Working as a Data Scientist in Sport

Professional Statisticians' Forum
21 November 2016
Dave Hastie
Director, Sporting Data Science

Introduction

Convincing you to invest the next hour of your life

- Statistical background
- Business background
- Talk outline

Statistical background

- MSci. Mathematics - University of Bristol - 1996 - 2000
  - 4 year undergraduate degree
  - Large focus on statistics and computing modules in the later years
- Ph.D. Statistics - University of Bristol - 2001 - 2004
  - Reversible jump Markov chain Monte Carlo
  - Methodological, statistical software development
  - No sporting applications
- Postdoctoral researcher - Imperial College London - 2010 - 2012
  - Computationally intensive non-parametric modelling
  - Epidemiology and biostatistics applications

Business background

- Smartodds - 2003 - 2010
  - Head of Quantitative Team
- Onside Analysis - 2011 - 2015
  - Co-Founder, CTO
- Stratagem Technologies - 2014 - 2015
  - Director, CTO
- CricViz - 2015 - present
  - Director (CTO)
- Sporting Data Science - 2015 - present
  - Founder, CEO

Talk outline

- Career highlights by case study
  - Smartodds
  - Onside Analysis
  - Stratagem Technologies
  - CricViz
- Common themes and best practices
- Common challenges
  - Empirical models
  - Model assessment
- My perspective on the future
  - Sporting Data Science

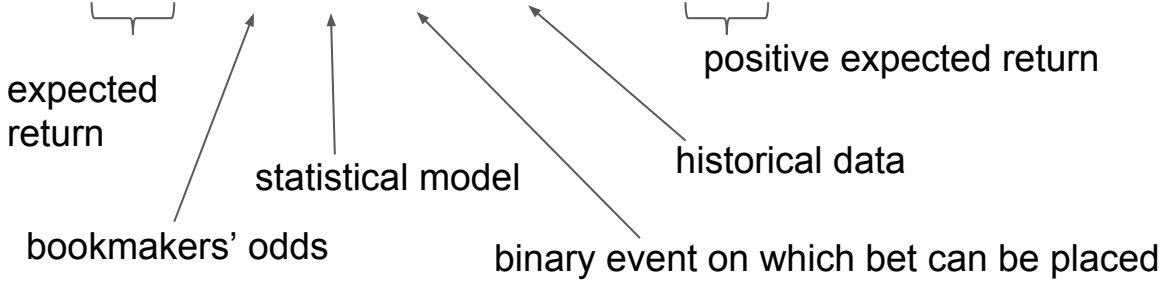
Career Highlights by Case Study

What (I like to tell people) I do

- Smartodds
- Onside Analysis
- Stratagem Technologies
- CricViz

Smartodds (2003 - 2010)

- What the company did [does]
  - Analysis (statistical modelling) of football [sport] using historical data to identify opportunities to place bets with positive expected return
  - $E[R] = X \; p(G \mid data) - 1 > 0$



Smartodds (cont.)

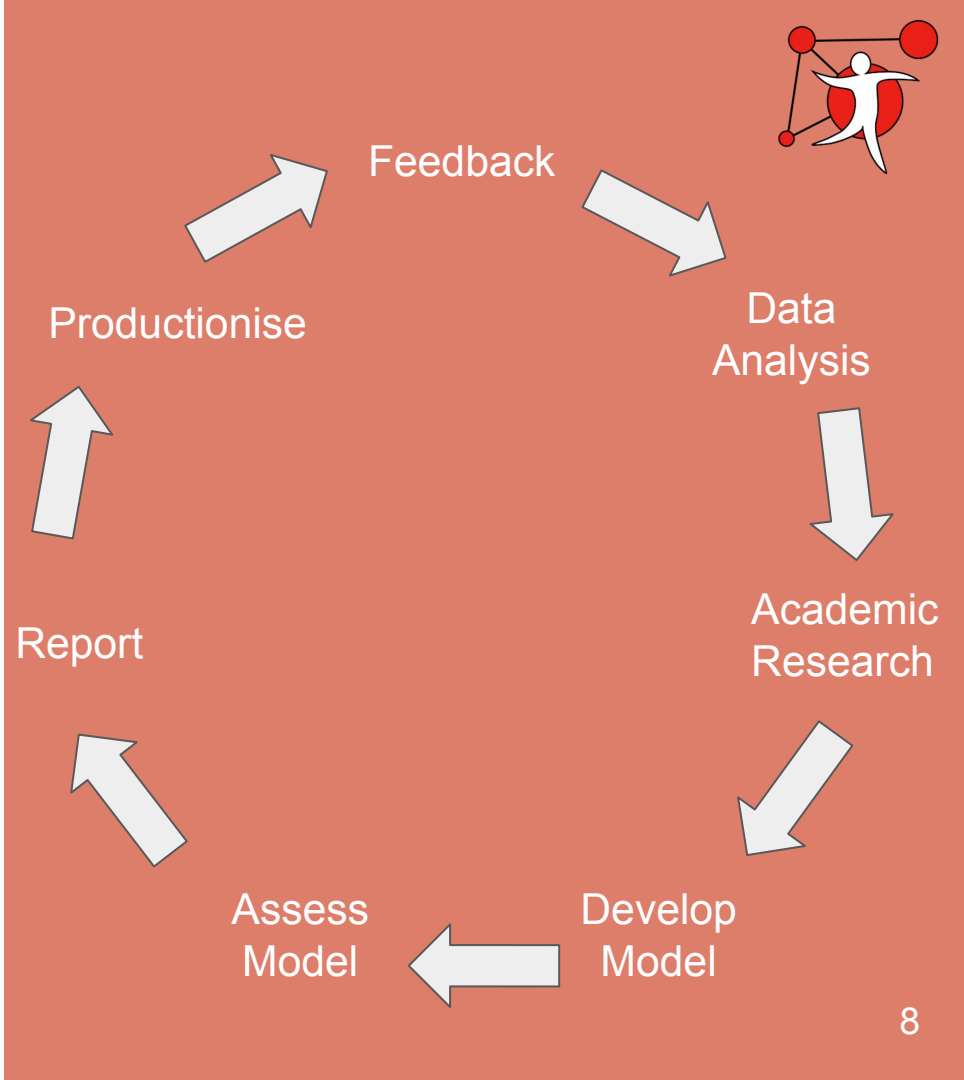
How was data science used?

Data processing

- Multiple sources, often unclean data
- Building pipelines

Model Development Cycle

- Key to reduce time around cycle



Smartodds (cont.)

Example model

$$\log \lambda_H = \mu + \gamma + \alpha_j + \beta_k$$
$$\log \lambda_A = \mu + \alpha_k + \beta_j$$
$$X \sim \text{Poisson}(\lambda_H)$$
$$Y \sim \text{Poisson}(\lambda_H)$$

Considerations:

- Temporal dependence
- Additional count data - shots, corners, chances
- Player data
- Environmental data - pitch, weather, distance

Onside Analysis (2011 - 2015)

- What the company did:
  - Analysis (including statistical modelling) of historical football data to inform and improve future decisions for multiple stakeholders in the football and connected industries.
    - Player assessment
      - Recruitment
      - Salary negotiation
    - Managerial assessment
      - Recruitment
    - Technical analysis
      - Coaching
    - Predictive modelling
      - Betting

Onside Analysis (cont.)

How was data science used?

- Data collection
- Data processing (multiple sources, often unclean)
  - Data pipelines
- Statistical modelling of sporting event data
  - Model development cycle
- Visualisations and reports
- Data delivery
  - API
  - Web dashboards

Onside Analysis (cont.)

Example - Evaluating the contribution of players

- Opta data - every touch of the ball is time stamped and coded
- Assessing contribution of player
  - Contribution by scoring goals
  - Contribution by helping team to score goals
  - Contribution by preventing opposing team from scoring goals
- Credibility challenges
  - Disconnect between analysts / C-Level and coaching / playing staff
  - Opta data is an imperfect data source
  - A single implausible model output can be enough for model to be dismissed

Onside Analysis (cont.)

Extract of report

the average leading minutes proportion.[1]. In the interests of protecting intellectual property, we omit the full details of this measure, but make the following observations about features of it's designs.

- Goals, times, line-ups and substitutions should be taken into account
- The measure should use the data to draw out increased signal about performance than is possible by simply considering the scoreline
- A player should be rewarded for obtaining positive player supremacy
- The greater the player supremacy, the greater the reward, but not all goals are equal
- Players should be rewarded more for an earlier increase in their player supremacy rather than a later change
- A player making an average contribution to a team should have a score of 0
- A player making a below average contribution should have a negative score
- A player making an above average contribution should have a positive score
- Scores should be comparable across team mates. Comparisons between players in separate teams should be possible, but with extreme caution recognising the limitations
- The measure should be in a form that can be generalised with access to additional high resolution player-action data
- The measure should be straightforward to test and refine as part of a statistical modelling process

The following sections provide some visualisations of OAPI.

Onside Analysis (cont.)

Extract of report

Premier League 2012/2013

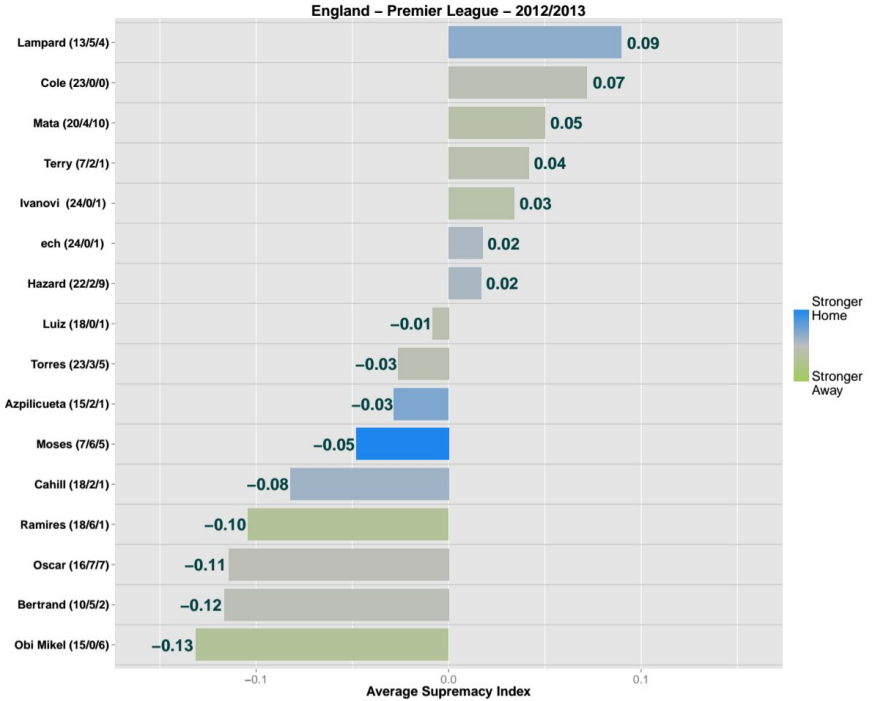


Figure 7: *Player index based on average time and supremacy whilst on pitch for 2012/2013 Premier League*

Stratagem Technologies (2014 - 2015)

- What the company did [does]:

"generating consistent edge in sports prediction trading - a highly liquid asset class with uncorrelated returns consistent edge in sports prediction trading - a highly liquid asset class with uncorrelated returns"

  - Similarities to Smartodds
  - Taking an approach more inspired by financial trading
    - Portfolio building
    - Trading strategies
    - Risk management
  - Trading platform
  - Proprietary trading
  - Managed fund

Stratagem Technologies (cont.)

How was data science used?

- Data pipelines (collection and processing unclean data)
- Statistical modelling of sporting event data
  - Model development cycle
- Statistical modelling of odds time series
  - Time series methods
  - Reinforcement learning
- Portfolio optimisation
- Visualisations and reports
- Image processing
  - Automated data collection

Stratagem Technologies (cont.)

Example application

- Image processing
  - Started out as a Masters student project in partnership with Imperial College London
  - Manual data collection is expensive, can we automate it using image processing
  - Difference categories of “shot” - can we identify them
  - Became a full-time research area

Stratagem Technologies (cont.)

Example application

CricViz
(2015 - present)

- What the company does:

"CricViz allows the user to understand cricket in greater detail than ever before. The unique CricViz computer model enables the prediction of match outcome, the interpretation of team and player performance and the anticipation of what is likely to happen next; cricket intelligence at the next level "

  - Very much interested in the modelling and understanding of the sport (cricket)
  - Providing services to cricket fans and enthusiasts
    - Building a data driven community
  - Providing data and analytics services
    - To broadcasters
    - To the professional cricket teams
    - For the betting industry

CricViz (cont.)

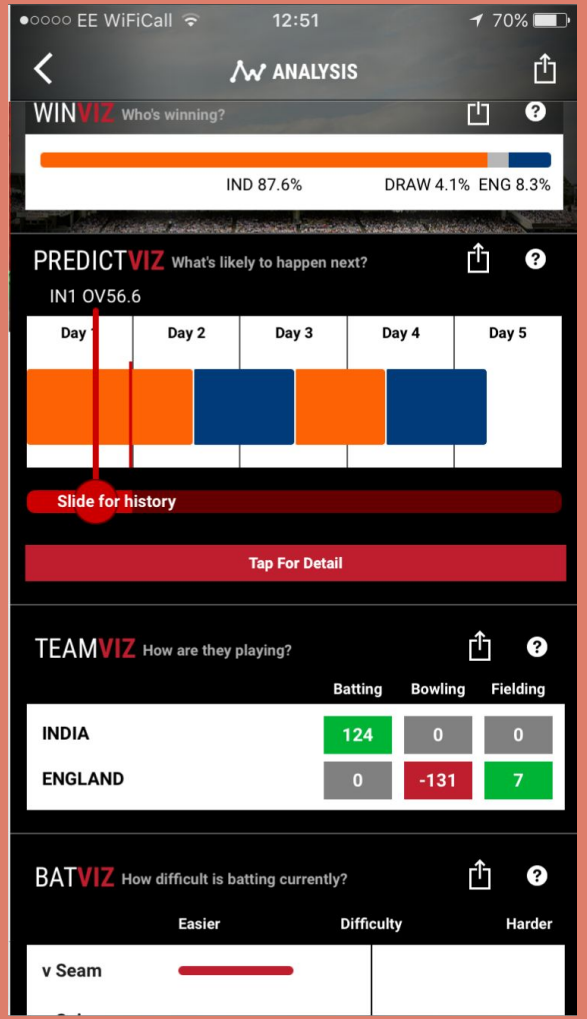
How is data science used?

- Data pipelines (collection and processing unclean data)
  - A large part of business model is built around the company aggregating and utilising multiple data sources that have never before been combined
  - Creating new derived data tailored to customer requirements
- Statistical modelling of sporting event data
  - Modelling outcomes of cricket games
  - Modelling player performance
- Visualisations and reports
- Data dissemination
  - APIs

CricViz (cont.)

Example - mobile app

- Probabilities of match outcome
- Expected length of innings
- Player performance
- Pitch difficulty based on ball tracking
- Demonstrates robust need of data science



CricViz (cont.)

Example - broadcast

Common Themes, Best Practices and Challenges

What I try to do and what makes my job interesting!

- Robust approach
- Pragmatism
- Domain credibility
- Data matching
- Incomplete data
- Model assessment
- Empirical modelling

Theme - Robust practices

Important to adopt robust approaches and utilise best practice

- Use databases
- Invest and take ownership of data processing
  - 50% of work
  - Rubbish in, rubbish out
- Model development framework
  - Repeatable and reproducible
  - Easy and quick to test hypothesis, document and iterate
- Record failed developments as well as successes
- Best software development practices

Theme - Pragmatism

Important to adopt an approach of pragmatic rigour

- Don’t discount expert knowledge just because it isn’t statistically robust
- Recognise the limitations and assumptions behind a lot of statistical theory
  - Assumptions may not hold in practice
  - Don’t discount something just because there is not a mathematical proof
- Be practically rigorous
  - Remember the final goal

Challenge - Domain credibility

- Historically athletes and coaches often typically not highly educated in an academic sense
  - Suspicious of what data can tell them compared to their own expertise
  - Not unique to pro-sports (e.g. betting, or music)
- One shot - unplausible output is punitive
- In professional sports, sample sizes can be very small
  - Moneyball has helped (and hindered)

Must adopt a professional attitude and obtain gradual buy in

Challenge - Data matching

- Multiple data sources with different identifiers for the same objects
- Linking data sources in real-time has historically been a significant challenge
  - Lots of bespoke algorithms
  - Frustrated by inconsistencies within source
  - Manual matching
- Opportunity for automated machine learning approaches

Challenge - Incomplete data

- Modelling sporting events is often complicated by incomplete data
  - Player / ball positions
    - Defenders non-actions
  - Subjective data
    - Shot quality
- Have to adapt parametric models to take account of these limitations
  - Introduce biases

Challenge
Model assessment

It is often hard to decide how to assess if one model is better than another

- Require a robust framework
  - Consistent data set for assessment
  - Reproducible research
  - Part of a wider model development pipeline
  - Document the failures as well as the successes
- How do we measure success?
  - Often lots of competing candidate measures
  - Can we isolate model from the ultimate outcome (e.g. betting)

Challenge - Model assessment (cont.)

Example

- Consider a model for predicting the outcome of a tennis match in-play
  - Simple version is a Markovian based model
    - Can be extended to be non-homogeneous / simulation based
  - Updates after every point
- How do we assess how good the model is
  - At the start?
  - At the end of each set?
  - At the end of each game?
  - Some combination?

Challenge - Model assessment (cont.)

Example (in-play tennis cont.)

- Suppose we can agree on a time frame - what do we measure?
  - RMSPE, not bias - but where is the tradeoff
  - Observed minus expected
    - win probabilities ?
    - set length predictions?
  - What about ultimate outcome?
    - e.g. betting: profit - how do we divorce from strategy?
    - e.g. pro: player rankings - what about outside influences?

Challenge - Model assessment (cont.)

- Rarely a single model that is best
- Must avoid sensitivity to data set selection
- Multiple criteria can lead to multiple models, depending on priority
  - Models can be inconsistent
- We have to be pragmatic
  - Academic theory is hugely useful and where possible should guide us, but often things are messier in practice
  - What is the most important thing that we are trying to measure?
  - Do inconsistent models matter?

Challenge
Empirical models

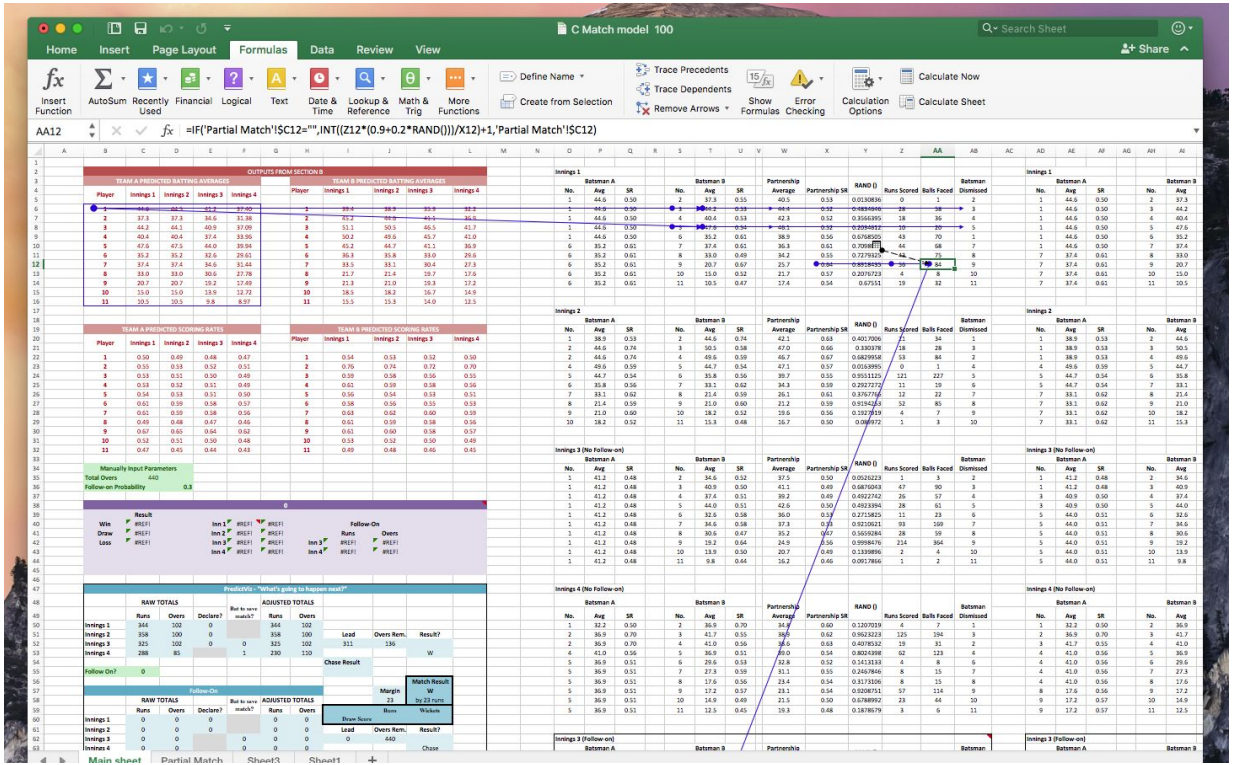
Combining academic statistical modelling theory and domain expertise can be difficulty

- Domain expertise is often built up through experience
  - Experts can find it hard to encode
- Where expertise is “encoded” it is often in the form of an empirical model, typically in a spreadsheet
  - Expert inputs
  - Black box constants
  - No “history” of development
  - Very hard to scale and assess
- BUT often very performant and contain valuable insight

Challenge - Empirical models (cont.)

Example - Empirical to Production

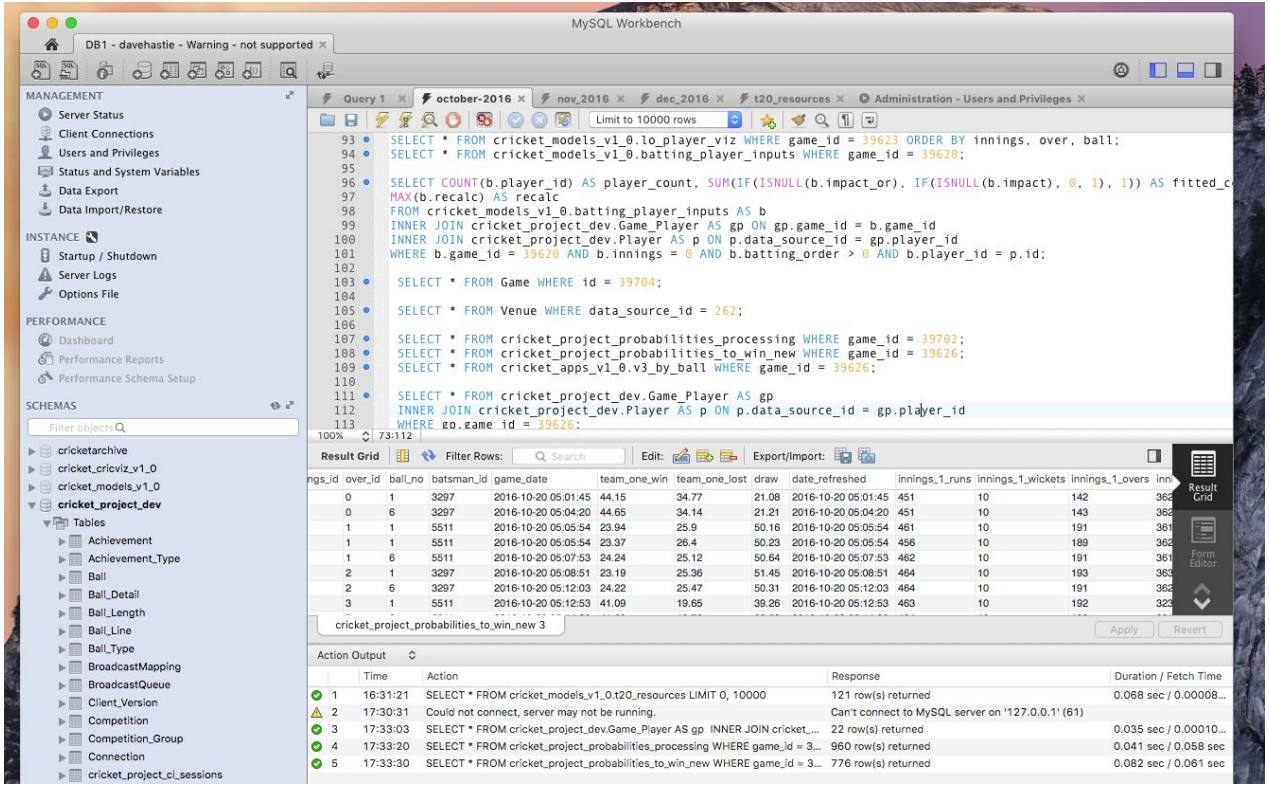
- The “expertise”



Challenge - Empirical models (cont.)

Example - Empirical to Production

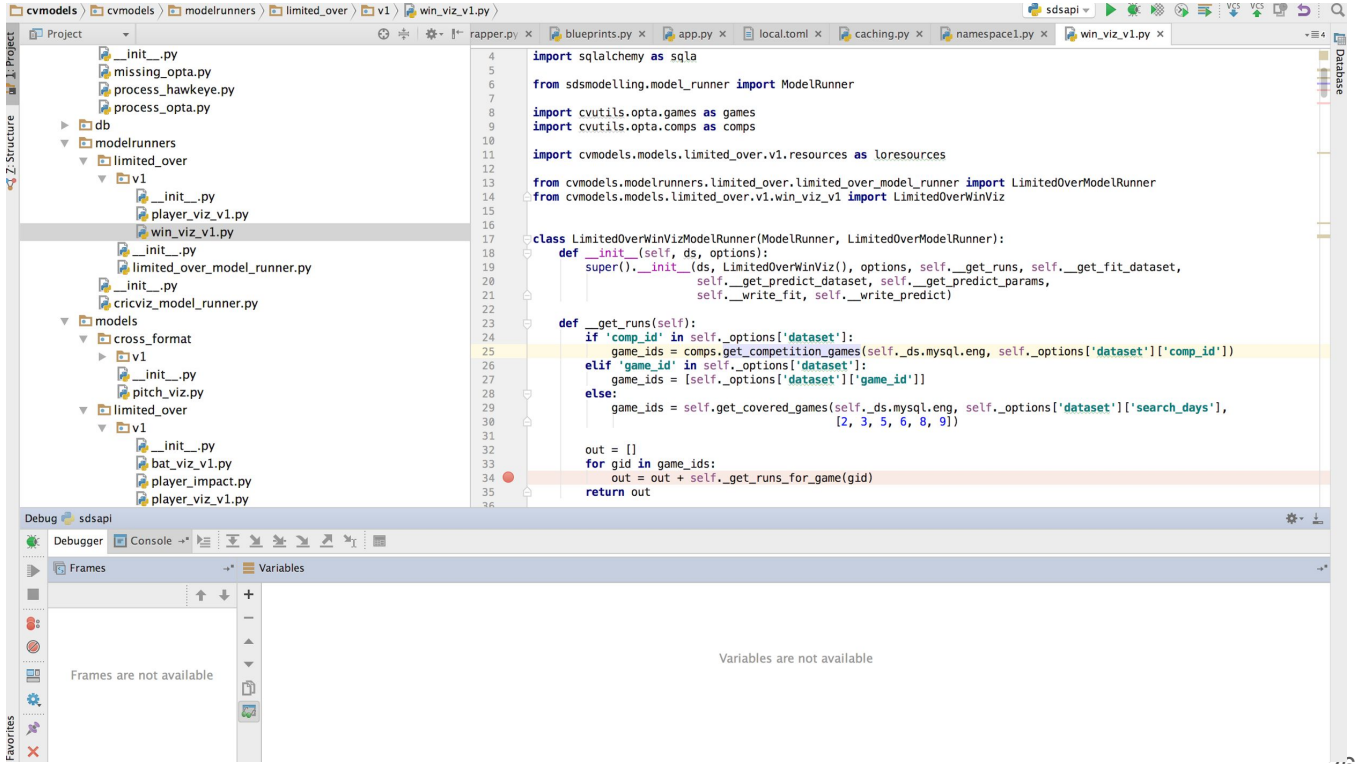
- Robust data
  - Parsing
  - Processing
  - Wrangling



Challenge - Empirical models (cont.)

Example - Empirical to Production

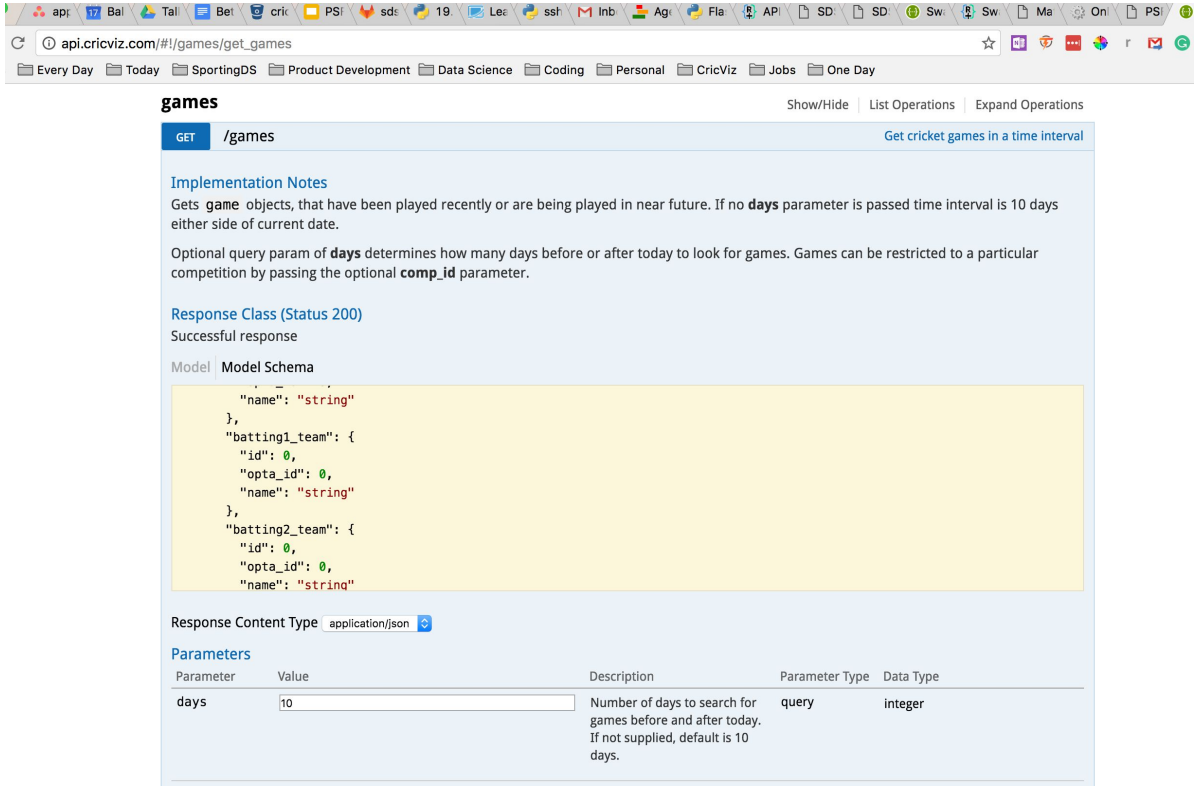
- Model development framework



Challenge - Empirical models (cont.)

Example - Empirical to Production

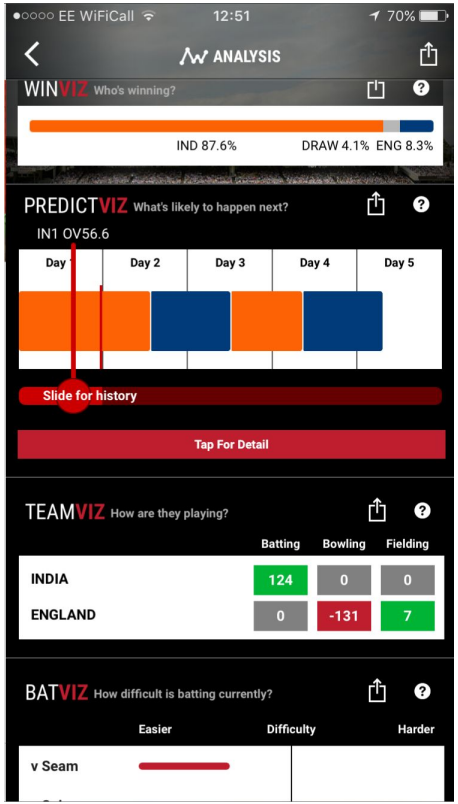
- An external interface



Challenge - Empirical models (cont.)

Example - Empirical to Production

- The customer view



Challenge - Empirical models (cont.)

Assessing empirical models is even more complicated

- All the usual challenges
- How do you account for the manually inputted domain expertise?
  - Cannot be populated post-hoc
  - Hard to build up a sample that is sufficiently large to robustly assess

Approach similarly to credibility challenges

The future

What we can look forward to

- Sporting Data Science
- Prevalence of data
- Predictive modelling
- Breadthening of scope

Sporting Data Science

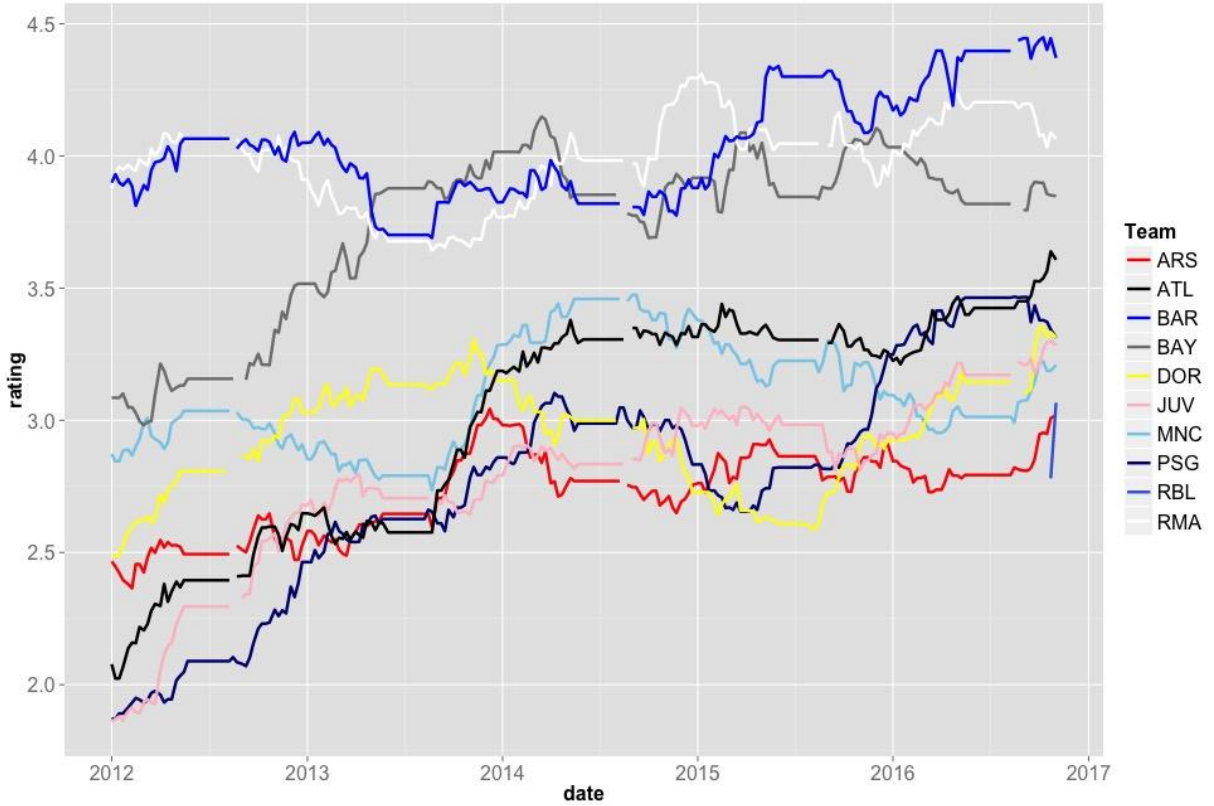
My latest project building on career experience to date

- Consulting to a number of companies
- Traditional predictive modelling for sporting events
- Helping clients with data assets generate a revenue stream from those assets
- Supporting general trend towards data engagement

Sporting Data Science

Example

Team ratings for fan engagement and brands



Prevalence of data

Data is now ubiquitous in sport, as throughout society

- Previously applications were limited by availability of data
- In sport, data has not traditionally been “big data”
  - This has changed with high frequency odds
  - Tracking (ball, player) data
- Big data sets lend themselves to different methods of analysis
- General increase in societal data literacy

Predictive modelling in sport

Role of predictive modelling is changing

- Typically parametric models have been used
  - Interpretable (critically important when data contains a lot of noise)
- Marginal gains in a lot of areas are much diminished but useful in other domains
  - e.g. Betting vs Broadcast
- Machine learning / artificial intelligence approaches can deal with correlations and data dimension better
  - Computer scientist vs statisticians
  - Challenge is to retain interpretability

Breadthening of scope

Changing nature of data means different applications

- Player and ball tracking data
  - Visualisations
  - Coaching applications
- InCrowd Sports
  - Fan engagement platform
  - Data regarding persons movements around match events
- Sportr
  - Natural language processing - summarising multiple news sources

Any questions?

dave.hastie@sportingds.com

- RSS Statistics in Sport section
- Thank you
  - Dixon and Coles, 1997
  - Smartodds
  - Stratagem Technologies
  - CricViz