Data, Inference & Applied Machine Learning

Course: 18-785

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

Fall 2022

ICT Center of Excellence
Carnegie Mellon University

Course Description

- Utilize data and quantitative models to automate predictive analytics and make improved decisions.
- Descriptive statistics to data analysis to machine learning
- Explain process of collecting, cleaning, interpreting, transforming, exploring, analysing and modeling data
- Goals: extract information, communicating insights and decision-making.
- Advantages and disadvantages of linear, nonlinear, parametric, nonparametric and ensemble methods will be discussed while exploring the challenges of both supervised and unsupervised learning.
- Quantify uncertainty, statistical hypothesis testing and communicating confidence in model results will be emphasized.
- Applications will include visualization, clustering, ranking, pattern recognition, anomaly detection, data mining, classification, regression, forecasting and risk analysis.

Learning Objectives

- Overview of the use and potential of data analysis and machine learning in research, business and government.
- Address practical challenges, e.g. segment customers, understand client behaviour or formulate a new strategy for optimising key performance indicators (KPIs), by applying data analysis techniques to real-world data.
- Plan, design and implement an empirical research project using statistical, computational and quantitative techniques.
- Communicate the outcomes from complicated analyses and build decision support tools.
- Discuss project design, data collection, data quality and techniques to cope with measurement errors, missing values and outliers.
- Combine theoretical aspects of data analysis with visual examples and demonstrations of how to construct and utilize statistical models in practice.
- Strong emphasis on highlighting the challenges of working with real-world data, avoiding over-fitting and the risks of relying on traditional assumptions.

Outcomes

- Design an empirical project in response to a specific objective
- Identify and collect relevant data for undertaking the project
- Acquire data and organize it into a structured format
- Visualize data, identify key characteristics and present a summary
- Describe the advantages and disadvantages of different models
- Decide which models are likely to work best for a given application
- Undertake feature construction and feature selection
- Apply machine learning techniques for estimation and evaluation
- Select an optimal model using statistical approaches
- Produce diagnostic information for investigating model properties
- Understand model weaknesses and where assumptions could fail
- Communicate model output and conclusions to end-users

Hurricane Classification

Saffir - Simpson hurricane scale

Category 1
- Winds 74-95 mph (119-153 km/h)
- Some damage and power cuts

Category 2
- Winds 96-110 mph (154-177 km/h)
- Extensive damage

Category 3
- Winds 111-129 mph (178-208 km/h)
- Well-built homes suffer major damage

Category 4
- Winds 130-156 mph (209-251 km/h)
- Severe damage to well-built homes, trees blown over

Category 5
- Winds 157+ mph (252+ km/h)
- Many buildings destroyed, major roads cut off

Hurricane Research

- How should one classify hurricanes?
- How does wind speed translate into insurance loss?
- What is the risk of damage?
- Risk = hazard x exposure x vulnerability

Suggested reading

- Chatfield, C., Statistics for technology: A course in applied statistics. An Introduction (3rd ed.), Chapman & Hall, CRC Press, 1983.

- Chatfield, C., The Analysis of Time Series: An Introduction (6th ed.), Chapman & Hall, CRC Press, 2004.

- Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2013.

  URL: http://statweb.stanford.edu/~tibs/ElemStatLearn/

Course outline

Week	Description
1	Measurement, data types, data collection, data cleaning
2	Data manipulation, data exploration, visualization techniques
3	Probability, statistical distributions, descriptive statistics
4	Statistical hypothesis testing, quantifying confidence
5	Time series analysis, autoregression, moving averages
6	Linear regression, parameter estimation, model selection, evaluation

Course outline continued

Week	Description
7	Statistical learning, scores, over-fitting, cross-validation
8	Regularization, feature selection, classification, prediction
9	Nonlinear techniques, nonlinear features, nonlinear models
10	Supervised learning, decision trees, non-parametric models
11	Unsupervised learning, dimensionality reduction, clustering
12	Ensemble approaches, model averaging, bootstrap, bagging

Weekly Timetable

Time	Tuesday	Thursday	Friday
08:30-10:20 US 14:30-15:20 Rwanda*	Class	Class	
08:00-09:00 US 14:00-15:00 Rwanda*			Lab/Recitation

*The times are anchored to US Eastern Time (Pittsburgh). Note the move one hour later after Nov, 7 due to the end of US daylight saving (and no change in Rwanda). Hence the time difference changes from 6 to 7 hours and the 14:30 start becomes 15:30 in Rwanda.

Detailed Timetable

Week	A	Pittsburgh	Kigali	B	Pittsburgh	Kigali
Week	A	Pittsburgh	Kigali	B	Pittsburgh	Kigali
1	Tues, Aug 30	08:30 ET	14:30 CAT	Thurs, Sep 01	08:30 ET	14:30 CAT
2	Tues, Sep 06	08:30 ET	14:30 CAT	Thurs, Sep 08	08:30 ET	14:30 CAT
3	Tues, Sep 13	08:30 ET	14:30 CAT	Thurs, Sep 15	08:30 ET	14:30 CAT
4	Tues, Sep 20	08:30 ET	14:30 CAT	Thurs, Sep 22	08:30 ET	14:30 CAT
5	Tues, Sep 27	08:30 ET	14:30 CAT	Thurs, Sep 29	08:30 ET	14:30 CAT
6	Tues, Oct 04	08:30 ET	14:30 CAT	Thurs, Oct 06	08:30 ET	14:30 CAT
7	Tues, Oct 25	08:30 ET	14:30 CAT	Thurs, Oct 27	08:30 ET	14:30 CAT
8	Tues, Nov 01	08:30 ET	14:30 CAT	Thurs, Nov 03	08:30 ET	14:30 CAT
9	Tues, Nov 08	08:30 ET	15:30 CAT	Thurs, Nov 10	08:30 ET	15:30 CAT
10	Tues, Nov 15	08:30 ET	15:30 CAT	Thurs, Nov 17	08:30 ET	15:30 CAT
11	Tues, Nov 22	08:30 ET	15:30 CAT	Tues, Nov 29	08:30 ET	15:30 CAT

Evaluation

- Participation (class, Piazza & Kahoot)
- Four assignments (D&I)
- Interim exam (multiple choice)
- Three assignments (AML)
- Project (Kaggle)
- Final exam (multiple choice)

Canvas and Piazza

- Canvas will be used for posting supplementary course materials and turning in assignments. Please familiarize yourself with navigating, uploading and downloading.
- Piazza will be used for questions and discussion among students, TAs and the instructor.

TA approach

- Open discussions on Piazza
- Zoom recitations (1h per session)
- Background material and information
- Agenda for recitations based on discussion and student questions
- Assigned TAs and telecon hours
- Solution walk through at recitations

TA Team

- Christian Iradukunda

ciradukunda@africa.cmu.edu

- Innocent Mukoki

imukoki@andrew.cmu.edu

- And other TAs being confirmed

Registration, Waitlist, Access

- Please contact Megan Oliver mvoliver@andrew.cmu.edu

  for any queries you may have on the course administration
- Contact your local assigned TA in the first instance or if you cannot access a communication channel or material.
- As course instructor, I am available should the above fail to resolve the issue.

Class components

Number	Description	Activity
1	Challenge	Presentation
2	Student discussion & Kahoot quiz	Debate
3	Historical case study	Presentation
4	Quantitative approach	Presentation
5	Demonstration in Matlab/Python	Code
6	Questions and feedback	Debate

Slido

- **Slido.com #608 556**

Patrick McSharry

Visiting Professor
ICT Center of Excellence
CMU

Life

23 years

16 years

1 year

2 years

7.5 years

Career

UNIVERSITY OF OXFORD

Research Fellow
11 years

MARIE CURIE ACTIONS

Research Fellow
2 years

ssee

Senior Research Fellow
5 years

OXFORD-MAN INSTITUTE

UNIVERSITY OF OXFORD

Faculty Member
1 year

Carnegie Mellon University Rwanda

Visiting Prof

Education

TRINITY COLLEGE DUBLIN

BA
in Theoretical Physics

TRINITY COLLEGE DUBLIN

Master of Science
in Engineering

UNIVERSITY OF OXFORD

DPhil in Mathematics

Patrick McSharry
Visiting Professor
ICT Center of Excellence
CMU

Research Areas

Forecasting

Construction of quantitative models for describing complex systems with the objective of forecasting, classification and decision-making.

Risk Management

Fusion of knowledge, science and data for pricing, transferring and sharing risk of extreme events.

Decision-making

Using mobile technology to collect data about medical disorders and develop tailor-made decision support tools for self-management.

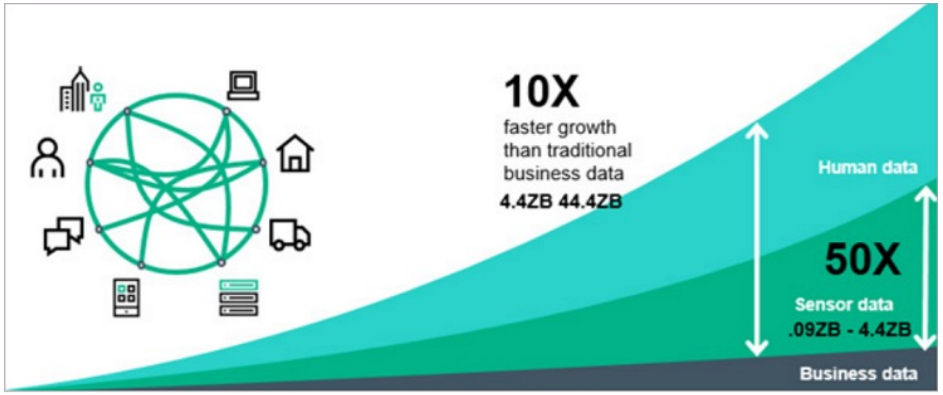
Big Data

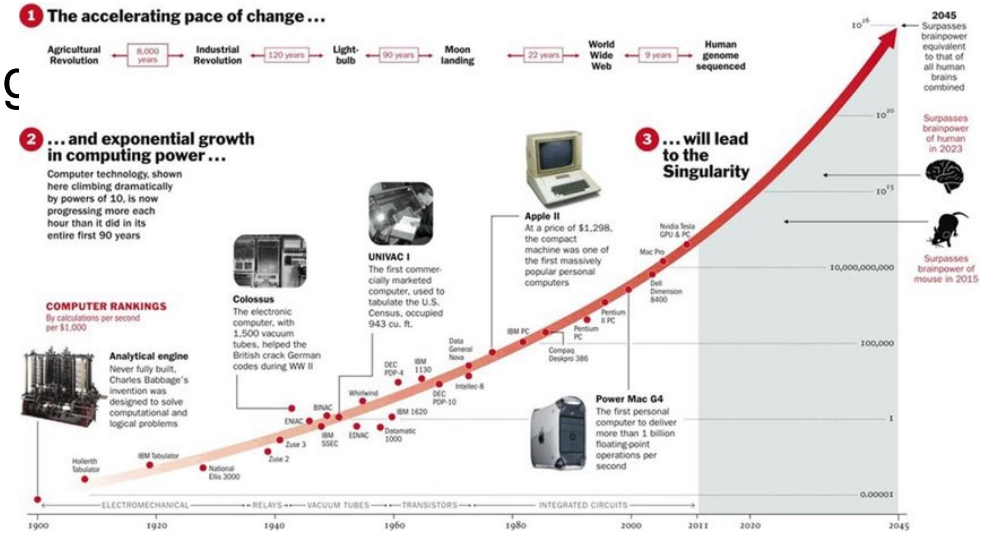
Deployment of quantitative expertise in public and private sectors to inform data-driven policy and strategy.

Artificial Intelligence (AI)

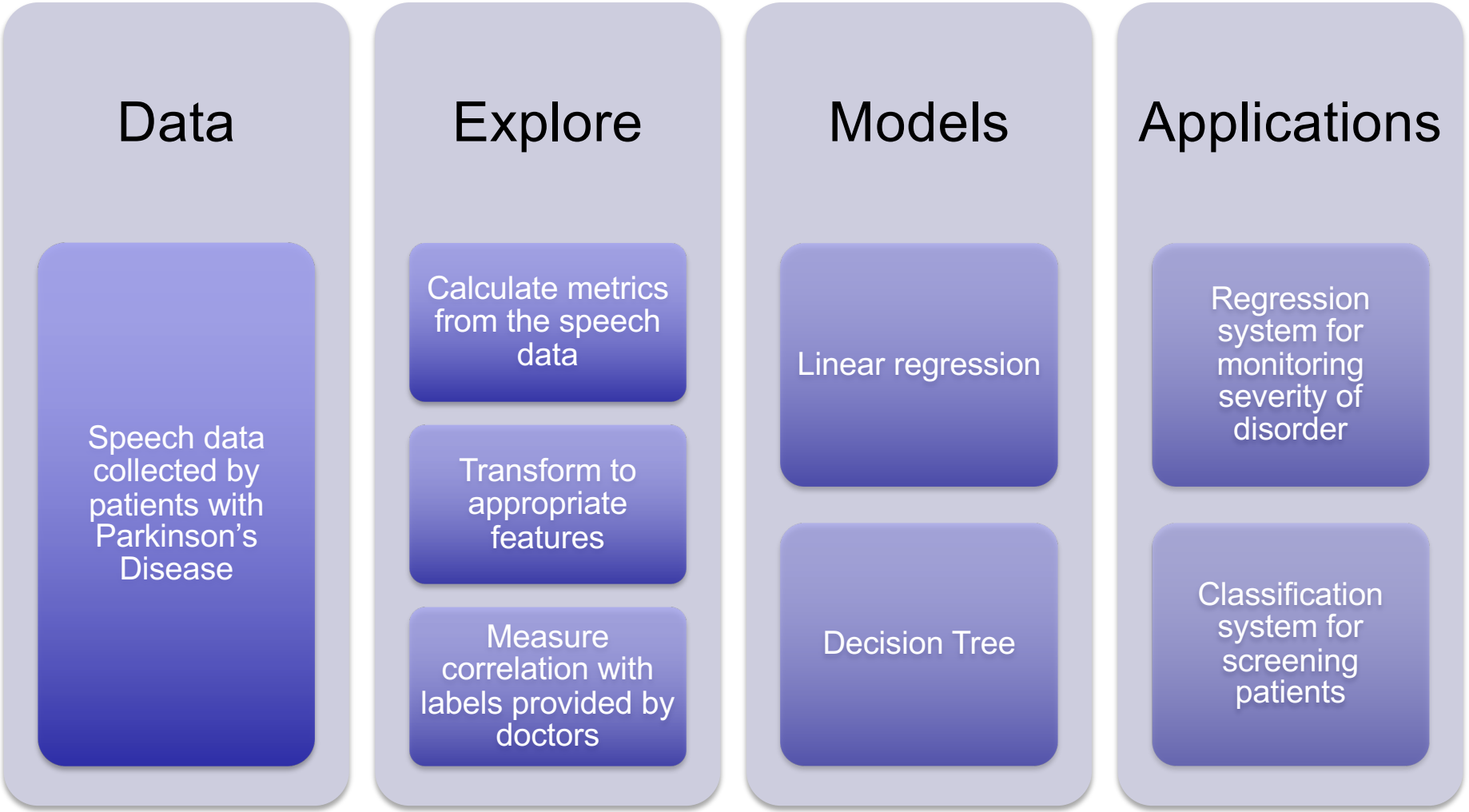
- "*AI is the science and engineering of making intelligent machines, especially intelligent computer programs*" – John McCarthy.

- AI is flourishing at present because of advances in computer power, availability of large amounts of digital information (big data, open data), and enhanced theoretical understanding.



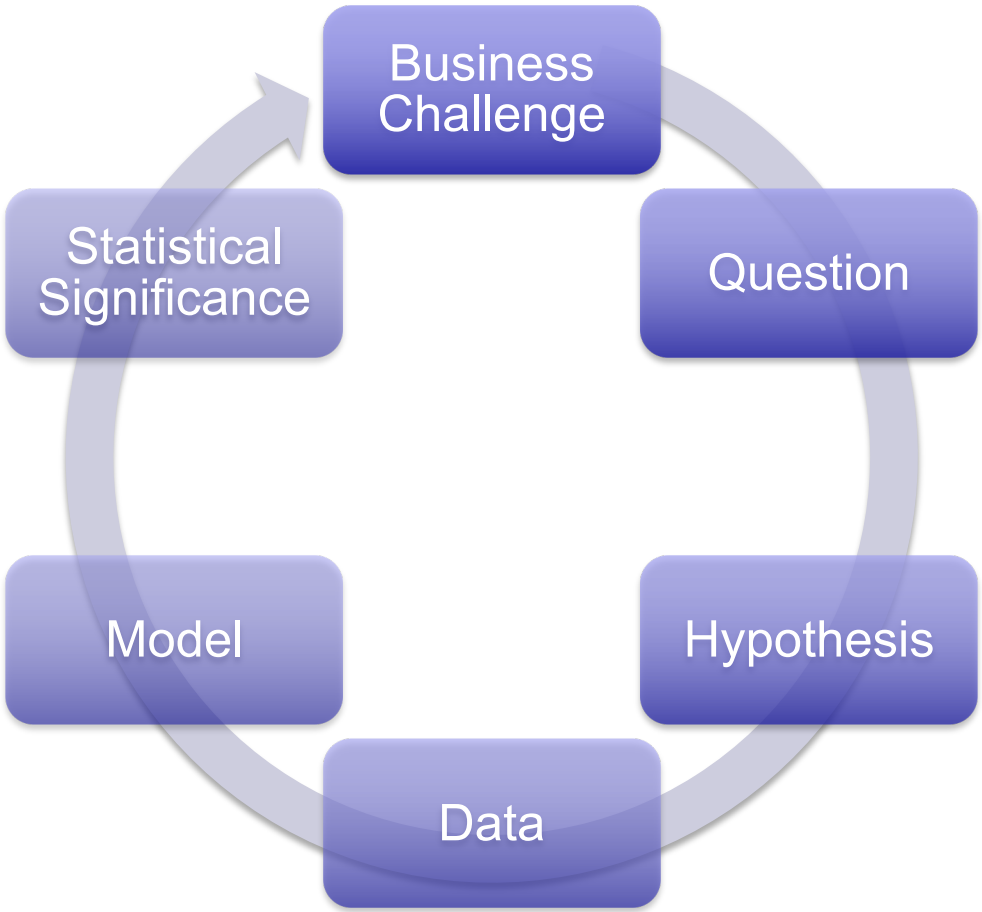


Workflow example



Roadmap

Data & Inference



Matlab

- MATLAB® is a high-level language and interactive environment for numerical computation, visualization, and programming.
- Using MATLAB, you can analyze data, develop algorithms, and create models and applications.
- The language, tools, and built-in math functions enable you to explore multiple approaches and reach a solution faster than with spreadsheets or traditional programming languages, such as C/C++ or Java™.
- Available from www.mathworks.com and also online.
- http://www.cmu.edu/computing/software/all/matlab/download.html

Other software

- There are a number of software platforms available for data analysis and machine learning
- Python: https://www.python.org
- R: https://www.r-project.org
- The course is designed around Matlab and TAs can only offer specific help with Matlab during the assignments.

Data

Statistics

- Statistics is the “science of making decisions in the face of uncertainty” – C. Chatfield.
- It is particularly useful for dealing with situations where there is measurement uncertainty.
- Data results from measurements and this is the starting point.

Initial questions

- Who owns the data?
- Is the data open access?
- Are there privacy concerns?
- Who is in charge of collecting the data?
- Who is responsible for quality control?
- Why is the data collected?
- What is the data meant to show?
- Who currently uses the data?
- What does the data tell us?

Example 1: Official statistics

- Governments routinely report official statistics about the state of the economy.
  - Gross domestic productivity
  - Inflation
  - Consumer price index
- The World Bank collects data on many indicators by country for each year.
- http://data.worldbank.org

Example 2: Development

- Human development index:
  - Economic activity
  - Life expectancy
  - Education
- Millenium Development Goals
- Sustainable Development Goals (SDGs)

Example 3: Industrial data

- Number of products manufactured
- Number of tickets sold
- Volume of sales
- Revenue generated
- Number of customers serviced
- Amount of airtime used

Example 3: ICT data

- Fixed-telephone subscriptions
- Mobile-cellular subscriptions
- Active mobile-broadband subscriptions
- Fixed (wired)-broadband subscriptions
- Households with a computer
- Households with Internet access at home
- Individuals using the Internet

Data types

Type	Description	Example
Nominal	Names	Identification numbers
Ordinal	Rank order	Income levels (low, medium, high)
Measure	Relative values	Average arrival rate
Count	Total values	Number of intrusions

Data representation

- In practice, we may want to measure continuous values
- On a computer, continuous recordings become integer values due to truncation
- This suggests that the observations will be quantised
- Suitable model structures may need to be selected to reflect this fact

Observational uncertainty

- Uncertainty influences quantitative modelling in many ways (observational, parametrical and structural)
- Observational uncertainty results from measurement errors, missing values, truncation effects, instrumental error
- Measurement apparatus may have changed over time
- External factors may vary with time

Data collection

- In the physical and medical sciences – data is usually collected by sensors
- In the social sciences – data collected using interview, surveys or focus groups
- Human activity can be measured using accelerometers in smart phones
- Human behaviour can be understood from internet searches, financial transactions and social media

formerly RDDC: Rebecca Davis Dance Company

MindLeaps

- Rebecca Davis Dance Company (RDDC):
- http://www.rebeccadavisdance.com/
- MindLeaps manages dance and education programs for street children in post-conflict and developing countries.
- Rebecca had the foresight to collect sufficient quantities of data with the hope of it being useful in the future.

MindLeaps Objectives

- Monitor performance of program over time
- Identify and understand the areas where the program is successful and unsuccessful
- Distinguish performance for individual skills
- Track performance of individual students

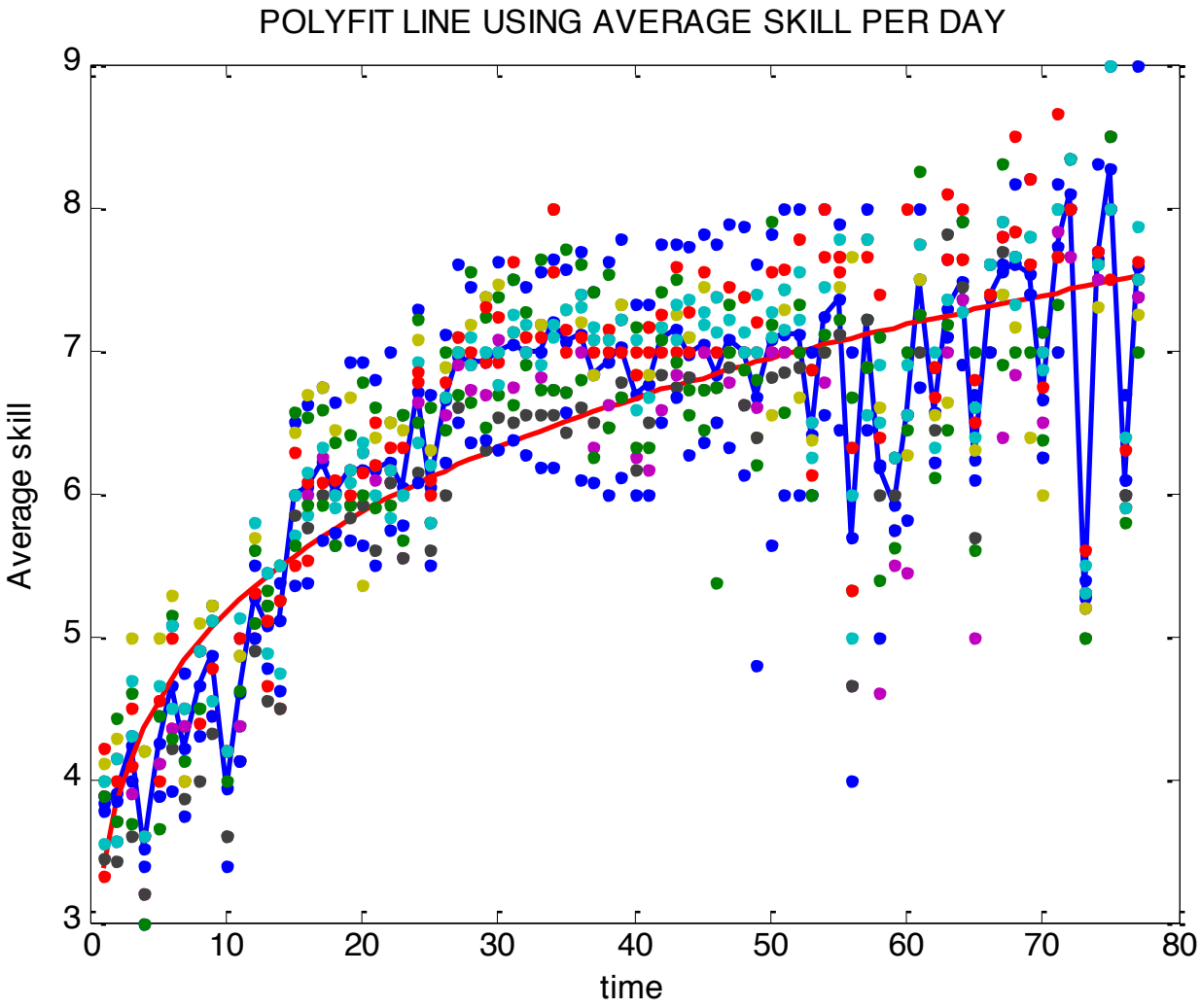
MindLeaps Data

- 11 cognitive skills measured each day
- These skill variables can be divided into three groups:
  - core,
  - behavior
  - expression

3D data structure

- 3D data structure containing:
  - N students
  - M variables
  - T days
- Visualization: think of data as a cuboid and analysis performed on slices or using averages
- Average skill per day; average over students and variables

Average Skill



MindLeaps Performance

A non-linear model providing a sigmoidal response was fit to the data to describe the learning process.

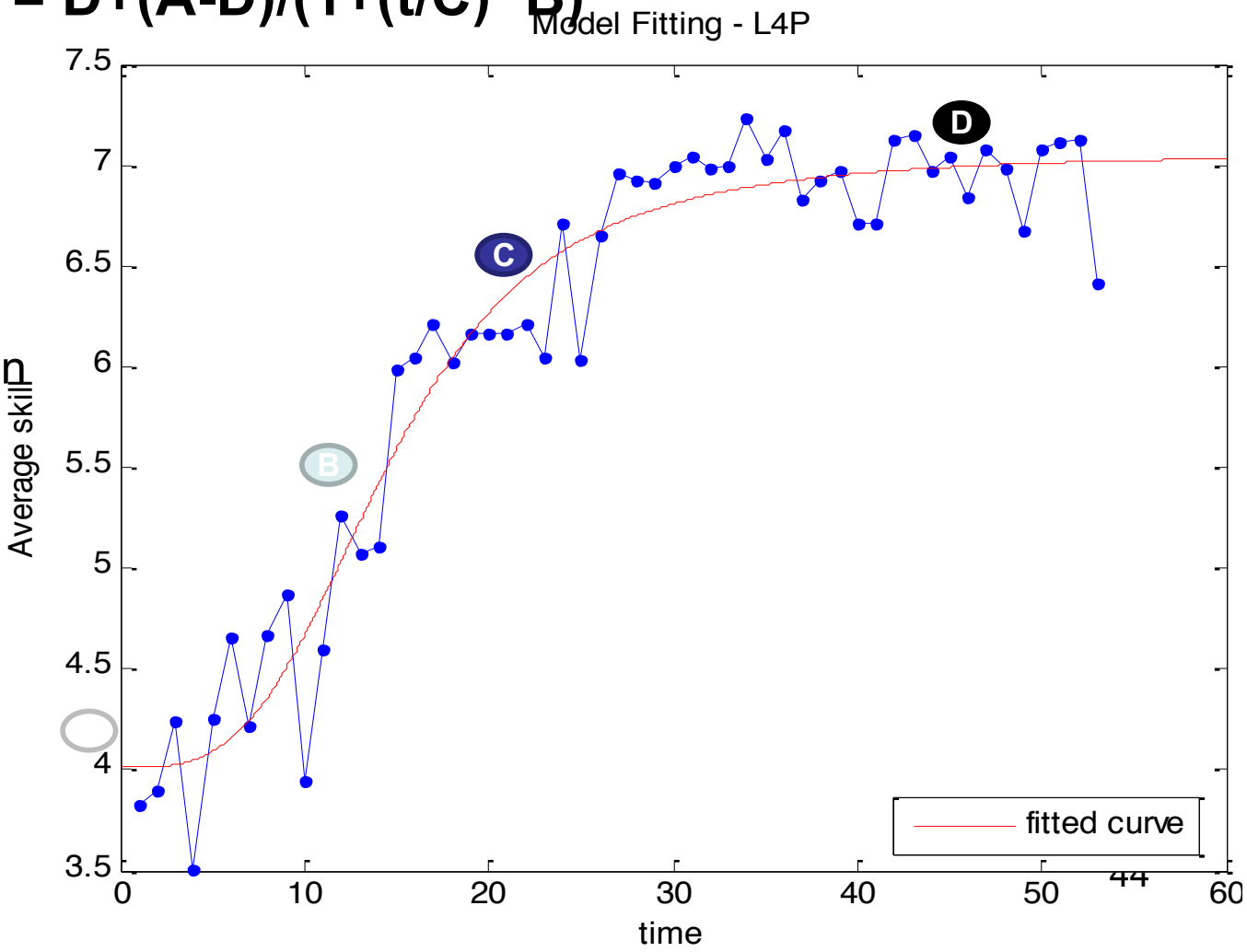
Model Equation: F(t) = D+(A-D)/(1+(t/C)^B)

A: Starting point

B: Rate of learning

C: Acceleration/deceleration

D: Saturation point



MindLeaps Discussion

- Data collection (paper or electronic)
- Issues with data (outliers, missing values)
- Analytics
- Insights

Data & Inference

WEEK 1B

Assignment 1

- Q1-Q4: Computing mathematical solutions using computer. e.g. bank accounts, loans
- Q5: assess borrowing for a start-up
- Q6: graph Ebola time series data
- Q7: determine average growth rate
- Q8: average ratio of Ebola deaths to cases
- Q9: graph time series for equities & bonds
- Q10: calculate average, min and max daily returns

Assignment 1

- Linear interpolation can be used when trying to understand the past
- Interpolation is dangerous when forecasting or developing early warning systems as data is leaked into the future
- Safest to take most recent observation in financial applications
- Daily Returns, r(t), can be calculated from daily closing prices, p(t) using:

$$r(t) = [p(t)/p(t-1)]-1$$

Today's Lecture

No.	Activity	Description	Time
1	Challenge	Capturing data	10
2	Discussion	How to obtain and aggregate data	10
3	Case study	True Colours – mood ratings	10
4	Analysis	APIs, Classification, stability	20
5	Demo	Loading in data	20
6	Q&A	Matlab questions and feedback	10

Systems modelling

- Why are we modelling the system?
- The motivation may be to better understand the system
  - The model should be parsimonious where each parameter has meaning
- Alternatively the objective may be to forecast, classify, control or optimise
  - Accuracy rather than understanding may be most important

System observation

- What variables should we measure?
- The system may be univariate, multivariate or spatiotemporal.
- What sampling time should we use?
- For what duration do we need to measure?
- If we know about fundamental frequencies, then we can select appropriate values.

System definition

- Is the system a closed circuit?
- Can we identify a sub-system that can be modelled in isolation?
- Does the system influence another system that is of interest?
- An appropriate definition of system boundaries can be important.

Obtaining Data

- Data capture refers to the process of obtaining data for a computer
- Historical data may need to be converted
- Examples include:
  - Digitization: convert paper records to electronic
  - Manual collection: taking measurements and typing them into a computer
  - Questionnaires, surveys and focus groups

Automated Data Capture

- An automated process can be used to collect data
- A direct connection is made between the measurement apparatus and the computer and avoids the need for a keyboard and manual labor
- Examples include sensors, log records, bar code readers and document scanners

Accessing Data

- In many situations, we simply want to access data that someone else has collected
- Big data relies on us being able to aggregate data from different sources
- The “Open Data” movement is helping to make it easier to access data

Inference & Decision-making

Environment

Telecoms

Data

Energy

Finance

Healthcare

http://www.amazon.com/Big-Data-Revolution-insurance-discovering/dp/1118943716

Recent trends

- Mobile sensors are cheaper than ever and increasingly available for measuring a large variety of variables
- Mobile penetration rates are almost equal to the global population at 95.5% worldwide, including a 90.2% penetration rate in developing countries (ITU, 2014)
- Both human activity and machines are generating big data that offers a means of obtaining a better understanding of complex systems
- Computers are capable of processing and analyzing large volumes of information

Discussion

- Data collection in the classroom
- How good are we at taking measurements?
- Counting Test
- https://app.sli.do/event/lgzpcd2u

- Join at **slido.com. #21420**

What did we learn?

- Uncertainty
- Biases
- Awareness

True Colours

What is True Colours?

- True Colours is an online self-management system that allows you to monitor your symptoms and experiences using text, email and the internet.
- By answering questionnaires you create a record of how you are feeling and can see how it changes over time.
- You can use this to help you to manage your own health and to share information with your family, friends or care team.
- Your data is stored on a secure computer system.

How can it help?

- Monitoring your wellbeing with True Colours will help you to notice when your feelings are changing.
- You can then act quickly to stop things from getting worse.
- This online record can also be annotated to note items such as changes in medication, changes in environmental stressors, and behavioral changes that might have happened.
- True Colours naturally lends itself to self-management, and is often used alongside integrated self-help programs.

Participant Testimonial

- *In a whole host of ways this has proved to be a godsend. The regularity with which the prompt text arrives - infallibly each Monday, before noon - acts as a support, and indeed comfort, at times when illness seems to bring only irregularity and uncertainty.*

Testimonial

- *From a clinical point of view the effect of the system has been dramatic. We feel much better informed about how well our patients are and about estimating the effects of changes in treatments.*
- - John Geddes, *Professor of Epidemiological Psychiatry*

Mood data

- The data used in this study are from 153 patients with bipolar disorder whose mood has been monitored remotely.
- Demographic data on individual patients' age are available for 119 patients and gender for 120 patients.
- Ages range from 20 to 75 years and for those for whom the gender was available,47 are male and 73 female.

Depression

- Mood data are returned approximately each week and comprises the answers to standard self-rating questionnaires for depression and mania.
- The rating scale used for depression is the *Quick Inventory of Depressive Symptomatology—Self Report* (*QIDS – SR*16) which comprises 16 questions.
- This scale assesses nine symptom domains for depression.

Depression symptoms

- Depression symptoms (Diagnostic and Statistical Manual of Mental Disorders, 4th edition, Text Revision) :
  - Sleep
  - Feeling sad
  - Appetite/weight
  - Concentration
  - Self-view
  - Death Suicide
  - General interest
  - Energy level
  - Slowed down/Restless

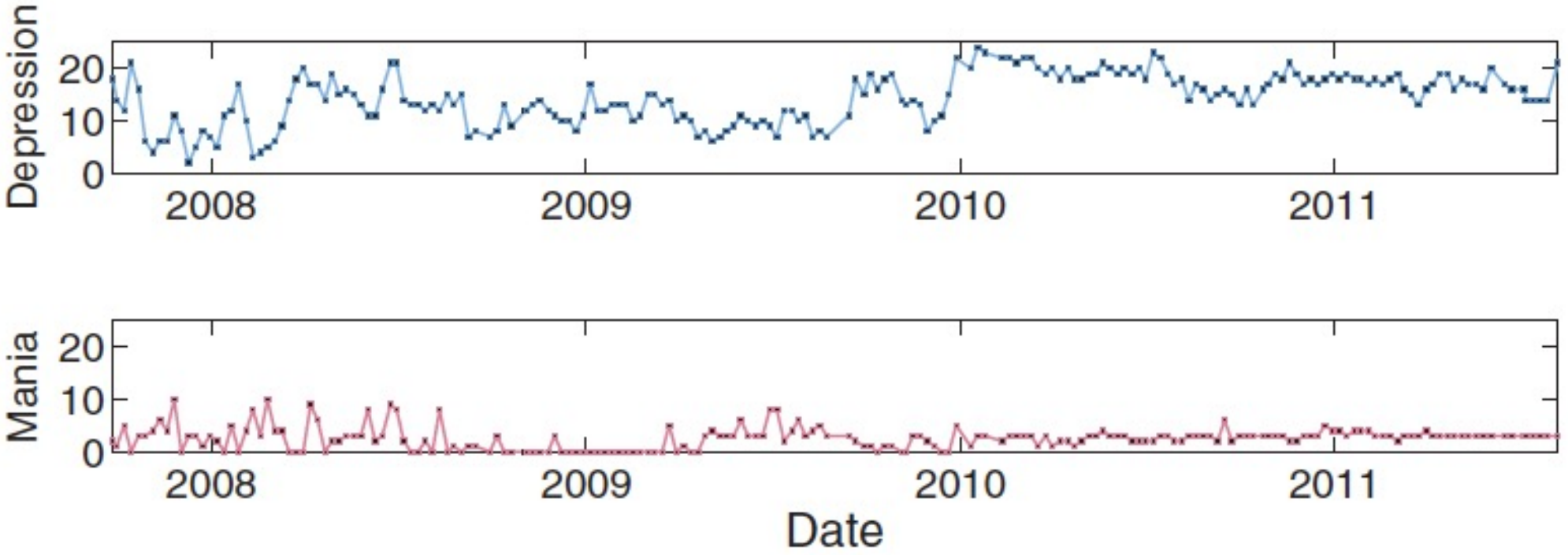
Depression - QIDS

- Each inventory category can contribute up to three points.
- The maximum score for each of the nine domains is totaled.
- This gives a total possible score of 27 on the QIDS scale.

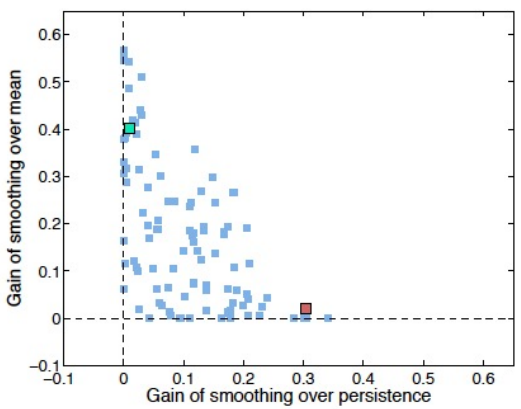
Mania - ASRM

- The severity of mania is quantified using the *Altman Self-Rating Mania Scale (ASRM)*.
- This scale consists of five items, each of which may can contribute up to four points.
- This gives a total possible score of 20 on the ASRM scale.

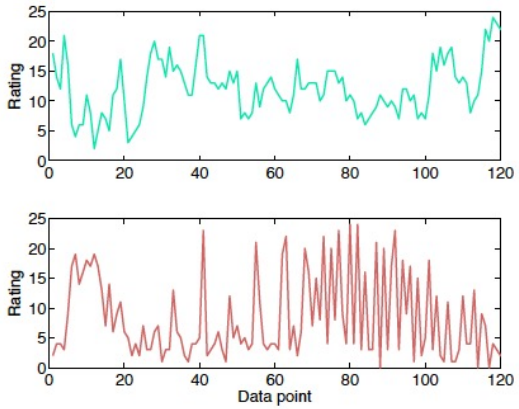
Time series


Depression
20
10
0
2008
2009
2010
2011
Mania
20
10
0
2008
2009
2010
2011
Date

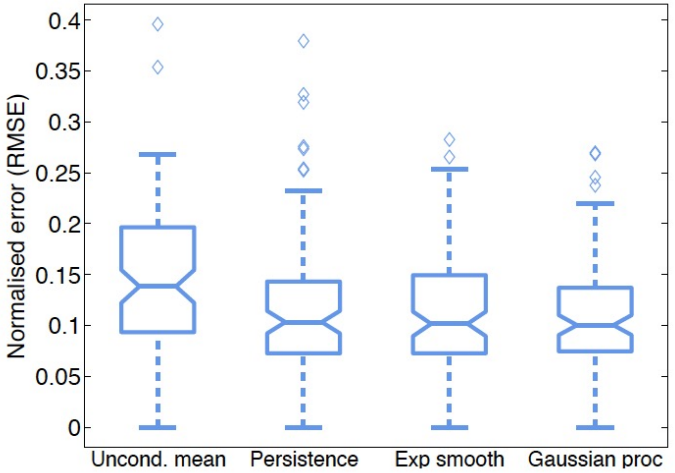
Forecasting depression in bipolar disorder



(c) Gain over baselines



(d) Examples

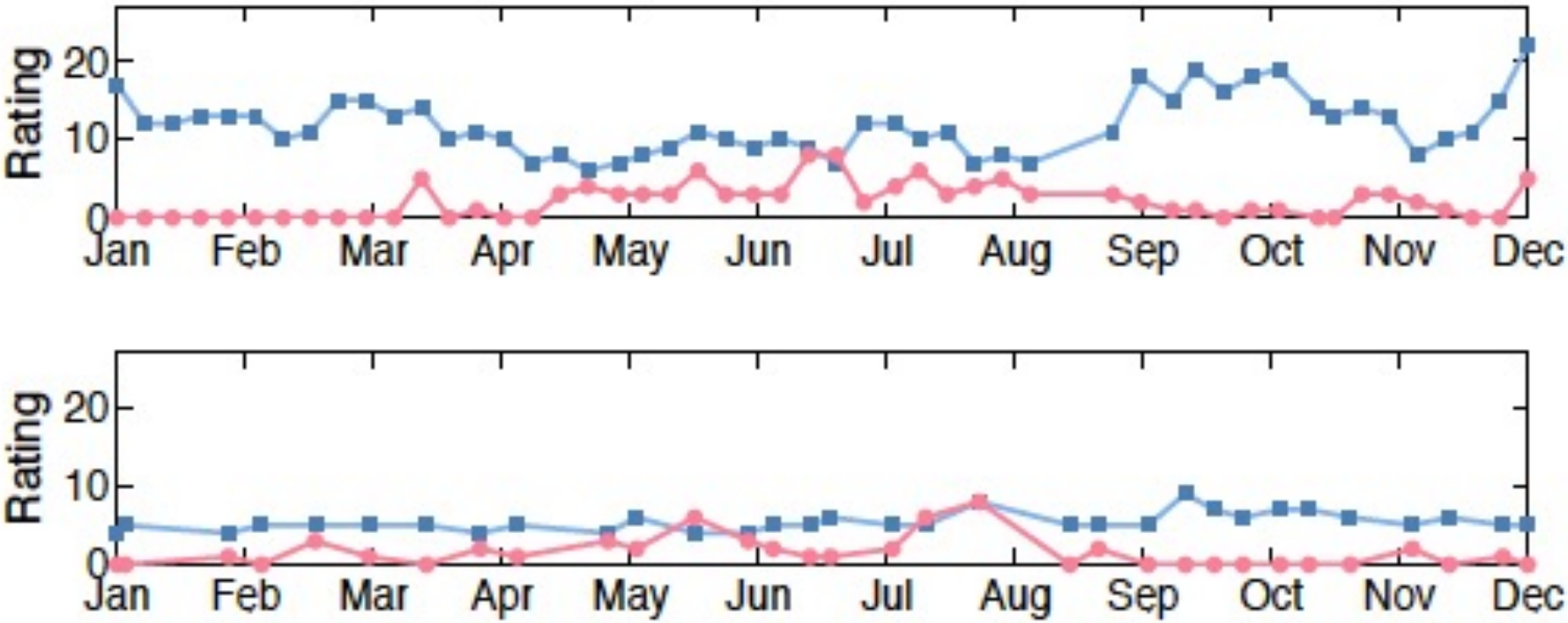


The left panel shows the gain of simple exponential smoothing compared with persistence (X-axis) and insample mean (Y-axis). Points in the top left region have a smoothing parameter close to 1, corresponding to a persistence forecast. Points in the bottom right region have a smoothing parameter close to 0, corresponding to a mean forecast.

- Telemedicine system give weekly SMS-based mood ratings
- Real data (especially telemedicine) is unevenly sampled and noisy.
- Metrics needed for measuring data quality (compliance and continuity).
- Testing hypothesis via multivariate predictive models.
- Actigraphy data will provide valuable information in the future.

Moore, Little, McSharry, Geddes & Goodwin (2012). IEEE Trans Biomed Eng.

Depression and mania



- Depression ratings (square markers) and mania ratings (circle markers). The maximum possible score on the QIDS depression scale is 27 and on the ASRM mania scale is 20.

Poll

- What do you need for constructing and operating a real-time daily prediction service?
  1) Static historical datasets
  2) Dynamic datasets from an API
  3) Both

slido.com #21420

Static Datasets

- Collect it yourself using questionaires, surveys or sensors
- Scrape data from a website
- Obtain from a third party source
- Download excel or csv file from a website

Dynamic data

- In reality, measurements are being continuously generated and we need to think about data streams
- For cutting edge applications, analytics should be generated in real-time
- The cost of delays could be substantial
- Real-time analytics facilitate immediate responses and timely decision-making

APIs

- Application Programming Interfaces (APIs) provide a means of obtaining live data in a structured format
- APIs are often used to access data allowing developers to build additional capability
- APIs provide internet bridges between different organizations
- Organizations such as Google, Yahoo and Twitter use APIs for data streams

API analogy

- You read the room service menu in a hotel.
- It lists all the drinks that are available.
- After selecting one, you place your order by telephone and wait for the staff to deliver it.
  - The menu acts like an API for the hotel
  - Ordering a drink is like executing an API call
  - The arrival of the drink is the requested response

The API/menu places structure on the process and ensures that the data/drink received is what you were expecting

Why are APIs useful?

- **Without an API:**
- An app finds the current weather in London by opening a website address, reading the webpage like a human and interpreting the content.
- **With an API:**
- An app finds the current weather in London by sending a message to the API in a structured format, such as XML, and waiting to receive a structured response.

Developing with APIs

- Websites may change the layout and structure of the information provided and therefore apps that have scraped information may stop working
- An API provides a robust means of exchanging data and information
- An API forces structure on the data-based exchanges

Exchange Rates

- Currency exchange rates are a good example of a quantity that changes over time
- At any given point in time we would like to know the latest exchange rate
- Openexchangerates.org provide an API for 165 world currencies, including historical data back to 1999.

Finance APIs

- Google finance (see googlesheets)
- Quandl (aggregator API)
- AlphaVantage: https://www.alphavantage.co
- Tiingo: https://api.tiingo.com
- Stooq: https://stooq.com
- DJI from Stooq:
- https://stooq.com/q/d/l/?s=^dji&i=d

Weather APIs

- Openweathermap.org have an API for historical data, current weather data, 5 and 16 day forecasts
- Wunderground.com has an API for weather data globally
- Forecast.io also has an API for weather and is specifically aimed at developers

API pricing model

- Many data-driven firms use the API as a distribution channel and the revenue stream is based around API calls.
- Forecast.io offers their API in both commercial and non-commercial applications.
  - The first thousand API calls you make every day are free, period.
  - Every API call after that costs $1 per 10,000 (that is, $0.0001).

Loading data into computer

- Numbers, Dates, Time;
- Strings and arrays;
- Indexing and subsets;
- Missing values and NaN
- Finding and matching values
- Using APIs for Google, Quandl

Poll on software

- Which software will you use for assignments?
  - Matlab
  - Python
  - R
  - Stata
  - SPSS
  - Other

  slido.com #21420

Matlab functions

- csvread, xlsread
- textread, fscan
- readtable
- datenum, datestr
- find, strcmp
- isnan