

CSCI 1070: Taming Big Data

Prof. Abby Stylianou

Logistics

- Are you on the waitlist? Talk to me after class!
- Class meetings: Tu/Th, 11AM–12:15PM, Ritter 115
 - Lecture, discussions, in class coding
 - Occasional quizzes/journaling about readings
 - Attendance not mandatory, but you're gonna have a bad time if you don't show up.
- Office Hours: Wed, 1:30-2:30PM, Ritter 107
- Class communications: Piazza! (*not email*)
- Website/schedule/materials: <https://cs.slu.edu/~stylianou/1070/>
- Technologies: Python, Jupyter Notebooks, Git

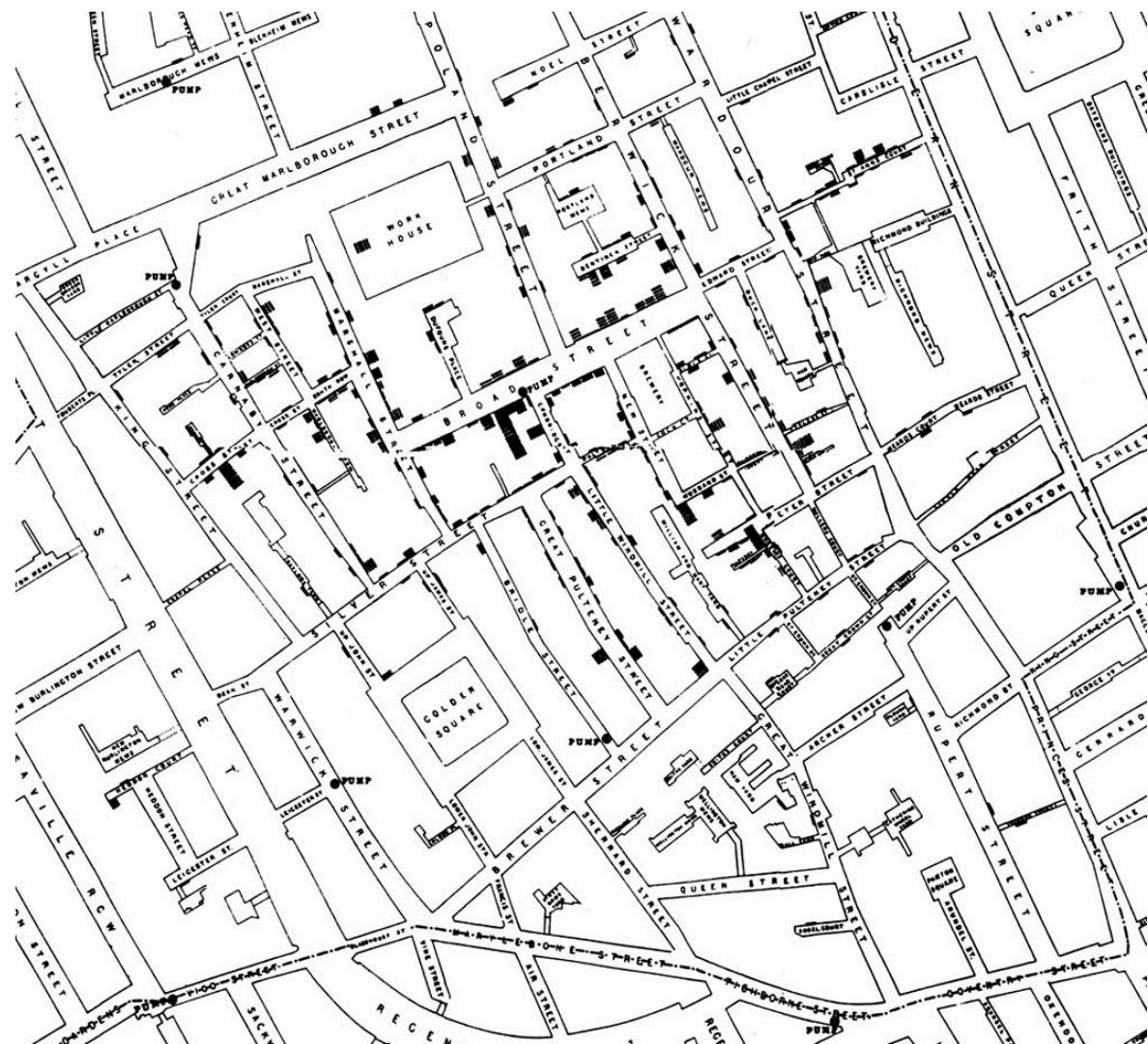
What is big data?

- 1663: John Graunt is the first person credited w/ statistical data analysis in his studies of the bubonic plague in Europe, dealing with what he referred to as “overwhelming amounts of information”

The Years of our Lord	1647	1648	1649	1650	1651	1652	1653	1654	1655	1656	1657	1658	1659	1660	1661	1662	1663	1664	1665	1666	1667	1668	1669	In 20 Years								
Alive and Still-born	335	329	327	351	389	381	384	433	483	419	463	457	421	544	499	439	410	445	503	475	52	1793	2005	1642	1582	1812	1247	8559				
Ail and Fever	916	835	889	780	824	824	822	974	743	875	869	1176	1000	1095	579	712	661	671	704	623	724	714	2475	2814	3310	3425	3682	4010	23784			
Alack and Suddenly	68	74	64	74	106	111	118	86	92	102	113	138	91	67	29	30	17	24	35	26	75	85	280	424	445	177	105	114	150			
Alack and Suddenly	4	1	1	3	7	6	6	4	5	5	3	8	13	8	10	13	6	4	4	54	14	5	12	14	16	90						
Alack and Suddenly	3	2	3	1	3	4	3	2	7	3	2	5	2	4	2	5	4	4	2	4	2	16	7	11	12	19	17	65				
Alack and Suddenly	155	176	802	289	833	762	200	386	168	363	362	235	346	251	449	428	352	348	278	512	346	330	387	1400	1422	2181	116	1597	7818			
Alack and Suddenly	3	6	10	5	11	8	5	7	10	5	7	4	6	6	3	10	7	3	3	12	3	25	19	24	31	26	19	123				
Alack and Suddenly	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	2	4	3	13					
Alack and Suddenly	26	29	31	19	31	53	36	37	73	31	24	35	63	52	20	14	23	25	27	30	24	30	85	112	105	157	150	114	609			
Alack and Suddenly	16	28	51	42	68	51	53	72	44	81	19	27	73	68	6	4	4	1	74	15	79	190	244	161	133	689						
Alack and Suddenly	161	106	114	117	200	213	158	192	177	201	236	235	226	194	150	157	171	132	143	163	133	390	606	498	769	839	490	3304				
Alack and Suddenly	1369	1254	1065	950	1237	1280	1050	1343	1089	1393	1162	1144	1250	1259	1278	2035	2268	2130	2315	2111	189	9277	8453	4078	1910	7788	4519	32106				
Alack and Suddenly	103	71	85	82	76	102	80	101	85	120	113	179	116	167	48	57	37	50	105	87	341	359	497	247	1389	508						
Alack and Suddenly	2423	2200	2388	1888	2530	210	2280	2863	2600	1845	2757	2610	2982	2414	1827	1910	1713	1797	1754	1555	2080	247	5157	8262	8999	9914	2157	7197	44487			
Alack and Suddenly	684	491	530	493	569	653	606	828	702	1027	807	841	742	1031	52	87	18	241	221	386	418	709	499	1734	1168	1056	317	1376	9073			
Alack and Suddenly	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
Alack and Suddenly	183	2	3	1	1	2	4	1	3	5	6	4	8	5	1	5	1	5	1	5	1	5	1	5	1	5	1	5				
Alack and Suddenly	47	40	30	27	49	50	53	30	43	49	63	60	57	48	43	33	29	34	37	32	32	45	159	147	144	182	255	130	827			
Alack and Suddenly	8	17	29	43	24	12	19	22	20	18	7	18	19	13	12	18	13	13	13	13	13	13	13	13	13	13	13					
Alack and Suddenly	3	2	3	1	3	1	1	2	4	1	3	5	6	4	8	5	1	5	1	5	1	5	1	5	1	5	1	5				
Alack and Suddenly	159	400	1190	184	52	1279	132	812	1089	821	833	835	409	152	354	74	40	58	531	73	1354	293	24	82	69	29	34	29	243			
Alack and Suddenly	6	6	9	8	7	2	14	4	9	11	2	20	23	25	53	51	17	12	12	12	12	17	12	12	12	12	12	12				
Alack and Suddenly	15	29	15	18	21	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20				
Alack and Suddenly	4	4	1	1	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2			
Alack and Suddenly	9	5	12	9	7	7	5	6	8	7	8	13	14	2	2	5	3	4	4	5	7	20	71	50	48	59	45	47	279			
Alack and Suddenly	12	13	16	7	17	14	11	14	17	10	13	13	14	16	24	18	11	13	20	11	14	17	15	14	16	17	16	15	111			
Alack and Suddenly	11	10	13	14	9	14	15	9	14	16	24	18	11	13	8	8	6	15	3	3	6	14	14	14	17	16	15	111				
Alack and Suddenly	57	39	49	41	43	57	71	61	41	46	77	102	76	47	59	35	43	33	45	54	11	47	35	62	5	428	228	1609				
Alack and Suddenly	75	61	65	58	80	105	79	90	92	122	80	134	105	96	58	76	73	74	10	60	62	73	10	60	62	201	201	201				
Alack and Suddenly	27	57	39	94	47	45	57	58	52	43	52	47	55	47	54	55	47	46	49	40	41	51	69	92	153	94	97	102	66	537		
Alack and Suddenly	27	26	22	19	22	20	26	27	27	28	25	28	24	28	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25			
Alack and Suddenly	3	4	3	4	4	4	3	10	9	4	6	2	6	4	1	2	2	3	2	3	2	3	2	3	2	3	2	3	2			
Alack and Suddenly	53	46	50	58	65	72	67	65	52	50	58	52	51	52	53	54	52	53	54	52	53	54	52	53	54	52	53	54	52	53		
Alack and Suddenly	12	18	6	11	2	11	9	12	6	9	9	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14		
Alack and Suddenly	12	13	5	8	6	6	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14		
Alack and Suddenly	51	92	5	33	23	60	9	52	11	133	25	80	6	74	42	2	3	80	21	33	37	18	107	13	155	252	57	757				
Alack and Suddenly	2	1	1	1	1	1	1	1	2	1	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Alack and Suddenly	2	2	7	5	4	3	3	3	6	6	5	7	7	20	1	3	7	8	6	5	4	3	4	16	123	245	66	529				
Alack and Suddenly	2	5	2	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
Alack and Suddenly	25	22	36	28	28	20	30	36	38	53	44	50	46	43	4	10	13	7	14	15	16	18	21	17	23	17	25	14	21	23	17	25
Alack and Suddenly	27	21	39	20	23	20	29	18	22	23	20	21	22	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17		
Alack and Suddenly	3	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Alack and Suddenly	3597	611	67	13	23	16	6	16	9	6	4	14	13	16	13	17	8	1	10400	112	90	89	89	89	89	89	89	89	89			
Alack and Suddenly	30	26	13	20	23	19	17	23	10	9	8	16	12	10	26	24	21	45	24	24	20	29	21	29	21	29	21	29	21	29		
Alack and Suddenly	52	20	21	21	21	29	43	41	44	103	71	82	82	95	12	5	7	9	5	0	25	33	34	94	132	300	115	93	20			
Alack and Suddenly	15	17	17	16	26	32	25	32	23	34	40	47	61	48	23	24	20	48	19	19	22	29	91	89	65	113	144	141				
Alack and Suddenly	15	17	17	16	26	32	25	32	23	34	40	47	61	48	23	24	20	48	19	19	22	29	91	89	65	113	144	141				
Alack and Suddenly	15	17	17	16	26	32	25	32	23	34	40	47	61	48	23	24	20	48	19	19	22	29	91	89	65	113	144	141				
Alack and Suddenly	15	17	17	16</td																												

What is big data?

- **1854:** John Snow maps London Cholera outbreaks and finds that they are clustered around a single pump; it was found that a cesspit was leaking into that pump



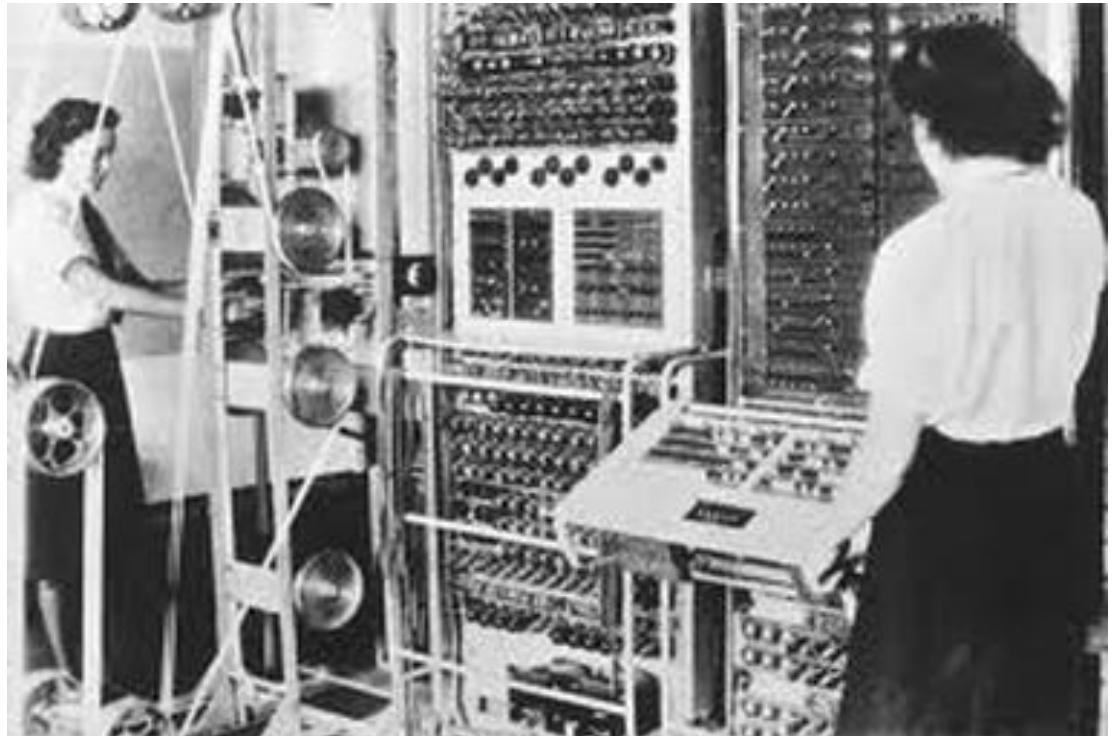
What is big data?

- **1880:** US Census Bureau estimates it will take eight years to process the data collected in the 1880 census, and over 10 years to process 1890 census data
- **Hollerith Tabulating Machine**
(punch card tabulation) reduces to ~3 months



What is big data?

- **WW2:** British invent the Colossus machine to scan for patterns in intercepted Nazi codes. Scans 5,000 characters a second, reducing workload from weeks to hours



What is big data?

- 1944: Librarian Fremont Rider @ Wesleyan estimates American university libraries doubling in size every 16 years:

*“the Yale Library in 2040 will have approximately 200,000,000 volumes,
which will occupy over 6,000 miles of shelves...
[requiring] a cataloging staff of over six thousand persons”*

What is big data?

- 1961: Derek Price shows # of new scientific journals growing exponentially rather than linearly

“each advance generates a new series of advances at a reasonably constant birth rate, so that the number of births is strictly proportional to the size of the population of discoveries at any given time”

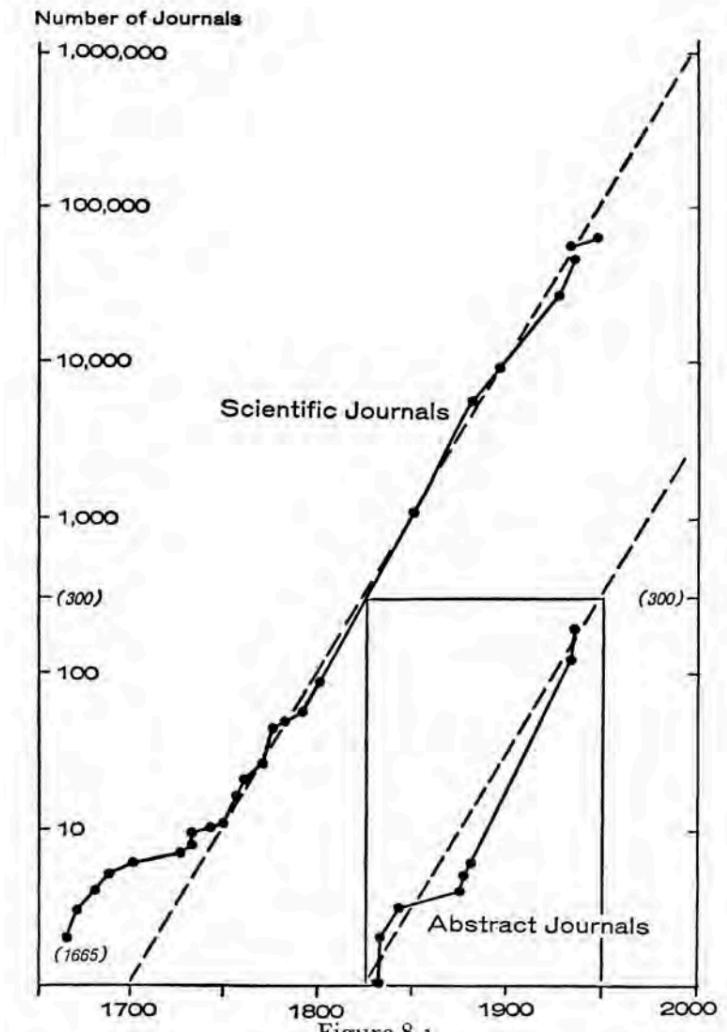


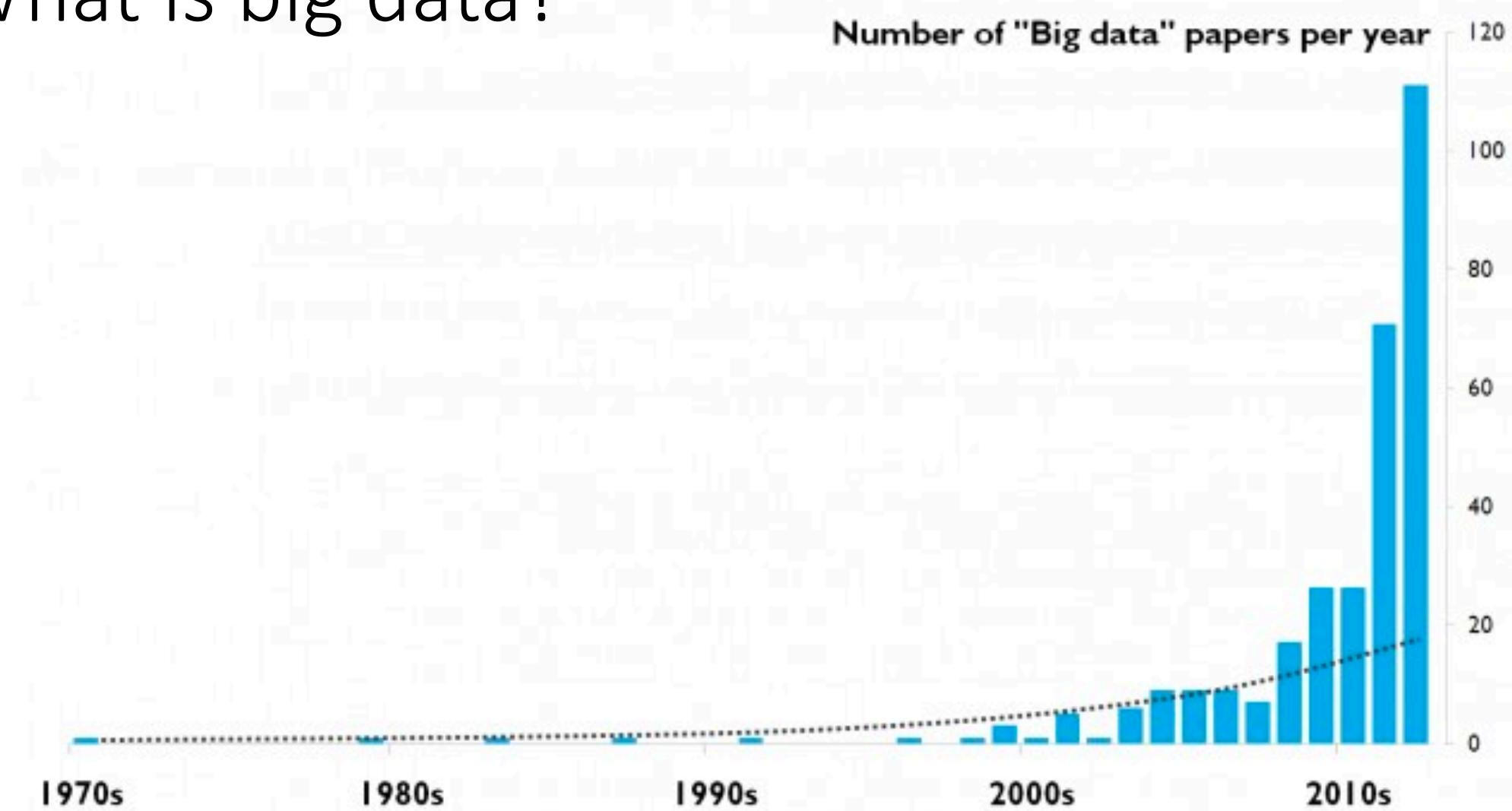
Figure 8.1

Number of journals founded (*not surviving*) as a function of date. The two uppermost points are taken from a slightly differently based list.

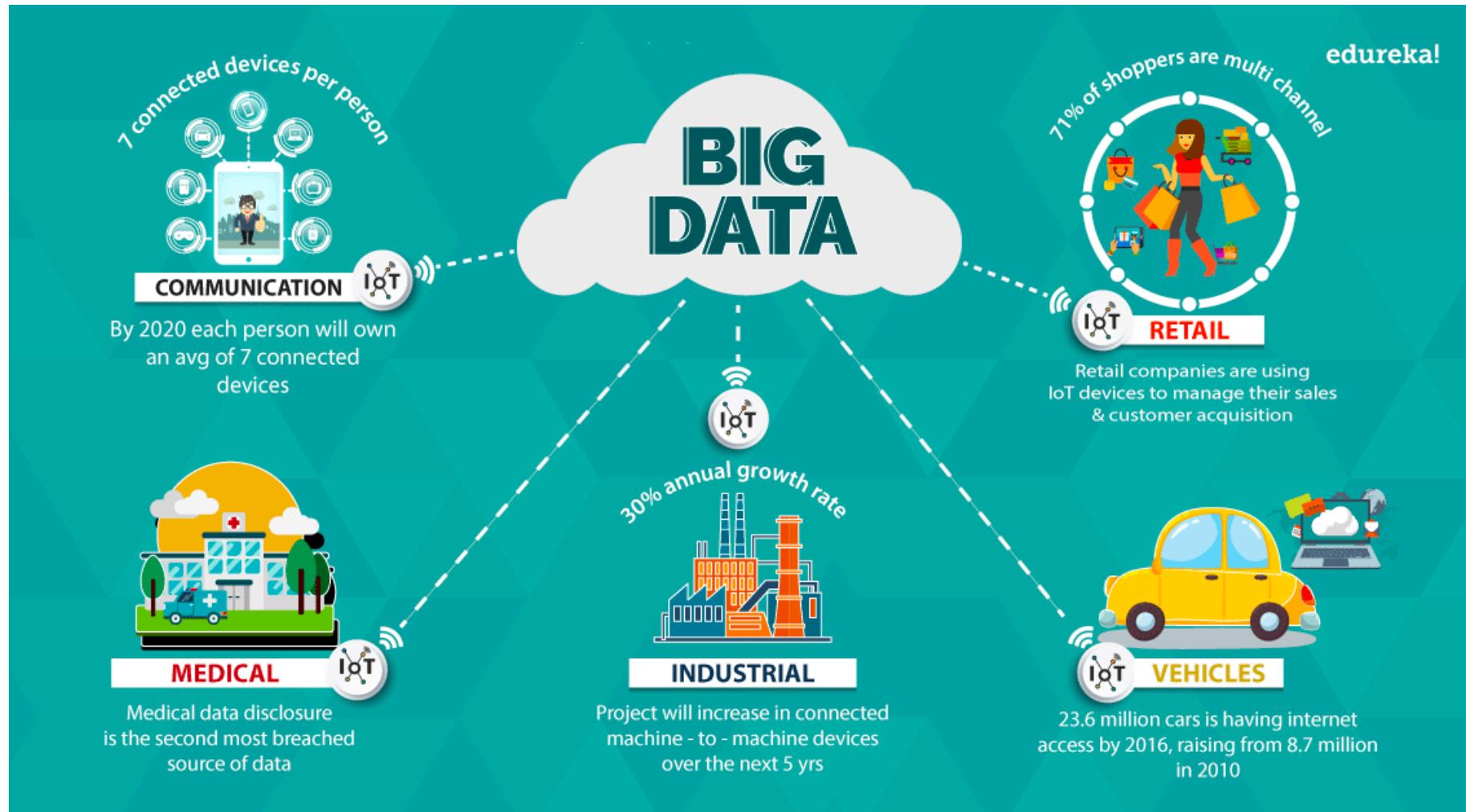
What is big data?

- **1969:** ARPANET kicks off the Internet
- **1977:** First personal computers come on the market
- **1989:** Tim Berners-Lee introduces the concept of the World Wide Web and the underlying protocols that support it (HTML, URL, HTTP)
- **1993:** CERN announces WWW will be free for everyone to develop and use
- **1990s– 2000s:** The explosion of the internet
- **2005:** Roger Moughalas coins term ‘Big Data’, referring to the scale of data that is nearly impossible to manage and process w/ available tools

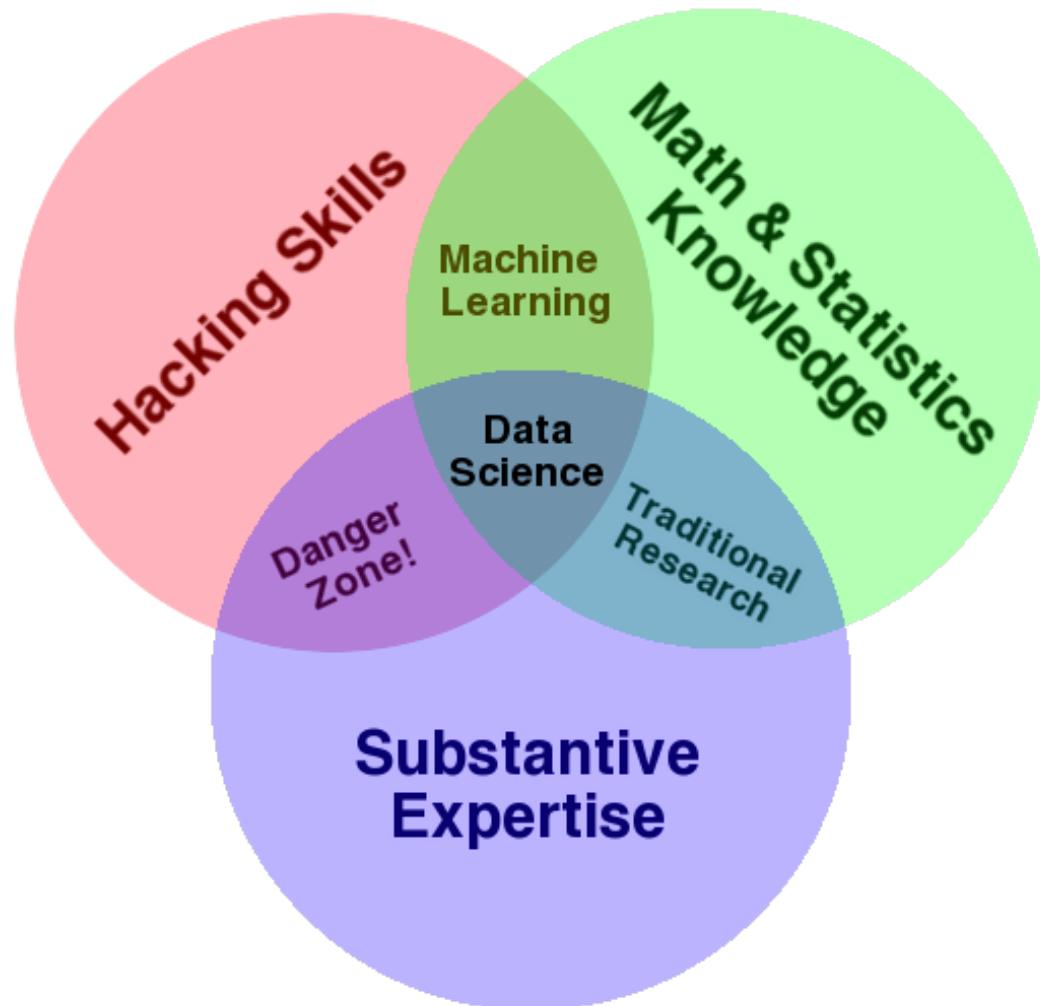
What is big data?



Where is big data?



What is data science?



What's the plan for this class?

- Learn the data science tools to collect, clean, understand, manipulate big data
- Exploration, Inference and Prediction
 - Exploration: Identify patterns in information
 - Tools: Visualization + Descriptive Statistics
 - Prediction: make informed guesses about what we want to know, based on the patterns that we identified
 - Tools: Machine Learning + Optimization
 - Inference: quantify our certainty about our predictions
 - Tools: Statistical Tests + Models

Login to Systems

- Class website: <https://cs.slu.edu/~stylianou/1070>
- JupyterHub: <http://cs1070.com>
 - Login: Your SLUNetID
 - Password: Set it the first time you login
- Piazza: <https://piazza.com/class/jzbep3yp13k3pr>
 - Should have gotten an email to active – let me know if you didn't!
- GitLab: http://git.cs.slu.edu/courses/fall19/csci_1070/<your-SLUNetID>
 - Login: SLUNetID
 - Password:
 - either you've done this already and should know it
 - or do a password reset the first time

Demo

Next class

- No class on Thursday 8/29 (Mass of the Holy Spirit)
- For next Tuesday, 9/3:
 - Assignment 0 ('Getting to Know You') due via git at beginning of class
 - Do readings that are listed for 8/27 and 8/29
 - We will discuss at the beginning of class