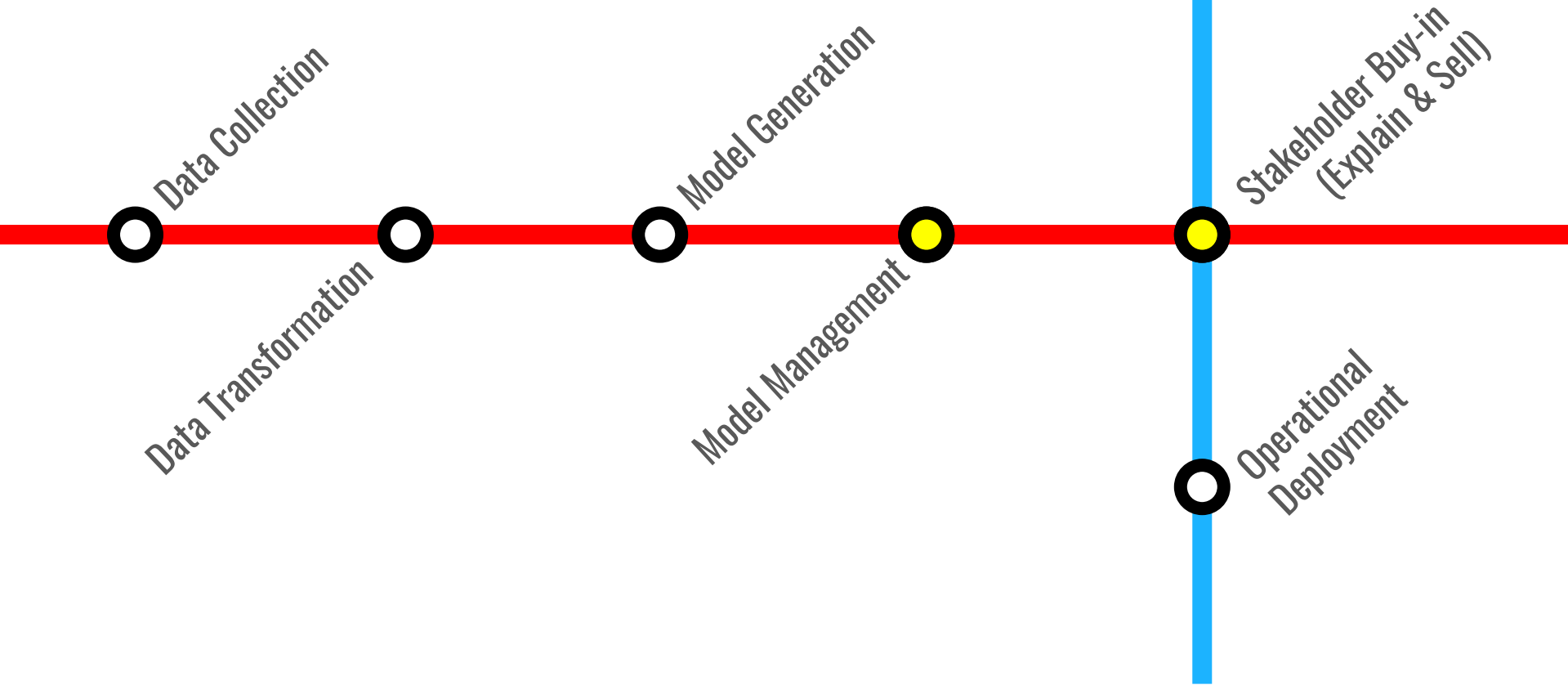


# **The Machine Learning Pipeline: Keeping Track of Models & Explaining Them (Visually)**

Weijing Tang & Erika Lee  
CS764  
Fall 2016

# A Typical Machine Learning Workflow



# Model Management



## *What*

- ★ Which algorithms, parameters, and features have been tried
- ★ How they each performed

## *Why*

- ★ Replicability
- ★ Minimize duplicate efforts

# Stakeholder Buy-in



## *What*

- ★ Present & explain what the best model does



## *Why*

- ★ Many machine-learning algorithms = black box
- ★ Need to convince decision makers & humans in the loop

# Data Transformation: Ava

★ Ava = A Machine Learning Chatbot developed by Rogers Jeffrey Leo John, Navneet Potti, Jignesh M. Patel @ UW Madison

The screenshot displays a Jupyter Notebook environment with a file explorer at the top showing 'Untitled2' and a last checkpoint from a minute ago. The notebook contains a code cell with the following content:

```
In [4]: loaded_data.describe(include="all").round(2)
```

The output of this cell is a detailed statistical summary of the data, including counts, unique values, top categories, frequency, mean, standard deviation, minimum, and various percentiles for each column. A 'Columns' dropdown menu is open, listing all available columns: Age, Workclass, fnlwgt, Education, Education\_Num, Marital\_Status, Occupation, Relationship, Race, Sex, and Capital\_Gain.

Below the code cell, another code cell is partially visible:

```
In [5]: print(loaded_data.columns)
```

On the right side of the notebook, there is a chatbot interface titled 'Ava: Your Data Science assistant'. The chat history shows:

- You:** hi
- Ava:** Hello! I am Ava, the data science chatbot. I am ready to chat.
- Hint:** You can instruct me to load data by typing "load data from filename", where filename is the full path to the file including the extension. e.g. "load data from Churn.xls".
- You:** load data from scratch/Desktop/adultdata.csv
- Ava:** LOADING DATA..
- Ava:** Data Loaded Successfully. There are 32561 rows and 15 columns in your data. Do you want to select columns as features?
- Hint:** Type yes to select columns as features.

# Model Generation: Ava

Jupyter Untitled2 Last checkpoint: 4 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Python [conda root]

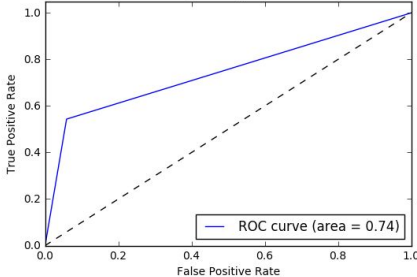
In [23]:

```
%matplotlib inline
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, auc
fpr, tpr, _ = roc_curve(y_test, dt_clf_gini_predicted_labels)
roc_auc = auc(fpr, tpr)
plt.figure()
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc='lower right')
plt.show()
```

Columns

- Age
- Workclass
- fnlwgt
- Education
- Education\_Num
- Marital\_Status
- Occupation
- Relationship
- Race
- Sex
- Capital\_Gain
- Capital\_Loss
- Hours\_per-week
- Country
- Income>50K

Receiver operating characteristic



— ROC curve (area = 0.74)

DEBUG:Ava:Ava,ROC generated.

Do you want to view any other plots?

Ava: Your Data Science assistant

- If None, then max\_features=number of features

You: 10

Ava: The accuracy after cross validation is 0.847.  
Do you want to run your model on testing data?

Hint: Type "yes" to see the accuracy of the algorithm on the test data.

You: yes

Ava: The testing accuracy is 0.846.  
What do you want to do next?

Hint: Type "show ROC plot" to see a ROC plot of the classifier you chose. You can also type "pick Column1, Column2, Column3 as features" to select a new set of features. If your current choice of classifier is decision trees type "show me the decision tree" to visualize the decision tree. Type "Run name-of-the-algorithm" to select an algorithm. e.g. "run logistic regression".

You: show ROC plot

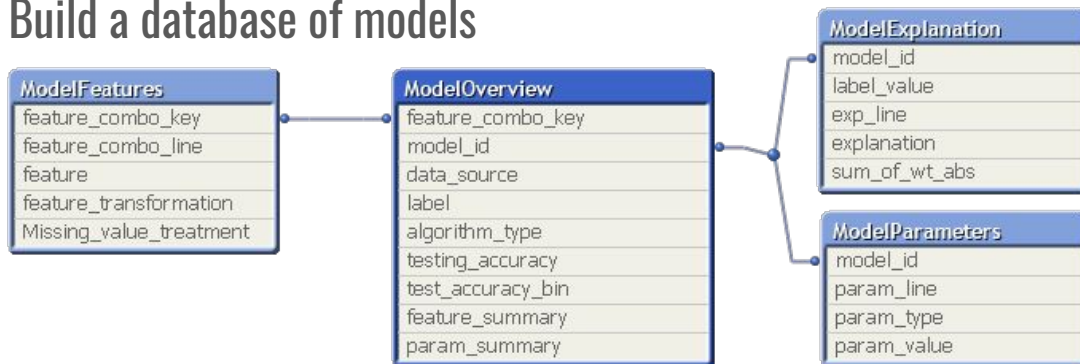
Ava: ROC generated.  
Do you want to view any other plots?

Hint: Type "show ROC plot" to see a ROC plot of the classifier you chose. If your current choice of classifier is decision trees type "show the decision tree" to visualize the decision tree. Type "No" to move on to the next conversation.

show the decision tree

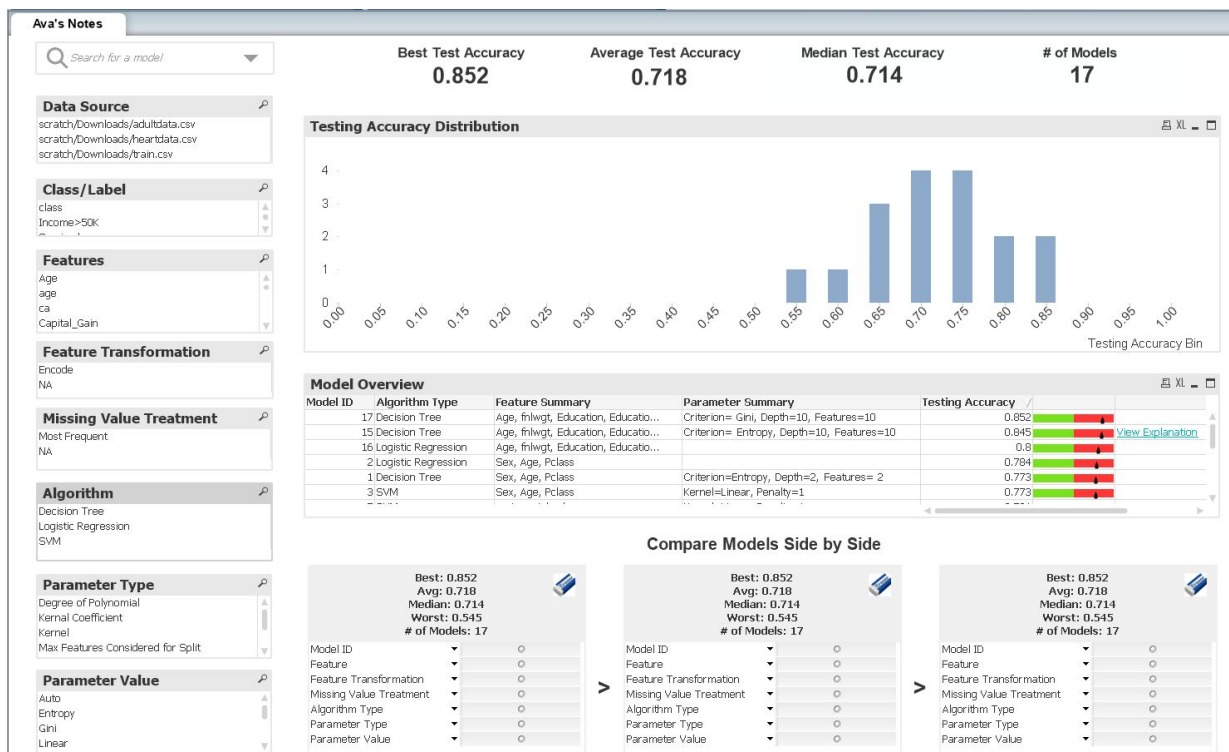
# Model Management: Ava's Notes

- ★ Extract discrete model information from Ava's logs
- ★ Build a database of models



- ★ Visualize it!

# Model Management: Ava's Notes

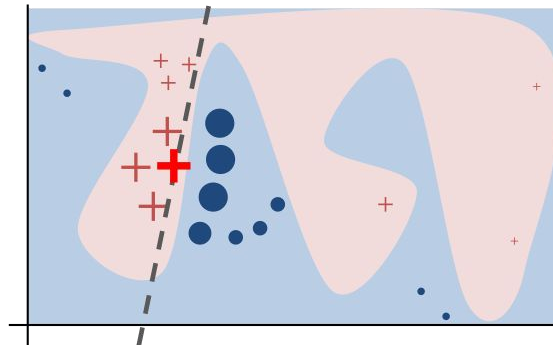




# Model Explanation: LIME

- ★ LIME = 'Local Interpretable Model-Agnostic Explanations' developed by Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin @ University of Washington

- ★ Intuition: learn a linear model approximating the actual model in the vicinity of a data point whose prediction we want to explain



# Model Explanation: LIME



## *Given*

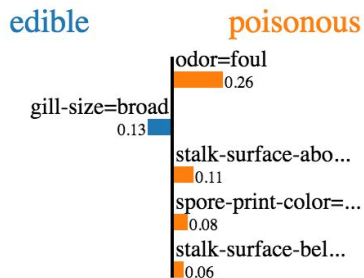
- ★ a data point whose prediction we want to explain
- ★ the predictions from a learned model



## *Output*

Prediction probabilities

edible	0.00
poisonous	1.00



Feature	Value
odor=foul	True
gill-size=broad	True
stalk-surface-above-ring=silky	True
spore-print-color=chocolate	True
stalk-surface-below-ring=silky	True

# Model Explanation: Globalizing LIME






- I. Aggregate the weight of “Distinguishing Features” identified at different data points
  - ★ Brute-force through all test data points?
  - ★ Randomly sample test data points?
  - ★ Cluster the test data points, use the centroids?
- II. Find the highest ranking “Distinguishing Features” for each unique label value



# Model Explanation: Globalizing LIME



Explanation of Model #15			  
Prediction	Most Distinguishing Features	Weight of Feature	
Income<=50K	Capital_Gain <= 0.00	54.16	<div></div>
	Marital_Status_Married-civ-spouse=0	22.81	<div></div>
	Occupation_Exec-managerial=0	13.05	<div></div>
	Education_Num <= 9.00	10.29	<div></div>
	Hours_per-week <= 40.00	6.85	<div></div>
Income>50K	Capital_Gain > 0.00	28.19	<div></div>
	Marital_Status_Married-civ-spouse=1	22.75	<div></div>
	Education_Num > 12.00	10.38	<div></div>
	Occupation_Other-service=0	2.36	<div></div>
	Occupation_Exec-managerial=1	2.19	<div></div>



# Model Explanation: What Next?



Further extension on LIME



Clustering + Discriminant Analysis



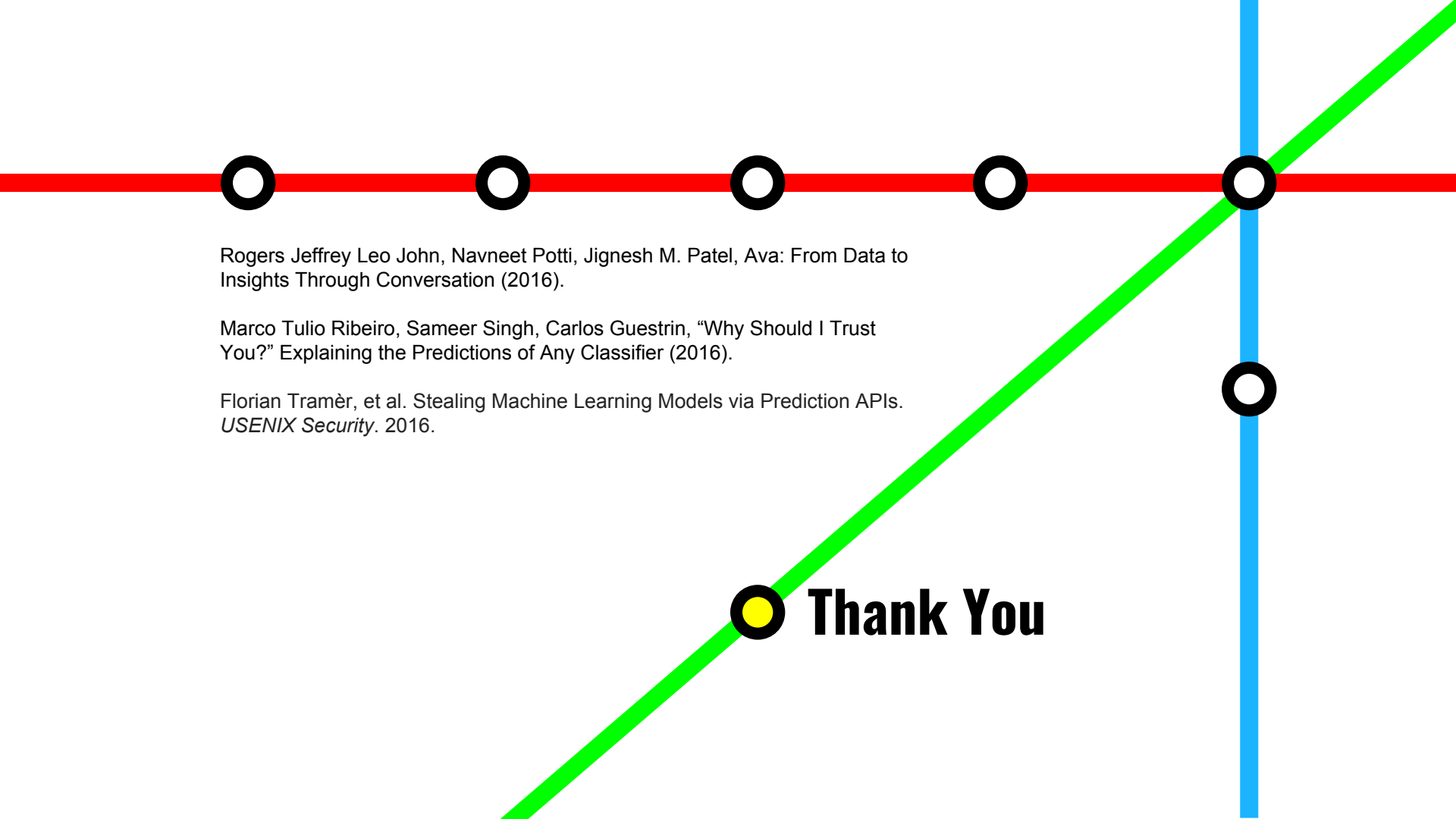
Principal Component Analysis/

Data Reduction Approach



“Hack”/reverse-engineer





Rogers Jeffrey Leo John, Navneet Potti, Jignesh M. Patel, Ava: From Data to Insights Through Conversation (2016).

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, “Why Should I Trust You?” Explaining the Predictions of Any Classifier (2016).

Florian Tramèr, et al. Stealing Machine Learning Models via Prediction APIs. *USENIX Security*. 2016.

**Thank You**