EXPERT
KNOW-HOW

Big Data, The Megatrends of ICT Industry!
메이저리그로 알아보는 알기 쉬운 빅데이터

+ **김희동** KT스카이라이프 대리

마이클 루이스의 '머니볼'

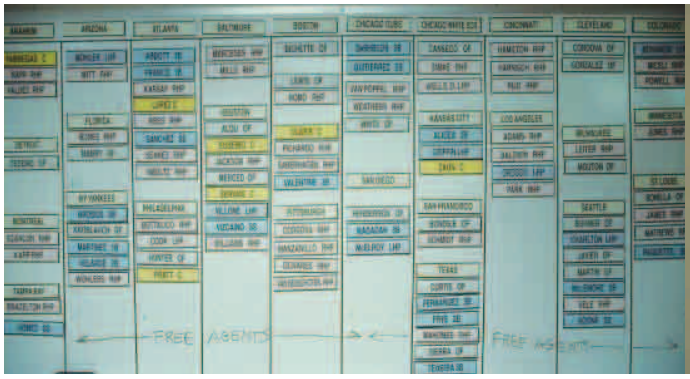
브래드 피트 주연의 '머니볼'

연일 ICT의 최대 화두는 빅데이터로 정부(공공부문-정부3.0)는 물론 많은 기업(민간부문-맵리듀스, 스마트 인사이트 등)들이 저렴한 하드웨어를 병렬로 연결하는 하둡 파일 시스템인 HDFS(Hadoop Distributed File System)와 분산 저장된 데이터를 연관성 여부로 분류하여 처리하는 맵(Map) 작업, 중복데이터를 제거하여 원하는 데이터를 추출하는 리듀스(Reduce) 작업을 통칭하는 MapReduce로 이루어진 하둡(Hadoop), 그리고 오픈소스 기반 분석엔진 'R' 등 대용량의 정보를 수집, 저장, 분석, 표현하기 위한 시스템 구축에 열을 올리고 있습니다.

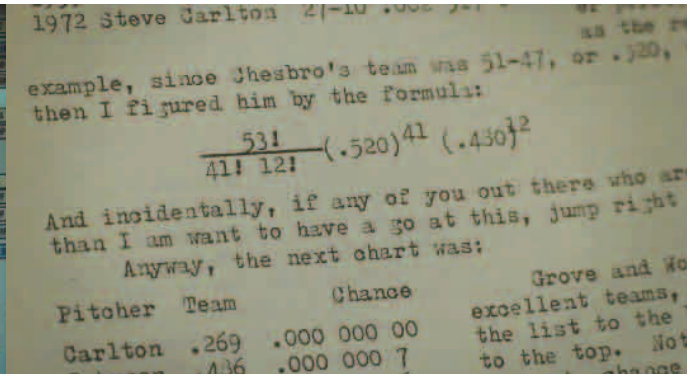
본고에서는 딱딱한 빅데이터의 정의, 개념, 구성 보다는 우리나라의 추신수, 류현진 선수가 활약하고 있는 메이저리그를 통해 구단은 어떤 식으로 빅데이터를 활용하고 있으며 MLB 사무국의 합자회사인 MLB Advanced Media를 통해 야구와 빅데이터, 기업과 빅데이터를 조명해 보도록 하겠습니다.

Moneyball, New Start-Up of Big Data!

머니볼은 2002년 오클랜드 애슬레틱스의 오직 데이터 기반의 선수기용에 의해 메이저리그 역사상 20연승을 이룬 단장 빌리 빈(Billy Beane)의 일화를 다룬 작품으로 2003년 발간된 마이클 루이스(Michael Lewis)의 Moneyball(subtitle-The Art of Winning Unfair Game)이 2011년 브래드 피트(Brad Pitt) 주연의 Moneyball로 영화화되어 국내에 널리 알려지게 되었습니다. 메이저리그 팀 최다 연승 기록은 1916년 뉴욕 자이언츠(지금의 샌프란시스코 자이언츠 전신)가 세운 26연승! 하지만 이때는 이른바 데드볼(Dead Ball) 시대로 공의 반발력이 떨어지며 스핏볼(Spit Ball-침과 같은 이물질을 공에 바르는 것이 허용)이 통용되어 통상 기록은 1920년 이후의 라이브볼(Live Ball) 시대부터 인정되어 2002년 오클랜드 애슬레틱스가 이룬 20연승이 아메리칸리그 공식 기록입니다.



구단별 자유계약 선수명단, LA 박찬호도 있음



철저한 Cost-Effective 기반, 통계적 분석데이터

Moneyball, The Art of Winning an Unfair Game! ($114,457,768 vs $39,722,689)

위 금액은 2002년 당시 뉴욕 양키스와 오클랜드 애슬레틱스의 선수 연봉 총합으로 전형적인 빅리그 내에서의 빅마켓(Big Market)과 스몰마켓 구단의 현실을 극단적으로 보여주고 있습니다. 하지만 그해 양 팀 모두 103승 59패라는 놀라운 성적으로 각각 동부지구와 서부지구 1위를 하며 포스트 시즌에 진출하게 됩니다. 오클랜드는 무슨 마법으로 기적과 같은 20연승 기록을 세우며 최고 승률을 이뤘을까요?

그 해답은 당시 구단주를 맡고 있던 빌리빈이 하버드 출신 경제학도 폴 데포데스터를 영입하며 홈런, 타율, 타점 등 흥행요소만 중시하는 기존 선수 영입 드래프트(Draft)를 버리고 야구 경기의 승리는 지난 데이터를 분석한 결과 홈런, 타율이 아닌 출루율과 장타율에 있다는 것에 착안하여 승리와 밀접한 출루율과 장타율의 합인 OPS(On-base Plus Slugging Percentage)를 기준으로 각각 선수들의 데이터를 수집, 비교, 분석하여 적재적소에 선수를 기용함으로써 경이로운 기록을 작성하게 됩니다. 어찌 보면 스몰마켓 구단이 적은 구단 운영금으로 취할 수 있는 유일한 방법이었는지도 모릅니다.

메이저리그의 어록 가운데 "홈런왕은 캐딜락을 타고, 타격왕은 포드를 탄다."는 말이 있을 정도로 홈런타자의 연봉은 천정부지로 올라 오클랜드와 같은 구단이 감당하기엔 무리일 수밖에 없습니다. 그에 반해 상대적으로 저평가된 출루율, 장타율 위주의 선수를 고용함으로써 철저하게 저비용 고효율(Cost-Effective) 전략으로 4년 연속 포스트 시즌에 진출하는 쾌거를 이룩하게 됩니다. 하지만 아메리칸리그 챔피언쉽 시리즈(ALCS) 우승은 물론 월드 시리즈에 오르지 못해 결국 실패한 전략이란 비판도 있지만 머니볼 이론이 포스트 시즌(홈런타자, 최정상 원-투 펀치 투수 보유 팀 유리)과 같은 단기전(5 또는 7경기)이 아닌 정규시즌(162경기)을 토대로 나온 데이터(*세이버메트릭스)라 성공 여부는 여러분께 맡기도록 하겠습니다.

세이버메트릭스(Sabermetrics)
SABR(Society for American Baseball Research)을 세이버(Saber)라고 읽고 측정기준, 계량화란 뜻의 접미어 metrics가 결합한 합성어로 객관적 증거, 즉 경기 중 기록된 통계를 통해 야구를 분석한다는 뜻으로 SABR는 통계를 뛰어넘어 야구를 분석하기 위한 수학적 도구의 용법(Usage of mathematical tools)으로 정의됨. 회원으로 빌제임스(Bill James)가 유명하며 세이버메트리션이라 부름.

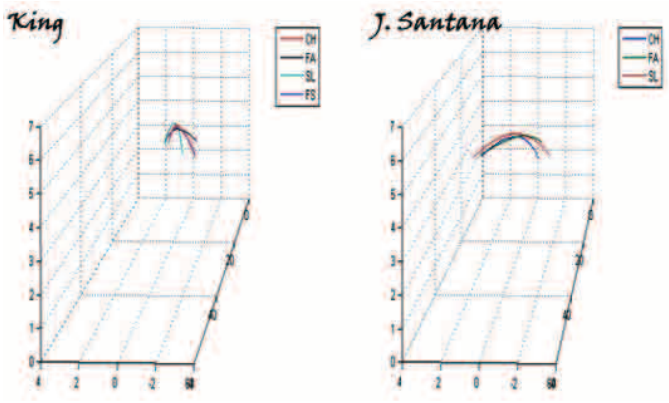
The Evolution of Big Data! (Structured vs Unstructured)

2009년 11월 국내 돌풍을 몰고 왔던 스마트폰과 같은 디지털 기기의 보급 확대로 2010년 인류가 쏟아낸 정보량은 1.2제타바이트(1조 2천억 기가바이트), 2011년에는 1.8제타바이트로 그전까지 인류가 생성한 총 데이터양이 5액사바이트(10억 기가바이트)인걸 감안하면 실로 그 양이 엄청나다는 것을 알 수 있습니다. 이는 단지 사용하는 기기의 변화뿐만이 아니라 다루는 데이터의 형태가 바뀌고 있다는 것을 의미합니다. 즉 과거에는 단순 사무정보(2진 데이터의 숫자로 표현 가능)와 같이 고정된 필드에 저장된 정형데이터(Structured)였다면 현재는 이미지, 음성, 동영상 스트리밍 등 숫자데이터와 달리 구조나 형태가 복잡하고 다양한 비정형데이터(Unstructured)가 주를 이루고 있습니다. IDC 조사에 의하면 매년 정형 데이터는 24% 증가하고 비정형 데이터는 55% 증가하여 2010년 정형데이터와 비정형데이터의 비율이 8:2이었다면 앞으로는 점점 비정형 데이터가 늘어나 1:9까지 될 것이란 전망이 나오고 있습니다. 이런 변화는 메이저리그가 다루는 데이터에도 그대로 적용되어 과거

단순 숫자로만 기록되었던 세이버메트릭스는 과학기술 발달에 힘입어 카메라, 분석 장비 등을 활용한 이미지, 동영상으로 저장되어 타자의 배트 스윙 메커니즘은 물론 투수의 공 궤적을 분석하는데 이용되고 있습니다.

	AL	NL
League avg.	4.82	4.53
Pitcher	Greinke	Lincecum
FIP	2.33	2.34
ERA	2.16	2.48
adjusted RA	2.53	2.52
Total Innings	229.33333	225.33333
Game	33	32
IP/G	6.95	7.04
Win(%)	0.756	0.734
Park Factor	1	1.01
WAR(pitching)	9.6	8.9
WAR(batting)	0	-0.7
WAR(overall)	9.6	8.2

2009년 양대 리그 사이영상 수상자 비교



King Felix와 Johan Santana의 구종별 궤적 Pitch f/x

표는 대체선수대비 승리기여도(기대승수)를 나타내주는 WAR(Wins Above Replacement)를 산출하는데 필요한 데이터입니다. 쉽게 말해 2009년 그레인키를 영입했다면 팀의 +9.6승이 추가될 수 있음을 의미합니다. 물론 정확한 WAR을 산출하기 위해서는 Runs per Win, Replacement Level 등과 같은 자료가 더 필요하지만 그러면 너무 세이버메트릭스에 치중하게 되는 것 같아 생략하였습니다. 리그 평균 방어율과 수비 조건에 무관한 독립적인 피칭(Fielding Independent Pitching)을 뜻하는 FIP(자책점 기준인 ERA와 달리 피홈런, 사사구, 탈삼진 등을 고려한 수치) 그리고 특정 구장을 홈으로 사용하는데 따르는 유, 불리를 보정해주기 위한 Park Factor를 고려하여 조정 평균 방어율(adjusted Run Average) 등으로 최종 WAR를 계산하게 됩니다. 류현진 선수는13승 6패 3.07/1.24/.257/173.0(ERA/WHIP/AVG/IP)로 2.9의 WAR를 기록하고 있습니다.(9월 10일 기준) 대부분의 선수가 -1~1 사이에 분포하는 걸 보면(6%의 선수만이 4를 넘음, 팀 내 투수인 커쇼의 경우 5.6) 데뷔 첫해 우수한 성적을 기록하고 있습니다.(신인왕은 말린스의 호세 페르난데스가 4.2로 유력) 야수인 경우에도 리그 평균과 비교하여 타격, 수비, 주루 등으로 WAR를 산출하는데 8월 당시 추신수 선수는 .290/.424/.469/.893(AVG/OBP/SLG/OPS)의 기록으로 4.7의 WAR를 기록하였습니다. 참고로 올해를 끝으로 FA(자유계약선수)를 맞는 추신수 선수의 경우 2370만 달러(1WAR를 대략 450만 달러 추정)의 값어치를 하고 있어 6년 9000만, 7년 1억 달러 내외로 예상됩니다. 물론 최종 기록에 따라 WAR는 달라지므로 앞으로 많은 분발이 필요한 시기입니다.

왼쪽 그림은 아메리칸리그 최고 투수인 펠릭스 에르난데스(2010년 사이영상 수상)와 서클 체인지업의 대명사로 불리는 요한 산타나(2004, 2006년 사이영상 수상)의 구종별 궤적을 보여주는 Pitch f/x(투구 추적 시스템)입니다. 구종별 상/하(V-Breaking), 좌/우(H-Breaking)의 값을 부여하여 상/하의 경우 아래로 변화하면 -로, 위로 변화하면 +값을 가지며, 좌/우의 경우는 홈플레이트를 기준으로 타자 몸쪽으로 변화하면 -로, 바깥쪽으로 변화하면 +의 값을 가지며 클수록 변화되는 무브먼트가 크다는 것을 알 수 있습니다. 그렇다면 누가 어떻게 Pitch f/x를 만드는 걸까요?

시기는 2006년 가을 MLB의 포스트시즌 축제가 한창일 무렵 스포트비전(Sportvision)이란 민간기업이 Pitch f/x란 기술(미사일 추적 시스템에서 착안)을 새롭게 고안하여 MLB 사무국에 승인을 받아 경기가 열리는 각 구장에 투수의 피치를 분석(스피드, 궤적 등)할 수 있는 카메라를 홈플레이트 뒤쪽과 내야, 외야에 설치하기 시작합니다. 시속 150km의 공이 18.44m의 거리를 날아오는데 걸리는 시간은 대략 0.44초! 이사이 각 카메라들은 초당 40프레임을 중앙 투구 추적 시스템(Central Pitch Track System)에 전송하게 되며 이를 실시간으로 분석하여 TV 중계는 물론 각 구단 분석실 및 전송매체에 보내지게 됩니다. 한 세트 가격이 10억이 넘어 아직 모든 구장에 설치가 안 돼(2010년 기준 잠실, 사직, 광주, 문학 설치) 국내에서는 스토커, 저그와 같은 스피드 건(Speed Gun)에 의존을 합니다. 하지만 스피드 건은 움직이는 공에 초음파를 발사하여 돌아오는 반사파의 진동수(주파수 변화)를 분석하여 측정 지점의 속도만 알 수 있는 1차원 시스템이라 측정 장소, 장비 특성에 따른 오차가 존재합니다.

스포트비전의 HIT f/x, Spray& Hit Angle

클리프 리와 맷 캐인의 포구반경, 5인치 차이

하지만 Pitch f/x는 세 개의 카메라가 공이 투수의 손을 떠나 포수의 미트에 들어갈 때까지 매 순간을 기록하기 때문에 초속, 중속, 종속의 변화는 물론 공의 궤적까지 분석할 수 있는 3차원 시스템입니다. 스포트비전의 도전은 투수뿐만이 아닌 타자를 분석하는데 까지 이어져 배트의 스윙속도는 물론 공의 컨택 포인트(*스윗스팟 여부), 배트를 맞고 난 후의 상승각, 속도, 필드 방향, 궤적 등을 기록하고 있습니다. 이 밖에도 2010 시즌부터는 포수의 콜에 얼마나 제구가 되는지를 알 수 있는 Command f/x를 도입하고 있습니다.

스윗스팟(Sweet Spot)
오디어 용어에서 청취 시 가장 이상적인 음을 들을 수 있는 위치(좌, 우 스피커와 정삼각형을 이루는 꼭짓점)를 뜻하는 스윗스팟은 탁구, 골프, 야구 등과 같은 스포츠 경기에서 가장 볼을 멀리 보낼 수 있는(최대 반발계수) 지점을 뜻함. 야구의 경우 배트 끝에서부터 5~10cm 정도

Leading Edge In Big Data! MLB Advanced Media (MLBAM)

메이저리그 사무국(the commissioner's office)과 Limited Partnership(합자회사)으로 2000년 6월 설립된 MLBAM은 미국 뉴욕에 본사를 두고 있으며 MLB 공식 웹 사이트와 이를 경유하여 총 30개 모든 구단의 웹 사이트를 운영하고 있습니다. 무료 가입자들은 기본적인 정보만 열람할 수 있으며 유료 가입자에게는 뉴스, 순위, 기록은 물론 실시간 오디오와 비디오 방송시청이 가능하며 특히 Pitch f/x와 같은 정보를 제공하고 있습니다. 2006년 영업이익 3억 달러에서 2012년 6억 2천만 달러로 두 배가 되는 데는 2012년 출시한 At Bat 12 애플리케이션의 역할이 주요했습니다. 애플 앱 스토어와 안드로이드 구글 플레이에서 다운로드하는 방식으로 시즌 9.99달러, 월 2.99달러의 요금을 받고 있으며 여기서 더 나아가 Xbox, Amazon Kindle Fire와 같은 새로운 플랫폼으로의 진출을 모색하고 있습니다.

MLBAM CEO인 Bob Bowman에 의하면 "우리는 투수가 던지는 공 하나하나를 기록하고 분석하고 있습니다. 지난 10년간 평균 투구 수가 이닝 당 16.345개에 달하며 총 30개의 메이저리그 구단이 하루에 15경기를 펼치며 1년에 162경기를 치러 총 4860게임, 약 70만 개의 공이 분석대상이며 이는 1.5페타바이트(백만 기가바이트)에 달하는 엄청난 양의 빅데이터입니다. 또한 실시간 분석팀은 15초 안에 모든 작업을 마쳐야 합니다. 왜냐하면 30초가 지나면 다른 공으로 새로운 게임(Ball in Play)이 시작되기 때문입니다. 실제로 하루 1200만 명의 방문객이 게임당 8~9테라바이트(천 기가바이트)에 달하는 정보량을 다운로딩하여 연간 1.5페타바이트에 이릅니다. 바로 이런 점에서 MLBAM의 Big Data Cloud Centre가 존재하는 이유입니다."

Nasty Factor
사전적 의미로 Nasty는 못된, 사악한, 지저분한이란 뜻으로 투수가 던진 공이 얼마나 타자가 치기 어려운 정도를 수치화한 값으로 0~100점으로 표기됨. 즉 공의 구종(Pitch Type)과 구질(Pitch Quality)을 분석하여 구위(Pitch Power)가 어느 정도인지를 보여줌. 평가 항목으로는 같은 구종을 던지는 리그 선수들의 속도범위와 비교한 Velocity, 공 배합(Mix up)을 보는 Sequence, 꽉 차는(Edge) 스트라이크존의 Location(스트라이크존을 벗어날수록 감점), 비슷한 구질을 던지는 리그 선수들의 범위와 비교하여 V-Break(수직변화), H-Break(수평변화) 정도를 보는 Movement로 이루어짐. 류현진 선수가 지난 6월 20일(뉴욕 양키스와의 선발 등판) 1사 만루의 위기에서 상대타자 버논 웰스에게 던진 126km의 서클 체인지업은 우타자 바깥쪽으로 날카롭게 떨어져 헛스윙을 유도, Nasty Factor 89점이 나옴.

MLBAM's Pitch f/x, 구종, 속도,*Nasty Factor

MLBAM's AT BAT 13 APP. 실시간 정보제공화면

The Core Business of Big Data!
(Record Analyzer vs Data Scientist)

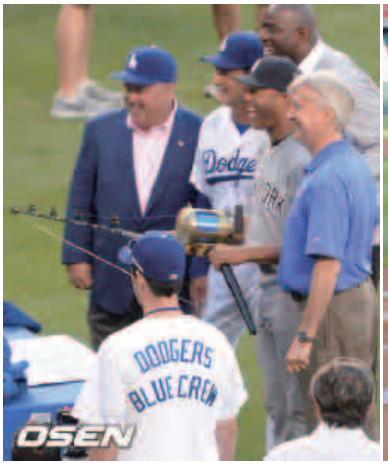
Record Analyzer의 역할이 타율, 타점, 방어율 등과 같은 숫자로 표현 가능한 정형데이터를 기록하고 분석하여 팀 승리에 기여하는 것이라면 Data Scientist의 그것은 정형데이터는 물론이며 이를 시각화(Data Visualization)하고 카메라 정보(Pitch, Bat, Field f/x 등)와 같은 다양하고 복잡한 이미지, 동영상의 비정형데이터를 쉽게 분석 가능한 형태로 가공하여 의미 있는 가치를 도출해내는 모든 과정이라 정의할 수 있습니다. 다시 말해 야구 기록뿐만이 아니라 날씨가 야구에 끼치는 영향에 대해서도 데이터과학자는 부상위험도, 에러확률, 수비범위 등을 선수 측면에서 분석하여 최고의 퍼포먼스를 이끌어 내고, 날씨별 관객 수와 판매상품 변화를 상황(우천, 더블헤더 Rain Check 등)에 따라 분석하여 마케팅 전략으로 이용되는 등 구단 운영 측면에서도 활용될 수 있어야 합니다. 그러기 위해서는 하둡과 같은 빅데이터 처리 플랫폼 운용은 기본이며 데이터를 효과적으로 전달할 수 있는 시각화와 스토리 텔링, 관련 업계의 배경적 지식 등 다양한 요소를 겸비하여야 데이터 과학자라 할 수 있습니다.

요즘 메이저리그는 전설적인 마무리 투수인 마리아노 리베라에게 각 구단 방문 시 어떤 은퇴 기념 선물을 주는지에 대해 관심이 뜨겁습니다. 파나마출신의 리베라는 1995년 양키스에 입단하여 1997년부터 마무리로 뛰면서 통산 1110게임에 출장해 81승 60패 651세이브라는 대기록을 세우고 있으며 올해를 끝으로 은퇴, 그의 등번호 42번 영구결번과 명예의 전당 헌액이 확실시되고 있습니다. 데이터 과학자 얘기를 하다가 갑자기 리베라를 언급하니 많이 당황하셨어요?(개콘 황해버전^^) 눈치 채신 분도 있겠지만 포인트는 바로 데이터 과학자 관점에서 바라본 선물의 가치입니다. 지금까지 선물을 들여다보면 액자, 유화그림, 서핑보드, 카우보이 모자와 부츠, 콜박스와 소방호스 노즐, 골드디스크 등이 있습니다. 마무리 투수를 소방수라고 부르니 콜박스와 노즐을, 리베라의 등장음악(메탈리카의 Enter Sandman)을 앞으로 들을 수 없으니 골드디스크 음반 등이 눈여겨 볼만 합니다.

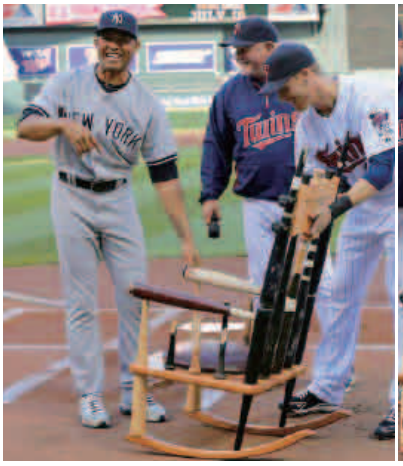
LA 다저스는 은퇴 후 선교사와 같은 자선사업을 꿈꾸는 리베라를 위해 낚싯대를 선물하였습니다. 단순히 은퇴에서 여가를 생각하고 낚싯대를 선물했을 것으로 생각할 수 있지만 절실한 기독교 신자인 리베라를 위해 '사람 낚는 어부'가 되어 은퇴 후의 꿈이 이루어지길 기도하는 의미도 함축하고 있습니다. 예수가 베드로(역대 교황들이 어부였던 베드로의 후계자로 여겨짐)에게 "내가 너를 사람 낚는 어부로 만들어 주겠다."에서 유례하여 성서에서 어부는 사람을 낚는 선교사로 여겨지는데서 착안한 의미 있는 선물인 것입니다.

미네소타 트윈스는 리베라의 은퇴를 휴식을 의미하는 의자와 그 재료로 쓰이는 나무를 배트 브레이커(Bat Breaker)의 별명을 가진 리베라의 커터(직구처럼 날아오다 홈플레이트 근처에서 좌타자의 몸쪽으로 급격하게 휘어져 나감)에 의해 부러진 배트로 사용해 이제 우리의 배트를 그만 부러뜨리고 의자에 앉아 쉬라는 경의를 표한 선물 또한 낚싯대 못지않게 의미 있는 선물이라 할 수 있습니다. 실제로 미네소타는 2002년을 끝으로 리베라의 양키스와 만난 포스트시즌에서 2승 12패로 단 한 번도 ALCS(포스트 시즌은 ALDS→ALCS→WS)에 오르지 못해 월드시리즈 우승의 꿈을 부러진 배트와 함께 접어야 했습니다. 그래서 선물한 의자에는 'Chair of Broken Dreams'라 쓰여 있습니다. 리베라의 포스트시즌 성적 8승 1패 42세이브 0.70의 방어율은 같은 리그의 미네소타 입장에서는 악몽과도 같았을 것입니다.

이렇듯 정보 과학자는 연관이 없어 보이는 데이터에서도 새로운 가치를 재탄생 시킬 수 있어야 합니다. 다시 말해 사전적 의미의 은퇴 → 낚싯대 → 여가를 함축적 의미의 은퇴 → 낚싯대 → 선교로 승화시킨 다저스의 선물이나 연관이 없는 의자, 커터 사이에 자신의 부러진 배트를 사용함으로써 의자 → 부러진 배트 → 커터로 리베라의 은퇴에 경의를 표한 트윈스의 의자 선물이 바로 무수히 많은 데이터 중에서 의미 있는 결과를 도출해 내야하는 정보과학자의 사명을 잘 나타낸 것이라 할 수 있습니다. 만약 리베라가 의자를 경매로 넘긴다면 그 가치는 얼마나 될까요? 소장하고 싶은 양키스 팬과 소각하고 싶은 트윈스 팬 사이의 동시호가가 이루어진다면(실제로 그럴 일은 희박하지만) 시



LA Dodgers로부터 받은 낚싯대 선물




Minnesota Twins로부터 받은 배트 의자 선물

애틀 매리너스와 콜로라도 로키스의 리베라 재단 5000$ 기부액보다 훨씬 더 큰 가치가 될 것입니다. 실례로 86년 만에(1918년 마지막 우승) 밤비노(어린 아이란 뜻의 이탈리아어로 여기서는 Babe Ruth을 일컬음) 저주를 풀게 해준 2004년 보스턴 레드삭스의 투수 커트 실링의 핏빛 양말(발목 인대 수술 후 역투의 흔적)이 9만 2613$, 우리 돈 1억이 넘는 금액에 낙찰된 것을 보면 의미 있는 물품의 값어치가 어느 정도인지를 알 수 있습니다.

Big Data! The More, The BATTER?

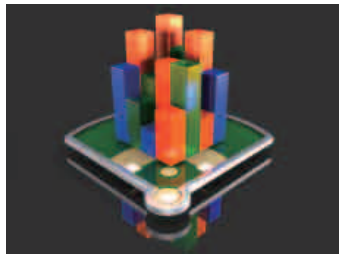
투수에 대한 정보(Pitch f/x 등)가 많다고 해서 반드시 타자(Batter)에게 더 좋은(Better) 걸까요? 그러면 오히려 너무 많은 준비를 해야 하니 역효과가 난다는 것은 자명한 사실일 겁니다. 여기서 메이저리그의 팀 구성을 간략히 설명해야 할 것 같습니다. 아시는 분도 있겠지만 지명타자 제도(보통 투수의 자리인 9번에 수비 없이 타격만 함)가 있는 아메리칸 리그와 없는 내셔널 리그로 나뉘며 각 리그는 동부, 중부, 서부의 3개의 디비전(지구)으로 이루어져 있습니다. 그래서 지구당 5개 팀으로 리그당 15개 팀이 되어 총 30개 팀이 시즌별 162게임을 갖게 됩니다. 같은 지구의 나머지 4개의 팀과는 가장 많은 19게임씩, 총 76게임을 하고 같은 리그 다른 지구의 나머지 10개의 팀과는 팀별로 6~7게임씩, 총 66게임을 하며 끝으로 인터리그라 하여 다른 리그 디비전팀(3개의 디비전팀이 매년 돌아감)과는 팀별로 3~4게임씩 16게임과 흥행을 위한 동일연고지 팀 간 갖는 *지역 라이벌전 4게임 도합 162(76+66+20)경기가 됩니다. 쉽게 말해 같은 지구 투수들과 가장 많은 상대를 하며 디비전과 리그가 달라질수록 적은 상대를(한 번도 상대하지 못하는 투수도 많음) 하게 됩니다. 그러면 당연히 자주 상대하는 선발 투수(Starter) 위주의 정보가 필요하며 특별한 상황에만 등판하는 중간계투(보통 원포인트 릴리프, 추신수의 경우 좌완 스페셜리스트)와 마무리(Closer) 투수 순으로 구종과 구질을 분석하여 대비하게 됩니다.

지역 라이벌 전

우리나라 LG와 두산의 잠실 라이벌전처럼 리그는 다르나 동일 연고지를 갖는데서 유래함. 뉴욕 메츠와 양키스의 지하철시리즈, LA 다저스와 에인절스의 프리웨이시리즈 등이 있음

Big Data! Big Profit?

그렇다면 이런 순서와 룰이 정해지지 않은 기업에는 지금 과연 무슨 일들이 벌어지고 있는 걸까요? 많은 기업들이 하둡 기반의 빅데이터 처리 인프라를 구축하고 있지만 정작 많은 데이터를 어떻게 어느 곳에 활용해야 할지에 대한 고민은 부족해 보이는 것이 현실입니다. 이를 증명 하듯이 시장 전문 분석기관인 가트너(Gartner)가 내놓은 '2012 기업 IT 전망'에서 포춘 500대 기업 가운데 85% 이상이 데이터를 쌓는 데만 치중해 빅데이터의 핵심인 분석으로 이어져 새로운 가치를 창출하는데 실패할 것으로 예상하였습니다. 실례로 야후의 경우 투자대비성(ROI)과 분석 없이 하둡 기반으로 클러스터 시스템을 구축한 결과 동일 데이터 처리 및 분석 작업에 50배 이상 돈이 더 낭비되어 큰 손해를 보았습니다. 여러 가지 요인이 있겠지만 15년의 서비스를 끝으로 작년 12월 31일 사라진 야후코리아는 시대의 흐름에 뒤처지는 순간 IT 기업이 겪게 되는 현실을 여실히 보여주고 있습니다.



페이스북은 어떨까요? 최근 주가가 30$대로 다시 오르긴 하였지만 공모가의 반도 안 되는 20$ 이하로 떨어진 적이 있는 페이스북의 데이터는 구글보다 10배가 많습니다. 하지만 수익 측면에서는 훨씬 뒤처져 있다는 건 새로운 가치는 데이터양이 아닌 분석의 질에서 결정이 난다고 할 수 있습니다.

최근 소셜 미디어업계의 블루칩은 페이스북을 2위로 밀어내고 가장 많은 트래픽을 유발시킨 스텀블어폰(StumbleUpon)이란 SNS 사이트입니다. 사용자 개인의 취향에 맞는 웹페이지, 이미지, 동영상 등을 기호 정보(I like it과 Thumbs Down으로 구분)를 통해 관심사를 자동으로 예측하고 SNS를 통하여 공유할 수 있는 서비스로 핵심기술 역시 비정형 데이터 분석에 있습니다. 그 밖에도 넷플릭스(Netflix)와 아마존(Amazon)의 소비 패턴을 분석해 반드시 구매할 의사가 있는 것들만 추천해주는 상품 추천 기술 역시 빅데이터 분석에서 나온 결과입니다.

바로 이런 점에서 앞서 언급한 바와 같이 데이터 과학자의 역할이 무엇보다 중요합니다. 기업은 데이터 과학자를 기업 내 빅데이터 클러스터를 운용하는 외주, 관리 용역 정도로 생각하여서는 안 됩니다. 데이터 과학자의 인프라 운용 능력은 물론 데이터로부터 새로운 가치를 끌어내기 위한 통찰력과 업계 전반에 걸친 히스토리, 기업문화 이해 없이는 창의적인 아이디어가 나오기 힘들기 때문입니다.

지금까지 메이저리그와 빅데이터 그리고 기업과 빅데이터에 대해 알아보았습니다. 민간 기업은 물론 스포츠분야에서도 빅데이터를 사용해 이제는 사회 전반에 걸쳐 정보의 홍수 속에 살아가고 있는 형국입니다. 정부도 정부3.0을 통한 빅데이터를 이용한다고 하니 본고에서는 언급하지 않았지만 잊혀질 권리(Right to be forgotten)와 같은 개인 정보 보호분야도 큰 이슈가 될 전망입니다. 누구나 그렇듯 조지오웰의 소설 '1984'처럼 빅브라더를 반기는 사람은 아무도 없을 것입니다. 주인공 윈스턴은 소설 속에서만 존재해야 하며 또 다른 에드워드 스노든 사건이 나오지 않길 희망하며...