Session 6-7. Big Data, Artificial Intelligence, Machine Learning .1

NATALIA MILOVANTSEVA, PHD
NRU HSE, FACULTY OF WORLD ECONOMY AND INTERNATIONAL AFFAIRS
MARCH 9, 2019

Plan

- Response to mid-course evaluation
- Big Data in social science research
- AI experiments
- History of AI
- A startup idea

Mid-course evaluation

1. Not clear what is the goal of the course.
   - Syllabus: "We will cover a little bit of material across key issues of the process of world's digital transformation, and the goal is to connect ideas together and engage in the digital economy through a group project"
2. The focus is unclear.
   - Syllabus: "We will jump around several topics where a common theme – digital transformation – is what holds everything together."
3. Material is not new. Language is too difficult / too simple. Expected Python learning. Don't understand evaluation criteria…
   - Syllabus: course "will not … include detailed technical discussions of each technology or aspire to achieve a comprehensive review of emerging technologies."
   - This is elective course. Are you in the right class?

Learning outcomes for sessions 6 (1st two)-7

- LO1: Investigate how big data could be used in social science research.
- LO2: Review the history of AI to understand the challenges of the first wave and how the second wave could be valuable to businesses.
- LO3: Identify increasing areas or functions where machines are simply better performers than humans.
- LO3: Investigate how machine learning could be used in a business context.

Data vs statistics

DATA

- Raw information from which statistics are derived
  - Datasets; machine-readable files

	year	id	wrkstat	hrs1	hrs2	evwork	occ
1	2006	1	1	35	-1	0	0
2	2006	2	1	40	-1	0	0
3	2006	3	5	-1	-1	1	0
4	2006	4	2	24	-1	0	0
5	2006	5	6	-1	-1	2	0
6	2006	6	1	37	-1	0	0
7	2006	7	1	40	-1	0	0
8	2006	8	4	-1	-1	0	0
9	2006	9	1	38	-1	0	0
10	2006	10	1	35	-1	0	0
11	2006	11	5	-1	-1	1	0
12	2006	12	8	-1	-1	1	0
13	2006	13	6	-1	-1	1	0
14	2006	14	1	43	-1	0	0
15	2006	15	7	-1	-1	1	0

STATISTICS

- Numbers that provide interpretation and summaries of data

Table 1206. **Adult Attendance at Sports Events by Frequency: 2007**

[**In thousands (2,343 represents 2,343,000), except percent. For fall 2007.** Based on survey and subject to sampling error; see source]

Event	Attend one or more times a month		Attend less than once a month		Event	Attend one or more times a month		Attend less than once a month	
	Number	Percent	Number	Percent		Number	Percent	Number	Percent
Auto racing—NASCAR	2,343	1.1	10,209	4.6	Weekend professional games	4,007	1.8	11,787	5.3
Auto racing—Other	2,384	1.1	7,443	3.4	Golf	1,499	0.7	6,122	2.8
Baseball	7,591	3.4	20,664	9.4	High school sports	10,850	4.9	10,557	4.8
Basketball:					Horse racing:				
College games	3,812	1.7	9,830	4.5	Flats, runners	1,279	0.6	5,860	2.7
Professional games	3,280	1.5	10,996	5.0	Trotters/harness	629	0.3	4,906	2.2
Bowling	1,602	0.7	5,460	2.5	Ice hockey	1,872	0.9	8,499	3.9
Boxing	990	0.5	5,012	2.3	Motorcycle racing	854	0.4	5,127	2.3
Equestrian events	475	0.2	5,177	2.3	Pro beach volleyball	403	0.2	4,729	2.1
Figure skating	391	0.2	5,044	2.3	Rodeo/bull riding	744	0.3	6,333	2.9
Fishing tournaments	740	0.3	4,933	2.2	Soccer	3,437	1.6	6,497	2.9
Football:					Tennis	901	0.4	5,527	2.5
College games	5,759	2.6	12,705	5.8	Truck and tractor pull/ mud racing	904	0.4	5,895	2.7
Monday night professional games	2,165	1.0	6,821	3.1	Wrestling—professional	943	0.4	5,562	2.5

Source: Mediamark Research, Inc., New York, NY, *Top-line Reports* (copyright). See also <http://www.mediamark.com/mri/docs/TopLineReports.html>.

The rise of big data in research and business

SOCIAL SCIENCE RESEARCH

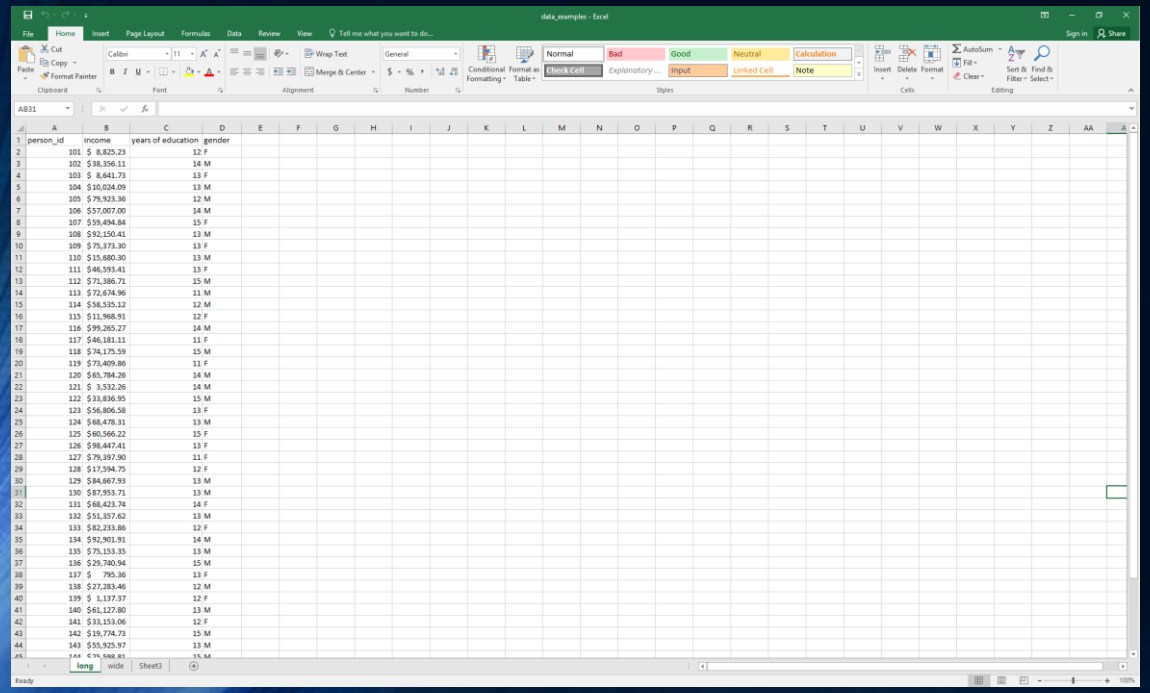
- Research in economics, political science, international relations, sociology, etc.
- Sources:
  - unstructured data: Twitter, newspapers
  - structured data: - government and corporate data

BUSINESS

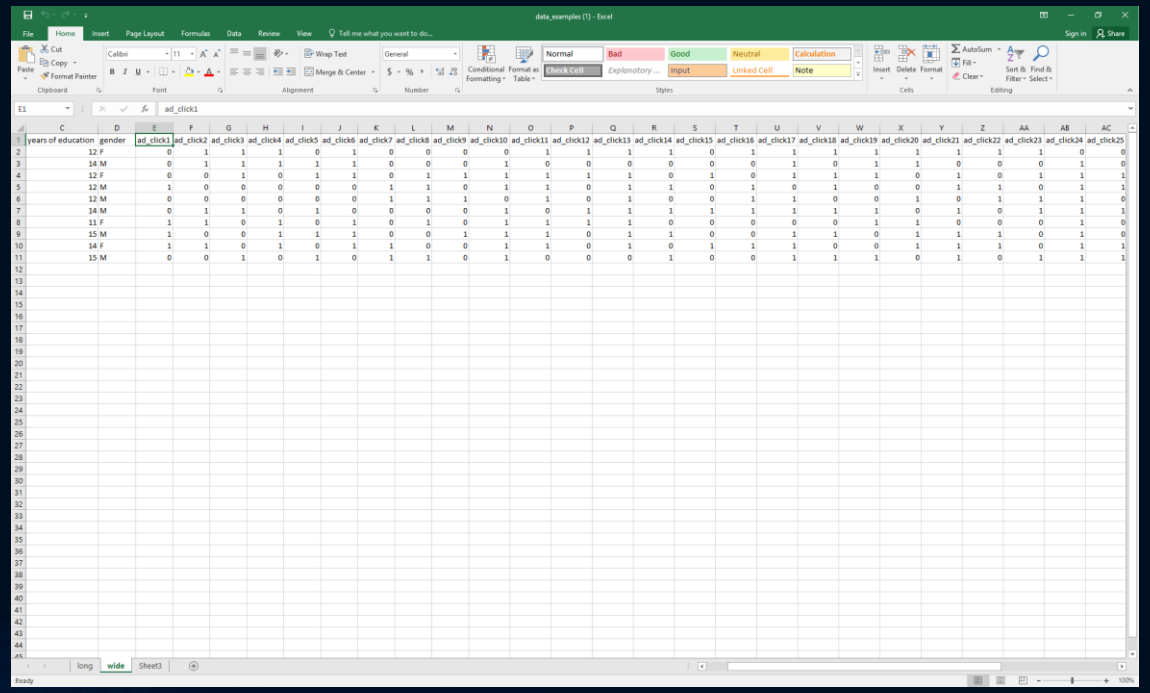
- Analyzed for "improving customer experiences"
- Analyzed for insights that lead to better decisions and strategic business moves (data-driven decision making)

Two types of big data

- "Long": many observations relative to variables (e.g., tax records)
- "Wide": few observations relative to variables (e.g. Amazon clicks)



person_id	income	years of education	gender
101	$ 8,825.23	12	F
102	$38,356.11	14	M
103	$ 8,641.73	13	F
104	$10,024.09	13	M
105	$79,923.36	12	M
106	$57,007.00	14	M
107	$59,494.84	15	F
108	$92,150.41	13	M
109	$75,373.30	13	F
110	$15,680.30	13	M
111	$46,593.41	13	F
112	$71,386.71	15	M
113	$72,674.96	11	M
114	$58,535.12	12	M
115	$11,968.91	12	F
116	$99,265.27	14	M
117	$46,181.11	11	F
118	$74,175.59	15	M
119	$73,409.86	11	F
120	$65,784.26	14	M
121	$ 3,532.26	14	M
122	$33,836.95	15	M
123	$56,806.58	13	F
124	$68,478.31	13	M
125	$60,566.22	15	F
126	$98,447.41	13	F
127	$79,397.90	11	F
128	$17,594.75	12	F
129	$84,667.93	13	M
130	$87,953.71	13	M
131	$68,423.74	14	F
132	$51,357.62	13	M
133	$82,233.86	12	F
134	$92,901.91	14	M
135	$75,153.35	13	M
136	$29,740.94	15	M
137	$ 795.36	13	F
138	$27,283.46	12	M
139	$ 1,137.37	12	F
140	$61,127.80	13	M
141	$33,153.06	12	F
142	$19,774.73	15	M
143	$35,925.97	13	M



years of education	gender	ad_click1	ad_click2	ad_click3	ad_click4	ad_click5	ad_click6	ad_click7	ad_click8	ad_click9	ad_click10	ad_click11	ad_click12	ad_click13	ad_click14	ad_click15	ad_click16	ad_click17	ad_click18	ad_click19	ad_click20	ad_click21	ad_click22	ad_click23	ad_click24	ad_click25
12	F	0	1	1	1	0	1	0	0	0	0	1	1	1	1	0	1	1	1	1	1	1	1	0	0
14	M	0	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	1	0	1	1	0	0	0	1	0
12	F	0	0	1	0	1	1	0	1	1	1	1	0	1	1	0	1	0	1	1	1	0	1	1	0	1
12	M	1	0	0	0	0	0	1	1	0	1	1	0	1	1	0	1	0	1	0	0	1	1	0	1	1
12	M	0	0	0	0	0	0	1	1	1	0	1	0	1	0	0	1	1	0	1	0	1	0	1	1	0
14	M	0	1	1	0	1	0	0	0	0	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	1
11	F	1	1	0	1	0	1	0	1	0	1	1	1	1	0	0	0	0	0	1	1	0	0	0	1	0
15	M	1	0	0	1	1	0	0	1	1	1	0	1	1	0	0	1	1	0	1	1	1	0	1	0	0
14	F	1	1	0	1	0	1	1	0	0	1	1	0	1	0	1	1	1	0	0	1	1	1	0	1	1
15	M	0	0	1	0	1	0	1	1	0	1	0	0	0	1	0	0	1	1	1	0	1	0	1	1	1

Applicability of types of data

- Social science has mainly focused on "long" data
  - Application: identifying causal effects
  - Example: effects of improving schools on income
  - Method: **regression analysis**

- Computer science has focused on "wide" data
  - Application: prediction
  - Example: predicting income to target ads
  - Method: **machine learning**

Regression analysis

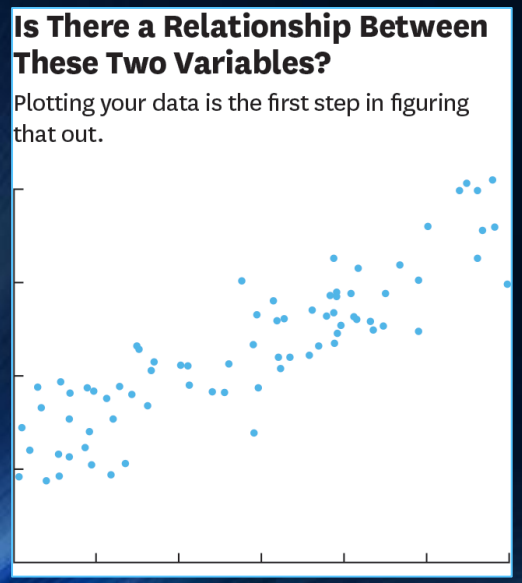
Answers the questions:

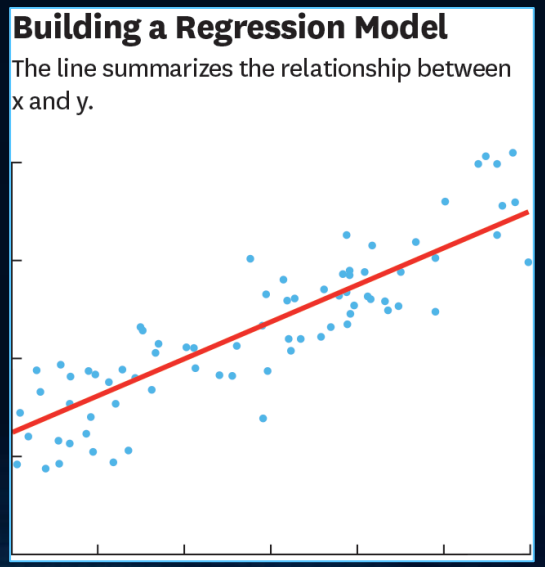
Which factors matter most?

Which can be ignored?

How do those factors interact with each other?

How certain are we about these factors?





Red line

- Computed using the method of least squares
- It is the best explanation of relationship between independent and dependent variables

Why is big data transforming social science?

- Economics has long been a theoretical field
  - Problem: untested theories lead to politicization

- Highly reliable data on a large scale
- Ability to measure new variables
- Almost universal coverage
  (can "zoom in" to subgroups)

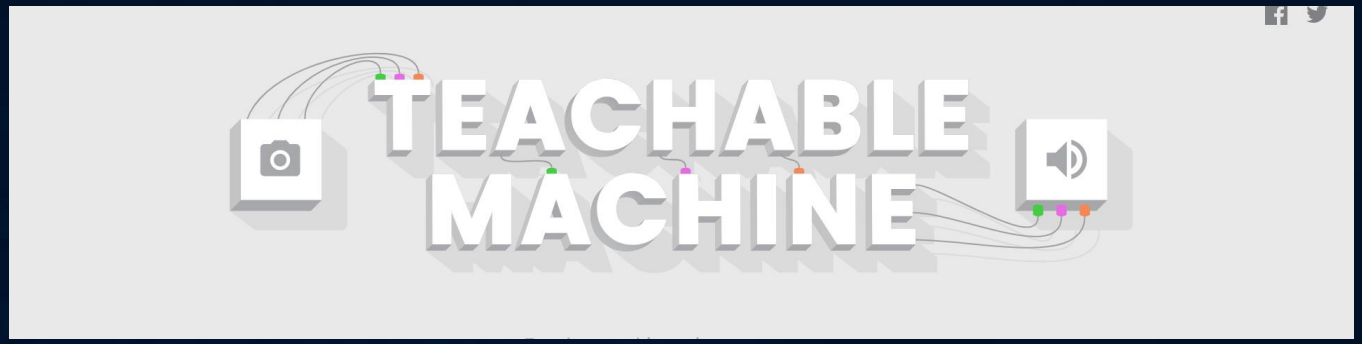
History of Artificial Intelligence

- AI is a general purpose technology (GPT)
  - GPTs affect entire economic system (examples: steam engine, electricity, Internet)
- The first wave of AI: a rule-based approach (*symbolic logic*)
  - logically codify what we know so that a computer reasons out an answer by logical deduction (`if … then …`)
  - limits of the rule-based approach: Polanyi's paradox ("*we know more than we can tell*")
  - explosion of interest in the 1980s, but only worked for a few narrow domains; followed by 20 years of "AI winter"

Overcoming Polanyi's paradox

- The second wave of AI
  - machine learning techniques enable machines to learn from examples
  - machines figure out the rules on their own
- Artificial general intelligence (AGI)
  - machines can do everything that humans can do
  - not to be confused with GPT
- Current progress with machine learning is economically significant and existentially trivial

Experience Artificial Intelligence



https://teachablemachine.withgoogle.com/

Turn up your sound

Startup example

The Databricks Unified Analytics Platform

https://www.youtube.com/watch?v=qosgI_uhBqM (2:02 min)