H2O for Marketing/CRM Applications

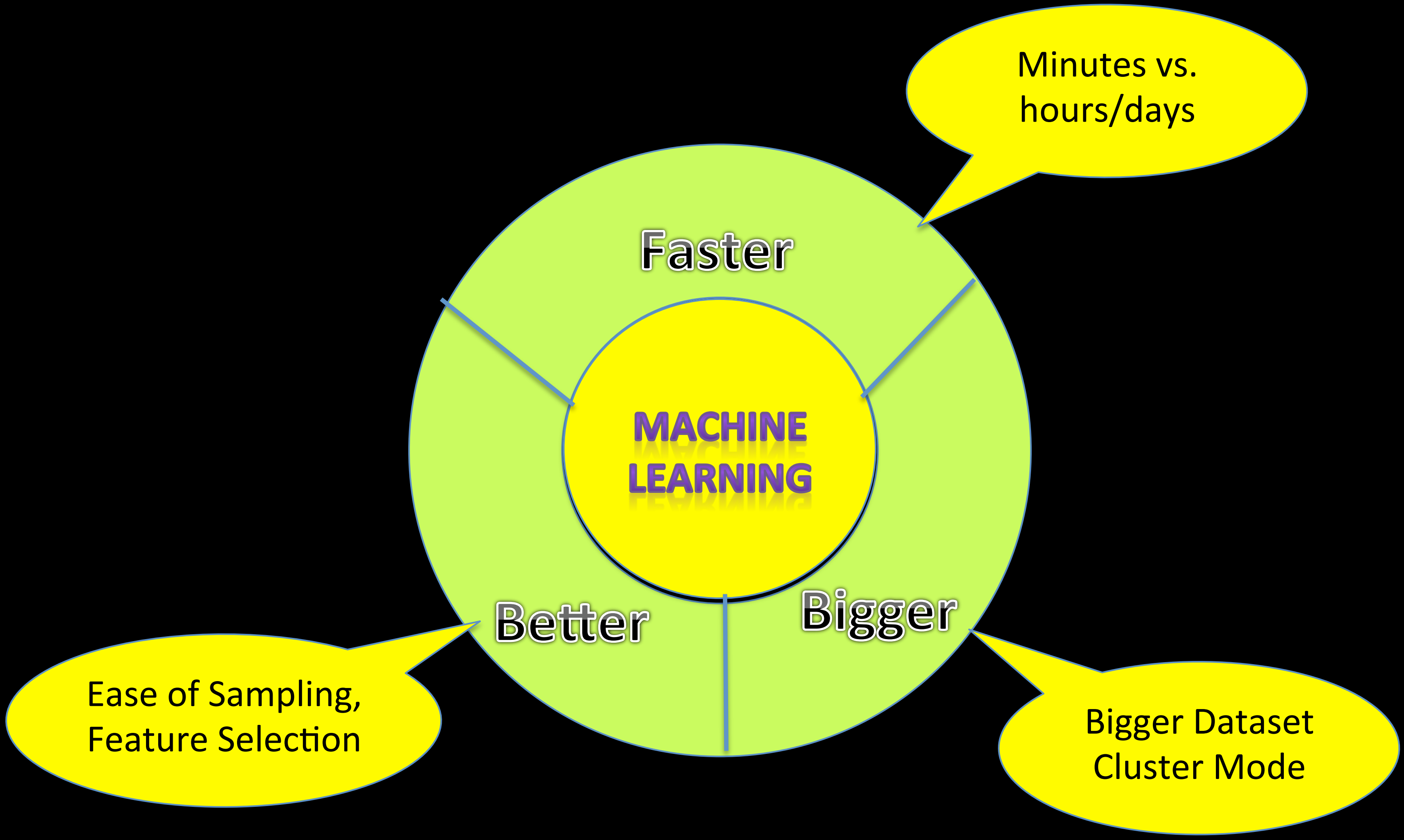
H2O: Faster, Bigger, Better

Yan Zou

Summary

- Why H2O?
- Marketing/CRM Applications
- KDDCup 1998
- Hands-on Training (R/H2O)
- Why H2O (Again)

H2O - Big Data Machine Learning Platform



Minutes vs. hours/days
Faster
MACHINE LEARNING
Better
Bigger
Ease of Sampling, Feature Selection
Bigger Dataset Cluster Mode

Marketing/CRM Applications

- CRM (Marketing, Sales, and Support) is the customer journey
- Intelligent CRM apps dominate corporate IT spending on analytics products
- The frontier of ML is to decipher customer behavior data

Intelligent CRM Apps

Yahoo!	Awareness	Ad Targeting
Google	Discovery	Keyword Bidding
B2B, B2C eCommerce	Consideration	Lead Scoring +
Amazon, Ebay, Macy's Salesforce	Sales	Recommendation / Forecast
	Support/ Retention	Churn Analysis
???	Loyalty	Cross-sell, Up-sell

Hands-on Example: KDDCup98

- Goal: to maximize the profit from fund-raising campaigns
- Dataset:
  - Training: 95412 samples, 481 attributes
    - 2 Target variables: TARGET_B and TARGET_D
  - Test: 96367 samples, 479 attributes
  - Cost per mail: $0.68
- Pre-processing (for this training)
  - ZIP = ZIP / 100
  - Cardinality: 19938 →199

KDDCup98: Using R vs. H2O

- R
  - Read Data
  - Selected Features
  - randomForest (Oops, too many missing values)
  - cforest (Oops, out of memory)
  - ZIP fixed (Oops, cforest still does not return)
  - Score

- H2O
  - Read Data
  - Big data RF
  - Score
    - Profit: $14,513 out-of-the-box
    - Ranked #3 in competition
    - #1: $14,712

KDDCup98: R

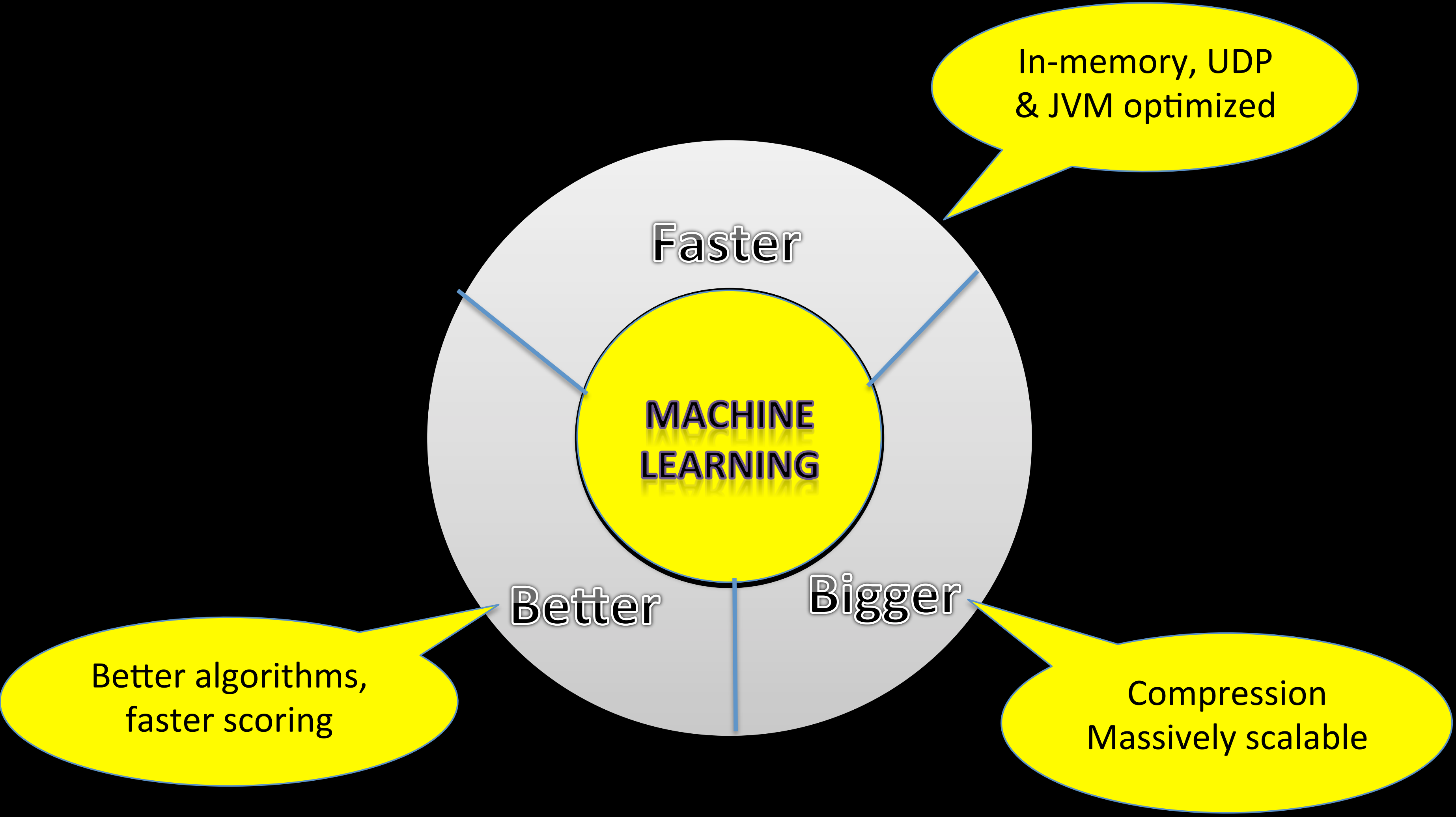
```
setwd("$PATH_TO_KDDCUP98/data/")
Kdd98 <- read.csv("cup98LRN_z.csv")

featureSet <- c("ODATEDW", "OSOURCE", "STATE", "ZIP", "PVASTATE", "DOB",
"RECINHSE", "MDMAUD", "DOMAIN", "CLUSTER", "AGE", "HOMEOWNR", "CHILD03",
"CHILD07",  "CHILD12", "CHILD18", "NUMCHLD", "INCOME", "GENDER", "WEALTH1",
"HIT", "COLLECT1", "VETERANS", "BIBLE", "CATLG", "HOMEE", "PETS", "CDPLAY",
"STEREO", "PCOWNERS", "PHOTO", "CRAFTS", "FISHER", "GARDENIN", "BOATS",
"WALKER", "KIDSTUFF", "CARDS", "PLATES", "PEPSTRFL”, "CARDPROM", "MAXADATE",
"NUMPROM", "CARDPM12", "NUMPRM12",  "RAMNTALL", "NGIFTALL", "CARDGIFT",
"MINRAMNT", "MAXRAMNT", "LASTGIFT", "LASTDATE", "FISTDATE", "TIMELAG",
"AVGGIFT", "HPHONE_D", "RFA_2F", "RFA_2A", "MDMAUD_R", "MDMAUD_F",
"MDMAUD_A", "CLUSTER2", "GEOCODE2", "TARGET_D")

kdd98 <- Kdd98[, setdiff(featureSet, c("CONTROLN", "TARGET_B"))]

library(randomForest)
rf <- randomForest(TARGET_D ~ ., data=kdd98)

library(party)
cf <- cforest(TARGET_D ~ ., data= kdd98, control = cforest_unbiased(mtry=2, ntree=50))
```

KDDCup98: H2O

- Training: (Web UI and R scripts)
- Scoring and Solution Evaluation

```
kdd98 <- read.csv("cup98VAL_z.csv")                    # read data
kdd_pred <- read.csv(“drf_predict.csv")                          # read prediction value
kdd_pred_val <- apply(kdd_pred,1,function(row) if (row[1] > 0.68) 1 else 0 )
kdd98_withpred <- cbind(kdd98, kdd_pred_val)
kdd98_withpred$yield <- apply(kdd98_withpred,1,function(row)
                    (as.numeric(row['TARGET_D']) - 0.68) * as.numeric(row[483]))
sum(kdd98_withpred$yield)                              # profit
max(kdd98_withpred$yield)                              # max donation
sum(kdd_pred_val)                                      # mails sent
```

H2O - Big Data Machine Learning Platform


In-memory, UDP & JVM optimized
Faster
MACHINE LEARNING
Better
Bigger
Better algorithms, faster scoring
Compression Massively scalable