Improving predictive accuracy using *Smart-Data* rather than *Big-Data*: A case study of soccer teams' evolving performance

Anthony Constantinou[1]

and

Norman Fenton[2]

1. *Post-Doctoral Researcher, School of EECS, Queen Mary University of London, UK.*
2. *Professor of Risk and Information Management, School of EECS, Queen Mary University of London, UK.*

Proceedings of the 13th UAI Bayesian Modeling Applications Workshop (BMAW 2016),
32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016), New York City, USA, June 29, 2016.

Introduction:

Smart-Data

What do we mean by *Smart-Data*?

- Big-data relies on automation based on the general consensus that relationships between factors of interest surface by themselves.
- Smart-data aims to improve the quality, as opposed to the quantity, of a dataset based on causal knowledge.

Introduction:

Smart-Data

What do we mean by *Smart-Data*?

- Big-data relies on automation based on the general consensus that relationships between factors of interest surface by themselves.
- Smart-data aims to improve the quality, as opposed to the quantity, of a dataset based on causal knowledge.

What does the '*quality' of a dataset* represent?

- The highest quality dataset represents the idealised information required for formal causal representation (e.g. simulated data).
- However big a dataset is, causal discovery is sub-optimal in the absence of a 'high quality' dataset.

Introduction:

Smart-Data

What do we mean by *Smart-Data*?

- Big-data relies on automation based on the general consensus that relationships between factors of interest surface by themselves.
- Smart-data aims to improve the quality, as opposed to the quantity, of a dataset based on causal knowledge.

What does the '*quality' of a dataset* represent?

- The highest quality dataset represents the idealised information required for formal causal representation (e.g. simulated data).
- However big a dataset is, causal discovery is sub-optimal in the absence of a 'high quality' dataset.

What do we propose?

- **Model engineering:** To engineer a simplified model topology based on causal knowledge.
- **Data engineering:** To engineer the dataset based on model topology such as to adhere to causal modelling (i.e. high quality) driven by what data we really require.

Introduction:

Soccer case study

Academic history

- Previous research focused on predicting the outcomes of individual soccer matches.

Introduction:

Soccer case study

Academic history

- Previous research focused on predicting the outcomes of individual soccer matches.

Our task?

- To predict a how a soccer team's performance evolves between seasons, without taking individual match instances into consideration.

Introduction:

Soccer case study

Academic history

- Previous research focused on predicting the outcomes of individual soccer matches.

Our task?

- To predict a how a soccer team's performance evolves between seasons, without taking individual match instances into consideration.

Why?

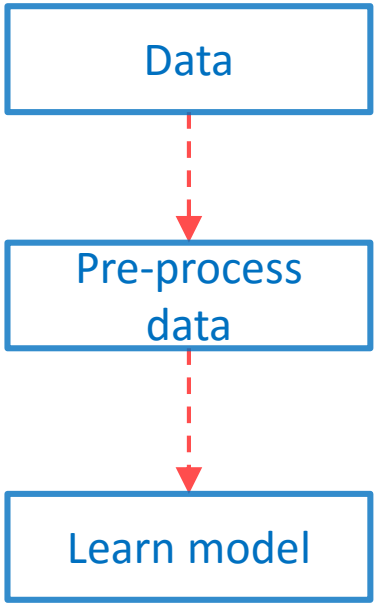
- Good case study to demonstrate the importance of a smart-data approach.
- No other model addresses this question, and which represents an enormous gambling market in itself (e.g. bettors start placing bets before a soccer season starts).

Model development process:

How does Smart-Data compare to Big-Data?

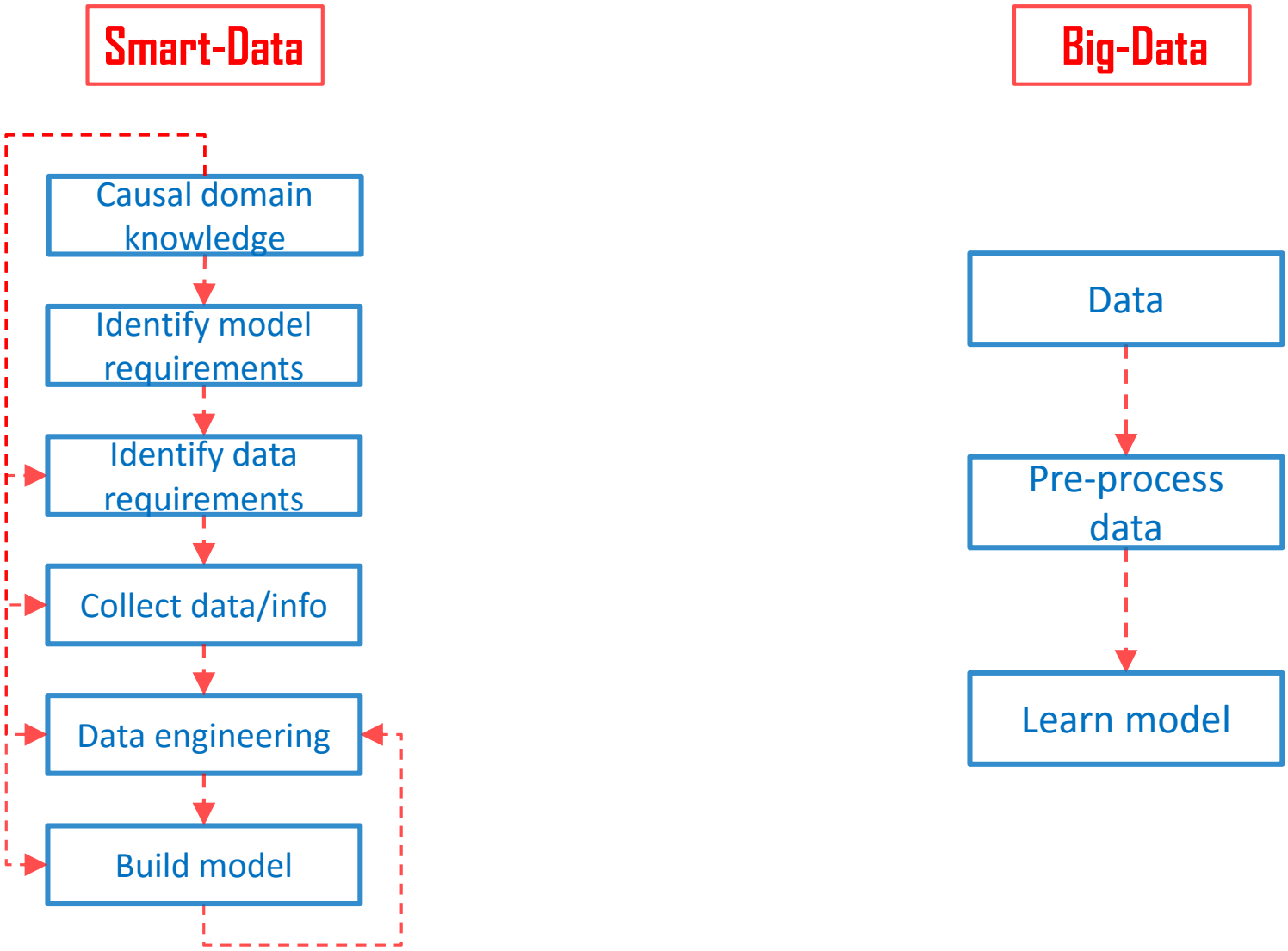
Smart-Data

Big-Data



Model development process:

How does Smart-Data compare to Big-Data?



Identifying model requirements

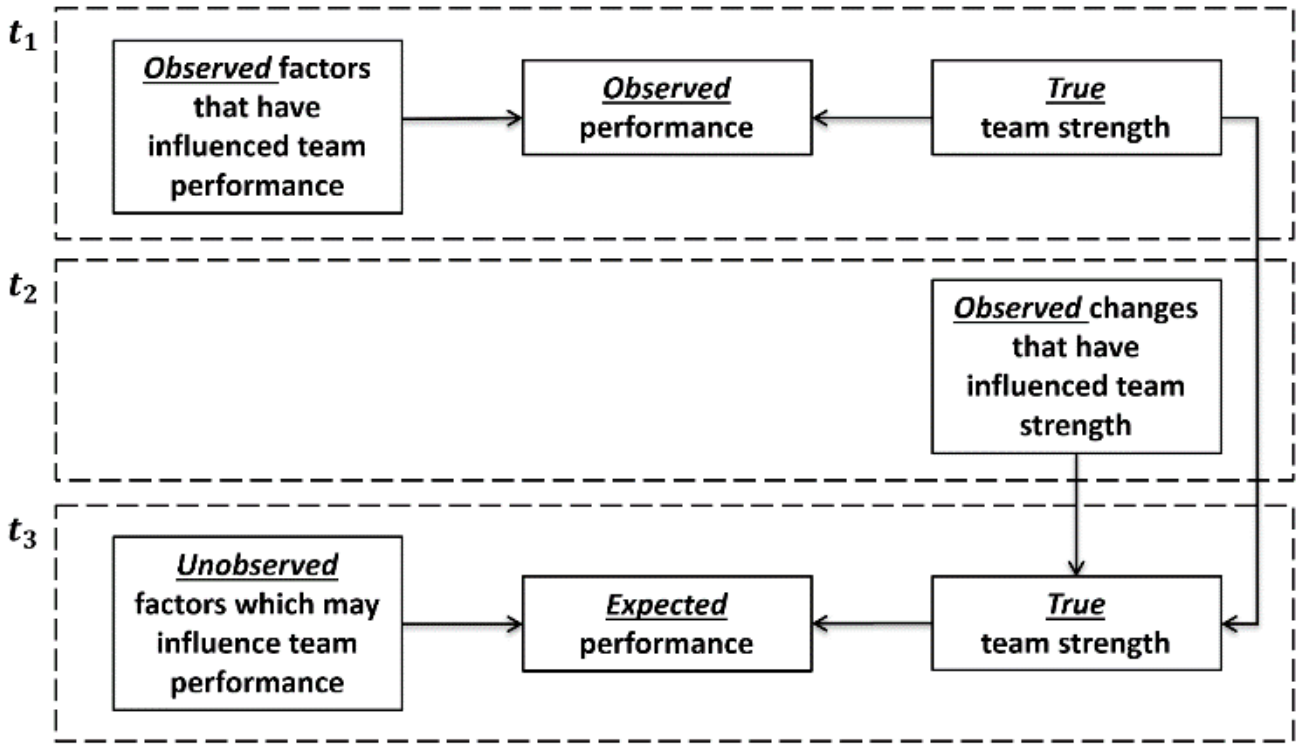


Figure 1. Simplified model topology of the overall Bayesian network model.

Where:

- t_1 is the previous season;
- t_2 is the summer break;
- t_3 is the next season

Identifying model requirements

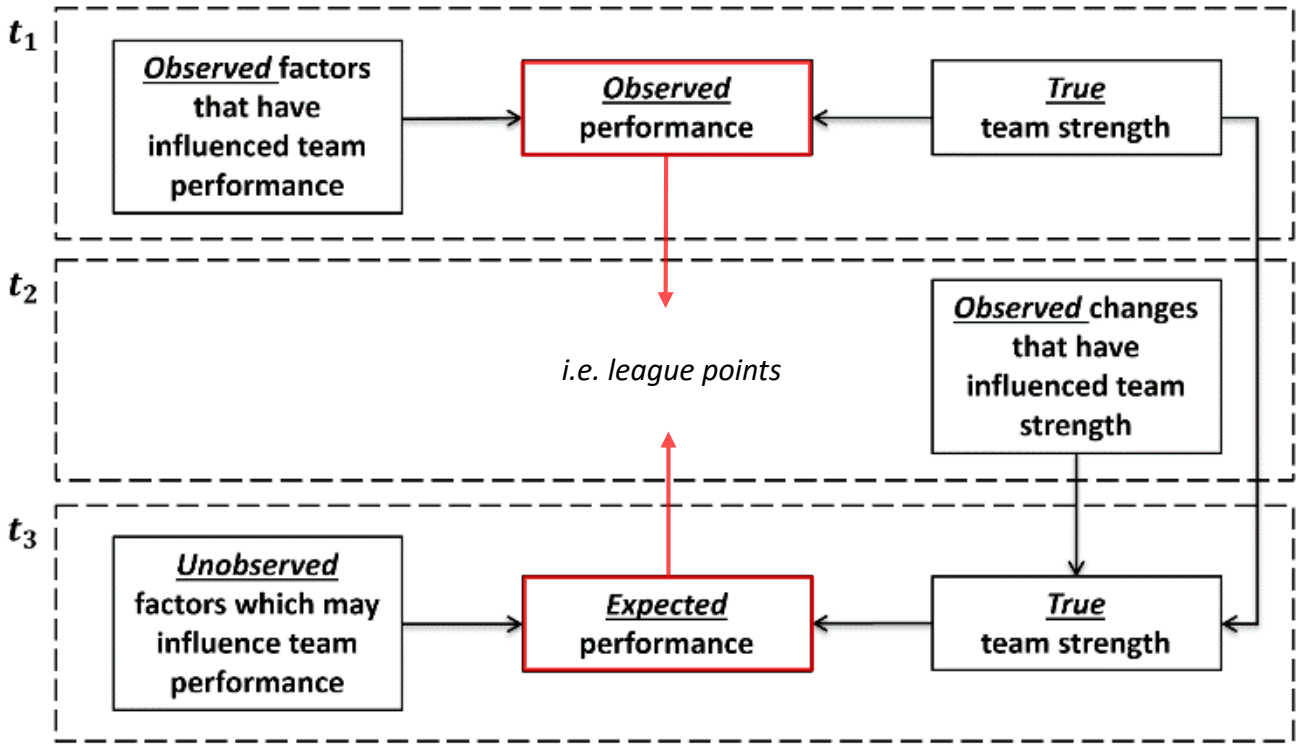


Figure 1. Simplified model topology of the overall Bayesian network model.

Where:

- t_1 is the previous season;
- t_2 is the summer break;
- t_3 is the next season

Identifying model requirements

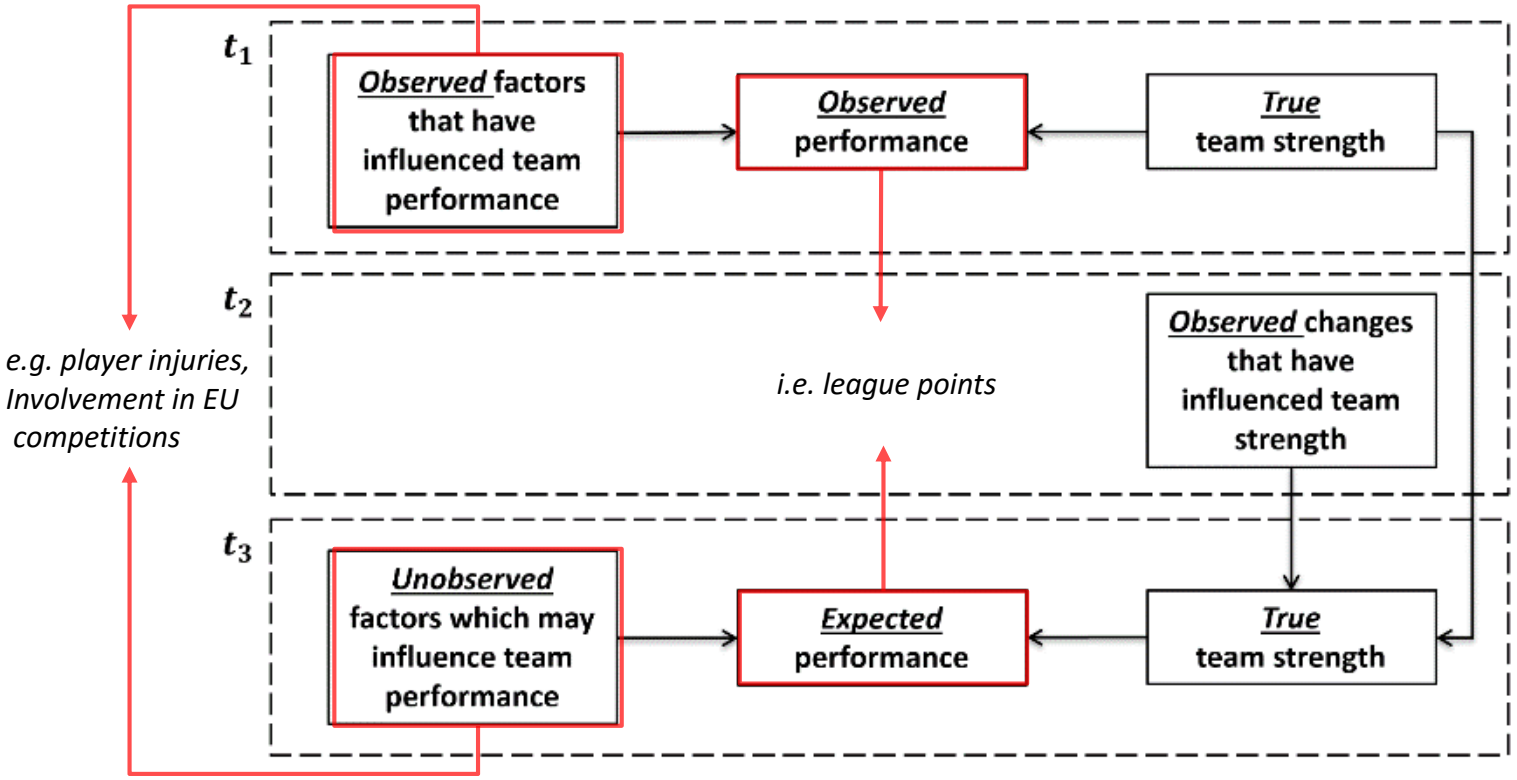


Figure 1. Simplified model topology of the overall Bayesian network model.

Where:

- t_1 is the previous season;
- t_2 is the summer break;
- t_3 is the next season

Identifying model requirements

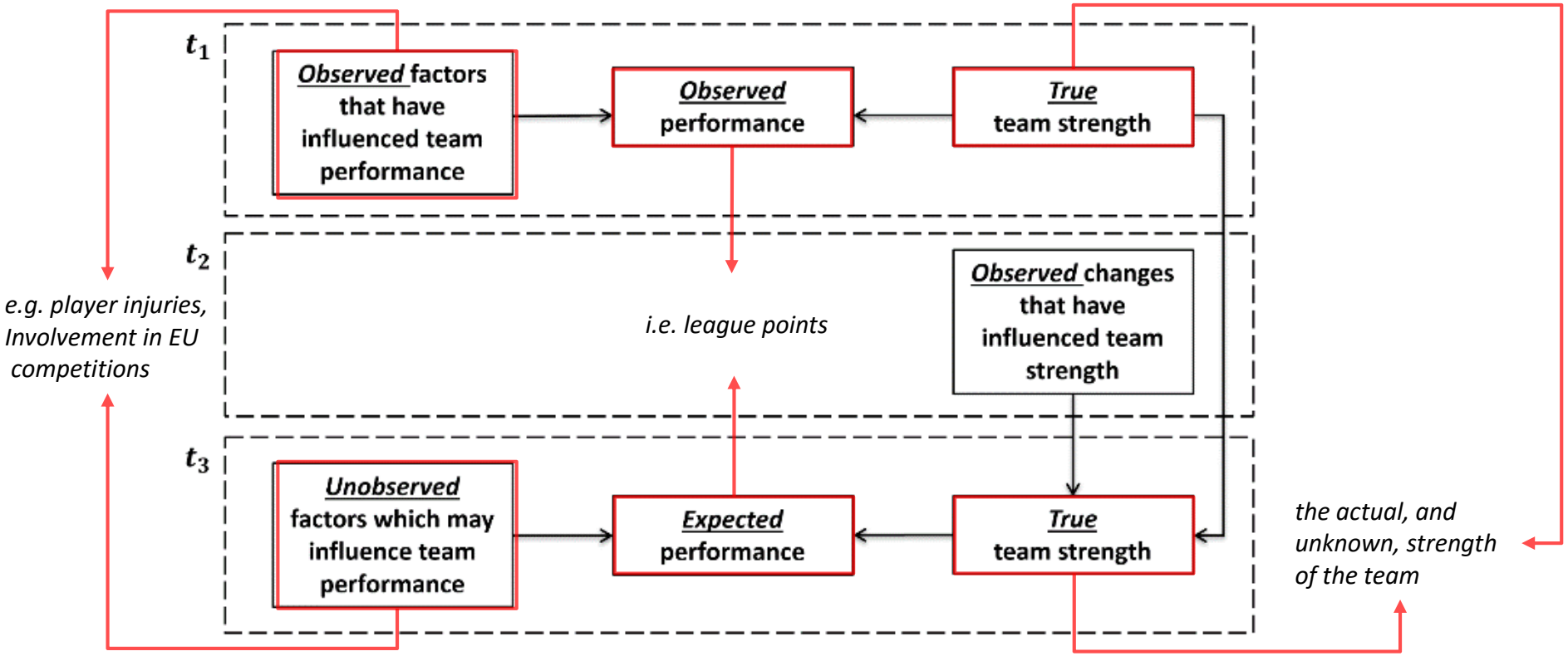


Figure 1. Simplified model topology of the overall Bayesian network model.

Where:

- t_1 is the previous season;
- t_2 is the summer break;
- t_3 is the next season

Identifying model requirements

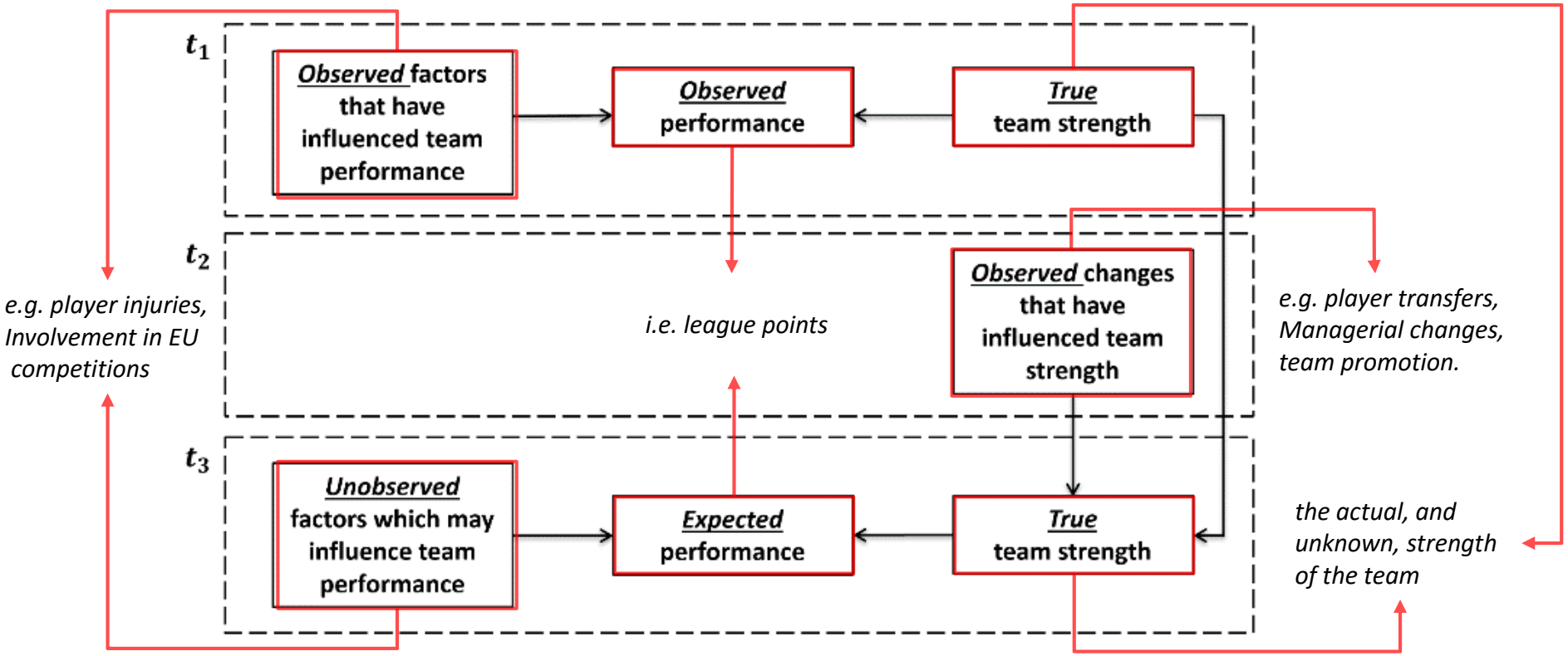
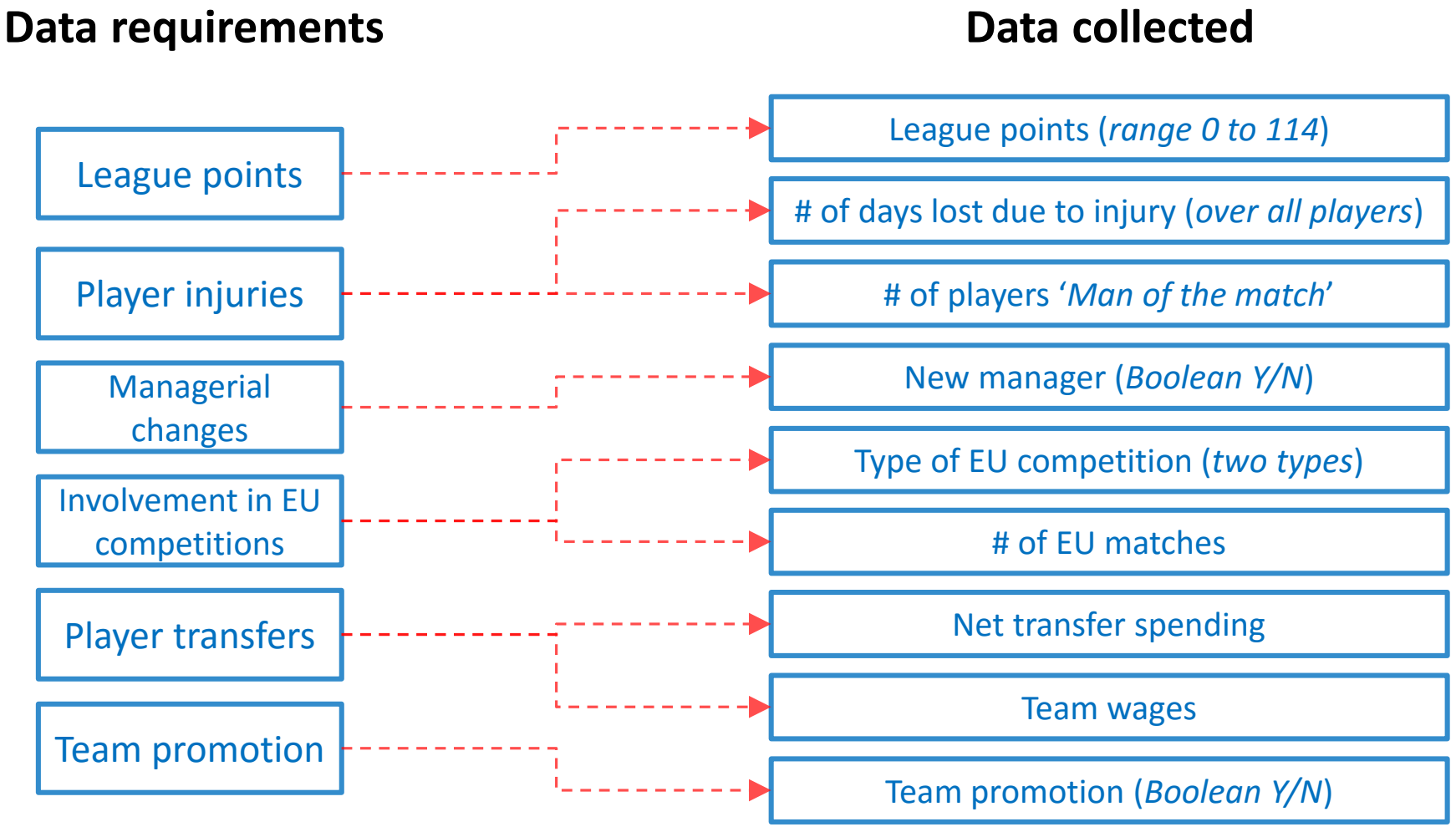


Figure 1. Simplified model topology of the overall Bayesian network model.

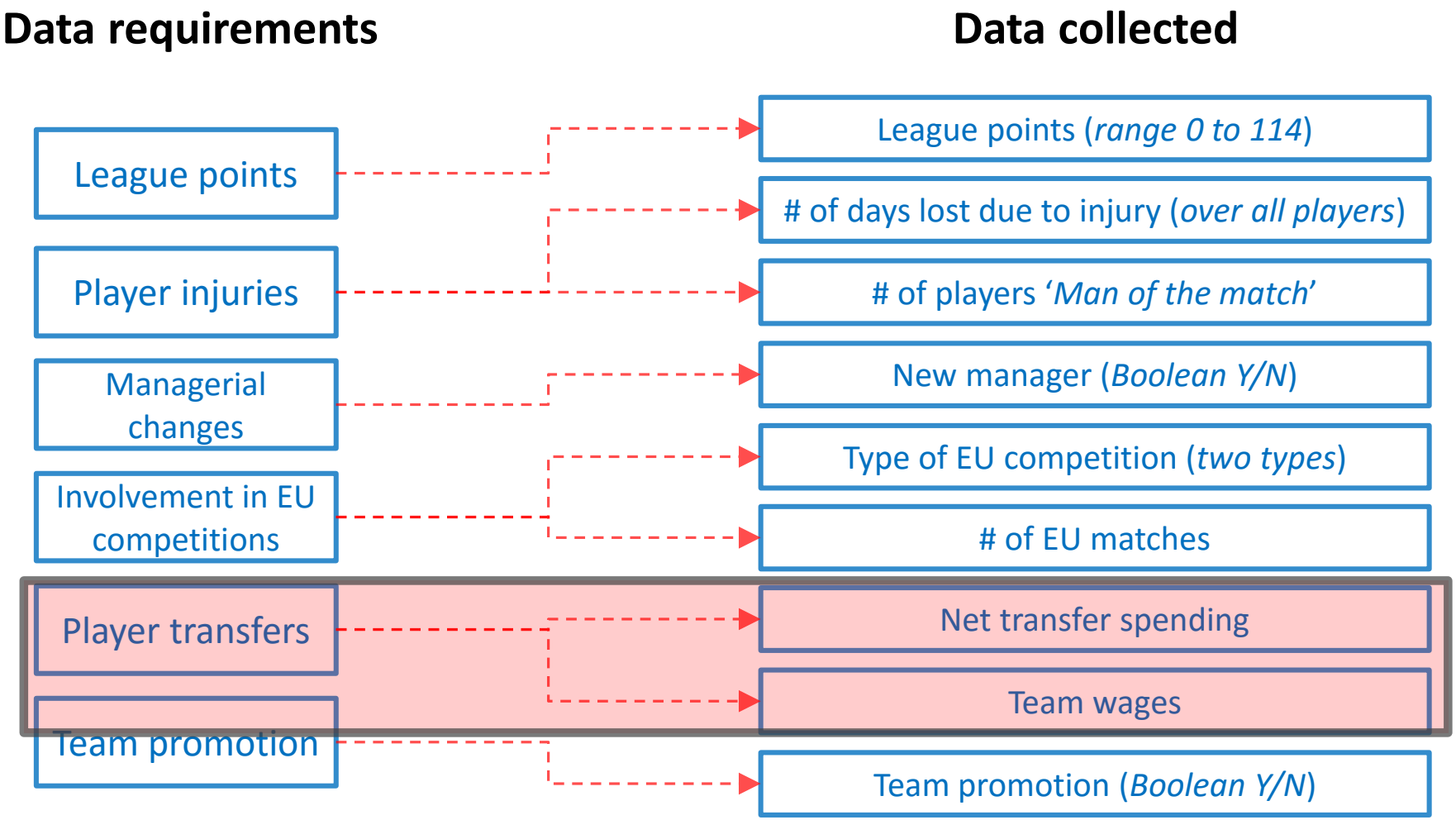
Where:

- t_1 is the previous season;
- t_2 is the summer break;
- t_3 is the next season

Collecting data

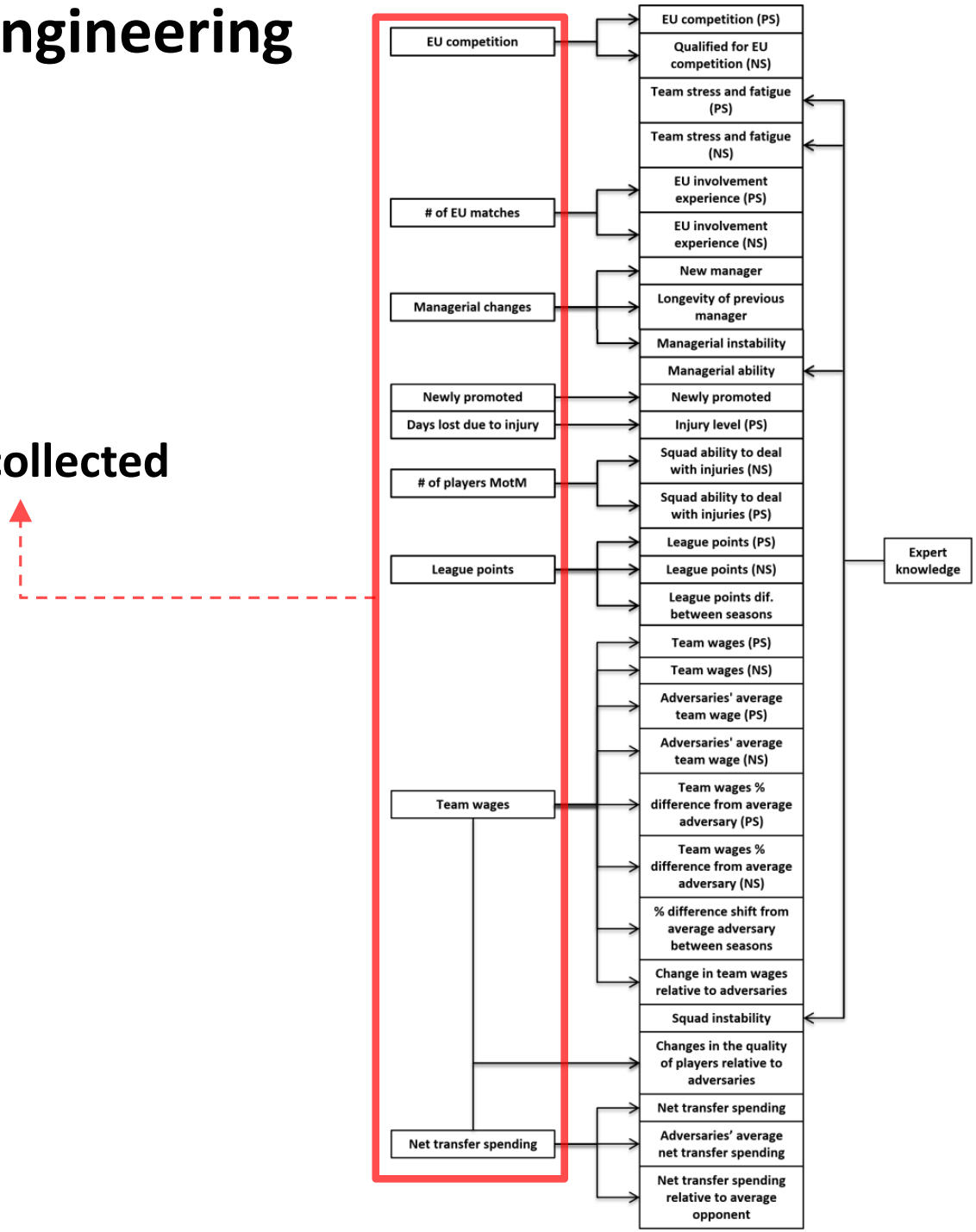


Collecting data



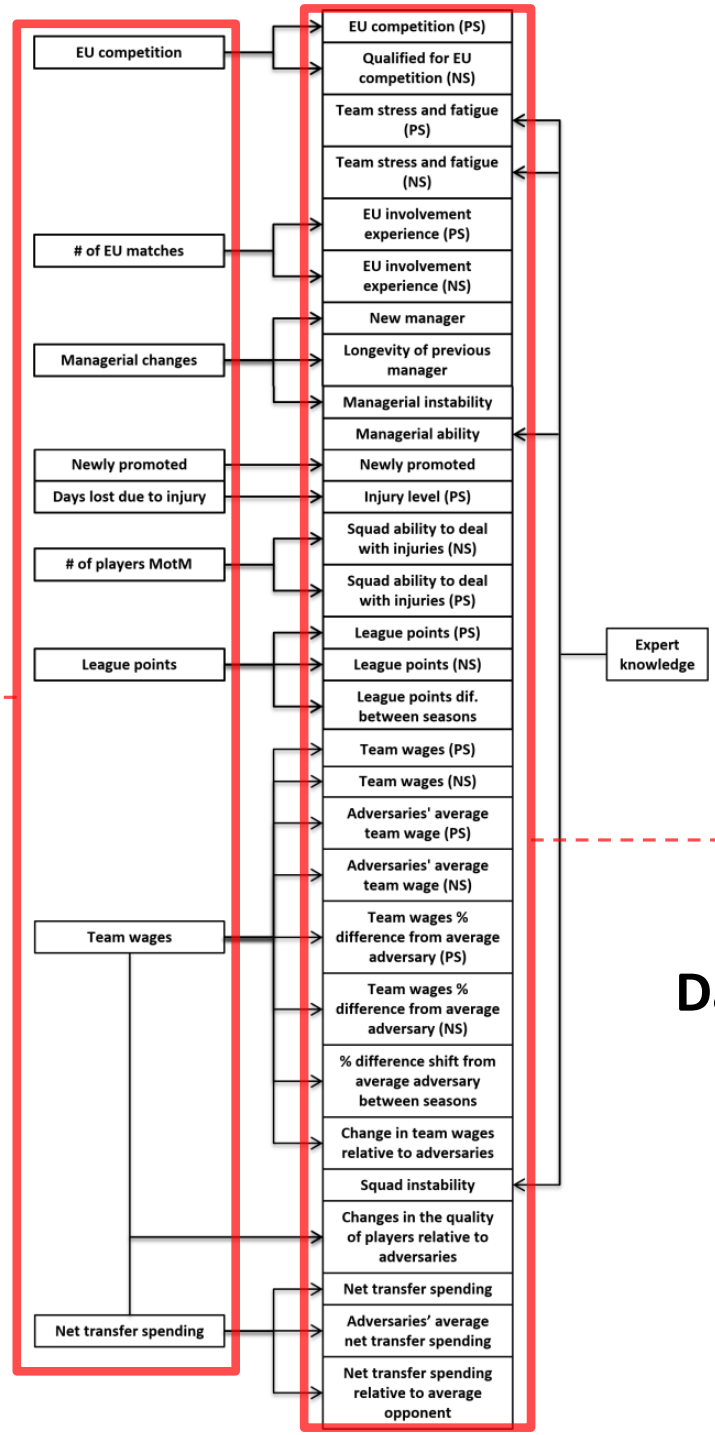
Data engineering

Data collected



Data engineering

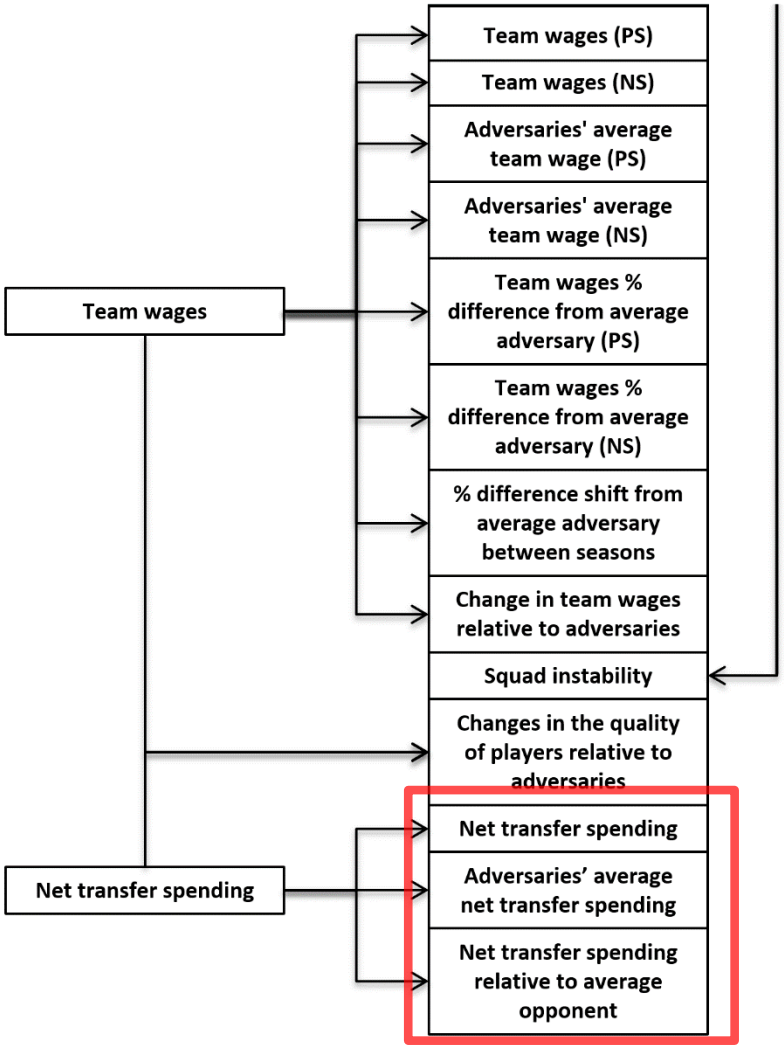
Data collected



Data restructured

Data engineering:

An example of how player transfers data are restructured

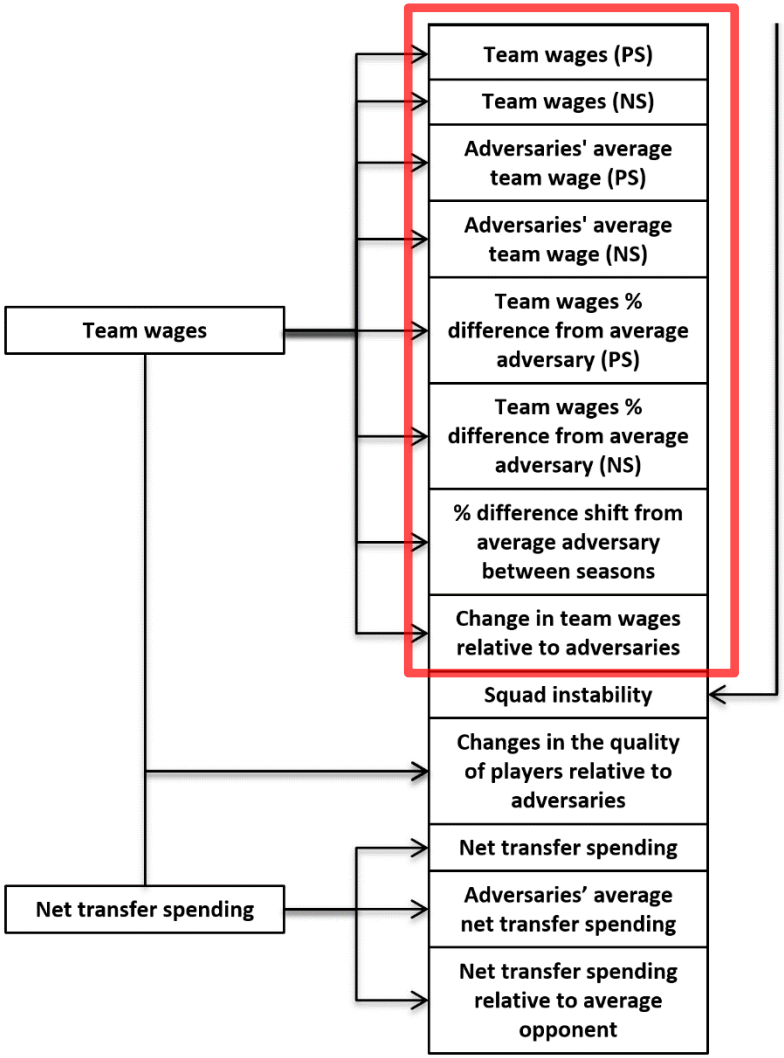


Restructuring the dataset this way, allowed the model to recognize:

- ***Relative additional spend:*** If a team invests $100m to buy new players for the upcoming season, then such a team's performance is expected to improve over the next season. If, however, every other team also spends $100m on new players, then any positive effect is diminished or cancelled.

Data engineering:

An example of how player transfers data are restructured



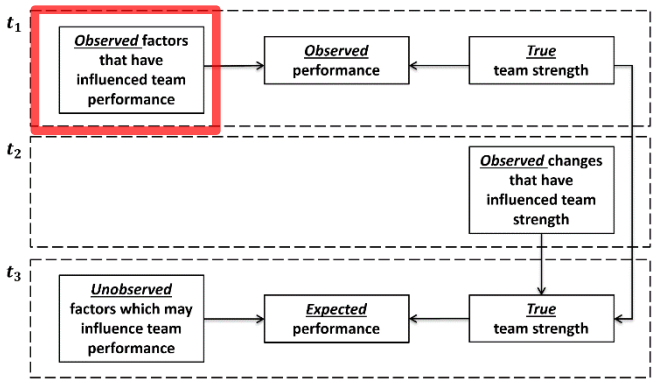
Restructuring the dataset this way, allowed the model to recognize:

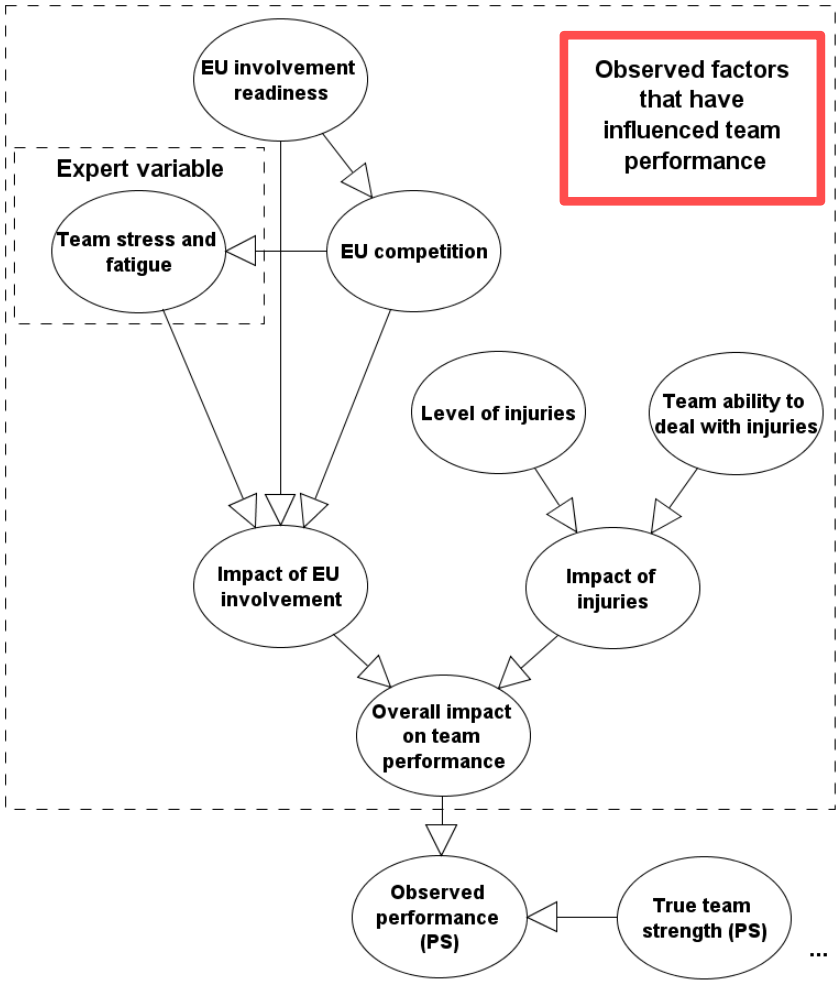
- ***Relative additional spend:*** If a team invests $100m to buy new players for the upcoming season, then such a team's performance is expected to improve over the next season. If, however, every other team also spends $100m on new players, then any positive effect is diminished or cancelled.

- ***Inflation of salaries and player values:*** Investing $100m to buy players during season 2014/15 is not equivalent to investing $100m to buy players during season 2000/01. The same applies to the wage increase of players over the years due to inflation.

The Bayesian network model:

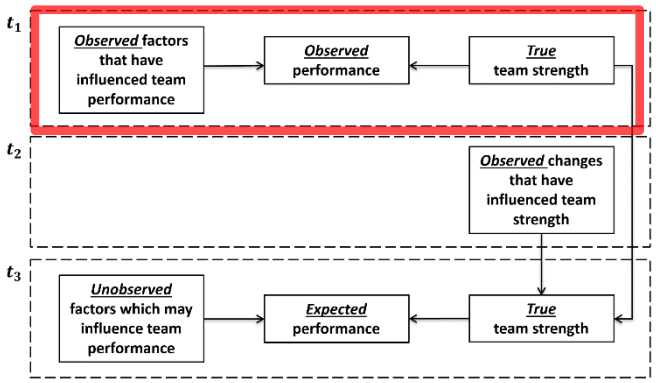
Component t_1

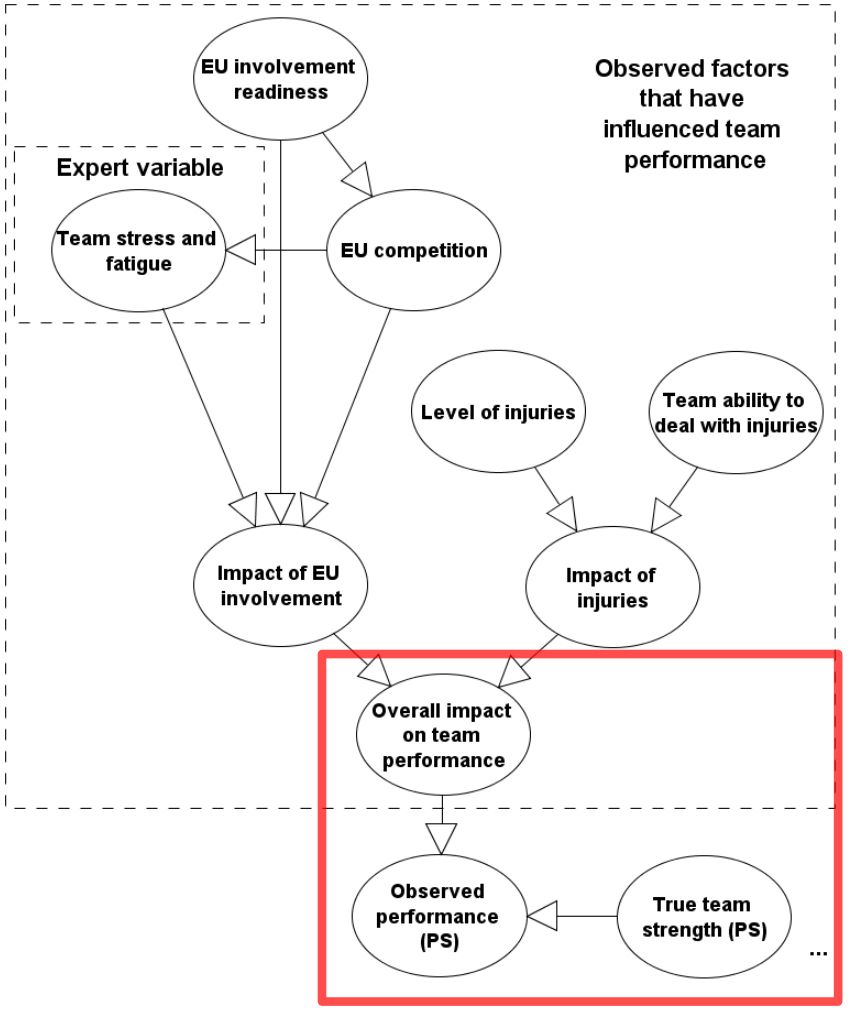




The Bayesian network model:

Component t_1

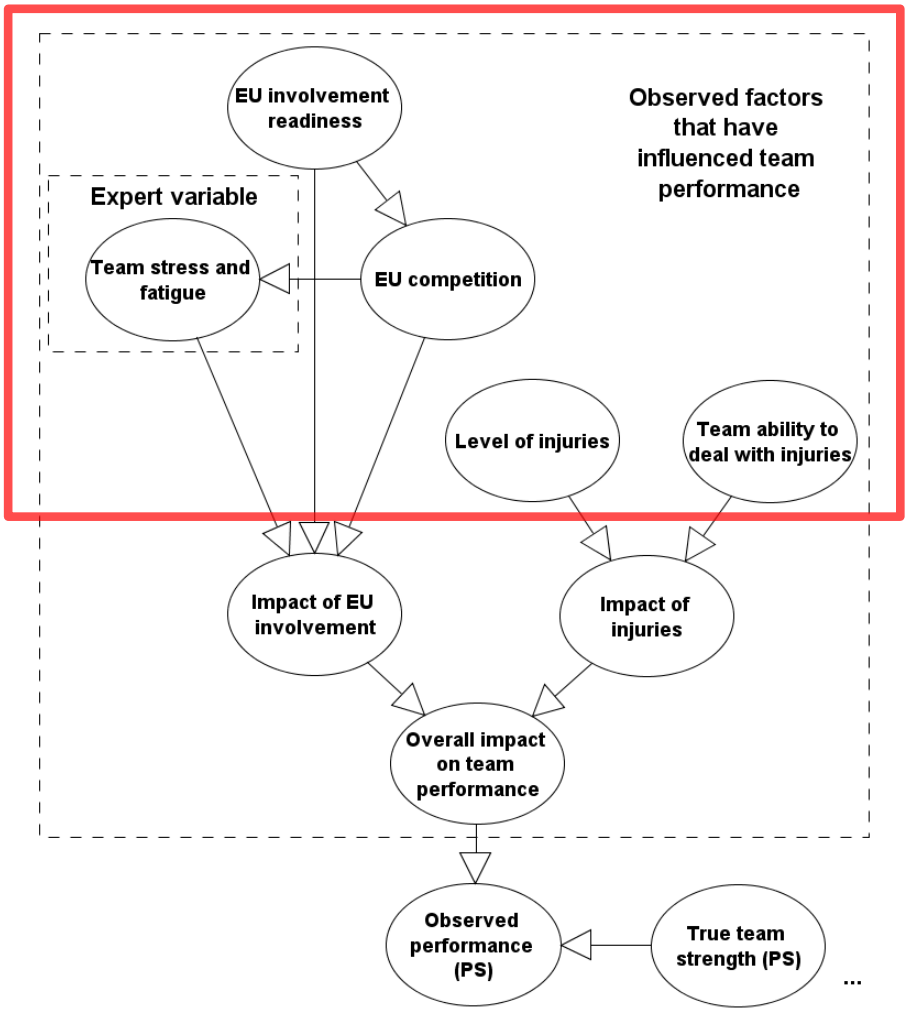




The Bayesian network model:

Component t_1

Discrete variables based on data or knowledge.

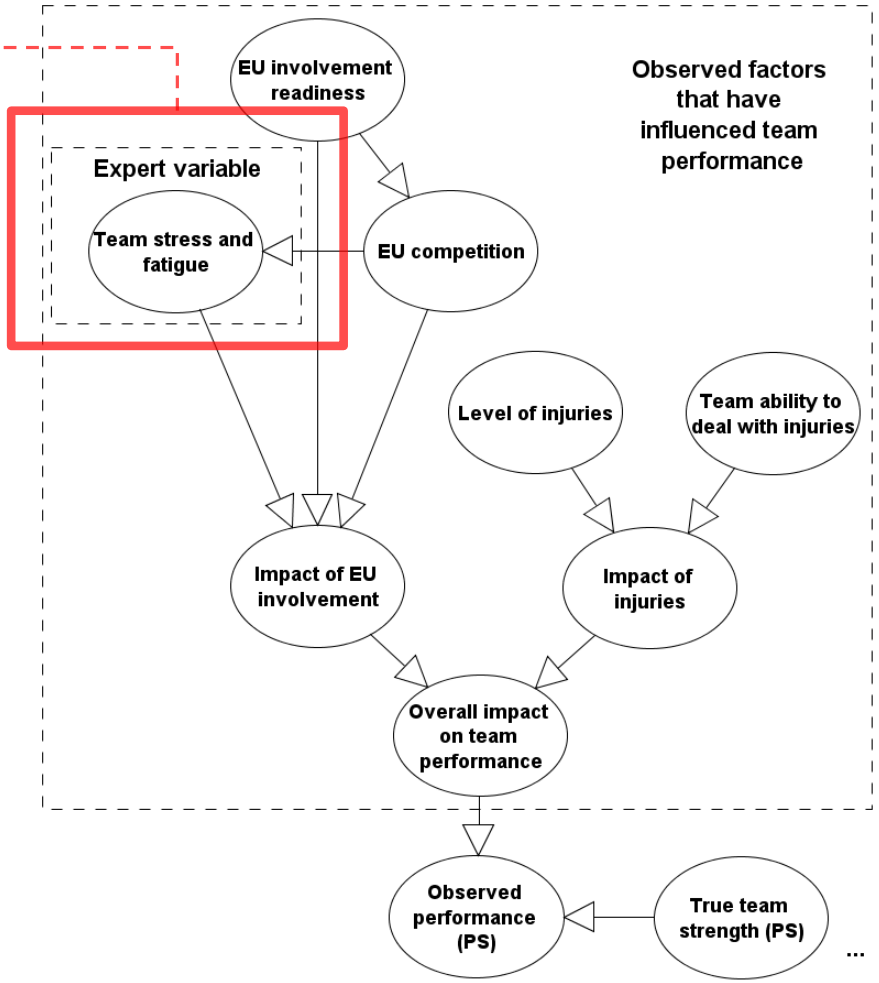


The Bayesian network model:

Component t_1

A few expert variables have been incorporated into the model and:

- do not influence data-driven expectations as long as they remain unobserved, based on the technique of [1];
- Are not taken into consideration for predictive validation;
- Are presented as part of a smart-data approach.



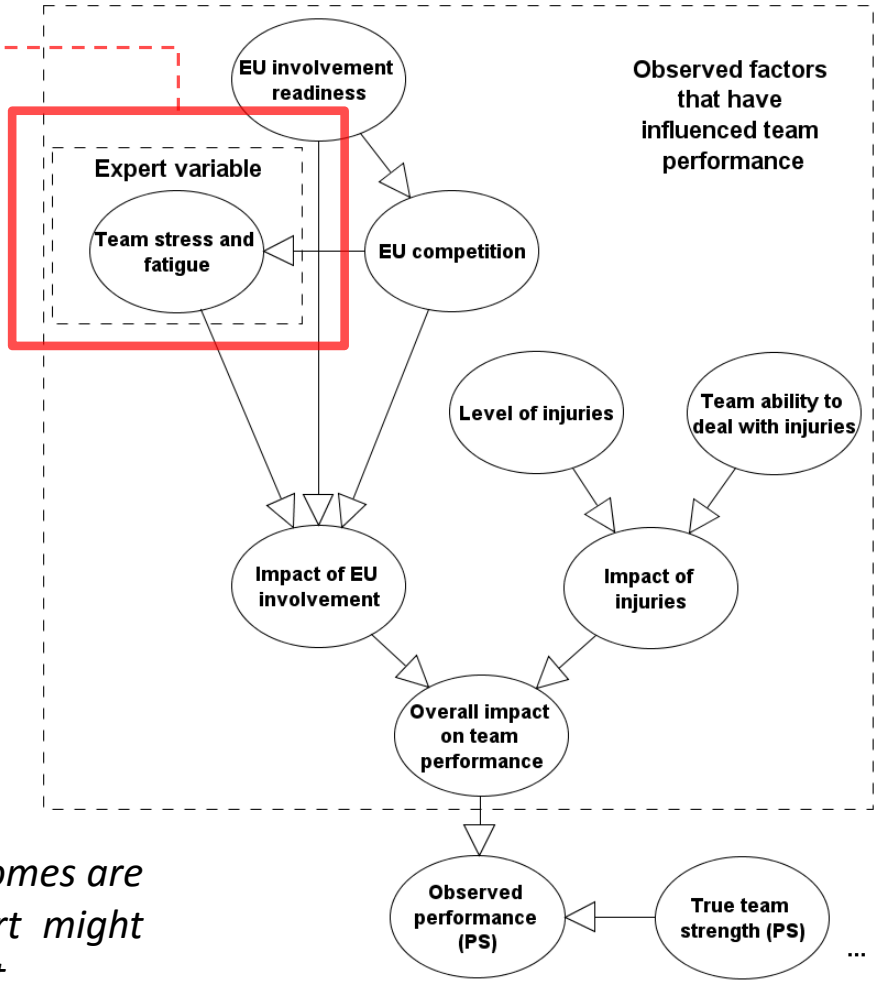
[1] Constantinou, A., Fenton, N., & Neil, M. (2016). Integrating expert knowledge with data in Bayesian networks: Preserving data-driven expectations when the expert variables remain unobserved. *Expert Systems with Applications*, 56: 197-208. [draft, DOI]

The Bayesian network model:

Component t_1

A few expert variables have been incorporated into the model and:

- do not influence data-driven expectations as long as they remain unobserved, based on the technique of [1];
- Are not taken into consideration for predictive validation;
- Are presented as part of a smart-data approach.



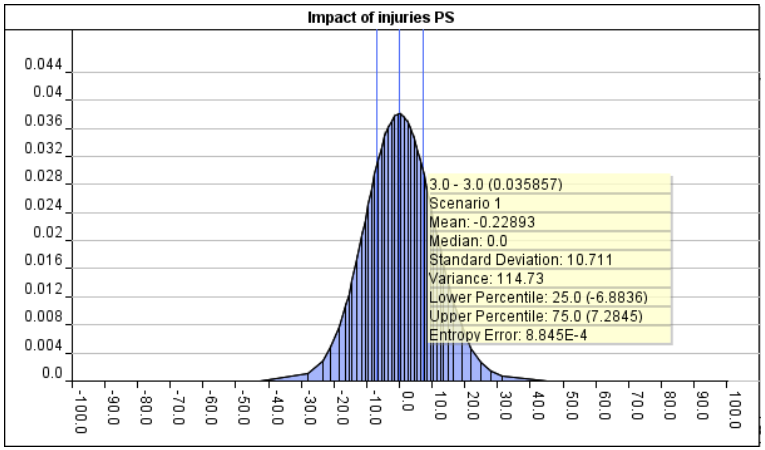
Based on the assumption the statistical outcomes are already influenced by the causes an expert might identify as variables missing from the dataset.

[1] Constantinou, A., Fenton, N., & Neil, M. (2016). Integrating expert knowledge with data in Bayesian networks: Preserving data-driven expectations when the expert variables remain unobserved. Expert Systems with Applications, 56: 197-208. [draft, DOI]

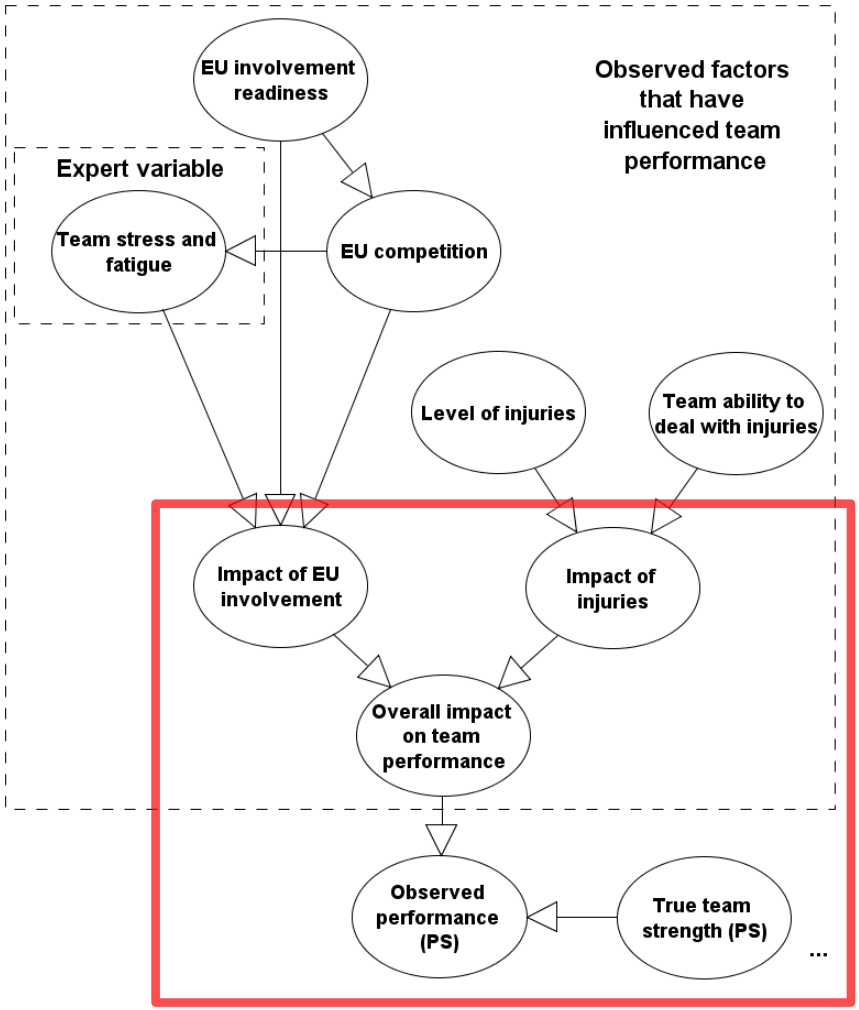
The Bayesian network model:

Component t_1

Normal, or a mixture of *Normal* distributions assessing team performance/strength in terms of league points.



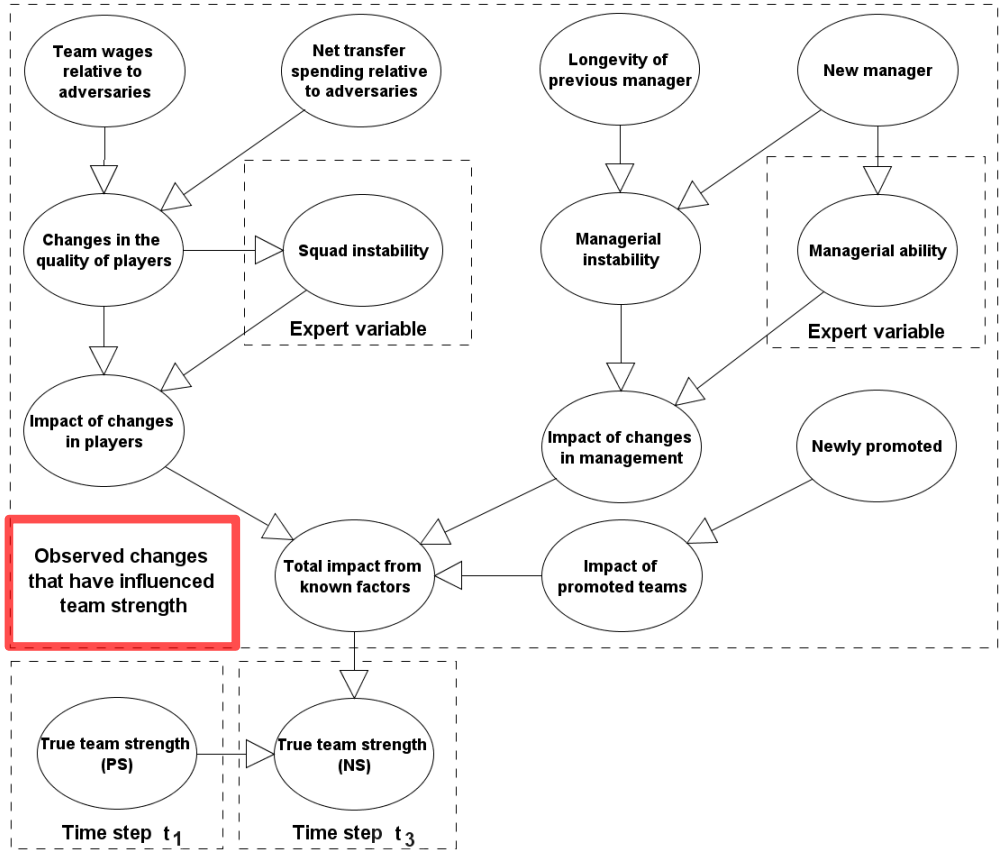
Continuous distributions are approximated with the Dynamic Discretization algorithm [2] implemented in the AgenaRisk BN software.

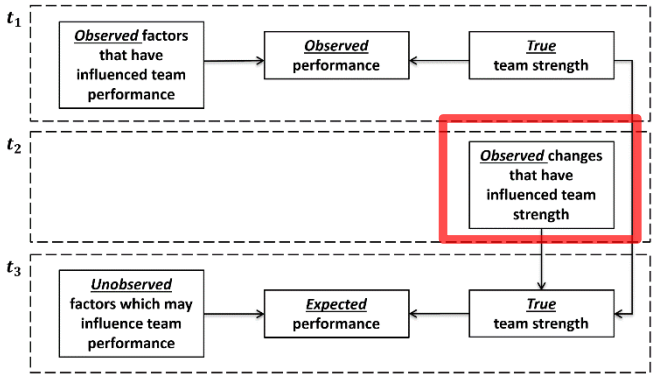


[2] Neil, M., Tailor, M. & Marquez, D. (2007). Inference in hybrid Bayesian networks using dynamic discretization. *Statistics and Computing,* 17, 219-233.

The Bayesian network model:

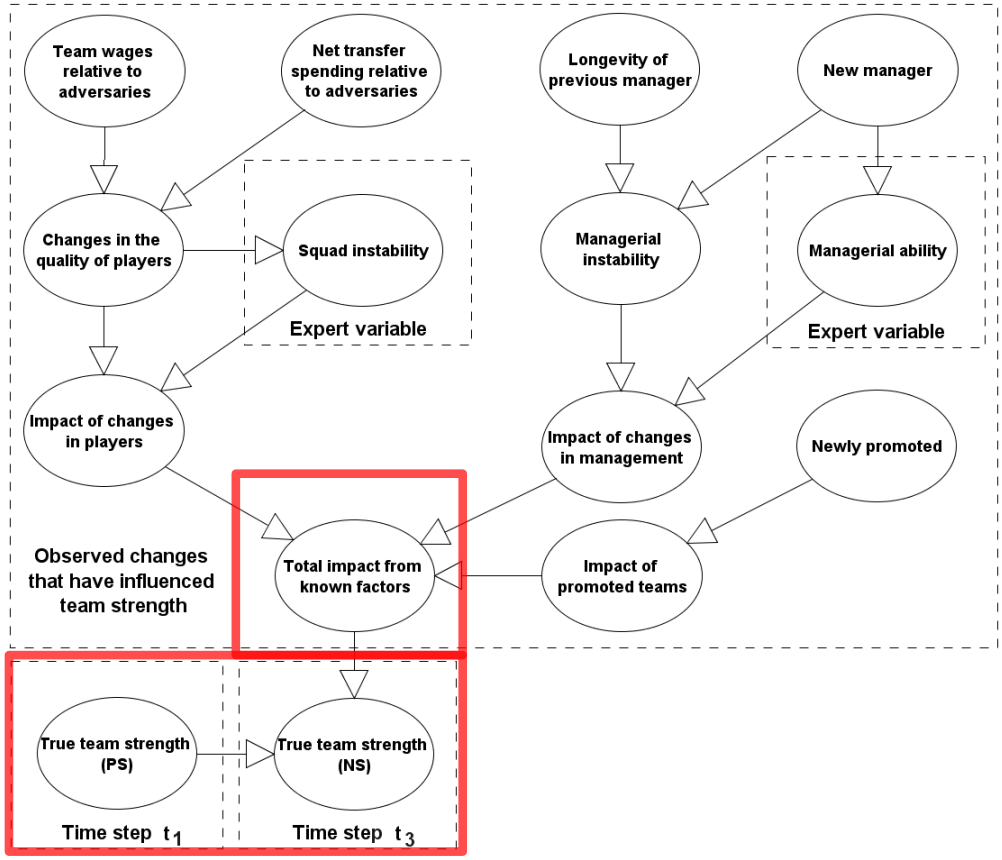
Component t_2

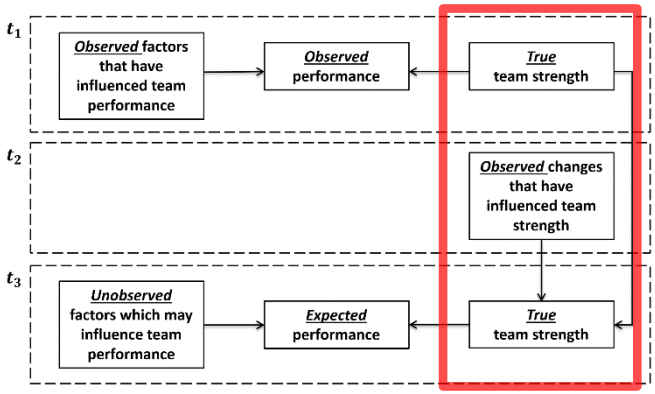




The Bayesian network model:

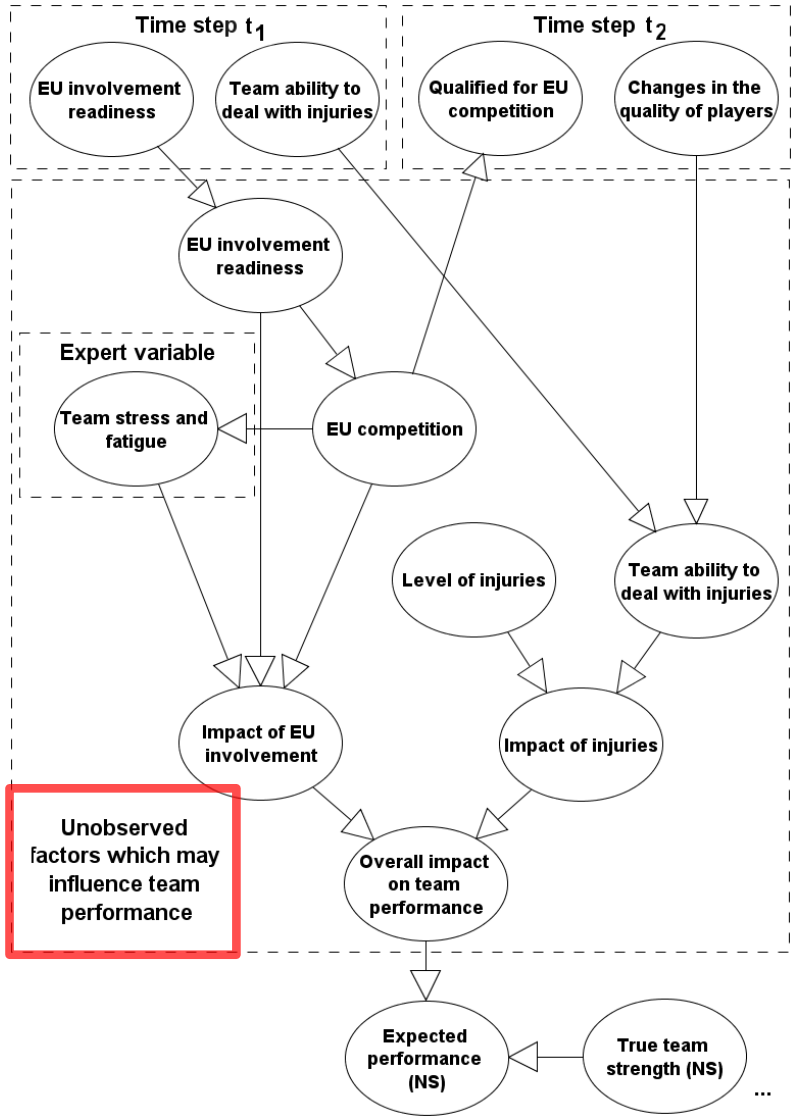
Component t_2

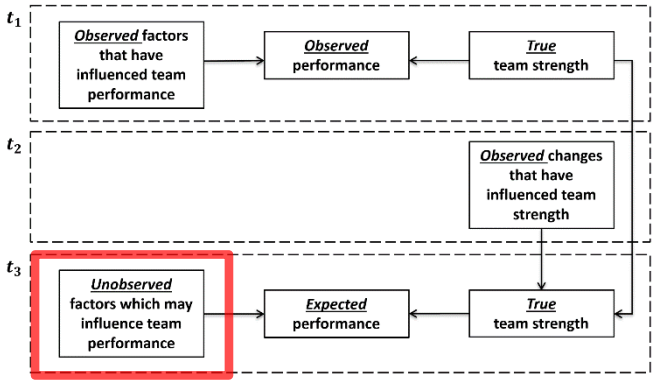




The Bayesian network model:

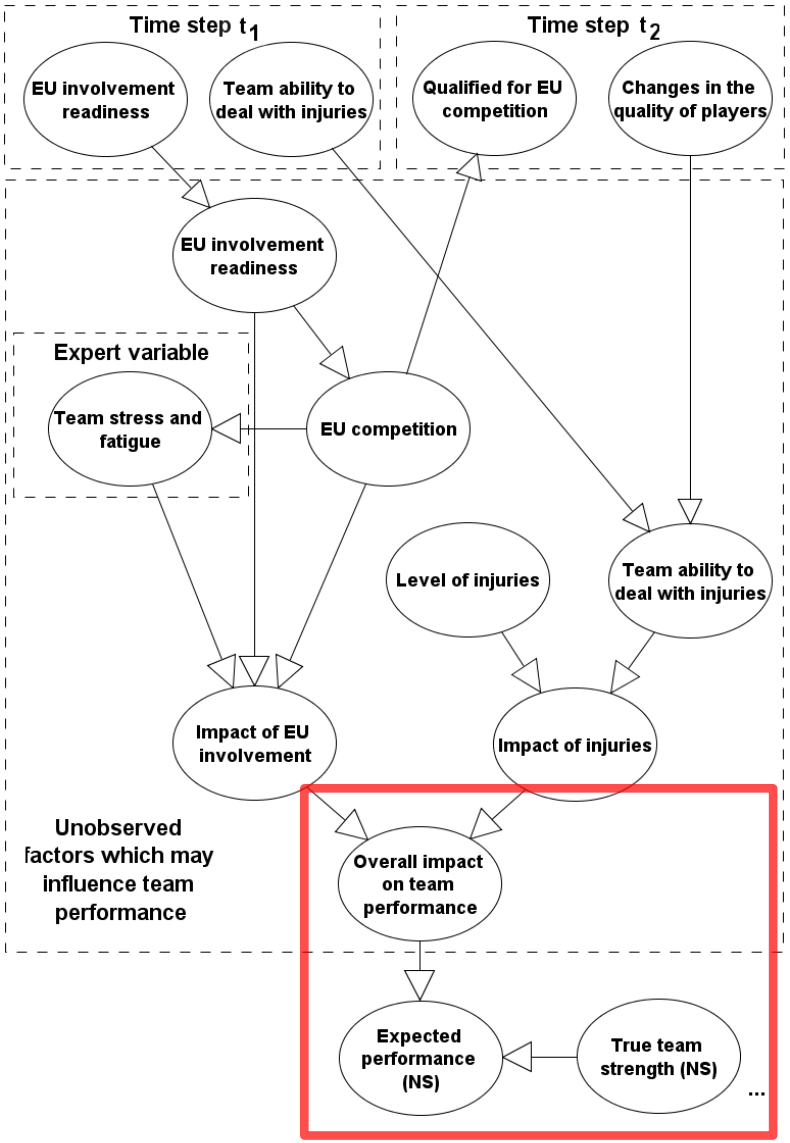
Component t_3

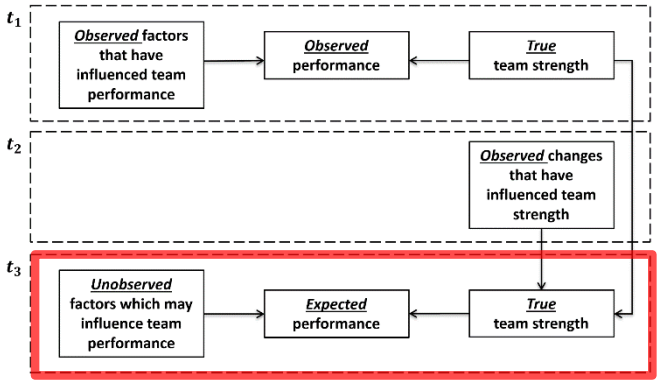




The Bayesian network model:

Component t_3





Results

The three basic ‘methods’ considered for comparison

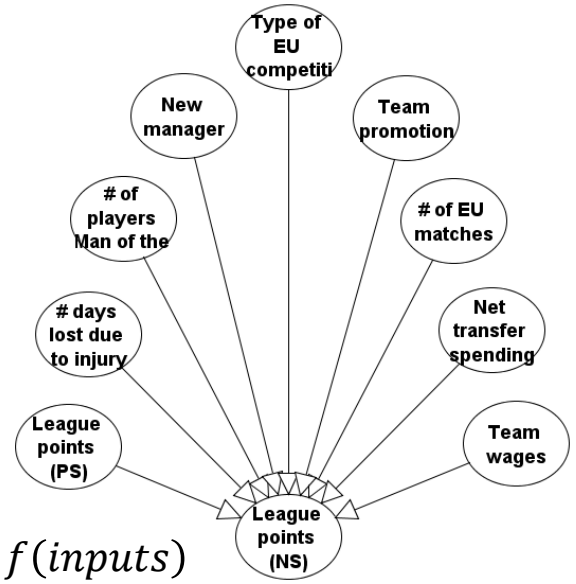
1. **No model (*NM*):** predicts the league points a team will accumulate at season $s + 1$ as the number of league points the team accumulated at season s;

Results

The three basic 'methods' considered for comparison

1. **No model (*NM*):** predicts the league points a team will accumulate at season $s + 1$ as the number of league points the team accumulated at season s;

2. **Regression 1 (*R1*):** Standard linear regression which predicts the points accumulated based on the data which was initially collected (i.e. before data engineering);



$$League\ points = f(inputs)$$

Results

The three basic 'methods' considered for comparison

1. **No model (*NM*):** predicts the league points a team will accumulate at season $s + 1$ as the number of league points the team accumulated at season s;

2. **Regression 1 (*R1*):** Standard linear regression which predicts the points accumulated based on the data which was initially collected (i.e. before data engineering);

3. **Regression 2 (*R2*):** Identical to *R1*, but with financial factors (i.e. team wages and net transfer spending) considered in relative terms and hence, the model predicts the change in points between seasons.

Results

Table 1. Average prediction error, along with standard error, for each model/method in terms of discrepancy between predicted and observed league points accumulated per team, over the 15 seasons (i.e., 300 cases). The range of league points in the EPL is 0 to 114.

Model	Prediction error	Standard error
NM	8.51	±0.3802
R1		
R2		
BN		

Results

Table 1. Average prediction error, along with standard error, for each model/method in terms of discrepancy between predicted and observed league points accumulated per team, over the 15 seasons (i.e., 300 cases). The range of league points in the EPL is 0 to 114.

Model	Prediction error	Standard error
NM	8.51	±0.3802
R1	7.27	±0.7957
R2		
BN		

Results

Table 1. Average prediction error, along with standard error, for each model/method in terms of discrepancy between predicted and observed league points accumulated per team, over the 15 seasons (i.e., 300 cases). The range of league points in the EPL is 0 to 114.

Model	Prediction error	Standard error
NM	8.51	±0.3802
R1	7.27	±0.7957
R2	7.3	±0.3301
BN		

Results

Table 1. Average prediction error, along with standard error, for each model/method in terms of discrepancy between predicted and observed league points accumulated per team, over the 15 seasons (i.e., 300 cases). The range of league points in the EPL is 0 to 114.

Model	Prediction error	Standard error
NM	8.51	±0.3802
R1	7.27	±0.7957
R2	7.3	±0.3301
BN	4.06	±0.1993

Results

Table 2. Time-series validation for teams which have demonstrated the most significant fluctuations in team strength, where *S* is the number of seasons a team participated (out of 15 taken into consideration), and E_{NM}, E_{R1}, E_{R2} and E_{BN} are the respective prediction errors generated for *NM*, *R1*, *R2*, and the BN models respectively.

Team	*S*	E_{NM}	E_{R1}	E_{R2}	E_{BN}
Liverpool	15				
Newcastle	14				
Blackburn	11				
West Ham	12				
Everton	15				
Man City	14				
Average	-				
Error increase (points)	-				

Results

Table 2. Time-series validation for teams which have demonstrated the most significant fluctuations in team strength, where *S* is the number of seasons a team participated (out of 15 taken into consideration), and E_{NM}, E_{R1}, E_{R2} and E_{BN} are the respective prediction errors generated for *NM*, *R1*, *R2*, and the BN models respectively.

Team	*S*	E_{NM}	E_{R1}	E_{R2}	E_{BN}
Liverpool	15	11.53			
Newcastle	14	11.64			
Blackburn	11	11.55			
West Ham	12	11.17			
Everton	15	9.8			
Man City	14	9.43			
Average	**-**	**10.81**			
Error increase (points)	**-**	**2.3**			

Results

Table 2. Time-series validation for teams which have demonstrated the most significant fluctuations in team strength, where *S* is the number of seasons a team participated (out of 15 taken into consideration), and E_{NM}, E_{R1}, E_{R2} and E_{BN} are the respective prediction errors generated for *NM*, *R1*, *R2*, and the BN models respectively.

Team	*S*	E_{NM}	E_{R1}	E_{R2}	E_{BN}
Liverpool	15	11.53	9.24		
Newcastle	14	11.64	10.65		
Blackburn	11	11.55	6.6		
West Ham	12	11.17	7.01		
Everton	15	9.8	9.34		
Man City	14	9.43	8.41		
Average	**-**	**10.81**	**8.73**		
Error increase (points)	**-**	**2.3**	**1.46**		

Results

Table 2. Time-series validation for teams which have demonstrated the most significant fluctuations in team strength, where *S* is the number of seasons a team participated (out of 15 taken into consideration), and E_{NM}, E_{R1}, E_{R2} and E_{BN} are the respective prediction errors generated for *NM*, *R1*, *R2*, and the BN models respectively.

Team	*S*	E_{NM}	E_{R1}	E_{R2}	E_{BN}
Liverpool	15	11.53	9.24	10.67	
Newcastle	14	11.64	10.65	9.22	
Blackburn	11	11.55	6.6	8.14	
West Ham	12	11.17	7.01	8.03	
Everton	15	9.8	9.34	9.66	
Man City	14	9.43	8.41	7.05	
Average	**-**	**10.81**	**8.73**	**8.69**	
Error increase (points)	**-**	**2.3**	**1.46**	**1.39**	

Results

Table 2. Time-series validation for teams which have demonstrated the most significant fluctuations in team strength, where *S* is the number of seasons a team participated (out of 15 taken into consideration), and E_{NM}, E_{R1}, E_{R2} and E_{BN} are the respective prediction errors generated for *NM, R1*, *R2*, and the BN models respectively.

Team	*S*	E_{NM}	E_{R1}	E_{R2}	E_{BN}
Liverpool	15	11.53	9.24	10.67	5.61
Newcastle	14	11.64	10.65	9.22	4.48
Blackburn	11	11.55	6.6	8.14	3.46
West Ham	12	11.17	7.01	8.03	3.41
Everton	15	9.8	9.34	9.66	3.65
Man City	14	9.43	8.41	7.05	4.64
Average	**-**	**10.81**	**8.73**	**8.69**	**4.27**
Error increase (points)	**-**	**2.3**	**1.46**	**1.39**	**0.21**

Results

Table 3: Model factors of interest and their impact on team performance, where *P* is the expected discrepancy in league points accumulated for the average subsequent season.

Factor/s	P
P(Net transfer spending…="*Much higher*"), and P(Team wages…="*Extreme increase*")	+8.49
P(Newly promoted="*Yes*")	+8.34
P(EU competition="*No*"), and P(EU readiness="*High*")	+5.17
P(Injury level="*High*"), and P(Squad ability to deal with injuries="*Low*")	-8.31
P(EU competition="*Both*"), and P(EU readiness="*No/Low*")	-16.52

Results

Table 3: Model factors of interest and their impact on team performance, where *P* is the expected discrepancy in league points accumulated for the average subsequent season.

Factor/s	P
P(Net transfer spending…="*Much higher*"), and P(Team wages…="*Extreme increase*")	+8.49
P(Newly promoted="*Yes*")	+8.34
P(EU competition="*No*"), and P(EU readiness="*High*")	+5.17
P(Injury level="*High*"), and P(Squad ability to deal with injuries="*Low*")	-8.31
P(EU competition="*Both*"), and P(EU readiness="*No/Low*")	-16.52

Conclusions and implications:

Application domain

1. First study to present a soccer model for time-series forecasting in terms of how the strength of soccer teams evolves over adjacent soccer seasons, without the need to generate predictions for individual matches.

Conclusions and implications:
Application domain

1. First study to present a soccer model for time-series forecasting in terms of how the strength of soccer teams evolves over adjacent soccer seasons, without the need to generate predictions for individual matches.

2. Previously published match-by-match prediction models which fail to account for the external factors influencing team strength, are prone to an error of 8.51 league points accumulated per team between seasons (assuming EPL league).

Conclusions and implications:
Application domain

1. First study to present a soccer model for time-series forecasting in terms of how the strength of soccer teams evolves over adjacent soccer seasons, without the need to generate predictions for individual matches.

2. Previously published match-by-match prediction models which fail to account for the external factors influencing team strength, are prone to an error of 8.51 league points accumulated per team between seasons (assuming EPL league).

3. Studies which assess the efficiency of the soccer gambling market may find the BN model helpful in the sense that it could help in explaining previously unexplained fluctuations in gambling market odds.

Conclusions and implications:

Smart-Data

1. Further evidence that seeking ‘bigger’ data is not always the path to follow. The model presented in this study is based on just 300 data instances.

Conclusions and implications:

Smart-Data

1. Further evidence that seeking ‘bigger’ data is not always the path to follow. The model presented in this study is based on just 300 data instances.

2. Standard non-linear statistical regression models, which are still the preferred method for real-world prediction in many areas of social and medical sciences, failed to achieve predictive accuracy similar to the smart-data BN model.

Conclusions and implications:

Smart-Data

1. Further evidence that seeking ‘bigger’ data is not always the path to follow. The model presented in this study is based on just 300 data instances.

2. Standard non-linear statistical regression models, which are still the preferred method for real-world prediction in many areas of social and medical sciences, failed to achieve predictive accuracy similar to the smart-data BN model.

3. The paper supports the development of a smart-data method which aims to improve the quality, as opposed to the quantity, of a dataset driven by model requirements.

Conclusions and implications:
Smart-Data

1. Further evidence that seeking ‘bigger’ data is not always the path to follow. The model presented in this study is based on just 300 data instances.

2. Standard non-linear statistical regression models, which are still the preferred method for real-world prediction in many areas of social and medical sciences, failed to achieve predictive accuracy similar to the smart-data BN model.

3. The paper supports the development of a smart-data method which aims to improve the quality, as opposed to the quantity, of a dataset driven by model requirements.

4. Attempted to highlight the importance of developing models based on what data we really require for inference, rather than based on what (big) data are available.

Conclusions and implications:

Smart-Data

1. Further evidence that seeking ‘bigger’ data is not always the path to follow. The model presented in this study is based on just 300 data instances.

2. Standard non-linear statistical regression models, which are still the preferred method for real-world prediction in many areas of social and medical sciences, failed to achieve predictive accuracy similar to the smart-data BN model.

3. The paper supports the development of a smart-data method which aims to improve the quality, as opposed to the quantity, of a dataset driven by model requirements.

4. Attempted to highlight the importance of developing models based on what data we really require for inference, rather than based on what (big) data are available.

5. Demonstrated that inferring knowledge from data imposes further challenges and requires skills that merge the quantitative as well as the qualitative aspects of data.

Conclusions and implications:
Smart-Data

1. Further evidence that seeking ‘bigger’ data is not always the path to follow. The model presented in this study is based on just 300 data instances.

2. Standard non-linear statistical regression models, which are still the preferred method for real-world prediction in many areas of social and medical sciences, failed to achieve predictive accuracy similar to the smart-data BN model.

3. The paper supports the development of a smart-data method which aims to improve the quality, as opposed to the quantity, of a dataset driven by model requirements.

4. Attempted to highlight the importance of developing models based on what data we really require for inference, rather than based on what (big) data are available.

5. Demonstrated that inferring knowledge from data imposes further challenges and requires skills that merge the quantitative as well as the qualitative aspects of data.

6. Invites examination of the impact of a smart-data method on processes of causal discovery.

Thank you

This study was part of the "Effective Bayesian Modelling with Knowledge Before Data (BAYES-KNOWLEDGE)", funded by the European Research Council (ERC), Grant reference number ERC-2013-AdG339182-BAYES_KNOWLEDGE. We also acknowledge Agena Ltd for Bayesian Network software support.

Thank you for listening.

…any questions?

Queen Mary
University of London