2013-07-03 DataEdge debrief

This week featured three brief presentations summarizing talks at DataEdge, with related discussion.

Chris Hoffman:
- Kevin Koy - Geospatial Innovation Facility
  - Geospatial - a way to convey complex findings, part of data science toolkit
  - How has biodiversity changed over short/long period of time?
  - Berkeley Ecoinformatics Engine (ecoengine.berkeley.edu) - people can add own data sets, API access
  - Public annotations, but still figuring out what you can do with the information (e.g. 10 observations on same photograph)
  - Humbling for collection owners to realize that there's some better experts out there than them (e.g. tank identification in WWI photographs)

Steve Masover:
- Kate Crawford - MS Research
  - Myths about big data: new, objective, won't discriminate, makes cities smart, anonymous, can opt out
  - Objectivity of results-- selection biases big data questions, mining 20 million tweets (but mostly from Manhattan, and the part with power)
  - Making cities smart: only works if analysts are smart (auto pothole detection-- neighborhood differences in who's carrying around smartphones)
  - Anonymity: DOB, gender and zipcode sufficient to identify people

Last year: more enthusiasm/hype, this year -- more considered
Are reservations due to lack of understanding about machine learning?
Most of work in data science relates to cleaning, harmonizing, aligning -- a lot of things can happen in these stages

David Greenbaum:
- Teaching data science, what is the meaning of data science
- Interesting industry conversation going on
- Rachel Schutt -- statisticians now have "cool jobs"
- Invited people from NY area to be guest lecturers in data science class
- Students from a variety of disciplines
- Students tend to know something about statistics, something about data science, something about domain
- Data scientist: "Person who is better at statistics than any software engineer and better at software engineering than any statistician"