

Visualization for Machine Learning

Cmpt 733 – Big Data Programming 2

Steven Bergner

sbergner@sfu.ca

Analysis of Big Data via ML

Outline

- Machine learning tasks
- Vis for ML
- ML for Vis
- Google Compute
- Best of: Weather model vis

Tasks performed **by** and **for** ML

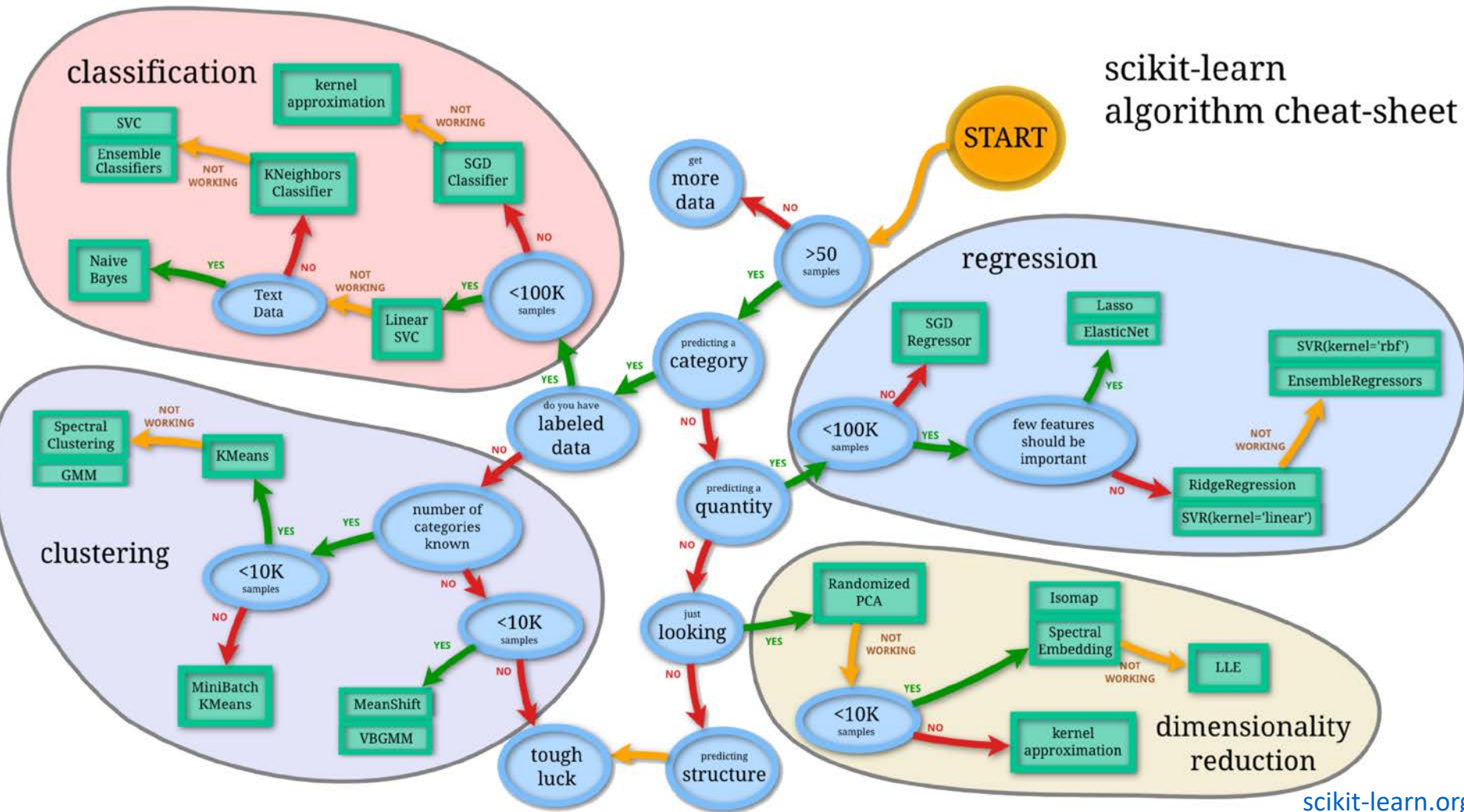
ML Task Categories: By Output

- Classification
- Regression
- Clustering
- Association rules
- Forecasting
- Dimensional reduction
- Density estimation

ML Task categories: By Training

- **Supervised learning**
 - Inputs and desired outputs are given
 - Find rule that maps unseen inputs to outputs
- **Semi-supervised learning**
 - Supervised learning with only few of the target outputs given
- **Active learning**
 - Data is not given, but asked for
- **Unsupervised learning**
 - No labels given
 - Discover hidden patterns and structure in inputs
 - Feature learning
- **Reinforcement learning**
 - Rewards/punishments given as feedback to actions in dynamic environment

scikit-learn algorithm cheat-sheet



Tasks **for** Machine Learning

For a given problem

- Choose a model class & estimator, loss function, optimization method
- Determine the right training data (attributes, distribution)

For a given model class

- Estimate model parameters to fit with given observations
- Make predictions, determine and **communicate uncertainty**
- Analyse model behaviour over a region of inputs
- **Understand how model works, explain its decision making**
- Validate fit of training assumptions vs operating conditions

All of the above: data-driven

Some: **human-in-the-loop**

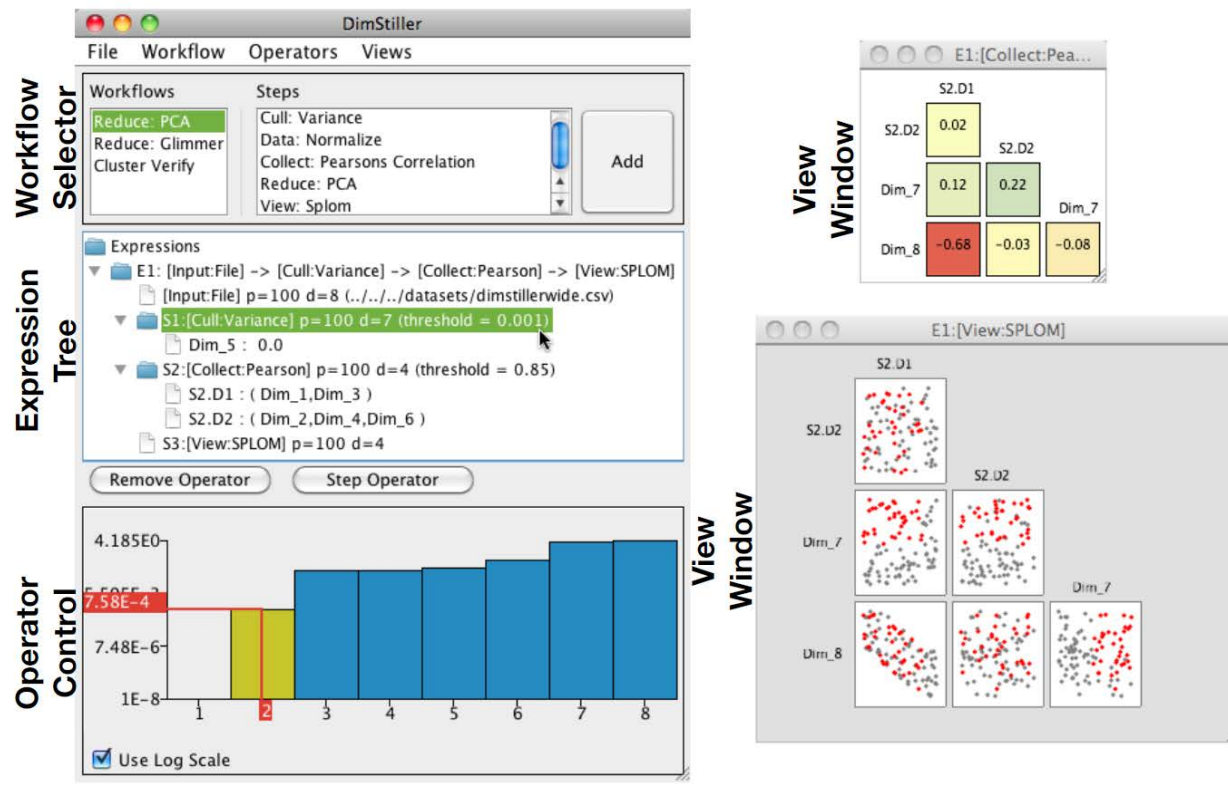
Cluster Analysis

Cluster Visualization

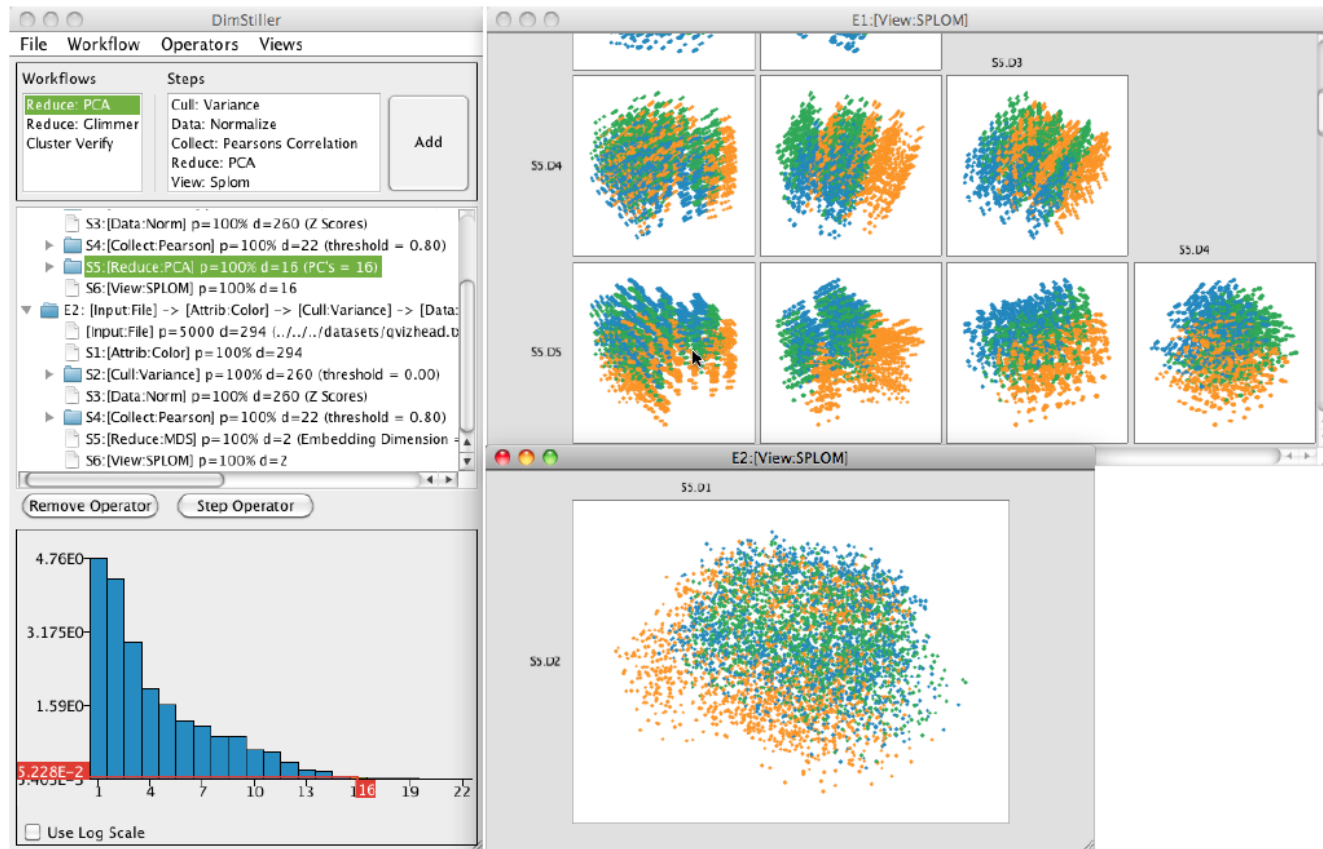
- Treat cluster label as categorical variable
- Multi-variate vis technique with encoding for label attribute

Example: DimStiller Workflows

- Goal: Understand and transform input data
- Chain operators together into pipelines



DimStiller Cluster Analysis Example



Dimension Reduction

Dimension Reduction

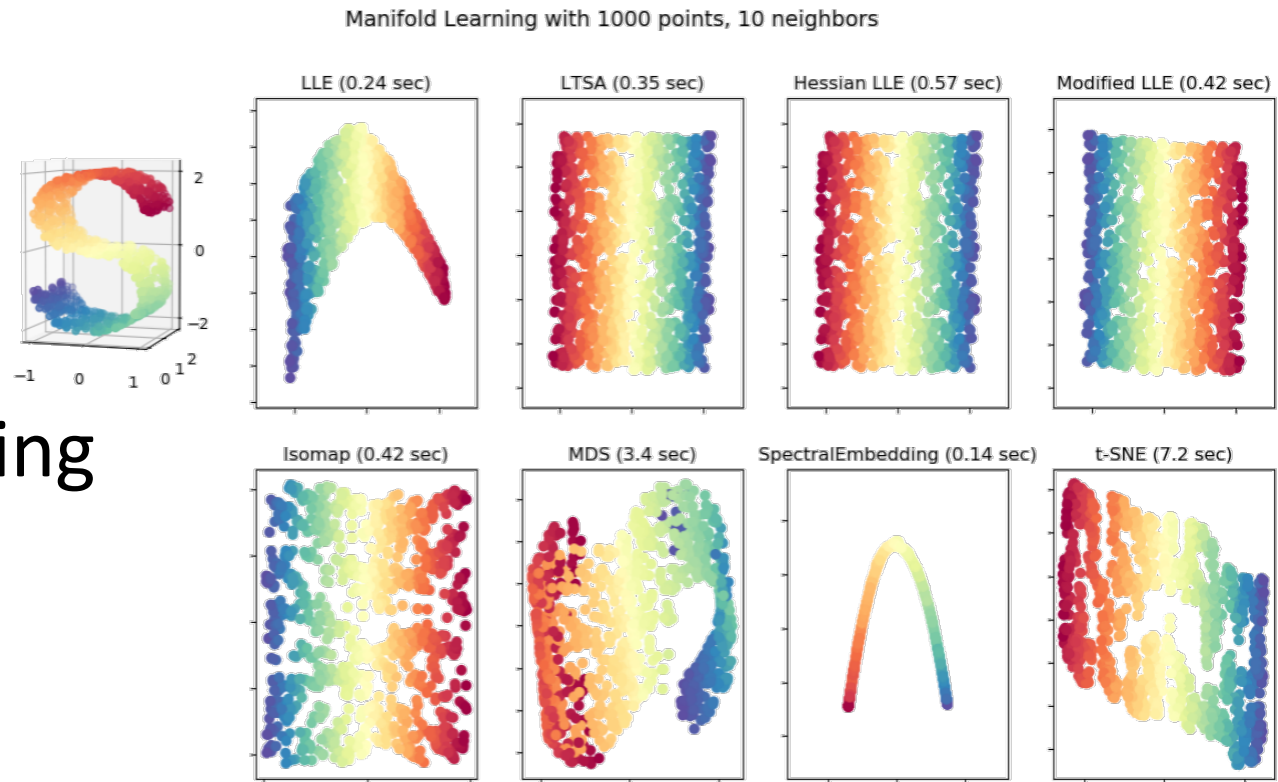
- PCA

- ICA

- Manifold Learning

 - t-SNE

 - LLE



[\[scikit-learn.org\]](https://scikit-learn.org)

Model Explanation

ML model explainability

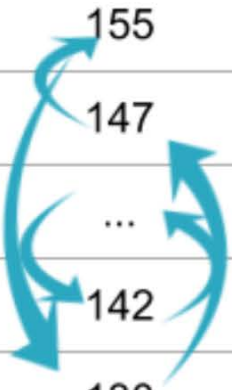
What features have the biggest impact?

- Debugging
- Informing feature engineering
- Directing future data collection
- Informing human decision-making
- Building Trust

Explain: Permutation importance

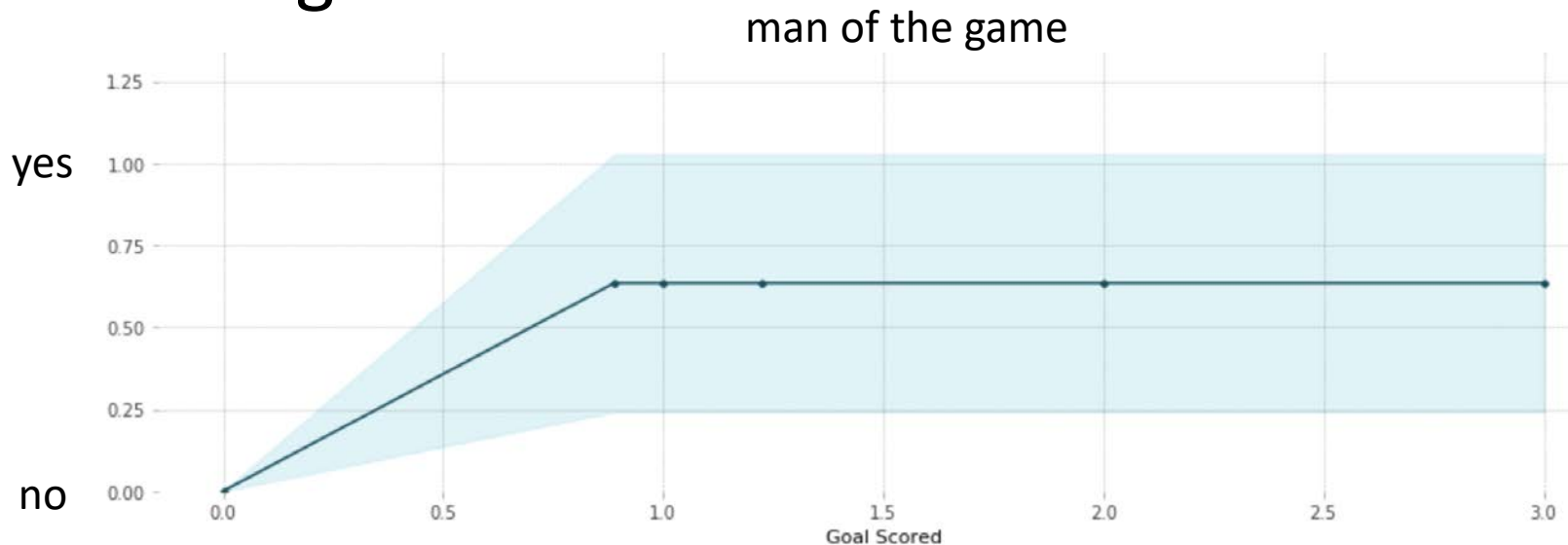
- For a trained model, per column of the data:
 - Randomly shuffle column values
 - Performance deterioration as feature importance

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...
156	142	...	8
153	130	...	24



Explain: Partial dependence plots

- Use trained model
- Show prediction as feature in one row is varied
- Average over all rows



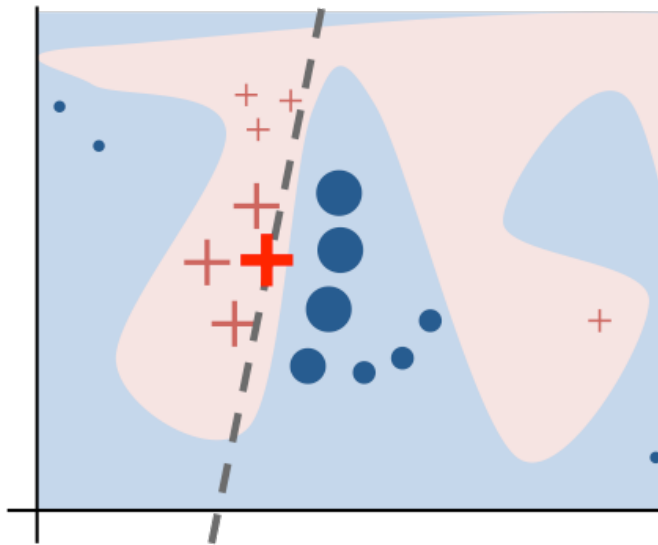
Explain: SHAP values

- How much does a particular prediction change, if a feature is reset to a base value?
- Additive: SUM of (SHAP values for all features) = $prediction - prediction_for_baseline_values$
- “Man of the match” example:



Model explanation via local linear approximation (LIME)

- Lime: Explaining the predictions of any ML classifier
- SHAP values are a generalization of this

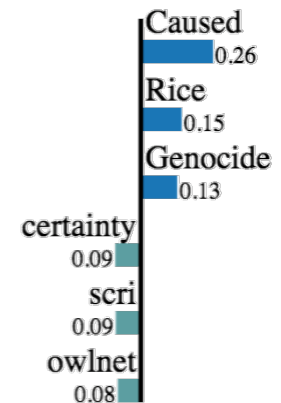


Prediction probabilities

atheism	0.50
christian	0.43
religion.misc	0.05
mideast	0.02
Other	0.00

NOT atheism

atheism



Deep Learning / Neural Networks

Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers

Fred Hohman, *Member, IEEE*, Minsuk Kahng, *Member, IEEE*, Robert Pienta, *Member, IEEE*,
and Duen Horng Chau, *Member, IEEE*

Abstract—Deep learning has recently seen rapid development and received significant attention due to its state-of-the-art performance on previously-thought hard problems. However, because of the internal complexity and nonlinear structure of deep neural networks, the **underlying decision making processes for why these models are achieving such performance are challenging** and [...]

- [[TVCG 2018](#)] [Web version](#)

Interrogative Questions

Interpretability & Explainability
Debugging & Improving Models
Comparing & Selecting Models
Teaching Deep Learning Concepts

WHY

Model Developers & Builders
Model Users
Non-experts

WHO

Computational Graph & Network Architecture
Learned Model Parameters
Individual Computational Units
Neurons in High-dimensional Space
Aggregated Information

WHAT

Node-link Diagrams for Network Architecture
Dimensionality Reduction & Scatter Plots
Line Charts for Temporal Metrics
Instance-based Analysis & Exploration

HOW

Interactive Experimentation
Algorithms for Attribution & Feature Visualization

During Training
After Training

WHEN

Publication Venue

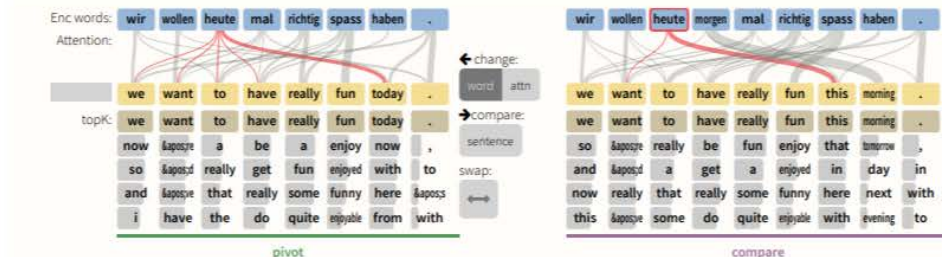
WHERE

[Web version](#)

Sequence model visualization

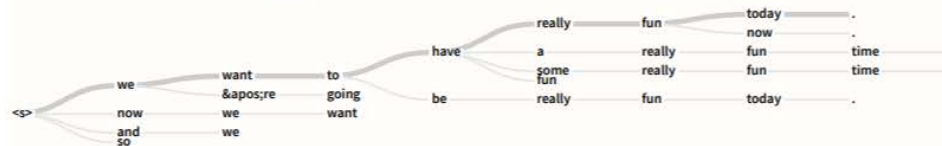
- Examine model decisions
- Connect decisions to previous examples
- Test alternative decisions
- See [[Video](#)] at <https://seq2seq-vis.io/>

Translation View (a)

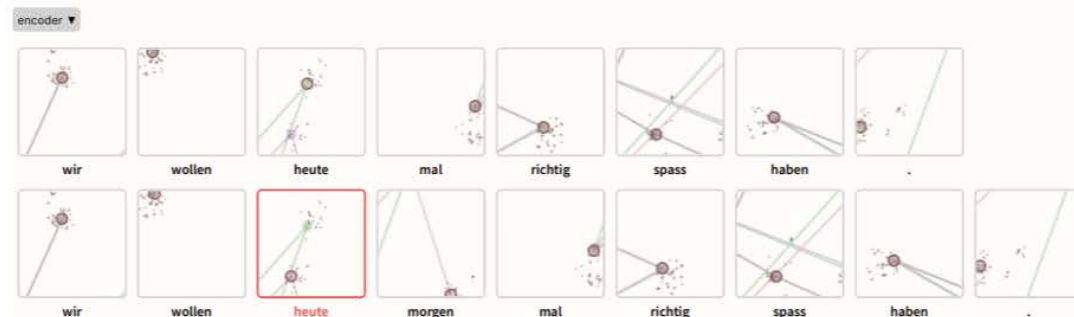


(c) Attention Vis

(d) TopK List

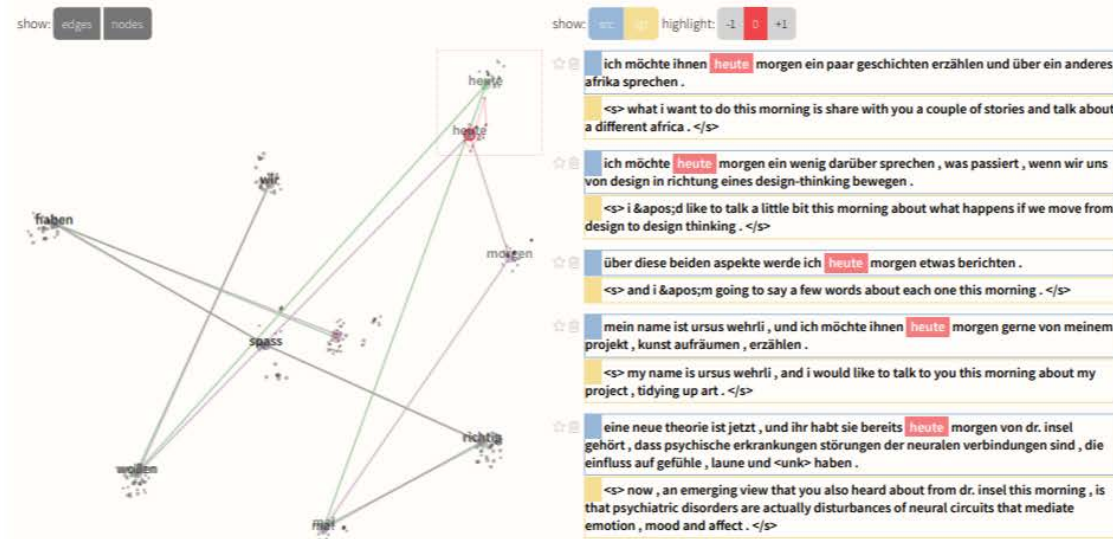


(e) Beamsearch Tree



(f) Trajectory Pictograms

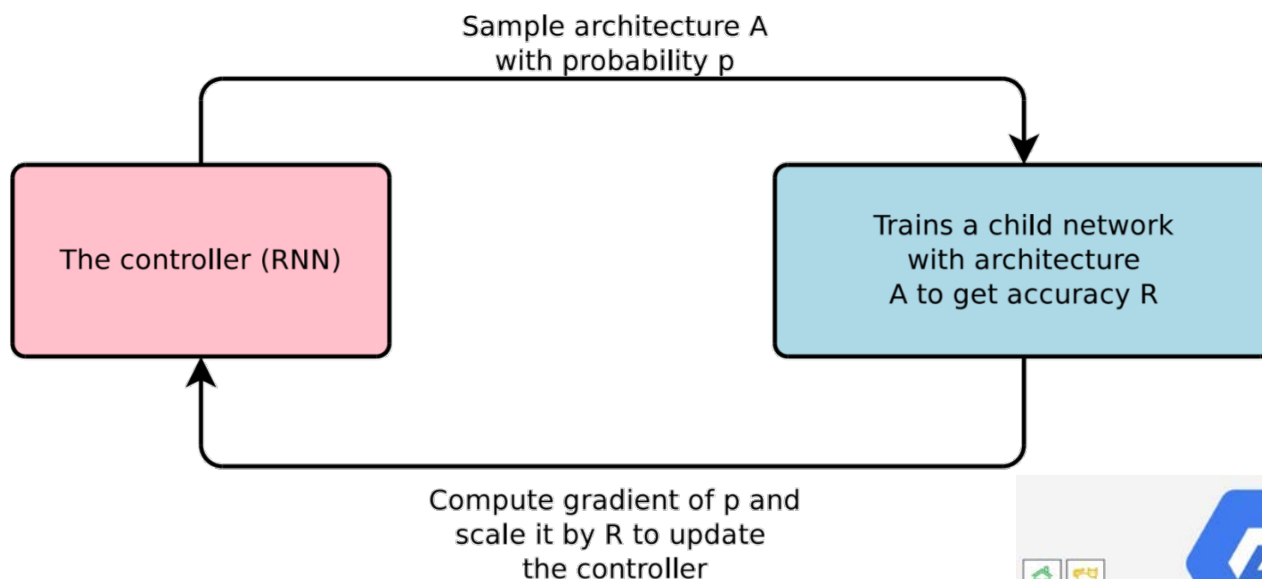
Neighborhood View (b)



(g) left: State Trajectories

(h) right: Neighbor List

AutoML - Neural Architecture Search



- Given a dataset, produce Model
- Serve model output via REST API

How:

- Controller suggests architecture for better R
- Combined with evolution



Further directions

- **Debugging tools**

- [Tensorboard: Visualizing Learning](#)
- [Visdom](#) (only supports (Py)Torch and numpy)

- **Explainables**

- [R2D3](#)
- [Tensorboard Playground](#)

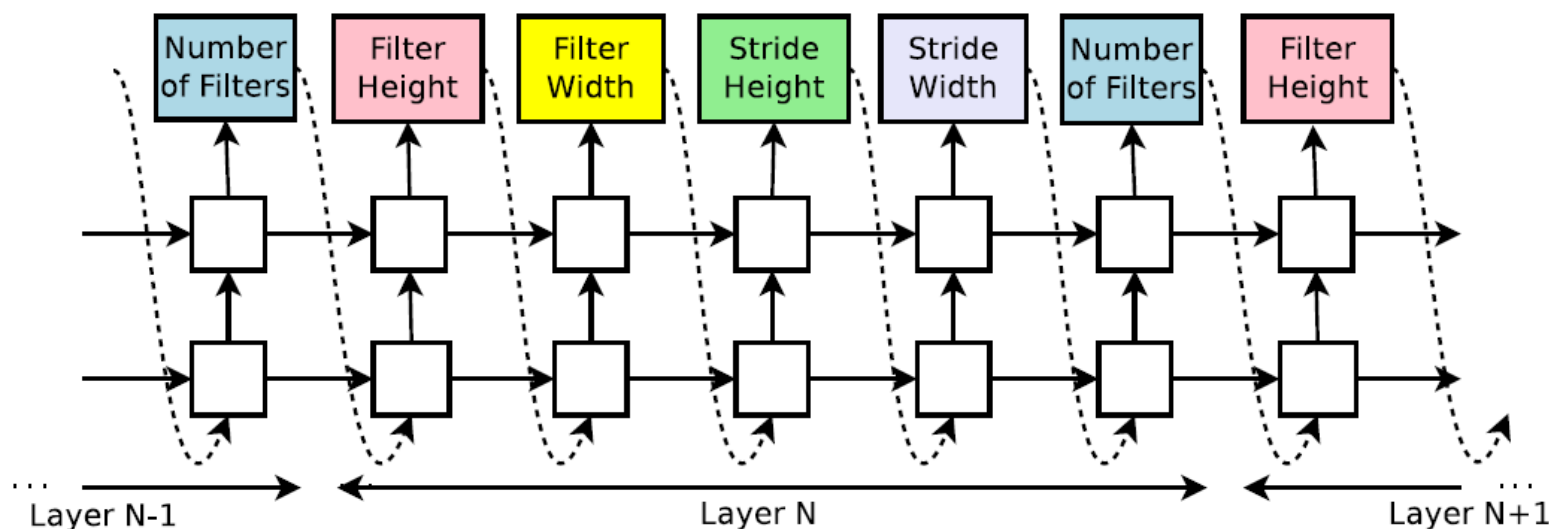
Model visualization

- Seq2Seq Vis:
<http://lstm.seas.harvard.edu/client/index.html>
- [Building blocks of interpretability](#)

Google Compute

AutoML

- Controller is implemented as RNN
- Generates string to define architecture
- Speed-up: Transfer learning for architecture and weights



Google Compute Platform

- Big Query: Cloud data warehouse with ML
 - Beware of the pricing
- Vis via Data Studio (free)
 - <https://cloud.google.com/bigquery/docs/visualize-data-studio>

Google Facets

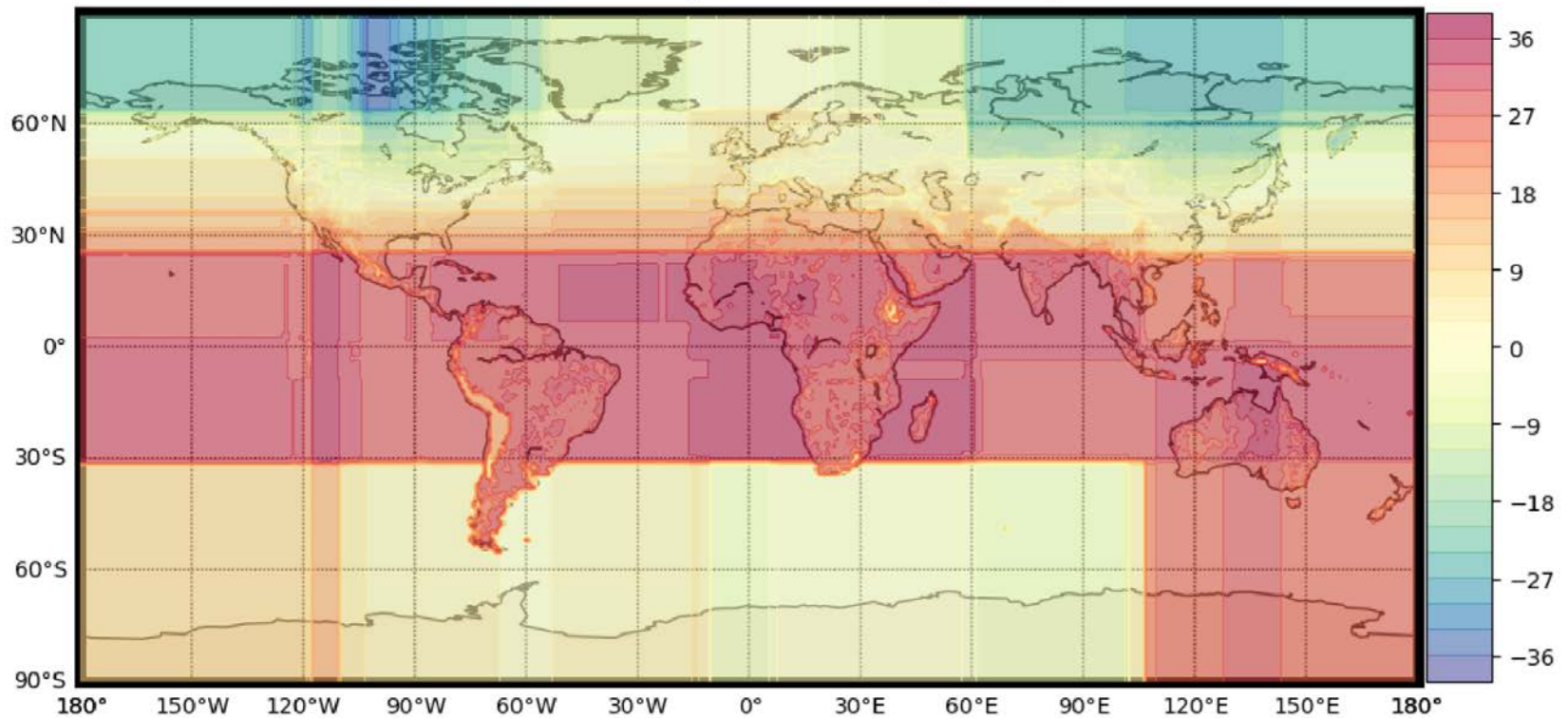
- <https://pair-code.github.io/facets/>
- Application to financial data

Weather Model Visualizations

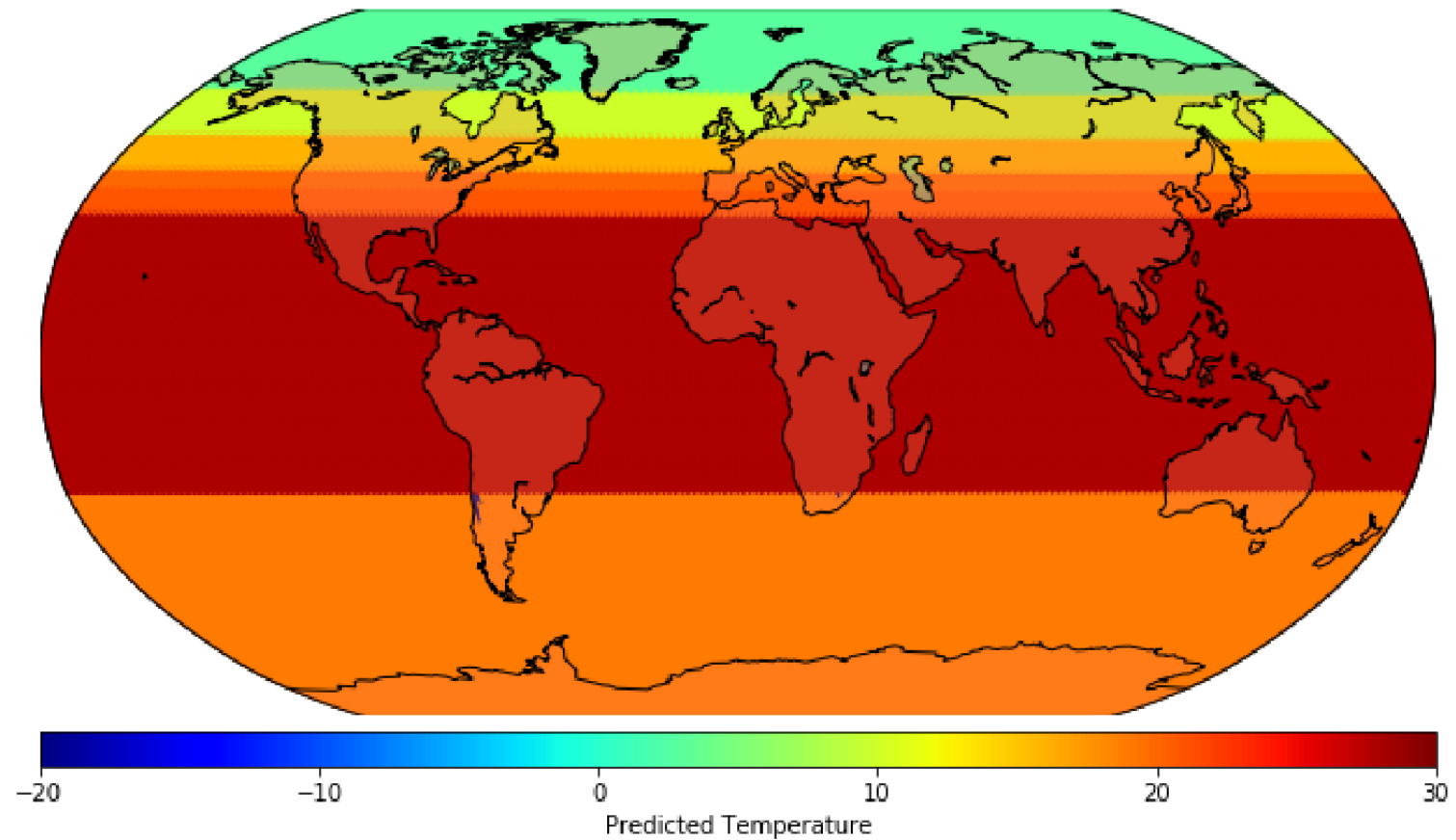
Selected solutions from Assignment 3 – Task 2 B1

Anurag Bejju – GBT Model

Predicted Global Temperatures for January Month

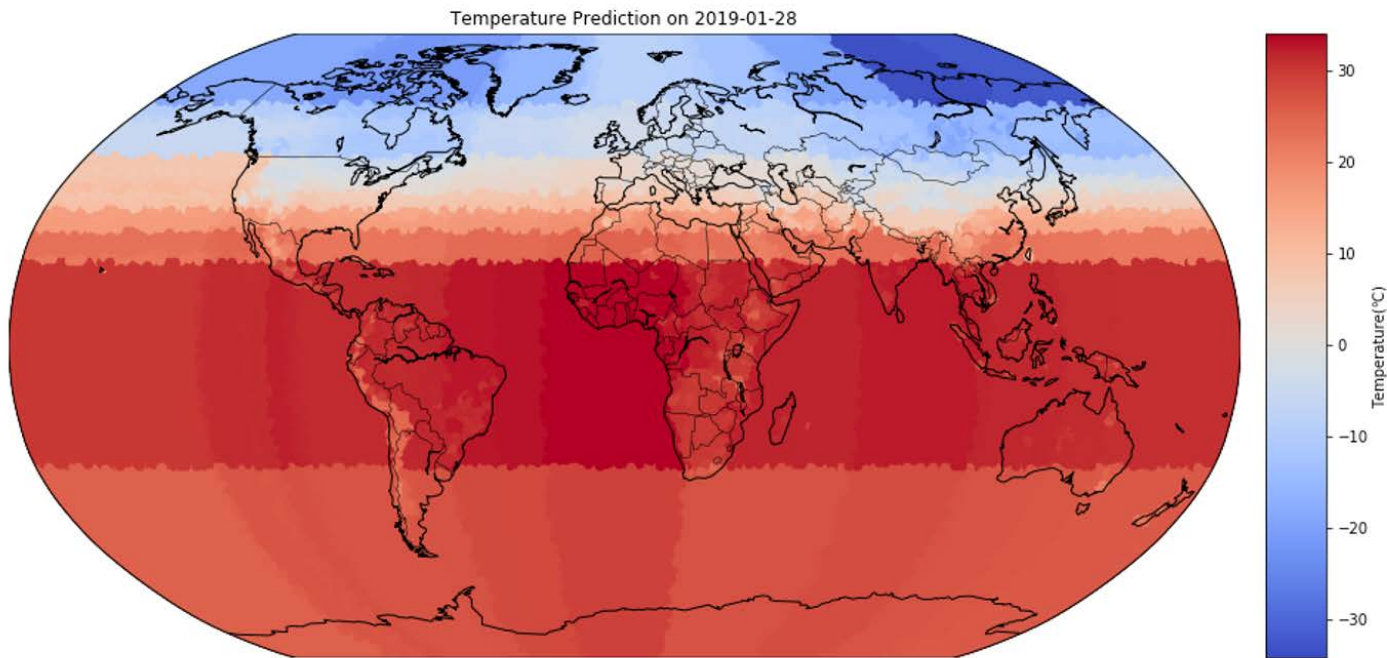


Aisuluu Alymbekova



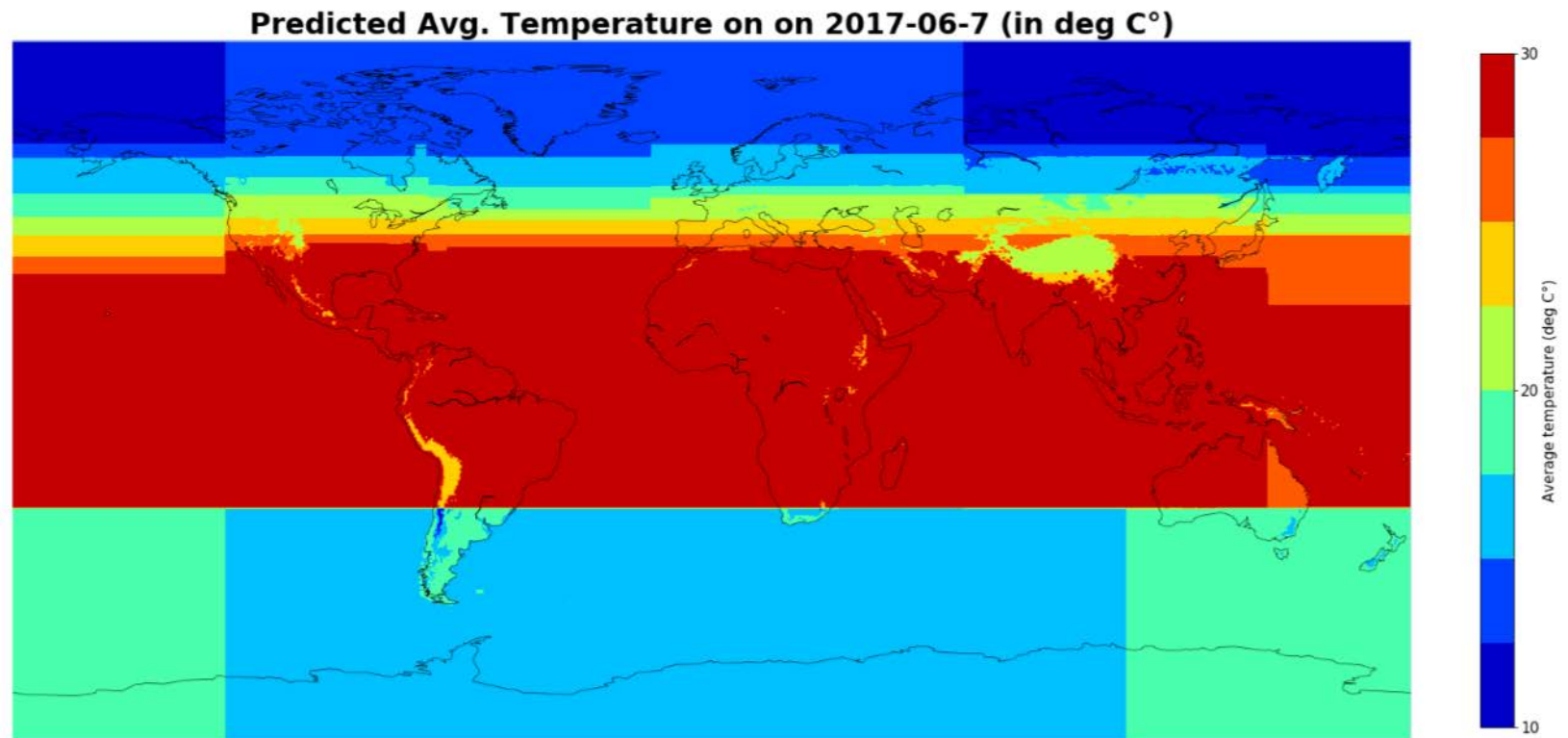
Andong (Anton) Ma

Fg3 is the dense heat-map generated by my model. Which we can easily evaluate as not good, because it shows that the temperature of southern earth is all above 0 C°, even in the South Pole.

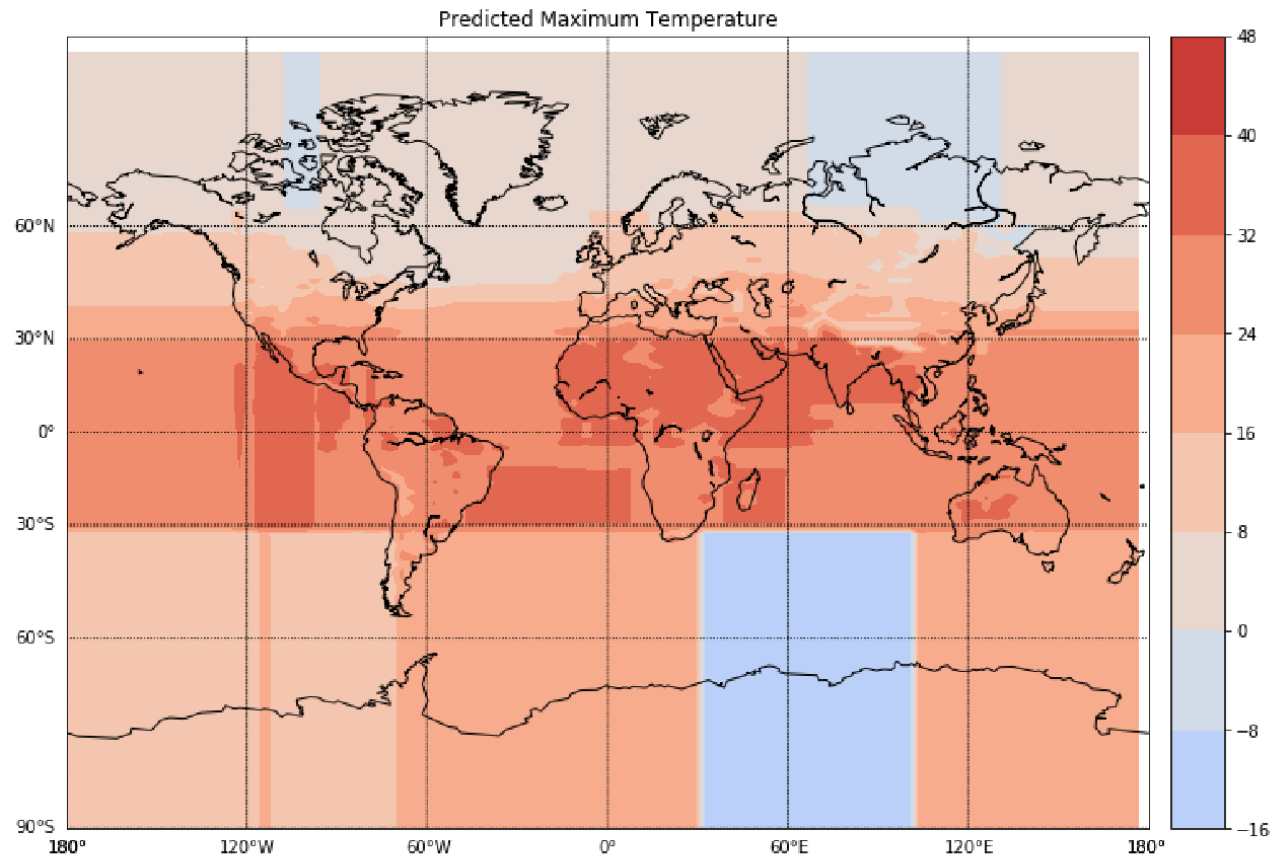


Fg 3

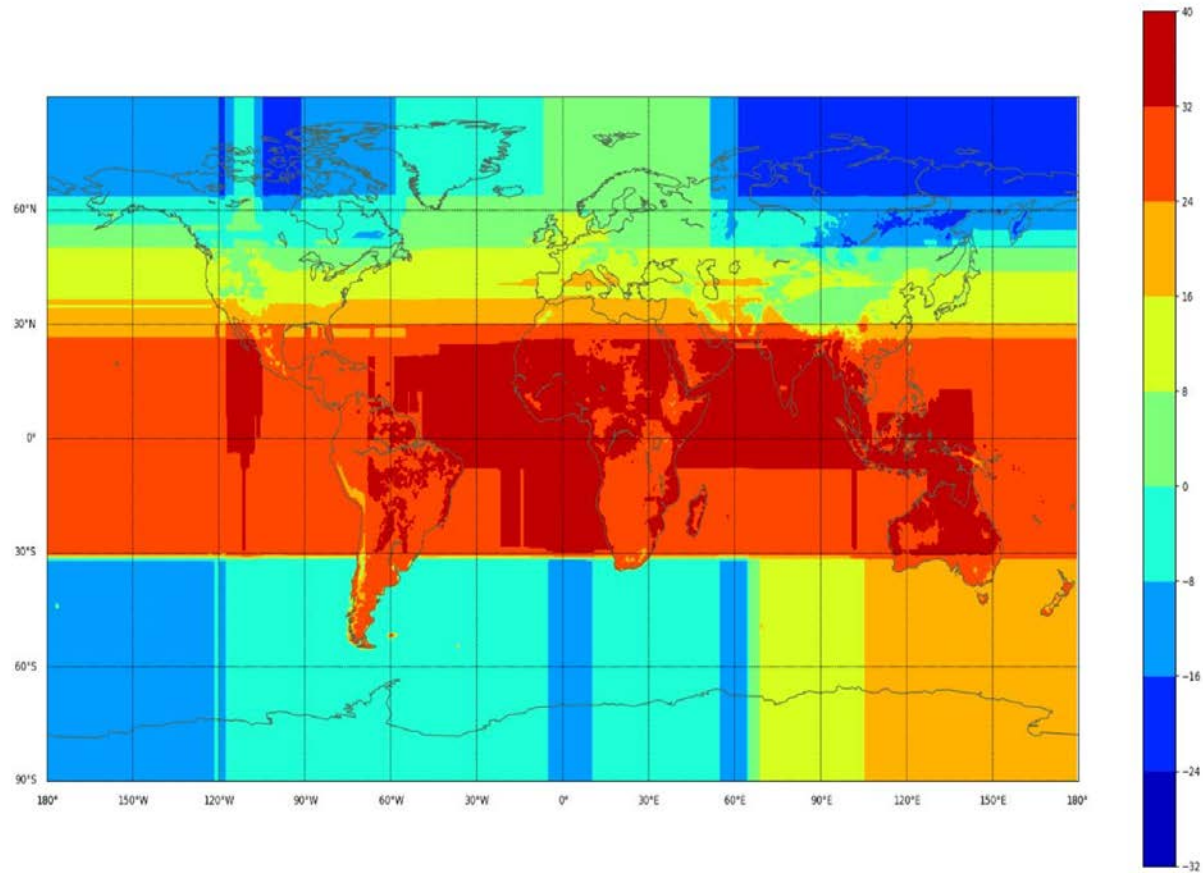
Abhishek Sunnak



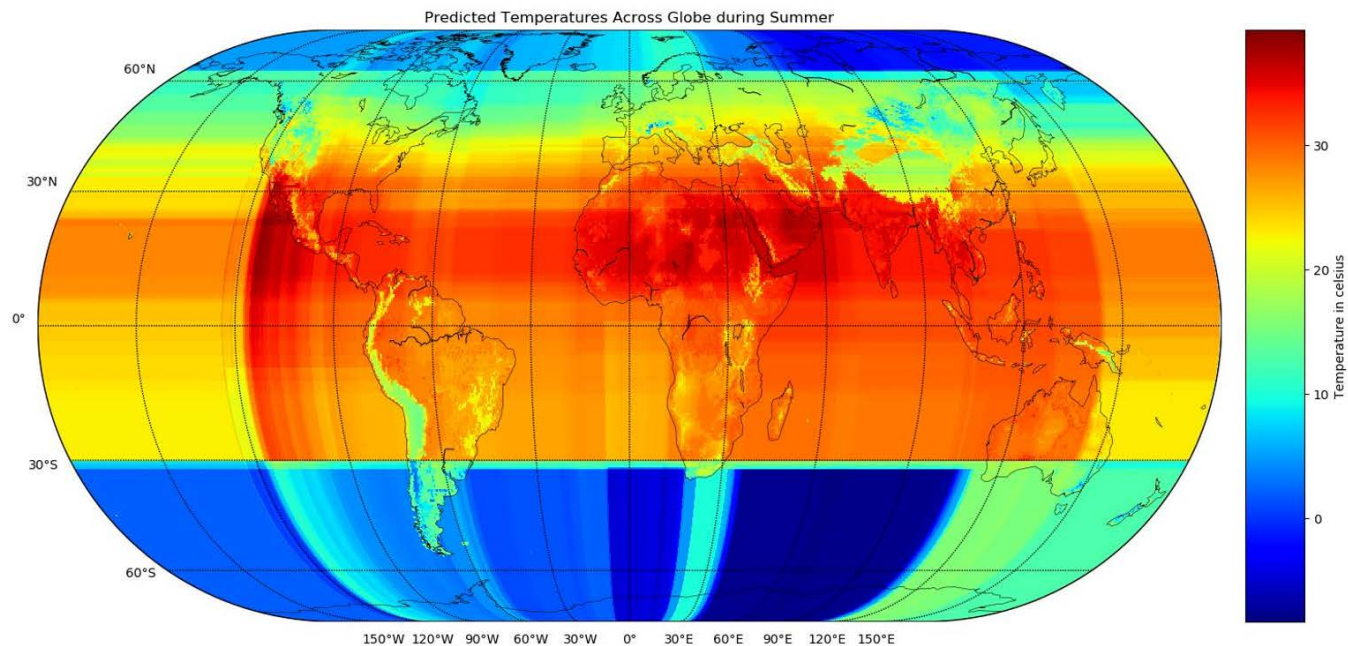
Andrew Wesson – Decision Tree



Manjur Patowary

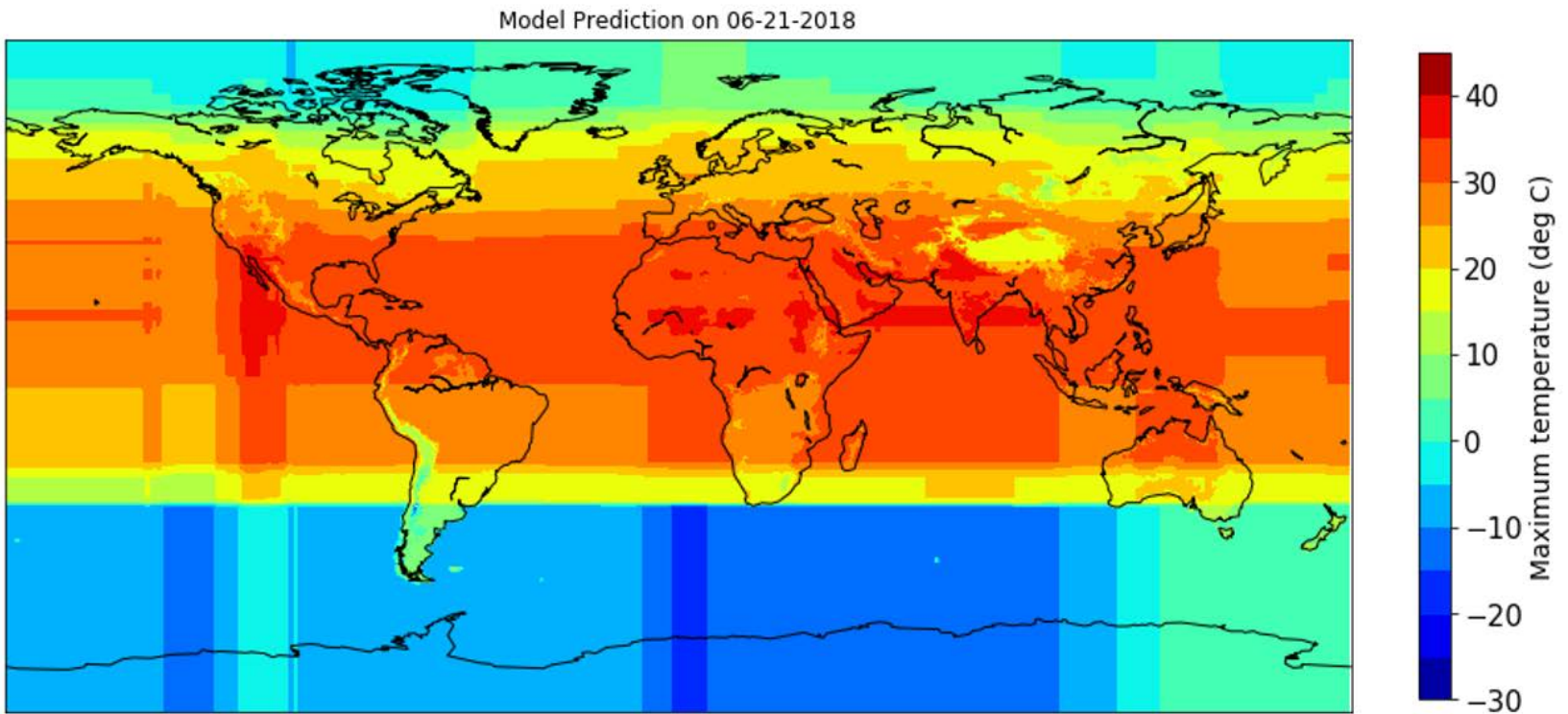


Venkata Sai Pavan Kumar Kosaraju

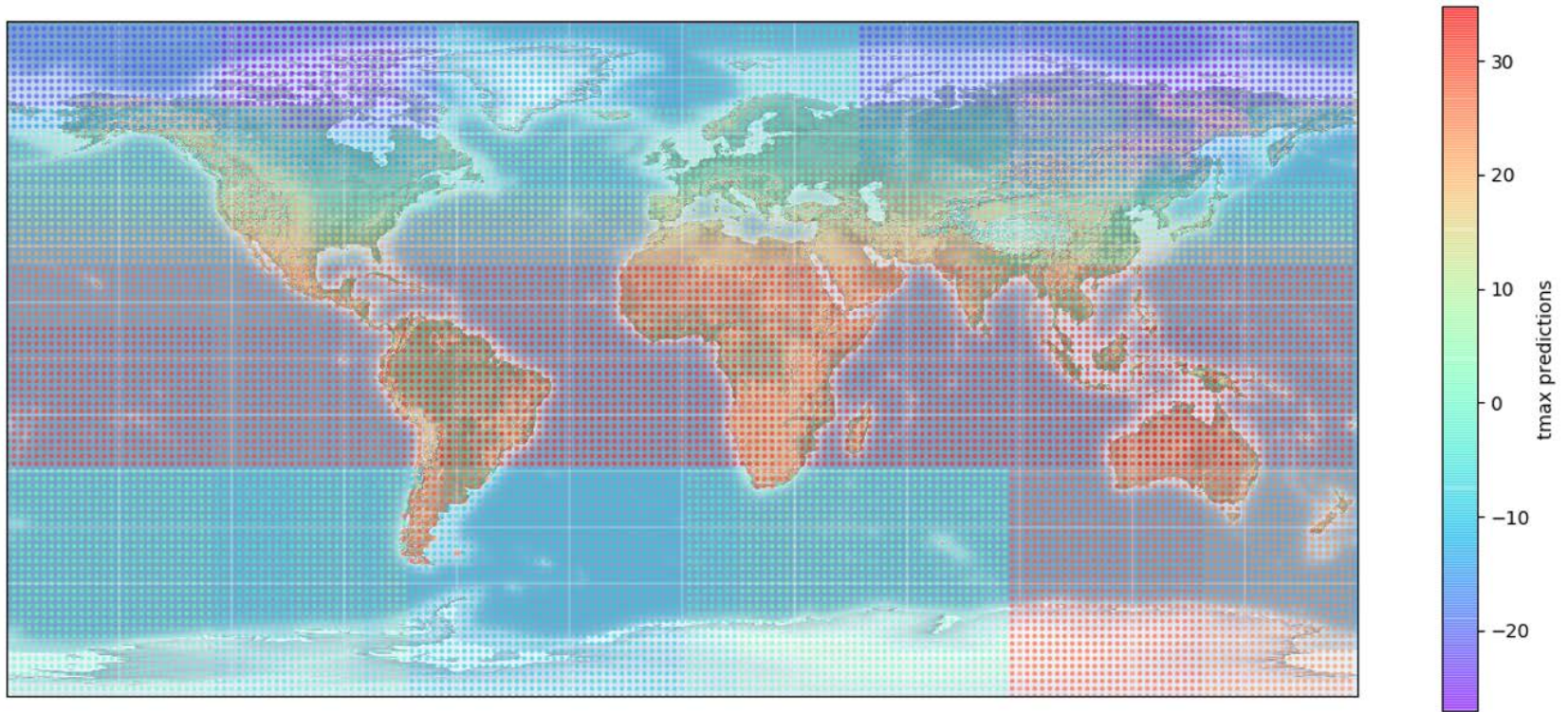


Interpretation: Predicted temperatures on 27th May 2019 across globe by the model. Temperatures may reach as high as 35 degrees near equator region while in Northern and Southern hemisphere it will be cool.

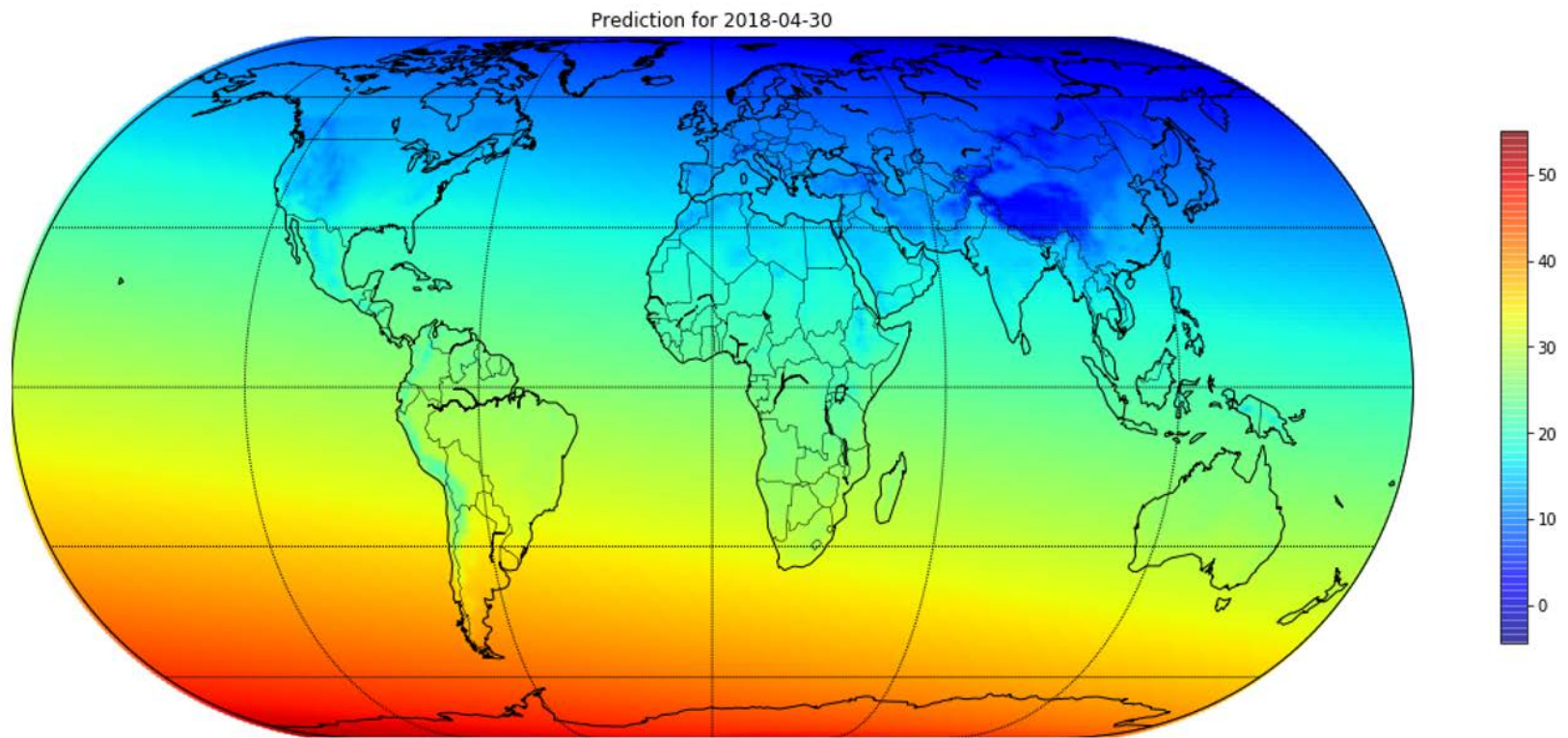
Prashanth Rao - Great discussion!



Neda Zolaktaf

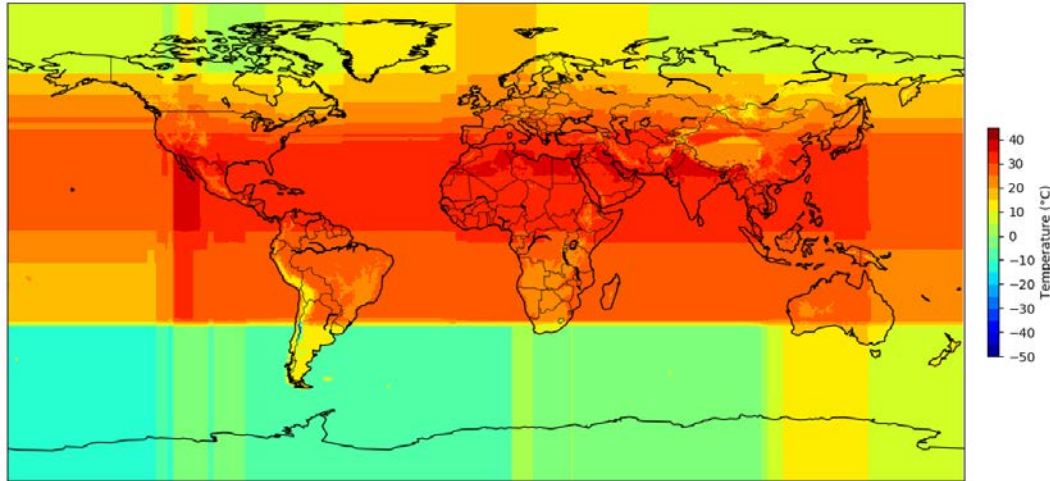


Linear Model?

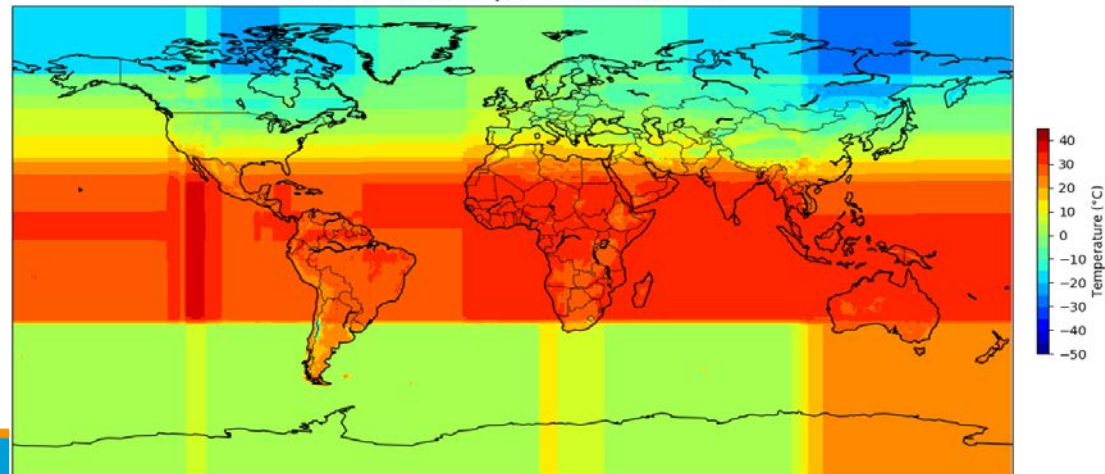


Oluwaseyi (Seyi) Talabi

Predicted Dense Temperatures for 2018-08-28



Predicted Dense Temperatures for 2019-01-28



Thank you for your attention!

