ACEMS/QAS/AIS Workshop
Stats and Sport

Methods for Big Data

Kerrie Mengersen

Overview

1. Bayesian Methods
2. Spatio-temporal Methods
3. Linking diverse data: Performance and Health
4. Wearables

Using data to update understanding or predictions

I have a prior estimate, and obtain some data, then update my estimate given the data.

Prior + Data > Posterior

Using data to update understanding or predictions

I have a prior estimate, and obtain some data, then update my estimate given the data.

Prior + Data > Posterior

I have a model with some parameters, and ask “what else do I know about the model and parameters”, and include this info in the parameter estimates.

Model + Priors > Posterior

Using data to update understanding or predictions

I have a prior estimate, and obtain some data, then update my estimate given the data.

Prior > Data > Posterior

I have a model with some parameters, and ask “what else do I know about the model and parameters”, and include this info in the parameter estimates.

Model > Priors > Posterior

"In the past ten years, it's hard to find anything that doesn't advocate a Bayesian approach." -Nate Silver

Example 1

Suppose you believe a team has a 50% chance to win.

You receive information that it's going to rain for their next game. What is the chance of them winning?

Suppose you have the following information:

- the chance that it rains when this team *plays* is 10%
- the chance that it rains when this team *wins* is 15%

Can we say that the chance the team wins given that it rains is 50%?

10%

15%?

Example 1 (cont.)

A: Team wins:
Pr(Win) = 0.5

B: Rains when this team plays
Prob(Rain) = 0.10

B|A: Rains when this team wins:
Prob(Rain|Win) = 0.15.

We want: Prob(Win given Rain)
= Prob(A|B)

https://www.pinnacle.com/en/betting-articles/Betting-Strategy/bayesian-analysis-and-sports-betting/5VM2U95MX696RGXP
http://www.sports-management-degrees.com/baseball/

Example 1 (cont.)

A: Prob(Win) = 0.5

B: Prob(Rain) = 0.1

B|A: Prob(Rain given Win) = 0.15.

Prob(Win given it rains) = Prob(A|B)

Use Bayes' Theorem

$$P(A|B) = P(A) * P(B|A) / P(B)$$

$$= 50\% * 15\% / 10\%$$

$$= 75\%.$$

The chance the team wins given that it rains is 75%.

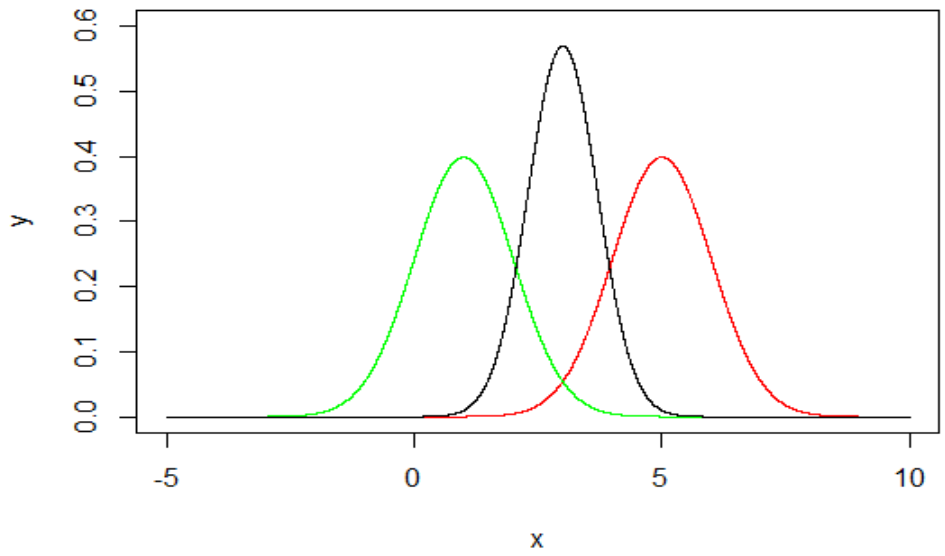
https://www.pinnacle.com/en/betting-articles/Betting-Strategy/bayesian-analysis-and-sports-betting/5VM2U95MX696RGXP
http://www.sports-management-degrees.com/baseball/

Example 2

What is the expected change in haemoglobin mass (HM) after training regime?

Prior > Data > Posterior

$\mu = 5 \pm 1$ > $y|\mu = 1 \pm 1$ > $\mu|y = 3 \pm 0.7$



Bayesian Models for Big Data

"raw data, no matter how extensive, are useless without a model" – Nate Silver

What is the expected no. goals scored by a team?

Model: $y \sim \text{Poisson}(\mu)$

$\log(\mu) = home + att + def$

Priors: *what do we know about home, att, def?*

is there structure in the data that we can model?

“Big Bayesian Learning”

Models:

- Probability
- Uncertainty
- Regularisation
- Flexibility
- Adaptivity

Algorithms:

- Scalability (parallelisation, distributed computing)
- Subsampling
- Approximations (eg. ABC)

Zhu J, Chen J, Hu W, Zhang B (2017) Big Learning with Bayesian Methods
arXiv: 1411.6370v2 https://arxiv.org/pdf/1411.6370.pdf

Big Data in Sports Science

- Sports performance data (wearables)
- Spatio-temporal data
- Personalised health

Wearables

- Extremely rapid growth in the wearable technology industry
  - 2015: Wearable sales: $14bn
  - 2017: $9.3bn revenue from smartwatches alone
  - 2019: estimated 148 million units shipped (compound annual growth of 35%)
  - 2020: expected sales $34.2bn
  - 2021: total shipments for all wearable devices will grow to 560m, revenue $95.3 bn
  - 2023: market for wearable technology over $100bn
  - 2026: market over $150bn

- Increasing range of wearables:

  Tommy Hilfiger – solar powered jackets as portable chargers

  Future Interfaces Group – skin as touchpad (SkinTrack)

https://delta2020.com/blog/132-analysis-of-wearable-technology-the-future-of-sports

https://www.smartinsights.com/digital-marketing-strategy/wearables-statistics-2017/

Wearable technology in sport

Catapult Sport: OptimEye S5

acceleration, direction, position, impact of collisions

Steph Curry: Halo Sport Headphone

stimulates brain to increase effectiveness in training session.

https://delta2020.com/blog/132-analysis-of-wearable-technology-the-future-of-sports

From Big Data to Little Data

Mengersen KL, Drovandi CC, Robert CP, Pyne DB, Gore CJ (2016) Bayesian Estimation of Small Effects in Exercise and Sports Science.

PLoS ONE 11(4): e0147311.

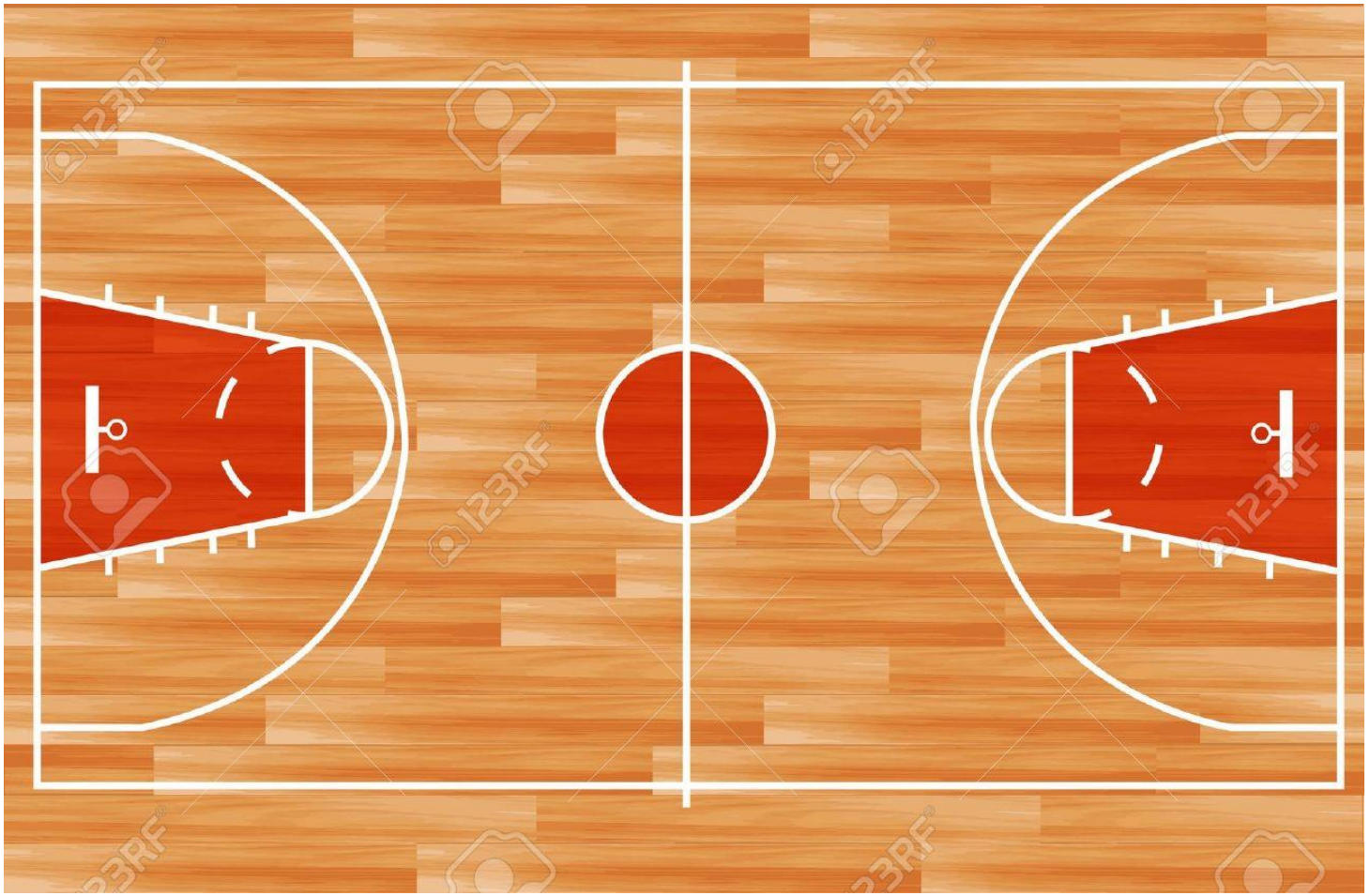
https://doi.org/10.1371/journal.pone.0147311

Overview

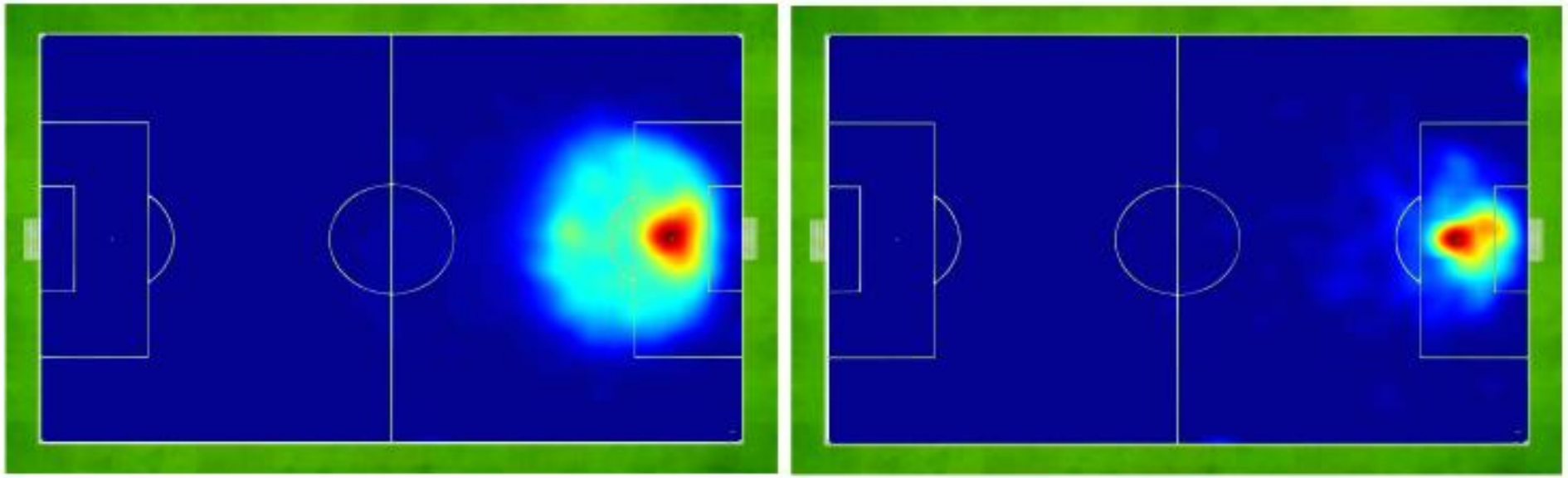
1. Bayesian Methods
2. Spatio-temporal Methods
3. Linking diverse data: Performance and Health
4. Wearables

Why spatio-temporal?

"things that are closer in space and/or time are more similar"

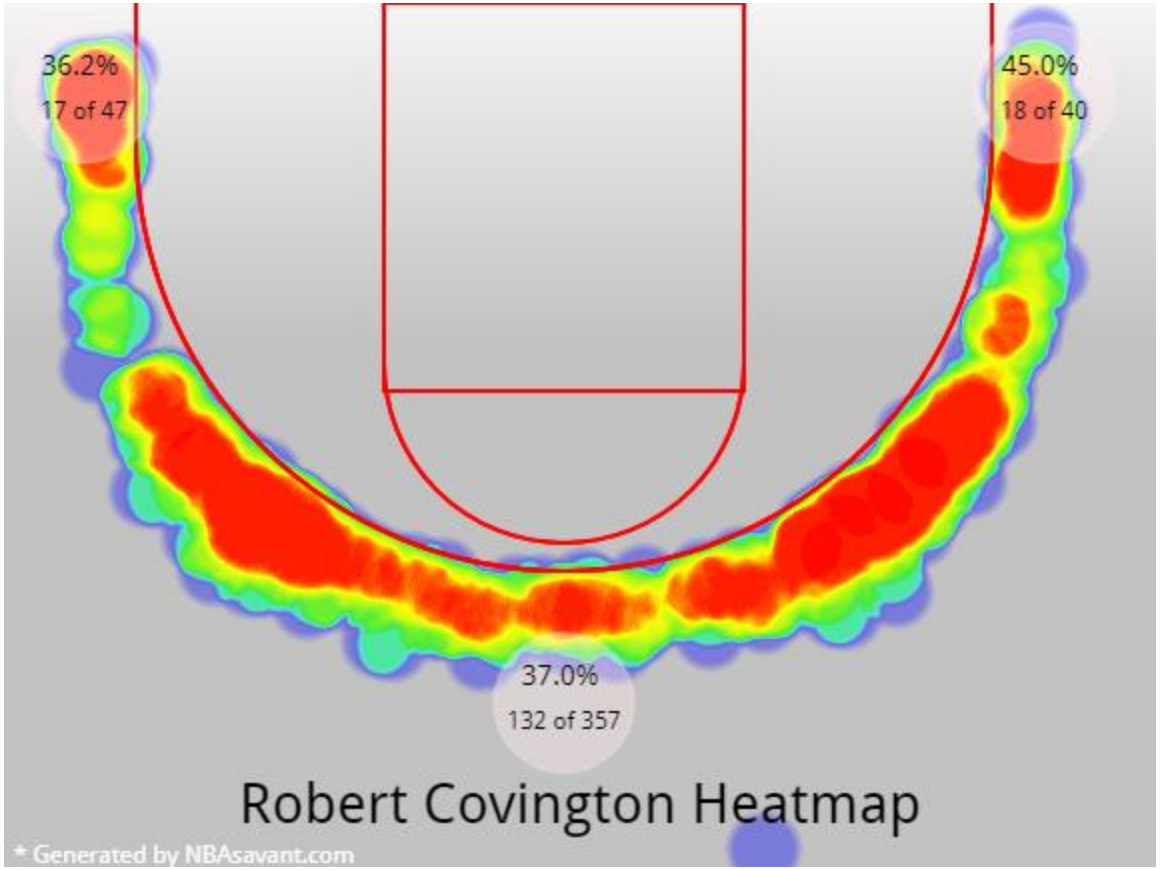


Spatio-temporal analyses in sport: VRES projects (Yu Yi Yu)



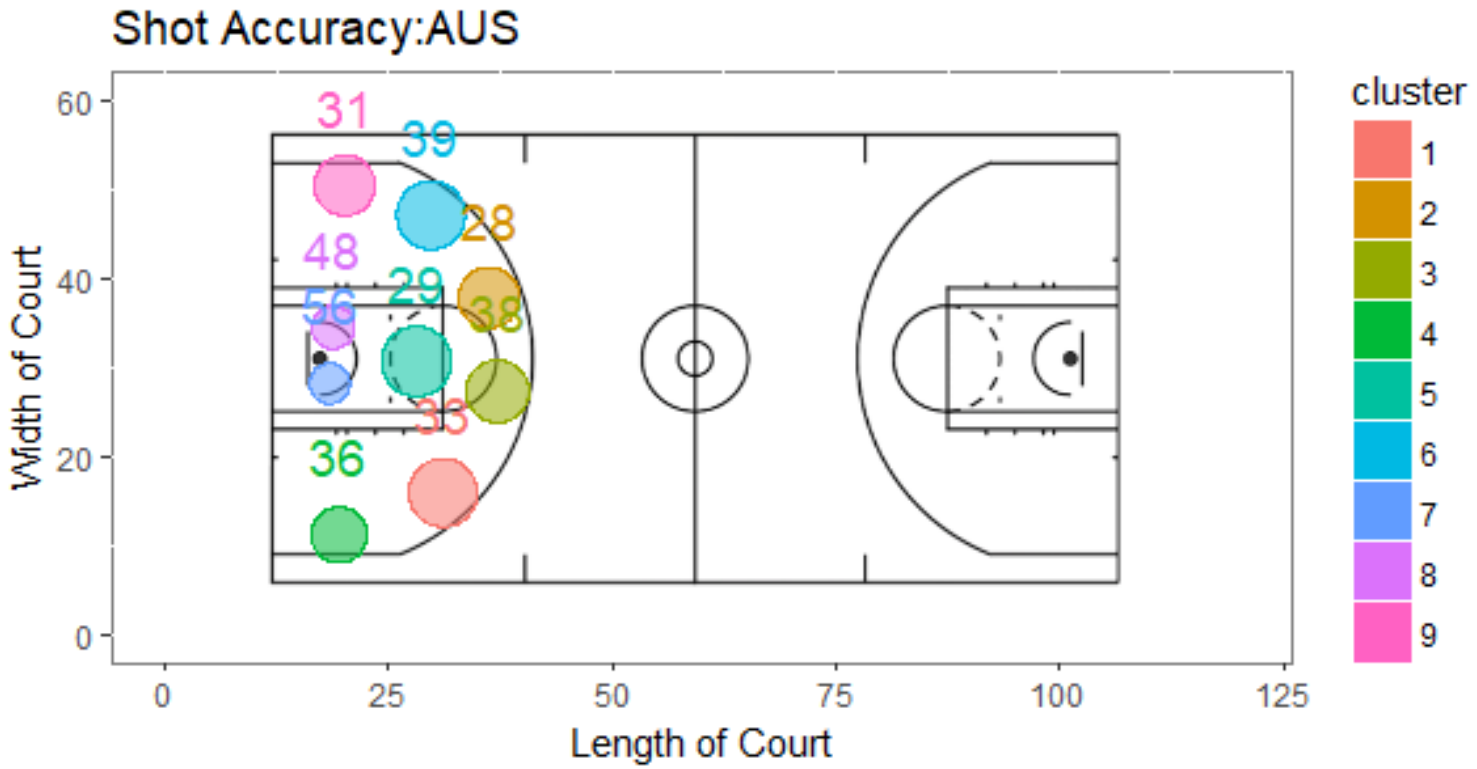
Heat map of Prominent Scoring Areas (Lucey, Bialkowski, Monfort, Carr, & Matthews, 2015)

Spatio-temporal analyses in sport: VRES projects (Yu Yi Yu)

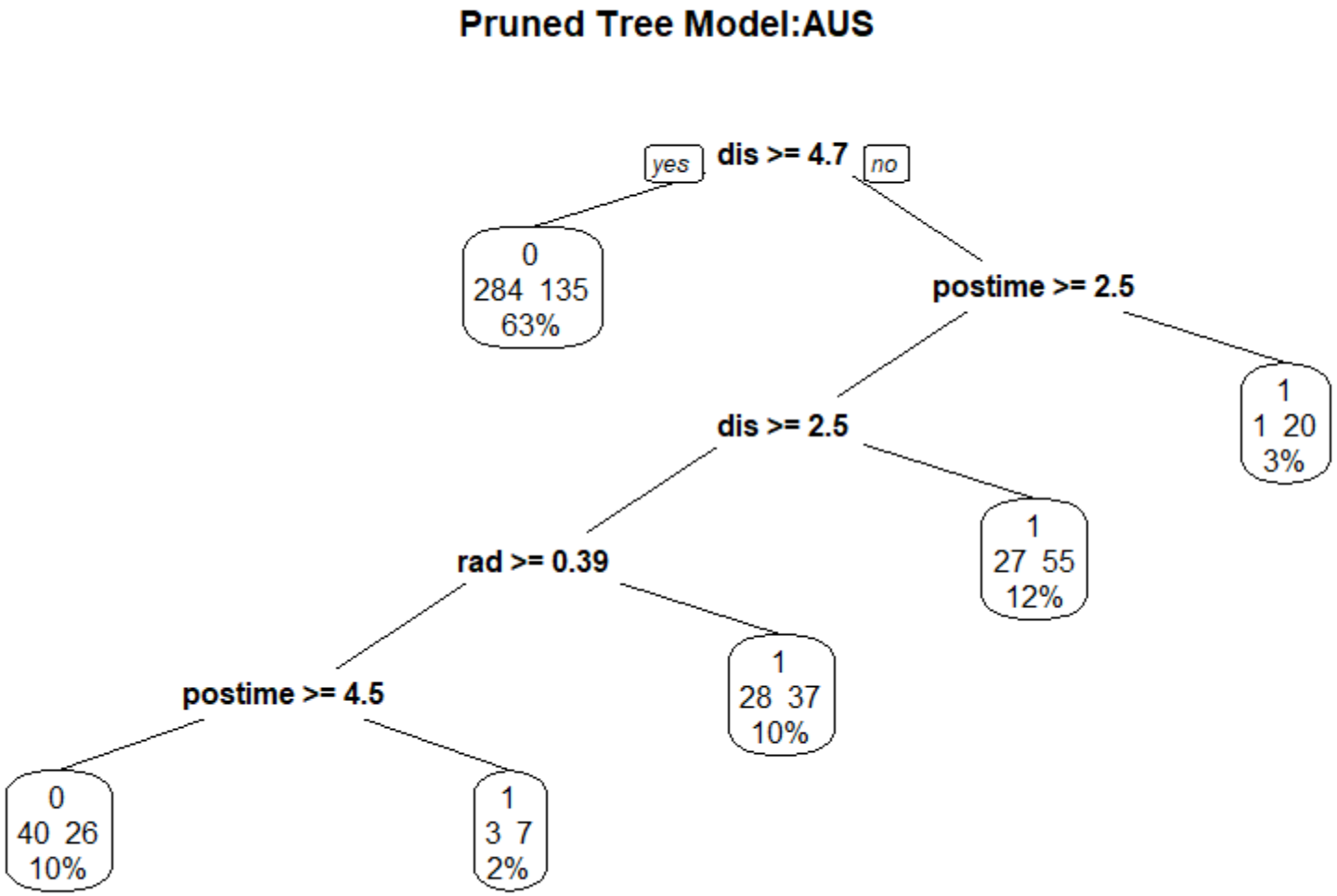


Heat map of Shot Probability for Robert Covington (Willman, 2014)

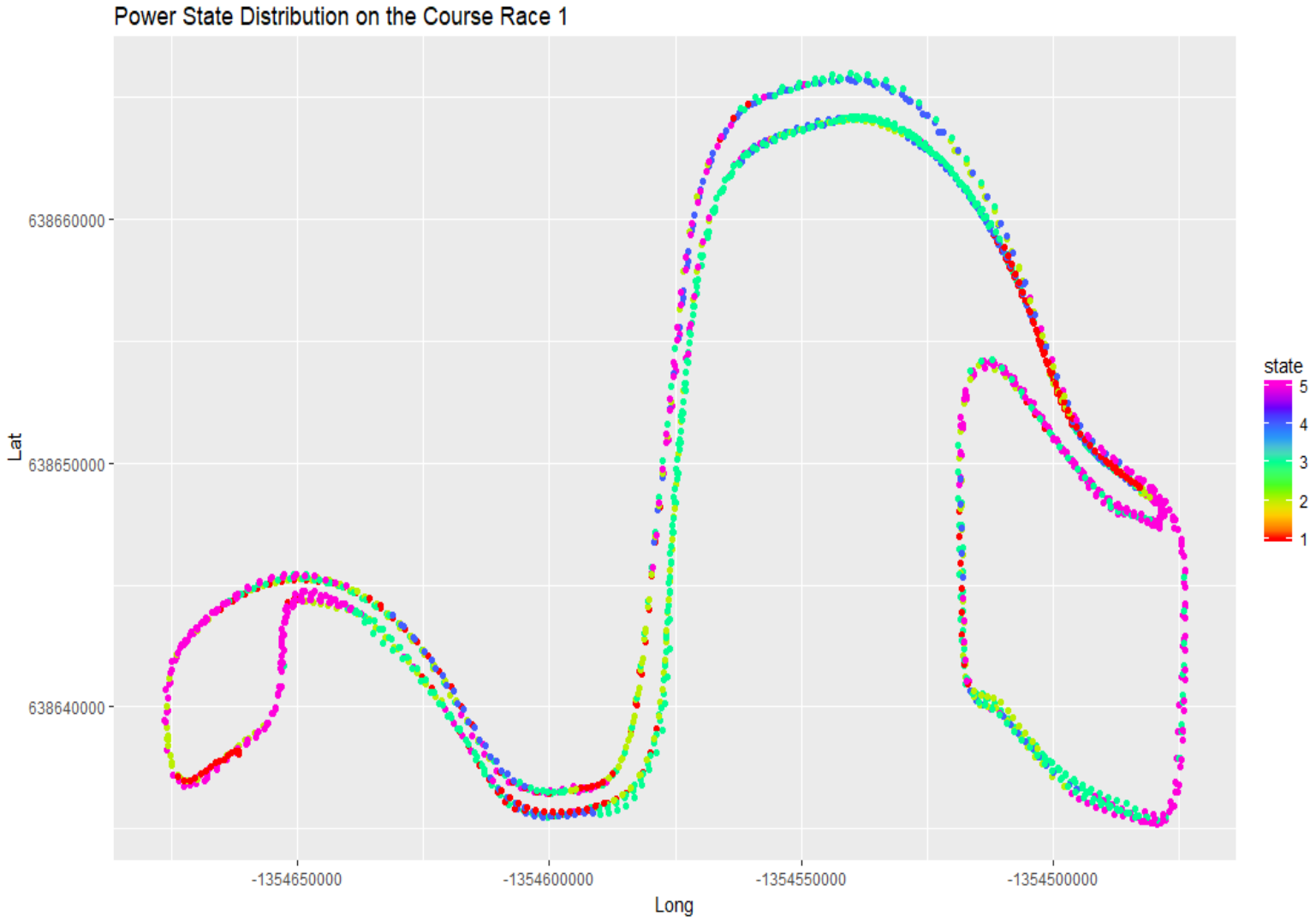
Spatio-temporal analyses in sport: VRES projects (Yu Yi Yu)



Spatio-temporal analyses in sport: VRES projects (Yu Yi Yu)



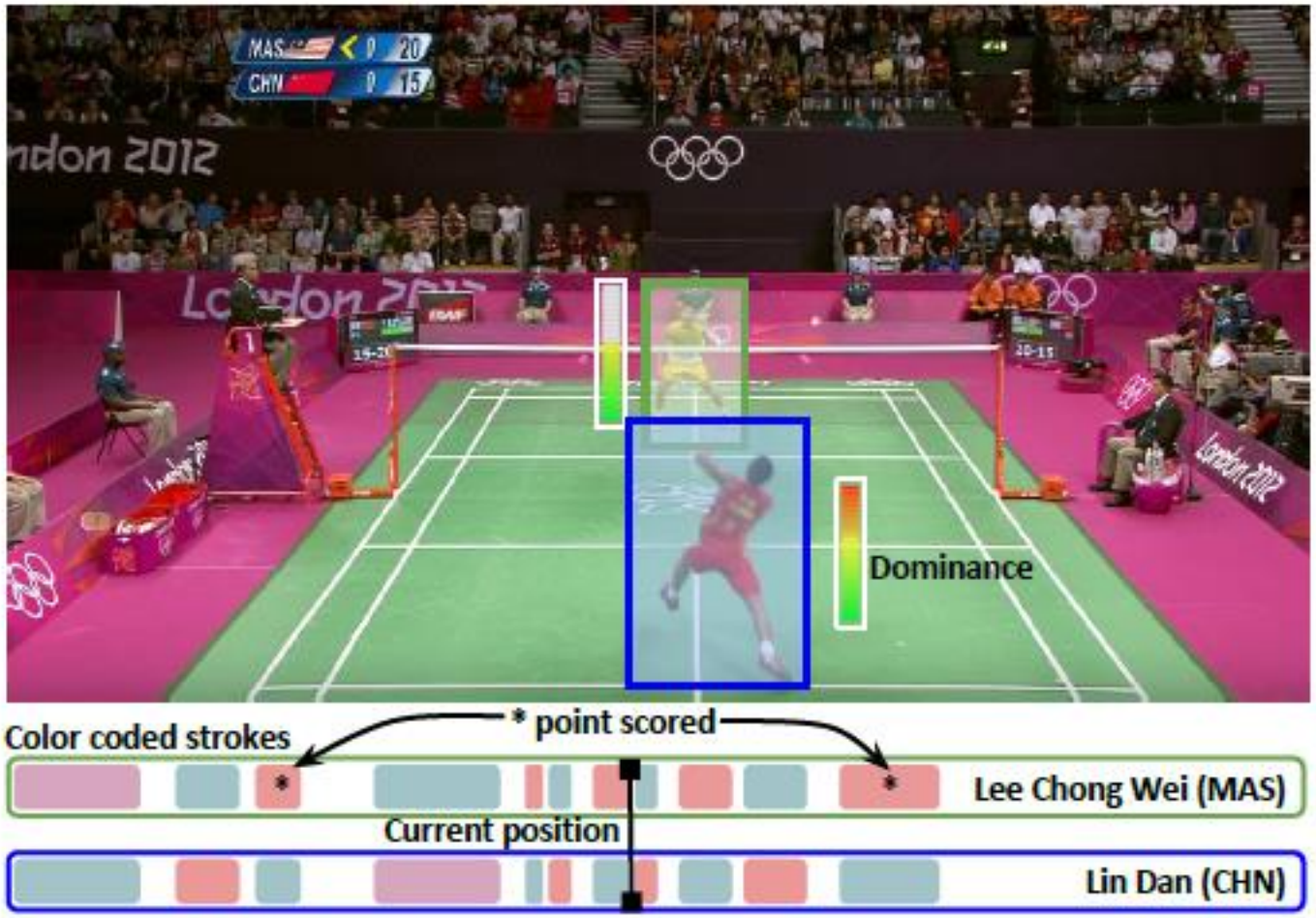
Spatio-temporal analyses in sport: VRES projects (Lawrence Garufi)



Example: Cervone (2016)

- Basketball games evolve continuously in space and time as players constantly interact with their teammates, the opposing team, and the ball.
- However, current analyses of basketball outcomes rely on discretized summaries of the game that reduce such interactions to tallies of points, assists, and similar events.
- Instead, use optical player tracking data to estimate, in real time, the expected number of points obtained by the end of a possession (EPV).
- Model at multiple levels of resolution: continuous movements of players, and discrete events such as shot attempts and turnovers.
- Estimate transitions using hierarchical spatiotemporal models that share information across players.
- Reveal novel insights on players' decision-making tendencies as a function of their spatial strategy.

Example: Ghosh et al. (2017)



Example: Wei et al. (QUT, 2016)

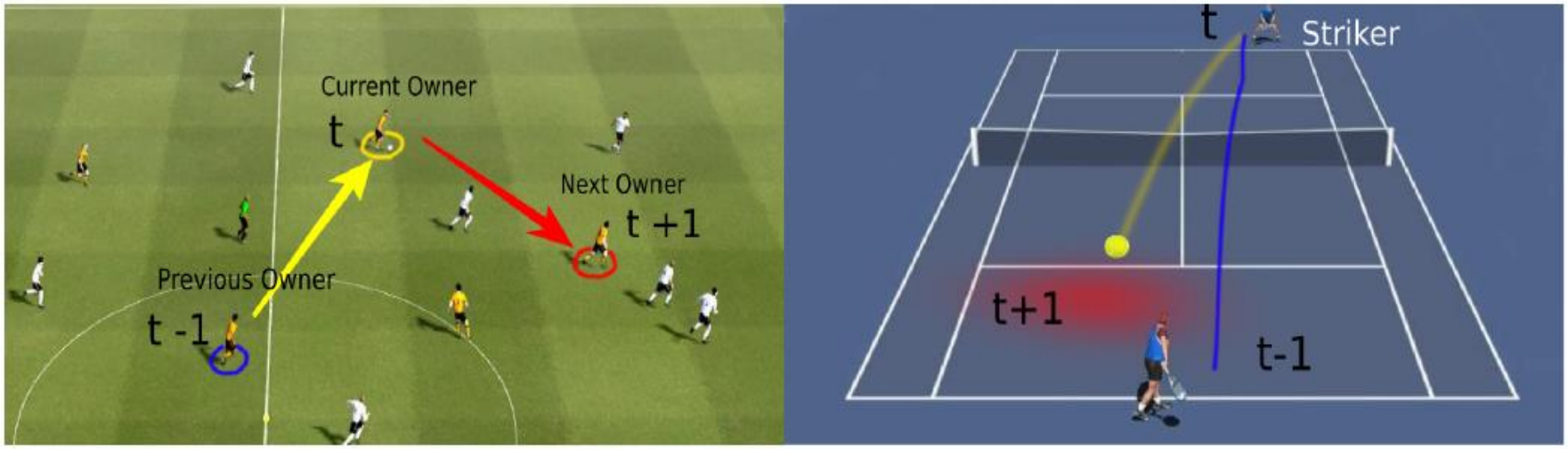


Fig. 1. In this paper, we use our a-HCRF method to: (left) predict the next pass in soccer, and (right) predict the location of the next shot in tennis.

Example: Wei et al. (QUT, 2016)

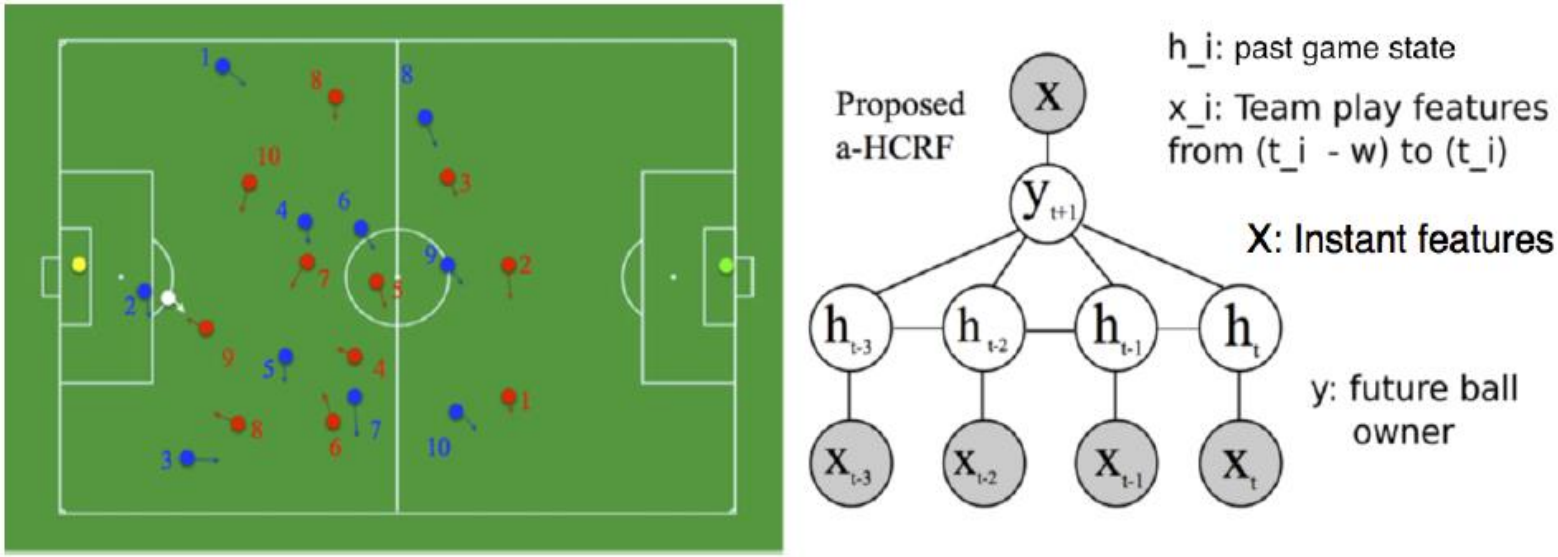


Fig. 3. (Left) In each frame, we extract speed, position and moving direction for each player. (Right) Model Representation for future ball owner prediction

Overview

1. Bayesian Methods
2. Spatio-temporal Methods
3. Linking diverse data: Performance and Health
4. Wearables

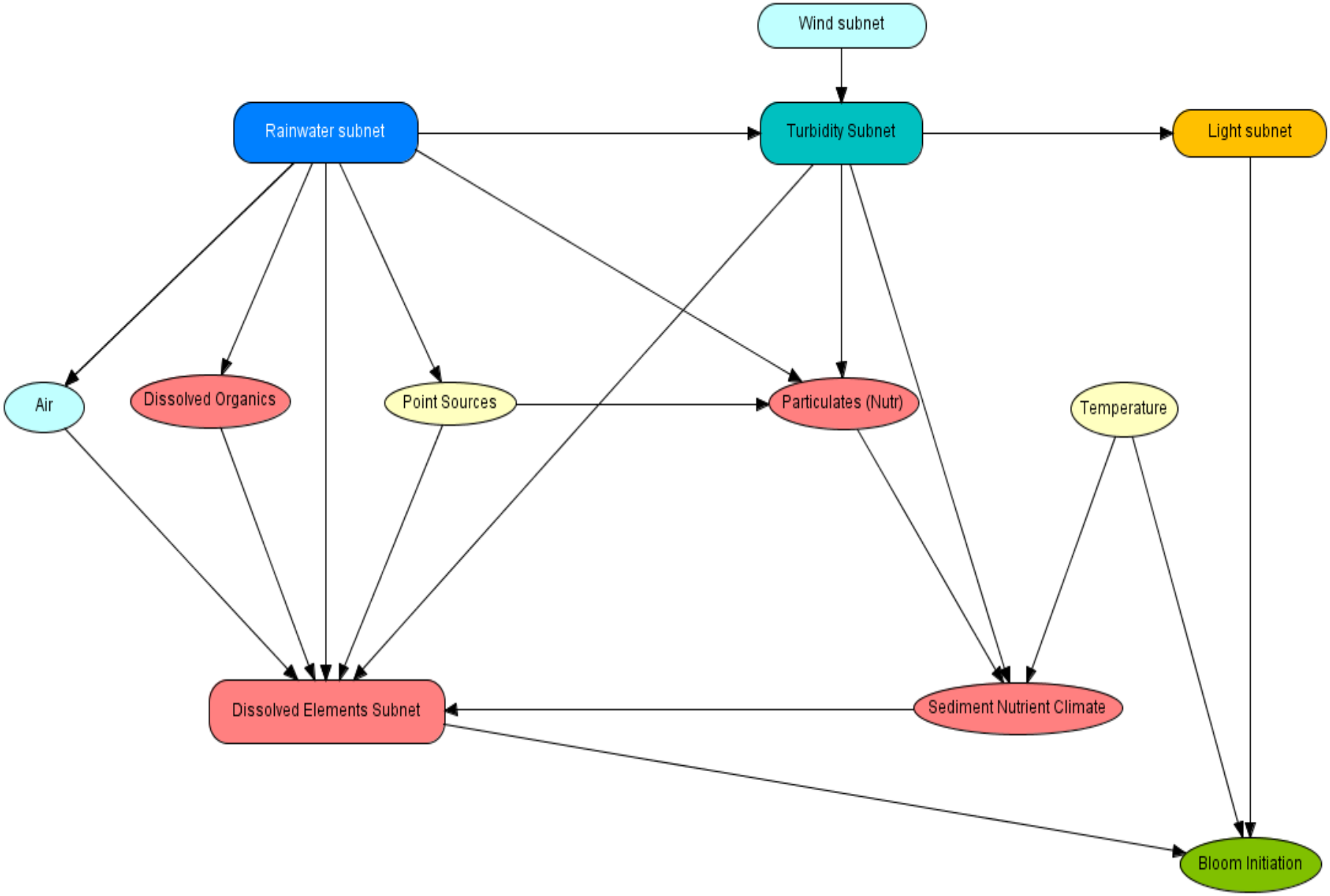
Combining Diverse Data

- **Bayesian Models**
  - Use priors
  - Hierarchical models
- **Bayesian Networks (BN)**
  - Create a conceptual graphical model of the system
  - Quantify with probabilities.

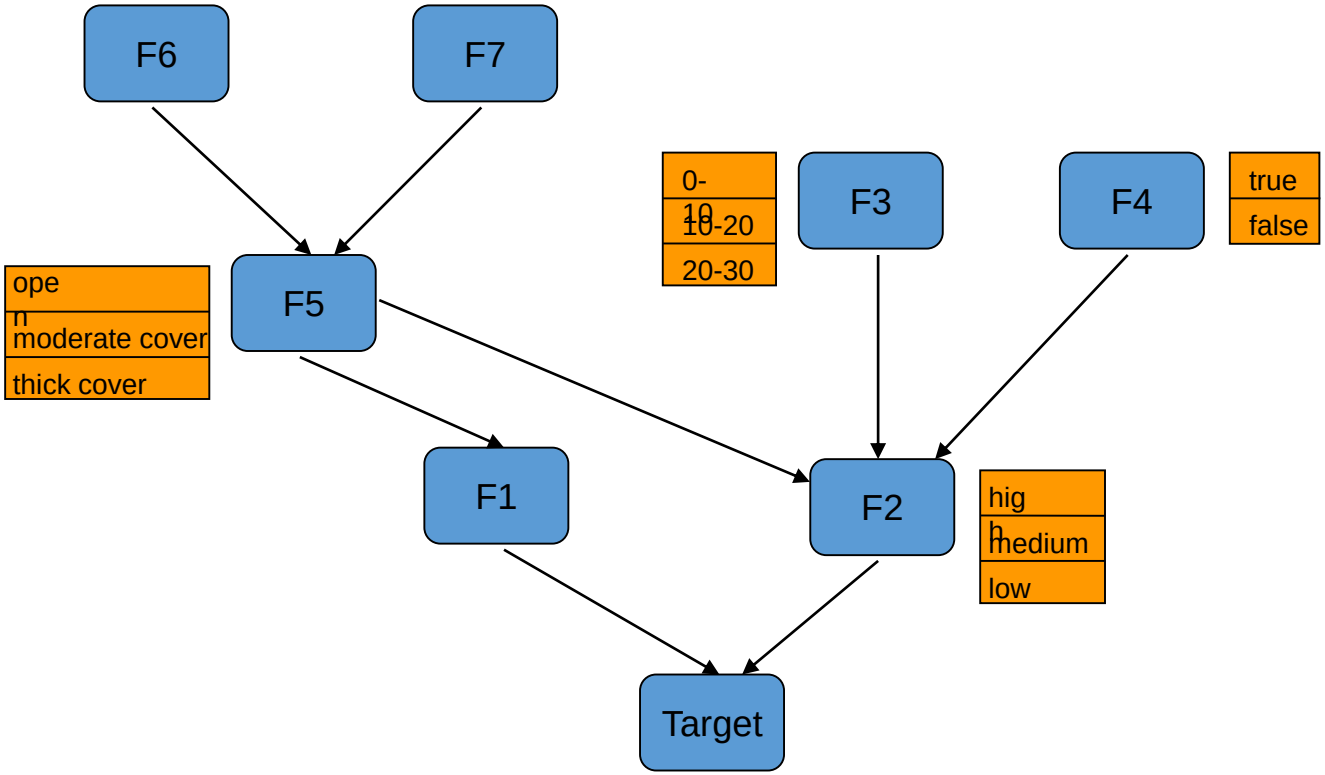
Case Study: Bayesian Networks
From science to management

What is the overall scientific consensus about the drivers of lyngbya?

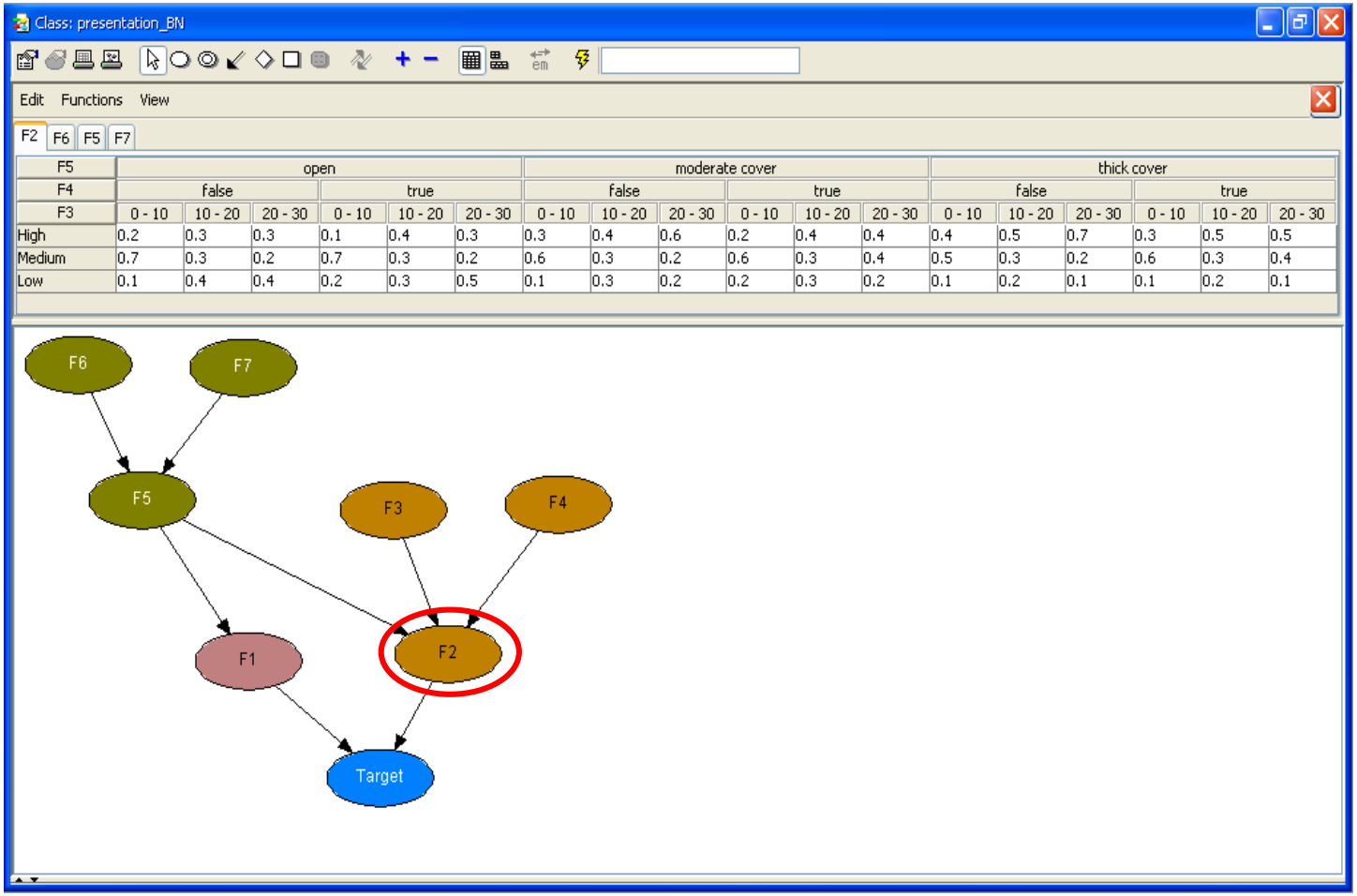
What management actions should be taken to reduce lyngbya in Moreton Bay, Australia?


Wind subnet
Rainwater subnet
Turbidity Subnet
Light subnet
Air
Dissolved Organics
Point Sources
Particulates (Nutr)
Temperature
Dissolved Elements Subnet
Sediment Nutrient Climate
Bloom Initiation

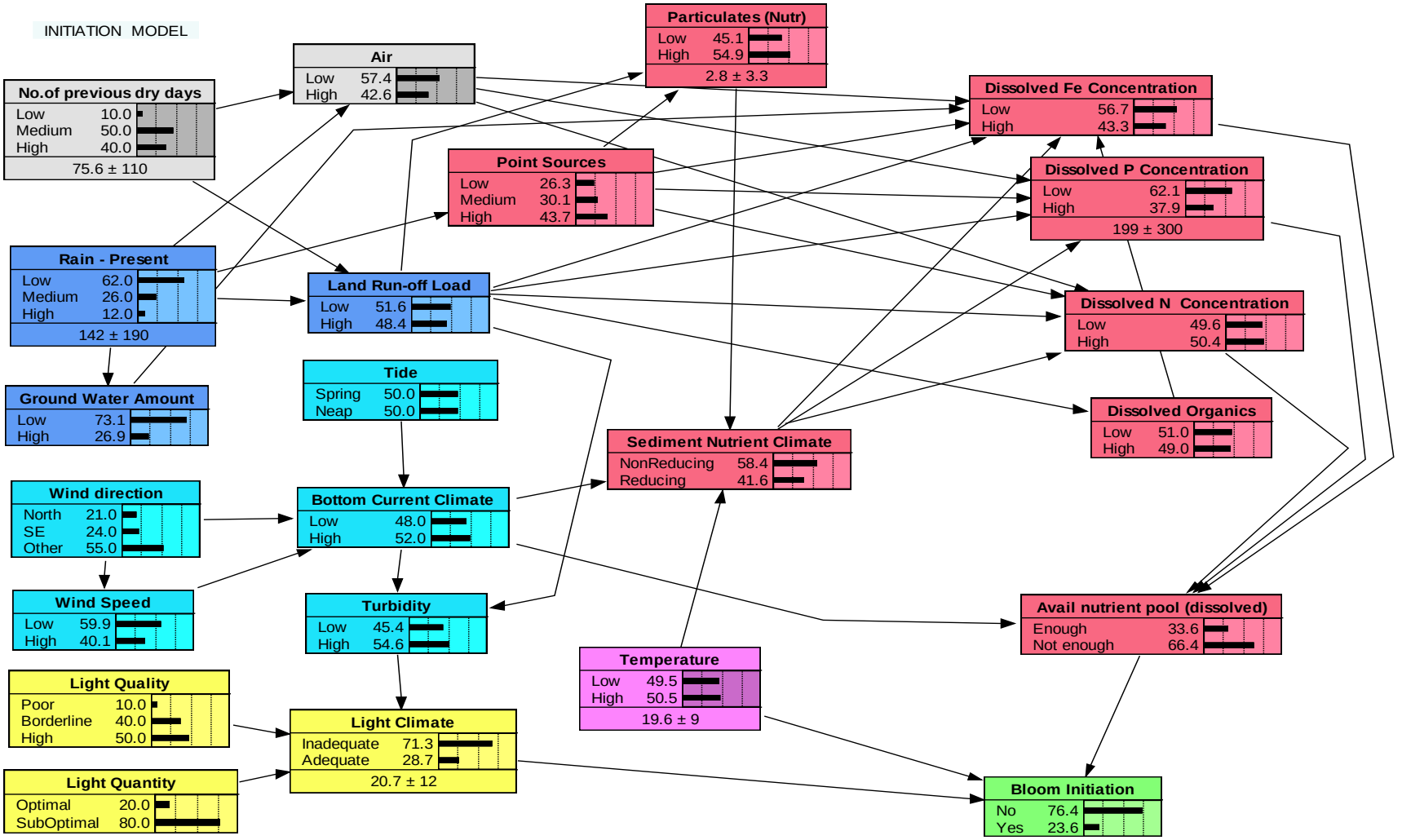
From concept to quantification



Constructing a BN - CPTs



F5	open						moderate cover						thick cover					
F4	false			true			false			true			false			true		
F3	0 - 10	10 - 20	20 - 30	0 - 10	10 - 20	20 - 30	0 - 10	10 - 20	20 - 30	0 - 10	10 - 20	20 - 30	0 - 10	10 - 20	20 - 30	0 - 10	10 - 20	20 - 30
High	0.2	0.3	0.3	0.1	0.4	0.3	0.3	0.4	0.6	0.2	0.4	0.4	0.4	0.5	0.7	0.3	0.5	0.5
Medium	0.7	0.3	0.2	0.7	0.3	0.2	0.6	0.3	0.2	0.6	0.3	0.4	0.5	0.3	0.2	0.6	0.3	0.4
Low	0.1	0.4	0.4	0.2	0.3	0.5	0.1	0.3	0.2	0.2	0.3	0.2	0.1	0.2	0.1	0.1	0.2	0.1


INITIATION MODEL
No.of previous dry days
Low 10.0
Medium 50.0
High 40.0
75.6 ± 110
Air
Low 57.4
High 42.6
Particulates (Nutr)
Low 45.1
High 54.9
2.8 ± 3.3
Dissolved Fe Concentration
Low 56.7
High 43.3
Point Sources
Low 26.3
Medium 30.1
High 43.7
Dissolved P Concentration
Low 62.1
High 37.9
199 ± 300
Rain - Present
Low 62.0
Medium 26.0
High 12.0
142 ± 190
Land Run-off Load
Low 51.6
High 48.4
Dissolved N Concentration
Low 49.6
High 50.4
Tide
Spring 50.0
Neap 50.0
Ground Water Amount
Low 73.1
High 26.9
Dissolved Organics
Low 51.0
High 49.0
Sediment Nutrient Climate
NonReducing 58.4
Reducing 41.6
Wind direction
North 21.0
SE 24.0
Other 55.0
Bottom Current Climate
Low 48.0
High 52.0
Wind Speed
Low 59.9
High 40.1
Turbidity
Low 45.4
High 54.6
Avail nutrient pool (dissolved)
Enough 33.6
Not enough 66.4
Temperature
Low 49.5
High 50.5
19.6 ± 9
Light Quality
Poor 10.0
Borderline 40.0
High 50.0
Light Climate
Inadequate 71.3
Adequate 28.7
20.7 ± 12
Light Quantity
Optimal 20.0
SubOptimal 80.0
Bloom Initiation
No 76.4
Yes 23.6

Most influential factors

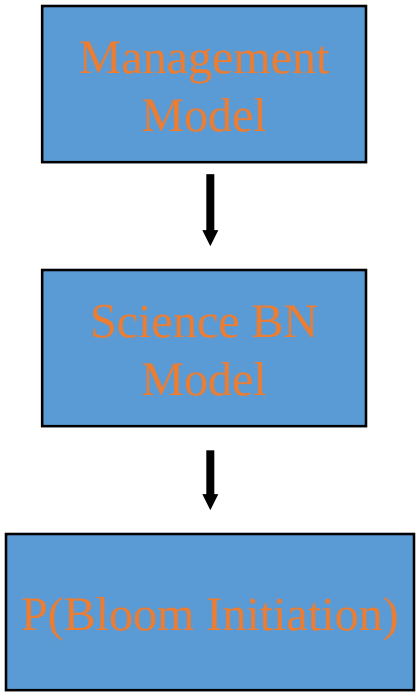
1. Available Nutrient Pool
2. Bottom Current Climate
3. Sediment Nutrients
4. Dissolved Iron
5. Dissolved Phosphorous
6. Light
7. Temperature

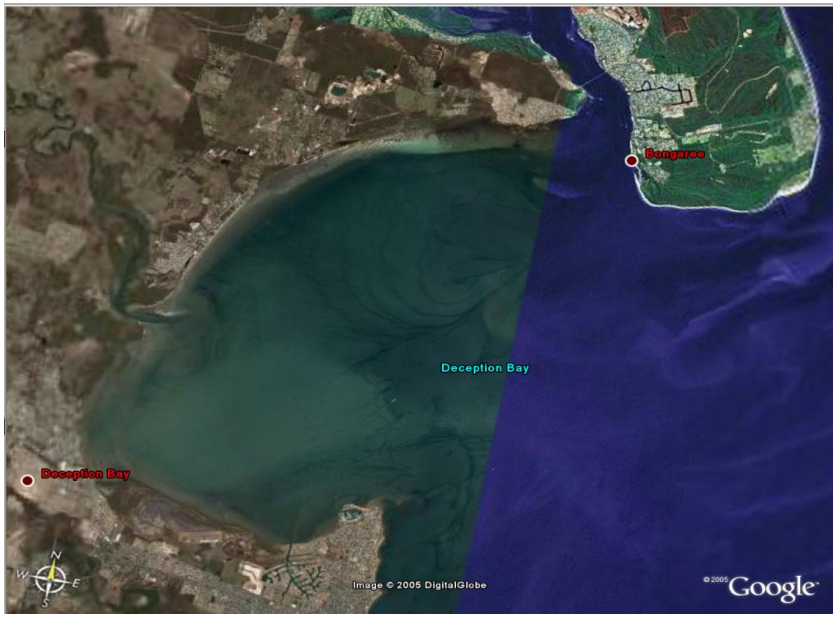
MANAGEMENT ACTIONS

“What-if” scenarios

Factor	Change in P(Bloom) (%)
Available Nutrient Pool	77 (3% - 80%)
Bottom Current Climate	28 (15% - 43%)
Sediment Nutrient Climate	17 (21% - 38%)
Dissolved Fe	16 (21% - 37%)
Dissolved P	15 (23% - 38%)
Light Climate	14 (18% - 32%)
Temperature	14 (21% - 35%)
Dissolved N	13 (22% - 35%)
Rain – present	10 (25% - 35%)
Light Quantity	9 (21% - 30%)

From Science to Management



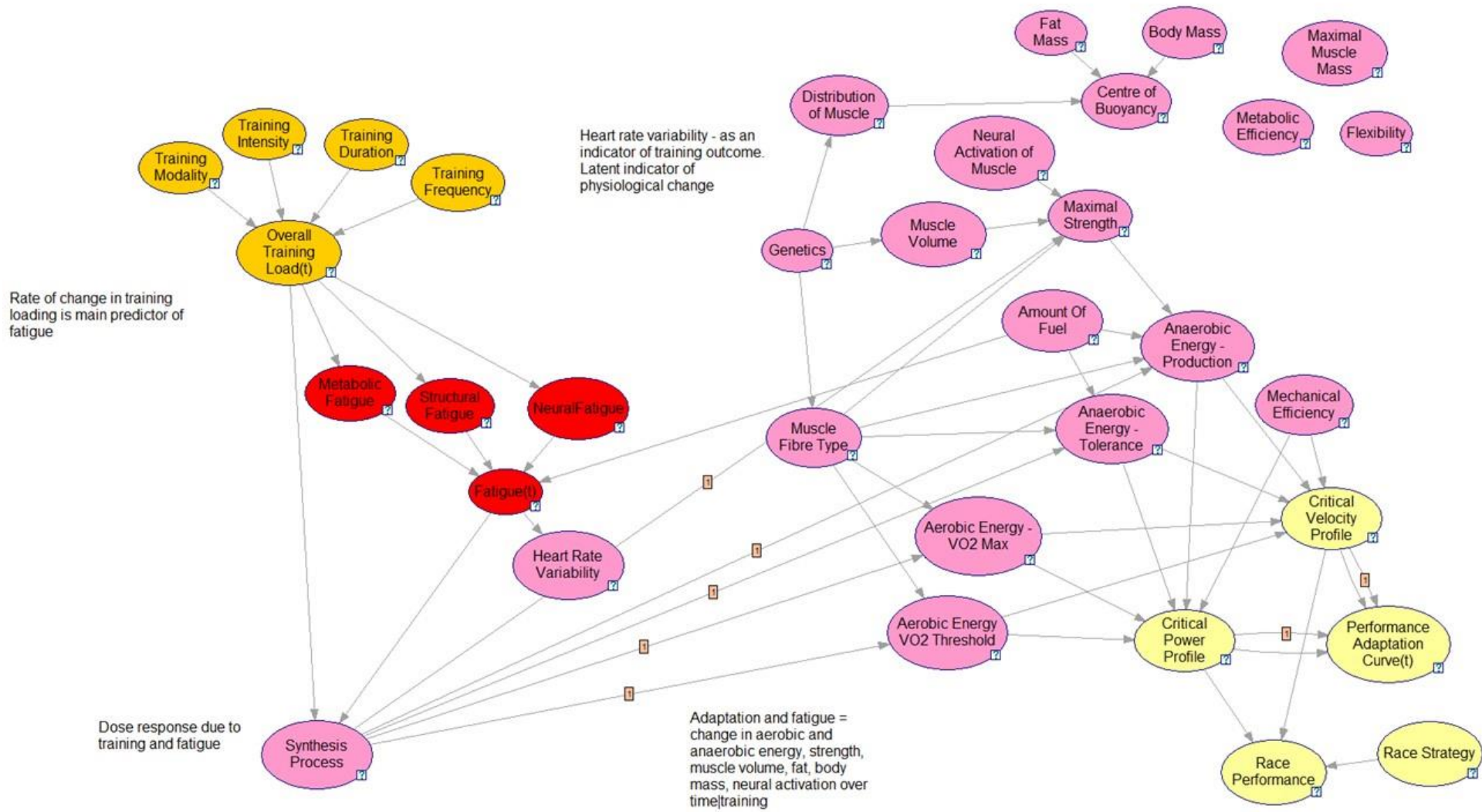


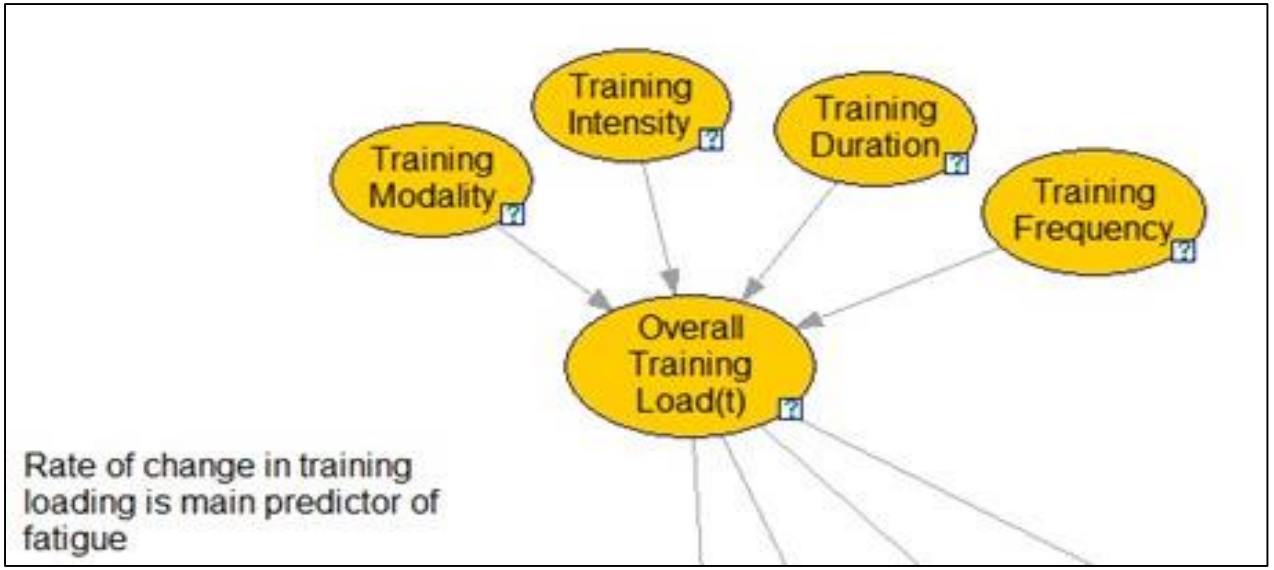
Evaluation of factors, scenario assessment
Integration of information, adaptive updates

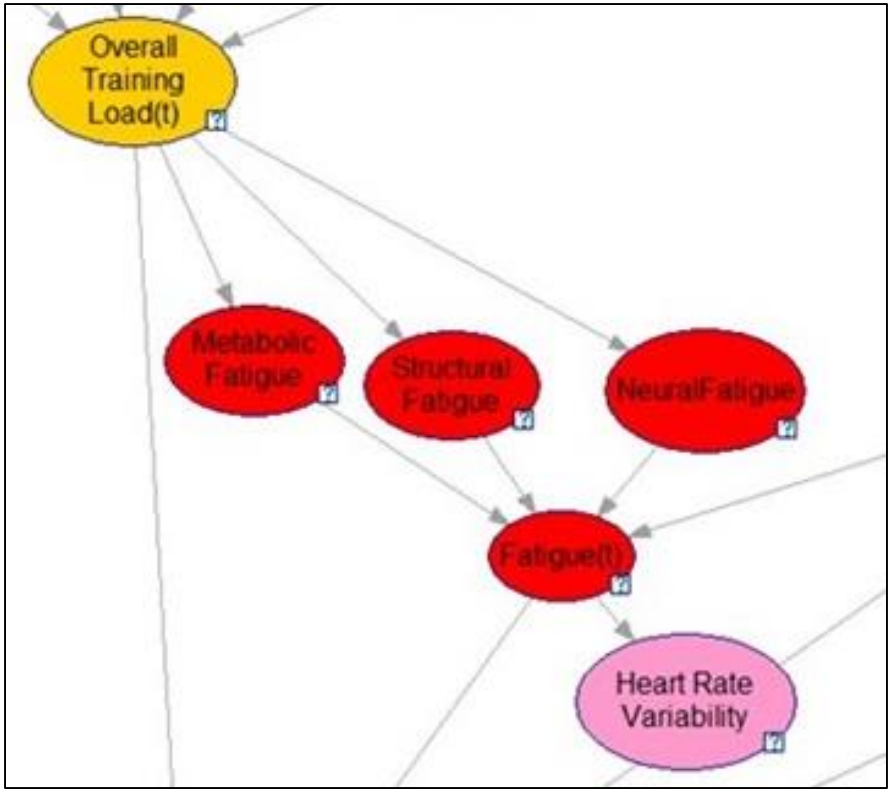
Other Applications of BNs

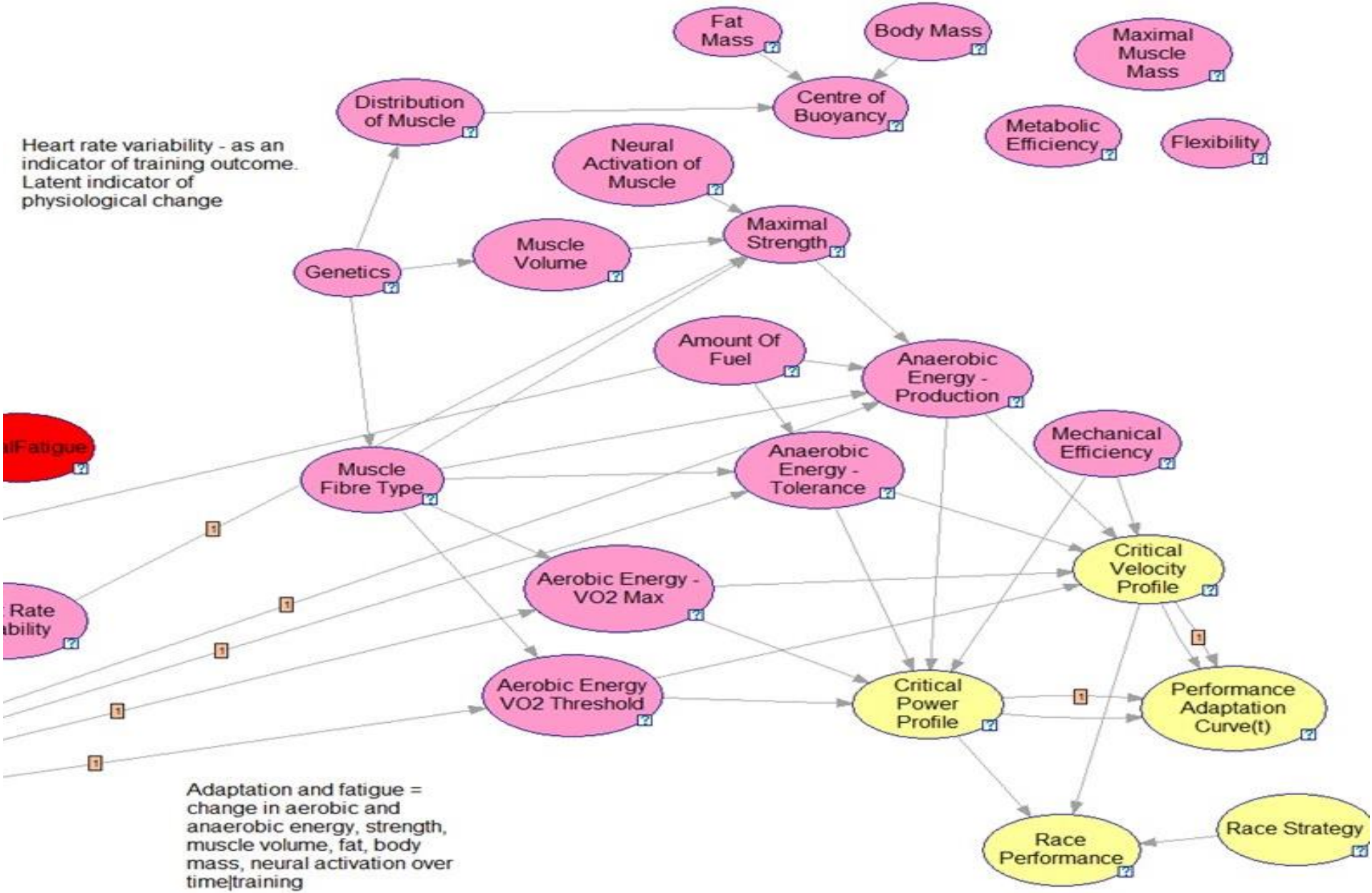
- Cheetah conservation in Southern Africa
- Airports
- Integrated asset management
- Resource management
- Recycled water and health
- Import risk
- Hospital infection
- PhD completion

Enhancing the performance and resilience of sportspeople

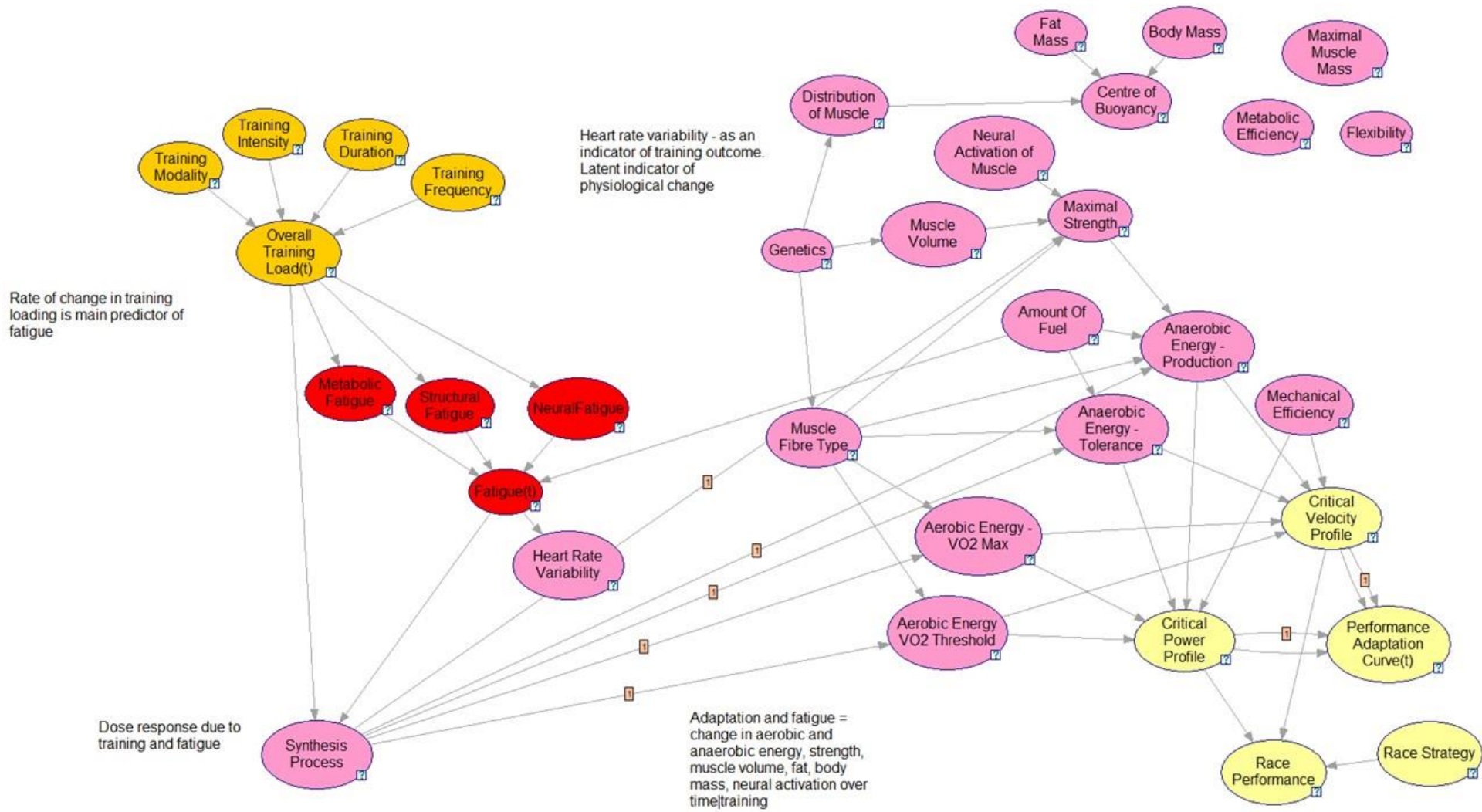



Training Intensity
Training Duration
Training Modality
Training Frequency
Overall Training Load(t)
Rate of change in training loading is main predictor of fatigue


Overall Training Load(t)
Metabolic Fatigue
Structural Fatigue
NeuralFatigue
Fatigue(t)
Heart Rate Variability


Fat Mass
Body Mass
Maximal Muscle Mass
Distribution of Muscle
Centre of Buoyancy
Metabolic Efficiency
Flexibility
Heart rate variability - as an indicator of training outcome. Latent indicator of physiological change
Neural Activation of Muscle
Maximal Strength
Muscle Volume
Genetics
Amount Of Fuel
Anaerobic Energy - Production
Mechanical Efficiency
Muscle Fibre Type
Anaerobic Energy - Tolerance
Critical Velocity Profile
Aerobic Energy - VO2 Max
Aerobic Energy VO2 Threshold
Critical Power Profile
Performance Adaptation Curve(t)
Adaptation and fatigue = change in aerobic and anaerobic energy, strength, muscle volume, fat, body mass, neural activation over time|training
Race Performance
Race Strategy

Enhancing the performance and resilience of sportspeople



Overview

1. Bayesian Methods
2. Spatio-temporal Methods
3. Linking diverse data: Performance and Health
4. Wearables

Finally…

Future research needs a multi-disciplinary approach: performance analysts, exercise scientists, biomechanists, practitioners, statisticians and computer scientists hold the key

Rein R, Memmert D (2016) Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. SpringerPlus 5:1410. https://doi.org/10.1186/s40064-016-3108-2

https://springerplus.springeropen.com/articles/10.1186/s40064-016-3108-2

Reading: Bayesian methods

- John Kruschke
  Book: “Doing Bayesian Data Analysis”
  Video: Bayesian Methods Interpret Data Better, Monday, November 12, 2012
  http://doingbayesiandataanalysis.blogspot.com.au/2012/11/video-bayesian-methods-interpret-data.html
  Goal: estimate the underlying probability of getting a hit by each player, based on their hits H, at bats AB, and primary position

- Guo S, Sanner S, Graepel T, Buntine W. Score-based Bayesian Skill Learning
  http://research.microsoft.com/pubs/184298/sbsl_ecml2012.pdf
  Extends ‘TrueSkill’ using Bayesian networks

Reading: Bayesian methods

- Red Sox as an illustration of Bayesian probability theory.
  http://www.bankers-anonymous.com/blog/red-sox-as-an-illustration-of-bayesian-probability-theory/

- Baio G, Blangiardo MA Bayesian hierarchical model for the prediction of football results.
  http://www.statistica.it/gianluca/Research/BaioBlangiardo.pdf
  http://danielweitzenfeld.github.io/passtheroc/blog/2014/10/28/bayes-premier-league/

- Karlis D, Ntzoufras J. Bayesian and Non-Bayesian Analysis of Soccer Data using Bivariate Poisson Regression Models.
  Analysis of sports data by using bivariate Poisson models. The Statistician (2003) 52, 381–393.http://www2.stat-athens.aueb.gr/~karlis/Bivariate%20Poisson%20Regression.pdf

- https://www.pinnacle.com/en/betting-articles/Betting-Strategy/bayesian-analysis-and-sports-betting/5VM2U95MX696RGXP

Reading: Spatio-temporal methods

- Bialkowski AN, Lucey PJ, Carr P, Matthews I, Sridharan S, Fookes CB, (2016) Discovering team structures in soccer from spatiotemporal data, *IEEE Transactions on Knowledge and Data Engineering* p2596-2605.

- Cervone D, D'Amour A, Bornn L, Goldsberry K (2014). A multiresolution stochastic process model for predicting basketball possession outcomes. *arXiv*:1408.0777v1.

- Franks AM, Miller A, Bornn L, Goldsberry K (2015) Characterizing the spatial structure of defensive skill in professional basketball. *Annals of Applied Statistics* 9 (1), 94-121.

- Ghosh A, Singh S, Jawahar CV (2017) Towards structured analysis of broadcast badminton videos. *arXiv*: 1712.08714v1.

- Wei X, Lucey PJ, Morgan SW, Sridharan S, (2016) Forecasting the next shot location in tennis using fine-grained spatiotemporal tracking data, *IEEE Transactions on Knowledge and Data Engineering* p2988-2997.

Reading: Other

- Yuan L-H, Liu A, Yeh A et al. (2016) A mixture-of-modelers approach to forecasting NCAA tournament outcomes. *Journal of Quantitative Analysis in Sports* 2015; 11(1): 13–27.

- Franks AM, D'Amour A, Cervone D, Bornn L (2016) Meta-analytics: tools for understanding the statistical properties of sports metrics. *Journal of Quantitative Analysis in Sports* 12 (4), 151-165.

- Saraivaa E, Suzuki A, Filhob CAO, Louzadab F (2016) Predicting football scores via Poisson regression model: applications to the National Football League. Communications for Statistical Applications and Methods 2016, Vol. 23, 297–319. http://dx.doi.org/10.5351/CSAM.2016.23.4.297

- http://www.sports-management-degrees.com/baseball/

- Hoopdata.com, basketballvalue.com, and apbr.org.