Understanding the Data Science Technical Landscape

NIST Data Science Symposium

Elham Tabassi + Hoa Dang

March 5, 2014

Today
Wild west but not Wild wild west!

- Data science in use in many disparate communities
    - Such as life sciences, bioinformatics, cyber security, etc.
    - Coming together as a community of scientists to build a common infrastructure and usable tools
- Some paradigms for data science are proposed, but not widely known or accepted
    - Some areas such as data curation, pattern detection and meta data are more well understood in some disciplines than others.
- Many tools and methods exist but they would require too much adaptation to meet individual data science problems
    - Challenges: Scalability, visualization, decision making, etc.

What are the gaps in data science technologies and methods?

Gaps

- Need Common language
  - Definition of data science, its scope, communities, etc.
  - Increase awareness of importance of data preservation and re-usability
- To identify the known unknowns learn the known knowns
  - Bound the problem space
- Standards

Challenges

- Data preservation and documentation
  - To facilitate reuse of data for repeatability and other uses not envisioned initially
- Infrastructure + Analytics + Visualization
- Standards
  - For Creation, storage, format to facilitate use, reuse, visualize and analysis of data

How do you think NIST can help

- Building Data Science **Community**
  - a forum to find point of contact in different domain and agencies
- **Taxonomy** and reference data sets
- Catalogue of existing tools and methods with guidance on how to use and which problems they are suitable for
- Algorithm verification, validation and compliance
- **Standards** for data collection; **data storage**, meta data fro data, format data types, and **data preservation**