Advanced Data Management (CSCI 640/490)

Introduction

Dr. David Koop
NYC Taxi Data

[Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance, T. W. Schneider]
NYC Taxi Data: Day analysis

Number of Trips for the years of 2011, and 2012

[Graph showing number of trips by month for 2011 and 2012]

[Several maps showing taxi movements in Manhattan, with different colors indicating different years or times]

[Image reference: Ferreira et al., 2013]
NYC Taxi Data: Region analysis

6.2 Exploring Movement: Transportation Hubs

By analyzing the data, we can observe the movement patterns for airports and train stations. For example, examining the average cost of trips per mile is lower for Harlem, and thus, there is less economic incentive for taxis to be in that area. The higher tips may be a means to reward drivers that go to Harlem.

Another interesting question is where passengers go. The choropleth (Fig. 8) shows that most people go to Midtown (the darkest region), followed by the Upper West Side. This is accentuated during rush hour on weekdays, when trips take much longer (see Fig. 9). Further analysis also showed that the fare per mile is lower for Harlem, and thus, there is less economic incentive for taxis to be in that area. The higher tips may be a means to reward drivers that go to Harlem.

6.3 Studying Behavior over Time

Studying how taxi demand varies over time can be useful to understand city dynamics. For taxi companies, this is a common practice when the destination is JFK.

An interesting observation is that the number of trips originating at the airports (blue outline) is updated to show the number of pickups in the two regions. Note that there are many more pickups around the train stations. Another key transportation hub in NYC is the Grand Central. By analyzing taxi movement to and from these locations, we can obtain insights into how people move into and out of the city. To compare the number of trips originating at and from these locations, we can obtain insights into how people move.

Using the summary view, we can further explore features of the destination-discrimination. For example, by examining the average cost of trip per mile, we can observe how the cost varies across different neighborhoods.

For taxis to be in that area, the economic incentive must be higher. The higher tips may be a means to reward drivers that go to Harlem.

The plot clearly shows that their discomfort is well justified. There is over one order of magnitude difference in the number of trips to/from Harlem compared to other more affluent neighborhoods. The heat map also shows that while people take taxis to Harlem, there are barely any pickups there. Exploring other neighborhoods, we find that there are many more pickups around the train stations.

The user first selects the time slices of interest. This can be done using the time selection widgets (Fig. 10). In the regular selection mode, the list of time ranges is already expressed and generated by the system. In the recurrent selection mode, TaxiVis provides a list of time ranges, each corresponding to a Sunday in the month of May, 2011. Given a list of time ranges, the result of a time space exploration is a multi-view visualization displaying one map per time interval, and a data summary view that aggregates the results for the entire period.

By examining the average cost of trip per mile, we can observe how the cost varies across different neighborhoods. This problem is well studied in the literature (e.g., [Ferreira et al., 2013]). In the bottom, we refine the query to compare trips starting at the two major airports in NYC: JFK and La Guardia. In the top, we examine trips starting at the two major airports in NYC: JFK and La Guardia. In the bottom, we refine the query to compare trips starting at the train stations, Penn Station and Grand Central.
Baseball Data

As with pitch data, we can aggregate all plays to generate heatmaps of the hit ball trajectories. The heat map shown in Figure 9 demonstrates the varying ball and fielder trajectories for hits that arise from different pitch types. 

With new tracking abilities, we can better track fielding performance. Coaches and players can visually determine where the ball landed and what player was responsible for. Each field point is assigned to the player that made the most catches in that location during the 2013 season. The availability of actual fielding data may help to determine the regions each player is responsible for. Each field point is assigned to the player that made the most catches in that location during the 2013 season. The availability of actual fielding data may help to determine the regions each player is responsible for.

One interesting direction that uses filtering is to combine pitch characteristics with hit ball trajectories and fielder trajectories. If a manager wants to induce a double-play, for example, he may signal in a particular type of pitch. Analysis of the hit trajectories of that type of pitch might enhance the understanding of how effective such strategies are. Note that a successful double-play depends on the quality of the pitcher's initial result, the hitter's decision, the ball trajectory, the fielders' initial movements, and the execution of all of the plays that follow. 

Two interesting directions that use filtering are to combine pitch characteristics with hit ball trajectories and fielder trajectories. If a manager wants to induce a double-play, for example, he may signal in a particular type of pitch. Analysis of the hit trajectories of that type of pitch might enhance the understanding of how effective such strategies are. Note that a successful double-play depends on the quality of the pitcher's initial result, the hitter's decision, the ball trajectory, the fielders' initial movements, and the execution of all of the plays that follow.

We could use Baseball4D to evaluate the different trajectories and speeds. Figure 1 presents a step towards improving our understanding of such complicated plays. In the previous era of fielding ability, the later tracking approaches (such as the zone rating of STATS Inc. and Baseball Info Solutions) were based on a discretization of the field into zones. The zones helped the reporters (zone rating operators) to visually determine where the ball landed and what player was supposed to field it (zones were assigned to specific players). The assignment of zones to specific fielders was put to discussion later, however, as with pitch data, we can aggregate all plays to generate heatmaps of the hit ball trajectories. The heat map shown in Figure 9 demonstrates the varying ball and fielder trajectories for hits that arise from different pitch types.
Baseball Data

[Deitrich et al., 2014]
Real-time Analysis

• Want to have results now
• How?
  - Faster machines
  - Clusters
  - Progressive techniques
What's involved in dealing with data?

Data Acquisition
- Structured data
- Unstructured data
- Event processing
- Sensor networks
- Protocols
- Real-time
- Data streams
- Multimodality

Data Analysis
- Stream mining
- Semantic analysis
- Machine learning
- Information extraction
- Linked Data
- Data discovery
- ‘Whole world’ semantics
- Ecosystems
- Community data analysis
- Cross-sectorial data analysis

Data Curation
- Data Quality
- Trust / Provenance
- Annotation
- Data validation
- Human-Data Interaction
- Top-down/Bottom-up
- Community / Crowd
- Human Computation
- Curation at scale
- Incentivisation
- Automation
- Interoperability

Data Storage
- In-Memory DBs
- NoSQL DBs
- NewSQL DBs
- Cloud storage
- Query Interfaces
- Scalability and Performance
- Data Models
- Consistency, Availability, Partition-tolerance
- Security and Privacy
- Standardization

Data Usage
- Decision support
- Prediction
- In-use analytics
- Simulation
- Exploration
- Visualisation
- Modeling
- Control
- Domain-specific usage

[Big Data Value Chain, Curry et al., 2014]
Data to Knowledge

[D. Somerville, based on H. McLeod’s original]
Data to Knowledge

[D. Somerville, based on H. McLeod’s original]
Data to Knowledge

Require People?

[D. Somerville, based on H. McLeod’s original]
Data to Knowledge

What can computers help with?

[D. Somerville, based on H. McLeod’s original]
FINDINGS

we got about the future of the data science, the most salient takeaway was how excited our respondents were about the evolution of the field. They cited things in their own practice, how they saw their jobs getting more interesting and less repetitive, all while expressing a real and broad enthusiasm about the value of the work in their organization.

As data science becomes more commonplace and simultaneously a bit demystified, we expect this trend to continue as well. After all, last year's respondents were just as excited about their work (about 79% were "satisfied" or better).

How a Data Scientist Spends Their Day

Here's where the popular view of data scientists diverges pretty significantly from reality. Generally, we think of data scientists building algorithms, exploring data, and doing predictive analysis. That's actually not what they spend most of their time doing, however.

As you can see from the chart above, 3 out of every 5 data scientists we surveyed actually spend the most time cleaning and organizing data. You may have heard this referred to as "data wrangling" or compared to digital janitor work. Everything from list verification to removing commas to debugging databases— that time adds up and it adds up immensely. Messy data is by far the more time-consuming aspect of the typical data scientist's work flow. And nearly 60% said they simply spent too much time doing it.

How do data scientists spend their time?

What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

[CrowdFlower Data Science Report, 2016]
Why That's a Problem

Simply put, data wrangling isn't fun. It takes forever. In fact, a few years back, the New York Times estimated that up to 80% of a data scientist's time is spent doing this sort of work. Here, it's necessary to point out that data cleaning is incredibly important. You can't do the sort of work data scientists truly enjoy doing with messy data. It needs to be cleaned, labeled, and enriched before you can trust the output.

The problem here is two-fold. One: data scientists simply don't like doing this kind of work, and, as mentioned, this kind of work takes up most of their time. We asked our respondents what was the least enjoyable part of their job. They had this to say:

Note how those last two charts mirror each other. The things data scientists do most are the things they enjoy least. Last year, we found that respondents far prefer doing the more creative, interesting parts of their job, things like predictive analysis and mining data for patterns. That's where the real value comes. But again, you simply can't do that work unless the data is properly labeled. And nobody likes labeling data.

Do Data Scientists Have What They Need?

With a shortage of data scientists out there in the world, we wanted to find out if they thought they were properly supported in their job. After all, when you need more data scientists, you'll often find a single person doing the work of several.

What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

[CrowdFlower Data Science Report, 2016]
Example: Compare public transit in Chicago and NYC
<table>
<thead>
<tr>
<th>station_id</th>
<th>stationname</th>
<th>date</th>
<th>datatype</th>
<th>rides</th>
</tr>
</thead>
<tbody>
<tr>
<td>40350</td>
<td>UIC-Halsted</td>
<td>01/01/2001</td>
<td>U</td>
<td>273</td>
</tr>
<tr>
<td>41130</td>
<td>Halsted-Orange</td>
<td>01/01/2001</td>
<td>U</td>
<td>306</td>
</tr>
<tr>
<td>40700</td>
<td>Granville</td>
<td>01/01/2001</td>
<td>U</td>
<td>1,050</td>
</tr>
<tr>
<td>40070</td>
<td>Jackson/Dearborn</td>
<td>01/01/2001</td>
<td>U</td>
<td>649</td>
</tr>
<tr>
<td>40320</td>
<td>Damen-Brown</td>
<td>01/01/2001</td>
<td>U</td>
<td>411</td>
</tr>
<tr>
<td>40550</td>
<td>Damen/Milwaukee</td>
<td>01/01/2001</td>
<td>U</td>
<td>870</td>
</tr>
<tr>
<td>40720</td>
<td>East 63rd-Cottage Grove</td>
<td>01/01/2001</td>
<td>U</td>
<td>301</td>
</tr>
<tr>
<td>41260</td>
<td>Austin-Lake</td>
<td>01/01/2001</td>
<td>U</td>
<td>359</td>
</tr>
<tr>
<td>40230</td>
<td>Cumberland</td>
<td>01/01/2001</td>
<td>U</td>
<td>788</td>
</tr>
<tr>
<td>41120</td>
<td>35 Bronzeville-ET</td>
<td>01/01/2001</td>
<td>U</td>
<td>448</td>
</tr>
<tr>
<td>40810</td>
<td>Medical Center</td>
<td>01/01/2001</td>
<td>U</td>
<td>479</td>
</tr>
<tr>
<td>40330</td>
<td>Grand/Slake</td>
<td>01/01/2001</td>
<td>U</td>
<td>2,542</td>
</tr>
<tr>
<td>41050</td>
<td>Linden</td>
<td>01/01/2001</td>
<td>U</td>
<td>176</td>
</tr>
<tr>
<td>40140</td>
<td>Slagle</td>
<td>01/01/2001</td>
<td>U</td>
<td>0</td>
</tr>
<tr>
<td>40450</td>
<td>95th/Dan Ryan</td>
<td>01/01/2001</td>
<td>U</td>
<td>3,948</td>
</tr>
<tr>
<td>40460</td>
<td>Noyes</td>
<td>01/01/2001</td>
<td>U</td>
<td>72</td>
</tr>
<tr>
<td>40150</td>
<td>Pulaski-Cermak</td>
<td>01/01/2001</td>
<td>U</td>
<td>0</td>
</tr>
<tr>
<td>40650</td>
<td>Dempster</td>
<td>01/01/2001</td>
<td>U</td>
<td>177</td>
</tr>
<tr>
<td>40460</td>
<td>Merchandise Mart</td>
<td>01/01/2001</td>
<td>U</td>
<td>185</td>
</tr>
<tr>
<td>40840</td>
<td>South Boulevard</td>
<td>01/01/2001</td>
<td>U</td>
<td>202</td>
</tr>
<tr>
<td>41260</td>
<td>Jefferson Park</td>
<td>01/01/2001</td>
<td>U</td>
<td>1,302</td>
</tr>
<tr>
<td>40150</td>
<td>Stid</td>
<td>01/01/2001</td>
<td>U</td>
<td>364</td>
</tr>
<tr>
<td>40870</td>
<td>Francisco</td>
<td>01/01/2001</td>
<td>U</td>
<td>156</td>
</tr>
<tr>
<td>40710</td>
<td>Chicago/Franklin</td>
<td>01/01/2001</td>
<td>U</td>
<td>384</td>
</tr>
<tr>
<td>40740</td>
<td>Western-Cermak</td>
<td>01/01/2001</td>
<td>U</td>
<td>0</td>
</tr>
<tr>
<td>40550</td>
<td>Irving Park-Chiare</td>
<td>01/01/2001</td>
<td>U</td>
<td>731</td>
</tr>
</tbody>
</table>

Public Transit Ridership Data
Cool Machine Learning Model & Pretty Visualizations

D. Koop, CSCI 640/490, Spring 2023
Wait... how do we actually get those results?
Processing the data

• Data Ingestion
  - Need to understand format of the data
  - Need to understand what the data is (types and semantics)

• Data Wrangling
  - Get the data into a meaningful state
  - Check for errors in the data
  - Check for missing data and deal with it

• Data Integration
  - Make it so we can actually compare the data
  - Put the datasets together
Cool Machine Learning Model & Pretty Visualizations

[Image of a machine learning model with graphs and a visualization of a train system boardings and ridership]

Note: Boardings include transfers.

OVERALL TRAIN SYSTEM BOARDINGS
- Blue Line (Downtown Subway): 229.9 million
- Blue Line (Forest Park): 238.1 million
- Blue Line (O’Hare): 241.7 million

OVERALL BUS SYSTEM
- Total Ridership: 1.2 million

LINES WITH THE LARGEST INCREASES
- Union/Olivet Express (417): 62.7%
- West 63rd (416): 62.1%

LINES WITH THE LARGEST DROPS
- 49th (210): -91.0%
- 57th (208): -68.3%

Segment (14) includes all riders who entered those stations and may have taken the Red, Brown or Purple lines.

1. Blue Line (Downtown Subway): 13.8%
2. Blue Line (Forest Park): 1.9%
3. Blue Line (O’Hare): 9.5%
4. Brown Line: 0.9%
5. Green Line (Ashland/3rd Branch): -11.7%
6. Green Line (East 63rd Branch): -0.9%
7. Green Line (Lake Street): 6.4%
8. Green Line (South Elevated): 16.2%
9. Loop (Brown, Orange, Pink, Purple, Green): 0.1%
10. Orange Line: 3.3%
11. Pink Line: 9.2%
12. Purple Line (Extravision): 1.8%
13. Red Line (Dan Ryan): 3.2%
14. Red Line (North Side): 4.2%
15. Red Line (State Street Subway): 10.5%
16. Yellow Line*: -49.2%

*Between May 17 and Oct. 30, 2015, there was no train service between Dempster-Skokie and Howard because of repairs after an embankment collapse.

[C. Bronner (left), Chicago Business (right)]

D. Koop, CSCI 640/490, Spring 2023
Lots of topics related to this
Data Wrangling

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transaction Date</strong></td>
<td><strong>Customer Name</strong></td>
<td><strong>Phone Numbers</strong></td>
<td><strong>Address</strong></td>
</tr>
<tr>
<td><strong>1</strong></td>
<td><strong>2</strong></td>
<td><strong>3</strong></td>
<td><strong>4</strong></td>
</tr>
<tr>
<td>Wed, 12 Jan 2011</td>
<td>John K. Doe Jr.</td>
<td>(609)-993-3001</td>
<td>2196 184th Ave NE, Redmond, 98052</td>
</tr>
<tr>
<td>Thu, 15 Sep 2011</td>
<td>Mr. Doe, John</td>
<td>609.993.3001 ext 2001</td>
<td>4297 148th Avenue NE, Bellevue, 98007</td>
</tr>
<tr>
<td>Mon, 17 Sep 2012</td>
<td>Jane A. Smith</td>
<td>+1-4250013981</td>
<td>2720 N Mesa St, El Paso, 79902, USA</td>
</tr>
<tr>
<td>2011-Jan-11 02:27:21</td>
<td>Smith, Jane</td>
<td>tel: 4250013981</td>
<td>4740 N 132nd St Apt 417, Omaha, 68164</td>
</tr>
<tr>
<td>2011-Jan-12</td>
<td>Anthony Von Fange</td>
<td>650-384-9911</td>
<td>10508 Prairie Ln, Oklahoma City</td>
</tr>
<tr>
<td>2010-Dec-24</td>
<td>Mr. Peter Tyson</td>
<td>(405)123-3981</td>
<td>525 1st St, Marysville, WA 95901</td>
</tr>
<tr>
<td>9/22/2011</td>
<td>Dan E. Williams</td>
<td>1-650-1234183</td>
<td>211 W Ridge Dr, Waukon, 52172</td>
</tr>
<tr>
<td>7/11/2012</td>
<td>James Davis Sr.</td>
<td>+1-425-736-9999</td>
<td>13120 Five Mile Rd, Brainerd</td>
</tr>
<tr>
<td>2/12/2012</td>
<td>Mr. James D. Davis</td>
<td>425.736.9999 x 9</td>
<td>602 Highland Ave, Shinnston, 26431</td>
</tr>
<tr>
<td>12/31/2013</td>
<td>Donald Edward Miller</td>
<td>(206) 309-8381</td>
<td>840 W Star St, Greenville, 27834</td>
</tr>
<tr>
<td>6/1/2009 12:01</td>
<td>Miller, Donald</td>
<td>206 309 8381</td>
<td>25571 Elba, Redford, 48239</td>
</tr>
<tr>
<td>2/26/2007 18:37</td>
<td>Rajesh Krishnan</td>
<td>206 456 8500 extension 1</td>
<td>539 Co Hwy 48, Silkenton, USA</td>
</tr>
<tr>
<td>1/4/2011 14:33</td>
<td>Daniel Chen</td>
<td>425 960 3566</td>
<td>1008 Whitlock Ave NW, Marietta, 30064</td>
</tr>
</tbody>
</table>

- **C**: Transaction Date
- **D**: Output
  - John K. Doe Jr.
  - Mr. Doe, John
  - Jane A. Smith
  - Ms. Jane Smith
  - Smith, Jane
  - Dr Anthony R Von Fange II
  - Peter Tyson
  - Dan E. Williams
  - James Davis Sr.
  - James J. Davis
  - Mr. Donald Edward Miller

- **Address**:
  - 2196 184th Ave NE Apt 417, Redmond, 98052
  - 4297 148th Avenue NE L105, Bellevue, WA 98007
  - 2720 N Mesa St, El Paso, 79902, USA
  - 352 W Shore Rd APT 1002, Warwick, 02886
  - 4740 N 132nd St, Omaha, 68164
  - 10508 Prairie Ln, Oklahoma City
  - 525 1st St, Marysville, WA 95901
  - 211 W Ridge Dr, Waukon, 52172
  - 602 Highland Ave, Shinnston, 26431
  - 840 W Star St, Greenville, 27834

**[Y. He et al., 2018]**
## Data Wrangling

### Contract End Details

<table>
<thead>
<tr>
<th>IMSI</th>
<th>CONTRACT_END</th>
<th>CONTRACT_START</th>
<th>SUBSCRIBER_AGE</th>
<th>STATUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>310780268122721</td>
<td>Jan 2013 - Dec 2016</td>
<td>7/29/199</td>
<td>0 - 15</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>310808907667609</td>
<td>3/28/15</td>
<td>16/6/13</td>
<td>1</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>3108054622541</td>
<td>9/23/16</td>
<td>1/6/97</td>
<td>7</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>310805432849230</td>
<td>5/29/9</td>
<td>14/6/81</td>
<td>13</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>310802693721005</td>
<td>9/11/15</td>
<td>18/9/10</td>
<td>4</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>310806281466952</td>
<td>8/27/15</td>
<td>3/13/86</td>
<td>8</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>310784824724601</td>
<td>1/16/16</td>
<td>5/11/84</td>
<td>1</td>
<td>ACTIVE</td>
</tr>
</tbody>
</table>

### Hide Example Values

- 9/10/15
- 6/13/15
- 5/21/15
- 12/13/15
- 1/16/16

### Pattern Details

- **IMSI**: Unique identifier for each record
- **Contract End**: Date the contract ended
- **Contract Start**: Date the contract started
- **Subscriber Age**: Age of the subscriber
- **Status**: Active or Inactive

---

D. Koop, CSCI 640/490, Spring 2023

[Trifacta]
Data Cleaning/Standardization (Aliases)

- 'google brain resident': 'google',
- 'google brain': 'google',
- 'google inc.': 'google',
- 'google inc.': 'google',
- 'google research nyc': 'google',
- 'google research': 'google',
- 'google, inc.': 'google',
- 'deepmind @ google': 'deepmind',
- 'deepmind technologies': 'deepmind',
- 'google deepmind': 'deepmind',

- 'ibm research - china': 'ibm',
- 'ibm research': 'ibm',
- 'ibm research, ny': 'ibm',
- 'ibm research, usa': 'ibm',
- 'ibm t.j. watson research center': 'ibm',
- 'ibm t.j. watson research center': 'ibm',
- 'ibm t.j. watson research center': 'ibm',
- 'ibm t.j. watson research center': 'ibm',
- 'ibm t.j. watson research center': 'ibm',
- 'ibm thomas j. watson research center': 'ibm',
- 'ibm tj watson research center': 'ibm',

- 'microsoft research cambridge': 'microsoft',
- 'microsoft research india': 'microsoft',
- 'microsoft research maluuba': 'microsoft',
- 'microsoft research new england': 'microsoft',
- 'microsoft research': 'microsoft',
- 'microsoft research, redmond, w': 'microsoft',
- 'microsoft research, redmond, wa': 'microsoft',
- 'microsoft research': 'microsoft',
- 'microsoft research': 'microsoft',

[ICML, NIPS, EMNLP, NAACL, EACL, ACL]

0 50 100 150 200 250 300

Georgia Institute of Technology
Chinese Academy of Sciences
Harbin Institute of Technology
Peking University
Tsinghua University
Cornell University
University College London
University of Toronto
University of Maryland
INRIA
University of Pennsylvania
Princeton University
UT Austin
Johns Hopkins University
University of Cambridge
University of Edinburgh
IBM
University of Washington
Massachusetts Institute of Technology
UC Berkeley
Columbia University
Google
Stanford University
Microsoft
Carnegie Mellon University

INRIA
Harvard University
Microsoft Research
Google Brain
Stanford University
Microsoft
Carnegie Mellon University

[D. Koop, CSCI 640/490, Spring 2023]
Anahí Giovanna Puente Portilla (born 14 May 1983), known mononymously as Anahí, is a Mexican actress and singer. In 1986, she started her acting career when she was cast on Chiquilladas. After working on many successful telenovelas produced by Televisa, including Alondra (1995), Vivo por Elena (1998), El Diario de Daniela (1998) and Mujeres Engañadas (1999), her first leading role was in Pedro Damián's production, Primer Amor... A Mi por Hora (2000). In 2003, she joined the cast in Clase 406. Anahí reached international success in 2004 after starring in Rebelde and being part of the twice-nominated for a Latin Grammy Award group RBD, who sold over 15 million records worldwide. In 2011, she starred in Dos Hogares, her last telenovela to date.

In 1993, at the age of 10, Anahí released her debut self-titled studio album.[2] In 1996, she released her second album, Hoy Es Mariana!, which featured the singles "Corazón de Bombón", "Por Volverte a Ver" and "Descontroladote". She went on to record two additional albums titled Anclado en Mi Corazón (1997) and Baby Blue (2000). In 2009, Anahí released her fifth album, Mi Delirio, which sold 500,000 copies worldwide.[3] The album debuted at number two on Billboard's Latin Pop Albums chart and number four on Billboard's Top Latin Albums chart,[4] and was certified Gold in Brazil.[5] Mi Delirio World Tour was her first worldwide tour. According to Billboard, Mi Delirio World Tour was the seventh most profitable tour of 2010.[6] Anahí's sixth studio album, Inesperado (2016), was preceded by the singles "Rumbal", "Eres" and "Amnesia", "Rumbal", a collaboration with reggaeton singer Wisin, peaked at number one on Billboard's Tropical Songs chart.[7] The album debuted on Billboard's Latin Pop Albums and Top Latin Albums charts,[8] while it was number one on Billboard Brazil, with her being the first Mexican artist to achieve that.[9] She has sold over five million albums worldwide in her career as a solo artist.[10]
Anahí Giovanna Puente Portilla (born 14 May 1983), known mononymously as Anahí, is a Mexican actress and singer. In 1986, she started her acting career when she was cast on Chiquilladas. After working on many successful telenovelas produced by Televisa, including Alondra (1995), Vivo por Elena (1998), El Diario de Daniela (1998) and Mujeres Engañadas (1999), her first leading role was in Pedro Damiani's production, Primer Amor... A Mi por Hora (2000). In 2003, she joined the cast in Clase 406. Anahí reached international success in 2004 after starring in Rebelde and being part of the twice-nominated for a Latin Grammy Award group RBD, who sold over 15 million records worldwide. In 2011, she starred in Dos Héroes, her last telenovela to date.

In 1993, at the age of 10, Anahí released her debut self-titled studio album. In 1996, she released her second album, Hoy Es Marta, which featured the singles "Corazón de Bombón", "Por Volverte a Ver" and "Desconfortadote". She went on to record two additional albums titled Anclado en Mi Corazón (1997) and Baby Blue (2000). In 2009, Anahí released her fifth album, Mi Delirio, which sold 500,000 copies worldwide. The album debuted at number two on Billboard's Latin Pop Albums chart and number four on Billboard's Top Latin Albums chart, and was certified Gold in Brazil. Mi Delirio World Tour was her first worldwide tour. According to Billboard, Mi Delirio World Tour was the seventh most profitable tour of 2010.

Anahí's sixth studio album, Inesperado (2016), was preceded by the singles "Rumba", "Eres" and "Amnesia". "Rumba", a collaboration with reggaeton singer Wisin, peaked at number one on Billboard's Tropical Songs chart. The album debuted on Billboard's Latin Pop Albums and Top Latin Albums charts; while it was number one on Billboard Brazil, with her being the first Mexican artist to achieve that. She has sold over five million albums worldwide in her career as a solo artist, being the first Mexican artist to achieve this milestone.

Anahí was born in Mexico. She's had roles in Tu y Yo, in which she played a 17 year old girl while she was 13, and Vivo Por Elena, in which she played Talita, a naive and innocent teenager. Anahí lives with her mother and sister name Marychelo. She hopes to become a fashion designer one day, and is currently...
Data Integration: Entity Resolution

- **Google Thinks I’m Dead**
  (I know otherwise.) [R. Abrams, NYTimes, 2017]

- **Not only Google, but also Alexa:**
  - "Alexa replies that Rachel Abrams is a sprinter from the Northern Mariana Islands (which is true of someone else)."
  - "He asks if Rachel Abrams is deceased, and Alexa responds yes, citing information in the Knowledge Graph panel."
Anahí Giovanna Puente Portilla (born 14 May 1983), known mononymously as Anahí, is a Mexican actress and singer. In 1986, she started her acting career when she was cast on Chiquilladas. After working on many successful telenovelas produced by Televisa, including Alondra (1995), Vivo por Elena (1998), El Diario de Daniela (1998) and Mujeres Enamoradas (1999), her first leading role was in Pedro Damiani's production, Primer Amor... A Mi por Hora (2000). In 2003, she joined the cast in Classe 406. Anahí reached international success in 2004 after starring in Rebelde and being part of the twice-nominated for a Latin Grammy Award group RBD, who sold over 15 million records worldwide. In 2011, she starred in Dos Hogares, her last telenovela to date.

In 1993, at the age of 10, Anahí released her debut self-titled studio album. In 1996, she released her second album, Hoy Es Martana, which featured the singles "Corazón de Bomboón", "Por Volverte a Ver" and "Descontrolándote". She went on to record two additional albums titled Anclado en Mi Corazón (1997) and Baby Blue (2000). In 2009, Anahí released her fifth album, Mi Delirio, which sold 500,000 copies worldwide. The album debuted at number two on Billboard's Latin Pop Albums chart and number four on Billboard's Top Latin Albums chart, and was certified Gold in Brazil. Mi Delirio World Tour was her first worldwide tour. According to Billboard, Mi Delirio World Tour was the seventh most profitable tour of 2010.

Anahí's sixth studio album, Inesperado (2016), was preceded by the singles "Rumba", "Eres" and "Amnesia", "Rumba", a collaboration with reggaeton singer Wisin, peaked at number one on Billboard's Tropical Songs chart. The album debuted on Billboard's Latin Pop Albums and Top Latin Albums charts, while it was number one on Billboard Brazil, with her being the first Mexican artist to achieve that. She has sold over five million albums worldwide in her career as a solo artist.
Anahí Giovanna Puente Portilla (born 14 May 1983), known mononymously as Anahí, is a Mexican actress and singer. In 1986, she started her acting career when she was cast on Chiquititas. After working on many successful telenovelas produced by Televisa, including Alondra (1995), Vivo por Elena (1998), El Diario de Daniela (1998) and Mujeres Engañadas (1999), her first leading role was in Pedro Damián’s production, Primer Amor... A Mi por Hora (2000). In 2003, she joined the cast in Clase 406. Anahí reached international success in 2004 after starring in Rebelde and being part of the twice-nominated for a Latin Grammy Award group RBD, who sold over 15 million records worldwide.[3] In 2011, she starred in Dos Hogares, her last telenovela to date.

In 1993, at the age of 10, Anahí released her debut self-titled studio album.[4] In 1996, she released her second album, Hoy Es Martana?, which featured the singles "Corazón de Bambón", "Por Volverte a Ver" and "Desconocíandote". She went on to record two additional albums titled Anclado en Mi Corazón (1997) and Baby Blue (2000). In 2009, Anahí released her fifth album, Mi Delirio, which sold 500,000 copies worldwide. The album debuted at number two on Billboard’s Latin Pop Albums chart and number four on Billboard’s Top Latin Albums chart,[5] and was certified Gold in Brazil. Mi Delirio World Tour was her first worldwide tour. According to Billboard, Mi Delirio World Tour was the seventh most profitable tour of 2010.[6]

Anahí's sixth studio album, Inesperado (2016), was preceded by the singles "Rumba", "Eres" and "Amnesia". "Rumba", a collaboration with reggaeton singer Wisin, peaked at number one on Billboard's Tropical Songs chart.[7] The album debuted on Billboard's Latin Pop Albums and Top Latin Albums charts,[8] while it was number one on Billboard Brazil, with her being the first Mexican artist to achieve that.[9] She has sold over five million albums worldwide in her career as a solo artist.[10]
## Data Integration: Data Fusion

### Source Observations

<table>
<thead>
<tr>
<th>Source</th>
<th>River</th>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>KG</td>
<td>Mississippi River</td>
<td>Length</td>
<td>2,320 mi</td>
</tr>
<tr>
<td>KG</td>
<td>Missouri River</td>
<td>Length</td>
<td>2,341 mi</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Mississippi River</td>
<td>Length</td>
<td>2,202 mi</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Missouri River</td>
<td>Length</td>
<td>2,341 mi</td>
</tr>
<tr>
<td>USGS</td>
<td>Mississippi River</td>
<td>Length</td>
<td>2,340 mi</td>
</tr>
<tr>
<td>USGS</td>
<td>Missouri River</td>
<td>Length</td>
<td>2,540 mi</td>
</tr>
</tbody>
</table>

### True Facts

<table>
<thead>
<tr>
<th>River</th>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mississippi River</td>
<td>Length</td>
<td>?</td>
</tr>
<tr>
<td>Missouri River</td>
<td>Length</td>
<td>?</td>
</tr>
</tbody>
</table>

**Goal:** Find the **latent** true value of facts.

[X. L. Dong & T. Rekatsinas, 2019]
Data Storage

**SQL DATABASES**

- **Relational**

**NoSQL DATABASES**

- **Column**
- **Graph**
- **Key-Value**
- **Document**

[V. Wilkinson]
Distributed Databases

Site 1

Site 2

Site 3

DB 1

DB 1

DB 2

DB 1
CAP Theorem
Scaling Dataframes

New Data Source

Prototyping

Exploring

Testing

Production

Laptop/Workstation

Small Cluster

Large Cluster

MODIN

Feedback

[D. Petersohn]
Figure 2: Visualizing the entire compression algorithm. For this example, 48 bytes of values and time stamps are compressed to just under 21 bytes/167 bits.

The key specified in the monitoring data is used to uniquely identify a time series. By sharding all monitoring data based on these unique string keys, each time series dataset can be mapped to a single Gorilla host. Thus, we can scale Gorilla by simply adding new hosts and tuning the sharding function to map new time series data to the expanded set of hosts. When Gorilla was launched to production 18 months ago, our dataset of all time series data inserted in the past 26 hours fit into 1.3TB of RAM evenly distributed across 20 machines. Since then, we have had to double the size of the clusters twice due to data growth, and are now running on 80 machines within each Gorilla cluster. This process was simple due to the share-nothing architecture and focus on horizontal scalability.

Gorilla tolerates single node failures, network cuts, and entire datacenter failures by writing each time series value to two hosts in separate geographic regions. On detecting a failure, all read queries are failed over to the alternate region ensuring that users do not experience any disruption.

4.1 Time series compression

In evaluating the feasibility of building an in-memory time series database, we considered several existing compression schemes to reduce the storage overhead. We identified techniques that applied solely to integer data which didn't meet our requirement of storing double precision floating point values. Other techniques operated on a complete dataset but did not support compression over a stream of data as was stored in Gorilla [7, 13]. We also identified lossy time series approximation techniques used in data mining to make the problem set more easily fit within memory [15, 11], but Gorilla is focused on keeping the full resolution representation of data.

Our work was inspired by a compression scheme for floating point data derived in scientific computation. This scheme leveraged XOR comparison with previous values to generate a delta encoding [25, 17].

Gorilla compresses data points within a time series with no additional compression used across time series. Each data point is a pair of 64 bit values representing the time stamp and value at that time. Timestamps and values are compressed separately using information about previous values. The overall compression scheme is visualized in Figure 2, showing how time stamps and values are interleaved in the compressed block.

Figure 2.a illustrates the time series data as a stream consisting of pairs of measurements (values) and time stamps. Gorilla compresses this data stream into blocks, partitioned by time. After a simple header with an aligned time stamp (starting at 2 am, in this example) and storing the first value in a less compressed format, Figure 2.b shows that time stamps are compressed using delta-of-delta compression, described in more detail in Section 4.1.1. As shown in Figure 2.d the time stamp delta of delta is 2. This is stored with a two bit header ('10'), and the value is stored in seven bits, for a total size of just 9 bits. Figure 2.c shows floating-point values are compressed using XOR compression, described in more detail in Section 4.1.2. By XORing the floating point value with the previous value, we find that there is only a single meaningful bit in the XOR. This is then encoded with a two bit header ('11'), encoding that there are eleven leading zeros, a single meaningful bit, and the actual value ('1'). This is stored in fourteen total bits.
Representing Graph Data

**INPUT GRAPH:**

- \( n \): number of vertices
- \( m \): number of edges
- \( d \): maximum graph degree

**Adjacency Matrix**

An \( n \times n \) matrix

- Unweighted graph: a cell is one bit
- Weighted graph: a cell is one integer

**Adjacency List**

- The number of all elements in the adjacency data structure: \( 2m \) (undirected graph) and \( m \) (directed graph)
- Neighbor-hoods can be arbitrary structures, e.g., arrays or lists
- Neighborhoods can't be sorted or unsorted
- Pointers from vertices to their neighborhoods

**Edge List (sorted & unsorted)**

- Offset array is optional
- Pointers from vertices to their neighborhoods
- Number of tuples: \( 2m \) (undirected), \( m \) (directed)
Data Cubes

Time

Product

Measures

2015 May
2015 Apr
2015 Mar
2015 Feb
2015 Jan

Accessories
Bikes
Clothing
Components

Units Sold
Gross Revenue
Cost of Goods Sold
Net Revenue
Average Sale Price
Visualization and Databases

- brushes in the precomputed view
- serves requests from a data cube


- interacts with a new view
- query for new data cubes

D. Mortiz et al.
Spatial Data
Provenance and Reproducibility

Data Management

Computation

Visualization

DATA

DATA

Visualization

Paper

DATA

DATA

Computation

Visualization

Provenance and Reproducibility
Provenance and Reproducibility

Data Management

Visualization

Provenance

Computation

Paper
About Me

• Research Interests
  - Visualization
  - Computational Provenance
  - Geospatial Analysis

• Research Projects
  - Dataflow Notebooks
  - Geospatial Trajectory Data
  - Provenance for Web Applications

• See my web page for more information
  - [http://faculty.cs.niu.edu/~dakoop/](http://faculty.cs.niu.edu/~dakoop/)
About You

• Research Papers?
• Data Science?
• Python?
• Database Experience?
• Analytics Experience?
• Cloud Computing Experience?
• Anything you want to see covered?
About this course

• Course web page is authoritative:
  - http://faculty.cs.niu.edu/~dakoop/cs640-2023sp/
  - Schedule, Readings, Assignments will be posted online
  - Check the web site before emailing me

• Lectures planned for in-person

• Course is meant to be more "cutting edge"
  - Focus on building skills related to data management
  - Tune into current research and tools

• Requires student participation: readings and discussions
About this course

- Balance of techniques and research ideas
- Background (Python & Relational DB) followed by topic areas & readings
- Programming assignments (4-5)
- Two tests + final exam: Please check these dates now
- Topic areas:
  - Data Wrangling
  - Data Cleaning, Integration, & Fusion
  - Scalable Data Management (Databases & Dataframes) + Cloud
  - Spatial, Graph, Time Series Data
  - Provenance and Reproducibility
About this course

• Course Registration:
  - Make sure you have registered for the course
  - Email me if you are not registered but are interested in taking the course

• Undergraduate (CS 490) and Graduate (CS 640)
  - Grad students have extra reading, exam questions, assignment tasks

• Review of course policies:
  - Plagiarism and academic honesty
  - If you have any concerns or questions, please email me as soon as possible

• If you are not sure if this course is a good fit, please email me or talk to me
Course Material

- Helpful Books:
  - *Python for Data Analysis*, W. McKinney
  - *Effective Pandas*, M. Harrison
  - *Intro to Python*, Deitel & Deitel
  - *Python Data Science Handbook*, J. VanderPlas

- Research papers
- Many websites
Course Material

• Software:
  - Anaconda Python Distribution (https://www.anaconda.com/distribution/): makes installing python and python packages easier
  - JupyterLab: Web-based interface for interactively writing and executing Python code
  - JupyterHub: Access everything through a server
Course Material

- Pandas:
  - Python library for data analysis
  - Many operations available
  - Efficient
- DuckDB
- Ibis
- Trifacta Wrangler
- More…
Office Hours & Email

• Scheduled office hours:
  - M: 2-3:15pm, W: 1-2:15pm, or by appointment
• You do not need an appointment to stop in during scheduled office hours
• If you need an appointment outside of those times, please email me with specific details about what you wish to discuss
• Many questions can be answered via email. Please do not schedule an appointment to ask a question that could be answered via email
Next Class

- Introduction to/review of Python
- Download and install anaconda distribution (Python 3.10):
  - [https://www.anaconda.com/distribution/](https://www.anaconda.com/distribution/)