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

[Fig. 2. Taxis as sensors of city life. The plot on the top shows how the number of trips varies over 2011 and 2012. While some patterns are regular, others are not. For example, there is a substantial drop in the number of trips during the Five Boro Bike Tour on May 1st, 2011.]

[Fig. 3. Maps showing the density of taxis across Manhattan from 7am to 11am on May 1st, 2011. Notice that there are few taxis along 6th Avenue, indicating the traffic was blocked.]

[Fig. 4. Maps showing the density of taxis across Manhattan from 7am to 11am on May 1st, 2011. Notice that there are fewer taxis along 6th Avenue compared to the previous year, indicating the traffic was blocked.]

[Fig. 5. Maps showing the density of taxis across Manhattan from 7am to 11am on May 1st, 2011. Notice that there are fewer taxis along 6th Avenue compared to the previous year, indicating the traffic was blocked.]

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

- Using the summary view, we can further explore features of the transportation hubs.
- By hovering the mouse over a neighborhood, the system displays the exact number of trips ending in that neighborhood.
- The heat map also shows that while people take the nightlife on weekends is very lively in Downtown.
- Note the increased number of dropoffs in Downtown.
- The figure also highlights the fact that Harlem is underserved by taxis.

[Image of heat map showing NYC neighborhoods]

[Fig. 8. Comparison of taxi pickups (left) and dropoff (right) in different neighborhoods over the first week of May 2011. The plots show that Midtown is the most active area, followed by the Upper West Side.]

[Fig. 10. Comparing movement across NYC transportation hubs. On the top, we refine the query to compare trips starting at JFK on most days. Another interesting question is where passengers go. The choropleth (Fig. 11) shows that most people go to Midtown (the darkest region), followed by the Upper West Side.]

[Ferreira et al., 2013]
Baseball Data

[Deitrich et al., 2014]
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?

<table>
<thead>
<tr>
<th>Data Acquisition</th>
<th>Data Analysis</th>
<th>Data Curation</th>
<th>Data Storage</th>
<th>Data Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Structured data</td>
<td>• Stream mining</td>
<td>• Data Quality</td>
<td>• In-Memory DBs</td>
<td>• Decision support</td>
</tr>
<tr>
<td>• Unstructured data</td>
<td>• Semantic analysis</td>
<td>• Trust / Provenance</td>
<td>• NoSQL DBs</td>
<td>• Prediction</td>
</tr>
<tr>
<td>• Event processing</td>
<td>• Machine learning</td>
<td>• Annotation</td>
<td>• NewSQL DBs</td>
<td>• In-use analytics</td>
</tr>
<tr>
<td>• Sensor networks</td>
<td>• Information extraction</td>
<td>• Data validation</td>
<td>• Cloud storage</td>
<td>• Simulation</td>
</tr>
<tr>
<td>• Protocols</td>
<td>• Linked Data</td>
<td>• Human-Data Interaction</td>
<td>• Query Interfaces</td>
<td>• Exploration</td>
</tr>
<tr>
<td>• Real-time</td>
<td>• Data discovery</td>
<td>• Top-down/Bottom-up</td>
<td>• Scalability and Performance</td>
<td>• Visualisation</td>
</tr>
<tr>
<td>• Data streams</td>
<td>• ‘Whole world’ semantics</td>
<td>• Community / Crowd</td>
<td>• Data Models</td>
<td>• Modelining</td>
</tr>
<tr>
<td>• Multimodality</td>
<td>• Ecosystems</td>
<td>• Human Computation</td>
<td>• Consistency,</td>
<td>• Control</td>
</tr>
<tr>
<td></td>
<td>• Community data analysis</td>
<td>• Curation at scale</td>
<td>Availability,</td>
<td>• Domain-specific usage</td>
</tr>
<tr>
<td></td>
<td>• Cross-sectorial data analysis</td>
<td>• Incentivisation</td>
<td>Partition-tolerance</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Automation</td>
<td>• Security and Privacy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Interoperability</td>
<td>• Standardization</td>
<td></td>
</tr>
</tbody>
</table>

[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.

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.

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>daytype</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>40760</td>
<td>Granville</td>
<td>01/01/2001</td>
<td>U</td>
<td>1,050</td>
</tr>
<tr>
<td>40075</td>
<td>Jackson/Dearborn</td>
<td>01/01/2001</td>
<td>U</td>
<td>649</td>
</tr>
<tr>
<td>40660</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>41720</td>
<td>East Eltm Cottage Grove</td>
<td>01/01/2001</td>
<td>U</td>
<td>391</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>23 Gronnville-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/State</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>Skokie</td>
<td>01/01/2001</td>
<td>U</td>
<td>0</td>
</tr>
<tr>
<td>40450</td>
<td>95th/Oan Ryan</td>
<td>01/01/2001</td>
<td>U</td>
<td>2,948</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>Pilsen-Cermak</td>
<td>01/01/2001</td>
<td>U</td>
<td>0</td>
</tr>
<tr>
<td>40660</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>40130</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>
Cool Machine Learning Model & Pretty Visualizations

OVERALL TRAIN SYSTEM BOARDINGS

<table>
<thead>
<tr>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>229.1 million</td>
<td>238.1 million</td>
</tr>
<tr>
<td>2014</td>
<td>241.7 million</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>203.7 million</td>
<td>201.8 million</td>
</tr>
</tbody>
</table>

Note: Boardings include transfers.

BUS RIDERSHIP

OVERALL BUS SYSTEM

<table>
<thead>
<tr>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>380.1 million</td>
<td>276.1 million</td>
</tr>
<tr>
<td>2014</td>
<td>274.3 million</td>
<td></td>
</tr>
</tbody>
</table>

LINES WITH THE LARGEST INCREASES

- Union/Powell/King Express (362)
- West Side (616)

LINES WITH THE LARGEST DROPS

- Aurora Express (500)
- Lincoln (111)

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) 1.0%
3. Blue Line (O'Hare) 9.9%
4. Brown Line 0.9%
5. Green Line (Ashland/3rd Branch) -11.1%
6. Green Line (East 56th 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 ( Evanston) 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 11 and Oct. 30, 2015, there was no train service between Dempster–Skokie and Howard because of repairs after an embankment collapse.
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

D. Koop, CSCI 640/490, Spring 2024

C. Bronner (left), Chicago Business (right)
Lots of topics related to this
Data Wrangling

D. Koop, CSCI 640/490, Spring 2024
# Data Wrangling

## Trifacta

### D. Koop, CSCI 640/490, Spring 2024

![Trifacta](image)

**Customer Analysis** > **Customer**

**Preview**

<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>3107084351</td>
<td>Jan 2013 - Dec 2016</td>
<td>Jan 2000 - Dec 2014</td>
<td>0 - 15</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>3103121111</td>
<td>6/4/16</td>
<td>7/29/89</td>
<td>2 Categories</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>31018226812727</td>
<td>3/28/15</td>
<td>18/6/13</td>
<td>1</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>31095863825407</td>
<td>9/23/16</td>
<td>1/9/87</td>
<td>7</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>31005433849230</td>
<td>5/29/15</td>
<td>2/14/81</td>
<td>13</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>310826937211005</td>
<td>9/11/15</td>
<td>9/18/10</td>
<td>4</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>31062914646952</td>
<td>8/27/15</td>
<td>3/13/86</td>
<td>8</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>31018448724861</td>
<td>7/16/16</td>
<td>5/11/84</td>
<td>8</td>
<td>ACTIVE</td>
</tr>
</tbody>
</table>

**Pattern Details**  
**CONTRACT_END**

**Hide Example Values**

- 9/10/15  
- 6/13/15  
- 5/21/15  
- 12/13/15  
- 11/16/16  

**Data Wrangling**

19 Columns  
20,000 Rows  
8 Data Types

Show only affected  
Rows

---

D. Koop, CSCI 640/490, Spring 2024
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': 'ibm',
- 'ibm t.j. watson research center': 'ibm',
- 'ibm t.j.watson research center': 'ibm',
- 'ibm thomas j. watson research center': 'ibm',
- 'ibm t.j.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',
- 'ml microsoft research': 'microsoft',

Carnegie Mellon University
Microsoft
Stanford University

ICML
NIPS
EMNLP
NAACL
EACL
ACL

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

D. Koop, CSCI 640/490, Spring 2024

[NLP Publishing Stats, M. Rei & R. Allen]
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. In 1996, she released her second album, Hoy Es Marta, which featured the singles "Corazón de Bombón", "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 "Rumbal", "Eres" and "Amnesia". "Rumbal", 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 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...
Anahí Giovanna Puente Portilla (born 14 May 1983), known professionally 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 Damían'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.[1] 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.[2] In 1996, she released her second album, Hoy Es Marta, which featured the singles "Corazón de Bombón", "Por Volverte a Ver" and "Descontróndote". 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 "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: 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 Enaguadas (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 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 Bombón", "Por Volverte a Ver" and "Descontrólante". She went on to record two additional albums titled Anda and 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 "Rumbal", "Eres" and "Amnesia". "Rumbal", 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 best-selling RBD member in the United States. In 2010, she released a compilation album titled, ÁnBHAnahí, which featured the singles "Entretenimiento" and "Anahí: Dividida". They have two sons: Manuel, born on 17 January 2017 and Emiliano, born on 2 April 2018.

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...
Anahí Giovanna Puente Portilla, born May 14, 1983, as 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 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 Mariana?, which featured the singles "Corazón de Bombón", "Por Volverte a Ver" and "Desconstrómate". 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.
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</td>
<td>Length</td>
<td>2,320 mi</td>
</tr>
<tr>
<td>KG</td>
<td>Missouri</td>
<td>Length</td>
<td>2,341 mi</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Mississippi</td>
<td>Length</td>
<td>2,202 mi</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Missouri</td>
<td>Length</td>
<td>2,341 mi</td>
</tr>
<tr>
<td>USGS</td>
<td>Mississippi</td>
<td>Length</td>
<td>2,340 mi</td>
</tr>
<tr>
<td>USGS</td>
<td>Missouri</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>

Fact's true value

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

Source reports a value for a fact

Conflicting value

[Fact: Source reports a value for a fact]

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

SQL DATABASES

NoSQL DATABASES

Relational

Column

Graph

Key-Value

Document
Distributed Databases

Site 1

DB 1

Site 2

DB 1

DB 2

Site 3

DB 1
CAP Theorem

Consistency

Partition-Tolerance

Availability

redis
mongoDB
HBASE
Amazon Redshift
MySQL
Vertica
SAP S/4 HANA
neo4j
PostgreSQL
DynamoDB
Cassandra
CouchDB

[B. Aunkofer]
Scaling Dataframes

- New Data Source
- New spec
- New requirements

- Prototyping
- Exploring
- Testing
- Production

Laptop/Workstation | Small Cluster | Large Cluster

Feedback

[D. Petersohn]
Representing Time Series Data

![Diagram of data compression algorithm]

a) Data stream:
- March 24, 2015 02:01:02 Value: 12
- 02:02:02 Value: 12
- 02:03:02 Value: 24

b) Compressed data:
- Header: March 24, 2015 02:00:00
- Bit length: 64, 14, 64, 9, 1, 1, 2 + 5 + 6 + 1

- N-2 timestamp: 02:00:00
- N-1 timestamp: 02:01:02 Delta: 62
- Timestamp: 02:02:02 Delta: 60

- Delta of deltas: -2

<table>
<thead>
<tr>
<th>Previous Value</th>
<th>12.0</th>
<th>0x4028000000000000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>24.0</td>
<td>0x4038000000000000</td>
</tr>
<tr>
<td>XOR</td>
<td>-</td>
<td>0x0010000000000000</td>
</tr>
</tbody>
</table>

11 leading zeros, # of meaningful bits is 1

[Pelkonen et al.]
Representing Graph Data

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

**Adjacency Matrix**
- An \( n \times n \) matrix
- A 0 indicates no edge
- A 1 indicates an edge

**Adjacency List**
- The number of all elements in the adjacency data structure: \( 2m \) (undirected graph) and \( m \) (directed graph)
- Neighbor-edges can be arbitrary structures, e.g., arrays or lists
- Neighborhoods can be sorted or unsorted
- Unweighted graph: a cell is one bit
- Weighted graph: a cell is one integer

**Edge List (sorted & unsorted)**
- One tuple corresponds to one edge
- Pointers from vertices to their neighborhoods
- Number of tuples: \( 2m \) (undirected), \( m \) (directed)
Data Cubes

Time
- 2015 May
- 2015 Apr
- 2015 Mar
- 2015 Feb
- 2015 Jan

Product
- Accessories
- Bikes
- Clothing
- Components

Measures
- 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
Spatial Data
Provenance and Reproducibility

DATA

Data Management

Computation

Visualization

Paper

Our research has been funded by the National Science Foundation grants. The workflow is available at http://www.paraview.org. The original workflow is available at http://www.research.ibm.com/dx. The design and realisation of the virtual research environment for social sharing of workflows is an image that was generated by embedding a workflow diagram. For more details, see the blog at http://www.swivel.com. This is an image that was created using VIEGAS.
Provenance and Reproducibility

DATA

Data Management

Computation

Provenance

Visualization

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/
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/cs640-2024sp/
  - 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 (5-6)
• 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

- Polars:
  - Pandas in Rust, more efficient, different syntax…

- DuckDB
- Ibis
- Trifacta Wrangler
- More…
Office Hours & Email

• Scheduled office hours:
  - M: 1:45-3:00pm, W: 10:45am-12:00pm, 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

• Review of Python
• Download and install anaconda distribution (Python 3.11 or 3.12):
  - https://github.com/conda-forge/miniforge#miniforge3