FACILITATING ANALYTIC QUERIES OVER GEOSPATIAL TIME-SERIES DATA

Background

- Distributed Hash Table
  - O(1)
- Graph-based index
- Dataset: NOAA NAM
  - Spatiotemporal
  - ~100 dimensions: climate observations
  - Wind speed, temperature, humidity…
  - Very large

Dataset Issues (1/2)

- Each composite ‘observation’ from NOAA contains information for every location in N. America
- Observations are generated at set intervals, ordered by time
- Great for things like predicting the weather
- Not great for other types of analysis…

Dataset Issues (2/2)

- Q: What happens if you want to create a summary climate chart for Colorado in 2014?
  - A: You parse about 1 TB of files
- Q: What if you decide to do the same thing, except with Wyoming?
  - A: You parse about 1 TB of files
- Q: What if…
  - Hey, it's time to go home!

Analytic Queries

- Key aspects:
  - Speed (Avoid disk I/O!)
  - Managing Trade-offs (timeliness, accuracy…)
  - Higher-level functionality
- Three flavors:
  - Approximate
  - Exploratory
  - Predictive

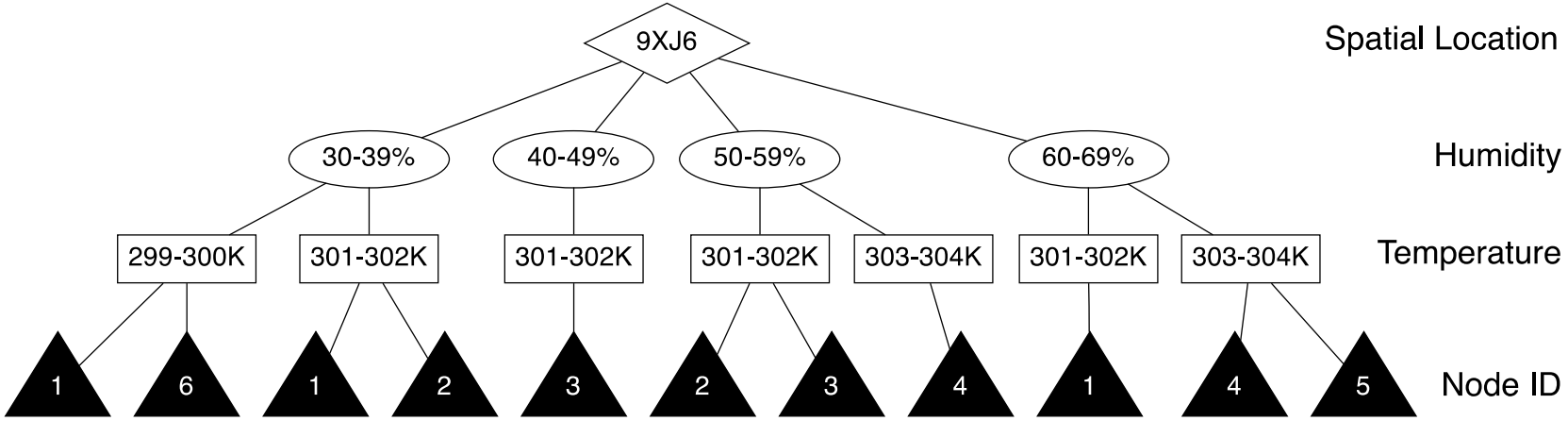
Analytic Queries

- Rather than retrieving ***data***, retrieve ***insights***
- Exploratory Analytics:
  - What relationships exist in the dataset? How do the variables interact?
  - What is the weather usually like in Colorado?
- Predictive Analytics:
  - What is the likelihood it will snow today?
  - Will this change to our supply chain improve sales?

Components

- The Index (Metadata Graph)
  - Store coarse-grained information about the data
  - Online summary statistics
  - Autonomous reconfiguration at runtime
- Queries
  - User restrictions on time and accuracy
  - Predictive Models
  - Sampling
- MapReduce

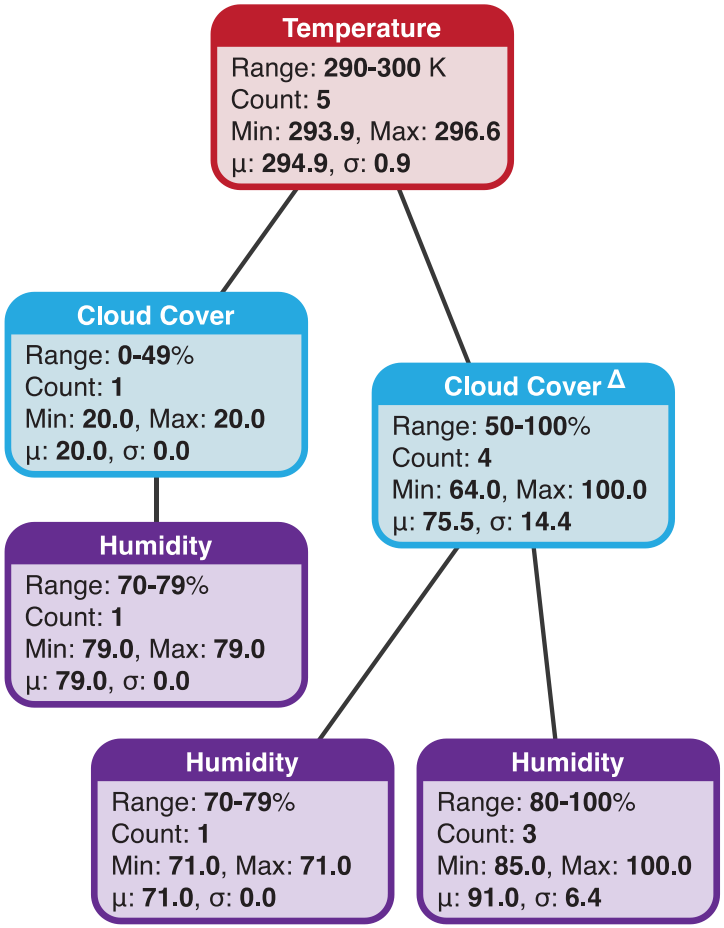
An Example Metadata Graph



(Conversions: 0C = 273K = 32F)

The Index

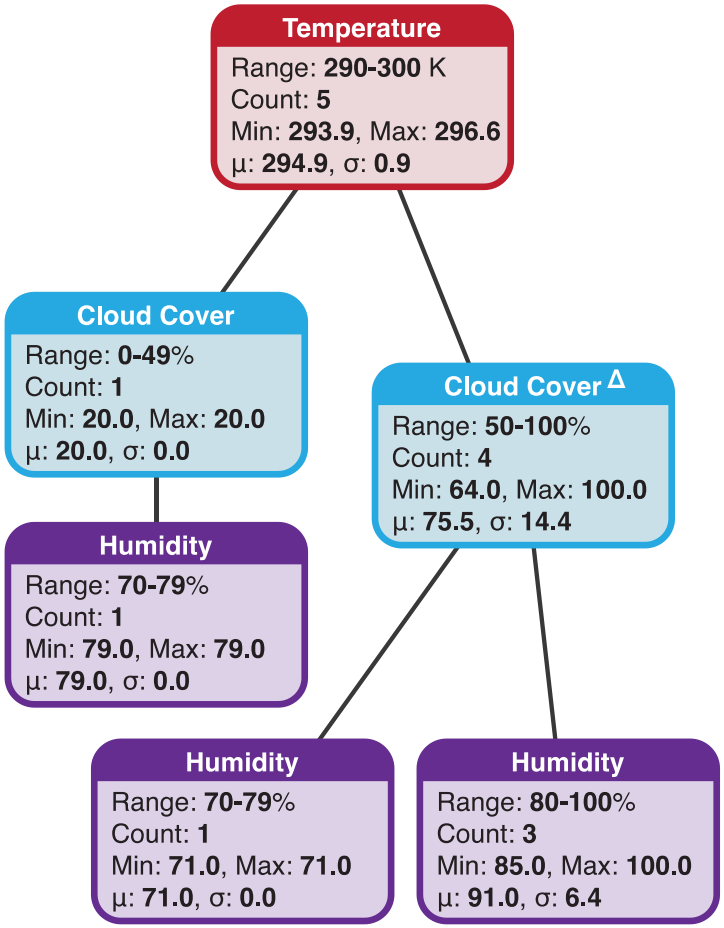
- Track statistics to facilitate analytics functions
- Benefits:
  - Avoids slow disk accesses
  - Incremental creation of models
  - Fast query turnaround times



Inserting Paths

Path	Temperature	Cloud Cover	Humidity
A	295.4 K	66.0%	88.0%
B	296.6 K	64.0%	100.0%
C	294.5 K	72.0%	71.0%
D	293.9 K	20.0%	79.0%
E	294.1 K	100.0%	85.0%

Statistic	A	B	C	D	E
Count	1	2	3	3	4
Min	66.0	64.0	64.0	64.0	64.0
Max	66.0	66.0	72.0	72.0	100.0
Mean (μ)	66.0	65.0	67.3	67.3	75.5
St. Dev. (σ)	0.0	1.0	3.3	3.3	14.4



Welford's Method

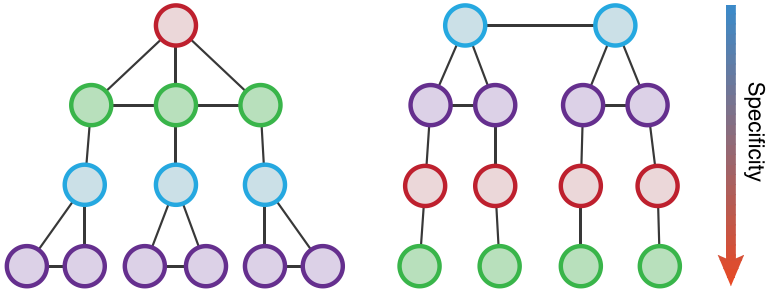
- Vertices stay up to date via Welford's method
  - Allows computing variance, mean, etc. in a single pass
- New data is transformed into a hierarchical graph ***path***
- If a path touches a vertex, its statistics are updated
- Extension: for each vertex, keep cross-feature statistics (2D)

Welford Performance

Operation	**Time (μs)**	σ **(μs)**
Add Data Point	1.489	0.044
Calculate Correlation	0.723	0.023
Calculate r^2	0.101	0.003
Predict y	0.381	0.016
Merge 2D Instances	0.919	0.103

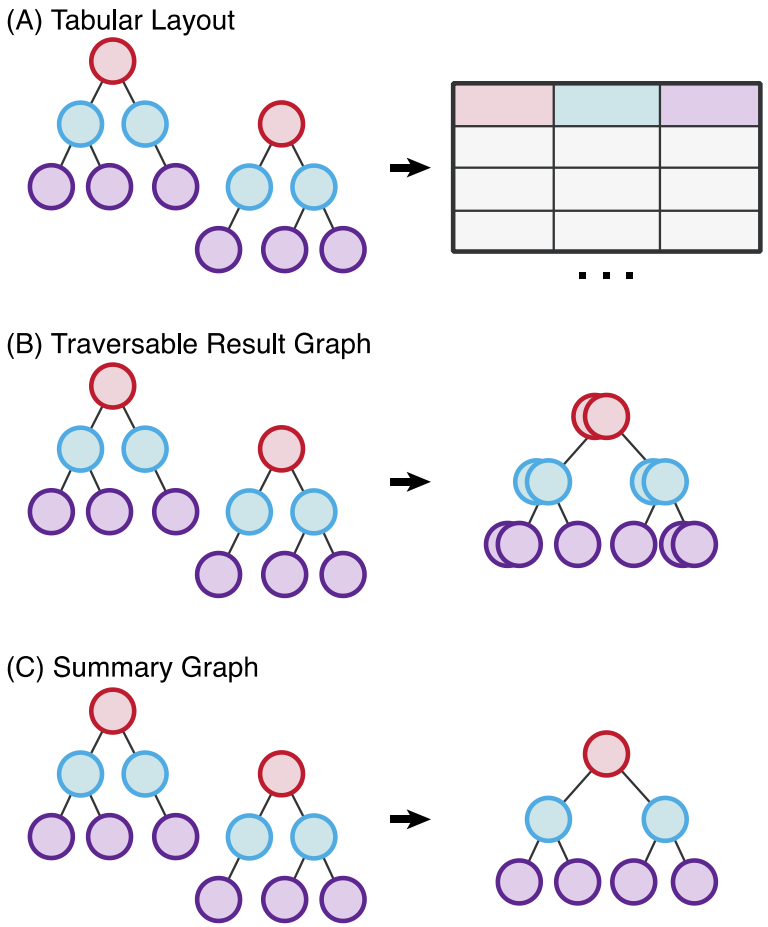
Vertex Specificity

- Graph orientations influence specificity
- General insights: near the top of the graph
- Specific Insights: near the bottom
- User can choose, or system choses automatically



Result Datasets

- Tabular: streaming, fast, huge
- Summary graph: merge all vertices across machines to create one graph
- Result graph: middle ground



Analytics Functionality

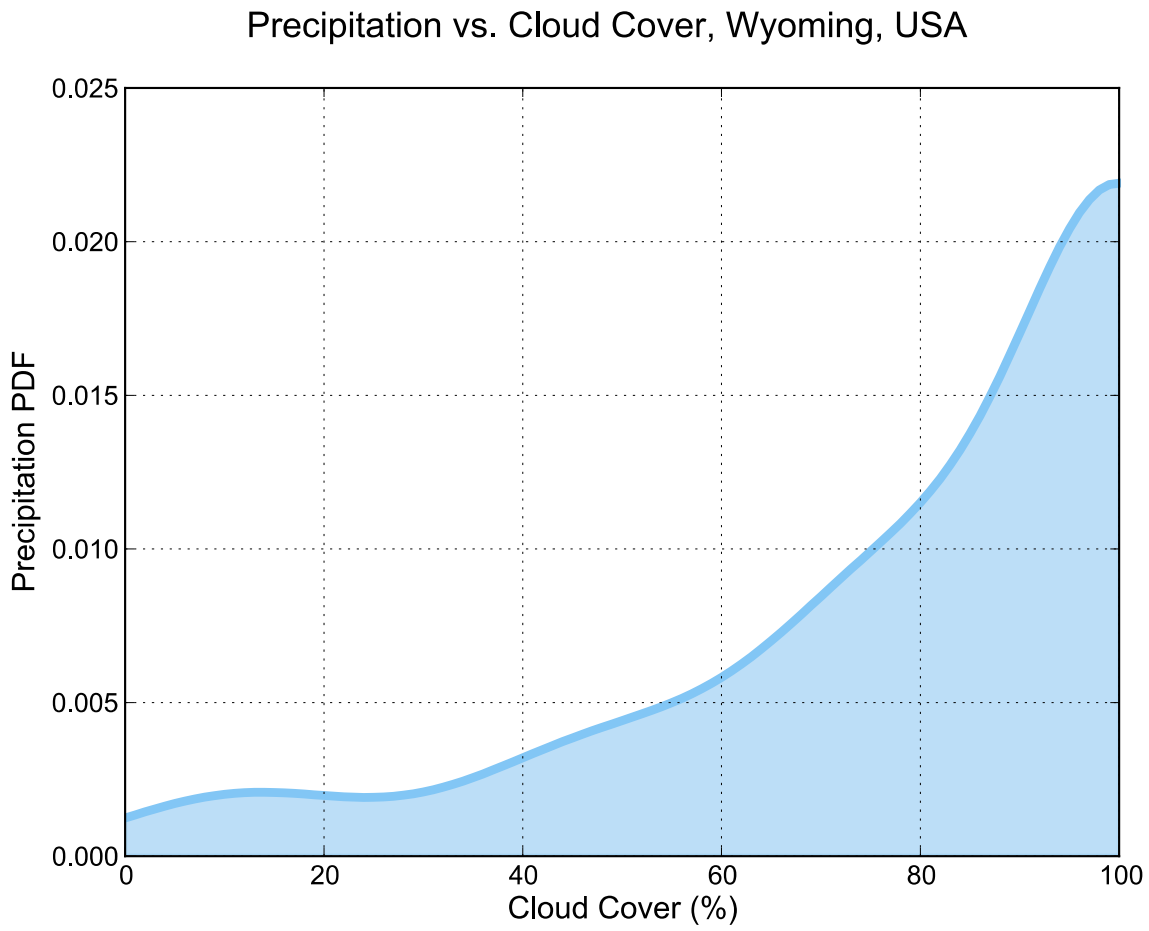
- Exploratory
  - Correlations between dimensions
  - Probability Densities
- Predictive
  - Significance testing, hypothesis testing
  - Predictive models: multiple linear regression, neural networks, ARIMA
  - Bayesian Classification

Creating Predictive Models

- Issue a query that selects some subset of the graph to use for training data
  - This may require disk accesses if historical data is necessary
- When new data points arrive that fall within the model scope, update it
  - The models we consider operate in a streaming mode, or support batch ingest of new data

Applications

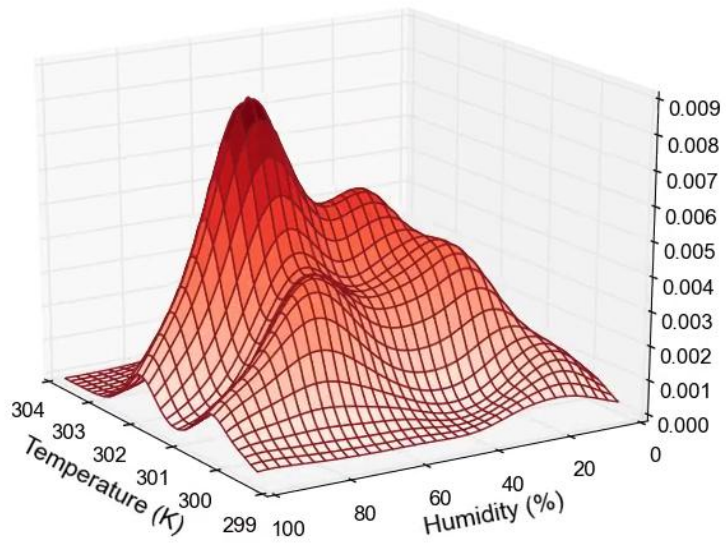
PDF Queries



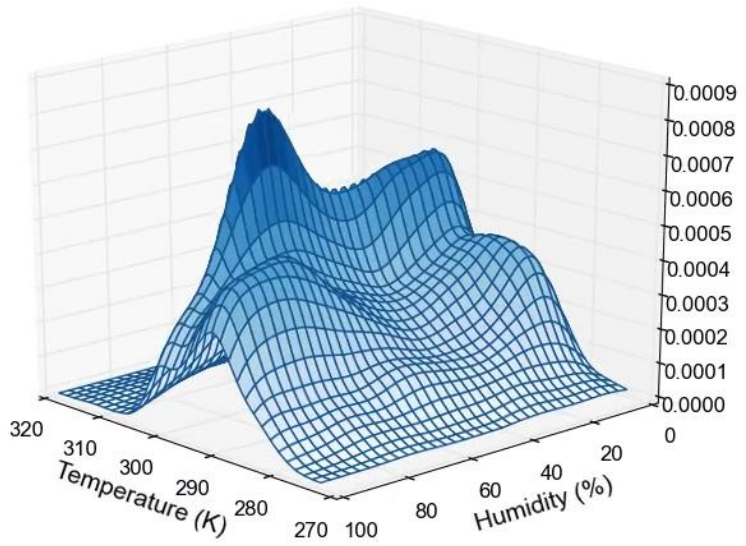
PDF(Cloud_Cover WHERE Precipitation > 0)

Interactions: Temperature and Humidity

PDF(Temperature ∩ Humidity): Florida, USA



PDF(Temperature ∩ Humidity): Continental United States

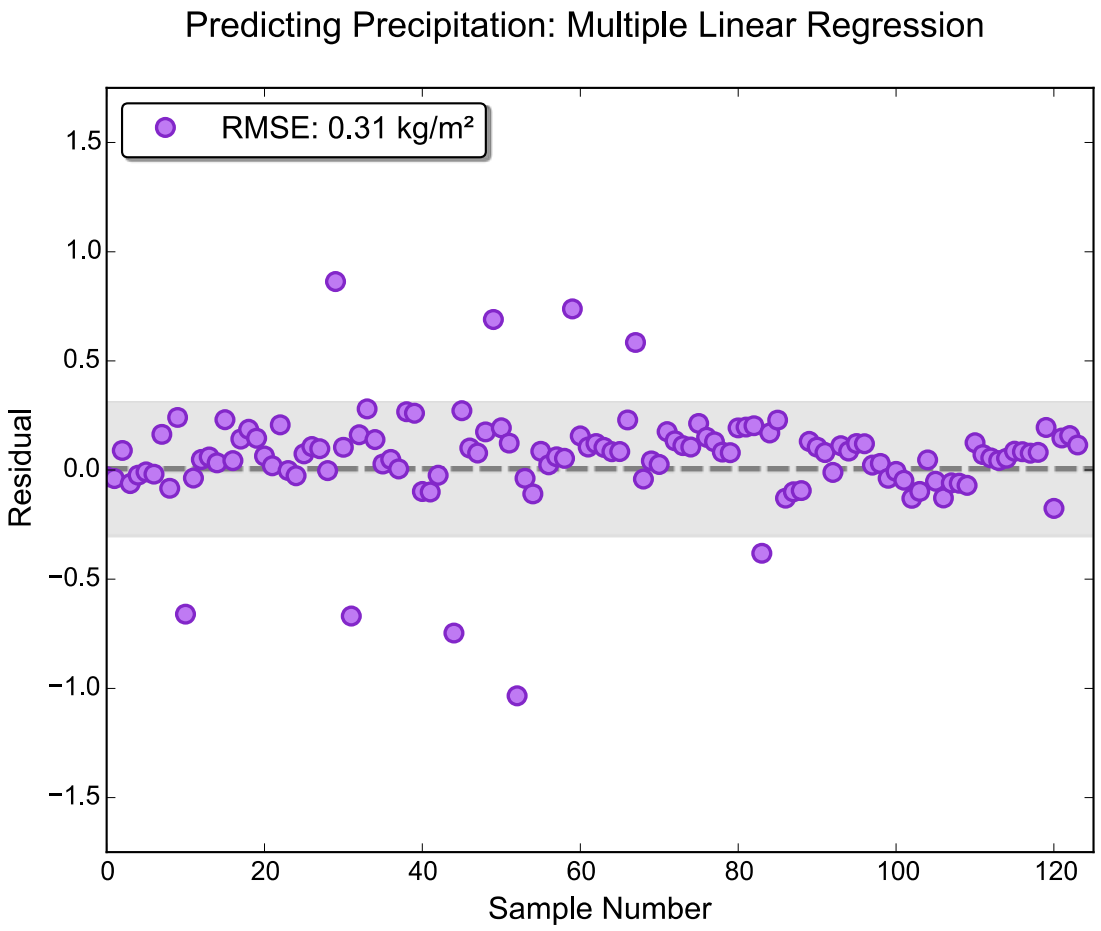


Feature Correlation Coefficients

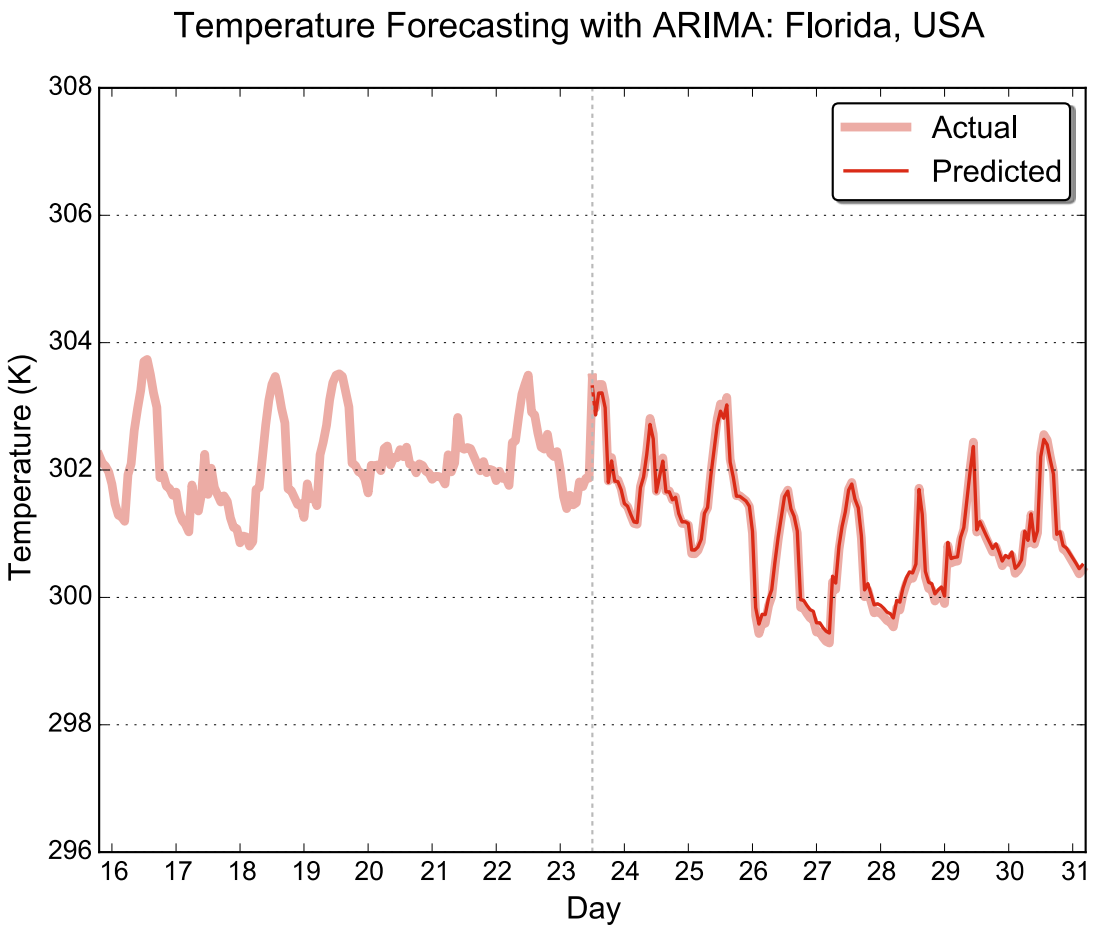
Feature A	Feature B	Correlation	p-value
Precipitation	Visibility	-0.49	3.39E-39
Humidity	Precipitation	0.37	3.08E-22
Pressure	Visibility	0.36	1.61E-20
Vegetation	Temperature	-0.06	0.12
Temperature	Snow Depth	0.0	1.0

Source: NOAA NAM July 2013, Wyoming

Binary Classifier: Precipitation

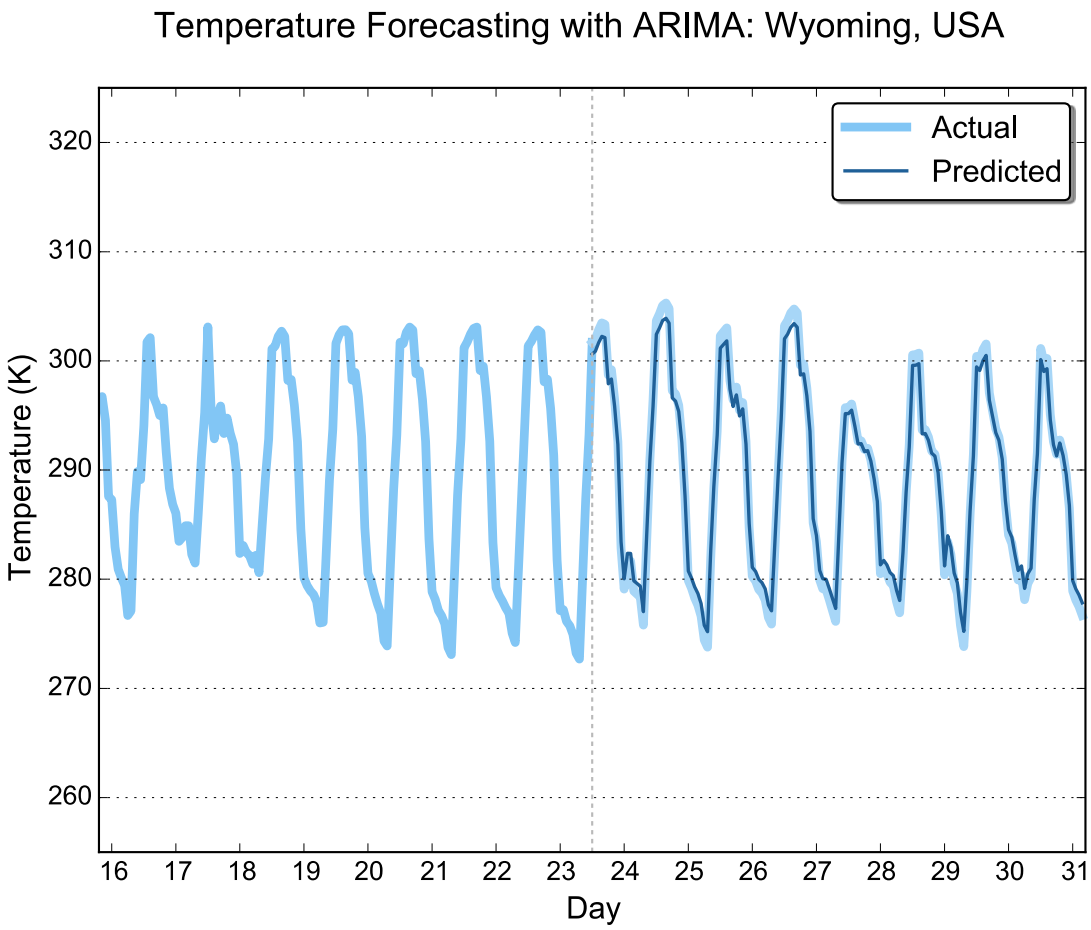


ARIMA Temperature Forecast: FLA



RMSE:
0.077

ARIMA Temperature Forecase: WY



RMSE:
0.818

Future Research Directions

- The index optimizes itself autonomously, but there are many avenues for improving this
  - Use summary statistics to guide splitting/merging of vertices
- Autonomously create models
  - Distant future: predict answers to queries?
- Causality analysis, Bayesian networks

Questions?