

Challenges and Applications of Data Science

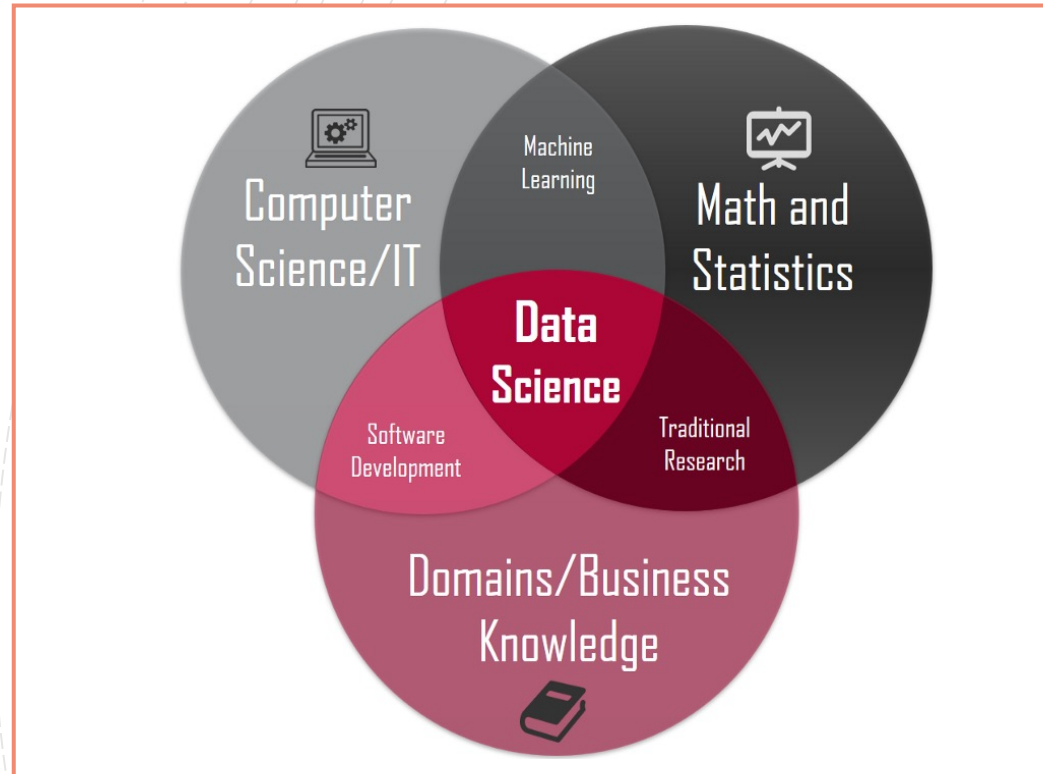
Yun-Ju Chen

UT Dallas

Jun 20 2019 CEDAR Santa Fe NM



Data Science



Workflow

- Initial data exploratory
- Quality control or cleaning
- Handling the missing values
- Figuring out a good way to analyzing (select the algorithms based on your task)
- Interpreting and visualizing data
- Documentation and Writing

Initial Data Exploratory – Get to Know the Dataset

- ✓ Hints to data cleaning and to deal with missing values
- ✓ Ideas for developing an approach to solving problems
- ✓ Help you to interpret your results
- Quick look of numerical distribution (ex. histograms, boxplots ...)
- Relationships between variables

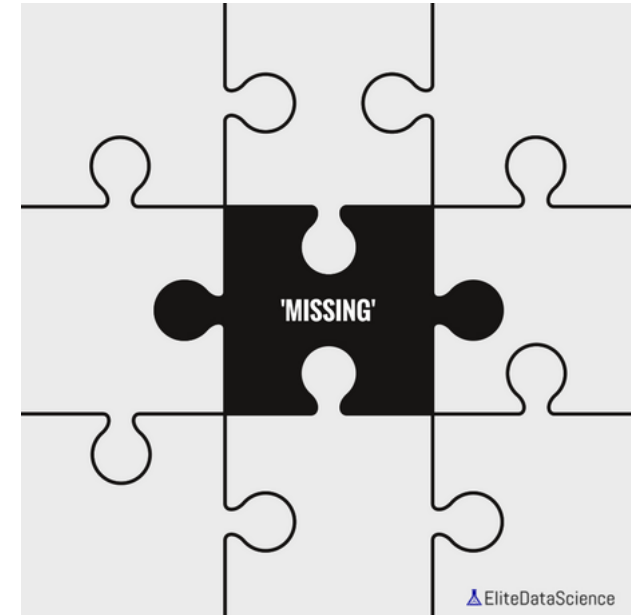
Quality control or cleaning

- Remove unwanted observation (duplicate or irrelevant observations)
- Remove the unwanted outliers (\neq big numbers)
- Remove the unwanted noise
- Flag the data that don't make sense in the moment
- *Whenever removing data, we lost information.*



Missing data

- **Deletion**
 - Drop the nearby observation (in time or space)
 - Drop the variable if missing data $> 50\%$
- **Imputation**
 - Mean, median, mode or most frequent values
 - Interpolation
 - Linear Regression
 - K Nearest Neighbors or other machine learning techniques
 - Multiple Imputation
- *Filling in the missing value, we modify the information in the data.*



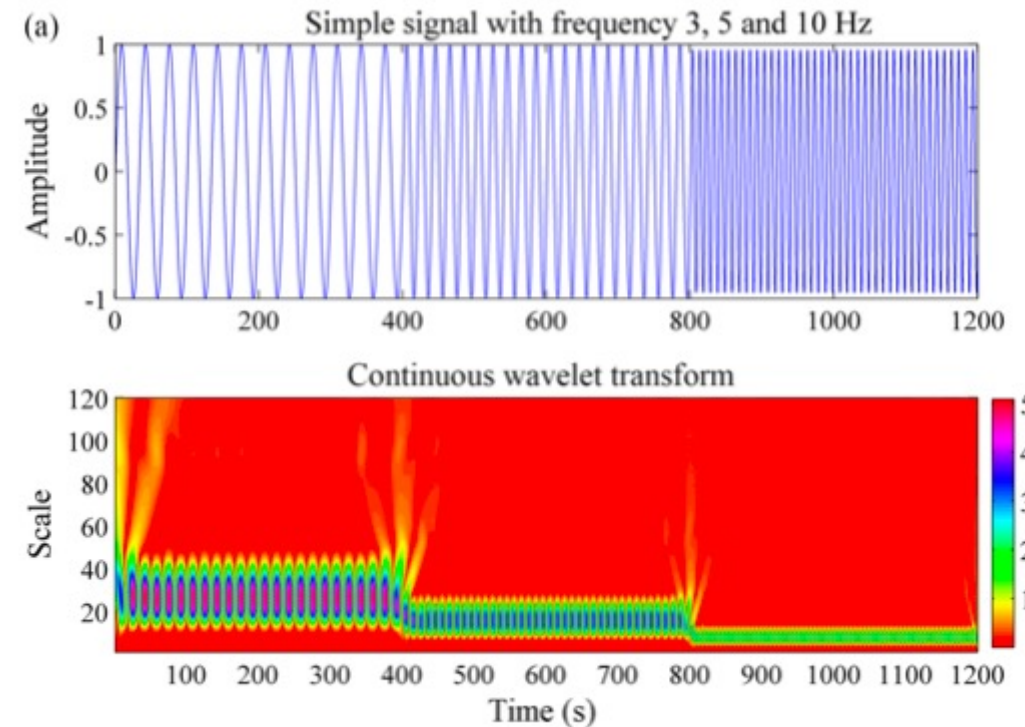
Algorithm Selection – way to approach your task

- Q : Is it a good way to compare the structures in time-series data and also take care of the shape of the structures in signal? Where are they and are they similar in the shape?

Wavelet Transformation

$$W(a, b) = \frac{1}{a^n} \int_{-\infty}^{\infty} f(x) \varphi\left(\frac{x-b}{a}\right) dx,$$

Explore how spectral features evolve over time, identify common patterns in two signals, and perform time-localized filtering

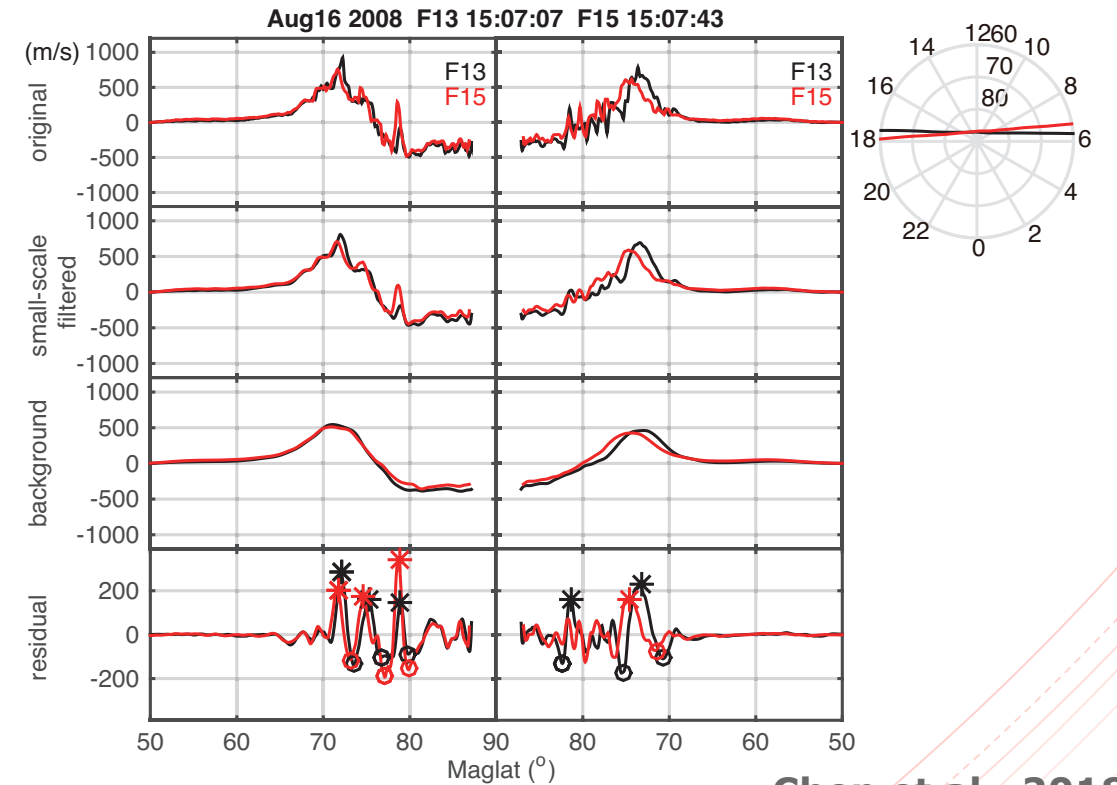
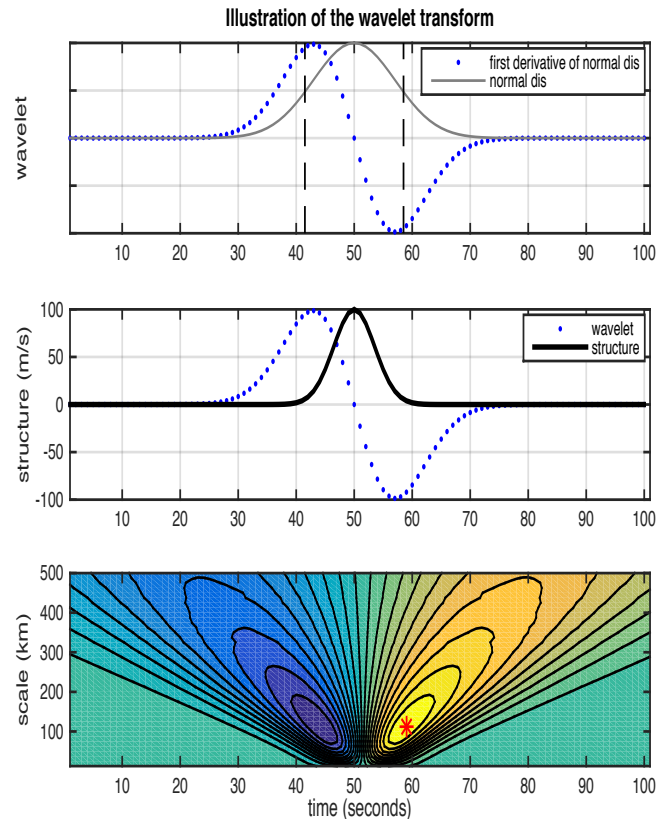


Upendra K. Singh et al., 2018

Algorithm Selection – way to approach your task

- Applied Wavelet Transformation – scaling and shifted!!!

Ex.



Chen et al., 2018

Algorithm Selection – way to approach your task

Machine learning

→ Supervised

- Each observation must be labeled with predetermined answer

$$Y=f(x)$$

- Learning mapping function from the input (x) to the output (Y)
- Often used as advanced form of predictive modeling
- Classification : categorical data
- Regression : numerical data

→ Unsupervised

- Observation has no predetermined labels

$$? = \{ x_j \}$$

- Algorithm learn patterns from data without supervision
- Often used as automated data extraction and feature discovering
- Clustering : inherent grouping

Algorithm Selection – way to approach your task

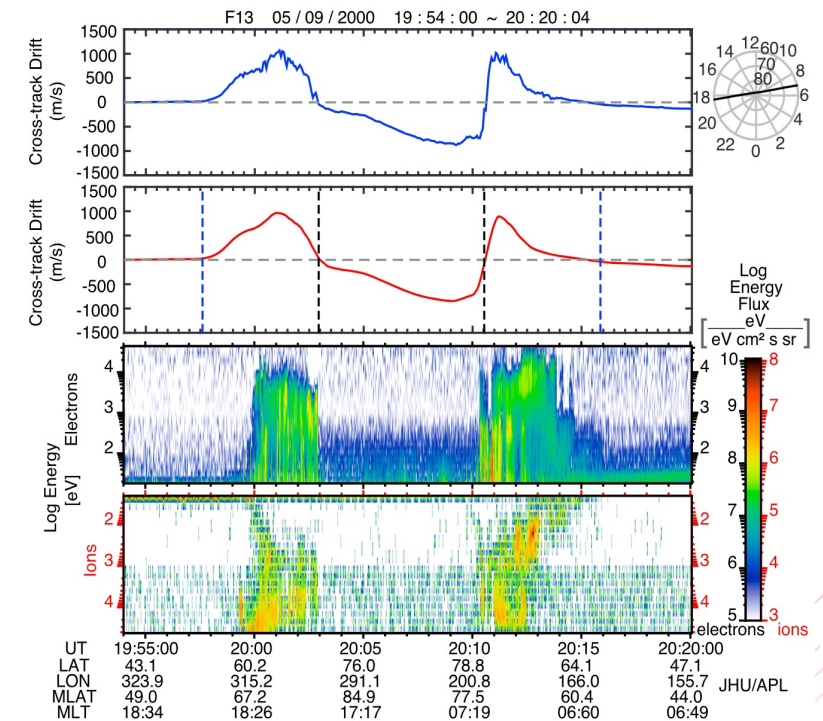
→ Unsupervised

■ Clustering : inherent grouping

- K-medoid : commonly used in domains that require robustness to outlier data, **arbitrary distance metrics**, or ones for which the **mean or median does not have a clear definition**
- The center of the subset is a member of the subset, called a medoid.

Ex. DMSP Satellite flow measurements

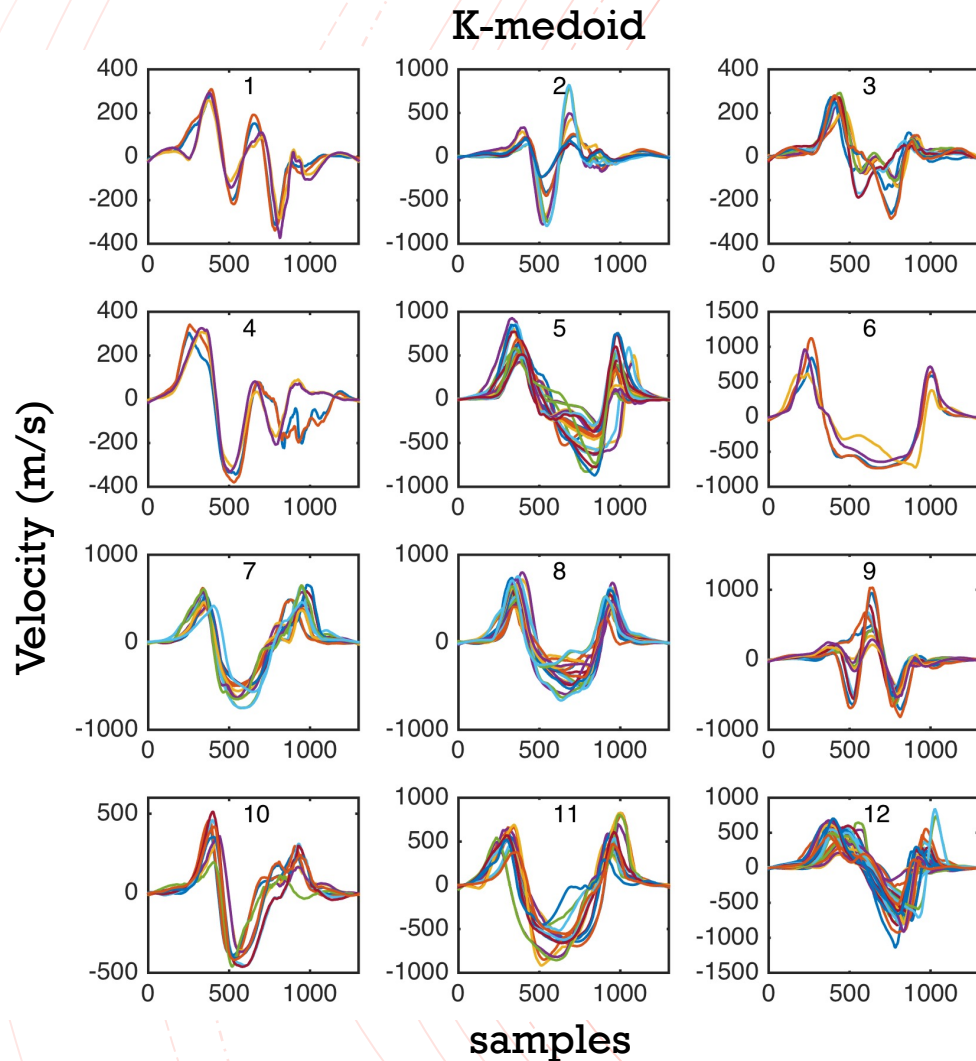
Group into categories based on the plasma flow profile



Algorithm Selection – way to approach your task

Ex. DMSP Satellite flow measurements

Group into categories based on the plasma flow profile

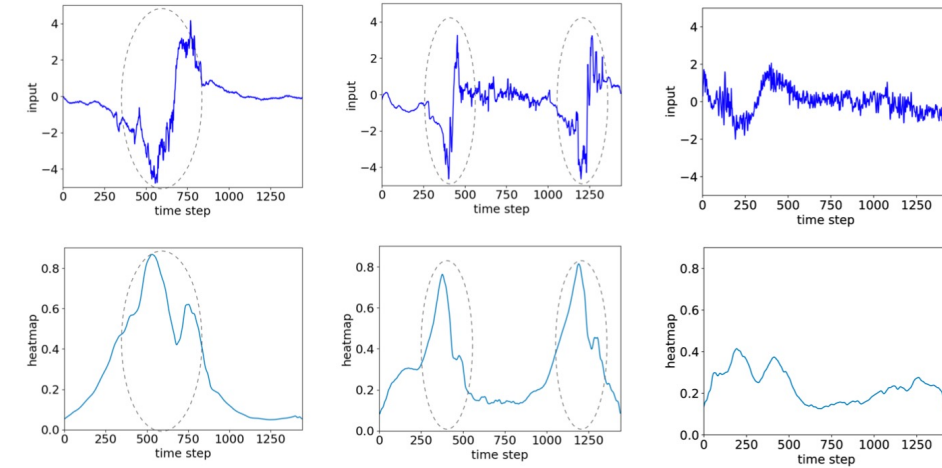


- Divide a set of observations into k clusters so that the subsets minimize the sum of *distances* between measurements and cluster medoids
- *Distance* : the similarity or dissimilarity
- Classified based on the major difference in profile and magnitude
- Process involves random draws from input dataset to initiate the clustering process

Wrap-up

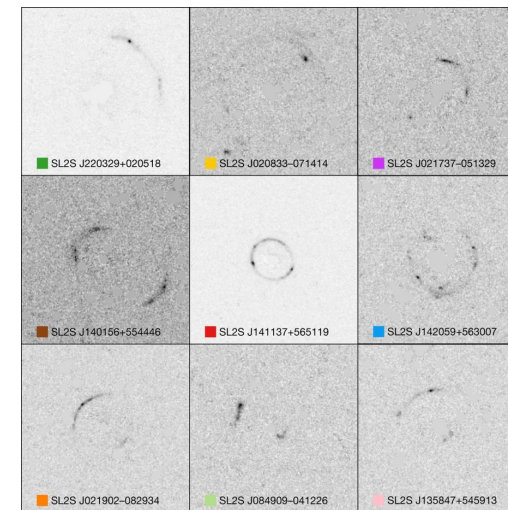
- Data quality control
- Feature or event detection
- Pre-processing for building model
- Predictive model/time-series forecasts
- *Unbiased data source*
- *Wisely choose algorithms*
- *Computational expensive*

FTE from MMS magnetometer data



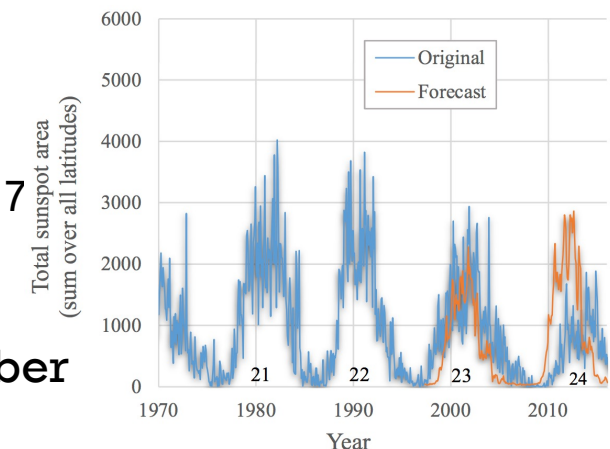
Naveen Sai Madiraju et al. 2018

Strong gravitational lensing images detection

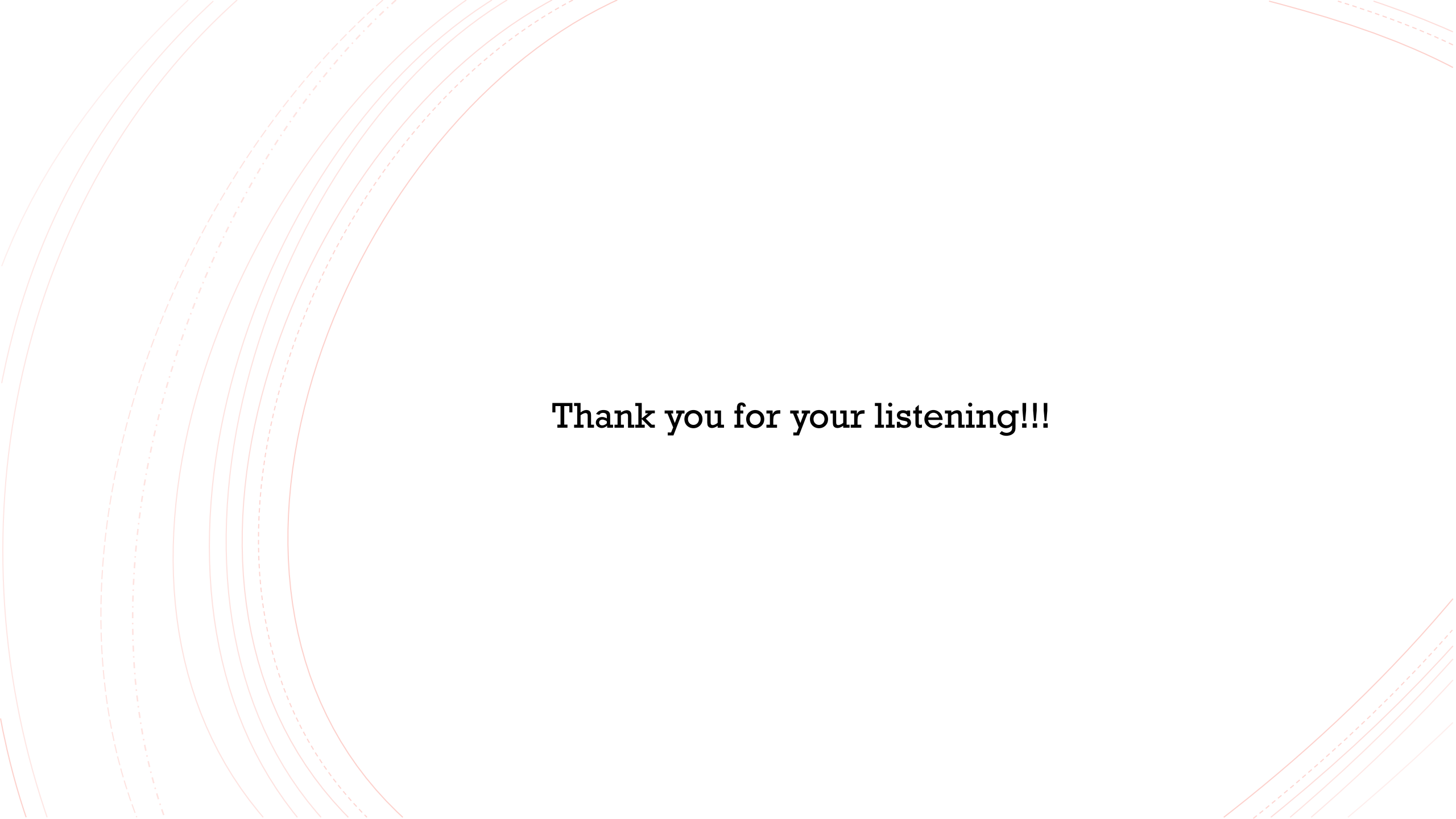


Yashar D. Hezaveh et al. 2017

Sun spot number prediction



Eurico Covas et al. 2019

The background features several concentric, curved lines in a light red or pinkish hue. These lines are solid on the left side and become dashed as they curve towards the right. The overall effect is a sense of motion or a stylized wave pattern.

Thank you for your listening!!!