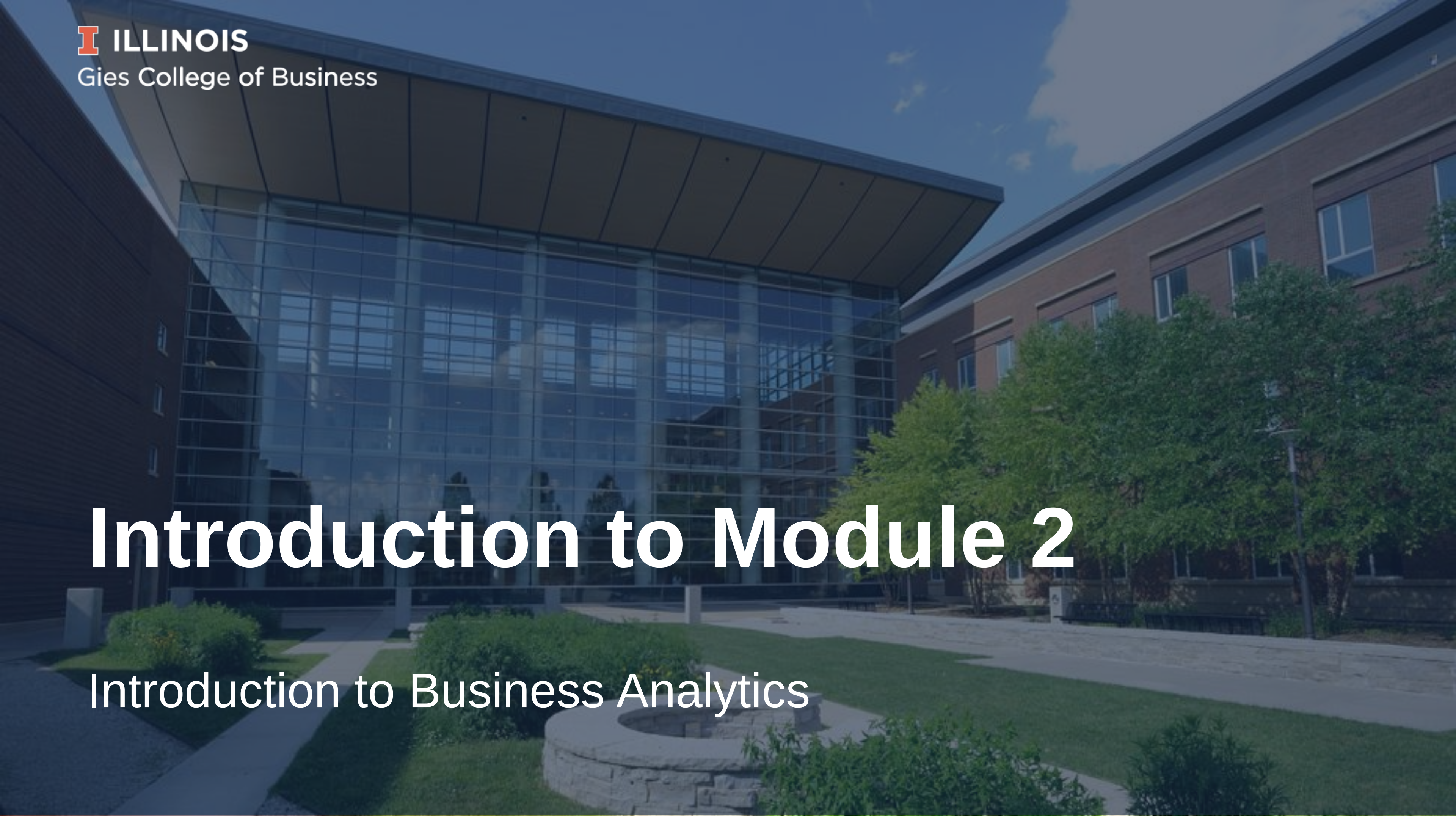
ILLINOIS
Gies College of Business

Introduction to Business Analytics

Ron Guymon

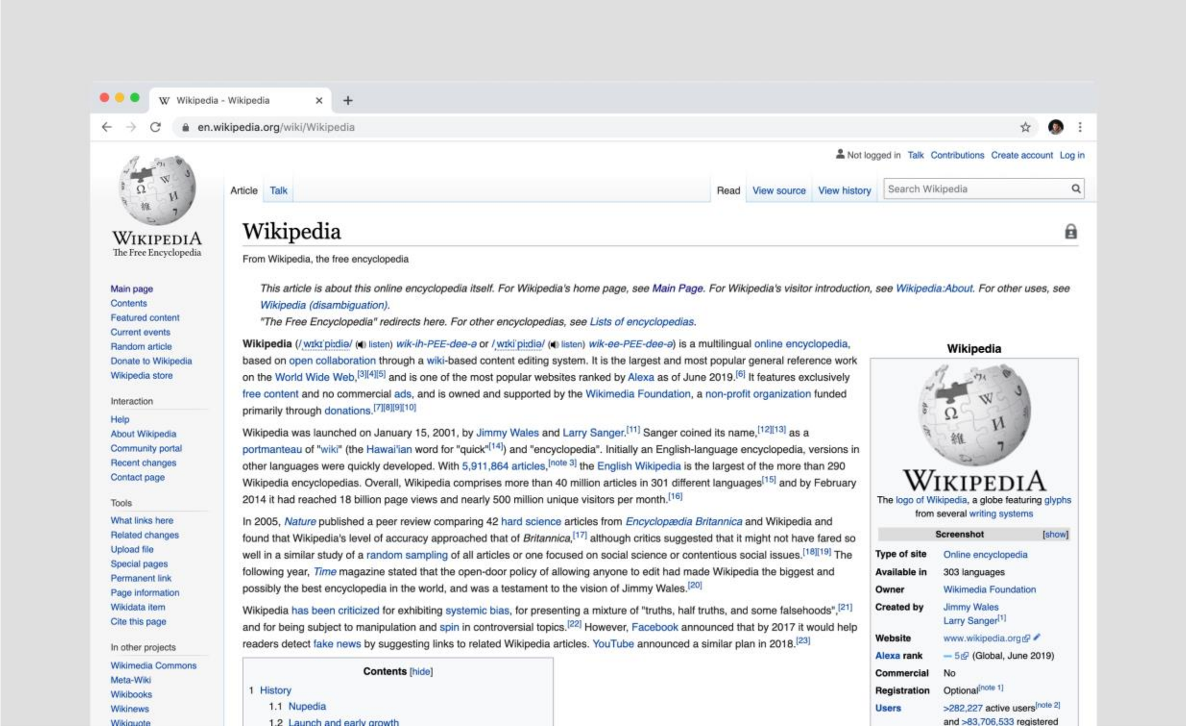
ILLINOIS
Gies College of Business



Introduction to Module 2

Introduction to Business Analytics

What Is Data?



Wikipedia

From Wikipedia, the free encyclopedia

This article is about this online encyclopedia itself. For Wikipedia's home page, see Main Page. For Wikipedia's visitor introduction, see Wikipedia:About. For other uses, see Wikipedia (disambiguation).

"The Free Encyclopedia" redirects here. For other encyclopedias, see Lists of encyclopedias.

Wikipedia (/ˌwɪkɪˈpiːdiə/ (listen) *wik-ih-PEE-dee-ə* or /ˌwɪkiˈpiːdiə/ (listen) *wik-ee-PEE-dee-ə*) is a multilingual online encyclopedia, based on open collaboration through a wiki-based content editing system. It is the largest and most popular general reference work on the World Wide Web,[3][4][5] and is one of the most popular websites ranked by Alexa as of June 2019.[6] It features exclusively free content and no commercial ads, and is owned and supported by the Wikimedia Foundation, a non-profit organization funded primarily through donations.[7][8][9][10]

Wikipedia was launched on January 15, 2001, by Jimmy Wales and Larry Sanger.[11] Sanger coined its name,[12][13] as a portmanteau of "wiki" (the Hawai'ian word for "quick"[14]) and "encyclopedia". Initially an English-language encyclopedia, versions in other languages were quickly developed. With 5,911,864 articles,[note 3] the English Wikipedia is the largest of the more than 290 Wikipedia encyclopedias. Overall, Wikipedia comprises more than 40 million articles in 301 different languages[15] and by February 2014 it had reached 18 billion page views and nearly 500 million unique visitors per month.[16]

In 2005, *Nature* published a peer review comparing 42 hard science articles from *Encyclopædia Britannica* and Wikipedia and found that Wikipedia's level of accuracy approached that of *Britannica*,[17] although critics suggested that it might not have fared so well in a similar study of a random sampling of all articles or one focused on social science or contentious social issues.[18][19] The following year, *Time* magazine stated that the open-door policy of allowing anyone to edit had made Wikipedia the biggest and possibly the best encyclopedia in the world, and was a testament to the vision of Jimmy Wales.[20]

Wikipedia has been criticized for exhibiting systemic bias, for presenting a mixture of "truths, half truths, and some falsehoods",[21] and for being subject to manipulation and spin in controversial topics.[22] However, Facebook announced that by 2017 it would help readers detect fake news by suggesting links to related Wikipedia articles. YouTube announced a similar plan in 2018.[23]

Contents [hide]

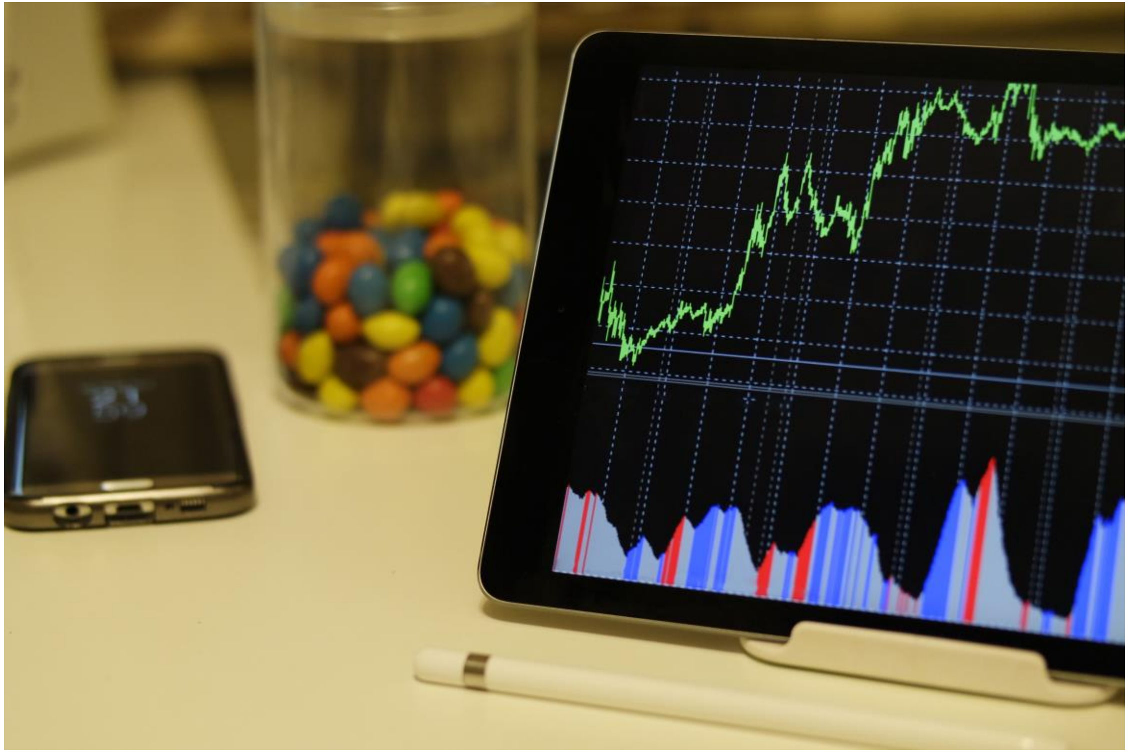
Wikipedia

The logo of Wikipedia, a globe featuring glyphs from several writing systems

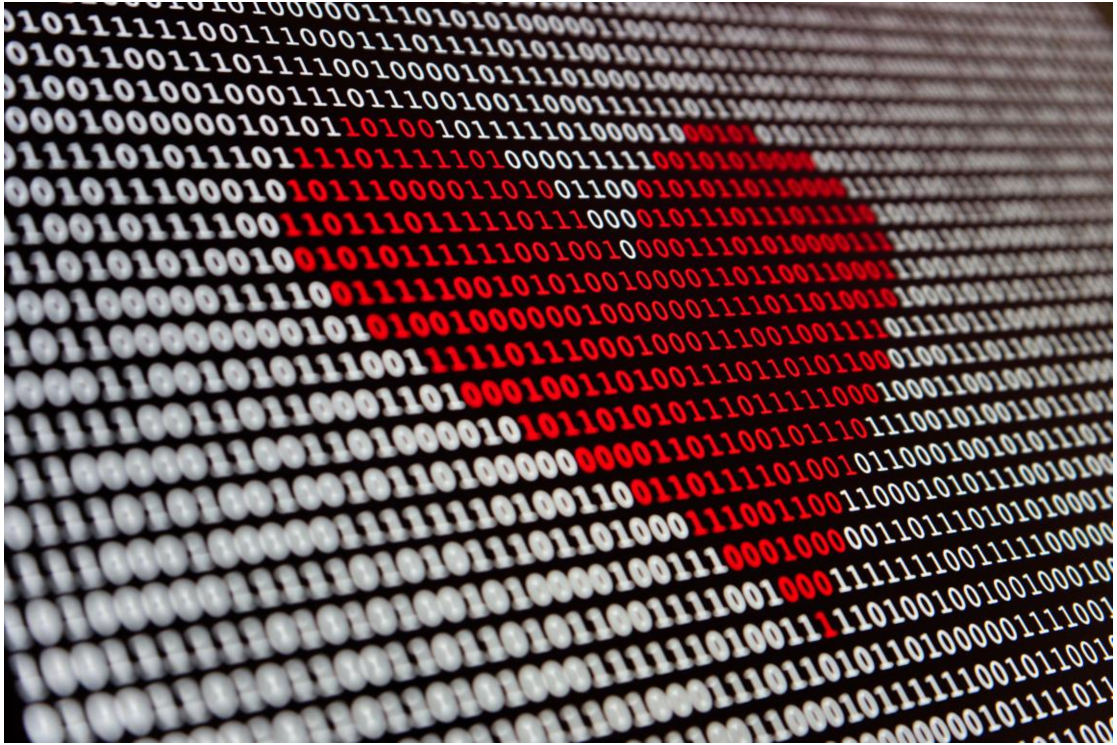
Screenshot [show]

Type of site	Online encyclopedia
Available in	303 languages
Owner	Wikimedia Foundation
Created by	Jimmy Wales Larry Sanger[1]
Website	www.wikipedia.org
Alexa rank	— 5 (Global, June 2019)
Commercial	No
Registration	Optional[note 1]
Users	>282,227 active users[note 2] and >83,706,533 registered

What Is Data?



What Is Data?



What Is Data?

What Is Data?

“Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation”

https://www.merriam-webster.com/dictionary/data

Module 2 Objectives

- Appreciate some common issues associated with assembling data.
- Investigate how assembling data is related to framing a question and calculating the results.
- Execute some fundamental data assembly tasks.

ILLINOIS

Gies College of Business

Framing a Question

Introduction to Business Analytics

FACT Framework

- Frame the question
- Assemble the data
- Calculate the results
- Tell others the results

Framing a Question Influences:

- Whether or not you will spend time supporting the organization’s goals
- Whether you even need to use data
- If you do could benefit from data, then it will influence the data that is assembled.

Frame a Question That

- Considers organization’s goals
- Is informed by domain knowledge
- Leads to action

Bad Examples of Framing a Question

- Why are things bad?

- Why are sales down?

Good Examples of Framing a Question

- In what regions are sales lower than expected?
- What factors caused a decline in sales for regions in which sales are lower than expected?

Great Examples of Framing a Question

- What factors are most influential in identifying regions that have sales that are statistically lower than expected?

- What factors cause a statistically significant decline in sales for regions in which sales are lower than expected?

ILLINOIS
Gies College of Business

Assembling Data

Introduction to Business Analytics

FACT Framework

- Frame the question
- Assemble the data
- Calculate the results
- Tell others the results

Steps for Assembling Data

- Finding data
- Extract, transform, load (ETL)
- Data wrangling

Chief Data Officers “have been chartered with improving the efficiency and value-generating capacity of their organization’s information ecosystem. That is, they’ve been asked to lead their organization in treating and leveraging information with the same discipline as its other, more traditional assets.”

Laney, Doug. *Infonomics*, p. 9

A Few Data Sources on the Web

www.data.gov

www.google.com/publicdata/directory

aws.amazon.com/opendata/public-datasets/

docs.microsoft.com/en-us/azure/sql-database/sql-database-public-data-sets

https://www.kaggle.com/datasets

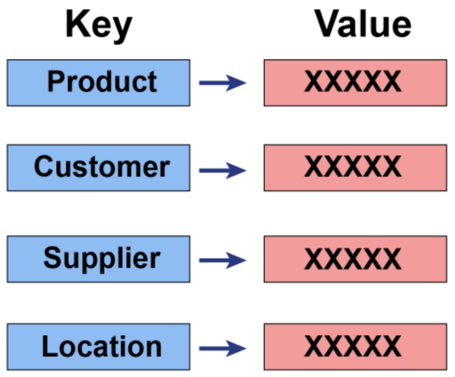
https://data.world/

https://www.gapminder.org/data/

ETL

- Extract
- Transform
- Load

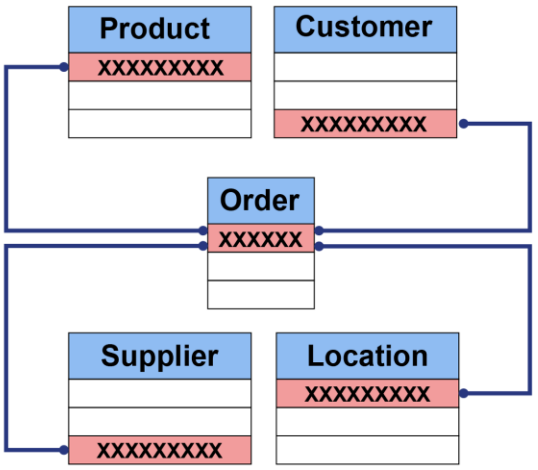
Key-Value Database



Key		Value
Product	→	XXXXX
Customer	→	XXXXX
Supplier	→	XXXXX
Location	→	XXXXX

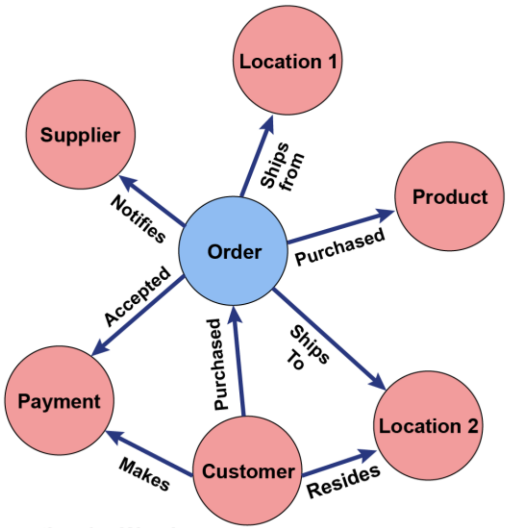
- Highly fluid schema/no schema
- High performance for simple transactions
- Poor performance deep analytics

Relational Database



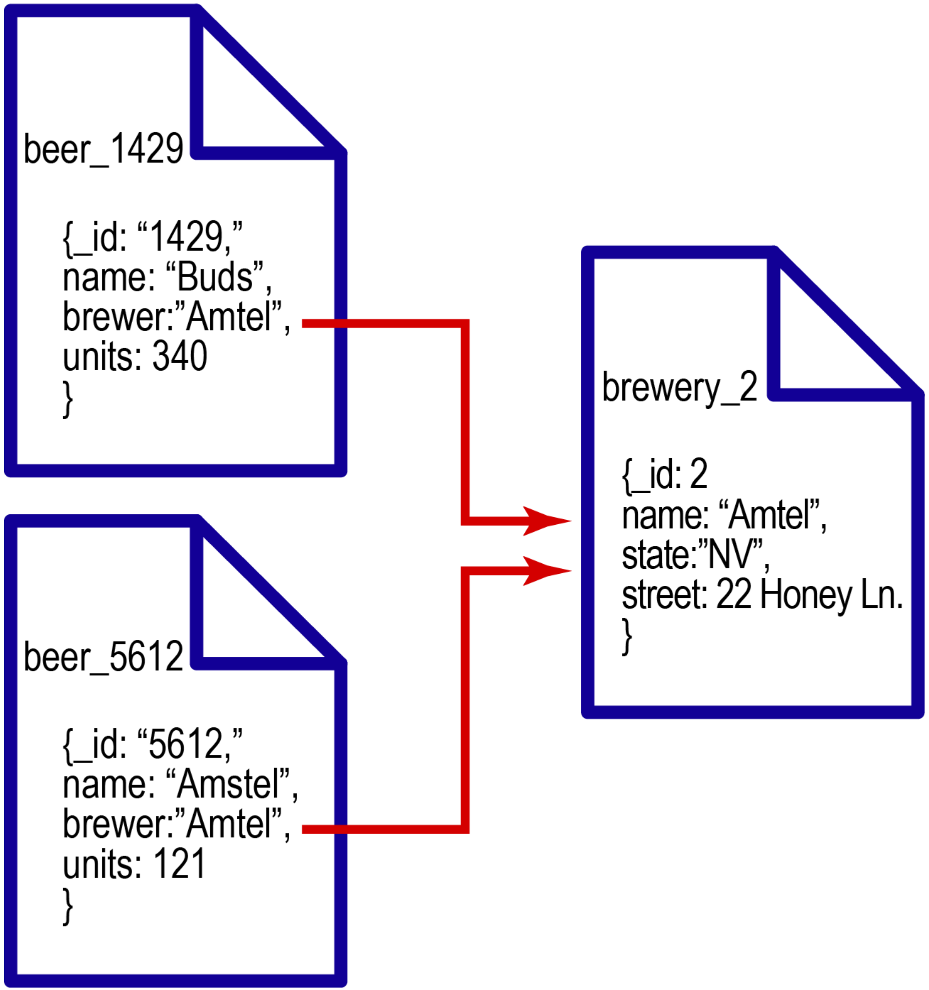
- Rigid schema
- High performance for transactions
- Poor performance for deep analytics

Graph Database



Location 1 = Warehouse
Location 2 = Deliver Location

- Flexible schema
- High performance for complex transactions
- High performance for deep analytics



beer_1429

```
{_id: “1429,”
name: “Buds”,
brewer:”Amtel”,
units: 340
}
```

beer_5612

```
{_id: “5612,”
name: “Amstel”,
brewer:”Amtel”,
units: 121
}
```

brewery_2

```
{_id: 2
name: “Amtel”,
state:”NV”,
street: 22 Honey Ln.
}
```

```
imestamp":"2017-06-03T18:42:18.018", "deltaStartMillis
lass":"com.orgmanager.handlers.RequestHandler", "method":"handle
izeChars":"5022", "message":"Duration Log", "durationMillis":"508"}{
ebURL":"/app/page/analyze", "webParams":"null", "class":"com.orgmanager
equetID":"8249868e-afd8-46ac-9745-839146a20f09", "sessionID":"144o2n620jm9trnd3s3n7
urationMillis":"36"}{"timestamp":"2017-06-03T18:43:335.030", "deltaStartMillis":"211
ebParams":"file=chartdata_new.json", "class":"com.orgmanager.handlers.RequestHandler
essionID":"144o2n620jm9trnd3s3n7wg0k", "sizeChars":"48455", "message":"Duration Log
eltaStartMillis":"0", "level":"INFO", "webURL":"/app/page/report", "webParams":"null
equestID":"789d89cb-bfa8-4e7d-8047-498454af885d", "sessionID":"144o2n620jm9trnd3s3n7w
urationMillis":"7"}{"timestamp":"2017-06-03T18:46:921.000", "deltaStartMillis":"0",
lass":"com.orgmanager.handlers.RequestHandler", "method":"handle", "requestID":"7ac6
izeChars":"10190", "message":"Duration Log", "durationMillis":"10"}{"timestamp":"2017
ebURL":"/app/rest/json/file", "webParams":"file=chartdata_new.json", "class":"com.org
equestID":"7ac6ce95-19e2-4a60-88d7-6ead86e273d1", "sessionID":"144o2n620jm9trnd3s3n7w
urationMillis":"23"}{"timestamp":"2017-06-03T18:42:18.018", "deltaStartMillis":"0",
lass":"com.orgmanager.handlers.RequestHandler", "method":"handle", "requestID":"b886
izeChars":"5022", "message":"Duration Log", "durationMillis":"508"}{"timestamp":"2017
ebURL":"/app/page/analyze", "webParams":"null", "class":"com.orgmanager.handlers.Requ
equetID":"8249868e-afd8-46ac-9745-839146a20f09", "sessionID":"144o2n620jm9trnd3s3n7w
urationMillis":"36"}{"timestamp":"2017-06-03T18:43:335.030", "deltaStartMillis":"211
ebParams":"file=chartdata_new.json", "class":"com.orgmanager.handlers.RequestHandler
essionID":"144o2n620jm9trnd3s3n7wg0k", "sizeChars":"48455", "message":"Duration Log
eltaStartMillis":"0", "level":"INFO", "webURL":"/app/page/report", "webParams":"null
equestID":"789d89cb-bfa8-4e7d-8047-498454af885d", "sessionID":"144o2n620jm9trnd3s3n7
urationMillis":"7"}{"timestamp":"2017-06-03T18:46:921.000", "deltaStartMillis
rgmanager.handlers.RequestHandler", "method":"handle
```

Data Wrangling

- Cleaning data
- Combining data with other data
- Cleaning data again
- Combining data with other data again
- Cleaning again
- Then changing its shape

ILLINOIS
Gies College of Business

Calculate the Results

Introduction to Business Analytics

FACT Framework

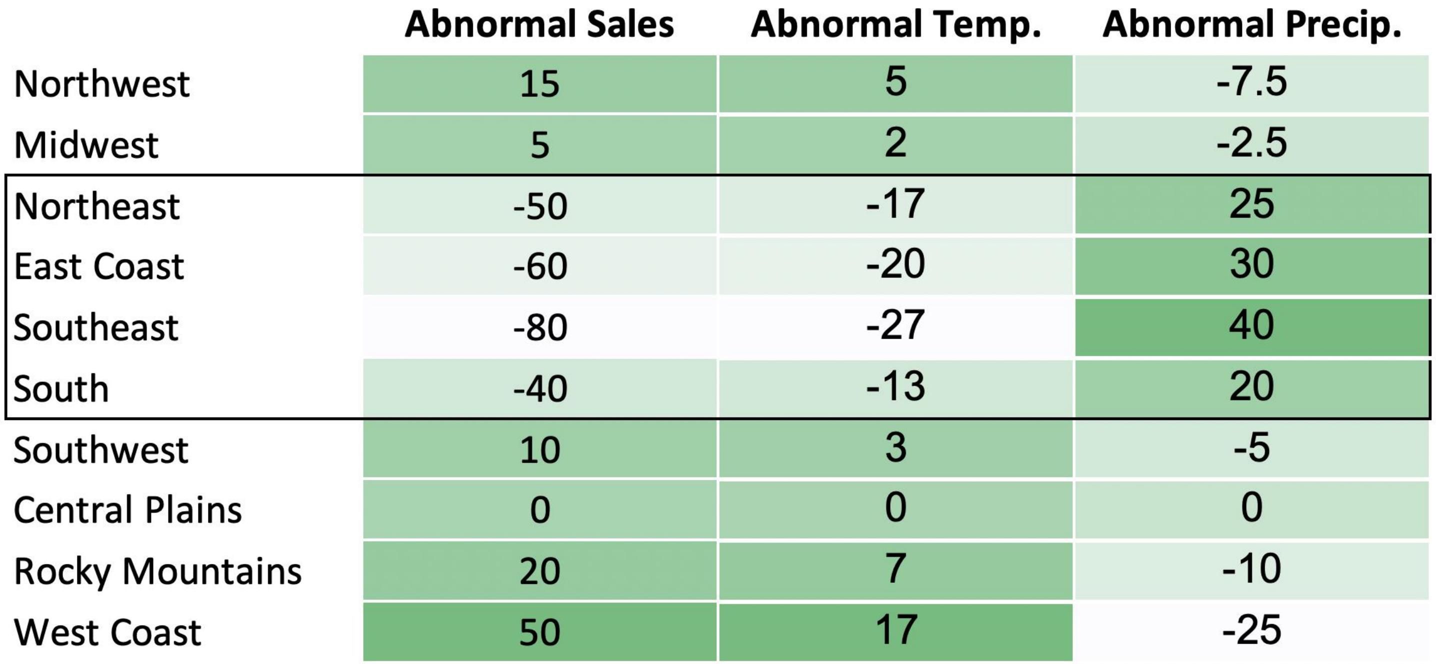
- Frame the question
- Assemble the data
- Calculate the results
- Tell others the results

Summary Statistics = Descriptive Statistics

Examples of Summary Statistics:

- Minimum
- Mean
- Median
- Mode
- Maximum
- Variance
- Standard Deviation
- Range

Sprinkler Sales

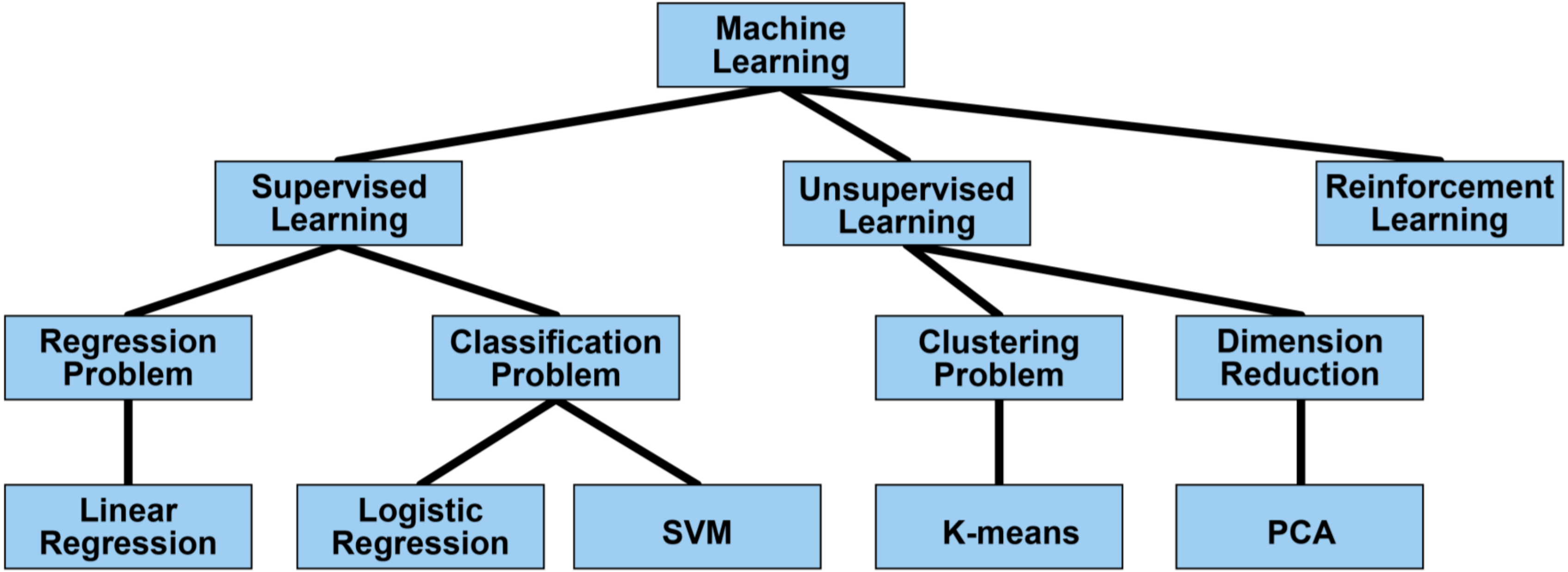


	Abnormal Sales	Abnormal Temp.	Abnormal Precip.
Northwest	15	5	-7.5
Midwest	5	2	-2.5
Northeast	-50	-17	25
East Coast	-60	-20	30
Southeast	-80	-27	40
South	-40	-13	20
Southwest	10	3	-5
Central Plains	0	0	0
Rocky Mountains	20	7	-10
West Coast	50	17	-25

Sprinkler Sales Model

Weekly Sales = 250 – 500*Precip + 250*Temp – 46*Wind

Categories of Machine Learning Algorithms



Considerations for Selecting an Algorithm

- Tradeoff between accuracy and the time to create the model
- How well does the algorithm deal with:
  - Nonlinearity
  - High dimensionality
  - Multicollinearity
  - Constant stream of new data
- Ability to explain the model to others

How Is Model Accuracy Evaluated

- Using a “loss” function that measures the difference between the right answers and the predicted answer

- Creating a model on a “training” dataset and then evaluating the performance on a separate “testing” dataset

ILLINOIS

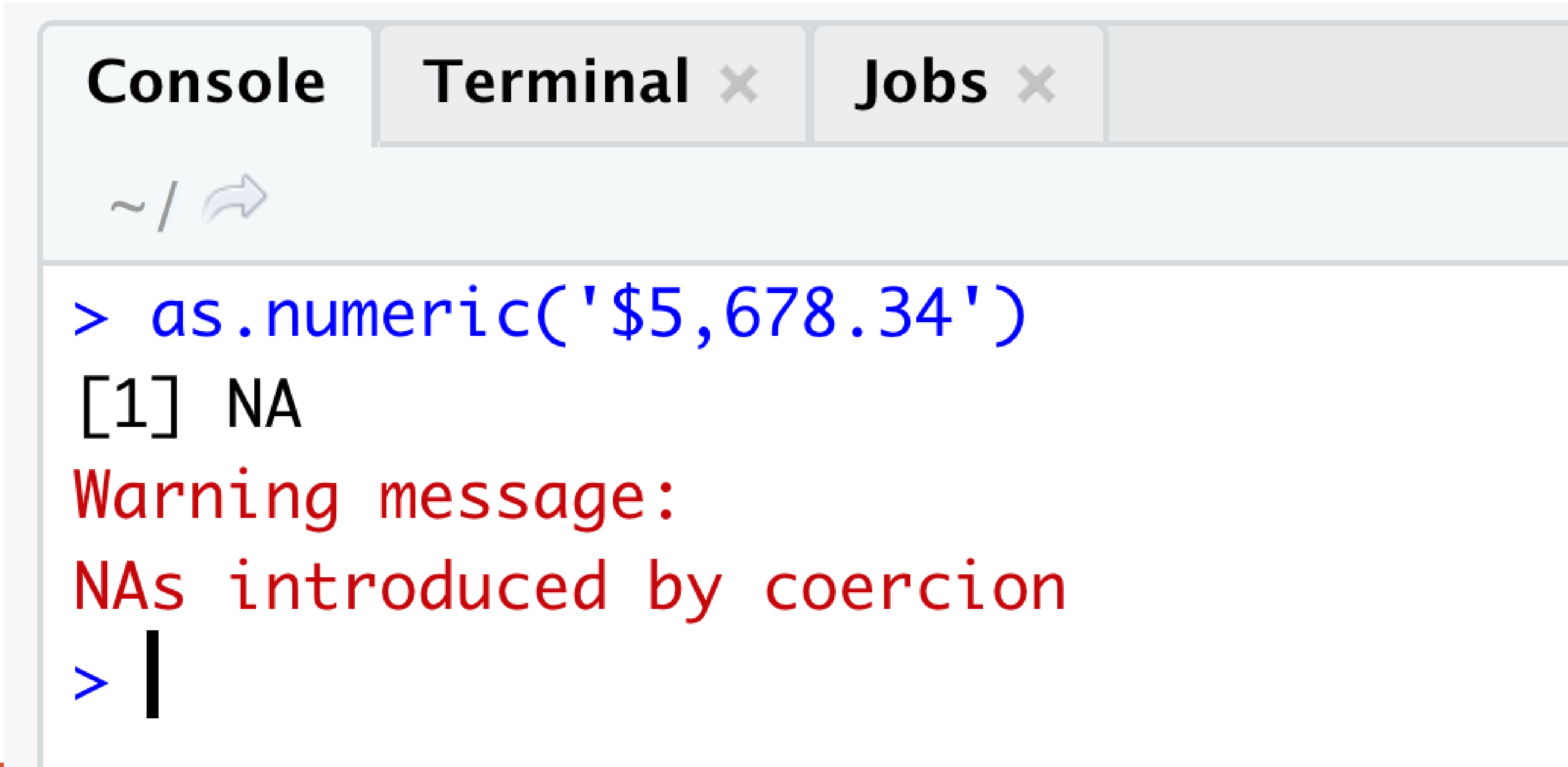
Gies College of Business

Data Types

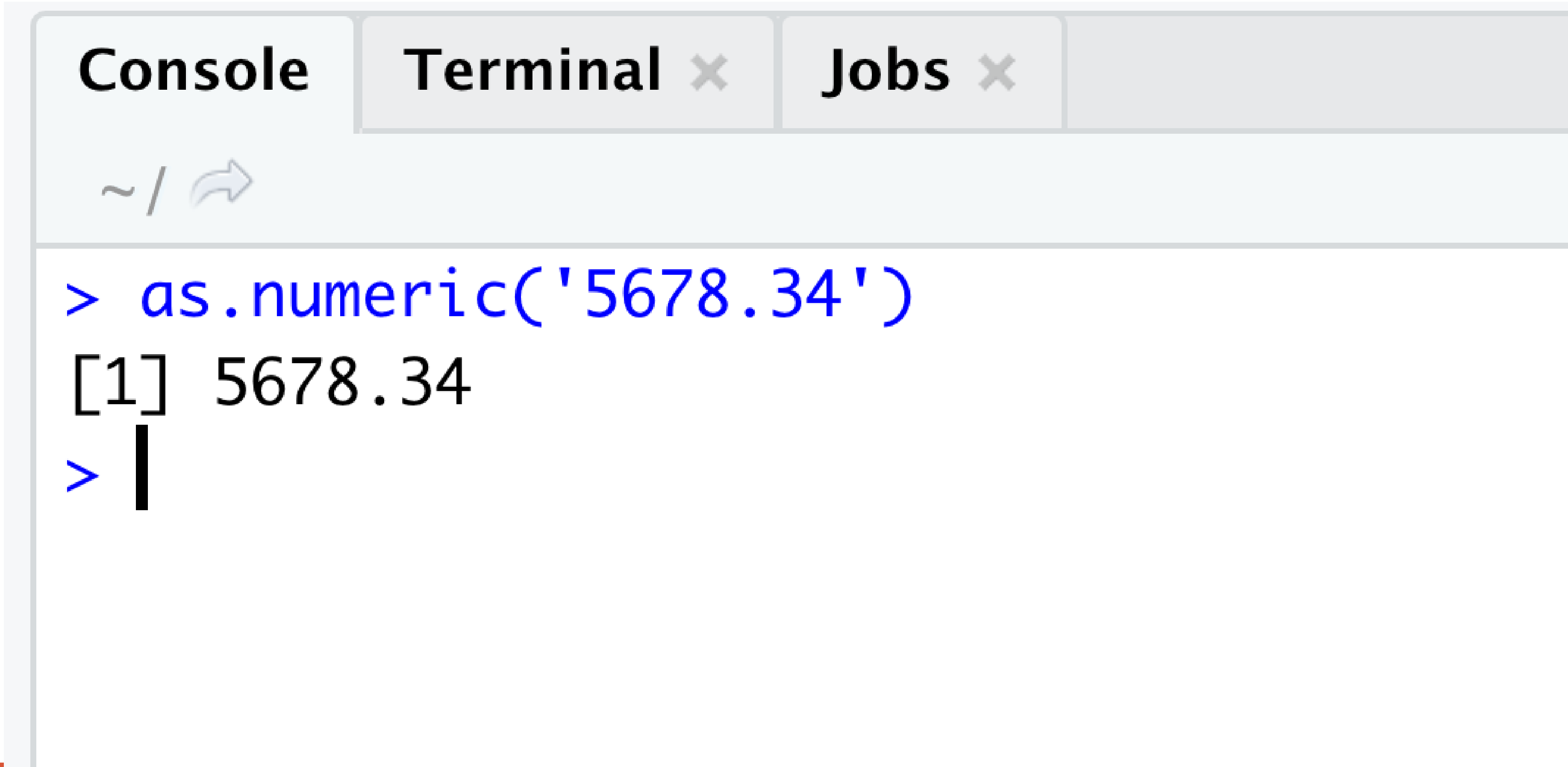
Introduction to Business Analytics

Data Types in This Lesson

- Character strings
- Numeric
- Datetimes
- Factors



```
> as.numeric('$5,678.34')
[1] NA
Warning message:
NAs introduced by coercion
> 
```



```
> as.numeric('5678.34')
[1] 5678.34
> 
```

Dates and Times

- Stored in a special integer format that displays them as dates
- Number of days that have passed since the start of a specific epoch
- March 3, 2005 is stored as the integer 38,415 in Excel.
- Timestamps represent the number of seconds that have passed since the start of an epoch.

What does 03/04/05 represent?

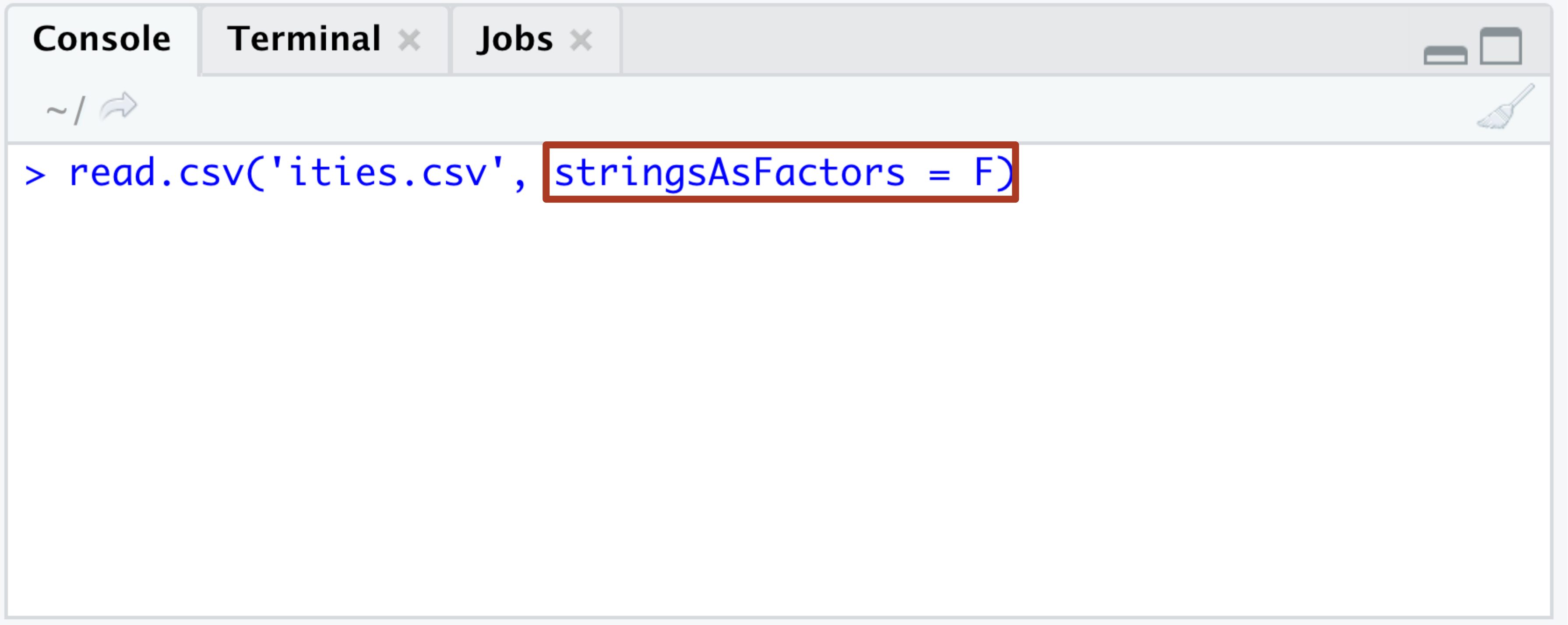
March 4, 2005

April 3, 2005

April 5, 2003

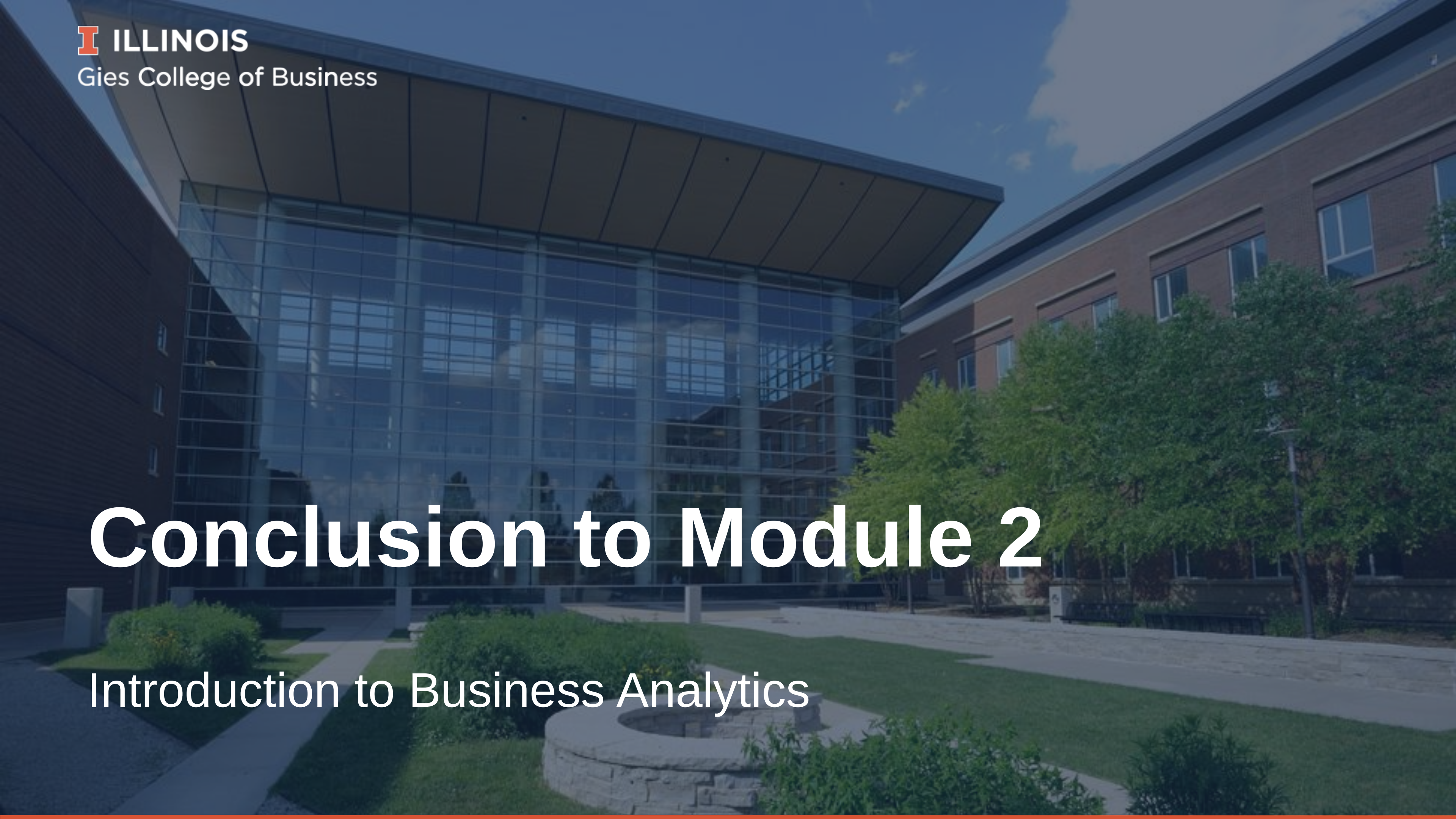
Additional Considerations for Dates and Times

- Time zones
- Daylight savings time
- Leap year
- Number of work days in a quarter
- Weekends
- Holidays



```
> read.csv('ities.csv', stringsAsFactors = F)
```

ILLINOIS
Gies College of Business



Conclusion to Module 2

Introduction to Business Analytics

MEAT BALLS
IN 3D

1:28