Practical Aspects of Database Design

L1 - Introduction

Stevens Institute of Technology

About FE513

Instructor Information

Xingjia Zhang
PhD student working on systemic risk with agent-based model
Desk: Babbio 5th fl, opposite the room 547
Email: xzhang21@stevens.edu
Office Hours: Tuesdays 3:00pm - 5:00pm, or by appointment.

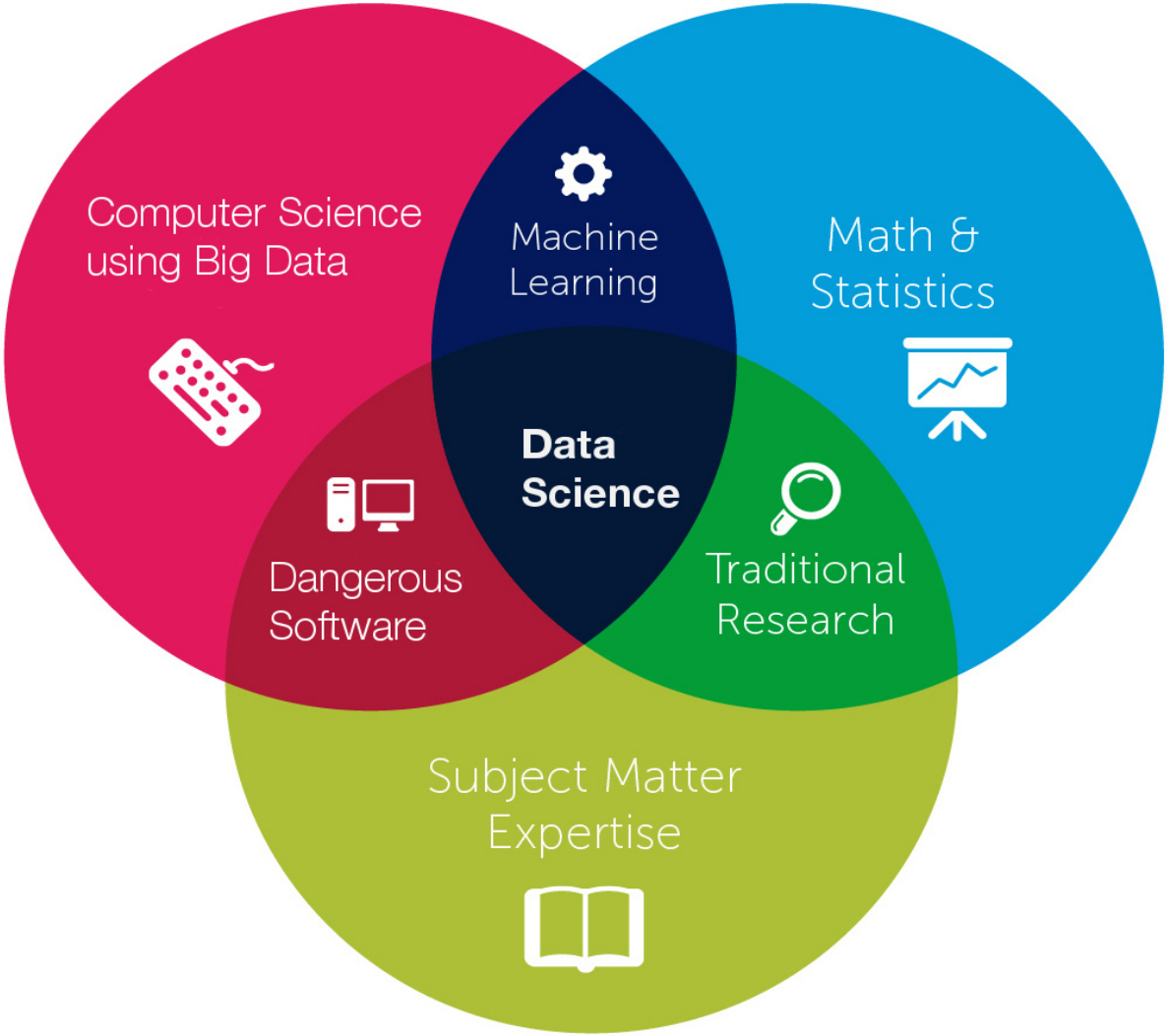
Course Policies

Your final grade will be determined based on your performance on each of the following items

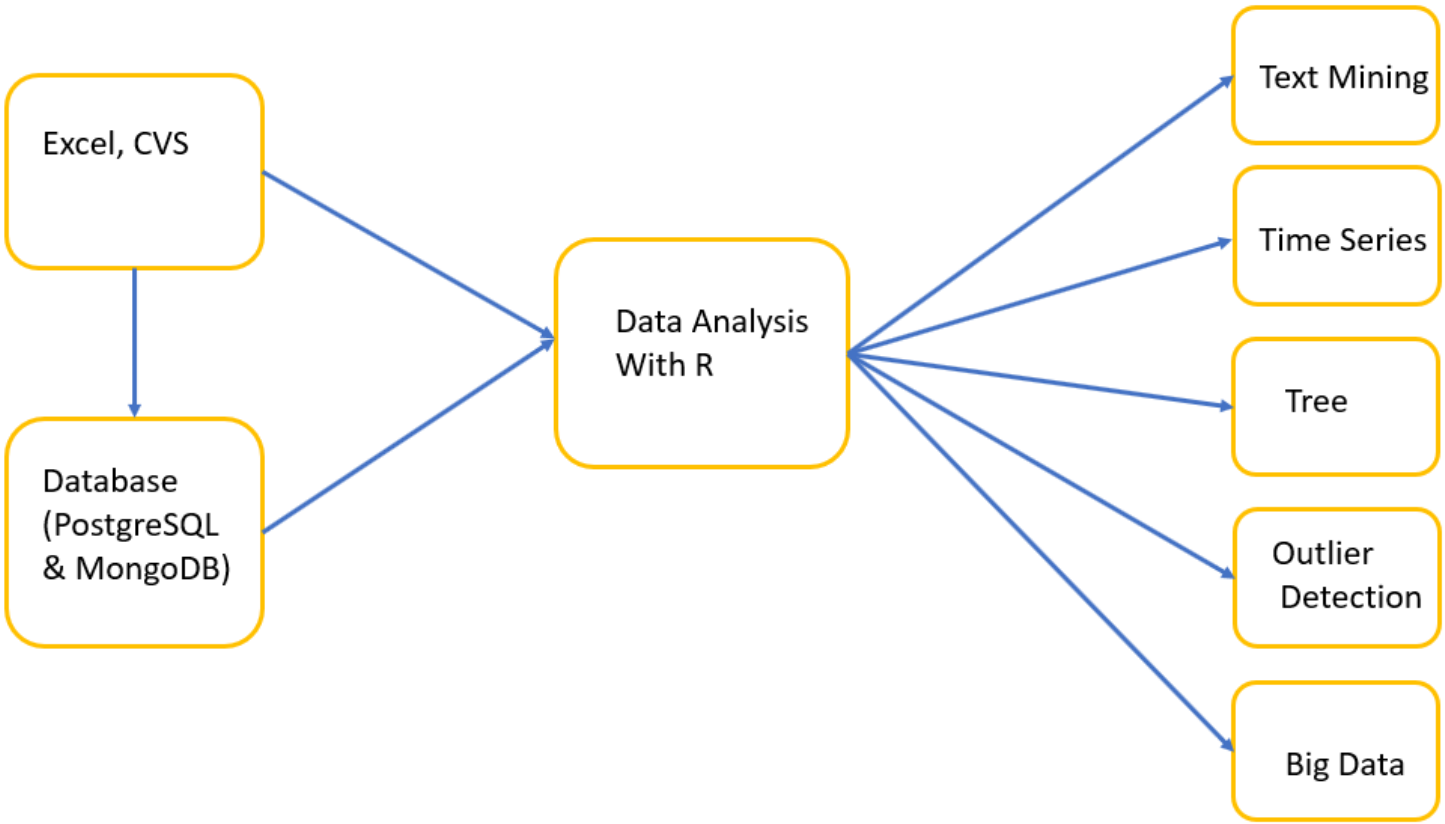
- Assignments (60%)
- Final exam (40%)

Grading scale may be adjusted depending on class average

Course Objectives



Course Outline



Recommended Readings

For R

- For beginner: interactive web tutorial: `https://www.datacamp.com/courses/free-introduction-to-r`
- Book "The Art of R Programming"
- Book "Cookbook for R"

For SQL

- For beginner: w3school web tutorial: `http://www.w3schools.com/sql/`
- Book "A First Course in Database Systems" (3rd edition) by Ullman and Widom (U/W)

For Data Science

- Book "Data Mining: The Textbook" by Charu C. Aggarwal

Recommended Practice

Kaggle

- Competitions: `https://www.kaggle.com/competitions`
- Tutorial for beginner: `https://www.kaggle.com/kanncaa1/data-sciencetutorial-for-beginners`
- Discussion: `https://www.kaggle.com/questions-and-answers?sortBy=top&group=all&page=1&pageSize=20&category=all&kind=all`

Software in FE513

R

download: `https://cloud.r-project.org/`

Rstudio

download: `https://www.rstudio.com/products/rstudio/download2/`

PostgreSQL

- windows: `https://www.enterprisedb.com/downloads/postgres-postgresql-downloads#windows`
- mac: `https://www.enterprisedb.com/downloads/postgres-postgresql-downloads#macosx`

Introduction to Financial data

Quantitative Data

- Structured, numeric data, such as stock return, ROA ratio, option price, etc.
- Analytical models can be easily applied on quantitative data.

Qualitative Data

- Unstructured, textual data, such as earnings announcement, news, etc.
- Categorical data can be easily transferred into numbers.
- More complex qualitative data needs additional procedure.

Stock daily data

- open/high/low/trading volume
- close (for daily): normally we use "adjust close" for modeling stock data in order to avoid the effects from dividend, stock split, etc.. It is the field we used to calculate stock return, variance.
- Note: we normally download stock data by TICKER. However, different database may have different symbol for stock.

Alphabet Inc. (GOOG)
NASDAQ Global Select - NASDAQ Global Select Real Time Price. Currency in USD

Add to watchlist

922.80 +16.14 (+1.78%)
As of 3:25PM EDT. Market open.

Summary | Conversations | Statistics | Profile | Financials | Options | Holders | Historical Data | Analysts

Time Period: Aug 22, 2016 - Aug 22, 2017 Show: Historical Prices Frequency: Daily Apply

Currency in USD

Download Data

Date	Open	High	Low	Close*	Adj Close**	Volume
Aug 22, 2017	912.72	924.74	911.48	922.81	922.81	838,977
Aug 21, 2017	910.00	913.00	903.40	906.66	906.66	934,900

Option data

- For one stock, there are always many options (call, put) with different strike price and expiration date.
- Historical option data are not available in Hanlon Lab due to its volume and complexity
- Financial websites and Bloomberg provide option data for current day.

Stock:	GE	Price	34.80			Expires:	1 month
	Call Option				Put Option		
Intrinsic Value	Time Value	Option Price	Strike Price	Intrinsic Value	Time Value	Option Price	
4.80	0.20	5.00	30.00	0.00	0.01	0.01	
0.00	0.53	0.53	35.00	0.20	0.40	0.60	
0.00	0.01	0.01	40.00	5.20	0.80	6.00	
		= In the Money					

Stock:	GE	Price	34.80			Expires:	9 months
	Call Option				Put Option		
Intrinsic Value	Time Value	Option Price	Strike Price	Intrinsic Value	Time Value	Option Price	
4.80	2.05	6.85	30.00	0.00	1.17	1.17	
0.00	3.70	3.70	35.00	0.20	2.62	2.82	
0.00	1.75	1.75	40.00	5.20	0.35	5.55	
		= In the Money					

Financial statements

- Public companies' data are available on SEC.gov. Normally in html or txt format.
- Accounting numbers can be found on bloomberg and WRDS
- Some commercial database have private companies' data.

Example Company
Balance Sheet
December 31, 2016

ASSETS		LIABILITIES	
Current assets		Current liabilities	
Cash	$ 2,100	Notes payable	$ 5,000
Petty cash	100	Accounts payable	35,900
Temporary investments	10,000	Wages payable	8,500
Accounts receivable - net	40,500	Interest payable	2,900
Inventory	31,000	Taxes payable	6,100
Supplies	3,800	Warranty liability	1,100
Prepaid insurance	1,500	Unearned revenues	1,500
Total current assets	89,000	Total current liabilities	61,000
Investments	36,000	Long-term liabilities	
		Notes payable	20,000
Property, plant & equipment		Bonds payable	400,000
Land	5,500	Total long-term liabilities	420,000
Land improvements	6,500		
Buildings	180,000		
Equipment	201,000	Total liabilities	481,000
Less: accum depreciation	(56,000)		
Prop, plant & equip - net	337,000		
Intangible assets		STOCKHOLDERS' EQUITY	
Goodwill	105,000	Common stock	110,000
Trade names	200,000	Retained earnings	220,000
Total intangible assets	305,000	Accum other comprehensive income	9,000
		Less: Treasury stock	(50,000)
Other assets	3,000	Total stockholders' equity	289,000
Total assets	$ 770,000	Total liabilities & stockholders' equity	$ 770,000

Introduction to Financial data

Quantitative Data

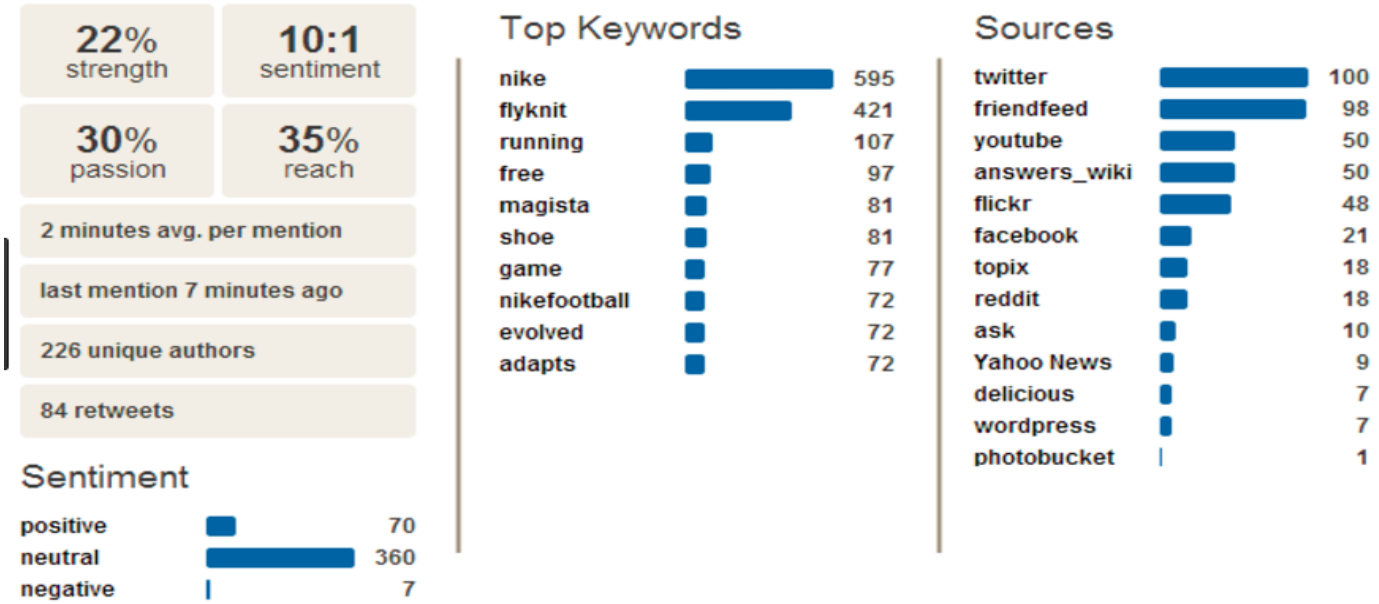
- Structured, numeric data, such as stock return, ROA ratio, option price, etc.
- Analytical models can be easily applied on quantitative data.

Qualitative Data

- Unstructured, textual data, such as earnings announcement, news, etc.
- Categorical data can be easily transferred into numbers.
- More complex qualitative data needs additional procedure.

News

- News are always public online which requires technique skills to download and clean. Some news website may provide API to download data.
- Thomson Reuters News Analytics(TRNA) database provides sentiment, highlight title and content.



Resources in Hanlon Lab

Databases

- Bloomberg: real-time tick data of financial markets
- Thomson Reuters Tick History (TRTH) : historical data of financial markets
- Thomson Reuters News Analytics (TRNA): PostgreSQL database of finanical news sentiment from Thomson Reuters (score calculated by Thomson Reuters)
- Wharton Research Data Services (WRDS): access to COMPUSTAT,

Detailed information

`https://fsc.stevens.edu/hfsc/?q=hfsl-resources/hfsc-databases`

Lab account request

`https://web.stevens.edu/hfslwiki/images/d/dd/HFSL_AccessRequestForm_Apr2016.pdf`

Why R

- R and Python can handle most requests. They are good for statistical analysis, data/text mining analysis.
- C++ and JAVA are for more advanced use (e.g. option pricing algorithm, complexity optimization).
- MATLAB can be used for data analysis (handles matrix very well), and optimization.
- VBA is used by some trading support team. It can build plug-ins for Excel which are easy for everyone.
- SAS is used to handle large size of data and complicated analysis, especially in traditional companies.
- Other big data techniques (e.g. HADOOP, SCALA) become popular especially in technique companies.

* Compared with SAS and Matlab, R is free and easy to install & learn;
* Compared with SPSS, R is more flexible;
* Python is very similar to R.

Introduction to R

- Creating, listing and deleting the objects in memory
- Data types
  - There are four main data types: numeric, character, complex7, and logical. Other modes exist but they do not represent data, for instance function or expression.
  - The length is the number of elements of the object.
- Data type conversion

Conversion to	Function	Rules
numeric	`as.numeric`	`FALSE` → 0
		`TRUE` → 1
		`"1"`, `"2"`, ... → 1, 2, ...
		`"A"`, ... → `NA`
logical	`as.logical`	0 → `FALSE`
		other numbers → `TRUE`
		`"FALSE"`, `"F"` → `FALSE`
		`"TRUE"`, `"T"` → `TRUE`
		other characters → `NA`
character	`as.character`	1, 2, ... → `"1"`, `"2"`, ...
		`FALSE` → `"FALSE"`
		`TRUE` → `"TRUE"`

Object Types

Vector	1 dimension	All elements have the same data types Data types: numeric, character logic, factor
Matrix	2 dimensions	
Array	2 or more dimensions	
Data frame	2 dimensions	table-like data object allowing different data types for different columns
List	Collection of data objects, each element of a list is a data object