Analyzing NIH Funding Patterns with Statistical Text Analysis

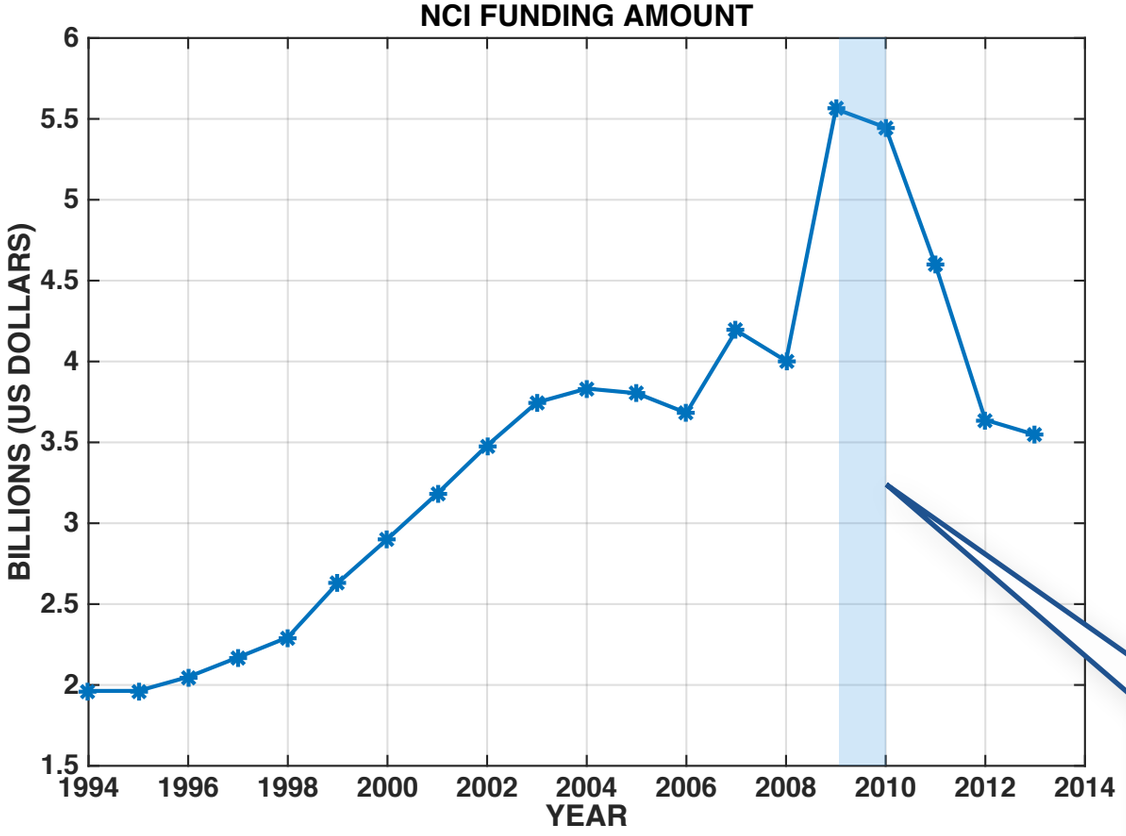
Margaret Blume-Kohout
New Mexico Consortium

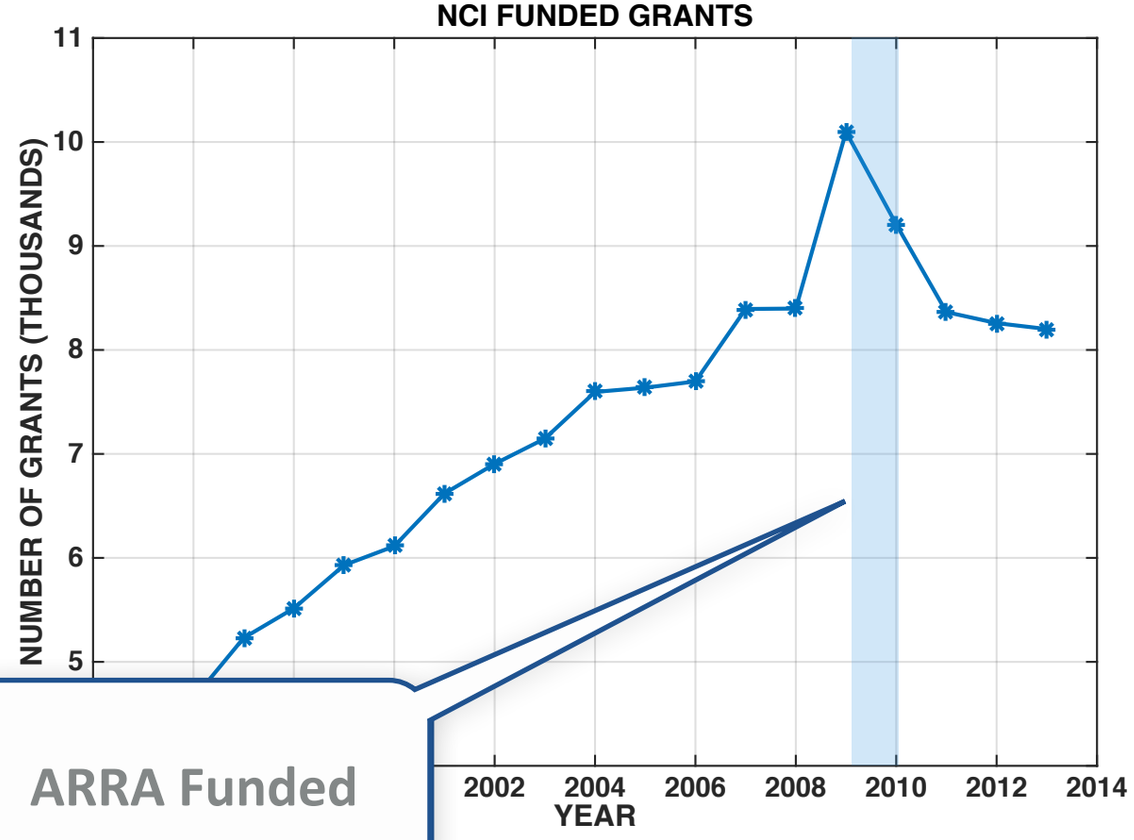
Jihyun Park
Eric Nalisnick
Padhraic Smyth
Dept. Of Computer Science
University Of California, Irvine

Ralf Krestel
Web Science Research Group
Hasso-Plattner-Institut

Measuring the Impact of NIH(National Institute of Health) Funding

- NIH invests over $30 billion each year
- Can we gain insight into this process using text and metadata?
- Our approach is to use statistical topic modeling
- We used grants data from NCI (National Cancer Institute)





ARRA Funded

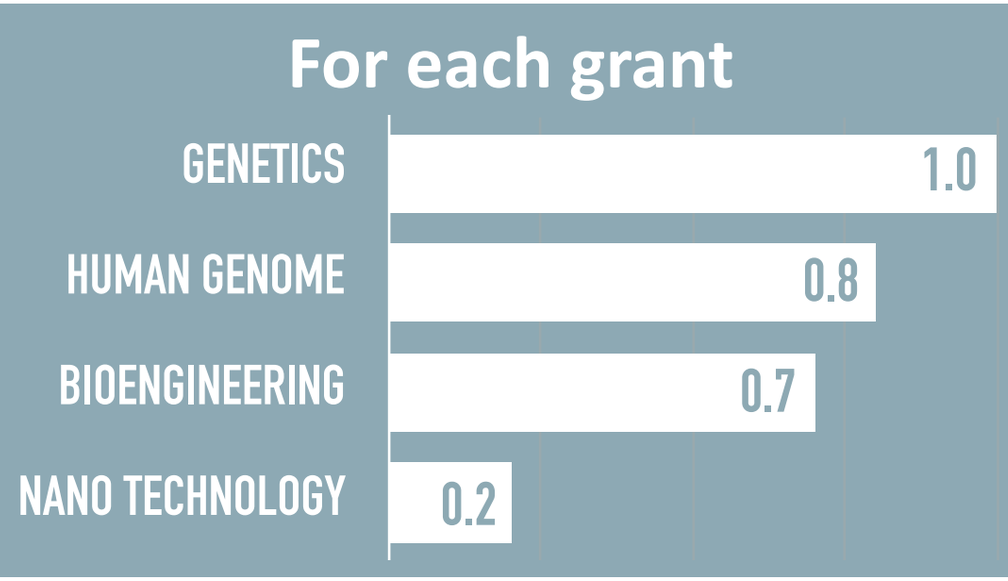
Overview

NCI Data

PROJECT ID	GRANT ABSTRACT	RCDC Labels	FUNDING	YEAR	. . .
		. . .			

Text Classification Techniques

For each grant



Probability of each label being associated with the grant

Funding + Year information

Funding patterns over time for each area

NCI (National Cancer Institute) Data

- **Grant abstracts from 1994 through 2013**

- **Text Processing**
  - BOW representation
  - Removed 500 common stopwords
  - Extracted noun-phrase terms using a NLP parser

- **BOW Data**
  - Total 149,901 documents
  - Number of documents with labels (training data) : 31,628 (2008~2011)
  - Number of documents without labels : 118,273
  - Size of vocabulary (W) : 29,713

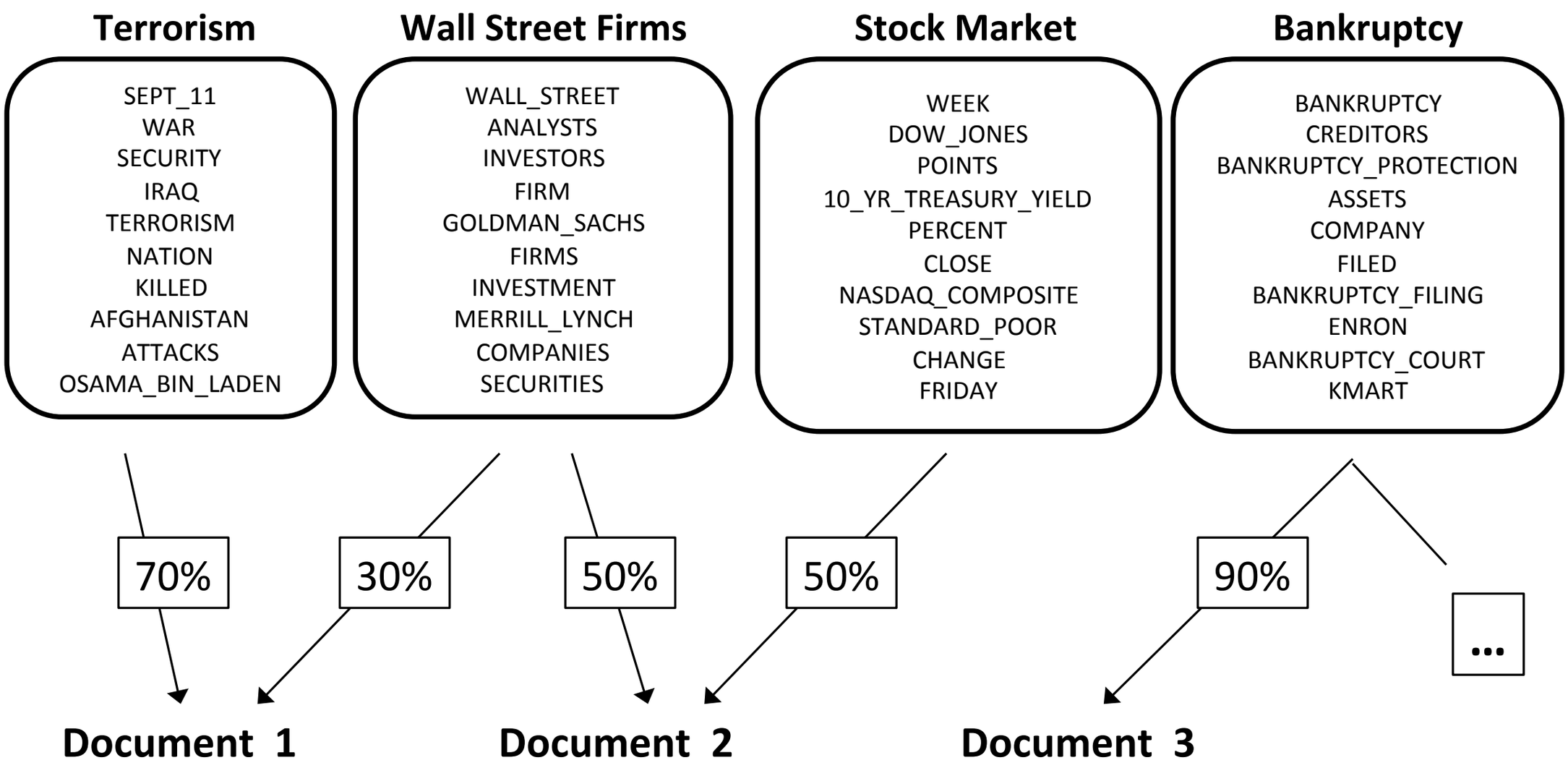
LDA: Topics are Represented as Distributions over Words

Figures from Mark Steyvers

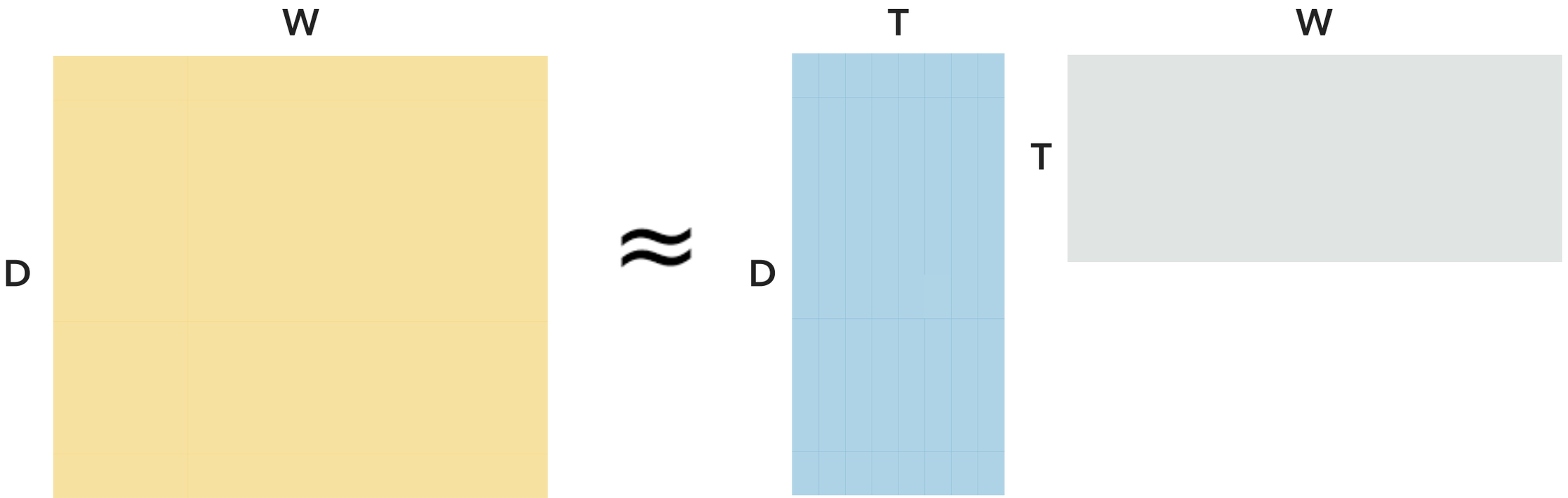
Terrorism	Wall Street Firms	Stock Market	Bankruptcy
SEPT_11	WALL_STREET	WEEK	BANKRUPTCY
WAR	ANALYSTS	DOW_JONES	CREDITORS
SECURITY	INVESTORS	POINTS	BANKRUPTCY_PROTECTION
IRAQ	FIRM	10_YR_TREASURY_YIELD	ASSETS
TERRORISM	GOLDMAN_SACHS	PERCENT	COMPANY
NATION	FIRMS	CLOSE	FILED
KILLED	INVESTMENT	NASDAQ_COMPOSITE	BANKRUPTCY_FILING
AFGHANISTAN	MERRILL_LYNCH	STANDARD_POOR	ENRON
ATTACKS	COMPANIES	CHANGE	BANKRUPTCY_COURT
OSAMA_BIN_LADEN	SECURITIES	FRIDAY	KMART

LDA: Documents are Represented as Combinations of Topics

Figures from Mark Steyvers

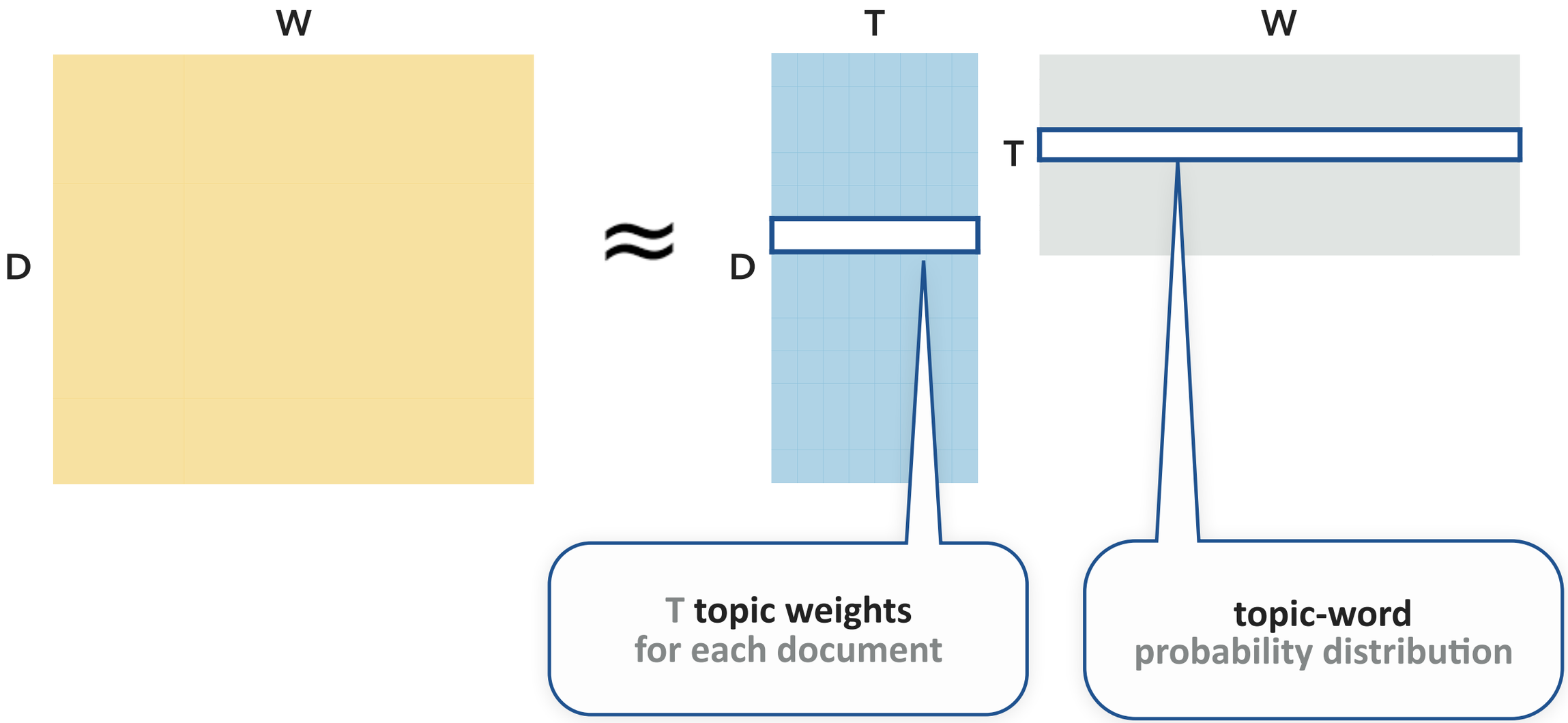


LDA (Latent Dirichlet Allocation)



LDA (Latent Dirichlet Allocation)

- Topic Models as Factor Analysis for Count Data



NIH Data Representation for L-LDA

Words or Terms

Documents	brain	lung_cancer	women	obesity	children	mice	experiment	hbv	quality	glioma	researcher
doc1	3		1				1		1		1
doc2		1		1							
doc3					2						1
doc4			1				1		1		
doc5				1				1			
doc6									2		
doc7		1				1					
doc8	2									1	
doc9					1	1	1			2	
doc10			2								

NIH Data Representation for L-LDA

Words or Terms

Documents	brain	lung_cancer	women	obesity	children	mice	experiment	hbv	quality	glioma	researcher
doc1	3		1				1		1		1
doc2		1		1							
doc3					2						1
doc4			1				1		1		
doc5				1				1			
doc6									2		
doc7		1				1					
doc8	2									1	
doc9					1	1	1			2	
doc10			2								

Codes or Labels

Documents	brain cancer	breast cancer	kidney disease	lung cancer	mind and body
doc1	1				
doc2				1	
doc3					1
doc4		1		1	
doc5			1		
doc6				1	
doc7			1		
doc8	1				
doc9	1				1
doc10		1			

NIH Data Representation for L-LDA

Words or Terms

Documents	brain	lung_cancer	women	obesity	children	mice	experiment	hbv	quality	glioma	researcher
doc1	3		1				1		1		1
doc2		1		1							
doc3					2						1
doc4			1				1		1		
doc5				1				1			
doc6									2		
doc7		1				1					
doc8	2									1	
doc9					1	1	1			2	
doc10			2								

Codes or Labels

Documents	brain cancer	breast cancer	kidney disease	lung cancer	mind and body	Background 1	Background 2
doc1	1					1	1
doc2				1		1	1
doc3					1	1	1
doc4		1			1	1	1
doc5			1			1	1
doc6					1	1	1
doc7				1		1	1
doc8	1					1	1
doc9	1				1	1	1
doc10		1				1	1

Examples of Topics from NCI Abstracts (5 out of 98)

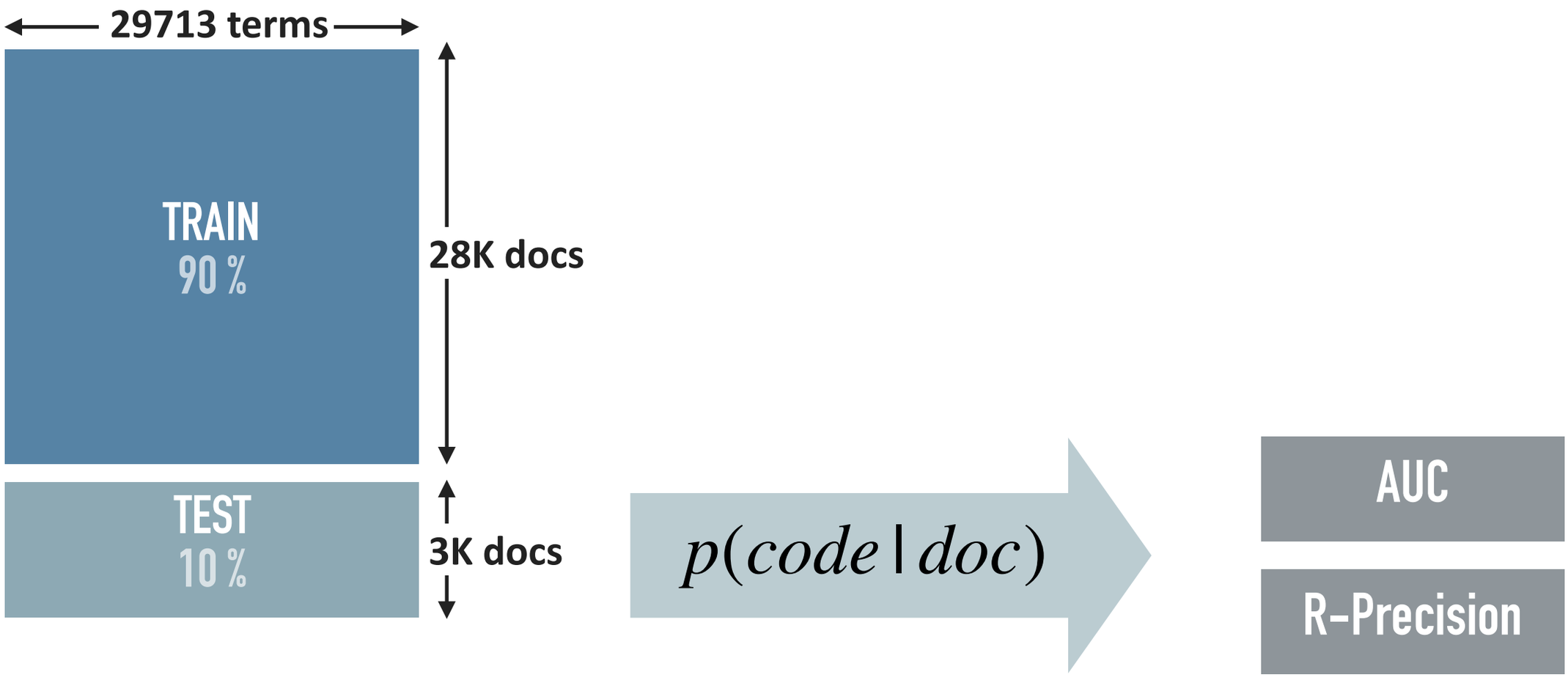
Brain Cancer	Breast Cancer	Kidney Disease
glioma	breast cancer	rcc
brain tumor	women	kidney cancer
gbm	breast cancer cell	renal cell carcinoma
malignant glioma	breast	vhl
glioblastoma	breast cancer patient	renal cancer
brain	brca1	pvhl

Background 1	Background 7
program	model
trainee	mice
university	work
training	experiment
candidate	human
field	mouse model

88 Topics from RCDC labels
10 Background topics

Evaluation

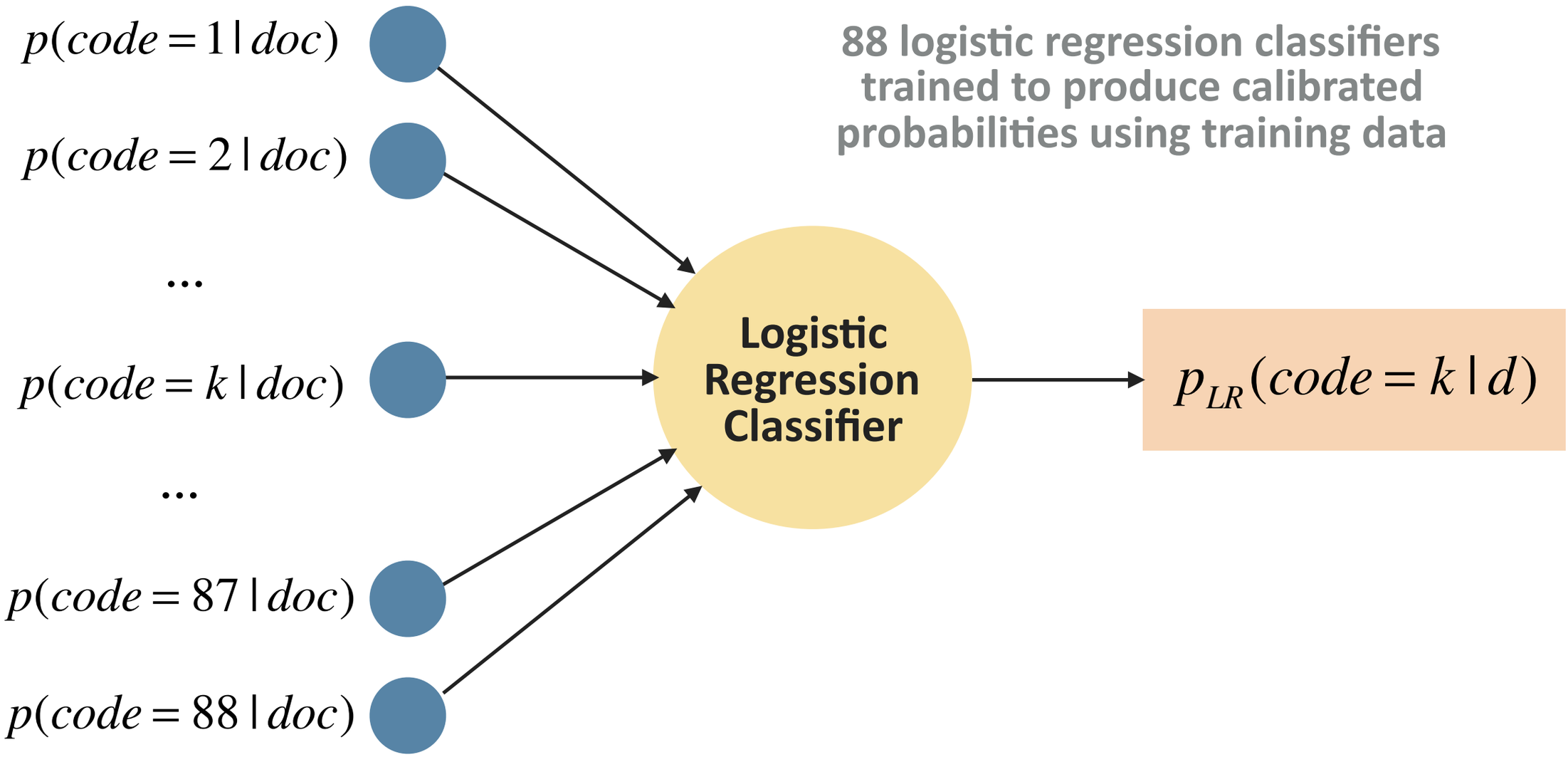
Grants with RCDC labels (31,628 documents)



$$p(code \mid doc) \propto \sum_i p(code \mid word_i, doc)$$

Sampling probabilities were averaged over the words in a document to calculate **AUC** and **R-precision** scores

Logistic Regression Classifier



88 logistic regression classifiers trained to produce calibrated probabilities using training data

L-LDA TOPIC PROBABILITY

CALIBRATED TOPIC PROBABILITY

Evaluation Result

	$p(c \mid d)$	$p_{LR}(c \mid d)$
	L-LDA	**L-LDA + Logistic Regression**
AUC	**0.80**	**0.89**
R-Precision	**0.56**	**0.64**

Analyzing Funding Patterns over Time

- Fractionally assign the funds in direct proportion to the probabilities from the logistic regression classifiers $p_{LR}(code \mid doc)$

$$w_{cd} = \frac{p_{LR}(c \mid d)}{\sum_k p_k(c = k \mid d)} \qquad c = 1,2,...88$$

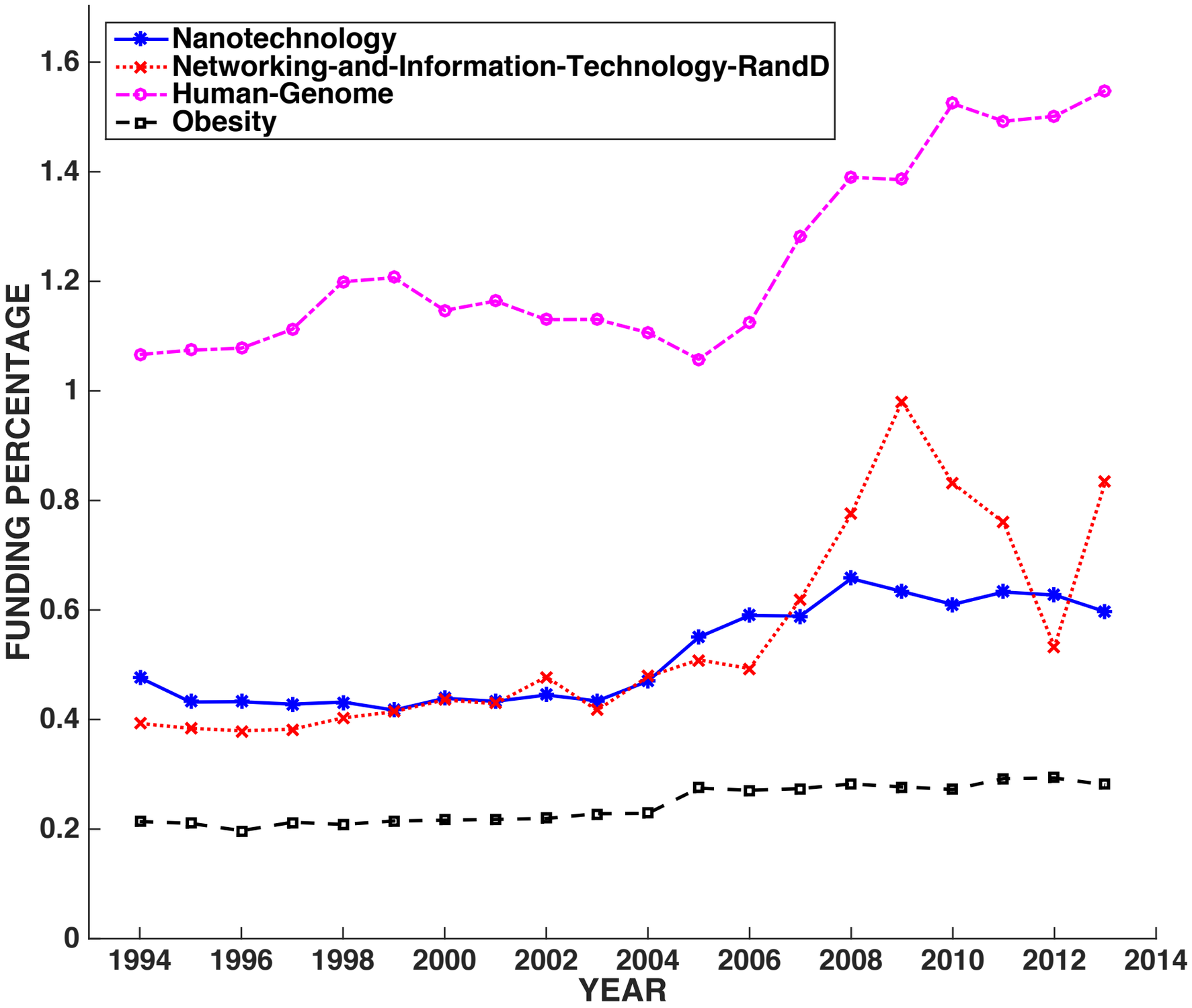
$$F_c^y = \sum_{d:y_d=y} w_{cd} x_d$$

w_{cd} : weight for the category c for document d

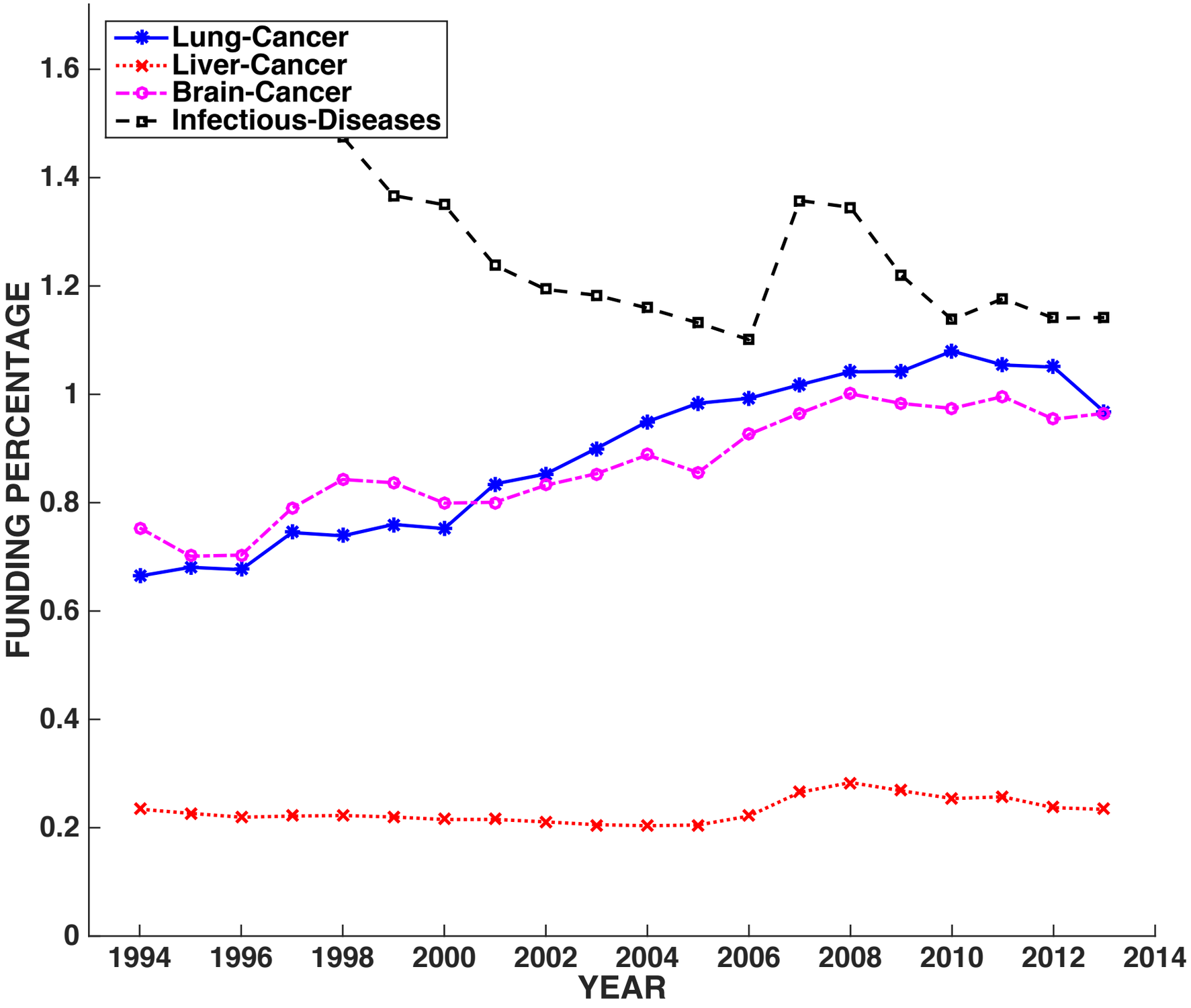
x_d : amount of funding for document d (considered inflation)

y_d : year when document d was funded

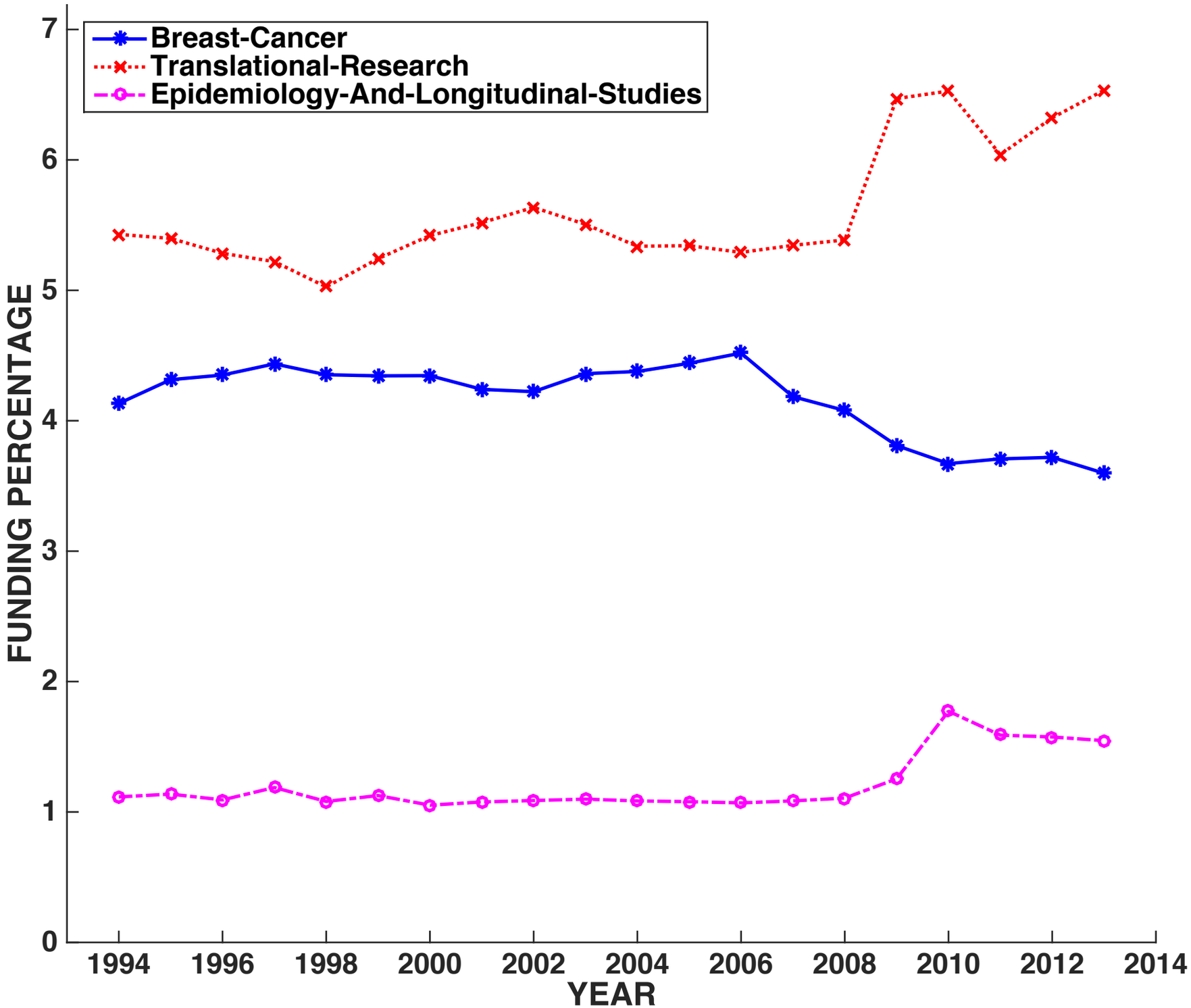
F_c^y : total estimated amount of funding for category c in year y

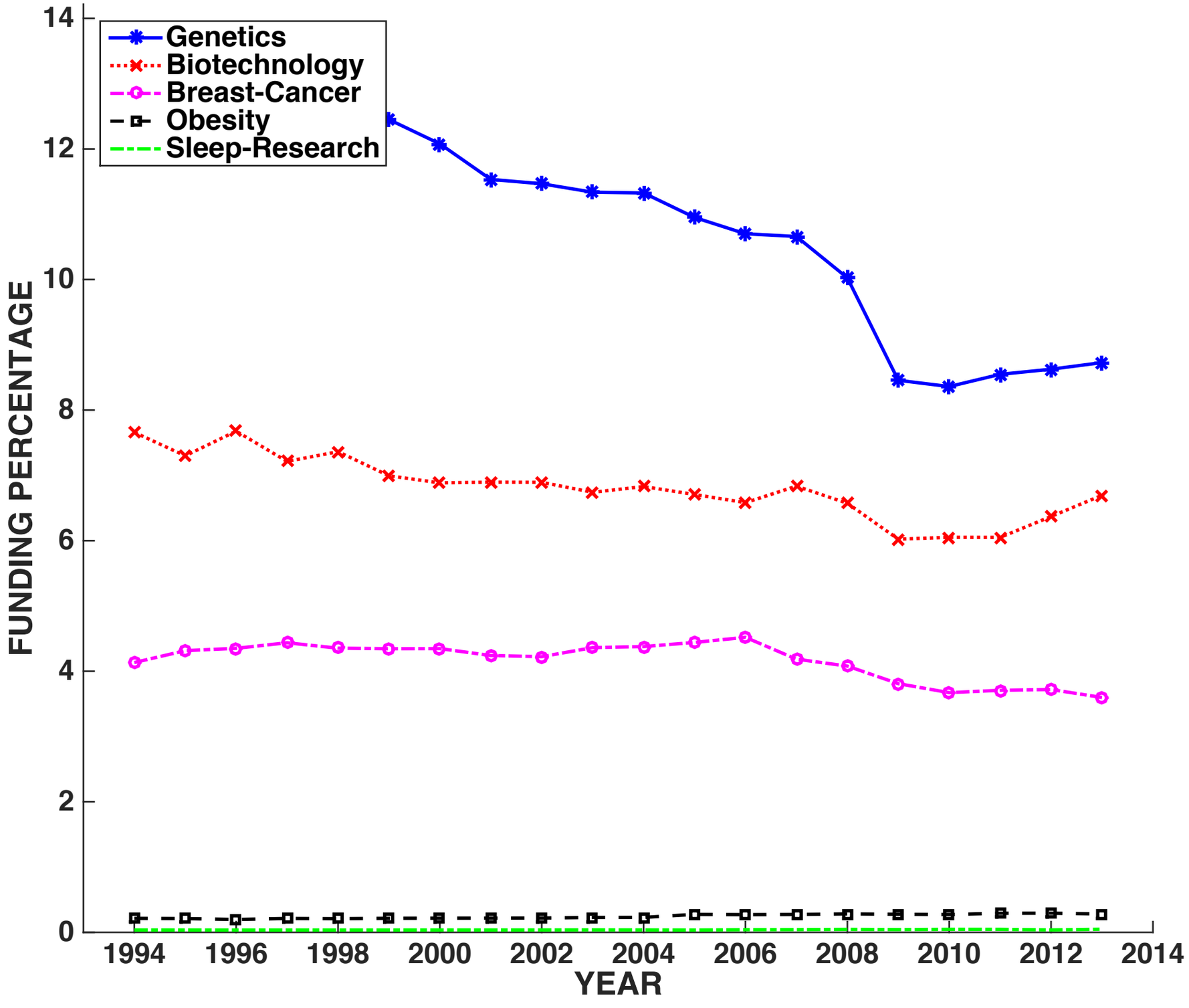


Estimated percentage of funding allocated to 4 general RCDC categories



Estimated percentage of funding allocated to 4 specific RCDC disease categories


Breast-Cancer
Translational-Research
Epidemiology-And-Longitudinal-Studies
FUNDING PERCENTAGE
YEAR
0
1
2
3
4
5
6
7
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014


Genetics
Biotechnology
Breast-Cancer
Obesity
Sleep-Research
FUNDING PERCENTAGE
YEAR
0
2
4
6
8
10
12
14
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014

Conclusions

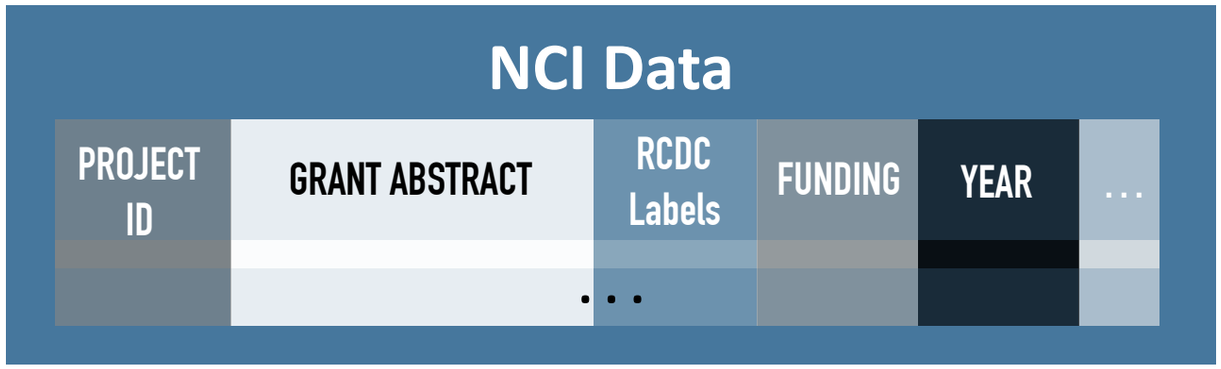
- **Summary**
  - Labeled topic modeling and logistic classifiers can be combined to analyze NIH grant funding data
  - Statistical topic modeling allows linking of text with metadata in a quantifiable manner

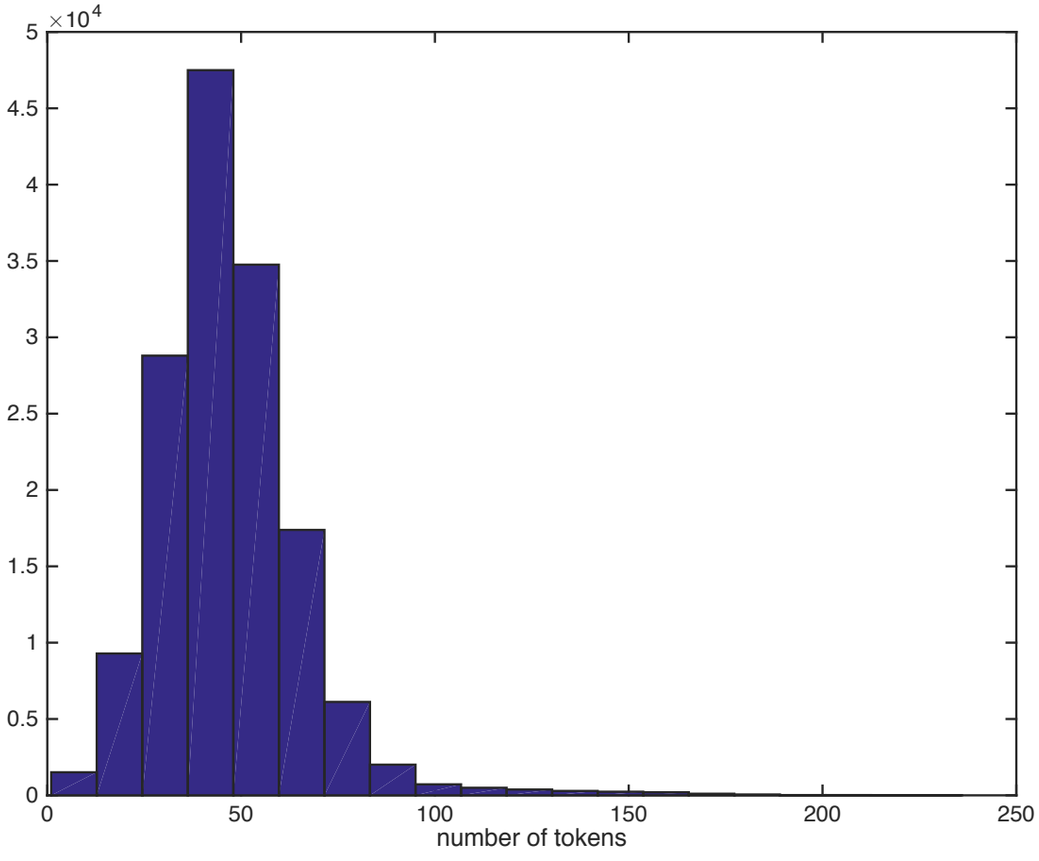
- **Future Work**
  - Jointly analyze grants and scientific articles related to the grants (ongoing)
  - Broader analysis of the economic and policy implications
  - Improvements on topic model
    - How to best calibrate
    - Selecting the right Hyper parameters
    - Methods such as using seed words

THANK YOU

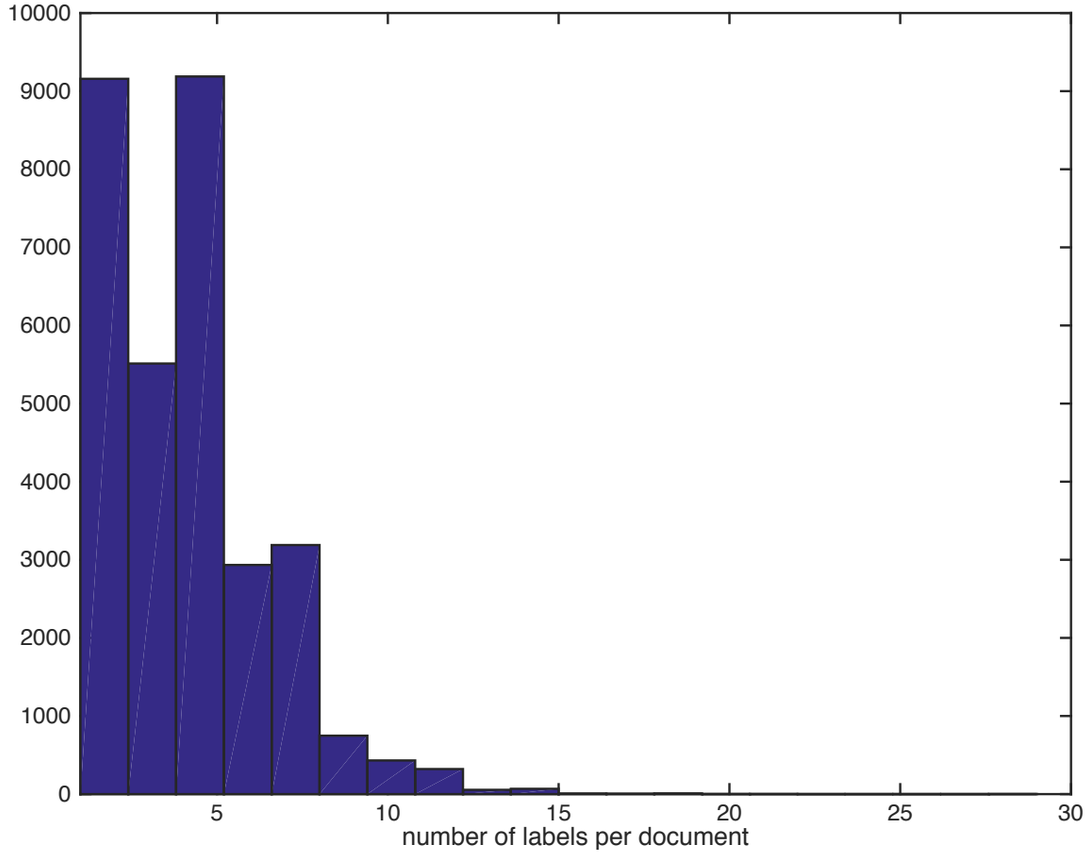
BACKUP SLIDES

NCI (National Cancer Institute) Data





Number of tokens in a document



Number of labels in a document

Examples of Topics from NCI Abstracts (5 out of 88)

Brain Cancer	Breast Cancer	Kidney Disease
glioma	breast cancer	rcc
brain tumor	women	kidney cancer
gbm	breast cancer cell	renal cell carcinoma
malignant glioma	breast	vhl
glioblastoma	breast cancer patient	renal cancer
brain	brca1	pvhl

Hepatitis	Lung Cancer
hcv	gliomalung cancer
hbv	nsclc
liver cancer	lung
hepatitis virus	leading cause
hbv infection	cancer death
hbv replication	egfr

Labeled-LDA for NIC Grants

- 88 Topics (RCDC Codes)
- 10 Background Topics

- Hyper parameters
  - Dirichlet prior for word-topic distribution
    - $\beta_w = 0.01$
  - Dirichlet prior for doc-topic distribution
    - Used proportional alphas

$$\sum_{c=1}^{88} \alpha_c = 5 \qquad \sum_{b}^{B} \alpha_b = 1$$

Improving Gibbs Sampling Predictions on Unseen Data for Latent Dirichlet Allocation

Yannis Papanikolaou YPAPANIK@CSD.AUTH.GR
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece

Timothy N. Rubin TIMRUBIN@INDIANA.EDU
Cognitive Computing Laboratory
Indiana University
Bloomington, IN, USA

Grigorios Tsoumakas GREG@CSD.AUTH.GR
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece

3.2 Averaging the conditional probabilities

Another approach we consider is averaging over the conditional probabilities from Equation 1 for every document. As stated before, p expresses the probability of assigning to a word token a given topic, conditioned on the word, the document, and the other assignments of words to topics in both the given document (through n_{dl}) and in the whole corpus (through n_{wl} during training and ϕ during prediction). The idea is to average these word-level probabilities for every document and obtain a total probability $p(l|d)$ such that:

$$\theta_p(d,l) = p(l|d,\cdot) = \sum_{w=1}^{N_d} p(z_{w_i} = l|\, w_i,\, d) = \sum_{w=1}^{N_d} \phi_l(w) \times \frac{n_{d,l} + \alpha}{\sum_{d=1}^{D} (n_{d,l} + \alpha)} \tag{6}$$

The motivation for this approach is that the conditional probabilities should provide a richer representation of the document-topics distribution. We note here, that as the probability $p(z_{w_i} = l|\, w_i,\, d)$ is unnormalized, an additional step of normalization is required so that $\sum_{l=1}^{L} p(z_{w_i} = l|\, w_i,\, d) = 1$. We will refer to this approach as CGS_p.

Analyzing Funding Patterns over Time

- Fractionally assign the funds in direct proportion to the probability

$$w_{cd} = \frac{p_l(c \mid d)}{\sum_k p_k(c = k \mid d)} \qquad c = 1,2,...88$$

$$F_c^y = \sum_{d:y_d=y} w_{cd} x_d$$

NCI (National Cancer Institute) Data

- **149,901 grants** in total
- for **FY1994 ~ FY2013**

- **Number of grants with labels :** 31,628 (2008~2011)
- **Number of grants without labels :** 118,273

- **Size of vocabulary (W)** : 29,713