San Jose, June 2015

GitPro

Predictive Modeling in Health Care

Taposh Dutta-Roy

Introduction

Health Care industry

Analytics in Health Care

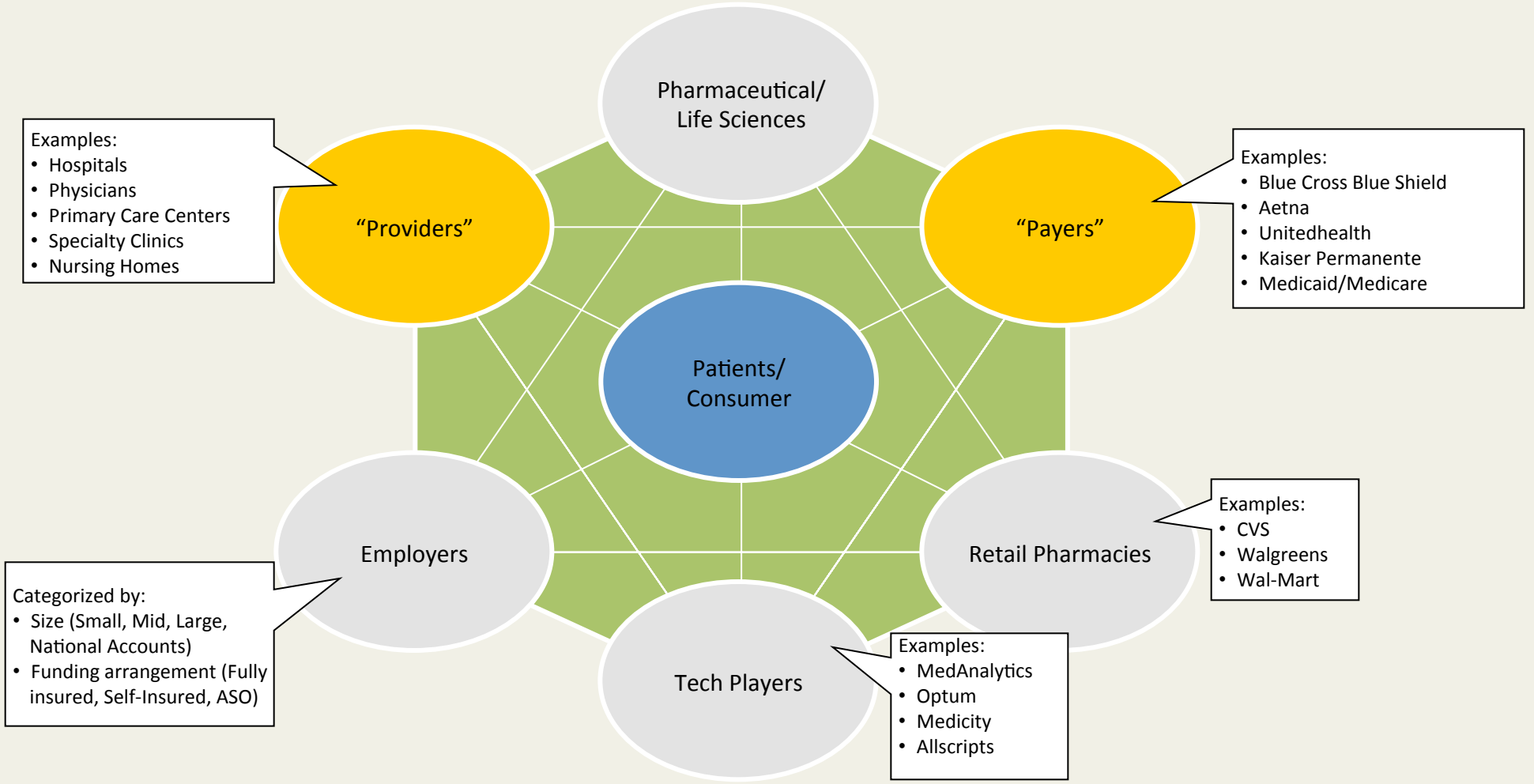
Examples of Predictive Models in Health Care

Predictive Modeling Process & Tools

Kaggle – Bike Sharing Predictive Modeling

Health Care industry

Healthcare Ecosystem



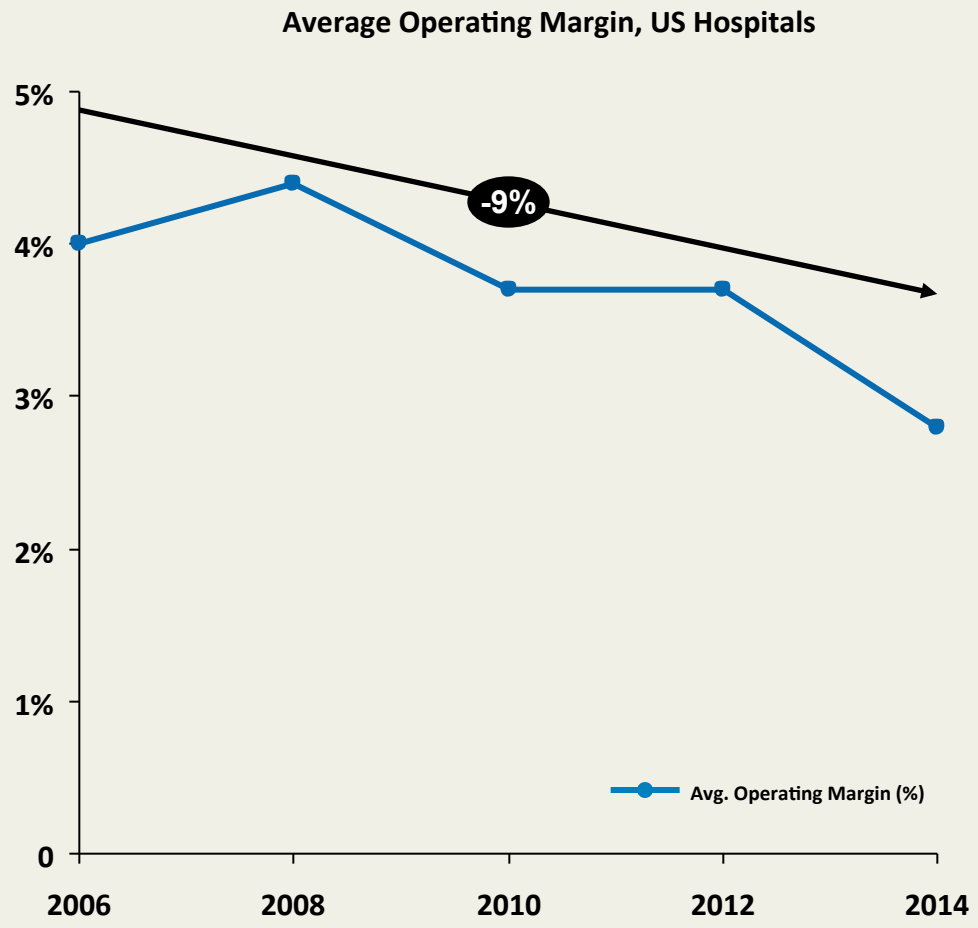
http://www.beckershospitalreview.com/hospital-management-administration/the-new-healthcare-ecosystem-5-emerging-relationships.html

“The healthcare ecosystem is made up of many players – some misaligned incentives exist across stakeholders”

Source : Booz & Company

Recent Trends in Health Care industry

U.S. hospitals face increasing financial pressures in the marketplace, threatening operating margins

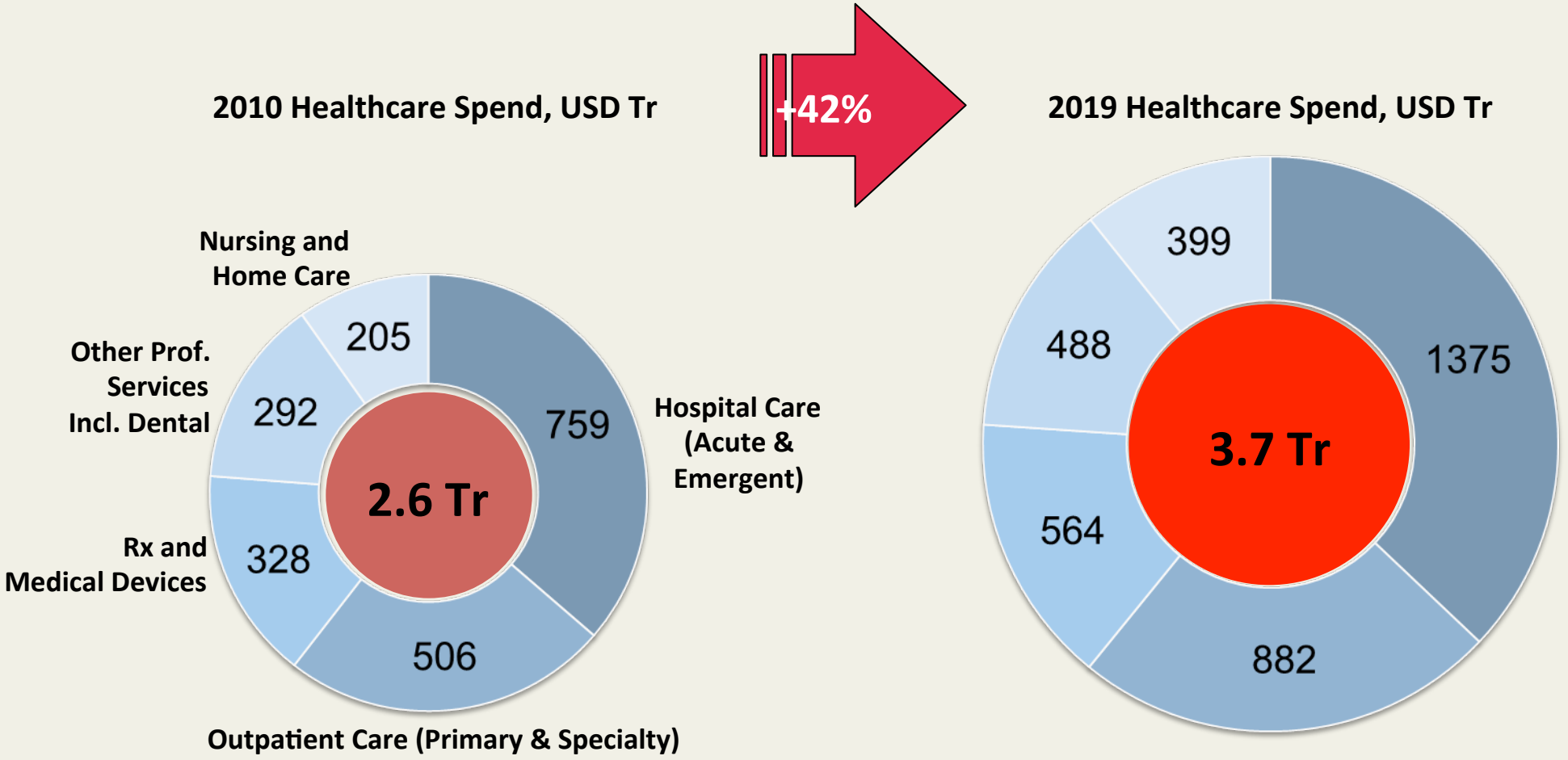


Contributing Factors

- Uncompensated (i.e. charity) and low reimbursement (i.e. Medicaid) care levels growing faster than increase in net patient revenue
- Misaligned financial incentives between physicians and hospitals contribute to overutilization and excessive labor/non-labor costs

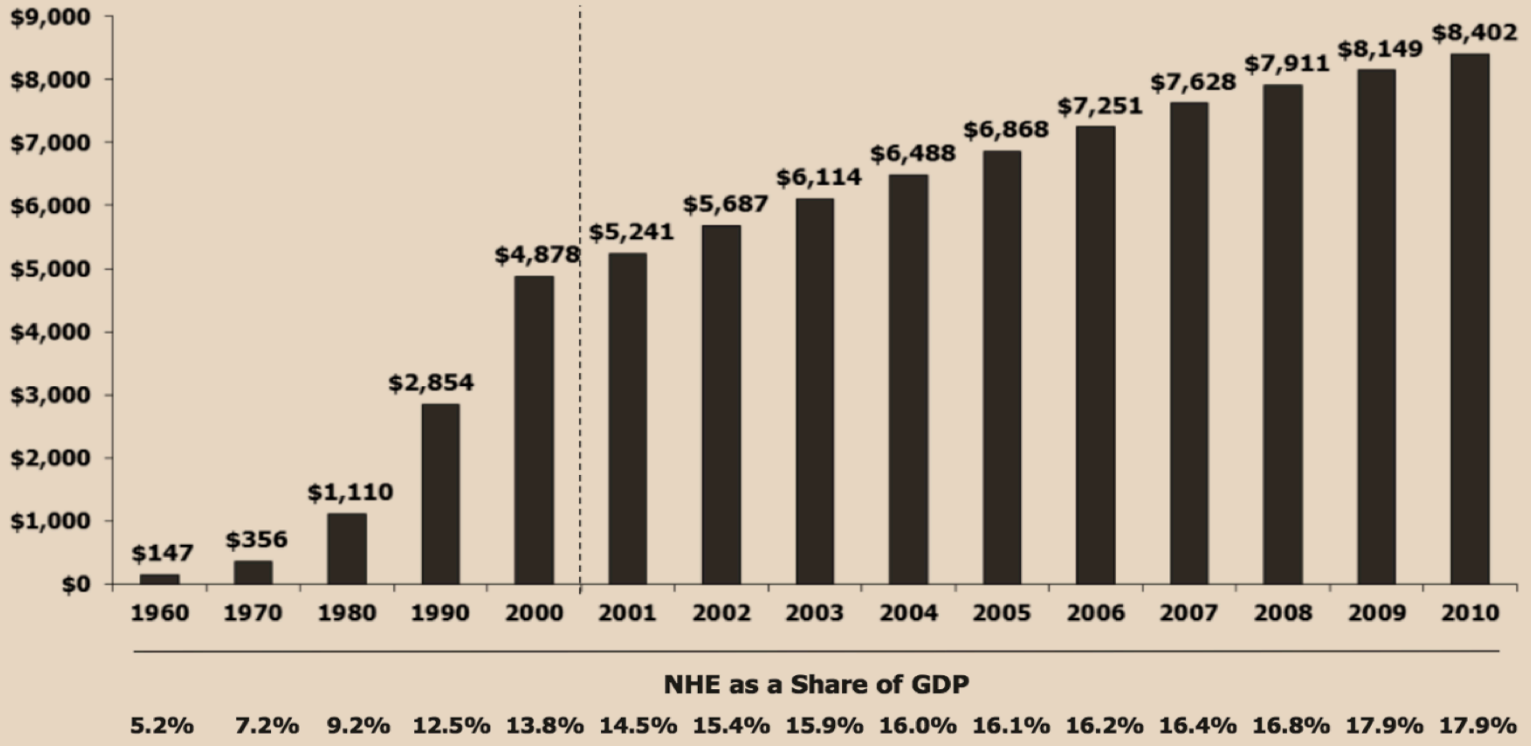
Source: http://healthcare.thomsonreuters.com/thought-leadership/assets/HospOperatingTrendsQuarterlyMarch2011.pdf

Healthcare spend is set to grow almost 50% by 2019 – hospital spend and nursing home spend are fastest growing



Source: CMS http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/index.html?redirect=/NationalHealthExpendData/

Figure 1: National Health Expenditures per Capita, 1960-2010



Notes: According to CMS, population is the U.S. Bureau of the Census resident-based population, less armed forces overseas.

Source: Centers for Medicare and Medicaid Services, Office of the Actuary, National Health Statistics Group, at http://www.cms.hhs.gov/NationalHealthExpendData/ (see Historical; NHE summary including share of GDP, CY 1960-2010; file nhegdp10.zip).

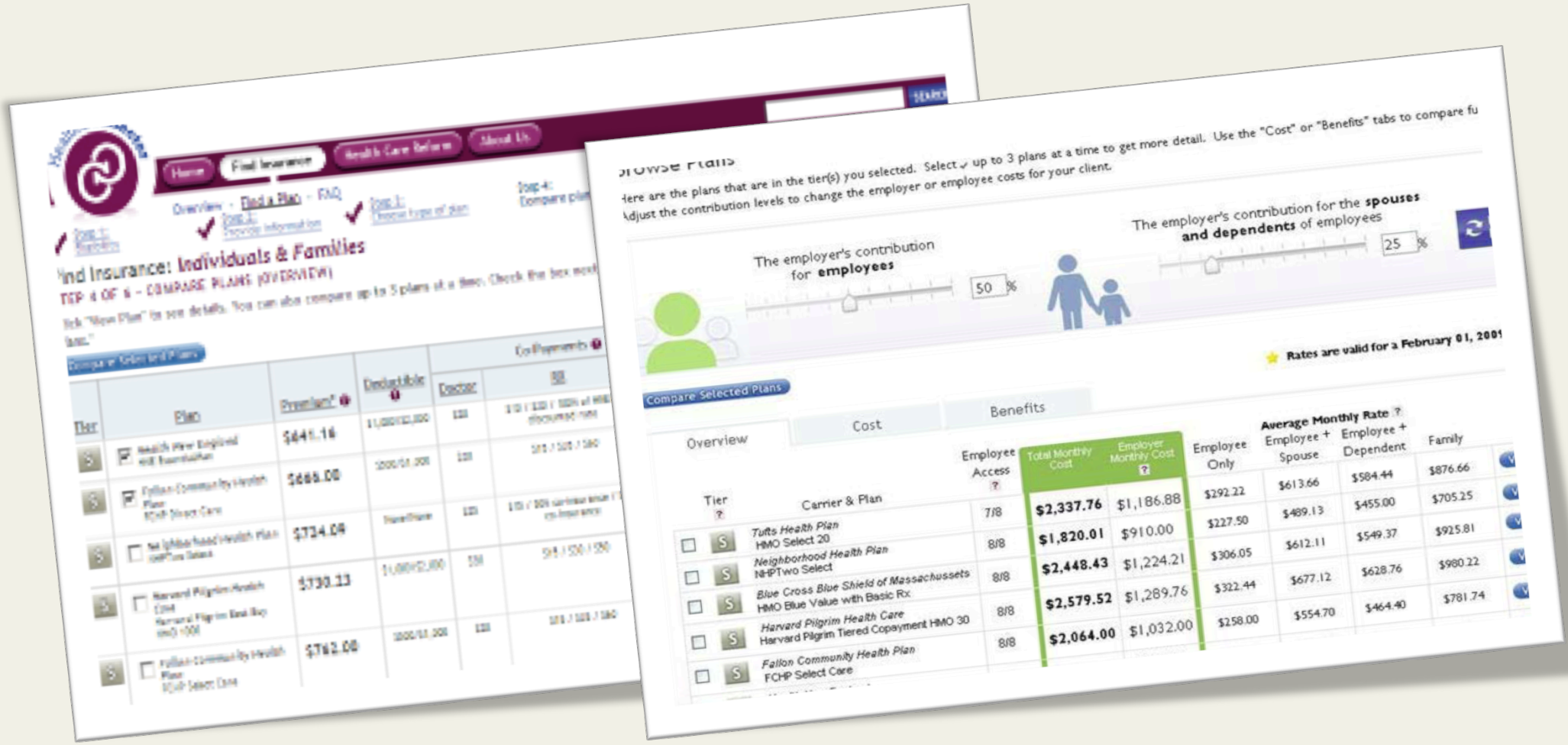
“The share of economic activity (gross domestic product, or GDP) devoted to health care has increased from 7.2% in 1970 to 17.9% in 2009 and 2010”

Key Facts

- In 2010, the U.S. spent $2.6 trillion on health care, an average of $8,402 per person.
- Half of health care spending is used to treat just **5%** of population.
- Although only 10% of total health expenditures, spending on prescription drugs has received considerable attention because of its rapid growth (114% from 2000 to 2010).
- Many policy experts believe new technologies and the spread of existing ones account for a large portion of medical spending & its growth.

Source : https://kaiserfamilyfoundation.files.wordpress.com/2013/01/7670-03.pdf

As Health Exchanges are established, the market will increasingly move toward a more competitive, transparent, retail environment



Analytics in Health Care

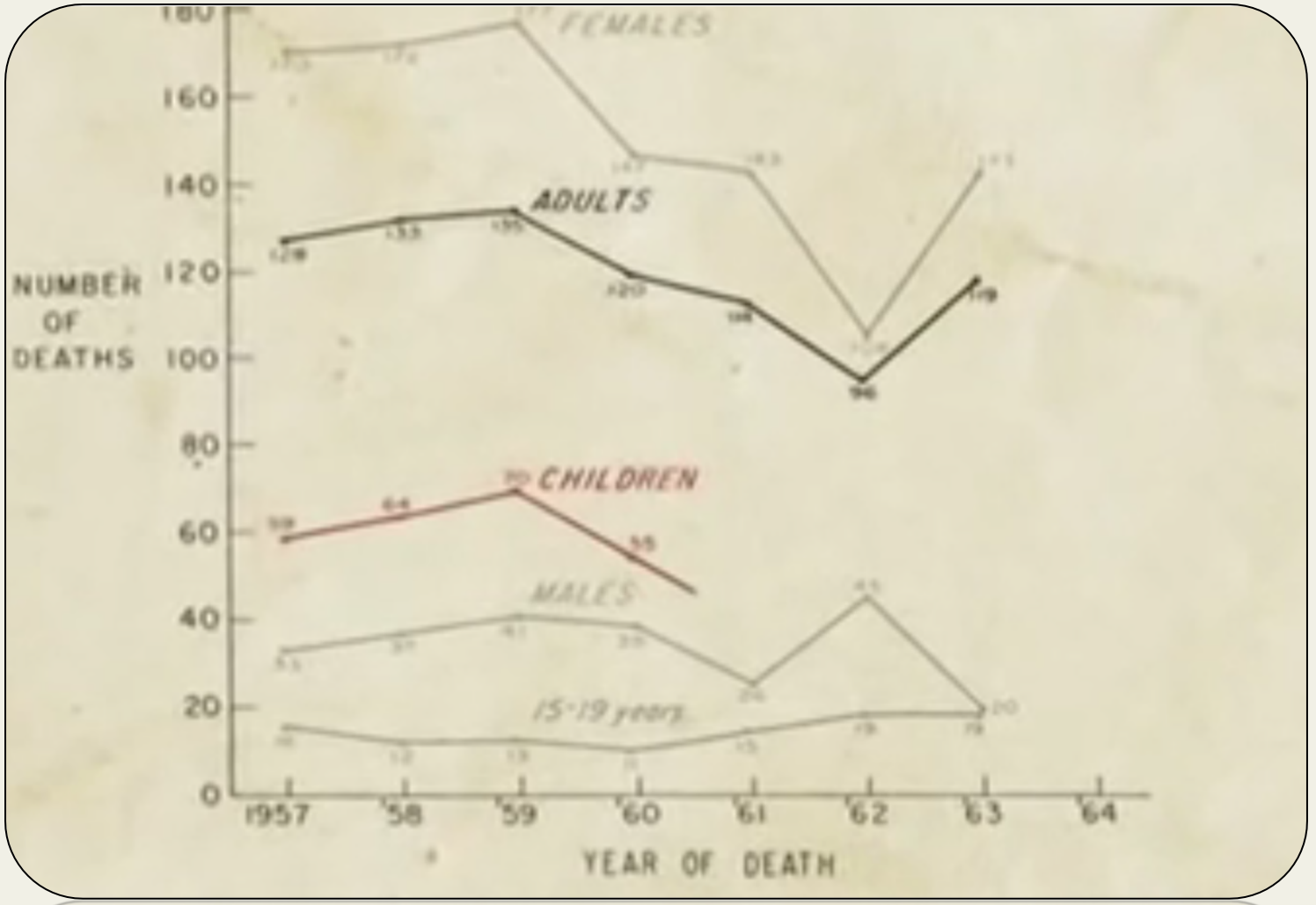
Trends in analytics used in Health Care

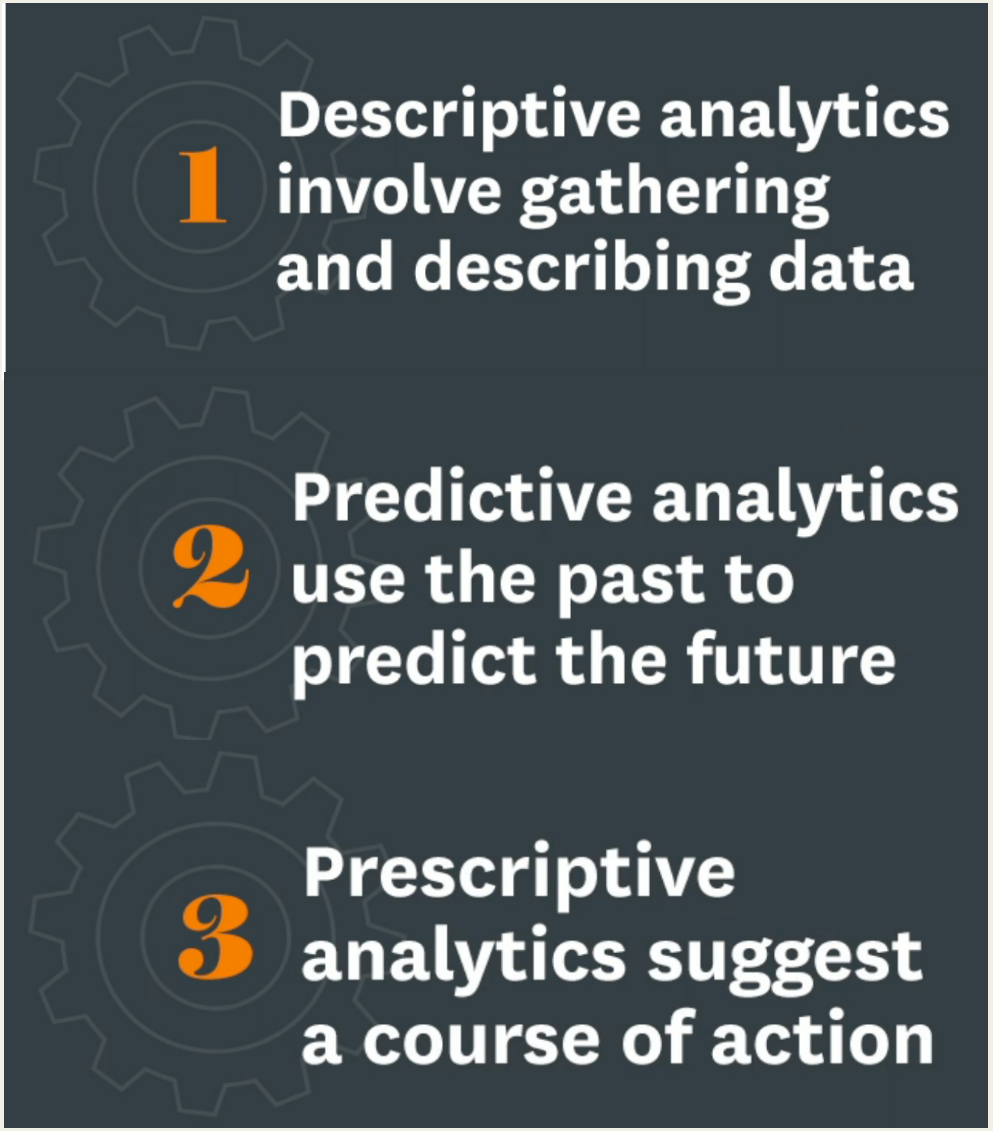
Analytics in HealthCare has been used for a very long time.

(Circa 1950) Basic Statistics Uncovered Kuru : A medical detective Story

Source : https://www.youtube.com/watch?v=ZHOy6hiUcyA

(1950-1964) Dr. Michael Alpers saved data for decades to get to a final realization, what causes Kuru using Data Analysis.





“Traditional Reports”

“Correlation”

“Correlation & Causation”

Source: https://hbr.org/video/2386816175001/business-analytics-defined

“Predictive analytics has been part of healthcare for several decades. Its applying what doctors have been doing on a bigger scale. What has changed with big data is our ability to process more data(quantity), combine variety of data and measure faster.”

Common Use Cases of Predictive Modeling:

1. Clinical Decision Support
2. Readmissions
3. Chronic Disease management
4. Patient Matching

Clinical Decision Support – Sepsis Alert

Use Case:
Very common use case. Implementation of a cloud based alerting system for variety of inpatient activities – care, Labs.

KPI: Better quality of care and sending less patients to emergencies. Cost reduction for patients and hospitals.

Clinical Decision Support Reduces Sepsis

Implementation of a cloud-based alerting system and change management were associated with a 53% drop in sepsis mortality in a single-site study.

http://www.medscape.com/viewarticle/845051

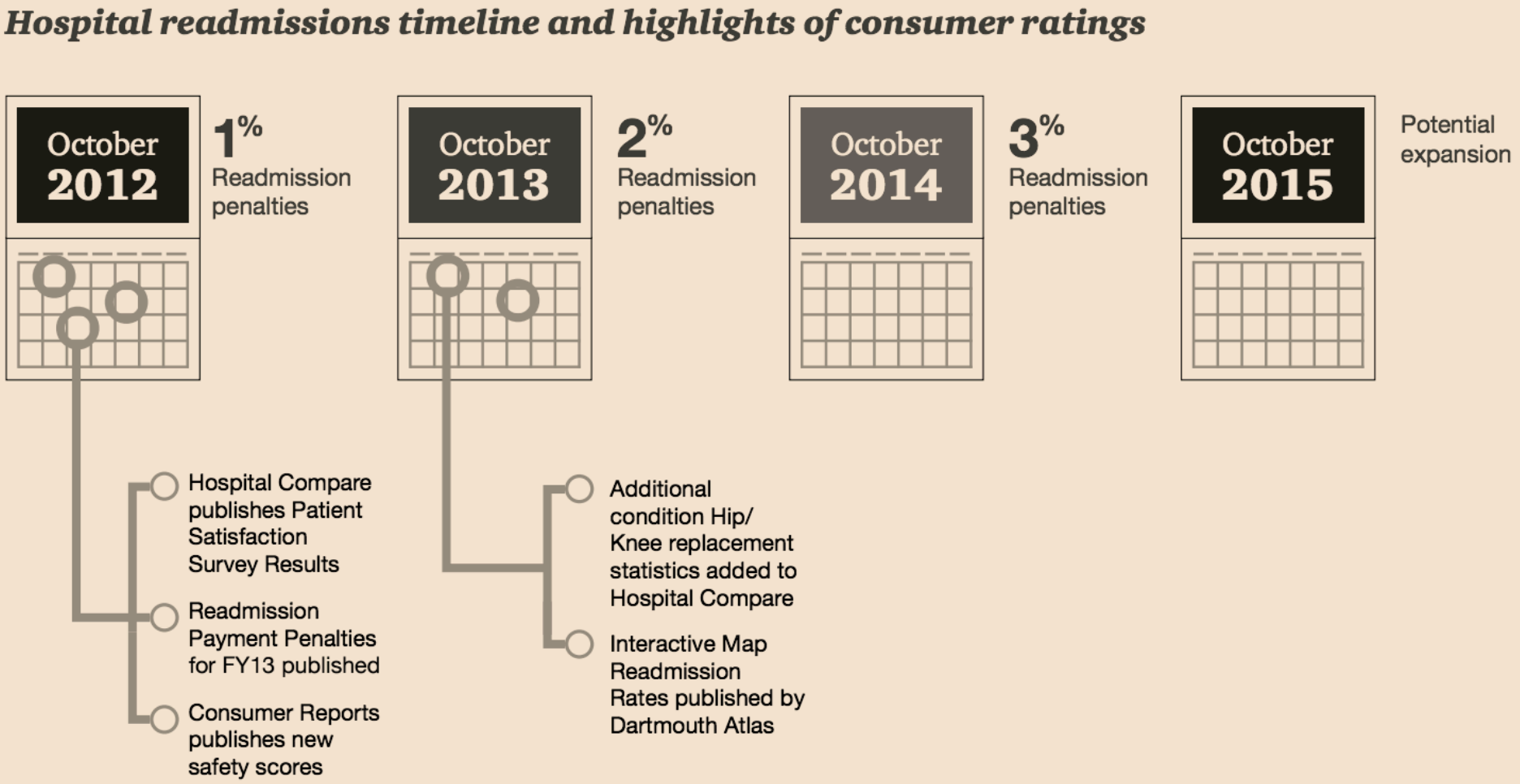
Readmissions

Use Case: Recent changes to Centers for Medicare and Medicaid Services (CMS) compensation will no longer cover hospital expenses for congestive heart failure (CHF) patients who are readmitted within 30 days of discharge. In addition to the risk of degraded health outcomes, the hospital faces financial loss for preventable CHF readmits.

KPI: An improved understanding of the prime drivers affecting readmission will benefit both the quality of care and business revenues.

http://www.commonwealthfund.org/~/media/Files/Publications/Case%20Study/2011/Apr/1473_SilowCarroll_readmissions_synthesis_web_version.pdf

Figure 4. Hospital readmission penalties increase along with publicly reported results



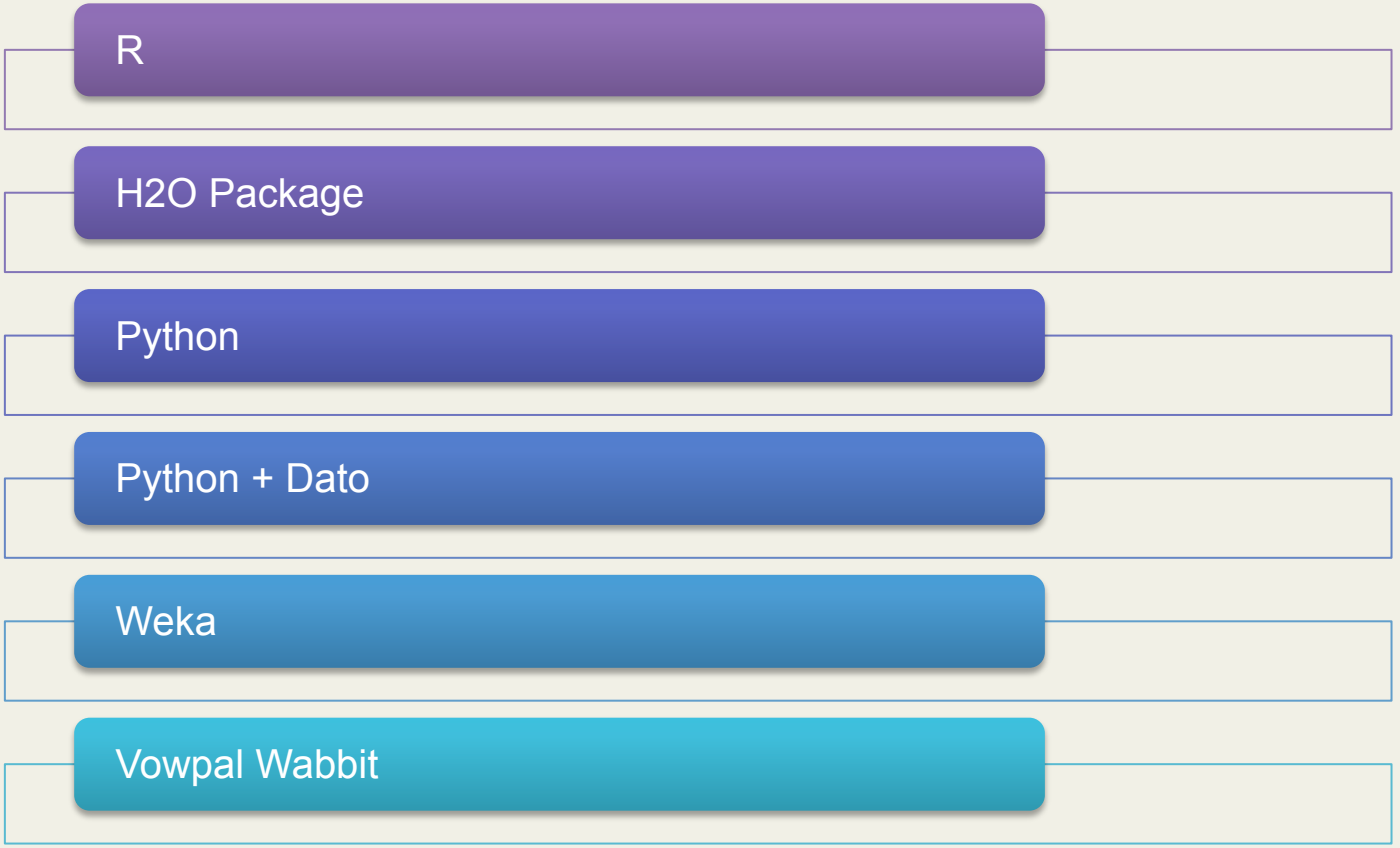
Source: PwC Health Research Institute[26,27,28,29,30]

Predictive Analytics Process & Tools

Predictive Modeling Process

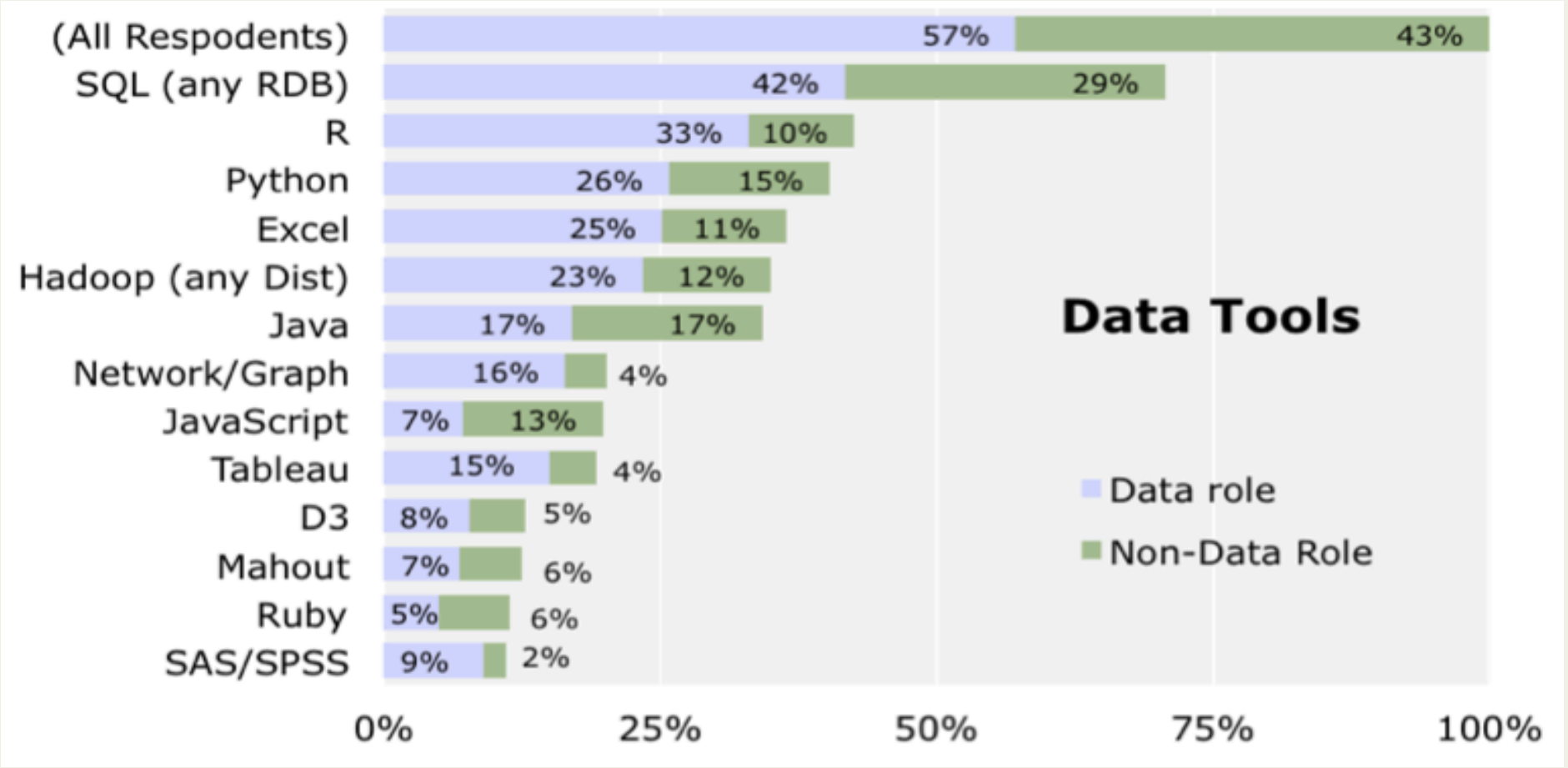
- Collect Data
- Visualizing & Understanding Data
- Determine Metrics – ROC, AUC (C-stat), Sensitivity/Specificity
- Dividing Data into Training, Validation, & Test
- Factor Engineering
- Applying Algorithm(s)
- Reviewing Training Results.
- Scoring
- Feedback new data into collection set

Tools



Survey on top tools used by Data Scientist

http://blog.revolutionanalytics.com/2014/01/in-data-scientist-survey-r-is-the-most-used-tool-other-than-databases.html



Example – Kaggle Bike Sharing

About Bike Share

Competition:
http://www.kaggle.com/c/bike-sharing-demand

Challenge:
Forecast use of a city's bike share system

Source : Bike-Share Data

Publication :
Fanaee-T, Hadi, and Gama, Joao, Event labeling combining ensemble detectors and background knowledge, Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg.

Data Ingest

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
1	2011-01-01 00:00:00	1	0	0	1	9.84	14.395	81	0.0000	3	13	16
2	2011-01-01 01:00:00	1	0	0	1	9.02	13.635	80	0.0000	8	32	40
3	2011-01-01 02:00:00	1	0	0	1	9.02	13.635	80	0.0000	5	27	32
4	2011-01-01 03:00:00	1	0	0	1	9.84	14.395	75	0.0000	3	10	13
5	2011-01-01 04:00:00	1	0	0	1	9.84	14.395	75	0.0000	0	1	1
6	2011-01-01 05:00:00	1	0	0	2	9.84	12.880	75	6.0032	0	1	1

The goal is to predict counts either based on sum of casual & registered or directly

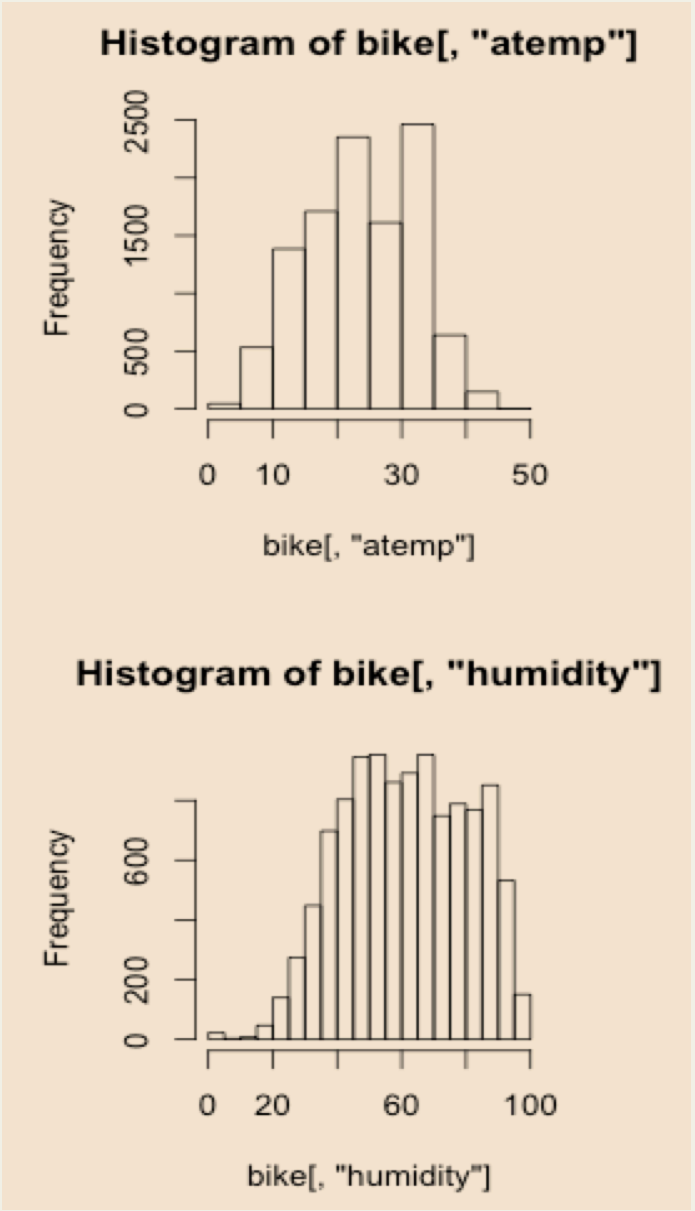
Data Fields

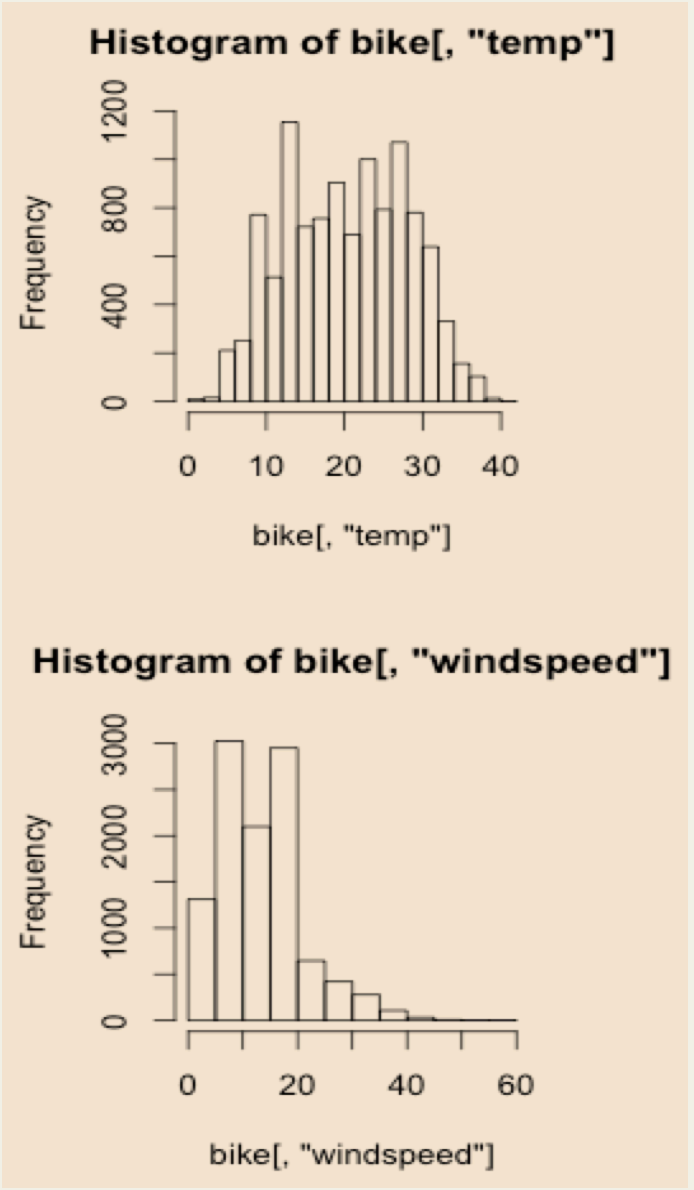
Datetime : hourly date + timestamp
Season : 1 = spring, 2 = summer, 3 = fall, 4 = winter
Holiday : whether the day is considered a holiday
Workingday : whether the day is neither a weekend nor holiday
Weather :

1: Clear, Few clouds, Partly cloudy, Partly cloudy
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

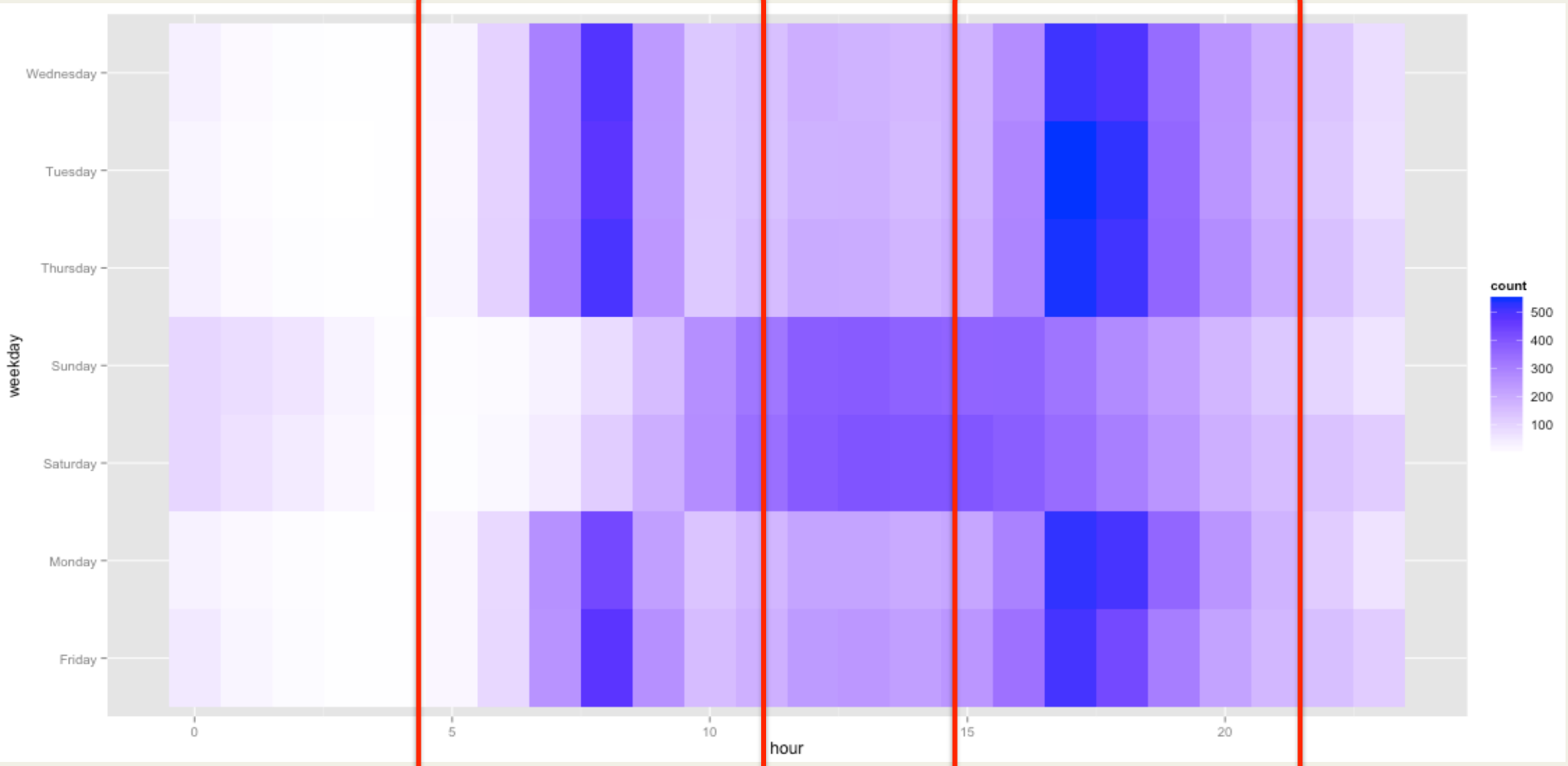
Temp : temperature in Celsius
Atemp : "feels like" temperature in Celsius
Humidity : relative humidity
Windspeed : wind speed
Casual : number of non-registered user rentals initiated
Registered : number of registered user rentals initiated
Count : number of total rental

Data - Continuous

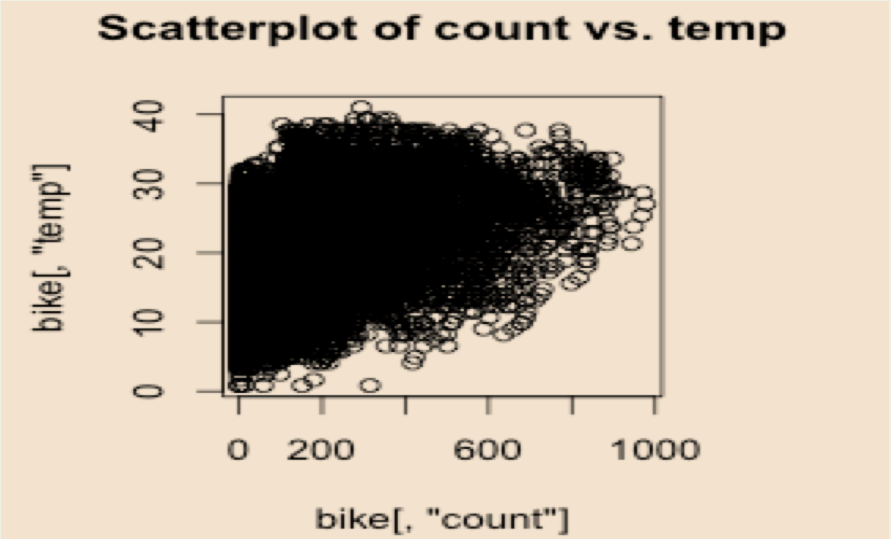


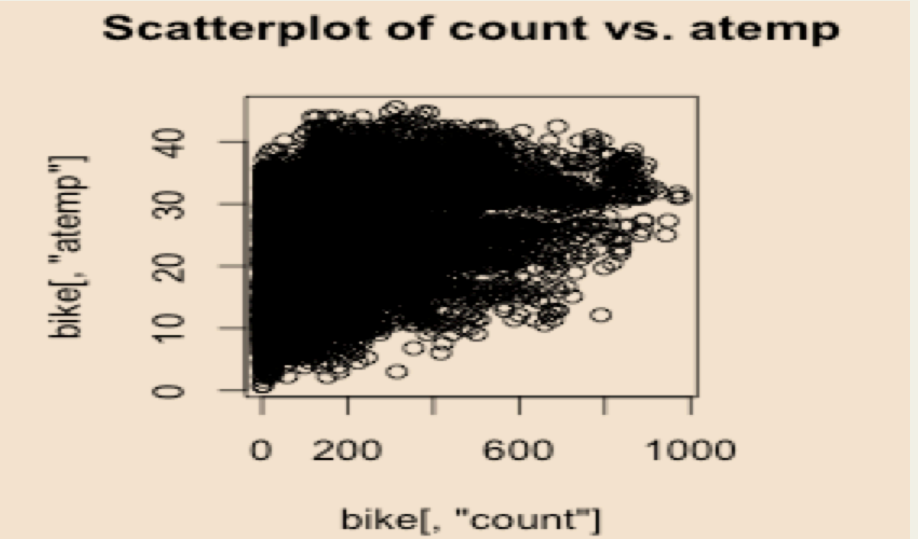


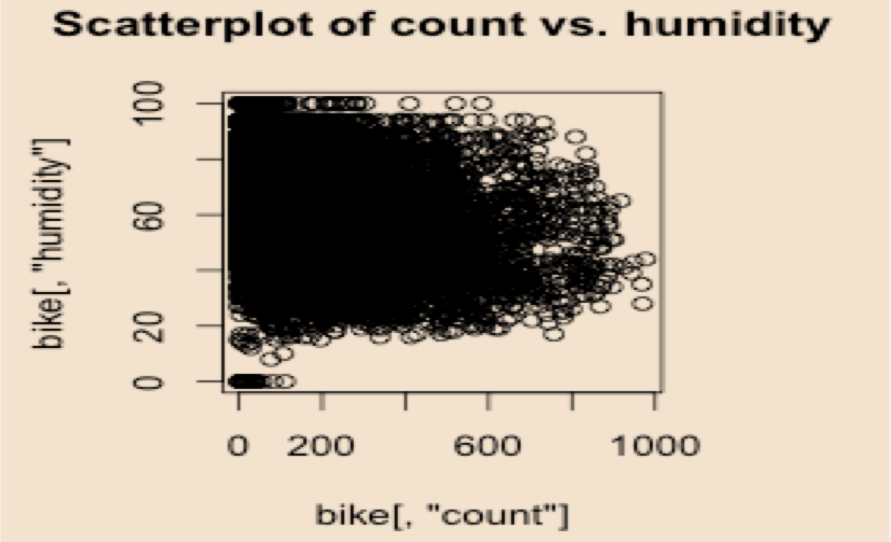
Workday busy hours

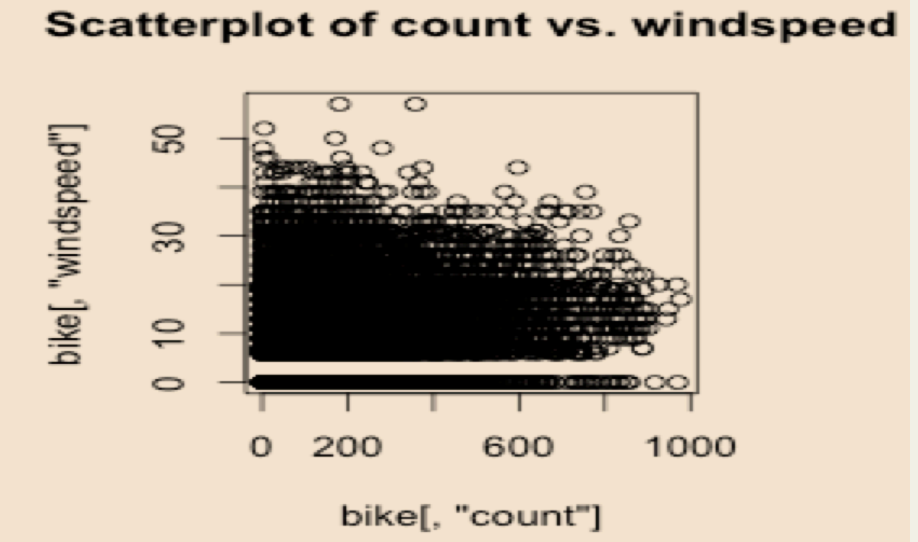


Data

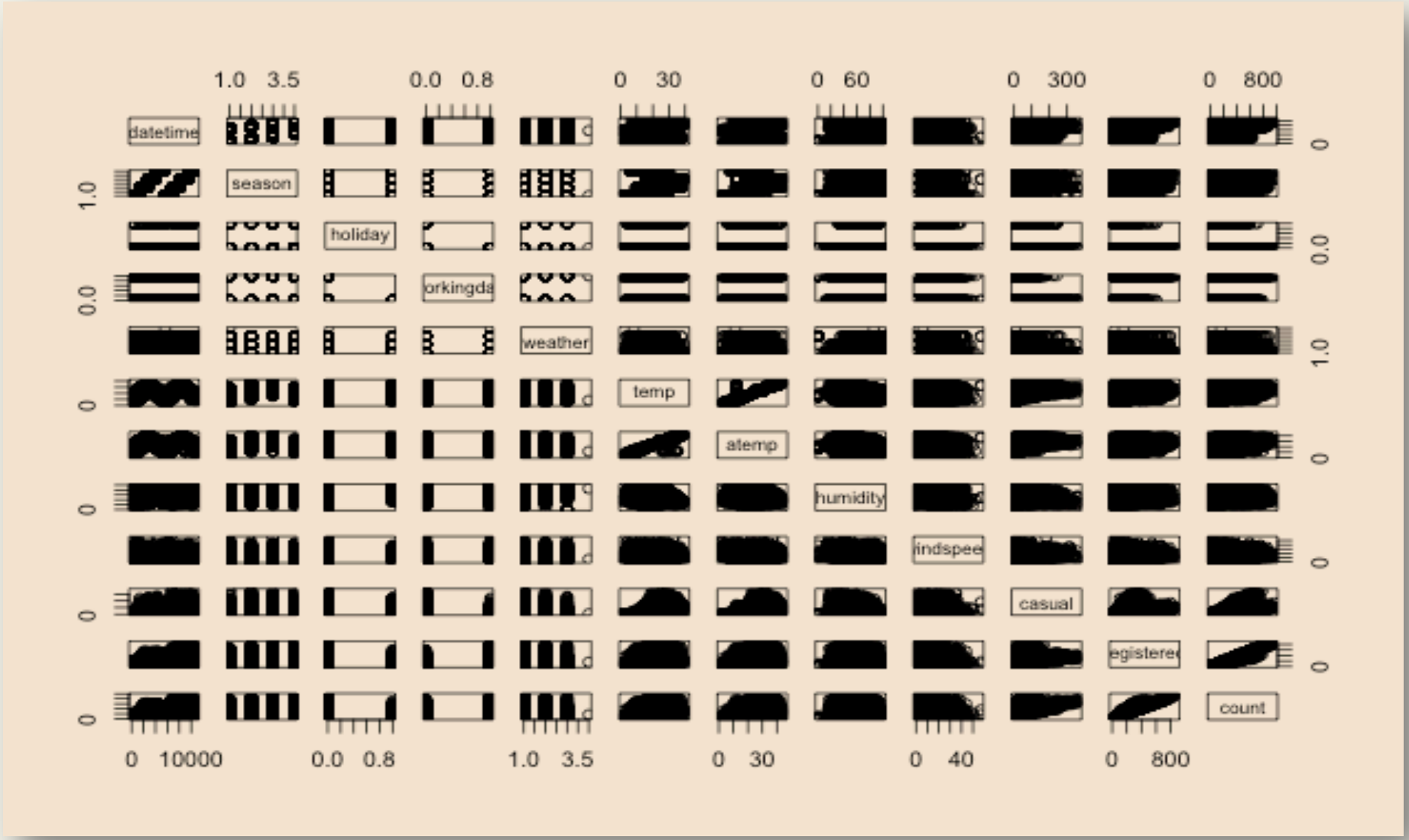

Scatterplot of count vs. temp
bike[, "temp"]
0
10
20
30
40
0
200
600
1000
bike[, "count"]


Scatterplot of count vs. atemp
bike[, "atemp"]
0
10
20
30
40
0
200
600
1000
bike[, "count"]


Scatterplot of count vs. humidity
bike[, "humidity"]
0
20
60
100
0
200
600
1000
bike[, "count"]


Scatterplot of count vs. windspeed
bike[, "windspeed"]
0
10
30
50
0
200
600
1000
bike[, "count"]

Pairs Command - R



Using R

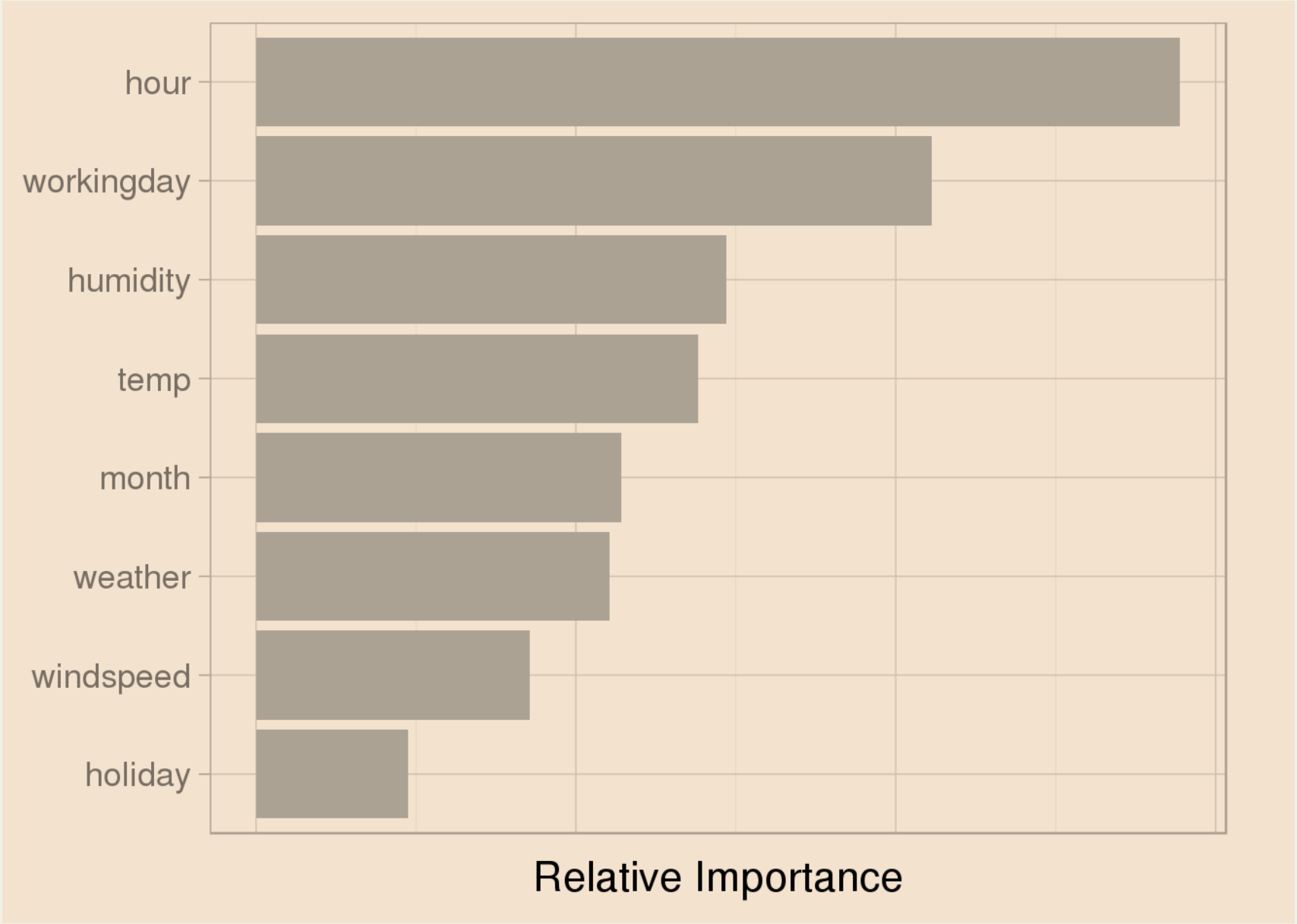
```r
#------------------------------------------------------------
#  Main Program for kaggle-bike-sharing
#  Taposh Roy @taposh_dr
#------------------------------------------------------------
setwd("/Users/taposh/workspace/kaggle/bikeshare/")
#sink the output
sink("bikeshare.log", split = T)
#source the libraries
source("mylibraries.R")
#input data
source("inputdata.R")
#View the data
head(bike)
#Visualize the data
source("/Users/taposh/workspace/kaggle/bikeshare/visualize.R")
#Utility Functions
source("/Users/taposh/workspace/kaggle/bikeshare/utils.R")

####################################
### Method #1 Matrix Computation
####################################
#Factorengineering
source("/Users/taposh/workspace/kaggle/bikeshare/factorengineering_bike.R")
source("matrix.R")
####################################
### Method #2 Models
####################################
#Factorengineering
source("/Users/taposh/workspace/kaggle/bikeshare/factorengineering_v1.R")
#colnames(bike)
#colnames(test)
source("models_nongrid.R")
####################################
### Method #2 Models
####################################
#Factorengineering
source("models.R")

```

Feature Engineering

```r
#-----------------------------------------------------
#  Factor Engineering for kaggle-bike-sharing
#  Taposh Roy
#  @taposh_dr
#-----------------------------------------------------

#Seperating out the outputs from training
actual<-c()
countresult <-c()
causal<-c()
registered<-c()
countresult<-cbind("count"=countresult,bike[,"count"])
actual<-cbind("count"=actual,bike[,"count"])
causal<-cbind(causal,bike[,"casual"])
registered<-cbind(registered,bike[,"registered"])

causal<-log(causal)
causal[causal[,1]<0,1]<-0
```

Models

```r
#-----------------------------------------------------
#  Factor Engineering for kaggle-bike-sharing
#  Taposh Roy
#  @taposh_dr
#-----------------------------------------------------

#Seperating out the outputs from training
actual<-c()
countresult <-c()
causal<-c()
registered<-c()
countresult<-cbind("count"=countresult,bike[,"count"])
actual<-cbind("count"=actual,bike[,"count"])
causal<-cbind(causal,bike[,"casual"])
registered<-cbind(registered,bike[,"registered"])

causal<-log(causal)
causal[causal[,1]<0,1]<-0
```


hour
workingday
humidity
temp
month
weather
windspeed
holiday
Relative Importance

Thank You !!