

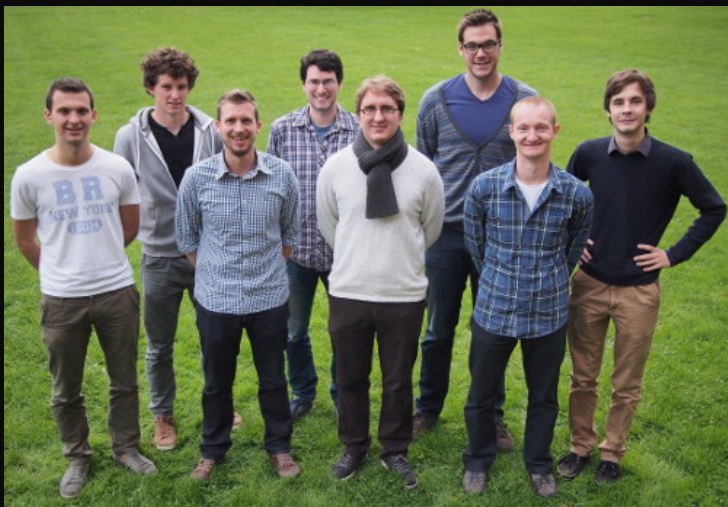


Mining soccer data

Vladimir Dzyuba

DTAI Seminar, KU Leuven

March 21, 2016



Analysing tracking data requires reasoning about relations

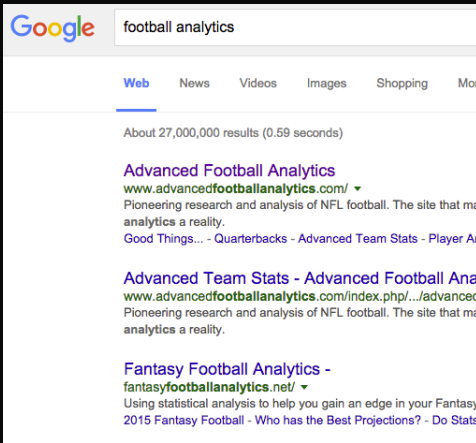
Inductive logic programming is a convenient tool for tackling it, but not the most efficient

Let's start with the most pressing issue...

Football vs. soccer: Google

3/44

"Football analytics" dominated by American football



Google

football analytics

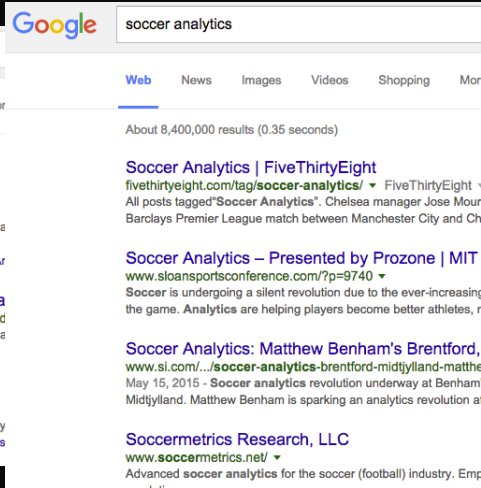
Web News Videos Images Shopping More

About 27,000,000 results (0.59 seconds)

Advanced Football Analytics
www.advancedfootballanalytics.com/ ▼
Pioneering research and analysis of NFL football. The site that ma
analytics a reality.
Good Things... - Quarterbacks - Advanced Team Stats - Player Ar

Advanced Team Stats - Advanced Football Ana
www.advancedfootballanalytics.com/index.php/.../advanced
Pioneering research and analysis of NFL football. The site that ma
analytics a reality.

Fantasy Football Analytics -
fantasyfootballanalytics.net/ ▼
Using statistical analysis to help you gain an edge in your Fantasy
2015 Fantasy Football - Who has the Best Projections? - Do Stats



Google

soccer analytics

Web News Images Videos Shopping More

About 8,400,000 results (0.35 seconds)

Soccer Analytics | FiveThirtyEight
fivethirtyeight.com/tag/soccer-analytics/ ▼ FiveThirtyEight ▼
All posts tagged "Soccer Analytics". Chelsea manager Jose Mour
Barclays Premier League match between Manchester City and Ch

Soccer Analytics – Presented by Prozone | MIT
www.sloansportsconference.com/?p=9740 ▼
Soccer is undergoing a silent revolution due to the ever-increasing
the game. Analytics are helping players become better athletes, r

Soccer Analytics: Matthew Benham's Brentford,
www.si.com/.../soccer-analytics-brentford-midtylland-matthe
May 15, 2015 - Soccer analytics revolution underway at Benham
Midtylland. Matthew Benham is sparking an analytics revolution at

Soccermetrics Research, LLC
www.soccermetrics.net/ ▼
Advanced soccer analytics for the soccer (football) industry. Emp
revolution

Two key drivers:

- ▶ Availability of data
- ▶ Abundance of high-profile success stories
Oakland A's (Moneyball), FC Midtjylland, etc.

Ultimate goal:

Use data analysis to improve player/team performance

Advanced data collection: Major driver 5/44

High-frequency tracking data: Coordinates every 0.1s

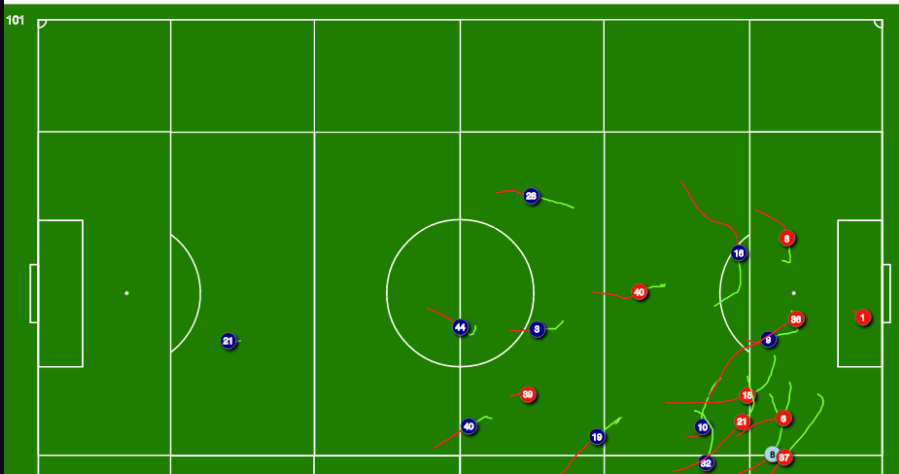


Image: Prozone

Event streams: Manually annotated video feeds

▶ || Play: 54 Tracer: 20 Speed: 5X Timer: 101 Max: 199

101



Dataset: Games of professional Belgian team 7/44

70 games in three competitions

59 - league, 2 - cup, 9 - Europa league

2600 events per game

40+ event types, including *pass*, *run*, *receive*, *clear*...

Each event has a number of attributes

Most importantly, *time*, *coordinates*, and *player(s)*

Motivation: Tactical patterns that create shots 8/44

Goals are rare and subject to luck

Shots are more frequent and correlate to long-term success

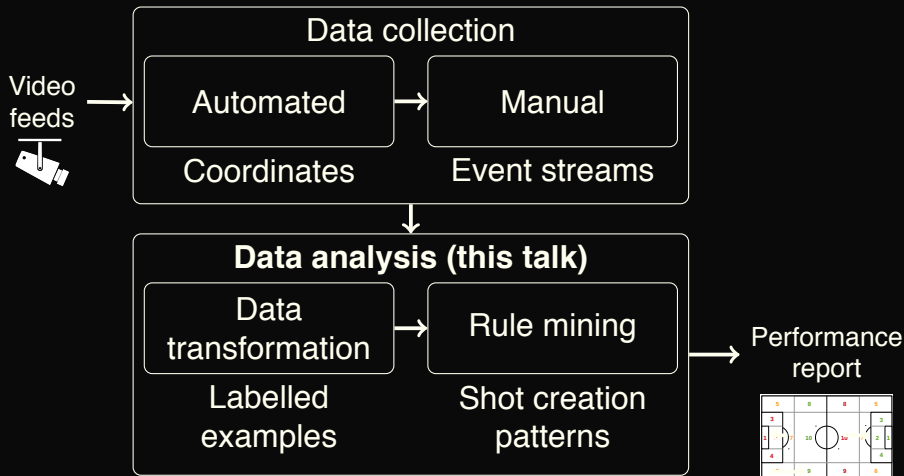
Team creating shots will eventually score

Given Game phases, labelled as *positive*,
i.e. resulted in a shot, or *negative*

Identify Features of phases
that make shots more likely

Sounds like

Rule learning, supervised pattern mining, subgroup discovery...



Outline

- Data modelling & analysis
- Experimental results
- Further extensions
- Conclusions

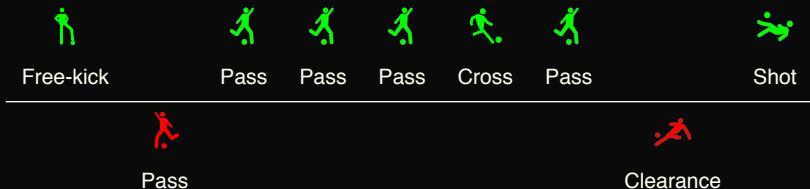
Constructing examples: Game phases

11/44

Game phases are event sequences, as determined by annotators:

Start *Throw-in, goal kick, or free kick*

End *Ball goes out of play or foul committed*



Only consider phases where *our team* is dominant

Most events in the phase are by its players

Icons: Freepik

Phase is a sequence of events

12/44

Phase label: ?

Time	Player	Team	Event	X,Y
47:35	p13	Us	 Free kick	(-1470, 2570)
47:36	p21	Them	 Reception	(-3200, 2630)
47:38	p21	Them	 Pass	(-3200, 2630)
47:40	p1	Us	 Pass	(540, 510)
47:42	p6	Us	 Pass	(-270, -1370)
47:45	p4	Us	 Pass	(-820, -2640)
47:46	p6	Us	 Reception	(-1830, -2630)
47:47	p6	Us	 Running with ball	(-1830, -2630)
47:48	p6	Us	 Running with ball	(-1830, -2630)
47:49	p6	Us	 Running with ball	(-1830, -2630)
47:51	p6	Us	 Cross	(-1830, -2630)
47:51	p26	Them	 Reception	(-4330, -770)
47:53	p6	Us	 Pass	(-4470, -2190)
47:55	p26	Them	 Clearance	(-4270, -890)
47:56	p8	Us	 Reception	(-3350, -1640)
47:57	p8	Us	 Shot not on target	(-3390, -1820)

Step 1: Only keep shots and ball movement

13/44

Phase label: ?

Time	Player	Team	Event	X,Y
47:35	p13	Us	 Free kick	(-1470, 2570)
47:36	p21	Them	 Reception	(-3200, 2630)
47:38	p21	Them	 Pass	(-3200, 2630)
47:40	p1	Us	 Pass	(540, 510)
47:42	p6	Us	 Pass	(-270, -1370)
47:45	p4	Us	 Pass	(-820, -2640)
47:46	p6	Us	 Reception	(-1830, -2630)
47:47	p6	Us	 Running with ball	(-1830, -2630)
47:48	p6	Us	 Running with ball	(-1830, -2630)
47:49	p6	Us	 Running with ball	(-1830, -2630)
47:51	p6	Us	 Cross	(-1830, -2630)
47:51	p26	Them	 Reception	(-4330, -770)
47:53	p6	Us	 Pass	(-4470, -2190)
47:55	p26	Them	 Clearance	(-4270, -890)
47:56	p8	Us	 Reception	(-3350, -1640)
47:57	p8	Us	 Shot not on target	(-3390, -1820)

Step 1: Only keep shots and ball movement 13/44

Phase label: ?

Time	Player	Team	Event	X,Y
47:35	p13	Us	 Free kick	(-1470, 2570)
47:36	p21	Them	 Reception	(-3200, 2630)
47:38	p21	Them	 Pass	(-3200, 2630)
47:40	p1	Us	 Pass	(540, 510)
47:42	p6	Us	 Pass	(-270, -1370)
47:45	p4	Us	 Pass	(-820, -2640)
47:46	p6	Us	 Reception	(-1830, -2630)
47:51	p6	Us	 Cross	(-1830, -2630)
47:51	p26	Them	 Reception	(-4330, -770)
47:53	p6	Us	 Pass	(-4470, -2190)
47:55	p26	Them	 Clearance	(-4270, -890)
47:56	p8	Us	 Reception	(-3350, -1640)
47:57	p8	Us	 Shot not on target	(-3390, -1820)

Step 2: Discard timestamps and opposing players

14/44

Phase label: ?

Time	Player	Team	Event	X,Y
47:35	p13	Us	 Free kick	(-1470, 2570)
47:36	p21	Them	 Reception	(-3200, 2630)
47:38	p21	Them	 Pass	(-3200, 2630)
47:40	p1	Us	 Pass	(540, 510)
47:42	p6	Us	 Pass	(-270, -1370)
47:45	p4	Us	 Pass	(-820, -2640)
47:46	p6	Us	 Reception	(-1830, -2630)
47:51	p6	Us	 Cross	(-1830, -2630)
47:51	p26	Them	 Reception	(-4330, -770)
47:53	p6	Us	 Pass	(-4470, -2190)
47:55	p26	Them	 Clearance	(-4270, -890)
47:56	p8	Us	 Reception	(-3350, -1640)
47:57	p8	Us	 Shot not on target	(-3390, -1820)


Phase label: ?

Player	Event	X,Y
p13	 Free kick	(-1470, 2570)
Opponent	 Reception	(-3200, 2630)
Opponent	 Pass	(-3200, 2630)
p1	 Pass	(540, 510)
p6	 Pass	(-270, -1370)
p4	 Pass	(-820, -2640)
p6	 Reception	(-1830, -2630)
p6	 Cross	(-1830, -2630)
Opponent	 Reception	(-4330, -770)
p6	 Pass	(-4470, -2190)
Opponent	 Clearance	(-4270, -890)
p8	 Reception	(-3350, -1640)
p8	 Shot not on target	(-3390, -1820)

Step 3: Label phases based on shot occurrences

15/44



Phase label: ?

Player	Event	X,Y
p13	 Free kick	(-1470, 2570)
Opponent	 Reception	(-3200, 2630)
Opponent	 Pass	(-3200, 2630)
p1	 Pass	(540, 510)
p6	 Pass	(-270, -1370)
p4	 Pass	(-820, -2640)
p6	 Reception	(-1830, -2630)
p6	 Cross	(-1830, -2630)
Opponent	 Reception	(-4330, -770)
p6	 Pass	(-4470, -2190)
Opponent	 Clearance	(-4270, -890)
p8	 Reception	(-3350, -1640)
p8	 Shot not on target	(-3390, -1820)

Step 3: Label phases based on shot occurrences

15/44

Phase label: \oplus

Player	Event	X,Y
p13	 Free kick	(-1470, 2570)
Opponent	 Reception	(-3200, 2630)
Opponent	 Pass	(-3200, 2630)
p1	 Pass	(540, 510)
p6	 Pass	(-270, -1370)
p4	 Pass	(-820, -2640)
p6	 Reception	(-1830, -2630)
p6	 Cross	(-1830, -2630)
Opponent	 Reception	(-4330, -770)
p6	 Pass	(-4470, -2190)
Opponent	 Clearance	(-4270, -890)
p8	 Reception	(-3350, -1640)

Step 4: Match passes with receptions

16/44










Phase label: \oplus

Player	Event	X,Y
p13	 Free kick	(-1470, 2570)
Opponent	 Reception	(-3200, 2630)
Opponent	 Pass	(-3200, 2630)
p1	 Pass	(540, 510)
p6	 Pass	(-270, -1370)
p4	 Pass	(-820, -2640)
p6	 Reception	(-1830, -2630)
p6	 Cross	(-1830, -2630)
Opponent	 Reception	(-4330, -770)
p6	 Pass	(-4470, -2190)
Opponent	 Clearance	(-4270, -890)
p8	 Reception	(-3350, -1640)

Step 4: Match passes with receptions

16/44

Phase label: \oplus

Event		From		To	
 Free kick	p13	(-1470, 2570)	Opponent	(-3200, 2630)	
 Pass	Opponent	(-3200, 2630)	p1	(540, 510)	
 Pass	p1	(540, 510)	p6	(-270, -1370)	
 Pass	p6	(-270, -1370)	p4	(-820, -2640)	
 Pass	p4	(-820, -2640)	p6	(-1830, -2630)	
 Cross	p6	(-1830, -2630)	Opponent	(-4330, -770)	
 Pass	Opponent	(-4330, -770)	p6	(-4470, -2190)	
 Pass	p6	(-4470, -2190)	Opponent	(-4270, -890)	
 Pass	Opponent	(-4270, -890)	p8	(-3350, -1640)	

Hard to represent as fixed-length feature vector

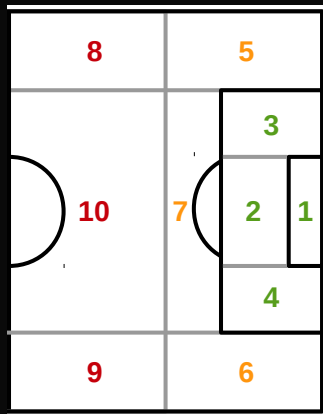
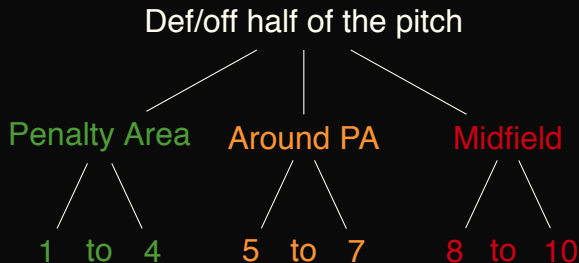
Solution: *Inductive logic programming*

Raw data is overly fine-grained:

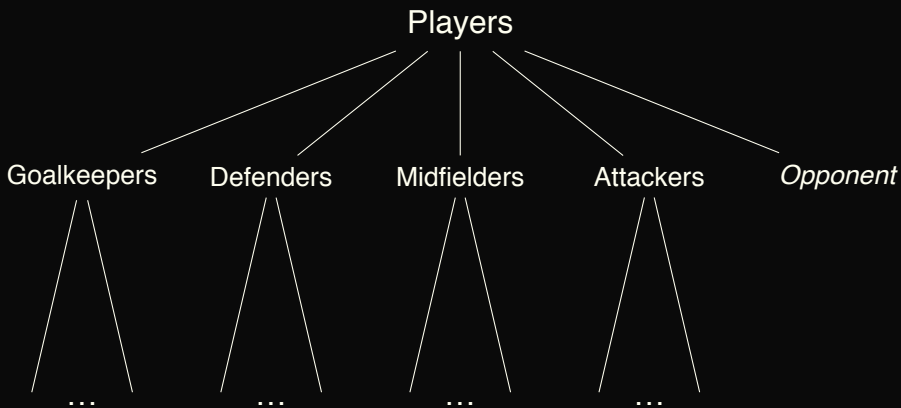
- ▶ Most players are interchangeable
- ▶ Precise coordinates hardly matter

Solution: *Hierarchical background knowledge*

Zone hierarchy divides pitch into 6 zones and 20 subzones 18/44












Player hierarchy divides players by positions 19/44












- ▶ Predicates *shot/1* (target), *pass/5*, *cross/5*, and *setpiece/5* + hierarchies
- ▶ Seed-based rule search
Each example as a seed
- ▶ Guide by m -estimate, smoothed estimate of rule precision
- ▶ Eliminate redundant rules via post-processing

Phase label: \oplus

Event		From		To
 Free kick	p13	(-1470, 2570)	Opponent	(-3200, 2630)
 Pass	Opponent	(-3200, 2630)	p1	(540, 510)
 Pass	p1	(540, 510)	p6	(-270, -1370)
 Pass	p6	(-270, -1370)	p4	(-820, -2640)
 Pass	p4	(-820, -2640)	p6	(-1830, -2630)
 Cross	p6	(-1830, -2630)	Opponent	(-4330, -770)
 Pass	Opponent	(-4330, -770)	p6	(-4470, -2190)
 Pass	p6	(-4470, -2190)	Opponent	(-4270, -890)
 Pass	Opponent	(-4270, -890)	p8	(-3350, -1640)

Target: *shot(ex1)*

Event	From				To
 <i>setpiece(ex1,</i>	<i>p13,</i>	<i>zO9,</i>	<i>opp,</i>	<i>zO6)</i>	
 <i>pass(ex1,</i>	<i>opp,</i>	<i>zO6,</i>	<i>p1,</i>	<i>zD10)</i>	
 <i>pass(ex1,</i>	<i>p1,</i>	<i>zD10,</i>	<i>p6,</i>	<i>zO10)</i>	
 <i>pass(ex1,</i>	<i>p6,</i>	<i>zO10,</i>	<i>p4,</i>	<i>zO8)</i>	
 <i>pass(ex1,</i>	<i>p4,</i>	<i>zO8,</i>	<i>p6,</i>	<i>zO8)</i>	
 <i>cross(ex1,</i>	<i>p6,</i>	<i>zO8,</i>	<i>opp,</i>	<i>zO2)</i>	
 <i>pass(ex1,</i>	<i>opp,</i>	<i>zO2,</i>	<i>p6,</i>	<i>zO5)</i>	
 <i>pass(ex1,</i>	<i>p6,</i>	<i>zO5,</i>	<i>opp,</i>	<i>zO2)</i>	
 <i>pass(ex1,</i>	<i>opp,</i>	<i>zO2,</i>	<i>p8,</i>	<i>zO7)</i>	

p13 is a midfielder

$pass(E, pMID, Z1, P2, Z2) \Leftarrow pass(E, p13, Z1, P2, Z2)$

$pass(E, p1, Z1, pMID, Z2) \Leftarrow pass(E, P1, Z1, p13, Z2)$

zO3 is in the offensive penalty area

$cross(E, P1, zOPA, P2, Z2) \Leftarrow cross(E, P1, zO2, P2, Z2)$

$cross(E, P1, Z1, P2, zOPA) \Leftarrow cross(E, P1, Z1, P2, zO2)$

$shot(ex1) \Leftarrow$

$setpiece(ex1,$	$p13,$	$zO9,$	$opp,$	$zO6)$
$pass(ex1,$	$opp,$	$zO6,$	$p1,$	$zD10)$
$pass(ex1,$	$p1,$	$zD10,$	$p6,$	$zO10)$
$pass(ex1,$	$p6,$	$zO10,$	$p4,$	$zO8)$
$pass(ex1,$	$p4,$	$zO8,$	$p6,$	$zO8)$
$cross(ex1,$	$p6,$	$zO8,$	$opp,$	$zO2)$
$pass(ex1,$	$opp,$	$zO2,$	$p6,$	$zO5)$
$pass(ex1,$	$p6,$	$zO5,$	$opp,$	$zO2)$
$pass(ex1,$	$opp,$	$zO2,$	$p8,$	$zO7)$

Generalise: Replace identifier with a variable 25/44

shot(*X*) \Leftarrow

<i>setpiece</i> (<i>X</i> ,	<i>p13</i> ,	<i>zO9</i> ,	<i>opp</i> ,	<i>zO6</i>)
<i>pass</i> (<i>X</i> ,	<i>opp</i> ,	<i>zO6</i> ,	<i>p1</i> ,	<i>zD10</i>)
<i>pass</i> (<i>X</i> ,	<i>p1</i> ,	<i>zD10</i> ,	<i>p6</i> ,	<i>zO10</i>)
<i>pass</i> (<i>X</i> ,	<i>p6</i> ,	<i>zO10</i> ,	<i>p4</i> ,	<i>zO8</i>)
<i>pass</i> (<i>X</i> ,	<i>p4</i> ,	<i>zO8</i> ,	<i>p6</i> ,	<i>zO8</i>)
<i>cross</i> (<i>X</i> ,	<i>p6</i> ,	<i>zO8</i> ,	<i>opp</i> ,	<i>zO2</i>)
<i>pass</i> (<i>X</i> ,	<i>opp</i> ,	<i>zO2</i> ,	<i>p6</i> ,	<i>zO5</i>)
<i>pass</i> (<i>X</i> ,	<i>p6</i> ,	<i>zO5</i> ,	<i>opp</i> ,	<i>zO2</i>)
<i>pass</i> (<i>X</i> ,	<i>opp</i> ,	<i>zO2</i> ,	<i>p8</i> ,	<i>zO7</i>)

$shot(X) \Leftarrow$

$setpiece(X, p13, zO9, opp, zO6)$
 $pass(X, opp, zO6, p1, zD10)$
 $pass(X, p1, zD10, p6, zO10)$
 $pass(X, p6, zO10, p4, zO8)$
 $pass(X, p4, zO8, p6, zO8)$

$shot(X) \Leftarrow$

$setpiece(X,$	$pMID,$	$zO9,$	$opp,$	$zO6)$
$pass(X,$	$opp,$	$zO6,$	$p1,$	$zD10)$
$pass(X,$	$p1,$	$zD10,$	$p6,$	$zO10)$
$pass(X,$	$p6,$	$zO10,$	$p4,$	$zO8)$
$pass(X,$	$p4,$	$zO8,$	$p6,$	$zO8)$

Generalise: Replace constants with variables 28/44

$shot(X) \Leftarrow$

$setpiece(X, pMID, zO9, opp, zO6)$

$pass(X, opp, zO6, p1, zD10)$

$pass(X, p1, zD10, p6, zO10)$

$pass(X, p6, zO10, Y, zO8)$

$pass(X, Z, zO8, p6, zO8)$

$$m(C) = \frac{|C^+| + mp}{|C| + p}$$

$|C|$ number of covered examples

$|C^+|$ number of covered *positive* examples (shots)

m prior probability of positive label

p strength of prior beliefs (pseudocount)

Outline

- Data modelling & analysis

- Experimental results

- Further extensions

- Conclusions

- ▶ 70 games
- ▶ 3,803 phases
 - ▶ 526 positive examples / shots (13.8%)
 - ▶ 3,277 negative examples
- ▶ 26,338 ball movement events
- ▶ 7 events per phase on average
6.5 passes, 0.3 crosses, 0.2 set pieces

Full: Players & zones

$$\textit{shot}(X) \Leftarrow \textit{pass}(X, p1, p2, z3, z4)$$

Players only

$$\textit{shot}(X) \Leftarrow \textit{pass}(X, p1, p2)$$

Zones only

$$\textit{shot}(X) \Leftarrow \textit{pass}(X, z3, z4)$$

Overview of results

32/44

	Setup	Hierarchy	Rules	m-est. of prec. (top 10) Maximum	Average	Time (min.) per seed
/3	Spatial		276	0.7396	0.6638	0.002
		+	323	0.7396	0.7065	0.840
/3	Players		91	0.7396	0.4855	0.006
		+	257	0.7396	0.6606	5.250
/5	Full	+				Timeout

Background knowledge increases runtime

33/44

526 seeds (shots)

	Setup	Hierarchy	Rules	m-est. of prec. (top 10) Maximum	Average	Time (min.) per seed
/3	Spatial		276	0.7396	0.6638	0.002
		+	323	0.7396	0.7065	0.840
/3	Players		91	0.7396	0.4855	0.006
		+	257	0.7396	0.6606	5.250
/5	Full	+				Timeout

Background knowledge improves rule quality 34/44

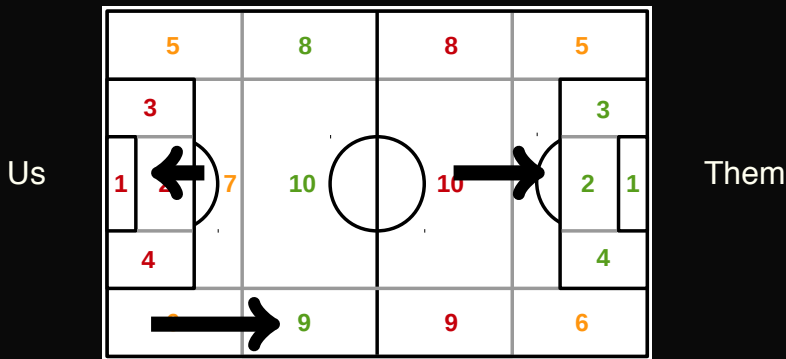
	Setup	Hierarchy	Rules	m-est. of prec. (top 10) Maximum	Average	Time (min.) per seed
/3	Spatial		276	0.7396	0.6638	0.002
		+	323	0.7396	0.7065	0.840
/3	Players		91	0.7396	0.4855	0.006
		+	257	0.7396	0.6606	5.250
/5	Full	+				Timeout

Counter-attack via right flank and middle

35/44

Top-scoring spatial rule

$$\begin{array}{ccc} |C| & |C^+| & m\text{-}est \\ 5 & 5 & 0.74 \end{array}$$



$$WRAcc(C) = \underbrace{\frac{|C|}{|D|}}_{\text{rule size}} \times \underbrace{\left(\frac{|C^+|}{|C|} - \frac{|D^+|}{|D|} \right)}_{\text{bonus accuracy}}$$

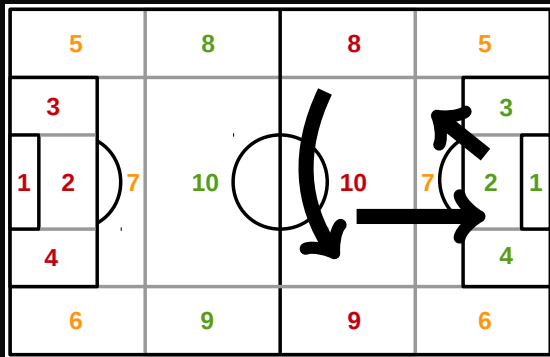
$|D|, |D^+|$ total number of (positive) examples in the data

$|C|, |C^+|$ number of covered (positive) examples

WRAcc gives more frequent/less pure rules 37/44

$|C|$ $|C^+|$ $m\text{-est}$ $WRAcc$
 $62 \gg 5$ $18 \gg 5$ $0.275 \ll 0.74$ 0.025

Us



Them

Outline

■ Data modelling & analysis

■ Experimental results

■ Further extensions

■ Conclusions

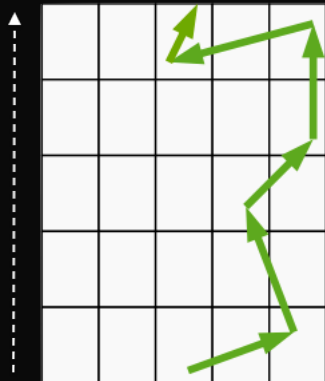
I only contributed to the previous part

Van Haaren, J., Dzyuba, V., Hannosset, S., & Davis, J. (2015).
Automatically Discovering Offensive Patterns in Soccer Match Data. Proceedings of IDA (pp. 286-297).

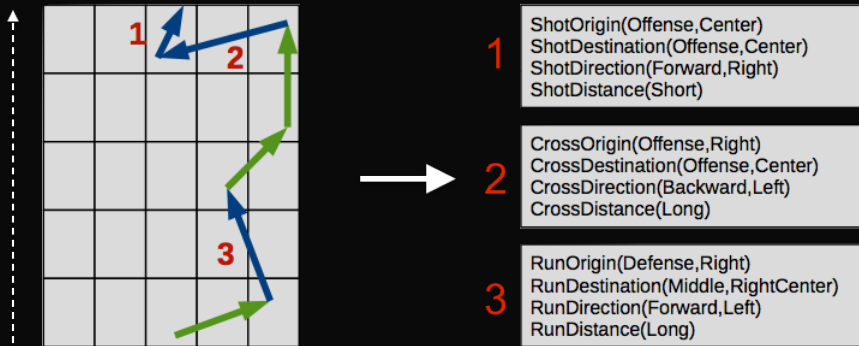
- ▶ Only few event types are considered
- ▶ Event order is discarded
- ▶ Player tracking data is unused

- ▶ Cluster phases based on *all events*
- ▶ Mine *sequential* patterns within each cluster
- ▶ Only few event types are considered
- ▶ Event order is discarded
- ▶ Player tracking data is unused

Clustering considers frequencies of all event types 41/44

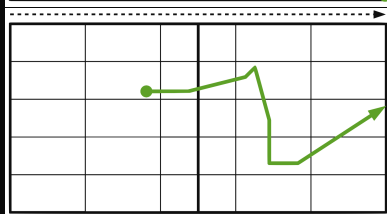
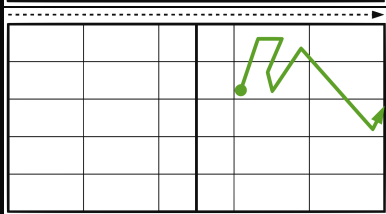
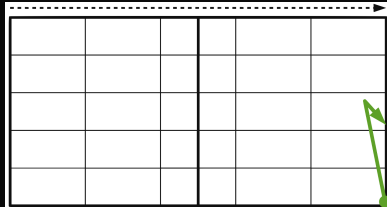
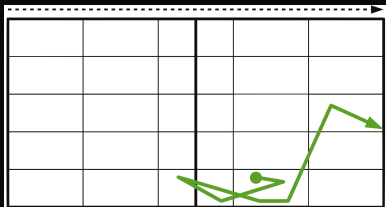


0	2	7	4	5
1	5	2	3	4
0	1	1	2	7
1	4	6	5	3
0	0	2	1	0



Tactics discovered in data-driven way

43/44



Outline

- Data modelling & analysis
- Experimental results
- Further extensions
- Conclusions

ILP is a useful tool for modelling sports data

- ▶ Naturally relational problem
- ▶ Hard to represent with fixed-length feature vectors
- ▶ Hierarchical background knowledge

Efficiency requires several modifications:

- ▶ Clustering to partition examples
- ▶ Sequence mining to account for order information

Mining soccer data

Thank you for your attention!

May I answer any questions?