Data Science at MSU and Beyond: A Paradigm Shift

The program will begin at 9:15 a.m.

Presented by the Office of Research & Economic Development and the Data Science Program

Mississippi State University

Dr. Philip Bourne
Dean of the School of Data Science
University of Virginia

Keynote: Is Data Science a Paradigm Shift?

Panel: Data Science in Higher Education & the Private Sector

Dr. Dan Gadke

Dr. Jing Liu

Mr. Danny Merchant

Dr. Jesse Smith

Dr. Jonathan Barlow

Business Information Systems

Dr. Merrill Warkentin

Professor, Management & Information Systems

What is Management Information Systems? (MIS), AKA "Information Systems" (IS)

- Study of technology, people, and organizations (+ their relationships)
- Organizations use IS to collect, process, & store data, then ...
- Turn data into information
- Design & implement IS to increase efficiency and effectiveness

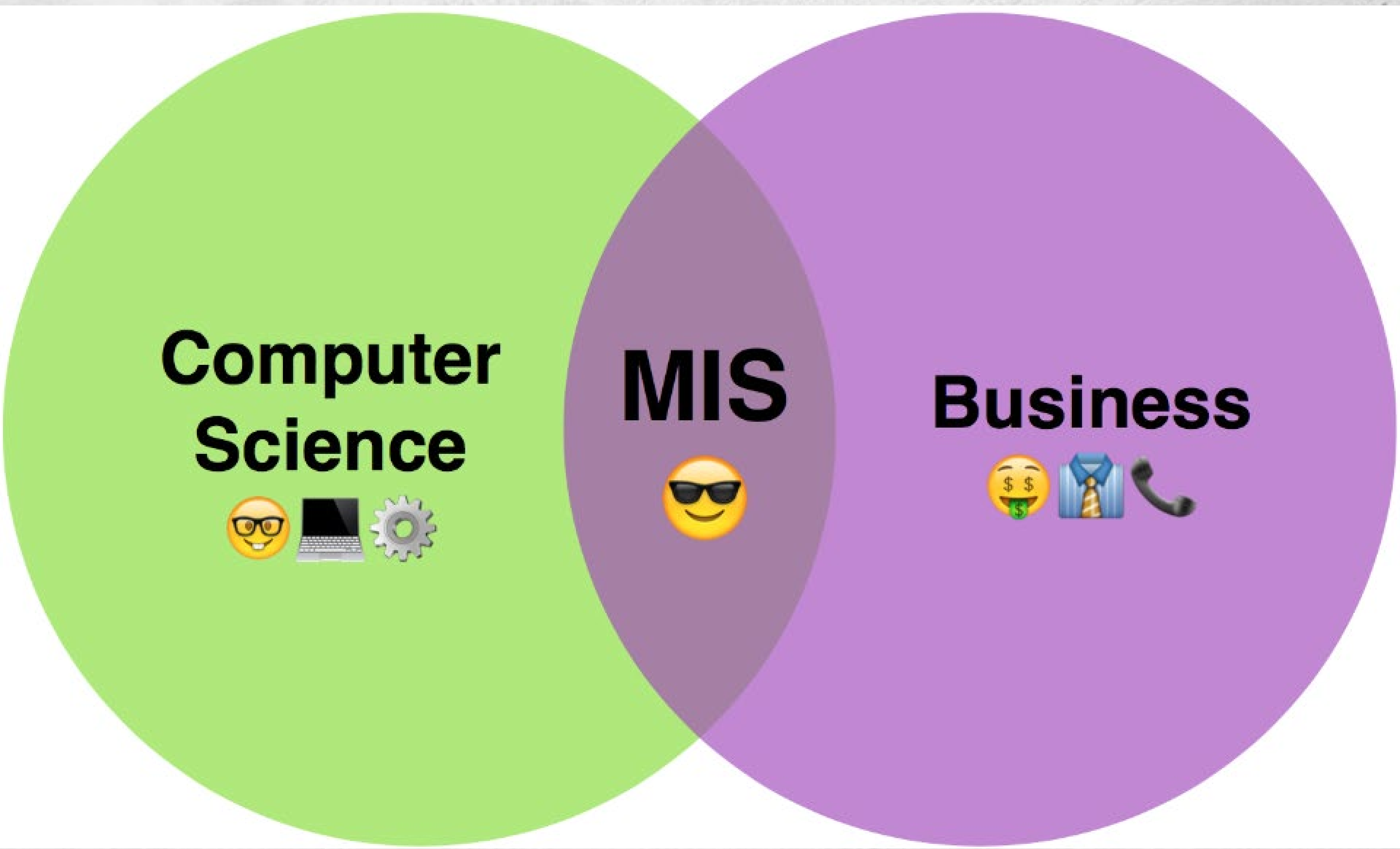

What Is MIS
Right Information
At Right Place
In the Right Form
To Right Person
At Right Time
At Right Cost

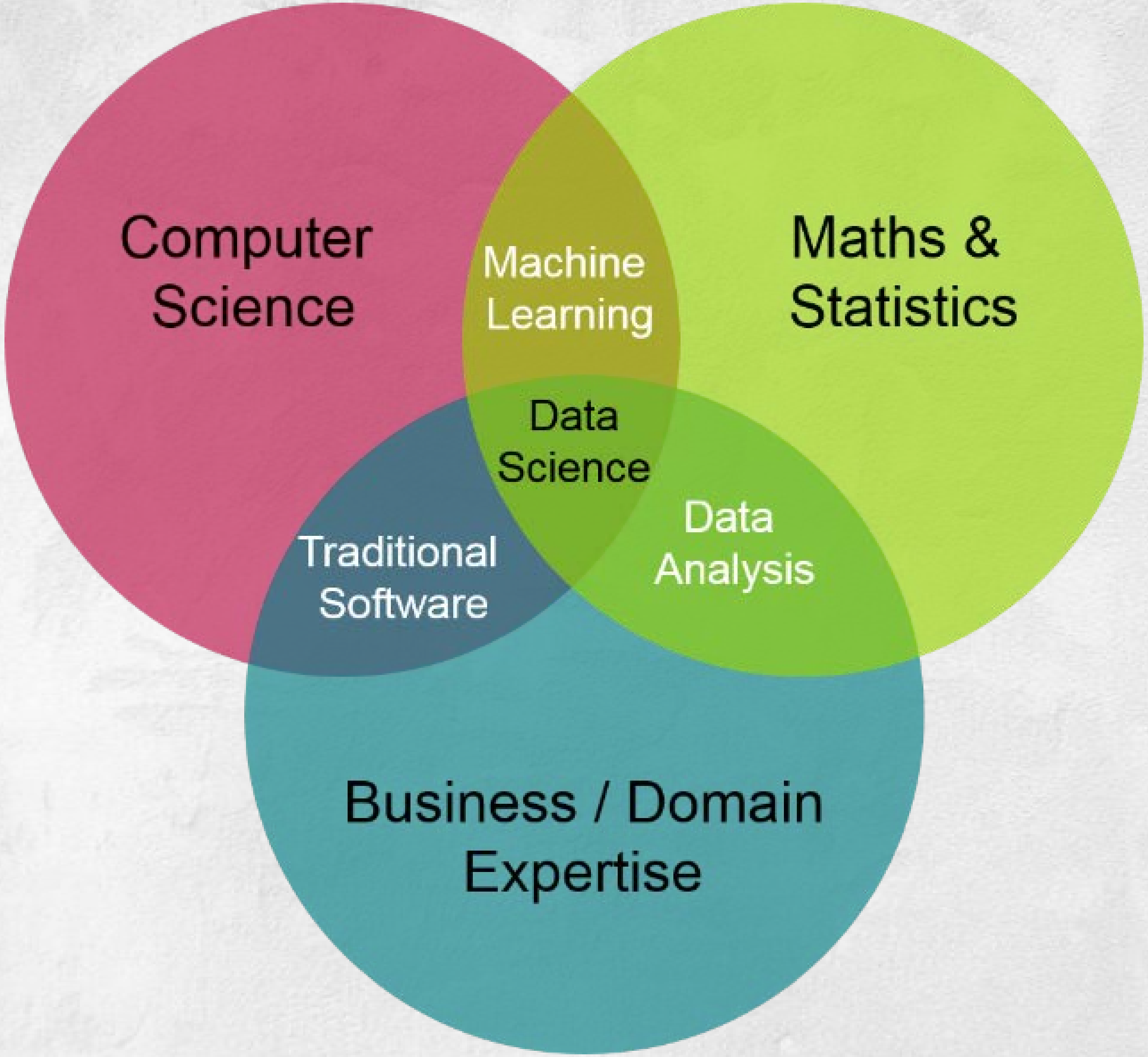
Is MIS the same as CS?

- MIS is to CS, ...

  as Pharmacy is to Chemistry

- An applied field focused on systems that solve problems and improve individuals, organizations, and society


Computer Science
MIS
Business


Computer Science
Machine Learning
Maths & Statistics
Data Science
Traditional Software
Data Analysis
Business / Domain Expertise

MIS focal areas

- Database management
  - I/O, queries, reporting, securing
- Systems analysis & design
- Networking
- Decision support systems
- Artificial intelligence

Business Intelligence (BI)

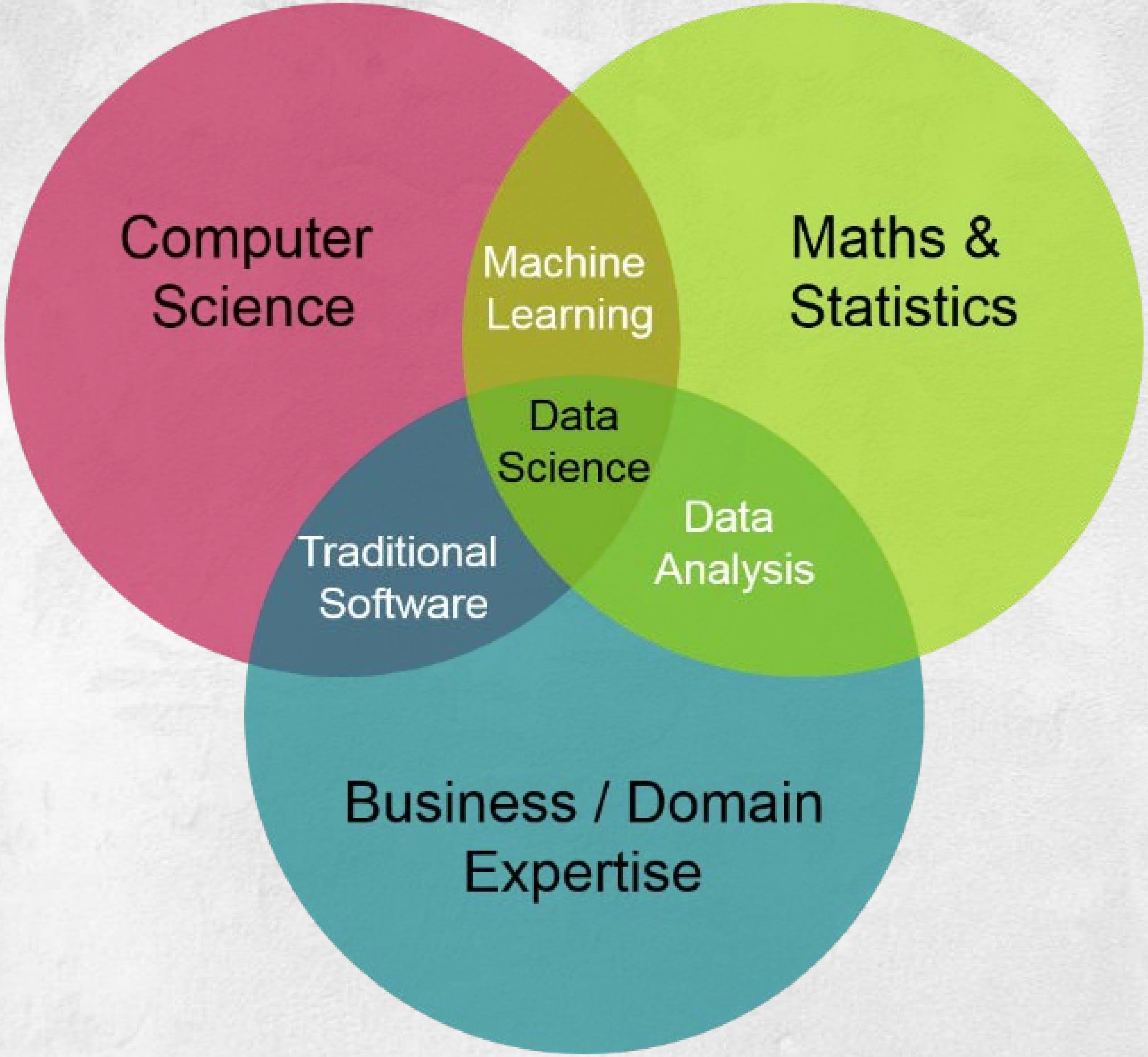
- Human decision makers (at all levels) need “better” managerial information
- Better = detailed, faster, visualized, easier to read, interactive, etc.
- Better info = better decision outcomes
- BI makes humans more powerful
- But must match domain!

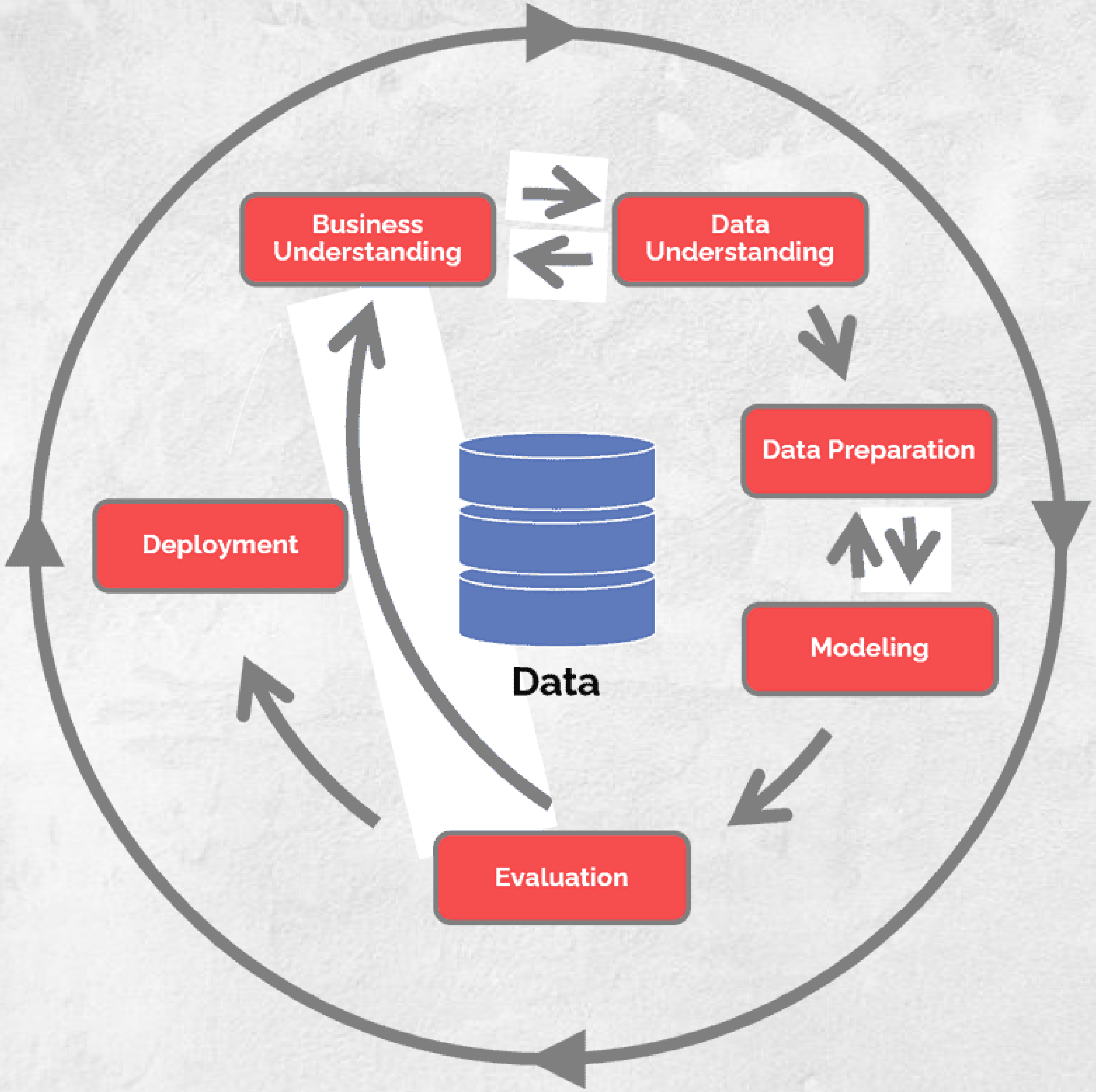
Data Science Applications in MIS

- Electronic commerce (incl. SEO)
- Fraud detection
- Human resource analytics
- Security and privacy management
- Recommendation systems
- Optimization (of routes, systems, etc.)

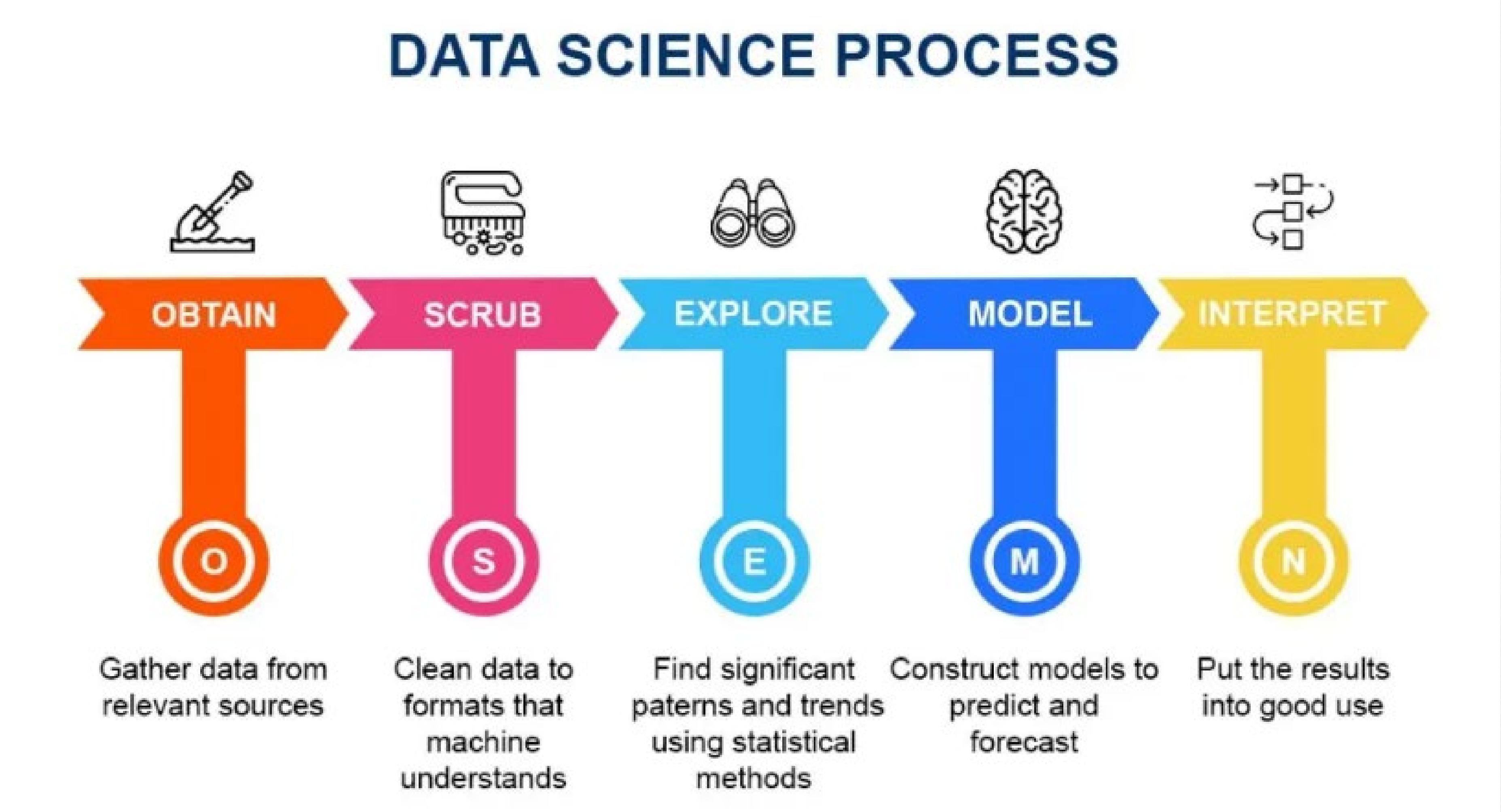
Research Topics (DSCI – MIS)

- Data governance
- Algorithm bias (+ ethics, dig divide)
- Future of work (tension between AI augmentation vs. automation)
- Innovations from AI/ML to enhance managerial decision making
- Cybersecurity, risk management
- Geospatial applications


Computer Science
Machine Learning
Maths & Statistics
Data Science
Traditional Software
Data Analysis
Business / Domain Expertise


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Contextualizing Data Science



MISSISSIPPI STATE UNIVERSITY

Journal of the Association for Information Systems

Editorial ISSN: 1536-9323

Big Data Research in Information Systems: Toward an Inclusive Research Agenda

Ahmed Abbasi
McIntire School of Commerce, University of Virginia
abbasi@comm.virginia.edu

Suprateek Sarker
McIntire School of Commerce, University of Virginia
School of Business, Aalto University
suprateek.sarker@comm.virginia.edu

Roger H. L. Chiang
Carl H. Lindner College of Business, University of Cincinnati
roger.chiang@uc.edu

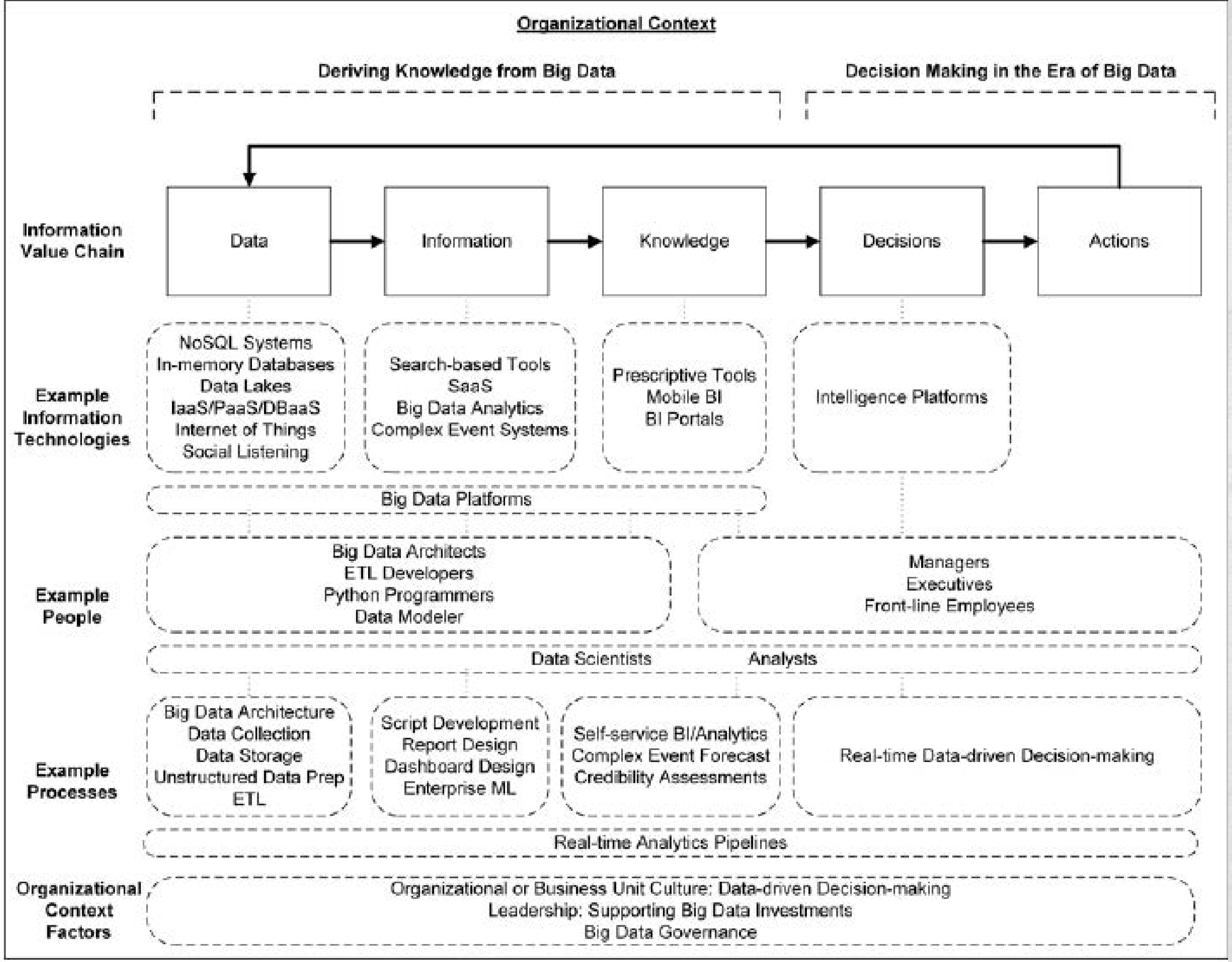


Figure 3. The Big Data Information Value Chain and Examples of Related People, Processes, and Technologies

Thank you!

Any Questions?

Computational Agriculture & Natural Resources

Dr. Will Davis

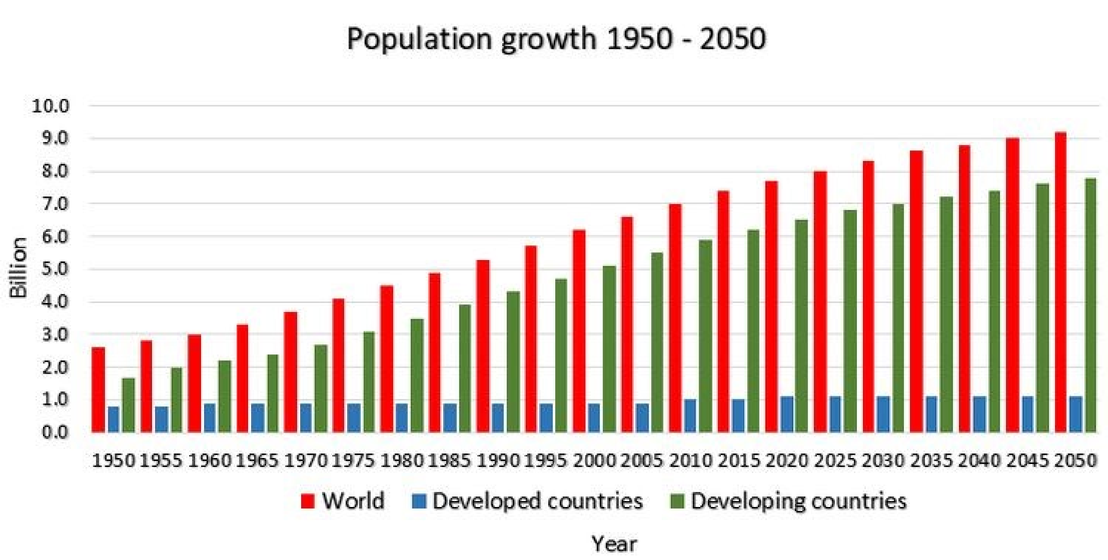
Assistant Professor, Agricultural Economics

Data Science Research in Agriculture and Natural Resources

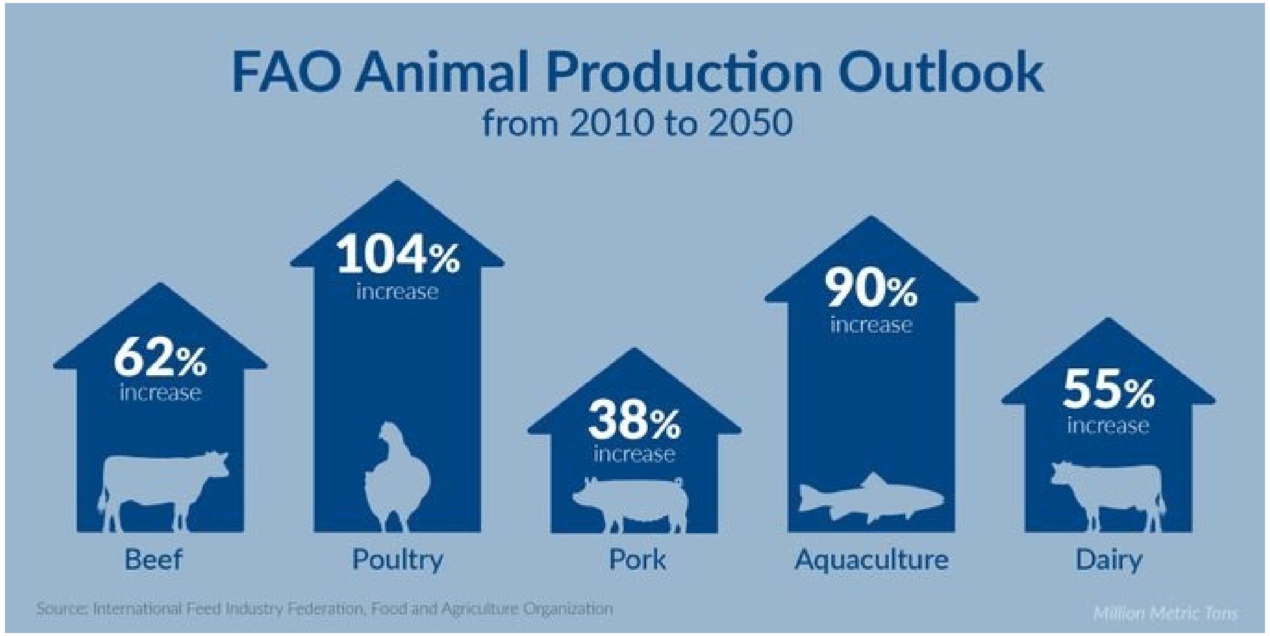
Will Davis, Ph.D.
Assistant Professor
Department of Agricultural Economics
Mississippi State University

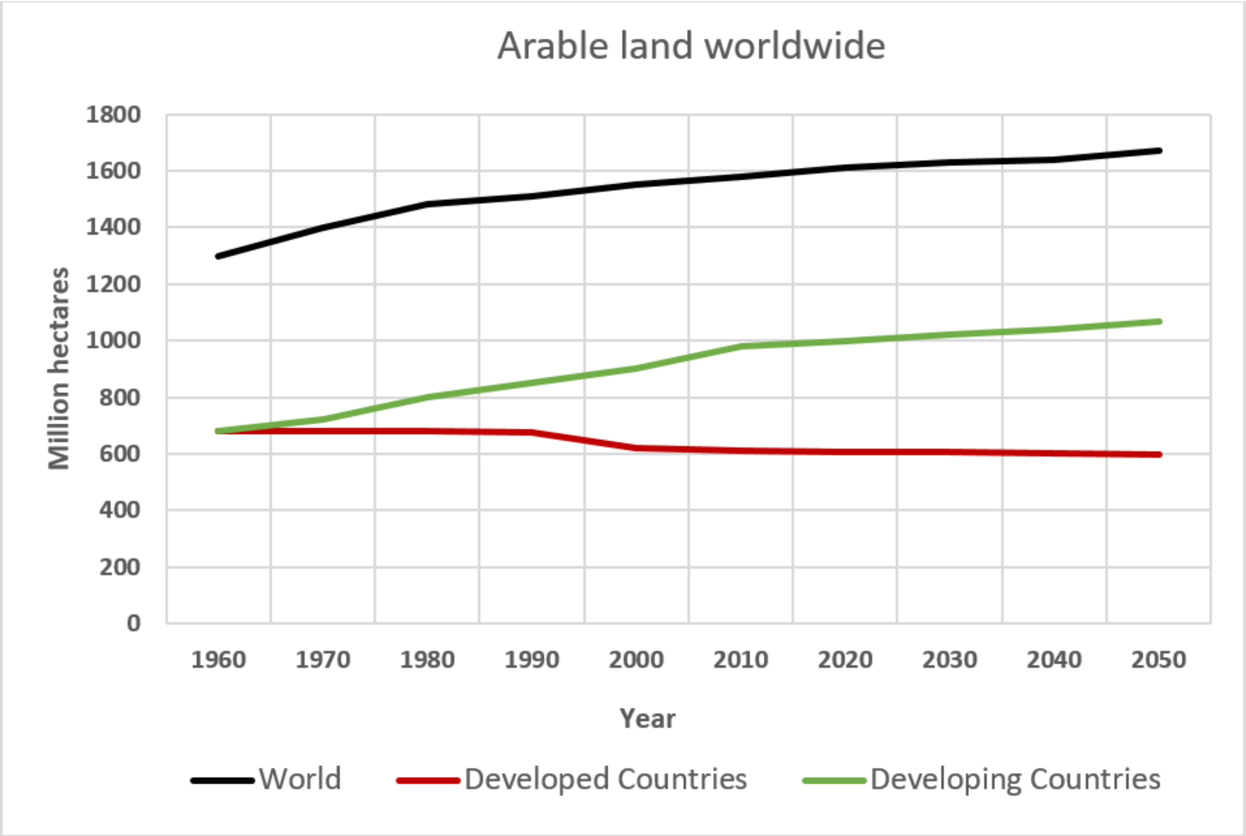
How do we Feed a Growing World?

- From 2009 to 2050, the world's population is expected to grow by 2.3 billion people.[1]
- FAO predicts that food production must increase by 70% over the same period.[1]
- Meeting this 70% goal is not guaranteed.
  - Annual agricultural productivity growth dropped by nearly 1/3rd in 2011-2020 compared to 2001-2010.[2]



Source: Silva (2018), Michigan State University Extension





Source: Silva (2018), Michigan State University Extension

The Data-Driven Future of Agriculture

- Modern agriculture relies on massive amounts of data to power technology and smart systems (AI).
- So called "Smart Farming" or "Farming 4.0" represents this technological evolution of agriculture.[3]
  - Incorporating new technologies like the Internet of Things, Autonomous Systems, AI, Machine Learning, and Big Data.
  - Allows us to do more with our limited resources.

The Data-Driven Future of Agriculture

- Not only allows us to produce more food, but in a more sustainable and environmentally responsible way.[3]
  - The largest known economic effect of climate change is on agriculture.[4]
  - Roughly 10% of all greenhouse gas emissions come from agricultural production.[5]

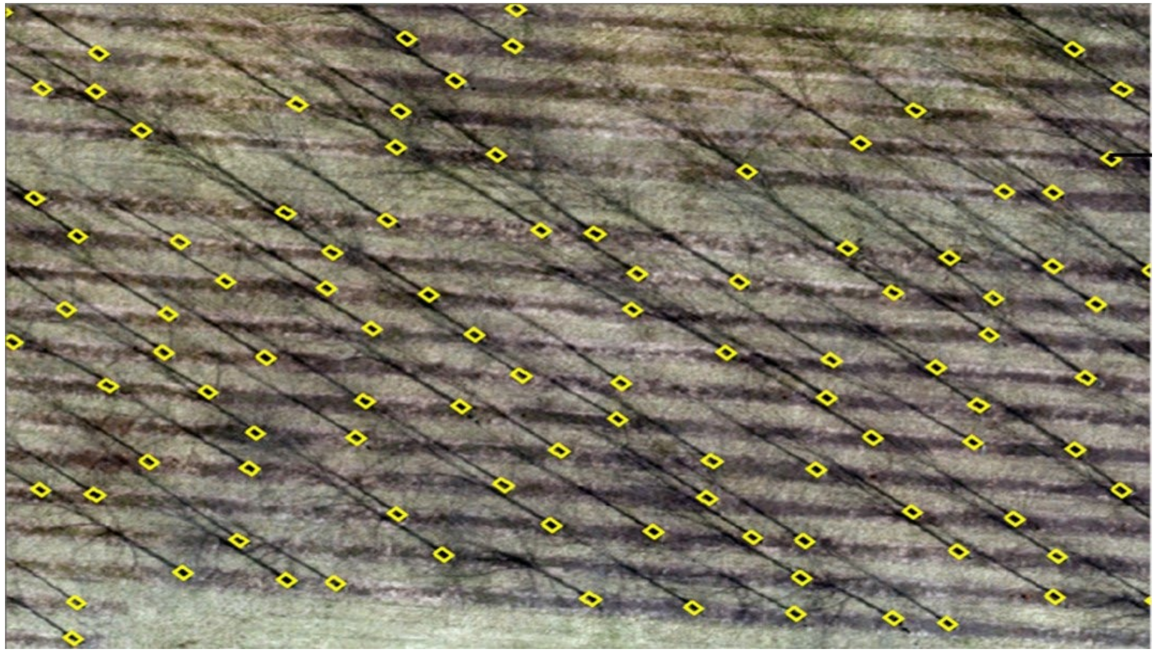
Source: Vegetable Growers News (Midwest's MightyVine doubles tomato greenhouses)

The Role of Data Science Research in Agriculture and Natural Resources

- The path towards more efficient agricultural and natural resource production is marked by Big Data and AI.
- Where there is need, there are opportunities for researchers both within and outside of academia.
  - Precision Agriculture.
  - AI and Autonomous Systems development.
  - Machine Learning and Big Data tools.

Source: John Deere UK (Autonomous Tractor Concept)

Source: John Deere UK (Large Spraying Drone)



Source: Purdue University (Tree Inventory with Aerial Remote Sensing)

Source: UGA Research (LiDAR Image of Tree Stand)

Opportunities for Data Science Research in Agriculture and Natural Resources at MSU

- Resources and faculty expertise at MSU well positioned for data science research in ag and natural resources.
- Faculty in CALS and CFR are currently working on cutting edge data-driven research.
  - AI, Autonomous Agriculture, Agricultural Production and Consumption, Human and Animal Health, Agricultural Supply Chains, Forestry, and Ecology to name a few.

Opportunities for Data Science Research in Agriculture and Natural Resources at MSU

- Need interdisciplinary approaches to solve complex problems.
  - Modern data-driven research in ag and natural resources stretches across disciplinary boundaries.
  - Must combine subject matter and technological expertise.

Source: Team Tweaks(The Future of Smart Farming using IOT))

Opportunities for Data Science Research in Agriculture and Natural Resources at MSU

- Access to key resources and institutions at MSU.
  - ORED.
  - High Performance Computing Collaboratory (HPC2) for GPU and CPU computing.
  - Mississippi State University Library system.
  - Predictive Analytics and Technology Integration (PATENT) Laboratory.
  - Potential for collaboration with groups like CAVs, SSRC, and NSPARC.

Source: Mississippi State University(Mitchel Memorial Library)

Opportunities for Data Science Research in Agriculture and Natural Resources at MSU

- Significant funding opportunities for data science research in agriculture and natural resources.
  - USDA, NIFA, NSF, the US Forest Service, and more.
- As of 4/18/22, grants.gov listed:
  - 409 matches for precision agriculture.
  - 345 matches for autonomous agriculture.
  - 559 matches for digital forestry.
- The funding is out there!

Opportunities for Data Science Research in Agriculture and Natural Resources at MSU

- Seeing movement towards large interdisciplinary and multi-institution grants.
  - Finding collaborators at MSU and other institutions is more important than ever to secure competitive funding.
  - Shifting focus towards collaboration between academia and industry.

Source:: EFMD Global (Multi- and Interdisciplinarity Empowered and Entailed by Business Schools' Digitalization)

Opportunities for Data Science Research in Agriculture and Natural Resources at MSU

- Growing access to agriculture and natural resource data.
  - Publicly available secondary data sources from government agencies like the USDA.
  - Privately collected data from corporations like John Deere.
  - Data collected directly from agricultural producers and natural resource firms.
- Current efforts at the University to improve ag and natural resource data access for faculty and students.

Anticipating the Future of Research in Agriculture and Natural Resources

- Data science is rapidly transforming the way we think and solve problems.
  – Nowhere is this truer than agriculture and natural resources.
- Need to anticipate the current and future applications of data science in these fields.

Source:: United Nations (Data-Science: Challenges and Opportunities)

Anticipating the Future of Research in Agriculture and Natural Resources

- Data science is also opening new opportunities that affect research, teaching, and extension/outreach.
  - How can we use data science to better achieve the Land Grant mission?
  - Need to identify new, transformative approaches to extension/outreach like digitally enabled extension services.
- Across all three pillars, MSU is positioned to lead the way towards a smarter tomorrow.

Thank You!

Will Davis, Ph.D.
gwd53@msstate.edu

References

1. FAO (2009), “2050: A third more mouths to feed”(https://www.fao.org/news/story/en/item/35571/icode/).
2. Morgan et al. (2022), “World Agricultural Output Growth Continues to Slow, Reaching Lowest Rate in Six Decades”, USDA Amber Waves.
3. Javaid et al. (2022), “Enhancing smart farming through the application of Agriculture 4.0 technologies”, International Journal of Intelligent Networks.
4. Mendelsohn (2009), “The Impact of Climate Change on Agriculture in Developing Countries”, Journal of Natural Resources Policy Research.
5. EPA (2023), “Sources of Greenhouse Gas Emissions” (https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions).

Geoinformatics

Dr. John Rodgers

Department Head, Geosciences

Data Science in Geosciences

Data Science in Geosciences

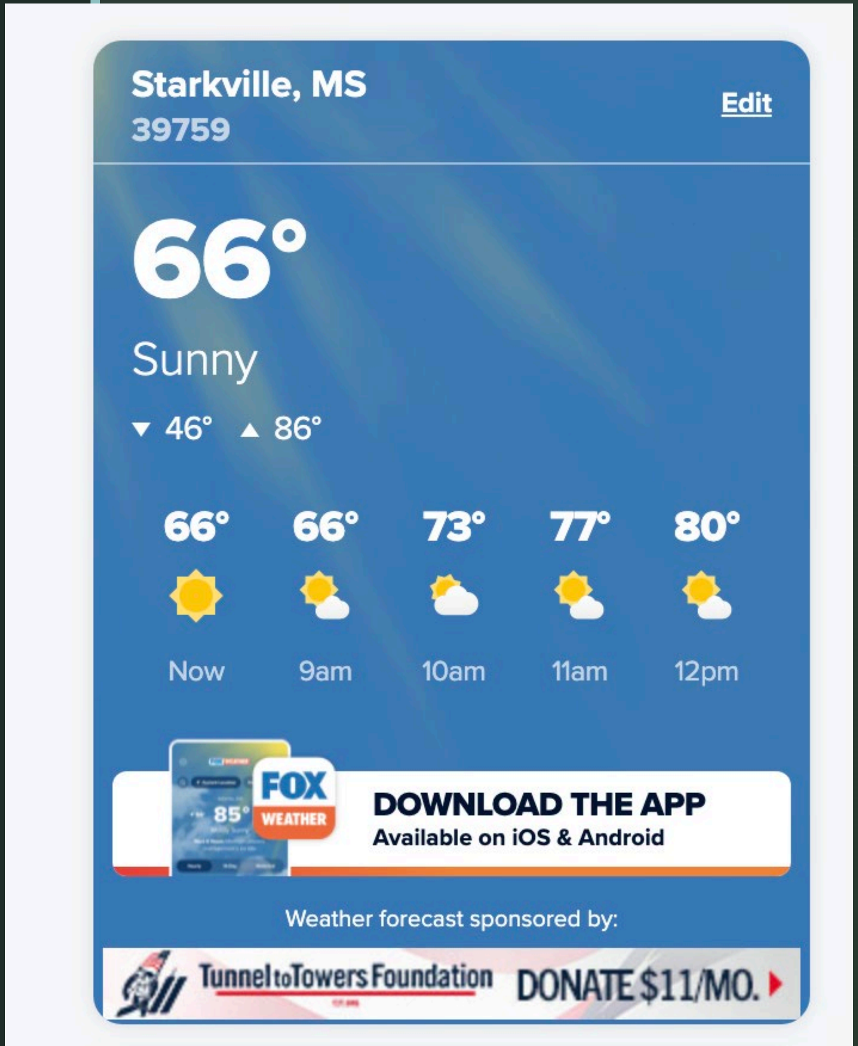
- How is Data Science transforming Geosciences?
- How is Data Science helping us engage in interdisciplinary research?
- Why Mississippi State University is at the vanguard?

What are the Geosciences?

Google Earth

- Study of the Earth
  - Geology
  - Meteorology
  - Geography
  - Geospatial
  - Human-Environment

Data Science and Meteorology



Improved Forecasting through Machine Learning

Intersection of
- Meteorology
- Computer Science
- Math and Stats

Informed Decisions about Severe Weather and Winter Weather

- Decisions to shelter?

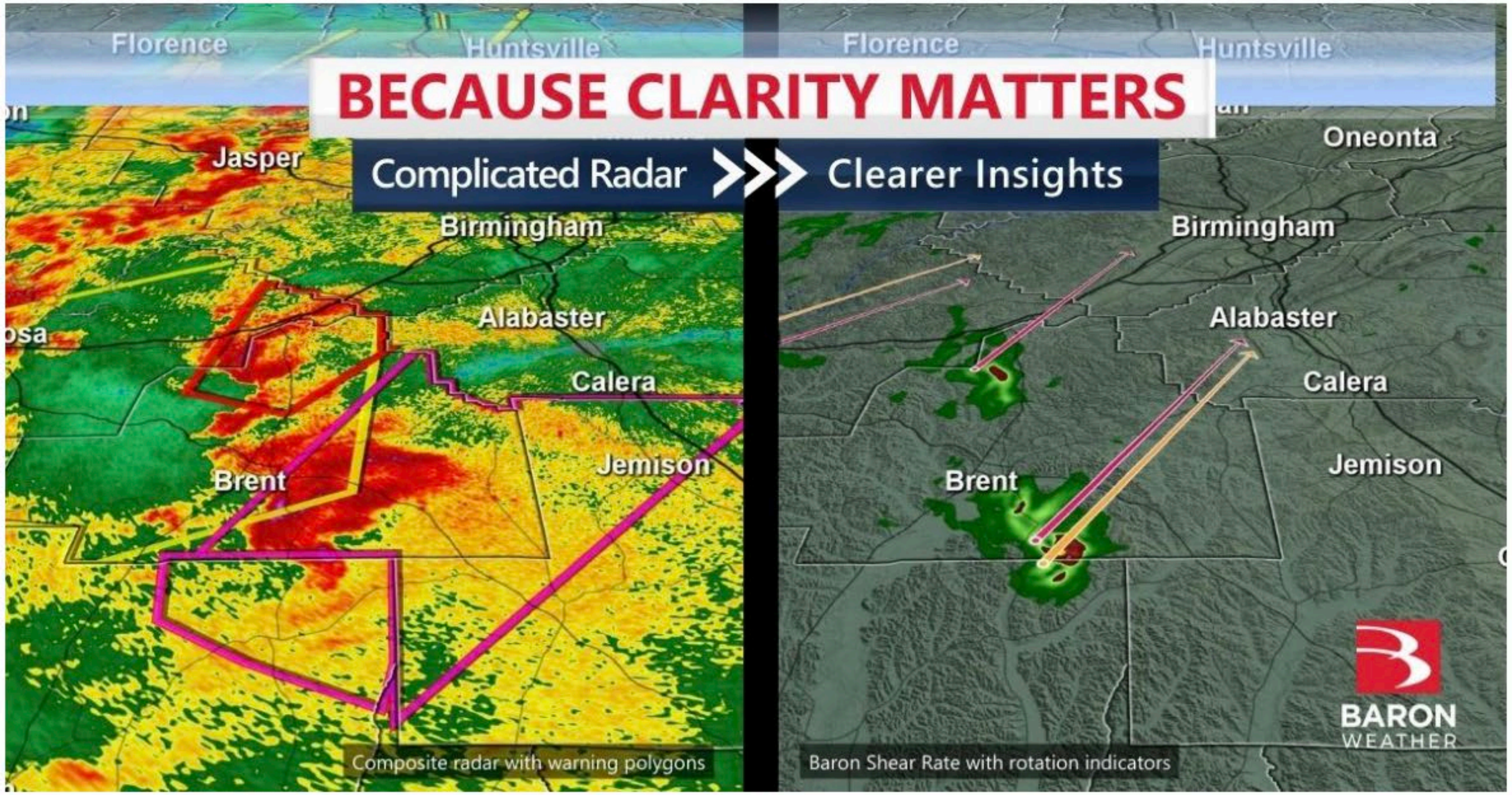
Is the forecast believable?

Probability of Detection Vs. False Alarm Ration

Intersection of

- Meteorology
- Social and Behavioral Sciences

warnings. This was apparent with the supercells during this event across central and s
multiple types of polygon warnings from the National Weather Service covered similar



Radar reflectivity and warning polygons from the National Weather Service (left) in contrast with Baron's exclusive Shear Rate and rotation indicators (right), captured at 10:36 pm CDT on March 30, 2022

Tropical Cyclone Rapid Intensification Applications

- Rapid intensification – increase in max wind of 30 kt in 24 hours (6% of TCs)
- Unsupervised learning to identify unique cyclone structures characterizing rapid intensification (Fig. 1)
- Operational product to predict tropical cyclone rapid intensification (Fig. 2)

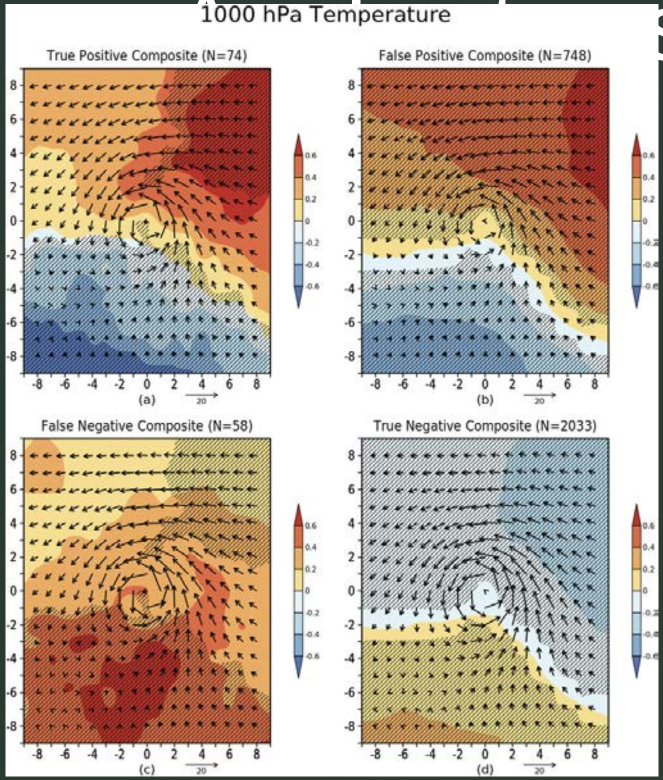


Fig. 1. 1000 mb temperature anomalies and winds for different RI forecasts

```
** MSU experimental Atlantic RI AI ensemble prediction: AL012020 Arthur 05/19/20 00UTC **

Global probability for RI (30kt/24h): 2.9%
This AI ensemble run is NOT EXPECTING RI (RI prob <= 18%)

Individual Ensemble Member Forecasts:
   Member Name     RI/no RI     RI prob
   AI1             no RI        4.1%
   AI2             no RI        4.5%
   AI3             no RI        0.0%
```

Fig. 2. Sample output from AI-based RI prediction system tested operationally with National Hurricane Center for 2020 Hurricane Arthur

Warm-Season Precipitation Forecasting

- Warm season (June-August) convective precipitation inherently difficult to model
- AI-based climatology of warm-season Southeast US (SEUS) rainfall (Figs. 3-4)
- Links established between climatological precipitation and El Niño/La Niña
- Future work will entail building weather model postprocessing prediction scheme for warm season rainfall using AI

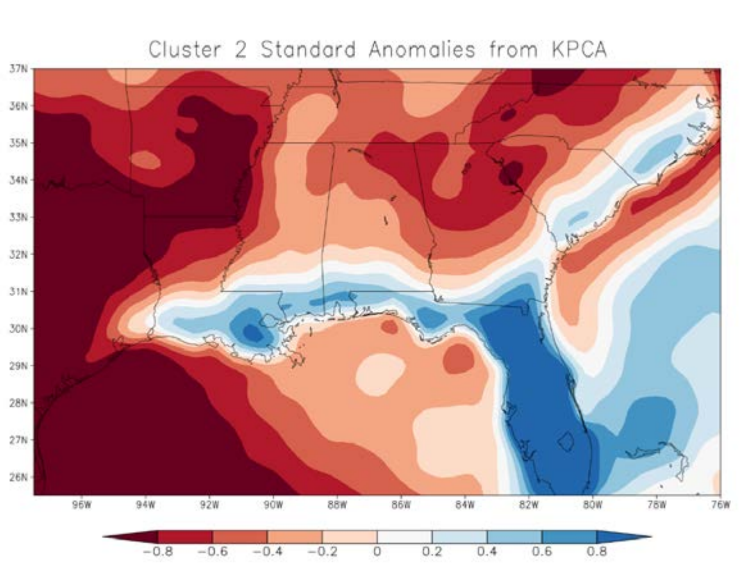


Fig. 3. Dry SEUS Pattern derived from kernel PCA

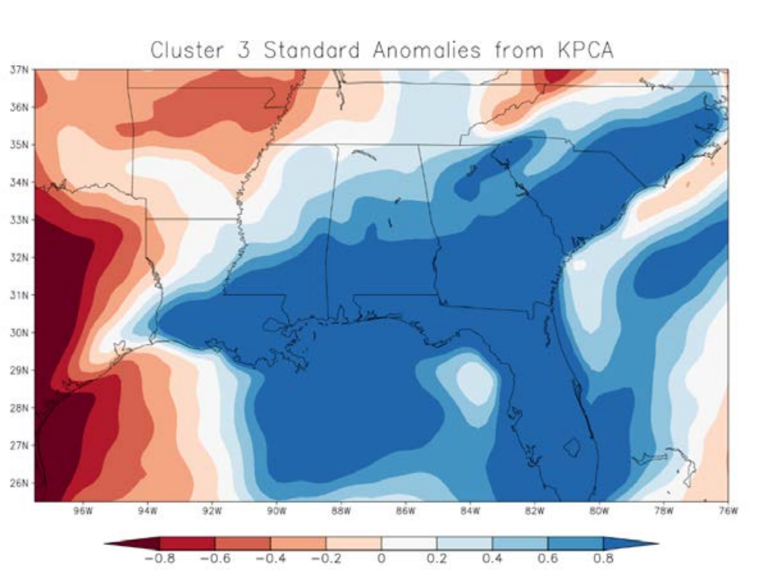


Fig. 4. Associated wet pattern from kernel PCA

fortune.com/2022/07/22/new-generation-data-scientists-heat-wave-climate-change-environment-weiwei-pan/

Notifications may include alerts, sou
and icon badges.

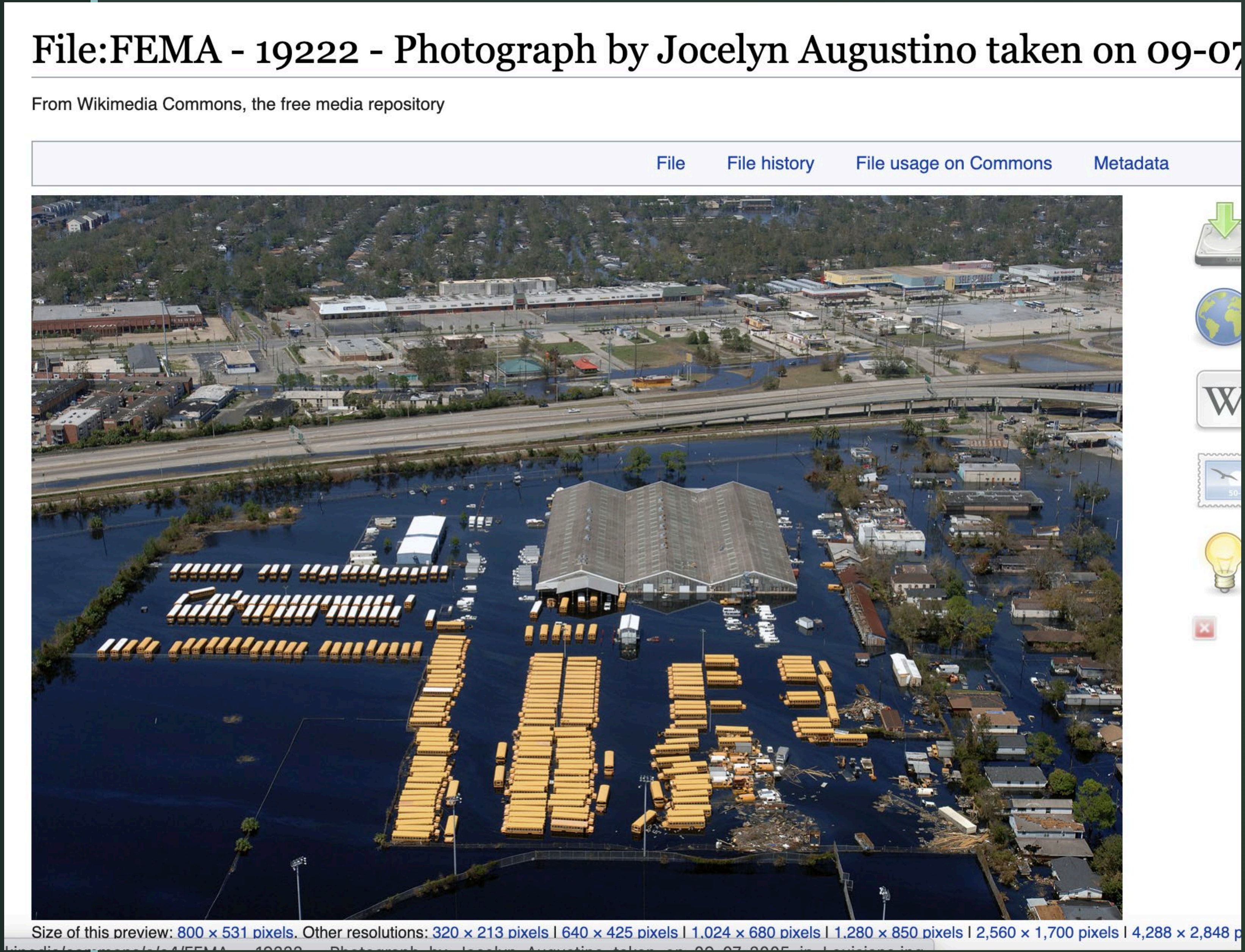
FORTUNE RANKINGS MAGAZINE NEWSLETTERS PODCASTS MORE SEARCH SIGN IN Subscribe Now

COMMENTARY · HEAT WAVE

A new generation of data scientists could be our best weapon against climate change

BY WEIWEI PAN

July 22, 2022 at 8:25 AM CDT



Preparation and Response to Disasters

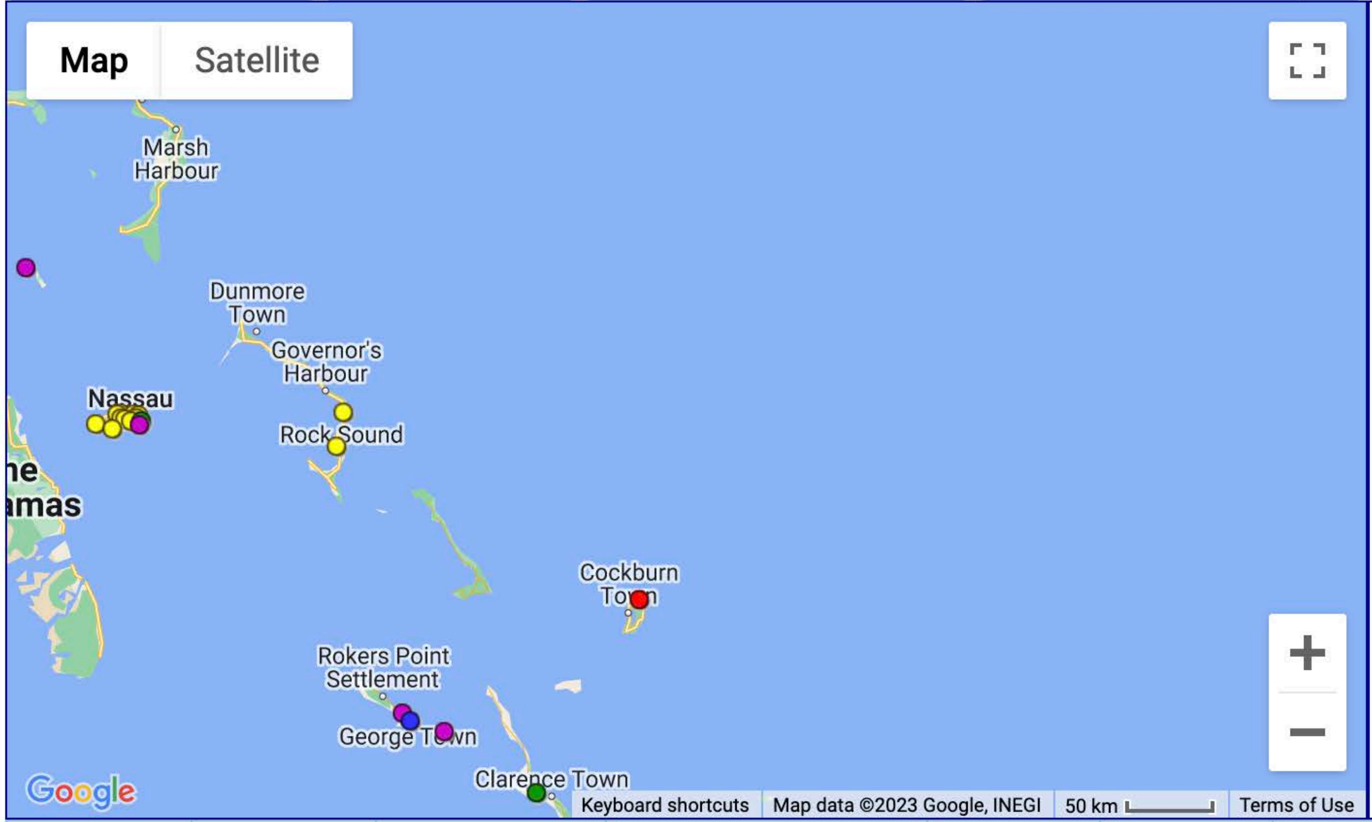
Intersection of

- Meteorology
- Computer Science
- Emergency Mngt.
- Urban Studies
- Sociology

Precipitation Study in The Bahamas

data.cocorahs.org/cocorahs/maps/?country=bhs

Precipitation ▾ | Bahamas ▾ | All Districts ▾ | 4/18/2023 | US Units ▾ | Update



CoCoRaHS Precipitation Map

Date: 04/18/2023
Country: BHS
District:
Units: US Units

- Zero
- Trace
- 0.01 - 0.11 in.
- 0.12 - 0.22 in.
- 0.23 - 0.54 in.
- 0.55 - 1.28 in.
- 1.29 - 1.92 in.
- 1.93 - 2.11 in.

Show US Active Fire Perimeters

Source: GEOMAC. GEOMAC wildfire data layers courtesy of the U.S. Geological Survey.

Obs Date	Obs Time	Station Number	Station Name	Precip (in.)	Snowfall (in.)	Snow Depth (in.)	Snow Depth SWE (in.)
2023-04-18	07:00 AM	BHS-NP-11	Nassau 4.1 ESE	0.94	NA	NA	NA
2023-04-18	06:00 AM	FL-WT-23	De Funiak Springs 7.3 NNE	0.00	0.0	NA	NA
2023-04-18	07:00 AM	FL-SR-26	Milton 2.9 NW	0.00	0.0	NA	NA
2023-04-18	07:00 AM	PR-PC-6	Ponce 2.3 NE	0.03	NA	NA	NA
2023-04-18	07:00 AM	FL-BY-22	Panama City Beach 4.4 ESE	0.00	0.0	NA	NA
2023-04-18	08:10 AM	FL-OR-61	Oakland 1.5 WSW	0.48	NA	NA	NA
			berdeen 3.7 WNW	0.12	NA	NA	NA

https://data.cocorahs.org/cocorahs/maps/?country=bhs#

Precipitation and Water Resources in the Caribbean

St Kitts Rainfall Data 2023

File Edit View Insert Format Data Tools Extensio

100% $ % .0 .00 123 Defau

A1 Date: 2023

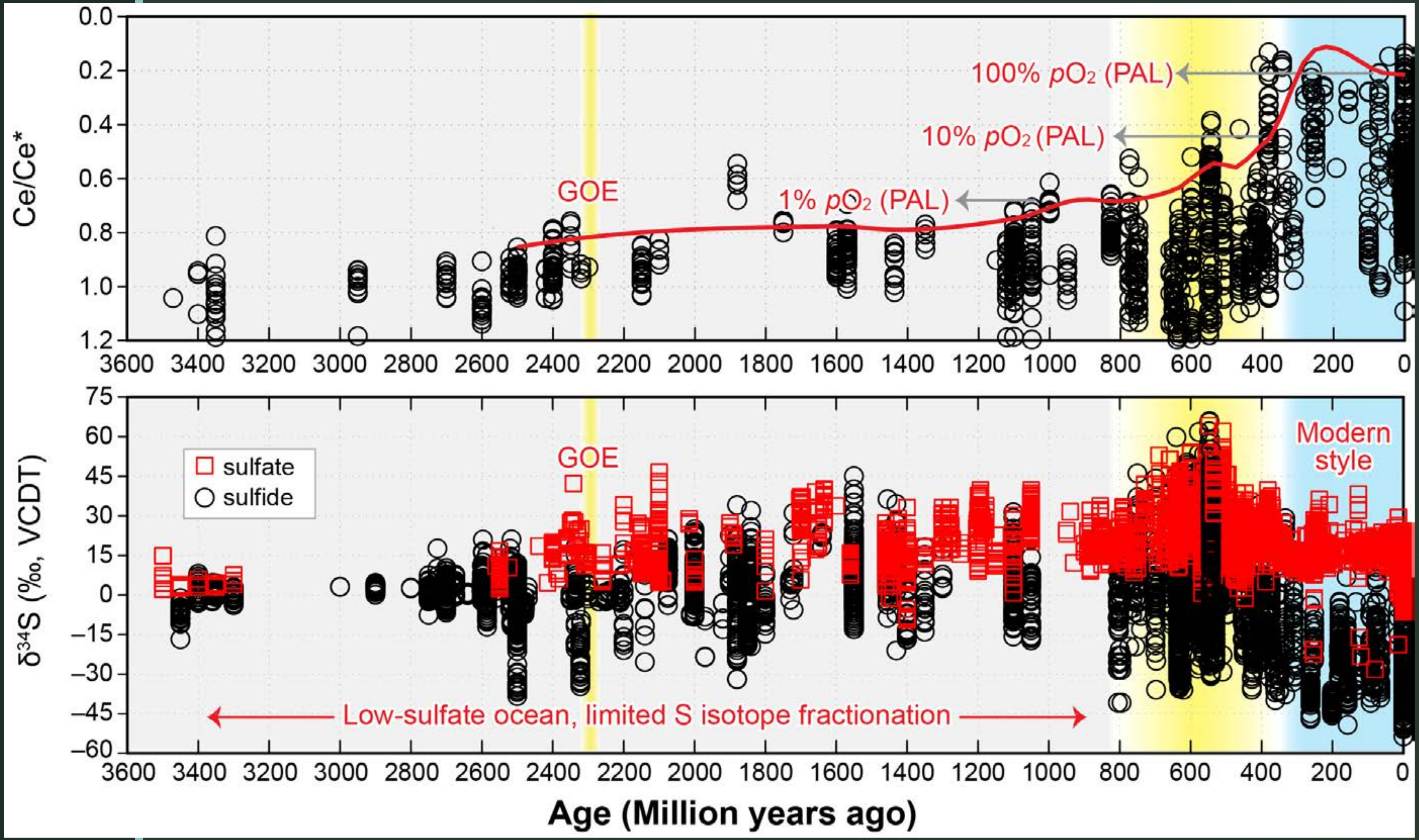
	B	C	D	E
1	Salt Pond	Olivees	St. Paul	Dieppie Bay
2	0	0.04	0.09	0.12
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0
11	0	0	0	0
12	0	0	0	0
13	0	0	0	0
14	0	0	0	0
15	0	0	0	0
16	0	0	0	0
17	0	0	0	0
18	0	0	0	0
19	0.18	0.49	0.42	0.11
20	0.08	0.08	0	0
21	0	0	0.05	0.04
22	0	0	0	0
23	0.06	0.06	0.42	0.11

Future

- Improved forecasting
  - accuracy and # days out
- Improved warnings
- Improved responses
- Potential Funding: NOAA

Data Science Geology

Reconstructing the evolution of ancient biogeochemical cycles by compiling geochemical signals throughout the Earth history



AGU ADVANCING EARTH AND SPACE SCIENCE

Geophysical Research Letters

RESEARCH LETTER
10.1029/2020GL088726

Global Marine Isochore Estimates Using Machine Learning

Taylor R. Lee[1,2], Benjamin J. Phrampus[1], Jeffrey Obelcz[1], Warren T. Wood[1], and Adam Skarke[2]

[1]U.S. Naval Research Laboratory, Stennis Space Center, MS, USA, [2]Department of Geosciences, Mississippi State University, Mississippi State, MS, USA

Key Points:

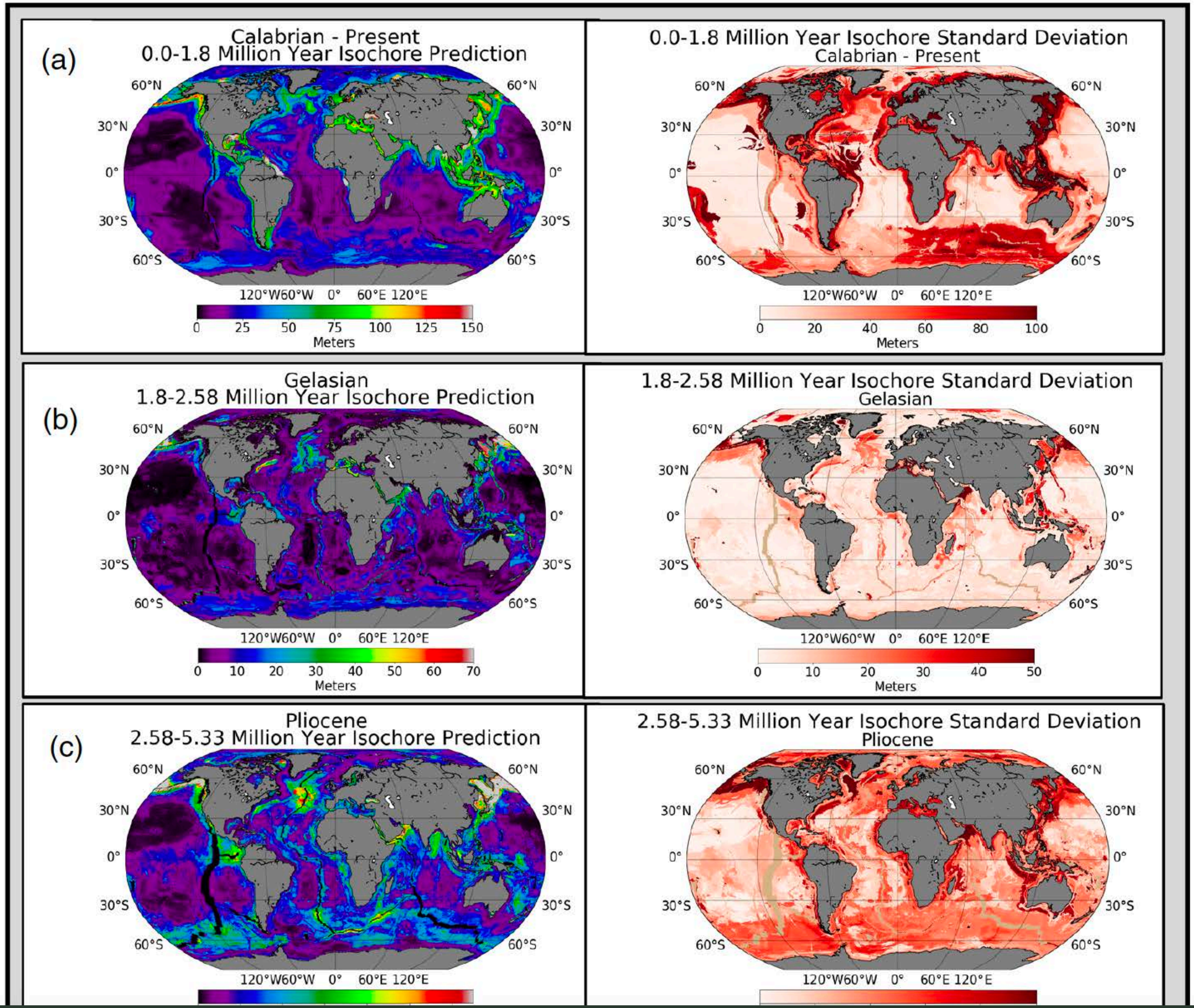
- We use a machine learning approach to produce the first global estimates of isochore thicknesses for the present to middle Miocene
- Given minimal training data, we produce global and temporally variable isochores with lower error than modeled isochores
- Results provide first order constraints on deposition with geologic time, which is of importance for assessing climate variability

Supporting Information:

- Supporting Information S1

Correspondence to:
T. R. Lee,
taylor.lee@nrlssc.navy.mil

Abstract The thickness normal to deposition (isopachs) and vertical thickness (isochores) of geological units is important for assessing various geologic processes. We present the first marine global sediment isochore estimates for five geological periods dating from middle Miocene (15.97 Ma) to present. We use sparsely distributed sediment depth vs. age observations from the Deep Sea Drilling Project and global maps of biological, oceanographic, geographic, and geological variables as training features in a k-nearest neighbor regressor to estimate isochores. Results are compared to isochore estimates generated by applying a constant depositional rate from recent estimates of global total sediment thicknesses. Both models of isochore thickness exhibit consistent error. Results from a machine learning approach show major advantages, including results that are quantitative, easily updatable, and accompanied with uncertainty estimation. Final predictions can provide first-order constraints on sediment deposition with geologic time, which is of timely importance for assessing past climate variability.


(a)
Calabrian - Present
0.0-1.8 Million Year Isochore Prediction
0 25 50 75 100 125 150
Meters
0.0-1.8 Million Year Isochore Standard Deviation
Calabrian - Present
0 20 40 60 80 100
Meters
(b)
Gelasian
1.8-2.58 Million Year Isochore Prediction
0 10 20 30 40 50 60 70
Meters
1.8-2.58 Million Year Isochore Standard Deviation
Gelasian
0 10 20 30 40 50
Meters
(c)
Pliocene
2.58-5.33 Million Year Isochore Prediction
2.58-5.33 Million Year Isochore Standard Deviation
Pliocene
60°N
30°N
0°
30°S
60°S
120°W60°W 0° 60°E 120°E

Check for updates

Global Biogeochemical Cycles

RESEARCH ARTICLE
10.1029/2021GB007248

Key Points:

- We use machine learning to produce several inputs to deterministic models that estimate subsurface carbon degradation and methane generation
- We estimate 0.8–2.2 and 1.1–3.0 × 10^6 Pg of carbon and methane, respectively, generated for subsurface sediments
- Results yield global geospatial estimates of the maximum amount of methane possibly generated *in-situ* and sequestered in marine sediments

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

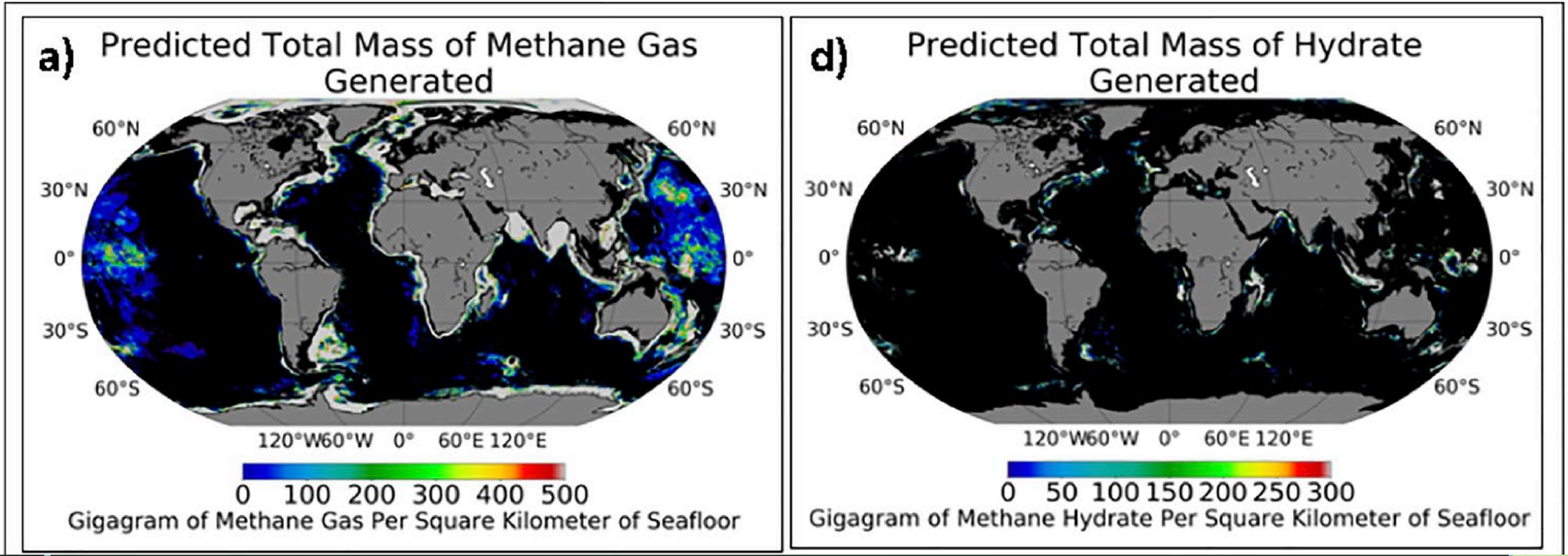
T. R. Lee,
taylor.lee@nrlssc.navy.mil

Global Estimates of Biogenic Methane Production in Marine Sediments Using Machine Learning and Deterministic Modeling

T. R. Lee[1,2], B. J. Phrampus[1], A. Skarke[2], and W. T. Wood[1]

[1]U.S. Naval Research Laboratory, Stennis Space Center, MS, USA, [2]Department of Geosciences, Mississippi State University, Starkville, MS, USA

Abstract We have developed a model of geospatially estimating carbon accumulation and methanogenesis in seabed sediments that uses more accurate and sophisticated inputs to models than used in previous estimates. Using this hybrid stochastic and deterministic model, we estimate the maximum carbon available for methanogenesis in the global seabed, and subsequent microbial methane generated as a function of location and depth (including the gas hydrate stability zone). Global integration over present and previously microbially reactive sediments column yields total carbon and methane to be ~0.8–2.2 × 10^6 and 1.1–3.0 × 10^6 Pg C and CH_4, respectively. Our improvements to accuracy include using geospatially machine learned estimates of seafloor inputs to which the methanogenesis modeling is most sensitive (e.g., total organic carbon, heat flux, porosity). Our improvements to model sophistication include geospatially dependent modeling (on a 5 × 5 arc-minute grid), a new model of sediment compaction (allowing for non-linear geothermal gradients), and variable age versus depth at each grid cell. A carbon reservoir of the magnitude we estimate here is


a) Predicted Total Mass of Methane Gas Generated
60°N
30°N
0°
30°S
60°S
120°W60°W 0° 60°E 120°E
0 100 200 300 400 500
Gigagram of Methane Gas Per Square Kilometer of Seafloor
d) Predicted Total Mass of Hydrate Generated
60°N
30°N
0°
30°S
60°S
120°W60°W 0° 60°E 120°E
0 50 100 150 200 250 300
Gigagram of Methane Hydrate Per Square Kilometer of Seafloor

Data Science and Planetary Geology: From MSU geology professor

“NASA is obviously interested in going back to the Moon. I’m currently using very high resolution datasets to generate maps of places we may be able to drill/access water ice that could be used for in situ resource utilization. On Earth, my use of data science is a bit different because I actually go to the Earth analogue sites to collect data. We generate large amounts of rock composition, presence/absence, and characteristic data to reconstruct 3D models of locations on Earth that can then be used to help us interpret the data we get from Moon, Venus, Mars, name your planet of interest.” Dr. Kelsey Crane, MSU Geology Professor

Data Science and Oceanography: From MSU geology professor

Water Quality Monitoring using Unmanned Aerial Systems Imagery and a Novel Autonomous Surface Vessel (Dash et al.)

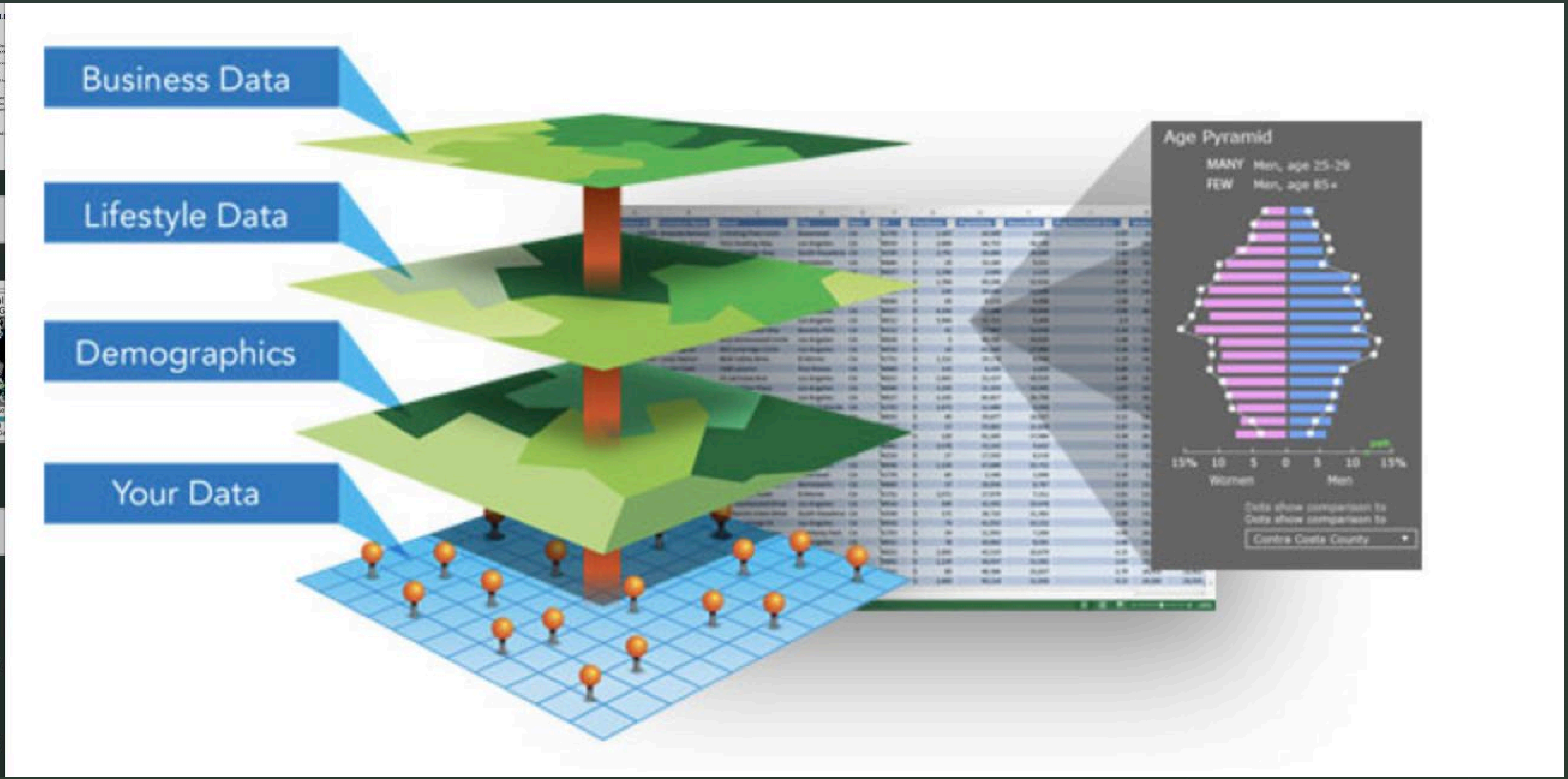
- UAS and ASV data helped develop robust algorithms, using which the impacts of water quality on the oyster reef could be evaluated.

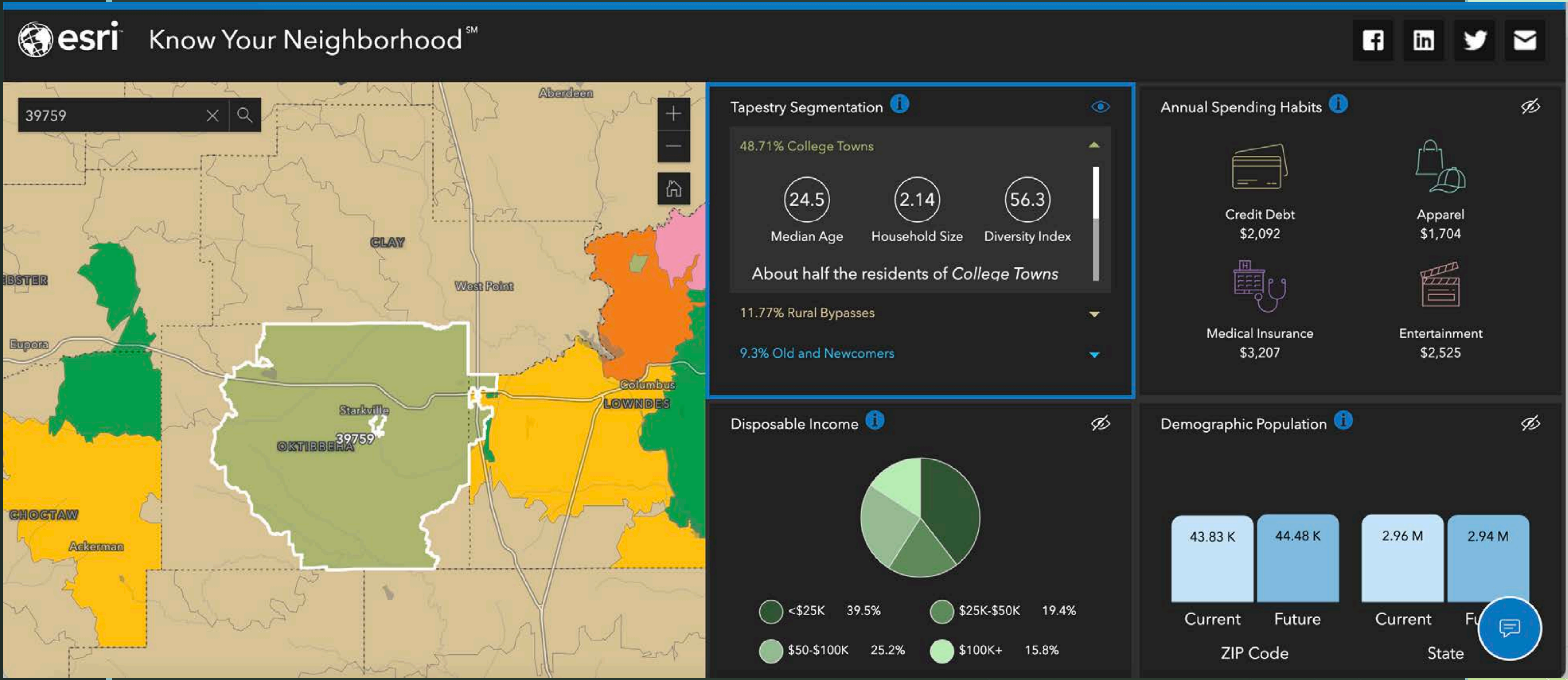
Future

- Exploration and quantification of resources
  - Energy
  - Minerals
  - Water
- Understanding of the Earth and Solar System

Intersection of
- Geology/Oceanography
- Chemistry
- Computer Science
- Math and Stats

Data Science and Geospatial




esri
Know Your Neighborhood SM
39759
Aberdeen
CLAY
West Point
Eupora
Columbus
LOWNDES
Starkville
OKTIBBEHA
39759
CHOCTAW
Ackerman
Tapestry Segmentation
48.71% College Towns
24.5
Median Age
2.14
Household Size
56.3
Diversity Index
About half the residents of College Towns
11.77% Rural Bypasses
9.3% Old and Newcomers
Annual Spending Habits
Credit Debt
$2,092
Apparel
$1,704
Medical Insurance
$3,207
Entertainment
$2,525
Disposable Income
<$25K 39.5%
$25K-$50K 19.4%
$50-$100K 25.2%
$100K+ 15.8%
Demographic Population
43.83 K
44.48 K
2.96 M
2.94 M
Current
Future
Current
ZIP Code
State

World of Change: Amazon Deforestation

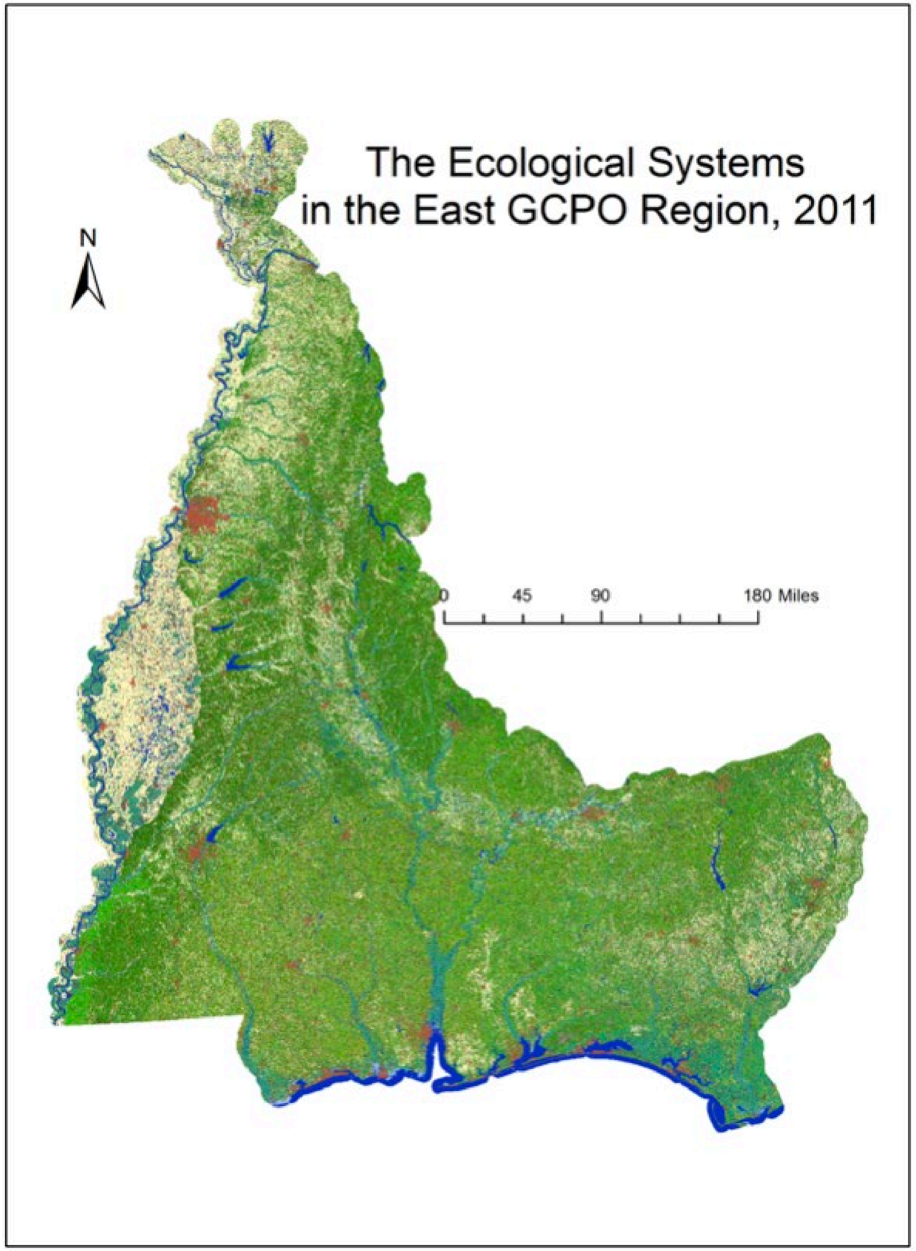


Fig. 6 The ecological systems in the east GCPO region, 2011

- Atlantic Coastal Plain Fall-line Sandhills Longleaf Pine Woodland - Open Understory
- Atlantic Coastal Plain Fall-line Sandhills Longleaf Pine Woodland - Scrub/Shrub Understory
- Atlantic Coastal Plain Upland Longleaf Pine Woodland
- Atlantic Coastal Plain Xeric River Dune
- East Gulf Coastal Plain Interior Upland Longleaf Pine Woodland - Open Understory Modifier
- East Gulf Coastal Plain Interior Upland Longleaf Pine Woodland - Scrub/Shrub Modifier
- West Gulf Coastal Plain Upland Longleaf Pine Forest and Woodland
- East Gulf Coastal Plain Limestone Forest
- East Gulf Coastal Plain Maritime Forest
- East Gulf Coastal Plain Southern Loess Bluff Forest
- East Gulf Coastal Plain Southern Mesic Slope Forest
- Southern Coastal Plain Dry Upland Hardwood Forest
- Southern Coastal Plain Oak Dome and Hammock
- Atlantic Coastal Plain Fall-Line Sandhills Longleaf Pine Woodland - Loblolly Modifier
- Deciduous Plantations
- East Gulf Coastal Plain Interior Upland Longleaf Pine Woodland - Loblolly Modifier
- East Gulf Coastal Plain Interior Upland Longleaf Pine Woodland - Offsite Hardwood Modifier
- East Gulf Coastal Plain Near-Coast Pine Flatwoods - Offsite Hardwood Modifier
- Evergreen Plantation or Managed Pine
- Allegheny-Cumberland Dry Oak Forest and Woodland - Hardwood
- Allegheny-Cumberland Dry Oak Forest and Woodland - Pine Modifier
- East Gulf Coastal Plain Black Belt Calcareous Prairie and Woodland - Woodland Modifier
- East Gulf Coastal Plain Northern Dry Upland Hardwood Forest
- East Gulf Coastal Plain Northern Loess Plain Oak-Hickory Upland - Hardwood Modifier
- East Gulf Coastal Plain Northern Loess Plain Oak-Hickory Upland - Juniper Modifier
- Lower Mississippi River Dune Woodland and Forest
- Northeastern Interior Dry Oak Forest - Mixed Modifier
- Northern Atlantic Coastal Plain Dry Hardwood Forest
- Southern Interior Low Plateau Dry-Mesic Oak Forest

Future

- Use of data science to answer spatial questions, “Where?”
  - Military Intelligence
  - Retail Location
  - Wildlife Management
  - Urban Planning
  - Epidemiology
- Funding – governmental sources, federal, state, and local

Conclusions

- Data Science has helped transform Geosciences
  - Improved our ability to study and understand the Earth
  - Improved our ability to answer questions
  - Improved our ability to find and manage resources
- Data Science helps make Geosciences part of an interdisciplinary research team
  - College of Arts & Sciences, BCOE, GRI, NGI, SSRC, nSPARC
- Mississippi State University is a leader in Data Science
  - Computer Resources + Faculty Expertise + Outstanding Students

Marketing Analytics

Dr. Stephen France
Associate Professor
Quantitative Analysis

Classification and Visualization in Marketing Analytics

Stephen L. France
Associate Professor of Quantitative Analysis
Department of Marketing, Quantitative Analysis, and Business Law
College of Business
Mississippi State University

It Starts in the 1960s…………

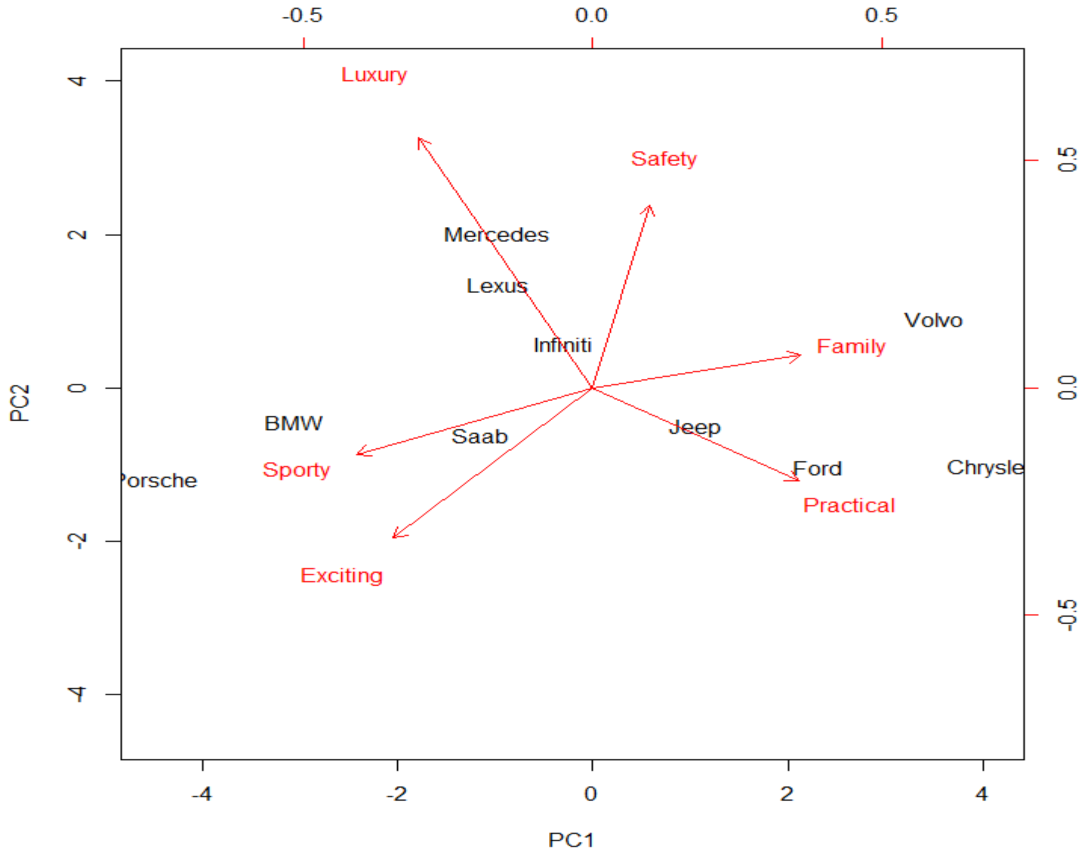
- Increasing use of data processing and availability of data in industry in the 1950s and 1960s.
- The general growth of consumerism and consumer marketing in the go-go era of the 1960s.
- Ford Foundation initiative to provide quantitative training to management school researchers in US.

Initial Methods

- Mostly perceptual mapping applications for product positioning and design
- Some customer segmentation and mapping applications.
- Work published in The Journal of Marketing, Journal of Marketing Research
- Visualization methods
  - Factor Analysis
  - Correspondence Analysis
  - Principal Components Analysis
  - Multidimensional Scaling
- Classification/clustering methods
  - Partitioning Clustering
  - Hierarchical Clustering
- I'll give a very simple example………………….

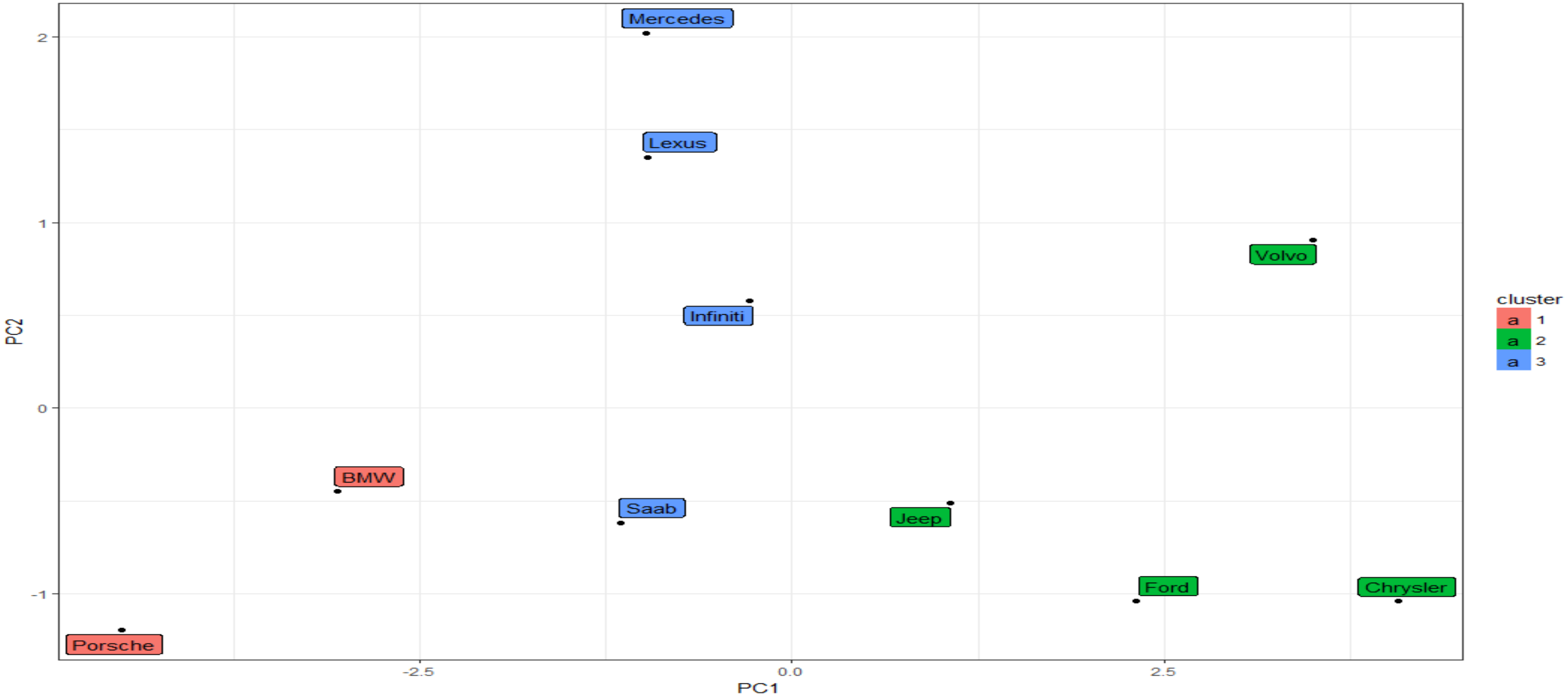
PCA Biplot

```
#The biplot for PCA
biplot (car.pr, scale =0,xlabs=CarData[,1])
```



PCA + K-means 3 Cluster Solution

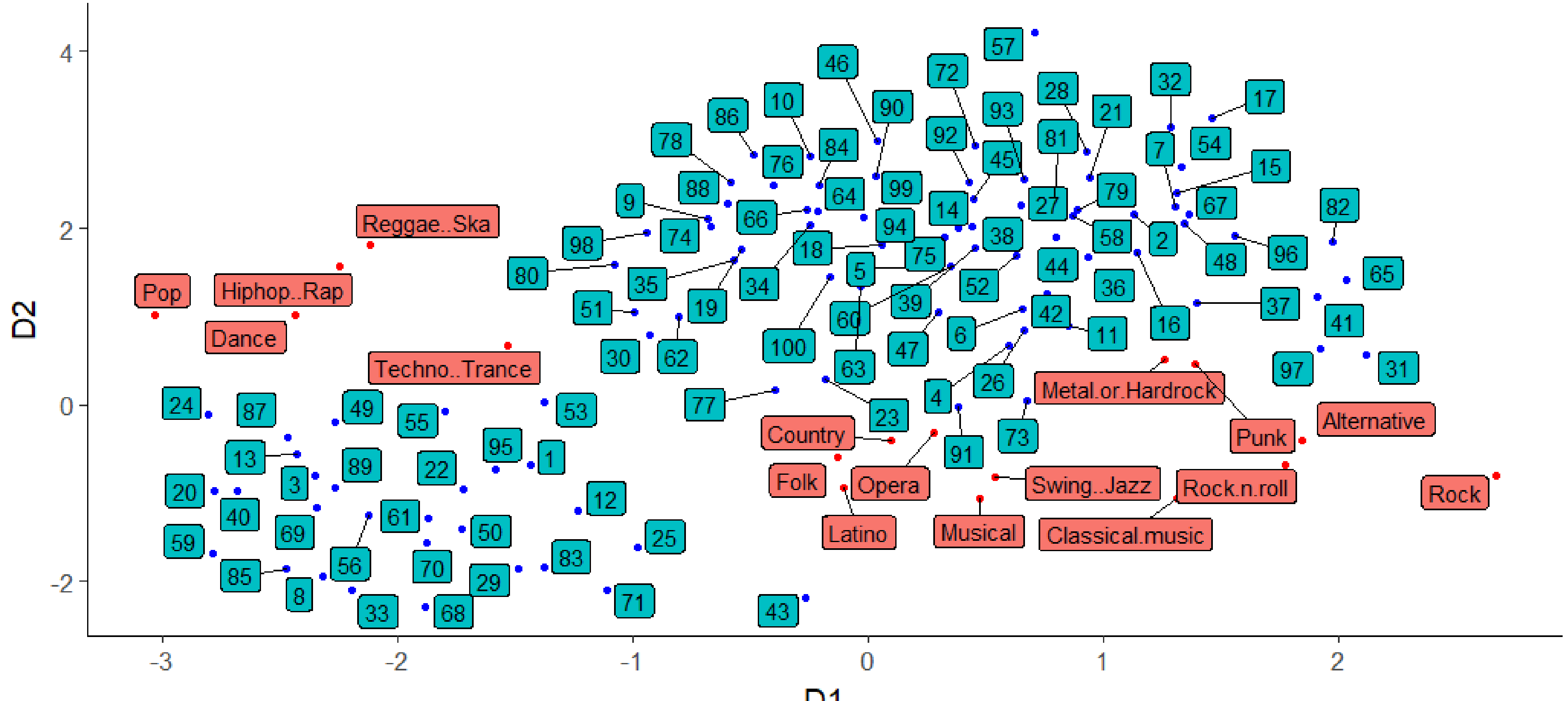
```
km.car<-kmeans(CarData[, 2:7],centers=3, nstart=20)
#Let's add the clustering solution to the previously created PCA scatterplot
NewConfig$cluster<-as.factor(km.car$cluster)
#Now plot using ggplot
ggplot(NewConfig, aes(x=PC1, y=PC2)) +
  geom_point() + theme_bw() +
  geom_label_repel(aes(fill = cluster,label=Brand))
```



1970s-1990s

- More complex methodologies
- Still relatively small datasets
- Joint space methods
- More complex clustering (probabilistic, overlapping, etc.)
- Work mostly published in The Journal of Marketing Research and Marketing Science

Example: Youth Preference Dataset (Joint Space Visualization)



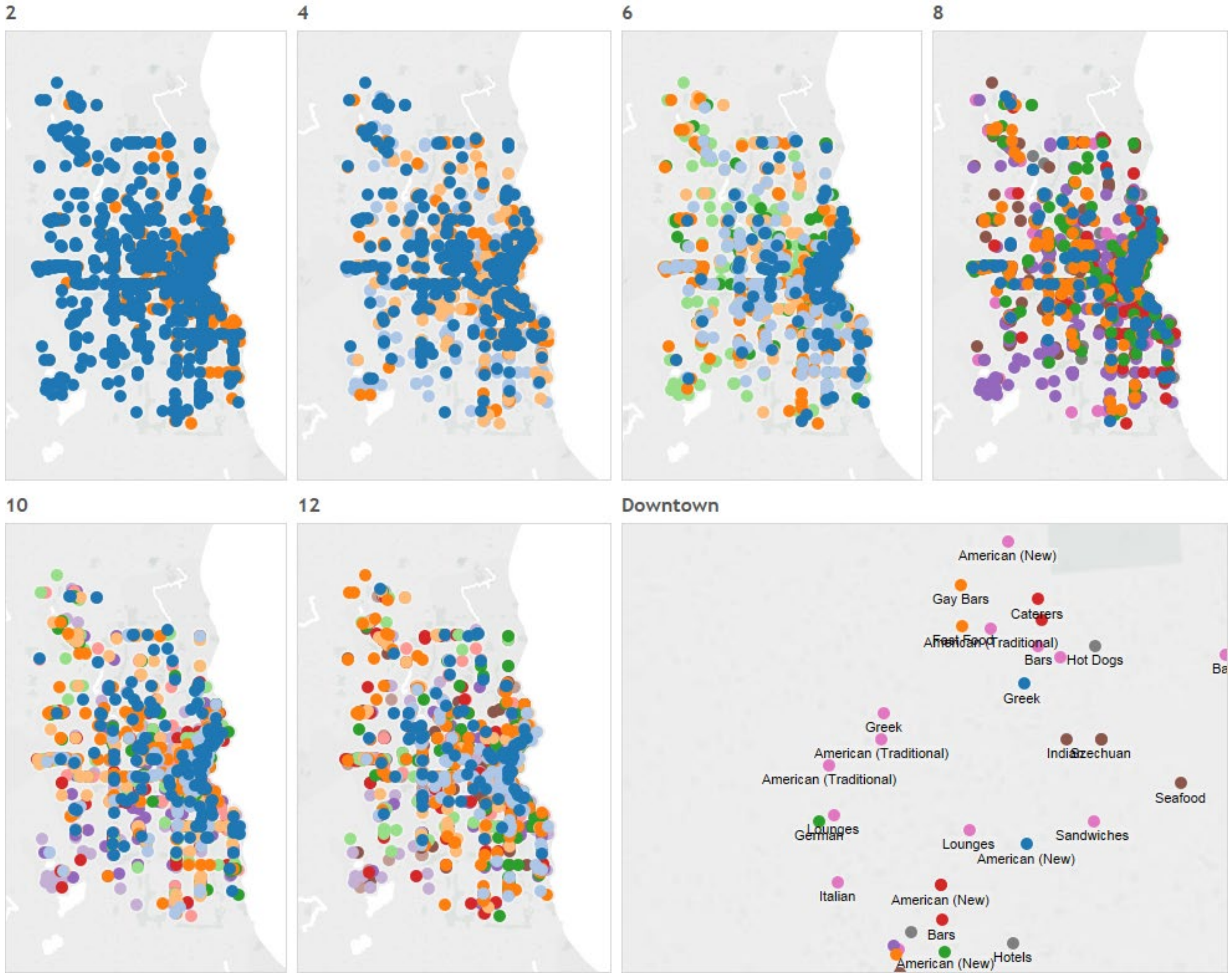
2000s

- The big data era (Volume, velocity, variety, veracity, value)
- Large, unstructured, datasets
- Marketing data from social media, websites, blogs, etc.
- Most marketing work is mostly mid sized
- Work mostly published in The Journal of Marketing Research and Marketing Science, but also Information Systems/Computer Science

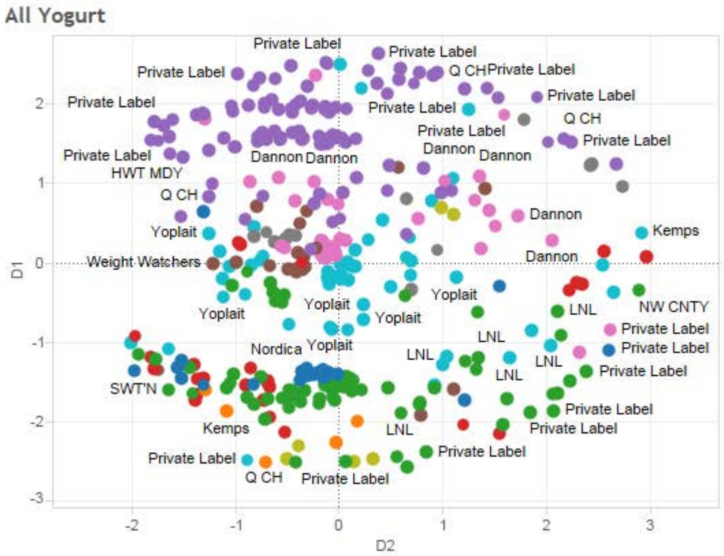
Some papers

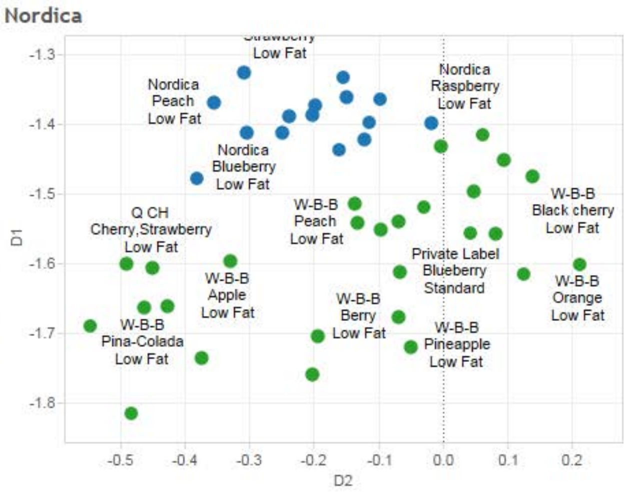
- France, S. L., & Ghose, S. (2016). An analysis and visualization methodology for identifying and testing market structure. *Marketing Science*, *35*(1), 182-197.
- Ringel, D. M., & Skiera, B. (2016). Visualizing asymmetric competition among more than 1,000 products using big search data. *Marketing Science*, *35*(3), 511-534.
- Ringel, D. M. (2023). Multimarket membership mapping. *Journal of Marketing Research*, 00222437221110460.
- Malhotra, P., & Bhattacharyya, S. (2022). Leveraging cofollowership patterns on social media to identify brand alliance opportunities. *Journal of Marketing*, *86*(4), 17-36.
- France, S. L., & Akkucuk, U. (2021). A review, framework, and R toolkit for exploring, evaluating, and comparing visualization methods. *The Visual Computer*, *37*(3), 457-475.

Example: France and Ghose (2016)



Example: France and Ghose (2016), cont.





BrandsCluster

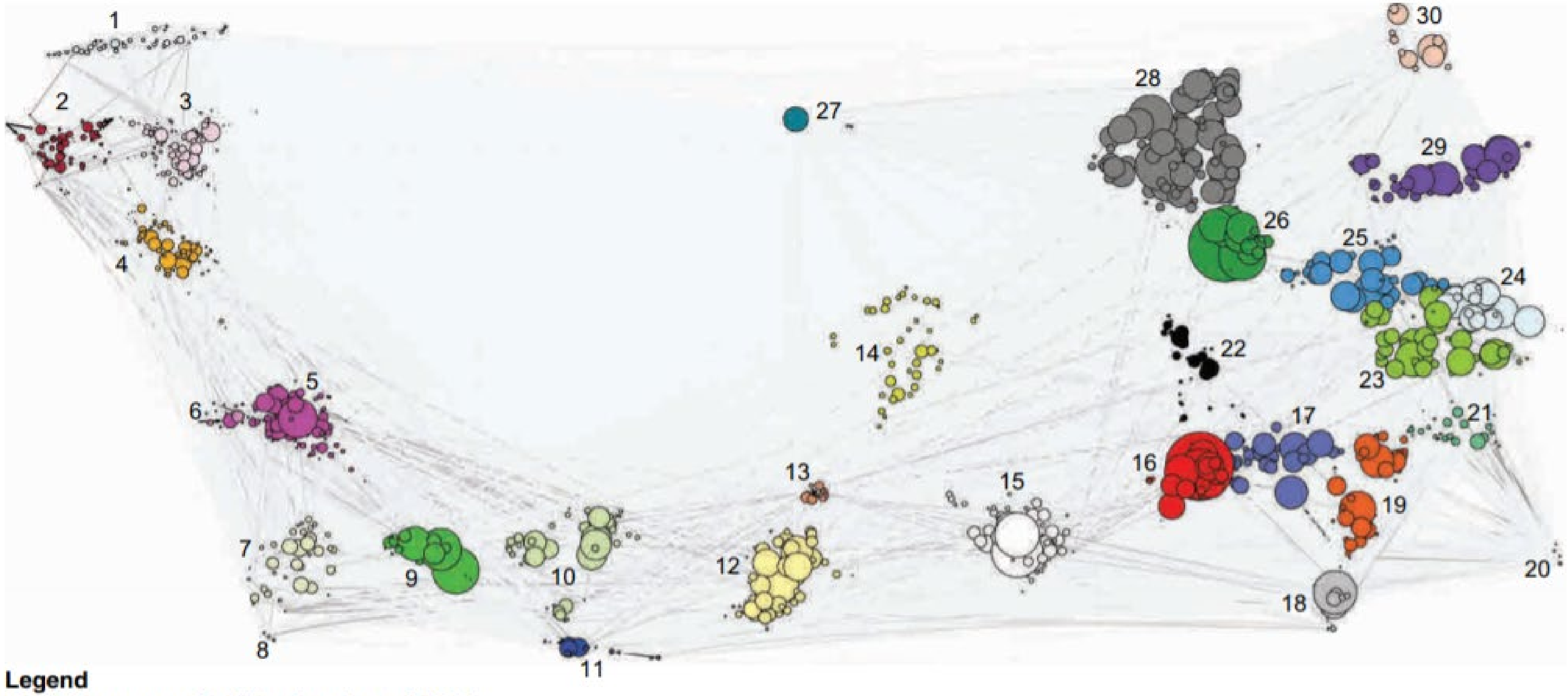
Brand	Cl10 1	2	3	4	5	6	7	8	9	10
Breyers	1		2			2				2
Dannon	1			3	16	6	25		1	8
Elm Grove					5					
Hiland Dairy					9					
HWT MDY					6					
Kemps		4	1	3						1
LNL			6							4
Nordica	16			3						
NW CNTY			3							
Private Label	1		21	1	50		5	4	4	4
Q CH	2	1	19	1	5	1				
Swiss Valley..			1	9						
SWT'N	3		3							
W-B-B			14	1						
Weight Watc..						10	1			
Whitney's				4						1
Yami Yoghurt					10					
Yoplait	1		3	1			5	9	1	36

TypeCluster

Type	Fat	Cl10 1	2	3	4	5	6	7	8	9	10
Custard	Low Fat					1		1	1		2
	Standard										8
Frozen	Low Fat					2		3			7
	Standard			1				2			
Frozen,Fruit on the Bottom	Low Fat			3							1
Fruit on the Bottom	Low Fat	10		6		10		19		1	1
	Non Fat								6		
	Standard	1		3		7	2				8
Soft whip	Low Fat	8		1	9	13					
	Standard			2	1	1				3	
Soft whip,Fruit o..	Low Fat					5					
Standard	Low Fat	1	5	35	10	40	7	3	1	2	6
	Non Fat	3		3	1	2	10	1			
	Standard	2		19	5	20		7	5		23

Example: Ringel and Skiera (2016)

Figure 5 **Visualization of Asymmetric Competitive Market Structure Map of 1,124 LED-TVs**



Legend
Bubbles represent individual products (SKUs)
Bubble color indicates submarket membership
Bubble size indicates global competitive asymmetry (consideration frequency)
Arrows represent local competitive asymmetry and point at competitors of the product they originate in
Arrow weight indicates how intense a competitive relationship is: the darker and thicker the arrow, the more intense the relationship
Submarkets are numbered 1 through 30

Example: Ringel (2023)

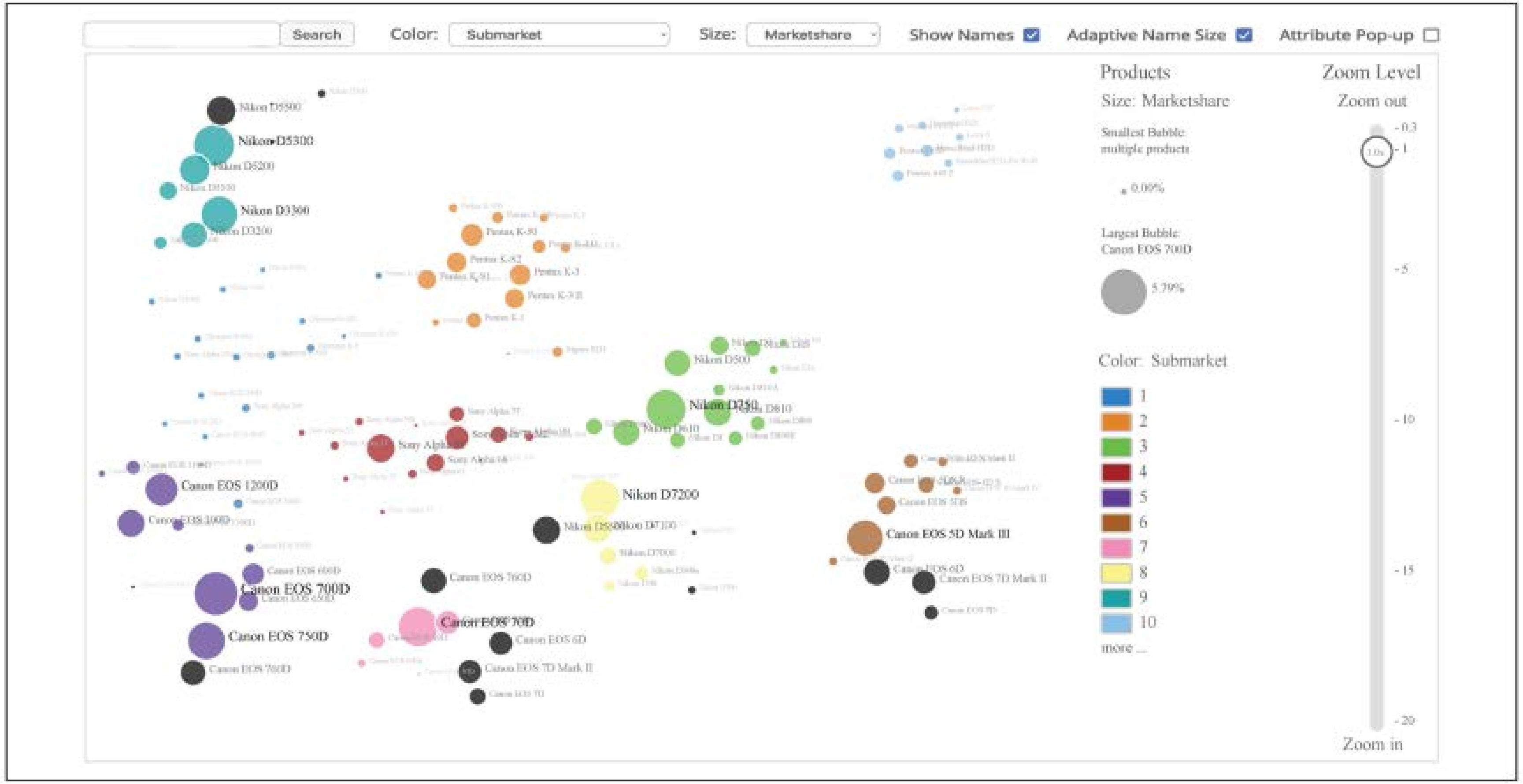


Figure 6. Multimarket Membership Map of the DSLR Camera Market Visualized in mapXP.
Notes: Axes are on the same scale. MMPs (indicated in black) are Nikon D5500, Canon EOS 760D, Nikon D3, Canon EOS 7D, Canon EOS 450D, Nikon D2X, Canon EOS 6D, Canon EOS 7D Mark II, and Nikon D700.

Example: Malhotra and Bhattacharyya (2022)

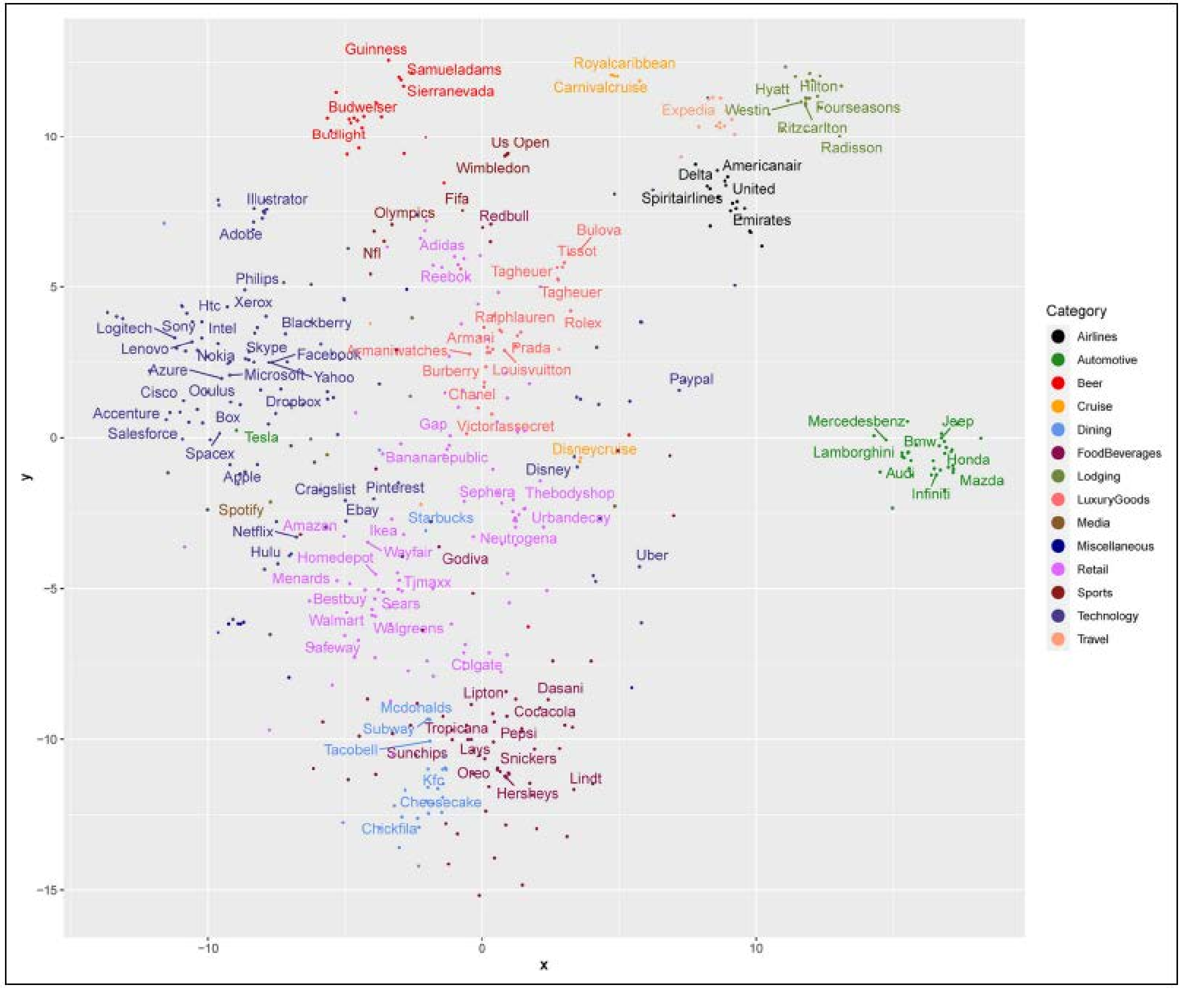
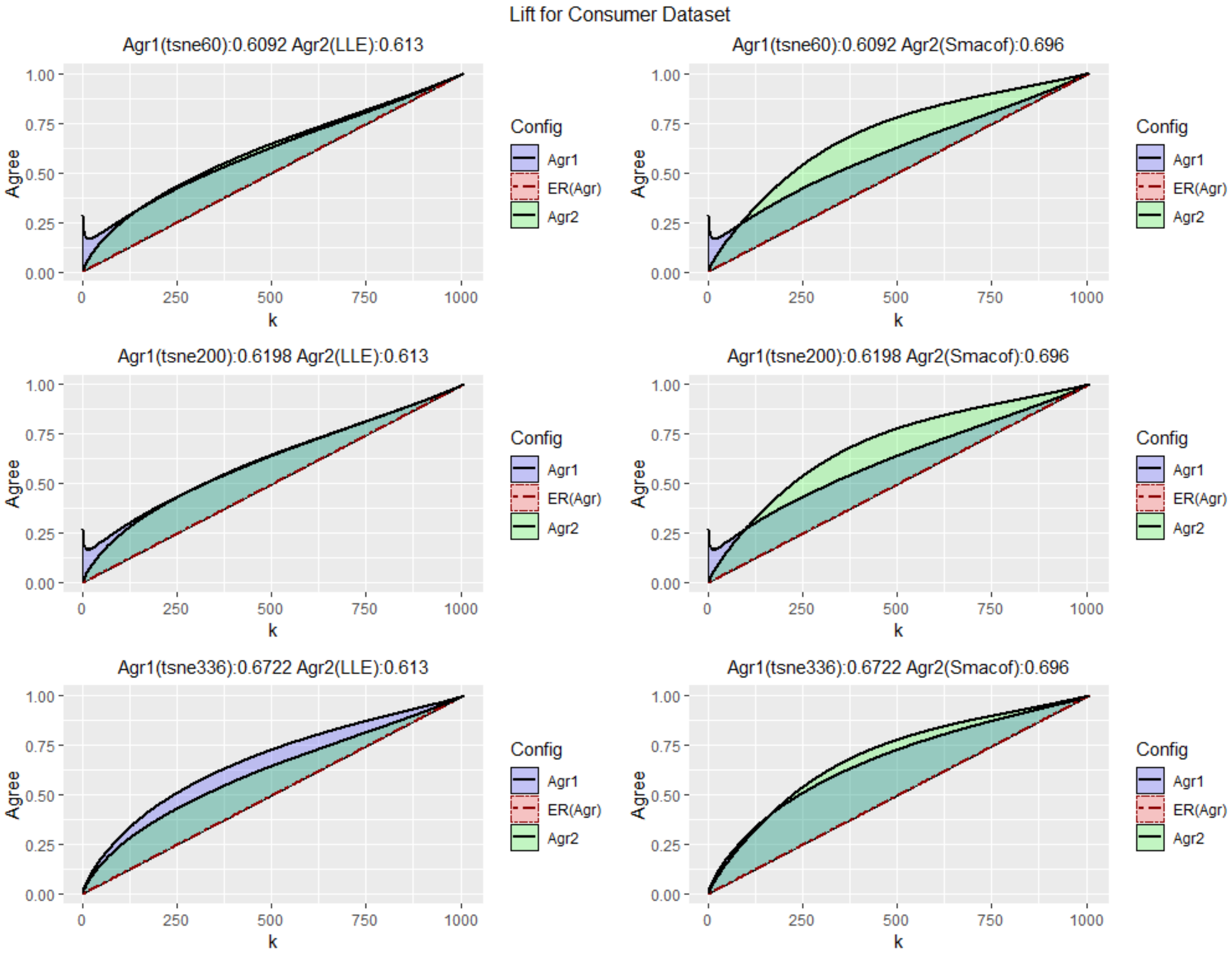


Figure 2. t-SNE map of the undirected brand network.[4]

Example: QVizViz (France and Akkucuk, 2021)



Some Issues

- Disconnect between academics in marketing and practitioners.
- Much applied marketing visualization and segmentation work is not done in marketing.
- Companies such as Google, Amazon, Facebook, etc. hire top Ph.D. graduates in machine learning data science/often ahead of academia.
- Yankelovich and Meer (2006) described a survey of 200 senior executives, where 59% reported carrying out a major segmentation exercise within the last two years, but only 14% reported that any real business value resulted from the exercise.

Ideas

- More User Testing
  - How do users use visualizations and segmentation solutions?
  - Which soultions give most insight?
- Improved evaluation of unsupervised learning
  - Confidence intervals for points (Iacobucci, Grisaffe, & DeSarbo, 2017)
  - Framework for analyzing visualization quality (France and Akkucuk, 2021)
  - Standard benchmarking testbeds
- Disseminate Widely
  - Integrate with Tableau and PowerBI
  - User Feedback
  - Make R packages available
  - Attend practitioner conferences

Psychoinformatics

Dr. Jarrod Moss

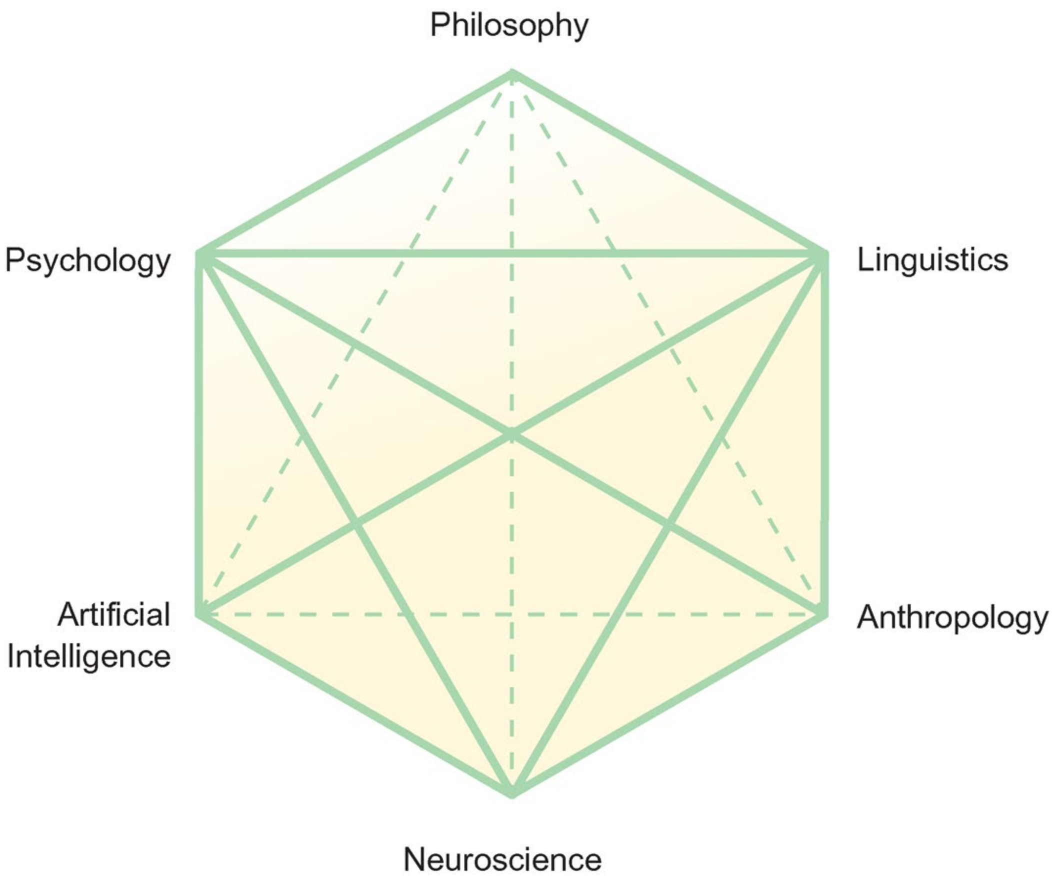
Associate Professor, Psychology

Psychoinformatics

Jarrod Moss
Department of Psychology

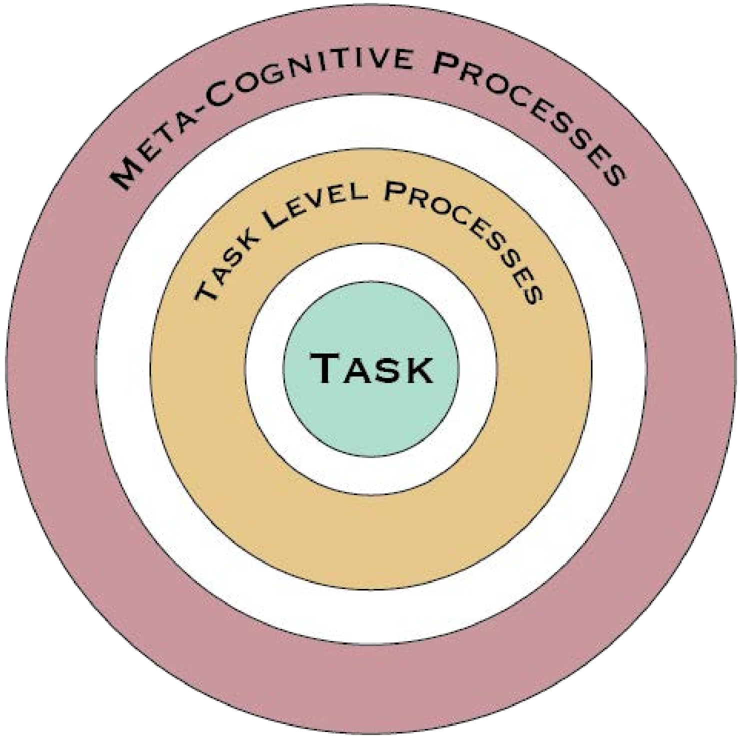
Psychoinformatics and Cognitive Science

- Harnessing rich data sources and informatics methods in ways that reshape the ways we collect, organize, and analyze data (Yarkoni, 2012)
- Aren't we already doing that?
  - Google Scholar - 990 hits for psychoinformatics vs. 1.8 million hits for bioinformatics
  - Cognitive science has been an interdisciplinary field
- What is changing?
  - Pace of development
  - The fields that are a part of this development



How do people develop effective strategies?

- Strategy is a coherent set of steps to carry out a task in which different sets of steps are possible
- Task complex enough to support multiple strategies
- Track strategy use
  - Explicit reports?

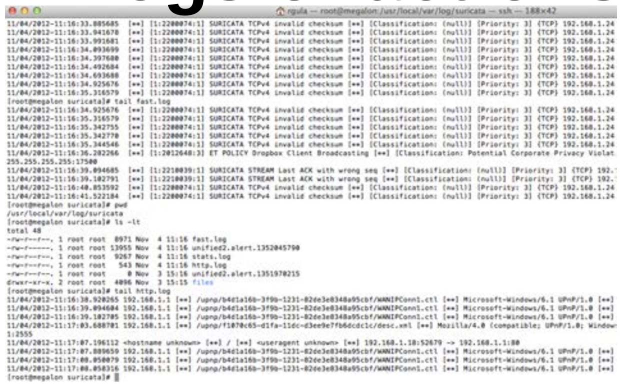


Strategies in a complex decision-making task

- What do you work on next?
- Tasks have different
  - Rewards/value
  - Time requirements
  - Deadlines
  - Penalties for not meeting the deadline

Using machine learning to infer strategies

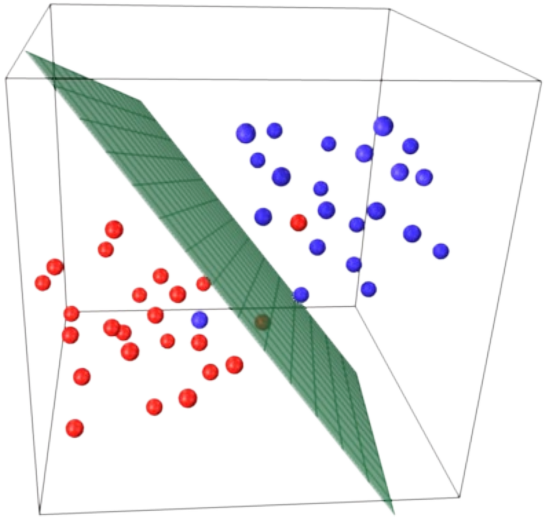
Logs of actions



Decisions

Object Name	Time/ Deadline	Points	Penalty
DAX	001/060	203	-45
FUB	047/120	402	-100
TIR	029/030	103	-38
WOH	015/020	153	-10
JIQ	100/110	237	-57

Classifiers



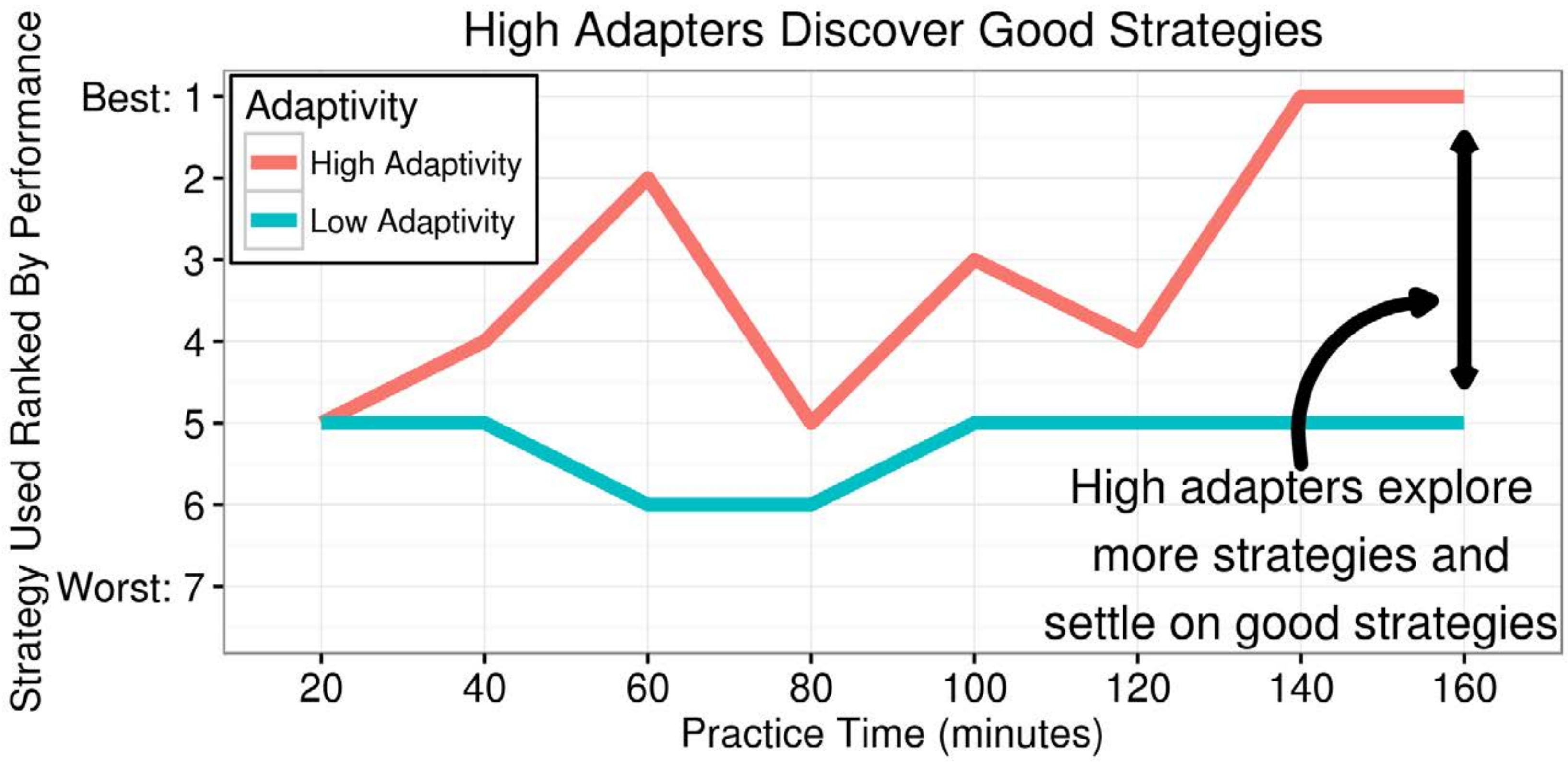
Strategy Elements

Attribute	Importance
Points	.8
Deadline	.07
Priority	.7
Position in Queue	.01

Moss, J., Wong, A. Y., Durriseau, J. A., & Bradshaw, G. L. (2022). Tracking strategy changes using machine learning classifiers. *Behavior Research Methods*, 1-23.

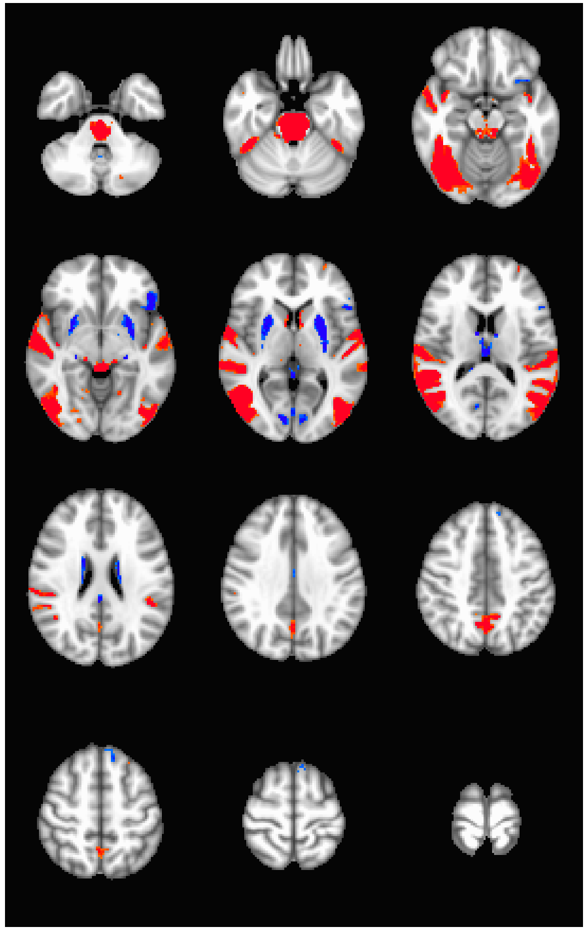
Are strategies important?

- Adapting strategy to improve performance
- In response to task changes or current performance



Using machine learning to infer strategies

- Explicit reports vs. classifier-based predictions
  - 51% people predicting their own decisions
  - 86% ML-based predictions
- What questions can we address now?
  - How and when do people incorporate task features into their strategies?
  - Are there ability differences that are related to strategy adaptivity?
  - What are effective training methods for improving strategy adaptivity?
  - Are there brain network interactions associated with strategy adaptivity?



Stepping back

- Our computational models (often) generate high-density data
- Our data has historically been either low-density or high-density but difficult to map onto models/theories
- New tools/techniques enable us to use high-quantity, high-density data effectively

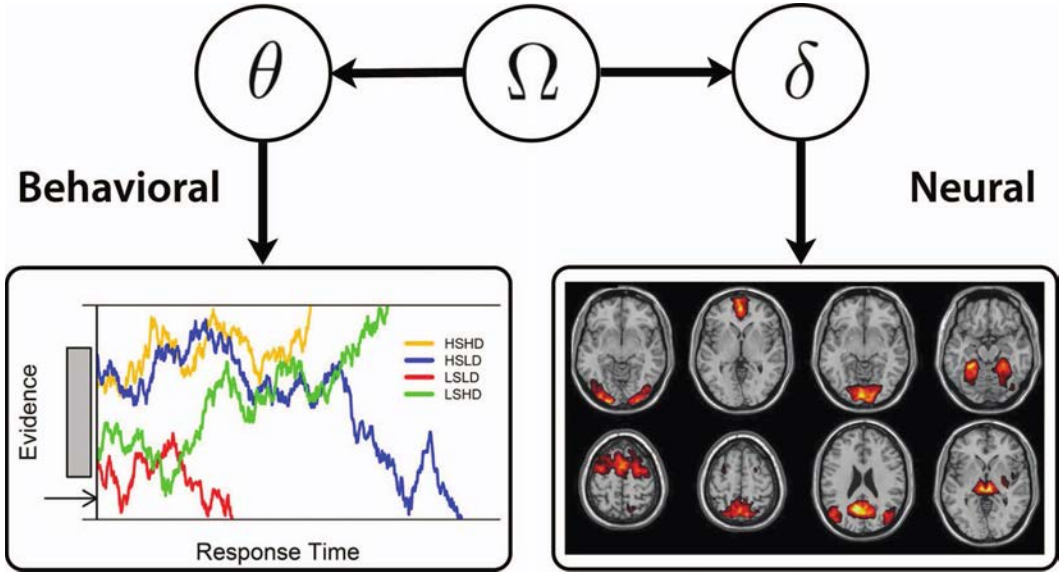


Figure from Turner et al. (2015)

Facilitating transformative science

- Data sources
  - Open science, large quantities of available data (HCP)
  - Diverse data sources - traditional in-person lab, online studies, eye tracking, neuroimaging - functional magnetic resonance imaging, electroencephalography
- Computational resources
- Training - students need background in psychological theory, models, and data manipulation
- Forums to facilitate development and transfer of techniques

Computational Intelligence

Dr. Shahram Rahimi
Department Head,
Computer Science & Engineering

Quantum Computing

Dr. Yaroslav Koshka
Professor, Electrical & Computer Engineering

Quantum Computing
and
Quantum ML

Yaroslav Koshka

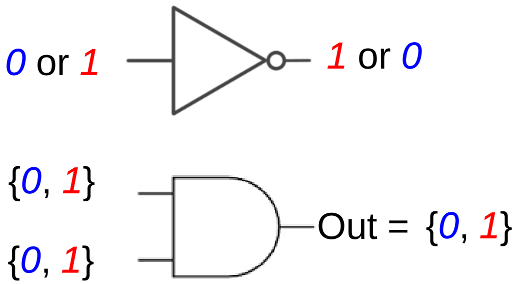
Dpt of Electrical and Computer Engineering, MSU

MISSISSIPPI STATE UNIVERSITY

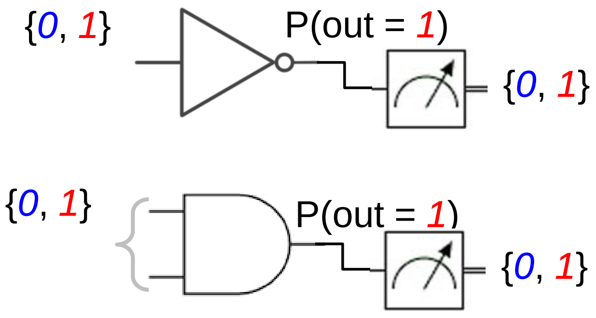
Apples-to-oranges or Apples-to-Bees:

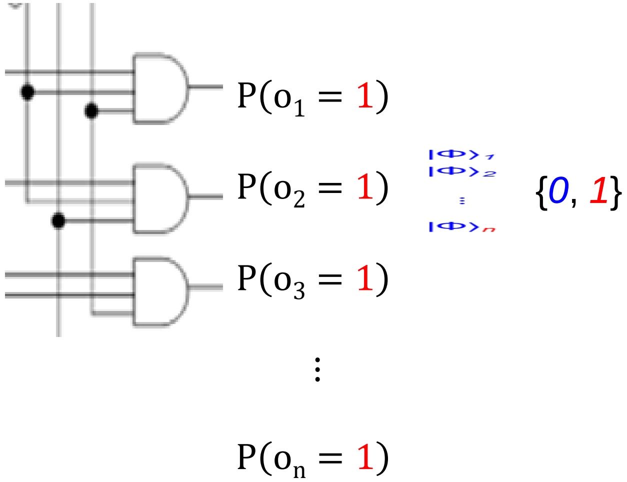
Classical, classical Probabilistic and Quantum computers

Classical gates

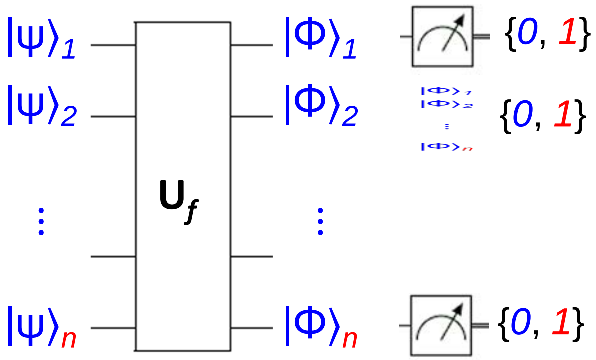


Classical Probabilistic gates

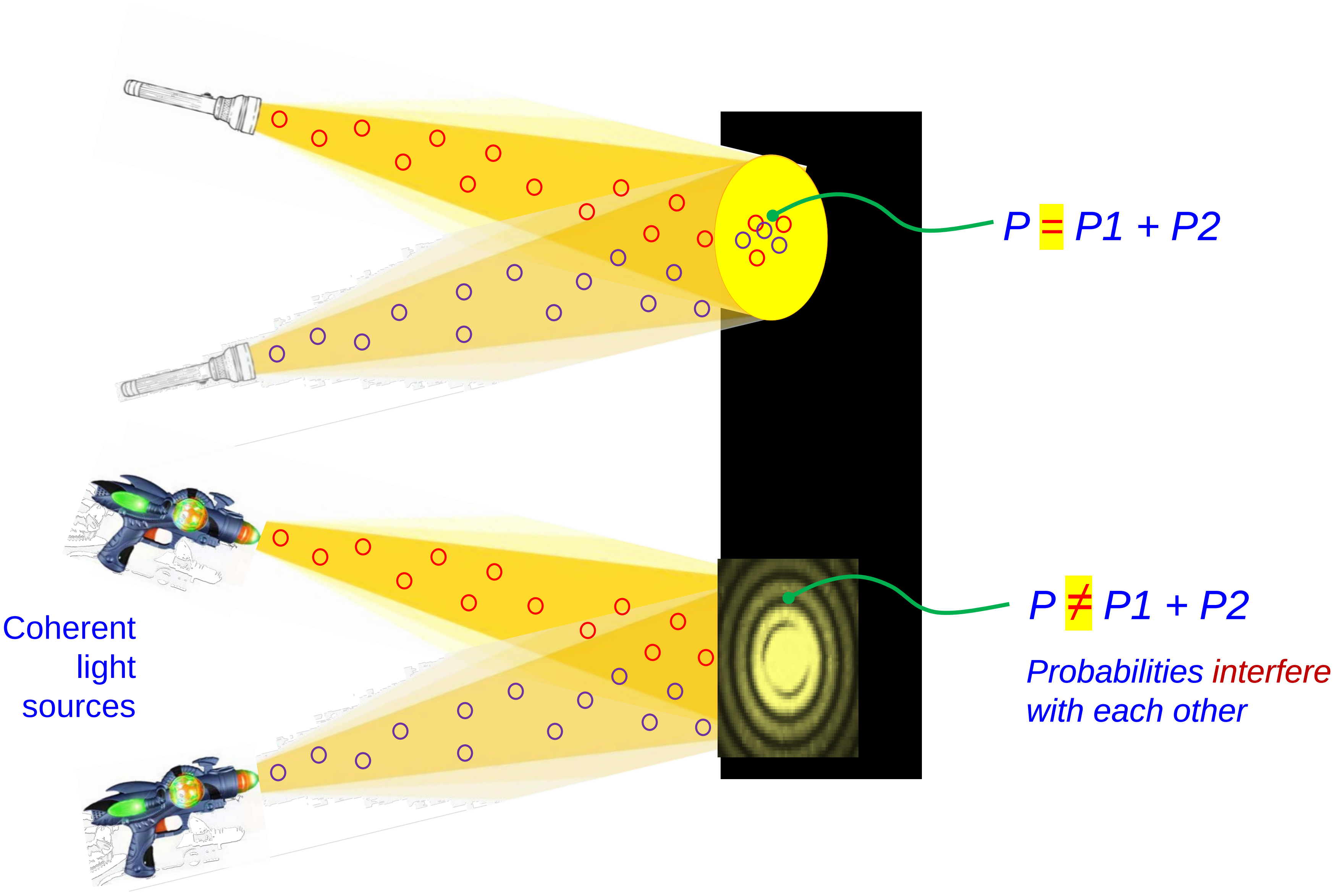




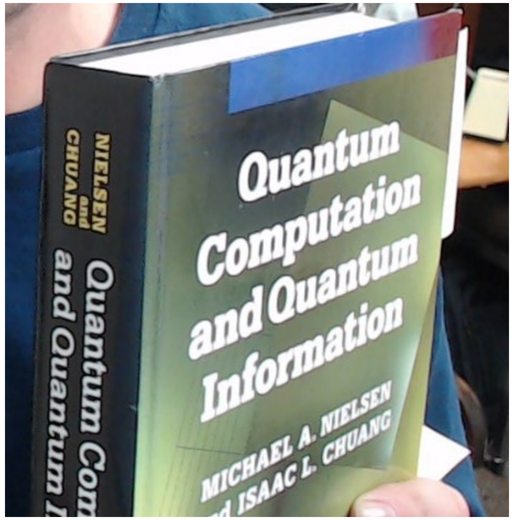
Quantum gages



Classical vs "quantum" probabilities



Developing Quantum Algorithms

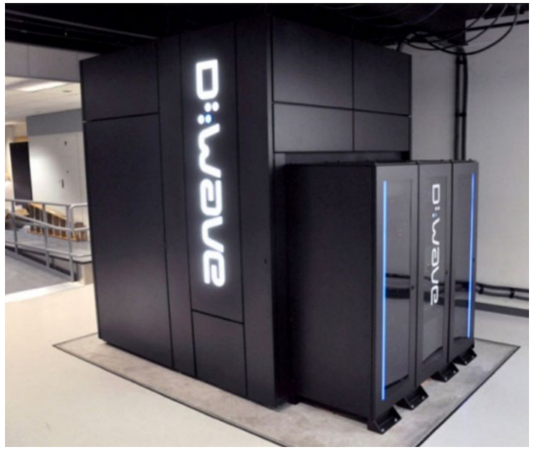


- ❑ A QC "bible", 23 yeas old, 700 plus pages
- ❑ Many more books and review articles.
- ❑ Comprehensive lists of Q algorithms.

?: Should we do nothing until good-enough QC arrives?

- ➢ Noisy Intermediate Scale Quantum (NISQ) – need different algorithms, now is the time to develop them.
- ➢ ML promises to be a killer up for QC and Q Algorithms (Q ML – a new field).
- ➢ Quantum Annealers – were not even mentioned in the "QC bible" but became the first ever commercial QCs.

Q Annealers (D-Wave Inc.)



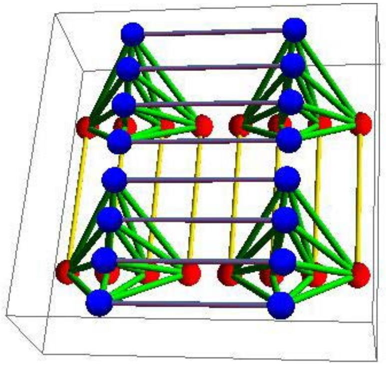
~2 years ago:

- *Chimera architecture – 2048 superconducting flux qubits*

Today:

- *Pegasus architecture – 5000 qubits*

2 x 2 Chimera graph (4 unit cells shown)



Only Two main applications:

- *Optimization (find the global extremum)*
- *Sampling from probability distributions.*

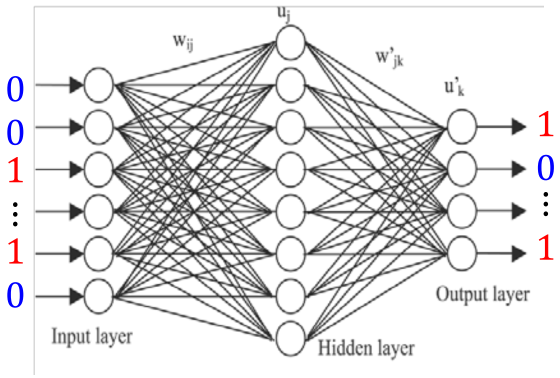
Benefits for non-physics G/UG research:

- *Can use as a black box; can survive without knowledge of QM / QC.*
- *Straightforward engineering applications.*

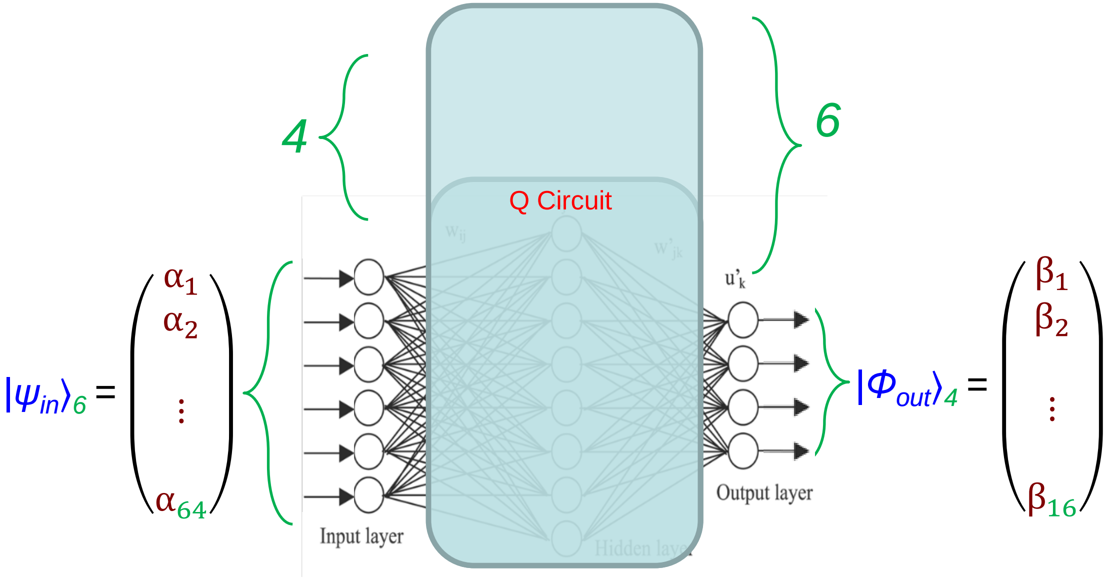
Gated QCs

- ❑ Expectation of a general-purpose QC.
- ❑ IBM Gated: only 127 qBits in 2021, 400 announced *(D-Wave QA: 5000 qBits)*
- ❑ Requires understanding of the Q Circuit Model of computation (both for QC and Q ML)

Classica feedforward NN



“Quantum feedforward NN”



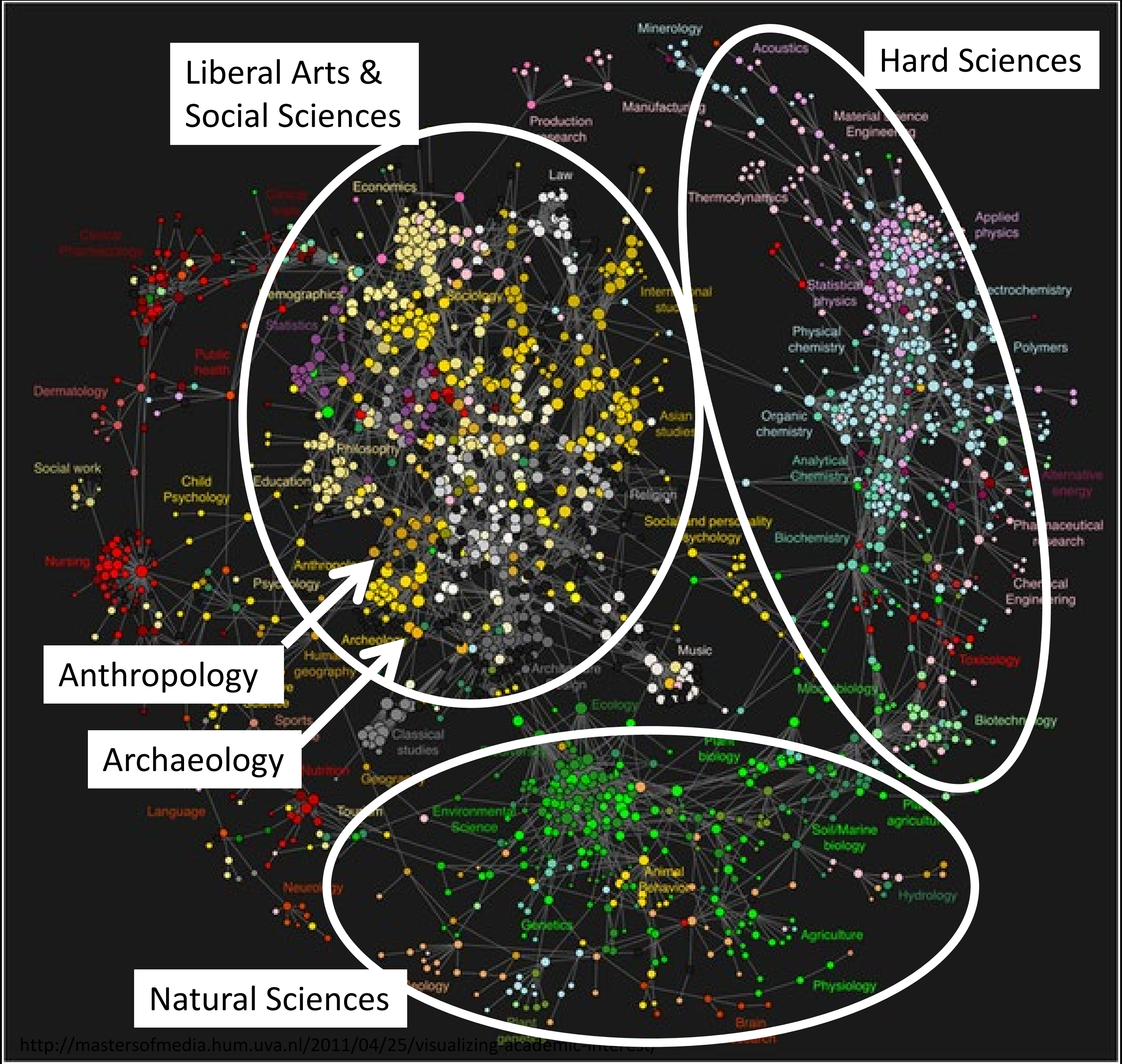
Can MSU participate in QC materials / hardware research?

- We do not have many capabilities to fabricate and test qBits.
- However, many big NSF-funded materials/device projects require multidisciplinary teams involving computational and data scientists.
- AI/ML and data science has become common in multidisciplinary proposals in materials/devices.

Social Data Analytics

Dr. Shane Miller

Associate Professor,
Anthropology & Middle Eastern Cultures

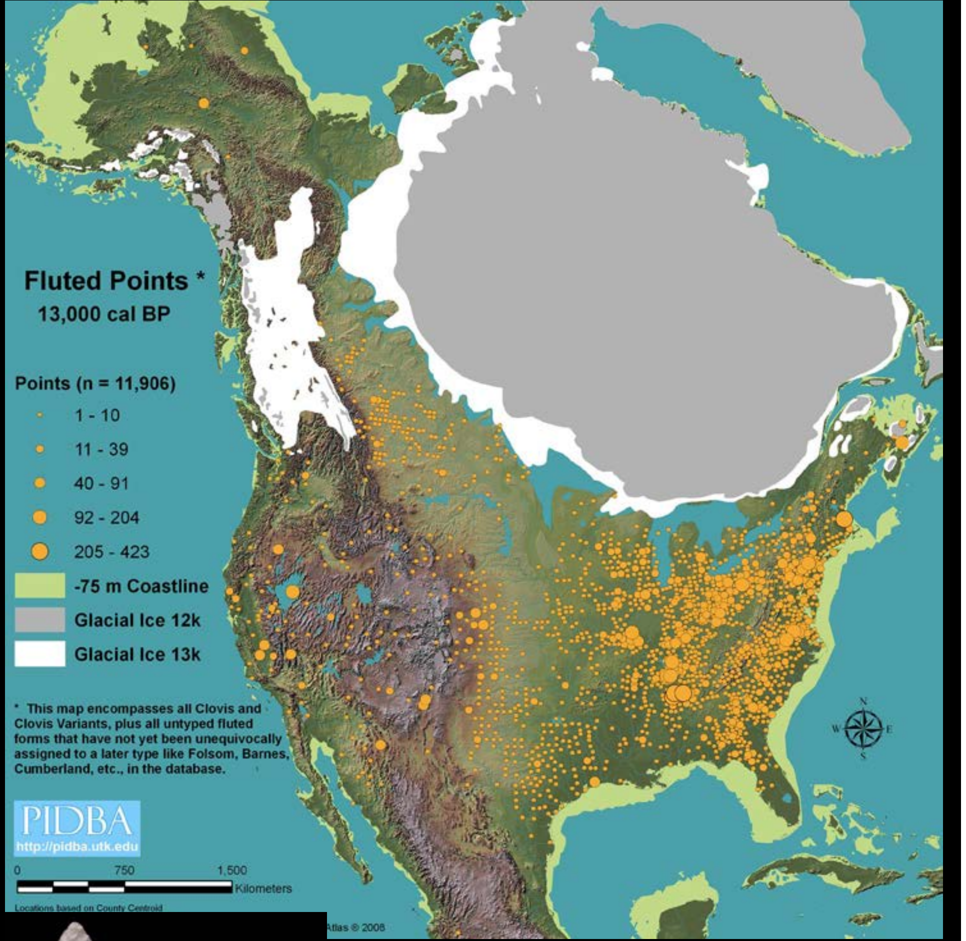

Liberal Arts & Social Sciences
Hard Sciences
Natural Sciences
Anthropology
Archaeology
Economics
Law
Demographics
Statistics
Philosophy
Education
Religion
Music
Classical studies
Sports
Language
Nutrition
Tourism
Social work
Child Psychology
Nursing
Dermatology
Public health
Minerology
Acoustics
Production research
Thermodynamics
Applied physics
Physical chemistry
Organic chemistry
Analytical Chemistry
Biochemistry
Polymers
Toxicology
Biotechnology
Ecology
Environmental Science
Soil/Marine biology
Animal Behavior
Genetics
Agriculture
Physiology
Hydrology
http://mastersofmedia.hum.uva.nl/2011/04/25/visualizing-academic-interest/

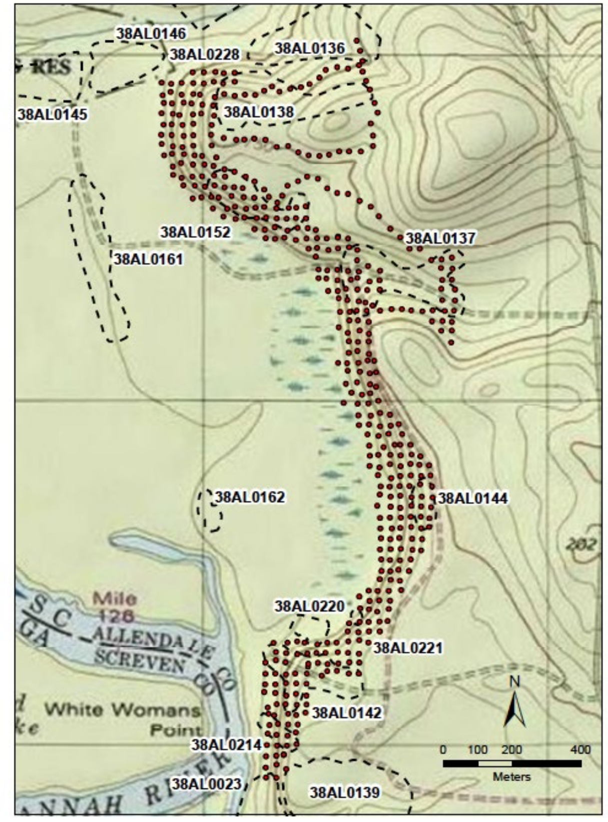
Topper, SC

Blackwater Draw, NM

Fin del Mundo, SON

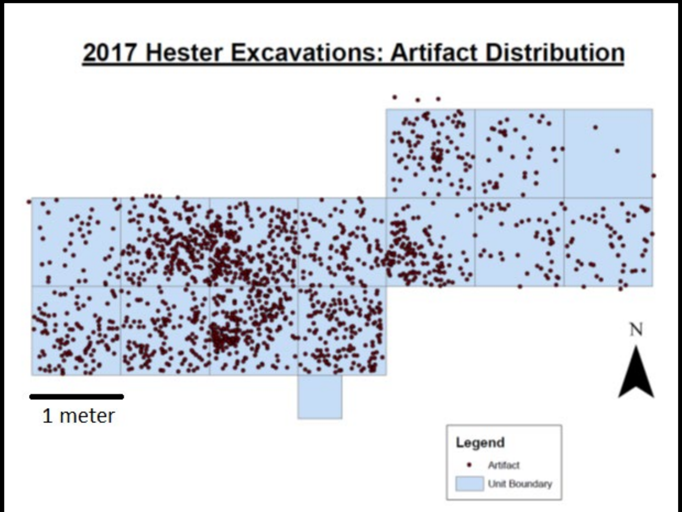
Cumberland River, TN





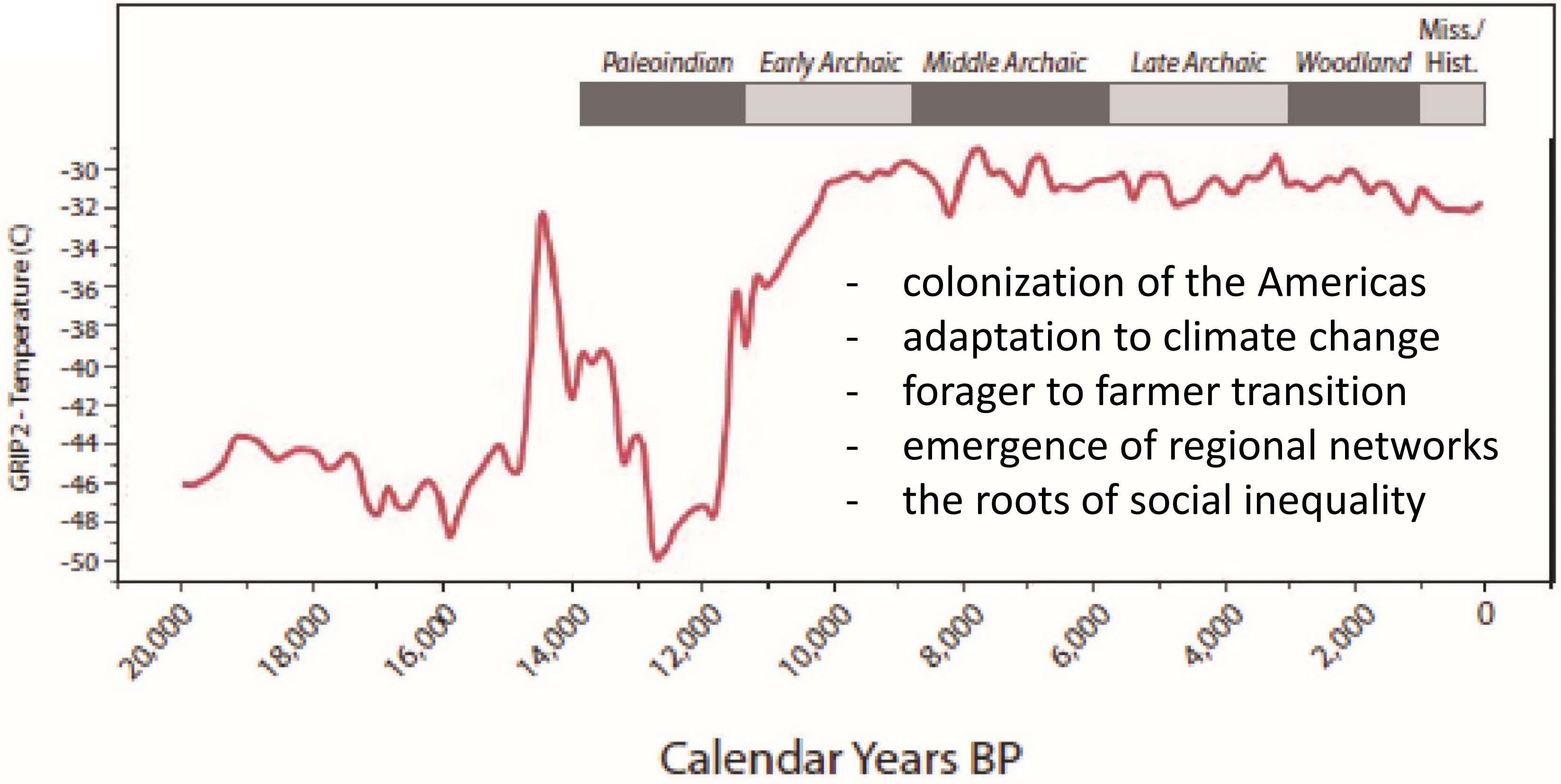
Clovis Points – Blackwater Draw, NM

C. Vance Haynes
Cast Collection
http://www.argonaut.arizona.edu/projects/castcollection.htm

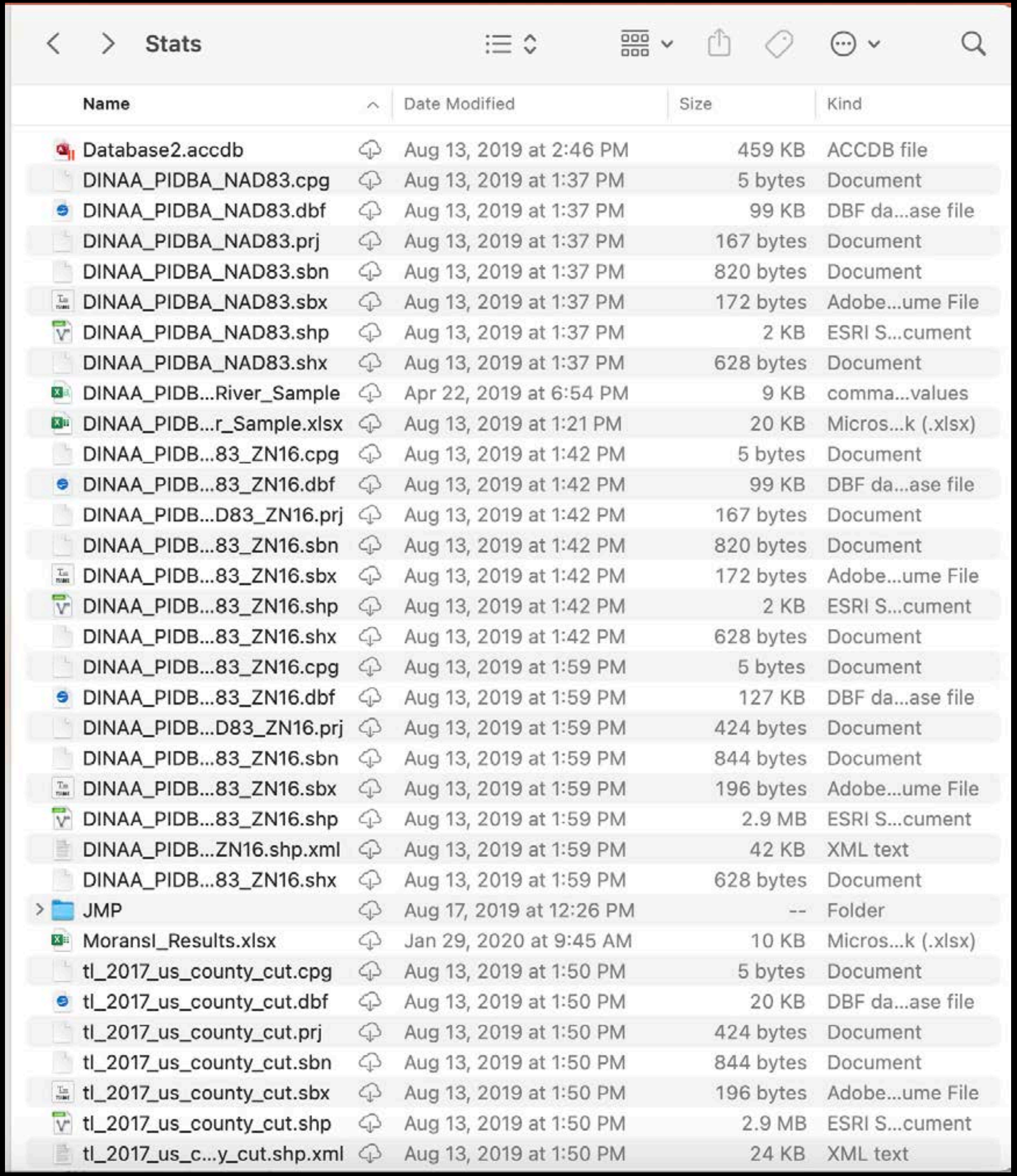


GRAND CHALLENGES FOR ARCHAEOLOGY

Keith W. Kintigh, Jeffrey H. Altschul, Mary C. Beaudry, Robert D. Drennan, Ann P. Kinzig, Timothy A. Kohler, W. Fredrick Limp, Herbert D. G. Maschner, William K. Michener, Timothy R. Pauketat, Peter Peregrine, Jeremy A. Sabloff, Tony J. Wilkinson, Henry T. Wright, and Melinda A. Zeder

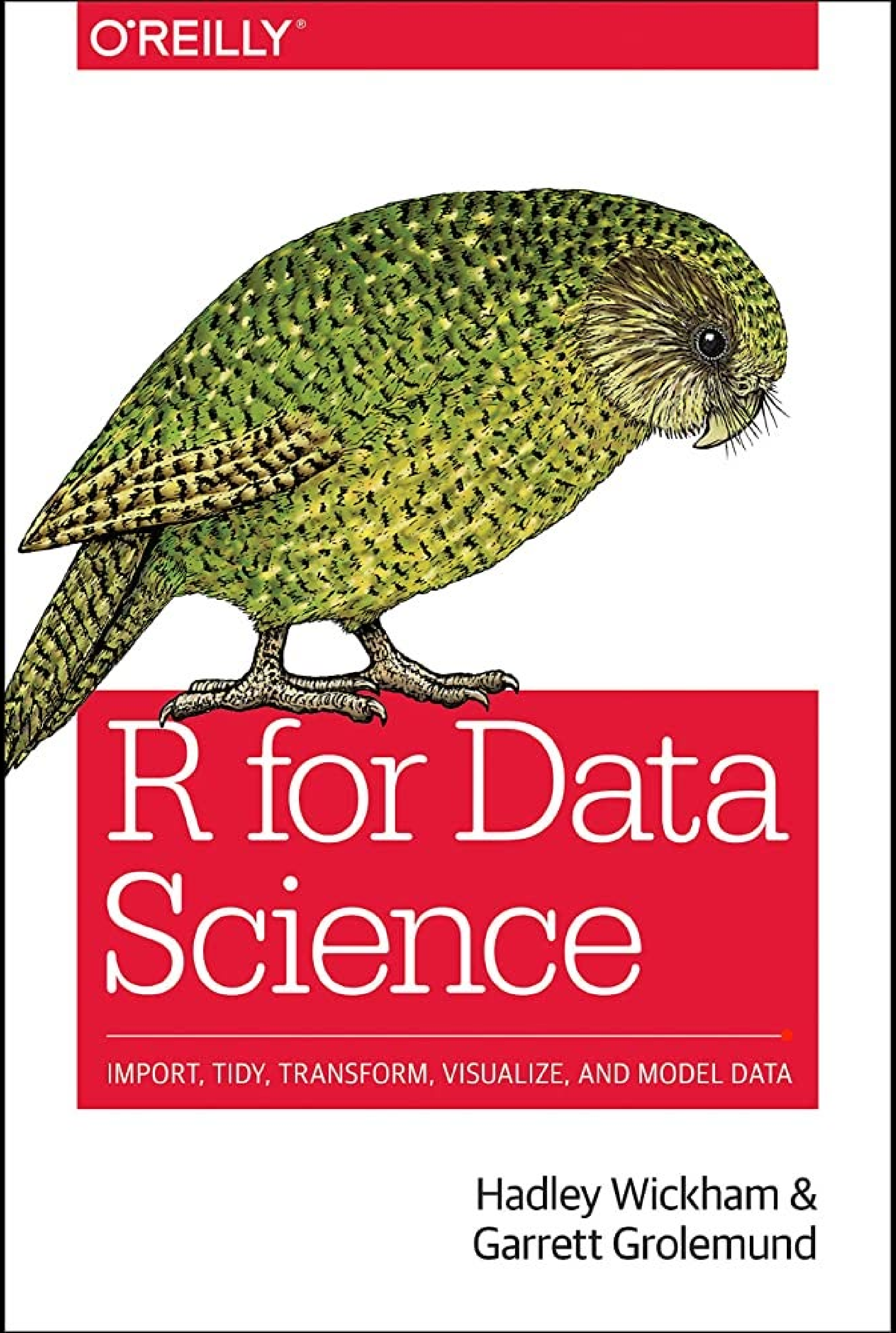


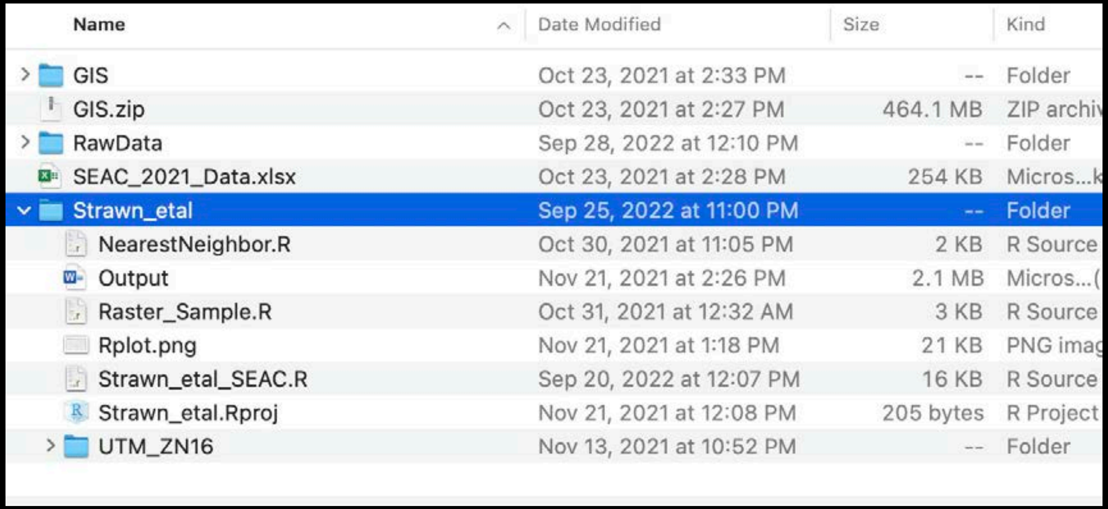
- colonization of the Americas
- adaptation to climate change
- forager to farmer transition
- emergence of regional networks
- the roots of social inequality

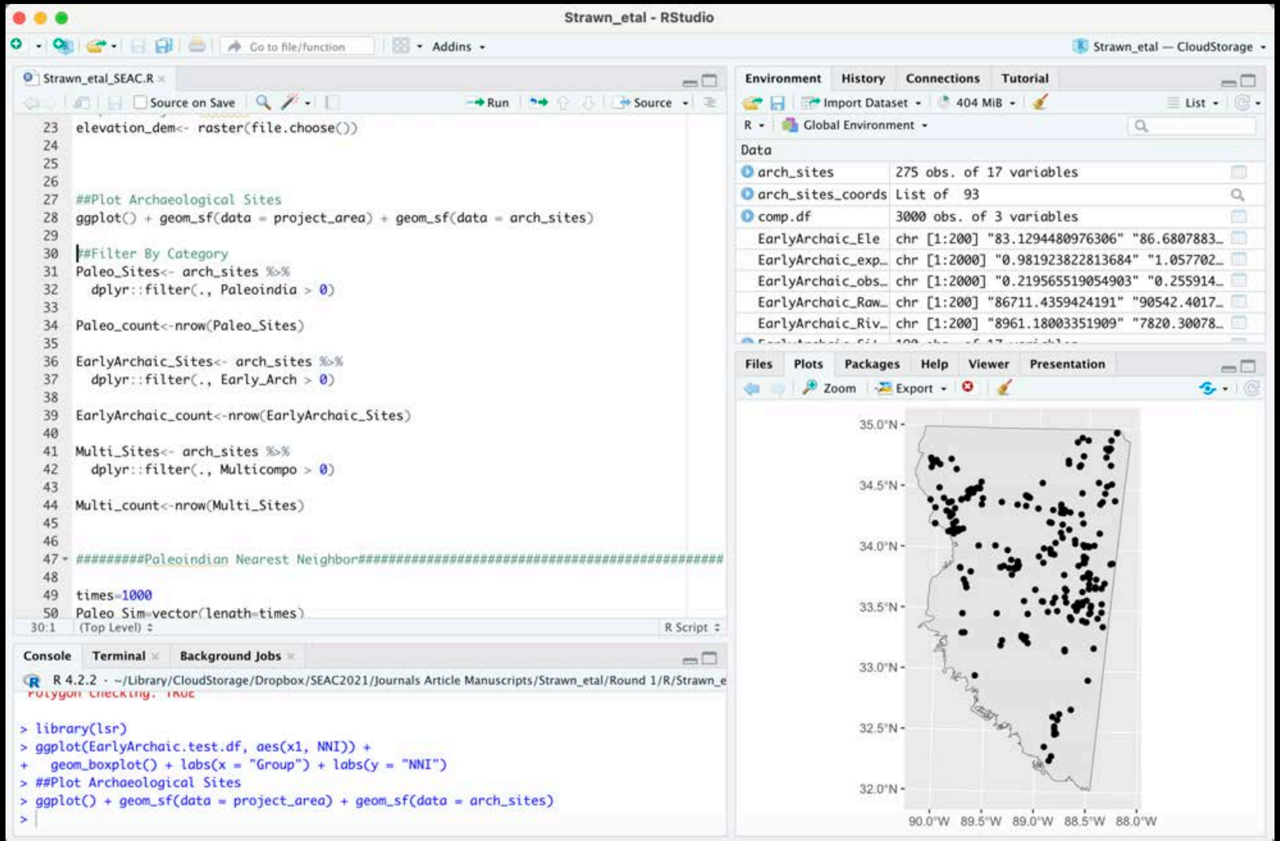


Stats

Name	Date Modified	Size	Kind
Database2.accdb	Aug 13, 2019 at 2:46 PM	459 KB	ACCDB file
DINAA_PIDBA_NAD83.cpg	Aug 13, 2019 at 1:37 PM	5 bytes	Document
DINAA_PIDBA_NAD83.dbf	Aug 13, 2019 at 1:37 PM	99 KB	DBF da...ase file
DINAA_PIDBA_NAD83.prj	Aug 13, 2019 at 1:37 PM	167 bytes	Document
DINAA_PIDBA_NAD83.sbn	Aug 13, 2019 at 1:37 PM	820 bytes	Document
DINAA_PIDBA_NAD83.sbx	Aug 13, 2019 at 1:37 PM	172 bytes	Adobe...ume File
DINAA_PIDBA_NAD83.shp	Aug 13, 2019 at 1:37 PM	2 KB	ESRI S...cument
DINAA_PIDBA_NAD83.shx	Aug 13, 2019 at 1:37 PM	628 bytes	Document
DINAA_PIDB...River_Sample	Apr 22, 2019 at 6:54 PM	9 KB	comma...values
DINAA_PIDB...r_Sample.xlsx	Aug 13, 2019 at 1:21 PM	20 KB	Micros...k (.xlsx)
DINAA_PIDB...83_ZN16.cpg	Aug 13, 2019 at 1:42 PM	5 bytes	Document
DINAA_PIDB...83_ZN16.dbf	Aug 13, 2019 at 1:42 PM	99 KB	DBF da...ase file
DINAA_PIDB...D83_ZN16.prj	Aug 13, 2019 at 1:42 PM	167 bytes	Document
DINAA_PIDB...83_ZN16.sbn	Aug 13, 2019 at 1:42 PM	820 bytes	Document
DINAA_PIDB...83_ZN16.sbx	Aug 13, 2019 at 1:42 PM	172 bytes	Adobe...ume File
DINAA_PIDB...83_ZN16.shp	Aug 13, 2019 at 1:42 PM	2 KB	ESRI S...cument
DINAA_PIDB...83_ZN16.shx	Aug 13, 2019 at 1:42 PM	628 bytes	Document
DINAA_PIDB...83_ZN16.cpg	Aug 13, 2019 at 1:59 PM	5 bytes	Document
DINAA_PIDB...83_ZN16.dbf	Aug 13, 2019 at 1:59 PM	127 KB	DBF da...ase file
DINAA_PIDB...D83_ZN16.prj	Aug 13, 2019 at 1:59 PM	424 bytes	Document
DINAA_PIDB...83_ZN16.sbn	Aug 13, 2019 at 1:59 PM	844 bytes	Document
DINAA_PIDB...83_ZN16.sbx	Aug 13, 2019 at 1:59 PM	196 bytes	Adobe...ume File
DINAA_PIDB...83_ZN16.shp	Aug 13, 2019 at 1:59 PM	2.9 MB	ESRI S...cument
DINAA_PIDB...ZN16.shp.xml	Aug 13, 2019 at 1:59 PM	42 KB	XML text
DINAA_PIDB...83_ZN16.shx	Aug 13, 2019 at 1:59 PM	628 bytes	Document
JMP	Aug 17, 2019 at 12:26 PM	--	Folder
MoransI_Results.xlsx	Jan 29, 2020 at 9:45 AM	10 KB	Micros...k (.xlsx)
tl_2017_us_county_cut.cpg	Aug 13, 2019 at 1:50 PM	5 bytes	Document
tl_2017_us_county_cut.dbf	Aug 13, 2019 at 1:50 PM	20 KB	DBF da...ase file
tl_2017_us_county_cut.prj	Aug 13, 2019 at 1:50 PM	424 bytes	Document
tl_2017_us_county_cut.sbn	Aug 13, 2019 at 1:50 PM	844 bytes	Document
tl_2017_us_county_cut.sbx	Aug 13, 2019 at 1:50 PM	196 bytes	Adobe...ume File
tl_2017_us_county_cut.shp	Aug 13, 2019 at 1:50 PM	2.9 MB	ESRI S...cument
tl_2017_us_c...y_cut.shp.xml	Aug 13, 2019 at 1:50 PM	24 KB	XML text


O'REILLY
R for Data Science
IMPORT, TIDY, TRANSFORM, VISUALIZE, AND MODEL DATA
Hadley Wickham &
Garrett Grolemund


Name
Date Modified
Size
Kind
GIS
Oct 23, 2021 at 2:33 PM
--
Folder
GIS.zip
Oct 23, 2021 at 2:27 PM
464.1 MB
RawData
Sep 28, 2022 at 12:10 PM
--
Folder
SEAC_2021_Data.xlsx
Oct 23, 2021 at 2:28 PM
254 KB
Strawn_etal
Sep 25, 2022 at 11:00 PM
--
Folder
NearestNeighbor.R
Oct 30, 2021 at 11:05 PM
2 KB
R Source
Output
Nov 21, 2021 at 2:26 PM
2.1 MB
Raster_Sample.R
Oct 31, 2021 at 12:32 AM
3 KB
R Source
Rplot.png
Nov 21, 2021 at 1:18 PM
21 KB
Strawn_etal_SEAC.R
Sep 20, 2022 at 12:07 PM
16 KB
R Source
Strawn_etal.Rproj
Nov 21, 2021 at 12:08 PM
205 bytes
R Project
UTM_ZN16
Nov 13, 2021 at 10:52 PM
--
Folder


Strawn_etal - RStudio
Strawn_etal_SEAC.R
Source on Save
Run
Source
elevation_dem<- raster(file.choose())
##Plot Archaeological Sites
ggplot() + geom_sf(data = project_area) + geom_sf(data = arch_sites)
##Filter By Category
Paleo_Sites<- arch_sites %>%
dplyr::filter(., Paleoindia > 0)
Paleo_count<-nrow(Paleo_Sites)
EarlyArchaic_Sites<- arch_sites %>%
dplyr::filter(., Early_Arch > 0)
EarlyArchaic_count<-nrow(EarlyArchaic_Sites)
Multi_Sites<- arch_sites %>%
dplyr::filter(., Multicompo > 0)
Multi_count<-nrow(Multi_Sites)
########Paleoindian Nearest Neighbor########
times=1000
Console
Terminal
Background Jobs
> library(lsr)
> ggplot(EarlyArchaic.test.df, aes(x1, NNI)) +
+ geom_boxplot() + labs(x = "Group") + labs(y = "NNI")
> ##Plot Archaeological Sites
> ggplot() + geom_sf(data = project_area) + geom_sf(data = arch_sites)
Environment
History
Connections
Tutorial
Import Dataset
Global Environment
Data
arch_sites
275 obs. of 17 variables
arch_sites_coords
List of 93
comp.df
3000 obs. of 3 variables
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
35.0°N
34.5°N
34.0°N
33.5°N
33.0°N
32.5°N
32.0°N
90.0°W
89.5°W
89.0°W
88.5°W
88.0°W

Early Holocene landscape use in the upper Tombigbee River valley

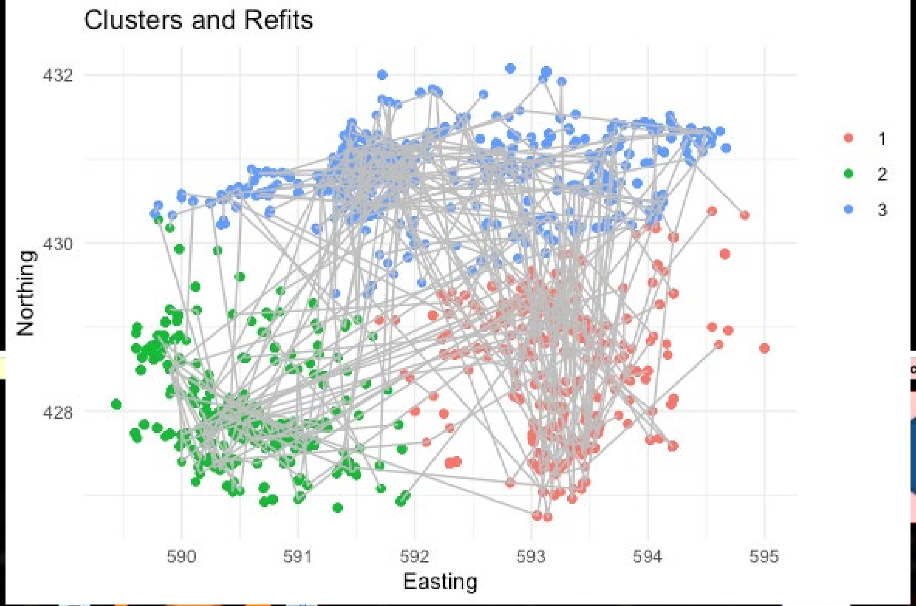
James L. Strawn [a,b], D. Shane Miller [b,c], Derek T. Anderson [b] and Stephen B. Carmody [d]

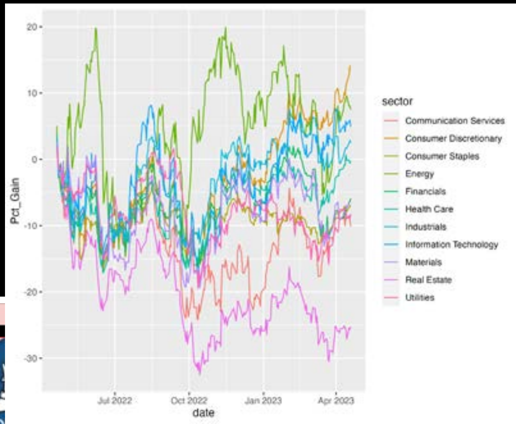
[a]Department of Anthropology, University of Georgia, Athens, USA; [b]Cobb Institute of Archaeology, Mississippi State University, Starkville, USA; [c]Department of Anthropology and Middle Eastern Cultures, Mississippi State University, Starkville, USA; [d]Department of Anthropology, Sociology, and Criminology, Troy University, Troy, USA

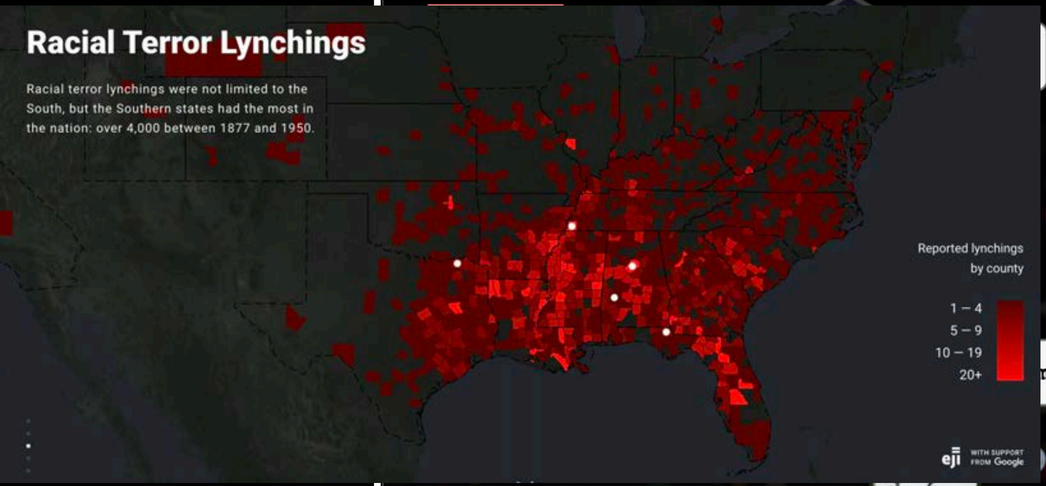
Fire on the Mountain: The Ideal Free Distribution and Early Hunter-gatherer Demography in the Tennessee River Drainage, USA

D. Shane Miller [a] and Stephen B. Carmody [b]

[a]Department of Anthropology and Middle Eastern Cultures, Mississippi State University, Starkville, MS, USA; [b]Department of Social Sciences, Troy University, Troy, AL, USA



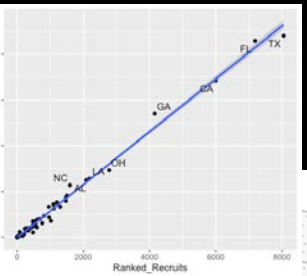


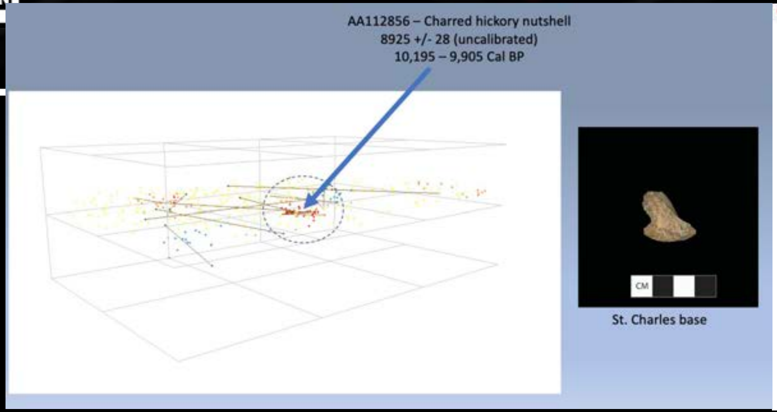


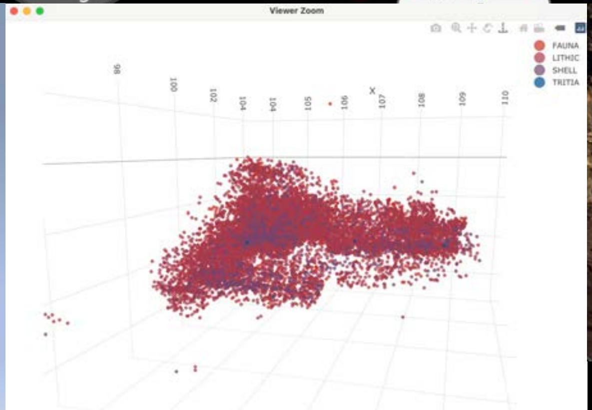
Understanding migration to protected area buffer zones in Costa Rica utilizing cultural consensus analysis

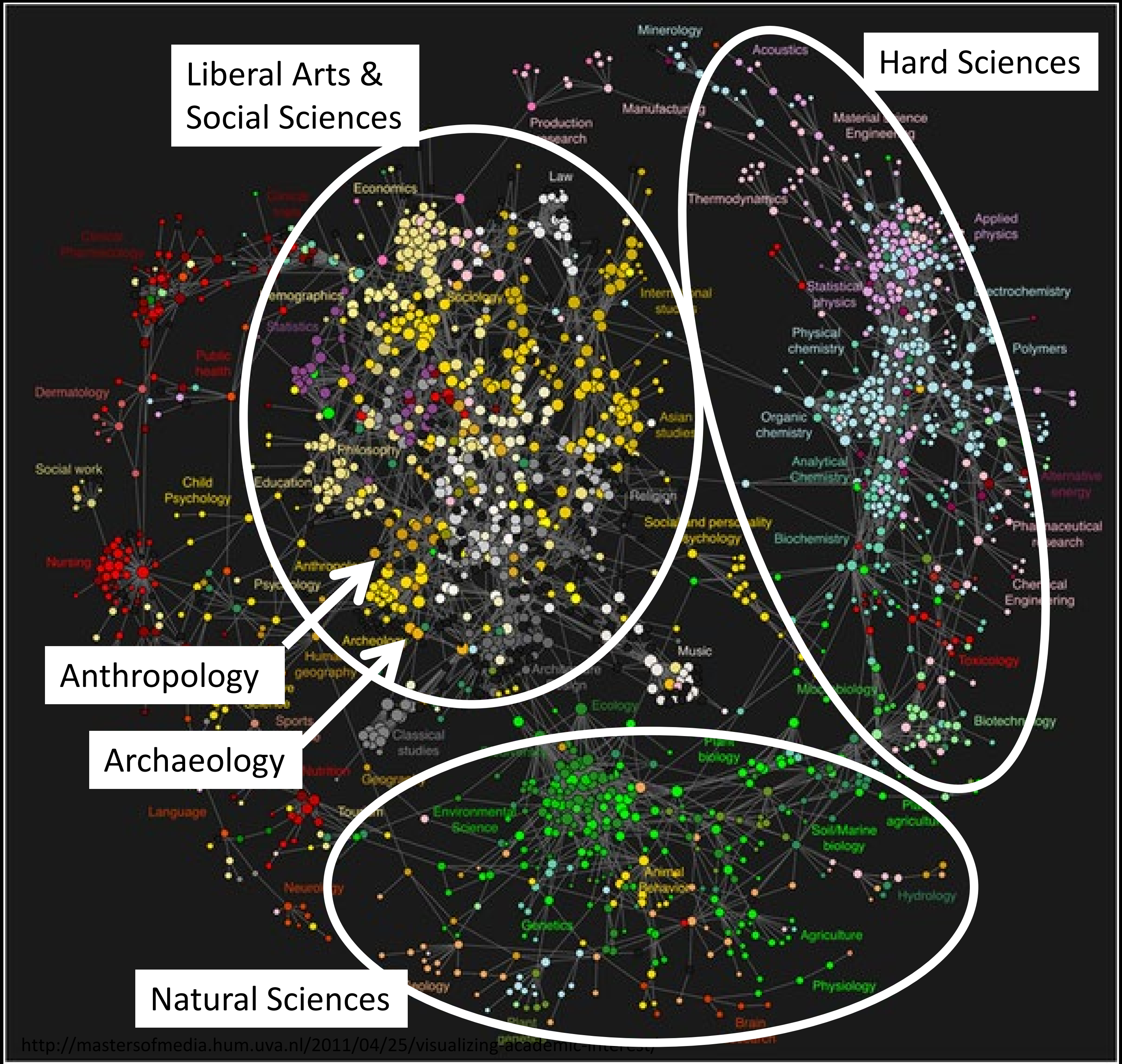
David M. Hoffman, Agustin Gomez-Melendez, Jessy Arends, Sallie Dehler, D. Shane Miller

https://doi.org/10.5751/ES-13529-270416









Liberal Arts & Social Sciences
Hard Sciences
Anthropology
Archaeology
Natural Sciences
Minerology
Acoustics
Production research
Economics
Law
Thermodynamics
Applied physics
Polymers
Physical chemistry
Organic chemistry
Analytical Chemistry
Biochemistry
Toxicology
Dermatology
Social work
Child Psychology
Education
Philosophy
Religion
Nursing
Music
Classical studies
Sports
Language
Ecology
Environmental Science
Soil/Marine biology
Animal Behavior
Genetics
Hydrology
Agriculture
Physiology
http://mastersofmedia.hum.uva.nl/2011/04/25/visualizing-academic-interest/

Sports Science

Dr. JohnEric Smith

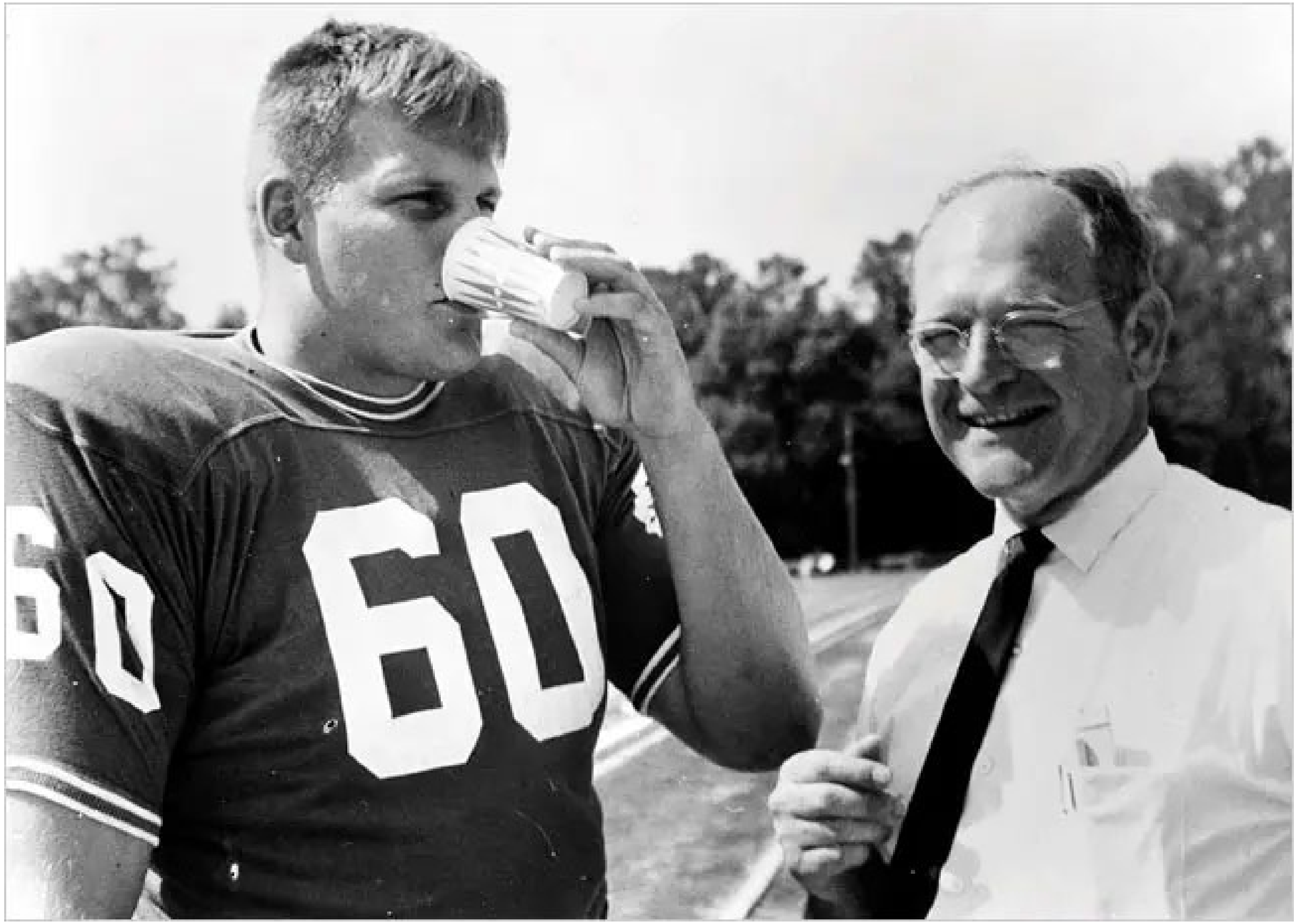
Associate Professor, Kinesiology

Sports Science

MISSISSIPPI STATE UNIVERSITY
DEPARTMENT OF KINESIOLOGY

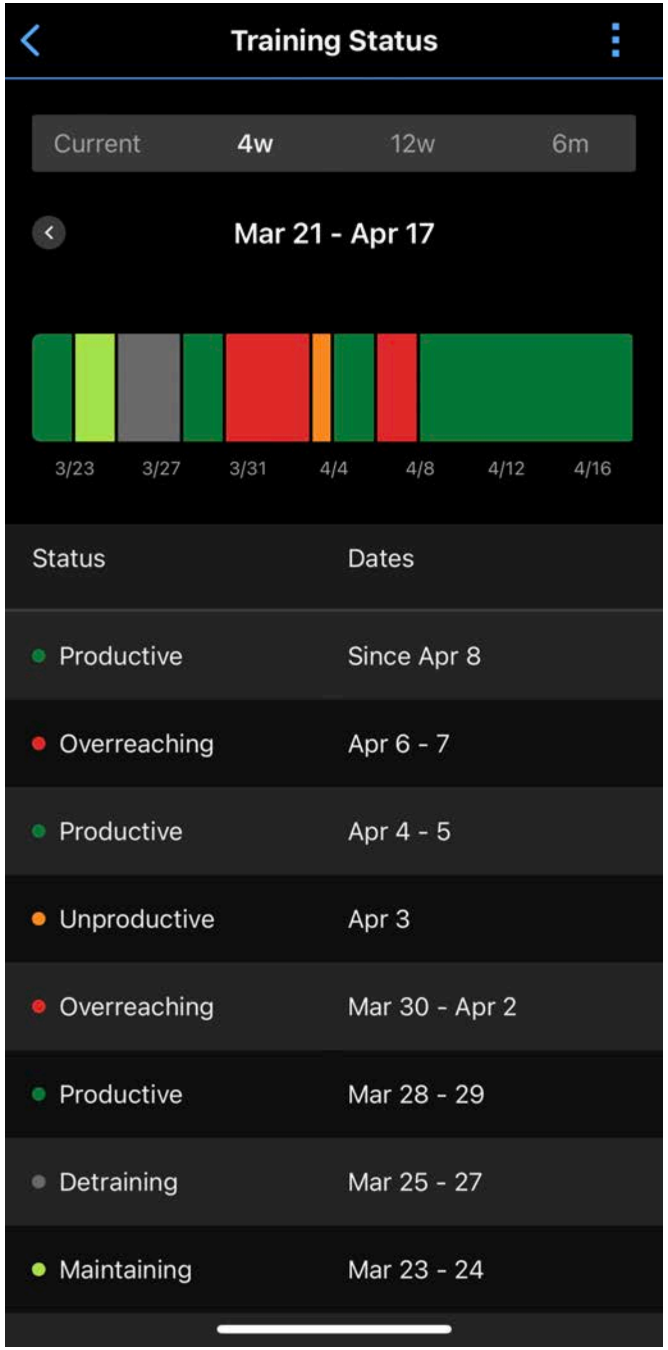
ATHLETE ENGINEERING AT MSU

Stokely
VAN CAMP'S
Finest
GATORADE



Images from: https://1000logos.net/gatorade-logo/

Images From: https://www.nytimes.com/2007/11/28/business/28cade.html


Training Status
Current
4w
12w
6m
Mar 21 - Apr 17
3/23
3/27
3/31
4/4
4/8
4/12
4/16
Status
Dates
Productive
Since Apr 8
Overreaching
Apr 6 - 7
Productive
Apr 4 - 5
Unproductive
Apr 3
Overreaching
Mar 30 - Apr 2
Productive
Mar 28 - 29
Detraining
Mar 25 - 27
Maintaining
Mar 23 - 24


23 Oct
12
58
133
40 km
10
10
Fri 01
GARMIN
11:00AM 5K TRAINING

01:59:40.2

Image from: https://believeintherun.com/shoe-reviews/a-breakdown-of-the-nike-kipchoge-prototype/


Road to Athlete
Performance

Road to Athlete Performance

Statistical Modeling

Dr. Mohammad Sepehrifar
Associate Professor, Mathematics & Statistics

"Data are becoming the new raw material of business." - Craig Mundie

Figure: Craig James Mundie (born July 1, 1949 in Cleveland, Ohio) is Senior Advisor to the CEO at Microsoft and its former Chief Research and Strategy Officer

Title: Data Science and Statistical Modeling: Extracting Insights from Data

Mohammad Sepehrifar
Department of Mathematics and Statistics
msepehrifar@math.msstate.edu

Spring 2023

”Data Science as an evolution of Statistical Modeling, or fundamentally different approaches to analyzing data?”

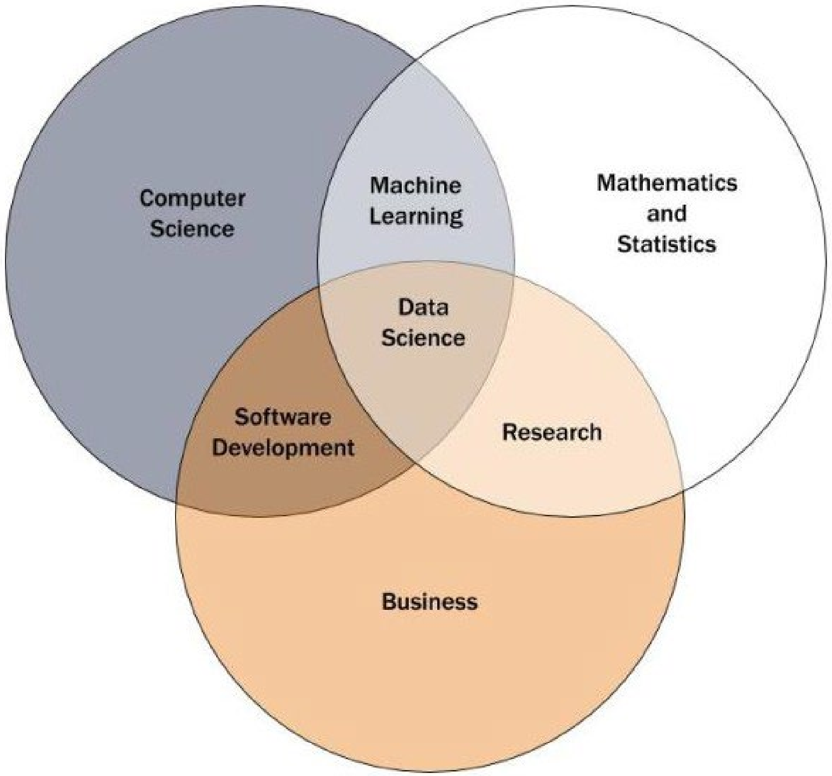


Figure:

1- Data Science as an evolution of Statistical Modeling

- **Statistical modeling** is the process of building mathematical models that represent the relationships between variables in a dataset. These models can then be used to make predictions, identify patterns, or test hypotheses.

- **Data science:**
  - is a more holistic approach that draws from a wide range of disciplines, including statistics, computer science, and domain expertise,
  - extract insights from large and complex datasets by using a variety of tools and techniques, such as machine learning, data mining, and data visualization
  - ability to handle large and unstructured data sets, such as those found in social media, web logs, and sensor data.

3- Are we over-relying on traditional Statistical Modeling techniques in the age of Big Data and Machine Learning, or is there still value in using these established methods?

Statistical Modeling:

- While newer techniques like deep learning and neural networks have gained popularity, traditional statistical modeling methods still have a place in data analysis.
- statistical models can provide a clear and interpretable explanation of the relationship between variables.
- This is particularly important in fields such as medical research, where it is essential to understand the underlying mechanisms behind observed patterns.

- In addition, traditional statistical methods are often better suited to hypothesis testing and theory development. They are designed to answer specific research questions, and they can help researchers test hypotheses and develop theories based on empirical data.
- However, it is also true that traditional statistical methods have some limitations. For example, they may not be able to handle large or complex data sets, and they may not be as accurate as some machine learning techniques in certain contexts. Thus, researchers need to carefully consider the strengths and limitations of different methods and select the most appropriate one for their specific research question and data set.

Data Science:

- Data science involves the entire process of collecting, cleaning, processing, analyzing, and visualizing data to gain insights and make data-driven decisions. It also involves machine learning and artificial intelligence techniques, which go beyond traditional statistical modeling methods.
- In essence, while statistical modeling is a critical component of data science, data science is a more comprehensive field that encompasses a broader range of techniques and tools. Therefore, data scientists need to have a broader skill set than just statistical modeling to be effective in their roles.

Highlights the importance of data science versus statistical modeling

- Data science and statistical modeling are both important fields in the world of data analysis, but they have different strengths and applications.
- Statistical modeling is a branch of statistics that focuses on developing mathematical models to describe and analyze relationships between variables. Statistical models are typically based on a set of assumptions and use a variety of techniques, such as regression analysis and hypothesis testing, to make inferences about data.

- Data science, on the other hand, is a broader field that encompasses a wide range of techniques for extracting insights and knowledge from data, including statistical modeling. Data science typically involves the use of machine learning algorithms, data visualization, and other techniques to uncover patterns and relationships in large, complex datasets.
- While statistical modeling is important for understanding the underlying structure of data and making precise, probabilistic predictions, data science is more focused on the practical application of data analysis techniques to solve real-world problems. Data science techniques can be used to develop predictive models, optimize business processes, and drive decision making in a wide range of industries.

2- Are they fundamentally different approaches to analyzing data?

- Statistical modeling and data science are not fundamentally different approaches to analyzing data, but rather different methodologies within the same overarching discipline of data analysis.
- They both aim to extract insights and knowledge from data, but differ in their approach, tools, and techniques.

Projected growth of the data science industry

- According to the U.S. Bureau of Labor Statistics, employment of computer and information research scientists (which includes data scientists) is projected to grow 19 percent from 2020 to 2030, much faster than the average for all occupations.
- Furthermore, a survey by IBM found that 59% of organizations believe that big data and analytics will be a key source of competitive advantage in their industries in the coming years. This highlights the growing importance of data science in helping businesses make better decisions and gain a competitive edge.

Visualization & Visual Analytics for Built Environment

Dr. Bimal Balakrishnan

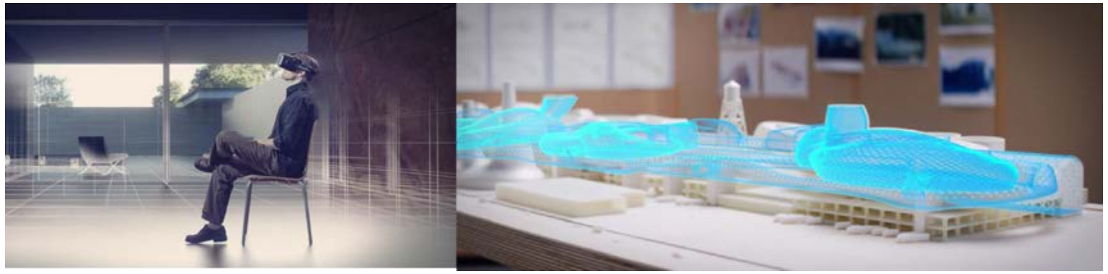
Associate Dean for Research and Professor,
College of Architecture, Art & Design

Visualization and Visual Analytics for the Built Environment: Integrating Design + Data

Bimal Balakrishnan, Ph.D.
Associate Dean for Research
College of Architecture, Art and Design
Mississippi State University

Images generated by Leonardo.Ai

Design and construction industry at the cusp of data-driven disruption…..



DESIGN

Building Information Modeling (BIM)
Virtual Prototyping and Simulations
Emerging Visualization Platforms – VR/AR/MR
Digital Twins

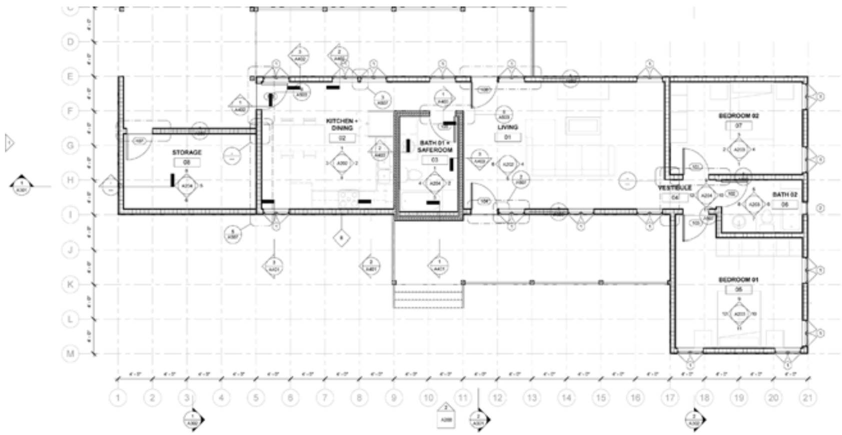
CONSTRUCTION

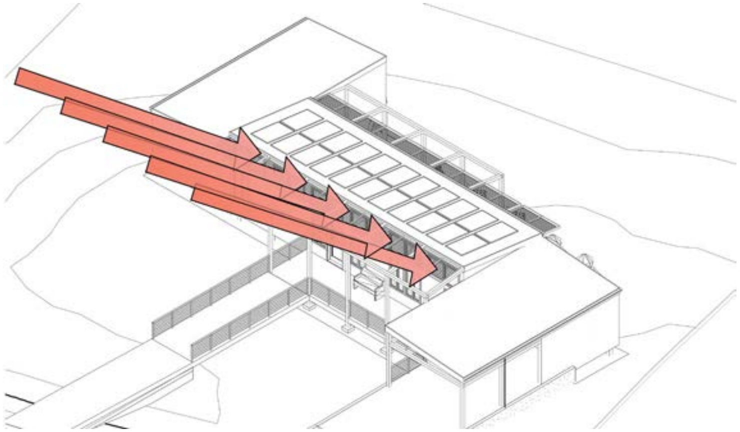
Digital Fabrication
Intelligent Construction Sites

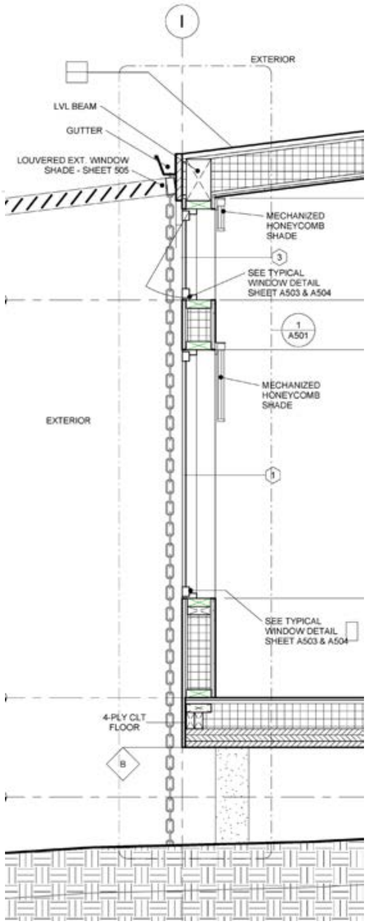
OPERATION

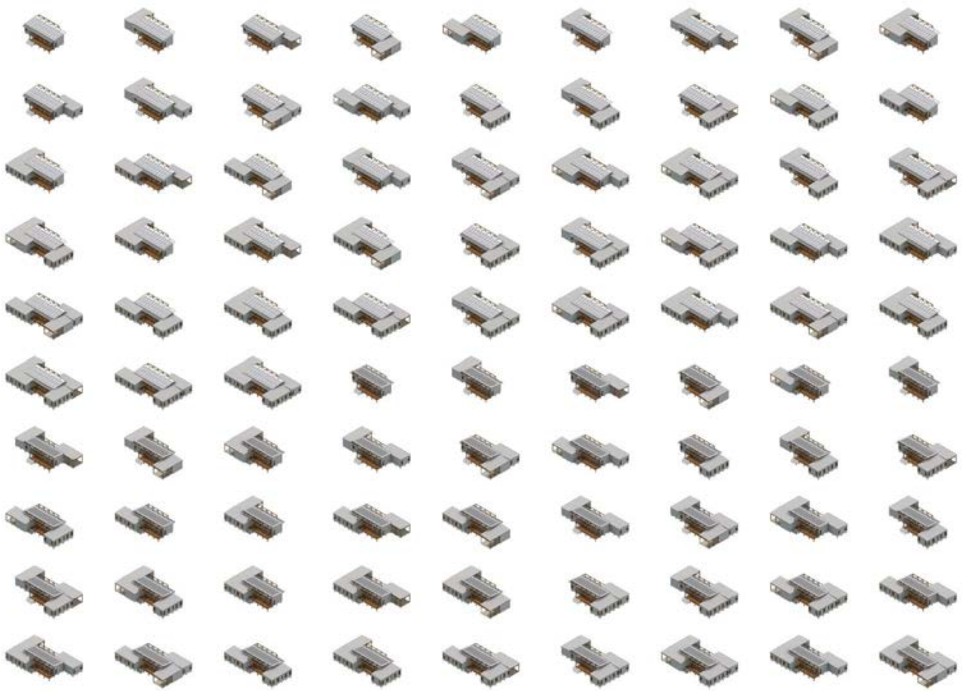
Cameras, LiDAR, IoT sensors etc. are producing rich, data streams that can enhance performance and sustainability

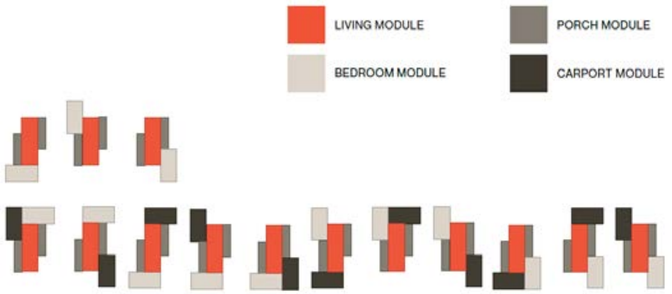
Parametric / Generative Design & Building Information Modeling





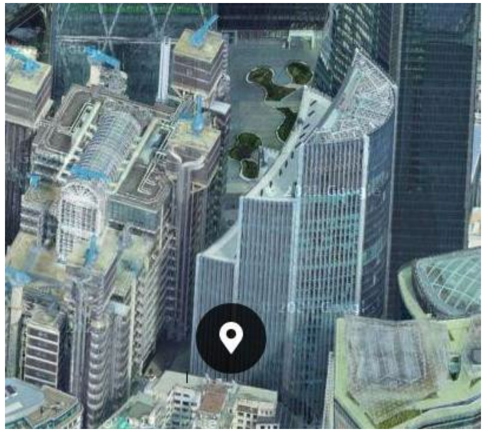


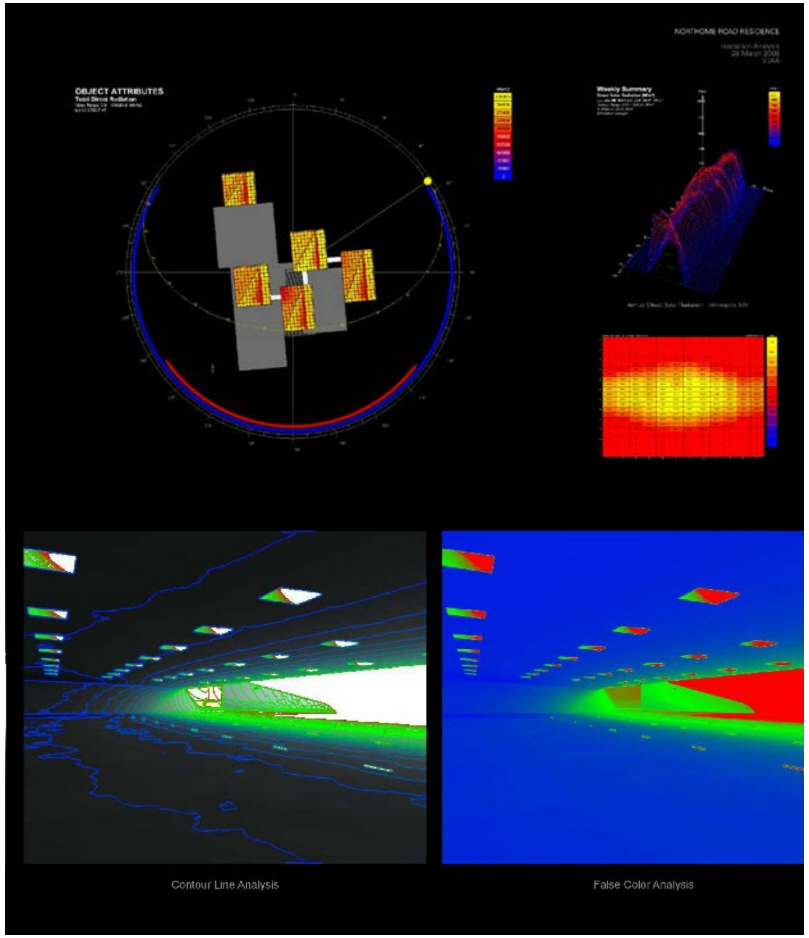


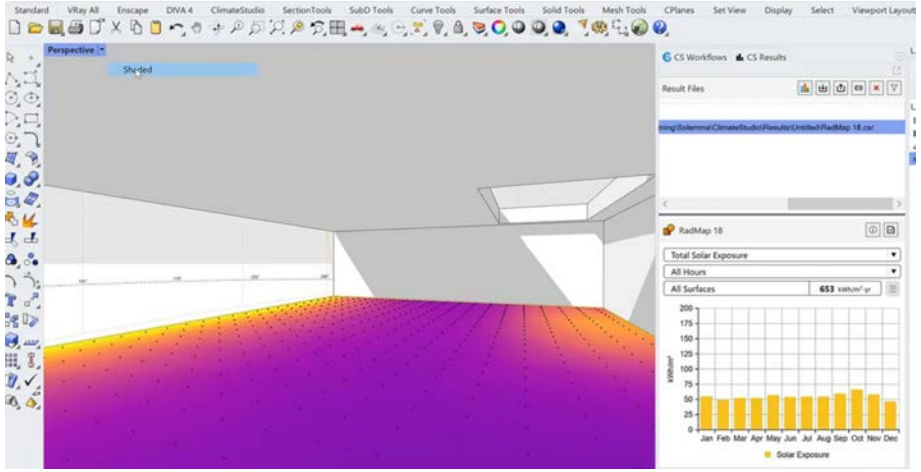


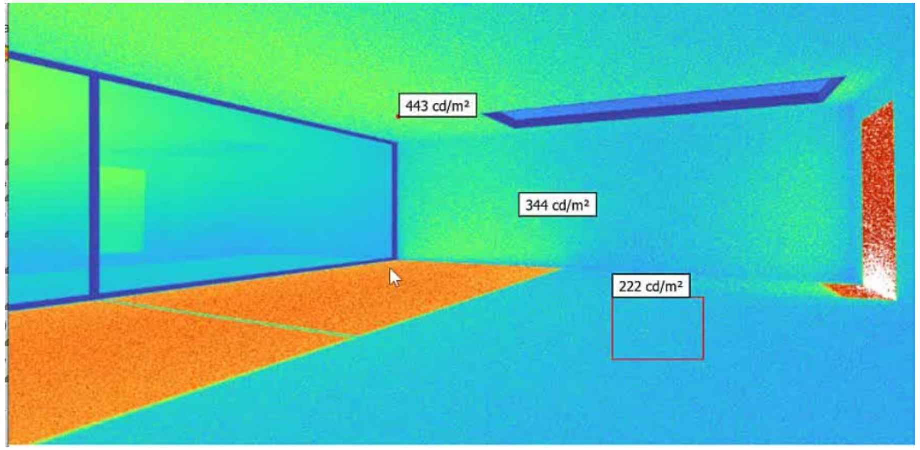
CAAD/ MSU Faculty: **Duane McLemore** & Solar Decathlon Competition Student Team

Building Performance Analysis



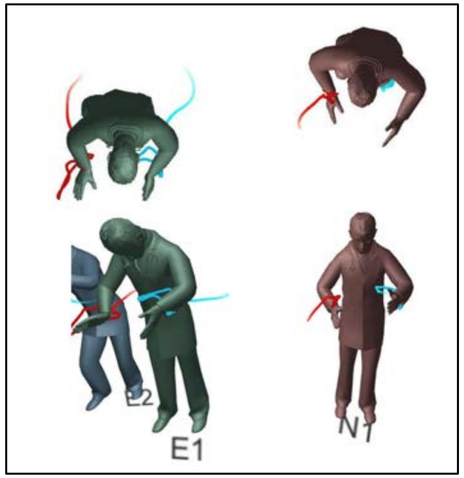


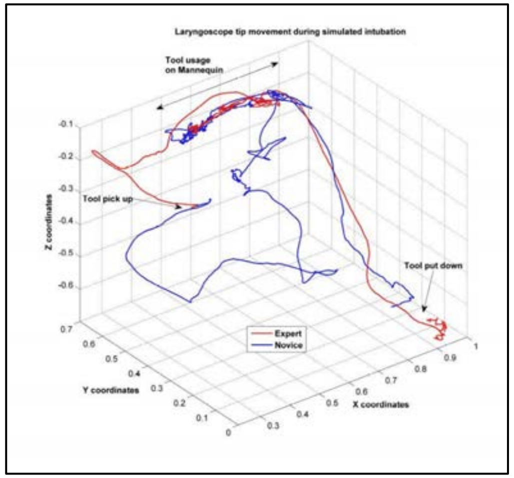




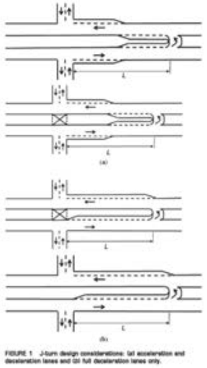
CAAD/ MSU Faculty: **John Ross**
Professional work with: Foster & Partners, Cannon Design

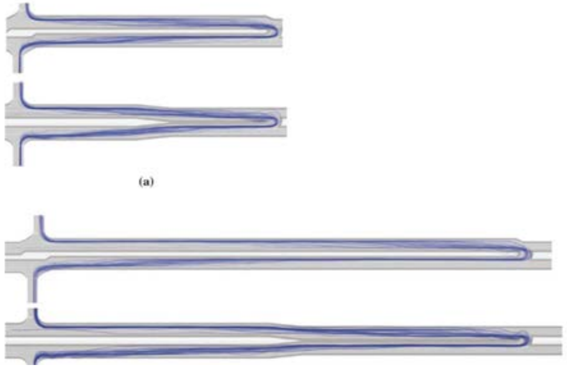
Human Performance & Behavioral Simulations

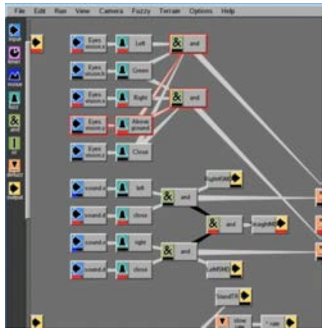


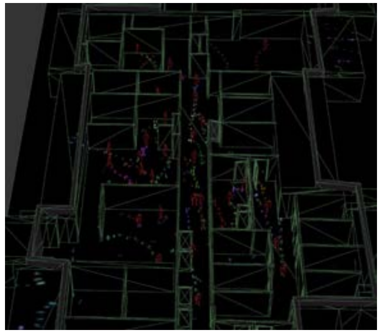


CAAD/ MSU Faculty: **Bimal Balakrishnan**
Collaborators: Julie Marshall (Univ. of Missouri – School of Medicine) Mark Hoffman (Children’s Mercy Research Institute), Greg King (Univ. of Missouri – Kansas City)









CAAD/ MSU Faculty: **Bimal Balakrishnan**
Collaborators: Praveen Edara, Carlos Sun (Univ. of Missouri)

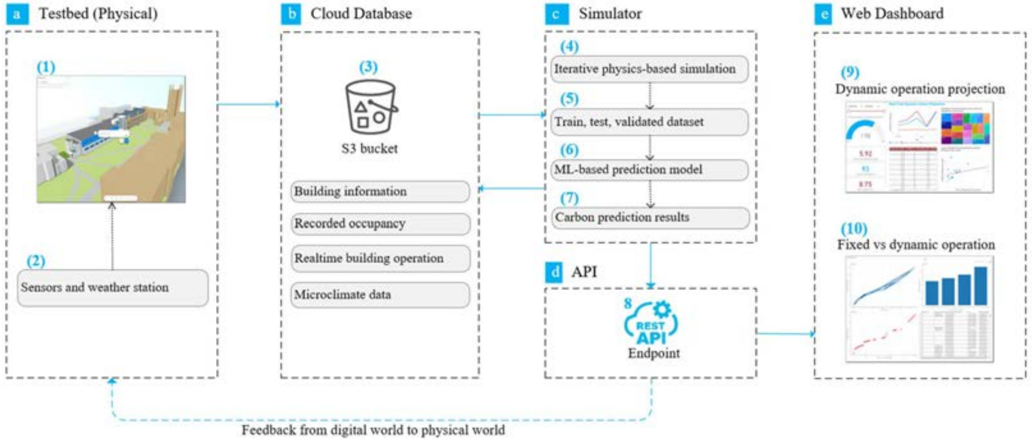
Digital Fabrication

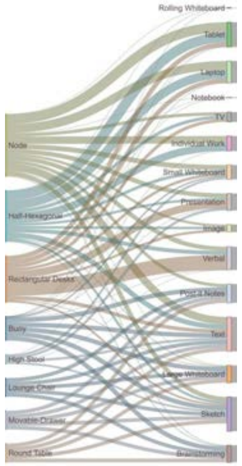
3D Printed Ceramics
CAAD/ MSU Faculty: **Duane McLemore**
(with students)

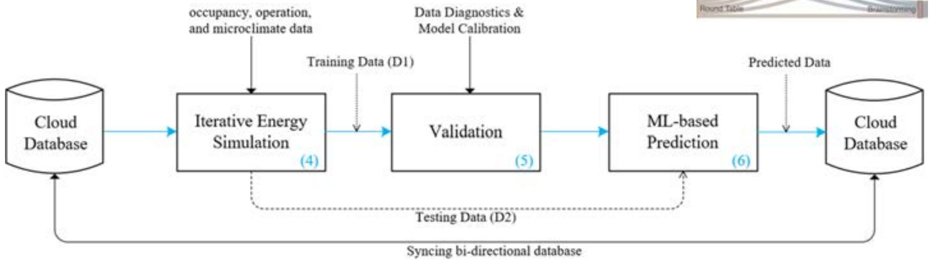
Outdoor Adventures area at the Sanderson Center
CAAD/ MSU Faculty: **John Ross and Silvina Lopez Barrera**
(with students)

Digital Twins and Facility Management

Fig. 3. Sensor equipped lighting (left) and WAC (right)







5G Enabled Digital Twin (Univ. of Missouri)

Partnership with AT&T, Cooper Lighting & Steelcase

Kim, Wang, Khanna, Balakrishnan et al (2023)

CAAD/ MSU Faculty: **Bimal Balakrishnan**

Collaborators: Jong Bum Kim, Fang Wang, Sanjeev Khanna (Univ. of Missouri)

Digital Twins for Infrastructure

Digital Twin of Stonecutters Bridge, Hong Kong

Digital representation

Digital representation of a physical asset, process or system

Continuously surveyed

Continuously synchronized from multiple sources, sensors

Generates insights

Predictability and performance optimization

CAAD/ MSU Faculty: **Afshin Hatami**
Professional work with: Ove Arup & Partners and Bentley Systems Bridge Team

Current Initiatives

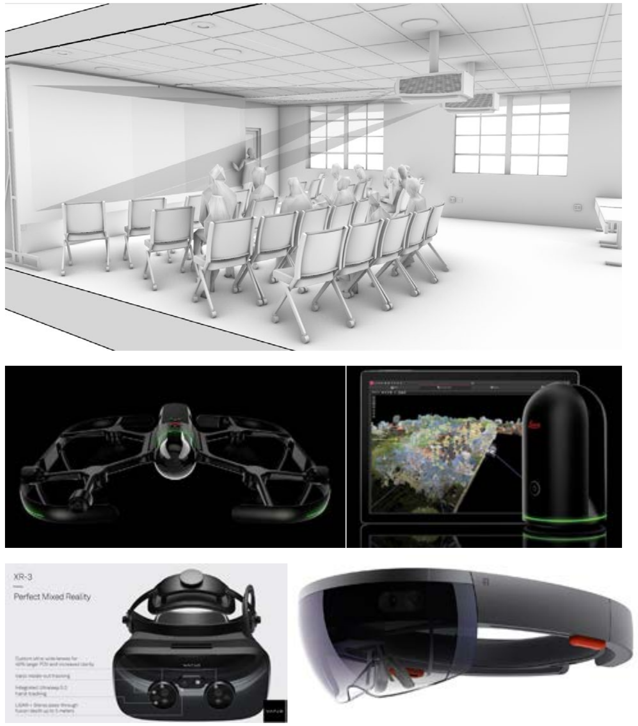
Collaborate with Data Science: Curriculum & Joint Faculty Hire

CAAD Concentration with course work in 3 areas

- **Design, Visualization & Simulation**
- **Design Computing + BIM**
- **Building Performance Simulation**

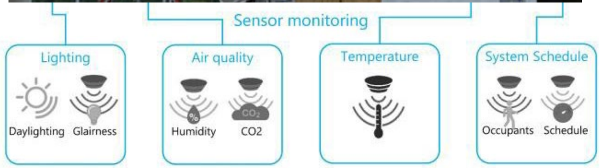
Investment in Infrastructure

Simulation & Interactive Visualization Applications Lab



Research Testbeds for Construction & Operation

Create Living Learning Labs: Modular Housing



CAAD/ MSU Faculty: **Lee Carson, Afshin Hatami, Saeed Rokooei**

Thank You !

bbalakrishnan@caad.msstate.edu