Big Data and Privacy

Gordon Hull

Director, Center for Professional and Applied Ethics Assoc. Prof. of Philosophy and Public Policy





Center for Professional & Applied Ethics

What is "Big Data?"



What is "Big Data?"



Center for Professional & Applied Ethics

The term is contested, but most generally:

- It's not the amount of data that matters
- It's the structure,
 - How it's collected, and
 - How it's processed



Traditional Databases



- <u>Structure</u>: Hierarchical
- <u>Collection</u>: Data collected specifically for a given purpose
- <u>Analyzed</u> by humans, looking for specific information, using statistical tools



Hierarchical Structure



\$0.00	\$2.07	\$2.00	\$1.50	\$0.00	\$0.90	8.97	5720.79	6.47	16.52	14.21	30.73	33.69	\$1,010.80	
2.50	\$3.77	\$2.50	\$1.00	\$1.40	\$0.00	15.17	5735.96	11.2	17.00	14.02	31.02	32.55	\$976.43	
1.50	\$9.07	\$3.50	\$0.00	\$0.70	\$0.30	33.07	5768.73	15.1	17.65	14.09	31.74	33.15	\$994.37	
9.00	\$10.47	\$0.50	\$1.50	\$1.40	\$0.00	46.87	5815.60	22.9	18.14	14.38	32.52	34.08	\$1,022.43	
2.50	\$4.44	\$2.00	\$1.50	\$3.15	\$0.00	33.84	5849.44	13.6	17.49	14.39	31.88	33.58	\$1,007.52	
5.50	\$13.72	\$1.50	\$0.50	\$4.20	\$0.90	44.57	5893.11	26.3	17.76	14.34	32.10	33.45	\$1,003.55	
2.00	\$4.72	\$2.50	\$0.00	\$1.05	\$0.00	37.27	5930.38	10.3	16.40	15.13	31.52	33.09	\$992.62	
2.00	\$1.56	\$0.00	\$0.00	\$0.35	\$0.60	10.51	5940.29	4.51	16.16	15.23	31.39	31.51	\$945.38	
1.00	\$4.00	\$0.00	\$1.00	\$1.05	\$0.00	14.05	5954.34	7.05	16.08	15.27	31.35	29.91	\$897.27	
3.50	\$13.42	\$7.50	\$0.50	\$0.70	\$0.30	48.07		25.9	16.85	15.71	32.56	31.06	\$931.67	
8.00	\$8.23	\$4.00	\$2.50	\$1.05	\$0.00	42.68	6044.79	23.8	17.44	15.65	33.09	32.04	\$961.09	
4.00	\$7.90	\$2.00	\$1.25	\$4.20	\$0.30	43.95	6088.44	19.7	17.12	16.17	33.28	32.28	\$968.37	
3.00	\$6.78	\$2.50	\$2.50	\$1.05	\$0.00	44.63	6133.07	15.8	16.78	16.94	33.72	32.34	\$970.29	
2.00	\$6.20	\$1.50	\$1.50	\$0.70	\$0.00	36.50	6169.57	11.9	15.31	17.55	32.87	31.79	\$953.70	
1.00	\$1.26	\$0.00	\$0.25	\$0.35	\$0.00	10.96	6180.53	2.86	15.06	17.95	33.01	30.95	\$928.44	
0.00	\$3.83	\$1.50	\$2.00	\$1.40	\$0.90	18.63	6198.26	9.63	14.95	18.31	33.26	30.91	\$927.25	
8.50	\$9.74	\$3.00	\$0.00	\$1.40	\$0.00	42.44	6240.70	22.6	15.49	18.44	33.93	31.78	\$953.46	~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~
3.50	\$9.17	\$2.00	\$1.50	\$1.40	\$1.20	41.87	6281.37	18.8	15.19	18.38	33.57	32.81	\$984.22	
3.00	\$9.73	\$5.50	\$0.50	\$5.25	\$0.00	49.78	6331.15	24	15.94	18.77	34.71	33.70	\$1,011.02	
4.00	\$11.00	\$3.00	\$2.00	\$21.00	\$0.00	72.20	6403.35	41	16.99	19.70	36.68	34.91	\$1,047.20	RAMEWORK BEFINITION USING SOCIAL YEARS BOY SYSTEMS SOCIAL YEARS MANNE INTERATION VOLUME MANAGE MENTION NATABASES PROCESSING MANAGEMENT SOCIAL YEARS MANNE INTERATION NATABASES SOFTWARE MANNE INTERATION NATABASES SOFTWARE MANNE INTERATION NATABASES SOFTWARE MANNE INTERATION NATABASES SOFTWARE MANNE INTERATION NATABASES SOFTWARE MANNE INTERATION NATABASES SOFTWARE MANNE INTERATION NATABASES SOFTWARE MANNE INTERATION NATABASES SOFTWARE MANNE INTERATION NATABASES SOFTWARE SOFTWARE SOFTWARE NATABASES SOFTWARE NATABASES SOFTWARE NATABASES SOFTWARE NATABASES SOFTWARE NATABASES SOFTWARE NATABASES SOFTWARE SOFTWARE NATABASES SOFTWARE SOFTWA
1.00	\$7.12	\$1.00	\$1.50	\$3.50	\$0.00	40.52	6443.87	14.1	17.26	19.65	36.91	34.83	\$1,044.93	
													Per citati	
														AND
														HYPOTHESIS SCIENCE
														PEOPLE TRAFFIC BE RELATED TIME SCHOOK B NOW

Collection



Center for Professional & Applied Ethics



-> Notice that getting consent is easy conceptualize in this context





Center for Professional & Applied Ethics

RELEVANT



-> The analysis run is chosen by the analyst, guided by the data and it structure, and often attempts causal explanation

Big Data negates all these



Center for Professional & Applied Ethics



Data is often <u>unstructured and continuously updated</u>: maybe from aggregating what people "like" on Facebook, or what words get typed into Google searches or public birth records

Big Data negates all these



Center for Professional & Applied Ethics



Data is more <u>harvested</u> than collected:

some of it may be from explicit questions that users consensually answer – but a lot is from "digital footprints" left behind by web searches, site visits, phone calling metadata, etc.

Big Data negates all these



Center for Professional & Applied Ethics



And the analysis often uses <u>machine learning</u>: The algorithm adjusts and seeks correlations: (Canadian Tire example)

In short:



- 1. The goal is correlations (not causes) that are predictive enough for whatever task is at hand (when it is causal (e.g., in medicine) it's expressed as risk factors apparently caused by genetic variations)
- 2. It needs to be sufficiently accurate for task at hand (ex. delivering "relevant" information to Google user)
- 3. Continuously updated (search for airline tickets twice, you'll see)

UNC CHARLOTTE College of Liberal Arts & Sciences

Center for Professional & Applied Ethics

What is Privacy?

• Literature full of definitions

- How about: "rules for how we deal with information flows"
- Big Data creates populations of "relevantly" "similar" people



Privacy Problems - example



Center for Professional & Applied Ethics



Target was right – she was indeed pregnant...



Center for Professional & Applied Ethics

Here's some potential kinds of problems...

1. Big Data makes "consent" meaningless

- Doesn't bind data brokers
- Can combine with other information about you "out there"
- 3rd party rule (*U.S. v. Maryland*, 1979)



2. Spurious correlations/bad grouping

- Same problem as studying "disease clusters:" statistically, these happen, and mean nothing.
- Or, info may just be wrong (FICO scores for those with no actual Wells Fargo accounts)
- And you may not know if the correlation is spurious..





3. Downward Ratcheting



Center for Professional & Applied Ethics

-> You might change your driving behavior, too (when, where, etc.)





4. Reflecting/Perpetuating Social Prejudice

- Those who downloaded Grindr from Android Market were recommended a sex-offender location app
- If you typed "she invented" in Google, it used to say:
 - "do you mean 'he invented'"
 - That's an accurate but unfortunate reflection of history, but...



5. Changes "who" you are



- Not just *more* info but *different*
- We modify our behavior if we know we're being recorded subjectivity needs personal space to develop (you might not read that controversial book)



