Data Science in the 21st Century

chair: Barbara Ryder, Virginia Tech

co-chairs: Lise Getoor, UC Santa Cruz

Steve Heller, Two Sigma

Snowbird CRA Conference
July 19 , 2016

Introductions

- Panelists:
  - David Culler, UC Berkeley
  - Rahel Jhirad, Hearst
  - Rayid Ghani, UChicago
  - Rob Rutenbar, UIUC

Roadmap

- Introduction - Lise
- Panelist Remarks - David, Rahel, Rayid, Rob
- Q&A

http://bit.ly/cradsp

Questions

- i. What is data science and what is its relation to computer science?

- ii. What do students have to learn about data science to become practitioners? Should learning data science require technical skill, and if so, what is the minimum skill set?

- iii. What are the big research questions in data science?

Data Science Panel: Introduction

Lise Getoor

UC Santa Cruz

Snowbird Panel on Data Science
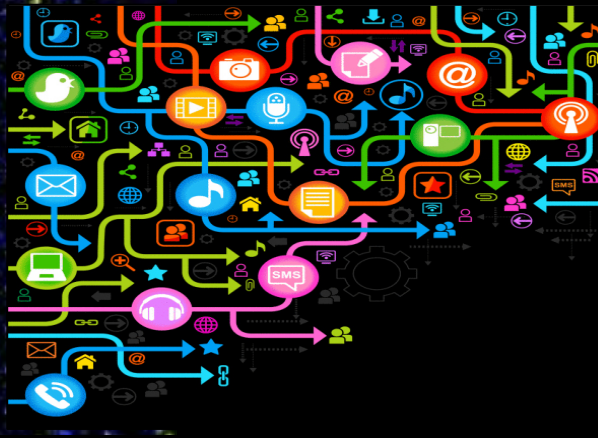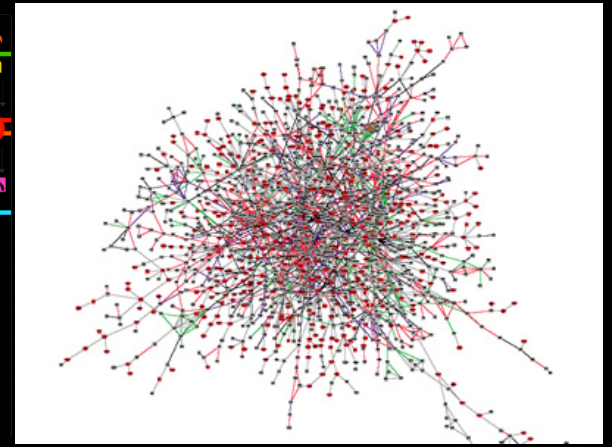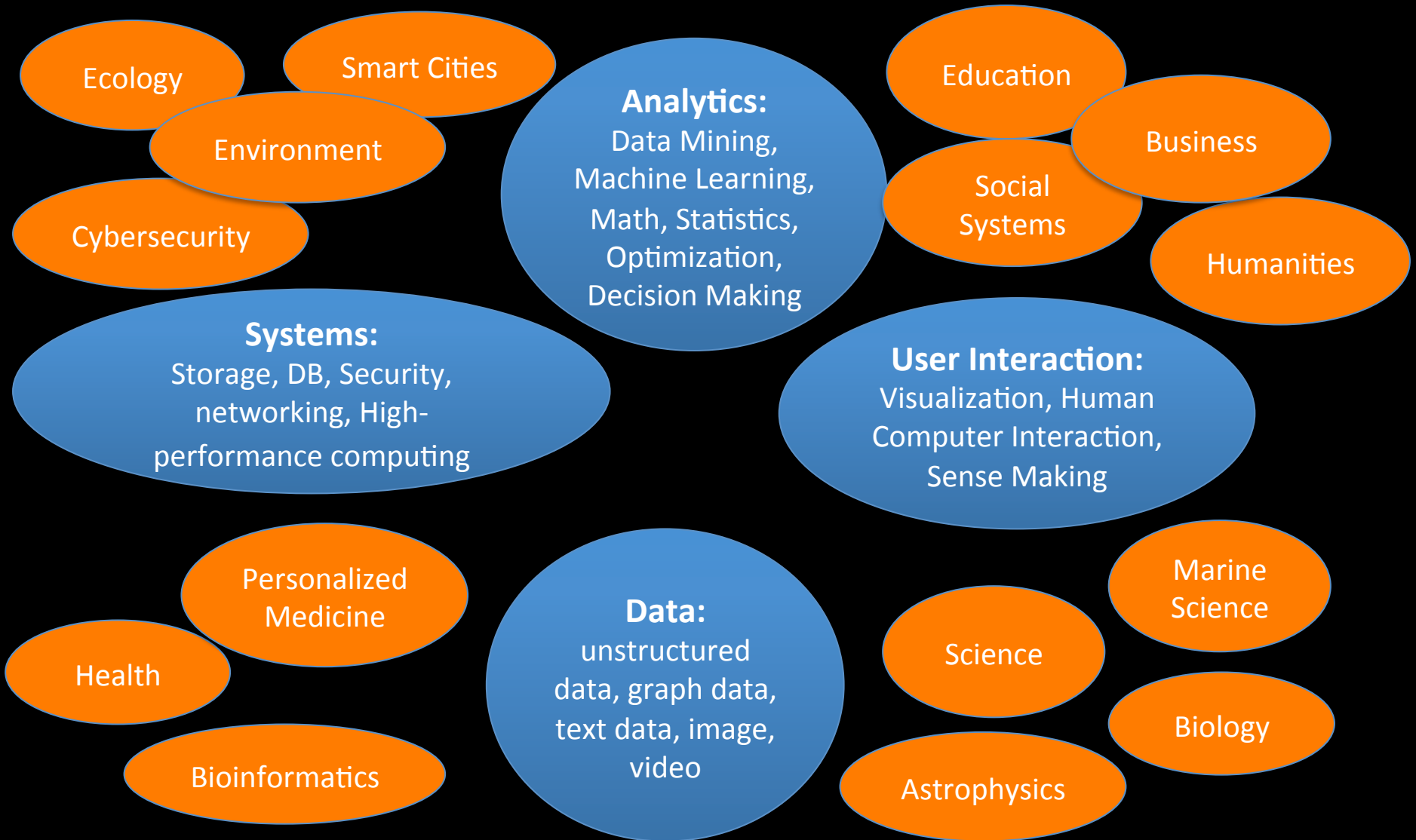July 19 , 2016

What is Data Science?

buzz

magic

old

huge　　　　heterogeneous　　linked & noisy

Definition: Turning data into knowledge,

and knowledge into action
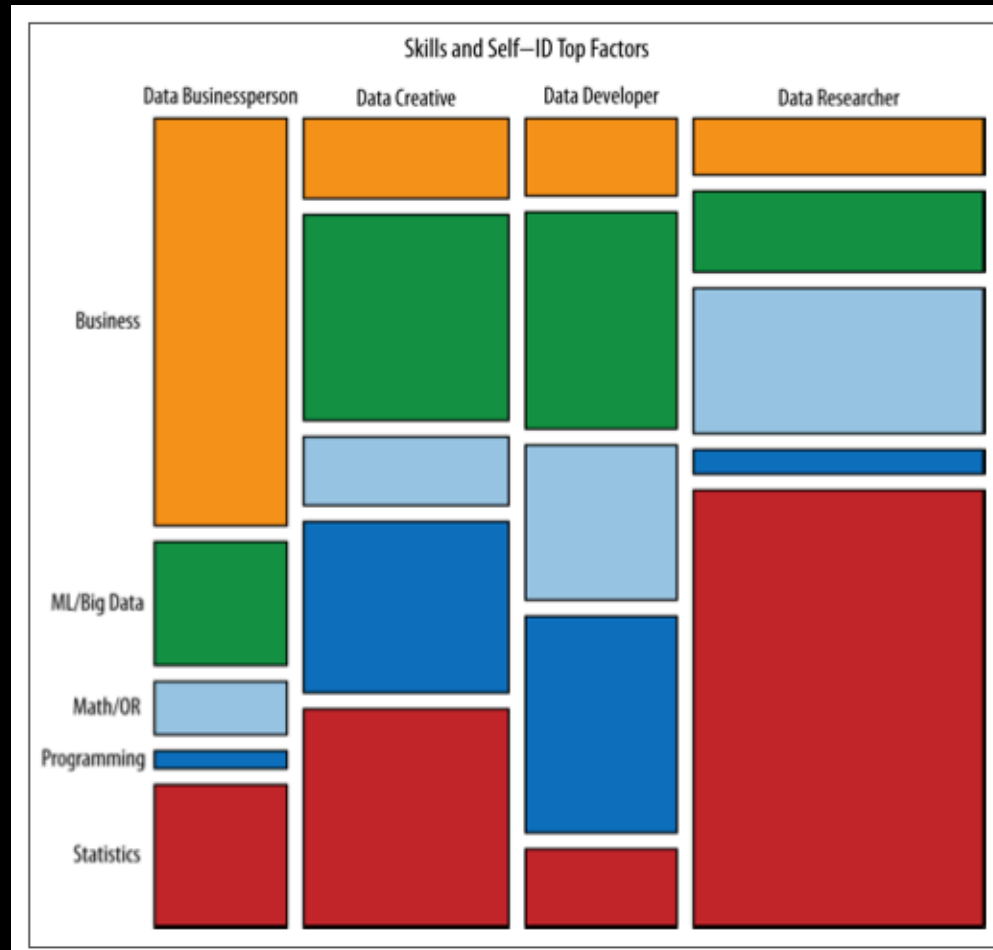
Data Science is Interdisciplinary

What do data scientists do?

Courtesy: Peter Skomoroch

Four types of data scientist

source: "Analyzing the Analyzers" O'Reilly
Media

Why?

Two Views

- **Pessimist:** No choice – data is transforming academic disciplines and we need to participate in defining the discipline of data science to not be left behind

- **Optimist:** Opportunity to develop new computational paradigms for data-driven and model-based methods and to train students to do *responsible* data science

How?

Resources

NSF Report on Data Science Education
- http://bit.ly/DSEW2015_DraftReport

- NSF CISE AC Report on Data Science
  - Outline, expected out this fall

- NTRD Report on Federal Big Data Research and Development Plan

- NAS report on Frontiers of Massive Data Analytics, 2013

- NSF Workshop Report on the Social, Econonic and Workforce Implications of Big Data Analytics and Decision Making

- ASA Statement on Role of Statistics in Data Science

- **CRA Statement on Data Science**

Academic Audience

- CS departments fall into following categories:
  1. **Leading** efforts at their university around DS
  2. Part of an **integrated** effort across campus
  3. One of many **independent** efforts across campus
  4. Not involved, but **wondering** whether to get involved, as DS efforts are being under taken by other departments, and they are not included
  5. Not involved and **not interested**

Academic Audience

- CS departments fall into following categories:
  1. **Leading** efforts at their university around DS
  2. Part of an **integrated** effort across campus
  3. One of many **independent** efforts across campus
  4. Not involved, but **wondering** whether to get involved, as DS efforts are being under taken by other departments, and they are not included
  5. Not involved and **not interested**

CRA Statement Goals

1. Communicate the centrality of computer science research and education to the emerging discipline of data science. Emphasize the need for computer science knowledge in data science and vice-versa.

2. Highlight the need and opportunities for collaboration around data science research between academia and industry.

3. Commit to developing a program of responsible data science research, which acknowledges the limitations of traditional computational and statistical thinking for many of big data science problems and highlights the importance of security, privacy, interpretability, ethics and fairness.

CS Research Opportunities

- From a computational point of view, data science requires a much deeper understanding and representation of how data is acquired, stored and accessed. Data lineage, data quality, quality assurance, data integration, storage and security all need to be rethought.

- The traditional approach of acquisition and storage, then processing often does not work for **Data Science ≠ Hadoop + R** linear algorithms; a **Data Science ≠ Hadoop + R** istic, stochastic and online approaches

It will require foundational new research into computation, systems, machine intelligence, and user interaction

new techniques that can cope with heterogeneity and biased sampling. causality

- The challenges in scale and heterogeneity also fundamentally change how users interact with data, how the data is visualized, what algorithms are needed to support understanding and interpretation of the results of data science models, and how user feedback is incorporated

Panel Questions

i. What is data science and what is its relation to computer science?

ii. What do students have to learn about data science to become practitioners? Should learning data science require technical skill, and if so, what is the minimum skill set?

iii. What are the big research questions in data science?

http://bit.ly/cradsp

Your Turn

- At your university:
  - **What** is being done?
    - New degrees, departments, etc.?
  - **Who** is doing it?
    - Which departments are leading the efforts?
  - **How** is the effort being supported?
    - Who is championing the effort?

Closing Remarks

- Let's work together to share experiences and resources

- Please see and share your resources here:

    http://bit.ly/cradsp