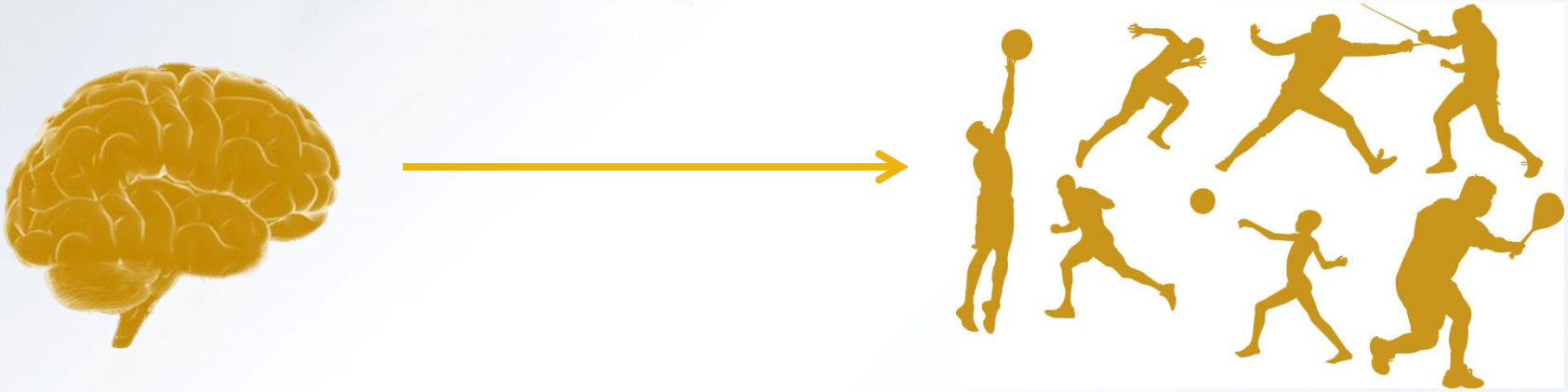
Incorporating human complexities in predictive sports analytics



Joel Sokol

Director, MS in Analytics (On-Campus and Online)
Associate Professor, Stewart School of ISyE
Georgia Institute of Technology

Georgia Tech Master of Science in Analytics

Georgia Tech ISyE H. Milton Stewart School of Industrial and Systems Engineering

Overview

- Why is predictive sports analytics different?
- Examples: what's needed?
  - Solutions: what's being done?
  - Research frontiers: what's needed?

Humans have little/no effect

Weather forecasting

Crop yield

Manufacturing productivity

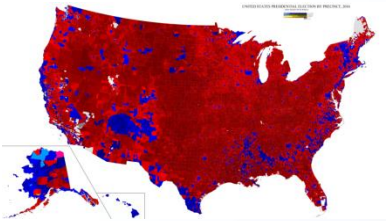
Human effects are collective

(Large *n* → law of large numbers)

Insurance

Marketing

Demand Forecasting



Election Forecasting

Loans

Social Media

Sports analytics: individuals matter!

- Game strategy
  - How will Player/Team X perform (against Player/Team Y) *today*?

- Player selection (Draft, free agency, contracts)
  - How well will Player X perform next year?
  - Over the next k years?

- Training/preparation
  - How will Player X respond to training regimen R?

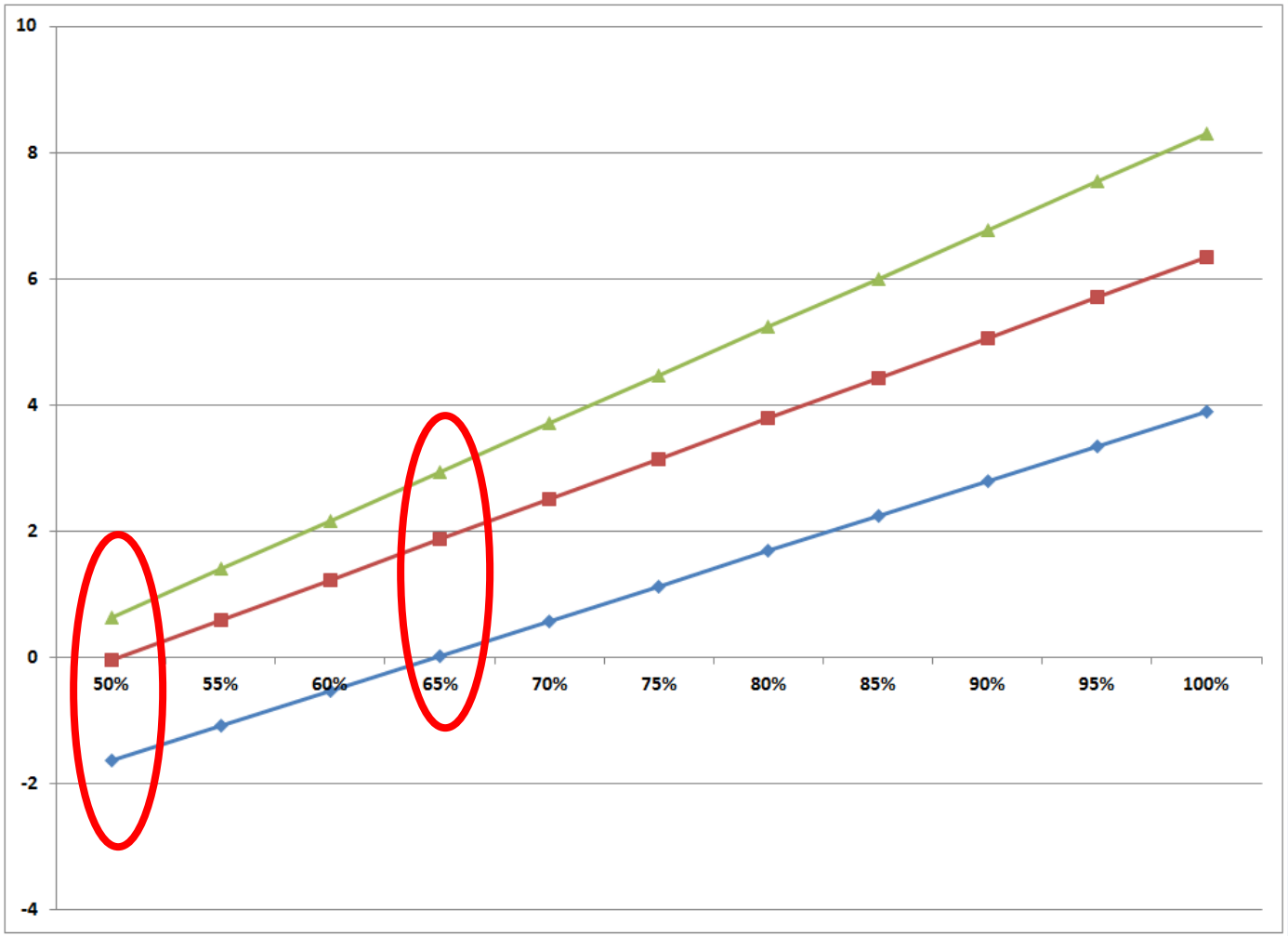
Game strategy example

"Batters behave like a tabletop baseball game…"

- Pick two very similar batters
  - Same team
  - Same hand, other characteristics
  - Played almost every day

- If we knew with probability *p* which has better outcome each day…

Paired batter studies



- Millions of dollars in impact
  - even from 50% to 65%
- Need new types of data
  - Biometric
  - Physiological
  - Psychological
  - Side question: measuring without disturbing

Approx. $6M

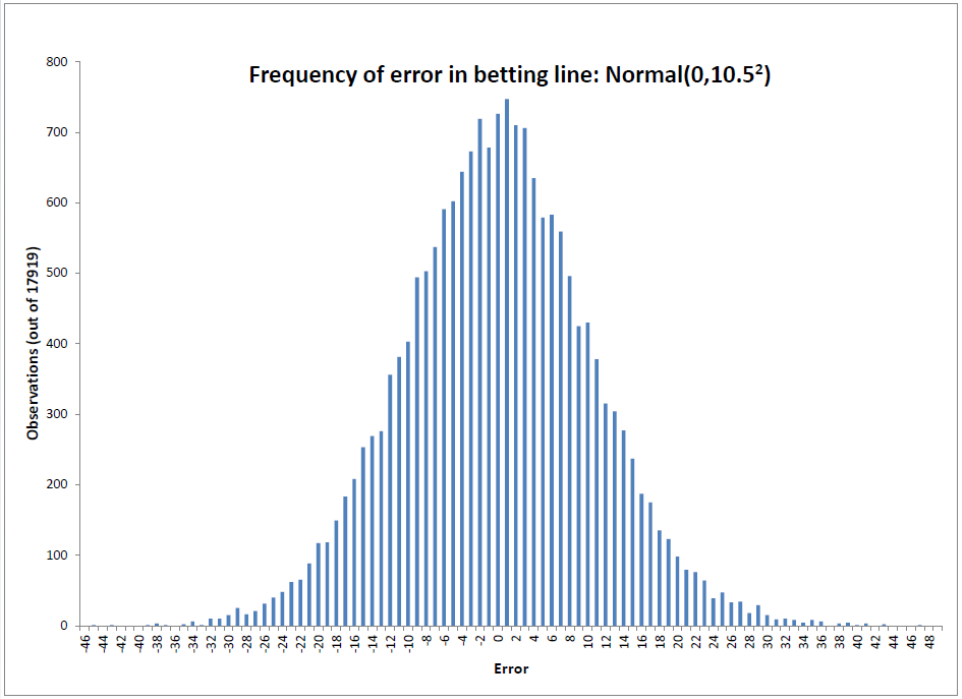
Game prediction

- Who will win the match?

- Who will win the tournament?
  - Complexities
    - Reliance on small number of events
    - Reliance on rare events (e.g., soccer goals)
    - Importance of individual performance
      - Physiology: Time zones, elevation, circadian rhythm
      - Other physiological/psychological factors?
      - Luck

Where does error come from?

"Why haven't you improved LRMC lately?"



- *Randomness*:
  - Does the ball bounce in or out?
  - Does the player trip or not?
  - Etc.
- *Uncertainty*:
  - Errors in estimate of team ability
  - (Also other estimation errors: home court advantage, etc.)
- Frontier of sports analytics:
  - Identify and quantify uncertainties separate from randomness
  - Better insight & better value

Where does error come from?

"Why haven't you improved LRMC lately?"

For significant improvement:
"Randomness" → "Uncertainty"

Usually that means "new data", not "better models or algorithms"

Uncertainty in team strength

Randomness in game events

Curry & Sokol (forthcoming)

Player selection

Player prediction *and* expert behavior

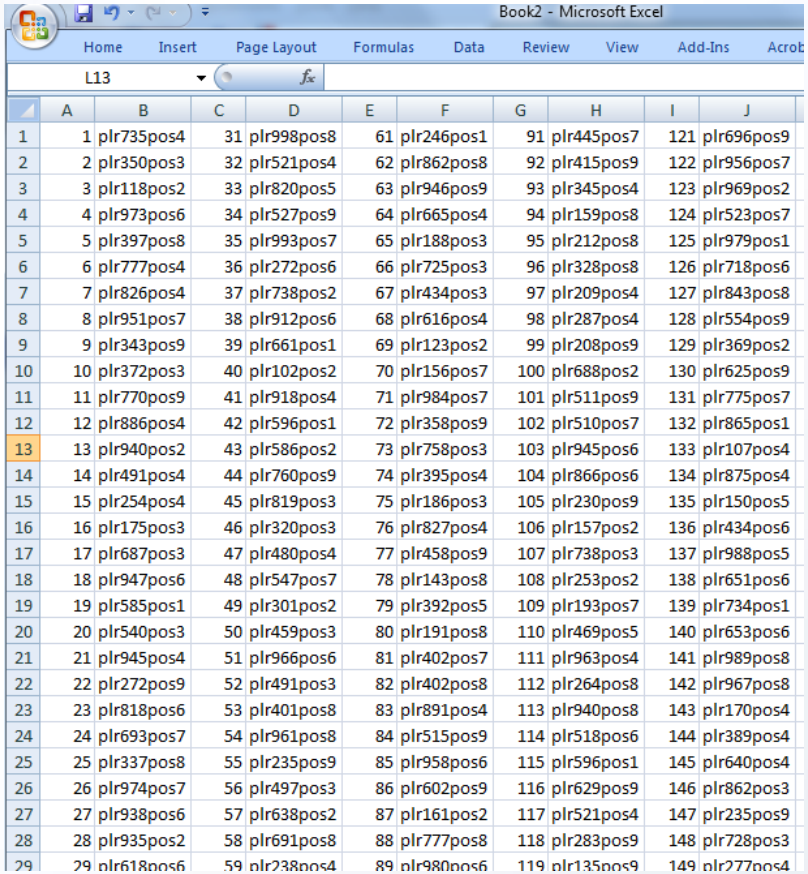
- Example: Major League Baseball draft
  - (Teams take turns selecting a player)

- Statistical models can predict…
  - Expected player performance (with lots of variance)
  - Probability of getting to Major Leagues
    - Little apparent difference between consecutive players

- ...but teams need to pick a single player
  - Unsuccessful choices cost wins, jobs

Second layer of complexity: uncertainty

- 1000+ players, 25+ scouts
- Each scout ranks a subset of players
  - Better statistical models → emphasis on psychology and motivation
  - Scouts have significant contradiction
  - Uncertainty is inherent in scouts' evaluations



Book2 - Microsoft Excel

Home Insert Page Layout Formulas Data Review View Add-Ins Acrob

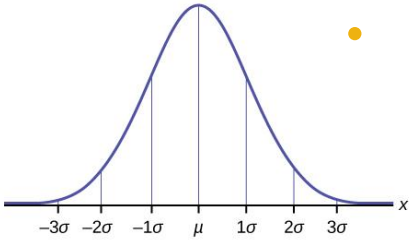
L13

	A	B	C	D	E	F	G	H	I	J
1	1	plr735pos4	31	plr998pos8	61	plr246pos1	91	plr445pos7	121	plr696pos9
2	2	plr350pos3	32	plr521pos4	62	plr862pos8	92	plr415pos9	122	plr956pos7
3	3	plr118pos2	33	plr820pos5	63	plr946pos9	93	plr345pos4	123	plr969pos2
4	4	plr973pos6	34	plr527pos9	64	plr665pos4	94	plr159pos8	124	plr523pos7
5	5	plr397pos8	35	plr993pos7	65	plr188pos3	95	plr212pos8	125	plr979pos1
6	6	plr777pos4	36	plr272pos6	66	plr725pos3	96	plr328pos8	126	plr718pos6
7	7	plr826pos4	37	plr738pos2	67	plr434pos3	97	plr209pos4	127	plr843pos8
8	8	plr951pos7	38	plr912pos6	68	plr616pos4	98	plr287pos4	128	plr554pos9
9	9	plr343pos9	39	plr661pos1	69	plr123pos2	99	plr208pos9	129	plr369pos2
10	10	plr372pos3	40	plr102pos2	70	plr156pos7	100	plr688pos2	130	plr625pos9
11	11	plr770pos9	41	plr918pos4	71	plr984pos7	101	plr511pos9	131	plr775pos7
12	12	plr886pos4	42	plr596pos1	72	plr358pos9	102	plr510pos7	132	plr865pos1
13	13	plr940pos2	43	plr586pos2	73	plr758pos3	103	plr945pos6	133	plr107pos4
14	14	plr491pos4	44	plr760pos9	74	plr395pos4	104	plr866pos6	134	plr875pos4
15	15	plr254pos4	45	plr819pos3	75	plr186pos3	105	plr230pos9	135	plr150pos5
16	16	plr175pos3	46	plr320pos3	76	plr827pos4	106	plr157pos2	136	plr434pos6
17	17	plr687pos3	47	plr480pos4	77	plr458pos9	107	plr738pos3	137	plr988pos5
18	18	plr947pos6	48	plr547pos7	78	plr143pos8	108	plr253pos2	138	plr651pos6
19	19	plr585pos1	49	plr301pos2	79	plr392pos5	109	plr193pos7	139	plr734pos1
20	20	plr540pos3	50	plr459pos3	80	plr191pos8	110	plr469pos5	140	plr653pos6
21	21	plr945pos4	51	plr966pos6	81	plr402pos7	111	plr963pos4	141	plr989pos8
22	22	plr272pos9	52	plr491pos3	82	plr402pos8	112	plr264pos8	142	plr967pos8
23	23	plr818pos6	53	plr401pos8	83	plr891pos4	113	plr940pos8	143	plr170pos4
24	24	plr693pos7	54	plr961pos8	84	plr515pos9	114	plr518pos6	144	plr389pos4
25	25	plr337pos8	55	plr235pos9	85	plr958pos6	115	plr596pos1	145	plr640pos4
26	26	plr974pos7	56	plr497pos3	86	plr602pos9	116	plr629pos9	146	plr862pos3
27	27	plr938pos6	57	plr638pos2	87	plr161pos2	117	plr521pos4	147	plr235pos9
28	28	plr935pos2	58	plr691pos8	88	plr777pos8	118	plr283pos9	148	plr728pos3
29	29	plr618pos6	59	plr238pos4	89	plr980pos6	119	plr135pos9	149	plr277pos4

Current approaches

Difficulty in predicting individuals

- Rely on “law of small numbers”
  - Acquire lots of good-probability players
  - Likely that some (enough?) will turn into good players



- Shift in scouting (facilitated by analytics/data science)
  - Statistics to predict range of outcomes (including injuries)
  - Scouts analyze psychology
    - Motivation, drive, self-control
    - “Big-market effect”

Current approaches

Include scout certainty estimates in player rankings

Team confidence in scout *v* ranking of players *i* and *j*

$$\sum_{v \in V} \sum_{(i,j) \in C_V} \left((r_v(i) - r_v(j)) \sqrt[a]{n_v} \sqrt[b]{t_{v,i} t_{v,j}} \sqrt[c]{u_{v,i} u_{v,j}} \right) f(p_i, p_j)$$

Scout *v* confidence in his ranking of players *i* and *j*

- Has been used by >15% of Major League Baseball teams

Streib, Young, & Sokol (2012)

Personalized training regimens

- Get player to maximum readiness at right time
  - Large *n* can take over?
    - Daily performance data (training/practice)
    - Millions of people with wearable devices
      - But only outlier performance matters!
  - Psychology still matters
    - Less data; more research

  - *Many* professional (and college) teams are working on this!

In-game seat upgrades

- Seat upgrades at sporting events
  - …as early as possible (*before* game time?)
  - …low probability of error

- Is individual fan behavior predictable?
  - With the right data…
    - Ticket scan times
  - …and the right probabilistic and statistical models
    - Sorry, NDA

Takeaways

- Predictive sports analytics is different
  - Individual human behavior: small *n* (or even *n*=1)
- Important effects under investigation
  - Daily performance variation
  - Physiological & psychological indicators
  - How to incorporate judgement uncertainty
  - Lots of interesting cross-disciplinary questions
    - “Analytics plus…”