CLÉMENT RIEDEL
clemriedel
clement.riedel

Natural Language Quantification

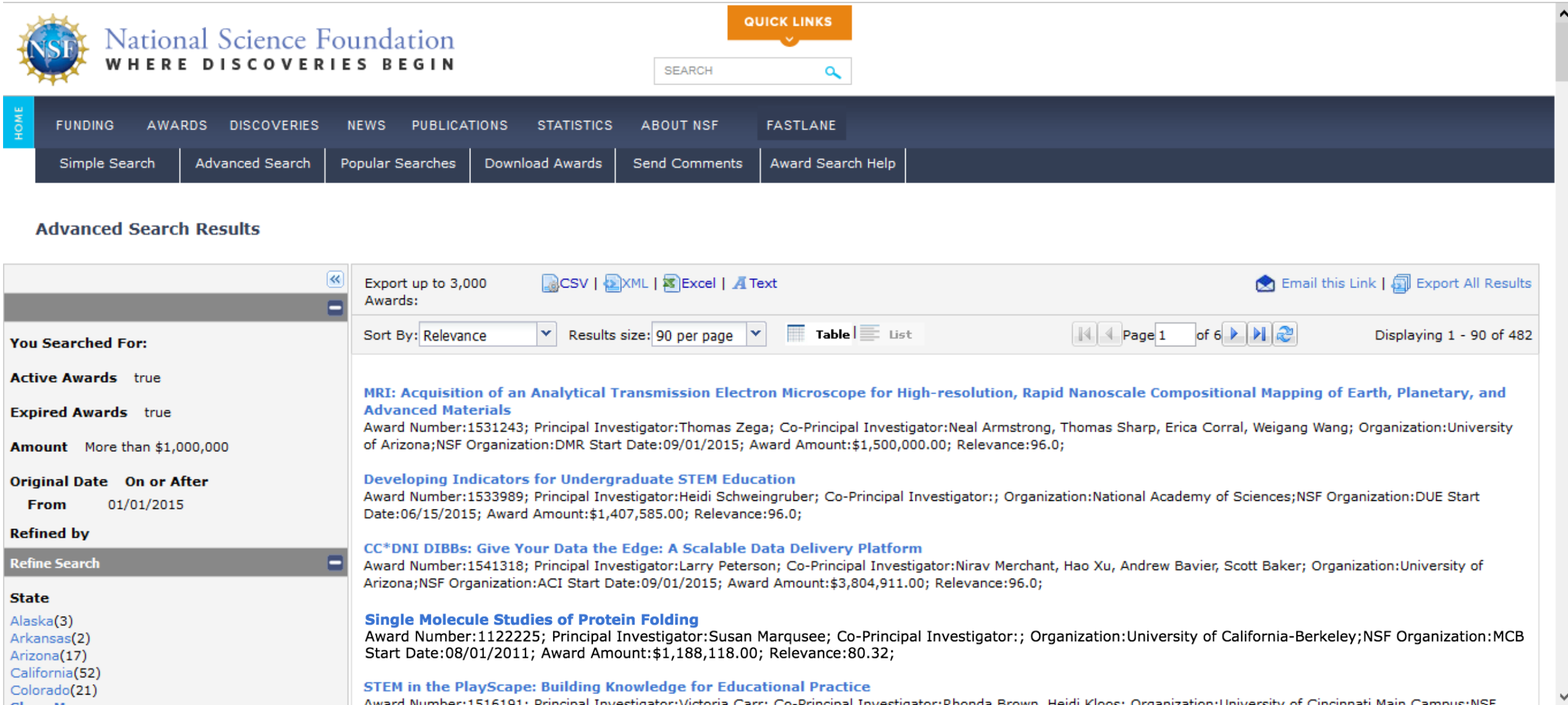
ds4all.io

1/ Find: patterns, topics, concepts

2/ Quantify: money

3/ Create: high-impact documents

Case study: Topic modeling of NSF grants



Data

Award Abstract #1122225

Single Molecule Studies of Protein Folding

NSF Org:	MCB Div Of Molecular and Cellular Bioscience
Initial Amendment Date:	June 20, 2011
Latest Amendment Date:	July 1, 2015
Award Number:	1122225
Start Date:	August 1, 2011
Awarded Amount to Date:	$1,188,118.00
Investigator(s):	Susan Marqusee marqusee@berkeley.edu (Principal Investigator)
Sponsor:	University of California-Berkeley Sponsored Projects Office BERKELEY, CA 94704-5940 (510)642-8109
NSF Program(s):	Molecular Biophysics

Number
M = 1.2 M$

ABSTRACT

The objective of this project is to investigate the molecule studies. Protein folding, the mechanism the energy landscape of a protein (the structures all of the accessible conformations), remains a m biology. Without a better understanding of how t sequence, the expanding sequence databases co important questions. The project aims to develop a single protein molecule using tools such as an monitoring the conformation of the protein by fo in the protein. The results will yield observables studies of protein folding and the design of the e iterative collaboration with theoreticians in the field. Mechanical stress plays a ubiquitous role in protein function and biology, and therefore, in addition to providing needed insight into the protein-folding problem this work will also provide much needed information about the mechanical stability of proteins.

In addition to its impact in protein folding and macromolecular biophysics, the research in this project requires training at the intersection of physics and biology: the work is a collaborative effort between a physics and a biochemistry lab. The work is not just interdisciplinary; it is actually interdependent, requiring state-of-the-art experiments in both physics and molecular biology. The work requires a combination of students and postdoctoral trainees with diverse backgrounds to work closely together. The project also involves the help of undergraduates, particularly those with a background in engineering and physics, giving them direct exposure to the biological sciences. In sum, this work offers a unique educational experience for the students and post-doctoral fellows involved in the project, and will help to train a new generation of scientists fluent in both the biological and quantitative sciences.

Finding topics

Blei, Ng & Jordan: Latent Dirichlet Alocation, *JMLR* 3 (2003) 993-1022

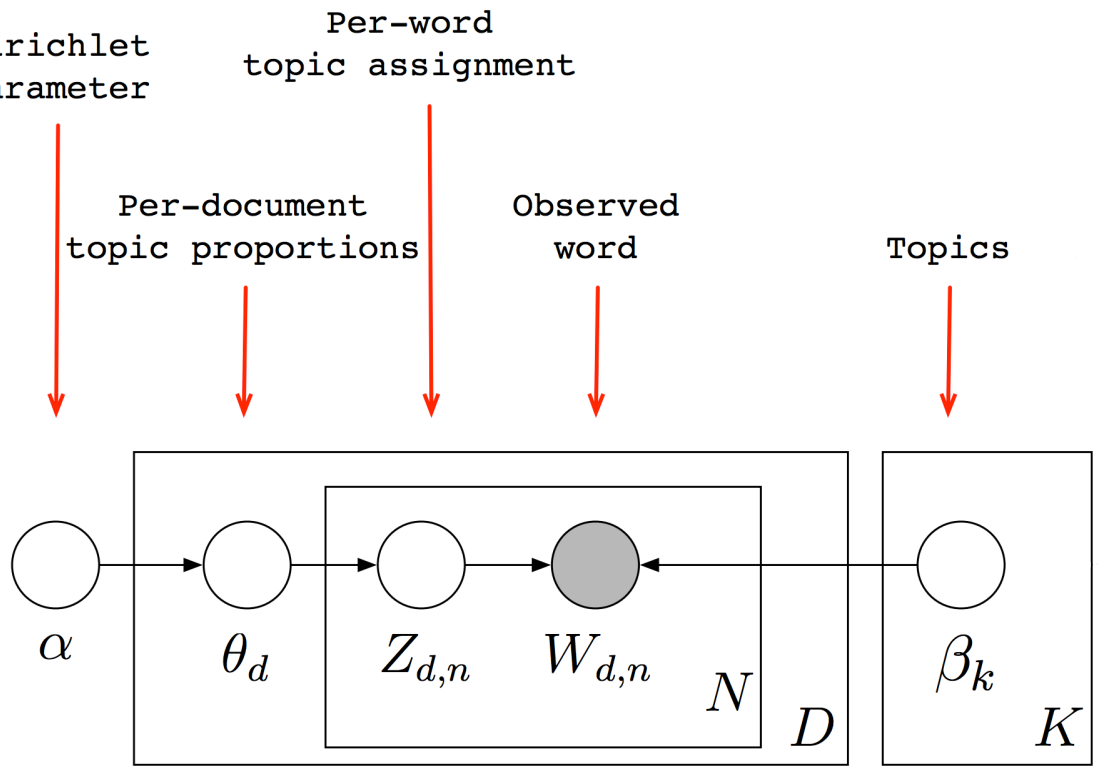


Plate notation

ABSTRACT

The objective of this project is to investigate the molecule studies. Protein folding, the mechanism the energy landscape of a protein (the structures all of the accessible conformations), remains a m biology. Without a better understanding of how t sequence, the expanding sequence databases co important questions. The project aims to develop a single protein molecule using tools such as an monitoring the conformation of the protein by fo in the protein. The results will yield observables studies of protein folding and the design of the e iterative collaboration with theoreticians in the field. Mechanical stress plays a ubiquitous role in protein function and biology, and therefore, in addition to providing needed insight into the protein-folding problem this work will also provide much needed information about the mechanical stability of proteins.

In addition to its impact in protein folding and macromolecular biophysics, the research in this project requires training at the intersection of physics and biology: the work is a collaborative effort between a physics and a biochemistry lab. The work is not just interdisciplinary; it is actually interdependent, requiring state-of-the-art experiments in both physics and molecular biology. The work requires a combination of students and postdoctoral trainees with diverse backgrounds to work closely together. The project also involves the help of undergraduates, particularly those with a background in engineering and physics, giving them direct exposure to the biological sciences. In sum, this work offers a unique educational experience for the students and post-doctoral fellows involved in the project, and will help to train a new generation of scientists fluent in both the biological and quantitative sciences.

Topic composition

Number of awards > 5k
Number of words > 1M

Biological Science
Received after 2015

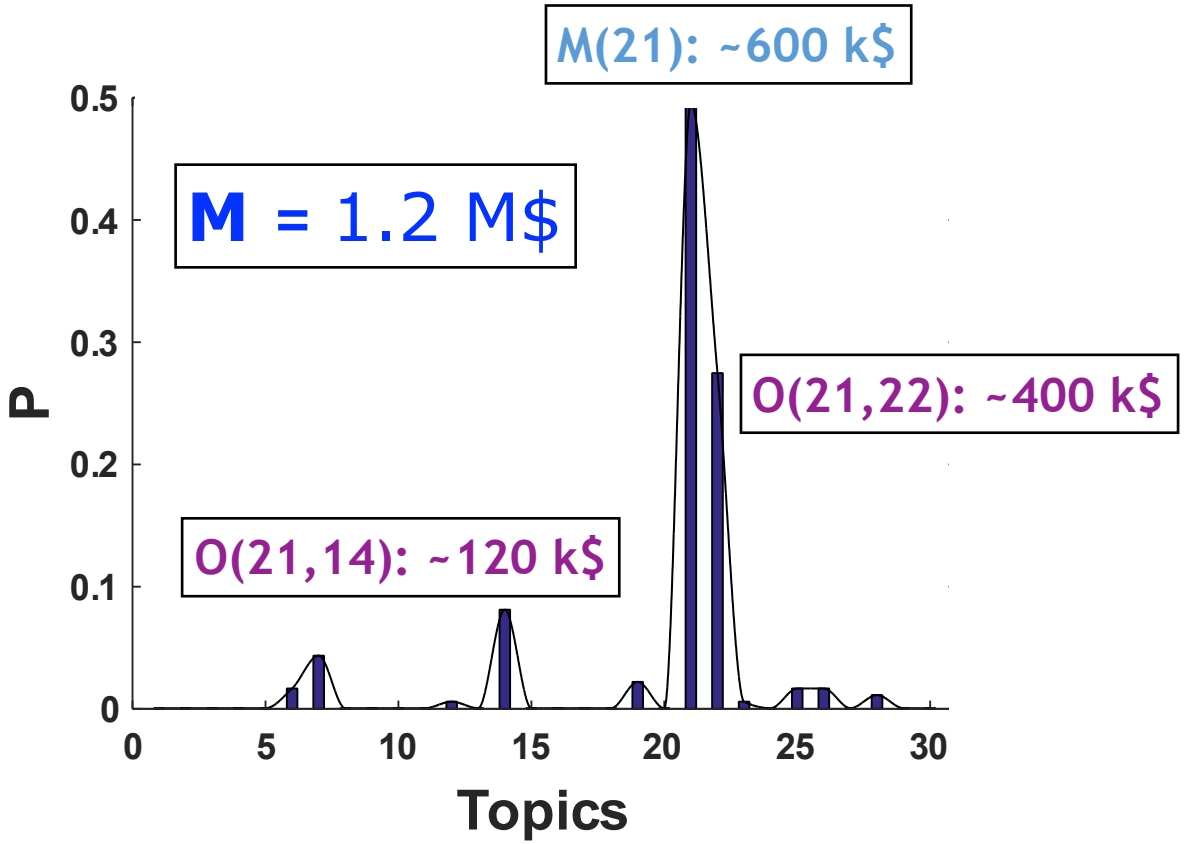
Topic 1: behavior social female male animal species individuals
Topic 2: cell membrane protein transport molecular cellular biology
Topic 3: fellowship biology animal training plan fellow include
Topic 4: bacteria metabolism microbes production energy pathways cell
Topic 5: cell development signaling protein gene molecular mechanisms
Topic 6: data develop model methods tools computational biology
Topic 7: researchers science network biology community scientific
Topic 8: fungi soil plant microbes carbon ecosystem decomposition
Topic 9: gene evolution species population variation different change
Topic 10: pollen habitat bees plant pollination change sperm
Topic 11: forest climate change ecosystem model tree data
Topic 12: students program reu biology edu training pi
Topic 13: gene dna chromatin epigene expression methylation
Topic 14: biology model systems network develop program division
Topic 15: species population plant diversity ecological effects
Topic 16: enzyme protein iron amino acid reaction function
Topic 17: university state students researchers system center
Topic 18: marine ecosystem ocean sea fish animal algae
Topic 19: gene rna protein expression transcription regulatory
Topic 20: cell imaging biology high resolution instrument develop
Topic 21: protein structure molecular interactions binding function folding
Topic 22: students undergraduate science school high graduate biology
Topic 23: dna gene chromosome repair cell protein bacteria
Topic 24: host disease pathogen pathogens immune bacteria plant
Topic 25: species diversity evolution data life tree phylogene
Topic 26: change environmental stress conditions climate responses
Topic 27: plant gene crop arabidopsis growth development production
Topic 28: brain neurons behavior neural animal system sensory
Topic 29: collection data specimens change resource biology available
Topic 30: nitrogen carbon stream nutrient water dioxide fixation

Topic proportion for one doc:

Award Abstract #1122225

Single Molecule Studies of Protein Folding

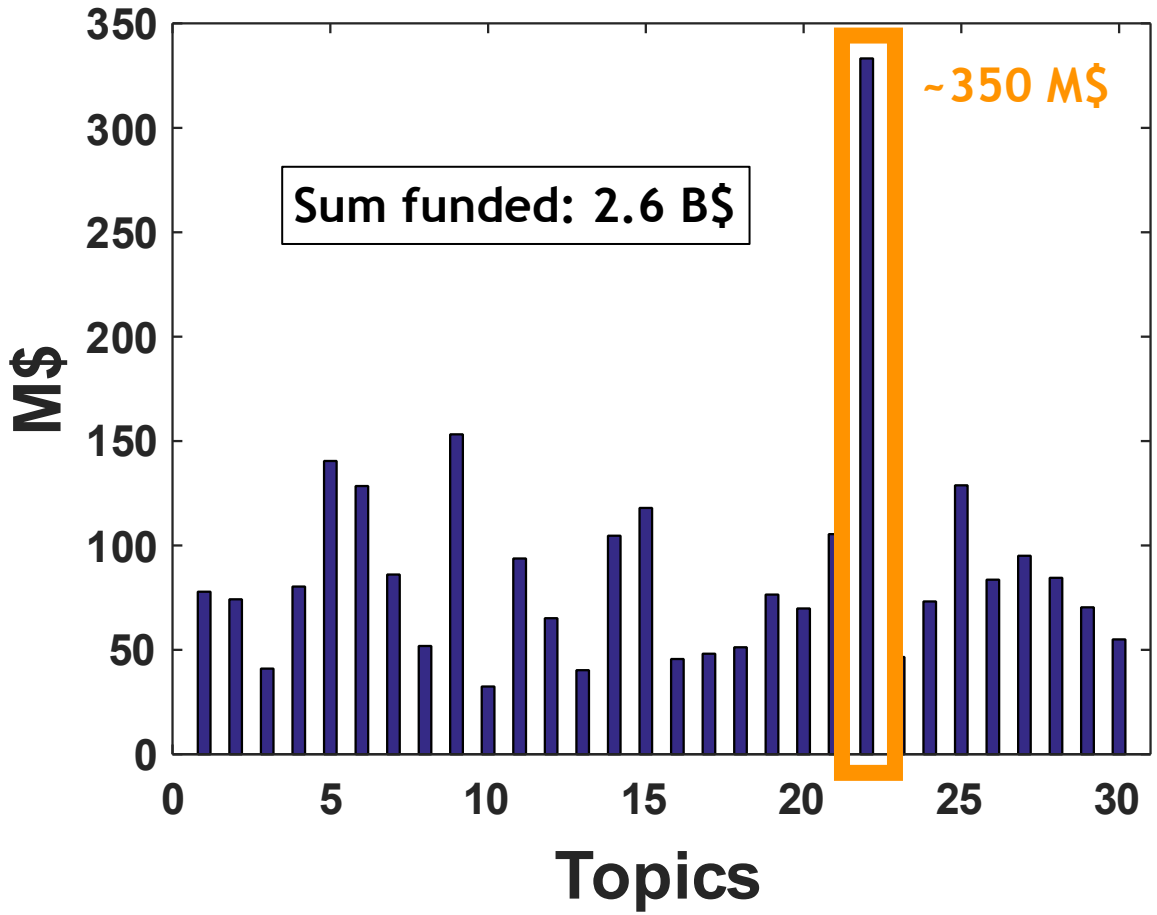
Topic 21: protein structure molecular interactions binding function folding
Topic 22: students undergraduate science school high graduate biology
Topic 14: biology model systems network develop program division



Topic 1: behavior social female male animal species individuals
Topic 2: cell membrane protein transport molecular cellular biology
Topic 3: fellowship biology animal training plan fellow include
Topic 4: bacteria metabolism microbes production energy pathways cell
Topic 5: cell development signaling protein gene molecular mechanisms
Topic 6: data develop model methods tools computational biology
Topic 7: researchers science network biology community scientific
Topic 8: fungi soil plant microbes carbon ecosystem decomposition
Topic 9: gene evolution species population variation different change
Topic 10: pollen habitat bees plant pollination change sperm
Topic 11: forest climate change ecosystem model tree data
Topic 12: students program reu biology edu training pi
Topic 13: gene dna chromatin epigene expression methylation development
Topic 14: biology model systems network develop program division
Topic 15: species population plant diversity ecological effects communities
Topic 16: enzyme protein iron amino acid reaction function
Topic 17: university state students researchers system center california
Topic 18: marine ecosystem ocean sea fish animal algae
Topic 19: gene rna protein expression transcription regulatory regulation
Topic 20: cell imaging biology high resolution instrument develop
Topic 21: protein structure molecular interactions binding function folding
Topic 22: students undergraduate science school high graduate biology
Topic 23: dna gene chromosome repair cell protein bacteria
Topic 24: host disease pathogen pathogens immune bacteria plant
Topic 25: species diversity evolution data life tree phylogene
Topic 26: change environmental stress conditions climate responses
Topic 27: plant gene crop arabidopsis growth development production
Topic 28: brain neurons behavior neural animal system sensory
Topic 29: collection data specimens change resource biology available
Topic 30: nitrogen carbon stream nutrient water dioxide fixation

Money generated by each topic over all the grants

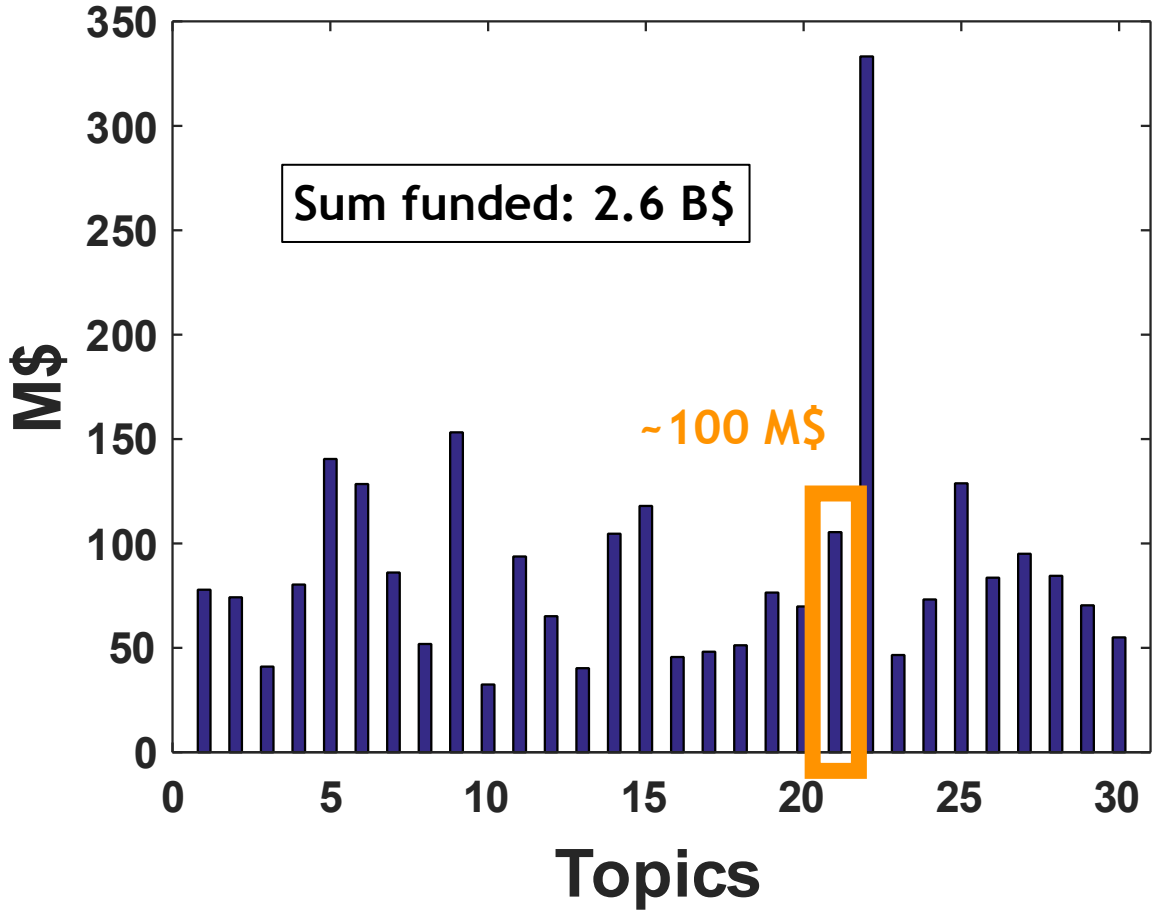
Biological Science (BIO) after 2015, above 100 k$



Topic 1: behavior social female male animal species individuals
Topic 2: cell membrane protein transport molecular cellular biology
Topic 3: fellowship biology animal training plan fellow include
Topic 4: bacteria metabolism microbes production energy pathways cell
Topic 5: cell development signaling protein gene molecular mechanisms
Topic 6: data develop model methods tools computational biology
Topic 7: researchers science network biology community scientific
Topic 8: fungi soil plant microbes carbon ecosystem decomposition
Topic 9: gene evolution species population variation different change
Topic 10: pollen habitat bees plant pollination change sperm
Topic 11: forest climate change ecosystem model tree data
Topic 12: students program reu biology edu training pi
Topic 13: gene dna chromatin epigene expression methylation development
Topic 14: biology model systems network develop program division
Topic 15: species population plant diversity ecological effects communities
Topic 16: enzyme protein iron amino acid reaction function
Topic 17: university state students researchers system center california
Topic 18: marine ecosystem ocean sea fish animal algae
Topic 19: gene rna protein expression transcription regulatory regulation
Topic 20: cell imaging biology high resolution instrument develop
Topic 21: protein structure molecular interactions binding function folding
Topic 22: students undergraduate science school high graduate biology
Topic 23: dna gene chromosome repair cell protein bacteria
Topic 24: host disease pathogen pathogens immune bacteria plant
Topic 25: species diversity evolution data life tree phylogene
Topic 26: change environmental stress conditions climate responses
Topic 27: plant gene crop arabidopsis growth development production
Topic 28: brain neurons behavior neural animal system sensory
Topic 29: collection data specimens change resource biology available
Topic 30: nitrogen carbon stream nutrient water dioxide fixation

Money generated by each topic over all the grants

Biological Science (BIO) after 2015, above 100 k$



Top 10 topics generating money

Topic 22: “Education” 333 M$: students undergraduate science school high graduate biology

Topic 9: “Genetics” 153 M$: gene evolution species population variation different change

Topic 5: “Cell” 140 M$: cell development signaling protein gene molecular mechanisms

Topic 25: “Evolution” 129 M$: species diversity evolution data life tree phylogene

Topic 6: “Modelisation” 128 M$: data develop model methods tools computational biology

Topic 15: “Ecology” 118 M$: species population plant diversity ecological effects communities

Topic 21: “Protein” 105 M$: protein structure molecular interactions binding function folding

Topic 14: “General” 105 M$: biology model systems network develop program division

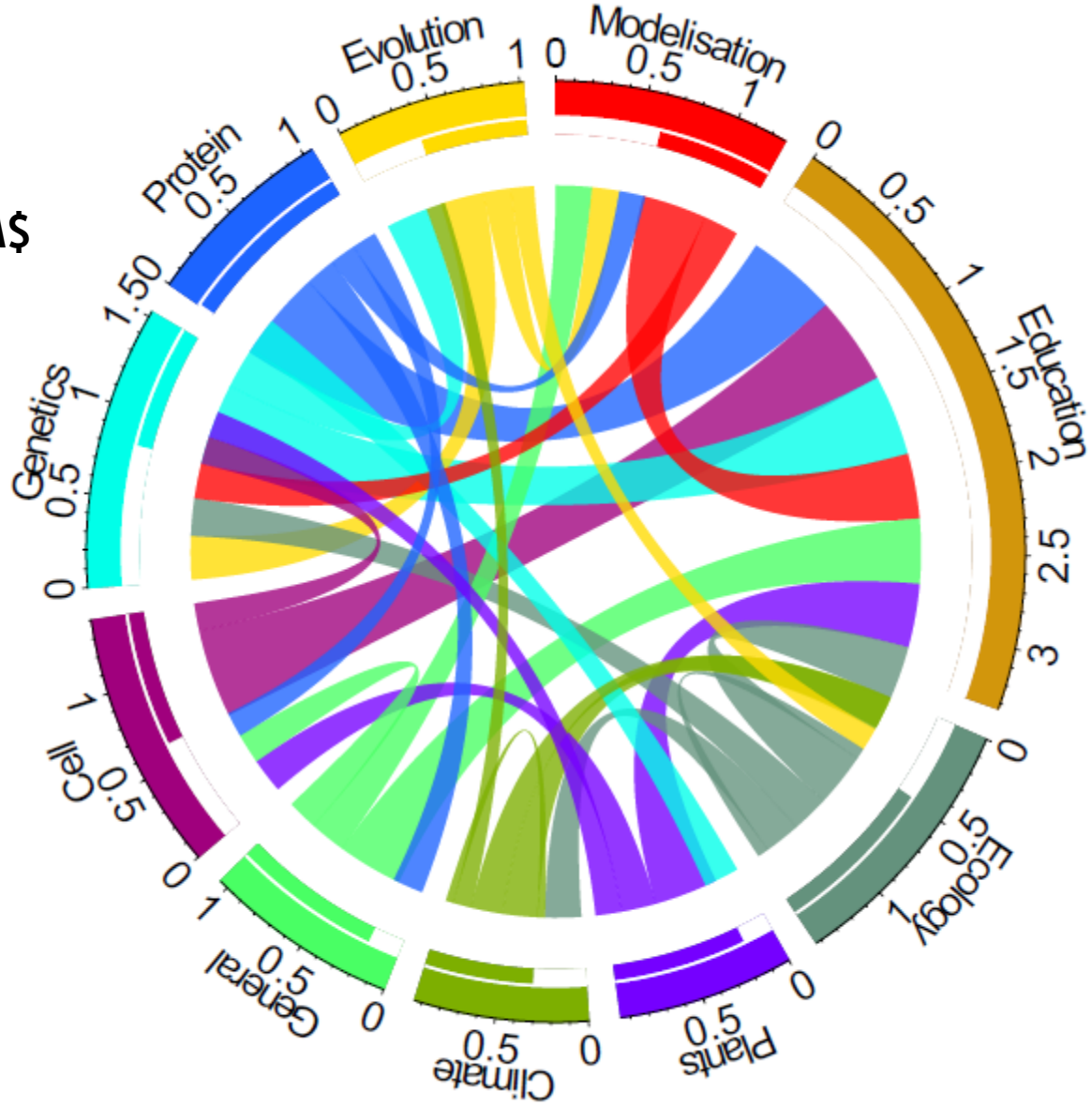
Topic 27: “Plants” 95 M$: plant gene crop arabidopsis growth development production

Topic 11: “Climate” 94 M$: forest climate change ecosystem model tree data

Circular plot to represent topics interaction

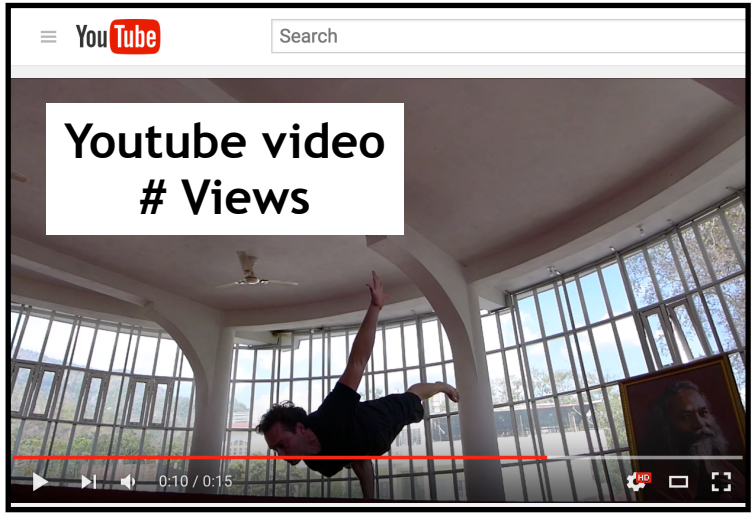
Each unit represents 100 M$
Only major flux are plotted.

"Protein" has generated **105 M$**
and overlapped with:
"Education" for **50 M$**
"Modelisation" for **10 M$**
"General" **for 10 M$**
"Cell" for **10M$**

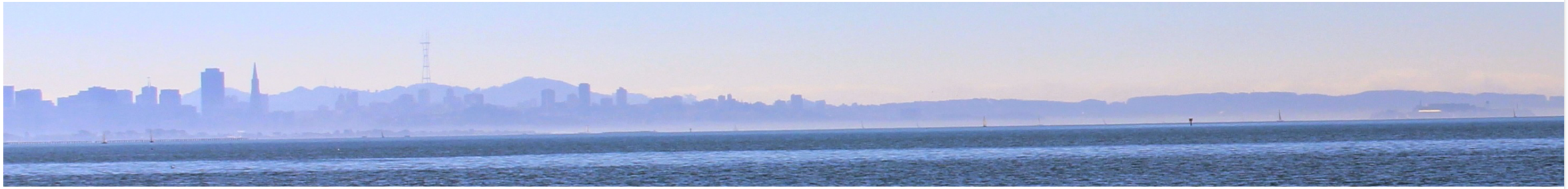


1/ Find topics
2/ Quantify money
3/ High impact

"Data are everywhere and they are for everyone"



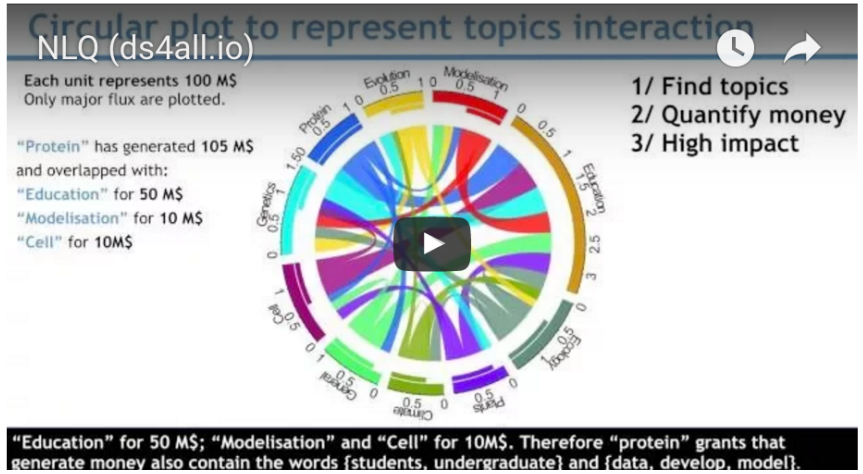
Any activity described by words and quantified by numbers.



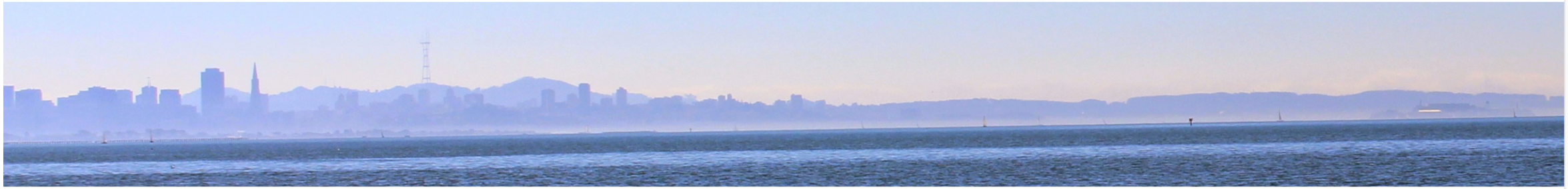
Welcome to 'Data Science for all' ! Our mission is to empower people by providing a simple and free access to data science. Natural Language Quantification (NLQ) is the first product of ds4all.io, it can find patterns (or topics) describing an activity and quantify how much is being created by the pattern (for instance in the National Science Fundation grants, the topic "Protein" has generated 100 M$...). You can also check a different prodcut: a full end to end data science pipeline (click on the ds4all dropdown menu on the upper left, or here).

Natural Language Quantification

You can simply upload your data, run the algorithm and get your report. Check the 'Tutorial' if you would like more information about how to perform this process. Inspiration and examples of databases (use them and get your first results in seconds) can be found in the 'Projects'. Click on 'Education' to learn more about data science & get guidance to build your own algorithm from scratch. This short (3 min) video summarizes everything:



ds4all.io



NLQ Report: NSF - Politics

Topic 1: benefit cost individual group provide good effect incentive

Topic 2: analysi maker actor condition action result processe outcome

Topic 3: citie acros area researcher resident investigator force consequence

Topic 4: state leader investigator researcher factor crisi les scholar

Topic 5: interest state rule congres legislator group legislature member

Topic 6: attitude citizen individual value survey difference question experiment

Topic 7: religiou group interview politic analysi way identitie processe

Topic 8: analysi effect acros measure policie change program variou

Topic 9: women policie right movement state statu succes politic

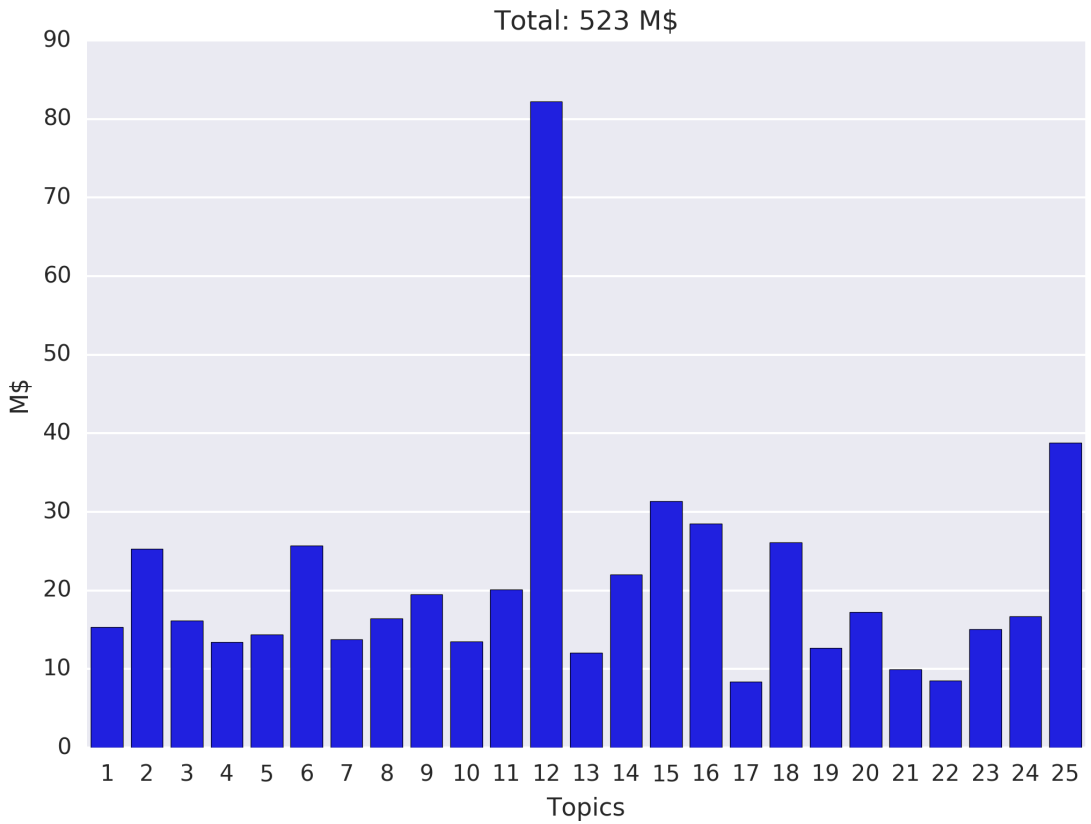
Topic 10: preference policie government decision agencie countrie firm analysi

Topic 11: scientist system proces problem method politic u way

Topic 12: student scholar scientist participant issue activitie method provide

Topic 13: election voter candidate partie campaign citizen effect position

Topic 14: practice processe analysi interview studie organization technologie activitie



clement.riedel@gmail.com

clemriedel clement.riedel

‘Data Science for all’

ds4all.io

Natural Language Quantification

Thank you

Clement Riedel

Nod to the canon

NSF promotes the progress of science; advances the national health, prosperity, and welfare; secures the national defense; provides open access to data.

Blei, **Ng** & **Jordan**: Latent Dirichlet Alocation, *Journal of Machine Learning Research* **3** (2003) 993-1022

Python is powerful... and fast; plays well with others; runs everywhere; is friendly & easy to learn; is Open.

Numpy the multi-dimensional container of generic data.

Allen B. Riddell "Topic modeling with latent Dirichlet allocation in Python"

Octave is a high-level interpreted language, primarily intended for numerical computations. It' free.

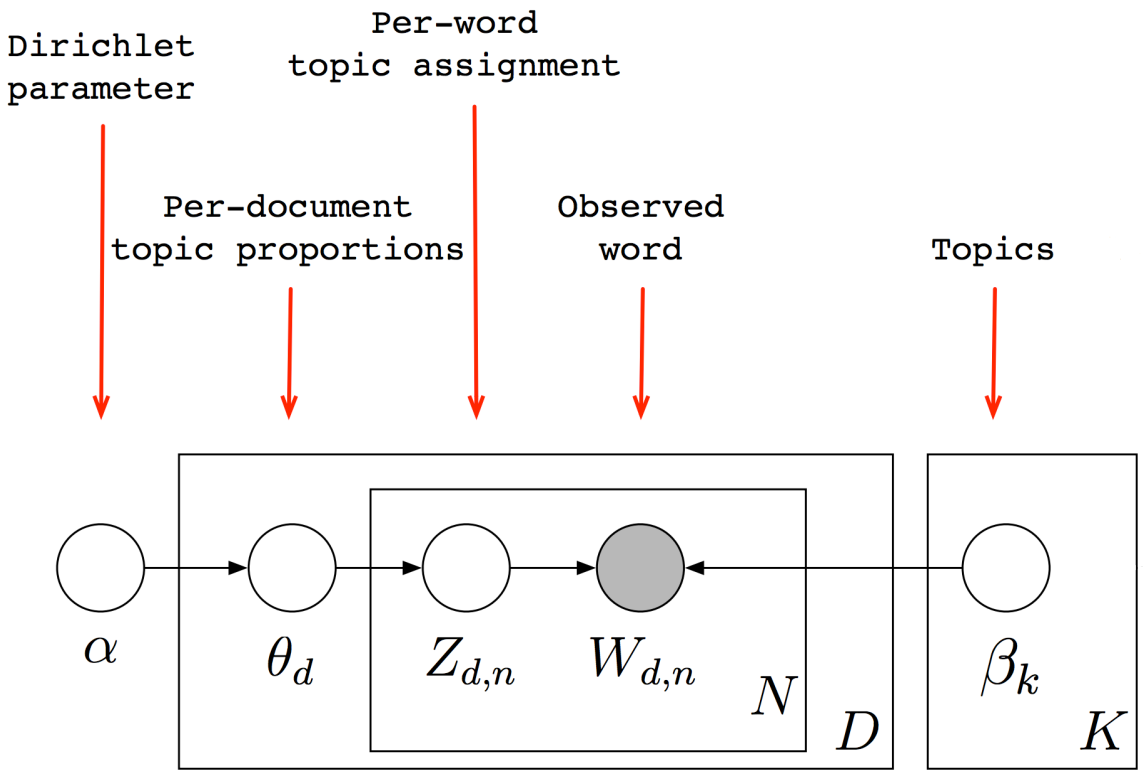
R is a language and environment for statistical computing and graphics. It is open. **Zuguang Gu** (Circlize package). **GJ Abel, N Sander** "Quantifying global international migration flows". – *Science* (2014), **343**, 6178, p. 1520. for the polar plots.

Galvanize DSI. Everyone I saw since I started April 2016.

Dav and Susan for their support.

LDA

Blei, Ng & Jordan: Latent Dirichlet Alocation, *JMLR* 3 (2003) 993-1022



$$\frac{p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}{\int_\theta p(\theta \mid \alpha) \prod_{n=1}^{N} \sum_{z=1}^{K} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}$$

ABSTRACT

The objective of this project is to inves
molecule studies. Protein folding, the r
the energy landscape of a protein (the
all of the accessible conformations), re
biology. Without a better understandin
sequence, the expanding sequence dat
important questions. The project aims
a single protein molecule using tools su
monitoring the conformation of the pro
in the protein. The results will yield obs
studies of protein folding and the desig
iterative collaboration with theoreticiar
role in protein function and biology, an
into the protein-folding problem this work will also provide much needed information about
the mechanical stability of proteins.

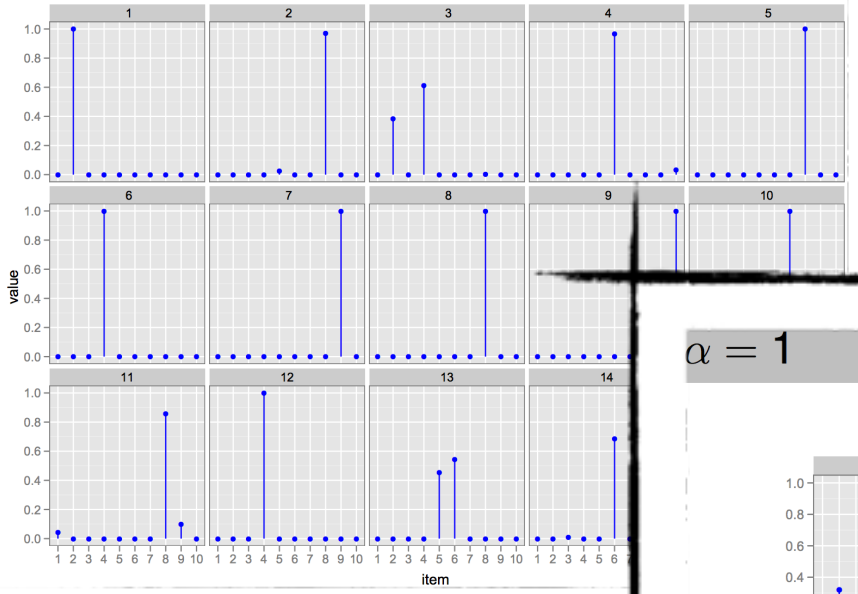
In addition to its impact in protein folding and macromolecular biophysics, the research in this project requires training at the intersection of physics and biology: the work is a collaborative effort between a physics and a biochemistry lab. The work is not just interdisciplinary; it is actually interdependent, requiring state-of-the-art experiments in both physics and molecular biology. The work requires a combination of students and postdoctoral trainees with diverse backgrounds to work closely together. The project also involves the help of undergraduates, particularly those with a background in engineering and physics, giving them direct exposure to the biological sciences. In sum, this work offers a unique educational experience for the students and post-doctoral fellows involved in the project, and will help to train a new generation of scientists fluent in both the biological and quantitative sciences.

α

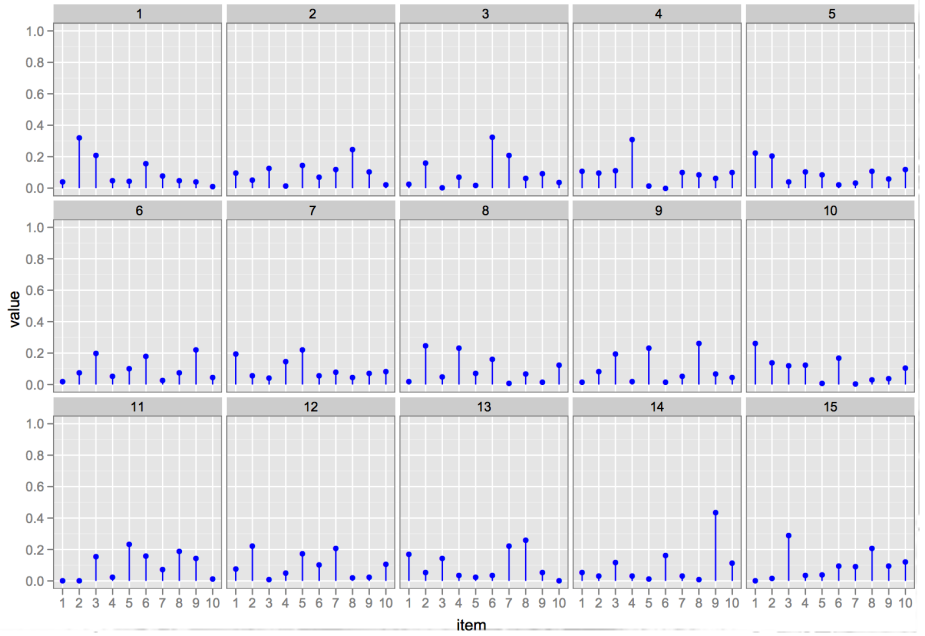
Dirichlet parameter

plots from **David Blei for KDD-2011**

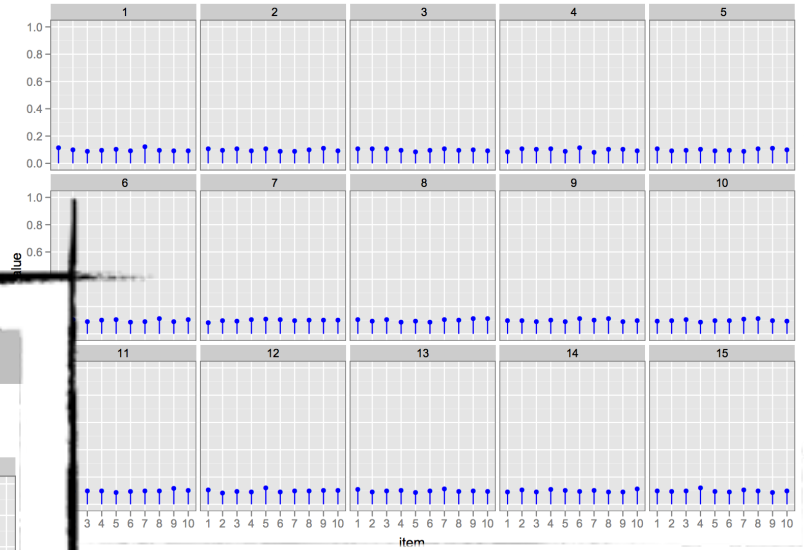
$\alpha = 0.01$



$\alpha = 1$



$\alpha = 100$



Input: Term document matrix

(~300 MB)

Document 1: "The objective of this project is to investigate the protein-folding problem using single molecule studies. Protein folding, the mechanism by which the amino acid sequence directs the energy landscape of a protein (the structures, energies…."

⋮

Document 5284: "Each year honey bees contribute billions of dollars of value to the United States economy by pollinating food crops. During the past decade, North American honey bee populations have declined, primarily because of the death of colonies during winter…"

1/ Remove stop words & stemming

2/ Add words to the term document matrix

3/ Solve posterior using collapsed Gibbs sampling

4/ Obtain: Topics composition & Topics proportion per document

	objective	project	investigate	protein	folding	….	bees	pollen	honey	….
Doc 1	1	5	1	12	5	….	0	0	0	….
⋮	⋮	⋮	⋮	⋮	⋮	….	⋮	⋮	⋮	….
Doc 5284	1	0	0	0	0	….	12	1	1	….

Number of awards

Number of unique words

Output: List of topics & topic table (csv)

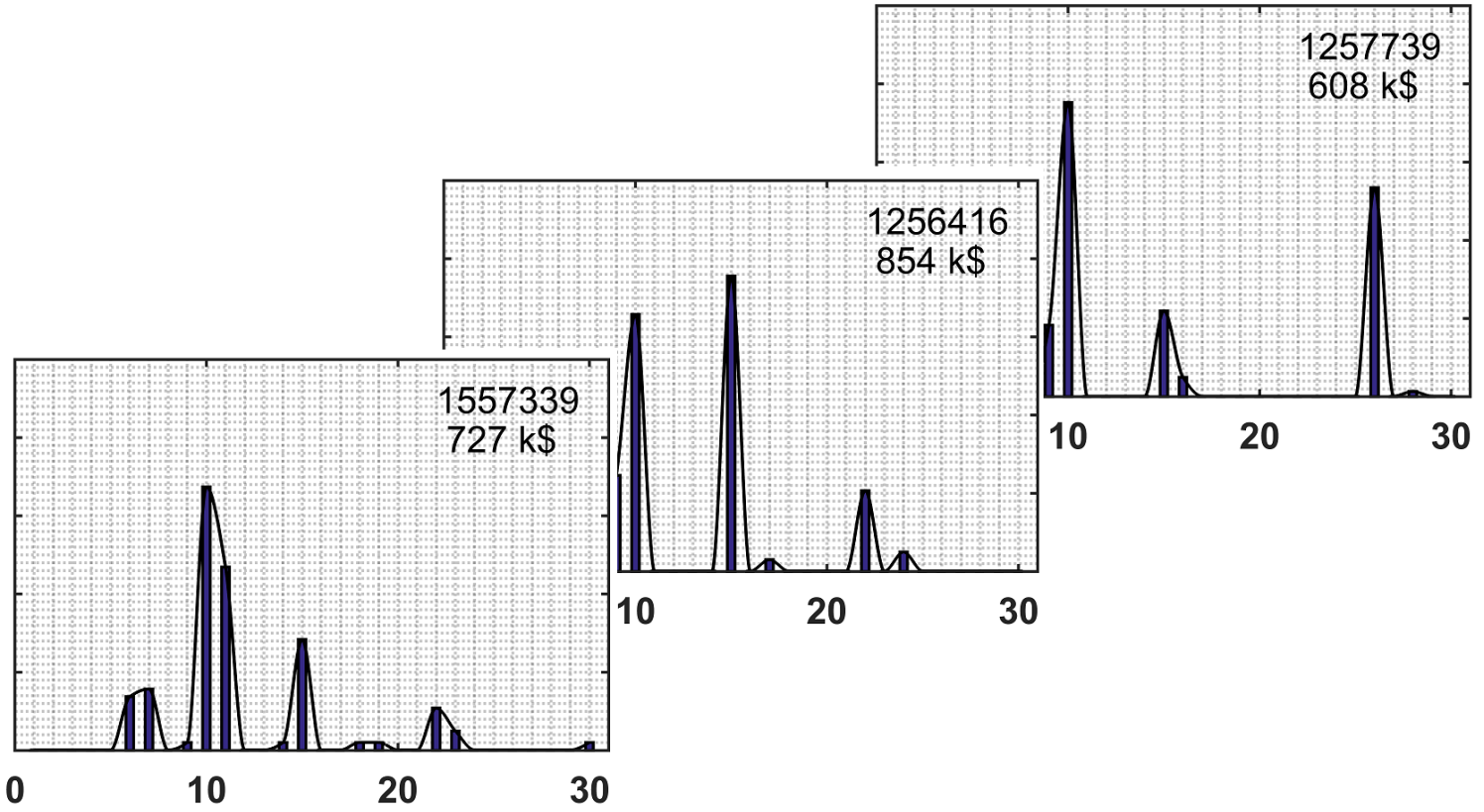
(~1.5 MB)

List of topics:

Topic 1: behavior social female male animal species
Topic 2: cell membrane protein transport molecular
Topic 3: fellowship biology animal training plan

⋮

Topic 29: collection data specimens change resource
Topic 30: nitrogen carbon stream nutrient water



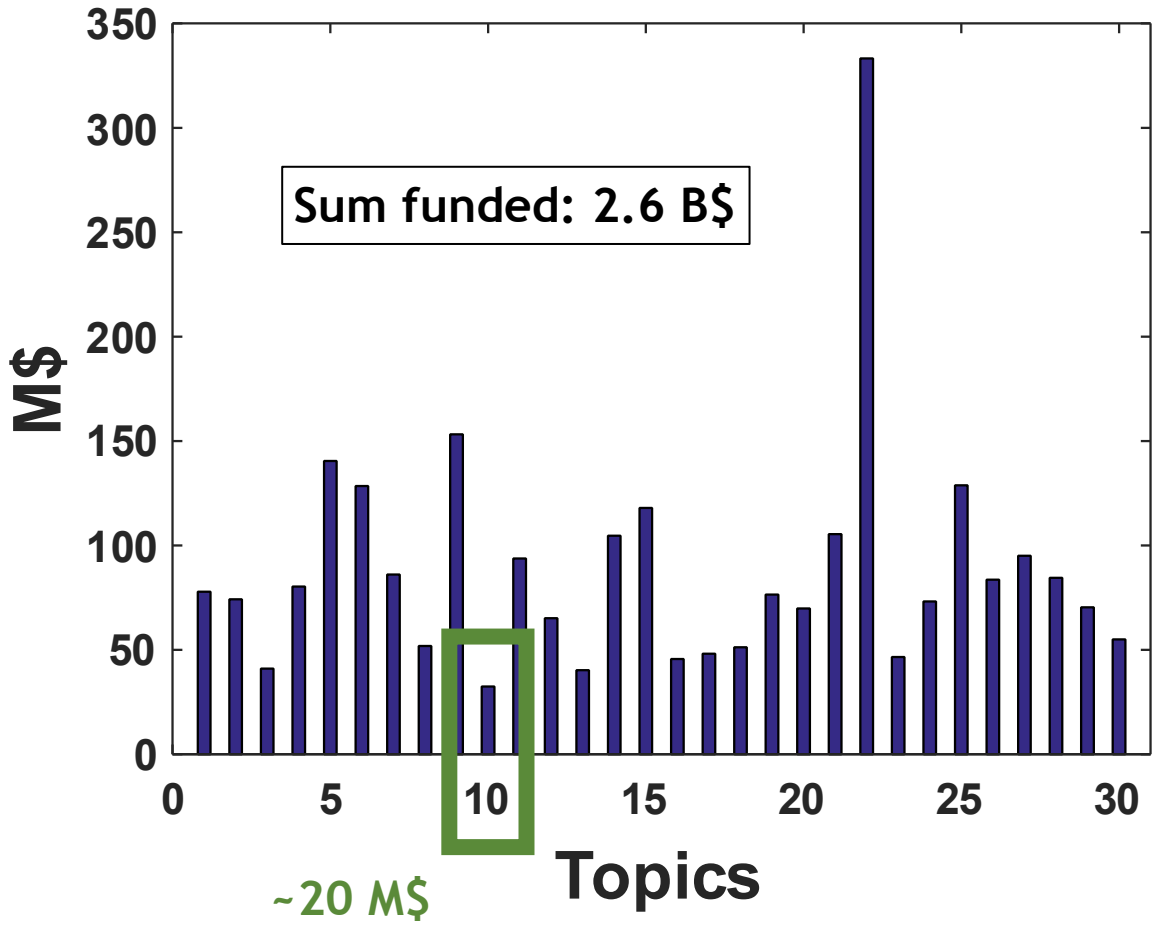
Topic table:

	Topic 1	Topic 2	Topic 3	...	Topic 29	Topic 30
Doc 1	0.2	0.1	0.02	...	0.1	0.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Doc 5284	0.3	0.1	0.01	...	0.1	0.1

Topic 1: behavior social female male animal species individuals
Topic 2: cell membrane protein transport molecular cellular biology
Topic 3: fellowship biology animal training plan fellow include
Topic 4: bacteria metabolism microbes production energy pathways cell
Topic 5: cell development signaling protein gene molecular mechanisms
Topic 6: data develop model methods tools computational biology
Topic 7: researchers science network biology community scientific
Topic 8: fungi soil plant microbes carbon ecosystem decomposition
Topic 9: gene evolution species population variation different change
Topic 10: pollen habitat bees plant pollination change sperm
Topic 11: forest climate change ecosystem model tree data
Topic 12: students program reu biology edu training pi
Topic 13: gene dna chromatin epigene expression methylation development
Topic 14: biology model systems network develop program division
Topic 15: species population plant diversity ecological effects communities
Topic 16: enzyme protein iron amino acid reaction function
Topic 17: university state students researchers system center california
Topic 18: marine ecosystem ocean sea fish animal algae
Topic 19: gene rna protein expression transcription regulatory regulation
Topic 20: cell imaging biology high resolution instrument develop
Topic 21: protein structure molecular interactions binding function folding
Topic 22: students undergraduate science school high graduate biology
Topic 23: dna gene chromosome repair cell protein bacteria
Topic 24: host disease pathogen pathogens immune bacteria plant
Topic 25: species diversity evolution data life tree phylogene
Topic 26: change environmental stress conditions climate responses
Topic 27: plant gene crop arabidopsis growth development production
Topic 28: brain neurons behavior neural animal system sensory
Topic 29: collection data specimens change resource biology available
Topic 30: nitrogen carbon stream nutrient water dioxide fixation

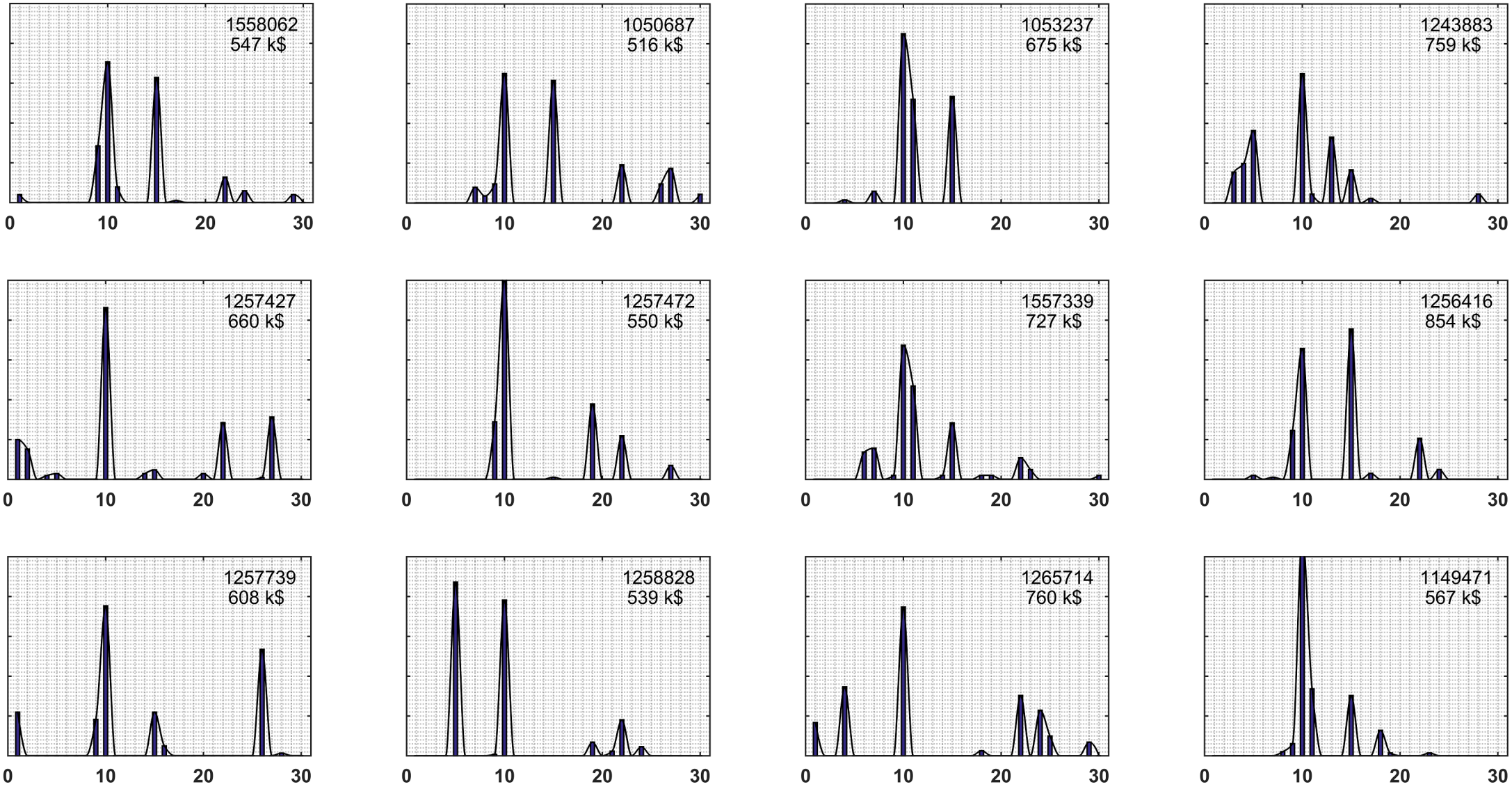
Money generated by each topic over all the grants

Biological Science (BIO) after 2015, above 100 k$



Don’t believe it? Check the profiles

Human verification is an essential part in any probabilistic, non deterministic process.
Topic 10 ranks 1st in 60 documents.



Don’t believe it? Check it in your browser

Human verification is an essential part in any probabilistic, non deterministic process.
Topic 10 ranks 1st in 60 documents.

Award Abstract #1457748

Evolution of dispersal and pollination in ecologically dominant grasses

Awarded Amount to Date: $639,217.00

Award Abstract #1457753

Functional, Genomic, and Evolutionary Analysis of Chemical Courtship Signals in Euglossine Bees

Awarded Amount to Date: $735,887.00

Award Abstract #1257455

Control of lipid metabolism and muscle hypertrophy by PPARs in Gray catbird annual life cycle

Awarded Amount to Date: $592,051.00

Award Abstract #1147165

Functional roles of FERONIA, LORELEI and relative proteins in regulating pollen-pistil interaction

Awarded Amount to Date: $686,000.00

⋮

Award Abstract #1457098

Intercellular Signaling in Pollen Tube Reception

Awarded Amount to Date: $600,000.00

Award Abstract #1344288

INSPIRE Track 1: What is Normal Milk? Sociocultural, Evolutionary, Environmental, and Microbial Aspects of Human Milk Composition

Awarded Amount to Date: $950,000.00

Award Abstract #1256992

Resin to Propolis: Biological origins and role in honey bee social immunity and health

Awarded Amount to Date: $635,976.00

Award Abstract #1452386

The role of post-pollination interactions in structuring co-flowe plant communities

Awarded Amount to Date: $654,641.00

⋮

How do topics overlap?

For each documents we sum the overlapping (minimum) probability between the considered topic and all the other topics. We then sum over all the documents.

$$O_{i,j} = \sum_{j,d} \min(P_i, P_j) * m_d$$

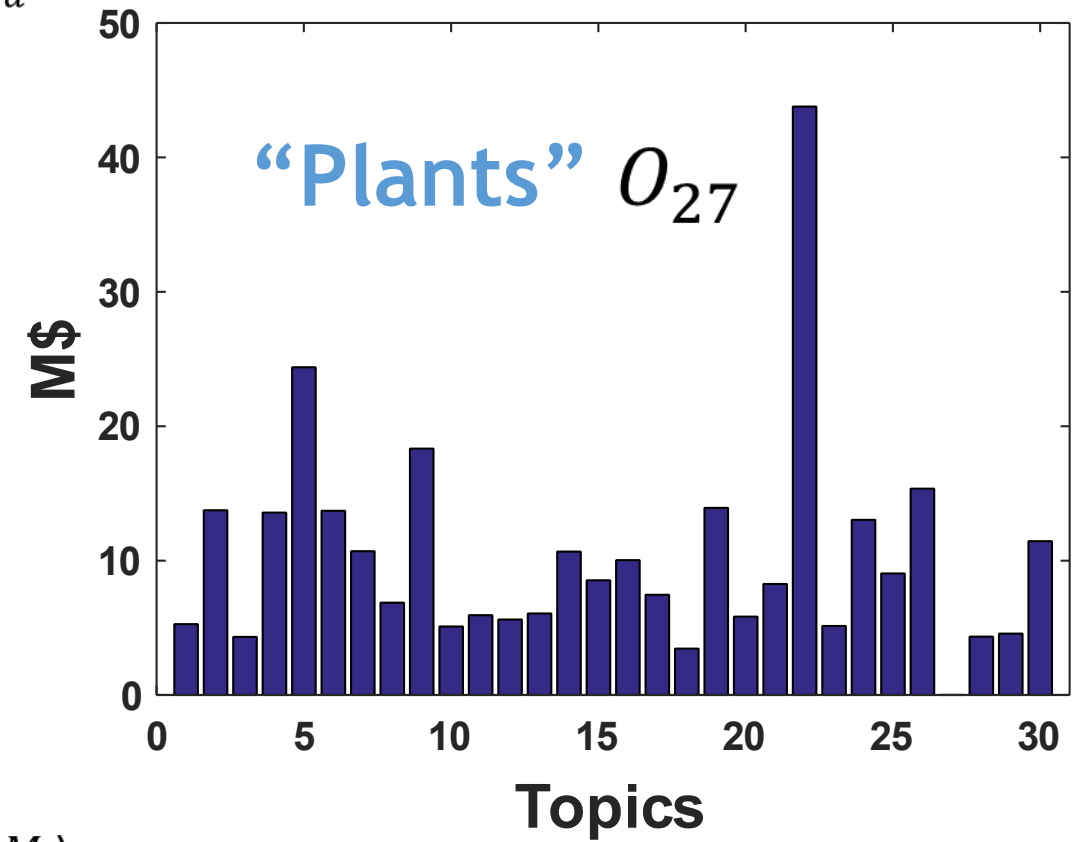
Topic 27: "Plants" has generated **95 M$ overlapped with:**

Topic 22: "Education" for **44 M$**

Topic 5: "Cell" for **24 M$**

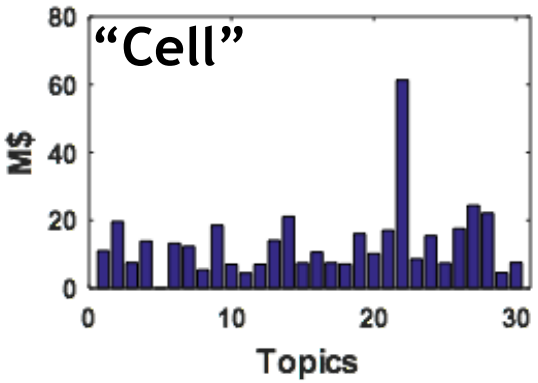
Topic 9: "Genetics" for **18 M$**

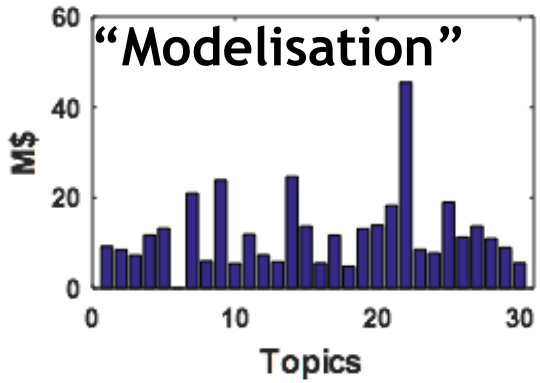
⋮

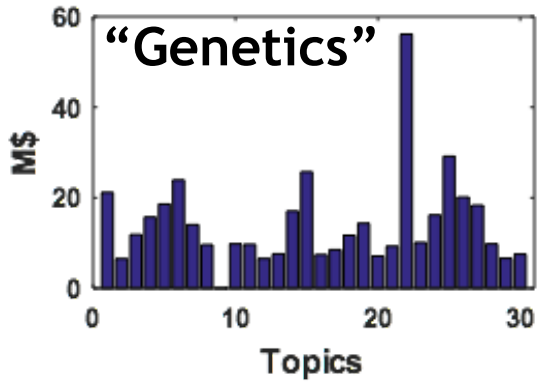


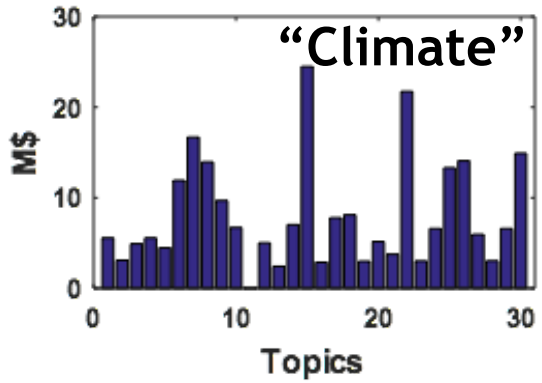
Note: $O_{i,i}$ is the total amount of money generated by the topic i. ($O_{i,i} = M_i$)
We set this quantity to zero for a better display of the overlapping topics.

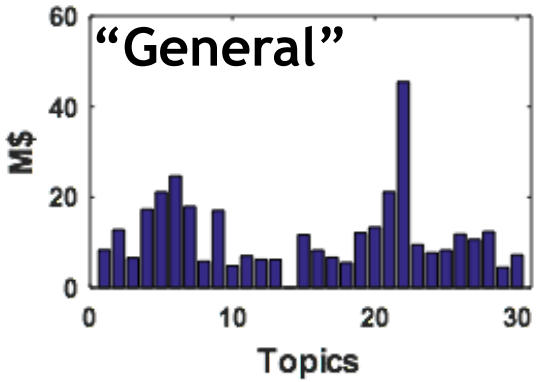
Topics overlap

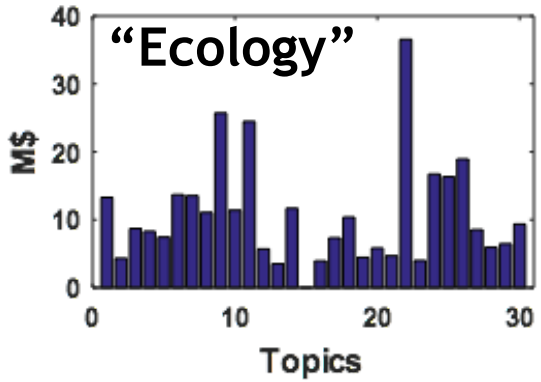


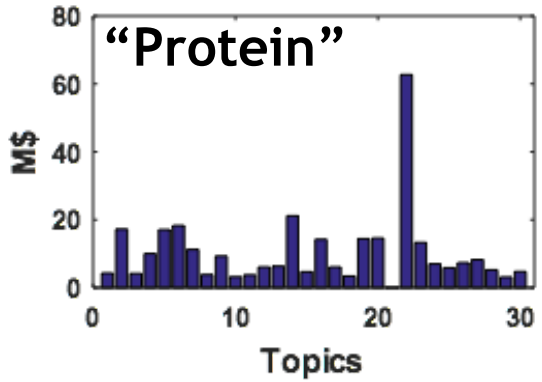


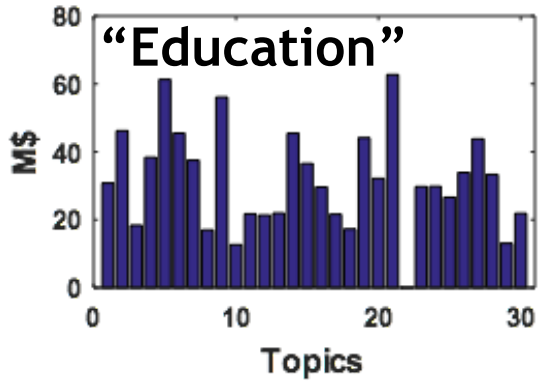


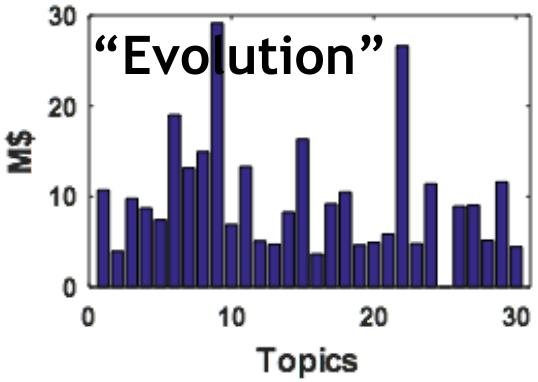


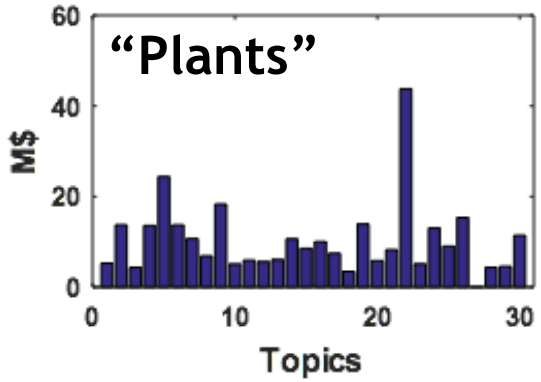












Topics overlap primary with "Education" and then between each other.