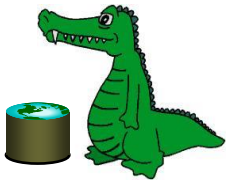
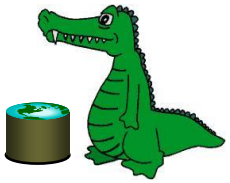


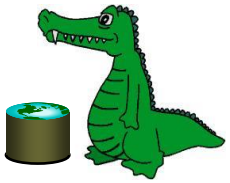
Logistics

- Next week: 1nd announced pop quiz coming Monday and in-class labs Wed. (unix) and Friday (python/panda) – future pop quiz may not be announced
- Prepare Lab by installing the virtual machine on course website
- If you are still looking to register, contact Todd Best (CSE405) or Adrienne Cook after 3pm before 4:45pm today: https://www.cise.ufl.edu/people/staff/admin/tbest

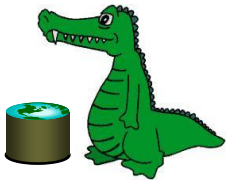


Review

- Why: fourth paradigm, BI/ML, data deluge, unreasonable effectiveness of data, new trends in health, politics, information, financial industries
- Where: Big data
- What: definitions and comparison to database, BI, DM and ML

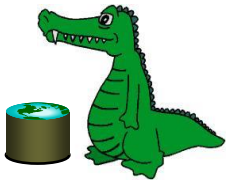


How to do Data Science?



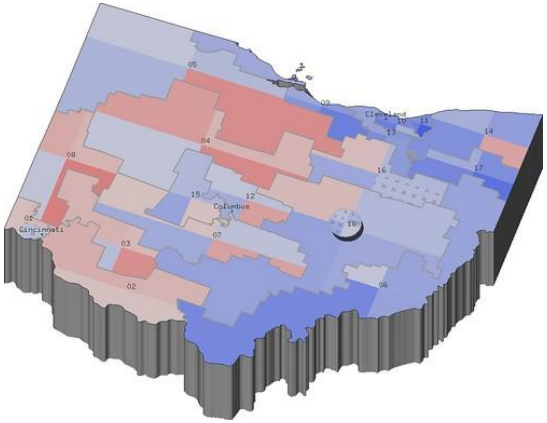
Ben Fry's Model

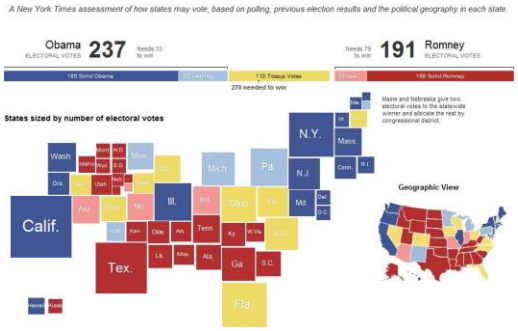
1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact

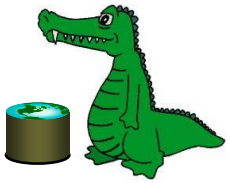


Jeff Hammerbacher's Model

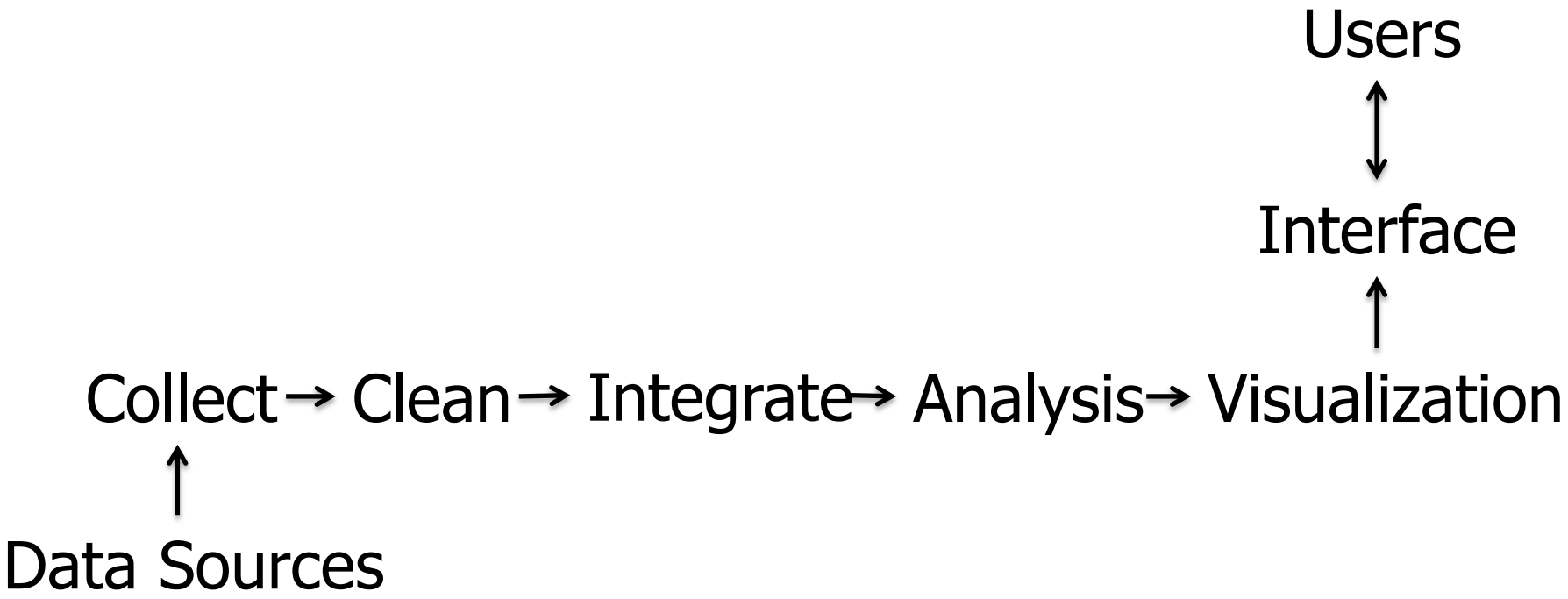
1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

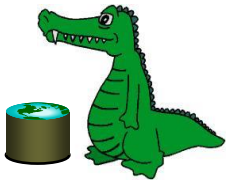






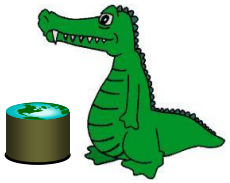
The Life of Data (Dr. Wang's Model, 2011)





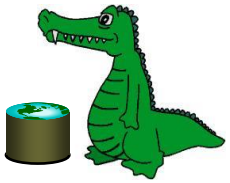
Challenges in Data Science

- Preparing Data (Noisy, Incomplete, Diverse, Streaming …)
- Analyze Data (Scalable, Accurate, Real-time, Advanced Methods, Probabilities and Uncertainties ...)
- Represent Analysis Results (i.e. data product) (Story-telling, Interactive, explainable…)



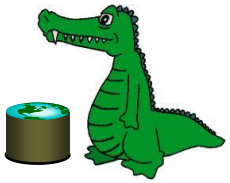
What's Hard about Data Science

- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Pitfall of Overgeneralizing
- Communication
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototype → Production transitions
- Data pipeline complexity



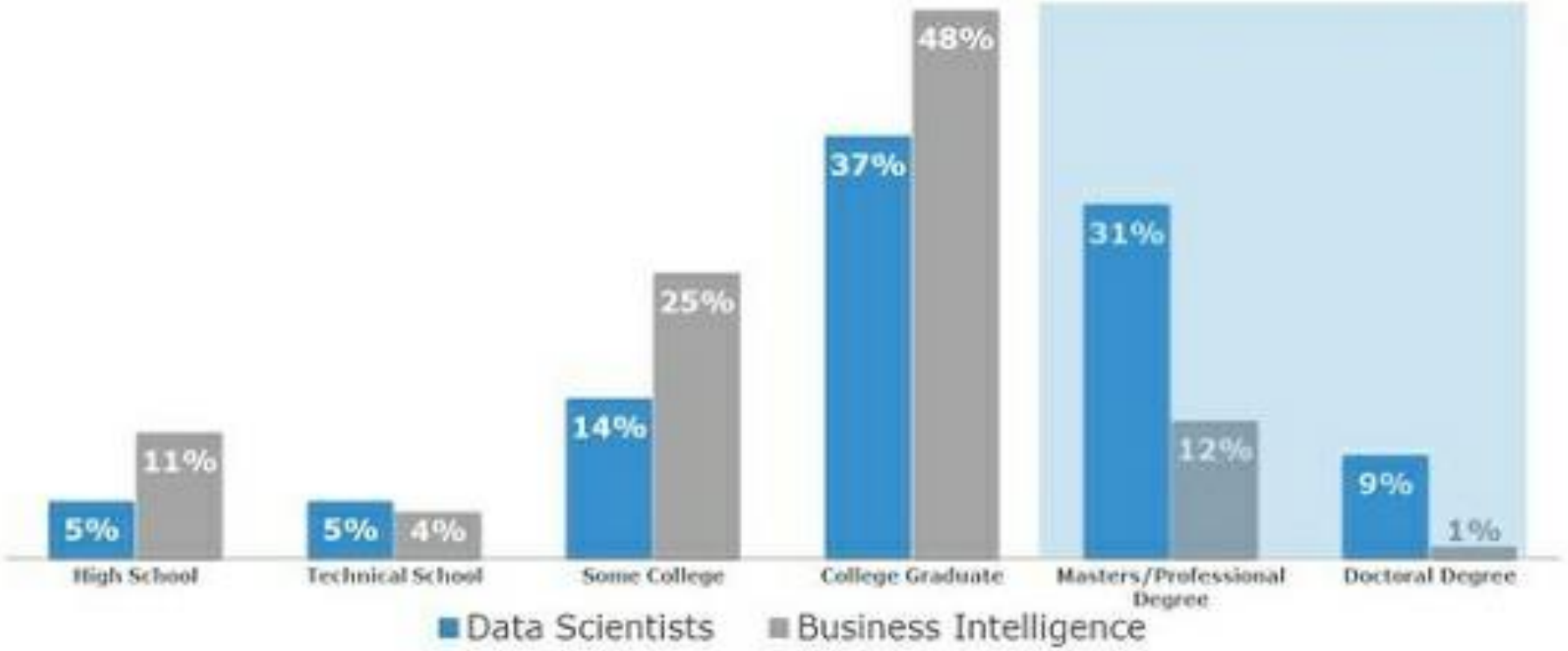
Skill Set of a Data Scientist

- Data Management
  - Data collection, storage, cleaning, filtering, integration …
- Large-scale Parallel Data Processing
  - Parallel computing
- Statistics and Machine Learning
  - Data modeling, inference, prediction, pattern recognition …
- Interface and Data Visualization
  - HCI design, visualization, story-telling …



Who are Data Scientists?

Data science requires greater education



40% of data science professionals have an advanced degree – and nearly one in ten have a doctorate. In contrast, less than 1% of BI professionals have a PhD.

EMC²

Analyzing the Analysts

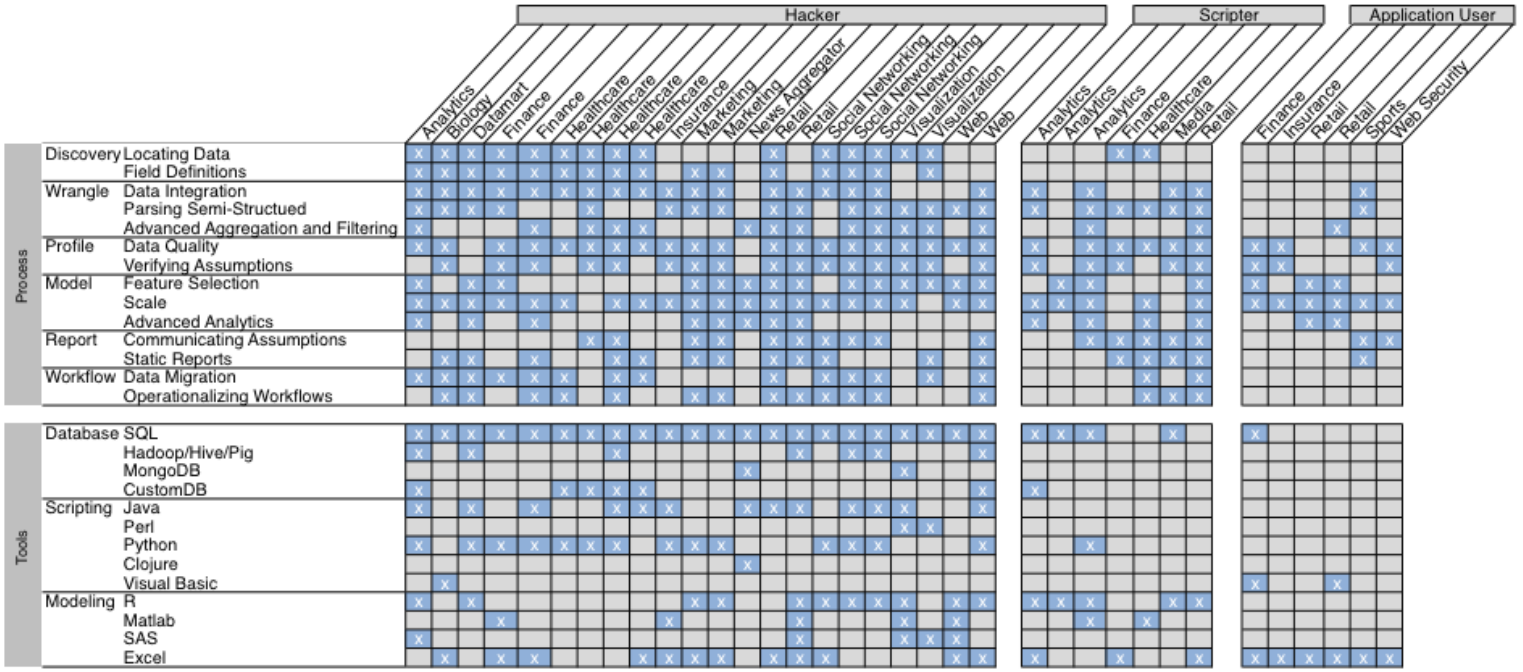
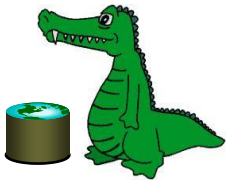


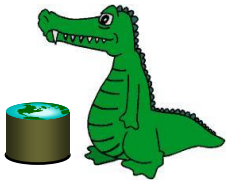
Fig. 1. Respondents, Challenges and Tools. The matrix displays interviewees (grouped by archetype and sector) and their corresponding challenges and tools. *Hackers* faced the most diverse set of challenges, corresponding to the diversity of their workflows and toolset. *Application users* and *scripters* typically relied on the IT team to perform certain tasks and therefore did not perceive them as challenges.

From Kandel, Paepcke, Hellerstein and Heer, "Enterprise Data Analysts and Visualization: An Interview Study", IEEE VAST 2012



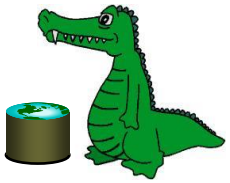
Kandel et. al. Data Analysis Process Model

- **Discover** data necessary to complete an analysis tasks.
- **Wrangle** data into a desired format.
- **Profile** data to verify its quality and its suitability for the analysis tasks.
- **Model** data for summarization or prediction.
- **Report** procedures and insights to consumers of the analysis.
- Additional Challenge in **Workflow** Management



"Big Data" to "Data Science"

- "… the sexy job in the next 10 years will be statisticians,"
  - Hal Varian, Google Chief Economist, 2009
- the U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018.
  - McKinsey Global Institute's June 2011
- New Data Science institutes being created or repurposed – NYU, Columbia, Washington, UCB,...
  - Berkeley, UW, NYU collaborate on $37.8M data science initiative from the Gordon and Betty Moore Foundataion, Alfred P. Sloan Foundation, 2013
- New degree programs, courses, boot-camps:
  - Fellowships to training courses of Data Scientists: http://www.skilledup.com/articles/list-data-science-bootcamps
  - MS in "Big Data Science", "Data Science and Analytics" …



Summary

- Why now: Dawn of Big Data, Need for Advanced Analytics and Cloud Computing
- What is it: Data → Data Product, many examples incl. Google, Netflix, Splunk, LinkedIn
- How to become: Data management, parallel computing and data processing, statistical machine learning, and visualization skills
  - Life/Workflow of Data Analytics
- Who are data scientists: Data Scientists are in great demands, from industry to government to science. Go Data Science!