Introduction to Data Science

Sara Bates & Diana Pfeil

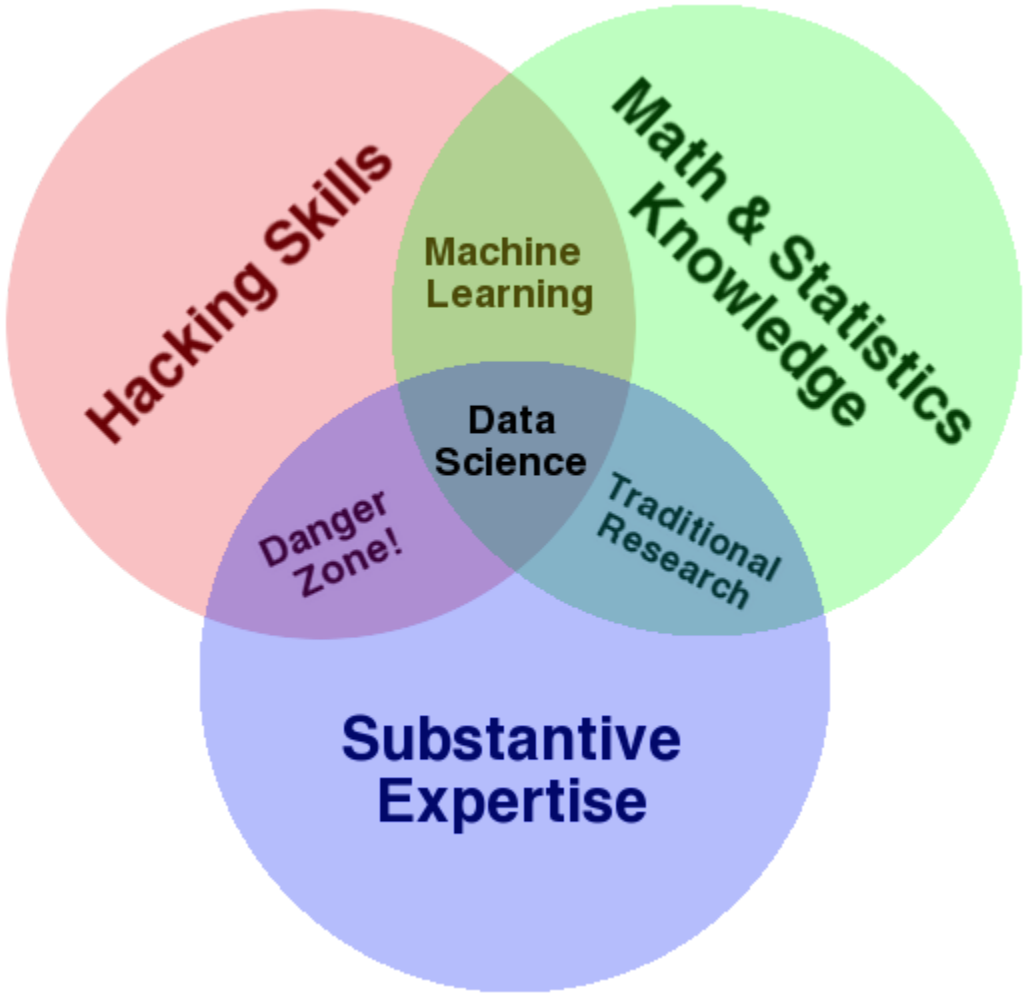
Girl Develop It Boulder, March 2015

About Us

Sara Bates, @BatesSara

Diana Pfeil, @dianam

What is data science?



Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

...data scientists [need] to communicate in language that all their stakeholders understand—and to demonstrate the special skills involved in storytelling with data, whether verbally, visually, or — ideally — both.

- DJ Patil and Thomas Davenport, Data Scientist: The Sexiest Job of the 21st Century, Harvard Business Review

...the dominant trait among data scientists is an intense curiosity — a desire to go beneath the surface of a problem, find the questions at its heart, and distill them into a very clear set of hypotheses that can be tested.

- DJ Patil and Thomas Davenport, Data Scientist: The Sexiest Job of the 21st Century, Harvard Business Review

Why is data scientist "The Sexiest Job of the 21st Century"?

What is Big Data?

What do data scientists do?

- Exploratory data analysis
- Predictive analytics and machine learning
- Testing and experimentation
- Data communication and visualization

A Typical Toolkit

- Python with pandas, numpy, scipy
- R, RStudio
- Unix utilities
- JVM (aka Java-based) tools: Hadoop, Spark

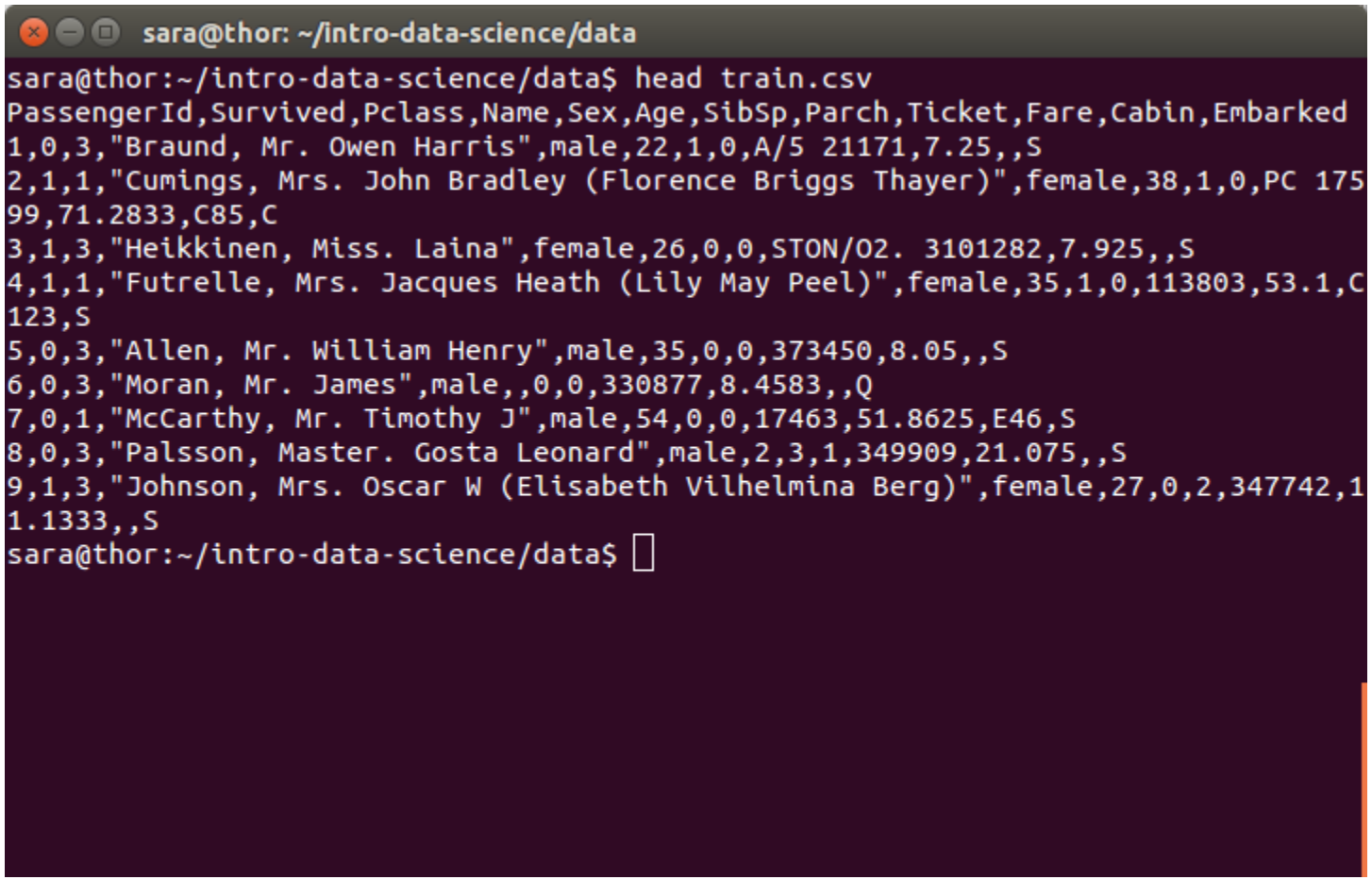
Exploratory Data Analysis

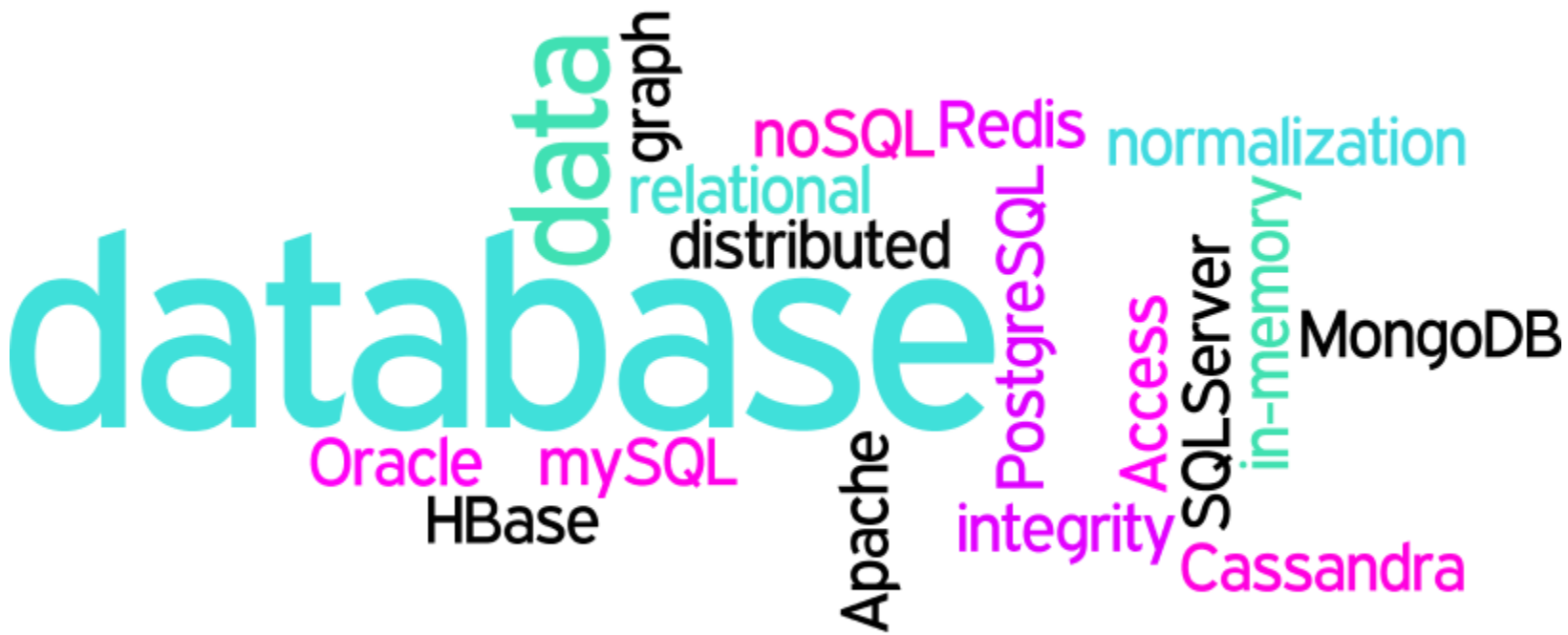
Titanic Survivors

Data Legend

survived	Survival (0 = No; 1 = Yes)
Pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Data formats




database
data
graph
noSQL
Redis
normalization
relational
distributed
PostgreSQL
Access
SQLServer
in-memory
MongoDB
Oracle
mySQL
HBase
Apache
integrity
Cassandra

Dataframes

```
import pandas as pd

df = pd.read_csv("../data/train.csv")
print df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	NaN	S

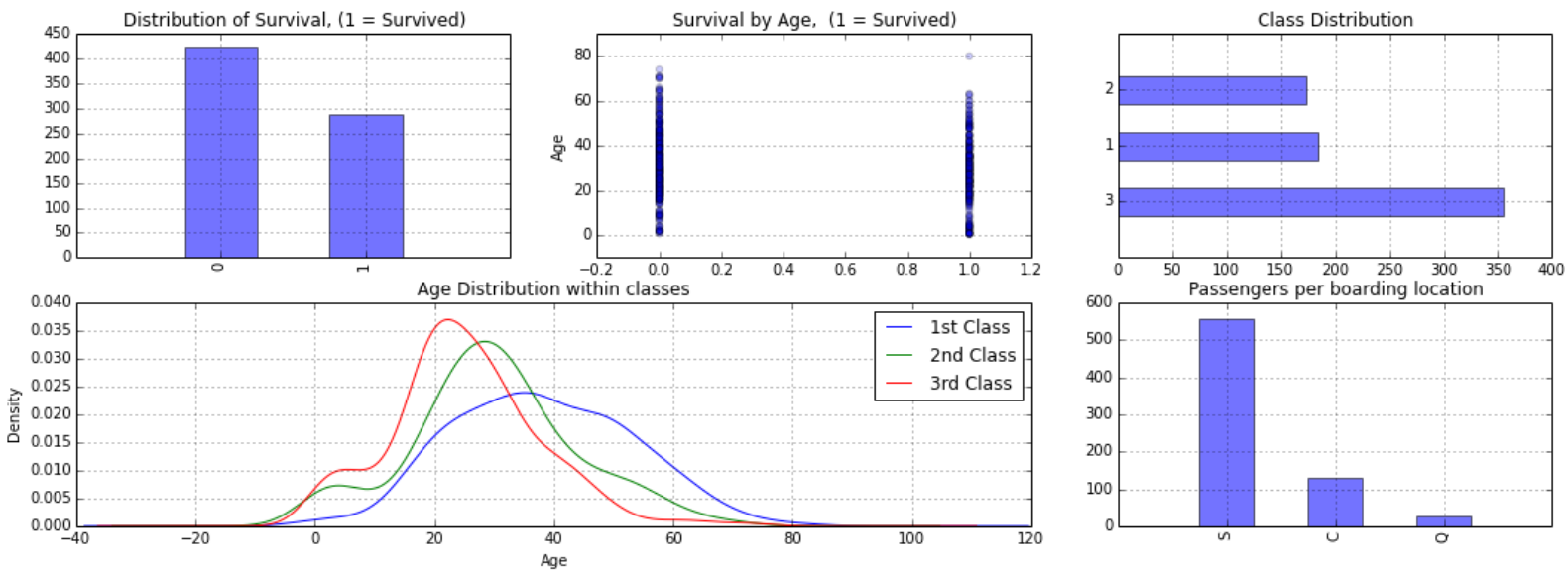
Summary Statistics

```
print df.dtypes
print df['Age'].describe()
```

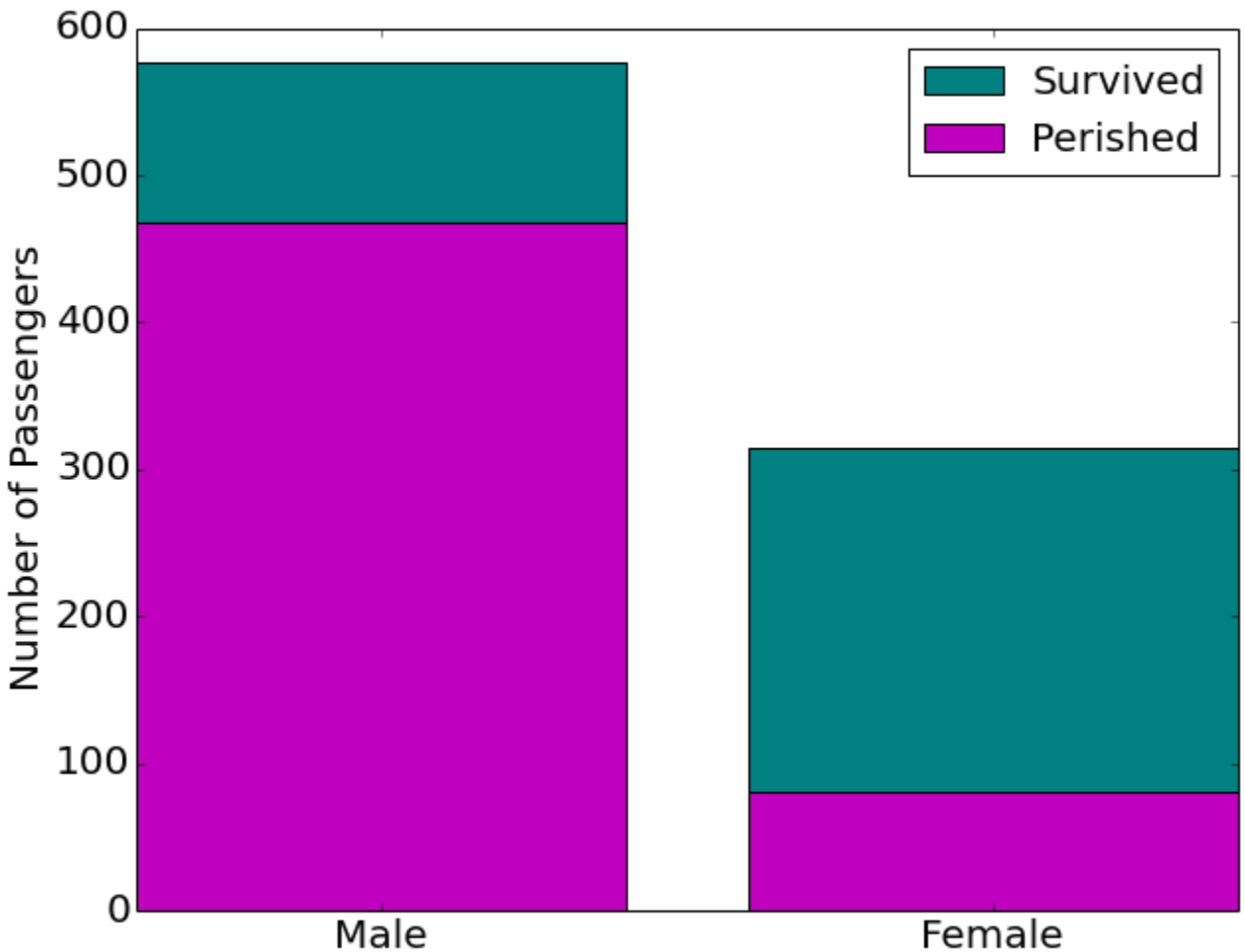
```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

```
count    714.000000
mean      29.699118
std       14.526497
min        0.420000
25%       20.125000
50%       28.000000
75%       38.000000
max       80.000000
dtype: float64
```

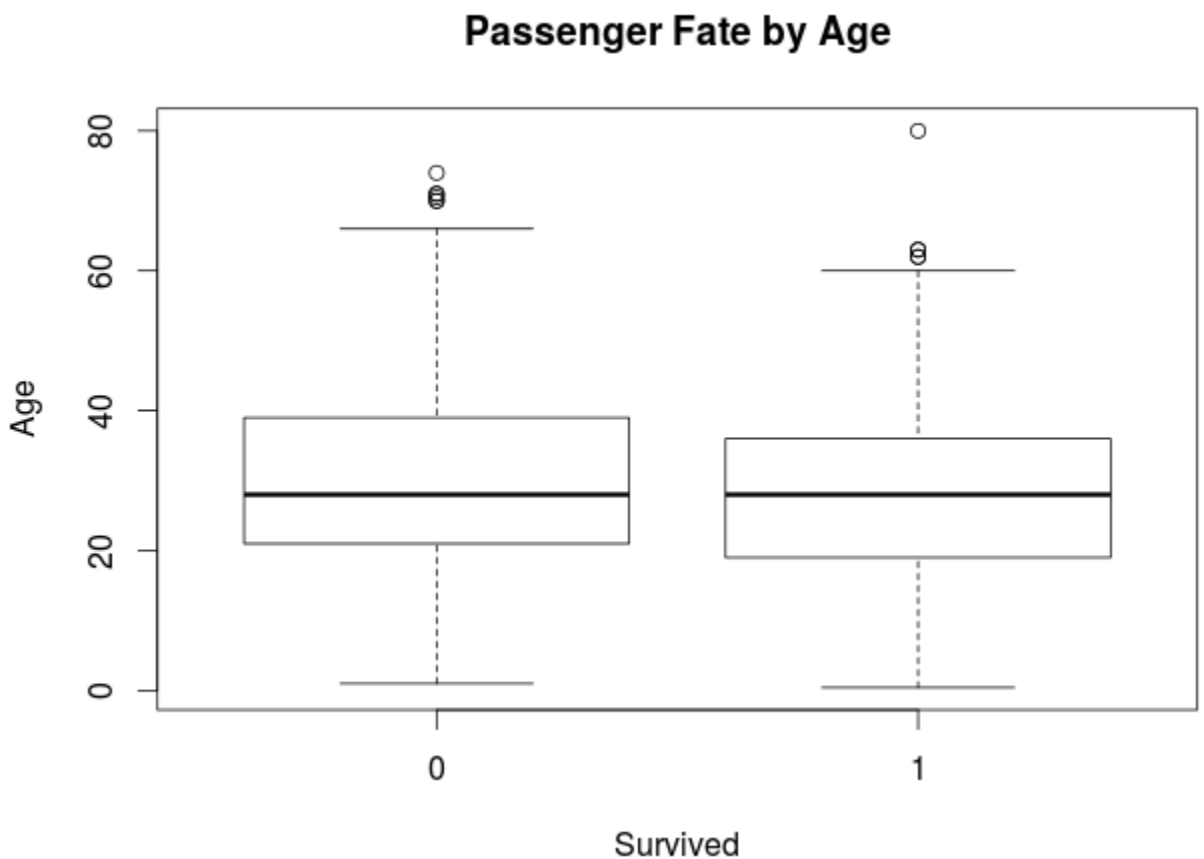
Plotting


Distribution of Survival, (1 = Survived)
Survival by Age, (1 = Survived)
Class Distribution
Age Distribution within classes
Density
Age
1st Class
2nd Class
3rd Class
Passengers per boarding location
S
C
Q

Women and children first?



Women and children first?



Data Issues

- Typos
- Missing data
- Redundant data
- Formatting
- Outliers

Example Data

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	NaN	C
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7000	G6	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.5500	C103	S
12	13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.0500	NaN	S
13	14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.2750	NaN	S
14	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542	NaN	S
15	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16.0000	NaN	S
16	17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.1250	NaN	Q
17	18	1	2	Williams, Mr. Charles Eugene	male	NaN	0	0	244373	13.0000	NaN	S
18	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vande...	female	31	1	0	345763	18.0000	NaN	S
19	20	1	3	Masselmani, Mrs. Fatima	female	NaN	0	0	2649	7225.0000	NaN	C
20	21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26.0000	NaN	S
21	22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13.0000	D56	S
22	23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292	NaN	Q
23	24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5000	A6	S
24	25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	349909	21.0750	NaN	S

Missing Data

```
df = df.drop(['Cabin'], axis=1)

med_age = df.Age.median()
df.Age = df.Age.fillna(med_age)
```

Formatting & Typos

```
df['Fare'][19] = 7.2250

df = df.drop(['Ticket'], axis=1)
```

Predictive Analytics and Machine Learning

Machine Learning

Supervised Learning

Unsupervised Learning

Statistical Modeling

Descriptive, Predictive, and Prescriptive Analytics

Supervised Learning

$\mathbf{x_i}$	features (input variables)
y_i	target (output variable)
$(x_i, y_i), i = 1, \ldots, m$	training set

Goal: learn a function

$$h : \mathcal{X} \to \mathcal{Y}$$

such that $h(x)$ is a good predictor of y on **new** data

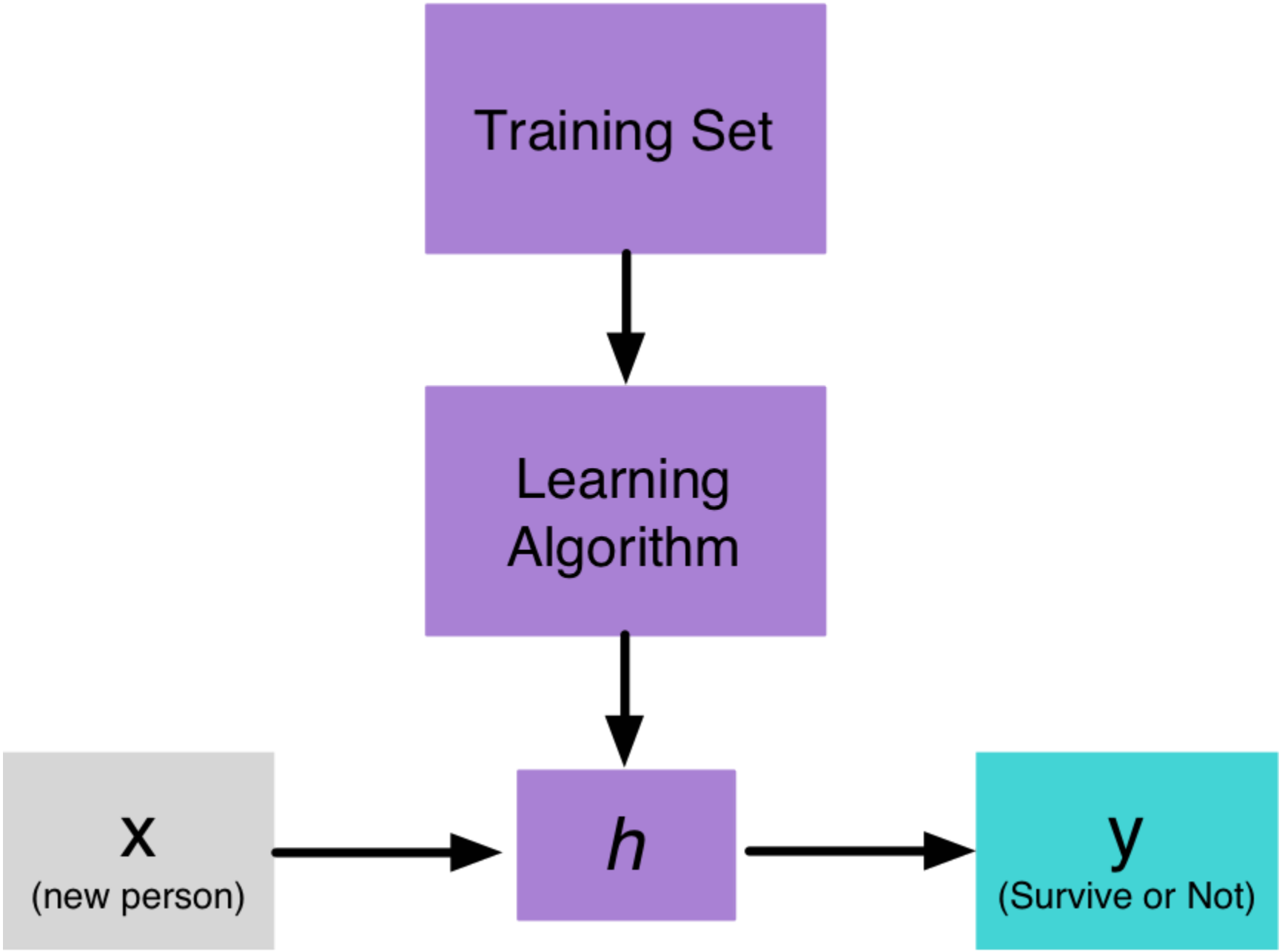
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2	<NA>	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.3	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9	<NA>	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.1	<NA>	S
6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.5	<NA>	Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.9	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.1	<NA>	S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1	<NA>	S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.1	<NA>	C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.6	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.1	<NA>	S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.3	<NA>	S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.9	<NA>	S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16.0	<NA>	S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.1	<NA>	Q
18	1	2	Williams, Mr. Charles Eugene	male	NA	0	0	244373	13.0	<NA>	S
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18.0	<NA>	S
20	1	3	Masselmani, Mrs. Fatima	female	NA	0	0	2649	7.2	<NA>	C
21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26.0	<NA>	S
22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13.0	D56	S
23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0	<NA>	Q
24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6	S
25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	349909	21.1	<NA>	S

features x can be

numeric/metric	Age: 14, 56, 1
ordinal	Ranking: 1st, 2nd, 3rd
categorical/nominal	Sex: male/female

target y can be

continuous (regression)	Housing Price: 500K, 150K, 2MM
categorical (classification)	Survival: Perish, Survive


Training Set
Learning Algorithm
x
(new person)
h
y
(Survive or Not)

But where do we find this h?

This is the process of doing predictive modeling

Predicting Survival on the Titanic

1. Clean and explore the data
2. Come up with new features
3. Split data into training and test
4. Tune the model and parameters using cross-validation
5. Compare model results

Step 2: Feature Engineering

```
> summary(titanic)
  PassengerId  Survived Pclass                                      Name         Sex          Age
 Min.   :  1   0:549    1:216   Abbing, Mr. Anthony                    :  1   female:314   Min.   : 0
 1st Qu.:224   1:342    2:184   Abbott, Mr. Rossmore Edward            :  1   male  :577   1st Qu.:21
 Median :446             3:491   Abbott, Mrs. Stanton (Rosa Hunt)       :  1                Median :30
 Mean   :446                     Abelson, Mr. Samuel                    :  1                Mean   :29
 3rd Qu.:668                     Abelson, Mrs. Samuel (Hannah Wizosky):  1                3rd Qu.:35
 Max.   :891                     Adahl, Mr. Mauritz Nils Martin         :  1                Max.   :80
                                 (Other)                                :885
     SibSp           Parch             Ticket         Fare               Cabin     Embarked
 Min.   :0.0   Min.   :0.0   1601     :  7   Min.   :  4   B96 B98        :  4   C:168
 1st Qu.:0.0   1st Qu.:0.0   347082   :  7   1st Qu.:  8   C23 C25 C27:  4   Q: 77
 Median :0.0   Median :0.0   CA. 2343:  7   Median : 14   G6             :  4   S:646
 Mean   :0.5   Mean   :0.4   3101295  :  6   Mean   : 33   C22 C26        :  3
 3rd Qu.:1.0   3rd Qu.:0.0   347088   :  6   3rd Qu.: 31   D              :  3
 Max.   :8.0   Max.   :6.0   CA 2144  :  6   Max.   :512   (Other)        :186
                             (Other)  :852                 NA's           :687
```

Extract honorific from the Name feature

```
titanic$Title <- gsub(".*\\, ([A-Za-z ]+)\\..*", "\\1", titanic$Name)
titanic$Title <- as.factor(titanic$Title)

unique(titanic$Title)
 [1] "Mr"           "Mrs"          "Miss"
 [4] "Master"       "Don"          "Rev"
 [7] "Dr"           "Mme"          "Ms"
[10] "Major"        "Lady"         "Sir"
[13] "Mlle"         "Col"          "Capt"
[16] "the Countess" "Jonkheer"
```

More New Features

Family combines siblings and spouses with parents and children

```
data$Family <- data$SibSp + data$Parch
```

Fare.pp attempts to adjust group purchases by size of family

```
data$Fare.pp <- data$Fare/(data$Family + 1)
```

First character in Cabin number represents the Deck

```
data$Deck <- substring(data$Cabin, 1, 1)
data$Deck[ which( is.na(data$Deck ))] <- "UNK"
data$Deck <- as.factor(data$Deck)
```

Odd-numbered cabins were reportedly on the port side of the ship, Even-numbered cabins on the starboard side

```
cabin.last.digit <- str_sub(data$Cabin, -1)
data$Side <- "UNK"
data$Side[which(isEven(cabin.last.digit))] <- "port"
data$Side[which(isOdd(cabin.last.digit))] <- "starboard"
```

Step 3: Split data into training and test

```
require('caret')
set.seed(99)

trainI <- createDataPartition(y = titanic$Fate, p = .80, list = FALSE)
training <- titanic[ trainI,]
test <- titanic[-trainI,]
```

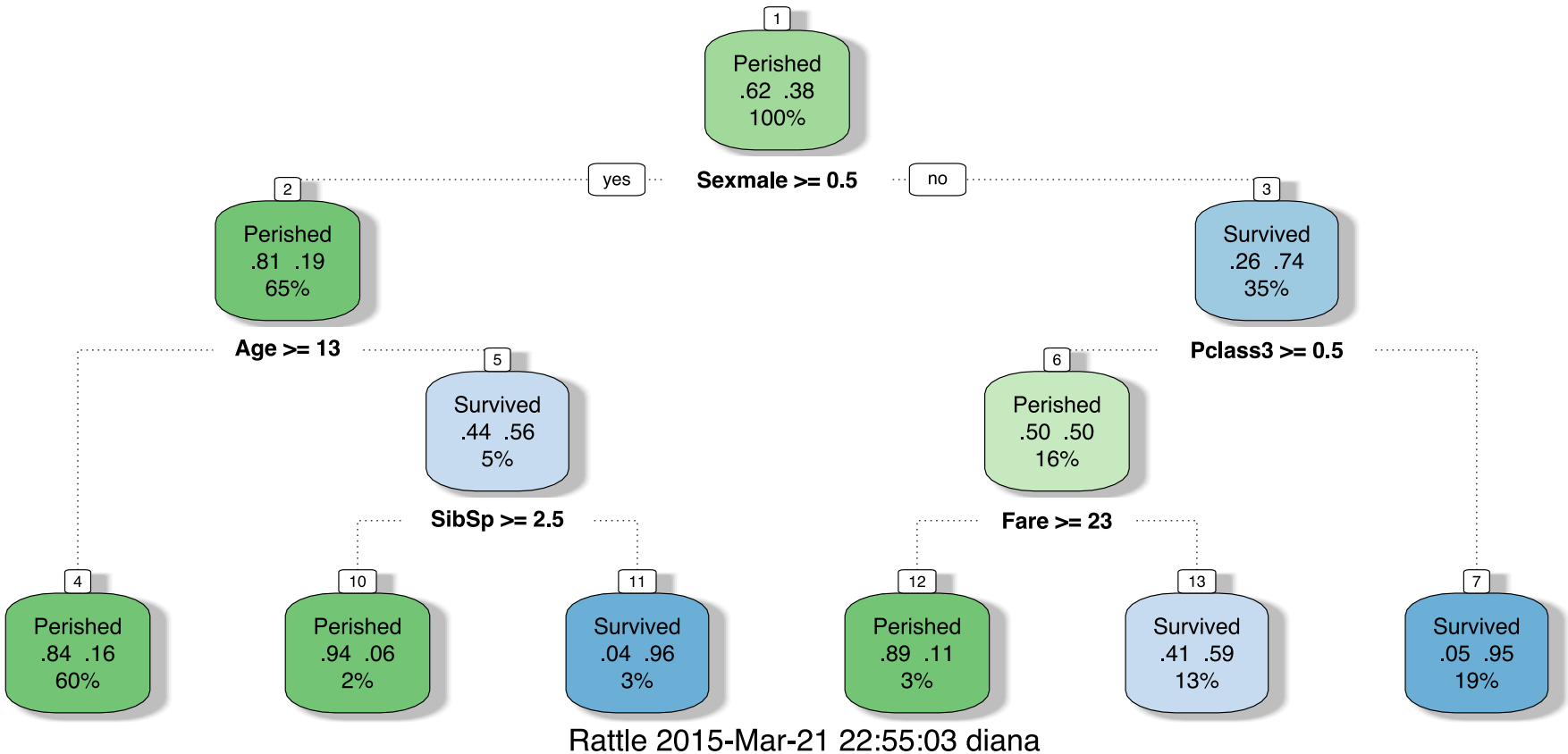
Step 4: Train the model using cross-validation

Which model?

Classification Tree	Regression Tree
Random Forest	Linear Regression
Support Vector Machine	Logistic Regression
Boosting	K-Nearest Neighbors
Naive Bayes	Neural Network

Train a Decision Tree

```
ctrl <- trainControl(method = "cv",
                     classProbs = TRUE,
                     summaryFunction = twoClassSummary)
tree1 <- train(Fate ~ ., data = training,
                     method = "rpart",
                     metric = "ROC",
                     trControl = ctrl)
```


1
Perished
.62 .38
100%
yes
Sexmale >= 0.5
no
2
Perished
.81 .19
65%
3
Survived
.26 .74
35%
Age >= 13
Pclass3 >= 0.5
5
Survived
.44 .56
5%
6
Perished
.50 .50
16%
SibSp >= 2.5
Fare >= 23
4
Perished
.84 .16
60%
10
Perished
.94 .06
2%
11
Survived
.04 .96
3%
12
Perished
.89 .11
3%
13
Survived
.41 .59
13%
7
Survived
.05 .95
19%
Rattle 2015-Mar-21 22:55:03 diana

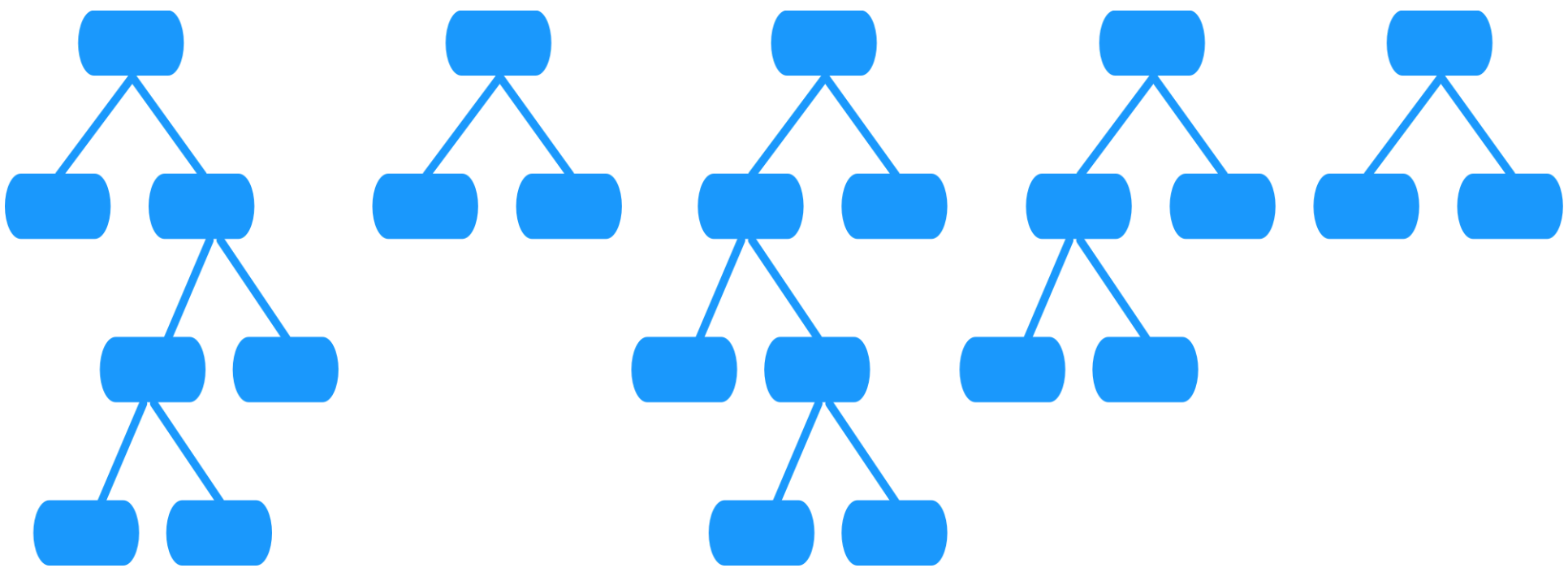
k-fold cross-validation

Confusion Matrix

	Truth Positive (Perish)	Truth Negative (Survive)
Predict Positive (Perish)	True Positive	False Positive
Predict Negative (Survive)	False Negative	True Negative

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Random Forest



Train a Random Forest

```
rf.grid <- data.frame(.mtry = c(2, 3))

randomForest <- train(Fate ~ .,
                      data = training,
                      method = "rf",
                      ntree = 1000,
                      tuneLength = 5,
                      metric = "ROC",
                      tuneGrid = rf.grid,
                      trControl = ctrl)
```

Step 5: Let's Compare the Models

Classification Tree

```
pred_class <- predict(dtree, newdata=testdata, type="raw")
truth_class <- testdata$Fate
confusionMatrix(pred_class, truth_class)

          Confusion Matrix and Statistics

          Reference
Prediction Perished Survived
  Perished      100       17
  Survived        9       51

               Accuracy : 0.853

            Sensitivity : 0.917
            Specificity : 0.750
```

Random Forest

```
pred_class <- predict(randomForest, newdata=test, type="raw")
truth_class <- test$Fate
confusionMatrix(pred_class, truth_class)

          Confusion Matrix and Statistics

          Reference
Prediction Perished Survived
  Perished      101       12
  Survived        8       56

               Accuracy : 0.887

            Sensitivity : 0.927
            Specificity : 0.824
```

Supervised Learning Summary

1. Clean and explore the data
2. Come up with new features
3. Split data into training and test
4. Tune the model and parameters using cross-validation
5. Compare model results

Testing and Experimentation

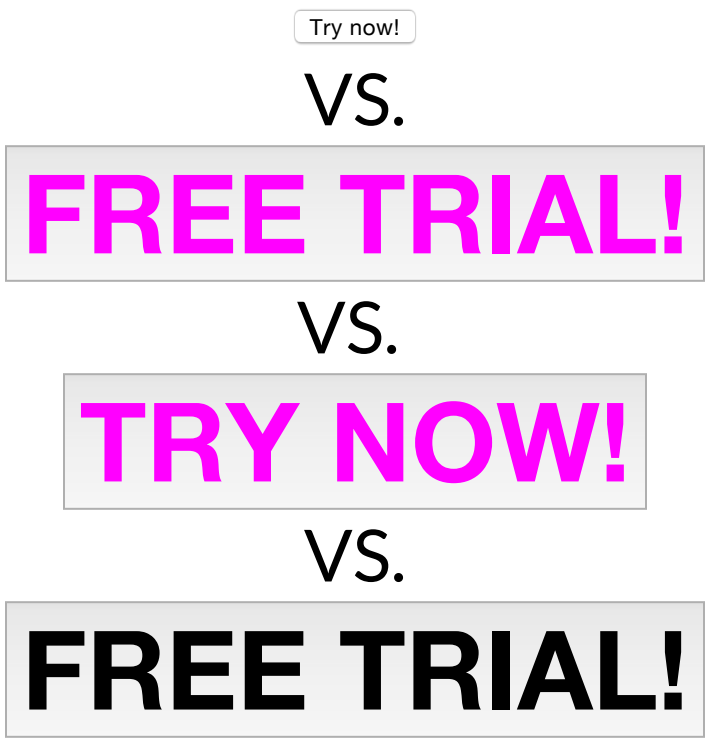
"[But] the best data-driven companies don't just passively store and analyze data, they actively generate actionable data by running experiments. The secret to getting value from data is testing, and if you're looking to grow your online business, implementing well-executed, consistent A/B testing is a necessity."

- Wyatt Jenkins, A/B Testing and the Benefits of an Experimentation Culture, Harvard Business Review

What is A/B testing?

Multi-Armed Bandit

Multivariate Testing



What NOT to do

No peeking!

Data Communication and Visualization

Beware of Model Complexity

If the model does not get used, it does not add value

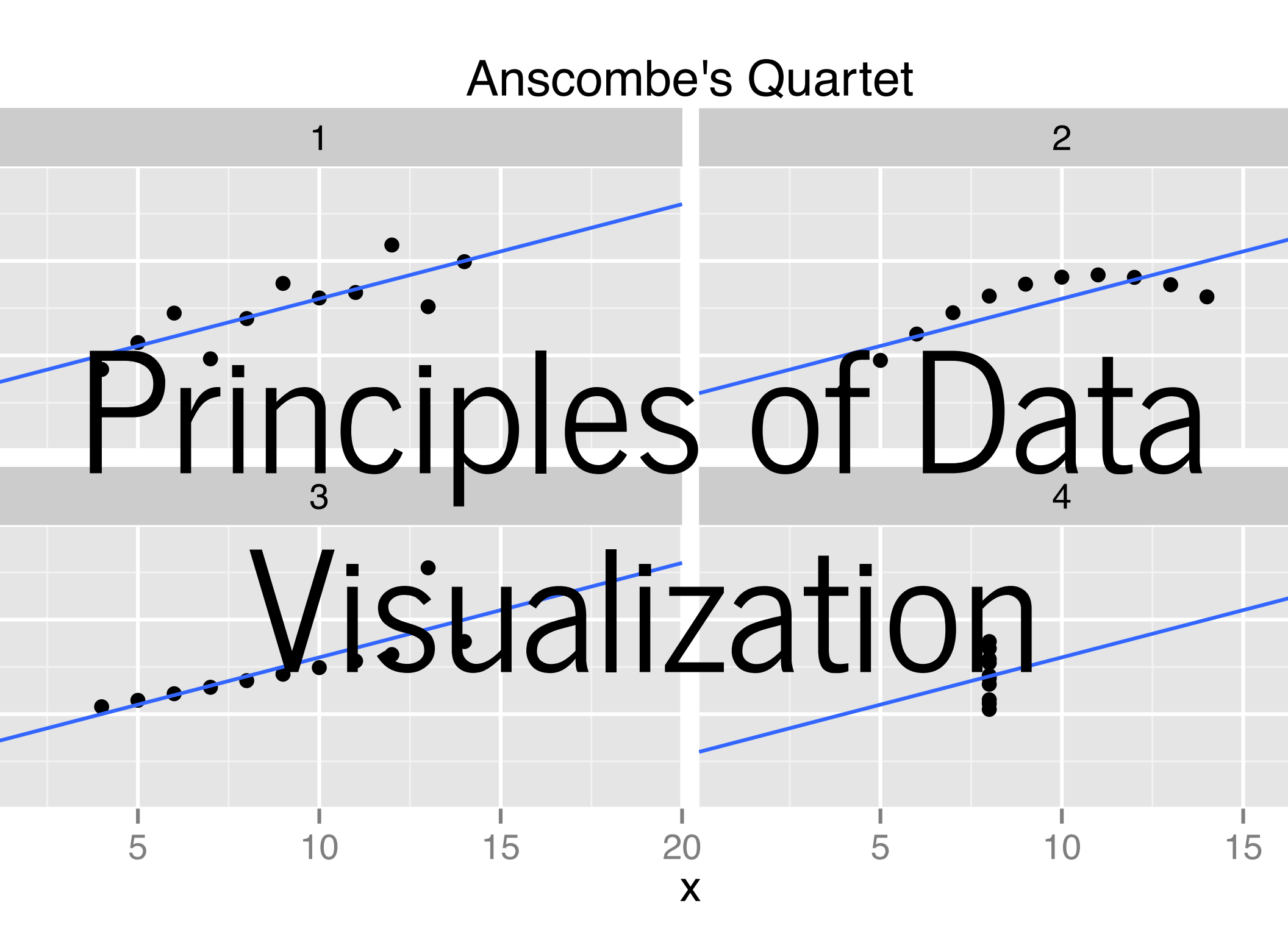
Translate Model Results into Business Value/ROI

Create Great Charts

Often the most effective way to describe, explore, and summarize a set of numbers - even a very large set - is to look at a picture of those numbers.

-Edward Tufte
The Visual Display of Quantitative Information

1		2		3		4	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

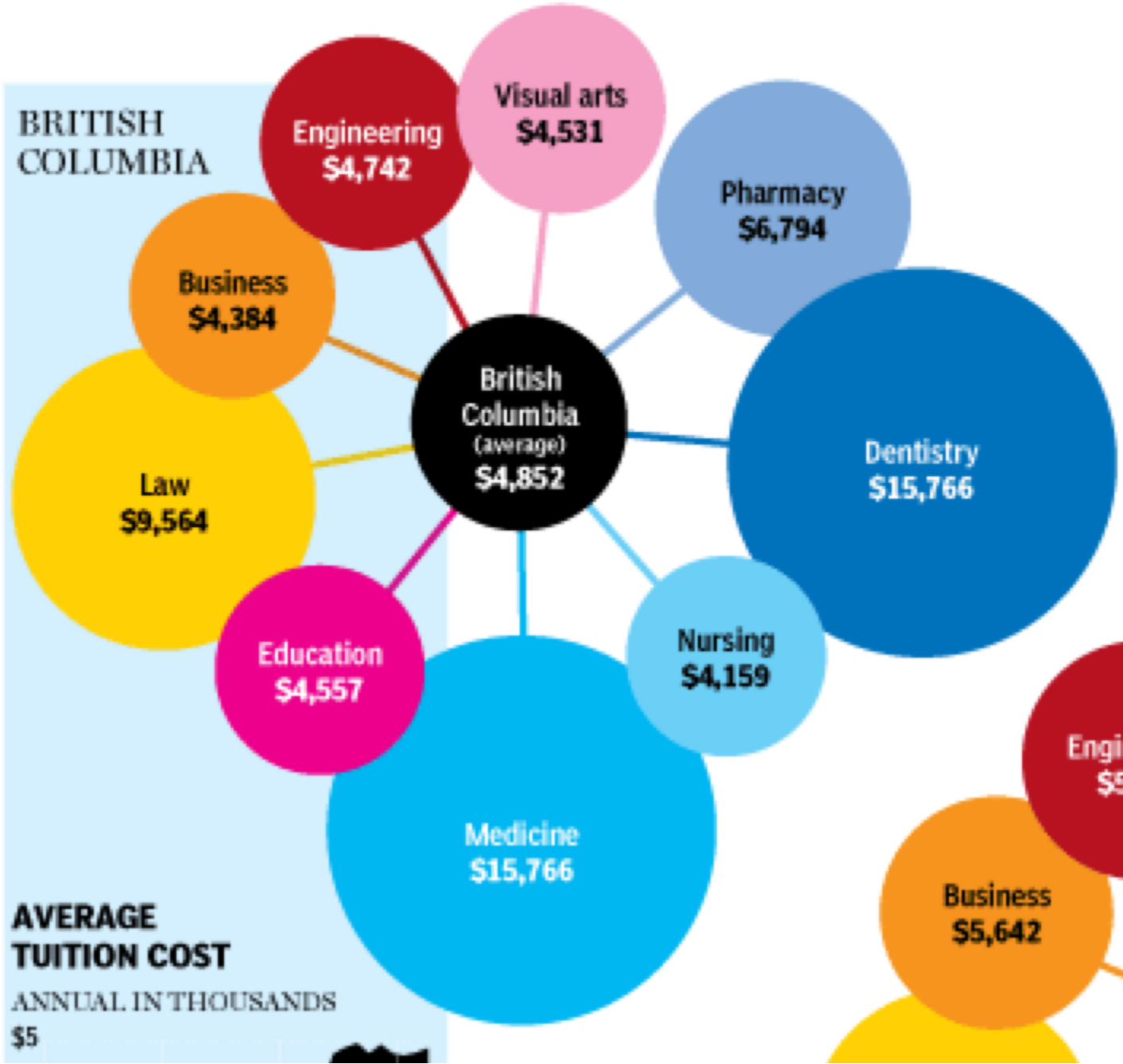


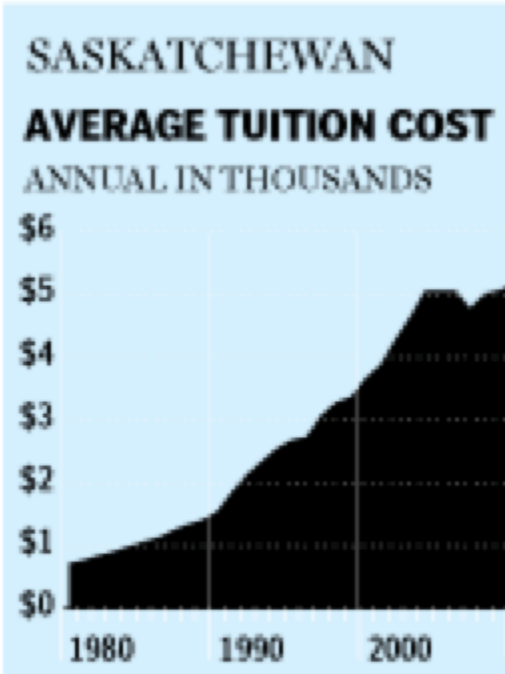
Principles of Data Visualization

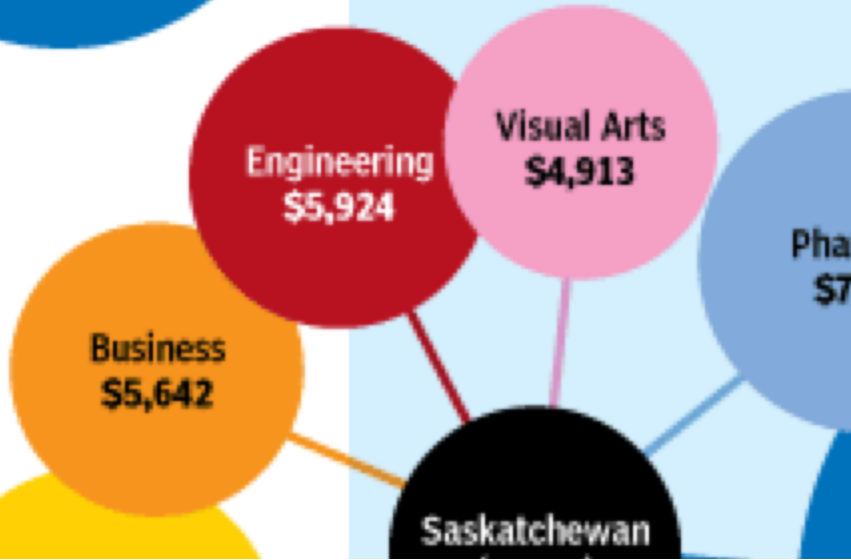
Above all else, show the data

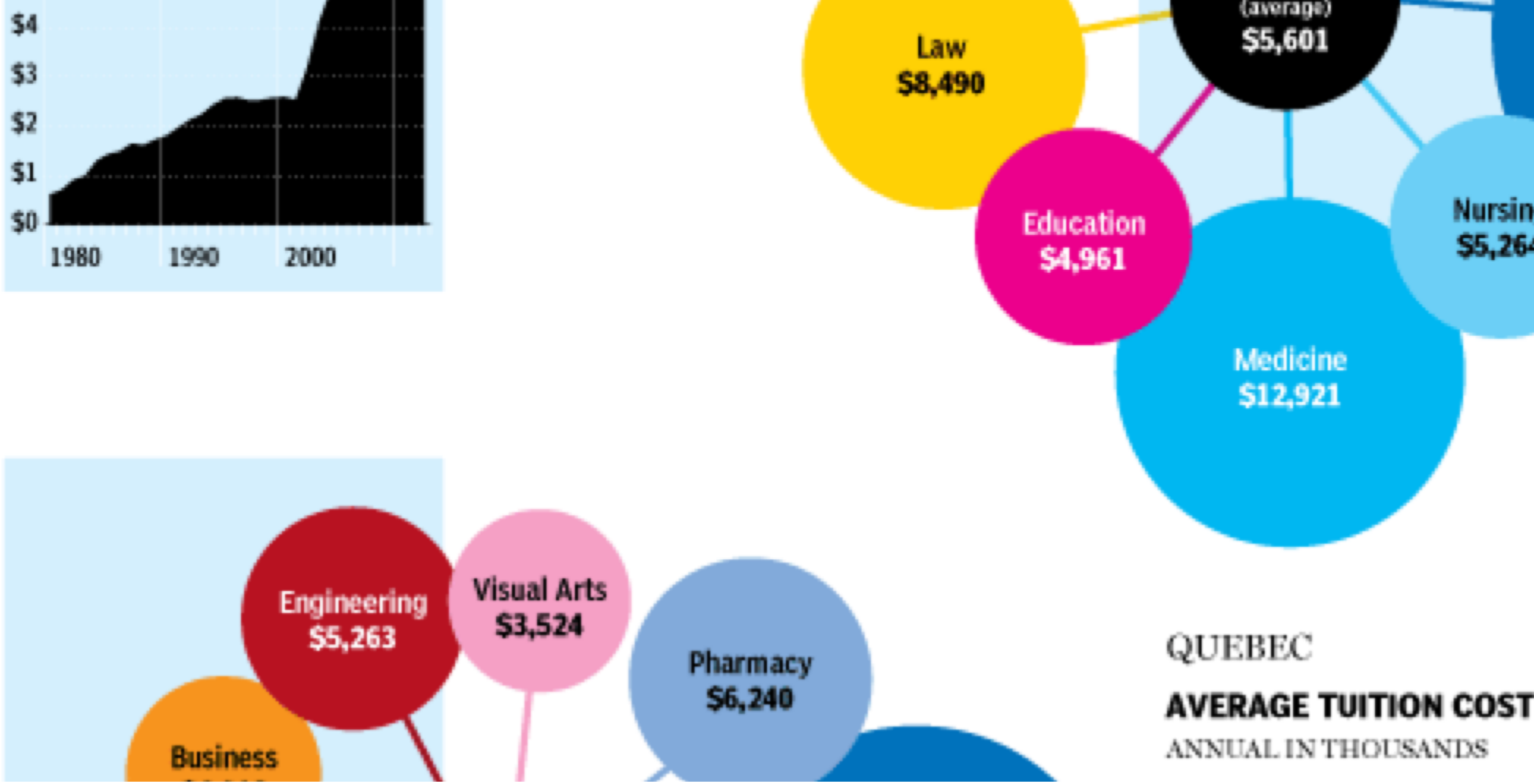
Draw attention to the data, not the visualization

AVERAGE ANNUAL TUITION COSTS BY PROVINCE BY AREA OF STUDY FOR 2011-2

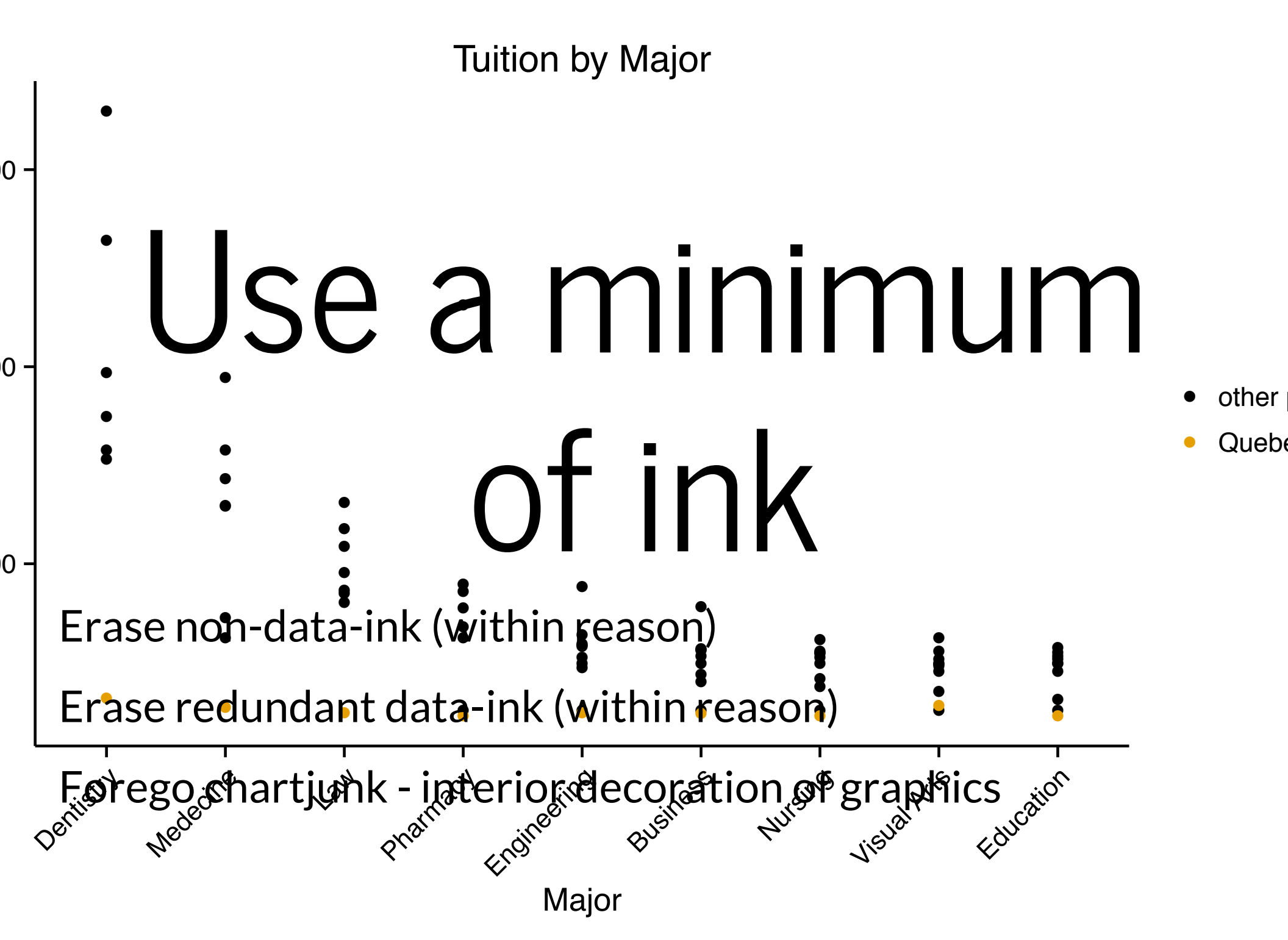


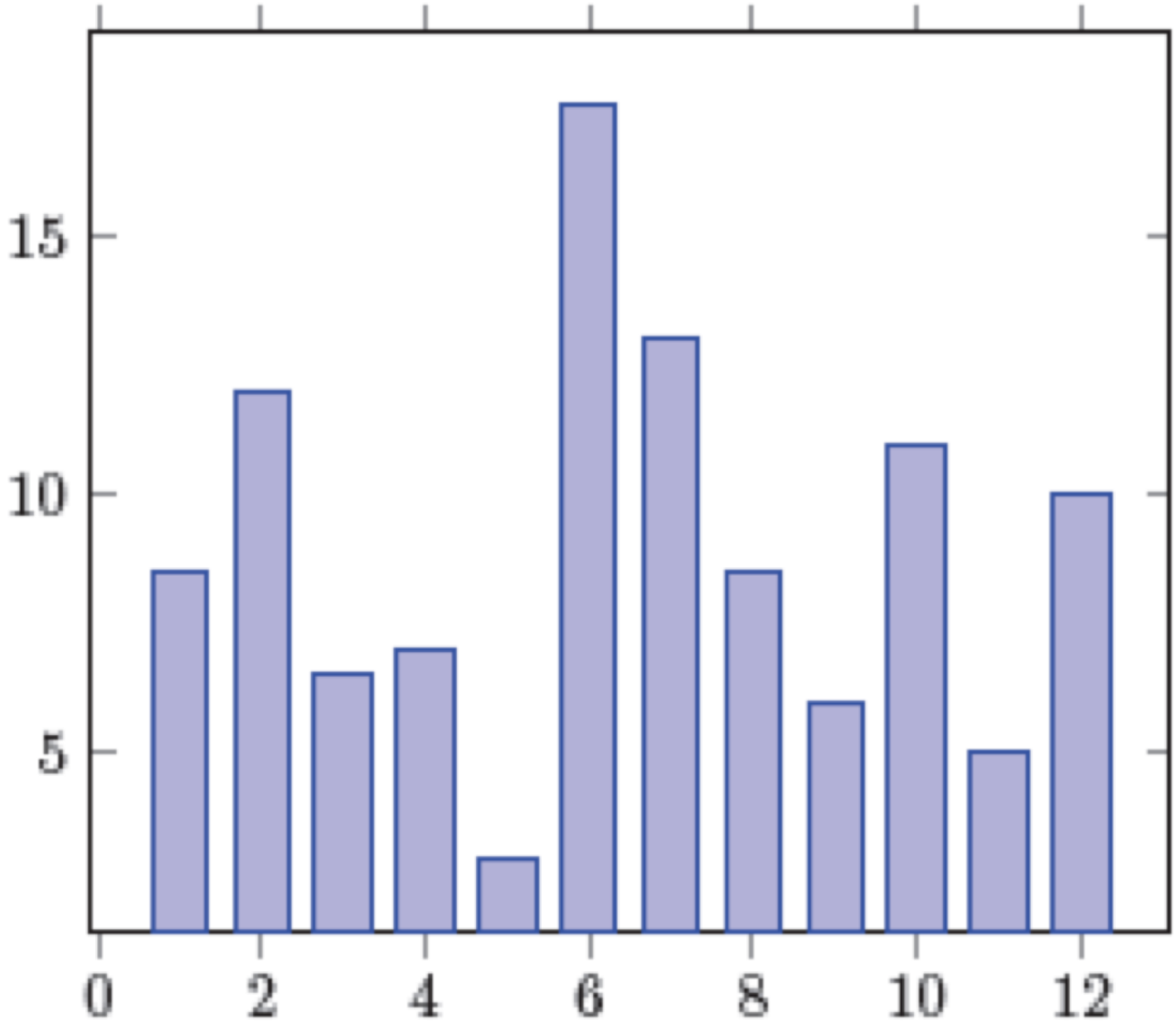


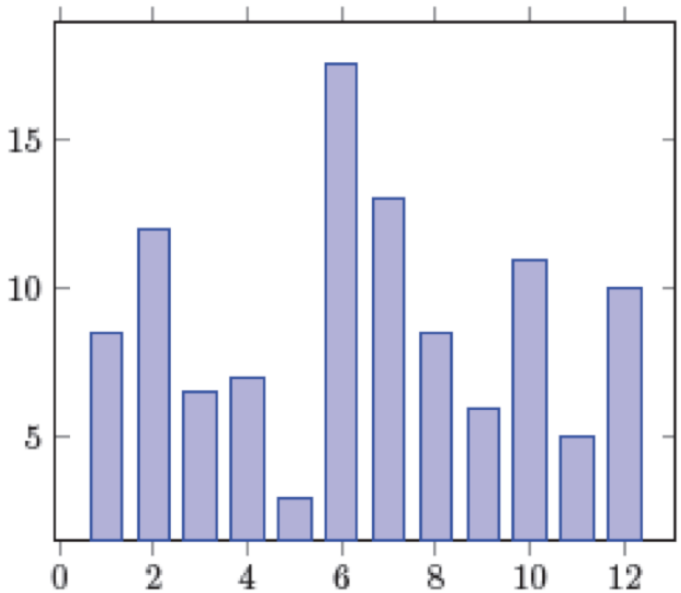


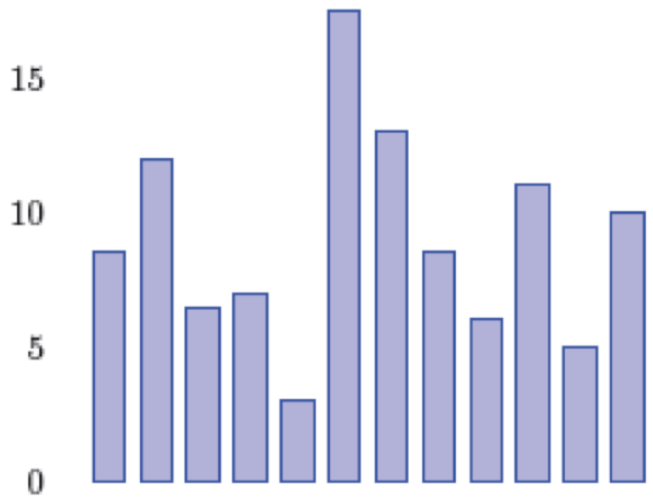


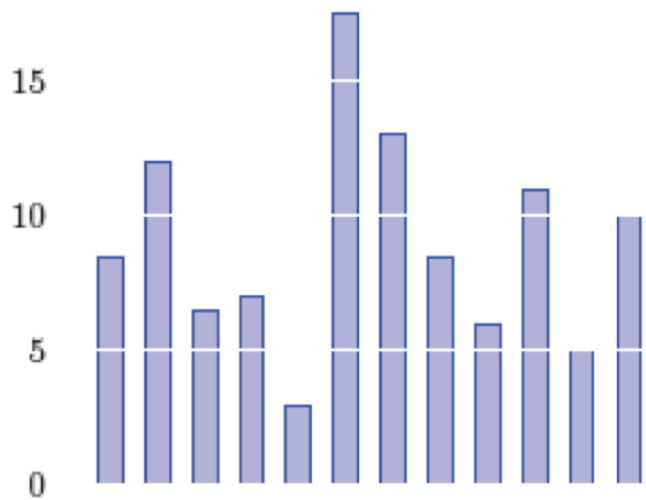
Source: http://news.nationalpost.com/2012/05/18/how-quebecs-tuition-price-tags-match-up-to-the-rest-of-canada-graphic/

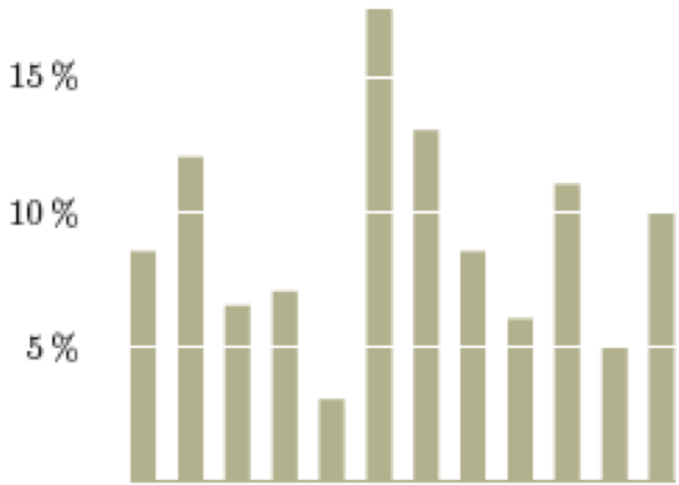

Tuition by Major
Use a minimum of ink
Erase non-data-ink (within reason)
Erase redundant data-ink (within reason)
Forego chartjunk - interior decoration of graphics
Major
other
Quebe

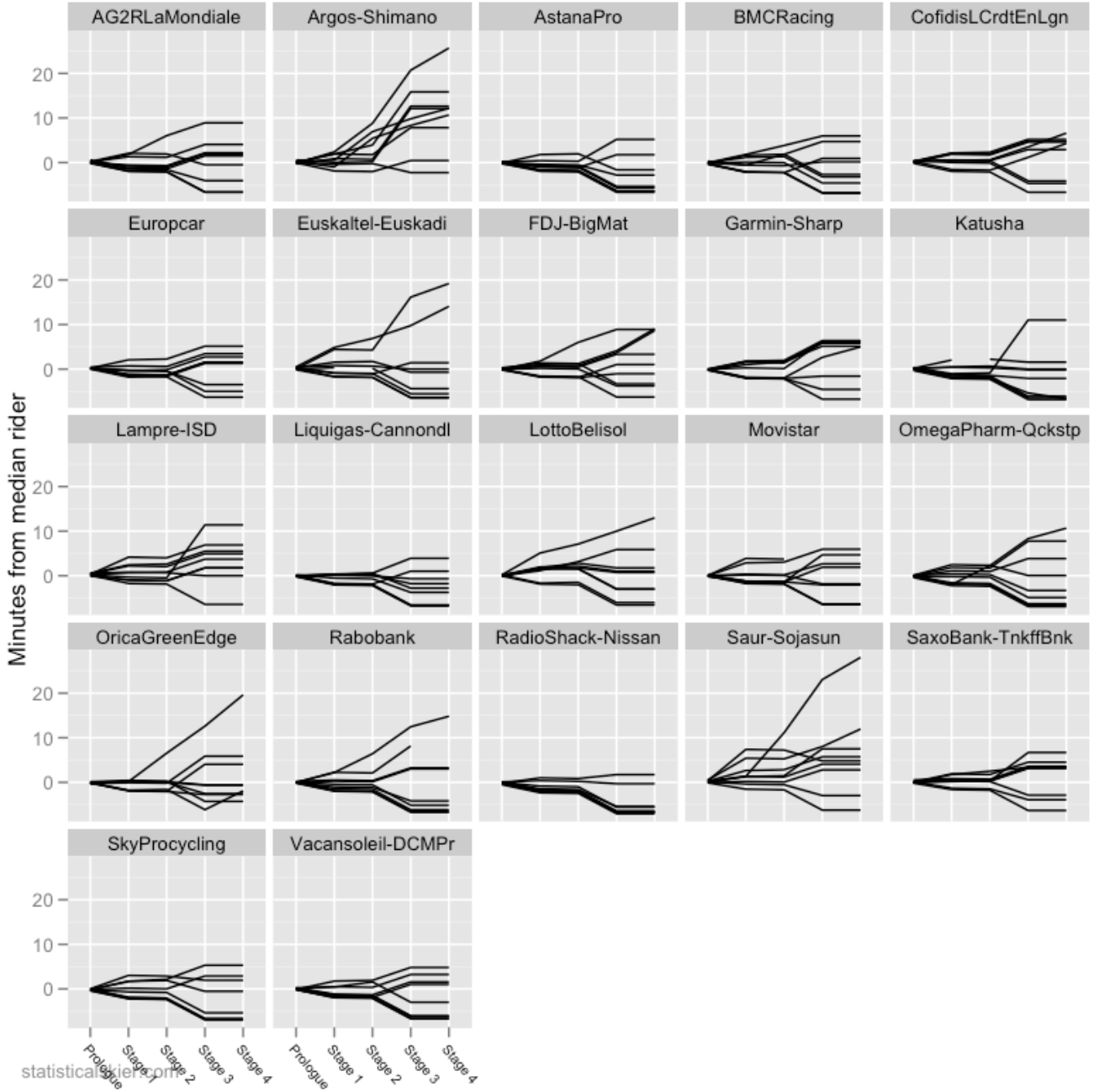

15
10
5
0
2
4
6
8
10
12


15
10
5
0
2
4
6
8
10
12

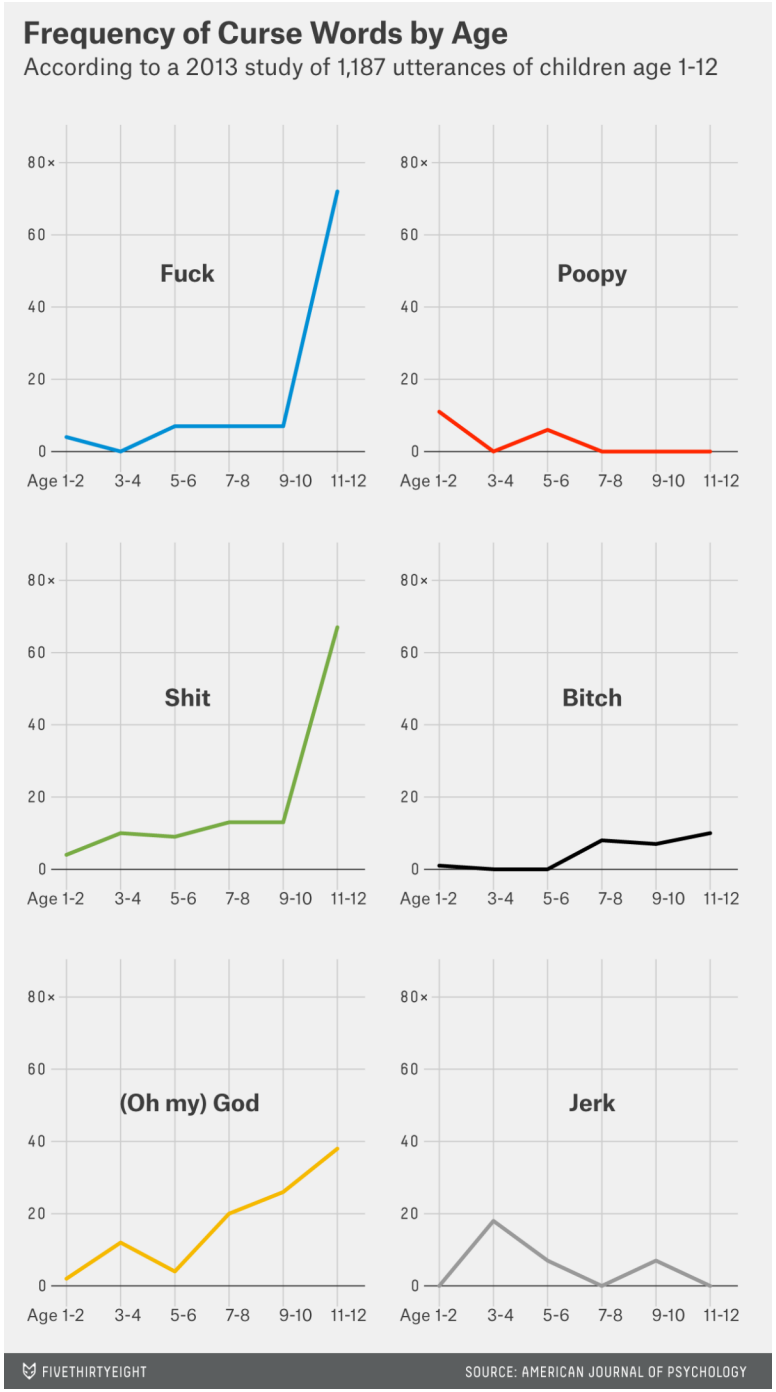

15
10
5
0


15
10
5
0


15 %
10 %
5 %


AG2RLaMondiale
Argos-Shimano
AstanaPro
BMCRacing
CofidisLCrdtEnLgn
Europcar
Euskaltel-Euskadi
FDJ-BigMat
Garmin-Sharp
Katusha
Lampre-ISD
Liquigas-Cannondl
LottoBelisol
Movistar
OmegaPharm-Qckstp
OricaGreenEdge
Rabobank
RadioShack-Nissan
Saur-Sojasun
SaxoBank-TnkffBnk
SkyProcycling
Vacansoleil-DCMPr
Minutes from median rider
20
10
0
Prologue
Stage 1
Stage 2
Stage 3
Stage 4
statisticalanalysis.com

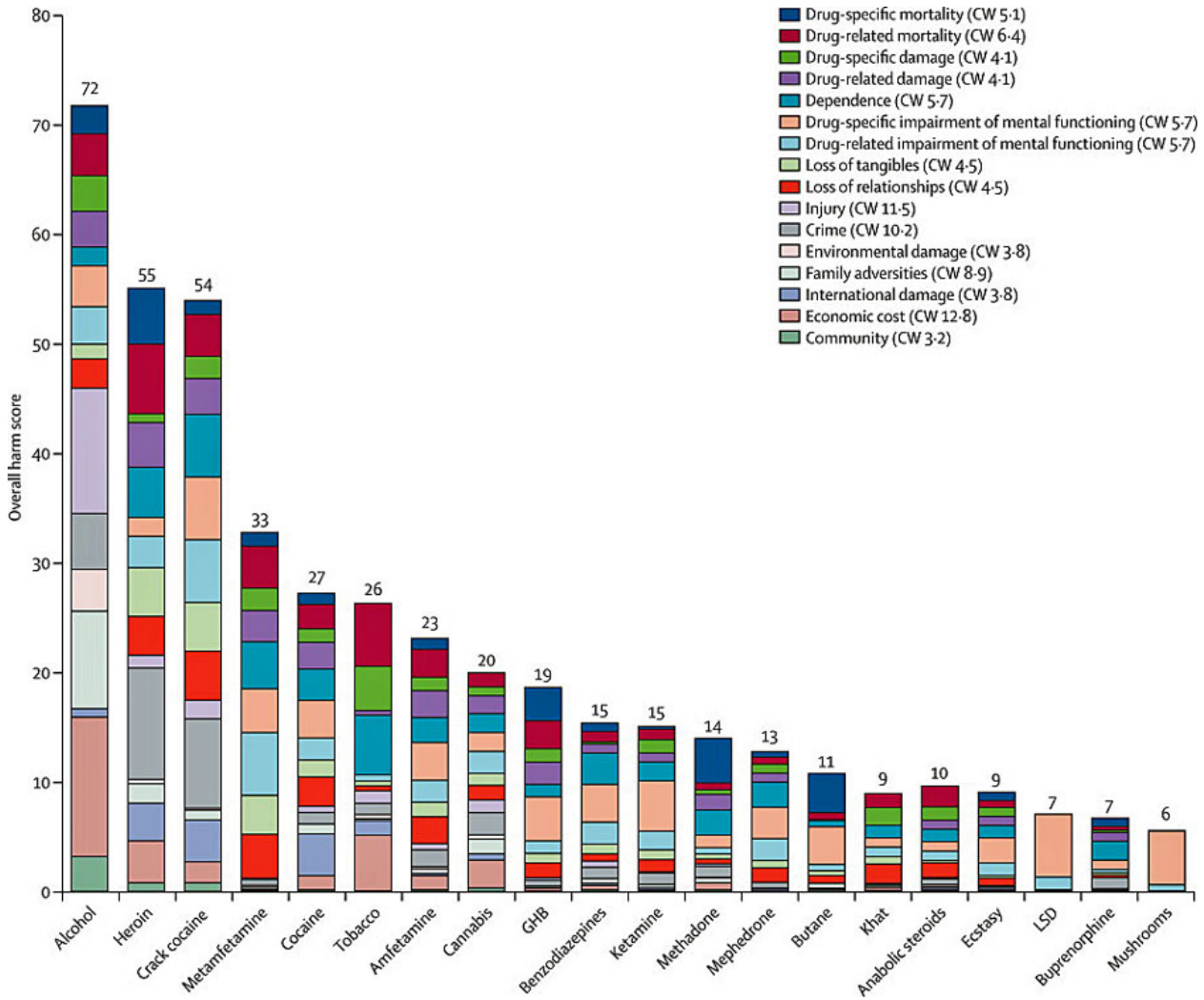
Source: http://www.statisticalskier.com/2012/07/tour-de-france-stage-4/



Source: http://fivethirtyeight.com/datalab/babys-first-profanity/

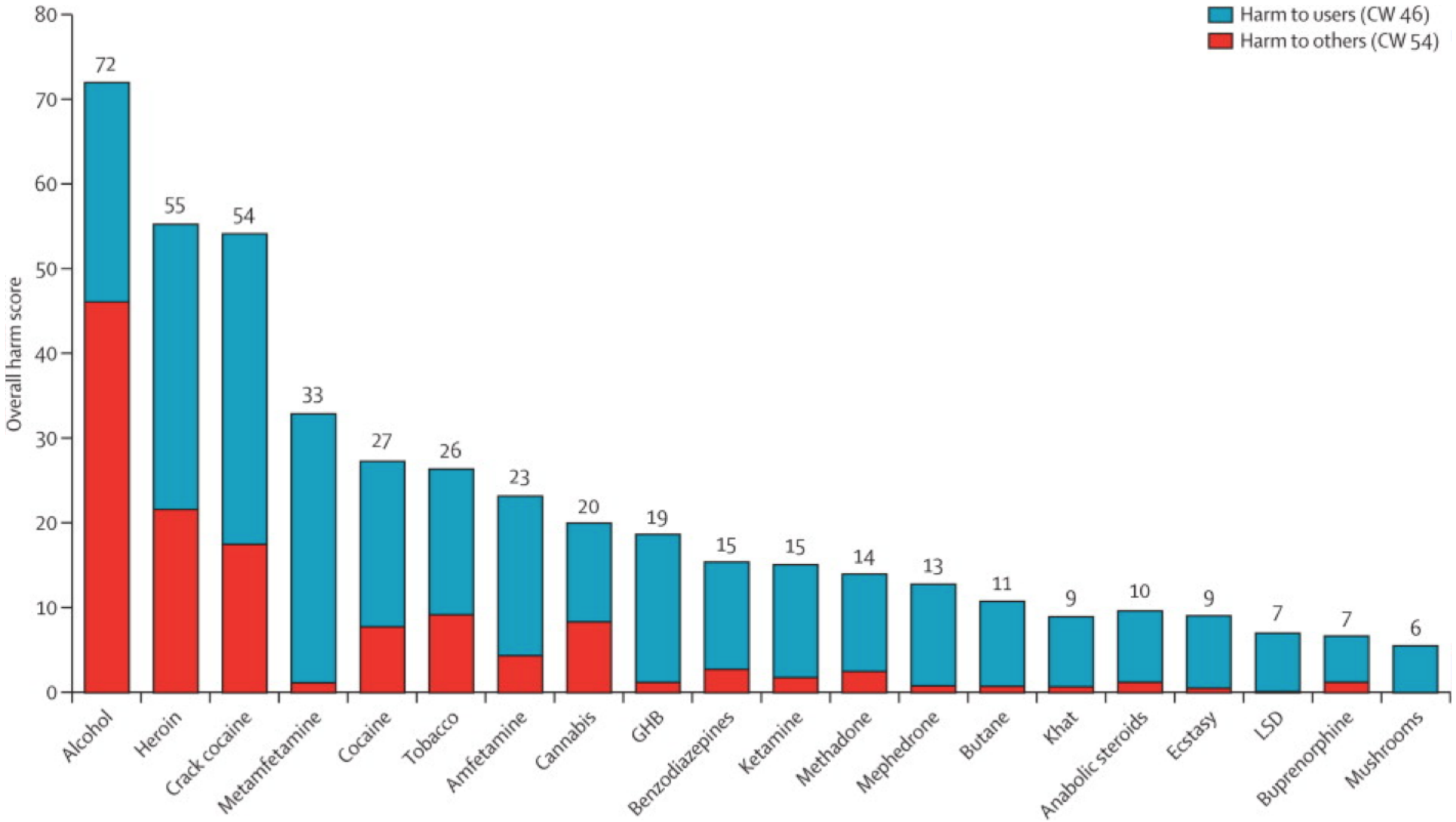
Avoid creating graphical puzzles

Why Alcohol is More Dangerous than Heroin



Source: http://www.businessinsider.com/alcohol-more-harmful-heroin-2012-7

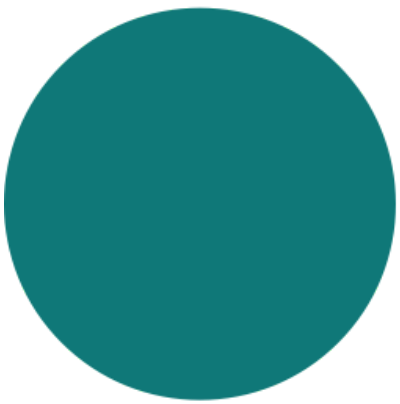
Why Alcohol is More Dangerous than Heroin

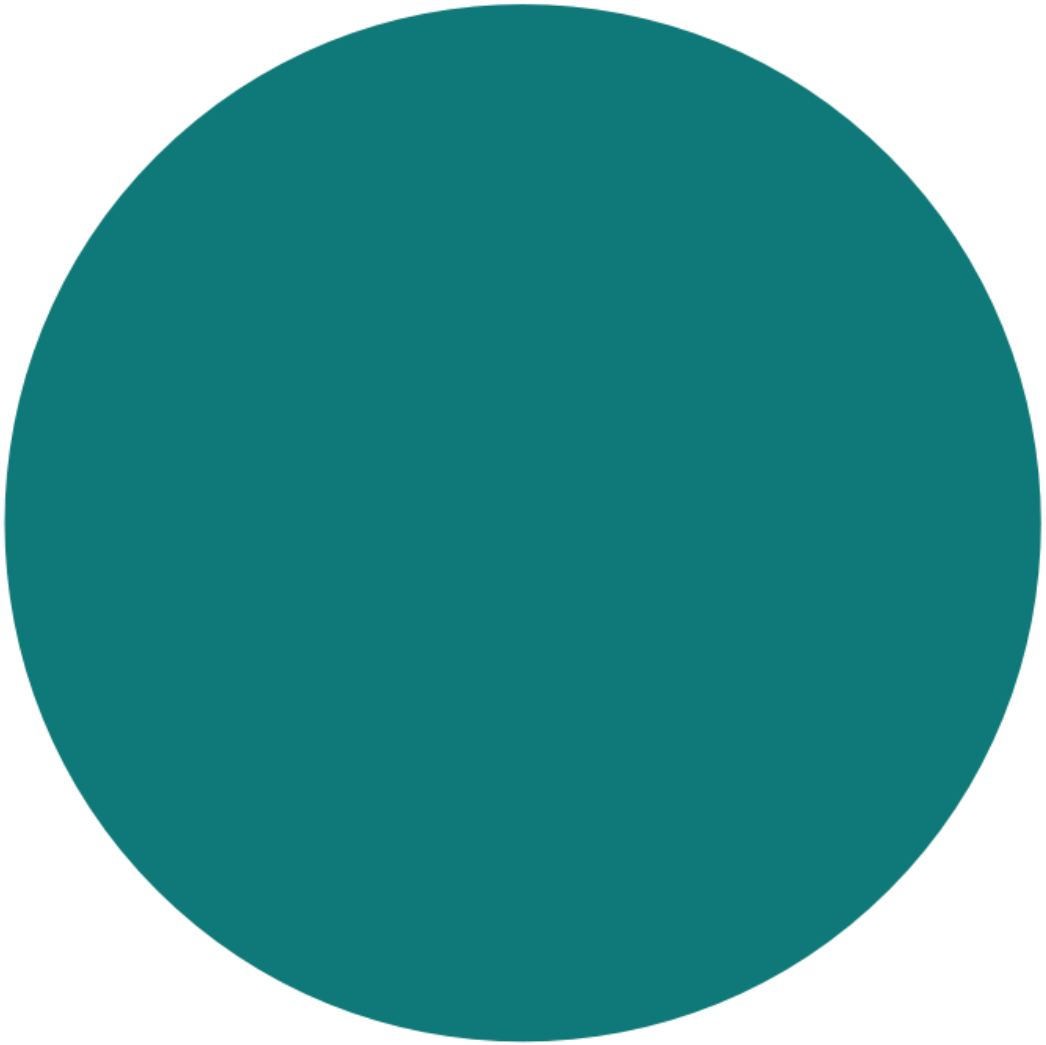


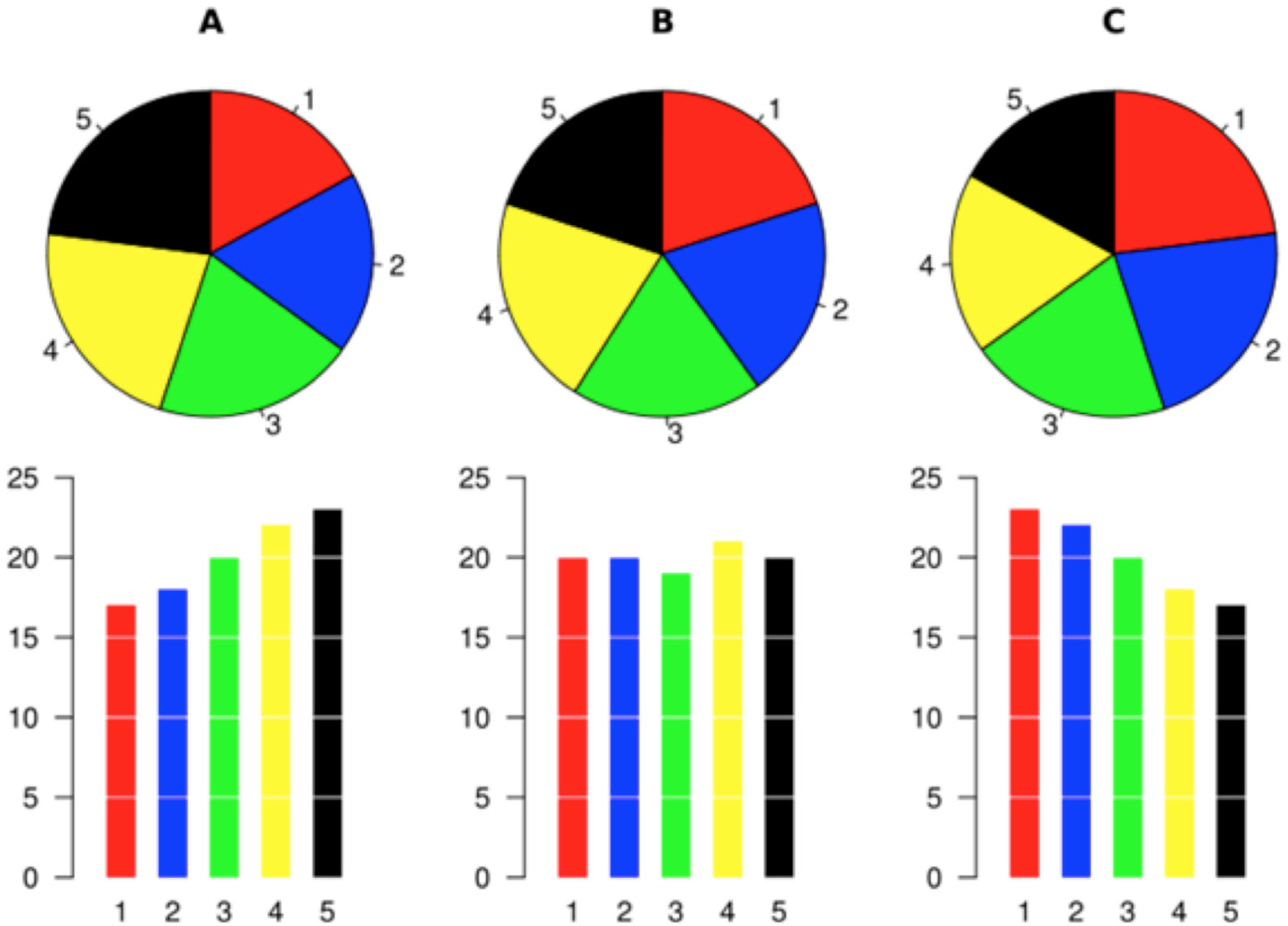
Source: http://fivethirtyeight.com/datalab/babys-first-profanity/

Do not distort the data

Visual representation should be consistent with the numerical representation

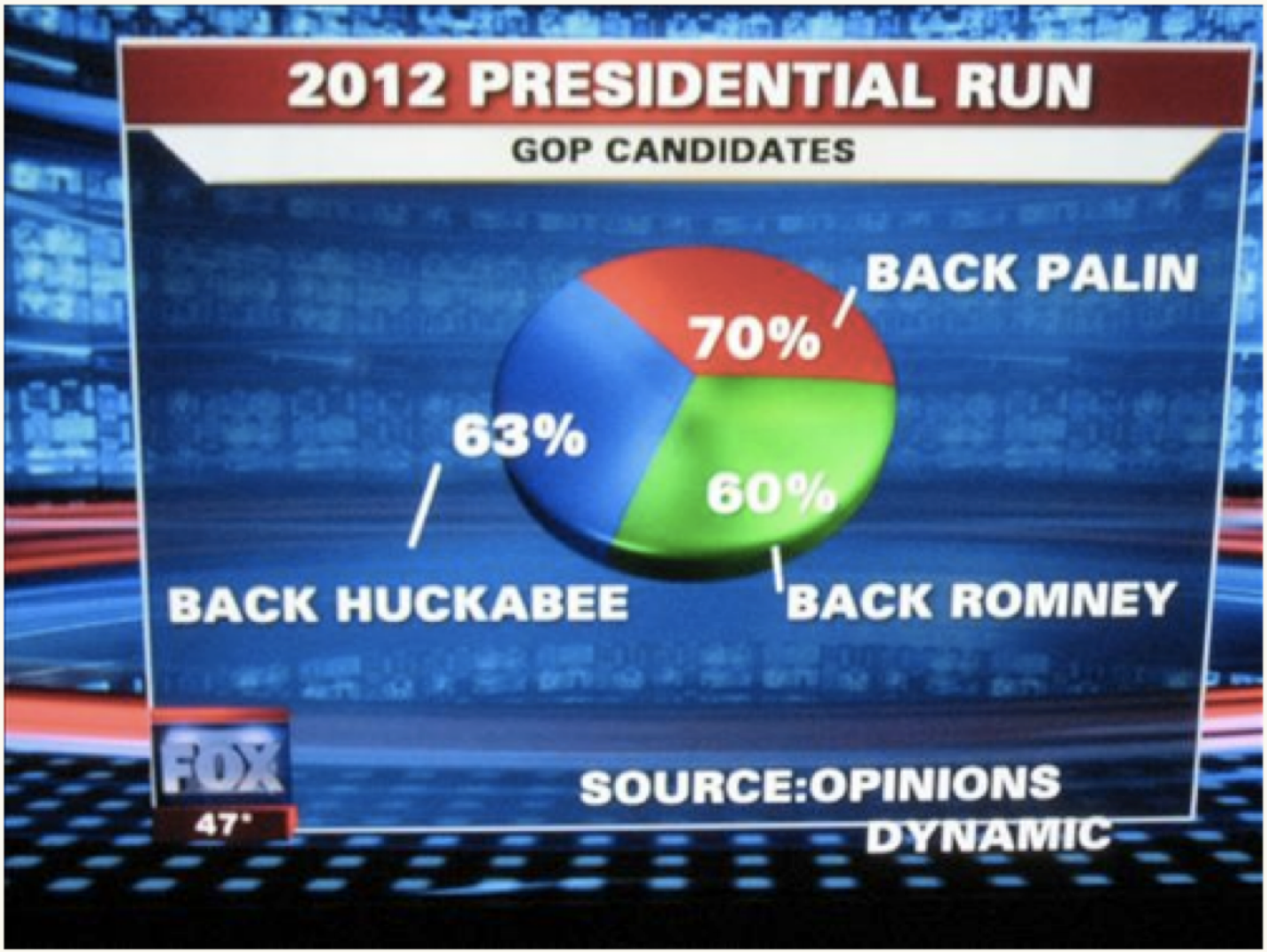






Source: https://blogs.oracle.com/experience/entry/pie_charts_just_dont_work_when_comparing_data_-_number_10_of_top_10_reasons_to_never_ever_use_a_pie

November 25th, 2009



Source: http://chartjunk.karmanaut.com/?p=45

Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.

-Edward Tufte

Source: https://c2.staticflickr.com/2/1162/1415120191_2aef20cb08_b.jpg

Resources for becoming a data scientist

- Programming
  Intro to Python GDI class, SQL GDI Class
- Statistics / machine learning / math
  MOOCs, Coursera
- Domain knowledge
  Think, talk to people, be proactive, learn as you go
- Kaggle

Happy Data-ing!