Public interest reporting defined

- Information of importance to the public that powerful institutions would prefer to be kept hidden or secret.
- A method of comparing what ought to happen with what actually happens
  - Breaking laws or rules
  - Taking advantage of the helpless
- Comforting the afflicted and afflicting the comfortable.

Common traits

- Reporting shares some traits with other investigative or exploratory fields:
  - Use of multiple sources with varying accuracy and collection methods
  - Practitioners have little or no technical ability or interest
  - Much of the data is in the form of unstructured text
  - Start with broad, ill-defined questions, not discrete goals
  - Combines field work and lab work (street reporting and data or document analysis)

Uncommon traits

- Rarely spend much time on a given data source or question
- Reporters look for tips and examples, not evidence or statistical patterns.
- Work from primary sources
- Analysis is not mission-critical
- Unique identifiers are censored.
- Little or no control over the form of the data: it’s up to the source
Typical data sources

- Interview notes, both on the record and for background only. Very little is taped, especially early in a story.
- Taped or videotaped hearings, sometimes transcribed already
- Calendars, correspondence, investigative and inspection reports
- Emails and press releases
- Excel spreadsheets created for printing, not analysis
- Structured databases
- Court documents, complaints, indictments, settlements, discovery.
- Online document collections, such as regulatory enforcement actions or audits.
- Handwritten forms, such as inspections or police incident reports, usually heavily redacted

Transparency myths

- Government records are easy to get in a usable form? Not so much.
- Examples, even when it is:
  - Recovery Act
  - Campaign finance
  - Crime
- Most government information is still hidden, secret or difficult to access
- data.gov Promise of 100,000 datasets. Currently have 728, half of which are annual state-level cuts of existing EPA data that has been available for years. No records, mostly just aggregated statistics. Almost nothing on it that wasn’t already readily available.

Harvesting Cash

Sources:
- People
- Database of 200 million farm subsidy payment
- Crop estimates
- Weather records
- Emergency records
- Property ownership records (local)
- Letters from members of Congress

Forced Out

Sources:
- File cabinets (some missing)
- Spreadsheet where the color meant different things.
- 200,000 or so housing code complaints and violations
- Paper records from court
- People (residents, activists, government officials)
- Physical inspection of properties
- Landlord-tenant court disputes
- Property assessments and deeds
Rove's Asset Deployment

Until the 2004 election, the White House policy team directed applications to key grants,ickle decisions, and other substantive actions in Washington, D.C., media markets. Top officials also traveled to nine congressional districts with close congressional races.

The Washington Post analyzed the travel of two cabinet members and their immediate deputies. More than three-quarters of the trips in the final five months of the election were in areas identified by the White House as key areas of interest, according to documents obtained by Congressional investigators. Many of the rest were in areas with close congressional races.
Resulting graphics at: http://www.washingtonpost.com/wp-srv/politics/obama/100days/

Sources: Original presidential documents, mainly in press release or pdf form.

Sources: Daily press releases and "pool" reports from the White House press corps, hand-coded.

Source: "Plum Book" list of jobs, original research on each person and congressional records.
## Theories of open data

- Governments prefer to lead you to the “right” answer, so they like colorful and “user-friendly” websites that convey their message.
- Transparency advocates look for government to first “wholesale” data, then work on their own website. They’ve had little luck.

## New directions in newsgathering

- What’s New? Smart scraping and text mining — indictments, settlements and regulatory actions across agencies; lawmaker releases and grant announcements.
- The Real Story: Anti-aggregation of stories and news from blogs, news sites, RSS feeds, government agencies grouped by story, not by source.
- Chronologies and social network tools to organize notes and see new connections
- Text mining of government documents: Jigsaw, Meandre, Document Cloud
- Audio and video analysis

## Opportunities for Computer Science students

- DukeEngage project (look on the DeWitt Wallace Center’s website for a link) to create newsgathering tools
- “Middle layer” of sense-making on unexplored local and national datasets
  - Stimulus
  - Contracts and grants
  - Dispersed records (calendars, etc.)
- Entity extraction, geocoding tools
- Visualizations that say something new