SPARK ML WITH HIGH DIMENSIONAL LABELS

Michael Zargham, Director Data Science
Stefan Panayotov, Senior Data Engineer
Cadent

Cadent: Data Empowered Television Advertising

Data Technology Company specializing in Television Advertising

Cadent has built a **bicoastal** data science and engineering team

Vision

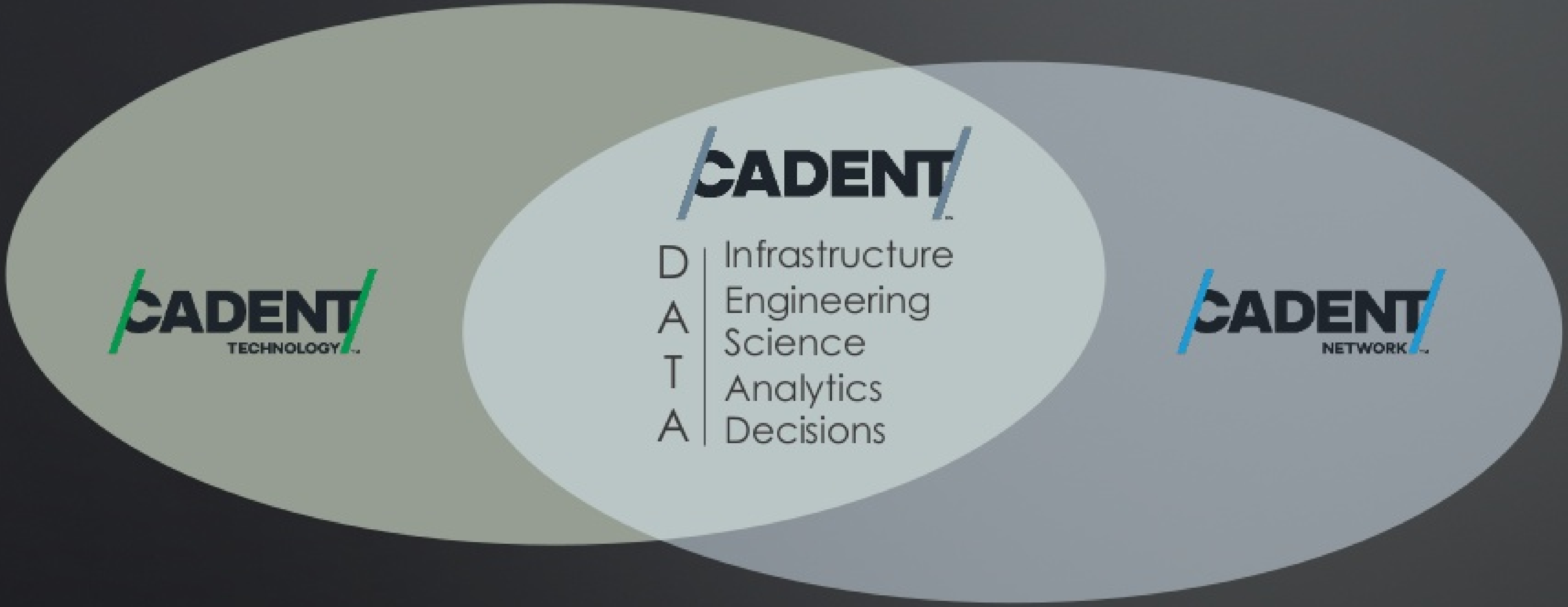
Unified Self Service Media Monetization Platform for all TV inventory

Our Team has cutting edge expertise

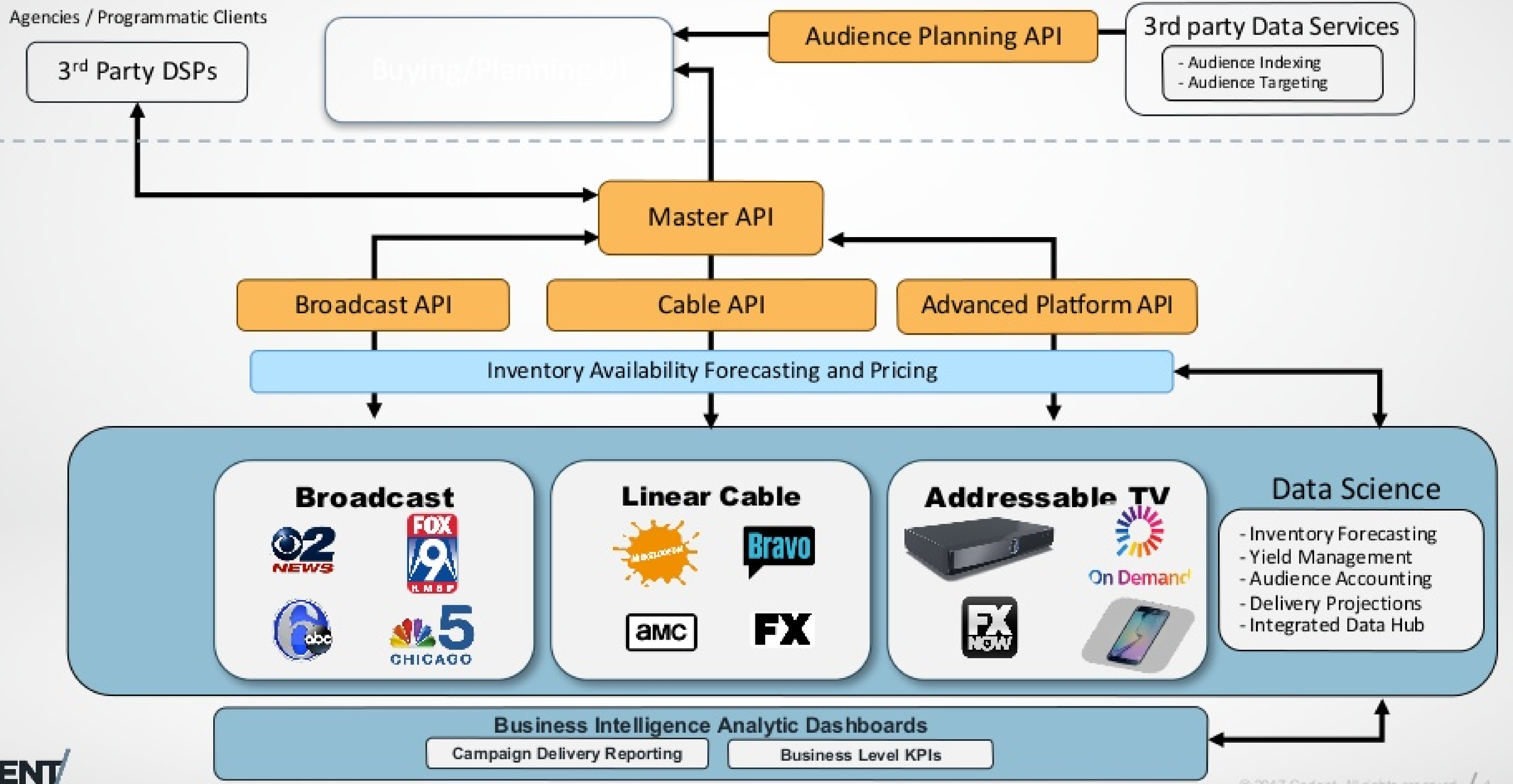
- Hybrid cloud Apache Spark infrastructure
- Analytical rather than rule driven algorithms
- Experience with Machine Learning APIs and custom mathematics in decision optimization

Cadent: Data Empowered Television Advertising

Data Technology Company specializing in Television Advertising



Cadent TV Buying Platform

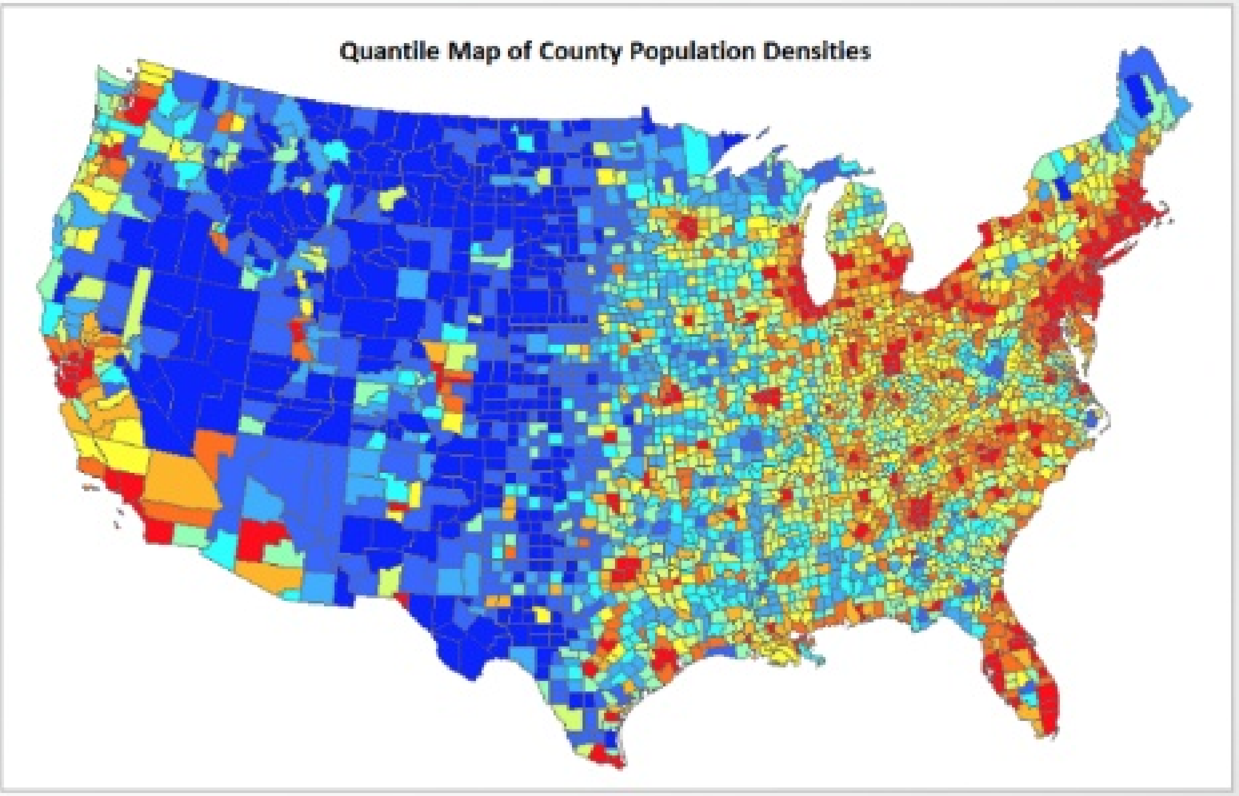


CADENT

Zoom In: Local Linear Cable

- Agency
  - Buys Media to Run ads for National advertisers
- Impressions
  - “eyeballs” currency of brand advertising
- Cable Operator
  - Media Companies providing TV service
  - Loosely segregated Geographically
- Subscribers
  - Consumers of Cable Operator Services
- Ratings
  - Fraction of Subscribers tuned in
  - Not known until after the fact
  - Ill conditioned: log-scale variance
  - O(10k) dimensions: variation in Demographics, geography and television content

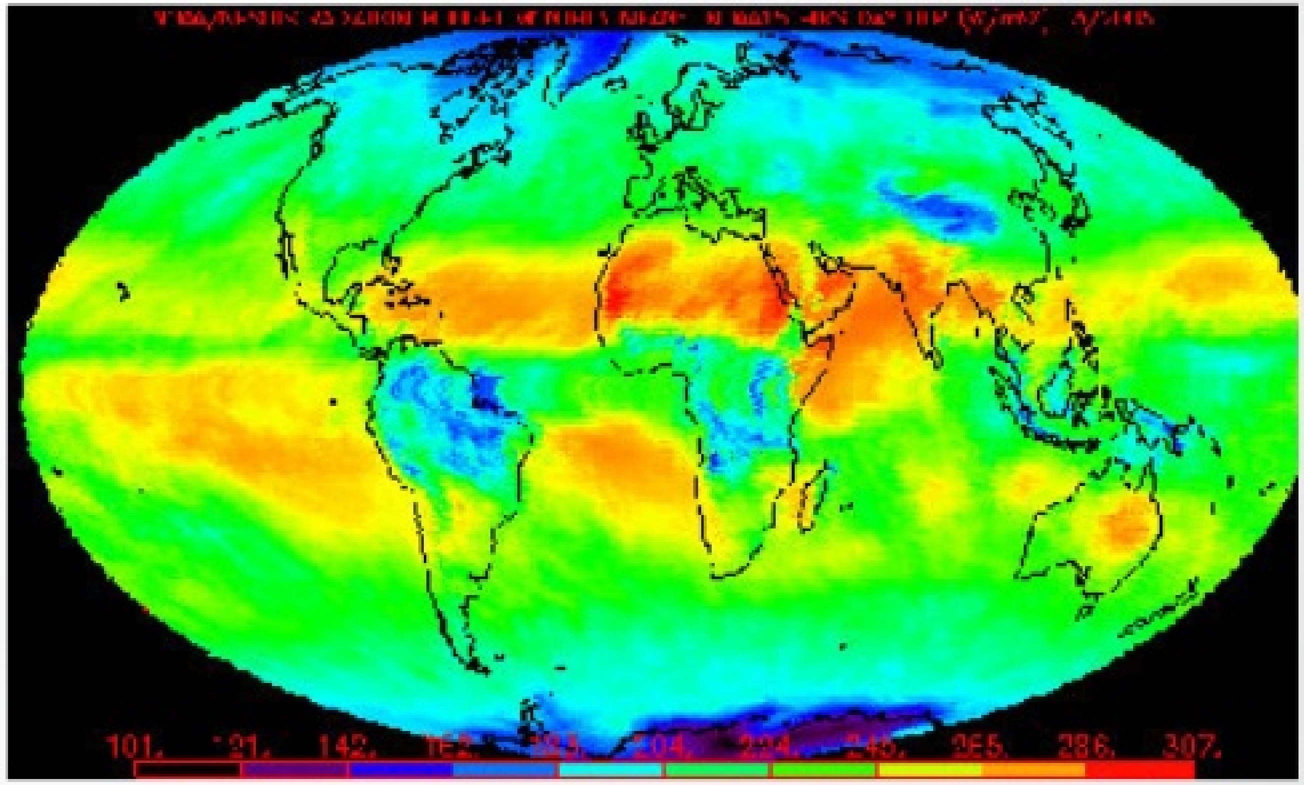
$$I_a = \sum_{g \in Geo} \sum_{n \in Net} \sum_{t \in time} R_{g,n,t,a} * S_{g,n,t,a}$$



CADENT

Models for Television Ratings

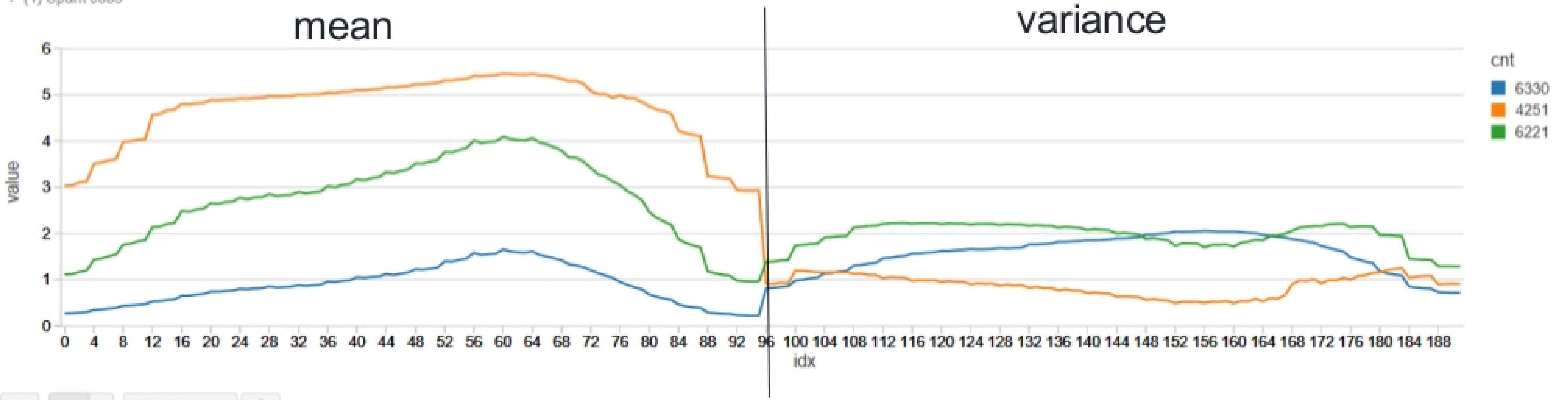
- Relevant Time Scales
  - Weather-like View
    - Shows
    - Twitter trends
    - Spectacle Events
  - Climate-like View
    - Seasonality
    - Subscriber trends
    - Daypart Variation
- Why High Dimensional?
  - In the climate view the features represent days but there significant intraday patterns
  - Pivot the daily pattern into vectors so that ML models can directly capture statistical correlations



A sense of Daily Patterns: Log Mean & Variance

```
> %sql select * from Vis_Grouping order by groupID, idx
```

▸ (1) Spark Jobs



Plot Options...

Command took 2.63 seconds -- by spanayotov@crossmw.com at 12/20/2016 11:53:27 AM on cte-spark2

Values shown in Log-like coordinate system:

value 0 = rating 0
value 3 = rating 10^(-5)
value 5 = rating 10^(-3)

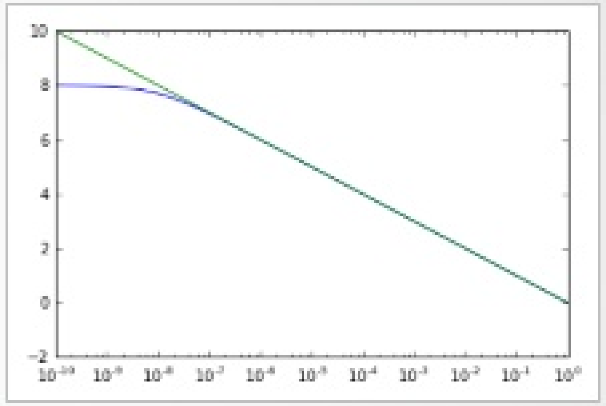
CADENT

An Intuitive Coordinate System for Human Interpretation of Ratings Data

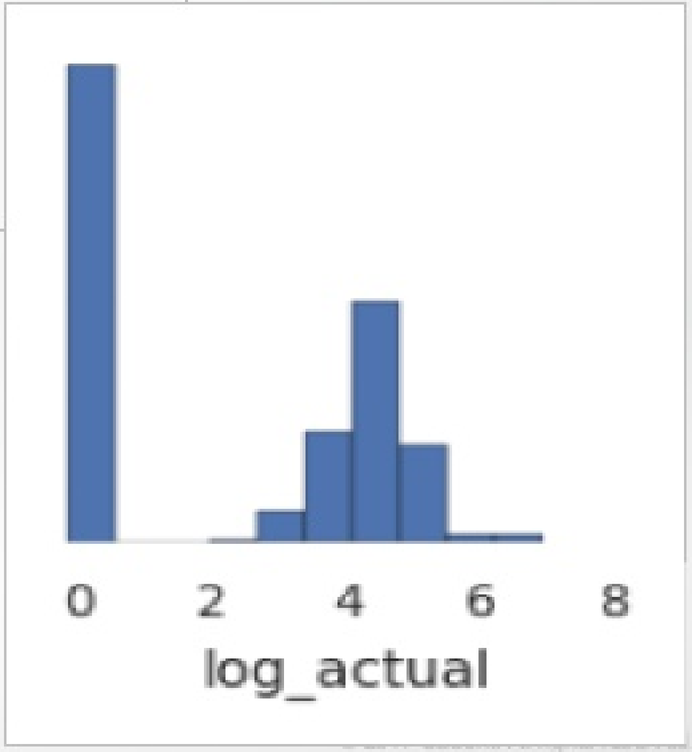
```
// *** This creates the log functions ***
def symlogGenerator(base: Double, offset: Double, flipSign: Boolean = true) = {
  val flipInt = flipSign match {case true => 1 case _ => 0}
  val sign = 1 - 2 * flipInt // value is always 1 or -1
  def logbase(lbase: Double, num: Double) = log(num)/log(lbase)
  val logOffset = - logbase(base, offset) * flipInt // log(offset, base)

  def transform(x: Double) = logOffset - sign * logbase(base, min(1,max(0,x)) + offset) // log(x + offset, base)
  def inverseTransform(y: Double) = pow(base, -(min(logOffset,max(0,y)) - logOffset) * sign) - offset

  (transform(_), inverseTransform(_))
}
// This line actually creates functions from the generator
val (t,it) = symlogGenerator(10, pow(10.0, -8.0), true)
sqlContext.udf.register("T",t)
sqlContext.udf.register("IT",it)
```



This coordinate system is used to eliminate bias in error metrics,
In the domain the errors in large value ratings swamp those of small value ratings



CADENT

DISCUSSION OUTLINE

01 THEORY

- Math

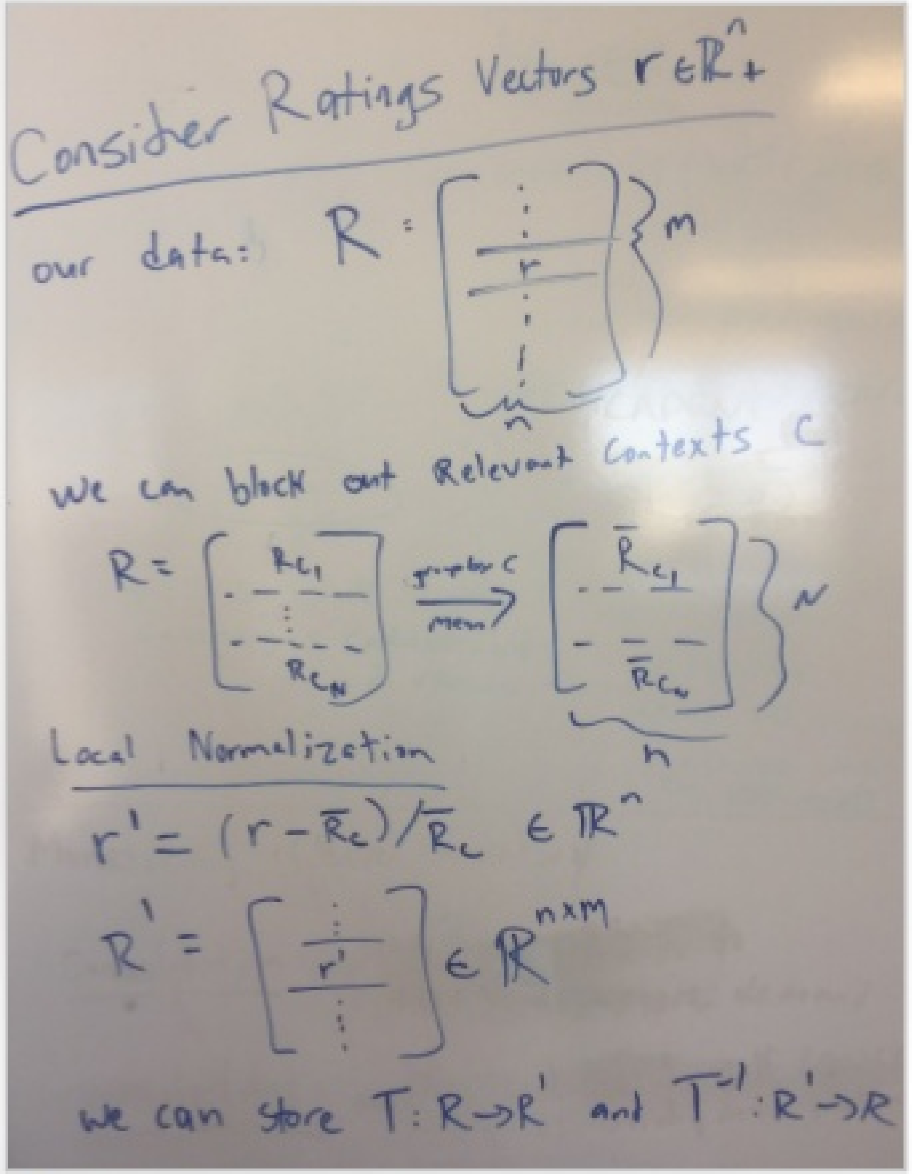
02 PRACTICE

- Code

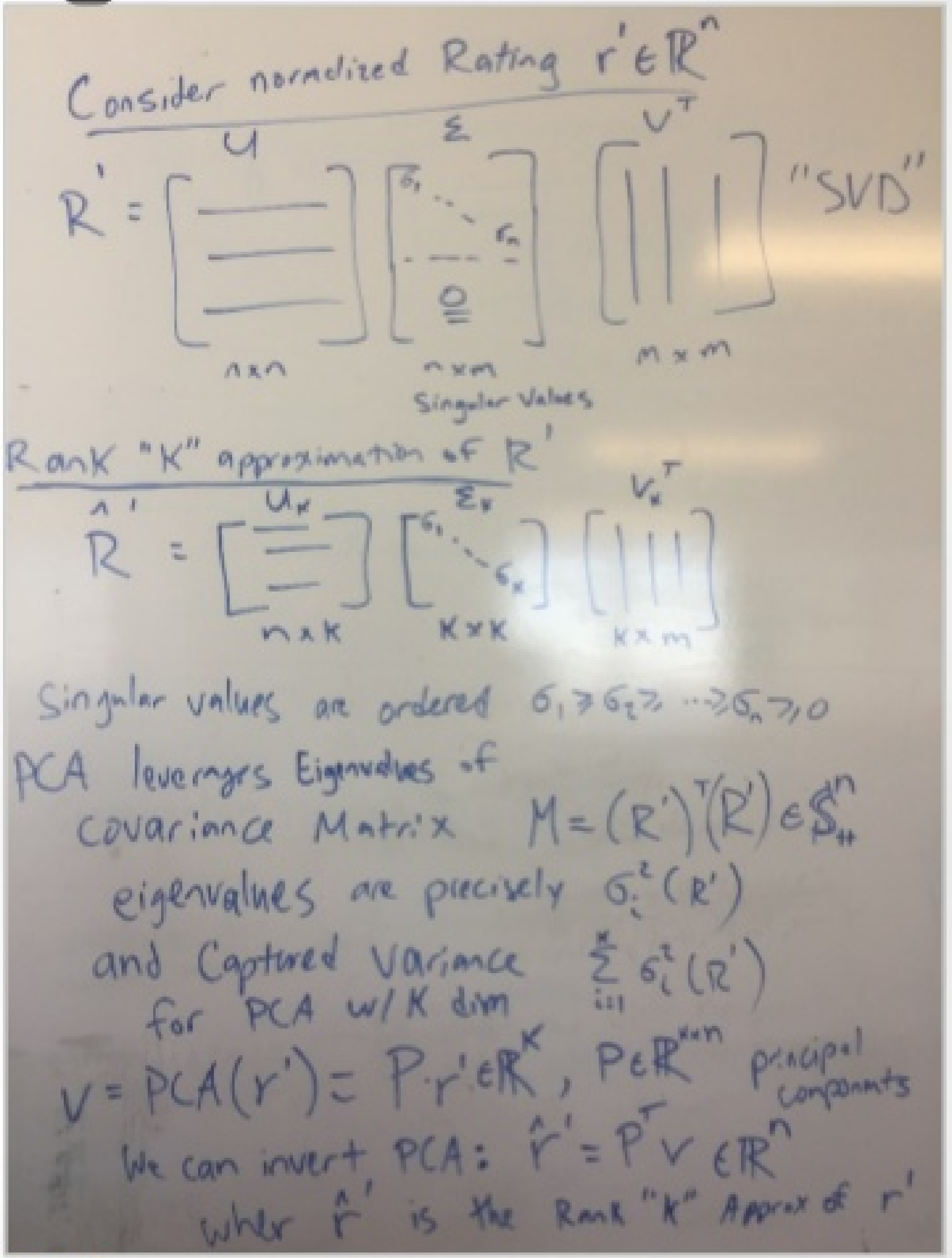
THEORY: Reframe the Problem with Math

Mathemagic… AKA Linear Algebra

Local Means as a Linear Transform

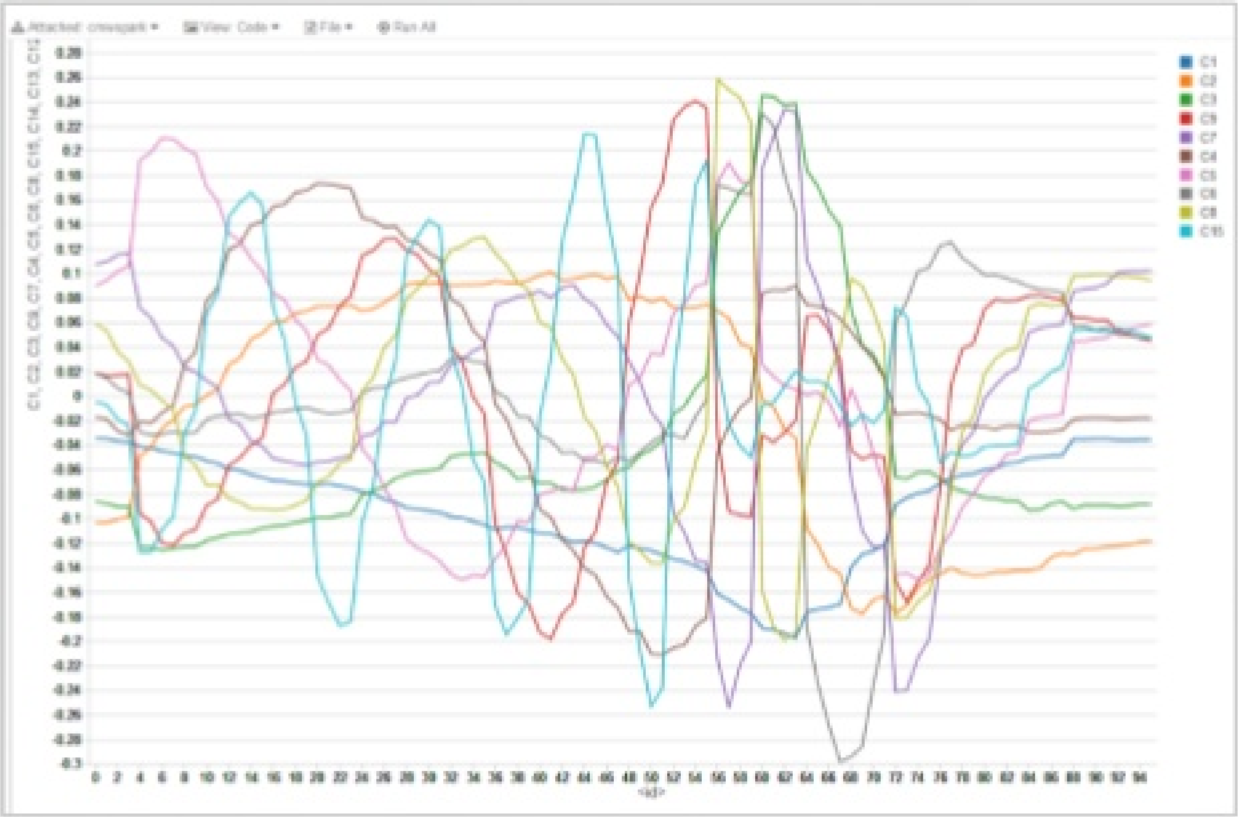


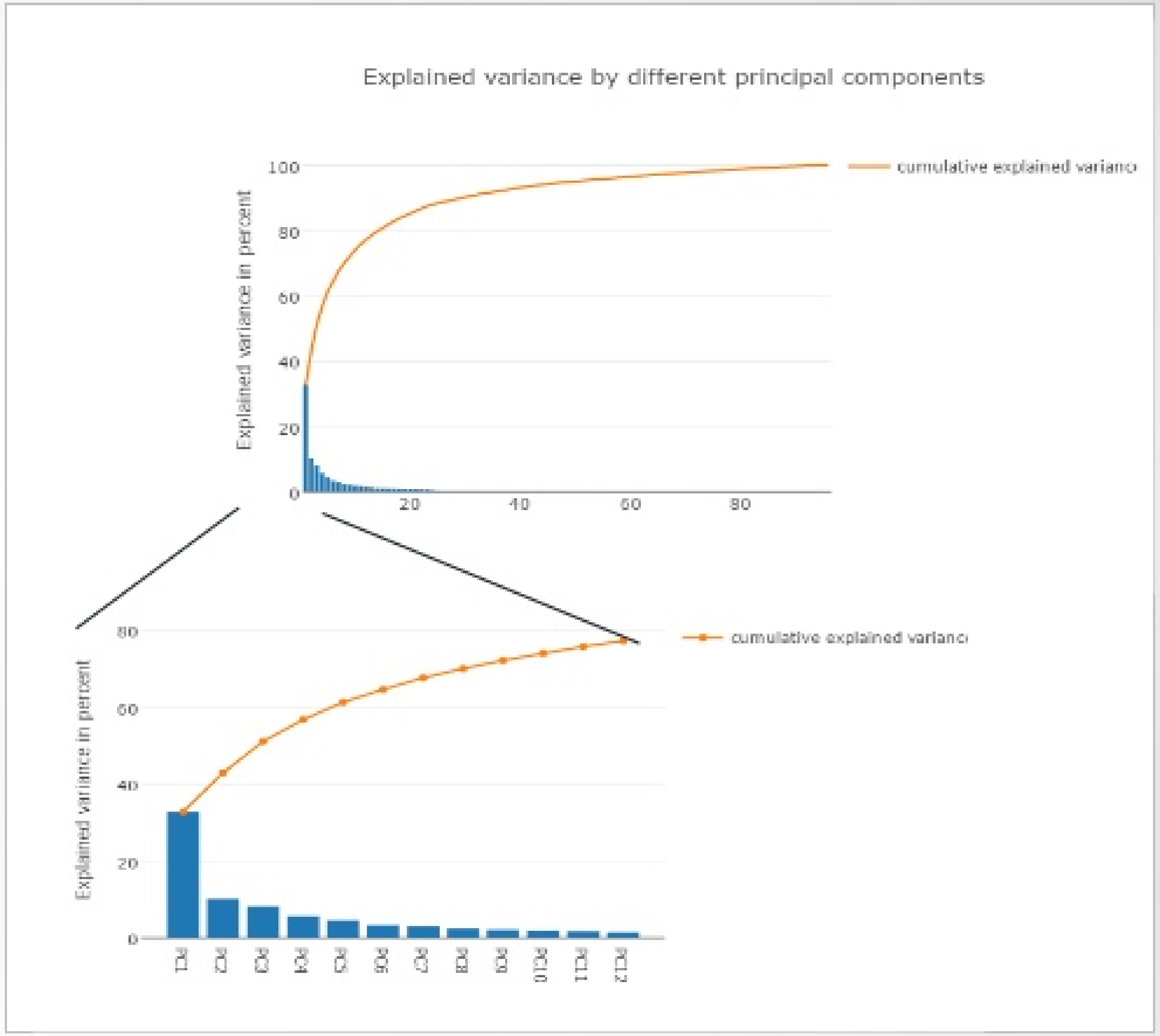
Interpreting PCA as Change of Basis



PCA reference: https://arxiv.org/pdf/1404.1100.pdf

Principal Components & Captured Variance





Warning:

Uncaptured Variance is strictly lost from the predictive model

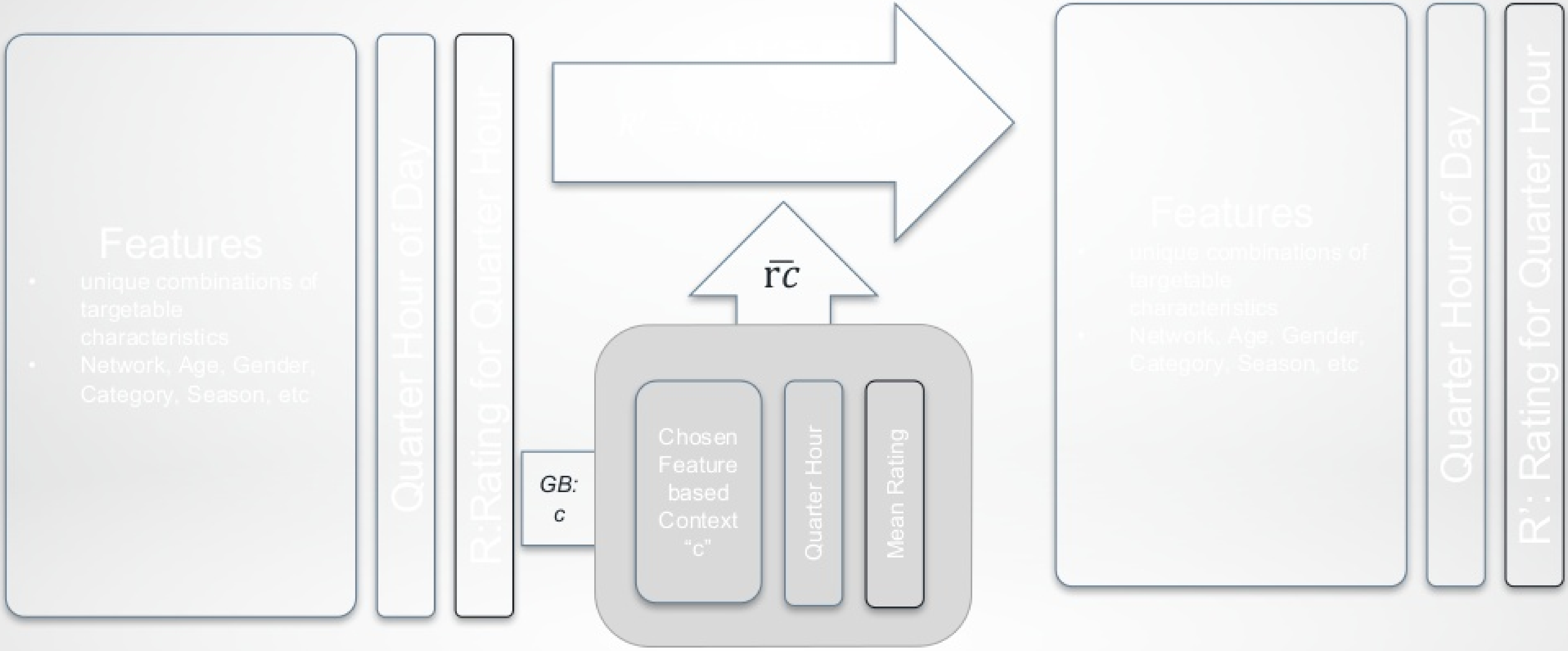
CADENT

Why Reduce Label Dimension?

- The noise reduction and correlations between values captured by reducing to principal components adds more value than variance lost
- Apache Spark ML API doesn't support n-Dimensional regression so k dimensional regression is computationally efficient for k<<n
- Since the 95% of the variance is captured by only the first few principal components there is little to no loss in modelling accuracy (we'll come back to this)
- Independent Component Analysis (ICA) would be even better than PCA because value chained regressors are treated as independent variables. ICA not yet available in Spark ML

CADENT

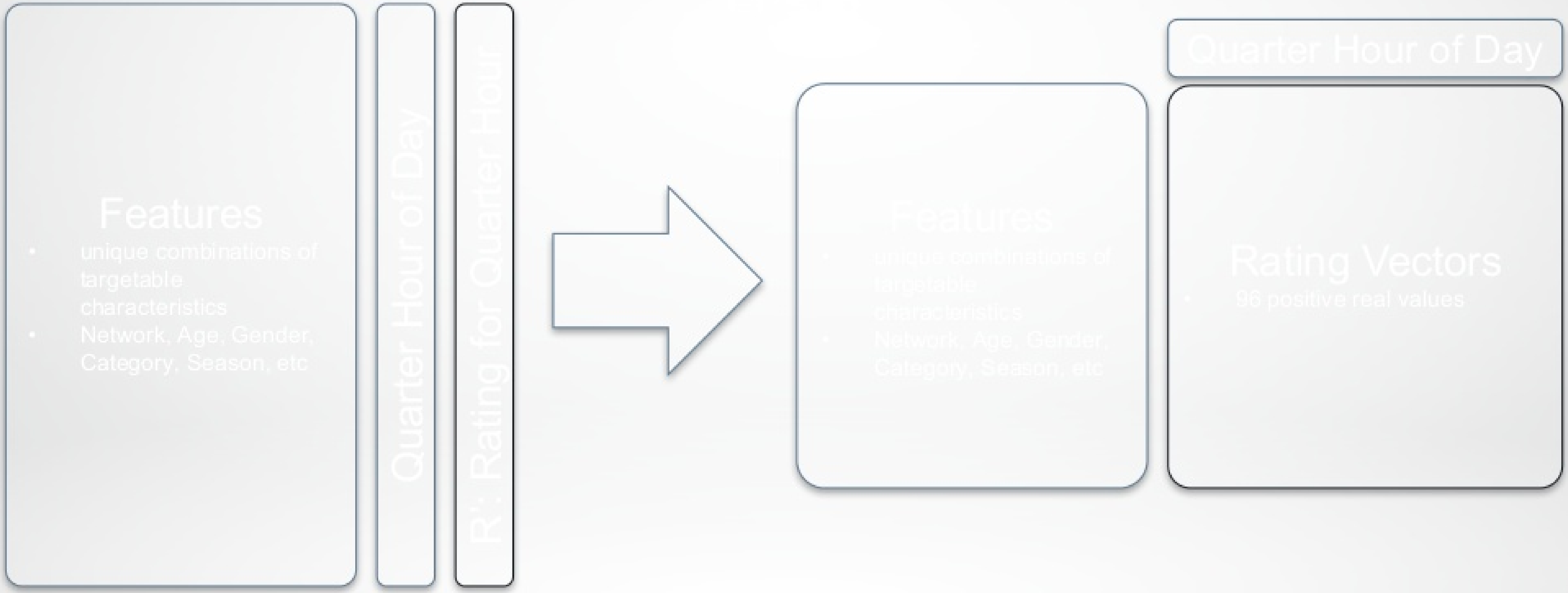
Local Means Coordinate Transform



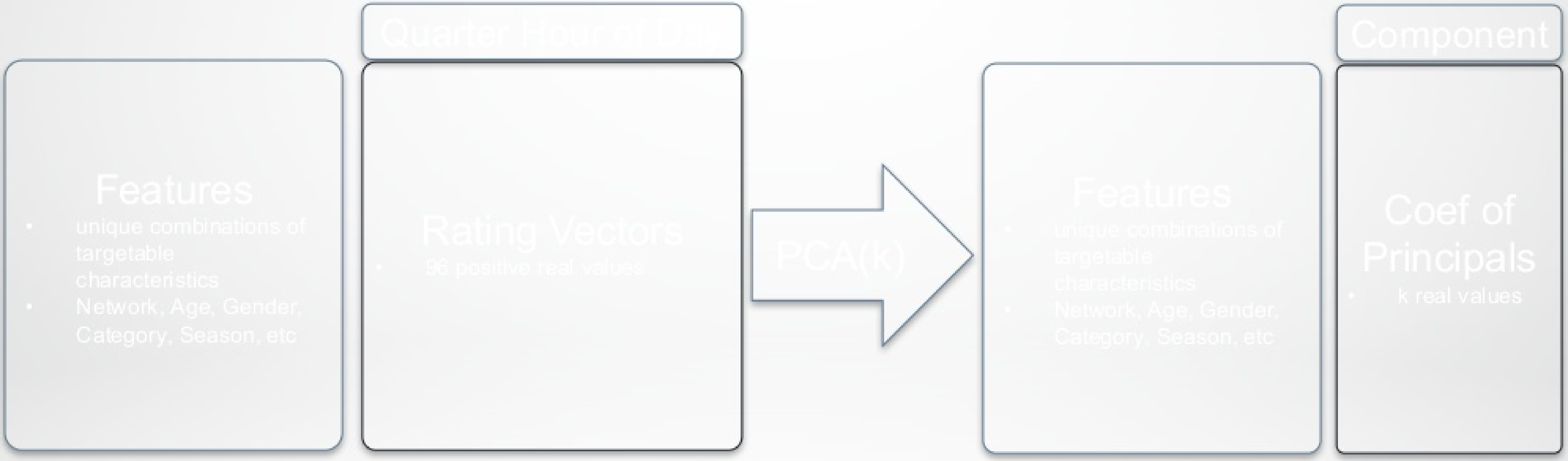
Store values $c: \overline{\mathrm{r}c}$

CADENT

Pivot our Data into to Vectors (day profiles)

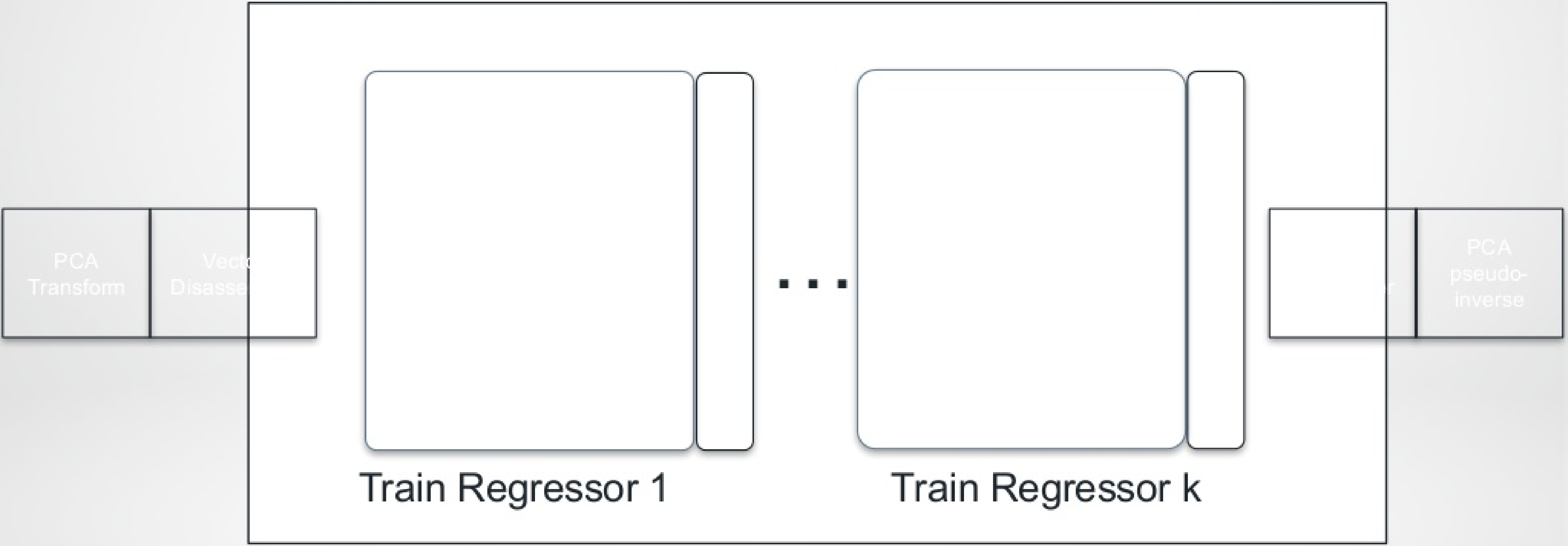


Label Dimensionality Reduction



CADENT

Principal Component Analysis (PCA): to reduce the dimensionality of the problem

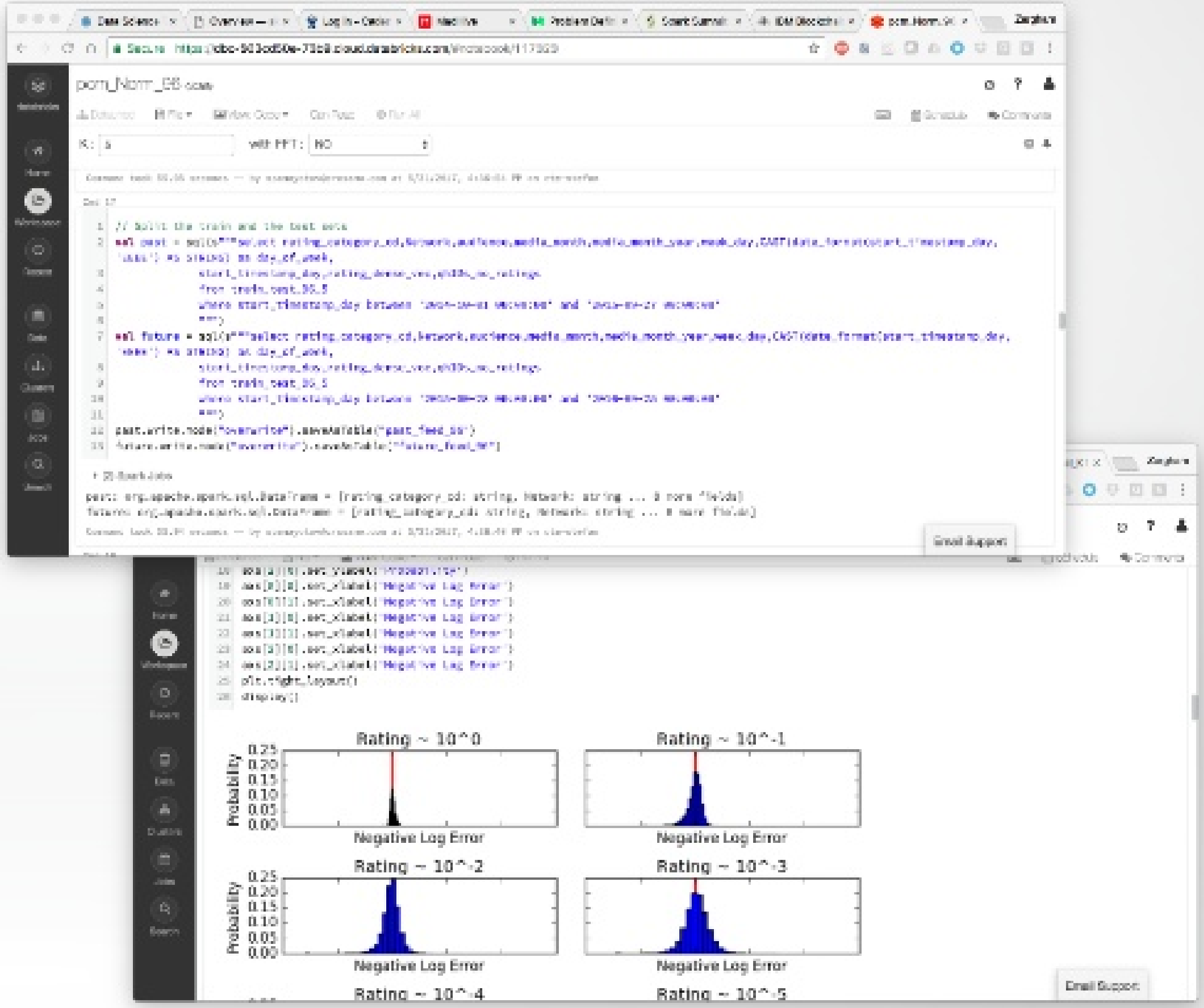


Pipeline of k single label regression models

PRACTICE: Code Demonstration

Technical Breakout to DataBricks Notebook

- Data preparation
  - Ratings Local Coordinate Transformation
  - Vectorization
- ML Pipeline creation and Execution
  - Custom Transformers and Estimators
    - DropColumnsStage
    - PCA2 (show the Scala and pyspark versions)
  - Custom UDFs
    - Used for Vector Disassembler
    - Used for Pseudo-inverse PCA
  - Train & Test
    - Undo custom Coordinate transforms to evaluate
- Results
  - Show Code for Model evaluation
  - Review some Graphical results
- Aside:
  - PySpark Version



CADENT

Preliminary Experimental Results

Data= ~6 Million (96 dimensional Vectors) and associated features

Data Size: 6.6 GB

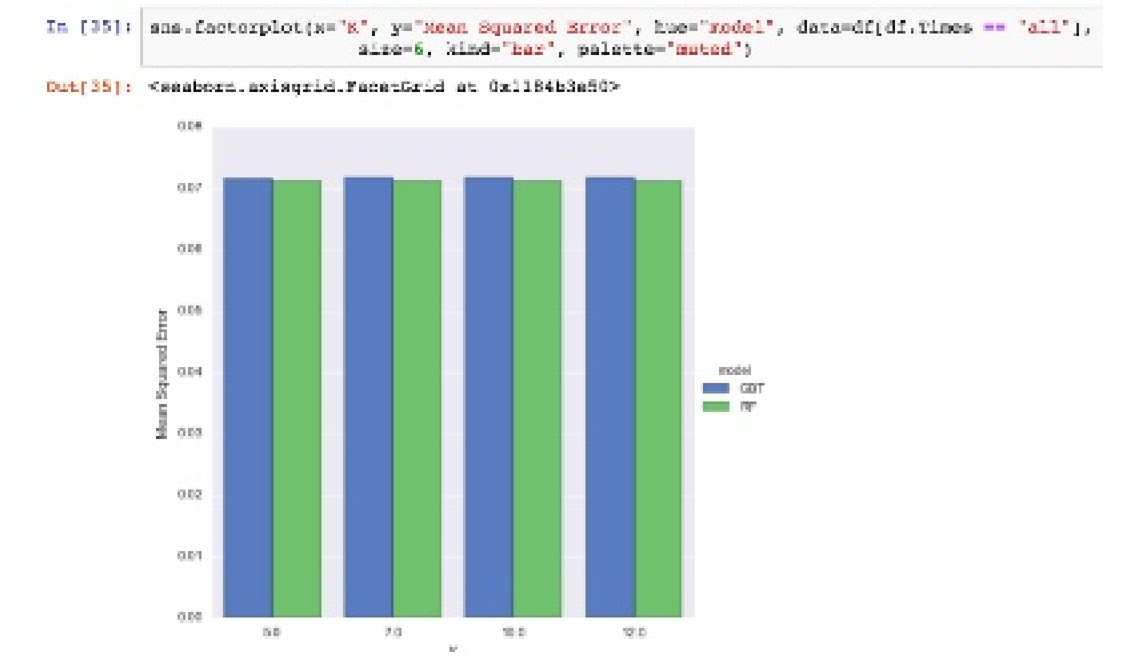
- 20 workers each with 60 GB RAM and 8 cores
- Auto-scaling on

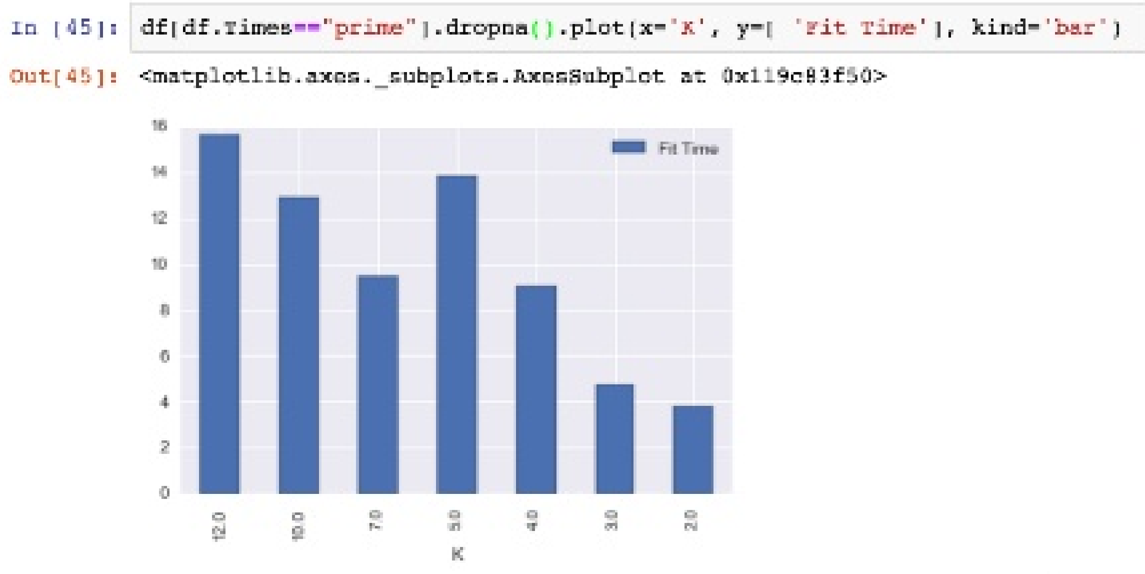
1	Type of model	Output table	Mean Absolute Error	Mean Squared Error	Fit run time	Save results
10	Prime time GBT K=12 net='TOTUS'	sdp_pcm_scatter_totusK12	0.051550236282	0.004443057429	5.84 min	13.59 sec
11	Prime time GBT K=12 norm	sdp_pcm_scatter2_normK12	0.148781770718	0.046845610136	15.66 min	28.36 sec
12	Prime time GBT K=10 norm	sdp_pcm_scatter2_normK10	0.148794481149	0.046891953196	12.95 min	26.36 sec
13	Prime time GBT K=7 norm	sdp_pcm_scatter2_normK7	0.148803329706	0.046886333718	9.47 min	20.08 sec
14	Prime time GBT K=5 norm	sdp_pcm_scatter2_normK5	0.148810023187	0.046904766631	13.88 min	17.24 sec
15	Prime time GBT K=4 norm	sdp_pcm_scatter2_normK4	0.148701796246	0.046768072815	9.02 min	24.31 sec
16	Prime time GBT K=3 norm	sdp_pcm_scatter2_normK3	0.148707072380	0.046826602934	4.73 min	11.05 sec
17	Prime time GBT K=2 norm	sdp_pcm_scatter2_normK2	0.148578846543	0.046695984573	3.80 min	12.09 sec
18	Prime time GBT norm no PCA	sdp_pcm_scatter2_normNoPCA	0.149369025754	0.047247526585	16.36 min	1.06 min
19	All time GBT norm K=12	sdp_pcm_scatter2_normAllK12	0.182586621908	0.071583207654	18.28 min	55.34 sec
20	All time GBT norm K=10	sdp_pcm_scatter2_normAllK10	0.182568372666	0.071576049525	13.52 min	30.75 sec
21	All time GBT norm K=7	sdp_pcm_scatter2_normAllK7	0.182582849778	0.071581692237	9.76 min	25.66 sec
22	All time GBT norm K=5	sdp_pcm_scatter2_normAllK5	0.182533610345	0.071509864673	7.33 min	23.81 sec
23	All time RF norm K=12	sdp_pcm_scatter2_normRFAllK12	0.182181487460	0.070974080041	10.41 min	2.09 min
24	All time RF norm K=10	sdp_pcm_scatter2_normRFAllK10	0.182190370610	0.070988125093	9.59 min	1.76 min
25	All time RF norm K=7	sdp_pcm_scatter2_normRFAllK7	0.182155097088	0.070941472845	5.86 min	45.28 sec
26	All time RF norm K=5	sdp_pcm_scatter2_normRFAllK5	0.182228954649	0.070978910337	3.87 min	1.17 min

CADENT

Insights from Initial Results

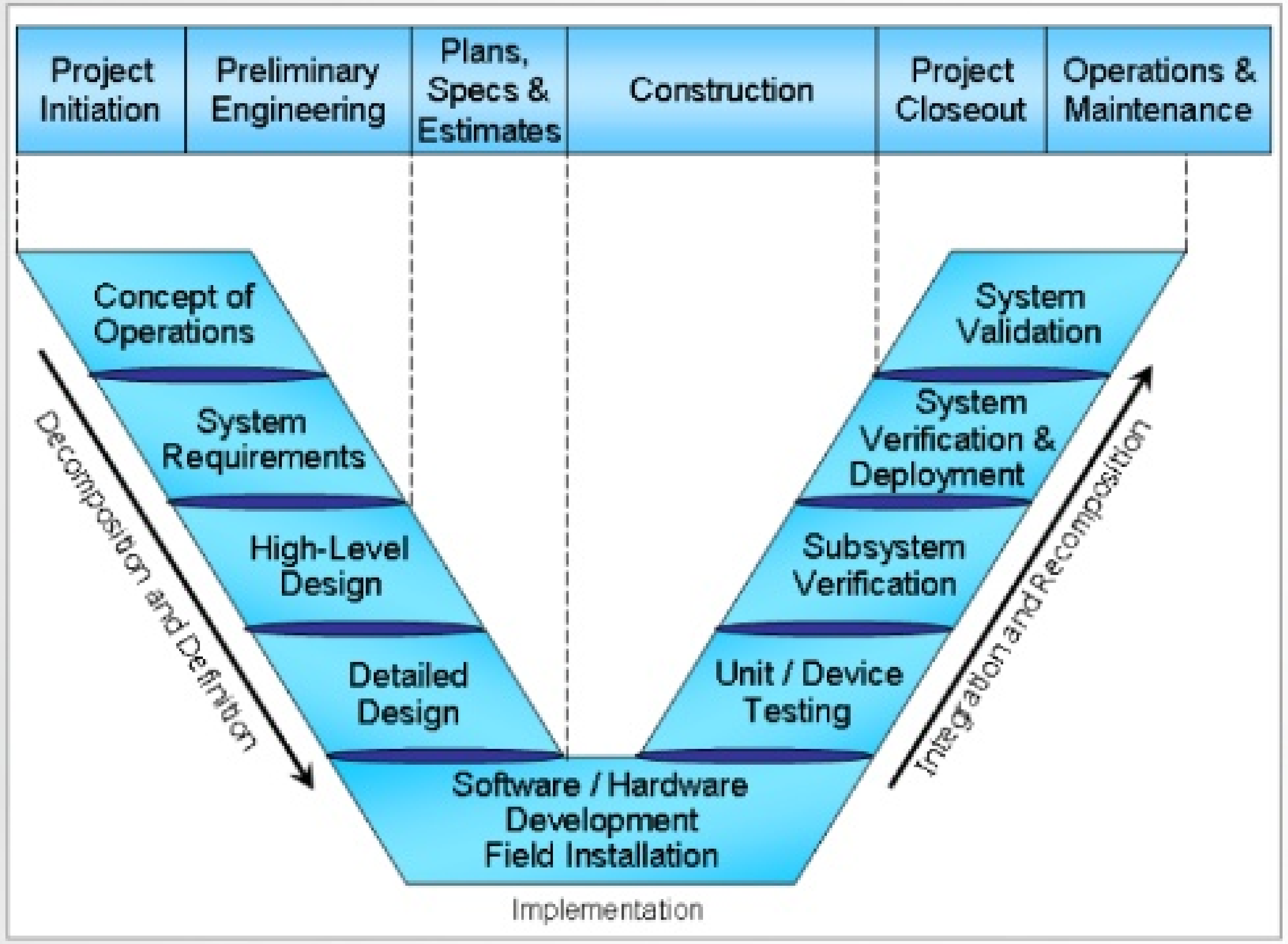
- Support for arbitrary Regression model class
  - Comparable results with Gradient Boosted Trees and Random Forests
- Daily Ratings Phenomenon is actually low Rank
  - No Measureable loss in accuracy as we reduce from K=12 to K=5 for the n=96 dimensional version
  - No Measureable loss in accuracy as we reduce from K=12 to K=2 for the n=12 dimensional version
- Reducing the number of dimensions saves runtime
  - More experiments are needed but preliminary results are still significant
- Inclusion of PCA when K=n provides no measureable improvement
  - Tested with K=n=12
  - Would like to test with ICA instead of PCA
- Next Steps:
  - More exhaustive repeated trials for run times
  - Further reduce K until we start to see performance degrade





Monitoring & Maintaining Machine Learning Models

Validation and Verification



IEEE 1012-2012 standard definition of Validation and Verification

Validation: The assurance that a product, service, or system meets the needs of the customer and other identified stakeholders. It often involves acceptance and suitability with external customers.

What did it do?

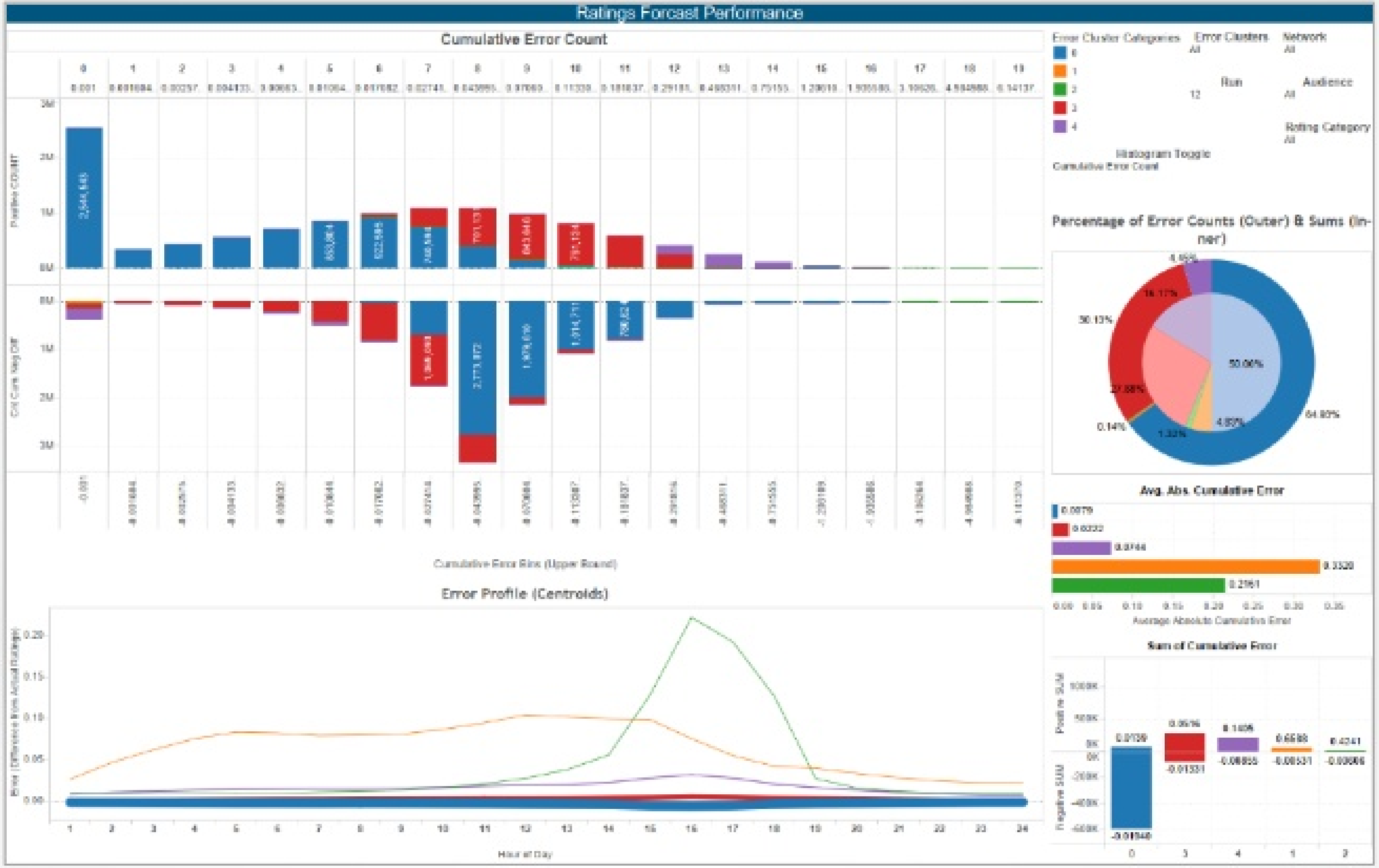
Verification: The evaluation of whether or not a product, service, or system complies with a regulation, requirement, specification, or imposed condition.
It is often an internal process.

How well did it do it?

System design, implementation and test (QA) are only part of the system acquisition process

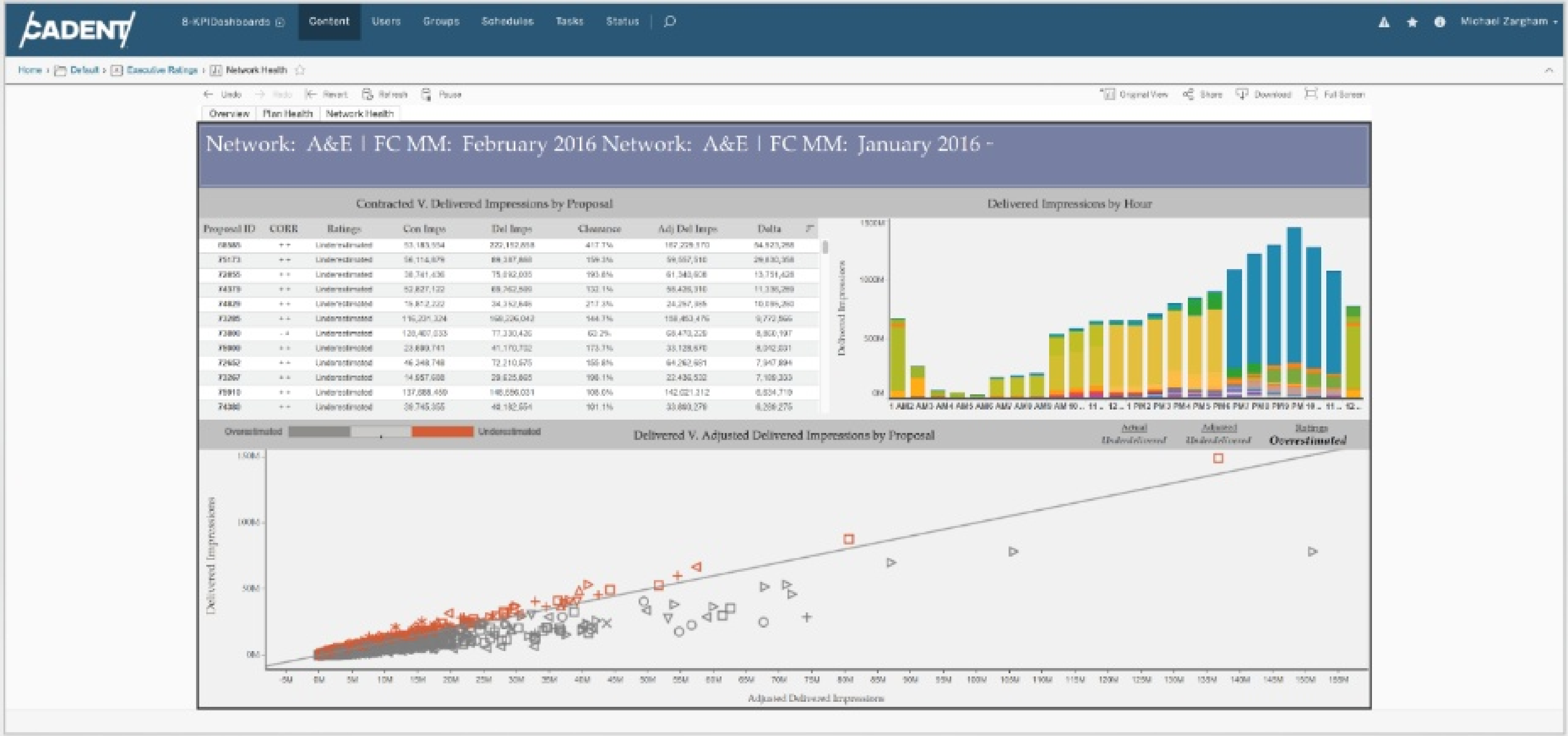
CADENT

Validation of Ratings via Performance Dashboard



CADENT

Verification of Ratings via Delivery KPI Dashboard



CADENT

WRAP UP

Other Projects

- Video on Demand Campaign Management
  - Forecasting Supply, Demand and Competition
  - Yield Management: Dynamic Inflight Optimization (feedback controller)
  - Pricing and Packaging of Inventory
- Extending Linear Advertising
  - Targeted Advertising Insertions in Linear Cable
  - Addressable Backfill for Linear Cable
  - Multicast Advertising on Broadcast Stations
- Unified Unicast/Multicast Advertising
  - Cross Platform Audience Based Planning
  - Flexible Hybrid-cloud Data Platform

CADENT

Thank You.

Michael Zargham mzargham@cadent.tv
Stefan Panayotov spanayotov@cadent.tv

SPARK SUMMIT 2017