9PM

2AM in London (GMT), 11AM in Tokyo (GMT+9)

Multiscale Models

Moderator: Katy Börner, *Indiana University*

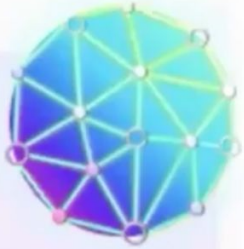
Presenters:

- Maria Brbic, *Swiss Federal Institute of Technology Lausanne, Switzerland*
- Filipi N. Silva, *Indiana University*

Maria Brbic, *Swiss Federal Institute of Technology Lausanne*

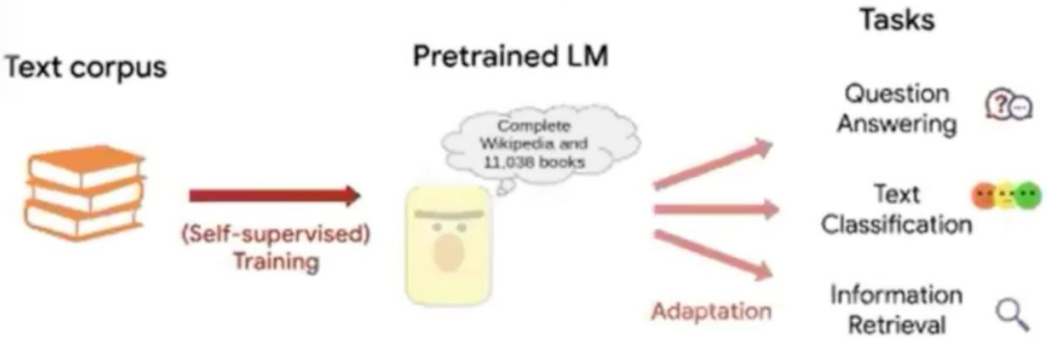
AI Revolution

Generative AI paradigm and the era of foundation models



Foundation Model

Model that has learnt features transferable to a wide range of downstream tasks



How can we leverage these AI advances in single cell biology?

What are their current limitations for biomedical applications?

Single-cell Data Is Challenging for Today's AI

1. Heterogenous experiments
2. Novel and unknown phenomena
3. Different modalities with different challenges

1. Heterogenous experiments
2. Novel and unknown phenomena
3. Different modalities

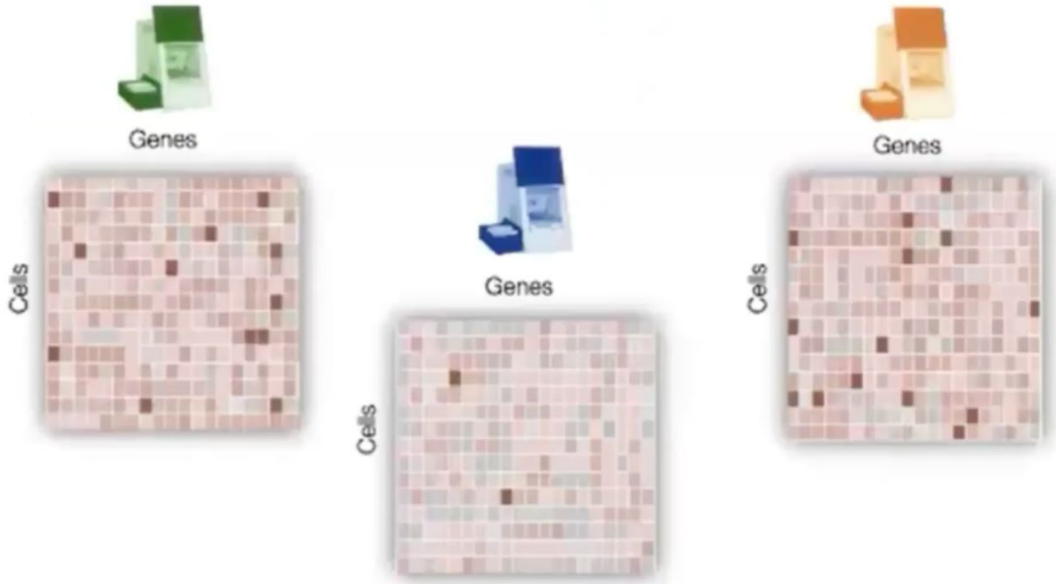
Today's talk: How to overcome some of these challenges

On Heterogeneity

Discovering Cell Types Across Tissues, Disease States & Species

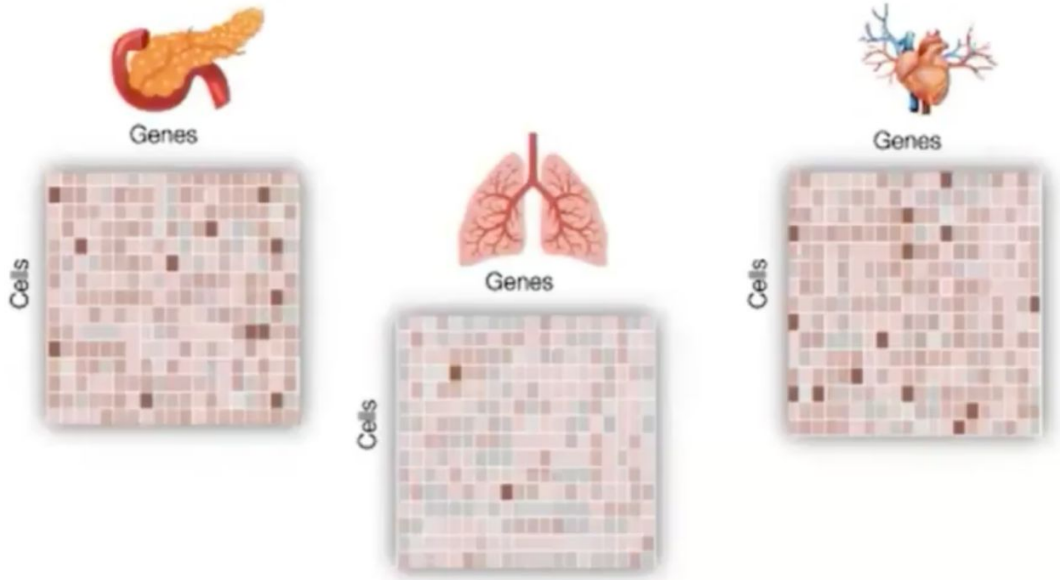
Data with Large Heterogeneity

different labs…



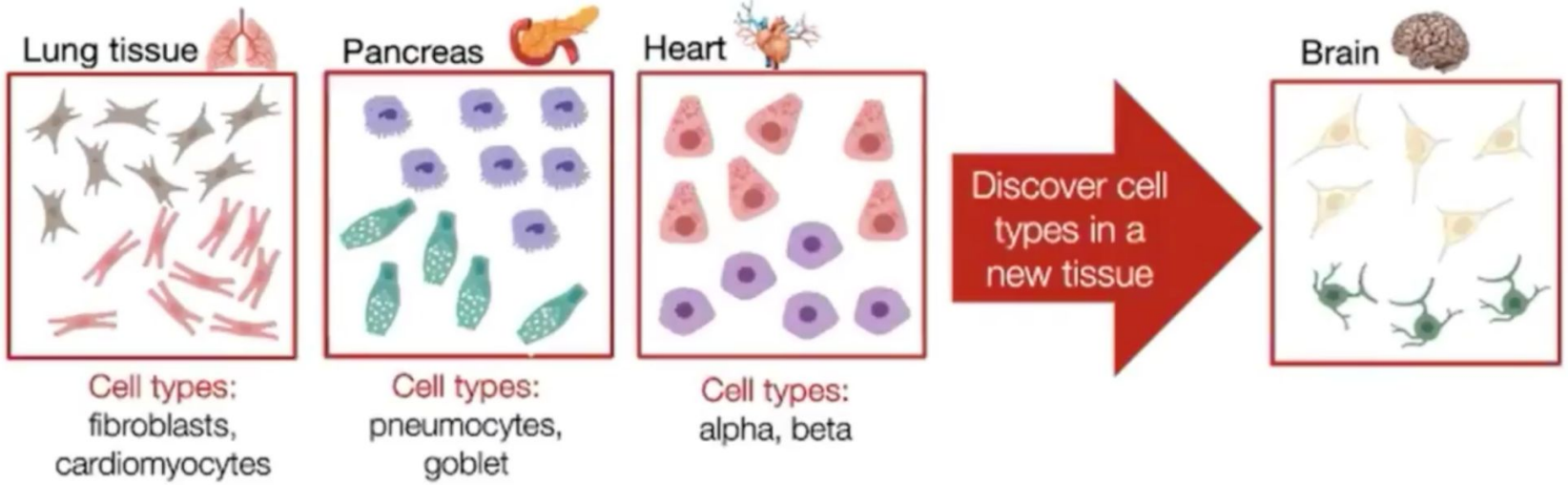
Data with Large Heterogeneity

different tissues…



How do we jointly analyze and gain new insights from these heterogenous datasets?

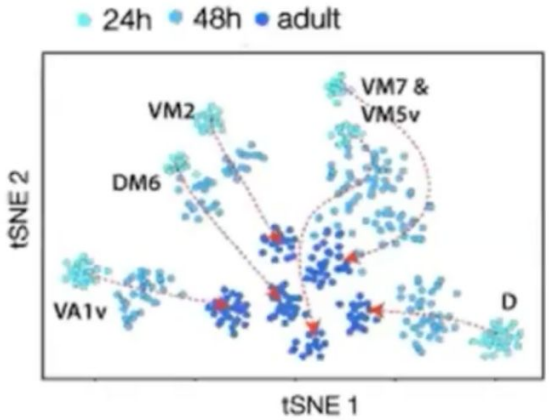
MARS: Learn Cell Embeddings to Discover Novel Cell Types



Brbic et al. *Nature Methods*' 20

Cell Type Discovery across Experiments

Across tissues of the Mouse Cell Atlas

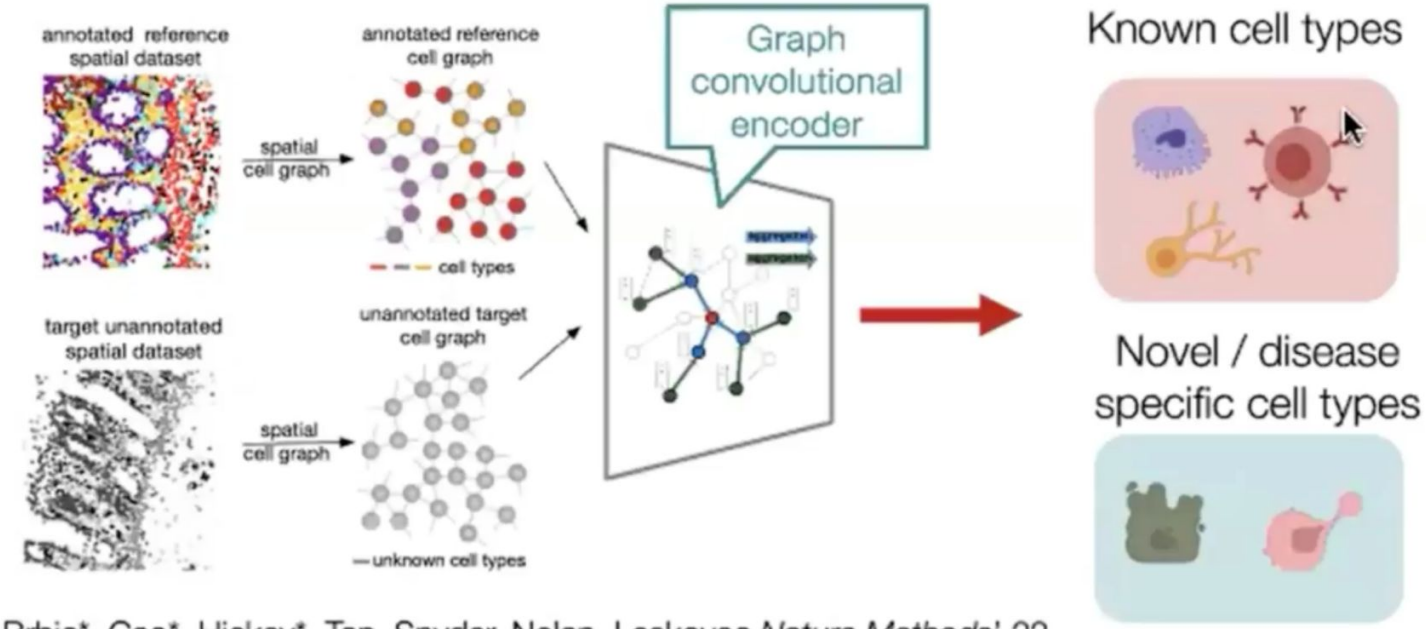


Xie*, Brbic* et al. *eLife* '21

Fly Cell Atlas

Li*, Janssens* et al. *Science* '22

STELLAR: Novel Cell Type Discovery Across Conditions



Brbic*, Cao*, Hickey*, Tan, Snyder, Nolan, Leskovec *Nature Methods*' 22

Towards Universal Cell Embeddings

Can we create cell embeddings for any species, any set of genes?

Tabula Muris
Nature '18, '20

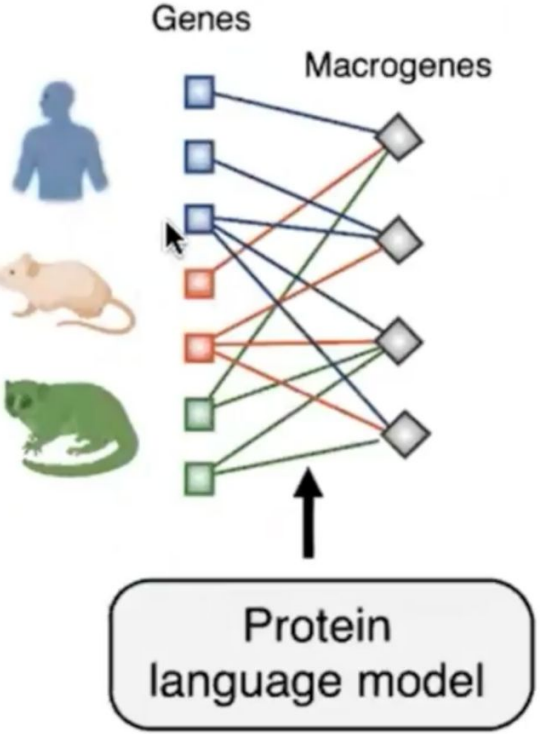
Fly Cell Atlas
Science '22 '23

Tabula Sapiens
Science '22

Our Approach: SATURN

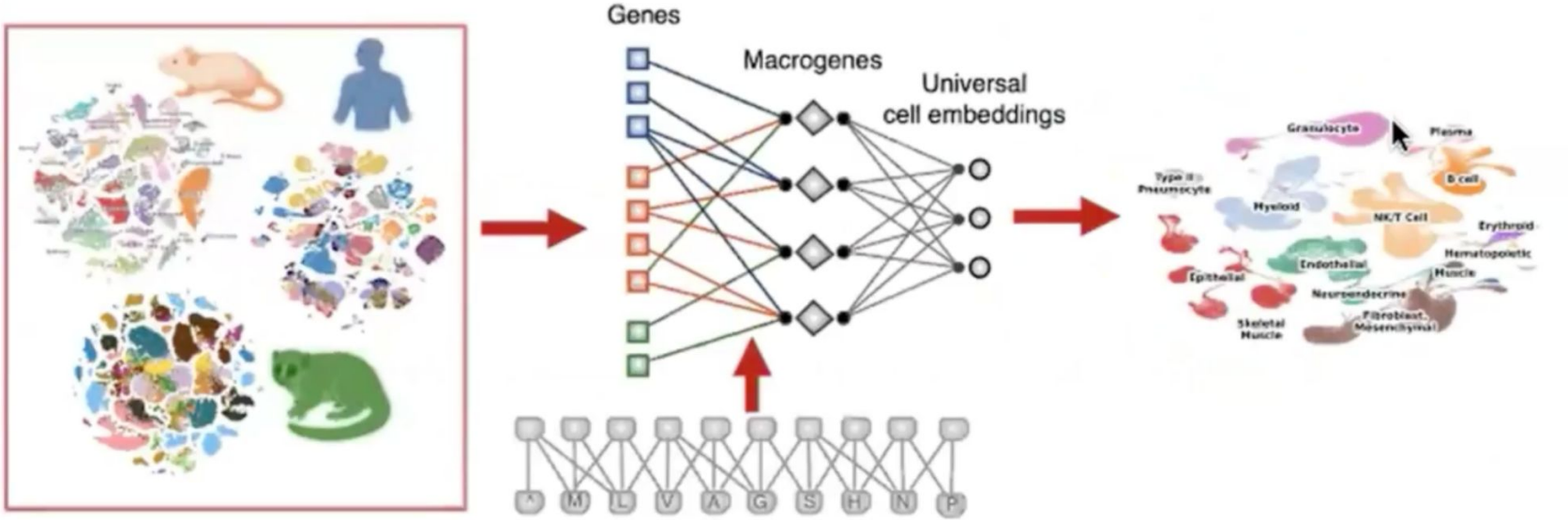
Key Idea: Map diverse sets of genes in the joint space of macrogenes

- Macrogenes: groups of functionally related genes
- Learn macrogene space using protein embeddings from language models



Rosen*, Brbic*, Roohani*, Swanson, Li, Leskovec. *Nature Methods* '24

SATURN: Integrating Datasets across Species

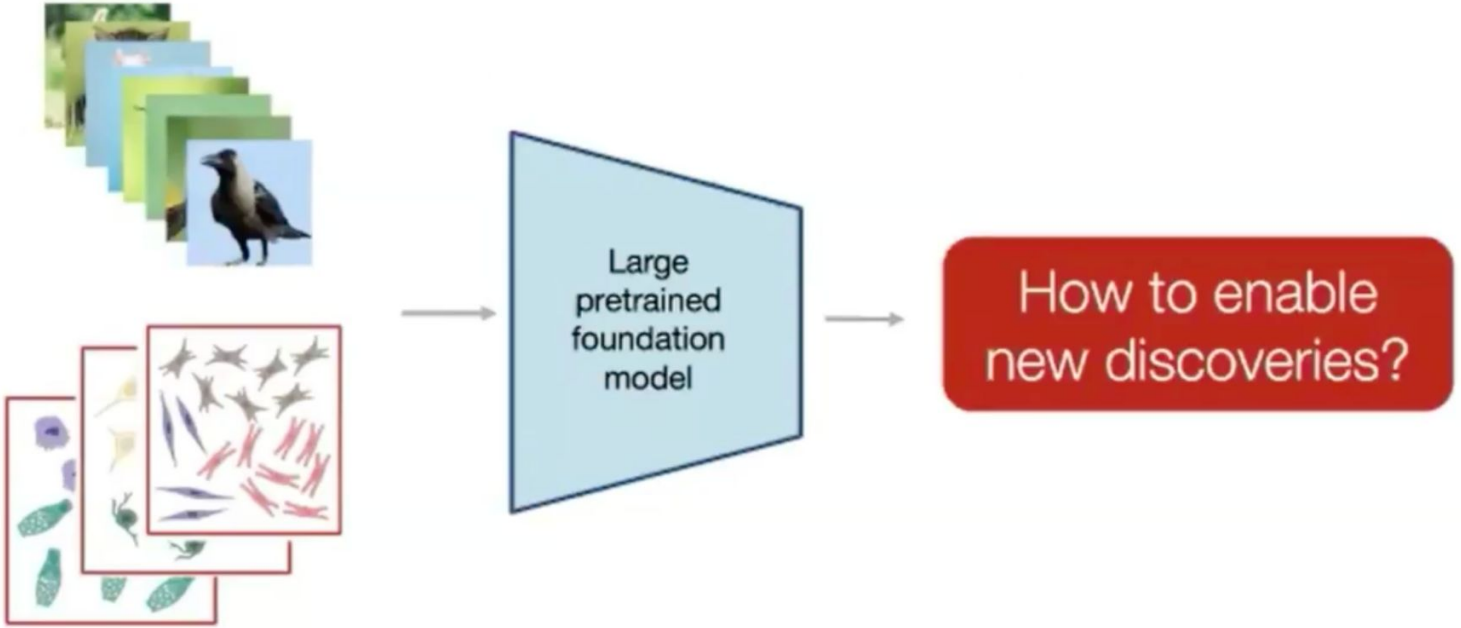


Rosen*, Brbic*, Roohani*, Swanson*, Li, Leskovec *Nature Methods*' 24

On Discovery

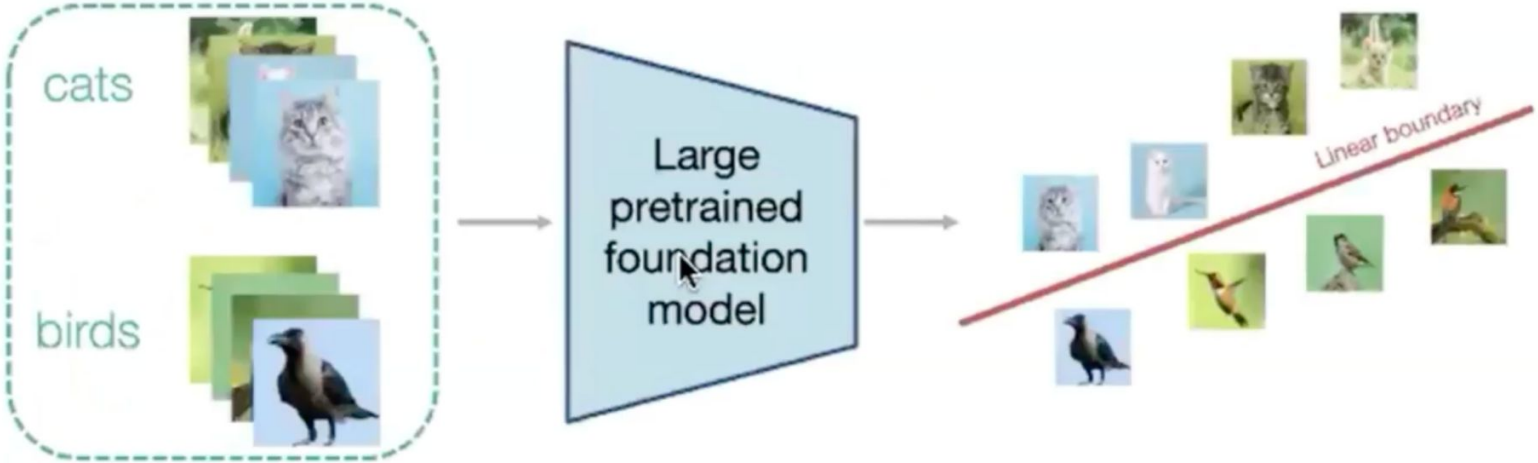
Enabling Discovery from Foundations Models

How To to Enable New Discoveries from Foundation Models?



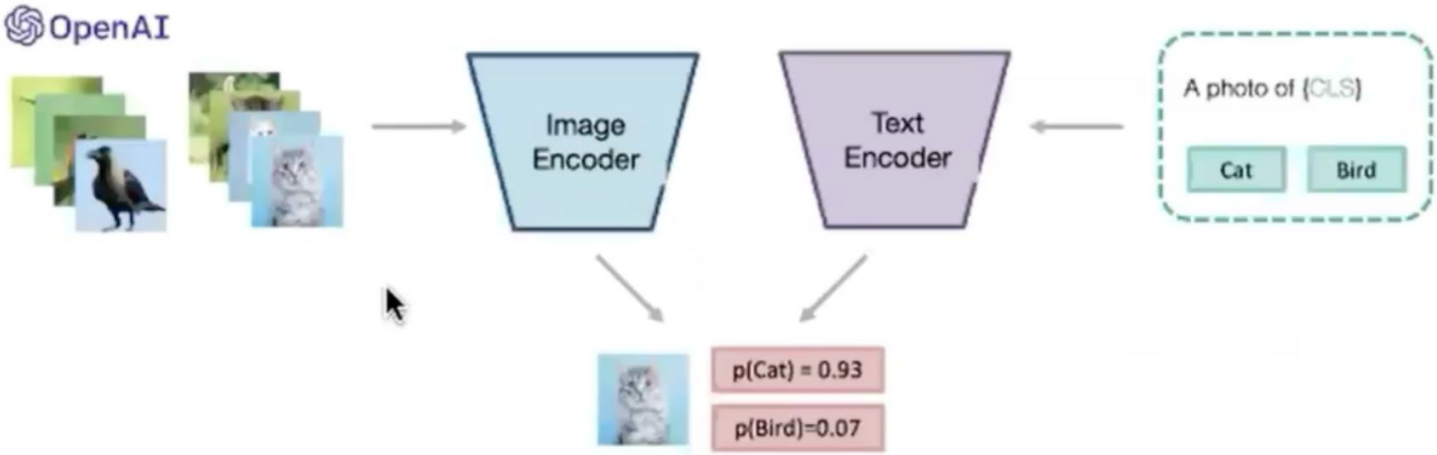
Current Paradigms Still Require Supervision

- Current paradigms:
  1. Fine-tune on the task of interest using labeled data



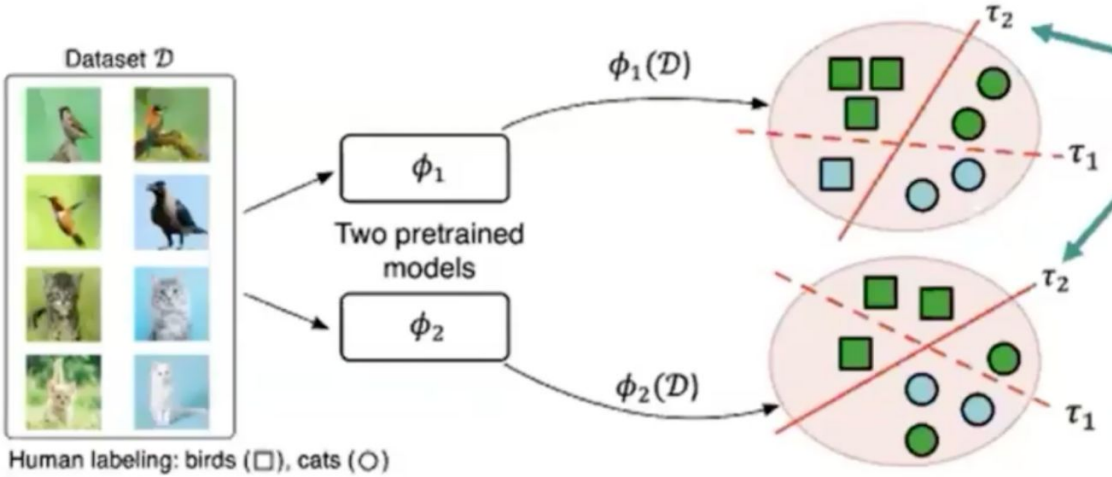
Current Paradigms Still Require Supervision

- Current paradigms:
  2. Zero-shot transfer on the task of interest using instructions



How to Infer Labeling without Any Supervision?

Key idea: Search for a labeling such that linear models will generalize well in different representation spaces



Labeling with a lower generalization error

Gadetsky, Brbić. *NeurIPS* '23 (spotlight)
Gadetsky*, Jiang*, Brbić. *ICML* '24

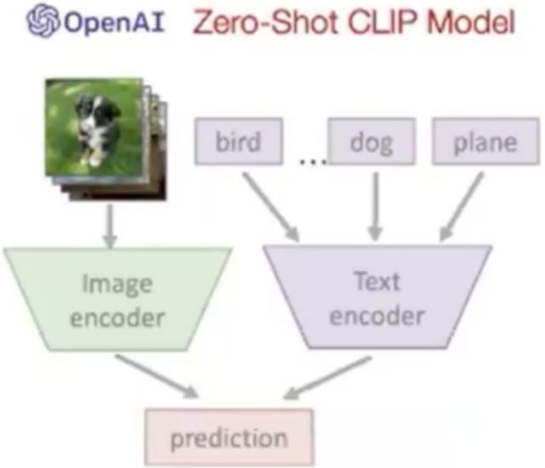
Artyom Gadetsky

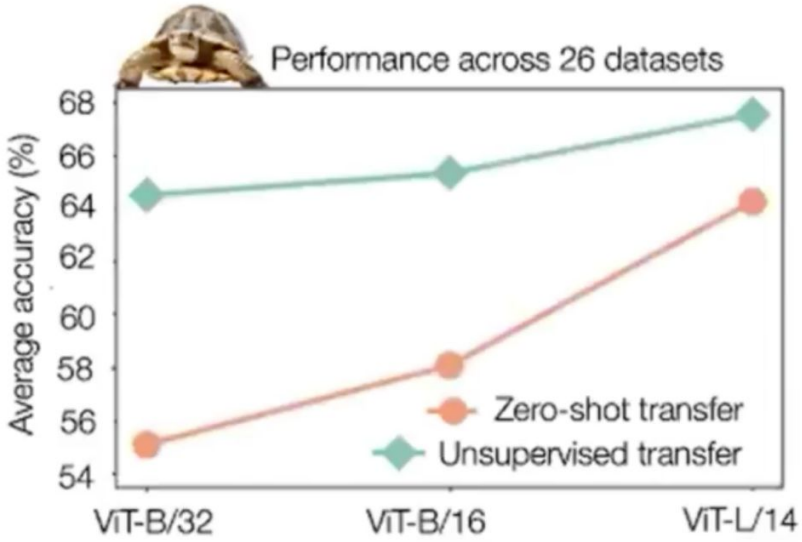
Yulun Jiang

Unsupervised Transfer Outperforms Zero-Shot

SOTA unsupervised performance

- 26 datasets benchmark from CLIP

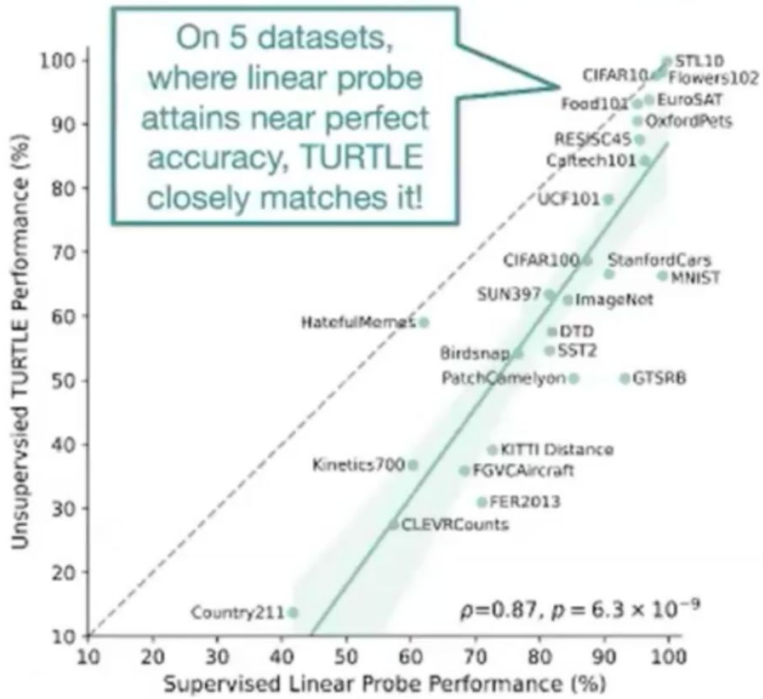




TURTLE is fully unsupervised!

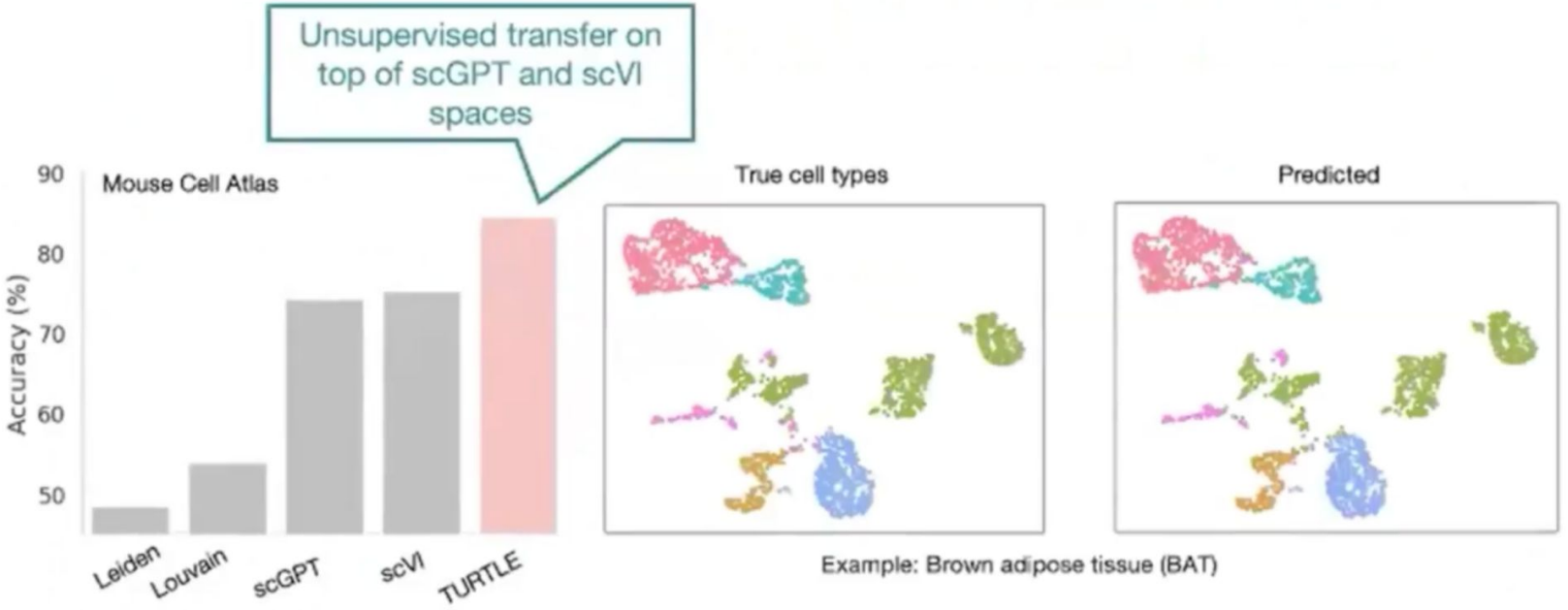
Gadetsky*, Jiang*, Brbić. *ICML '24*

TURTLE's Performance Is Correlated to Linear Probe



TURTLE can infer "optimal" classifier without supervision given high-quality representations

Application to Single-Cell Data



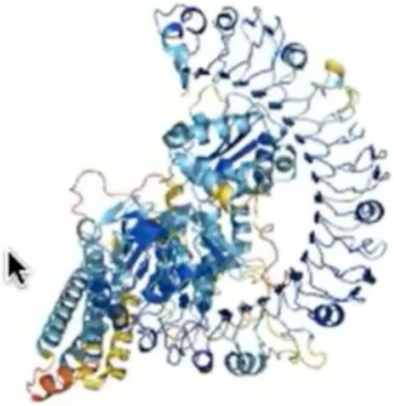
On Multi-modality

Generating Complex Tissue Structures from Gene Expressions

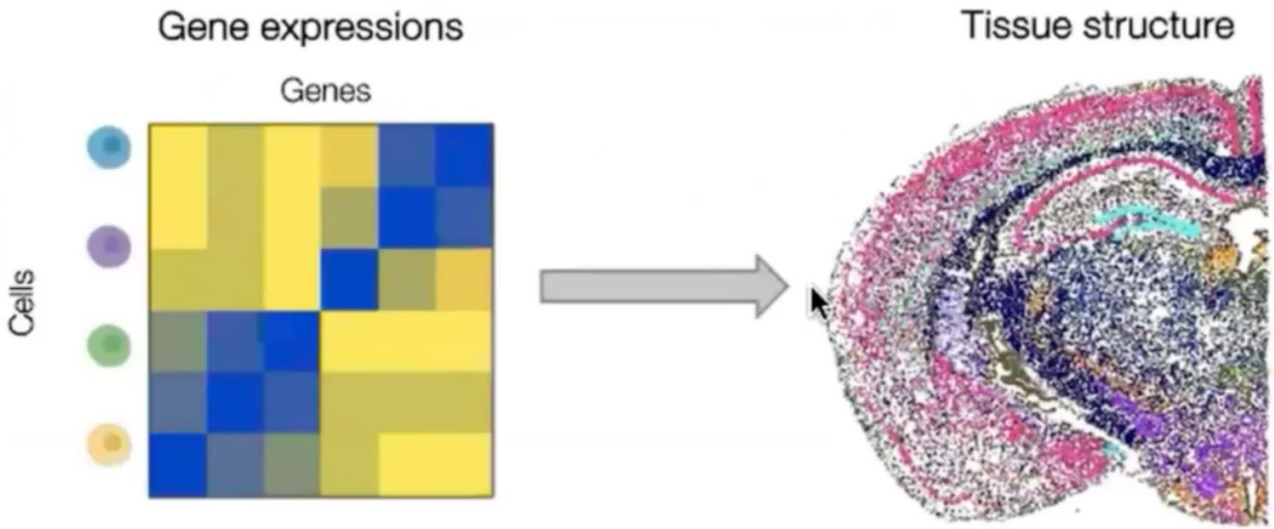
AlphaFold For Cells

Protein structure

Protein sequence



AlphaFold For Cells

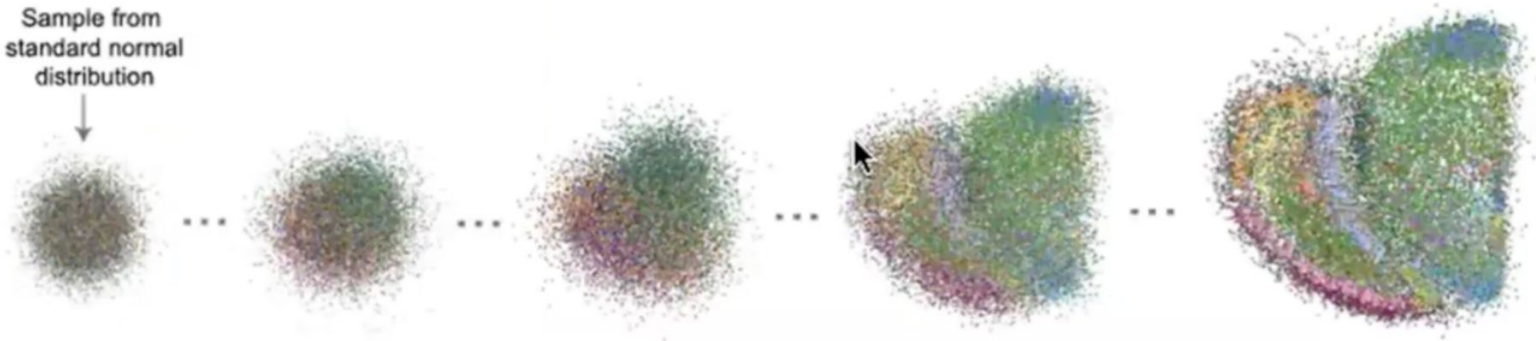


LUNA: From Cells to Locations

A generative model for mapping cells to their locations and generating tissue structures

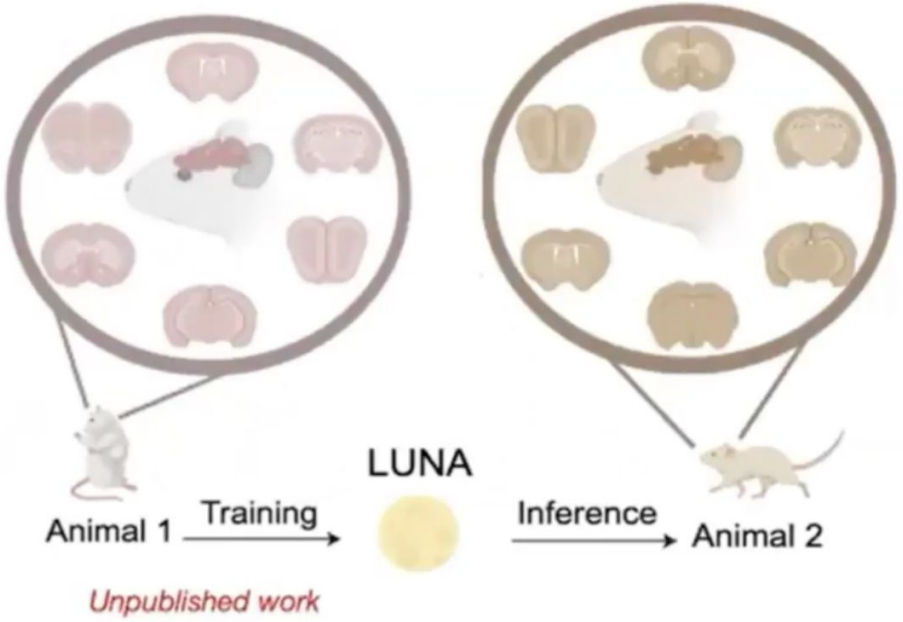
Yist Yu

Chanakya Ekbote



Unpublished work

Reconstruction of Whole Mouse Brain MERFISH Atlas



Dataset:

- MERFISH Mouse Brain Atlas with over 4 million cells

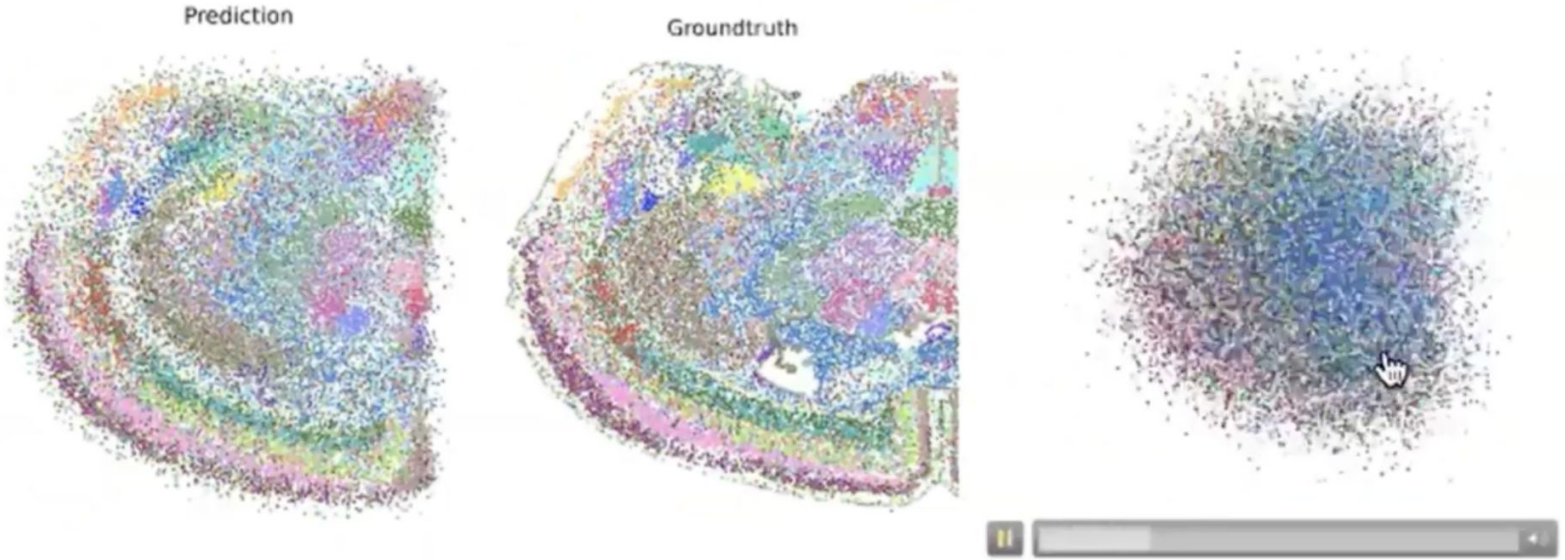
Training dataset:

- 2.85 million cells across 147 slices from one mouse

Target unlabeled dataset:

- 1.23 million cells across 66 slices from another mouse

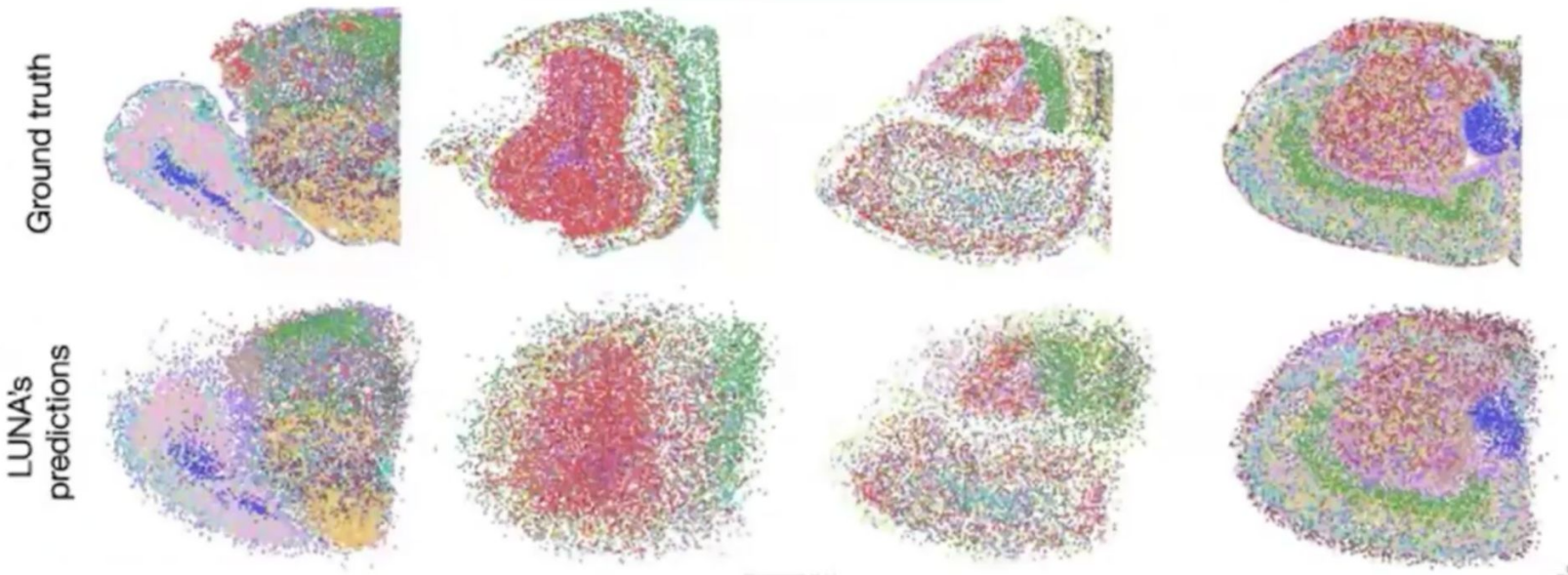
Reconstruction of Whole Mouse Brain MERFISH Atlas



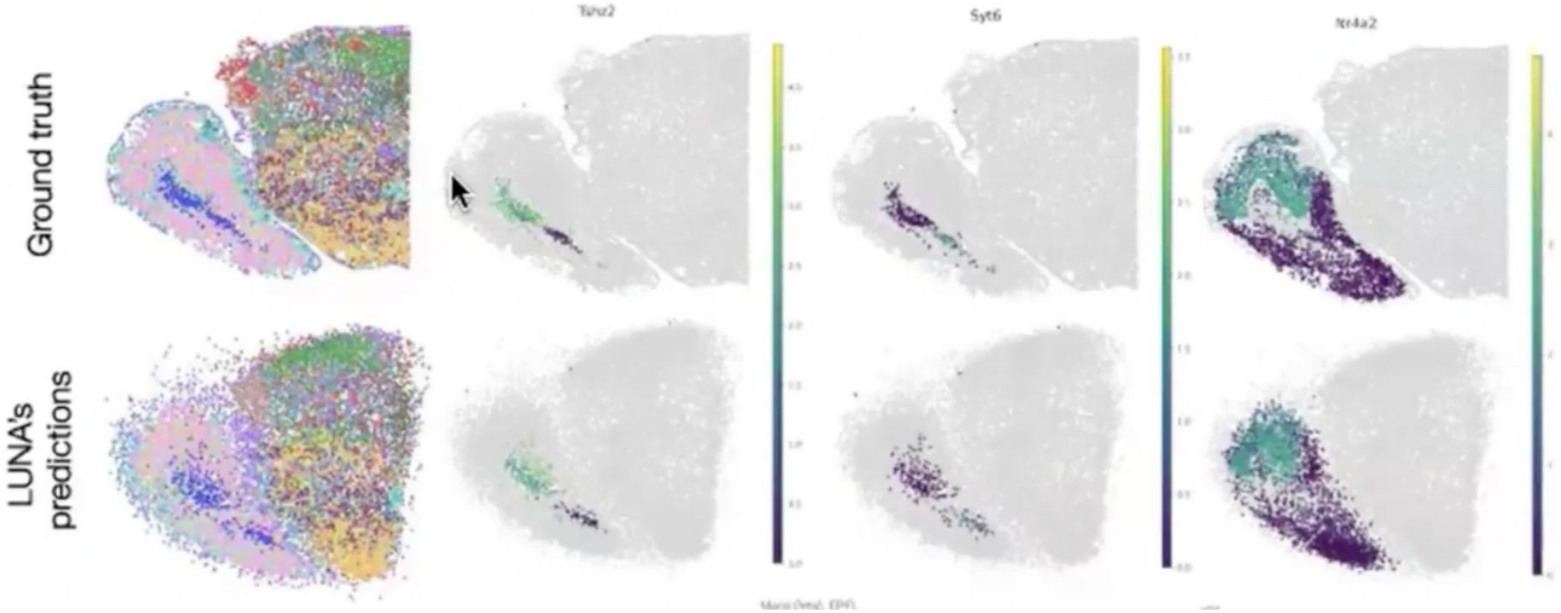
Unpublished work

Reconstruction of Whole Mouse Brain MERFISH Atlas

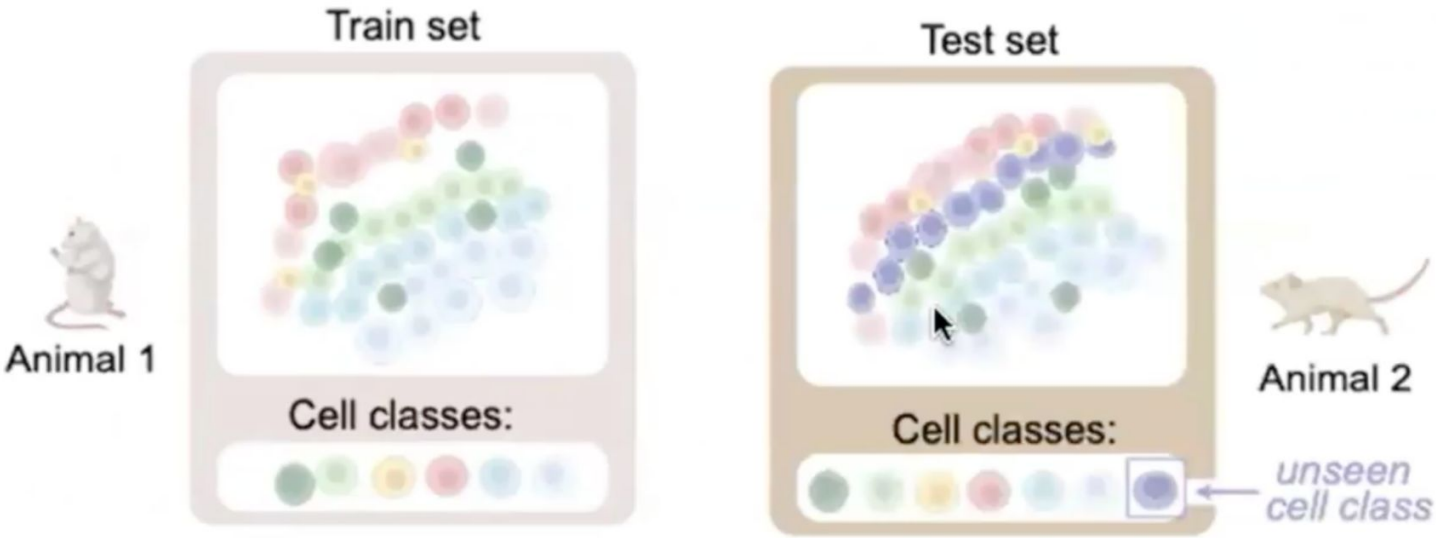
338 different subclasses!



Reconstruction of Whole Mouse Brain MERFISH Atlas

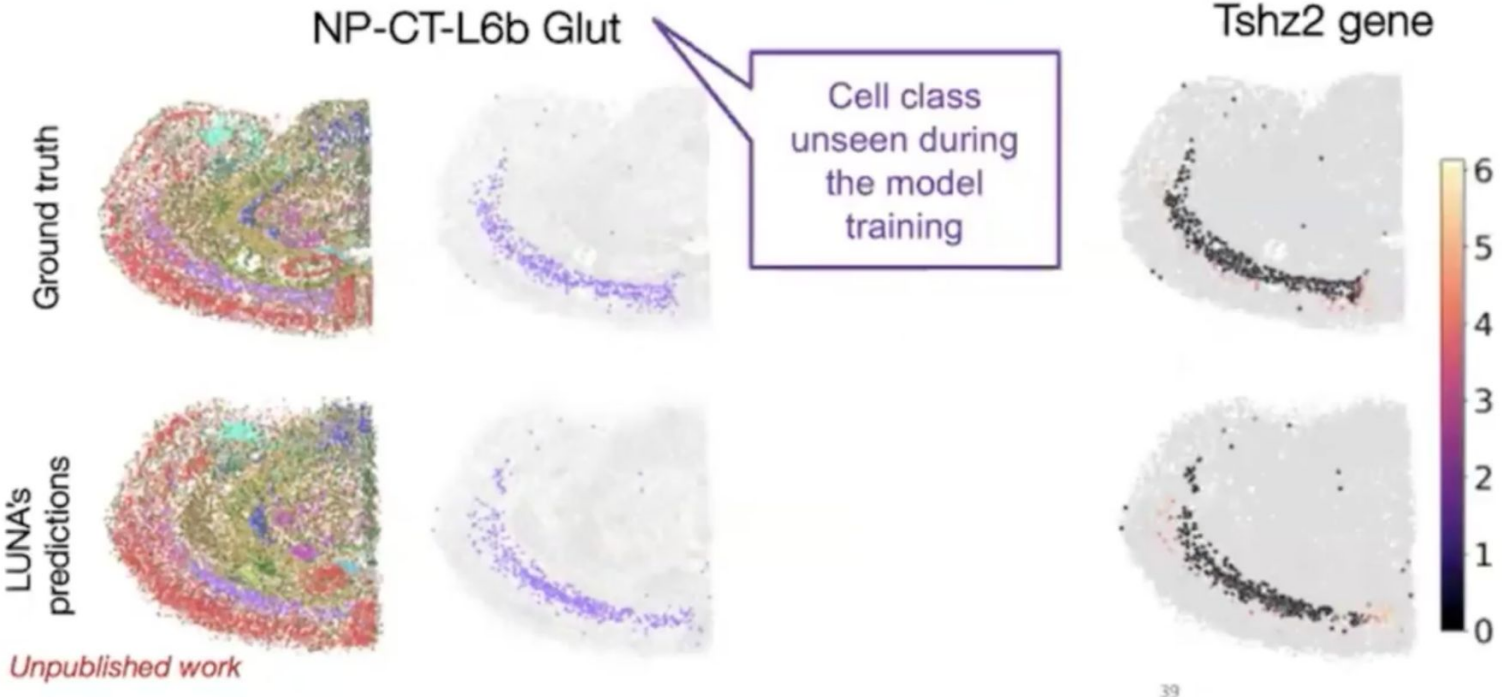


LUNA: Zero-Shot Setting



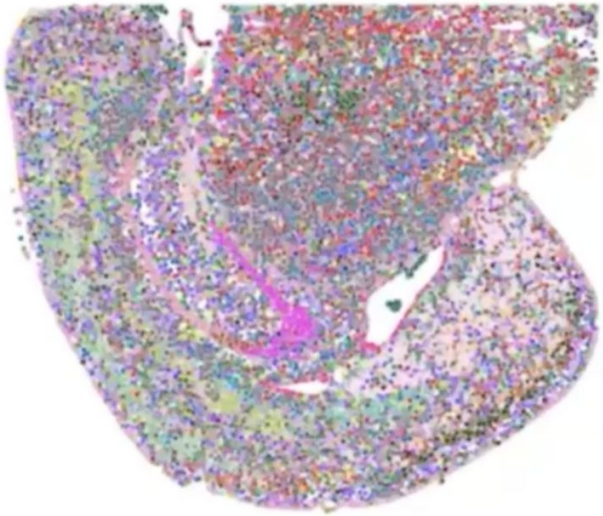
Unpublished work

Zero-Shot Generalization to Unseen Cell Types

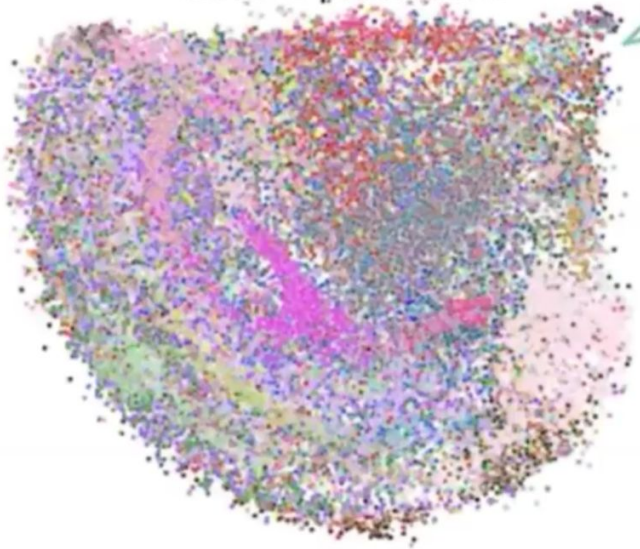


De Novo Reconstruction of CNS ScRNA-seq Atlas

Estimated ground truth



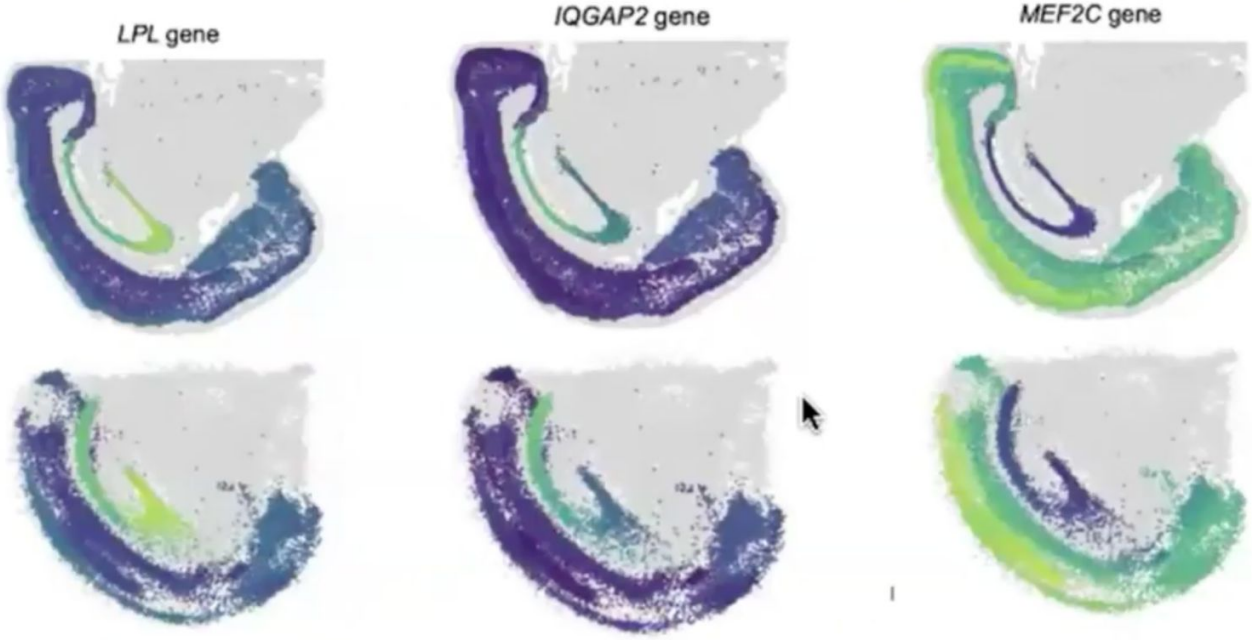
LUNA's predictions



216 cell classes

Unpublished work

De Novo Reconstruction of CNS ScRNA-seq Atlas



Unpublished work

Acknowledgements

PhD students
Artyom Gadetsky
Shuo Wen
Tingyang Yu
Yulun Jiang
Siba Panigrahi

Postdocs:
Ramon Vinās Tórne
Myeongho Jeon

Research engineers
Jeremy Goumaz

Assistant:
Marie Künzle

Machine Learning

MLBio lab@EPFL

Biomedicine

Collaborators:
Jure Leskovec, Stanford
Yanay Rosen, Stanford
Yusuf Roohani, Stanford
Kaidi Cao, Stanford
John Hickey, Duke
Hongjie Li, Baylor College of Medicine
Yuqi Tan, Stanford
Liqun Luo, Stanford
Michael Snyder, Stanford
Garry Nolan, Stanford
Pascal Frossard, EPFL
Chanakya Ekbote, MIT

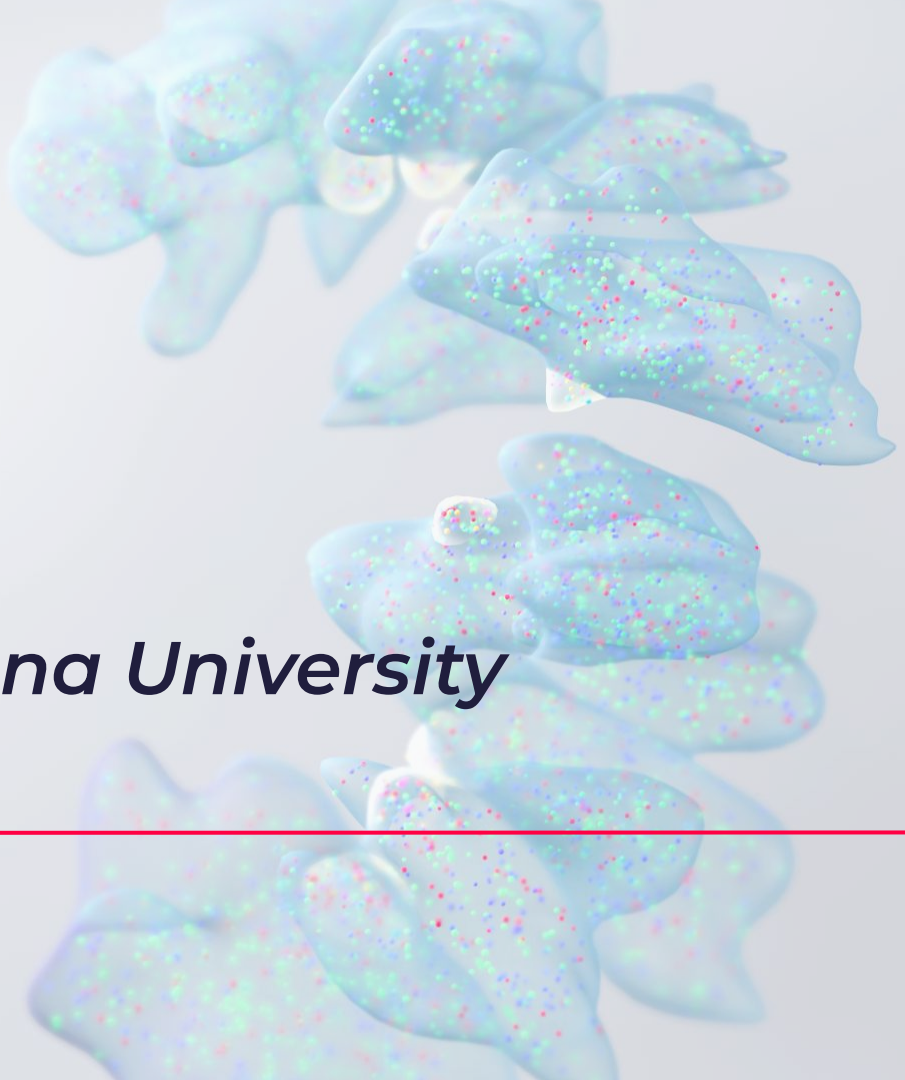
Swiss National Science Foundation

NIH ImmGen

Generative AI for Decoding Single-Cell Complexity

Maria Brbić

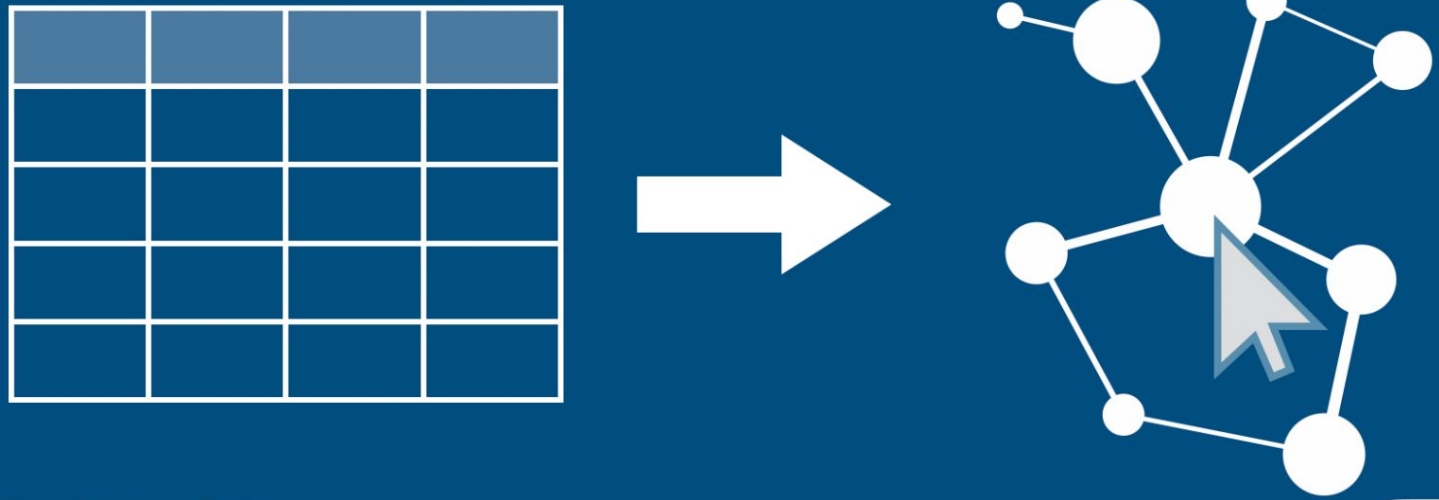
EPFL



Filipi Silva, *Indiana University*

Charting Complexity: Interactive real-time visualizations of large-scale networks and embeddings with Helios-Web

https://www.icloud.com/keynote/018HYK4JjBOCRA8N5NGjU7dVg#Embeddings_Visualization_reduced

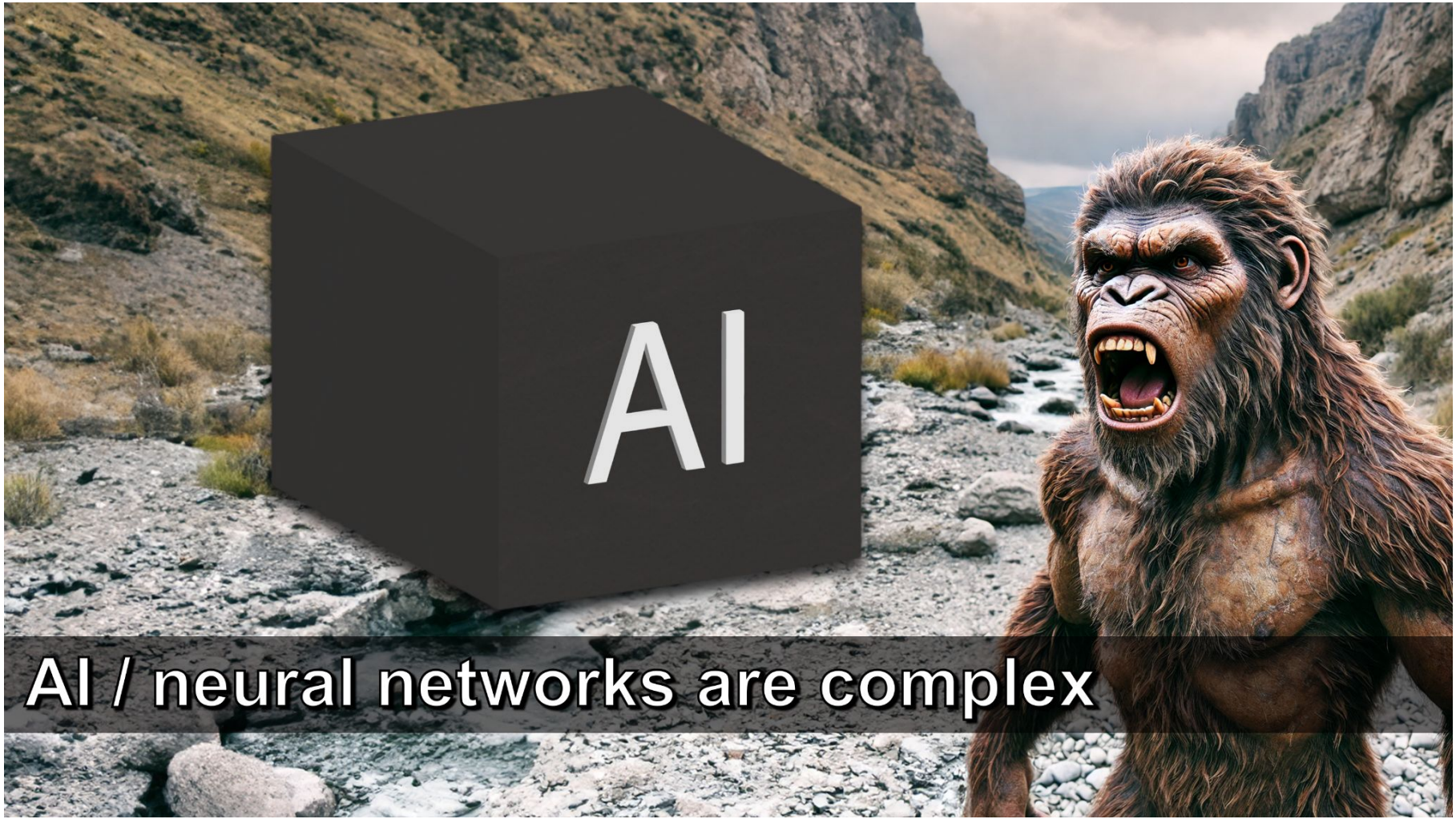


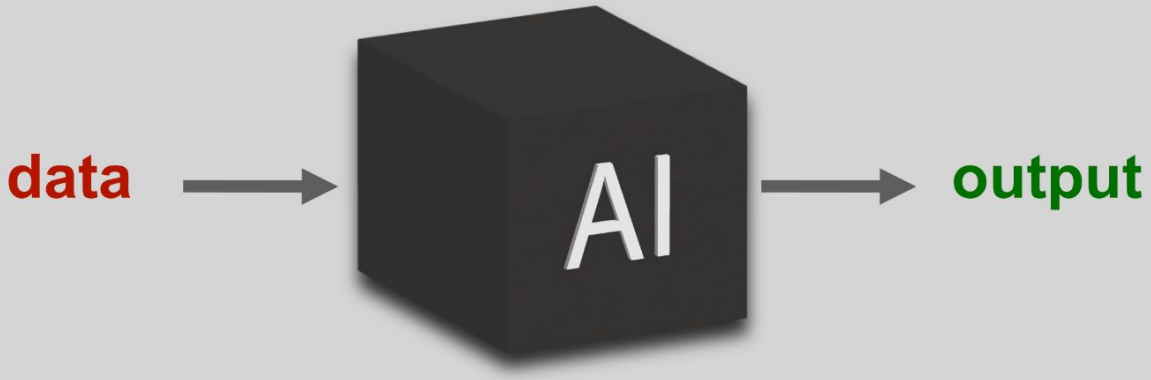
Filipi N. Silva
research scientist - Indiana University
filipinascimento.github.io • filsilva@iu.edu

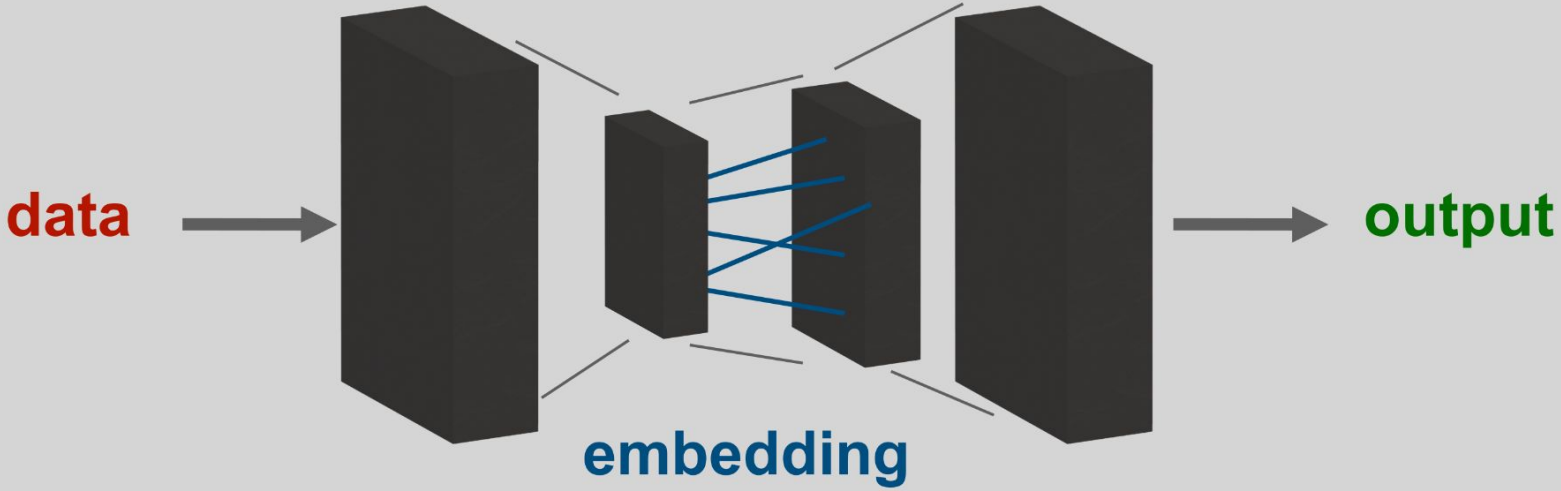
LUDDY

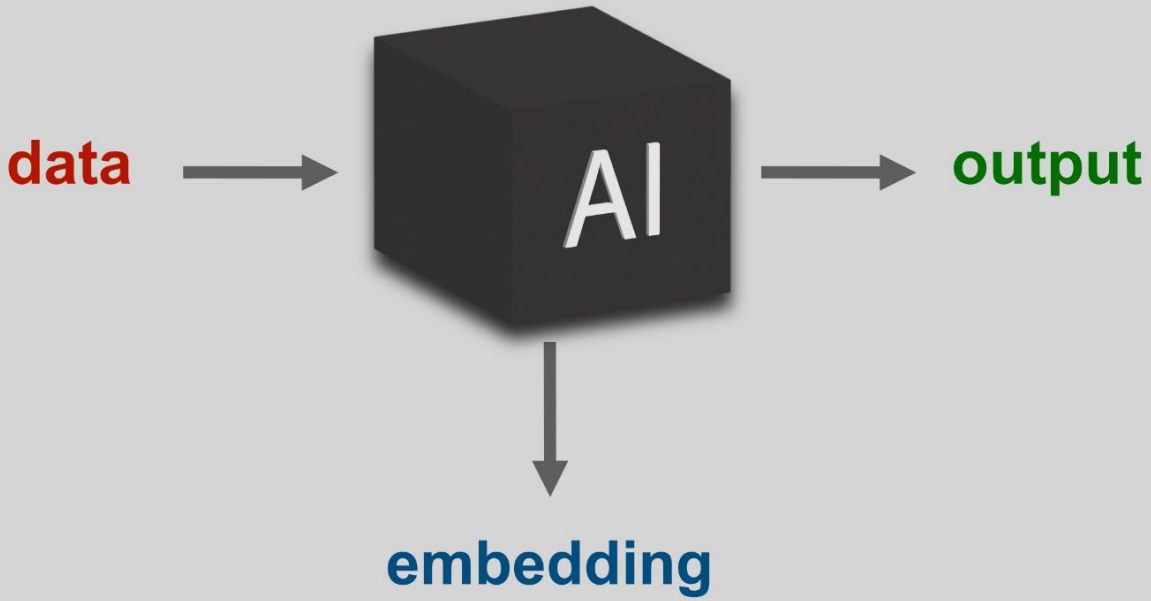
OSoMe

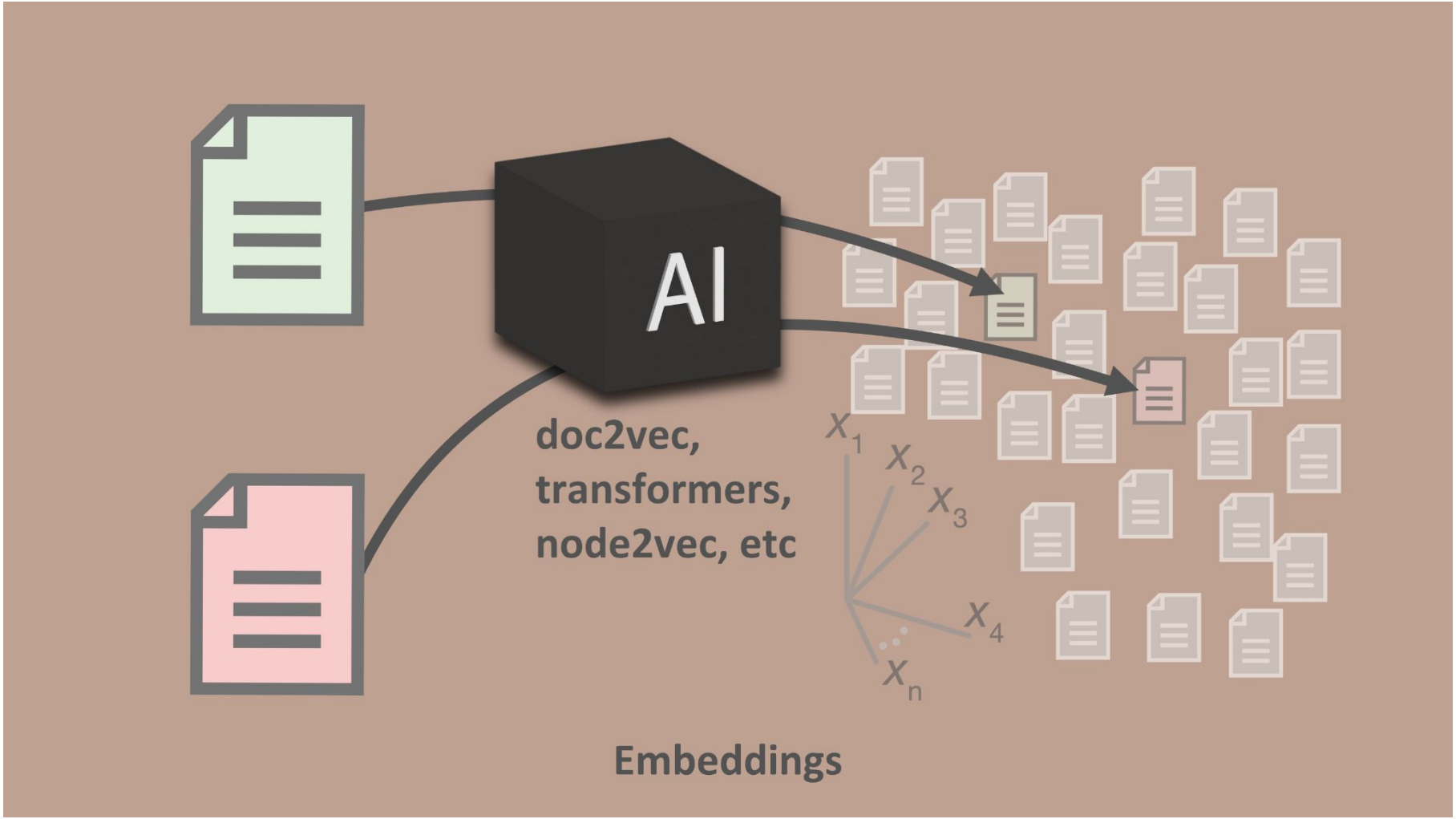
web Helios
heliosweb.io

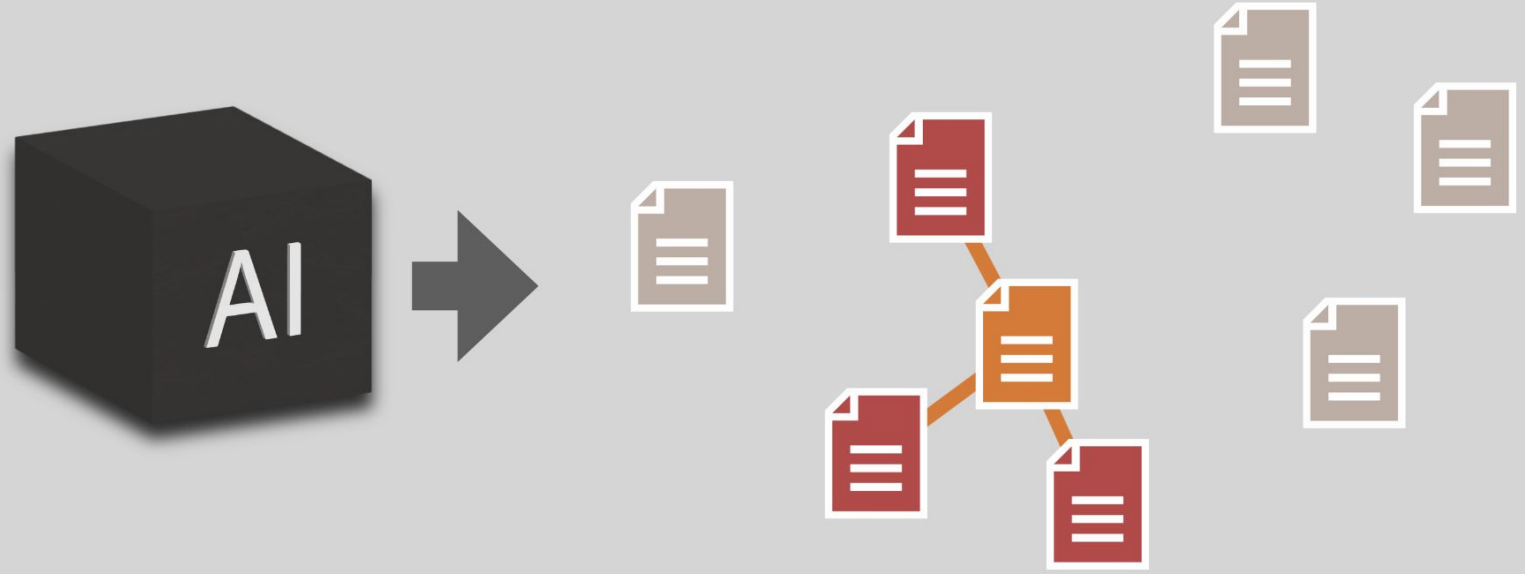

AI
AI / neural networks are complex


data
AI
output


data
output
embedding


data
AI
output
embedding


AI
doc2vec,
transformers,
node2vec, etc
X_1
X_2
X_3
X_4
X_n
Embeddings



Recommendation systems

Databases/Search engine

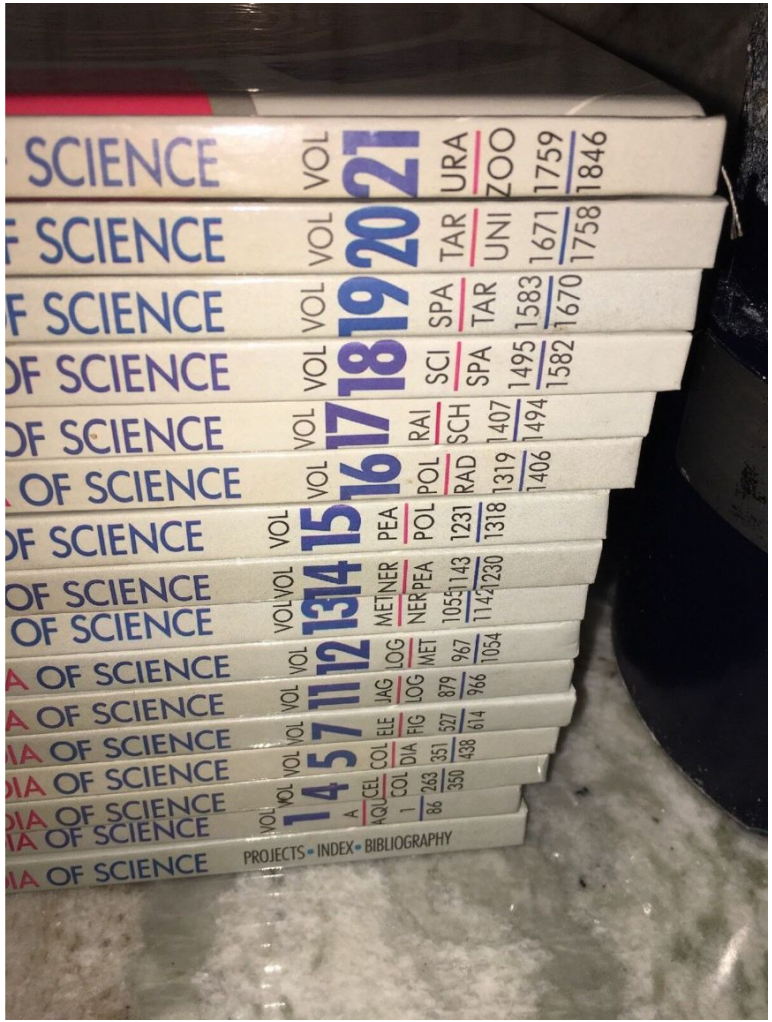
Retrieval-Augmented Generation (RAG)

Anomaly Detection

...


Data is also complex

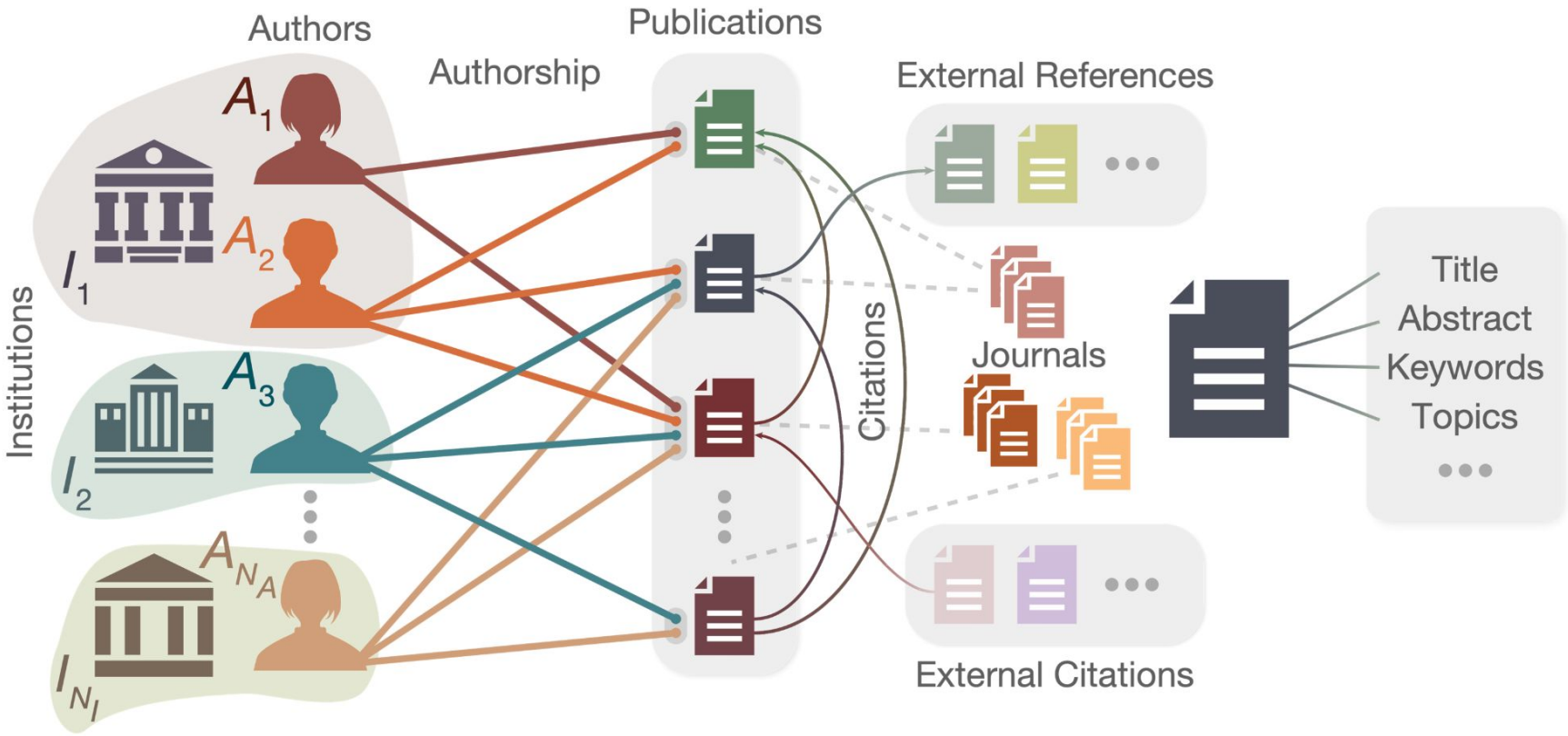
by Chico Camargo (Twitter)
https://twitter.com/evoluchico/status/1388137531552718860

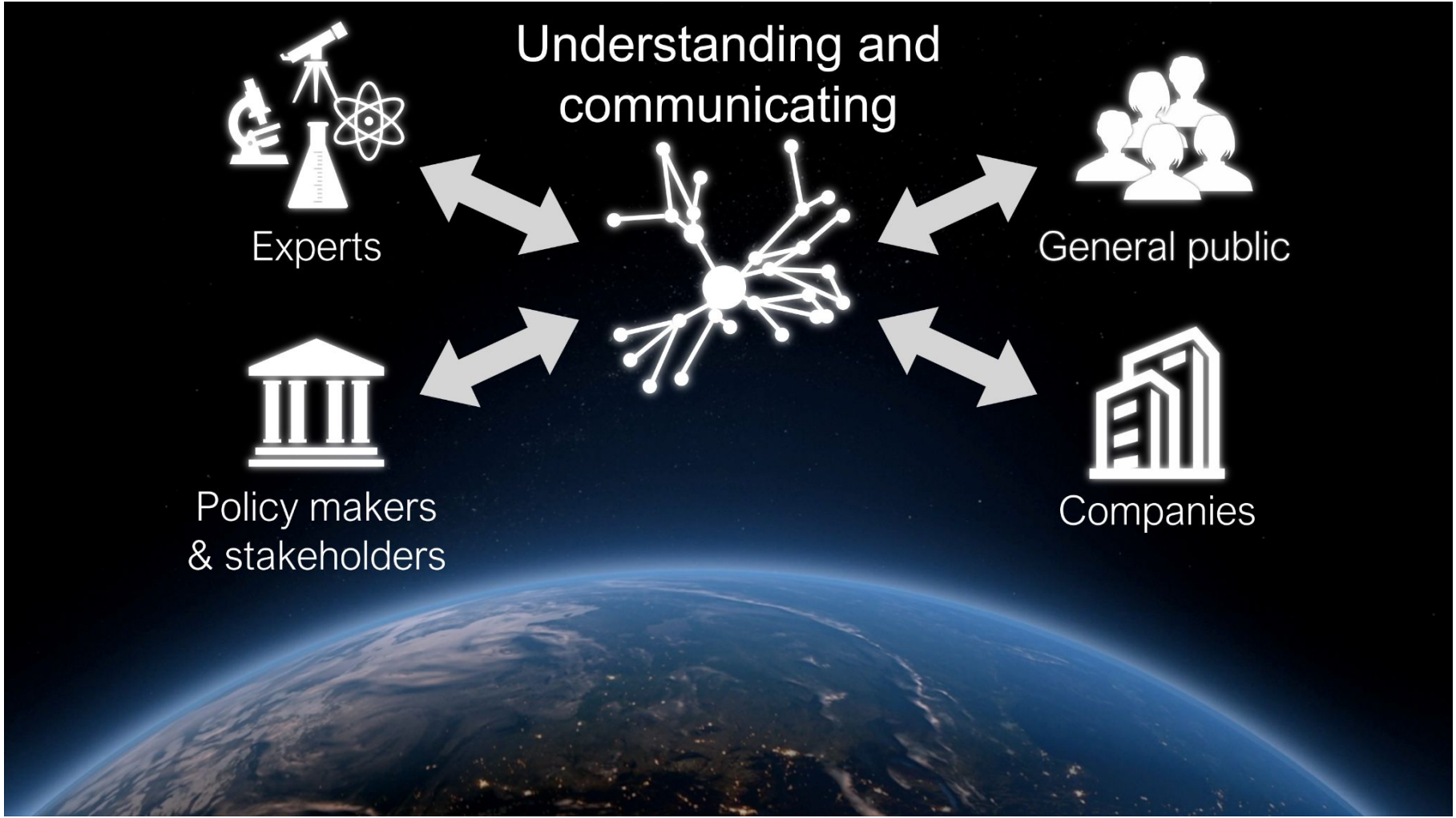


Science of Science

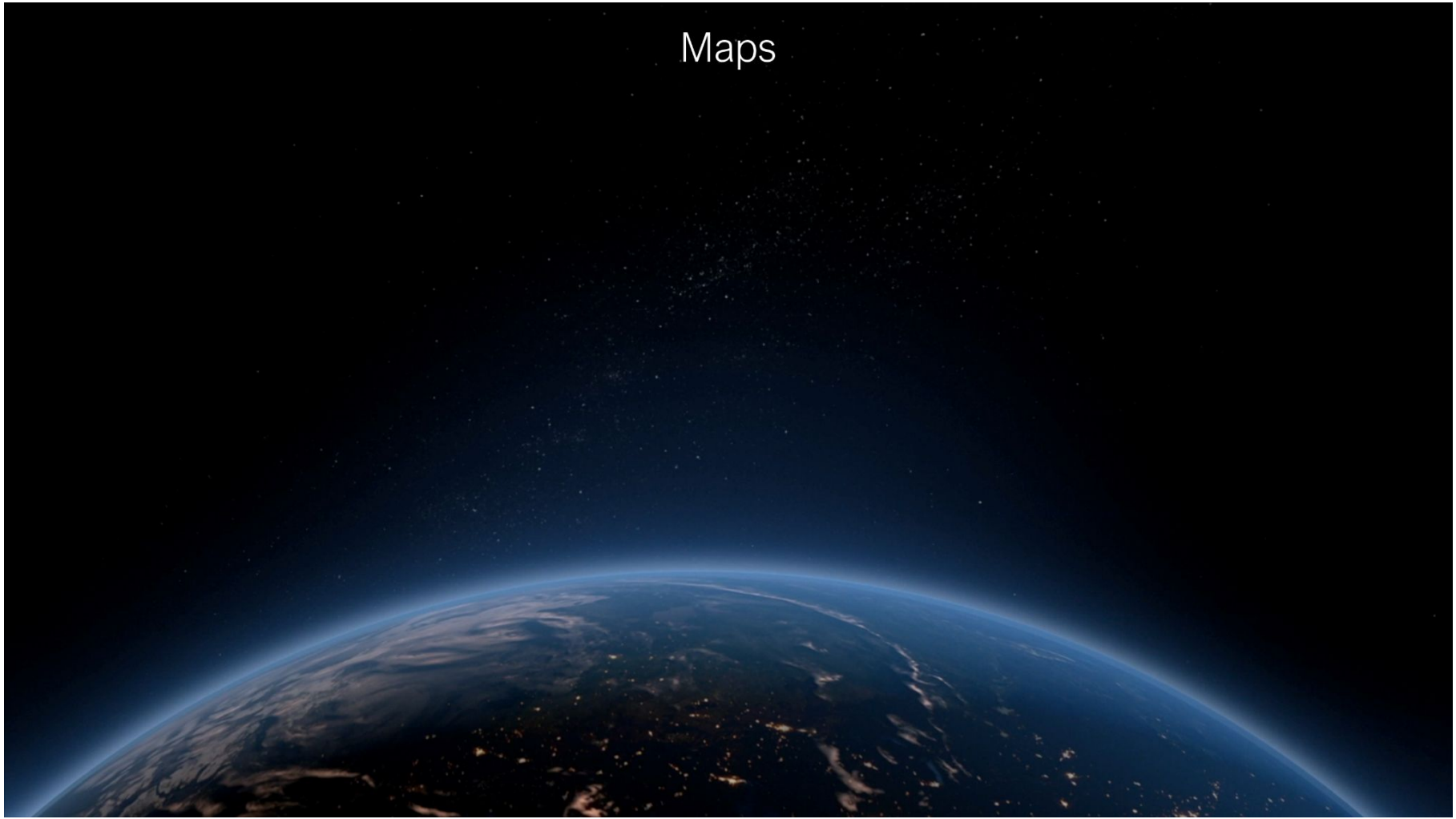
- How science is evolving?
- How researcher teams are formed?
- Is science becoming more interdisciplinary?
- Can we predict success in science?
- How to properly evaluate researchers? journals? papers?
- Can tools/approaches accelerate the scientific development?
- Can we predict the benefits of implementing a policy?
- ...

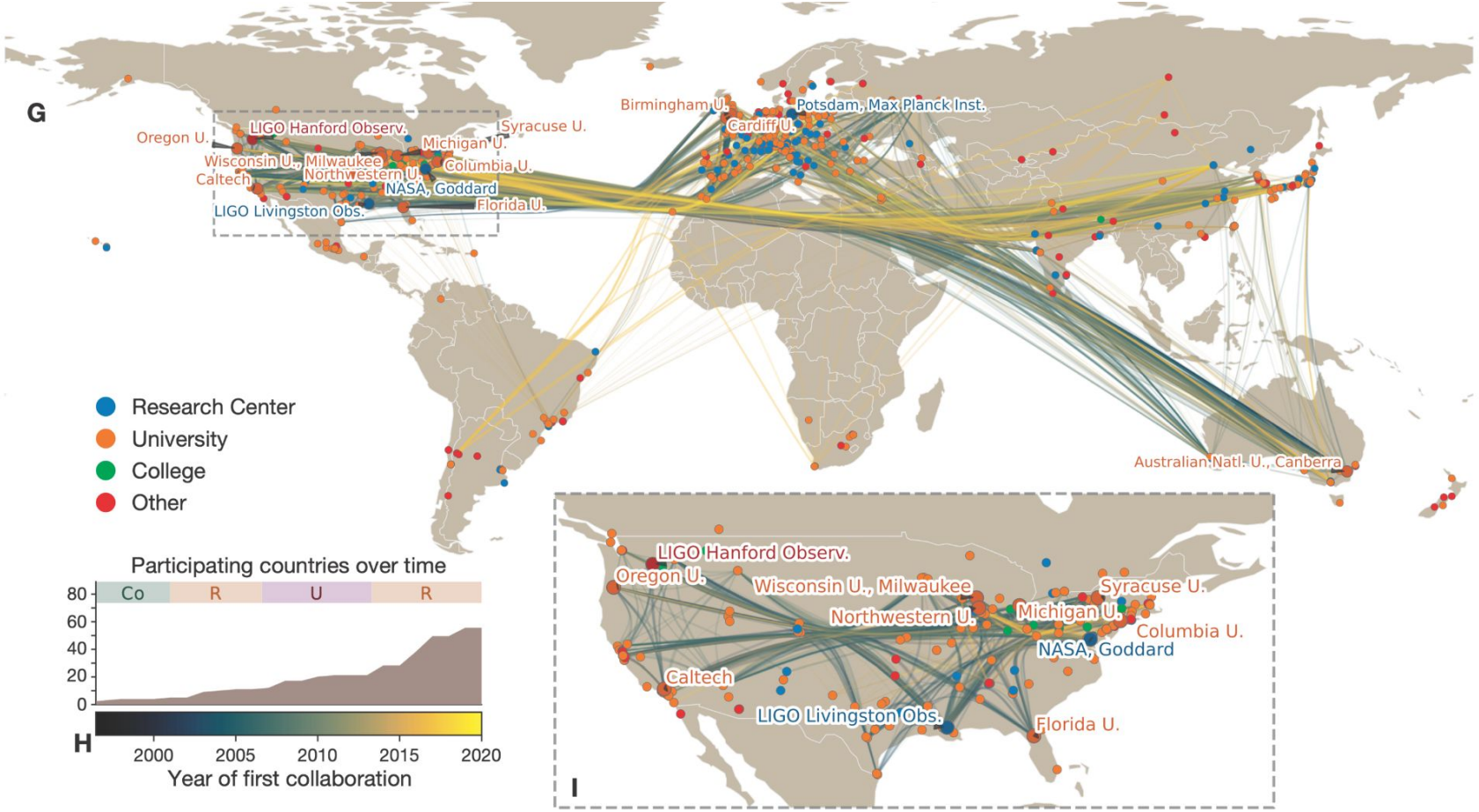
Publications


Authors
Authorship
Publications
External References
Institutions
A_1
A_2
A_3
A_{N_A}
I_1
I_2
I_{N_I}
Citations
Journals
Title
Abstract
Keywords
Topics
External Citations


Understanding and communicating
Experts
General public
Policy makers & stakeholders
Companies

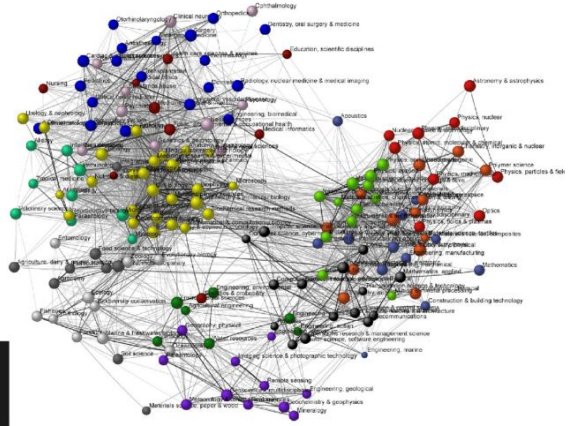
Maps



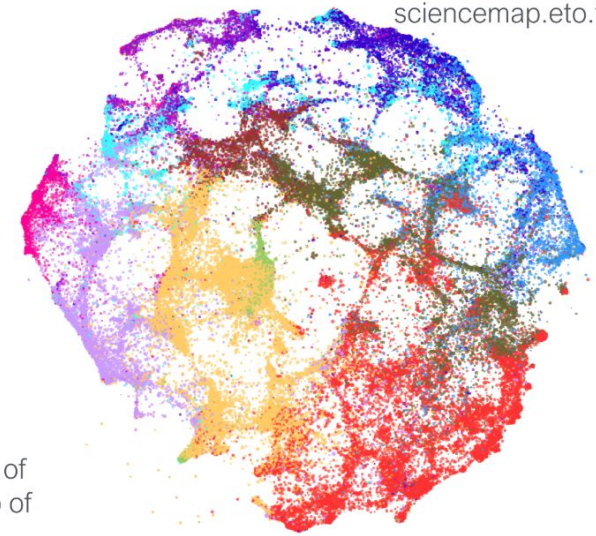


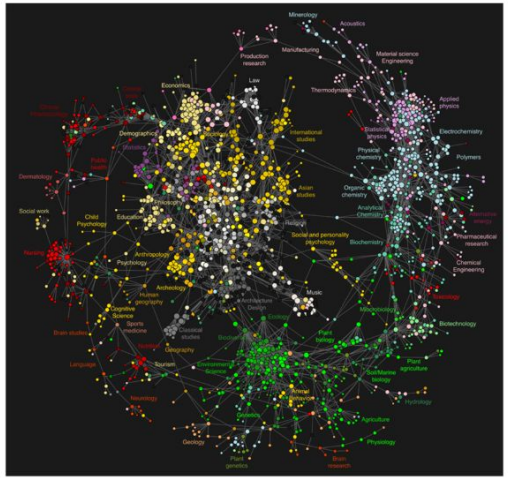
Börner, K., Silva, F. N., & Milojević, S. (2021).
Visualizing big science projects. Nature Reviews Physics, 3(11), 753-761.

Leydesdorff, Loet, and Ismael Rafols. "A global map of science based on the ISI subject categories." Journal of the American Society for Information Science and Technology 60.2 (2009): 348-362.



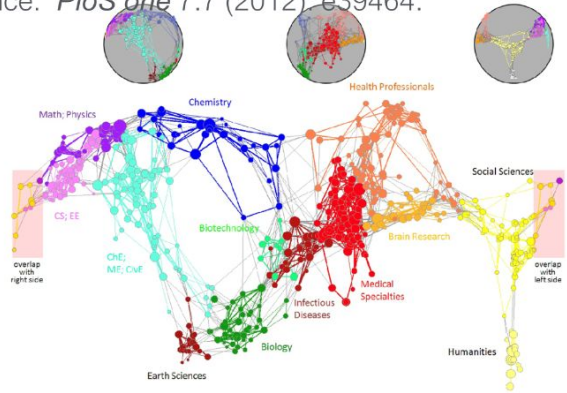
ETO Map of Science
sciencemap.eto.tech

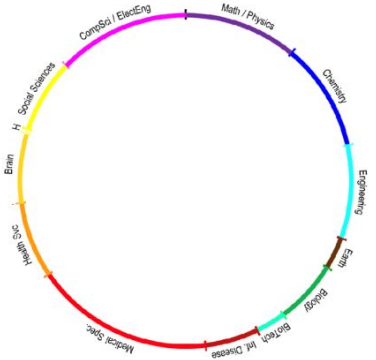


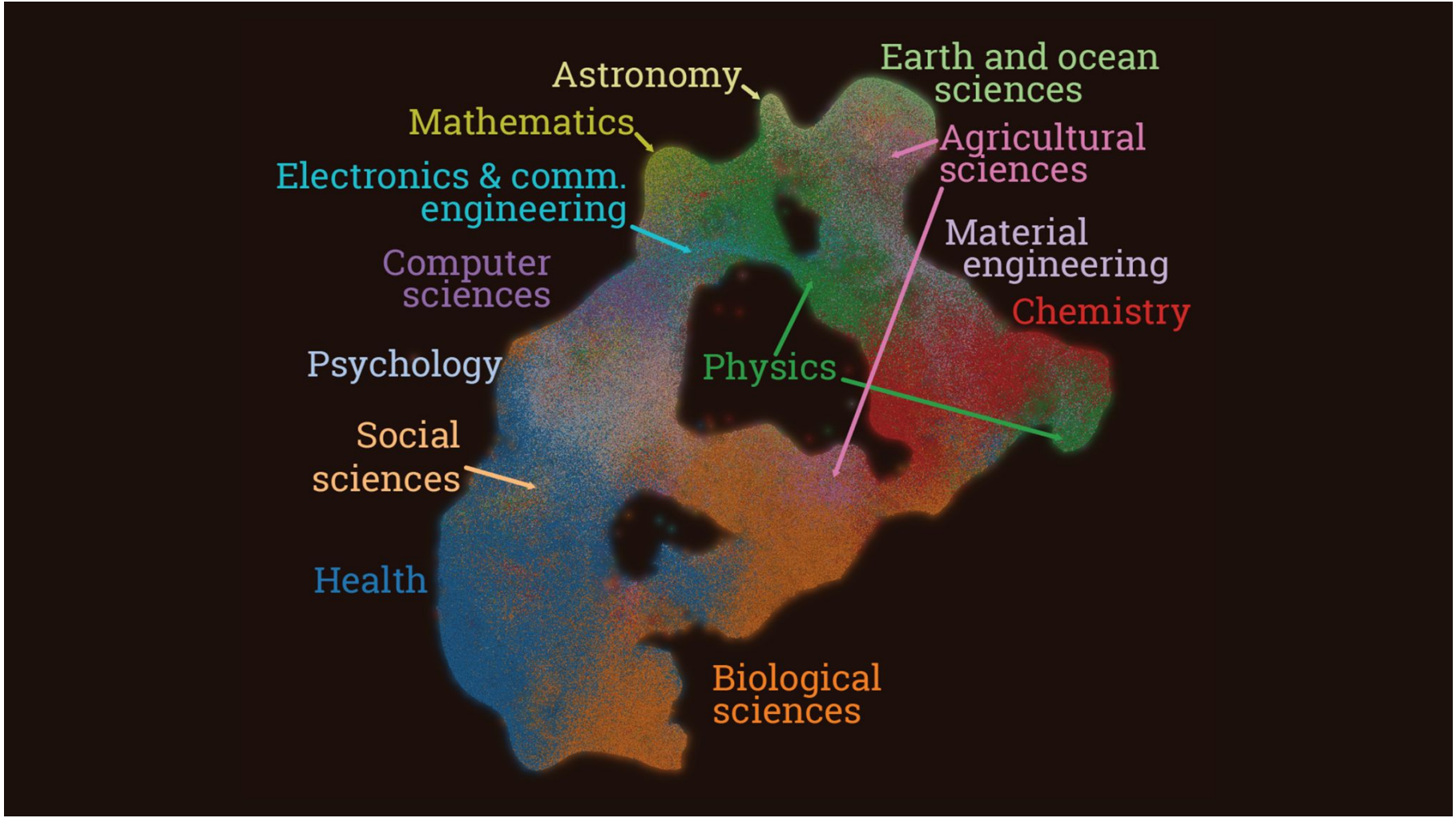


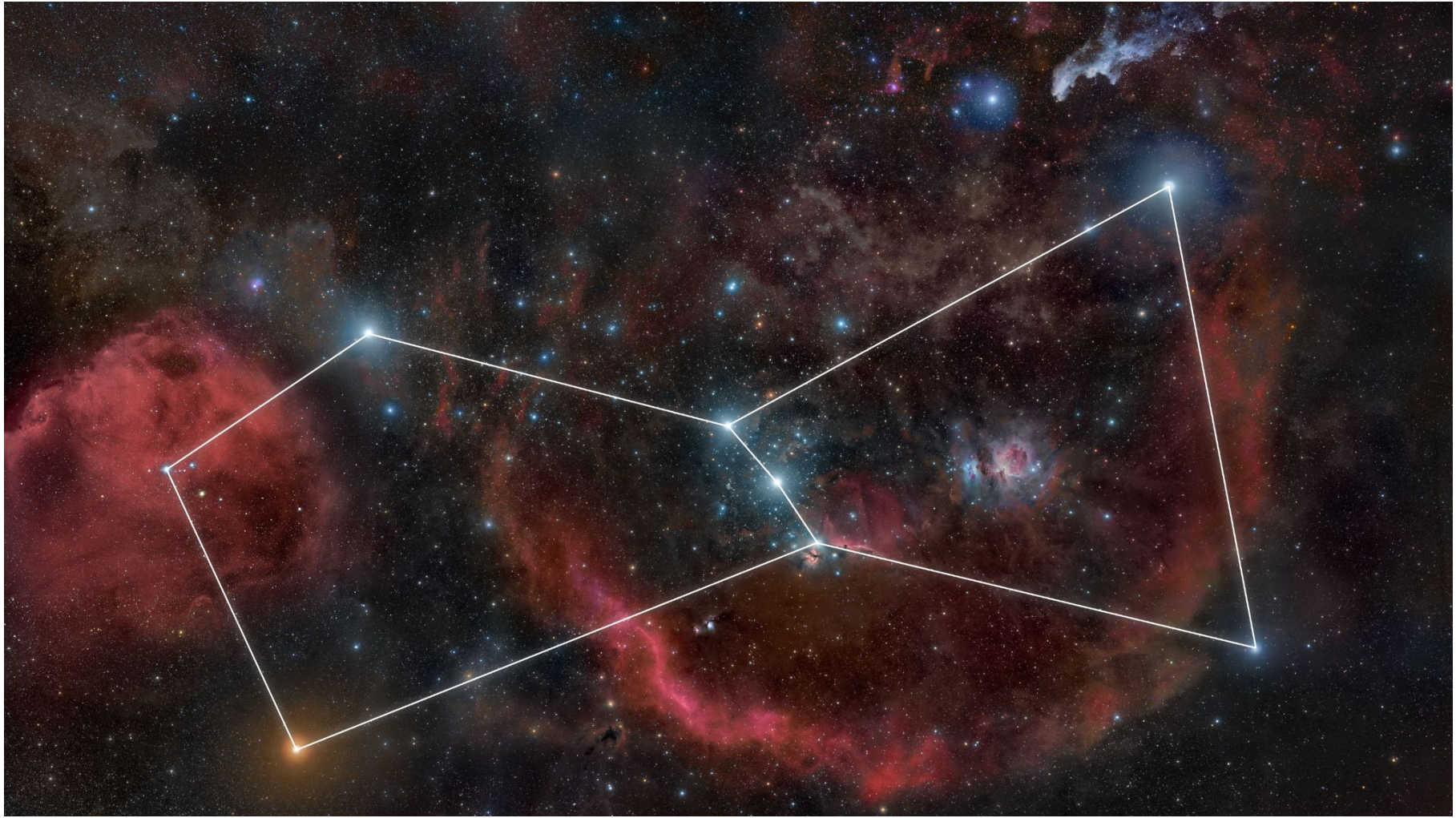
Bollen, Johan, et al. "Clickstream data yields high-resolution maps of science." PloS one 4.3 (2009): e4803.

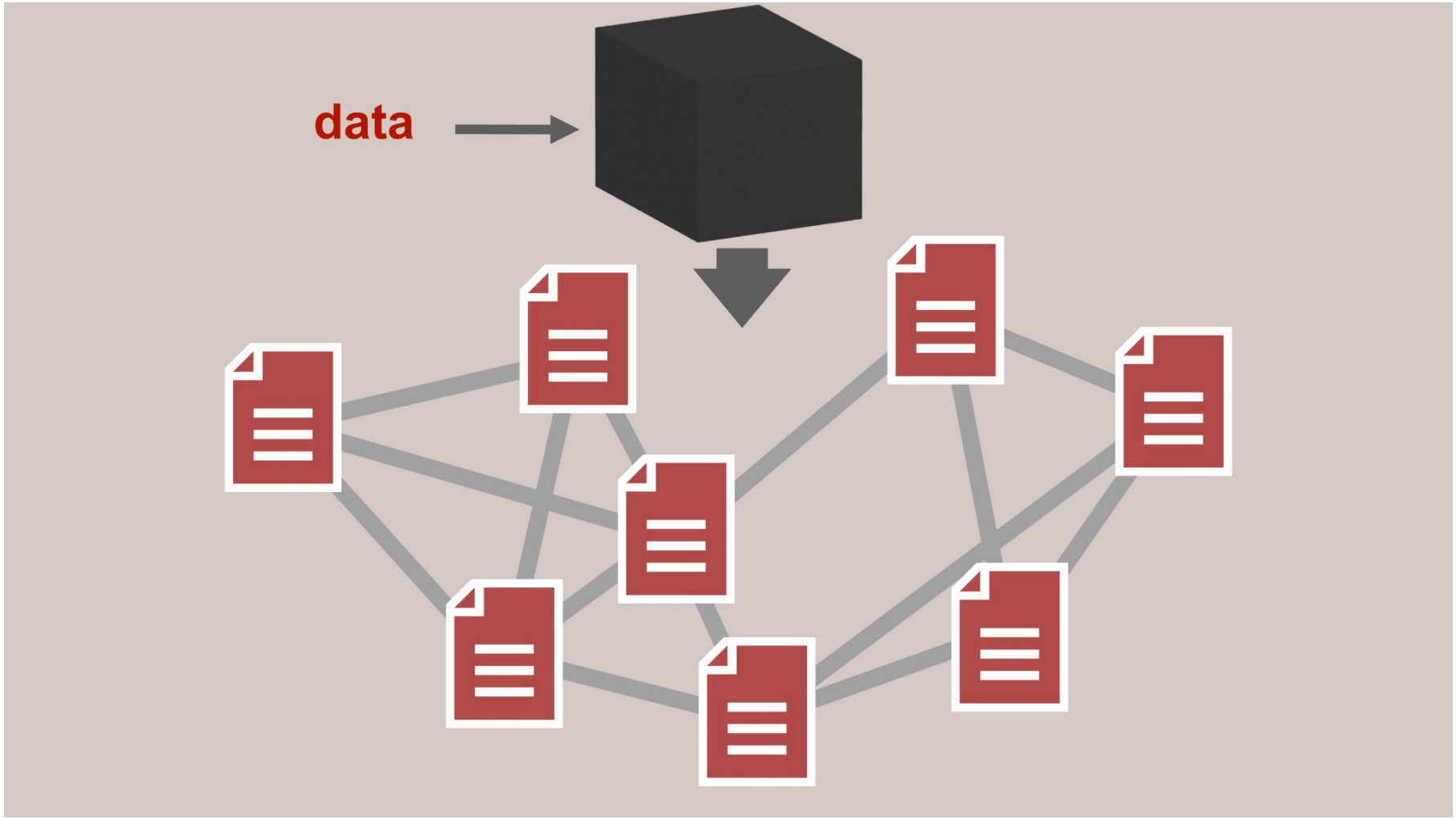
Börner, Katy, et al. "Design and update of a classification system: The UCSD map of science." *PloS one* 7.7 (2012): e39464.

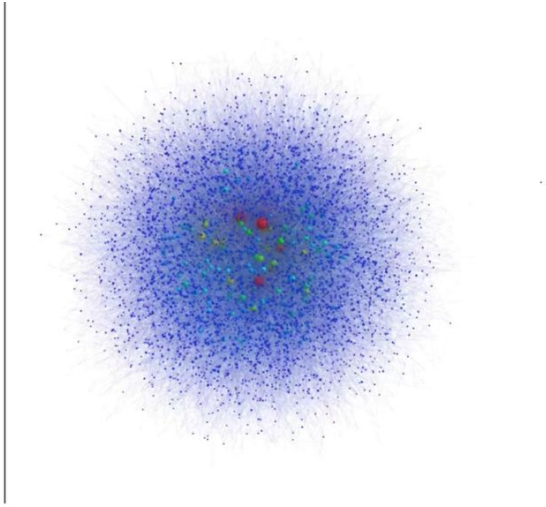
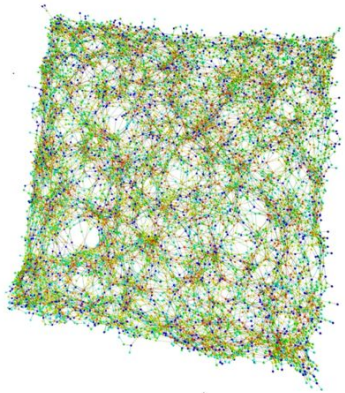
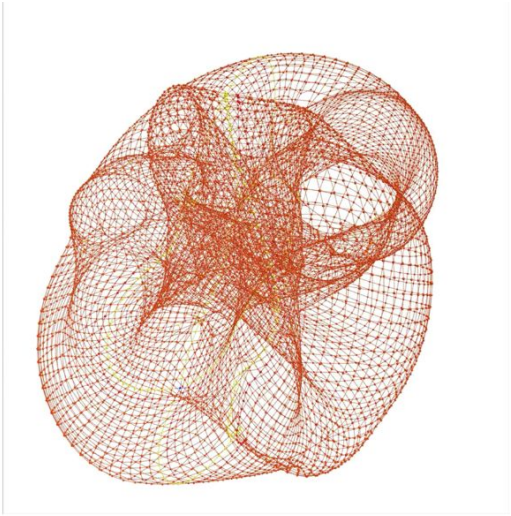
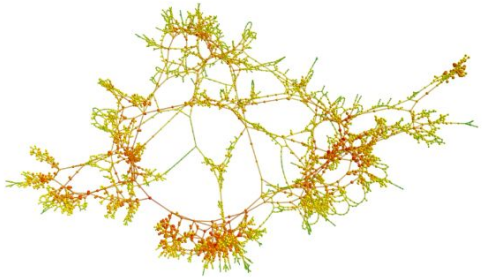
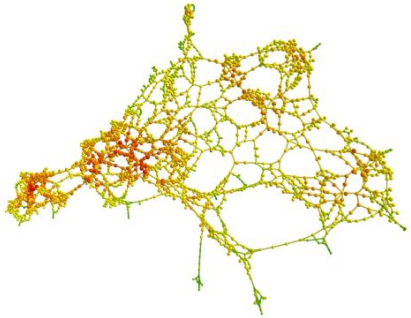
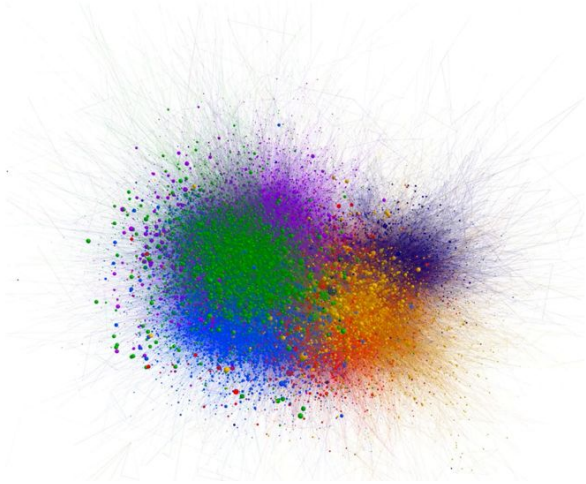


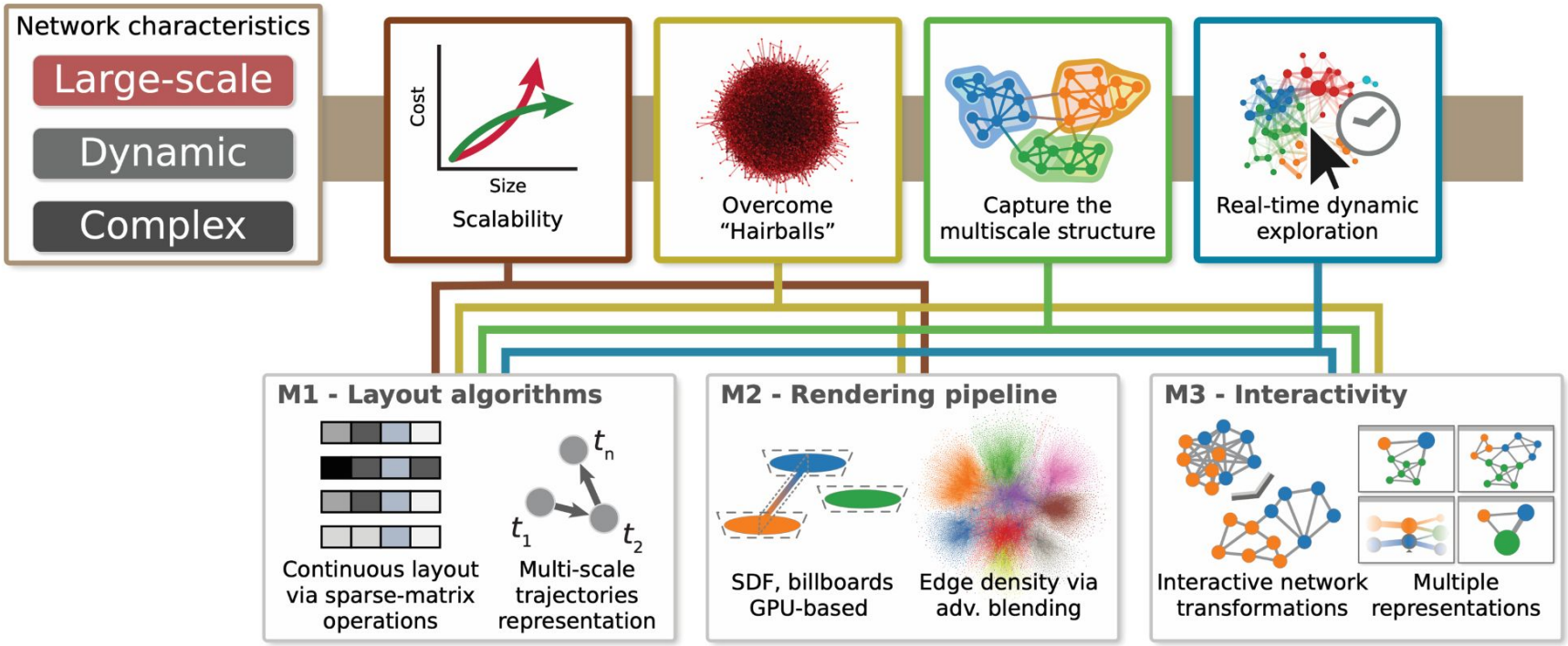



Astronomy
Earth and ocean sciences
Mathematics
Agricultural sciences
Electronics & comm. engineering
Material engineering
Computer sciences
Chemistry
Psychology
Physics
Social sciences
Health
Biological sciences




data


Network characteristics
Large-scale
Dynamic
Complex
Cost
Size
Scalability
Overcome
“Hairballs”
Capture the
multiscale structure
Real-time dynamic
exploration
M1 - Layout algorithms
t_n
t_1
t_2
Continuous layout
via sparse-matrix
operations
Multi-scale
trajectories
representation
M2 - Rendering pipeline
SDF, billboards
GPU-based
Edge density via
adv. blending
M3 - Interactivity
Interactive network
transformations
Multiple
representations

Katy Borner

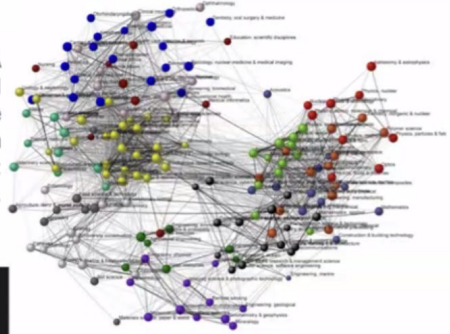
Katy Borner

Filipi N. Silva, IU, OSoMe

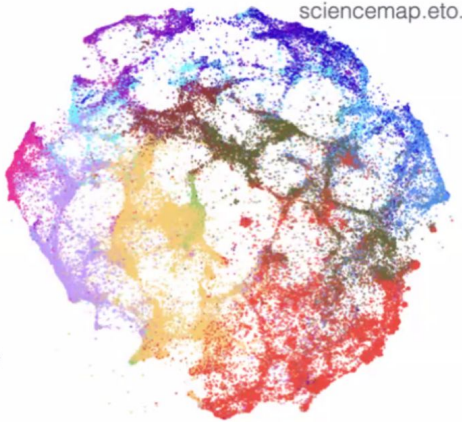
Meryl Sarah Jac...

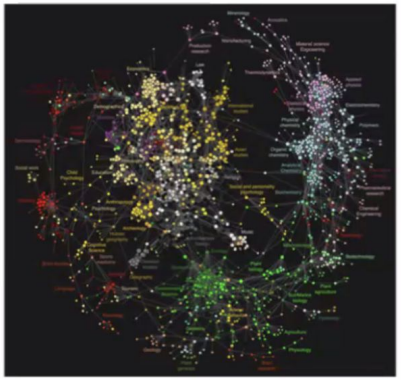
Meryl Sarah Jacob

Leydesdorff, Loet, and Ismael Rafols. "A global map of science based on the ISI subject categories." Journal of the American Society for Information Science and Technology 60.2 (2009): 348-362.



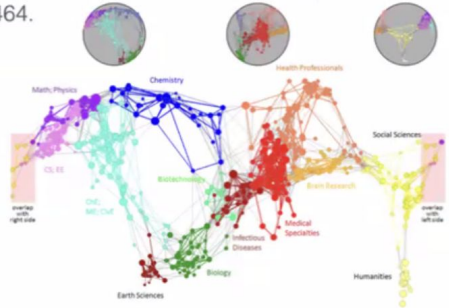
ETO Map of Science
sciencemap.eto.tech

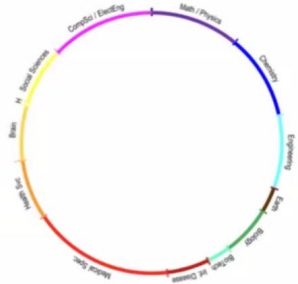


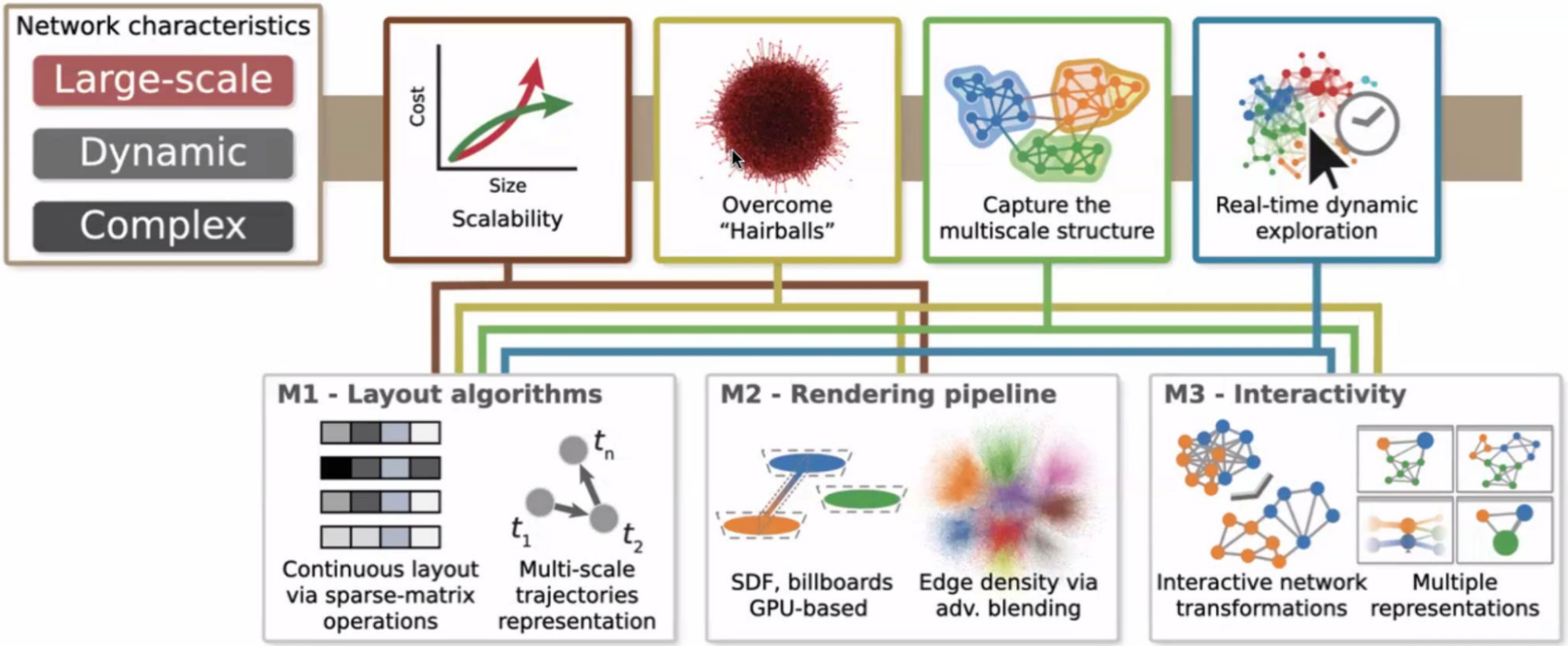


Bollen, Johan, et al. "Clickstream data yields high-resolution maps of science." PloS one 4.3 (2009): e4803.

Börner, Katy, et al. "Design and update of a classification system: The UCSD map of science." *PloS one* 7.7 (2012): e39464.






Network characteristics
Large-scale
Dynamic
Complex
Cost
Size
Scalability
Overcome "Hairballs"
Capture the multiscale structure
Real-time dynamic exploration
M1 - Layout algorithms
Continuous layout via sparse-matrix operations
Multi-scale trajectories representation
M2 - Rendering pipeline
SDF, billboards GPU-based
Edge density via adv. blending
M3 - Interactivity
Interactive network transformations
Multiple representations


web
Helios
An open-source visualization library for the web
heliosweb.io

Open-source web framework

can be integrated in websites, portals, dashboards ...

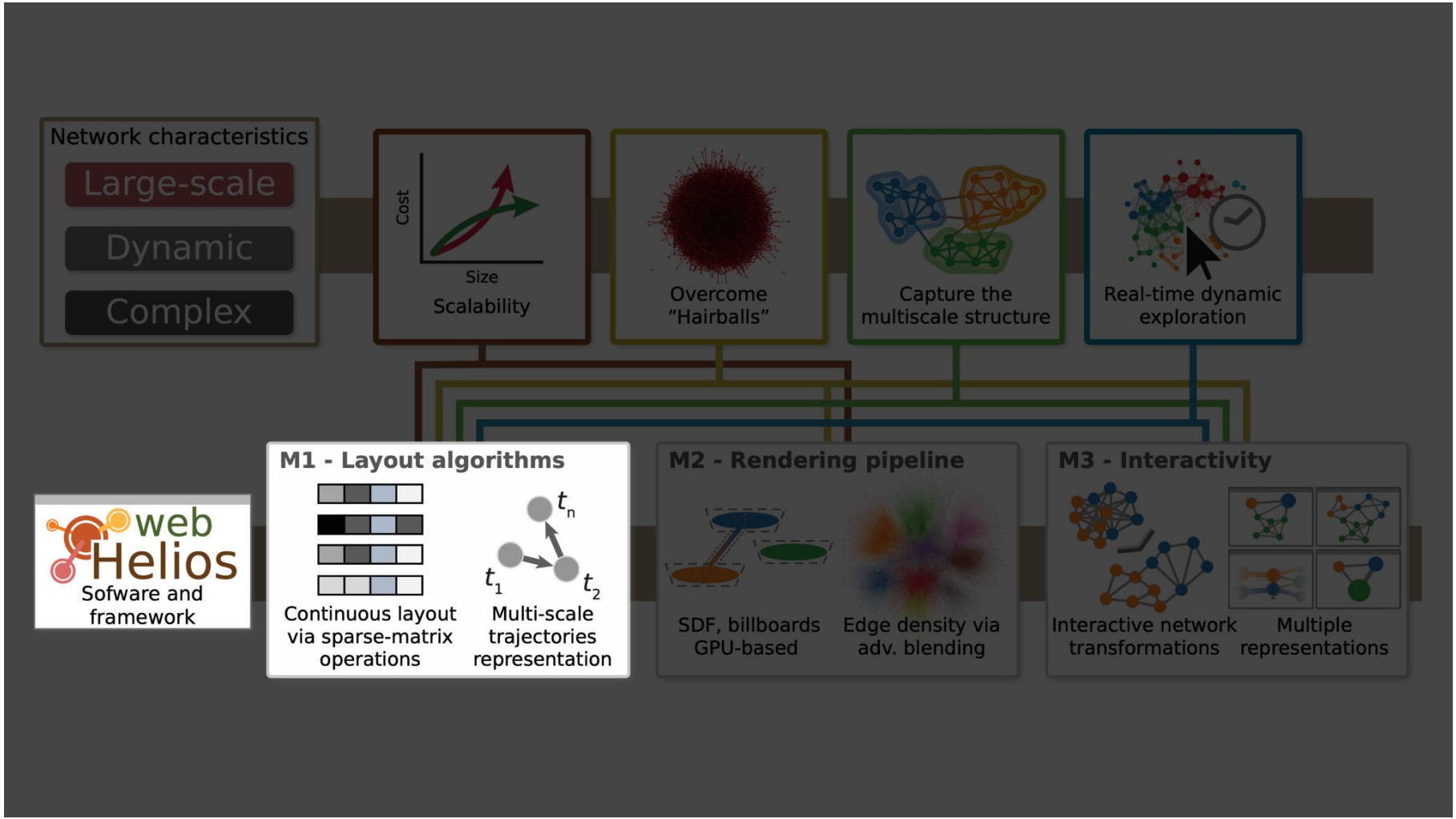
Optimized rendering and layouts

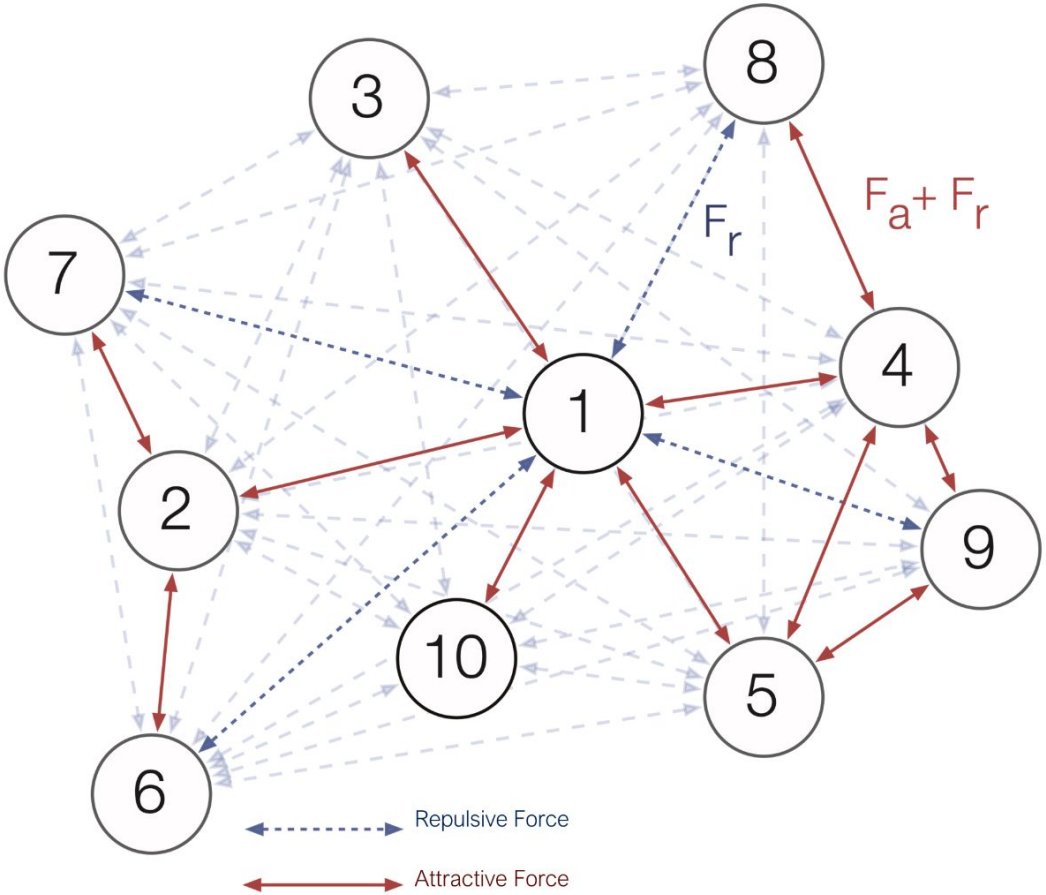
can visualize large networks, high-quality rendering ...

Interactivity*

allows picking, filtering, navigation, multi-representations ...

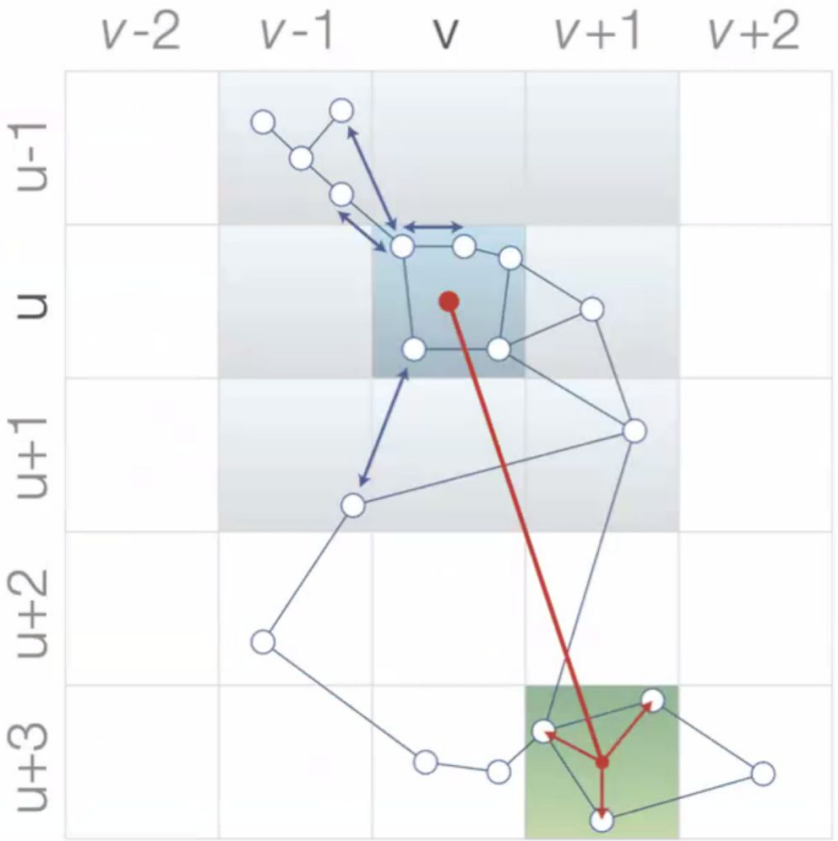
*in development


Network characteristics
Large-scale
Dynamic
Complex
Cost
Size
Scalability
Overcome "Hairballs"
Capture the multiscale structure
Real-time dynamic exploration
web Helios
Sofware and framework
M1 - Layout algorithms
Continuous layout via sparse-matrix operations
t_n
t_1
t_2
Multi-scale trajectories representation
M2 - Rendering pipeline
SDF, billboards GPU-based
Edge density via adv. blending
M3 - Interactivity
Interactive network transformations
Multiple representations


8
3
$F_a + F_r$
F_r
7
4
1
2
9
10
5
6
Repulsive Force
Attractive Force

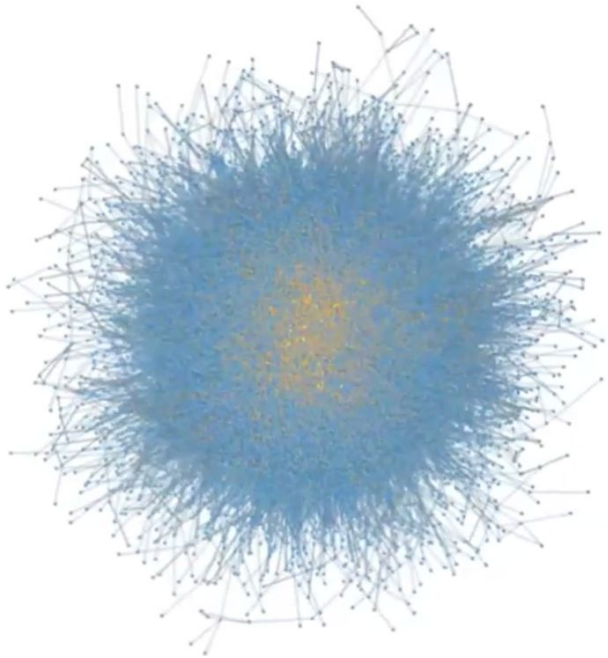
Layout optimizations

- Molecular dynamics simulation is $O(N^2)$.
- We can use multipole expansion (FM3):
- Segment the space
- Real-time continuous layout

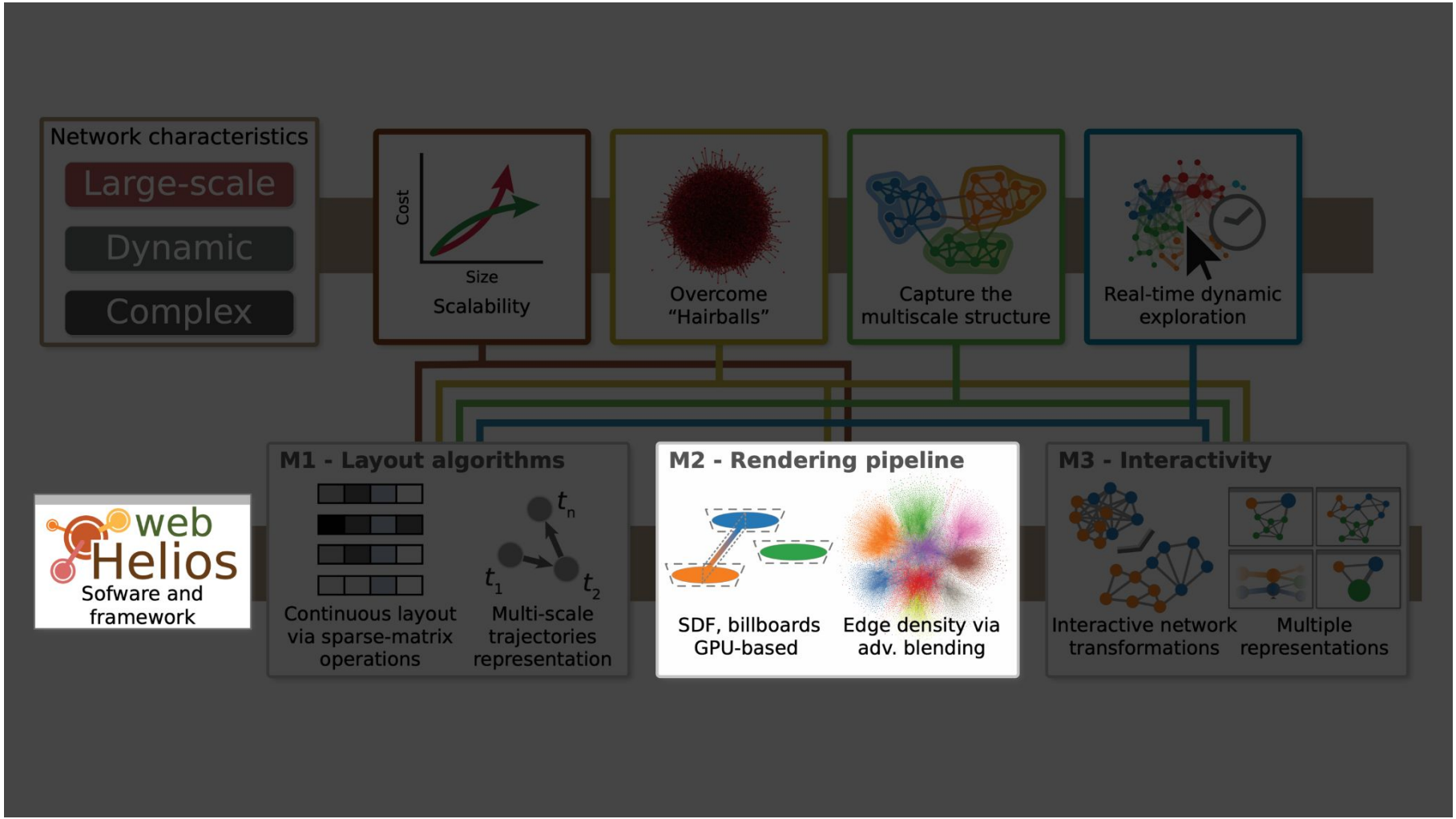


* S. Hachuland M. Jünger. Drawing large graphs with a potential-field-based multilevel algorithm. In International Symposium on Graph Drawing, pp. 285–295. Springer, 2004. doi: 10.1007/978-3-540-31843-9_29

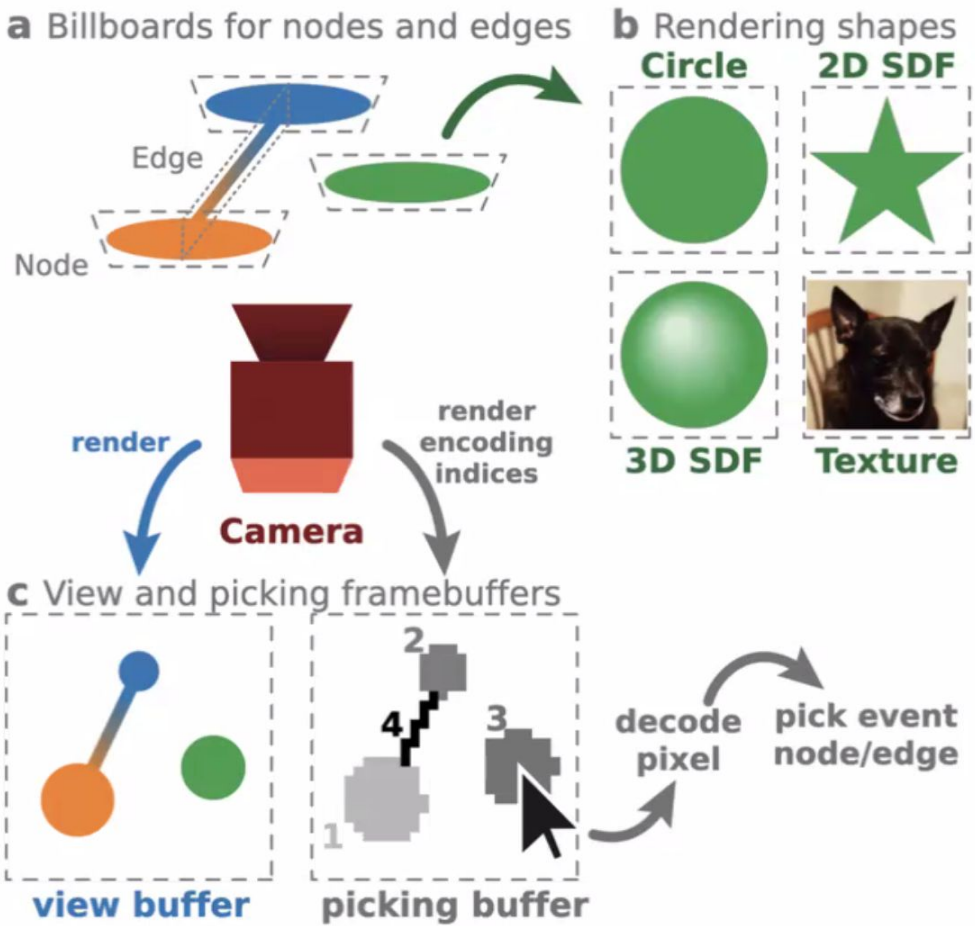
Wiki Medicine and Mathematics

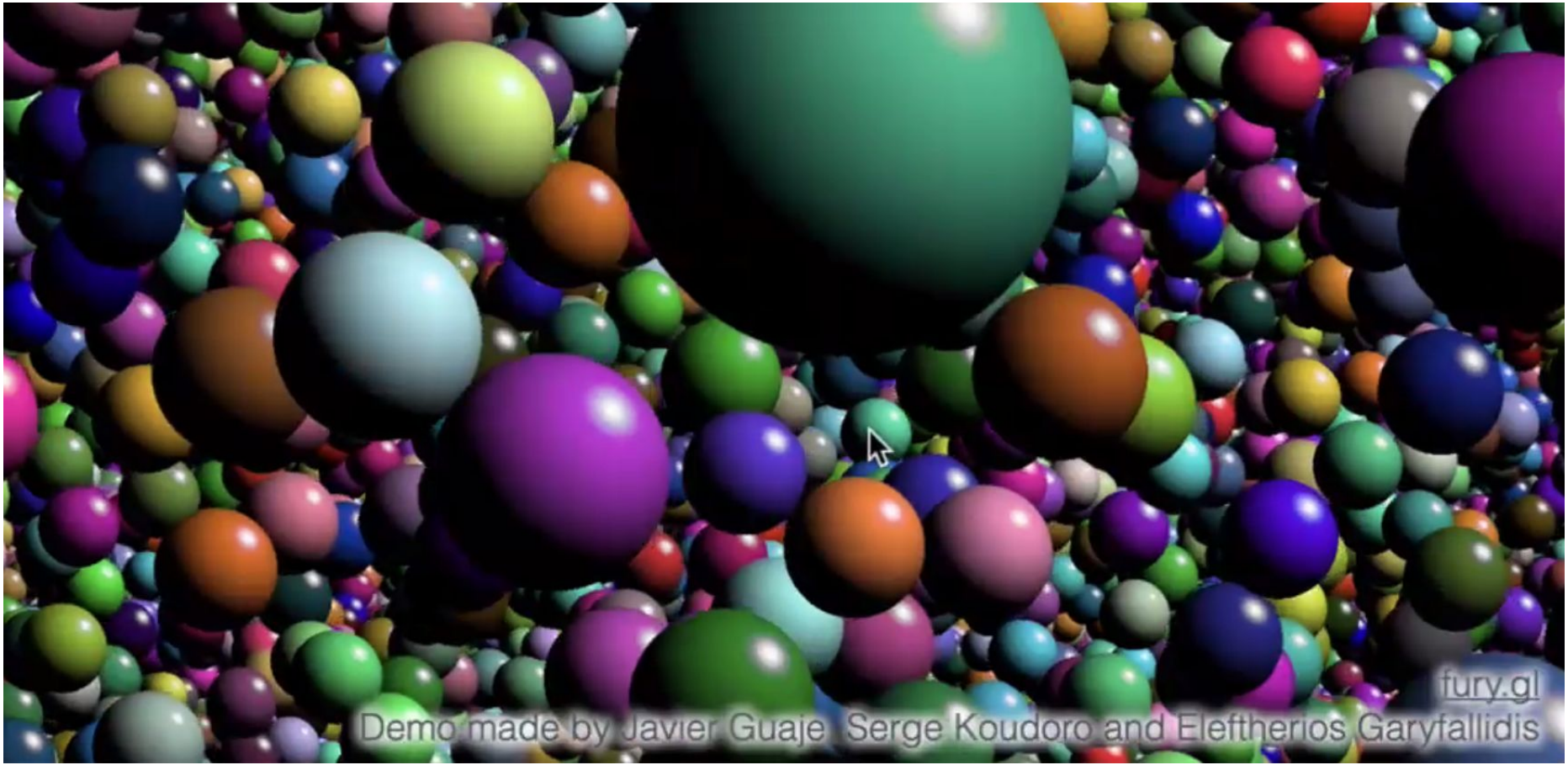


Visualizing Complex Networks (CDT-5)
Silva, F. N. and Costa, L. da F.
http://dx.doi.org/10.13140/RG.2.2.21310.74567/1

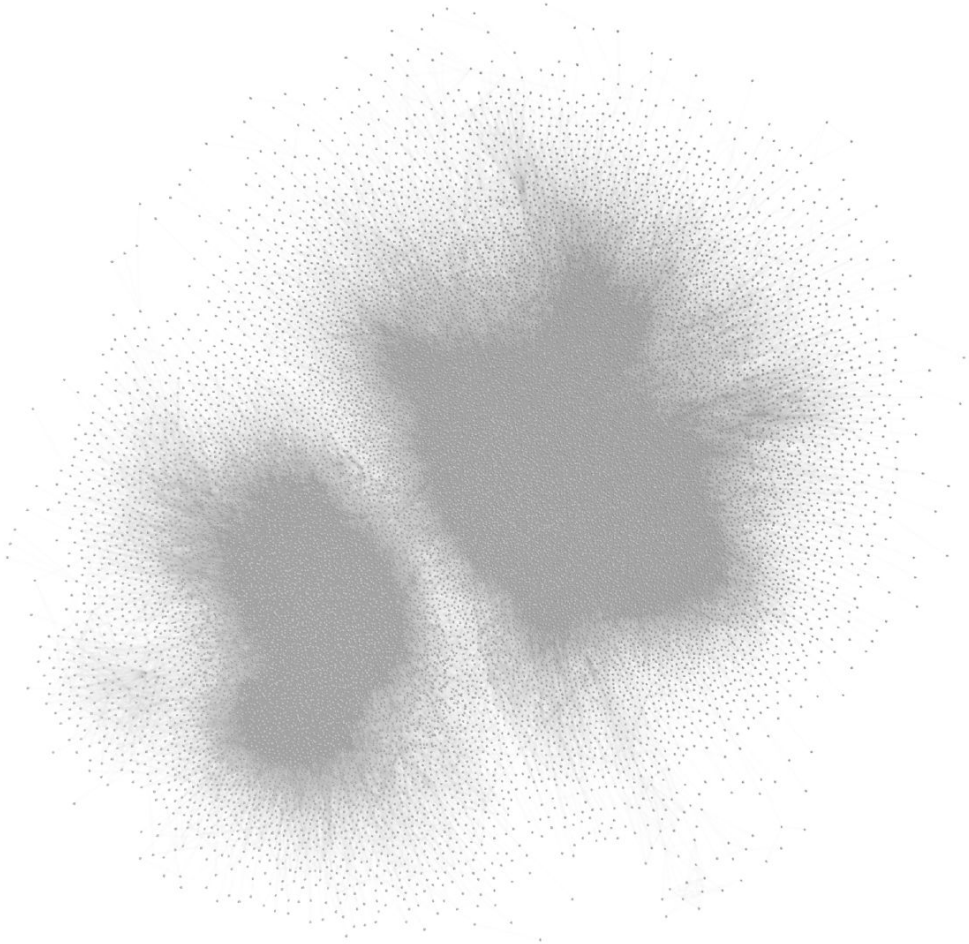

Network characteristics
Large-scale
Dynamic
Complex
Cost
Size
Scalability
Overcome "Hairballs"
Capture the multiscale structure
Real-time dynamic exploration
web Helios
Sofware and framework
M1 - Layout algorithms
Continuous layout via sparse-matrix operations
Multi-scale trajectories representation
M2 - Rendering pipeline
SDF, billboards GPU-based
Edge density via adv. blending
M3 - Interactivity
Interactive network transformations
Multiple representations

Rendering in the GPU

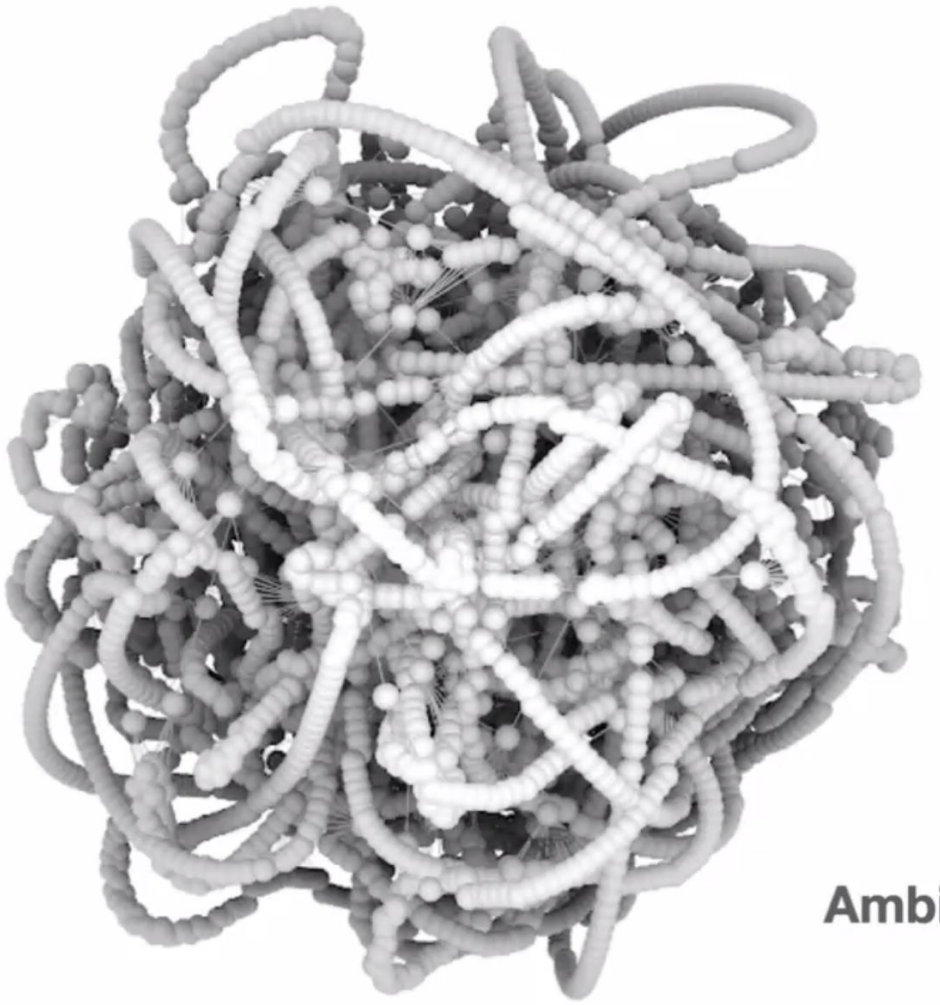



fury.gl
Demo made by Javier Guaje, Serge Koudoro and Eleftherios Garyfallidis

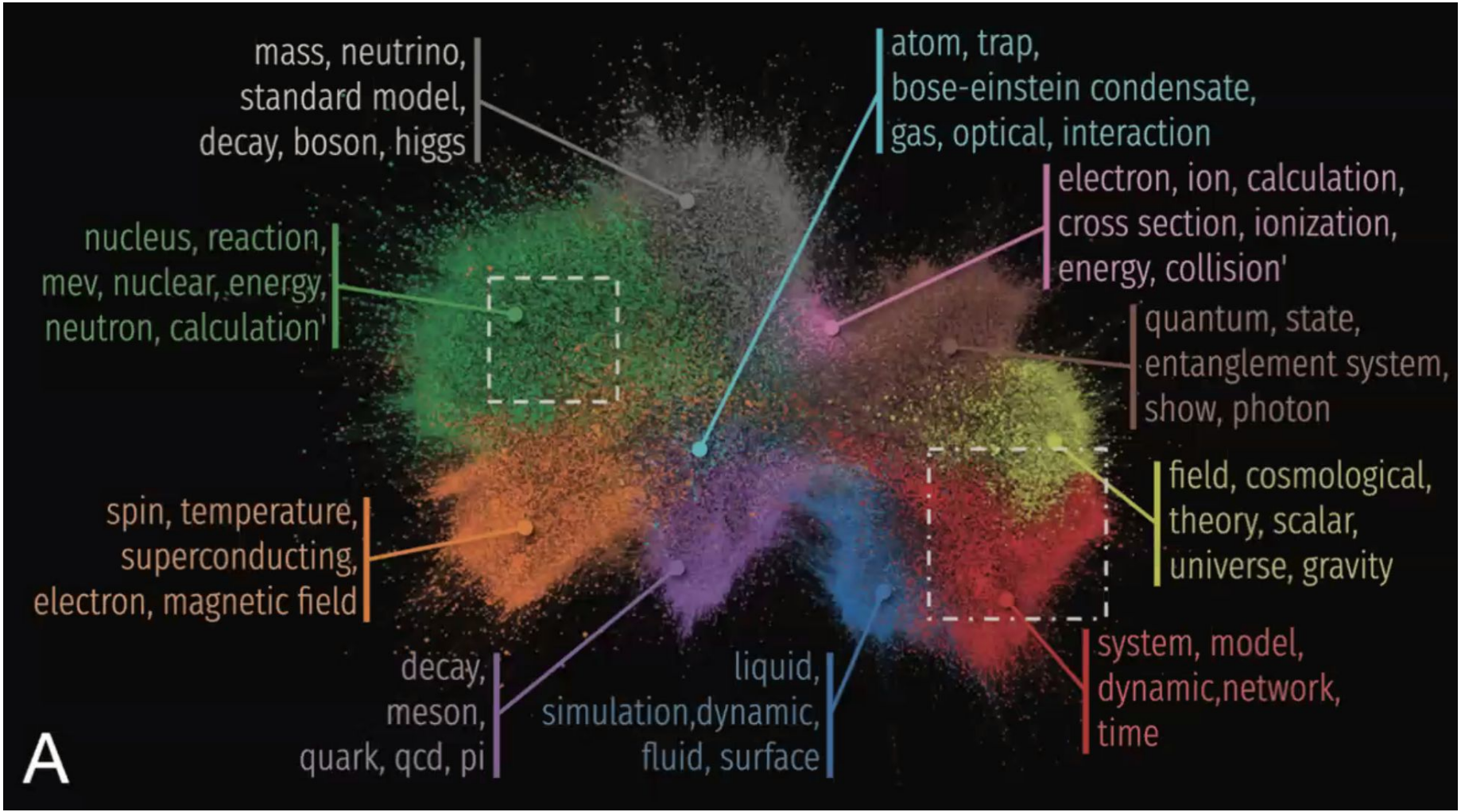
Rendering



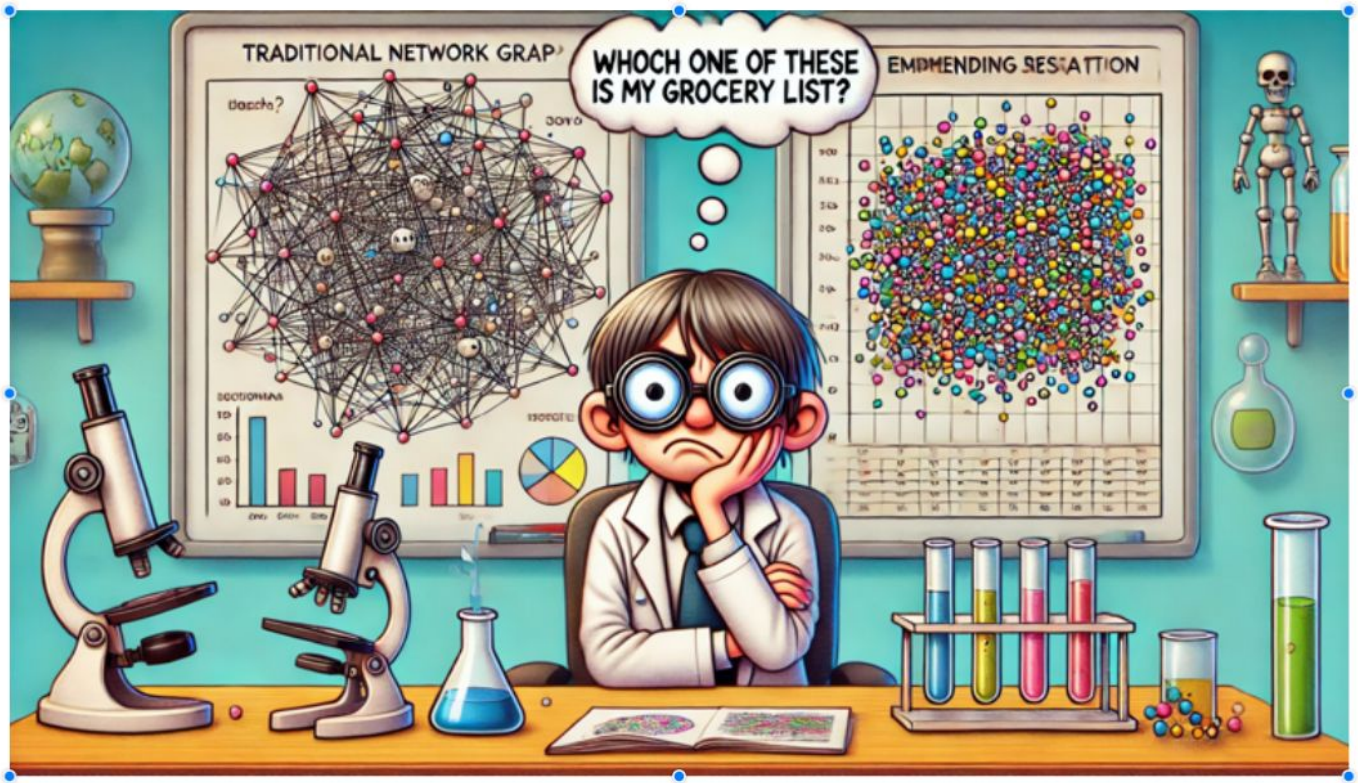
Edge density



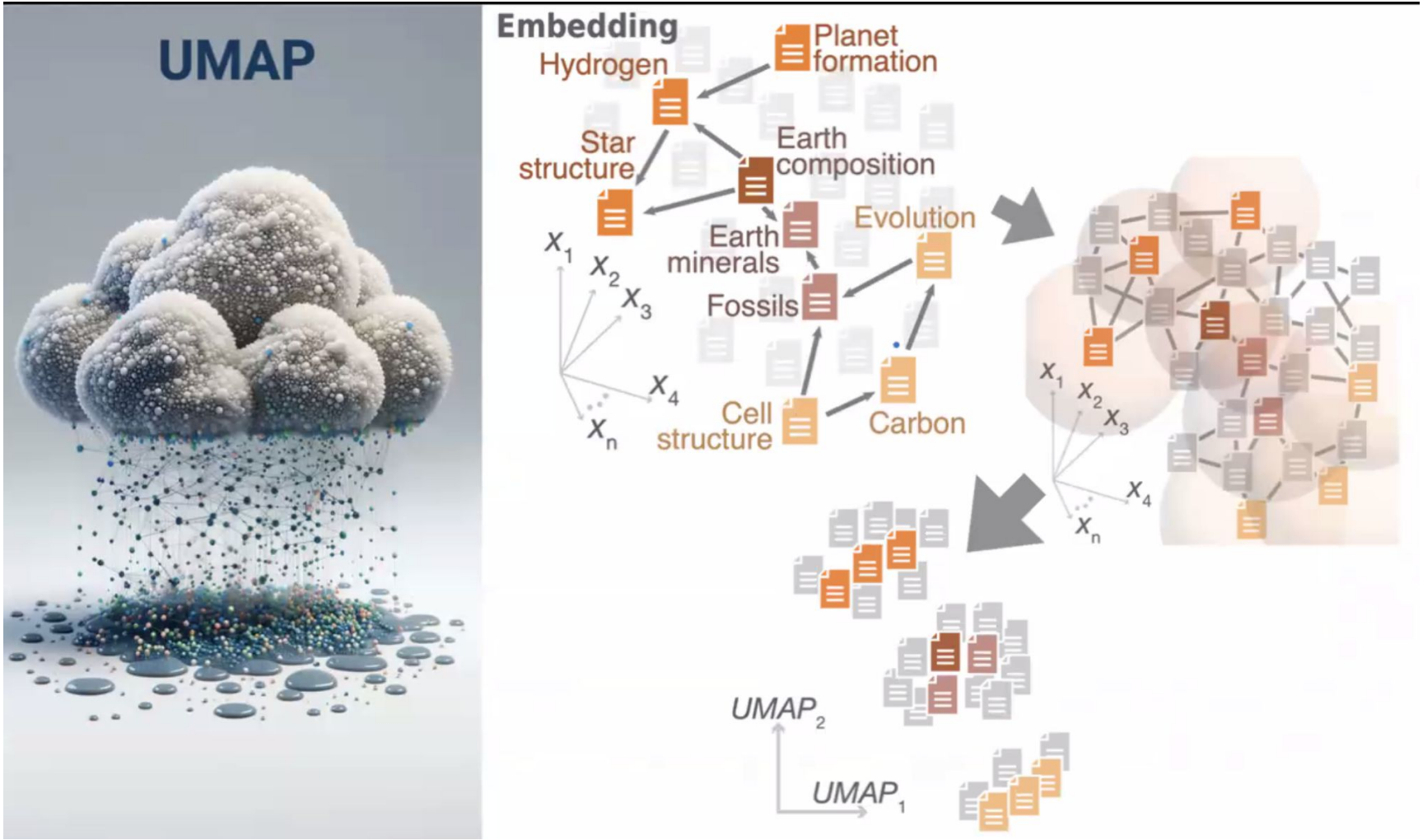
Ambience Occlusion


mass, neutrino,
standard model,
decay, boson, higgs
atom, trap,
bose-einstein condensate,
gas, optical, interaction
electron, ion, calculation,
cross section, ionization,
energy, collision'
nucleus, reaction,
mev, nuclear, energy,
neutron, calculation'
quantum, state,
entanglement system,
show, photon
field, cosmological,
theory, scalar,
universe, gravity
spin, temperature,
superconducting,
electron, magnetic field
system, model,
dynamic,network,
time
decay,
meson,
quark, qcd, pi
liquid,
simulation,dynamic,
fluid, surface
A

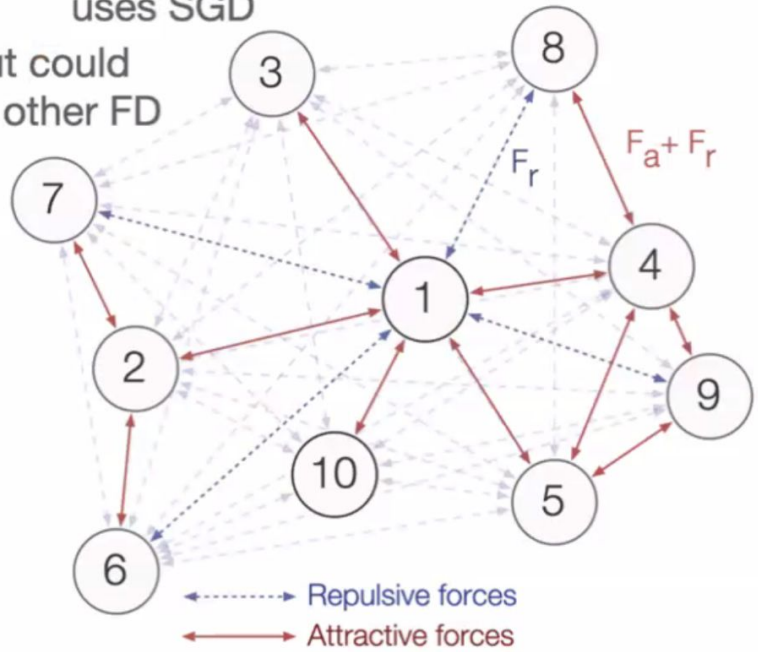

A - artificial neural network, simulation, system, pape...
B - convolutional neural network, propose, learning, t...
C - kalman filter, estimation, estimate, state, measur...
D - support vector machine, classification, classifier, ...
E - signal, brain, independent component analysis, e...
F - image, patient, classification, diagnosis, %, disea...
G - hidden markov model, feature, paper, speech rec...
H - linear regression, study, conclusion, associate, a...
I - kriging, spatial, soil, study, sample, area, concentr...
J - principal component analysis, sample, componen...
K - principal component analysis, image, feature, faci...
L - image, classification, area, spatial, spectral, data, ...
M - neural network, invention disclose, accord, meth...
N - neural network, neuron, spike, synaptic, brain, ac...
O - probability, probabilistic, inference, model, struct...
P - fault diagnosis, base, monitoring, propose, signal,...
Q - molecular, compound, descriptor, regression, rel...
R - linear regression model, estimator, estimate, nonl...
Other
Export
Size
Color
cluster name (level1
Category18
Edges



Ok, but what about embeddings?


UMAP
Embedding
Hydrogen
Planet formation
Star structure
Earth composition
Earth minerals
Evolution
Fossils
Cell structure
Carbon
x_1
x_2
x_3
x_4
x_n
$UMAP_2$
$UMAP_1$

UMAP uses SGD but could use other FD



For large embeddings: >10M points

Negative samples rate : Number of repulsive interactions to update for each positive.

increased to 10 (default = 5)

Epochs: Number of iterations

increased to 200000 (default = 200) !!!

Number of neighbors: Number of neighbors in the NN graph

increased to 30 (default = 15) !!!

umap-learn.readthedocs.io

attractive repulsive

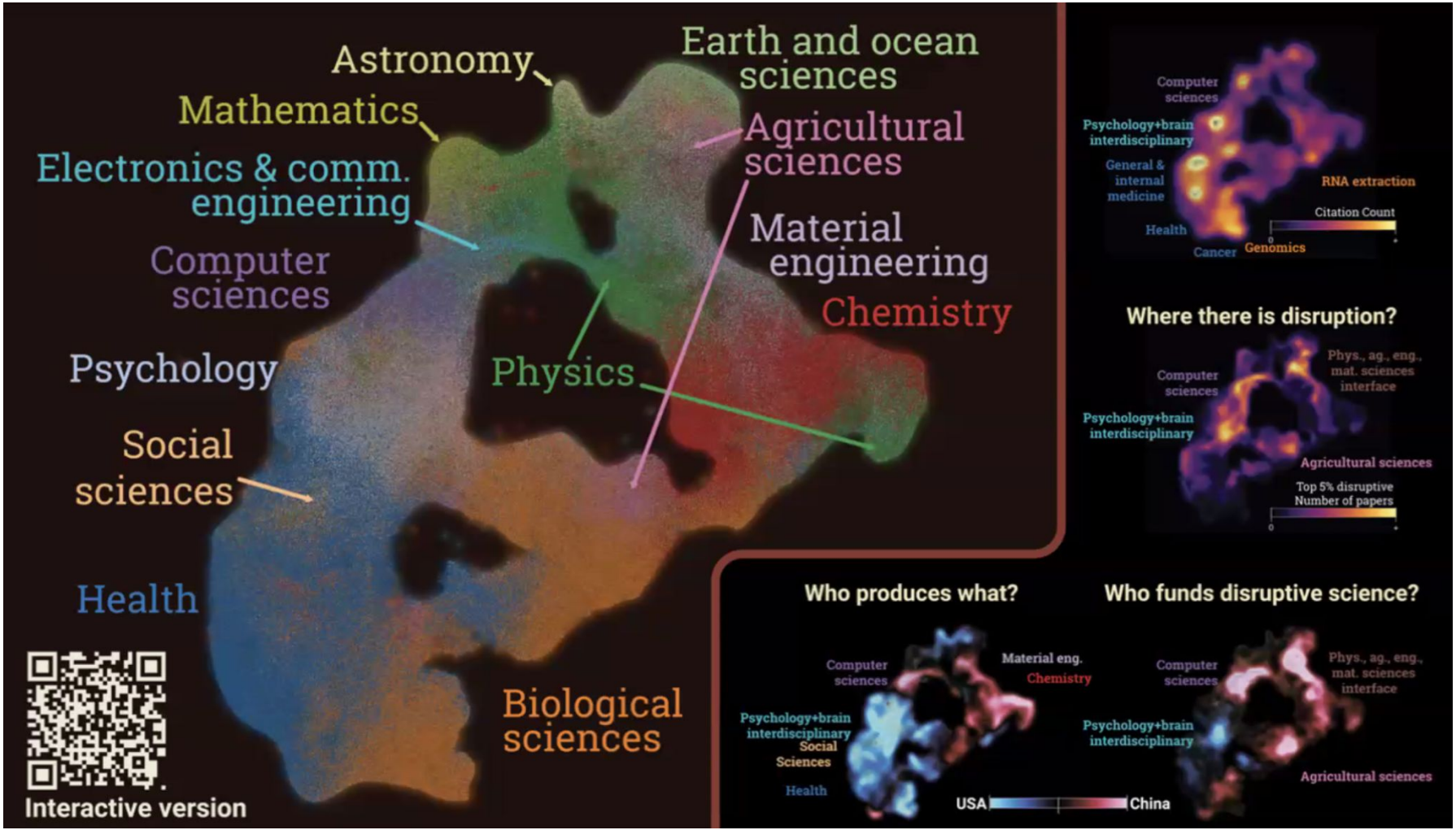
$$\frac{\partial \mathcal{L}_{\text{UMAP}}(\gamma)}{\partial \mathbf{y}_i} \sim \sum_j v_{ij} w_{ij} (\mathbf{y}_i - \mathbf{y}_j) - \gamma \sum_j \frac{1}{d_{ij}^2 + \epsilon} w_{ij} (\mathbf{y}_i - \mathbf{y}_j).$$

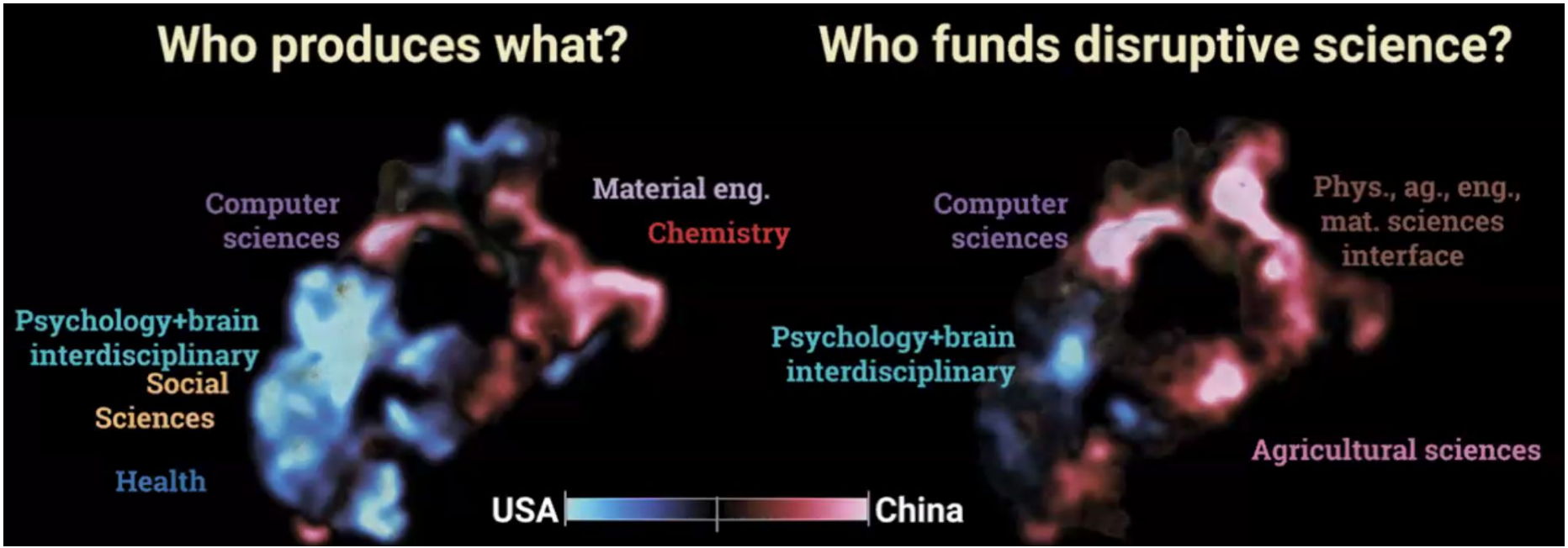
GPU version (much faster!)

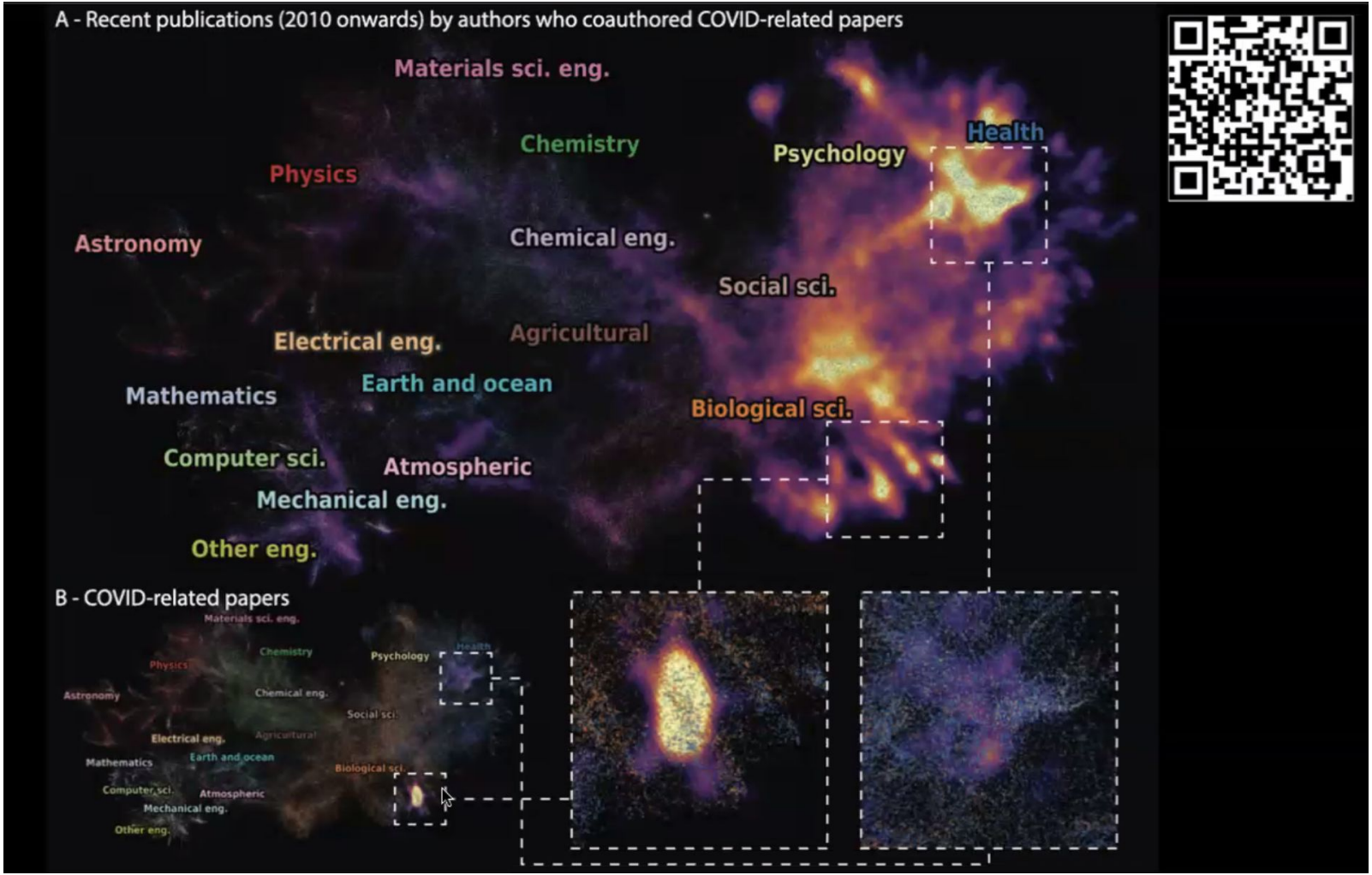
beware: it has bugs that lead to bad projections

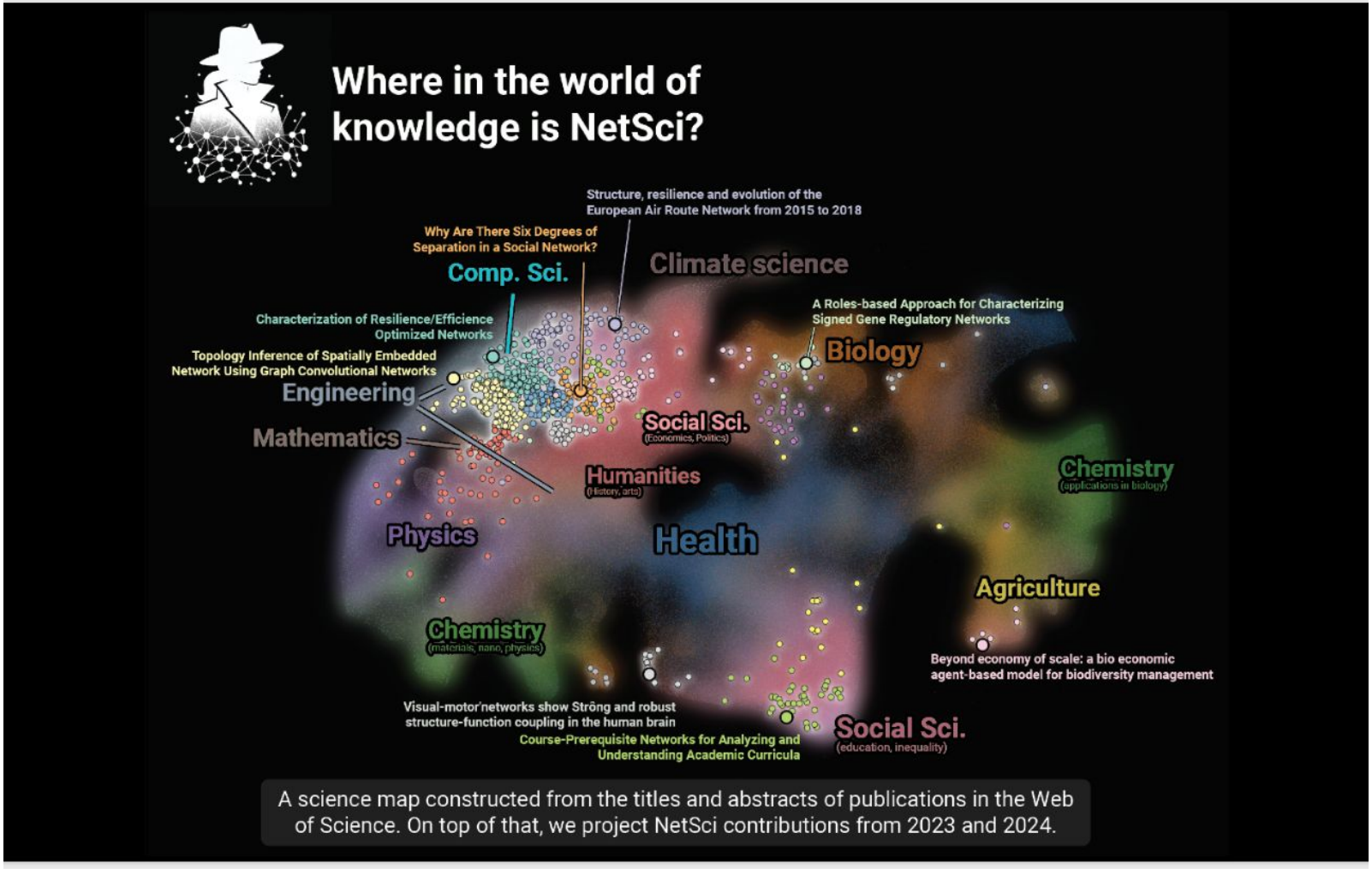
https://docs.rapids.ai/api/cuml/stable/api/#umap

Böhm, J. N., Berens, P., & Kobak, D. (2022). Attraction-repulsion spectrum in neighbor embeddings. The Journal of Machine Learning Research, 23(1), 4118-4149.

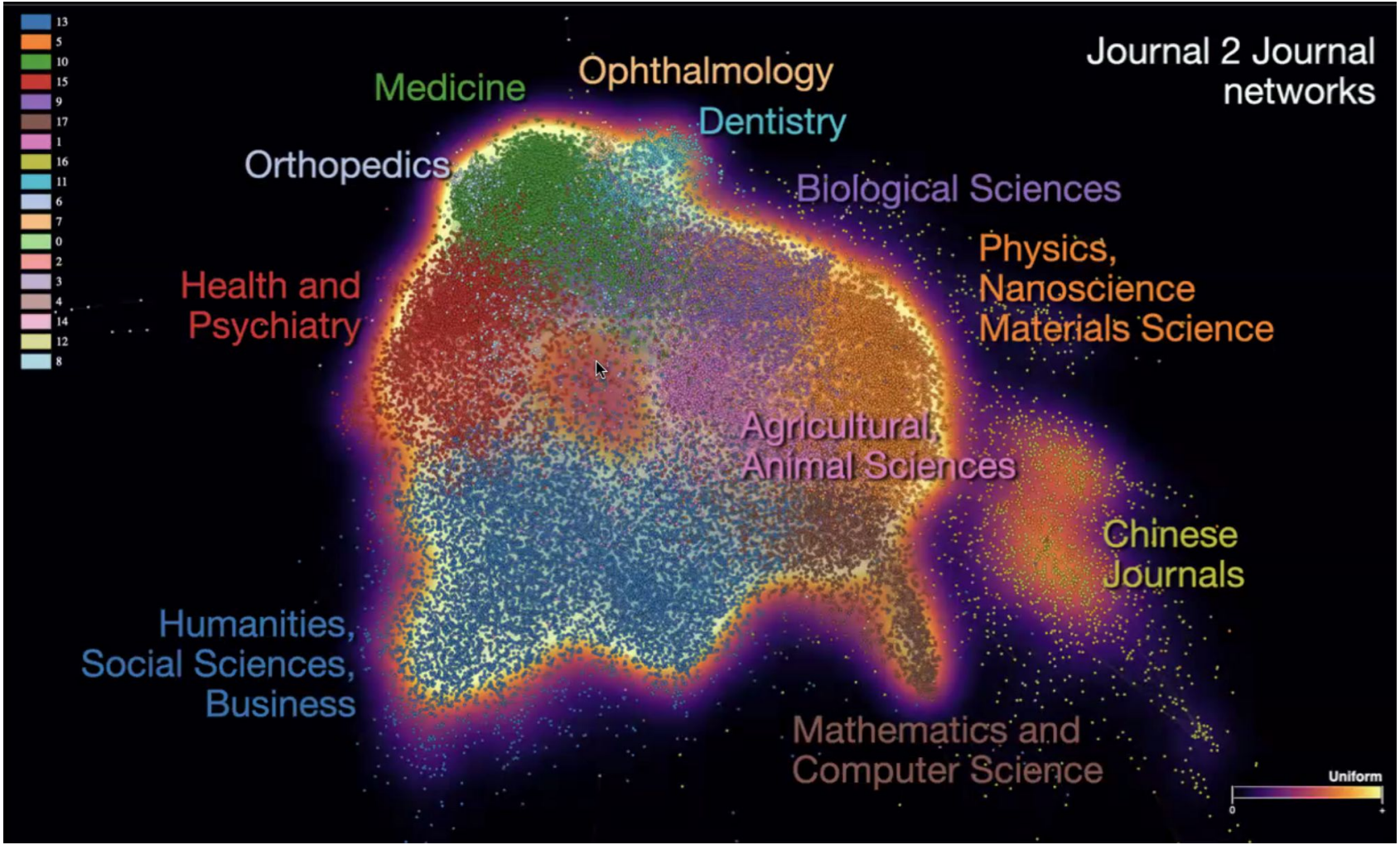

Astronomy
Earth and ocean sciences
Mathematics
Agricultural sciences
Electronics & comm. engineering
Material engineering
Computer sciences
Chemistry
Psychology
Physics
Social sciences
Health
Biological sciences
Interactive version
Computer sciences
Psychology+brain interdisciplinary
General & internal medicine
RNA extraction
Citation Count
Health
Cancer
Genomics
Where there is disruption?
Computer sciences
Phys., ag., eng., mat. sciences interface
Psychology+brain interdisciplinary
Agricultural sciences
Top 5% disruptive Number of papers
Who produces what?
Computer sciences
Material eng.
Chemistry
Psychology+brain interdisciplinary
Social Sciences
Health
USA
China
Who funds disruptive science?
Computer sciences
Phys., ag., eng., mat. sciences interface
Psychology+brain interdisciplinary
Agricultural sciences

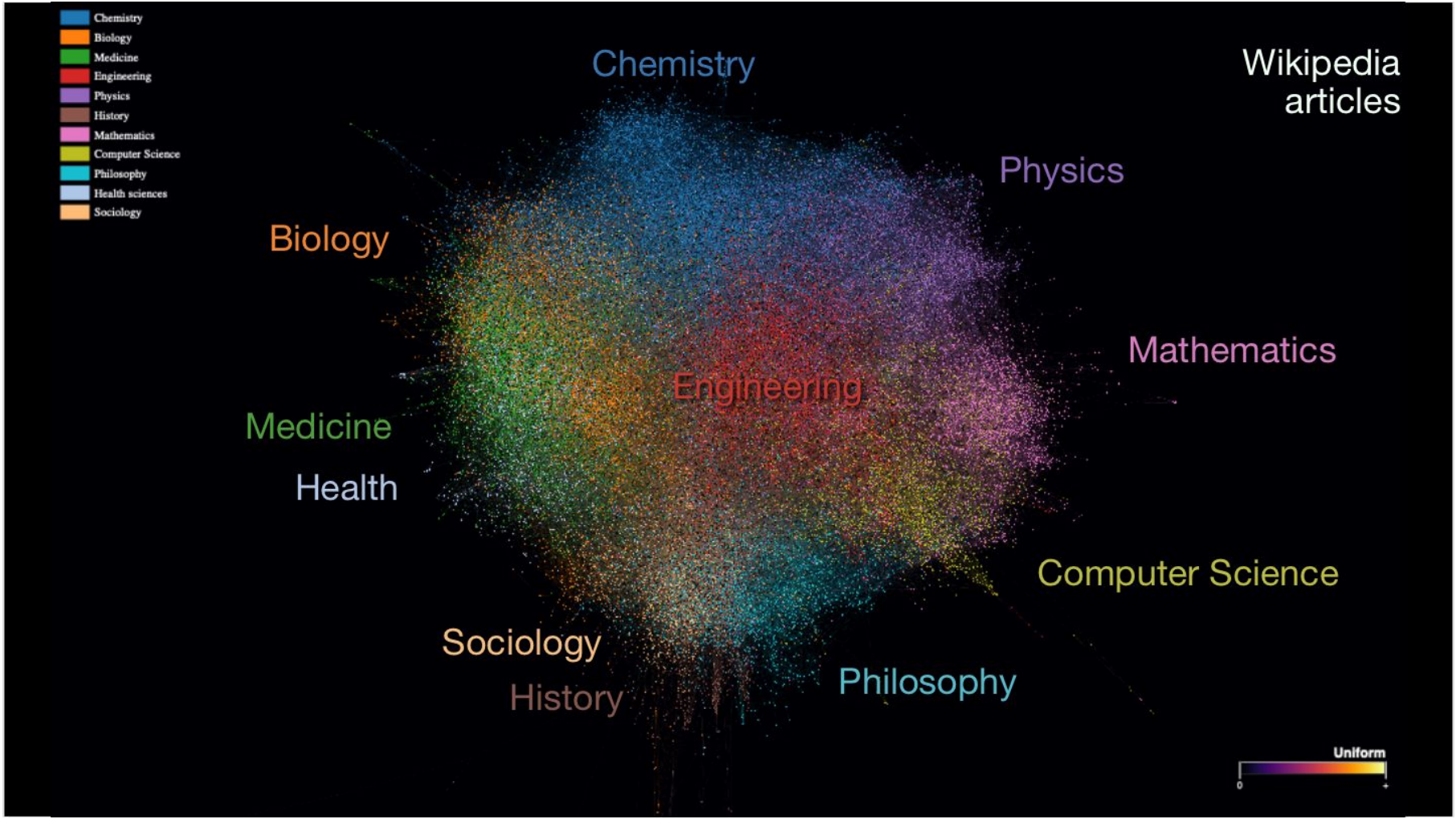

Who produces what?
Computer sciences
Material eng.
Chemistry
Psychology+brain interdisciplinary
Social Sciences
Health
USA
China
Who funds disruptive science?
Computer sciences
Phys., ag., eng., mat. sciences interface
Psychology+brain interdisciplinary
Agricultural sciences

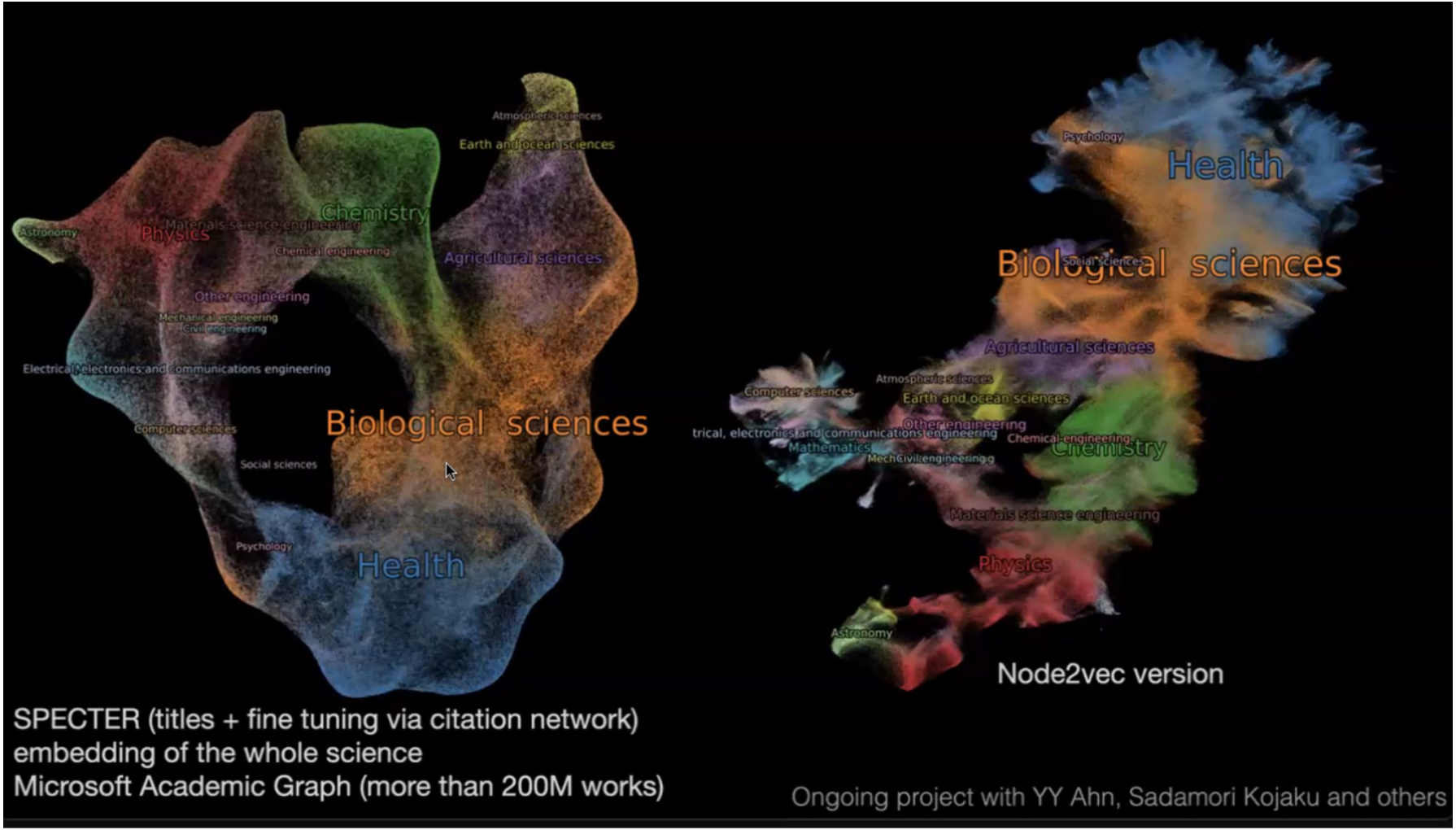

A - Recent publications (2010 onwards) by authors who coauthored COVID-related papers
Materials sci. eng.
Chemistry
Health
Psychology
Physics
Astronomy
Chemical eng.
Social sci.
Electrical eng.
Agricultural
Earth and ocean
Mathematics
Biological sci.
Computer sci.
Atmospheric
Mechanical eng.
Other eng.
B - COVID-related papers

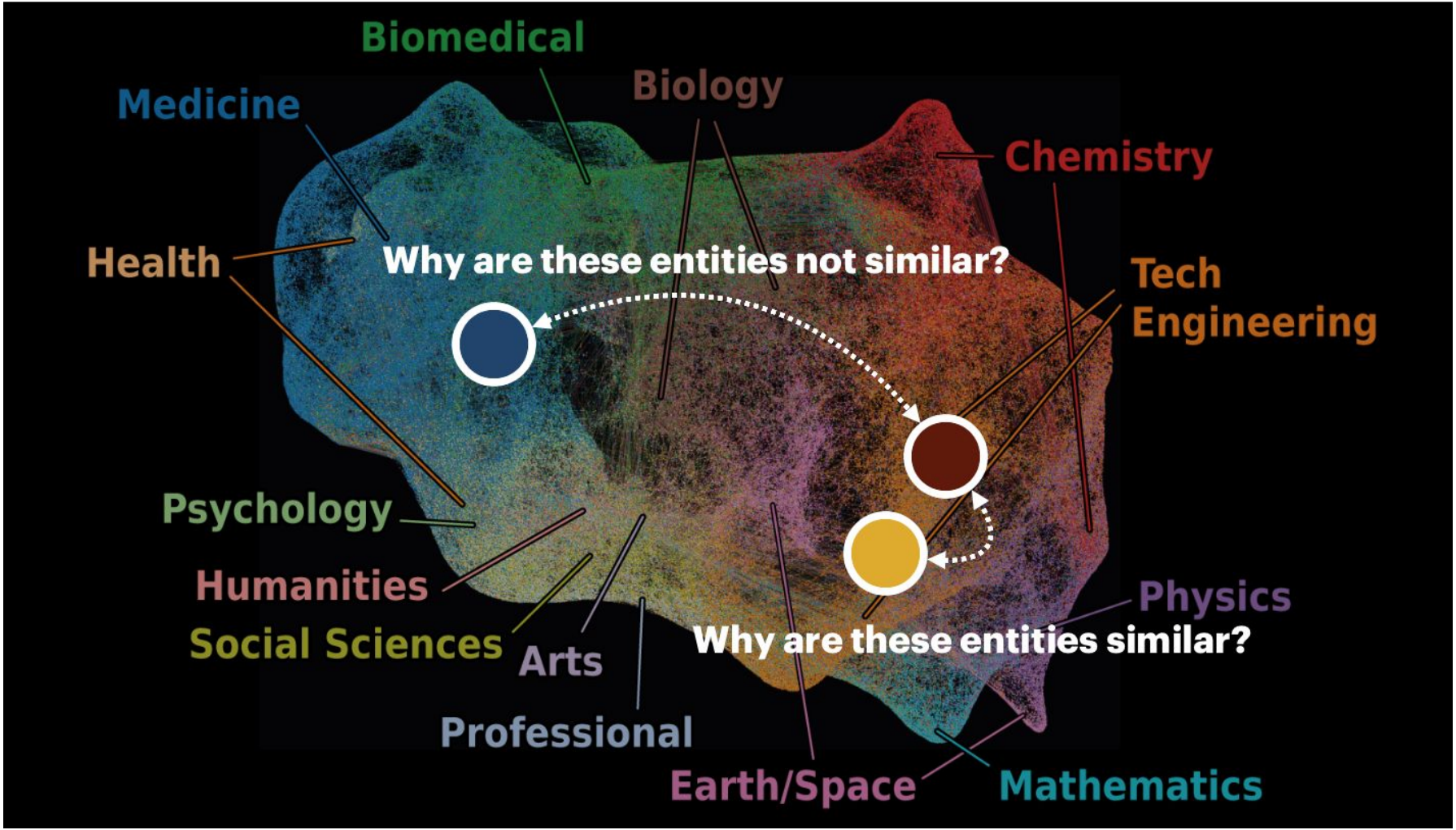


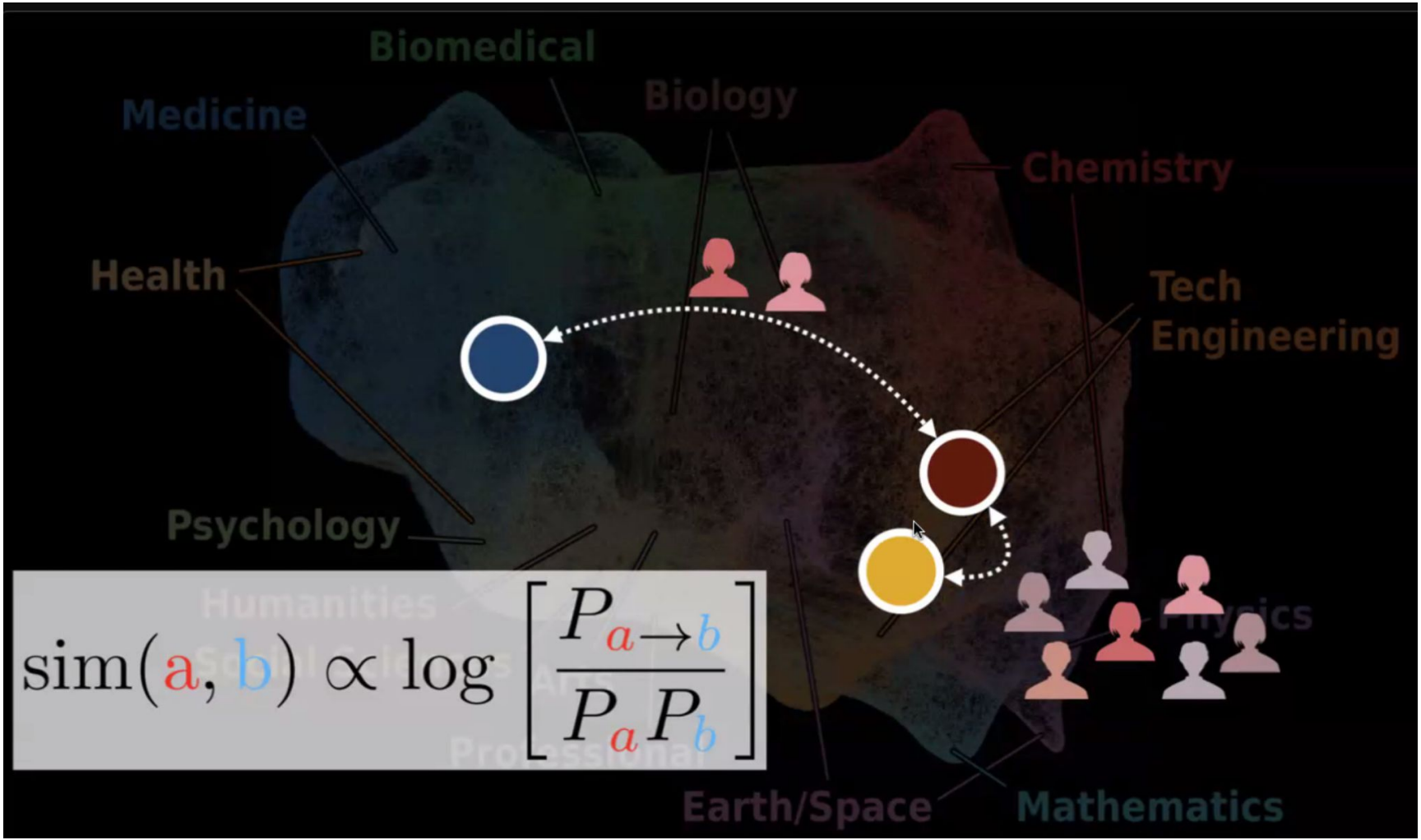
A science map constructed from the titles and abstracts of publications in the Web of Science. On top of that, we project NetSci contributions from 2023 and 2024.

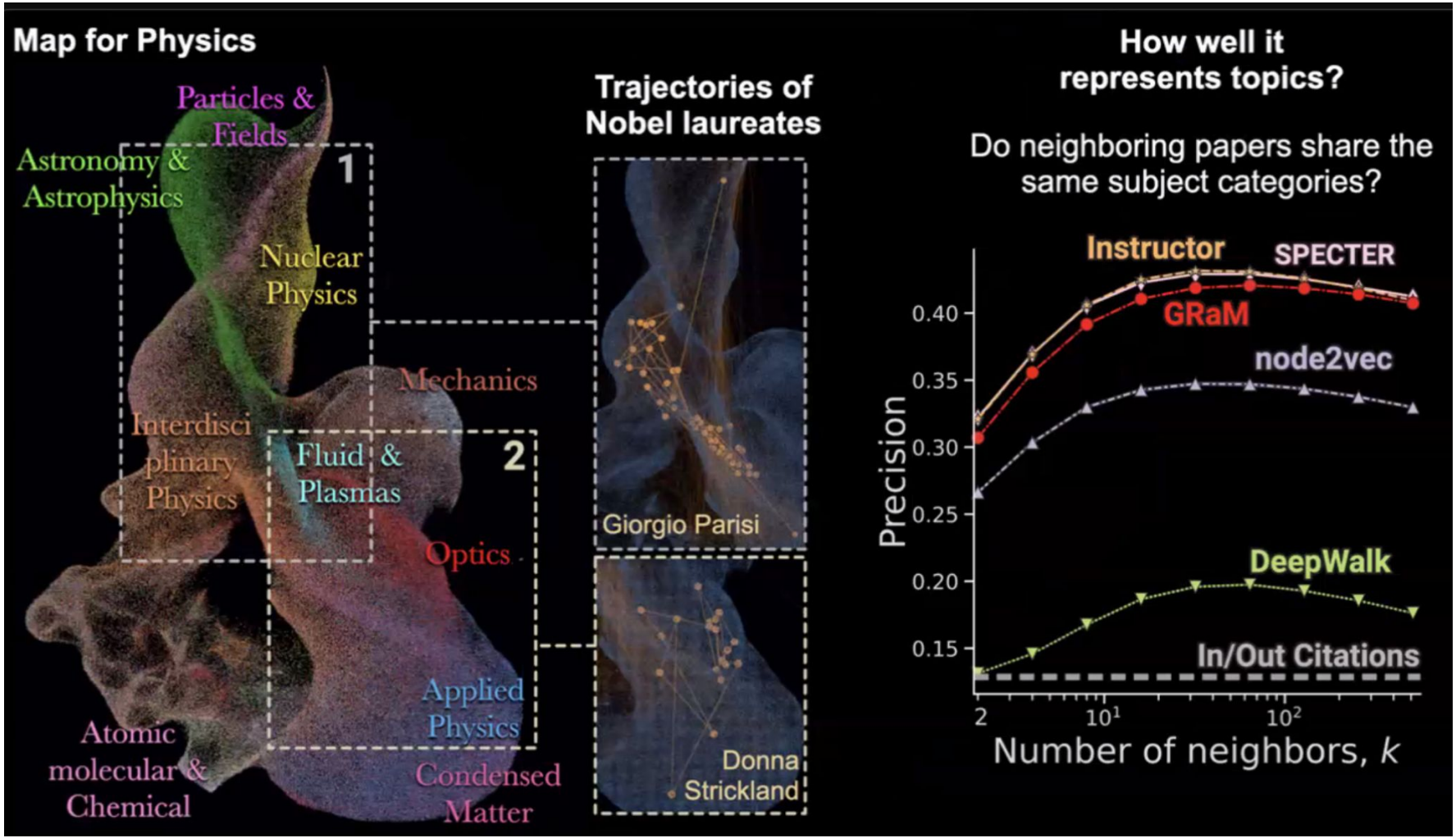

Journal 2 Journal networks
Medicine
Ophthalmology
Dentistry
Orthopedics
Biological Sciences
Physics, Nanoscience Materials Science
Health and Psychiatry
Agricultural, Animal Sciences
Chinese Journals
Humanities, Social Sciences, Business
Mathematics and Computer Science
Uniform
0
+


Chemistry
Biology
Medicine
Engineering
Physics
History
Mathematics
Computer Science
Philosophy
Health sciences
Sociology
Chemistry
Wikipedia articles
Physics
Biology
Mathematics
Engineering
Medicine
Health
Computer Science
Sociology
History
Philosophy
Uniform
0
+


Atmospheric sciences
Earth and ocean sciences
Chemistry
Astronomy
Physics
Materials science engineering
Chemical engineering
Agricultural sciences
Other engineering
Mechanical engineering
Civil engineering
Electrical, electronics and communications engineering
Computer sciences
Biological sciences
Social sciences
Psychology
Health
SPECTER (titles + fine tuning via citation network)
embedding of the whole science
Microsoft Academic Graph (more than 200M works)
Psychology
Health
Biological sciences
Social sciences
Agricultural sciences
Atmospheric sciences
Computer sciences
Earth and ocean sciences
Other engineering
trical, electronics and communications engineering
Chemical engineering
Mathematics
Chemistry
Mech Civil engineering
Materials science engineering
Physics
Astronomy
Node2vec version
Ongoing project with YY Ahn, Sadamori Kojaku and others

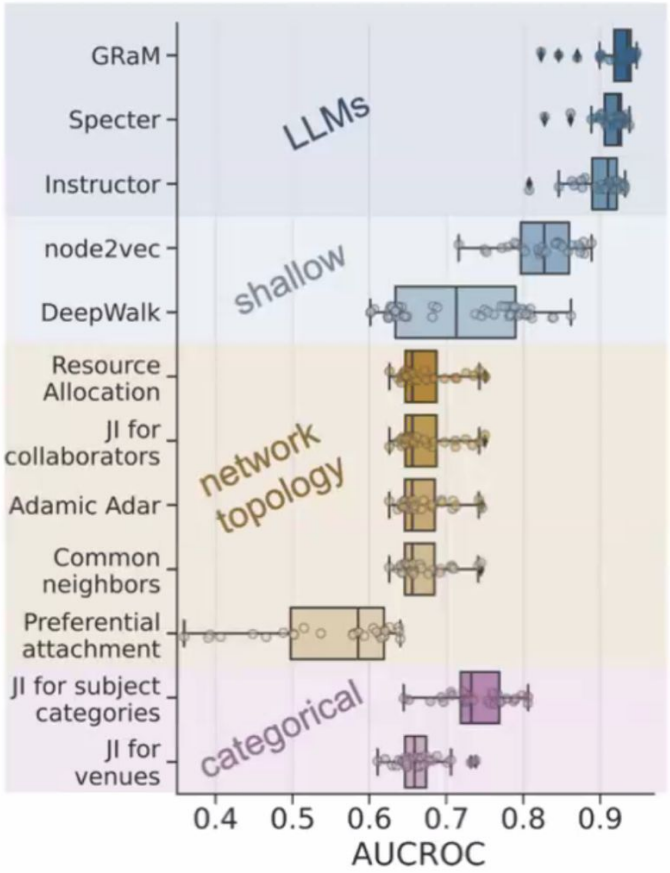

Biomedical
Medicine
Biology
Chemistry
Health
Why are these entities not similar?
Tech
Engineering
Psychology
Humanities
Social Sciences
Arts
Professional
Physics
Why are these entities similar?
Earth/Space
Mathematics

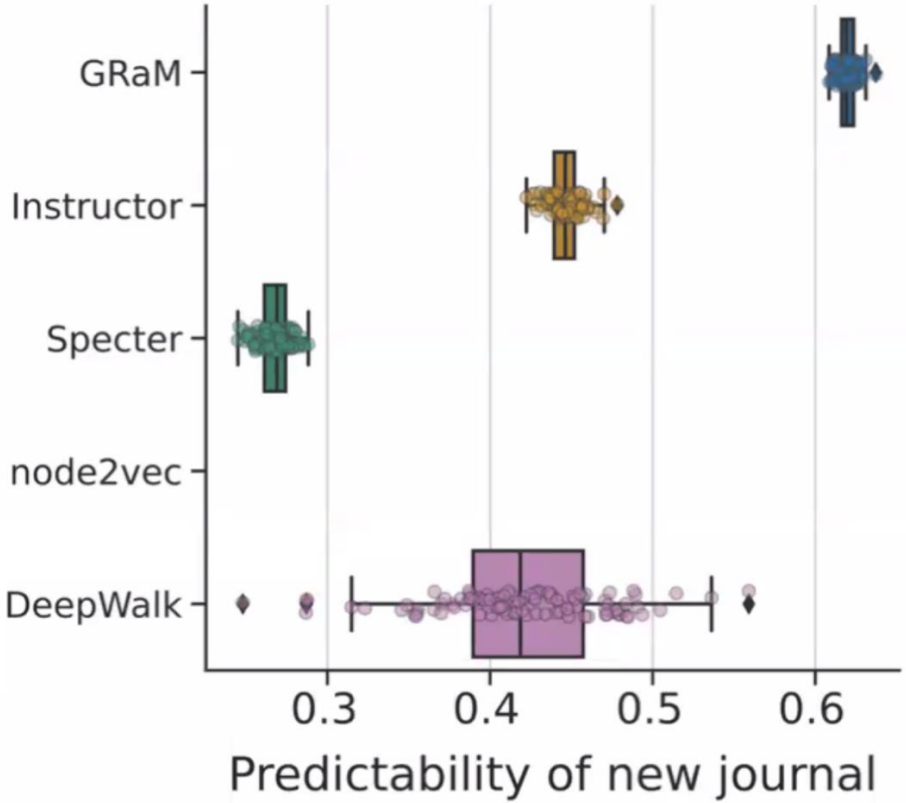

Biomedical
Medicine
Biology
Chemistry
Health
Tech
Engineering
Psychology
Physics
Earth/Space
Mathematics
$\text{sim}(a, b) \propto \log \left[\frac{P_{a \to b}}{P_a P_b} \right]$

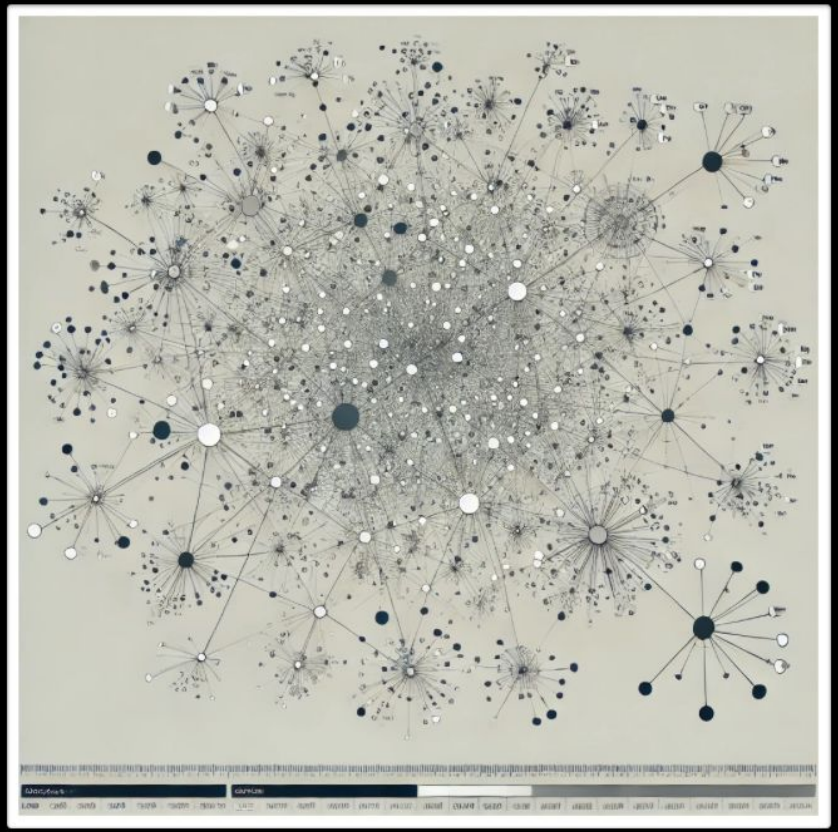

Map for Physics
Particles & Fields
Astronomy & Astrophysics
1
Nuclear Physics
Mechanics
Interdisci plinary Physics
Fluid & Plasmas
2
Optics
Applied Physics
Atomic molecular & Chemical
Condensed Matter
Trajectories of Nobel laureates
Giorgio Parisi
Donna Strickland
How well it represents topics?
Do neighboring papers share the same subject categories?
Instructor
SPECTER
GRaM
node2vec
DeepWalk
In/Out Citations
Precision
0.40
0.35
0.30
0.25
0.20
0.15
2
10^1
10^2
Number of neighbors, k

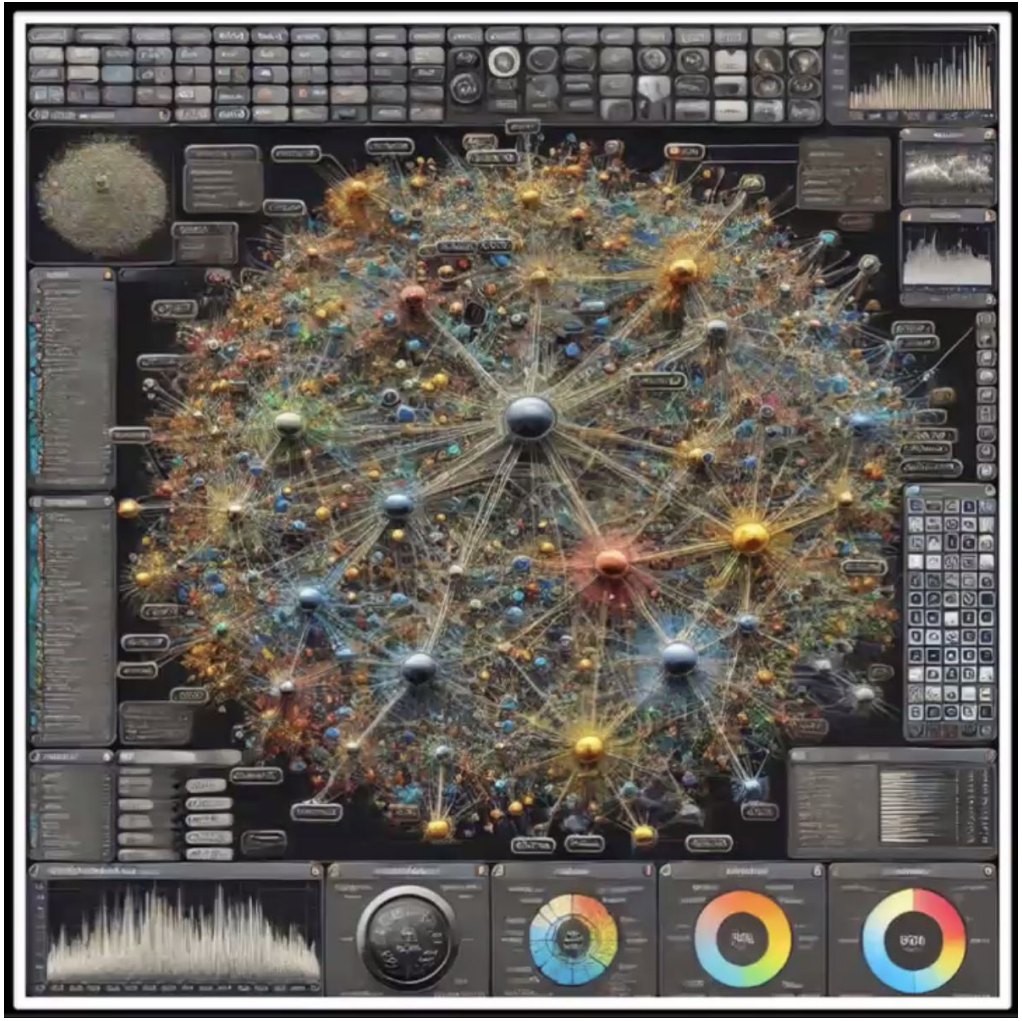
Predictive power of the GRaM

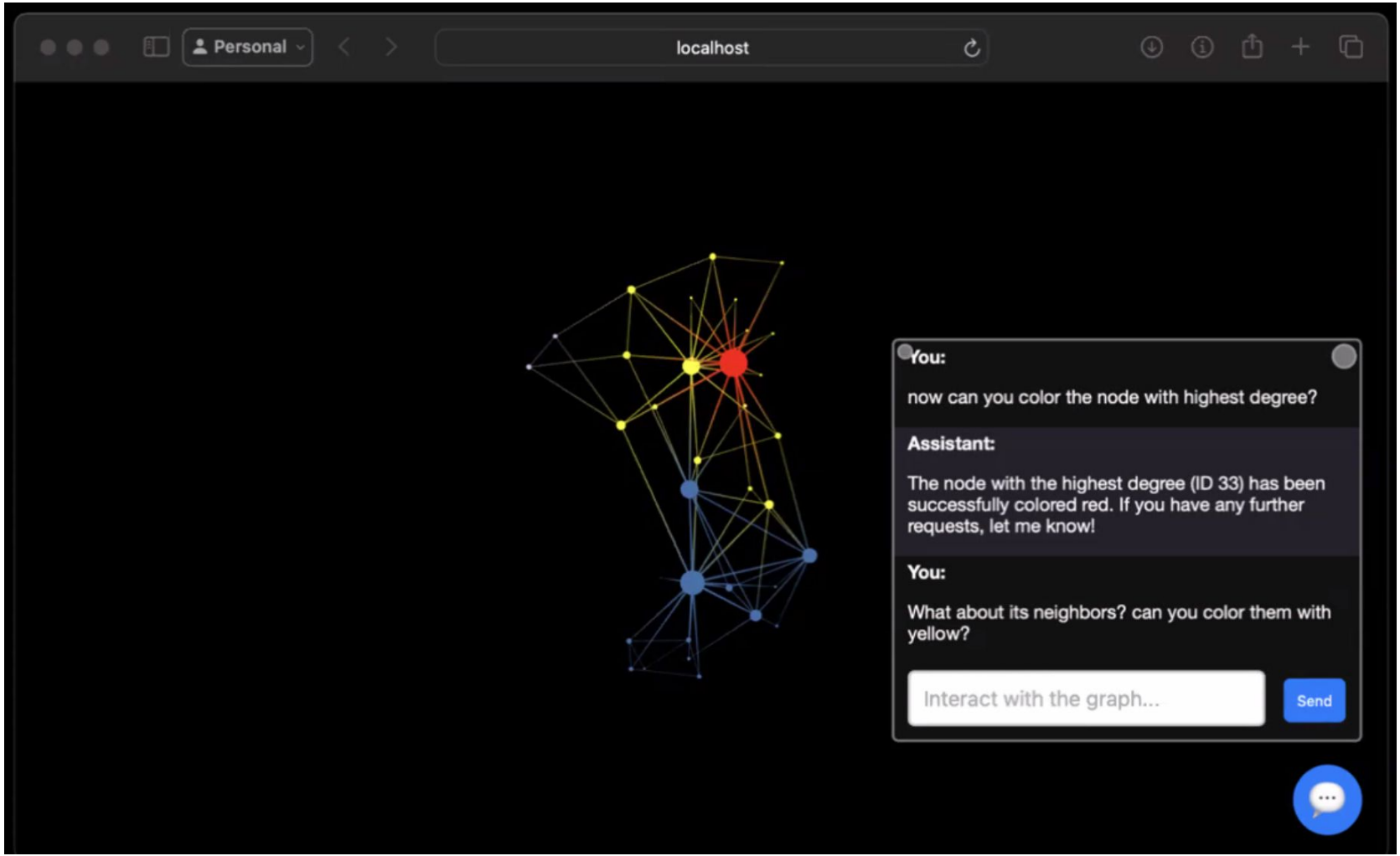
New collaborations

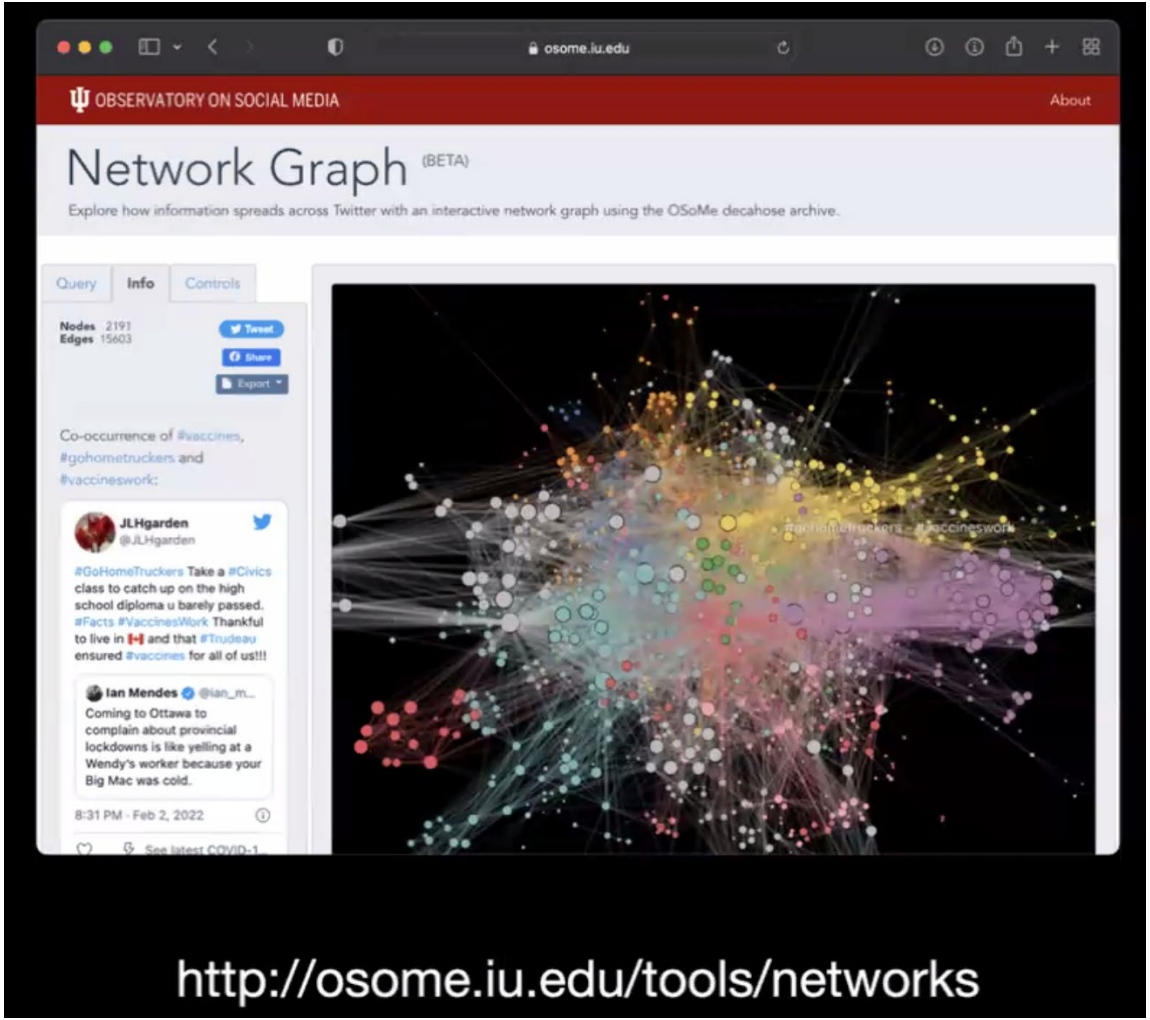




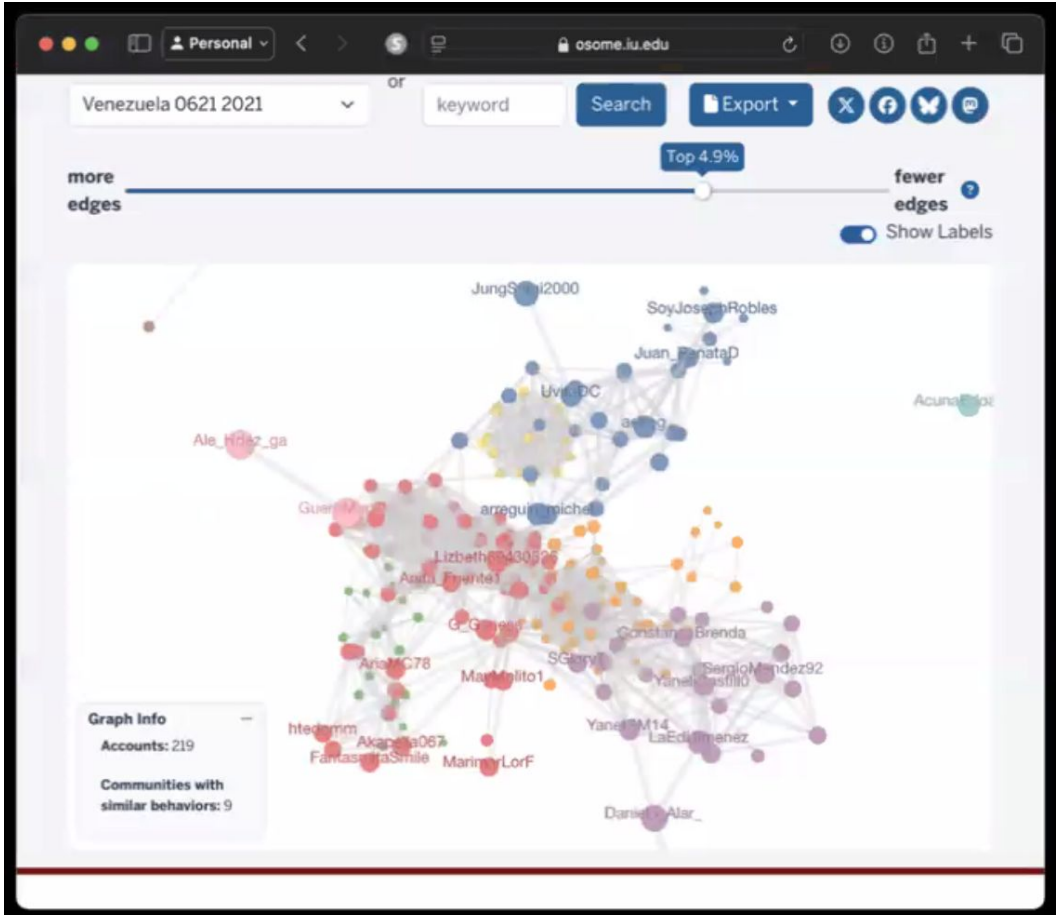





Personal
localhost
You:
now can you color the node with highest degree?
Assistant:
The node with the highest degree (ID 33) has been successfully colored red. If you have any further requests, let me know!
You:
What about its neighbors? can you color them with yellow?
Interact with the graph...
Send

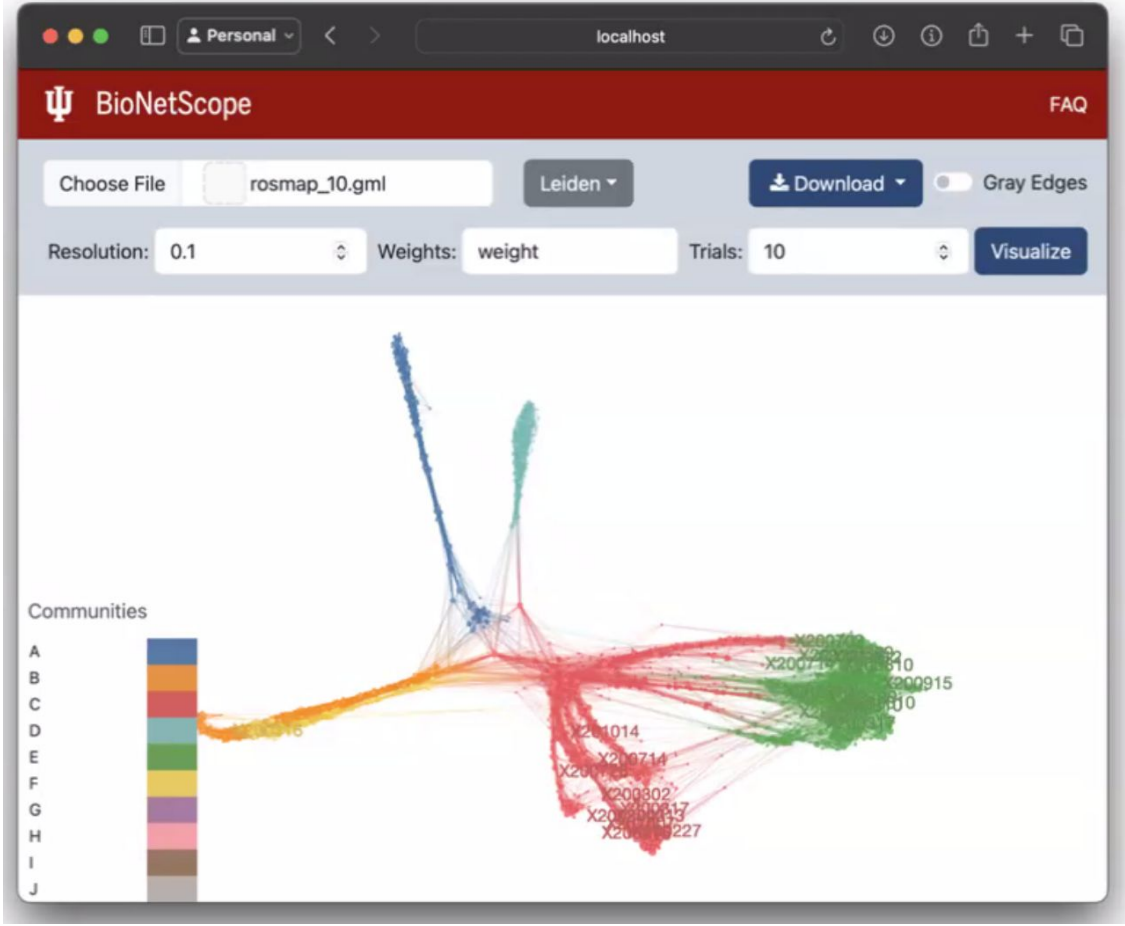


http://osome.iu.edu/tools/networks



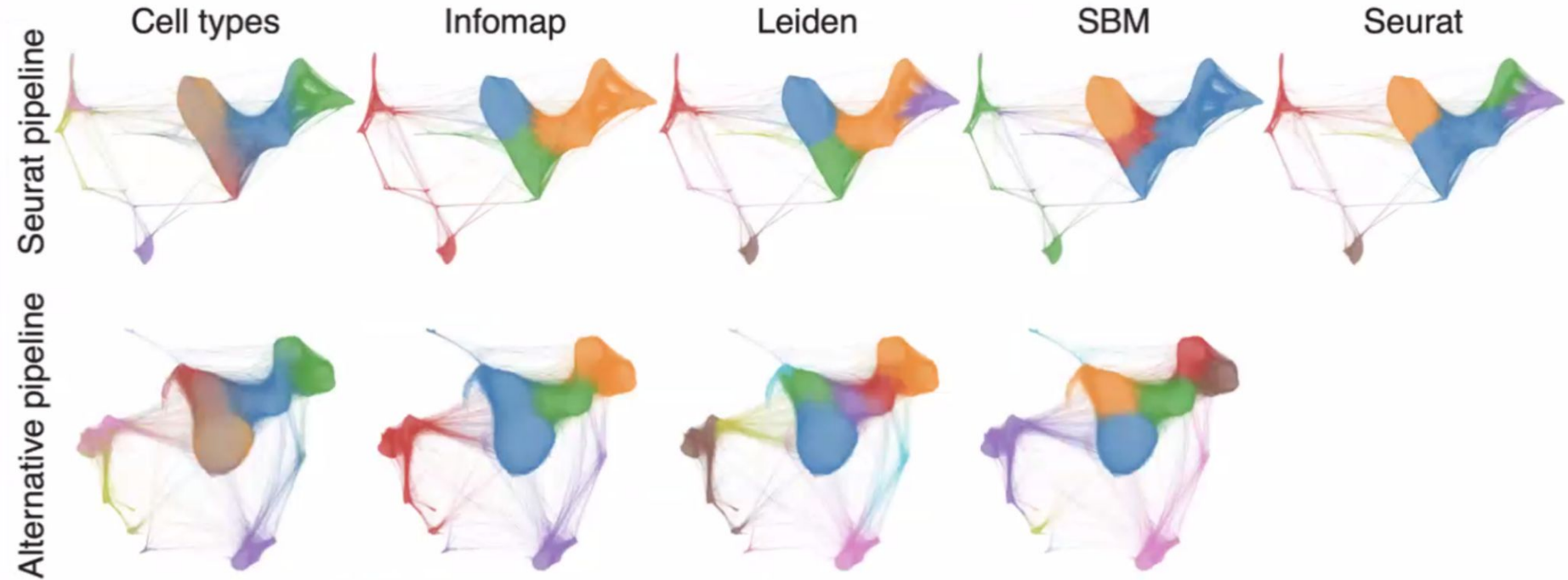
https://osome.iu.edu/tools/coordiscope

Adapting these models and tools for biomedical research



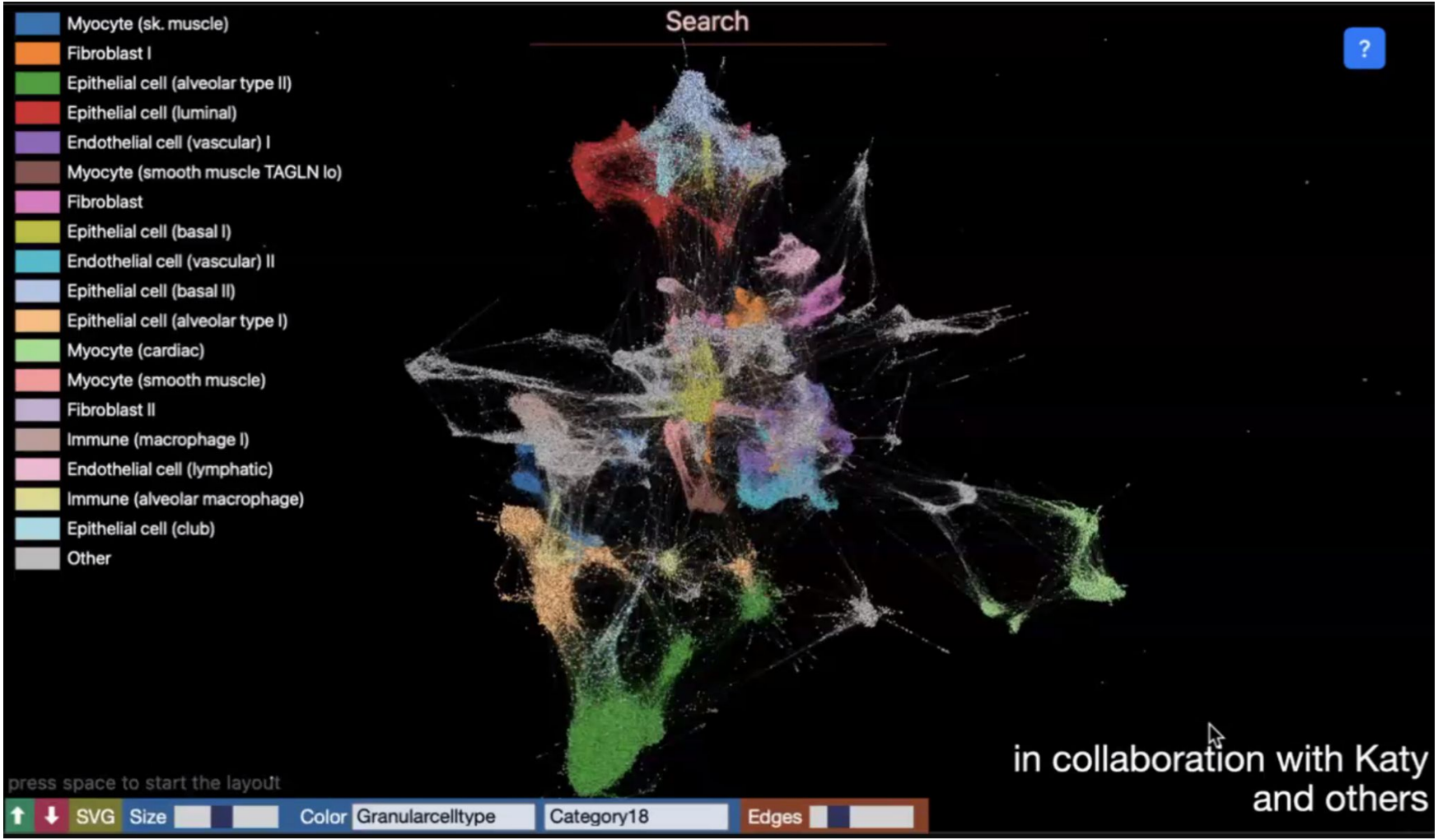
- Single cell RNA-seq data
- Microbiome associations with tissue gene expression
- Brain networks across age
- Metabolomics

Cell Type Differentiation Using Community Detection



68k human Peripheral Blood Mononuclear Cells (PBMCs) scRNA-seq dataset

Fatemi Nasrollahi, F. S., Silva, F. N., Liu, S., Chaudhuri, S., Yu, M., Wang, J., ... & Fortunato, S. (2024). Cell Type Differentiation Using Network Clustering Algorithms. bioRxiv, 2024-12.


Search
?
Myocyte (sk. muscle)
Fibroblast I
Epithelial cell (alveolar type II)
Epithelial cell (luminal)
Endothelial cell (vascular) I
Myocyte (smooth muscle TAGLN lo)
Fibroblast
Epithelial cell (basal I)
Endothelial cell (vascular) II
Epithelial cell (basal II)
Epithelial cell (alveolar type I)
Myocyte (cardiac)
Myocyte (smooth muscle)
Fibroblast II
Immune (macrophage I)
Endothelial cell (lymphatic)
Immune (alveolar macrophage)
Epithelial cell (club)
Other
press space to start the layout
SVG
Size
Color
Granularcelltype
Category18
Edges
in collaboration with Katy and others

Thanks

Indiana University U.S.

Staša Milojević
Yong-Yeol "YY" Ahn
Santo Fortunato
Filippo Menczer
Alessandro Flammini
Attila Varga
Lili Miao
Katy Börner
Andy Saykin
Thomas M O'Connell
Vijay R. Ramakrishnan
Filippo Radicchi

George Mason US

Henrique F. de Arruda
Sandro Reia

Univ. São Paulo Brazil

Diego R. Amancio
Osvaldo N. de Oliveira Jr.
Ana C. Medeiros

Binghamton U. U.S.

Sadamori Kojaku

Q&A

https://humanatlas.io/events/2024-24h

Questions

How do we define a Multiscale Human?

How do we map a Multiscale Human?

How do we model a Multiscale Human?

How can LLMs or RAGs be used to advance science and clinical practice?

Thank you