Data Science & Visualization

Working definition:

Science of working with data, interactive & communicative pipeline from data to discovery to decision to presentation to action
Overarching Themes

• Making meaning out of data
  • Modeling data science user intent, understand process of data science

• **Human-in-the-loop analysis methods**: interactive visual analytics, interpretable Machine Learning, understanding rationale and steer
  • interactive data science vs traditional batch model

• **Automated analysis methods**: Data representations, storage, cleaning, preprocessing, in prep for visualization vs. in situ visualization
  • Guidelines for Visual representations of diverse data
  • Guidelines for Analysis methods of diverse data

• Scalable interactive visualization strategies
  • Immersive analytics, walls, VR, AR

• Accessibility:
  • Vis-4-all, ML-4-all, Data democratization, visual literacy
  • Tools for data science productivity, Accessible DS & visualization APIs
  • Taxonomy, common language, Translating methods across domains

• **Data science education**: curriculum, interdisciplinary collaboration
Recent Successes

• Data science education programs have been developed and become in high demand; CS enrollments up; increased resources
• Growing industry in data science, employment demand
• Expanding use of VR (virtual reality) in many industries (e.g., medical training)
• High-performance computing has made big data visualization much more feasible; in-situ vis
• Public reception of progress on problems is greatly aided by visualization
• Software libraries for data science & visualization
Major Obstacles

- Disciplines and Scientific Domains have different language and terminology for describing similar methods --- across computer Science, different physical sciences, social sciences

- When developing courses and curricula for teaching data science and visualization, it is challenging to form any kind of standard content to include in that course or curriculum

- “Turf wars” in who will teach and claim as theirs Data Science -- -- computer science, statistics, domains, ...

- Data science & visualization has the algorithm side and the cognitive side ---- there needs to be better collaboration and sharing between researchers on each side

- Need more people, students, partnerships with industry

- collaboration between methods and applications
  - Don’t call me a “domain” person, call me a data scientist too
  - Don’t call me the “data” person, call me a scientist too
Areas of Neglect

• Funding specifically for data science & visualization
  • Raise data science to the level of an actual science
  • Methods scientists often added to projects as service provider

• Speeding up data science process:
  • Collaborative tools ---- need a pipeline from data to extracting meaning in data science
  • Data cleaning, feature engineering

• Data sharing and curation, accessible repositories
  • Public data, gold standards, benchmarks
  • “Dark data” ---- lost or hidden data, not extracted and made into electronic form

• Communication between domains and methods people
  • Domain scientists are often neglected in developing tools

• Domains often neglected in data science education
Strategic Priorities & Investments

• Promote **data science education standards** ---- e.g. workshops for educators to compare curricula, share examples, guidelines
• Data science + X, or X + data science; capstones that go beyond just methods
• Develop more interdisciplinary thinking, beginning with education of students coming from different majors ---- Data science is not **just** computer science or statistics
• Bring together different components of data science pipeline
• Funding for interactive data science and visualization specifically ---- e.g. revisit FODAVA
• Develop more ways to connect people developing methods to people in substantive domains that could benefit
• Promote data standards
• Promote collaboration between industry and academia in data science & visualization