

Visualization of Unstructured Sports Data - An Example of Cricket Text Commentary

Vijaya Saradhi

IIT Guwahati

Thu, 21th Nov 2019

Collaborators

Rule Construction work in ICMLA2019

- Swarup Ranjan Behera, PhD student, IIT Guwahati
- Parag Agrawal, TechLead, LinkedIn, USA
- Dr. Amit Awekar, Dept. of CSE, IIT Guwahati and
- Dr. Vijaya Saradhi, Dept. of CSE, IIT Guwahati

Collaborators

Visualization work

- Swarup Ranjan Behera, PhD student, IIT Guwahati
- Dr. Vijaya Saradhi, Dept. of CSE, IIT Guwahati

A Little Bit About You & Cricket

Domain Knowledge

- This work requires some understanding of the cricket domain
- Are you aware of the cricket game?
- Formats of the game?
- Have you played cricket?
- Were you a batsman? Were you a bowler? Were you wicket keeper
- Were you a captain?
- Have you coached any team?
- Do not worry even if the answer is NO to all of the above questions

A Little Bit About Cricket

About Cricket

- Two teams
- Eleven players each
- Three formats
 - Test Cricket 90 overs per day spanning for five days
 - One Day Cricket 50 overs per team game
 - 20-20 Cricket 20 Overs per team game
- Winning criteria? When a team score more runs than the opponent team
- One day and 20-20 are result oriented - that is match should result in win or loose for a team
- In test cricket the result could be win, loose or draw (neither team has own or lost)

Test Cricket

About test cricket

- Spans over five days
- Played at the same venue
- Played on the same pitch
- Each team gets to play twice in turns
- There will not be any changes to the team players during the five day play

First innings Team A is bats first. When all batsmen are out, team B gets to bat

Second innings Team A once again bats followed. Depending on the total score of Team A and team B's target is decided

Result when team B reaches the target set by team A

OR all the five days are completed and team B still short of the target set by team A.

Test Cricket - Data Generated

Generated Data

Box data Score card data

Tracking data ball & player tracking data

Meta data Non-game related data











Box Data

Score cards - offline






MATCH		v ENGLAND		PLAYED AT		COMPETITION		ASHES	
BETWEEN AUSTRALIA		v ENGLAND		AT MULLONGMEAD		COMPETITION		ASHES	
1st or 2nd INNINGS OF AUSTRALIA (1st)		TOSS WON BY AUSTRALIA		DATE 25-NOV-1994		GRADE TEST		ROUND NO. 1	
1	9	141							
2	19	152							
3	19	10							
4	29	14							
5	26	192							
6	27	198							
7	27	106							
8	27	106							
9	31	188							
10	27	188							
11	27	188							
12	44	212							
13	45	212							
14	45	212							
15	45	212							
16	50	238							
17	55	240							
18	58	250							
19	61	259							
20	62	261							
21	63	261							
22	70	276							
23	76	277							
24	77	285							
25	86	293							
26	91	297							
27	92	300							
28	94	311							
29	95	323							
30	99	331							
31	107	348							
32	112	353							
33	117	359							
34	118	359							
35	124	372							
36	125	383							
37	126	393							
38	127	393							
39	128	393							
40	132	408							
41	132	408							
42	139	408							
43	140	408							

Box Data

Score cards - online

India 2nd innings		R	M	B	4s	6s	SR
 G Gambhir	c Warner b Siddle	83	217	142	11	0	58.45
 V Sehwag	c Warner b Hifenhaus	4	14	8	1	0	50.00
 R Dravid	b Hifenhaus	29	100	73	6	0	39.72
 SR Tendulkar	c Hussey b Clarke	80	213	141	9	0	56.73
 VVS Laxman	b Hifenhaus	66	130	119	7	0	55.46
 V Kohli	lbw b Pattinson	9	35	24	2	0	37.50
 MS Dhoni*†	c & b Hifenhaus	2	11	11	0	0	18.18
 R Ashwin	c Lyon b Hifenhaus	62	102	76	9	1	81.57
 Z Khan	c Marsh b Siddle	35	36	26	5	1	134.61
 I Sharma	lbw b Lyon	11	38	35	2	0	31.42
UT Yadav	not out	0	17	14	0	0	0.00
Extras	(b 6, lb 3, w 2, nb 8)	19					
Total	(all out; 110.5 overs)	400				(3.60 runs per over)	

Fall of wickets 1-18 (Sehwag, 3.3 ov), 2-100 (Dravid, 26.3 ov), 3-168 (Gambhir, 48.2 ov), 4-271 (Tendulkar, 78.2 ov), 5-276 (Laxman, 82.1 ov), 6-286 (Dhoni, 84.6 ov), 7-286 (Kohli, 85.5 ov), 8-342 (Khan, 93.5 ov), 9-384 (Sharma, 105.1 ov), 10-400 (Ashwin, 110.5 ov)

Bowling	O	M	R	W	Econ	0s	4s	6s
 JL Pattinson	23	4	106	1	4.60	101	13	0 (3nb, 1w)
 BW Hifenhaus	32.5	8	106	5	3.22	159	20	0 (1w)
 PM Siddle	24	8	88	2	3.66	110	13	1 (1nb)
 NM Lyon	20	2	64	1	3.20	88	6	1
 MJ Clarke	9	0	22	1	2.44	36	0	0
MEK Hussey	2	0	5	0	2.50	7	0	0

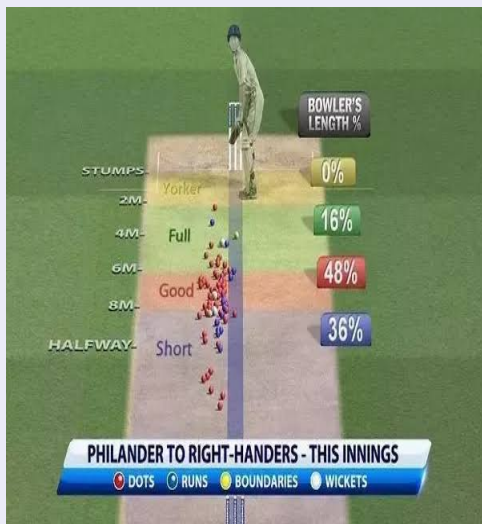
Tracking Data

Wagon Wheel



Tracking Data

Pitch Map



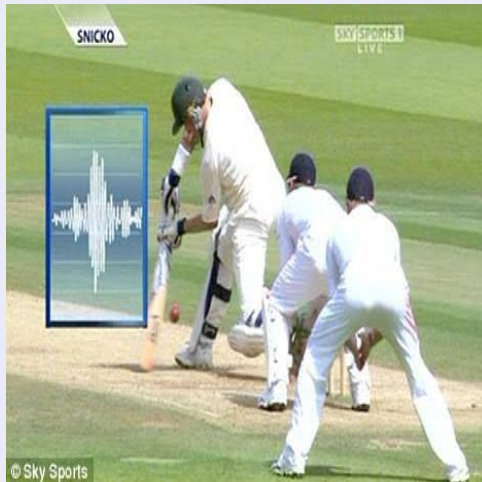
Tracking Data

Fielding Position Map



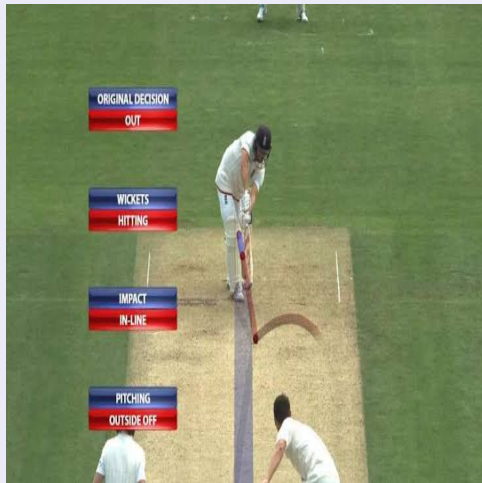
Tracking Data

Snikometer



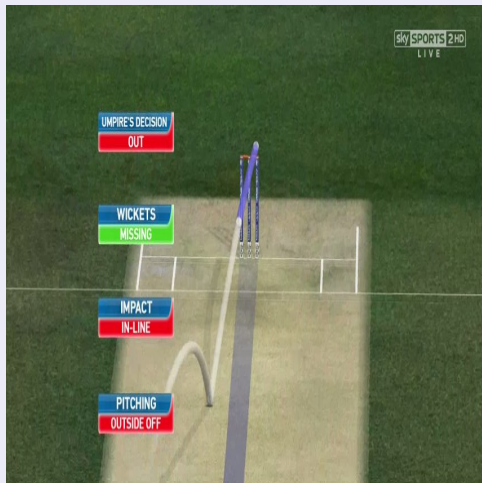
Tracking Data

Hawk Eye



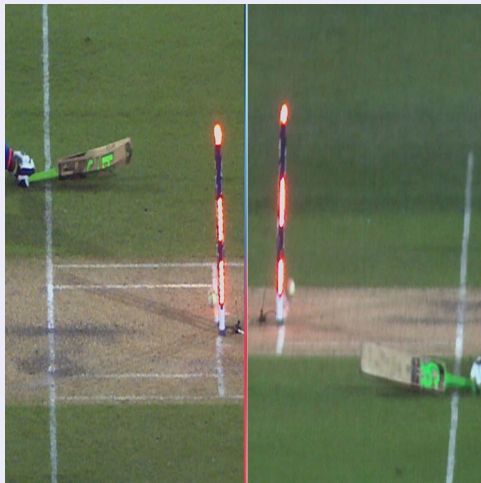
Tracking Data

Hawk Eye



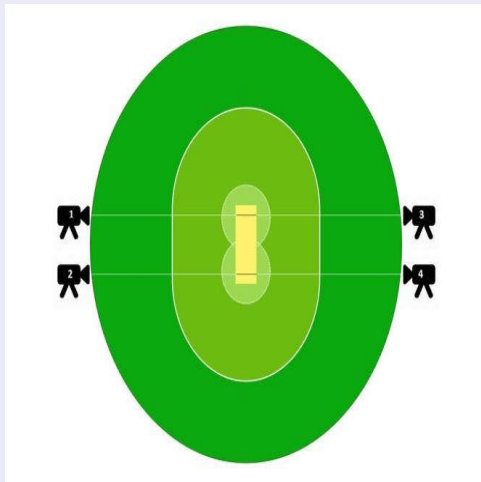
Tracking Data

Camera Placements - Run out



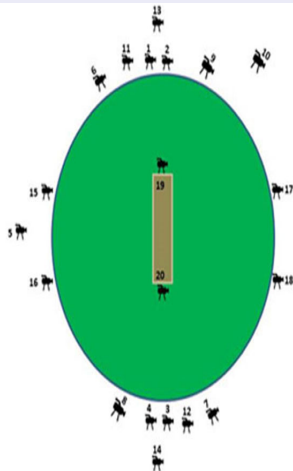
Tracking Data

Camera Placements - No ball



Tracking Data

Camera Placements - Complete ground



Background

Box Score Data

- First kind of data that is recorded
- The earliest box score data dates back to 1772 in EspnCricInfo
- Today's box score data differs only in few places with the first version
- Fall of wickets, number of balls faced, number of balls bowled, number of extras etc.
- 247 years and changes to this form of data reflected in very few places

Background

Tracking Data

- Uses advanced data capture techniques
- Hawk-eye technology introduced in 2001
 - Tracks ball trajectory
 - Speed of the delivery
 - Position of the pitched ball
 - Bounce of the ball
 - Used by third umpires for taking decisions
- UltraEdge or Snicko meter introduced in 2013
 - Analyses sound graphically and shows the noise frequency
 - Used in finding whether ball edged the bat
- HotSpot introduced in 2006
 - Heat-sensitive infrared cameras installed
 - Used in decision about the ball struck batsman on bat or pad

Tracking & Meta Data

Publicly Not Available

- Several sensors deployed on the cricket field to collect the data
- Used predominantly in making informed decisions by third umpire
- This form of data is not available publicly

Background

Meta Data

- Account for conditions that are outside the context of the game
- Example: [Moisture condition](#), [pitch condition](#)
- Venue is home ground of the hosting team or not
- New ball or old ball

Unstructured Data

Text Data

- Unstructured text data is witnessing substantial growth every year
- Sources include: [blogs](#), [microblogs](#), [news articles](#) and [social media posts](#)
- The text data is utilized in various applications
 - Information retrieval
 - Clustering
 - Classification
 - Event detection
 - Pattern Analysis

Motivation

Unstructured Data - in Sports

- Unstructured text data also percolated in sports domain
- **Sources** Same as described above
- Visualization based on unstructured data is not fully explored

Motivation

Cricket Commentary Example

106.1, Anderson to Smith, 1 run, England have drawn a false shot from Smith! well done. good length, angling in, straightens away, catches the outside edge but does not carry to Cook at slip. 106.1, Anderson to Smith, 1 run, England have drawn a false shot from Smith! well done. good length, angling in, straightens away, catches the outside edge but does not carry to Cook at slip.

Motivation

Cricket Commentary - Color Coding

- Red color stand for **Structured data**
- Blue color stand for **Unstructured data**
- Unstructured data contains minute details that differ significantly from the structured part
- Batsman or bowler exhibit weakness on the any delivery and not necessarily on the delivery in which he got out or attacked to four or six

Motivation

Central Idea

To propose a set of **new** visualization methods for cricket using **unstructured cricket short text commentary data**

The Problem

Propose a Computational Definition

For expressing strength and weakness rules

Computational Model

- Propose a computational model to find strength and weakness rules of individual players
- Example: Bowl to Sehwag outside the off stump a bouncer ball to get the wicket

Visualization

To visualize the constructed strength and weakness rules

Non-trivial Problem

Why it is a difficult problem?

- Intuitively stating one strength (or one weakness rule) may be possible
- Only for particular players one will be able to explain such a rule
- Listing several strengths and several weaknesses for one player is quite difficult task
- It is not guaranteed that there will be agreement on the proposed strength/weakness rule
- What might be a weakness from one expert's perspective may not be weakness from second expert's perspective
- There is no **universally** agreeable definition of strength and weakness rule
- Validating the obtained rules is a challenge as we do not have **ground truth**
- Graphically representing these rules is complex one

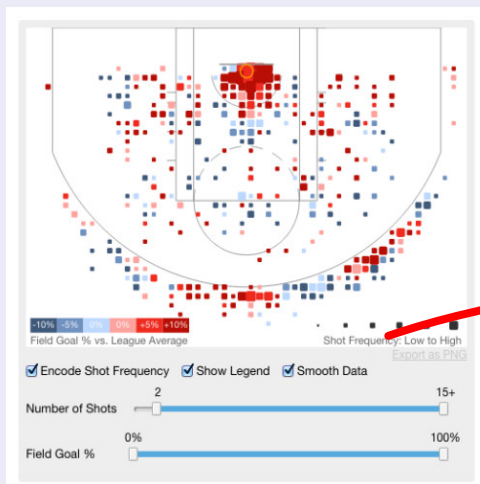
Non-trivial Problem

Why it is a difficult problem?

- There is no agreeable **definition** for strength or weakness of a player
- There is no **computational model** that exists for identifying such rules

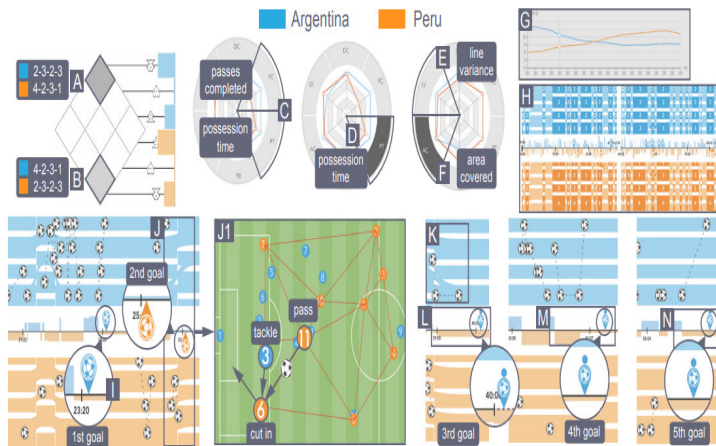
Structured Data In Other Sports Domains

Basket Ball



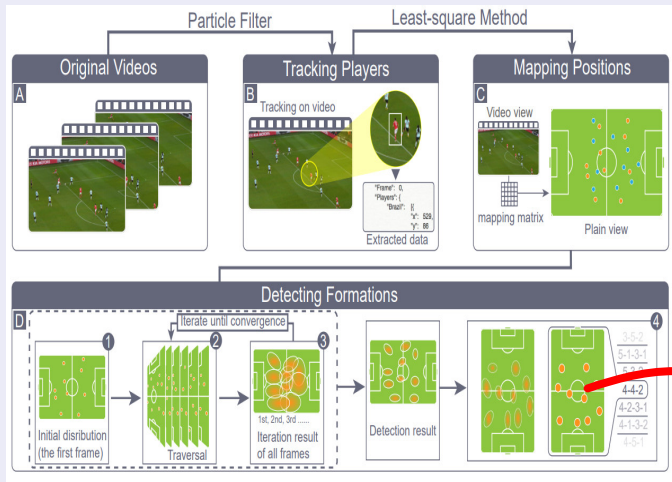
Structured Data In Other Sports Domains

Football - 01



Unstructured Data In Other Sports Domains

Football - 02



Unstructured Data In Other Sports Domains

Football - 03 - Trend Analysis

- Uses number of tweets pertaining to a specific football match
- Provided the visualization of exciting moments during soccer games
- Peak number of tweets reflect exciting moments

Text Visualization

Classification

- Single document
- Document collection
- Document collection over time

Single Document

Tag Cloud

TagCrowd is a [web application for visualizing word frequencies in any text](#) by creating what is popularly known as a word cloud, text cloud or tag cloud.

Visualization - Tag Cloud



Single Document

Wordle Cloud

A web-based tool for visualizing text. It improves visual area for obtaining the word cloud

Wordle Cloud

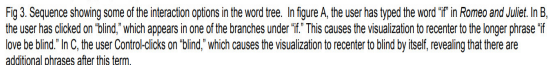


Single Document

Word Tree

- The word frequency-based methods cannot convey semantic relationships among the words
- Word Tree is proposed to overcome this limitation
- Text document is represented as a tree (suffix tree) with nodes representing words and branches linking sequential words

Word Tree



Document Collection

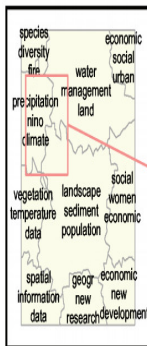
Cartographic Visualization

- A cartographic approach to mapping non-geographic information
- It helps to manage graphic complexity
- It requires same cognitive skills we use when viewing geographic maps
- Applied to collection of abstracts

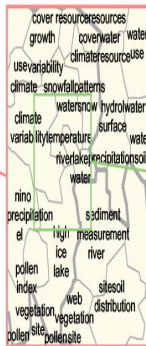
Document Collection

Cartographic Visualization

6 Three different zoom levels in a visualization of conference abstracts: (a) complete map shown in a 10-cluster solution and map portions for (b) a 100-cluster and (c) 800-cluster solution. Higher level boundaries are accentuated to provide context during zoom operations.



(a)



(b)



(c)

complexity. When we apply this principle to nongeographic data via hierarchical clustering, it provides for

tions. In this experiment I decided to feed raw term counts to the SOM training procedure. This makes for

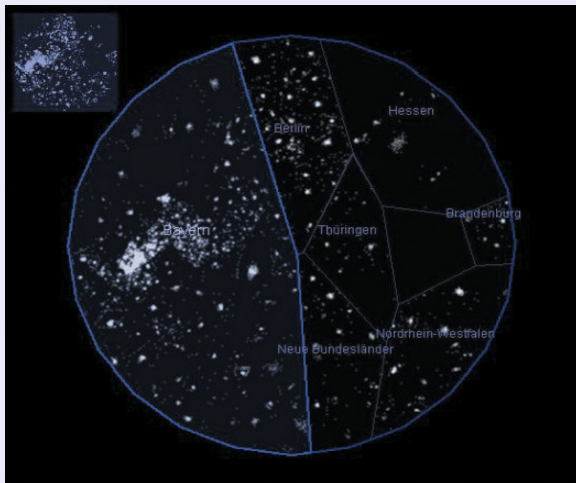
Document Collection

InfoSky Visualization

- Represent both the hierarchical organization of documents and inter-document similarity within a single visualization
- Integrate both global and local view of the information space into one seamless visualization

Document Collection

InfoSky Visualization



Document Collection Over Time

SparkCloud Visualization

- Tag clouds do not explicitly represent trends
- They do not support comparisons
- In spark cloud, spark lines are integrated into tag cloud

Document Collection

SparkCloud Visualization



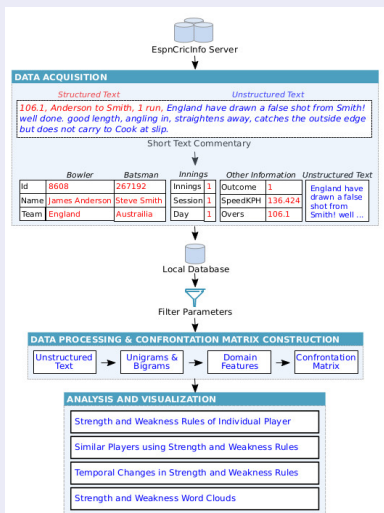
Requirements From Cricket Text Commetnary

Research Questions

- RQ1 What are the strength and weakness rules of a player? Can these rules be *visualized*?
- RQ2 How do the strength and weaknesses of a player change over time?
- RQ3 Which players have **similar** strengths and **similar** weaknesses?
- RQ4 Are traditional text visualization techniques answer the above questions?

System Over View

Proposed System



Issues With Cricket Commentary

Issues

Technical words in cricket are stop words in conventional information retrieval

off	on	room	across	behind	back	out
good	great	into	away	up	down	long
turn	point	from	further	under	full	open

Issues With Cricket Commentary

Issues

- Objective is neither to store text commentary data nor to retrieve text commentary data
- Traditional text representation models cannot be directly employed
- Term frequency is significantly less in every document
- Inverse document frequency dominates
- Technical words lose their significance in the TFIDF representation

Issues With Cricket Commentary

Issues

Word	Off	On	Stump	Outside	Leg	Length
Frequency	34043	30953	17604	16407	15907	12490
Word	Ball	Short	Defend	Line	Full	
Frequency	12419	11146	10266	6593	6388	

Issues With Cricket Commentary

Issues

- Technical words **alone** cannot distinguish type of ball bowled or the way batsman played

Issues With Cricket Commentary

Issues

- Technical words **alone** cannot distinguish type of ball bowled or the way batsman played
- 151.2 Flintoff to Mishra, no run, **Swings in from outside off**, well left in the end as it scoots past off stump.

Issues With Cricket Commentary

Issues

- Technical words **alone** cannot distinguish type of ball bowled or the way batsman played
- 151.2 Flintoff to Mishra, no run, **Swings in from outside off**, well left in the end as it scoots past off stump.
 - Swing signify the type of ball

Issues With Cricket Commentary

Issues

- Technical words **alone** cannot distinguish type of ball bowled or the way batsman played
- 151.2 Flintoff to Mishra, no run, **Swings in from outside off**, well left in the end as it scoots past off stump.
 - Swing signify the type of ball
- 122.4 Flintoff to Dhoni, 1 run, Short ball over middle stump, **Dhoni swings into a pull** and takes it down to fine leg.

Issues With Cricket Commentary

Issues

- Technical words **alone** cannot distinguish type of ball bowled or the way batsman played
- 151.2 Flintoff to Mishra, no run, **Swings in from outside off**, well left in the end as it scoots past off stump.
 - Swing signify the type of ball
- 122.4 Flintoff to Dhoni, 1 run, Short ball over middle stump, **Dhoni swings into a pull** and takes it down to fine leg.
 - Swing signifies the way batsman played the ball

Issues With Cricket Commentary

Issues

- Meaning of technical words change when combined with other technical words

Issues With Cricket Commentary

Issues

- Meaning of technical words change when combined with other technical words
- **Short** refers to the type of ball

Issues With Cricket Commentary

Issues

- Meaning of technical words change when combined with other technical words
- **Short** refers to the type of ball
- When combined with **leg**, **midwicket** or **cover** refers to fielding position

Issues With Cricket Commentary

Issues

- Meaning of technical words change when combined with other technical words
- **Short** refers to the type of ball
- When combined with **leg**, **midwicket** or **cover** refers to fielding position
 - **Short on the body**, he gets up and nicely plays it to square leg but declines the single.

Issues With Cricket Commentary

Issues

- Meaning of technical words change when combined with other technical words
- **Short** refers to the type of ball
- When combined with **leg**, **midwicket** or **cover** refers to fielding position
 - **Short on the body**, he gets up and nicely plays it to square leg but declines the single.
 - Full outside off, Dhoni reaches out and pushes it to **short cover**

Comprehensive List of Features

Feature Name	Description	Values
Batting features		
Outcome	Number of runs scored in a particular delivery	0, 1, 2, 3, 4, 5, 6(also 6+) runs and out
Response	Response of the batsman on each delivery	Beaten (exhibits imperfection), Defended (blocks or leaves the ball), Attacked (plays aggressive shots)
Footwork	Technique a batsman employs when facing a delivery	Front foot (ball is played in front of the batsman), Back foot (ball is played behind the batsman's wicket)
Shot area	Region where shot is played by batsman	Third man, Square off, Long off, Long on, Square leg, Fine leg
Bowling features		
Length	How far down the pitch the ball bounces	Short (closer to the bowler), Good (on an optimal length, in between short and full), Full (nearer the batsman)
Line	How far to the left or right of the wicket the ball is travelling w.r.t. the line from wicket to wicket	Off (on or outside off-stump), Middle (on middle-stump), Leg (on or outside leg-stump)
Type	Nature of the delivery	Spin (slow deliveries which turn sharply after pitching), Swing (fast deliveries which have movement in the air)
Speed	Speed of the ball after it is released	Fast (medium: 60-80 mph, fast: 80+ mph), Slow (40-60 mph)
Movement	Movement of the ball as it approaches batsman	Move-in (towards batsman), Move-away (away from batsman)

Features - Batting

Attack: A well executed shot or display aggressive play.

Attack	
boundary	deep
cut	sweep
pull	hook
flick	third man
long on	long off
down ground	Negative
down track	leave
whip	defend
punch	deep crease
whack	block
four	leave alone
six	down under
deep	edge
nice timing	shoulder arm
great timing	push back
drive	miss

Features - Batting

Defense: Intension of not scoring any run or the delivery is a difficulty one to play a shot.

Defense	
leave	shoulder arm
defend	push back
defense	play back
block	tap back
leave alone	down under

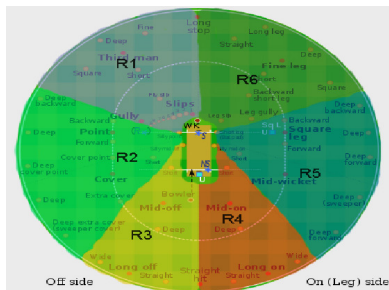
Features - Batting

Shot Type: Various orthodox shots offered.

Shot Type	
drive driven cut	pull hook

Features - Batting

Shot Area: Region where the ball is hit.



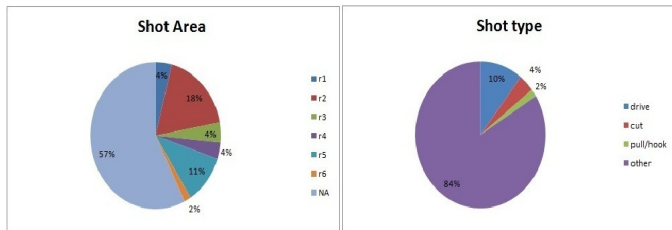
Features - Batting

Shot Area: Region where the ball is hit.

Shot Area					
R1	R2	R3	R4	R5	R6
third man late cut cut gully Negative square cut point	square cover point square cut square drive cover drive Negative square leg	mid off long off off drive straight drive	mid on long on on drive	square leg sweep hook pull flick midwicket leg gully	fine leg long leg leg glance

Features - Batting

Shot Area: Region where the ball is hit.



Features - Batting

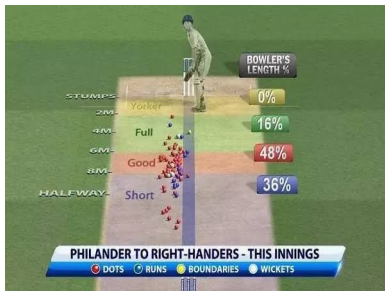
Weakness: Exhibiting imperfection on a delivery.

Weakness	
miss	struck pad
beat	wrong line
edge	lbw
plumb	poor shot
hit stump	bad shot
knock stump	out mistimed
miss stump	out bowled
uproot stump	out pad
expose stump	out stump
crash stump	out lbw
crash glove	

Outcome: Result from batting perspective for a particular delivery.

Bowling Related Features

Line: Indicates the direction of the ball with respect to batsman's wickets.



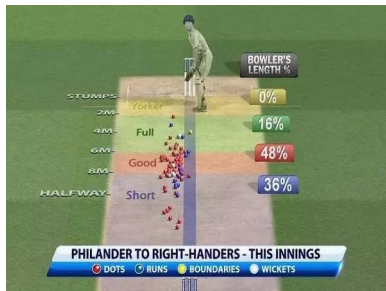
Bowling Related Features

Line: Indicates the direction of the ball with respect to batsman's wickets.

Line		
Off	Middle	Leg
outside off wide off pitch off off stump from off	bowl straight straight ball straight line on middle into middle pitch middle middle stump Negative straight bat straight drive	down leg wide leg outside leg leg stump Negative fine leg square leg short leg long leg

Bowling Related Features

Length: Area where the ball is pitched.



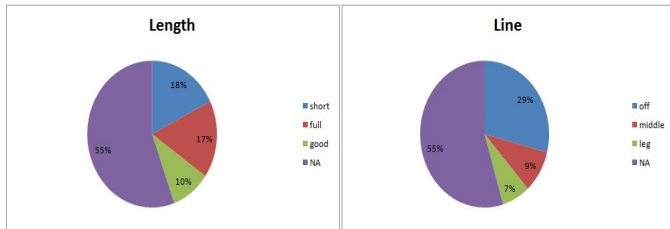
Bowling Related Features

Length: Area where the ball is pitched.

Length		
Short	Full	Good length
short bounce bouncer bumper shortpitch back length Negative short midwicket short cover short leg	full fuller overpitch full toss half volley toss up blockhole	length good length length delivery Negative short length full length back length fuller length shorten length

Bowling Related Features

Line and Length Coverage: Majority of the balls are short placed outside off stump.



Bowling Related Features

Type: Refers to swing, spin or a normal delivery.

Quality: Ball is good or bad as described by commentator.

Spin	Swing	Good ball	Poor Ball
spin	swing away	good bowling	poor bowling
turn	swing in	good ball	poor ball
googly	swing back	great ball	poor line
doosra	late swing	superb ball	bad line
topspin	inswing	well bowled	bad delivery
legspin	outswing	tight line	bad ball
offspin	reverse swing	good delivery	
		superb delivery	
		jaffa	

Bowling Related Features

Speed and Movement: whether ball is moving into the batsman or moving away.

Speed		Moment	
Fast	Slow	Away	Into
fast	slow	move away	move in
pace	slower	swing away	swing in
quick		turn away	turn in
quicker		seam away	seam in
		angle away	angle in

Selection of Text Commentary

Filter Tuple

- Has three parameters

Player Whose text commentary to be extracted

Opponent player Against whom the player in question batted (or bowled)

Time duration involved (per over, per session, per day, ..., entire career)

- Given the above three parameters, extract subset of the text commentary
- Example: $\langle \text{Steve Smith, All opponent players, entire career} \rangle$
- Will extract text commentary for the batsman Steve Smith played against all the opponennt players for the entire career of Steve Smith

Confrontation Matrix

Example

Table 1: Confrontation Matrix for Batsman Steve Smith

	good	short	full	off	leg	middle	spin	swing	fast	slow	move-in	move-out
run0	1331	1461	1933	3720	304	805	1018	196	4830	1704	660	297
run1	157	522	387	497	156	243	278	23	997	617	122	17
run2	47	117	110	134	38	68	55	9	295	109	33	5
run3	5	40	45	48	9	9	11	4	89	36	5	1
run4	40	215	209	214	42	72	56	26	484	124	30	19
run5	1	3	2	1	1	0	0	0	6	0	0	0
run6	3	6	5	10	0	2	5	2	11	14	0	0
out	14	21	18	38	7	27	33	10	60	23	18	12
attacked	269	955	925	1160	204	419	449	63	1917	1088	208	68
defended	862	674	930	2179	61	361	435	91	2506	737	332	134
beaten	106	135	159	296	68	78	163	44	522	163	112	61
front foot	119	36	311	358	11	77	156	14	347	372	41	20
back foot	180	318	80	346	19	91	171	12	418	287	95	21
thirdman	25	85	36	100	1	6	14	4	172	21	7	10
square off	273	591	523	1073	20	96	230	38	1367	427	79	78
long off	30	35	199	193	4	18	37	10	237	75	12	9
long on	48	69	269	208	11	126	91	12	316	225	40	1
square leg	217	553	455	464	348	462	364	42	1210	747	218	20
fine leg	17	93	35	39	35	34	34	7	191	38	27	6

Confrontation Matrix

Construction

- 1: **for** Every commentary **do**
- 2: Get all the unigrams and bigrams from text commentary

Confrontation Matrix

Construction

```
1: for Every commentary do
2:   Get all the unigrams and bigrams from text commentary
3:   for Every unigram/bigram  $y$  do
4:     for  $i$  in batting features do
5:       if  $y \in FD_i$  then
6:          $bat = bat \cup \{i\}$ 
7:       end if
8:     end for
```

Confrontation Matrix

Construction

```
1: for Every commentary do
2:   Get all the unigrams and bigrams from text commentary
3:   for Every unigram/bigram  $y$  do
4:     for  $i$  in batting features do
5:       if  $y \in FD_i$  then
6:          $bat = bat \cup \{i\}$ 
7:       end if
8:     end for
9:     for  $j$  in bowling features do
10:      if  $y \in FD_j$  then
11:         $bowl = bowl \cup \{j\}$ 
12:      end if
13:    end for
14:  end for
```


Confrontation Matrix

Construction

```
1: for Every commentary do
2:   Get all the unigrams and bigrams from text commentary
3:   for Every unigram/bigram  $y$  do
4:     for  $i$  in batting features do
5:       if  $y \in FD_i$  then
6:          $bat = bat \cup \{i\}$ 
7:       end if
8:     end for
9:     for  $j$  in bowling features do
10:      if  $y \in FD_j$  then
11:         $bowl = bowl \cup \{j\}$ 
12:      end if
13:    end for
14:  end for
15:  for Every  $a \in bat$  and  $b \in bowl$  do
16:     $CM[a, b] = CM[a, b] + 1$ 
17:  end for
```

Confrontation Matrix

Construction

```
1: for Every commentary do
2:   Get all the unigrams and bigrams from text commentary
3:   for Every unigram/bigram  $y$  do
4:     for  $i$  in batting features do
5:       if  $y \in FD_i$  then
6:          $bat = bat \cup \{i\}$ 
7:       end if
8:     end for
9:     for  $j$  in bowling features do
10:      if  $y \in FD_j$  then
11:         $bowl = bowl \cup \{j\}$ 
12:      end if
13:    end for
14:  end for
15:  for Every  $a \in bat$  and  $b \in bowl$  do
16:     $CM[a, b] = CM[a, b] + 1$ 
17:  end for
18: end for
19: return Confrontation Matrix (CM)
```

Batting & Bowling Features Relationship

Example

	good
run0	1331
run1	157
run2	47
run3	5
run4	40
run5	1
run6	3
out	14
attacked	269
defended	862
beaten	106
front foot	119
back foot	180
thirdman	25
square off	273
long off	30
long on	48
square leg	217
fine leg	17

Batting & Bowling Features Relationship

Example

	good
run0	1331
run1	157
run2	47
run3	5
run4	40
run5	1
run6	3
out	14
attacked	269
defended	862
beaten	106
front foot	119
back foot	180
thirdman	25
square off	273
long off	30
long on	48
square leg	217
fine leg	17

Interpretation

On **good length** deliveries Steve Smith gets **beaten** 106 times

In terms of Probability

$P(\text{Good length delivery AND Steve Smith gets beaten}) = 0.0556 \left(\frac{106}{1907} \right)$

In terms of Probability

- $P(\text{Good length}) = 0.061 \left(\frac{3744}{61522} \right)$
- $P(\text{Beaten}) = 0.297 \left(\frac{106}{3744} \right)$
- Is $P(\text{Good length} \cap \text{Beaten}) = P(\text{Good length}) * P(\text{Beaten})$?
- is $0.055 = 0.061 * 0.297$?

Strength/Weakness

Definition

- Every rule must contain two features
- One bowling feature and one batting feature
- These two features must be dependent on each other
- That is when $P(a \cap b) = P(a) \times P(b)$ then a and b are independent
- When a and b are dependent then $P(a \cap b) = \alpha \times P(a) \times P(b)$ where $\alpha < 1$
- The amount of deviation is measured by the present method

Strength/Weakness

Definition

Correspondence Analysis

Definition

$$\alpha_{ab} = \frac{P(a \cap b)}{P(a) \times P(b)} \text{ for all } a \text{ and } b \text{ in CM}$$

Correspondence Analysis

Definition

$$\alpha_{ab} = \frac{P(a \cap b)}{P(a) \times P(b)} \text{ for all } a \text{ and } b \text{ in CM}$$

Deviation Matrix

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mn} \end{pmatrix}$$

Correspondence Analysis

Definition

$$\alpha_{ab} = \frac{P(a \cap b)}{P(a) \times P(b)} \text{ for all } a \text{ and } b \text{ in CM}$$

Deviation Matrix

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mn} \end{pmatrix}$$

Factorize

Subject the above matrix to decomposition

Correspondence Analysis

Definition

Require: A confrontation matrix N of size $I \times J$

- 1: Matrix sum: $n = \sum_{i=1}^I \sum_{j=1}^J N_{ij}$
- 2: Row masses(r): $r_i = \frac{N_{i.}}{n}, i = 1, 2, \dots, I$
- 3: Diagonal matrix: $D_r = \text{diag}(r_1, r_2, \dots, r_I)$
- 4: Column masses(c): $c_j = \frac{N_{.j}}{n}, j = 1, 2, \dots, J$
- 5: Diagonal matrix: $D_c = \text{diag}(c_1, c_2, \dots, c_J)$
- 6: Correspondence matrix: $P = \frac{1}{n}N$

Correspondence Analysis

Definition

Require: A confrontation matrix N of size $I \times J$

- 1: Matrix sum: $n = \sum_{i=1}^I \sum_{j=1}^J N_{ij}$
- 2: Row masses(r): $r_i = \frac{N_{i.}}{n}, i = 1, 2, \dots, I$
- 3: Diagonal matrix: $D_r = \text{diag}(r_1, r_2, \dots, r_I)$
- 4: Column masses(c): $c_j = \frac{N_{.j}}{n}, j = 1, 2, \dots, J$
- 5: Diagonal matrix: $D_c = \text{diag}(c_1, c_2, \dots, c_J)$
- 6: Correspondence matrix: $P = \frac{1}{n}N$
- 7: Standardized residuals: $A = D_r^{-\frac{1}{2}}(P - rc^T)D_c^{-\frac{1}{2}}$

Correspondence Analysis

Definition

Require: A confrontation matrix N of size $I \times J$

- 1: Matrix sum: $n = \sum_{i=1}^I \sum_{j=1}^J N_{ij}$
- 2: Row masses(r): $r_i = \frac{N_{i.}}{n}, i = 1, 2, \dots, I$
- 3: Diagonal matrix: $D_r = \text{diag}(r_1, r_2, \dots, r_I)$
- 4: Column masses(c): $c_j = \frac{N_{.j}}{n}, j = 1, 2, \dots, J$
- 5: Diagonal matrix: $D_c = \text{diag}(c_1, c_2, \dots, c_J)$
- 6: Correspondence matrix: $P = \frac{1}{n}N$
- 7: Standardized residuals: $A = D_r^{-\frac{1}{2}}(P - rc^T)D_c^{-\frac{1}{2}}$
- 8: Singular value decomposition: $A = U\Sigma V^T$
- 9: Standard coordinates of rows: $\Phi = D_r^{-\frac{1}{2}}U$
- 10: Standard coordinates of columns: $\Gamma = D_c^{-\frac{1}{2}}V$

Correspondence Analysis

Definition

Require: A confrontation matrix N of size $I \times J$

- 1: Matrix sum: $n = \sum_{i=1}^I \sum_{j=1}^J N_{ij}$
- 2: Row masses(r): $r_i = \frac{N_{i.}}{n}, i = 1, 2, \dots, I$
- 3: Diagonal matrix: $D_r = \text{diag}(r_1, r_2, \dots, r_I)$
- 4: Column masses(c): $c_j = \frac{N_{.j}}{n}, j = 1, 2, \dots, J$
- 5: Diagonal matrix: $D_c = \text{diag}(c_1, c_2, \dots, c_J)$
- 6: Correspondence matrix: $P = \frac{1}{n}N$
- 7: Standardized residuals: $A = D_r^{-\frac{1}{2}}(P - rc^T)D_c^{-\frac{1}{2}}$
- 8: Singular value decomposition: $A = U\Sigma V^T$
- 9: Standard coordinates of rows: $\Phi = D_r^{-\frac{1}{2}}U$
- 10: Standard coordinates of columns: $\Gamma = D_c^{-\frac{1}{2}}V$
- 11: Principal coordinates of rows: $F = \Phi\Sigma$
- 12: Principal coordinates of columns: $G = \Gamma\Sigma$
- 13: return F and G

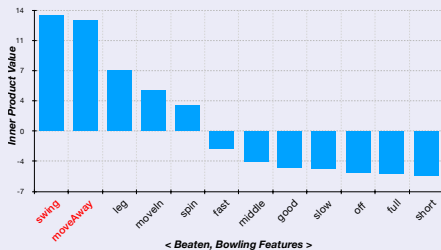
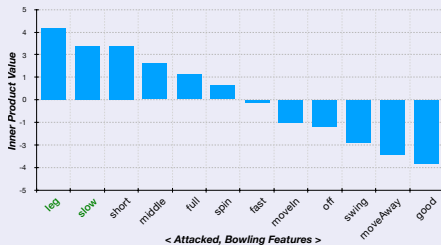
Correspondence Analysis

Definition

- F and G retains the batting and bowling feature factors
- Plot row vector of F and column vector of G on a two dimensional plot (biplot)
- The inner product of the above two vectors result in dependence between batting and bowling features
- The more the value, the higher the dependence and hence the stronger the rule is

Inner Products

F and G inner products



Rules

Smith's Strengths

- 1 Steve Smith **attacks** the deliveries that are bowled on the **leg** stump
- 2 He **attacks** the **slow** deliveries

Rules

Smith's Strengths

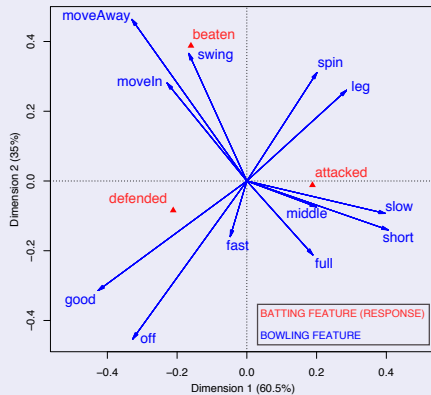
- ① Steve Smith **attacks** the deliveries that are bowled on the **leg** stump
- ② He **attacks** the **slow** deliveries

Smith's Weaknesses

- ① Steve Smith gets **beaten** on the deliveries that are **swinging**
- ② He gets **beaten** on the deliveries that are **moving away** from him

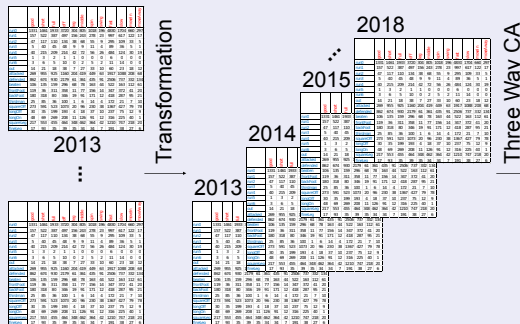
Biplot - Visualization

Smith's Response



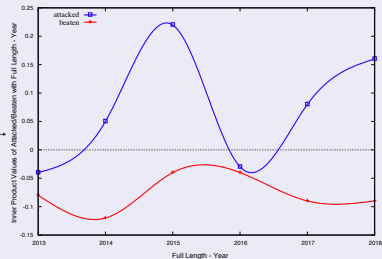
Confrontation Tensor

Several CM's and tensor



2018
Yearwise
Confrontation
Matrices

Three Dimensional
ConfrontationTensor



Temporal Changes in Strength
and Weakness Using Line Chart

Three way CA

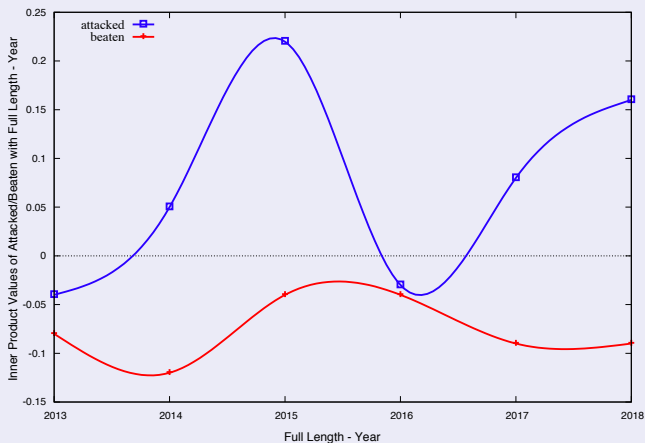
Algorithm

Require: A three dimensional confrontation tensor N of size $I \times J \times K$

- 1: Tensor sum: $n = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K N_{ijk}$
- 2: Correspondence tensor: $P = \frac{1}{n} N$
- 3: Univariate marginal relative frequencies: $p_{i..} = \sum_{j=1}^J \sum_{k=1}^K p_{ijk}$,
 $p_{.j.} = \sum_{i=1}^I \sum_{k=1}^K p_{ijk}$, and $p_{..k} = \sum_{i=1}^I \sum_{j=1}^J p_{ijk}$
- 4: Bivariate marginal relative frequencies: $p_{ij.} = \sum_{k=1}^K p_{ijk}$, $p_{i.k} = \sum_{j=1}^J p_{ijk}$, and
 $p_{.jk} = \sum_{i=1}^I p_{ijk}$
- 5: Deviations from the three way independence: $\pi_{ijk} = \frac{p_{ijk} - p_{i..} p_{.j.} p_{..k}}{p_{i..} p_{.j.} p_{..k}}$
- 6: Tucker3 tensor decomposition: $Tucker3(\Pi_P) = AG(B^T \otimes C^T)$
- 7: Standard coordinates of rows: $F' = A$
- 8: Standard coordinates of column-tubes: $H' = (B \otimes C)$
- 9: Principal coordinates of rows: $F = AG_{(P \times QR)}$
- 10: Principal coordinates of column-tubes: $H = (B \otimes C) G_{(QR \times P)}$
- 11: **return** F and H

Three way CA

Result



Similar Players

With Strength/Weakness

- F and G contains the batting and bowling related vectors for one player
- Obtain F and G matrices for all the players
- Take the strongest strength rule (or strongest weakness rule) of each player
- Concatenate the i^{th} row vector of F with j^{th} column vectors of G and notate it as FG_{ij}^p for player p
- Compute all the strongest strengths (weakness) of every player
- Each vector is a 31 dimensional vector (19+12)
- Employ t-SNE plot to visualize players with similar strengths/similar weaknesses

SNE Motivation

- Define **new notion** of **embedding**
- This notion is based on **probable neighbors**
- Optially preserve **neighborhood identity**

Basic SNE Algorithm

- For each data point i , and each **potential neighbor**, j compute **assymetric probabilities**

$$p_{j|i} = \frac{\exp(-d_{ij}^2)}{\sum_{k \neq i} \exp(-d_{ik}^2)}$$

where $d_{ij}^2 = \frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_i^2}$; σ_i is a chosen parameter

- Let \mathbf{y}_i 's be the data points **after** projecting \mathbf{x}_i in lower dimensional space
- Induced** probability $q_{j|i}$
- Point i picks point j as its neighbor is given by

$$q_{j|i} = \frac{\exp(-\|\mathbf{y}_i - \mathbf{y}_j\|^2)}{\sum_{k \neq i} \exp(-\|\mathbf{y}_i - \mathbf{y}_k\|^2)}$$

Objective Function

- SNE's objective is to find **that** neighborhood distribution in the lower dimension which **minimize** the **loss of information**
- Minimize the distance between the neighbor distributions for all i

$$\begin{aligned}\text{minimize } C &= \sum_i D_{KL}(P_i || Q_i) \\ &= \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}\end{aligned}$$

- An unconstrained optimization problem. Solve the formulation using gradient descent method.
- Gradient wrt \mathbf{y}_i is computed as

$$\frac{\partial C}{\partial \mathbf{y}_i} = 2 \sum_j (\mathbf{y}_i - \mathbf{y}_j)(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})$$

t-distribution SNE

- Uses **symmetric** version of SNE

$$\begin{aligned}\text{minimize } C &= KL(P||Q) \\ &= \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}\end{aligned}$$

where $p_{ii} = q_{ii} = 0$; and

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

- Computation of exponent is expensive;
- t-SNE replaces Gaussian distribution with Students' t-distribution with one degree of freedom

$$p(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

t-distribution SNE

- When $\nu = 1$ we have:

$$p(x) = \frac{1}{\pi(1+x^2)}$$

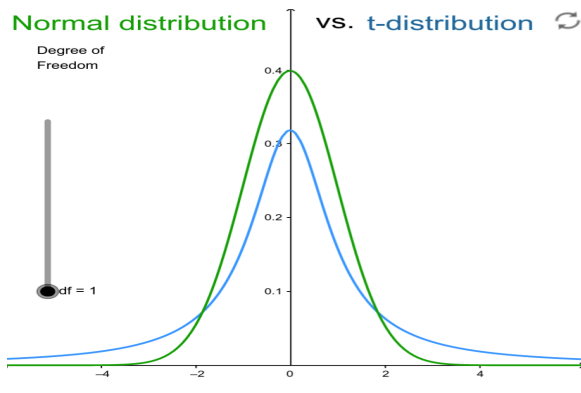


Figure 1: Gaussian Distribution and Students' t-distribution

t-distribution SNE

- Using this density function, the probabilities in the high dimensional and low dimensional space are defined as

$$p_{ij} = \frac{(p_{j|i} + p_{i|j})}{2n}$$

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

$$p_{ii} = q_{ii} = 0$$

- Let $d_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|$; $Z = \sum_{k \neq l} (1 + d_{kl}^2)^{-1}$ Re-write q_{ij} as:

$$q_{ij} = \frac{(1 + d_{ij}^2)^{-1}}{Z}$$

t-distribution SNE

- Objective is to **minimize** the KL divergence:

$$\begin{aligned}\text{minimize } C &= KL(P||Q) \\ &= \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}\end{aligned}$$

- If \mathbf{y}_i changes, only distances that changes are d_{ij} and d_{ji} for all j .
- Gradient of C wrt \mathbf{y}_i is given by

$$\begin{aligned}\frac{\partial C}{\partial \mathbf{y}_i} &= \sum_j \left(\frac{\partial C}{\partial d_{ij}} + \frac{\partial C}{\partial d_{ji}} \right) (\mathbf{y}_i - \mathbf{y}_j) \\ &= 2 \sum_j \frac{\partial C}{\partial d_{ij}} (\mathbf{y}_i - \mathbf{y}_j)\end{aligned}$$

t-distribution SNE - Gradient derivation

- The gradient $\frac{\partial C}{\partial d_{ij}}$ is derived from KL divergence definition:

$$\begin{aligned}
 \frac{\partial C}{\partial d_{ij}} &= - \sum_{k \neq l} p_{kl} \frac{\partial \log q_{kl}}{\partial d_{ij}} \\
 &= - \sum_{k \neq l} p_{kl} \frac{\partial (\log q_{kl} Z - \log Z)}{\partial d_{ij}} \\
 &= - \sum_{k \neq l} p_{kl} \left(\frac{\partial \log q_{kl} Z}{\partial d_{ij}} - \frac{\partial \log Z}{\partial d_{ij}} \right) \\
 &= - \sum_{k \neq l} p_{kl} \left(\frac{1}{q_{kl} Z} \frac{\partial (q_{kl} Z)}{\partial d_{ij}} - \frac{1}{Z} \frac{\partial Z}{\partial d_{ij}} \right) \\
 &= - \sum_{k \neq l} p_{kl} \left(\frac{1}{q_{kl} Z} \frac{\partial ((1+d_{kl}^2)^{-1})}{\partial d_{ij}} - \frac{1}{Z} \frac{\partial Z}{\partial d_{ij}} \right) \\
 &= 2 \frac{p_{ij}}{q_{ij} Z} (1 + d_{ij}^2)^{-2} - 2 \sum_{k \neq l} p_{kl} \frac{(1+d_{ij}^2)^{-2}}{Z}
 \end{aligned}$$

t-distribution SNE - Gradient derivation

- The gradient $\frac{\partial C}{\partial d_{ij}}$ is derived from KL divergence definition:

$$\frac{\partial C}{\partial d_{ij}} = 2 \frac{p_{ij}}{q_{ij}Z} (1 + d_{ij}^2)^{-2} - 2 \sum_{k \neq l} p_{kl} \frac{(1 + d_{ij}^2)^{-2}}{Z}$$

- Note that $q_{ij}Z = (1 + d_{ij}^2)^{-1}$
- First term of $\frac{\partial C}{\partial d_{ij}}$ is then given by

$$2p_{ij}(1 + d_{ij}^2)^{-1}$$

t-distribution SNE - Gradient derivation

- Multiply **second term** with q_{ij} in numerator and denominator

$$\begin{aligned}
 & \sum_{k \neq l} p_{kl} \frac{q_{ij} (1 + d_{ij}^2)^{-2}}{q_{ij} Z} \\
 &= \sum_{k \neq l} p_{kl} \frac{q_{ij} (1 + d_{ij}^2)^{-2}}{(1 + d_{ij}^2)^{-1}} \\
 &= \sum_{k \neq l} p_{kl} q_{ij} (1 + d_{ij}^2)^{-1} \\
 &= q_{ij} (1 + d_{ij}^2)^{-1} \sum_{k \neq l} p_{kl} \\
 &= 2q_{ij} (1 + d_{ij}^2)^{-1}
 \end{aligned}$$

t-distribution SNE - Gradient derivation

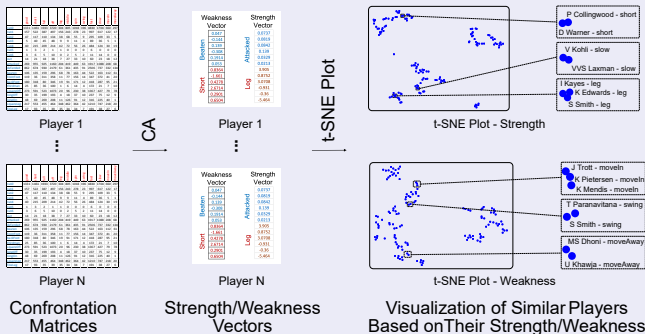
- Putting things together we have

$$\begin{aligned}\frac{\partial C}{\partial d_{ij}} &= 2p_{ij}(1 + d_{ij}^2)^{-1} - 2q_{ij} \left(1 + d_{ij}^2\right)^{-1} \\ &= 2(p_{ij} - q_{ij})(1 + d_{ij}^2)^{-1}\end{aligned}$$

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_j (p_{ij} - q_{ij})(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}(\mathbf{y}_i - \mathbf{y}_j)$$

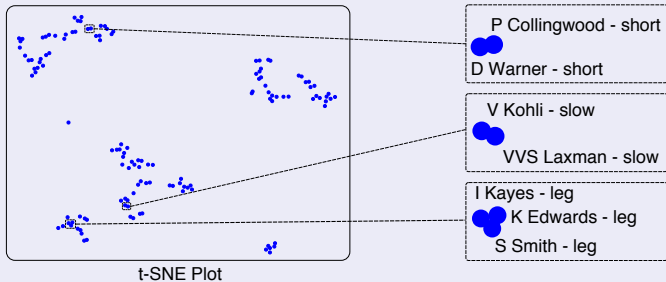
Similar Players

With Strength/Weakness



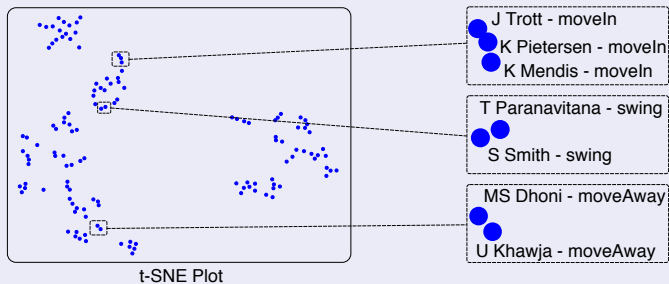
Similar Strengths

Strength Comparison



Similar Weaknesses

Weakness Comparison



Word Cloud Based Rule Construction

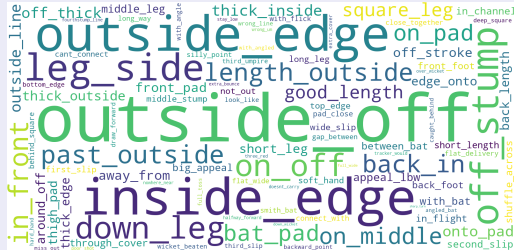
Are They Useful?

- The strength rule and weakness rules are constructed based on **the frequency** of occurrence
- Extract the commentary involving a specified feature (batting/bowling)
- Obtain the bigrams in the commentary
- Represent them in the **word cloud**

Strength Word Cloud



Weakness Word Cloud



Word Cloud Based Rule Construction

Observations

- Strength rule: Steve Smith attacks deliveries that are bowled on *outside off stump*
- Weakness rule: Steve Smith gets beaten on the deliveries that are bowled on *outside off stump*
- Contradictory rules are obtained using word clouds
- Domain experts immediately discards any rule that involve simple presence of *outside off stump* as majority of the balls bowled by bowlers are of this form
- Strength and weakness rule construction from our method do not find contradicting rules
- The proposed method also discards simple presence of *outside off stump* kind of delivers
- They are obtained only when there is statistically significant dependence with the batting feature

Rule Validation

Challenges

- There is no ground truth rule set available
- There is subjectivity involved in validating rules by humans
- As the relation strengths weaknes for a strength rule or weakness rule, bias and errors are inevitable
- There is no measurable loss function that can captures quantitatively the deviation from ground truth

Rule Validation

Methods

Direct Match the rules obtained by our method with expert given rules

In-direct Use statistical measures to validate the rules

Direct Rule Validation

Sanjay Manjrekar's Video Blog

- Published a video with ESPNCricInfo titled *What is Steve Smith's weakness?"* on 1st June, 2017
- One strenth rule of Steve Smith
- One wekness rule of Steve Smith are described
- Strength rule: video time 0.15"
- Weakness rule: video time 0.49"

Direct Rule Validation

Sanjay Manjrekar: Strengths of Steve Smith

Expert strength rule (Video time 0.15"): "Bowlers think he is very vulnerable to the ball pitching on the leg stump, ..., Bowlers tend to attack him on the stump (*leg or middle stump* line), but then his wonderful angle of the bat carves everything on the leg side. Gets lot of runs on the leg side".

Direct Rule Validation

Sanjay Manjrekar: Strengths of Steve Smith

Expert strength rule (Video time 0.15''): "Bowlers think he is very vulnerable to the ball pitching on the leg stump, ..., Bowlers tend to attack him on the stump (*leg or middle stump* line), but then his wonderful angle of the bat carves everything on the leg side. Gets lot of runs on the leg side".

Sanjay Manjrekar: Strengths of Steve Smith

In other words, Steve Smith scores lot of runs on the leg side for balls bowled on the leg stump (or middle stump).

Direct Rule Validation

Sanjay Manjrekar: Strengths of Steve Smith

Expert strength rule (Video time 0.15''): "Bowlers think he is very vulnerable to the ball pitching on the leg stump, ..., Bowlers tend to attack him on the stump (*leg or middle stump* line), but then his wonderful angle of the bat carves everything on the leg side. Gets lot of runs on the leg side".

Sanjay Manjrekar: Strengths of Steve Smith

In other words, Steve Smith scores lot of runs on the leg side for balls bowled on the leg stump (or middle stump).

CA Strengths of Steve Smith

"Steve Smith attacks deliveries that are bowled on the *leg stump*".

Direct Rule Validation

Sanjay Manjrekar: Weakness of Steve Smith

“Bowl *seamers* as much as possible”.

Direct Rule Validation

Sanjay Manjrekar: Weakness of Steve Smith

“Bowl *seamers* as much as possible”.

CA Weakness of Steve Smith

“Steve Smith gets beaten on the move away (*seam*) deliveries”.

Indirect Rule Validation

Method

- For every player, hold out the last one year of cricket commentary data from the player's career
- The above step results in two sets of data (i) training and (ii) testing
- Apply CA on (i) training data and (ii) test data
- Obtain F and G from (i) training data and (ii) test data
- Compare how close are two F_{trg} and F_{tst} using Procrustes analysis

Indirect Rule Validation

Procrustes Algorithm

Require: Training confrontation matrix $(N_{train})_{n \times k}$ and Test confrontation matrix $(N_{test})_{n \times k}$

1: Apply Correspondence analysis and obtain the row coordinates

$(F_{trg})_{n \times k}$ contains row coordinates for N_{train}

$(F_{tst})_{n \times k}$ contains row coordinates for N_{test}

Indirect Rule Validation

Procrustes Algorithm

Require: Training confrontation matrix $(N_{train})_{n \times k}$ and Test confrontation matrix $(N_{test})_{n \times k}$

1: Apply Correspondence analysis and obtain the row coordinates

$(F_{trg})_{n \times k}$ contains row coordinates for N_{train}

$(F_{tst})_{n \times k}$ contains row coordinates for N_{test}

2: Rescale and center the row coordinates

$$F_{trg_{scl}} = \frac{(I-P)F_{trg}}{\sqrt{\text{tr}[(I-P)F_{trg}^t(I-P)]}},$$

$$F_{tst_{scl}} = \frac{(I-P)F_{tst}}{\sqrt{\text{tr}[(I-P)F_{tst}^t(I-P)]}}, \text{ where}$$

I is a $n \times n$ identity matrix and P is a $n \times n$ matrix with all elements $1/n$

Indirect Rule Validation

Procrustes Algorithm

Require: Training confrontation matrix $(N_{train})_{n \times k}$ and Test confrontation matrix $(N_{test})_{n \times k}$

1: Apply Correspondence analysis and obtain the row coordinates

$(F_{trg})_{n \times k}$ contains row coordinates for N_{train}

$(F_{tst})_{n \times k}$ contains row coordinates for N_{test}

2: Rescale and center the row coordinates

$$F_{trg_{scl}} = \frac{(I-P)F_{trg}}{\sqrt{\text{tr}[(I-P)F_{trg}^t(I-P)]}},$$

$$F_{tst_{scl}} = \frac{(I-P)F_{tst}}{\sqrt{\text{tr}[(I-P)F_{tst}^t(I-P)]}}, \text{ where}$$

I is a $n \times n$ identity matrix and P is a $n \times n$ matrix with all elements $1/n$

3: Perform rotation by fixing the F_{trg} and rotating F_{tst} until the sum-of-squared residuals (Δ_{12}^2) is minimized between F_{trg} and F_{tst} . This is achieved by performing SVD on the matrix $F_{trg_{scl}}^t F_{tst_{scl}}$ i.e. $SVD(F_{trg_{scl}}^t F_{tst_{scl}}) = U\Sigma V^t$. The sum-of-squared residuals is:

$$\Delta_{12}^2 = 2(1 - (\text{trace } \Sigma)^2)$$

Indirect Rule Validation

Procrustes Algorithm

Require: Training confrontation matrix $(N_{train})_{n \times k}$ and Test confrontation matrix $(N_{test})_{n \times k}$

1: Apply Correspondence analysis and obtain the row coordinates

$(F_{trg})_{n \times k}$ contains row coordinates for N_{train}

$(F_{tst})_{n \times k}$ contains row coordinates for N_{test}

2: Rescale and center the row coordinates

$$F_{trg_{scl}} = \frac{(I-P)F_{trg}}{\sqrt{\text{tr}[(I-P)F_{trg}^t(I-P)]}},$$

$$F_{tst_{scl}} = \frac{(I-P)F_{tst}}{\sqrt{\text{tr}[(I-P)F_{tst}^t(I-P)]}}, \text{ where}$$

I is a $n \times n$ identity matrix and P is a $n \times n$ matrix with all elements $1/n$

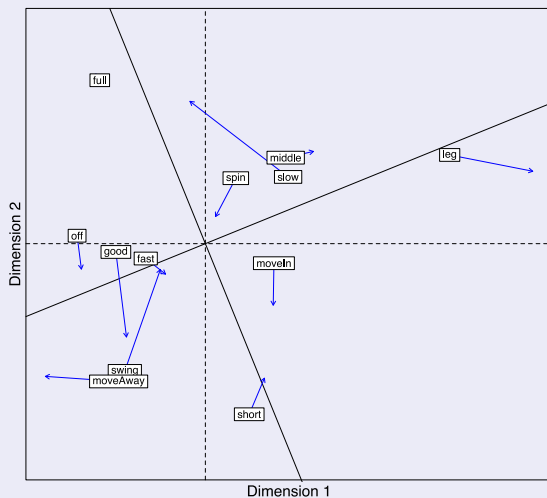
3: Perform rotation by fixing the F_{trg} and rotating F_{tst} until the sum-of-squared residuals (Δ_{12}^2) is minimized between F_{trg} and F_{tst} . This is achieved by performing SVD on the matrix $F_{trg_{scl}}^t F_{tst_{scl}}$ i.e. $SVD(F_{trg_{scl}}^t F_{tst_{scl}}) = U\Sigma V^t$. The sum-of-squared residuals is:

$$\Delta_{12}^2 = 2(1 - (\text{trace } \Sigma)^2)$$

4: **return** Sum-of-squared residual Δ_{12}^2

Indirect Rule Validation

Procrustes Algorithm



Indirect Rule Validation

Procrustes Results

Table 2: Procrustes Analysis

Batsman	$\#Balls_{train}$	$\#Balls_{test}$	Δ_{12}^2
Joe Root	8376	2421	0.09
Dimuth Karunaratne	4898	1831	0.11
Steve Smith	9250	1948	0.17
Cheteshwar Pujara	7947	1546	0.27
Dean Elgar	4475	2597	0.28
Virat Kohli	8085	1482	0.30
David Warner	6999	1557	0.47
Kane Williamson	10165	439	0.47

Summary

Unstructured data visualization in sports

- Sports visualization predominantly centers around box score data and tracking data
- For the first time, the utility of the unstructured data is demonstrated
- The analysis is focused on advanced cognitive computation namely weakness and strength rule identification
- Visualization of the obtained rules for each player through biplots is presented
- In this work, CA is shown to be a suitable method for computation of such tasks
- The constructed rules are validated using the intrinsic and extrinsic methods
- The visualizations will be helpful for analysts, coaches and team management in building game strategies

Disclaimer

Weaknesses

- We, as team of collaborators, acknowledge the achievements of each and every individual player
- We have high regard to each and every player of this game
- This work in no way should be viewed as our criticism with respect to these individuals
- Our findings must be viewed from the prism of scientific understanding and technique improvization, if any.

Thank You!

Everyone Who Helped

- Thank you for your valuable time and listening
- Thank [Dr. Amit Awekar](#) and [Dr. Ashish Anand](#) for providing this opportunity
- Thank the organizers of the course [Deep Learning for NLP](#)
- Thank all my collaborators
- Questions welcome