INFO 290T

Human-Centered Data Management

Today's Agenda

- The essentials
- Bird's eye view of the class material
- Getting to know you

The Essentials I

- Instructor: Aditya Parameswaran
- Associate Professor, I School and EECS
- Office: 212 South Hall
- Email: adityagp@berkeley.edu
  - **Mention "INFO290" in the title!** (will help ensure a response)
  - Don't expect a quick response
- Meeting slots: Tu/Th 9.30-11am @ 210 South Hall
  - Class slides will be posted after class
- Office hours: Th 1-2pm @ 212 South Hall
  - Also on demand if you have a conflict

R
RIGGS
BEER COMPANY

The Essentials II

- Reader: Chanwut (Mick) Kittivorawong
- PhD Student, EECS
- Email: chanwutk@berkeley.edu
  - **Mention "INFO290" in the title!** (will help ensure a response)
- Office Hours: TBA
- Website: https://info290.github.io

The Essentials: Prerequisites

- *"Students taking the class should have taken a database or data engineering class, at the level of INFO 258 / DATA 101 / COMPSCI 186, and/or have experience working with database or data engineering tools."*
- That said, I am happy to have people from different backgrounds
  - Talk to me if you're not sure
  - Be prepared to read up on concepts outside of class
- The class may not be for you if you have never heard of these terms:
  - Materialized view maintenance
  - Join reordering
  - Nested queries
- The class may not be for you if you have limited programming experience, since you will be undertaking a research project

What sort of course is this?

- **Graduate paper reading class**
  - Reading and discussing research papers from the top venues in data management and HCI/visualization

- A focus on
  - concepts, rather than tools (which change)
  - trade-offs, often no clear "winners"
  - critical evaluation

Goals of the Class

1. Read and critically evaluate research papers in data management and HCI/visualization
2. State-of-the-art techniques and understanding across multiple aspects:
   - Modern Human-Data **Interfaces**
   - **Scalability** techniques focused on humans
   - **Human factors**: personas, behavior, constraints, perceptual limits
3. Process of design, development, validation, and evaluation of ideas in this space

Please Beware of Hiccups!

- Four challenges:
  - A new class
  - A class on new unpolished research material (papers!)
  - A new role-playing format
  - The first time I'm teaching in two years!
- So expect problems to arise! (And please be tolerant!)

- Please feel free to provide feedback to Mick and/or me at any time
- From our end, we will be as tolerant as we can

Grading Breakdown

- Class Participation & Presentation = 50%
  - Paper Reviews = 15%
  - Class Participation = 10%
  - Paper Presentation (Role Playing) = 25%
- Research Project = 50%

Paper Reviews (15%)

- Due day before class at midnight

- Must submit >10 reviews through the semester
- Lightly graded
- Can't "double-dip" with presentations
- Must cover the following aspects:
  - what is the problem?
  - why is it important?
  - what sets it apart from previous work?
  - what are the key technical ideas?
  - what are the main areas of improvement and open issues?
- Limit: 500 words (not strictly enforced)

Class Participation (10%)

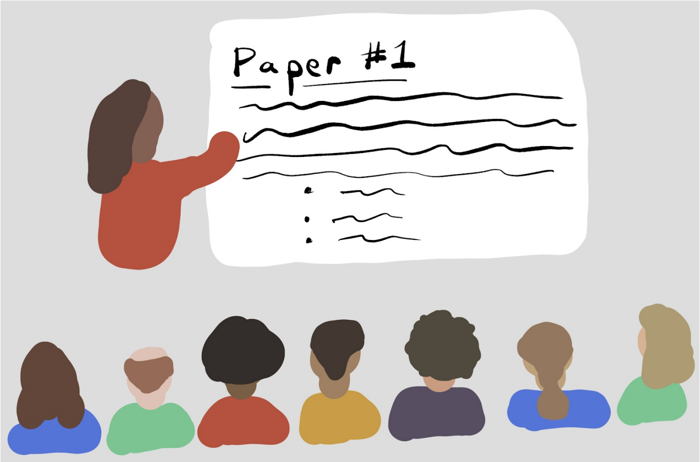
- Goals: assess understanding, get feedback, make the class more lively
- Not essential that
    - you ask "good" questions
    - answer questions "correctly"
    - make "intelligent" points
    - you have to attend every class
    - you have to ask a question every class
- Any participation is good participation!

- Hard to come up with a rubric for class participation, but I promise to not be dogmatic about this, nor grade strictly…
- The goal is to just ensure that you are engaging with the class material even when you're not one of the presenters

Role Playing Activity (25%)

- Pioneered by Colin Raffel (UNC) and Alec Jacobson (Toronto)
    - Our adaptation: Kexin Rong (Georgia Tech)

- Flipping class format from one-to-many to many-to-many
- Each class has multiple presenters, embodying various roles



Role Playing

Overall, over the semester:

- 1-2 paper author roles
  - 15-20 minute presentation
- 2-5 accessory roles
  - 5 minute presentation each

A given person can only play one role per class

Your role is primarily for your presentation, rather than for the subsequent discussion

- Though you may continue to embody the same role throughout if you'd like!

Role: Paper Author

- 15-20 minutes (~1 slide per minute)
  - ~15 minutes if one presenter
  - ~20 minutes if two presenters
- Imagine you are the author of the paper who is presenting your work at a conference.
- In your talk, you should probably address the following:
  - Why should people care about your work?
  - What are the key technical challenges and solutions?
  - How did you evaluate your hypothesis?
  - What are the main takeaways?
- Can reference authors' slides, but don't use them without modification!

Accessory roles x 4

5 min

always start with a one-slide summary

Roles available

peer reviewer (1-2)
archaeologist
academic researcher
industry practitioner

Role: Peer Reviewer (1-2)

- The paper has not been published yet and is currently submitted to a top conference where you've been assigned as a peer reviewer.
- Complete a full review of the paper based on prompts of the official review form of the top venue in this research area (e.g.,VLDB, SIGMOD, CHI,VIS, …):
  - Overall evaluation: {Accept, Weak Accept, Weak Reject, Reject}
  - Summary of contribution
  - Describe in detail all strong points, labeled S1, S2, S3, etc.
  - Describe in detail all opportunities for improvement, labeled O1, O2, O3, etc.

Role: Archaeologist

This paper was found buried under ground in the desert. You're an archeologist who must determine where this paper sits in the context of previous and subsequent work.

Find and report on one older paper cited within the current paper that **substantially influenced** the current paper and one newer paper that **substantially builds** on this current paper.

Role: Academic Researcher

You’re a researcher who is working on a new project in this area.

Propose an imaginary follow-up project not just based on the current paper but only possible due to the existence and success of the current paper.

Could be the start of your own project!

Role: Industry Practitioner

You work at a company or organization developing an application or product of your choice.

Bring a convincing pitch for why you should be paid to implement the method in the paper, and discuss at least one positive and negative impact of this application. Integration and adoption challenges welcome!

Grading System

Credit system

every 5min presentation = 1 credit = 5% of grade
paper author role: 2~3 credits
(i.e., three if solo, two if paired up)
accessory roles: 1 credit

Rules

need >=5 credits over the semester
need to be in the paper author role at least once
paper author presenters for the first 3 papers get 1 extra credit
accessory role presenters for the first 3 papers get 0.5 extra credit
(maximum 1 extra credit per person)

Examples

1 solo paper author role + 2 accessory roles
1 shared paper author role in the first 3 papers + 2 accessory roles
1 shared paper author role + 3 accessory roles

Process for Signing Up

- We will send the link soon, along with instructions …

Grading Breakdown

- Class Participation & Presentation = 50%
    - Paper Reviews = 15%
    - Class Participation = 10%
    - Paper Presentation (Role Playing) = 25%
- Research Project = 50%

Research Project

- 50% of grade
- Main criteria:
  - Build/design/test something new and cool!
  - Also relevant to class topics
- Project milestones
  - 09/14: project proposal 5%
  - 10/17: intermediate report 10%
  - 11/28: final report and presentation 35%

Research Project Goals

- Ideally: something "publishable" (ish) for a data management or HCI venue
- Spectrum of contributions:
  - Mainly algorithmic
  - Mainly system or tool-building oriented
  - Mainly interface or human factors
  - Or all of them
- Important to build on state of the art and extend it
  - Simply reproducing existing work not sufficient (unless it is a new benchmark)
  - Evaluating why it is "better" than prior work is needed
    - Usability, Performance, Accuracy, Expressiveness (or multiple)
  - Getting your hands dirty is a must!
- At the very least, we'd want you to ***develop something new and evaluate it***

Usual Process

- Rough phases:
  - Identify problem (we can help!)
  - Explore related work/tools
  - Prototype
  - Build & Evaluate
  - Write "paper"

Examples of Research Projects

- An new touch-based interface for dataset search and discovery
- A spreadsheet-notebook hybrid
- An approximate query processing engine for Matplotlib/Vega-lite
- A tool to visualize differences between two versions of the dataset
- A human-in-the-loop tool to convert json to relations
- A "model browser" that helps explore areas where an ML model is doing well vs. not
- Starting from a new application: e.g., a tool for exploring large genomics datasets
- Extending an existing algorithm to a new setting or domain: e.g., building a search engine for legal documents

Many public data and workflows

- Kaggle
- UCI Machine Learning Repository
- Github
- StackOverflow
- OpenAIRE Research Graph
- Data.gov
- Data World

Talk to us!

One “Weird Trick” to Find a Research Project

- Take a data processing problem:
  - Data cleaning, data extraction, data transformation, data visualization, data manipulation, data discovery & search, machine learning, …
- Take the state of the art approach for it
- And see if you can get “close” with LLMs!
  - Either out of the box or fine-tuned
- Will need to think about:
  - Error handling, guardrails, hallucination, ambiguity
  - Whether chat is indeed the right interface always
  - “Multiple iterations” with humans of confirmation, validation, …
  - Incorporating context and domain-knowledge

Different flavors of projects

Benchmarking/new datasets/user study

    Show: new insights and understanding

Tool/system/interface

    Show: ease of use, scalability, design novelty

Algorithm

    Show: novelty, correctness, scalability

Reproduce and extend

    Show: assumptions/contexts that have changed

Medical Disruptions

- We are entering a Fall/Winter season of Flu, COVID, …

- We don’t require you to attend every class
- If you’re unwell and aren’t one of the presenters, it’s *OK* to skip class to safeguard others
  - If you are ending up skipping a lot of classes, please talk to us
- If you are slated to present, let us know ASAP
  - Unless in case of sudden medical emergencies, we will need 24+ hour notice to make alternate plans
  - If you miss a presentation slot with no advance notice, you will receive negative credit

Any Questions?

- Getting to know you
- An overview of class content

Getting to know you

- Your name & preferred pronouns?
- Which department/program are you in?
- Fun fact about you?
- What are you hoping to get from the class?

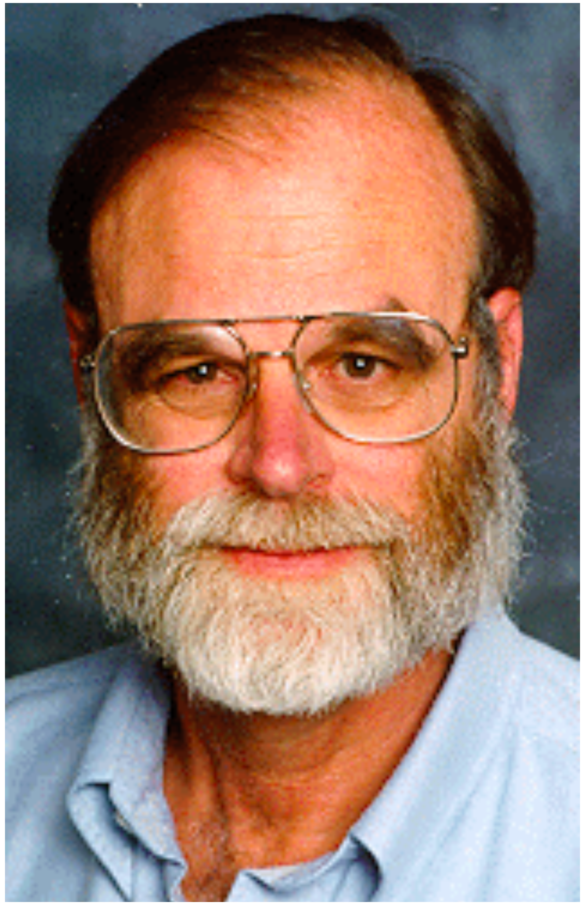
A Brief History of Data Management

- The field of data management began in the 60s with "database systems"…
- Really hit its stride in the 70s thanks to a clean new "data model" – a way of representing and thinking about data
  - The relational model: essentially a model of relations (or sets)
  - Led to the field of relational databases, with several commercial and open-source database systems
- Led to several software systems "powered" by relational databases
  - Banking systems
  - Flight reservation systems
  - Payment systems
- Relational databases are still at the core of the technology stack of most companies today.
- This includes the so-called *modern data stack*

Classical Data Management Systems: A Brief Berkeley Side Note…

The last two Turing awards for databases came from Berkeley!

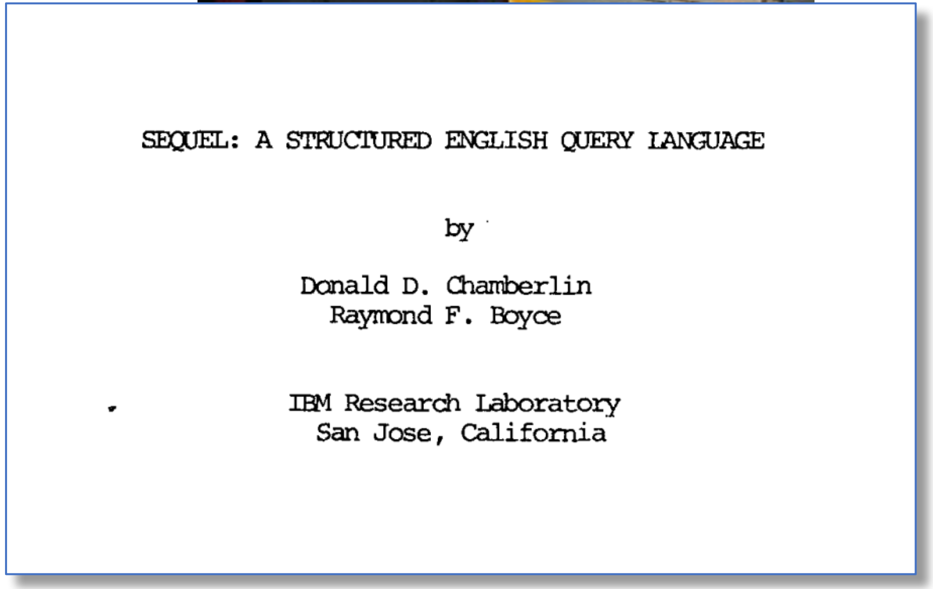
- Jim Gray (‘98) was the first CS PhD student from Berkeley; pioneer of transaction processing
- Mike Stonebraker (‘14) was faculty for many years at Berkeley; influenced or developed many popular database systems



SQL: Intergalactic Dataspeak

- The language spoken by relational databases
- Don Chamberlin: designed SQL with a "walk up and read it" property

- NoSQL came and went, but SQL stayed
- Truly stood the test of time



SEQUEL: A STRUCTURED ENGLISH QUERY LANGUAGE

by

Donald D. Chamberlin
Raymond F. Boyce

IBM Research Laboratory
San Jose, California

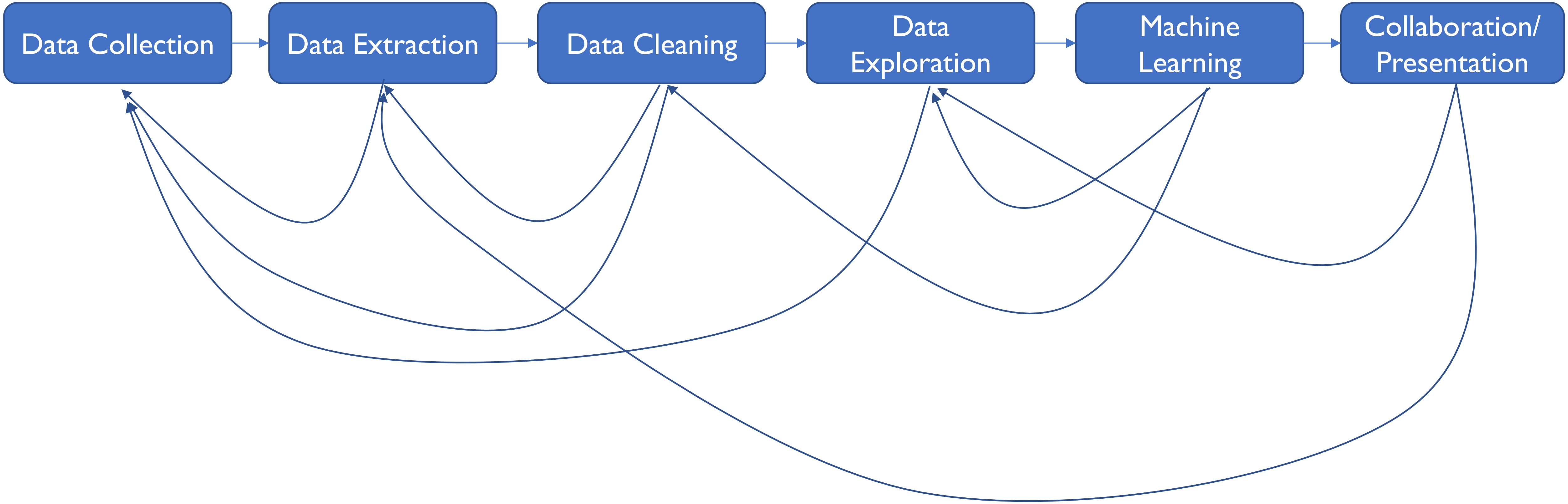
- Despite the readability intent, SQL has its limitations in terms of usability …

Today…

SQL and Relational Databases are still important …

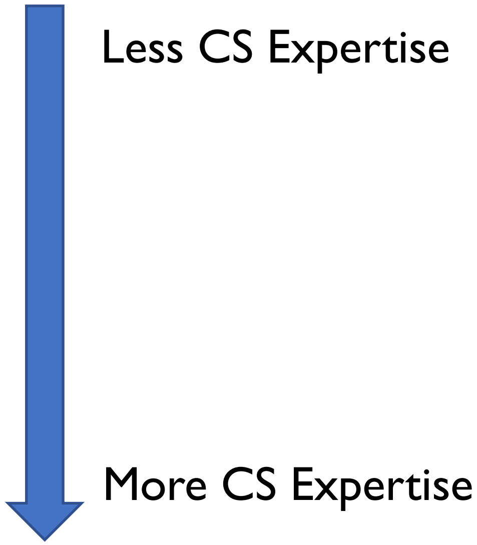
But form a small fraction of "data work"

Modern Data Lifecycle



Modern Data Roles

- Data Consumers
- Business Analysts
- Data Analysts
- Data Scientists
- Data/ML Engineer



Less CS Expertise

More CS Expertise

Tools used also vary by role!

A “Human-Centered” Perspective

Why the fuss about humans?

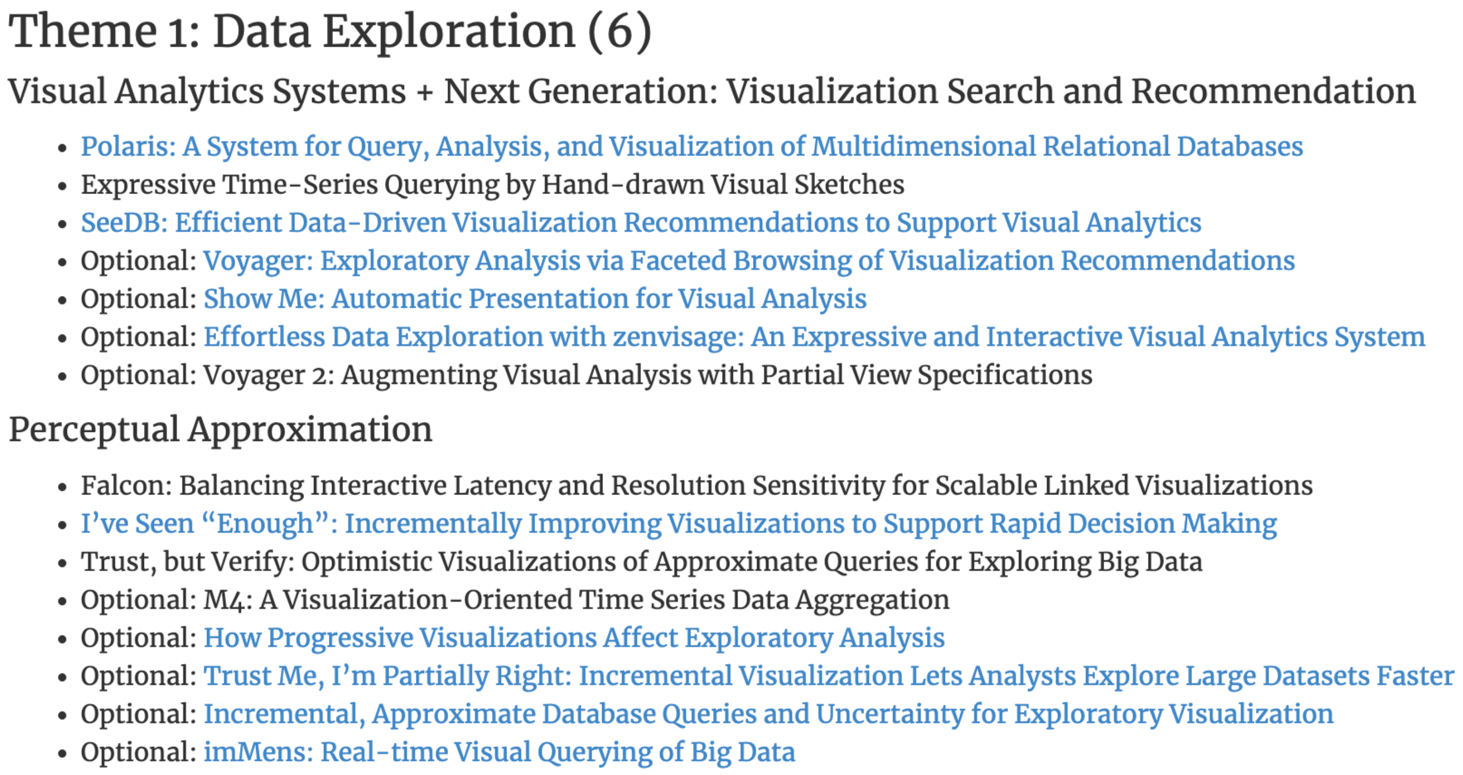
As we consider (Lifecycle stage x Role) — developing the next generation of tools requires human-centered design

- Humans are the ones doing the data work
- Reasoning about their needs and constraints is crucial
- Traditional data management research and practice ignores the human aspects!

Theme 1: Data Exploration

Visual analytics systems are a huge portion of the data analysis market: Tableau, PowerBI, …

How do we make visual analytics systems (and visual data exploration) easier and faster?



Theme 1: Data Exploration (6)

Visual Analytics Systems + Next Generation: Visualization Search and Recommendation

- Polaris: A System for Query, Analysis, and Visualization of Multidimensional Relational Databases
- Expressive Time-Series Querying by Hand-drawn Visual Sketches
- SeeDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics
- Optional: Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations
- Optional: Show Me: Automatic Presentation for Visual Analysis
- Optional: Effortless Data Exploration with zenvisage: An Expressive and Interactive Visual Analytics System
- Optional: Voyager 2: Augmenting Visual Analysis with Partial View Specifications

Perceptual Approximation

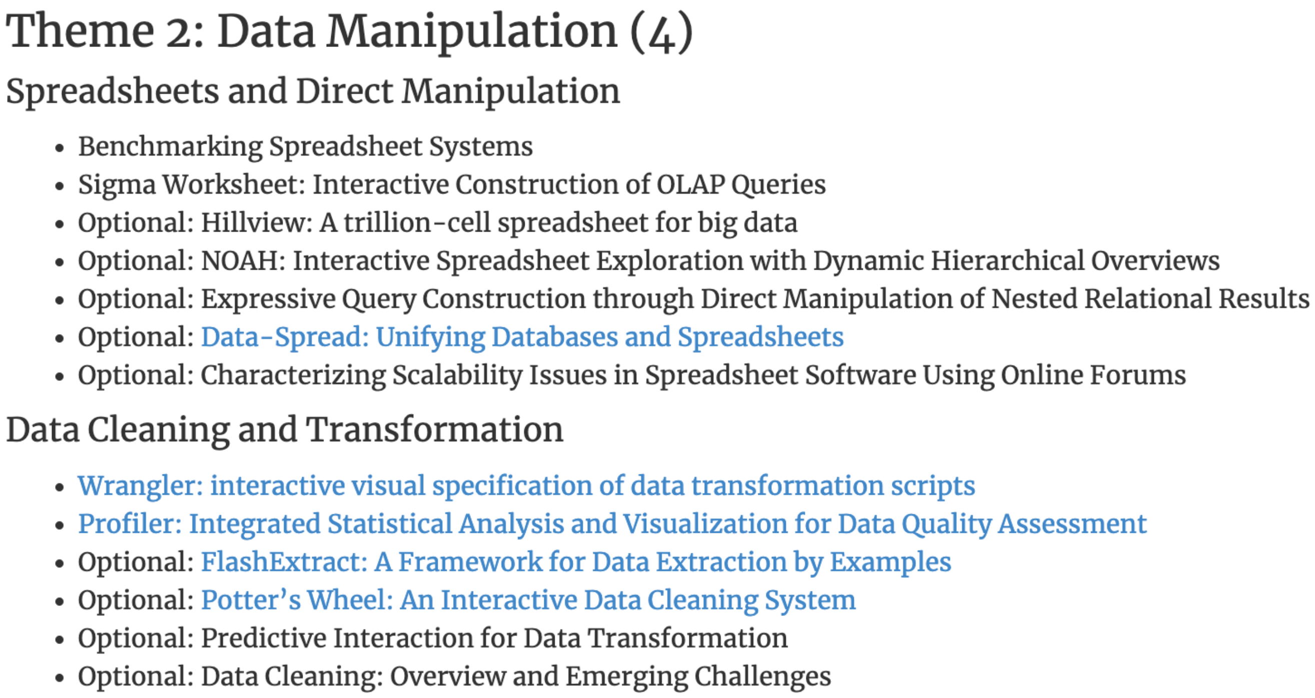
- Falcon: Balancing Interactive Latency and Resolution Sensitivity for Scalable Linked Visualizations
- I've Seen "Enough": Incrementally Improving Visualizations to Support Rapid Decision Making
- Trust, but Verify: Optimistic Visualizations of Approximate Queries for Exploring Big Data
- Optional: M4: A Visualization-Oriented Time Series Data Aggregation
- Optional: How Progressive Visualizations Affect Exploratory Analysis
- Optional: Trust Me, I'm Partially Right: Incremental Visualization Lets Analysts Explore Large Datasets Faster
- Optional: Incremental, Approximate Database Queries and Uncertainty for Exploratory Visualization
- Optional: imMens: Real-time Visual Querying of Big Data

Theme 2: Data Manipulation

Spreadsheets: used by 20% of the world's population

+

Data prep & cleaning tools: prior to + during analysis



Theme 2: Data Manipulation (4)

Spreadsheets and Direct Manipulation

- Benchmarking Spreadsheet Systems
- Sigma Worksheet: Interactive Construction of OLAP Queries
- Optional: Hillview: A trillion-cell spreadsheet for big data
- Optional: NOAH: Interactive Spreadsheet Exploration with Dynamic Hierarchical Overviews
- Optional: Expressive Query Construction through Direct Manipulation of Nested Relational Results
- Optional: Data-Spread: Unifying Databases and Spreadsheets
- Optional: Characterizing Scalability Issues in Spreadsheet Software Using Online Forums

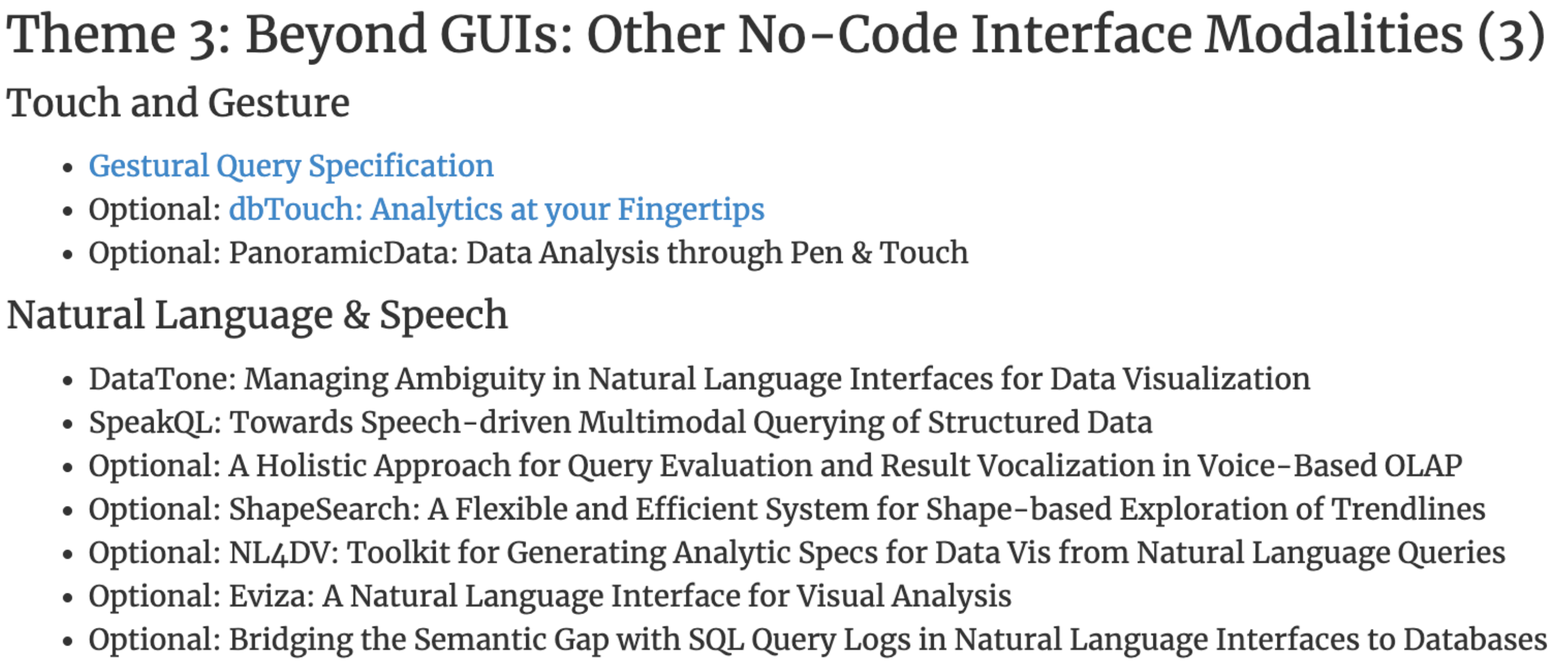
Data Cleaning and Transformation

- Wrangler: interactive visual specification of data transformation scripts
- Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment
- Optional: FlashExtract: A Framework for Data Extraction by Examples
- Optional: Potter's Wheel: An Interactive Data Cleaning System
- Optional: Predictive Interaction for Data Transformation
- Optional: Data Cleaning: Overview and Emerging Challenges

Theme 3: Other Interface Modalities

GUIs are not always better

Especially for non-programmers



Theme 3: Beyond GUIs: Other No-Code Interface Modalities (3)

Touch and Gesture

- Gestural Query Specification
- Optional: dbTouch: Analytics at your Fingertips
- Optional: PanoramicData: Data Analysis through Pen & Touch

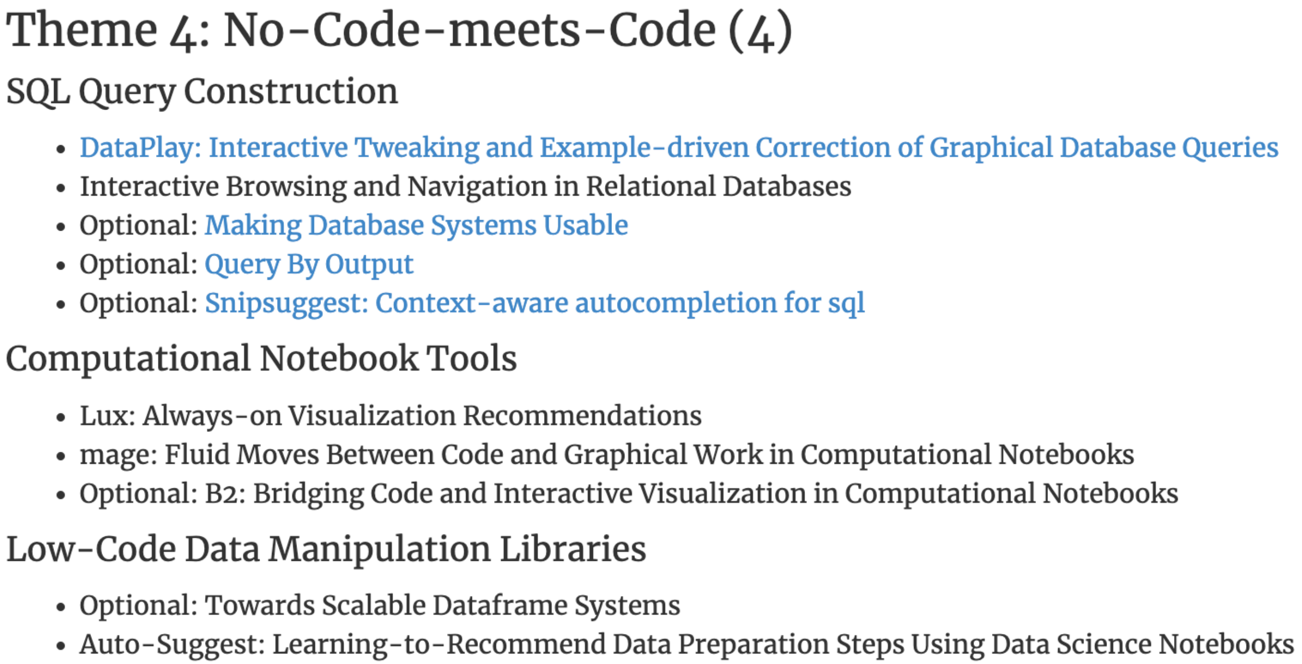
Natural Language & Speech

- DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization
- SpeakQL: Towards Speech-driven Multimodal Querying of Structured Data
- Optional: A Holistic Approach for Query Evaluation and Result Vocalization in Voice-Based OLAP
- Optional: ShapeSearch: A Flexible and Efficient System for Shape-based Exploration of Trendlines
- Optional: NL4DV: Toolkit for Generating Analytic Specs for Data Vis from Natural Language Queries
- Optional: Eviza: A Natural Language Interface for Visual Analysis
- Optional: Bridging the Semantic Gap with SQL Query Logs in Natural Language Interfaces to Databases

Theme 3: No-Code Meets Code

Some roles (eg business analysts) can write code — but not proficiently

Can we get best of code-based and interactive modalities?



Theme 4: No-Code-meets-Code (4)

SQL Query Construction

- DataPlay: Interactive Tweaking and Example-driven Correction of Graphical Database Queries
- Interactive Browsing and Navigation in Relational Databases
- Optional: Making Database Systems Usable
- Optional: Query By Output
- Optional: Snipsuggest: Context-aware autocompletion for sql

Computational Notebook Tools

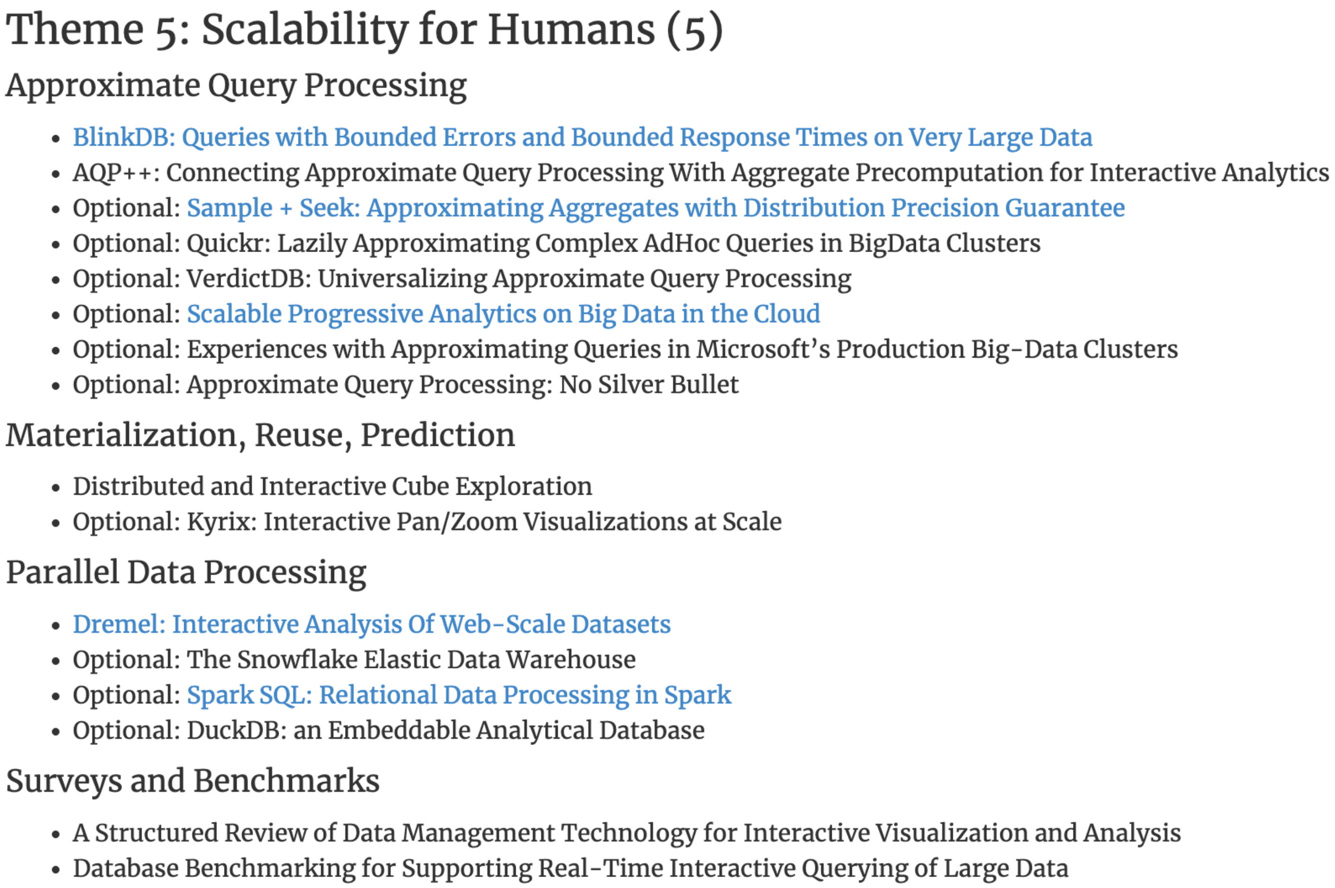
- Lux: Always-on Visualization Recommendations
- mage: Fluid Moves Between Code and Graphical Work in Computational Notebooks
- Optional: B2: Bridging Code and Interactive Visualization in Computational Notebooks

Low-Code Data Manipulation Libraries

- Optional: Towards Scalable Dataframe Systems
- Auto-Suggest: Learning-to-Recommend Data Preparation Steps Using Data Science Notebooks

Theme 5: Scalability for Humans

A variety of scalability techniques targeted at data analysis



Theme 5: Scalability for Humans (5)

Approximate Query Processing

- BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data
- AQP++: Connecting Approximate Query Processing With Aggregate Precomputation for Interactive Analytics
- Optional: Sample + Seek: Approximating Aggregates with Distribution Precision Guarantee
- Optional: Quickr: Lazily Approximating Complex AdHoc Queries in BigData Clusters
- Optional: VerdictDB: Universalizing Approximate Query Processing
- Optional: Scalable Progressive Analytics on Big Data in the Cloud
- Optional: Experiences with Approximating Queries in Microsoft's Production Big-Data Clusters
- Optional: Approximate Query Processing: No Silver Bullet

Materialization, Reuse, Prediction

- Distributed and Interactive Cube Exploration
- Optional: Kyrix: Interactive Pan/Zoom Visualizations at Scale

Parallel Data Processing

- Dremel: Interactive Analysis Of Web-Scale Datasets
- Optional: The Snowflake Elastic Data Warehouse
- Optional: Spark SQL: Relational Data Processing in Spark
- Optional: DuckDB: an Embeddable Analytical Database

Surveys and Benchmarks

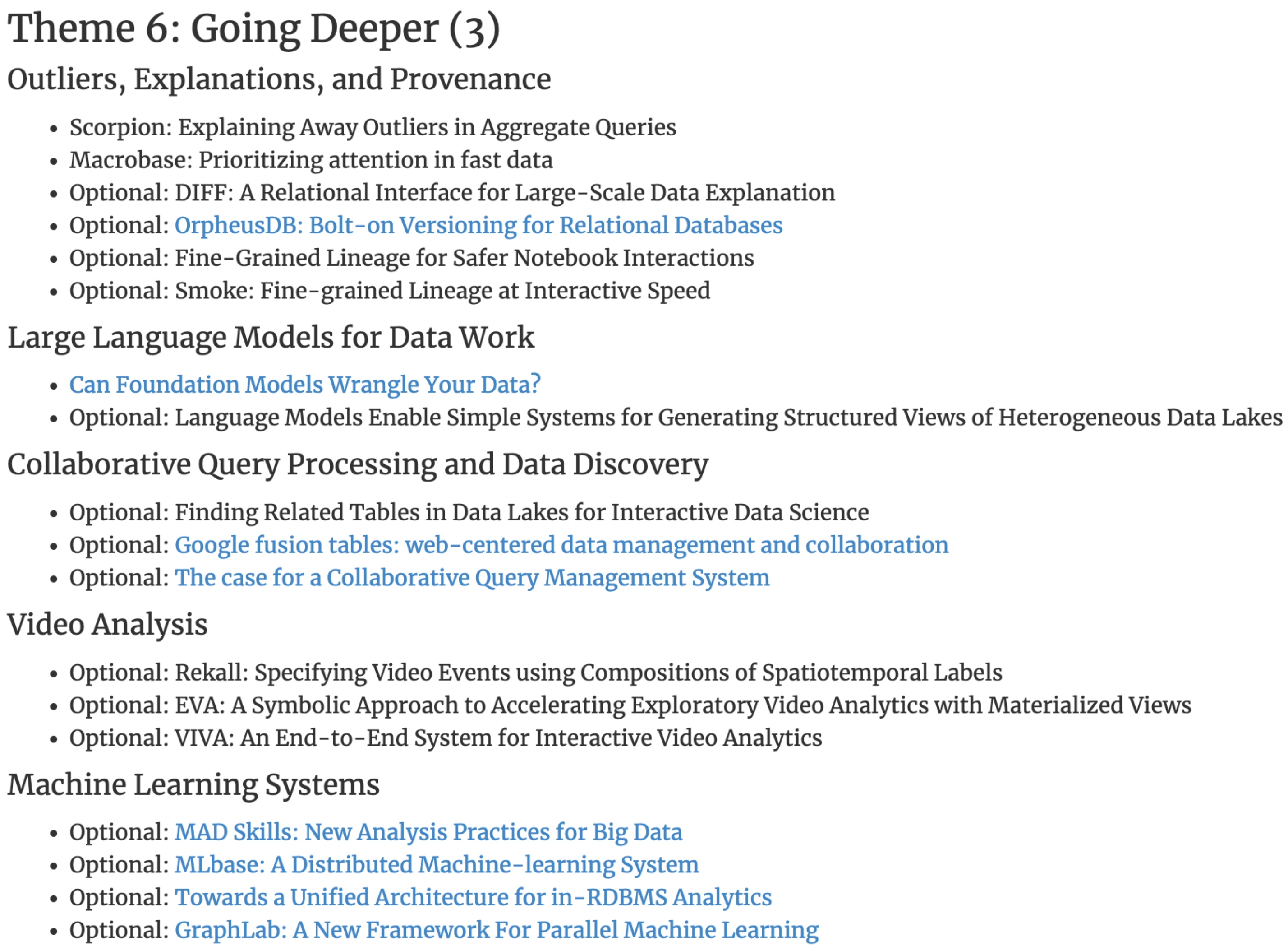
- A Structured Review of Data Management Technology for Interactive Visualization and Analysis
- Database Benchmarking for Supporting Real-Time Interactive Querying of Large Data

Theme 6: Going Deeper

Asking deeper (causal) questions

Working with text and video

Sharing results with others



Theme 6: Going Deeper (3)

Outliers, Explanations, and Provenance

- Scorpion: Explaining Away Outliers in Aggregate Queries
- Macrobase: Prioritizing attention in fast data
- Optional: DIFF: A Relational Interface for Large-Scale Data Explanation
- Optional: OrpheusDB: Bolt-on Versioning for Relational Databases
- Optional: Fine-Grained Lineage for Safer Notebook Interactions
- Optional: Smoke: Fine-grained Lineage at Interactive Speed

Large Language Models for Data Work

- Can Foundation Models Wrangle Your Data?
- Optional: Language Models Enable Simple Systems for Generating Structured Views of Heterogeneous Data Lakes

Collaborative Query Processing and Data Discovery

- Optional: Finding Related Tables in Data Lakes for Interactive Data Science
- Optional: Google fusion tables: web-centered data management and collaboration
- Optional: The case for a Collaborative Query Management System

Video Analysis

- Optional: Rekall: Specifying Video Events using Compositions of Spatiotemporal Labels
- Optional: EVA: A Symbolic Approach to Accelerating Exploratory Video Analytics with Materialized Views
- Optional: VIVA: An End-to-End System for Interactive Video Analytics

Machine Learning Systems

- Optional: MAD Skills: New Analysis Practices for Big Data
- Optional: MLbase: A Distributed Machine-learning System
- Optional: Towards a Unified Architecture for in-RDBMS Analytics
- Optional: GraphLab: A New Framework For Parallel Machine Learning

Goals of the Class: Recap

1. Read and critically evaluate research papers in data management and HCI/visualization
2. State-of-the-art techniques and understanding across multiple aspects:
   - Modern Human-data **Interfaces**
   - **Scalability** techniques focused on humans
   - **Human factors**: personas, behavior, constraints, perceptual limits
3. Process of design, development, validation, and evaluation of ideas in this space

Any Questions?

- Getting to know you
- An overview of class content