Rensselaer

why not change the world?®

Data Mining I

Ahmed Eleish

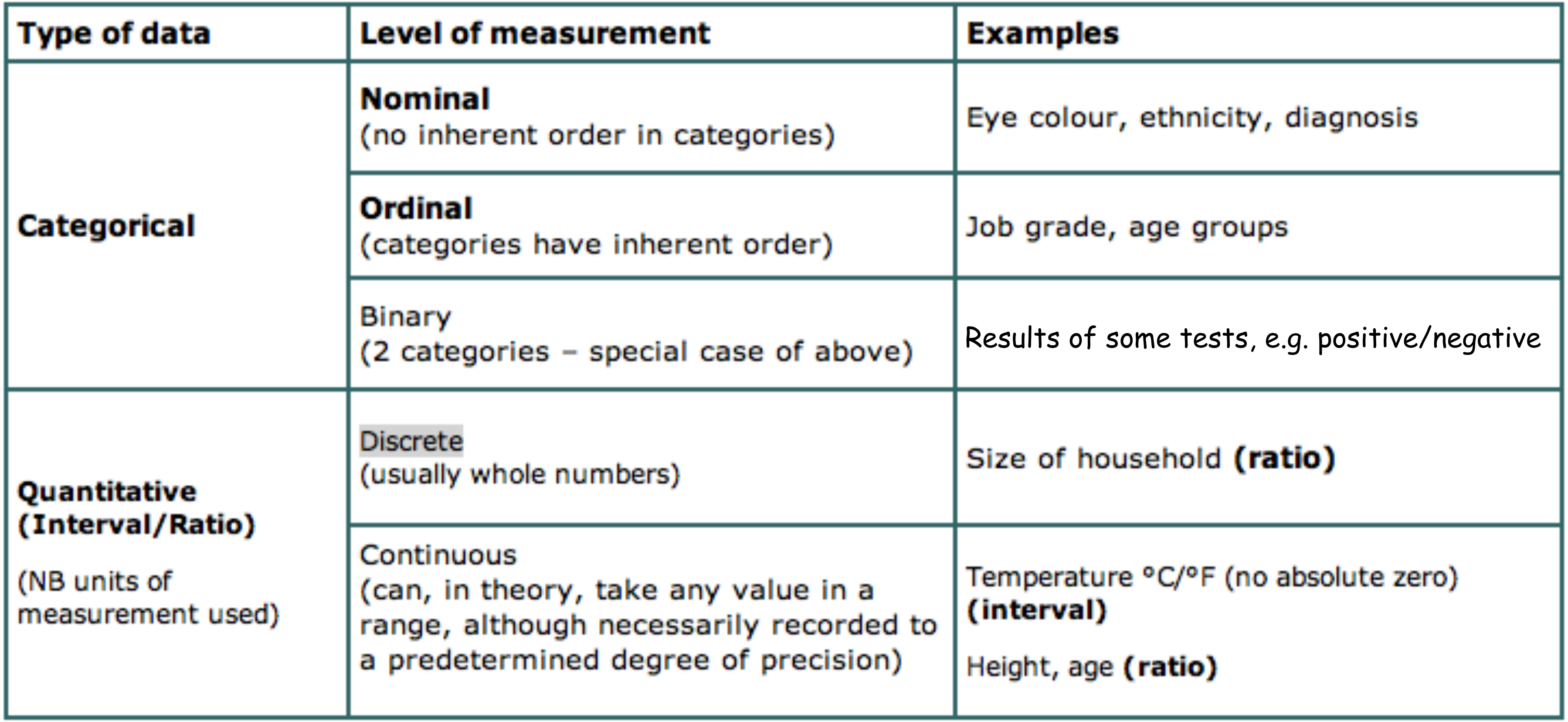
Data Science – ITWS/CSCI/ERTH-4350/6350 Module 4, October 10th, 2024

Tetherless World Constellation
Rensselaer Polytechnic Institute

Contents

- Data Mining what it is, is not, types
- Distributed applications – modern data mining
- Science example(s)
- A specific toolkit and two examples
    - Classifier
    - Image analysis – clouds
- Week 9 reading – note is PRE-READING (only two articles)

Types of Data



Type of data	**Level of measurement**	**Examples**
Categorical	**Nominal** (no inherent order in categories)	Eye colour, ethnicity, diagnosis
	Ordinal (categories have inherent order)	Job grade, age groups
	Binary (2 categories – special case of above)	Results of some tests, e.g. positive/negative
Quantitative (Interval/Ratio) (NB units of measurement used)	Discrete (usually whole numbers)	Size of household **(ratio)**
	Continuous (can, in theory, take any value in a range, although necessarily recorded to a predetermined degree of precision)	Temperature °C/°F (no absolute zero) **(interval)** Height, age **(ratio)**

Data Mining - What it is

- Extracting knowledge from large amounts of data
- Motivation
  - Our ability to collect data has expanded rapidly
  - It is impossible to analyze all of the data manually
  - Data contains valuable information that can aid in decision making
- Uses techniques from:
  - Pattern Recognition
  - Machine Learning
  - Statistics
  - High Performance Database Systems
  - OLAP (Online Analytical Processing) i.e Financial systems, sensor systems related to weather data.
- Data mining methods must be efficient and scalable (8~10 years ago, data mining could not be done on your Laptop).

Data Mining – What it isn’t

- Small Scale
    - Data mining methods are designed for large data sets
    - Scale is one of the characteristics that distinguishes data mining applications from traditional machine learning applications
- Foolproof
    - Data mining techniques will discover patterns in any data
    - The patterns discovered may be meaningless
    - It is up to the user to determine how to interpret the results
    - “Make it foolproof and they’ll just invent a better fool”
- Magic
    - Data mining techniques cannot generate information that is not present in the data
    - They can only find the patterns that are already there

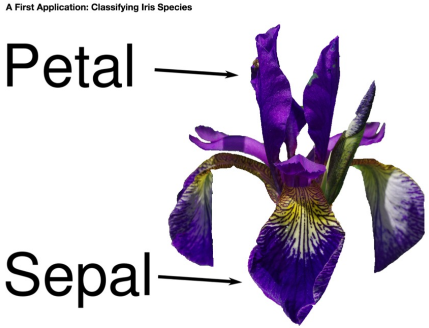
Data Mining - Types of Mining

- Classification (Supervised Learning)
  - Classifiers are created using labeled training samples
  - Training samples created by ground truth / experts
  - Classifier later used to classify unknown samples
- Clustering (Unsupervised Learning)
  - Grouping objects into classes so that similar objects are in the same class and dissimilar objects are in different classes
  - Discover overall distribution patterns and relationships between attributes
- Association Rule Mining
  - Initially developed for market basket analysis
  - Goal is to discover relationships between attributes
  - Uses include decision support, classification and clustering
- Other Types of Mining
  - Outlier Analysis
  - Concept / Class Description
  - Time Series Analysis

Science Motivation: Botany

- Classifying Iris Species
- Let’s assume a botanist is interested in distinguishing the species of some iris flowers that she has found. She has collected some measurements associated with each iris: length and width of the petals and length and width of sepals.
- She also has the measurements of some irises that have been previously identified by an expert botanist as belonging to the species
    - Setosa
    - Versicolor
    - Virginica

For these measurements, she can be certain of which species each iris belongs to.



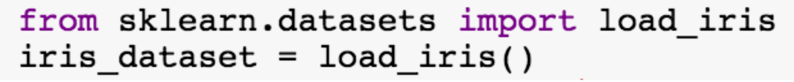
Our goal is to build a machine learning model that can

Classification Problem..

- Because we have measurements for which we know the correct species of iris, this is a *supervised* learning problem.
- In this problem, we want to predict one of several options (species of iris). This is an example of “*Classification Problem*”.
- The possible outputs (different species of irises) are called “*Classes*”.
- Every iris in the dataset belongs to one of three classes, so the problem is “*three-class classification problem*”.
- Desired output for a single data point (an iris) is the species of this flower.
- For a particular data point, the species it belongs to is called its “*label*”.

Meet the Data (Iris dataset)

- Data we will use for this example is Iris dataset (a famous dataset in machine learning and statistics ☺)
- It is included in “*Scikit-Learn*” in the datasets.
- We can load the dataset by using the load_iris function.



```
from sklearn.datasets import load_iris
iris_dataset = load_iris()
```

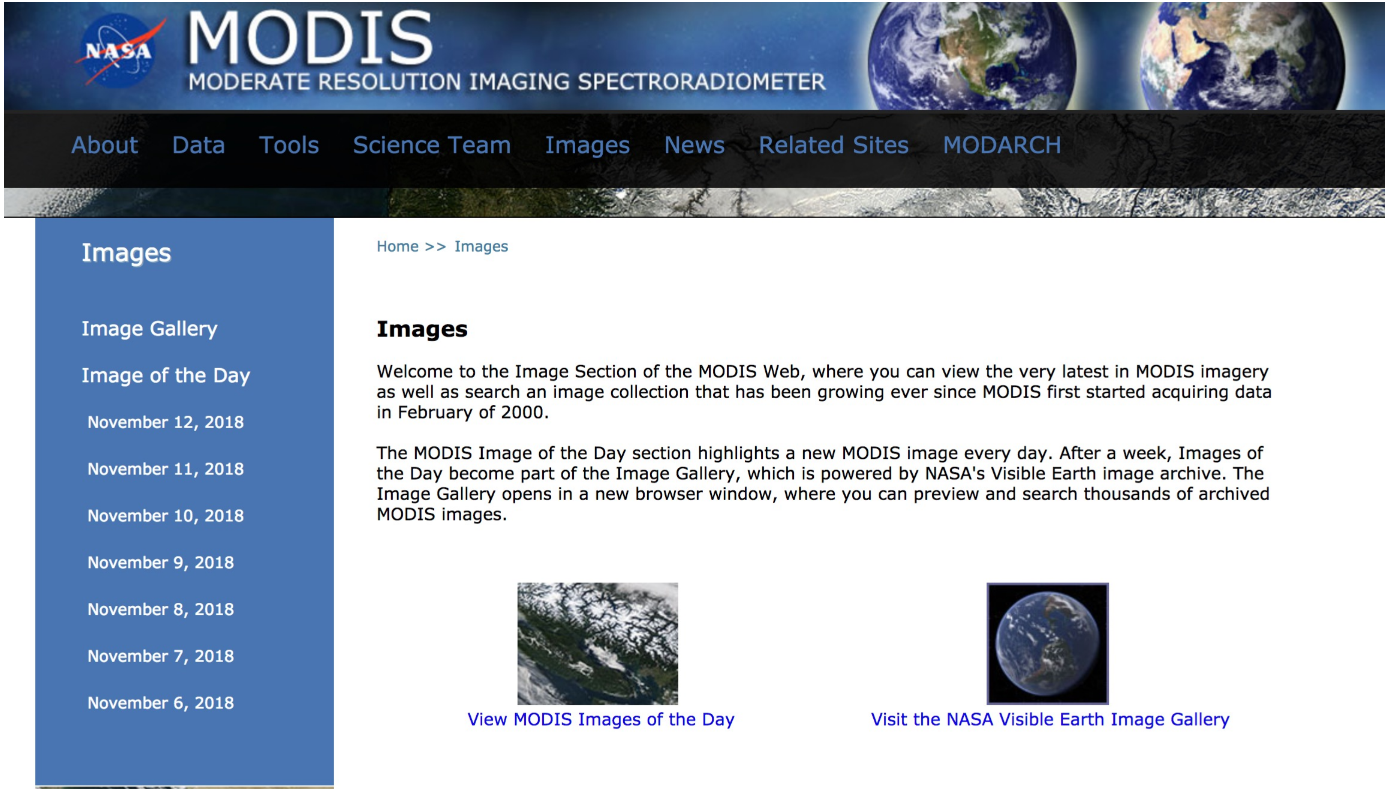
- The iris object that is returned by load_iris is a Bunch object, which is very similar to a dictionary, it contains Keys and values.

http://bit.ly/datascience2018-class8-code-examples

Science Motivation: Health

- Breast Cancer Prediction
- Wisconsin breast cancer dataset includes clinical measurements of breast cancer tumors.
- Each tumor is labeled
  - "benign" (for harmless tumors) or
  - "malignant" (for cancerous tumors)
- The task is to learn to predict whether a given tumor is malignant based on the measurements of the tissues.

Moderate Resolution Imaging Spectroradiometer



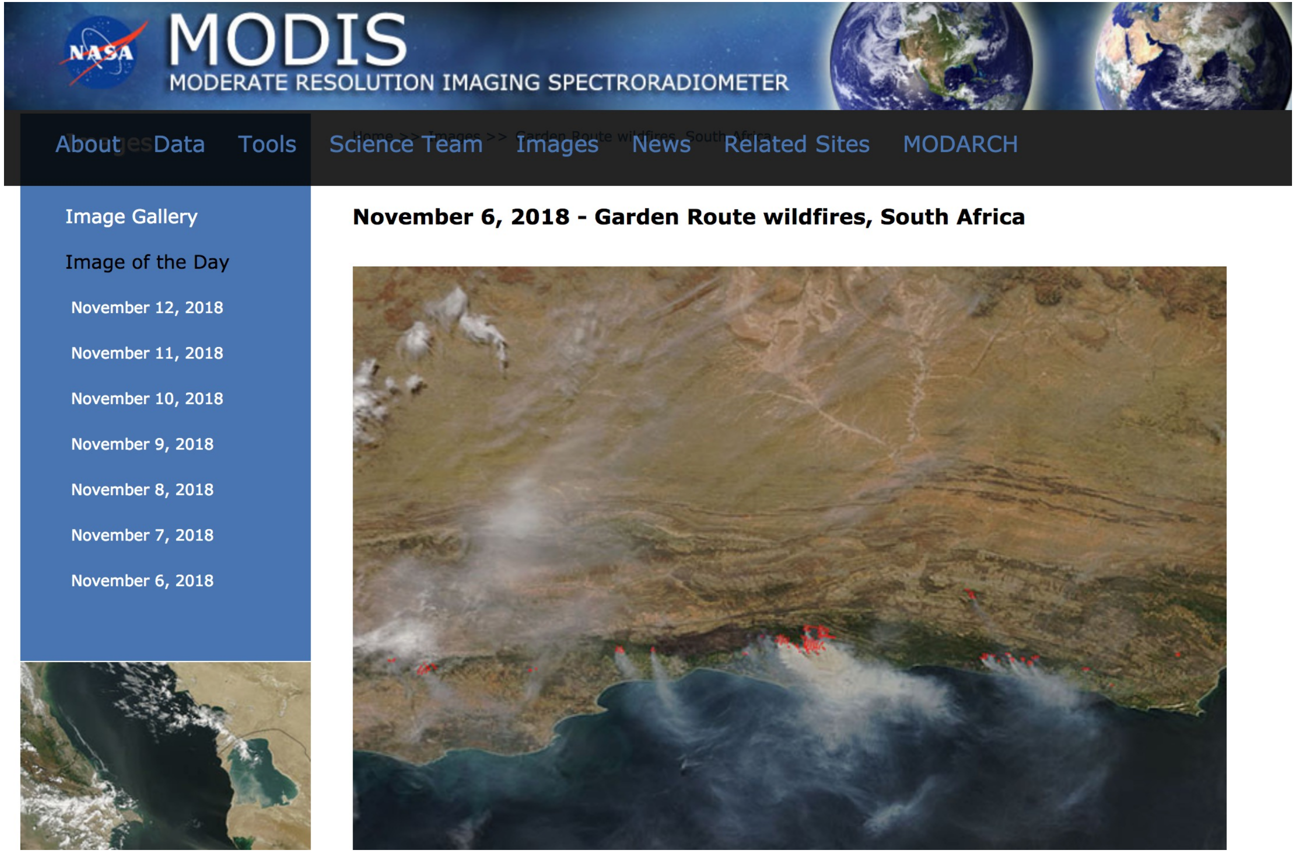
Images

Welcome to the Image Section of the MODIS Web, where you can view the very latest in MODIS imagery as well as search an image collection that has been growing ever since MODIS first started acquiring data in February of 2000.

The MODIS Image of the Day section highlights a new MODIS image every day. After a week, Images of the Day become part of the Image Gallery, which is powered by NASA's Visible Earth image archive. The Image Gallery opens in a new browser window, where you can preview and search thousands of archived MODIS images.

https://modis.gsfc.nasa.gov/gallery/

Moderate Resolution Imaging Spectroradiometer



November 6, 2018 - Garden Route wildfires, South Africa

https://modis.gsfc.nasa.gov//gallery/individual.php?db_date=2018-11-06

Science Motivation

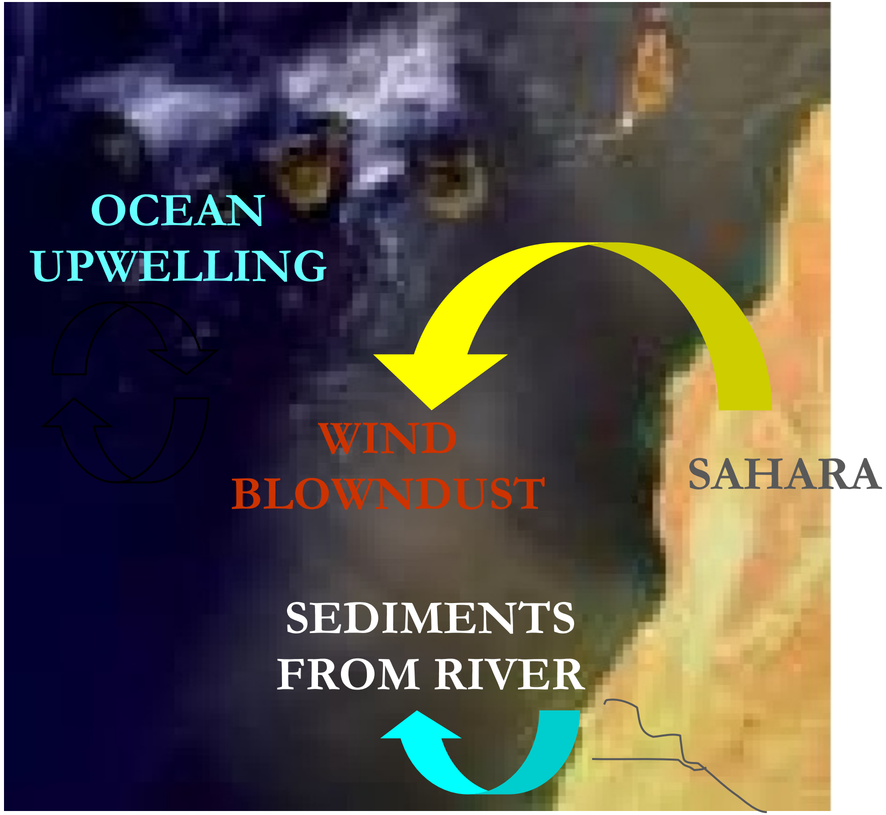
- Study the impact of natural iron fertilization process (such as a dust storm) on plankton growth and subsequent dimethyl sulfide (DMS) production
  - Plankton plays an important role in the carbon cycle
  - Plankton growth is strongly influenced by nutrient availability (Fe/Ph)
  - Dust deposition is important source of **Fe** over ocean
  - Satellite data is an effective tool for monitoring the effects of dust fertilization

Hypotheses

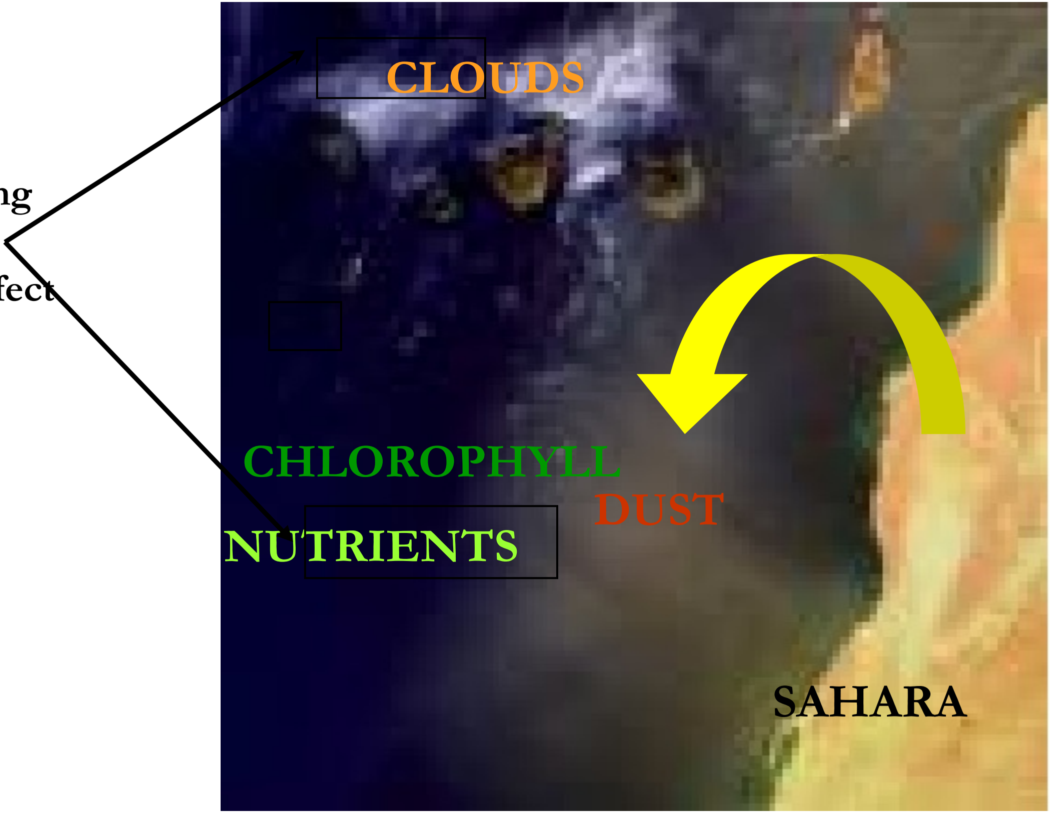
- In remote ocean locations there is a positive correlation between the area averaged atmospheric aerosol loading and oceanic chlorophyll concentration

- There is a time lag between oceanic dust deposition and the photosynthetic activity

Primary source of ocean nutrients



Factors modulating dust-ocean photosynthetic effect



Objectives

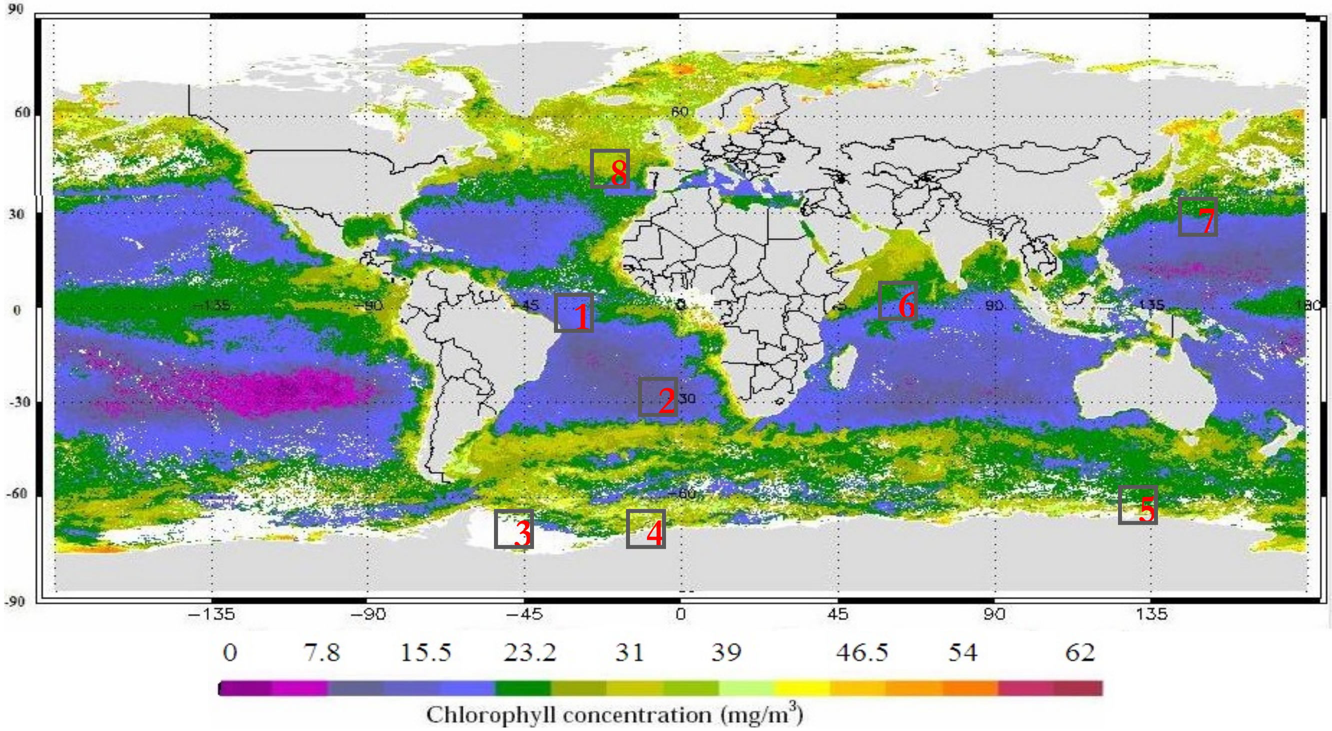
- Use satellite data to determine, if atmospheric dust loading and phytoplankton photosynthetic activity are correlated.
- Determine physical processes responsible for observed relationship

Data and Method

- Data sets obtained from two instruments: SeaWiFS and MODIS during 2000 – 2006 are employed

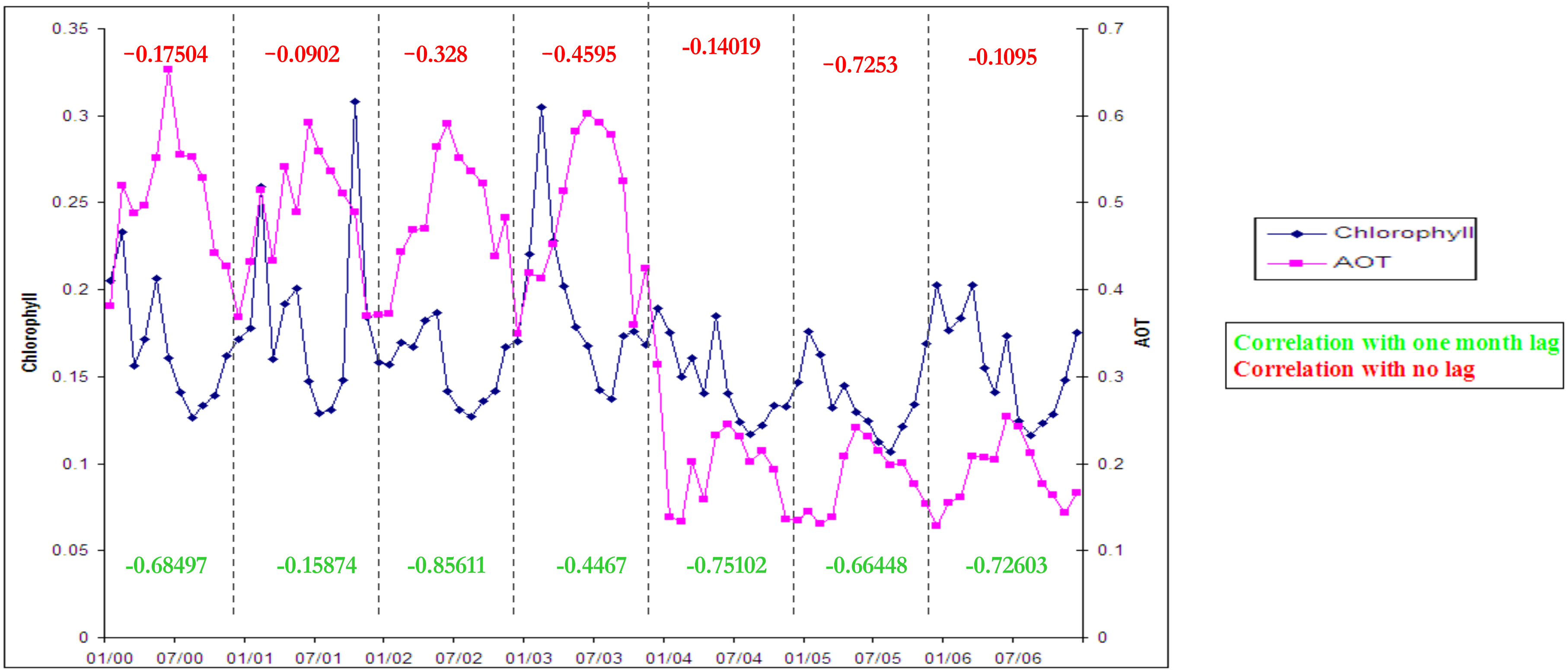
- MODIS derived AOT (Aerosol Optical Thickness)
  - SeaWIFS - Sea-Viewing Wide Field-of-View Sensor
  - MODIS – Moderate resolution Imaging Spectrometer
  - AOT – Aerosol Optical Thickness

*Figure: annual SeaWiFS chlorophyll image for 2001

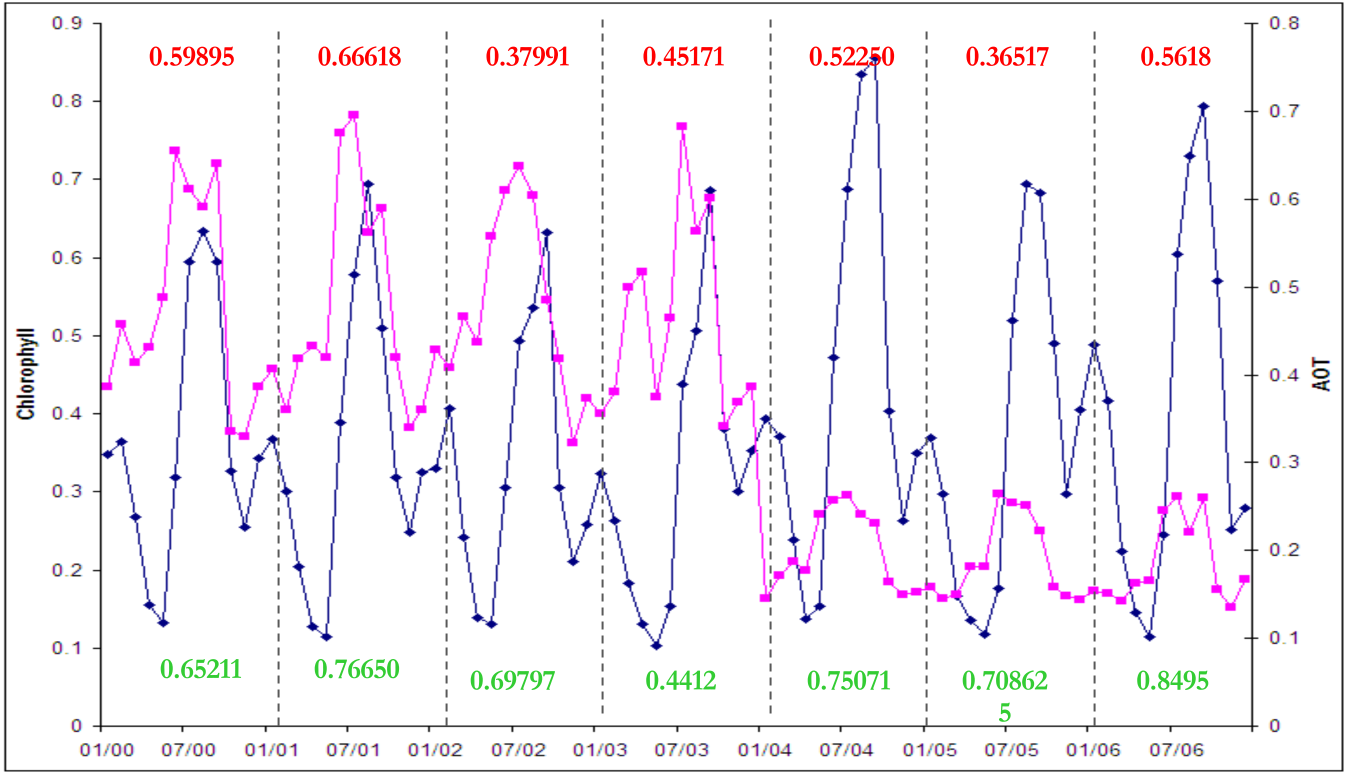


1-Tropical North Atlantic Ocean 2-West coast of Central Africa 3-Patagonia 4-South Atlantic Ocean 5-South Coast of Australia 6-Middle East 7- Coast of China 8-Arctic Ocean

Tropical North Atlantic Ocean → dust from Sahara Desert



Arabian Sea → Dust from Middle East



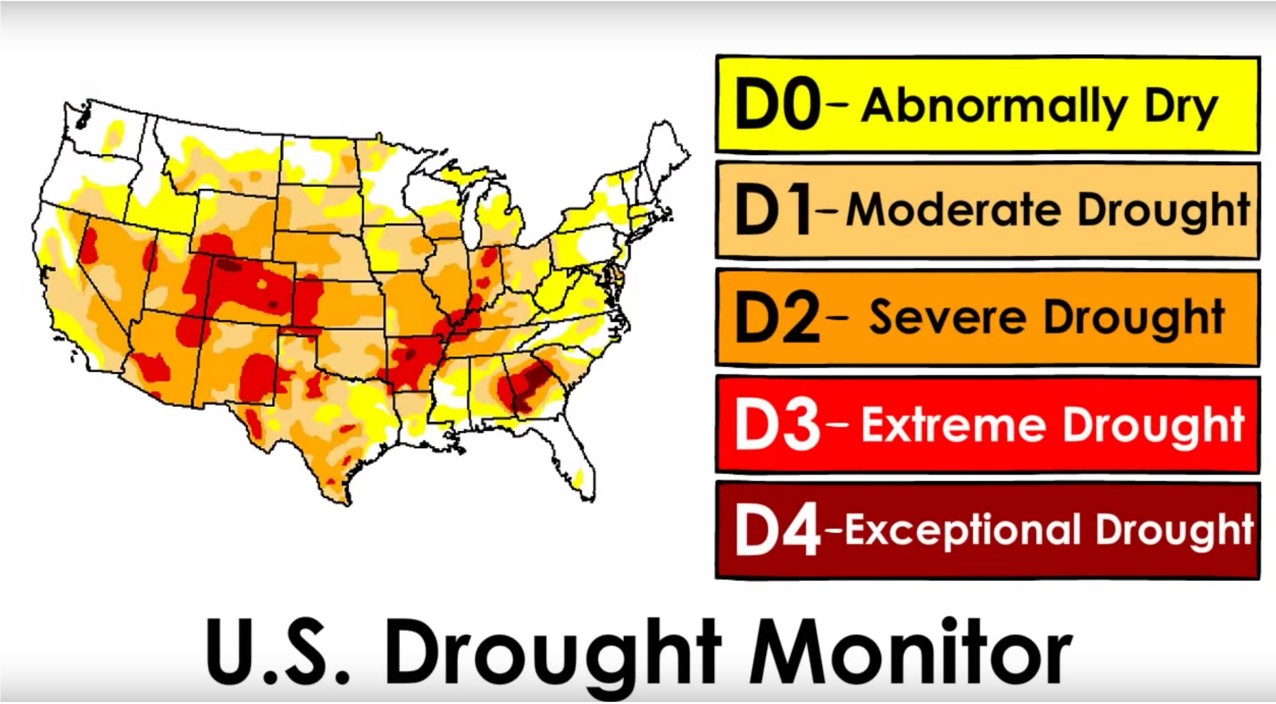
Correlation with one month lag

Correlation with no lag

Summary …

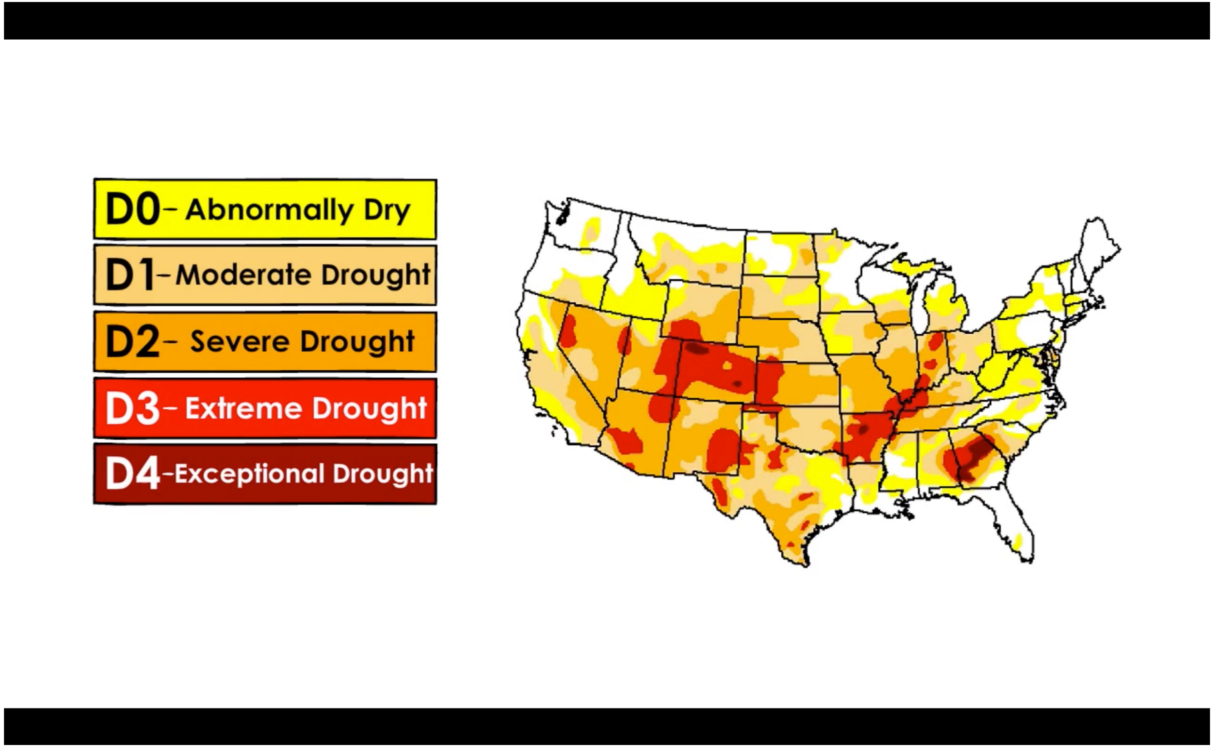
- Dust impacts oceans photosynthetic activity, positive correlations in some areas NEGATIVE correlation in other areas, especially in the Saharan basin
- Hypothesis for explaining observations of negative correlation: In areas that are not nutrient limited, dust reduces photosynthetic activity
- But also need to consider the effect of clouds, ocean currents. Also *need to isolate the effects of dust*. MODIS AOT product includes contribution from dust, Dimethyl Sulfide (DMS), biomass burning etc.

Drought Categories



Assessing Drought Maps in the United States

Drought Categories



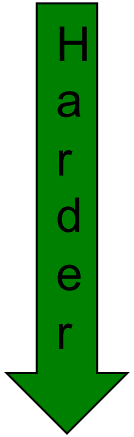
https://pmm.nasa.gov/education/videos/assessing-drought-united-states

Data Mining - What it is

- Extracting knowledge from large amounts of data
- Motivation
  - Our ability to collect data has expanded rapidly
  - It is impossible to analyze all of the data manually
  - Data contains valuable information that can aid in decision making
- Uses techniques from:
  - Pattern Recognition
  - Machine Learning
  - Statistics
  - High Performance Database Systems
  - OLAP (Online Analytical Processing) i.e Financial systems, sensor systems related to weather data.
- Data mining methods must be efficient and scalable (8~10 years ago, data mining could not be done on your Laptop).

Models/ types

- Trade-off between Accuracy and Understandability
- Models range from “easy to understand” to incomprehensible
  - Decision trees
  - Rule induction
  - Regression models
  - Neural Networks



Harder

Thanks!

Form your teams!