Bio-geo analytics in the face of real data: sparse, heterogeneous, multi-dimensional – network analysis to the rescue?

Prof. Peter Fox

(pfox@cs.rpi.edu, @taswegian, #twcrpi, ORCID: 0000-0002-1009-7163)

Tetherless World Constellation Chair, Earth and Environmental Science/ Computer Science/ Cognitive Science/ IT and Web Science

Rensselaer Polytechnic Institute, Troy, NY USA

And many people to be named as we go…

C3DIS, Canberra, ACT, AUSTRALIA, May 7, 2019

What to expect…

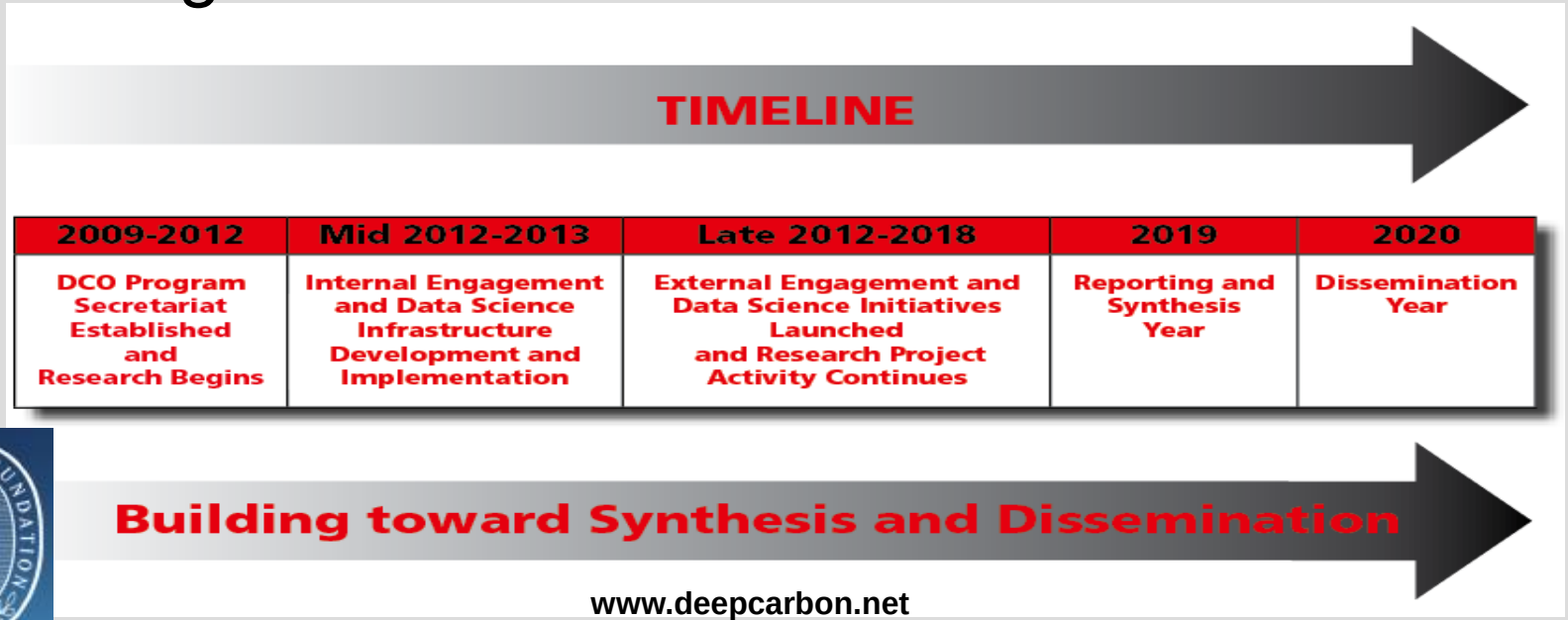
- Bio-geo sciences data – sparse, heterogeneous, …
- Network science
- Minerals and Fossils over <when> and <where>
- Serpentinization
- Lakes and water quality
- Data science as a socio-technical undertaking
- What is the needed future for network analytics?

DATA SCIENCE

Rensselaer

Deep Carbon Observatory (DCO) …

- "We are dedicated to achieving transformational understanding of carbon's chemical and biological roles in Earth."



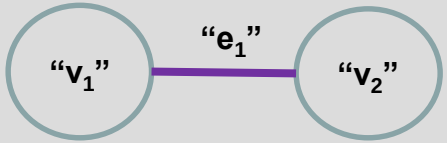
DCO Data Legacies: Geoscience Data Journal Special Issue, Oct. 2019

- Group 1: Super-sized Sample Inventory, Inventory of “Deep Carbon” Instrumentation, Inventory of “Deep Carbon” Field Sites
- Group 2: Census of Deep Microbial Life, Thermodynamic Parameters for High-Pressure, High-Temperature Physics and Chemistry Modeling, Global map of Carbonate Lithologies of Earth
- Group 3: Global Earth Minerals Inventory, Global Abiotic Fluid Distribution, Inventory of Dynamics and Physics of Deep Fluids, Inventory of Geochemical Models, Geo Sample Curation
- Group 4: Inventory of Diamonds with Inclusions + Derived products, Carbon Cycle, Flux of Volcanic Systems (magmatic, ...), State of High P and T Carbon and Related Materials

https://deepcarbon.net/page/dco-open-access-and-data-policies

Network Analysis

- Academic field which studies complex networks considering distinct elements represented by **nodes** (or vertices) and the connections between the elements or actors as **links** (or edges).



- Successfully applied in statistical physics, particle physics, computer science, biology, economics, finance, climatology and sociology. But **these methods had not been fully leveraged in many areas of Earth Science**.
- Using networks can help **view existing geological and biological information systems from a purely mathematical perspective, and infer new relationships** or new information about existing relationships.
- Cf. neural networks (stay tuned)

Courtesy: Eleish, Morrison, Prabhu et al.

What is encoded vs What is seen

Encoded	Seen/Inferred/Calculated
Nodes	Patterns in the Network Geometry
Links	Sub-Communities formed in the Network
Layout (Mostly Force Directed)	Important Hubs in the Network
Additional Parameters for Nodes (Optional) and Links	Additional metrics that explain the complexity of the environment (assortativity, betweenness, centrality etc.)

Comparison of how different networks change = understand the given environment. I.e. network “evolution”!

Mineral Networks (http://dtdi.carnegiescience.edu)

- Mineral occurrence data

mineral_name	locality_id	parent_id	mindat_id	is_bottom_level	is_meteorite	locality_name
Malachite	268	255	227207	False	False	Miramor District
Malachite	307	291	226717	True	False	Unnamed W Occurrence (1)
Malachite	313	310	227375	True	False	Chohe-Arusi Cu Occurrence (Chohe-Hrusi)
Azurite	313	310	227375	True	False	Chohe-Arusi Cu Occurrence (Chohe-Hrusi)
Malachite	315	310	227383	True	False	Ghuri-Safed Cu Occurrence
Calcite	315	310	227383	True	False	Ghuri-Safed Cu Occurrence
Azurite	315	310	227383	True	False	Ghuri-Safed Cu Occurrence
Malachite	340	336	227393	True	False	Rod-e Duzd Cu Occurrence (Rode-Duzd)
Azurite	340	336	227393	True	False	Rod-e Duzd Cu Occurrence (Rode-Duzd)
Magnesite	360	359	226925	False	False	Ajristan District
Magnesite	361	360	226583	True	False	Unnamed Magnesite Occurrence
Malachite	505	504	226850	False	False	Baghran District
Azurite	505	504	226850	False	False	Baghran District
Aragonite	515	504	226855	False	False	Garmser District (Garmsir District)
Malachite	533	520	227328	False	False	Shaida Cu-Zn Deposit
Azurite	533	520	227328	False	False	Shaida Cu-Zn Deposit
Azurite	536	533	227331	True	False	Shaida No. 3 Occurrence
Malachite	538	520	227338	True	False	Unnamed Cu Occurrence (1)
Azurite	538	520	227338	True	False	Unnamed Cu Occurrence (1)
Calcite	585	582	227337	True	False	Zinda Jan Baryte Occurrence (Zindajan; Zandadshon)
Malachite	609	587	226902	False	False	Khaki Jabar District (Khaki Jabbar District)
Azurite	609	587	226902	False	False	Khaki Jabar District (Khaki Jabbar District)

- Mineral coexistence matrix

	Chalcopyrite	Malachite	Chalcocite	Bornite	Azurite	Tetrahedrite	Covell
Chalcopyrite	25179	6949	3935	4376	3298	3654	32
Malachite	6949	11439	2920	2437	4603	1564	20
Chalcocite	3935	2920	5330	2468	1706	1034	19
Bornite	4376	2437	2468	5197	1414	1094	16
Azurite	3298	4603	1706	1414	5197	1071	12
Tetrahedrite	3654	1564	1034	1094	1071	5010	11
Covellite	3215	2089	1935	1695	1294	1122	37

- Symmetric adjacency matrix
- Rows and column names represent mineral species
- Values represent **co-occurrence** of 2 minerals

Mineral Networks – data structures

- Links

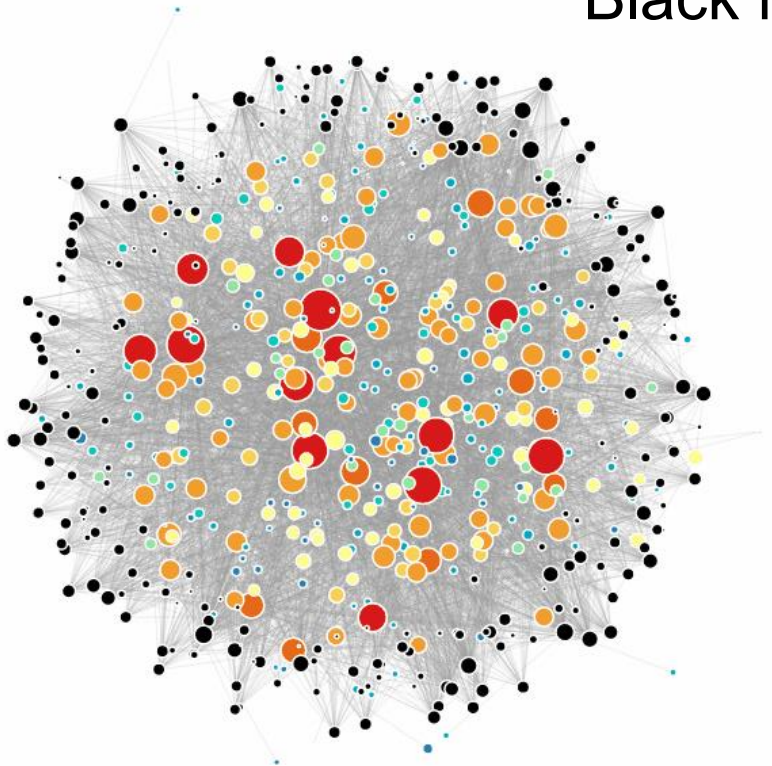
source	target	value
Abenakiite-(Ce)	Adamsite-(Y)	0.83333333
Agricolaite	Albrechtschraufite	0.83333333
Adamsite-(Y)	Alexkhomyakovite	0.85714286
Adamsite-(Y)	Alstonite	0.94444444
Alexkhomyakovite	Alstonite	0.85714286
Aerinite	Alumohydrocalcite	0.96
Alloriite	Alumohydrocalcite	0.83333333
Alstonite	Alumohydrocalcite	0.97468354
Abenakiite-(Ce)	Ancylite-(Ce)	0.83333333
Adamsite-(Y)	Ancylite-(Ce)	0.83333333
Agricolaite	Ancylite-(Ce)	0.85714286
Albrechtschraufite	Ancylite-(Ce)	0.83333333
Alexkhomyakovite	Ancylite-(Ce)	0.85714286

- Nodes

Name	Name (plain)	Cu redox	color	luster	hard
Abswurmbachite	Abswurmbachite	2	black	1	6.5
Agaite	Agaite	2	blue	3	2.5
Agardite-(Ce)	Agardite-(Ce)	2	green	2	3.5
Agardite-(La)	Agardite-(La)	2	green	2	3.5
Agardite-(Nd)	Agardite-(Nd)	2	green	2	3.5
Agardite-(Y)	Agardite-(Y)	2	green	2	3.5
Aikinite	Aikinite		black	1	2.5
Ajoite	Ajoite	2	blue	2	
Aktashite	Aktashite		black	1	3.5
Aldridgeite	Aldridgeite	2	blue	2	3
Algodonite	Algodonite		gray	1	4
Allochalcoselite	Allochalcoselite	1,2	brown	3	3.5

Carbon mineral bi-partite network

Colour nodes=carbon mineral species
Black nodes=”locations”



- R
  - igraph
  - ggnetwork
  - Network
  - SNA
  - d3Network
- JS
  - D3js
  - Threejs

Two types of network metrics:

Local

(few nodes)

- How “important” is one node?
- Does one node “communicate” between two distinct groups?

Global

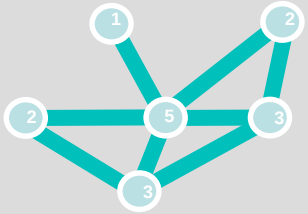
(entire network)

- Is the network highly interconnected?
- Does the network form distinct groups or clusters?

Please make a note: metrics focus on dominance and uniformity, some nodes or all … is this sufficient?

Metrics: local (left) and global (right)

Degree is the number of links connected to a given node.

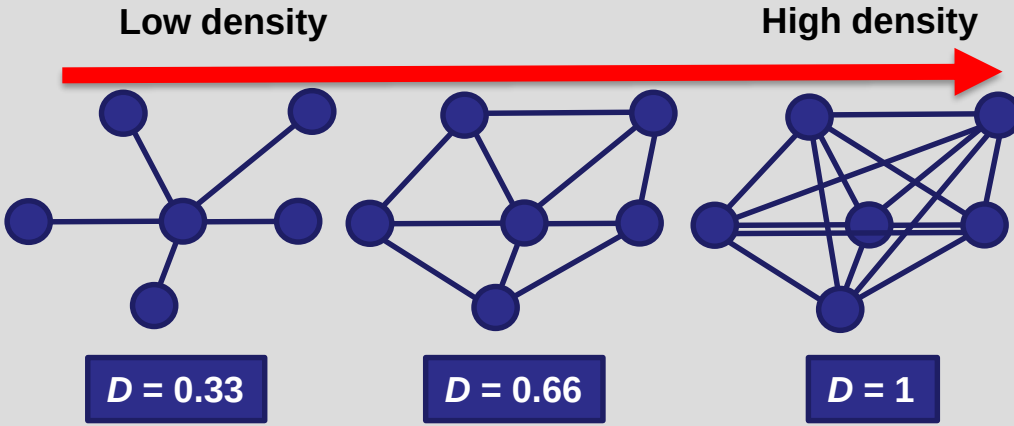


Distance is the geodesic (shortest) between any two nodes.

Betweenness is a measure of the number of geodesic paths that pass through a given node.

Density, *D*, is the no. of links divided by the no. of possible links

$$D = \frac{2L}{N(N-1)}$$



Diameter: largest geodesic distance in a network (the shortest path between the two most separated nodes)

Mean Distance: average “degree of separation” in a network

Metrics: Global

Centralization:

A measure of how central a network's "most central" node is relative to how central all the other nodes are.

(really important for terrorist and social networks ;-)

- Degree centralization: number of links to each node
  - Are there many highly interconnected nodes?

- Betweenness centralization: number of shortest paths through each node
  - Are there a few key "broker" nodes?

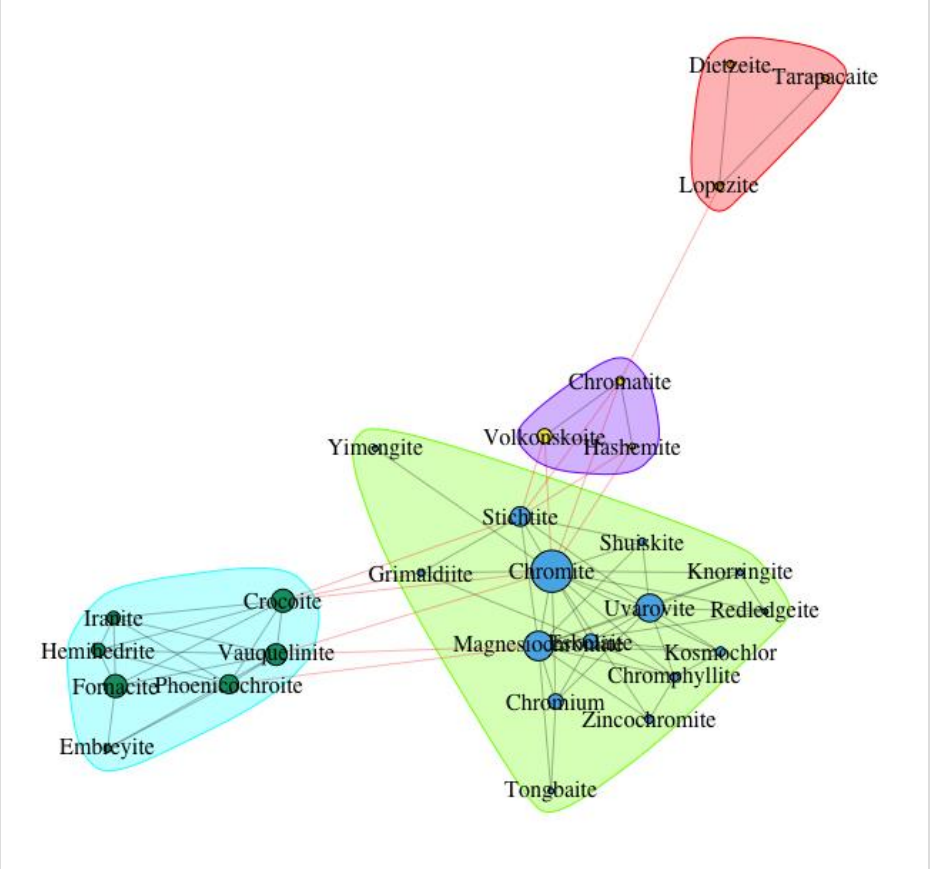
Mineral Global Metrics

Metric	 Earth	 Mars	 Moon	 Vesta
Density	0.64	0.27	0.69	0.78
Maximum Network Diameter	2	3	3	2
Mean Network Diameter	1.36	1.69	1.26	1.22
Degree Central.	0.34	0.62	0.23	0.22
Betweenness Centralization	0.02	0.09	0.03	0.02

Mineral Co-occurrence and “clustering”

Walktrap Community Detection

- Groups identified to correspond to Paragenetic Mode, i.e. how and when the Minerals were formed.
- They also represent a network!



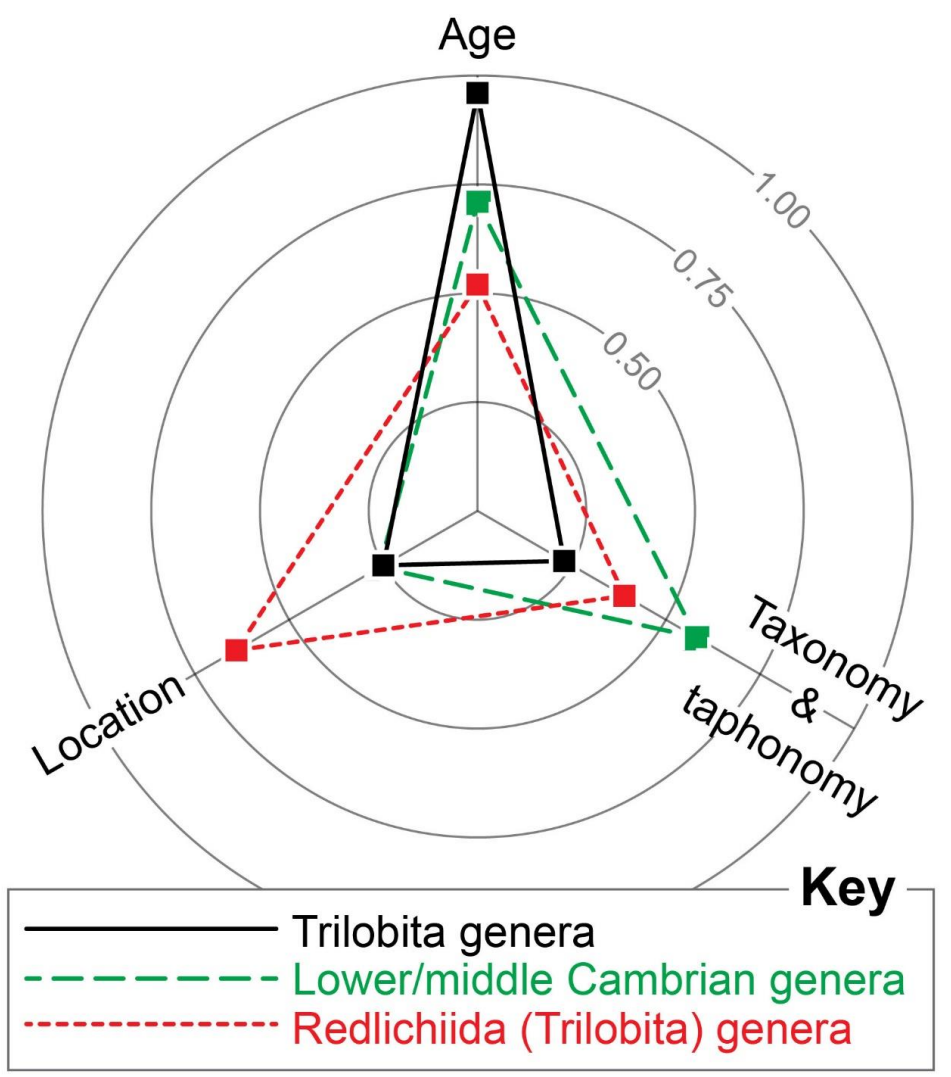
Morrison SM, Liu C, Eleish A, Prabhu A, Li C, Ralph J, Downs RT, Golden JJ, Fox P, Hummer DR, Meyer MB, and Hazen RM (2017) Network analysis of mineralogical systems. American Mineralogist 102

Paleontology

Assortativity (Homophily)

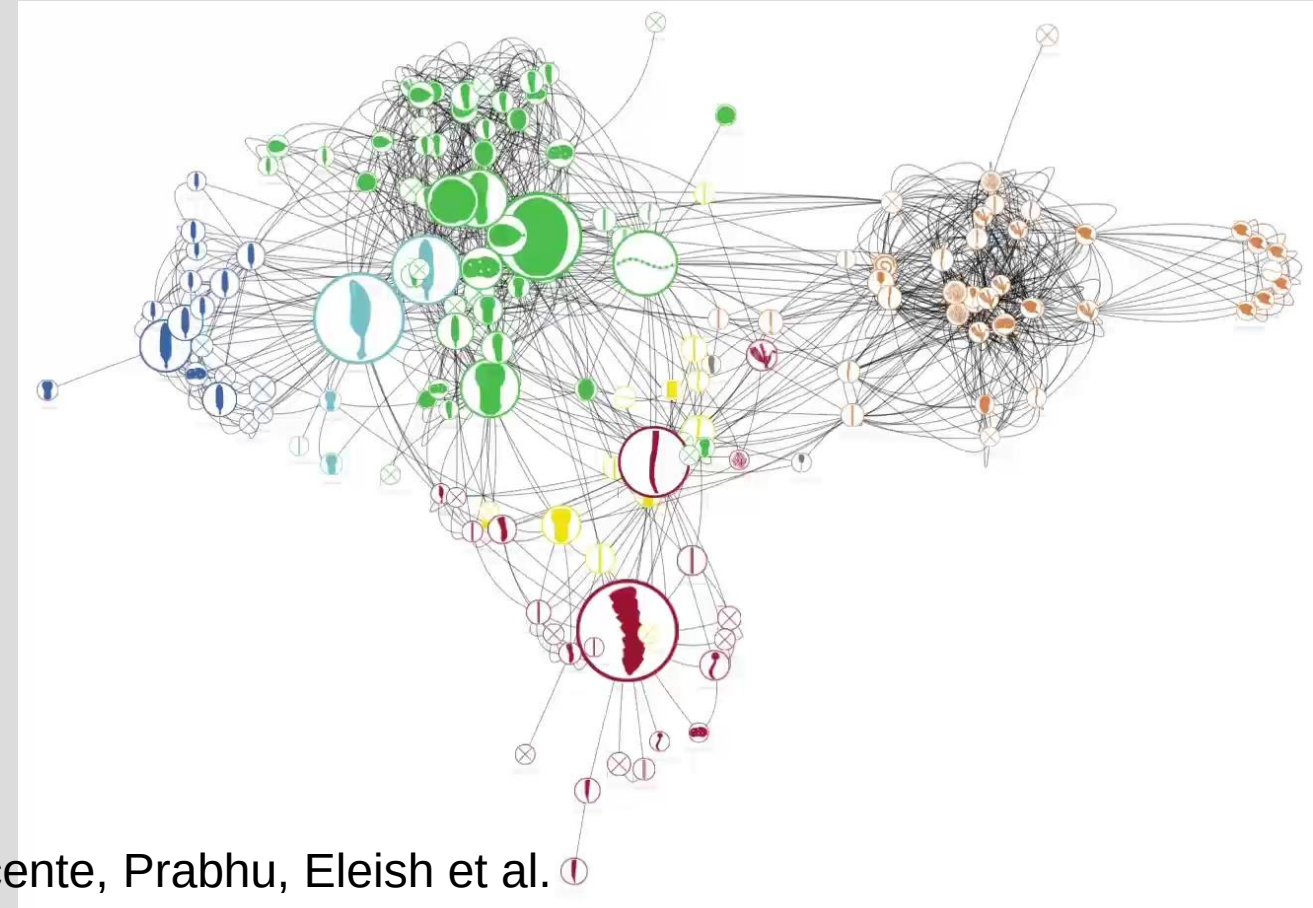
Network equivalent of Pearson correlation coefficient

Values between 1 & -1
1 = similarity favors connections
0 = non-assortative
-1 = opposites attract



•Muscente AD, Prabhu A, Zhong H, Eleish A, Meyer M, Fox P, Hazen R, and Knoll A (2017) The network paleoecology of mass extinctions. (Proceedings of National

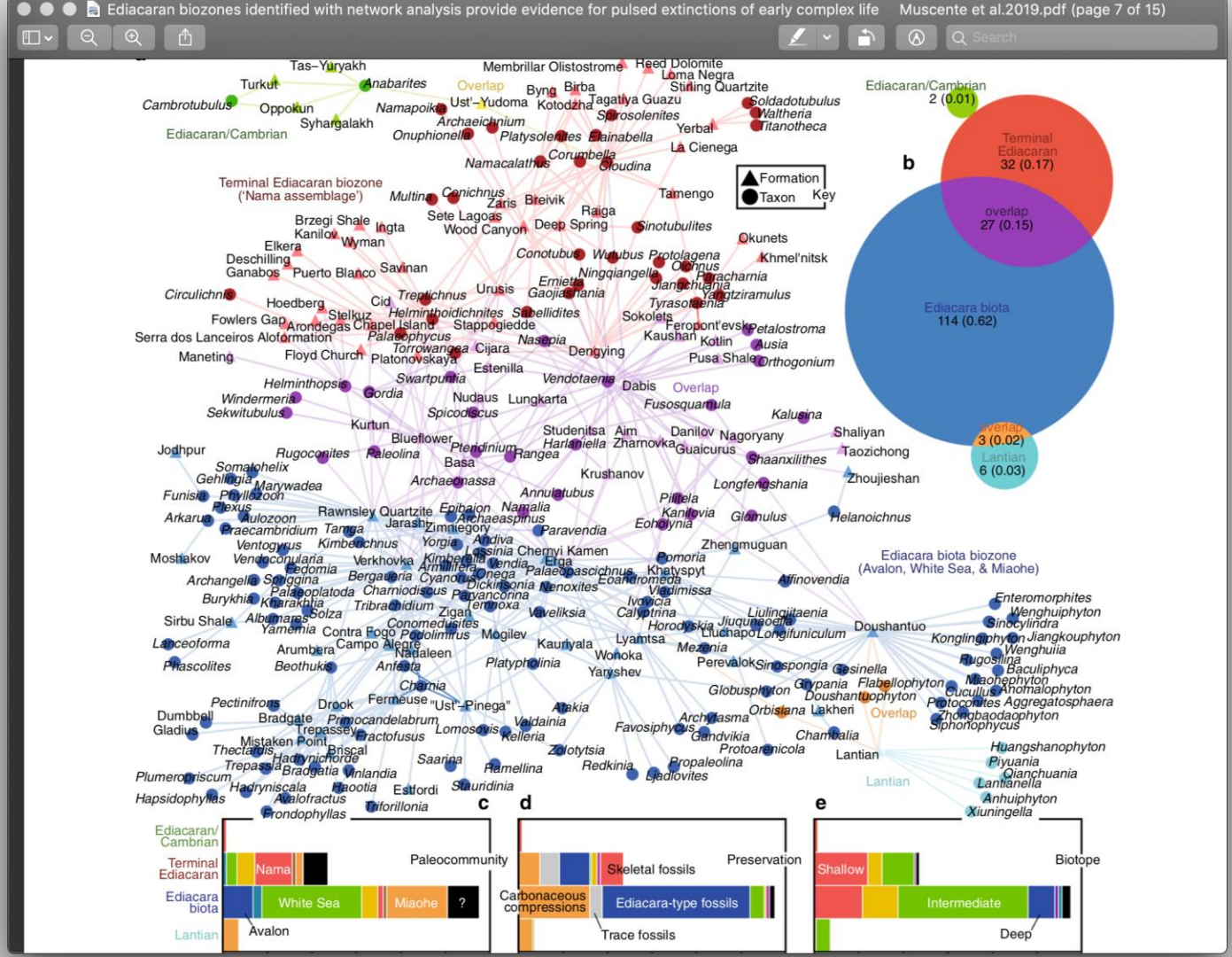
Networks and time, and ..



Courtesy: Muscente, Prabhu, Eleish et al.

Pulsed extinction in Ediacaran marine fossils

Muscente et al. 2019 Nature Comm .



Muscente (Nature Comm.): "Nama and White Sea fauna are different facies, whereas a mass extinction occurred after the Avalonian." – science hypothesis

Serpentinization and hydrocarbon formation from experiments…

Fang Huang [a], Samuel Barbier [b,c], Muriel Andreani [b], Renbiao Tao [b], Jihua Hao [b], Osama Minhas [a], Kathy Fontaine [a], Peter Fox [a] and Isabelle Daniel [b]

a. Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, 12180; b. Univ Lyon, UCBL, Ens Lyon, CNRS, Laboratoire de Géologie de Lyon UMR 5276, Villeurbanne, France; c. Total CSTJF, Avenue Larribau, F-64018 Pau, France;

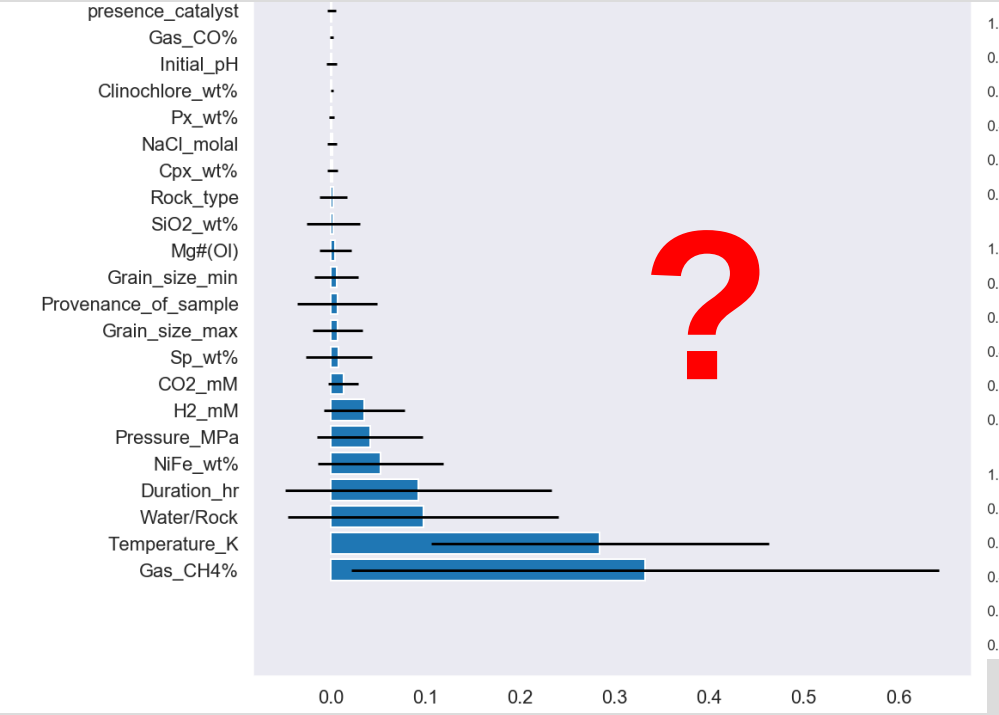
- Serpentinization is a hydrous alteration of ultramafic rocks in hydrothermal systems, which generates H2 and organic species, and can potentially contribute to the origin of life.
- In past decades, serpentinization has been extensively studied in labs, producing methane in highly variable amounts.
- Such a large experimental variability could not be explained => retrieval of “data” from experimental literature == meta-data science!
- **We have used random forest and network analysis**

mber of ividual eriments	Temperatur e / Pressure	Duration at each analyses	Rock type (not precise if less than 5% different)	Mineralogy composition	Initial knows catalyst (s)	Degree of alteration	Grainsize	Surfac e area	Blank or mineral free	Solution composition	Initial pH	Mass solids and liquid	W/R ratio	Gase compositio n	Origin of the carbon source	Reactor type	Reactor composition	State of initial phases adds	Intermediate sampling
N°	**Kelvin / MPa**	**Hours**	Name or description, and origine	Olivine (Mg#(Ol)), Pyroxene (Cpx, Opx), Spinelle, Magnetite, Chromite, Serpentine, Brucite, SiO2, Talc, oxides	Spinel, Magnetite, Alloys ... **yes or no**	Fresh, Medium, or Highly alterated	*min and MAX* ***µm***	**XXXX cm2 / gram of rock**	**Yes or No**	NaCl, CO2 CH4 CO (dissolved), NaHCO3, HCOOH, KCL, CaCl2, MgCl2,... **mol-mmol/kg**	X	X **g**	X	N2, CO2, H2, Ar,... **%**	Solid, liquid and or gases phase(s)	Flexible, Parr type, Glass bottle,...	Gold, Titanium, alloys, Teflon, Borosilicates, ...	Solids, and/or Liquids, and/or Gases	**Yes or No**

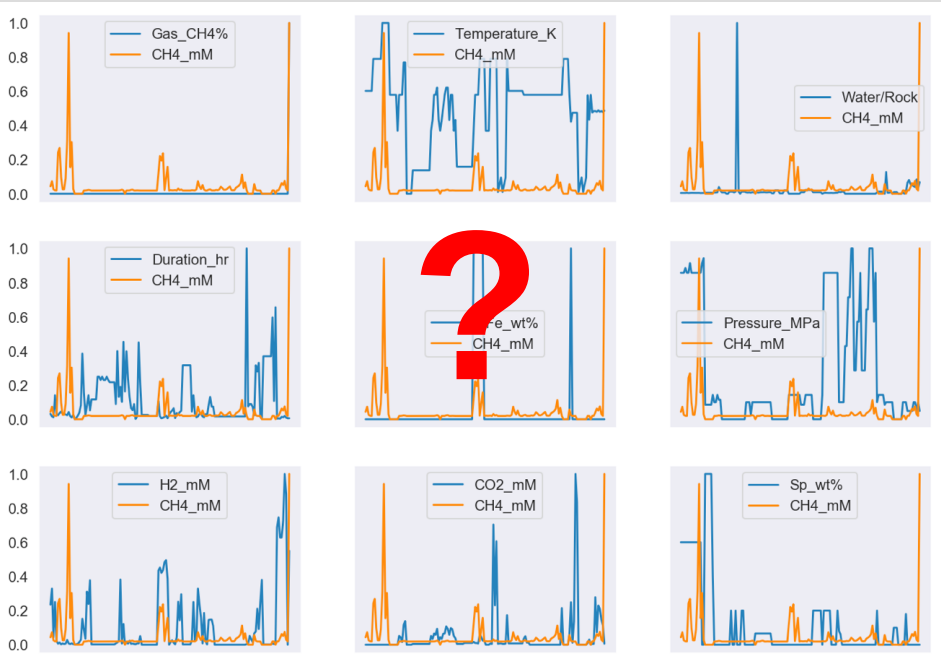
Random Forest Results

We use **all** experimental parameters as model features and the amount of CH4 production as model target to train a random forest model. The model gives importance scores to each features (see below).

Top factors influencing CH4 production

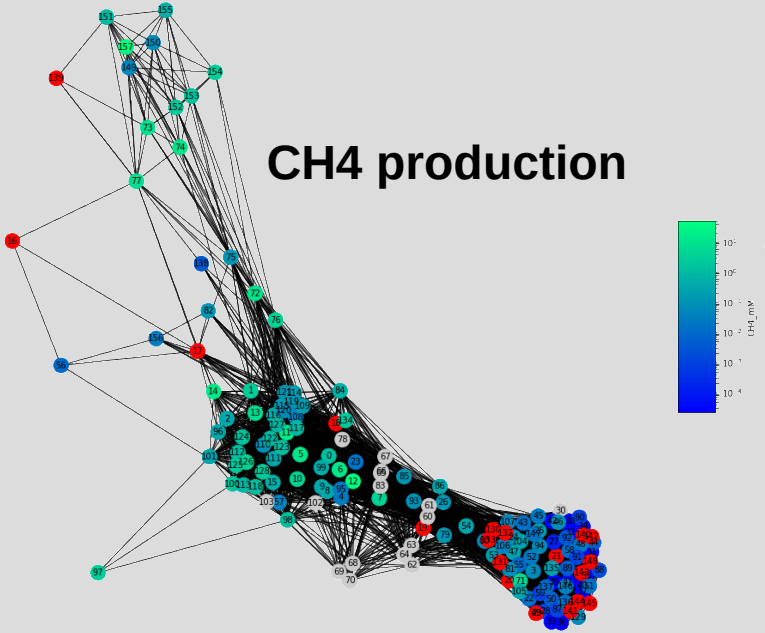


Correlations of individual features and CH4 production

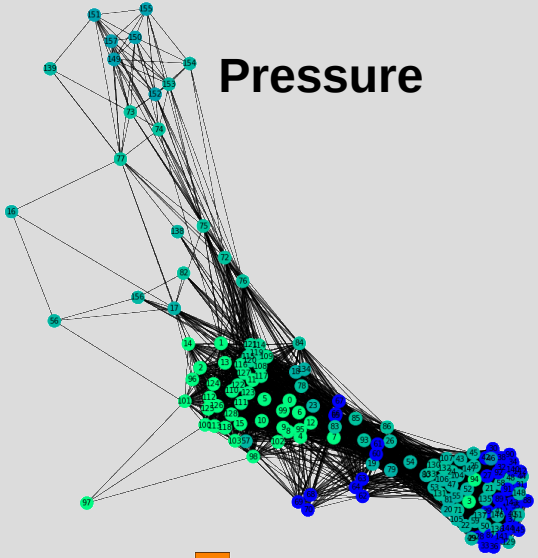


Network Analysis Results

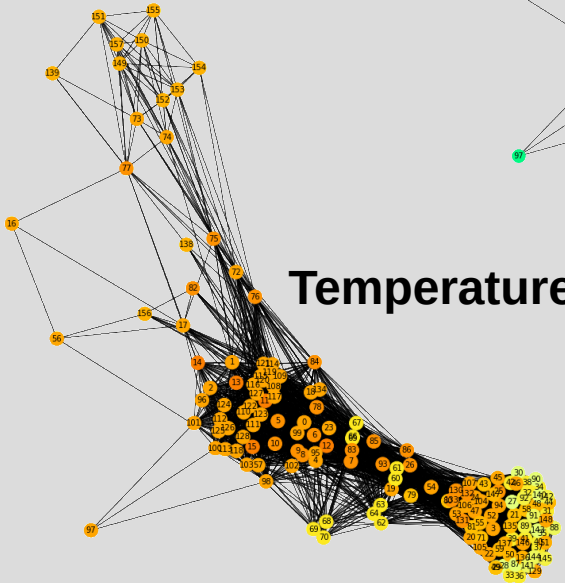
Network built by calculating the cosine similarity of each experiment with all features, including experimental parameters and the amount of hydrocarbon produced.



CH4 production



Pressure

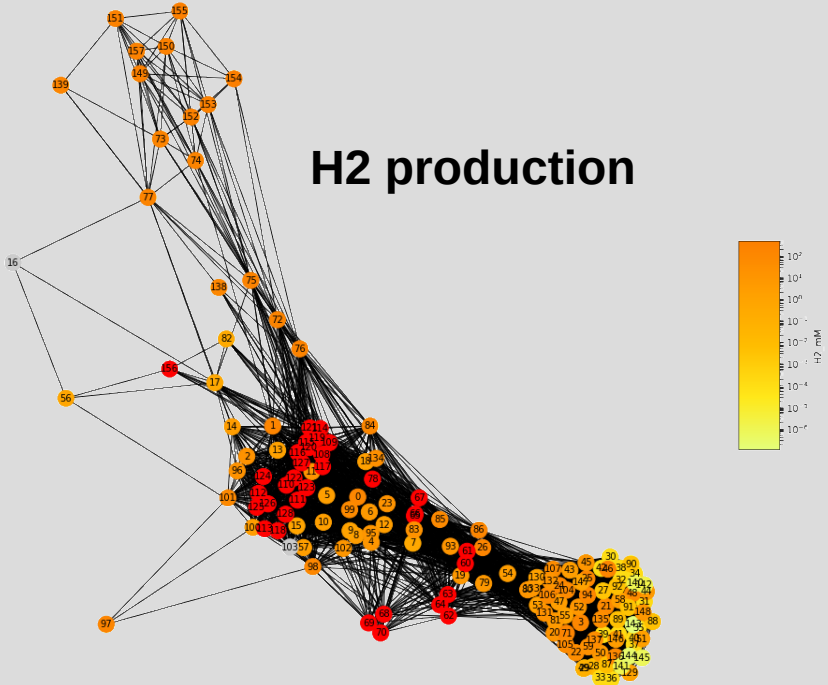


Temperature

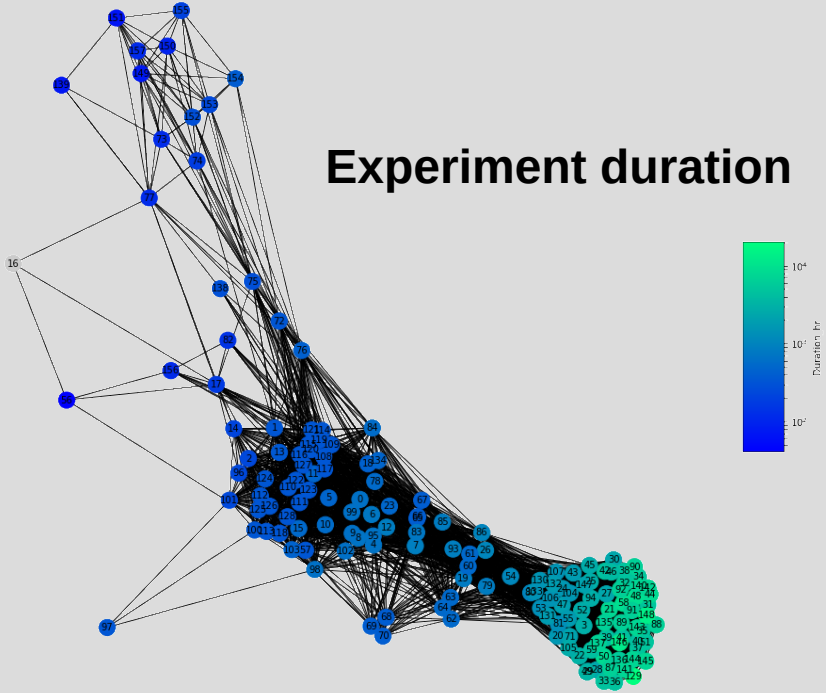
DATA SCIENCE

Network Analysis Results

Network built by calculating the cosine similarity of each experiment with all features, including experimental parameters and the amount of hydrocarbon produced.



H2 production



Experiment duration

Lake research *challenges*

Pis: Kevin Rose, Peter Fox (RPI), Harry Kolar, Mike Kelly (IBM)
PhD: Ahmed Eleish, Anirudh Prabhu

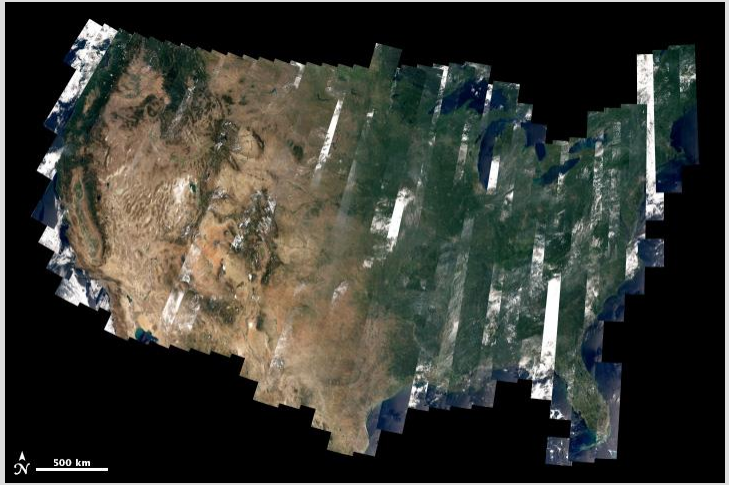
Spatially -
What watershed features most influence water quality? Does spatial configuration matter? And, by extension, how can competing interests between land use demands and freshwater quality be optimized?

Temporally -
What watershed features best predict long-term changes in water quality? Does spatial configuration matter?

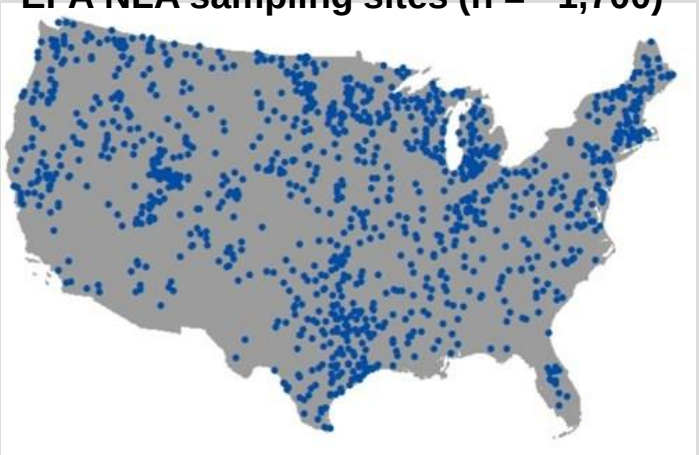
Data

- Multispectral Landsat satellite imagery
  - ~2500 images of US lakes and their watersheds over 2013-2018, but goes back decades for previous satellites.
  - 9 bands per image.
  - 30m pixel size, 16 day return time.
- Water quality measurements
  - *In situ* measurements including water clarity, algal biomass, dissolved organic carbon.
  - Many other physical, chemical, and biological measurements.
- All data sources are publicly available.

First Landsat 8 composite image of the continental US.



EPA NLA sampling sites (n = ~1,700)

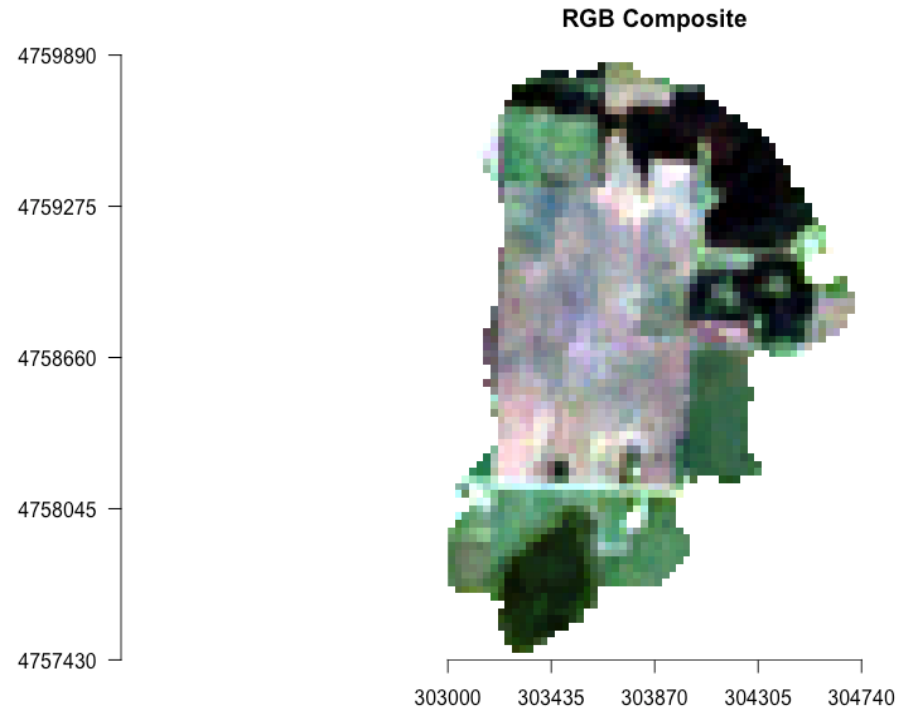


Our approach

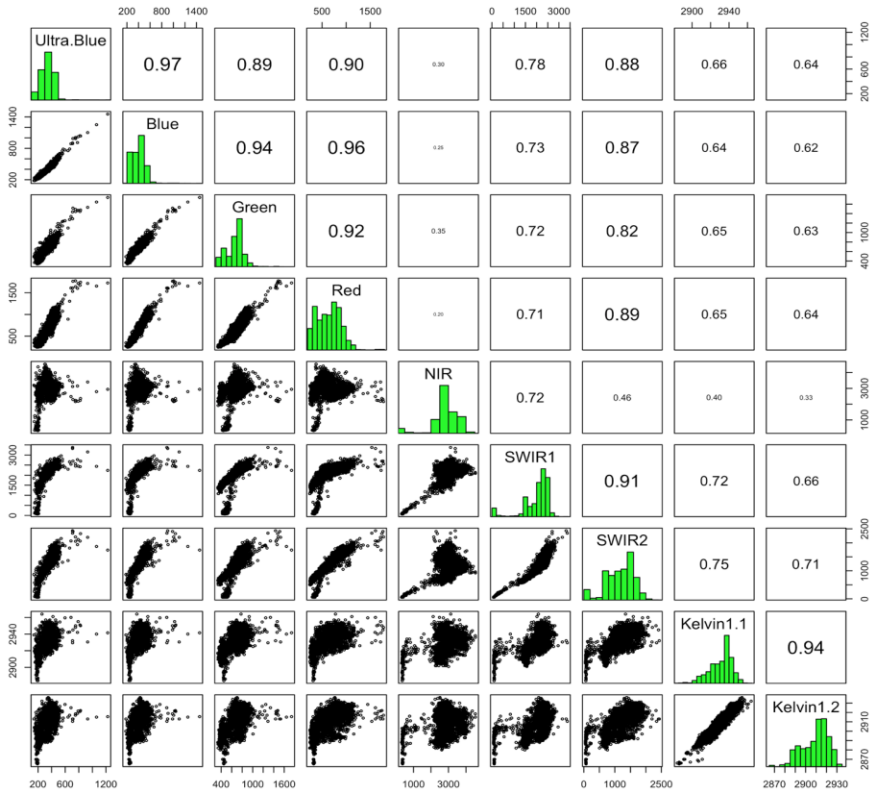
- Satellite imagery processing
  - Compute mean, median, and variance image for each lake across time interval.
  - Using median images compute remote sensing indices, e.g. normalized difference vegetation index (NDVI), normalized difference water index, and modified soil-adjusted vegetation index.
- Analysis
  - Examine patterns in **relationships** between various features.
  - Using identified trends to construct and test various machine learning models that **relate** data extracted from the satellite imagery to water quality measurements.

Lake Barney, Wisconsin

Lake and watershed RGB composite

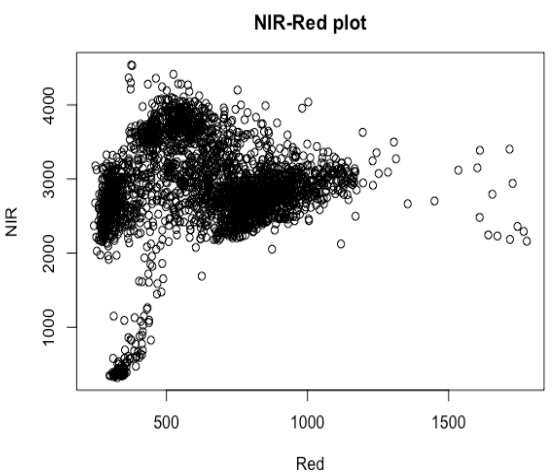


Waveband cross-correlation

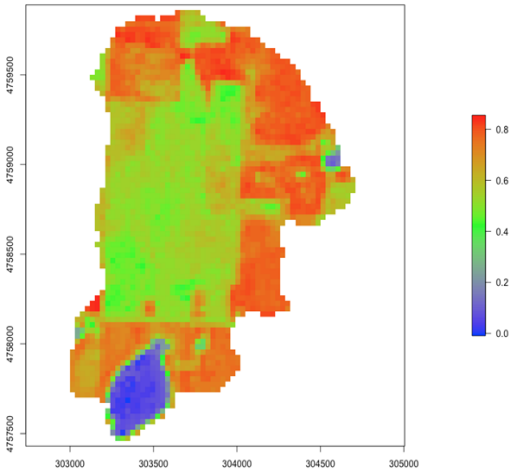


Lake Barney, Wisconsin

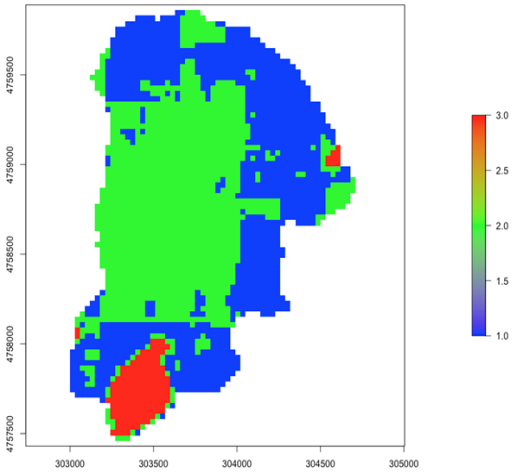
- Normalized Difference Vegetation Index

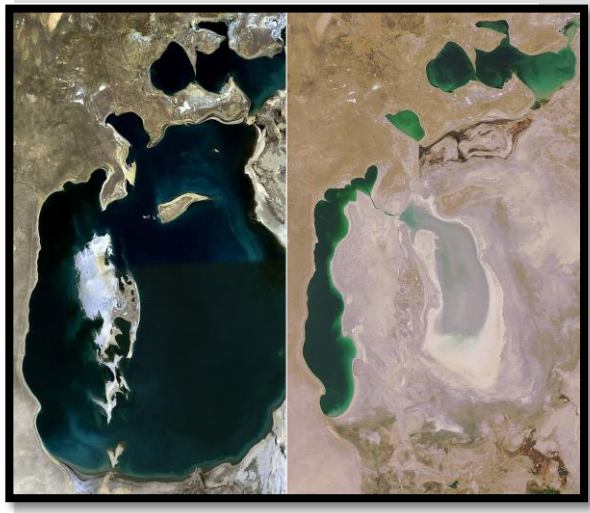


NDVI



K_{means} clusters (k=3)

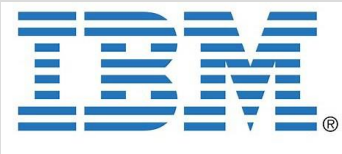




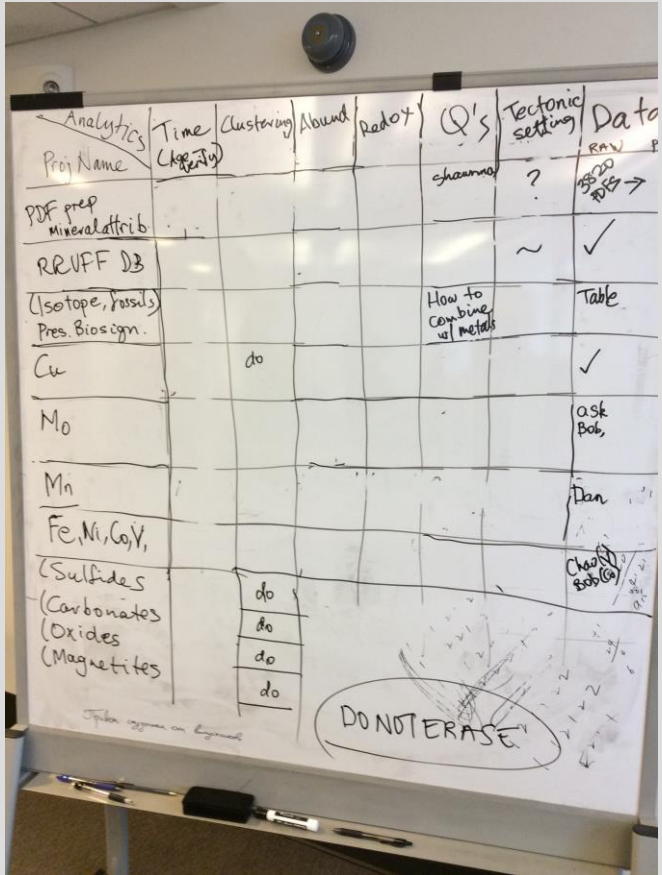
NDVI may be a useful approach to water feature detection..

Of course: network analysis…

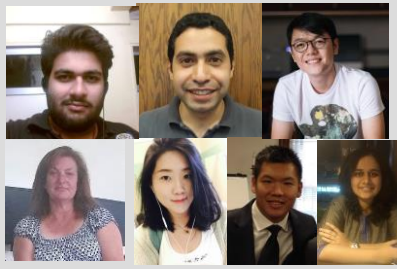
This page is intentionally left blank.

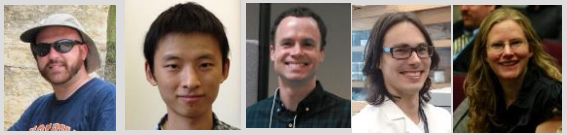


Data Science as-a Socio-Technical System



Schematic for Deep Time Virtual Laboratory

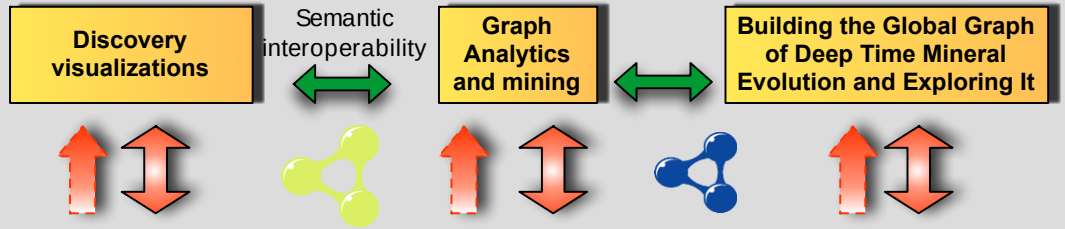




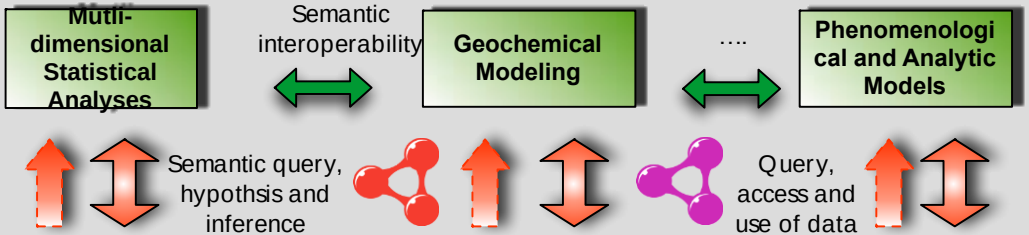
Integrated Applications

Software, Tools & Apps

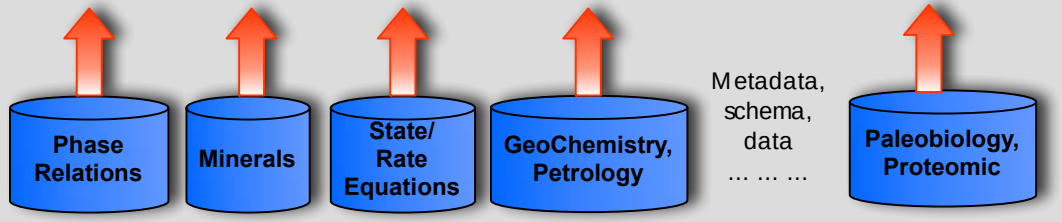
Data Repositories

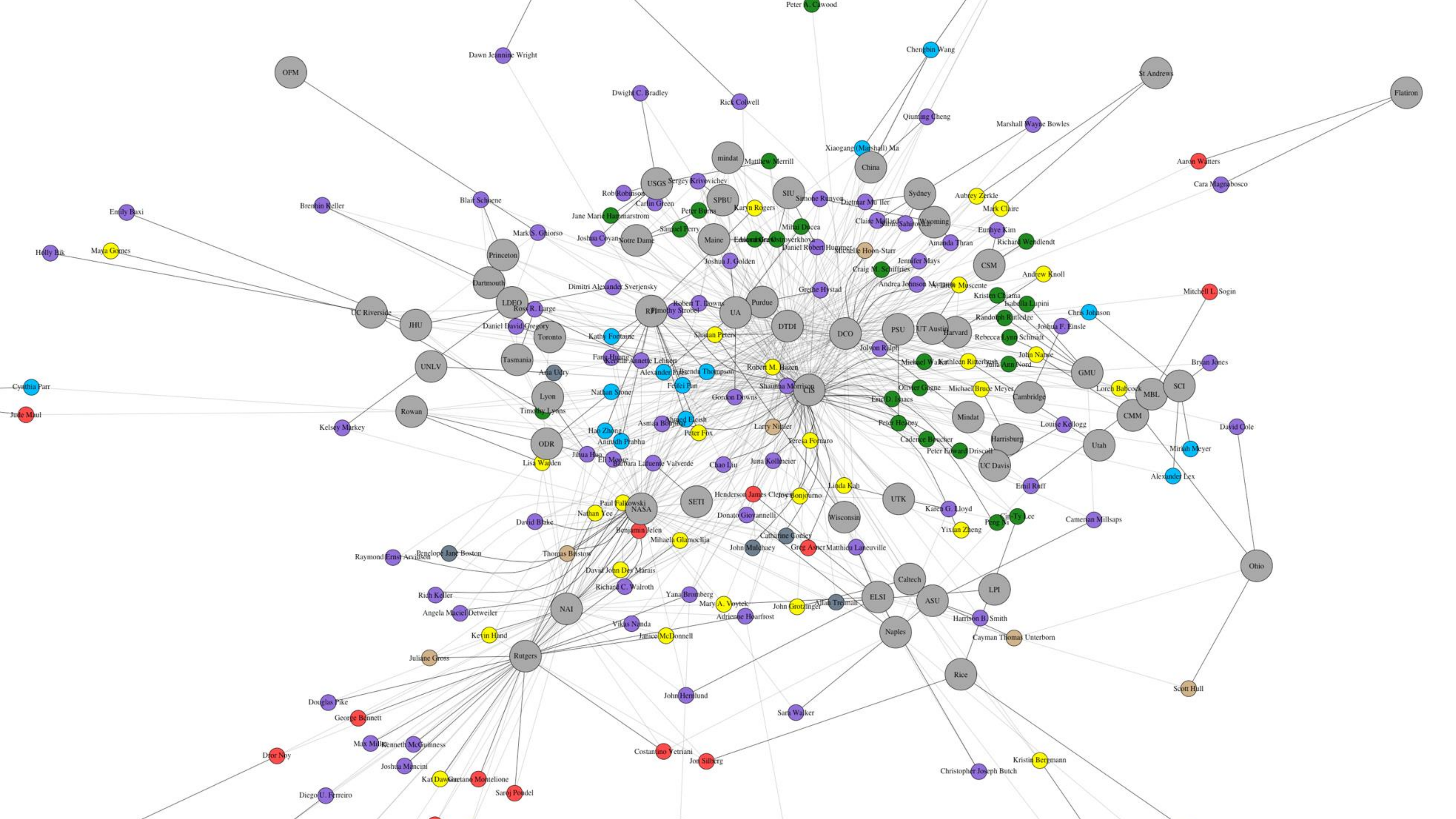


Application-level mediation: vocabulary, mapping to science and data terms, graph algorithms, hypothesis formulation, testing



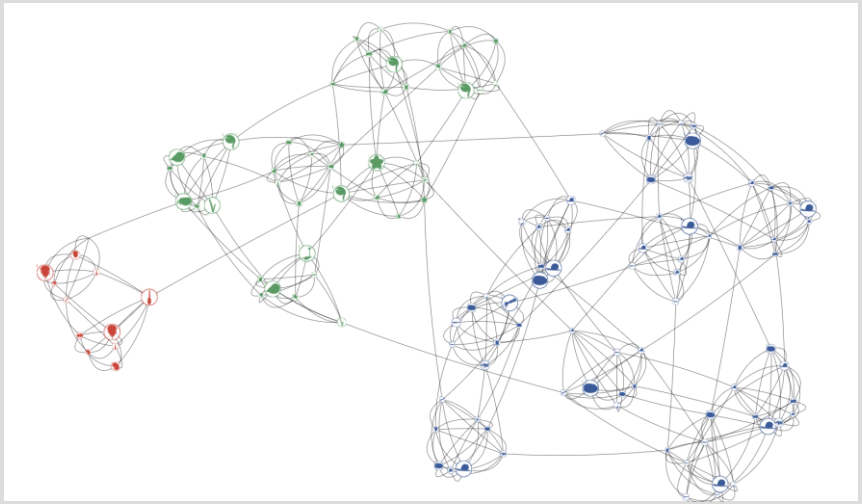
Semantic mediation: physics, chemistry, mineral, composition data, into graphs



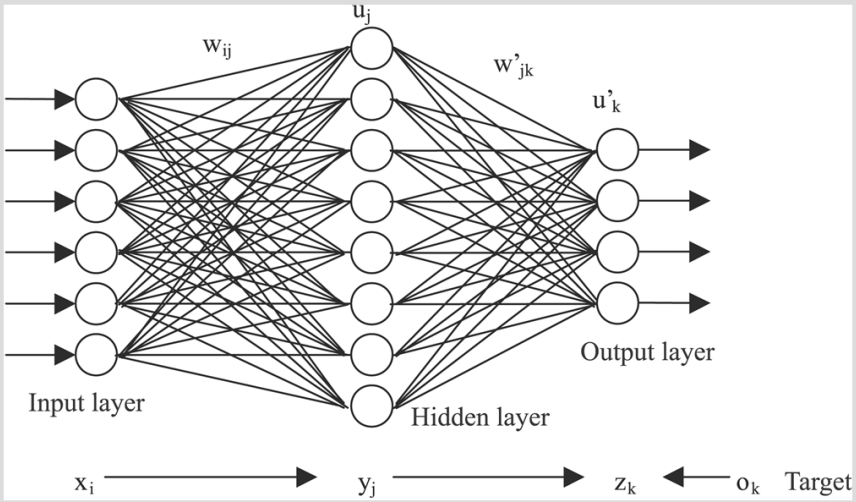


Peter A. Cawood
Chengbin Wang
Dawn Jeannine Wright
OFM
St Andrews
Flatiron
Dwight C. Bradley
Rick Colwell
Qiuming Cheng
Marshall Wayne Bowles
Xiaogang (Marshall) Ma
mindat
Matthew Merrill
China
Aaron Watters
Cara Magnabosco
USGS
Sergey Krivovichev
Rob Robinson
SIU
Sydney
Aubrey Zerkle
Carlin Green
SPBU
Simone Runyon
Dietmar Müller
Mark Claire
Brennin Keller
Blair Schoene
Jane Marie Hammarstrom
Peter Burns
Karyn Rogers
Emily Baxi
Samuel Perry
Mihai Ducea
Wyoming
Eunhye Kim
Mark S. Ghiorso
Joshua Goyan
Notre Dame
Maine
Richard Wendlendt
Amanda Thran
Holly Rik
Maya Gomes
Princeton
Daniel Robert Hummer
Michelle Hoon-Starr
Joshua J. Golden
Jennifer Mays
CSM
Craig M. Schiffries
Andrew Knoll
Dartmouth
Andrea Johnson Mangum
Dimitri Alexander Sverjensky
Grethe Hystad
Kristen Chiama
Mitchell L. Sogin
LDEO
Robert T. Downs
Isabella Lupini
UC Riverside
Ross R. Large
RPI
Timothy Strobel
UA
Purdue
Chris Johnson
JHU
DTDI
DCO
PSU
UT Austin
Harvard
Randolph Rutledge
Daniel David Gregory
Joshua F. Einsle
Toronto
Kathy Fontaine
Shanan Peters
Rebecca Lynn Schmidt
Tasmania
Jolyon Ralph
John Nance
Bryan Jones
Michael Walter
Kathleen Ritterbush
Julia Ann Nord
UNLV
Alexander Prabhu
Brenda Thompson
Robert M. Hazen
GMU
Ana Udry
Cynthia Parr
Feifei Pan
Shaunna Morrison
CIS
Olivier Gagne
Michael Bruce Meyer
Lorch Babcock
MBL
SCI
Lyon
Nathan Stone
Gordon Downs
Eric D. Isaacs
Cambridge
Jule Maul
Rowan
Timothy Lyons
Mindat
CMM
Kelsey Markey
Ahmed Eleish
Larry Nittler
Peter Heaney
Louise Kellogg
David Cole
Hao Zhong
Asmaa Boujibar
Peter Fox
Teresa Fornaro
Cadence Boucher
Harrisburg
Utah
ODR
Anirudh Prabhu
Peter Edward Driscoll
Mirah Meyer
Jihua Hao
Eli Moore
Barbara Lafuente Valverde
Chao Liu
Juna Kollmeier
UC Davis
Lisa Warden
Alexander Lex
Linda Kah
Emil Ruff
Henderson James Cleaves
Joy Bonjourno
Paul Falkowski
Nathan Yee
NASA
SETI
UTK
Karen G. Lloyd
Donato Giovannelli
Wisconsin
Camerian Millsaps
David Blake
Benjamin Jelen
Peng Ni
Yixian Zheng
Catherine Conley
Mihaela Glamoclija
John Mulchaey
Greg Asner
Matthieu Laneuville
Ohio
Raymond Ernst Arvidson
Penelope Jane Boston
Thomas Bristow
David John Des Marais
Richard C. Walroth
Caltech
Rich Keller
Yana Bromberg
Mary A. Voytek
John Grotzinger
Allan Treiman
ELSI
LPI
ASU
NAI
Angela Maciel Detweiler
Adrienne Hoarfrost
Harrison B. Smith
Vikas Nanda
Kevin Hand
Janice McDonnell
Naples
Cayman Thomas Unterborn
Rutgers
Juliane Gross
Rice
Scott Hull
John Hedlund
Douglas Pike
Sara Walker
George Bennett
Max Miller
Kenneth McGuinness
Dror Noy
Costantino Vetriani
Jon Silberg
Kristin Bergmann
Joshua Mancini
Christopher Joseph Butch
Kat Dawson
Gaetano Montelione
Saroj Poudel
Diego U. Ferreiro

Future for network analytics?



v.



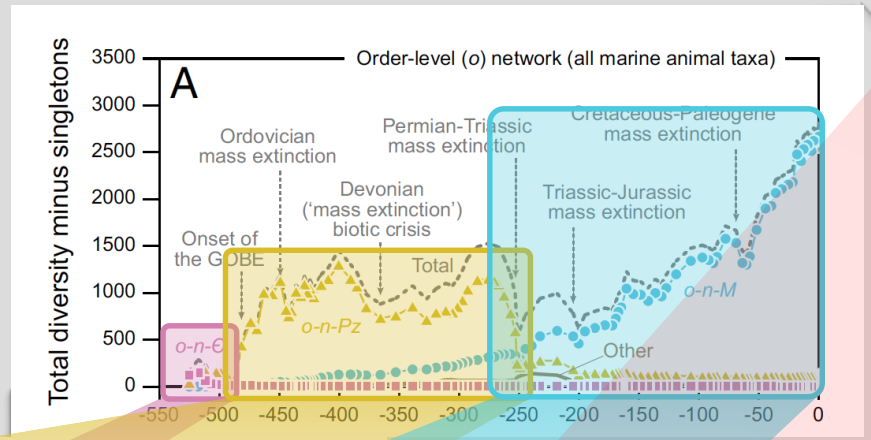
- Multi-layer networks, and network “difference” (not just a number)
- Characterizing heterogeneity and non-uniformity
- Networks over “time” (or some other parameter)
- Overcome my dissatisfaction with network metrics being “a number”…

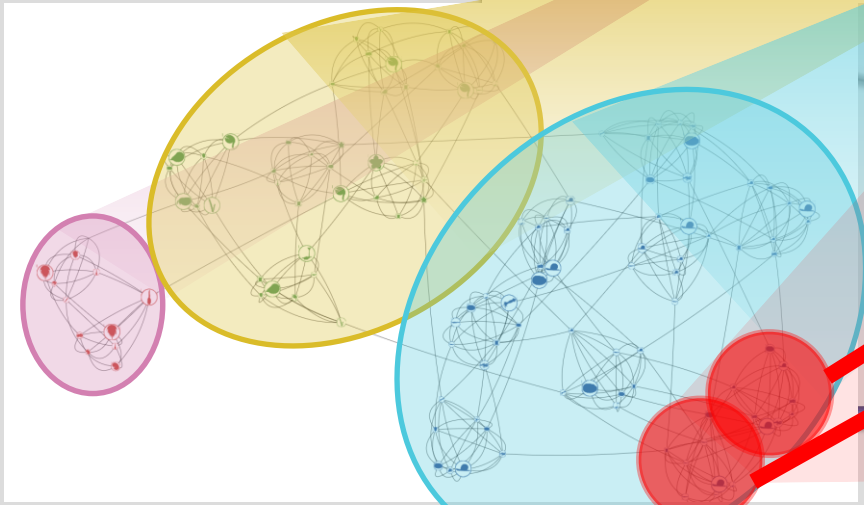
Big-sparse-heterogeneous-small

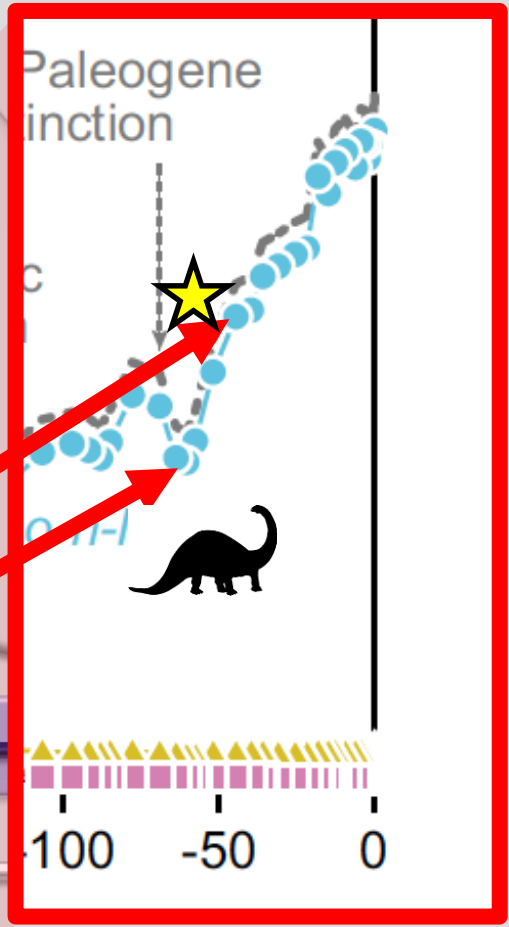
- Please take a look at the Science of Data Science paper (Fox and Hendler, Big Data. June 2014, 2(2): 68-70. doi:10.1089/big.2014.0011
  - Look at the call to action

- Thanks. pfox@cs.rpi.edu

Marine Fossil Occurrence Through Time







3D Network Geometry