ITM 6285 Data Mining

Instructor:
Peng Xie

Department of Management

Agenda

- Self Introduction
- Syllabus
- Assignment and lab session
- Overview of analytics

About Me

- 2012: Bachelor and Master in Finance
  - Shanghai Jiaotong University
- 2017: PhD, Gerogia Tech
  - Business administration and Management
  - ITM group
- Currently
  - Assistant Professor at Dept. Management
- Research
  - Social Media Analytics, Digital Currency, Text Mining

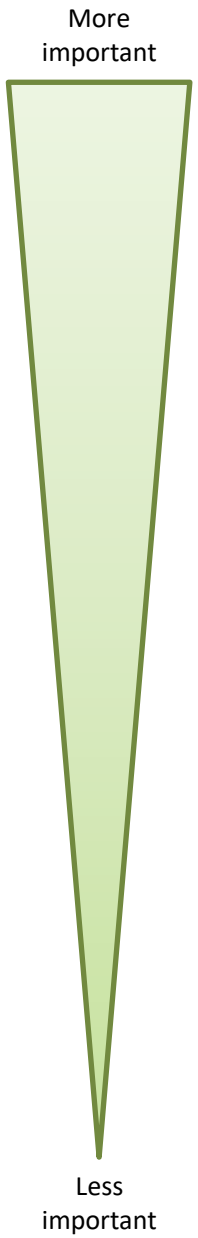
Who Should Take This Course?

- Pre-requisites:
  - Calculus, Probabiltiy Theory, Statistics, Linear Algebra
  - Excel
  - like working with data

- Not for Advanced Data Analysts:
  - ***Not*** a math course
  - ***Not*** a algorithm design course
  - ***Not*** a programming course

What Should We Care

- Purpose of data mining techniques
  - the problems it can solve
- How to carry out the analysis in R
- Results interpretation
  - what you learn from the result outputs
- The details of the algorithms
- Programming



Course Information

- Important Dates
  - Oct 23: midterm
  - Change of the schedule, please download the newest syllabus

About Lab Session

- Lab Session
  - Get your hands dirty
  - You will not learn data mining just by listening to lectures
  - Bring your Laptop or Mac ***Every Class***
  - Do some real data analyzing **together**
  - I will provide step-by-step guide
  - Open everything
  - Feel free to ask questions
  - And communicate with your classmates
  - Submissions will be graded (class participation included)

Course Materials

- Primary learning reference
  - Slides and code snippets
  - Notes on the codes

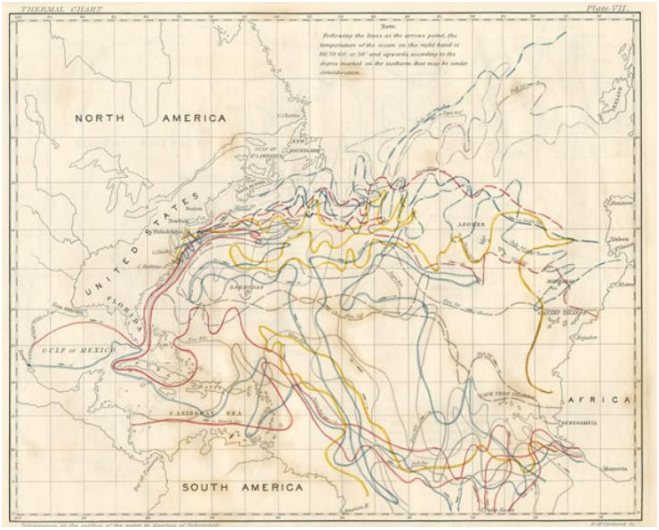
- Recommended reference book
  - Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition
  - There is 2nd uploaded on Blackboard
  - However the book does not cover regression analysis

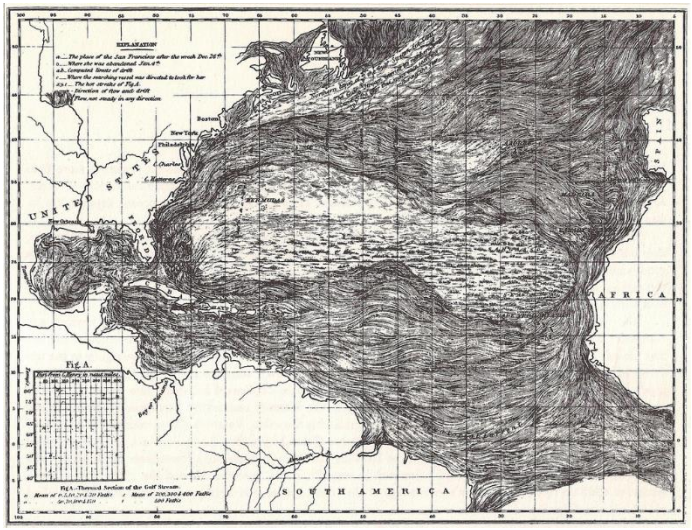
Grading

Criterion	Weight	Comment
Lab Session	20%	The students shall follow the lab sessions instructions to complete the data analysis tasks. The instructor is available to answer any question the students may have during the lab session. At the end of the class, the students shall submit their works to Blackboard for evaluation. Screenshots are needed to show the computed results. The 20% should be ***free points*** if the students show up.
Homework	20%	There is no late make-up homework.
Midterm exam	30%	There are no late make-up midterms. Early midterm exams, if accepted, must be for reasons beyond the student's control.
Comprehensive final exam	30%	There are no early or make-up final exams.
Total	100%	

Where is the value of the data

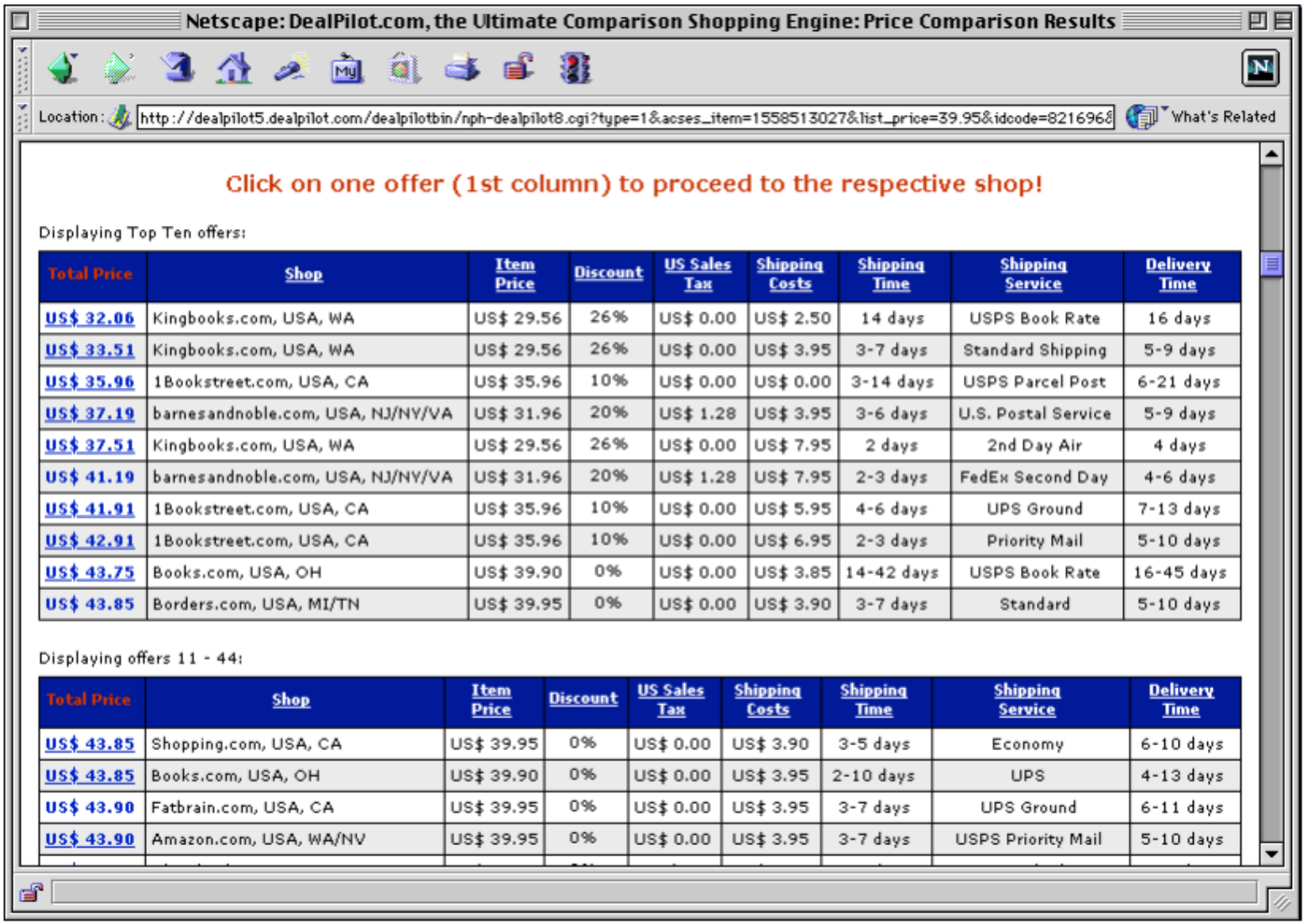
- Matthew Fontaine Maury Navigation Chart: 1842-1861
  - Collect data from captains
  - Create six charts: Track, wind, pilot, thermal, storm and rain, whale
  - And then sell the charts, or give away charts in exchange of new data





Where is the value of the data

- Example of book pricing


Netscape: DealPilot.com, the Ultimate Comparison Shopping Engine: Price Comparison Results

Location: http://dealpilot5.dealpilot.com/dealpilotbin/nph-dealpilot8.cgi?type=1&acses_item=1558513027&list_price=39.95&idcode=821696& What's Related

Click on one offer (1st column) to proceed to the respective shop!

Displaying Top Ten offers:

Total Price	Shop	Item Price	Discount	US Sales Tax	Shipping Costs	Shipping Time	Shipping Service	Delivery Time
US$ 32.06	Kingbooks.com, USA, WA	US$ 29.56	26%	US$ 0.00	US$ 2.50	14 days	USPS Book Rate	16 days
US$ 33.51	Kingbooks.com, USA, WA	US$ 29.56	26%	US$ 0.00	US$ 3.95	3-7 days	Standard Shipping	5-9 days
US$ 35.96	1Bookstreet.com, USA, CA	US$ 35.96	10%	US$ 0.00	US$ 0.00	3-14 days	USPS Parcel Post	6-21 days
US$ 37.19	barnesandnoble.com, USA, NJ/NY/VA	US$ 31.96	20%	US$ 1.28	US$ 3.95	3-6 days	U.S. Postal Service	5-9 days
US$ 37.51	Kingbooks.com, USA, WA	US$ 29.56	26%	US$ 0.00	US$ 7.95	2 days	2nd Day Air	4 days
US$ 41.19	barnesandnoble.com, USA, NJ/NY/VA	US$ 31.96	20%	US$ 1.28	US$ 7.95	2-3 days	FedEx Second Day	4-6 days
US$ 41.91	1Bookstreet.com, USA, CA	US$ 35.96	10%	US$ 0.00	US$ 5.95	4-6 days	UPS Ground	7-13 days
US$ 42.91	1Bookstreet.com, USA, CA	US$ 35.96	10%	US$ 0.00	US$ 6.95	2-3 days	Priority Mail	5-10 days
US$ 43.75	Books.com, USA, OH	US$ 39.90	0%	US$ 0.00	US$ 3.85	14-42 days	USPS Book Rate	16-45 days
US$ 43.85	Borders.com, USA, MI/TN	US$ 39.95	0%	US$ 0.00	US$ 3.90	3-7 days	Standard	5-10 days

Displaying offers 11 - 44:

Total Price	Shop	Item Price	Discount	US Sales Tax	Shipping Costs	Shipping Time	Shipping Service	Delivery Time
US$ 43.85	Shopping.com, USA, CA	US$ 39.95	0%	US$ 0.00	US$ 3.90	3-5 days	Economy	6-10 days
US$ 43.85	Books.com, USA, OH	US$ 39.90	0%	US$ 0.00	US$ 3.95	2-10 days	UPS	4-13 days
US$ 43.90	Fatbrain.com, USA, CA	US$ 39.95	0%	US$ 0.00	US$ 3.95	3-7 days	UPS Ground	6-11 days
US$ 43.90	Amazon.com, USA, WA/NV	US$ 39.95	0%	US$ 0.00	US$ 3.95	3-7 days	USPS Priority Mail	5-10 days

Where is the value of the data

- Modeling result

Variable	Coefficient
Price	-0.194 (0.001)
Shipping cost	-0.368 (0.002)
Sales tax	-0.432 (0.014)
Delivery time	-0.019 (0.001)
Sold at Amazon	0.477 (0.020)
Sold at BN	0.177(0.023)
Sold at Border	0.266 (0.02)

Smith, M., E. Brynjolfsson. 2001. Customer Decision Making at an Internet Shopbot: Brand Still Matters. The Journal of Industrial Economics, 49(4) 541-558.

Where is the value of the data

- Modeling result
  - Consumers are sensitive to shipping and tax.
  - Free shipping everywhere

Where is the value of the data

- Modeling result
  - Different level of brand recognition
  - Consumers are willing to pay higher premium for Amazon books

Sold at Amazon	**0.477 (0.020)**
Sold at BN	0.177(0.023)
Sold at Border	0.266 (0.02)

The Iceberg of Data

- Without further analysis, we only see a tip
  - The average production level
  - The past trend
  - A few pie charts and bar charts
  - Earnings per share growth
  - ……

The Iceberg of Data

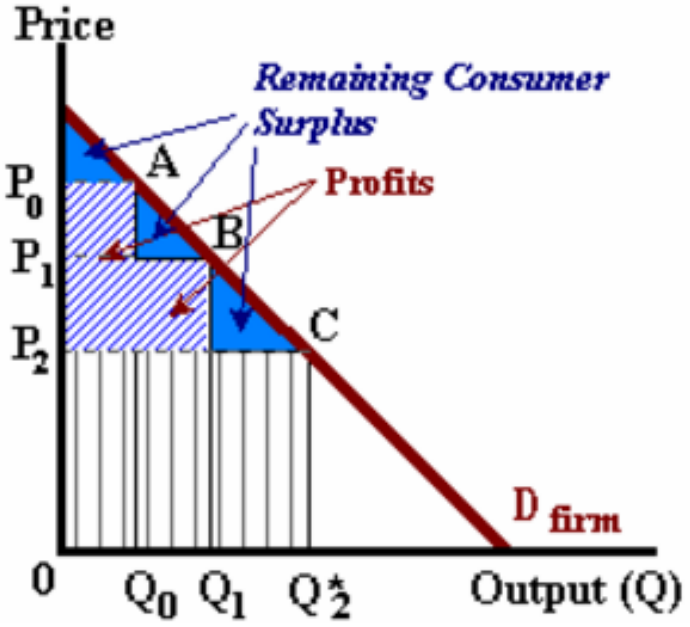
- With further analysis
  - How and why
  - Maybe the causal link
  - Counter intuitive results caused by a hidden mechanism.

- A research from my advisor in GT
  - Sending catalogs to customers. Does it worth the cost?
  - Intuition: Yes, it is useful.
  - Sometimes the data speaks louder than the opinions.

Data-Driven Personalized Pricing?

- Retailers are watching your every move
  - Amazon did this before
  - After logging out, the price changed
  - Information asymmetry: uniform price
  - Retailer knows your preference, gender, purchasing frequency, return frequency etc..

You buy a pair of shoes from a popular online retailer. Across town, another shopper buys the exact same pair of shoes from the same retailer. Except he paid $10 less than you did. Not because he used a coupon or discount code, but because the store decided to offer him a lower price than you.



The Big Data Era

- Data-fy everything
  - Matthew Fontaine Maury Navigation Charts
  - Social media discussions (Facebook, Youtyube……)
  - Location change (Google Timeline, check out my timeline)
  - And everthing

The Big Data Charateristics

- Volume
- Velocity
- Variety

The Big Data Characteristics

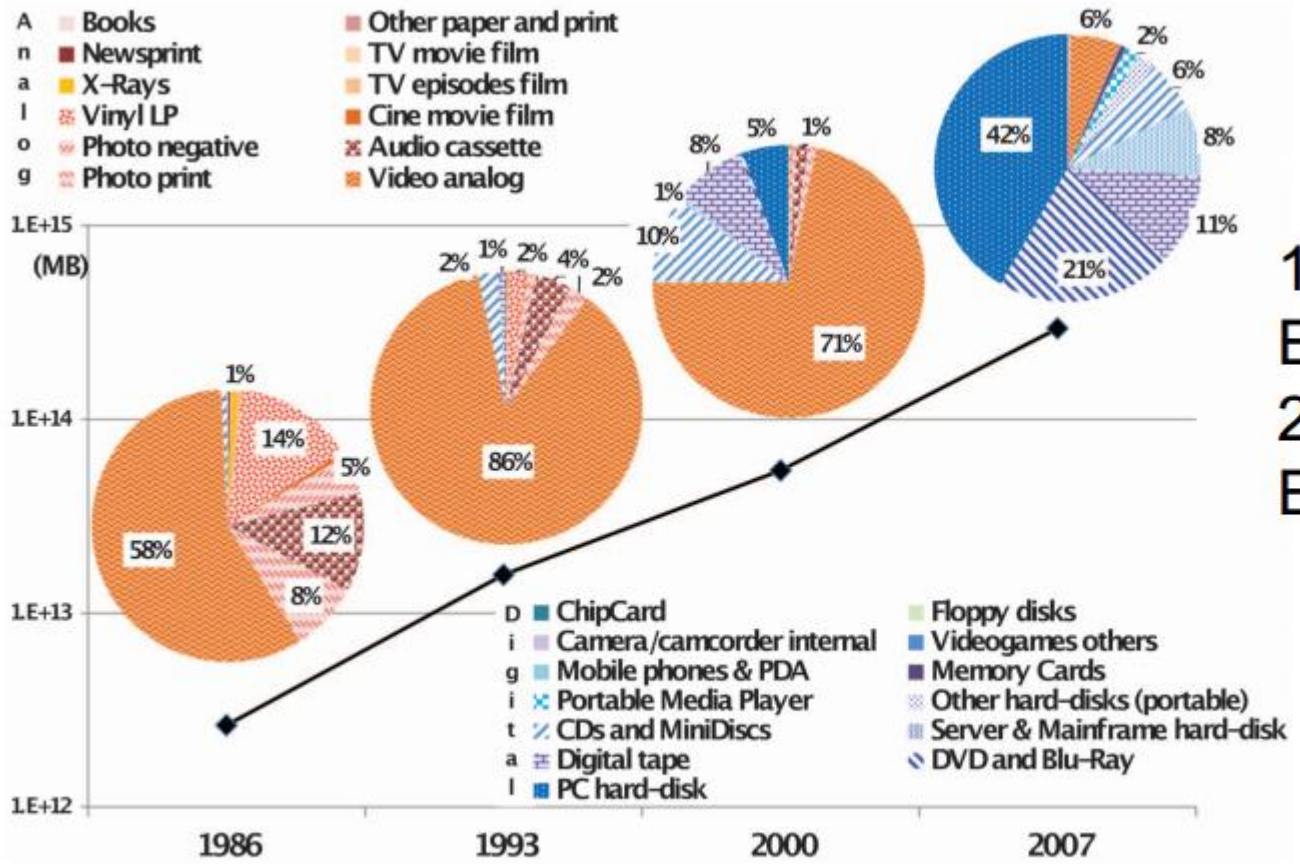
- Volume, Velocity, Variety
  - Mega:10^6
  - Giga:10^9
  - Tera:10^{12}
  - Peta:10^{15}
  - Exa:10^{18}
  - Zetta:10^{21}

2017: Seagate 60 TB SSD drive

2000: Seagate 18 GB Cheetah X15

The Big Data Characteristics

- **Volume**, Velocity, Variety

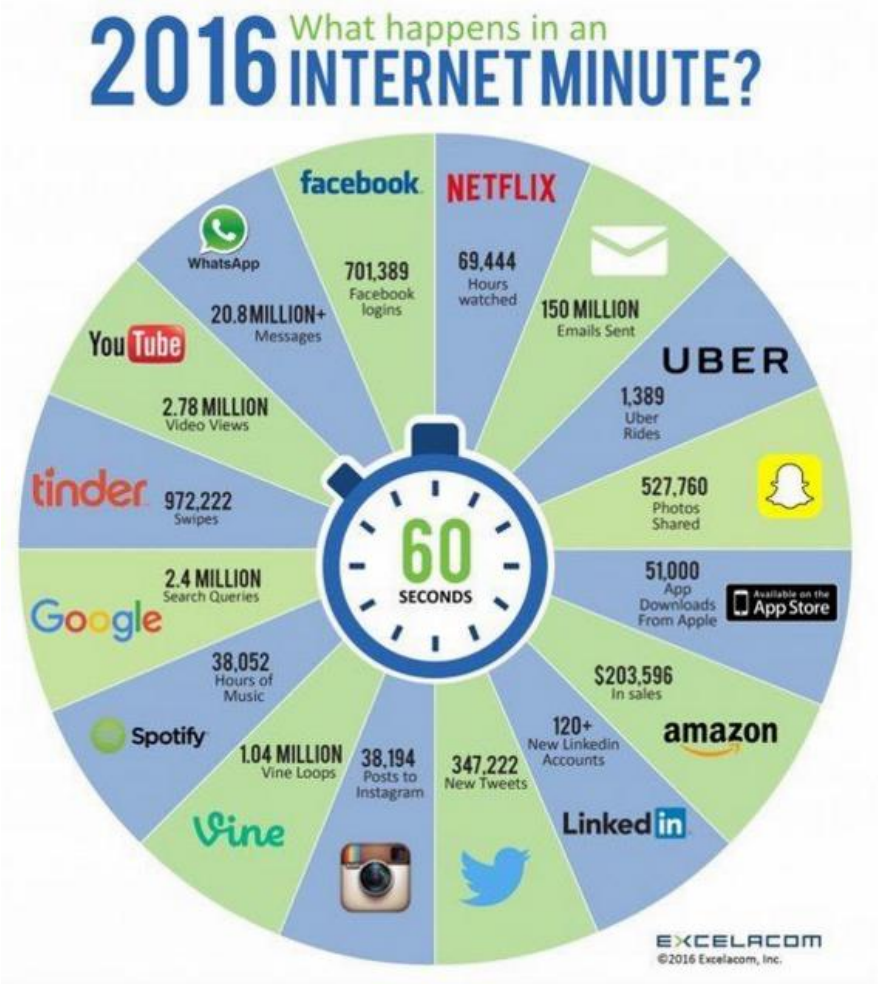


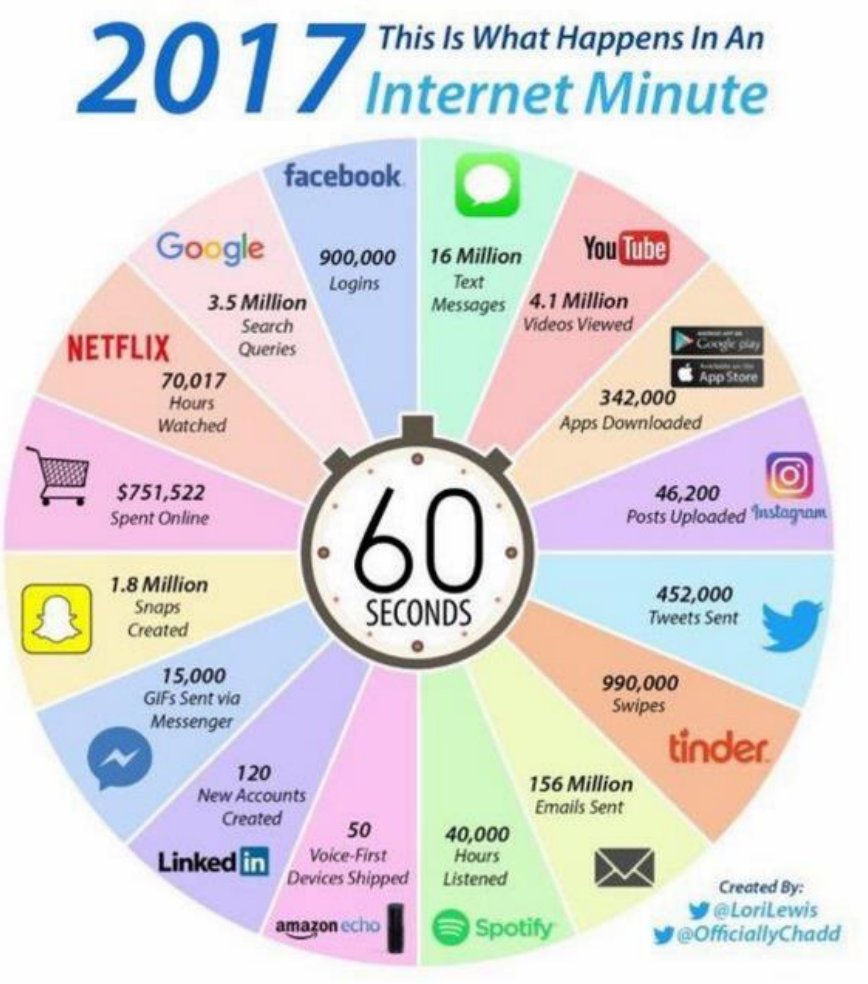
1986: 2.6 Exabyte
2007: 295 Exabyte

Fig. 2. World's technological installed capacity to store information (table SA1) (*16*).

The Big Data Characteristics

- Volume, Velocity, Variety





The Big Data Characteristics

- Volume, Velocity, **Variety**
  - Old: Structured data
  - Numbers, categorical variables (gender……)

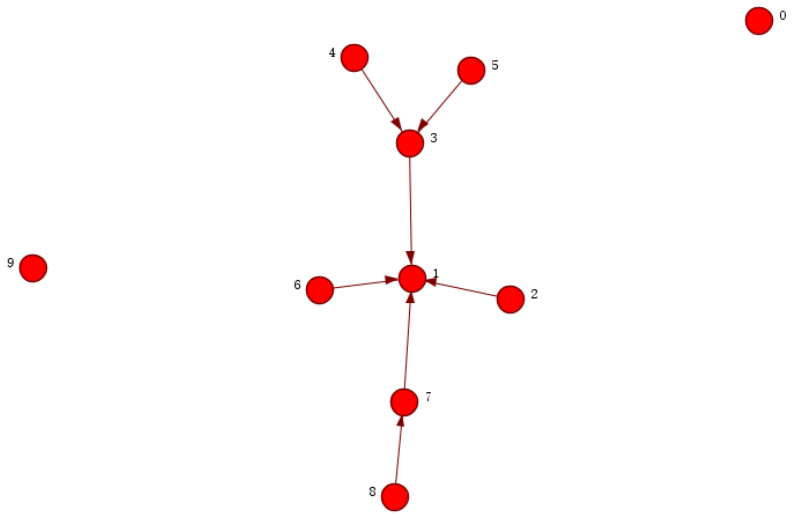
Example of Data Variety

I was looking forward to my trip to Atlanta. On my way to Atlanta, I discovered that I would be arriving earlier than expected so I called to see if I could arrive before check-in. The woman who answered the phone **rudely** cut me off before I finished speaking to tell me that it would not be possible for me to arrive early, so I waited until the appropriate time to check-in. Upon arriving at the hotel, the front desk agent made an **unnecessary** and **insulting** comment that I looked too young to check-in alone.

Sentiment: the percentage of words expressing negative feelings

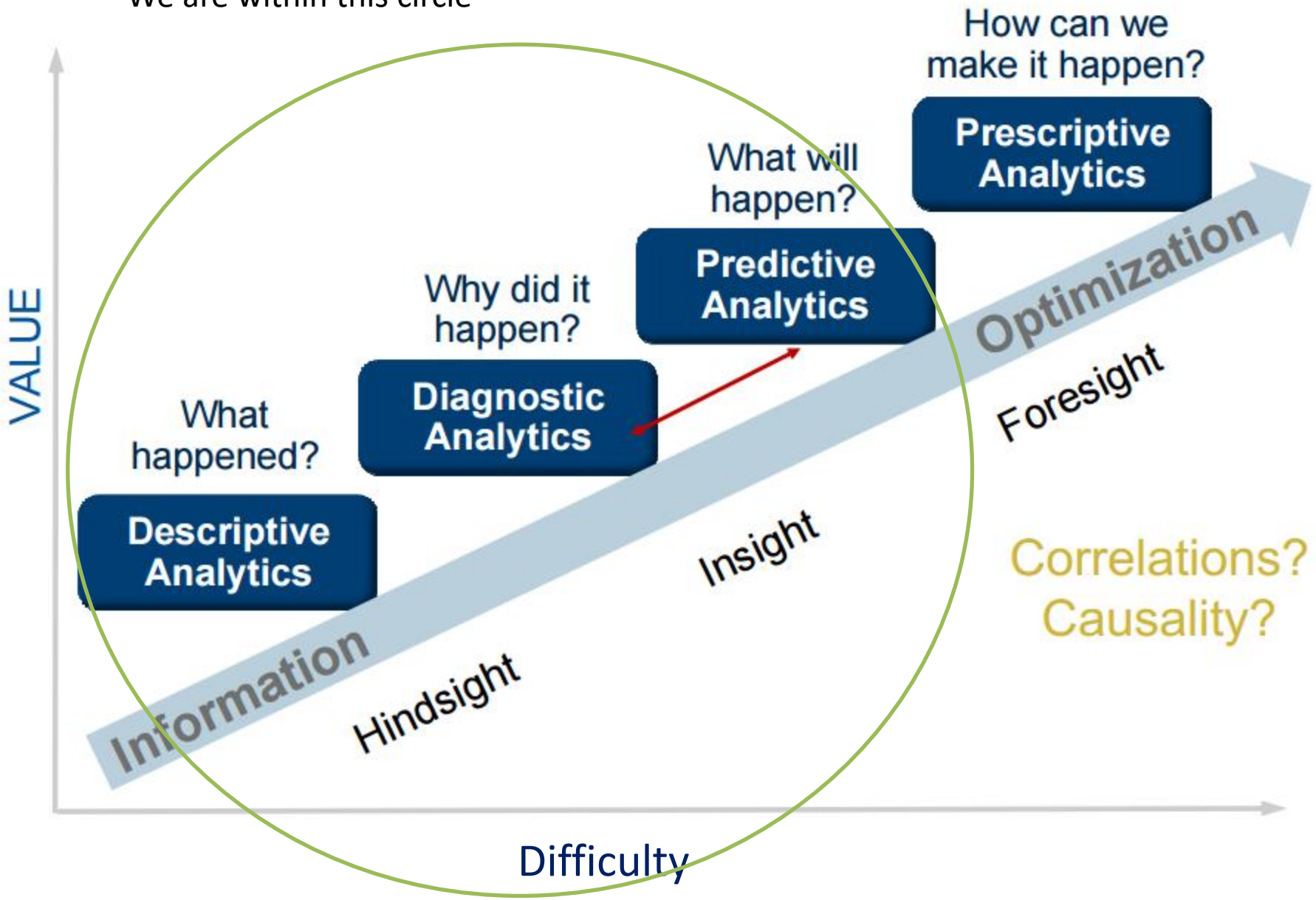
Example of Data Variety

- Relationship between email network structure and salary
  - Annual salary positively related to network in-degree centrality and betweeness
  - 1 has high in-degree centrality and 3 has high betweeness centrality.



Different Types of Analysis

We are within this circle



Different Types of Analysis

- The highest place in the ladder is causal analysis
  - What causes the increased number crimes
  - You did predictive analysis and find out the number of polices is highly related to the number of crimes
  - Hiring more police increases crime??
  - And the “go to doctor example”.
  - The dif-in-dif method is used to establish causal relationship.