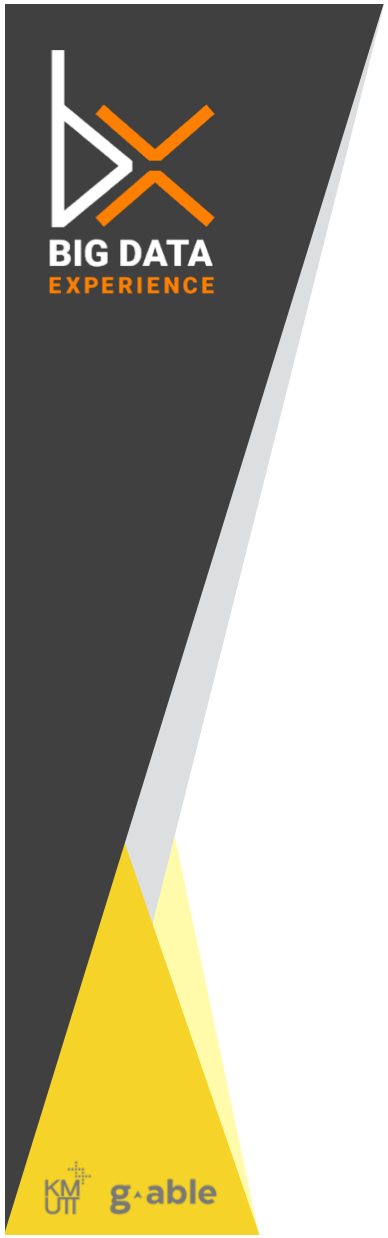


Practical R for Data Science
Workshop

28 February 2018, 8.00 – 12.00

Asst. Prof. Dr. Santitham Prom-on

Department of Computer Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi

"DATA IS THE NEW OIL."

Coined in 2006 by Clive Humby, a British data commercialization entrepreneur, this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered **data to be an economic asset,** like oil.

From the beginning of recorded time until 2003, we created **5 exabytes** (5 billion gigabytes) **of data.**

In 2011 the same amount was created every two days.

By 2013, it's expected that the time will shrink to 10 minutes.

There are nearly as many bits of information in the digital universe as there are stars in our actual universe.

As of August 2012, there were just over **4 million** articles in the English Wikipedia.

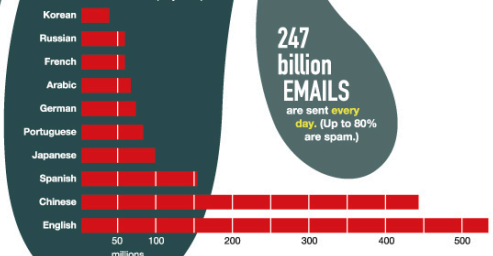
There are **133 million BLOGS** on the web.

80% of all humans own a mobile phone of some sort. Out of 5 billion mobiles, 1 billion are smartphones. (In Singapore, 54% of citizens are smartphone users.)

247 billion EMAILS are sent every day. (Up to 80% are spam.)

English is the dominant language of the web. But by 2014 it will be **Chinese,** if its current rate of increase continues.

Top languages used on the web (May 2011):



Every hour, we create enough Internet traffic to fill **7 billion DVDs.** Side by side, that's that's seven times the height of Everest.

Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, **high-frequency traders,** with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.

These specialized algorithms make split-second decisions to buy or sell a commodity. New cable being laid under the Atlantic will shave **5 milliseconds** from the current 65 milliseconds it takes for trading instructions to travel between New York City and London.

With new fiber-optic cable, the round-trip time between New York and London will be 59.6 milliseconds.

This 5-millisecond saving is worth many millions of dollars to the trading firms who use the cable (and who will pay millions to do so).

60% of all humans (5.4 billion people) are active texters. In 2010, 193,000 text messages were sent every second.

10% of all photos ever taken were taken in 2011.

50% of 5-year-old kids in the U.S. are given access to a smartphone.

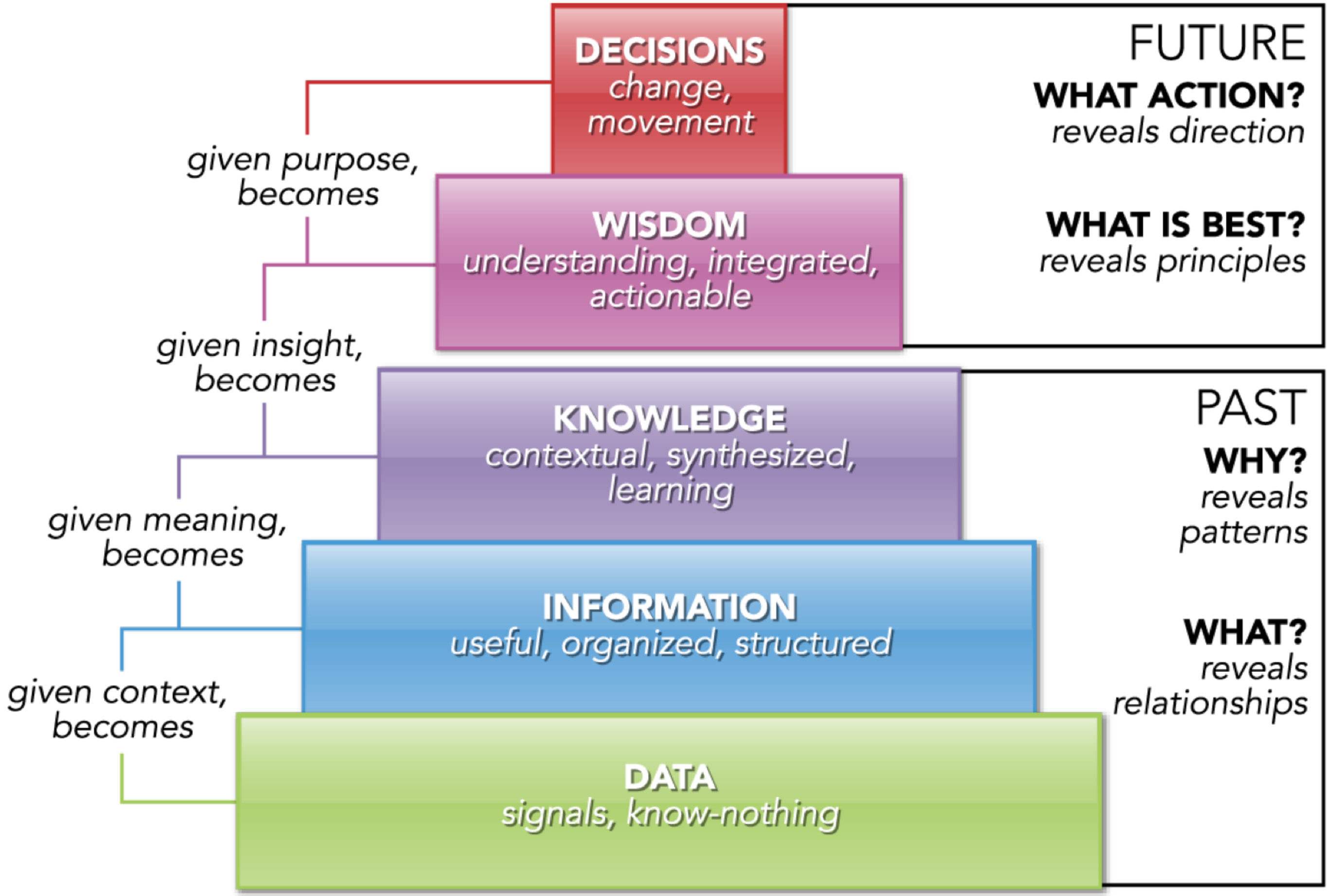
How they save 5 milliseconds

The depth of the Atlantic Ocean varies.

The new cable will lie on areas of the ocean floor that are up to 1,000 feet shallower than the current fastest cable. By taking a different route, the new cable is shorter, meaning that the time it takes for messages to travel along it is shortened.

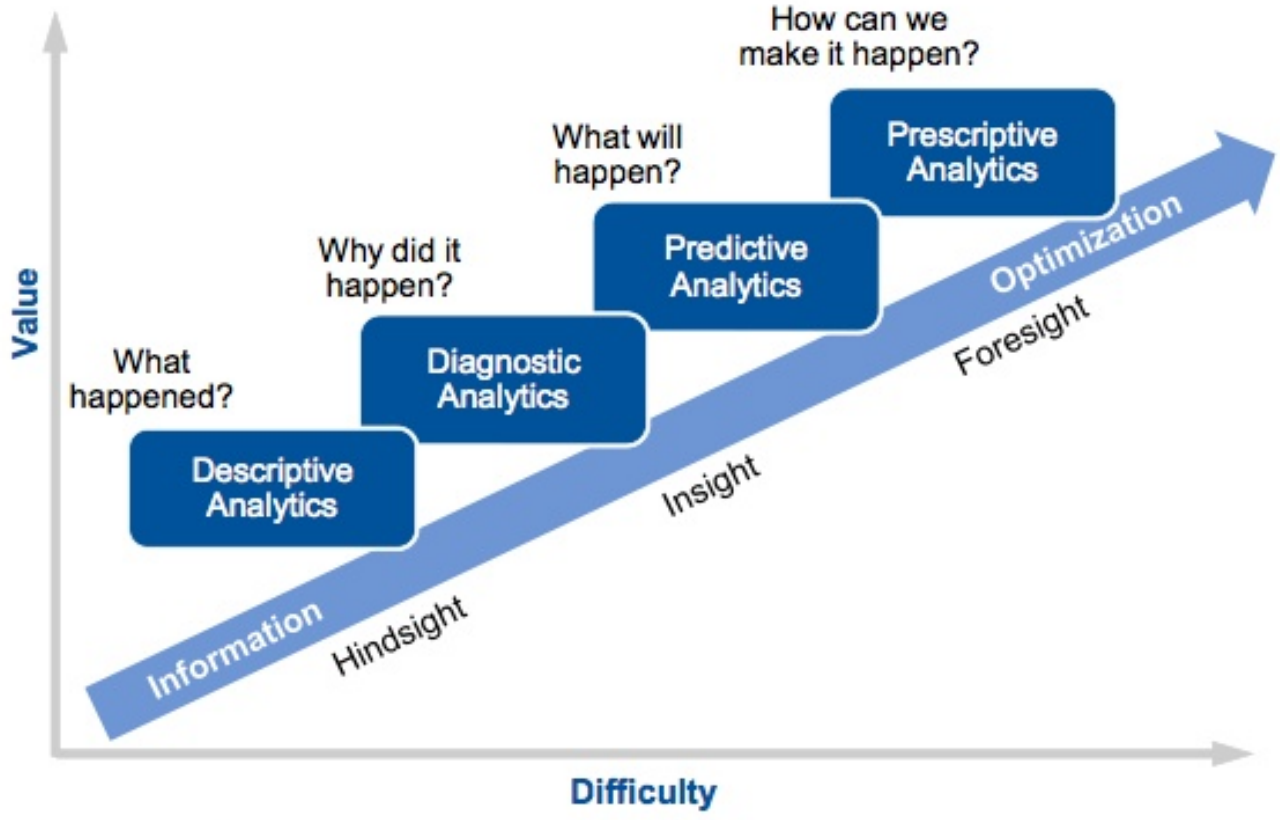
Information Graphics by Nigel Holmes

DIKW (D) Pyramid



BIG DATA EXPERIENCE

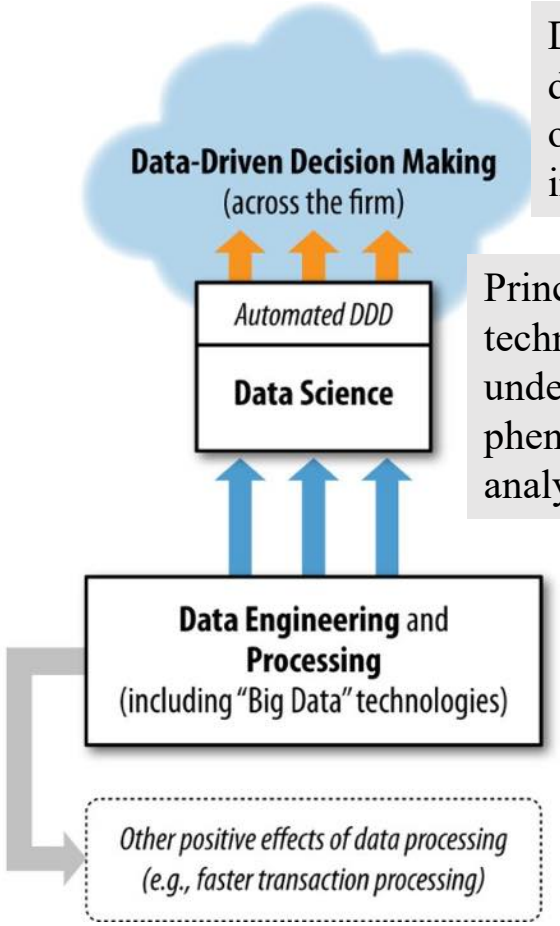
From descriptive … to prescriptive



Source: Gartner (March 2012)

KM UTT g·able

Data Engineering and Data Science



DDD = practice of basing decision on the analysis of data, rather than intuition

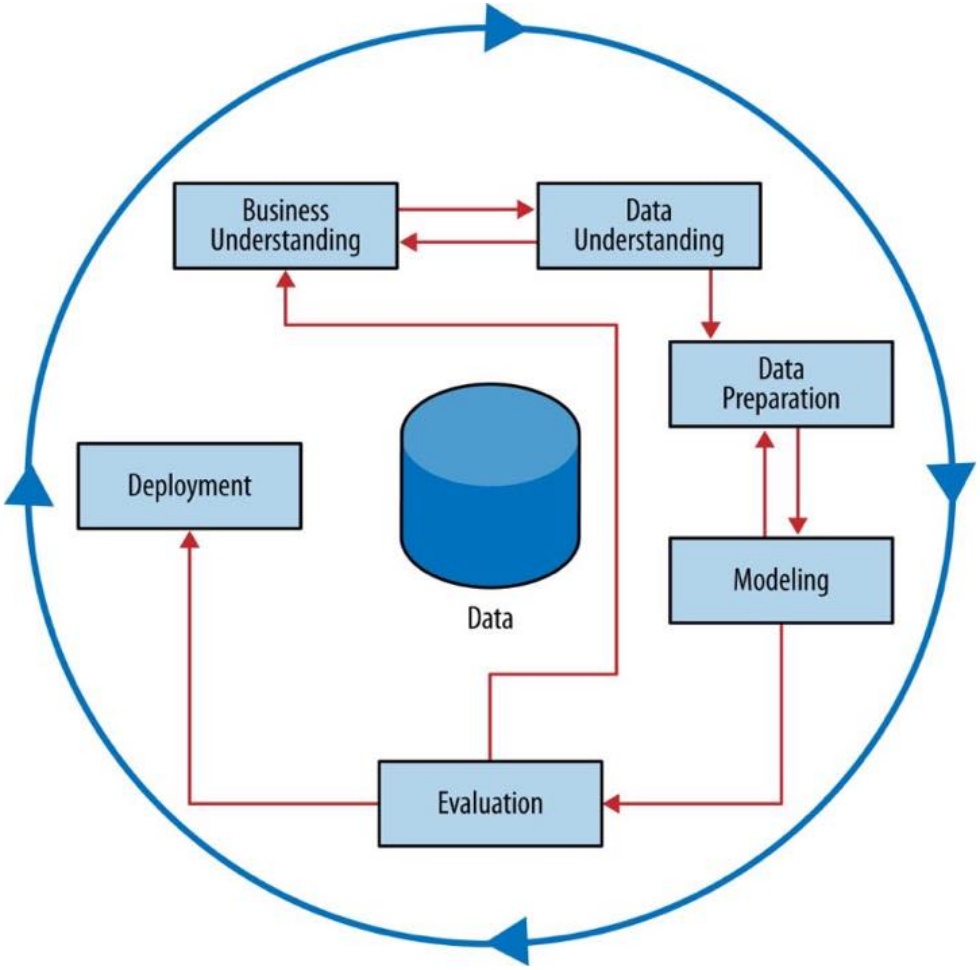
Principles and techniques for understanding phenomena via the analysis of data.

Accessing and processing of massive-scale data flexibly and efficiently with Big Data technologies

BIG DATA EXPERIENCE

Data science is an iterative process

The entire process is an exploration of the data over and over



KMUTT g-able

Data Science Process

Service data

Monitoring data

Operation data

Data Preprocessing

Cleaning
Tokenization
Joining
Aggregating
etc.

PCA
Factor Analysis
Distribution
etc.

Exploratory Data Analysis

Visualize Insight
Communicate Finding

Recommendation
Make decision

Data Modeling

Fault Tree Analysis
Markov Model
Simulation
etc.

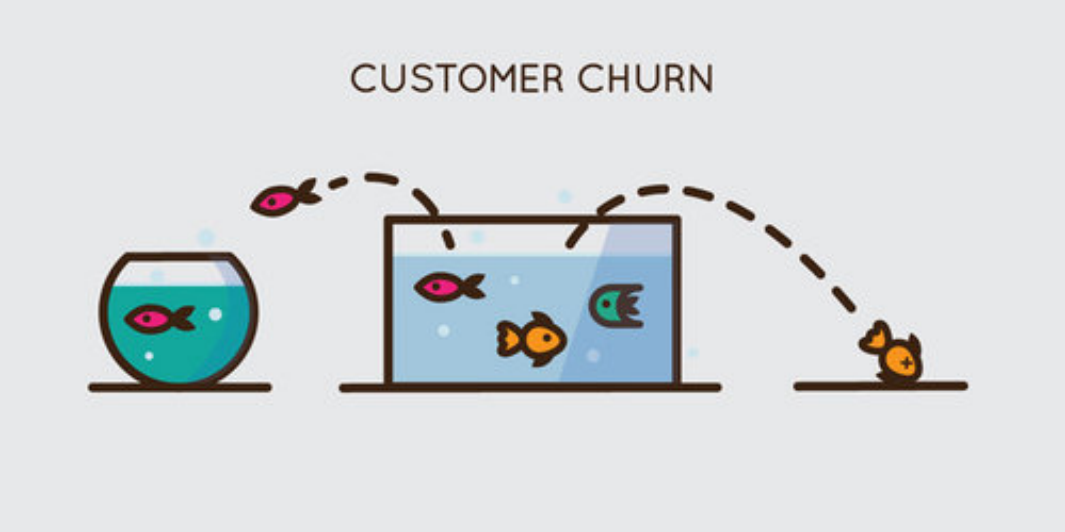
Data Services

Fault Prediction
Alert and Classification
Optimization
Planning
etc.

BIG DATA EXPERIENCE

Churn

- Churn is an event of individuals or items moving out of subscriptions or regular usage.



KMUTT g-able

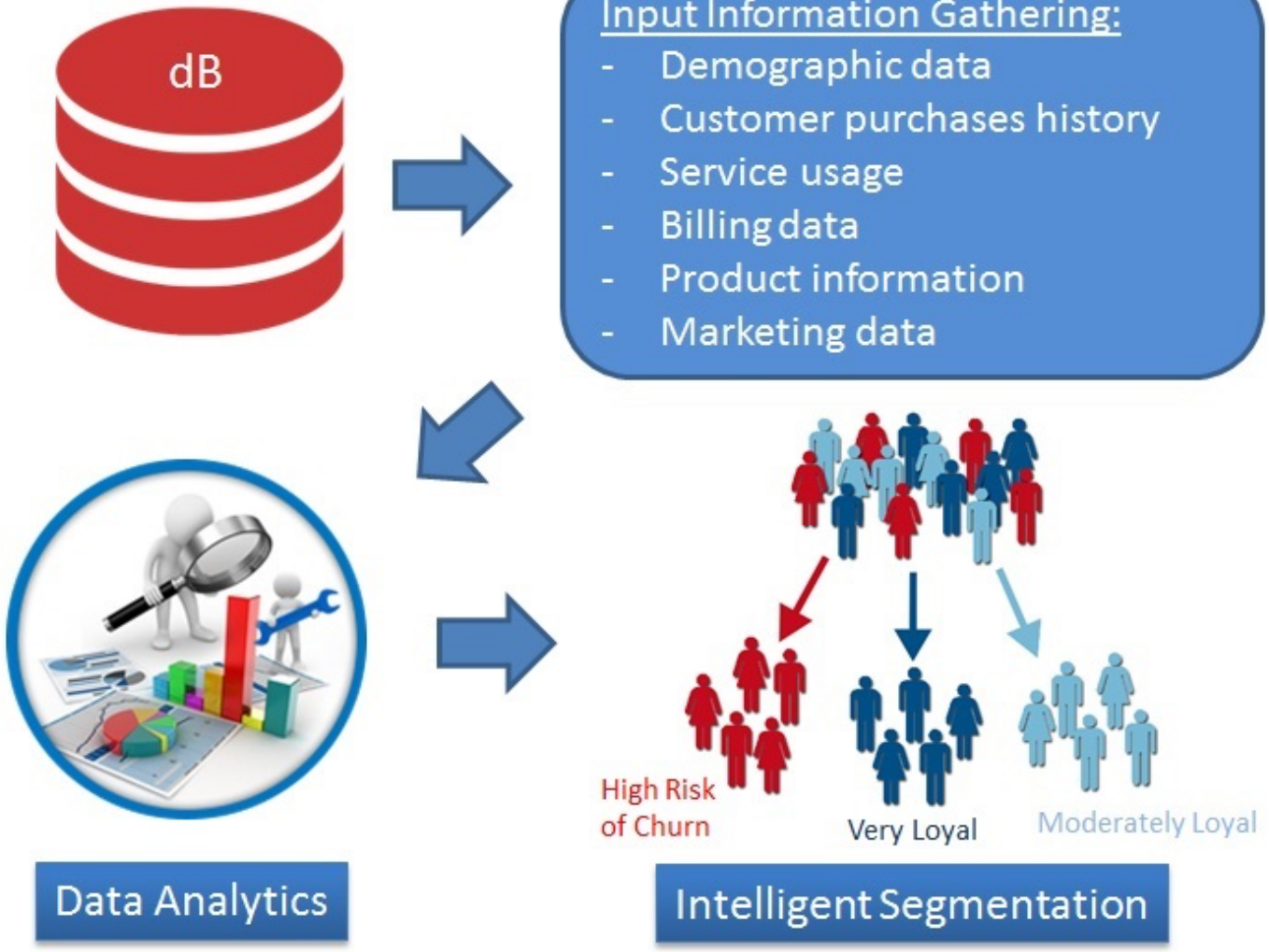
BIG DATA EXPERIENCE

Dealing with churn

- Why?
  - Cost of Acquiring New Clients >> Keep Client
- Predict churn
  - One model for performance (short term)
    - Special offer
  - One model for understanding (long term)
    - Fix the process

KMUTT g-able

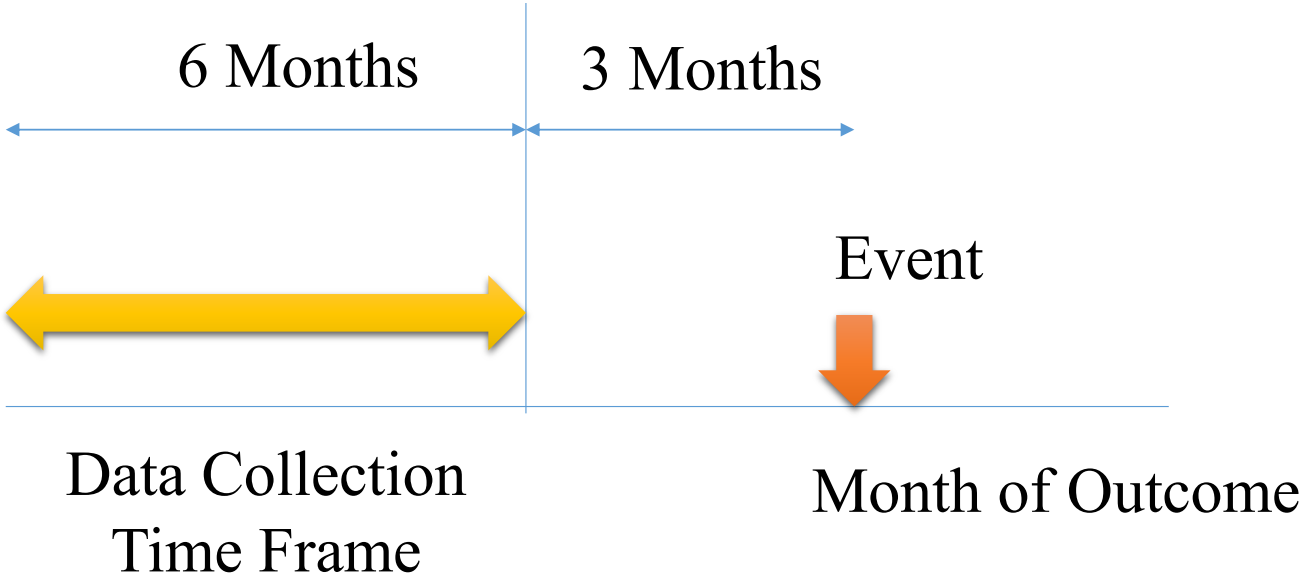
BIG DATA
EXPERIENCE



KMUTT g-able

BIG DATA EXPERIENCE

Acquiring data

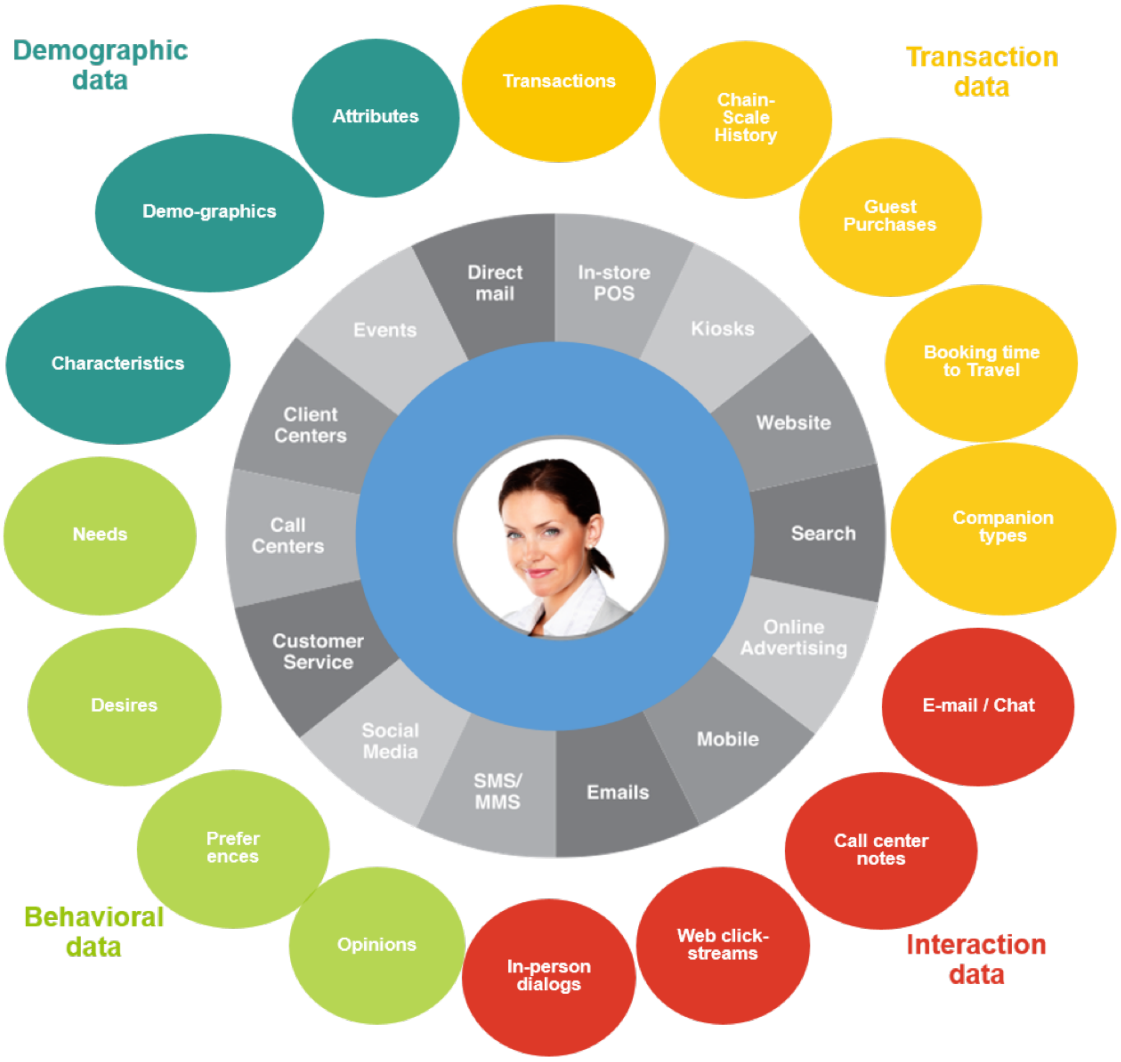


KM UTT g·able

BIG DATA EXPERIENCE

Training data

- Training data contains
  - Label
  - Attribute data
    - Demographics
    - Transactions
    - Interactions
    - Behavioral



BIG DATA EXPERIENCE

Load Data

```
churnData <- read.csv("Telco-Churn.csv")
head(churnData)
```

```
  customerID gender SeniorCitizen Partner Dependents tenure
1 7590-VHVEG Female             0     Yes         No      1
2 5575-GNVDE   Male             0      No         No     34
3 3668-QPYBK   Male             0      No         No      2
4 7795-CFOCW   Male             0      No         No     45
5 9237-HQITU Female             0      No         No      2
6 9305-CDSKC Female             0      No         No      8
  PhoneService    MultipleLines InternetService OnlineSecurity
1           No No phone service             DSL             No
2          Yes               No             DSL            Yes
3          Yes               No             DSL            Yes
4           No No phone service             DSL            Yes
5          Yes               No     Fiber optic             No
6          Yes              Yes     Fiber optic             No
```

KMUTT g-able

BIG DATA EXPERIENCE

View general distribution

```
summary(churnData)
```

```
                     PaymentMethod  MonthlyCharges    TotalCharges
Bank transfer (automatic):1544   Min.   : 18.25   Min.   :  18.8
Credit card (automatic)  :1522   1st Qu.: 35.50   1st Qu.: 401.4
Electronic check         :2365   Median : 70.35   Median :1397.5
Mailed check             :1612   Mean   : 64.76   Mean   :2283.3
                                 3rd Qu.: 89.85   3rd Qu.:3794.7
                                 Max.   :118.75   Max.   :8684.8
                                                  NA's   :11

Churn
No :5174
Yes:1869
```

KMUTT g-able

BIG DATA EXPERIENCE

Cleansing
NA

- Why are there NAs?
- What to do with NAs?

```
library(tidyr)
library(dplyr)
churnData %>% filter(is.na(TotalCharges))
```

KMUTT g-able

BIG DATA EXPERIENCE

Cleansing NA

- Since NAs here means customers have not been charged, we will replace them with zero

```
churnData[is.na(churnData$TotalCharges),] <- 0
```

KMUTT g-able

BIG DATA EXPERIENCE

Cleansing
Misrepresentation & Remove IDs

- Column SeniorCitizen should be converted into a factor

```
churnData %>%
   mutate(SeniorCitizen = factor(
    ifelse(SeniorCitizen == 1,
           "Yes", "No"))) -> churnData.1
```

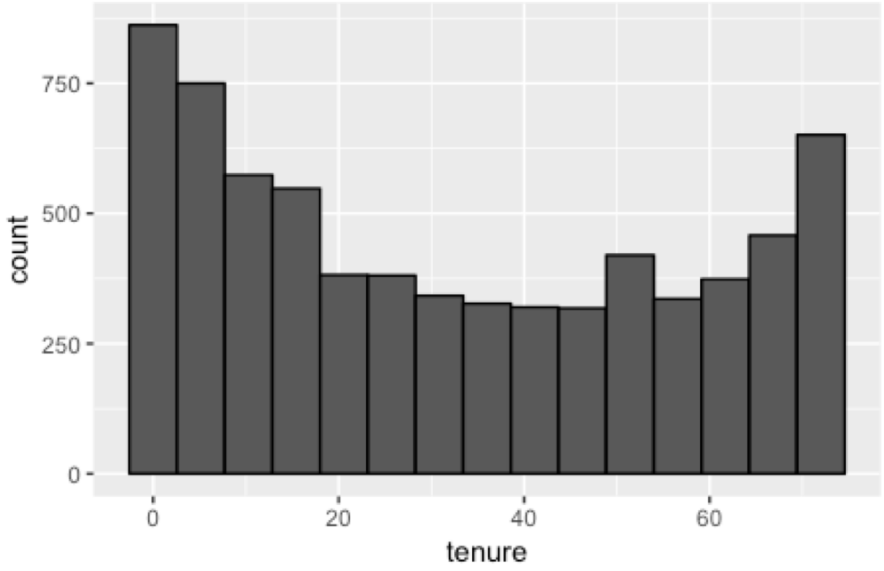
Outliers
Checking numeric variables

tenure	MonthlyCharges	TotalCharges
Min. : 0.00	Min. : 18.25	Min. : 0.0
1st Qu.: 9.00	1st Qu.: 35.50	1st Qu.: 398.6
Median :29.00	Median : 70.35	Median :1394.5
Mean :32.37	Mean : 64.76	Mean :2279.7
3rd Qu.:55.00	3rd Qu.: 89.85	3rd Qu.:3786.6
Max. :72.00	Max. :118.75	Max. :8684.8

BIG DATA EXPERIENCE

Distribution

```
library(ggplot2)
ggplot(churnData.1, aes(x = tenure)) +
  geom_histogram(bins = 15, color = "black")
```



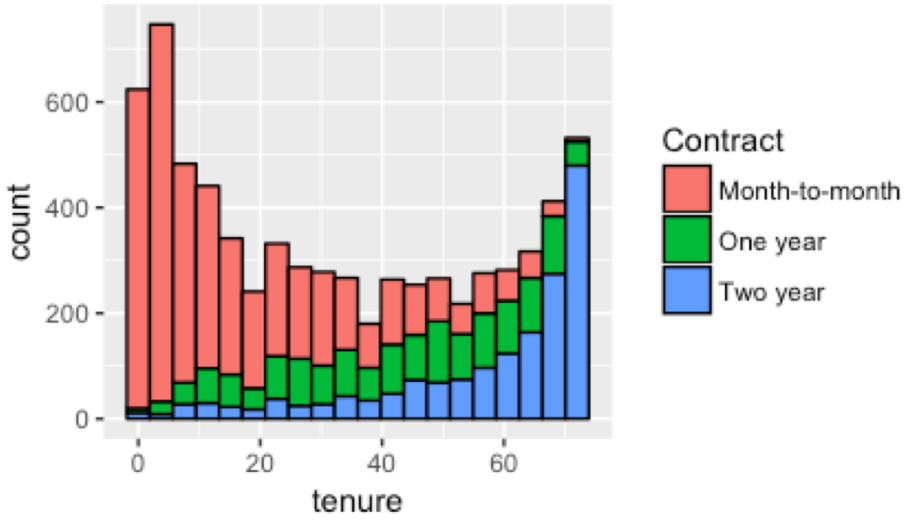
Why is the distribution of tenure look like this?

KMUTT g-able

BIG DATA EXPERIENCE

Distribution
Why tenure look like this?

```
p <- ggplot(churnData.1,
            aes(x = tenure, fill = Contract)) +
   geom_histogram(bins = 20, color="black")
```

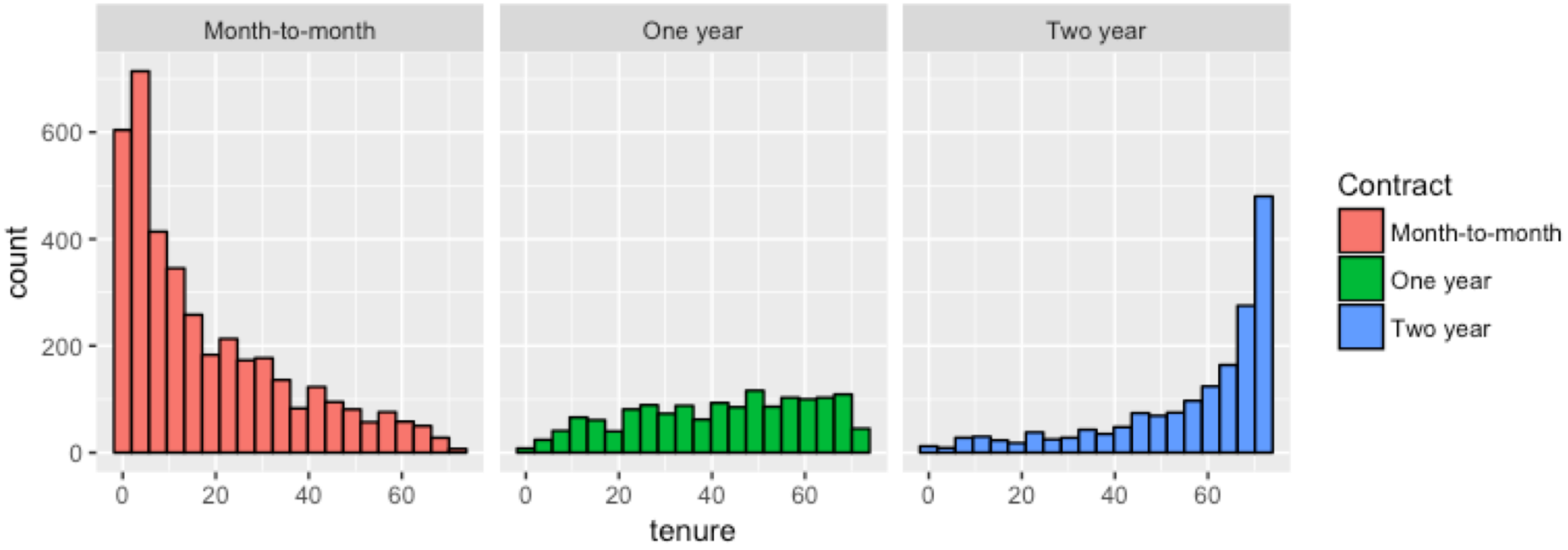


KMUTT g-able

BIG DATA EXPERIENCE

It is the Contract

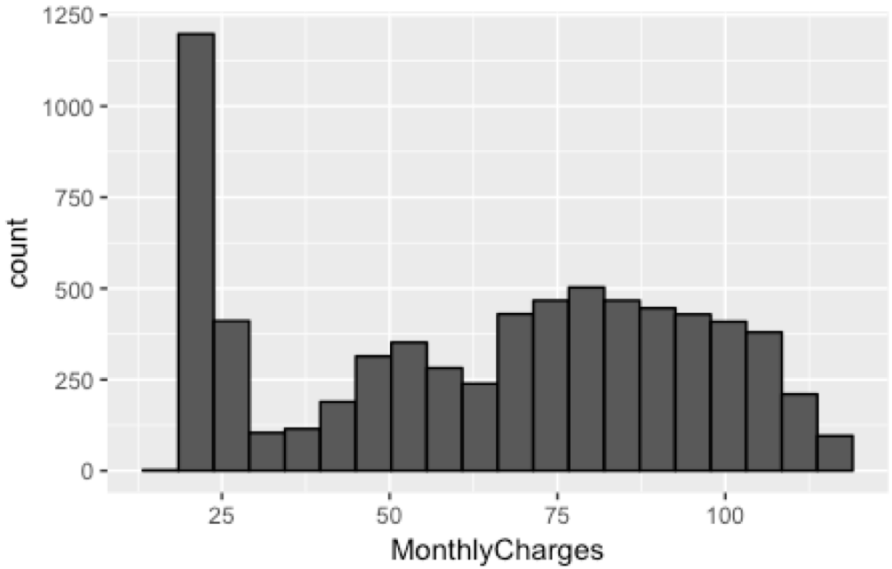
```
p + facet_grid(. ~ Contract)
```



KMUTT g-able

Distribution

```
ggplot(churnData.1, aes(x = MonthlyCharges)) +
  geom_histogram(bins = 20,
                 color = "black")
```



Why are there two peak?

BIG DATA EXPERIENCE

Use lm to assess the relationship

```
fit <- lm(MonthlyCharges ~ ., churnData.1)
summary(fit)
```

```
MultipleLinesNo phone service              NA         NA       NA       NA
MultipleLinesYes                     4.997e+00  2.987e-02  167.319  < 2e-16 ***
InternetServiceFiber optic           2.489e+01  3.855e-02  645.691  < 2e-16 ***
InternetServiceNo                   -2.501e+01  4.944e-02 -505.911  < 2e-16 ***
OnlineSecurityNo internet service          NA         NA       NA       NA
OnlineSecurityYes                    4.984e+00  3.290e-02  151.511  < 2e-16 ***
OnlineBackupNo internet service            NA         NA       NA       NA
OnlineBackupYes                      4.958e+00  3.120e-02  158.902  < 2e-16 ***
DeviceProtectionNo internet service        NA         NA       NA       NA
DeviceProtectionYes                  4.993e+00  3.198e-02  156.143  < 2e-16 ***
TechSupportNo internet service             NA         NA       NA       NA
TechSupportYes                       5.002e+00  3.345e-02  149.534  < 2e-16 ***
StreamingTVNo internet service             NA         NA       NA       NA
StreamingTVYes                       9.939e+00  3.299e-02  301.247  < 2e-16 ***
StreamingMoviesNo internet service         NA         NA       NA       NA
StreamingMoviesYes                   9.931e+00  3.308e-02  300.223  < 2e-16 ***
```

KMUTT g-able

BIG DATA EXPERIENCE

Filter specific group to see who they are

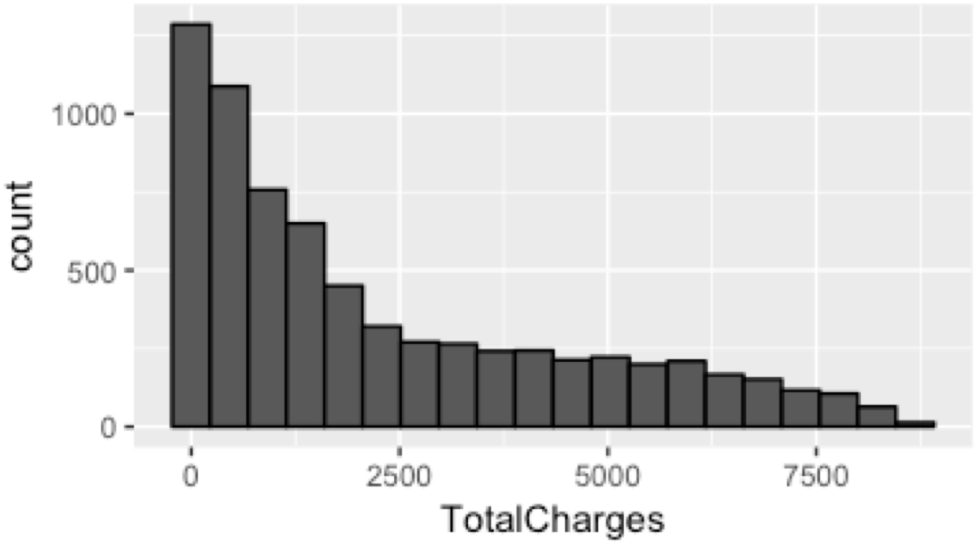
```
summary(churnData.1 %>% filter(MonthlyCharges<26))
```

```
   gender    SeniorCitizen Partner   Dependents     tenure      PhoneService
Female:764   No :1510      No :838   No :932    Min.   : 0.00   No :  79
Male  :812   Yes:  66      Yes:738   Yes:644    1st Qu.: 6.00   Yes:1497
                                                Median :23.00
                                                Mean   :29.18
                                                3rd Qu.:51.00
                                                Max.   :72.00
         MultipleLines     InternetService             OnlineSecurity
No                :1184   DSL         :  79   No                  :  79
No phone service:  79   Fiber optic:   0   No internet service:1497
Yes               : 313   No          :1497   Yes                 :   0
```

BIG DATA EXPERIENCE

Distribution

```
ggplot(churnData.1, aes(x = TotalCharges)) +
  geom_histogram(bins = 20,
                 color = "black")
```



Poisson distribution?
Really skewed

KMUTT g-able

BIG DATA EXPERIENCE

What variable associate with churn? *** at the base level

```
library(FSelector)
options(scipen = T)
ig <- information.gain(Churn ~ ., churnData.1)
```

	attr_importance
gender	0.000037
SeniorCitizen	0.010577
Partner	0.011454
Dependents	0.014467
tenure	0.073362
PhoneService	0.000072
MultipleLines	0.000801
InternetService	0.055574
OnlineSecurity	0.064677
OnlineBackup	0.046792
DeviceProtection	0.043917

KMUTT g-able

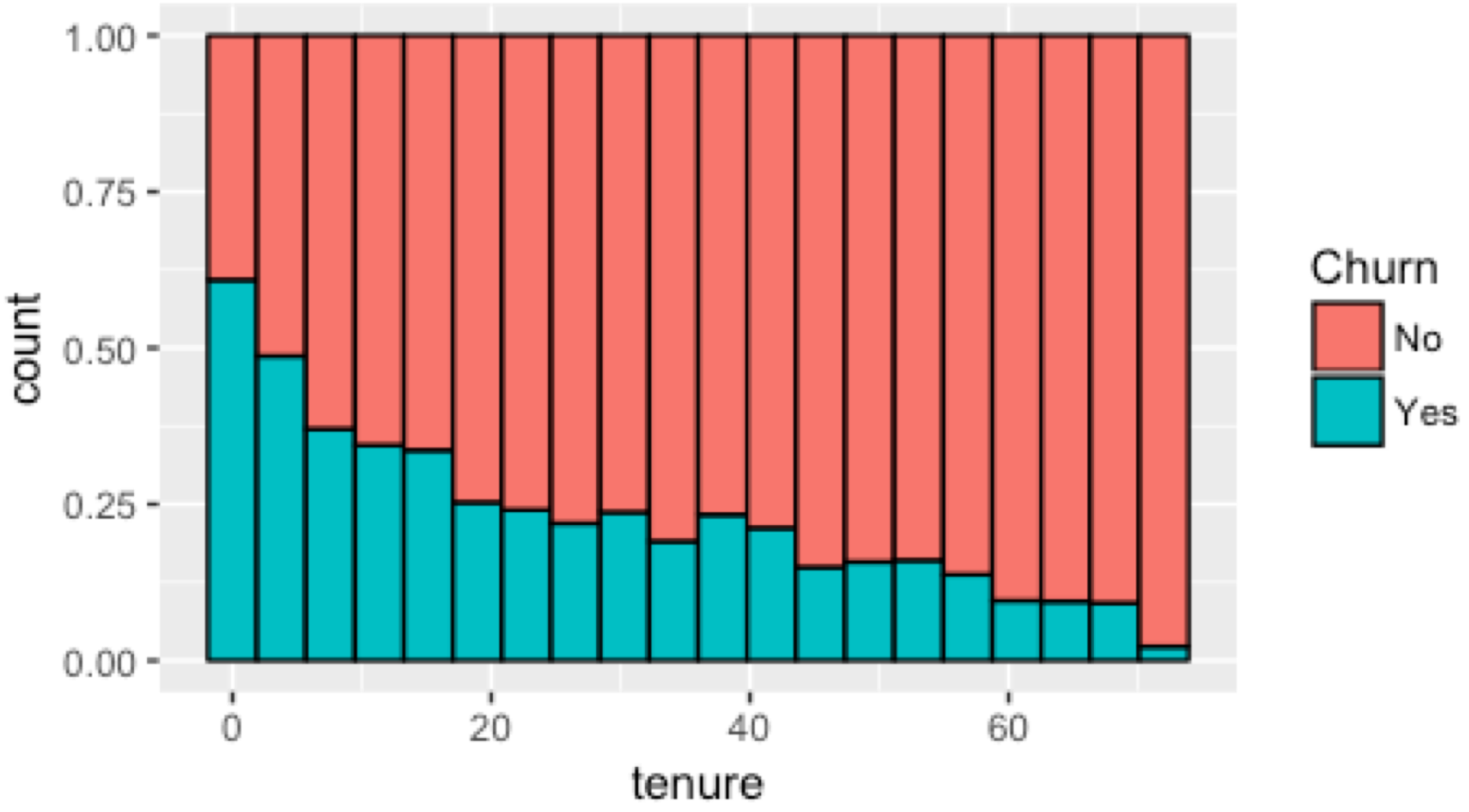
BIG DATA EXPERIENCE

Try tenure

```
ggplot(churnData.1,
       aes(x = tenure, fill = Churn)) +
   geom_histogram(bins = 20,
       color = "black", position = "fill")
```

KMUTT g-able

BIG DATA
EXPERIENCE



g·able

BIG DATA EXPERIENCE

Model for understanding: CART
Create hold-out dataset

```
set.seed(555)
train_pos <- sample(nrow(churnData.1),
                    0.6*nrow(churnData.1))
churnData.train <- churnData.1[train_pos,]
churnData.test <- churnData.1[-train_pos,]
```

KM UTT g·able

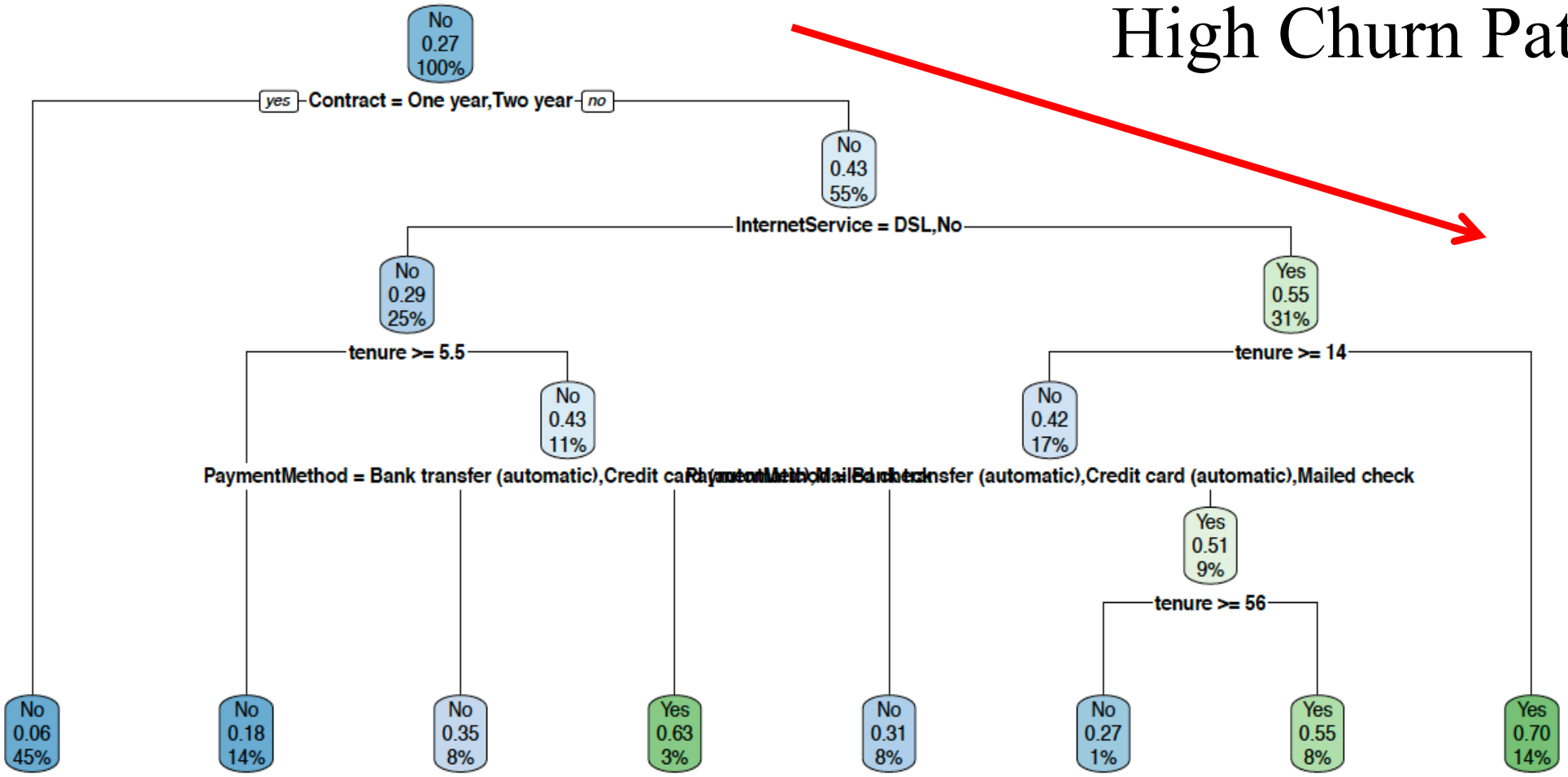
BIG DATA EXPERIENCE

Building CART

```
library(rpart)
library(rpart.plot)
tree <- rpart(Churn ~ ., churnData.train)
pdf("tree.pdf")
rpart.plot(tree)
dev.off()
```

KMU g·able

High Churn Path



BIG DATA EXPERIENCE

Supervised segmentation

- We can get the data of high churn node

```
table(tree$where)
```

2	5	7	8	11	13	14	15
1889	586	334	124	326	48	324	594

KMU g-able

BIG DATA EXPERIENCE

Supervised segmentation

- Get node 15

```
cond <- which(tree$where == 15)
high_churn_node <- churnData.train[cond,]
summary(high_churn_node)
```

KM UTT g-able

BIG DATA EXPERIENCE

More microsegmentation

- We can control the parameter cp to increase the tree complexity
- By setting cp to low value the tree will be more complex
- We can then locate the node that have high churn rate

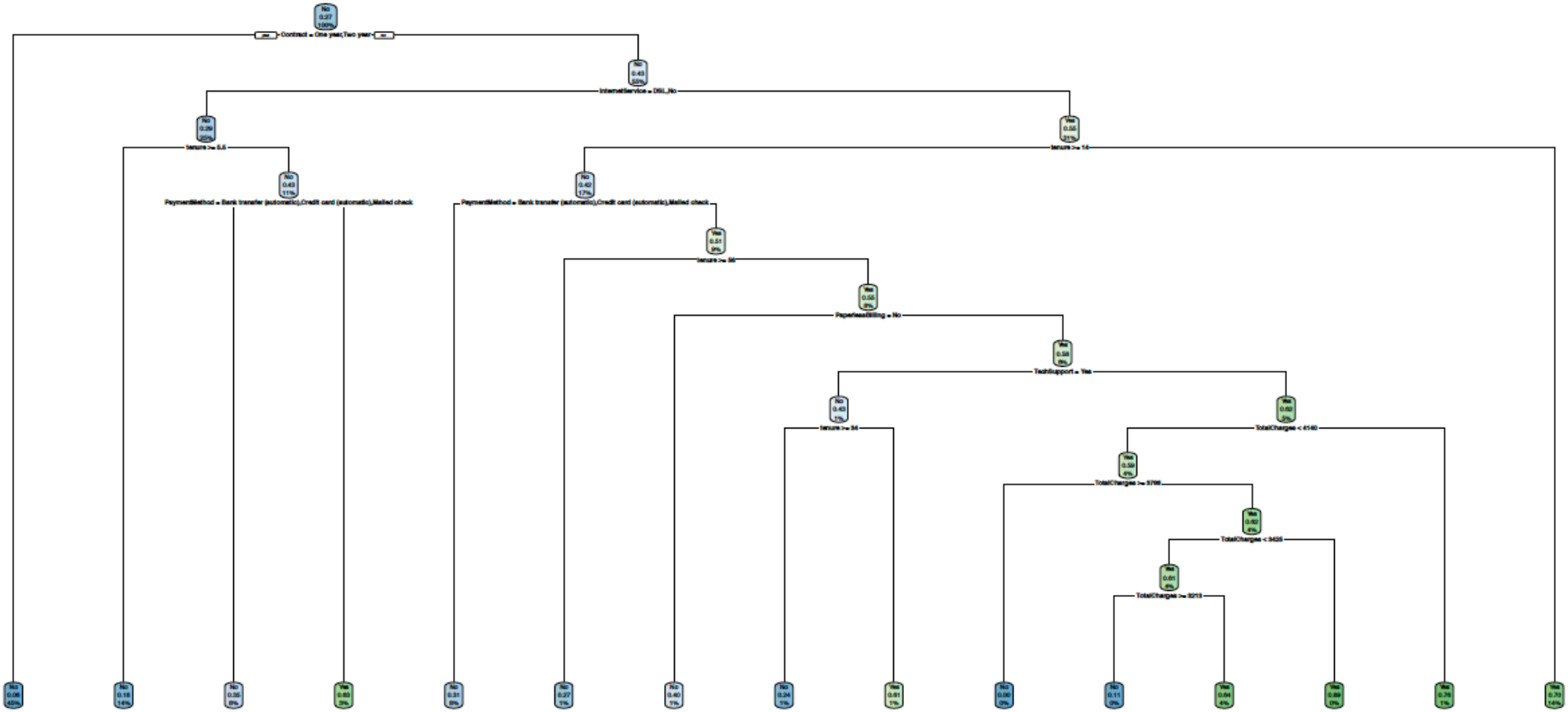
KMUTT g-able

BIG DATA EXPERIENCE

Adjusting cp parameter

```
tree <- rpart(Churn ~ ., churnData.train,
      control = rpart.control(cp = 0.003))
pdf("tree.pdf",width=15)
rpart.plot(tree)
dev.off()
```

KM UTT g·able


No
0.27
100%
Contract = One year,Two year
no
No
0.43
69%
InternetService = DSL,No
No
0.29
10%
Yes
0.55
31%
tenure >= 0.6
tenure >= 14
No
0.43
11%
No
0.42
17%
PaymentMethod = Bank transfer (automatic),Credit card (automatic),Mailed check
PaymentMethod = Bank transfer (automatic),Credit card (automatic),Mailed check
Yes
0.51
9%
tenure >= 56
Yes
0.55
8%
PaperlessBilling = No
Yes
0.58
6%
TechSupport = Yes
No
0.45
1%
tenure >= 54
Yes
0.52
3%
TotalCharges < 4140
Yes
0.58
4%
TotalCharges >= 5799
Yes
0.62
4%
TotalCharges < 3425
Yes
0.61
4%
TotalCharges >= 3213
No
0.06
40%
No
0.18
14%
No
0.35
8%
Yes
0.63
3%
No
0.31
8%
No
0.27
1%
No
0.40
1%
No
0.24
1%
Yes
0.61
1%
No
0.00
0%
No
0.11
0%
Yes
0.64
4%
Yes
0.69
0%
Yes
0.75
1%
Yes
0.70
14%

BIG DATA EXPERIENCE

Assess baseline prediction

```
library(caret)
res <- predict(tree,
               churnData.test,
               type = "class")
confusionMatrix(res,
                churnData.test$Churn,
                mode = "prec_recall",
                positive = "Yes")
```

KM UTT g·able

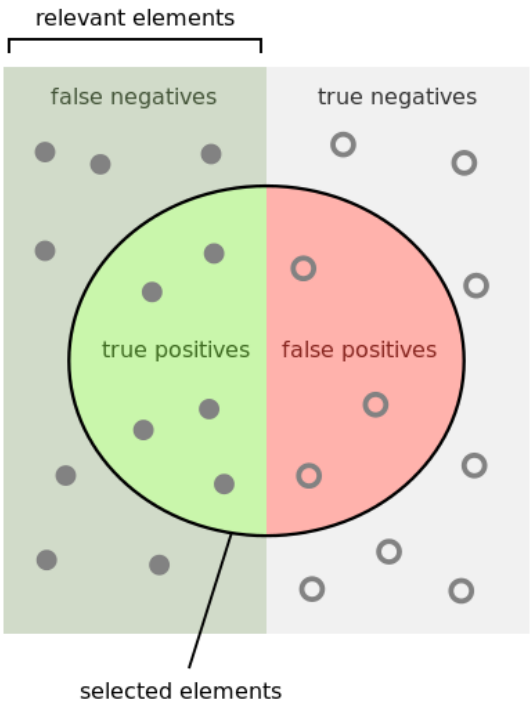
Accuracy, Precision and Recall

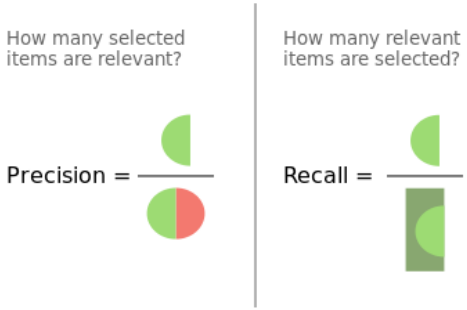
	Actual Positive (p)	Actual Negative (n)
The model says "Yes" = positive (y)	True positives	False positives
The model says "No" = not positive (n)	False negatives	True negatives

- Accuracy = (TP + TN)/(TP + FP + TN + FN)
- Recall (Completeness) = true positive rate = TP/(TP + FN)
- Precision (Exactness) = the accuracy over the cases predicted to be positive, TP/(TP + FP)
- F-measure = the harmonic mean of precision and recall

  = the balance between recall and precision

  $= 2 . \frac{precision * recall}{precision + recall}$





BIG DATA
EXPERIENCE

```
          Reference
Prediction   No  Yes
       No  1785  320
       Yes  292  421

               Accuracy : 0.783
                 95% CI : (0.767, 0.798)
    No Information Rate : 0.737
    P-Value [Acc > NIR] : 1.02e-08

                  Kappa : 0.433
 Mcnemar's Test P-Value : 0.275

              Precision : 0.590
                 Recall : 0.568
                     F1 : 0.579
```

KMU g·able

BIG DATA EXPERIENCE

Lift analysis

- When evaluating machine learning models there is a plethora of possible metrics to assess performance.
- They can also be misleading or don't answer the question at hand very well.
- Accuracy falls short if the classes are unbalanced, e.g. 98% active

```
lift = (predicted rate / average rate)
```

KMU g·able

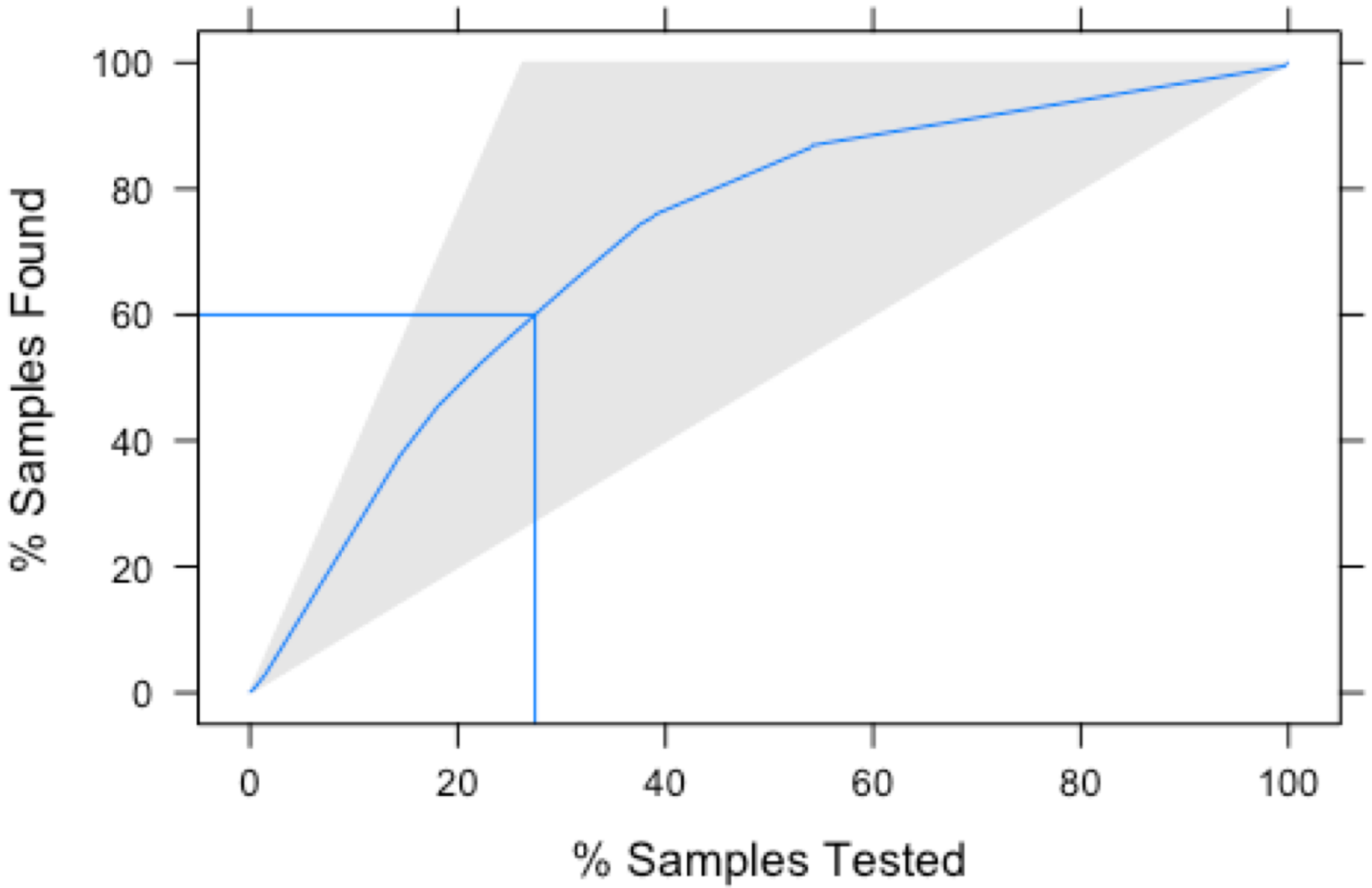
BIG DATA EXPERIENCE

R: Lift calculation (caret)

```
res.p <- predict(tree, churnData.test)[,"Yes"]
lift_result <- data.frame(
  prob =  res.p,
  y = churnData.test$Churn)
lift_obj <- lift(y ~ prob,
                 data = lift_result,
                 class="Yes")
plot(lift_obj, values = 60)
```

KM UTT g·able

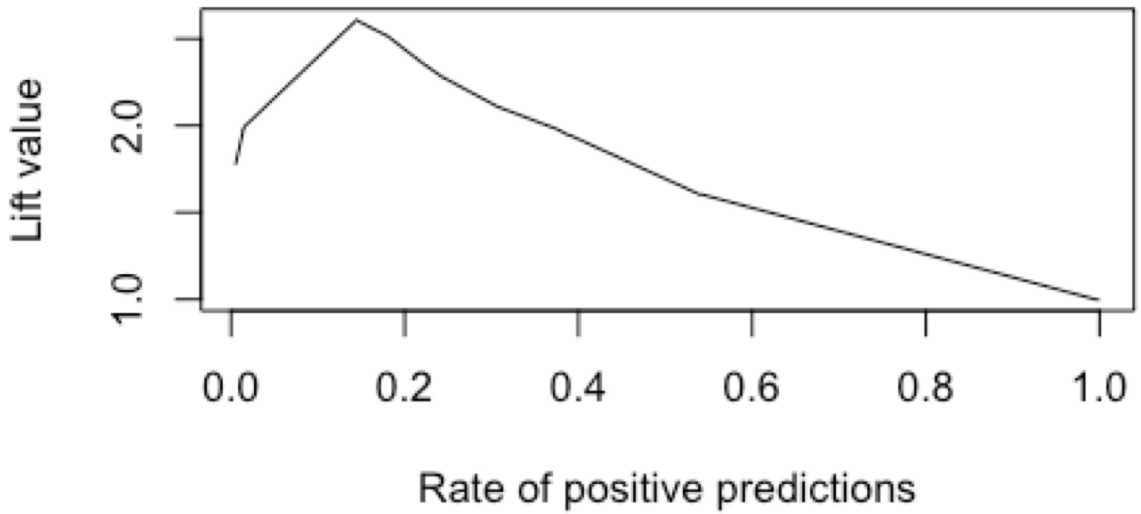
BIG DATA
EXPERIENCE



KMUTT g-able

Lift (ROCR)

```
library(ROCR)
pred <- prediction(res.p, churnData.test$Churn)
perf_lift <- performance(pred, "lift", "rpp")
plot(perf_lift)
```



Lift at 10%

```
library(lift)
TopDecileLift(res.p,
              1-as.integer(churnData.test$Churn))


# [1] 2.5
#
# This 2.5 means 2.5 times random chance
```

BIG DATA EXPERIENCE

How to increase performance?

- Increase complexity
- Ensemble (Bagging -> Random forest)
- Other types of models (not this time…)

KMUTT g-able

BIG DATA EXPERIENCE

Tuning model & cross validation

```
library(randomForest)
train_control = trainControl(method = "cv",
                             number = 5)
metric = "Accuracy"
model <- train(Churn ~ .,
               churnData.train,
               method = "rf",
               metric = metric,
               trControl = train_control)
```

KMUTT g-able

BIG DATA
EXPERIENCE

```
Random Forest

4225 samples
  19 predictor
   2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 3380, 3380, 3380, 3380, 3380
Resampling results across tuning parameters:

  mtry  Accuracy  Kappa
   2    0.79      0.36
  16    0.79      0.43
  30    0.79      0.42

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was mtry = 2.
```

BIG DATA EXPERIENCE

Test the tuned parameters

```
rf <- randomForest(Churn ~ .,
                   churnData.train,
                   mtry = 2)
res <- predict(rf,
                  churnData.test)
confusionMatrix(res,
                 churnData.test$Churn,
                 mode = "prec_recall",
                 positive = "Yes")
```

KMU g·able

BIG DATA
EXPERIENCE

```
          Reference
Prediction   No  Yes
       No  1887  359
       Yes  190  382

               Accuracy : 0.805
                 95% CI : (0.79, 0.82)
    No Information Rate : 0.737
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.458
 Mcnemar's Test P-Value : 7.5e-13

              Precision : 0.668
                 Recall : 0.516
                     F1 : 0.582
```

KMUTT g-able

BIG DATA EXPERIENCE

Cost-benefit of churn prevention

- The model gives us the churn propensity score $P_C(x)$
- Expected benefit/saving can be calculated by using expected value

$$EV = P_C(x) \bullet v_C(x) + (1 - P_R(x))\, v_{NC}(x)$$

- $v_C(x)$ is the benefit/saving of churn
- $v_{NC}(x)$ is the value of not saving churn (negative for cost)

KM UTT g·able

BIG DATA
EXPERIENCE

Thank you

BIG DATA
EXPERIENCE
CENTER

KMU g·able