Statistical Methods in Business

Lecture 1: Introduction & Applications

Yuan Qu
Rutgers University 09/05/2019

Course Info.

RUTGERS
Rutgers Business School
Newark and New Brunswick

- Session: 33:136:185:09.
- Class: Thursday, 6:40 pm -- 9:40 pm.
- Classroom: BRR, Room 5073.
- Instructor: Yuan Qu (or, for convenience, Ryan Q)
- E-mail: yuan.qu@rutgers.edu
- The 26th of November 2019 follows a Thursday Schedule.
- Final Exam: 12/19/2019: 8:00 PM - 11:00 PM (???)

Survey

RUTGERS
Rutgers Business School
Newark and New Brunswick

- I will upload all documents ASAP when I can access the blackboard.
- Including a survey…
  - Course starting time
  - Materials that you have learned (especially for statistic)
  - Materials that you want to learn (practical exercise, theory, pure math)
  - Programming languages that you want to learn (R, Python…)
  - About grade weights

What we will cover

RUTGERS
Rutgers Business School
Newark and New Brunswick

- Math
- Model
- Application
- Excel exercise
- Maybe R and Python exercise
- Topic related to Business Analysis

Content

RUTGERS
Rutgers Business School
Newark and New Brunswick

- Analysis of Variance (Homework 1)
- Linear Regression & Analysis (Homework 2)
- Linear Regression Model Design
- **Mid-term Exam**
- Time Series
- Chi-square Test (Homework 3)
- Business Analysis & Data Mining
- **Final Exam (Cumulative)**

Example: New Drink

RUTGERS
Rutgers Business School
Newark and New Brunswick

Suppose your company want to develop a new drink, with 4 different colors, black, orange, pink and green.

Here are sales data from 3 different grocery stores.

Store	Black	Orange	Pink	Green
1	26.5	27.9	31.2	30.8
2	28.7	25.1	28.3	29.6
3	25.1	28.5	30.8	32.4

- Question: **Did color affect sale?**

Analysis of Variance

RUTGERS
Rutgers Business School
Newark and New Brunswick

Store	Black	Orange	Pink	Green
1	26.5	27.9	31.2	30.8
2	28.7	25.1	28.3	29.6
3	25.1	28.5	30.8	32.4
Mean	26.77	27.17	30.10	30.93
Variance	3.29	3.29	2.47	1.97
Total Mean	28.74	Total Variance	5.56	

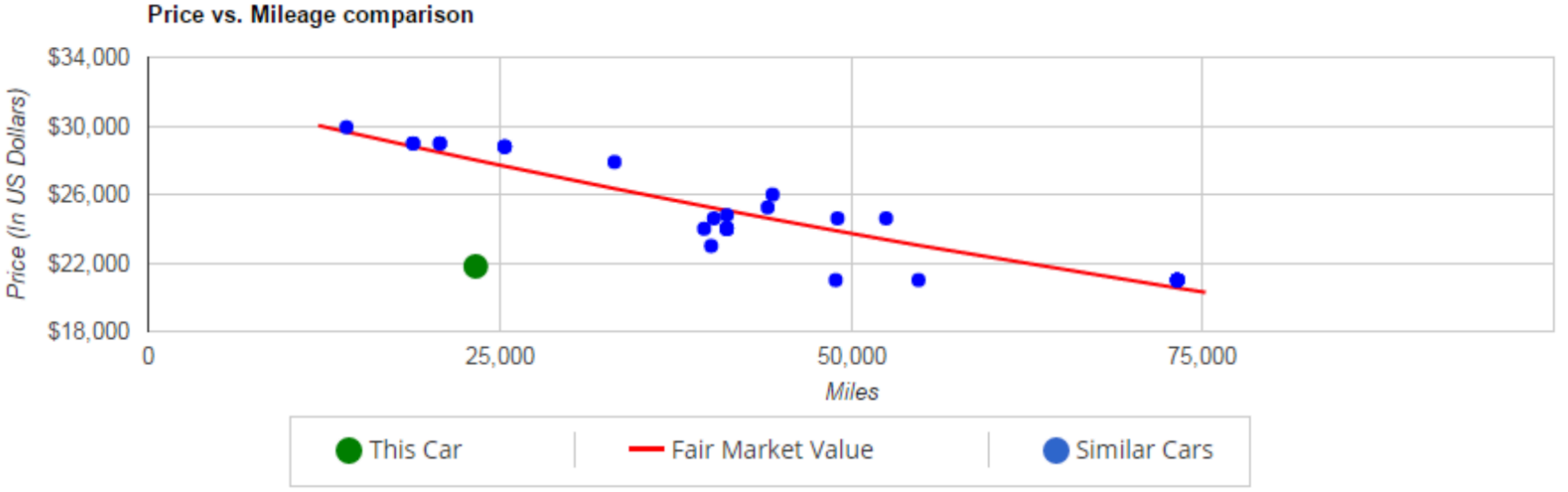
Question: **Did color affect sale?**

Example: Buy a used car

- www.iseecars.com
- Now you see the following model:
  - 2013 BMW 3 Series 328i xDrive AWD 4dr Sedan
  - 23,302 miles
  - Price: $21,795
- How does Mileage affect Price?

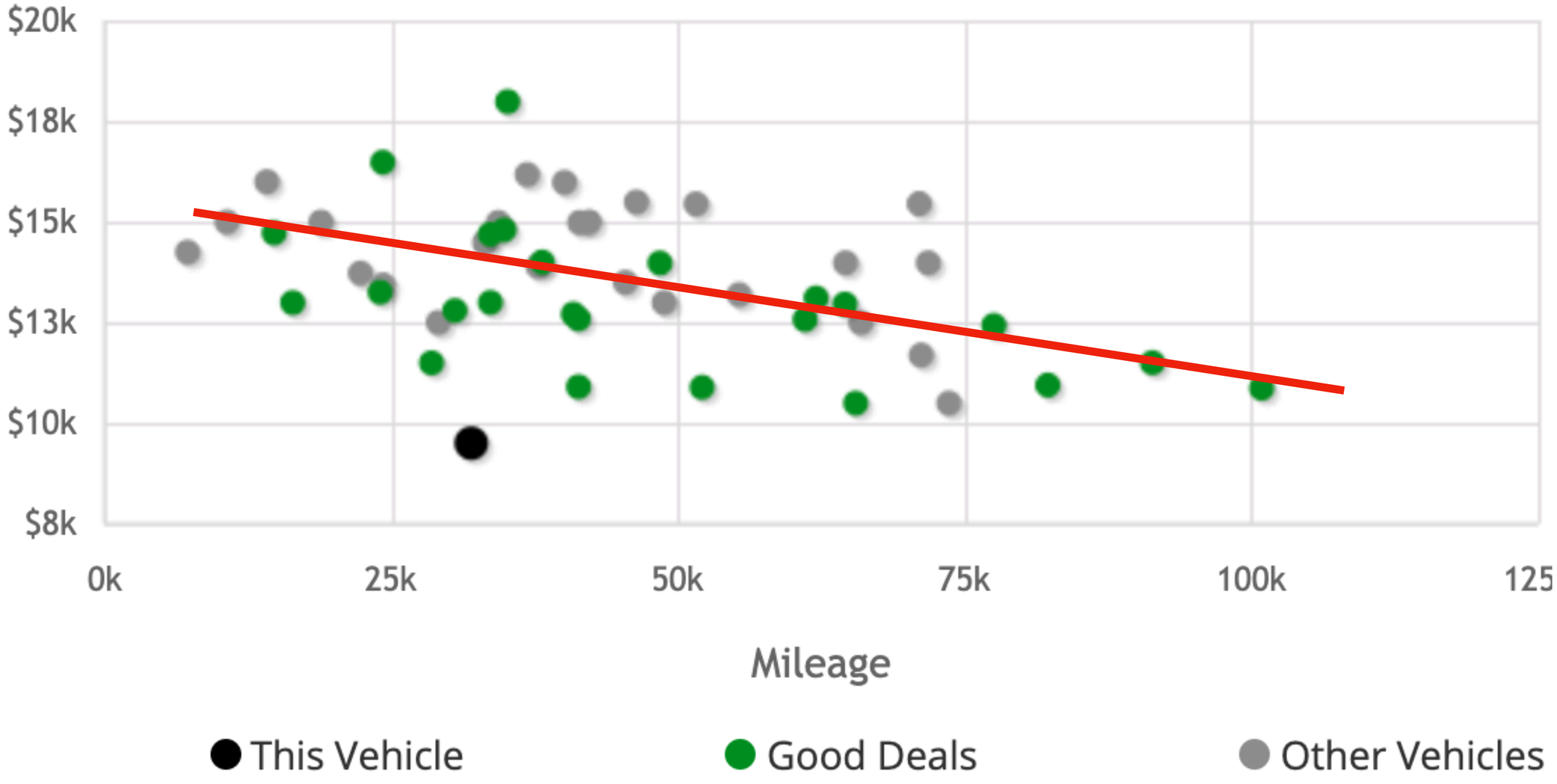
Price

Price:	$21,795
Estimated market value:	$27,801 The current price is $6,006 below market value.



- **The green dot on the chart is this car.**
- **The blue dots are similar cars for sale.**
- **The red line is the market value for vehicles like this one.**

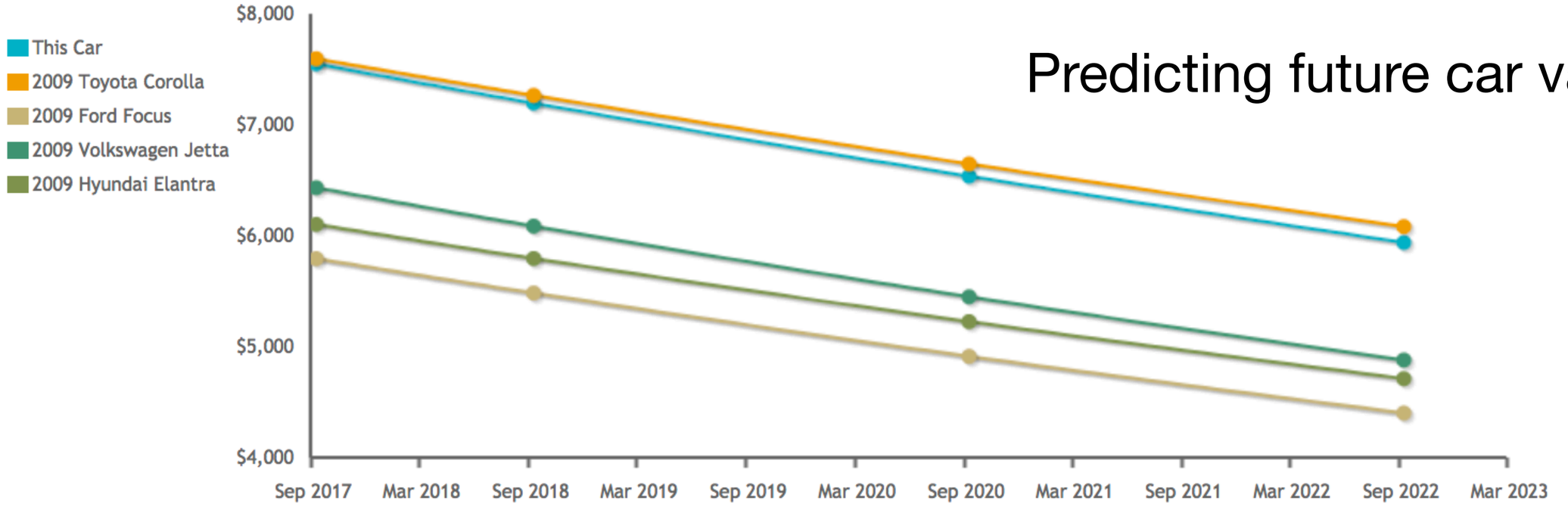
2015 Honda Civic



What is Regression?

- Regression analysis allows scientists to quantify how the average of one variable systematically varies according to the levels of another variable.
- The former variable is often called a **dependent variable** or outcome variable and the latter an **independent variable**, predictor variable, or explanatory variable.
- If we can fit a line to these data points (a.k.a regression line), then we could predict the future car value.

	Current Value	1 Year from Now	3 Years from Now	5 Years from Now
This Car	$7,548	$7,194 (-$354)	$6,536 (-$1,012)	$5,938 (-$1,610)
Typical Similar Cars:				
2009 Toyota Corolla	$7,595	$7,265 (-$330)	$6,647 (-$948)	$6,082 (-$1,513)
Competitor Cars:				
2009 Ford Focus	$5,793	$5,483 (-$310)	$4,912 (-$881)	$4,401 (-$1,392)
2009 Volkswagen Jetta	$6,432	$6,086 (-$346)	$5,449 (-$983)	$4,879 (-$1,553)
2009 Hyundai Elantra	$6,102	$5,794 (-$308)	$5,225 (-$877)	$4,711 (-$1,391)



Predicting future car value

Example: Titanic

RUTGERS
Rutgers Business School
Newark and New Brunswick

- The ship Titanic sank in 1912 with the loss of most of its passengers
- 809 of the 1,309 passengers and crew died
- Death rate = 61.8%

- Question: **Did class (of travel) affect survival?**

Hypothesis Test

RUTGERS
Rutgers Business School
Newark and New Brunswick

- If we assume:
  - **Null**: There is NO association between class and survival
  - **Alternative**: There IS an association between class and survival
  - Overall Death rate = 61.8%

Count		Survived?		Total
		Died	Survived	
Class	1st	123	200	323
	2nd	158	119	277
	3rd	528	181	709
Total		809	500	1,309

If Null is true

RUTGERS
Rutgers Business School
Newark and New Brunswick

- **Null**: There is NO association between class and survival

Same proportion of people would have died in each class!

Count		Survived?		Total
		Died	Survived	
Class	1st	123	200	323
	2nd	158	119	277
	3rd	528	181	709
Total		809	500	1,309

Class	Total	DiedRate	Exp.Rate
1st	323	38.1%	61.8%
2nd	277	57.0%	61.8%
3rd	709	74.5%	61.8%
Total	1,309	61.8%	61.8%

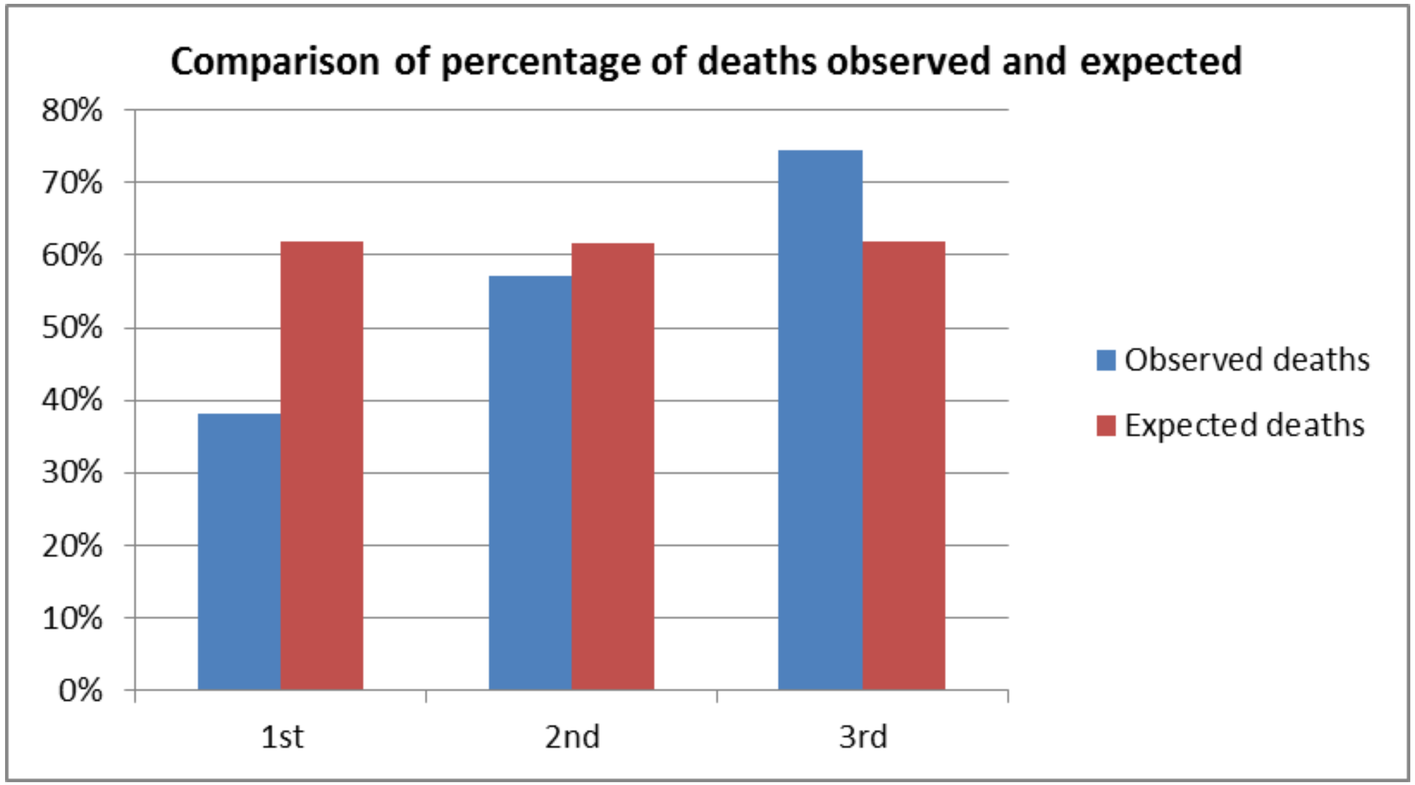
If Null is true

RUTGERS
Rutgers Business School
Newark and New Brunswick

- **Null**: There is NO association between class and survival

  Same proportion of people

  would have died in each class!



Chi-square Test

RUTGERS
Rutgers Business School
Newark and New Brunswick

- The chi-square test is used when we want to see if two categorical variables are related
- The test statistic for the Chi-square test uses the sum of the squared differences between each pair of observed (O) and expected values (E)

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Example: Simpson's paradox

- UC Berkeley gender bias

Men		Women	
Applicants	Admitted	Applicants	Admitted
8442	44%	4321	35%

- The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance

Actually…

RUTGERS
Rutgers Business School
Newark and New Brunswick

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	**825**	62%	108	**82%**
B	**560**	63%	25	**68%**
C	325	**37%**	**593**	34%
D	417	33%	375	**35%**
E	191	**28%**	**393**	24%
F	373	6%	341	**7%**

- In fact, the pooled and corrected data showed a "small but statistically significant bias in favor of women".
- the top two departments by number of applicants for each gender bolded.

Example: Base rate fallacy

- The ridiculous question:
- Ariana loves music, she plays piano at home everyday
- So, what's Amy's job?
  - A. Musician
  - B. Engineer
- If presented with related base rate information (i.e. generic, general information) and specific information (information pertaining only to a certain case), the mind tends to ignore the former and focus on the latter

Example: Disease Test

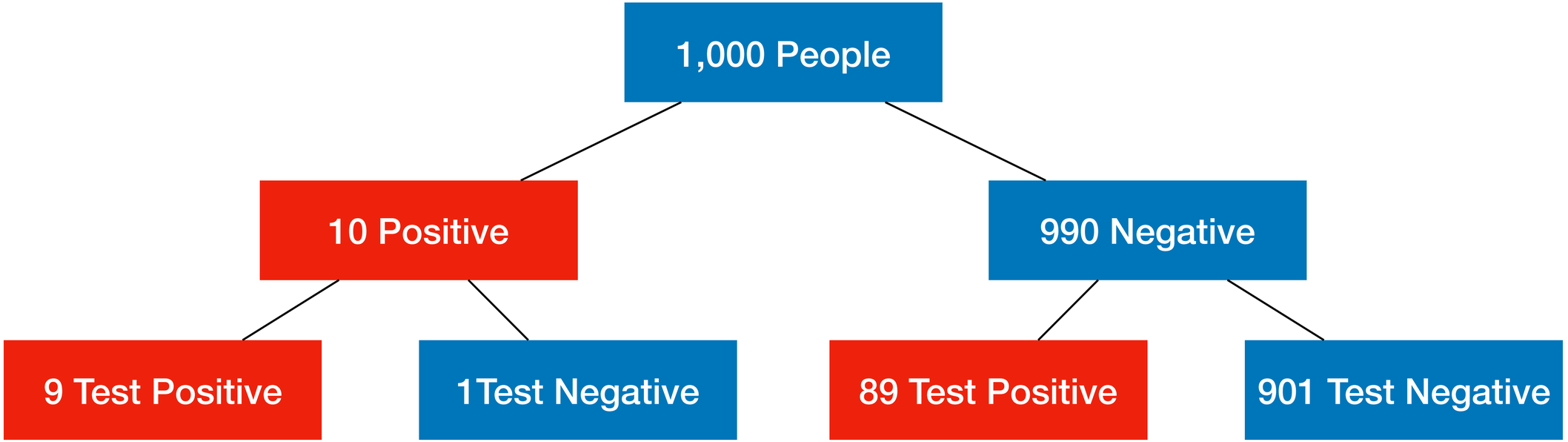
RUTGERS
Rutgers Business School
Newark and New Brunswick

- According to statistic, the probability that a person gets positive result in “MM” disease test is 1%.
- The error rate of the disease test is 9%.
- One patient, A, he got “positive” in test.
- What’s the real probability that the inspection is correct?

- 91%?

Example: Disease Test

RUTGERS
Rutgers Business School
Newark and New Brunswick



$$P = \frac{9}{9+89} = \frac{9}{98} = 9.18\,\%$$

Why Data Analytics?

RUTGERS
Rutgers Business School
Newark and New Brunswick

- Data analytics is the driving force behind the new way of doing business.
- Data analytics has enabled modern organizations to make tremendous strides in productivity.
- Data analytics has opened new market and created new product and service opportunities.

Why Data Analytics?

RUTGERS
Rutgers Business School
Newark and New Brunswick

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at grocery stores
  - Bank/Credit Card transactions
- Computers have become cheaper and more powerful

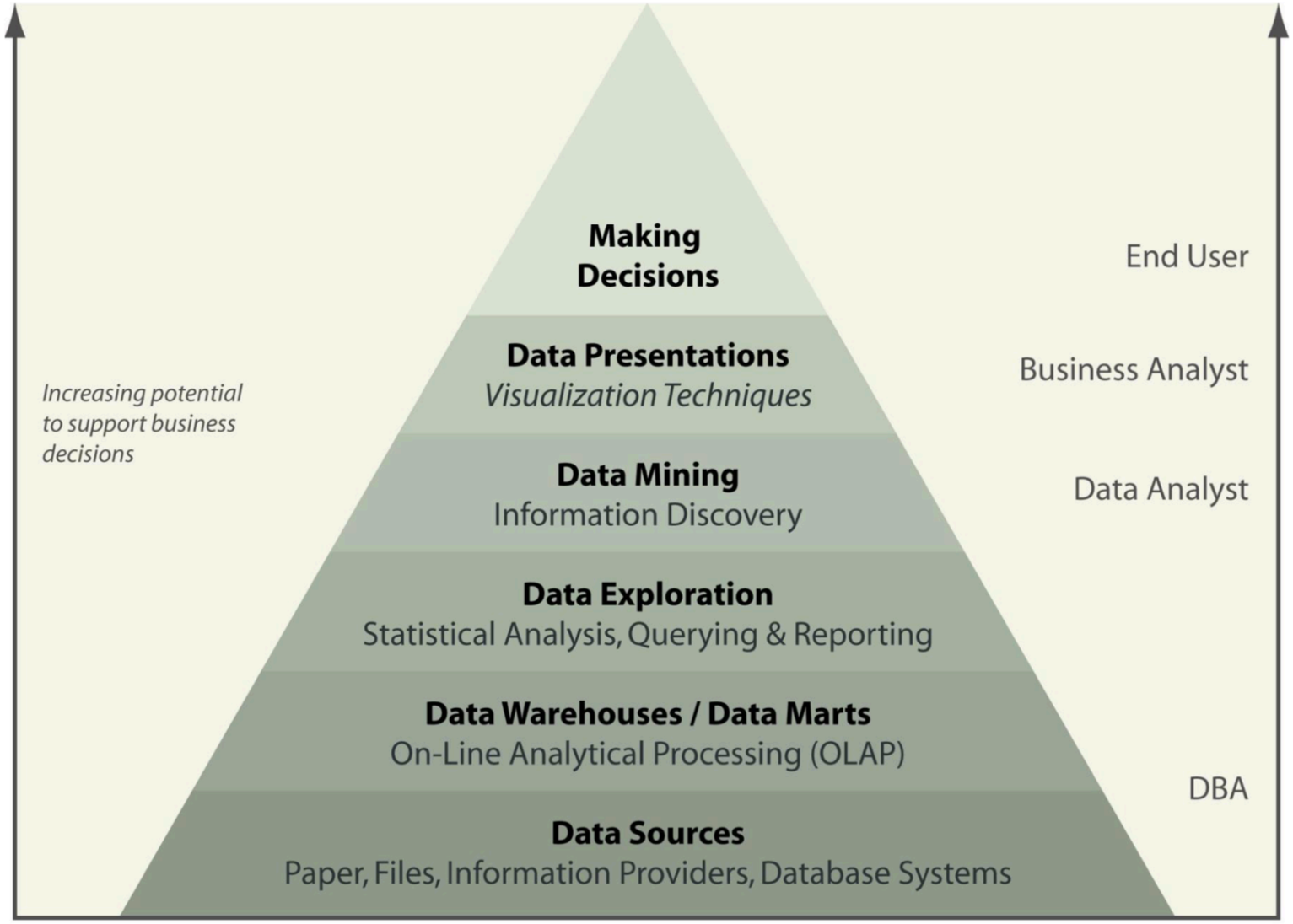
What is Data Mining?

RUTGERS
Rutgers Business School
Newark and New Brunswick

- Many definitions
- One definition is: “A non-trivial extraction of implicit, previously unknown and potentially useful information from data”
- An iterative process
  - raw data -> transformed data -> preprocessed data ->

    data mining -> post-processing -> knowledge

DM & Business Intelligence



Data Mining Tasks

RUTGERS
Rutgers Business School
Newark and New Brunswick

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.

Data Mining Tasks

RUTGERS
Rutgers Business School
Newark and New Brunswick

[Predictive]
- Classification
- Regression
- Anomaly Detection

[Descriptive]
- Clustering
- Association Rule Discovery
- Sequential Pattern Discovery