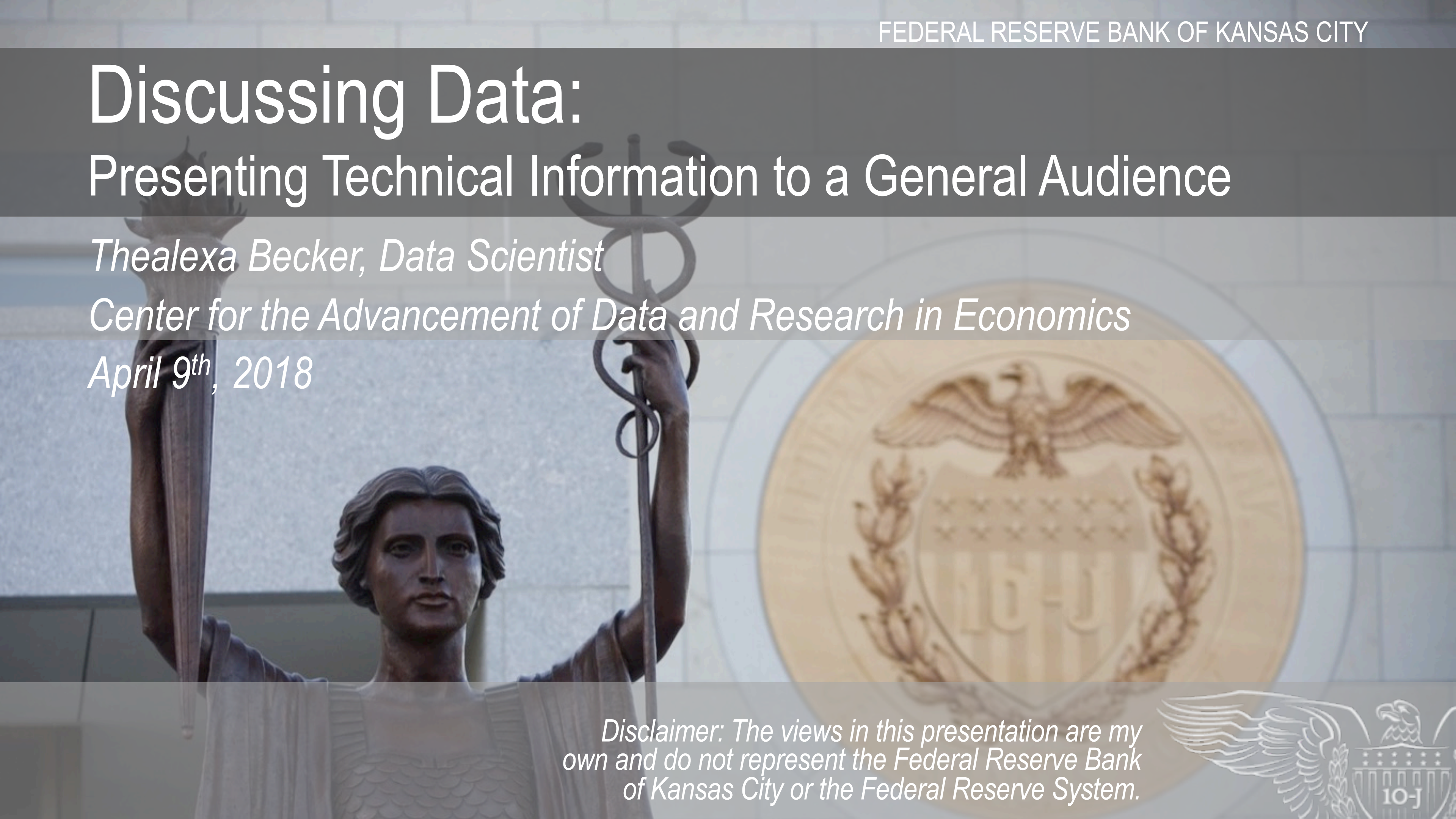


FEDERAL RESERVE BANK OF KANSAS CITY

Discussing Data:

Presenting Technical Information to a General Audience

Thealexa Becker, Data Scientist

Center for the Advancement of Data and Research in Economics

April 9th, 2018

Disclaimer: The views in this presentation are my own and do not represent the Federal Reserve Bank of Kansas City or the Federal Reserve System.

Overview

1. Data Fluency
2. Origins of the Data Museum
3. The Data Museum Today
4. The Future of the Data Museum

What is Data Fluency?

The ability to speak about, write about, and visualize data in a clear, coherent way.

- Think about this as learning a language…
  - Recognizing indivual words -> understanding specific data sets
  - Reading a newspaper -> understanding someone else’s data product
  - Conversational -> discussing someone else’s data product with them
  - Fluent -> being able to communicate your own ideas verbally, visually, and orthographically.

Why is this important?

- Data are **everywhere.**
- Understanding data can be arduous…
- ...Communicating data can be even harder.
- New standards for graphics and charts.
- High-skill jobs are becoming more data intensive – all different technical professions need to be able to talk to each other in a meaningful way.

Who should care about being data fluent?

- Two questions:

  1.Do you work with data?

  2.Do you interact with data from outside sources (news, etc.)?

- If you answered ”yes”, then you need to be data fluent!!
  - The next frontier for high-tech employers is interdisciplinary skills and data fluency – they need someone who can do the work and then tell other people what work they have done.

Data fluency is not a monolith

- Data Fluency looks different for different people in different roles. It depends on…
  - …who your audience is.
  - …what data you use.
  - …what data you consume.
- The goal is to be knowledgeable about the data that you and others in your organization work with.
  - Everyone is an ambassador for the work done by their department, and the more confidently you can talk about it the better.

Get to know your data

- Core characteristics of data that all users should know.
- Collate that information into a short summary of data to better explain to new users or a presentation/paper audience.
  - Think of this as an “elevator pitch.”
- Once you know your data, learn to compare similar data sets.
- Goal is to build a foundation of knowledge about data sets that are commonly used in any line of business.
- In CADRE, one application of this data fluency is the Data Museum.

The Elevator Pitch

- Explain your data efficiently and effectively to a general audience.
- Aim to create a pitch that can be delivered in 30-45 seconds in a presentation or brief paragraph in writing.
- Goal is to highlight seven broad characteristics of your data.
- Use the Data Worksheet to fill out this info.

Seven Core Characteristics

1. Sample Population or Universe
2. Collection Method
3. Frequency and Timeframe
4. Notion of Dimension
5. Purpose and Main Content
6. Access and Use Information
7. Producer and Publisher

Example: Current Population Survey

The Current Population Survey is a monthly address-based survey of U.S. households that gathers geographic, demographic, and employment status information. These data are used to calculate several national indicators of employment, such as the unemployment rate and labor force participation. In its current iteration, data on over 150,000 respondents are collected monthly to create around 400 variables. The CPS survey data are collected by the Census on behalf of the Bureau of Labor Statistics and are publically available dating back to 1976.

Comparing Data

- Now that you can describe your data, how do you compare datasets?
- Data must have at least one characteristic in common in order to compare.
- Data Comparison worksheet:
  - Start with Data Profile foundation
  - List seven characteristics side by side
  - See which dimension is the source of the most difference
  - Dive deeper

Example: Cleveland Sportsball

- Cleveland has three major league teams: Browns (football), Indians (baseball), and Cavaliers (basketball).
- Question: What is the best major league sports franchise in Cleveland?
  - Hint: If you say the Browns, you are clearly not a sports fan…or you're super duper optimistic.
- How can we compare the data on the teams?

Disclaimer: I am not a sports person, please don't be offended if I overly simplify something for the sake of an example.

Example: Cleveland Sportsball - Browns

- Sample Population or Universe
  - Players on the Browns and their performance during active play.
- Collection Mechanism
  - Observation of players/team during games.
- Frequency and Timeframe
  - 16 games in regular season (also post-season, but who are we kidding).
  - Browns have been in the NFL since 1950.

Sportsball - Browns

- Notion of Dimension
  - Offensive/defensive exclusive players.
  - Defensive: tackles/sack, penalties, interceptions…
  - Offensive: attempts, yards per carry, completed pass, fumbles…
  - Special players: Quarterback, kicker
  - Players: 11 players in a game
  - Scoring: Touchdowns, P-A-T, field goals
  - Game: 4 Quarters (60 minutes), clear separation of offense/defense play
- Purpose and Main Content
  - Want to track the performance of players and team during game play.
  - Contains statistics and metrics about each player on the roster and their performance during football games.
- Access and Use
  - Data can be found on the Browns website.
- Publisher/Producer
  - Cleveland Browns franchise

Example: Cleveland Sportsball - Indians

- Sample Population or Universe
  - Players on the Indians and their performance during active play.
- Collection Mechanism
  - Observation of players/team during games.
- Frequency and Timeframe
  - 162 games in regular season (also post-season, if applicable)
  - Indians have been in the MLB since 1901.

Example: Cleveland Sportsball - Indians

- Notion of Dimension
  - Players have offensive and defensive stats
  - Offensive: Batting average, RBI, On-base percentage, slugging…
  - Defensive: Fielding, errors, chances, put-outs…
  - Special players: pitchers, designated hitter
  - Players: 9 players (offense and defense)
  - Scoring: Runs scored (home runs, RBI)
  - Game: 9 innings (no time constraint), clear separation of offense/defense play
- Purpose and Main Content
  - Want to track the performance of players and team during game play.
  - Contains statistics and metrics about each player on the roster and their performance during baseball games.
- Access and Use
  - Data can be found on the Indians website.
- Publisher/Producer
  - Cleveland Indians franchise

Example: Cleveland Sportsball - Cavaliers

- Sample Population or Universe
  - Players on the Cavs
- Collection Mechanism
  - Observation of players/team during games.
- Frequency and Timeframe
  - Possible 82 games in regular season (possible post-season)
  - Cavs have been in the NBA since 1970.

Example: Cleveland Sportsball - Cavaliers

- Notion of Dimension
  - No offensive/defensive separation of players, stats on both
  - Offense: Attempts made, field goals, rebounds, assists…
  - Defense: defensive proficiency…
  - Special players: none
  - Players: 5 on court at a time
  - Scoring: baskets (2, 3-pts), free throws (1-pt)
  - Game: 4 quarters (48 minutes), rapid offense/defense switch
- Purpose and Main Content
  - Want to track the performance of players and team during game play.
  - Contains statistics and metrics about each player on the roster and their performance during basketball games.
- Access and Use
  - Data can be found on the Cavs website.
- Publisher/Producer
  - Cleveland Cavs franchise

Foundations of the Data Museum

- Cannon (2015) provides the arguments for and outline of a Data Museum.
- Way to make finding and learning about data easier for researchers.
- Floor Plan for Museum and Exhibits
  - Identification
    - Purpose, description of collection mechanism, terms of use
  - Relevance
  - Access and Use
  - Supplementary Information

Mission Statement

The goal of the Data Museum is to act as a stepping stone for new users of data sets that have a learning curve. In many cases, important and oft used data sets may be available to the public, but using them is a burden for researchers due to dense technical documentation or difficulty compiling the data into a useable form, among other concerns.

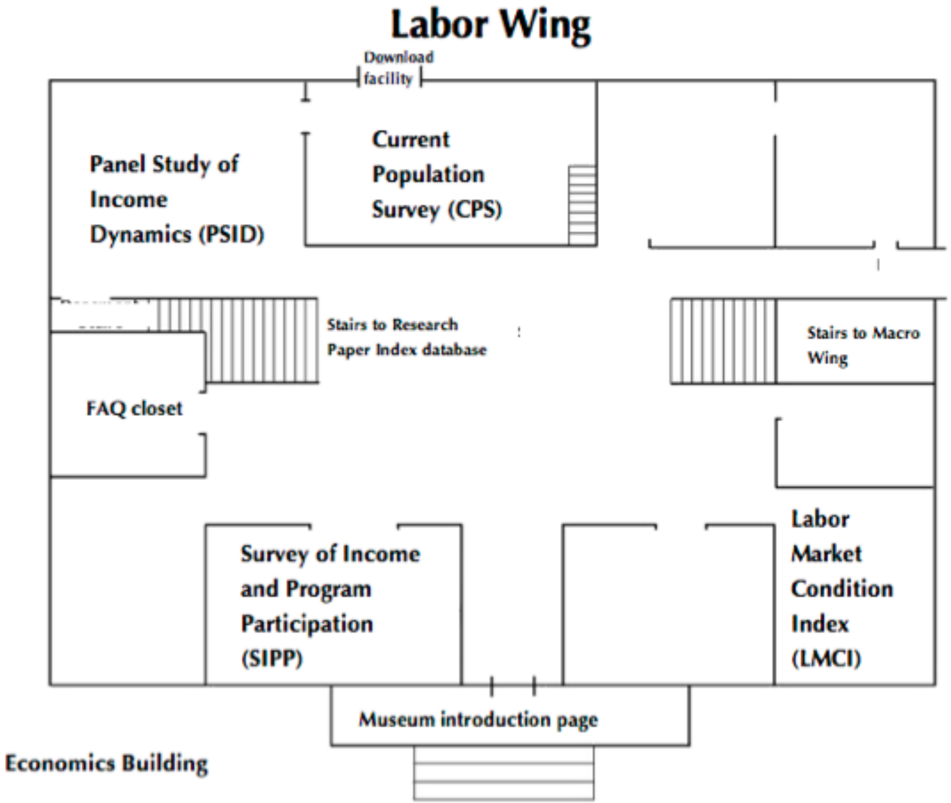
History of the Data Museum at KC Fed

- The concept of a Data Museum was well defined, needed a test subject.
- The Current Population Survey (CPS) is the most frequently used micro data set in the ER Department.
- CPS was identified at a conference on the data as a good starting point for several reasons:
  1. Frequently used, many resources available.
  2. Widely known, good opportunity for feedback.
  3. Publicly available.
- Staff assigned to start project (one Data Scientist, one Data Engineer).

The CPS is a good starting candidate

- The source for the monthly Employment Situation Report.
- Foundation for a good chunk of labor research at KC Fed.
- Provides a wealth of geographic and demographic information at a high frequency for representative sample of U.S.
- Because it’s a “known” entity, opportunity to build a template for museum.

A Guided Tour of the Data Museum


Labor Wing
Download facility
Panel Study of Income Dynamics (PSID)
Current Population Survey (CPS)
Stairs to Research Paper Index database
Stairs to Macro Wing
FAQ closet
Survey of Income and Program Participation (SIPP)
Labor Market Condition Index (LMCI)
Museum introduction page
Economics Building

Survey of Income & Program Participation (SIPP)

- CADRE had a SIPP conference in 2017, active user base.
- Large panel data set with wealth of information on social program participation.
- More difficult to work with than other data.

The Survey of Income and Program Participation (SIPP) is a continuous series of nationally representative address-based panels of U.S. households that range from 2.5 to 4 years in length. Beginning in 1984, these panels collect details on demographic, geographic, labor force, income, and social program participation information of respondents. The latest iteration of the survey contains around 53,000 households who provide data on a monthly frequency that generates over 1000 variables for analysis. The SIPP data are collected by the Census and are publically available.

SIPP vs. CPS

- SIPP and CPS seem very similar at first blush.
  - Both public, collected by Census, labor data, demographics, geographic info.
- They differ fairly significantly in three ways:
  1. Purpose: SIPP wants to get at program participation, CPS wants snapshots of labor market.
  2. Collection Mechanism: SIPP surveys participants for entire panel (~4 yrs), CPS for only 16 month period.
  3. Notion of Dimension: Survey structure and difference purpose causes SIPP to have over 1000 variables, CPS has only around 400.

Categorically Linked Timeline (CLINT)

- It began with a green notebook…
- A database of economic events that are mirrored in economic time series and affect movements of those series.
- First (and only) data of its kind.
  - Great deal of interest from places like SF Fed, Bloomberg, Pew Research, U. of Colorado, Claremont, George Washington University.

Dataset Utilization Open Source (DUOS)

- Collaboration with colleagues at Loyola with a Sloan Foundation grant.
- Goal is to help identify research that uses a given dataset by automatically extracting information from the source.
- Use cases: Literature review, dataset utilization, approaching new datasets, generating new research applications.

Benefits of the Data Museum

- To the public:
  - Lowers barriers to entry for data use.
  - Encourages use of data that were previously more difficult to access.
  - Lowers training time for new users or employees who would need data for research.
  - Makes use of data quicker and more efficient.
- To CADRE:
  - Gathers institutional knowledge.
  - Provides data dissemination services to a broader audience.
  - Demonstrates expertise in data and content curation.

What the Data Museum is NOT

- A replacement for other access options.
- A data discovery tool.
- A host for modified data used in a particular research project.
- A host for data that is confidential or exclusive in some way.

How do we compare?

- A few other sites offer similar data access.
- IPUMS (U. Minnesota Population Center):
  - How are they similar?: Offer a platform to access CPS/ACS data with documentation. Can analyze data online.
  - How are they different?: They synchronize data to Census Population Tables. Documentation is also more researcher use based than educational.
- National Bureau of Economic Research (NBER):
  - How are they similar?: Provide access to CPS/SIPP data files with some links to documentation.
  - How are they different?: Links are to technical documentation, which is impenatrable for average user. Data files are not cleaned up.
- Census (Data Ferrett):
  - How are they similar?: Provides access to all Census data with technical documentation.
  - How are they different?: Documentation not user friendly, nor is platform to pull data. You need to download a client for Data Ferrett, not for Data Museum.

The Future of the Data Museum

- The CPS Museum Room was a “pilot” program, or proof of concept.
  - Established a work flow for staff.
  - Provided concrete illustration of Cannon (2015) formula for creating a room.
  - Ideas for expanded content (Technical Working Papers).
  - Was a successful pilot, good feedback.
- Where does the Data Museum go from here?
  - Addition of CLINT Content
  - Develop SIPP Room
  - Fall 2018 “Grand Opening”
  - Data Visualization
  - Expansion Plan
  - External Collaboration Opportunities

The Future - SIPP and CLINT

- CLINT is almost ready to go!
  - Data access platform is beta testing.
  - Summer intern will help fine tune data before launch.
  - Content is being organized for the website.
- SIPP Room
  - The room nominally exists and active curation is starting.
  - Possibility to adopt new techniques for curating the data.
  - No one else has done this work.
  - Research project underway using SIPP data.
    - Feature selection project, looks at identifying smaller datasets for users depending on their analysis needs.

The Future – Fall 2018 Launch

- CPS Room ”launched” last May
- Continued improvements to website.
- CLINT content available by end of summer.
- SIPP Room will have additional documentation, but not data (yet).
- A few smaller galleries to complement larger rooms.

The Future – Data Visualization

- Data Visualization is becoming an important focus.
  - Increased Tableau, Rshiny use.
  - Integration with data science.
- Use CPS Room to pilot some data viz of our own.
  - Summer intern with data viz experience will help design some features for the CPS.
  - Experts in this area, opportunity to see how and where data viz enhances existing content.
  - Develop format for future data viz use in other galleries.

The Future – Expansion Plans

- Three avenues for expansion of the Data Museum:
  - Expansion plans will begin as early as Q2 2018.
- Three broad categories of exhibits:

1. Main galleries
   - These are rooms like the SIPP, CPS, CLINT, that are managed by CADRE staff.
   - Large in size, will feature access to the data.
2. Smaller galleries
   - Rooms featuring other publicly available data sets but do not provide access to the data.
   - Documentation and education focused.
3. External collaborations
   - Opportunity to work with other organizations where they curate rooms on data they have expertise in.

The Future – Small Gallery Expansion

- Collaboration between Data Science and Library teams in CADRE.
- Designed to be short term investments with a big payoff .
- Smaller rooms – pilot room is JOLTS data.
- How it works:
  - CADRE staff identify data sets.
  - Library staff conduct literature review and metadata.
  - Data Science staff create or expand on documentation for the data set’s use.

The Future – External Collaboration Expansion

- The Data Museum invites and encourages external collaboration to capitalize on their expertise.
- How it would work:
  - Connect with an external collaborator (associated with an institution, including KC Fed).
  - Collaborator identifies a data set they would like to create a room for.
  - Collaborator is responsible for all content generated and for any updates and additions to content.
  - CADRE staff will manage updates to Museum Room.
- On our minds…
  - Patent and Trademark
  - Call Report Data
  - American Community Survey or other large labor data sets.

Conclusion

- The Data Museum is an innovative and unique project that increases transparency, education about data, and informed use of public data.
- The CPS Room is a strong foundation that provides a blueprint for meaningful and efficient expansion.
- Gearing for a Fall 2018 Grand Opening with additional content from the SIPP and CLINT.

- Thank you! Questions?

References

Cannon, San. "Content Curation for Research: A Framework for Building a “Data Museum”." International Journal of Digital Curation. Vol 10, no. 2 (2015). doi:https://doi.org/10.2218/ijdc.v10i2.355.

Gemignani, Zach, Chris Gemignani, Richard Galentino, and Patrick Schuermann. *Data Fluency: Empowering Your Organization with Effective Data Communication*. Indianapolis, IN: Wiley, 2014.