BRAINWORKS

Mohammad M. Ghassemi, Ph.D.

Assistant Professor, Computer Science, Michigan State University
Data and Technology Advancement National Service Scholar, NIH

Career straddles academia, industry, and entrepreneurship

Interested in both the theoretical development, and practical deployment, of technology

Mohammad Ghassemi, Ph.D.

Is a assistant professor of computer science at MSU. He holds graduate degrees from MIT, and Cambridge (UK). He was a director of data science at S&P Global, and a strategic consultant with BCG. He has over ten years of technical and strategic consulting experience for many of the world's largest brands.

Human Augmentation and AI Lab

We develop tools and systems that combine human and machine intelligence (A.I.) to solve problems that neither humans nor machines can solve as effectively alone. More specifically, we develop new theoretical knowledge and practical tools for *Augmented Intelligence (A-I)*: the enhancement of individual or collective cognitive function through the use of technology and social/environmental factors.

Data scholar project supports mission of the BRAIN Initiative

BRAIN supports technologies that promote a dynamic understanding of the brain

Deliver a functional proof-of-concept of the "*BRAIN initiative Workspace to ORganize the Knowledge Space*" platform ("BRAIN WORKS" hereafter) – a tool for the discovery of more comprehensive theories of brain function through knowledge integration.

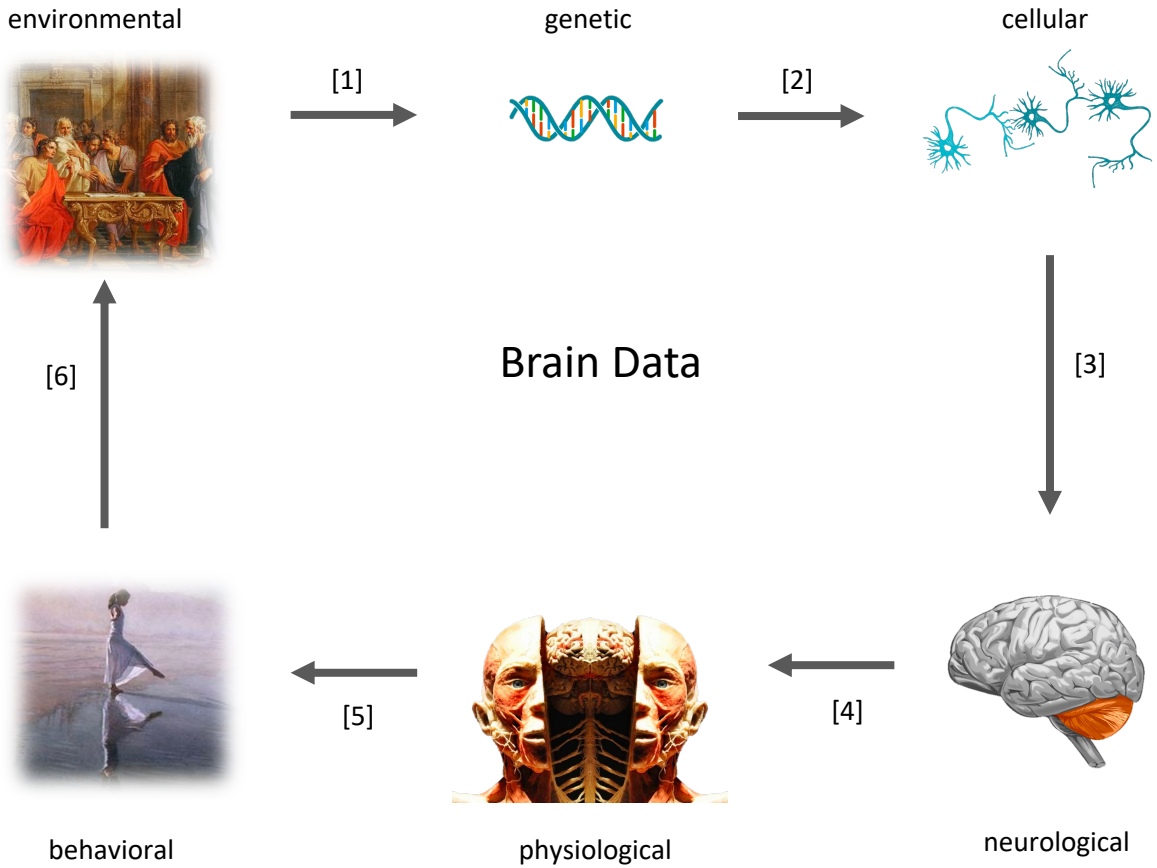
BRAINWORKS will support discovery of new brain theories

Discovery through integration of multi-modal, multi-scale brain data and knowledge

Brain Data is Heterogeneous: exists in multiple modalities, scales and levels of resolution. The heterogeneity of brain data requires thoughtful approaches to data storage, analysis and representation.

Discovery Requires Integration: next generation brain theories require a holistic approach to the brain where the plurality of contexts that impact brain structure and function are accounted for.

Data Science is Pivotal: by automating tasks, augmenting investigator capabilities we may assist with the discovery of holistic theories of brain function, and definition of new research horizons.



[1-6]: See References Slide

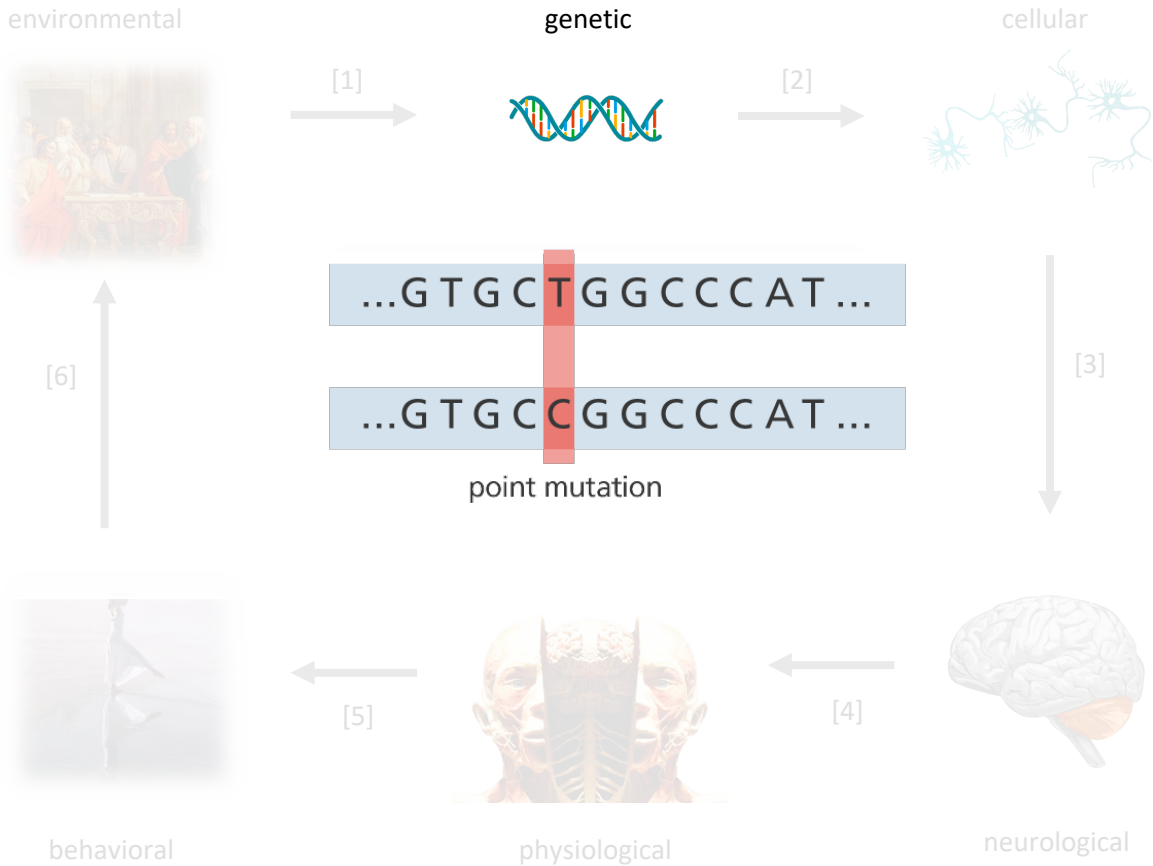
BRAINWORKS will support discovery of new brain theories

Discovery through integration of multi-modal, multi-scale brain data and knowledge

Brain Data is Heterogeneous: exists in multiple modalities, scales and levels of resolution. The heterogeneity of brain data requires thoughtful approaches to data storage, analysis and representation.

Discovery Requires Integration: next generation brain theories require a holistic approach to the brain where the plurality of contexts that impact brain structure and function are accounted for.

Data Science is Pivotal: by automating tasks, augmenting investigator capabilities we may assist with the discovery of holistic theories of brain function, and definition of new research horizons.



[1-6]: See References Slide

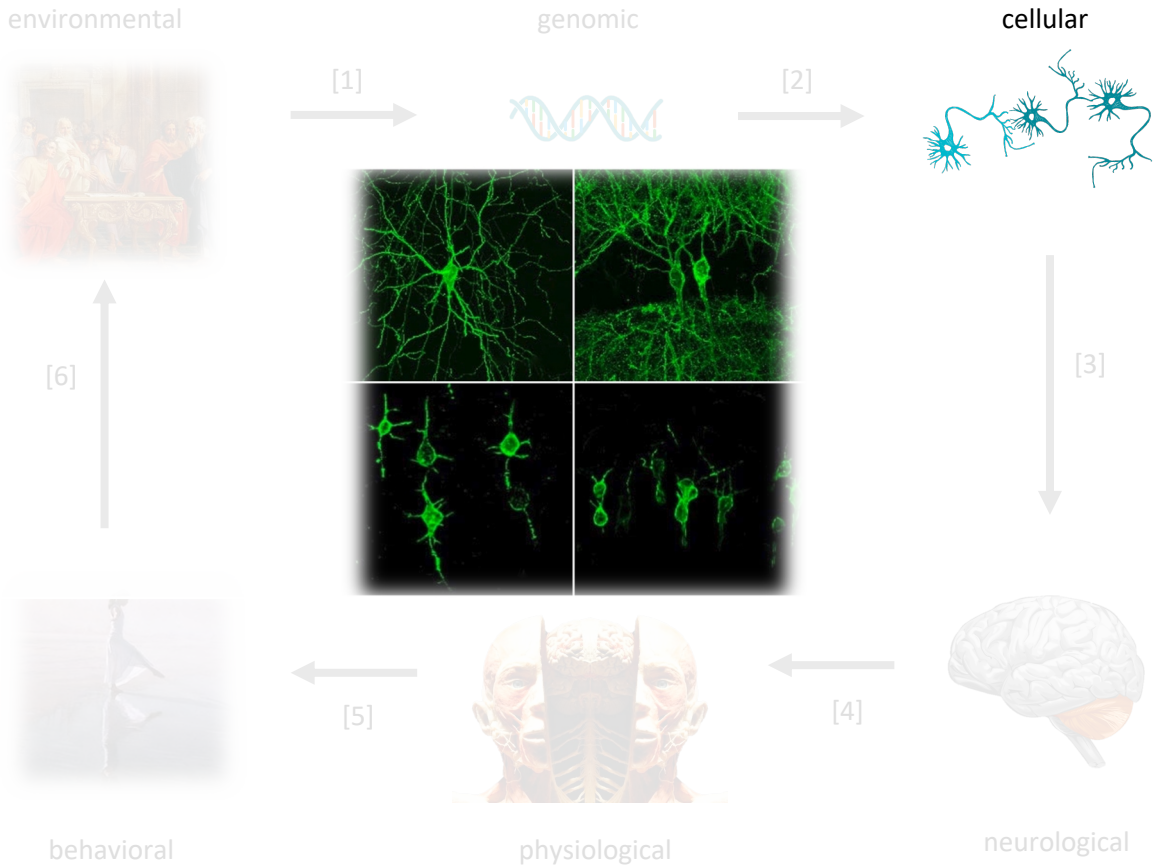
BRAINWORKS will support discovery of new brain theories

Discovery through integration of multi-modal, multi-scale brain data and knowledge

Brain Data is Heterogeneous: exists in multiple modalities, scales and levels of resolution. The heterogeneity of brain data requires thoughtful approaches to data storage, analysis and representation.

Discovery Requires Integration: next generation brain theories require a holistic approach to the brain where the plurality of contexts that impact brain structure and function are accounted for.

Data Science is Pivotal: by automating tasks, augmenting investigator capabilities we may assist with the discovery of holistic theories of brain function, and definition of new research horizons.



[1-6]: See References Slide

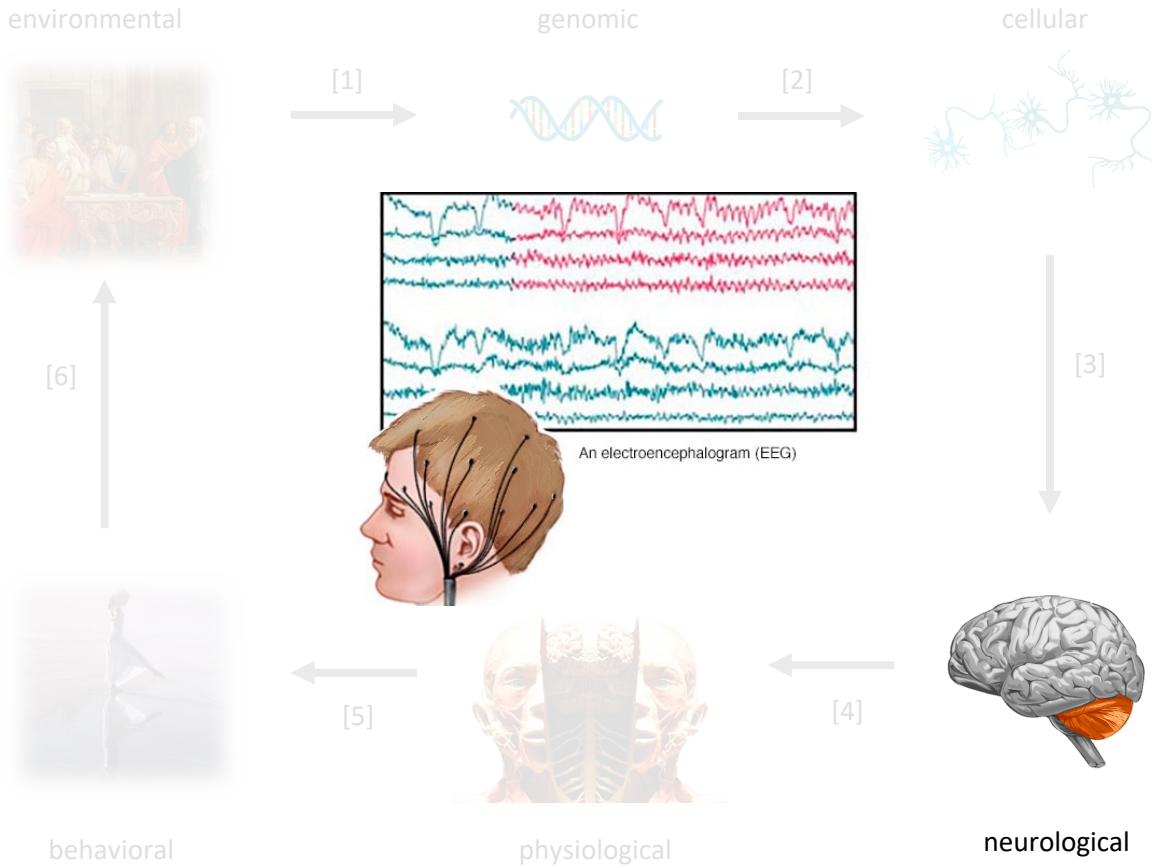
BRAINWORKS will support discovery of new brain theories

Discovery through integration of multi-modal, multi-scale brain data and knowledge

Brain Data is Heterogeneous: exists in multiple modalities, scales and levels of resolution. The heterogeneity of brain data requires thoughtful approaches to data storage, analysis and representation.

Discovery Requires Integration: next generation brain theories require a holistic approach to the brain where the plurality of contexts that impact brain structure and function are accounted for.

Data Science is Pivotal: by automating tasks, augmenting investigator capabilities we may assist with the discovery of holistic theories of brain function, and definition of new research horizons.



[1-6]: See References Slide

BRAINWORKS will support discovery of new brain theories

Discovery through integration of multi-modal, multi-scale brain data and knowledge

Brain Data is Heterogeneous: exists in multiple modalities, scales and levels of resolution. The heterogeneity of brain data requires thoughtful approaches to data storage, analysis and representation.

Discovery Requires Integration: next generation brain theories require a holistic approach to the brain where the plurality of contexts that impact brain structure and function are accounted for.

Data Science is Pivotal: by automating tasks, augmenting investigator capabilities we may assist with the discovery of holistic theories of brain function, and definition of new research horizons.

[1-6]: See References Slide

BRAINWORKS will support discovery of new brain theories

Discovery through integration of multi-modal, multi-scale brain data and knowledge

Brain Data is Heterogeneous: exists in multiple modalities, scales and levels of resolution. The heterogeneity of brain data requires thoughtful approaches to data storage, analysis and representation.

Discovery Requires Integration: next generation brain theories require a holistic approach to the brain where the plurality of contexts that impact brain structure and function are accounted for.

Data Science is Pivotal: by automating tasks, augmenting investigator capabilities we may assist with the discovery of holistic theories of brain function, and definition of new research horizons.

[1-6]: See References Slide

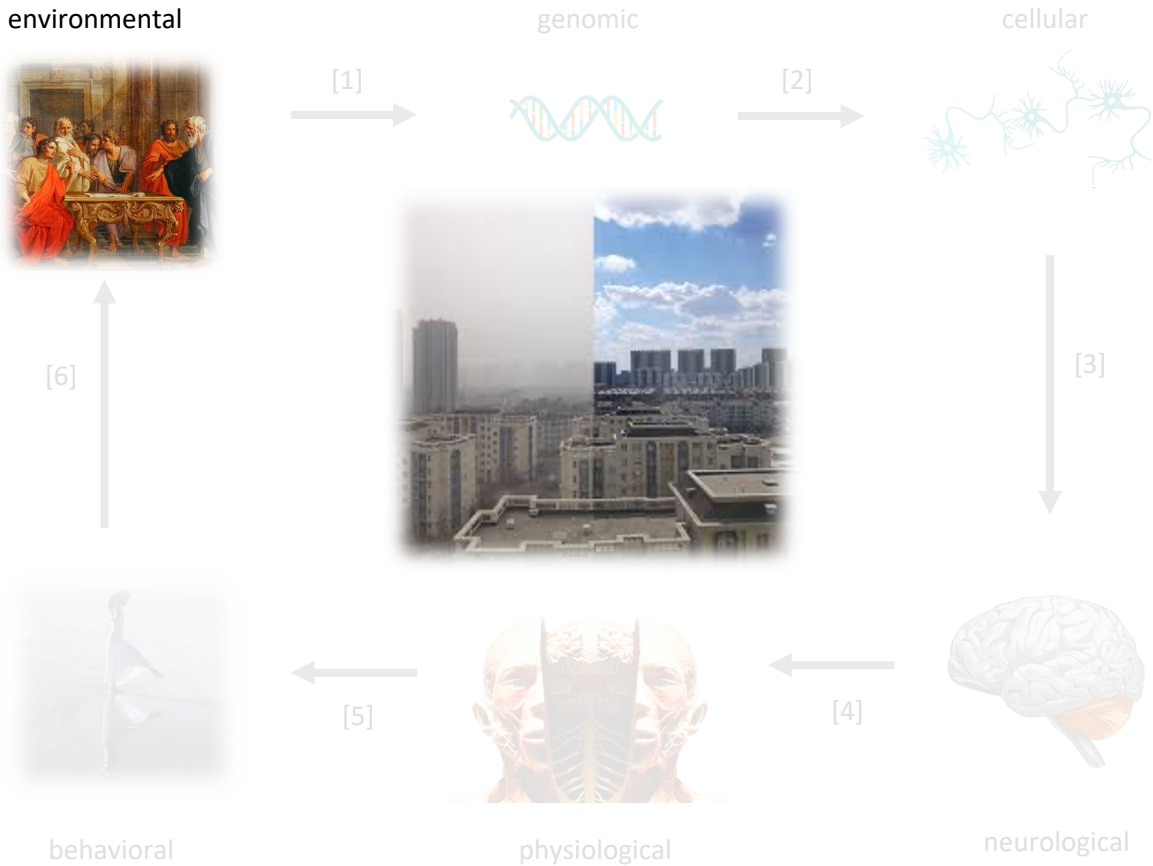
BRAINWORKS will support discovery of new brain theories

Discovery through integration of multi-modal, multi-scale brain data and knowledge

Brain Data is Heterogeneous: exists in multiple modalities, scales and levels of resolution. The heterogeneity of brain data requires thoughtful approaches to data storage, analysis and representation.

Discovery Requires Integration: next generation brain theories require a holistic approach to the brain where the plurality of contexts that impact brain structure and function are accounted for.

Data Science is Pivotal: by automating tasks, augmenting investigator capabilities we may assist with the discovery of holistic theories of brain function, and definition of new research horizons.



[1-6]: See References Slide

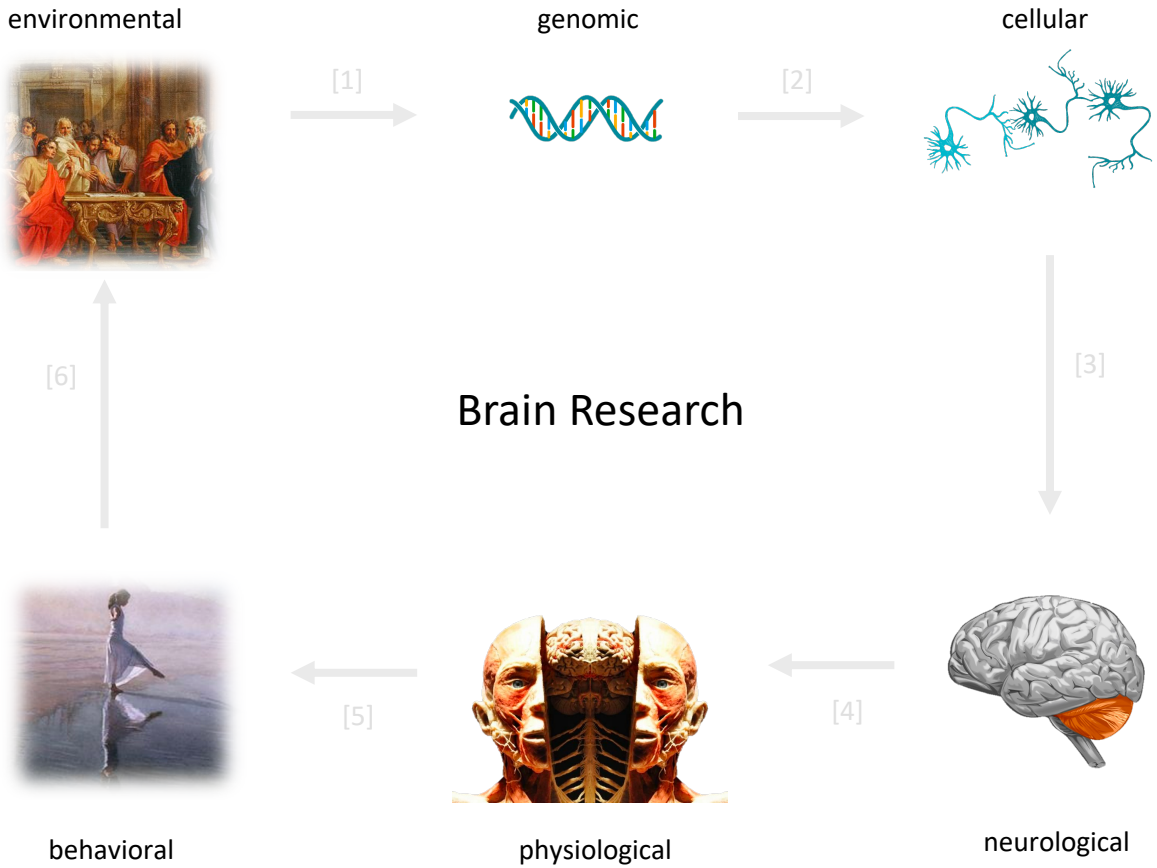
BRAINWORKS will support discovery of new brain theories

Discovery through integration of multi-modal, multi-scale brain data and knowledge

Brain Data is Heterogeneous: exists in multiple modalities, scales and levels of resolution. The heterogeneity of brain data requires thoughtful approaches to data storage, analysis and representation.

Discovery Requires Integration: next generation brain theories require a holistic approach to the brain where the plurality of contexts that impact brain structure and function are accounted for.

Data Science is Pivotal: by automating tasks, augmenting investigator capabilities we may assist with the discovery of holistic theories of brain function, and definition of new research horizons.



[1-6]: See References Slide

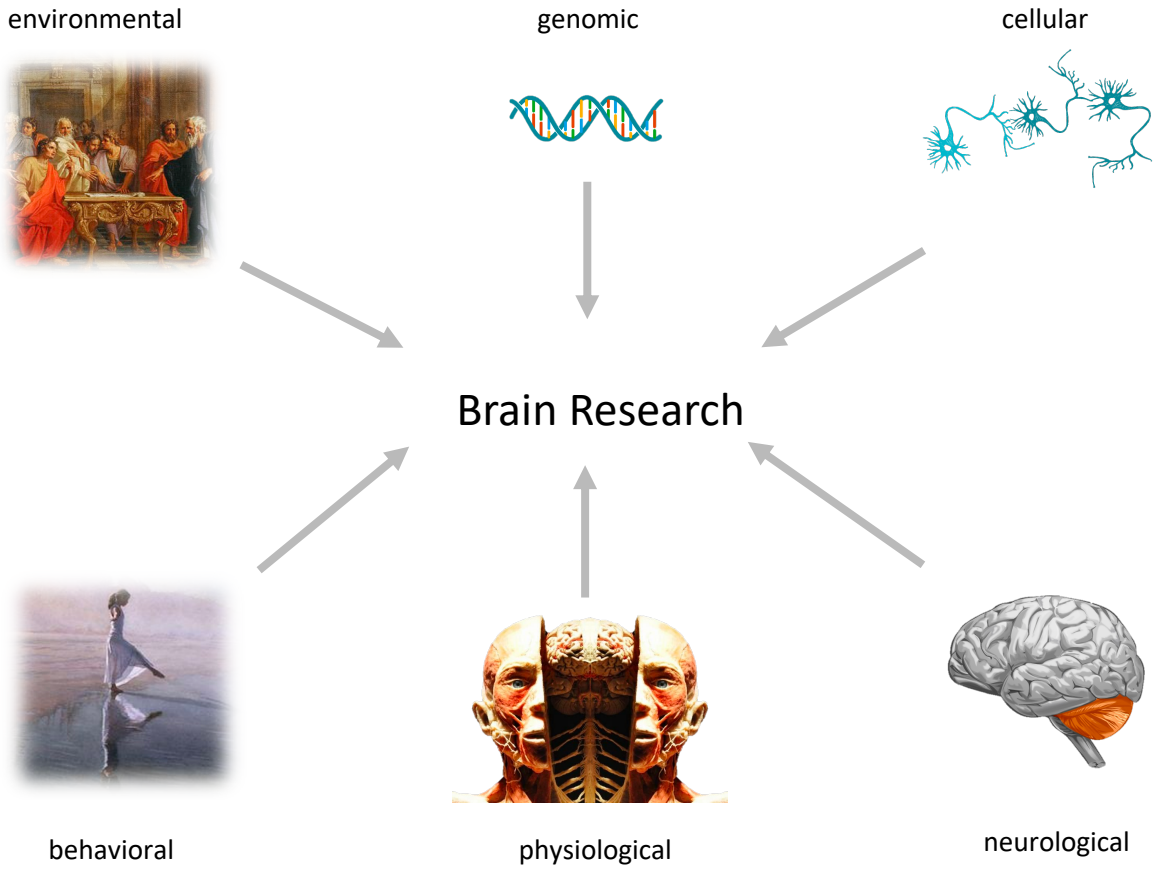
BRAINWORKS will support discovery of new brain theories

Discovery through integration of multi-modal, multi-scale brain data and knowledge

Brain Data is Heterogeneous: exists in multiple modalities, scales and levels of resolution. The heterogeneity of brain data requires thoughtful approaches to data storage, analysis and representation.

Discovery Requires Integration: next generation brain theories require a holistic approach to the brain where the plurality of contexts that impact brain structure and function are accounted for.

Data Science is Pivotal: by automating tasks, augmenting investigator capabilities we may assist with the discovery of holistic theories of brain function, and definition of new research horizons.



[1-6]: See References Slide

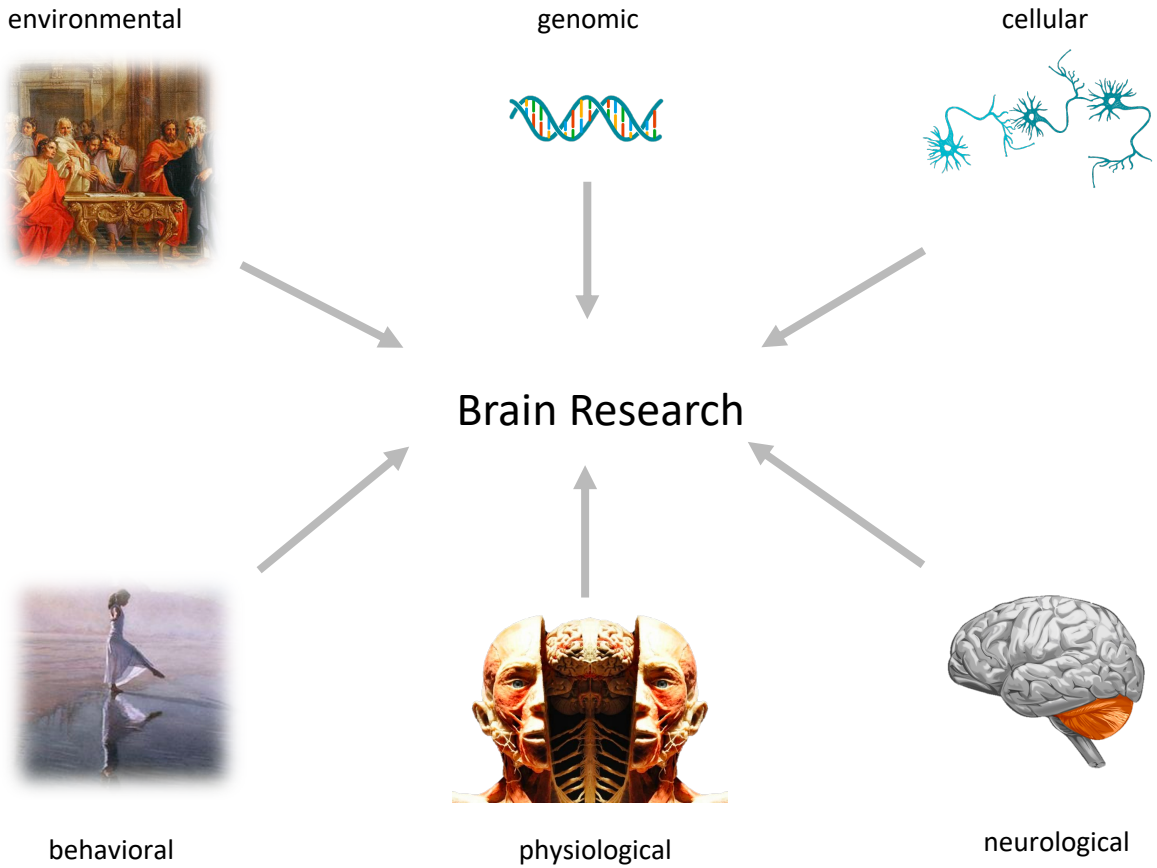
BRAINWORKS will support discovery of new brain theories

Discovery through integration of multi-modal, multi-scale brain data and knowledge

Brain Data is Heterogeneous: exists in multiple modalities, scales and levels of resolution. The heterogeneity of brain data requires thoughtful approaches to data storage, analysis and representation.

Discovery Requires Integration: next generation brain theories require a holistic approach to the brain where the plurality of contexts that impact brain structure and function are accounted for.

Data Science is Pivotal: by automating tasks and augmenting investigator capabilities we may assist with the discovery of holistic theories of brain function and define new research horizons.



BRAIN WORKS uses a *value-centered* approach to data science

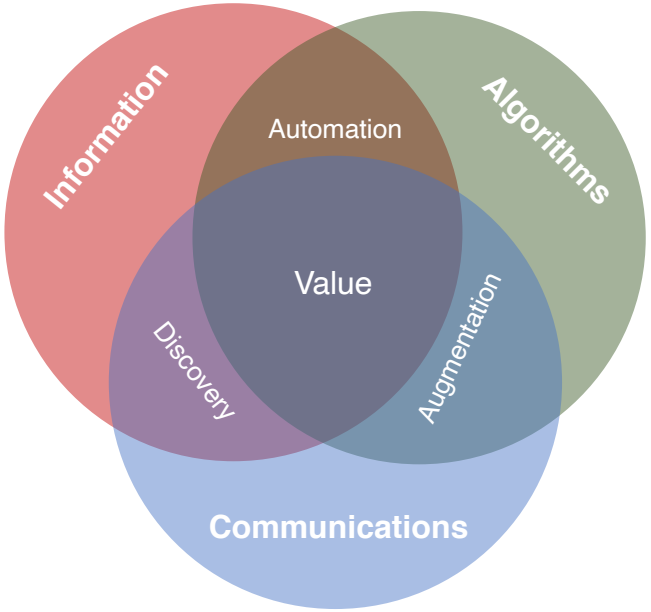
Value lies at the confluence of specific *outcomes* and robust *foundations*

Data Science Foundations

Information: is created when *data* is processed into clean usable form.

Algorithms: are tools that *predict* one kind of information using another.

Communications: are methods for *representing* complex information in simpler forms.



Data Science Outcomes

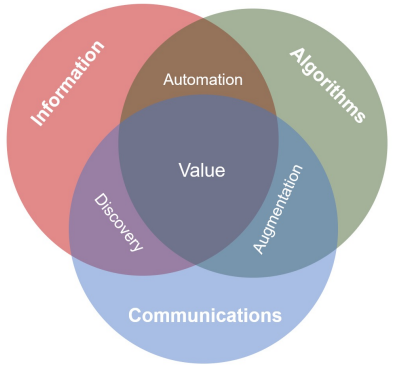
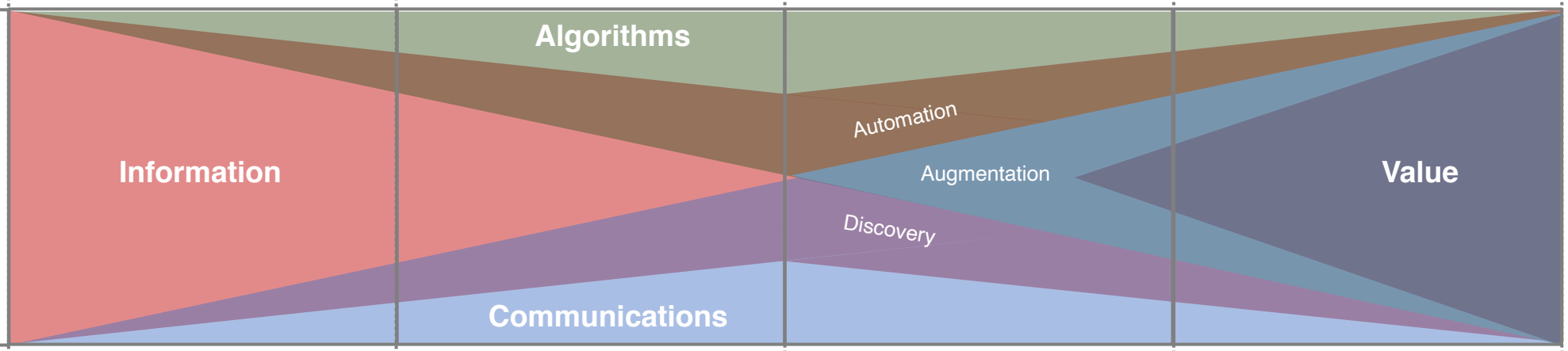
Automation: is the use of *algorithms* to convert low-value *information* into high-value information.

Augmentation: is the combination of human and machine capabilities to create more value than either human or machines could alone.

Discovery: is the process of *communicating* information so humans can extract value on their own.

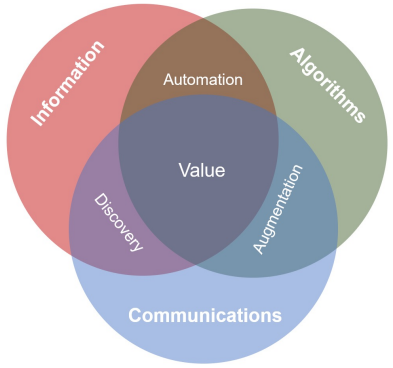
Project developed in five phases over a twelve-month period

Initial efforts focus on data preparation transitioning to outcome generation and value demonstration.

1. Specification *January, 2021*	2. Data to Information *February – May, 2021*	3. Algorithms and Comms. *June – August, 2021*	4. Augmentation *September – November, 2021*	5. Value *December, 2021*
				
1.1 Define value 1.2 Define outcomes 1.3 Define Foundations	2.1 Data characterization 2.2 Data centralization 2.3 Data cleansing	3.1 Automation outcomes 3.2 Communication outcomes 3.3 Performance Evaluation	4.1 Augmentation outcomes 4.2 Integration of novel data 4.3 Performance Evaluation	5.1 Demonstration 5.2 Documentation 5.3 Hand-off

To start, we spoke with several internal and external experts

Definition of the value, data science outcomes, and required foundations

1. Specification *January, 2021*	2. Data to Information *February – May, 2021*	3. Algorithms and Comms. *June – August, 2021*	4. Augmentation *September – October, 2021*	5. Value *December, 2021*
				
1.1 Define value 1.2 Define outcomes 1.3 Define Foundations	2.1 Data characterization 2.2 Data collection 2.3 Data centralization	3.1 Automation outcomes 3.2 Communication outcomes 3.3 Performance Evaluation	4.1 Augmentation outcomes 4.2 Integration of novel data 4.3 Performance Evaluation	5.1 Demonstration 5.2 Documentation 5.3 Hand-off

Requirements captured by synthesizing needs of NIH & scientific community.

Acknowledgements

Grace Peng
Program Director
NIBIB

Jim Gnadt
Program Director
NINDS

Karen David
Program Director
NINDS

Christina Fang
Program Manager
NINDS

Michele Ferrante
Program Director
NIMH

Amy Poremba
Program Director
NIDCD

Susan Wright
Program Director
NIDA

Tatiana Pasternik
Scientific Review Officer
NINDS

Elizabeth Buffalo
Professor
Washington

Ilya Nemenman
Professor
Emory

Edmund Talley
Program Director
NINDS

Olivier Burton
Program Director
NIDA

Susan Volman
Program Director
NIDA

Martha Flanders
Program Director
NEI

Merav Sabri
Program Director
NCCIH

Holly Moore
Program Officer
NIDARossi, Andrew

John Ngai
BRAIN Director
NINDS

Elizabeth Powell
Program Director
NIAAA

David Kleinfeld
Professor
UCSD

Mark Reimers
Professor
MSU

William Lytton
Professor
SUNY Downstate

Fidel Santamaria
Professor
UTSA

Carina Curto
Professor
Penn State

Brent Dorion
Professor
Pittsburgh

Larry Abbot
Professor
Columbia

Marc Howard
Professor
BU

Greg DeAngelis
Professor
Rochester

Ivan Soltesz
Professor
Stanford

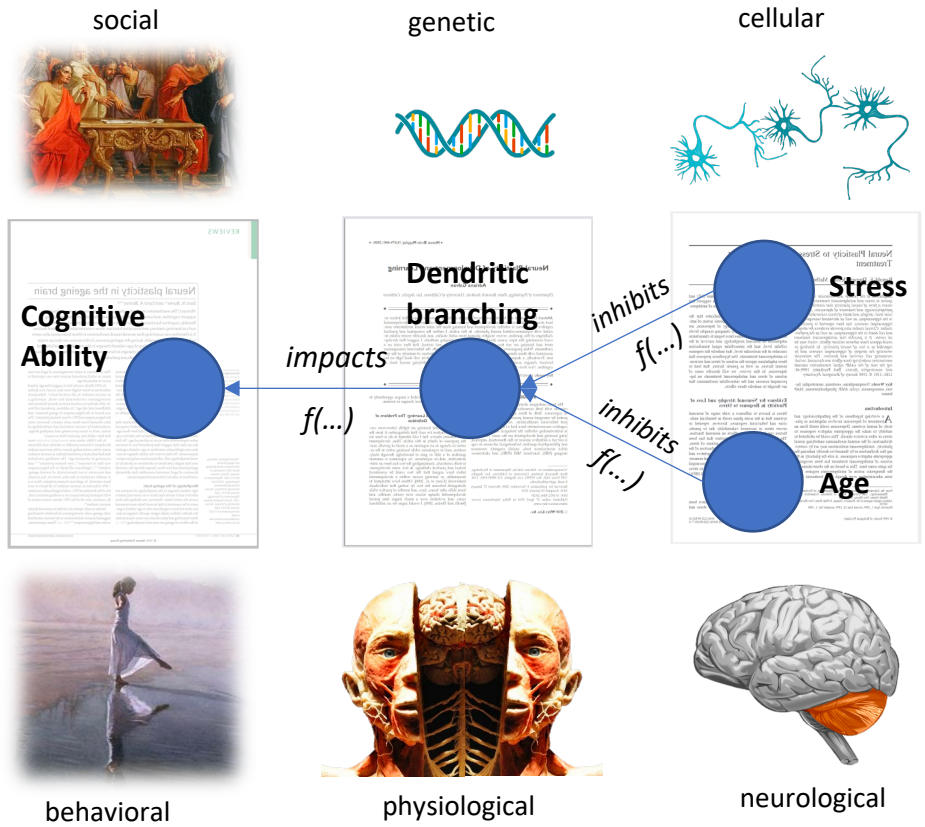
Jan Drugowitsch
Professor
Harvard

Florian Engert
Professor
Harvard

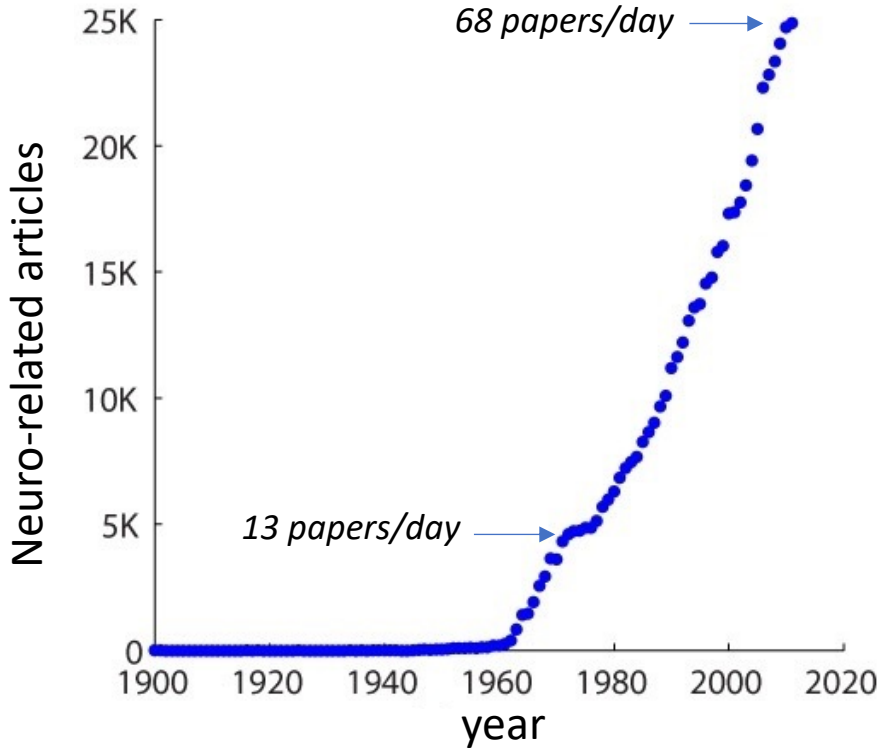
BRAINWORKS converts scientific papers into knowledge graphs

Allows for exploration of scientific literature, and integration of findings across papers

BRAINWORKS: is an web application being developed in 2021 that uses AI to organize the neuroscience literature as an intuitive and interactive knowledge graph.

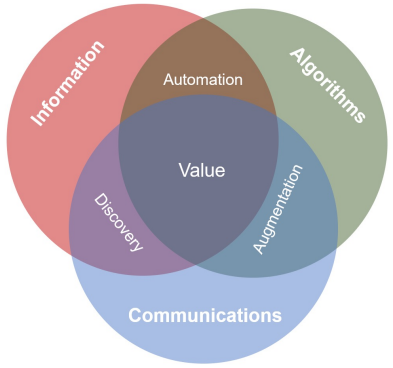
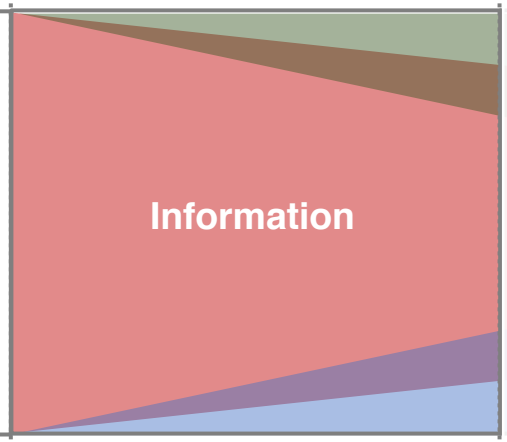


Motivation: Neuroscience theories and knowledge are distributed across a complex, rapidly evolving scientific landscape that no one person can fully master.



Our next step was to collect data to build the knowledge graph

We focused on data cleansing, characterization, and centralization

1. Specification	2. Data to Information	3. Algorithms and Comms.	4. Augmentation	5. Value
January, 2021	*February – May, 2021*	*June – August, 2021*	*September – October, 2021*	*December, 2021*
				
1.1 Define value	2.1 Data characterization	3.1 Automation outcomes	4.1 Augmentation outcomes	5.1 Demonstration
1.2 Define outcomes	2.2 Data collect	3.2 Communication outcomes	4.2 Integration of novel data	5.2 Documentation
1.3 Define Foundations	2.3 Data centralization	3.3 Performance Evaluation	4.3 Performance Evaluation	5.3 Hand-off

Collect data to enable the platform realization.

Data was collected from three publicly available NIH resources

RePORTER, PubMed, and MeSH served as the primary data sources

Grants

- The Research Portfolio Online Reporting Tools (RePORT) provides access to reports, data, and analyses of NIH research activities.
- RePORTER provides access to NIH-funded research projects and their affiliated publications and patents.

Papers

- PubMed provides 32 million citations from life science journals , books, and other sources of biomedical literature.
- Citations include abstracts and links to full text content from PubMed Central and publisher web sites.

Topics

- MeSH (Medical Subject Headings) provides a hierarchically organized thesaurus of continually reviewed and updated medical concepts
- Each PubMed paper represented with a set of MeSH terms that describe its subject topic.

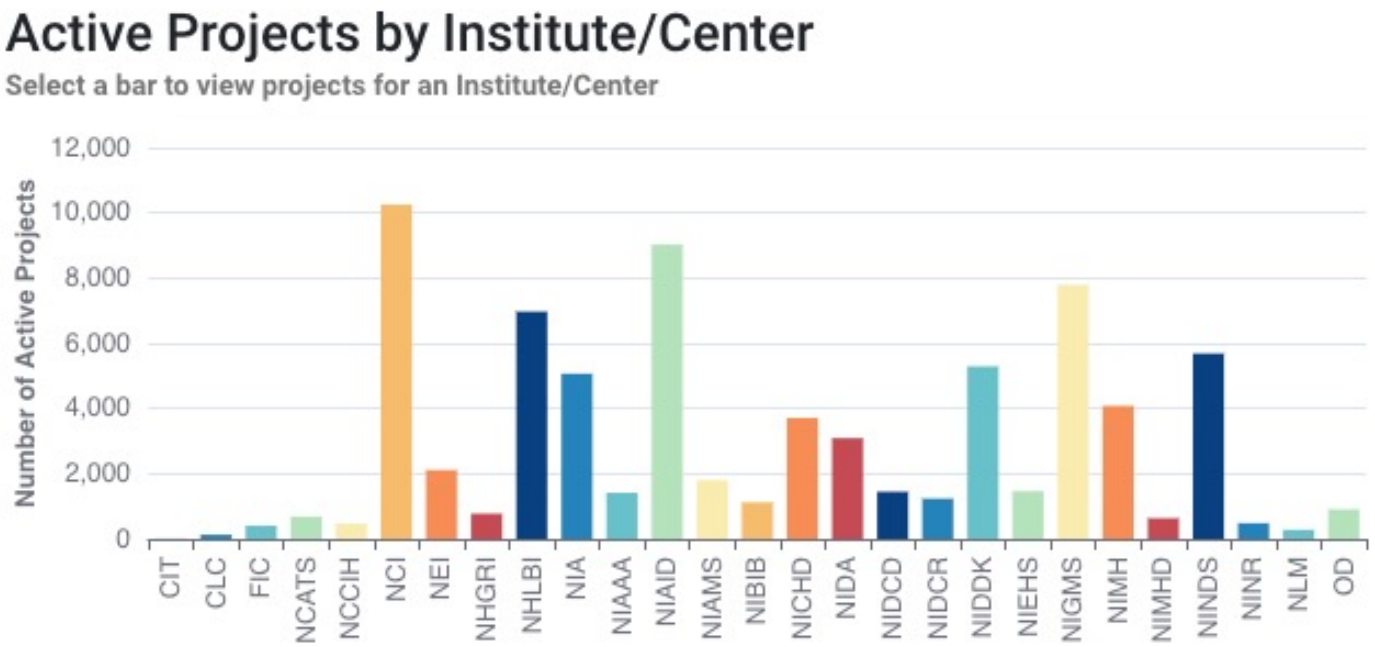
Data was collected from three publicly available NIH resources

RePORTER, PubMed, and MeSH served as the primary data sources

Grants

NIH RePORT
Research Portfolio Online Reporting Tools

- The Research Portfolio Online Reporting Tools (RePORT) provides access to reports, data, and analyses of NIH research activities.
- RePORTER provides access to NIH-funded research projects and their affiliated publications and patents.

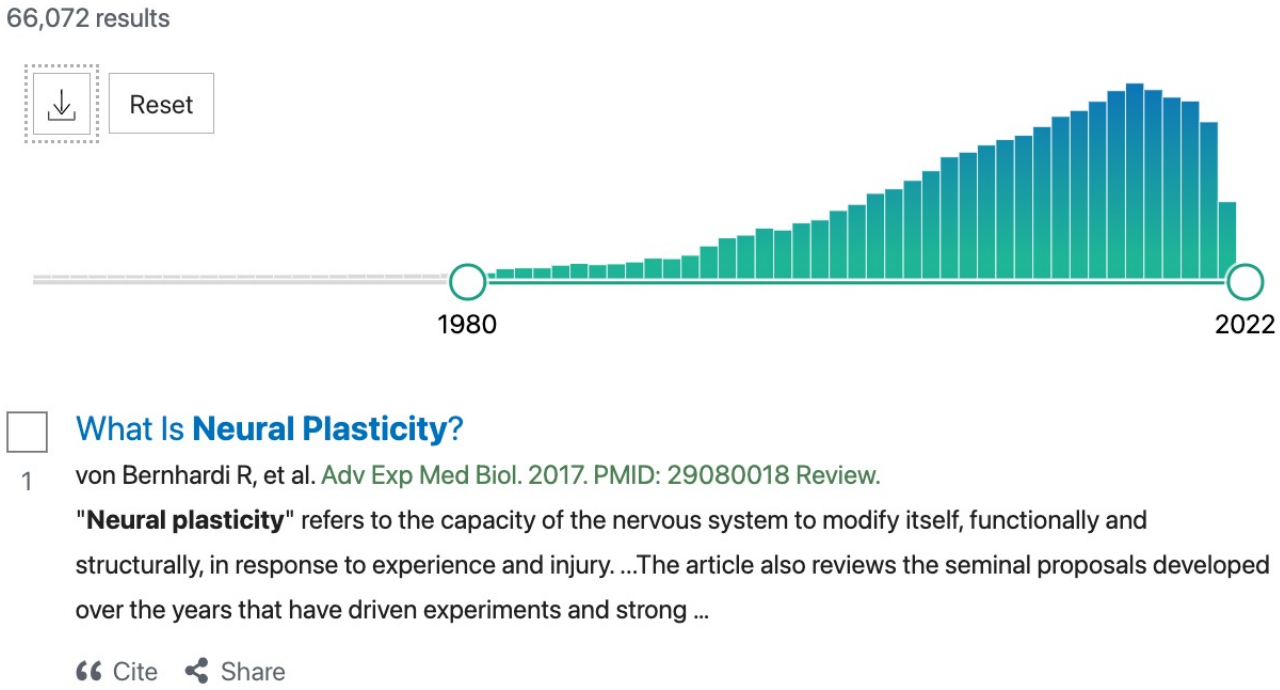


Data was collected from three publicly available NIH resources

RePORTER, PubMed, and MeSH served as the primary data sources

Papers

- PubMed provides 32 million citations from life science journals , books, and other sources of biomedical literature.
- Citations include abstracts and links to full text content from PubMed Central and publisher web sites.



Data was collected from three publicly available NIH resources

RePORTER, PubMed, and MeSH served as the primary data sources

Topics

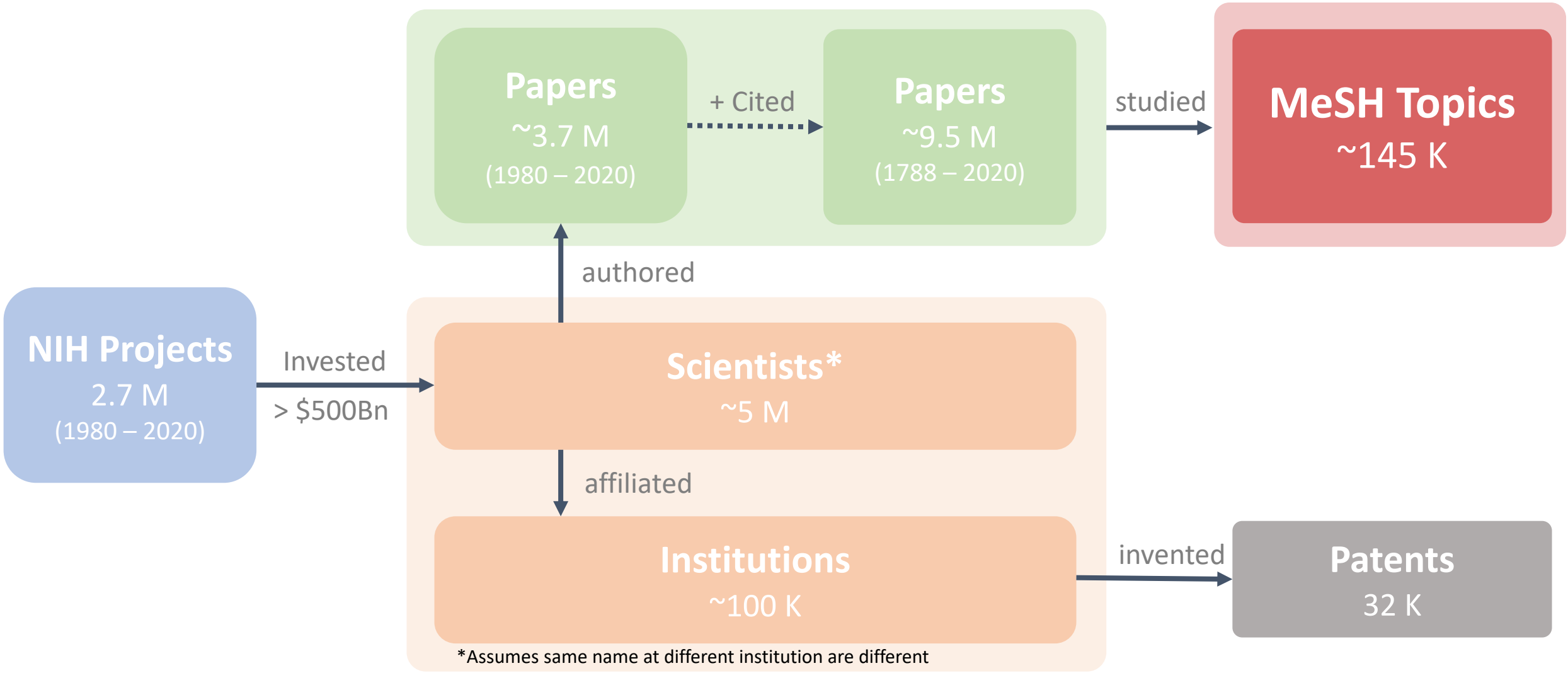
MeSH

- MeSH (Medical Subject Headings) provides a hierarchically organized thesaurus of continually reviewed and updated medical concepts
- Each PubMed paper contains a set of MeSH terms that describe its content.

Neurons [A11.671]
- Adrenergic Neurons [A11.671.050]
- Axons [A11.671.137]
  - Axon Initial Segment [A11.671.137.170]
  - Growth Cones [A11.671.137.340]
  - Mossy Fibers, Hippocampal [A11.671.137.560]
  - Presynaptic Terminals [A11.671.137.750]
- Cholinergic Neurons [A11.671.188]
  - Cholinergic Fibers [A11.671.188.500]
    - Autonomic Fibers, Preganglionic [A11.671.188.500.060]
    - Parasympathetic Fibers, Postganglionic [A11.671.188.500.700]

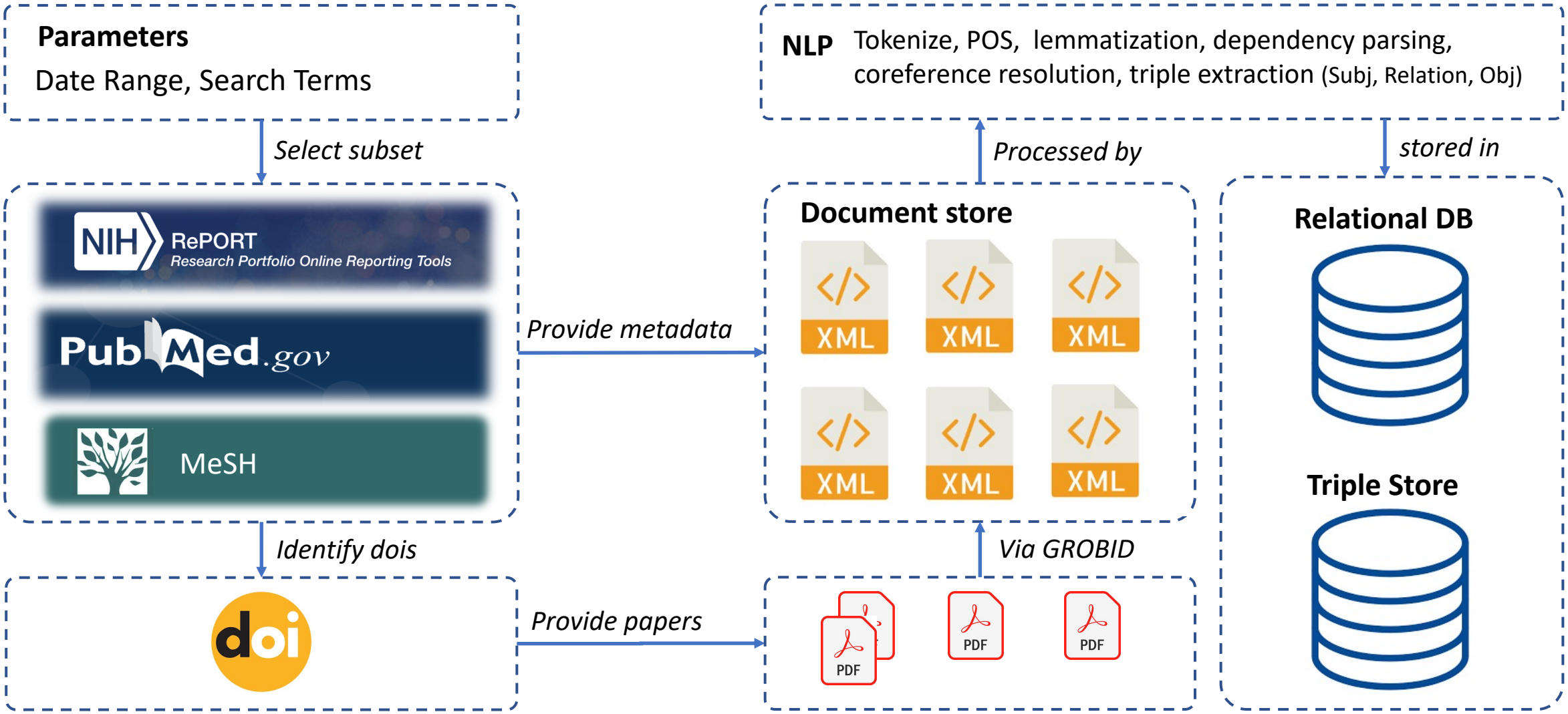
Combined data provides view of science in more complete context

Data unites information on grants, papers, topics, authors, institutions, and patents



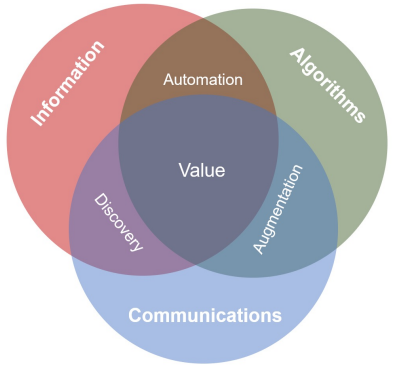
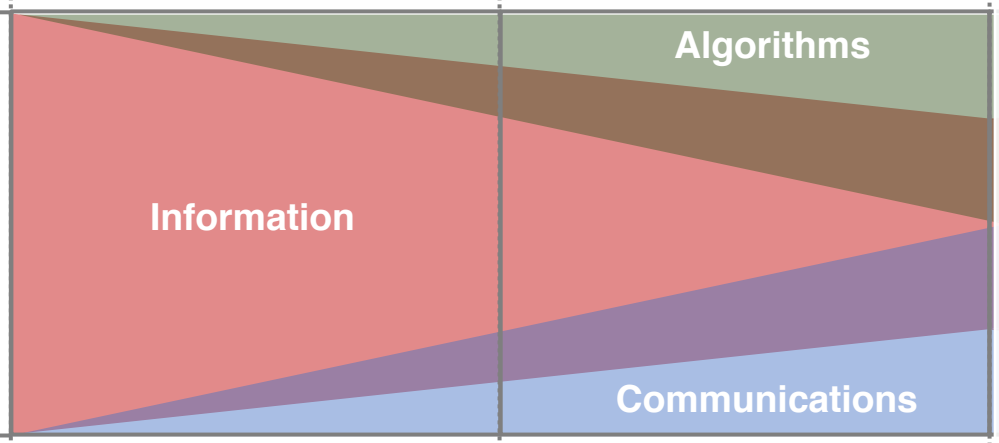
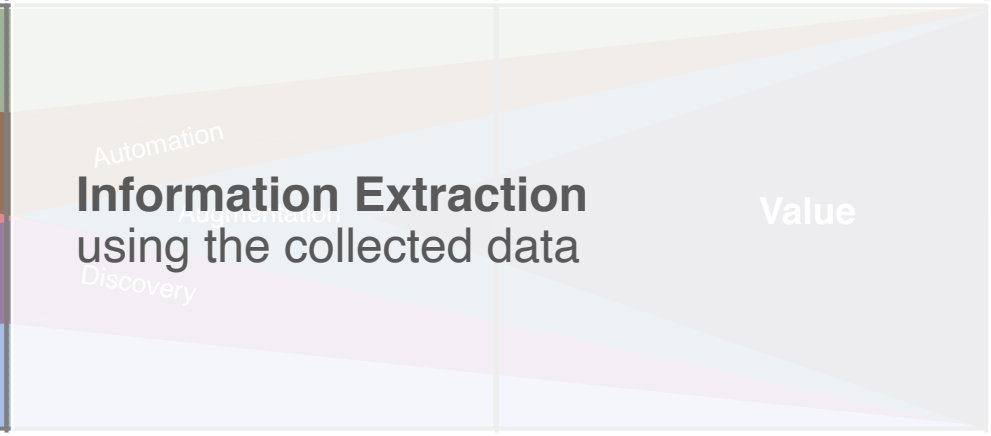
Information collection pipelines accommodate other use cases

Tools enable on-demand download and ingestion of source data based on date, search terms



Next, we developed tools to extract information from documents

Specifically, we focused on NLP processing of documents

1. Specification *January, 2021*	2. Data to Information *February – May, 2021*	3. Algorithms and Comms. *June – August, 2021*	4. Augmentation *September – October, 2021*	5. Value *December, 2021*
				
1.1 Define value 1.2 Define outcomes 1.3 Define Foundations	2.1 Data characterization 2.2 Data collect 2.3 Data centralization	3.1 Automation outcomes 3.2 Communication outcomes 3.3 Performance Evaluation	4.1 Augmentation outcomes 4.2 Integration of novel data 4.3 Performance Evaluation	5.1 Demonstration 5.2 Documentation 5.3 Hand-off

Information Extraction
using the collected data

NLP Pipeline transforms text documents into structured data

Each component of pipeline is modular and can be extended, replaced, or modified

NLP Tokenize, POS, lemmatization, dependency parsing, coreference resolution, triple extraction, entity recognition

Mary had a little lamb whose fleece was black as coal.

She had it last year.

Dolly was a female Finnish Dorset sheep.

She was cloned by associates of the “Roslin Institute” in Scotland.

NLP Pipeline transforms text documents into structured data

Each component of pipeline is modular and can be extended, replaced, or modified

NLP Tokenize, POS, lemmatization, dependency parsing, coreference resolution, triple extraction, entity recognition

Mary had a little lamb whose fleece was black as coal.

She had it last year.

Dolly was a female Finnish Dorset sheep.

She was cloned by associates of the “Roslin Institute” in Scotland.

NLP Pipeline transforms text documents into structured data

Each component of pipeline is modular and can be extended, replaced, or modified

NLP Tokenize, POS, lemmatization, dependency parsing, coreference resolution, triple extraction, entity recognition

Mary had a little lamb whose fleece was black as coal .

She had it last year .

Dolly was a female Finnish Dorset sheep .

She was cloned by associates of the “ Roslin Institute " in Scotland .

NLP Pipeline transforms text documents into structured data

Each component of pipeline is modular and can be extended, replaced, or modified

NLP Tokenize, POS, lemmatization, dependency parsing, coreference resolution, triple extraction, entity recognition

Mary had a little lamb whose fleece was black as coal

She had it last year

Dolly was a female Finnish Dorset sheep

She was cloned by associates of the Roslin Institute in Scotland

NLP Pipeline transforms text documents into structured data

Each component of pipeline is modular and can be extended, replaced, or modified

NLP Tokenize, POS, lemmatization, dependency parsing, coreference resolution, triple extraction, entity recognition

Mary had a little lamb whose fleece be black as coal

She have it last year

Dolly be a female Finnish Dorset sheep

She be clone by associate of the Roslin Institute in Scotland

NLP Pipeline transforms text documents into structured data

Each component of pipeline is modular and can be extended, replaced, or modified

NLP Tokenize, POS, lemmatization, dependency parsing, coreference resolution, triple extraction, entity recognition

Mary had [a little lamb] [whose fleece] be black as coal

She have it last year

Dolly be [a female Finnish Dorset sheep]

She be clone by associate of [the Roslin Institute] in Scotland

NLP Pipeline transforms text documents into structured data

Each component of pipeline is modular and can be extended, replaced, or modified

NLP Tokenize, POS, lemmatization, dependency parsing, coreference resolution, triple extraction, entity recognition

Mary had [a little lamb] [whose fleece] be black as coal

Mary have [a little lamb] last year

Dolly be [a female Finnish Dorset sheep]

Dolly be clone by associate of [the Roslin Institute] in Scotland

NLP Pipeline transforms text documents into structured data

Each component of pipeline is modular and can be extended, replaced, or modified

NLP Tokenize, POS, lemmatization, dependency parsing, coreference resolution, triple extraction, entity recognition

Subject	Relation	Object
Mary	had	[a little lamb]
Mary	Last year have	[a little lamb]
[a little lamb]	be	black
[a little lamb]	be	Black as coal
Dolly	be	[female Finnish Dorset sheep]
Dolly	be	clone
Dolly	be	Clone by [the Roslin Institute]
[the Roslin Institute]	in	Scotland

NLP Pipeline transforms text documents into structured data

Each component of pipeline is modular and can be extended, replaced, or modified

NLP Tokenize, POS, lemmatization, dependency parsing, coreference resolution, triple extraction, entity recognition

Subject	Relation	Object
Mary	had	[a little lamb]
Mary	Last year have	[a little lamb]
[a little lamb]	be	black
[a little lamb]	be	Black as coal
Dolly	be	[female Finnish Dorset sheep]
Dolly	be	clone
Dolly	be	Clone by [the Roslin Institute]
[the Roslin Institute]	in	Scotland

NLP Pipeline transforms text documents into structured data

Each component of pipeline is modular and can be extended, replaced, or modified

NLP Tokenize, POS, lemmatization, dependency parsing, coreference resolution, triple extraction, entity recognition

Subject	Relation	Object
Mary	Last year have	[a little lamb]
[a little lamb]	be	Black as coal
Dolly	be	[female Finnish Dorset sheep]
Dolly	be	Clone by [the Roslin Institute]
[the Roslin Institute]	in	Scotland

Unified Medical Language System (UMLS)

The UMLS integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems.

NLP Pipeline transforms text documents into structured data

Each component of pipeline is modular and can be extended, replaced, or modified

NLP Tokenize, POS, lemmatization, dependency parsing, coreference resolution, triple extraction, entity recognition

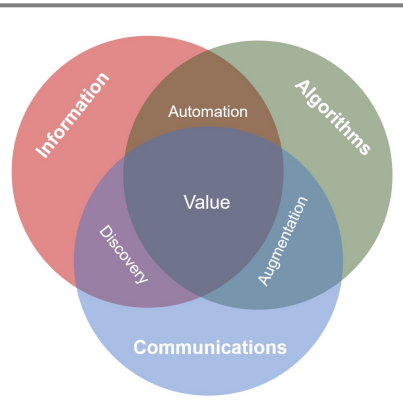
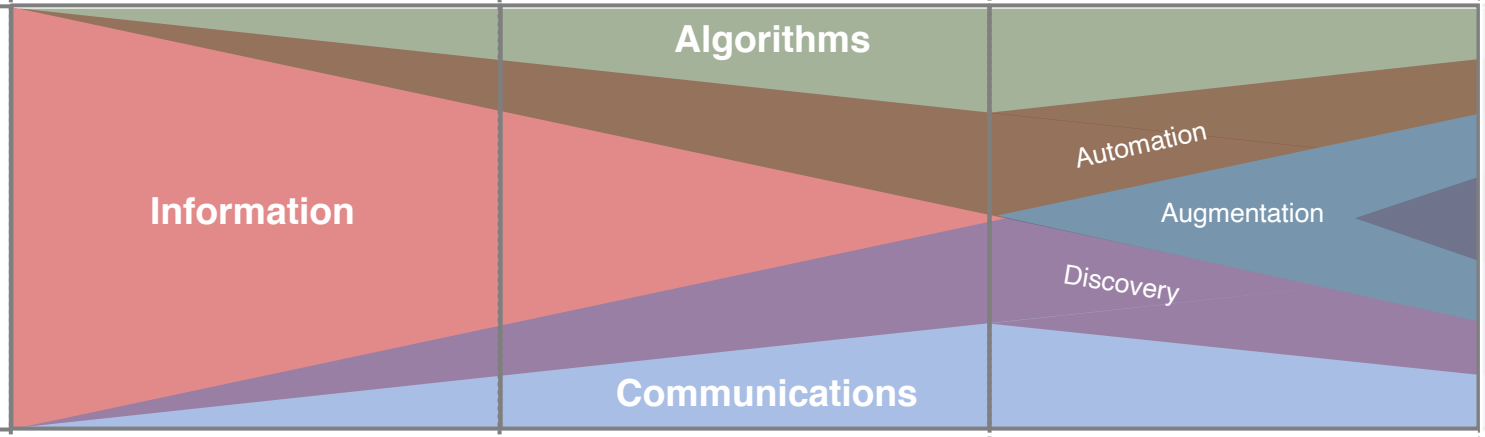
Subject	Relation	Object
[a little lamb]	be	Black as coal
[the Roslin Institute]	in	Scotland

Unified Medical Language System (UMLS)

The UMLS integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems.

Next, we developed tools to represent and visualize the data

Specifically, we generated dynamic networks visualizations

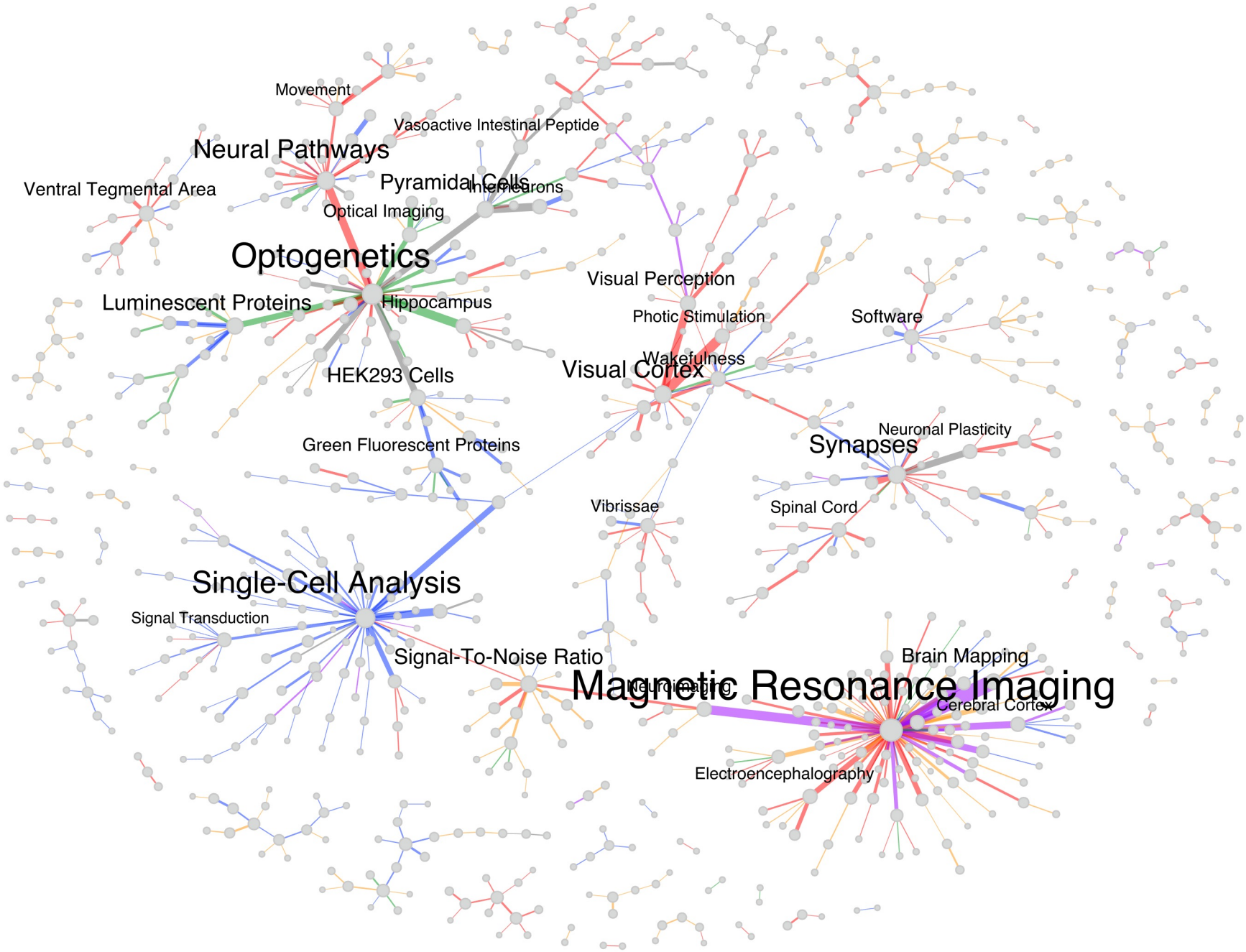
1. Specification *January, 2021*	2. Data to Information *February – May, 2021*	3. Algorithms and Comms. *June – August, 2021*	4. Augmentation *September – November, 2021*	5. Value *December, 2021*
				**Generate Dynamic Networks** using the collected data
1.1 Define value 1.2 Define outcomes 1.3 Define Foundations	2.1 Data characterization 2.2 Data collect 2.3 Data centralization	3.1 Automation outcomes 3.2 Communication outcomes 3.3 Performance Evaluation	4.1 Augmentation outcomes 4.2 Integration of novel data 4.3 Performance Evaluation	5.1 Demonstration 5.2 Documentation 5.3 Hand-off

Publication topic graph of BRAIN awardees (2014 – 2020)

Nodes: topics in papers
- Size: number of papers published

Edges: topic co-occurrence in papers
- Color: NIH BRAIN Team
- Gray: increased inter-group output

Exclusion Criteria:
- Edges: only rank 1 edges are shown.

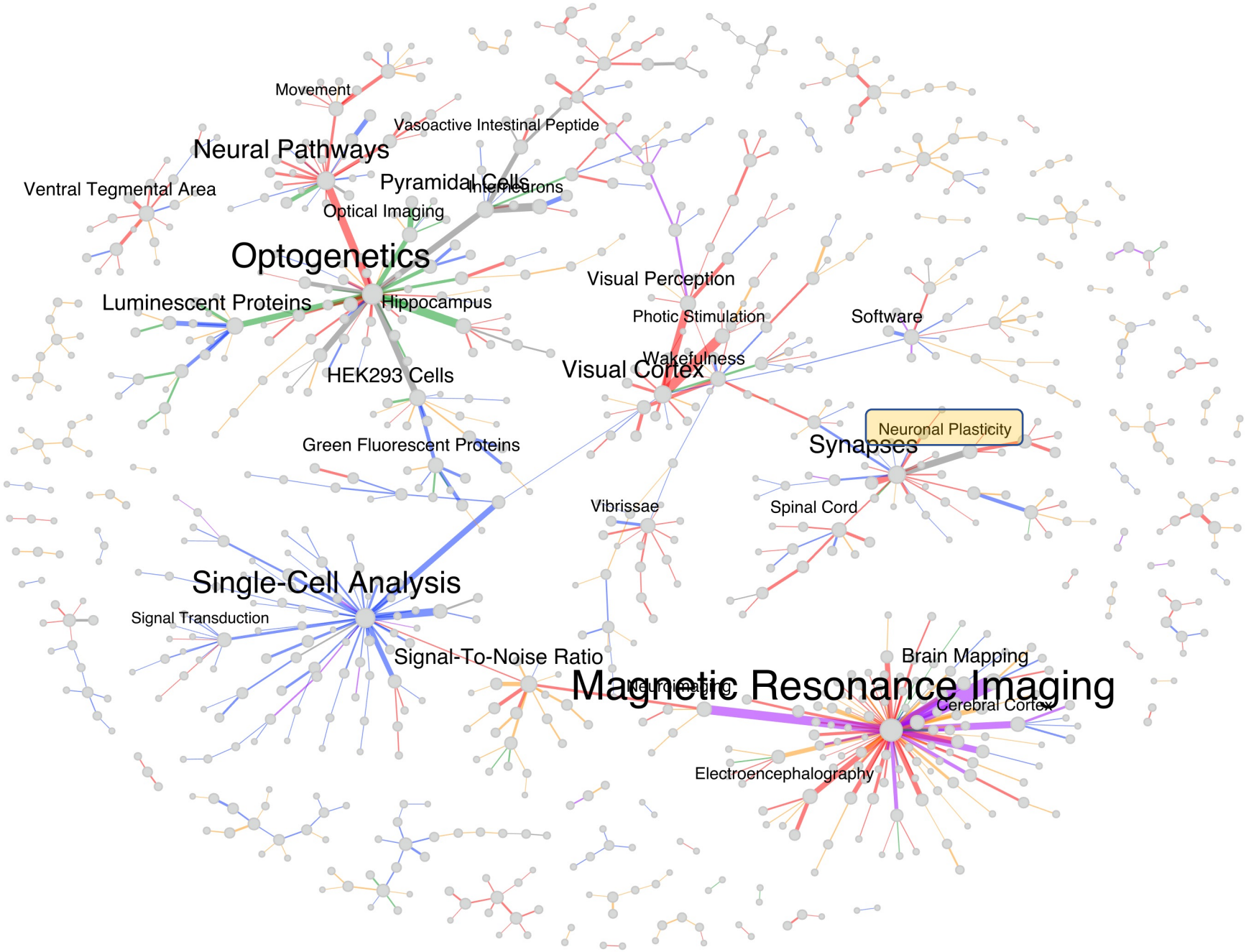


Publication topic graph of BRAIN awardees (2014 – 2020)

Nodes: topics in papers
- Size: number of papers published

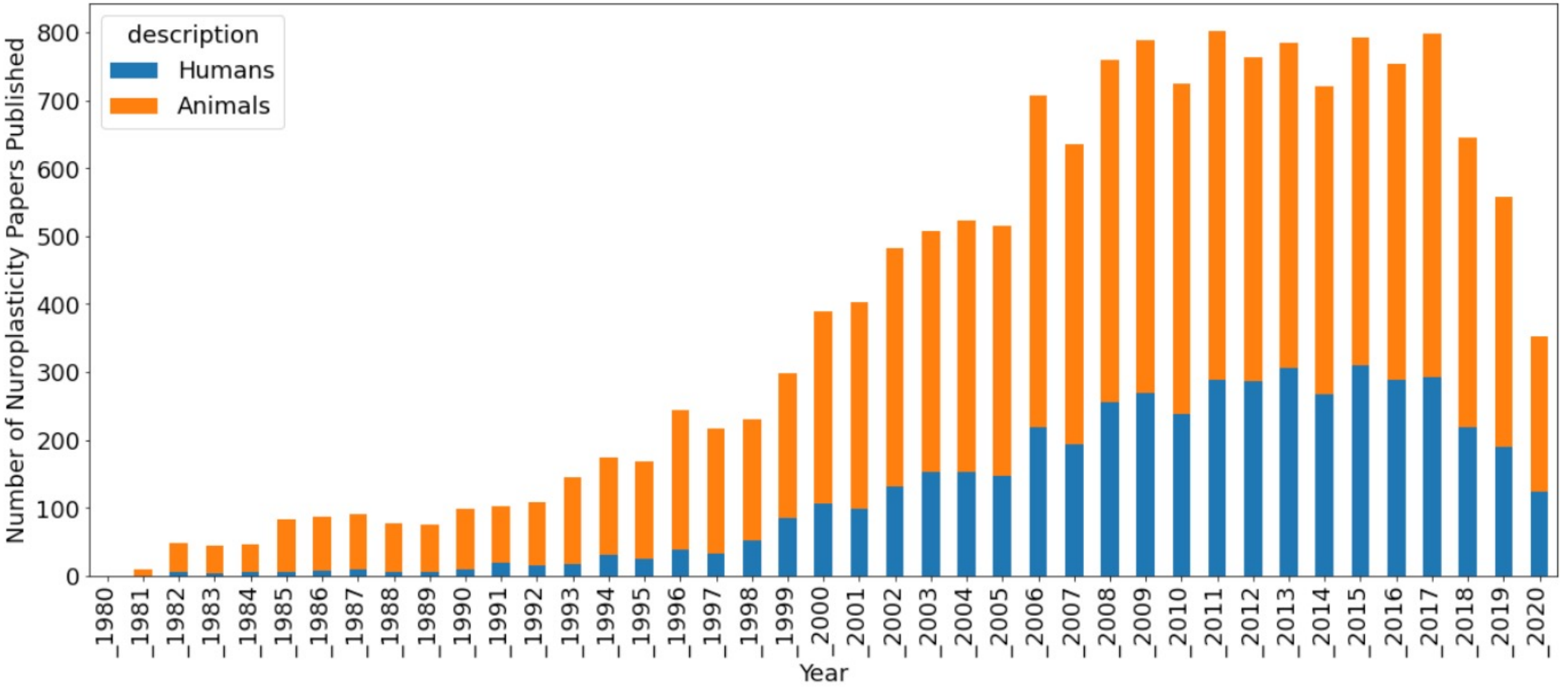
Edges: topic co-occurrence in papers
- Color: NIH BRAIN Team
- Gray: increased inter-group output

Exclusion Criteria:
- Edges: only rank 1 edges are shown.



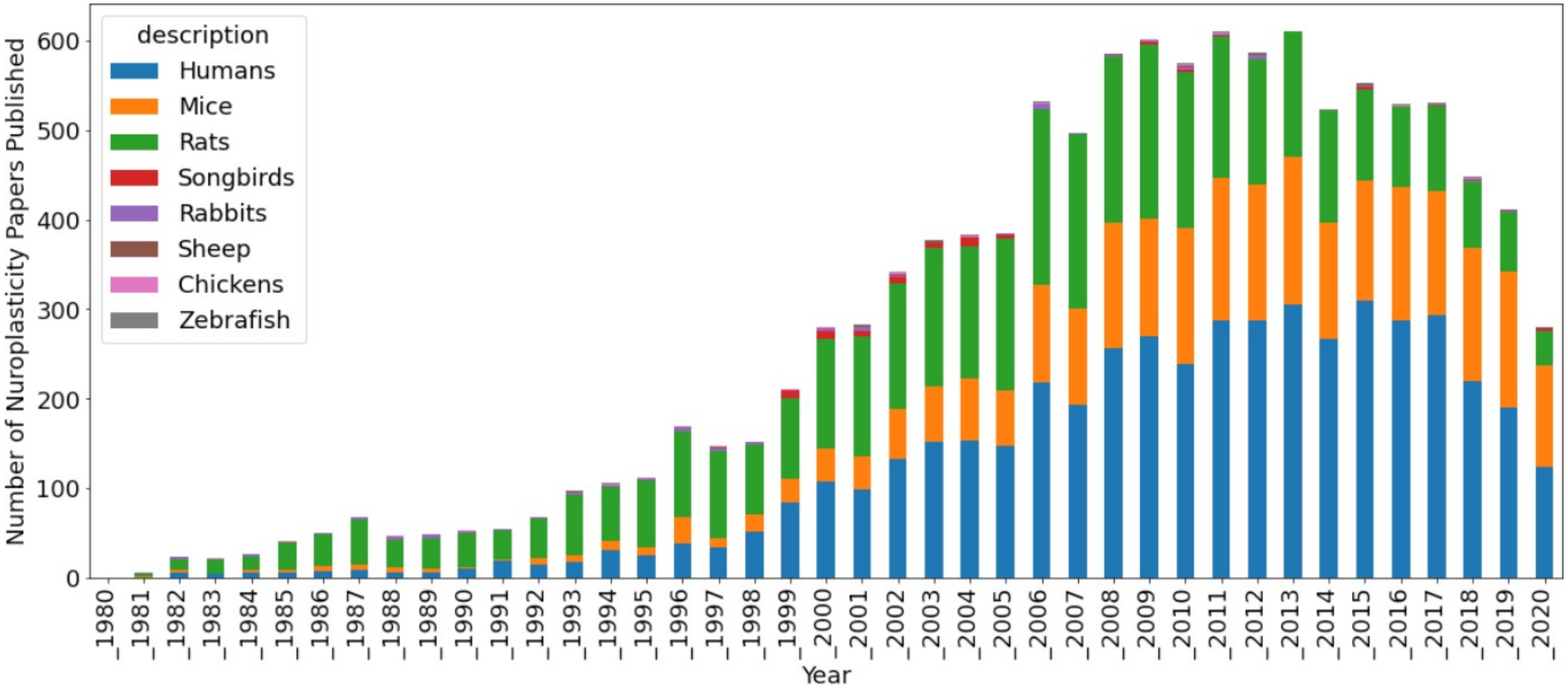
Co-occurrence of topics in publications provides research trends

Trends in ‘Neuroplasticity’ research may be understood by topics that covary with it



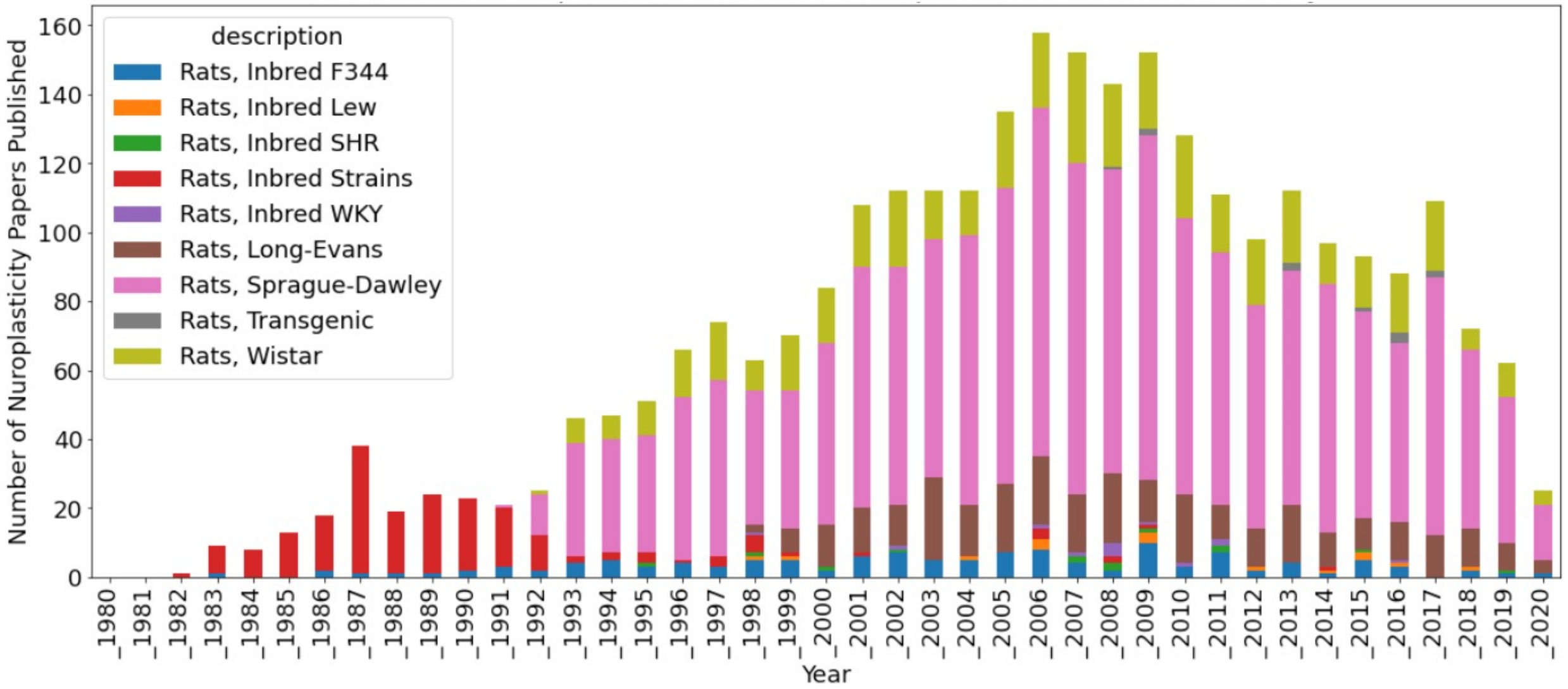
Co-occurrence of topics in publications provides research trends

Trends in ‘Neuroplasticity’ research may be understood by topics that covary with it



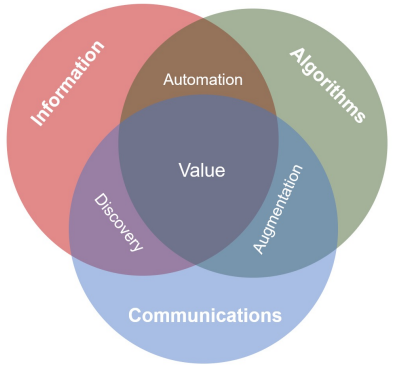
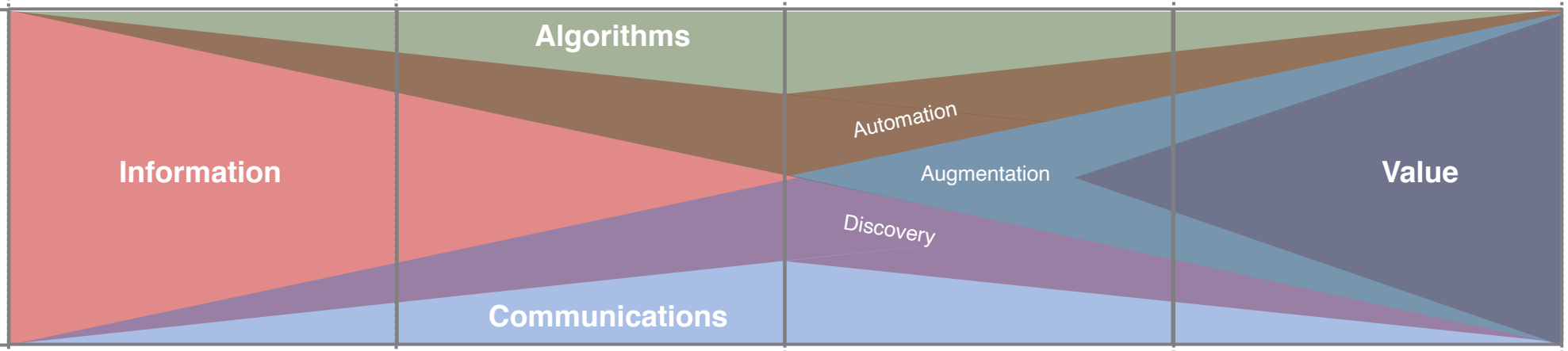
Co-occurrence of topics in publications provides research trends

Trends in ‘Neuroplasticity’ research may be understood by topics that covary with it



Finally, we have leveraged the tools in several ways to create value

Specifically, developed tools for graph search, representation, and group comparison

1. Specification *January, 2021*	2. Data to Information *February – May, 2021*	3. Algorithms and Comms. *June – August, 2021*	4. Augmentation *September – November, 2021*	5. Value *December, 2021*
				
1.1 Define value 1.2 Define outcomes 1.3 Define Foundations	2.1 Data characterization 2.2 Data collect 2.3 Data centralization	3.1 Automation outcomes 3.2 Communication outcomes 3.3 Performance Evaluation	4.1 Augmentation outcomes 4.2 Integration of novel data 4.3 Performance Evaluation	5.1 Demonstration 5.2 Documentation 5.3 Hand-off

Several use cases were explored to validate the tool

BRAINWORKS provides information, algorithms and visualization tools that can be used in other contexts

- **Information**
  - Normalize research impact by time and topic areas.
  - Disambiguate authors and organizations in citations.
  - Extract and prioritize knowledge triples (subject, relation, object) from free text

- **Algorithms**
  - Predict impact of interventions on graph structure
  - Predict prospective graph structure from graph dynamics
  - Identify and map scientific entities in free text to established ontologies (UMLS).

- **Interactive Visualizations**
  - The evolution of scientific theories (subject, relation, object) in time.
  - The hierarchical relationship of knowledge in graphs.

How do collaborative projects impact author networks and papers?

How do collaborative projects impact author networks and papers?

- Retrospective population:
  - ~1,500 Researchers affiliated with ~60 lab:
  - Intervention: ~30 labs received an award in 2017
  - Control: ~30 labs with best record for R01 acquisition.

- Study period:
  - 2013 – 2016: 4 years before the intervention group received their award
  - 2017 – 2020: 4 years following the intervention group receiving their award

- Characteristics of interest
  - *Changes* in publication output and collaborations

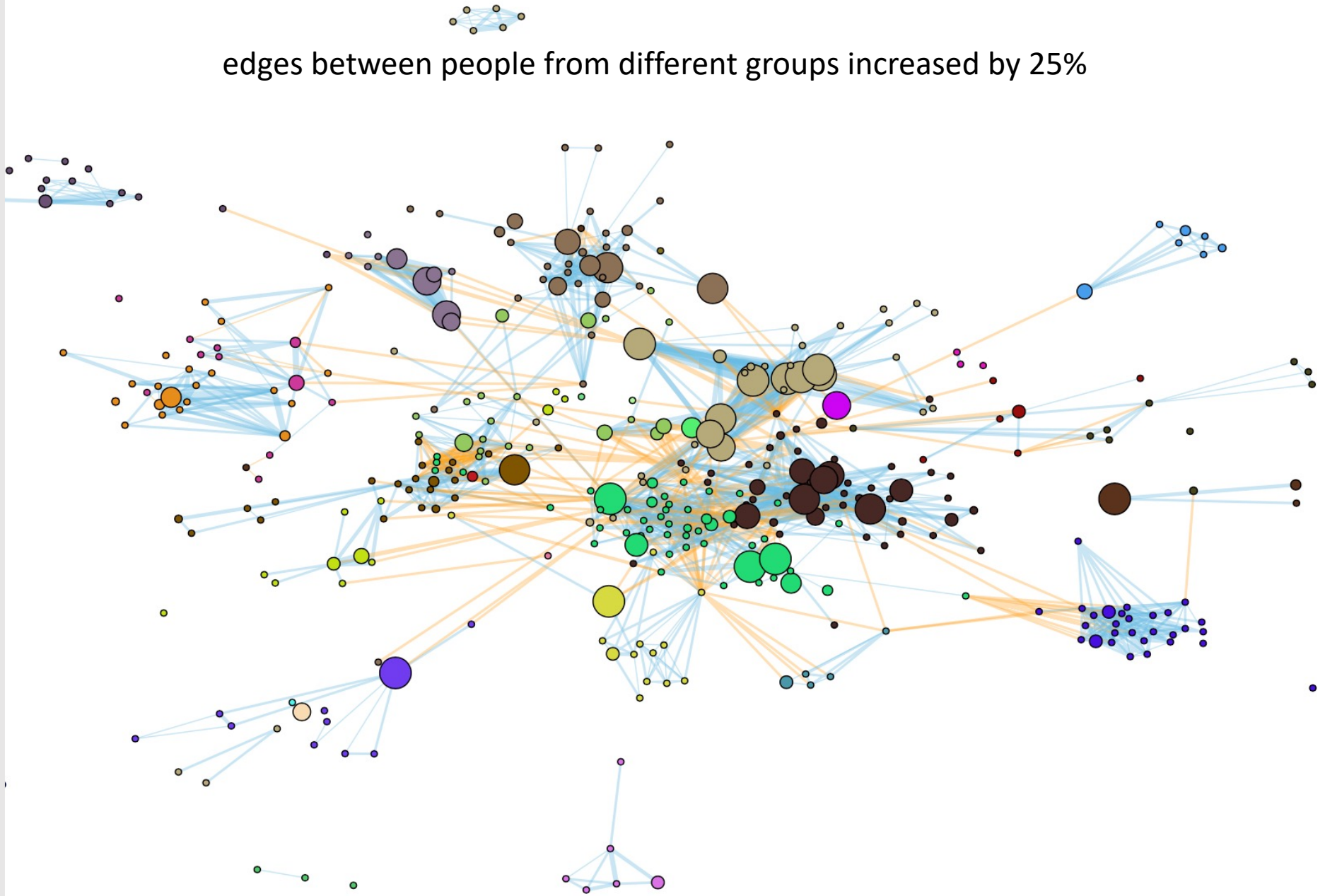
Changes in Neuroscience collaborations

Random Control
(4 year contrast)

Nodes: individual authors
- Color: investigator group
- Size: increase in publication output from [2013-2016] to [2017-2020]

Edges: author collaboration
- Orange: increased intra-group output
- Blue: increased inter-group output

Exclusion Criteria:
- Nodes: authors must have authored in both periods
- Nodes: authors with decreased or stable collaborations not shown.
- Edges: reductions in collaborations not shown.



Changes in Neuroscience collaborations

Intervention Group (4 year contrast)

Nodes: individual authors

- Color: investigator group
- Size: increase in publication output from [2013-2016] to [2017-2020]

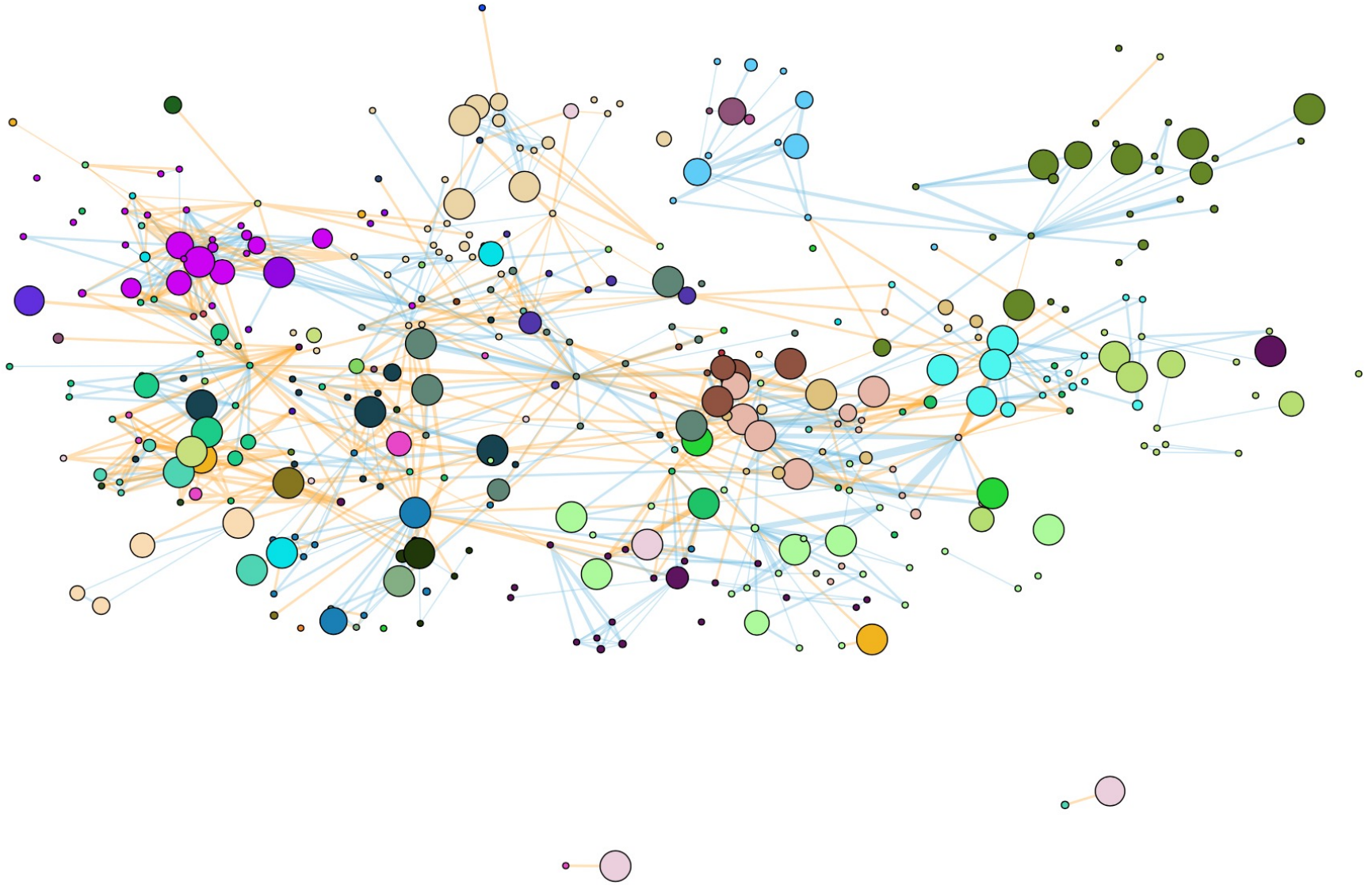
Edges: author collaboration

- Orange: increased intra-group output
- Blue: increased inter-group output

Exclusion Criteria:

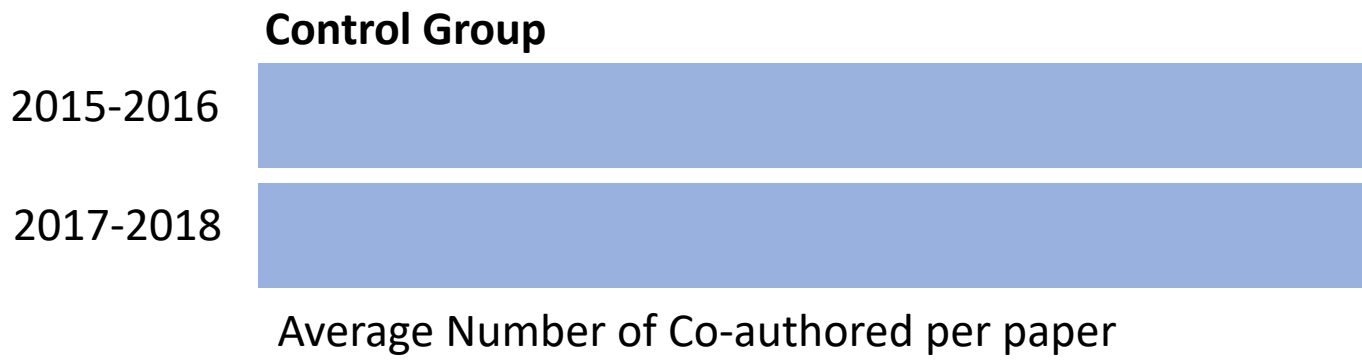
- Nodes: authors must have authored in both periods
- Nodes: authors with decreased or stable collaborations not shown.
- Edges: reductions in collaborations not shown.

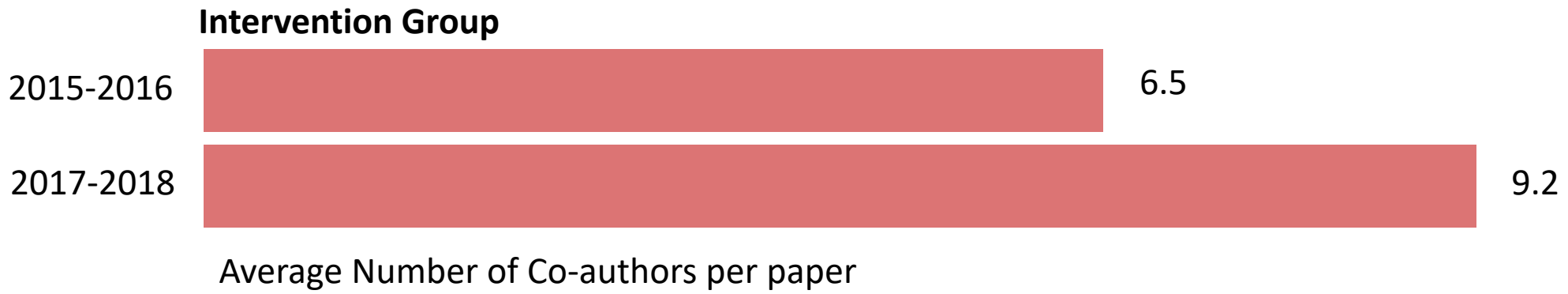
edges between people from different groups increased by 2x



Associativity = 0.49

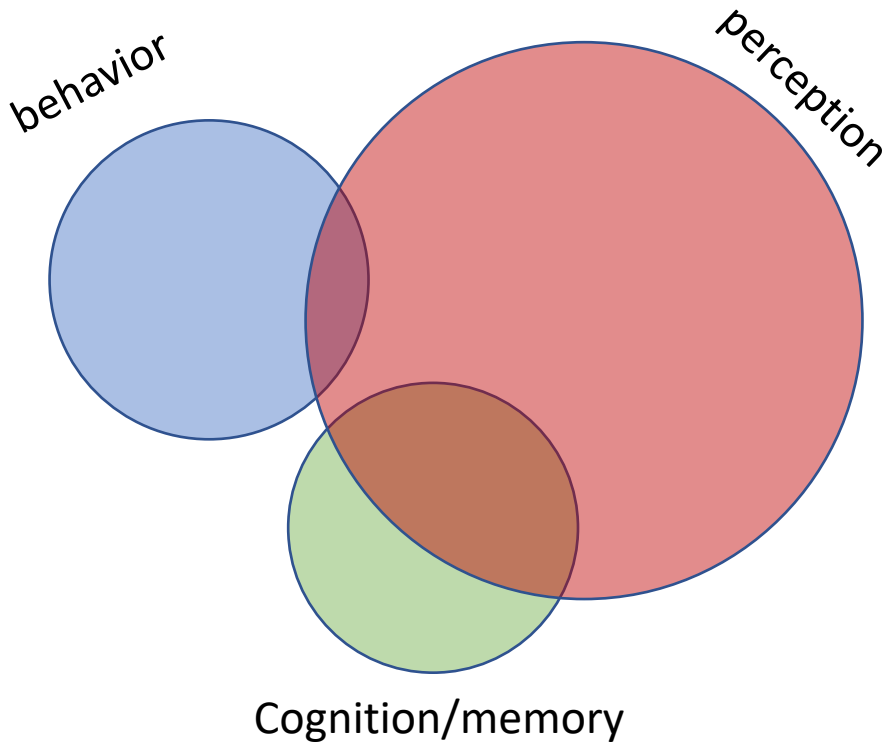
How do collaborative projects impact author networks and papers?



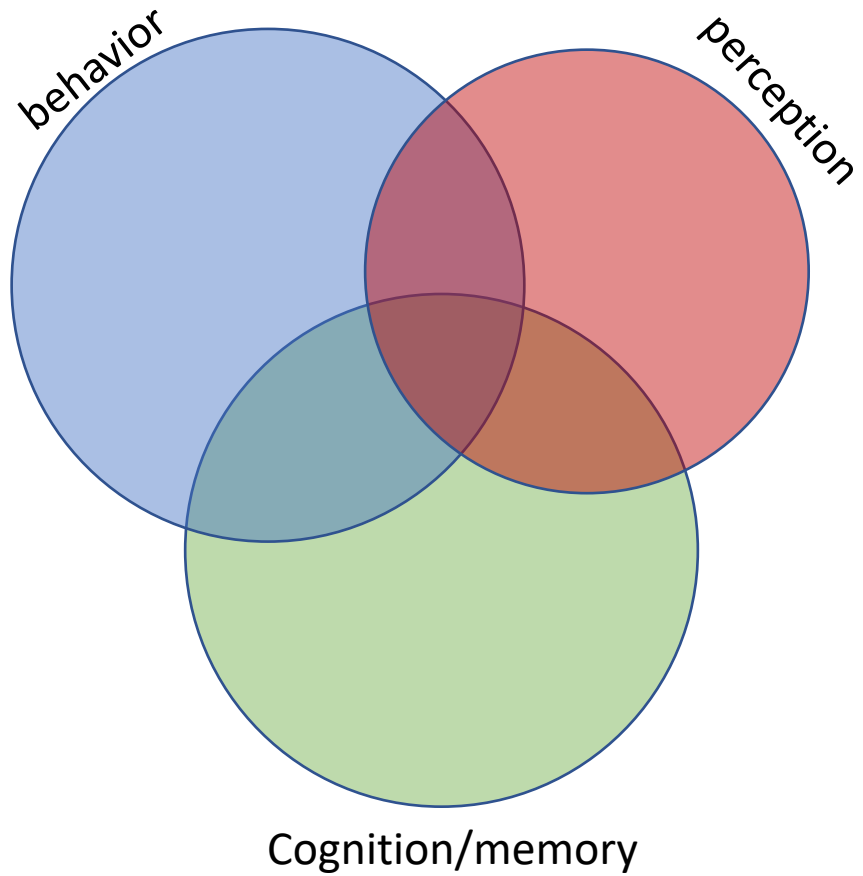


How do collaborative projects impact author networks and papers?

Control Group – 2017/2018



Intervention Group – 2017/2018



Can we predict future graph edges via graph dynamics?

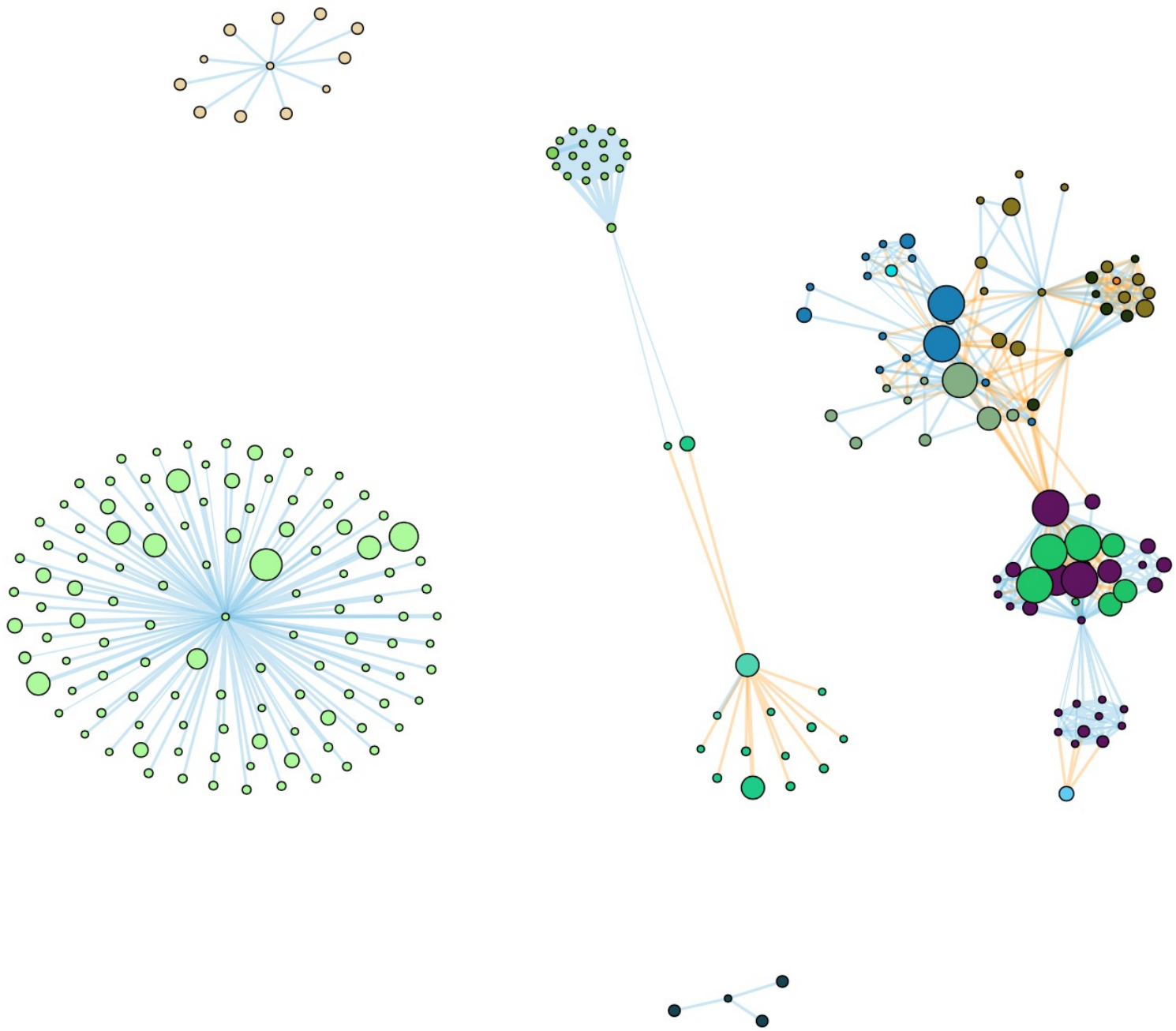
Changes in Neuroscience collaborations

Intervention Group
(one cluster – 2 year contrast)

Nodes: individual authors
- Color: investigator group
- Size: increase in publication output from [2015/2016] to [2017/2018]

Edges: author collaboration
- Orange: increased intra-group output
- Blue: increased inter-group output

Exclusion Criteria:
- Nodes: authors with decreased or stable collaborations not shown.
- Edges: reductions in collaborations not shown.



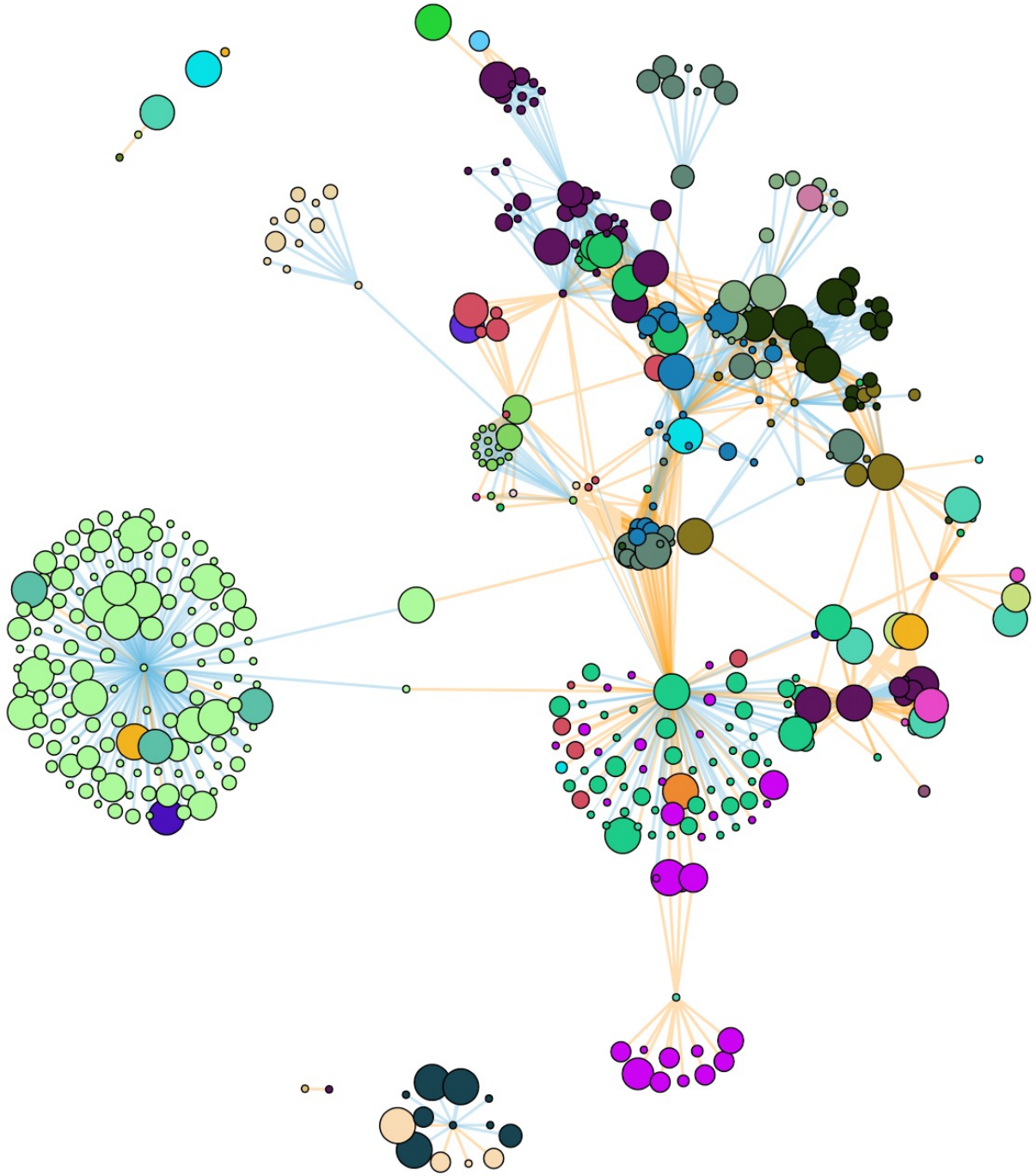
Changes in Neuroscience collaborations

Intervention Group
(one cluster – 4 year contrast)

Nodes: individual authors
- Color: investigator group
- Size: increase in publication output from [2013-2016] to [2017-2020]

Edges: author collaboration
- Orange: increased intra-group output
- Blue: increased inter-group output

Exclusion Criteria:
- Nodes: authors with decreased or stable collaborations not shown.
- Edges: reductions in collaborations not shown.



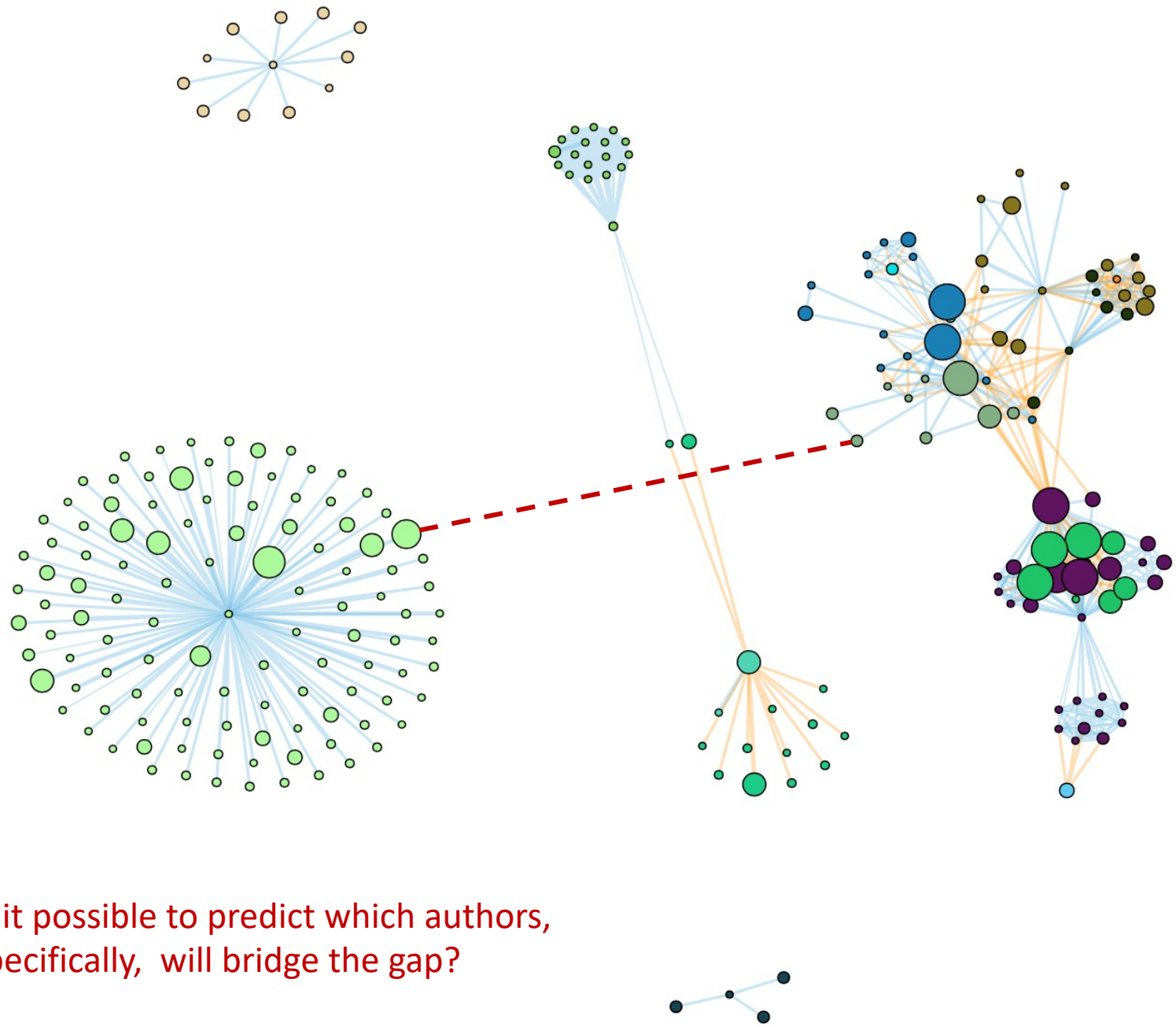
Changes in Neuroscience collaborations

Intervention Group
(one cluster – 2 year contrast)

Nodes: individual authors
- Color: investigator group
- Size: increase in publication output from [2015/2016] to [2017/2018]

Edges: author collaboration
- Orange: increased intra-group output
- Blue: increased inter-group output

Exclusion Criteria:
- Nodes: authors with decreased or stable collaborations not shown.
- Edges: reductions in collaborations not shown.

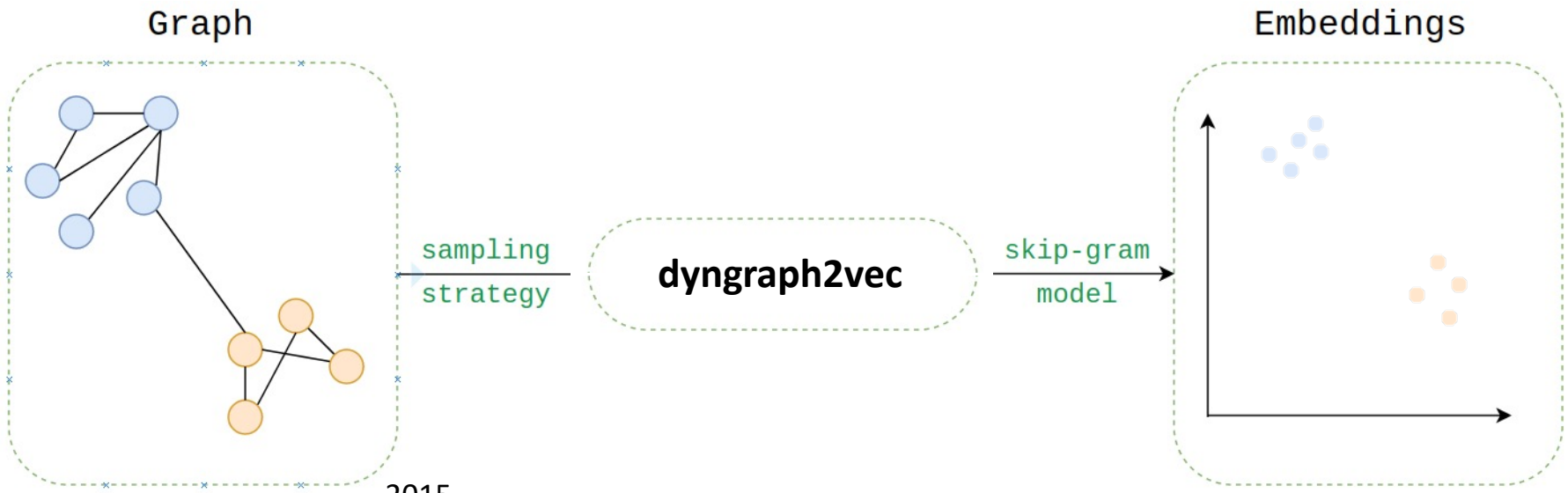


Is it possible to predict which authors, specifically, will bridge the gap?

Can we predict future graph edges via graph dynamics?

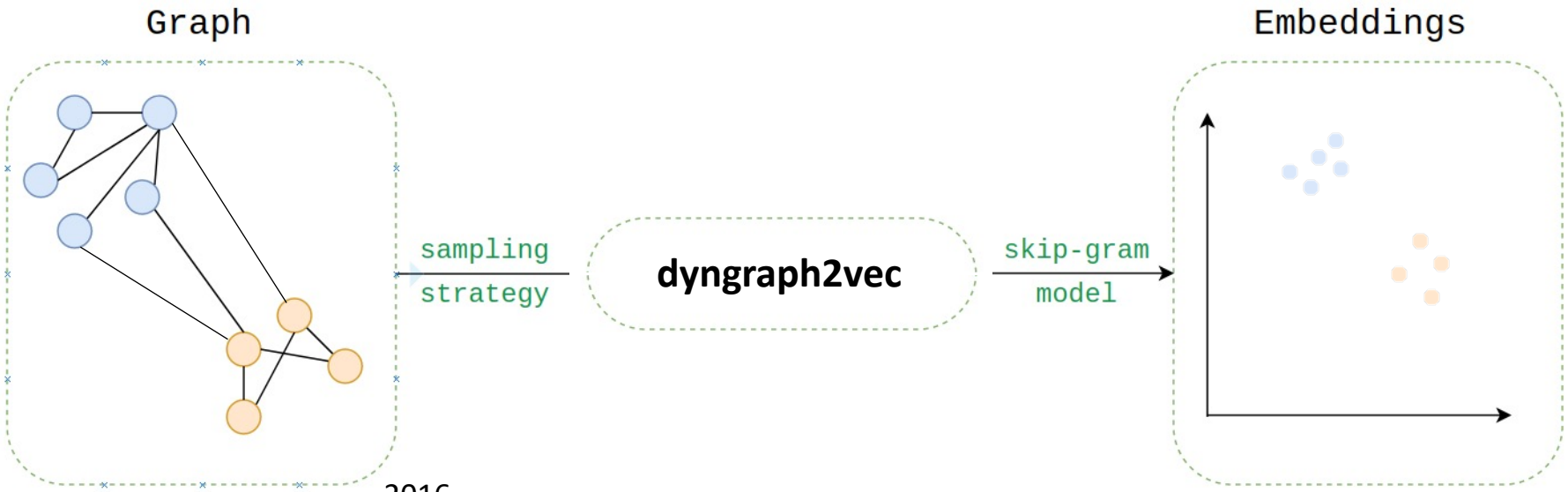
- Retrospective population:
  - ~62,000 published scientists - nodes
  - ~600,000 coauthorship - edges
- Study period:
  - 2015 – 2019: 5 years
- Characteristics of interest
  - Do co-authorship networks in 2015-2018 predict links between the nodes of the network in 2019?

Can we predict future graph edges via graph dynamics?



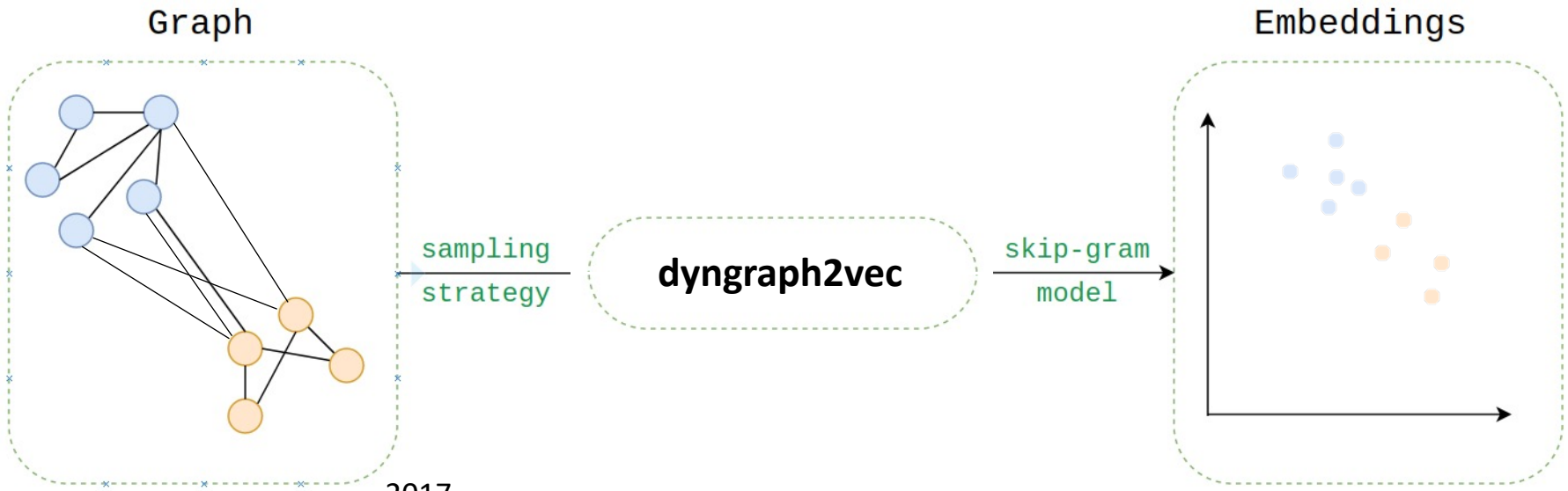
2015

Can we predict future graph edges via graph dynamics?



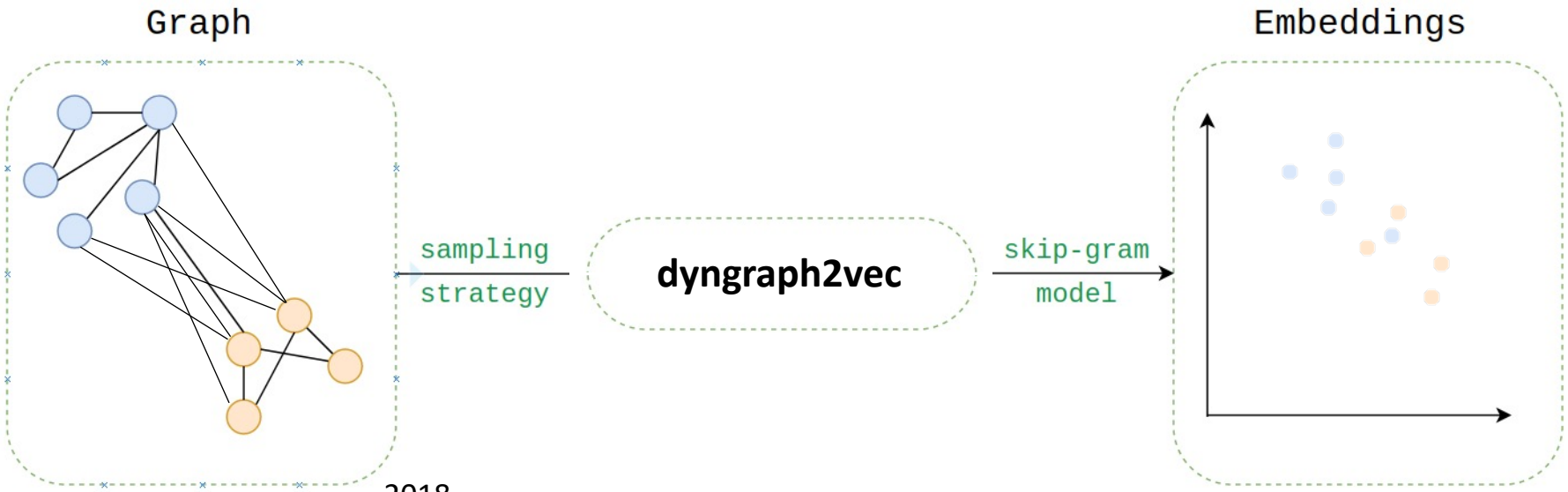
2016

Can we predict future graph edges via graph dynamics?



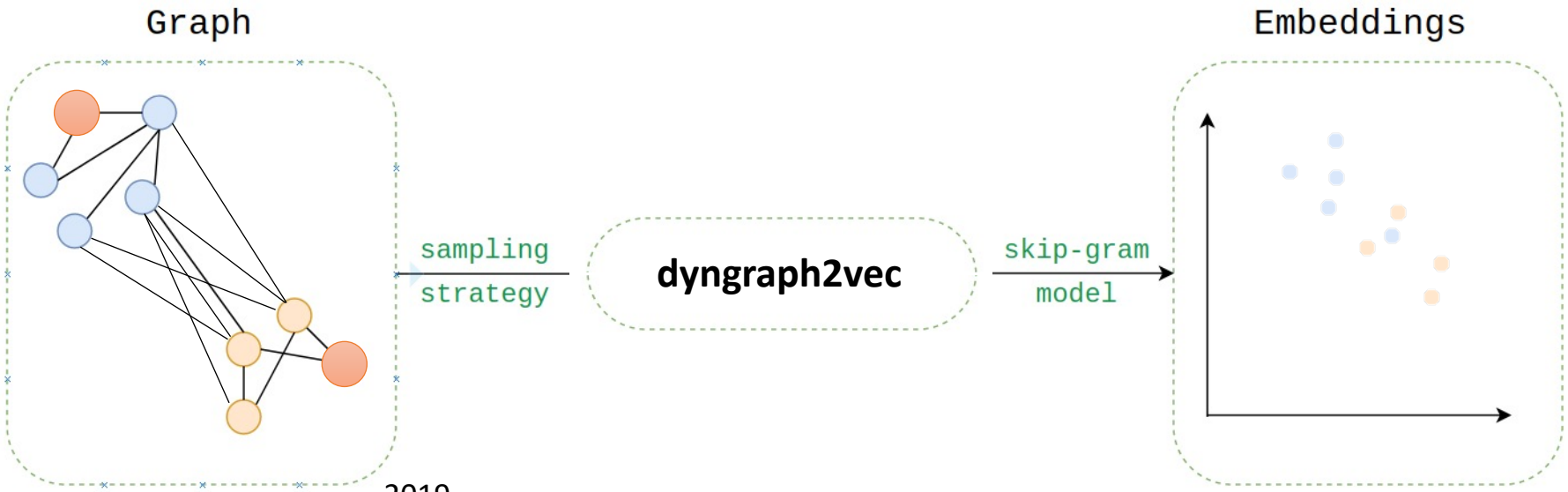
2017

Can we predict future graph edges via graph dynamics?



2018

Can we predict future graph edges via graph dynamics?



2019

Given the position, and dynamics (velocity and acceleration) of the embeddings,
we can predict the future propensity of an edge between two nodes in the future with higher fidelity than baseline approaches: **F1 Score (Binary) 0.18 vs. 0.13 literature baseline**.

Contact Information

Mohammad Ghassemi, PhD

mohammad.ghassemi@nih.gov