Smart Scalable Feature Reduction with Random Forests

Erik Erlandson
Red Hat, Inc.

SPARK SUMMIT 2017

Erik Erlandson

- Software Engineer
- Radanalytics.io community
- Apache Spark on OpenShift
- Intelligent Applications in the cloud

Talk

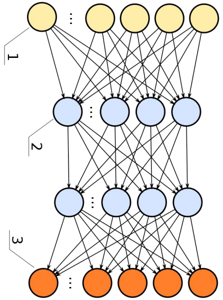
- Motivate Feature Reduction
- Random Forest Clustering
- T-Digest Feature Sketching
- RF Feature Reduction
- Example: Tox21 Assay Data

Features

Measurable Properties!

2.3	1.0	0.0	1.0	3.1	4.2	6.9	0.0	7.3

Model Training



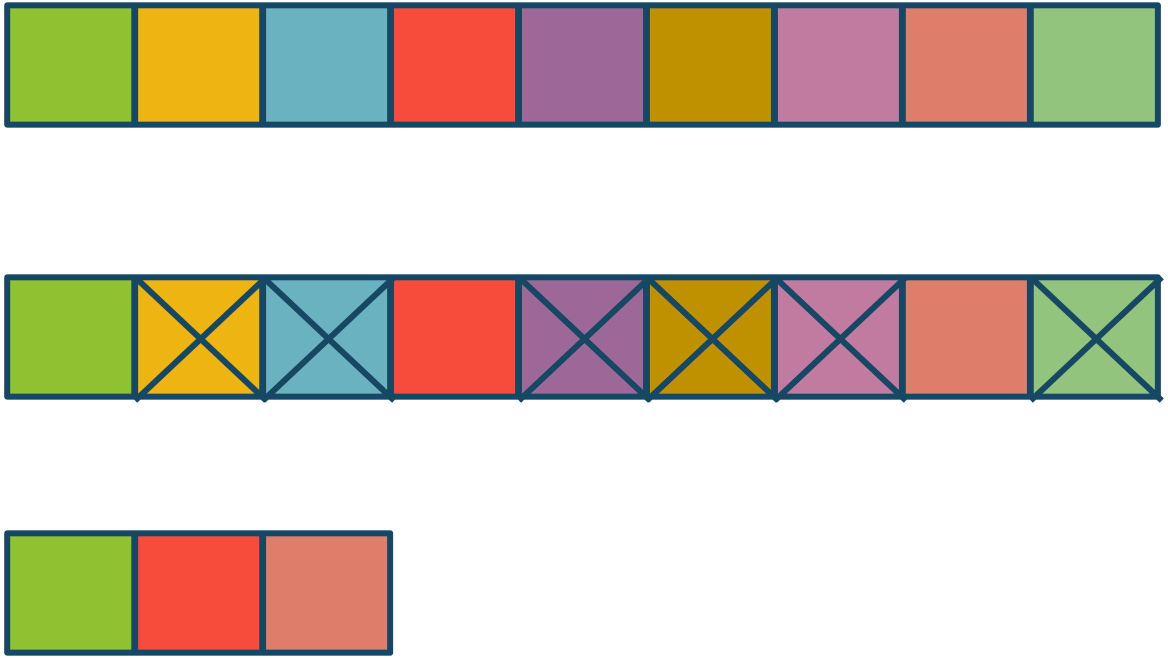
Evaluation

Feature Reduction

Full Feature Set

Identify Useful Features

Reduced Feature Set



SPARK SUMMIT 2017

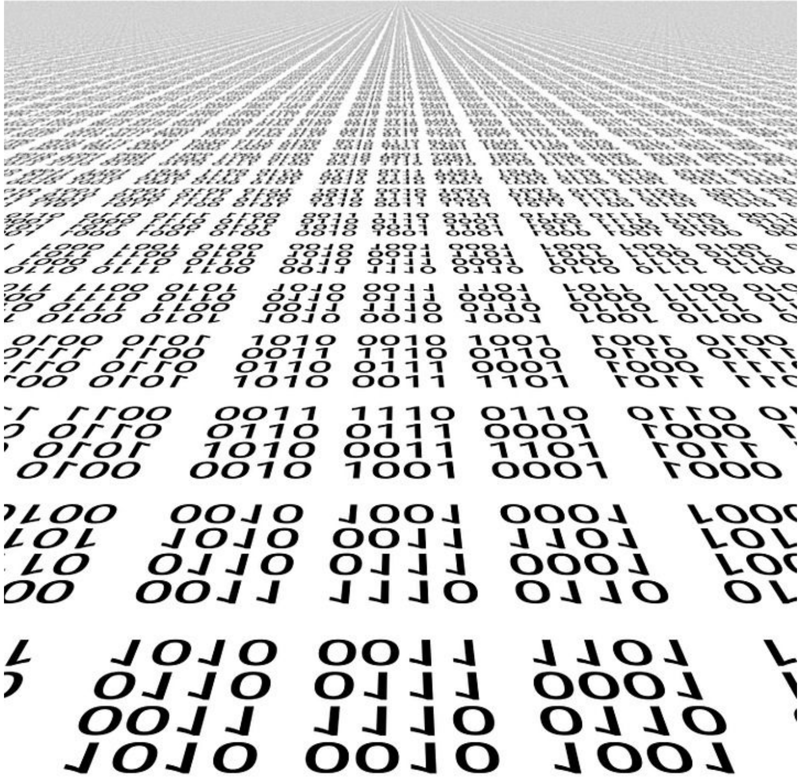
Feature Sets Can Be Very Large

hundreds

thousands

...

millions



Features Cost Resources

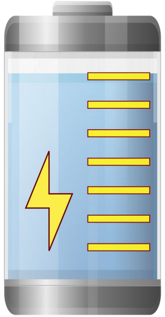
Memory

Network

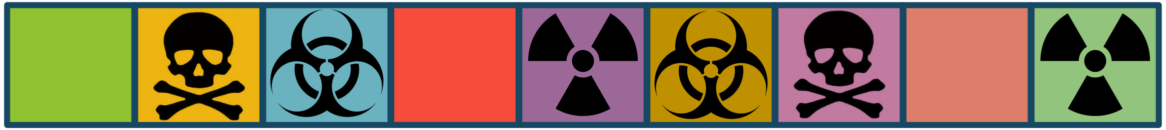
Time

Disk

Energy



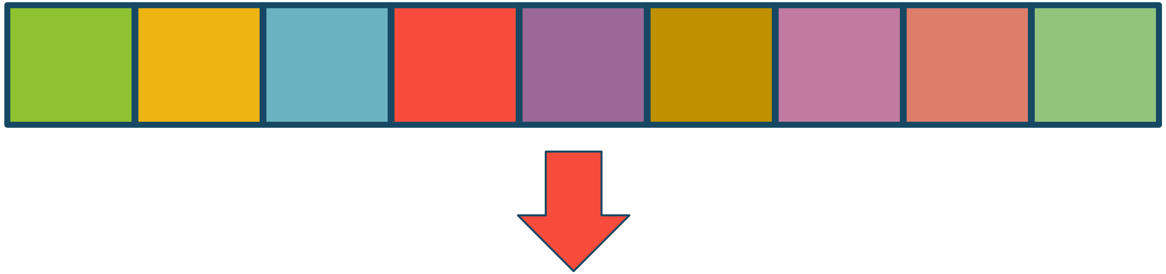
Features Inject Noise

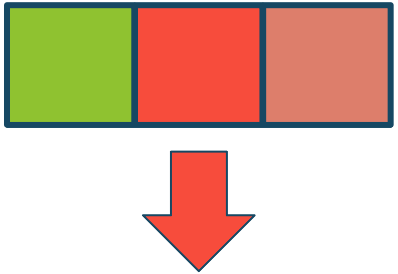


Model Training
Without Reduction

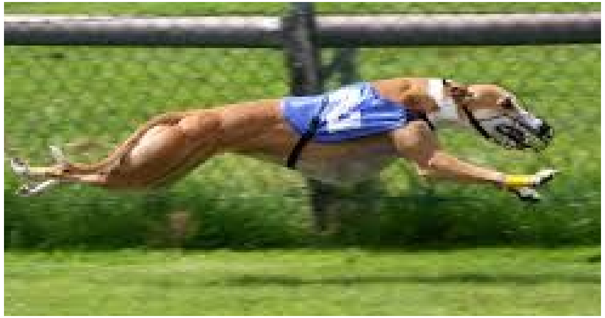
With Feature
Reduction

Features Impact Model Size



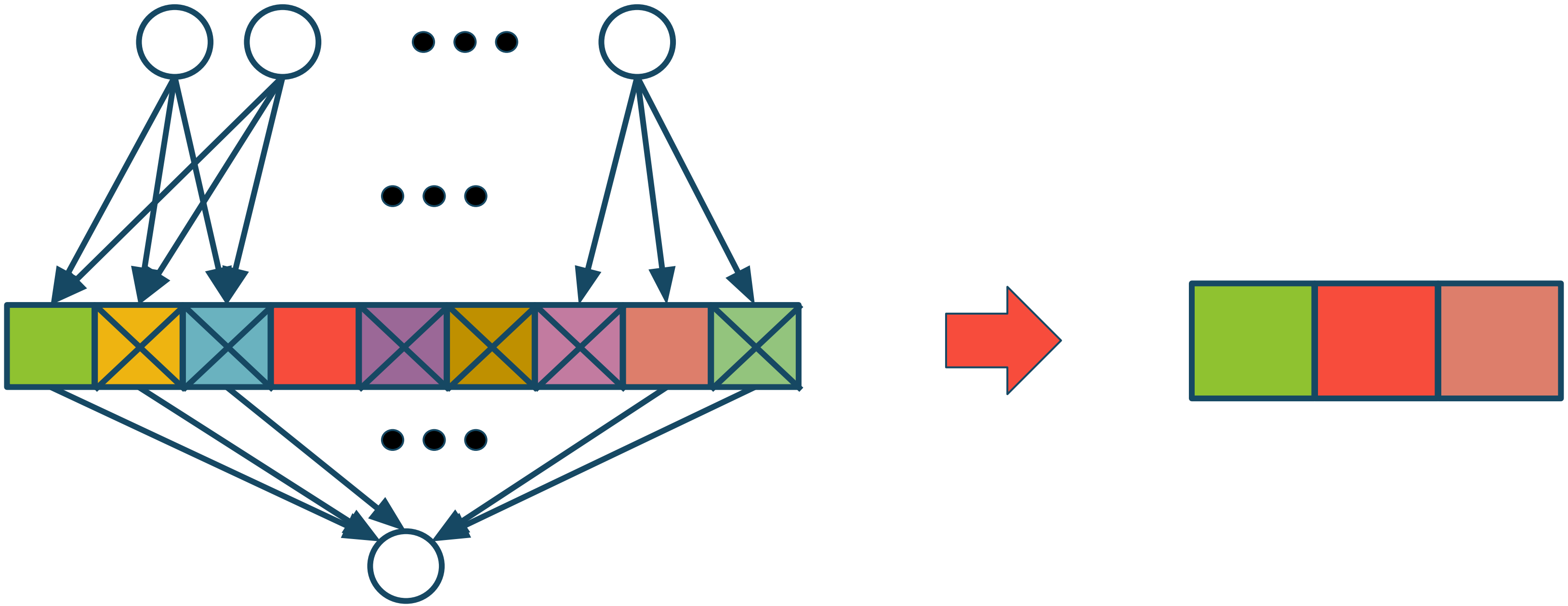


Model Without Reduction

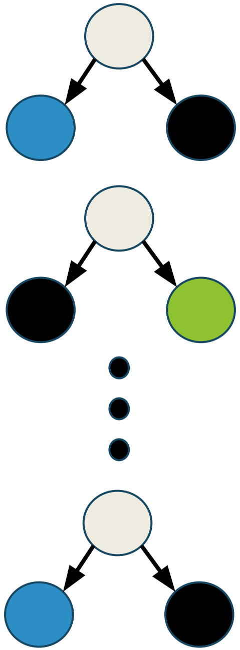


With Reduction

Representation & Transfer Learning



Random Forests



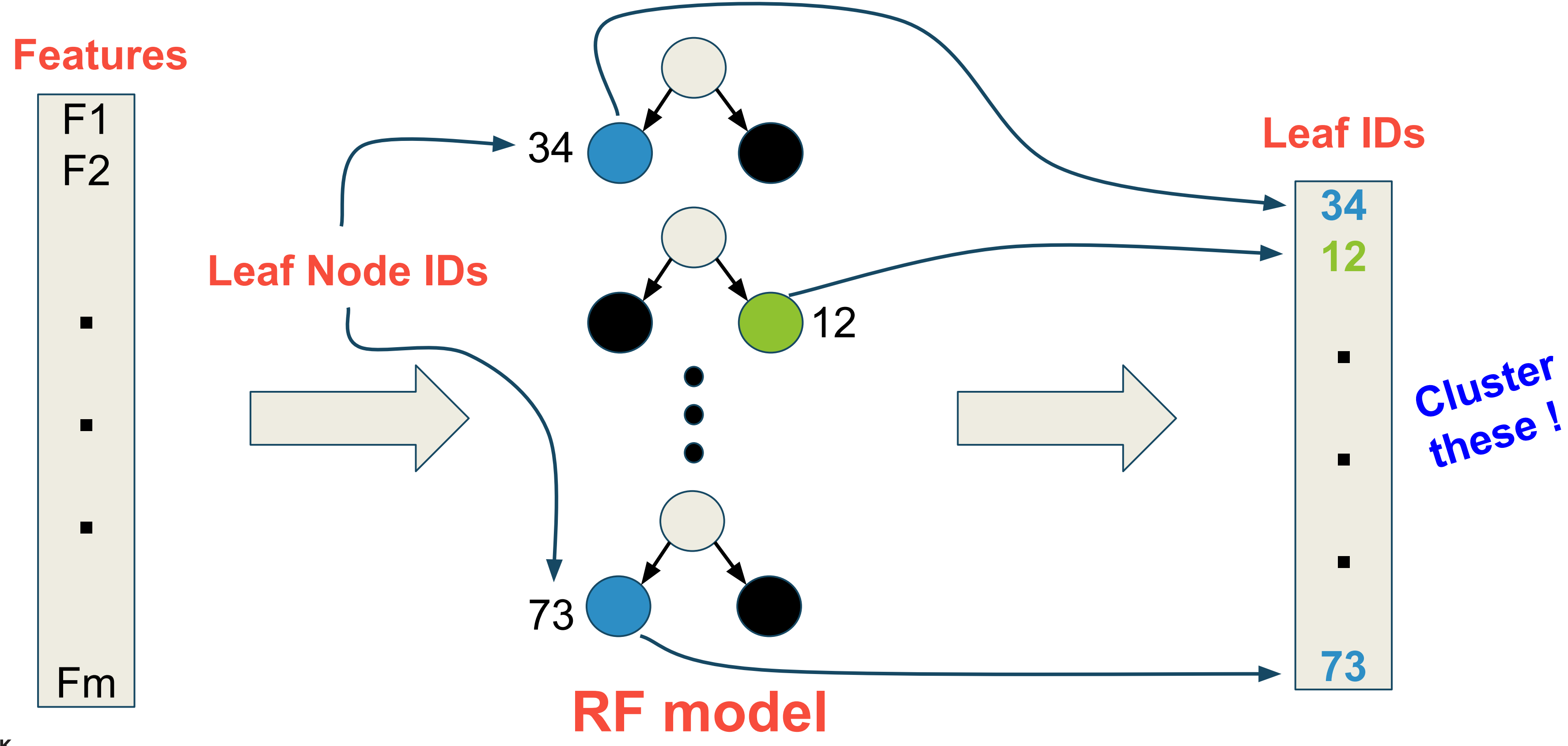
Leo Breiman (2001)

Ensemble of Decision Tree Models

Each tree trains on random subset of data

Each split considers random subset of features

Random Forest Clustering

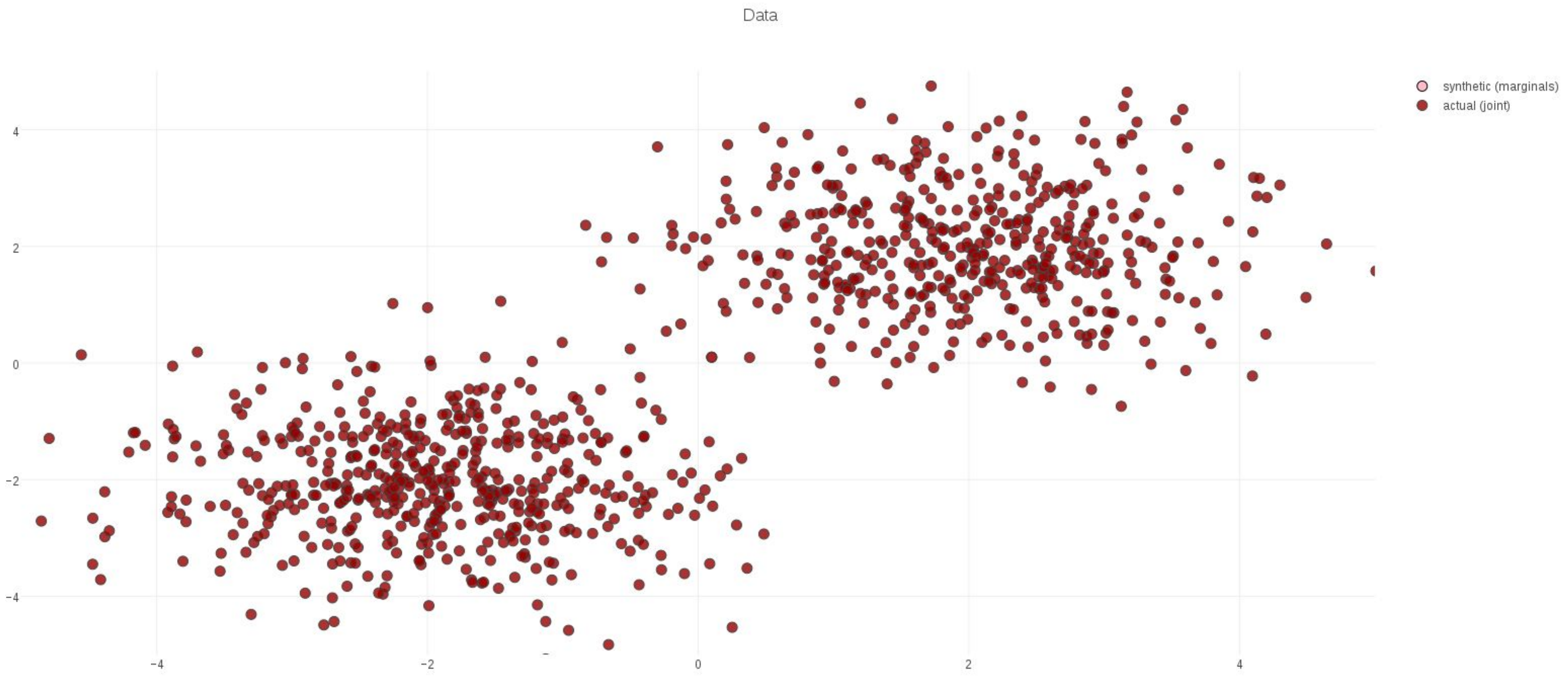


2 Key Benefits of RF Clustering

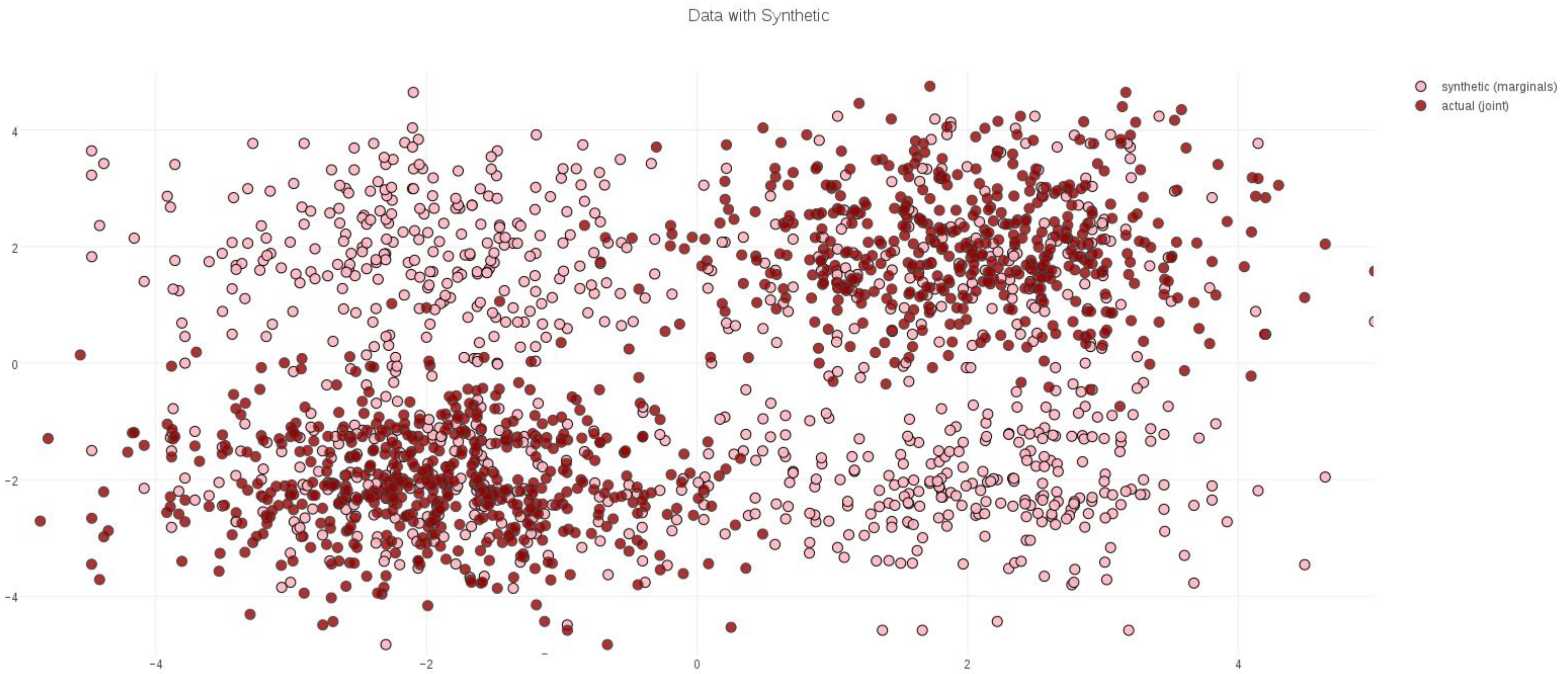
RF Training ignores unhelpful features

Features Used by RF Model << Full Feature Set

Data with a Joint Distribution in R^2

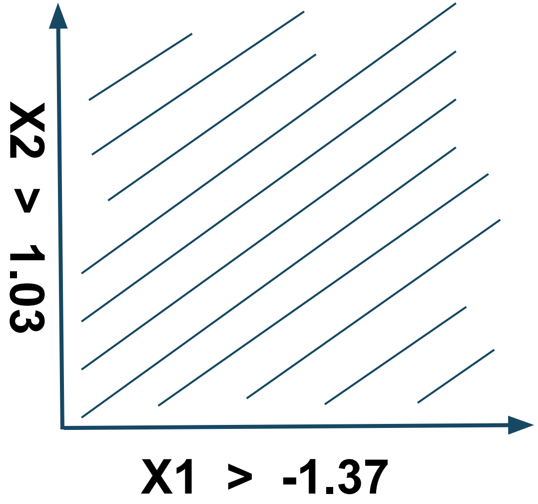


Data with Synthetic

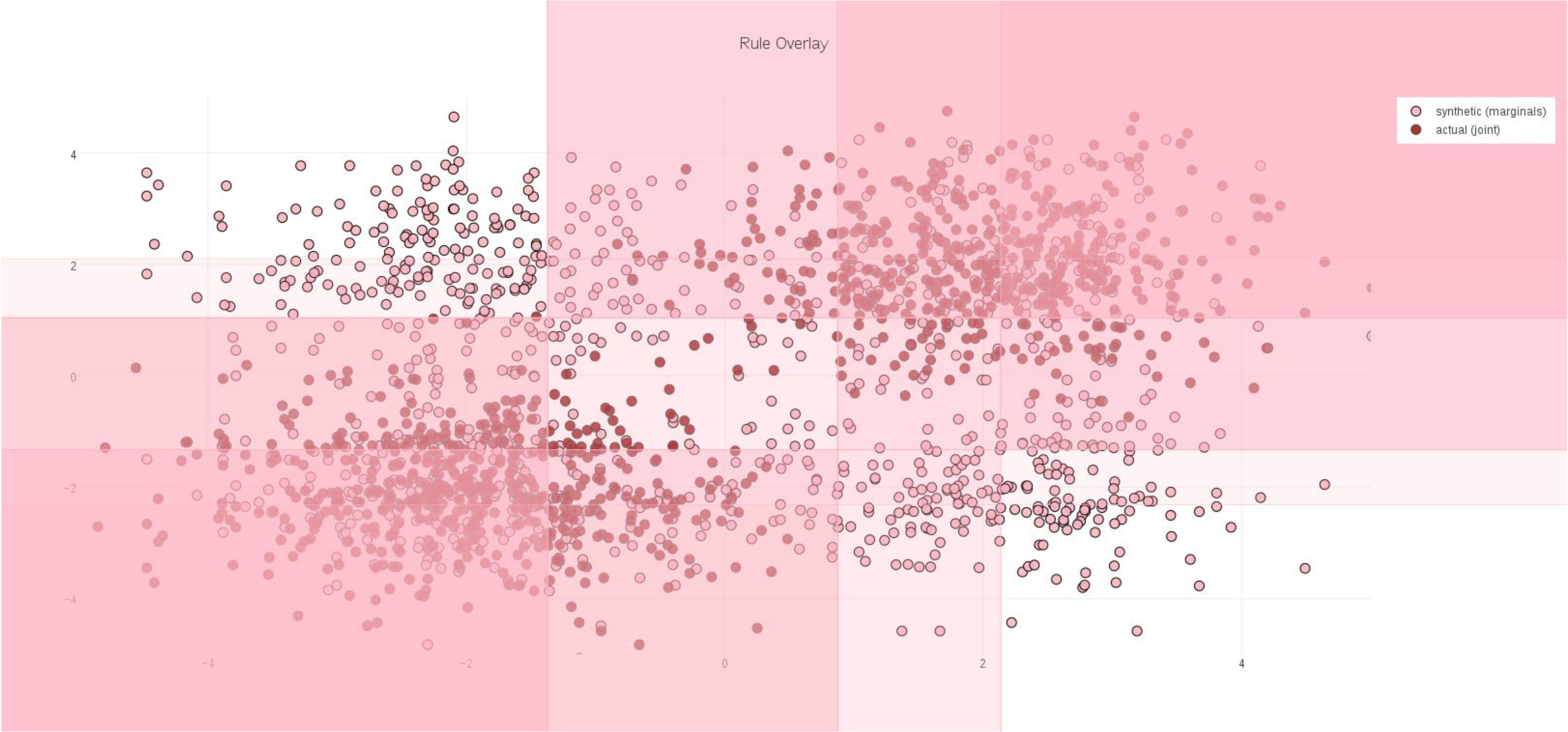


RF Rules for Data (non-synthetic)

```
List((x2 <= -1.32), (x1 <=  0.87))
List((x1  > -1.37), (x2  >  1.03))
List((x2 <=  2.09), (x1 <=  0.87))
List((x1 <=  2.13), (x2 <= -1.32))
List((x2 <= -2.31), (x1 <=  0.87))
```



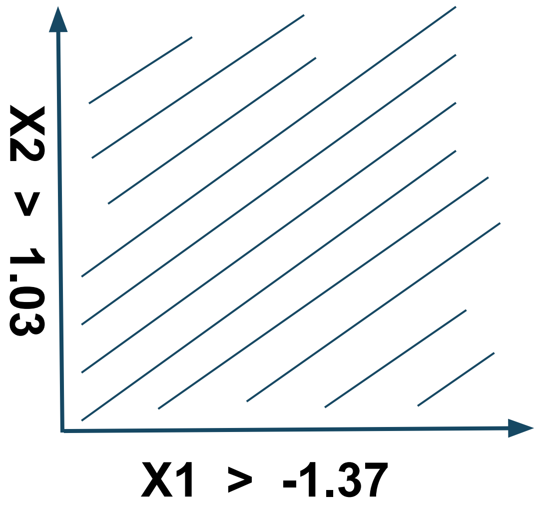
RF Rules in Feature Space



What Features Did the RF Use?

```
List((x2 <= -1.32), (x1 <=  0.87))
List((x1  > -1.37), (x2  >  1.03))
List((x2 <=  2.09), (x1 <=  0.87))
List((x1 <=  2.13), (x2 <= -1.32))
List((x2 <= -2.31), (x1 <=  0.87))

reduced = {"x1", "x2"}
```



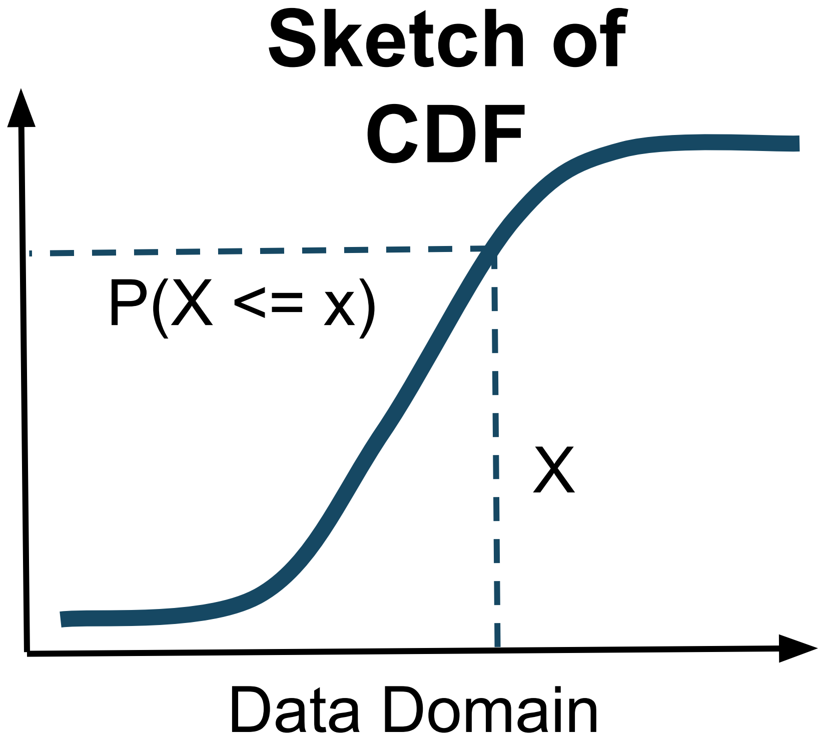
T-Digest Sketches a Distribution

3.4

6.0

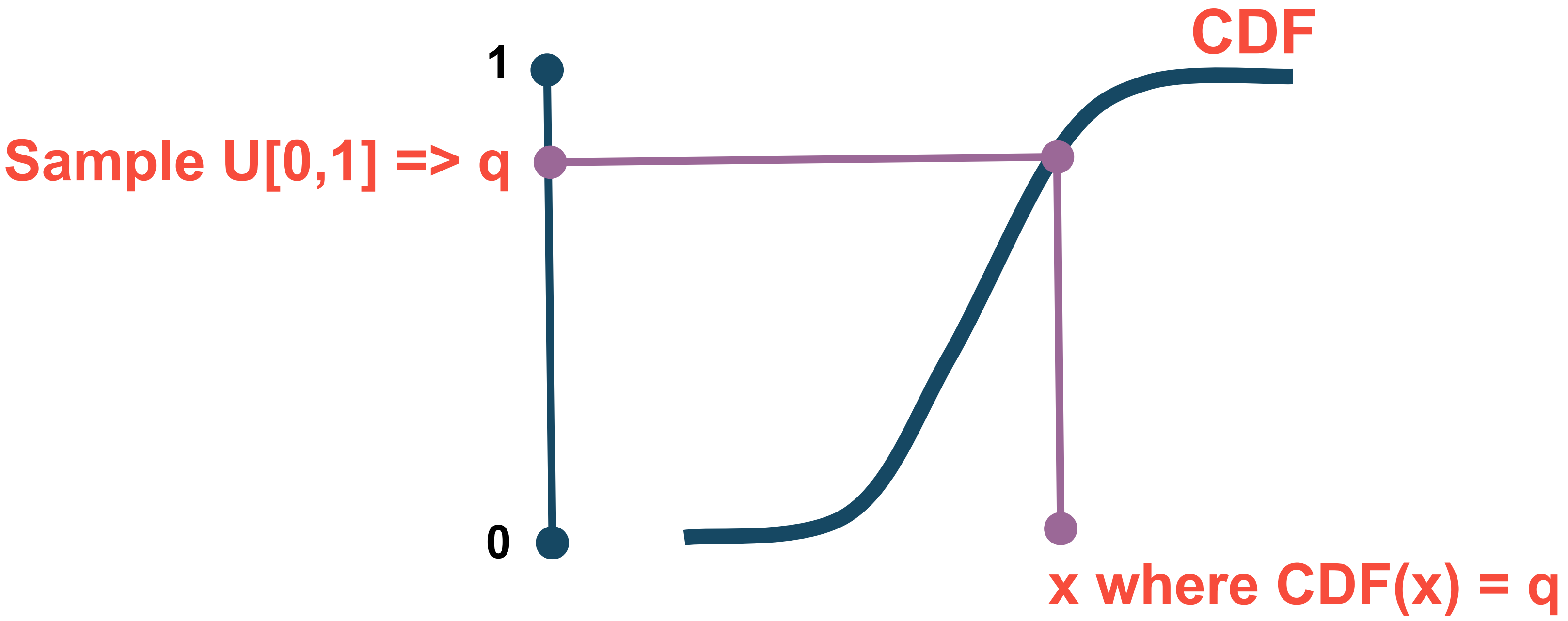
2.5

⋮

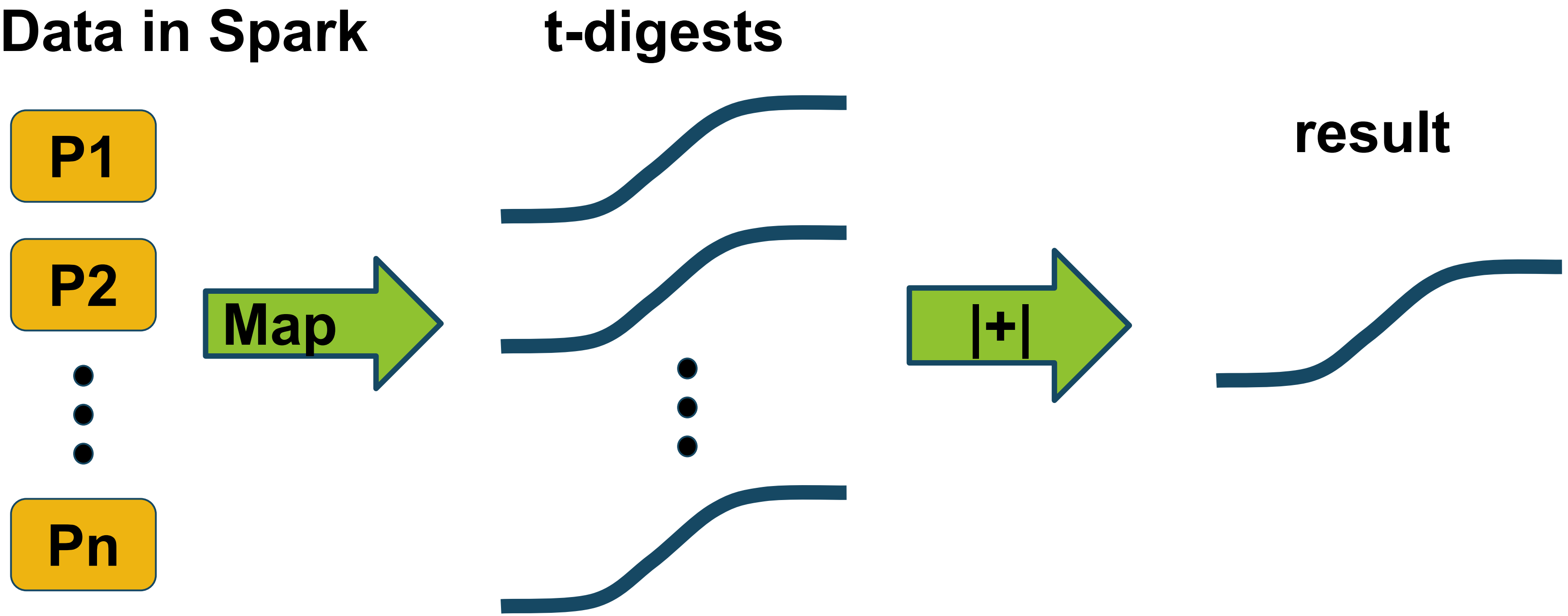


SPARK SUMMIT 2017

Inverse Transform Sampling



T-Digests Can Aggregate



Sketching a Feature

```
feature.aggregate(TDigest.empty())(
   (td, x) => td + x,
   (td1, td2) => td1 ++ td2
)
```

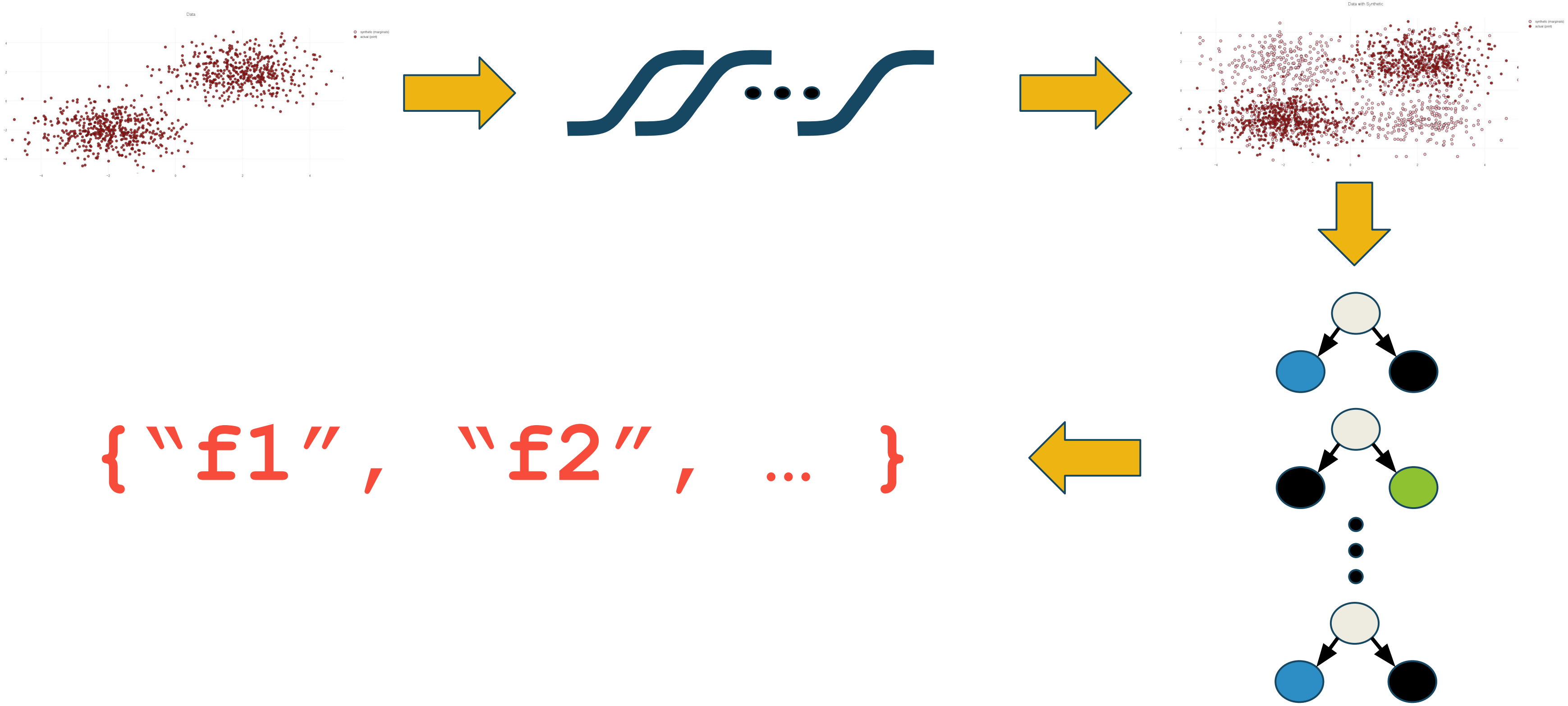
Synthesizing Data from TDigests

```
def synthesize(tdVec: Vector[TDigest],
                     n: Int) = {
  val tdVecBC = sc.broadcast(tdVec)
  sc.parallelize(1 to n).map { _ =>
    tdVecBC.value.map(_.sample)
  }
}
```

Random Forest Training Data

```
val fvSketches = sketchFV(trainFV)
val synthFV = synthesize(fvSketches, 48000)
val trainLab = trainFV.map(_.toLabeledPoint(1.0))
val synthLab = synthFV.map(_.toLabeledPoint(0.0))
val trainFR = trainLab ++ synthLab
```

Random Forest Feature Reduction



Tox21 Data Challenge

National Institute of Health (2014)

12 Toxicity Assays

12060 compounds + 647 hold-out

https://tripod.nih.gov/tox21/challenge/index.jsp

DeepTox

Johannes Kepler University Linz

Institute of Bioinformatics

http://bioinf.jku.at/research/DeepTox/tox21.html

[Mayr2016] Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2016). DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science*, **3**:80.

[Huang2016] Huang, R., Xia, M., Nguyen, D. T., Zhao, T., Sakamuru, S., Zhao, J., Shahane, S., Rossoshek, A., & Simeonov, A. (2016). Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science*, **3**:85.

Tox21 Data

I used these

801 Dense Features

272K Sparse Features

Each assay represented on a different subset

compound	NR.AhR	NR.AR	NR.AR.LBD	NR.Aromatase	...
NCGC00261900-01	0	1	NA	0	
NCGC00260869-01	0	1	NA	NA	.
NCGC00261776-01	1	1	0	NA	.
NCGC00261380-01	NA	0	NA	1	.
NCGC00261842-01	0	0	0	NA	
NCGC00261662-01	1	0	0	NA	
NCGC00261190-01	NA	0	0	NA	

Experiment

Train models on all 12 assays

Perform Random Forest Feature Reduction

Train similar models on reduced feature set

Compare models on each assay

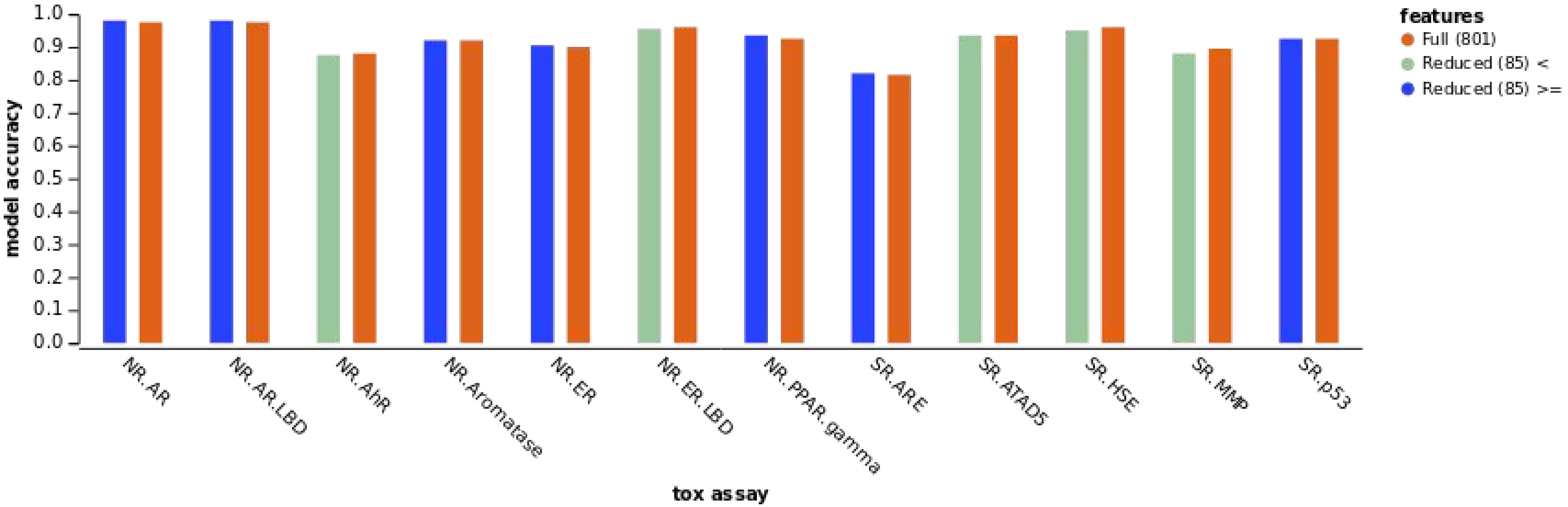
85 of 801 Features Were Used

Features

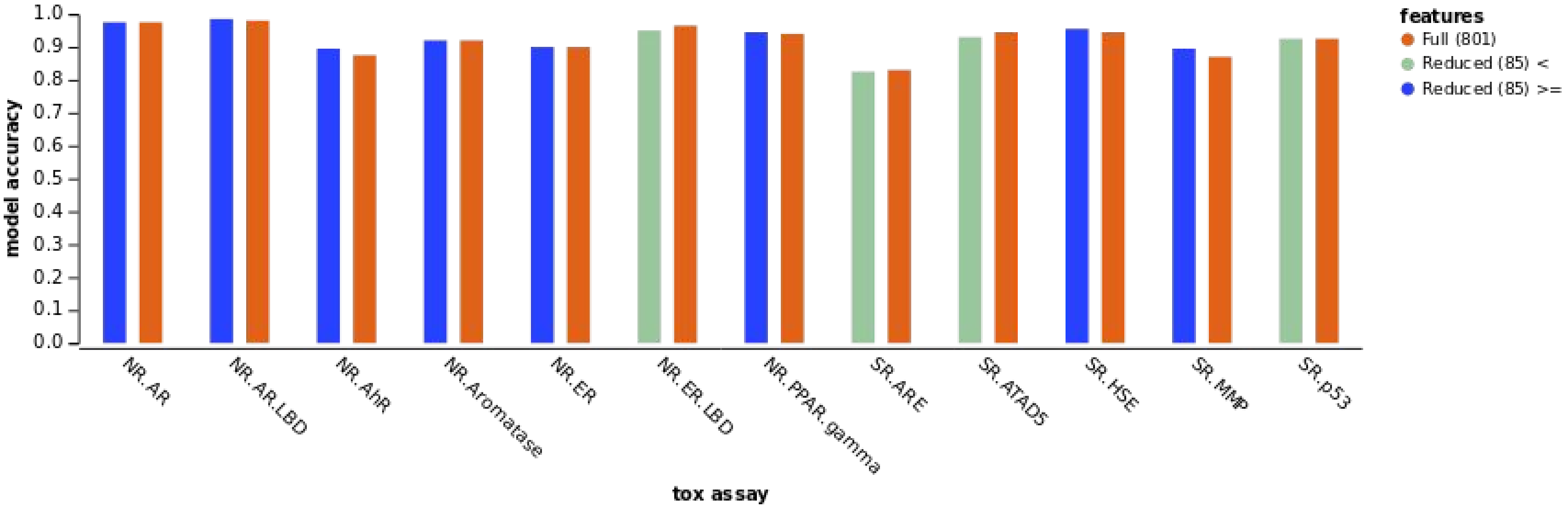
Number trees used

RNCS	21
MRVSA7	20
VSAEstate2	19
VSAEstate3	18
slogPVSA8	18
VSAEstate0	17
slogPVSA6	16
RDFM29	12
slogPVSA3	12
RDFM30	12

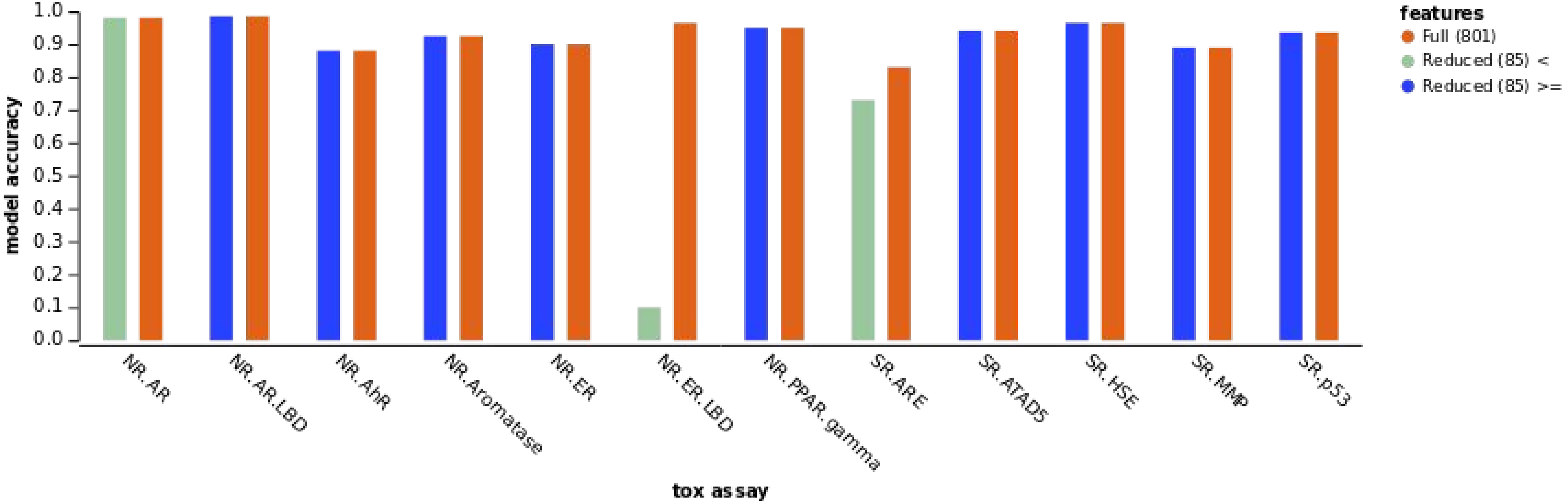
Full vs Reduced (Logistic Reg)



Full vs Reduced (Boosted DTE)



Full vs Reduced (SVM)



Training Times

(times in seconds)	Full (801)	Reduced (85)
Logistic Regression	68.5	46.8
SVM	35.3	33.8
GB Tree Ensemble	247	65.0

Evaluation Times

(times in seconds)	Full (801)	Reduced (85)
Logistic Regression	32.1	3.88
SVM	0.59	0.23
GB Tree Ensemble	1.33	0.88

Thank You

SPARK SUMMIT 2017

Erik Erlandson
eje@redhat.com
@manyangled
https://github.com/erikerlandson/feature-reduction-talk