Human-in-the-Loop Data Management

CMPT 884, FALL 2016

JIANNAN WANG

https://sfu-db.github.io/cmpt884-fall16

Introduce Yourself

What's your name?

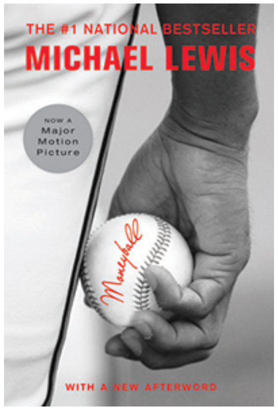
Where are you from?

M.Sc. or Ph.D.? Which year?

What do you want to get out of the course?

A Problem That Everybody Cares About!

How to manage data and extract value from it?



Key Resources

Algorithms

- Machine Learning, Statistical Methods
- Prediction, Business Intelligence

Machines

- Clusters and Clouds
- Warehouse Scale Computing

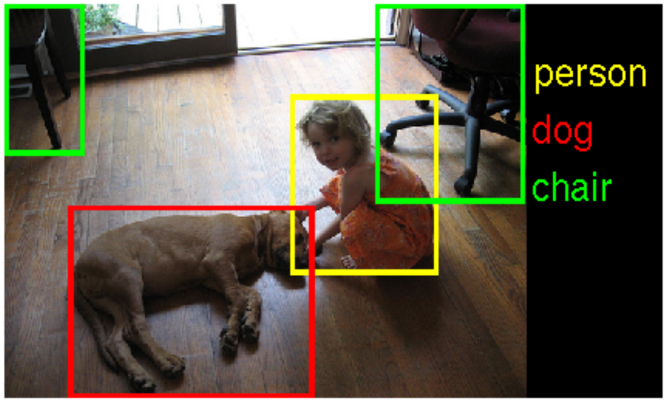
People

- Crowdsourcing, Human Computation
- Data Scientists, Analysts

From Prof. Michael Franklin's talk

An Example of Using Three Resources

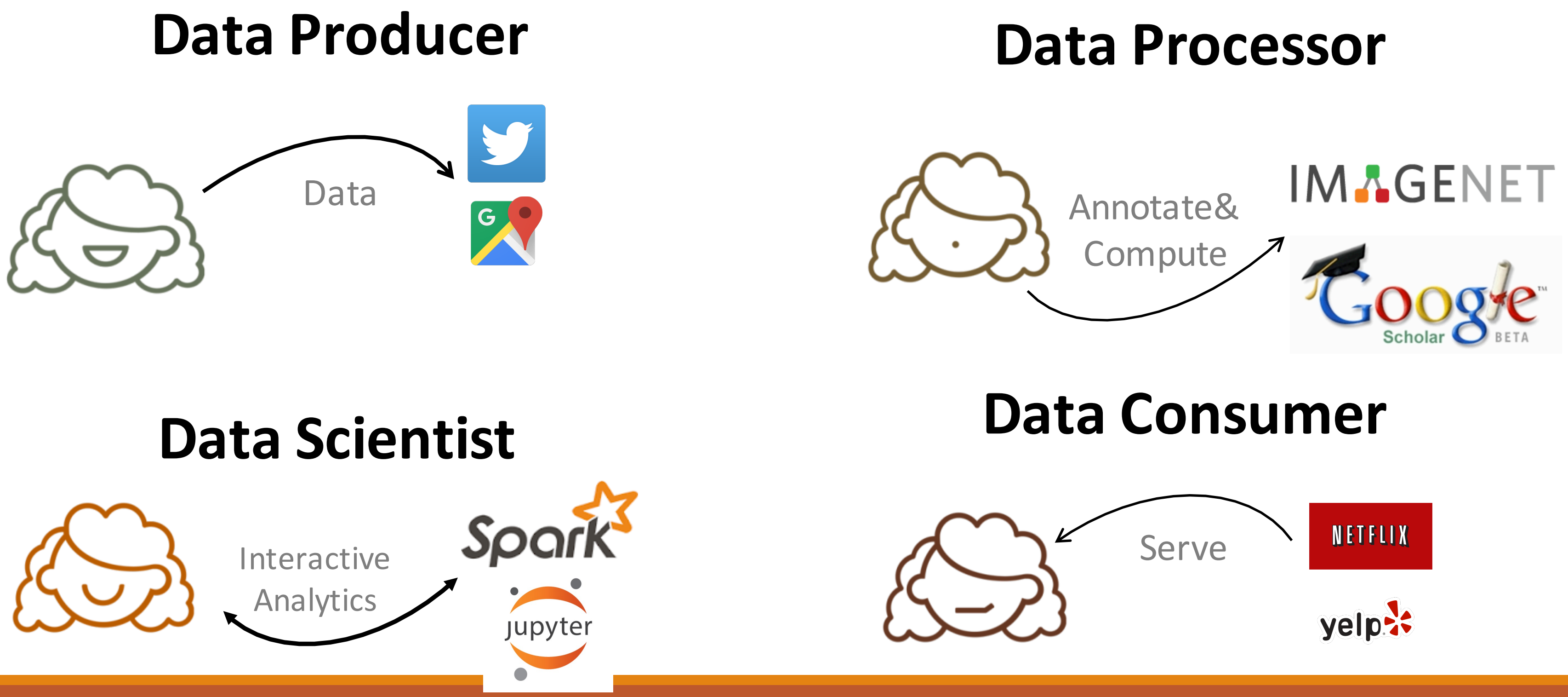
What are in the image?



How to solve the problem?

Deep Learning (Algorithms)
GPU Cluster (Machines)
ImageNet (People)

Human-in-the-loop Data Management



Data Producer

Data

Data Processor

Annotate& Compute

Data Scientist

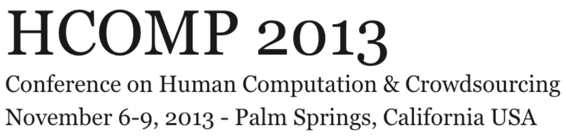
Interactive Analytics

Data Consumer

Serve

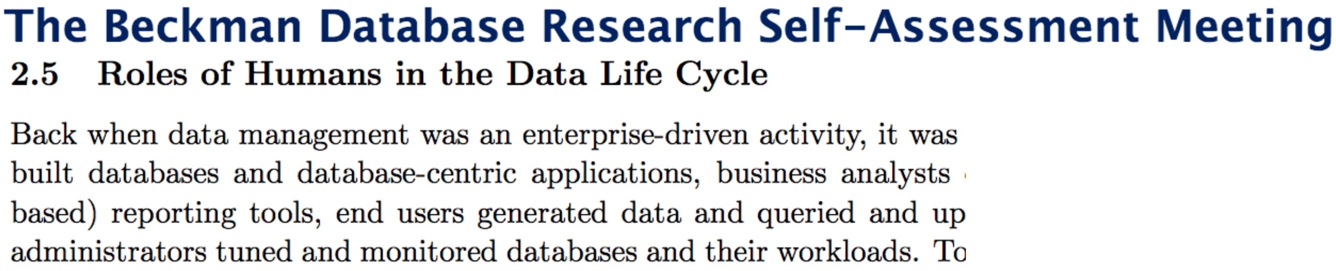
Icons created by Clara Joy from Noun Project

A very hot topic



HCOMP 2013

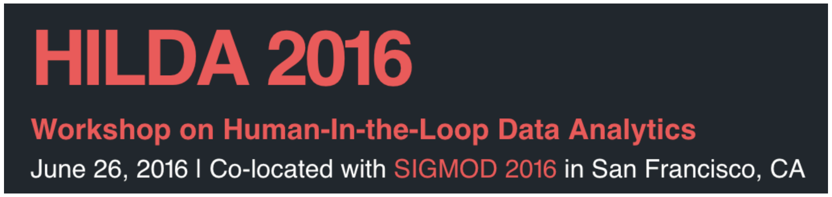
Conference on Human Computation & Crowdsourcing
November 6-9, 2013 - Palm Springs, California USA



The Beckman Database Research Self-Assessment Meeting

2.5 Roles of Humans in the Data Life Cycle

Back when data management was an enterprise-driven activity, it was
built databases and database-centric applications, business analysts
based) reporting tools, end users generated data and queried and up
administrators tuned and monitored databases and their workloads. To



HILDA 2016

Workshop on Human-In-the-Loop Data Analytics

June 26, 2016 I Co-located with SIGMOD 2016 in San Francisco, CA

Course Objectives

- Introducing students **the cutting-edge research on Human-in-the-loop Data Management**

Part 1: Crowdsouced Data Management

(Human as Data Processor, 13 papers)

Part 2: Interactive Analytics

(Human as Data Scientist, 17 papers)

30 papers

Part 1: Crowdsouced Data Management

(Human as Data Processor, 13 papers)

Machine-based

👎 Quality

👍 Time

👍 Money

Hybrid
Human and Machine

👍 Quality

👍 Time

👍 Money

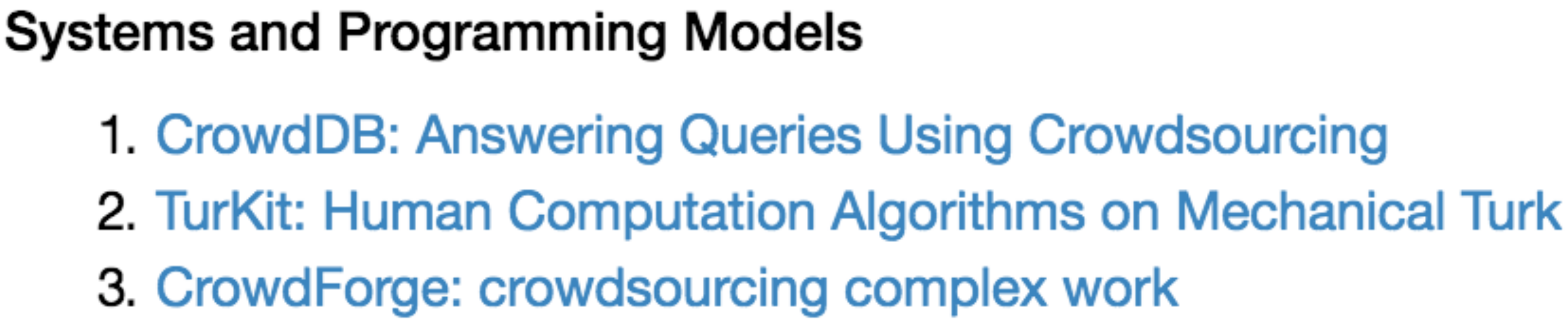
Human-based

👍 Quality

👎 Time

👎 Money

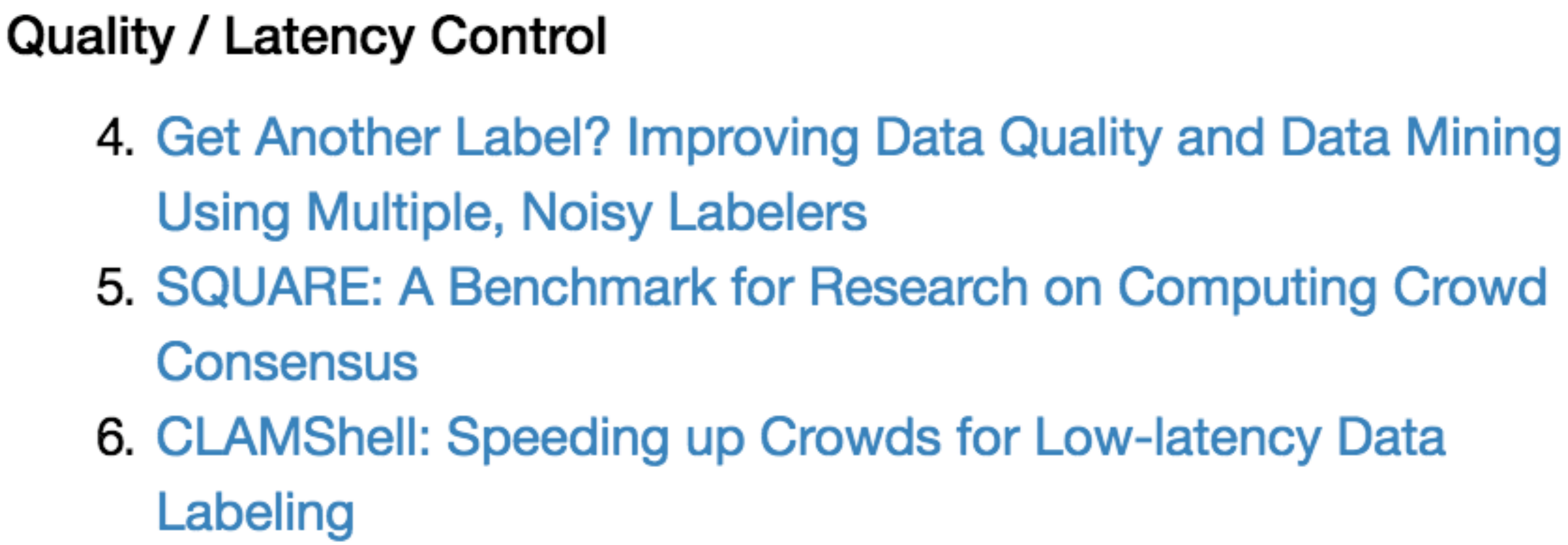
Part 1: Crowdsouced Data Management
(Human as Data Processor, 13 papers)



Systems and Programming Models

1. CrowdDB: Answering Queries Using Crowdsourcing
2. TurKit: Human Computation Algorithms on Mechanical Turk
3. CrowdForge: crowdsourcing complex work

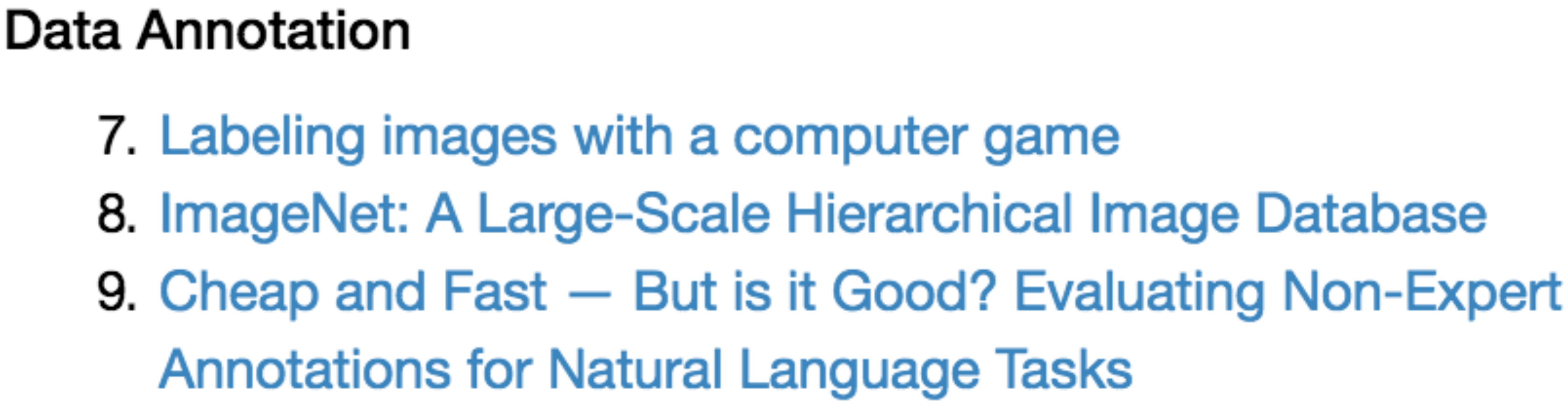
Part 1: Crowdsouced Data Management
(Human as Data Processor, 13 papers)



Quality / Latency Control

4. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers
5. SQUARE: A Benchmark for Research on Computing Crowd Consensus
6. CLAMShell: Speeding up Crowds for Low-latency Data Labeling

Part 1: Crowdsouced Data Management
(Human as Data Processor, 13 papers)

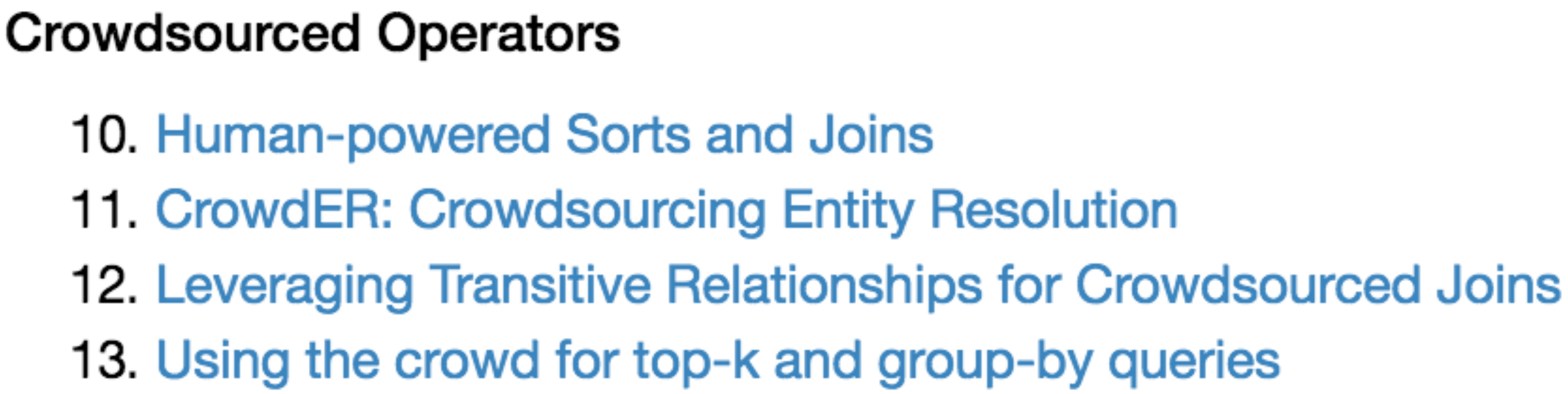


Data Annotation

7. Labeling images with a computer game
8. ImageNet: A Large-Scale Hierarchical Image Database
9. Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks

Part 1: Crowdsouced Data Management

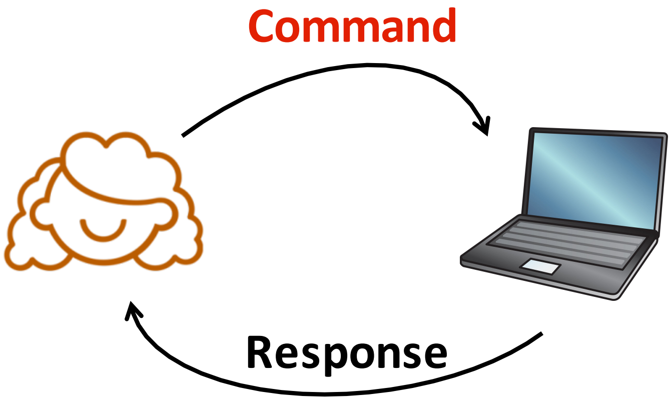
(Human as Data Processor, 13 papers)



Crowdsourced Operators

10. Human-powered Sorts and Joins
11. CrowdER: Crowdsourcing Entity Resolution
12. Leveraging Transitive Relationships for Crowdsourced Joins
13. Using the crowd for top-k and group-by queries

Part 2: Interactive Analytics
(Human as Data Scientist, 17 papers)



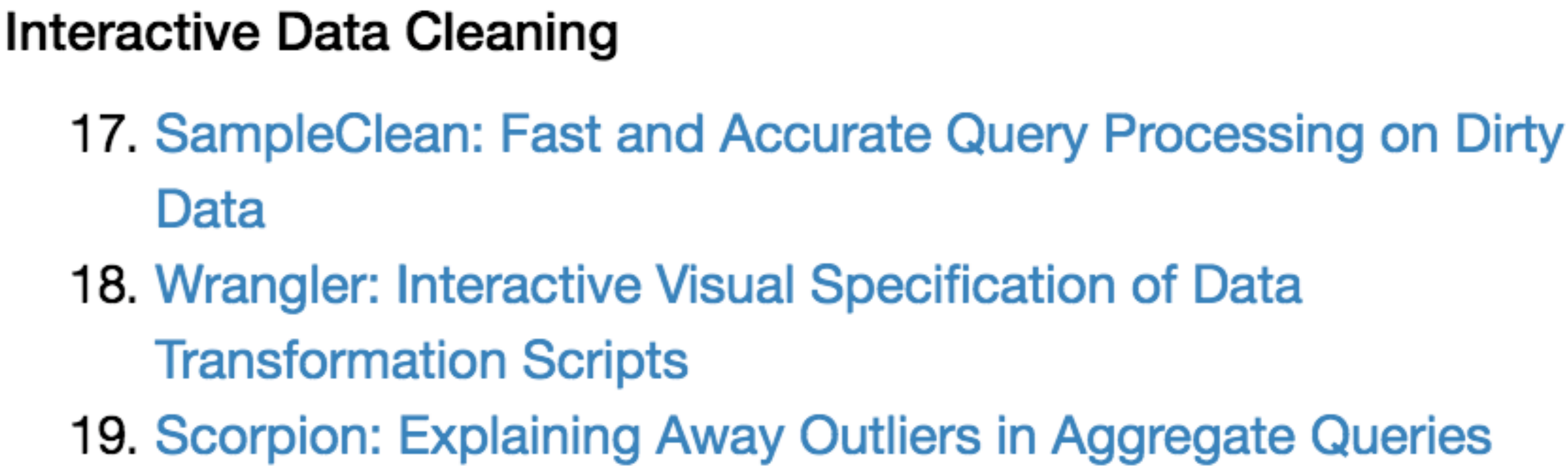
Interactive **Data Cleaning**
Interactive **Visualization**
Interactive **Machine Learning**
Interactive **SQL Analytics**

Part 2: Interactive Analytics
(Human as Data Scientist, 17 papers)

Background

14. Enterprise data analysis and visualization: An interview study
15. The Emerging Role of Data Scientists on Software Development Teams
16. IPython: A System for Interactive Scientific Computing

Part 2: Interactive Analytics
(Human as Data Scientist, 17 papers)



Interactive Data Cleaning

17. SampleClean: Fast and Accurate Query Processing on Dirty Data
18. Wrangler: Interactive Visual Specification of Data Transformation Scripts
19. Scorpion: Explaining Away Outliers in Aggregate Queries

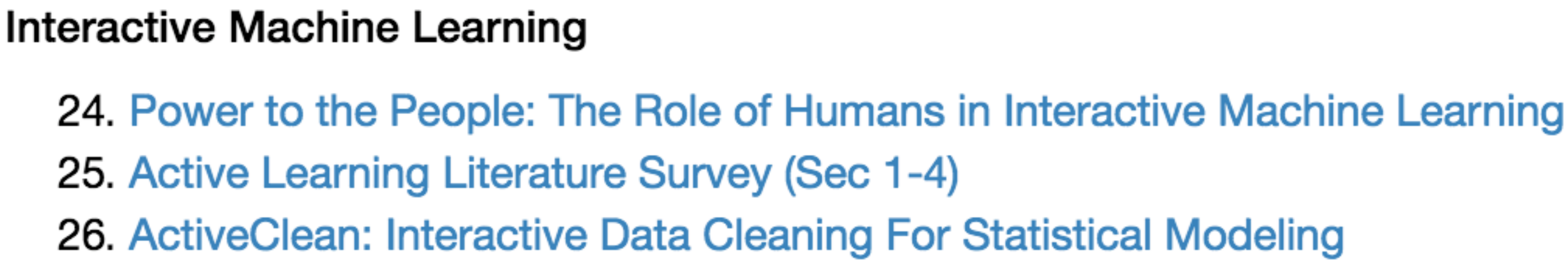
Part 2: Interactive Analytics
(Human as Data Scientist, 17 papers)

Interactive Visualization

20. Polaris: A System for Query, Analysis, and Visualization of Multidimensional Relational Databases
21. Prefuse: a toolkit for interactive information visualization
22. SEEDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics
23. imMens: Real-time Visual Querying of Big Data

Part 2: Interactive Analytics

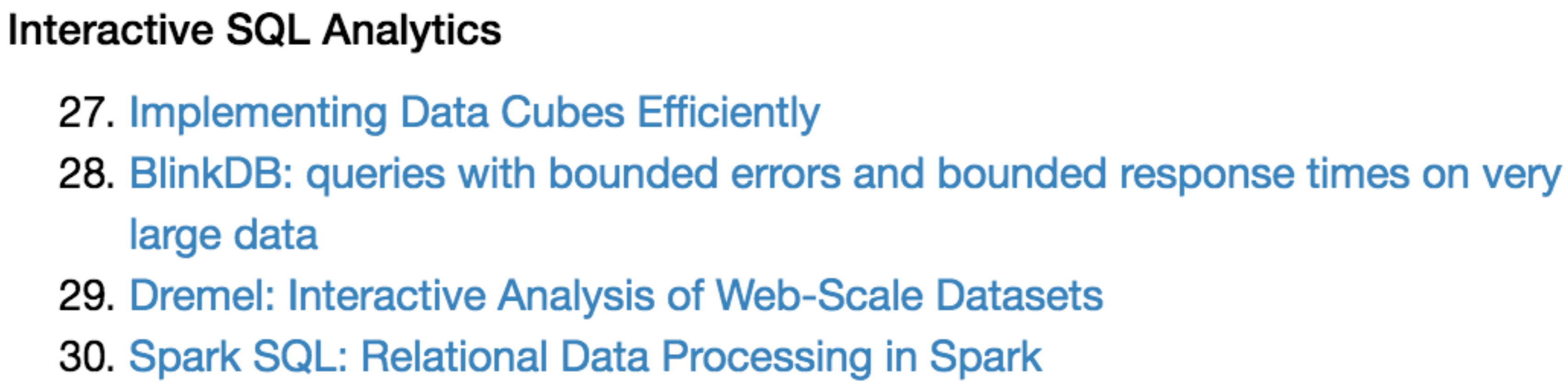
(Human as Data Scientist, 17 papers)



Interactive Machine Learning

24. Power to the People: The Role of Humans in Interactive Machine Learning
25. Active Learning Literature Survey (Sec 1-4)
26. ActiveClean: Interactive Data Cleaning For Statistical Modeling

Part 2: Interactive Analytics
(Human as Data Scientist, 17 papers)



Interactive SQL Analytics

27. Implementing Data Cubes Efficiently
28. BlinkDB: queries with bounded errors and bounded response times on very large data
29. Dremel: Interactive Analysis of Web-Scale Datasets
30. Spark SQL: Relational Data Processing in Spark

Course Objectives

1. Introducing students **the cutting-edge research on Human-in-the-loop Data Management**

2. Training students to master **basic skills for being a researcher**

Skills

Reading Papers

Giving Talks

Reviewing Papers

Asking Questions

How you will be trained

Reading 27+3 Papers

- A quick scan of 27 papers
- A virtual reimplementation of 3 papers

Giving 1 Talk

- Choosing 1 paper to present (35min+15 min Q&A)

Writing 2 reviews

- One from Part 1 and the other from Part 2

Asking 10 Questions

- Asking at least 10 questions in the Q&A sessions

Grading

Paper Presentation: 25%

Questions: 10%

Paper Review: 15%

Assignments: 15%

Final Project: 35% (5% proposal + 10% presentation + 20% report)

What’s next

Fill in the form by the end of Sunday 9/11

https://goo.gl/forms/FPEXVnosd00CCpDj1