One-Pass Data Science in Apache Spark with Generative T-Digests

Erik Erlandson, Red Hat

#EUds11

SPARK SUMMIT

Landscape

Features & Feature Randomization

3 Applications

T-Digests & Generative Sampling

3 Applications: Reprise

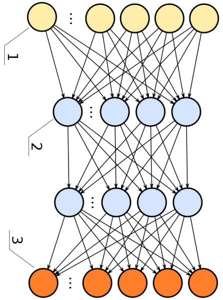
Feature Importance Demo

Features

Measurable Properties!

2.3	1.0	0.0	1.0	3.1	4.2	6.9	0.0	7.3

Model Training

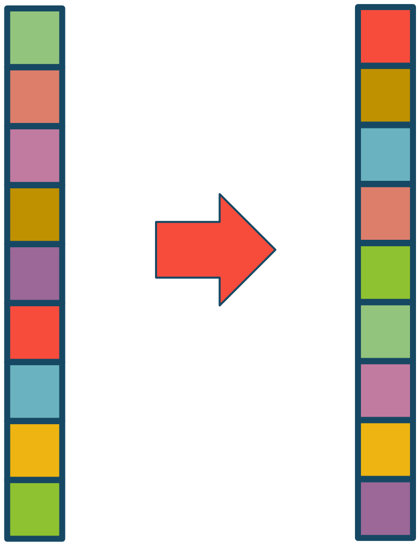


Evaluation

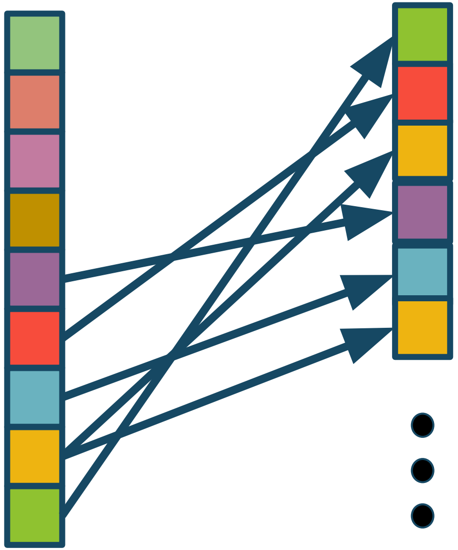
Feature Randomization

Randomization Methods

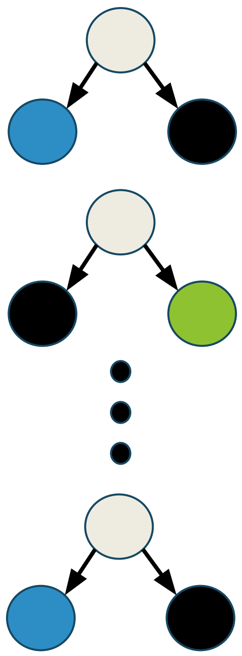
Permutation



Selection



Random Forests



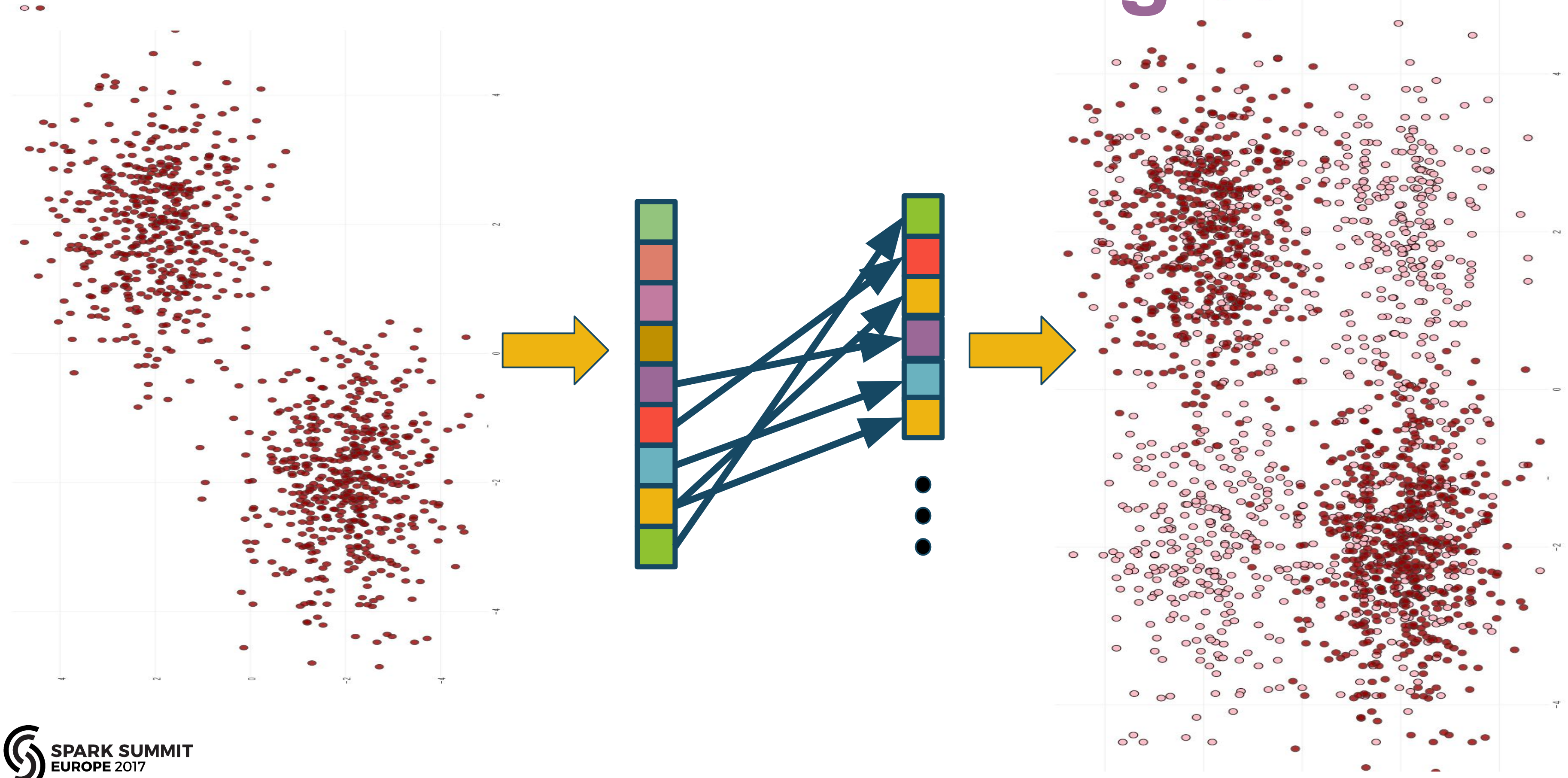
Leo Breiman (2001)

Ensemble of Decision Tree Models

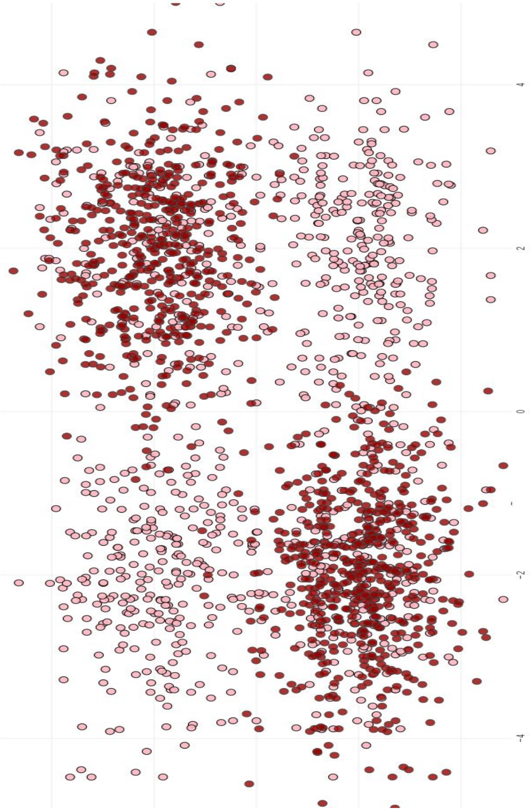
Each tree trains on random subset of data

Each split considers random subset of features

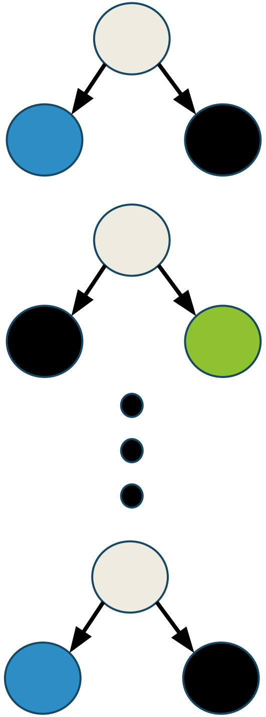
Random Forest Clustering



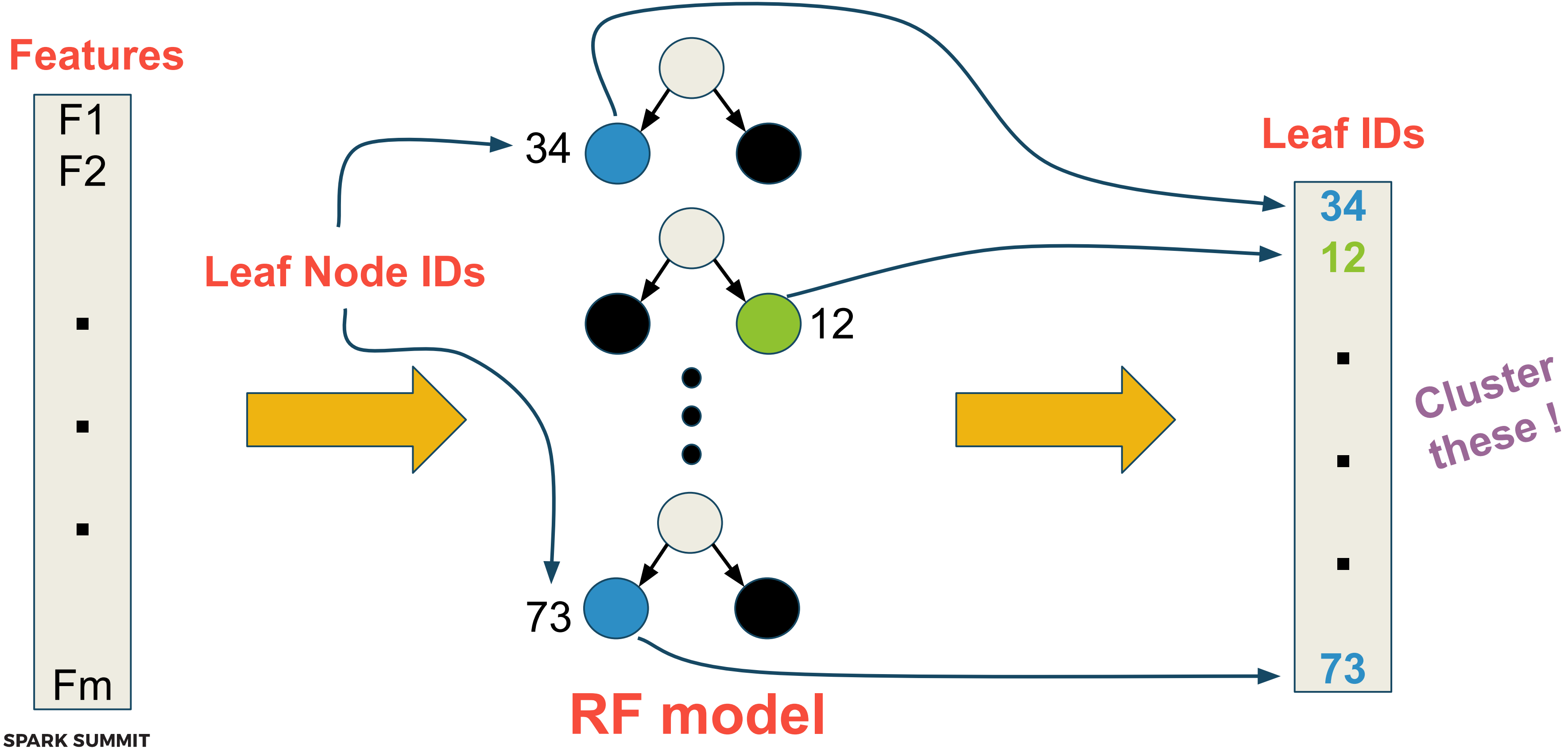
Random Forest Clustering



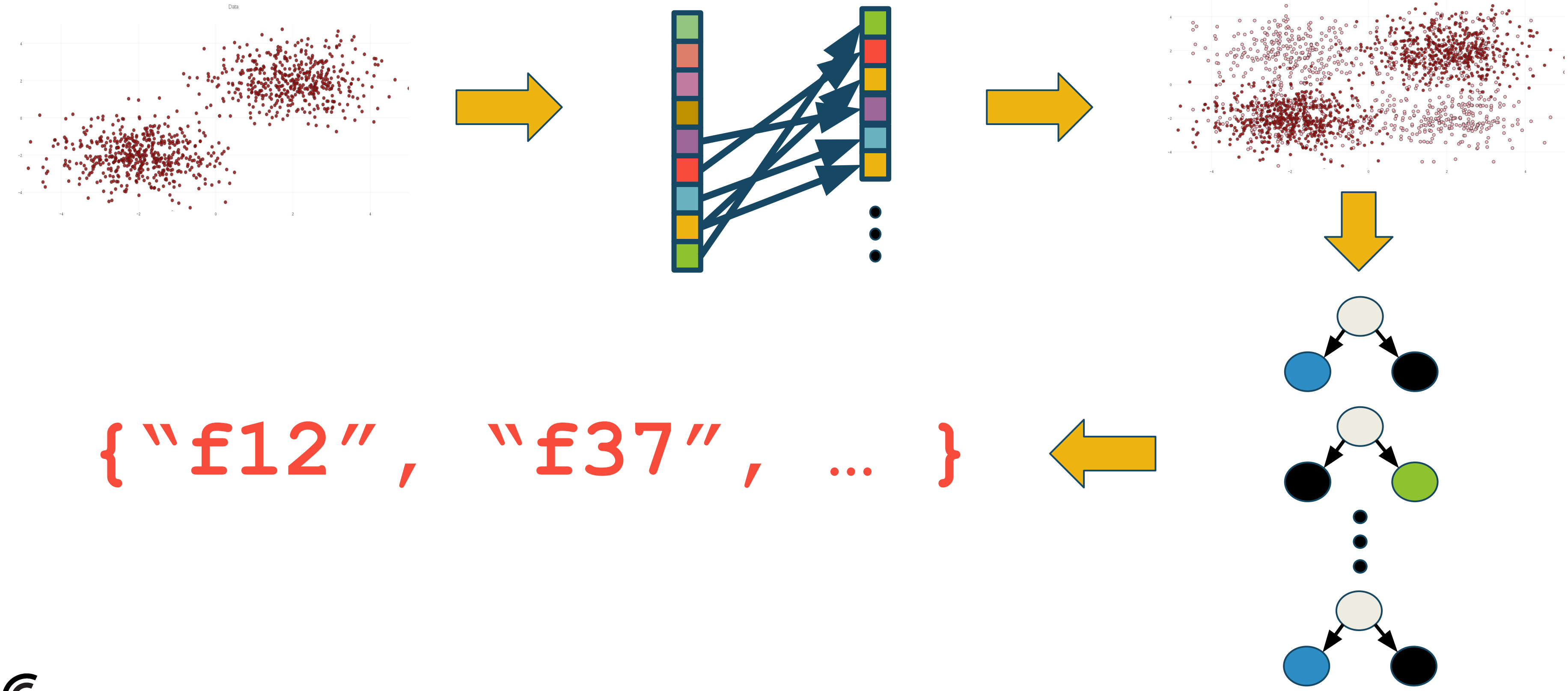
Learn Real vs Fake!



Random Forest Clustering

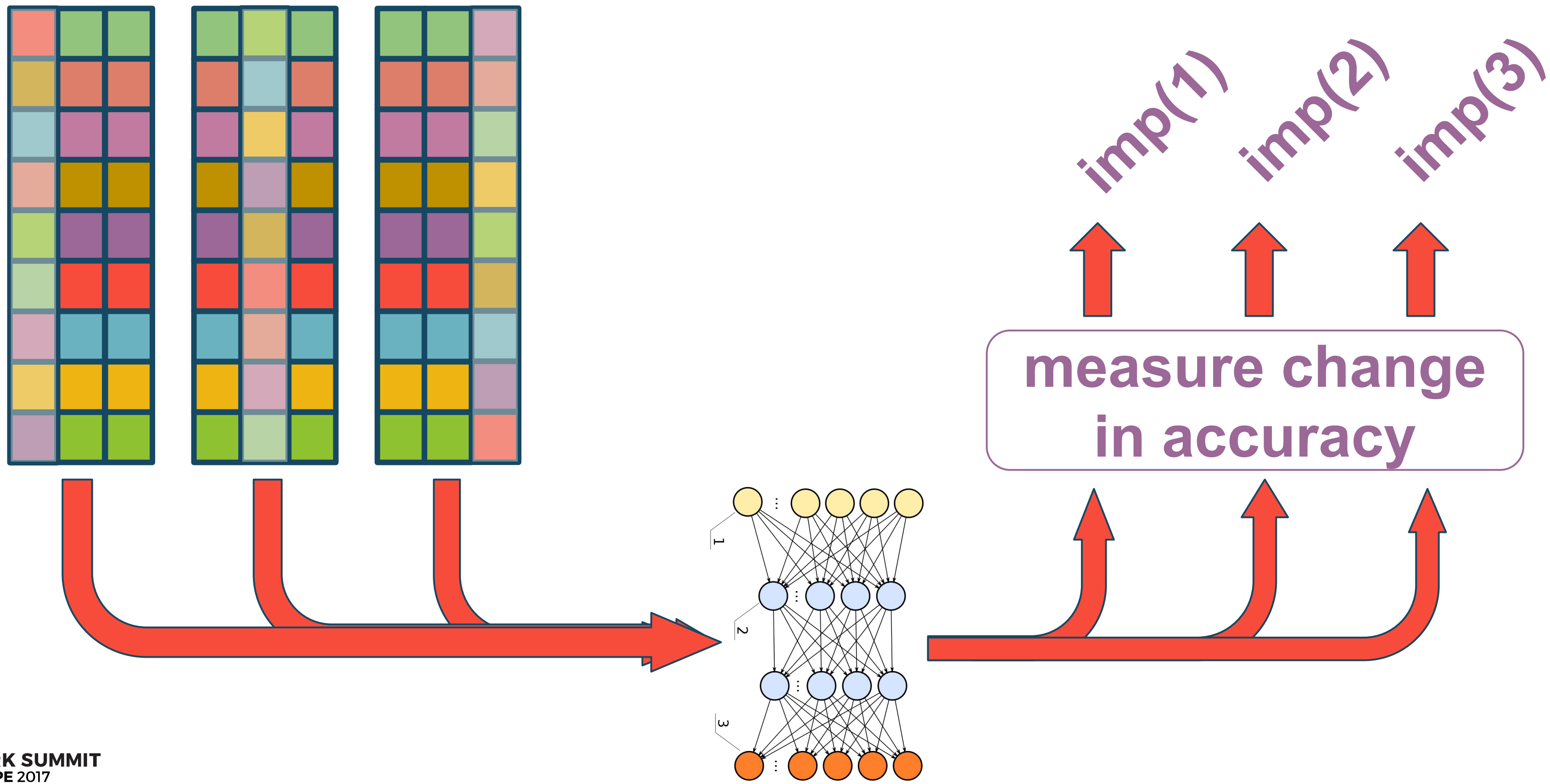


Feature Reduction

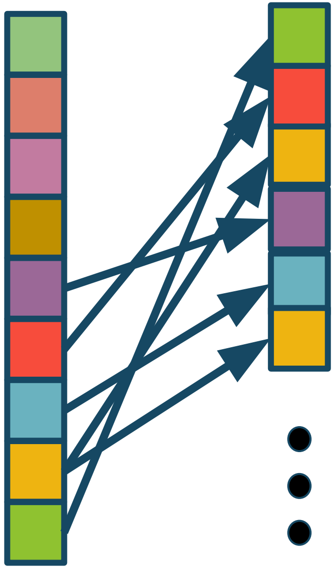


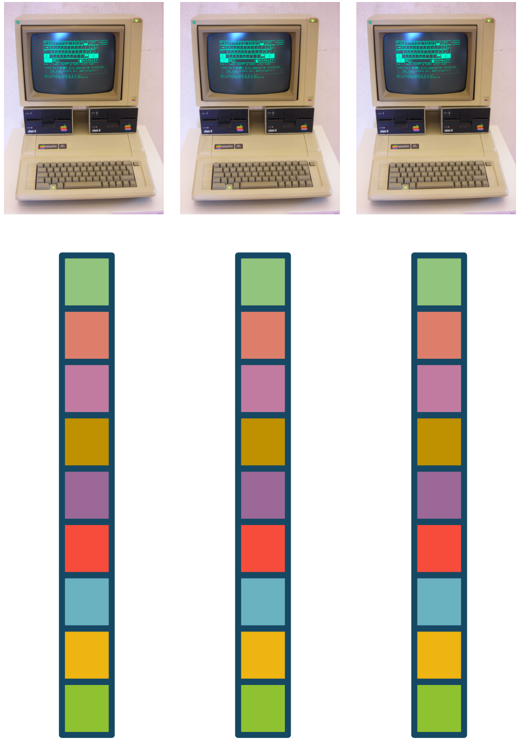
{"f12", "f37", … }

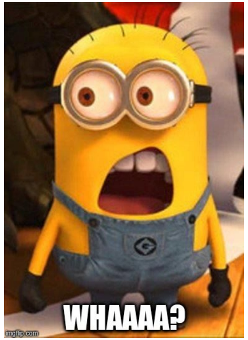
Feature Importance



What If Data Is Partitioned?







T-Digest

- **Computing Extremely Accurate Quantiles Using t-Digests**
- Ted Dunning & Omar Ertl
- https://github.com/tdunning/t-digest
- Implementations in Java, Python, R, JS, C++ and Scala
- UDAFs packaged for Spark and PySpark

What is T-Digest Sketching?

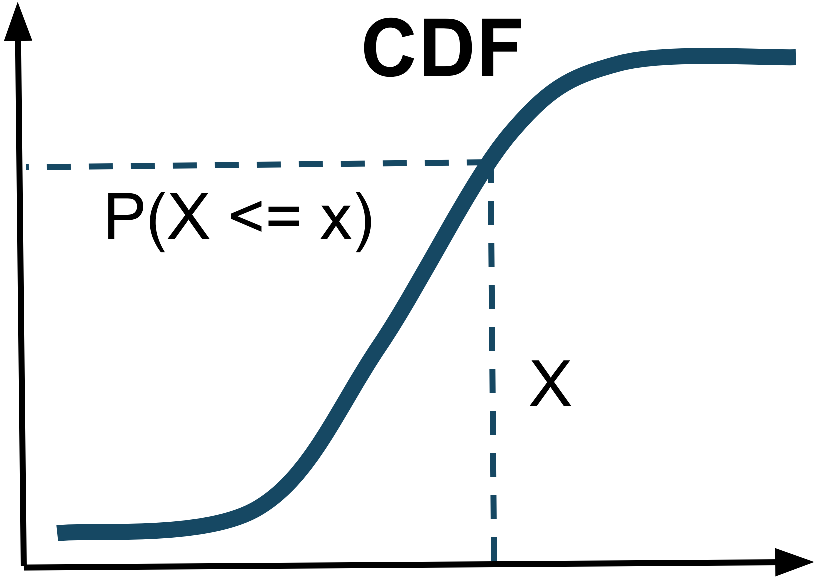
3.4

6.0

2.5

⋮

Sketch of CDF



Data Domain

Incremental Updates

Current T-Digest

+

x

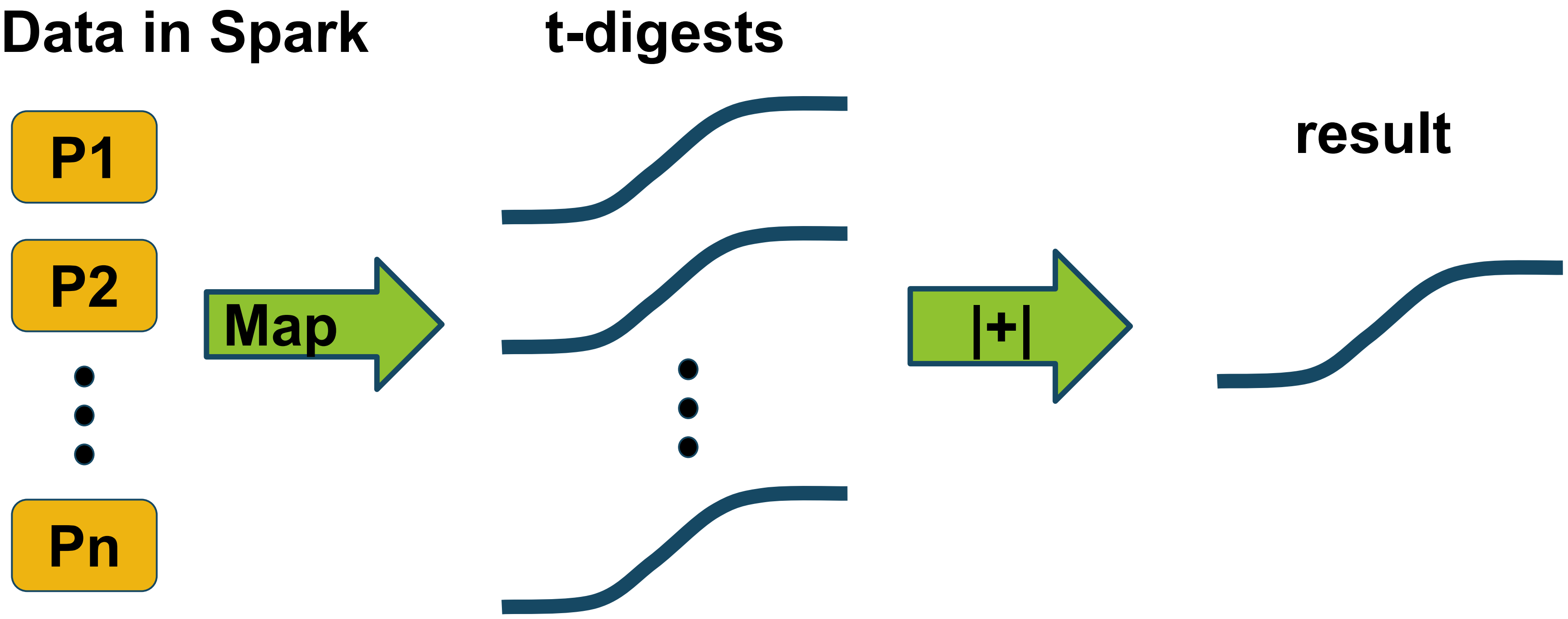
=

Updated T-Digest

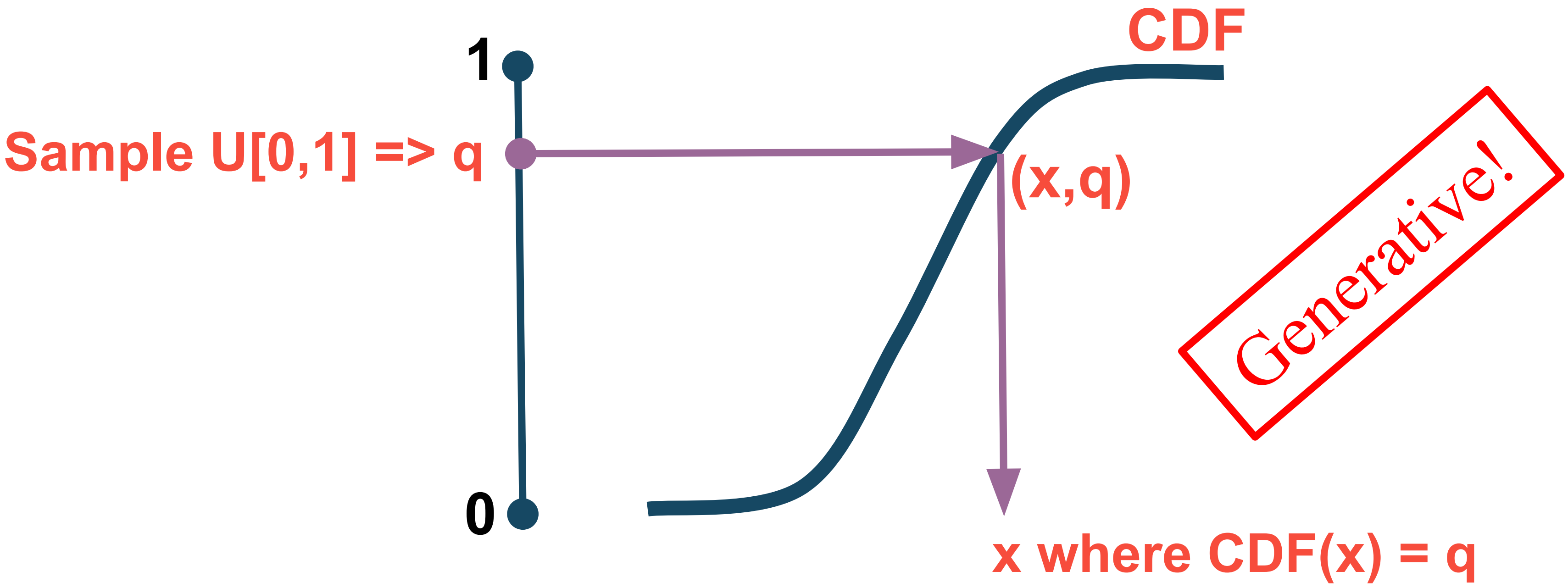
Large or Streaming Data

Compact "Running" Sketch

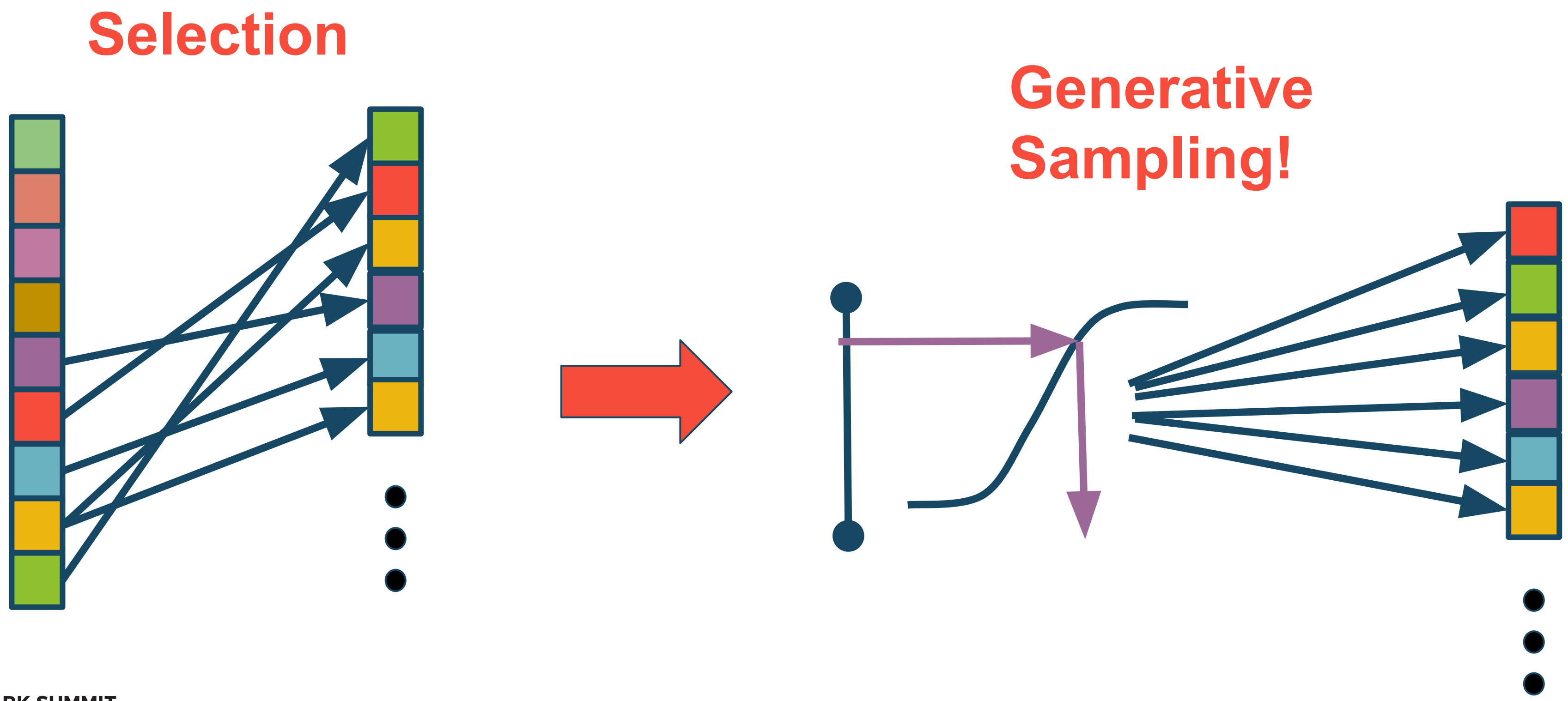
T-Digests Can Aggregate



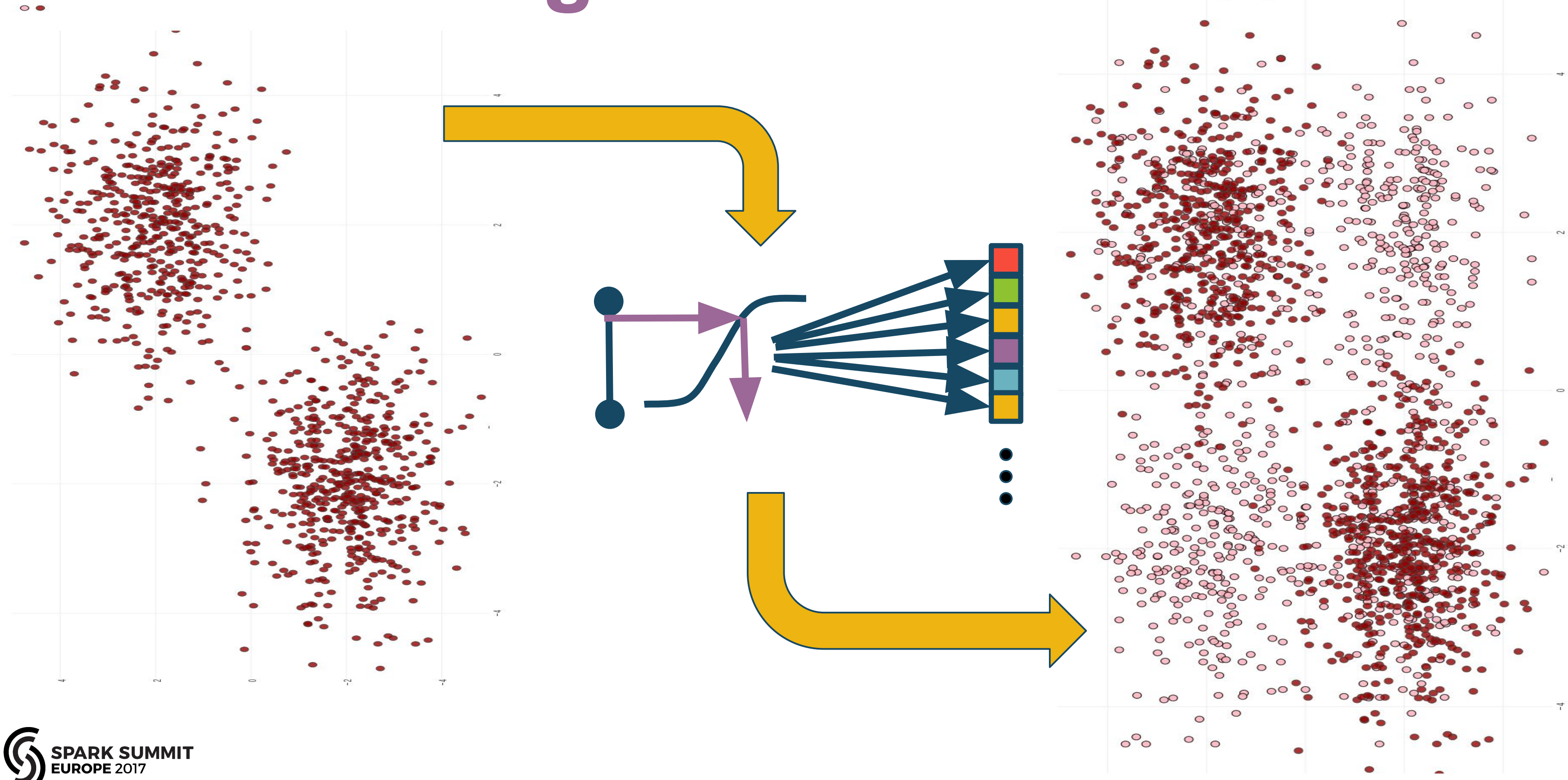
Inverse Transform Sampling (ITS)



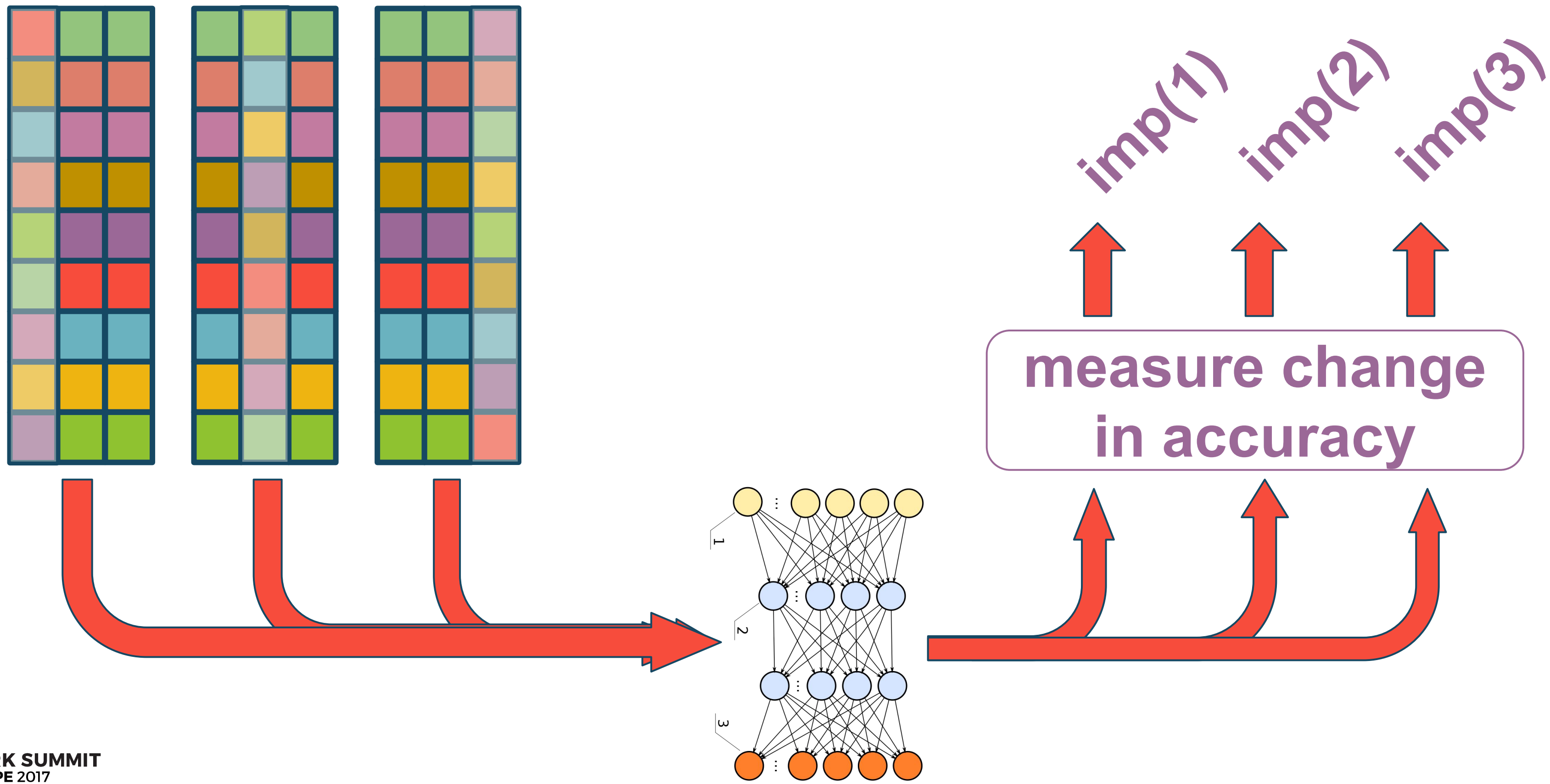
Random Selection => ITS



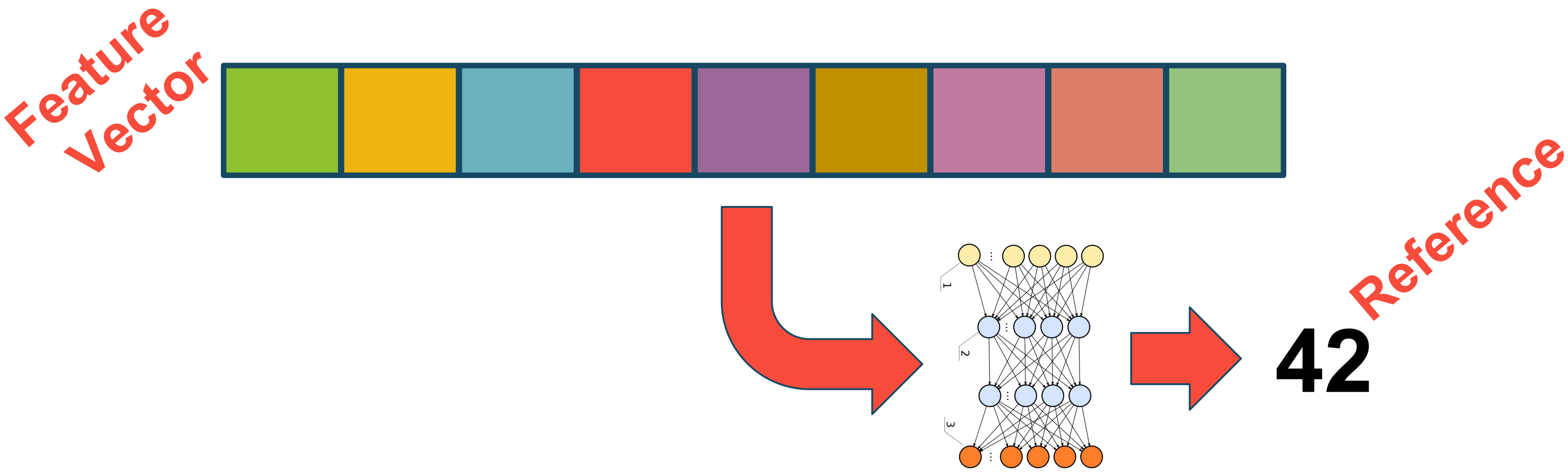
RF Clustering & Feature Reduction



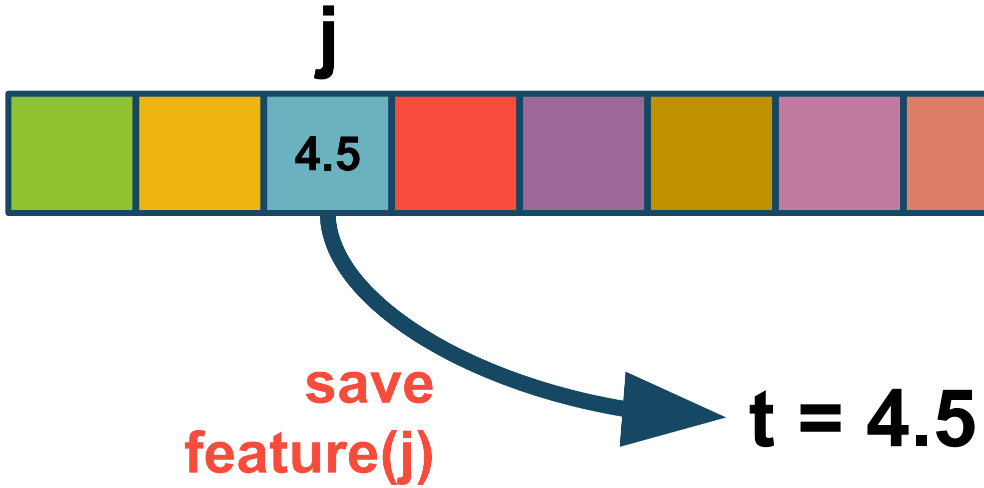
Feature Importance



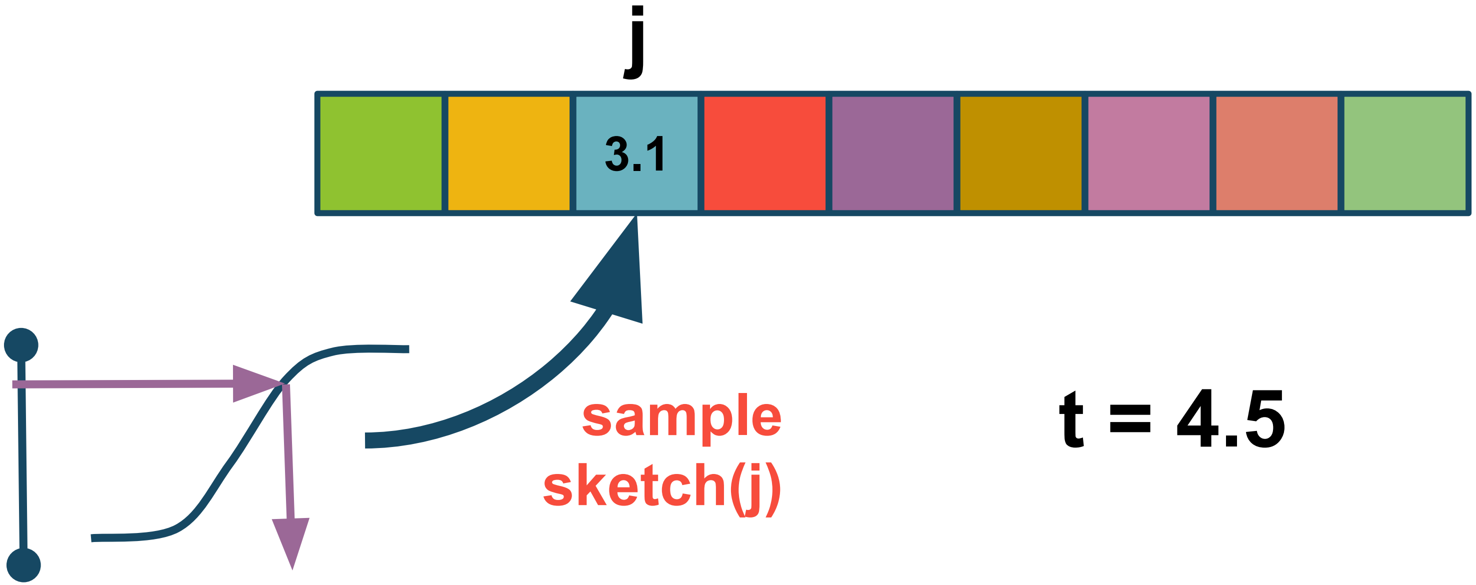
Feature Importance



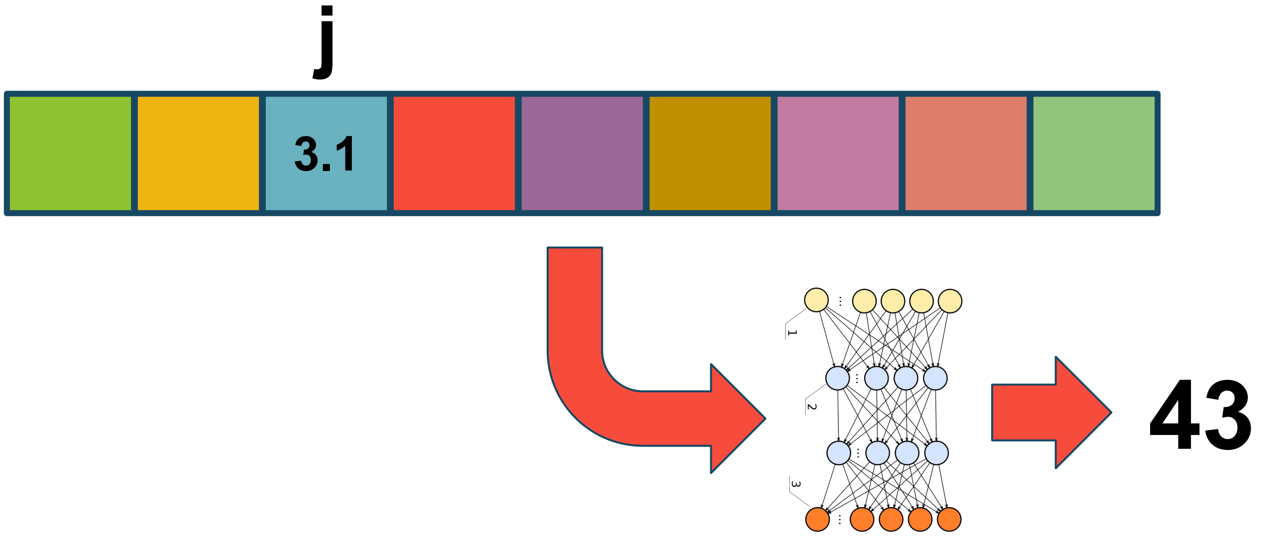
Feature Importance



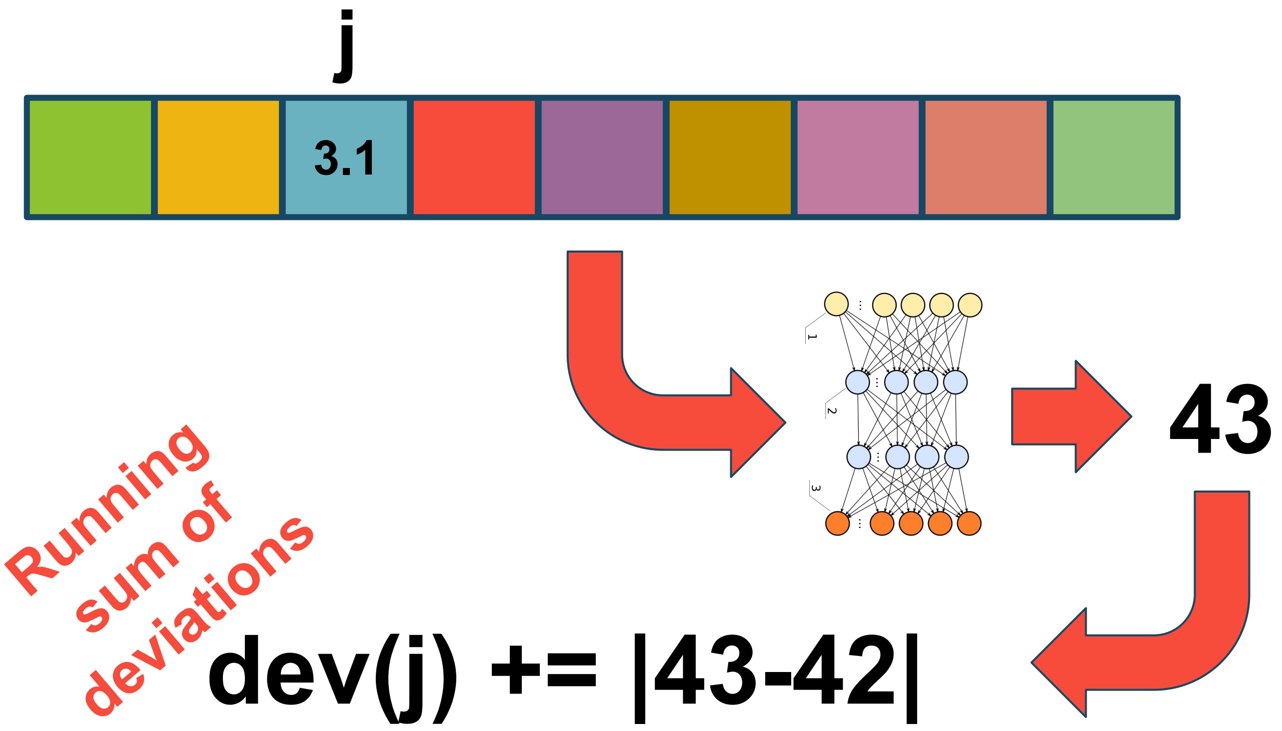
Feature Importance



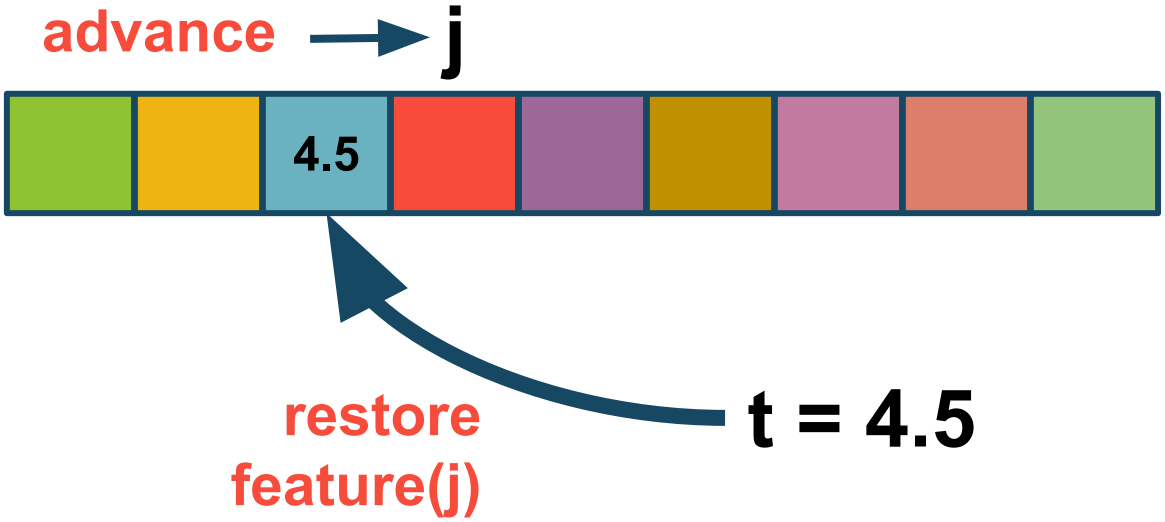
Feature Importance



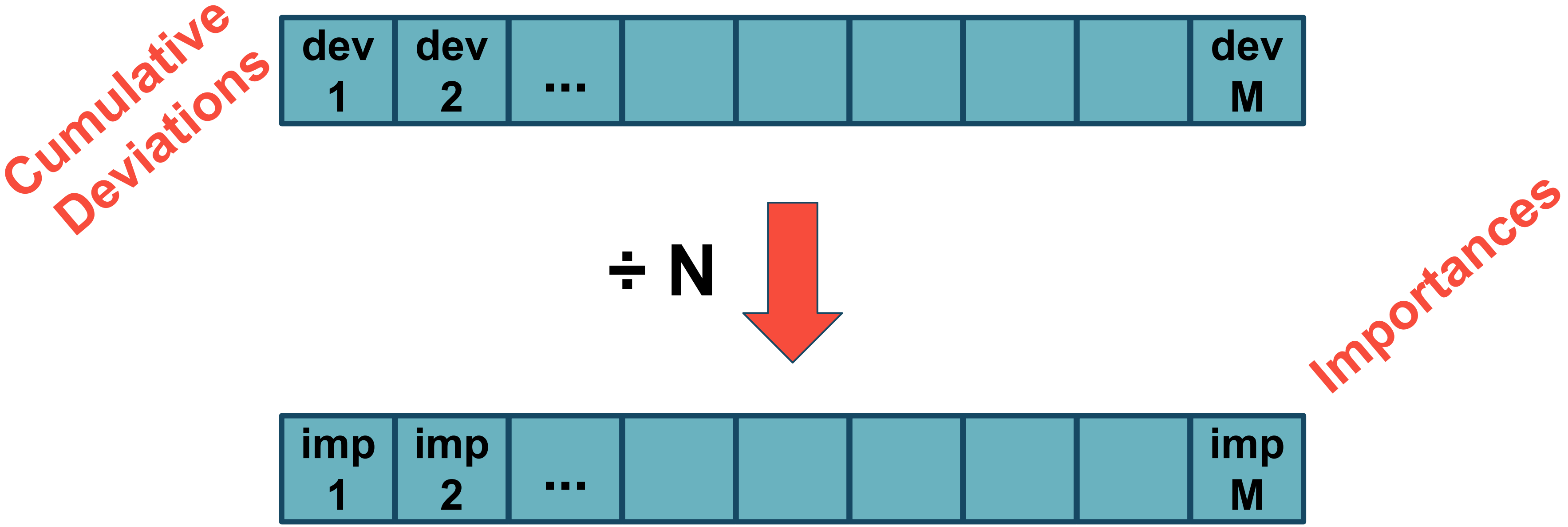
Feature Importance



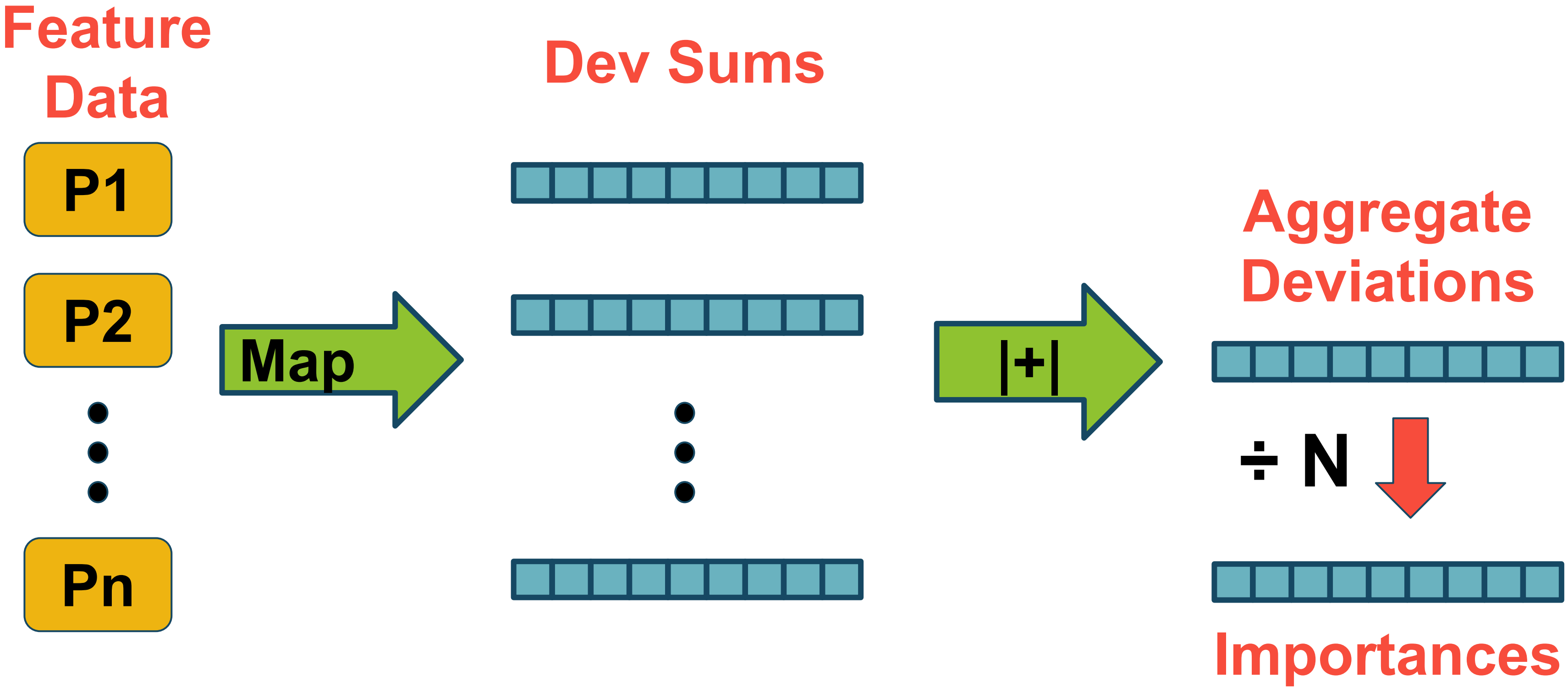
Feature Importance



Sum of Dev ÷ N = Importance



Deviations can Aggregate



One-Pass Feature Importance

Linear in Samples and Features

Single Pass over the Feature Data

Parallel over Data Partitions

Tox21 Data

National Institute of Health (2014)

12 Toxicity Assays, 800 “dense” features

12060 compounds + 647 hold-out

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

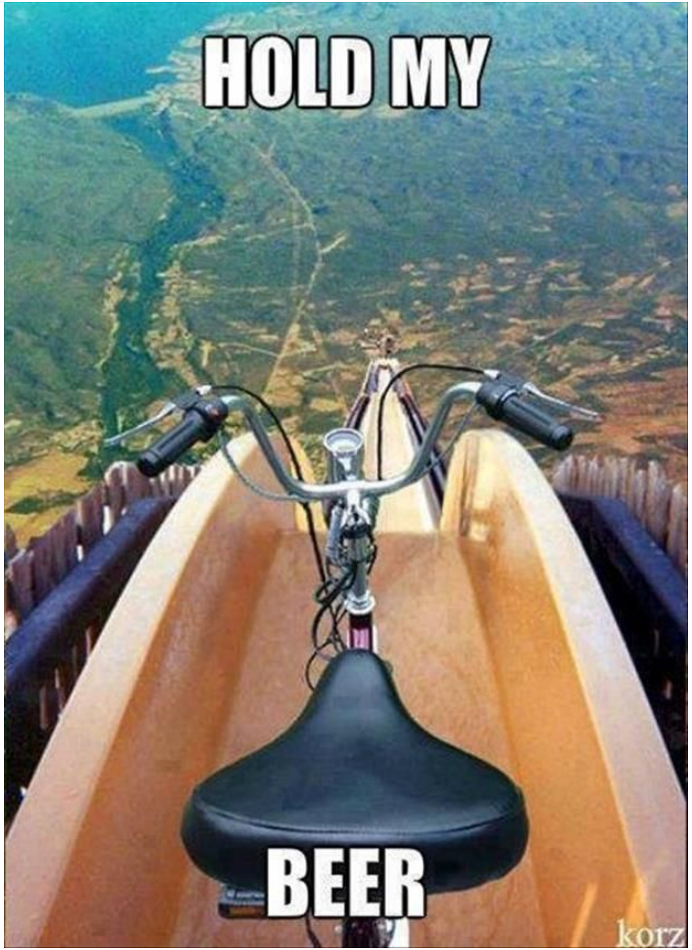
Johannes Kepler University Linz

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.

Demo



Explore

- Building ML Algorithms on Apache Spark
- Sketching With T-Digests
- Random Forest Feature Reduction
- Random Forest Clustering for Spark
- T-Digests and Feature Importance for Spark
- Demo Notebook for This Talk

Thank You!

eje@redhat.com
@manyangled
https://github.com/isarn/isarn-sketches-spark

#EUds11