Who's Bigger? A Quantitative Analysis of Historical Fame

Steven Skiena[1] Charles Ward[1]

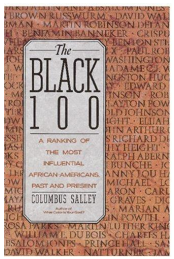
[1] Department of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400

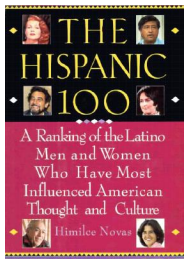
June 1, 2012

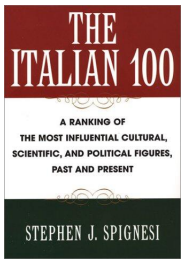
Outline

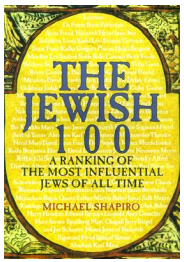
Who's Bigger: Historical Rankings

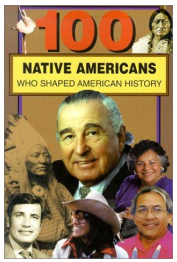
Numerical ratings/rankings provide a way to focus greater attention on the best/most important things.

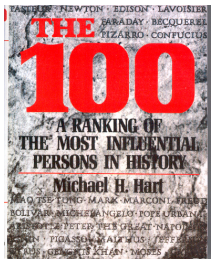








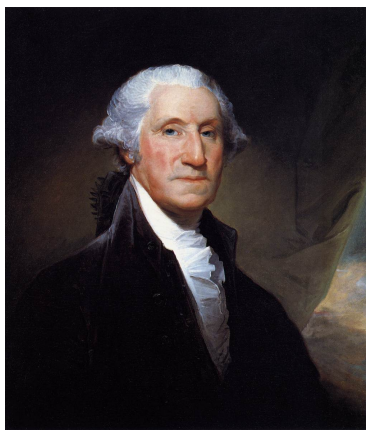


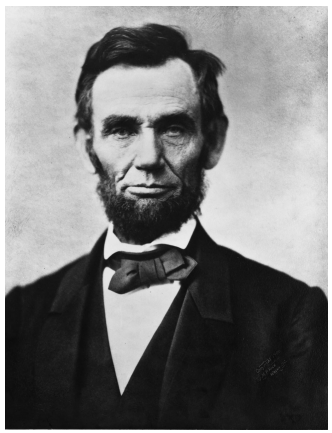


Rankings are highly subjective and culturally biased, yet rankings provide a popular mix of education and entertainment.

We seek algorithms to construct informative and meaningful historical rankings of all the people described in Wikipedia.

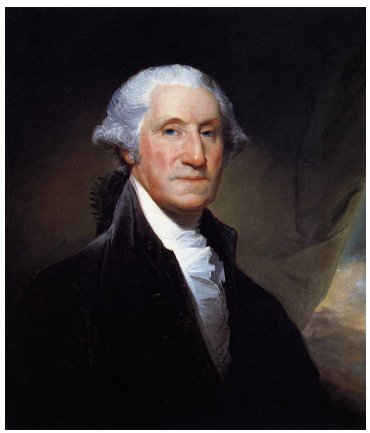
Who's Bigger? (Presidents)

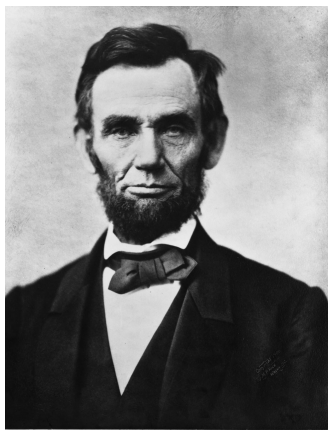




Battle: George Washington vs. Abraham Lincoln

Who's Bigger? (Presidents)

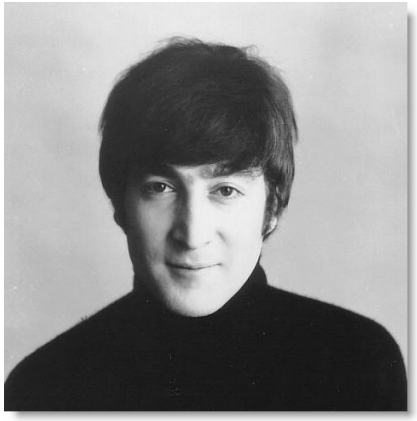




Battle: George Washington vs. Abraham Lincoln

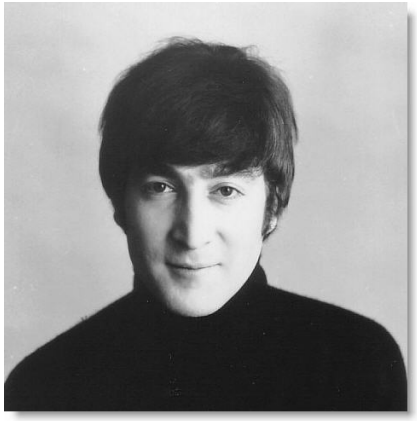
- *George Washington (1732–1799)* [6]
- *Abraham Lincoln (1809–1865)* [5] *
- *Chester A. Arthur (1829–1886)* [490]

Are the Beatles "Bigger than Jesus"?



Battle: Jesus vs. John Lennon

Are the Beatles "Bigger than Jesus"?



Battle: Jesus vs. John Lennon

- *Jesus (7 B.C.–30 A.D.)* [1] *
- *John Lennon (1940–1980)* [141]
- *Paul McCartney (1942–)* [399]
- *George Harrison (1943–2001)* [615]
- *Ringo Starr (1940–)* [1729]

Roll Over Beethoven?

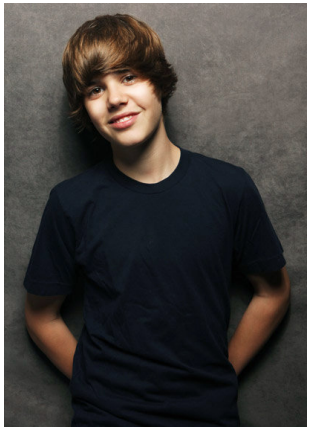
Battle: Elvis Presley vs. Ludwig van Beethoven

Roll Over Beethoven?

Battle: Elvis Presley vs. Ludwig van Beethoven

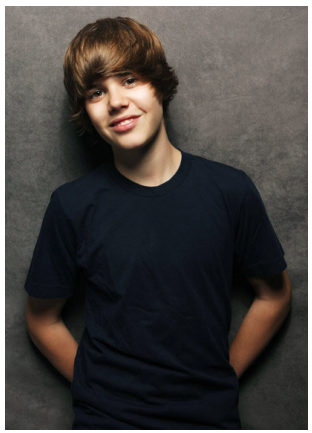
- *Ludwig van Beethoven (1770–1827)* [28] *
- *Elvis Presley (1935–1977)* [58]
- *Pyotr Ilyich Tchaikovsky (1840–1893)* [57]
- *Franz Liszt (1811–1886)* [108]

Justin Time or Forever?



Battle: Justin Bieber vs. Eli Whitney

Justin Time or Forever?

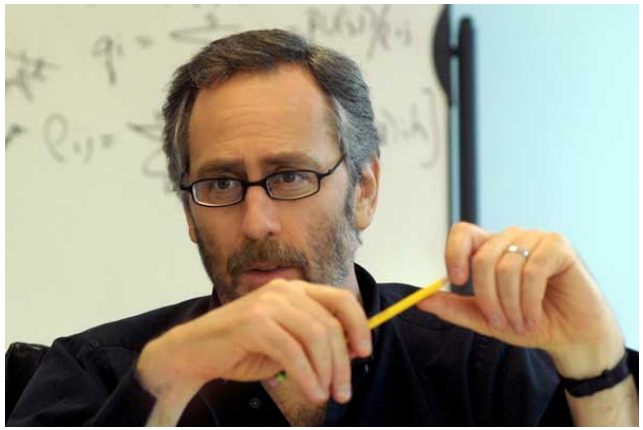


Battle: Justin Bieber vs. Eli Whitney

- *Justin Bieber (1994–)* [7718] (fame rank 1479)
- *Eli Whitney (? – ?)* [857] * (fame rank 2423)
- *Frankie Avalon (1940–)* [17125] (fame rank 12693)

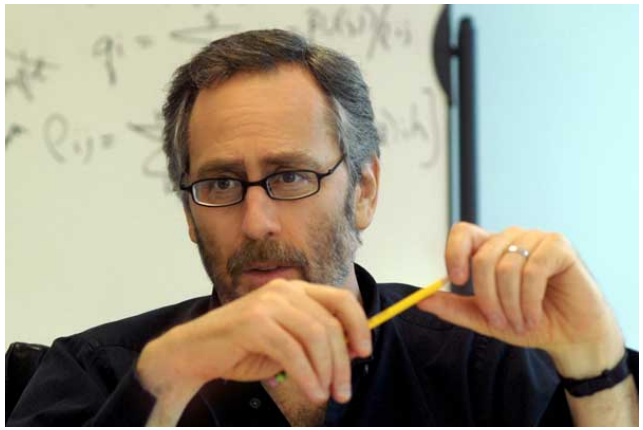
Bieber is more famous but less historically significant.

Who's Bigger? (Hedge Fund Directors)



Battle: Jim Simons vs. D. E. Shaw

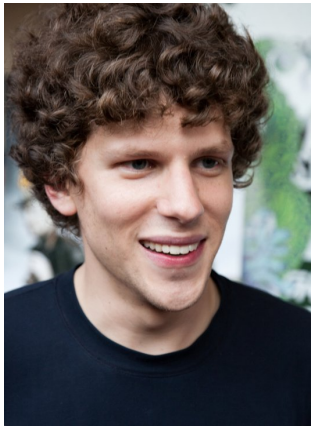
Who's Bigger? (Hedge Fund Directors)



Battle: Jim Simons vs. D. E. Shaw

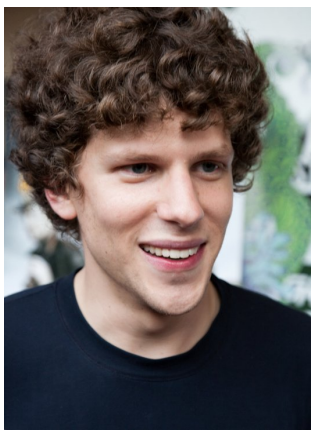
- *James Harris Simons (1938–)* [70874] *
- *David E. Shaw (1951–)* [230986]

Who's Bigger? (Facebook)



Battle: Mark Zuckerberg vs. Jesse Eisenberg

Who's Bigger? (Facebook)



Battle: Mark Zuckerberg vs. Jesse Eisenberg

- *Mark Zuckerberg (1984–)* [7468] * (fame rank 1427)
- *Jesse Eisenberg (1983–)* [41324]

The Historical Top 20

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)

Culturenomics

The Big Data revolution is changing how research is done, including the humanities and social sciences.

New data sets drawn from massive text corpora let us watch history unfold, and measure seemingly unquantifiable aspects of fame.

We work with sociologists to apply large-scale news/text analysis to better understand the processes of fame and reputation.

One aspect of this work has been the development of methods to measure people's historical significance...

Not Just a Toy

Applications of significance ranking include:

- Klout score beyond Twitter.
- Entity disambiguation: *Larry Page (1973–)* [11665] vs. *Larry Page (? – ?)* [116064]
- Esoteric content detection for measuring document readability.
- Objective reference standard for bias detection.

Outline

Ranking Methodologies

Ranking documents by significance lies at the heart of Internet search engines like Google.

Ranking people by merit is the goal of college admissions and the job hiring process.

Subjective rankings of historical figures appear frequently in books / magazines, when awarding prizes / honors, and in populating textbooks.

Traditional approaches to ranking include:

- Expert Polls
- Public Voting
- Single Variable models
- Multiple Factor models

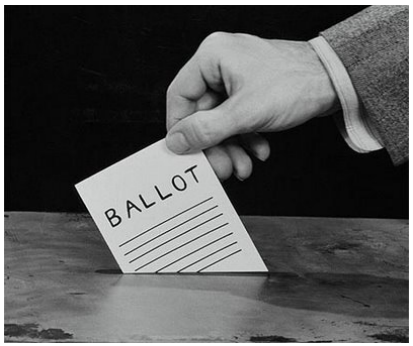
Expert Polls

Examples: UPI/Associated Press Top 20 College Football rankings, the Academy Awards/Oscars; Historian Presidential rankings.

Strengths: Seems fair; knowledgable people usually make sound decisions; safety in numbers.

Weaknesses: Group-think from lack of diversity and second-hand knowledge; political bias, difficult to compare across varied / less popular domains.

Public Voting



Examples: Democratic elections; All-star team selection; the Internet Movie Database (IMDb) *STARmeter*.

Strengths: The wisdom of crowds hypothesis dictates that large, diverse groups can make better decisions than individual experts.

Weaknesses: The public has very limited historical knowledge; organized special-interest campaigns can skew selection.

How often does the public select the right candidate in an election?

Single Factor Models

Examples: The Forbes 400 rankings by wealth; University rankings by SAT scores or acceptance percentage.

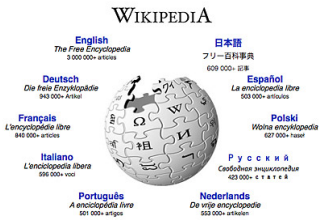
Strengths: Simplicity and clarity.

Weaknesses: Single statistics paint a one-dimensional picture of reputation; individual statistics are often quite easy to game.

Ironically, artistic achievements are often ranked by money: bestselling books, highest grossing films, top-40 record countdown.

We will employ models built from multiple factors.

Our Historical Universe: Wikipedia



We rank all the people with pages in the English edition of Wikipedia, a population roughly equal to that of San Francisco.

The least significant person is *Dejan Paji (1989–)* [771384], a Serbian sprint canoer who won a bronze medal in the K-2 500 meter event at the 2010 World Championships.

Several measures of fame/significance can be found in Wikipedia...

PageRank

Wikipedia pages link to other Wikipedia pages through the text of articles, defining a network.

Links from important people to your page means you are probably important.

Google's PageRank algorithm measures the centrality of a vertex/page in a network of links.

High/Low PageRank Individuals

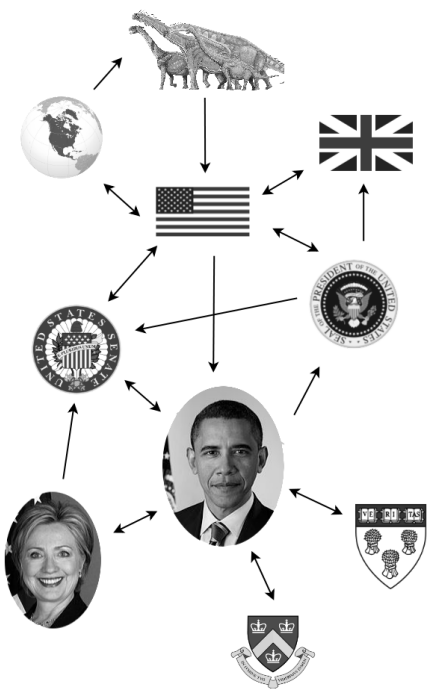
Person	PR1	Sig
Napoleon	1	2
George W. Bush	2	29
Carl Linnaeus	3	46
Jesus	4	1
Barack Obama	5	91
Aristotle	6	8
William Shakespeare	7	3
Elizabeth II	8	125
Adolf Hitler	9	7
Bill Clinton	10	101

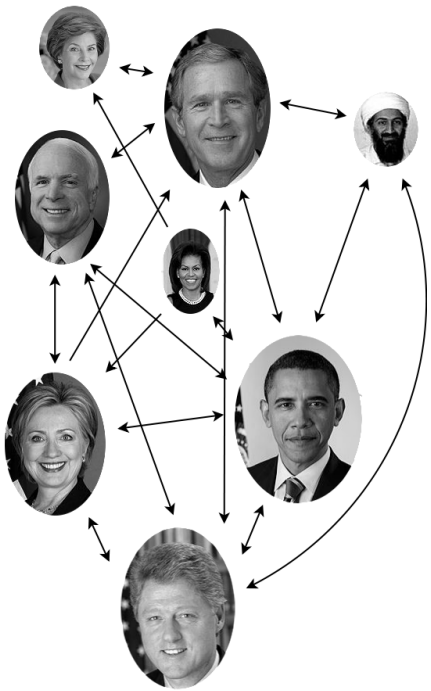
Person	PR1	Sig
Vijay	16463	4269
Dave Batista	16262	4297
CM Punk	14784	4988
Daniel Radcliffe	12219	6462
Jesse McCartney	11704	3556
Randy Orton	10966	3551
Ashley Tisdale	10184	3992
Ashanti	9990	3923
Edge	9956	2308
Brock Lesnar	9296	4003

Famous low PageRank people include young celebrities.

Why does *Carl Linnaeus (1707–1778)* [46] have such high PageRank?

Should Dinosaurs Vote?





Restricting vertices in the graph to people yields a different network to compute PageRank on.

High/Low People PageRank

Person	PR2	Sig
George W. Bush	1	29
Bill Clinton	2	101
William Shakespeare	3	3
Ronald Reagan	4	27
Adolf Hitler	5	7
Barack Obama	6	91
Napoleon	7	2
Richard Nixon	8	78
Franklin D. Roosevelt	9	41
Elizabeth II	10	125

Person	PR2	Sig
Richard Stallman	16693	4831
John Cabot	14282	389
Ashlee Simpson	14171	3678
Jimmy Wales	13307	2150
Vijay	12720	4269
Dave Batista	12298	4297
Ashley Tisdale	12285	3992
Jacques Cartier	12175	360
Jesse McCartney	11846	3556
Nicole Scherzinger	10687	5460

Explorers and programmers are propped up by their organizations.

We use both PageRanks in our final computation.

Hits



An orthogonal measure of a person's significance is how frequently readers visit their Wikipedia page.

More famous/interesting people should have their pages read more frequently than lesser lights.

Hits measures the number of Wikipedia readers, while PageRank depends upon actions by the authors of Wikipedia pages.

High/Low Hit Individuals

Person	Hits	Sig
Eminem	3	755
Lady Gaga	5	2142
Adolf Hitler	6	7
Lil Wayne	7	1707
Katy Perry	9	4647
Rihanna	10	1089
Barack Obama	12	91
Michael Jackson	13	136
Kanye West	15	1373
Miley Cyrus	16	1719

Person	Hits	Sig
Gough Whitlam	11391	1047
Paul Martin	9332	2989
Pope Leo XIII	9154	406
Charles Sanders Peirce	8469	229
Brian Mulroney	7966	2415
Joseph Priestley	7725	331
George Galloway	7594	4343
Lester B. Pearson	7525	1180
Suharto	7521	1738
Lil Kim	7366	3991

Adolf Hitler (1889–1945) [7] is the only non-contemporary figure on the frequently read list.

Gough Whitlam (1916–) [1047] is an important / controversial former Australian Prime Minister.

Article Length

Wikipedia article length provides a natural measure of fame: more significant people merit longer articles.

Article length is not the hard constraint of printed texts, yet clear social pressures by the Wikipedia community favor conciseness.

Over 100,000 people have suffered the ignominy of having their (usually autobiographical) articles removed from Wikipedia.

High/Low Article Length Individuals

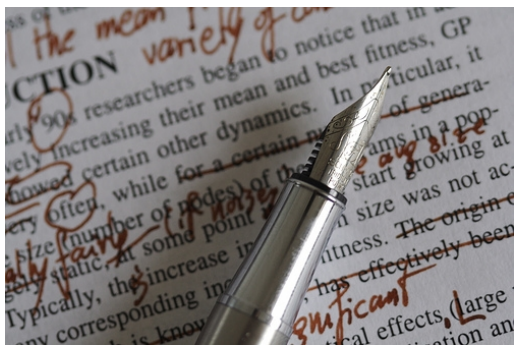
Person	Words	Sig
Adolf Hitler	5	7
Stanley Kubrick	12	1811
Elvis Presley	14	58
Joseph Stalin	18	18
L. Ron Hubbard	21	1045
Che Guevara	25	429
Paul Robeson	27	1235
Janet Jackson	28	461
Michael Jackson	29	136
Douglas MacArthur	30	285

Person	Words	Sig
Euclid	37084	152
Tony Hawk	35701	3632
Hugh Hefner	35681	3494
Vijay	31906	4269
Sean Hannity	31509	4340
Ja Rule	31313	1241
Fergie	29737	3045
Euripides	27290	355
Will Smith	25942	2611
John Travolta	24479	1573

Controversial people get long articles, while certain contemporary celebrities have few accomplishments to write about.

The longest prominent article (Hitler) runs 29341 words, where the shortest (*Euclid (? – ?)* [152]) contains only 1542.

Page Edits



The Wikipedia collaborative model empowers thousands to contribute their knowledge to the world.

Famous/important people have more refined articles than lesser personages, because more readers will have both desire and information to contribute.

High/Low Page Edit Individuals

Person	Edits	Sig
George W. Bush	1	29
Michael Jackson	2	136
Jesus	3	1
Britney Spears	4	566
Adolf Hitler	5	7
Barack Obama	6	91
Muhammad	7	4
Elvis Presley	8	58
Roger Federer	11	746
Mariah Carey	12	531

Person	Edits	Sig
Tacitus	7184	287
Francis I	6777	344
Pope Leo XIII	6713	406
Toyotomi Hideyoshi	6708	365
Plutarch	6638	236
Josephus	5910	347
Friedrich Engels	5820	402
Jerome	5506	298
Aristophanes	5380	407
George II	5377	335

High-edit people tend to be contemporary, or religious figures with active constituencies.

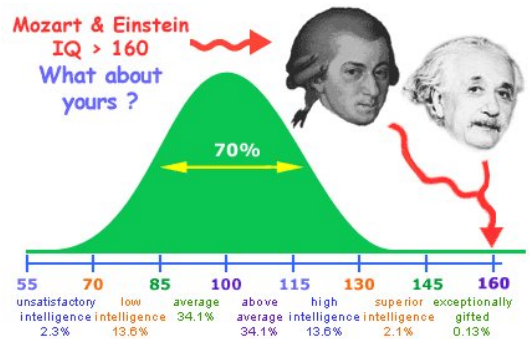
Prominent low-edit people tend to be dead for over 1,000 years.

Statistical Factor Analysis

Factor analysis is a statistical tool used to describe the communality of variables in terms of a small set of unobserved variables, or "factors."

Getting fit for a tux requires measurements of height, weight, shoe size, inseam length, waist size, neck size, jacket length.

Yet two factors explain most of them: girth and span.



Charles Spearman (1863–1945) [15988] developed factor analysis to prove one factor underlies performance on many types of intelligence (IQ) tests.

Factor Score Loadings

Variable	F1 Loading	F2 Loading
Full Pagerank	0.403	0.912
Person Pagerank	0.401	0.630
Pagehits	0.697	0.485
No. of Revisions	0.829	0.395
Article Length	0.360	0.184
News Hits	0.376	0.167

Two factors pop out, each explaining roughly the same proportion of variance (31% and 28%).

Celebrity vs. Gravitas

These factors have natural interpretations as celebrity (F1) and gravitas (F2).

Gravitas loads primarily on the two forms of PageRank.

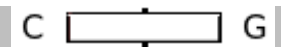
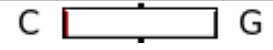
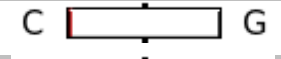
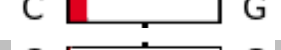
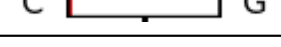
Celebrity loads heavily on page hits, revisions, and article lengths.

These factors distinguish popular personalities from lower-profile people with heftier achievements.

We define significance to be the sum of the celebrity and gravitas factors.

High Gravitas/Celebrity Individuals

Person	Grav.	Celeb/Grav
Napoleon	6	
Carl Linnaeus	21	
Aristotle	26	
William Shakespeare	27	
Plutarch	29	
F. D. Roosevelt	31	
Charles II	33	
Elizabeth II	34	
Pliny the Elder	36	
Tacitus	37	

Person	Celeb	Celeb/Grav
Vijay	3	
Edge	5	
Kane	7	
John Cena	9	
Triple H	17	
Rey Mysterio	19	
Roger Federer	22	
Britney Spears	25	
Dave Batista	26	
Ashley Tisdale	32	

Professional wrestlers tend to have almost all their significance explained by celebrity.

The Ravages of Time

> *"My name is Ozymandias, king of kings:*
> *Look on my works, ye mighty, and despair!"*
> – Percy Bysshe Shelley (1792–1822) *[324]*

Contemporary figures are substantially overrated by uncorrected factor scores, with 28 of the 100 most famous individuals still alive.

Uncorrected significance does serve as an effective proxy to measure fame.

Britney Spears (1981–) [20] ranks ahead of *Aristotle (384–322 B.C.)* [24] in uncorrected significance!

Modeling Reputation Decay

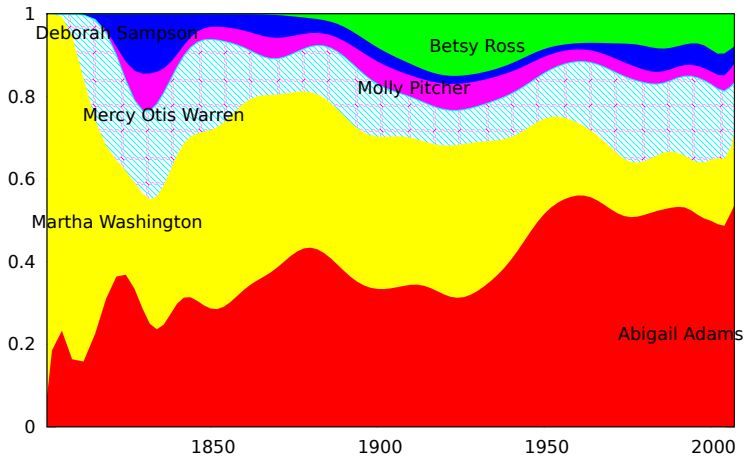
Historical figures do not have a "half-life", or else all ancient figures would have already been forgotten.

Estimating the decay rate is essential to appropriately compare the significance of current individuals with older figures long dead.

There are two distinct processes at work here: first the lapse from living memory inherent in the passage of generations, and second a more contemporary bias due to the advent of Wikipedia.

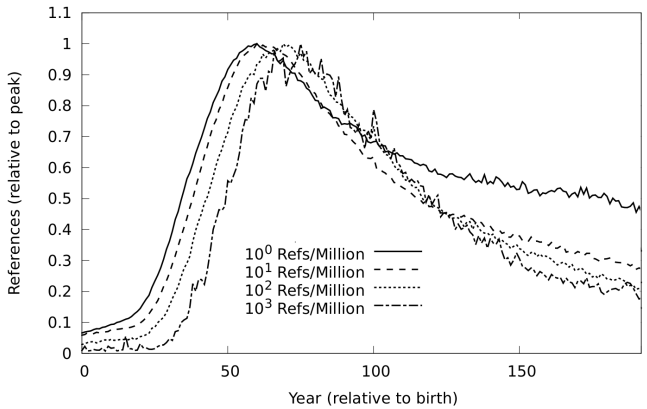
Google Book Ngrams: http://books.google.com/ngrams

Data to calibrate a historical reputation decay model comes from reference counts in millions of scanned books.



Betsy Ross (1752–1836) [2346] didn't exist in historically until 1870.

The Decay Rate of Fame

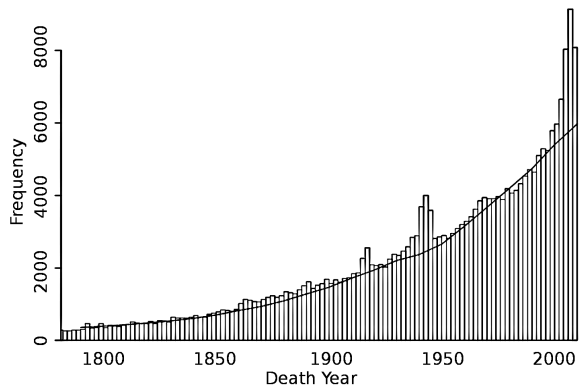


Reputations peak between age 60 and 75 and decay relatively slowly. Decay eventually stops, which is why we can still recall prominent ancient figures.

These decay rates permit us to project the reputation of contemporary figures 170 years into the future, to permit fair comparisons.

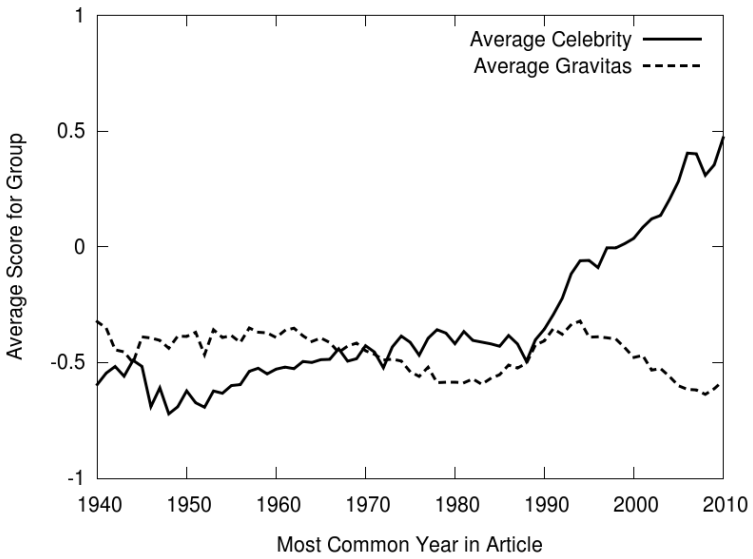
The Wikipedia Generation

For most of the past 200 years, the people in Wikipedia tracks the global (U.S. plus world) population *amazingly* well.



More famous people died during war years, as well as in the post-Wikipedia era.

The Wikipedia Effect: A Plague of Celebrity



Gravitas has held basically constant over the past seventy years, while the level of celebrity has exploded since 1990.

We correct for this by subtracting a fixed amount of celebrity from each contemporary person as a function of peak activity year, yielding our final measure of significance.

Validation: Gold Standards

We have assembled a collection of 35 published rankings (averaging about 100 people each) over a wide variety of domains, in history, sports, and entertainment, including:

- *1,000 years, 1,000 People: Ranking the Men and Women Who Shaped the Millennium*
- *AFI Screen Legends*, compiled by AFI historians.
- Internet polls from `www.thebest100lists.com`, ranking the top 100 athletes, authors, and movie directors.
- The Associated Press Top 100 athletes of the century, voted by a 16-member panel.
- Baseball star rankings by experts at `www.baseballevolution.com`.
- *The 100: A Ranking of the Most Influential Persons in History* (Hart)
- The Rolling Stone magazine top 100 singers.
- IMDB STARmeter rankings of actors, determined by search history.
- The *Time 100* social networking ranks, based on Twitter followers and Facebook connections.
- The United States Presidency Center's expert poll rankings of U.S. presidents.

These serve as gold standards to assess how well our rankings correspond to expert evaluation of fame/significance.

Performance: Rank correlations with published rankings

ListGroup	NL	ILA	NPR	PPR	PH	NR	F1	F2	F1+F2	dPR	dF1+F2
Actors	3	0.412	0.499	0.556	0.273	0.327	0.178	0.411	0.466	0.501	**0.523**
Actresses	3	0.419	0.491	0.514	0.272	0.389	0.226	0.349	0.501	0.509	**0.546**
Authors	3	0.353	0.419	0.426	0.415	0.358	0.189	0.358	0.436	0.429	**0.458**
Directors	5	0.491	0.586	0.562	0.431	0.502	0.364	0.466	0.576	0.600	**0.608**
Musicians	3	N/A	0.648	0.621	0.572	0.569	0.416	0.413	0.618	0.638	**0.672**
Individual Sports	10	0.280	0.459	0.457	0.408	0.406	0.316	0.381	0.453	0.462	**0.463**
General Athletics	3	0.569	0.489	**0.571**	0.467	0.463	0.369	0.323	0.537	0.497	0.554
US Presidents	5	0.909	0.576	0.490	0.625	0.549	0.386	0.532	0.580	0.623	**0.655**
General Historical	3	N/A	0.434	0.388	0.499	0.489	0.420	0.324	0.482	0.458	**0.511**
Overall	35	0.490	0.511	0.509	0.440	0.450	0.318	0.395	0.517	0.524	**0.554**

Time-corrected significance correlates best (0.554) with validation lists in essentially all categories.

Better than the Experts?

The experts did not all agree on their relative rankings of shared people.

Authors (0.363) and sports figures (0.280) showed particularly weak correlation, while there was consensus on the ranking presidents (0.909).

The inter-list correlation of 0.490 is substantially lower than the correlation (0.554) with our corrected significance measure.

This provides evidence that our measure is, overall, substantially superior to human rankings.

Outline

Trends Analysis

Our significance rankings provide an objective measuring stick to quantify and interpret historical phenomena, opening up a variety of applications. These can be used to tease out historical trends which might otherwise be difficult to pin down:

- The decline of the great scientist
- The evolution of the Papacy

The Decline of the Great Scientist...

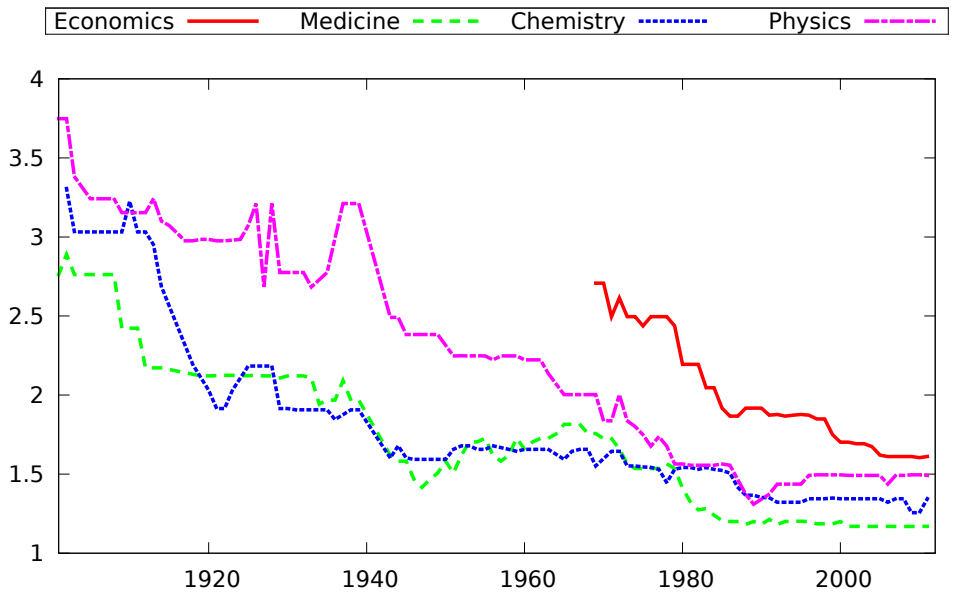


Figure: The Significance of Nobel Prize Winners in the Science and Economics

Economic laureates now out-rank the hard sciences, but all are declining.

While Literature and Humanity Endure

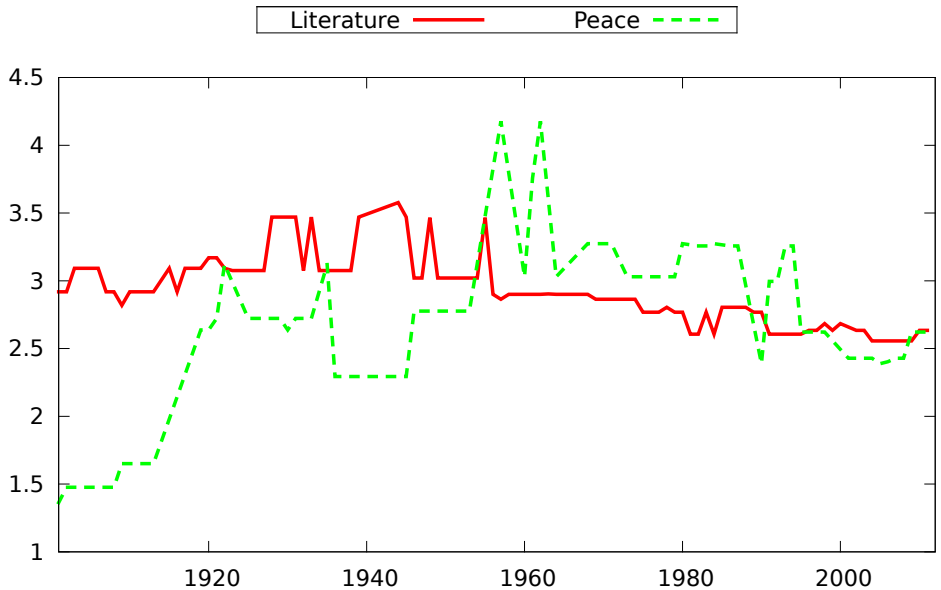
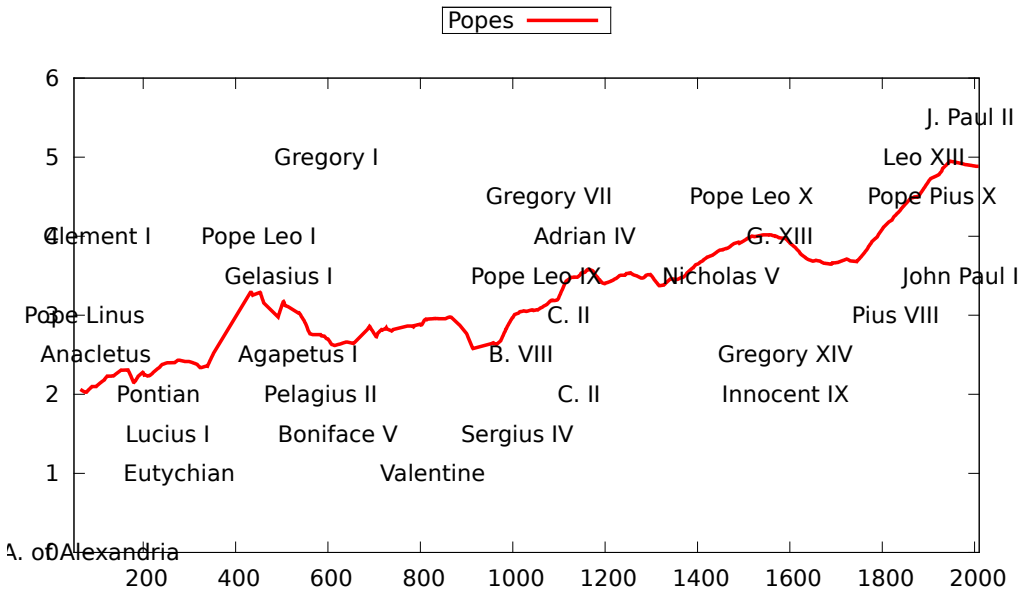


Figure: The Significance of Nobel Peace and Literature Prize Winners

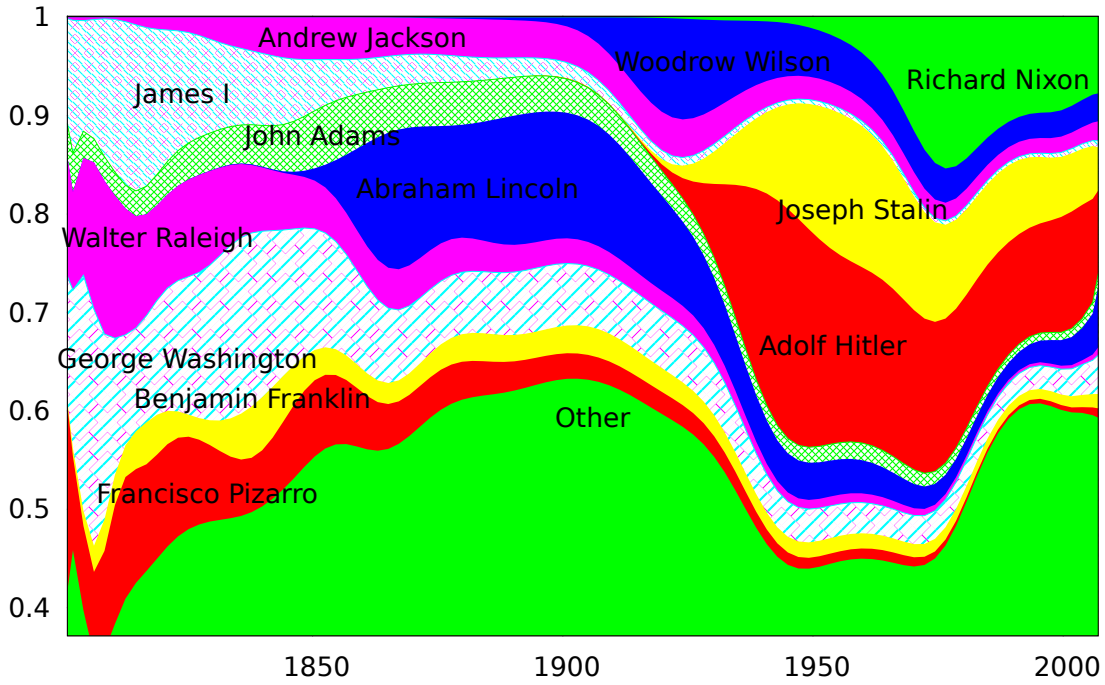
The Evolution of the Papacy



Papal significance rose with Roman Christianity, fell with Rome, rose with the Crusades/Renaissance, fell with the Enlightenment, and have continued to rise through modern times.

Analyzing History Textbooks and Teaching Standards

Do the figures canonized in history textbooks truly reflect the historical significance of those figures, or are other forces at work?



Who Belongs in the History Book?

ID	Name / Rank
1	Mary Antin
2	Nat Love
3	George Shima
4	Benjamin Pap Singleton
5	Luzena Wilson

ID	Description
A	Cowboy and author of a popular autobiography.
B	Entrepreneur who ran hotels and restaurants during the California Gold Rush.
C	Immigrant from Japan who became known as the "Potato King" for his successful farming of potatoes in California
D	Immigrant from Russian who published a popular autobiography called "The Promised Land".
E	Leader of African American homesteaders known as Exodusters.

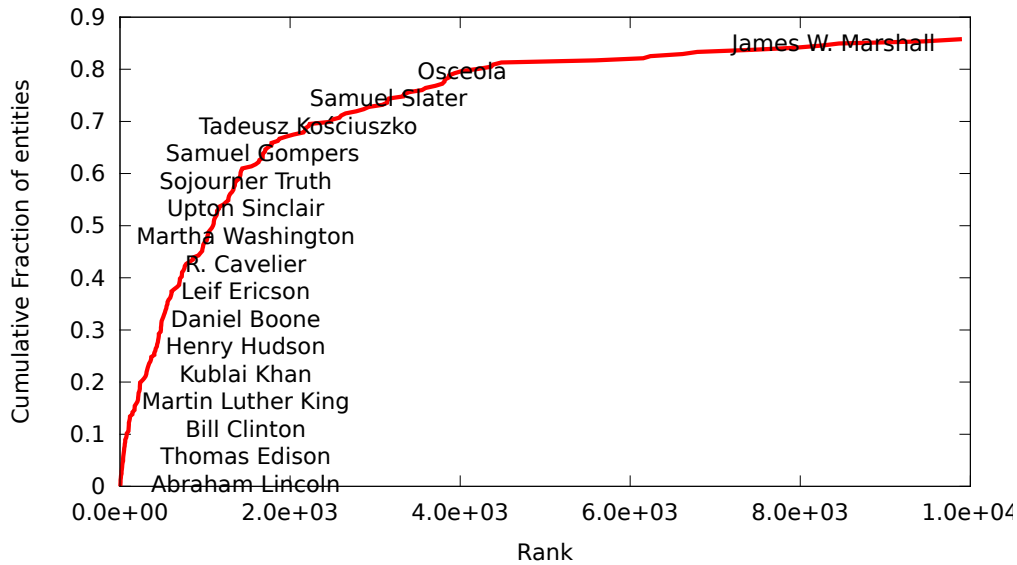
Correct Answers [1D, 2A, 3C, 4E, 5B]

The Most/Least Significant People Highlighted in a Fifth Grade History Textbook[1]

Sig.	Person	Sig.	Person
5	Abraham Lincoln	410,979	B. P. Singleton
6	George Washington	212,518	Luzena Wilson
7	Adolf Hitler	158,240	George Shima
10	Thomas Jefferson	106,237	Mary Antin
12	Elizabeth I	101,015	Venture Smith
18	Joseph Stalin	50,230	H. Theyanoguin
19	Theodore Roosevelt	41,536	Dolores Huerta
22	C. Columbus	38,653	Nat Love
25	Ulysses S. Grant	28,243	Joseph Cinquè
27	Ronald Reagan	28,048	Peter Salem

[1] Pearson Scott Foresman's *Social Studies: The United States* (Gold Edition), 2008.

Cumulative Distribution by Significance



Almost half of the 246 textbook figures appear in our top 1000 most historically significant entities.

But the bottom 35 figures all rank below $10,000$; idiosyncratic choices inappropriate for a fifth grade textbook.

Political Correctness?

Very few of the questionable choices appear in any of the 16 state teaching standards we analyzed, so the states are not to blame.

Of the 50 least significant people in the text, 11 were Native American and 13 more were African-American.

This seems unnecessary: 9 of the 100 *most* significant figures in the text were African-American, and 16 Native-Americans in the text ranked in the top 5000 in historical significance.

For each ethnic group, we can identify excluded people who are substanially more significant, and hence worthy of inclusion.

Evaluating Human Selection Processes

Human selection processes have critical implications with respect to job hiring, college admissions, sports drafts, and democratic elections.

Meritocracies rest on the precision with which society can make accurate judgements about the accomplishments and potential of people.

To what extent can people be trusted to get these decisions correct?

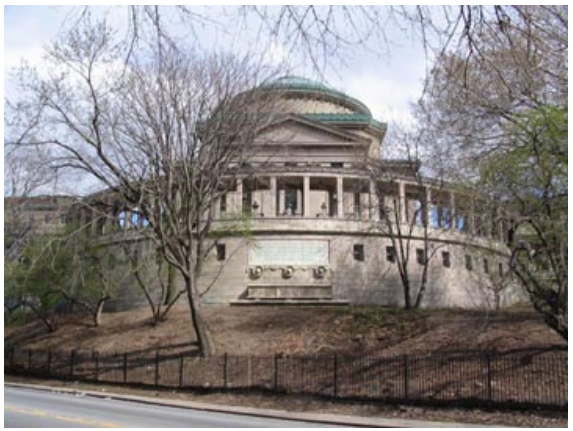
Studies of Hall of Fame elections provide an interesting laboratory to measure the extent to which experts can recognize historical significance.

HoF elections seek to recognize *achivement*, which should be substantially easier than evaluating *potential* (e.g. sports drafts or college admissions).

Michael Jordan (1963–) [1018] vs. *Sam Bowie (1961–)* [33949]?

The Hall of Fame for Great Americans

For over 70 years starting in 1900, prominent electors from this Bronx institution voted every five years to select the greatest Americans.



How well did they do?

The Most/Least Significant Great Americans

Sig	Person	Dates
5	Abraham Lincoln	(1809–1865)
6	George Washington	(1732–1799)
10	Thomas Jefferson	(1743–1826)
19	Theodore Roosevelt	(1858–1919)
25	Ulysses S. Grant	(1822–1885)
36	Benjamin Franklin	(1706–1790)
38	Thomas Edison	(1847–1931)
41	F. D. Roosevelt	(1882–1945)
45	Alexander Hamilton	(1755–1804)
48	Woodrow Wilson	(1856–1924)

Sig	Person	Dates
110648	Mark Hopkins	(1802–1887)
64747	A. Freeman Palmer	(1855–1902)
32162	C. Saunders Cushman	(1816–1876)
27754	Lillian Wald	(1867–1940)
24265	James Buchanan Eads	(1820–1887)
23445	Rufus Choate	(1799–1859)
21450	John Lothrop Motley	(1814–1877)
21013	Sylvanus Thayer	(1785–1872)
16398	James Kent	(1763–1847)
15815	Emma Willard	(1787–1870)

The Best Man/Women Does not Always Win

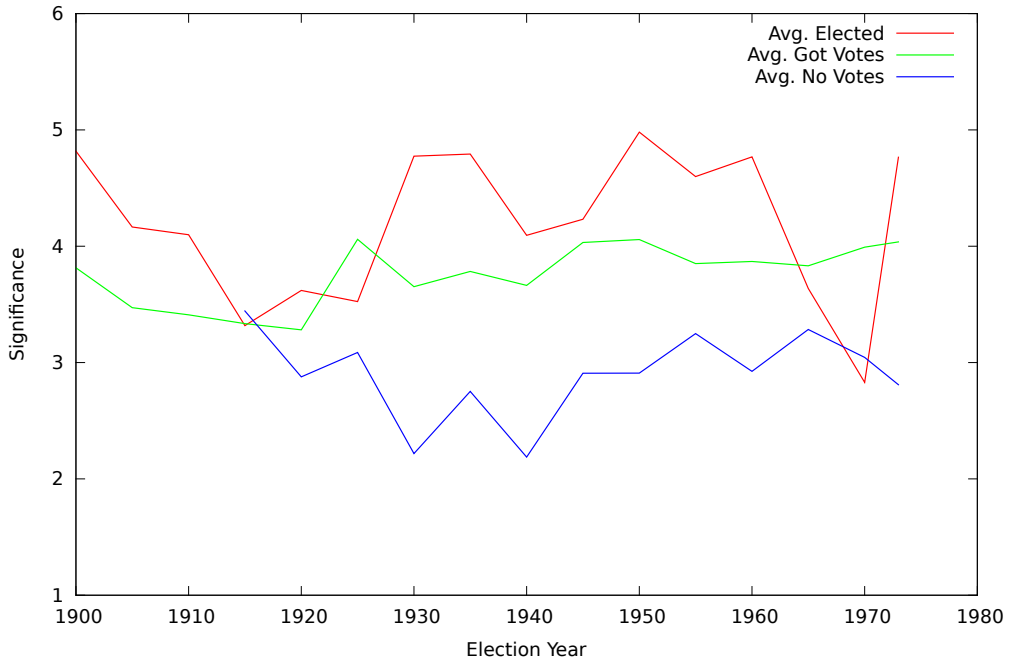
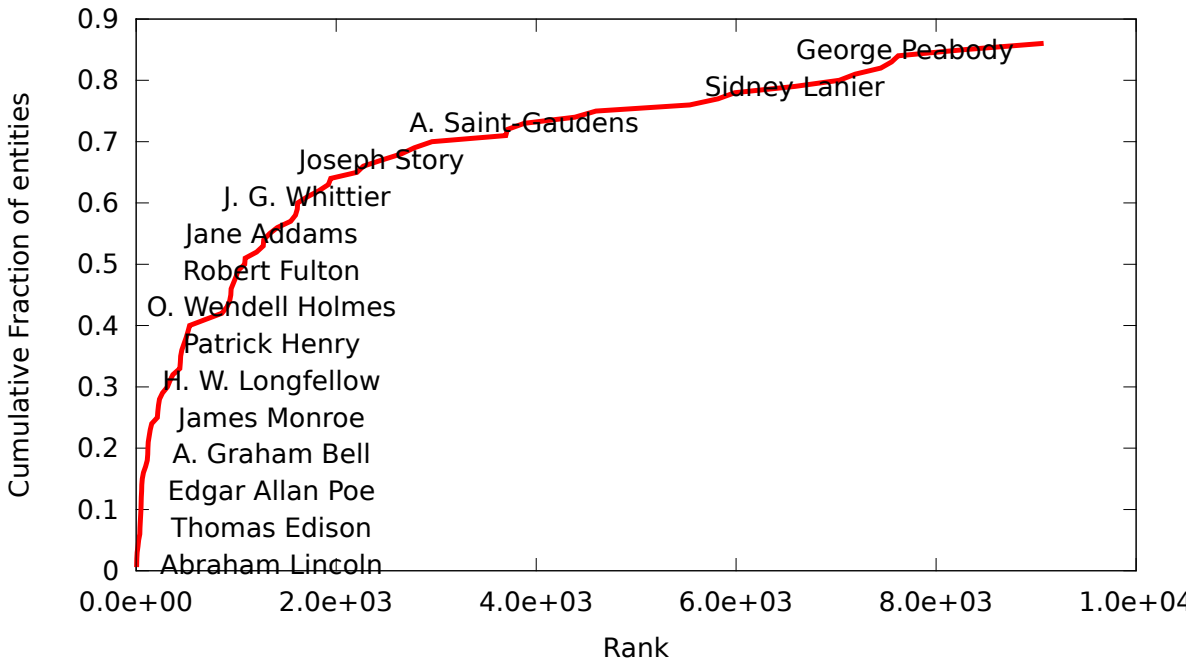


Figure: Significance of elected, losing, and unvoted-for Hall of Fame candidates in each election.

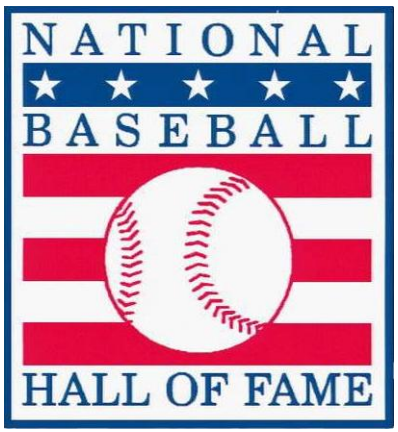
The Weakness of the Tail



A cumulative distribution plot of Hall of Fame members by significance rank shows that the top 65% are substantially stronger than the rest.

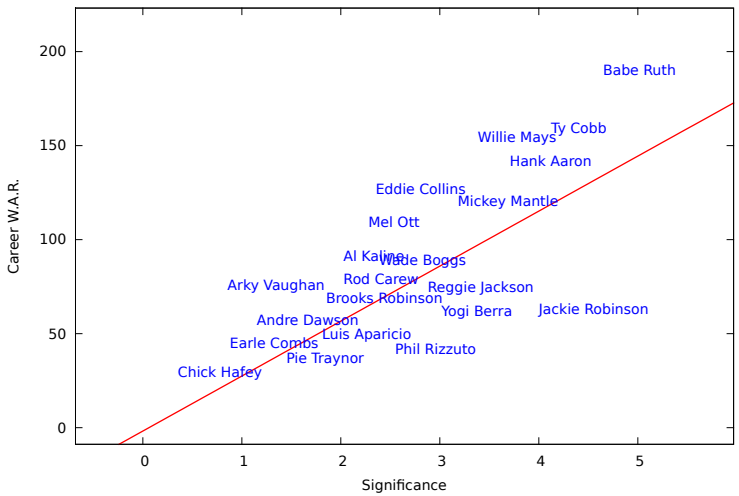
The Baseball Hall of Fame

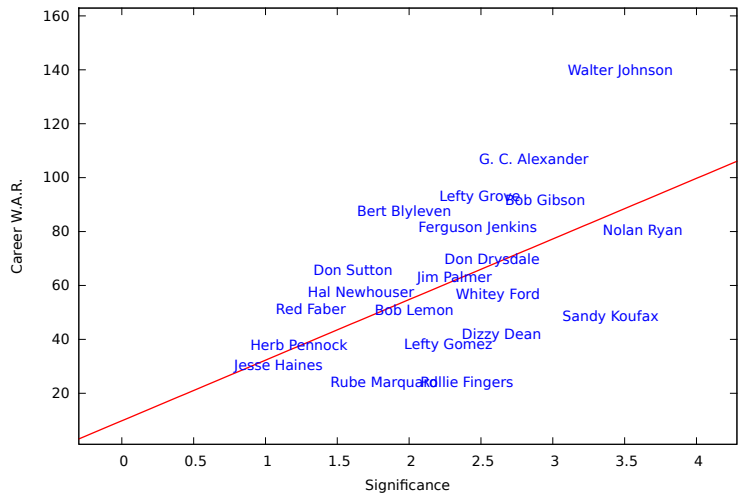
Since 1936, this popular Cooperstown NY institution has held annual elections to honor the greatest figures in baseball history.



Baseball players leave an meaningful statistical measure of accomplishment, enabling comparison between historical significance, statistics, and Hall of Fame voting records.

Performance (Wins Above Replacement) vs. Significance





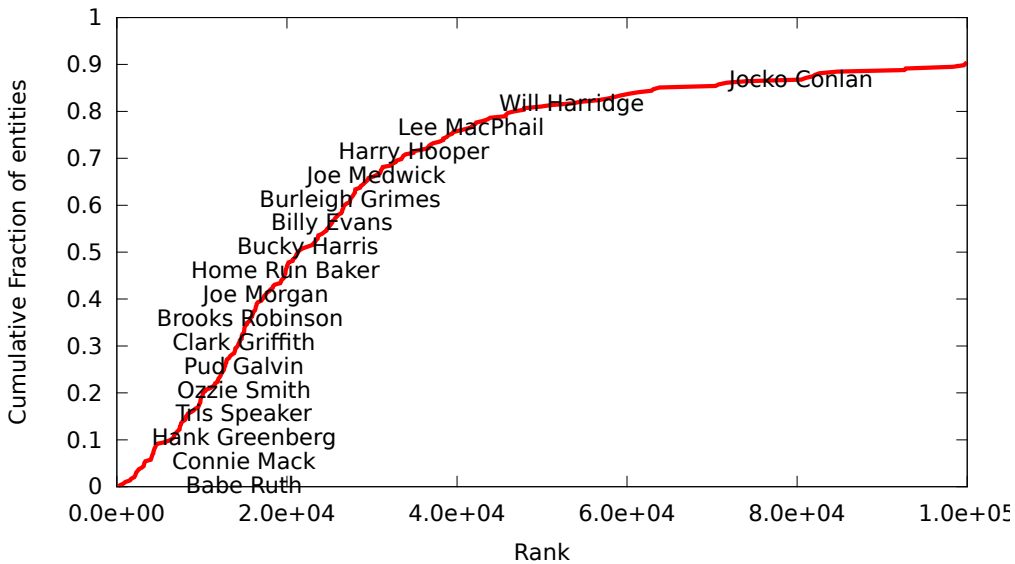
Historical significance is strongly corrolated with statistical performance (Wins Above Replacement) for position players (left) and pitchers (right).

The Most/Least Significant Members of the Baseball Hall of Fame

Sig	Person	Dates
397	Babe Ruth	(1895–1948)
800	Jackie Robinson	(1919–1972)
1010	Ty Cobb	(1886–1961)
1550	Hank Aaron	(1934–)
1688	Lou Gehrig	(1903–1941)
2044	Ted Williams	(1918–2002)
2145	Joe DiMaggio	(1914–1999)
2223	Cap Anson	(1852–1922)
2277	Cy Young	(1867–1955)
2449	Honus Wagner	(1874–1955)

Sig	Person	Dates
246381	Alex Pompez	(1890–1974)
171864	Ray Brown	(1908–1965)
169077	Bill McGowan	(1896–1954)
167887	Andy Cooper	(1898–1941)
166440	Louis Santop	(1890–1942)
165474	Turkey Stearnes	(1901–1979)
162565	Al Barlick	(1915–1995)
162506	Nestor Chylak	(1922–1982)
158675	J. L. Wilkinson	(1878–1964)
153094	Hilton Smith	(1907–1983)

The Weakness of the Tail



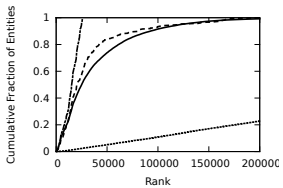
The top 70% of the Baseball Hall of Fame are much stronger choices than the rest.

Explaining Weak Selections

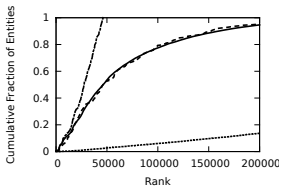
In these and other domains, expert panels consistantly failed to identify strong candidates for roughly 30% of its selections.

These results are consistant with a Gaussian measurement error, where significant numbers of relatively average individuals will be vastly overvalued and therefore selected

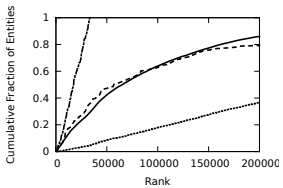
Baseball



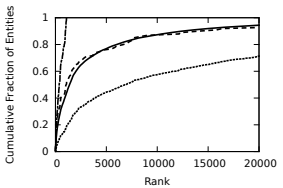
Football



Basketball



Great Americans



Similar evaluation errors presumably occur with job hiring and college admissions.

Look to your left and right – one of you probably doesn't belong here!

Are Women Underrepresented in Wikipedia?

There are far more Wikipedia articles about men than women.

Women make up only 15% of Wikipedia contributors and 8.5% of its editors, so there may be systematic bias against them.

But should there be more women in Wikipedia?

Have important women's achievements have been forgotten? Or alternately, perhaps more marginal women have been added to correct for perceived bias?

How can we tell?

Assessing Missing People

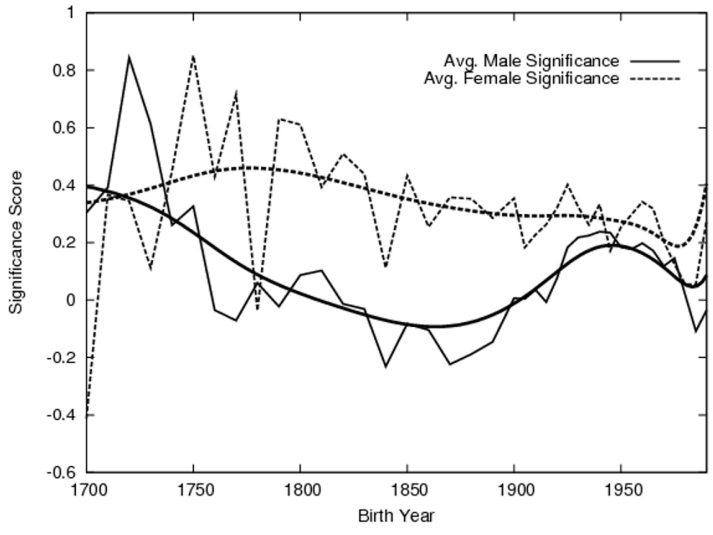
Famous people are outliers in the world population, at the very highest levels of significance.

Thus there must exist many candidates (for each gender) just below the Wikipedia standard for inclusion.

Thus if men in Wikipedia score as more significant than women, either men *should* be admitted or women excluded to maintain equal standards.

By partitioning people into cohorts based on birth year, we can study how the effect of gender varies over time.

Male/Female Significance by Birthyear



Women have long required far greater achievement levels (by over 0.25σ, analogous to 4 IQ points in the mean) than men to get equally noted for posterity.

Thus women *are* underrepresented in the historical record.

Philanthropy to Win

Some of you may achieve a financial position sufficient to make a large donation to a worthy cause.

Lorenzo de Medici (1449–1492) [541] is more renown today for his sponsorship of Renaissance artists than for ruling the Florentine Republic.

How can you invest this money to bring you the greatest historical glory?

Buildings crumble, so we recommend the sponsorship of living institutions..

Endow a Prize?


- The Nobel Prizes – *Alfred Nobel (1833–1896)* [611]
- The Fields Medal – *John Charles Fields (1863–1932)* [39159]
- The Pritzker Architecture Prize – *Jay Pritzker (1922–1999)* [72735]
- The MacArthur "Genius" Fellowship – *John D. MacArthur (1897–1978)* [56614] and *Catherine T. MacArthur (1908–1981)* [203172]

A meaningful prize can be endowed in the $10-20 million dollar range.

Take Over a University

- *Ezra Cornell (1807–1874)* [10996] – founder of Western Union, donated $400K in 1865.
- *John Harvard (1607–1638)* [5833] – bequethed 780 £ plus 320 books in 1638.
- *Johns Hopkins (1795–1873)* [1300] – bequeathed $7 million in 1876.
- *Elihu Yale (1649–1721)* [9471] – goods worth 800 £ in 1718.
- *Henry Rowan (1923–)* [182874] – Glassboro State College became Rowan University after a $100 million donation in 1997.

Perhaps $1 billion is sufficient to get a good state university (Stony Brook?) named after you.

I am willing to help broker such a donation if needed.

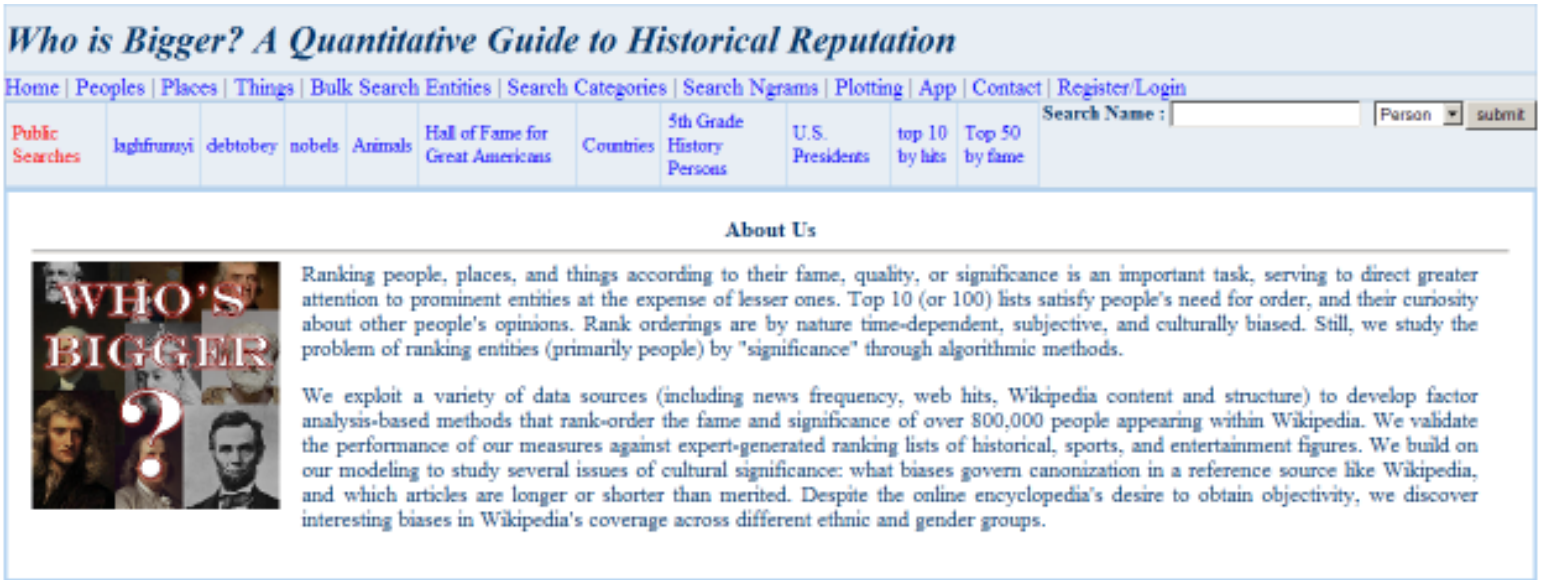
Outline

Future Work

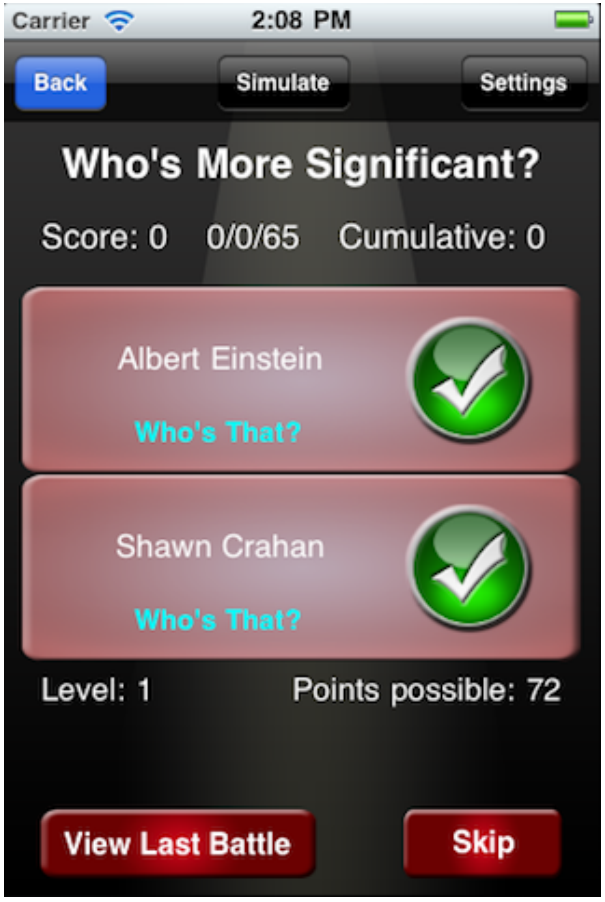
- Cross-cultural comparisons based on non-English Wikipedia analysis.
- Sociological studies of cumulative advantage and gender bias in the news/historical record.
- Ngram assembly and sentiment analysis.
- Web-scale entity ranking.
- *Who's Bigger*: The book

Who's Bigger: The Resource

Check out our analysis of any historical figure in Wikipedia at `http://www.whoisbigger.com`



Who's Bigger: The App

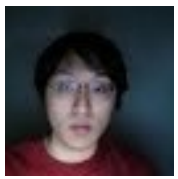
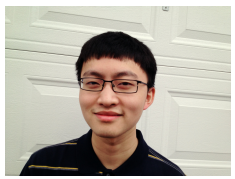


http://itunes.apple.com/us/app/whos-bigger/id437080657

Thanks





- Charles Ward (Wingman)
- Bala Mundiam, Goutham Bhat, Ajeesh Elikkottil (graduate students)
- Vincent Tsuei, Qi Chou (undergraduate students)
- Arnout van de Rijt, Eran Shor (Sociology)

This work was partially supported by NSF Grants DBI-1060572 and IIS-1017181.

Thanks!

Pairwise Feature Correlation

	FPR	PPR	PH	NR	AL	NF
FPR	1.00	0.68	0.67	0.63	0.27	0.30
PPR		1.00	0.50	0.49	0.24	0.29
PH			1.00	0.71	0.27	0.37
NR				1.00	0.34	0.36
AL					1.00	0.20
NF						1.00

Several other measures were also considered as inputs to our analysis, but were rejected.

For instance, the estimated number of page hits returned by the Google and Bing search engines were found to be largely erroneous.

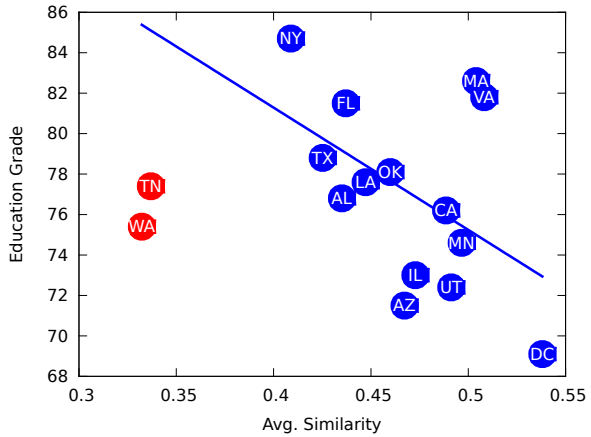
Other graph measures were subsumed by PageRank, and sentiment data from news data proved to be largely uninformative.

Higher Celebrity/Gravitas Professions

High Celebrity	F1-F2	High Gravitas	F2-F1
Wrestler	1.76	Architect	0.69
Fashion Model	1.55	Conductor (Music)	0.66
Serial Killer	0.73	Political Scientist	0.61
Comedian	0.66	Anthropologist	0.59
TV Anchorperson	0.63	Judge	0.47

State Teaching Standards

We carefully analyzed all figures appearing in 16 state teaching standards.



The genericness of state education standard appears to correlate negatively with education outcomes. Tennessee and Washington (red points) are notable outliers. differing greatly from other standards, but having average educational quality.
Fully 63 of the weakest textbook figures appeared in zero state standards.
Thus the standards are not to blame for the textbook.