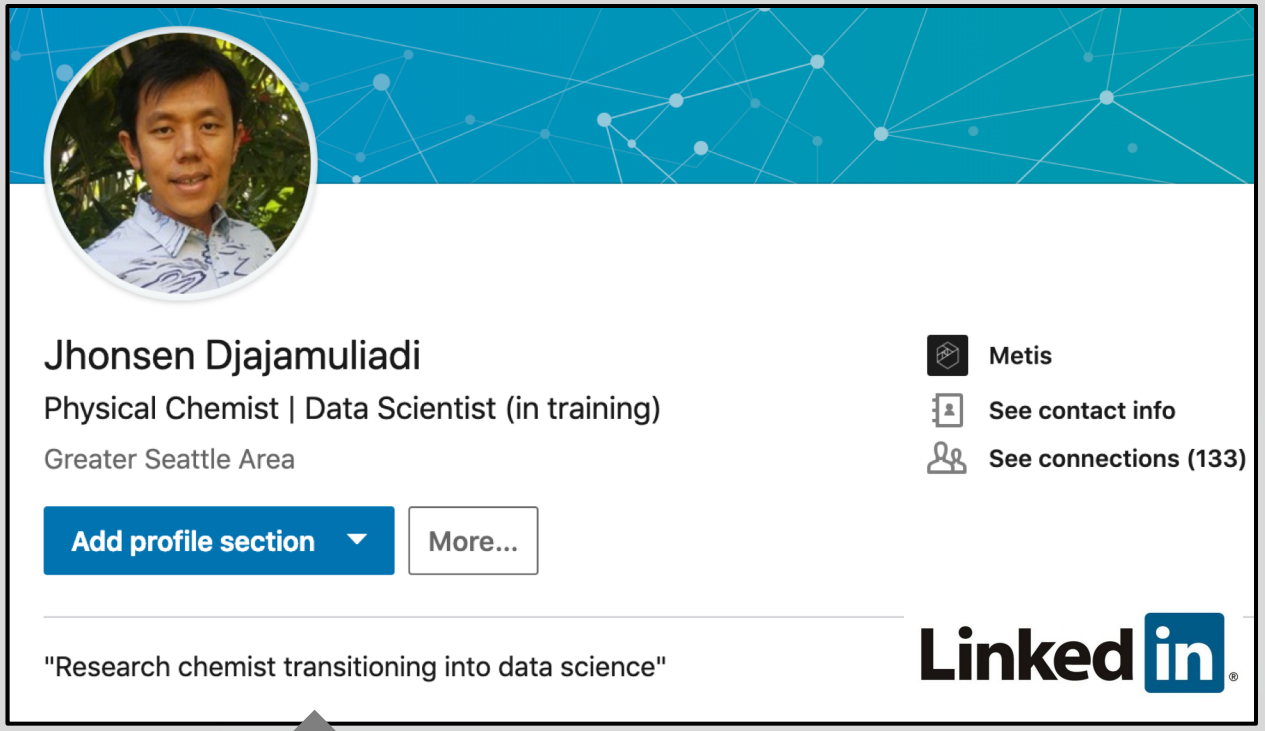
"Am I a Data Scientist?"

Finding My Identity Through Unsupervised Learning

Project-4 @ **METIS**

Jhonsen Djajamuliadi

My Journey in Becoming a Data Scientist



"Soft Skills"

- How should I describe myself?
- How should I network?
- How should I negotiate?

"Hard Skills"

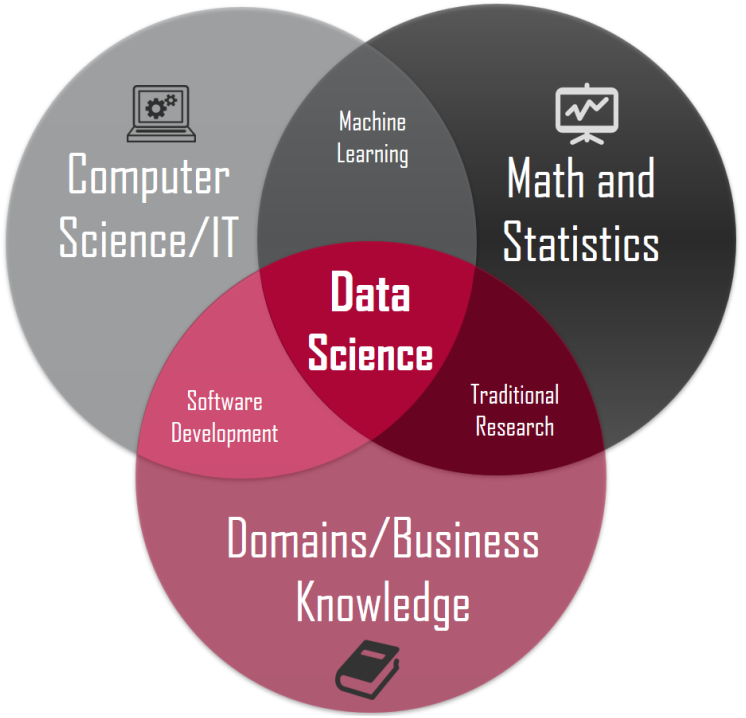


Figure from https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

Seeking My Identity in Wikipedia Articles

- *What is a **data scientist**?*
  - *Is it similar to other **scientists**?*

Data Collection

Natural Language Processing

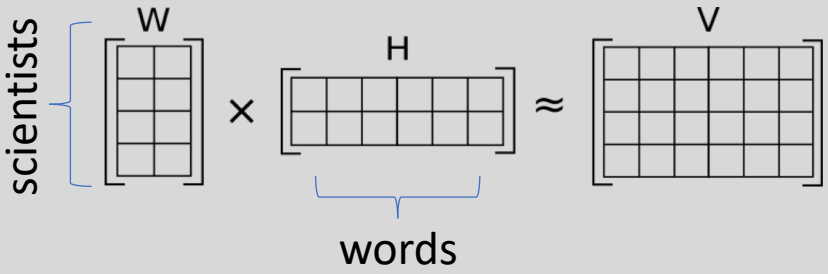
Unsupervised Learning

Deployed Application

Category: Scientists

~15,000 articles x 1.2 million words

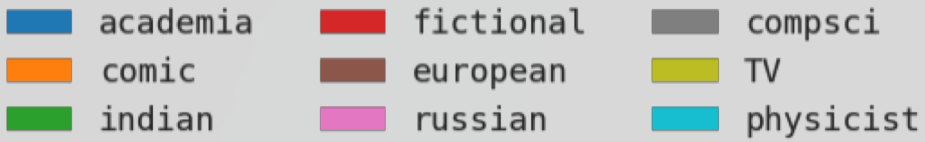
Non-negative **M**atrix **F**actorization



- ✓ Dimensionality reduction
- ✓ ***"Topics"*** in documents

Finding Topics in Summaries

Topic Modeling



university
professor
department
institute
director
born
research
science

soviet
russian
academy
member
sciences
academician
academy sciences
chinese

characters
star trek
trek
star
created
created star
fictional
created star trek

characters television
character
television
series
fictional
characters
characters television series
television series

scientist
computer
institute
director
computer science
computing
science
computer scientist

german
known
april
august
french
february
work
august
march

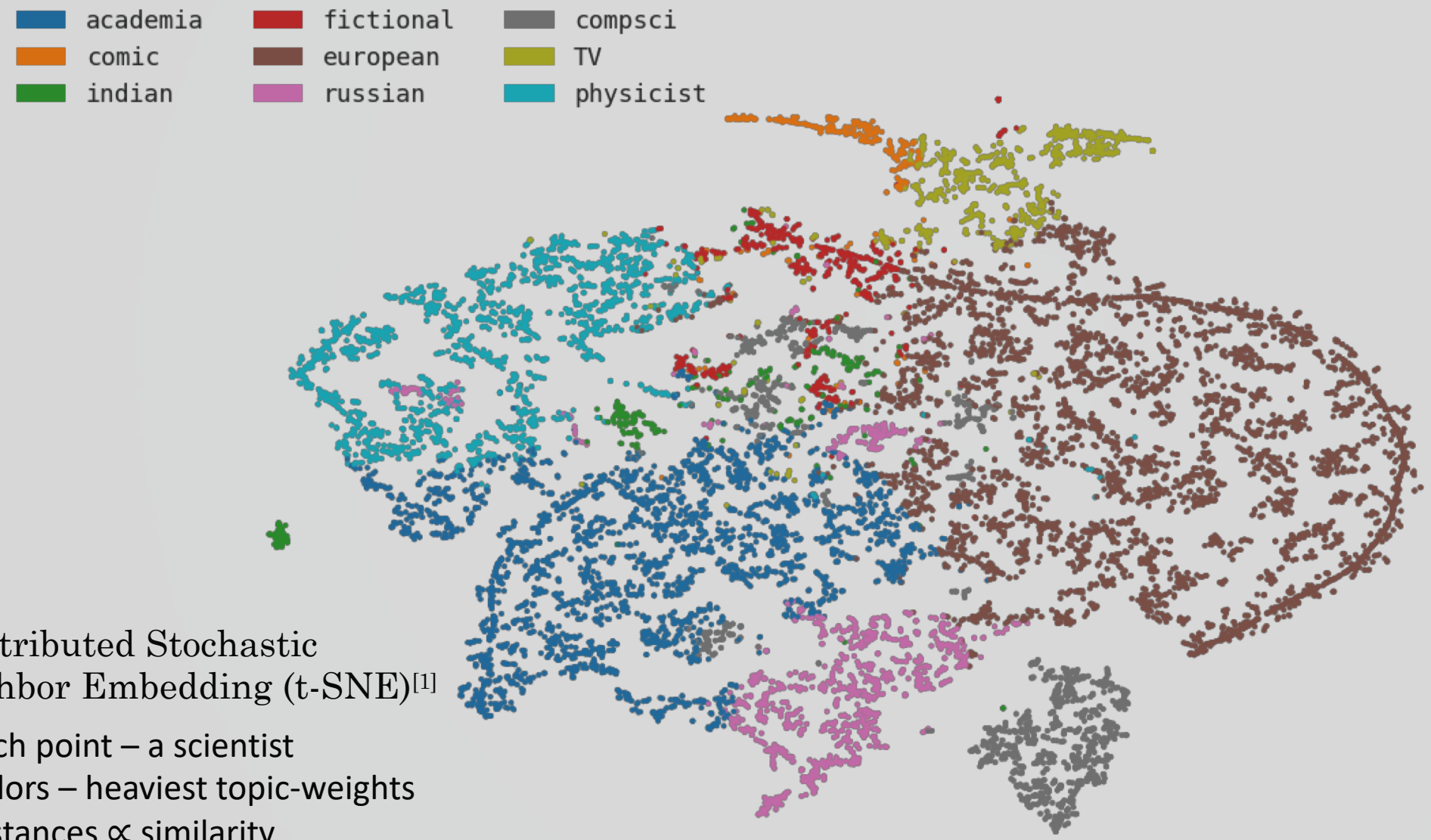
indian
science
academy
india
research
academy sciences
indian science
sciences
scientific

comic books published
comics
marvel
comic
books published
comic books
american comic books
appearing american comic

physics
theory
quantum
prize
nobel prize
theoretical
nobel
physicist

Finding Structure Among Scientists

Topic Modeling and 2D-Embedding with T-SNE



t-Distributed Stochastic Neighbor Embedding (t-SNE)[1]

- Each point – a scientist
- Colors – heaviest topic-weights
- Distances ∝ similarity

[1] T-SNE, https://lvdmaaten.github.io/tsne/

Finding Structure Among Scientists

Topic Modeling and 2D-Embedding with T-SNE

academia
comic
indian
fictional
european
russian
compsci
TV
physicist

Hadley Wickham

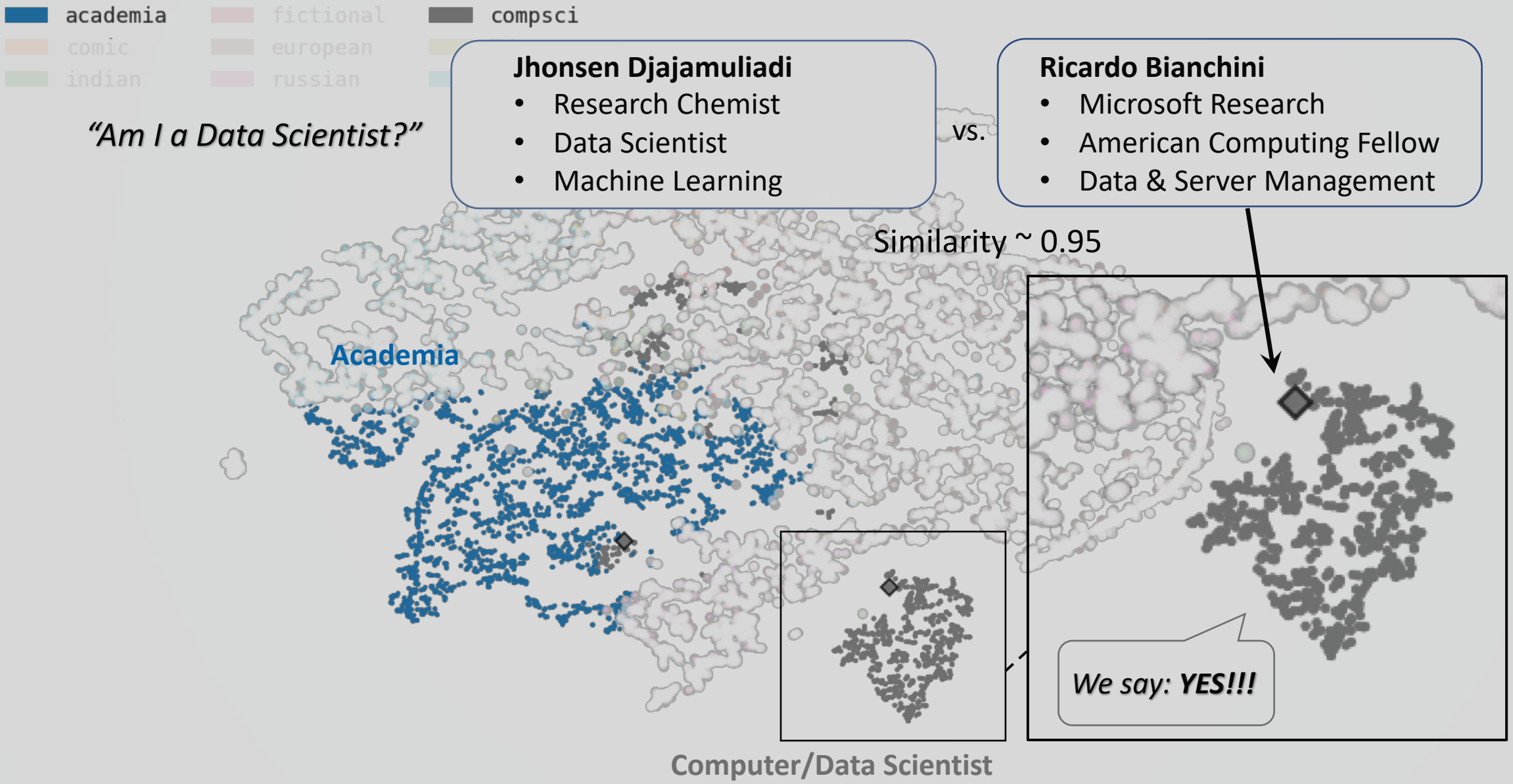
- Chief Scientist, Rstudio
- R Developer
- Adjunct Professor

Academia

Computer/Data Scientist

Finding (*Cosine*)-Similarity with Others

Topic Modeling and 2D-Embedding with T-SNE



Web Application in Development

Input:

- LinkedIn Summary

"Research chemist transitioning into data science"

I have an academic background in spectroscopy and molecular modeling, which are the science of "extracting signals out of the noise" and "computationally recreating or simulating their interactions".

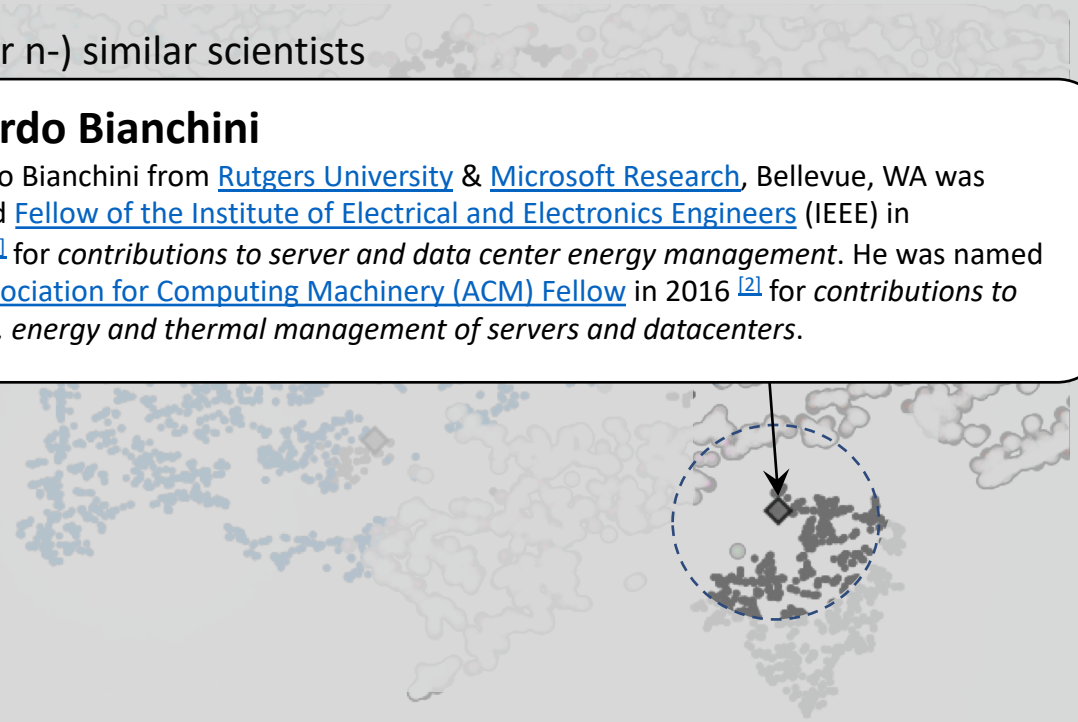
I love integrating experimental techniques with computational approaches, to find actionable insights and most consistent answer to research questions. "Hands-on" laboratory work is fun, but "in silico" computational projects have always peaked my interest. The latter has led me to this exciting field of data science. I'm currently taking a deeper dive into machine learning and AI for industrial applications, especially, in the biomedical and healthcare fields.

I'm also passionate about science communication, which is a way of "storytelling using data". I find it enjoyable to decompose technical concepts, and convey them to a mixed audience, e.g., scientists in different disciplines or others without scientific backgrounds.

Submit

Output:

- **Top 10** (or n-) similar scientists



Future Work:

- "Cleaner data", **LinkedIn**
- Target data science domain

"Am I a Data Scientist?"

Web Application in Development

Input:

- LinkedIn Summary

"Research chemist transitioning into data science"

I have an academic background in spectroscopy and molecular modeling, which are the science of "extracting signals out of the noise" and "computationally recreating or simulating their interactions".

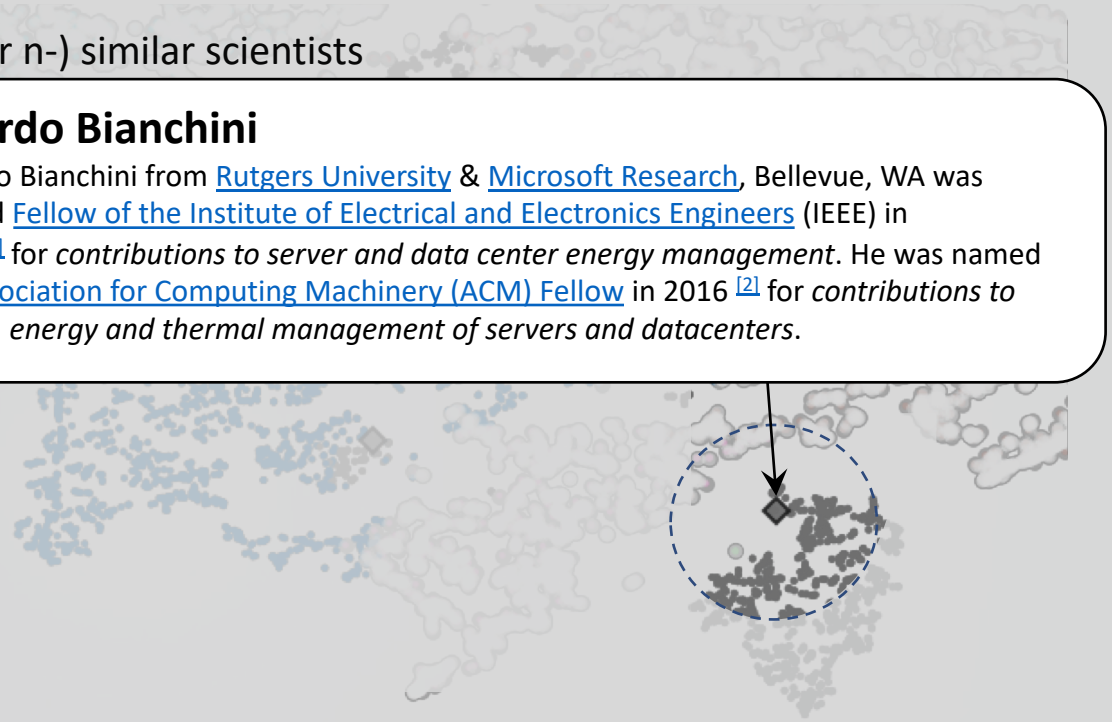
I love integrating experimental techniques with computational approaches, to find actionable insights and most consistent answer to research questions. "Hands-on" laboratory work is fun, but "in silico" computational projects have always peaked my interest. The latter has led me to this exciting field of data science. I'm currently taking a deeper dive into machine learning and AI for industrial applications, especially, in the biomedical and healthcare fields.

I'm also passionate about science communication, which is a way of "storytelling using data". I find it enjoyable to decompose technical concepts, and convey them to a mixed audience, e.g., scientists in different disciplines or others without scientific backgrounds.

Submit

Output:

- **Top 10** (or n-) similar scientists



Future Work:

- "Cleaner data", **LinkedIn**
- Target data science domain

"Are you a good fit?"

Application to Industry:

- Tool for company recruiters
  - *"New feature"* on **LinkedIn**

Candidates Employers

Thank You

Physical Chemist | Research Scientist | Data Analyst | Data Scientist

jhonsen

jhonsen

jhonsen_dj