



Data Science for Researchers and Scholars

Vasant G. Honavar

Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics,
Public Health Sciences and Neuroscience
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

vhonavar@psu.edu
<http://faculty.ist.psu.edu/vhonavar>
<http://ailab.ist.psu.edu>

Introductions

- Instructor
 - Dr. Vasant Honavar
 - Professor, Data Sciences, BG, Neuroscience, CSE, Public Health Sciences
 - Director, Artificial Intelligence Research Lab
 - Director, Center for Artificial Intelligence Foundations and Scientific Applications (CENSAI)
 - E335 Westgate Building
 - vhonavar@psu.edu
 - <http://faculty.ist.psu.edu/vhonavar>
- Teaching Assistant
 - Zhimeng Guo
 - PhD Student, Informatics
 - zgz5107@psu.edu
- Students?



What I do

- **Machine learning:** Statistical, information theoretic, linguistic and structural approaches to machine learning; learning predictive relationships from sequential, graph-structured, multi-relational, multimodal, partially specified, partially labeled, distributed data, linked data
- **Causal Inference:** Causal inference from disparate experimental and observational studies, causal inference from relational data, causal inference from temporal data
- **Knowledge Representation and Inference:** Logical, probabilistic, and decision-theoretic knowledge representation and inference; federated knowledge bases; selective information sharing; federated services; representing and reasoning about qualitative preferences
- **Applied Informatics**
 - **Bioinformatics:** Prediction of macromolecular (protein-protein, protein-RNA, and protein-DNA) interaction networks, interfaces, and complexes; immune networks; microbiomes etc.
 - **Health Informatics:** Predictive and causal modeling of health outcomes from patient (health records, genomics, socio-economic, environmental) data
 - **Brain Informatics:** Modeling and analysis of structure and dynamics of brain networks
 - **Materials Informatics:** Predicting material properties from structure and composition
- **Algorithmic Discovery**
 - Algorithmic abstractions of scientific domains
 - Representations of scientific artifacts (experiments, data, models, assumptions, hypotheses, theories ...)
 - Infrastructure for computationally mediated collaborative science

PennState
Institute for Computational
and Data Sciences

Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory

PennState
Clinical and Translational
Science Institute

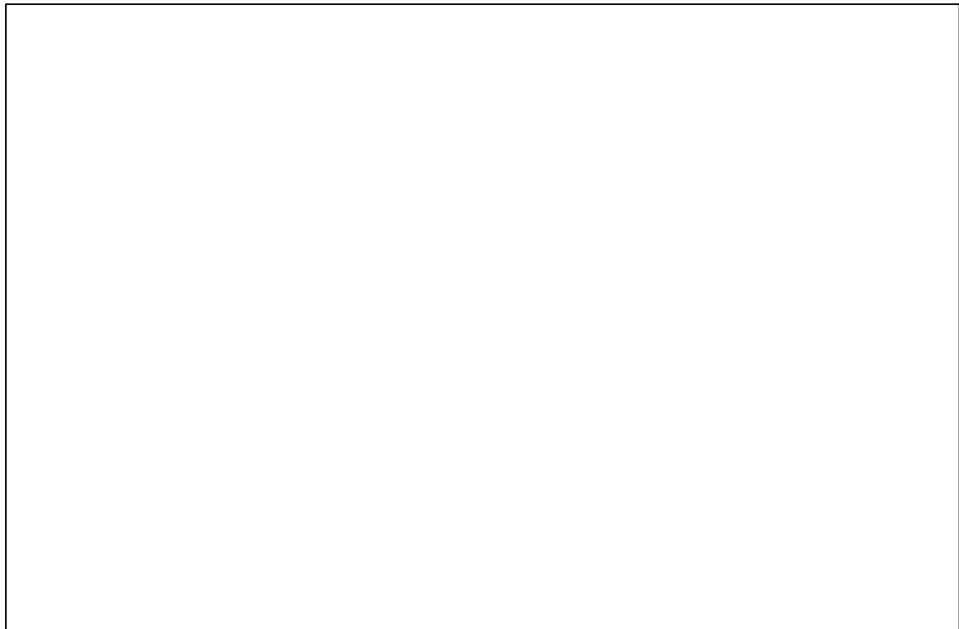
Computing, Artificial Intelligence, Informatics, and Data Sciences

- **Computation is the best formalism we have for describing how information is encoded, stored, communicated and used by natural as well as synthetic systems**
- **Computation plays in many sciences a role that is analogous to what calculus played in transforming physics from a descriptive science (pre Newton) into a predictive science (post Newton)**
 - Computation: Cognitive sciences / AI :: Calculus : Physics
 - **Computation: Life sciences :: Calculus : Physics**
 - Computation: Social sciences :: Calculus : Physics
- **Algorithms as theories:** We understand a phenomenon when we have an algorithm that models it at the desired level of detail
- **Computing offers an exploratory apparatus for science:** To the extent that science is about acquiring, organizing, integrating, analyzing, and reasoning with **information**, computing, science of information processing, provides exploratory apparatus for science

PennState
Center for Artificial Intelligence Foundations & Scientific Applications

Data Science for Researchers and Scholars

Vasant Honavar, Fall 2023



Data and Science



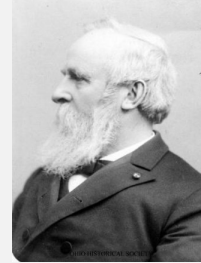
Data = plural of datum – resulting from measurement or observation



Science = systematic study of the structure and behavior of the physical, biological, cognitive, social, and engineered systems through observation and experimentation

Data, Philately, Physics

- All science is either stamp collecting or physics – Rutherford



Journey from Philately to Physics



- Brahe gathered 20 years of extremely accurate astronomical measurements: positions of the stars and planets: **data**



- Kepler, working for Brahe, fit the data in every way imaginable to discover laws of planetary motion: **big data analytics and machine learning**



- Newton and Leibnitz invented calculus – a language for expressing and reasoning about physical laws – transforming natural philosophy into physics

What is Data Science?



A discipline concerned with the collection and analyses of data to make sense of the world around us.



Data may be generated by **humans** (surveys, logs, etc.) or **machines** (weather data, road vision, etc.).



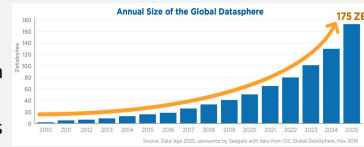
Data may be in different **formats** (text, audio, video, augmented or virtual reality, etc.).



Analyses can be descriptive, predictive, or causal

Fast forward 200 years

- Our instruments of observation have become highly sophisticated
- Data acquisition is increasingly automated
- Exponential increase in
 - Volume or quantity
 - Velocity or rate of acquisition
 - Variety
 - Multiple sources, owners
 - Multiple views, uses, users
 - Multiple data types, representations, semantics
 - Multiple levels of abstraction, granularity
 - Multiple contexts, scope
- Data sources include databases, literature, images, social media, sensors, scientific data, simulation results..



Two notions of data

- Scientific data
 - Observational or experimental
- Digital data
 - Input to, or output of, a computation
 - Scientific data
 - Meta data
 - Analyses results
 - Hypotheses
 - Models
 - Simulation results
- “Data” in “data science” often refers to both
- What we mean by data should be clear from context

Fast forward 200 years: Data, data everywhere!

Biology

Humanities

Behavioral and Social Sciences

Personal **Public** **Social**

Health Sciences

Policy Environment

- Healthcare
- Education
- Environment
- Transportation
- Energy
- Urban Planning
- Public Safety

Built and Natural Environment

- Healthcare
- Education
- Environment
- Transportation
- Energy
- Urban Planning
- Public Safety

Social Environment

- Healthcare
- Education
- Environment
- Transportation
- Energy
- Urban Planning
- Public Safety

Individual

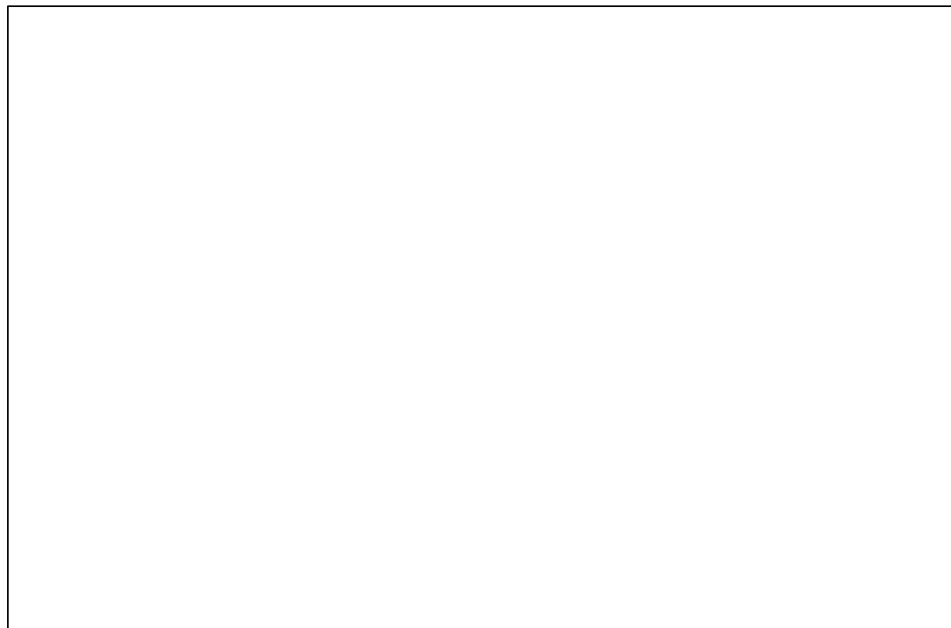
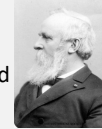
- Healthcare
- Education
- Environment
- Transportation
- Energy
- Urban Planning
- Public Safety

Source: Keith Marzullo



Data, Philately, Physics

- All science is either stamp collecting or physics – Rutherford
- Data revolution has dramatically accelerated the volume, velocity, variety of stamp collection!
- Computing revolution enables us to sort, label, organize, catalog and annotate the resulting stamp collection more efficiently than ever before!
- Connectivity revolution allows us to share, interconnect our disparate stamp collections!
- But without data science methods, all we are left with are exquisite stamp collections!
- The journey from philately to physics, from collecting data to understanding of the world around us is impossible without Data Science methods and tools



Transformative potential of data science

- Understanding physical, biological, cognitive, social and engineered systems – materials, cells, brains, individuals, organizations, societies
- Improving population health
- Anticipating and responding to crises
- Personalizing teaching and learning
- Defending critical infrastructure and services
- Making better decisions, e.g., public policy
- Making cities and communities smarter
- Improving food, energy, and water security
-

PennState Institute for Computational and Data Sciences Center for Artificial Intelligence Foundations & Scientific Applications Artificial Intelligence Research Laboratory PennState Clinical and Translational Science Institute

Transdisciplinary foundations of data science

Computer Science using Big Data Machine Learning Math & Statistics
Dangerous Software Data Science Traditional Research
Subject Matter Expertise

Source: Drew Conway

PennState Clinical and Translational Science Institute Data Science: for Researchers and Scholars Vasant Honavar, Fall 2023

Data Science is a close cousin of Informatics

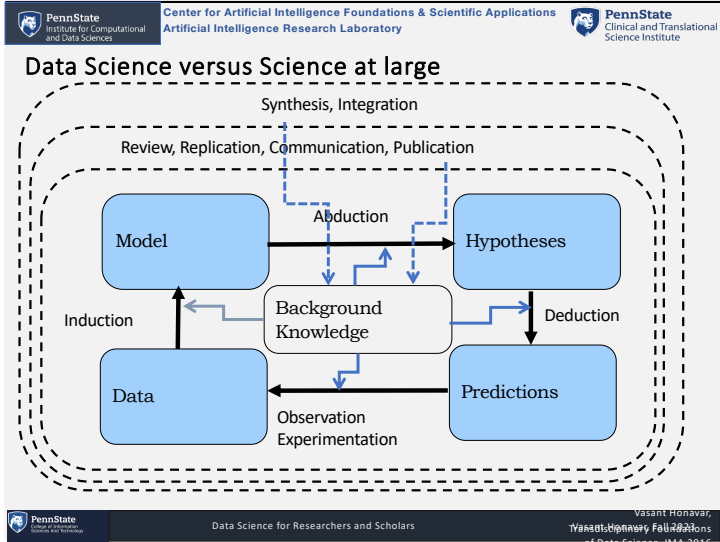


- Informatics is the study of the **structure, behavior, and interactions** of natural and engineered computational systems, including genomes, cells, brains, computers, organisms, societies.
- Data science shares the goal of understanding physical, biological, cognitive, social, and engineered systems
- Informatics is concerned with the **representation, processing, and communication of information** in such systems.
- Data science is concerned with formulating descriptive, predictive, and causal questions about the world around us and answering them using data







Data science

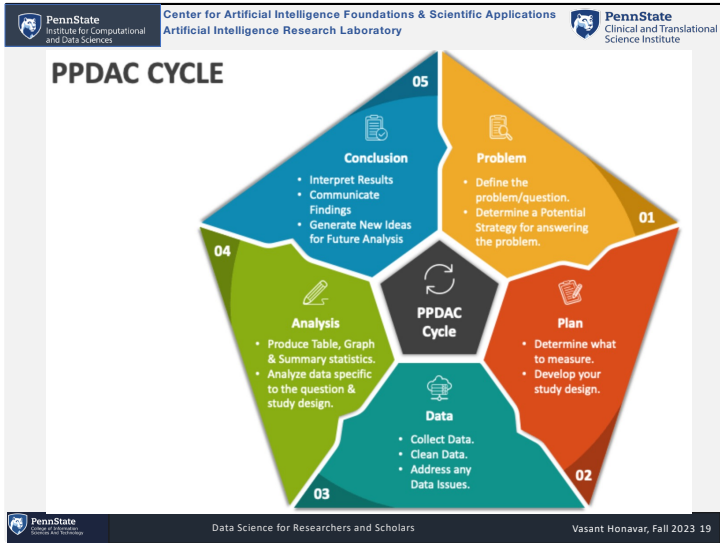
- Data science uses computers and algorithms for data analyses, yet Data Science \neq Computer science
 - Data scientists are driven by questions, data, and assumptions
 - Computer scientists are driven by problems and the algorithms that solve them
 - Software developers are hired to produce code
 - Data scientists are hired to produce insights
- Data science uses mathematics, yet Data Science \neq Mathematics
- Data science uses statistical reasoning, yet Data Science \neq Statistics!
- **Science probes; it does not prove** – Gregory Bateson
- **Corollary: Data science probes; it does not prove**





Ingredients of data science practice

-  Scientific mindset
-  Mathematical and statistical reasoning
-  Data literacy
-  Algorithmic Thinking
-  Scientific communication
-  Data Ethics



Data

Origins of data can differ

- Data from a designed study
 - Observational study
 - Experimental study
- Data in the wild
 - Documents on the internet
 - Electronic health records of patients
 - Interactions on Facebook
- The story behind the data matters!
- Data often has bias and errors
- Formulating questions, answering them, and interpreting the answers requires creativity, skill, and rigor
- Doing data science is rarely is simply a matter of blindly applying a tool

Scientific mindset and curiosity

Doing good data science requires


- Curiosity about the domain and the story behind the data
- Asking important questions
 - Whose answers matter for science or society
- Pragmatic approach to assembling or acquiring the data needed to answer the questions
- Tolerance of uncertainty and ignorance
- Ethics

Warm up – asking questions

Who, What, Where, When, and Why on the following datasets:

- Baseball-reference.com
- International Movie Database (IMDb)
- Google ngrams
- NYC taxi cab records

Baseball-Reference.com



[Home](#) | [About](#) | [Contact](#) | [Privacy](#) | [Terms](#) | [Advertise](#) | [Help](#) | [Feedback](#) | [Site Map](#) | [Mobile](#) | [RSS](#) | [Print](#) | [Search](#)


[Home](#) | [About](#) | [Contact](#) | [Privacy](#) | [Terms](#) | [Advertise](#) | [Help](#) | [Feedback](#) | [Site Map](#) | [Mobile](#) | [RSS](#) | [Print](#) | [Search](#)

Transactions

July 9, 1916: Purchased with [Eddie Stange](#) and [Red Faber](#) by the [Boston Red Sox](#) from [Baltimore \(Oriental\)](#) for more than \$25000, more than \$25000.
December 26, 1916: Purchased by the [Boston Red Sox](#) from the [Boston Red Sox](#) for \$10,000.
February 26, 1920: Released by the [New York Yankees](#).
February 26, 1920: Signed as a Free Agent with the [Boston Braves](#).

The biographic information used here was obtained from [Baseball Reference](#) and is copyrighted by [Baseball Reference](#). All rights reserved.

Babe Ruth


George Herman Ruth (Nick: The Bambino or The Sultan Of Swat)
 Positions: Outfielder and Pitcher
 Bats: Left, Throws: Left
 Height: 6' 7", Weight: 215 lb.

Birth: February 6, 1895 in Baltimore, MD
High School: St. Mary's HS (Baltimore, MD) (Graduated)
Debut: August 29, 1915 in New York, NY (Age 20.192)
Final Game: October 6, 1935 in Baltimore, MD
Inducted: into the Hall of Fame in 1936 (1936 ballgame induction ceremony in Cooperstown)
The Babe Ruth Page at the Baseball Hall of Fame (photos, photos, videos)
Baseball Reference contains more than 100,000 baseball players and teams.

[View Player Page](#) | [View Player Page](#) | [View Player Page](#) | [View Player Page](#) | [View Player Page](#)

Salaries

Salaries may not be complete (especially pre 1955) and may not include some earned bonuses.

Year	Age	Team	Salary	Sign/Trade/Contract	Source	Notes/Other Source
1916	21	Baltimore Red Sox	\$2,000	F	1st year contract signed	Contract also came with 100 shares
1917	22	Baltimore Red Sox	\$3,500	F	1st year contract signed	
1918	23	Baltimore Red Sox	\$5,000	F	Contract of 1 year	
1919	24	Baltimore Red Sox	\$8,000	F	Contract of 1 year	ESPN: 100,000 shares of stock
1920	25	New York Yankees	\$13,000*	F	Player request transfer of AIP contracts; signed an AIP contract	Player stock, 100,000 shares of stock
1921	26	New York Yankees	\$15,000*	F	Player request transfer of AIP contracts; signed an AIP contract	
1922	27	New York Yankees	\$20,000*	F	Player request transfer of AIP contracts; signed an AIP contract	
1923	28	New York Yankees	\$25,000*	F	Player request transfer of AIP contracts; signed an AIP contract	
1924	29	New York Yankees	\$30,000*	F	Player request transfer of AIP contracts; signed an AIP contract	
1925	30	New York Yankees	\$35,000*	F	Player request transfer of AIP contracts; signed an AIP contract	
1926	31	New York Yankees	\$40,000*	F	Player request transfer of AIP contracts; signed an AIP contract	
1927	32	New York Yankees	\$45,000*	F	Player request transfer of AIP contracts; signed an AIP contract	
1928	33	New York Yankees	\$50,000*	F	Player request transfer of AIP contracts; signed an AIP contract	
1929	34	New York Yankees	\$55,000*	F	Player request transfer of AIP contracts; signed an AIP contract	
1930	35	New York Yankees	\$60,000*	F	Player request transfer of AIP contracts; signed an AIP contract	
1931	36	New York Yankees	\$65,000*	F	Player request transfer of AIP contracts; signed an AIP contract	
1932	37	New York Yankees	\$70,000*	F	Player request transfer of AIP contracts; signed an AIP contract	
1933	38	New York Yankees	\$75,000*	F	Player request transfer of AIP contracts; signed an AIP contract	
1934	39	New York Yankees	\$80,000*	F	Player request transfer of AIP contracts; signed an AIP contract	
1935	40	New York Yankees	\$85,000*	F	Player request transfer of AIP contracts; signed an AIP contract	
1936	41	New York Yankees	\$90,000*	F	Player request transfer of AIP contracts; signed an AIP contract	
1937	42	New York Yankees	\$95,000*	F	Player request transfer of AIP contracts	

*Denote to date (not by historical) \$1,000,000

Baseball statistics

Yearly summary
statistics of batting,
pitching, and
fielding record, with
teams and awards.

Penn State Nittany Lion		1924-1926			1927-1930			1931-1934			1935-1938			1939-1942									
Standard Batting		More Stats			Summary - Show More Stats			Splits			Career			1924-1926									
Player	Age	Pos	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	OBP	SLG	OPS	WAR	WAR*	WAR ⁺	WAR	Pace	Awards
1924	19	SS	118	101	11	21	5	0	0	0	0	0	4	68	.286	.389	.675	0.9	-	-	-	-	-
1925	20	SS	42	35	10	20	4	1	0	0	1	1	23	21	.324	.514	.838	0.9	-	-	-	-	-
1926	21	SS	47	42	10	21	5	2	1	1	1	1	22	22	.333	.524	.857	1.0	-	-	-	-	-
1927	22	SS	52	44	16	30	7	2	1	1	1	1	30	20	.352	.584	.936	1.2	-	-	-	-	-
1928	23	SS	70	60	17	30	8	1	1	1	1	1	40	24	.350	.583	.933	1.3	-	-	-	-	-
1929	24	SS	120	102	22	30	13	2	1	1	1	1	50	30	.333	.583	.916	1.4	-	-	-	-	-
1930	25	SS	152	134	28	35	8	1	1	1	1	1	60	40	.324	.583	.907	1.5	-	-	-	-	-
1931	26	SS	152	134	28	35	8	1	1	1	1	1	60	40	.324	.583	.907	1.5	-	-	-	-	-
1932	27	SS	152	134	28	35	8	1	1	1	1	1	60	40	.324	.583	.907	1.5	-	-	-	-	-
1933	28	SS	152	134	28	35	8	1	1	1	1	1	60	40	.324	.583	.907	1.5	-	-	-	-	-
1934	29	SS	152	134	28	35	8	1	1	1	1	1	60	40	.324	.583	.907	1.5	-	-	-	-	-
1935	30	SS	152	134	28	35	8	1	1	1	1	1	60	40	.324	.583	.907	1.5	-	-	-	-	-
1936	31	SS	152	134	28	35	8	1	1	1	1	1	60	40	.324	.583	.907	1.5	-	-	-	-	-
1937	32	SS	152	134	28	35	8	1	1	1	1	1	60	40	.324	.583	.907	1.5	-	-	-	-	-
1938	33	SS	152	134	28	35	8	1	1	1	1	1	60	40	.324	.583	.907	1.5	-	-	-	-	-
1939	34	SS	152	134	28	35	8	1	1	1	1	1	60	40	.324	.583	.907	1.5	-	-	-	-	-
1940	35	SS	152	134	28	35	8	1	1	1	1	1	60	40	.324	.583	.907	1.5	-	-	-	-	-
1941	36	SS	152	134	28	35	8	1	1	1	1	1	60	40	.324	.583	.907	1.5	-	-	-	-	-
1942	37	SS	152	134	28	35	8	1	1	1	1	1	60	40	.324	.583	.907	1.5	-	-	-	-	-

Baseball Questions

- How to best measure individual player's skill, value or performance?
- How fair do trades between teams work out?
- What do the performance trajectories of players look like as they mature and age?
- To what extent does batting performance correlate with the position played?

Demographic Questions

- Do left-handed players have shorter careers than right-handers?
- Do player salaries reflect past, present, or future performance?
- Are heights and weights of players increasing over the years?

IMDb: Movie Data

IMDb Find Movies, TV shows, Celebrities and more... AI

Movies, TV & Showtimes Celebs, Events & Photos News & Community Watchlist



It's a Wonderful Life (1946) Top 5000

Approved 138 min - Drama / Family / Fantasy - 7 January 1947 (USA)

Your ratings: ★★★★★ /10
Rating: 8.7/10 from 302,343 users
Reviews: 632 user | 187 critic

An angel helps a compassionate but despairingly frustrated businessman by showing what life would have been like if he never existed.

Director: Frank Capra
Writers: Frances Goodrich (screenplay), Albert Heckelt (screenplay), 4 more credits >
Stars: James Stewart, Donna Reed, Lionel Barrymore | See full cast and crew >

Home at DVD/Blu-ray >

+ Watchlist Watch Trailer Share...

Details Edit

Country: USA
Language: English
Release Date: 7 January 1947 (USA) See more >
Also Known As: The Greatest Gift See more >
Filming Locations: California, USA See more >

Box Office

Budget: \$3,180,000 (estimated)
Opening Weekend: \$19,845 (UK) (19 December 2008)
Gross: \$682,222 (UK) (24 December 2018)
See more >


Company Credits

Production Co: Liberty Films (E) See more >
Show detailed company contact information on IMDbPro >

Technical Specs

Runtime: 130 min | 118 min (DVD edition)
Sound Mix: Mono (DCA Sound System)
Color: Color (colorized) Black and White
Aspect Ratio: 1.37 : 1
See full technical specs >

IMDb: Actor Data




James Stewart (1) (1906–1997)
Actor | Soundtrack | Director

James Maitland Stewart was born on 22 May 1906 in Indiana, Pennsylvania, where his father owned a hardware store. He was educated at a local prep school, Marsburg Academy, where he was a keen athlete (football and track), musician (singing and accordion playing), and sometime actor. In 1929 he won a place at Princeton, where he studied ... [View full bio](#) >

Born: James Maitland Stewart
May 20, 1908 in Indiana, Pennsylvania, USA

Died: July 2, 1997 (age 89) in Los Angeles, California, USA













250 photos | 42 videos | 1180 news articles >

Won 1 Oscar. Another 25 wins & 19 nominations. [See more awards](#) >

Cast Edit

Cast overview, first billed only:

	James Stewart	...	George Bailey
	Donna Reed	...	Mary Hatch
	Lionel Barrymore	...	Mr. Potter
	Thomas Mitchell	...	Uncle Billy
	Henry Travers	...	Clarence
	Beulah Bondi	...	Mrs. Bailey
	Frank Faylen	...	Ernie
	Ward Bond	...	Bert
	Gloria Grahame	...	Violet
	H.B. Warner	...	Mr. Gower

Movie Questions

- Can we predict how well people will like a movie? What about its earnings?
- What does the social network of actors look like? (Six degrees of Kevin Bacon)
- What is the age distribution of actors and actresses in films?
- Do stars live longer or shorter lives than bit players or public?

Google N-grams

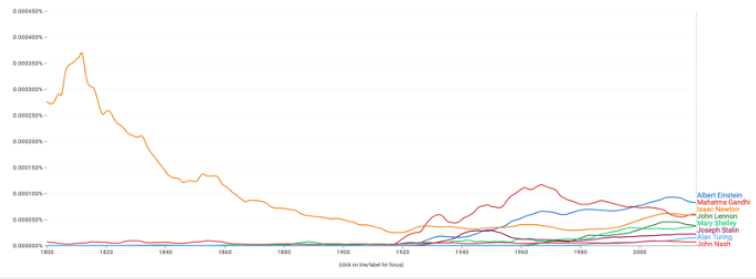
- Presents an yearly time series of the frequency of every “popular” word/phrase with 1 to 5 words occurs in scanned books.
- “Popular” word/phrase is one which appears > 40 times total.
- Google has scanned about 15% of all books ever published
- Caveat: The books scanned by google may or may not be a representative sample of the books ever published

Google N-gram Viewer

Google Books Ngram Viewer


Albert Einstein, Mahatma Gandhi, John Lennon, Isaac Newton, Mary Shelley, Joseph

1900 - 2019 English (2019) Case-insensitive Smoothing




N-gram Questions

- How has the use of slang words changed over time?
- What is the lifespan of fame and technologies? Is it increasing/decreasing?
- How often do new words emerge? How long do they stay in common usage?
- What words are associated with other words?



PennState
Institute for Computational
and Data Sciences

Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory



PennState
Clinical and Translational
Science Institute

US Congressional voting records

CONGRESS.GOV

Advanced Searches Browse

Search Tools Support Help

Current Congress
 117th Congress (2019-2020)

Topics: All Issues Health Care

WORK OPTIONS

Help Center

- Getting Started
- Tracking Legislative Activity
- About Alerts
- Floor Calendars
- Appropriations and Budget
- Votes in the House and Senate**
- Creating and Using Congress.gov Email Alerts
- Congress.gov Collections
- Searching
- Other Resources
- Contact

Votes in the House and Senate

Congress.gov Vote Information
Actions taken to record floor votes from legislation, nominations, and treaties
EXAMPLES - House bill (look for "Roll no."); Senate bill (look for "Report Vote Number")

Congressional Record - All recorded floor votes are published in the Congressional Record
EXAMPLES - House committee vote (look for "VOTES OF THE COMMITTEE"); Senate committee vote (look for "VOTES OF THE COMMITTEE")

Roll Call Votes for the U.S. Congress - Internet report provides links to House and Senate floor votes from the 17th Congress to the present.

House Votes
Click on the table website provides information about Roll Call Votes, Conference Calendar Sessions and Calendar Matters.

Senate Votes
Roll Call Votes, Roll Call Votes

Details for this Roll Call: [100-100](#) or [100-100](#)

[Calendar Notifications](#)

[CRS](#)

[Senate Court Notifications](#)

[The Votes](#)


[Vote Analysis](#)

Related

[Vote Resources](#)

[Senate Resources](#)


CONGRESS.GOV




PennState
Institute for Computational
and Data Sciences

Data Science for Researchers and Scholars


Vasant Honavar, Fall 2023

 PennState
 Institute for Computational and Data Sciences

Center for Artificial Intelligence Foundations & Scientific Applications
 Artificial Intelligence Research Laboratory

 PennState
 Clinical and Translational Science Institute

US congressional voting records

 CLERK
UNITED STATES HOUSE OF REPRESENTATIVES
FIND YOUR REPRESENTATIVE


[LEGISLATIVE INFORMATION](#)
[WORKERS INFORMATION](#)
[COMMITTEE INFORMATION](#)
[DISCLOSURES](#)
[ABOUT THE CLERK](#)

ALL VOTES

Name: Party: State: Votes:

Search by name: All Parties: All States: All States:

REPRESENTATIVE	PARTY	STATE	VOTE
Adams	Democratic	North Carolina	YEA
Adenfelt	Republican	Alabama	NAY
Agular	Democratic	California	YEA
Allen	Republican	Georgia	NAY
Allred	Democratic	Texas	YEA
Almouzni	Republican	Nevada	NAY
Armstrong	Republican	North Dakota	NAY
Auerhahn	Republican	Texas	NAY
Auzan	Democratic	Massachusetts	YEA
Aune	Democratic	Iowa	YEA
Babin	Republican	Texas	NAY

 PennState
Office of the Clerk, U.S. House of Representatives

Data Science for Researchers and Scholars

Vasant Honavar, Fall 2023

US Congressional voting questions

- Is political party affiliation predictive of votes on specific measures?
- Has the extent of bipartisan support for bills increased/decreased over the years?
- How does the voting record of senate members correlate with the demographics of their constituencies?

NYC Taxi Cab Data

- Gives driver/owner, pickup/drop-off location, and fare data for every taxi trip taken.
- Data obtained from NYC via Freedom of Information Act Request (FOA)

trip_id	pickup_datetime	dropoff_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	store_fare	tip	tolls	total_fare	pickup_location	dropoff_location
1	2013-01-01 00:00:00	2013-01-01 00:05:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
2	2013-01-01 00:05:00	2013-01-01 00:10:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
3	2013-01-01 00:10:00	2013-01-01 00:15:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
4	2013-01-01 00:15:00	2013-01-01 00:20:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
5	2013-01-01 00:20:00	2013-01-01 00:25:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
6	2013-01-01 00:25:00	2013-01-01 00:30:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
7	2013-01-01 00:30:00	2013-01-01 00:35:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
8	2013-01-01 00:35:00	2013-01-01 00:40:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
9	2013-01-01 00:40:00	2013-01-01 00:45:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
10	2013-01-01 00:45:00	2013-01-01 00:50:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
11	2013-01-01 00:50:00	2013-01-01 00:55:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
12	2013-01-01 00:55:00	2013-01-01 01:00:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
13	2013-01-01 01:00:00	2013-01-01 01:05:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
14	2013-01-01 01:05:00	2013-01-01 01:10:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
15	2013-01-01 01:10:00	2013-01-01 01:15:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
16	2013-01-01 01:15:00	2013-01-01 01:20:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
17	2013-01-01 01:20:00	2013-01-01 01:25:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
18	2013-01-01 01:25:00	2013-01-01 01:30:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
19	2013-01-01 01:30:00	2013-01-01 01:35:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square
20	2013-01-01 01:35:00	2013-01-01 01:40:00	-74.00597	40.71285	-74.00597	40.71285	10.00	0.00	0.00	10.00	Times Square	Times Square

Taxicab Questions

- How much do drivers make each night?
- How far do they travel?
- How much slower is traffic during rush hour?
- Where are people traveling to/from at different times of the day?
- Do faster drivers get tipped better?
- Where should drivers go to pick up their next fare?

Data Science: Descriptive, Predictive, Causal

- **Descriptive data science**
 - What is the gender distribution of STEM PhD recipients?
 - How is gender correlated with income among physicians?
- **Predictive data science**
 - Can we predict degree completion based on student demographics, high school GPA, academic major?
 - Can we predict voting preferences in presidential elections based on race, education, income, religious affiliation, marital status, and state of residence? If so, which variables are the most predictive of voting preferences?
- **Causal data science**
 - Does smoking cause cancer?
 - How do diet and exercise modulate the risk of heart disease among those with a family history of heart disease?

About the course

- Course rationale
- What can you expect to learn in the course?
- Course mechanics

Course rationale

- Progress in many fields, including sciences and humanities, is increasingly enabled by our ability to acquire, share, integrate and analyze disparate types of data.
- Advances in data science methods and tools, coupled with large data sets, are leading to breakthroughs in many sciences.
- Consequently, there is a need for researchers, scholars, and practitioners, regardless of their disciplinary background and interests, to become proficient in applying modern data science methods and tools to gain useful insights from data.

Who is this course for?

- The course is designed for graduate students from a wide range of disciplinary backgrounds and interests in informatics, physical, biological, cognitive, social, and health sciences , public policy and humanities.
- The course is **not** intended for students with strong prior exposure to computational and data sciences (including computer science, statistics) or engineering disciplines.

What are the course prerequisites?

- Graduate standing at Penn State
- Familiarity with, or at least willingness to learn the relevant topics in
 - Probability and statistics
 - Mathematics
 - Programming in Python
- Scientific mindset and curiosity
- Ability to communicate effectively

Course objectives

- Introduce data science methods and tools needed to formulate and answer research questions in sciences and humanities using large and complex data sets.
 - Descriptive, predictive, and causal analyses to answer research questions from data.
 - Laboratory assignments that offer hands-on experience with the application and evaluation of common data science methods.
 - A term project focused around assembling and using data to answer research questions of interest to the student

Learning outcomes

Upon completion of this course, students should be able to:

- Demonstrate broad understanding of the principles and practice of data sciences
- Assess the feasibility of answering chosen research questions using available data and methods
- Choose the right method(s) in a given setting
- Validate analyses
- Ensure reproducibility of analyses
- Responsibly handle sensitive data
- Assess data and algorithmic bias
- Critically evaluate research and scholarly studies
- Effectively communicate the results of analyses to technical and non-technical audiences

Tentative outline of topics

- What is data science?
- Mathematical preliminaries
- Python for Data Science
- Data Representation
- Descriptive statistics
- Exploratory Analysis: Clustering
- Data Visualization
- Predictive Modeling
- Linear Algebra for Data Analysis
- Probabilistic (generative) models)
- Evaluating predictive models
- Decision Trees
- Logistic Regression
- Multi-Class Extensions
- Kernel Machines and Kernel Trick
- Kernel Machines: Kernel Design
- Ensemble Methods: Random Forests
- Deep Learning
- Representation Learning
- Responsible Data Science
- Analysis of Sensitive Data
- Important ideas in machine learning
- Causal Inference from data and assumptions
- Causal Inference from causal models
- Causal Inference using Machine Learning
- Review

Course materials

- Recommended Books

- Shah, Chirag (2020). A Hands-On Introduction to Data Science, Cambridge University Press
- Skiena, S. (2017). Data Science Design Manual, Springer. Available for download by Penn State Students.
- Behrman, K. (2022). Foundational Python for Data Science.
- Watt, J., Borhani, R., Katsagellos, A. (2020). Machine Learning Refined. Cambridge University Press. Available through Penn State Libraries online.
- Deisenroth, M.P., Faisal, A., and Ong, C.S. (2018) Math for Machine Learning. Cambridge University Press. Available through Penn State Libraries online.
- Vanderplas, J. (2017). Python Data Science Handbook. O'Reilly. Freely available for online reading.
- Lecture slides and readings to be posted on the study guide on the course web page
 - <https://faculty.ist.psu.edu/vhonavar/Courses/dsmethods/homepage.html>
- Labs will be shared using google colab
- Canvas will be used for everything else – announcements, course-related emails, communicating with the instructor and the TA

Assignments and projects

- Weekly reading assignments
- Approximately biweekly ungraded problem sets
- Approximately biweekly graded laboratory assignments
- 1-3 graded projects
 - Formulate questions
 - Identify and assemble the relevant data
 - Apply data science methods to answer questions
 - Write up your results in the form of a paper

Labs

- We will use google colab: <https://colab.research.google.com>
- To access google colab:
 - Sign into your google account using your Penn State email
 - Go to <https://colab.research.google.com>
 - If you have multiple google accounts, please make sure that you switch to the account associated with your Penn State email address
 - We will share python notebooks on google colab with you using your Penn State email address

What to expect

- Lectures cover concepts, relevant math, methods
- Assigned readings reinforce the material covered in the class
- Lab assignments will provide hands-on experience with data science methods
- Projects give you experience with formulating and answering research questions using data
- Rule of thumb: For each hour of lecture, expect to spend three hours outside class to master the material

Grading

- Lab Assignments: 30%
 - Projects: 60%
 - Class participation: 10%
- 93% - 100% A
 - 90% - 93% A-
 - 87% - 90% B+
 - 83% - 87% B
 - 80% - 83% B-
 - 77% - 80% C+
 - 70% - 77% C
 - 60% - 70% D
 - 0% - 60% F

Please consult course policies regarding late assignments, and projects

Other policies

- Academic misconduct
- Responsible data science
- Disability accommodation
- Educational equity and non-discrimination
- Pandemic guidelines
- Emergency notifications