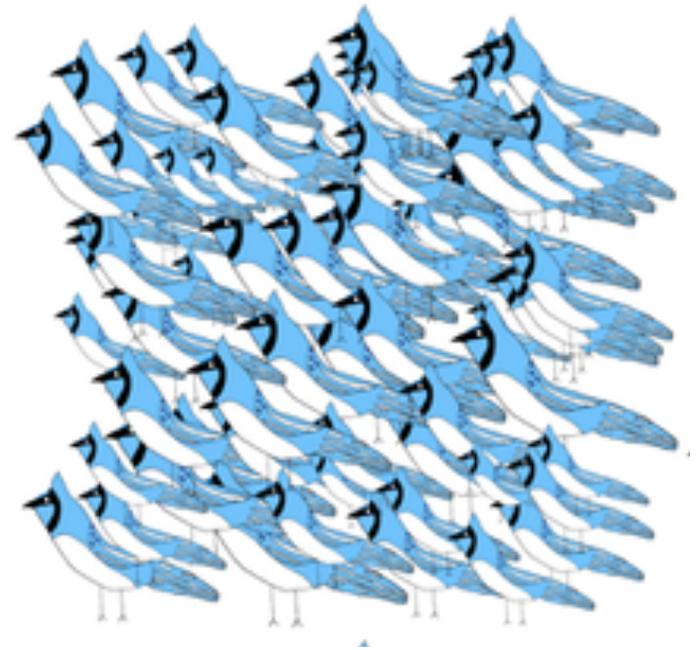
`{dyogatama,nasmith}@cs.cmu.edu`



Dani Yogatama

Summary

- Words of a feather (should) flock together
- Idea: use linguistic structure to define *feathers* (flocks) instead of *features*
- Math: sparse group lasso regularization
- Results: text classification (sentiment, forecasting, topic)



Text Classification

this film is one big joke : you have all the basics elements of romance (love at first sight , great passion , etc .) and gangster flicks (brutality , dagerous machinations , the mysterious don , etc.) , but it is all done with the crudest humor .

it ' s the kind of thing you either like viserally and immediately " get " or you don ' t .

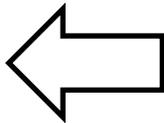
that is a matter of taste and expectations .

i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes .

the acting is very good , if a bit obviously tongue - in - cheek .

Bag of Words

1	acting
1	at
1	back
1	basics
1	big
1	bit
1	brutality
1	but
1	cheek
1	crudest
1	dagerous
	:
6	the
	:



this film is one big joke : you have all the basics elements of romance (love at first sight , great passion , etc .) and gangster flicks (brutality , dagerous machinations , the mysterious don , etc.) , but it is all done with the crudest humor . it ' s the kind of thing you either like viserally and immediately " get " or you don ' t . that is a matter of taste and expectations . i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes . the acting is very good , if a bit obviously tongue - in - cheek .

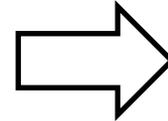
Bag of Words

1	acting
1	at
1	back
1	basics
1	big
1	bit
1	brutality
1	but
1	cheek
1	crudest
1	dagerous
	⋮
6	the
	⋮

Linear Classifier

1	acting	w_{acting}
1	at	w_{at}
1	back	w_{back}
1	basics	w_{basics}
1	big	w_{big}
1	bit	w_{bit}
1	brutality	$w_{brutality}$
1	but	w_{but}
1	cheek	w_{cheek}
1	crudest	$w_{crudest}$
1	dagerous	$w_{dagerous}$
	⋮	⋮
6	the	w_{the}
	⋮	⋮

$$\text{sign}(\mathbf{f}(\text{document}) \cdot \mathbf{w})$$


$$\hat{y}$$

Text is Not a Bag of Words!

- Sentences

this film is one big joke : you have all the basics elements of romance (love at first sight , great passion , etc .) and gangster flicks (brutality , dagerous machinations , the mysterious don , etc.) , but it is all done with the crudest humor .

it ' s the kind of thing you either like viserally and immediately " get " or you don ' t .

that is a matter of taste and expectations .

i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes .

the acting is very good , if a bit obviously tongue - in - cheek .

Text is Not a Bag of Words!

- Sentences
- Phrases

this film is one big joke : you have all the basics elements of romance (love at first sight , great passion , etc .) and gangster flicks (brutality , dagerous machinations , the mysterious don , etc.) , but it is all done with **the crudest humor** .

it ' s the kind of thing you either like viserally and immediately " get " or you don ' t .

that is a **matter of taste and expectations** .

i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes .

the acting is very good , if a bit obviously tongue - in - cheek .

Text is Not a Bag of Words!

- Sentences
- Phrases
- Fine-grained syntactic classes

this film is one **big** joke : you have all the **basics** elements of romance (love at **first** sight , **great** passion , etc .) and gangster flicks (brutality , **dagerous** machinations , the **mysterious** don , etc.) , but it is all done with the **crudest** humor .

it ' s the kind of thing you either like viserally and immediately " get " or you don ' t .

that is a matter of taste and expectations .

i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes .

the acting is very **good** , if a bit obviously tongue - in - cheek .

Text is Not a Bag of Words!

- Sentences
- Phrases
- Fine-grained syntactic classes
- Thematic topics

(and many more!)

this film is one big **joke** : you have all the basics elements of romance (love at first sight , great passion , etc .) and gangster flicks (brutality , dagerous machinations , the mysterious don , etc.) , but it is all done with the crudest **humor** .
it ' s the kind of thing you either like viserally and immediately " get " or you don ' t .
that is a matter of taste and expectations .
i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes .
the acting is very good , if a bit obviously **tongue** - in - **cheek** .

Learning the Weights \mathbf{w}

“fit the data”

(e.g., log-likelihood of y_n given d_n ,
hinge loss, ...)

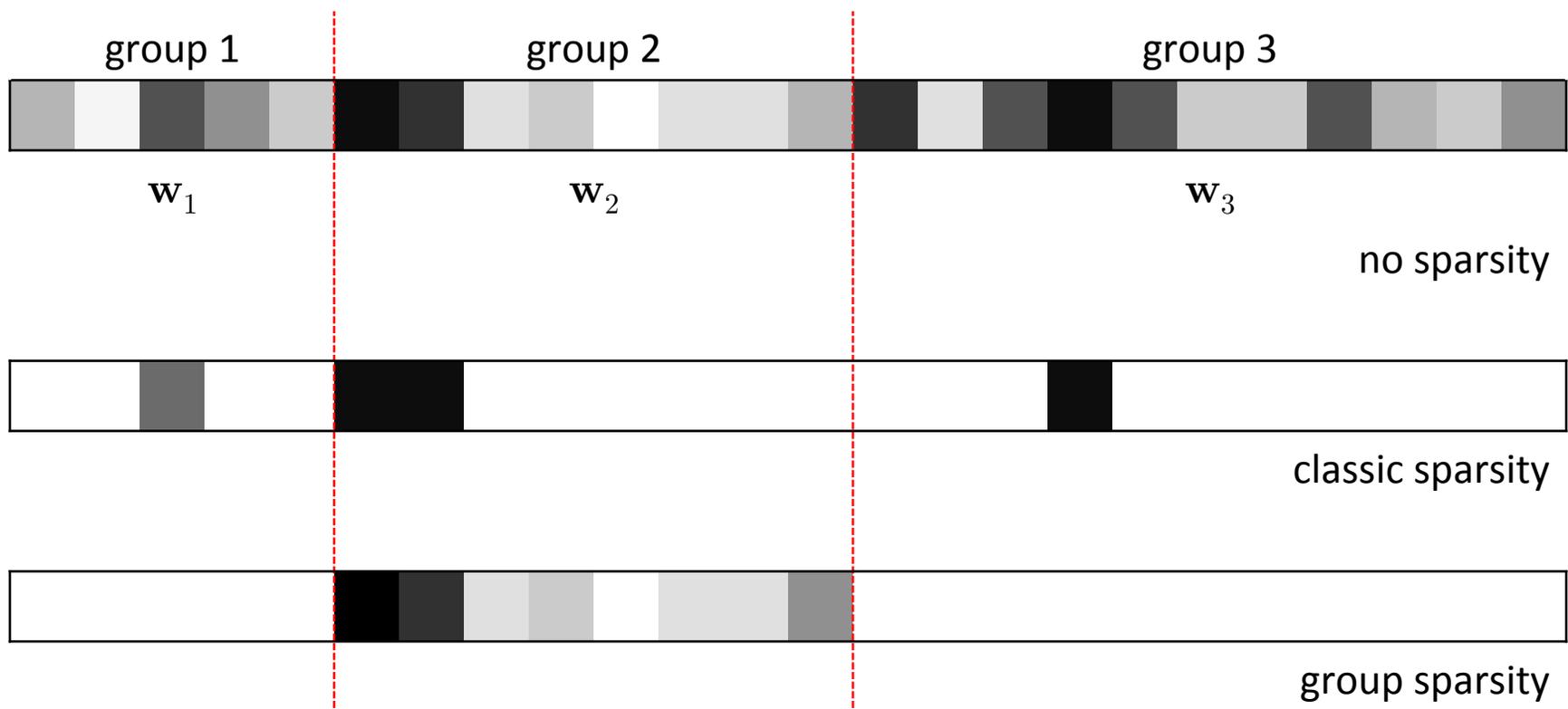
$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w}) + \underline{R(\mathbf{w})}$$

“generalize”

(e.g., $\lambda \|\mathbf{w}\|_2^2$;
 $\lambda \|\mathbf{w}\|_1$)

Group Lasso (Yuan & Lin '06)

$$R(\mathbf{w}) = \sum_g \lambda_g \|\mathbf{w}_g\|_2$$

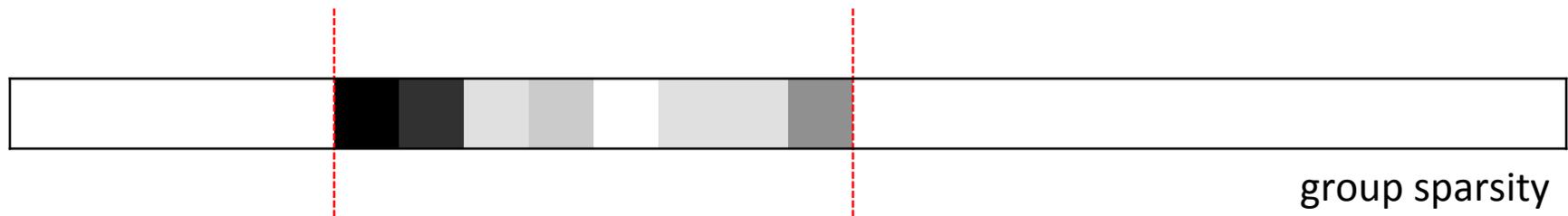


Group Lasso (Yuan & Lin '06)

$$R(\mathbf{w}) = \sum_g \lambda_g \|\mathbf{w}_g\|_2$$

In NLP:

- chunking and parsing (Martins et al., 2011)
- language modeling (Nelakanti et al., 2013)



Learning the Weights \mathbf{w}

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w}) + R(\mathbf{w})$$

Learning the Weights \mathbf{w}

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w}) + R(\mathbf{w})$$

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w})$$

$$\text{s.t. } R(\mathbf{w}) \leq \tau$$

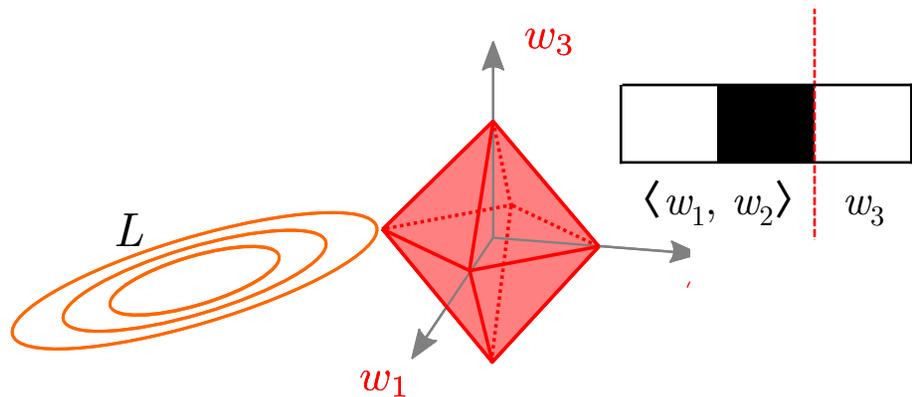
“Tikhonov” regularization



“Ivanov” regularization



Lasso vs. Group Lasso

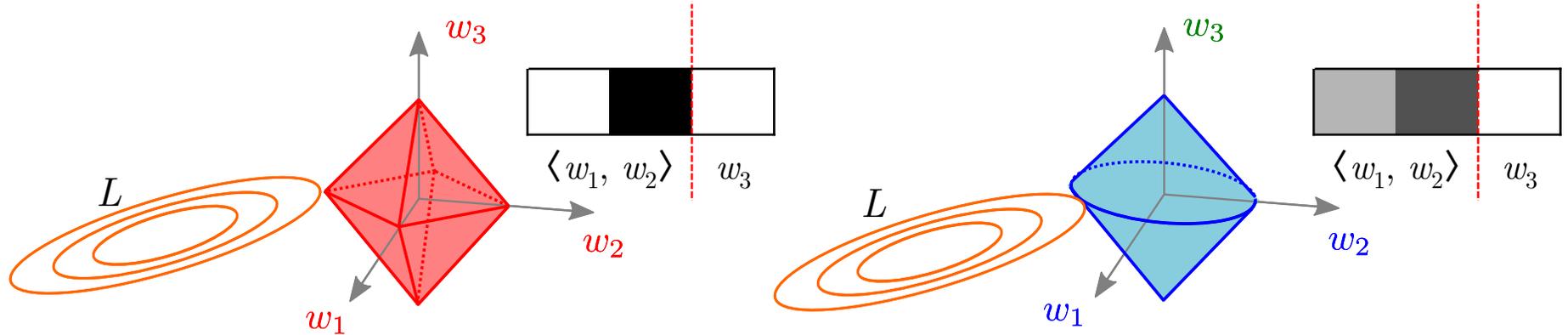


$$R(\mathbf{w}) = |w_1| + |w_2| + |w_3|$$

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w})$$

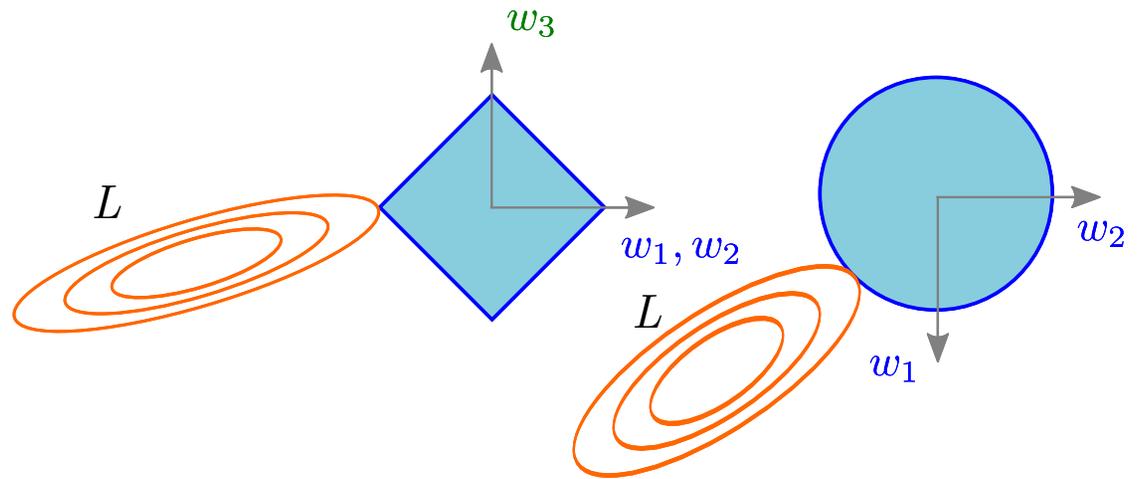
s.t. $R(\mathbf{w}) \leq \tau$

Lasso vs. Group Lasso



$$R(\mathbf{w}) = |w_1| + |w_2| + |w_3|$$

$$R(\mathbf{w}) = \|\langle w_1, w_2 \rangle\|_2 + |w_3|$$



Whence Groups?

Back to NLP ...

Sentence Regularizer

$$R(\mathbf{w}) = \sum_{n=1}^N \sum_{s=1}^{S_n} \lambda_{n,s} \|\mathbf{w}_{n,s}\|_2$$

- Every sentence s in every document n gets a group.
- If $\mathbf{w}_{n,s}$ can be driven to zero, that means the sentence is irrelevant to the task.
- Many overlapping groups!

Group for Sentence 1

1	acting
1	at
1	back
1	basics
1	big
1	bit
1	brutality
1	but
1	cheek
1	crudest
1	dagerous
	:
6	the
	:

this film is one big joke : you have all the basics elements of romance (love at first sight , great passion , etc .) and gangster flicks (brutality , dagerous machinations , the mysterious don , etc.) , but it is all done with the crudest humor .

it ' s the kind of thing you either like viserally and immediately " get " or you don ' t .

that is a matter of taste and expectations .

i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes .

the acting is very good , if a bit obviously tongue - in - cheek .

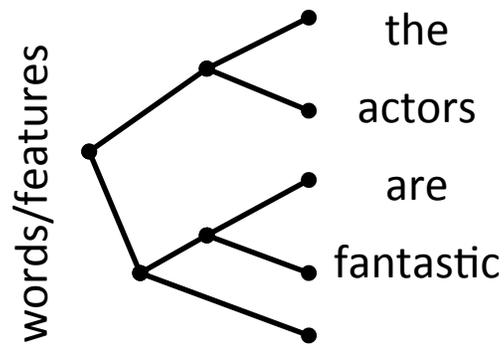
Group for Sentence 5

1	acting
1	at
1	back
1	basics
1	big
1	bit
1	brutality
1	but
1	cheek
1	crudest
1	dagerous
	:
6	the
	:

this film is one big joke : you have all the basics elements of romance (love at first sight , great passion , etc .) and gangster flicks (brutality , dagerous machinations , the mysterious don , etc.) , but it is all done with the crudest humor . it ' s the kind of thing you either like viserally and immediately " get " or you don ' t . that is a matter of taste and expectations . i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes . the acting is very good , if a bit obviously tongue - in - cheek .

More Linguistic Structure Regularizers

- Parse tree regularizer

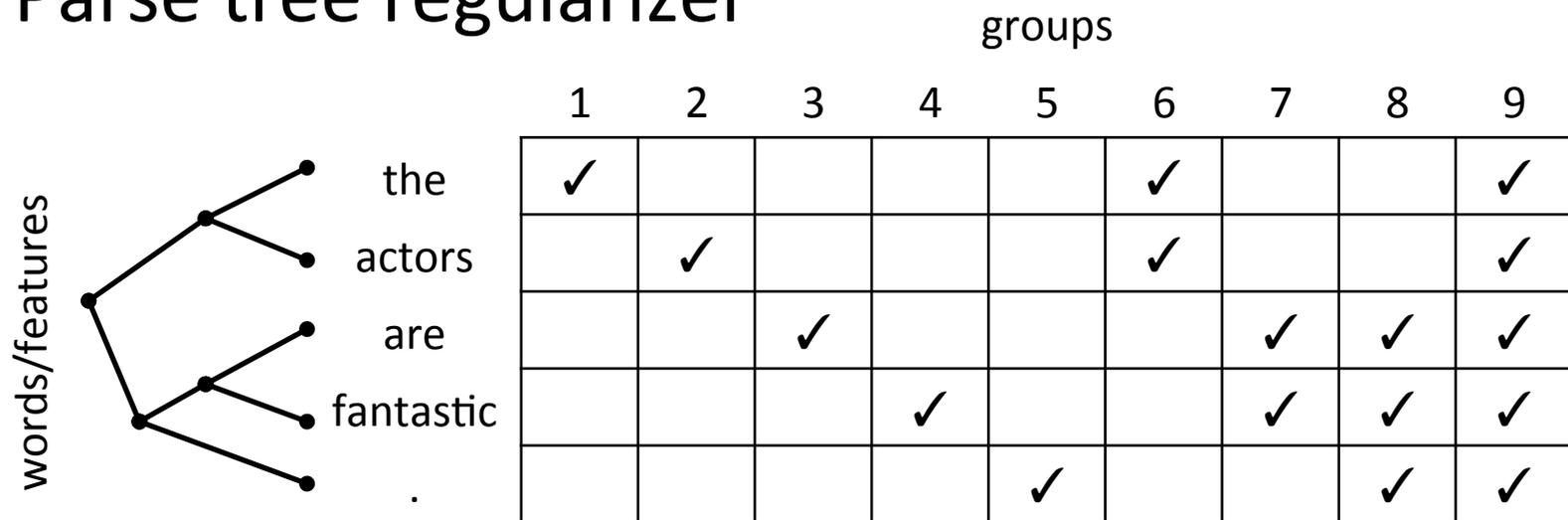


groups

	1	2	3	4	5	6	7	8	9
the	✓					✓			✓
actors		✓				✓			✓
are			✓				✓	✓	✓
fantastic				✓			✓	✓	✓
.					✓			✓	✓

More Linguistic Structure Regularizers

- Parse tree regularizer



- Each of 5,000 hierarchical Brown clusters

Sparse Group Lasso

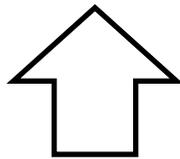
$$\min_{\mathbf{w}} R(\mathbf{w}) + \lambda \|\mathbf{w}\|_1 + \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w})$$

Optimization

$$\min_{\mathbf{w}} R(\mathbf{w}) + \lambda \|\mathbf{w}\|_1 + \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w})$$

Optimization

$$\begin{aligned} \min_{\mathbf{w}, \mathbf{v}} R(\mathbf{v}) + \lambda \|\mathbf{w}\|_1 + \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w}) \\ \text{s.t. } \mathbf{v} = \mathbf{M}\mathbf{w} \end{aligned} \left. \vphantom{\begin{aligned} \min_{\mathbf{w}, \mathbf{v}} R(\mathbf{v}) + \lambda \|\mathbf{w}\|_1 + \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w}) \\ \text{s.t. } \mathbf{v} = \mathbf{M}\mathbf{w} \end{aligned}} \right\} \begin{array}{l} \text{separate } \mathbf{w} \text{ from "copies" } \mathbf{v}, \\ \text{constraint forces agreement} \end{array}$$



$$\min_{\mathbf{w}} R(\mathbf{w}) + \lambda \|\mathbf{w}\|_1 + \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w})$$

Optimization

$$\begin{aligned} \min_{\mathbf{w}, \mathbf{v}} R(\mathbf{v}) + \lambda \|\mathbf{w}\|_1 + \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w}) \\ \text{s.t. } \mathbf{v} = \mathbf{M}\mathbf{w} \end{aligned} \quad \left. \vphantom{\min_{\mathbf{w}, \mathbf{v}}} \right\} \begin{array}{l} \text{separate } \mathbf{w} \text{ from "copies" } \mathbf{v}, \\ \text{constraint forces agreement} \end{array}$$

Optimization

$$\min_{\mathbf{w}, \mathbf{v}} R(\mathbf{v}) + \lambda \|\mathbf{w}\|_1 + \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w})$$

s.t. $\mathbf{v} = \mathbf{M}\mathbf{w}$

separate \mathbf{w} from “copies” \mathbf{v} ,
constraint forces agreement

$$\min_{\mathbf{w}, \mathbf{v}} \max_{\mathbf{u}} R(\mathbf{v}) + \lambda \|\mathbf{w}\|_1 + \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w}) + \mathbf{u} \cdot (\mathbf{v} - \mathbf{M}\mathbf{w}) + \frac{\rho}{2} \|\mathbf{v} - \mathbf{M}\mathbf{w}\|_2^2$$

“augmented Lagrangian”

Optimization

$$\min_{\mathbf{w}, \mathbf{v}} R(\mathbf{v}) + \lambda \|\mathbf{w}\|_1 + \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w})$$

s.t. $\mathbf{v} = \mathbf{M}\mathbf{w}$

} separate \mathbf{w} from “copies” \mathbf{v} ,
constraint forces agreement

$$\min_{\mathbf{w}, \mathbf{v}} \max_{\mathbf{u}} R(\mathbf{v}) + \lambda \|\mathbf{w}\|_1 + \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w}) + \mathbf{u} \cdot (\mathbf{v} - \mathbf{M}\mathbf{w}) + \frac{\rho}{2} \|\mathbf{v} - \mathbf{M}\mathbf{w}\|_2^2$$

**ADMM: Alternating
Directions**

alternating, blockwise updates of \mathbf{w} and \mathbf{v}

**Method of
Multipliers**

a “faster” version of dual ascent for solving the
augmented Lagrangian (Hestenes '69; Powell '69)

(Glowinski & Marroco '75; Gabay & Mercier '76)

“Blockwise” Updates

\mathbf{w} update \approx loss minimization with elastic net regularization (Zou & Hastie '05)

$$\min_{\mathbf{w}, \mathbf{v}} \max_{\mathbf{u}} R(\mathbf{v}) + \lambda \|\mathbf{w}\|_1 + \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w}) + \mathbf{u} \cdot (\mathbf{v} - \mathbf{M}\mathbf{w}) + \frac{\rho}{2} \|\mathbf{v} - \mathbf{M}\mathbf{w}\|_2^2$$


constant

“Blockwise” Updates

$$\min_{\mathbf{w}, \mathbf{v}} \max_{\mathbf{u}} \underbrace{R(\mathbf{v}) + \lambda \|\mathbf{w}\|_1 + \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w})}_{\text{w updates: proximal operator for each group:}} + \underbrace{\mathbf{u} \cdot (\mathbf{v} - \mathbf{M}\mathbf{w})}_{\text{dual update}} + \underbrace{\frac{\rho}{2} \|\mathbf{v} - \mathbf{M}\mathbf{w}\|_2^2}_{\text{proximal operator for each group:}}$$

\mathbf{v} updates: proximal operator for each group:

$$\mathbf{z}_{n,s} = \mathbf{M}_{d,s} \mathbf{w} - \frac{\mathbf{u}_{d,s}}{\rho}$$
$$\mathbf{v}_{n,s} = \begin{cases} \mathbf{0} & \text{if } \|\mathbf{z}_{n,s}\|_2 \leq \tau \\ \frac{\|\mathbf{z}_{n,s}\|_2 - \tau}{\|\mathbf{z}_{n,s}\|_2} \mathbf{z}_{n,s} & \text{otherwise} \end{cases}$$

“Blockwise” Updates

$$\min_{\mathbf{w}, \mathbf{v}} \max_{\mathbf{u}} R(\mathbf{v}) + \lambda \|\mathbf{w}\|_1 + \sum_{n=1}^N L(\mathbf{f}(d_n), y_n; \mathbf{w}) + \underbrace{\mathbf{u} \cdot (\mathbf{v} - \mathbf{M}\mathbf{w})}_{\text{simple dual update } \mathbf{u}} + \frac{\rho}{2} \|\mathbf{v} - \mathbf{M}\mathbf{w}\|_2^2$$

Implications

- Group sparsity and strong sparsity
- Model class is still a (fast) bag of words ...
but somehow “informed” by structure
- Learning is more expensive ... but still convex
- A new kind of **interpretability** ...

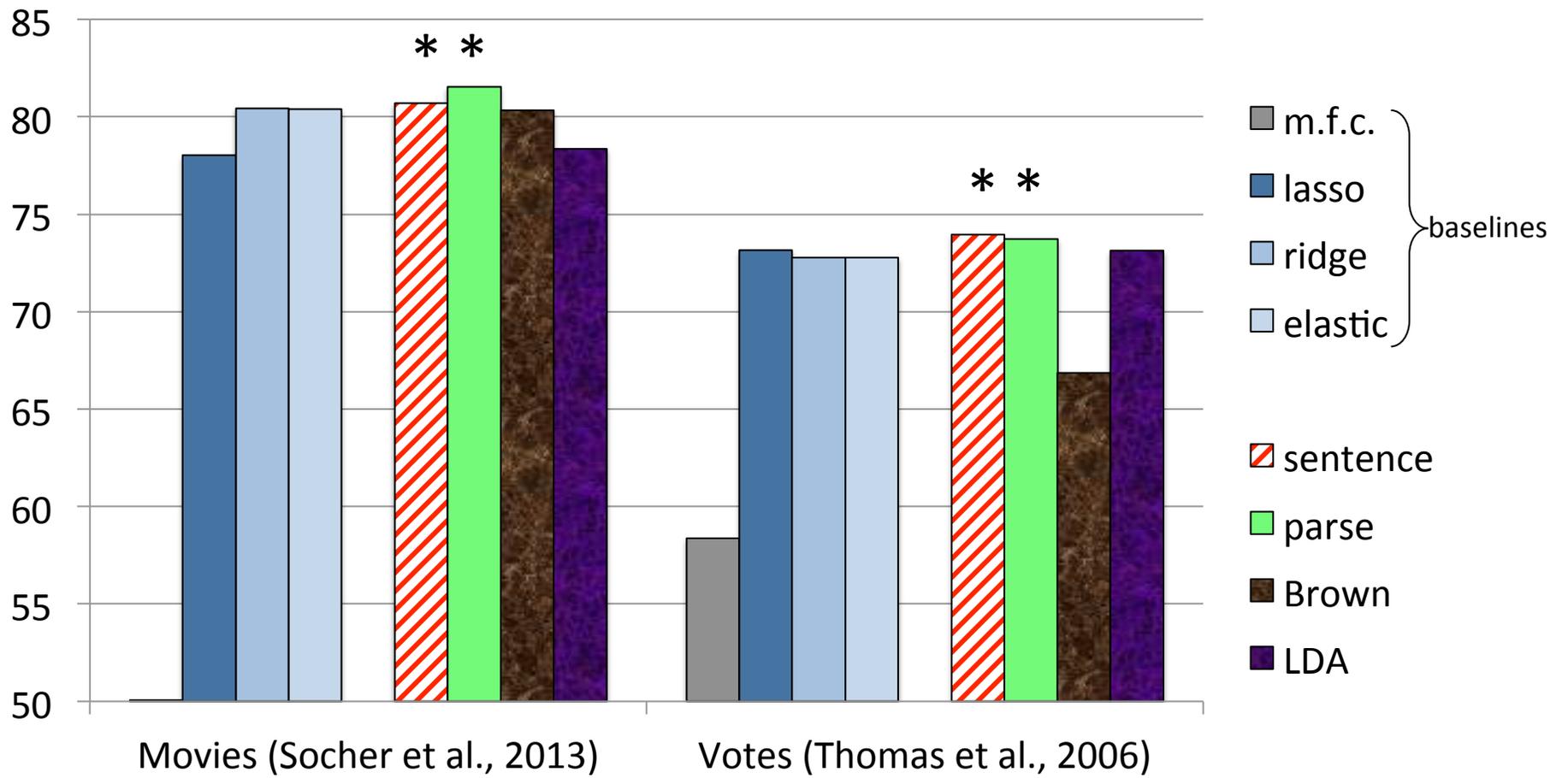
$$\frac{p(y = 1 | d)}{p(y = 1 | d \setminus s)}$$

- 1.52 this film is one big joke : you have all the basics elements of romance (love at first sight , great passion , etc .) and gangster flicks (brutality , dangerous machinations , the mysterious don , etc.) , but it is all done with the crudest humor .
- 1.01 it ' s the kind of thing you either like viserally and immediately " get " or you don ' t .
- 1.01 that is a matter of taste and expectations .
- 1.02 i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes .
- 1.00 the acting is very good , if a bit obviously tongue - in - cheek .

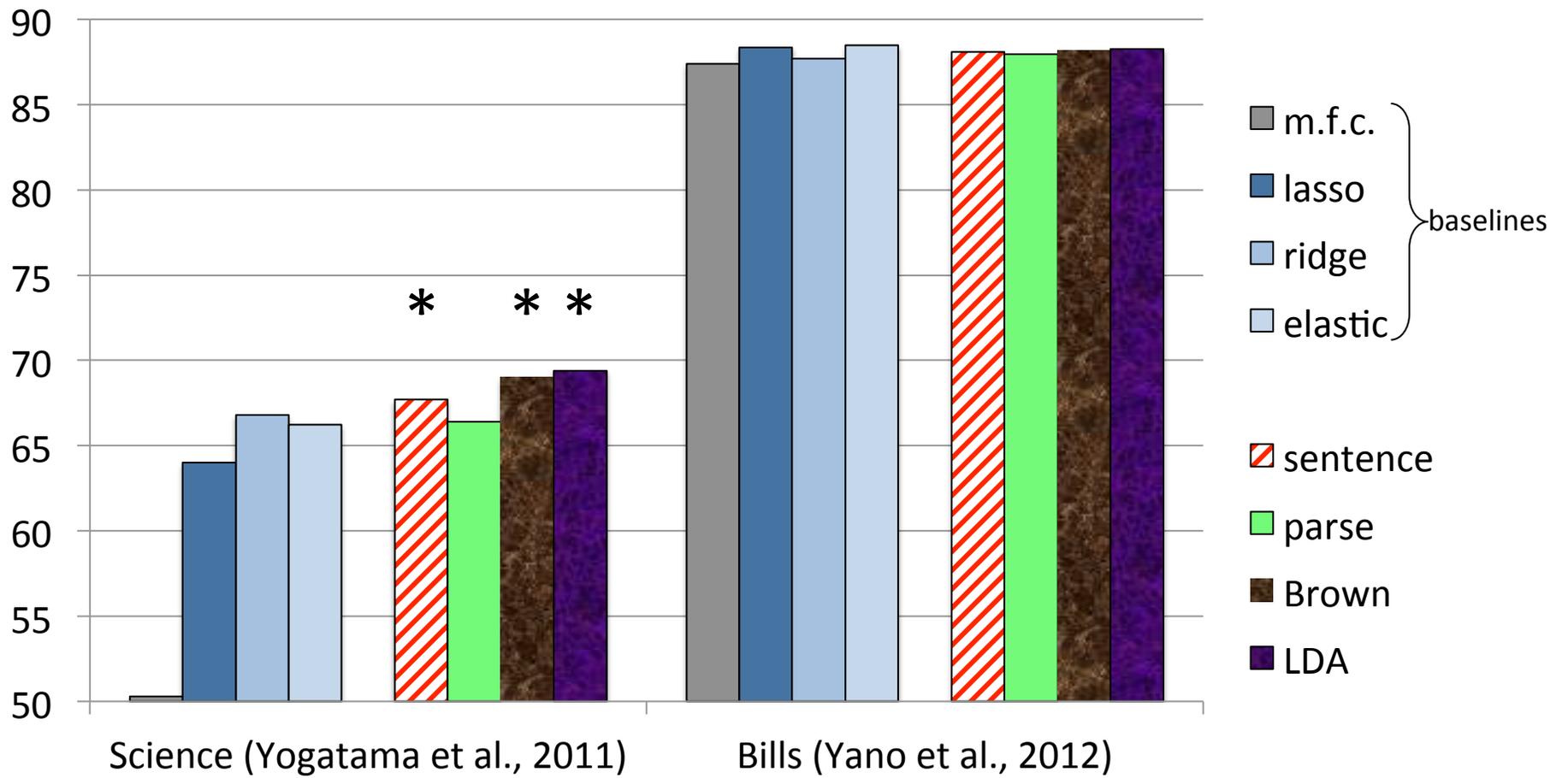
Classification Experiments

- *L*: Bag of words logistic regression
- Baselines: m.f.c., lasso, ridge, elastic
- Eight datasets

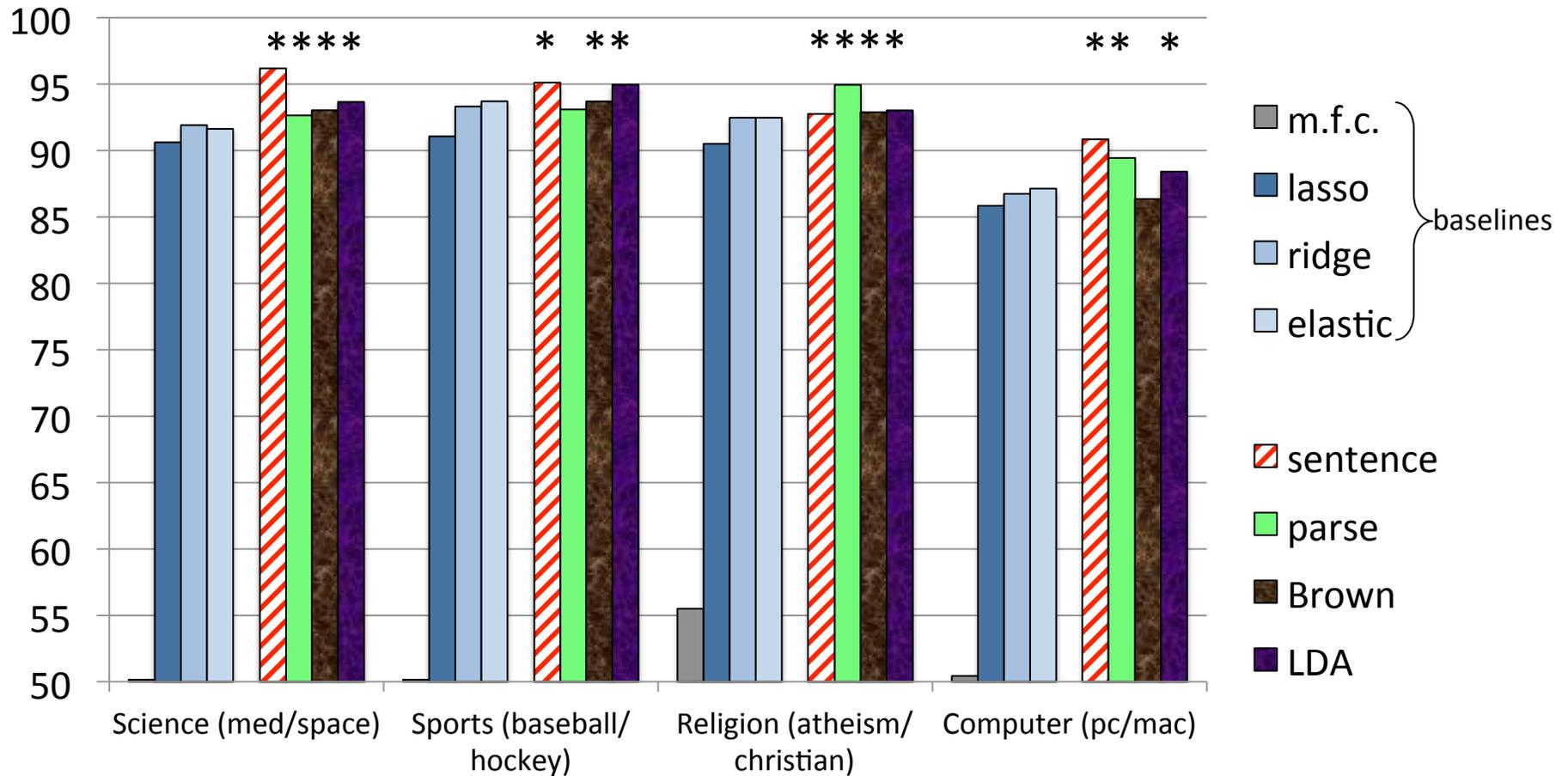
Sentiment



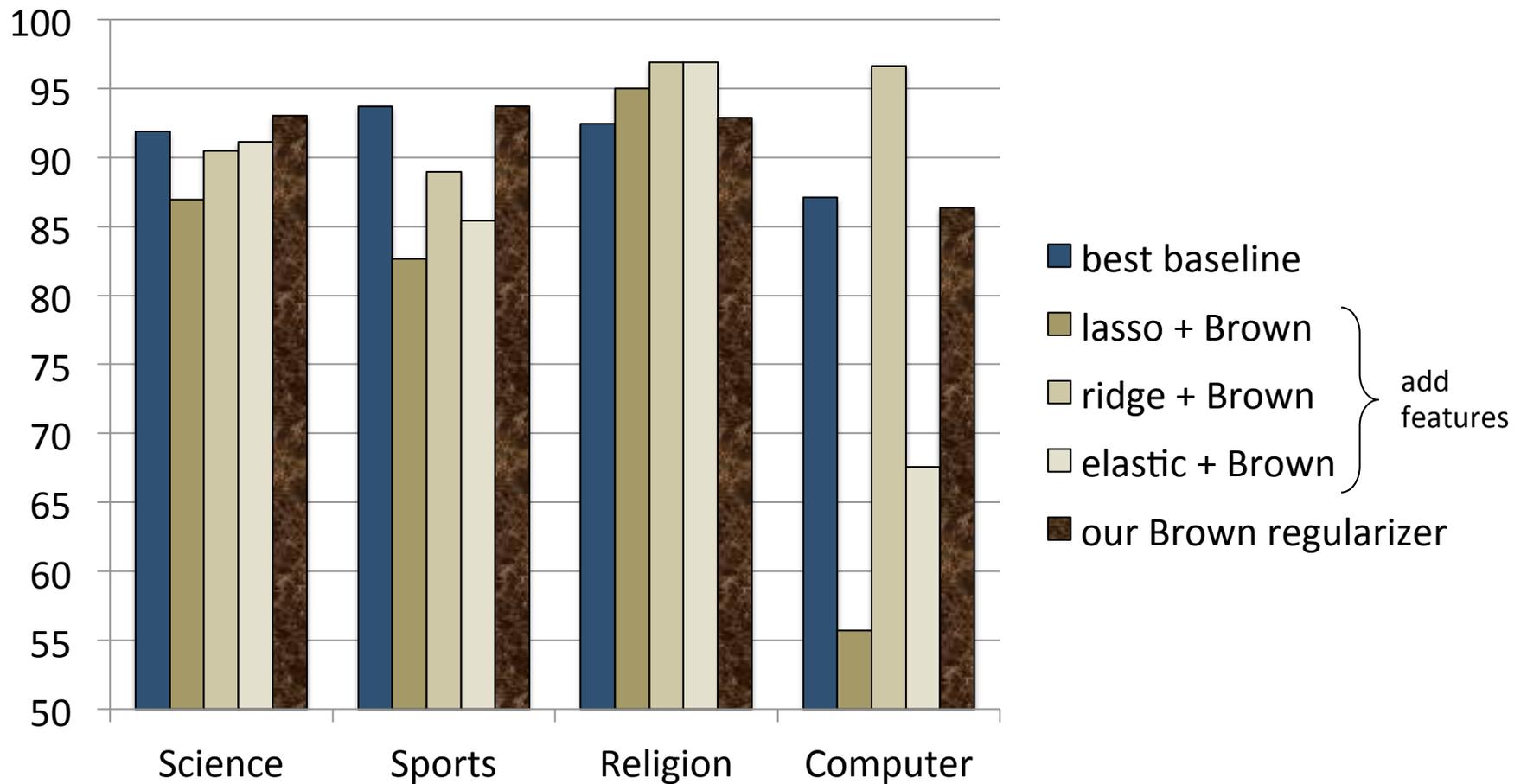
Forecasting



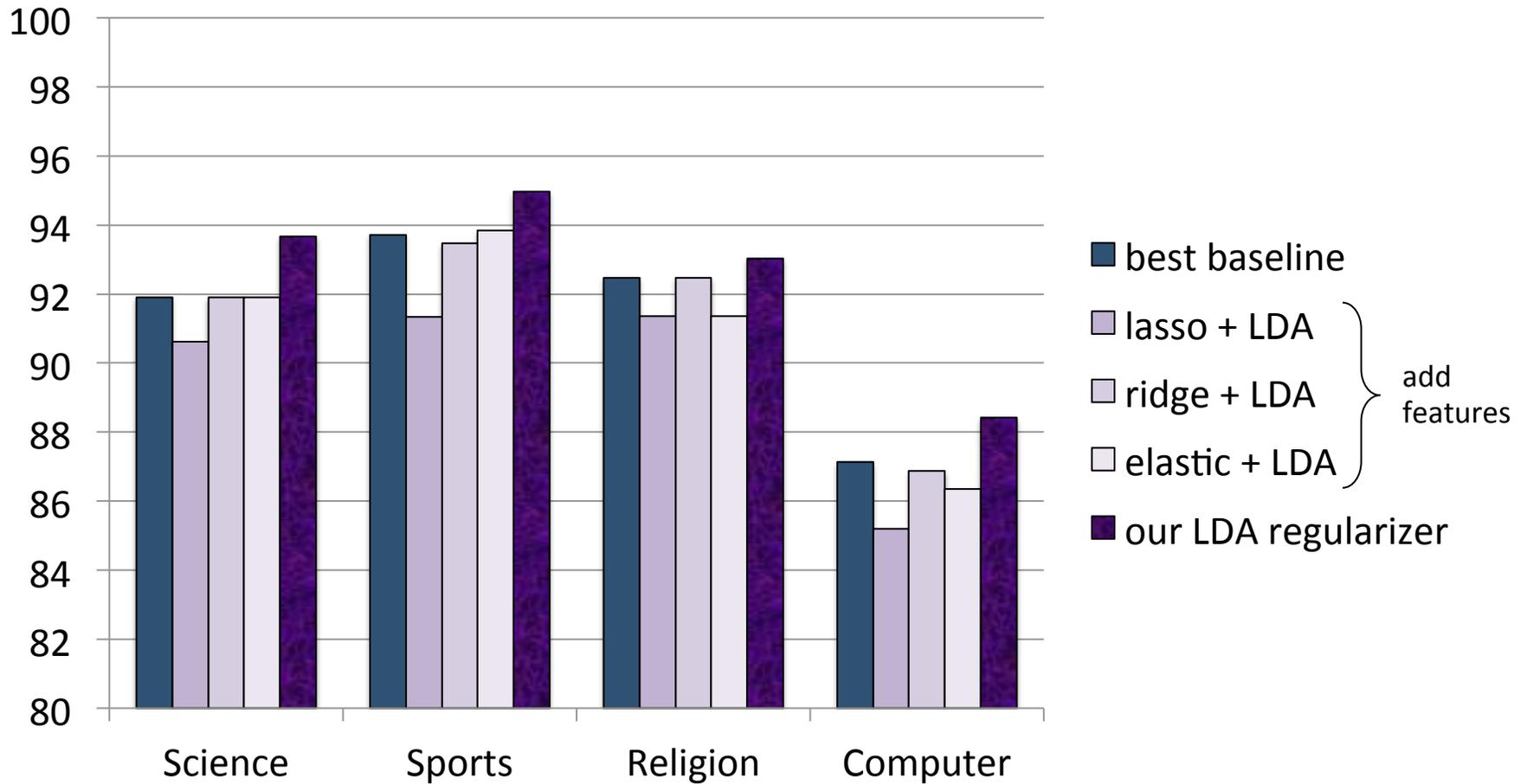
20 Newsgroups Binary Tasks



Brown as features or regularizer?

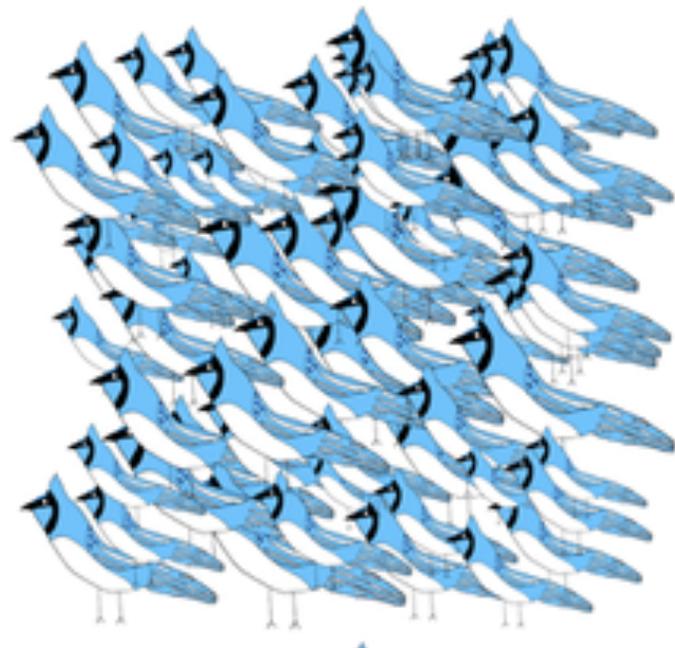


LDA as features or regularizer?



Summary

- Words of a feather (should) flock together
- Idea: use linguistic structure to define *feathers* (flocks) instead of features
- Math: sparse group lasso regularization
- Results: text classification (topics, sentiment, forecasting)



Acknowledgments: Google, IARPA, Pittsburgh Supercomputing Center

Processing Text for HealthCare

Presentations by

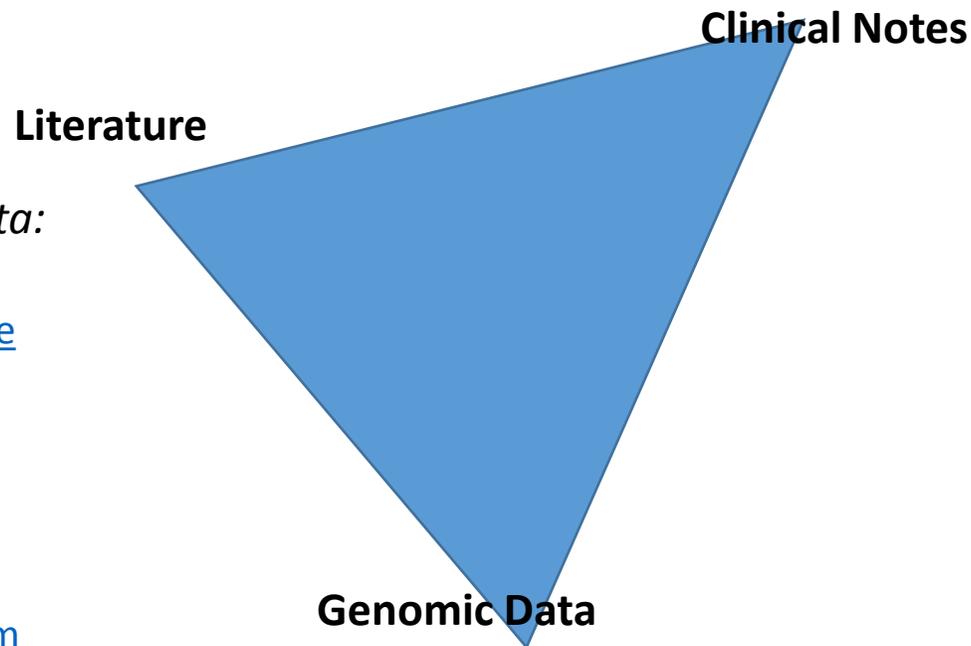
Hoifung Poon and Lucy Vanderwende,

Microsoft Research

Processing Text for HealthCare

- Introduction
- The Structure of Free Text in Clinical Records
 - Lucy Vanderwende
- Machine Reading for Cancer Panomics
 - Hoifung Poon

Introduction



Needs patient data:

[Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study](#)

[Learning Data-Driven Patient Risk Stratification Models for Clostridium difficile](#)

[On-time clinical phenotype prediction based on narrative reports](#)

Leveraging PubMed and shared task data:

[Literome: PubMed-Scale Genomic Knowledge Base in the Cloud](#)

[Distant Supervision for Cancer Pathway Extraction from Text](#)

[Joint Inference for Knowledge Extraction from Biomedical Literature](#)

Big Mechanisms DARPA Program

[Quantifying the uncertainty in heritability](#)

[An Exhaustive Epistatic SNP Association Analysis on Expanded Wellcome Trust Data](#)

[Correction for hidden confounders in the genetic analysis of gene expression](#)

The Structure of Free Text in Clinical Records

Lucy Vanderwende

Affiliate Associate Professor,
Biomedical and Health Informatics, UW Medicine
University of Washington

also:

*Senior Researcher,
Microsoft Research*

Electronic medical records

- Structured data: problem lists, lab results, pharmacy orders, discharge diagnoses, ...
- Unstructured data (Free-form text): radiology reports, operative notes, discharge summaries, ...
- More than 80% of data is in the form of free-text

An example of discharge summary

HISTORY OF PRESENT ILLNESS:

The patient is a 68 year old with acute leukemia.

The patient was in her usual state of health until about three weeks prior to admission when she began to notice increased weakness and bruising.

She presented to a Wood Emergency Department six days prior to admission.

Platelets were 9,000, hemoglobin 9.5, temperature was 100.4.

The patient had a smear there consistent with ALL.

The patient was transferred to Norri Hospital.

REVIEW OF SYSTEMS:

No headache, no nausea, vomiting or diarrhea.

Some shortness of breath with allergies, particularly cats.

No chest pain.

The patient had been doing aerobics three times a week until a couple of weeks before admission.

PAST MEDICAL HISTORY:

The patient's past medical history is significant for allergies, depression and anxiety, pleural thickening / asbestosis.

ALLERGIES:

The patient's allergies include a questionable penicillin reactions; however, the patient tolerated Ampicillin well.

The patient does not recollect what her reaction to penicillin was.

The patient also had a history of platelet reaction.

FAMILY HISTORY:

The patient's family history was significant for a brother with colon cancer.

discharge summary: keywords + sections

HISTORY OF PRESENT ILLNESS:

acute leukemia
admission
weakness bruising
Platelets hemoglobin temperature
smear ALL

REVIEW OF SYSTEMS:

headache, nausea, vomiting diarrhea
shortness of breath allergies
chest pain

admission

PAST MEDICAL HISTORY:

asbestosis allergies depression anxiety pleural thickening /

ALLERGIES:

allergies penicillin reactions Ampicillin
penicillin
platelet reaction

FAMILY HISTORY:

colon cancer

discharge summary: keywords + sections + assertions

HISTORY OF PRESENT ILLNESS:

acute leukemia.
increased weakness and bruising. until about three weeks prior to admission
Platelets hemoglobin temperature
a smear consistent with ALL. six days prior to admission.

Negation: no headache,
no nausea,
no vomiting, no diarrhea

REVIEW OF SYSTEMS:

No headache, no nausea, vomiting or diarrhea.
Some shortness of breath with allergies, particularly cats.
No chest pain.

until a couple of weeks before admission.

PAST MEDICAL HISTORY:

asbestosis. past medical history is significant for allergies, depression and anxiety, pleural thickening /

ALLERGIES:

The patient's allergies include a questionable penicillin reactions; however, the patient tolerated Ampicillin well.

a history of platelet reaction.

FAMILY HISTORY:

The patient's family history was significant for a brother with colon cancer.

discharge summary: keywords + sections + assertions

HISTORY OF PRESENT ILLNESS:

acute leukemia.
increased weakness and bruising. until about three weeks prior to admission
Platelets hemoglobin temperature
a smear consistent with ALL. six days prior to admission.

REVIEW OF SYSTEMS:

No headache, no nausea, vomiting or diarrhea.
Some shortness of breath with allergies, particularly cats.
No chest pain.

Specific information:
particularly cats

until a couple of weeks before admission.

PAST MEDICAL HISTORY:

asbestosis. past medical history is significant for allergies, depression and anxiety, pleural thickening /

ALLERGIES:

The patient's allergies include a questionable penicillin reactions; however, the patient tolerated Ampicillin well.

a history of platelet reaction.

FAMILY HISTORY:

The patient's family history was significant for a brother with colon cancer.

discharge summary: keywords + sections + assertions

HISTORY OF PRESENT ILLNESS:

acute leukemia.

increased weakness and bruising. until about three weeks prior to admission

six days prior to admission.

Platelets hemoglobin temperature
a smear consistent with ALL.

Temporal information, building a timeline

REVIEW OF SYSTEMS:

No headache, no nausea, vomiting or diarrhea.

Some shortness of breath with allergies, particularly cats.

No chest pain.

until a couple of weeks before admission.

PAST MEDICAL HISTORY:

asbestosis. past medical history is significant for allergies, depression and anxiety, pleural thickening /

ALLERGIES:

The patient's allergies include a questionable penicillin reactions; however, the patient tolerated Ampicillin well.

had a history of platelet reaction.

FAMILY HISTORY:

The patient's family history was significant for a brother with colon cancer.

discharge summary: keywords + sections + assertions

HISTORY OF PRESENT ILLNESS:

acute leukemia.

increased weakness and bruising. until about three weeks prior to admission

six days prior to admission.

Platelets hemoglobin temperature
a smear consistent with ALL.

REVIEW OF SYSTEMS:

No headache, no nausea, vomiting or diarrhea.

Some shortness of breath with allergies, particularly cats.

No chest pain.

until a couple of weeks before admission.

PAST MEDICAL HISTORY:

asbestosis. past medical history is significant for allergies, depression and anxiety, pleural thickening /

ALLERGIES:

The patient's allergies include a questionable penicillin reactions; however, the patient tolerated Ampicillin well.

a history of platelet reaction.

FAMILY HISTORY:

The patient's family history was significant for a brother with colon cancer.

Uncertainty:
questionable penicillin
reactions

discharge summary: keywords + sections + assertions

HISTORY OF PRESENT ILLNESS:

acute leukemia.

until about three weeks prior to admission

increased weakness and bruising.

six days prior to admission.

Platelets hemoglobin temperature
a smear consistent with ALL.

REVIEW OF SYSTEMS:

No headache, no nausea, vomiting or diarrhea.

Some shortness of breath with allergies, particularly cats.

No chest pain.

until a couple of weeks before admission.

PAST MEDICAL HISTORY:

asbestosis. past medical history is significant for allergies, depression and anxiety, pleural thickening /

ALLERGIES:

The patient's allergies include a questionable penicillin reactions; however, the patient tolerated Ampicillin well.

a history of platelet reaction.

FAMILY HISTORY:

The patient's family history was significant for a brother with colon cancer.

Change of State

discharge summary: keywords + sections + assertions + tests

HISTORY OF PRESENT ILLNESS:

acute leukemia.

until about three weeks prior to admission

increased weakness and bruising.

six days prior to admission.

Platelets were 9,000, hemoglobin 9.5, temperature was 100.4.

The patient had a smear there consistent with ALL.

Medical tests

REVIEW OF SYSTEMS:

No headache, no nausea, vomiting or diarrhea.

Some shortness of breath with allergies, particularly cats.

No chest pain.

until a couple of weeks before admission.

PAST MEDICAL HISTORY:

past medical history is significant for allergies, depression and anxiety, pleural thickening / asbestosis.

ALLERGIES:

The patient's allergies include a questionable penicillin reactions; however, the patient tolerated Ampicillin well.

The patient does not recollect what her reaction to penicillin was.

a history of platelet reaction.

FAMILY HISTORY:

The patient's family history was significant for a brother with colon cancer.

Medical treatments

Tool #1 - Section Chunking

History of present illness	HISTORY OF PRESENT ILLNESS: This is an 85-year-old man initially admitted to the Plastic Surgery Service for evaluation of a left facial mass. Subsequently, CMED/CCU was consulted and he was transferred to our Service postoperatively.
Past medical history	MEDICAL HISTORY: His past medical history is significant for prostate cancer, benign prostatic hypertrophy, hypothyroidism, status post radiation for non-Hodgkin's lymphoma, chronic painless hematuria, degenerative joint disease and history of a murmur. Last colonoscopy, five years ago. Dementia.
Allergies	ALLERGIES: No known drug allergies.
Medications	MEDICATIONS: 1. Levothyroxine. 2. Lasix. 3. Proscar. 4. Aeroseb. 5. Ancef.
Physical Examination	PHYSICAL EXAMINATION: On examination, he is afebrile. Vital signs, stable. Elderly man, somewhat cachectic. Head, eye, ears, nose and throat, polypoid lesion just inferior to the left zygoma, elevated superiorly, with visible bone. No exudate. Minimal bleeding. Regular rate and rhythm. Clear to auscultation. Nontender, nondistended.
Hospital Course	HOSPITAL COURSE: He was initially admitted to CMED for resection and repair of this left facial lesion. He also had consults from Urology for his hematuria as well as Medicine preoperatively and CMED/CCU. He went to the Operating Room on 2016-03-10 with Urology for hematuria where he had a cystoscopy transurethral resection of prostate placement. He then went to the Operating Room on 2016-03-14 where he had...

Tool #1 - Section Chunking

M. Tepper, D. Capurro, F. Xia, L. Vanderwende, M. Yetisgen-Yildiz. [Statistical Section Segmentation in Free-Text Clinical Records](#). Proceedings of the International Conference on Language Resources and Evaluation (LREC), Istanbul. May, 2012.

section chunking improves accuracy tagging UMLS concepts

Baseline: MetaMap identifies medical concepts in discharge summaries when checking concepts for comorbidities

Comorbidity	System	Prec/Rec/F1
Asthma	Baseline	82.8/ 84.1/ 83.5
	with Sections	89.5/ 81.0/ 85.0
Diabetes	Baseline	88.8/ 75.8/ 81.8
	with Sections	92.2/ 75.8/ 83.2

Tool #2 - Assertion analysis*

- The patient was then followed in the cardiac critical care unit where he had evidence of **anoxic encephalopathy**. (**present**)
- Heart was regular with a I/VI systolic ejection murmur without **jugular venous distention**. (**absent**)
- He does become **slightly short of breath** when lifting furniture. (**conditional**)
- If you have **fevers** please contact your PCP or return to the emergency room. (**hypothetical**)
- The patient was continued on antibiotics for possible **pneumonia**. (**possible**)
- Father had **coronary artery disease**. (**not patient**)

*enabling: 2010 Informatics for Integrating Biology and the Bedside (i2b2)/Veteran's Affairs (VA) shared-task challenge

Tool #2 - Assertion analysis

C.A. Bejan, L. Vanderwende, F. Xia, M. Yetisgen-Yildiz. [Assertion modeling and its role in clinical phenotype identification](#). J Biomed Inform, 2013. 46(1):68-74.

System configuration	<u>Absent</u>		<u>Not patient</u>		<u>Conditional</u>		<u>Hypothetical</u>		<u>Possible</u>		<u>Present</u>		<u>Overall</u>	
	<i>P</i>	<i>R</i>	<i>P</i>	<i>R</i>	<i>P</i>	<i>R</i>	<i>P</i>	<i>R</i>	<i>P</i>	<i>R</i>	<i>P</i>	<i>R</i>	macro <i>F</i>	micro <i>F</i>
<i>Training set</i>														
Basic	95.77	95.66	85.48	57.61	76.19	31.07	94.26	88.33	79.36	64.67	95.05	97.81	77.93	94.48
+Section	96.20	95.78	92.68	82.61	73.33	32.04	95.36	91.55	79.73	65.42	95.50	97.89	81.65	94.96*
+Category specific	96.50	95.78	92.94	85.87	80.39	39.81	95.51	91.55	84.87	72.34	95.97	98.16	84.55	95.55*
+Assertion focus	96.87	96.37	95.18	85.87	82.35	40.78	95.54	92.17	87.03	74.02	96.20	98.31	85.42	95.89*

Section Chunking, UMLS Concept Mapper, Assertion Tool applied to EMR

```
RandomSets - mysql - 123x42
```

filename	sectionType	CUI	conceptName	semanticTypes	assertValue
100.txt	3_History_of_Present_Illness	C0683519	history of disease	fndg	present
100.txt	3_History_of_Present_Illness	C0158266	degenerative disc	dsyn	present
100.txt	3_History_of_Present_Illness	C0011164	Degenerative	patf	present
100.txt	3_History_of_Present_Illness	C0730226	history of disease	fndg	present
100.txt	3_History_of_Present_Illness	C0024031	Low Back Pain	sosy	present
100.txt	3_History_of_Present_Illness	C0700594	Radiculopathy	dsyn	present
100.txt	3_History_of_Present_Illness	C0264180	Spondylolisthesis, grade 1	dsyn	present
100.txt	3_History_of_Present_Illness	C0009814	Stenosis	patf	present
100.txt	3_History_of_Present_Illness	C0947637	Stenosis	patf	present
100.txt	3_History_of_Present_Illness	C1261287	Stenosis	anab	present
100.txt	3_History_of_Present_Illness	C0038454	CVA, NOS	dsyn	present
100.txt	3_History_of_Present_Illness	C0011847	Diabetes	dsyn	present
100.txt	3_History_of_Present_Illness	C0011849	Diabetes	dsyn	present
100.txt	3_History_of_Present_Illness	C1457887	Symptoms	sosy	present
100.txt	3_History_of_Present_Illness	C1444648	Offered	fndg	present
100.txt	3_History_of_Present_Illness	C1576875	Write	fndg	present
100.txt	4_Discharge_Diagnoses	C0838438	Spondylolisthesis, lumbar region	dsyn	present
100.txt	6_Consultation	C1363945	Therapy	fndg	present
100.txt	6_Diagnostic_Studies_Results	C2825142	Result	fndg	present
100.txt	6_Diagnostic_Studies_Results	C0011900	Diagnostic	fndg	present
100.txt	6_Diagnostic_Studies_Results	C0683954	study results	fndg	present
100.txt	6_Diagnostic_Studies_Results	C0009566	Complication	patf	absent
100.txt	6_Diagnostic_Studies_Results	C0580859	change in position	fndg	absent
100.txt	6_Diagnostic_Studies_Results	C0237053	adnexal lesion	acab	present
100.txt	6_Diagnostic_Studies_Results	C0011649	Dermoid Cyst	neop	possible
100.txt	6_Diagnostic_Studies_Results	C0589120	Follow-up	fndg	present
100.txt	6_Diagnostic_Studies_Results	C2825142	Result	fndg	present
100.txt	6_Diagnostic_Studies_Results	C2825142	Result	fndg	present
100.txt	6_Diagnostic_Studies_Results	C2825142	Result	fndg	present
100.txt	6_Hospital_Course	C0489547	Hospital course	fndg	present
100.txt	6_Hospital_Course	C0043194	WAS	dsyn	present
100.txt	6_Hospital_Course	C1509143	Physical	fndg	present
100.txt	6_Hospital_Course	C0012634	condition	dsyn	present
100.txt	6_Hospital_Course	C0311392	SIGNS	fndg	present
100.txt	6_Hospital_Course	C0030193	Pain	sosy	present

35 rows in set (0.00 sec)

```
mysql>
```

Section Chunking, UMLS Concept Mapper, Applied to EMR

“Spondylolisthesis – Grade 1” - present
*The patient is a 56-year-old female who was found to have **grade 1 L4-L5 spondylolisthesis** and lateral recess stenosis at L4-L5 and L5-S1.*

```
RandomSets - mysql - 123x42
```

I	conceptName	semanticTypes	assertValue
683519	history of disease	fndg	present
158266	degenerative disc	dsyn	present
0011164	Degenerative	patf	present
0730226	history of disease	fndg	present
0024031	Low Back Pain	sosy	present
0000594	Radiculopathy	dsyn	present
0264180	Spondylolisthesis, grade 1	dsyn	present
0009814	Stenosis	patf	present
0947637	Stenosis	patf	present
1261287	Stenosis	anab	present
0038454	CVA, NOS	dsyn	present
0011847	Diabetes	dsyn	present
0011849	Diabetes	dsyn	present
1457887	Symptoms	sosy	present
1444648	Offered	fndg	present
1576875	Write	fndg	present
0838438	Spondylolisthesis, lumbar region	dsyn	present
1363945	Therapy	fndg	present
2825142	Result	fndg	present
0011900	Diagnostic	fndg	present
0683954	study results	fndg	present
0009566	Complication	patf	absent
0580859	change in position	fndg	absent
0237053	adnexal lesion	acab	present
0011649	Dermoid Cyst	neop	possible
0589120	Follow-up	fndg	present
2825142	Result	fndg	present
2825142	Result	fndg	present
2825142	Result	fndg	present
0489547	Hospital course	fndg	present
0043194	WAS	dsyn	present
1509143	Physical	fndg	present
0012634	condition	dsyn	present
0311392	SIGNS	fndg	present
0030193	Pain	sosy	present

35 rows in set (0.00 sec)

```
mysql>
```

Section Chunking, UMLS Concept Mapper, Assertion Tool applied to EMR

“Complication” - absent
 ... without evidence of **complication** or significant change in position or alignment.

filename	sectionType	CUI	conceptName	semanticTypes	assertValue
100.txt	3_History_of_Present_Illness	C0683519	history of disease	fndg	present
100.txt	3_History_of_Present_Illness	C0158266	degenerative disc	dsyn	present
100.txt	3_History_of_Present_Illness	C0011164	Degenerative	patf	present
100.txt	3_History_of_Present_Illness	C0730226	history of disease	fndg	present
100.txt	3_History_of_Present_Illness	C0024031	Low Back Pain	sosy	present
100.txt	3_History_of_Present_Illness	C0700594	Radiculopathy	dsyn	present
100.txt	3_History_of_Present_Illness	C0264180	Spondylolisthesis, grade 1	dsyn	present
100.txt	3_History_of_Present_Illness	C0009814	Stenosis	patf	present
100.txt	3_History_of_Present_Illness	C0947637	Stenosis	patf	present
100.txt	3_History_of_Present_Illness	C1261287	Stenosis	anab	present
100.txt	3_History_of_Present_Illness	C0038454	CVA, NOS	dsyn	present
100.txt	3_History_of_Present_Illness	C011847	Diabetes	dsyn	present
100.txt	3_History_of_Present_Illness	C011849	Diabetes	dsyn	present
100.txt	3_History_of_Present_Illness	C0457887	Symptoms	sosy	present
100.txt	3_History_of_Present_Illness	C0444648	Offered	fndg	present
100.txt	3_History_of_Present_Illness	C0576875	Write	fndg	present
100.txt	3_History_of_Present_Illness	C0838438	Spondylolisthesis, lumbar region	dsyn	present
100.txt	3_History_of_Present_Illness	C0363945	Therapy	fndg	present
100.txt	3_History_of_Present_Illness	C02825142	Result	fndg	present
100.txt	6_Diagnostic_Studies_Results	C0011900	Diagnostic	fndg	present
100.txt	6_Diagnostic_Studies_Results	C0683954	study results	fndg	present
100.txt	6_Diagnostic_Studies_Results	C0009566	Complication	patf	absent
100.txt	6_Diagnostic_Studies_Results	C0580859	change in position	fndg	absent
100.txt	6_Diagnostic_Studies_Results	C0237053	adnexal lesion	acab	present
100.txt	6_Diagnostic_Studies_Results	C0011649	Dermoid Cyst	neop	possible
100.txt	6_Diagnostic_Studies_Results	C0589120	Follow-up	fndg	present
100.txt	6_Diagnostic_Studies_Results	C2825142	Result	fndg	present
100.txt	6_Diagnostic_Studies_Results	C2825142	Result	fndg	present
100.txt	6_Diagnostic_Studies_Results	C2825142	Result	fndg	present
100.txt	6_Hospital_Course	C0489547	Hospital course	fndg	present
100.txt	6_Hospital_Course	C0043194	WAS	dsyn	present
100.txt	6_Hospital_Course	C1509143	Physical	fndg	present
100.txt	6_Hospital_Course	C0012634	condition	dsyn	present
100.txt	6_Hospital_Course	C0311392	SIGNS	fndg	present
100.txt	6_Hospital_Course	C0030193	Pain	sosy	present

35 rows in set (0.00 sec)

mysql>

Section Chunking, UMLS Concept Mapper, Assertion Tool applied to EMR

filename	sectionType	CUI	conceptName	semanticTypes	assertValue
100.txt	3_History_of_Present_Illness	C0683519	history of disease	fndg	present
100.txt	3_History_of_Present_Illness	C0158266	degenerative disc	dsyn	present
100.txt	3_History_of_Present_Illness	C0011164	Degenerative	patf	present
100.txt	3_History_of_Present_Illness	C0730226	history of disease	fndg	present
100.txt	3_History_of_Present_Illness	C0024031	Low Back Pain	sosy	present
100.txt	3_History_of_Present_Illness	C0700594	Radiculopathy	dsyn	present
100.txt	3_History_of_Present_Illness	C0264180	Spondylolisthesis, grade 1	dsyn	present
100.txt	3_History_of_Present_Illness	C0009814	Stenosis	patf	present
100.txt	3_History_of_Present_Illness	C0947637	Stenosis	patf	present
100.txt	3_History_of_Present_Illness	C1261287	Stenosis	anab	present
100.txt	3_History_of_Present_Illness	C0038454	CVA, NOS	dsyn	present
100.txt	3_History_of_Present_Illness	C0011847	Diabetes	dsyn	present
100.txt	3_History_of_Present_Illness	C0011849	Diabetes	dsyn	present
100.txt	3_History_of_Present_Illness	C1457887	Symptoms	sosy	present
100.txt	3_History_of_Present_Illness	C1444648	Offered	fndg	present
100.txt	3_History_of_Present_Illness	C1576875	Write	fndg	present
100.txt	4_Discharge_Diagnoses	C0838438	Spondylolisthesis, lumbar region	dsyn	present
100.txt	6_Consultation	C1363945	Therapy	fndg	present
100.txt	6_Diagnostic_Studies_Results	C2825142	Result	fndg	present
100.txt	6_Diagnostic_Studies_Results	C0011900	Diagnostic	fndg	present
100.txt	6_Diagnostic_Studies_Results	C0683954	study results	fndg	present
100.txt	6_Diagnostic_Studies_Results	C0009566	Complication	patf	absent
100.txt	6_Diagnostic_Studies_Results	C0580859	change in position	fndg	absent
100.txt	6_Diagnostic_Studies_Results	C0237053	adnexal lesion	acab	present
100.txt	6_Diagnostic_Studies_Results	C0011649	Dermoid Cyst	neop	possible
100.txt	6_Diagnostic_Studies_Results	C0009120	Follow-up	fndg	present
100.txt	6_Diagnostic_Studies_Results	C2825142	Result	fndg	present
100.txt	6_Diagnostic_Studies_Results	C2825142	Result	fndg	present
100.txt	6_Diagnostic_Studies_Results	C2825142	Result	fndg	present
100.txt	6_Diagnostic_Studies_Results	C489547	Hospital course	fndg	present
100.txt	6_Diagnostic_Studies_Results	C043194	WAS	dsyn	present
100.txt	6_Diagnostic_Studies_Results	C509143	Physical	fndg	present
100.txt	6_Diagnostic_Studies_Results	C012634	condition	dsyn	present
100.txt	6_Diagnostic_Studies_Results	C311392	SIGNS	fndg	present
100.txt	6_Diagnostic_Studies_Results	C0030193	Pain	sosy	present

“dermoid cyst” - possible
 This is incompletely characterized on the current study and may represent a dermoid cyst.

35 rows in set (0.00 sec)

mysql>

Extracting Structured Information from Free Text

M. Yetisgen, P. Klassen, P. Tarczy-Hornoch. [Automating Data Abstraction with Natural Language Processing in a Surgical Quality Improvement Platform](#). Proceedings of the American Medical Informatics Association Clinical Research Informatics Summit (AMIA CRI'15) (to appear), San Francisco. March, 2015.

D. Capurro, M. Yetisgen, E. van Eaton, R. Black, P. Tarczy-Hornoch. [Availability of Structured and Unstructured Clinical Data for Comparative Effectiveness Research and Quality Improvement: A Multi-Site Assessment](#). eGEMs, 2014. 2(1).

E.B. Devine, E. van Eaton, A. Devlin, N.D. Yanez, M. Yetisgen-Yildiz, D. Capurro, R. Alfonso-Cristancho, D.R. Flum, P. Tarczy-Hornoch. [Preparing electronic clinical data for quality improvement and research: The CERTAIN validation project](#). eGEMs, 2014. 1(1).

C.A. Bejan, L. Vanderwende, H.L. Evans, M.M. Wurfel, M. Yetisgen-Yildiz. [On-time clinical phenotype prediction based on narrative reports](#). Proceedings of the American Medical Informatics Association Fall Symposium (AMIA'13), Washington DC. November, 2013. (Distinguished Paper Award)

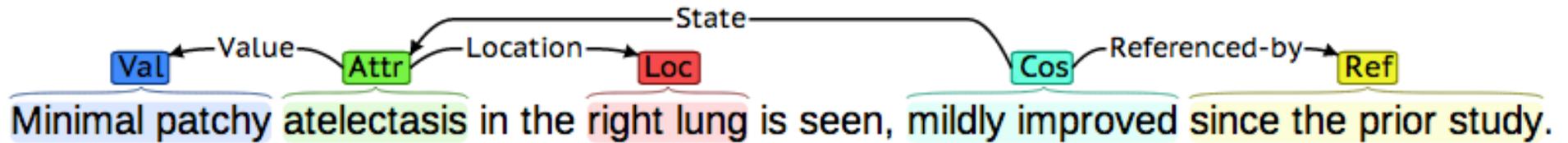
M. Yetisgen-Yildiz, M.L. Gunn, F. Xia, T.H. Payne. [A Text Processing Pipeline to Extract Recommendations from Radiology Reports](#). J Biomed Inform, 2013. 46(2):354-362.

Next tool being built:

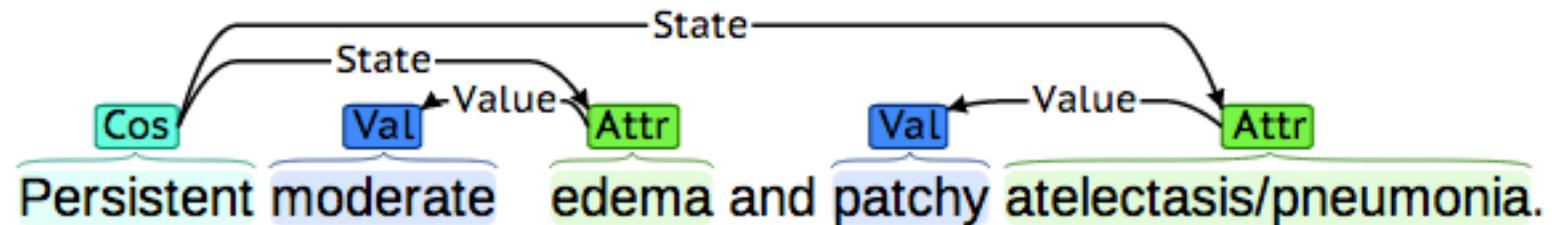
Events with change of state

- ICU Day #1: Diffuse lung opacities consistent with pulmonary edema.
- ICU Day #1: **No change** in diffuse lung opacities consistent with pulmonary edema.
- ICU Day #2: Diffuse lung opacities consistent with pulmonary edema have **worsened**.
- ICU Day #3: There has been **gradual improvement** of diffuse lung opacities consistent with pulmonary edema.

Example annotations



A snippet featuring an event annotation connecting all five fields of the COS tuple.



A snippet featuring shared entities between events

Corpus

- 1008 sentences from 1344 chest x-ray notes
 - 7173 entities
 - 4128 relations
 - 2101 event tuples
- Agreement:
 - 3 annotators annotated 100 snippets
 - Entity annotation: Kappa = 0.902
 - Event annotation: Kappa = 0.716

M. Yetisgen, P. Klassen, L. Vanderwende, F. Xia. [A New Corpus for Clinical Events with Change of State](#). Proceedings of the American Medical Informatics Association Clinical Research Informatics Summit (AMIA CRI'14), San Francisco, CA. April, 2014.

P.Klassen, F. Xia, L. Vanderwende, M. Yetisgen. [Annotating Clinical Events in Text Snippets for Phenotype Detection](#). Proceedings of International Conference on Language Resources and Evaluation (LREC). Reykjavik, Iceland, May, 2014.

Conclusion

There is rich structure in EMRs far beyond keywords and/or UMLS concepts

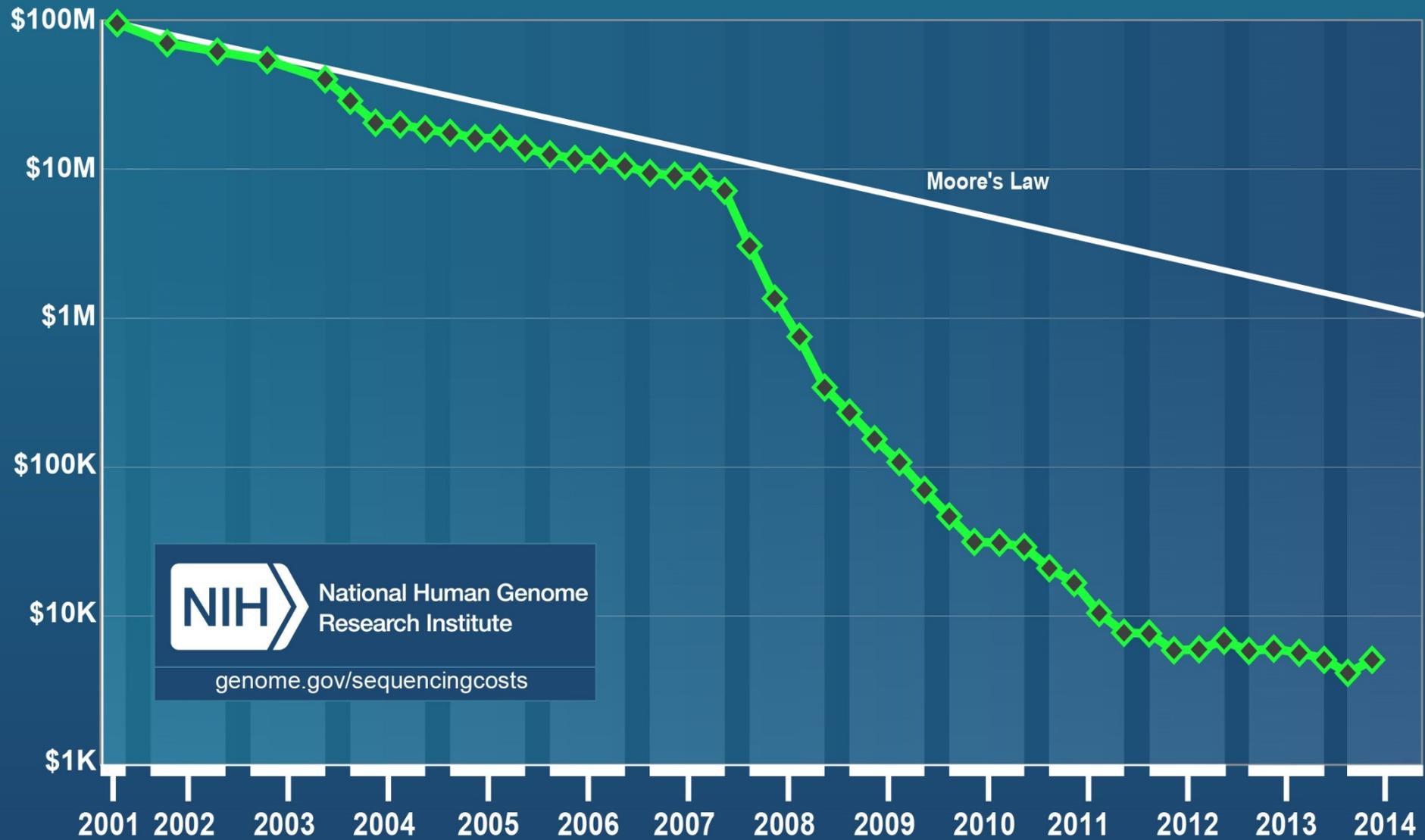
We can leverage NLP to make accessible vast amounts of clinical data available in electronic medical records (EMR)

We extract rich information with the goal of improving clinical research and patient care

Machine Reading for Cancer Panomics

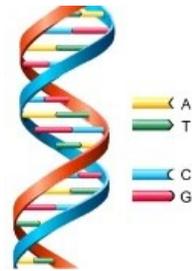
Hoifung Poon

Cost per Genome

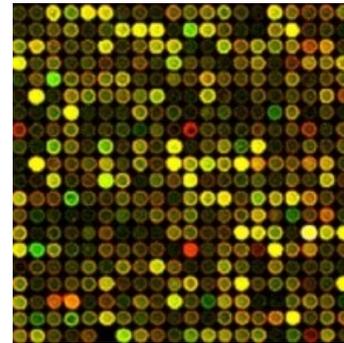


Panomics

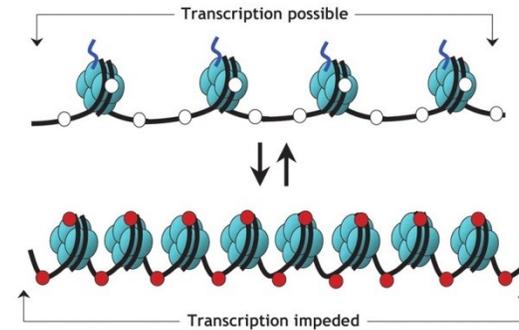
... ATTCGGATATTTAAGGC ...



Genome



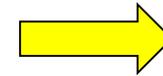
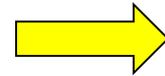
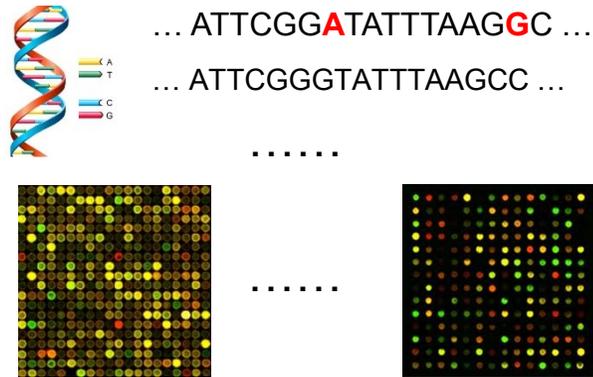
Transcriptome



Epigenome

.....

Genotype → Phenotype



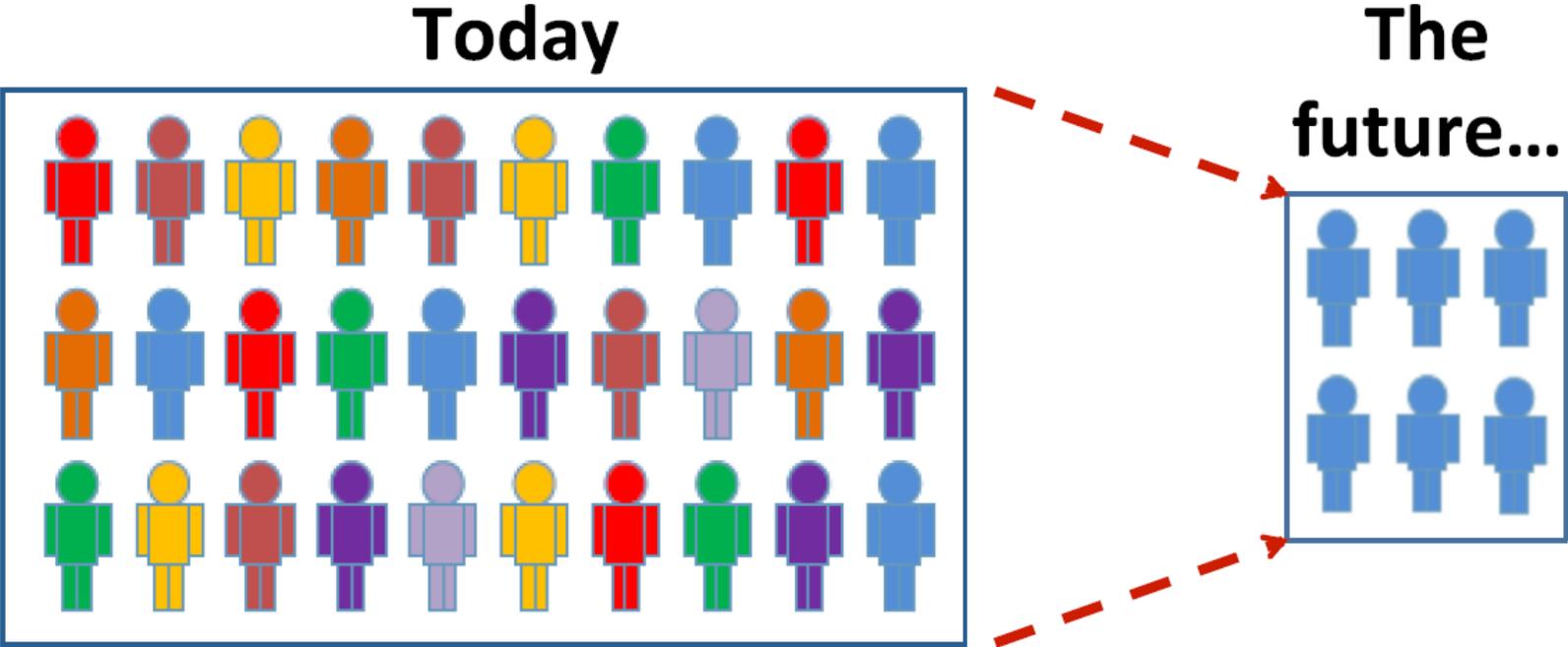
High-Throughput Data

Disease Genes

Drug Targets

.....

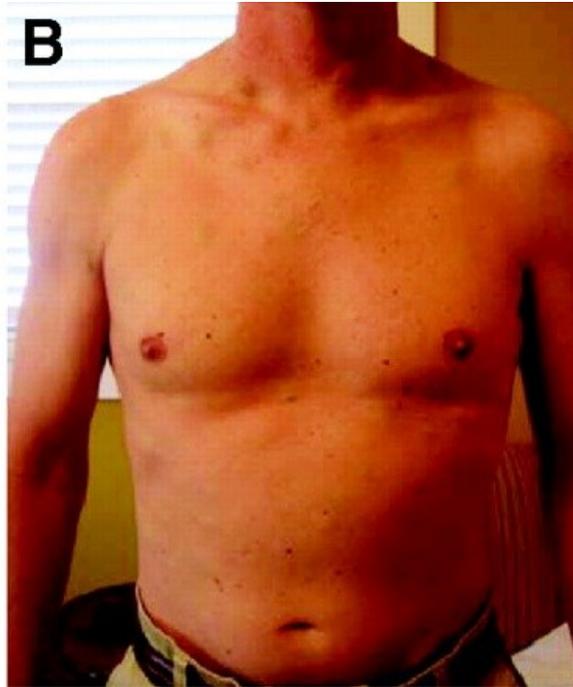
Precision Medicine



Vemurafenib on BRAF-V600 Melanoma



Before Treatment

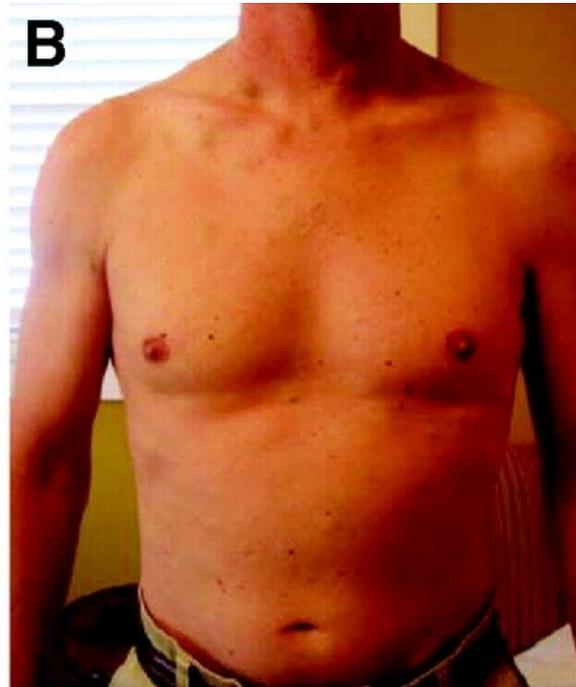


15 Weeks

Vemurafenib on BRAF-V600 Melanoma



Before Treatment



15 Weeks

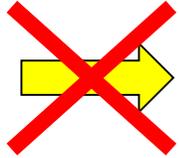
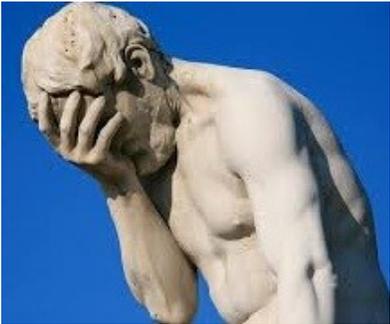
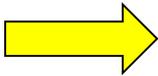


23 Weeks

Cancer Panomics



... ATTCGG**A**TATTTAAG**G**C ...
... ATTCGGGTATTTAAGCC ...
... ATTCGG**A**TATTTAAG**G**C ...
... ATTCGGGTATTTAAGCC ...
... ATTCGG**A**TATTTAAG**G**C ...
... ATTCGGGTATTTAAGCC ...



High-Throughput Experiments

Discovery

Bottleneck #1: Knowledge

Bottleneck #2: Reasoning

Example: Tumor Molecular Board



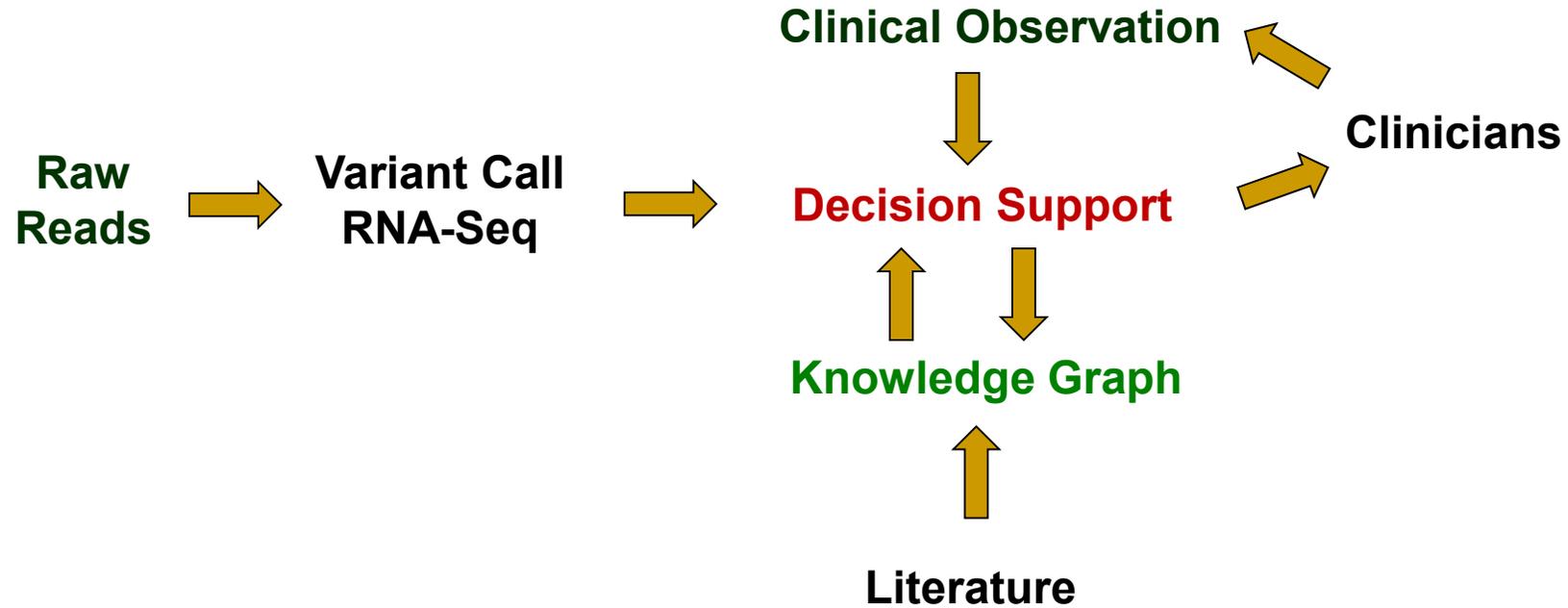
Example: Tumor Molecular Board

- 10-20 highly trained specialists
- Tens of hours on each patient
- **Problem: Hard to scale**

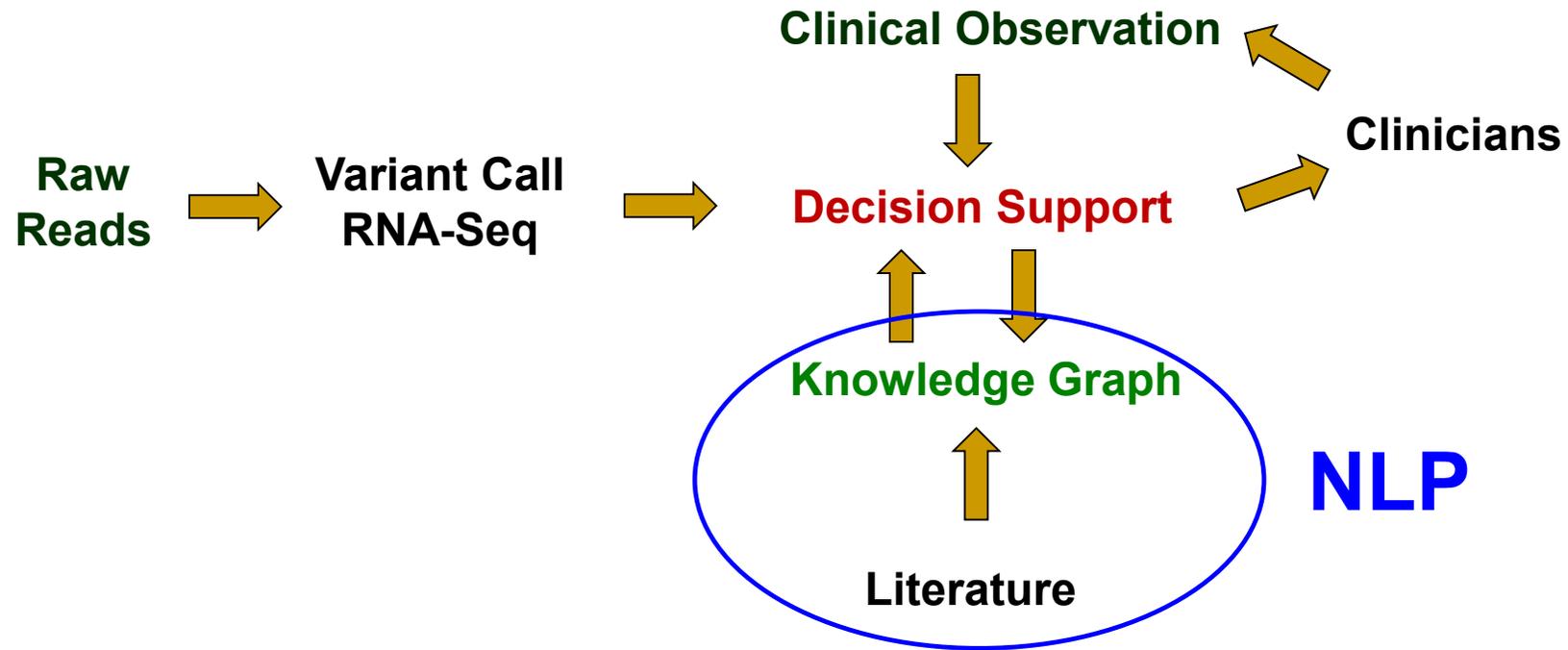
U.S. 2014: 1.6 million new cases, 585K deaths

- **Wanted: Decision support for clinical genomics**

Decision Support for Clinical Genomics

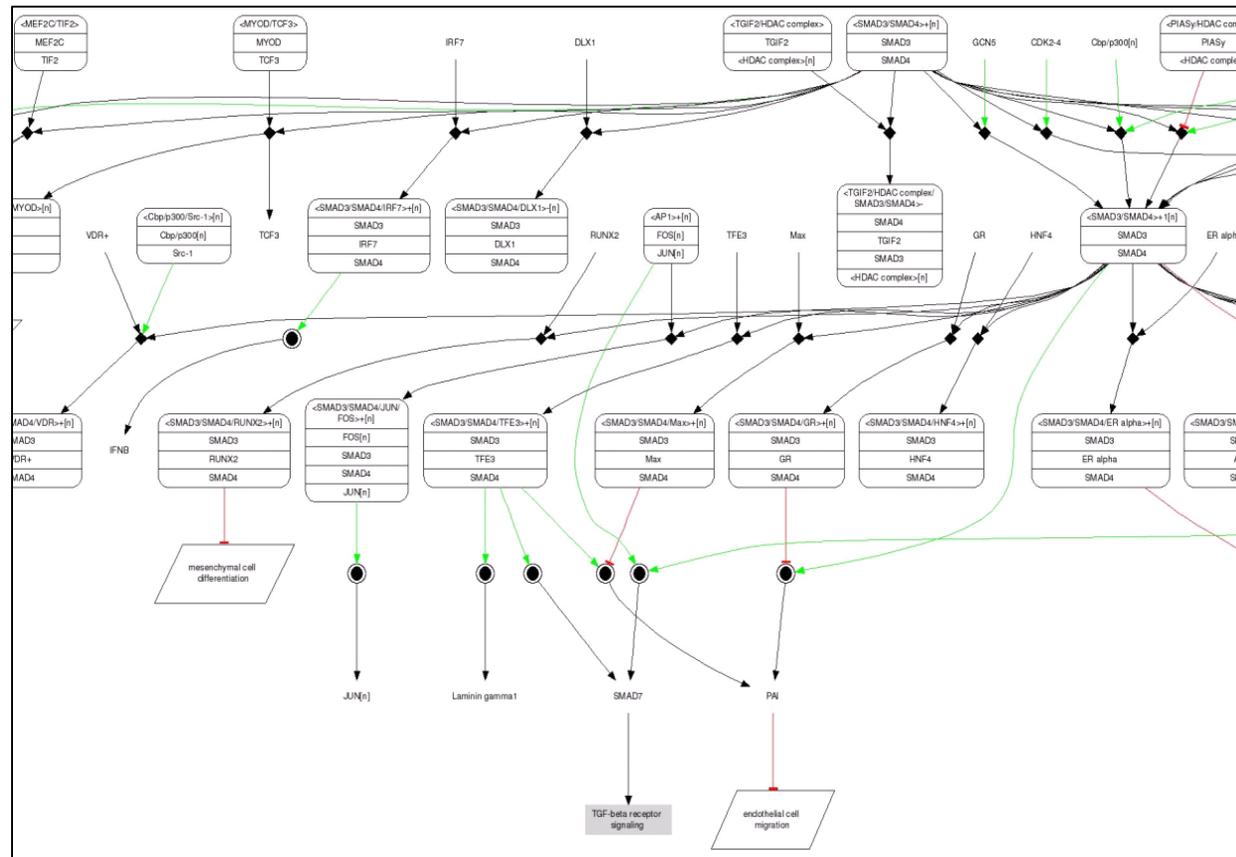


Decision Support for Clinical Genomics



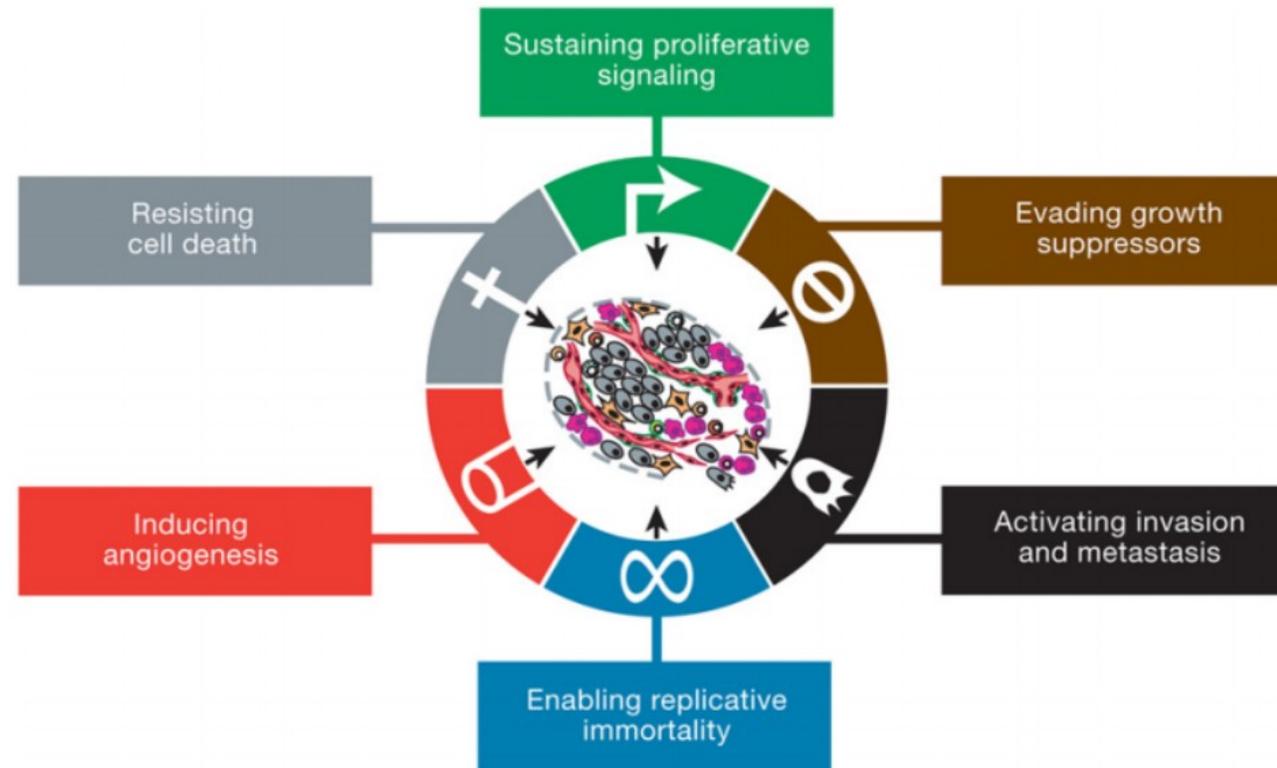
Pathway Knowledge

Genes work synergistically in pathways



Why Hard to Identify Drivers?

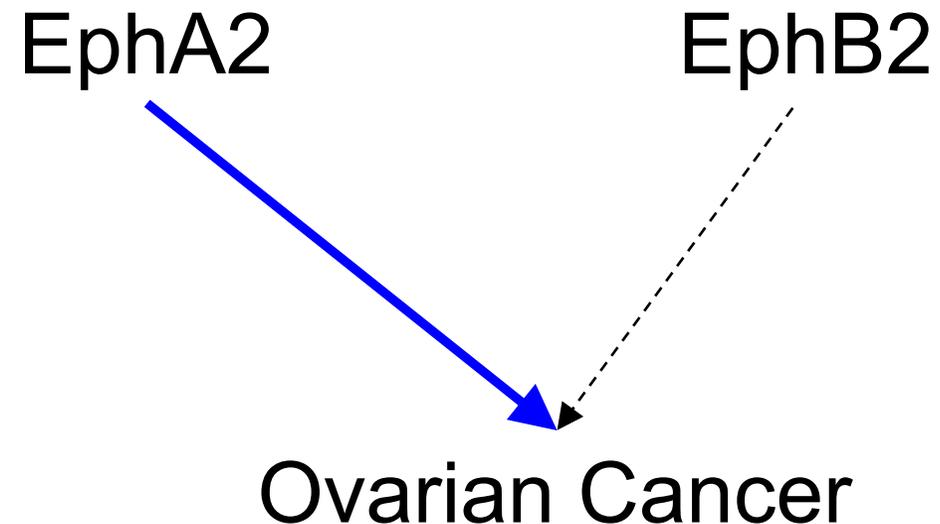
Complex diseases ← Perturb multiple pathways



Hanahan & Weinberg [Cell 2011]

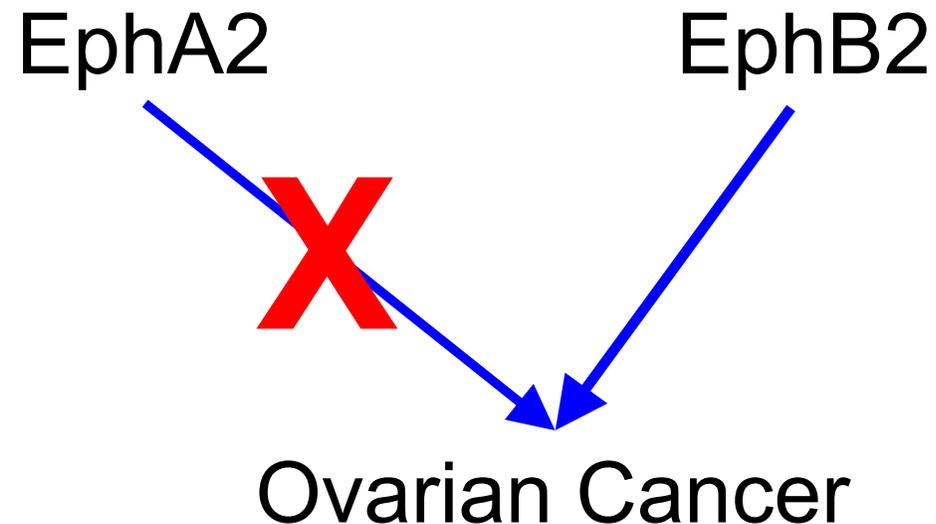
Why Cancer Comes Back?

- Subtypes with alternative pathway profile
- Compensatory pathways can be activated

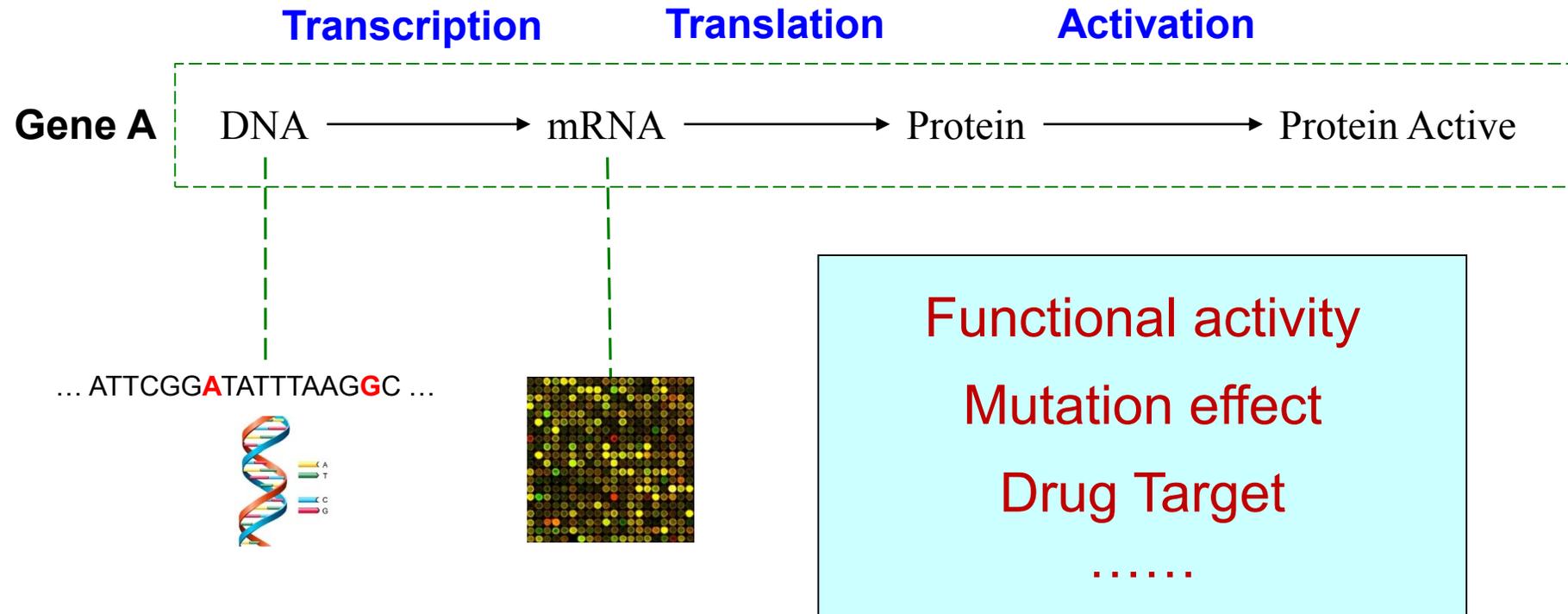


Why Cancer Comes Back?

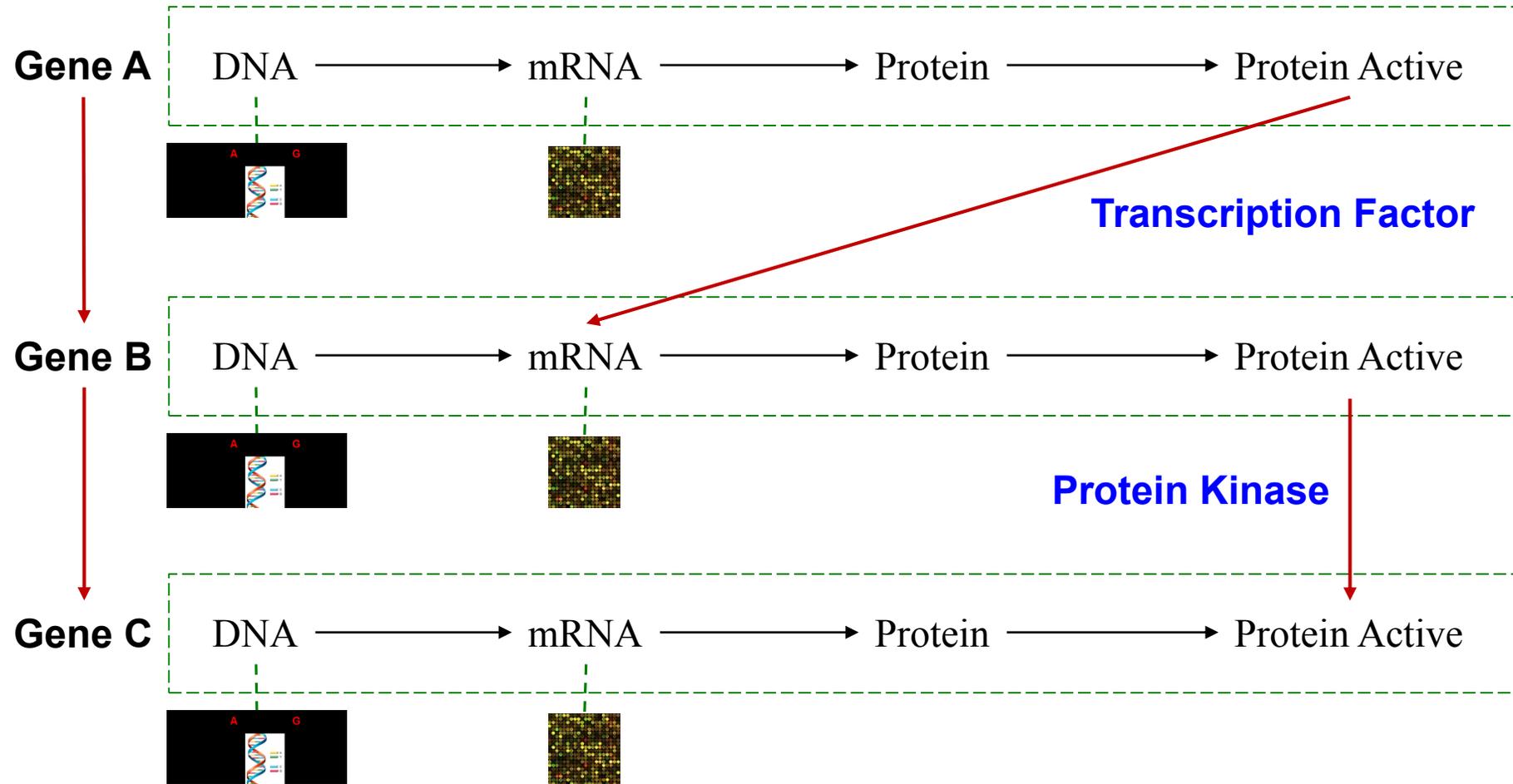
- Subtypes with alternative pathway profile
- Compensatory pathways can be activated



Cancer Systems Modeling

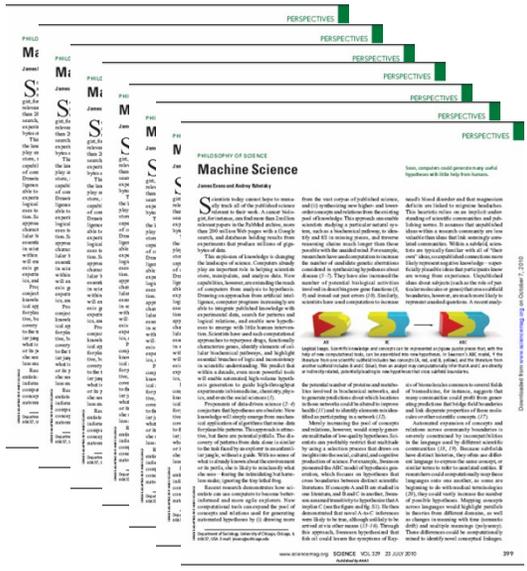


Knowledge → Model



PubMed

- 24 millions abstracts
- Two new abstracts every minute
- Adds over one million every year



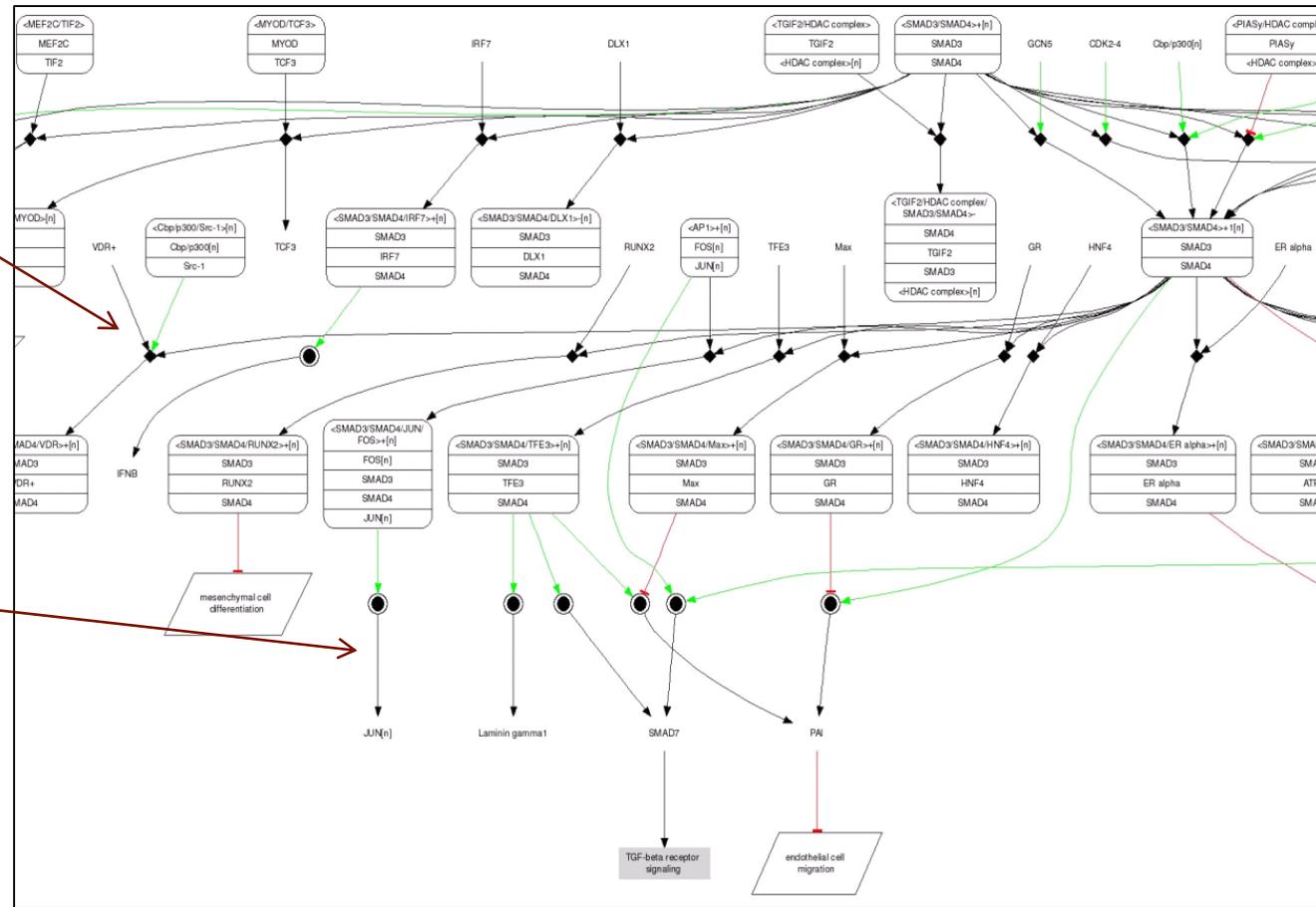
Machine Reading

PMID: 123

...
VDR+ binds to
SMAD3 to form
...

PMID: 456

...
JUN expression
is induced by
SMAD3/4
...



Machine Reading

Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...

Machine Reading

Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...

IL-10
PROTEIN

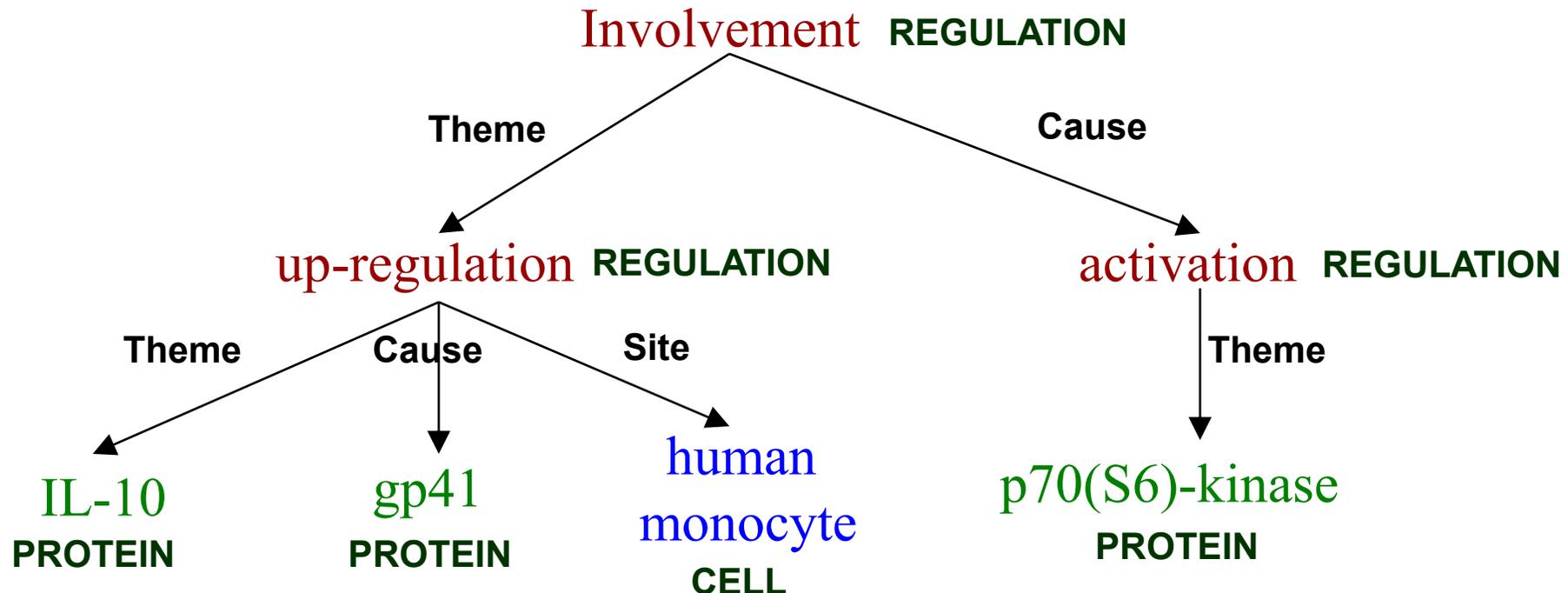
gp41
PROTEIN

human
monocyte
CELL

p70(S6)-kinase
PROTEIN

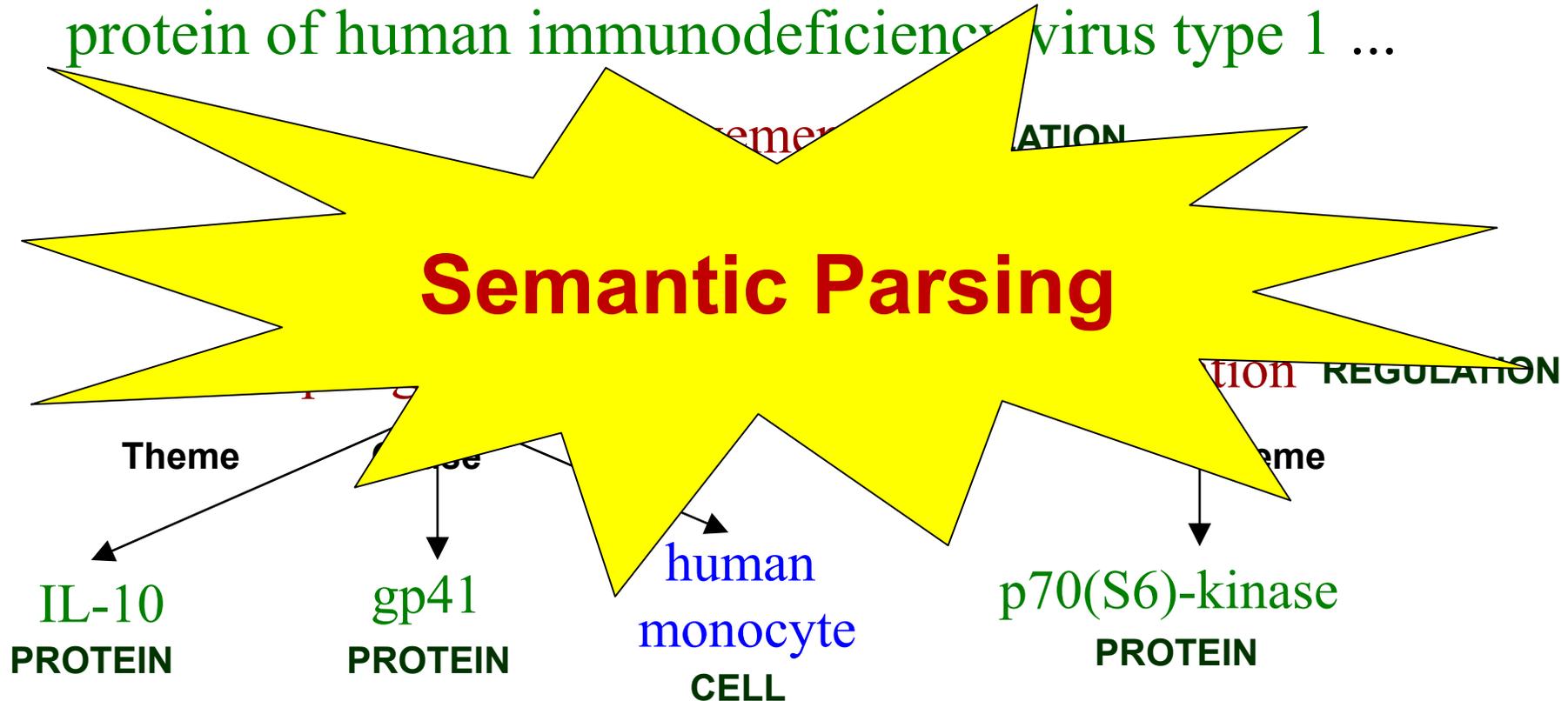
Machine Reading

Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...



Machine Reading

Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...



Long Tail of Variations

TP53 inhibits BCL2.

Tumor suppressor P53 down-regulates the activity of BCL-2 proteins.

BCL2 transcription is suppressed by P53 expression.

The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 ...

.....

Bottleneck: Annotated Examples

- GENIA (BioNLP Shared Task 2009-2013)
 - 1999 abstracts
 - MeSH: human, blood cell, transcription factor
- Challenge for “supervised” machine learning
- Can we breach this bottleneck?

Free Lunch: Existing KBs

- Many KBs available
 - Gene/Protein: GeneBank, UniProt, ...
 - Pathways: NCI, Reactome, KEGG, BioCarta, ...
- Indirect supervision

Relation Extraction

**NCI-PID
Pathway KB**

Regulation	Theme	Cause
Positive	A2M	FOXO1
Positive	ABCB1	TP53
Negative	BCL2	TP53
...

TP53 inhibits BCL2.

Tumor suppressor P53 down-regulates the activity of BCL-2 proteins.

BCL2 transcription is suppressed by P53 expression.

The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 ...

.....

Relation Extraction

NCI-PID
Pathway KB

Regulation	Theme	Cause
Positive	A2M	FOXO1
Positive	ABCB1	TP53
Negative	BCL2	TP53
...

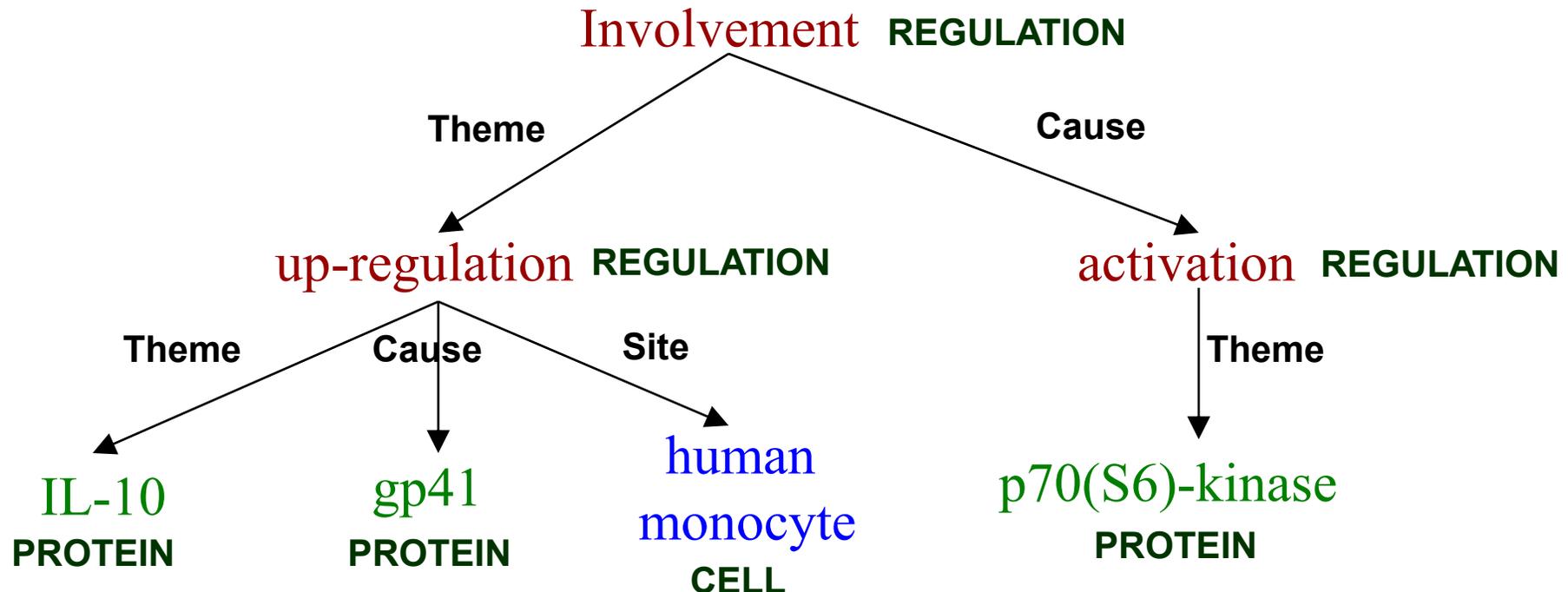
TP53 inhibits BCL2.

Tumor suppressor P53 downregulates BCL2 transcription is suppressed in B-cell CLL/Lymphoma. The inhibition of B-cell CLL/Lymphoma by TP53 is mediated by BCL2 downregulation.



Nested Events

Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...



Grounded Semantic Parsing

- **Generalize distant supervision to extracting nested events**
- **Prior:** Favor semantic parse grounded in KB
- **Outperformed 19 out of 24 participants in GENIA Shared Task [Kim et al. 2009]**

Parikh, Poon, Toutanova. “Grounded Semantic Parsing for Complex Knowledge Extraction”, *NAACL-15*.

Literome

The Literome Project

Welcome **charlie**
change to user id

Microsoft
Research

filter by ABC*

Genes: ABCA1, ABCA2, ABCA3, ABCA4, ABCA5 (1 - 50 of 5498)

genes	<input checked="" type="checkbox"/> ABCA1	ABCA1	Abacavir	PMID: 15327972 Improved antiviral activity of the aryloxymethoxyalaninyl phosphoramidate (APA) prodrug of abacavir (ABC) is due to the formation of markedly increased carbovir 5'-triphosphate metabolite levels.	... of abacavir (ABC: 1 -(1S,4R) -4-[2-amino-6-(cyclopropylamino)-9H-purin-9-yl]-2-cyclopentene-1-methanol) ... (details)
snps	<input type="checkbox"/> ABCA10				
diseases	<input type="checkbox"/> ABCA11P				
	<input type="checkbox"/> ABCA12				
drugs	<input type="checkbox"/> ABCA13				
	<input type="checkbox"/> ABCA17P				
	<input checked="" type="checkbox"/> ABCA2		Abetalipoproteinemia	PMID: 16569910	of ABCA1 with

Poon *et al.*, “Literome: PubMed-Scale Genomic Knowledge Base in the Cloud”, *Bioinformatics-14*.

<http://literome.azurewebsites.net>

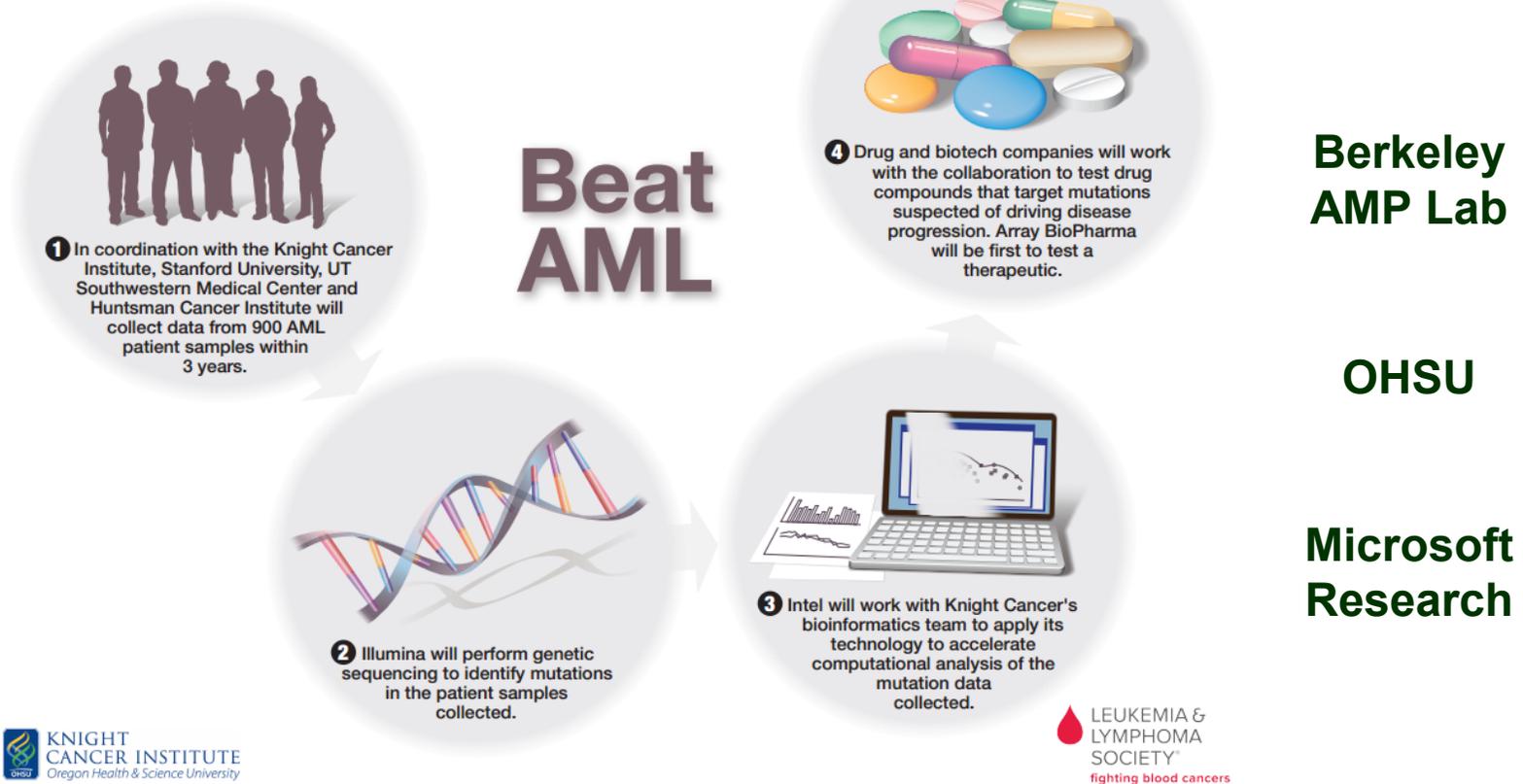
PubMed-Scale Extraction

- Preliminary pass:
 - 1.5 million instances
 - 13,000 genes, 838,000 unique regulations
- Applications:
 - UCSC Genome Browser, MSR Interactions Track
 - Expression profile modeling
 - Validate *de novo* pathway prediction
 - Etc.

Poon, Toutanova, Quirk, "Distant Supervision for Cancer Pathway Extraction from Text". *PSB-15*.

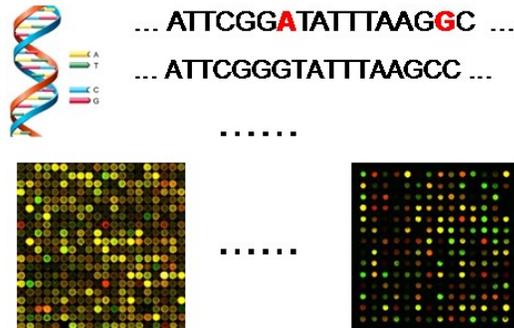
Personalized medicine approach to treating AML

The Leukemia & Lymphoma Society (LLS) and the Knight Cancer Institute at Oregon Health & Science University are leading a pioneering collaboration to develop a personalized medicine approach to improve outcomes for patients with acute myeloid leukemia (AML), a particularly devastating cancer of the blood and bone marrow. LLS provided \$8.2 million to fund Beat AML and here is how the collaboration will work:



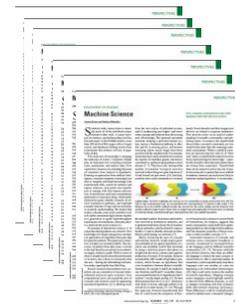
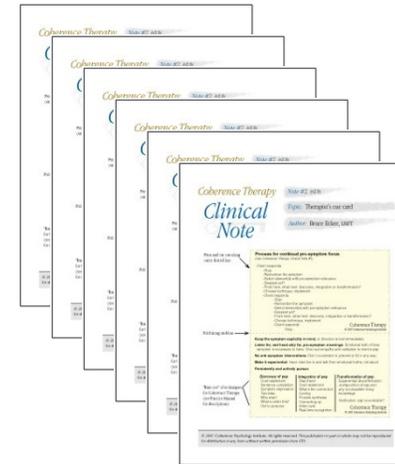
Future: VA MVP

Panomics



EHR

Precision Medicine

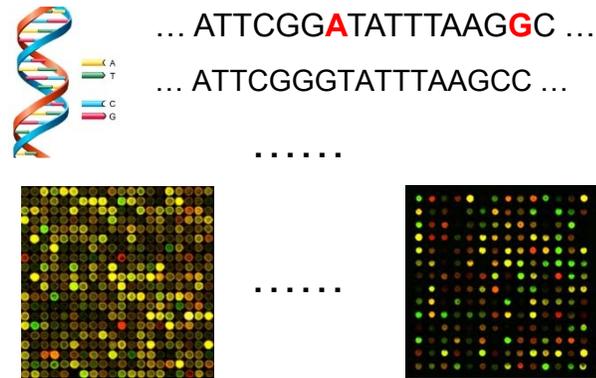


Literature

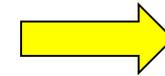
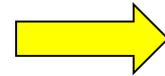
Collaborators

- **Chicago:** Andrey Rzhetsky, Kevin White
- **OHSU:** Brian Drucker, Jeff Tyner
- **Berkeley AMP Lab:** David Patterson
- **Wisconsin:** Mark Craven, Anthony Gitter
- **Microsoft Research:** Chris Quirk, Kristina Toutanova, David Heckerman, Scott Yih, Lucy Vanderwende, Bill Bolosky, Ravi Pandya

Summary



High-Throughput Data



Disease Genes

Drug Targets

.....

