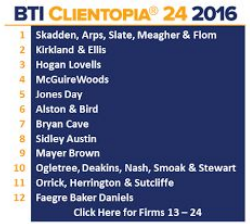
CSE 519: Data Science
Steven Skiena
Stony Brook University

Lecture 8: Scores and Rankings

Scores and Rankings

Scoring functions are measures that reduce multi-dimensional data to a single value, to highlight some particular property.

Rankings order items, usually by sorting scores.



Assigning Grades

Course grades get assigned by scoring functions. Observe that grading systems have:

- Degrees of arbitrariness: each teacher differs.
- Lack of validation data: there is no *right* grade.
- General robustness: students tend to get similar grades in all their classes anyway.

Calling scores *statistics* lends them more dignity.

Scoring vs. Regression

The critical issue in designing scoring functions is that there is no gold standard/right answer.

Machine learning techniques like linear regression can learn a scoring function from features if you had training data, which generally you don't.

Google's ranking algorithms train on click data.

The Body-Mass Index (BMI)

BMI is a score designed to capture whether your weight is under control:

$$BMI = \frac{mass}{height^2}$$

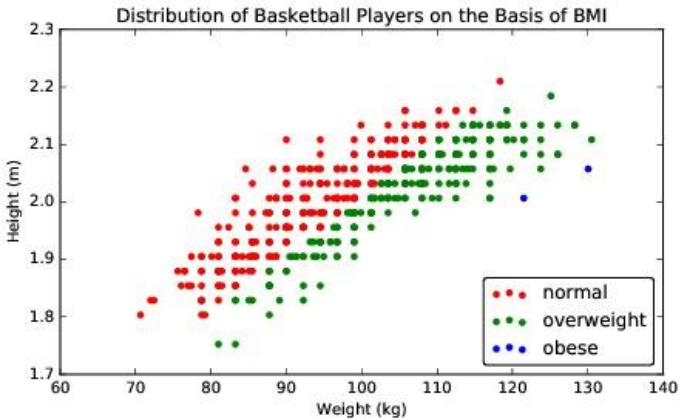
Mass is in kg and height in meters

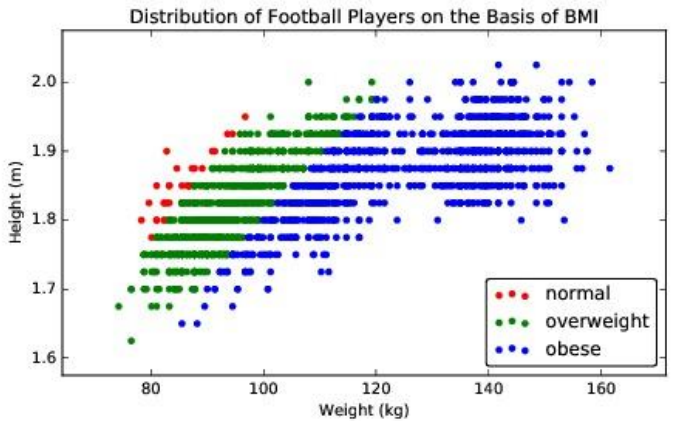
- Underweight: below 18.5
- Normal: 18.5 to 25
- Overweight: 25 to 30
- Obese: over 30

$$Skiena = \frac{68.0}{(1.727^2)} = 22.8$$

BMI: Pro Basketball and US Football

BMI is easy to interpret, and correlates with body fat. Mass should scale with the square (or cube) of height.





Gold Standards and Proxies

Gold standards are labels or answers we trust to be correct, reflecting the scoring goal.

Proxies are available quantities correlated with what we want to measure.

Your GPA or SAT/GRE is a proxy for how you should do in my class.

Scoring vs. Machine Learning

When you have a gold standard, you can train a regression function to accurately predict things.

When all you have are proxies, all you can do is evaluate your scoring function.

“Weapons of *Math* Destruction” happen when you confuse proxies for gold standards: e.g. student test scores for teaching quality.

Scores vs. Rankings

Which is more interpretable depends on:

- Will the numbers be presented in isolation?

Stony Brook ranks 111th of 351 teams RPI=39.18

- What is the distribution of scores?

How much better is #1 than #2?

- Do you care about the middle or extremes?

Small changes in score can cause big rank diffs

Recognizing Good Scoring Functions

- Easily computible
- Easily understandable
- Monotonic intepretation of variables
- Produces satisfying results on outliers
- Uses systematically normalized variables

Normalization and Z-scores

It is critical to normalize different variables to make their range/distribution comparable.

Z-scores are computed: $Z_i = (X_i - \bar{X})/\sigma$

Z-scores of height measured in inches is the same as height measured in miles.

Your biggest analysis sins will come in using unnormalized variables for analysis!

Z-score Examples

Z-scores have mean 0 and sigma=1.

Thus Z-scores of different variables are of comparable magnitude.

The sign identifies if it is above/below the mean.

$$\mu(B) = 21.9 \quad \sigma(B) = 1.92$$
$$\mu(Z) = 0 \quad \sigma(Z) = 1$$

B	19	22	24	20	23	19	21	24	24	23
Z	-1.51	0.05	1.09	-0.98	0.57	-1.51	-0.46	1.09	1.09	0.57

Advanced Ranking Techniques

Linear combinations of normalized values generally yield reasonable scores, but other techniques include:

- Elo rankings
- Merging rank orderings
- Directed graph orderings
- PageRank

Binary Comparisons

Rankings are often formed by analyzing series of binary comparisons:

- Team A beats team B
- Expert votes for A instead of B
- Student choses university A over B

Vote counts fail to pick the best when different teams face different levels of competition.

Elo Rankings

After starting equally ranked, scores are then adjusted to reflect the surprise of each match.

$$r'(A) = r(A) + k(S_A - \mu_A)$$

S is the actual score (-1,1) for *A*, with *mu* the expected score from the previous *r(A)* and *r(B)*.

Parameter *k* modulates the maximum possible swing in any one match.

What is the Expected Match Score?

If P(A>B) estimates the probability that A beats B, then:

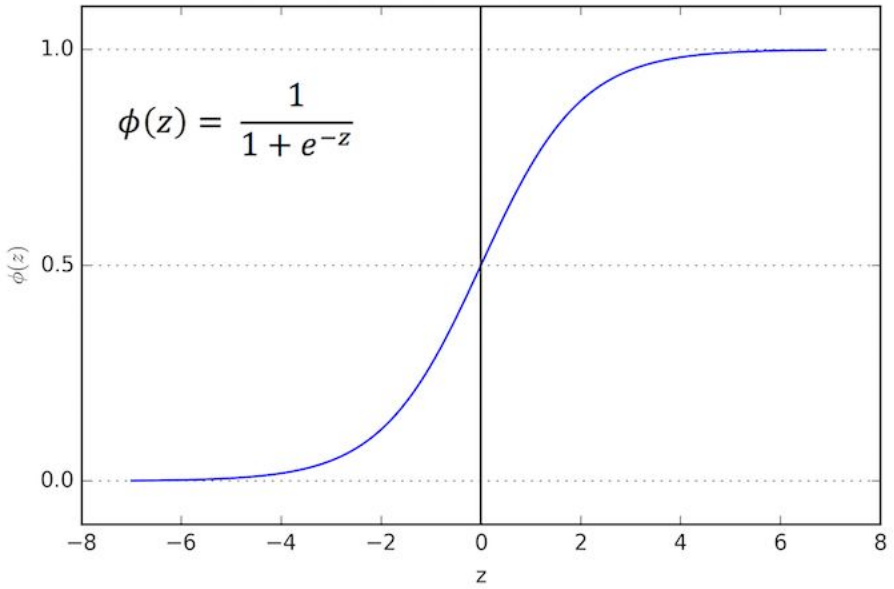
$$\mu_A = 1 \cdot P_{A>B} + (-1) \cdot (1 - P_{A>B})$$

If the ranking system is meaningful, this probability should be a function of the difference between the scores *r(A)* and *r(B).*

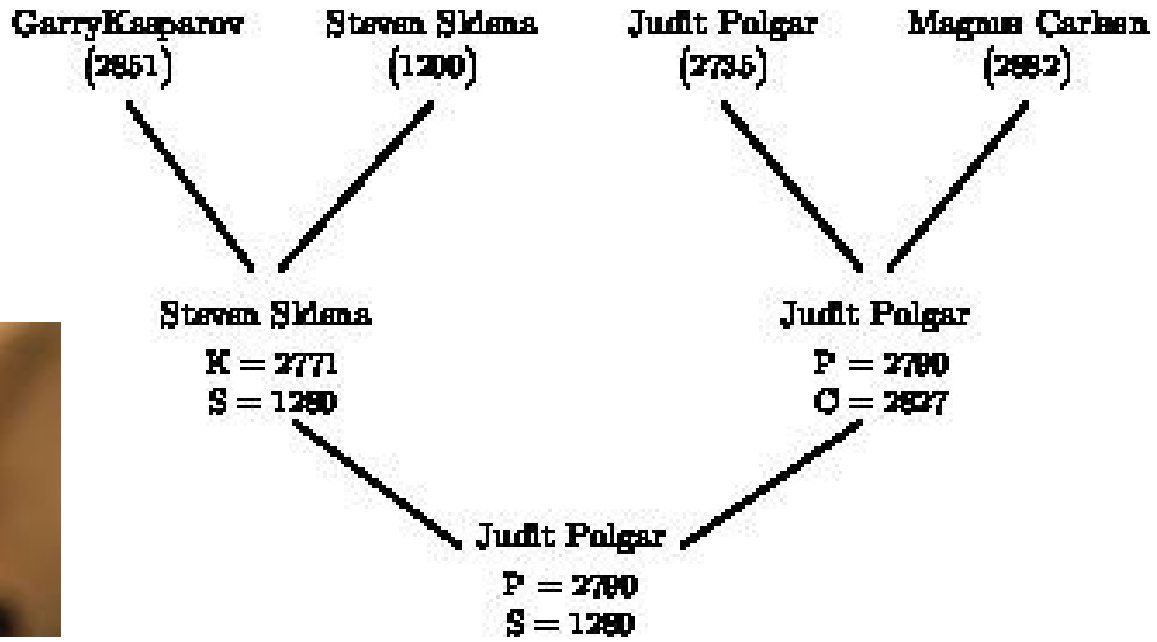
The Logit Function

We need a function f(x) that takes x and yields a probability:

- f(0) = ½
- f(infty) = 1
- f(-infty) = 0



Elo Chess Ranking Example



Merging Rankings / Votes

Consider determining the winner of an multiparty election where each voter ranks the candidates in order of preference.

1. Stony Brook 2. MIT 3. Illinois 4. ….

Equivalently, consider merging rankings independently drawn on different features.

Borda's Method

By assigning an increasing score per position, the resulting point total ranks the items:

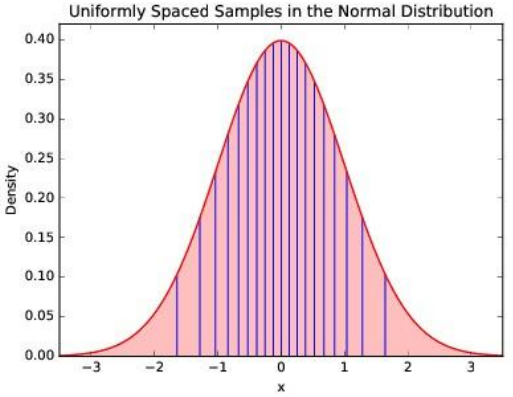
1	A	B	A	A	A:5
2	C	A	B	B	B:8
3	B	C	C	D	C:12
4	D	D	E	C	D:16
5	E	E	D	E	E:19

Four voters, each ranking five items.

Weights for Borda's Method

Linear position weights make sense when we have equal confidence across all positions.

But we presumably trust our distinctions among the best/worst more than the middle elements, suggesting normally distributed weights.



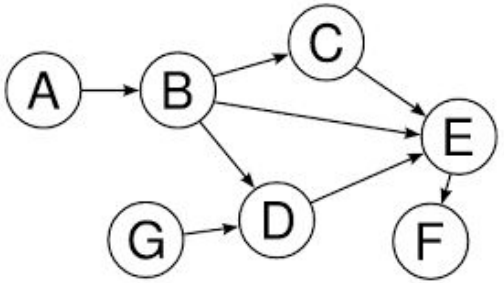
Directed Graph Orderings

Treating the vote (A>B) as an edge (A,B) yields a directed graph.

If there are no inconsistencies, we get a directed acyclic graph (DAG).

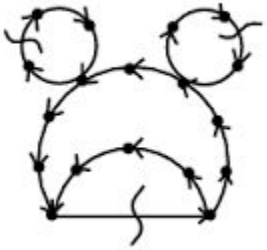
Topologically sorting this DAG gives a reasonable order, like

ABCGDEF or GABCDEF



Ranking General Digraphs

For general directed graphs, we seek the order minimizing the number of ``wrong way'' edges.



Cutting the minimum number of edges to leave a DAG is NP-complete.

But reasonable heuristics start by sorting by the difference between in/out degree.

Arrow's Impossibility Theorem

There is no ranking system that satisfies all desirable properties:

- The system should be complete: given A and B it must pick one or say equal preference.
- The system must be transitive, meaning that if A>B and B>C then A>C.
- If every voter prefers A to B, then A wins over B.
- Preferences cannot depend only on one dictator.
- The preference of A to B should be independent of preferences for all other candidates.

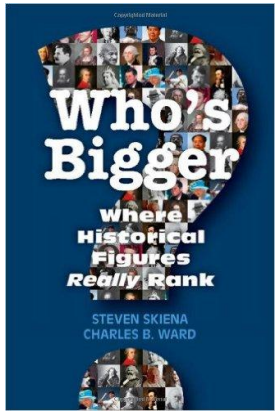
Voter	Red	Green	Blue
X	1	2	3
Y	2	3	1
Z	3	1	2

Red beats Blue, Green beat Blue, Blue beats Red

Ranking Example: Who's Bigger?

Analyzed Wikipedia to extract measures of historical significance: PageRank, length, hits…

- Mapped values to normal distributions
- Use linear combination (factor analysis)
- Corrected for time by decaying modern figures.
- Separate scores for celebrity/gravitas.



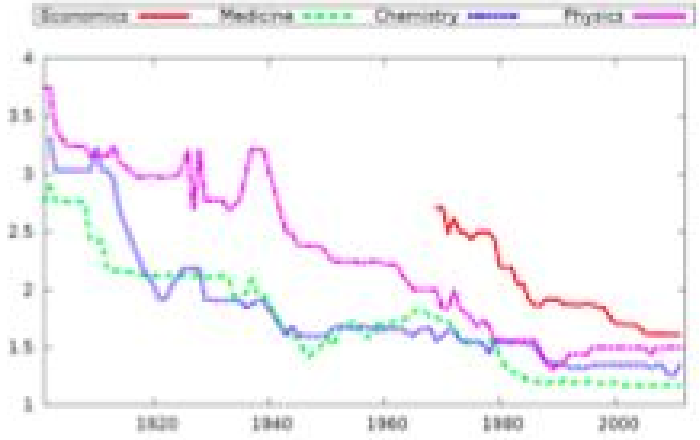
Who's Biggest?

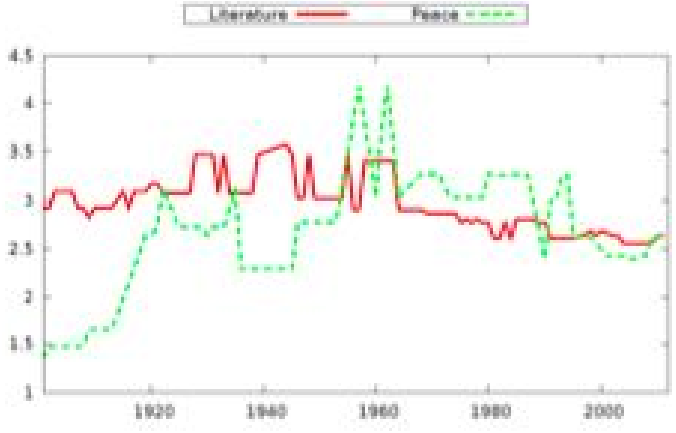
Here are the top 20 most significant historical figures among over 800,000 in the English Wikipedia:

Rank	Name	Dates	Description
1	Jesus	(7 B.C.–30 A.D.)	Central figure of Christianity
2	Napoleon	(1769–1821)	French military leader and emperor
3	William Shakespeare	(1564–1616)	English playwright ("Hamlet")
4	Muhammad	(570–632)	Founder of Islam
5	Abraham Lincoln	(1809–1865)	16th U.S. President (Civil War)
6	George Washington	(1732–1799)	1st U.S. President (Revolution)
7	Adolf Hitler	(1889–1945)	Fuehrer of Nazi Germany (WW II)
8	Aristotle	(384–322 B.C.)	Greek philosopher and scientist
9	Alexander the Great	(356–323 B.C.)	World conqueror (Greek)
10	Thomas Jefferson	(1743–1826)	3rd U.S. Pres. (Decl. of Independence)
11	Henry VIII	(1491–1547)	King of England (6 Wives)
12	Elizabeth I	(1533–1603)	Queen of England (The Virgin Queen)
13	Julius Caesar	(100–44 B.C.)	Roman general and statesman (Et tu, Brute?)
14	Charles Darwin	(1809–1882)	Scientist (Theory of Evolution)
15	Karl Marx	(1818–1883)	Philosopher ("Communist Manifesto")
16	Martin Luther	(1483–1546)	Protestant Reformation (95 Theses)
17	Queen Victoria	(1819–1901)	British Queen (Victorian Era)
18	Joseph Stalin	(1878–1953)	Russian leader (World War II)
19	Theodore Roosevelt	(1858–1919)	26th President (Spanish-American War)
20	Albert Einstein	(1879–1955)	Physicist (Theory of Relativity)

The Decline of the Great Scientist

The magnitude of Nobel Prize scientists is declining, but not literature/peace winners...





What Can You Learn from Rankings?

- Women are underrepresented in Wikipedia.
- Halls of Fame / textbooks do not always pick the strongest figures.
- Certain fields (e.g. poetry) are not producing historically significant figures as in the past.