Classification and Modeling of Evolving Solar Features

David C. Stenning

Statistics Section, Department of Mathematics, Imperial College London

New Paradigm for Solar Imaging Processing

- Current solar observatories are generating an enormous volume of high-resolution solar image data.
- Manual identification, classification and tracking of sunspots and other solar features is becoming increasingly laborious.
- Studying images "by eye" limits the types of analyses that can be performed—interesting features must be extracted and propagated in machine-readable form if they are to be utilized in a sophisticated statistical procedure.
- Automated data processing = reproducible science

More data is <u>not</u> just more data...*more is different!*
(K. Borne, Computational Astrostatistics 2010)

Sunspots

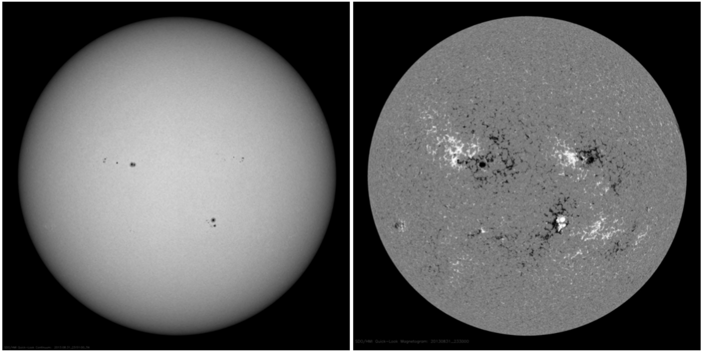
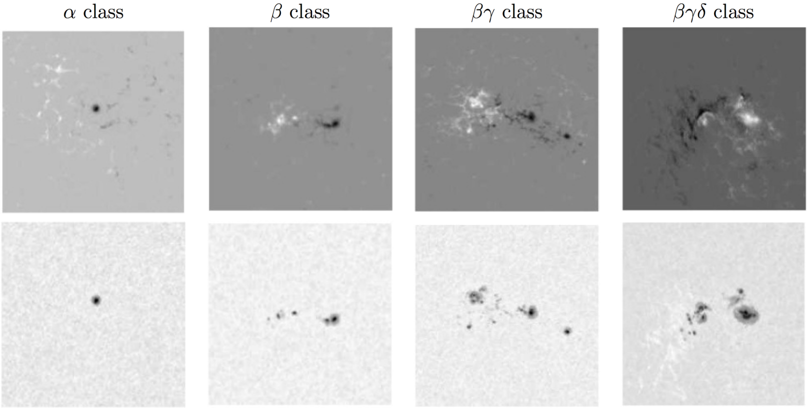


Image Credit: NASA/SDO

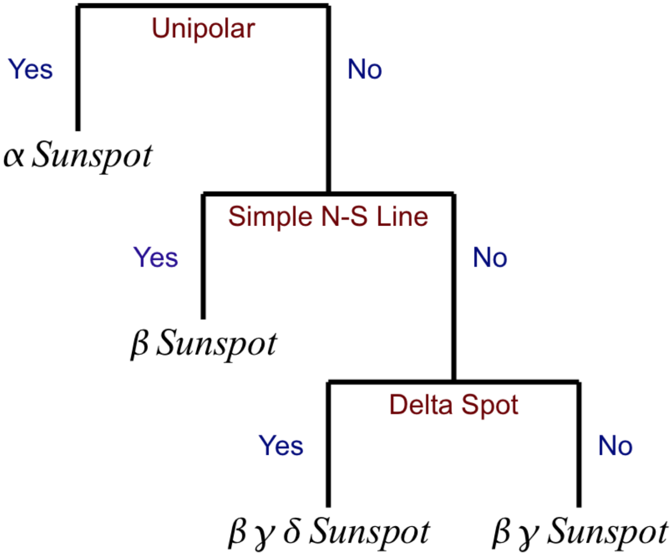
- *Sunspots* form when intense magnetic fields inhibit convection.
    - Show up as dark spots on the Sun's photosphere in *white-light images* (left image).
    - Classified based on the complexity of magnetic flux distribution as seen in *magnetograms* (right image).

Mount Wilson Classification

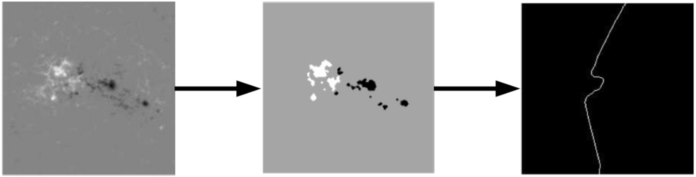


Four broad classes— α, β, $\beta\gamma$, and $\beta\gamma\delta$—based on the complexity of magnetic flux distribution. *Top row:* magnetograms. *Bottom row:* white-light images.

Mount Wilson Classification Rules: Decision Tree



Science-Driven Feature Extraction

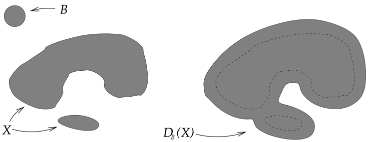


- Classification is predictive of solar activity (e.g., solar flares)
- Use Mt. Wilson rules to guide feature selection → **science-driven feature extraction**
  - Physically meaningful and interpretable features
- Features from **mathematical morphology**
- Capture relevant information in more informative manner vs. manual classification
- Amenable to statistical analyses: **model sunspot evolution**

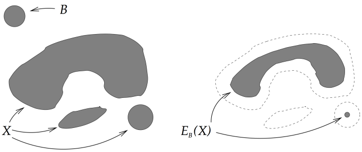
By crafting numerical features that are motivated by knowledge of the underlying physical processes, we are attempting to steer "black-box" classification algorithms with science.

Basic Morphological Operations

Dilation and Erosion



Example of dilation

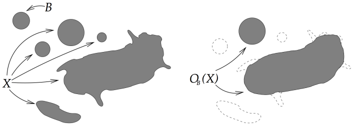


Example of erosion

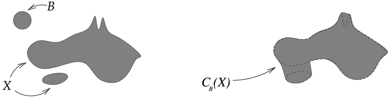
- The two fundamental operations in mathematical morphology are *dilation* and *erosion*.
- They use a structuring element (SE) B to probe and alter the shapes of the objects inside an image X.
  - The *dilation* of X by B is the set of points z such that B *hits* X when the origin of B is placed at z.
  - The *erosion* of X by B is the set of points z such that B *fits* wholly inside X when the origin of B is at z.

Basic Morphological Operations

Opening and Closing



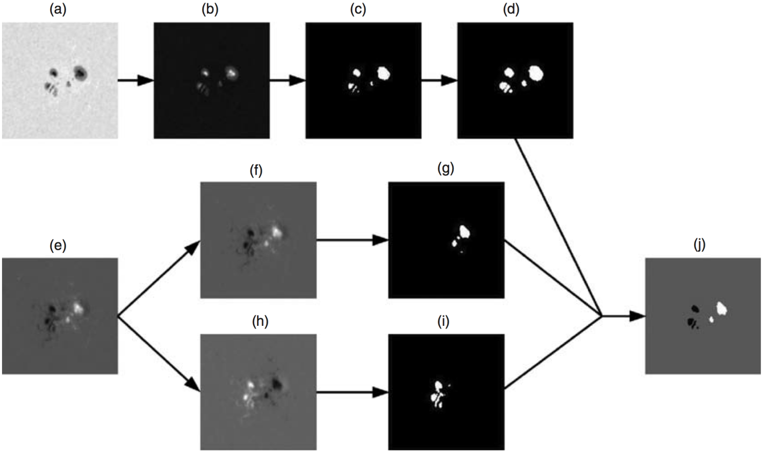
Example of opening



Example of closing

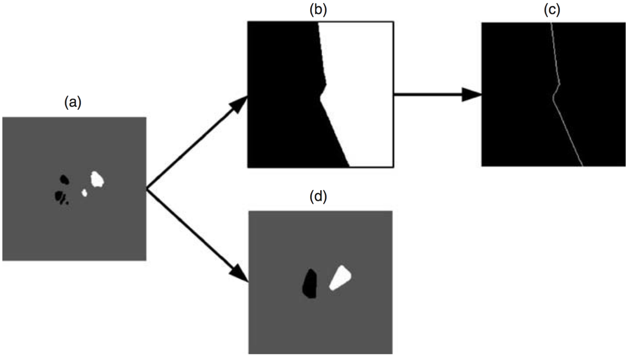
- Dilation and erosion are combined to form the two most common morphological operations: *opening* and *closing*.
  - Morphological opening is an erosion of the image with a SE, followed by a dilation with the same SE.
    - Smoothes features from the interior and removes noise.
  - Morphological closing is a dilation followed by an erosion.
    - Smoothes out the image and fills in gaps without degrading or distorting the salient features.

Feature Extraction Routine I: Active Region Identification



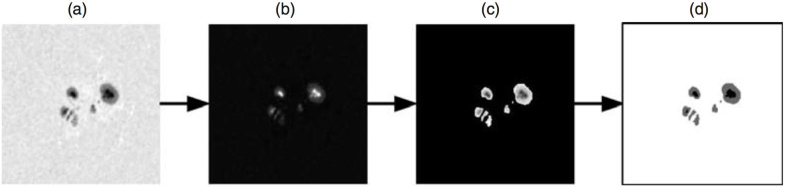
- Using MM to take a white-light image, image (a), and corresponding magnetogram, image (e), to produce a simple "trinary" representation of the active region, image (j).

Feature Extraction Routine II: Numerical Summaries



- From (a) we calculate the *ratio of the number of opposite polarity pixels* and the *amount of scattering* of the pixels for each polarity.
- A seeded region growing algorithm applied to (a) yields (b), from which we obtain the polarity inversion line (c). We then calculate the *polarity inversion line curvature*.
- Convex hulls around the pixels of opposite polarity in (a) yields (d), from which we calculate the *polarity mixture*.

Feature Extraction Routine III: Delta Spots



- We return to the white light image, image (a) above, and use MM to identify the *umbrae* and *penumbrae* pixels.
- Image (d) above, when combined with the trinary active region representation, is used to determine the *number of delta spots* and the *total size of delta spots*.

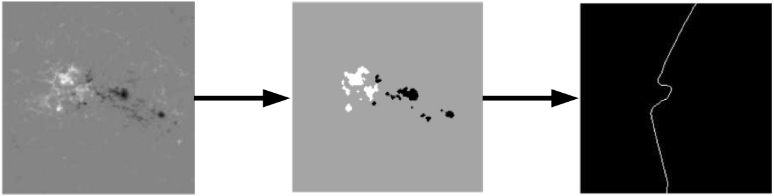
Numerical Summaries Summary

We use our morphological representation of sunspot groups and active regions to obtain scientifically based numerical features:

- The *ratio* of pixels of opposite polarities.
- The *amount of scattering* of the pixels for each polarity.
- Polarity inversion line *curvature*.
- Area of *mixture* for the convex hulls around each polarity region.
- The *number and size of delta spots*.

Science-Driven Feature Extraction: Examples

$\beta\gamma$ sunspot group:



β sunspot group:

Machine Learning

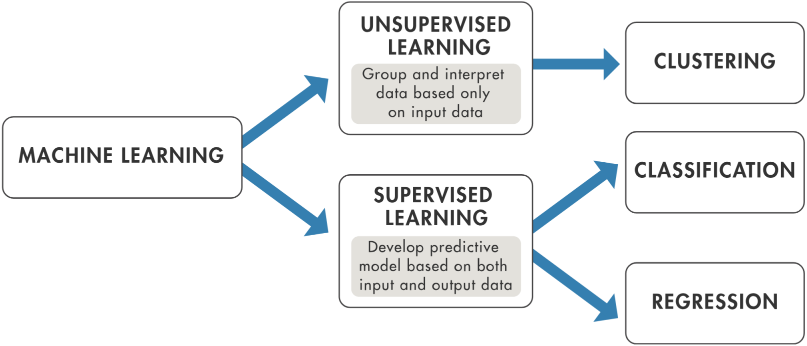


Image Credit: https://uk.mathworks.com/discovery/machine-learning.html

Decision Trees (for Classification)

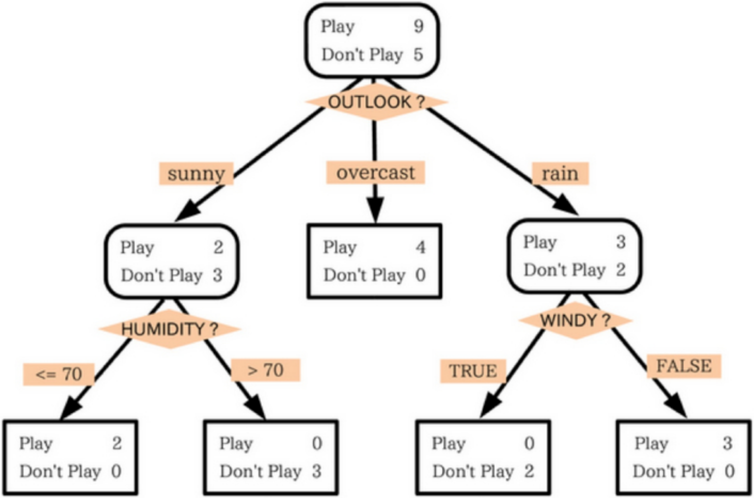


Figure Credit: http://gautam.lis.illinois.edu/monkmiddleware/public/analytics/decisiontree.html

Decision Boundaries

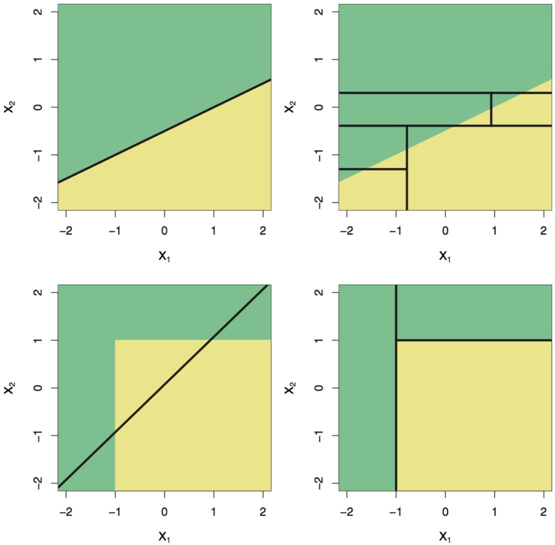


Figure from ISLR (Figure 8.7, pg 315)

Advantages and *Disadvantages* of Trees

Adapted from ISLR (pgs 315-316):

- **Easy to explain. (Easier than linear regression!)**
- **Mirror human decision-making. (Maybe? Seems to be the case for MW classification!)**
- **Can be displayed graphically.(Easy for non-experts!)**
- **Easily incorporate qualitative predictors. (No dummy variables needed!)**
- *Predictive accuracy can be poor compared to other methods.*
- *Non-robust. Small change in data typically results in large change in final tree.*

Random Forest (RF)

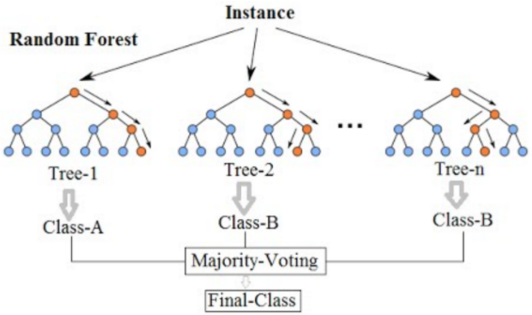
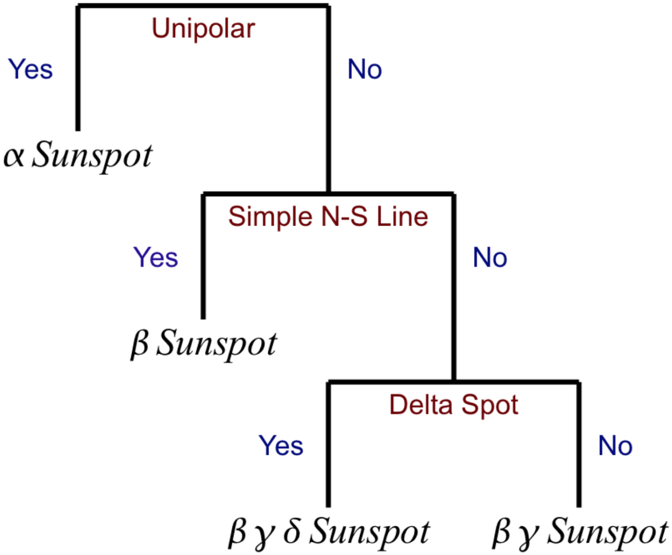


Figure: https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

- An RF is an *ensemble* of decorrelated decision trees
- With N cases in a training set and p features, each tree in the (RF) is constructed by
  - sampling $n = N$ cases from the training set with replacement
  - randomly selecting $\sqrt{p}$ features to make a decision at each node, and growing tree to completion
- Resulting classifications are decided are by majority vote

Mount Wilson Classification Rules: Decision Tree



Classifying Sunspot Groups with Random Forests

- The features we have derived—pixel ratio, amount of scattering, separating line curvature, polarity mixture, and number and size of delta spots—are used as inputs to an RF.
- Scientific validity of the numerical features is determined by a satisfactory level of agreement between the manual and automatic classifications.
- RF well-suited to this particular problem:
  - features were crafted to make “if-then-else” type decisions
  - ”soft” classifications
  - can easily incorporate new features
  - easy to use software (e.g., `randomForest` package in R)

Random Forest Results

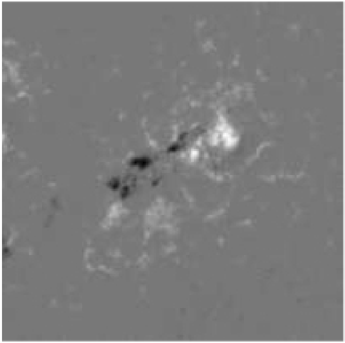
- Data are 119 magnetogram and white light image pairs
- Because the training set for a particular tree in the RF is a bootstrap sample, the cases not included form an "out-of-bag" (OOB) test set for that tree.
- We can thus evaluate the RF's performance based on prediction on OOB data.
- Using a RF with 1000 trees we obtain:

		Manual Classification			
		α	β	$\beta\gamma$	$\beta\gamma\delta$
Automatic Classification	α	25	1	0	0
	β	2	63	5	0
	$\beta\gamma$	0	1	11	1
	$\beta\gamma\delta$	0	0	2	8

Classification Disagreements

- Perfect classification is not necessarily the gold standard when automating a manual classification that is artificial and subjective.
- Classification "by eye" is prone to error and inconsistencies.
  - Two experts looking at the same images will not have 100% agreement.
- Nevertheless, results suggest that the numerical summaries we derived capture salient scientific information.
  - In particular, all disagreements are over adjacent classes.

Example: $\beta\gamma/\beta\gamma\delta$ Disagreement

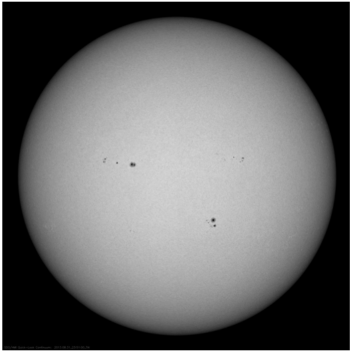


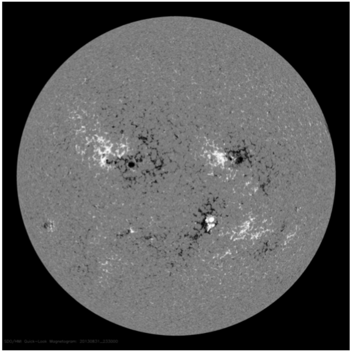
This active region has a manual classification of $\beta\gamma$ and was given a classification of $\beta\gamma\delta$ by the random forest classifier. The presence of a δ spot in the center of the active region is ambiguous.

Beyond Discrete Classification

- Manual classification routines must necessarily rely on a discrete number of classes, but automatic routines need not be likewise hindered.
- Continuous numerical features allow us to better describe the continuum of sunspot group/active region morphology.
- By tracking particular sunspots/active regions over time, we will be able to model the evolution of the magnetic field structure.
- This will hopefully allow for better prediction of dramatic solar events.

High-Cadence SOHO Data





- Have 14 years of SOHO data with images taken every few hours
- Numerical features that were used for classification will be extracted for all active regions, creating a *time series* of features
  - Useful for predicting solar flares?

Other Data to Consider?

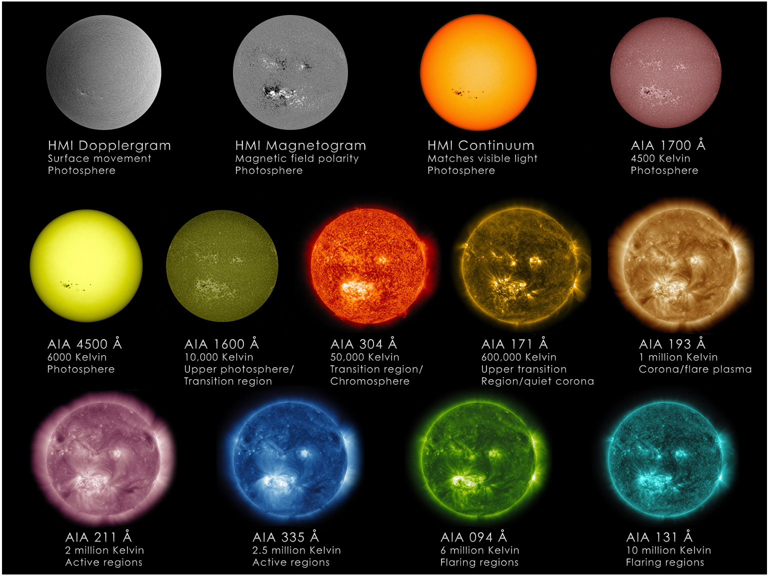


Image Credit: NASA/SDO/GSFC

Thanks!

Collaborators:

- David A. van Dyk (Imperial College London)
- Vinay Kashyap (Smithsonian Astrophysical Observatory)
- Thomas C.M. Lee (UC Davis)
- C. Alex Young (NASA)

Also, thanks to Imperial College London and the CHASC International Astro-Statistics Collaboration!

Any questions? (I have plenty for you!)

For Further Reading I

Stenning et al.
Morphological Image Analysis and Its Application to Sunspot Classification.
Statistical Challenges in Modern Astronomy V, Springer, 2012.

Stenning et al.
Morphological Feature Extraction for Statistical Learning with Applications to Solar Image Data.
Statistical Analysis and Data Mining, August, 2013.

James, Witten, Hastie and Tibshirani.
An Introduction to Statistical Learning with Applications in R (ISLR), Springer, 2013.