Introduction to Big Data

Modeling the Analytic Problem

YORK UNIVERSITÉ UNIVERSITY

Review Previous Lesson

Review Concepts from Day 1

The following major topics were introduced last class.

- Big Data
- Data Analytics
- Business Questions
- Big Data and Analytics
- Analytical Models
- Statistical Methods
- Python Jupyter Notebooks

Lesson Review from Day 1

- The following questions provide a review from last class.
  - Describe the key components to transform data into business value
  - Describe how Big Data is different from traditional Small Data
  - Identify some of the historical events since the year 2000 that have catalysts for the Big Data era
  - Differentiate data from information
  - Explain the major purposes of analytics
  - Describe how business problems need to be translated and framed to take advantage of analytics including the essential components of a framed problem
  - Describe how data analytics relates to business performance management
  - Describe the main role of statistics in describing data
  - Describe the main set of analytics capabilities that are in common use

Day 2 – New Topics Introduced

The following major topics are introduced this Day.

- Business Problem Definition
- Analytics Problem Definition
- Influence Diagramming
- Stakeholder & Analytics Teams
- Data Characteristics
  - Data Structure
  - Data Format
  - Data Granularity
  - Data Latency
  - Data Security
- Variable Types
- Data Visualization Basics

Learning Objectives for Day 2

Day 2 - Learning Objectives

During Day 2 you will learn to:

- Translate a business problem into an analytics problem statement that defines purpose, scope and requirements of a proposed analytical model.
- Describe a proposed solution to the analytics problem statement in terms of stakeholders and team management.
- Further differentiate Big Data from Small Data in terms of key characteristics including structure, format, granularity, latency and security requirements
- Describe different types of variables encountered in measurement
- Describe some basic concepts about data visualizations
- Apply and use a Python Jupyter Notebook

Business Problem Statement

Business Problem Statement

- Business problems must be well framed before analytics can be applied
- The problem must be understood in terms of input and output variables

- Problems may be either:
  - Already well defined and clearly understood
    - Problem symptom (s) are clearly known and described by an output variable
    - Decision variables are well known and understood
    - Assumptions are consistent
    - Other input variables (parameters) are well known
  - Poorly defined and not well understood
    - The actual problem symptom must be discovered and related to an output variable that is understood
    - Decision variables must be discovered and evaluated
    - Assumption variables must be aligned with different points of view
    - Other relevant input variables (parameters) must be discovered and validated.

Business Problems that Are Well Framed

Necessary conditions for a well framed business problem

1. The symptom (output) of the problem has been identified , is well understood and can be described by a variable that can be measured

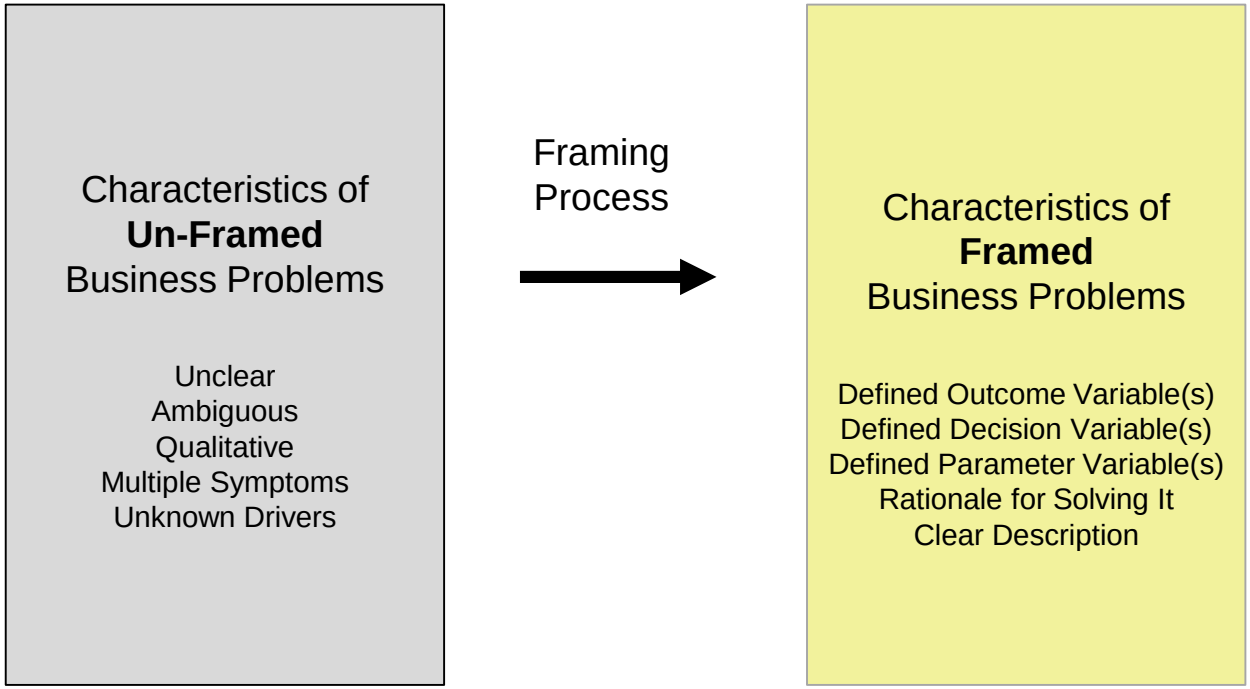
2. The decision (input) variables to the problem have been discovered & identified. They are well understood and can be adjusted, manipulated and measured

3. Additional parameters(input) variables to the problem have been discovered & identified. They can be assigned values that are assumed according to consensus.

4. The importance of solving the problem has been determined and agreed to.

5. The problem has been clearly described in a few sentences.

Business Problems Framing Process



Business Problem Example

- A pizza restaurant is facing new competition from the recent opening of more dining options in its neighborhood.
- The average number of daily customers has started to slowly decline over the past 12 months. The monthly average revenue has also started to decline.
- Lets assume that the problem symptom is framed as “Margin” level. The symptom will be reduced if the “Margin” can be raised to a desired level Margin is defined as Monthly Revenue minus Month Expenses.
- Decision variables that influence margin are price per pizza sold, number of pizza’s sold and various types of operating expenses.
- Parameter variables include local demographic data such as population size and income levels/
- Note that the Margin level was chosen as the framed problem output variable instead of customer count and quantity of pizza’s sold.

Business Problems – Next Step

Problem Framing concludes with:

- Outcome (output) variables
- Input variables
  - Decision variables
  - Parameter variables
- Description of the problem and why it is important to solve

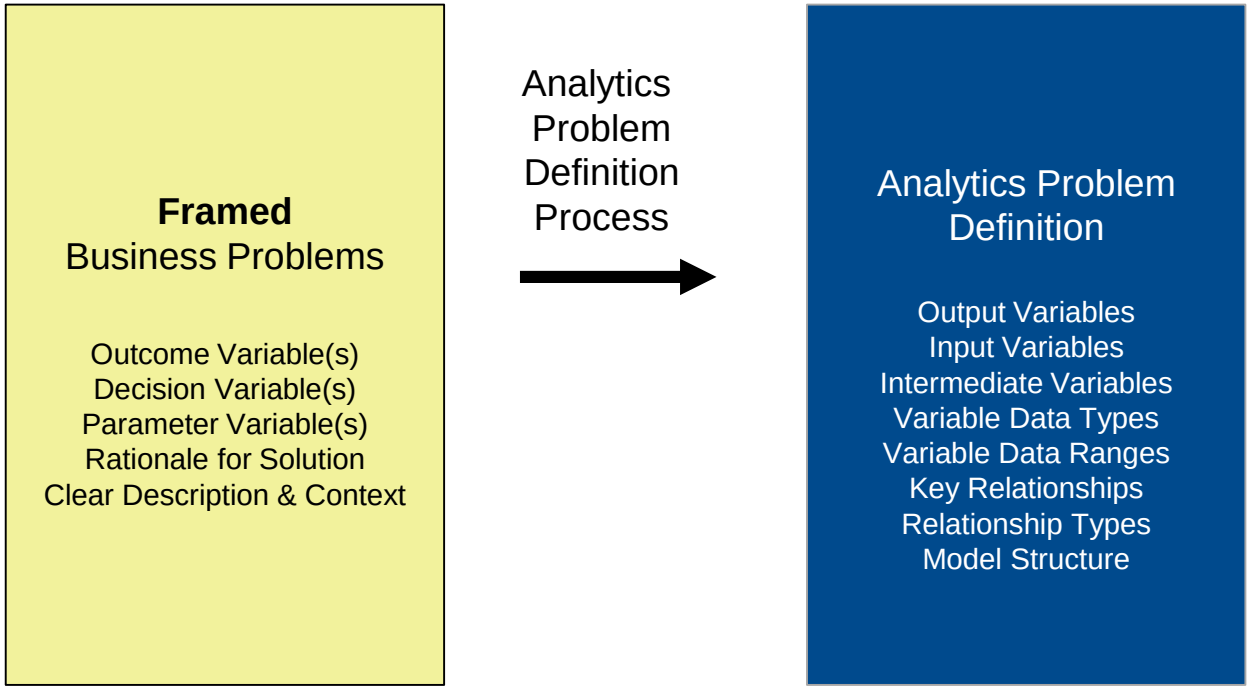
The next step determines how the input variables are related to the output variables.

- This is called the Analytics Problem Statement

Analytics Problem Statement

Analytics Problem Statement

From Business Problem to Analytics Problem



Analytics Problem Statement

Major Components of an Analytics Problem Statement

Component	Purpose	Description
Output Variables	Calculated by an analytics model	Measures improvement or problem resolution
Input Variables	Provided as input to the model	Includes constants, decision variables and estimated parameters
Intermediate Variables	Calculated internally by the model	Provides additional information to the decision maker
Variable Data Types	Determines how the model can process them	Variables are either qualitative (descriptive) or quantitative (measures)
Relationship	The logic that transforms inputs into outputs	Expressed as rules, formulae or algorithms
Analytics Approach	Classifies the type of model to be built	Diagnostic, Descriptive, Predictive or Prescriptive

Influence Diagramming

Influence Diagramming

Influence diagrams allow you to create a logical diagram that describes your current understanding of the analytics problem. It uses a graphical technique to define the key variables and how they influence each other. This is an iterative process and it evolves as your understanding of the analytics problem becomes more clear.

- There are 4 types of variables in an influence diagram
  - Outcome variables
  - Decision variables
  - Parameter variables
  - Intermediate variables

- The variables are connected by an “influence” line

- This technique enables you to identify
  - Key variables
  - Expected relationships between variables

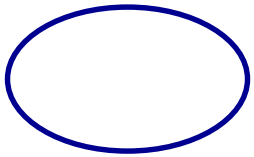
When the analytics solution is eventually designed each “influence” line identifies the need for a formula, heuristic or algorithm.

Influence Diagramming Symbols

Outcome variable calculated by the model

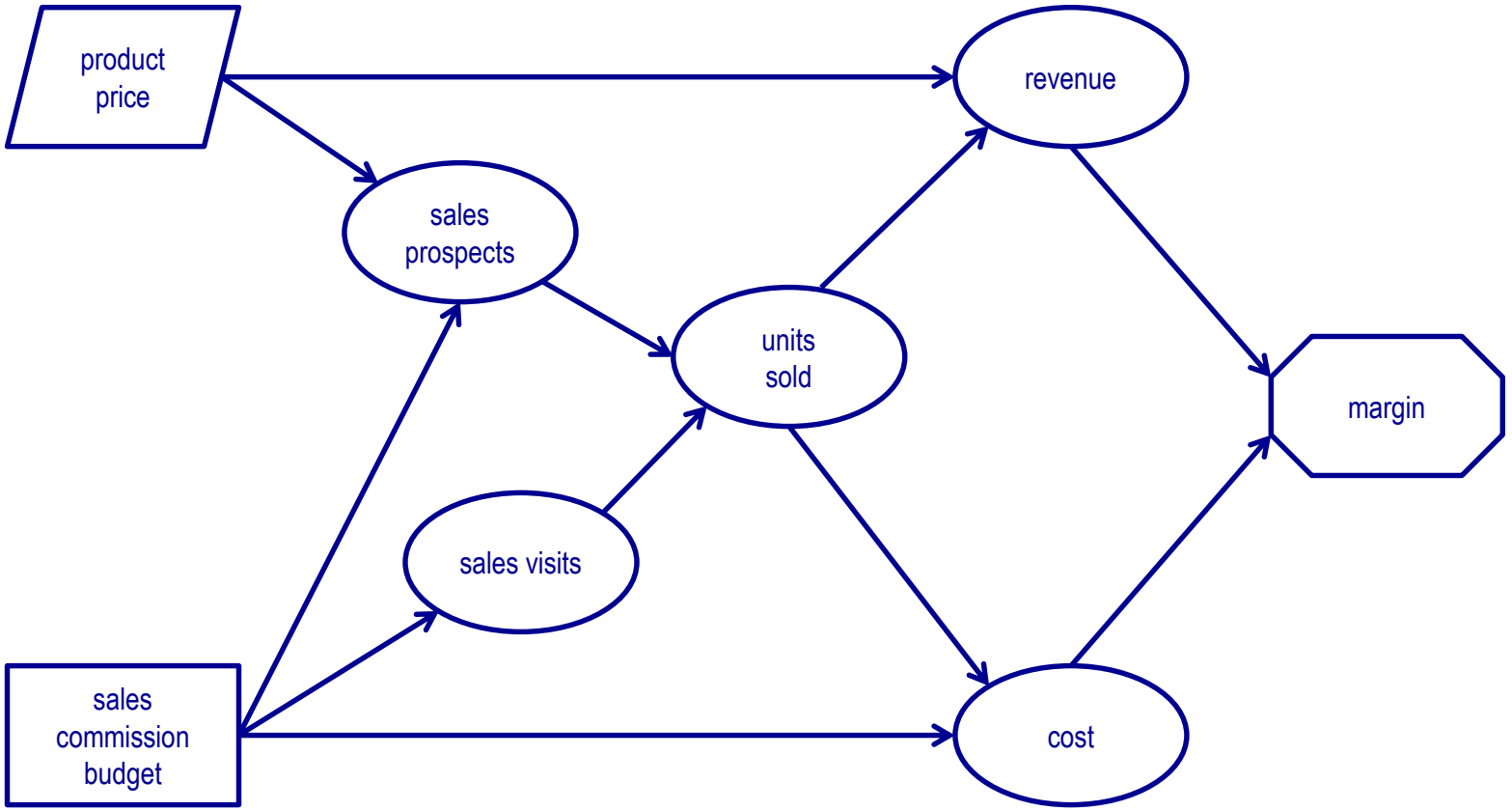
Decision variable input to the model

Parameter variable input to the model

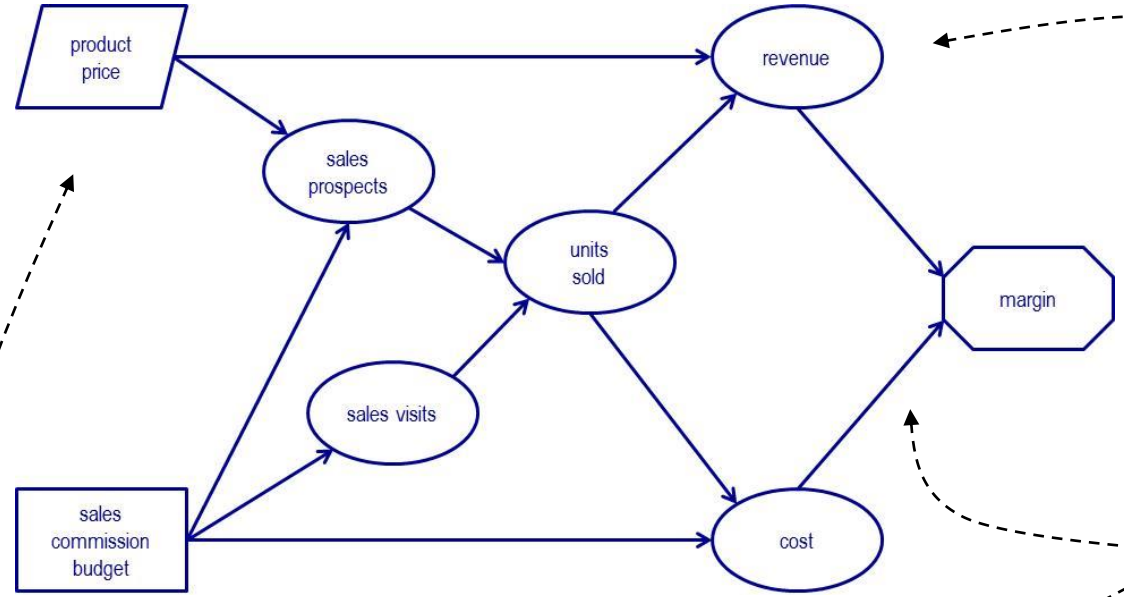


Intermediate variable calculate by the model

Influence Diagramming Symbols



Influence Diagramming Use



Variables
to be Calculated

Influence Lines Show
Relationships to be defined as

1. Rules
2. Formulae
3. Algorithms

Discover Key Variables and Relationships
to Specify the Analytics Problem

Variables
to be Input

Stakeholder and Analytics Teams

Stakeholder and Analytics Teams

Data Analyst

Analytics Modeller

Domain Expert

Business Analyst

Decision Maker

Analytics Team Key Roles

Generate Value from Analytics
Collaborative
Virtual Structure
Based on Trust
Cross Functional

Suppliers

Customers

Regulators

Shareholders

Other Functional Departments

Leadership Teams

Stakeholder Groups

Benefit from Value Generated
Transparent
Based on Trust

Data Characteristics

Data Characteristics

- Data exists in a wide range of diverse formats and structures.
- The ability to analyse data, requires a common foundation of key characteristics of it in order that an integrated view is available.
- Data will be acquired from many different disparate sources, technologies and processes. A common foundation is needed to facilitate the analysis
- Think of data characteristics as different properties possessed by a data set that must be managed to provide a uniform and integrated view
- This section introduces some key characteristics of data that make it challenging to provide the necessary uniform view.

Data Characteristics

The following characteristics are described in this section.

- Structure
- Usage
- Data Type
- Latency
- Movement
- Quality
- Security

Restrictions and opportunities for the Analytics Problem

Data Structure

Structure

- Describes how the data content is organized.
- Structured data describes content that individual fields are organized in a tabular style
- Unstructured data describes content where the content organization is not at a field level
  - It generally unknown.
  - The structure must be discovered before analytics can play a role
- Multi-structured data describes content where data has known but diverse organizations
  - Combinations of structured and unstructured may exist

Data Usage

Usage

- Data usage puts context around the roles that data elements play in helping to create the information used to answer question that were framed in the Business Problem Defintion.

- Facts
  - An answer to a question is a “fact”. The fact is a data element that may be quantitative or qualitative depending on the question.

    Example:
    - Who is my largest customer?
    - The answer to this question is a customer name. It is qualitative
  - How many customers are in my “high value segment”?
  - The answer to this question is a customer count. It is quantitative.

- Qualifiers
  - Additional data elements are qualifiers found in a business question. They provide context and grouping criteria for the fact.

    Example
    - How many customers exist by sales region, by product, by month? The fact is customer count and the qualifiers are sales region, product and month.

Data Type

Type

Data type refers to how data is stored and it defines the types of processing that are supported. Common examples of data types include:

1. Fixed Character Strings
2. Variable Length Character Strings
3. Floating Point
4. Integers
5. Binary Values
6. Date
7. Binary Large Objects (BLOBS)

Refer to the following link for more details about data types

http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=Geodatabase_field_data_types

Data Latency

Latency

Latency refers to a time perspective of data. It describes how frequently it is updated and how much of a time delay exists between an actual event happening and the related data base is made aware of it.

1. Real Time
   - low latency describes the condition when there is a “small” amount of time between a real event and the data is updated to reflect it. Depending on the application, low latency may be sub second in an automation environment and several minutes in a business environment.

2. Right Time
   - the latency in the data adequately reflects the requirements of the decision maker

3. Near Time
   - refers to the latency that is less than a day but greater than an hour.
   - high latency describes the condition when the delay is larger and typically measured in hours or days.

Data Movement

Movement

Data movement is an important concept to understand because it restricts the types of analysis that can be supported. It is related to latency but it also refers to having historical data available that may be necessary for some analytical applications.

At Rest

- Data that is stored in a data platform of some kind and is waiting to be accessed is “at rest”
- Data at rest is found in a data base or some other form of repository
- Historical data is typically available

In Motion

- Data that is flowing through an environment using messages and is never actually stored is “in motion”
- Data in motion is found in a message stream also called a data pipeline.
- Historical data is typically not-available

Data Quality

Quality

- Data Quality refers to the condition when various conditions in the data sufficiently meet the requirements of those conditions by a given analytical use case
- High quality data is able to meet the needs of more use cases than low quality data.
- Examples of conditions of data that need to be evaluated and assessed.
- Correctness Conditions – based on content
  - Accuracy
  - Completeness
  - Granularity
  - Latency
  - Consistency
  - etc
- Integrity Conditions – based on structure and relationships
  - Unique records (no duplicates)
  - Parent records exist for all child records
  - Domain range (valid values based on defined domain set)
  - etc

Data Security

Security

- Update, access and delete rules are defined based on regulations and corporate policies.

- They are implemented in two categories

- Authentication Rules
  - Validates that who you are and that you are allowed to be in the given system or application.
- Authorization Rules
  - Define what actions you are allowed to carry out with the data given that you are on the system.

Quantitative Variable Types Based on Measurement Scales

Quantitative Variable Types

- The concept of measurement and measurement scales applies to data that is quantitative and used in an Analytics Solution as part of a measurement process

- Quantitative data elements are used to measure the property of something and relate its value to a standard scale based on the type of variable and its chosen units of measure

Examples of Measurement Data Used to Quantify Properties

speed of a car	100 km/hr
expected hotel service level	4 stars
size of a customer segment	1500 high value customers
current outdoor temperature	55 deg C

Describing Measurement Scales

Variable Type Based on Measurement Scales	Description	Example
Ratio	Whole numbers including fractional values. The “zero” position on the ratio scale is meaningful as the “absence” of something	Comparing A which is $10 to B which is $5 we can conclude 2 things 1. A is five units bigger than B (delta) 2. A is twice as large as B (proportion)
Interval	Whole numbers including fractional values. The “zero” position is arbitrary and does not mean the “absence” of something.	Compare two temperature readings. A is 20 Deg C and B is 10 Deg C. We can only conclude that A is 20 deg warmer than B. It is incorrect to conclude the proportion between A and B
Ordinal	Integer numbers used to rate some thing in order from best to worst.	A ranked list of hotels rated by the Star system provides a list of best to worst hotels based on service levels. The only valid conclusion is that a “4 star” hotel is better than a “2 star” hotel but you cannot infer how much better
Nominal	By aggregating individual counts into a higher level total provides a nominal measure of size of the importance of a given category	Assume a company has 3 categories of risk for their equipment. They are Low, Medium and High. Nominal variables show the count of equipment in each category. Number of equipment items in each category Low Risk = 100 units Medium Risk = 250 units High Risk = 50 units

Data Visualization Basics

What is Information Visualization?

“Transformation of the symbolic into the geometric”

(McCormick et al., 1987)

“... finding the artificial memory that best supports our natural means of perception."

(Bertin, 1983)

The depiction of information using spatial or graphical representations, to facilitate comparison, pattern recognition, change detection, and other cognitive skills by making use of the visual system.

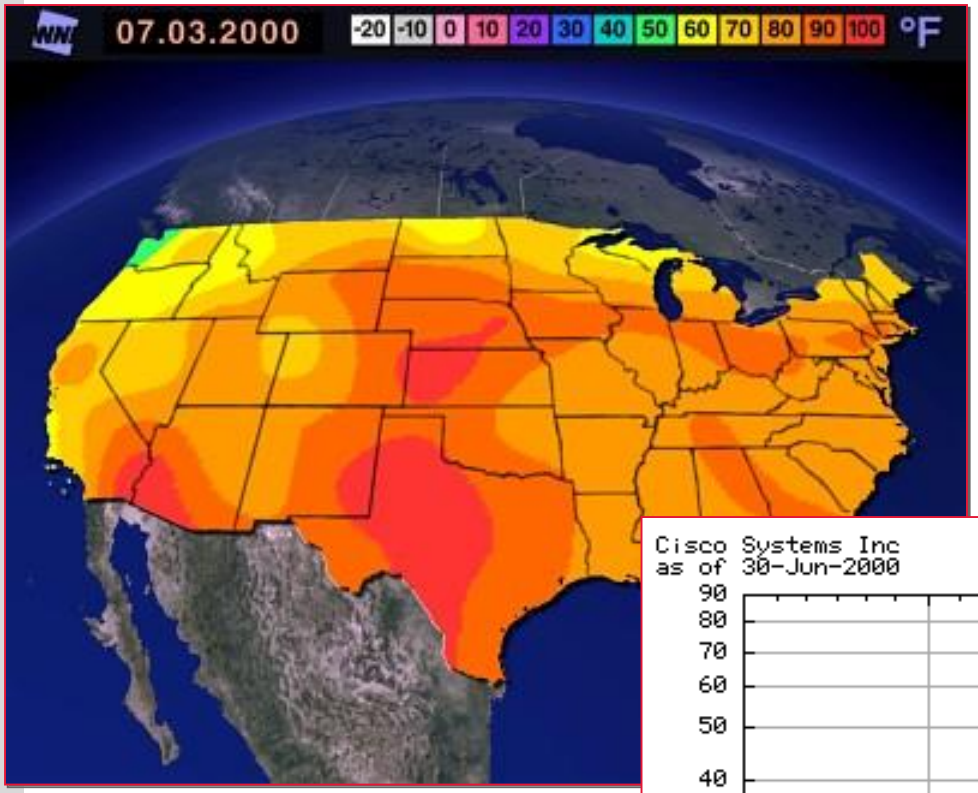
Information Visualization

- Problem
  - Big datasets: How to understand them?
- Solution
  - Take better advantage of human perceptual system
  - Convert information into a graphical representation.
- Issues
  - How to convert abstract information into graphical form?
  - Do visualizations do a better job than other methods?

Goals of Information Visualization

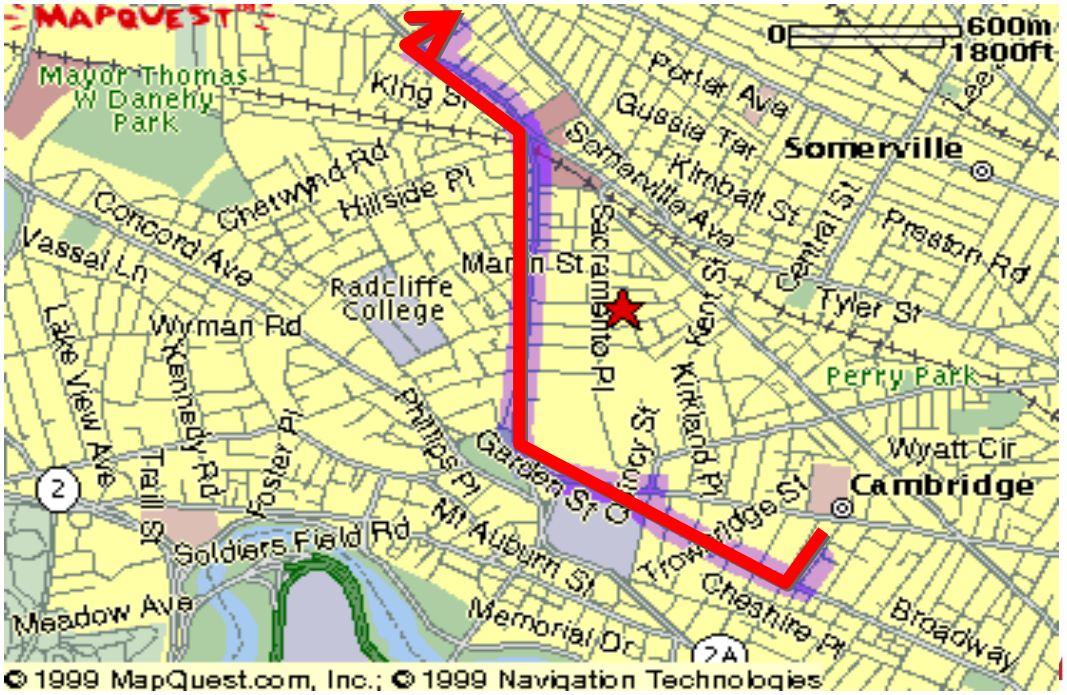
- More specifically, visualization should:
  - Make large datasets coherent
    (Present huge amounts of information compactly)
  - Present information from various viewpoints
  - Present information at several levels of detail
    (from overviews to fine structure)
  - Support visual comparisons
  - Tell stories about the data

Visualization Success Stories



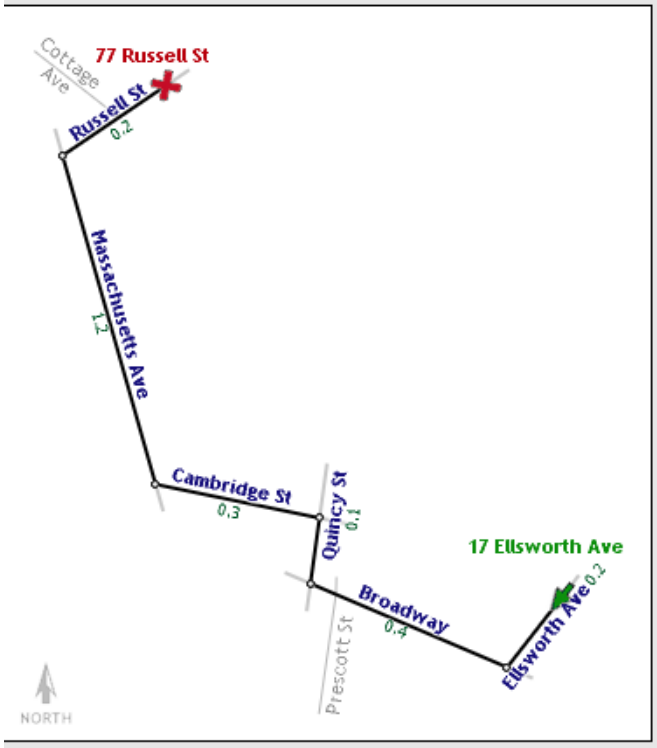
The Power of Visualization

1. Start out going Southwest on ELLSWORTH AVE
   Towards BROADWAY by turning right.
2: Turn RIGHT onto BROADWAY.
3. Turn RIGHT onto QUINCY ST.
4. Turn LEFT onto CAMBRIDGE ST.
5. Turn SLIGHT RIGHT onto MASSACHUSETTS AVE.
6. Turn RIGHT onto RUSSELL ST.



The Power of Visualization

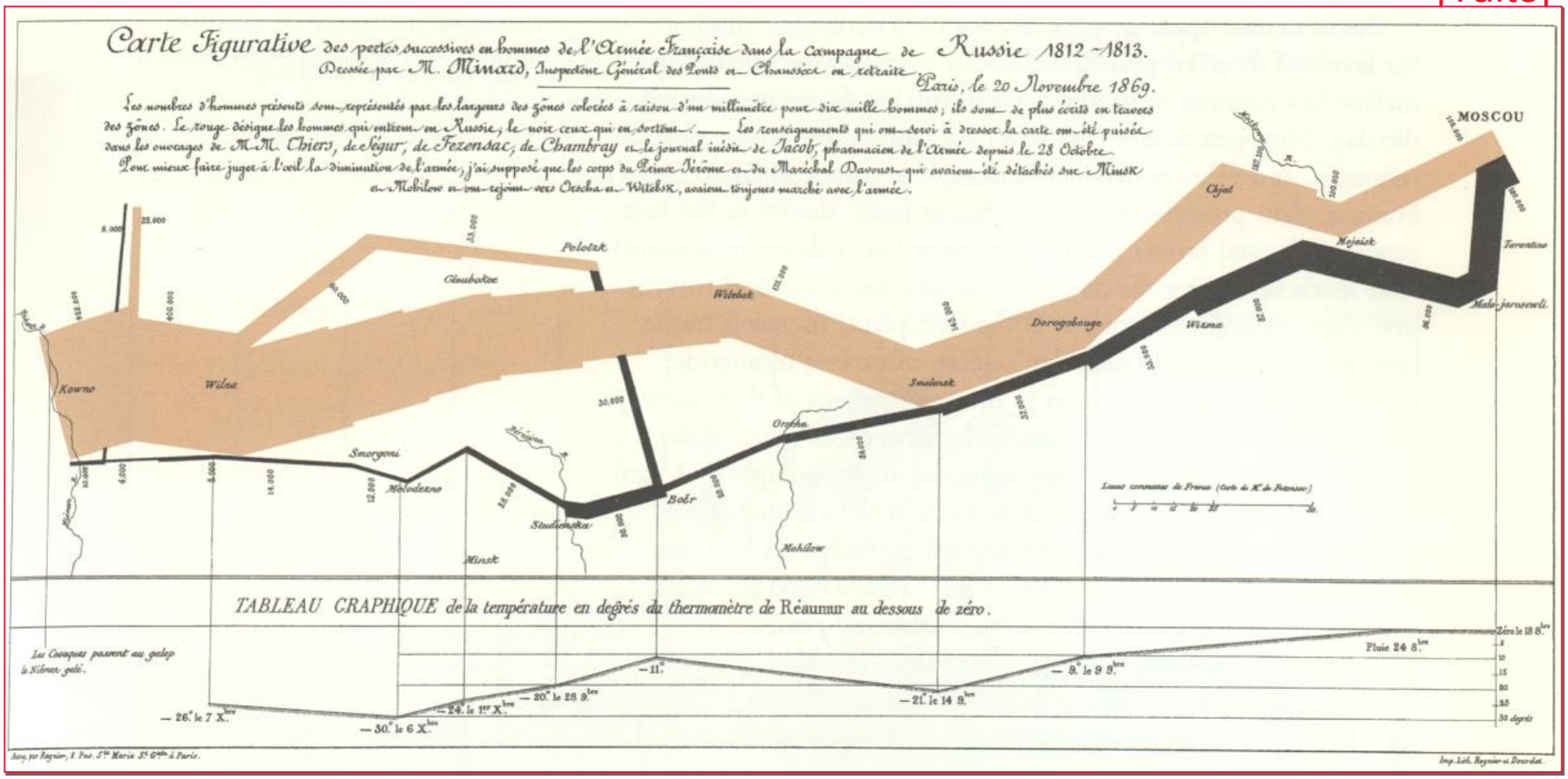
The estimated travel time is 5 minutes for 2.16 miles of travel, total of 6 steps.



	Directions	Elapsed Distance
1	Begin at **17 Ellsworth Ave** on **Ellsworth Ave** and go Southwest for 500 feet	0.1
2	Turn right on **Broadway** and go Northwest for 0.4 miles	0.5
3	Turn right on **Quincy St** and go North for 200 feet	0.5
4	Turn left on **Cambridge St** and go West for 0.3 miles	0.8
5	Bear right on **Massachusetts Ave,Mass Ave,RT-2A** and go North for 1.2 miles	2.0
6	Turn right on **Russell St** and go Northeast for 1000 feet to **77 Russell St**	2.2

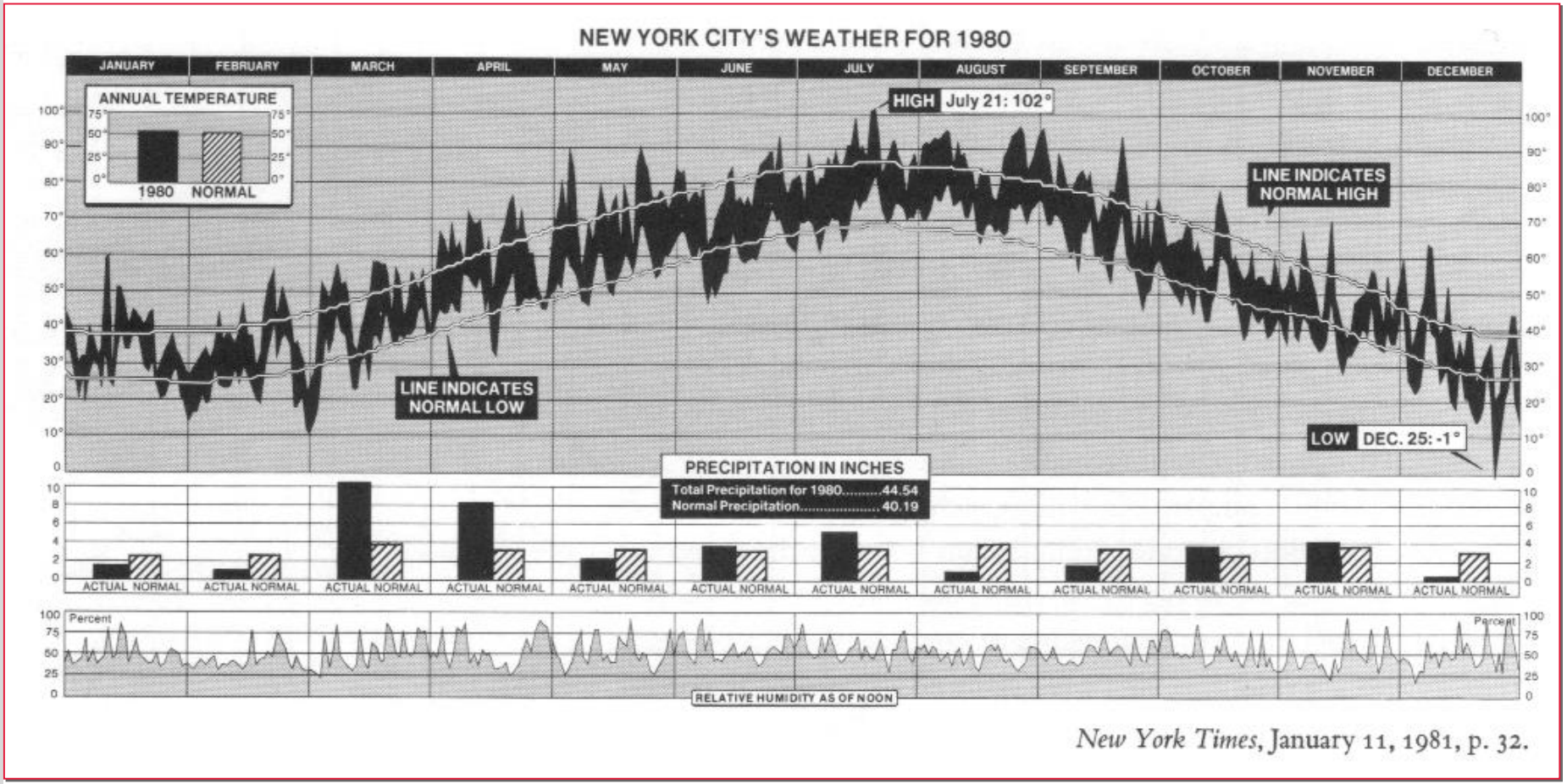
Napoleon's 1812 March by Charles Joseph Minard

[Tufte]



Variables shown: •size of army •temperature •direction •date •latitude •longitude

NYC Weather



New York Times, January 11, 1981, p. 32.

2220 numbers

Visualization Success Story

Mystery: what is causing a cholera epidemic in London in 1854?

Visualization Success Story

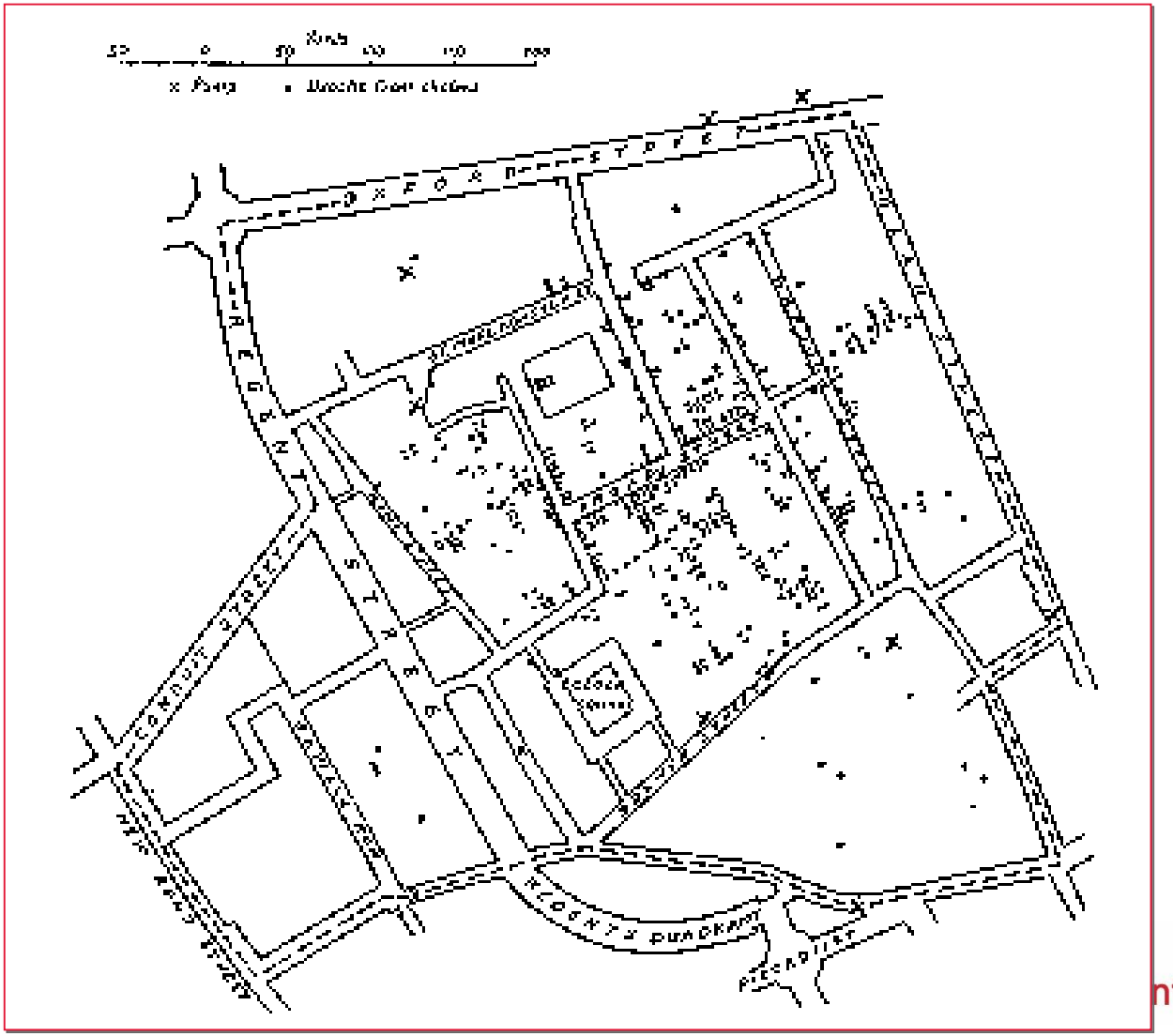
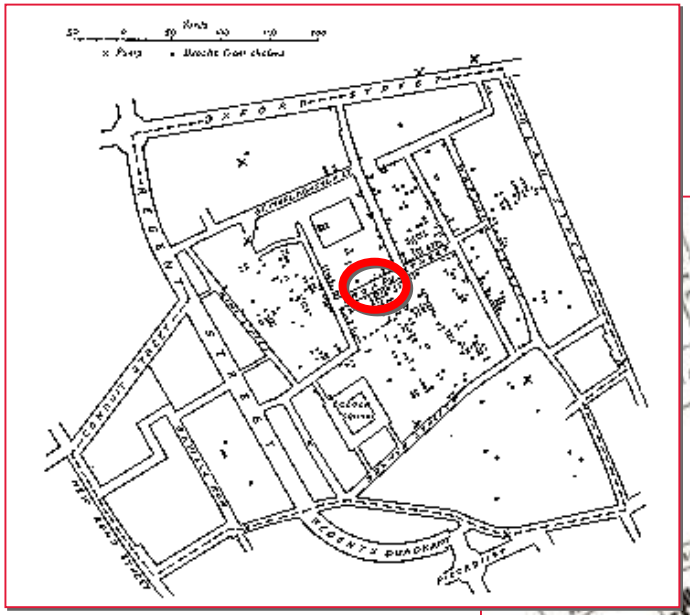
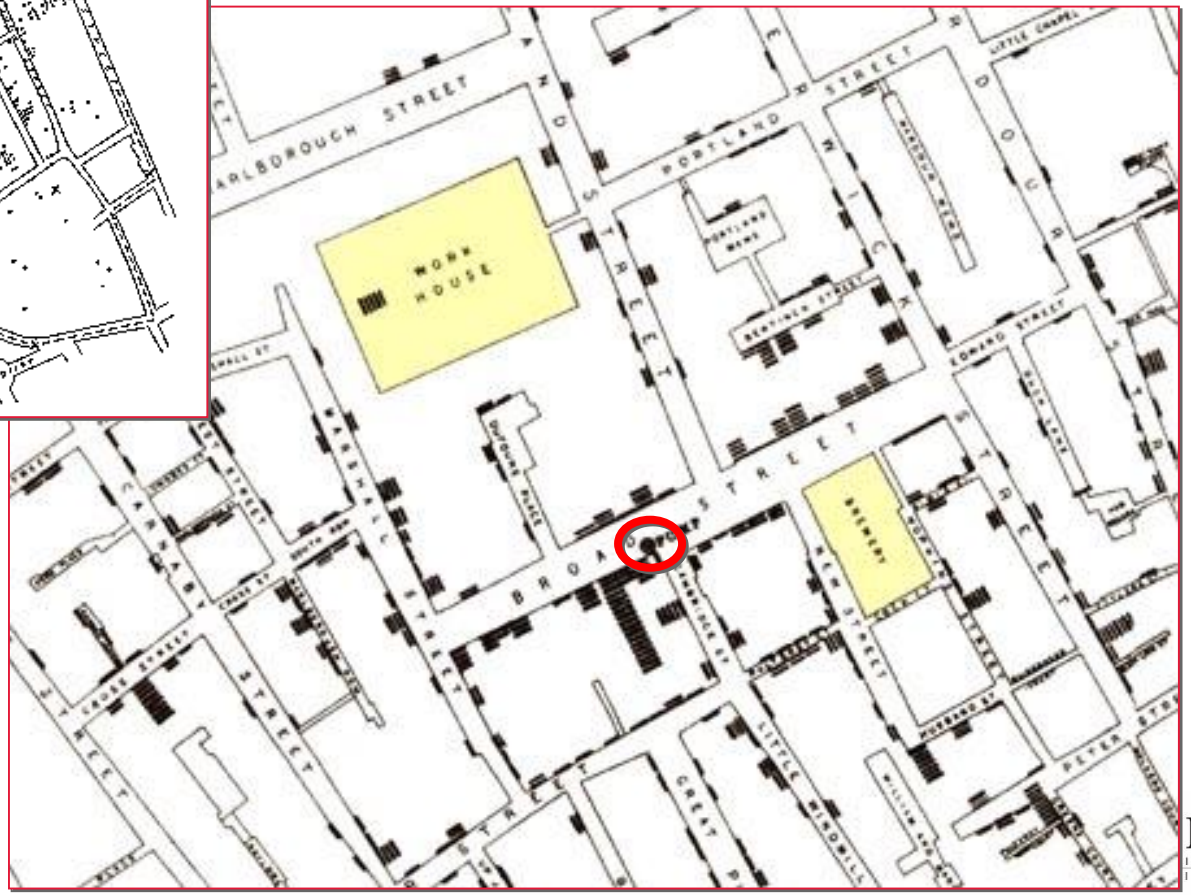


Illustration of John Snow's deduction that a cholera epidemic was caused by a bad water pump, circa 1854.

Horizontal lines indicate locations of deaths.

Visualization Success Story





Data Visualization

- Become acquainted with basic data visualization concepts by starting off with two videos from the Ted Talks series
  - https://www.ted.com/talks/hans_rosling_at_state
  - https://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization
- Get introduced to creating visualizations using Python on the following website
  - https://www.datasciencecentral.com/group/tutorials/forum/topics/cheat-sheet-data-visualisation-in-python

Data Visualization

- Read the following about the concepts related data visualization.

- Become familiar with the following online resource that provides broad range of visualisation techniques on the following. Browse the site and explore how it can provide you with interesting approaches for visualising your data.
  - https://datavizcatalogue.com/

- Read the following online resource. It provides you with a useful perspective on how to select an appropriate visualisation technique for your application.

  - https://blog.hubspot.com/marketing/data-visualization-choosing-chart

Lesson Review

Lesson Review

Consider the following questions that you should be able to answer by completing Day 2.

- What are the main items found in a framed business problem?
- What are the main items found in a defined analytics problem?
- What is the purpose of an influence diagram?
- What are the critical roles found in an analytics team?
- What does data granularity refer to?
- What are the different types of data latency?
- What are some examples of ordinal data?
- What are some restrictions placed on interval and ordinal data?
- What criteria could you use when selecting a visualization?
- What are the main building blocks of a SQL Select statement?

Lesson Summary

Lesson 2 Summary

During Day 2 you learned to:

- Translate a business problem into an analytics problem statement that defines purpose, scope and requirements of a proposed analytical model.
- Describe a proposed solution to the analytics problem statement in terms of stakeholders and team management.
- Further differentiate Big Data from Small Data in terms of key characteristics including structure, format, granularity, latency and security requirements
- Describe different types of variables encountered in measurement
- Describe some basic concepts about data visualizations
- Describe the basics of a SQL query
- Apply and use a Python Jupyter Notebook