Inspiring Data Science Research through Collaboration with University Operations

Nick Duffield

Director, Texas A&M Institute of Data Science

Royce E. Wisenbaker Professor I, Department of Electrical and Computer Engineering

duffieldng@tamu.edu https://tx.ag/duffield https://tamids.tamu.edu

Thanks for the invitation to speak today!

Thanks for the invitation to speak today!

(Confession: I haven't been so active in Sigmetrics recently. The chairs indicated this was a feature, not a bug)

Universities are Large and Complex Organizations!

Example: Texas A&M University

Academic

- **73,284** students
- **4,922** faculty
- **17** colleges and schools
- **130+** undergraduate programs
- **268** graduate & professional programs

Universities are Large and Complex Organizations!

Example: Texas A&M University

Geographic

- **Campuses:** 2 US + 1 international
- **Campus area:** 22 km^2
- **4,922** faculty
- **17** colleges
- **130+** undergrad programs
- **268** graduate & prof programs

Universities are Large and Complex Organizations!

Example: Texas A&M University

Infrastructure

- **Transportation:**
  - Cars, Parking, Bus Services, Bicycles, Pedestrians
- **Buildings:**
  - Environment, Repair State
- **Utilities:**
  - Water, Heat, AC, Electricity

Universities are Large and Complex Organizations!

Example: Texas A&M University

Research

- **Over $1B** research expenditures
- **Over 150** centers and institutes

This Talk: Operational Data Science

- Using Data Science to enhance university operations and administration
- Benefits for operations, faculty, students
- Three project in Texas A&M Operation Data Science Lab
- Comparing Operational Data Science in universities vs. industry
- comparing universities and in

Operational Data Science Lab

- Universities as a Living Laboratory for Data Science
  - Data Science as an organizing theme between related operations
- Opportunity: improve operations through Data Science:
- Direct Operational Projects
  - Partnering with TAMU operational units to collaboratively solve their problems
- Cognate Operational Projects
  - Collaborations in domains adjacent to and informing operational problems
- Linked by recurrent & reusable set of system components
  - Databases, dashboard interfaces, visualization, statistical modeling, ML / AI

Direct Operational Projects

- RDASH: Research Dashboard for Organizational Intelligence
  - Graph analysis of authorship and citations to identify research strength and opportunities
- Lighthouse: Research Compliance
  - Automated assist for risk categorization of research proposal (animal, human, privacy, …)
- Computer Vision for Game-Day Traffic Measurement
  - Measure traffic traversing road junctions & parking garages; support traffic planning and operations
- Energy Consumption Analysis
  - Automating detection of anomalies form time-series of building energy consumption
- Bikeshare Effectiveness and Operations
  - Using real-time bike locations to characterize and predict demand and optimize operational rebalancing

Cognate Operational Projects

- Data Driven Expert Systems for Irrigation Planning
  - Texas AgriLife Research Corpus Christi & AgriLife Extension
- Creating Public Safety Radio Research Data Set
  - Texas A&M Telecommunications Academy (Magnussen) *[NIST]*
- Rapid Disaster Damage Prediction from Geo-tagged Social Media
  - With Texas A&M College of Architecture *[Microsoft AI for Humanitarian Action]*

Attraction to Faculty of Operational Data Science

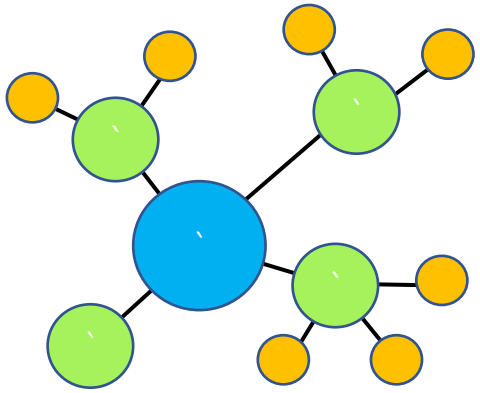
- Operational projects tend to be:
  - Multidisciplinary
  - Integrate vertically: foundations, systems, practice
- Improves positioning for funding solicitations
  - Interdisciplinary initiatives; now also vertical initiatives
  - US: NSF Convergence Research / Accelerator
- Improves positioning for industry outreach
- Career benefits to students from interdisciplinary / integrative work

External vs. In-house vs. Researcher Development

- Operational units typically:
  - Have their own development teams
  - Buy commercial software for many tasks
    - Costly, well-supported, new features slow to be developed
- Researchers
  - Bring state-of-the-art ideas and capabilities
    - Faculty, postdocs, students
  - Can respond quickly to develops prototype
    - Hand off to development team or become feature requests to vendors

Three Projects in Two Areas

- Research Administration
  - RDASH: Organization Intelligence Platform for Institutional Research\
  - Lighthouse: Red Flag AI Tool for Research Compliance
- Infrastructure
  - Computer Vision for Game-Day Traffic Management
- More projects at: https://tamids.tamu.edu/op-data-sci/
  - Energy Consumption Analysis
  - Bikeshare Effectiveness and Operations



RDash: An Organizational Intelligence Platform for Institutional Research

Division of Research | Institute of Data Science
University Libraries | Department of Visualization

Jack Baldauf, Nick Duffield, Costas Georghiades, Bruce Herbert, Dongjoon Lee, Revanth Reddy Male, Sree Kiran Prasad Vadaga, Jian Tao

TEXAS A&M UNIVERSITY Division of Research

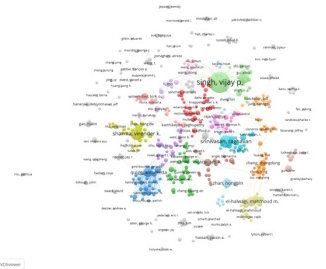
TEXAS A&M Institute of Data Science

LIBRARIES TEXAS A&M UNIVERSITY

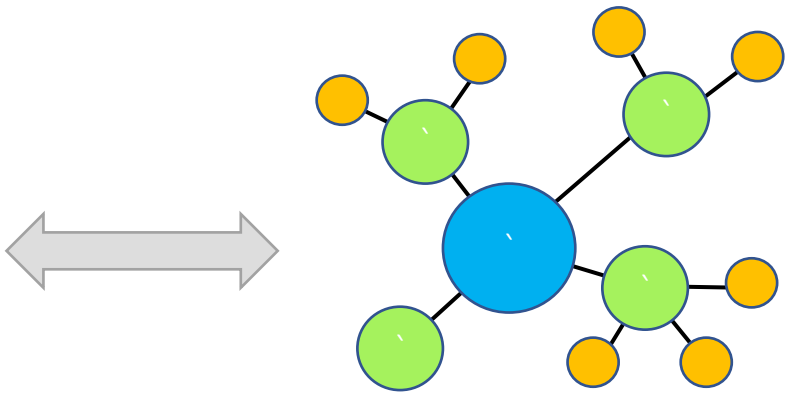
TEXAS A&M UNIVERSITY Visualization

The Networks of Research

- **Question: now can research better address societal grand challenges**
  - Regional, national, and global priorities, funding support
  - Energy, sustainability, climate change, health, poverty, education,…
- **Progress on complex problems will need multidisciplinary teams**
  - How to identify emerging problems and the research teams to match them?
- **Approach: network analysis**



External Researcher Networks



Internal Researcher Network
researchers, topics, collaborations

Research Impacts
Projects, Outcomes
Media, Public Discourse

Research Agendas
Funding Areas, Solicitations
Strategy & Expert Reports

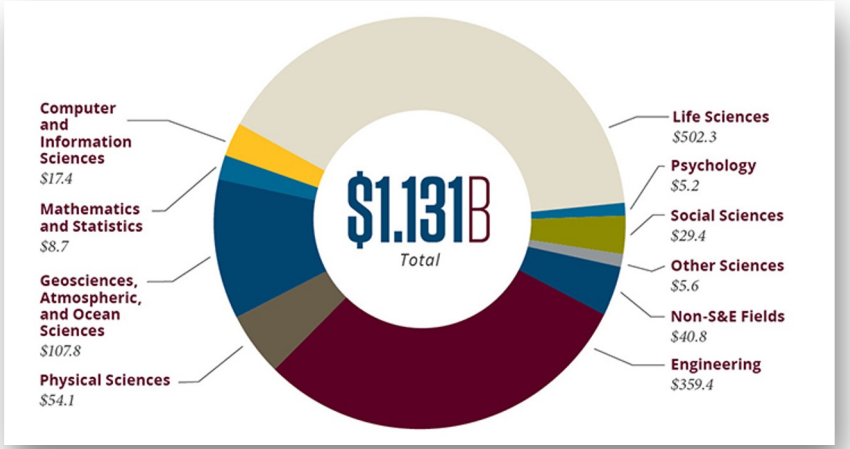
Decision Support for Research Administration

Motivation

- Identify research capacity
  - Areas of faculty expertise and research clusters
- Interdisciplinary research
  - How does it evolve, who drives it?
- Strategic research development
  - Map capacity to funding opportunities and priorities

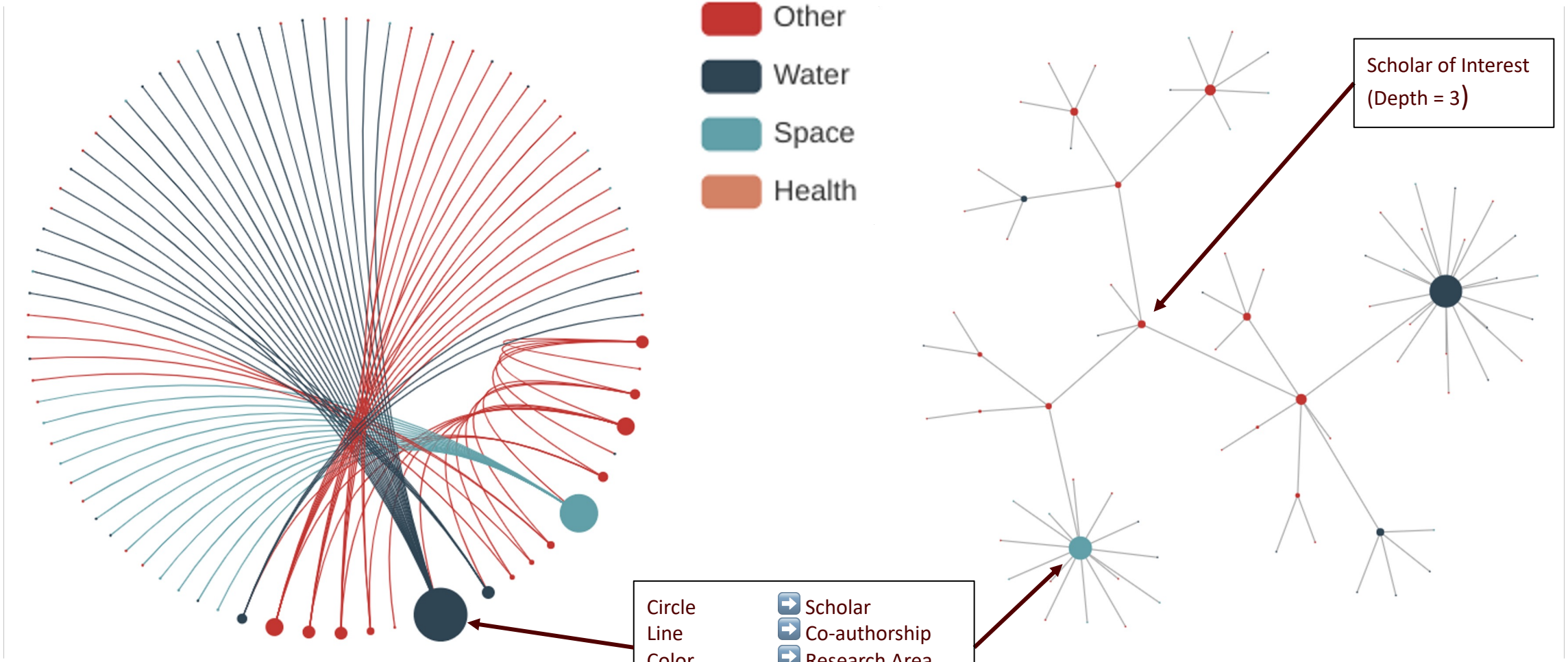
Approach

- Develop profilers
  - Researcher expertise, solicitations, policy, discourse,…
- Topic-specific similarity
  - Match (researchers to each other, and to solicitations)
  - Predict (new collaborations)

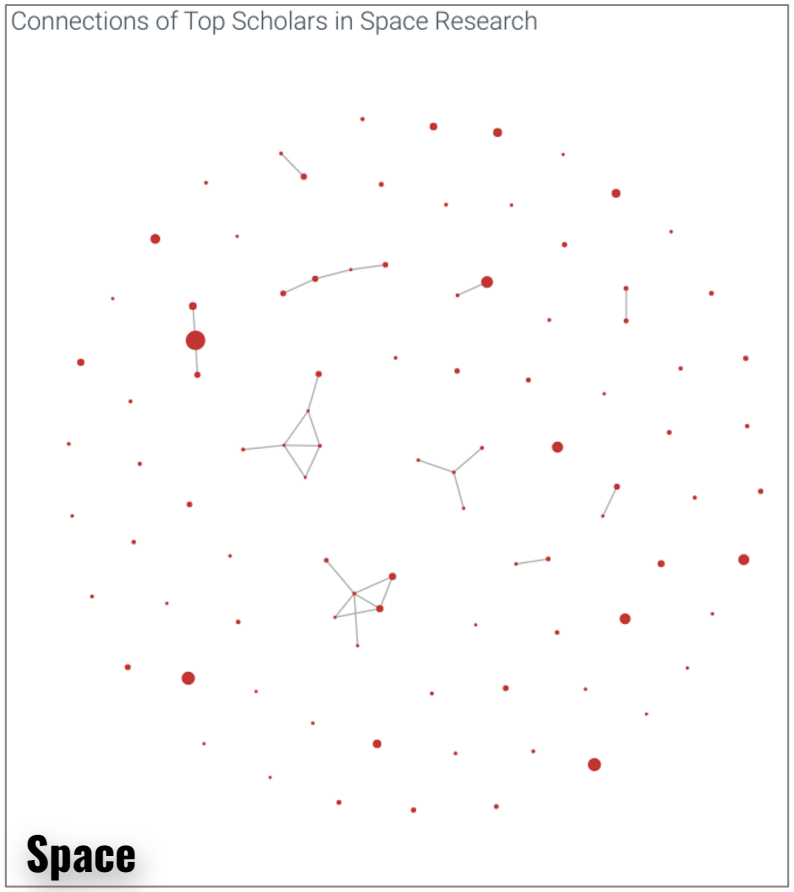


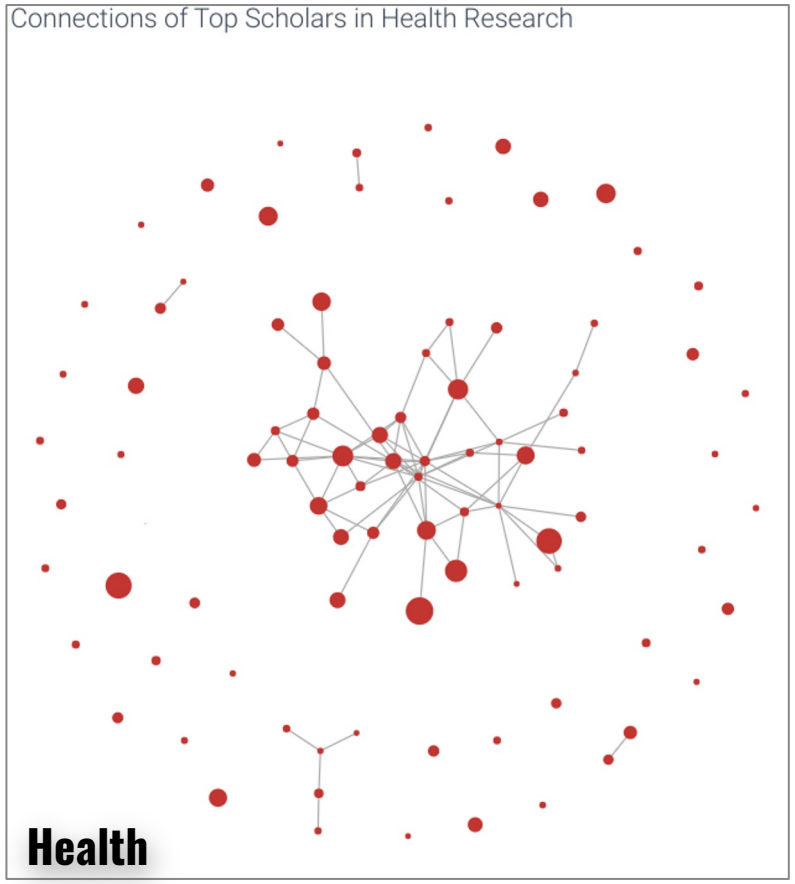
Texas A&M **$1.131B** Research Expenditures (FY2020)

Co-authorship Graph

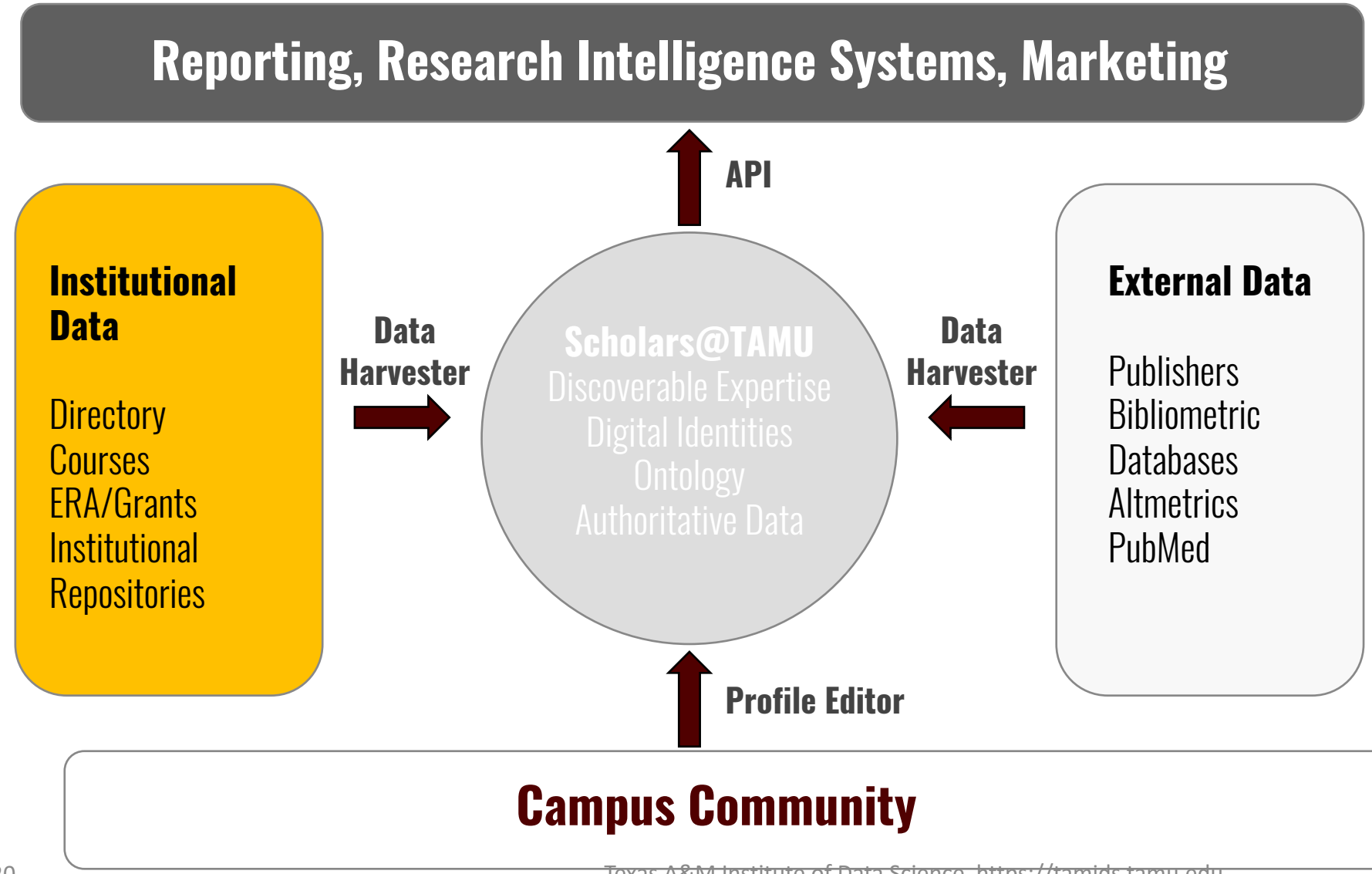


Co-authorship of Scholars in Different Areas (Top 50)





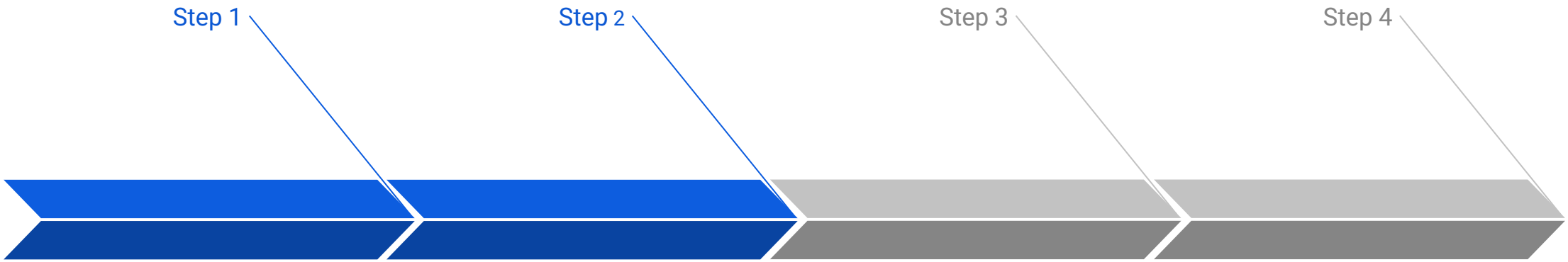
Research Intelligence with Scholars@TAMU



Authoritative Data

Affiliations & Title
Education
Research Blurb
Discipline keywords
Publications
Repository Documents
Grants
Theses & Dissertations
Teaching
Awards
Citations/Altmetrics

Similarity Analysis to Identify Research Areas of a Scholar



Define Keyword Lists for Research Areas

There are three preliminary keyword lists defined: **water, space, and health**.

Water: Water policy Watershed Climate Contamination, saltwater Ice Hydrogeology Hydrobiology Freshwater desalination

Read and Clean up Keywords for Scholars

A sparse document-term matrix is created to describe the frequency of terms that occur in a collection of documents (i.e., scholar profile).

Profile of scholar A: Ice modeling, computational sciences, water management, climate modeling

TF-IDF Analysis

The Term Frequency-Inverse Document Frequency (TF-IDF) analysis is done to identify representative terms for each scholar profile.

TF-IDF entry of scholar A: [0,0,0,0,0,1,0,1,0,1,1,0,0,0,0,0,0, 0,1,...]

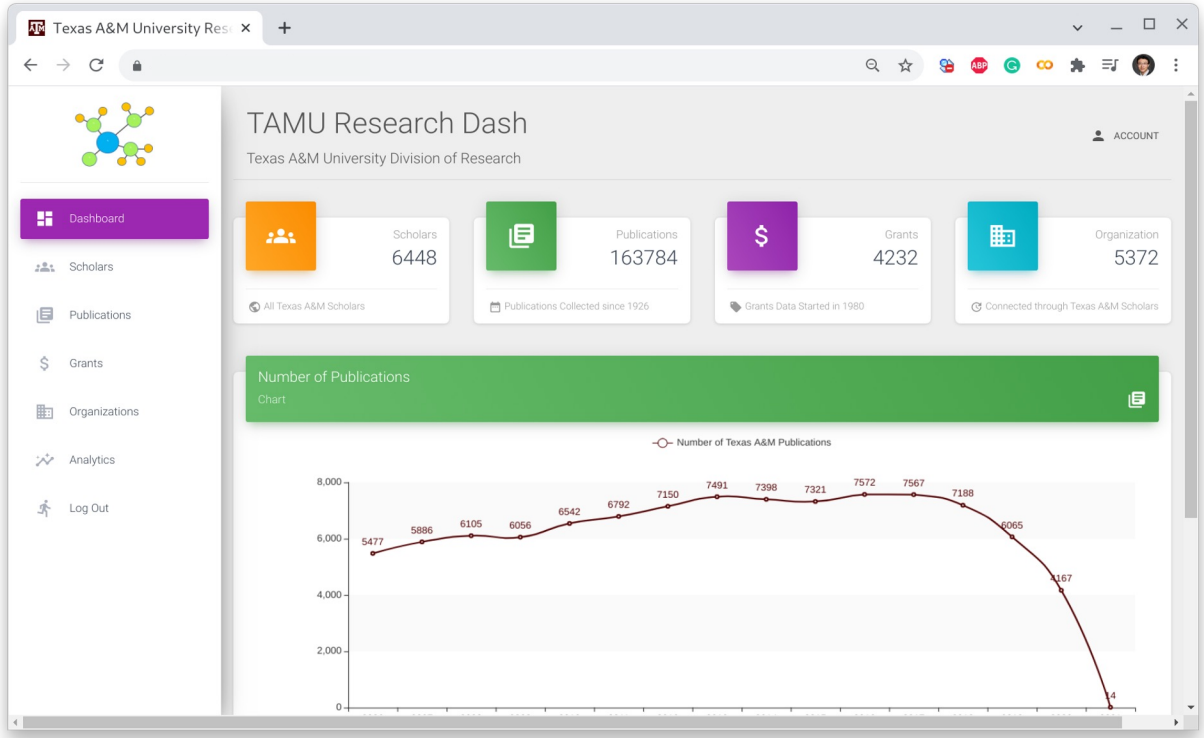
Similarity Analysis

Carry out similarity analysis between the predefined keyword lists and scholar profile to identify research areas for each scholar.

Similarity index of scholar A on water-related researcher: 0.01

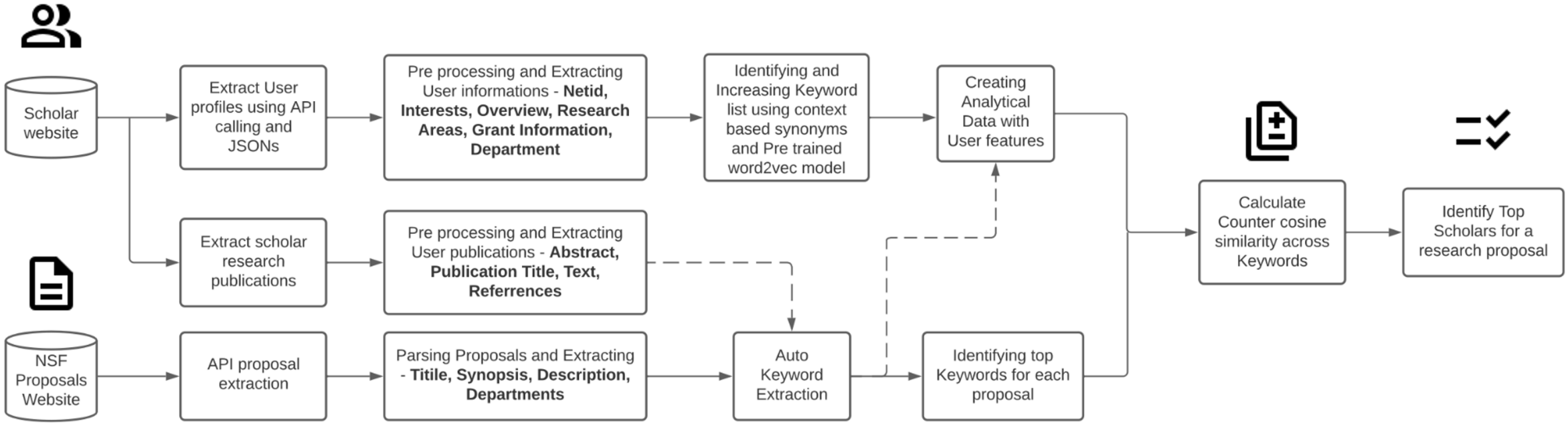
Design & Implementation

- **Dashboard interface**
  - *Python-based open-source web programming framework integrated to analysis backend*
- **Analysis backend**
  - *Graph databases built on relational DB*
  - *Precomputed reports and support user queries for drill-down and exploration*
- **System Design**
  - *Model-View-Controller design facilitates future development & **related reuse***



TEXAS A&M Institute of Data Science

Research Proposal Recommendation Process Flow



Lighthouse Project: Red Flag AI Tool

Division of Research – Research Compliance
Institute of Data Science | Department of Visualization

Nick Duffield, Tiffany Inbody, Revanth Reddy Male,
Katherine Rojo del Busto, Sree Kiran Prasad Vadaga, Jian Tao

TEXAS A&M UNIVERSITY
Division of Research

TEXAS A&M
Institute of
Data Science

TEXAS A&M UNIVERSITY
Visualization

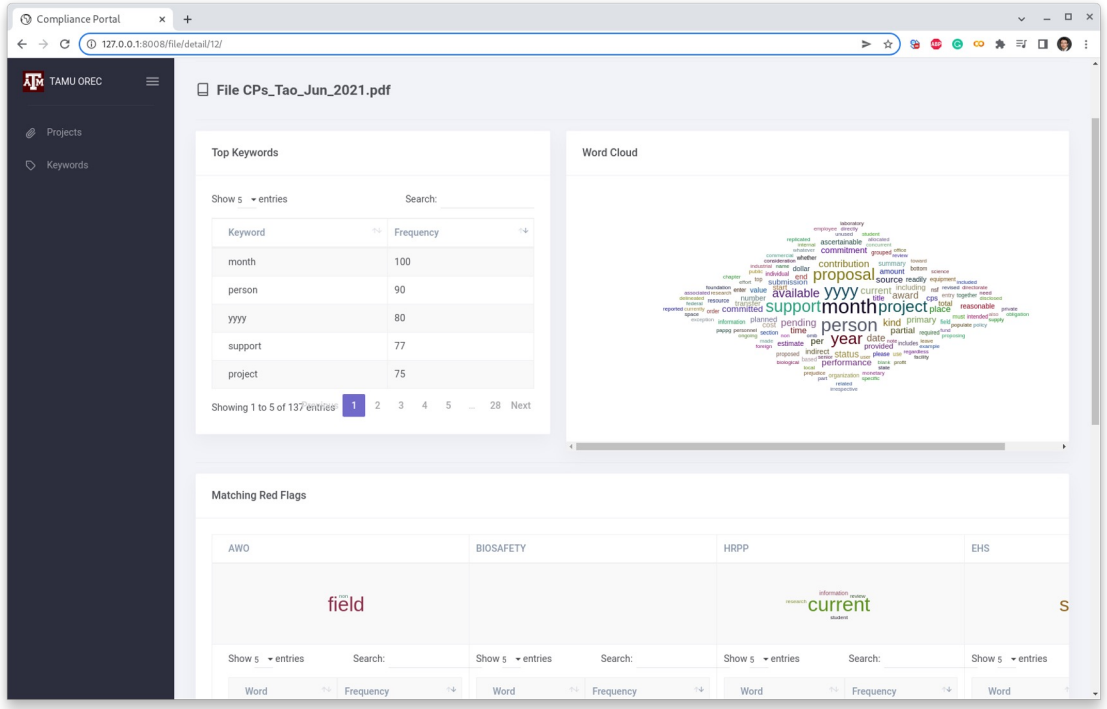
Setting: Research Regulatory Compliance

- Regulations apply to research in some areas
  - Biosafety, animal subjects, human subject, …
- Universities have responsibility to identify
  - Label “red-flag” areas at the proposal stage
  - Provide information and guidance to proposers
- Current model
  - Proposers indicate categories on submission
  - Staff experts review and refine categories

Image courtesy of the Texas A&M Division of Research

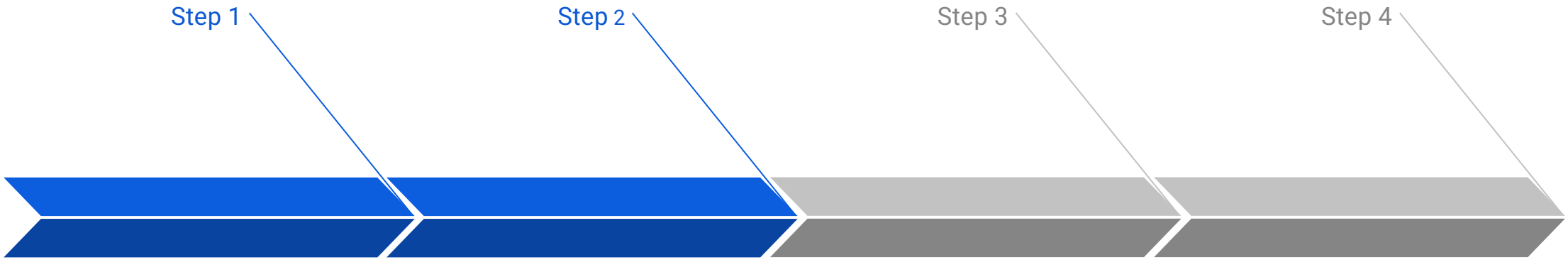
Automation with Natural Language Processing

- Enhance proposal review with NLP
  - Generate flag categories from document
  - Indicate text associated with category
- Implement previous framework
  - Web-based interactive front end
  - Document analysis backend



TEXAS A&M Institute of Data Science

Implementation - Backend with Natural Language Processing



Define Red Flag Lists

There are 6 red flag lists: AWO, Biosafety, HRPP, EHS, Export Control, and Privacy.

AWO: Animal Vertebrate Mice Mouse Rat Rodent Cattle Bovine Dog Canine Cat Feline Rabbit Guinea Pig Hamster Pig ...

Read in PDFs

Read in and transform PDF files into plain text files for further analysis.

Proposal A: Ice modeling, computational sciences, water management, climate modeling, mammal, surviving ...

TF-IDF Analysis

The Term Frequency Inverse Document Frequency (TF-IDF) analysis is done to identify representative terms for each proposal.

TF-IDF entry of proposal A: [0,0,0,0,0,1,0,1,0,1,1,0,0,0,0,0,0,0,1,...]

Similarity Analysis

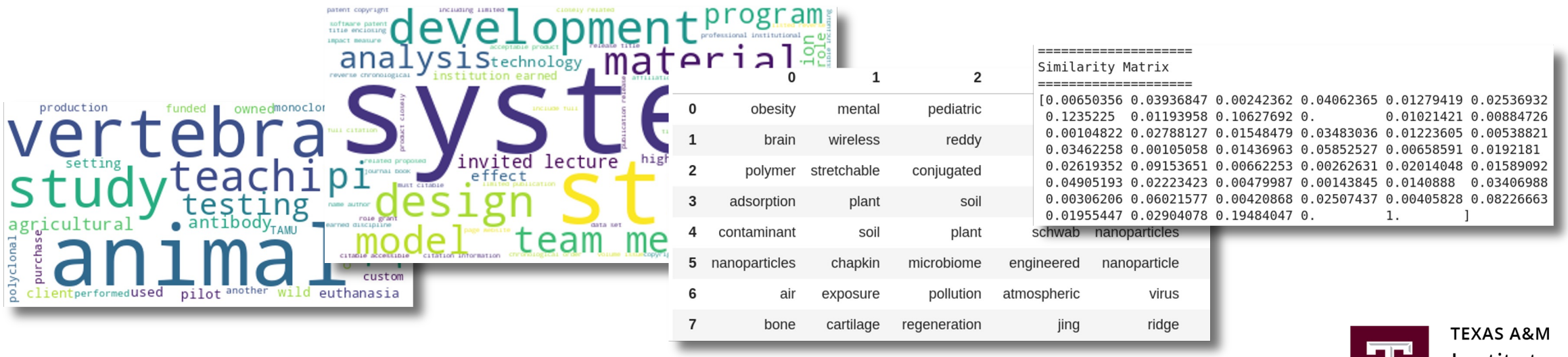
Carry out similarity analysis between the reg flags and parsed proposals to identify compliance issues.

Similarity index of proposal A on AWO: 0.01

Implementation - Backend with Natural Language Processing

Step 1 → Step 2 → Step 3 → Step 4

Define Red Flag Lists | **Read in PDFs** | **TF-IDF Analysis** | **Similarity Analysis**



	0	1	2		
0	obesity	mental	pediatric		
1	brain	wireless	reddy		
2	polymer	stretchable	conjugated		
3	adsorption	plant	soil		
4	contaminant	soil	plant	schwab	nanoparticles
5	nanoparticles	chapkin	microbiome	engineered	nanoparticle
6	air	exposure	pollution	atmospheric	virus
7	bone	cartilage	regeneration	jing	ridge

```
====================
Similarity Matrix
====================
[0.00650356 0.03936847 0.00242362 0.04062365 0.01279419 0.02536932
 0.1235225  0.01193958 0.10627692 0.         0.01021421 0.00884726
 0.00104822 0.02788127 0.01548479 0.03483036 0.01223605 0.00538821
 0.03462258 0.00105058 0.01436963 0.05852527 0.00658591 0.0192181
 0.02619352 0.09153651 0.00662253 0.00262631 0.02014048 0.01589092
 0.04905193 0.02223423 0.00479987 0.00143845 0.0140888  0.03406988
 0.00306206 0.06021577 0.00420868 0.02507437 0.00405828 0.08226663
 0.01955447 0.02904078 0.19484047 0.         1.         ]
```


TEXAS A&M Institute of Data Science

Lessons, limitations, and work in progress

- Keywords
  - We found no comprehensive set of **detailed** key words an phrases
  - Can have depth or breadth, but not both
- Topic modeling
  - Currently supervised by domain experts on small set of field; doesn't scale
  - Topic modelling for emerging interdisciplinary fields challenging
  - Plan to augment with semi-supervised approach
- Interface and Visualization
  - Opportunity to incorporate advanced design for information presentation

Game-Day Traffic Measurement

Transportation Institute | Institute of Data Science
Department of Construction Science

Amir Behzadan, Nick Duffield, Tim Lomax, Yalong Pi

Texas A&M Transportation Institute

TEXAS A&M Institute of Data Science

TEXAS A&M UNIVERSITY Construction Science

Game-day Traffic

- College Football Games at Texas A&M
  - Kyle Field: 6th largest capacity worldwide
  - Largest traffic events of the year
  - 120,000+ attendees & workers
- Traffic Management Aims
  - #1 Help people travel away in safe and timely manner after the game!
- More Detailed Version
  - Soothe local pinch points: eliminate localized long delays
  - Coordinate flows of pedestrians and cars to minimize interaction
  - Provide timely information on traffic and pedestrian flows to local police
- Approach
  - Estimate latent vehicle demand for planning
  - Vehicle counts in roadways, junction ingress-egress, parking garage egress

City of College Station, https://blog.cstx.gov/tag/game-day-traffic/

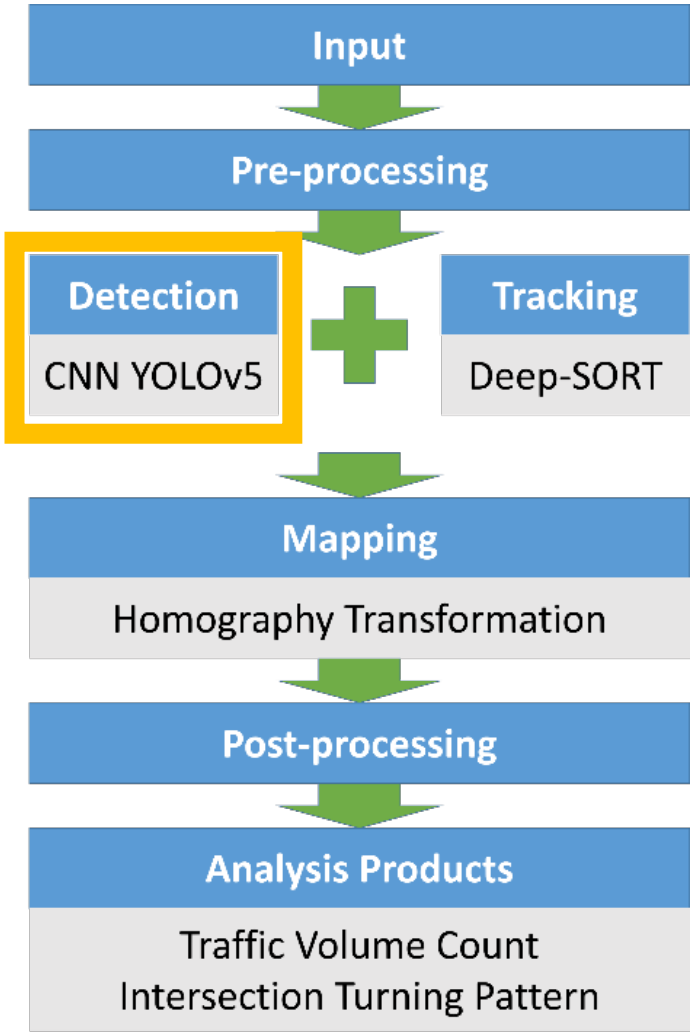
Computer Vision in Traffic Measurement

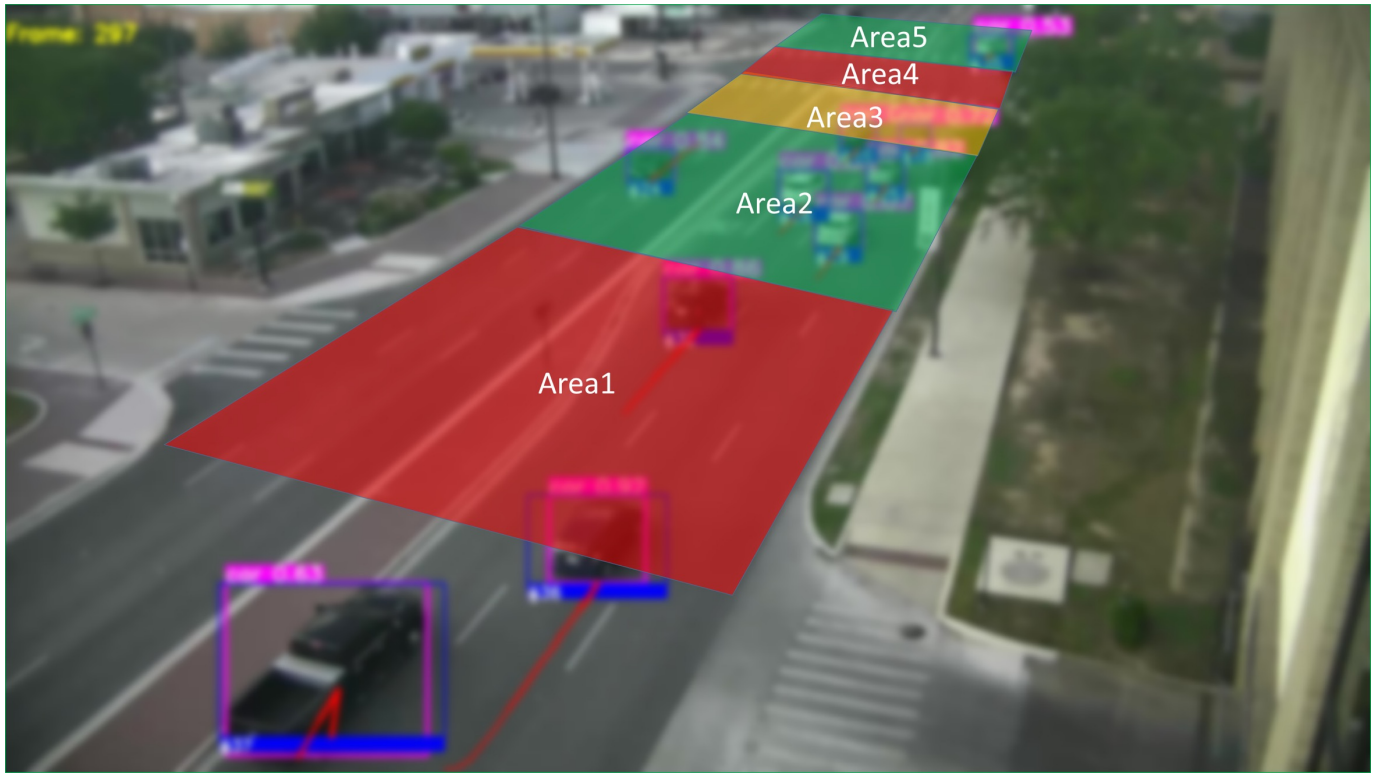
- Computer Vision based vehicle counting
  - Data collection capability already available
  - Texas A&M Transportation Service video monitoring

- Less resource demanding that traditional methods
  - Human counters
  - Pressure hoses
  - Electric sensors

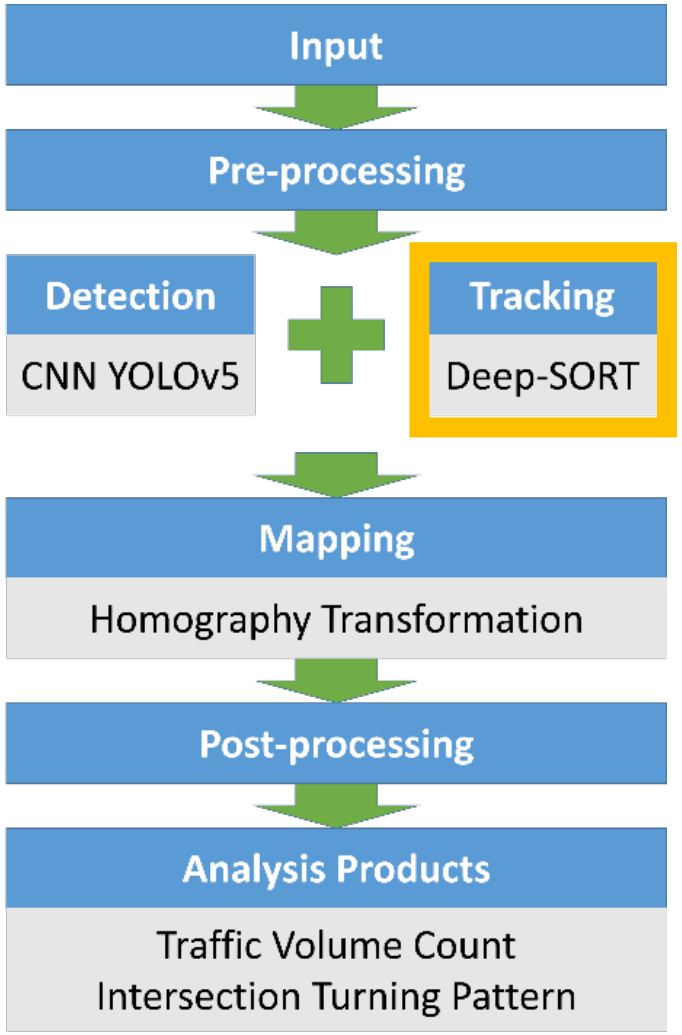
- GPS / Smartphone location reporting service
  - Insufficient detail on counts, insufficiently low latency

WorkFlow: Detection

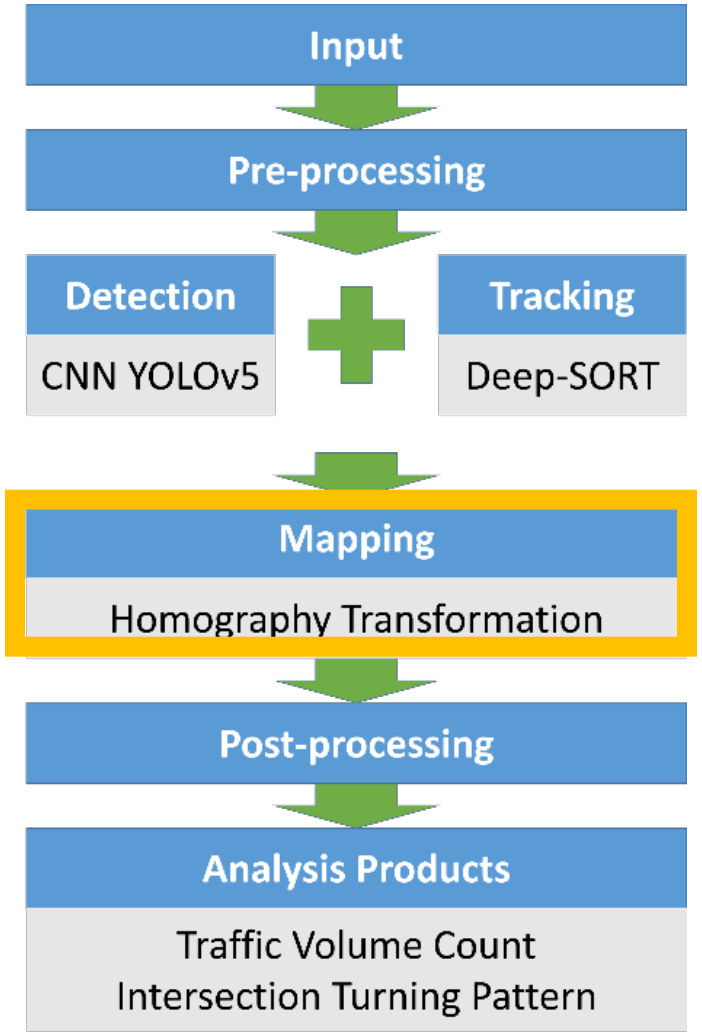


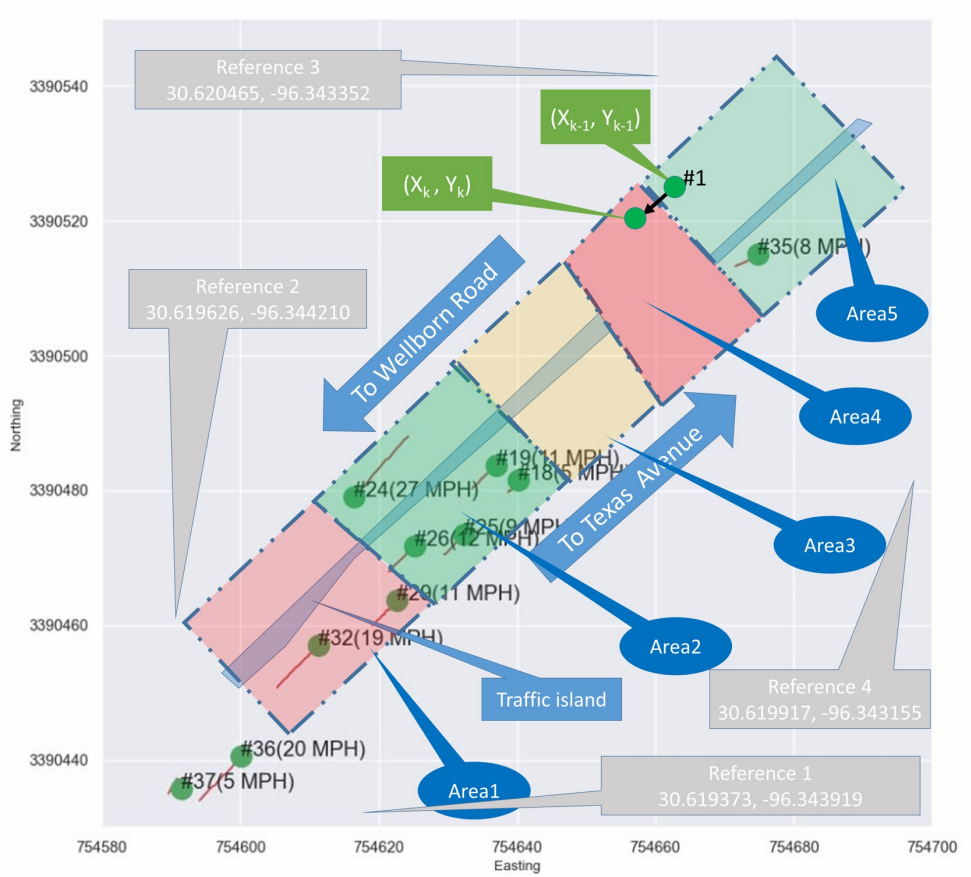


WorkFlow: Tracking



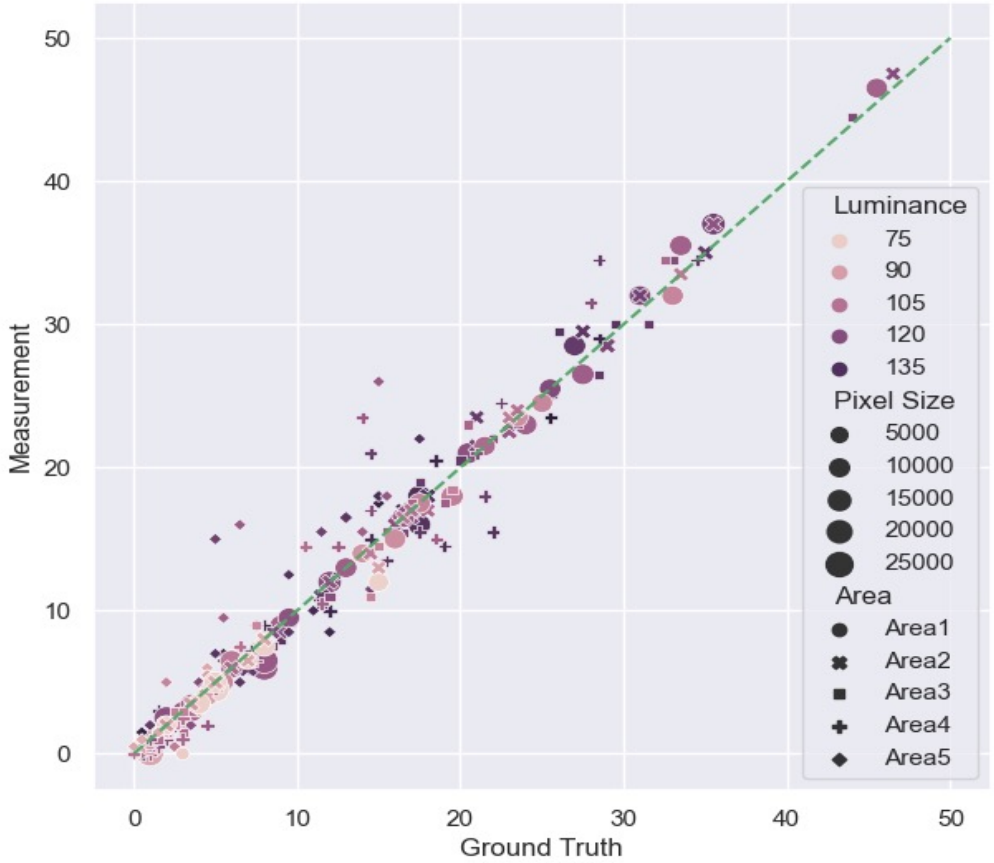
WorkFlow: Mapping





TEXAS A&M
Institute of
Data Science

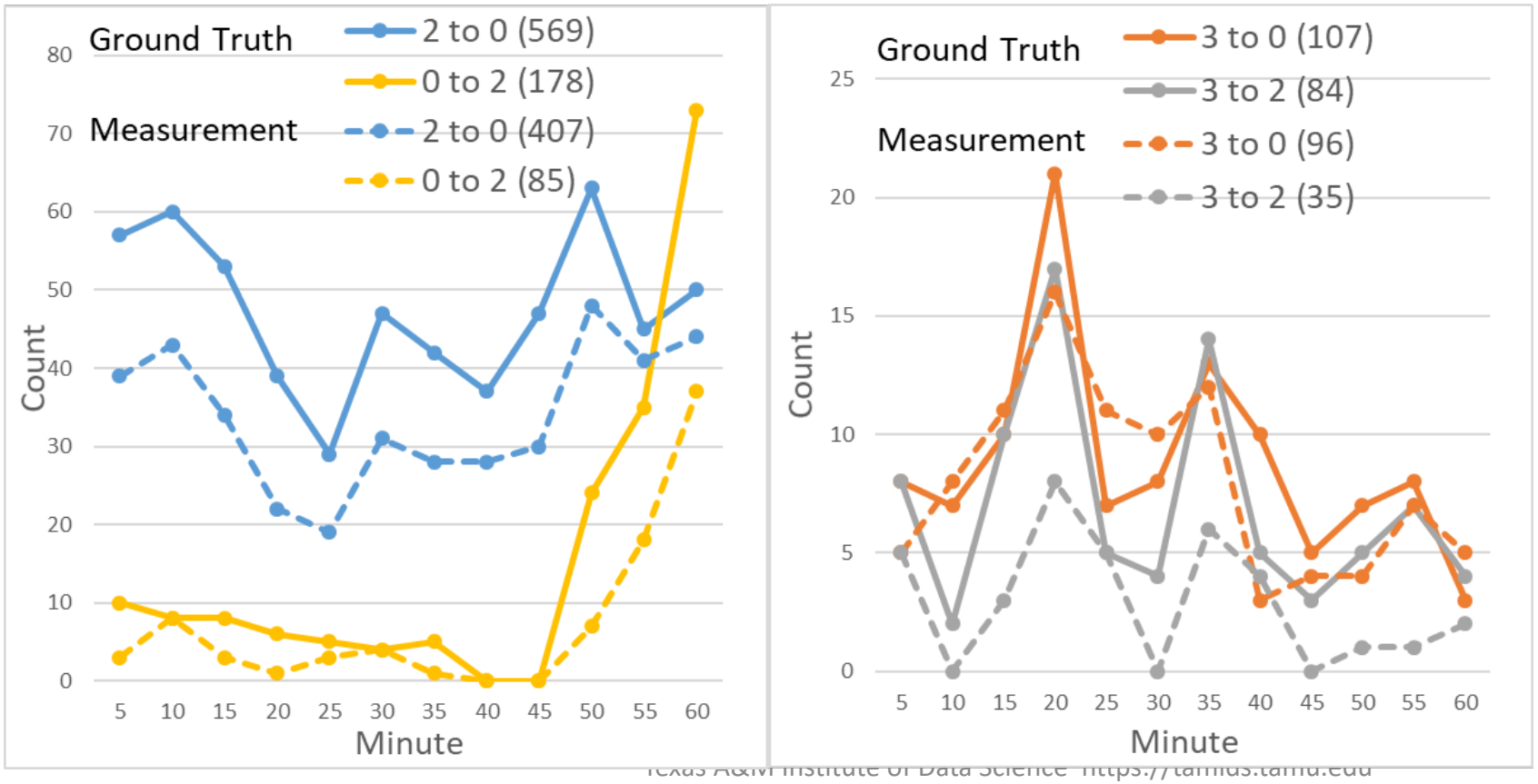
Traffic Count Precision

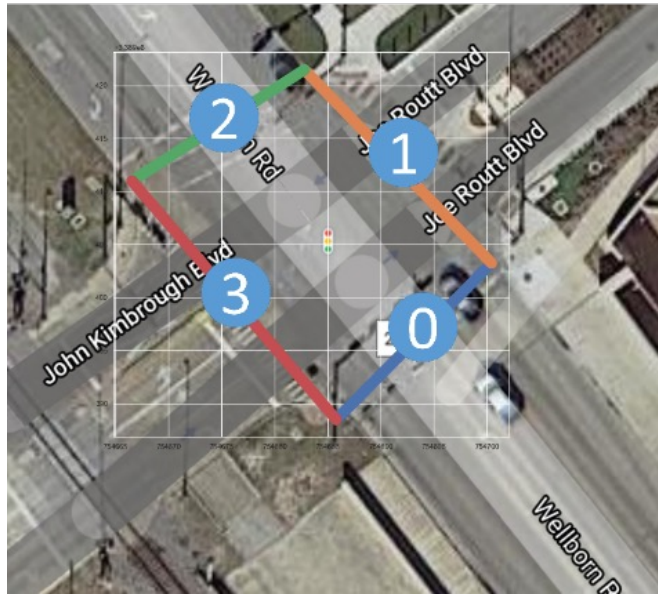


- 60 x 1-minute sample, manually annotated
- Multiple Object Tracking Accuracy 61%
- Better accuracy (unsurprisingly)
  - Closer, larger, objects are smaller angle to vertical

Turning Map Application

- Count vehicle ingress-egress traversals of junction
- 12 x 5 mnute samples during 1 hour in a game day





TEXAS A&M Institute of Data Science

Lesson learned and work in progress

- Need to isolate inconsistencies in imaging
  - Manual camera panning by operators
  - Light conditions, weather, road surface variation, impact accuracy
  - Detect / flag / omit, or potentially adapt
- Egress counts from videos at parking garages:
  - No counts from exit gate: these are kept up to speed exits
- Pedestrian counts on routes from stadium to garage
  - Learn / predict detailed car egress volumes from varying pedestrian counts
- Implement on commodity computing platform (PC + lightweight YOLO)
  - Want additional per-camera cost to be nearly negligible

Follow-on Computer Vision Projects

- Construction Risk Monitoring
  - CV-based speed tracking of vehicle speeds in construction zones
- Emergency Management
  - Rapid Disaster Damage Prediction from Geo-tagged Social Media
- Animal Science
  - CV-based tracking of individual livestock food consumption

Operational Data Science @ Texas A&M aims:

- What:
  - Improve student experience and outcomes at the university
  - Provide decision support to academic leaders and researchers
  - Increase efficiency and responsiveness of infrastructure
- How:
  - Engage with practitioners to understand their problems
  - Use access and insights to abstract longer-term research agenda
  - Work across disciplines to find solutions that integrate science and systems

Operational Data Science @ Texas A&M aims:

- What:
  - Improve **student** experience and outcomes at the **university**
  - Provide decision support to **academic leaders** and **researchers**
  - Increase efficiency and responsiveness of infrastructure
- How:
  - Engage with practitioners to understand their problems
  - Use access and insights to abstract longer-term research agenda
  - Work across disciplines to find solutions that integrate science and systems

(Prior to Texas A&M, I worked in AT&T Labs-Research)

Operational Data Science @ Texas A&M aims:

- What:
  - Improve **student** experience and outcomes at the **university**
  - Provide decision support to **academic leaders** and **researchers**
  - Increase efficiency and responsiveness of infrastructure
- How:
  - Engage with practitioners to understand their problems
  - Use access and insights to abstract longer-term research agenda
  - Work across disciplines to find solutions that integrate science and systems

Operational Data Science @ AT&T aims:

- What:
  - Improve **customer** experience and outcomes in the **network**
  - Provide decision support to **business leaders** and **engineers**
  - Increase efficiency and responsiveness of infrastructure
- How:
  - Engage with practitioners to understand their problems
  - Use access and insights to abstract longer-term research agenda
  - Work across disciplines to find solutions that integrate science and systems
- Difference between two contexts?
  - Just the people involved

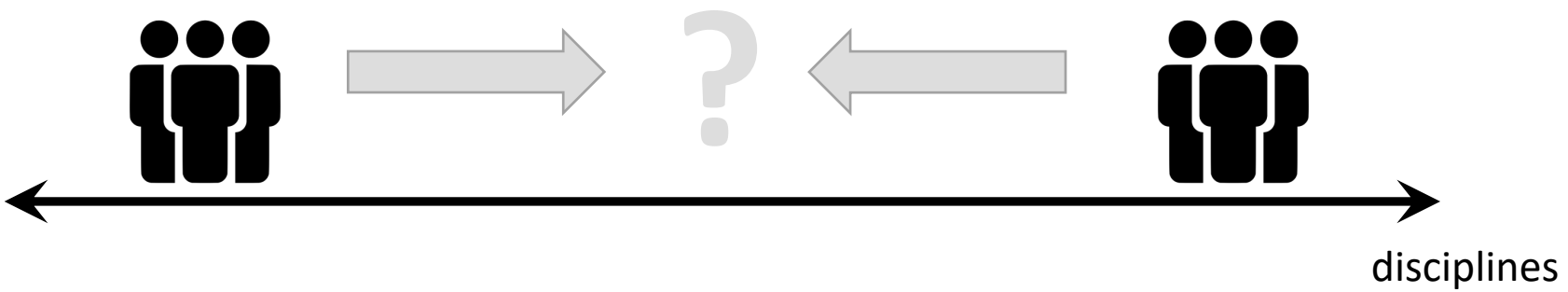
Operational Data Science:

What are similarities and differences between the university and the industry contexts?

What are the ramifications?

Academic Models for Connecting Disciplines

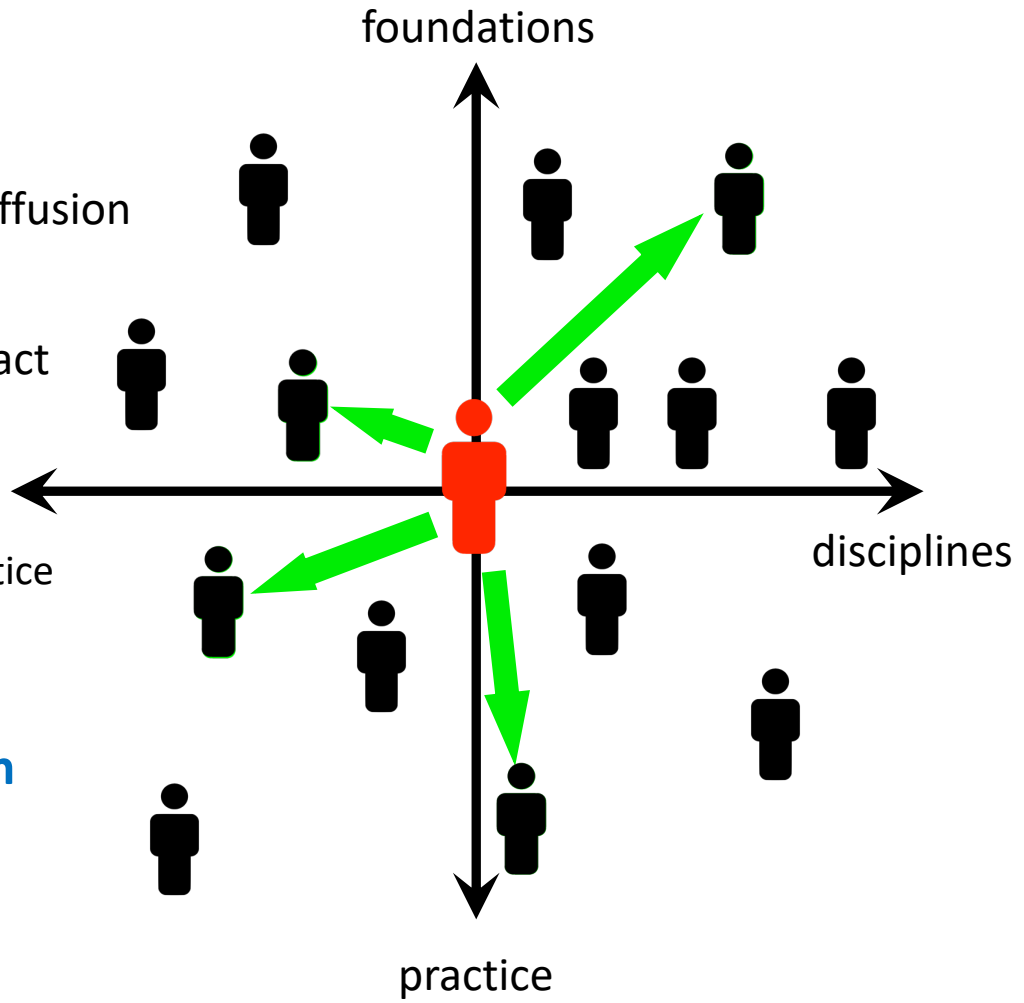
- There are superb people working within different disciplines
  - Academic career structure incentivizes narrow specialization
- Problem: How to enable people to work together?



- Responses: bring people together in multi-disciplinary work to solve specific problems
- Agencies such as NSF increasingly soliciting proposals for multi-disciplinary work
- Many universities have been cultivating responses to prepare faculty
  - Interdisciplinary seed grant programs
- Problem: how to sustain engagement beyond funding period?

Industrial Model for Connecting Disciplines

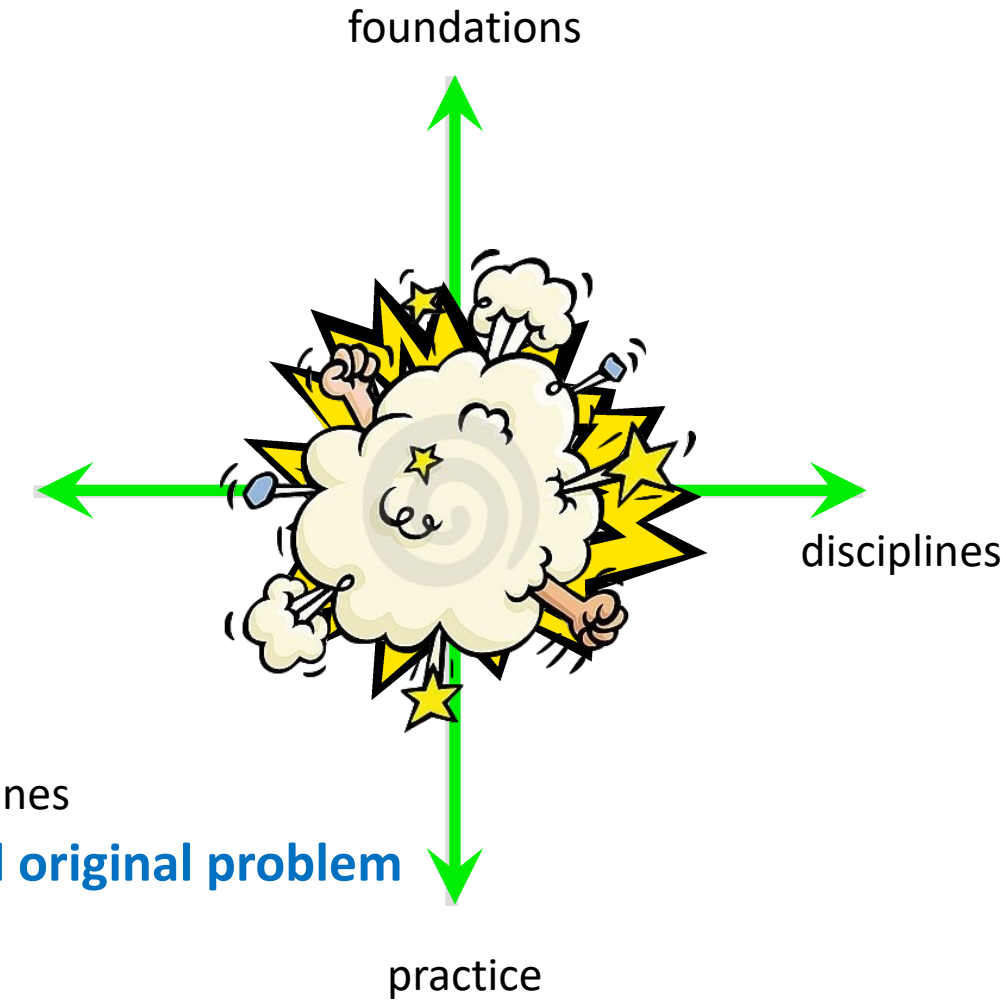
- There are superb people working within and across disciplines
  - Career incentives favor collaboration, knowledge acquisition, skill diffusion
- There are superb people working across foundations and practice
  - Career incentives to invest time on engagement, with payoff of impact
- Crucial industrial role for **central people**
  - **Connect others** who have greater specialisms
  - Have knowledge and experience to bridge disciplines, foundations, and practice
  - Enables team to maintain diversity of knowledge and approach
  - Fits Carver Mead's model of the *Tall Thin Engineer**
- **Low barriers in industry to exploratory multidisciplinary research**
  - Organizational imperative: the constant search for the next secret sauce
- **Outcome**
  - Build **long-lasting research projects & teams** combining science with impact



*"one who becomes accomplished in all aspects of chip design, from algorithm creation to layout, from concept to chip"

Interdisciplinarity: Benefits and Challenges

- Inbound benefits to the center (what happens now)
  - Everyone brings their strengths to meet in the center
  - Assemble diverse parts to accomplish specific goal
- Outbound benefits from the center (need more of this)
  - New foundational work stimulated by practical requirements
  - Practice adjusts to make best use of foundational work
  - Disciplines refreshed by importing new ideas and methods
- Hardest challenge: joining everything in the center
  - Less certainty than in single disciplinary work
    - Less clear how diverse components fit together
    - Harder to anticipate problems
  - Friction means exploratory work take longer across disciplines
    - Different languages, levels of abstraction, notions of what constitutes research
  - Challenges of working in the center often under-appreciated at the edge
    - Seen as mere “applications” or ”theory” (depending on point of view)
- Risks & potential rewards are inherently higher than for single disciplines
- Refreshment of disciplines crucial to **sustain momentum beyond original problem**
  - **Want long-lasting research venture not just one-off solutions**



TEXAS A&M Institute of Data Science

Interdisciplinarity: Recommendations

- Universities should hire more people in the center
  - From industry, agencies, national labs
  - Mixed roles OK: research active, team building, management, admin
  - Give them enough job security to enable them to take risks
- Incentivize and support interdisciplinary activities
  - Sufficient to make projects interdisciplinary, not necessarily people
  - Even if (established) faculty are not-interdisciplinary, students can become so
- Develop capacity for rapid interdisciplinary and vertical integration
  - Faculty-to-Faculty connections take a long time to develop
  - Interdisciplinary faculty: incentivize departments to hire them
  - Data Science: statistical consultants, systems specialists, developers

Operational Data Science: Summary

- Universities are awash with opportunities for Data Science R&D
- Collaborations help all partners
  - Operations:
    - rapid prototype prototype embodying state-or-art idea
  - Faculty:
    - develop credentials for outreach, vertical integration
    - new contexts can be source of further research questions
  - Students
    - Enhance interdisciplinary profile for employment
- Some similarities with industry practice
  - University hiring practices can learn from these

Thank you!

Questions?

Shoutout to recent **student** collaborators in networking

- S. Panda, **Yixiao Feng**, S. G. Kulkarni, K. K. Ramakrishnan, N. Duffield and L. N. Bhuyan (2021), *SMARTWATCH: Accurate Traffic Analysis and Flow-State Tracking for Intrusion Prevention using SmartNICs*, CoNEXT'21
- **Yixiao Feng** & N. Duffield (2002), *Multiscale Energy Network Tomography*, IFIP Networking 2002
- **Yunhong Xu**, K. He, R. Wang, M. Yu, N. Duffield, H. Wassel, S. Zhang, L. Poutievski, J. Zhou, A. Vahdat (2022), *Hashing Design in Modern Networks: Challenges and Mitigation Techniques*, USENIX ATC 2022