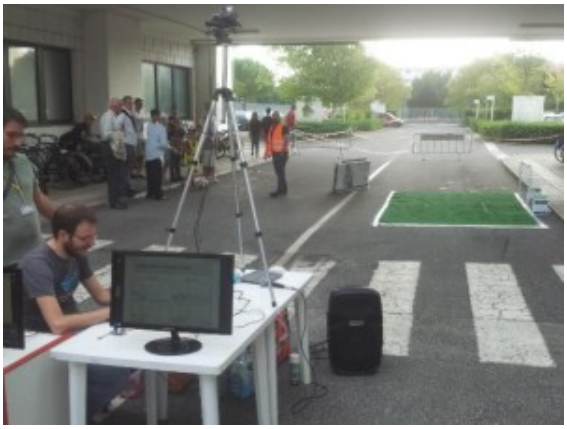
O' Big Data è meglio e' Pelè - BRIGHT 2015

Tra i ventiquattro Paesi europei e le venti città italiane impegnate nella notte dei ricercatori, c'era anche Pisa. Area della ricerca del CNR inclusa. E' qui che ieri, 25 settembre, ci siamo divertiti a provare "sul campo" le tecniche di acquisizione dati dai campi di calcio. Al Kddlab dell'ISTI-CNR siamo abituati a lavorare sui dati dei calciatori professionisti, ma la rivoluzione dei dati si allargherà anche alle serie minori. Per questo, abbiamo invitato i visitatori di Bright a giocare sul nostro mini campo, allestito sul piazzale del CNR.

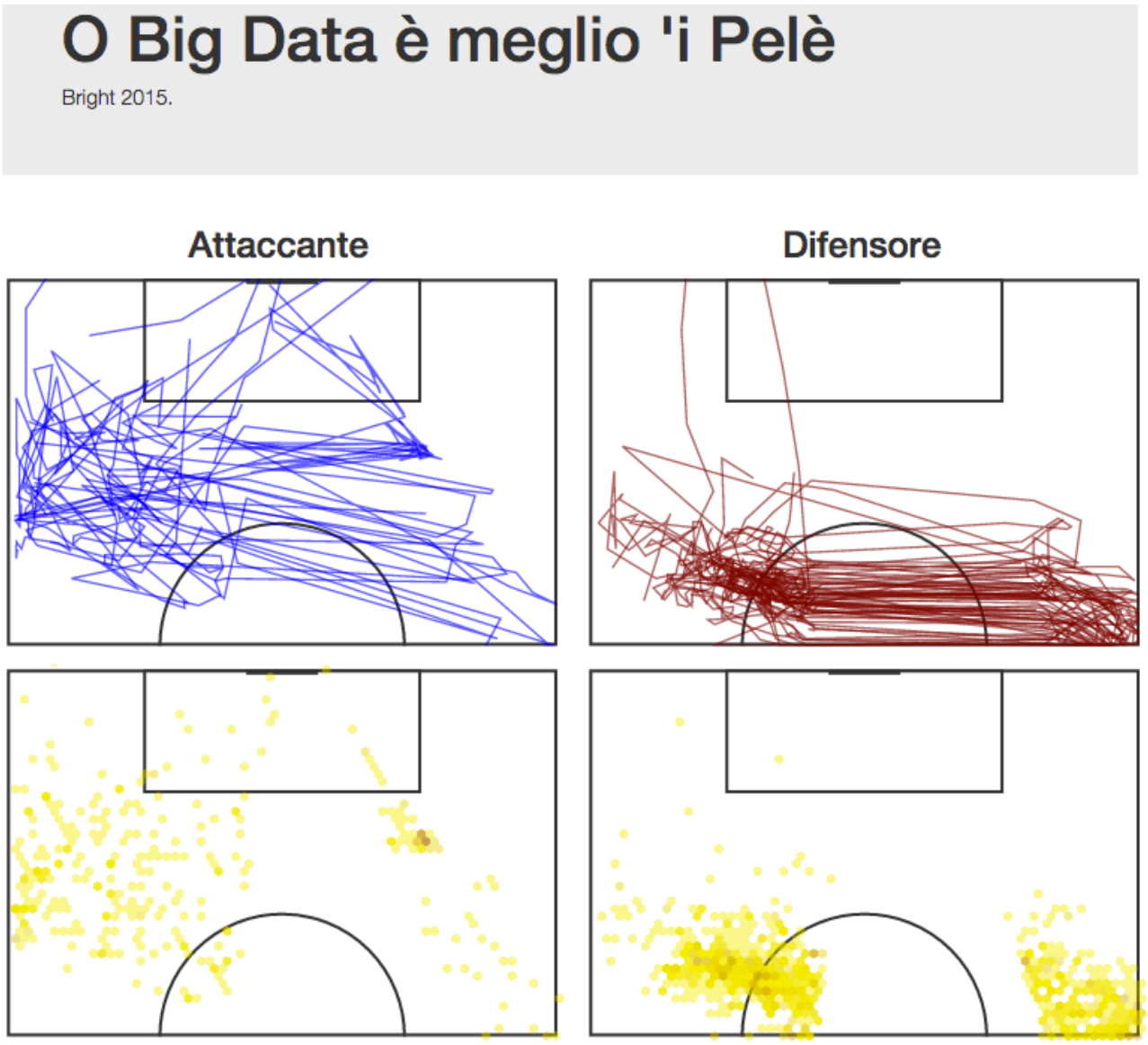


Bright 2015: tracking di giocatori in real time

Abbiamo proposto due sfide, a grandi e piccoli: segnare un gol affrontando il più arcigno dei difensori del CNR, al secolo Salvo Rinzivillo, e segnare da calcio piazzato, con tanto di barriera. La differenza, rispetto al calcio da strada con cui siamo cresciuti, è che le sfide uno contro uno sono state tracciate tramite videocamera e apposito software sviluppato da noi stessi, con le traiettorie e relative velocità mostrate in tempo reale. Nei calci piazzati, invece, un pallone con sensori e trasmettitore bluetooth ha registrato i dati sul

colpo tirato dai partecipanti.

Dei tanti, bellissimi, stand presenti a Bright, siamo probabilmente gli unici a poter dire dove e come sono andati i nostri visitatori:



Bright: le traiettorie dei nostri visitatori

Dai due grafici in alto si vede la mobilità, limitata tra l'altro, del nostro Rinzivillo, comparata soprattutto con quella dei ben più giovani attaccanti che si sono succeduti nell'affrontarlo. Le velocità medie parlano chiaro: 11 km/h per gli attaccanti, 8 km/h per il difensore. Nei due plot in basso, invece, la densità di gioco in base alla zona del (mini) campo. La maggioranza di attaccanti di piede destro è

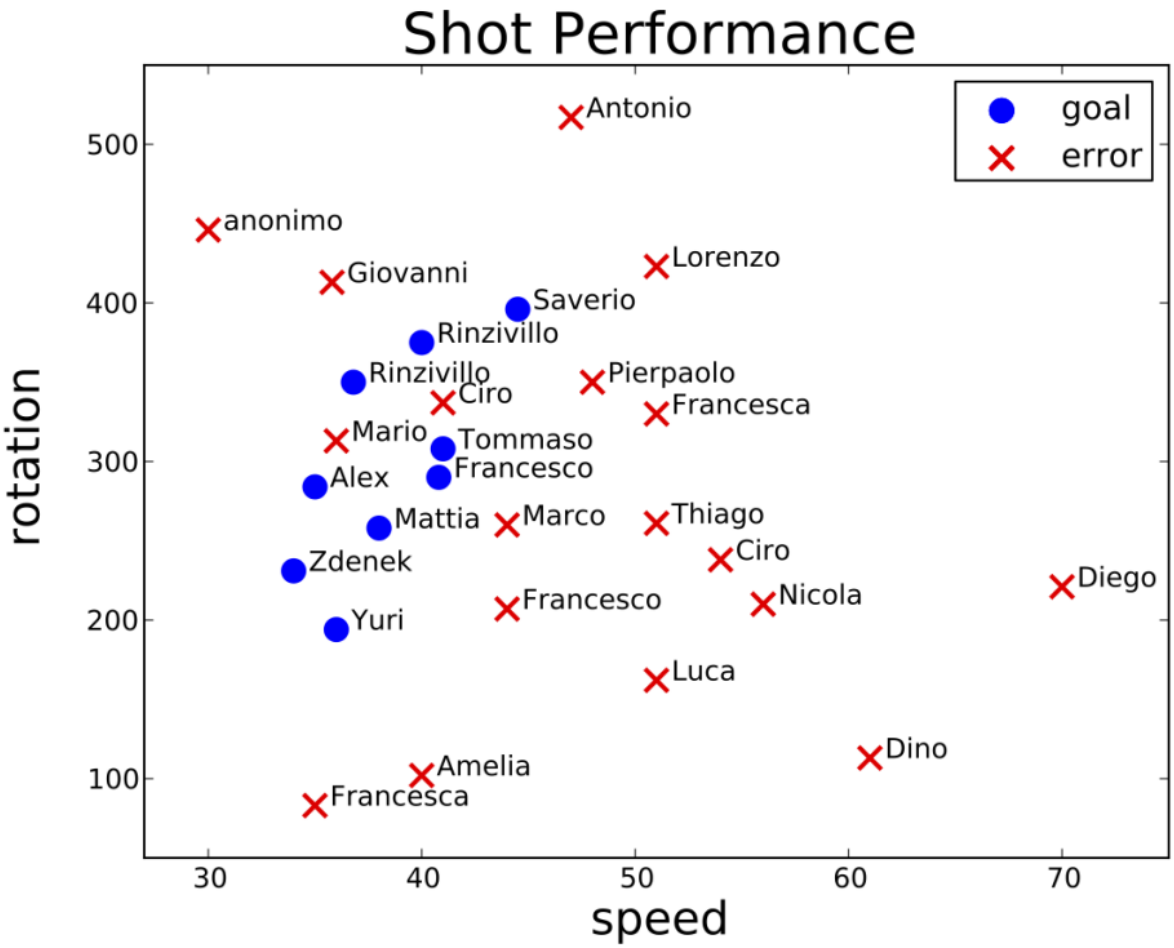
evidente dal grafico relativo: la zona più densa è quella, appunto, da dove calcia in porta un destro. Il nostro Rinzivillo, invece, ha difeso la (mini) porta con spostamenti prevalentemente laterali. Tanto per avere un confronto, un giocatore professionista arriva ad oltre 30 km/h, nei suoi scatti più intensi.

Sfida: battere la barriera dei brasiliani del CNR

La sfida sui calci piazzati è stata un po' più difficile. Se i bambini più piccoli hanno ancora poco controllo della palla, già dai 10 anni in su abbiamo potuto apprezzare dei piedini raffinati. La nostra smart ball ha registrato i dati di velocità della palla in km/h e di rotazione in giri al minuto.

La palla è stata sempre calciata dallo stesso punto, con tanto di barriera fissa. Il risultato più interessate è l'emergere di un pattern abbastanza preciso: per quanto la situazione fosse più un esempio giocattolo che un caso reale – distanza 5 metri, porta di 1,20×0.8 metri -, dal grafico finale si vede come la maggior parte dei gol siano stati segnati calciando ad una velocità tra i 37 e i 39 km/h. La velocità giusta per far abbassare la palla dietro la barriera, insomma. Altri numeri, per fortuna delle finestre del CNR, rispetto agli oltre 100 km/h di Pirlo. Dal grafico si nota anche la precisione del nostro Rinzivillo: da fine data scientist quale è, ha usato i giorni di preparazione a Bright per provare la palla bluetooth e sfruttare al meglio i dati raccolti. Altre due edizioni di Bright e sarà pronto per

diventare il numero 10 del Pozzallo.



La bravura dei tiratori di Bright

Bright è stato un evento di successo, in tanti hanno visitato gli stand dei tanti laboratori presenti nell'are del CNR. Certo, l'organizzazione non è stata una passeggiata e al termine della giornata ci è sembrato di aver corso per 90 minuti di fila, ma d'altronde, come cantavano a Napoli:

> *O' Big Data è meglio 'e Pelè, c'amm fatto 'o mazzo tanto pe l'avé. (*)*

*: Si ringrazia Luca Pappalardo per la consulenza sul lessico napoletano

In basso a destra, Diego e David (anpas Ponsacco) mentre attendono di osservare il risultato dei loro test

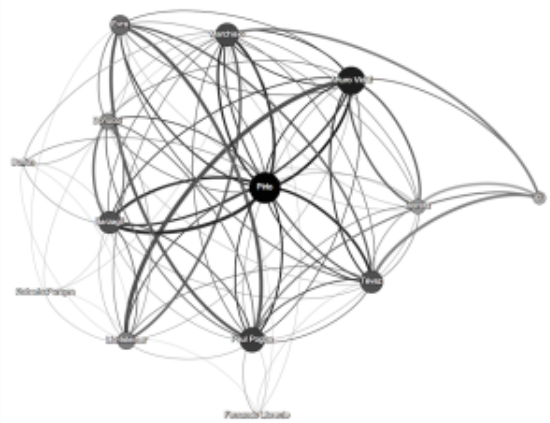
Post Scriptum: ai nostri esperimenti si sono prestati anche Diego e David, due calciatori dell'Anpas Ponsacco, venuti anche per ricordare il loro impegno nella partita del cuore. Partita che, l'indomani, hanno vinto. L'allenamento mirato a suon di dati è servito!

Chi ha infranto la dura legge matematica del gol?

Il calcio porta con sé tanti misteri. Squadre che vincono senza quasi tirare in porta, amicizie sincere che si rovinano per un rigore negato rivisto alla moviola, città paralizzate dopo una partita vinta. Studiare il calcio, in fondo, significa cercare di svelare almeno una piccola parte di questi fenomeni al limite del paranormale. Messa così, è una sfida a cui un giallista non direbbe mai di no, soprattutto se il giallista è anche uno scienziato.

L’avvento dei Big Data del calcio è l’occasione giusta per poter spingere un po’ più in là la nostra comprensione del fenomeno calcistico. I dati sono gli indizi che servono alla nostra indagine. Fino a ieri, conoscevamo poche regole, perlopiù validate in modo empirico: il calcio è un gioco basato su episodi casuali che nel 99% dei casi sono favorevoli alla Juventus. Il mare di dati che oggi accompagna ogni partita, però, è in grado di farci capire qualcosa di più.

Prendiamo, ad esempio, la rete dei passaggi tra i giocatori di una squadra. Come per due amici in una rete sociale, due giocatori sono nodi di una rete che entrano in relazione tra di loro nel momento in cui uno passa la palla all’altro. In più, muovere la palla cercando di portarla dentro la porta avversaria è lo scopo primario del gioco del calcio. Il passaggio, inoltre, è l’evento più ricorrente durante una partita: mediamente uno squadra ne effettua 400 ogni partita. Mentre il goal, al contrario, è l’evento più raro.

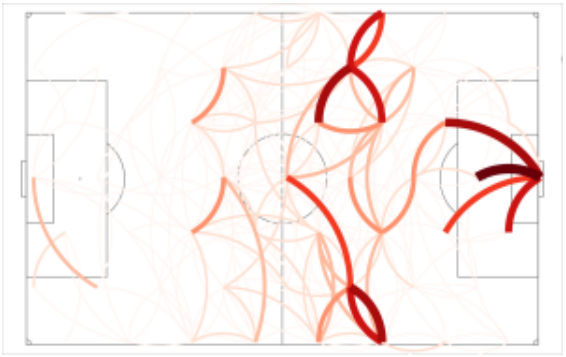


Juventus-Barcellona, giugno 2015. La rete della Juventus è incentrata su Pirlo.

Neanche Marco Malvaldi, lo scienziato giallista, ha saputo resistere quando gli abbiamo proposto di indagare il calcio. Niente morti ammazzati a Calambrone, né vecchietti ficcanaso: qui i delitti sono punti in classifica non meritati e i

ficcanaso siamo noi. Gli indizi stavolta sono un po’ di più della solita impronta digitale: 600mila passaggi, divisi in 148 partite di 4 campionati diversi. Per ogni partita possiamo costruire due reti, che esprimono *chi* effettua i passaggi e *dove* i passaggi vengono effettuati.

Di queste due reti, analizziamo la distribuzione: la media di passaggi effettuata da ogni nodo e la relativa varianza. Possiamo esprimere due caratteristiche del gioco di una squadra: il volume di gioco e la sua imprevedibilità. Le caratteristiche spaziali sono invece analizzabili usando come nodi le zone del campo in cui avvengono i passaggi. Possono le caratteristiche di questa rete darci qualche informazione in più?



Rete dei passaggi del Barcellona durante l’ultima finale di Champions League. Il gioco si sviluppa su due zone precise, quelle di Messi e Neymar

L’ipotesi c’è, dunque. Ora servono le prove. Il fiuto di chi ha risolto decine di casi contorti sa dove andarle a cercare: analizziamo le reti di tutte le squadre dei campionati europei principali, durante una stagione intera. In pratica, simuliamo il campionato: per ogni partita calcoliamo volume e imprevedibilità delle reti di ogni squadra, li combiniamo in un’unica misura calcolandone la media armonica. Per comodità la chiamiamo *H*. Quindi, per ogni partita assegniamo i 3 punti

alla squadra con l'*H* più alto. Se la differenza di due squadre è ragionevolmente vicina, allora il risultato è un pareggio. Alla fine, la classifica finale secondo la misura *H* ha "confermato" l'ipotesi: le reti ci danno qualcosa in più. La correlazione tra la classifica simulata e quella reale è superiore all'80%. Arriva all'89% nel caso del campionato tedesco.

simulated ranking		real ranking	
Bayern	95	**Bayern**	90
Dortmund	75	Dortmund	71
Wolfsburg	62	Schalke	64
Leverkusen	59	Leverkusen	61
Augsburg	54	Wolfsburg	60
Hoffenheim	54	Mönchengladbach	55
Hannover	49	Mainz	53
Schalke	47	Augsburg	52
Hertha	43	Hoffenheim	44
Mönchengladbach	42	Hannover	42
Mainz	40	Hertha	41
Hamburg	40	Werder	39
Stuttgart	38	Freiburg	36
Frankfurt	34	Frankfurt	36
Nürnberg	29	Stuttgart	32
Braunschweig	26	Hamburg	27
Freiburg	24	Nürnberg	26
Werder	22	Braunschweig	25

Bundesliga 2013/2014, classifica reale e simulata secondo la misura H.

In generale, ciò che emerge è che l'analisi delle reti riesce a descrivere la performance di una squadra. Considerando che stiamo parlando solo di reti di passaggi, senza aver ancora incluso altri tipi di eventi (dribbling, tackle, etc.). Tuttavia, c'è comunque un errore da migliorare. E per raggiungere la perfezione, vanno scomodati gli dèi. Il calcio l'hanno spiegato già in tanti prima di noi, e qualcuno ha anche provato a formularne le leggi che lo governano. Uno di questi è Max Pezzali. Già, la dura legge del gol.

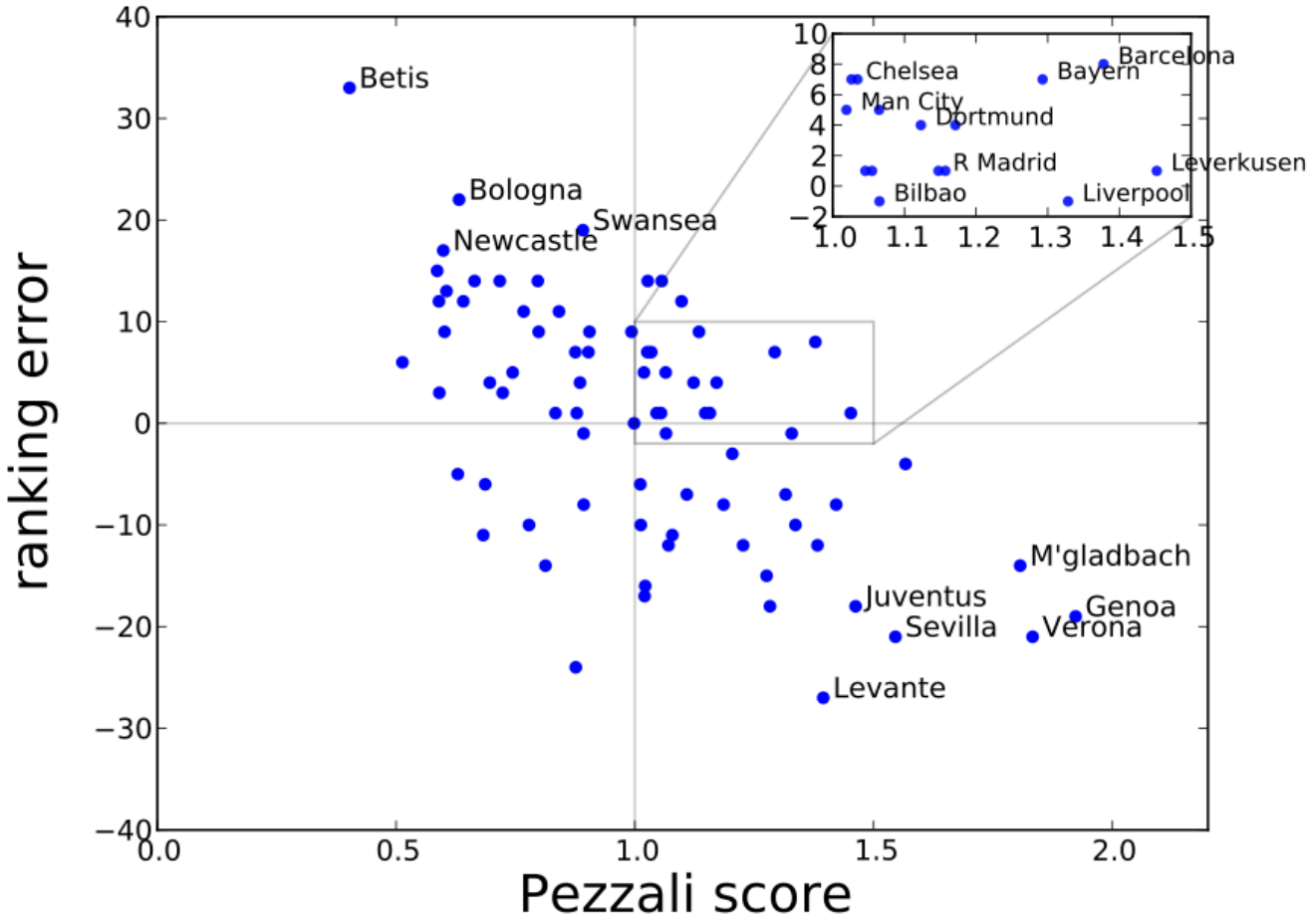
Osservando le squadre per cui la nostra simulazione sbagliava di più, abbiamo notato che sono sia quelle che giocano in contropiede, che quelle che difendono male. Queste, però, sono osservazioni empiriche, la cui formalizzazione risulta difficile. Si possono invece analizzare gli eventi in campo ispirandosi a Pezzali. Ad esempio, una squadra dalla difesa debole subisce la dura legge del gol:

> *E' la dura legge del gol, fai un gran bel gioco però, se non hai difesa gli altri segnano e poi vincono.*

Mentre ci sono squadre che questa legge la impongono:

> *Loro stanno chiusi ma, alla prima opportunità, salgon subito e la buttan dentro a noi.*

A livello numerico, abbiamo definito il Pezzali score: il rapporto tra i tiri in porta e i gol fatti moltiplicato per il rapporto tra i gol subiti e i tiri fatti dall'avversario. Quando il Pezzali score è basso, la squadra è un po' come l'Inter, gioia e dolore dell'autore stesso: tanto gioco, ma agli avversari di turno basta un tiro per fare un gol. Quando, invece, il valore è alto, la squadra ottiene il massimo dal minimo sforzo offensivo. Per capire quanto la dura legge del gol sia decisiva nel dirci dove sbaglia la nostra simulazione, abbiamo confrontato il Pezzali score di ogni squadra con l'errore in termini di punti finali che la nostra simulazione assegna. Il risultato (vedi sotto) conferma quanto la visione di Pezzali, pur poetica e non scientifica, fosse azzeccata. Se con l'H siamo in grado di valutare quanto sia efficace il gioco di una squadra, la dura legge del gol ci riporta alla realtà: il calcio è un gioco dove esiste una componente casuale. Non basta il gran bel gioco per vincere, insomma.



La dura legge del gol: l'errore maggiore (Betis, Verona, etc) corrisponde a valori estremi del Pezzali score.

Il mistero non è ancora risolto, perché in fondo il giallo non è ancora finito. Con buona pace di Malvaldi e dei suoi vecchietti ficcanaso che risolvono casi di cronaca nera, c'ancora tanto da studiare, e da scrivere, prima di riuscire a spiegare il gioco più bello del mondo. E anche se la dura legge del gol continuerà a valere nei secoli dei secoli, proveremo fino in fondo ad infrangerla.

I nostri risultati in tema di analisi dei dati calcistici li trovate qui:

The harsh rule of the goals: data-driven performance

indicators for football teams
P.Cintia, L. Pappalardo, D. Pedreschi, F. Giannotti, M. Malvaldi
2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA'15), Paris, France

Ne abbiamo parlato anche a radio3 scienza e a Zona Cesarini (Radio1):

Ringraziamo: Marco Malvaldi, per averci prestato i vecchietti indagatori. Max Pezzali, perché siamo cresciuti negli anni '90. Mariano Tredicini e la Tim, per aver supportato la nostra ricerca.

Calcio, Big Data e il sol dell'avvenire

Un computer a disposizione di un allenatore era un'utopia, nel 1973. La storia dei dati e del calcolo applicato al calcio non poteva che cominciare in un posto dove di utopia, nel 1973, se ne intendevano: l'Unione Sovietica. In quell'anno il colonnello Lobanovs'kyj iniziò ad allenare la Dinamo Kiev, e come prima richiesta nel suo staff volle uno statistico ed un computer. Di calcolatori se ne vedevano talmente pochi che il KGB lo mise subito sotto controllo. "Tutto è un numero" profetizzava il colonnello. Addestrò la sua squadra alla ripetizione di schemi elaborati al computer. Fu il primo ad introdurre il calcio totale e portò la Dinamo Kiev e la nazionale russa ai massimi livelli del calcio europeo.

Quando invece i computer erano ancora solo prototipi e i Big

Data una visione fantascientifica, il ragionier Charles Reep riempiva taccuini su taccuini con i dati delle partite a cui assisteva. Quindi li elaborava, rigorosamente a mano. Cominciò nel 1950, quando durante un partita dello Swindon Town si era annoiato talmente tanto da cominciare ad annotare tutto ciò che succedeva in campo. La sua scoperta è ancora alla base di tanti sistemi di gioco: la probabilità di sbagliare un passaggio aumenta con il numero di passaggi consecutivi. Altro che Tiki-Taka, Reep dimostrò numericamente l'efficacia del contropiede: portando prima possibile la palla nell'area avversaria si massimizza il numero dei gol. Questa teoria, conosciuta come "Teoria della palla lunga", venne pubblicata sul prestigioso Journal of Royal Statistical Society e ha ispirato più o meno tutto il calcio inglese dagli anni '60 in poi.

Il frutto più recente della rivoluzione dei dati è il Midtjylland, squadra fondata nel 1999 e fresca vincitrice del campionato danese. L'Ad della società, il trentaduenne Rasmus Ankersen, ha lanciato una provocazione fantascientifica: "l'algoritmo è più importante della classifica". Il padrone del Midtjylland è Matthew Benham, uno scommettitore incallito che con i proventi delle vincite acquista squadre di calcio. La scommessa di Benham e Ankersen si è rivelata poco azzardata: il loro scouting algebrico li ha portati a selezionare i giocatori giusti durante il calciomercato, facendoli poi rendere al meglio delle loro possibilità.

L'occhio (poco attento) di Ferguson su Stam

I dati possono anche portare a prendere una cantonata, come successe a Ferguson nel 2001. Le statistiche evidenziavano un calo dei tackle di Stam, suo difensore al Manchester United, perciò sir Alex decise di vendere il giocatore, considerandolo in calo. Stam restò su alti livelli di rendimento per altri sei anni, prima alla Lazio e poi al Milan. Ferguson non aveva considerato il "principio di Maldini": Paolo Maldini è stato uno dei più grandi difensori della storia pur facendo, in media, un tackle ogni due partite.

La ricerca dell'alchimia che si nasconde dietro una vittoria non poteva che affascinare anche il mondo della scienza. Undici giocatori che ne affrontano altrettanti, ognuno libero di muoversi nello spazio ma con un obiettivo comune a tutti e indipendente (o quasi) dal singolo. Una sfida, più che un problema da risolvere. Uno dei lavori più interessanti è quello di Taki e Hasegawa, con la loro definizione di 'regione dominante'. Analizzando i dati di tracking dei giocatori, cioè la loro posizione in campo registrata ogni decimo di secondo, i giapponesi hanno elaborato un modello geometrico in grado di calcolare, ad ogni istante, l'area che un singolo giocatore può raggiungere prima di tutti gli altri. Sebbene onerosa in termini di calcolo, questa misura si è rivelata molto interessante: la strategia di attacco una squadra può essere valutata in base alla capacità di massimizzare le regioni dominanti dei suoi giocatori. Un team di ricercatori australiani guidati dal prof. Horton ha invece coinvolto dieci allenatori nella costruzione di un classificatore di passaggi: ogni allenatore ha visionato e valutato una serie di passaggi (intelligente, scontato, etc); sulla base di questi giudizi è stato sviluppato un algoritmo in grado di riprodurre il ragionamento degli allenatori, automatizzando tutto il processo.

Il santo Graal della Data Science applicata al calcio è proprio la valutazione dell'intelligenza di una mossa. Quanto incide la decisione sul dove e a chi passare la palla? Quanto

è efficace uno scatto in profondità o una rincorsa dell'avversario a metà campo? Il fiuto nel rispondere a queste domande è ciò che fino ad oggi ha reso grandi gli allenatori. Lo scenario sta però cambiando, e i presidenti sanno già che il loro prossimo tecnico dovrà essere anche un bravo Data Scientist.

Mattarella e il complotto dei dati

Accessi alla pagina del Presidente della Repubblica su wikipedia

A partire da dicembre un flusso di notizie, pettegolezzi, opinioni e retroscena ci ha sommerso, come per ogni elezione del presidente della Repubblica che si rispetti. Questa volta la portata della piena è stata maggiore, perché il periodo è tra i più contorti della storia contemporanea. L'ultima elezione del Presidente è avvenuta meno di due anni fa, anziché i sette convenzionali. Il Presidente uscente è dimissionario (raro) e al secondo mandato (mai successo prima): la riconferma di Napolitano arrivò al termine di votazioni in cui successe di tutto, dalla silurazione di Prodi alla rivolta contro Marini, passando per le quirinarie dei 5stelle dove arrivò prima una giornalista (Gabanelli) e secondo un medico (Gino Strada). Con questi presupposti, l'elezione del Presidente della Repubblica si preannunciava come una specie di thriller.

Eppure, nel mare magnum della rete la soluzione all'enigma era già lì, pronta due giorni prima dell'ultimo scrutinio. Il 29 gennaio Renzi sciolse la sua riserva. Durante la direzione PD comunicò chi aveva scelto come candidato da far votare al suo partito: Sergio Mattarella. I dati di accesso a Wikipedia erano chiari: il 29 gennaio, ben due giorni prima dell'elezione, la partita poteva già dirsi chiusa. Questi dati hanno una particolarità: indicano il numero di accessi alla pagina, ma comprendono anche gli accessi fatti quando il nome viene cercato su google e compare il box a destra con le informazioni principali tratte da wikipedia. Sono dunque una valutazione più che consona dell'interesse che c'è intorno a quel nome.
Oggi sappiamo già tutto di Mattarella, dalla storia personale agli hobby, ma prima di quel giorno non era proprio così. Prima dell'annuncio, nessuno lo conosceva. I dati di Wikipedia sono inequivocabili: all'annuncio del premier, tutta la rete è andata alla ricerca della storia dell'ex giudice costituzionale-ex ministro-ex democristiano designato da Renzi come candidato unico alla presidenza.

Mattarella chi?

Nel frattempo, il secondo partito italiano ha organizzato le consuete Quirinarie, per chiedere ai suoi iscritti di scegliere il candidato da votare. L'annuncio dei 5stelle è arrivato lo stesso giorno, il 29 gennaio, ma l'interesse della rete è su scala decisamente minore. Il risultato si può interpretare in vari modi: Imposimato era già nome noto e popolare, oppure intorno al candidato dei 5stelle, famoso o meno che fosse, non c'era una grande curiosità. Lo stesso si può dire per il candidato dell'estrema destra Vittorio Feltri, sebbene partisse da una popolarità maggiore. Per meglio comprendere il fenomeno, vale la pena osservare anche l'interesse attorno al nome di Romano Prodi: ad un certo punto la minoranza del PD sembrava volesse votarlo, la rete sembra

confermare che delle mosse della minoranza PD non gli importa poi chissà quanto. Anzi, "se i dati fossero confermati" - come si dice al termine delle elezioni - alla rete, delle intenzioni della minoranza PD, non gliene può fregare di meno.

Wikipedia anticipa i risultati delle elezioni. Il complotto dei dati.

Al termine della cavalcata trionfale di Mattarella, vale la pena dare uno sguardo a cosa è successo nelle retrovie. Come ad ogni elezione che si rispetti, a volte dalle urne sbucano fuori nomi poco consueti. Al quarto scrutinio, Walter Veltroni (PD), Roby Facchinetti (Pooh) e Lino Banfi (BA) hanno ottenuto lo stesso numero di voti: 1. Francesco Guccini li ha surclassati, ottenendone il doppio. Ancora prima, Giancarlo Magalli aveva vinto a mani basse il sondaggio de Il Fatto Quotidiano. Quanta curiosità c'era intorno a questi nomi? In quanti li hanno cercati su Google? Vittoria schiacciante di Magalli, idolo di internet e trasformato da conduttore televisivo a opinionista politico. A gennaio ha inaugurato il suo blog all'interno de L'Espresso.

Il gennaio d'oro di Magalli

Il turbine di emozioni previsto per le elezioni del Presidente della Repubblica non si è rivelato quello previsto. Sergio Mattarella è stato votato senza affanni né tradimenti. Non c'era nessun altro in campo, ma soprattutto i 101 franchi tiratori del PD sono stati buoni questa volta. Bontà forse dettata dalla disciplina di partito - oggi il segretario è Renzi, nel 2013 c'era Bersani - o dalla consapevolezza che quando scende in campo la balena bianca, non c'è partita per nessuno.

Muovere il mouse sulla figura per scoprire cosa fa la DC ai franchi tiratori