HPCC Systems workshop – Day 2

OCTOBER 2022

H. Watanuki
LexisNexis RISK Solutions

Day 2 – October 17th 5:00 pm (GMT-3)

1. Two-hour workshop via MS Teams (invites to be sent by the instructor).

2. Resources:
    1. **PC and Github (https://github.com) account**
    2. **HPCC Systems cluster**

3. Core content:

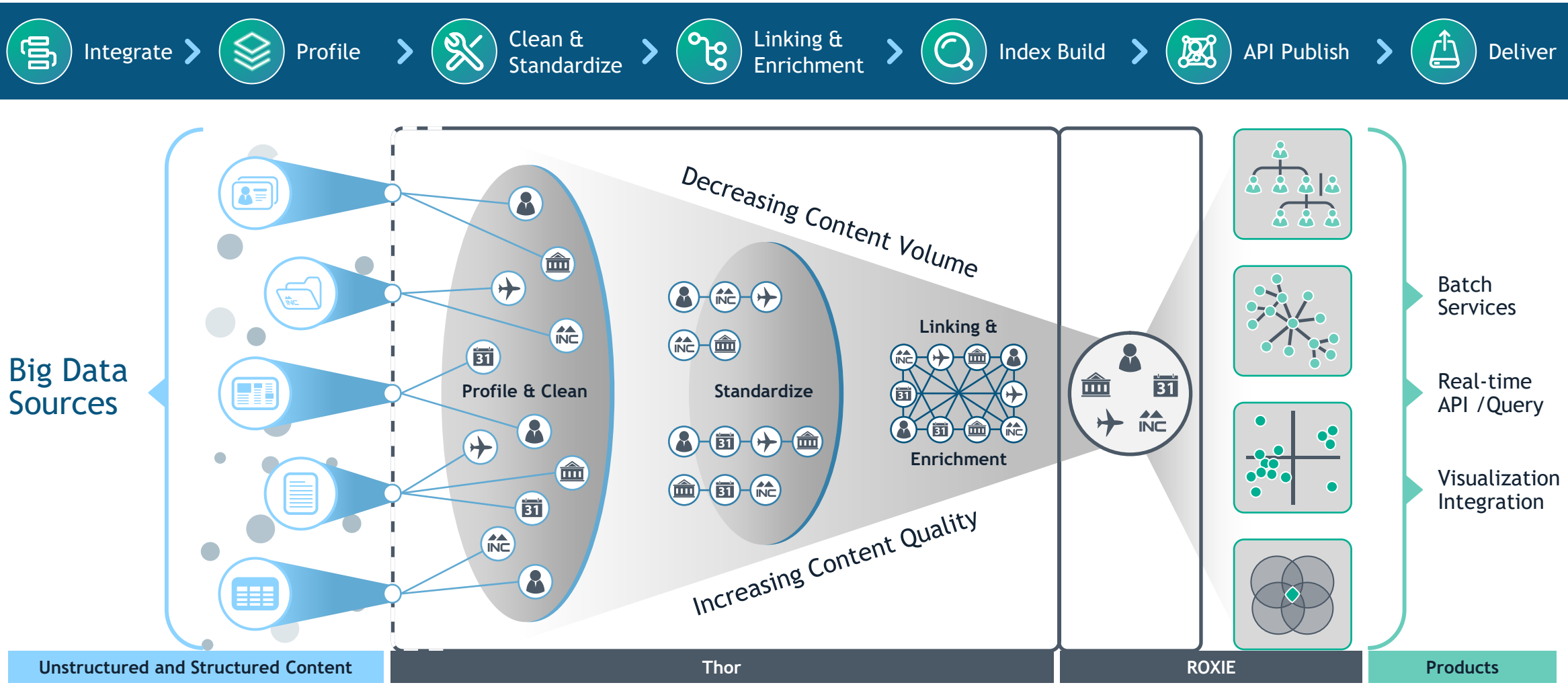
    1. **Recap fundamentals of HPCC Systems:**
        1. **Background and System overview: 15 mins**

    2. **Machine Learning with HPCC Systems:**
        1. **Main ECL primitives for data preparation (data frame): 45 mins**
        2. **Model training and assessment: 45 mins**

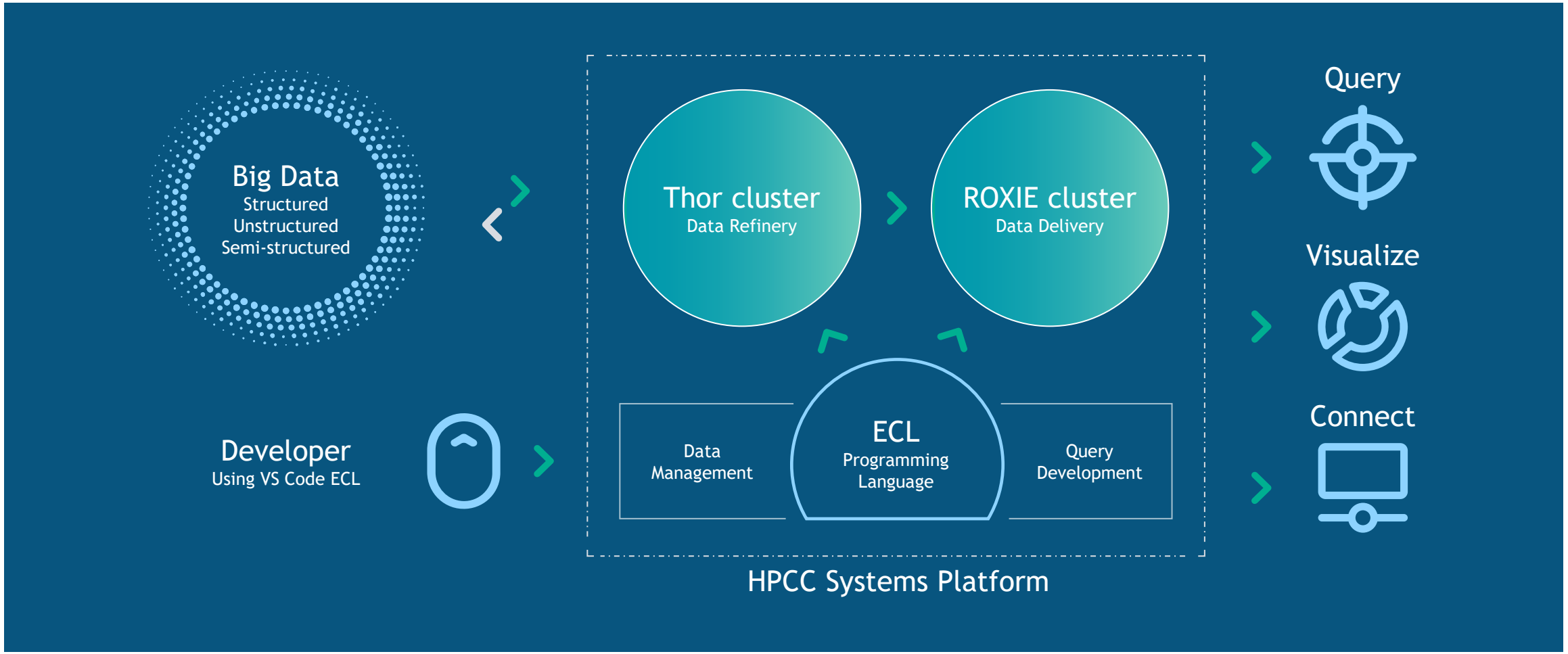
    3. **Q&A (if possible, to be submitted upfront)**
        1. **Specific questions from audience: 15 mins**

Recap: HPCC Systems overview

HPCC Systems Data Enrichment Pipeline



The HPCC Systems Engines



Machine Learning with HPCC Systems

It’s a Machine Learning World

Classical Machine Learning

Unsupervised

Clustering
DBSCAN
K-Means

Pattern Search
Text Vectors

Dimension Reduction
PCA

Supervised

Classification
SVM
Decision Trees
Logistic Regression
Classification Forest

Regression
Linear Regression
Regression Forest

Neural Nets & Deep Learning

Autoencoders/Perceptrons

Convolutional Neural Networks

Recurrent Neural Networks

Generalized Neural Networks

Ensemble Methods

Random Forest

Gradient Boosted Forest

Boosted Trees

Machine Learning Basics

Terminology overview:

Given a set of data samples:

```
Record1: Field1, Field2, Field3, … , FieldM
Record2: Field1, Field2, Field3, … , FieldM
…
RecordN: Field1, Field2, Field3, …, FieldM
```

Independent Variables

Note: The fields in the independent data are also known as “features” of the data

And a set of target values,

```
Record1: TargetValue
Record2: TargetValue
…
RecordN: TargetValue
```

Dependent Variables

Learn how to predict target values for *new* samples.

- The set of Independent and Dependent data is known as the **Training Set**
- The encapsulated learning is known as the **Model**
- Each model represents a **Hypothesis** regarding the relationship of the Independent to Dependent variables.

Machine Learning Conceptual Example

Given the following data about trees in a forest:

Height	Diameter	Altitude	Rainfall	Age
50	8	5000	12	80
56	9	4400	10	75
72	12	6500	18	60
47	10	5200	14	53

Learn a Model that approximates Age (i.e., the *Dependent* Variable) from Height, Diameter, Altitude, and Rainfall (i.e., the *Independent* Variables).

Machine Learning Pipeline

1. Problem definition
2. Data extraction
3. Data preparation
4. Data conversion and split
5. Model training and assessment
6. Model implementation

Demo: Machine Learning pipeline

1. Problem definition

“Given a set of attributes from a property (address, sq. ft., year built), how to predict its commercial value?"

propertyid	house_numbe	house_nu	predir	street	streett	postdir	apt	city	state	zip	total_value	assessed_value	year_acquired	land_square_foot	living_square_fee	bedrooms	full_baths
828195	144			MCKIERNAN	DR			WALNUT CREEK	CA	94597	62614	62614	2006	20418	2485	3	2
1144455	281			CENTER	ST			BALTIMORE	MD	21136	105500	10550	2007	4807	1368	0	0
1494347	483			NEWTON	RD			FLAGSTAFF	AZ	86011	2220	2220	0	5654	1011	3	1
1910847	802			HATCHERY	CT			WOODLAND	WA	98674	356000	356000	0	6094	0	2	1
4267562	5007		E	ROY ROGERS	RD			TROY	MI	48085	327253	327253	2007	3484	0	3	0
4888602	7607			PEBBLESTONE	DR		000009	KERNVILLE	CA	93238	732179	732179	2010	19597	6132	6	6
48725	4			LONG	AVE			SUNRISE	FL	33323	271000	271000	2008	6880	2392	4	2
83528	6			TRILLUM	LN			WAYLAND	MA	02193	79889	79889	2007	7657	1657	4	1
94604	7			PARMENTER	AVE			PLYMOUTH	MN	55441	23800	23800	2005	19994	1754	3	2
220326	17			TIMBER	RD			LOS ANGELES	CA	90063	89000	89000	2008	7840	954	3	1
994609	212			FREYER	DR	NE		PHILOMONT	VA	20131	59800	59800	2009	11199	1241	3	0
1836173	724			EASTER	ST			ALLENTOWN	PA	18102	191600	191600	0	9100	2534	4	2
2910797	1903			SADDLE BROOK	DR			CLIO	CA	96106	61610	61610	2007	0	0	0	0
3083959	2158			RIVERSIDE	DR			UPPER MORELA…	PA	19006	90300	0	0	0	1235	3	2
3952189	4040			GRAND VIEW	BLVD		000054	RIO LINDA	CA	95673	0	0	0	2700720	0	0	0
4186238	4726			LAS PALMAS	CT			WAELDER	TX	78959	18816	18816	2009	2159	1320	0	0
4597143	6213			WILSON	RD			ZOLFO SPRINGS	FL	33890	72600	0	0	8496	0	3	1
4624905	6321			STONEWALL	LN			PATERSON	NJ	07514	139880	139880	2008	10454	1391	4	2
92326	7			KNOLLCREST	DR			NARANJA	FL	33032	76214	76214	2008	4800	930	2	0
1792852	704			ERIN	DR			TRABUCO	CA	92678	28010	28010	2007	5200	0	3	1
1843977	728		S	ARLINGTON HE…	RD			BLOOMING GRO…	TX	76626	130400	130400	2007	36154	1629	3	1
4214872	4821			MYRTLE OAK	DR		000025	SAN BERNARDI	CA	92376	33250	0	2007	93654	0	0	0

2. Data Extraction

RECORD structure and DATASET declaration

```
EXPORT File_Property := MODULE
     EXPORT Layout := RECORD
          INTEGER8      propertyid;
          STRING5       streettype;
          STRING40      city;
          STRING2       state;
          STRING5       zip;
          UNSIGNED4     total_value;
          UNSIGNED4     assessed_value;
          UNSIGNED2     year_acquired;
          UNSIGNED4     land_square_footage;
          UNSIGNED4     living_square_feet;
          UNSIGNED2     bedrooms;
          UNSIGNED2     full_baths;
          UNSIGNED2     half_baths;
          UNSIGNED2     year_built;
     END;
     EXPORT File := DATASET('~online::hmw::AdvECL::property',Layout,THOR);
END;
```

```
EXPORT MLProp := RECORD
          UNSIGNED8 PropertyID;
          UNSIGNED3 zip;                          //categorical
          UNSIGNED4 assessed_value;
          UNSIGNED2 year_acquired;
          UNSIGNED4 land_square_footage;
          UNSIGNED4 living_square_feet;
          UNSIGNED2 bedrooms;
          UNSIGNED2 full_baths;
          UNSIGNED2 half_baths;
          UNSIGNED2 year_built;
          UNSIGNED4 total_value;       //Dependent variable
     END;
```

2. Data Extraction

##	personid	propertyid	house_number	house_number_suffix	predir	street	streettype	postdir	apt	city	state	zip	total_value
1	187522928604396	828195	144			MCKIERNAN	DR			WALNUT CREEK	CA	94597	62614
2	187522928604396	1144455	281			CENTER	ST			BALTIMORE	MD	21136	105500
3	187522928604396	1494347	483			NEWTON	RD			FLAGSTAFF	AZ	86011	2220
4	187522928604396	1910847	802			HATCHERY	CT			WOODLAND	WA	98674	356000
5	187522928604396	4267562	5007		E	ROY ROGERS	RD			TROY	MI	48085	327253
6	187522928604396	4888602	7607			PEBBLESTONE	DR		000009	KERNVILLE	CA	93238	732179
7	1258313199446079	48725	4			LONG	AVE			SUNRISE	FL	33323	271000
8	1258313199446079	83528	6			TRILLUM	LN			WAYLAND	MA	02193	79889
9	1258313199446079	94604	7			PARMENTER	AVE			PLYMOUTH	MN	55441	23800
10	1258313199446079	220326	17			TIMBER	RD			LOS ANGELES	CA	90063	89000
11	1258313199446079	994609	212			FREYER	DR	NE		PHILOMONT	VA	20131	59800
12	1258313199446079	1836173	724			EASTER	ST			ALLENTOWN	PA	18102	191600
13	1258313199446079	2910797	1903			SADDLE BROOK	DR			CLIO	CA	96106	61610
14	1258313199446079	3083959	2158			RIVERSIDE	DR			UPPER MORELAND	PA	19006	90300
15	1258313199446079	3952189	4040			GRAND VIEW	BLVD		000054	RIO LINDA	CA	95673	0

3. Data Preparation

PROJECT(), RANDOM() and SORT()

```
// Clean the data and assign a random number to each record

CleanFilter := Property.zip <> '' AND Property.assessed_value <> 0 AND Property.year_acquired <> 0
               AND Property.land_square_footage <> 0 AND Property.living_square_feet <> 0
               AND  Property.bedrooms <> 0 AND Property.year_Built <> 0;

MLPropExt := RECORD(ML_Prop)
     UNSIGNED4 rnd; // A random number
END;

EXPORT myDataE := PROJECT(Property(CleanFilter), TRANSFORM(MLPropExt,
                                                           SELF.rnd := RANDOM(),
                                                           SELF.Zip := (UNSIGNED3)LEFT.Zip,
                                                           SELF := LEFT)) ;

// Shuffle your data by sorting on the random field
SHARED myDataES := SORT(myDataE, rnd);

// Treat first 5000 as training data.  Transform back to the original format.
EXPORT myTrainData := PROJECT(myDataES[1..5000], ML_Prop);

// Treat next 2000 as test data
EXPORT myTestData  := PROJECT(myDataES[5001..7000], ML_Prop);
```

3. Data Preparation

##	propertyid	zip	assessed_value	year_acquired	land_square_footage	living_square_feet	bedrooms	full_baths	half_baths	year_built	total_value
1	79784	33424	76440	2015	4299	1255	3	2	0	2010	76440
2	3924129	20601	95900	2013	11224	1468	3	2	1	2007	95900
3	413843	8803	76000	2015	57000	1858	3	2	0	1970	76000
4	608224	98370	39340	2012	7405	1066	3	1	1	1967	39340
5	942963	72032	278400	2008	9600	2459	3	2	0	1963	278400
6	2237271	79935	143600	2011	8430	1008	2	1	1	1961	143600
7	4443742	84065	166934	2013	9317	1700	4	2	0	1991	166934
8	3834707	66227	348350	2012	15300	2663	4	2	1	2002	348350
9	3592739	19606	54000	2015	15060	2292	4	2	1	1980	90000
10	2916349	34639	119050	2015	6947	1709	3	2	0	2009	140950

4. Data Conversion and Split

ML_Core.ToField()

```
IMPORT $;
IMPORT ML_Core;

myTrainData := $.Prep01.myTrainData;
myTestData  := $.Prep01.myTestData;

//Numeric Field Matrix conversion
ML_Core.ToField(myTrainData, myTrainDataNF);
ML_Core.ToField(myTestData, myTestDataNF);

EXPORT Convert02 := MODULE
  EXPORT myIndTrainDataNF := myTrainDataNF(number < 10); //

  EXPORT myDepTrainDataNF := PROJECT(myTrainDataNF(number = 10),
                                     TRANSFORM(RECORDOF(LEFT),
                                               SELF.number := 1,
                                               SELF := LEFT));


  EXPORT myIndTestDataNF := myTestDataNF(number < 10);

  EXPORT myDepTestDataNF := PROJECT(myTestDataNF(number = 10),
                                    TRANSFORM(RECORDOF(LEFT),
                                              SELF.number := 1,
                                              SELF := LEFT));

END;
```

wi	id	number	value
1	160350	1	20706
1	160350	2	18020
1	160350	3	2007
1	160350	4	4610
1	160350	5	2594
1	160350	6	2
1	160350	7	2
1	160350	8	0
1	160350	9	1916
1	82569	1	60527
1	82569	2	78477
1	82569	3	2007
1	82569	4	6098
1	82569	5	1032
1	82569	6	3
1	82569	7	2
1	82569	8	0
1	82569	9	1992

wi	id	number	value
1	160350	1	185000
1	82569	1	78477
1	2192898	1	79290
1	2223942	1	45511
1	4648854	1	39900
1	2367580	1	108610
1	607178	1	31072
1	1584497	1	88284
1	3615520	1	341400
1	2103806	1	58520
1	2209348	1	87610
1	1298734	1	66175
1	1310023	1	301644
1	2840506	1	94200
1	3600022	1	262700
1	131449	1	16500
1	4649661	1	84000
1	1042629	1	38740
1	1732698	1	197700

4. Data Conversion and Split

##	wi	id	number	value
1	1	79784	1	33424.0
2	1	79784	2	76440.0
3	1	79784	3	2015.0
4	1	79784	4	4299.0
5	1	79784	5	1255.0
6	1	79784	6	3.0
7	1	79784	7	2.0
8	1	79784	8	0.0
9	1	79784	9	2010.0
10	1	3924129	1	20601.0

##	wi	id	number	value
1	1	79784	1	76440.0
2	1	3924129	1	95900.0
3	1	413843	1	76000.0
4	1	608224	1	39340.0
5	1	942963	1	278400.0
6	1	2237271	1	143600.0
7	1	4443742	1	166934.0
8	1	3834707	1	348350.0
9	1	3592739	1	90000.0
10	1	2916349	1	140950.0

5. Model training and assessment

GetModel() and Predict()

```
IMPORT LearningTrees AS LT;
IMPORT ML_Core;
IMPORT $;

//Training and Test data
XTrain := $.Convert02.myIndTrainDataNF;
YTrain := $.Convert02.myDepTrainDataNF;
XTest  := $.Convert02.myIndTestDataNF;
YTest  := $.Convert02.myDepTestDataNF;

//Train Boosted Forest model on Property data
myLearner    := LT.BoostedRegForest(,,,,,[1]); // Make the zipcode field a nominal (categorical) field.
myModel      := myLearner.GetModel(XTrain,YTrain);
OUTPUT(myModel,,'~mymodelXXX',NAMED('TrainedModel'),OVERWRITE); //Replace XXX by your initials

//Test Boosted Forest model on Property data
MyPredict := myLearner.Predict(myModel,XTest);
OUTPUT(MyPredict, NAMED('PredictedValues'));//workitem,uniqueid,field number, dependent value

//Assess Boosted Forest model on Property data
assessmentR2   := ML_Core.Analysis.Regression.Accuracy(MyPredict,YTest);
OUTPUT(assessmentR2, NAMED('Accuracy'));
```

5. Model training and assessment

	wi	value	indexes	fileposition
			Item	
1	0	4356.0	3	0
			10	
			1	
2	0	2812.0	3	27
			10	
			2	
3	0	2476.0	3	54
			10	
			3	
4	0	1244.0	3	81
			10	
			4	
5	0	1082.0	3	108
			10	
			5	
6	0	4085.0	3	135
			10	
			6	

##	wi	id	number	value
1	1	3634	1	59055.31318837311
2	1	5840	1	126151.3283316611
3	1	12721	1	150876.4676173128
4	1	47045	1	233897.4086392291
5	1	91757	1	111950.2604939628
6	1	117238	1	81157.13156934927
7	1	149746	1	75868.58107175257
8	1	239046	1	39961.17077444747
9	1	246517	1	128203.9088547347
10	1	252615	1	69009.47259550788

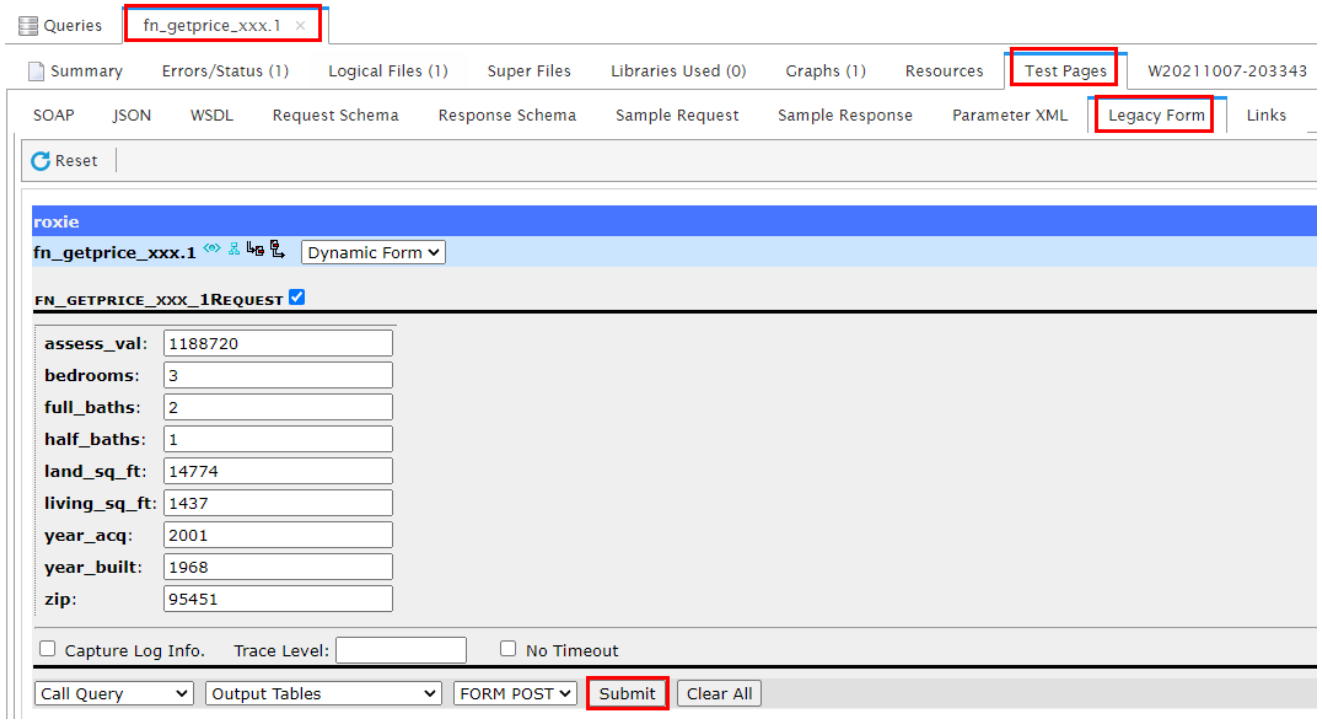
##	wi	regressor	r2	mse	rmse
1	1	1	0.7304899830671003	7982069594.129144	89342.4288573416

6. Model implementation

FUCTION() structure

```
IMPORT $;
IMPORT ML_Core;
IMPORT LearningTrees as LT;

EXPORT FN_GetPrice(Zip, Assess_val, Year_acq,
                           Land_sq_ft, Living_sq_ft, Bedrooms,
                           Full_baths, Half_baths, Year_built) := FUNCTION

     myInSet := [zip, assess_val, year_acq, land_sq_ft, living_sq_ft,
                                      bedrooms, full_baths, half_baths, year_built];

     myInDs := DATASET(myInSet, {REAL8 myInValue});

     ML_Core.Types.NumericField PrepData(RECORDOF(myInDS) Le, INTEGER C) := TRANSFORM
               SELF.wi          := 1,
               SELF.id                := 1,
               SELF.number := C,
               SELF.value       := Le.myInValue;
     END;
     myIndepData := PROJECT(myInDs, PrepData(LEFT,COUNTER));
     mymodel := DATASET('~mymodelXXX',ML_Core.Types.Layout_Model2,FLAT,PRELOAD);

     myLearner := LT.RegressionForest(10,,10,[1]);
     myPredictDeps := MyLearner.Predict(myModel, myIndepData);
     RETURN OUTPUT(myPredictDeps,{preco:=ROUND(value)});
END;
```

6. Model implementation



fn getprice xxx.1 Response

Dataset: Result 1

	preco
1	722902

Training and Support

https://hpccsystems.com/training/classes

IN-PERSON AS WELL AS
FREE ONLINE TRAINING

Learning tracks include:

- Core Platform
- Administration

Annual online Community Event including platform roadmap updates and training workshop

Monitored Stack Overflow channel for Support and Platform Issues

General Support available from Partners, including Clear Funnel, Infosys, and others

HPPC Systems is offering free courses to build your skills in working with ECL and big data!

INTRODUCTORY COURSES

Prerequisite for advanced courses

- Introduction to ECL (Part 1)
- Introduction to ECL (Part 2)

ADVANCED COURSES

$495 value each (FREE with promo)

- Advanced ECL (Part 1)
- Advanced ECL (Part 2)
- ROXIE ECL (Part 1)
- ROXIE ECL (Part 2)
- Applied ECL

HPCC SYSTEMS


LexisNexis®
RISK SOLUTIONS

Decision trees model

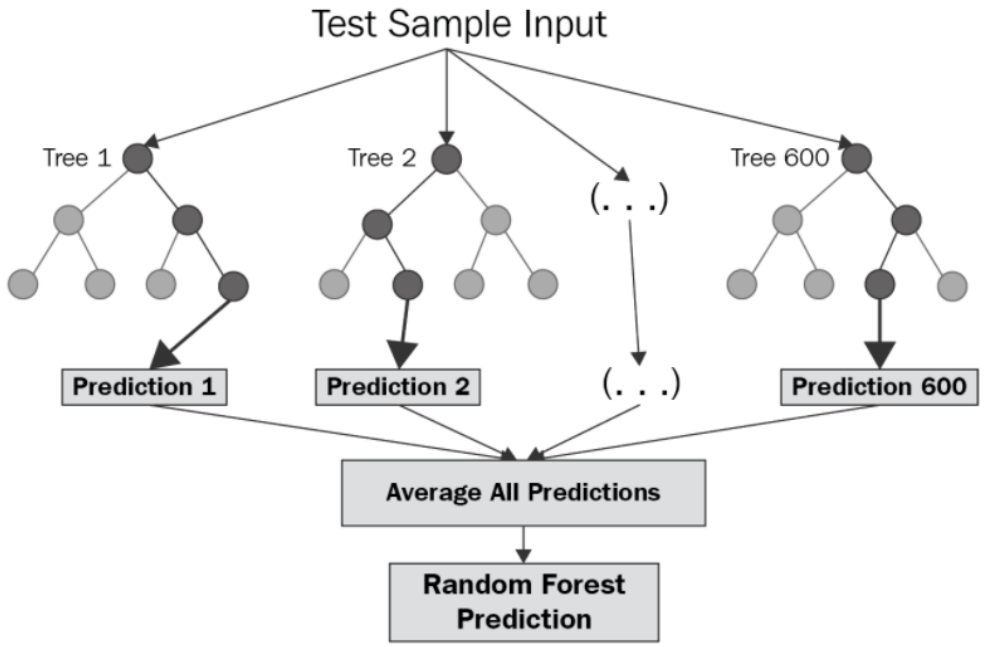
(https://hpccsystems.com/blog/learning-trees-guide-to-decision-tree-based-machine-learning)

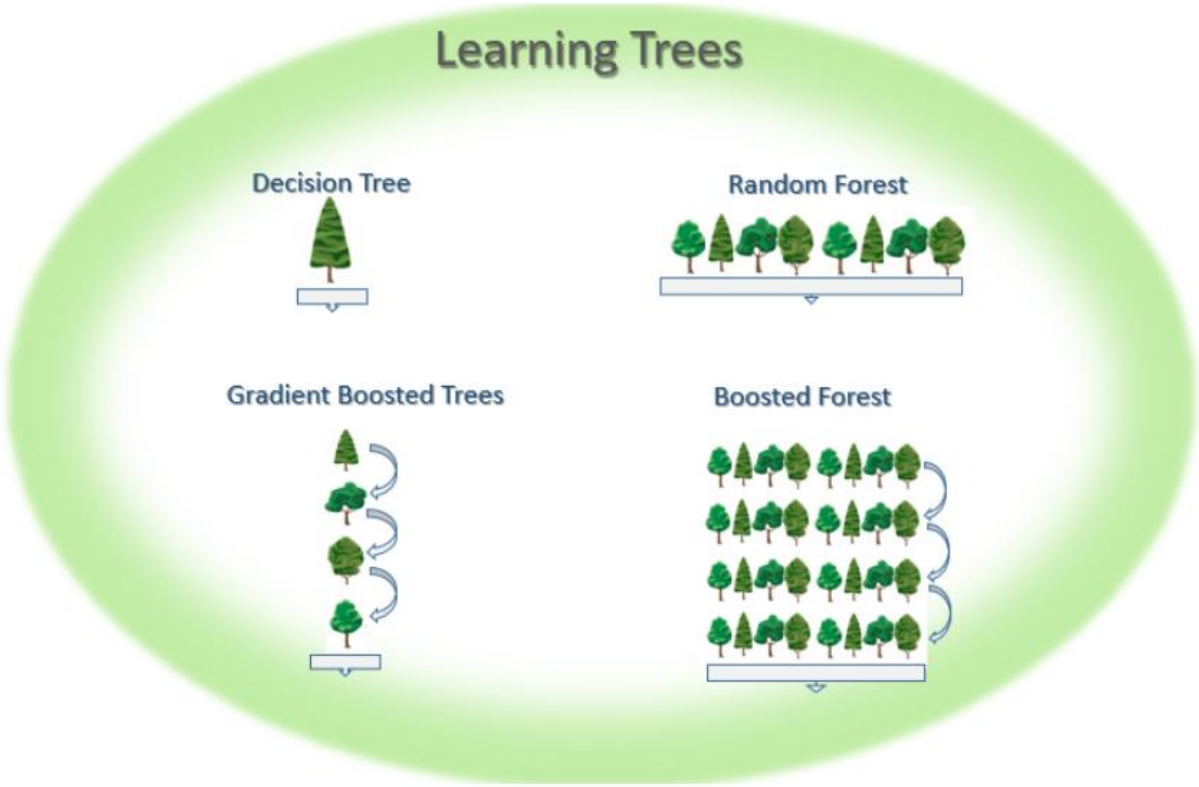
XOR Truth Table

Feature 1	Feature 2	Result
0	0	0
0	1	1
1	0	1
1	1	0

Ensemble methods

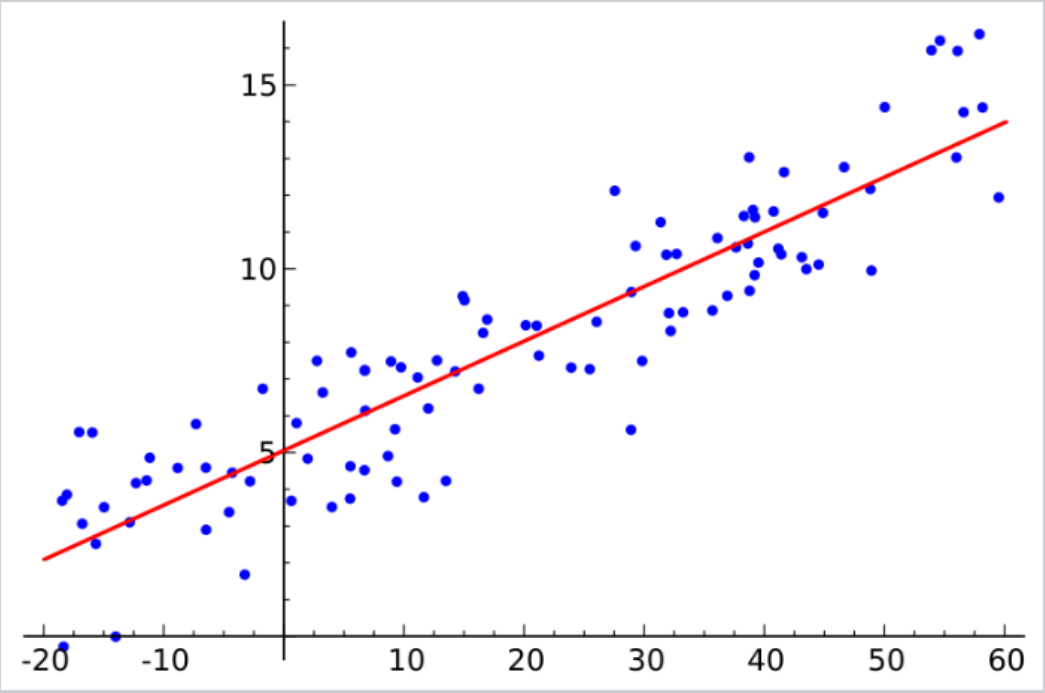
(https://hpccsystems.com/blog/learning-trees-guide-to-decision-tree-based-machine-learning)





Linear Regression

(https://github.com/hpcc-systems/LinearRegression.git)



Ref: https://en.wikipedia.org/wiki/Linear_regression

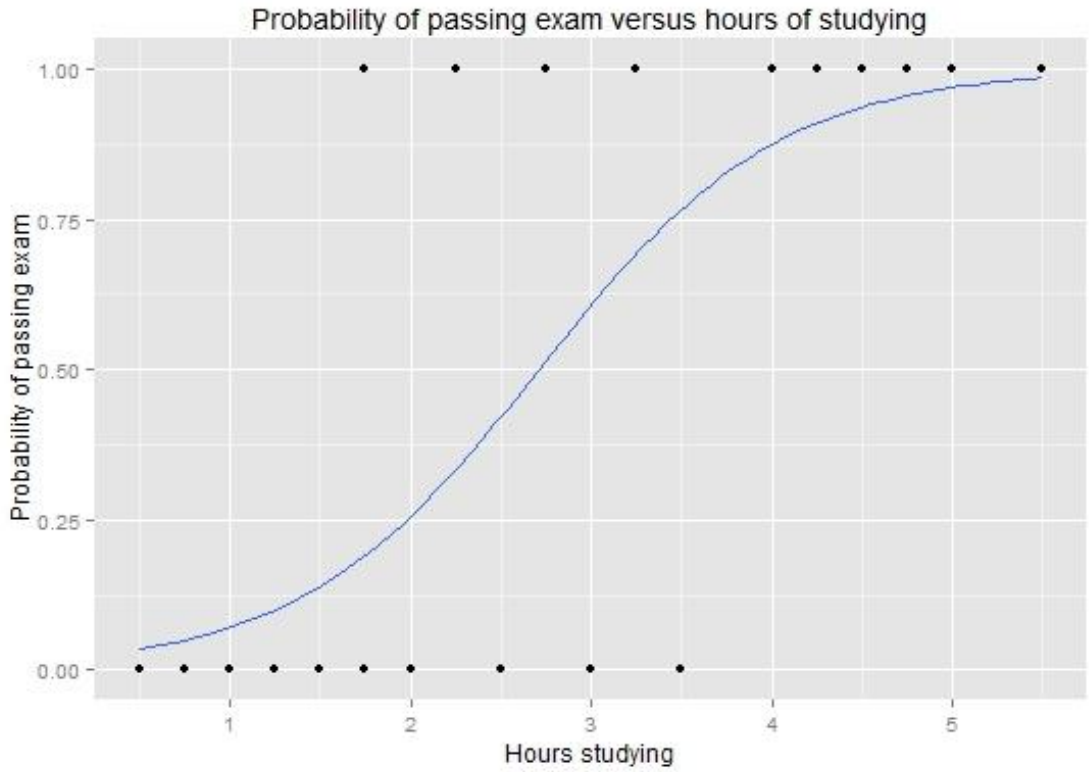
Example

“Given a set of attributes from a property (address, sq. ft., year built), how to predict its commercial value?"

propertyid	house_numbe	house_nu	predir	street	streett	postdir	apt	city	state	zip	total_value	assessed_value	year_acquired	land_square_foot	living_square_fee	bedrooms	full_baths
828195	144			MCKIERNAN	DR			WALNUT CREEK	CA	94597	62614	62614	2006	20418	2485	3	2
1144455	281			CENTER	ST			BALTIMORE	MD	21136	105500	10550	2007	4807	1368	0	0
1494347	483			NEWTON	RD			FLAGSTAFF	AZ	86011	2220	2220	0	5654	1011	3	1
1910847	802			HATCHERY	CT			WOODLAND	WA	98674	356000	356000	0	6094	0	2	1
4267562	5007		E	ROY ROGERS	RD			TROY	MI	48085	327253	327253	2007	3484	0	3	0
4888602	7607			PEBBLESTONE	DR		000009	KERNVILLE	CA	93238	732179	732179	2010	19597	6132	6	6
48725	4			LONG	AVE			SUNRISE	FL	33323	271000	271000	2008	6880	2392	4	2
83528	6			TRILLUM	LN			WAYLAND	MA	02193	79889	79889	2007	7657	1657	4	1
94604	7			PARMENTER	AVE			PLYMOUTH	MN	55441	23800	23800	2005	19994	1754	3	2
220326	17			TIMBER	RD			LOS ANGELES	CA	90063	89000	89000	2008	7840	954	3	1
994609	212			FREYER	DR	NE		PHILOMONT	VA	20131	59800	59800	2009	11199	1241	3	0
1836173	724			EASTER	ST			ALLENTOWN	PA	18102	191600	191600	0	9100	2534	4	2
2910797	1903			SADDLE BROOK	DR			CLIO	CA	96106	61610	61610	2007	0	0	0	0
3083959	2158			RIVERSIDE	DR			UPPER MORELA…	PA	19006	90300	0	0	0	1235	3	2
3952189	4040			GRAND VIEW	BLVD		000054	RIO LINDA	CA	95673	0	0	0	2700720	0	0	0
4186238	4726			LAS PALMAS	CT			WAELDER	TX	78959	18816	18816	2009	2159	1320	0	0
4597143	6213			WILSON	RD			ZOLFO SPRINGS	FL	33890	72600	0	0	8496	0	3	1
4624905	6321			STONEWALL	LN			PATERSON	NJ	07514	139880	139880	2008	10454	1391	4	2
92326	7			KNOLLCREST	DR			NARANJA	FL	33032	76214	76214	2008	4800	930	2	0
1792852	704			ERIN	DR			TRABUCO	CA	92678	28010	28010	2007	5200	0	3	1
1843977	728		S	ARLINGTON HE…	RD			BLOOMING GRO…	TX	76626	130400	130400	2007	36154	1629	3	1
4214872	4821			MYRTLE OAK	DR		000025	SAN BERNARDI	CA	92376	33250	0	2007	93654	0	0	0

Binomial logistic regression

(https://github.com/hpcc-systems/LogisticRegression.git)



Ref: By Michaelg2015 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=42442194

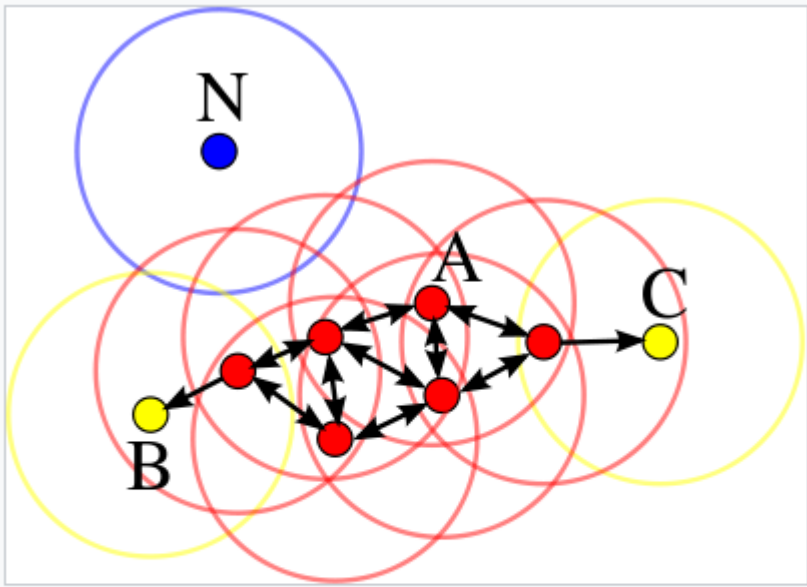
Example

“Given a set of attributes from marketing phone calls, how to predict if a client will decide to make an investment?"

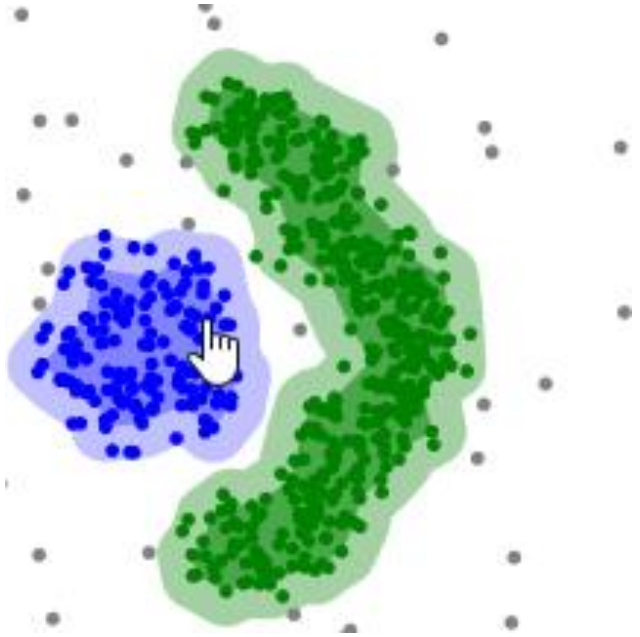
age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp_var_rate	cons_price_idx	cons_conf_idx	euribor3m	nr_employed	y
44	blue-collar	married	basic.4y	unknown	yes	no	cellular	aug	thu	210	1	999	0	nonexistent	1.4	93.444	-36.1	4.963	5228.1	0
53	technician	married	unknown	no	no	no	cellular	nov	fri	138	1	999	0	nonexistent	-0.1	93.2	-42	4.021	5195.8	0
28	management	single	university.degree	no	yes	no	cellular	jun	thu	339	3	6	2	success	-1.7	94.055	-39.8	0.729	4991.6	1
39	services	married	high.school	no	no	no	cellular	apr	fri	185	2	999	0	nonexistent	-1.8	93.075	-47.1	1.405	5099.1	0
55	retired	married	basic.4y	no	yes	no	cellular	aug	fri	137	1	3	1	success	-2.9	92.201	-31.4	0.869	5076.2	1
30	management	divorced	basic.4y	no	yes	no	cellular	jul	tue	68	8	999	0	nonexistent	1.4	93.918	-42.7	4.961	5228.1	0
37	blue-collar	married	basic.4y	no	yes	no	cellular	may	thu	204	1	999	0	nonexistent	-1.8	92.893	-46.2	1.327	5099.1	0
39	blue-collar	divorced	basic.9y	no	yes	no	cellular	may	fri	191	1	999	0	nonexistent	-1.8	92.893	-46.2	1.313	5099.1	0
36	admin.	married	university.degree	no	no	no	cellular	jun	mon	174	1	3	1	success	-2.9	92.963	-40.8	1.266	5076.2	1
27	blue-collar	single	basic.4y	no	yes	no	cellular	apr	thu	191	2	999	1	failure	-1.8	93.075	-47.1	1.41	5099.1	0
34	housemaid	single	university.degree	no	no	no	telephone	may	fri	62	2	999	0	nonexistent	1.1	93.994	-36.4	4.864	5191	0
41	management	married	university.degree	no	yes	no	cellular	aug	thu	789	1	999	0	nonexistent	1.4	93.444	-36.1	4.964	5228.1	0
55	management	married	university.degree	no	no	no	cellular	aug	mon	372	3	999	0	nonexistent	1.4	93.444	-36.1	4.965	5228.1	1
33	services	divorced	high.school	no	yes	no	cellular	may	tue	75	5	999	0	nonexistent	-1.8	92.893	-46.2	1.291	5099.1	0
26	admin.	married	high.school	no	no	yes	telephone	jun	mon	1021	1	999	0	nonexistent	1.4	94.465	-41.8	4.96	5228.1	0
52	services	married	high.school	unknown	yes	no	cellular	jul	thu	117	2	999	0	nonexistent	1.4	93.918	-42.7	4.962	5228.1	0

DBSCAN

(https://hpccsystems.com/blog/DBSCAN)



In this diagram, minPts = 4. Point A and the other red points are core points, because the area surrounding these points in an ε radius contain at least 4 points (including the point itself). Because they are all reachable from one another, they form a single cluster. Points B and C are not core points, but are reachable from A (via other core points) and thus belong to the cluster as well. Point N is a noise point that is neither a core point nor directly-reachable.



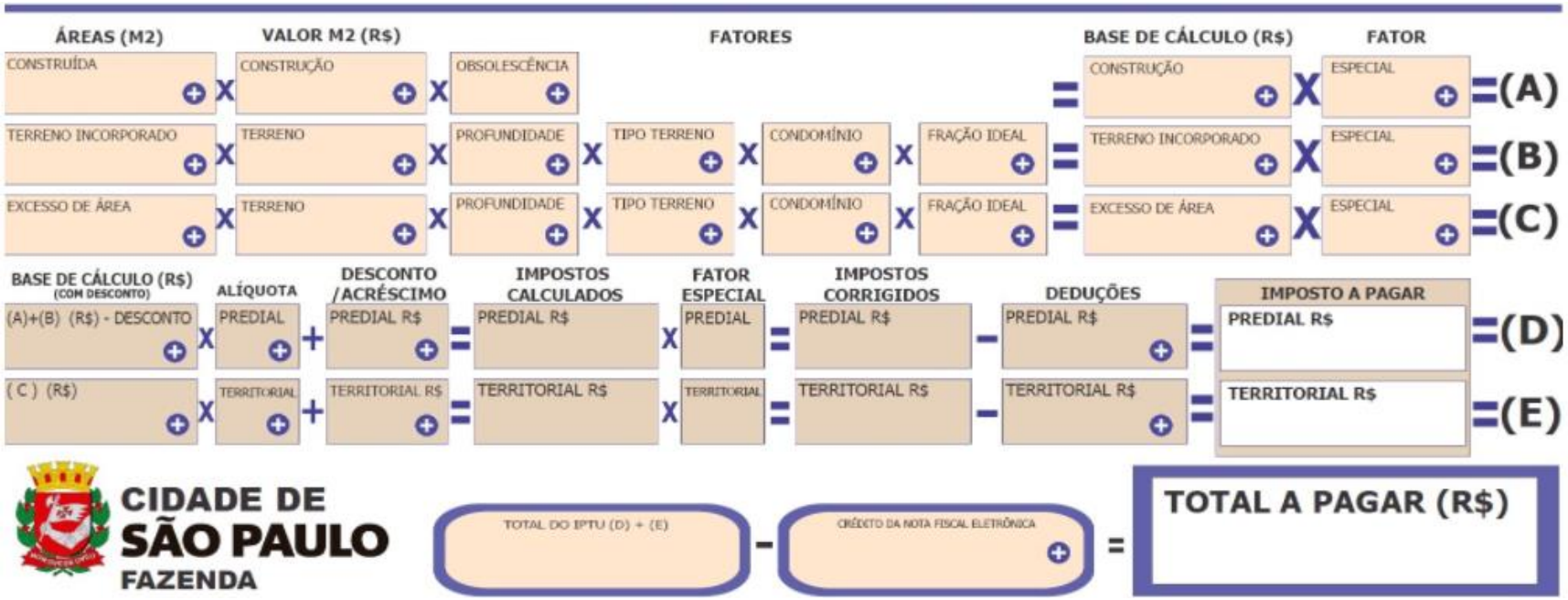
Ref: https://en.wikipedia.org/wiki/DBSCAN

Example

“Given a set of attributes from a property (address, sqr. ft., construction year), is it possible to group them?"

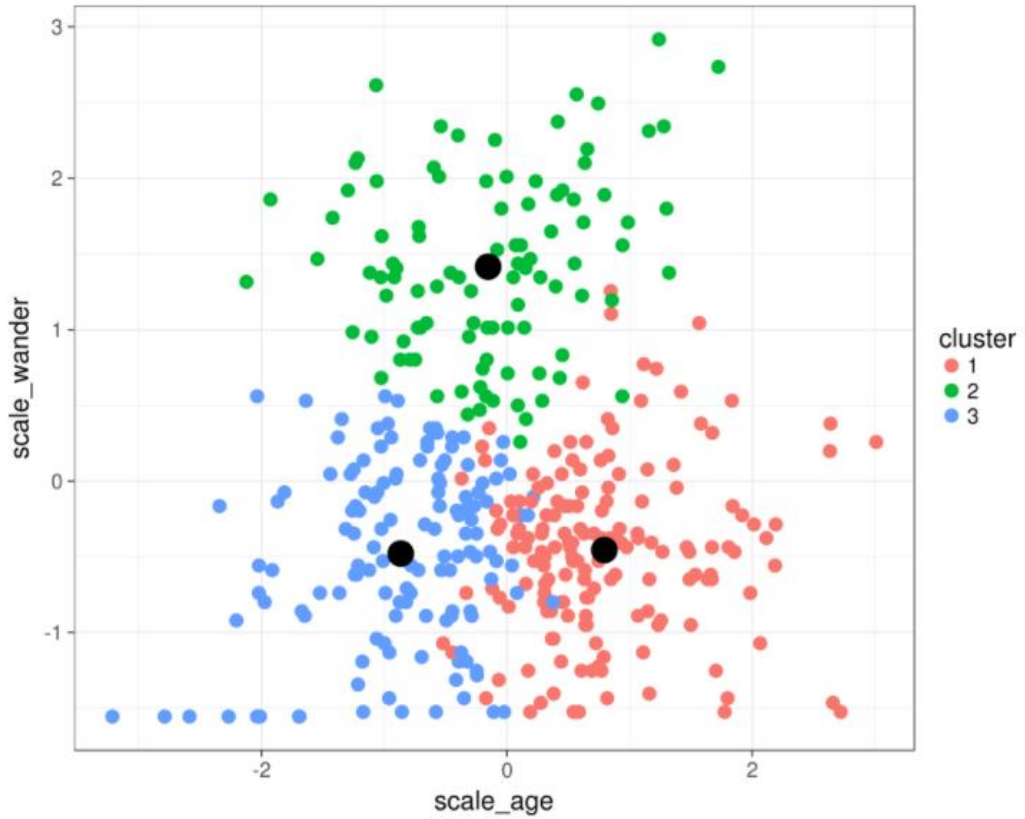
http://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx

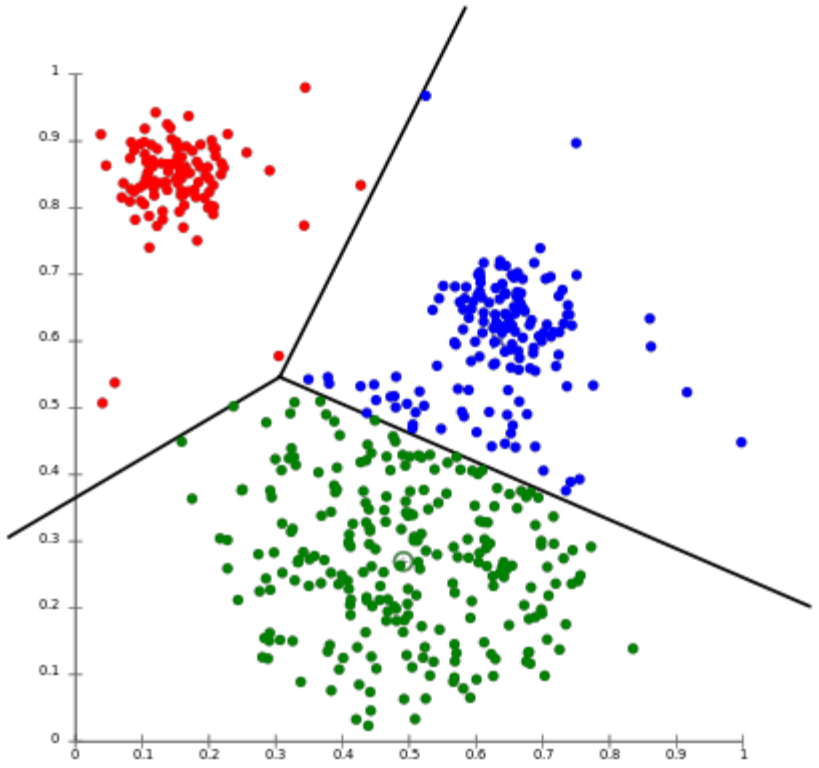
Property Tax formula: https://web1.sf.prefeitura.sp.gov.br/CartelaIPTU/



K-Means

(https://hpccsystems.com/blog/kmeans)





Ref: https://rpubs.com/cyobero/k-means

Example

“Given a set of attributes from a property (address, sqr. ft., construction year), is it possible to order its outliers?"

http://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx

Property Tax formula: https://web1.sf.prefeitura.sp.gov.br/CartelaIPTU/