Rapid Microbe Identification

A Machine Learning proof-of-concept

People love pets

- We don’t want our furry friends to be sick
- An average American spends $687 per year on their cat and over $1,200 on their dog
- Primary costs are food, followed by veterinary care

Speed of Identification

- Microorganisms can replicate very quickly
- Identification takes time
- Manual identification is subjective
- Identification drives treatment

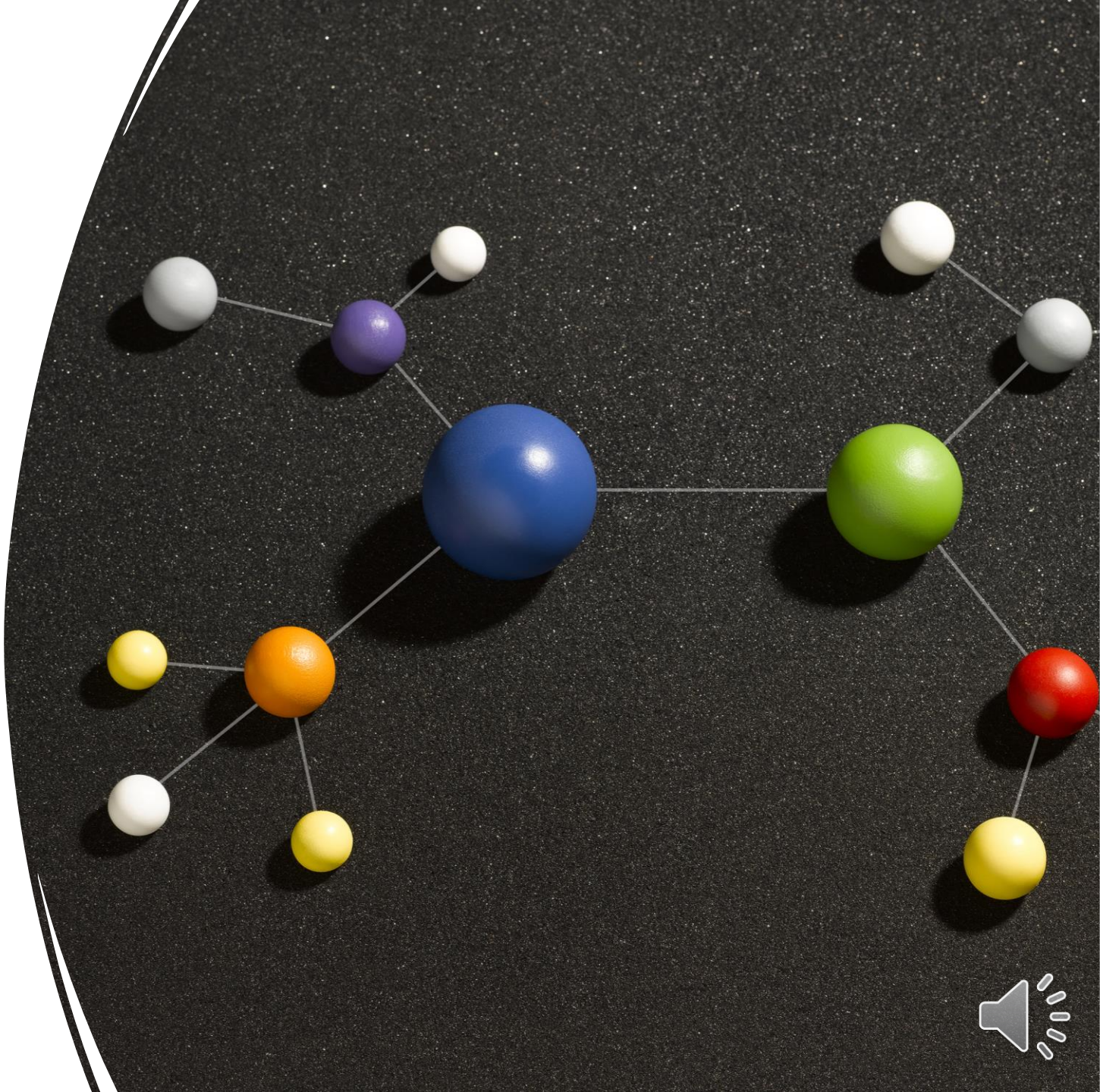
Microbes Dataset (Kaggle)

- 23 numeric features
  - Measurements generated from microscopic imagery
  - Many skewed distributions
- Target: names of organisms
  - 10 classes
  - Imbalanced
- Are the samples representative?
- Optimal number of features

https://upload.wikimedia.org/wikipedia/commons/f/f1/Diatoms.jpg

Comparing algorithms

- Best algorithms:
  - RandomForestClassifier - 98.2%
  - DecisionTreeClassifier - 98.0%
  - GradientBoostingClassifier - 87.5%
- Worst algorithms:
  - GaussianNB (Naïve Bayes) - 33.0%
  - AdaBoostClassifier - 30.2%
- After tuning:
  - RandomForestClassifier - 98.7%
  - DecisionTreeClassifier - 98.7%



Conclusion and next steps

- Selected model:
  - Decision Tree Classifier
- Recommendations:
  - Supplement data set
  - Expand to additional microbes
  - Consider new measurements
  - Train a neural network directly on microscope images