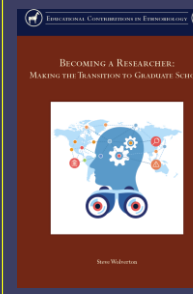
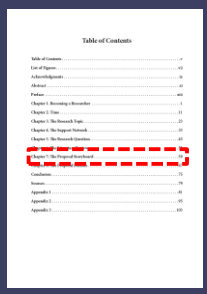
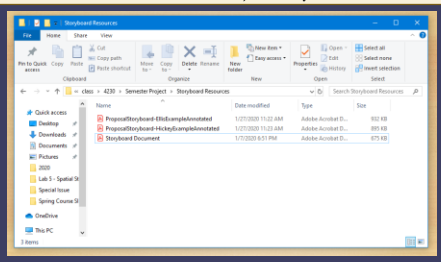




Check out the "Class > 4230 > Semester Project > Storyboard Resources" Folder



In summary, the first phase of your project is to create a storyboard document that gives your take on the class project and summarizes the key elements in your thinking that collectively form a plan for you to successfully complete your work and solve the central project problem(s).

Your storyboard should include these specific sections, with details as defined in this chapter:

1. **An introduction that defines the basics of the**
   a. Topic
   b. Research Questions
   c. Significance of the Research Questions
   d. Consequences of Not Doing the Work
   e. Preview of Specific Objectives Needed to Complete the Project
2. **A statement of your study objectives**
   a. Note that the objectives here are goals, not plans
3. **Definition of your proposed research methods**
   a. Study Area
   b. Data Requirements
   c. Analytical Methods
4. **An overall sketch of your research plan**
   a. Tasks
   b. Deadline Dates for Accomplishing Each Task

We will discuss your KFC project further on Thursday

Aim: continuing to develop the ideas we're discussing today

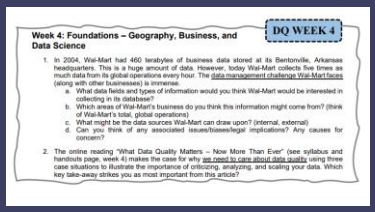
Please check out the revised syllabus, available on our course website

- Has an updated schedule reflecting the minor shuffle of topics we discussed in class last week

Also please sign up to reserve your project region

- Largest regions (Central and Southeast)
  - Set aside for pairs projects
- All other regions (Northeast, West)
  - For single student projects

And lastly, note that your week 4 DQ set is due today



Week 4: Foundations – Geography, Business, and Data Science — DQ WEEK 4

1. In 2004, Wal-Mart had 460 terabytes of business data stored at its Bentonville, Arkansas headquarters. This is a huge amount of data. However, today Wal-Mart collects five times as much data from its global operations every hour. The data management challenge Wal-Mart faces (along with other businesses) is immense.
   a. What data fields and types of information would you think Wal-Mart would be interested in collecting in its database?
   b. Which areas of Wal-Mart's business do you think this information might come from? (think of Wal-Mart's total, global operations)
   c. What might be the data sources Wal-Mart can draw upon? (internal, external)
   d. Can you think of any associated issues/biases/legal implications? Any causes for concern?
2. The online reading "What Data Quality Matters – Now More Than Ever" (see syllabus and handouts page, week 4) makes the case for why we need to care about data quality using three case situations to illustrate the importance of criticizing, analyzing, and scaling your data. Which key take-away strikes you as most important from this article?

Foundations: Geography, Business, and Data Science

Week 4

UNT UNIVERSITY OF NORTH TEXAS Discover the power of ideas

Big Picture for This Week

- **Overall theme:** exploring the connections that exist between geography, business, and data analysis practices employed in business
- To begin to develop these connections, this week we will
  - 1. Explore the connection between geographic analysis & business needs
  - 2. Define the principles of data science, big data, and data mining as foundations for business geographic analysis

Note: Class Discussion for this Week

- This week our discussion will be organized into two segments
  - ***1. Today:*** cover crucial connections between **data science** and **geography** and **business**
  - ***2. Thursday:*** general discussion of what these connections mean for your work on the **major semester project**
    - We will also set aside time on Thursday to address **broader issues** you might have with understanding the project and its two phases: bring other **project-focused questions**

- To clear out time for discussion on Thursday, I am leaving a section of this week's presentation content for your personal review only:
  - The section introduction slide at right marks the beginning of this content (slide 59 in this presentation)

Further Data Analytic Concepts

Data Acquisition and Data Quality Content for Personal Review and Reflection

Geography and Business

- Based on all of our discussion and GIS work up to now, we can see that geography is a central consideration for creating solid **business** plans
  - Recall in particular our week 1 discussion of location intelligence (LI) and its relationship with business intelligence (BI)
- General principle that applies to both LI and BI:
  - ***Data* and the *capability to extract useful knowledge from data*** should be regarded as key strategic assets for any business (leading to potential competitive advantage)

Geography and Business

- Great example: Wal-Mart's response to Hurricane Frances in 2004


The New York Times

BUSINESS DAY

What Wal-Mart Knows About Customers' Habits

HURRICANE FRANCES was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons, something that the company calls predictive technology.

Geography and Business

- Great example: Wal-Mart's response to Hurricane Frances in 2004
  - A week before the storm made landfall, Wal-Mart decided to try data analysis to improve their response to a wide range of emergency situations
  - Idea: **(1)** create sales forecasts based on a previous storm, **(2)** start predicting what's going to happen, instead of waiting for it to happen
  - **Q:** what role might geography play in this kind of analysis?

Geography and Business

- Great example: Wal-Mart's response to Hurricane Frances in 2004
  - At the time of the storm, Wal-Mart had 460 terabytes of business data at its Bentonville headquarters (at the time, the entire internet stored less than half that much)
  - **Side Note:** Today, Wal-Mart collects 2,500 terabytes of business data every hour

Geography and Business

- Great example: Wal-Mart's response to Hurricane Frances in 2004
  - At the time of the storm, Wal-Mart had 460 terabytes of business data at its Bentonville headquarters (at the time, the entire internet stored less than half that much)
  - **Q:** What kinds of data do you think this Wal-Mart database included?
    - Relating to which areas of Wal-Mart's business? (think of Wal-Mart's total operations)
    - What data sources? (internal, external)
    - Associated issues/biases/legal implications? Any causes for concern?

Geography and Business

- Great example: Wal-Mart's response to Hurricane Frances in 2004
  - Wal-Mart used its analysis of its databases to gain multiple insights
    - Products with *unusual* local demand before/after storm: not just obvious items (flashlights)
      - Top sellers it identified: Strawberry Pop-Tarts, Beer
    - Also, used GIS to track storm path and identify stores in likely impact area
    - Formulate a plan to stock those stores with needed items immediately following landfall

Much of the analysis and response had a direct connection to geography and location

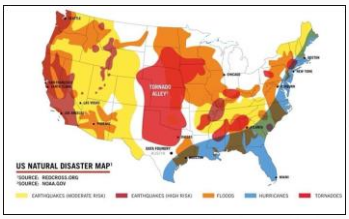
Wal-Mart's experience with Hurricane Frances in 2004 provided them with valuable experience they were able to use to great advantage with the coming of the even-more-destructive Hurricane Katrina in 2005.

Benefits:

- Strategic positioning for post-storm sales of needed supplies
- Immense goodwill from communities that were well-served by the retailer following the storm

Before we leave this topic completely, it is worthwhile to consider the variety of environmental hazards that face organizations of all kinds.

Q: What is the role for geographic analysis in planning for response to each of the disaster types represented here?



Geography and Business

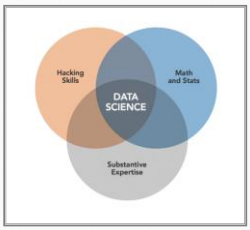
- Incorporation of geographic concepts, data, and analysis into standard operating procedures gives businesses an advantage over competitors who do not
  - Part of a truly "Data-Driven Decision-making" ("DDD") culture: the inclination to base business decisions on *analysis of objective data* rather than primarily on *intuition* and *experience*

Video Case Study: Organizational Application of Data Analysis

- Another case study is helpful in clarifying the organizational importance of effective analysis of a broad range of data sources
  - Video case study: geospatial data analysis at the University of Minnesota
  - Watch for the different datasets analyzed and the applications made possible through the use of GIS technology
  - **Q:** So based on the video, what uses of GIS technology did you see? How important are these applications?

Need for a Broad Data Discussion

- To go any further in developing a data-rich, location-aware perspective on business, we need to have a thorough understanding of data and issues associated with data use
- If we can't be confident about our data, we can't be confident about anything else we might do
- **Q:** what might be reasons to have a lack of confidence in our data?



Data Science

Developing a Basic Understanding

Data Science

- A modern framework for effective data use is provided by the emerging field of data science
  - **Q:** Data science is a common term that has grown much in popularity and acceptance in business and engineering circles, but what is it?
    - Have you encountered data science in your studies so far?
    - What does data science encompass?

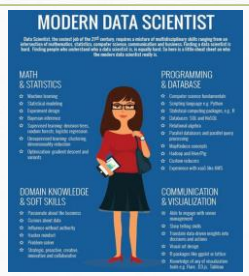
Data Science

- A modern framework for effective data use is provided by the emerging field of data science
  - There is no single, authoritative definition of "data science", but there is some convergence around basic answers to three basic questions: ***"Data science is…"***
    - **Doing What?** Identifying patterns and regularities
    - **Where?** In data of all sorts
    - **Why?** To advance scholarship, improve the human condition, and create commercial and social value

Data Science

- Thus "data science" provides an overall term that delimits a broad set of tools and approaches aimed at making sense of data for the good of society
- Data science has emerged because of a unique challenge
  - Businesses today are accumulating new data at a rate that far exceeds their capacity to extract value (think back to what we said about Walmart)
  - The question facing every organization: how to use data effectively
    - Not just their own data, but all data that are available and relevant to the organization's aims

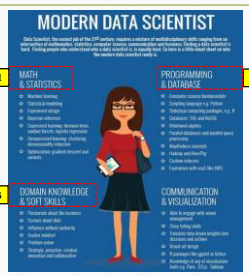
Another view of data science from the perspective of what a "data scientist" job entails



Data Science Central: @analyticbridge

Another view of data science from the perspective of what a "data scientist" job entails

Three elements of data science that are reasonably well-agreed-upon (we will see these again in a minute)

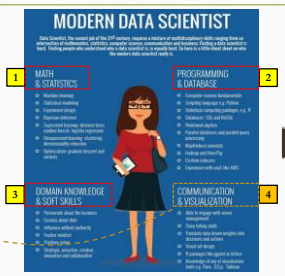


Data Science Central: @analyticbridge

Another view of data science from the perspective of what a "data scientist" job entails

Three elements of data science that are reasonably well-agreed-upon (we will see these again in a minute)

A fourth competency that I would also argue is central to the successful practice of data science (and location intelligence)



Data Science Central: @analyticbridge

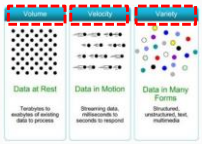
Aside: "Data Science" vs. "Big Data"

- Today we hear a lot about both **data science** and **big data**, and often the terms are used interchangeably
- **Q:** But are data science and big data the same thing?
  - **Short answer:** no
  - **Longer answer:** data science and big data are distinctive, but they can overlap

What is "Big Data"?

One conceptualization of the components that make "big data" unique: summarizing the field as the "4 Vs"

- Three Commonly Accepted Vs: Volume, Velocity, Variety



Data Science Central: @DataScienceCtrl

What is "Big Data"?

One conceptualization of the components that make "big data" unique: summarizing the field as the "4 Vs"

- Three Commonly Accepted Vs: Volume, Velocity, Variety
- One Addition: Veracity (getting at the idea of questioning your data; issue of data quality)

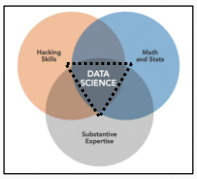
Data Science Central: @DataScienceCtrl

Aside: "Data Science" vs. "Big Data"

One more "Venn Diagram" definition of that combines our three previous Data Science elements

1. Hacking (Programming/Coding) Skills
2. Math and Stats Knowledge
3. Substantive (Domain Knowledge) Expertise

This diagram indicates that data science exists at the intersection of these three key elements

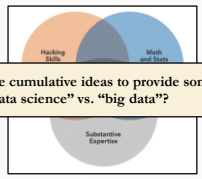


Aside: "Data Science" vs. "Big Data"

One more "Venn Diagram" definition of that combines our three previous Data Science elements

So how could we use these cumulative ideas to provide some comparison of "data science" vs. "big data"?

This diagram indicates that data science exists at the intersection of these three key elements



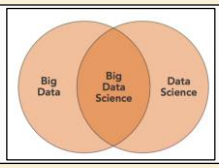
Aside: "Data Science" vs. "Big Data"

Aside: "Data Science" vs. "Big Data"

Aside: "Data Science" vs. "Big Data"

"Big Data Science" Combines All Six of These Attributes

Hacking
Math & Stats
Substantive Expertise



Data Volume
Data Velocity
Data Variety

Analytical Skills + Dataset Characteristics = Unique Power

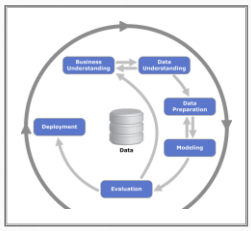
Data Science

- Geographic analysis finds a natural home within data science/big data science
  - Geography provides two unique contributions to the "pattern and regularity recognition"/"substantive expertise" objectives of data science
    - **1. Geographic analysis** (e.g., mapping, spatial statistics) uncovers a dimension of reality that is ignored when we do not consider space
    - **2. Geographic visualization** (e.g. maps and cartograms) enables decision-makers to gain new perspective on the problems impacting their organization

Data Science

- However, data science can also provide some important contributions to geographic analysis in return: attention to data science can help geographers
- Focus of contribution: improved consideration of the key dimensions of rigorous data practice
  - Data mining
  - Data acquisition
  - Data quality

Q: What meaning do these terms have to you?



Data Mining

Understanding Data

Data Mining

- One important distinction to make is the difference between:
  - 1. Mining data: understanding data, preparing data, processing data
  - 2. Using the results of data mining for our complete analytical purposes
- Therefore, one foundational task we need to complete before proceeding further: understanding the process of data mining

Data Mining

- Data mining is a diverse craft that involves elements of
  - Science and technology (quantitative/GIS analysis, computer hardware/software)
  - Art (skill, judgment, and experience)
- As with many crafts, data mining is guided by highly-developed and tested processes that define the best practices in the field
  - Following best practices contributes to consistency, repeatability, objectivity, and ultimately obtaining a successful result

Data Mining

- The Cross-Industry Standard Process for Data Mining ("CRISP-DM") is a useful process codification that places data mining in a broad, holistic perspective
  - CRISP-DM has been further developed by IBM as the "Analytics Solutions Unified Method for Data Mining/Predictive Analytics" ("ASUM-DM"), but for our purposes here we will simply refer to CRISP-DM (name still most widely used)

Data Mining

Cross Industry Standard Process for Data Mining (CRISP-DM)

Six data-centered tasks

Q: What do you understand about data mining based on this process representation?

https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

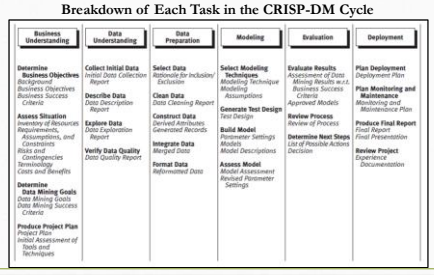
Data Mining

- Key feature of CRISP-DM: inherently iterative
  - Assumes multiple cycles are likely necessary before a final solution is reached
  - Cycling through one time without arriving at an acceptable solution is not a failure
    - CRISP-DM has a consistent focus on exploration and improvement (e.g. improving the problem statement, assumptions, data used, analytical methods employed)
  - Each cycle gains from what was learned before
  - End process when an acceptable solution has been found

Data Mining

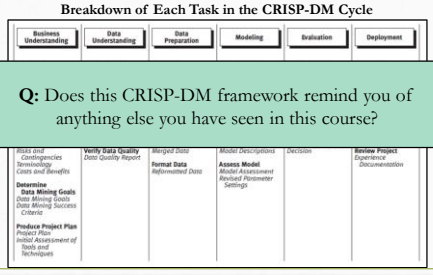
- We do not raise the subject of CRISP-DM here to suggest we must follow each of the six tasks in complete detail in every project
  - Indeed, the next slide provides a highly-itemized overview of CRISP-DM that might be discouraging: many considerations listed for each of the six tasks
  - Rather, following from geography's long-standing perspective of flexibility and adaptation to each situation, I recommend you take inspiration from CRISP-DM and apply what makes sense to any given problem

Breakdown of Each Task in the CRISP-DM Cycle



https://datascienceforall.wordpress.com/2016/08/06/first-blog-post/

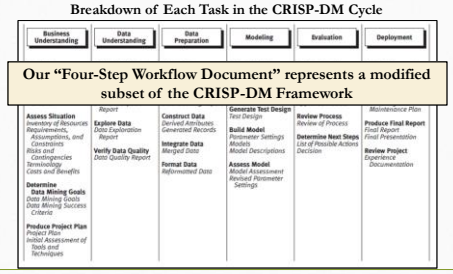
Breakdown of Each Task in the CRISP-DM Cycle



Q: Does this CRISP-DM framework remind you of anything else you have seen in this course?

https://datascienceforall.wordpress.com/2016/08/06/first-blog-post/

Breakdown of Each Task in the CRISP-DM Cycle



Our "Four-Step Workflow Document" represents a modified subset of the CRISP-DM Framework

https://datascienceforall.wordpress.com/2016/08/06/first-blog-post/

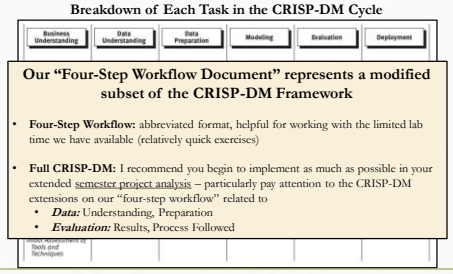
Breakdown of Each Task in the CRISP-DM Cycle

Our "Four-Step Workflow Document" represents a modified subset of the CRISP-DM Framework

- **Four-Step Workflow:** abbreviated format, helpful for working with the limited lab time we have available (relatively quick exercises)

https://datascienceforall.wordpress.com/2016/08/06/first-blog-post/

Breakdown of Each Task in the CRISP-DM Cycle



Our "Four-Step Workflow Document" represents a modified subset of the CRISP-DM Framework

- **Four-Step Workflow:** abbreviated format, helpful for working with the limited lab time we have available (relatively quick exercises)
- **Full CRISP-DM:** I recommend you begin to implement as much as possible in your extended semester project analysis – particularly pay attention to the CRISP-DM extensions on our "four-step workflow" related to
  - ***Data:*** Understanding, Preparation
  - ***Evaluation:*** Results, Process Followed

https://datascienceforall.wordpress.com/2016/08/06/first-blog-post/

Breakdown of Each Task in the CRISP-DM Cycle

Our "Four-Step Workflow Document" represents a modified subset of the CRISP-DM Framework

- Also: see the "Assessment Materials" link on the GEOG 4230/5230 course page for related points in the report and presentation assessment documents
- You will find parallels between the CRISP-DM framework and the report and presentation evaluation dimensions you see represented in those assessment documents

https://datascienceforall.wordpress.com/2016/08/06/first-blog-post/

Let's take a quick tour through the six tasks in CRISP-DM to appreciate the framework's contribution to a decision-focused analysis

Overview of CRISP-DM Tasks

- **1. Business Understanding**
  - Define the problem to be solved
  - Understand the business, its objectives, and how "success" is best defined in this specific situation
  - *Key Issues:* what resources are available, what constraints exist on problem solution, and what important features need to be present in the final solution?
  - **Q:** Why is accounting for these items first a helpful task?

Overview of CRISP-DM Tasks

- **2. Data Understanding**
  - Define the information that is available initially
  - What kinds of information could be obtained if needed?
  - What are the strengths and limitations of each data source?
  - Is data quality an issue with any dataset you will use?
  - What information gaps still need to be addressed?

Q: What would be example(s) of this kind of understanding?

We will further address more data-focused issues later in our data discussion

Overview of CRISP-DM Tasks

- **3. Data Preparation**
  - Define the form of data necessary for our analysis
  - Database selection: what is most suitable?
  - Might need data in a particular format
  - Need for data transformation/reformat

Q: What would be example(s) of this kind of preparation?

Again, we will further address more data-focused issues later in our data discussion

Overview of CRISP-DM Tasks

- **4. Modeling**
  - Define the analytical technique(s) to be used
  - List and evaluate the potential methodologies and select the methods that best fit the problem to be solved and the available data
    - In this course, our analysis will focus on selecting from a suite of GIS methodologies, but we should ultimately become familiar with a larger set of methodological options available in the broader BI/Data Science field (this process will go beyond our work this semester)
  - **Q:** Why do you think CRISP-DM calls for this to follow the previous two data steps?

Overview of CRISP-DM Tasks

- **5. Evaluation**
  - Assess the analytical results achieved: gain confidence that the output is valid and reliable, or see where it is not
  - Evaluate results against the initial problem definition (task 1) for good fit: do the results solve the problem given?
  - Consider assessment involving intermediate steps in the analysis and solution (not only the final results)
  - Involve multiple stakeholders, if possible

Overview of CRISP-DM Tasks

- **6. Deployment**
  - Implement the analytical results in the actual business setting
  - Where possible, implement in test cases first (e.g., for a retailer: in a small number of stores in a larger chain, rather than in all)
  - Track implementation: is it possible to roll out your solution as you specified, or do real-world issues create barriers that require an adjustment to your approach?
  - Evaluate field results and use feedback in future solutions

Further Data Analytic Concepts

Data Acquisition and Data Quality Content for Personal Review and Reflection

Data Acquisition & Data Quality

- In many ways, geographic data analysts have it very easy today:
  - Able to easily pull in a variety of data sources (in many cases, from the cloud)
  - Able to quickly assign geographic coordinates to addresses in a large database (easy "geocoding")
  - Able to try any of a wide range of analysis types within a few moments
- Perhaps because it is so easy to access and use a variety of data sets, it is more critical than ever that we pay close attention to data acquisition and data quality issues (especially relating to spatial data)

Data Acquisition

- "Being able to obtain the data you need to complete your project"
- **Real-world constraint:** we have to recognize that some data needs might not be possible to meet
  - **Q:** what might be some factors that could influence data availability?

Data Acquisition

- **Basic question here:** "Do the data I want actually exist, and if so, how do I get them?"
- The overall concern with data acquisition relates to detailed concern with issues included in the further issues of
  - 1. Data availability and
  - 2. Data accessibility

Data Acquisition

- **1. Data availability**
  - Refers to three data-related questions
    - (1) Have the data actually been collected (or could the data be collected)?
    - (2) Do the data refer to the geography in question (your study area)?
    - (3) Are the data able to be obtained from the data collector/provider?
  - ***"The data exist somewhere"***

Data Acquisition

- **2. Data accessibility**
  - Refers to the mechanics of actually locating, selecting, and retrieving the data you wish to use in your study
  - Are your needed data (kind of data, geographical coverage of data) truly accessible by you?
    - **Q:** What factors could impact whether an existing dataset can actually be accessed by you?
  - ***"I can actually get the data I need"***

Data Acquisition

- Church and Murray (2009) discuss data acquisition issues in detail
  - Their chapter 2 discussion is framed around the fairly narrow issue of GIS data that defines a specific geographic feature:
    - **Point files:** defining specific locations (store, factory, or warehouse sites)
    - **Line files:** defining specific travel corridors or other linear features (rivers, power lines)
    - **Area files:** defining political units (states, counties) or business zones (market areas)

Data Acquisition

- Church and Murray (2009) discuss data acquisition issues in detail
  - You should recognize that other data files/types matter to our business analysis:
    - One example: data about the features represented as points, lines, and areas
      - **Point:** data referring to a specific business location (e.g. store size, sales, employees)
      - **Line:** data referring to a characteristic of a road segment (e.g. number of lanes, speed limit)
      - **Area:** data referring to a particular census tract or zip code (e.g. population, average income, sales of a specific product in the zone)

Data Acquisition

- Church and Murray also refer to three types of data sources
  - **1. Existing sources:** data source options are many today, and are increasing
    - Cloud: many sources available on the web, often for free
      - Many are reputable, but the source and its credibility/bias needs to be assessed
      - All sources have issues, but some sources have more basic issues than others
      - How were the data collected, and what issues could impact dataset quality?

Data Acquisition

- Church and Murray also refer to three types of data sources
  - **1. Existing sources:** data source options are many today, and are increasing
    - Commercial providers: firms that specialize in selling different kinds of data
      - ***Transportation Data:*** TeleAtlas, NAVTEQ, HERE
      - ***Geodemographic Data:*** Claritas, Esri, Nielsen

Data Acquisition

- Church and Murray also refer to three types of data sources
  - **1. Existing sources:** data source options are many today, and are increasing
    - Government providers: some of which may be free (e.g. US Census, US Geological Survey)
      - However, in many countries governments sell census data and other specialized datasets
      - Even in the United States, various levels and departments of government may charge a cost-recovery fee for some geographic files
      - There will always be issues related to data quality and timeliness that should be considered (e.g. census data in the United States are collected every 10 years)

Data Acquisition

- Church and Murray also refer to three types of data sources
  - **2. Semiexisting sources:** data that are available in some form, but not necessarily in a GIS-compatible or digital format
    - Historical data often fall in this category: data produced before the digital era may exist in maps, directories, or other paper formats, but need to be transferred over into digital files to be useful for modern analysis
    - Even recent data might have this issue: might be distributed in a hard-copy format (or an electronic format that is not immediately useful: PDF files, JPG files)

Data Acquisition

- Church and Murray also refer to three types of data sources
  - **2. Semiexisting sources:** data that are available in some form, but not necessarily in a GIS-compatible or digital format
    - Even easily-readable electronic files (e.g., Excel files) might have issues: for example, need to locate addresses from a customer or store location file ("geocoding")
    - Further issues emerge with geocoding with addressed that cannot be immediately located (**Q:** what do you do when a specific street address cannot be found?)

Data Acquisition

- Church and Murray also refer to three types of data sources
  - **3. Surveying/Airborne Approaches:** directly encoding GIS data yourself via a variety of technologies
    - On-the-ground surveying
    - GPS (global positioning systems)
    - Aerial photography
    - Remote sensing (airplane/helicopter/satellite)

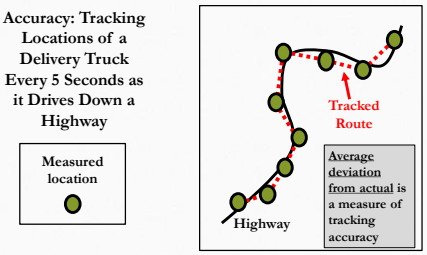
Data Quality

- Beyond data acquisition, we also need to be concerned about the condition of the data we use
  - ***"Does a given dataset have the specific characteristics we need to do the analysis we want to do?"***
  - To answer this question requires that we have some way of defining or understanding the components of data quality

Data Quality

- Spatial Data quality issues can be divided into five categories
  - **1. Data Accuracy:** the discrepancy between the actual attribute value and the attribute value that is coded

Accuracy: Tracking Locations of a Delivery Truck Every 5 Seconds as it Drives Down a Highway



Data Quality

- Spatial Data quality issues can be divided into five categories
  - **2. Data Precision:** relates to the degree of detail displayed in the database
    - Related terms: "resolution", "granularity"
    - How much detail is represented in the database?

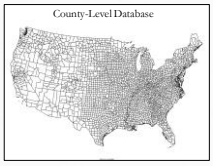
Low Resolution Analysis Framework

State-Level Database

vs.

Higher Resolution Analysis Framework

County-Level Database



"Is the detail I need in a database for my project purposes actually available?"

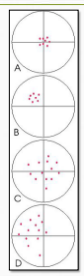
Accuracy versus Precision

Case A: Accurate and Precise

Case B: Not Accurate and Precise

Case C: Accurate but Not Precise

Case D: Not Accurate and Not Precise



Data Quality

- Spatial Data quality issues can be divided into five categories
  - **3. Data Completeness:** a measure of the totality of features included.
    - Are all of the features that can realistically be represented for a given phenomenon actually in the database?
    - A data set with no missing features can be thought of as "data-complete"

Fortune 500: "Data Complete"*

Rank	Name	City	State	Zip	Revenues ($b)	Profits ($b)	Employees	Public	Industry
1	Wal-Mart Stores	Bentonville	AR	72716	469.2	17.0	2,200,000	Yes	General Merchandisers
2	Exxon Mobil	Irving	TX	75039	449.9	44.9	88,000	Yes	Petroleum Refining
3	Chevron	San Ramon	CA	94583	233.9	26.2	62,000	Yes	Petroleum Refining
4	Phillips 66	Houston	TX	77042	169.6	4.1	13,500	Yes	Petroleum Refining
5	Berkshire Hathaway	Omaha	NE	68131	162.5	14.8	288,500	Yes	Insurance: Property, Casualty (Stock)
6	Apple	Cupertino	CA	95014	156.5	41.7	76,100	Yes	Computers, Office Equipment
7	General Motors	Detroit	MI	48265	152.3	6.2	213,000	Yes	Motor Vehicles and Parts
8	General Electric	Fairfield	CT	06828	148.9	13.6	305,000	Yes	Diversified Financials
9	Valero Energy	San Antonio	TX	78249	138.3	2.1	21,700	Yes	Petroleum Refining
10	Ford Motor	Dearborn	MI	48126	134.3	5.7	171,000	Yes	Motor Vehicles and Parts
11	AT&T	Dallas	TX	75202	127.4	7.3	241,800	Yes	Telecommunications
12	Fannie Mae	Washington	DC	20016	127.2	17.2	7,200	Yes	Diversified Financials
13	CVS Caremark	Woonsocket	RI	02895	123.1	3.9	164,500	Yes	Food and Drug Stores
14	McKesson	San Francisco	CA	94104	122.7	1.4	37,700	Yes	Wholesalers: Health Care
15	Hewlett-Packard	Palo Alto	CA	94304	120.4	-12.7	331,800	Yes	Computers, Office Equipment

* For the purposes of a particular business mapping project; you could certainly find other business data fields to add if you wish and need

Fortune 500: Incomplete Data*

Rank	Name	Employees	Industry
1	Wal-Mart Stores	2,200,000	General Merchandisers
2	Exxon Mobil	88,000	Petroleum Refining
3	Chevron	62,000	Petroleum Refining
4	Phillips 66	13,500	Petroleum Refining
5	Berkshire Hathaway	288,500	Insurance: Property, Casualty (Stock)
6	Apple	76,100	Computers, Office Equipment
7	General Motors	213,000	Motor Vehicles and Parts
8	General Electric	305,000	Diversified Financials
9	Valero Energy	21,700	Petroleum Refining
10	Ford Motor	171,000	Motor Vehicles and Parts
11	AT&T	241,800	Telecommunications
12	Fannie Mae	7,200	Diversified Financials
13	CVS Caremark	164,500	Food and Drug Stores
14	McKesson	37,700	Wholesalers: Health Care
15	Hewlett-Packard	331,800	Computers, Office Equipment

* Some business data fields we need for our mapping purposes (i.e. location) are not present here

Data Quality

- Spatial Data quality issues can be divided into five categories
  - **4. Data Consistency:** the absence of logical conflicts in a database
    - Points: only one customer data point occupying any given location
    - Lines: rivers are geocoded to flow into another river, lake, or ocean (they do not end in the middle of nowhere)
    - Areas: census tracts are surrounded by a complete set of boundaries

Data Quality

- Spatial Data quality issues can be divided into five categories
  - **5. Data Uncertainty:** to what degree does user error and/or measurement issues cloud the accuracy of the data
    - Are there systematic issues with the data that we can know and assess?
    - Knowledge of the data collection process itself

Data Discussion and This Course

- Before we end today, let's recap the conceptual resources we have discussed this week and the how they relate to our course work
  - Data mining: CRISP-DM (six tasks)
  - Data acquisition concerns
  - Data quality issues

Data Discussion and This Course

- The ideas we've surveyed in this module provide a foundation for us to bring forward into all of the analysis and problem-solving we will do in the remainder of this course
  - GIS Exercises
  - Semester Project
- Next week we will explore specific applications involving data for the improvement of business decisions in the retail and services domain