

CPS 296.1 (Spring 2012):
Project in Computational Journalism

Computational Journalism: Two Overview Papers

A high-level overview

- rev "Computational Journalism." Cohen, Hamilton, and Turner. *Communications of ACM*, 54(10), 2011.
 - Scope: **accountability reporting**, or **watchdog journalism**
 - Focus: generating stories

👉 What other aspects of CJ are not covered?

What changed investigative reporting?

- 50 years ago: copy machines
 - Ability to copy/archive data
- Late 1960's: statistical methods
 - E.g., systematic discrimination in mortgage lending
 - Ability to analyze data
- 1970's-1980's: relational databases
 - E.g., convicted criminals driving school buses
 - Ability to join, manipulate, and query data

👉 Personally, I see “data” written all over these changes!

Two challenges, intertwined

- Preparing data for analysis, and
- Analyzing data

👉 What's new or unique?

Preparing data

- Data cleansing
 - Not only fixing obvious typos and inconsistencies, but also identifying issues in the underlying data collection
- Information extraction
 - Distilling structured data from unstructured data with multiple modalities (text, image, a/v)
- Data integration
 - “Matching data sets never intended to be matched”
 - Heterogeneous datasets with no common identifiers

Analyzing data

- Reporters are more interested in finding “the unusual handful of individual items” than trends and patterns
- There is a strict limit on time and money spent
- Also
 - More data are in the public domain
 - Moderately high false positive rate is acceptable
 - Analysis can cut corners—it’s just a means, not an end

Five areas of opportunity

- Combining information from varied sources
- Information extraction
- Document exploration and redundancy
- Audio and video indexing
- Extracting data from forms and reports

👉 The majority of them are about preparing data; analysis is mostly exploratory

Info extraction (incl. a/v, forms, reports)

- Entities
 - Attributes
 - Relationships
 - Lots of tools are available, but they are expensive, difficult to use, or don't work in particular settings
 - Good solutions exist for some domains
 - E.g., comparison shopping
- 👉 But scope of information interesting to journalism is very open-ended

Information integration

- Example project: routinely combine press releases from all members of the Congress
 - How would you implement this?
 - How far we go in extracting information determines what questions we can ask
 - Stay on the level of documents/keywords?
 - “Price” + “Duke”
 - At least identify what are press releases?
 - Extract entities, attributes, relationships ?
 - Number of fund raising events attended by each in the past year

Exploratory analysis

- Visualizing relationships among documents or extracted entities
- Identifying redundancy among documents
 - A form of document clustering
- Spotting what's "interesting"
 - ☞ But how do we define interestingness?
- Better querying methods
 - Forcing a user to ask 500 queries is plain wrong
 - Any conjecture why they were needed?
 - Noisy, heterogeneous data and/or complex patterns?

Anything new or unique?

My thoughts:

- 👉 Humans likely have to remain in the solution loop
- The problem is very open-ended; it might be hard for automated methods to work well in general
- We must guard against “data determinism,” i.e., jumping to conclusions from incomplete and noisy data
- 👉 So, data preparation and analysis should be interleaved
- Automated analysis should guide costly data preparation efforts—focus on cleaning up/verifying data that lead to interesting findings

Tools/projects mentioned

- Data cleansing: Google Refine
- Speech recognition: Sphinx
- Visualization: ManyEyes, Tableau, Google Earth, TimeFlow, Jigsaw...
- Cloud/collaboration: DocumentCloud
- Crowdsourcing:
 - Talking Points Memo: users contribute story ideas
 - Guardian: users review MP expense reports
 - Public Insight Network: users serve as experts

👉 What would you add here?

Get more people interested

- Education
- Games!
- Research/project funding
- Awareness at computer science venues

So what do you think?

- Did the authors place too much faith on technology?
- Do you agree with the focus on preparing data?
 - Is there any low-hanging fruit left in that space?
- Should we focus on empowering the few journalists or the mass?
- Would you trust the crowd for investigative journalism?
- Who guards those who guard the guardians?
- Did you get any good project ideas?

A database-centric overview

- "Computational Journalism: A Call to Arms to Database Researchers." Cohen, Li, Yu, Yang. *CIDR* 2011.
 - A "cloud for crowd" vision, and
 - Some ideas for checking and finding facts
 - Less focus on data preparation
 - An appeal to computer science (database) researchers—journalism is not only a consumer of technology, but it can also drive computer science

Fact checking

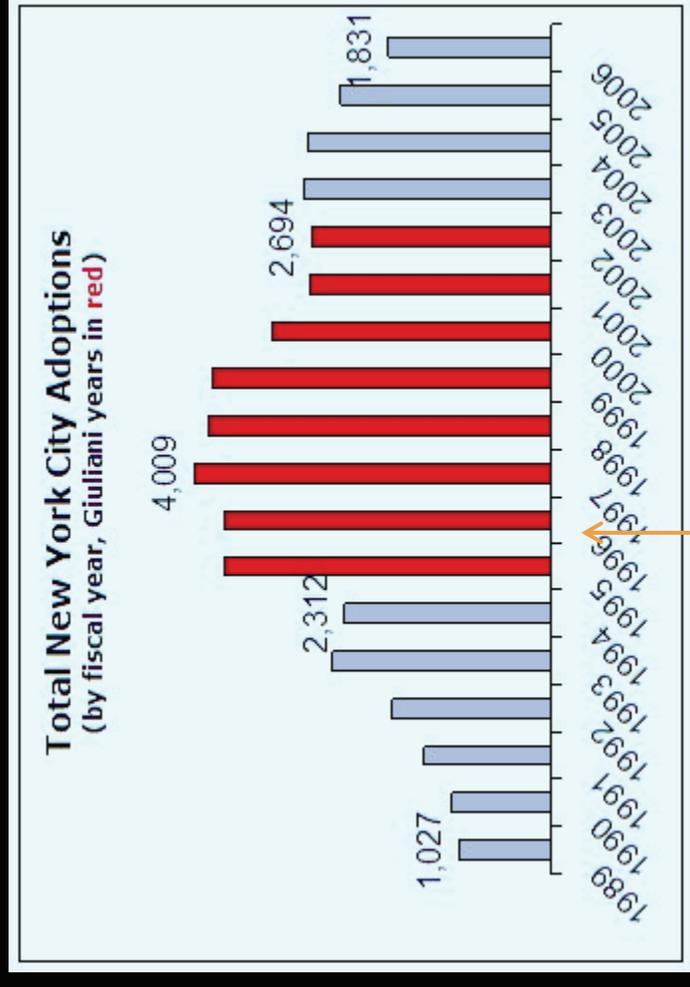
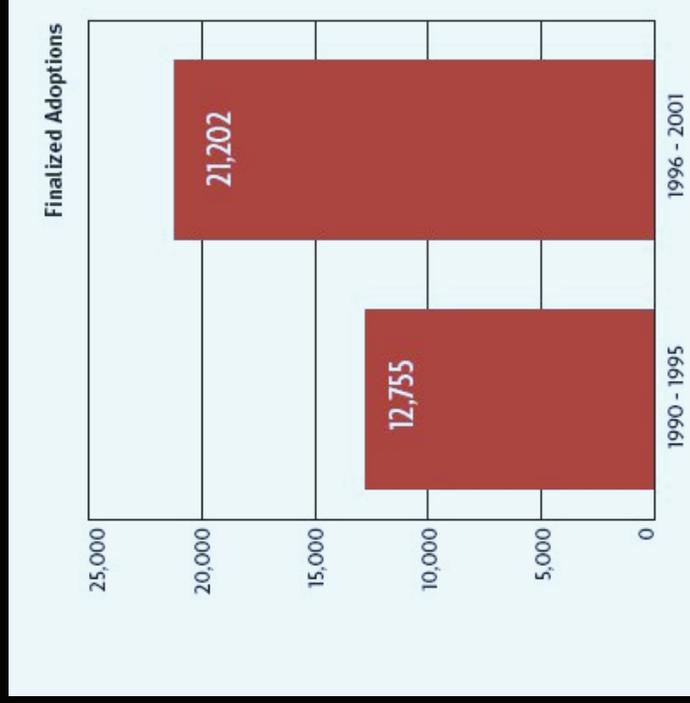
... (Lincoln) Davis voted with Nancy Pelosi 94 percent of the time...

... For 36 months in a row, our district has maintained the lowest unemployment rate among our neighboring five districts...

- Fact-checking is absurdly difficult, even if you know SQL and the databases are cleansed and documented
- A tool for checking claims in English, with no knowledge of SQL or database schema?
 - But is this simply natural language querying (NLQ)?

Example: Giuliani's adoption claim

- In the 2007 Republican presidential debate, Giuliani claimed that “adoptions went up 65 to 70 percent” in New York when he was in office



Administration for Children's Services was created in 1996

Fact checking \neq NLQ

- Claims often are vague and/or involve complex queries
- Users don't expect one-click fact-checking with instant gratification
- Clarifying a claim and tweaking the way it presents data are instructive in their own right
- ➔ An interactive interface that relies on user feedback
 - Suggest possible SQL queries for user to choose
 - To help user choose, show English translations, preview answers, ask questions...

Fact-check⁺

... For 36 months in a row, our district has maintained the lowest unemployment rate among our neighboring five districts...

- Test how robust a claim is
 - What's the margin? Did it change over time?
 - What if we compare with six instead of five districts?
- See if similar claims hold for different settings
 - How does my district do in a similar comparison?
 - How about median income instead of employment rate?
- Monitor a claim over time
 - What if we revisit the comparison a year later?
 - Can we get an alert when the streak is broken?
- ➡ Allow reuse of expertise/effort beyond a single story

Finding answers → finding questions

- A fact checking service would allow us to build up a “library” of datasets, queries leading to claims, and stories using them



- ⇒ “A reporters’ black box”
 - Learn “standard” query templates from the library and human experts
 - Run all templates on new/updated data to find claims that hold
 - Rank claims for further investigation by journalists

Vision: a cloud for the crowd

Cloud: aggregate/share computing resources

- Large-scale, real-time data analysis
- E.g., map/reduce for machine translation, information extraction, reporters' black box, etc.

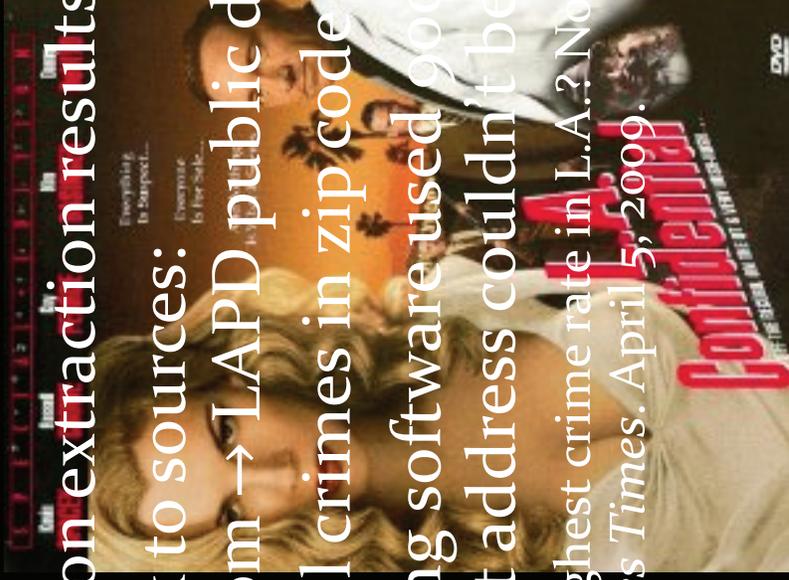
Crowd: aggregate/share data, tools, and insights

- Leverage the crowd in simpler and more effective ways
- An “optimizer” for the investigative process with crowdsourcing support

Example: crime-ridden LA City Hall?

Suppose many blogs seem to be talking about high crime rates around LA City Hall; what do you do?

- Verify information extraction results from blogs?
- Trace blogs back to sources:
 - EveryBlock.com → LAPD public database
- Check individual crimes in zip code 90012
- LAPD's geocoding software used 90012 as the default zip when a street address couldn't be mapped!
- ➔ Welsh and Smith. "Highest crime rate in L.A.? No, just an LAPD map glitch." *The Los Angeles Times*. April 5, 2009.



An investigative “optimizer”

- The investigative process is difficult to plan
- Can our system help plan it intelligently (incl. directing the crowd), in a goal-driven fashion?
 - Specify tasks declaratively
 - Identify mini-tasks that can be crowdsourced
 - Quantify cost-benefit of mini-tasks
 - Matching mini-tasks to users
 - Coordinate/reprioritize execution of mini-tasks
 - ...

So what do you think?

- Did authors convince you that there are interesting computer science research in CJ?
- How much of this sound like pie in the sky?
- Are they building the unrealistic assumption that data is complete and accurate?
- Did you get any good project ideas?

Reading for next Wednesday

- [rev](#) "News and Information as Digital Media Come of Age." Berkman Center for Internet and Society at Harvard University, 2008.
- "Shared Values, Clashing Goals." Cohen. *ACM Crossroads*, December 2011.
- Discussion leaders needed now!
 - I will always meet with leaders before class meeting