CIS 602-01: Scalable Data Analysis

Reproducibility

Dr. David Koop
## Analyzing Text: Common words in *Tom Sawyer*

<table>
<thead>
<tr>
<th>Word</th>
<th>Freq.</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>3332</td>
<td>determiner (article)</td>
</tr>
<tr>
<td>and</td>
<td>2972</td>
<td>conjunction</td>
</tr>
<tr>
<td>a</td>
<td>1775</td>
<td>determiner</td>
</tr>
<tr>
<td>to</td>
<td>1725</td>
<td>preposition, verbal infinitive marker</td>
</tr>
<tr>
<td>of</td>
<td>1440</td>
<td>preposition</td>
</tr>
<tr>
<td>was</td>
<td>1161</td>
<td>auxiliary verb</td>
</tr>
<tr>
<td>it</td>
<td>1027</td>
<td>(personal/expletive) pronoun</td>
</tr>
<tr>
<td>in</td>
<td>906</td>
<td>preposition</td>
</tr>
<tr>
<td>that</td>
<td>877</td>
<td>complementizer, demonstrative</td>
</tr>
<tr>
<td>he</td>
<td>877</td>
<td>(personal) pronoun</td>
</tr>
<tr>
<td>l</td>
<td>783</td>
<td>(personal) pronoun</td>
</tr>
<tr>
<td>his</td>
<td>772</td>
<td>(possessive) pronoun</td>
</tr>
<tr>
<td>you</td>
<td>686</td>
<td>(personal) pronoun</td>
</tr>
<tr>
<td>Tom</td>
<td>679</td>
<td>proper noun</td>
</tr>
<tr>
<td>with</td>
<td>642</td>
<td>preposition</td>
</tr>
</tbody>
</table>
### Zipf's Law in *Tom Sawyer*

<table>
<thead>
<tr>
<th>Word</th>
<th>Freq. ($f$)</th>
<th>Rank ($r$)</th>
<th>$f \cdot r$</th>
<th>Word</th>
<th>Freq. ($f$)</th>
<th>Rank ($r$)</th>
<th>$f \cdot r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>3332</td>
<td>1</td>
<td>3332</td>
<td>turned</td>
<td>51</td>
<td>200</td>
<td>10200</td>
</tr>
<tr>
<td>and</td>
<td>2972</td>
<td>2</td>
<td>5944</td>
<td>you’ll</td>
<td>30</td>
<td>300</td>
<td>9000</td>
</tr>
<tr>
<td>a</td>
<td>1775</td>
<td>3</td>
<td>5235</td>
<td>name</td>
<td>21</td>
<td>400</td>
<td>8400</td>
</tr>
<tr>
<td>he</td>
<td>877</td>
<td>10</td>
<td>8770</td>
<td>comes</td>
<td>16</td>
<td>500</td>
<td>8000</td>
</tr>
<tr>
<td>but</td>
<td>410</td>
<td>20</td>
<td>8400</td>
<td>group</td>
<td>13</td>
<td>600</td>
<td>7800</td>
</tr>
<tr>
<td>be</td>
<td>294</td>
<td>30</td>
<td>8820</td>
<td>lead</td>
<td>11</td>
<td>700</td>
<td>7700</td>
</tr>
<tr>
<td>there</td>
<td>222</td>
<td>40</td>
<td>8880</td>
<td>friends</td>
<td>10</td>
<td>800</td>
<td>8000</td>
</tr>
<tr>
<td>one</td>
<td>172</td>
<td>50</td>
<td>8600</td>
<td>begin</td>
<td>9</td>
<td>900</td>
<td>8100</td>
</tr>
<tr>
<td>about</td>
<td>158</td>
<td>60</td>
<td>9480</td>
<td>family</td>
<td>8</td>
<td>1000</td>
<td>8000</td>
</tr>
<tr>
<td>more</td>
<td>138</td>
<td>70</td>
<td>9660</td>
<td>brushed</td>
<td>4</td>
<td>2000</td>
<td>8000</td>
</tr>
<tr>
<td>never</td>
<td>124</td>
<td>80</td>
<td>9920</td>
<td>sins</td>
<td>2</td>
<td>3000</td>
<td>6000</td>
</tr>
<tr>
<td>Oh</td>
<td>116</td>
<td>90</td>
<td>10440</td>
<td>Could</td>
<td>2</td>
<td>4000</td>
<td>8000</td>
</tr>
<tr>
<td>two</td>
<td>104</td>
<td>100</td>
<td>10400</td>
<td>Applausive</td>
<td>1</td>
<td>8000</td>
<td>8000</td>
</tr>
</tbody>
</table>
Bigrams in *New York Times*

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Word 1</th>
<th>Word 2</th>
<th>POS pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>11487</td>
<td>New</td>
<td>York</td>
<td>A N</td>
</tr>
<tr>
<td>7261</td>
<td>United</td>
<td>States</td>
<td>A N</td>
</tr>
<tr>
<td>5412</td>
<td>Los</td>
<td>Angeles</td>
<td>N N</td>
</tr>
<tr>
<td>3301</td>
<td>last</td>
<td>year</td>
<td>A N</td>
</tr>
<tr>
<td>3191</td>
<td>Saudi</td>
<td>Arabia</td>
<td>N N</td>
</tr>
<tr>
<td>2699</td>
<td>last</td>
<td>week</td>
<td>A N</td>
</tr>
<tr>
<td>2514</td>
<td>vice</td>
<td>president</td>
<td>A N</td>
</tr>
<tr>
<td>2378</td>
<td>Persian</td>
<td>Gulf</td>
<td>A N</td>
</tr>
<tr>
<td>2161</td>
<td>San</td>
<td>Francisco</td>
<td>N N</td>
</tr>
<tr>
<td>2106</td>
<td>President</td>
<td>Bush</td>
<td>N N</td>
</tr>
<tr>
<td>2001</td>
<td>Middle</td>
<td>East</td>
<td>A N</td>
</tr>
<tr>
<td>1942</td>
<td>Saddam</td>
<td>Hussein</td>
<td>N N</td>
</tr>
<tr>
<td>1867</td>
<td>Soviet</td>
<td>Union</td>
<td>A N</td>
</tr>
<tr>
<td>1850</td>
<td>White</td>
<td>House</td>
<td>A N</td>
</tr>
<tr>
<td>1633</td>
<td>United</td>
<td>Nations</td>
<td>A N</td>
</tr>
<tr>
<td>1337</td>
<td>York</td>
<td>City</td>
<td>N N</td>
</tr>
<tr>
<td>1328</td>
<td>oil</td>
<td>prices</td>
<td>N N</td>
</tr>
<tr>
<td>1210</td>
<td>next</td>
<td>year</td>
<td>A N</td>
</tr>
<tr>
<td>1074</td>
<td>chief</td>
<td>executive</td>
<td>A N</td>
</tr>
<tr>
<td>1073</td>
<td>real</td>
<td>estate</td>
<td>A N</td>
</tr>
</tbody>
</table>
on Many Eyes, for instance, we would not have guessed at the popularity of religious analyses. Given the broad demand for text visualizations, however, it seems like a fruitful area of study.

ACKNOWLEDGEMENTS

The authors thank Frank van Ham, Jesse Kriss, Matt McKeon, Lea Byron, and Eric Gilbert for helpful suggestions. In addition, we are grateful to the users of Many Eyes for their creativity and willingness to provide feedback on an experimental visualization technique.

Fig 10: Word Tree showing all occurrences of "I have a dream" in Martin Luther King's historical speech.

Fig 9. Word tree of the King James Bible showing all occurrences of "love the."

Text Visualization: Word Tree

[Wattenberg & Viegas, 2007]
Positive or negative movie review?

Unbelievably disappointing

Full of zany characters and richly applied satire, and some great plot twists

This is the greatest screwball comedy ever filmed

It was pathetic. The worst part about it was the boxing scenes.
Positive or negative movie review?

👏 Unbelievably disappointing

Full of zany characters and richly applied satire, and some great plot twists

This is the greatest screwball comedy ever filmed

It was pathetic. The worst part about it was the boxing scenes.
Positive or negative movie review?

👍 Unbelievably disappointing

👎 Full of zany characters and richly applied satire, and some great plot twists

   This is the greatest screwball comedy ever filmed

   It was pathetic. The worst part about it was the boxing scenes.
Positive or negative movie review?

👎 Unbelievably disappointing
👍 Full of zany characters and richly applied satire, and some great plot twists
👍 This is the greatest screwball comedy ever filmed

It was pathetic. The worst part about it was the boxing scenes.
Positive or negative movie review?

👍 Unbelievably disappointing

👎 Full of zany characters and richly applied satire, and some great plot twists

👍 This is the greatest screwball comedy ever filmed

👎 It was pathetic. The worst part about it was the boxing scenes.
Twitter Mood Predicts the Stock Market

[J. Bollen et al., 2011]
Twitter Mood Predicts the Stock Market

CALM Predicts Dow Jones Industrial Average 3 days later

[J. Bollen et al., 2011]
Why sentiment analysis?

- Movie: is this review positive or negative?
- Products: what do people think about the new iPhone?
- Public sentiment: how is consumer confidence? Is despair increasing?
- Politics: what do people think about this candidate or issue?
- Prediction: predict election outcomes or market trends from sentiment
Sentiment Analysis Tasks

• Simplest task:
  - Is the attitude of this text positive or negative?

[D. Jurafsky]
Sentiment Analysis Tasks

- Simplest task:
  - Is the attitude of this text positive or negative?

- More complex:
  - Rank the attitude of this text from 1 to 5
Sentiment Analysis Tasks

• Simplest task:
  - Is the attitude of this text positive or negative?

• More complex:
  - Rank the attitude of this text from 1 to 5

• Advanced:
  - Detect the target, source, or complex attitude types
Features for Sentiment Classification

• How to handle negation
  - "I didn’t like this movie"
  - "I really like this movie"
  - Add NOT_ to words after negation and before punctuation
    - didn’t like this movie, but I
      → didn’t NOT_like NOT_this NOT_movie but I

• Which words to use?
  - Only adjectives
  - All words
  - All words turns out to work better on some data
Hard Cases

• Perfume review in "Perfumes: the Guide": “If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut.”

• Dorothy Parker on Katherine Hepburn: “She runs the gamut of emotions from A to B”

• “This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can’t hold up.”

• Well as usual Keanu Reeves is nothing special, but surprisingly, the very talented Laurence Fishbourne is not so good either, I was surprised.

[D. Jurafsky]
Scalable Sentiment Analysis

- Challenges:
  - Different lexicon
    - Twitter is not formal language
    - Existing lexicons don't consider misspellings
  - Volume of tweets
  - Time to process
- Use MapReduce

- Components:
  - Lexicon Builder
  - Sentiment Classifier

[Khuc et al., 2012]
Scalable Sentiment Analysis Architecture
Final Project

- Presentations: Saturday, Dec. 16 (8-11am), Sorry :(  
- Presentation (5-6 minutes)  
  - Brief Introduction to Dataset/Problem (1 minute)  
  - Scalability Challenges (1-2 minutes)  
  - Results (2-3 minutes)  
- Report  
  - Expand proposal into technical report (figures, tables, references)  
- Code  
  - Include README  
  - Source with build/run instructions  
  - Link to data (make subset available to me if not publicly available)
Reproducibility: Gene Names and Excel
Gene names and Excel

• SEPT2 → 9/2/2016, MARCH1 → 3/1/2016
• First cited in 2004
• Blog posts:
  - https://dontuseexcel.wordpress.com/2013/02/07/dont-use-excel-for-biological-data/
• Studied supplemental data from 18 journals, 35,175 Excel files
• Increased at an annual rate of 15% in the past five years
• Not just Excel: e.g. Apache OpenOffice Calc

["Gene name errors are widespread in the scientific literature", M. Ziemann et al., 2016,]
Affected Gene Lists

"Gene name errors are widespread in the scientific literature", M. Ziemann et al., 2016,
One could teach a course on reproducibility.
I did. (Fall 2016)
Reproducibility Overview

• Code Availability, Collaboration, and Version Control
• Data Availability, Citation, and Curation
• Virtual Machines and Containers
• Scientific Workflows
• Provenance
• Tools
• Numerical Reproducibility & Scalability
• Cultural, Ethical, and Legal Challenges
Importance of Reproducibility
Better Understanding of how to do Reproducible Work
Rules for Reproducible Computational Research

• Rule 1: For Every Result, Keep Track of How It Was Produced
• Rule 2: Avoid Manual Data Manipulation Steps
• Rule 3: Archive the Exact Versions of All External Programs Used
• Rule 4: Version Control All Custom Scripts
• Rule 5: Record All Intermediate Results, When Possible in Standardized Formats

[Sandve et al., 2013]
Rules for Reproducible Computational Research

• Rule 6: For Analyses That Include Randomness, Note Underlying Random Seeds
• Rule 7: Always Store Raw Data behind Plots
• Rule 8: Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
• Rule 9: Connect Textual Statements to Underlying Results
• Rule 10: Provide Public Access to Scripts, Runs, and Results

[Sandve et al., 2013]
Scalability Concerns?
Scalability Concerns?

• Rule 1: For Every Result, Keep Track of How It Was Produced
• Rule 2: Avoid Manual Data Manipulation Steps
• Rule 3: Archive the ExactVersions of All External Programs Used
• Rule 4: Version Control All Custom Scripts
• Rule 5: Record All Intermediate Results, When Possible in Standardized Formats

[Sandve et al., 2013]
Scalability Concerns?

- Rule 6: For Analyses That Include Randomness, Note Underlying Random Seeds
- Rule 7: Always Store Raw Data behind Plots
- Rule 8: Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
- Rule 9: Connect Textual Statements to Underlying Results
- Rule 10: Provide Public Access to Scripts, Runs, and Results

[Sandve et al., 2013]
Enabling Reproducibility in Big Data Research

V. Stodden
Reliable Scientific Conclusions in Big Data Era

• "Ubiquity of Error":
  - Mistakes and self-delusion can creep in absolute anywhere
  - Scientist's effort is expended in recognizing and noting that error

• "Computations are frequently of such a complexity that an explanation sufficiently detailed to enable others to replicate the results is not possible in a typical scientific publication."
Scale of Data (LSST)

- Image every 15 seconds
- 100PB over 10 years

[http://www.lsst.org]
Scalable Data Science
Scalability

• “Big Data”
  - What is “big”? For whom is it “big”?
  - variety, velocity, volume, …

• Lots of data that was big is not an issue now

• Understanding the scalability of techniques is important

• There will always be larger datasets, want to understand
  - how methods scale
  - performance bounds
  - storage constraints
Scalable Visualization Challenges

Visual clutter

How can we avoid visual clutters like overlaps and crossings?

Performance issues

How can we render the huge datasets in real time with rich interactions?

Limited cognition

How can users understand the visual representation when the information is overwhelming?
Progressive Visualization

Progressive visualizations are augmented with error metrics indicating that the current view is only an approximation of the final result. Even though confidence intervals in the example are rather small, but note that the visualization and how it changes over time. Progressive visualizations are augmented with error metrics indicating that the current view is only an approximation of the final result.

Upon startup, the system loads the selected dataset into memory. The system randomly shuffles the data points in order to avoid artifacts caused by the natural order of the data and to improve convergence of progressive computations. Depending on the visualization condition, we display a simple loading animation until the full result is available, or incrementally delayed so that the final accurate visualization is displayed as quickly as possible, with subsequent updates appropriately reflecting the new query (e.g., changing an axis). Fig. 4 (top) shows an example of a blocking visualization over time.

We implemented the progressive condition by processing the data and to improve convergence of progressive computations. Depending on the visualization condition, we display a simple loading animation until the full result is available, or incrementally delayed so that the final accurate visualization is displayed as quickly as possible, with subsequent updates appropriately reflecting the new query (e.g., changing an axis). Fig. 4 (top) shows an example of a blocking visualization over time.

Fig. 4 (bottom) shows an example of a progressive visualization condition, where we display a progress indication in the bottom left corner of the visualization (Fig. 3d). Progressive visualizations are augmented with error metrics indicating that the current view is only an approximation of the final result. Even though confidence intervals in the example are rather small, but note that the visualization and how it changes over time. Progressive visualizations are augmented with error metrics indicating that the current view is only an approximation of the final result.

We simulated the blocking condition by artificially delaying the system's random number generator differently for each session. We seeded our system and ran all sessions on a quad-core 3.60 GHz, 1,080 pixel display.

We recruited 24 participants from a research university in the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 graduate), all of whom had some experience with data exploration or analysis tools (e.g., Excel, R, Pandas). 21 of the participants were from the US. All participants were students (22 undergraduate, 2 grade...
Data Science Tools

KDnuggets Analytics, Data Science, Machine Learning Software Poll, top tools share, 2015-2017

- Python
- R language
- SQL language
- RapidMiner
- Excel
- Spark
- Anaconda
- Tensorflow
- scikit-learn
- Tableau
- KNIME

[2017 %share] [2016 %share] [2015 %share]
The Cloud: Scaling Up

PC  Server

[Haeberlen and Ives, 2015]
The Cloud: Scaling Up

PC  Server  Cluster

[Haeberlen and Ives, 2015]
The Cloud: Scaling Up

PC  Server  Cluster  Data center

[Haeberlen and Ives, 2015]
The Cloud: Scaling Up

PC

Server

Cluster

Data center

Network of data centers

[Haeberlen and Ives, 2015]
Cloud Workloads

Figure 2: Log-log scale inverted CDF of job durations. Only the duration for which the job runs during the trace time period is known; thus, for example, we do not observe durations longer than around 700 hours. The thin, black line shows all jobs; the thick line shows production-priority jobs; and the dashed line shows non-production priority jobs.

3.3 Job durations

Job durations range from tens of seconds to essentially the entire duration of the trace. Over 2000 jobs (from hundreds of distinct users) run for the entire trace period, while a majority of jobs last for only minutes. We infer durations from how long tasks are active during the one month time window of the trace; jobs which are cut off by the beginning or end of the trace are a small portion (<1%) of jobs and consist mostly of jobs which are active for at least several days and so are not responsible for us observing many shorter job durations. These come from a large portion of the users, so it is not likely that the workload is skewed by one particular individual or application. Consistent with our intuition about priorities correlating with job types, production priorities have a much higher proportion of long-running jobs and the 'other' priorities have a much lower proportion. But slicing the jobs by priority or 'scheduling class' (which the trace providers say should reflect how latency-sensitive a job is) reveals a similar heavy-tailed distribution shape with a large number of short jobs.

3.4 Task shapes

Each task has a resource request, which should indicate the amount of CPU and memory space the task will require. (The requests are intended to represent the submitter's predicted "maximum" usage for the task.) Both the amount of the resources requested and the amount actually used by tasks varies by several orders of magnitude; see Figures 3 and 4, respectively. These are not just outliers. Over 2000 jobs request less than 0.0001 normalized units of memory per task, and over 8000 jobs request more than 0.1 units of memory per task. Similarly, over 70000 jobs request less than 0.0001 units of CPU per task, and over 8000 request more than 0.1 units of CPU. Both tiny and large resource requesting jobs include hundreds of distinct users, so it is not likely that the particularly large or small requests are caused by the quirky demands of a single individual or service.

We believe that this variety in task "shapes" has not been seen in prior workloads, if only because most schedulers simply do not support this range of sizes. The smallest resource requests are likely so small that it would be difficult for any VM-based scheduler to run a VM using that little memory. (0.0001 units would be around 50MB if the largest machines in the cluster had 512GB of memory.) Also, any slot-based scheduler, which includes all HPC and Grid installations we are aware of, would be unlikely to have thousands of slots per commodity machine.

The ratio between CPU and memory requests also spans a large range. The memory and CPU request sizes are correlated, but weakly (linear regression $R^2 \approx 0.14$). A large number jobs request 0 units of CPU — presumably they require so little CPU they can depend on running in the 'left-over' CPU of a machine; it makes little sense to talk about the CPU:memory ratio without adjusting these. Rounding these requests to the next largest request size, the

[Reiss et al., 2012]
# MapReduce

## Input
Geographic feature list

<table>
<thead>
<tr>
<th>I-5</th>
<th>Lake Washington</th>
<th>WA-520</th>
<th>I-90</th>
<th>...</th>
</tr>
</thead>
</table>

## Map
Emit each to all overlapping latitude-longitude rectangles

<table>
<thead>
<tr>
<th>(N, I-5)</th>
<th>(S, I-5)</th>
<th>(N, Lake Wash.)</th>
<th>(S, Lake Wash.)</th>
<th>(N, WA-520)</th>
<th>(S, I-90)</th>
<th>...</th>
</tr>
</thead>
</table>

## Shuffle
Sort by key (key = Rect. Id)

<table>
<thead>
<tr>
<th>(N, I-5)</th>
<th>(N, Lake Wash.)</th>
<th>(N, WA-520)</th>
<th>...</th>
</tr>
</thead>
</table>

## Reduce
Render tile using data for all enclosed features

<table>
<thead>
<tr>
<th>(S, I-5)</th>
<th>(S, Lake Wash.)</th>
<th>(S, I-90)</th>
<th>...</th>
</tr>
</thead>
</table>

## Output
Rendered tiles

(Bucket pattern) (Parallel rendering)

[Google]
Spark and RDDs

Example of how Spark computes job stages. Boxes with solid outlines are RDDs. Partitions are shaded rectangles, in black if they are already in memory. To run an action on RDD G, we build stages at wide dependencies and pipeline narrow transformations inside each stage. In this case, stage 1’s output RDD is already in RAM, so we run stage 2 and then 3.

[Zaharia et al., 2012]
Data Cleaning By Example

**Input-example**

- **Input Example**
  - Raw Data

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bureau of U.A.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regional Director</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Nina C.</td>
<td>Tel: (600)45-8397</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Fax: (907)566-7552</td>
</tr>
<tr>
<td>5</td>
<td>Joan H.</td>
<td>Tel: (918)781-4600</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Fax: (918)781-4604</td>
</tr>
</tbody>
</table>

- **Output Example**

**Raw Data**

**Test Data**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bureau of U.A.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regional Director</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Nina C.</td>
<td>Tel: (600)45-8397</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Fax: (907)566-7552</td>
</tr>
<tr>
<td>5</td>
<td>Joan H.</td>
<td>Tel: (918)781-4600</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Fax: (918)781-4604</td>
</tr>
</tbody>
</table>

**Transformed Data**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

**System**

**Synthesized Program**

```
# Data Transformation

t = f_split(t, ",", "'")
t = f_delete(t, 2)
t = f_fill(t, 3)
t = f_unfold(t, 1)
```

**Synthesized Data Transformation Program in Python**

- **[Z. Jin et al., 2017]**
Data Cleaning Using Samples

A recent survey on duplicate detection has argued that the matching phase takes on the order of minutes for a dataset of thousands of tuples [39]. This is especially true in the context of blocking phase [13]. For instance, an evaluation of the popular duplicate detection technique [9] shows that the matching phase is typically much more expensive than the reference-based estimation, leading to an over-estimated result estimation to ensure that the estimate remains acceptable quality.

It is important to note, however, that compared to full crowdsourcing is used to get humans to match each tuple pair as duplicate or non-duplicate [9]. As a classification problem, and train a classifier to label tuples as duplicate or not, existing techniques typically model this problem as a binary classification.

Dirty Data

Data Cleaning

Sample Creation

Cleaned Sample

Dirty Sample

Aggregate Queries

Result Estimation (RawSC)

Result Estimation (NormalizedSC)

Results with Confidence Intervals

Results with Confidence Intervals

Due to the large (quadratic) cost of all-deduplication information is loaded back into the dataset. For instance, an evaluation of the popular duplicate detection technique [9] shows that the matching phase is typically much more expensive than the reference-based estimation, leading to an over-estimated result estimation to ensure that the estimate remains acceptable quality.

It is important to note, however, that compared to full crowdsourcing is used to get humans to match each tuple pair as duplicate or non-duplicate [9]. As a classification problem, and train a classifier to label tuples as duplicate or not, existing techniques typically model this problem as a binary classification.

Dirty Data

Data Cleaning

Sample Creation

Cleaned Sample

Dirty Sample

Aggregate Queries

Result Estimation (RawSC)

Result Estimation (NormalizedSC)

Results with Confidence Intervals

Results with Confidence Intervals

Due to the large (quadratic) cost of all-deduplication information is loaded back into the dataset. For instance, an evaluation of the popular duplicate detection technique [9] shows that the matching phase is typically much more expensive than the reference-based estimation, leading to an over-estimated result estimation to ensure that the estimate remains acceptable quality.

It is important to note, however, that compared to full crowdsourcing is used to get humans to match each tuple pair as duplicate or non-duplicate [9]. As a classification problem, and train a classifier to label tuples as duplicate or not, existing techniques typically model this problem as a binary classification.

Dirty Data

Data Cleaning

Sample Creation

Cleaned Sample

Dirty Sample

Aggregate Queries

Result Estimation (RawSC)

Result Estimation (NormalizedSC)

Results with Confidence Intervals

Results with Confidence Intervals

Due to the large (quadratic) cost of all-deduplication information is loaded back into the dataset. For instance, an evaluation of the popular duplicate detection technique [9] shows that the matching phase is typically much more expensive than the reference-based estimation, leading to an over-estimated result estimation to ensure that the estimate remains acceptable quality.

It is important to note, however, that compared to full crowdsourcing is used to get humans to match each tuple pair as duplicate or non-duplicate [9]. As a classification problem, and train a classifier to label tuples as duplicate or not, existing techniques typically model this problem as a binary classification.
Data Fusion: Less is More

Figure 1: Coverage of results.

Figure 3: Different integration models.
CAP Theorem

Scalability: CAP Theorem

- Availability: Remains accessible and operational at all times.
- Consistency: Commits are atomic across the entire distributed system.
- Partition Tolerance: Only a total network failure can cause the system to respond incorrectly.

Pick Two!

Traditional relational databases: PostgreSQL, MySQL, etc.

Voldemort, Riak, Cassandra, CouchDB, Dynamo-like systems

CP
- HBase
- MongoDB
- Redis
- Memcached
- BigTable-like systems

CA
- CA

AP
- AP

SQL

[E. Brewer]
NoSQL: Replication and Consistency Levels

DC 1

DC 2

Write

[R. Stupp]
Data Cubes and Nanocubes

\[ \ell_{\text{device}}(\bigcirc) = \text{Android} \]
\[ \ell_{\text{device}}(\bullet) = \text{iPhone} \]

Indexing Schema
\[ S = [([\ell_{\text{spatial1}}, \ell_{\text{spatial2}}], \ell_{\text{device}})] \]

1. Five Tweets: Location and Device
2. 3D visualization of spatial data
3. 3D visualization of temporal data
4. 3D visualization of device data
5. 3D visualization of combined data

[Lins et. al, 2013]
Scalable Machine Learning

Devices: Processes, Machines, GPUs, etc

[J. Dean, 2015]
Graph Analytics

Raw Wikipedia

Text Table

Hyperlinks

PageRank

Top 20 Pages

Term-Doc Graph

Topic Model (LDA)

Word Topics

Editor Graph

Community Detection

User Community

Community Topic

User Disc.

Table

[J. Gonzalez, 2014]
Internet of Things

Fig. 5. Services made possible thanks to the CloudIoT paradigm.
Source: http://siliconangle.com/

4. Applications

CloudIoT gave birth to a new set of smart services and applications, that can strongly impact everyday life (Fig. 5). Many of the applications described in the following (may) benefit from Machine-to-Machine communications (M2M) when the things need to exchange information among themselves and not only send them towards the cloud [42]. This represents one of the open issues in this field, as discussed in Section 7. In this section we describe the wide set of applications that are made possible or significantly improved thanks to the CloudIoT paradigm. For each application we point out the challenges – see Fig. 6 – which we discuss in detail in Section 5.

Healthcare.

The adoption of the CloudIoT paradigm in the healthcare field can bring several opportunities to medical IT, and experts believe that it can significantly improve healthcare services and contribute to its continuous and systematic innovation [43]. Indeed, CloudIoT employed in this scenario is able to simplify healthcare processes and allows to enhance the quality of the medical services by enabling the cooperation among the different entities involved. Ambient assisted living (AAL), in particular, aims at easing the daily lives of people with disabilities and chronic medical conditions. Through the application of CloudIoT in this field it is possible to supply many innovative services, such as: collecting patients’ vital data via a network of sensors connected to medical devices, [Botta et al., 2016]
Final Project

• Presentations: Saturday, Dec. 16 (8-11am), Sorry :(  
• Presentation (5-6 minutes)  
  - Brief Introduction to Dataset/Problem (1 minute)  
  - Scalability Challenges (1-2 minutes)  
  - Results (2-3 minutes)  
• Report  
  - Expand proposal into technical report (figures, tables, references)  
• Code  
  - Include README  
  - Source with build/run instructions  
  - Link to data (make subset available to me if not publicly available)