Course Introduction

BAN 5733

Descriptive Business Analytics

Agenda

- Teaching philosophy
- Expectations from students
- An overview of different degrees and certificates
  - MS in Business Analytics
  - Graduate Certificate in Business Data Mining
  - Graduate Certificate in Marketing Analytics

Teaching Philosophy

I listen, I forget.

I see, I may remember.

I do, I understand.

Expectations from Students

- Read Syllabus/Schedule carefully
- Due to the nature of this course it is very important that *you take personal responsibility* for keeping up with
  - Watching video lectures, following demonstrations, doing exercises,
  - Reading assigned/optional readings, **watching lab recordings**,..
- Understand the difference between a **graduate versus an undergraduate course**:
  - Taking a class versus learning marketable skills.
  - Importance of professional communication styles (not just technical skills).
  - Importance of retaining business focus in your write-up/discussion of exercises/assignments/cases.
- Understand the importance of **academic integrity** (unless stated otherwise, all work must be done individually!!)

Communication with Faculty or, TAs

- If you need clarification on any issues related to this class – please *first* try to use the class discussion bulletin board on class site or, ask in the lab
  - DL students consider asking questions/clarifications at the beginning of each lab via Go To meeting
  - *Except*, for any questions that involve personal information (such as your grade), please use email.
- Faculty and TAs will monitor and answer questions posted on this bulletin board.
  - If a question on the bulletin board is not answered within a reasonable time (**say 12 hours from posting a question**) then you can send email to faculty with your question.
  - If you are not satisfied with TA answers, then you may email faculty.
  - Avoid **asking direct questions about how to do an exercise or case**. Asking for **clarifications** about those topics or **direct questions** about software issues are ok.

Role of Teaching Assistant (TA)

- TA for the class will be announced in the first week of classes.
- TA will help during lab (primarily Stillwater students) or during the week (primarily DL students):
  - Monitor discussion board and answer questions
  - Help you with software related issues
  - Help you with conceptual issues
- TA will **NOT** help you with direct exercise/assignment related questions of how to solve it (unless the question is of clarification type)

Degrees and Certificates in Analytics : https://analytics.okstate.edu/

- MS in Business Analytics
  - Intended for students *with technical background* and/or have tolerance for programming
    - Minimum 33 credit hours for working professionals (online)
    - Minimum 37 credit hours for full-time students (on campus)
  - Possible to earn up to 3 SAS OSU Certificates
    - SAS and OSU Data Mining
    - SAS and OSU Predictive Analytics
    - SAS and OSU Marketing Data Science
- Graduate Certificate in Business Data Mining
  - Intended for working professionals *with technical background* and/or have tolerance for programming
    - Minimum 12 credit hours for working professionals (online)
    - Possible to **transfer all credit hours** to MS in Business Analytics
      - Best to do any transfer **before you** file for graduate certificate diploma
      - You can do it even after you finished all courses in the graduate certificate program
  - Possible to earn 1 SAS OSU certificate
    - SAS and OSU Data Mining

Introduction to Analytics

Agenda

- Non-industry specific and a high-level discussion of:
  - What is analytics?
    - History, terms and meanings
  - Different types of data and the types of questions they can answer
  - Progression of analytics over time

What’s in a name?

Analytics or,

Statistical Analysis
Data Analysis
Data Mining

Data Science
Machine Learning
Deep Learning

Business Analytics
Advanced Analytics
Big Data Analytics
Marketing Analytics
Customer Analytics
Web Analytics
Social Media Analytics
Financial Analytics
Fraud Analytics
Risk Analytics
HR Analytics
Operational Analytics
Supply Chain Analytics
Sports Analytics
Healthcare Analytics
……

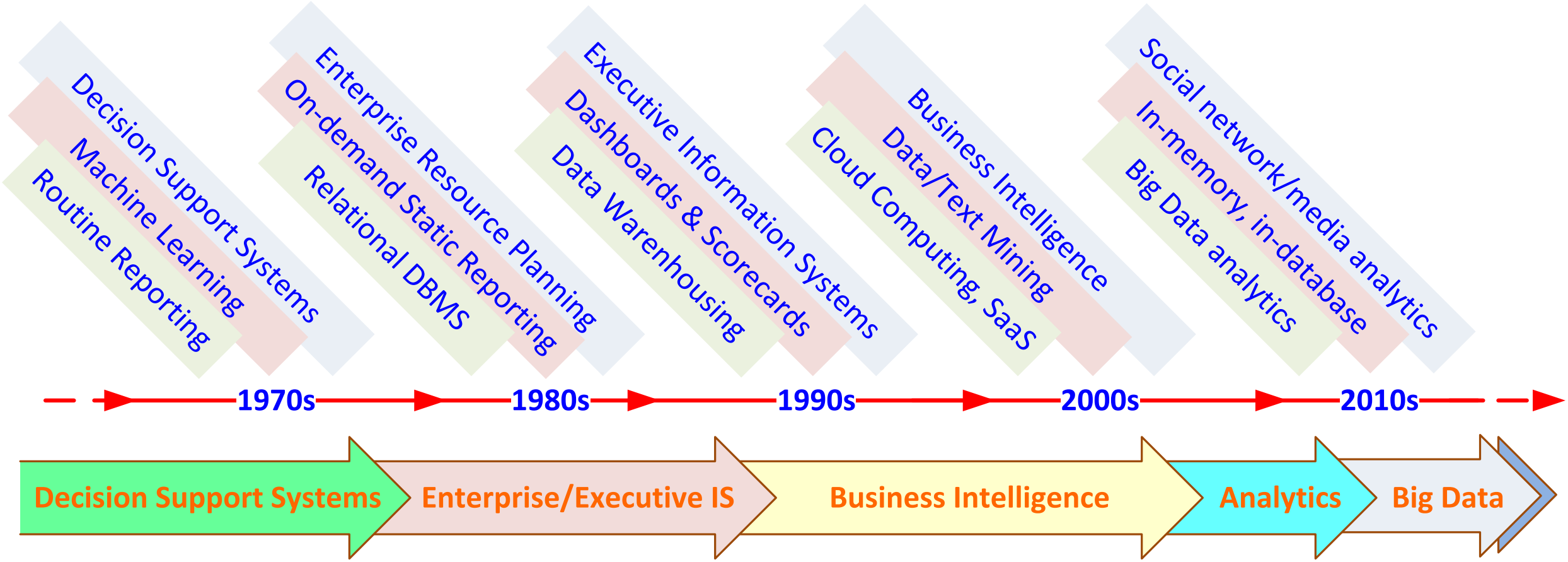
What is Analytics?

- It is not “IT”
- It is not “IS/CS”
- It is not “programming”
- It is not “statistics”
- It is not “optimization”
- But, it uses concepts and tools from all of the above plus more
- “Analytics is the scientific process for transforming data into insights for making better business decisions” (INFORMS, https://www.informs.org/)

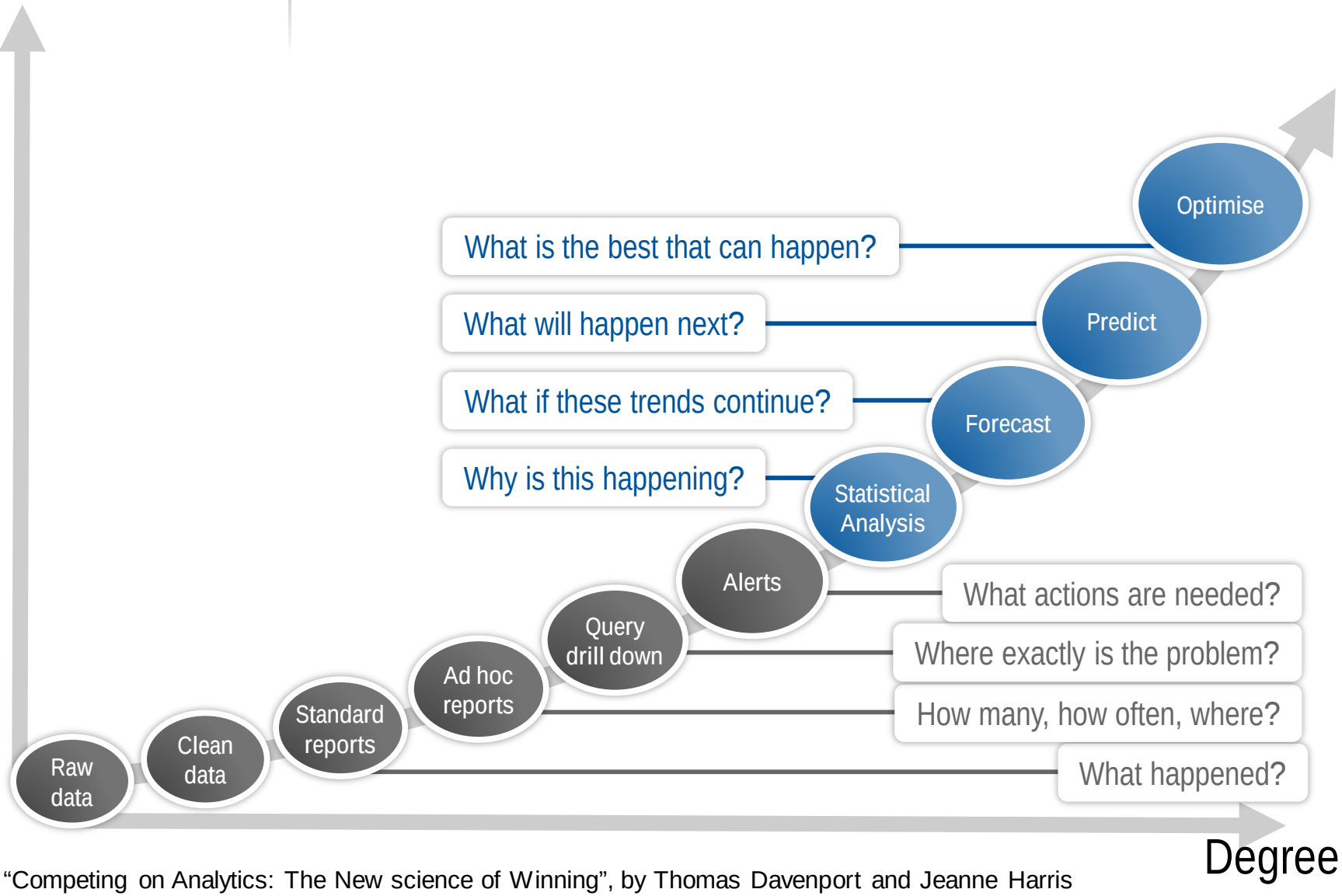
A great story…

- How-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/
  - https://shopping.yahoo.com/news/target-figured-teen-girl-pregnant-000000163.

Evolution of Analytics

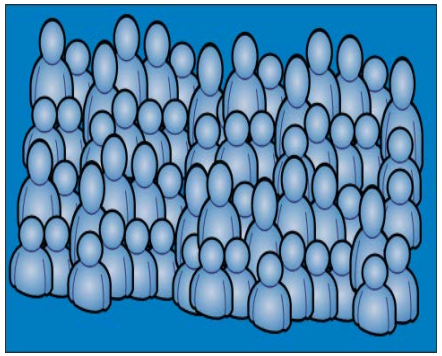


ANALYTICAL DECISION MAKING

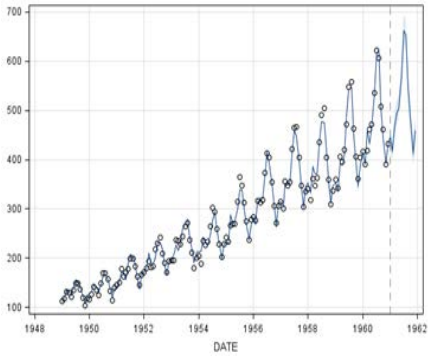


Adapted from the book "Competing on Analytics: The New science of Winning", by Thomas Davenport and Jeanne Harris

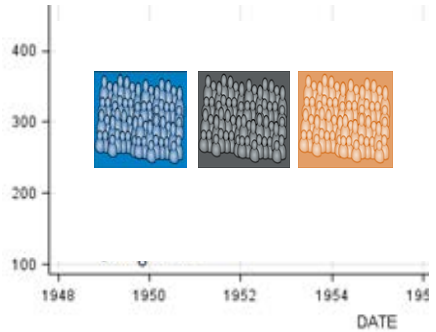
Data Galore



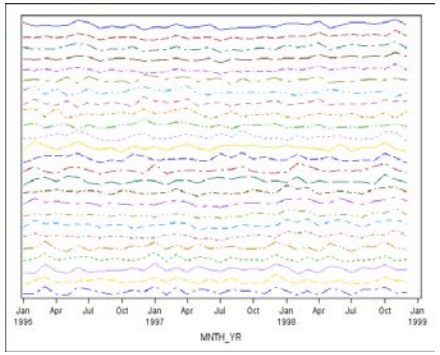
Cross-Sectional



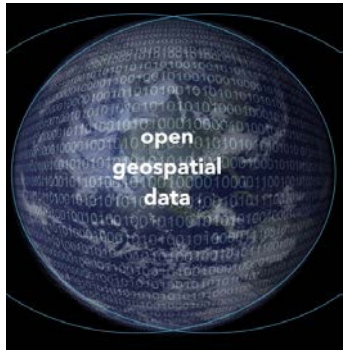
Time Series



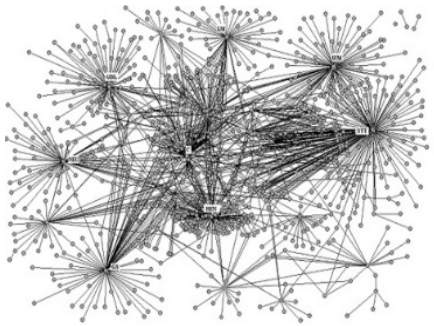
Panel



Streaming



Spatial



Network

2			4	5	2.94*
5		4			1
		5		2	2.48*
	1		5		4
		4			2
4	5		1		1.12*

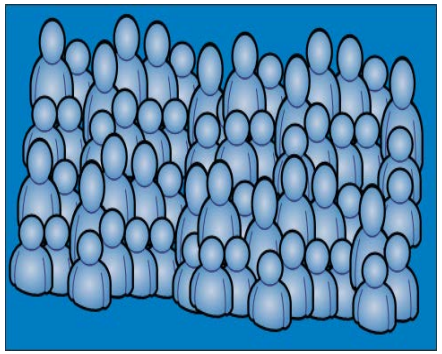
Link

Text

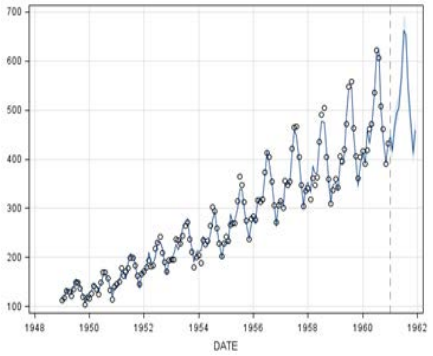
Sound

Image/Video

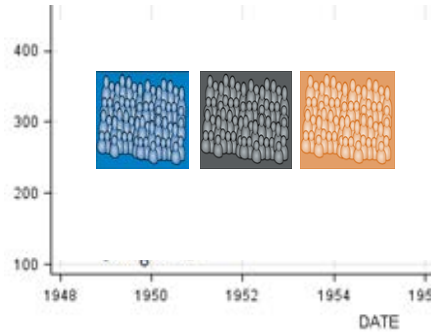
Business Questions



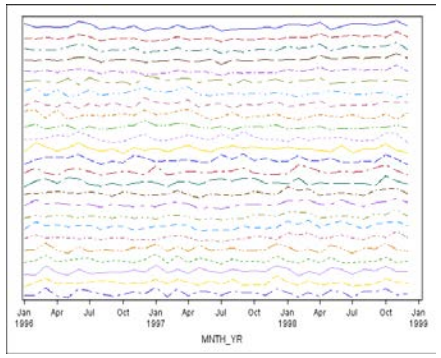
Who will respond to a campaign?



What will future demand look like?

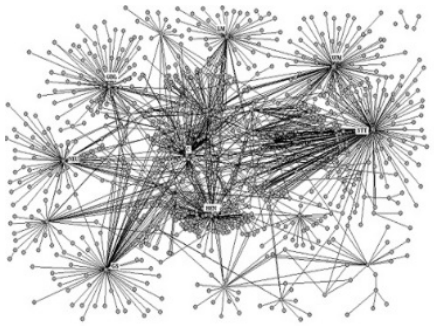


How does panel members' opinion change over time?



Are there anomalies?

Where should we look for oil?



Who are influencers?

2			4	5	2.94*
5		4			1
		5		2	2.48*
	1		5		4
		4			2
4	5		1		1.12*

Which item should I recommend?

How is our product perceived?

Emerging problem in our product?

Can we prevent insurance claims?

The Little “Analytics” Shop of Ice Cream

Descriptive Analytics (numbers)

Which flavors do customers prefer? **How does** that relate to customers’ demographics?

Text Analytics (Text)

What are **people saying** about our vanilla ice cream in social media?

Forecasting

How much vanilla ice cream will be purchased tomorrow, next day..?

Predictive Analytics

Which customers are likely to buy vanilla ice cream?

Prescriptive Analytics (optimization)

How to produce **enough** vanilla ice cream, taking production **constraints** into account?

Progression of Analytics Applications

Traditional analytics applications started in **marketing** and **finance**:

- Understanding and prediction of **buyer behavior** (predict who is likely to buy, predict churn, predict what customer will buy next, etc.)
- **Risk** prediction and risk management

Today analytics are being used in every functional areas such as:

- Human Resources (HR)
- Manufacturing, Production and Operations
- IT
- R&D
- Service
- and many more

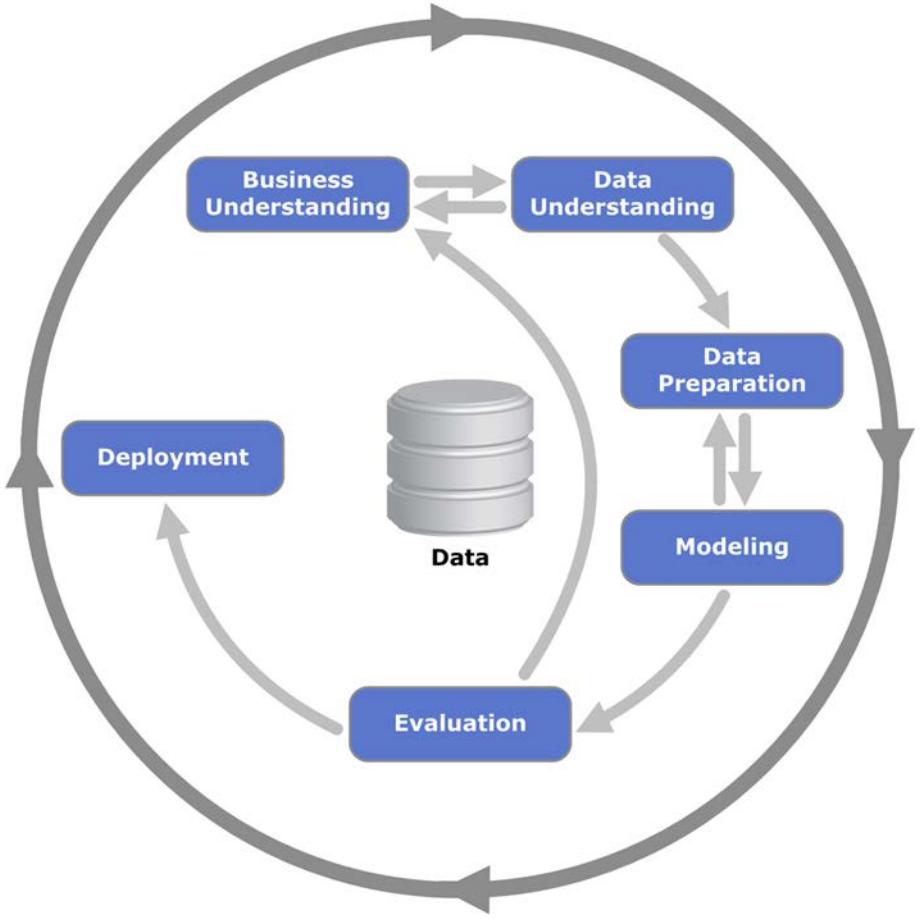
Introduction to Business Analytics Process

Analytics Process

- CRISP-DM
  - Cross Industry Process for Data Mining
- SEMMA
  - Sample, Explore, Modify, Model and Assess
- BAP
  - Business Analytics Process

Business Analytics Process and CRISP-DM

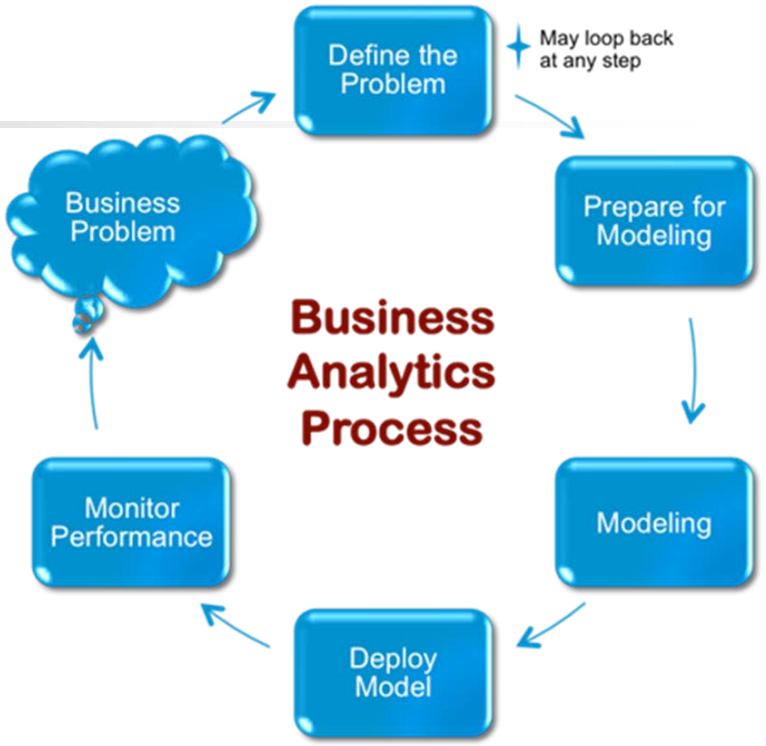
CRISP-DM



Adapted from the book "Building Better Models with JMP Pro" by Gayson, Gardner and Stephens and copyrighted by SAS

Define the Problem

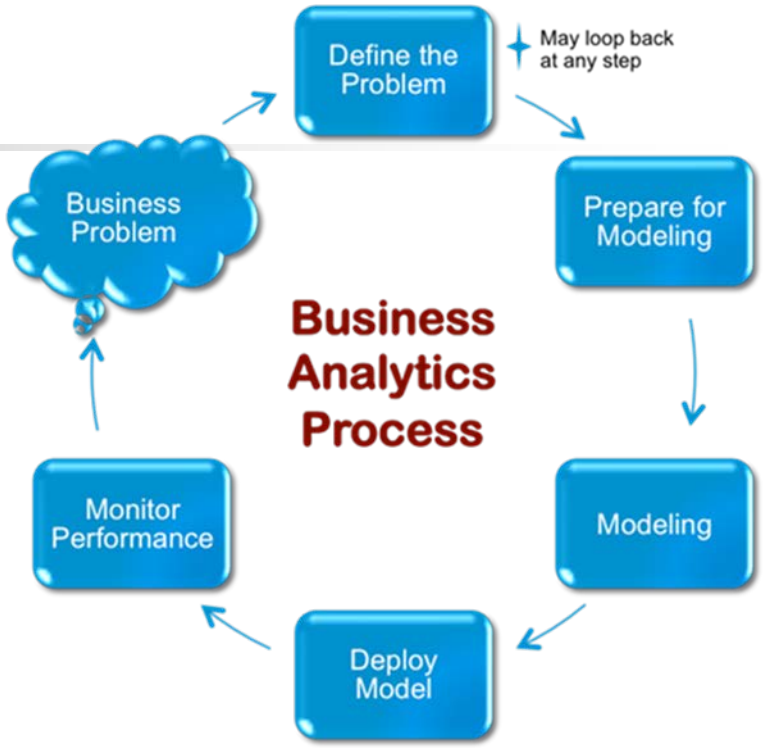
- Define the business problem
- Frame the analytics problem
- Define success
- Project charter is defined and approved



Adapted from the book "Building Better Models with JMP Pro" by Gayson, Gardner and Stephens and copyrighted by SAS

Prepare for Modeling

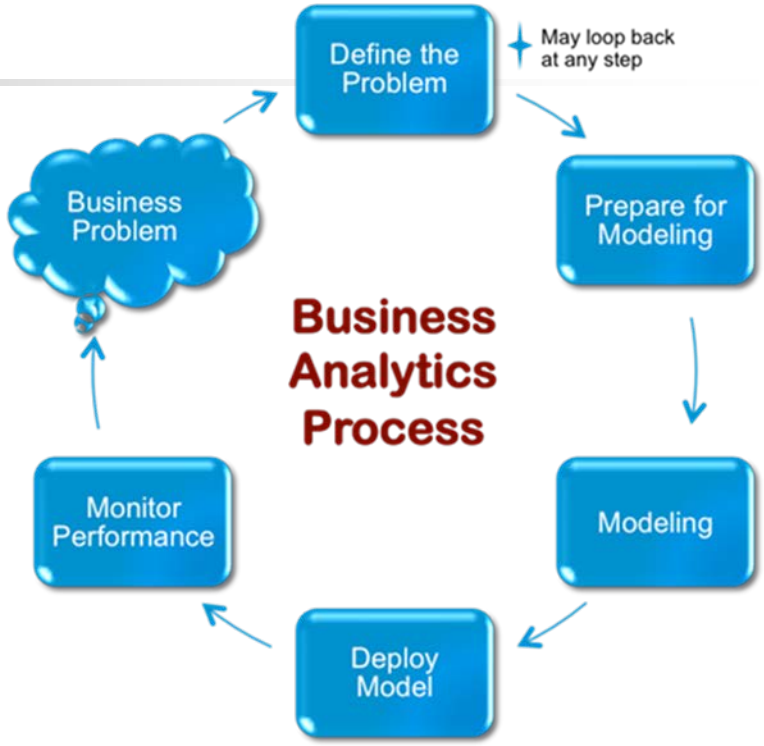
- Collect, clean, and transform data
- Define features
- Examine and understand data
- Data is ready for analysis and model-building



Adapted from the book "Building Better Models with JMP Pro" by Gayson, Gardner and Stephens and copyrighted by SAS

Build Model

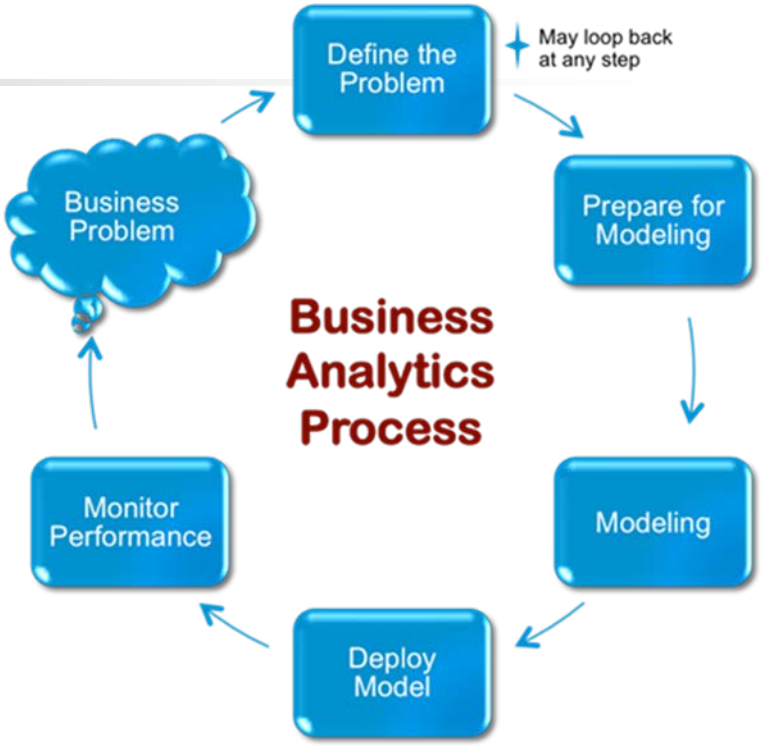
- Select, assess, and validate models
- Tests the best model
- Benchmark performance
- Validate the selected model



Adapted from the book "Building Better Models with JMP Pro" by Gayson, Gardner and Stephens and copyrighted by SAS

Deploy Model

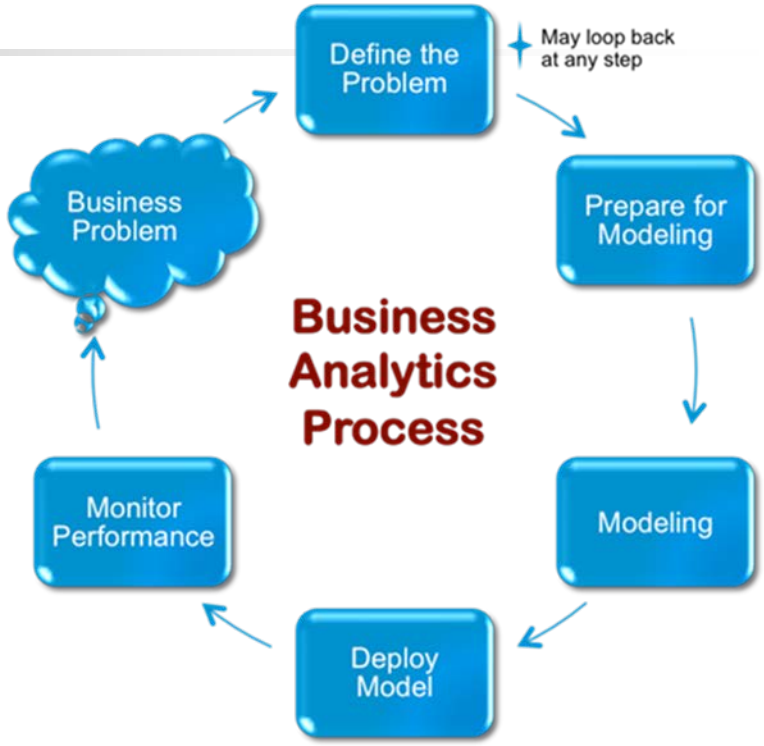
- Deliver the model to clients
- Assists in implementation of model
- Confirm performance
- Meet with sponsor and close the project



Adapted from the book "Building Better Models with JMP Pro" by Gayson, Gardner and Stephens and copyrighted by SAS

Monitor Performance

- Monitor model performance
- Refine model if needed
- Evaluate business benefit



Adapted from the book "Building Better Models with JMP Pro" by Gayson, Gardner and Stephens and copyrighted by SAS

Sources and Types of Data

Sources and Types of Data

- Internal versus External Sources
  - Primary versus Secondary
- Structured versus Unstructured

Internal vs. External

Internal Sources of Data

- Data sources that are within the firm (company)
  - Transaction data (accounting based)
  - Operations data
  - Call center data
  - Sensor data
  - …
- Relevant, cheap and easy to get to

External Sources of Data

- Data that does not exist within the firm (company) and has to be collected or obtained
  - **Primary data** – where data are collected by the analyst from customers, suppliers, dealers, employees, others through surveys and other methods
  - **Secondary data** – where data exists (already collected by someone else) and available (may be at a cost)
    - Government
    - Private companies
    - Trade organizations….

External Data Sources (A Few Exemplars)

- US Government (mostly free):
  - http://www.census.gov/
  - http://www.census.gov/data.html
  - http://www.census.gov/content/census/en/programs-surveys/surveys-programs.html/
  - http://www.bls.gov/tus/datafiles_2015.htm
  - And many more….
- Private companies (you have to pay)
  - http://www.experian.com/marketing-services/insource-demographics.html
  - http://www.equifax.com/compiled-data/en_us
  - http://www.myfico.com/consumer-division-of-fico.aspx
  - https://www.iriworldwide.com/en-US
  - And many more…

Structured vs. Unstructured Data

Customer	Age	Income	Gender	Response	…	Target
John	30	1200	M	No		0
Sarah	25	800	F	Yes		1
Sophie	52	2200	F	Yes		1
David	48	2000	M	No		0
Peter	34	1800	M	Yes		1

Top Customer Reviews

★★★★★ **This book is highly recommended**

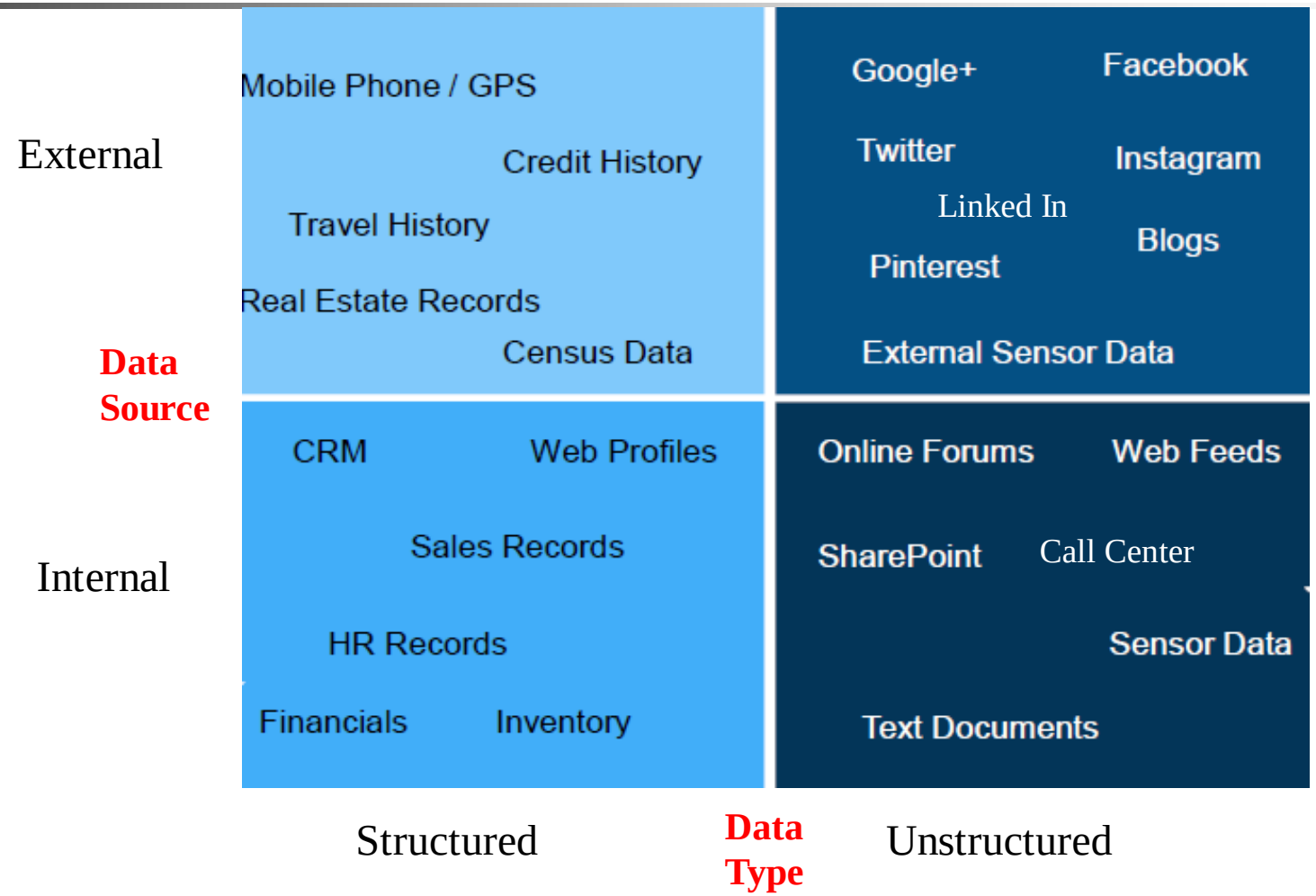
By Dr. Edi Shivaji on May 10, 2014

Format: Kindle Edition

This book is a splendid and valuable addition to this subject. The whole document is well written and I have no hesitation to recommend that this can be adapted as a text book for graduate courses in BIDM. The students will find it easy to read and follow, The style is lucid and explains the concepts and wholeness without going into any complicated maths.
The author has taken immense pains to keep the writing style simple and easy to follow. This book develops the intuition and generates an interest among the students in this field. Even a casual reader of the book will be left with a keen desire to learn more because this book nicely and clearly brings out the practical benefits. The Case exercises are very good and nicely planned. The book is very reasonably priced.

Comment | 12 people found this helpful. Was this review helpful to you? Yes No Report abuse

Analytics Opportunities in Data Space



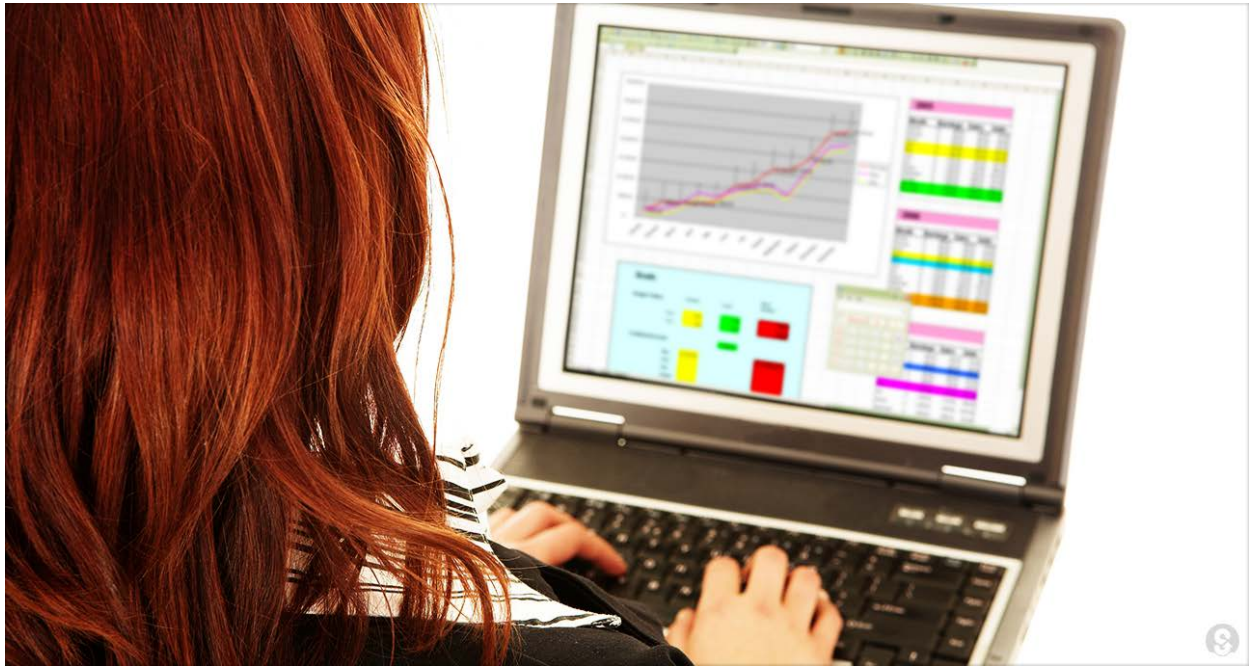
Source: Data and Technology Perspective by Ramesh Nair and Andy Narayanan

Different Types of Variables

Types of Variables

- Dependent versus Independent variable
- Categorical versus Continuous variable
  - Nominal, Binary, Ordinal, Interval and Continuous
- Reliability vs. Validity of measures
- Recommended analysis for each type of variable

Two Terms for the Same Thing



Generally, business people refer to data in Rows and Columns.

In analytics, we talk about Observations (records) and Variables

Dependent versus Independent

- Dependent (or, Target) is the variable of interest in predictive modeling.
  - Dependent variable *depends* on other independent variables
  - Example, sales of a product depends on its price among other variables
- Independent (or, Input) variables usually do not depend on other variables
- A predictive model is a representation of how the dependent variable relates to the independent variables
- **Question for you to think about**: Can an analytics project have **only** independent variables? What may be examples of such projects?
  - Answers will be discussed in the lab

Categorical Versus Continuous

- Before you run any analytics with any numbers in a data set, think about:
  - What those numbers are really representing?
  - How were the numbers created/generated/reported?

What Can we do with the Numbers in Column B?

	A	B	C	D
1	1	1	1	1
2	2	0	2	2
3	3	1	3	3
4	4	0	1	4

	A	B
1	ID	Gender
2	1	1
3	2	0
4	3	1
5	4	0

What Can we do with the Numbers in Column C?

	A	B	C	D
1	1	1	1	1
2	2	0	2	2
3	3	1	3	3
4	4	0	1	4

	A	B	C
1	ID	Gender	Race
2	1	1	1
3	2	0	2
4	3	1	3
5	4	0	1

What Can we do with the Numbers in Columns A and D?

	A	B	C	D
1	1	1	1	1
2	2	0	2	2
3	3	1	3	3
4	4	0	1	4

	A	B	C	D	E
1	ID	Gender	Race	Income group	Income
2	1	1	1	1	Less than 50,000
3	2	0	2	2	Between 50 - 75,000
4	3	1	3	3	Between 75-100,000
5	4	0	1	4	Above 100,00

What Can we do with the Numbers in Column F?

	A	B	C	D	E	F
1	ID	Gender	Race	Income group	Income	Income
2	1	1	1	1	Less than 50,000	38,750
3	2	0	2	2	Between 50 - 75,000	64,920
4	3	1	3	3	Between 75-100,000	75,140
5	4	0	1	4	Above 100,00	121,300

Variable Types

- Categorical (Discrete) Variable
  - Nominal (multiple categories)
    - Binary or, dichotomous (exactly two categories)
  - Ordinal (multiple categories with an ordering)
- Continuous Variable
  - Interval (measured on a continuum)
  - Ratio (measured on a continuum and 0 denotes absence of property being measured)
- **Question for you to think about**:
- Is it twice as hot when it is 80^0 F compared to when it is 40^0 F?
- Is an object weighing 80 lbs twice as heavy as an object weighing 40 lbs?
  - Answers will be discussed in the lab

Ambiguities in Variable Types

- Count data (such as number of times a customer has visited our website in last 6 months)?
- What about 5-point Likert type scale? Or, semantic-differential scale?

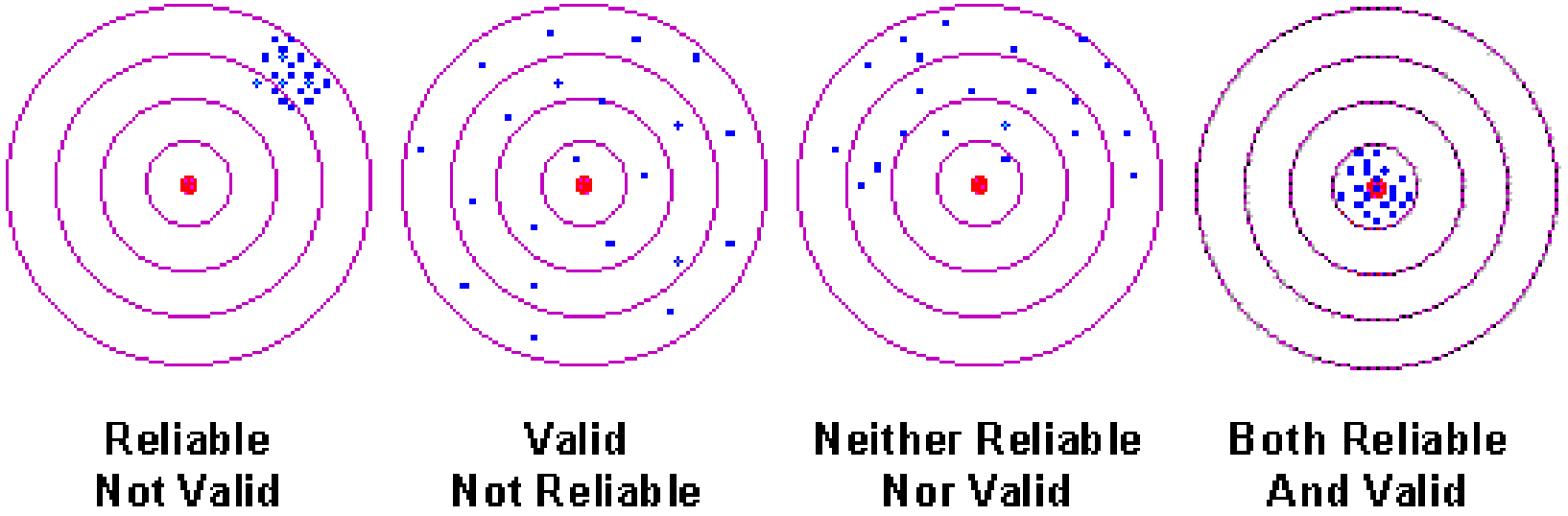
	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
The cashier was courteous.	○	○	○	○	○
The cashier was professional in appearance	○	○	○	○	○
I was given a receipt at the end of my transaction.	○	○	○	○	○

1. Please rate the President of the United States on the following traits:

Strong	○	○	○	○	○	Weak
Decisive	○	○	○	○	○	Indecisive
Good	○	○	○	○	○	Bad
Active	○	○	○	○	○	Passive
Industrious	○	○	○	○	○	Lazy
Happy	○	○	○	○	○	Sad

Construct Reliability and Validity

- **Construct validity** is "the degree to which something measures what it claims, or purports, to be measuring.
- **Reliability** in statistics and psychometrics is the overall consistency of a measure. A measure is said to have a high reliability if it produces similar results under consistent conditions.



Recommended Analysis for Each Variable Type

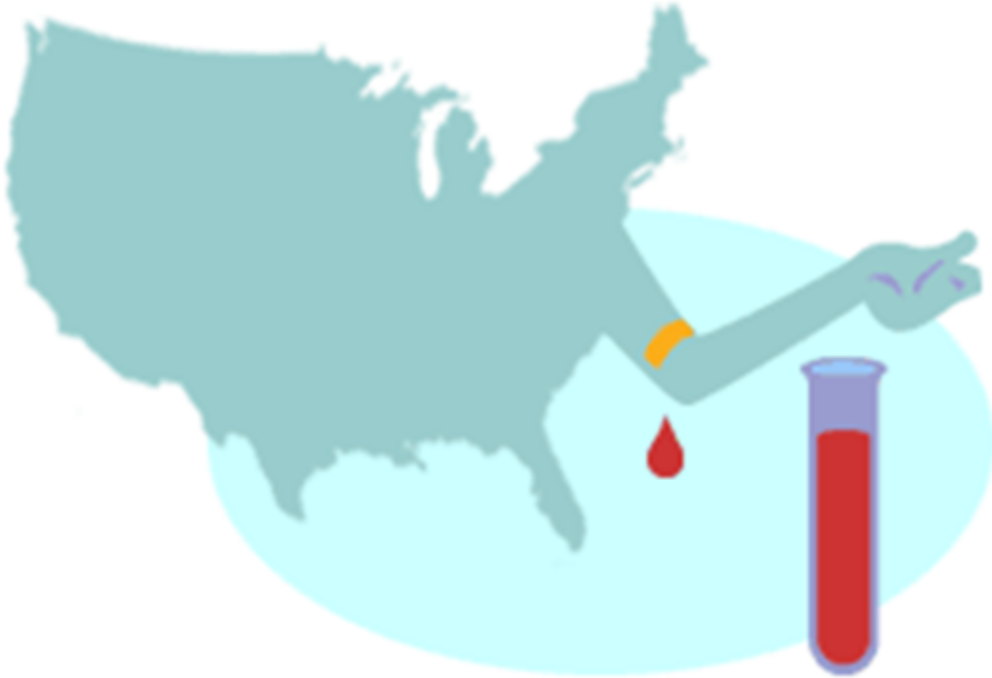
- Nominal, Binary or Ordinal:
  - Visualization (univariate): Bar chart, pie-chart, …
  - Visualization (bivariate/multivariate): Mosaic plot, Tile plot…
  - Statistics : Count (or, percentage of observations in each group), Mode (the most common group)
- Interval or Ratio:
  - Visualization (univariate): Histogram, box-plots, density plot…
  - Visualization (bivariate/multivariate): Scatter plot, Scatter plot with density ..
  - Statistics : Mean, median, variance/standard deviation, range, etc.

Sample, Population and Confidence Intervals

Agenda

- Sample and population
- Different sampling methods
- Confidence Intervals for descriptive statistics

Important Concepts & Terms



"If you don't believe in sampling, next time you go to the doctor for a blood test, have them take it all." -- ABCNEWS.com web site on polling

- Sample and Population
  - Always, try and take a *random* and *representative* sample from the population of interest

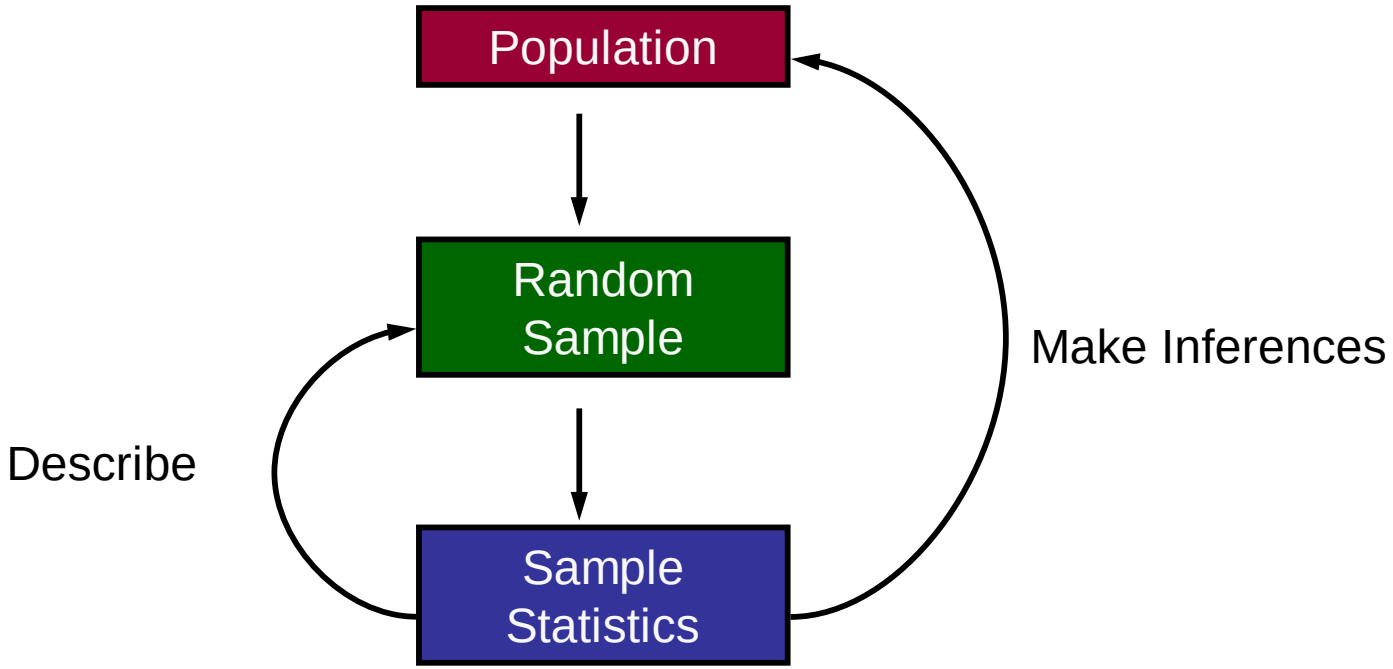
Important Concepts & Terms (contd.)

- Non- probability sampling:
  - Convenient sampling
  - Quota sampling
  - Other methods….
- Probability sampling:
  - Simple random sampling
    - Systematic sampling
  - Stratified sampling
  - Cluster sampling
  - Other methods…

Important Concepts & Terms (contd.)

- A big challenge for business analysts
  - *Impractical* to measure entire population
  - But we are interested in the population. What to do?
    - Take a sample from the population
    - Generalize the metrics from the sample to the population
  - Question : how confident are we in our results?
- Caution: statistical decision making is not an exact science!

Process of Statistical Data Analysis



Confidence Intervals

- Can be calculated for most business metrics
- Generally, we work with 95% confidence level
  - Sometimes we may consider 99% or 90%
- Layman's explanation of 95% CI: we are "95% sure the population business metric will be within the confidence interval calculated form the sample business metric"
  - Example: A company makes a new "BOGO" offer on its web site for one of its product. Runs the campaign for a day.
    - *Business metric*: % of web site visitors who took the offer
    - Sample business metric: sample conversion rate is 12.3%
    - A 95% CI based on that sample, works out to be between 10.1 to 14.5%
    - The 95% CI can be interpreted as …..
  - Demo of calculating CIs in labs

Why Statistics?

Why Study Statistics

- One reason:
  - Because if you had one day left to live, a statistics class could make it seem like eternity
- Other good reasons
  - For analytics/data science jobs, statistical modeling and analysis is a very important skill set
    - "The evolution of Analytics….." by Bowers, Camm and Chakraborty in *Interfaces*, an INFORMS journal
  - Making sense of numerical information, dealing with uncertainty (probability), extending results from a sample to a population, forecasting and prediction,…

Traditional View

- A set of techniques that are utilized to process numerical data.
- *Statistics* is the body of methods for the collection, analysis, presentation, interpretation, and use of data.
- Heavily influenced by way statistics used in the scientific process

Modern View

- Came about from
  - The process view of a business organization
  - Changes in Statistical Pedagogy
  - Changes in the V’s of data and advancement and challenges in IT technology for handling such data

The Process View

- Arose out of Deming's work and following movements of total quality management and six sigma
- A process is something that takes inputs and transforms them into outputs

Changes in Statistical Pedagogy

- *Statistical thinking* is a philosophy of learning and action based on three fundamental principles:
  - All work consists of a system of interconnected processes.
  - Variability exists in all processes.
  - Understanding and reducing variability are the keys to improving a process.

Variability

- Deviation of a value from an average or expected outcome.
- Two kinds of process variability
  - Common cause
  - Special cause
- Distinguishing between common and special cause variability is key to process control and improvement

Common Cause Variability

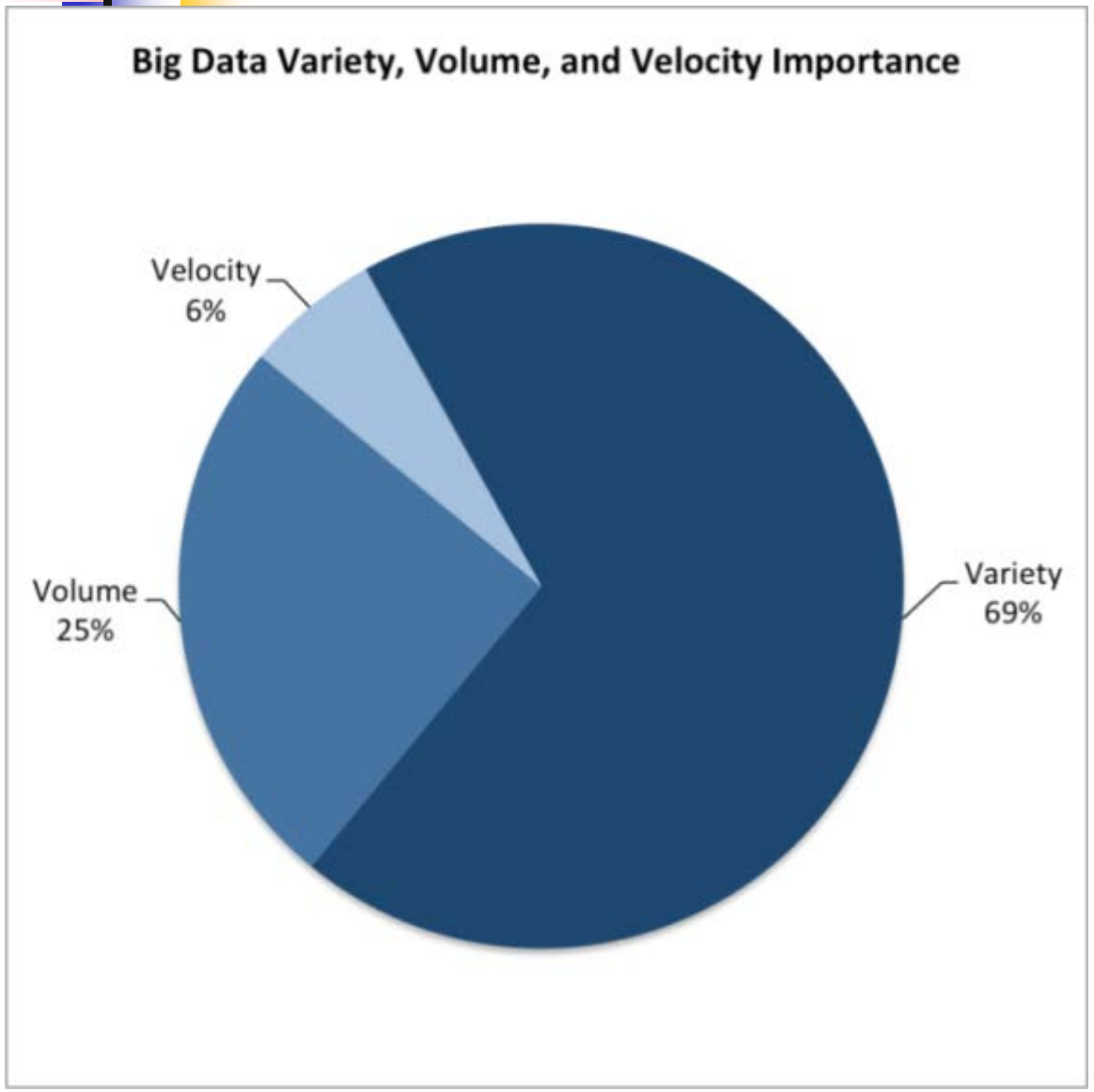
- Random in nature
- Due to many small causes
- Does not require special action
- Should be measured and tracked

Special Cause Variability

- Due to one or a few causes
- Indicates a problem or that something has changed
- Requires action

Challenges in IT for Managing the V's of Data

Volume, Velocity, Variety ……



http://sloanreview.mit.edu/article/variety-not-volume-is-driving-big-data-initiatives/?utm_source=twitter&utm_medium=social&utm_campaign=sm-direct

How is modern view different?

- Move from “affirmation/confirmation” to “discovery”
  - Many newer statistical software packages are oriented more toward discovery by empowering users with tools that are easy to use
- Move from heavy use of feature-engineering by users to automated feature-selection and predicting by complex models
  - Machine learning and deep learning methods at their core uses models developed by statisticians!