

Sports Data Mining



TECHNISCHE
UNIVERSITÄT
DARMSTADT

**Data Sources for Sports -
Daten und Datenquellen zur Nutzung für
DataMining - Michael Gleser**



“Football is a simple game; 22 men chase a ball for 90 minutes and at the end, the Germans always win.”

- Gary Lineker

Daten...

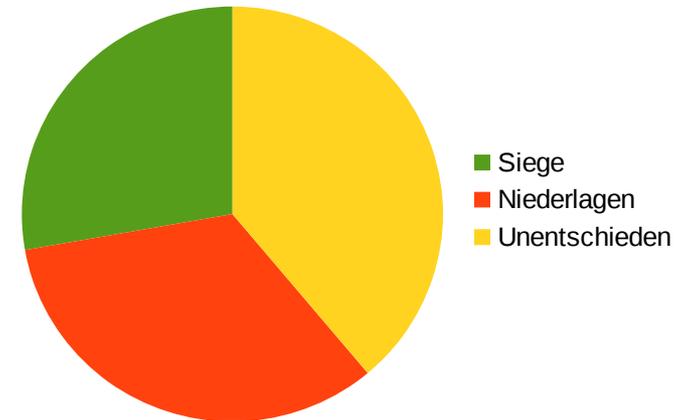


VS



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- Begegnungen im Fußball zwischen Deutschland und Italien
 - Insgesamt 18 Begegnungen (Pflichtspiele + Freundschaftsspiele)
 - 5 Siege Deutschland
 - 6 Siege Italien
 - 7 Unentschieden



Daten...

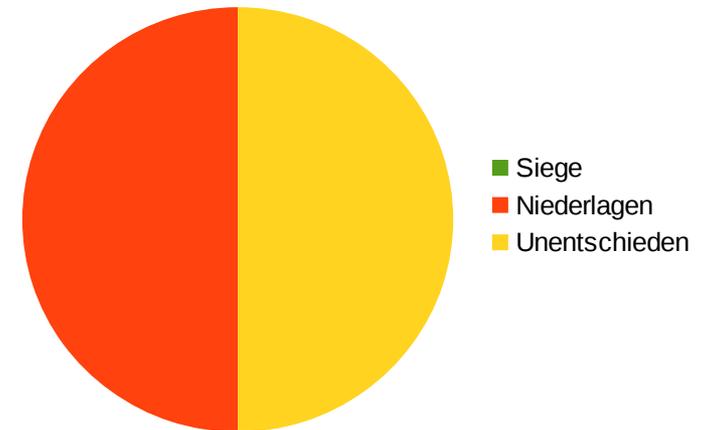


VS



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- Begegnungen im Fußball zwischen Deutschland und Italien
 - Jedoch nur 8 Begegnungen bei Turnieren
 - 0 Siege Deutschland
 - 4 Siege Italien
 - 4 Unentschieden

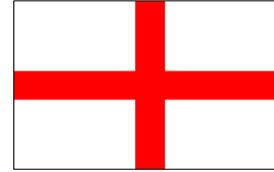


→ Deutschland sollte in einer K.O.-Runde besser nicht auf Italien treffen

Daten...



VS



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- Elfmeterschießen zwischen Deutschland und England
 - Elfmeterschießen wurde 1971 eingeführt
 - England hatte bisher 7 Elfmeterschießen bei Turnieren
 - 1 Sieg und 6 Niederlagen
 - Deutschland hatte bisher 6 Elfmeterschießen bei Turnieren
 - 5 Siege und 1 Niederlage
- England vs. Deutschland im Elfmeterschießen
 - 2 Siege Deutschland
 - 0 Siege England

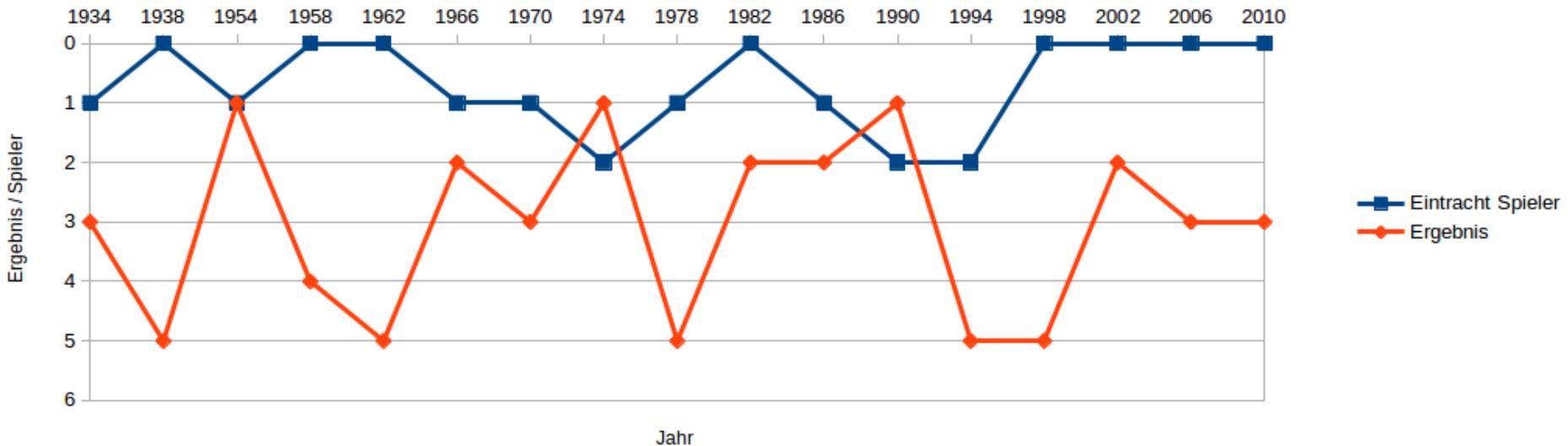
→ Elfmeterschießen gegen England scheinen eine sichere Sache zu sein

Daten...



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- Spieler von Eintracht Frankfurt in der deutschen Nationalelf bei Weltmeisterschaften



Daten...

- Für viele Sportarten sind Statistiken essentiell um den Sieger eines Wettbewerbes zu ermitteln:

TABELLE	HEIMTABELLE	AUSWÄRTSTABELLE	HINRUNDENTABELLE	RÜCKRUNDENTABELLE	FORM								
Platz	Club					Spiele	S*	U*	N*	Tore	TD*	Punkte	
1	 FC Bayern München					34	29	3	2	94:23	+71	90	CL*
2	 Borussia Dortmund					34	22	5	7	80:38	+42	71	CL*
3	 FC Schalke 04					34	19	7	8	63:43	+20	64	CL*
4	 Bayer 04 Leverkusen					34	19	4	11	60:41	+19	61	CL* Qual.
5	 VfL Wolfsburg					34	18	6	10	63:50	+13	60	EL*
6	 Borussia Mönchengladbach					34	16	7	11	59:43	+16	55	EL* Qual.
7	 1. FSV Mainz 05					34	16	5	13	52:54	-2	53	EL* Qual.
8	 FC Augsburg					34	15	7	12	47:47	0	52	
9	 1899 Hoffenheim					34	11	11	12	72:70	+2	44	

Daten...

- Die Halbfinal-Teilnehmer des BCS wurden bis 2013 anhand eines Rankings ermittelt. Hierfür wurden Statistiken herangezogen.
- Seit 2014 gibt es ein Expertengremium, welches die Teilnehmer bestimmt



Welche Daten existieren?



- Mannschaftsdaten
 - Spielergebnisse (+Datum)
 - Aufstellung
 - Akkumulierte Spielerleistungen

- Individualdaten
 - Pro Spieler
 - Meist schwer erfassbar

Welche Daten existieren?

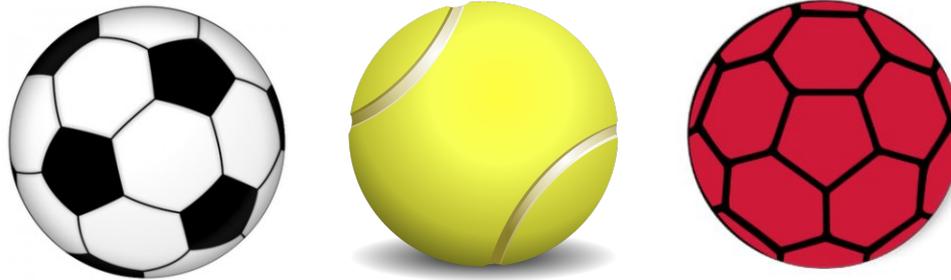
- Fokus der Bücher auf die großen amerikanischen Sportarten:



- Besonders bei den amerikanischen Sportarten lassen sich sowohl Individual- als auch Mannschaftsleistungen gut statistisch erfassen.

Welche Daten existieren?

- In Europa besonders beliebte Sportarten:



- Probleme beim Fußball / Handball
 - Individuelle Leistung der Sportler schwer erfassbar
 - Statistiken meist nicht aussagekräftig für Leistung eines Spielers

Welche Daten existieren?

- Beispiele für Individualdaten pro Sportler (nach Sportarten):
 - **Fußball:**
 - Pässe (kurz)
 - Pässe (lang)
 - Zweikämpfe
 - Fouls
 - Laufdistanz
 - Verwarnungen
 - **Football:** (für QB)
 - Completions
 - Pass Attempts
 - Yards per Pass
 - Longest Pass Play
 - Interceptions
 - Passing Touchdown

Welche Daten existieren?

- Beispiele für Individualdaten pro Sportler (nach Sportarten):
 - **Baseball:**
 - Runs (OF)
 - Hits (OF)
 - Homerun (OF)
 - Games (DF)
 - Assists (DF)
 - Errors (DF)
 - **Basketball:**
 - Points (OF)
 - Assists (OF)
 - Rebounds (OF)
 - Rebounds (DF)
 - Steals (DF)
 - Blocks (DF)

Wer sammelt Daten? - Privates Interesse

- Sportbegeisterte Personen mit Interesse an Lieblingsmannschaft
- Auf Daten basierende Hobbies
 - Fantasy Leagues

**→ Daten meist auf persönlichen Websites,
jedoch öffentlich zugänglich**

Wer sammelt Daten? - Kommerzielles Interesse

- Wettmarkt & Quotenberechnung
 - Riesiger Markt (approx. Volumen > 500 Mrd US-\$)¹
 - Berechnung von Quoten für Wetten müssen einen Gewinn für die eigene Gesellschaft abwerfen damit sich das Angebot lohnt
 - Berechnung InHouse
 - Wettanbieter muss selbst Daten sammeln
 - Aufwendig und teuer
 - Outsourcing der Quotenberechnung
 - Spezialisierte Firmen haben große Datenbasis
 - Berechnen die Quoten direkt für die Wettanbieter
 - Sportdaten als Geschäftsmodell²

→ Daten nicht frei verfügbar, da Geschäftsgrundlage

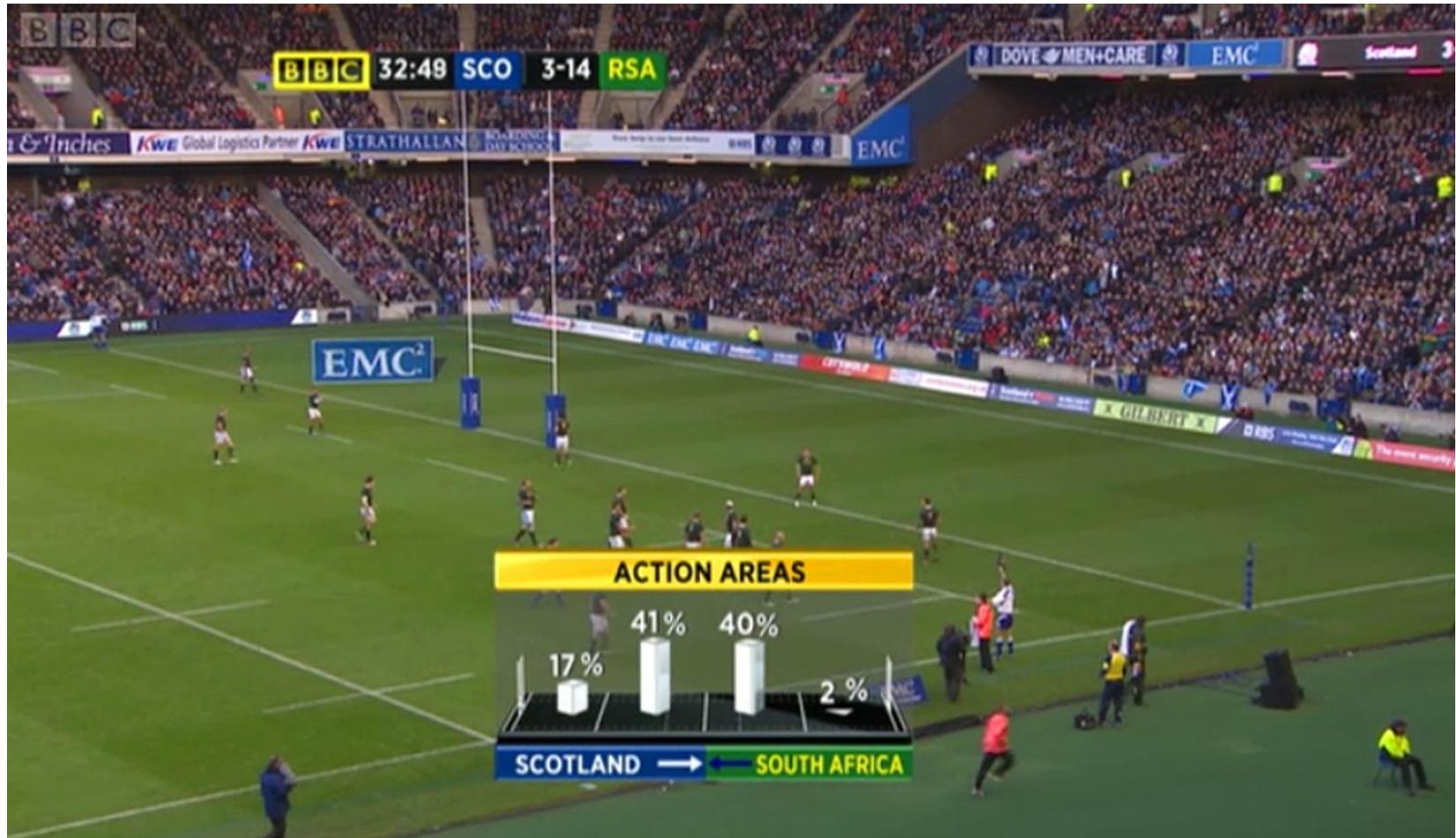
¹ <http://www.bbc.com/sport/0/football/24354124>

² <http://www.sportradar.com/>

Wer sammelt Daten? - Sportliches Interesse

- Professionalisierung des Sports
 - Taktische und sportliche Analyse und Verbesserung basierend auf Datenanalyse
 - Team-Geek
 - Echtzeitanalyse und Echtzeitreaktion
 - Mehrere kommerzielle Anbieter für Spielanalyse
 - Bspw. für Fußball
 - IMPIRE AG
 - Opta
 - Matchanalysis.com
- Analyse und Datenbereitstellung meist kostenpflichtig, da hoher technischer Aufwand zur Erfassung notwendig.**

Wer sammelt Daten? - Sportliches Interesse



Wer sammelt Daten? - Sportliches Interesse



Wer sammelt Daten? - Sportliches Interesse



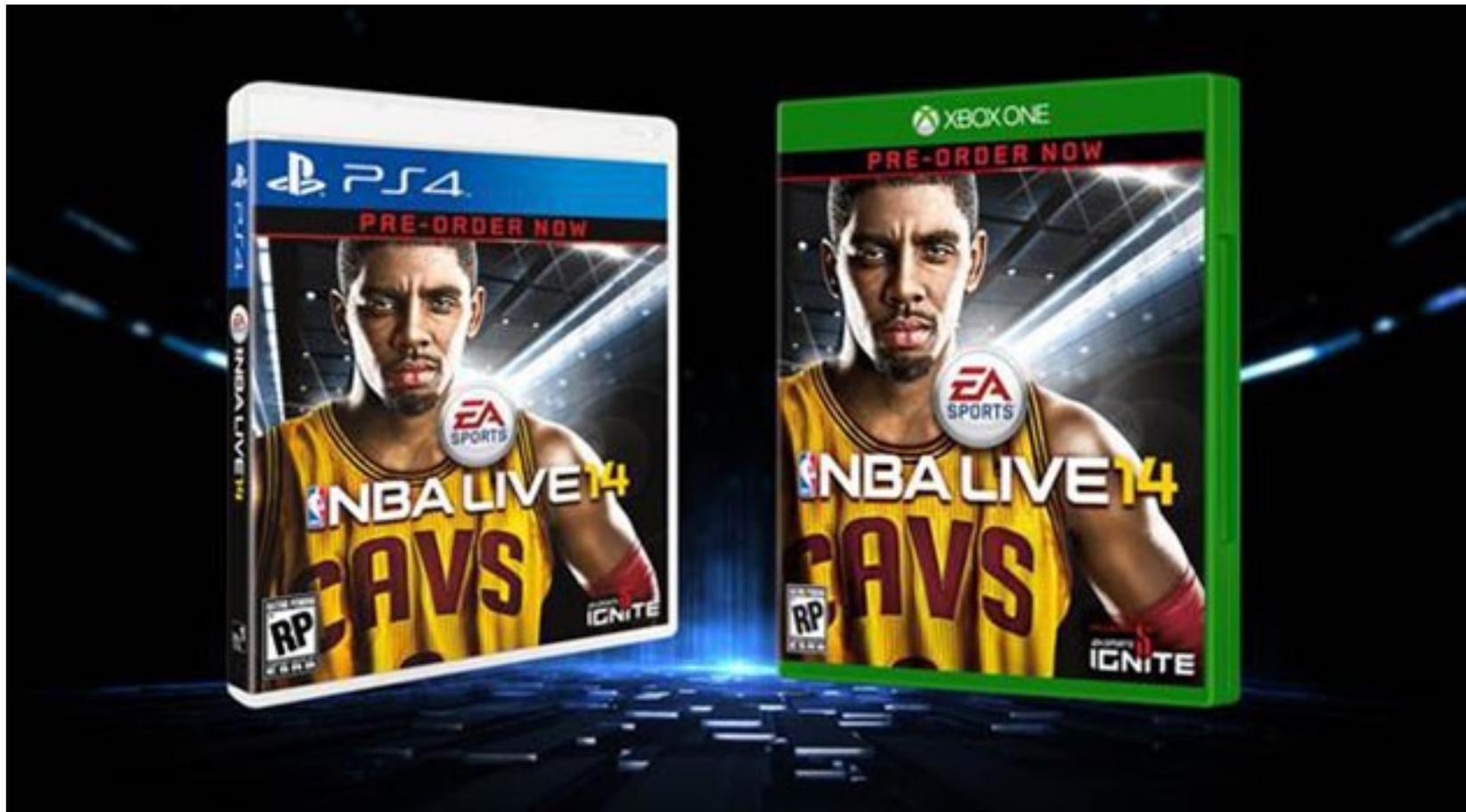
TECHNISCHE
UNIVERSITÄT
DARMSTADT



Wer sammelt Daten? - Sportliches Interesse



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Wer sammelt Daten? - Akademisches Interesse

- Sport als Anwendungsfach für Datenanalyse
- Beispiele des Buches „Who's #1 – The Science of Rating and Ranking“ basieren auf Daten von Prof. Massey

→ **Daten frei abrufbar unter:**

- <http://www.masseyratings.com/data.php>

Wer sammelt Daten? - Öffentliches Interesse

- Besonders in den USA haben sich mehrere sportspezifische Gesellschaften gegründet
 - Sammeln aller verfügbaren Daten der Sportart
 - Veröffentlichung von Analysen
 - Bereitstellung der Daten für Mitglieder zur Analyse
- **SABR - Society for American Baseball Research**
 - Sammelt und verwaltet alle relevanten Baseball Daten
 - Analysiert die Daten nach eigenen entwickelten Methodiken (SABRmetrics)
- **APBR - Association for Professional Basketball Research**
- **PFRA - Professional Football Researchers Association**

Wer sammelt Daten? - Öffentliches Interesse

- Übergreifende Gesellschaften für Sportdaten, stellen Daten für öffentliche Einrichtungen (z.B. IOC) und ihre Mitglieder bereit
 - **IACSS - International Association on Computer Science in Sport (IACSS)**
 - **IASI - International Association for Sports Information**

Wie kommt man an die Daten?

- Eine optimale Datenquelle...
 - ... besitzt qualitativ hochwertige Daten
 - ... ist maschinenlesbar
 - ... kann kostenlos abgerufen werden
 - ... ermöglicht Verknüpfungen zu anderen Datensätzen

Wie kommt man an die Daten?

- In der Realität jedoch sind Datenquellen...
 - ... kostenpflichtige Datenabonnements
 - ... unstrukturiert auf Websites
 - ... ohne standardisierte Form
 - ... nur für Mitglieder von Vereinigungen verfügbar

Wie kommt man an die Daten?

- Glücklicherweise jedoch...
 - **Massey's Datenbasis**
 - Geeignet für Datenanalysen rund um die amerikanischen Sportarten
 - Umfangreiche Datenbasis
 - Frei abrufbar
 - In tabellarischer Form

Wie kommt man an die Daten?

- Alternativ...

- **Scraping**

- „Abgreifen“ von Daten bestimmter Websites mithilfe von Skripten
- Praktische Fähigkeit für das beschaffen möglichst umfangreicher Daten
- Es existieren bereits kommerzielle Anbieter, die Scraping-Software anbieten

- BigData und In-Memory Computing in Sports
 - Aufzeichnung und automatische Bildanalyse
 - Sensoren
 - Hohe Datenqualität und Datendichte
 - Rechenintensive Auswertung

Bereits heute im Einsatz:

- Einsatz von SAP HANA im Motorsport um in Echtzeit Parameter des Rennwagens zu steuern¹

¹ <http://blog.sap-tv.com/2012/09/cnn-on-mclaren-f1-racing-team-using-sap/>