Modelling Species Distribution in the Great Smoky Mountains National Park

Lonnie Yu

Agenda

1. Background
2. Research Problems
3. Maximum Entropy Modeling
4. Legacy Data Pipeline
5. New Data Pipeline
6. Live Demo
7. Future Work

Background

- Species Occurrence Data from the Great Smoky Mountains National Park (GRSM)
  - Ecological and Public Research
- Species Distribution Modelling
  - Originally sequential runs
  - Scaled via High Performance Computing in Legacy Pipeline
- New Browser-based Species Mapping
  - Leaflet.js Map Application

Research Problems

1. Learning GRSM's requirements and recommending solutions (consulting)
2. Choosing the maximum number of simultaneous species/colors displayed (data visualization)
3. Learning legacy/new data pipeline (ML, data engineering)
4. Automating new data pipeline (data engineering)
5. Designing Front-end UI/UX (collaboration with UI/UX professor)
   - Researchers
   - General Public (e.g. 8th grade students)

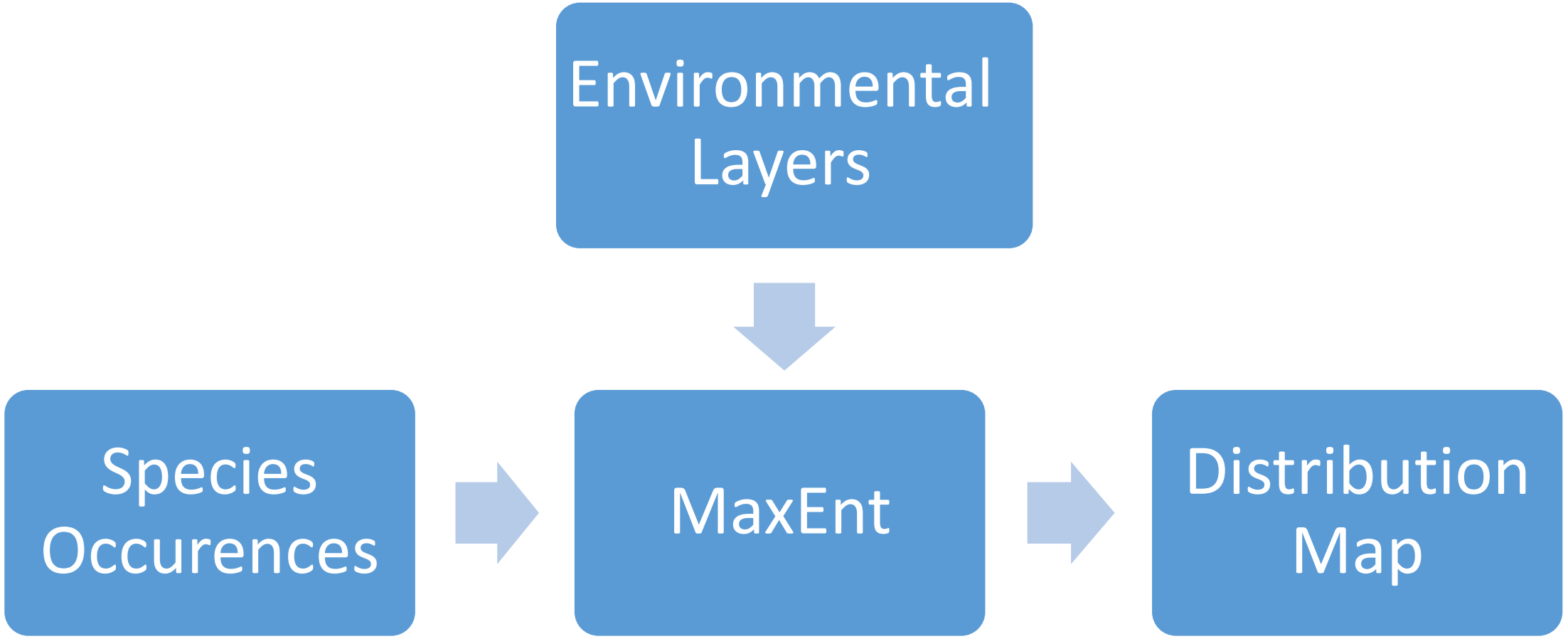
Maximum Entropy Modeling

- “A Maximum Entropy Approach to Species Distribution Modeling” (ICML 2004)
  - Skewed training data: too many positives, not few negatives.
  - “… estimate the target distribution by finding the distribution of maximum entropy … subject to the constraint that the expected value of each feature under this estimated distribution matches its empirical average.”
  - http://rob.schapire.net/papers/maxent_icml.pdf
- Maxent.jar (Java)
  - http://biodiversityinformatics.amnh.org/open_source/maxent/Maxent_tutorial2017.pdf

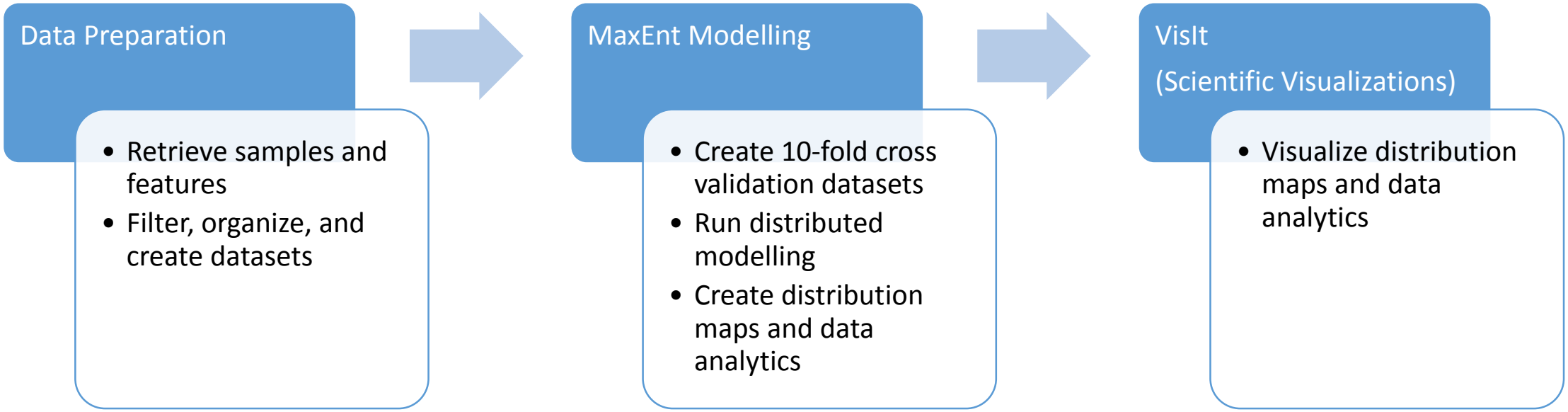
Maximum Entropy Modeling

- Samples: Species Occurrences
  - CSV format: “species name, longitude, latitude”
  - (>= 30) occurrences x ~900 species
- Features: Environmental Layers
  - ASCII Raster Grid format
  - Continuous or Categorical
  - ~50 environmental layers
- Output: Distribution Map
  - ASCII Raster Grid format
  - Pixel color corresponds to predicted distribution value

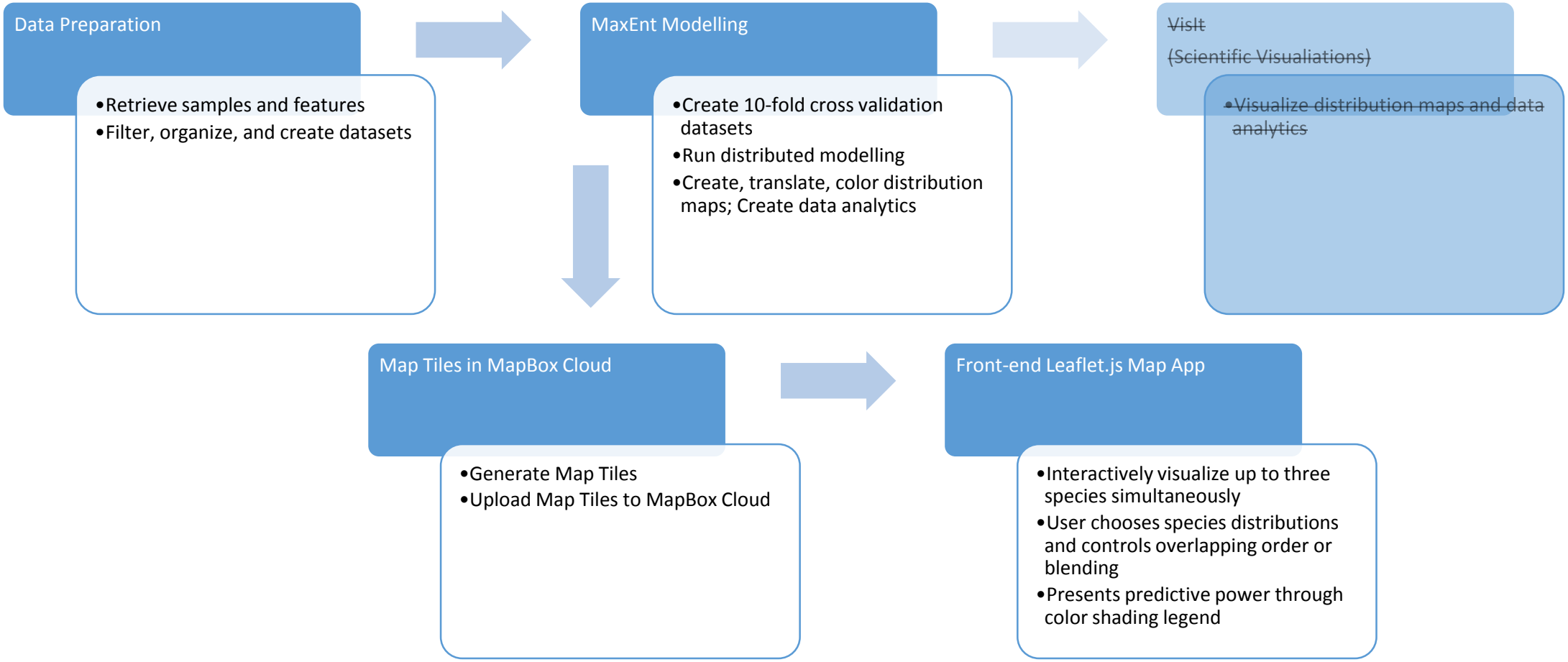
MaxEnt Workflow



Legacy Data Pipeline



New Data Pipeline



Live Demo

- https://science.nature.nps.gov/parks/grsm/species/

Future Work

- NPS Direct Control of Data Pipeline
- Tracking Historical Data
  - Analyzing/Forecasting Trends
- Public Mobile App