Big Data Science

Course: 18-788

Patrick McSharry
patrick@mcsharry.net
www.mcsharry.net
Twitter: @patrickmcsharry

Spring 2023

ICT Center of Excellence
Carnegie Mellon University

Big Data Science

WEEK 4A

Assignment 2 recap

- Define dependent variable has many options and as a researcher, this decision needs to be made (despite there being many correct options).
- Simple return: $r(t) = p(t)/p(t-1)-1$
- Log return: $\log(p(t)/p(t-1))$
- Risk premium is the excess return: remove risk free return

Assignment 2 recap

- Univariate analysis – correlation of x(t) and r(t+1) to understand relevance of each variable
- Emphasis was placed on using a rolling window model (updating model parameters with each new month)
- Rolling should be better than using a fixed training dataset as we would not incorporate most recent information
- This can be applied to any model specification (selection of variables) or model structure (linear or nonlinear)

Rolling window

- Forecast for dependent variable y at future time t+1 using information X available at t:

$$y^*(t+1) = F[X(t),a(t)]$$

- Simplest static approach is to train the model F by only considering data up to time $T<=t$ so that the model parameters are fixed $a = a(T)$.
- Rolling approach means that a(t) is continuously updated over time.

Model comparison

- In the end, it makes sense to have a hierarchy of models so that we can study the forecast performance to understand the benefits of:
- (1) using each variable separately;
- (2) a joint model consisting of all variables;
- (3) selecting a subset of variables;
- (4) linear models;
- (5) nonlinear models.

Course outline

Week	Lecture A	Lecture B
1	Weather & agriculture	Climate change
2	Climate scenarios	Catastrophe models
3	Social trends	Finance
4	Sentiment analysis	Health
5	Telemedicine	Mobile data
6	Data4Dev	Socioeconomic status

Today's Lecture

No.	Activity	Description	Time
1	Challenge	The meaning of words	10
2	Discussion	Quantifying words	10
3	Case study	Twitter and S&P500	10
4	Analysis	Twitter sentiment	20
5	Demo	Tweets and floods	20
6	Q&A	Questions and feedback	10

Meaning of words

- In 1957, Charles Osgood used semantic differential to classify words.
- Semantic differential is a type of a rating scale designed to measure the connotative (suggestive significance) meaning of objects, events, and concepts.
- The connotations are used to derive the attitude towards the given object, event or concept.
- The respondent is asked to choose where his or her position lies, on a scale between two bipolar adjectives (for example: "Adequate-Inadequate", "Good-Evil" or "Valuable-Worthless").
- Semantic differentials can be used to measure opinions, attitudes and values on a psychometrically controlled scale.

Osgood, C.E., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning.* Urbana, IL: University of Illinois Press.

Semantic differential rating scale

Bad						Good
Slow						Fast
Small						Large
Closed						Open
Old						Young
Unfriendly						Friendly
Unfair						Fair
Easy						Difficult
Firm						Flexible
Numb						Sensitive
Casual						Formal
Vague						Clear

A scale for describing a shop

Clean						Dirty
Bright						Dark
Low quality						High quality
Conservative						Innovative
Inconvenient						Convenient

Osgood’s three factors

- Osgood performed a factor analysis of large collections of semantic differential scales and found three recurring attitudes that people use to evaluate words and phrases: **evaluation**, **potency**, and **activity**.
- Evaluation loads highest on the adjective pair 'good-bad'.
- The 'strong-weak' adjective pair defines the potency factor.
- Adjective pair 'active-passive' defines the activity factor.
- These three dimensions of affective meaning were found to be cross-cultural universals in a study of dozens of cultures.

Quantifying words

- One approach to sentiment analysis is to simply match words with an appropriate dictionary of words.
- In this way, sentiment variables can be easily produced by keyword matching.
- One popular word list is the Harvard IV-4 categories.

Database	Positive	Negative	Strong	Weak	Active	Passive
Number	1045	1160	1902	755	284	911

http://www.wjh.harvard.edu/~inquirer/homecat.htm

Poll

- Which techniques or applications might be useful for measuring sentiment?

- **Slido.com #649 792**

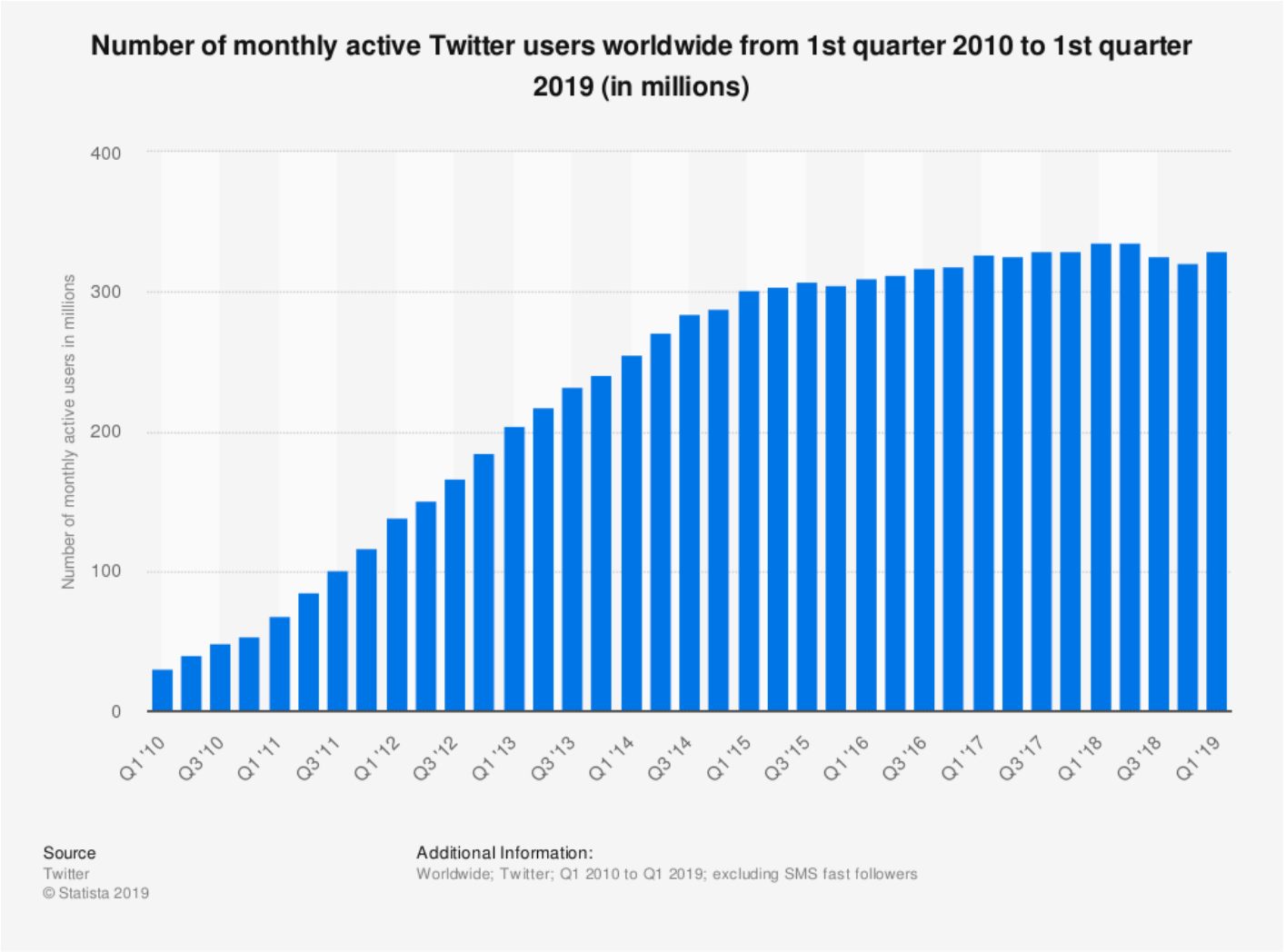
Types of sentiment analysis

- Emoticons
- Matching the words and emoticons with each tweet
- Linguistic analysis packages:
- OpinionFinder, LingPipe, Google POMS
- These may be slow and difficult to customize for different applications

Twitter

- Launched July 2006
- Micro-blog up to 140 characters: subject oriented
- Extended to 280 characters from November, 2017
- Individual experiences with real-time events; public opinion
- Real time stream / keyword search API available
- Gardenhose (up to 5% sample of the tweet stream)
- Firehose (20%, 30% or all)
- User location information
- Web user: registered location information in the profile
- Smart phone: actual geographic coordinate
- Follower/followee; tweet to (@); retweet (re); keyword (#);
- Directed social networks, not reciprocal (Facebook/MySpace)

Twitter users worldwide



There are 330 million monthly active **users** and 145 million daily active **users** on **Twitter**. 63 percent of all **Twitter users worldwide** are between 35 and 65.

Literature Review

- Detection
  - Earthquake detection (Kalman/particle filter)
  - Influenza epidemics (regression; moving average; Bayesian approach)
  - Online service outage (exponential smoothing), hot topic detection
- Online tools: education, communication, collaboration, Twitter search engine, news/rss feed recommendation, finding influentials
- Information diffusion: Iranian presidential election (using retweets), popularity of celebrities (Google PageRank vs Twitter), URL mentions
- Sentiment analysis: brand marketing
- Social network analysis: geographical and topological structures

Poll

- Do you think twitter can be used to predict the stockmarket?
  a) Yes
  b) No

- **Slido.com #649 792**

S&P 500 Stocks and Twitter Data

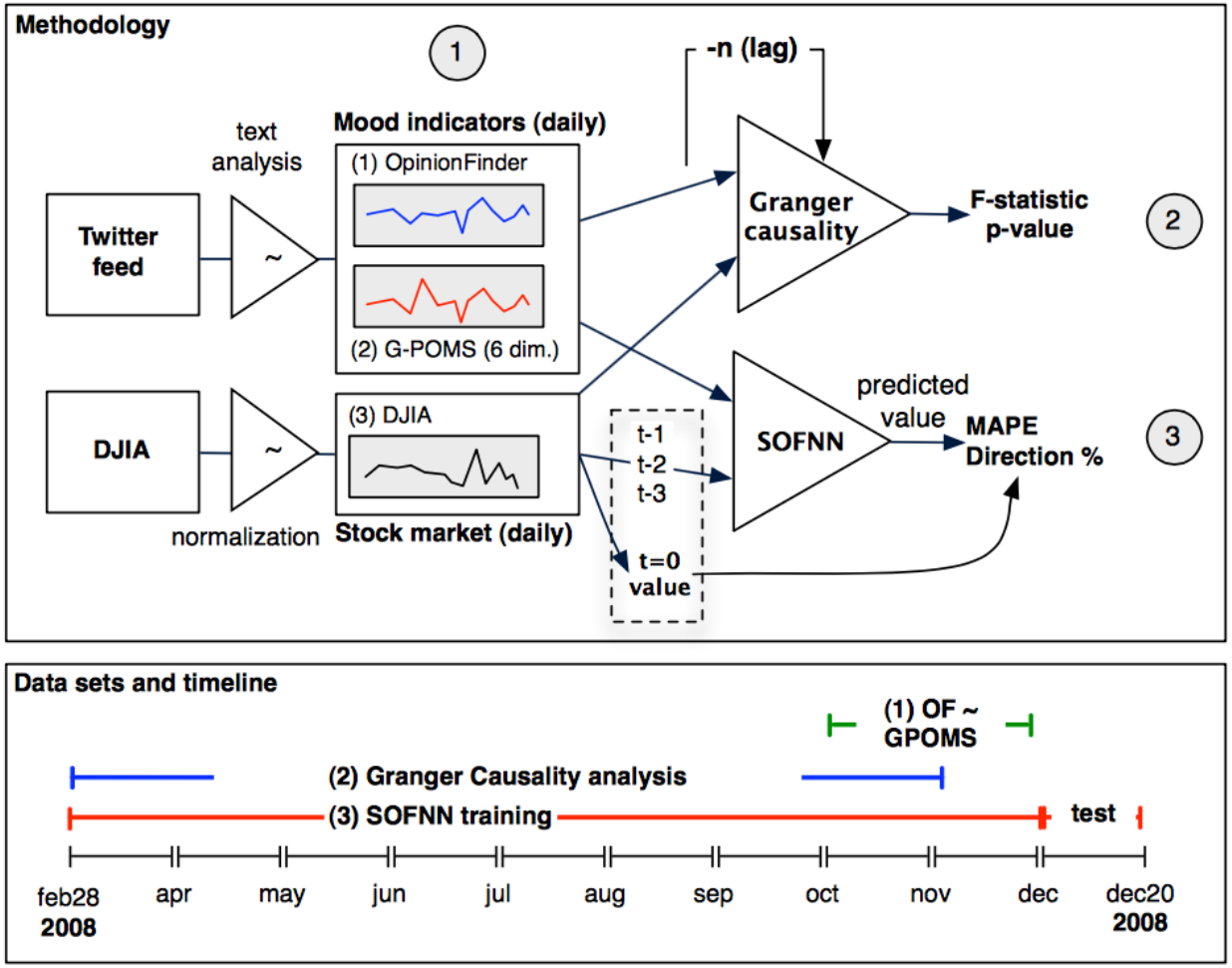
- Mao et al. look at number of tweets as a means of predicting the direction of the stockmarket.
- To study whether the daily number of tweets that mention Standard & Poor 500 (S&P 500) stocks is correlated with S&P 500 stock indicators (stock price and traded volume) at three different levels, from the stock market to industry sector and individual company stocks.
- Found a strong correlation between the number of daily tweets and daily traded volume for Apple.
- Accuracy in predicting the direction of change in daily traded volume is only 52%.
- They found a prediction accuracy of 68% for direction of the S&P500 index.
- Prediction accuracy of direction at sector level is also 68%.

Mao et al. Correlating S&P 500 Stocks with Twitter Data

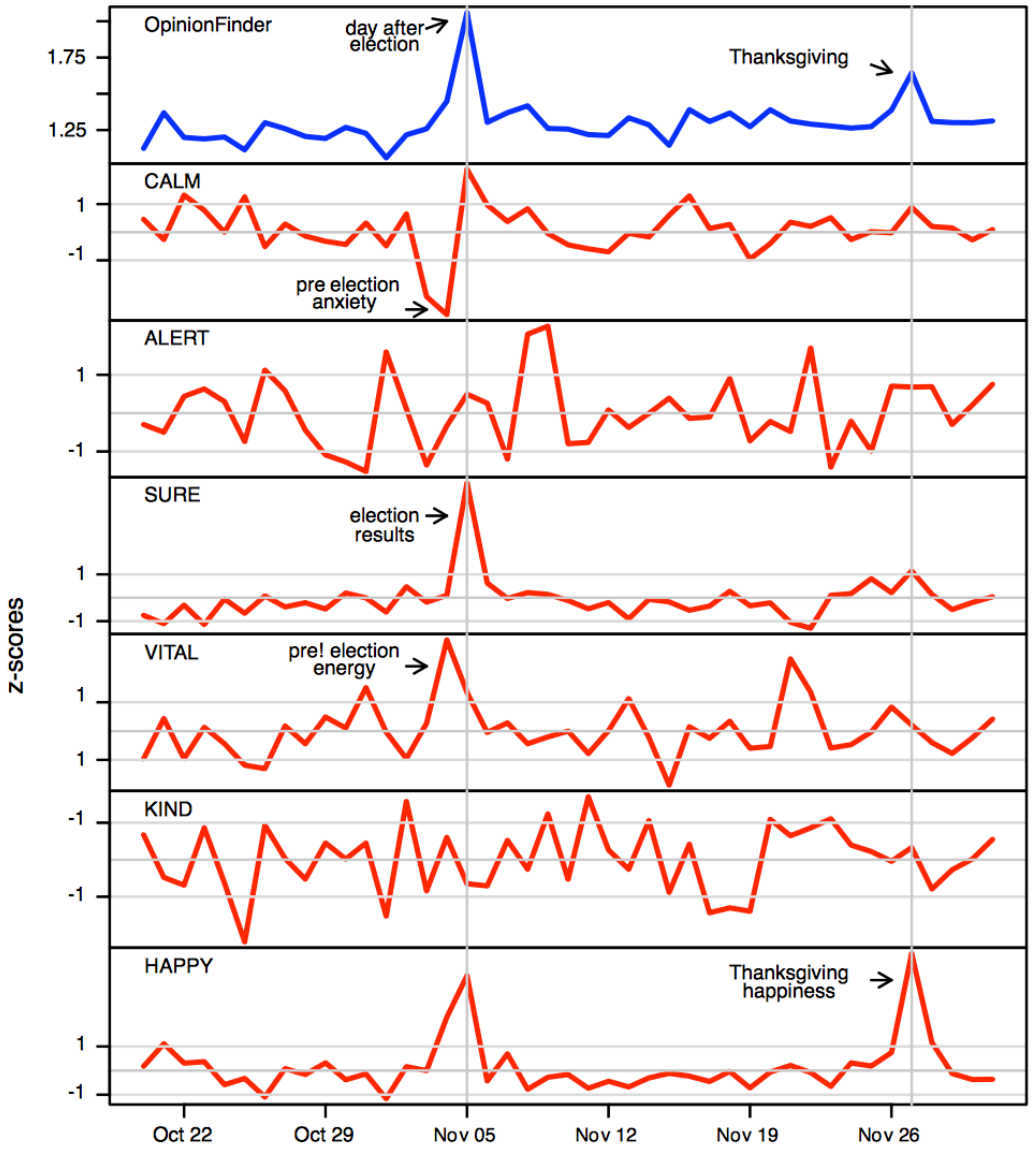
Twitter and stockmarket

- Bollen et al. (2010) studied impact of twitter on the stockmarket.
- Content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy).
- Accuracy of 87.6% in predicting the daily up and down changes in the closing values of the DJIA.
- Reduction of the Mean Average Percentage Error by more than 6%.
- OpinionFinder (lag 1 day) and Calm (lags 2-6 days) significant.

Bollen et al. methodology



Election and thanksgiving



Twitter and word frequency

- The most frequent words
  - Positive

:D	968
FREE	935
LIKE	918
:)	847
PRO	694
LOVE	597
GOOD	515
KNOW	503
HOME	486
BUY	405
MAIN	393
GREAT	369
BACK	334
BEST	334

- Negative

:(	474
SHOW	425
NEED	418
TOO	348
SERVICE	288
CHARGE	287
DRIVE	265
LOW	249
ALARM	247
BEAT	236
HARD	202
ERROR	200
MISS	195
SORRY	189
WAIT	176

  - Different application area might require different list of sentiment words

Twitter and AFINN

- AFINN database constructed by Finn Årup Nielsen.
- The benefit of the AFINN word list is that the words were collected and labelled for the purpose of monitoring tweets and that it includes the magnitude of the each sentiment word.
- Each sentiment word from the AFINN has a value between -5 (strongest negative) and 5 (strongest positive).
- By using this, strong positive or negative sentiment words such as 'fantastic' and 'worst' could be distinguished from weak sentiment words such 'good' and 'mediocre'.

http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

Poll

- Is it likely that Twitter may be best suited to predicting consumer behaviour?

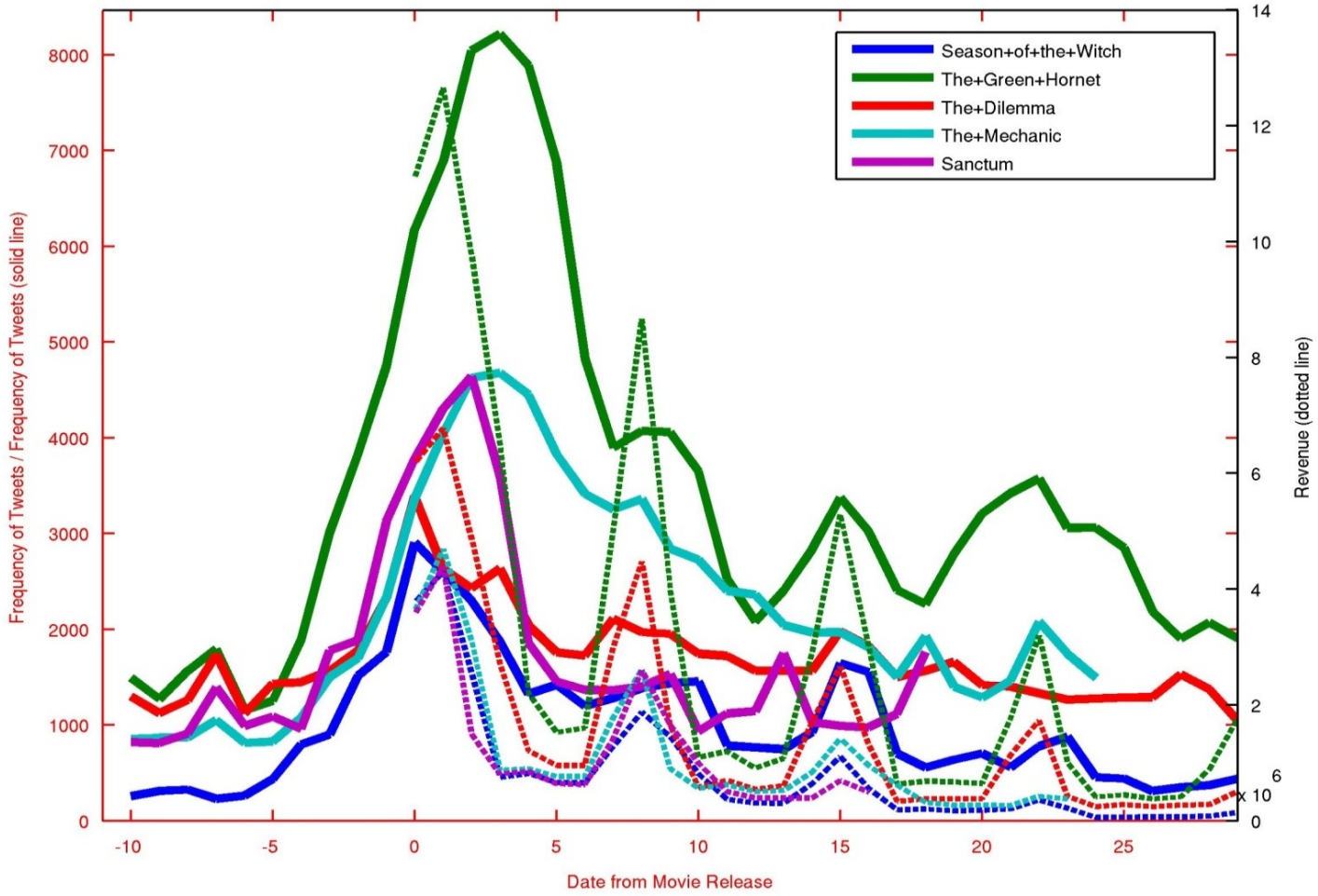
  a) Yes
  b) No

- **Slido.com #649 792**

Forecasting box office revenue

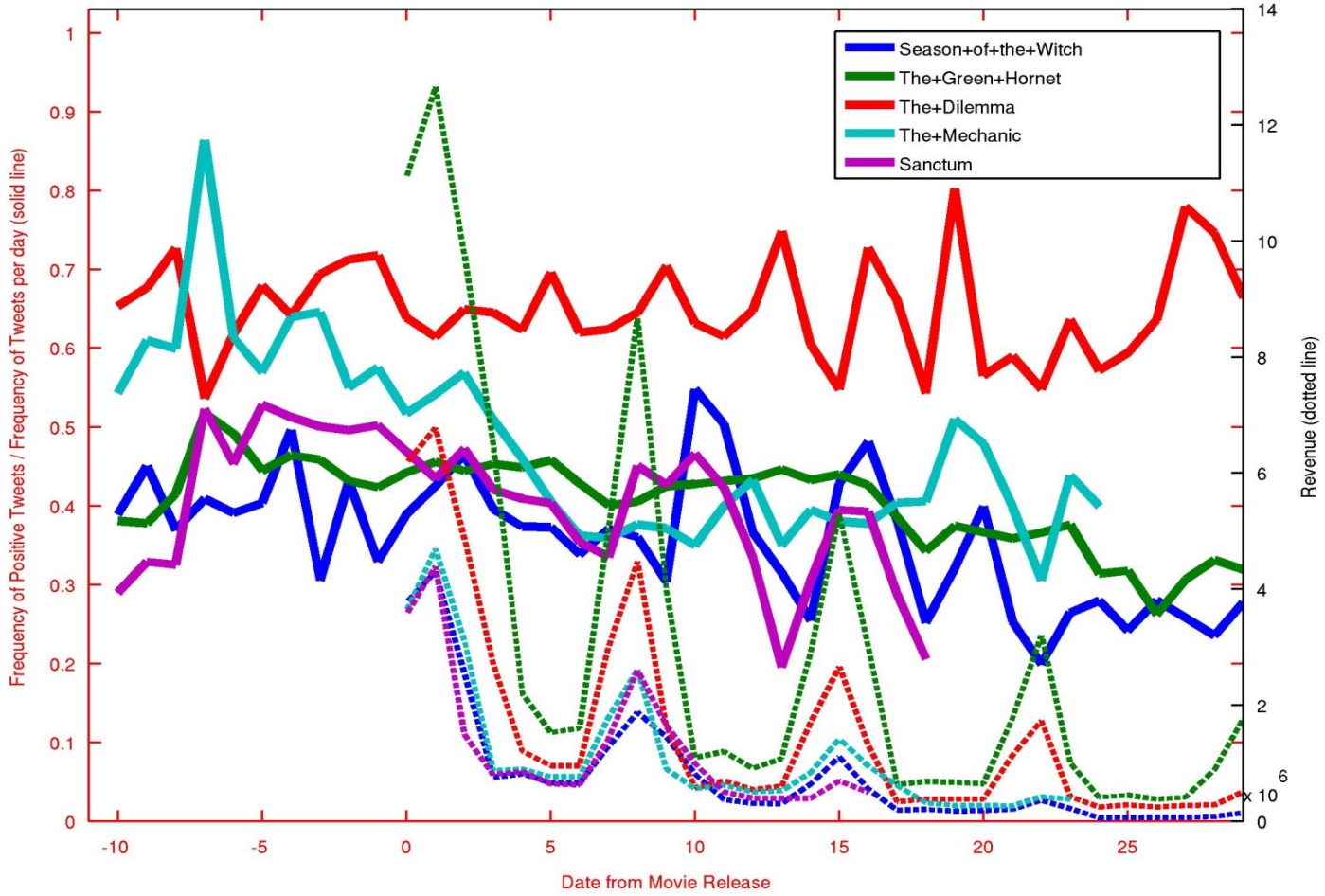
- Dependent variable:
  - total revenue of the first Friday, Saturday and Sunday
- Independent variables:
  - theater_openings
  - frequency
  - positive_frequency
  - negative_frequency
  - positive_over_negative

Frequency of mentions vs Revenue



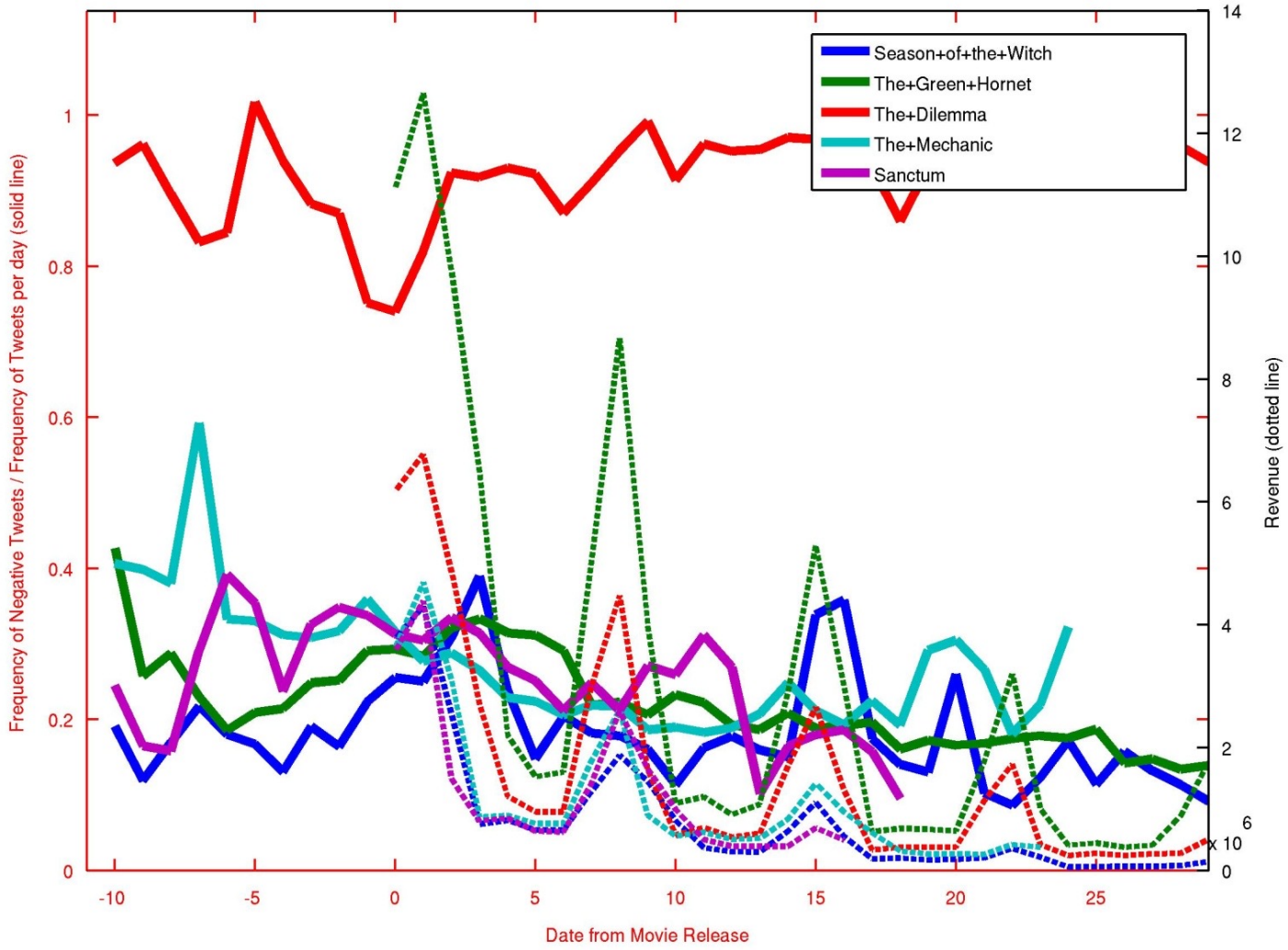
- Tweets between d-4 to d-1 vs Revenue between d-0 to d-2
- Peak tweets in the first week?

Positive Sentiment vs Revenue



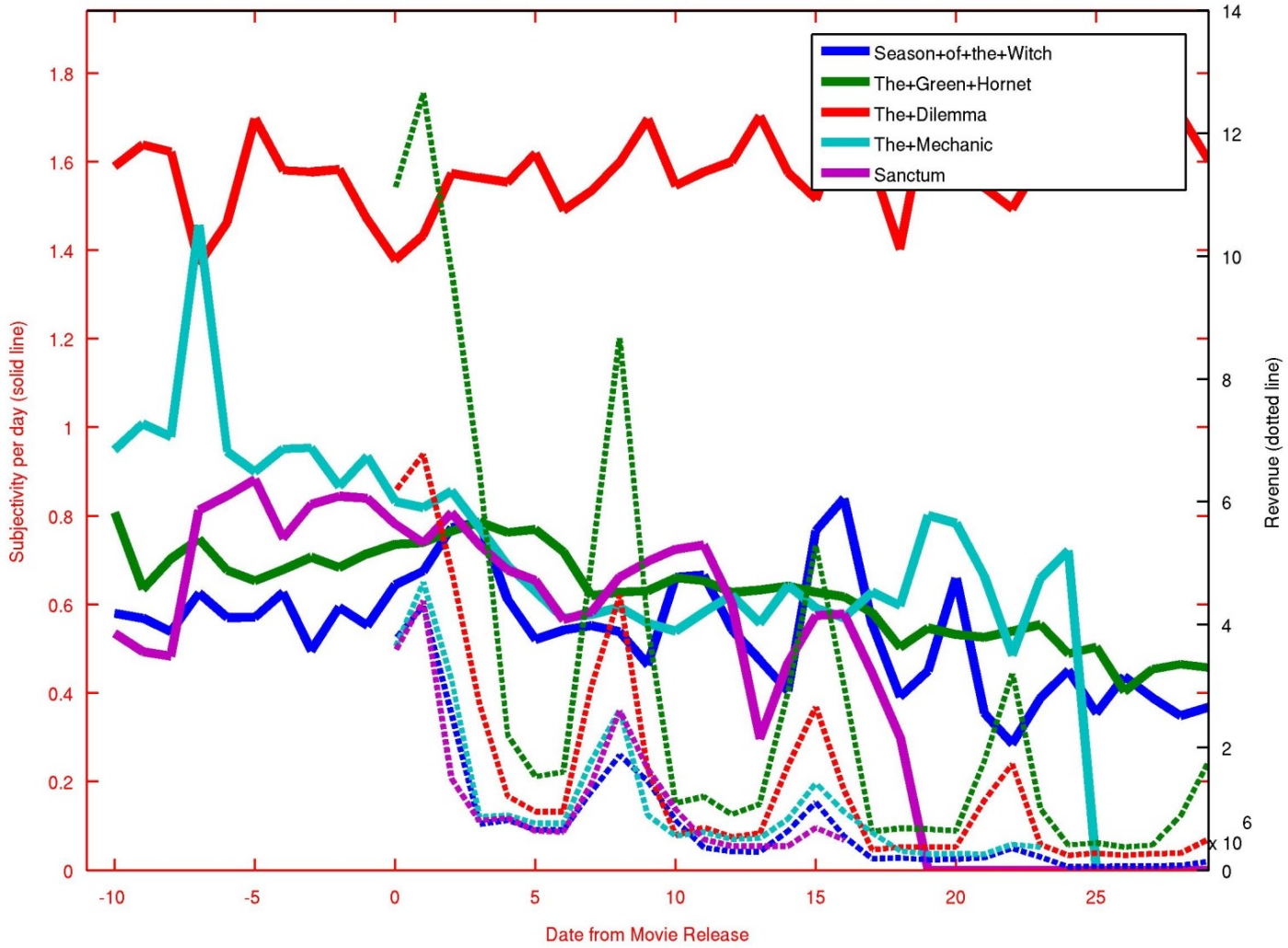
- Positive / frequency of tweets: decreasing after d-0
- Any Competing movies?

Negative sentiment vs Revenue



- Negative / frequency of tweets: decreasing after d-0

Subjectivity vs Revenue



- Subjectivity = (positive+negative)/frequency: neutralise after d-0

Movies and twitter

- Sentiment analysis based on tweets suggests that more extreme sentiment has more impact
- the more negative the tweets about a movie, the higher its revenue
- Data collected between 20 Dec, 2010 and 8 Jul, 2011
- tweets collected via Twitter Streaming API and Search API
- box office revenues and theater opening counts collected from The Numbers (http://the-numbers.com)
- The power of Twitter on predicting box office revenues. Jooyoung & McSharry, pp. 2-22 (2012)

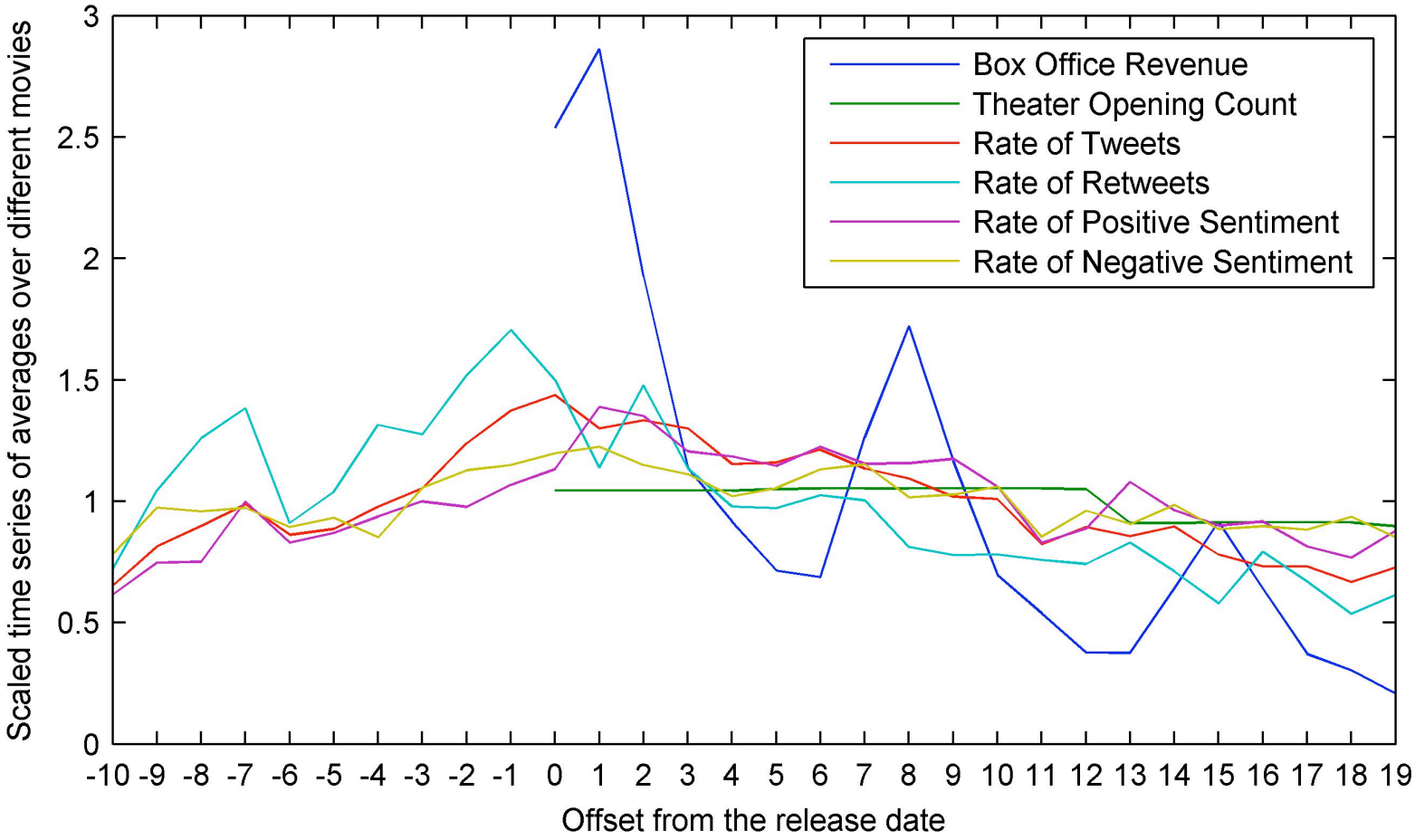
Box office revenue

- Simple bag-of-words approach (AFINN)
- Study shows that simple keyword matching with extreme sentiment words is predictive of revenue
- Demonstrates that increasingly extreme (negative) sentiment drives movie revenues
- “the expression ‘any publicity is good publicity’ holds true for movies”

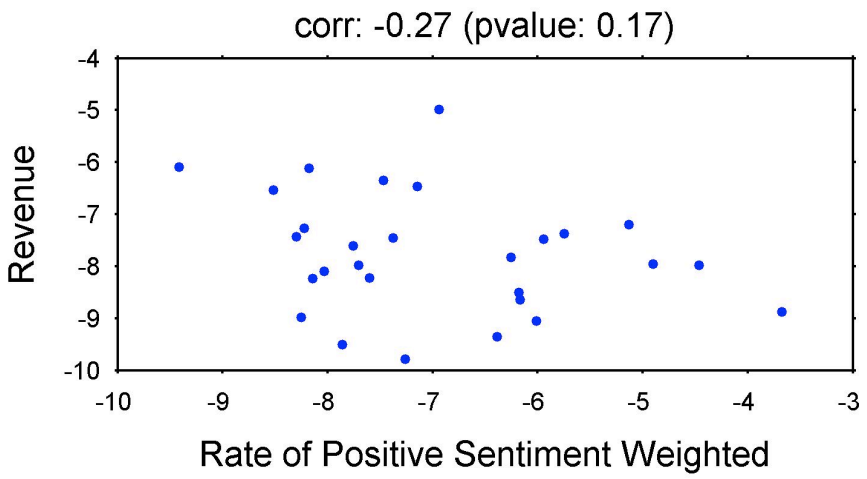
Movies

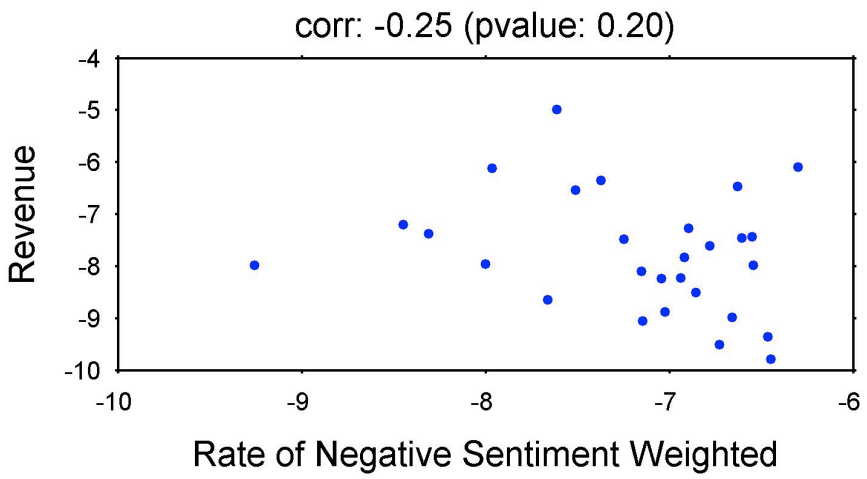
Movie	Opening date	Opening day of week
Sanctum	22/12/2010	Wed
Monte Carlo	22/12/2010	Wed
Season of the Witch	25/12/2010	Sat
The Adjustment Bureau	07/01/2011	Fri
Sucker Punch	14/01/2011	Fri
Source Code	14/01/2011	Fri
African Cats	28/01/2011	Fri
Rango	04/02/2011	Fri
Cars 2	04/02/2011	Fri
The Mechanic	11/02/2011	Fri
Hop	11/02/2011	Fri
The Green Hornet	18/02/2011	Fri
The Zookeeper	18/02/2011	Fri
I am Number Four	04/03/2011	Fri
Paul	04/03/2011	Fri
The Roommate	11/03/2011	Fri
Unknown	11/03/2011	Fri
The Eagle	18/03/2011	Fri
Red Riding Hood	25/03/2011	Fri
The Dilemma	01/04/2011	Fri
Mars Needs Moms	01/04/2011	Fri
Little Fockers	22/04/2011	Fri
True Grit	24/06/2011	Fri
Gullivers Travels	24/06/2011	Fri
The Rite	01/07/2011	Fri
Just Go With It	01/07/2011	Fri
Larry Crowne	08/07/2011	Fri

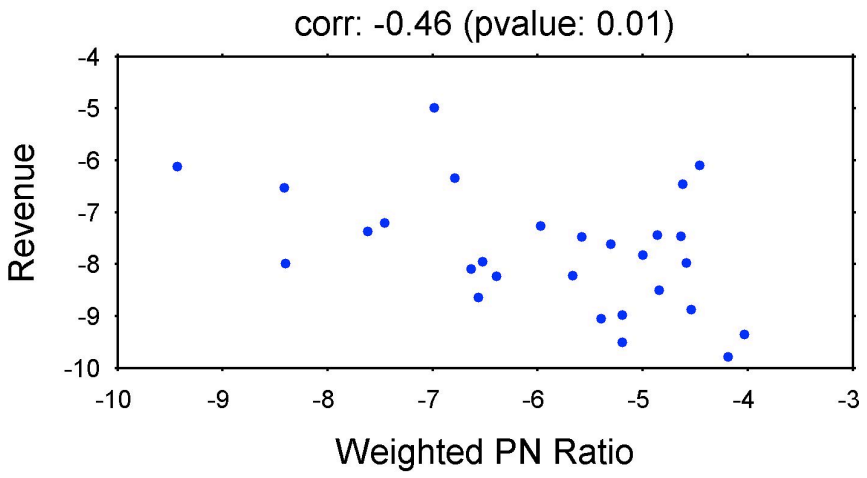
Box office revenue



Scatter plots of sentiment variables(*t*-1) against box office sales revenue(*t*).







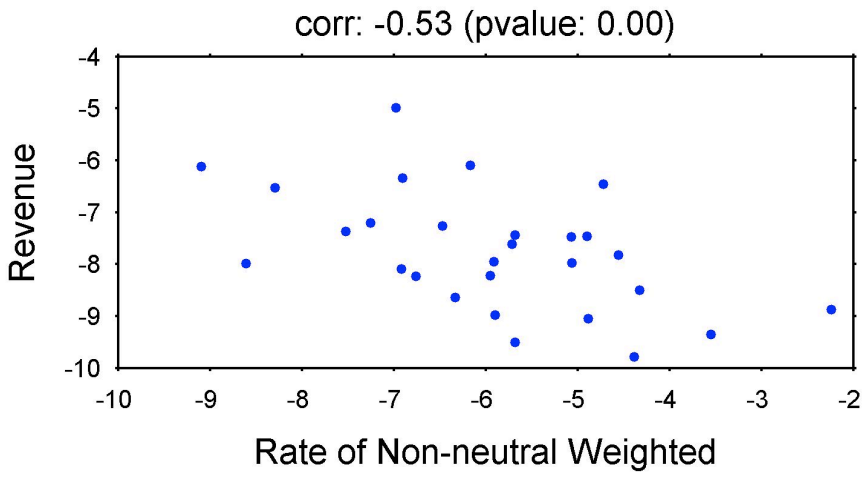


Table 2. MAPEs of regression models for the first weekend box office revenue prediction. Smaller values are better.

Variables used in regressions	1st weekend	2nd weekend
$\text{TheaterOpenCount}_{t-1}$	0.2878	0.2527
RateTweets_{t-1}	0.4592	0.4931
$\text{RatePositive}_{t-1}$	0.5034	0.5426
$\text{RateNegative}_{t-1}$	0.4825	0.5258
PNRatio_{t-1}	0.4274	0.5135
$\text{RateNonneutral}_{t-1}$	0.5046	0.5226
$\text{RatePositiveWgt}_{t-1}$	0.4957	0.5362
$\text{RateNegativeWgt}_{t-1}$	0.5012	0.5857
PNRatioWgt_{t-1}	0.4424	0.5162
$\text{RateNonneutralWgt}_{t-1}$	0.4258	0.5201
$\text{TheaterOpenCount}_{t} + \text{RateTweets}_{t-1}$	0.2813	0.2506
$\text{TheaterOpenCount}_{t} + \text{RatePositive}_{t-1}$	0.2981	0.2655
$\text{TheaterOpenCount}_{t} + \text{RateNegative}_{t-1}$	0.2926	0.2571
$\text{TheaterOpenCount}_{t} + \text{PNRatio}_{t-1}$	0.2878	0.2636
$\text{TheaterOpenCount}_{t} + \text{RateNonneutral}_{t-1}$	0.2967	0.2610
$\text{TheaterOpenCount}_{t} + \text{RatePositiveWgt}_{t-1}$	0.2924	0.2660
$\text{TheaterOpenCount}_{t} + \text{RateNegativeWgt}_{t-1}$	0.3005	0.2768
$\text{TheaterOpenCount}_{t} + \text{PNRatioWgt}_{t-1}$	0.2941	0.2484
$\text{TheaterOpenCount}_{t} + \text{RateNonneutralWgt}_{t-1}$	0.2797	0.2517
$\text{TheaterOpenCount}_{t} + \text{RateTweets}_{t-1} + \text{RatePositive}_{t-1}$	0.2924	0.2632
$\text{TheaterOpenCount}_{t} + \text{RateTweets}_{t-1} + \text{RateNegative}_{t-1}$	0.2883	0.2630
$\text{TheaterOpenCount}_{t} + \text{RateTweets}_{t-1} + \text{PNRatio}_{t-1}$	0.2666	0.2585
$\text{TheaterOpenCount}_{t} + \text{RateTweets}_{t-1} + \text{RateNonneutral}_{t-1}$	0.2921	0.2640
$\text{TheaterOpenCount}_{t} + \text{RateTweets}_{t-1} + \text{RatePositiveWgt}_{t-1}$	0.2905	0.2639
$\text{TheaterOpenCount}_{t} + \text{RateTweets}_{t-1} + \text{RateNegativeWgt}_{t-1}$	0.2908	0.2772
$\text{TheaterOpenCount}_{t} + \text{RateTweets}_{t-1} + \text{PNRatioWgt}_{t-1}$	0.2750	**0.2154**
$\text{TheaterOpenCount}_{t} + \text{RateTweets}_{t-1} + \text{RateNonneutralWgt}_{t-1}$	**0.2578**	0.2493

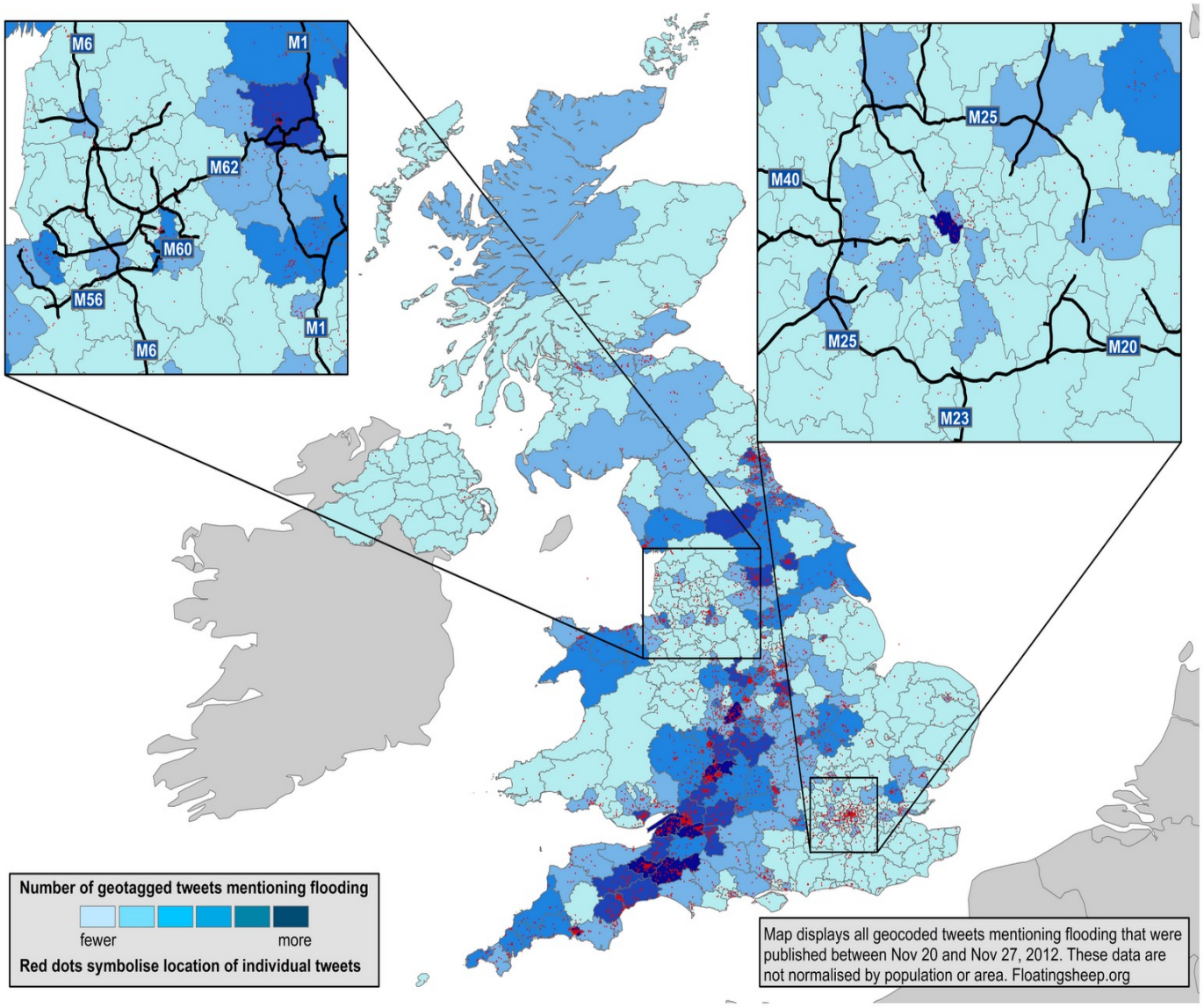
* MAPE is MAPE weighted averaged by the revenue of each movie.

Quiz

- Twitter is more likely to reflect
  a) Actual risks
  b) Future risks
  c) Emerging risks
  d) Perceived risks

- **Slido.com #649 792**

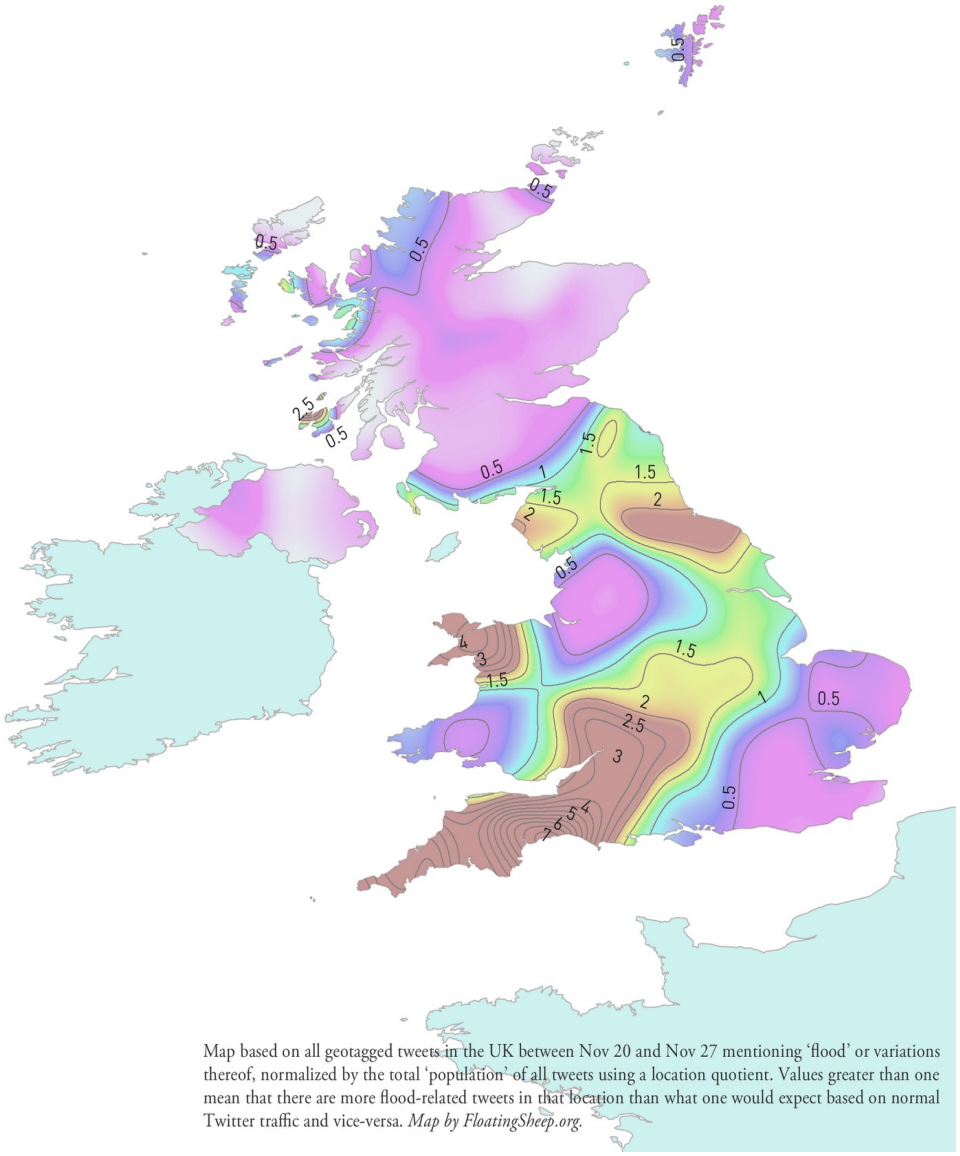
Flood tweets



Every geocoded tweet between Nov 20 and Nov 27, 2012 that mentioned the word "flood" (or variations like "flooded" or "flooding")

M. Graham. Oxford Internet Institute, University of Oxford. (www.floatingsheep.org)

Twitter flood map



Map based on all geotagged tweets in the UK between Nov 20 and Nov 27 mentioning 'flood' or variations thereof, normalized by the total 'population' of all tweets using a location quotient. Values greater than one mean that there are more flood-related tweets in that location than what one would expect based on normal Twitter traffic and vice-versa. *Map by FloatingSheep.org.*

Normalised data showing a location quotient where everything greater than 1 indicates that there are more tweets related to flooding than one would expect based on normal Twitter usage in that area

M. Graham. Oxford Internet Institute, University of Oxford. (www.floatingsheep.org)

Big Data Science

WEEK 4B

Course outline

Week	Lecture A	Lecture B
1	Weather & agriculture	Climate change
2	Climate scenarios	Catastrophe models
3	Social trends	Finance
4	Sentiment analysis	Health
5	Telemedicine	Mobile data
6	Data4Dev	Socioeconomic status

Today's Lecture

No.	Activity	Description	Time
1	Challenge	Big data and health	10
2	Discussion	Benchmarks	10
3	Case study	Linear or nonlinear	10
4	Analysis	Drug trials	20
5	Demo	Significance	20
6	Q&A	Questions and feedback	10

Poll

- What are the sources of data in healthcare?

- **Slido.com #091 208**

Healthcare

- Big data could transform the health-care sector, but the industry must undergo fundamental changes before stakeholders can capture its full value – McKinsey “Big data revolution in US healthcare”.
- Stakeholders have been opening their vast stores of health-care knowledge, including data from clinical trials and information on patients covered under public insurance programs.
- Now possible to collect and analyze information from multiple sources, e.g. hospitals, laboratories, and physician offices.

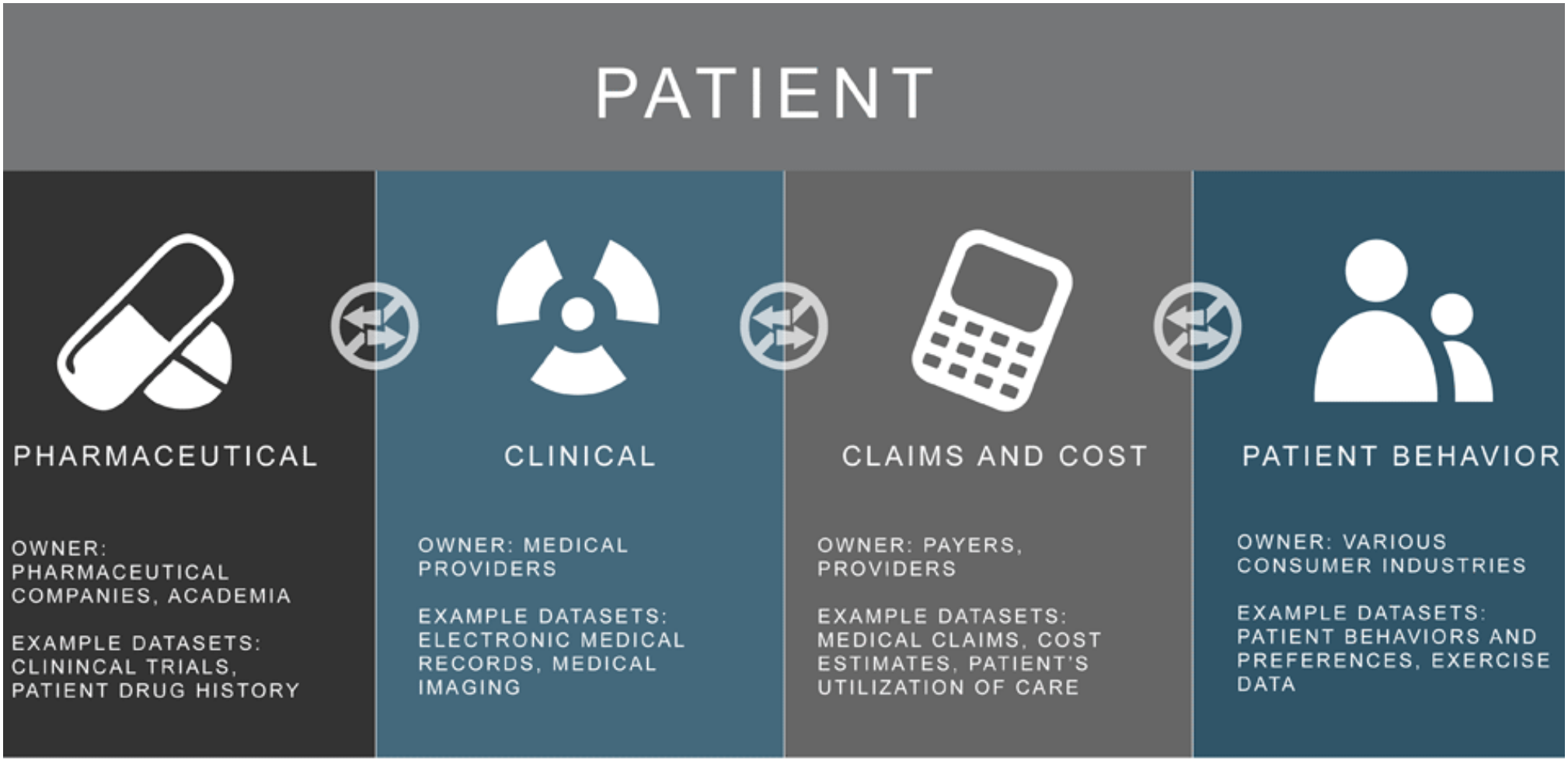
Opportunity and challenges

- According to McKinsey, the health care industry could potentially realize $300 billion in annual value by leveraging big data.
- However health data is unstructured.
- It is rarely standardized and often stored in legacy IT systems with incompatible formats.
- Moreover much of the information must be digitized and natural language processing is required to make sense made from the unstructured data.

Five value pathways

- **(1) Right living**: patients encouraged to play an active role in their own health.
- **(2) Right care**: patients should receive the most timely, appropriate treatment available.
- **(3) Right provider**: professionals treating patients must have strong performance records.
- **(4) Right value**: providers and payors should continually improve value while preserving or improving quality.
- **(5) Right innovation**: stakeholders focus on identifying new therapies and approaches to healthcare delivery.

Source of data



MAPR

www.healthworkscollective.com

Quiz

- Which of these variables is likely to change fastest during the day?

  a) Blood pressure
  b) Heart rate
  c) Temperature
  d) Weight

- **Slido.com #091 208**

Health variables

- Various medical variables can be measured by clinicians.
- Weight and height are relatively static.
- Others such as temperature, heart rate, blood pressure vary dramatically over time (biomedical signals).
- For example, heart rate changes as we lie down, sit, stand, walk, jog and run.

Medical variability

- For each variable, there will be a certain amount of natural variability.
- In order to detect early warnings of potential medical problems we need to establish normality and quantity the level of natural variability.
- This is typically done by studying a large group of people and defining a range based on the distribution of values observed.

Empirical decision support

- The big data revolution is providing data on both subjects with disorders and those that are healthy.
- This is a change to the normal paradigm where only data on those seeking healthcare has typically been available.
- This makes it possible not only to identify medical disorders using novelty detection but also to construct sophisticated models for quantifying severity.

Poll

- Which of the following would you recommend for providing medical advice?
  a) AI system
  b) Medical Doctor
  c) Medical Doctor supported by AI system

- **Slido.com #091 208**

IBM’s Watson Health

- Watson Health is powered by IBMs powerful Watson Analytics engine, which uses sophisticated machine learning algorithms based on natural-language processing.
- In theory users can ask it to perform complex analytical operations in simple human language, rather than having to understand code to communicate with the computer.
- It was originally (successfully) designed to become a champion at the game show *Jeopardy!* by answering questions asked of it in English.

Apple and IBM

- Apple and IBM formed a partnership which will allow iPhone and Apple Watch users to share data to IBM's Watson Health cloud-based healthcare analytics service.
- Potential for further advances in healthcare as IBM's Big Data-crunching engines gain access to real-time activity and biometric data from potentially millions of people who use Apple's devices around the world.

Healthcare Apps

- Smart devices linked to analytical applications such as the Apple/IBM service can also be used as remote monitoring devices, allowing doctors to offer intervention when they are alerted to warning signs by the app.
- IBM have estimated that almost 5 million people around the world are connected to their healthcare providers in this way.

Apple and Google team up to contact trace COVID-19

- Apple and Google are jointly developing technology to alert people if they have recently come into contact with others found to be infected with coronavirus (COVID-19).
- Their contact-tracing method would work by using a smartphone's Bluetooth signals to determine to whom the owner had recently been in proximity for long enough to have established contagion a risk.
- If one of those people later tested positive for the Covid-19 virus, a warning would be sent to the original handset owner.
- No GPS location data or personal information will be recorded, and cryptography specifications are available and along with an API.

https://www.bbc.co.uk/news/technology-52246319

AI and COVID-19

- An Israeli company, Diagnostic Robotics, analyzes the patient's clinical symptoms and underlying health status, generates a personalized, AI-based risk profile for COVID-19, in addition to providing next-step guidance.
- The information is delivered as "red flags" to health authorities, creating a "heat map" of coronavirus hot spots, which in turn helps medical services identify which regions need intensive care.
- Healthcare providers engage their clients with a simple symptoms questionnaire via a text message.
- This remote screening process, illustrated by a high-resolution epidemiologic heat map, enables health officials to gain a continuous, real-time, and large-scale assessment of the virus' spreading rate.
- By referring to such a heat map, decision makers can know immediately which geographic areas warrant immediate attention.

https://www.forbes.com/sites/startupnationcentral/2020/04/13/israeli-startups-artificial-intelligence-covid19-coronavirus/#70c077d45677

Big data and health

- Statistical learning is concerned with finding a relationship between explanatory variables (or features) and a response variable.
- In a clinical setting, we wish to determine the relationship between the observed characteristics of a subject and the diagnosis or clinical outcome.
- In most medical settings, the diagnosis or clinical outcome can take a small range of possible values.
- For example, the final diagnosis of a clinician may simply be a “yes” or “no” to a question (e.g. whether a subject has cancer).

Classification and regression

- This idea can be generalized to a more general setting, where for example a number of explanatory variables are used to assign subjects to different pathologies.
- The possible values that the response variable can have are simply known as *categories* or *classes*.
- When the response variable can take any number of finite classes the problem of predicting the response variable is known as *classification*.
- When the response variable can take any real value (any possible number from $-\infty$ to ∞), the problem is known as *regression*.

Clinical questions

- How can we associate **X** and **y**? That is, what is the relationship between the explanatory variables and the response variable
- Is there a convenient way to estimate the response variable when presented with the explanatory variables of a subject?
- Which of the explanatory variables are useful in actually determining the response variable?
- What are the relationships between the explanatory variables?
- Is it possible that some of the explanatory variables are redundant and need not be computed?

Data structure

- Data for medical applications can be represented as
- X is an NxM data matrix where each row contains the M explanatory variables for each of the N subjects
- For example, these rows might be age, gender…
- Y is an Nx1 vector containing the response variable.
- Classification is concerned with responses such as healthy (y=0) and disorder (y=1).

Parametric and non-parametric

- The challenge is to find a model f such that

  $Y \sim f(X)$

- **Parametric**: imposes a structure on the function form.
- A multivariate linear model is an example of a parametric model: $y_n = a_1 x_{n1} + \ldots + a_M x_{nM}$
- **Non-parametric**: the data itself determines the model structure.
- A decision tree is an example of a non-parametric approach.

Poll

- How much improvement would you expect when using nonlinear models instead of linear models for classifying medical disorders?
  a) 1%
  b) 5%
  c) 10%
  d) 20%
  e) 40%

- **Slido.com #091 208**

Random Forests

- Random Forests (RF) is a powerful *non-parametric* classifier, which can provide a model where the explanatory variables combine *non-linearly* to estimate the response variable.
- It is constructed by combining many base experts, the *trees* (by default 500 trees), and then uses majority voting from the trees to decide on the final output.
- The way trees are built is similar to the approach taken by the clinician when making a decision.
- There are successive binary splits of the data before reaching a conclusion on how to make a classification.
- Effectively, the tree partitions the data based on a single feature at each decision point (node).
- This approach could be compared to the method that clinicians use in deciding the course of optimal treatment for a patient.

Proposed methodology

- Apply statistical tests to the data and explore visually (correlations, density plots and scatter plots).
- Select relevant features.
- Consider both standard classifiers (logistic regression) and more complicated nonlinear models.
- Obtain results (classification accuracy) using ten-fold cross validation.

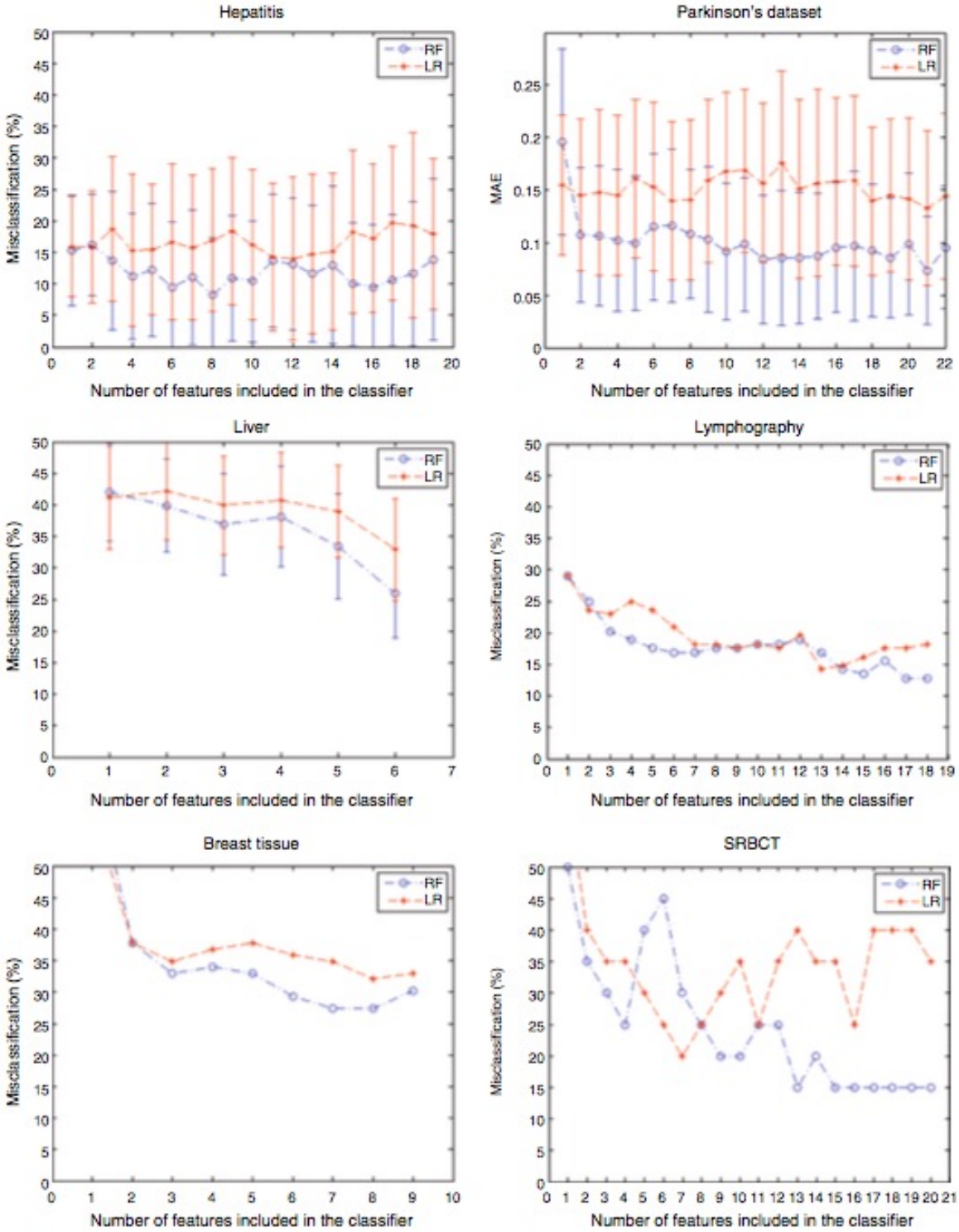
Summary of datasets

Dataset	Data Matrix	Associated Task	Feature type
Hepatitis	155 x 19	2 classes	C(17), D(2)
Parkinson's	195 x 22	2 classes	C(22)
Liver	345 x 6	2 classes	D(6)
Lymphography	148 x 18	4 classes	D(18)
Breast tissue	106 x 9	6 classes	C(9)
SRBCT tumours	63 x 2308	4 classes	C(2308)

Comparison of LR and RF (all features)

Dataset	LR MCR	RF MCR	MCR difference	Relative Improvement (%)	Validation Scheme
Hepatitis	17.25 ± 11.48	11.55 ± 11.03	5.70	33.04	10-fold CV
Parkinson's	14.75 ± 8.63	8.90 ± 6.65	5.85	39.66	10-fold CV
Liver	32.91 ± 8.14	25.94 ± 7.02	6.97	21.18	10-fold CV
Lymphography	18.24	12.83	5.41	29.66	LOO
Breast tissue	33.02	30.19	2.83	8.57	LOO
SRBCT Tumours	75.00	10.00	65.00	86.66	Test set

The misclassification difference between LR and RF was in all cases statistically significant ($p<0.001$) using the Mann–Whitney statistical hypothesis test.


Hepatitis
Misclassification (%)
Number of features included in the classifier
RF
LR
Parkinson's dataset
MAE
Number of features included in the classifier
RF
LR
Liver
Misclassification (%)
Number of features included in the classifier
RF
LR
Lymphography
Misclassification (%)
Number of features included in the classifier
RF
LR
Breast tissue
Misclassification (%)
Number of features included in the classifier
RF
LR
SRBCT
Misclassification (%)
Number of features included in the classifier
RF
LR

More data or better models?

- In many disciplines, the focus is often on obtaining new predictive variables.
- In medicine, the traditional model structure is a logistic regression.
- We use LR as a linear benchmark for this reason.
- A powerful nonlinear classifier, RF, is consistently superior to logistic regression, offering a relative improvement in performance of 36% in predicting the outcome across six different medical datasets.

Tsanas, Little, McSharry (2013). In “Handbook of systems and complexity in health”, Springer. New York.

Nonlinearity is important

- In settings where the number of features is larger than the number of samples the improvement with RF is even more impressive.
- Interestingly, in some datasets a lower number of features leads to a *lower* misclassification error.
- This is the manifestation of the curse of dimensionality, where additional features increase the signal-to-noise ratio in the data and are detrimental for the performance of the classifier.
- RF is fairly robust and LR is particularly sensitive to the number of features with respect to the number of samples.

Clinical trials

- Clinical trials are prospective biomedical or behavioral research studies on human participants.
- They are are designed to answer specific questions about biomedical or behavioral interventions, including new treatments (such as novel vaccines, drugs, dietary choices, dietary supplements, and medical devices) and known interventions that warrant further study and comparison.
- Clinical trials generate data on safety and efficacy.
- Only 10% of all drugs started in human clinical trials become an approved drug

Benchmark: placebo & control

- A placebo is a simulated or otherwise medically ineffectual treatment for a disease or other medical condition intended to deceive the recipient.
- It could be a pill, a shot, or some other type of "fake" treatment.
- This can be different to a control group who knowingly did not get a placebo.
- Sometimes patients given a placebo treatment will have a perceived or actual improvement in a medical condition, a phenomenon commonly called the placebo effect or placebo response.

Bipolar disorder

- Bipolar disorder is a psychiatric condition characterized by repeated episodes on elevated mood (mania) and low mood states (depression) (Anderson et al. 2012).
- It is ranked as the sixth cause of disability worldwide and bipolar spectrum disorders affect nearly 1% of adult population (Merikangas et al. 2007; Schmitt et al. 2014).
- During the mania episode, the subject's brain goes into a high energy state; experiencing extreme emotions, irritable mood and sometimes extreme happiness (euphoria).

Bipolar disorder

- It is also possible that the subject may become overconfident, very optimistic, engage in risky activities or tend to spend a lot of money (American Psychiatry Association, 2013).
- The subject may also get confused about what is real and what is not (Post et al. 2013).
- At the other extreme, during the depression episode, the subject is in a low mood state, has low self-esteem, loss of interest or pleasure, feelings of sadness and anxiety.
- Finally subjects may also experience disturbance in sleep and appetite, fatigue and loss of interest in activities that are normally enjoyable.

Lamotrigine

- Lamotrigine is an anticonvulsant (anti-seizure medication) used for the treatment of bipolar disorder, especially for depressive episodes.
- Using data collected through a randomized, placebo-controlled trial, we want to determine whether bipolar patients taking lamotrigine significantly differ from those on placebo.
- Equally important, we also want to establish how long it takes for the drug to provide statistically significant outcomes, thus estimating the minimum amount of time required to undertake trials.

Nzeyimana, Antoine; Saunders, Kate; Geddes, John; McSharry, Patrick (2018), JMIR Mental Health. 2018 Nov 22;5(4):e63

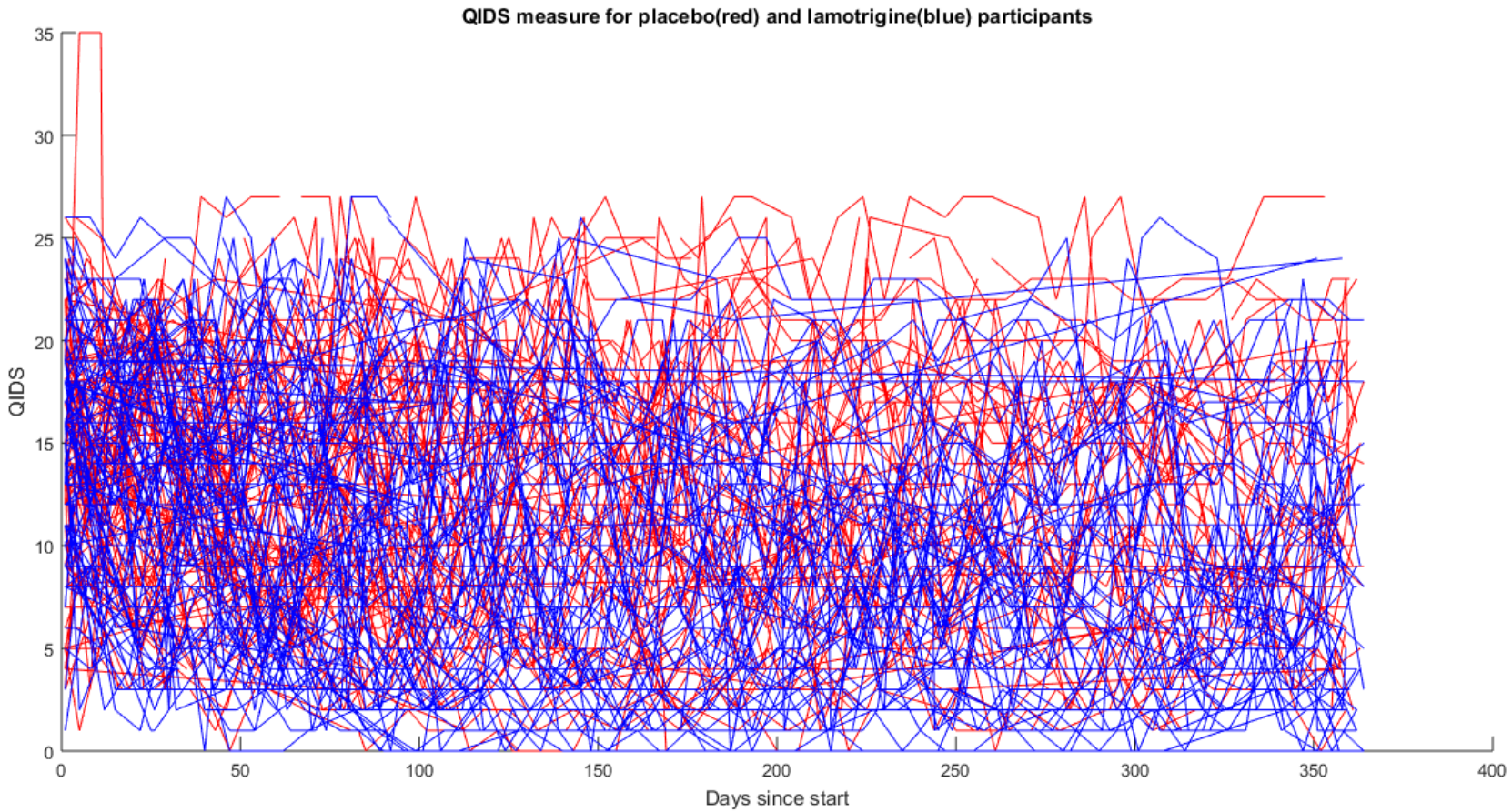
Bipolar disorder treatment

- Bipolar disorder: Psychiatric condition with episodes of depression and mania
- UK NHS True Colors program: Tele-monitoring system collecting self-rated mood scores.
- Data: Quick Inventory of Depressive Symptomatology (QIDS-SR16) scores(Max: 27), Altman Self-Rating Mania Scale (ASRM) scores(Max: 20)
- Two forms of treatment: Lamotrigine vs Placebo
- Research question: Can we differentiate the two groups? If so, how soon can we detect that difference?
- Methods: Classification, Time series analysis, Spectral analysis, Mainly using QIDS-SR data.

Data collection

- The data was collected over a period of 52 weeks and reported from 202 participants suffering from bipolar disorder
- 149 patients are classified as having bipolar type I disorder and 53 as having bipolar type II disorder.
- Among those participants, 90 are male and 112 are female.
- In terms of age, 36 participants are under 30 years of age, 45 are between 31 and 40, 66 are between 41 and 50 years old; we also have 55 subjects who are older than 50 years.

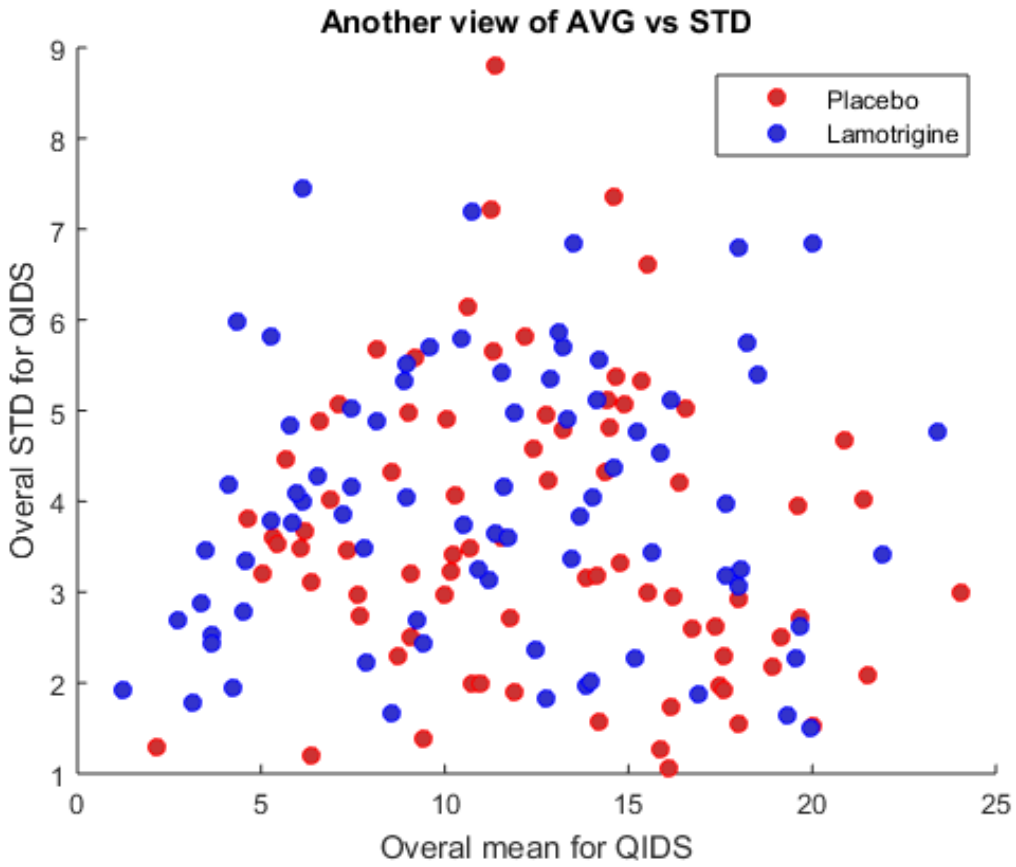
Data: Lamotrigine vs Placebo



Original time series(QIDS over 52 weeks): Not so obvious to differentiate the two groups by eye.

Challenge No 1: Non-uniformly sampled data

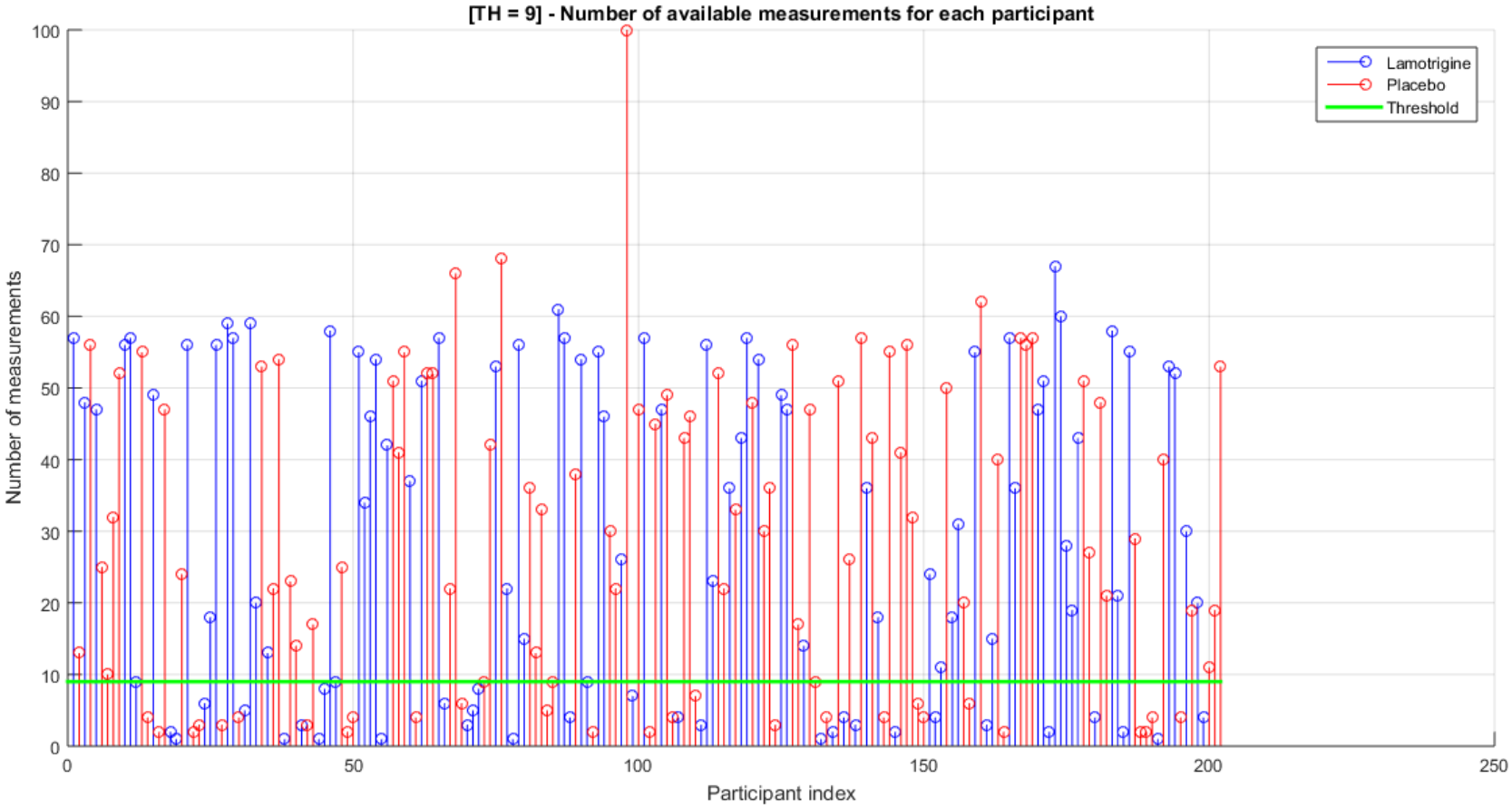
Feature design: Standard deviation vs Mean



QIDS - Standard deviation vs Mean

Challenge No 2: Not easy to draw a good decision boundary.

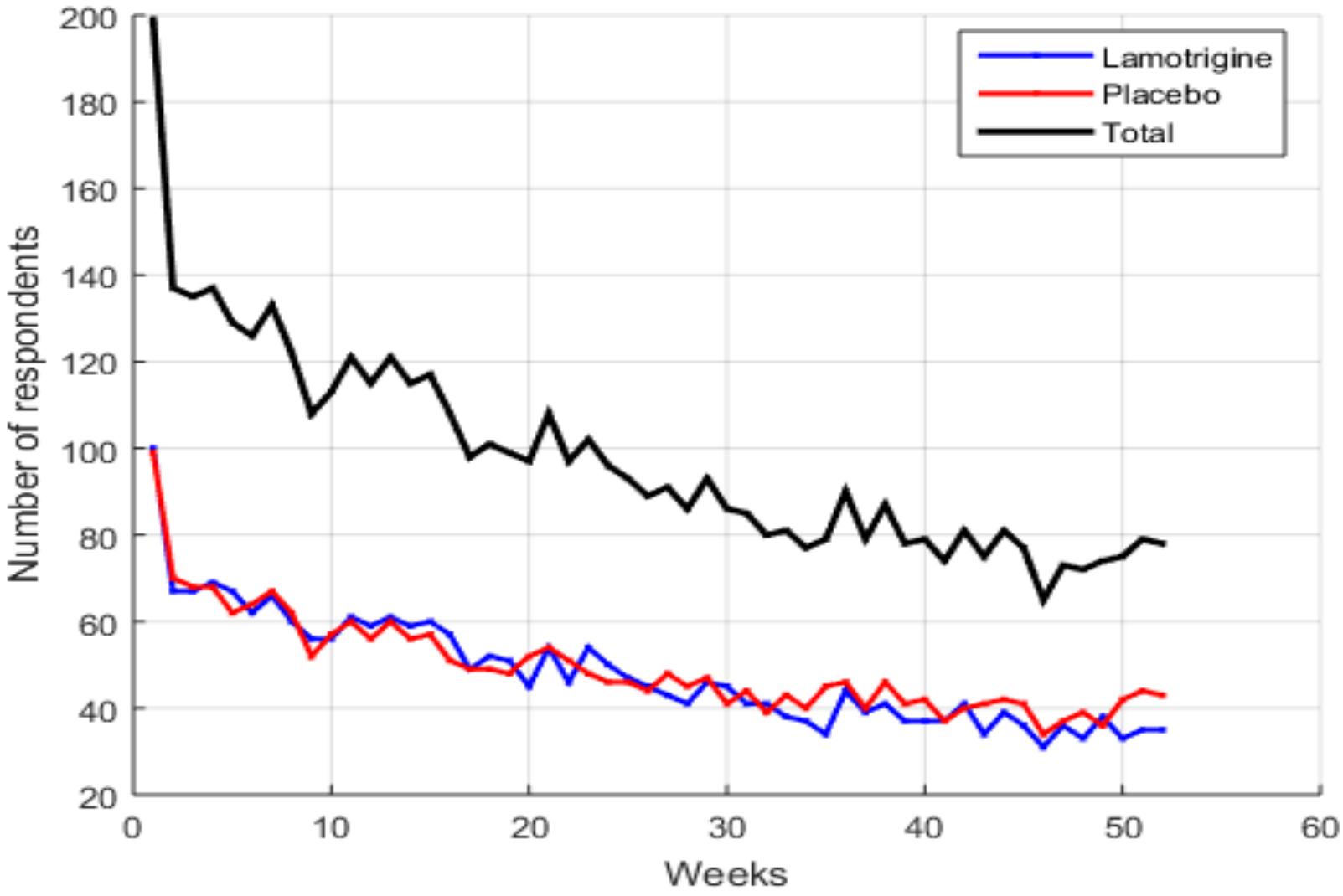
Data insights: Measurements per participant



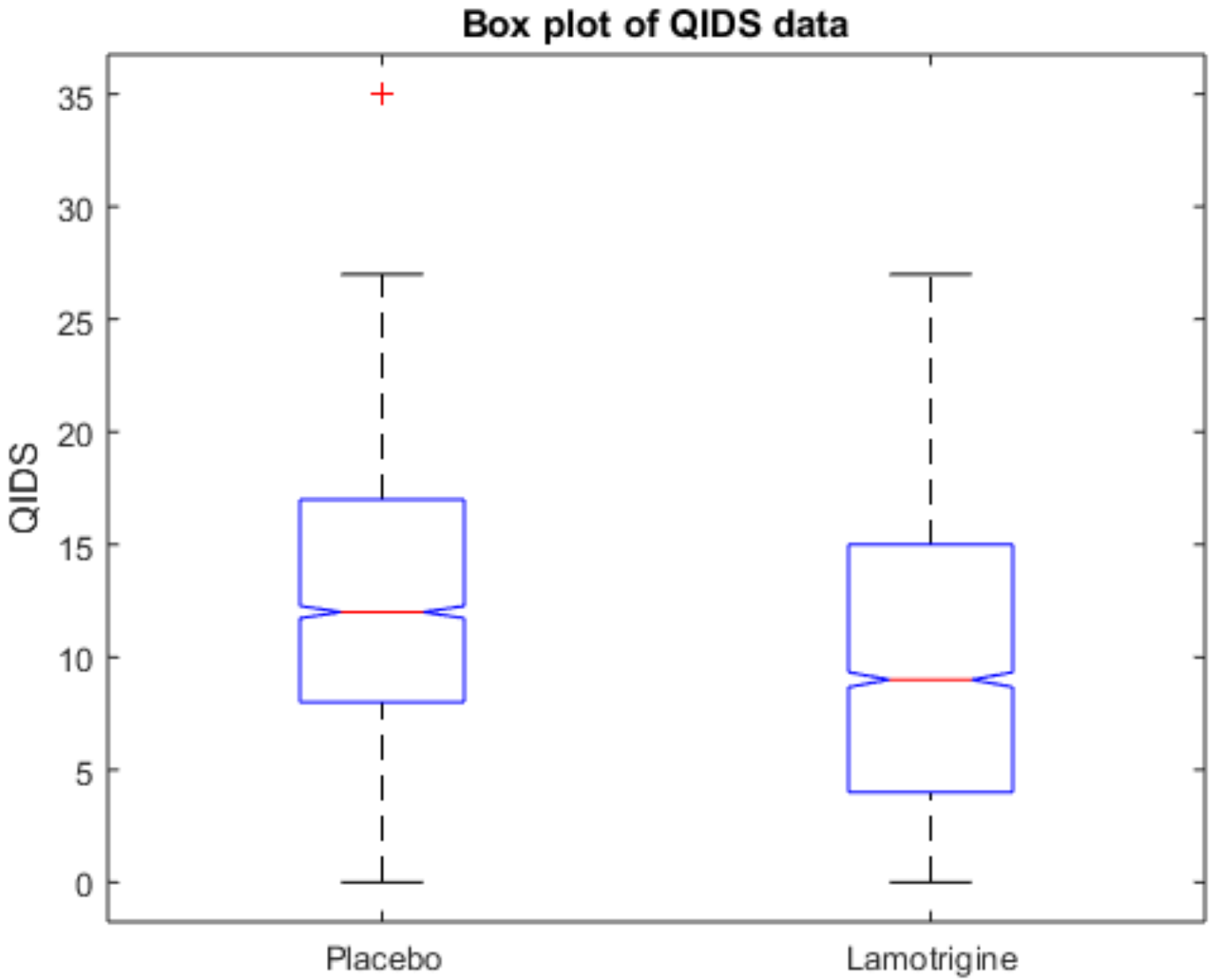
QIDS - Number of available measurements for each participant.

Challenge No 3: Fewer participants report for the entire period of 52 weeks.

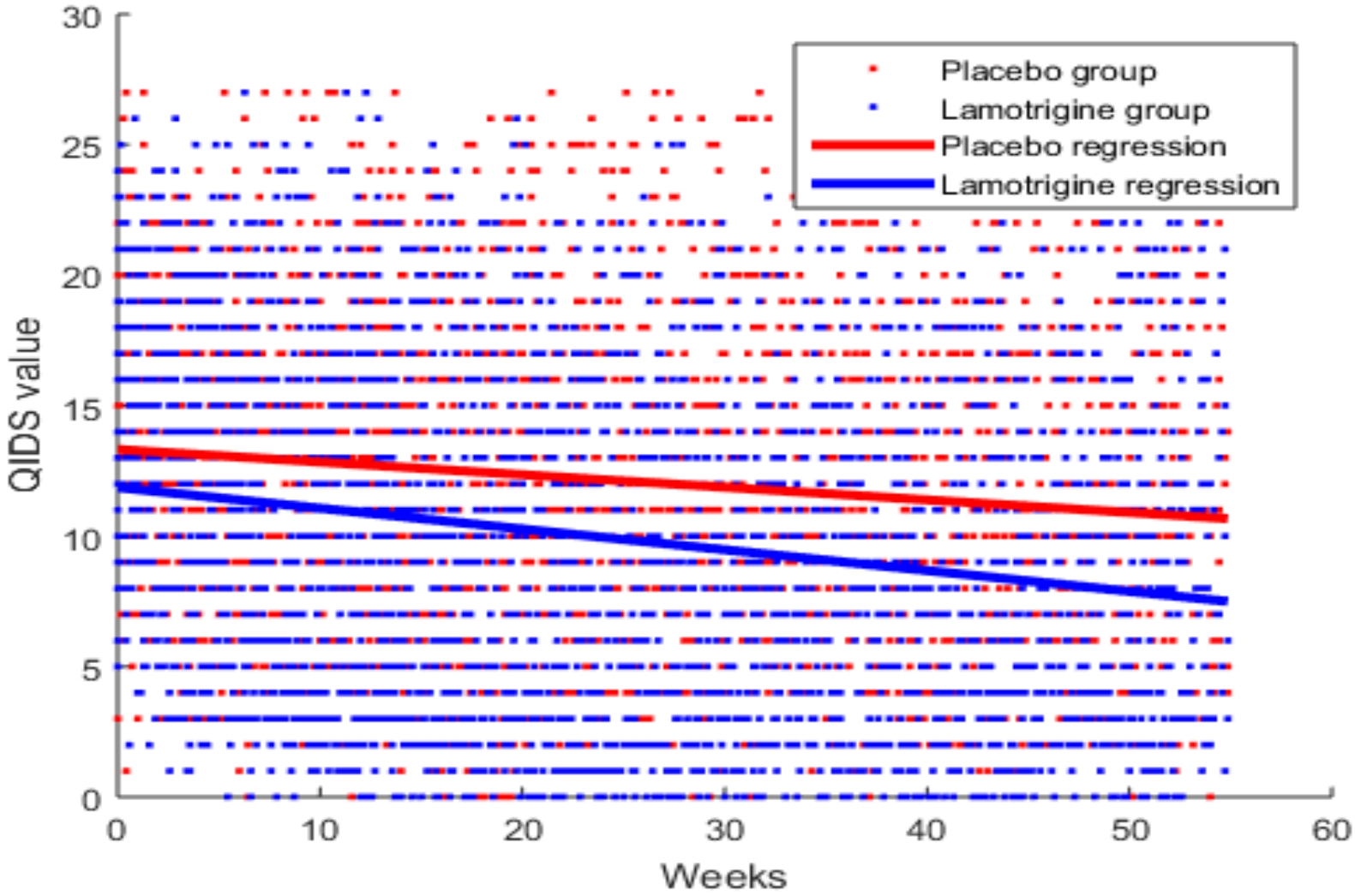
Compliance


Lamotrigine
Placebo
Total
Number of respondents
Weeks
200
180
160
140
120
100
80
60
40
20
0
10
20
30
40
50
60

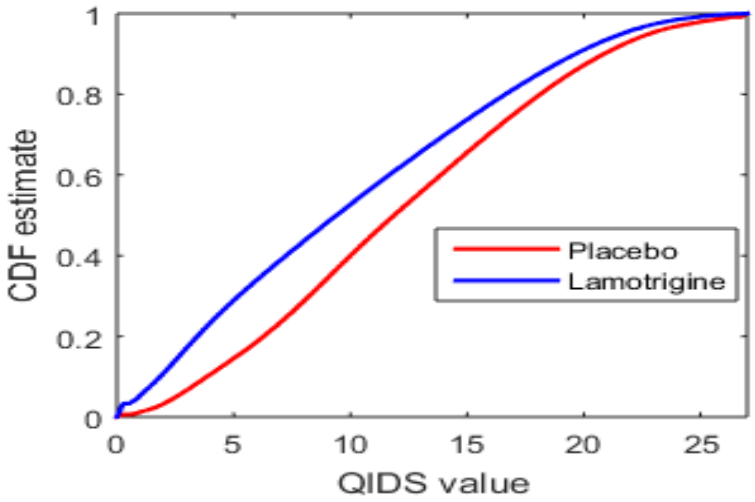
Distribution of QIDS

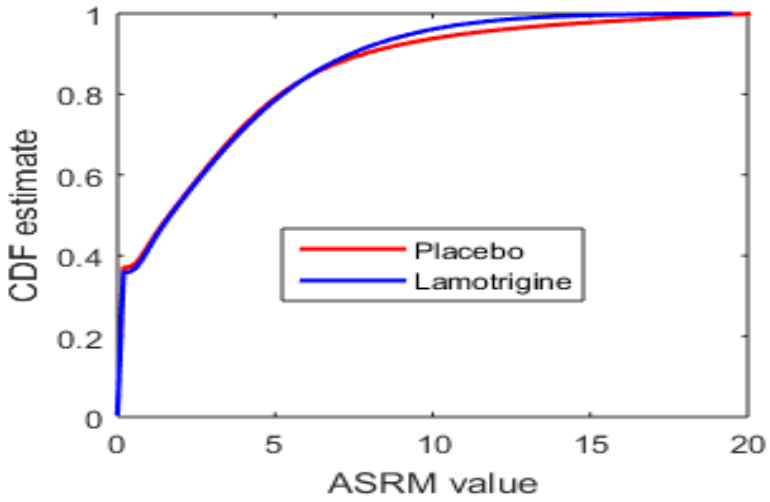


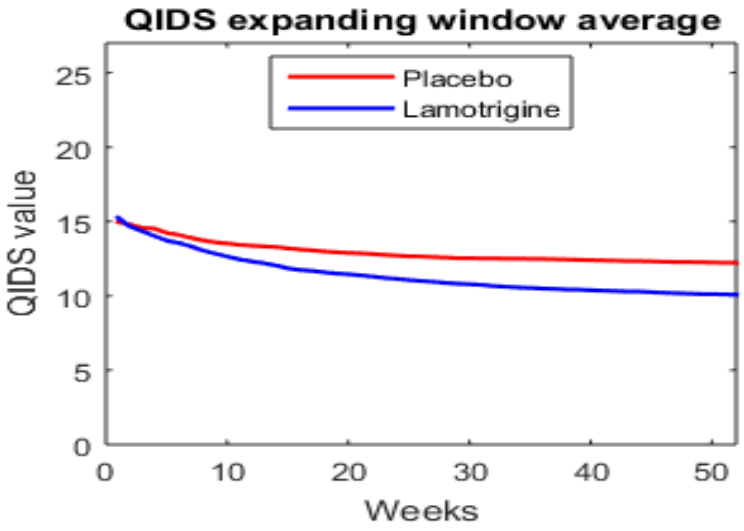
Trend analysis

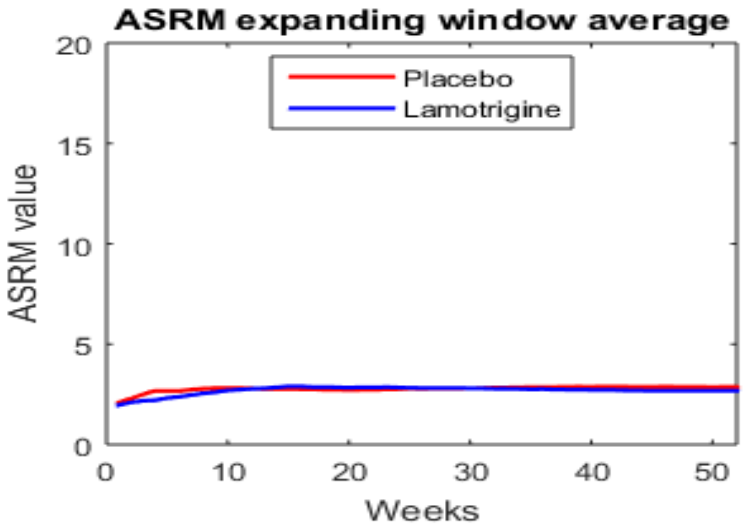


QIDS and ASRM









Features

- Subject attributes: age, gender, bipolar type
- Additional features derived from QIDS values
- Simple distribution summary statistics:
  - mean, μ
  - standard deviation, σ
  - coefficient of variation, σ/μ
  - kurtosis
  - skewness

Time series features

- We are interested in the variation in QIDS from one measurement to the next and the control mechanism.
- Slope of x(t) versus x(t-1) measures the basic lag autocorrelation structure.
- Correlation coefficient of x(t) versus x(t-1) measures the goodness of fit of a linear AR(1) model.

Lomb Periodogram

- Spectral information can be derived using the Lomb periodogram (Lomb, 1976; Scargle, 1982), also known as least-squares spectral analysis.
- It is an appropriate technique for estimating the power spectral density of unevenly sampled time series.
- We derived the amount of power in the low and high frequency bands
- Also estimated the ratio of these amounts of power.

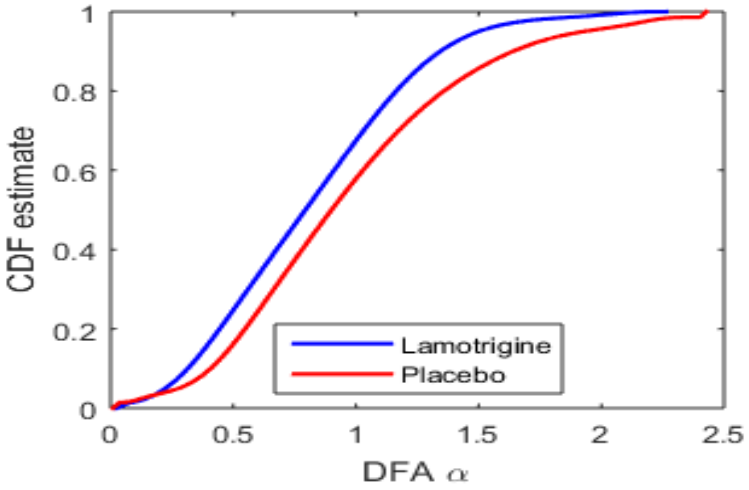
Detrended Fluctuation Analysis (DFA)

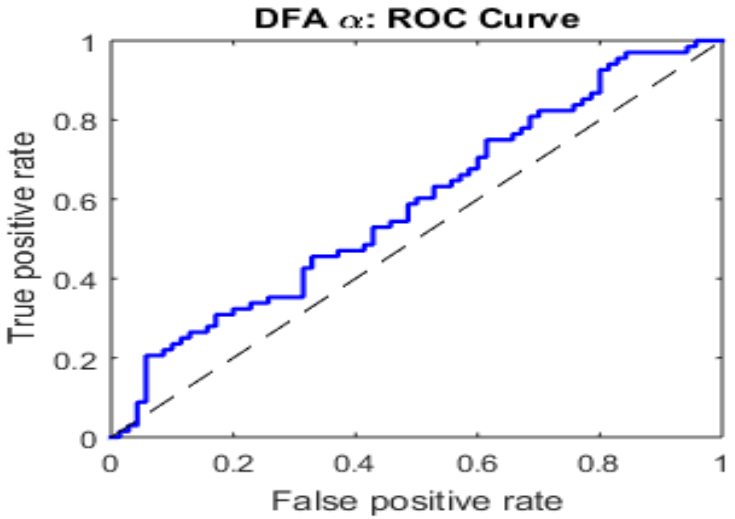
- Detrended Fluctuation Analysis (DFA) (Peng et al. 1994) is another commonly used technique for analyzing biomedical signals.
- DFA basically measures the statistical self-affinity of a signal.
- (McSharry et al., 2005) give details of the DFA algorithm and related methods.
- Essentially, DFA scaling exponent α measures how rough a time series is. For example, white noise which fluctuates a lot has α of 0.5; for pink noise α is 1 and random walk gets DFA α of 1.5.
- We compute the DFA exponent α for each individual subject and use it as another feature in our classification model.

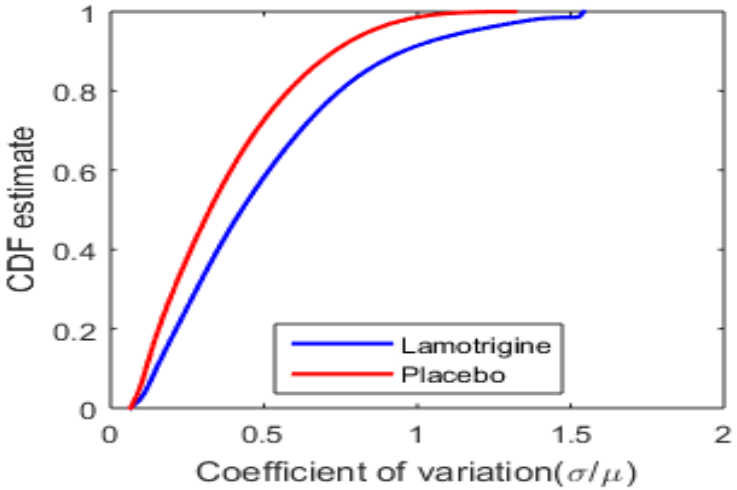
Classifiers

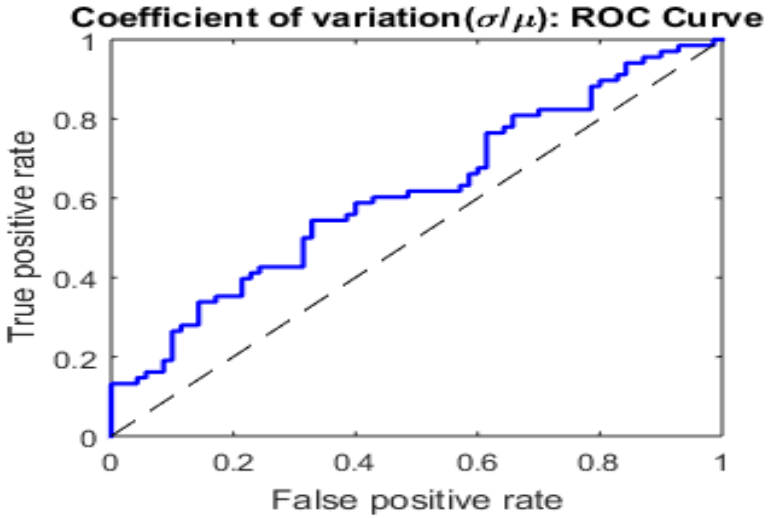
- A number of linear and nonlinear classifiers were compared.
- A linear classifier is classification algorithm whose objective function is a function of a linear combination of features.
- A binary linear classifier has a linear decision boundary.
- A non-linear classifier will have a non-linear decision boundary.
- The linear classifiers we investigated are logistic regression (Cox, 1958), linear discriminant analysis (LDA) (McLachlan, 2004) and linear support vector machines (SVM) (Cortes et al., 1995).
- The non-linear classifiers investigated include quadratic discriminant analysis (QDA), Gaussian kernel SVM and k-nearest neighbors (KNN)

CDF estimates and ROC curves









Classification accuracy

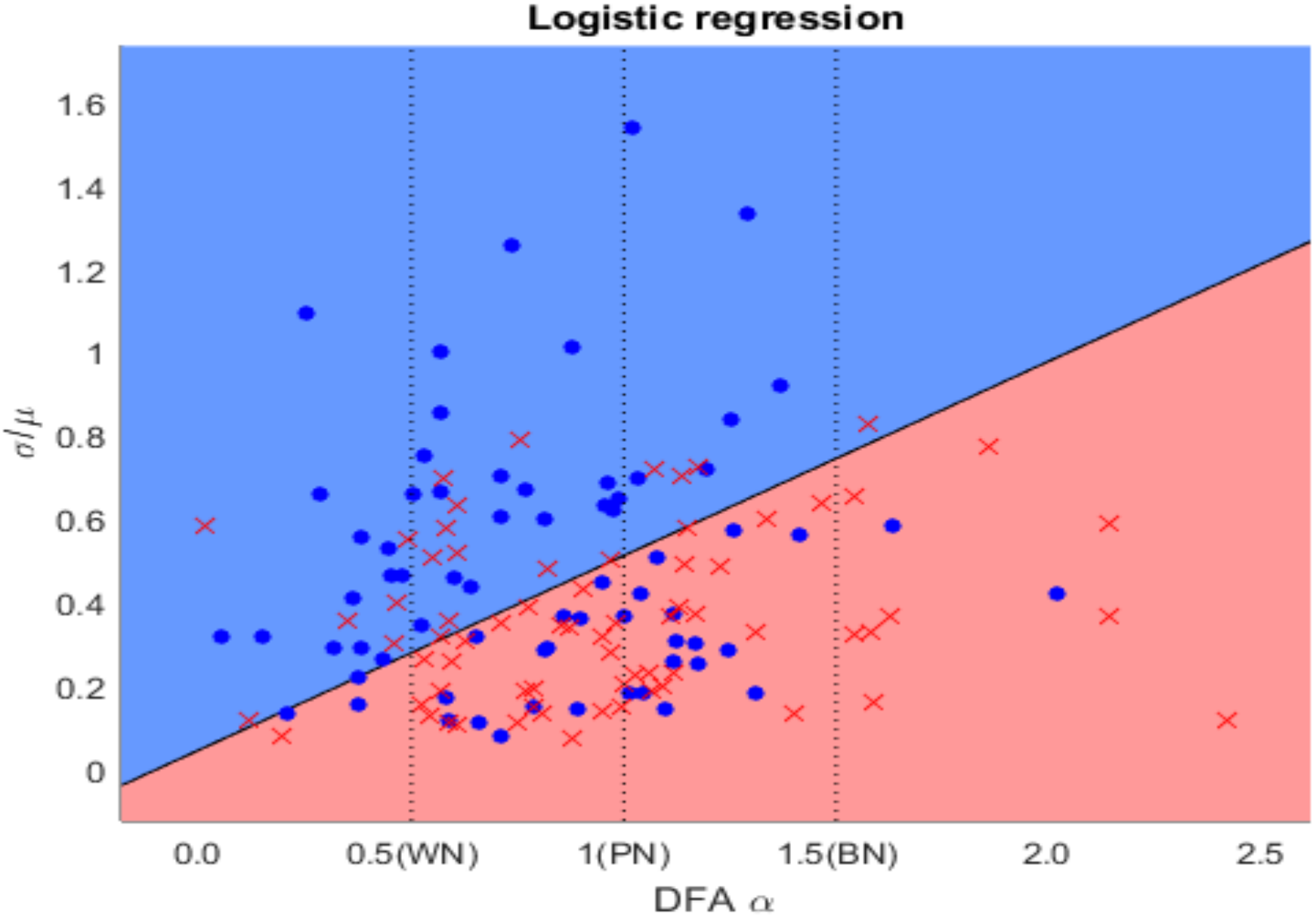
- The area under the curve in the ROC indicates the importance of a given feature.
- Complex models with many parameters may over-fit the data and produced over-optimistic results.
- To avoid this problem, our evaluation metric is therefore based on the classification accuracy obtained from out-of-sample data.
- We used 10 fold cross validation.
- The percentage of accurate classifications is reported.

FEATURE	AUC	IN SAMPLE ACCURACY						OUT OF SAMPLE ACCURACY					
		LR	LDA	QDA	LSVM	GSVM	KNN	LR	LDA	QDA	LSVM	GSVM	KNN
QIDS Mean(AVG)	0.587	0.594	0.558	0.565	0.558	0.573	0.696	0.517	0.558	0.580	0.551	0.536	0.536
QIDS Standard Deviation(STD)	0.552	0.558	0.544	0.551	0.544	0.558	1.000	0.406	0.536	0.500	0.551	0.500	0.478
Low Frequency Power(LF)	0.482	0.507	0.449	0.544	0.507	0.558	0.710	0.353	0.442	0.507	0.478	0.486	0.522
High Frequency Power(HF)	0.599	0.601	0.601	0.573	0.558	0.601	0.601	0.581	0.587	0.551	0.558	0.551	0.587
DFA Alpha	0.587	0.573	0.544	0.565	0.544	0.565	0.645	0.418	0.536	0.544	0.544	0.464	0.529
Coefficient of Variation(STD/AVG)	0.614	0.609	0.587	0.587	0.580	0.594	0.630	0.604	0.587	0.594	0.580	0.573	0.580
Power Ratio(HF/LF)	0.571	0.544	0.507	0.507	0.515	0.573	0.616	0.555	0.471	0.478	0.471	0.536	0.580
Skewness	0.537	0.558	0.565	0.573	0.544	0.580	0.645	0.480	0.529	0.551	0.464	0.522	0.500
Kurtosis	0.581	0.616	0.558	0.558	0.551	0.587	0.652	0.580	0.551	0.551	0.565	0.587	0.587
Slope of X_t vs X_{t-1}	0.528	0.536	0.536	0.522	0.536	0.536	1.000	0.448	0.522	0.515	0.464	0.500	0.580
Correlation coefficient of X_t vs X_{t-1}	0.521	0.536	0.529	0.522	0.529	0.536	0.725	0.433	0.515	0.500	0.457	0.515	0.529
Bipolar type	0.511	0.507	0.507	0.507	0.507	0.507	0.507	0.360	0.442	0.457	0.442	0.471	0.493
Age	0.545	0.551	0.551	0.551	0.551	0.551	0.551	0.551	0.551	0.544	0.515	0.551	0.544
Gender	0.500	0.507	0.507	0.507	0.507	0.507	0.507	0.303	0.370	0.442	0.420	0.420	0.493
All features combined	0.720	0.696	0.674	0.732	0.659	1.000	0.580	0.557	0.536	0.493	0.558	0.449	0.536
DFA Alpha, Coeff. Of Variation	0.665	0.659	0.630	0.630	0.630	0.659	0.638	0.619	0.609	0.623	0.587	0.616	0.594

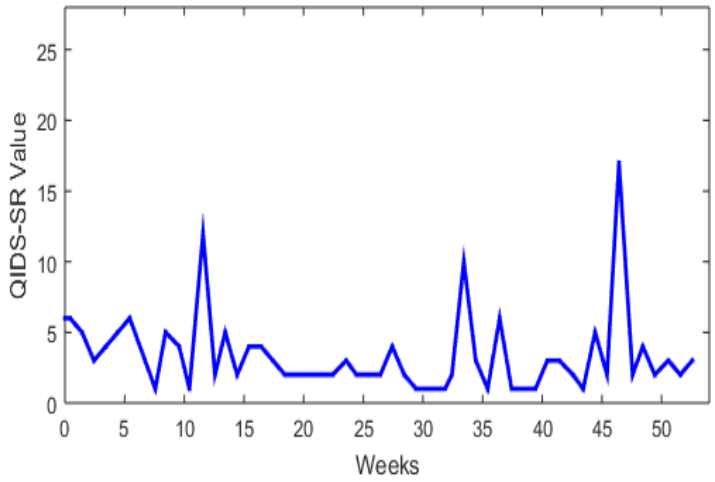
Implications

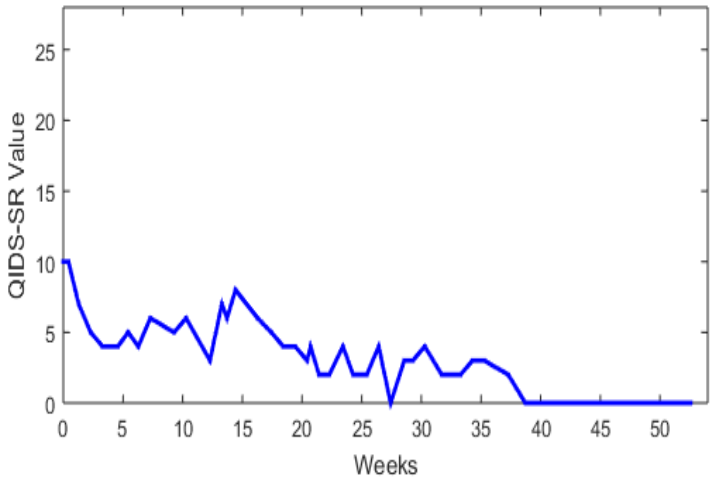
- The two features, the coefficient of variation and DFA exponent, imply that a two-dimensional visualization diagram and decision boundary can be constructed to better understand bipolar disorder and how participants are affected by lamotrigine.
- No empirical evidence for a nonlinear classification model.
- The selected features for classification suggest that lamotrigine increases the coefficient of variation (achieved by increasing the standard deviation or decreasing the mean of the QIDS time series).
- Subjects taking lamotrigine tend to have rougher time series, indicative of greater temporal instability in the time series.

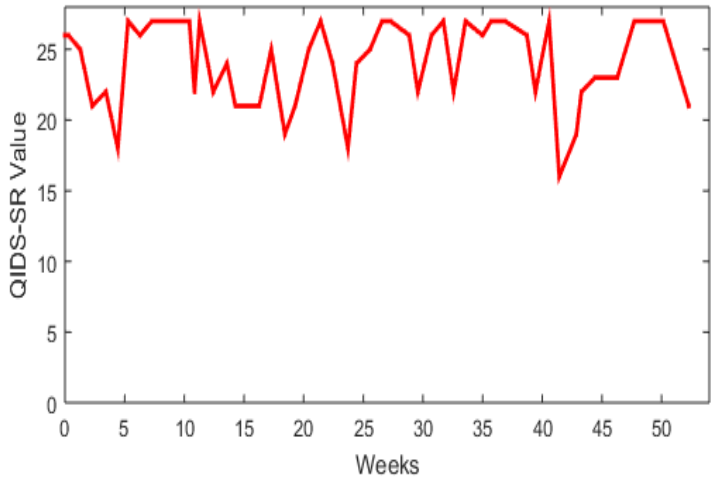
Classification decision boundary

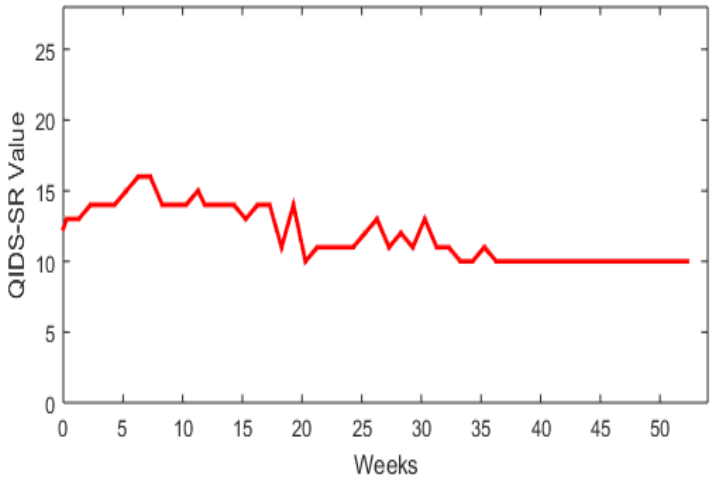


QIDS time series for subjects in corners

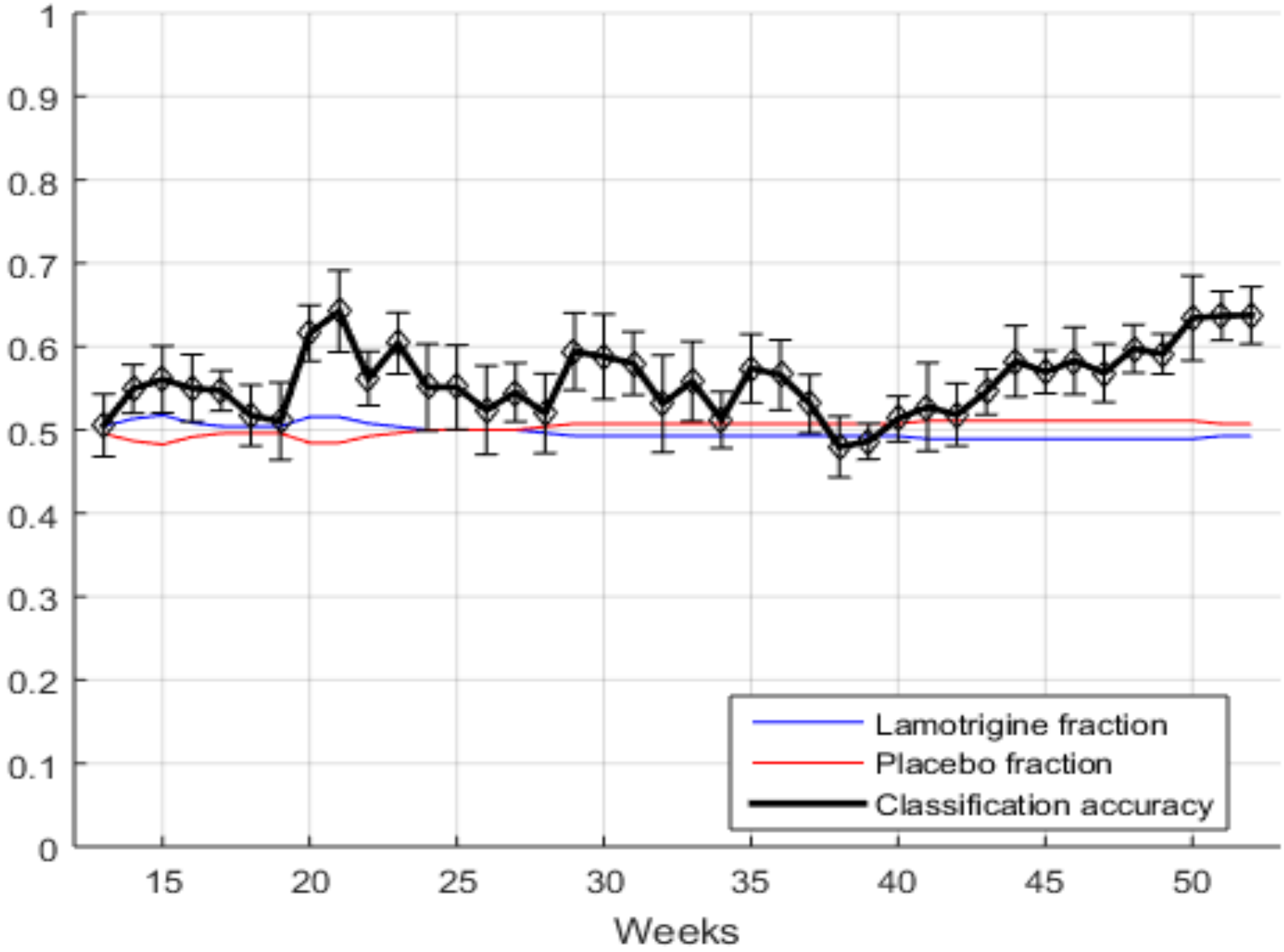








Significance of classification accuracy



Trial duration

- The maximum of the fractions of lamotrigine and placebo participants serves as a “no information” benchmark for the classifier.
- For the majority of periods from 20 weeks onwards, the classifier is outperforming the benchmark.
- The statistical significance of the classifications were evaluated and suggest that a trial of at least 44 weeks is required to distinguish between lamotrigine and placebo.