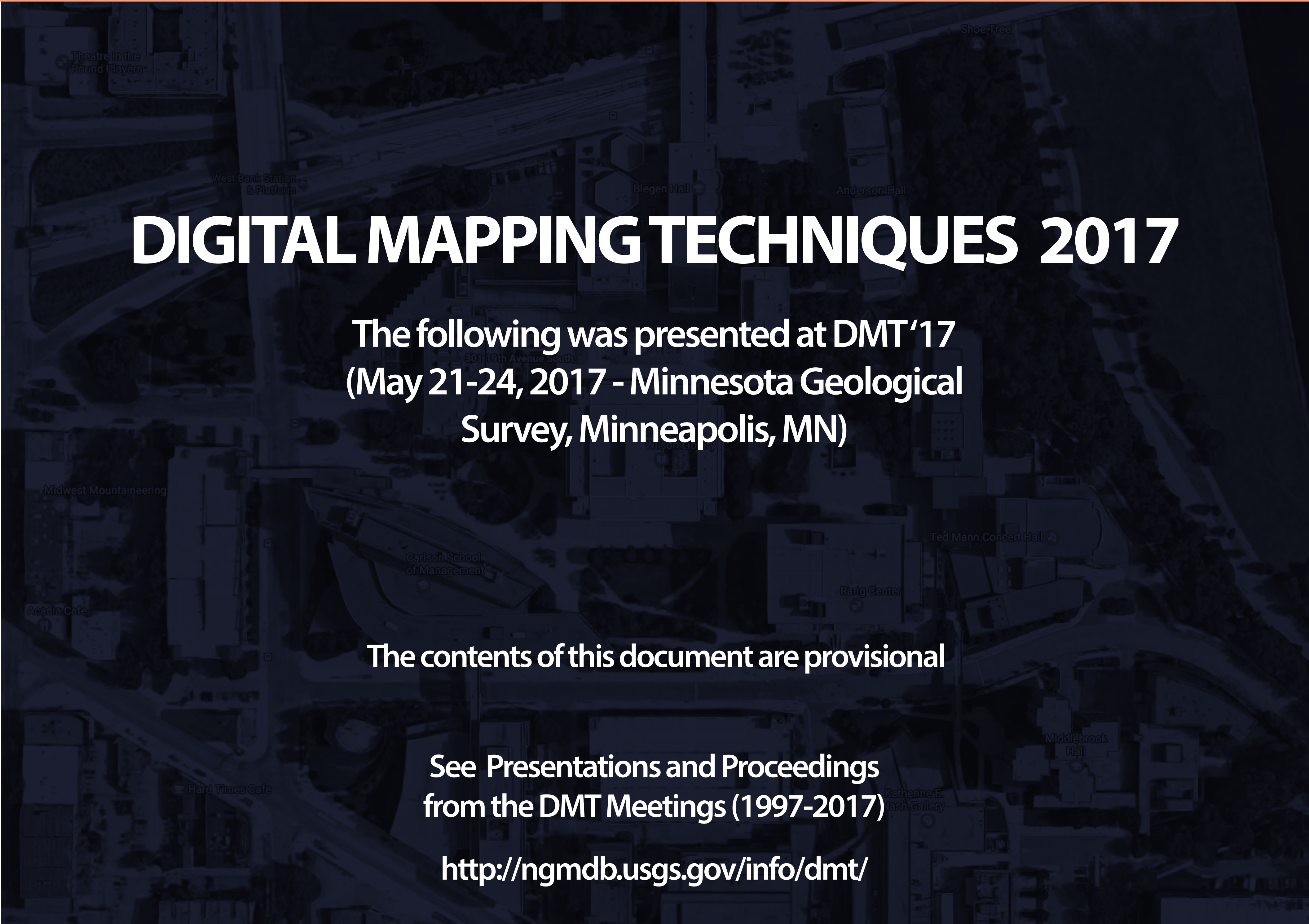
Association of American State Geologists

DMT 2017



DIGITAL MAPPING TECHNIQUES 2017

The following was presented at DMT'17
(May 21-24, 2017 - Minnesota Geological
Survey, Minneapolis, MN)

The contents of this document are provisional

See Presentations and Proceedings
from the DMT Meetings (1997-2017)

http://ngmdb.usgs.gov/info/dmt/

Teaching a Computer Geology: Automated Lithostratigraphic Classification Using Machine Learning Algorithms

By: Seth Willis Bassett, GISP
3000 Commonwealth Blvd,
Suite 1
Tallahassee, FL 32303
Telephone: (850)-617-0300
Email: Seth.Bassett@dep.state.fl.us

Presentation Summary

"Machine Learning" (ML) is an interdisciplinary field developed within the computer sciences **(Slide 2).** Although machine learning algorithms (MLAs) have been around for decades (formerly being called "Artificial Intelligence" or similar), there has been a renewed interest in machine learning in the past few years across a wide variety of domains. This interest has been driven by the development of new, robust algorithms and techniques capable of solving problems that were formally considered intractable in the computer sciences: vision and pattern recognition; time-series and high-dimensional data predictions; and non-linear or complex problems that are resistant to more traditional modeling and predictive approaches. Today, machine learning encompasses a pragmatic set of techniques and approaches to computational problems, and there is a distinct focus among practitioners with regards to developing practical algorithms that 'work' in real-time or near-real time over the development of 'theoretically pure' mathematical theory.

MLAs are new enough that they are only just now beginning to be applied to domain of geology, and ML approaches to modeling geologic epiphenomenon have only recently become a trending topic in the academic literature (**Slide 3).** Nevertheless, since 2014 there have been several outstanding and notable works in applying ML to geology and geologic mapping. ML and MLAs have been used to predict lithostratigraphic classifications from wireline geophysical logging data, automatically classify surficial geology from remotely sensed satellite data, and to effectively predict geohazard susceptibility.

This presentation reports back on a "proof of concept" study conducted to evaluate the effectiveness of ML techniques when used to predict stratigraphy and lithology from 144 geophysical logging files supplied by the St. John's Water Management District (SJRWMD). The proof of concept evaluated effectiveness of 3 machine learning algorithms and 3 "post-processing strategies" at 3 different stratigraphic classification tasks, for a total of 27 test cases. The author concludes that MLAs can be quickly, easily, and effectively applied to questions in geology and have substantial predictive power and are able to classify lithostratigraphy from geophysical logging data just as well as a human being for the simplest of tasks.

Summary of Slide Content

- Slide 4 discusses technical considerations for implementing MLAs by contrasting the two most popular languages that have active ML packages under development. Python is by far the easier language to learn and to use, but encapsulating content in a web-based format and parallelizing algorithms can be difficult for the beginner. R has the disadvantage that it is much more difficult to learn and work with than Python, but the MLAs available to the user tend to be more 'cutting edge' than those available in Python. R also has the advantage of being easy to implement in parallel by using the doMC package, while the Shiny package makes wrapping R analyses in dynamic web content relatively trivial. Whether these advantages outweigh the horror show that is R syntax is left to the user to decide.

- Slide 5 illustrates the four types of tasks MLAs can perform: regression, binary classifications, multiclass classifications, and ranking tasks.
- Slides 6-9 detail the three types MLAs used in this study:
    - Slide 6: Decision Trees are an algorithm that takes a 'divide and conquer' approach to predictive modeling. The decision tree algorithm uses recursive partitioning and binary IF-THEN logic to construct a logical decision tree that is used to make new predictions.
    - Slide 7: An example decision tree for classifying limestone using gamma ray geophysical logs. Reading the graph, we begin at the root (topmost) node and follow the decision tree down.
        - Each node shows the decision criteria at that node (depth < 174 ft for the root node, indicated by the orange overlay text and arrow) above the node box. Inside the node box, we see the decision class ('Not Limestone'; red overlay text and arrow), along with the class split at that node (62% 'Not Limestone, 38% Limestone; green overlay text and arrow) for all data points that fall within the node. The last line in each box represents the total percentage of the entire dataset that falls within that node (purple overlay text and arrow).
        - To reconstruct how this decision tree model determines a data point's classification, we follow each tree down to its terminal node, moving left if the node's decision criteria is true, and right if it is false.
        - For example: if we have a point in a gamma ray geophysical log that recorded a value of 10 CPS at a depth of 248 feet., we can use the decision tree to determine how the model will classify the point as either "Limestone" or "Not Limestone."
        - In the above case, the top decision node would be false ("No") the depth of our example point is *greater* than 174 feet, so we move along the right hand branch and down one level to the next node. At this node, our decision criteria – 'val >= 28' - is again false, so we again move along the right hand branch and down a level, arriving at the terminal node in the bottom right hand corner of the decision tree.
        - The bottom-most row of terminal nodes are the model's predictive classes for our data point ("Limestone"). In the above example, the class split at the terminal node indicates that 4% of our training data at this node were classified as "Not Limestone" and that 96% were classified as "Limestone." This node also indicates that 31% of all of our training data ultimately fell within this terminal decision node.
        - An animated example of the decision tree classification process can be seen by clicking on the decision tree model, which will follow a link to https://youtu.be/XmnenS9d3cA.
    - Slide 8: Random Forest (RF) is an extension of decision tree models first described by Brennan (2001). The Random Forest algorithm builds a large number of trees (N) by randomly subsampling the input training data. All of these decision trees are then used to classify each data point provided to the model to predict that data point's class, with

the majority prediction of all of the decision trees 'winning.' This technique reduces the sensitivity to variance in training data and provides for more robust predictions when applied to data previously unseen by the model.

- Support Vector Machines (SVMs) are a binary classification method that can be extended to multiple classes. These models use a hyperplane in an n-dimensional mathematical space to cluster input data into groups. SVMs have the advantage of working well with small training sets and only have two user-defined parameters, *gamma* and *cost.* Cross validation can be used to empirically determine the optimum setting for these parameters and reduce the subjective decisions required of the user.
- Slide 10: A humorous attempt to illustrate that while machine learning might seem complicated at first glance, there is a very well defined and concrete workflow that is comprehensible even to the most unsophisticated user.
- Slide 11: Training Workflow for Training ML models.
    - Step 1: Subset input data from the total bucket of raw data
    - Step 2: Label input data with known class labels.
    - Step 3: Split the labeled data into two groups by random assignment: the training and test sets. The training set is comprised of 60% of the labeled input data, and will be used to train the algorithm. The test set is comprised of 40% of the labeled data and will be held in reserve in order to assess model performance.
        - Classes should be balanced so that they have a 50/50 rate of occurance in both the training set and the test set. **Note that this was not done for this project.**
    - Step 4: Train model using the training set.
    - Step 5: Use the trained model to predict the classes of the points within the test data
    - Step 6: Use the test set to compare the model predictions against the known class labels in order to assess model performance and predictive accuracy.
- Slide 12 shows the labeled dataset used as training/test input in this study. This data consisted of 22 gamma radiation geophysical logs labeled with matching stratigraphic formation picks.
- Slides 14-15 show examples of the labeled data used in this study. In slide 15, yellows and oranges represent surficial sediments, greens represent the clay-bearing Hawthorn group, and blues represent consolidated rock in the form of limestone
- Slide 16 shows the 3 different classification tasks developed for this project, as well as the formula notation for the models. Task complexity is an important factor in model performance, with models performing better at simpler tasks than more complex tasks.
    - The pick2 task is the simplest task and is the primary focus of this presentation. This task asks the model to classify each input data point as "limestone" or "not limestone."
    - The pick3 task is slightly more complex, asking each model to predict whether a data point represents surficial sediments, the clay-bearing hawthorn group, or limestone rock.

    - The pick6 task is the most difficult task developed, and asks the models to predict the formal lithostratigraphic formation of each data point.
    - The model formula gives that the dependent (predictive) variable is the pick classes from each of the three tasks. The independent variables used to make this prediction are the gamma radiation intensity in Counts per Second (CPS) of each geophysical log data point, the elevation of each data point, the depth from land surface of each data point, and the projected X and Y location of each data point.
- Slide 19 shows the accuracy and kappa statistics measured against the test data for each of the three types of models across each of the three predictive tasks. As expected, tasks with fewer predictive classes were performed more accurately than tasks with more predictive classes. Significantly, the test accuracy of each of the models was above 94%, suggesting that overmodeling might be a problem.
- Slide 21 shows the workflow for applying trained models to novel data in order to generate predictions.
    - Step 1: Query all data
    - Step 2: Predict against all of the data using the trained models
    - Step 3: Post-process predicted classes in order to remove or mitigate problems with thin-section units.
    - Step 4: Take the maximum elevation of each predicted class as the "top" of that class
    - Step 5: Compare with known values for the top of the Floridan Aquifer System (FAS) and the top of the Intermediate Confining Unit (ICU). In this geographic area, the top of the FAS is equivalent to the top of limestone, and the top of the ICU is equivalent to the top of the Hawthorn group.
    - Step 6: Assess model accuracy
- Slide 22 shows the geographic distribution of the geophysical logging dataset that was used to generate predictions using the three, trained machine learning algorithms.
    - This dataset consisted of 144 gamma logs that were not used during the training/test process
    - The geophysical logs in this dataset had been assess by a SJRWMD geologist, who had assigned a 'Top of FAS' and 'Top of ICU' value by interpreting the gamma curve for every log.
    - Note that this geologist did not look at physical samples to determine these values; rather, the geologist relied only upon the geophysical logging data.
    - The MLAs outlined above were used to generate predictive classifications for every data point in every well within this dataset.
- Slide 23 illustrates the problem with 'thin-section classifications' and some approaches to correcting or mitigating their influence. Three machine learning algorithms engaging in 3 predictive tasks with three postprocessing strategies gives a total of 27 test cases to assess for accuracy.
    - The left-most gamma curve in the chart shows predictive classes generated by the random forest model.

  - The two thin-sections of limestone predictions between –200 and –250 ft MSL are problematic.
  - While thin sections of limestone are known to occur within the Hawthorn group, these predictions place the "top" of the limestone unit at –200 ft rather than at –500 ft when the "naïve maximum elevation' for the limestone class is taken from this prediction.
  - Two techniques were used to postprocess the model predictions and mitigate the influence of thin sections: changepoint clustering and a majority filter.
    - The changepoint algorithm clusters a continuous curve into groups, based on breakpoints detected within the curve. The second graph from the left labeled "Changepoint Clusters" shows the 6 clusters the changepoint algorithm found in this particular gamma log. Each cluster is then assigned to the same predictive class as the majority of the points that within the cluster – this is shown in the third graph from the left, labeled "RF + Changepoints" graph.
    - The majority filter takes each 5 ft interval of the gamma log and assigns a predictive class to the interval based on the majority predictive class within the interval. This method essentially downsamples the model predictions to a vertical resolution of 5 feet ft.
- Slide 24 shows the formula for calculating predictive error.
- Slide 25 & 26 show model performance and accuracy metrics.
  - The best performing model across all predictive tasks was the random forest algorithm.
  - The best overall performance was by the random forest algorithm in the pick2 classification task, using the changepoint postprocessing strategy. This model managed to pick the top of limestone with a median difference of 2.5 feet from what the human geologist picked. Additionally, the mean error was only 8.69 ft, and 75% of all model predictions for the top of limestone were within 8 ft of the Top of FAS pick made by the SJRWMD geologist.
- Slides 27-36 show the Random Forest + Changepoint predictions for the pick2 class, plotted against the actual Top of FAS values picked by the SJRWMD geologist.
  - The color of the gamma curve itself shows the model's predicted class for each data point within the curve
  - The dashed blue line illustrates the human geologist's pick for top of FAS for that well.
  - Agreement between the geologist and the machine learning algorithms is shown when the color break in the gamma curve occurs at or near the dashed blue line.
- Slides 37-39 show how applying machine learning to geophysical logs helps to increase the accuracy of geologic mapping. The accuracy of mapping stratigraphic structure is highly dependent upon the spatial density of the geologic data available for a region, with more data points producing a higher resolution map.
  - The primary structural feature in this study area is the Jacksonville Basin, a large, bowl shaped depression in the top of the bedrock limestone

  - Slide 37 shows an exact interpolation (Radial Basis Function/Spline) of the top of limestone for the Jacksonville basin using only the formation picks generated by the FGS STATEMAP team.
  - Slide 38 shows the same interpolation method used with both the STATEMAP geologic picks and the top of limestone picks made by the Random Forest + Changepoint method
  - Slide 39 compares these two interpolated surfaces side-by-side for convenience. It is evident that the inclusion of more data has produced a much higher resolution map of the structure and shape of the Jacksonville Basin.
- Slide 40 details some conclusions and lessons learned by the author during the course of the project

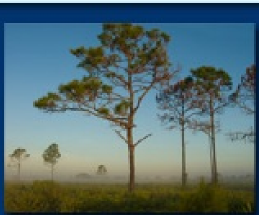
Florida Department of Environmental Protection
Florida Geological Survey

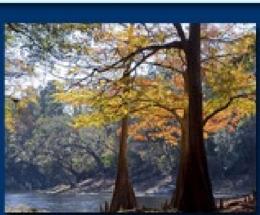
Teaching a Computer Geology:

Automated Lithostratigraphic Classification Using Machine Learning Algorithms

Seth Willis Bassett, GISP

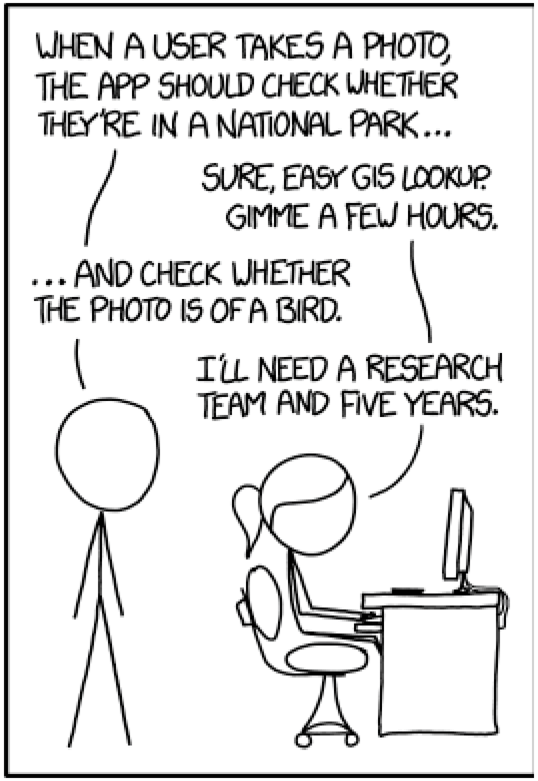
May 23, 2017





What is Machine Learning (ML)?

- Machine learning "gives computers the ability to learn without being explicitly programmed." (Arthur Samuel, 1959)
- A set of robust predictive modeling techniques that have been developed and popularized within the field of computer science
- Applications include 'hard' CS problems such as image recognition and feature detection
- Focus of machine learning is a **pragmatic**, multi-disciplinary approach to expert systems and predictive modeling
- Spatial autocorrelation improves the predictive power of many ML algorithms
- *Deep learning* is a further extension of ML that uses more complex modeling techniques


WHEN A USER TAKES A PHOTO, THE APP SHOULD CHECK WHETHER THEY'RE IN A NATIONAL PARK...
SURE, EASY GIS LOOKUP. GIMME A FEW HOURS.
...AND CHECK WHETHER THE PHOTO IS OF A BIRD.
I'LL NEED A RESEARCH TEAM AND FIVE YEARS.

IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

Prior Work in Geology

- **ML Classification of Lithostratigraphy Using Wireline Geophysical Logging Data**
  - Bhattacharya, Shuvajit, Timothy R. Carr, and Mahesh Pal. **2016**. "Comparison of Supervised and Unsupervised Approaches for Mudstone Lithofacies Classification: Case Studies from the Bakken and Mahantango-Marcellus Shale, {USA}." *Journal of Natural Gas Science and Engineering* 33: 1119–33. doi:http://dx.doi.org/10.1016/j.jngse.2016.04.055.
  - Sebtosheikh, Mohammad Ali, and Ali Salehi. **2015**. "Lithology Prediction by Support Vector Classifiers Using Inverted Seismic Attributes Data and Petrophysical Logs as a New Approach and Investigation of Training Data Set Size Effect on Its Performance in a Heterogeneous Carbonate Reservoir." *Journal of Petroleum Science and Engineering* 134: 143–49. doi:http://dx.doi.org/10.1016/j.petrol.2015.08.001.
  - Silversides, Katherine, Arman Melkumyan, Derek Wyman, and Peter Hatherly. **2015**. "Automated Recognition of Stratigraphic Marker Shales from Geophysical Logs in Iron Ore Deposits." *Computers & Geosciences* 77: 118–25. doi:doi:10.1016/j.cageo.2015.02.002.
  - Salehi, Seyyed Mohsen, and Bizhan Honarvar. **2014**. "Automatic Identification of Formation Lithology from Well Log Data: A Machine Learning Approach." *Journal of Petroleum Science Research* 3 (2): 73–82.
  - Ferraretti, Denis, Giacomo Gamberoni, and Evelina Lamma. **2012**. "Unsupervised and Supervised Learning in Cascade for Petroleum Geology." *Expert Systems with Applications* 39 (10): 9504–14. doi:http://dx.doi.org/10.1016/j.eswa.2012.02.104.
  - Gifford, Christopher M., and Arvin Agah. **2010**. "Collaborative Multi-Agent Rock Facies Classification from Wireline Well Log Data." *Engineering Applications of Artificial Intelligence* 23 (7): 1158–72. doi:10.1016/j.engappai.2010.02.004.
  - Goncalves, Carlos A. **1998**. "Lithologic Interpretation of Downhole Logging Data from the Cote D'Ivoire Tranform Margin: A Statistical Approach." In *Proceedings of the Ocean Drilling Program, Scientific Results*, 159:157–70. College Station, TX.
- **ML Classification of Surficial Geologic Maps Using Remotely Sensed Data**
  - Harvey, A. S., and G. Fotopoulos. 2016. "Geological Mapping Using Machine Learning Algorithms." *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLI-B8: 423–430. doi:10.5194/isprs-archives-XLI-B8-423-2016.
  - Harris, J. R., and E. C. Grunsky. 2015. "Predictive Lithological Mapping of Canada's North Using Random Forest Classification Applied to Geophysical and Geochemical Data." *Computers & Geosciences* 80: 9–25. doi:http://dx.doi.org/10.1016/j.cageo.2015.03.013.
  - Cracknell, Matthew J., and Anya M. Reading. 2014. "Geological Mapping Using Remote Sensing Data: A Comparison of Five Machine Learning Algorithms, Their Response to Variations in the Spatial Distribution of Training Data and the Use of Explicit Spatial Information." *Computers & Geosciences* 63: 22–33. doi:10.1016/j.cageo.2013.10.008.
  - Waske, Bjorn, Benediktsson, Jon Atli, Arnason, Kolbeinn, and Sveinsson, Johannes. 2009. "Mapping of Hyperspectral AVIRIS Data Using Machine-Learning Algorithms." *Candian Journal of Remote Sensing* 35 (Suppl. 1): S106–16.
- **ML & Geohazard Susceptibility**
  - Goetz, J. N., A. Brenning, H. Petschko, and P. Leopold. 2015. "Evaluating Machine Learning and Statistical Prediction Techniques for Landslide Susceptibility Modeling." *Computers & Geosciences* 81: 1–11. doi:http://dx.doi.org/10.1016/j.cageo.2015.04.007.

Technical Considerations

Python

Packages

- Scikit-learn
- Scikit-image
- Pandas
- Matplotlib

Advantages

- Easy to learn, interpret, use
- Import antigravity! (Python does anything)
- Direct ArcGIS integration via ArcPy

Disadvantages

- ML algorithms are less 'cutting edge' than in R
- Dynamic web dashboards require substantial amount of learning and effort
- Parallel implementation is difficult

R

Packages

- caret
- e1071
- randomForest
- rpart
- [dozens more]

Advantages

- Cutting edge ML algorithms
- Easy web dashboards using shiny-server
- DataTables support makes R extremely fast with large datasets
- Easy parallel implementation in caret and doMC

Disadvantages

- **It's R.**

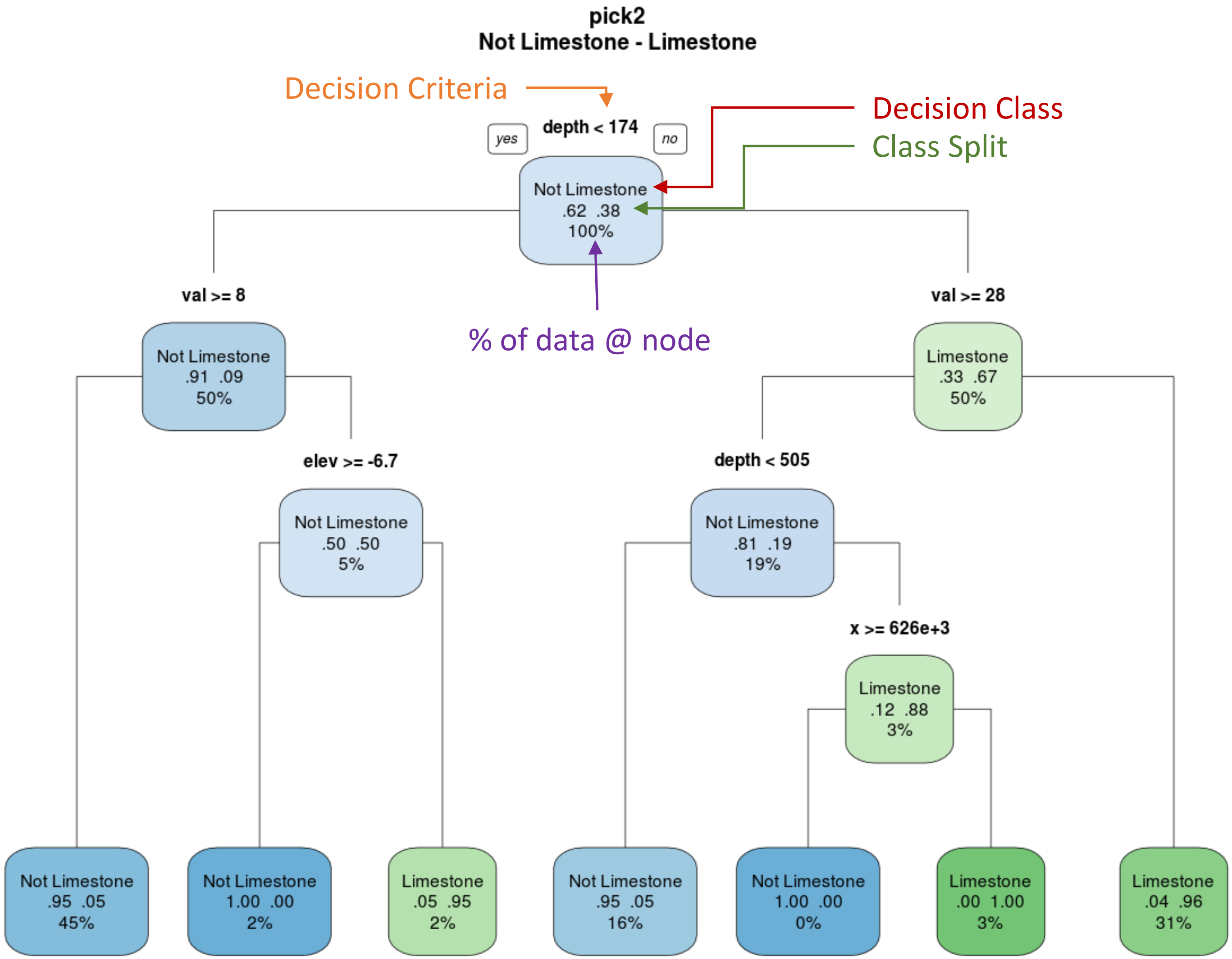
Machine Learning Capabilities

Regression	• Predict real values
Binary Classification	• Predict membership in 1 of 2 categories
Multiclass Classification	• Predict membership in 1 of 3 or more categories
Ranking	• Order objects according to relevance

Model Selection: Decision Trees

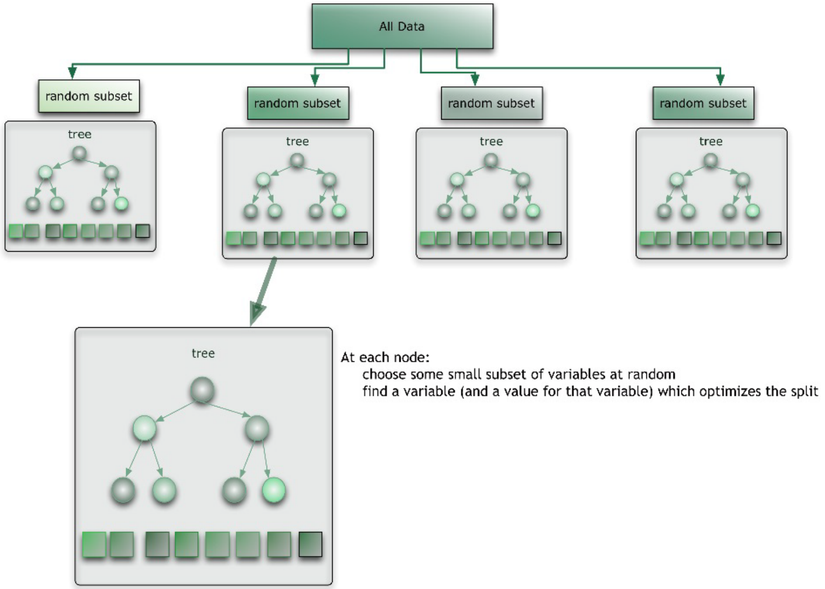
- Classic ‘divide and conquer’ approach to machine learning
- Uses recursive partitioning and binary IF/THEN logic to construct a logical flow chart (a decision tree) that is then used to make predictions against new data
- Decision trees have the advantage that the output is extremely simple to interpret and understand

Model Selection: Decision Trees



Model Selection: Random Forest

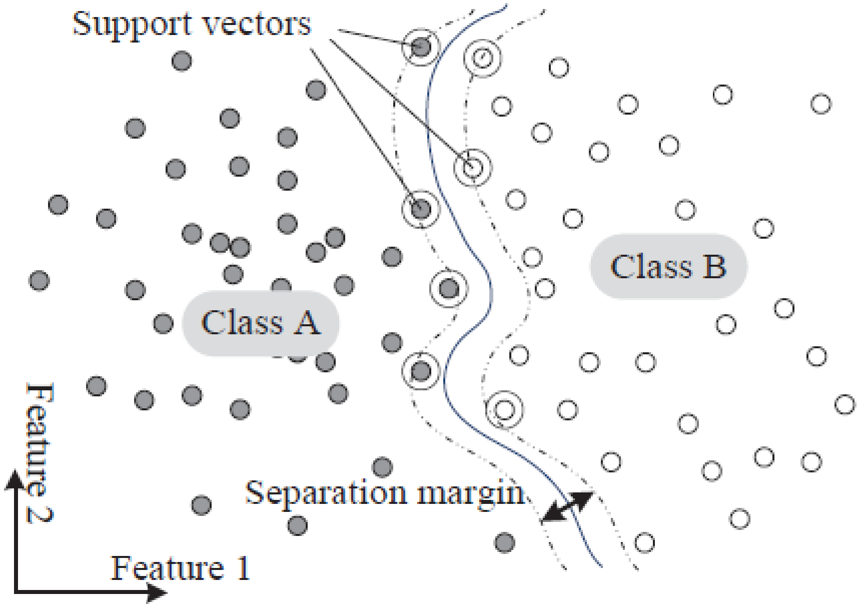
- Extension of Decision Trees by Brennan (2001)
- Uses a technique known as *bagging* to reduce variance in input data
  - Training data is randomly sampled (with replacement) and a decision tree is built
  - Process is repeated **N** times to generate **N decision trees**
  - **N** selected by user as a model parameter but usually 500
- **All N-trees** are used to classify a new data point, with the majority vote from all decision trees taken as the final value



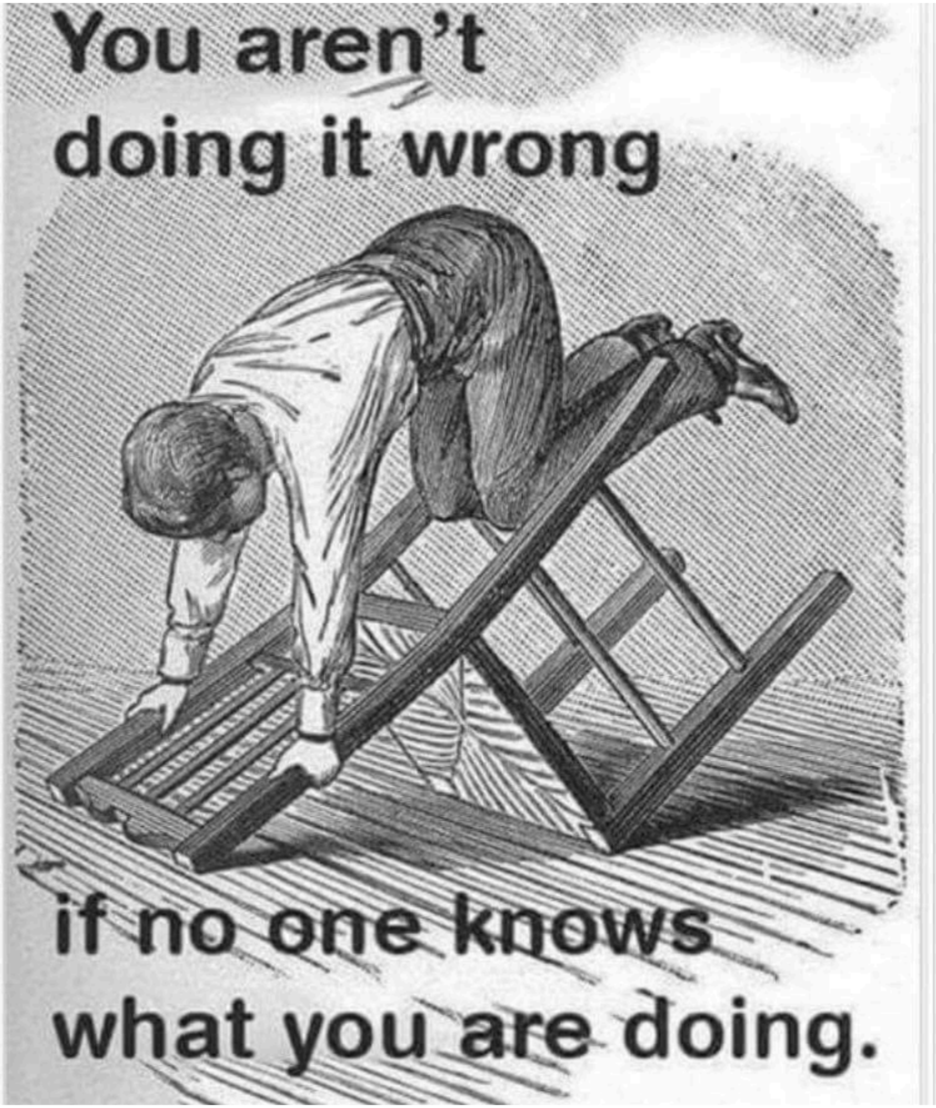
At each node:
choose some small subset of variables at random
find a variable (and a value for that variable) which optimizes the split

Model Selection: Support Vector Machines

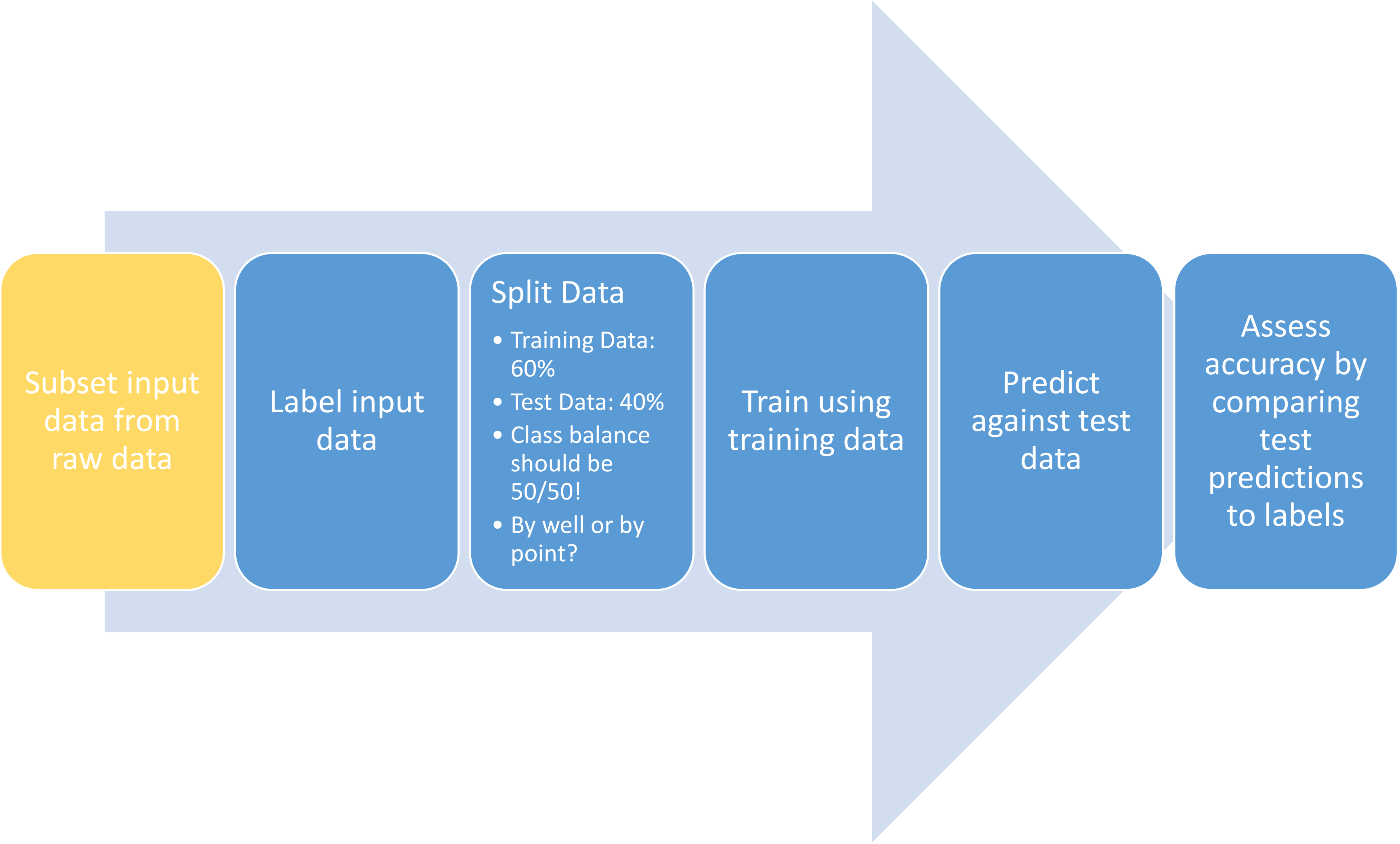
- Binary classification method that can be extended to multiple classes
- Uses a hyperplane to cluster data groups using a *n*-dimensional vector space
  - ‘Kernel trick’ allows SVMs to cluster groups that are not linearly separable
- Works well with small training sets
- User selected parameters (*gamma* and *cost*) can be searched empirically using cross validation to find the best values



Life Rule #1



Training Workflow



Subset input data from raw data

Label input data

Split Data

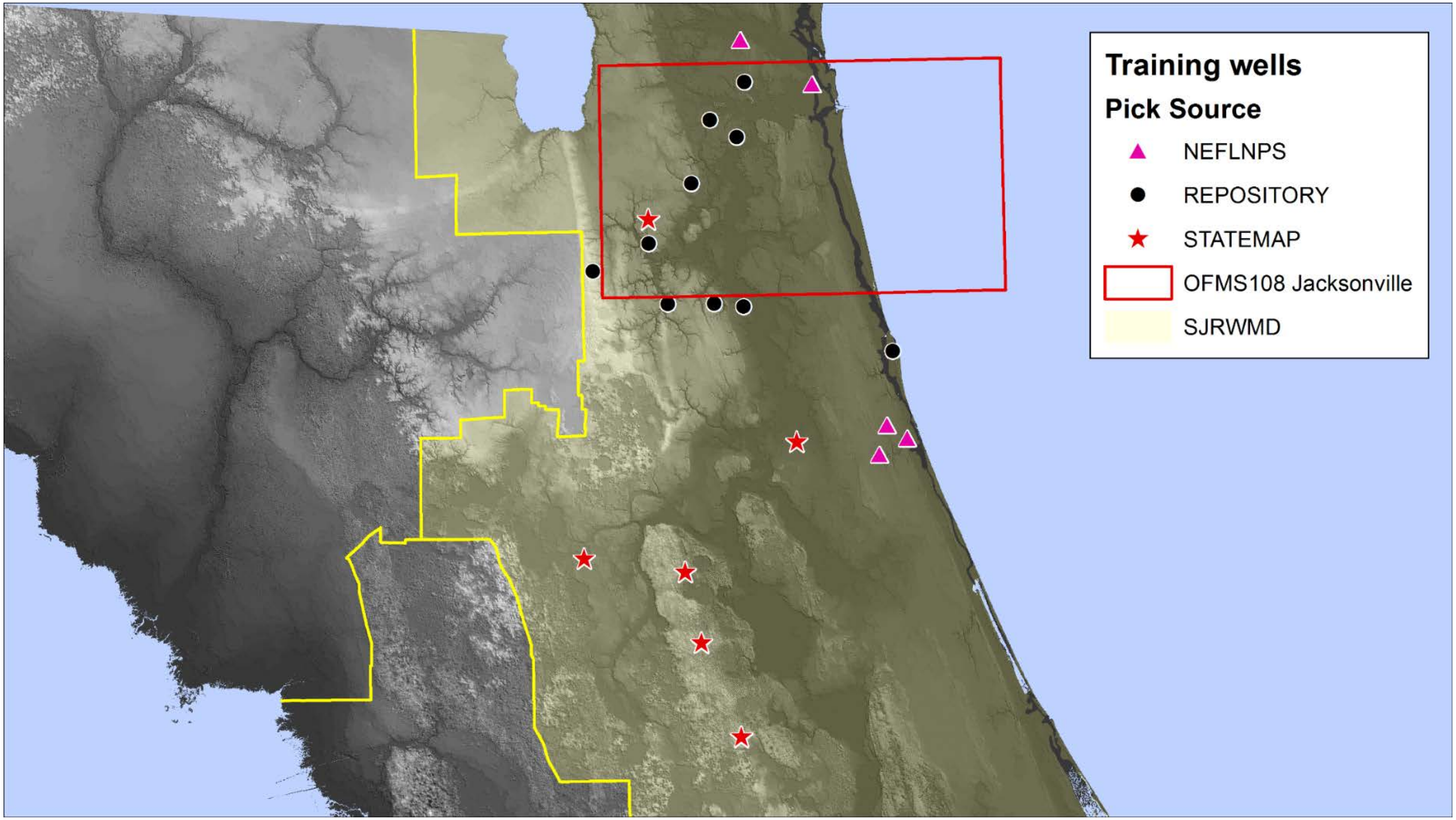
- Training Data: 60%
- Test Data: 40%
- Class balance should be 50/50!
- By well or by point?

Train using training data

Predict against test data

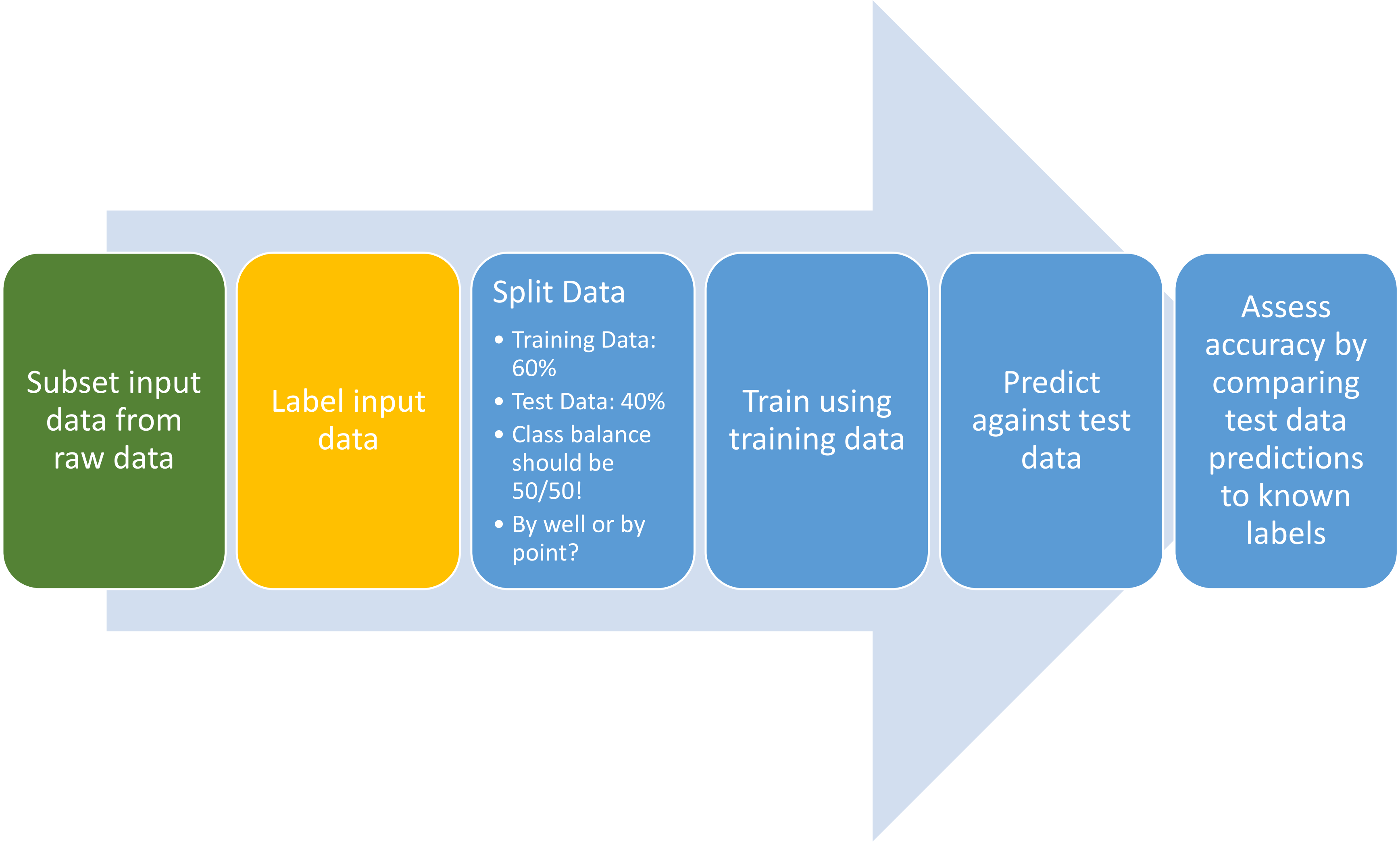
Assess accuracy by comparing test predictions to labels

Example ML Training Task



- 22 log-description combos from across SJRWMD
  - 11 historic descriptions
  - 11 wells drilled and described in the past 3-5 years

Training Workflow



Subset input data from raw data

Label input data

Split Data

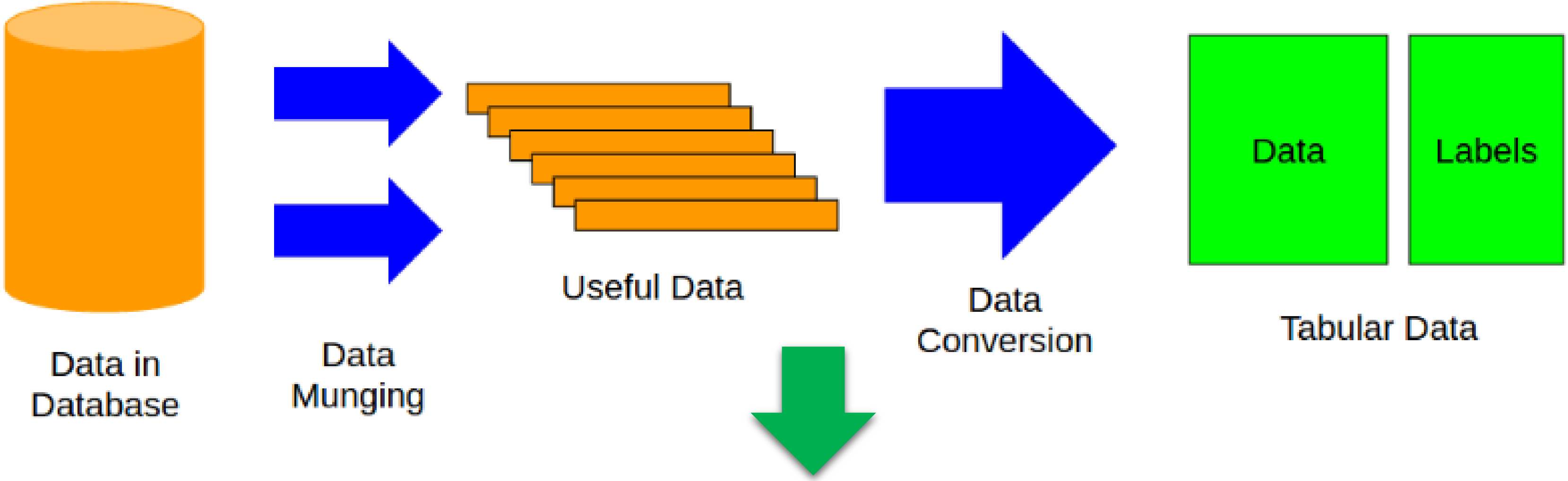
- Training Data: 60%
- Test Data: 40%
- Class balance should be 50/50!
- By well or by point?

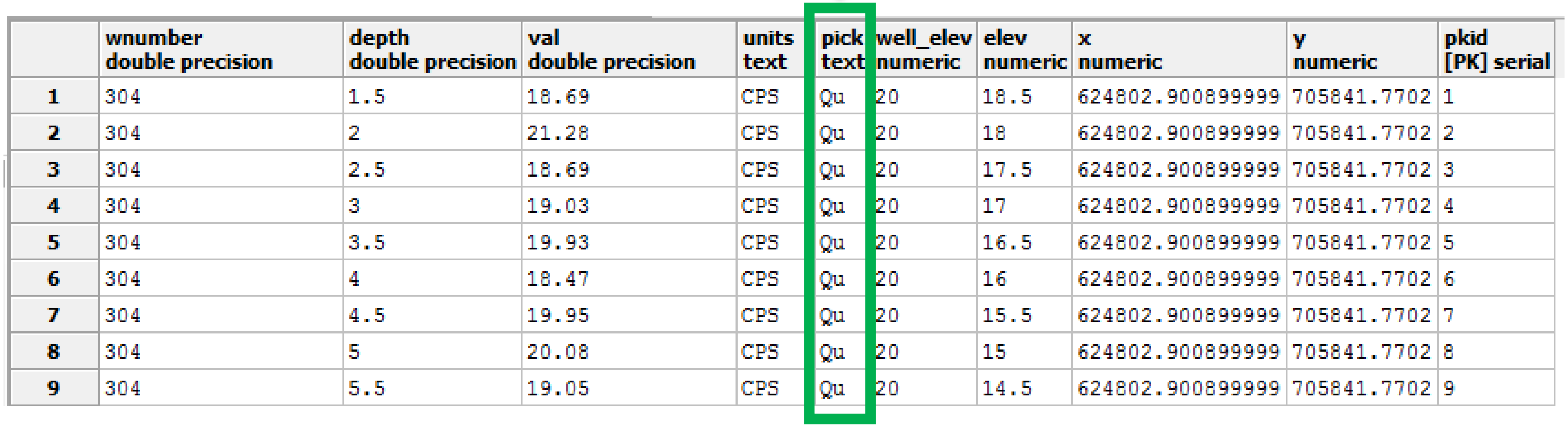
Train using training data

Predict against test data

Assess accuracy by comparing test data predictions to known labels

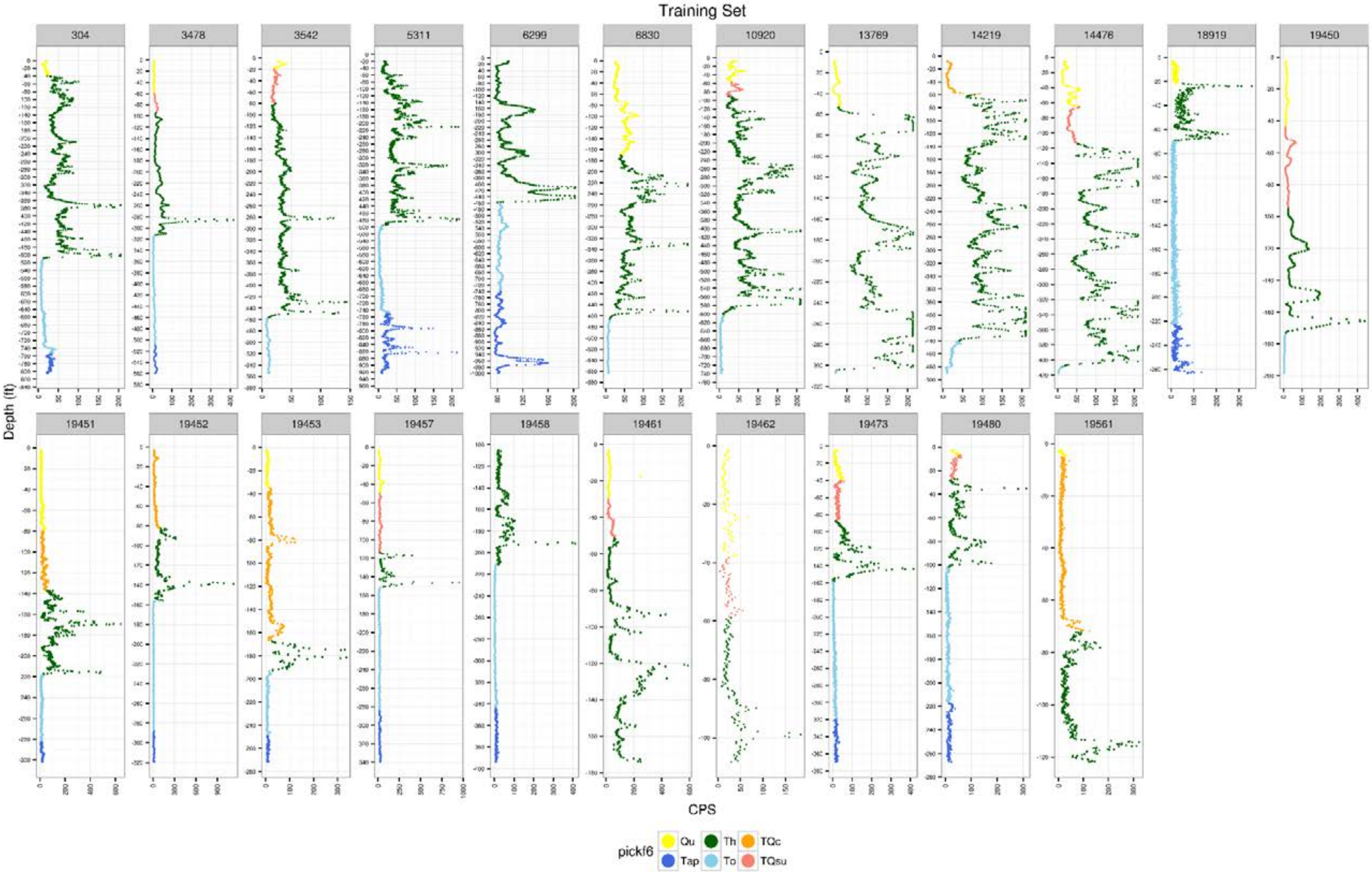
Example Labeled Data





	wnumber double precision	depth double precision	val double precision	units text	pick text	well_elev numeric	elev numeric	x numeric	y numeric	pkid [PK] serial
1	304	1.5	18.69	CPS	Qu	20	18.5	624802.900899999	705841.7702	1
2	304	2	21.28	CPS	Qu	20	18	624802.900899999	705841.7702	2
3	304	2.5	18.69	CPS	Qu	20	17.5	624802.900899999	705841.7702	3
4	304	3	19.03	CPS	Qu	20	17	624802.900899999	705841.7702	4
5	304	3.5	19.93	CPS	Qu	20	16.5	624802.900899999	705841.7702	5
6	304	4	18.47	CPS	Qu	20	16	624802.900899999	705841.7702	6
7	304	4.5	19.95	CPS	Qu	20	15.5	624802.900899999	705841.7702	7
8	304	5	20.08	CPS	Qu	20	15	624802.900899999	705841.7702	8
9	304	5.5	19.05	CPS	Qu	20	14.5	624802.900899999	705841.7702	9

Example Labeled Data



Model Formula: pick ~ CPS + elevation + depth + X + Y

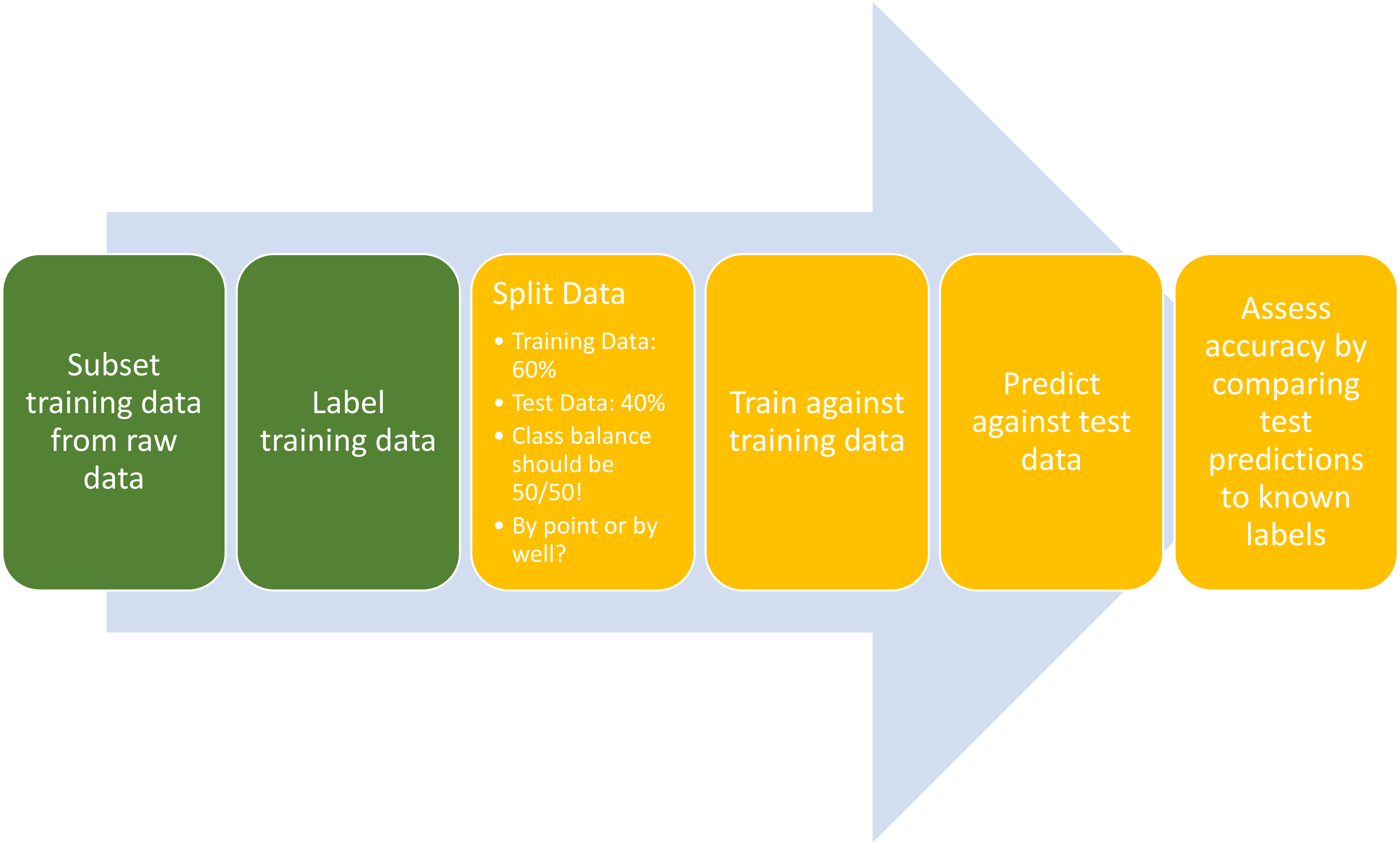
Example Classification Tasks

Predictive Class Relationships

pick2	pick3	pick6
Not Limestone	Sediments	Undifferentiated Quaternary Sediments
		Pliocene/Pleistocene Shelly Sediments
		Cypresshead Formation
	Hawthorn Group	Hawthorn Group
Limestone	Limestone	Ocala Limestone
		Avon Park Formation

pick ~ CPS + elevation + depth + X + Y

Training Workflow



Subset training data from raw data

Label training data

Split Data

- Training Data: 60%
- Test Data: 40%
- Class balance should be 50/50!
- By point or by well?

Train against training data

Predict against test data

Assess accuracy by comparing test predictions to known labels

Dramatic Recreation

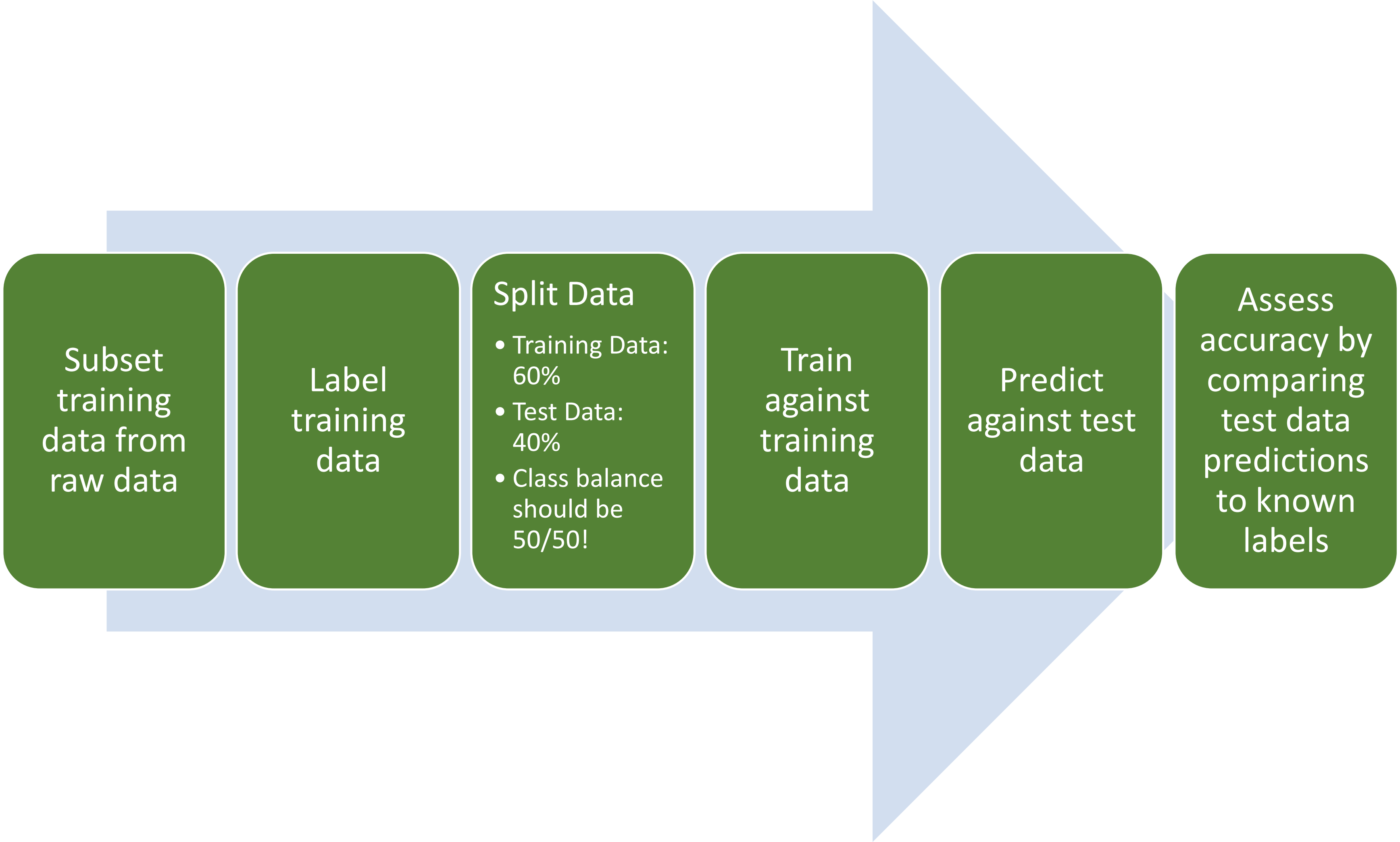
Example Test Accuracy

	Accuracy			Kappa		
	pick2	pick3	pick6	pick2	pick3	pick6
Decision Tree	0.994	0.97	0.9459	0.9871	0.9542	0.9272
Random Forest	0.9993	0.9987	0.9979	0.9986	0.998	0.9971
Support Vector Machine	0.9871	0.97	0.9459	0.9871	0.9542	0.9272

Observed accuracy decreases as task complexity increases

The Kappa statistic (or value) is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance). Kappa takes into account random chance (agreement with a random classifier), which generally means it is less misleading than simply using accuracy as a metric (an Observed Accuracy of 80% is a lot less impressive with an Expected Accuracy of 75% versus an Expected Accuracy of 50%).

Training Workflow Complete!



Subset training data from raw data

Label training data

Split Data

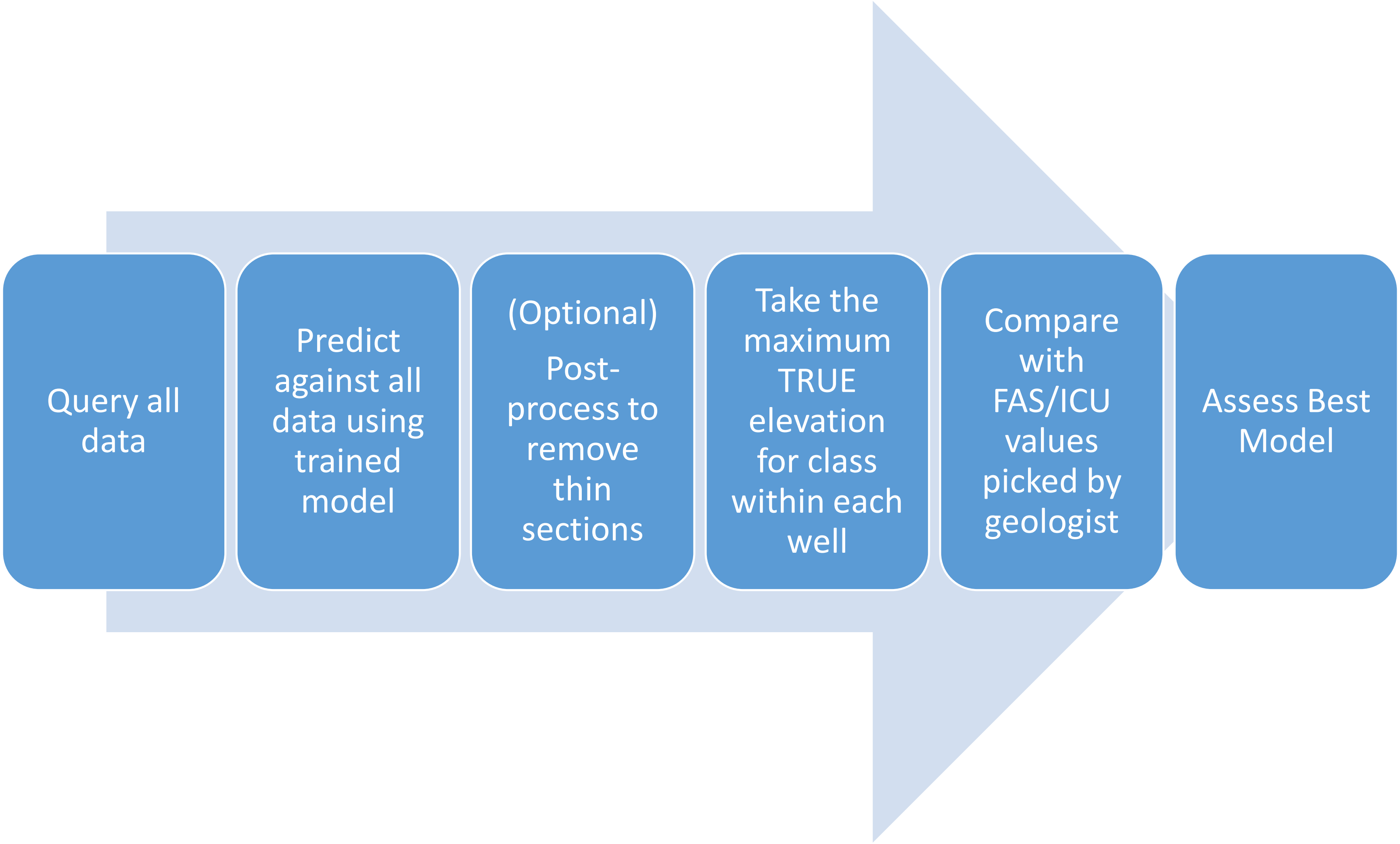
- Training Data: 60%
- Test Data: 40%
- Class balance should be 50/50!

Train against training data

Predict against test data

Assess accuracy by comparing test data predictions to known labels

Predictive Workflow



Query all data

Predict against all data using trained model

(Optional)

Post-process to remove thin sections

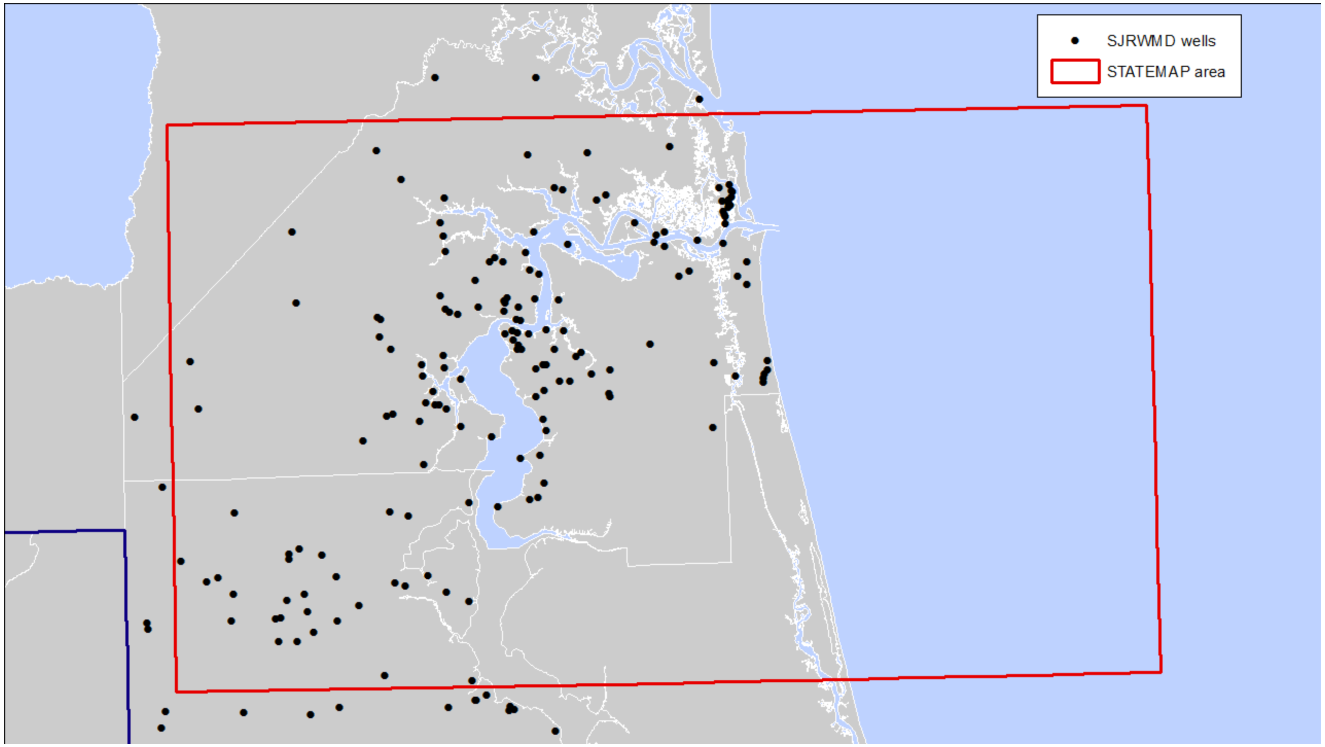
Take the maximum TRUE elevation for class within each well

Compare with FAS/ICU values picked by geologist

Assess Best Model

Example ML Predictive Task

- 144 St. Johns River WMD gamma logs **not used** in the labeled training/test data
  - Each log has a "Top of Floridan Aquifer System (FAS)" value picked by a SJRWMD geologist.
  - Top of FAS value is equivalent to top of limestone in this area.
- Models used to predict the classification of every data point in each gamma log
- Models had not "seen" this data prior to predictive classification



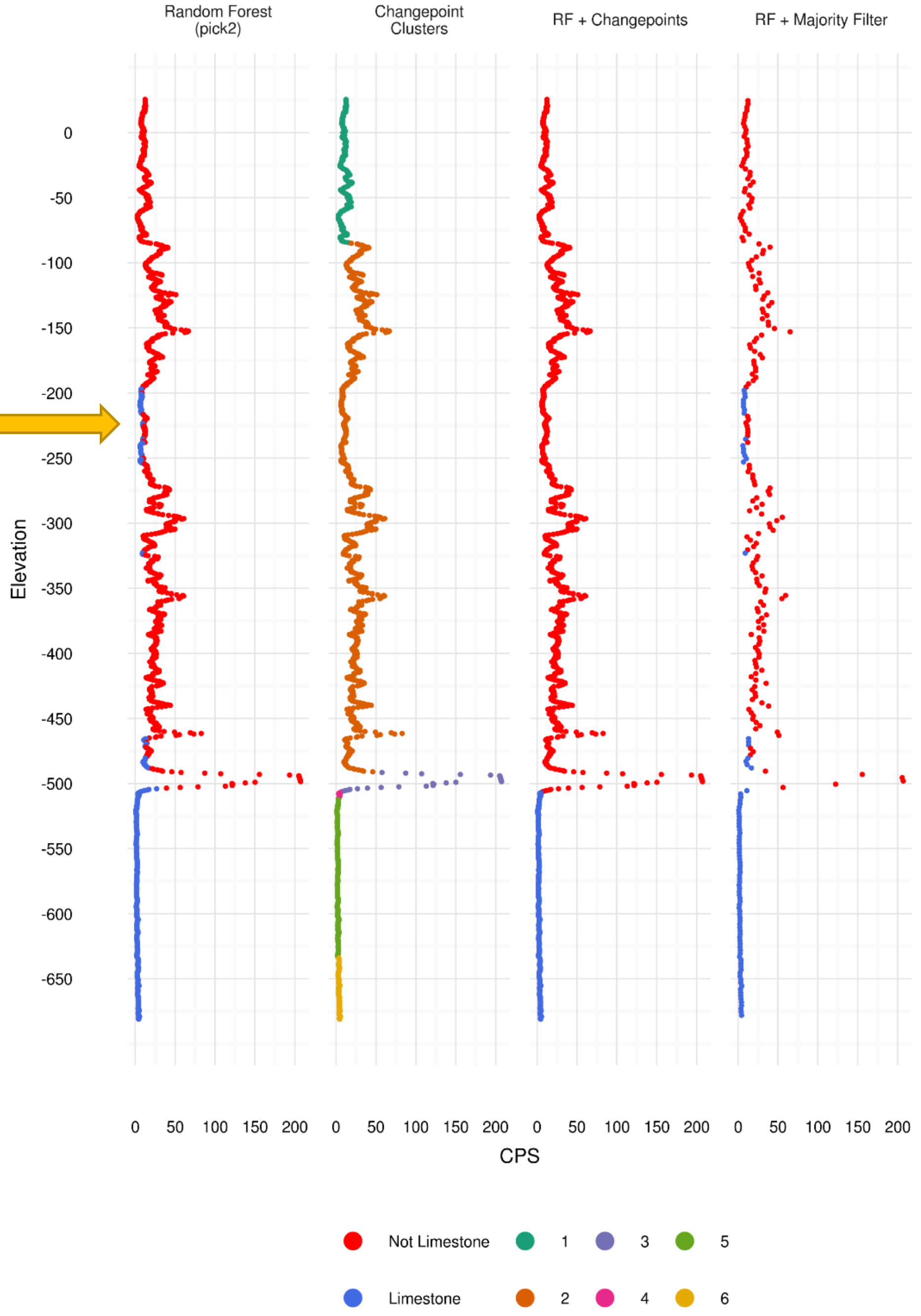
Postprocessing Strategies & Thin Sections

Naive Minimum
Changepoint Clustering
Majority Filtering

(3 Tasks)
x (3 Models)
x (3 PP methods)

= 27 Outcomes to assess

Thin Section Classifications

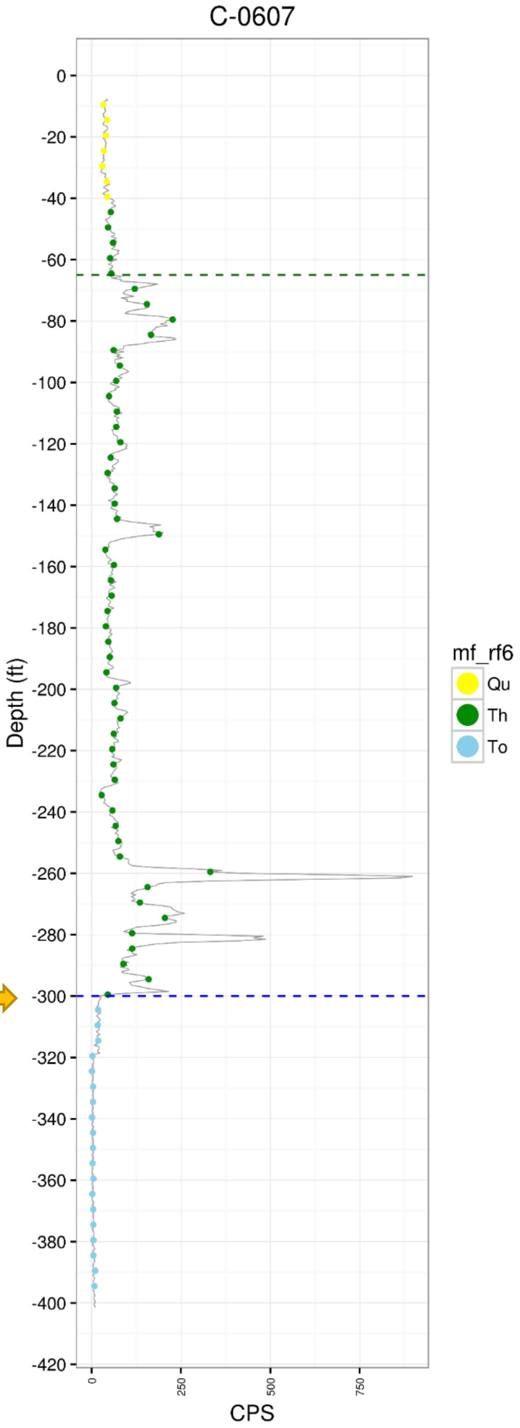


Example: Validation Error

(Absolute Error)$_{\text{WELL}}$ =

| (Top of FAS pick)$_{\text{WELL}}$ – (Maximum elevation of predicted TRUE class)$_{\text{WELL}}$ |

Dotted Blue Line = Geologist's Pick for Top of FAS



Example Validation Accuracy

Model Validation: Absolute Error

	Naive Maximum			Changepoint			Majority Filter		
Decision Tree									
	Median	**Mean**	**Q3**	**Median**	**Mean**	**Q3**	**Median**	**Mean**	**Q3**
Top of Limestone (pick2)	215.50	200.43	319.88	2	41.42	6.125	174.50	160.00	267.00
Top of Hawthorn (pick3)	19.50	23.99	31.88	13.75	31.19	47.88	**20.00**	**23.03**	**30.00**
Top of Hawthorn (pick6)	41.75	43.50	62.00	40.50	43.38	64.50	38.75	38.75	59.75
Random Forest									
	Naive Maximum			**Changepoint**			**Majority Filter**		
	Median	**Mean**	**Q3**	**Median**	**Mean**	**Q3**	**Median**	**Mean**	**Q3**
Top of Limestone (pick2)	6.50	22.08	15.62	**2.50**	**8.69**	**8.00**	8.00	16.95	15.13
Top of Hawthorn (pick3)	21.25	25.26	37.88	14.50	28.19	47.88	**19.00**	**24.01**	**34.00**
Top of Hawthorn (pick6)	24.50	28.30	41.88	25.00	35.74	62.00	**20.75**	**26.27**	**38.38**
Support Vector Machine									
	Naive Maximum			**Changepoint**			**Majority Filter**		
	Median	Mean	Q3	Median	Mean	Q3	Median	Mean	Q3
Top of Limestone (pick2)	11.00	30.23	44.25	**2.50**	**18.67**	**10.00**	11.50	30.34	35.50
Top of Hawthorn (pick3)	21.75	32.09	38.50	22.50	41.94	56.50	20.50	33.84	37.00
Top of Hawthorn (pick6)	23.50	29.64	35.50	23.00	37.99	56.25	**22.75**	**32.16**	**35.50**

Example Validation Accuracy

Decision Tree

	Naive Maximum			Changepoint			Majority Filter		
	RMSE	SD Error	MAD	RMSE	SD	MAD	RMSE	SD	MAD
Top of Limestone (pick2)	259.63	165.43	216.83	123.9	117.0668	2.2239	223.50	156.46	247.59
Top of Hawthorn (pick3)	33.18	21.98	17.79	60.16	47.35	19.64	**31.78**	**21.02**	**15.57**
Top of Hawthorn (pick6)	63.48	25.26	28.54	77.88	50.30	40.77	57.69	26.27	29.28

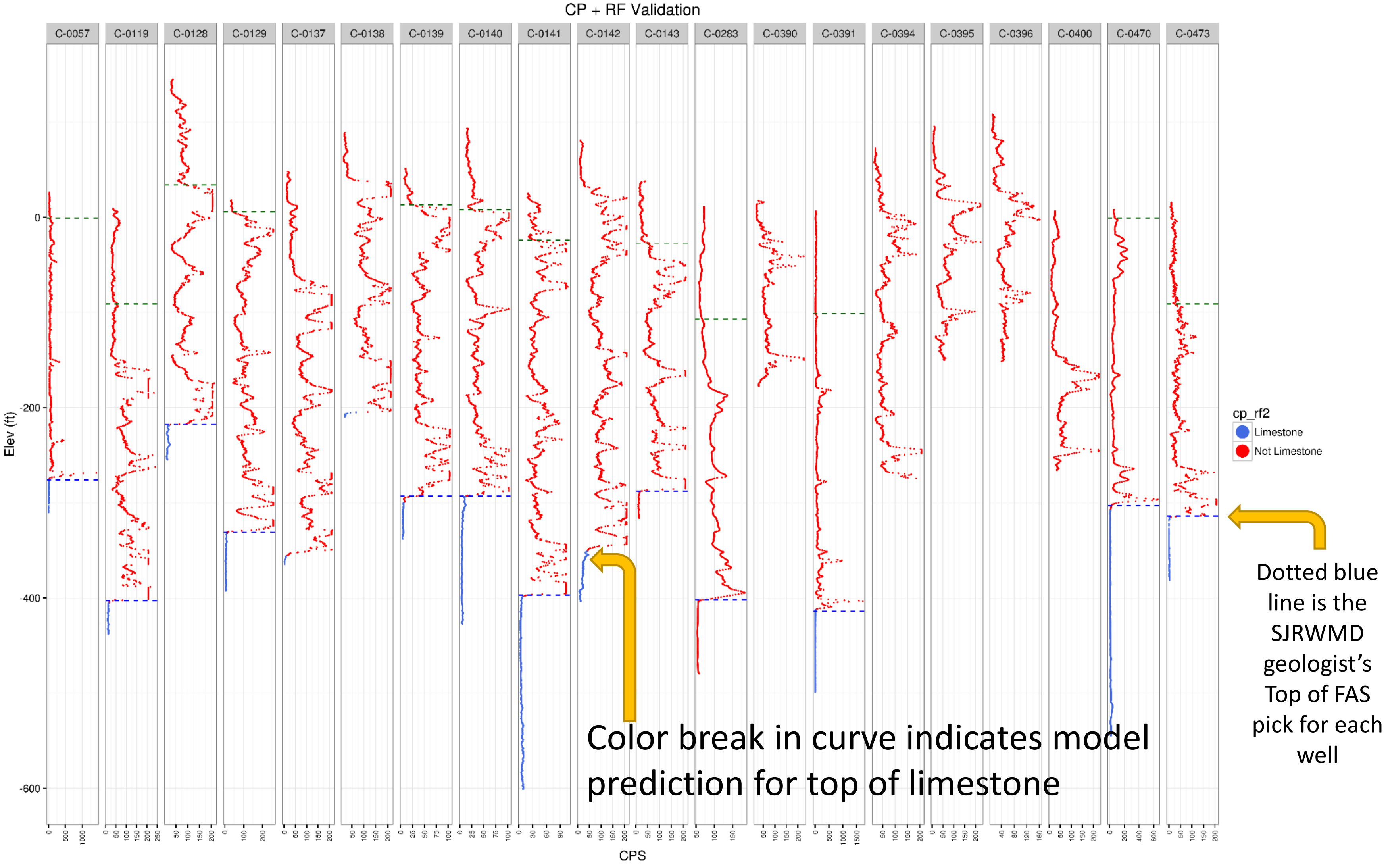
Random Forest

	Naive Maximum			Changepoint			Majority Filter		
	RMSE	SD	MAD	RMSE	SD	MAD	RMSE	SD	MAD
Top of Limestone (pick2)	54.09	49.55	7.78	**23.24**	**21.62**	**2.97**	39.91	36.26	6.67
Top of Hawthorn (pick3)	33.37	21.39	21.12	48.95	31.61	20.75	**31.44**	**19.90**	**17.79**
Top of Hawthorn (pick6)	37.95	24.08	22.98	58.79	35.84	34.84	**35.14**	**22.80**	**20.02**

Support Vector Machine

	Naive Maximum			Changepoint			Majority Filter		
	RMSE	SD	MAD	RMSE	SD	MAD	RMSE	SD	MAD
Top of Limestone (pick2)	54.33	39.88	14.08	**52.25**	**46.78**	**2.97**	53.66	39.29	13.34
Top of Hawthorn (pick3)	53.67	42.19	20.75	76.68	59.23	31.13	59.00	47.61	18.53
Top of Hawthorn (pick6)	43.00	29.66	17.04	67.12	48.61	32.24	**47.43**	**33.65**	**15.94**

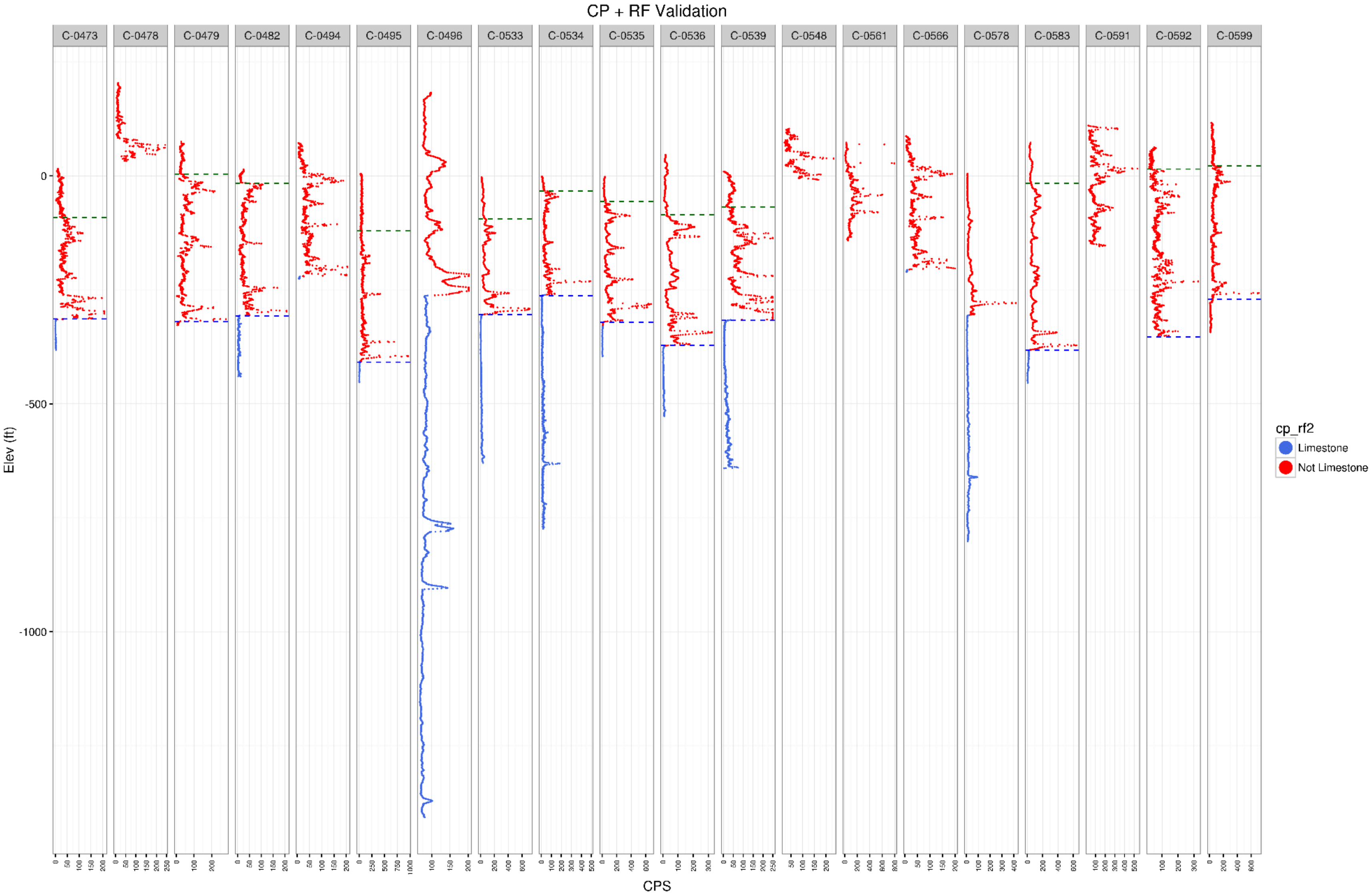
Seeing is Believing...



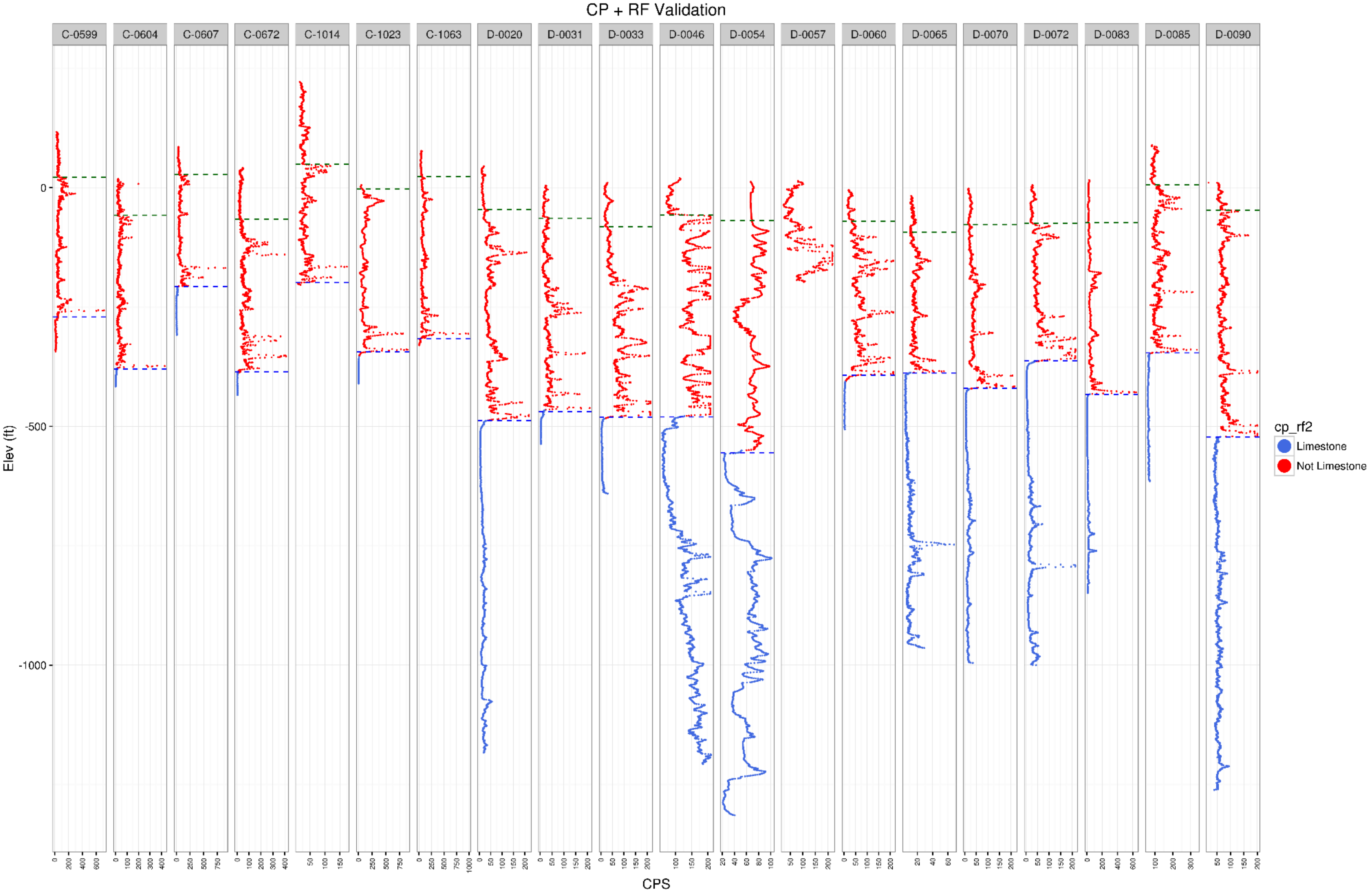
Color break in curve indicates model prediction for top of limestone

Dotted blue line is the SJRWMD geologist's Top of FAS pick for each well

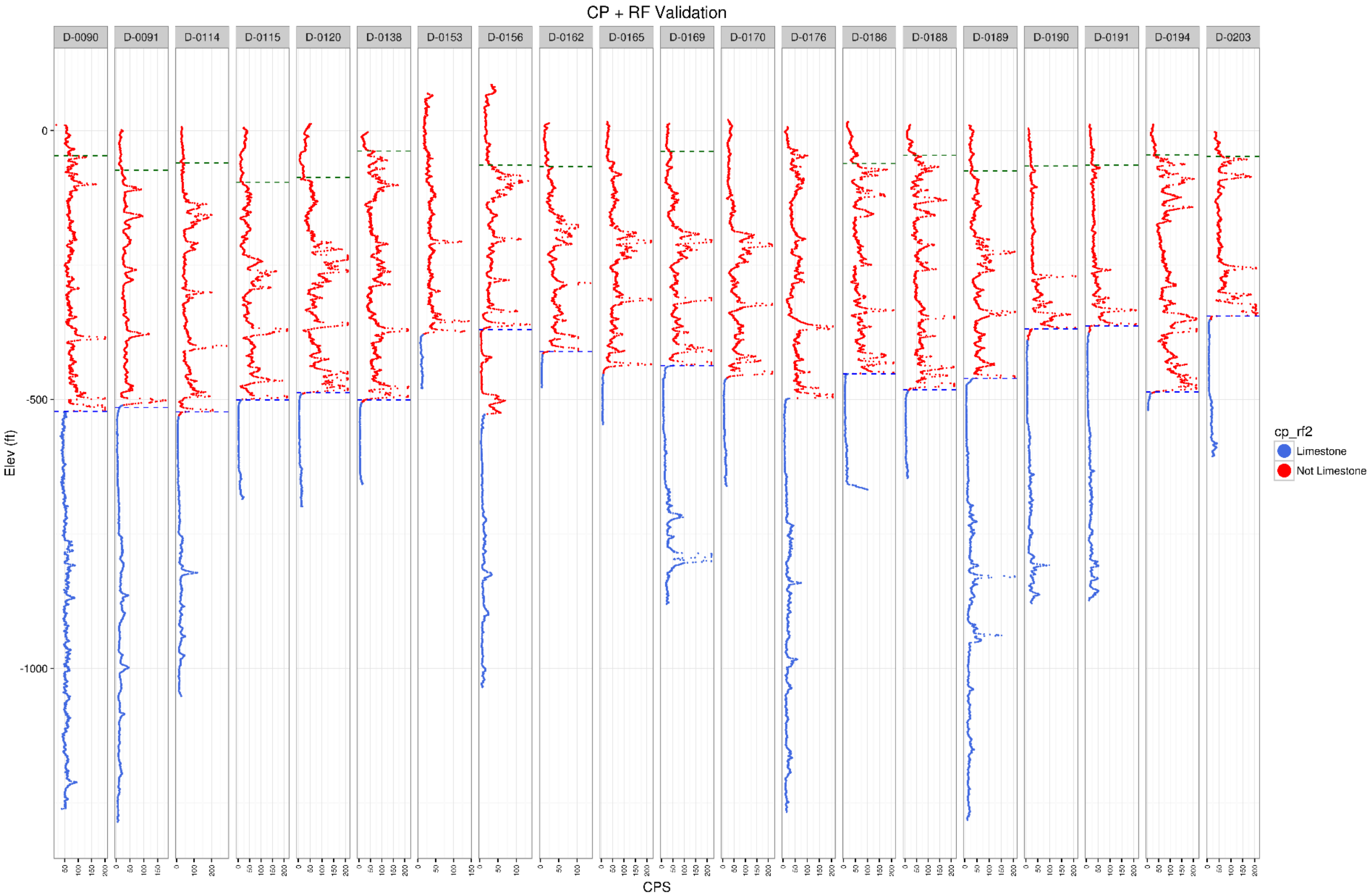
Seeing is Believing...



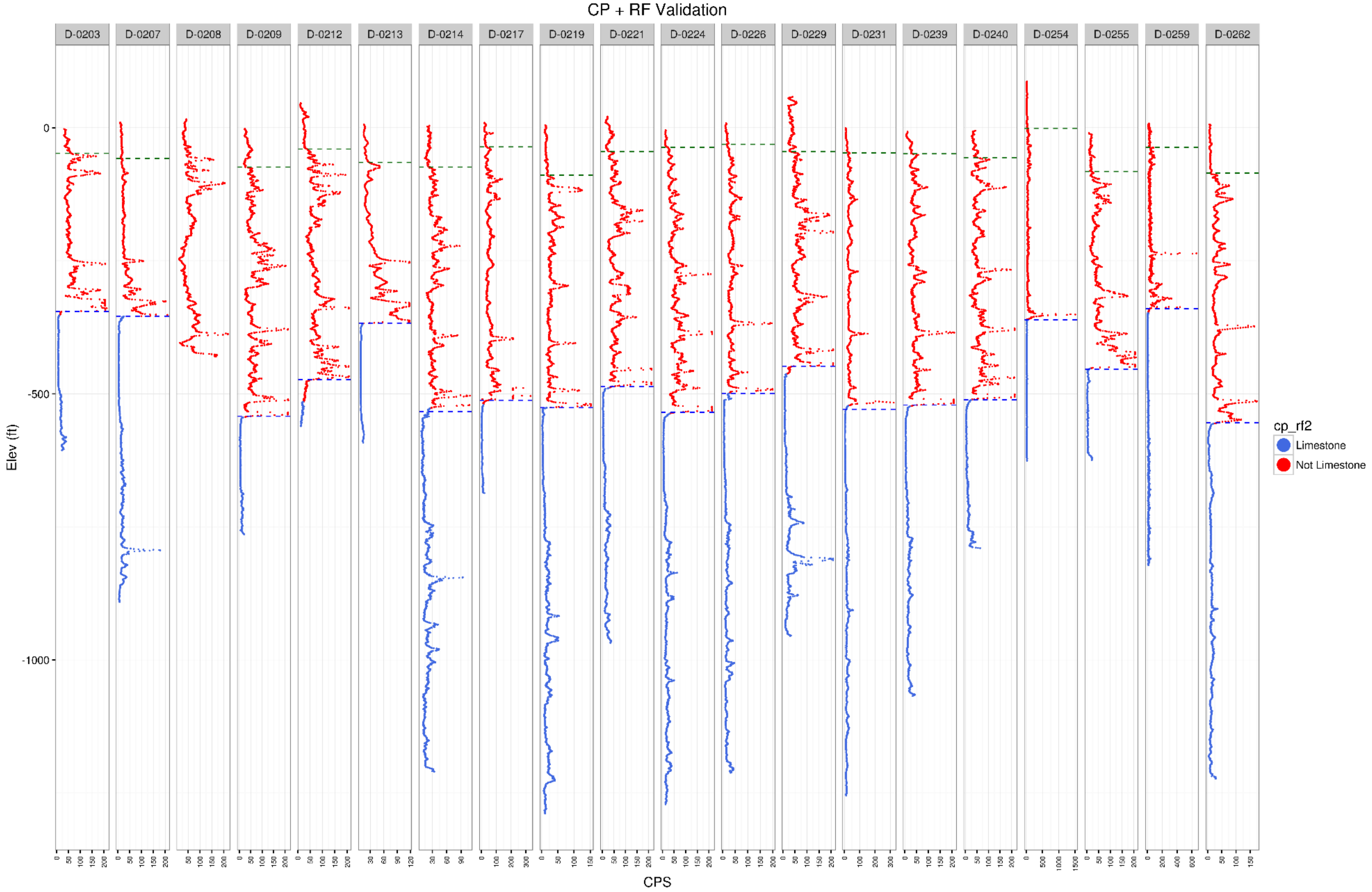
Seeing is Believing...



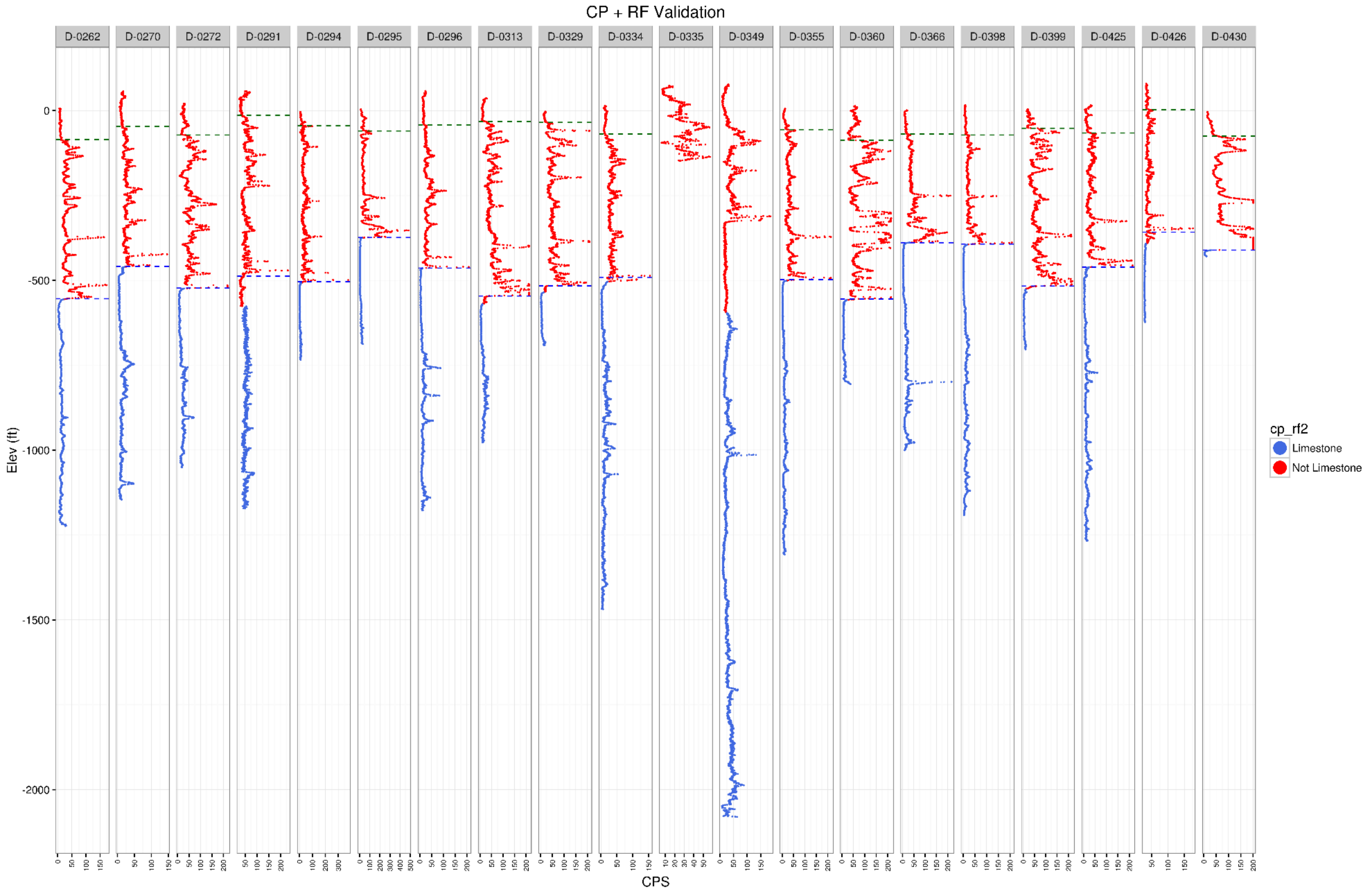
Seeing is Believing...



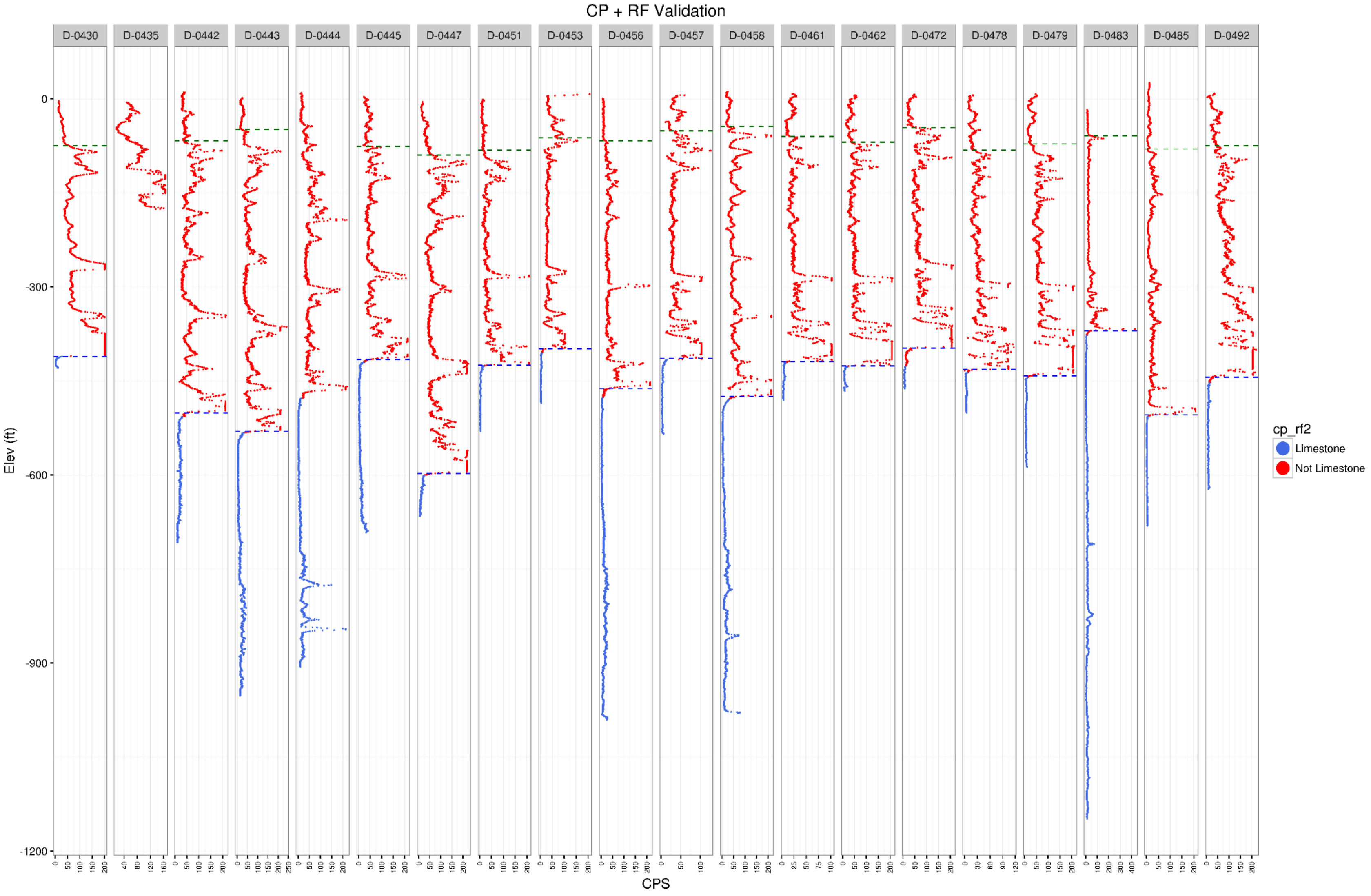
Seeing is Believing...



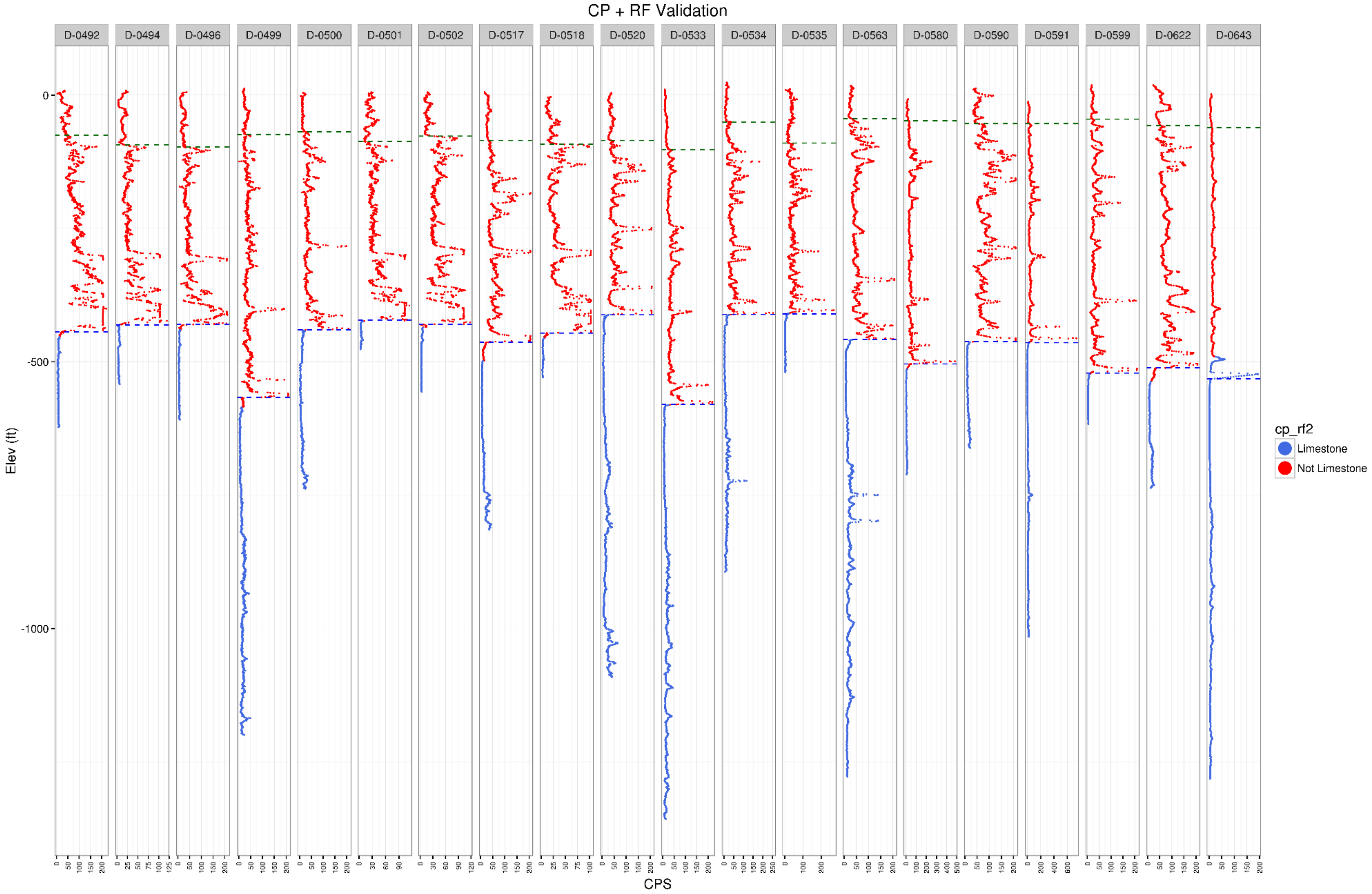
Seeing is Believing...



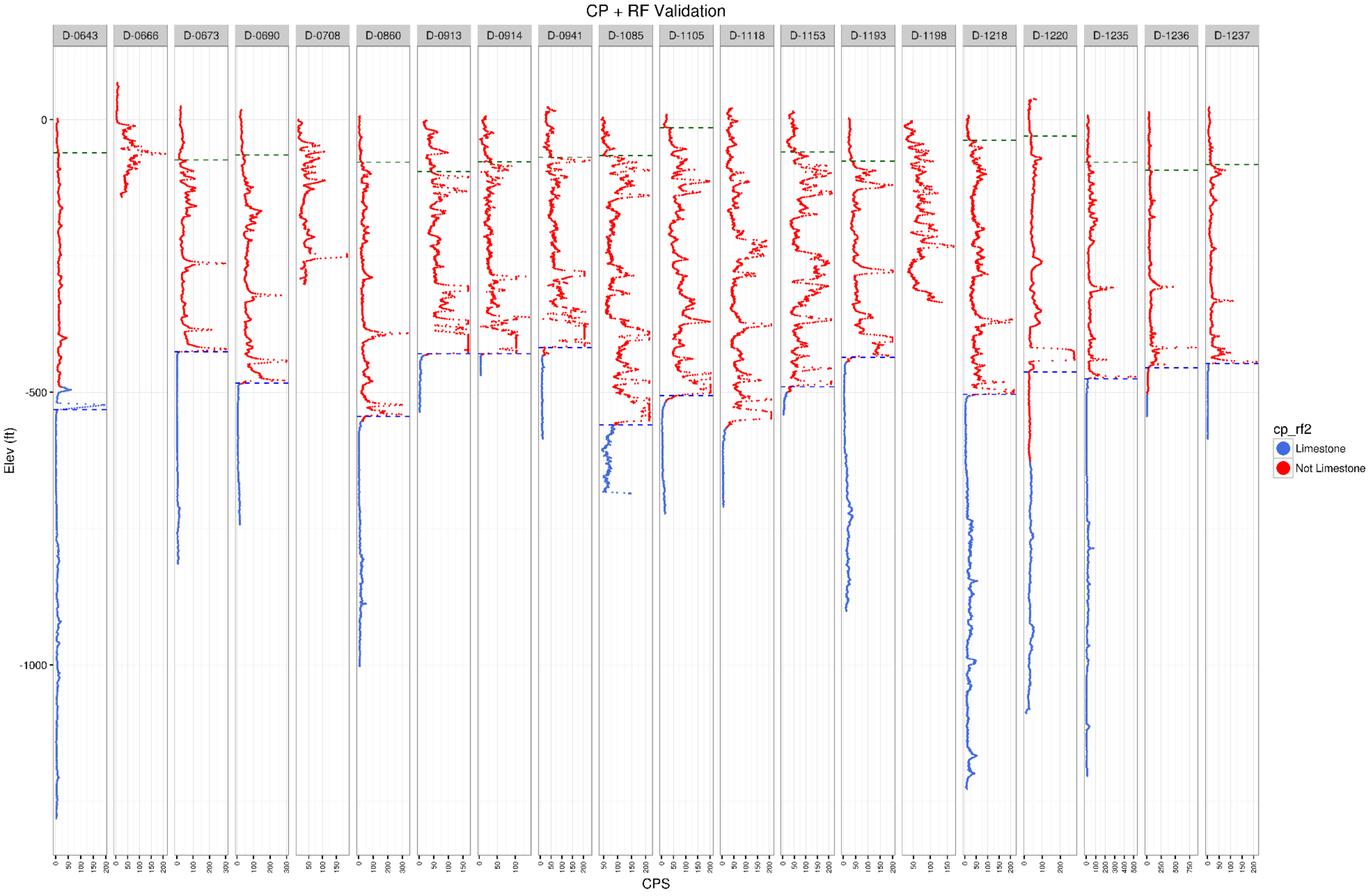
Seeing is Believing...



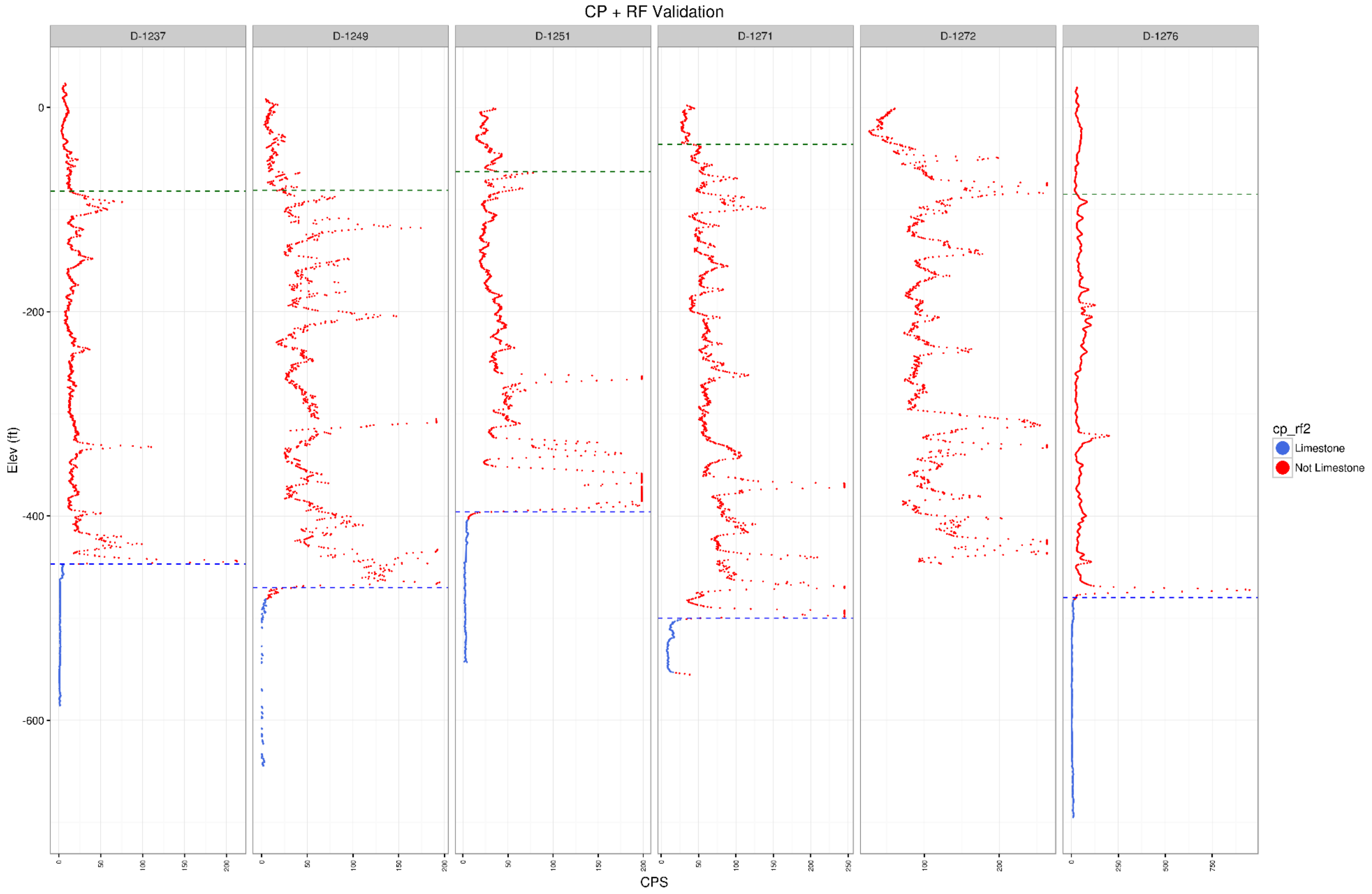
Seeing is Believing...



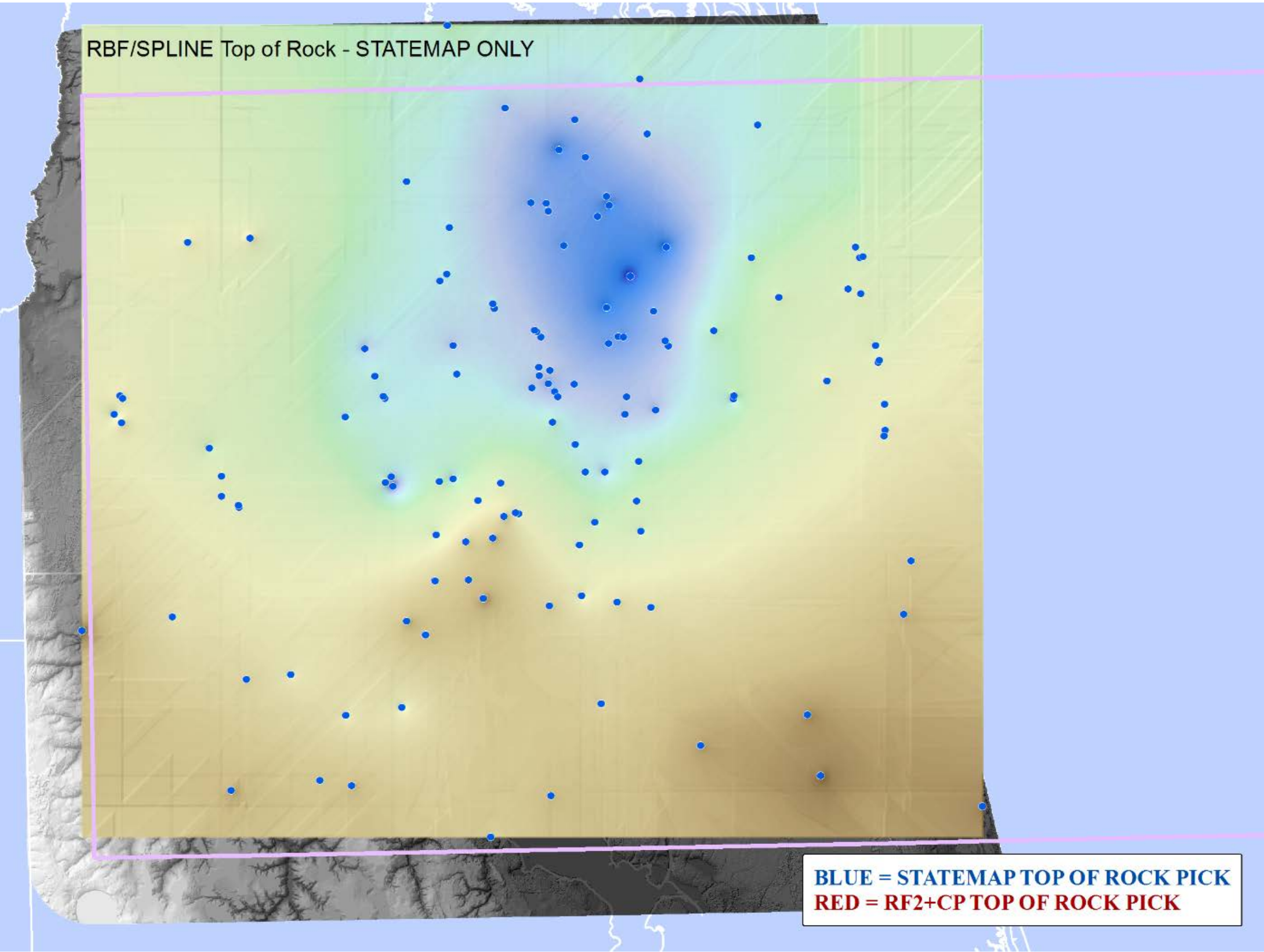
Seeing is Believing...



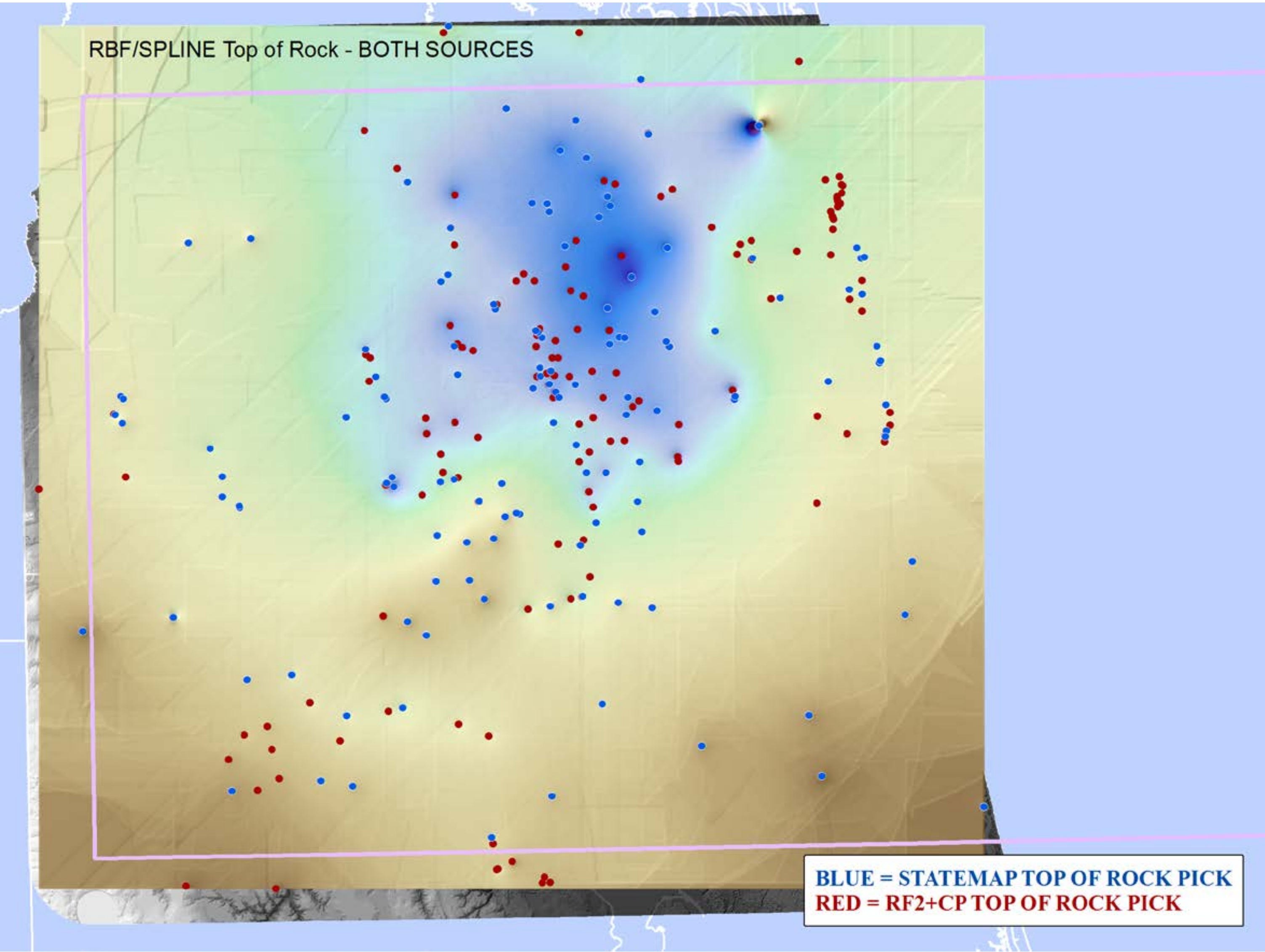
Seeing is Believing...



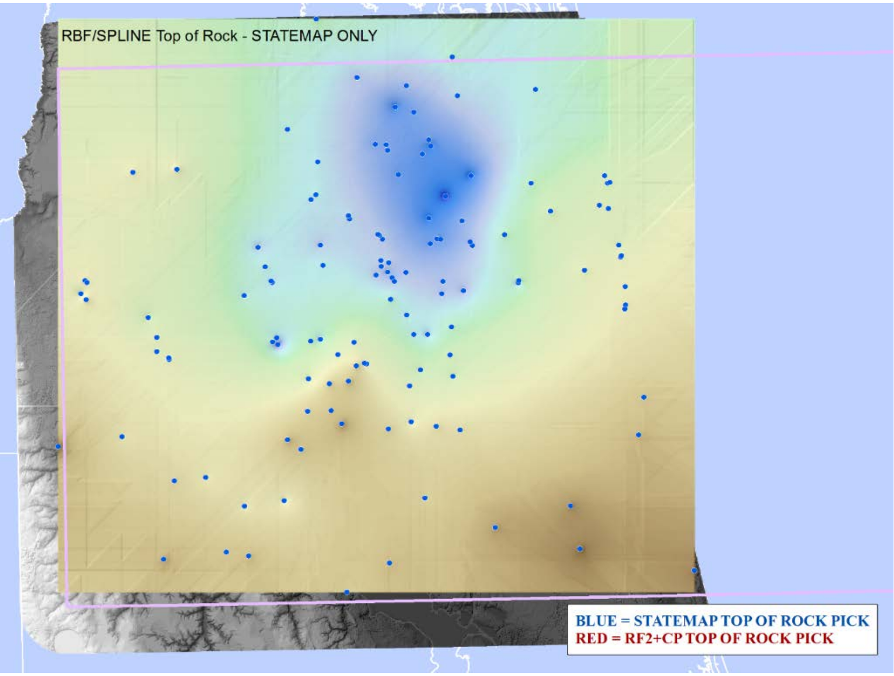
Ensemble Geospatial Modeling: The Jacksonville Basin

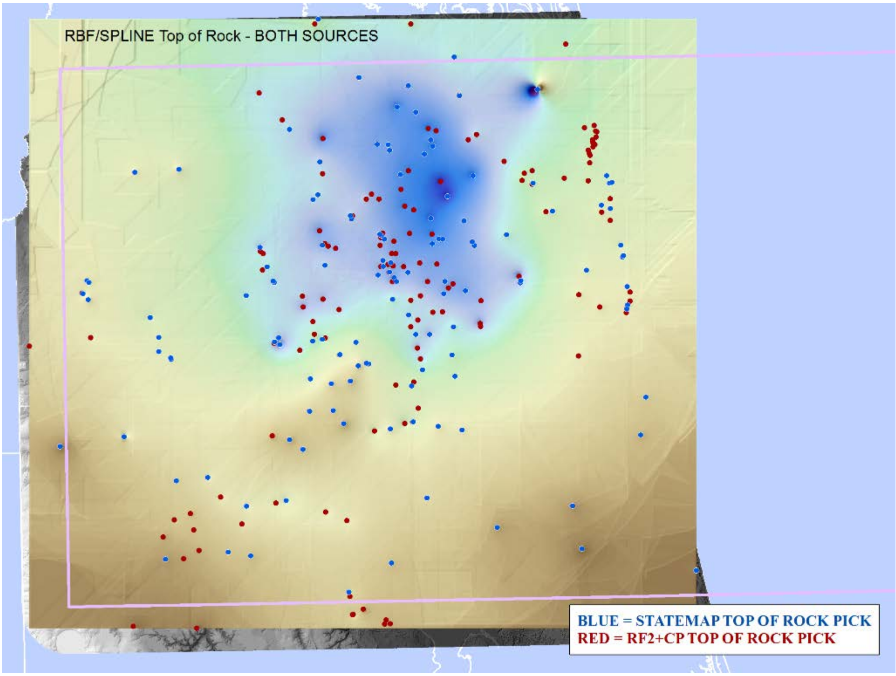

RBF/SPLINE Top of Rock - STATEMAP ONLY
BLUE = STATEMAP TOP OF ROCK PICK
RED = RF2+CP TOP OF ROCK PICK

Ensemble Geospatial Modeling: The Jacksonville Basin


RBF/SPLINE Top of Rock - BOTH SOURCES
BLUE = STATEMAP TOP OF ROCK PICK
RED = RF2+CP TOP OF ROCK PICK

Ensemble Geospatial Modeling: The Jacksonville Basin


RBF/SPLINE Top of Rock - STATEMAP ONLY
BLUE = STATEMAP TOP OF ROCK PICK
RED = RF2+CP TOP OF ROCK PICK


RBF/SPLINE Top of Rock - BOTH SOURCES
BLUE = STATEMAP TOP OF ROCK PICK
RED = RF2+CP TOP OF ROCK PICK

Conclusion:
We are living in the future

- Computers **can** “learn” to perform expert tasks like lithostratigraphic identification robustly and accurately
- Machine learning algorithms are effective at predicting lithostratigraphy using geophysical logging data
- Combining machine learning models and geospatial models produces a significant increase in mapping resolution due to increased data density
- Simpler/binary tasks are easier to model effectively than multiple classifications
- Machine learning algorithms will not replace all professional geologists in the workforce (GIGO)
  - However, one geologist supervising a computer **will** replace a geologists supervising a number of human staff in the next 15 years
  - Cf. drafting staff vs GIS personnel
- What about descriptions?
  - This ML workflow could be extended to automatically describe core samples based on continuous core images
- Other applications include hydrologic modeling, geohazard detection and prediction, automated surficial geologic mapping using remotely sensed data, and many more