Data Analytics

Course: 18-787

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

Spring 2023

ICT Center of Excellence
Carnegie Mellon University

Data Analytics

WEEK 1A

Today's Lecture

No.	Activity	Description	Time
1	Challenge	Human versus machine	10
2	Discussion	Course objectives and timeliness	10
3	Case study	Data scientist	10
4	Analysis	Statistician versus data scientist	20
5	Demo	Roadmap	20
6	Q&A	Questions and feedback	10

Course Overview

- An introduction to the challenges of modelling real world systems using data
- An understanding of how data analytics are used in organizations to support decision-making
- Case studies of data science applications in forecasting, classification and risk management

Course Objective

- The objective is to facilitate the development of quantitative models for **real-world data analytics**, forecasting, decision support and risk management.
- The emphasis will range from forecasting in an **operational** setting to providing actionable insights to detecting anomalous behaviour to automated risk management systems.
- Using real-world datasets the course will show how to construct quantitative models for addressing specific questions in a range of different applications domains.
- The course will be delivered as a set of **case studies** and students areinvited to explore similar data sets during lab sessions.
- The motivating questions and pathway to solutions will be carefully described in each case.

Learning Objectives

- The course will build on the methods introduced in the “DIAML” course and show how different questions are addressed in practice.
- The course will emphasise the **challenges** and **limitations** of each of the modelling approaches.
- Through the lab sessions, students will gain **hands-on experience** analyzing similar time series datasets to those used in the lectures.
- Students will develop the **skills and confidence** needed to develop data analytics applications in several different domains.

Course outcomes

- After completing this course, students should be able to:
  - Analyze time series and identify what modelling approach is best
  - Follow a systematic approach to select and estimate models
  - Evaluate models and compare with appropriate benchmarks
  - Describe and understand model flaws
  - Communicate model limitations to non-experts
  - Summarise and explain the benefits of the model to end-users

Course outline

Week	Lecture A	Lecture B
1	Data Analytics	Weather forecasting
2	Renewable energy	Wind energy
3	Solar energy	Demand forecasting
4	Risk	Extreme events
5	Health	Biomedicine
6	Early warning systems	Economic forecasting

Poll

- Which of the course topics most interest you?

slido.com #29600

Decision-making

Environment

Telecoms

Data

Energy

Finance

Healthcare

http://eu.wiley.com/WileyCDA/WileyTitle/productCd-1118943716.html

Context

- Mobile sensors are cheaper than ever and increasingly available for measuring a large variety of variables
- The number of smartphone users in the world today is 6.648 Billion, which translates to 83.96% of the world's population owning a smartphone.
- In total, the number of people that own a smart and feature phone is 7.26 Billion, making up 91.69% of the world's population
- 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

Mobile Technology

- Infrastructure to collect time series data
- Data about the environment
- Data about individuals
- This could facilitate innovation in data analytics with significant societal impact
- Mobile Apps and SMS allow us to target and deliver analytics to customers in real-time
- Mobile is enabling the big data revolution

Mobile broadband Quiz

- Increase of 10% in mobile broadband penetration in Africa would yield an increase of x% in GDP per capita:
- 1%
- 1.5%
- 2%
- 2.5%

slido.com #29600

Impact of mobile broadband

- A regional study of the majority of countries in the Africa region confirms that an increase of 10 per cent in mobile broadband penetration in Africa would yield an increase in 2.5 per cent in GDP per capita.
- In addition, it suggests that a 10% drop in mobile broadband prices will boost adoption of mobile broadband technology by more than 3.1%.

ITU (2019). Economic contribution of broadband, digitization and ICT regulation: Econometric modelling for Africa.

The business case

- Business has realized that data is the new oil which must be extracted and utilized to create valuable products and services
- Data not only provides a competitive edge but it is becoming intellectual property in and of itself
- Data science is the process of refining the raw data
- The value of the Big Five (Apple, Microsoft, Google, Amazon and Facebook) is more than the value of *the next 27* most valuable US companies put together
- Firms like Uber, AirBnB, Tesla & Deliveroo are using data science to disrupt the market

Human experts versus machines

- Traditionally important decisions have been made by human experts
- Predictive analytics due to availability of data, computational resources and machine learning/statistical techniques
- Accounting for uncertainty is key for quantifying the confidence underlying the decision-making process

Intelligent decision-support

- Predictive analytics help organisations understand trends and patterns in human activity
- Facilitating the identification of relationships between explanatory variables and key performance indicators (sales, profit, risk reduction, efficiency)
- Having impacts on finance, insurance, energy, healthcare, government,…

More data or better models?

- In many disciplines, the focus is often on obtaining new predictive variables
- In medicine, traditional model structure is a logistic regression
- A powerful nonlinear classifier 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.

Data stream analytics

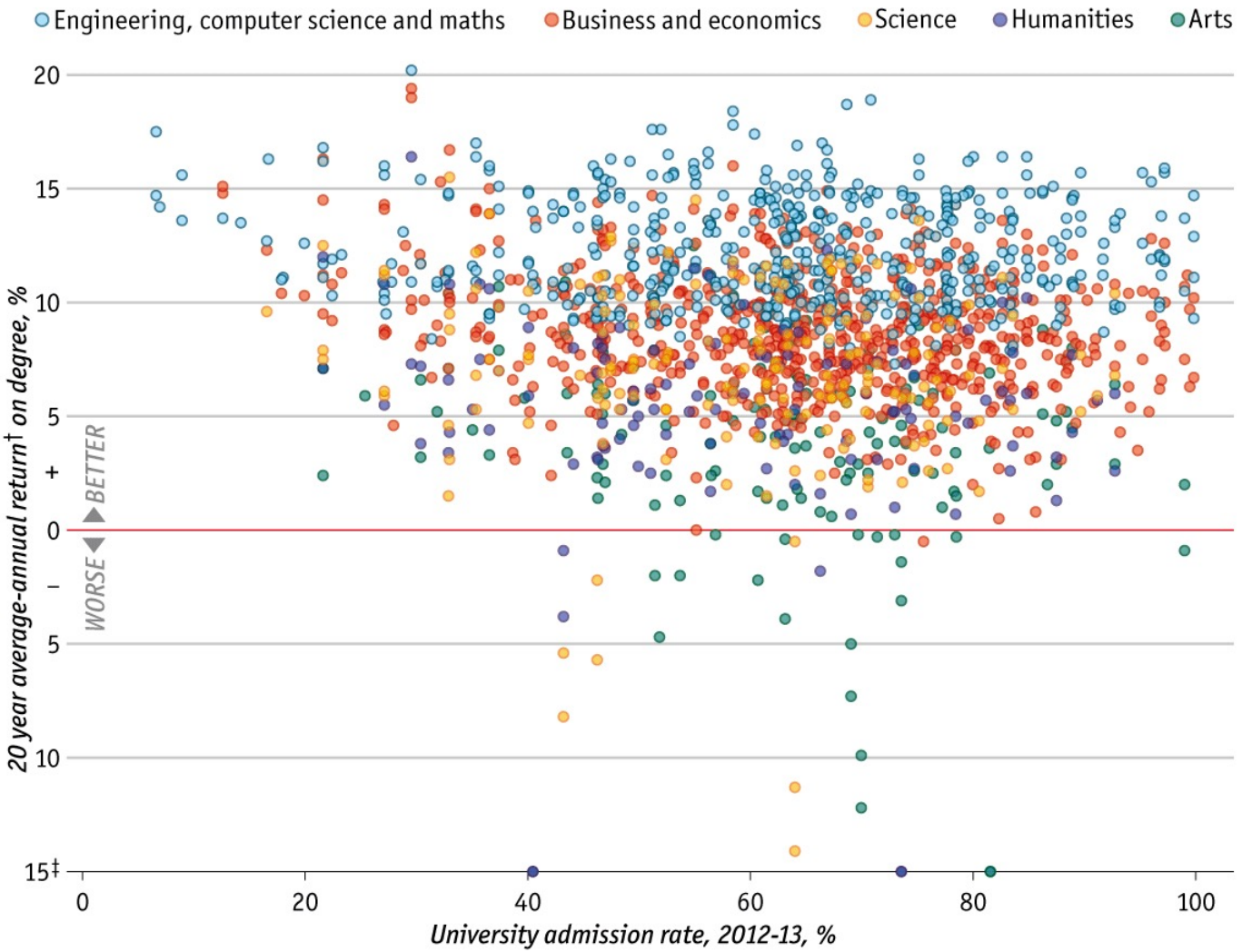
- Three key drivers:
- (1) information flow outstripping computational resources (since 2007, IDC)
- (2) a growing demand for intelligent decision support systems
- (3) reduced time available for decision-making in critical applications

McSharry, P.E. (2012). Stream analytics for forecasting. *Foresight: The International Journal of Applied Forecasting.*

Challenges/Opportunities

- Computation
  - Technical challenges to sift through large quantities of information
- Interpretation
  - Sentiment analysis, new quantitative/qualitative collaboration
- Transparency
  - Close the gap between data/models and policy
  - Open access/open source approaches
- Prediction
  - Evidence-based policy requires standards, evaluation protocols and appropriate benchmarks
- Scenarios
  - Probabilistic decision support for increasing competitiveness, sustainability, reducing risk and enhancing quality of life

Degrees with the best financial returns



Engineering and computer-science students earn most, achieving an impressive 20-year annualised return of 12% (the S&P 500 managed just 7.8%). Source: Economist, Mar 6, 2015.

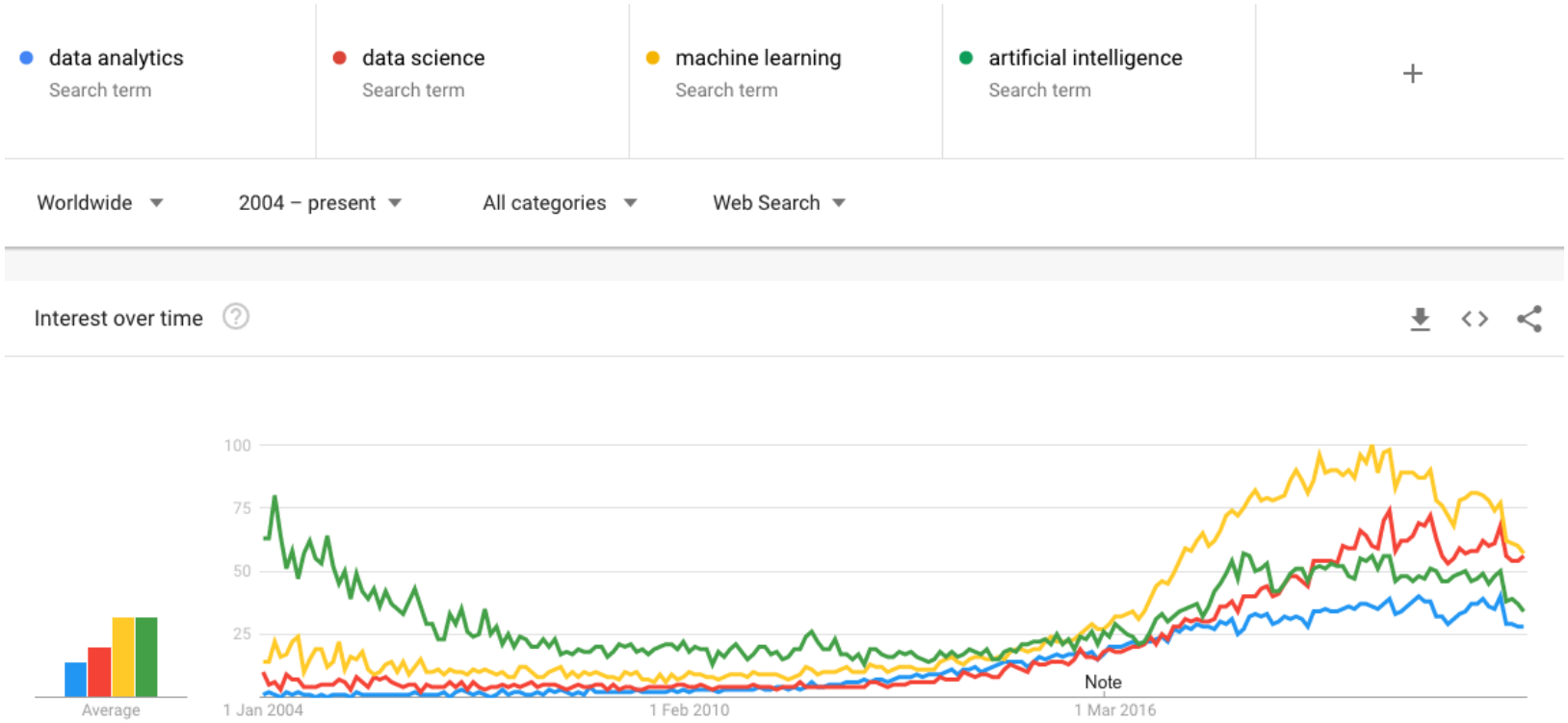
Poll

- Which of the following do you think is most popular?

1. Data analytics
2. Data science
3. Machine learning
4. Artificial intelligence

slido.com #29600

Interest in “Data Analytics”



Source: Google Trends, Worldwide, 2014 – 2022

Sexiest job of the 21st century

- Harvard Business Review called data scientists the sexiest job of the 21st century (Tom Davenport and D. J. Patil)

- Data scientist is "a high-ranking professional with the training and curiosity to make discoveries in the world of big data"

Motivation for students

- Data science sits at the interface of business, technology and science
- Opportunities in large established firms (Google, IBM, Amazon, Facebook) and start-ups
- Almost 50% of existing US jobs likely to be automated due to these technological developments (Economist, 2014)
- By 2025 the next wave of automation – turbocharged by the pandemic – will disrupt 85m jobs globally (WEF, 2020).

WEF (2020). Future of Jobs Report, Geneva, Switzerland.

Why you should become a Data Scientist ?

2,339 data science job listings from companies

Edmunds.com
PwC
Facebook
Fidelity Bank
Kohl's
Uber
Pinterest

Various Locations

California Boston New York Washington Atlanta Dallas Chicago

Average data science entry-level salary in

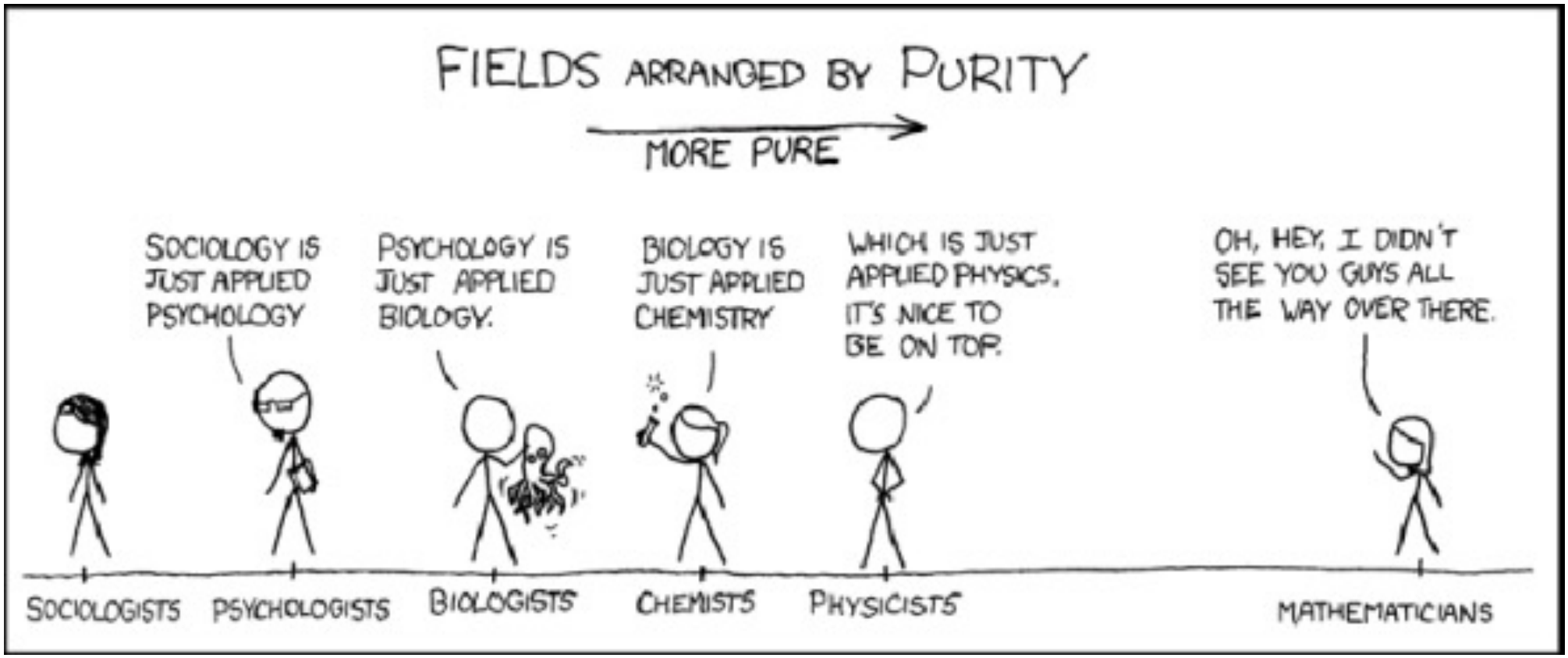
US
$91,000

Silicon Valley
$110,000

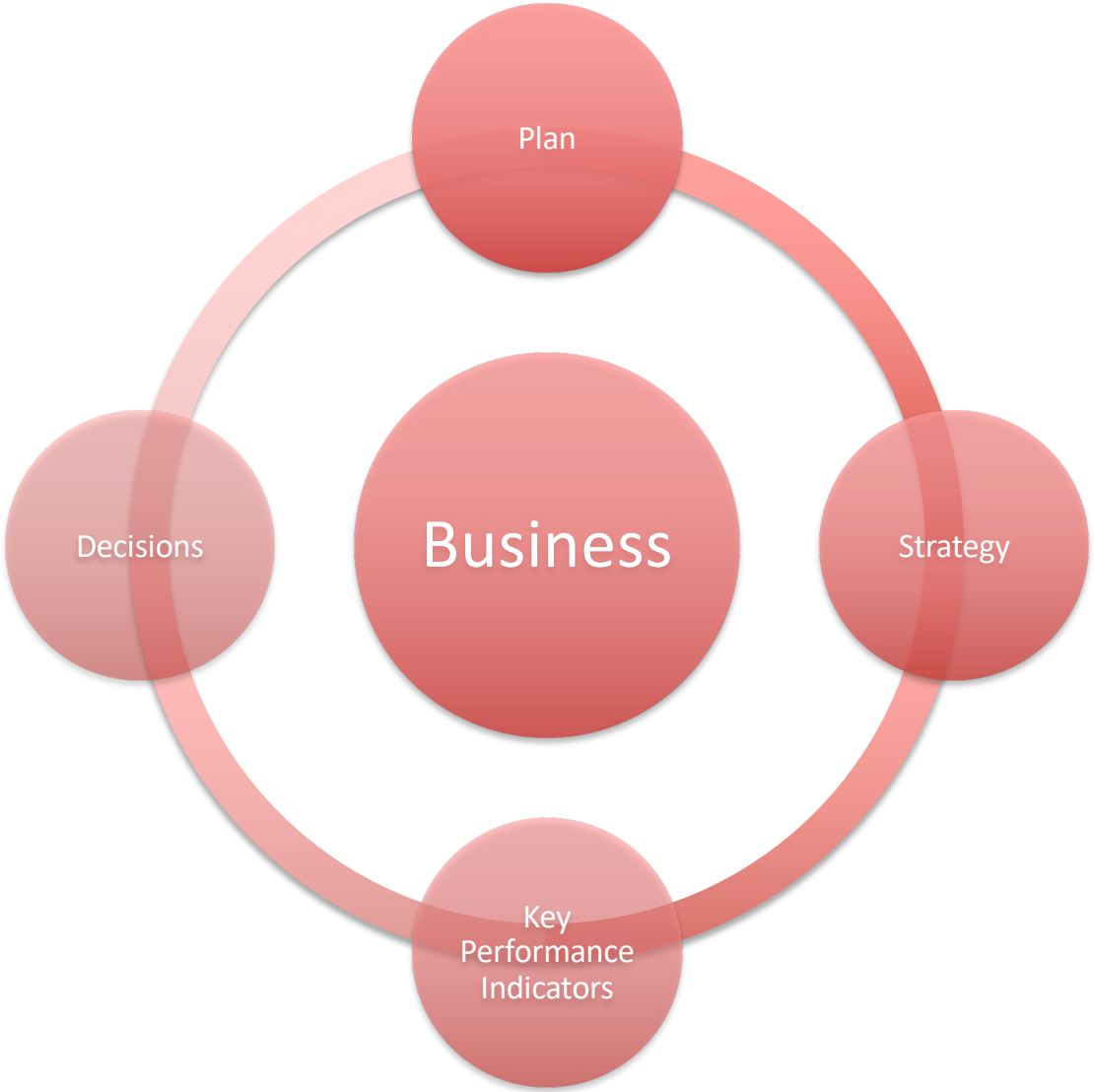
$140,000 Average data science managerial level salary ranges from **$240,000**

DeZyre

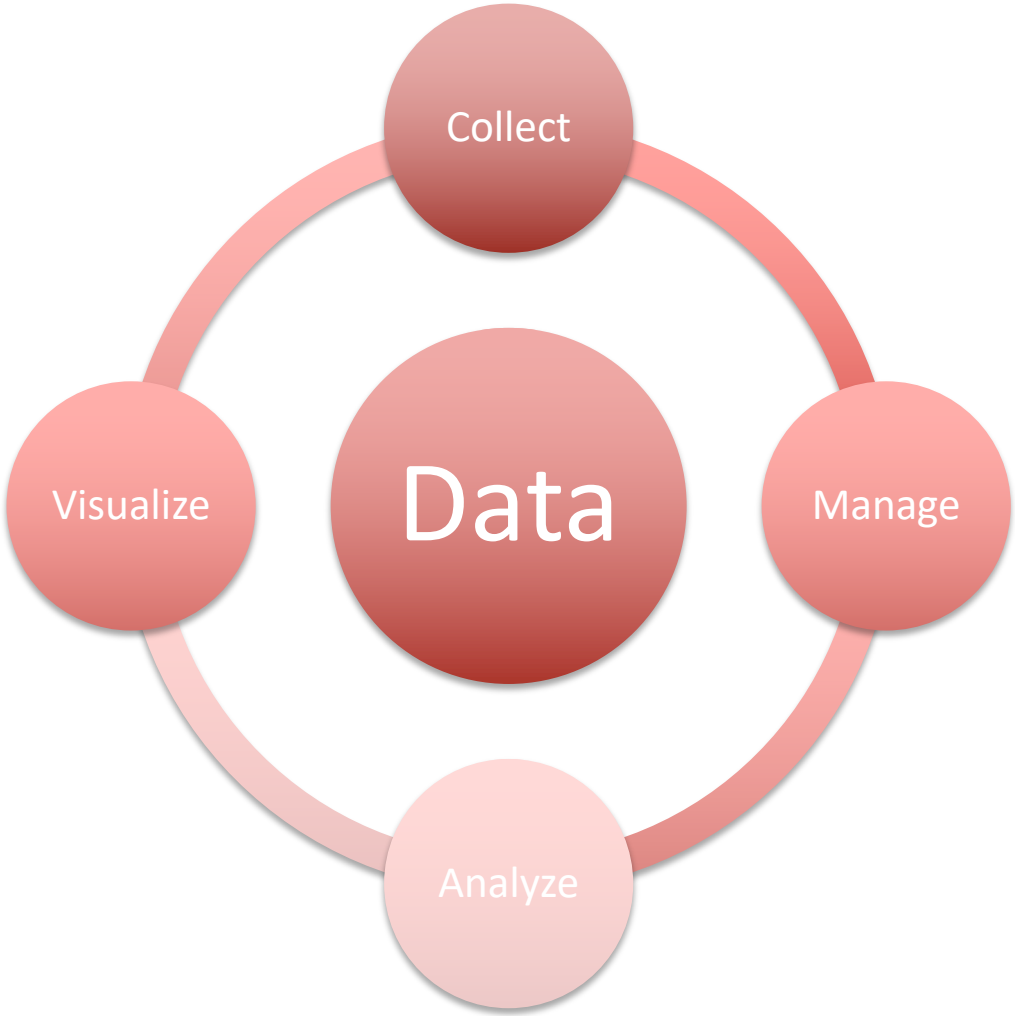
Applied Science



Business Objectives



Data Tasks



Data Scientist

- "Data scientists support critical business requirements by creating sophisticated methods of mining data for patterns that can be turned into business value"

- Source: Gartner (May 2012)

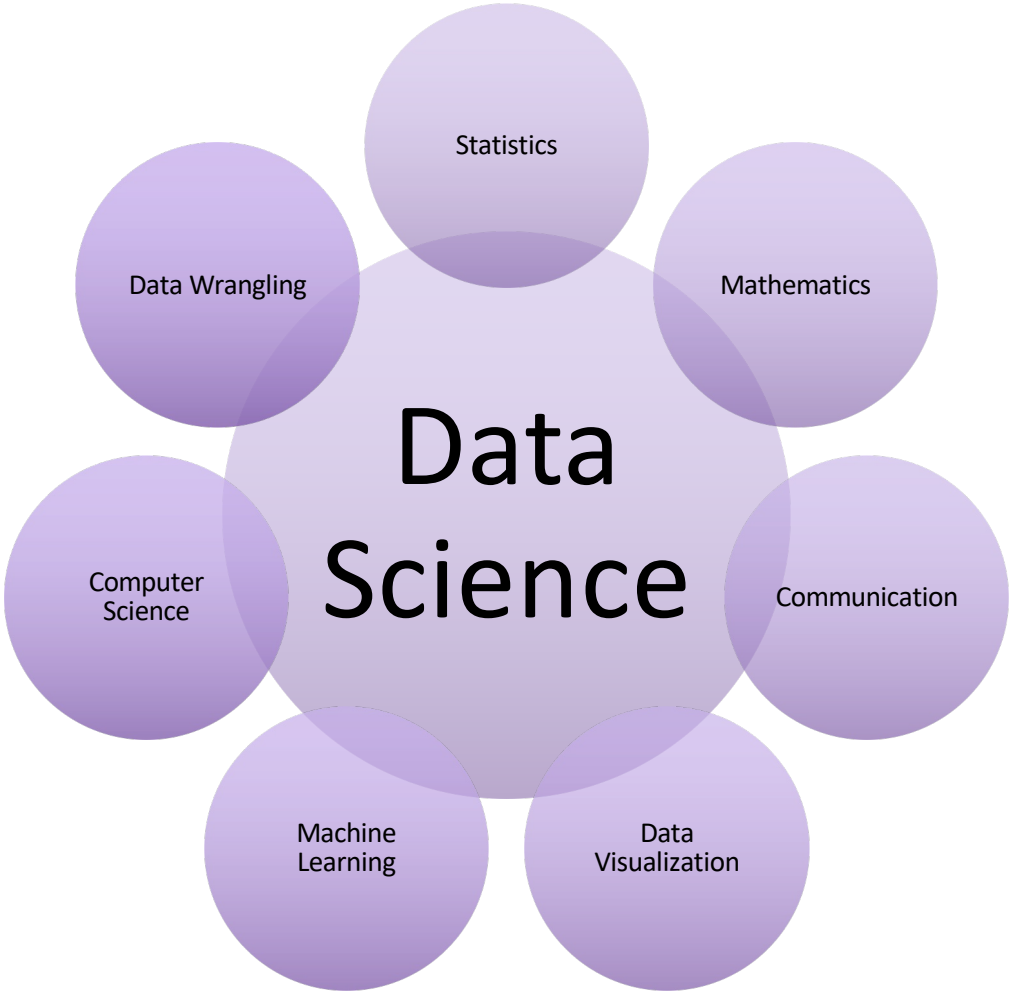
Intuition versus Data Science

Description	Intuition	Data Science
Amount of data	Small	Large
Understanding	High	Low
Transparency	Low	High
Visualization	Low	High
Repeatability	Low	High
Scalability	Low	High
Implementation Speed	Fast	Slow
Ease of Communication	High	Low

Statistician versus Data Scientist

Description	Statistician	Data Scientist
Data type	Clean and numerical	Structured & unstructured
Data size	Kilobytes	Gigabytes
Data frequency	Weekly to Yearly	Seconds to hours
Operation	Responsive	Consultative
Inputs	Data	Business challenge
Outputs	Tables	Graphs, visuals
Deliverable	Report	Algorithm, App or Product
Work preference	Alone	Multidisciplinary team
Software	Excel, SAS, SPSS, …	Matlab, R, Python, …

Data Science Skills



Poll

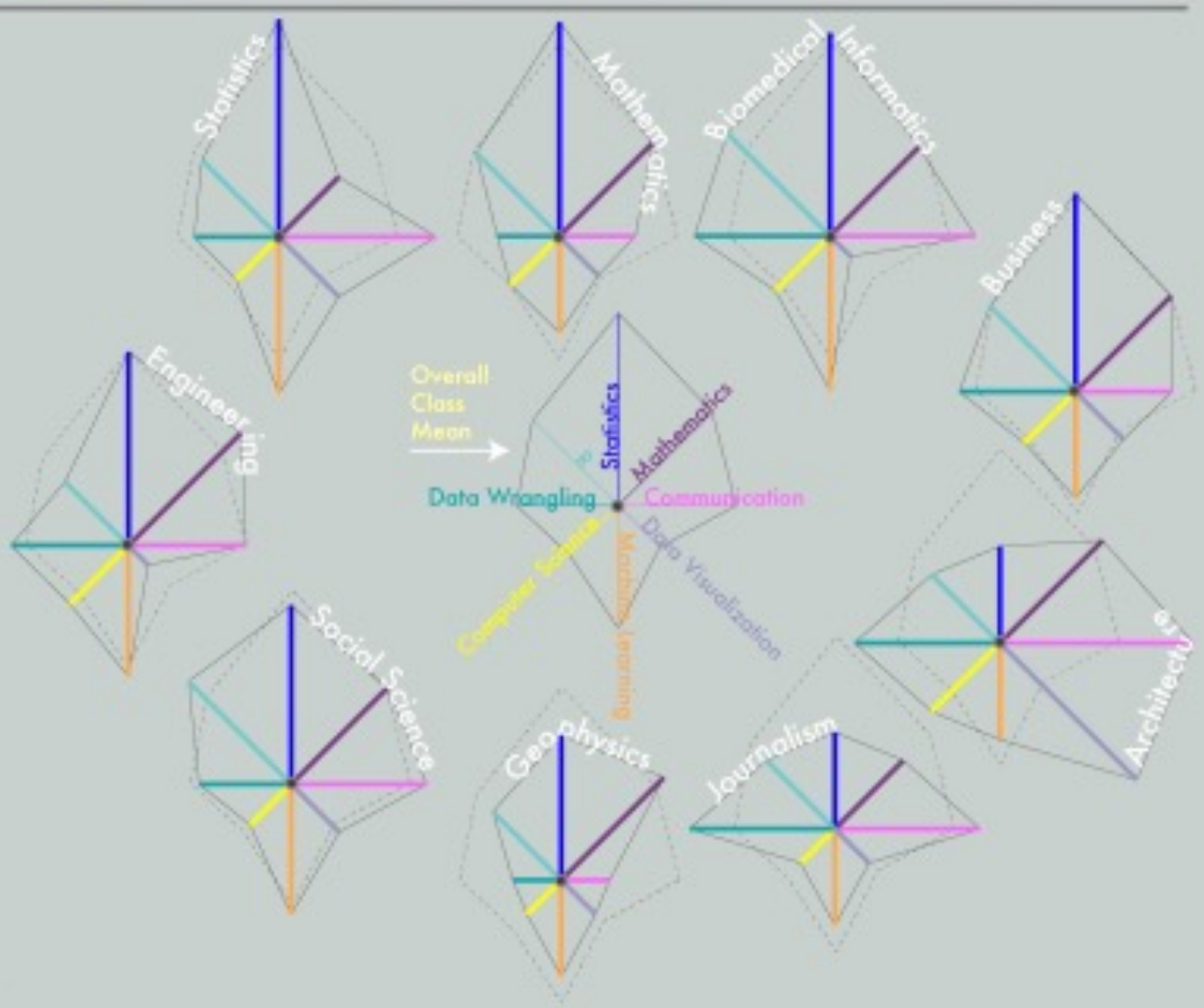
- Select the top three skills that you possess today:

1. Statistics
2. Mathematics
3. Communication
4. Data visualization
5. Machine learning
6. Computer Science
7. Data wrangling

slido.com #29600

The Stars of Data Science

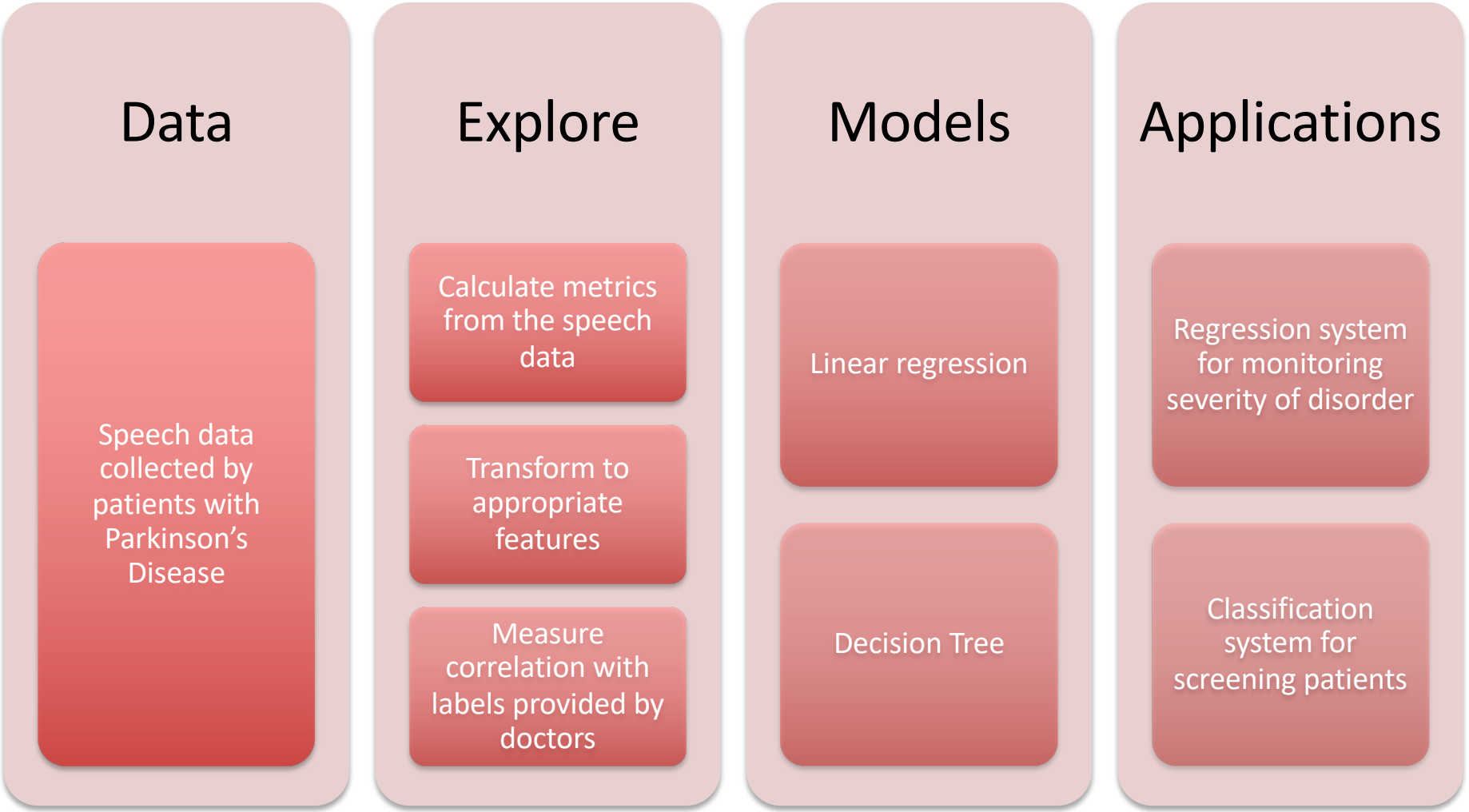
Students in Columbia's Introduction to Data Science course came from across the academic spectrum. Their skills are presented here in star charts with spokes representing their skill levels* across the data science skillset: R, **statistics**, **mathematics**, communication, data visualization, machine learning, **computer science**, and **data wrangling**. In addition to hovering in the center, the star chart of the overall class mean underlies each academic domain, so you can see students from each academic domain relative to the rest of the class. How would you compose your own intergalactic data science team?



*Skills were assessed by a survey written and administered by a subset of students in the class.

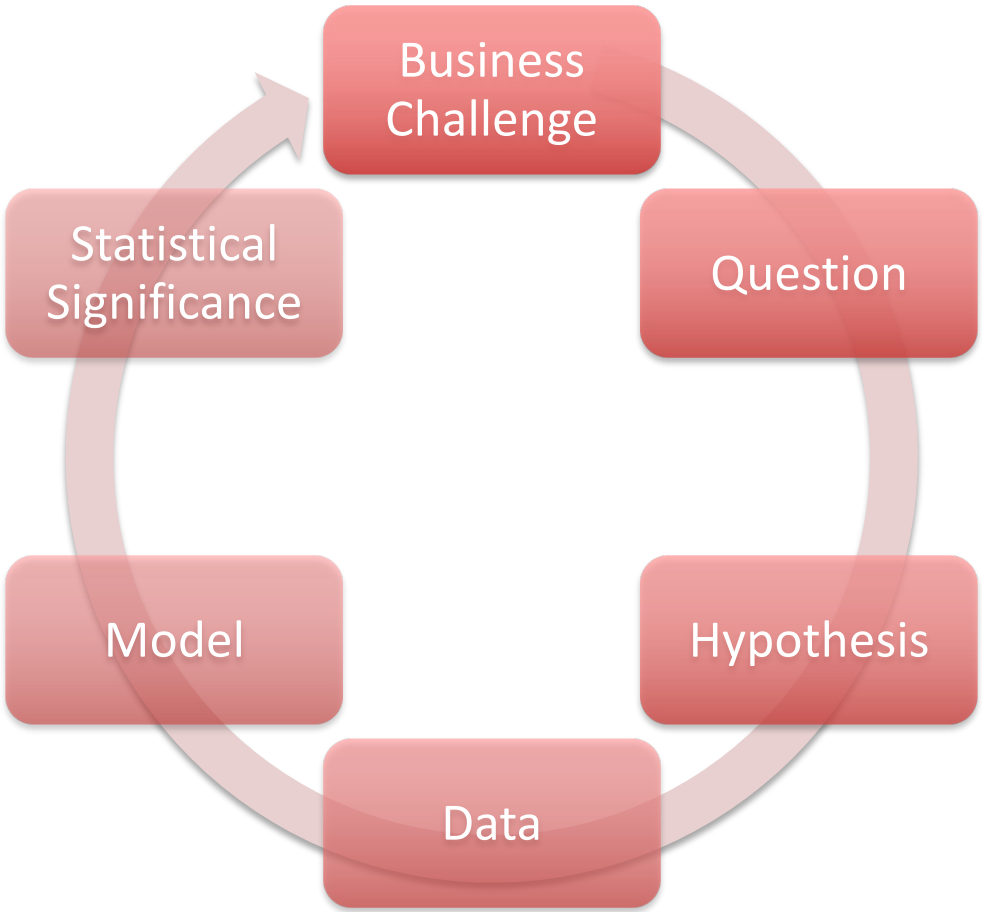

Data
• Structured numerical data: official statistics, signals (e.g. sales, demand, prices, weather)
• Big data (volume; velocity; variety): human activity (sensors, mobiles, accelerometers, movement, transactions, searches, webclicks); social media (SMS, emails, blogs, newsfeeds, voice, video, tweets, wikis); satellites, drones, …
Exploration
• Pre-processing; filtering; imputing missing values; observational uncertainty
• Statistical analysis, correlations, hypothesis testing
• Visualization tools for spatial and multidimensional data sets
Modelling
• Time series analysis; Linear stochastic model
• Dynamical models; nonlinear models
• Clustering (PCA, ICA, K-means)
• Machine learning (Neural Networks; KNN, SVM, CART, Random Forests, …)
Applications
• Forecasting, classification, decision-making, scenarios, policy-making, risk management and early warning systems
• Data mining: knowledge discovery (customer segmentation)
• Artificial intelligence: reproduce human knowledge, supervised learning, unsupervised learning, (e.g. pattern recognition; speech recognition)

Workflow example



Data Science Roadmap

Data Science Solution



What is big data?

- Three Vs:
  - Volume: fast growing volumes
  - Velocity: high frequency and real-time
  - Variety: structured and unstructured data (observations, transactions & interactions)
- Why now?
  - Storage and management of big data sets
  - Advances in computer science, math, statistics
  - Data about what people actually did and what they are doing in real-time

Needle in a haystack?

Assignment Structure

- Case study approach: preparation material (papers, reports, book chapters)
- Lectures – case study, discuss challenges and solution
- Focus on a paper to inspire assignment project with challenges and goals
- Recitation sessions:
  - Assignment explanation
  - Discuss project objectives
  - Discuss methodology
  - Discuss implementation
  - Questions and answers

Assignment components

- (1) analysis of a data set;
- (2) identification of statistically significant relationships;
- (3) hypothesis testing (specify the significance level);
- (4) clear explanation of the reason for selection particular quantitative techniques
- (5) construction of one or more quantitative models for a specific application; and
- (6) evaluation of the outcomes and comparison with appropriate benchmarks.

Assignment topics

Item	Description
Assignment 1	Weather dependent electricity demand
Assignment 2	Renewable energy generation
Assignment 3	Modelling intraday demand
Final Exam	Date/time to be confirmed

TA approach

- Recitations (<1h per session)
- Open discussions on Piazza & Canvas
- Agenda for recitations based on discussion and student questions
- Assigned TA hours for feedback

Submissions & responsibility

- Read CMU policy on academic integrity
- Manage time carefully: deadlines are strict with increasing penalties for being late
- Policy for consideration of unexpected events (sickness, accidents, technical) to be submitted copying faculty advisor with supporting evidence
- Submissions must be complete (all requested reports and files) with fully documented code (one line of text per line of code)
- Plagarism will not be tolerated: zero for assignment with any sign of plagarism detected

Q&A

Data Analytics

WEEK 1B

Course outline

Week	Lecture A	Lecture B
1	Data Analytics	Weather forecasting
2	Renewable energy	Wind energy
3	Solar energy	Demand forecasting
4	Risk	Extreme events
5	Health	Biomedicine
6	Early warning systems	Economic forecasting

Today’s Lecture

No.	Activity	Description	Time
1	Challenge	Weather forecasting	10
2	Discussion	History of weather forecasting	10
3	Case study	Michael Fish’s hurricane	10
4	Analysis	Communicating forecasts	20
5	Demo	Evaluating forecasts	20
6	Q&A	Questions and feedback	10

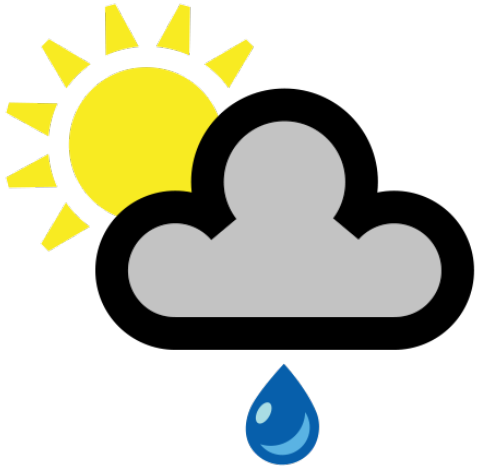
Assignment 1

- Weather data (daily): W
- Electricity consumption data (daily): C
- C = F(W,a) where F is a model to be constructed and a are parameters to be estimated
- What about time and recorded time stamps?
- Match dates in order to study the influence of W on C (data wrangling, any gaps, daylight savings)
- Correlation and scatterplots to study relationships
- C(t) related to W(t)?
- C(t) related to W(t-1)?

Weather

Source: BBC Weather

Weather



- Weather, n:
- The condition of the atmosphere (at a given place and time) with respect to heat or cold, quantity of sunshine, presence or absence of rain, hail, snow, thunder, fog, etc., violence or gentleness of the winds.

Source: Oxford English Dictionary

Weather and the economy

- Weather impacts many industrial sectors, including energy, agriculture, tourism, transport, retail & insurance
- Extreme weather events can have an adverse effect on revenues
- In developing countries, the impacts on agriculture can be devastating for the livelihoods of farmers

Weather forecasting

Source: BBC News (Prince Charles)

Weather variables

- Surface Air Temperature
- Surface Air MaxTemperature
- Surface Air Min Temperature
- Total precipitation
- 10 Meter wind speed
- Total cloud cover
- Sunshine duration

Weather maps

Source: BBC News (rain and snow)

Traditional weather forecasting

- Many traditional approaches are used for forecasting the weather.
- Farming communities have noticed that animals display certain characteristics before particular weather events.
- A famous old wives' tale suggests that a field full of cows lying down indicates a rainstorm is coming.

History of weather forecasting

- The Babylonians first attempted to use cloud patterns to forecast the weather in 650 BC.
- Aristotle wrote *Meteorologica* in 340 BC, which contains theories about weather phenomena such as rain, clouds and wind.

Renaissance

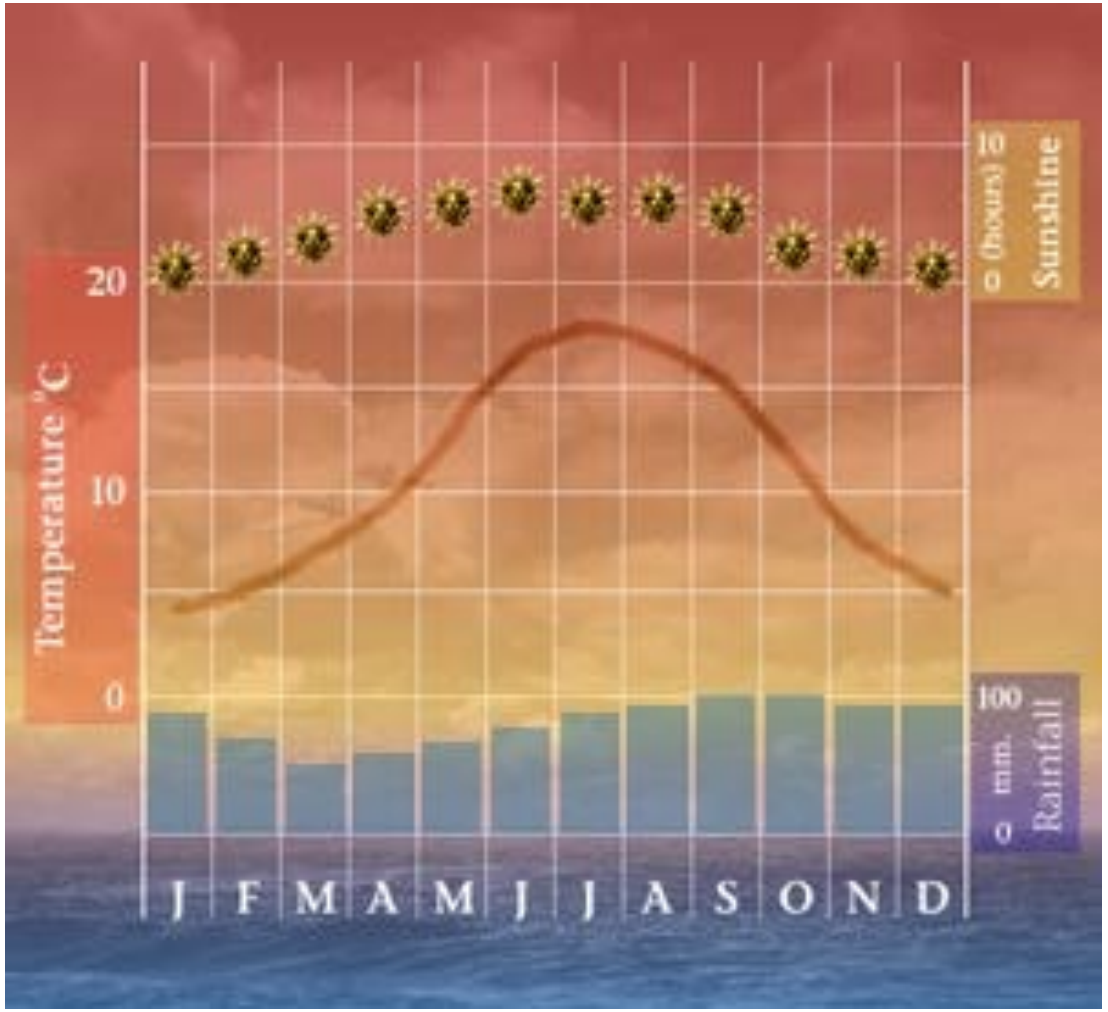
- It was not until the Renaissance period that Italian scientists invented instruments to measure the actual properties of the atmosphere.
- The thermometer was invented by Galileo Galilei around 1592.
- The barometer for measuring pressure was invented by Evangelista Torricelli in 1643.
- These instruments finally offered a means of observing and recording the weather.

Climate

- Climate is a measure of the average pattern of variation in temperature, humidity, atmospheric pressure, wind, precipitation, atmospheric particle count and other meteorological variables in a given region over long periods of time (e.g. 30 years).
- Climate is different from weather, in that weather only describes the short-term conditions of these variables in a given region.

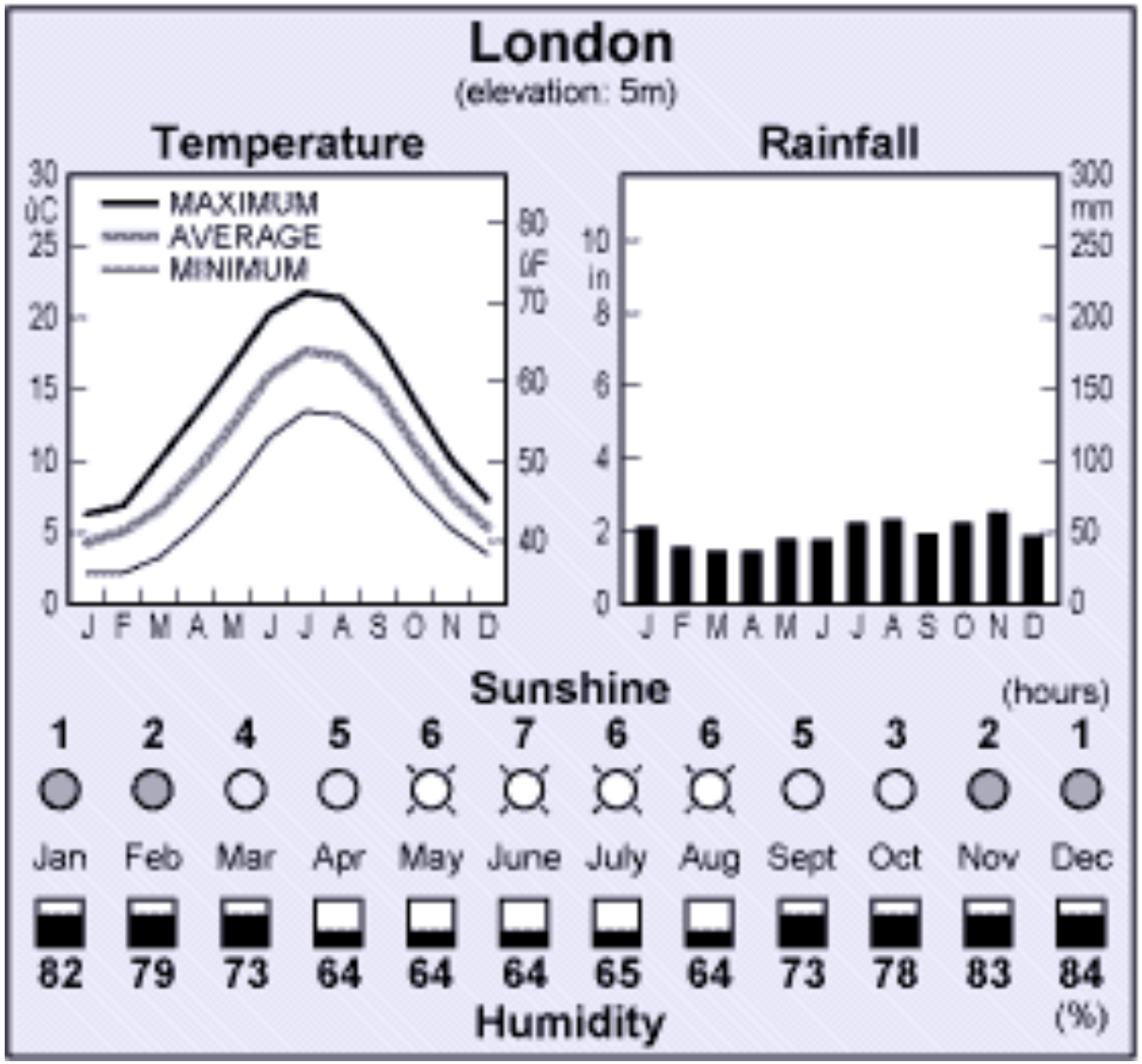
Source: wikipedia

Climate: London



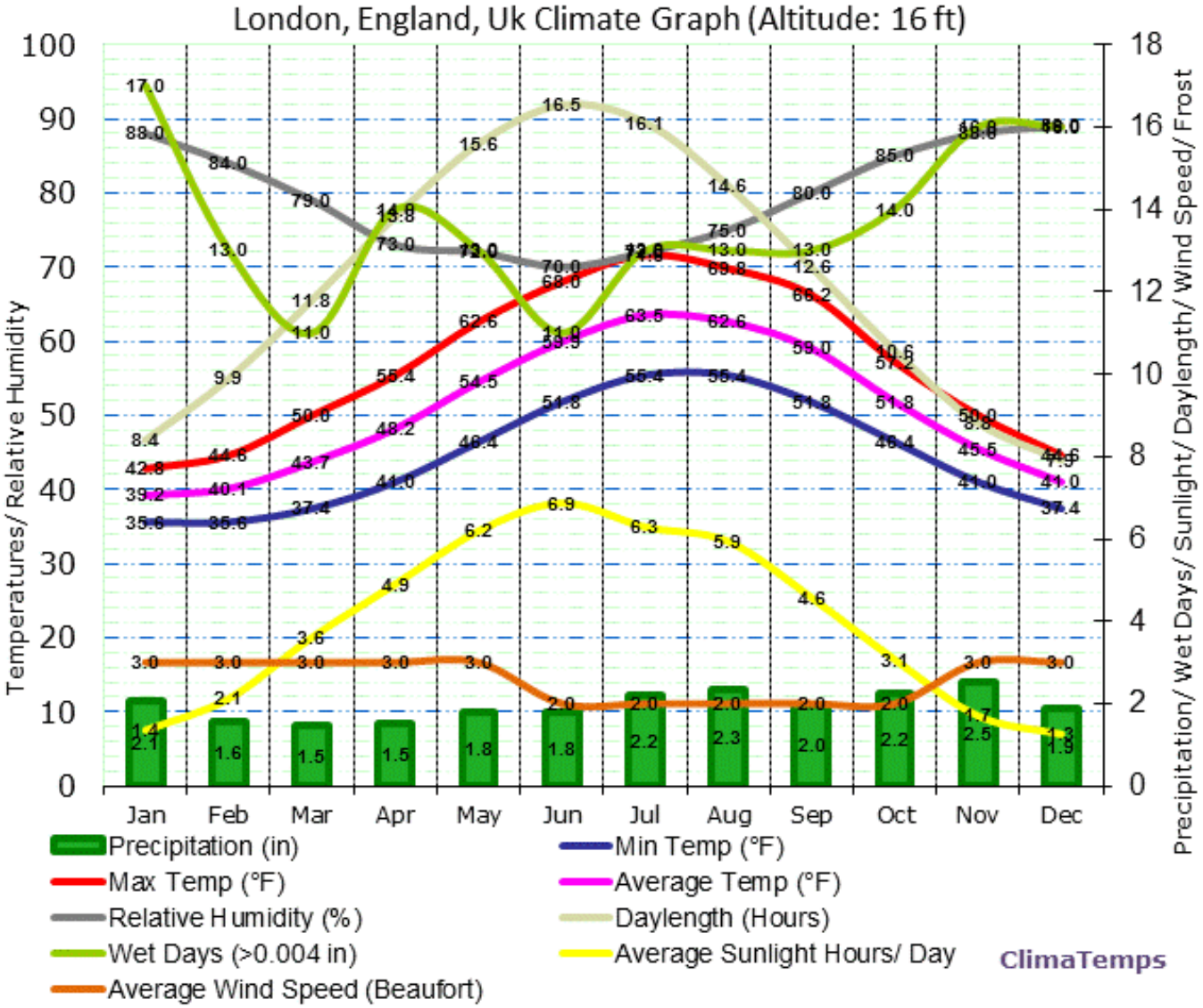
Source: www.travel.yaia.com

Climate: London



Source: www.travelpuppy.com

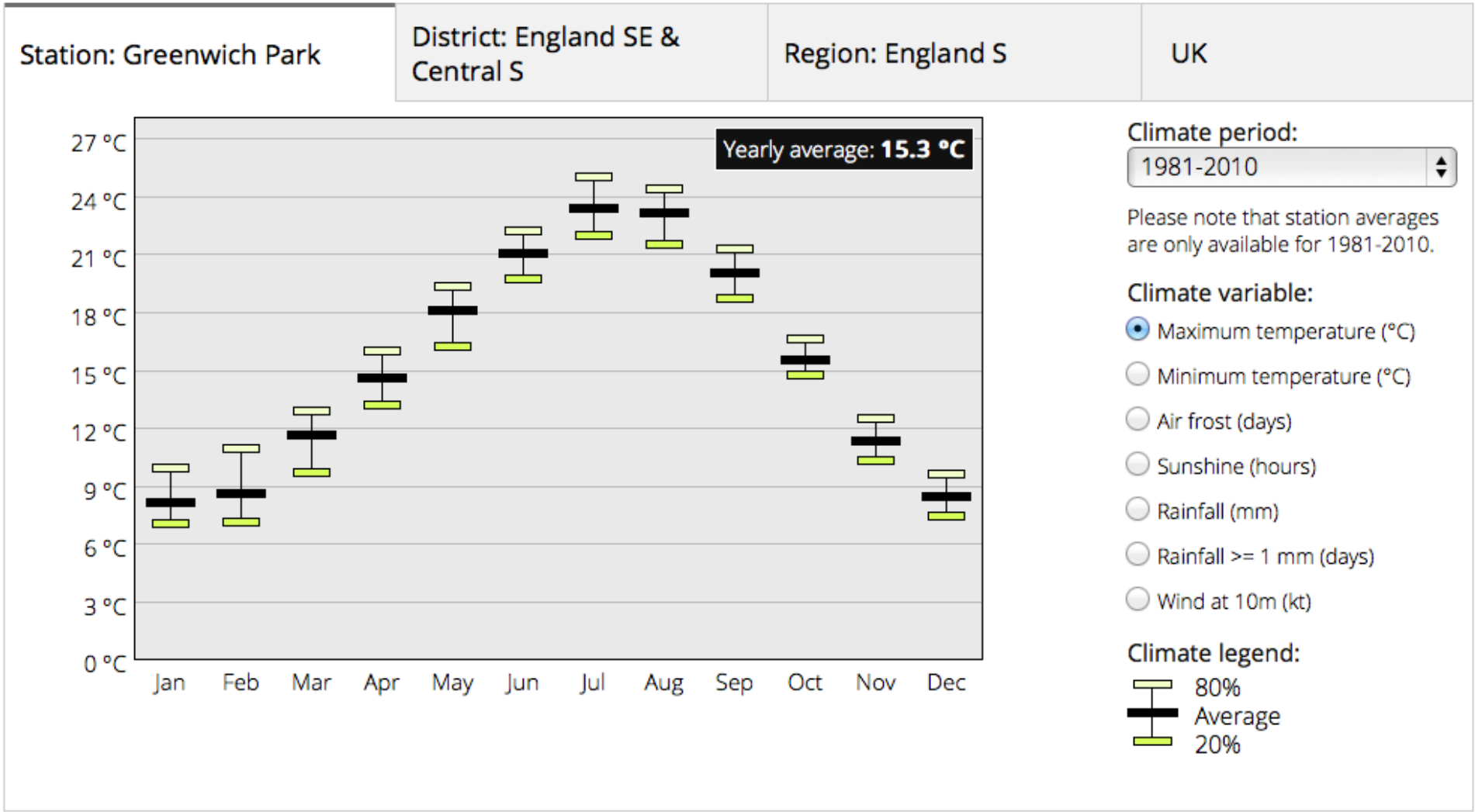
Climate: London



Source: www.climatemps.com

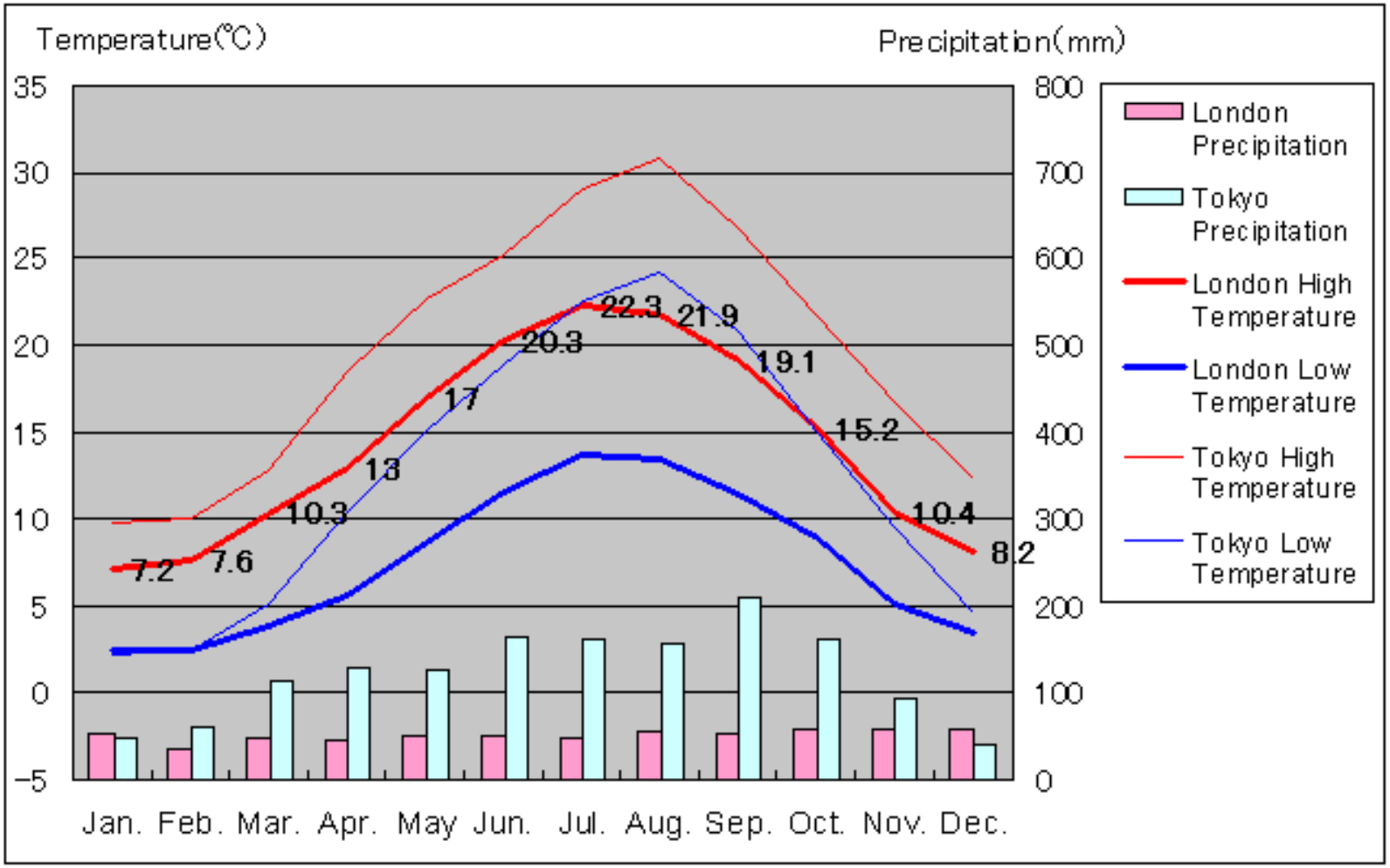
Climate: London

Greenwich Park (Nearest climate station to London)
Climate period: 1981-2010 - Maximum temperature (°C)



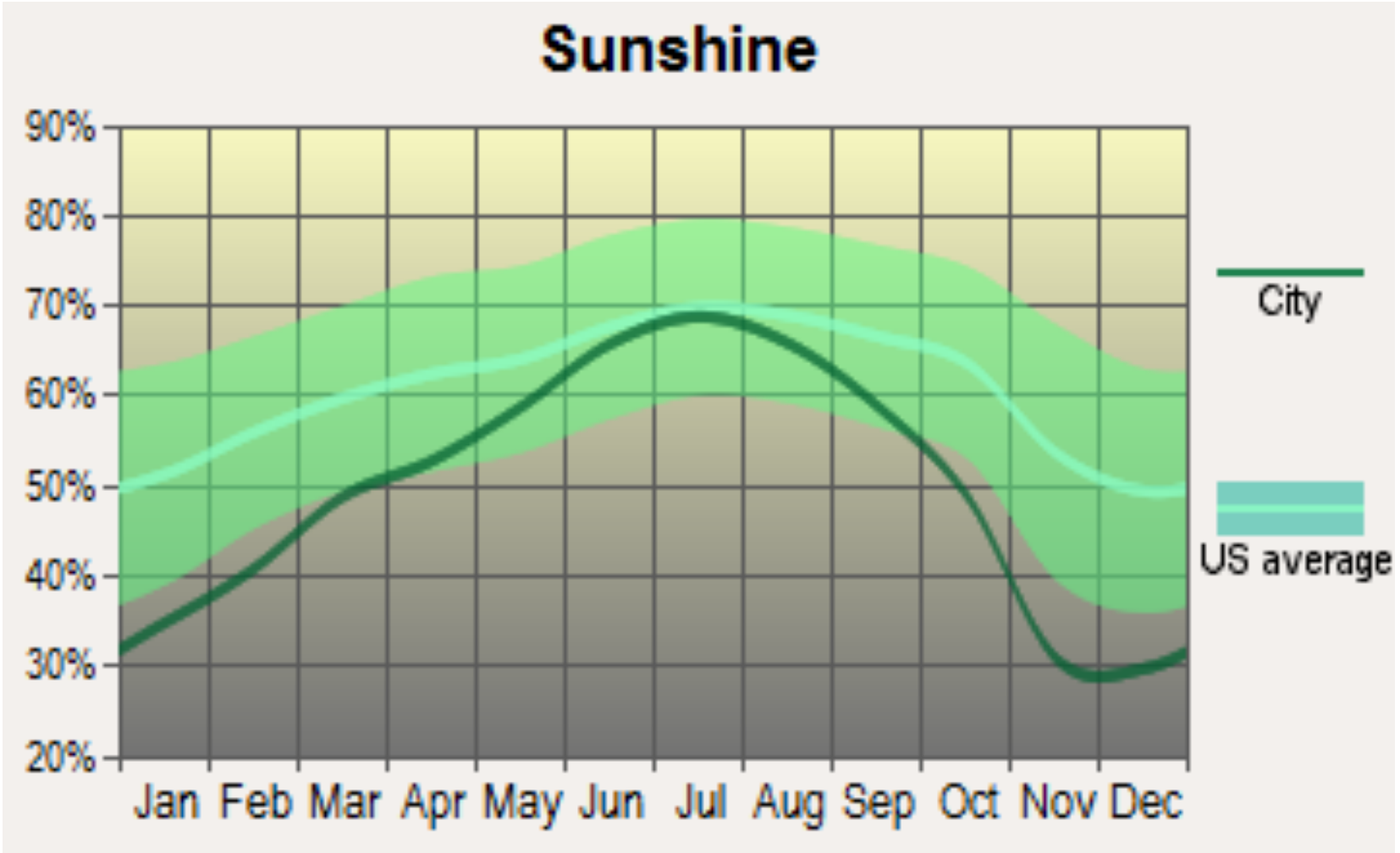
Source: http://www.metoffice.gov.uk/public/weather/climate/gcpvj0dup

Climate: London versus Tokyo



Source: www2m.biglobe.ne.jp

Comparing New York and US

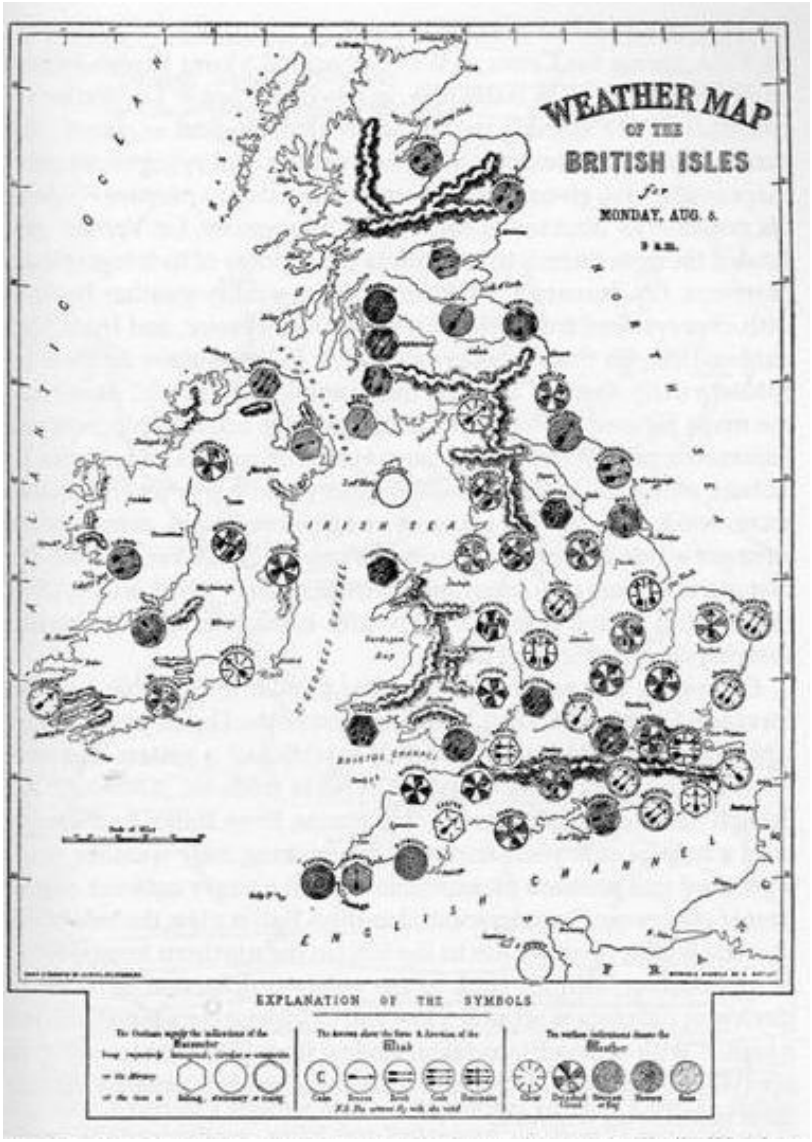


Source: www.city-data.com

Weather charts

- Two Royal Navy officers, Francis Beaufort and Robert FitzRoy, are widely acknowledged as transforming weather forecasting into a science.
- Their work led to the development of reliable tide tables and the collection of weather data at sea by ship captains as a service to mariners.
- It was the loss of the Royal Charter in 1859 due to a storm that inspired FitzRoy to develop charts for forecasting the weather.

Fitzroy daily weather map (1861)



Robert Fitzroy 1805-1865

Quiz

- When was numerical weather prediction first proposed?
  1. 1880s
  2. 1900s
  3. 1920s
  4. 1940s

Slido.com **#71725**

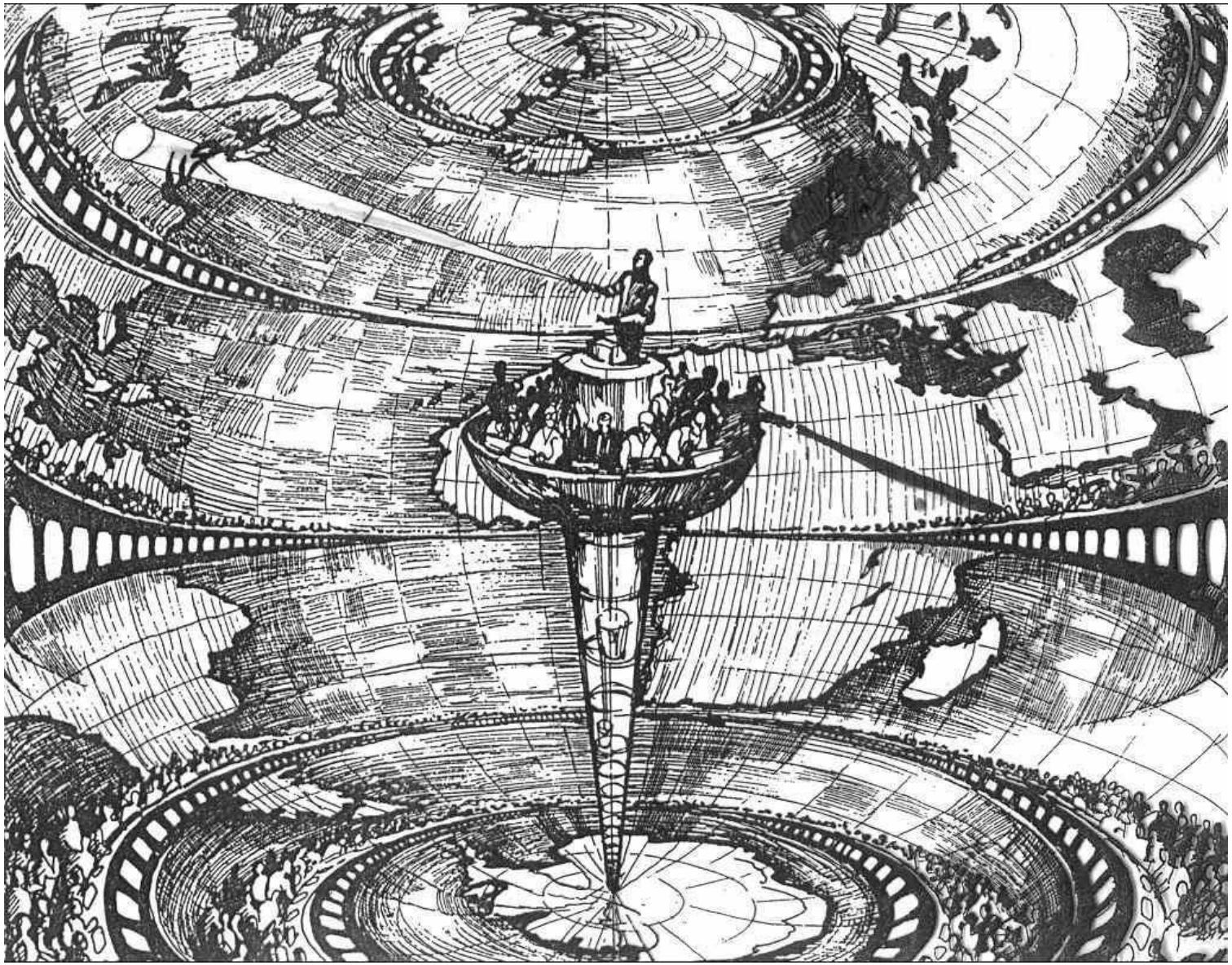
Twentieth century

- By the twentieth century, scientists had gained a deeper understanding of the atmosphere and were in a position to propose a mathematical model for describing how it changes over time.
- In 1922, the English scientist, Lewis Fry Richardson, published "Weather Prediction By Numerical Process" based on notes he had written while an ambulance driver during World War I.

Richardson's auditorium

- As Richardson's approach to numerical weather prediction came before computers were invented, he envisioned a large auditorium of thousands of people performing the calculations and passing them to others.
- It is not such a leap of the imagination to view this as the forerunner of modern data science.

Richardson’s auditorium



Richardson’s auditorium for weather forecasting

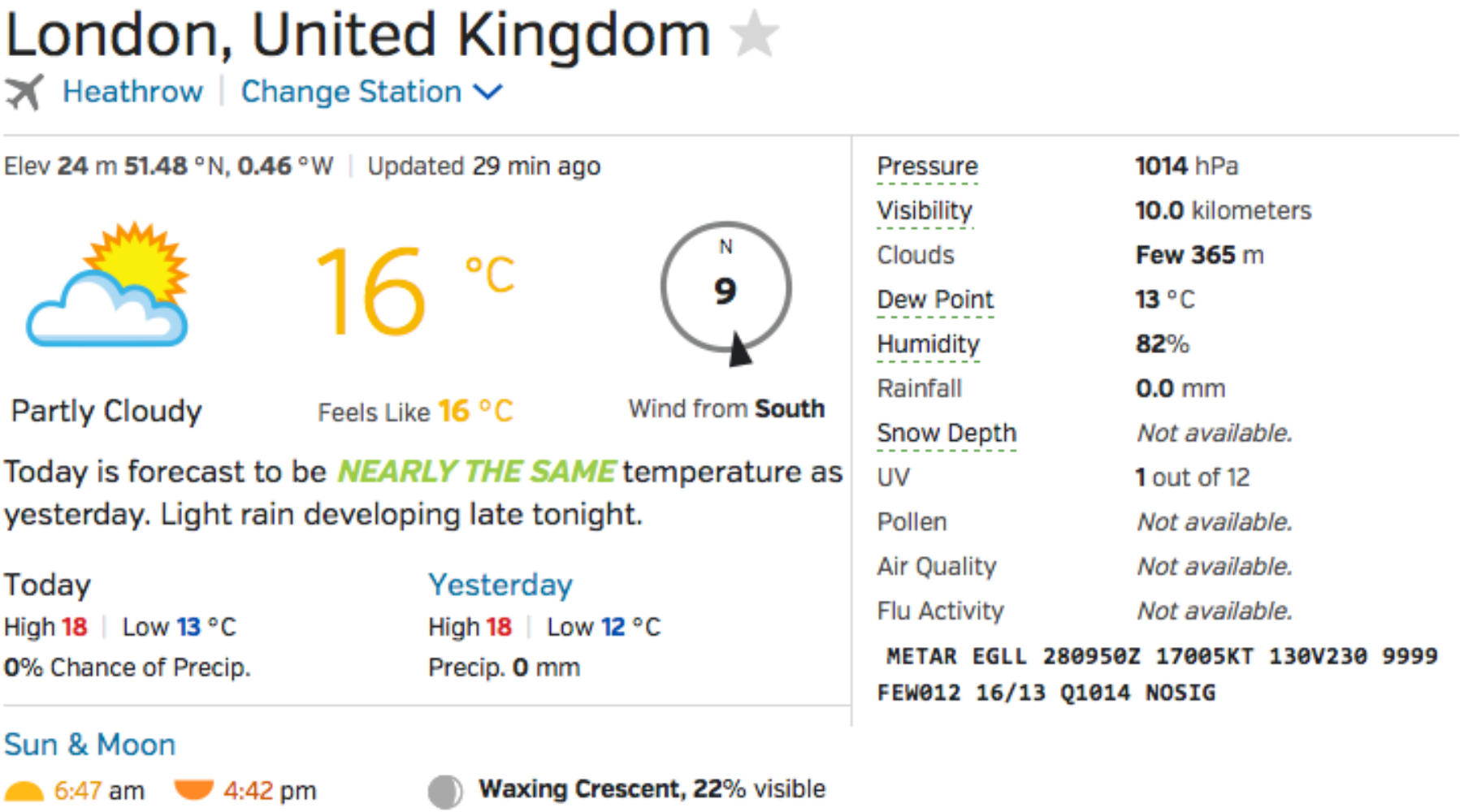
Weather Poll

- Which weather variable are you most interested in?

Select a variable name (rainfall, temperature, humidity, wind, sunshine, snow, pressure, …)

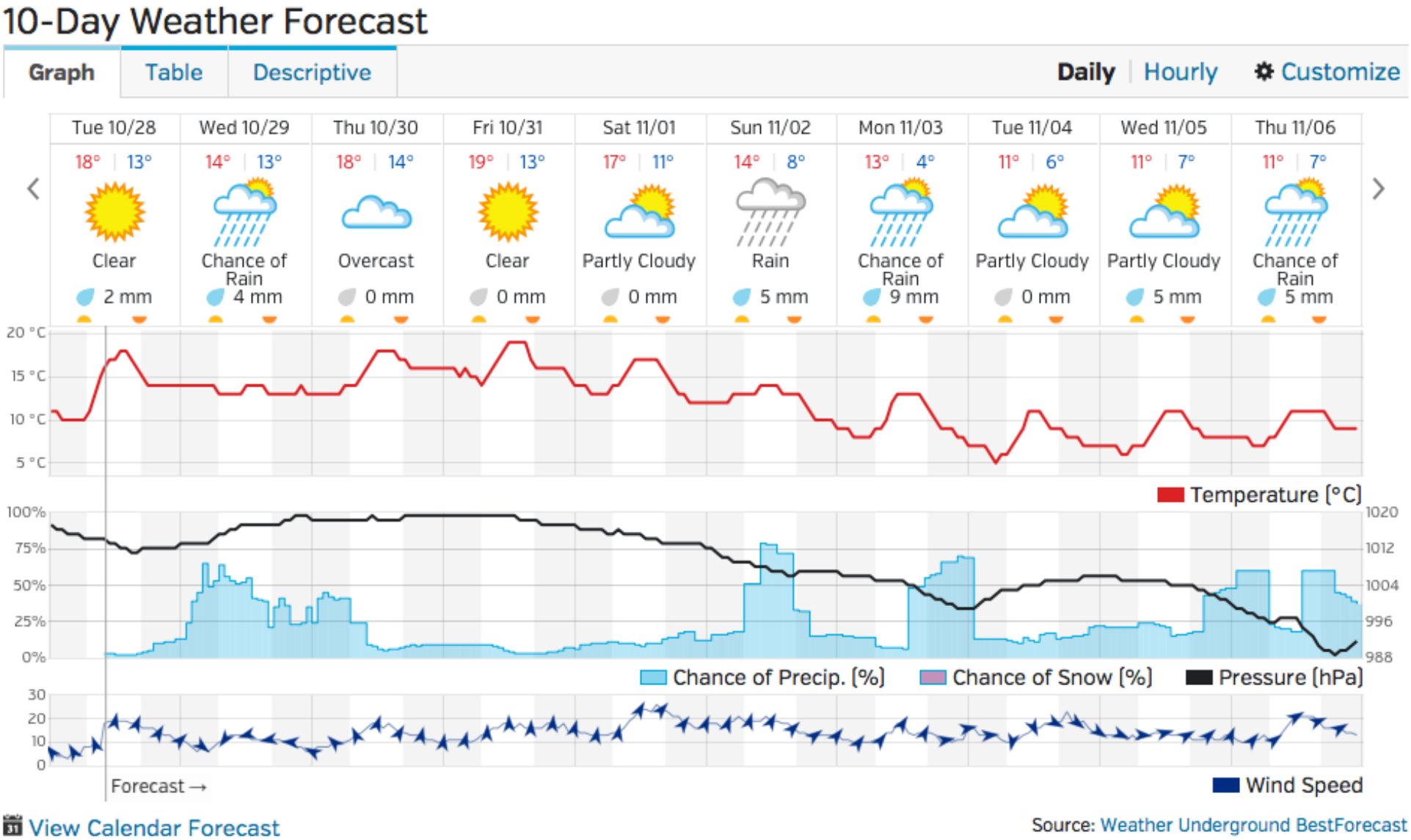
Slido.com **#71725**

Weather report



Source: www.wunderground.com

Weather Forecast



Weather forecast errors

- Weather forecasts certainly have their limitations.
- It is easier to produce forecasts than it is to undertake a serious scientific evaluation of their quality.
- Nobody assumes that weather forecasts will always be correct, but exactly how much error should we tolerate and how often should we expect to be let down?

Weather forecast accuracy Poll

- How many days ahead would you trust a weather forecast?

1. 1 day
2. 5 days
3. 10 days
4. 20 days

Slido.com **#71725**

Evaluating quality

- Forecasting the amount of rain falling in the Sahara desert is a relatively easy challenge and this goes someway to explain that forecast performance should be measured relative to the natural variability of the weather data of interest.
- Hence, we need to employ appropriate benchmarks to measure the quality of the weather forecast service provider

Remembering forecast failure

- The models utilized for numerical weather prediction suffer from insufficient observations across the globe and a lack of computational resources for propagating this initial information to future forecast horizons.
- Our view of how well a weather forecast service is doing is typically based on our memory of events associated with failure.

Focus on extremes

- If we use weather forecasts for making decisions and our actions lead to financial gains or penalties, then we can perform a cost-benefit analysis of a given forecasting service.
- For this reason, we often focus on extreme events as these are the days when public attention is at its peak.

Fish and the hurricane

- Michael Fish was a household name in the UK, a popular weather presenter for the BBC.
- Unfortunately, his words of reassurance on the 15th October 1987 are associated with one of the greatest forecast errors in the UK.
- He started by saying "Earlier on today a woman rang the BBC to say she'd heard there was a hurricane on the way," and then continued with "Well, if you're watching, don't worry there isn't."

The Great Storm

- Although not technically a hurricane, what followed was the worst storm to hit southeast England since 1703.
- This particular forecast error made the weather forecaster infamous as the man who failed to spot what became known as the Great Storm.

The Weather News

Source: www.youtube.com/watch?v=uqs1YXfdtGE

The aftermath

Devastation near Sevenoaks, Kent in the UK
Source: http://www.mirror.co.uk/news/uk-news/uk-weather-michael-fish-would-1377118

Michael Fish's response

- I used to say that if I had a pound for every time someone mentioned that broadcast to me, I'd be a millionaire.
- Since they showed the clip at the Olympic opening ceremony – watched by four billion people – I now say I'd be a multi-billionaire.
- I'd like every single copy of that tape buried with me when I die – but even then it will probably say on my tombstone: "Here lies Hurricane Fish."
- The British are obsessed with the weather and obsessed with failure. I suppose I unite the two.

Chaos

- It is of course easy to throw tomatoes at the scientists responsible and blame them for their inadequate modeling.
- What should we realistically expect from the models?
- Chaos theory plays an important role here as the atmosphere behaves like a chaotic system in that small initial errors tend to be amplified as the forecast horizon is increased.

Lorenz's butterfly effect

- Ed Lorenz, the physicist who discovered chaos in weather models, famously speculated that a butterfly flapping its wings in Brazil could eventually cause a hurricane in the US.
- So should we give up on forecasting the weather because of chaos?
- Definitely not, uncertainty is something that all data scientists have to live with but the methods used for accounting for uncertainty vary dramatically.

Ensemble forecasts

- Dealing with uncertainty in initial conditions, the starting values for numerical weather prediction, is relatively straight-forward.
- Multiple weather services providers such as the European Centre for Medium-Range Weather Forecasting (ECMWF) and the US National Centers for Environmental Prediction (NCEP) both use ensemble weather forecasts to manage this uncertainty.

Ensemble forecasts

- An ensemble weather forecast is simply obtained by running the model multiple times to obtain a collection of different scenarios, each origination with a different initial condition.
- Now the whole point of generating ensemble forecasts is to provide an improved forecasting approach and a better way of communicating to end-users.

Evaluation Poll

- In order to select the best forecast provider for weather forecasts, you need to decide on which of the following:

1. Weather variable
2. Forecast horizon
3. Performance metric (R^2, RMSE, MAE, …)
4. Back-test period
5. All of the above

Slido.com **#71725**

Communication

- Unfortunately, clear communication has not always been the result of using ensemble forecasts.
- The UK Met Office have been haunted by their use of seasonal weather forecasts.
- By definition, seasonal forecasting refers to horizons of up to six months ahead and this is pushing the boundaries of what is possible given current technical capability and computational resources.

Media sound bites

- Seasonal forecasts used an ensemble approach to evaluate whether futures temperatures were likely to be below, equal to or above typical values.
- The Met Office ran into difficulties in the 2009 when it forecasted a “barbecue summer” which did not materialize.
- Later in January 2011 its “mild winter” ended up being the coldest for 31 years.

The BBC

- As a result of the controversy around these seasonal forecasts, the BBC launched an attack on the UK Met Office, with a BBC Newsnight investigation into their failure to forecast the extremely cold conditions.
- It transpired that the model stated that there was a 50% probability of a “mild winter” and this was the most likely outcome of the three possible outcomes.

Impact

- The media sound bite from the Met Office presented this most likely outcome as the forecast and failed to discuss the other possibilities.
- In retrospect it seemed unfair that the Met Office got such bad publicity and it was arguably the communication that was at fault rather than the model.
- Later that year, however, the Met Office scrapped its seasonal forecasts after carrying out extensive customer research.

Lessons learned

- There is an important lesson here for all those hoping to reap the benefits of data analytics.
- Analytics by their very nature always depend on a mathematical model and this model will often be wrong due to observational uncertainty, difficulties in estimating parameters and failure to identify the correct model structure.
- For this reason the machine delivering the analytics is uncertain and the method used to communicate this uncertainty is crucial for building confidence.

Honest forecasts

- If forecasters are honest, they should express their forecasts using an expected value and a confidence interval.
- Forecasters with no confidence intervals are claiming perfect precision and this certainly does not imply that their forecasts are more accurate, but more likely that they are ignorant about potential forecast errors.

McSharry (2010). Revealing the real forecast. The Guardian, UK.

Confidence

- Unfortunately, human nature is such that the overconfident forecasters are sometimes viewed as having a better approach or simply knowing better than their competitors.
- Deploying statistical techniques to evaluate historical forecasts is the only way to figure out which forecaster is best.

Q&A