CERN

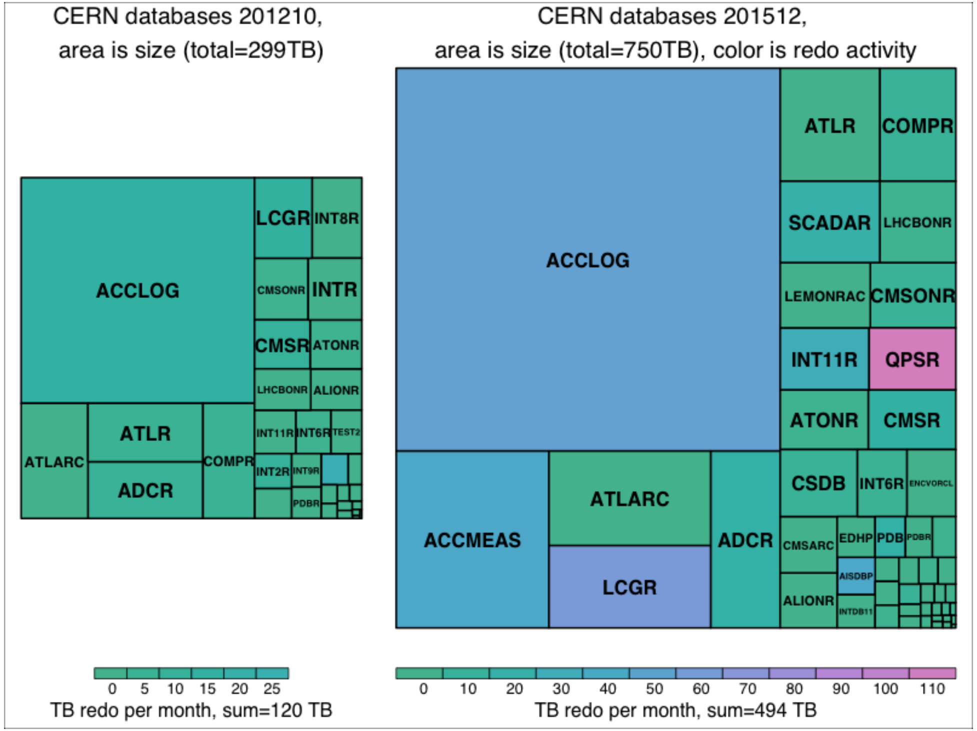
What's New for Machine Learning with Oracle Database and Hadoop

Eric Grancher
Manuel Martín Márquez

Outline

- CERN Data environment
- Introduction to Machine Learning
- Our Machine Learning path
  - Introduction to a real use case
  - R – prototyping and validating ideas
  - First Scalable attend – Oracle R Enterprise
  - Hadoop and the analytic transformation
  - Oracle Advance Analytics for Hadoop
  - Spark
  - TensorFlow + Spark

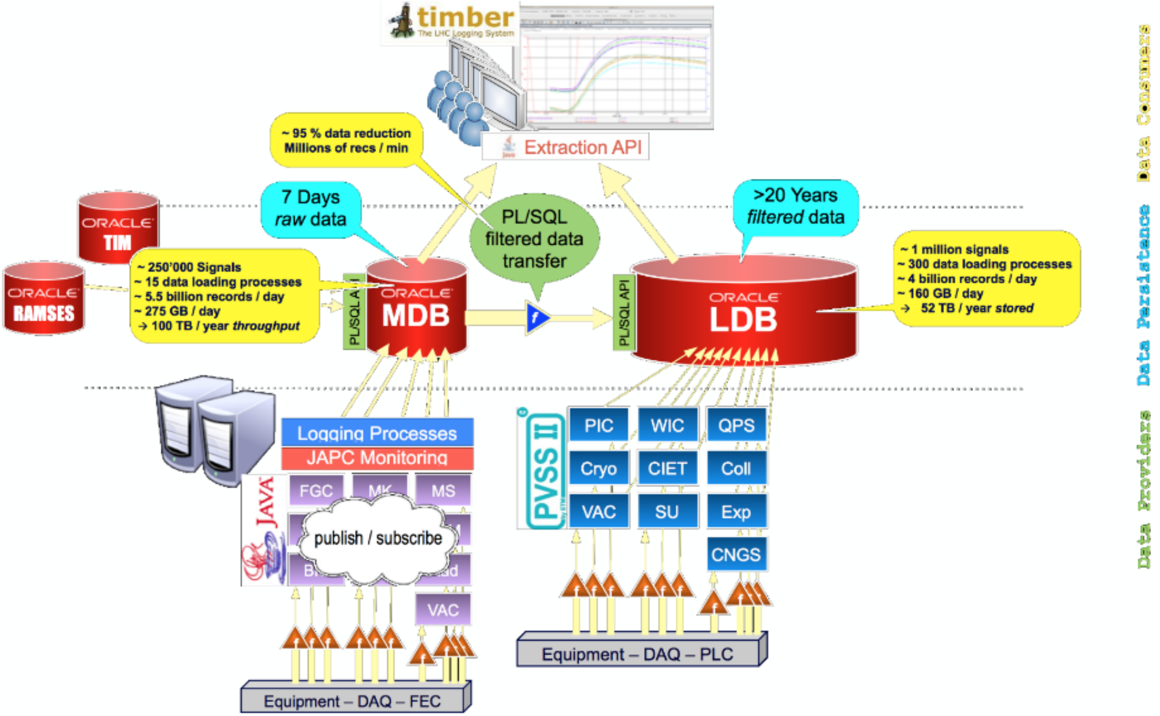
CERN Database Environment



	October 2012	December 2015
Max size	ACCLOG 136TB	ACCLOG 352TB
Max redo	ACCMEAS 27TB / month	QPSR 115TB / month

CERN Control Systems

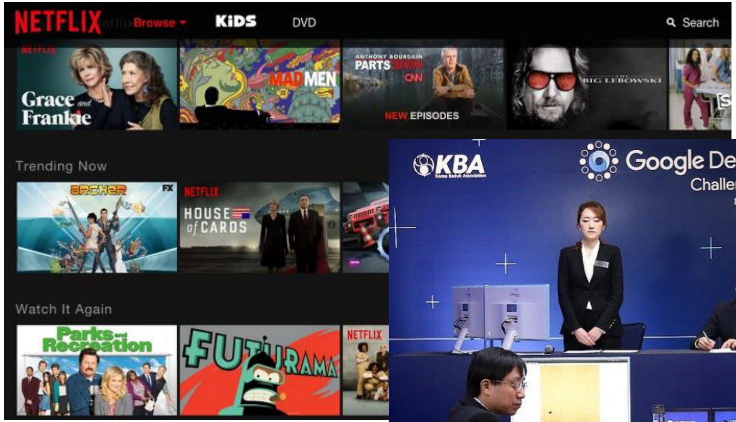
- IoT and Control Syste
  - **Cryogenics**
  - Vacuum
  - Machine Protection
  - Power Converters
  - QPS
- Accelerator Logging Service
  - ~ 275 GB/day
  - Storing more than 50 TB / year
  - Data acquisition
    - CERN accelerator complex
    - Related subsystems
    - Experiments
  - Around 1 million signals

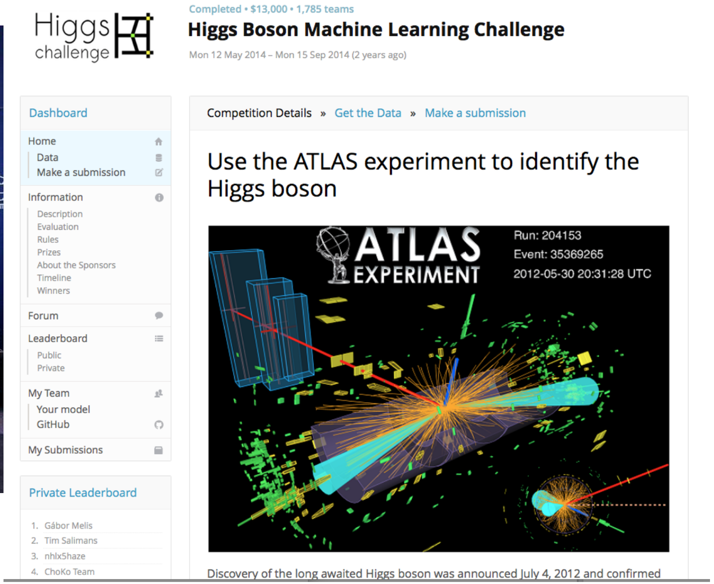


Machine Learning (ML)

- ML is a branch of artificial intelligence:
  - Uses computing based systems to make sense out of data
    - Extracting patterns, fitting data to functions, classifying data, etc
  - ML systems can learn and improve
    - With historical data, time and experience
  - Bridges theoretical computer science and real noise data.

ML in real-life

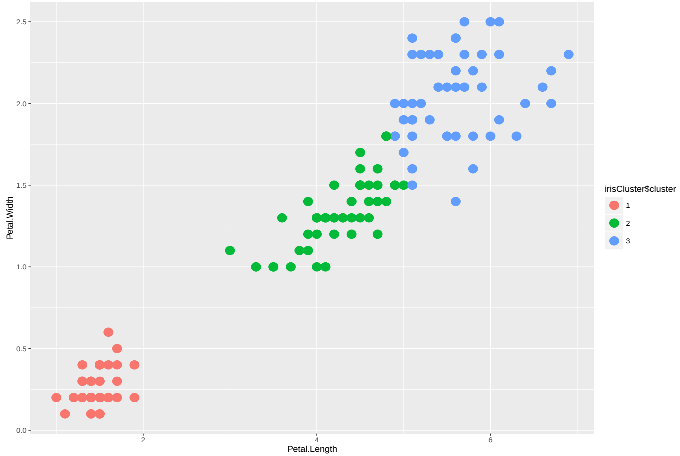




Supervised and Unsupervised Learning

- Unsupervised Learning
  - There are not predefined and known set of outcomes
  - Look for hidden patterns and relations in the data
  - A typical example: Clustering

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5.0	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1



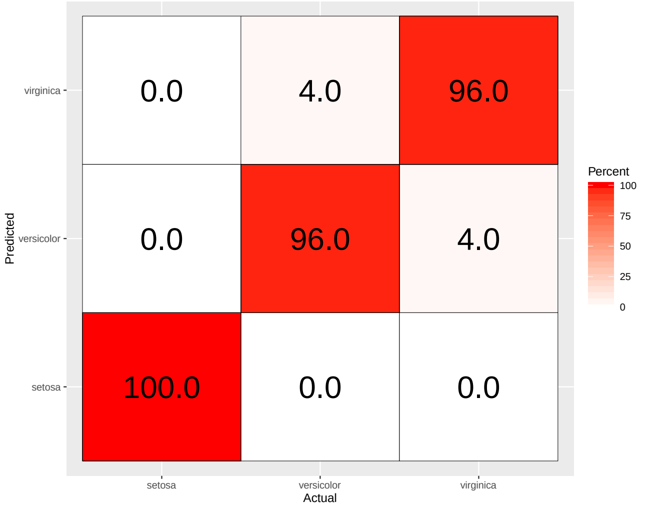
Supervised and Unsupervised Learning

- Supervised Learning
  - For every example in the data there is always a predefined outcome
  - Models the relations between a set of descriptive features and a target (Fits data to a function)
  - 2 groups of problems:
    - Classification
    - Regression

Supervised Learning

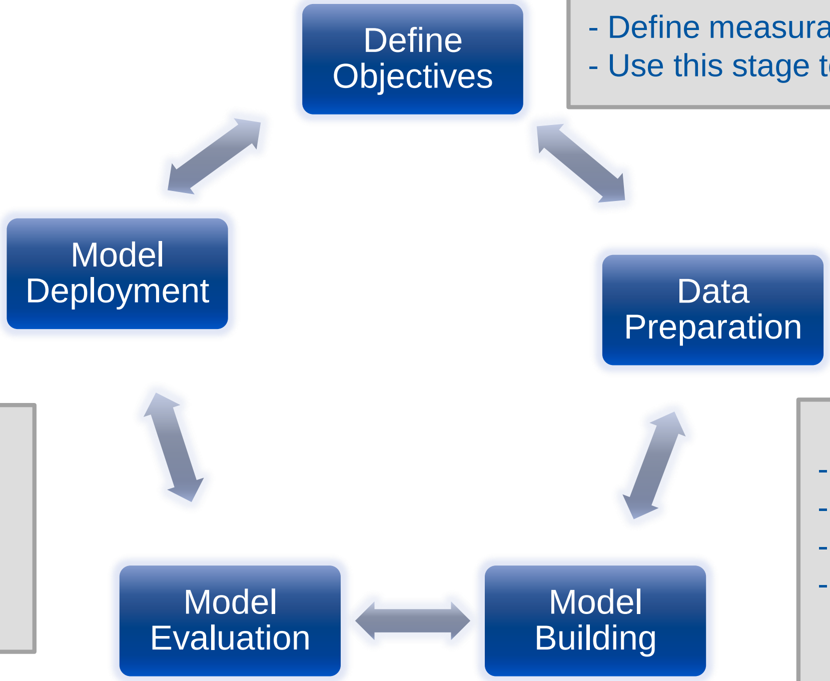
- Classification
  - Predicts which class a given sample of data (sample of descriptive features) is part of (**discrete value**).

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa



- Regression
  - Predicts continuous values.

Machine Learning as a Process

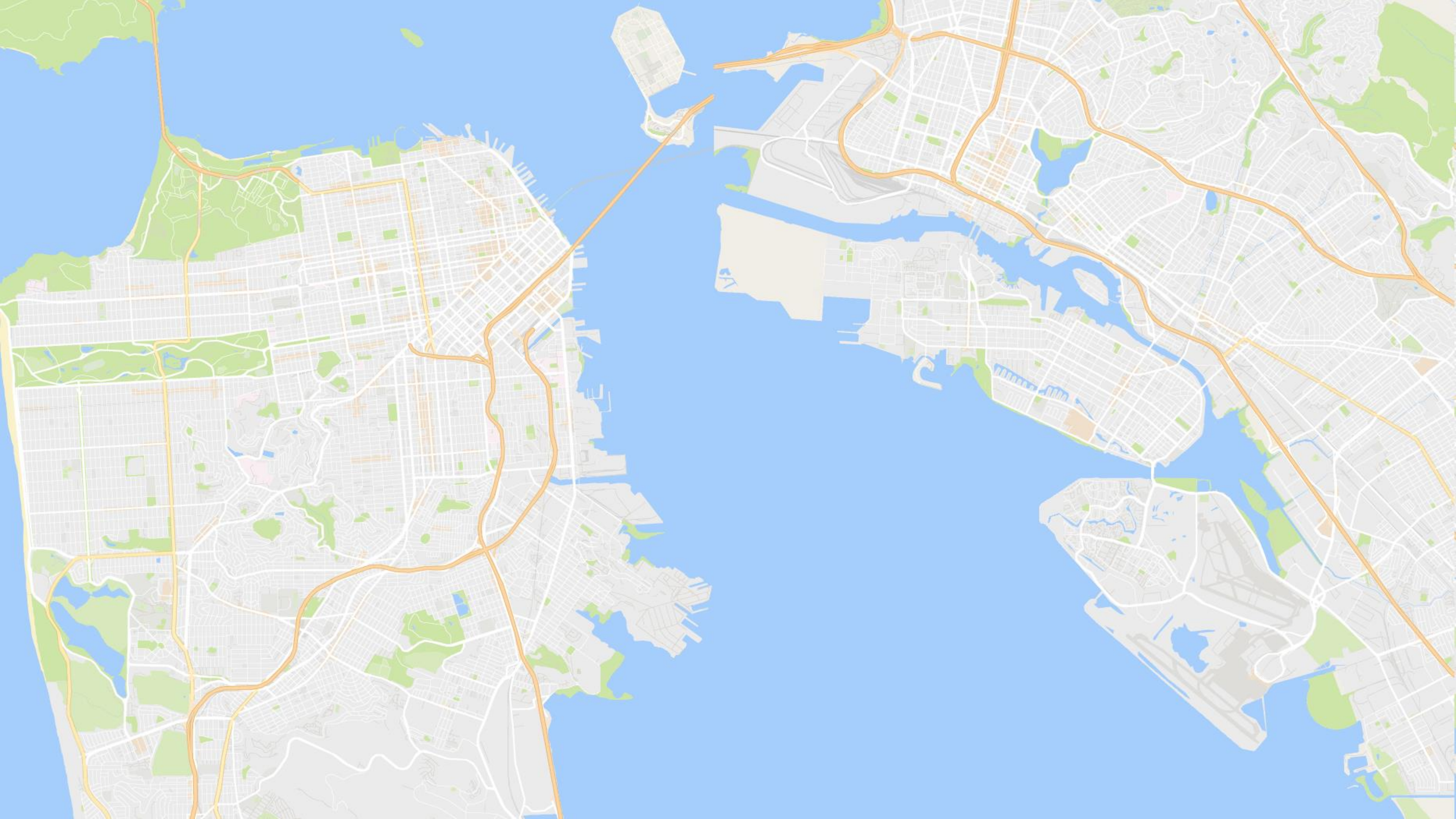


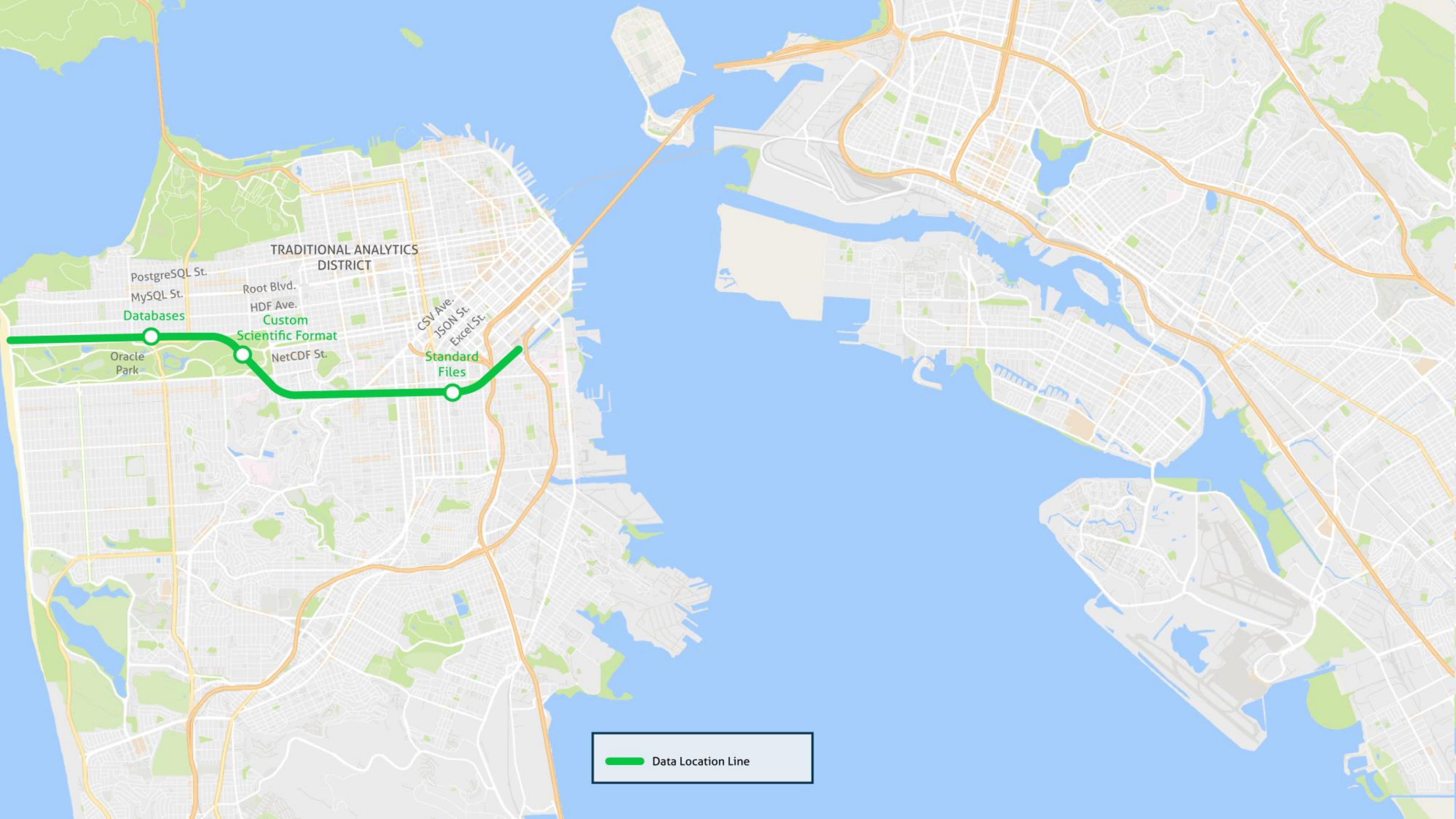
- Define measurable and quantifiable goals
- Use this stage to learn about the problem

- Normalization
- Transformation
- Missing Values
- Outliers

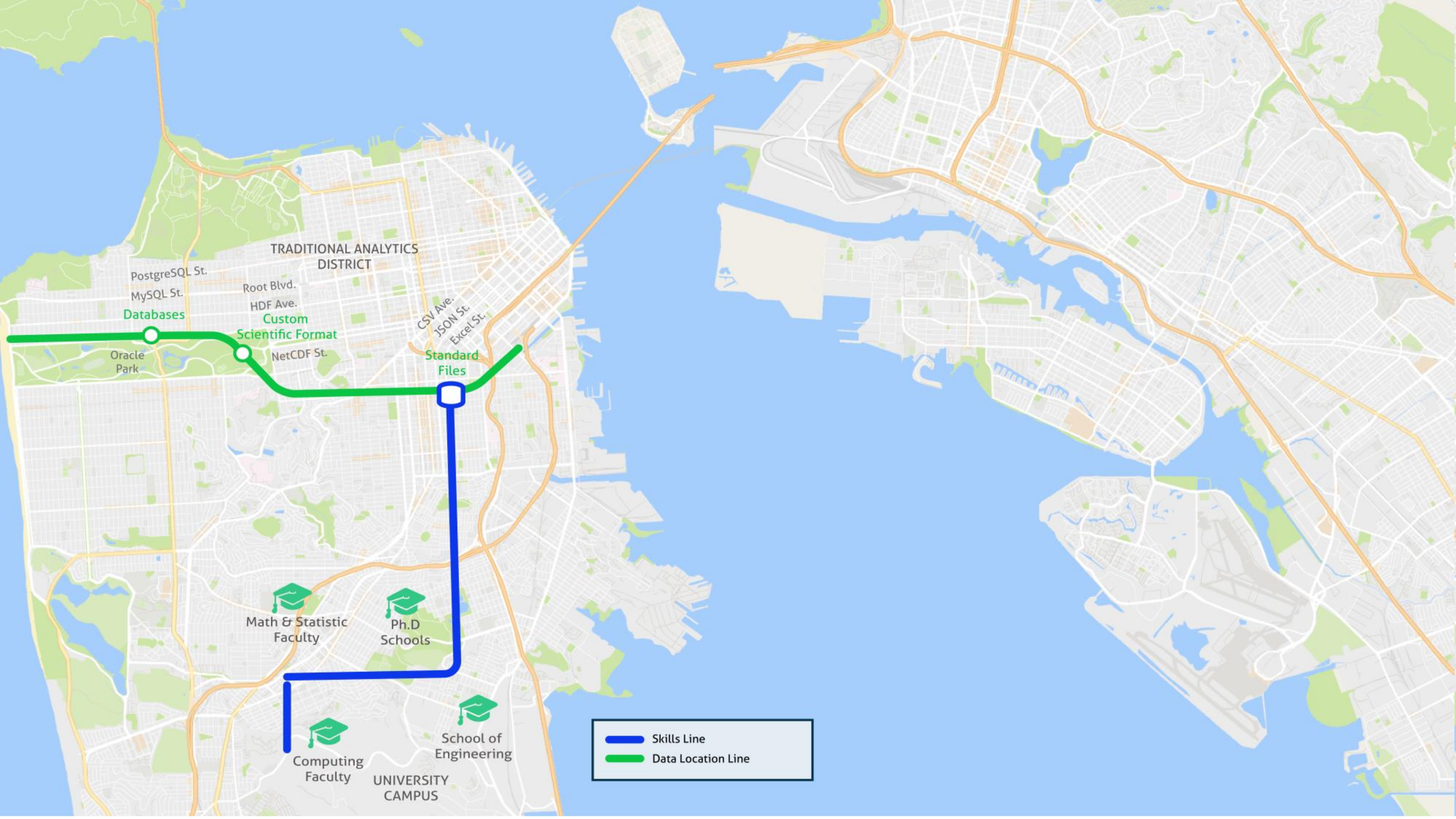
- Data Splitting
- Features Engineering
- Estimating Performance
- Evaluation and Model Selection

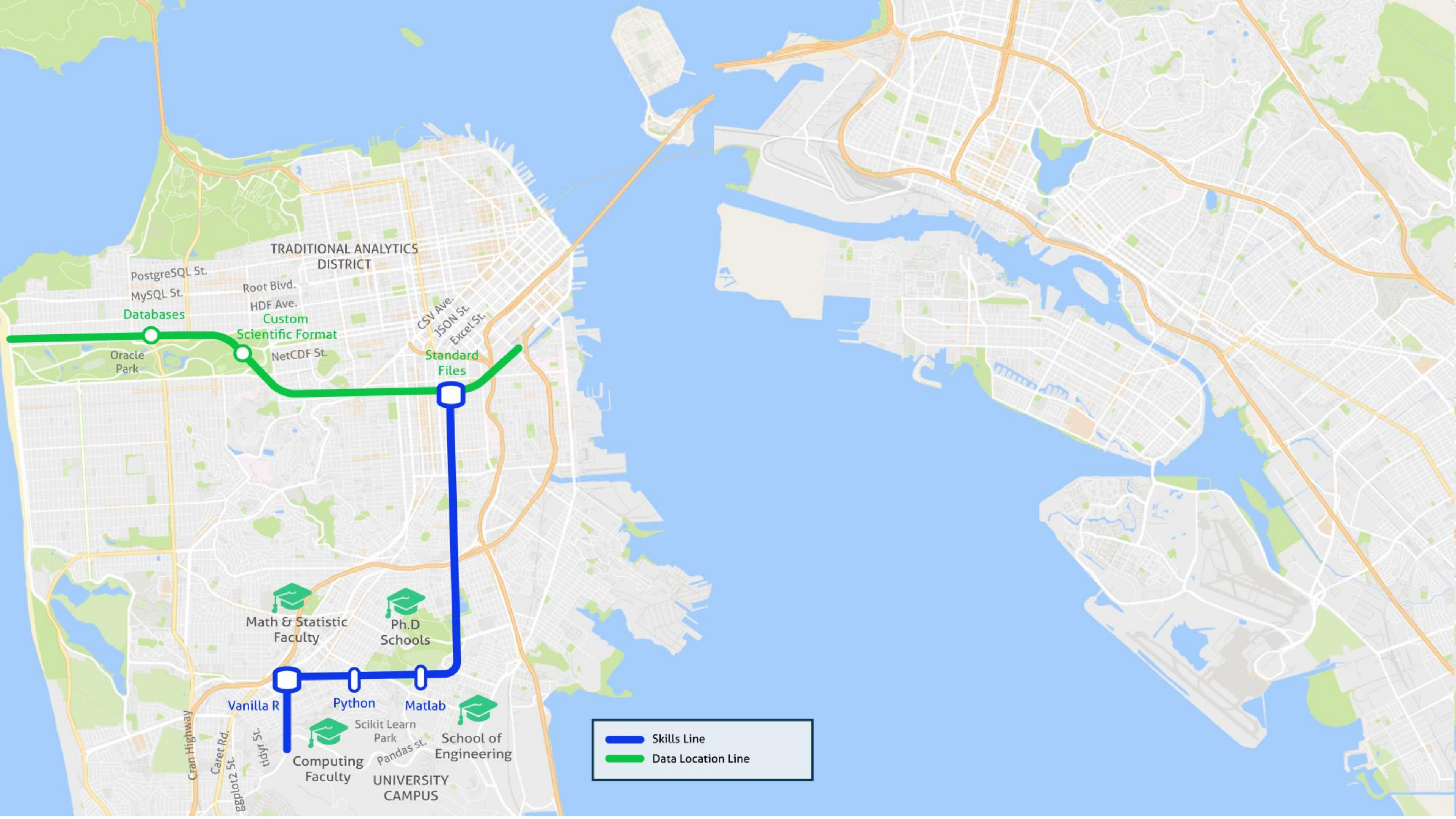
- Study models accuracy
  - Work better than the naïve approach or previous system
- Do the results make sense in the context of the problem

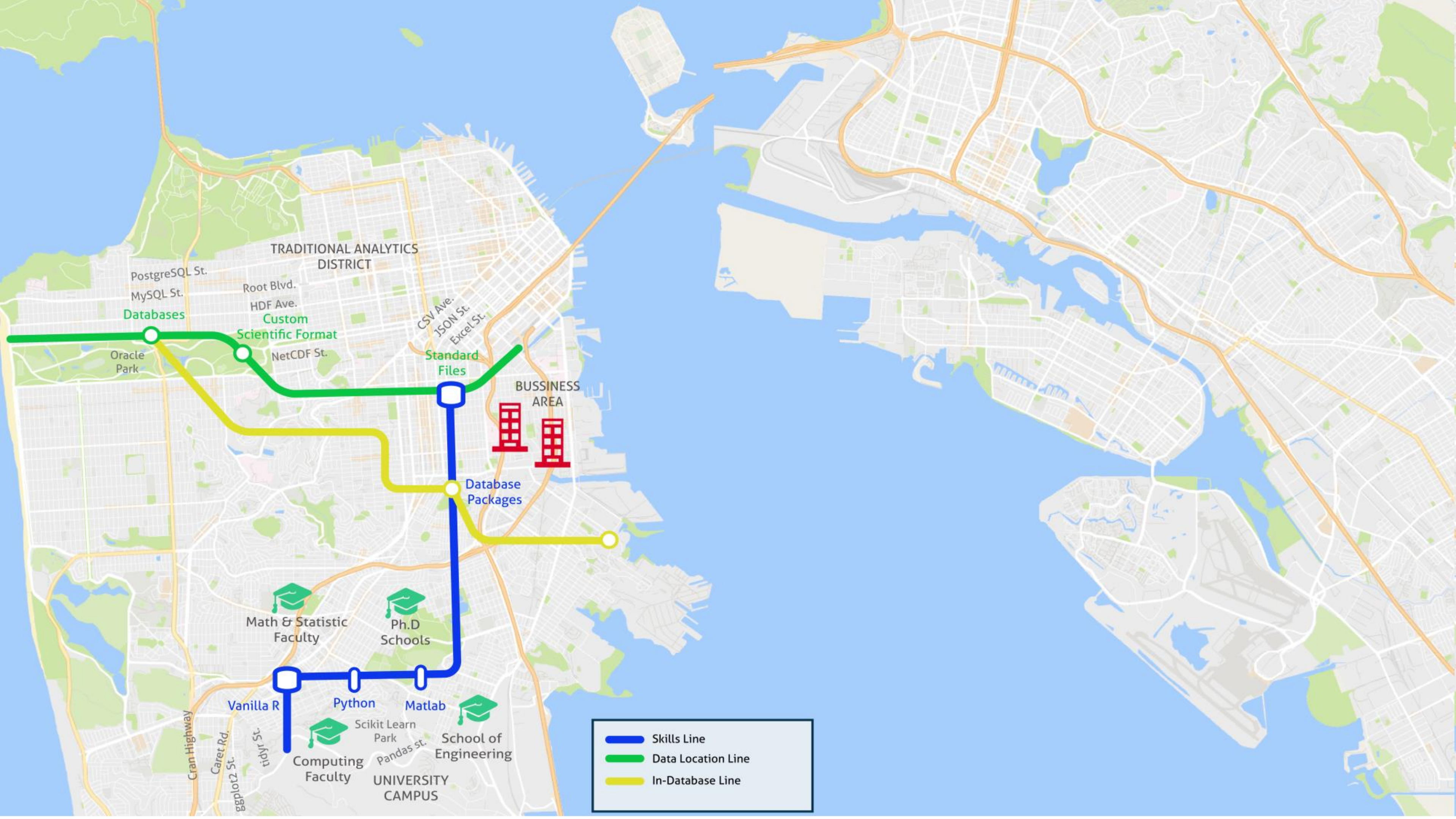




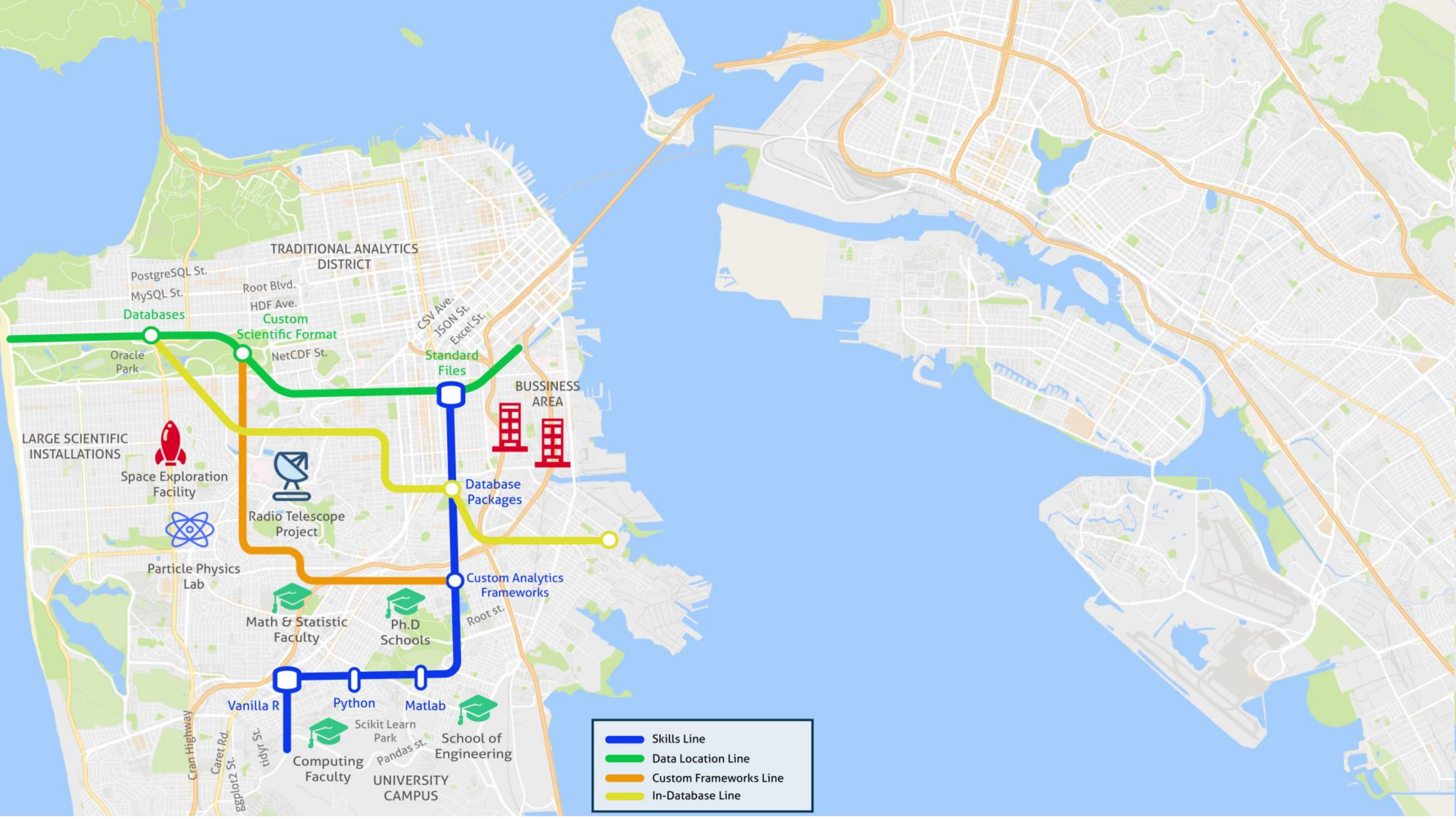
TRADITIONAL ANALYTICS DISTRICT
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
Databases
Custom Scientific Format
CSV Ave.
JSON St.
Excel St.
Oracle Park
NetCDF St.
Standard Files
Data Location Line

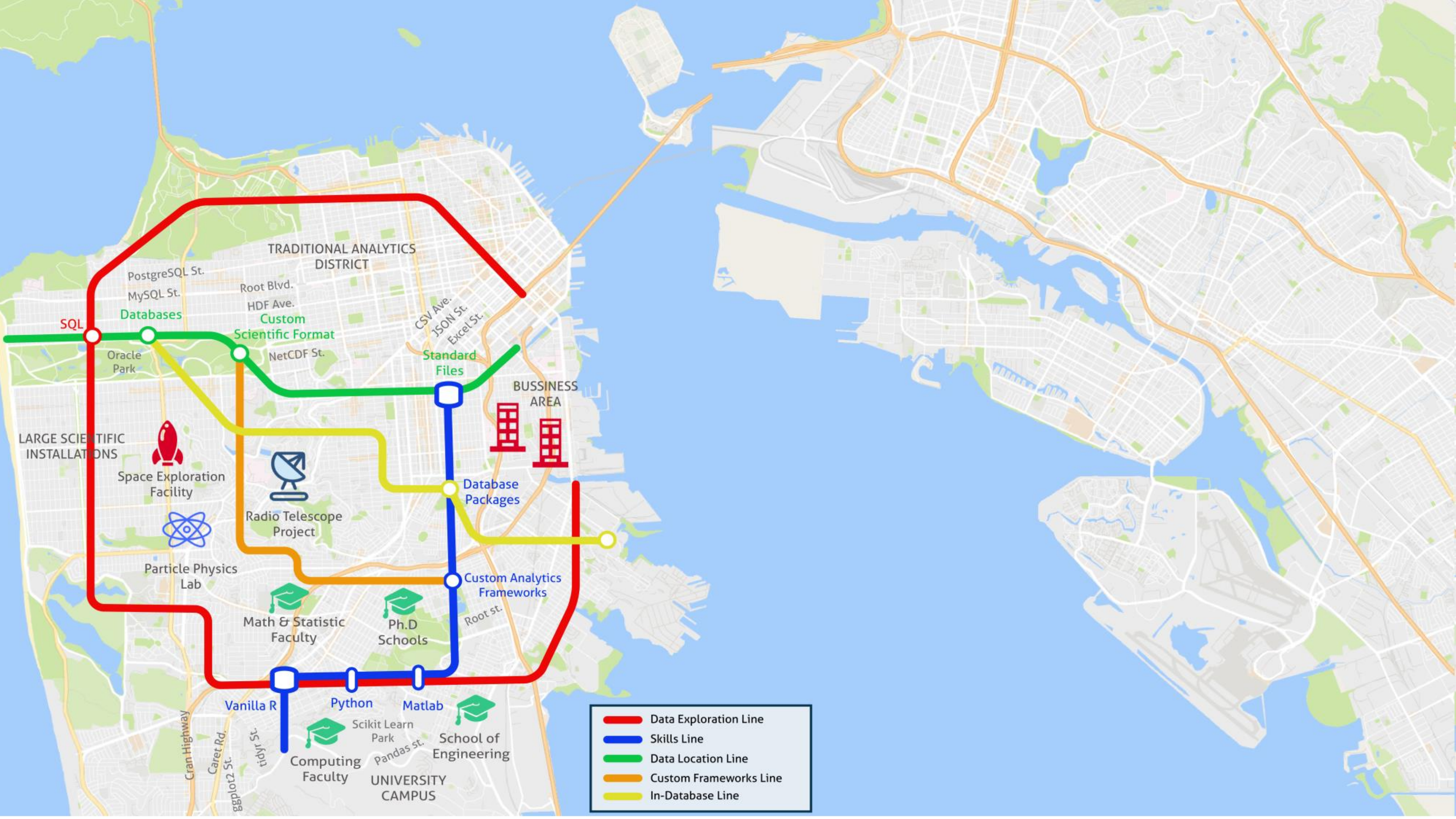

TRADITIONAL ANALYTICS DISTRICT
PostgreSQL St.
MySQL St.
Databases
Oracle Park
Root Blvd.
HDF Ave.
Custom Scientific Format
NetCDF St.
CSV Ave.
JSON St.
Excel St.
Standard Files
Math & Statistic Faculty
Ph.D Schools
School of Engineering
Computing Faculty
UNIVERSITY CAMPUS
Skills Line
Data Location Line


TRADITIONAL ANALYTICS DISTRICT
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
Databases
Custom Scientific Format
Oracle Park
NetCDF St.
CSV Ave.
JSON St.
Excel St.
Standard Files
Math & Statistic Faculty
Ph.D Schools
Vanilla R
Python
Matlab
Scikit Learn Park
Computing Faculty
Pandas St.
School of Engineering
UNIVERSITY CAMPUS
Cran Highway
Caret Rd.
ggplot2 St.
tidyr St.
Skills Line
Data Location Line



TRADITIONAL ANALYTICS DISTRICT
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
Databases
Custom Scientific Format
Oracle Park
NetCDF St.
CSV Ave.
JSON St.
Excel St.
Standard Files
BUSSINESS AREA
Database Packages
Math & Statistic Faculty
Ph.D Schools
Vanilla R
Python
Matlab
Scikit Learn Park
School of Engineering
Computing Faculty
Pandas st.
UNIVERSITY CAMPUS
Cran Highway
Caret Rd.
ggplot2 St.
tidyr St.
Skills Line
Data Location Line
In-Database Line


TRADITIONAL ANALYTICS DISTRICT
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
Databases
Custom Scientific Format
NetCDF St.
Oracle Park
CSV Ave.
JSON St.
Excel St.
Standard Files
BUSSINESS AREA
LARGE SCIENTIFIC INSTALLATIONS
Space Exploration Facility
Radio Telescope Project
Database Packages
Particle Physics Lab
Custom Analytics Frameworks
Math & Statistic Faculty
Ph.D Schools
Root st.
Vanilla R
Python
Matlab
Scikit Learn Park
School of Engineering
Pandas st.
Computing Faculty
UNIVERSITY CAMPUS
Cran Highway
Caret Rd.
ggplot2 St.
tidyr St.
Skills Line
Data Location Line
Custom Frameworks Line
In-Database Line

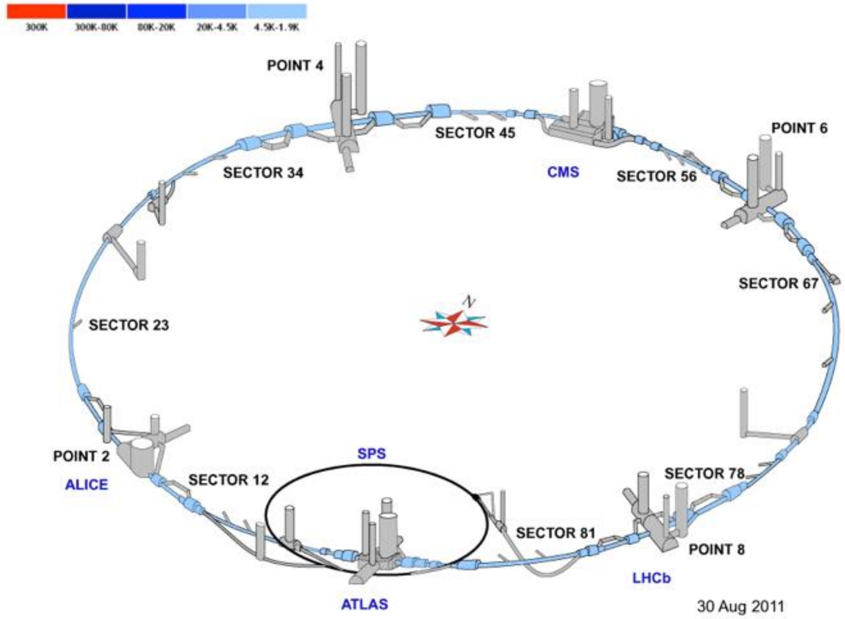


TRADITIONAL ANALYTICS DISTRICT
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
Databases
Custom Scientific Format
SQL
CSV Ave.
JSON St.
Excel St.
Oracle Park
NetCDF St.
Standard Files
BUSSINESS AREA
LARGE SCIENTIFIC INSTALLATIONS
Space Exploration Facility
Radio Telescope Project
Database Packages
Particle Physics Lab
Custom Analytics Frameworks
Math & Statistic Faculty
Ph.D Schools
Root st.
Vanilla R
Python
Matlab
Scikit Learn Park
Pandas st.
School of Engineering
Computing Faculty
UNIVERSITY CAMPUS
Cran Highway
Caret Rd.
ggplot2 St.
tidyr St.
Data Exploration Line
Skills Line
Data Location Line
Custom Frameworks Line
In-Database Line

Largest Cryogenics Installation

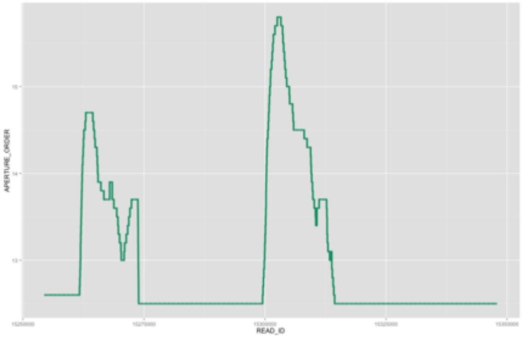
- 50k I/O, 11k actuators, ~5k control loops
- Control:
  - ~100 PLCs (Siemens, Schneider)
  - ~40 FECs (industrial PCs)
- Supervision: 26 SCADA servers

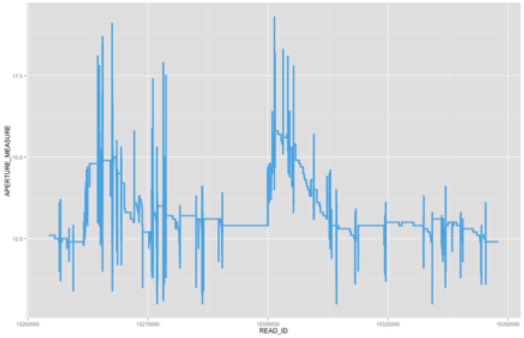
Instrument/Actuators	Total
Temperature [1.6 – 300 K]	10361
Pressure [0 – 20 bar]	2300
Level	923
Flow	2633
Control valves	3692
On/Off valves	1835
Manual valves	1916
Virtual flow meters	325
Controllers (PID)	4833

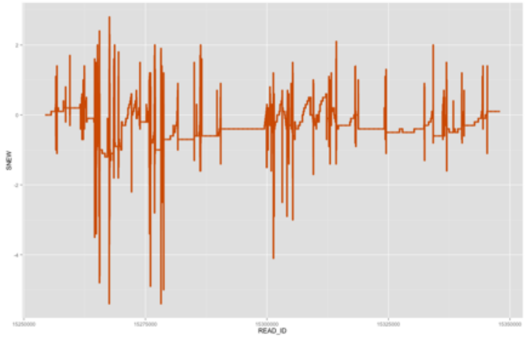


Use Case: Faulty Cryogenics Valve Detection

- What is the objective?
  - Predict faulty valves before they actually fail
- How?
  - Valve receive an aperture order value (**aperture order**)
  - Effective aperture realized by the valve (**aperture measured**)
  - Analyzing the difference between both (**S** = **aperture order** - **aperture measured**)







Faulty Cryogenics Valve Detection with R

- Signals used:
  - S = aperture order - aperture measured
- Features extractions based on S
  - Variance
  - Percentile 99.9
  - Rope distance – R(S)
  - Noise Band – B(S) (Pxx be the power spectrum of the signal S, from 0 to 0.5Hz, where S has been previously mean-centred).
- Automatic Faulty Valves Detection System
  - SVM - Support Vector Machine

$$R(S) = \frac{1}{N}\sum_{i=2}^{N}\left|S(i) - S(i-1)\right|$$

$$B(S) = \frac{\left|\sum_{k=1}^{N_{fft}/2} Pxx(k)\right|^2}{\sum_{k=1}^{N_{fft}/2} Pxx(k)^2}$$

Faulty Cryogenics Valve Detection with R

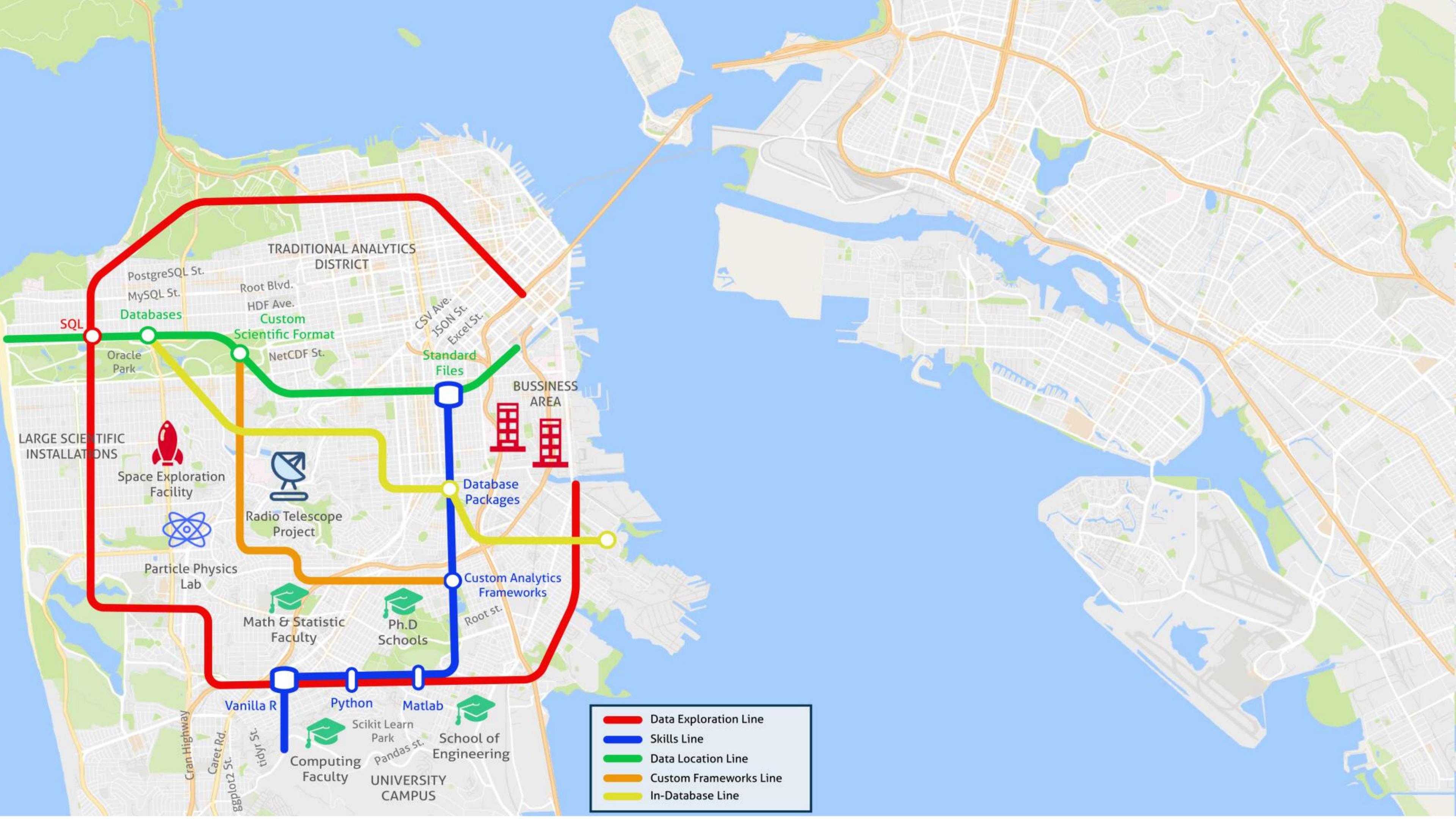
```r
#Libraries to be used
library(e1071)
library(ggplot2)
library(plyr)
library(bspec)
library(randomForest)

#Reading the dataset from a CVS file (about 200Mb - reduced version)
CV910_dataset <- read.csv("~/Documents/IT-DB-DBB/Projects/EN Department/POST MORTEM/Data/CV910_dataset.csv")

#Calculating the diference signal
CV910_dataset$s<-CV910_dataset$order-CV910_dataset$measure


#Calculating the features (agreedated values)
#Secuential Calculation
valve.features <- ddply(CV910_dataset, c("cycle_id"), summarize,
                          var = var(s),
                          max = max(s),
                          min = min (s),
                          rope_dist = sum(abs(diff(s)))/length(s),
                          bs=(sum(welchPSD(s-mean(s),seglength=512,two.sided = TRUE)$power)^2)/(2*512*sum(welchPSD(s-mean(s),seglength=512,two.sided = TRUE)$power^2)),
                          status = unique(status)
                        )

#Training the model - pretty fast training
svm_mod <- svm(status ~ var + max + min + rope_dist + bs, valve.features,na.action=na.omit)
rf_mod  <- randomForest(status ~ var + max + min + rope_dist + bs, valve.features, na.action=na.omit, ntree=50, norm.votes=FALSE)
```

Global Environment

Data	
CV910_dataset	5053287 obs. of 7 variables
valve.features	54 obs. of 7 variables

Our experience

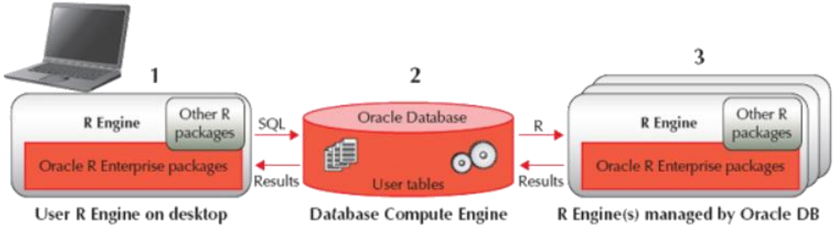
- Excellent for prototype potential solution or validate idea
  - Fast development using standard CRAN packages such as CARET etc.
  - Large number of models and statistic functions (+7500 packages) covering a wide range of fields
  - Data Exploration
- Use the existing skills
  - R is widely use in the domain
- Move the data is very expensive
  - The data need to be extracted from DB and generate files CSV
    - SQL, Java API, Custom Extraction Applications (Timber)
- Hard to deploy models in production and scale the solutions as the data grown
  - Data limited by memory size.
  - Few packages but very limited scalability - the models themselves do not scale
    - Foreach, Snow, Rmpi, BatchExperiments package (BatchJobs)


TRADITIONAL ANALYTICS DISTRICT
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
SQL
Databases
Custom Scientific Format
CSV Ave.
JSON St.
Excel St.
Oracle Park
NetCDF St.
Standard Files
BUSSINESS AREA
LARGE SCIENTIFIC INSTALLATIONS
Space Exploration Facility
Radio Telescope Project
Database Packages
Particle Physics Lab
Custom Analytics Frameworks
Root st.
Math & Statistic Faculty
Ph.D Schools
Vanilla R
Python
Matlab
Scikit Learn Park
School of Engineering
Cran Highway
Caret Rd.
ggplot2 St.
tidyr St.
Computing Faculty
Pandas st.
UNIVERSITY CAMPUS
Data Exploration Line
Skills Line
Data Location Line
Custom Frameworks Line
In-Database Line

TRADITIONAL ANALYTICS DISTRICT
SCALABLE ANALYTICS DISTRICT
BUSSINESS AREA
LARGE SCIENTIFIC INSTALLATIONS
UNIVERSITY CAMPUS
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
CSV Ave.
JSON St.
Excel St.
NetCDF St.
Oracle Park
SQL
Databases
Custom Scientific Format
Standard Files
Database Packages
Custom Analytics Frameworks
Vanilla R
Python
Matlab
Oracle Rac Ave.
Oracle R Entrerprise
Oracle Rac Ave.
ORE Rd.
Space Exploration Facility
Radio Telescope Project
Particle Physics Lab
Math & Statistic Faculty
Ph.D Schools
Root st.
Computing Faculty
Scikit Learn Park
Pandas st.
School of Engineering
Cran Highway
Caret Rd.
ggplot2 St.
tidyr St.
Data Exploration Line
Skills Line
Data Location Line
Custom Frameworks Line
In-Database Line
TRANSFER LINES
Super-Reduced Extra Cost

Why ORE? - ORE benefits

- A database-centric environment for analytical processes in R
  - Allows to use the database server to run R scripts (scalability & performance)
  - Eliminate memory constraint of client R engine



- Transparency Layer
  - Transparently analyze and use data in Oracle Database through R
  - Tables as R native data frames

- Enables users to take advantage of data-parallel and task-parallel execution through Oracle Database

Cryo Valves – Parallel Features Extraction in ORE

```r
#Funtion to calculate and extract the features
features <- function(dat) {
  #Load R Libraries
  library(bspec)

  #Calculate the signal to work with
  s<-dat$APERTURE_ORDER-dat$APERTURE_MEASURE

  #Features Calculations
  valve = unique(dat$VALVE)
  cycle = unique(dat$CYCLE_NUMBER)
  status = unique(dat$STATUS)

  var=var(s)
  max=max(s)
  min=min(s)

  rope_dist=sum(abs(diff(s)))/length(s)
  pxx<-welchPSD(s-mean(s),seglength=512,two.sided = TRUE)$power
  bs=(sum(pxx)^2)/(2*512*sum(pxx^2))

  #Return the features
  data.frame(Cycle = cycle,
             Valve = valve,
             Status = status,
             Var=var,
             Max=max,
             Min=min,
             Rope_dist = rope_dist,
             Bs = bs)

}
```

Instrument/Actuators	Total
Temperature [1.6 – 300 K]	10361
Pressure [0 – 20 bar]	2300
Level	923
Flow	2633
Control valves	**3692**
On/Off valves	1835
Manual valves	1916
Virtual flow meters	325
Controllers (PID)	4833

93600 points per cycle (about 24 hours)

CERN

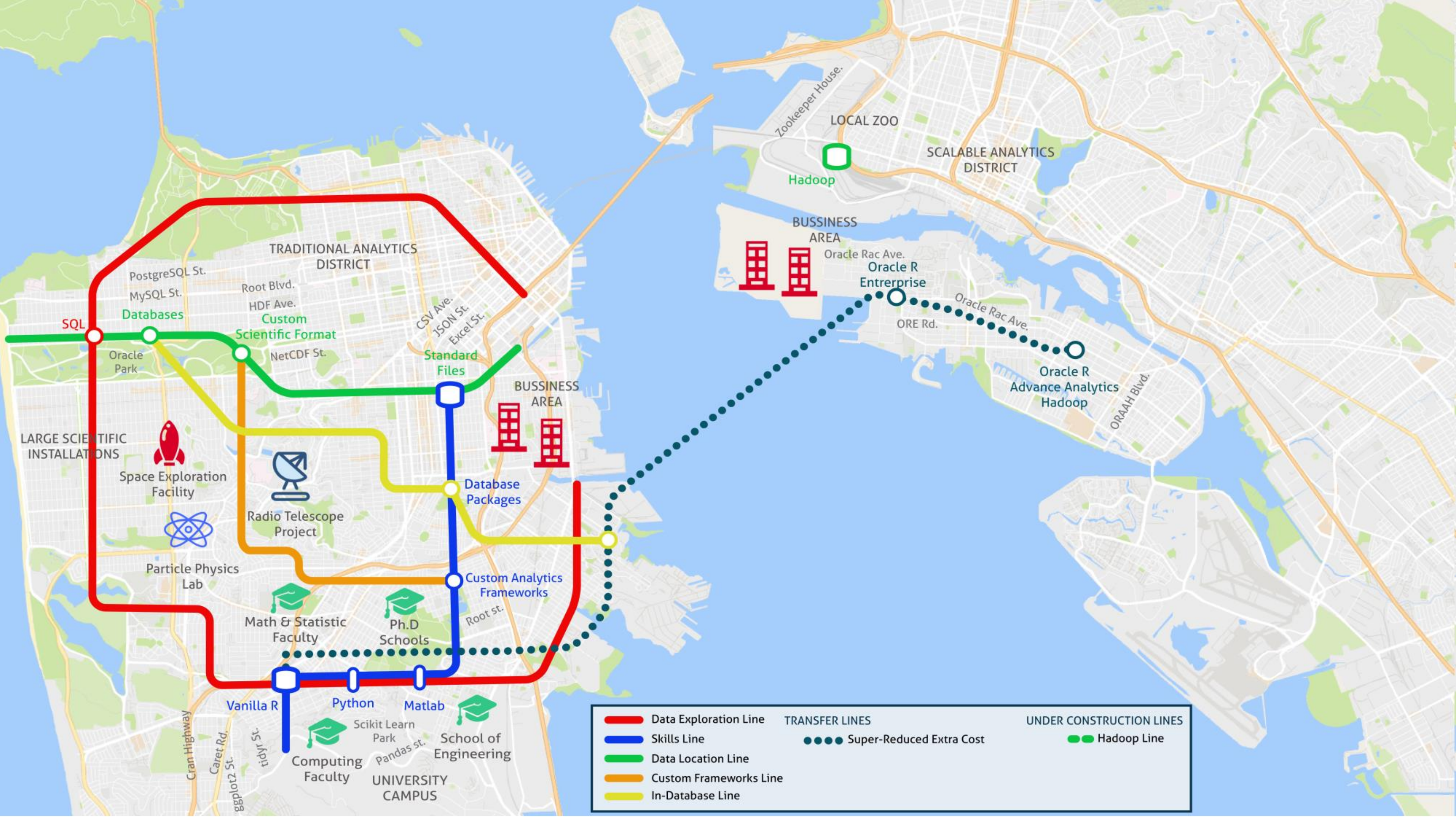
Cryo Valves – Parallel Features Extraction in ORE

```r
#Create in-db partitions and proccess them in parallel (VALVE_READS_CL_TRA is the name of the table in our db)
#Lazy evaluation!!!
ore.valve.features <- ore.groupApply(VALVE_READS_CL_TRA [,c('READ_ID','CYCLE_NUMBER','VALVE','APERTURE_ORDER','APERTURE_MEASURE','STATUS')],
                                     INDEX=VALVE_READS_CL_TRA$CYCLE_NUMBER,
                                     features,
                                     #Return value signature
                                     FUN.VALUE =
                                       data.frame(Cycle = numeric(),
                                                  Valve = character(),
                                                  Status = character(),
                                                  Var = numeric(),
                                                  Max = numeric(),
                                                  Min = numeric(),
                                                  Rope_dist = numeric(),
                                                  Bs = numeric())
                                     ,parallel=TRUE)

#Order the output
row.names(ore.valve.features)=ore.valve.features$Cycle

#Pulling the fatures from DB for local processing
valve.features<-ore.pull(ore.valve.features)

#Training the models locally
rf_mod <- svm(Status ~ Var + Max + Min + Rope_dist + Bs, valve.features,na.action=na.omit)
```

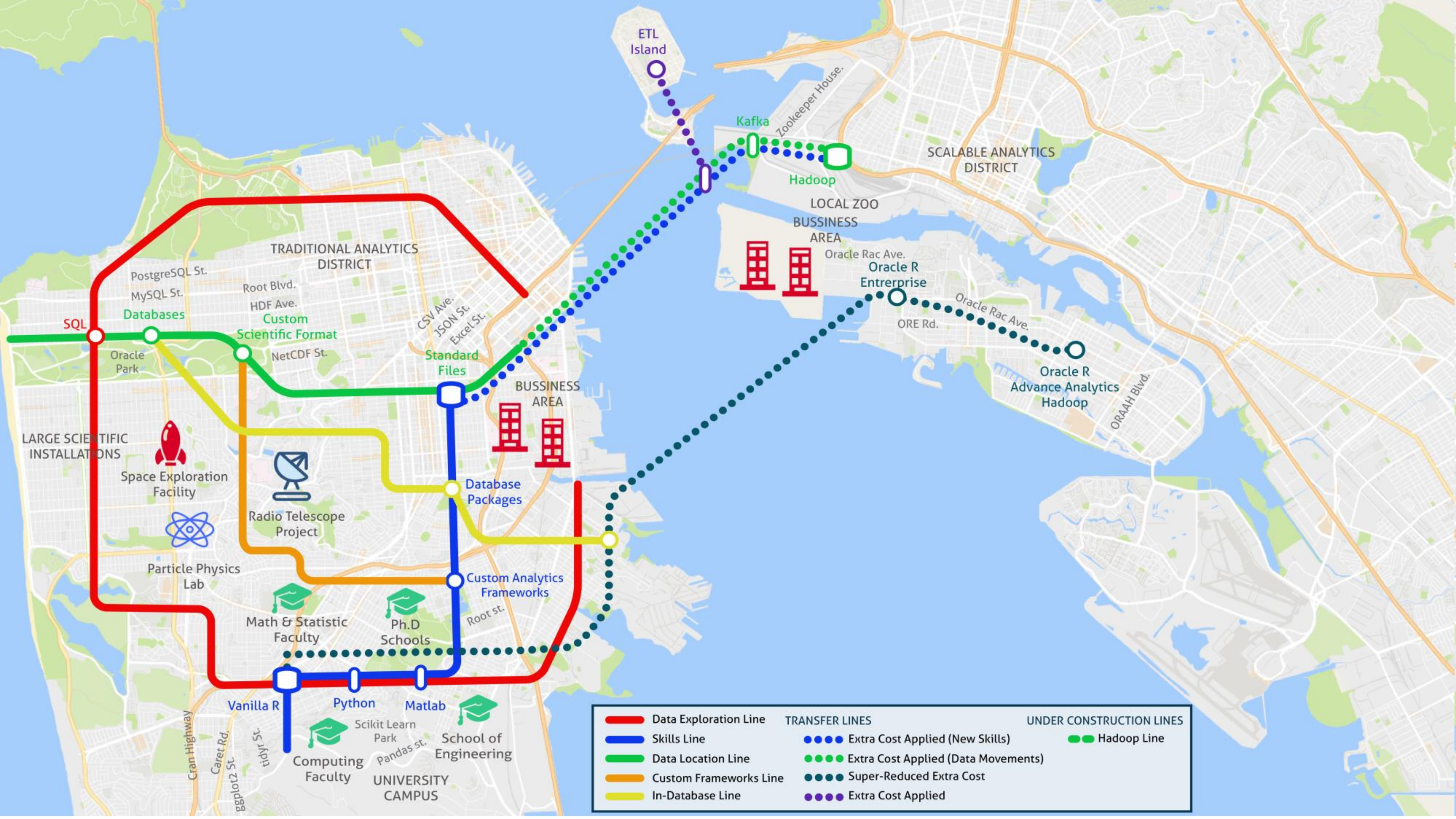
Our experience

- No need to move data
  - It is faster to process it in-DB with ORE using the appropriate degree of parallelism
- DB nodes already prepared for the workload
  - Simplifies the infrastructure
- Write/adapt R code is straight forward
  - Thanks to transparency layer and embedded R execution
  - Tables and Views as R dataframes
- Still problems scaling
  - Scalability determined by RAC installations
- Need to differentiate between production and analytics environments
  - Risk on affecting the production environment performance by running in-database analytics
- Analytics developments in-database
  - Risk on data security and resources competition

TRADITIONAL ANALYTICS DISTRICT
SCALABLE ANALYTICS DISTRICT
LARGE SCIENTIFIC INSTALLATIONS
BUSSINESS AREA
UNIVERSITY CAMPUS
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
CSV Ave.
JSON St.
Excel St.
NetCDF St.
SQL
Databases
Custom Scientific Format
Standard Files
Oracle Park
Database Packages
Custom Analytics Frameworks
Root st.
Space Exploration Facility
Radio Telescope Project
Particle Physics Lab
Math & Statistic Faculty
Ph.D Schools
Vanilla R
Python
Matlab
Scikit Learn Park
Pandas st.
Computing Faculty
School of Engineering
Cran Highway
Caret Rd.
ggplot2 St.
tidyr St.
Oracle Rac Ave.
Oracle R Entrerprise
Oracle Rac Ave.
ORE Rd.
Data Exploration Line
Skills Line
Data Location Line
Custom Frameworks Line
In-Database Line
TRANSFER LINES
Super-Reduced Extra Cost



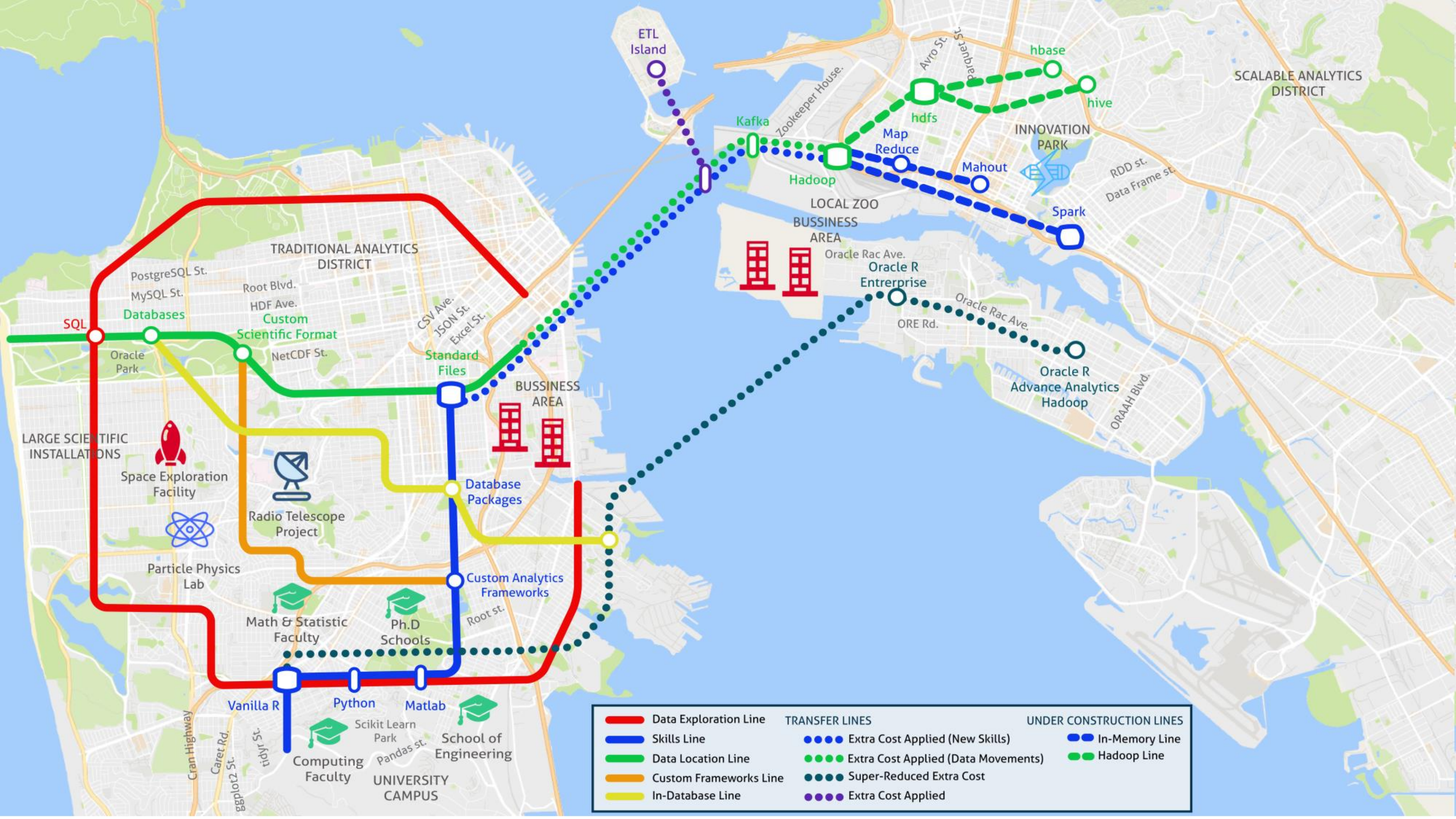
LOCAL ZOO
Zookeeper House.
Hadoop
SCALABLE ANALYTICS DISTRICT
BUSSINESS AREA
Oracle Rac Ave.
Oracle R Entrerprise
Oracle Rac Ave.
ORE Rd.
Oracle R Advance Analytics Hadoop
ORAAH Blvd.
TRADITIONAL ANALYTICS DISTRICT
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
Databases
Custom Scientific Format
SQL
Oracle Park
NetCDF St.
CSV Ave.
JSON St.
Excel St.
Standard Files
BUSSINESS AREA
LARGE SCIENTIFIC INSTALLATIONS
Space Exploration Facility
Radio Telescope Project
Database Packages
Particle Physics Lab
Custom Analytics Frameworks
Math & Statistic Faculty
Ph.D Schools
Root st.
Vanilla R
Python
Matlab
Scikit Learn Park
Pandas st.
School of Engineering
Computing Faculty
UNIVERSITY CAMPUS
Cran Highway
Caret Rd.
ggplot2 St.
tidyr St.
Data Exploration Line
Skills Line
Data Location Line
Custom Frameworks Line
In-Database Line
TRANSFER LINES
Super-Reduced Extra Cost
UNDER CONSTRUCTION LINES
Hadoop Line



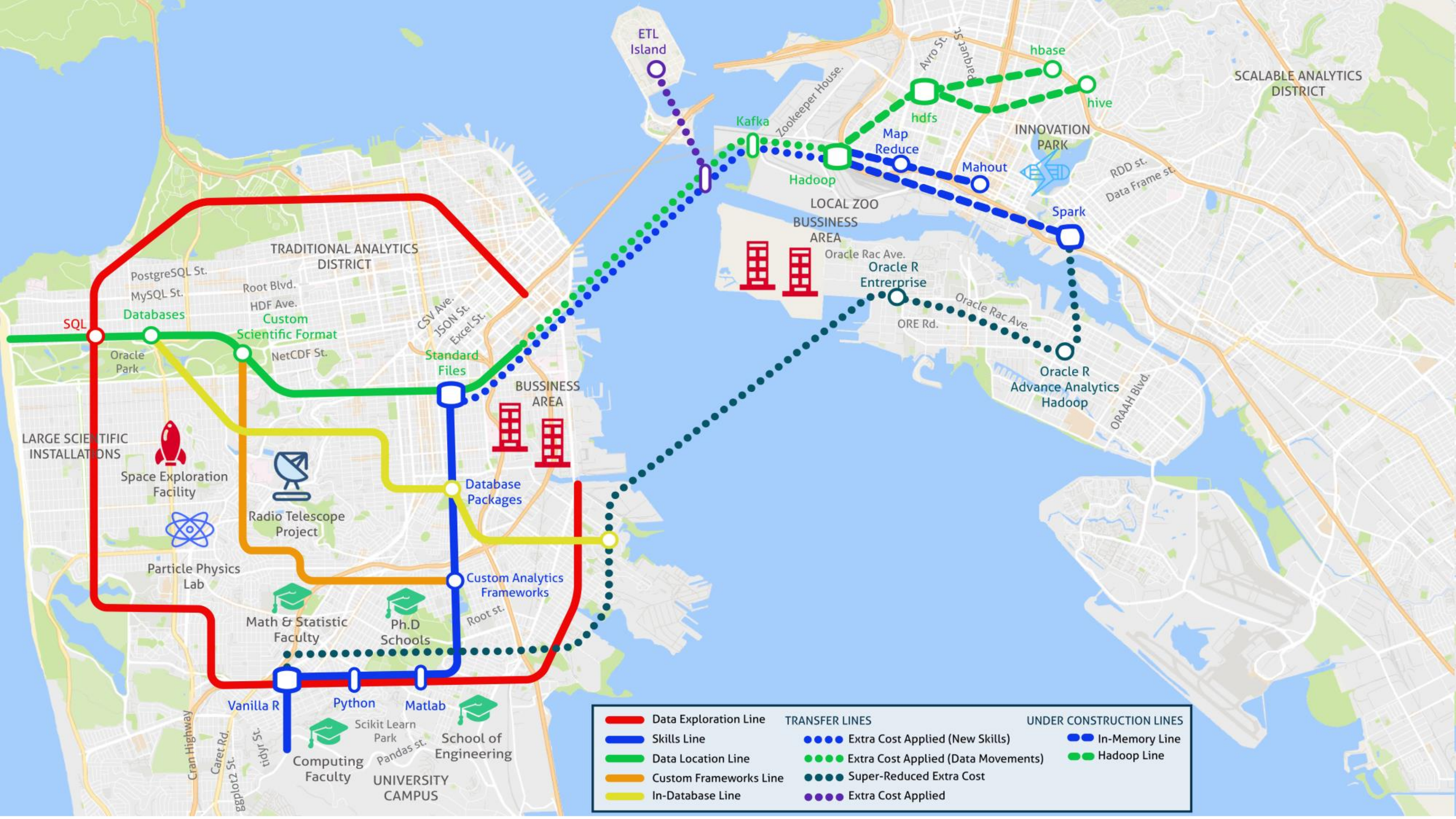
ETL Island
Kafka
Zookeeper House.
Hadoop
SCALABLE ANALYTICS DISTRICT
LOCAL ZOO
BUSSINESS AREA
Oracle Rac Ave.
Oracle R Entrerprise
Oracle Rac Ave.
ORE Rd.
Oracle R Advance Analytics Hadoop
ORAAH Blvd.
TRADITIONAL ANALYTICS DISTRICT
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
Databases
SQL
Custom Scientific Format
NetCDF St.
CSV Ave.
JSON St.
Excel St.
Standard Files
Oracle Park
BUSSINESS AREA
LARGE SCIENTIFIC INSTALLATIONS
Space Exploration Facility
Radio Telescope Project
Particle Physics Lab
Database Packages
Custom Analytics Frameworks
Math & Statistic Faculty
Ph.D Schools
Root st.
Vanilla R
Python
Matlab
Scikit Learn Park
Pandas st.
School of Engineering
Computing Faculty
UNIVERSITY CAMPUS
Cran Highway
Caret Rd.
ggplot2 St.
tidyr St.
Data Exploration Line
Skills Line
Data Location Line
Custom Frameworks Line
In-Database Line
TRANSFER LINES
Extra Cost Applied (New Skills)
Extra Cost Applied (Data Movements)
Super-Reduced Extra Cost
Extra Cost Applied
UNDER CONSTRUCTION LINES
Hadoop Line

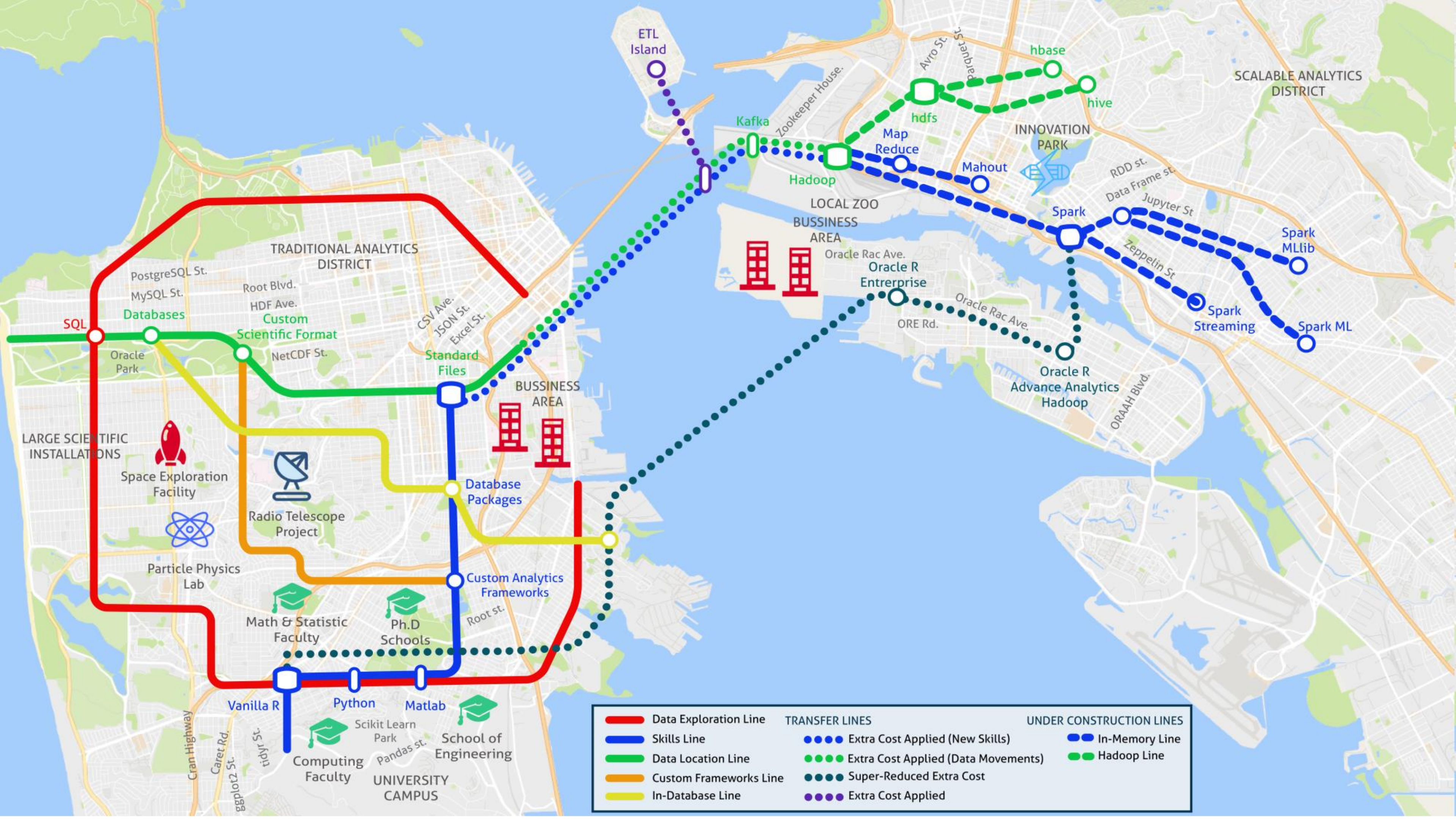
ETL Island
Kafka
Zookeeper House.
Hadoop
Avro St.
Parquet St.
hbase
hdfs
hive
SCALABLE ANALYTICS DISTRICT
INNOVATION PARK
LOCAL ZOO
BUSSINESS AREA
Oracle Rac Ave.
Oracle R Entrerprise
ORE Rd.
Oracle Rac Ave.
Oracle R Advance Analytics Hadoop
ORAAH Blvd.
TRADITIONAL ANALYTICS DISTRICT
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
SQL
Databases
Custom Scientific Format
NetCDF St.
CSV Ave.
JSON St.
Excel St.
Standard Files
Oracle Park
BUSSINESS AREA
LARGE SCIENTIFIC INSTALLATIONS
Space Exploration Facility
Radio Telescope Project
Database Packages
Particle Physics Lab
Custom Analytics Frameworks
Root st.
Math & Statistic Faculty
Ph.D Schools
Vanilla R
Python
Matlab
Scikit Learn Park
Pandas st.
School of Engineering
Computing Faculty
UNIVERSITY CAMPUS
Cran Highway
Caret Rd.
ggplot2 St.
tidyr St.
Data Exploration Line
Skills Line
Data Location Line
Custom Frameworks Line
In-Database Line
TRANSFER LINES
Extra Cost Applied (New Skills)
Extra Cost Applied (Data Movements)
Super-Reduced Extra Cost
Extra Cost Applied
UNDER CONSTRUCTION LINES
In-Memory Line
Hadoop Line

ETL Island
Kafka
Zookeeper House.
Avro St.
Parquet St.
hbase
hive
hdfs
SCALABLE ANALYTICS DISTRICT
INNOVATION PARK
Map Reduce
Mahout
Hadoop
LOCAL ZOO
BUSSINESS AREA
Oracle Rac Ave.
Oracle R Entrerprise
Oracle Rac Ave.
ORE Rd.
Oracle R Advance Analytics Hadoop
ORAAH Blvd.
TRADITIONAL ANALYTICS DISTRICT
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
SQL
Databases
Custom Scientific Format
CSV Ave.
JSON St.
Excel St.
NetCDF St.
Oracle Park
Standard Files
BUSSINESS AREA
LARGE SCIENTIFIC INSTALLATIONS
Space Exploration Facility
Radio Telescope Project
Database Packages
Particle Physics Lab
Custom Analytics Frameworks
Root st.
Math & Statistic Faculty
Ph.D Schools
Vanilla R
Python
Matlab
Scikit Learn Park
Pandas st.
School of Engineering
Computing Faculty
UNIVERSITY CAMPUS
Cran Highway
Caret Rd.
ggplot2 St.
tidyr St.
Data Exploration Line
Skills Line
Data Location Line
Custom Frameworks Line
In-Database Line
TRANSFER LINES
Extra Cost Applied (New Skills)
Extra Cost Applied (Data Movements)
Super-Reduced Extra Cost
Extra Cost Applied
UNDER CONSTRUCTION LINES
Hadoop Line



ETL Island
Kafka
Zookeeper House.
Avro St.
Parquet St.
hbase
hive
hdfs
SCALABLE ANALYTICS DISTRICT
INNOVATION PARK
Map Reduce
Mahout
Hadoop
LOCAL ZOO
RDD st.
Data Frame st.
Spark
BUSSINESS AREA
Oracle Rac Ave.
Oracle R Entrerprise
Oracle Rac Ave.
ORE Rd.
Oracle R Advance Analytics Hadoop
ORAAH Blvd.
TRADITIONAL ANALYTICS DISTRICT
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
Custom Scientific Format
CSV Ave.
JSON St.
Excel St.
SQL
Databases
Oracle Park
NetCDF St.
Standard Files
BUSSINESS AREA
LARGE SCIENTIFIC INSTALLATIONS
Space Exploration Facility
Radio Telescope Project
Database Packages
Particle Physics Lab
Custom Analytics Frameworks
Math & Statistic Faculty
Ph.D Schools
Root st.
Vanilla R
Python
Matlab
Scikit Learn Park
Pandas st.
School of Engineering
Computing Faculty
UNIVERSITY CAMPUS
Cran Highway
Caret Rd.
ggplot2 St.
tidyr St.
Data Exploration Line
Skills Line
Data Location Line
Custom Frameworks Line
In-Database Line
TRANSFER LINES
Extra Cost Applied (New Skills)
Extra Cost Applied (Data Movements)
Super-Reduced Extra Cost
Extra Cost Applied
UNDER CONSTRUCTION LINES
In-Memory Line
Hadoop Line

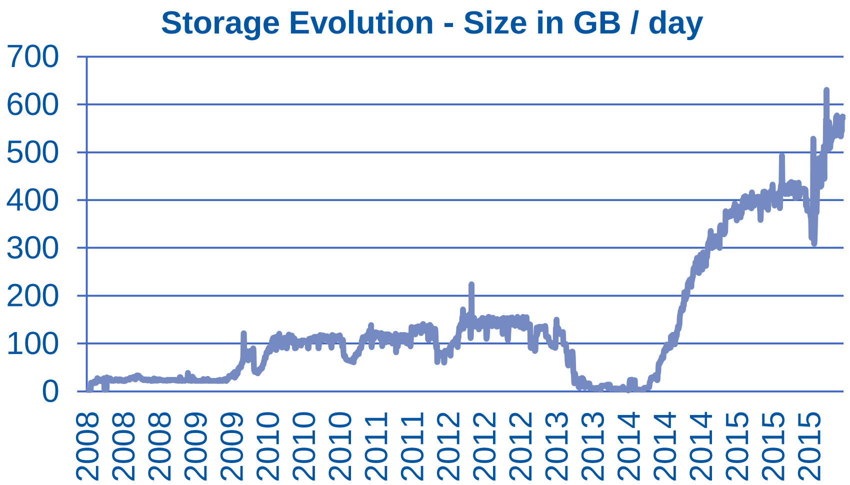

ETL Island
Kafka
Zookeeper House.
Avro St.
Parquet St.
hbase
hive
hdfs
SCALABLE ANALYTICS DISTRICT
INNOVATION PARK
Map Reduce
Mahout
Hadoop
RDD st.
Data Frame st.
LOCAL ZOO
BUSSINESS AREA
Spark
Oracle Rac Ave.
Oracle R Entrerprise
Oracle Rac Ave.
ORE Rd.
Oracle R Advance Analytics Hadoop
ORAAH Blvd.
TRADITIONAL ANALYTICS DISTRICT
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
SQL
Databases
Custom Scientific Format
CSV Ave.
JSON St.
Excel St.
Oracle Park
NetCDF St.
Standard Files
BUSSINESS AREA
LARGE SCIENTIFIC INSTALLATIONS
Space Exploration Facility
Radio Telescope Project
Database Packages
Particle Physics Lab
Custom Analytics Frameworks
Root st.
Math & Statistic Faculty
Ph.D Schools
Vanilla R
Python
Matlab
Scikit Learn Park
Pandas st.
School of Engineering
Computing Faculty
UNIVERSITY CAMPUS
Cran Highway
Caret Rd.
ggplot2 St.
tidyr St.
Data Exploration Line
Skills Line
Data Location Line
Custom Frameworks Line
In-Database Line
TRANSFER LINES
Extra Cost Applied (New Skills)
Extra Cost Applied (Data Movements)
Super-Reduced Extra Cost
Extra Cost Applied
UNDER CONSTRUCTION LINES
In-Memory Line
Hadoop Line



ETL Island
Kafka
Zookeeper House.
Avro St.
Parquet St.
hbase
hive
hdfs
SCALABLE ANALYTICS DISTRICT
INNOVATION PARK
Map Reduce
Mahout
Hadoop
LOCAL ZOO
BUSSINESS AREA
RDD st.
Data Frame st.
Jupyter St
Spark
Spark MLlib
Zeppelin St
Spark Streaming
Spark ML
Oracle Rac Ave.
Oracle R Entrerprise
Oracle Rac Ave.
ORE Rd.
Oracle R Advance Analytics Hadoop
ORAAH Blvd.
TRADITIONAL ANALYTICS DISTRICT
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
Custom Scientific Format
NetCDF St.
CSV Ave.
JSON St.
Excel St.
SQL
Databases
Oracle Park
Standard Files
BUSSINESS AREA
LARGE SCIENTIFIC INSTALLATIONS
Space Exploration Facility
Radio Telescope Project
Database Packages
Particle Physics Lab
Custom Analytics Frameworks
Root st.
Math & Statistic Faculty
Ph.D Schools
Vanilla R
Python
Matlab
Scikit Learn Park
Pandas st.
School of Engineering
Computing Faculty
UNIVERSITY CAMPUS
Cran Highway
Caret Rd.
ggplot2 St.
tidyr St.
Data Exploration Line
Skills Line
Data Location Line
Custom Frameworks Line
In-Database Line
TRANSFER LINES
Extra Cost Applied (New Skills)
Extra Cost Applied (Data Movements)
Super-Reduced Extra Cost
Extra Cost Applied
UNDER CONSTRUCTION LINES
In-Memory Line
Hadoop Line

CERN Accelerator Logging Service 2.0

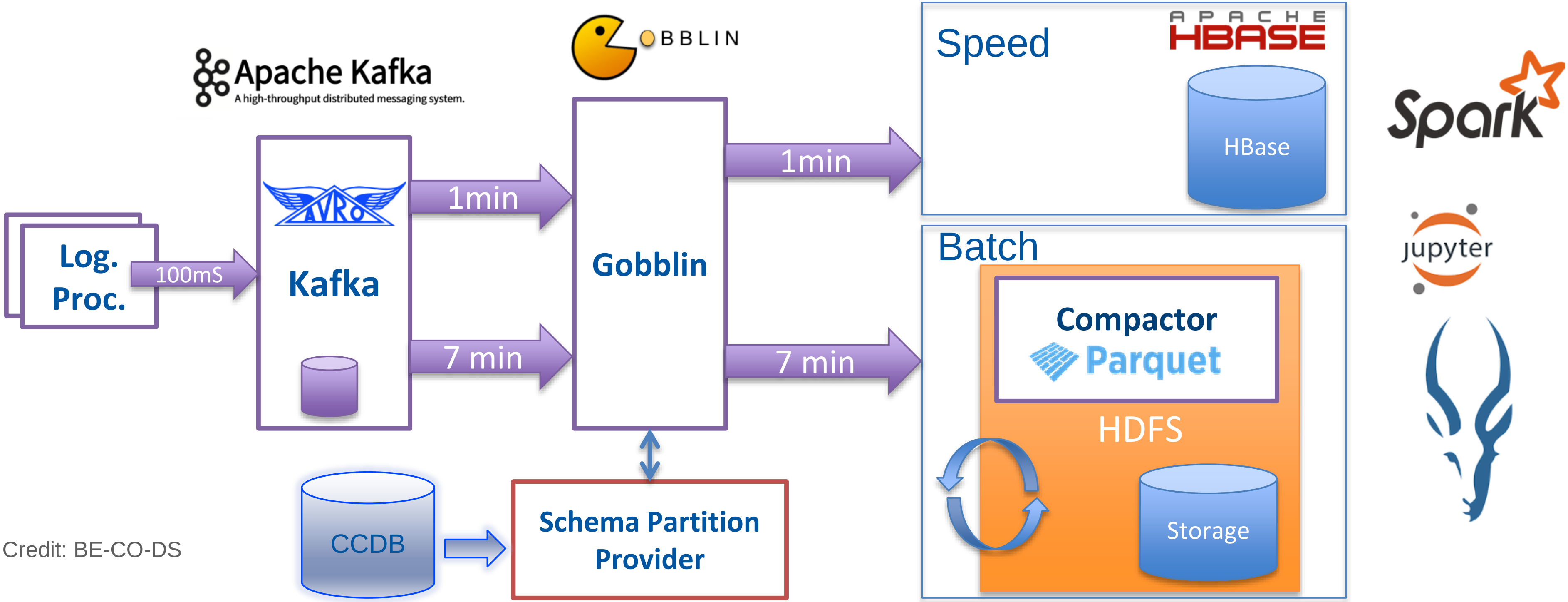
- New Landscape bring new challenges
  - Better Performance on bigger datasets
    - Big Data queries: Impala, Spark SQL
  - Leverage analytics capabilities
    - Spark Analytics: Python, ML, R
  - More heterogeneous data access models



QPS

Credit: BE-CO-DS

CERN Accelerator Logging Service



Credit: BE-CO-DS

Cryo Valves – ORAAH

Oracle R Advanced Analytics for Hadoop

```r
library(rpart)
library(ORCH)

#Setting up the connection to Spark installation
spark.connect(master="yarn-client",memory="8G",dfs.namenode="-----------")

#Preparing the data on HDFS
train  <- hdfs.put(training)
test <- hdfs.put(testing)

#Let focus now in model generation and training
#Spark MLlib algorithms available from R in ORAAH (source data can be CSVs in HDFS or HIVE tables)
model  <- orch.ml.svm(formula = status ~ rope_dist + bs + mean + var + max ,data = train)

#Predincting on test and writing back the results to HDFS
pred   <- predict(model, newdata = test)
hdfs.write(pred, outPath = "Prediction")
```

Our experience

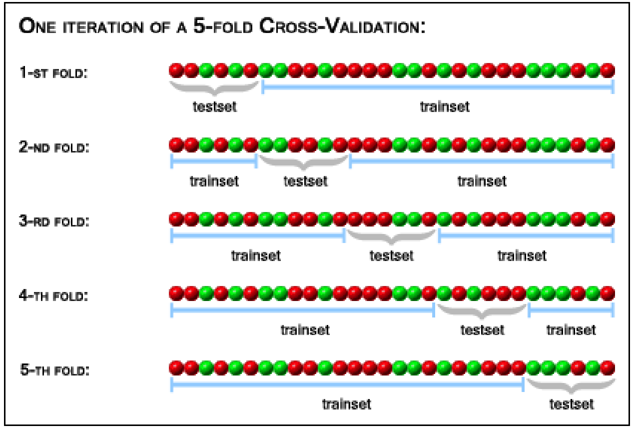
- No need to move data – the analysis in done where the data is
  - Access to database or Hive tables transparently
- Memory is not a problem anymore
  - The analysis is not anymore limited by the dataset size
  - Simplifies the infrastructure – transparent use of hadoop technologies
- Write/adapt R code is straight forward
  - Same principals than ORE – no need to acquire a new set of skills
  - Background use of Spark machine learning capabilities
- Limited functionality
  - When using ORAAH for Machine learning in scalable way the functionality is limited to Spark Machine Learning libraries
  - No as fast pace as Spark itself – Why do we need to wait
- Commercial VS open source

Machine Learning with Spark

- Why Apache Spark for Machine Learning
  - No memory limitations
  - Compatibility – Scala, R, Python
  - General purpose
    - Not only machine learning also **advanced data preparation, feature engineering, parameter tuning and model selection** etc.
- New Skills required
  - RDD-based API (spark.mllib)
  - DataFrame-based API (spark.ml)
  - Pipelines Concept (CARET package does in R)
    - Cross Validation
    - Parameter tuning using parameter grid
- Fast Pace Evolution

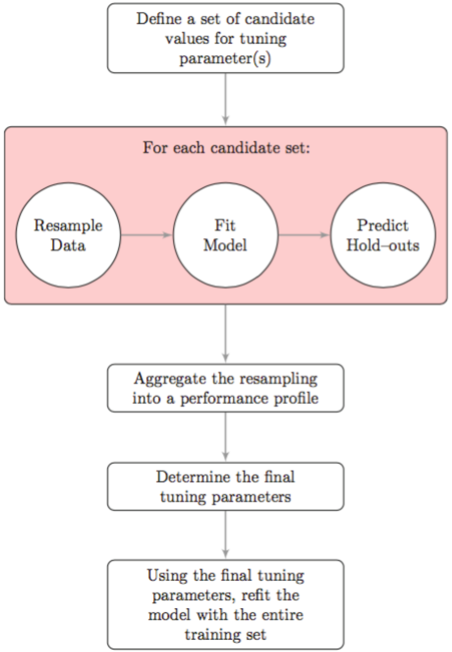
Machine Learning with Spark

- Cross Validation
  - Repeat the construction of the model on different subsets of the available training data and then evaluate the model only on data not seen during training



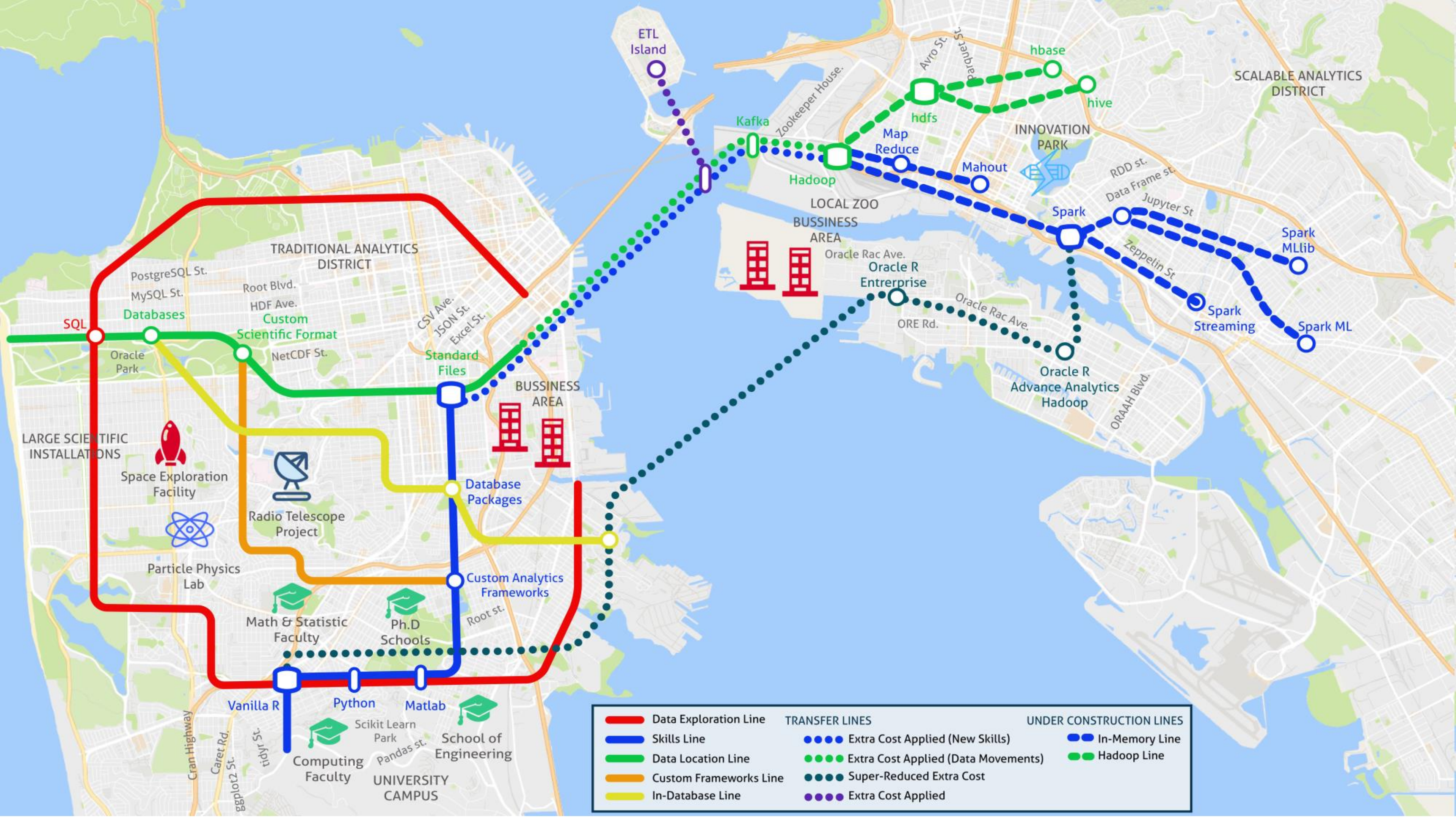
Machine Learning with Spark

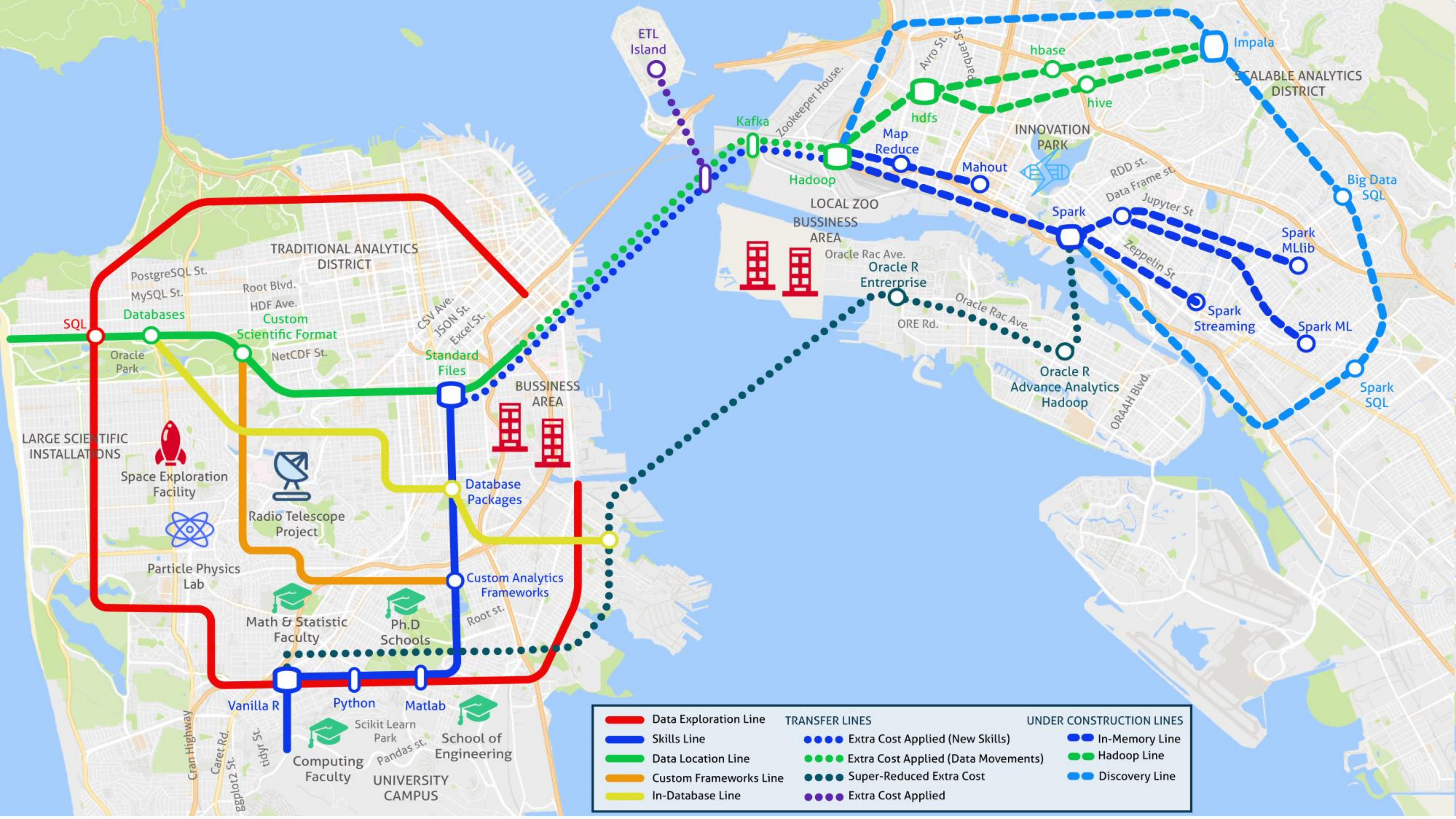
- ## Model Tuning
  - ML models have several parameter
    - there is no analytics formula to calculate appropriate values
  - These parameters control the complexity of the model
    - bad performance
    - over-fitting
    - etc.



Cryo Valves – Spark

```
In [ ]: from pyspark.ml.classification import RandomForestClassifier

        #select the features to used
        features = ['var, max, min, rope_dist, bs']


        #We have to shape our dataset before we can use some Spark ML algorithms with it.
        #First, we need to create a new column containing a Vector with all the features.
        #That vector will be the input features used by the ML for the model training.
        assembler = VectorAssembler(inputCols=features, outputCol="features")

        rfClassifierCV = RandomForestClassifier(labelCol="status", featuresCol="features")

        #Setting the pipeline
        rfPipelineCV = Pipeline(stages=[assembler, rfClassifierCV])


        #Prepare de evaluator, CV configuration and parameter tuning
        evaluatorCV = BinaryClassificationEvaluator(labelCol="indexedLabel")

        #3 different models will be evaluated 5,10 and 15 number of trees
        paramGrid = ParamGridBuilder().addGrid(rfClassifierCV.numTrees, [5,10,15]).build()

        #Definition of a the CV pofile
        crossValidator = CrossValidator(estimator=rfPipelineCV, estimatorParamMaps=paramGrid, evaluator=evaluatorCV, numFolds=3

        #Now we perform the computation all the prevoius code is lazy evaluated
        rfClassifierCV = crossValidator.fit(train)
        rfPredictionCV = cvModel.transform(test)
```


ETL Island
Kafka
Zookeeper House.
Avro St.
Parquet St.
hbase
hive
hdfs
SCALABLE ANALYTICS DISTRICT
INNOVATION PARK
Map Reduce
Mahout
Hadoop
LOCAL ZOO
BUSSINESS AREA
RDD st.
Data Frame st.
Jupyter St
Spark
Spark MLlib
Zeppelin St
Spark Streaming
Spark ML
Oracle Rac Ave.
Oracle R Entrerprise
ORE Rd.
Oracle Rac Ave.
Oracle R Advance Analytics Hadoop
ORAAH Blvd.
TRADITIONAL ANALYTICS DISTRICT
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
Custom Scientific Format
NetCDF St.
CSV Ave.
JSON St.
Excel St.
SQL
Databases
Oracle Park
Standard Files
BUSSINESS AREA
LARGE SCIENTIFIC INSTALLATIONS
Space Exploration Facility
Radio Telescope Project
Particle Physics Lab
Database Packages
Custom Analytics Frameworks
Root st.
Math & Statistic Faculty
Ph.D Schools
Vanilla R
Python
Matlab
Scikit Learn Park
Pandas st.
School of Engineering
Computing Faculty
UNIVERSITY CAMPUS
Cran Highway
Caret Rd.
ggplot2 St.
tidyr St.
Data Exploration Line
Skills Line
Data Location Line
Custom Frameworks Line
In-Database Line
TRANSFER LINES
Extra Cost Applied (New Skills)
Extra Cost Applied (Data Movements)
Super-Reduced Extra Cost
Extra Cost Applied
UNDER CONSTRUCTION LINES
In-Memory Line
Hadoop Line



ETL Island
Kafka
Zookeeper House.
Avro St.
Parquet St.
hbase
Impala
SCALABLE ANALYTICS DISTRICT
hdfs
hive
INNOVATION PARK
Map Reduce
Hadoop
Mahout
LOCAL ZOO
BUSSINESS AREA
RDD st.
Data Frame st.
Jupyter St
Big Data SQL
Spark
Spark MLlib
Zeppelin St
Spark Streaming
Spark ML
Oracle Rac Ave.
Oracle R Entrerprise
Oracle Rac Ave.
ORE Rd.
Oracle R Advance Analytics Hadoop
ORAAH Blvd.
Spark SQL
TRADITIONAL ANALYTICS DISTRICT
PostgreSQL St.
MySQL St.
Root Blvd.
HDF Ave.
Custom Scientific Format
NetCDF St.
CSV Ave.
JSON St.
Excel St.
SQL
Databases
Oracle Park
Standard Files
BUSSINESS AREA
LARGE SCIENTIFIC INSTALLATIONS
Space Exploration Facility
Radio Telescope Project
Database Packages
Particle Physics Lab
Custom Analytics Frameworks
Math & Statistic Faculty
Ph.D Schools
Root st.
Vanilla R
Python
Matlab
Scikit Learn Park
Pandas st.
School of Engineering
Computing Faculty
UNIVERSITY CAMPUS
Cran Highway
Caret Rd.
ggplot2 St.
tidyr St.
Data Exploration Line
Skills Line
Data Location Line
Custom Frameworks Line
In-Database Line
TRANSFER LINES
Extra Cost Applied (New Skills)
Extra Cost Applied (Data Movements)
Super-Reduced Extra Cost
Extra Cost Applied
UNDER CONSTRUCTION LINES
In-Memory Line
Hadoop Line
Discovery Line

Machine Learning with TensorFlow

- Why TensorFlow for Machine Learning
  - Spark machine learning capabilities are really limited
    - Number of models
    - Customization capabilities
- Overcome in term of performance any of the previous technologies
  - Spark is slow on training models
- State-of-the-art algorithms available
  - Deep-learning
- New skill need to be understood
  - Tensor concept
- Model freedom comes with a price
  - Coding

Machine Learning with TensorFlow+Spark

- No memory limitations
  - Bigger than memory datasets treated transparently
- Parallelization
  - Tensorflow profit from Spark partitioning concepts to improve the user control over parallelization