Topic Modeling with NIGMS Grants

Office of Program Planning, Analysis, and Evaluation (OPAE)

Alaz Sengul
Nathan Moore, Ph.D.

NIH National Institute of General Medical Sciences

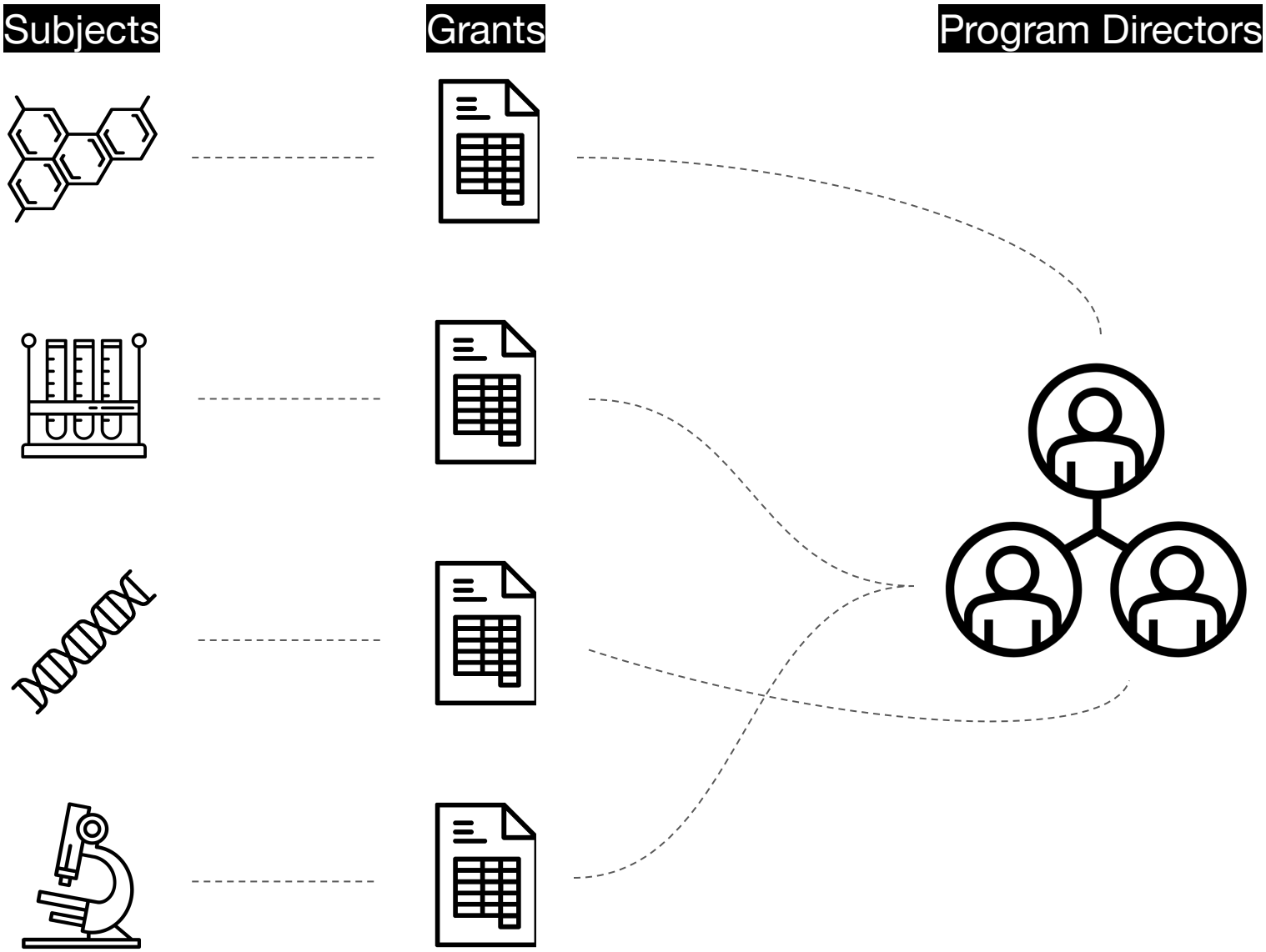
CIVIC DIGITAL FELLOWSHIP

National Institute of General Medical Sciences (NIGMS)

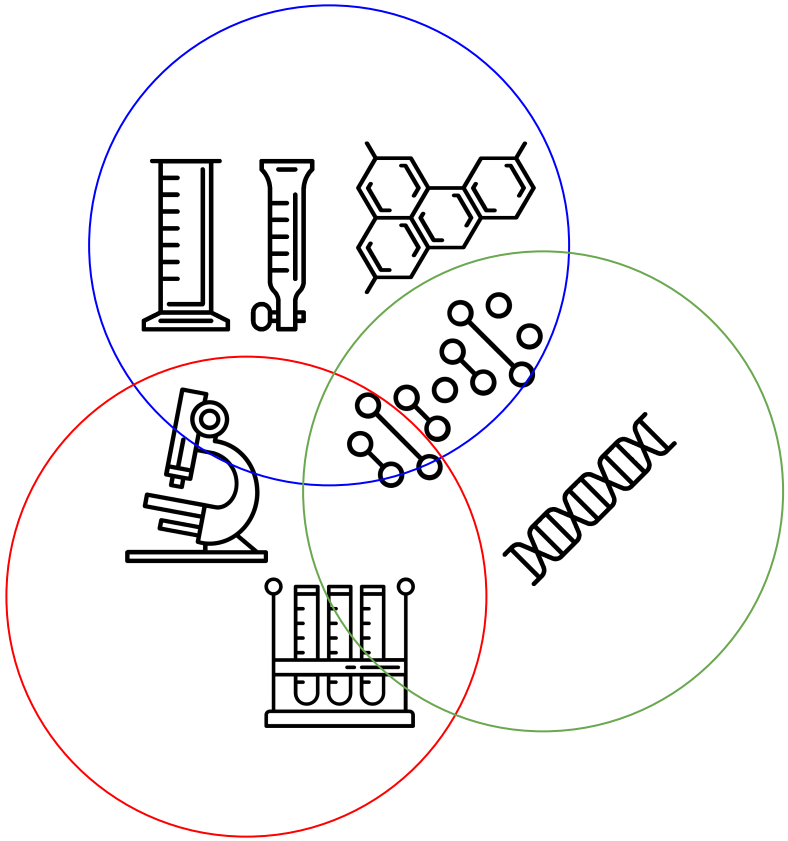
The National Institute of General Medical Sciences supports basic research that increases our understanding of biological processes and lays the foundation for advances in disease diagnosis, treatment, and prevention.

The Office of Program Planning, Analysis, and Evaluation's mission is to ensure that NIGMS has a cohesive, well-directed research program aimed at maximizing the impact of its resources and return on investment for taxpayer dollars.

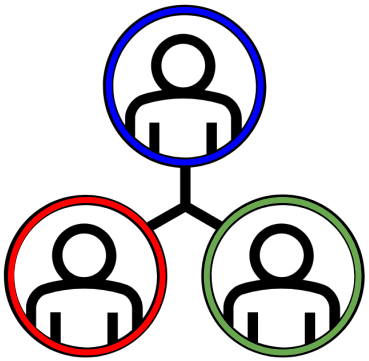
Office of Program Planning, Analysis, and Evaluation (OPAE)


Subjects
Grants
Program Directors

Portfolios



=

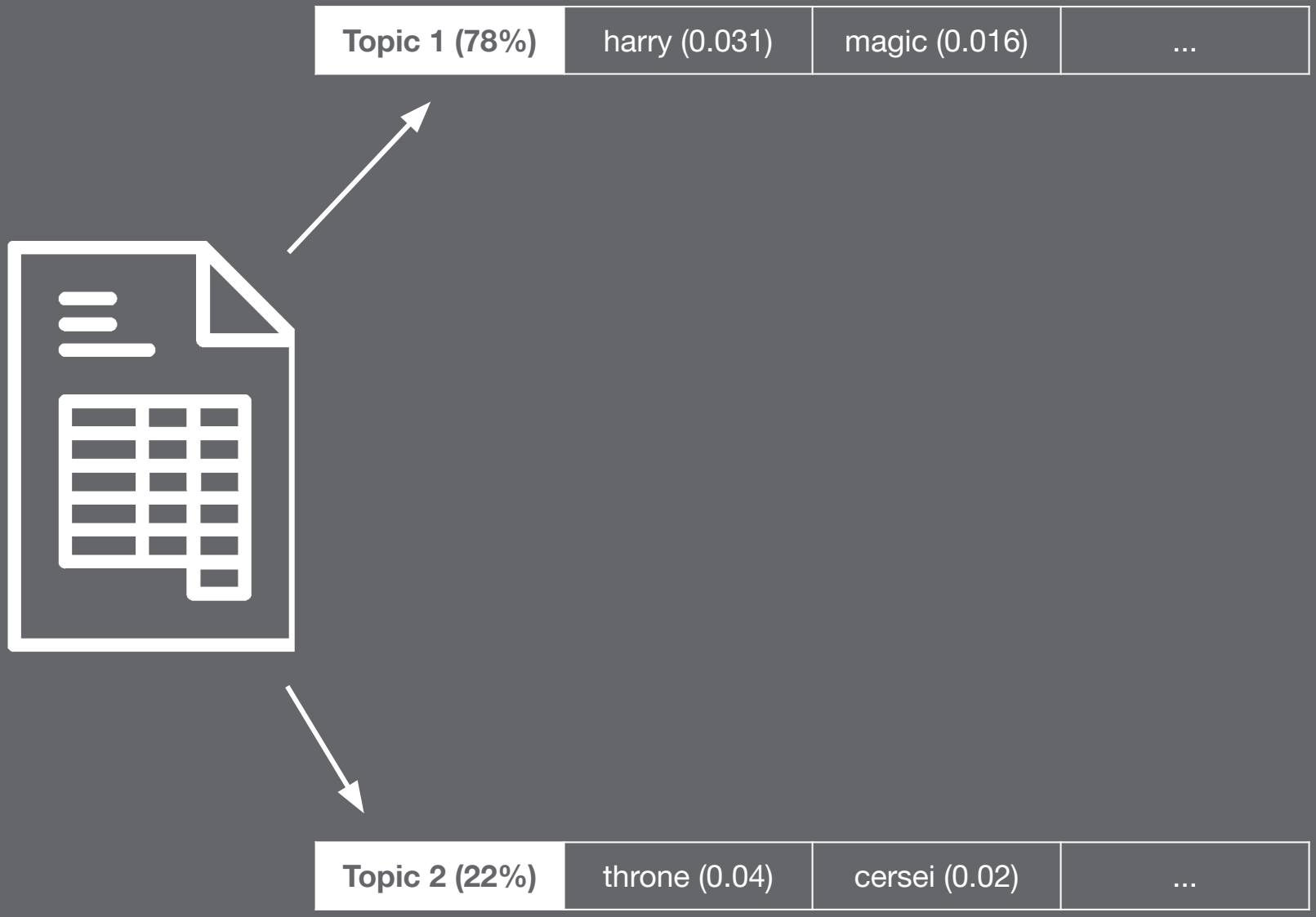


Topic Modeling

Topic 1	Topic 2	...
hogwarts	snow	...
harry	throne	...
wand	dragon	...
hermione	westeros	...
wizard	lannister	...
azkaban	stark	...

Latent Dirichlet Allocation (LDA)

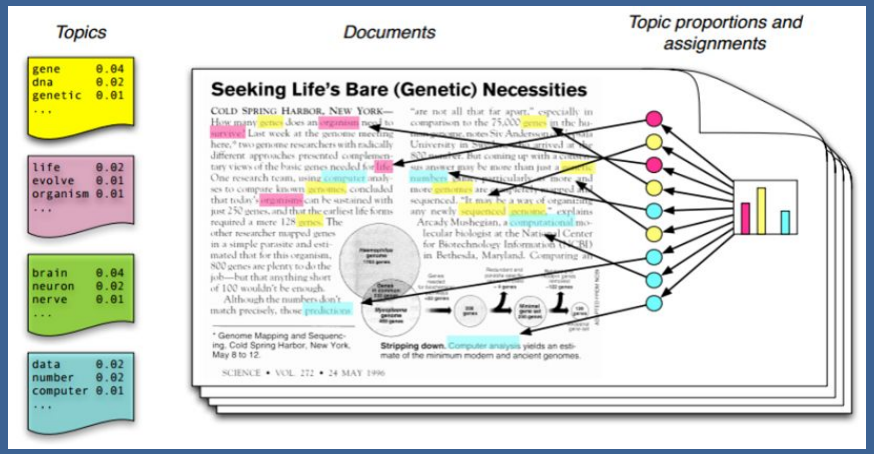
- unsupervised natural language processing
- generative statistical model
- each document is comprised of a probability distribution of topics
- each topic is comprised of a probability distribution of words
- two important steps: preprocessing and modeling

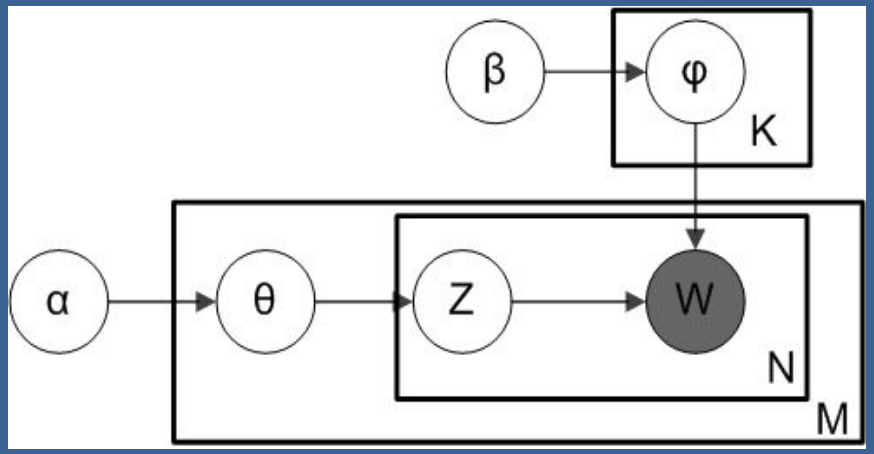

Topic 1 (78%)
harry (0.031)
magic (0.016)
...
Topic 2 (22%)
throne (0.04)
cersei (0.02)
...

Preprocessing

Insects that transmit disease will spread farther and transmit illness more quickly as the planet warms, according to experts at the Centers for Disease Control who monitor vector-borne diseases.	**RAW TEXT**
insects that transmit disease will spread farther and transmit illness more quickly as the planet warms according to experts at the centers for disease control who monitor vectorborne diseases	**LOWERCASE & REMOVE PUNCTUATION**
insects transmit disease spread farther transmit illness quickly planet warms according experts centers_for_disease_control monitor vectorborne diseases	**CHECK FOR n-GRAMS & REMOVE STOPWORDS**
insect transmit disease spread farther transmit illness quickly planet warms accord expert centers_for_disease_control monitor vectorborne disease	**LEMMATIZATION**

Modeling





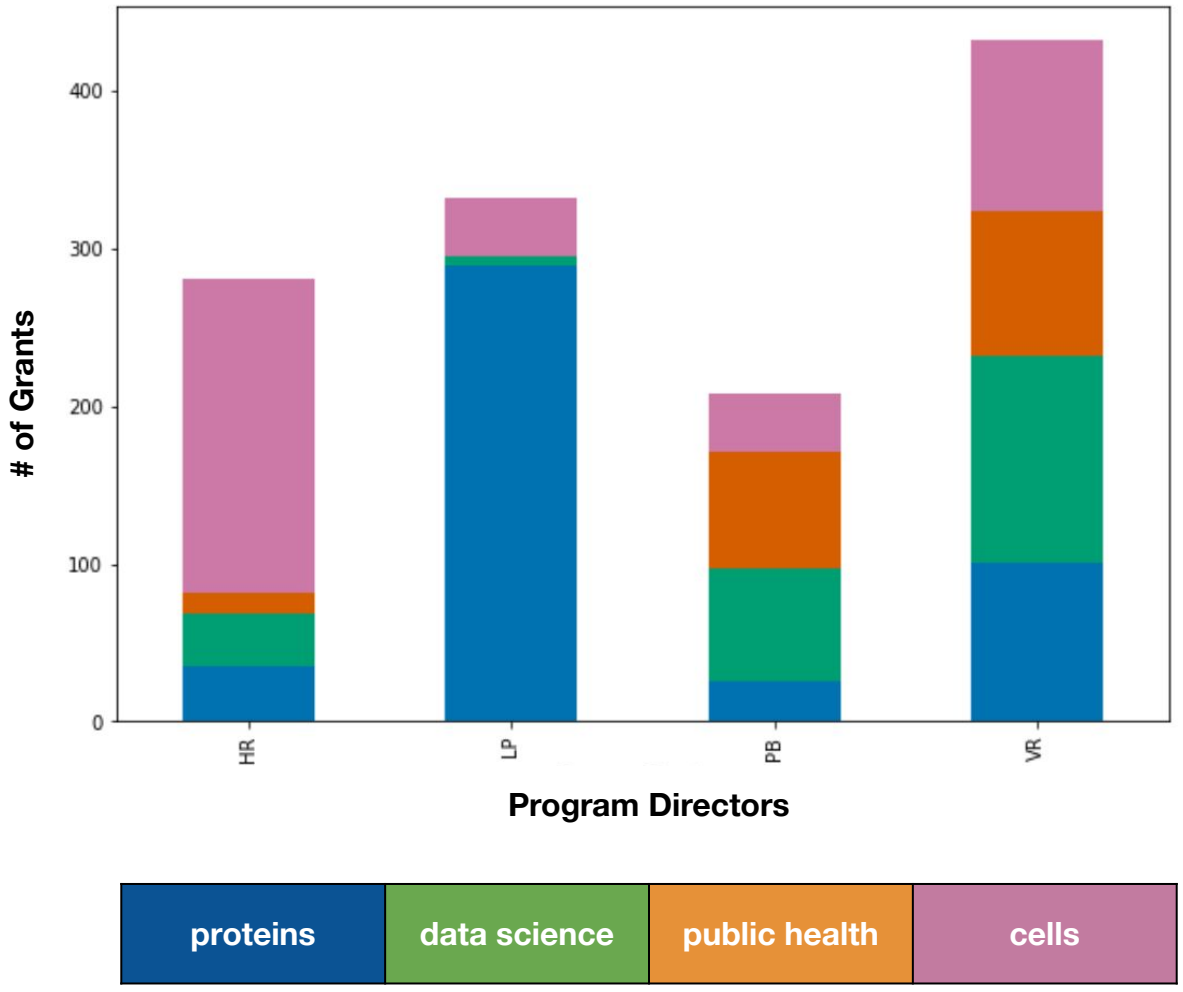
- number of topics (N)
- alpha (hyperparameter)
- beta (hyperparameter)
- perplexity (cohesiveness)

N = 4
alpha = 0.05
beta = 0.05

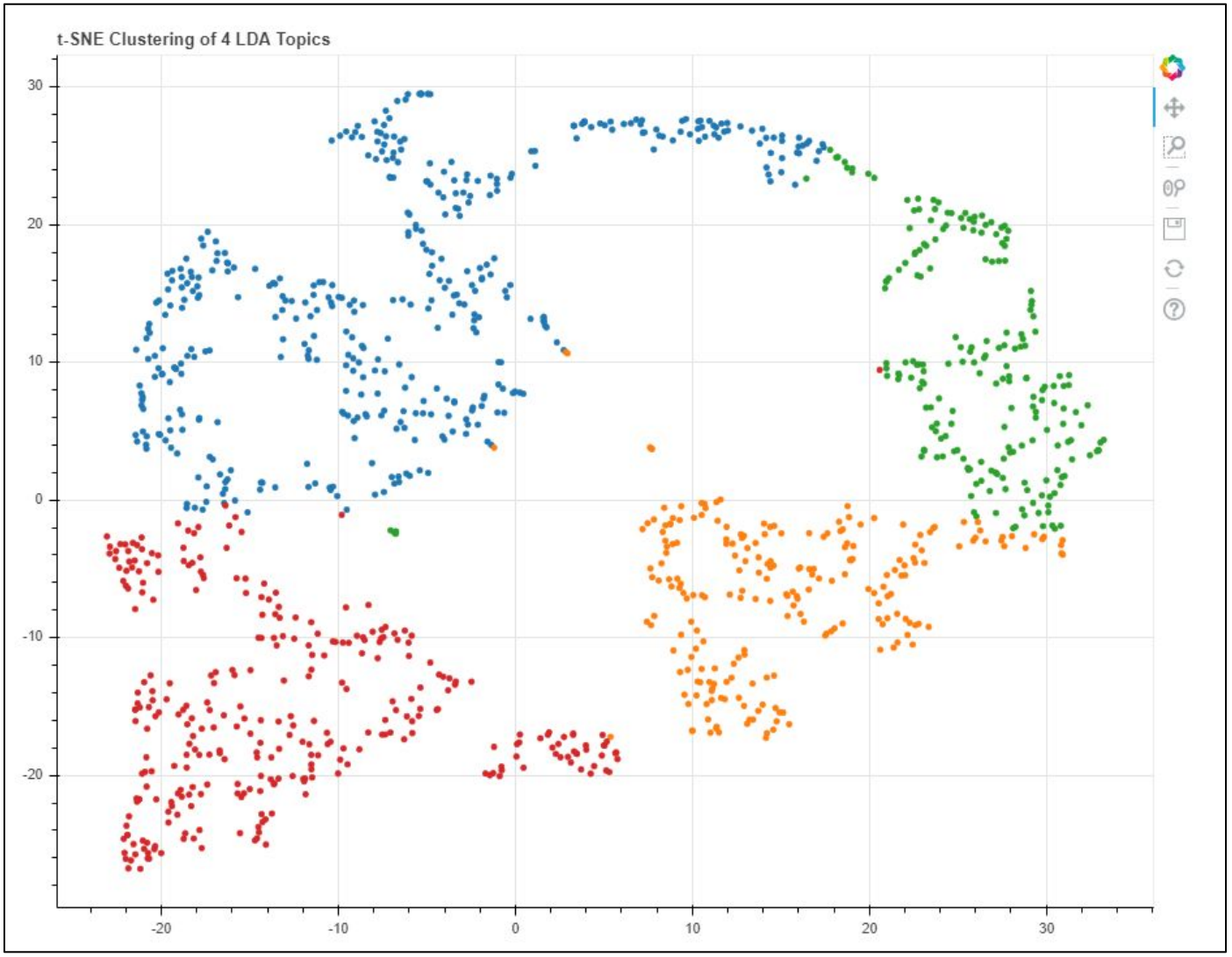
Bioinformatics Grants – Results

Topic 1	Topic 2	Topic 3	Topic 4
protein	datum	health	cell
structure	tool	disease	network
sequence	analysis	system	disease
bind	software	project	study
drug	technology	patient	gene
design	project	framework	mechanism
interaction	propose	population	dynamic
simulation	study	outcome	system
proteins?	**data science?**	**public health?**	**cells?**

Portfolio Distribution



Topic Spread Visualization



Bioinformatics Grants

Topic 1	Topic 2	Topic 3	Topic 4
protein	datum	health	cell
structure	tool	**disease**	network
sequence	analysis	system	**disease**
bind	software	project	study
drug	technology	patient	gene
design	project	framework	mechanism
interaction	propose	population	dynamic
simulation	study	outcome	system

?

Bioinformatics Grants – Word Vectors

Topic 3
disease
[('forecast'), ('incidence'), ('intervention'), ('understand'), ('measle'), ('response'), ('factor'), ('epidemic'), ('outbreak'), ('transmission'), ('vaccination'), ('spread'), ('forecasting'), ('level'), ('determine'), ('strategy'), ('pattern'), ('burden'), ('associate'), ('risk')]

Topic 4
disease
[('treatment'), ('cancer'), ('understand'), ('target'), ('neurodegenerative_disease'), ('drug'), ('understanding'), ('health',), ('lead'), ('understudy'), ('prevention'), ('therapy'), ('insight'), ('pathogenesis'), ('impact'), ('cause'), ('mechanism'), ('alzheimer'), ('heart'), ('diabetes')]

Questions?

Thank you to everyone at OPAE for such a warm welcome and an amazing summer!

Also, thank you of course to Jess and the Coding it Forward team for making it all happen!