Image Metadata Extraction
-- Drain the Data Swamp

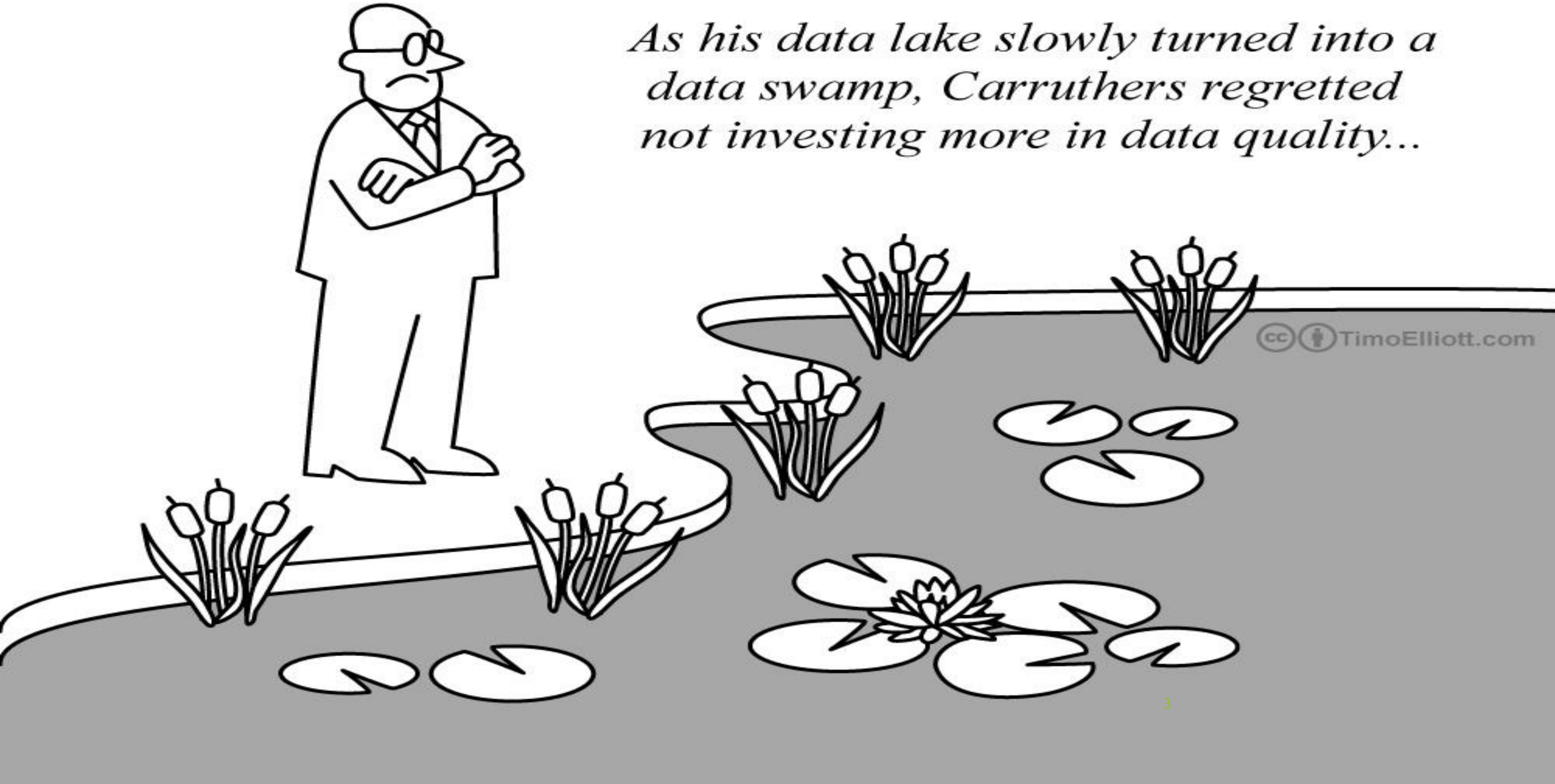
Chaofeng Wu

Advisor: Kyle Chard, Tyler J. Skluzacek, Ian Foster

Computational Institute

The University of Chicago

Motivation

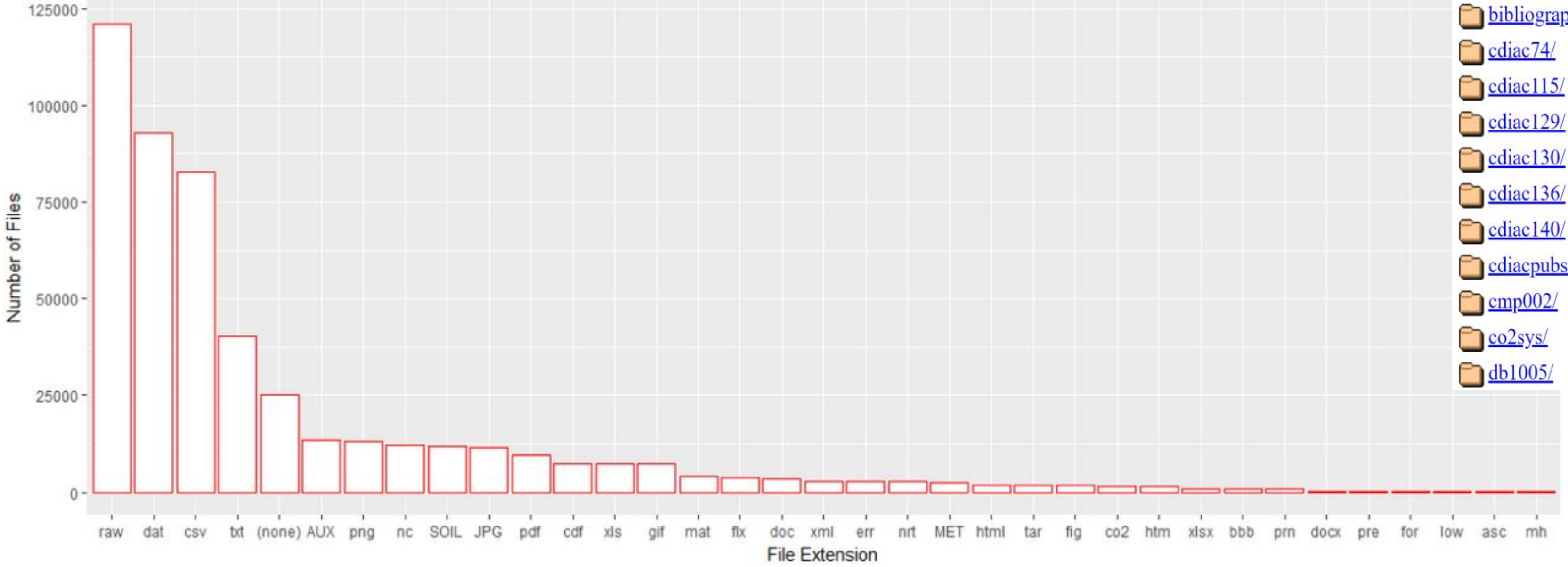

As his data lake slowly turned into a data swamp, Carruthers regretted not investing more in data quality...
TimoElliott.com

f23146.dat	06-Nov-2008 00:23	1.1M
f23205.dat	06-Nov-2008 00:23	1.7M
f23219.dat	06-Nov-2008 00:23	1.4M
f23405.dat	06-Nov-2008 00:23	2.4M
f23418.dat	06-Nov-2008 00:23	1.3M
f23472.dat	06-Nov-2008 00:23	944K
f23631.dat	06-Nov-2008 00:23	1.4M
f23711.dat	06-Nov-2008 00:23	2.5M
f23724.dat	06-Nov-2008 00:23	1.4M
f23804.dat	06-Nov-2008 00:23	2.5M
f23849.dat	06-Nov-2008 00:23	2.6M
f23884.dat	06-Nov-2008 00:23	1.4M
f23891.dat	06-Nov-2008 00:23	1.4M
f23921.dat	06-Nov-2008 00:23	1.5M
f23933.dat	06-Nov-2008 00:23	2.4M

Data in,
different formats,
different extensions,
different repositories,
with similar name

cdiac.ess-dive.lbl.gov/ftp/

Name	Last modified	Size	Description
Parent Directory		-	
ALE_GAGE_AGAGE_deep_archive.tar.gz	25-May-2017 21:10	2.5G	
Atul_Jain_etal_Land_Use_Fluxes/	16-Apr-2013 18:14	-	
CDIAC_UWG_Presentations_Sept2010/	30-Sep-2010 17:21	-	
CSEQ/	14-Aug-2003 20:46	-	
FACE/	23-Jan-2015 20:49	-	
GISS3-D/	19-Jul-2005 19:45	-	
Global_Carbon_Project/	25-May-2017 19:46	-	
HIPPO/	19-Dec-2012 21:26	-	
ICRCCM-radiative_fluxes/	04-Aug-2009 17:59	-	
Nassar_Emissions_Scale_Factors/	18-Aug-2014 18:42	-	
README	13-Sep-2017 22:15	17K	
Smith_Rothwell_Land-Use_Change_Emissions/	01-Aug-2014 19:07	-	
Tris_West_US_County_Level_Cropland_C_Estimates/	08-Jun-2009 17:45	-	
ale_gage_Agage/	07-Sep-2017 19:37	-	
ameriflux/	09-Mar-2017 20:39	-	
bibliography/	14-Feb-2000 21:27	-	
cdiac74/	19-Jul-2005 19:52	-	
cdiac115/	19-Jul-2005 19:48	-	
cdiac129/	19-Jul-2005 19:49	-	
cdiac130/	19-Jul-2005 19:50	-	
cdiac136/	19-Jul-2005 19:51	-	
cdiac140/	19-Jul-2005 19:51	-	
cdiacpubs/	19-Jul-2005 19:53	-	
cmp002/	10-Aug-2009 16:51	-	
co2sys/	28-Oct-2013 13:50	-	
db1005/	04-Aug-2009 18:10	-	



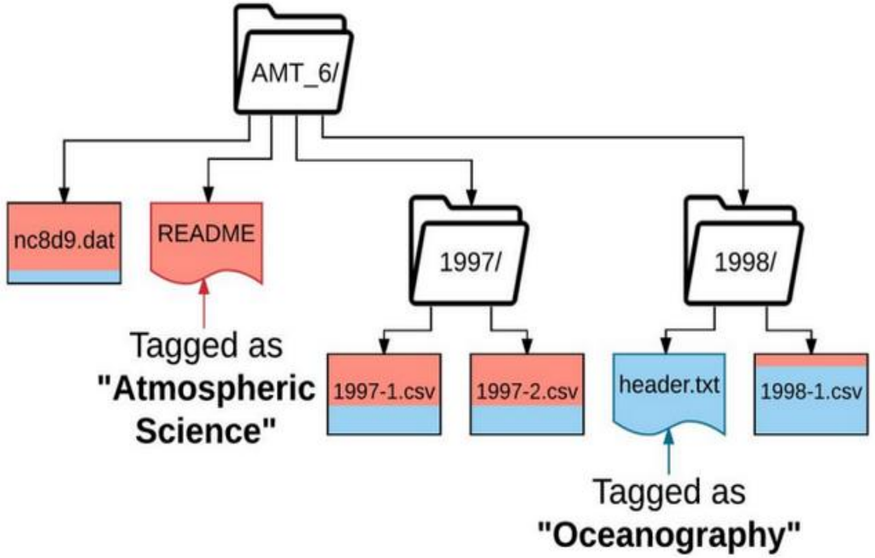
Histogram of top files on a scientific Data Lake

How to deal with all the mess? – Skluma!

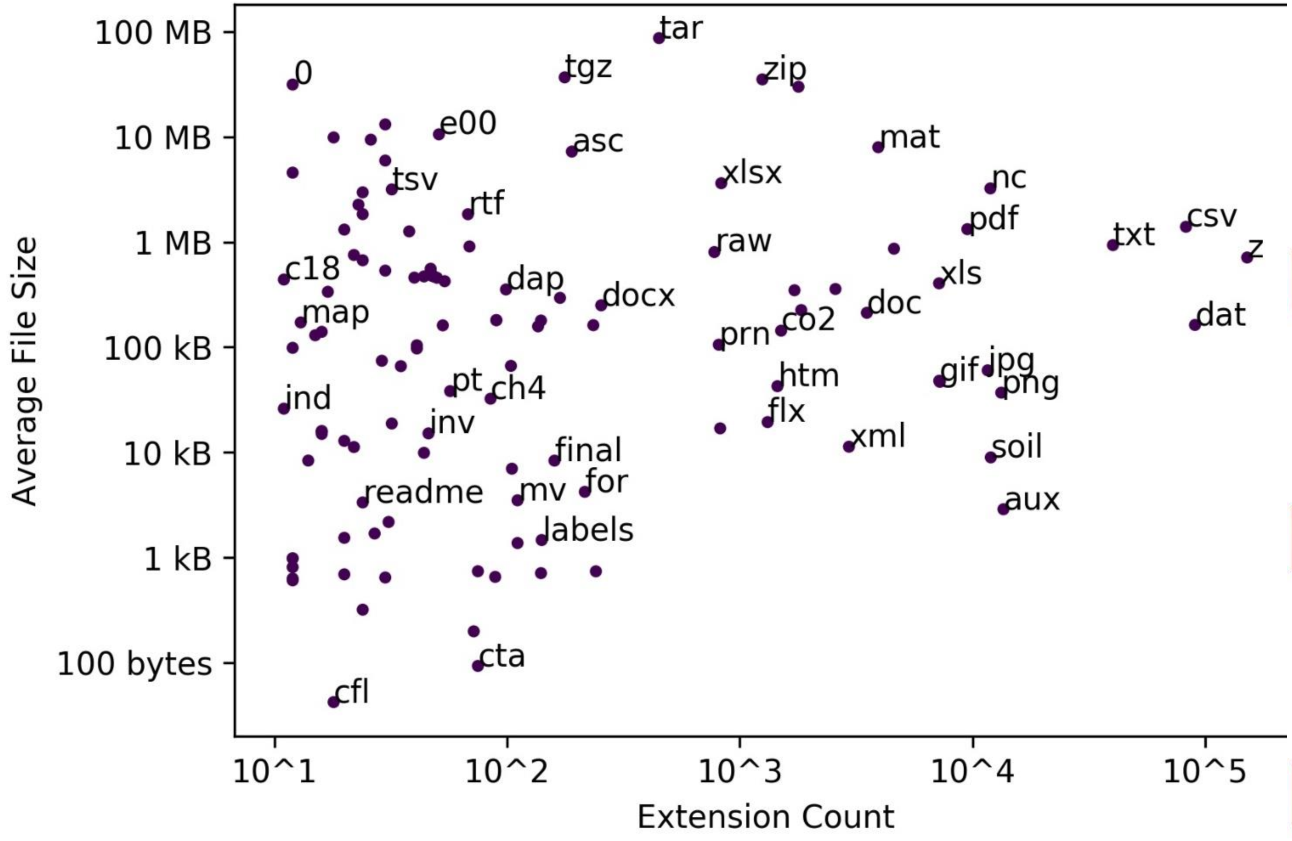
Text-based files → File metadata → Contextual relationship → Indexed searchable collection

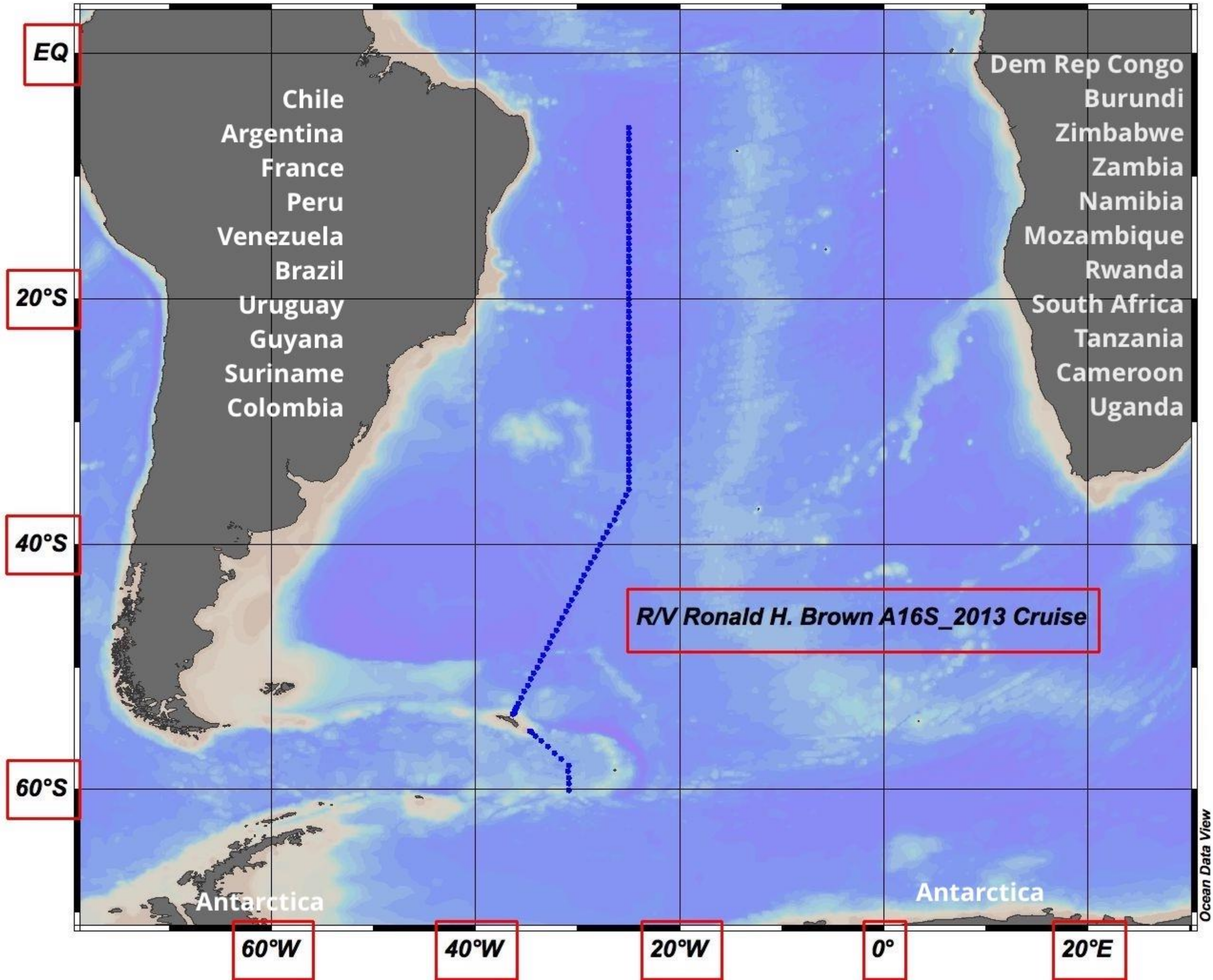
```
SUMMARY11.MONTHLY.ISOMASS.DAT
DOI: 10.3334/CDIAC/ffe.MonthlyIsomass.2009
 All values, except year, should be zero or negative.
 Units of minigrid and maxigrid are mass*del.
 Units of del 13 C are per mil.
year  minigrid maxigrid   del13C
1950   -69.75     0.00   -26.16
1951   -72.36     0.00   -26.15
1952   -71.14     0.00   -26.27
1953   -72.29     0.00   -26.33
1954   -72.61     0.00   -26.46
1955   -79.85     0.00   -26.48
1956   -81.25     0.00   -26.36
1957   -80.49     0.00   -26.51
1958   -79.62     0.00   -26.55
1959   -83.20     0.00   -26.62
```

CDIAC
2009
Year range
Minigrid
Maxigrid
…

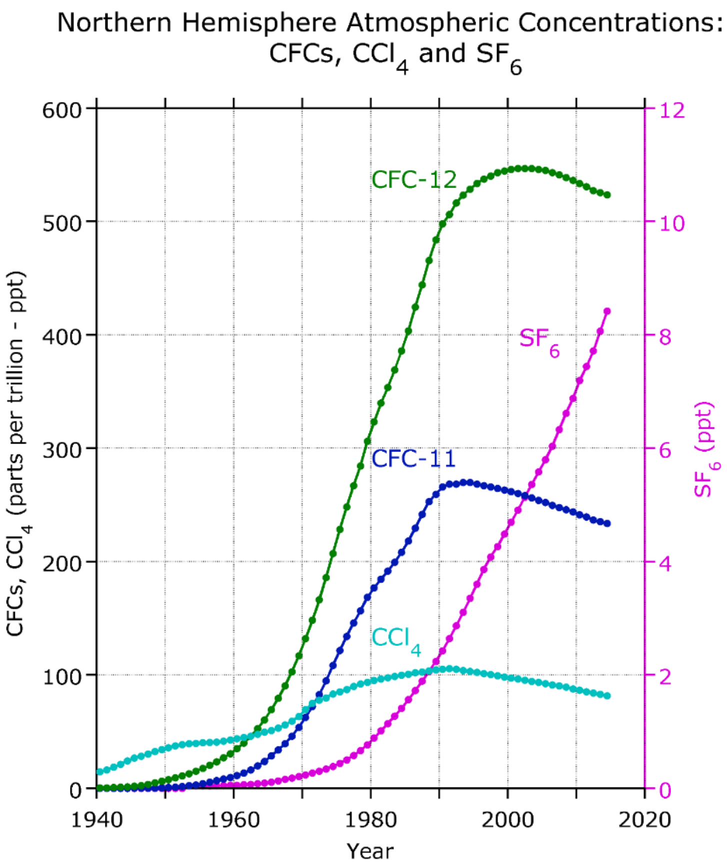


Goal: extract metadata from images

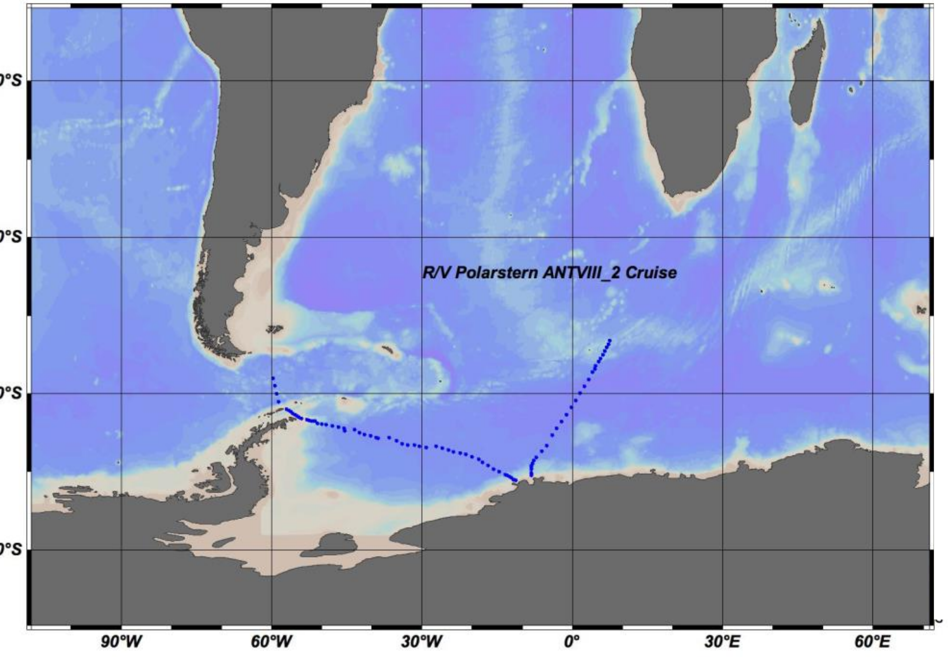




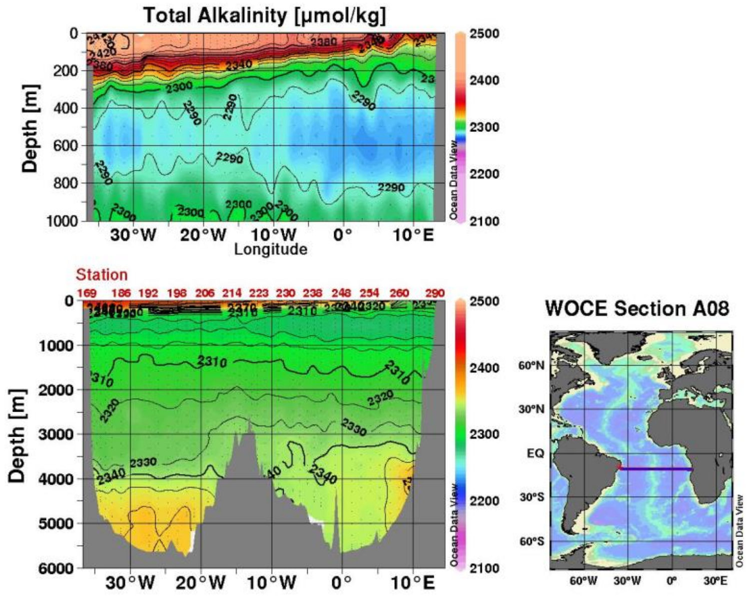
Designed Solution



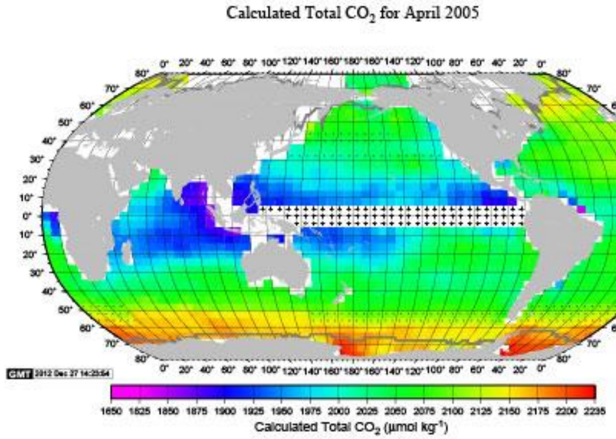
Northern Hemisphere Atmospheric Concentrations: CFCs, CCl4 and SF6
CFC-12
CFC-11
CCl4
SF6
CFCs, CCl4 (parts per trillion - ppt)
SF6 (ppt)
Year



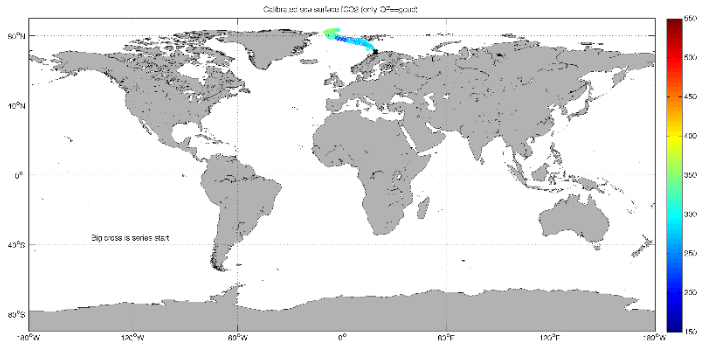
R/V Polarstern ANTVIII_2 Cruise
20°S
40°S
60°S
80°S
90°W
60°W
30°W
0°
30°E
60°E

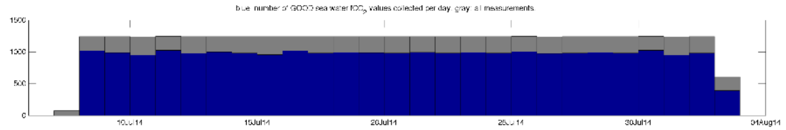


Total Alkalinity [µmol/kg]
Depth [m]
Longitude
Station
WOCE Section A08
Ocean Data View



Calculated Total CO2 for April 2005
Calculated Total CO2 (µmol kg-1)





Metadata

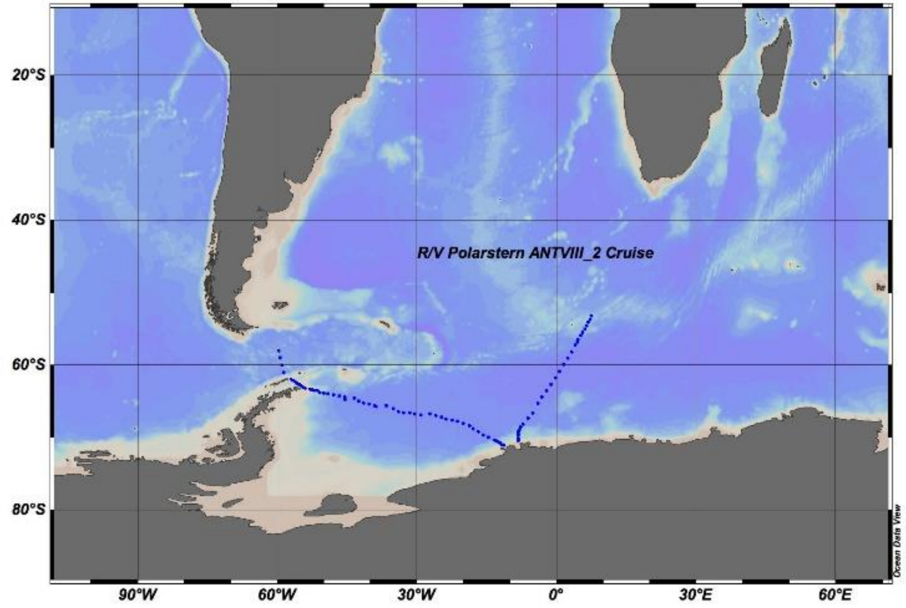
System metadata

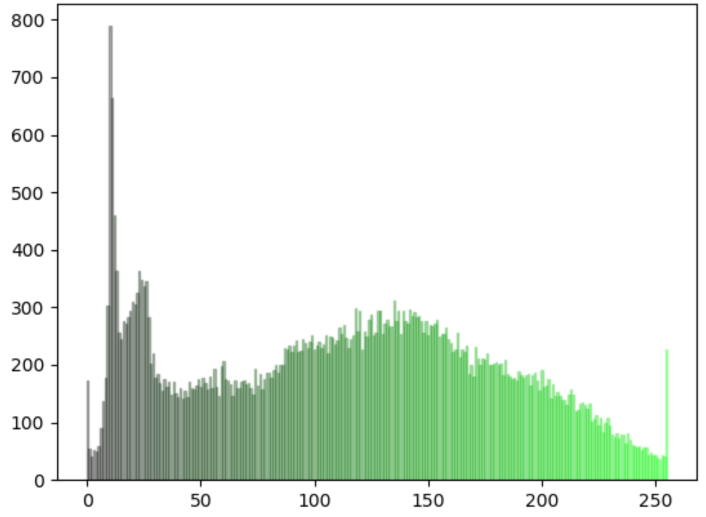
- Image name
- Image path
- Image extension (jpg, png…)
- Image file size (KB, MB…)
- Image size (1024*768…)
- Image color mode (RGB…)

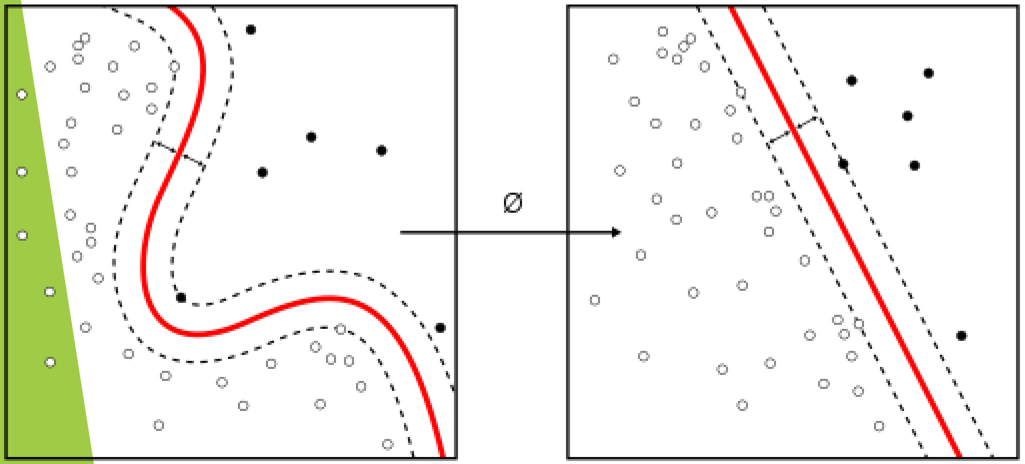
Image metadata

- A map?
- A plot?
- A figure?
- A photo?
- How to classify the image?

Process


20°S
40°S
60°S
80°S
R/V Polarstern ANTVIII_2 Cruise
90°W
60°W
30°W
0°
30°E
60°E
Ocean Data View


800
700
600
500
400
300
200
100
0
0
50
100
150
200
250

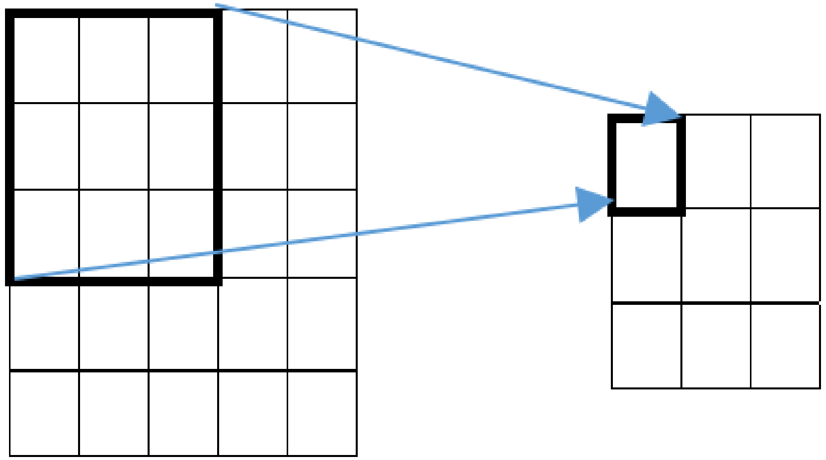

Ø

MAP

Features

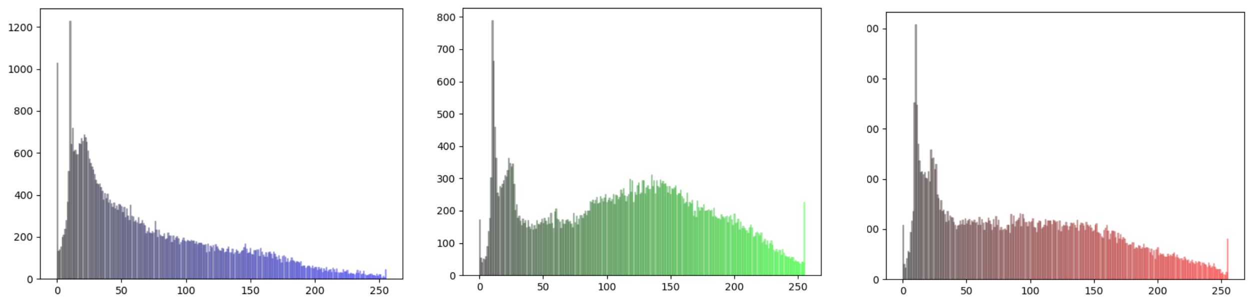
Mean square

- The average of small blocks in different part of original images
- Hope this can show some local features of image



Color histogram

- The color frequency in the images
- Hope this can show color features of image



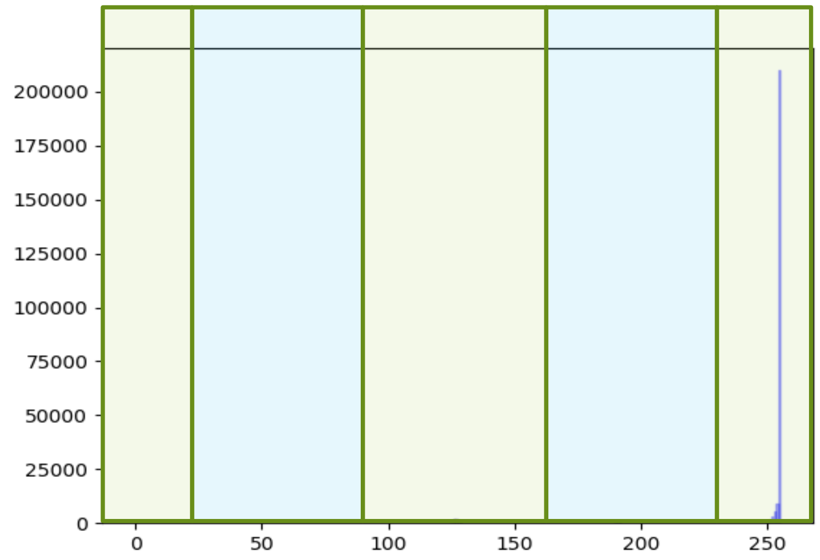
Reducing dimension of features

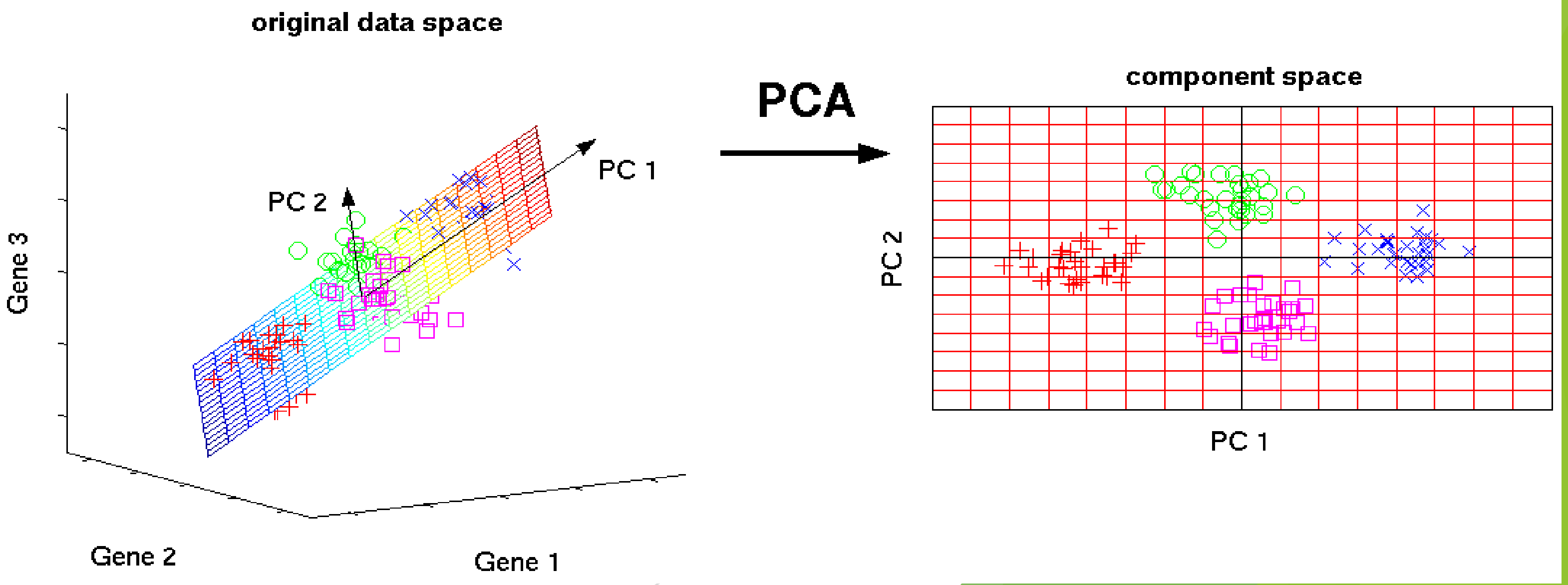
Naïve way

- Grouping color histogram in to small partitions
- Repeat the mean square for small square

Principal Component Analysis (PCA)

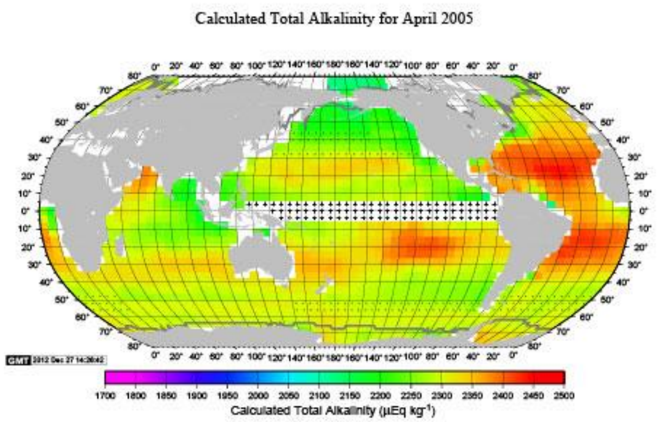
- Statistical procedure to reduce dimension of variables
- Keep the variable whose variance is large in result

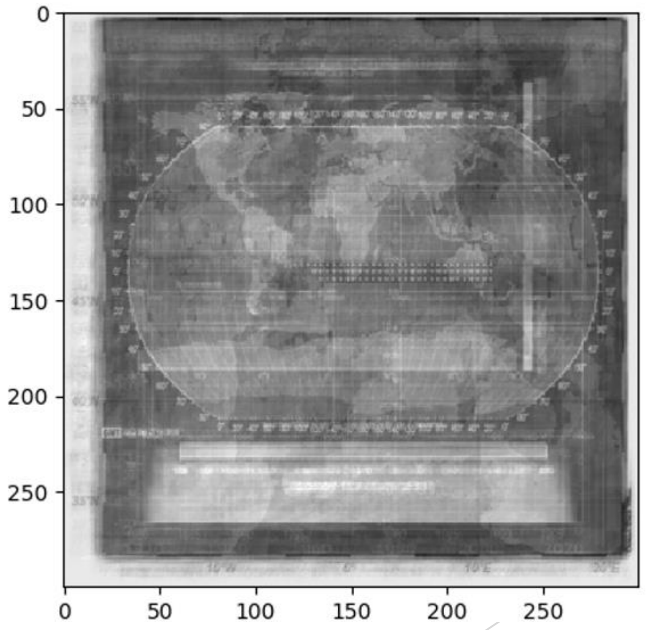




New feature by PCA

- Resized image
- Using PCA to form a basic set of images
- Using basic set to represent new images

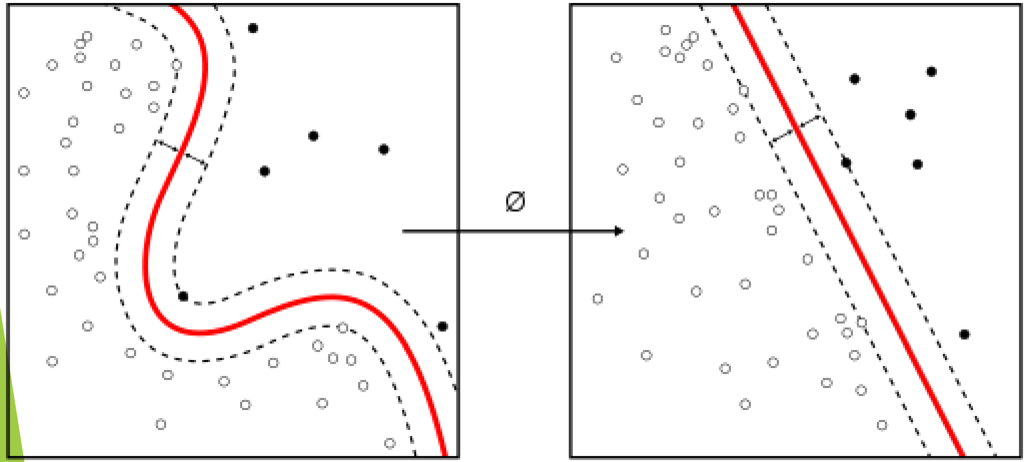




Models

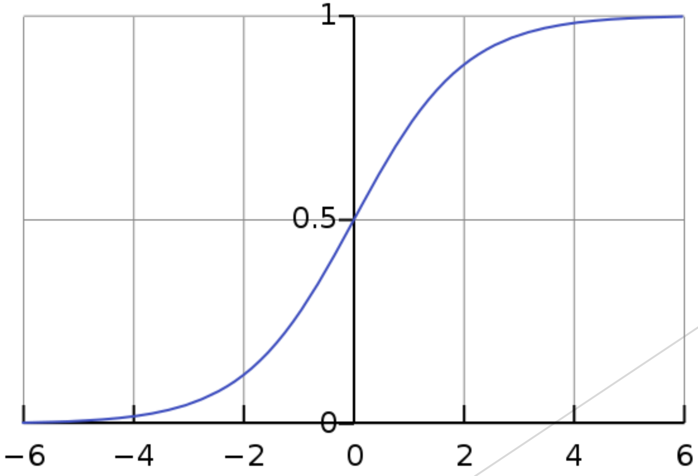
support vector machine(SVM)

- Supervised learning model used for classification and regression analysis
- SVM is applied to text categorization, image classification, and image segmentation



Logistic regression

- A classification method that generalizes logistic regression to multiclass problems
- Logistic regression models the probability of classes



Evaluation

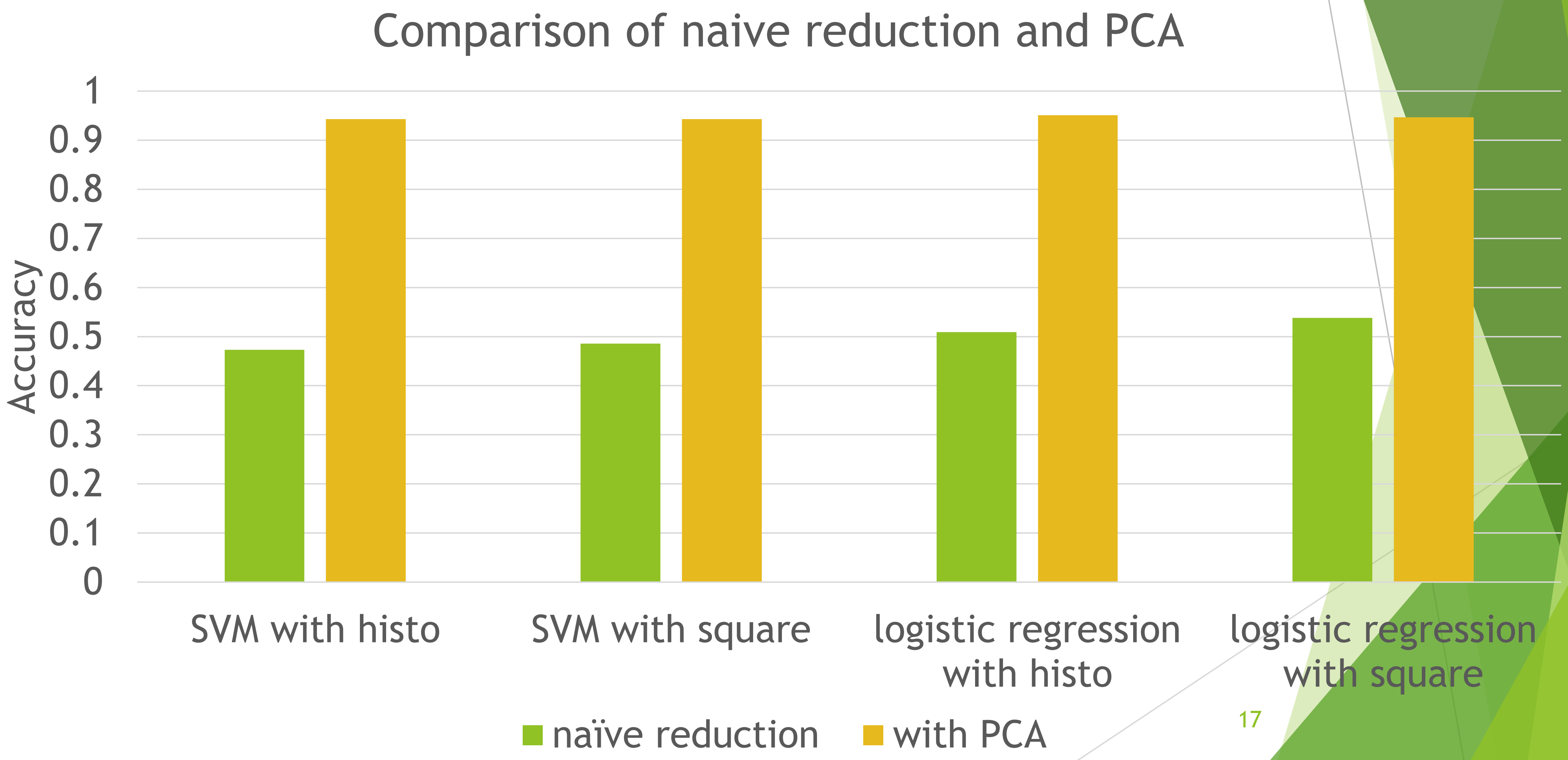
Dataset

- Using part of data from Carbon Dioxide Information Analysis Center (CDIAC)
- Labeling 532 images, half for training and half for testing

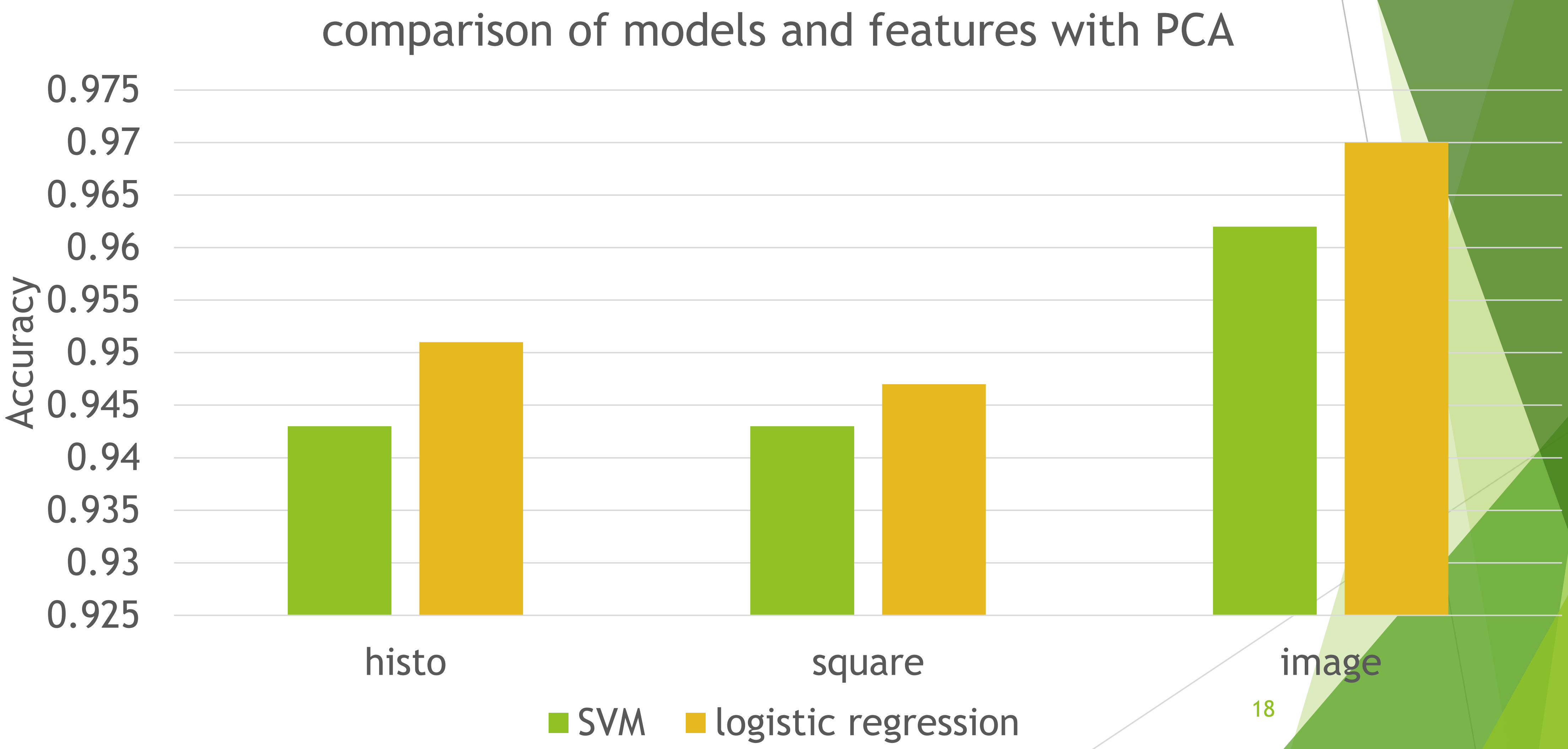
	Gif	Png	Jpg	Bmp	Total
Number	18	64	449	1	532

	Line plot	Map	Map&chart	Map&colorplot	Map&histogram	figure
Number	51	236	105	87	49	4

Comparison of reduction methods



Comparison of models and features with PCA

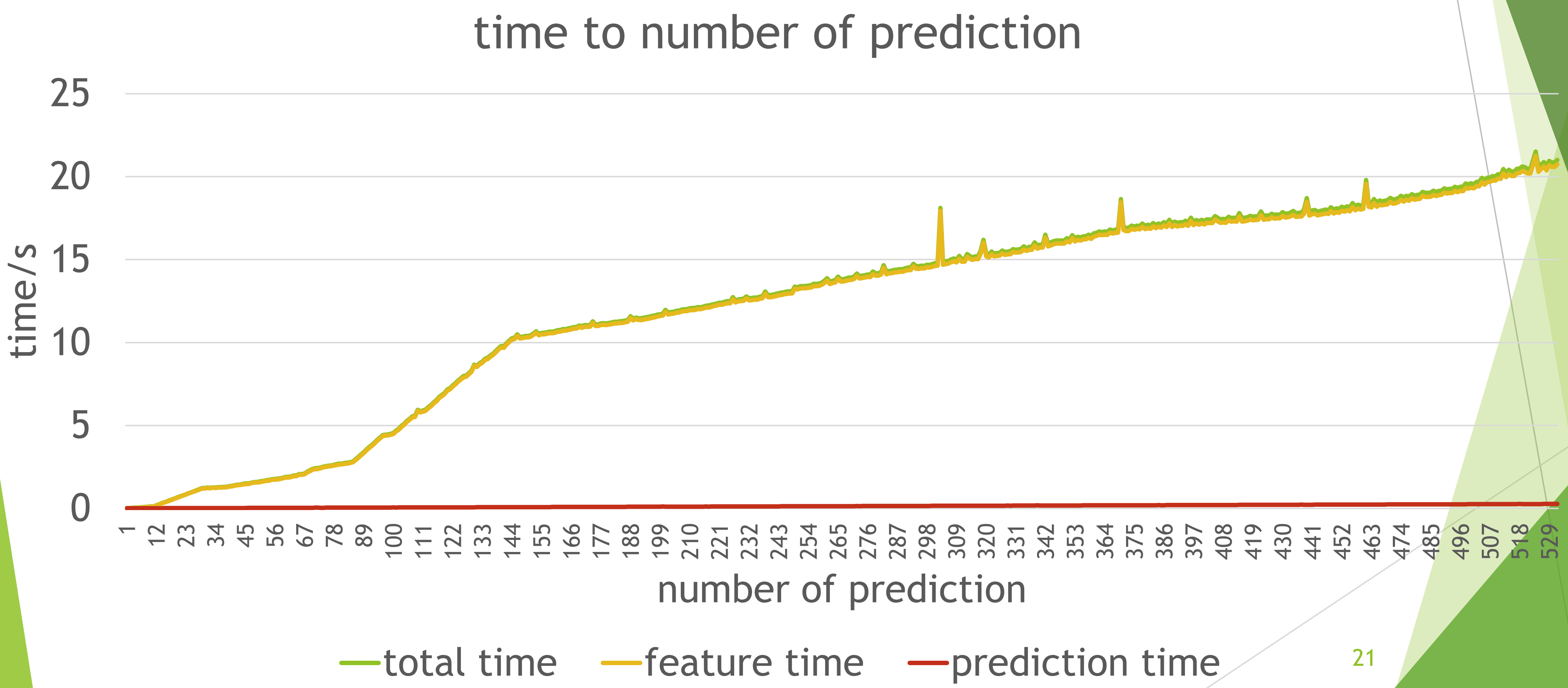


Time analysis across models and features

Detailed result analysis
-- logistic regression with image and PCA

	precision	recall	F-score	support
line plots	0.8947	1	0.9444	34
maps	0.9652	0.9823	0.9737	113
map&depth chart	1	1	1	44
map&colorplot	1	0.9565	0.9778	46
map&histogram	1	1	1	25
figures	0	0	0	4

Detailed prediction time analysis
-- logistic regression with image and PCA



conclusion

- Applying PCA to get the feature can get better accuracy in image classification
- SVM and logistic regression model work well in image classification
- Future work
  - Test current model on larger dataset
  - Try convolutional layer with PCA
  - Try convolutional neural network and see performance
  - Extract more info from image

Citation

- Chapelle, O., et al. “Support vector machines for histogram-Based image classification.” *IEEE Transactions on Neural Networks*, vol. 10, no. 5, 1999, pp. 1055–1064., doi:10.1109/72.788646.
- P. Beckman, T. J. Skluzacek, K. Chard, and I. Foster, “Skluma: A statistical learning pipeline for taming unkempt data repositories,” in Proceedings of the 29th International Conference on Scientific and Statistical Database Management, SSDBM ’17, (New York, NY, USA), pp. 41:1–41:4, ACM, 2017.
- U.S. Department of Energy, “Carbon dioxide information and analysis center,” 2018.
- Matthew A. Turk, Alex P.Pentland, “Face Recognition Using Eigenfaces.” MIT, IEEE 1991 https://www.cs.ucsb.edu/~mturk/Papers/mturk-CVPR91.pdf

Thanks!

Questions?

Eigenfaces

- Eigenfaces is the name given to a set of eigenvectors when they are used in the computer vision problem of human face recognition
- The eigenvectors are derived from the covariance matrix of the probability distribution over the high-dimensional vector space of face images
- The eigenfaces themselves form a basis set of all images used to construct the covariance matrix
- This produces dimension reduction by allowing the smaller set of basis images to represent the original training images
- Classification can be achieved by comparing how faces are represented by the basis set

Curse of dimensionality

- The curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience
- In machine learning problems that involve learning a "state-of-nature" from a finite number of data samples in a high-dimensional feature space with each feature having a range of possible values, typically an enormous amount of training data is required to ensure that there are several samples with each combination of values
- A typical rule of thumb is that there should be at least 5 training examples for each dimension in the representation. With a fixed number of training samples, the predictive power of a classifier or regressor first increases as number of dimensions/features used is increased but then decreases, which is known as Hughes phenomenon` or peaking phenomena