



The Business School  
for the World®

# DS(ML)B: Data Science (& Machine Learning) for Business

Profs. Anton Ovchinnikov, Theos Evgeniou, Spyros Zoumpoulis

Sessions 01-02

- Course Intro
- UDJ regression recall: “Sarah Gets a Diamond” case and competition
- Storytelling with data: visualizations in Tableau
- From Excel to R

# Why are we here?



“All-things-digital/data” (AI, Machine Learning, ...) is at the top of every leader’s agenda:

- **Forbes: “The Top 10 Business Trends That Will Drive Success...” AI is #1**  
<https://www.forbes.com/sites/ianaltman/2017/12/05/the-top-business-trends-that-will-drive-success-in-2018/#66beaf0701a>
- **Fortune “Five Big Business Trends to Watch...” – AI is #2** <http://fortune.com/2018/01/02/five-big-business-trends-to-watch-in-2018/>
- **Economist: “The world’s most valuable resource is no longer oil, but data”**  
<https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data>
- **WSJ, FT, ... – “data” regularly on the cover/1<sup>st</sup> page**

*“Companies have to race to build AI or  
they will be made uncompetitive.  
Essentially, if your competitor is racing  
to build AI, they will crush you.”*

**ELON MUSK**



Forbes:

- AI will generate **\$2.9 trillion** in business value and recover 6.2 billion hours of worker productivity by 2021.

<https://www.forbes.com/sites/louis columbus/2017/10/03/gartners-top-10-predictions-for-it-in-2018-and-beyond/#39ec316f45bb>

Forrester:

- AI-driven companies will take **\$1.2 trillion** from competitors by 2020.

[https://go.forrester.com/wp-content/uploads/Forrester\\_Predictions\\_2017\\_-Artificial\\_Intelligence\\_Will\\_Drive\\_The\\_Insights\\_Revolution.pdf](https://go.forrester.com/wp-content/uploads/Forrester_Predictions_2017_-Artificial_Intelligence_Will_Drive_The_Insights_Revolution.pdf)



*“The leaders in artificial  
intelligence will rule the world.”*

**VLADIMIR PUTIN**



# Why are we here?



- “All-things-digital/data” (AI, Machine Learning, ...) is at the top of every leader’s agenda



**Elon Musk**

@elonmusk

[Follow](#)



It begins ...



**The Verge** @verge

Putin says the nation that leads in AI ‘will be the ruler of the world’ [theverge.com/2017/9/4/16251...](http://theverge.com/2017/9/4/16251...)

2:11 AM - 4 Sep 2017

SPOTLIGHT ON BIG DATA

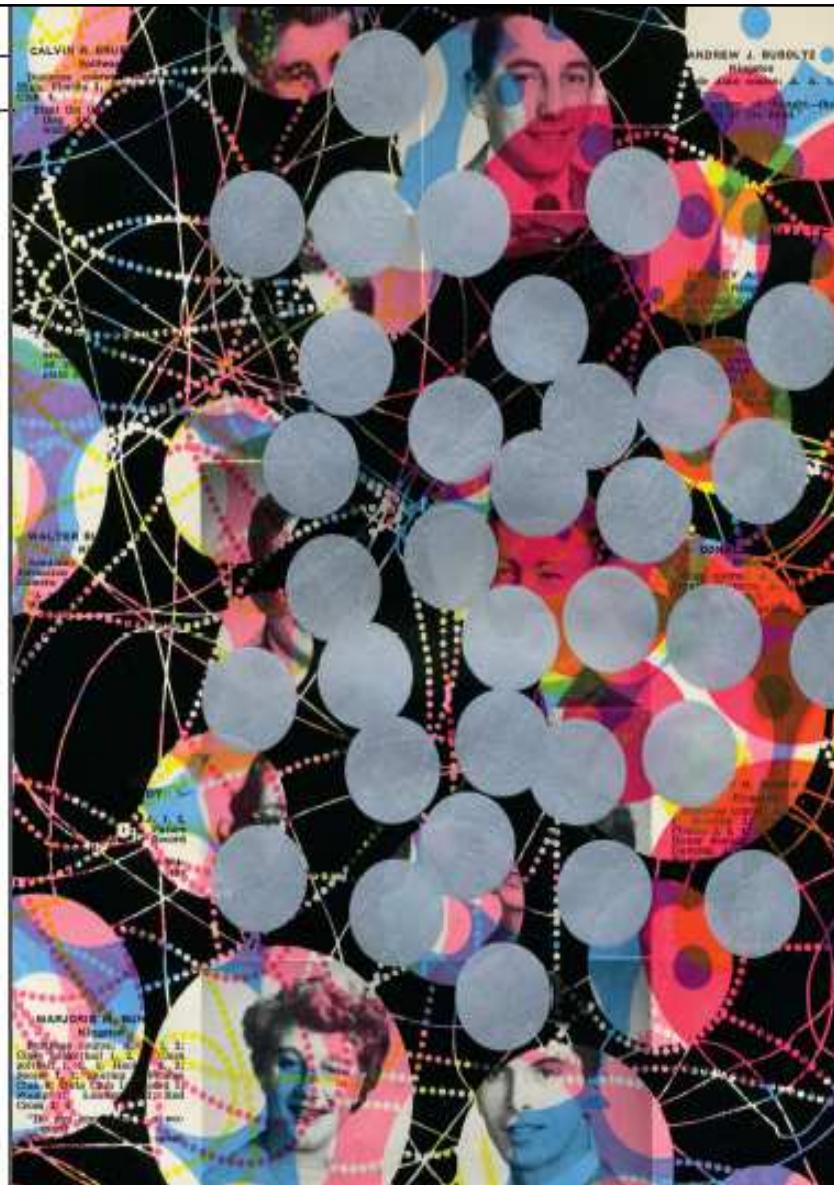
## Spotlight

ARTWORK Tavar Coker, Andrew J. Babiak  
Acrylic ink screen on a page from a high school  
yearbook, 10<sup>1/2</sup> x 11<sup>1/2</sup>

# Data Scientist: *The Sexiest Job of the 21st Century*

**Meet the people who can coax treasure out of messy, unstructured data.**  
by Thomas H. Davenport  
and D.J. Patil

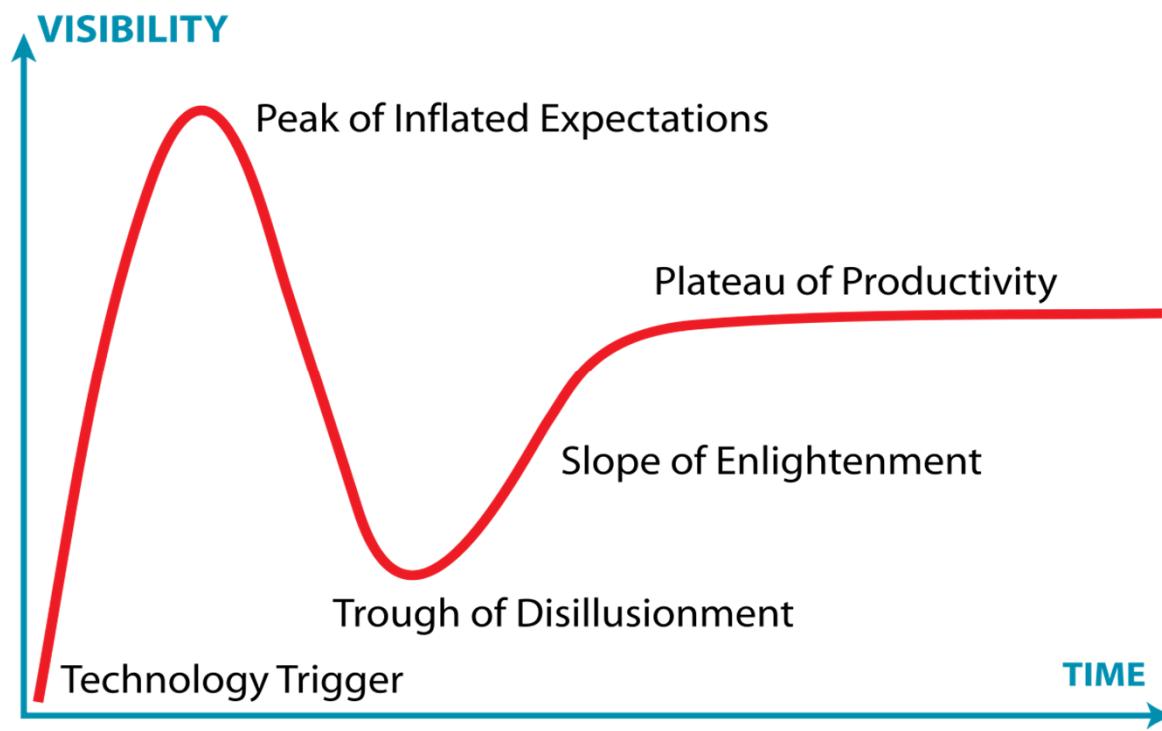
**W**hen Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 3 million members, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably know only"



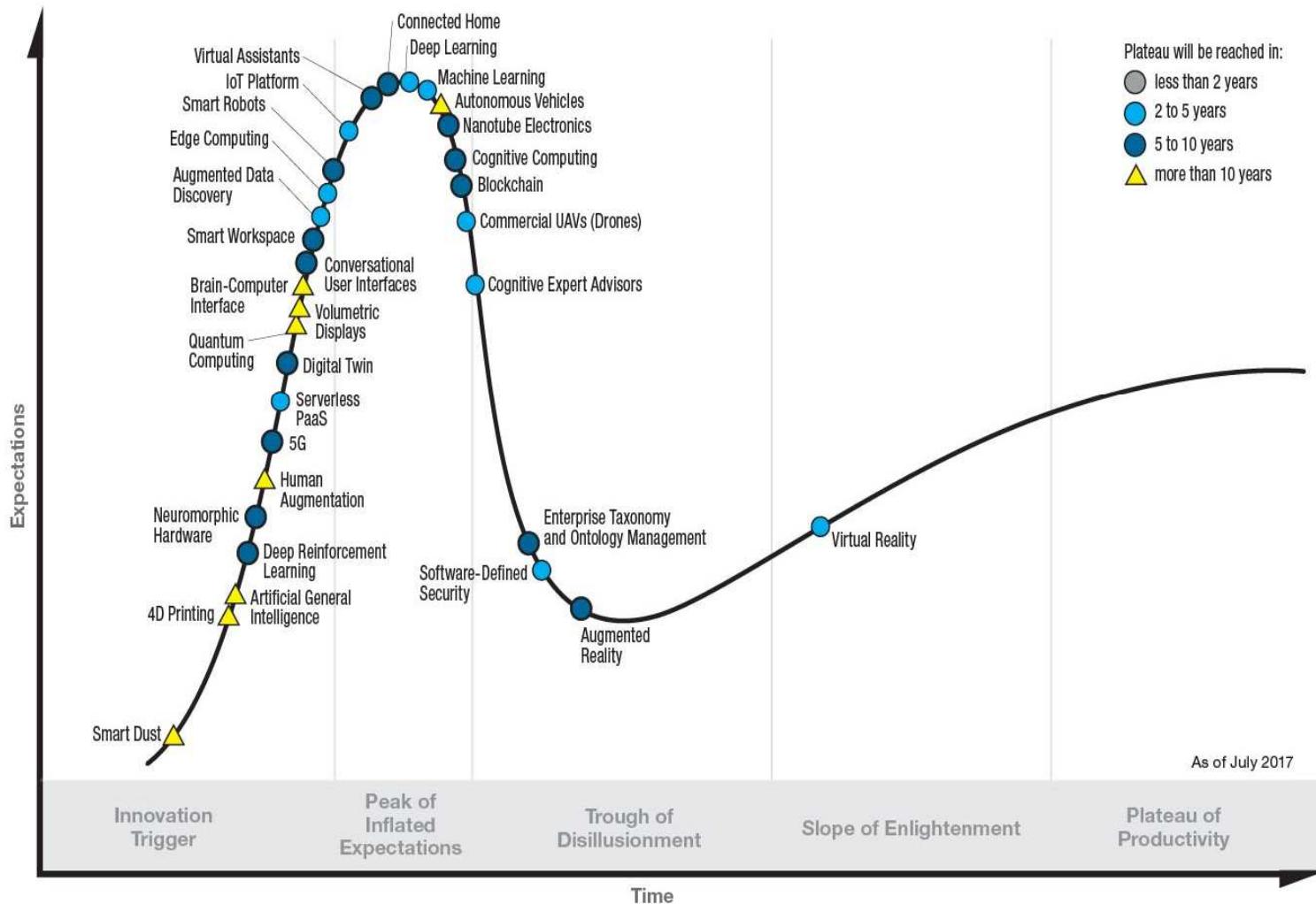
# Is “all-things-digital” just a hype?



The perception of highly publicized new technologies tends to follow a consistent pattern, Gartner hype cycle [www.gartner.com](http://www.gartner.com)



# Gartner Hype Cycle for Emerging Technologies, 2017



Source:

<https://www.gartner.com/doc/3783465?ref=SiteSearch&sthkw=Hype%20Cycle%20for%20Analyti cs%20and%20Business%20Intelligence&fnl=search&srcl=1-347892225#2048791703>

# Before we advance further: Some definitions/background



1. Descriptive vs Predictive vs Prescriptive Analytics
2. Big Data vs Smart Data
3. Data Science
4. AI vs Machine Learning vs Deep Learning
  - Intelligence – Learning – Data & Science
5. Supervised vs Unsupervised vs Reinforcement Learning
6. Why all this became so important NOW?

# Analytics: Descriptive vs Predictive vs Prescriptive

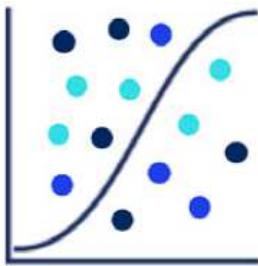


Descriptive



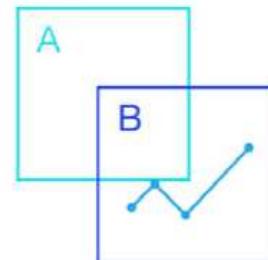
- Describe **what happened**
- Employed heavily across all industries

Predictive



- Anticipate **what will happen** (inherently probabilistic)
- Employed in data-driven organizations as a key source of insight

Prescriptive

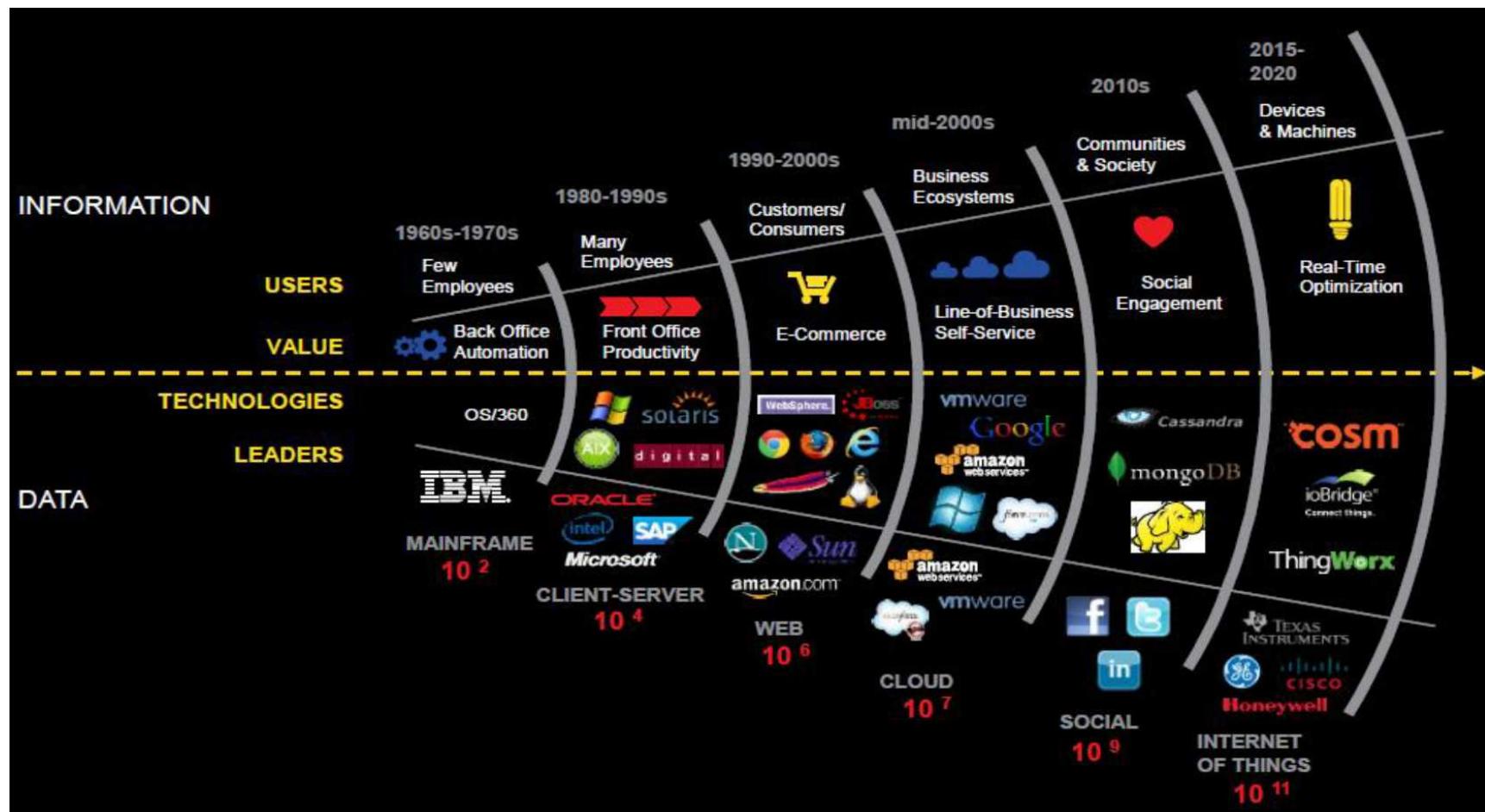


- Provide recommendations on **what to do** to achieve goals
- Employed heavily by leading data and Internet companies

Focus of machine learning

# Big Data vs Smart Data:

## What makes data “Big”? 3Vs: Volume, Variety, Velocity +



# Big Data vs Smart Data: What makes data “Smart”?



Smart data is:

- Data that is right for the decision
- Supports (and is supported by) analytics, expertise and machines
- Hits your key business drivers: customer acquisition, loyalty, growth, risk optimization, etc.
- “Big Data is **mostly** not about Data”

# Big Data vs Smart Data: Examples



## “Big data”

- Full-motion video feed from security cameras at a bank branch
- Real-time website click-stream data
- Raw twitter feed
- Your examples?

## “Smart data”

- Customer arrival patterns by time of day; security alert
- Purchase behavior segmentation
- Sentiment analyses
- Your examples?

# Big Data vs Smart Data: What makes data “Smart”?

Smart data is:

- Data that is right for the decision
- Supports (and is supported by) analytics, expertise and machines
- Hits your key business drivers: customer acquisition, loyalty, growth, risk optimization, etc.
- “Big Data is **mostly** not about Data”

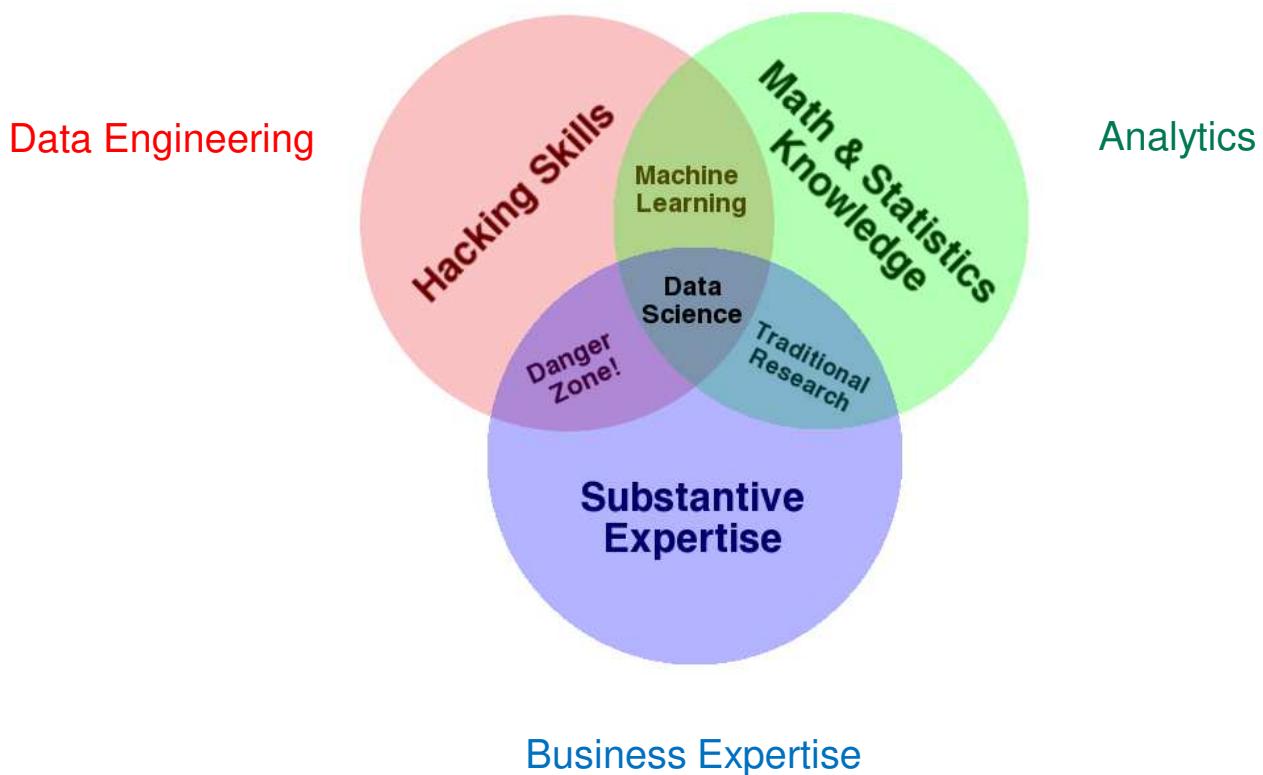
Data Engineering

Data Analytics

Business Expertise



# The driver of “Smart Data”: Data Science



# Making sense of AI and ML



ENCYCLOPÆDIA BRITANNICA

**Artificial intelligence**  
/ˌɑ:tɪ.flɪ̄.əl ɪn'tel.i.dʒəns /

the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings.



# Making sense of AI and ML



The ability for machines to take on human-like behavior and make decisions intelligently like humans

Artificial Intelligence

The ability for machines to intelligently learn something, without being specifically programmed to learn that

Machine Learning

Intelligence – Learning – Data & Science

Subset of Machine Learning that uses advanced Neural Networks with massive amounts of data to learn

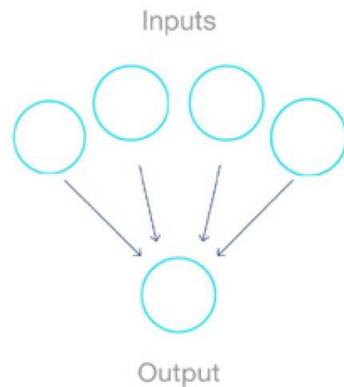
Deep Learning

"Sexy" rebranding of Neural Networks with certain clever structures/algorithms: Convolution NN, Recursive NN

# Learning: Supervised vs Unsupervised vs Reinforcement



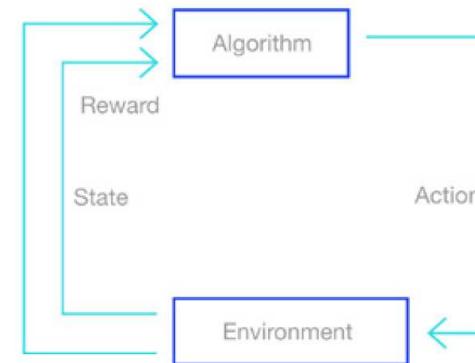
## Supervised learning



## Unsupervised learning



## Reinforcement learning



An algorithm uses training data and feedback from humans to learn the relationship of given inputs to a given output (eg, how the inputs “time of year” and “interest rates” predict housing prices)

**Regression:** predicting numbers  
**Classification:** predicting events

An algorithm explores input data without being given an explicit output variable (eg, explores customer demographic data to identify patterns)

**Clustering**  
**Anomaly detection**  
**Association rules**

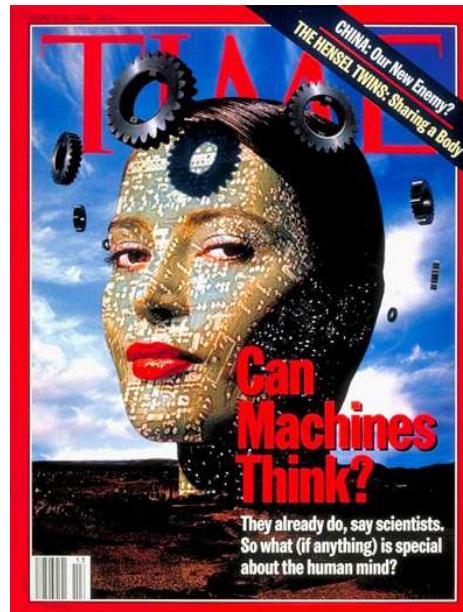
An algorithm learns to perform a task simply by trying to maximize rewards it receives for its actions (eg, maximizes points it receives for increasing returns of an investment portfolio)

**Learning to act based on feedback**  
Games (chess, Go), Driverless car [precise rules, stable environment]

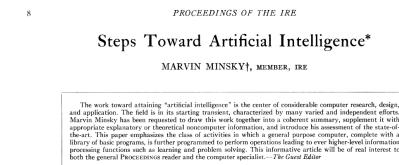
# Why this became so important NOW?



2017



1987



PROCEEDINGS OF THE IRE

Steps Toward Artificial Intelligence

MARVIN MINSKY†, MEMBER, I.

The work toward attaining "artificial intelligence" is the center of considerable computer research, design, and application. The field is in its starting transient, characterized by many varied and independent efforts. Marvin Minsky has been requested to draw this week together in his column summary, supplement it with his own comments, and then add his own thoughts on the present state of affairs and the direction he sees the field taking. This paper emphasizes the class of activities in which a general purpose computer, complete with its own library of programs, is further programmed to generate or receive higher-level information and process it further in the problem-solving activity. An informative article will be of real interest to both the general *PROCEEDINGS* reader and the computer specialist. —*The Guest Editor*

**Summarize:** The problems of heuristic programming—of making computers solve really difficult problems—are divided into five main categories: *Planning*, *Perception*, *Reasoning*, *Memory*, and *Induction*.

A computer can do many things, but only a few it is said to do well. One of the first things it does well is to know where certain parts of a scene are located. We can teach a machine to *Search* through some large space of possibilities, such as the possible ways of arranging a set of blocks, by an efficient process. With *Pattern-Recognition* techniques, efficiency can be increased. *Planning* is the ability to choose from among several methods to approach a problem. *Perception*, Recognition, and *Reasoning* are used together to solve a problem. *Memory* is used to accumulate further research. *Induction* is the ability to learn from experience. In the situation, after *Planning* methods, we usually find a *Random* search. This is a method of trial and error, which is often the most appropriate approach. To manage broad classes of problems, we must have a general scheme for *Induction*. This is the basic scheme for *Induction*.

The following discussion is supported by extensive citation of the literature and is composed of a few of the more interesting results.

**A**VISITOR TO our planet might be passed about the role of computers in our technology. On the one hand, he would read and hear about wondrous achievements in computer design and creation with prodigious intellectual performance. And for him it would be warned that these machines must be restrained from autonomy and initiative. For he would be told that even the evolution of truths too terrible to be borne. On the other hand, our visitor would find the analysis of the computer's potentialities for obediency, for innovative and interpretive, and for initiative, for autonomy or initiative; in short, for the development of problems. In this article, an attempt will be made to introduce the reader to the findings between some of these problems. Analysis will be supported with enough examples from the literature to serve the introductory purpose. The reader is invited to follow along with the relevant work not described here. This paper is highly

Our visitor might remain puzzled if he set out to find, and judge for himself, these monsters. For example:

"Received by the IRE, October 24, 1960. The author's work summarized here— which was done at Lincoln Lab., a center for research in applied science and technology, was supported in part by the U.S. Air Force, Navy, and Army and Air Force under Air Force contract AF-33(65)-13000, and by the Massachusetts Institute of Technology, Cambridge, Mass., which is supported in part by the U.S. Army Signal Research and Development Center, Fort Monmouth, N.J. It is based on early work done by the author as a junior Fellow of the Society of Industrial and Applied Mathematics, and it is based on work done by the author and his Companion, R. E. Labey, at the University of Michigan, Ann Arbor."

There is, of course, no generally accepted theory of "intelligence"; the analysis is our own and may be controversial. We regret that we cannot give full personal information about these matters here— suffice it to say that we have discussed these matters with almost every one of the cited authors.

of Electronics, M.I.T., Cambridge, Mass.

# Why this became so important NOW?



## Algorithms:

1960s: Rosenblatt (US), Ivakhnenko (UKR) - ANN  
1986: Hinton (CAN) – back-propagation  
1998: Brin (RUS/US) and Page (US) - pagerank  
2006: Hinton - “deep learning”

## Data:

1991: Internet  
1997: Google  
2000s: home PCs  
2004: Facebook  
2005: YouTube  
2007: iPhone [one]  
...

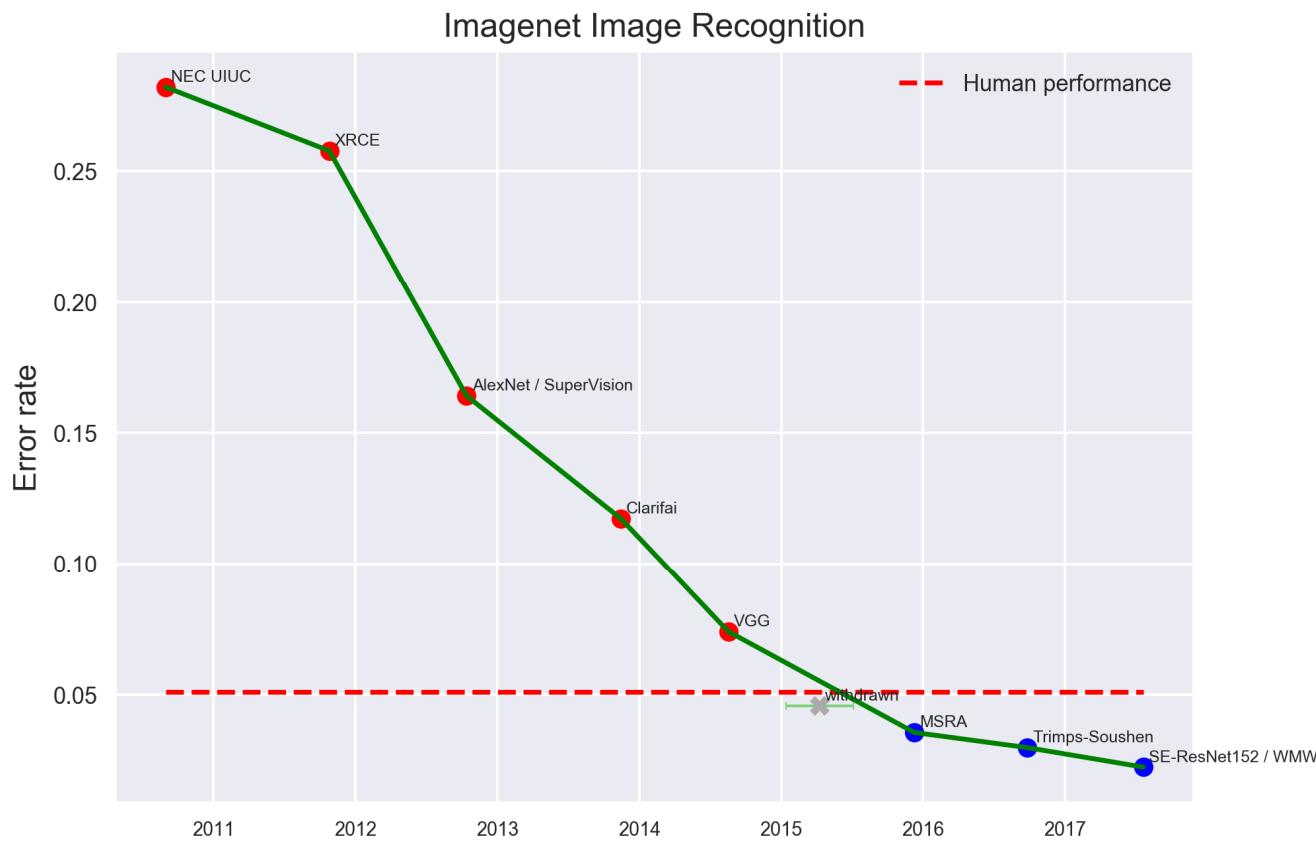
## Computing power:

1965: Moore’s law  
1999: GPU (Nvidia, Ng)  
2002: Amazon cloud  
2004: MapReduce  
2006: Hadoop (Yahoo)  
2009: Spark  
...

2015+

Par-human performance in various “intelligent” tasks  
Super-human performance in various situations with clearly defined simple rules (e.g., games)

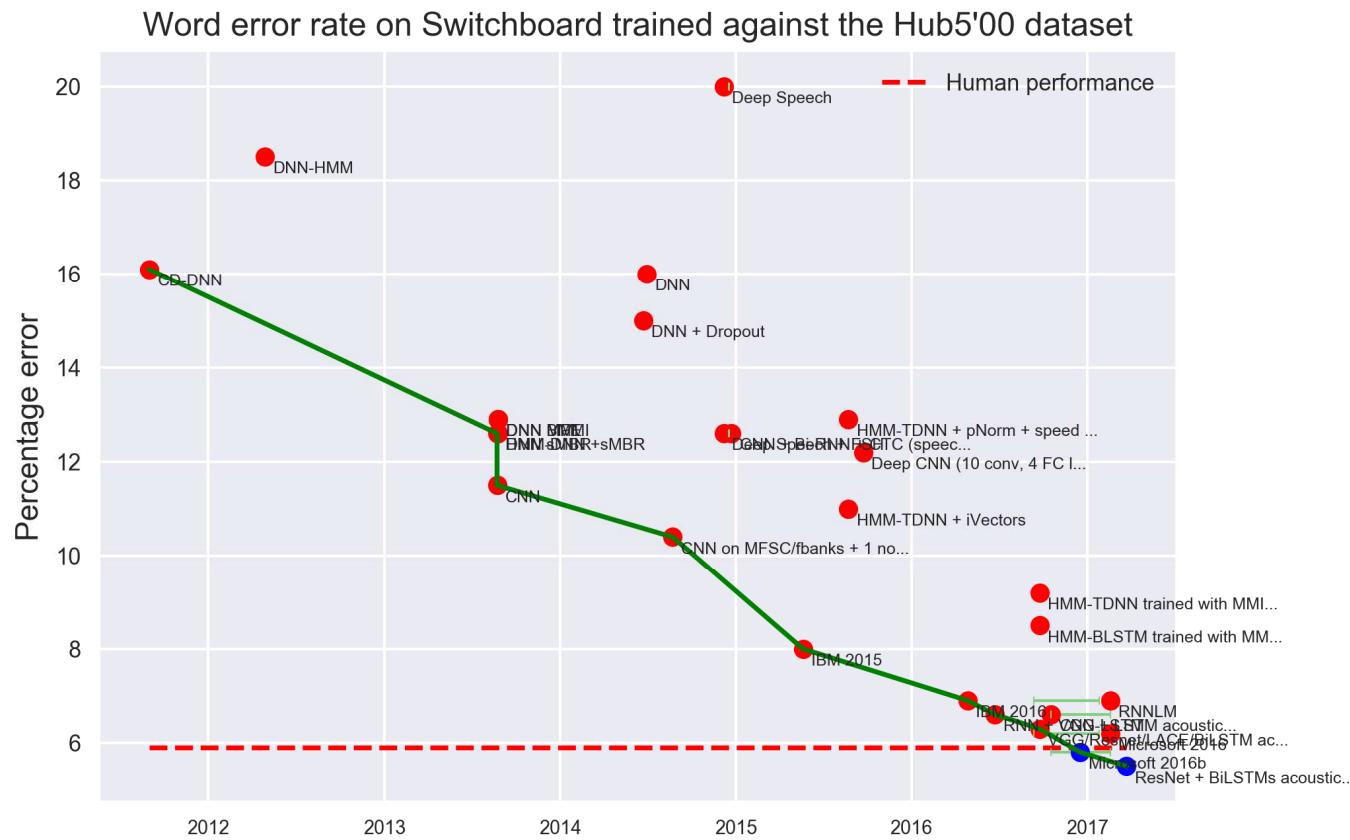
# Why this became so important NOW? Image recognition



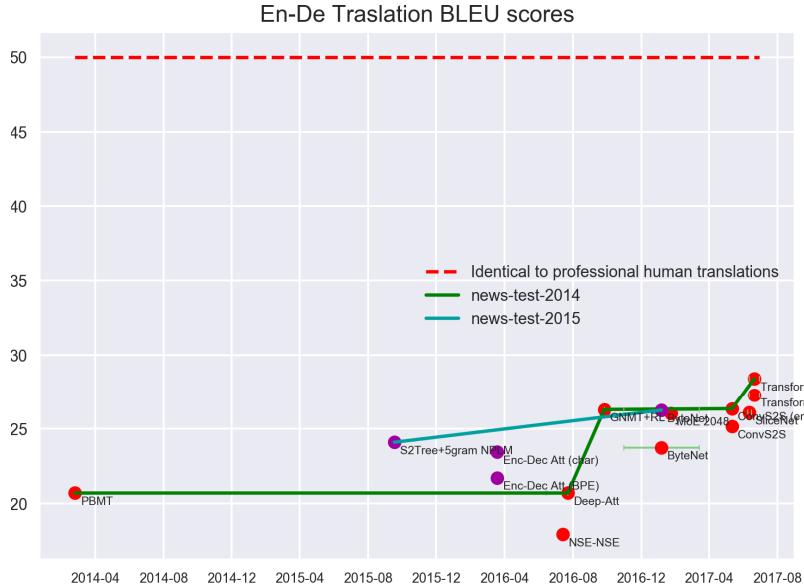
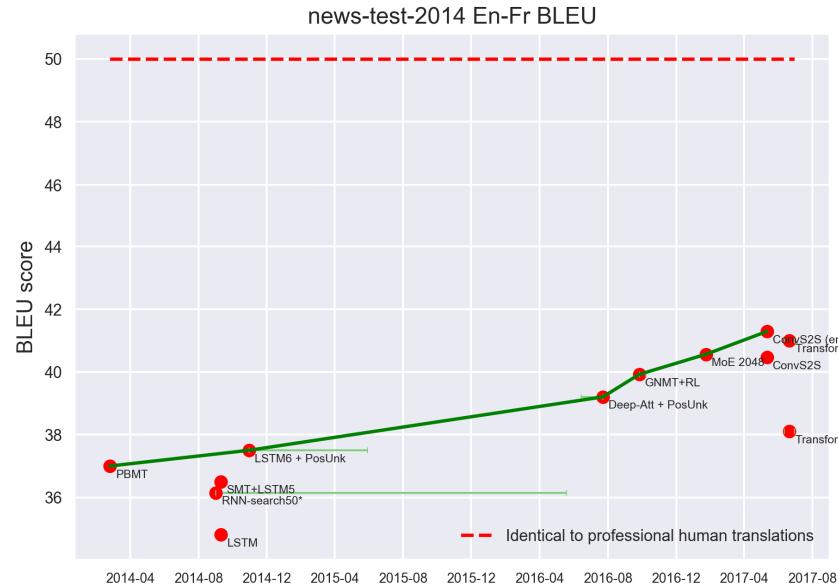
# Why this became so important NOW? Image recognition



# Why this became so important NOW? Voice-to-text



# Why this became so important NOW? Translations



Why worse than images or text?  
Why performance in French is better than in German?

More complex task  
More data (Canada's Parliament Records)

# Why this became so important NOW? “Games”



- Computer beats the best human chess players ever since IBM's Deep Blue defeated Kasparov in 1997.
- AlphaGo by DeepMind/Google beat best Go player in 2016 [Netflix documentary about it]
- AlphaZero has beaten the world's best chess-playing computer program, having **taught itself how to play in four hours** in 2017.
- How? Why?
  - **Reinforcement Learning:** works when clear “rules” are present and machines can play with themselves (create its own data)

THE DAILY NEWSLETTER  
Sign up to our daily email newsletter

## NewScientist

News Technology Space Physics Health Environment Mind | Travel Live Jobs

Home | Features | Technology



FEATURE 31 May 2017

## Human vs Machine: Five epic fights against AI

One by one, gaming champions are losing out to artificial intelligence. *New Scientist* takes a look at their glorious defeats – and stakes out the next battlefield



# Why this became so important NOW? How do these algorithms learn



Do they learn from / like humans?

Polanyi's paradox [https://en.wikipedia.org/wiki/Polanyi%27s\\_paradox](https://en.wikipedia.org/wiki/Polanyi%27s_paradox)



We know more  
than we can  
tell!



# Why this became so important NOW? How do these algorithms learn



Do they learn from / like humans?

Polanyi's paradox [https://en.wikipedia.org/wiki/Polanyi%27s\\_paradox](https://en.wikipedia.org/wiki/Polanyi%27s_paradox)



We know more  
than we can  
tell!



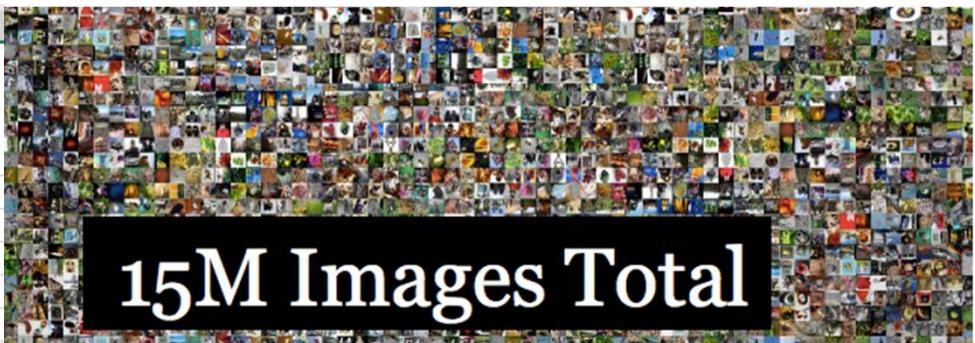
# Why this became so important NOW? How do these algorithms learn



Do they learn from / like humans?

- Both Yes, and No, and “We don’t know”
- 1. Data: “features” / variables that describe the situation
- Structured alpha-numerical data (transaction and customer characteristics)
- Unstructured: images (“ImageNet”), sound, network links

	A	B	C	D	E	F	M	N	S	T
1	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	BILL_AMT1	BILL_AMT2	PAY_AMT1	PAY_AMT2
2	1	20000	2	2	1	24	3913	3102	0	689
3	2	90000	2	2	2	34	29239	14027	1518	1500
4	3	50000	2	2	1	37	46990	48233	2000	2019
5	4	50000	1	2	1	57	8617	5670	2000	36681
6	5	50000	1	1	2	37	64400	57069	2500	1815



# Why this became so important NOW?

## How do these algorithms learn

Do they learn from / like humans?

- Both Yes, and No, and “We don’t know”

1. Data: “features” / variables that describe the situation
2. Feature Engineering: what information is in your data, but not captured by the existing variables?
  - “Retiree” ( $\text{age} > 65$ ), “single and male”, etc. → **VERY MANY**

#	Name	Description	Details
1	Feature_Binary_001	0/1	0/1
2	Feature_Binary_002	0/1	0/1
3	Feature_Binary_003	0/1	0/1
4	Feature_Binary_004	0/1	0/1
5	Feature_Binary_005	0/1	0/1
6	Feature_Binary_006	0/1	0/1
7	Feature_Binary_007	0/1	0/1
8	Feature_Binary_008	0/1	0/1
9	Feature_Binary_009	0/1	0/1
10	Feature_Binary_010	0/1	0/1
11	Feature_Binary_011	0/1	0/1
12	Feature_Binary_012	0/1	0/1
13	Feature_Binary_013	0/1	0/1
14	Feature_Binary_014	0/1	0/1
15	Feature_Binary_015	0/1	0/1
16	Feature_Binary_016	0/1	0/1
17	Feature_Binary_017	0/1	0/1
18	Feature_Binary_018	0/1	0/1
19	Feature_Binary_019	0/1	0/1
20	Feature_Binary_020	0/1	0/1
21	Feature_Binary_021	0/1	0/1
22	Feature_Binary_022	0/1	0/1
23	Feature_Binary_023	0/1	0/1
24	Feature_Binary_024	0/1	0/1
25	Feature_Binary_025	0/1	0/1
26	Feature_Binary_026	0/1	0/1
27	Feature_Binary_027	0/1	0/1
28	Feature_Binary_028	0/1	0/1
29	Feature_Binary_029	0/1	0/1
30	Feature_Binary_030	0/1	0/1
31	Feature_Binary_031	0/1	0/1
32	Feature_Binary_032	0/1	0/1
33	Feature_Binary_033	0/1	0/1
34	Feature_Binary_034	0/1	0/1
35	Feature_Binary_035	0/1	0/1
36	Feature_Binary_036	0/1	0/1
37	Feature_Binary_037	0/1	0/1
38	Feature_Binary_038	0/1	0/1
39	Feature_Binary_039	0/1	0/1
40	Feature_Binary_040	0/1	0/1
41	Feature_Binary_041	0/1	0/1
42	Feature_Binary_042	0/1	0/1
43	Feature_Binary_043	0/1	0/1
44	Feature_Binary_044	0/1	0/1
45	Feature_Binary_045	0/1	0/1
46	Feature_Binary_046	0/1	0/1
47	Feature_Binary_047	0/1	0/1
48	Feature_Binary_048	0/1	0/1
49	Feature_Binary_049	0/1	0/1
50	Feature_Binary_050	0/1	0/1
51	Feature_Binary_051	0/1	0/1
52	Feature_Binary_052	0/1	0/1
53	Feature_Binary_053	0/1	0/1
54	Feature_Binary_054	0/1	0/1
55	Feature_Binary_055	0/1	0/1
56	Feature_Binary_056	0/1	0/1
57	Feature_Binary_057	0/1	0/1
58	Feature_Binary_058	0/1	0/1
59	Feature_Binary_059	0/1	0/1
60	Feature_Binary_060	0/1	0/1
61	Feature_Binary_061	0/1	0/1
62	Feature_Binary_062	0/1	0/1
63	Feature_Binary_063	0/1	0/1
64	Feature_Binary_064	0/1	0/1
65	Feature_Binary_065	0/1	0/1
66	Feature_Binary_066	0/1	0/1
67	Feature_Binary_067	0/1	0/1
68	Feature_Binary_068	0/1	0/1
69	Feature_Binary_069	0/1	0/1
70	Feature_Binary_070	0/1	0/1
71	Feature_Binary_071	0/1	0/1
72	Feature_Binary_072	0/1	0/1
73	Feature_Binary_073	0/1	0/1
74	Feature_Binary_074	0/1	0/1
75	Feature_Binary_075	0/1	0/1
76	Feature_Binary_076	0/1	0/1
77	Feature_Binary_077	0/1	0/1
78	Feature_Binary_078	0/1	0/1
79	Feature_Binary_079	0/1	0/1
80	Feature_Binary_080	0/1	0/1
81	Feature_Binary_081	0/1	0/1
82	Feature_Binary_082	0/1	0/1
83	Feature_Binary_083	0/1	0/1
84	Feature_Binary_084	0/1	0/1
85	Feature_Binary_085	0/1	0/1
86	Feature_Binary_086	0/1	0/1
87	Feature_Binary_087	0/1	0/1
88	Feature_Binary_088	0/1	0/1
89	Feature_Binary_089	0/1	0/1
90	Feature_Binary_090	0/1	0/1
91	Feature_Binary_091	0/1	0/1
92	Feature_Binary_092	0/1	0/1
93	Feature_Binary_093	0/1	0/1
94	Feature_Binary_094	0/1	0/1
95	Feature_Binary_095	0/1	0/1
96	Feature_Binary_096	0/1	0/1
97	Feature_Binary_097	0/1	0/1
98	Feature_Binary_098	0/1	0/1
99	Feature_Binary_099	0/1	0/1
100	Feature_Binary_100	0/1	0/1
101	Feature_Binary_101	0/1	0/1
102	Feature_Binary_102	0/1	0/1
103	Feature_Binary_103	0/1	0/1
104	Feature_Binary_104	0/1	0/1
105	Feature_Binary_105	0/1	0/1
106	Feature_Binary_106	0/1	0/1
107	Feature_Binary_107	0/1	0/1
108	Feature_Binary_108	0/1	0/1
109	Feature_Binary_109	0/1	0/1
110	Feature_Binary_110	0/1	0/1
111	Feature_Binary_111	0/1	0/1
112	Feature_Binary_112	0/1	0/1
113	Feature_Binary_113	0/1	0/1
114	Feature_Binary_114	0/1	0/1
115	Feature_Binary_115	0/1	0/1
116	Feature_Binary_116	0/1	0/1
117	Feature_Binary_117	0/1	0/1
118	Feature_Binary_118	0/1	0/1
119	Feature_Binary_119	0/1	0/1
120	Feature_Binary_120	0/1	0/1
121	Feature_Binary_121	0/1	0/1
122	Feature_Binary_122	0/1	0/1
123	Feature_Binary_123	0/1	0/1
124	Feature_Binary_124	0/1	0/1
125	Feature_Binary_125	0/1	0/1
126	Feature_Binary_126	0/1	0/1
127	Feature_Binary_127	0/1	0/1
128	Feature_Binary_128	0/1	0/1
129	Feature_Binary_129	0/1	0/1
130	Feature_Binary_130	0/1	0/1
131	Feature_Binary_131	0/1	0/1
132	Feature_Binary_132	0/1	0/1
133	Feature_Binary_133	0/1	0/1
134	Feature_Binary_134	0/1	0/1
135	Feature_Binary_135	0/1	0/1
136	Feature_Binary_136	0/1	0/1
137	Feature_Binary_137	0/1	0/1
138	Feature_Binary_138	0/1	0/1
139	Feature_Binary_139	0/1	0/1
140	Feature_Binary_140	0/1	0/1
141	Feature_Binary_141	0/1	0/1
142	Feature_Binary_142	0/1	0/1
143	Feature_Binary_143	0/1	0/1
144	Feature_Binary_144	0/1	0/1
145	Feature_Binary_145	0/1	0/1
146	Feature_Binary_146	0/1	0/1
147	Feature_Binary_147	0/1	0/1
148	Feature_Binary_148	0/1	0/1
149	Feature_Binary_149	0/1	0/1
150	Feature_Binary_150	0/1	0/1
151	Feature_Binary_151	0/1	0/1
152	Feature_Binary_152	0/1	0/1
153	Feature_Binary_153	0/1	0/1
154	Feature_Binary_154	0/1	0/1
155	Feature_Binary_155	0/1	0/1
156	Feature_Binary_156	0/1	0/1
157	Feature_Binary_157	0/1	0/1
158	Feature_Binary_158	0/1	0/1
159	Feature_Binary_159	0/1	0/1
160	Feature_Binary_160	0/1	0/1
161	Feature_Binary_161	0/1	0/1
162	Feature_Binary_162	0/1	0/1
163	Feature_Binary_163	0/1	0/1
164	Feature_Binary_164	0/1	0/1
165	Feature_Binary_165	0/1	0/1
166	Feature_Binary_166	0/1	0/1
167	Feature_Binary_167	0/1	0/1
168	Feature_Binary_168	0/1	0/1
169	Feature_Binary_169	0/1	0/1
170	Feature_Binary_170	0/1	0/1
171	Feature_Binary_171	0/1	0/1
172	Feature_Binary_172	0/1	0/1
173	Feature_Binary_173	0/1	0/1
174	Feature_Binary_174	0/1	0/1
175	Feature_Binary_175	0/1	0/1
176	Feature_Binary_176	0/1	0/1
177	Feature_Binary_177	0/1	0/1
178	Feature_Binary_178	0/1	0/1
179	Feature_Binary_179	0/1	0/1
180	Feature_Binary_180	0/1	0/1
181	Feature_Binary_181	0/1	0/1
182	Feature_Binary_182	0/1	0/1
183	Feature_Binary_183	0/1	0/1
184	Feature_Binary_184	0/1	0/1
185	Feature_Binary_185	0/1	0/1
186	Feature_Binary_186	0/1	0/1
187	Feature_Binary_187	0/1	0/1
188	Feature_Binary_188	0/1	0/1
189	Feature_Binary_189	0/1	0/1
190	Feature_Binary_190	0/1	0/1
191	Feature_Binary_191	0/1	0/1
192	Feature_Binary_192	0/1	0/1
193	Feature_Binary_193	0/1	0/1
194	Feature_Binary_194	0/1	0/1
195	Feature_Binary_195	0/1	0/1
196	Feature_Binary_196	0/1	0/1
197	Feature_Binary_197	0/1	0/1
198	Feature_Binary_198	0/1	0/1
199	Feature_Binary_199	0/1	0/1
200	Feature_Binary_200	0/1	0/1
201	Feature_Binary_201	0/1	0/1
202	Feature_Binary_202	0/1	0/1
203	Feature_Binary_203	0/1	0/1
204	Feature_Binary_204	0/1	0/1
205	Feature_Binary_205	0/1	0/1
206	Feature_Binary_206	0/1	0/1
207	Feature_Binary_207	0/1	0/1
208	Feature_Binary_208	0/1	0/1
209	Feature_Binary_209	0/1	0/1
210	Feature_Binary_210	0/1	0/1
211	Feature_Binary_211	0/1	0/1
212	Feature_Binary_212	0/1	0/1
213	Feature_Binary_213	0/1	0/1
214	Feature_Binary_214	0/1	0/1
215	Feature_Binary_215	0/1	0/1
216	Feature_Binary_216	0/1	0/1
217	Feature_Binary_217	0/1	0/1
218	Feature_Binary_218	0/1	0/1
219	Feature_Binary_219	0/1	0/1
220	Feature_Binary_220	0/1	0/1
221	Feature_Binary_221	0/1	0/1
222	Feature_Binary_222	0/1	0/1
223	Feature_Binary_223	0/1	0/1
224	Feature_Binary_224	0/1	0/1
225	Feature_Binary_225	0/1	0/1
226	Feature_Binary_226	0/1	0/1
227	Feature_Binary_227	0/1	0/1
228	Feature_Binary_228	0/1	0/1
229	Feature_Binary_229	0/1	0/1
230	Feature_Binary_230	0/1	0/1
231	Feature_Binary_231	0/1	0/1
232	Feature_Binary_232	0/1	0/1
233	Feature_Binary_233	0/1	0/1
234	Feature_Binary_234	0/1	0/1
235	Feature_Binary_235	0/1	0/1
236	Feature_Binary_236	0/1	0/1
237	Feature_Binary_237	0/1	0/1
238	Feature_Binary_238	0/1	0/1
239	Feature_Binary_239	0/1	0/1
240	Feature_Binary_240	0/1	0/1
241	Feature_Binary_241	0/1	0/1
242	Feature_Binary_242	0/1	0/1
243	Feature_Binary_243	0/1	0/1
244	Feature_Binary_244	0/1	0/1
245	Feature_Binary_245	0/1	0/1
246	Feature_Binary_246	0/1	0/1
247	Feature_Binary_247	0/1	0/1
248	Feature_Binary_248	0/1	0/1
249	Feature_Binary_249	0/1	0/1
250	Feature_Binary_250	0/1	0/1
251	Feature_Binary_251	0/1	0/1
252	Feature_Binary_252	0/1	0/1
253	Feature_Binary_253	0/1	0/1
254	Feature_Binary_254	0/1	0/1
255	Feature_Binary_255	0/1	0/1
256	Feature_Binary_256	0/1	0/1
257	Feature_Binary_257	0/1	0/1
258	Feature_Binary_258	0/1	0/1
259	Feature_Binary_259	0/1	0/1
260	Feature_Binary_260	0/1	0/1
261	Feature_Binary_261	0/1	0/1
262	Feature_Binary_262	0/1	0/1
263	Feature_Binary_263	0/1	0/1
264	Feature_Binary_264	0/1	0/1
265	Feature_Binary_265	0/1	0/1
266	Feature_Binary_266	0/1	0/1
267	Feature_Binary_267	0/1	0/1
268	Feature_Binary_268	0/1	0/1
269	Feature_Binary_269	0/1	0/1
270	Feature_Binary_270	0/1	0/1
271	Feature_Binary_271	0/1	0/1
272	Feature_Binary_272	0/1	0/1
273	Feature_Binary_273	0/1	0/1

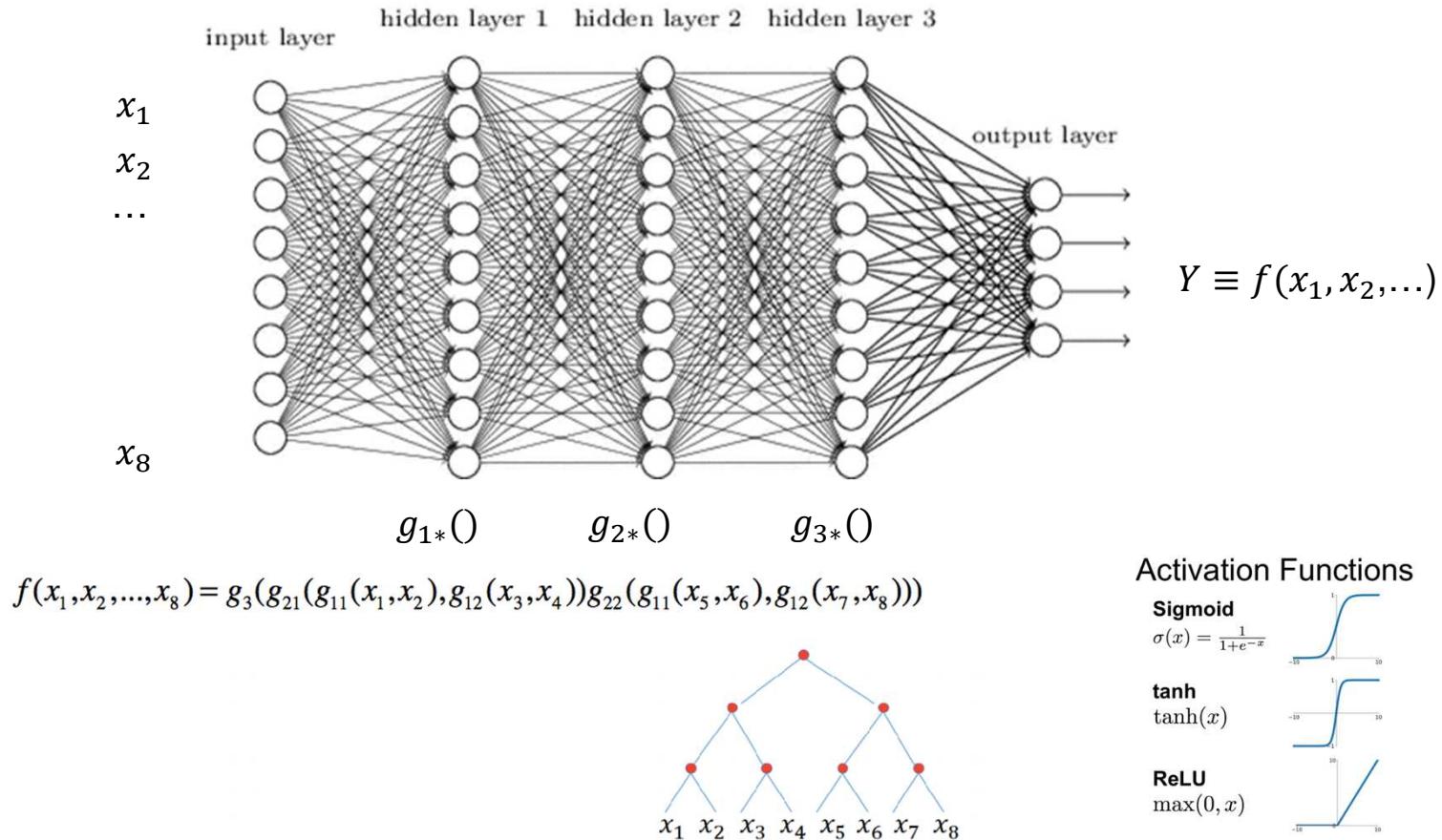
# Why this became so important NOW? How do these algorithms learn



Do they learn from / like humans?

- Both Yes, and No, and “We don’t know”
- 1. Data: “features” / variables that describe the situation
- 2. Feature Engineering: creating many new variables
- 3. Indirect (non-linear) functions / relationships / “representations”
  - Regression:  $Y = f(X) = a + b * X$
  - Modern ML:  $Y = \text{crazy complicated function (of functions (of other functions (of many-many Xs)))}$

## Deep neural network



### Activation Functions

#### Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

#### tanh

$$\tanh(x)$$

#### ReLU

$$\max(0, x)$$

Why is this so powerful? → Modern Machine Learning methods can “learn” (find) functions that cannot be expressed / explained with simple rules.

# Why this became so important NOW?

## How do these algorithms learn



Do they learn from / like humans?

- Both Yes, and No, and “We don’t know”
- 1. Data: “features” / variables that describe the situation
- 2. Feature Engineering: creating many new variables
- 3. Indirect (non-linear) relationships / “representations”
  - Regression:  $Y = f(X) = a + b * X$
  - Modern ML:  $Y = \text{crazy complicated function (of functions (of functions)) of many-many } Xs$
- 4. Complexity control: not letting the ML overfit (“learn” what’s in the data it knows, but may not generalize beyond)
  - Cross-fold validation, train-test-holdout, regularizations

# From the “fathers” of Deep Learning



## **ImageNet classification with deep convolutional neural networks**

Authors: [Alex Krizhevsky](#) [University of Toronto](#)  
[Ilya Sutskever](#) [University of Toronto](#)  
[Geoffrey E. Hinton](#) [University of Toronto](#)

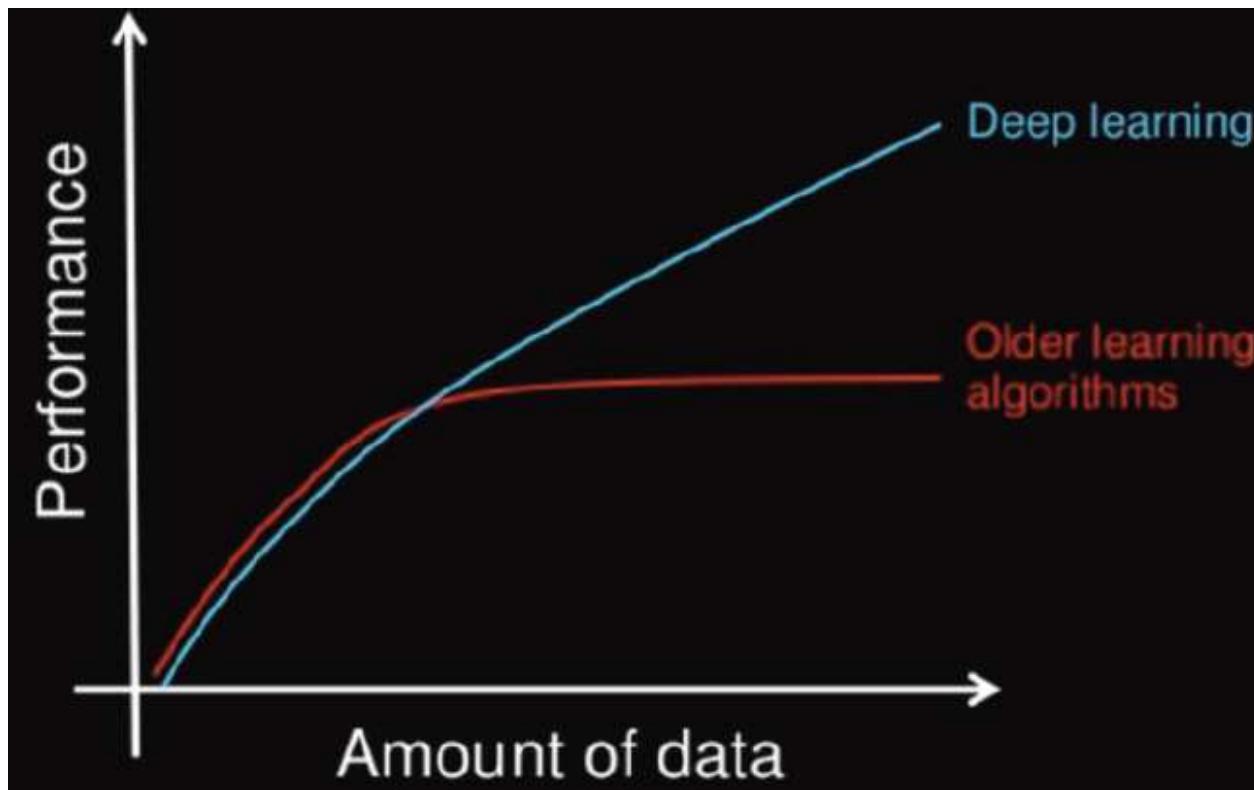
Published in:

- Proceeding  
NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1  
Pages 1097-1105

---

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called “dropout” that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

Note: DL is not best for all use-cases  
In many business applications other methods are better



These algorithms are  
“workhorse” for many  
firms:

- RandomForest
- Gradient Boosted  
Trees (xgboost)
- Support Vector  
Machines
- Regularized  
regressions  
(LASSO)

We will study  
all of these

# Sooo... Course objectives



- We are here because Machine Learning and Data Science are “the next big thing” in business and society
- Course objective: “To build your capability in data science so that you can effectively add value through the intelligent management and use of data in your organizations.”
- Three key elements: analytics techniques, business applications, basic coding/programming (in R).
- Three classes of techniques / “modules” of the course:
  - “Predicting quantities” (stats/ML, building on UDJ): Regression → Sessions 01-02, Time-series → Sessions 03-04
  - “Predicting events” (mostly ML): Classification → Sessions 05-06-07-08
  - Clustering and dimensionality reduction → Sessions 09-10

supervised learning

unsupervised learning

# Course work



- **Workshop format:**
  - Follow-along provided materials: ask questions, interruptions welcome!
  - Have you installed R and R-studio?
- **Three Case assignments [group]**
  - Quantities/Timeseries (Yahoo), Classification (Credit), Clustering (Boats)
- **Group projects**
  - Come up with a potential data science/analytics project
  - Ideally from your past or future workplace
  - If in trouble, check open datasets: Kaggle, Forbes
    - <https://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/>
  - Perform the analyses, draw the insights, prepare a 15-min presentation
  - All groups will present during classes 13-14

# What is Kaggle?



The screenshot shows the Kaggle homepage with a dark background. At the top, there is a navigation bar with the "kaggle" logo, a search bar, and links for Competitions, Datasets, Kernels, Discussion, Learn, and Sign In. Below the navigation bar, the text "The Home of Data Science & Machine Learning" is displayed in large white font, followed by the subtitle "Kaggle helps you learn, work, and play". To the right of the text, there is an illustration of an astronaut in space with a "jobs board" icon above them. A central image titled "What's Kaggle?" shows various icons related to data science and machine learning, such as a camera, a laptop, and social media symbols. At the bottom, there are two blue buttons: "Create an account" and "Host a competition", separated by the word "or".

# What is Kaggle?



- Kaggle is the world's largest community of data scientists, a two-sided platform

**Kaggle hosts (firms/organizations) post public competitions [with prizes, "Netflix challenge"]**

The screenshot shows a grid of competition cards. One card for the Home Depot Product Search competition is circled in red. Other visible cards include State Farm (Distracted Driver), Santander (Customer), BNP Paribas Cardif (Claims), and a digit recognizer challenge. Logos for GE, MasterCard, Allstate, Merck, Amazon, Facebook, Microsoft, and Neiman Marcus are at the bottom.

**Featured Competitions** [View All »](#)

MACHINE LEARNING CHALLENGES FOR EDUCATION, RESEARCH, AND INDUSTRY.

**State Farm**  
Distracted Driver  
\$65,000  
Can computer vision spot distracted drivers?

**Santander**  
Customer  
\$60,000  
Which customers are happy customers?

**HOME DEPOT**  
Product Search  
\$40,000  
Predict the relevance of search results on homedepot.com

**BNP PARIBAS CARDIF**  
Claims  
\$30,000  
Can you accelerate BNP Paribas Cardif's claims management process?

**Digit Recognizer**  
Classify handwritten digits using the famous MNIST data

**GE** **MasterCard** **Allstate** **MERCK** **amazon** **facebook** **Microsoft** **Neiman Marcus**

**Kagglers (individuals) compete and earn credits/medals**

The screenshot shows a LinkedIn profile for Rohit Gupta, a Product Manager at The Home Depot. It includes a photo, a summary, and a kaggle.com link. Below is a detailed competition history section.

**Rohit Gupta**  
Product Manager at The Home Depot  
Atlanta, Georgia | Computer Software

Current: The Home Depot  
Previous: The Home Depot, Hubcasa, McAfee (Solidcore)  
Education: University of Virginia Darden School of Business

**kaggle** Host Competitions Datasets Scripts Jobs Community Sign up Login

**RG** Verified account

**KAGGLER** **Highest!** **1001st** **Current!** **1099th** /515,996  
6,637.8 points Joined 3 years ago Ranking method changed 13 May 2015 (?)

Profile Results Scripts Forum

TOP 10% **160th/3303** TOP 10% **120th/1463** TOP 25% **211th/1326** TOP 25% **428th/2226** TOP 25% **123rd/587** **international airports** **amazon** **yelp** **13**

# Course materials



- Main resource: course portal: **LINKS TBA**
- No mandatory textbook; optional textbooks:
  - “Data Science for Business” (DSB)
  - “Forecasting Principles and Practice” (FPP), <https://www.otexts.org/book/fpp2>
  - “An Introduction to Statistical Learning” (ISL), <http://www-bcf.usc.edu/~gareth/ISL/>
  - Numerous online courses (datacamp, udemy, coursera, etc.)
- Multiple supplementary articles from Science, HBR, MGI, etc. **on portal** → Class Descriptions

# About your professor(s)



## **Anton Ovchinnikov (AO)**

Visiting Professor of Technology,  
Operations and Decision Sciences

Distinguished Professor of  
Management Analytics at Queen's  
University, Canada

[anton.ovchinnikov@insead.edu](mailto:anton.ovchinnikov@insead.edu)

Office: PMLS0.11 (FLB), TBA (SGP)

Grew up in Russia/Siberia, co-owned  
a design business before academia

Love sailing/windsurfing and skiing



## **Spyros Zoumpoulis (SZ)**

Assistant Professor of Decision  
Sciences

[spyros.zoumpoulis@insead.edu](mailto:spyros.zoumpoulis@insead.edu)

Office: PMLS0.11 (FLB), TBA (SGP)

Spyros teaches sessions 7-12



## **Theos Evgeniou (TE)**

Professor of Decision Sciences and  
Technology Management

[theodoros.evgeniou@insead.edu](mailto:theodoros.evgeniou@insead.edu)

Office: PMLS0.08 (FLB), TBA (SGP)

Theos teaches sessions 7-12

# Questions?



- On Course Objectives?
- Structure/schedule?
- Deliverables?
- Grading?

**Linear regression “recall” from UDJ**

**“Sarah Gets a Diamond” case**

# “Sarah Gets a Diamond” case



- Excel file “0102 Excel Data -- Sarah Gets a Diamond.xls” contains data about various attributes for >9000 diamonds. For the first 6000 you also have the price. The goal is to predict the prices of the remaining diamonds (i.e., those with IDs 6001 and above).
- Next – mini-competition:
  - **Use tools you’ve learned in UDJ to predict prices for IDs 6001 & above**
  - Groups of ~6 [from the **course admin**]
  - BORs (SW 1 to 12)
  - Create a spreadsheet containing your group name in cell A1 (be creative), your names in cells A2 to A7, and your forecasts of the prices of the diamonds with ID 6001 and above in cells A8 to A3149
- Email these spreadsheets to me ([anton.ovchinnikov@insead.edu](mailto:anton.ovchinnikov@insead.edu)) and the **course admin** by the beginning of the break (**TBA**); we will resume with the competition results after the break

# Competition mechanics



- **How will I determine which team did best/won?**
  - Training vs Prediction/Validation data
  - Metrics:
    - Mean Abolute Percentage Error (MAPE)
    - (Root) Mean Squared Error (MSE, RMSE)
- **Does any team know how well they did?**
  - This is a fundamentally important question for model building
  - The goal is not to build a model that would predict the existing data best (i.e., have the highest  $r^2$ ) but one that would predict well the data it has not seen yet
  - The notion of “Holdout sample” and its importance (TBD soon)

# Team names!



- “Regress my heart”
- “Every kiss begins with Jay”
- “Sarah gets a Diamond makes Spencer Nervous”
- “Gonna-Forward-This-To-My-Boyfriend-NOW“

TEAM	First men MAPE	TEAM	First men MAPE
1 Team Jetson	Nick Czuro 4.72%	70 Team Aspiring-Minds	Ashish Gup 10.68%
2 ERJ	JE&his tear 5.36%	71 LMM	10.83%
3 Team HERTz Analytics	Hasan Too 5.52%	72 I heard it was a CZ	10.84%
4 Team MDP2	Adam Liu 5.67%	73 Breakfast at Tiffany's	10.91%
5 Team GOAT (Greatest Team of All Time)	Kasturi Kur 5.70%	74 AubergineAnalytics 🍆	Alexander J 10.92%
6 DefaultH2O	SP&AO 6.07%	75 Team Waitlist	Fernando F 10.93%
7 Team YAY!	Steven Che 6.51%	76 Rohan's Five	Sehrish Sat 11.06%
8 Team Rita	Rita Wu 6.79%	77 The Dummy Variables	Ishan Kawk 11.44%
9 The One Hit Wonders		78 Gradient	11.62%
10 Team ABB		79 ffn(input your group name)	Ivan Gn 11.78%
11 Team "Gonna-Forward-This-To-My-Boyfriend-NOW"	Susan War 7.43%	80 Just the Two of Us, We can Make it if we Try	Tovey, Jos 11.85%
12 Team Sarah gets a Diamond makes Spencer Nervous		81 UNPREDICTABLEZ	Hill Pruskar 11.87%
13 The Dummy Variables		82 Team Name: Black Hole	Jay Bhatna 11.87%
14 We've Regressed		83 Team Y-Acating	Yeganeh G 12.82%
15 The Nostradamil of Diamonds		84 VWZ	13.55%
16 Fly Like A G6		85 Team Chloe	Chloe Chi 14.83%
17 Yes, I do.		86 MakeISPRedundantAgain	Bery Prana 15.38%
18 Two girls, one guy, and a lot of rocks		87 Wildcats	Nambaya C 17.53%
19 Juri dreams of Vodka	Jan Hendril 7.60%	88 Midnight lamps	20.07%
20 Sarah's Answer		89 Group 8 - AA	Giuditta Sa 20.07%
21 Real Gems		90 Naive single variable Log-Log	AO&CL 20.07%
22 Team Night Owl		91 BloodDiamond Regressors	20.07%
23 The Blood Diamonds		92 Luso-Mandarin	Xiaoyu 22.40%
24 Every Kiss Begins With Jay		93 Diamondsters	Bernardo A 22.56%
25 This is why you get cubic zirconia		94 Russ-ians	Darya Ber 24.11%
26 Team: Diamonds in the Rough		95 Crack R	Han Songyi 25.02%
27 Sorry for Partying		96 Okay-Lah	Chris Leuni 26.23%
28 World's Tiniest Handcuffs		97 Error 505	Shapour Sa 26.37%
29 Team Members	Rain Huang 7.62%	98 Pink Panther	TOREN Roy 26.50%
30 The Blood Diamonds		99 Shallow Un-Learning	Stefano Ta 27.44%
31 Team Splendites		100 Group R	Eszter Sant 27.94%
32 The HPs		101 Team Starkk	Sheng Jian 27.95%
33 She'll never notice it's zirconia?		102 DataHustlers	27.95%
34 Team KissMyDiamond	Poony Fer 7.64%	103 Group 9	Antoine 27.95%
35 DLS		104 TeamOfOne	Simeen Alz 27.96%
36 ShinyShinyShiny		105 Diamonds are forever	Dimitar Toi 28.00%
37 Team Victory	Parikshit Ci 7.77%	106 FiveGuys+1	Edwar 28.00%
38 The Stoners		107 Lucy in the Sky	Ishika Saha 28.05%
39 Vivek is a Slacker		108 Rinseed	Florian 28.41%
40 M-Cubed		109 BrazilianBus	Leonardo C 28.96%
41 Team Eglington	Jaime Lem 7.78%	110 Geeky Durians	Camille Am 29.00%
42 DS Wiz.	Rachit Agar 7.78%	111 G7	Gokalp 29.08%
43 TeamDataMaffia	Maxime 7.78%	112 Outliers	Tabriz Aliyy 29.12%
44 Gods of Numbers	Nadezhda 7.78%	113 G10	Archit 29.15%
45 Dummy	Patricia Du 7.79%	114 The Undatables	Dianne Gor 29.23%
46 GUS JOHNSON: ALL-AMERICAN HERO		115 Data Dumpsters	John Davie 29.42%
47 Darden Dummies		116 Uh-Oh Oreos	Felipe Arav 29.43%
48 Team Tornado Watch		117 Elon's Angels	29.62%
49 "Where is Nik?"	Gregory Fa 8.13%	118 like a q6	paul 29.99%
50 Diamonds Are Overpriced (DAO)		119 CP3	40.14%
51 Cubic Zirconia		120 Team Leonard	Sean O'Mai 45.66%
52 Team Individual	Gary Dhaliv 8.28%	121 Team Supermeura	Kenneth Br 48.13%
53 The Foxfielders		122 3L's Diamond Trade	55.88%
54 Bling, Bling		123 DAO MG	68.75%
55 PRICING FOR THE WIN	FINJA BOC 8.46%	124 Group 1	Clement Va 86.80%
56 Data Geeks	Kita Walser 8.47%	125 Machine Unlearning	Chi Sheng 90.28%
57 Kohinoor	Sachin Aga 8.59%	126 Team Guess	Joy Wang 91.67%
58 The Antwerp 4	Dikaios, Jo 8.59%	127 Team No Regresssts	Mitchell Po 96.90%
59 Drink Milk & Kick Ass	Lin Lin 8.59%	128 Swagatha Christie	Ross MacD 103.62%
60 Data Sharks	Nancy Srv 8.59%	129 Fantastic 5	Isabelle Pe 106.18%
61 Prestige Worldwide		130 Group 8	Mi Zhou 116.37%
62 SteakSauce	Shi Xian 8.61%	131 3 Girls, 3 Guys, No Diamonds	Eva Perrett 118.19%
63 Regress My Heart		132 Hypothetically Null	Flora Xu 120.56%
64 The Doug		133 Team Lone Solider	Mohib Rab 228.65%
65 Team Winning	Paras Gupta 8.70%	134 randomer	Cody Li 313.90%
66 TheSeven	Jan Focked 9.61%	135 Can Log La	David Gala 830.47%
67 Diamond District		136 DATA DORKS	Noemi #####
68 abracadata	Andrew Ko 10.02%	137 Diamond Hunters	Mariana Le #VALUE!
69 Diamond Data Miners	10.41%	138 G03 - Cube Datos	Nathalie Kc #VALUE!

300 Teams, 1000+ Members, 1000+ Solutions

# Next: going beyond Excel



- Intro to data analyses in R
  - What is R?
  - Why R?
  - “How to” R?: By following a simple code for the Sarah’s case you just did in Excel
- Session 0304: time series models in R
  - How? Again, by following a simple code that I will provide
- Assignment 1, Yahoo/Tumblr case: combining Excel and R for startup valuation.
  - How? By modifying the code from sessions 01-02 and 03-04
- **Before that, however:** data visualization in Tableau

# Why data visualization?

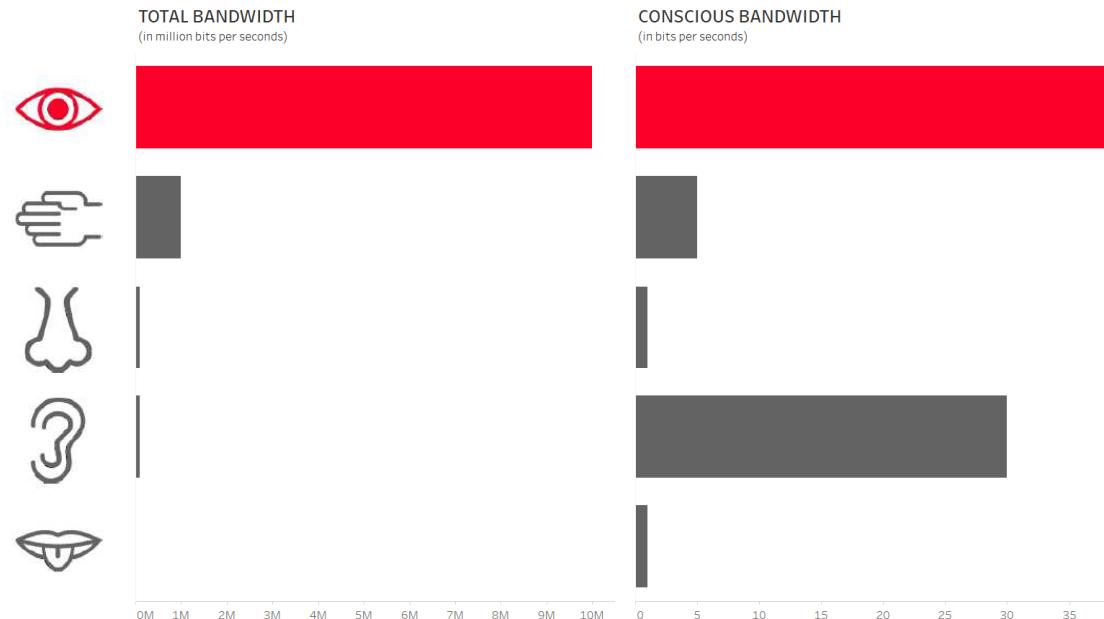


- **Q:** “What is really scarce in the Internet economy?”
- **A:** “Attention”

2009, Hal Varian [chief economist, Google; Professor, UC Berkeley]

- “A wealth of information creates a poverty of attention.”
  - 1978, Herb Simon [Nobel Prize in economics]
- “The best way to capture imagination is to speak to the eyes”
  - 1780, William Playfair [Economist, early “data visualizer”]

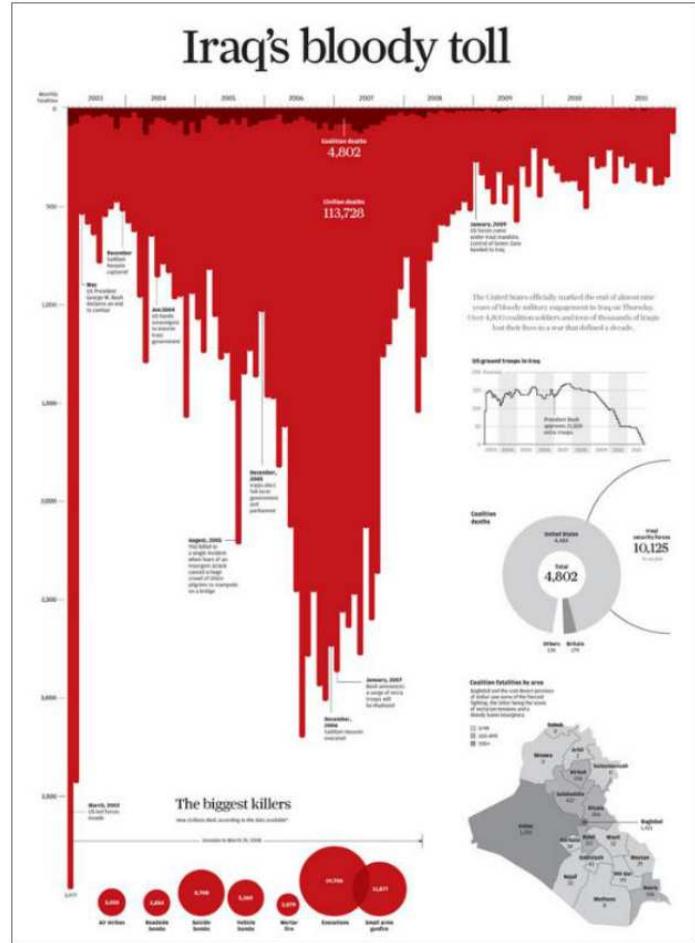
# Why data visualization?



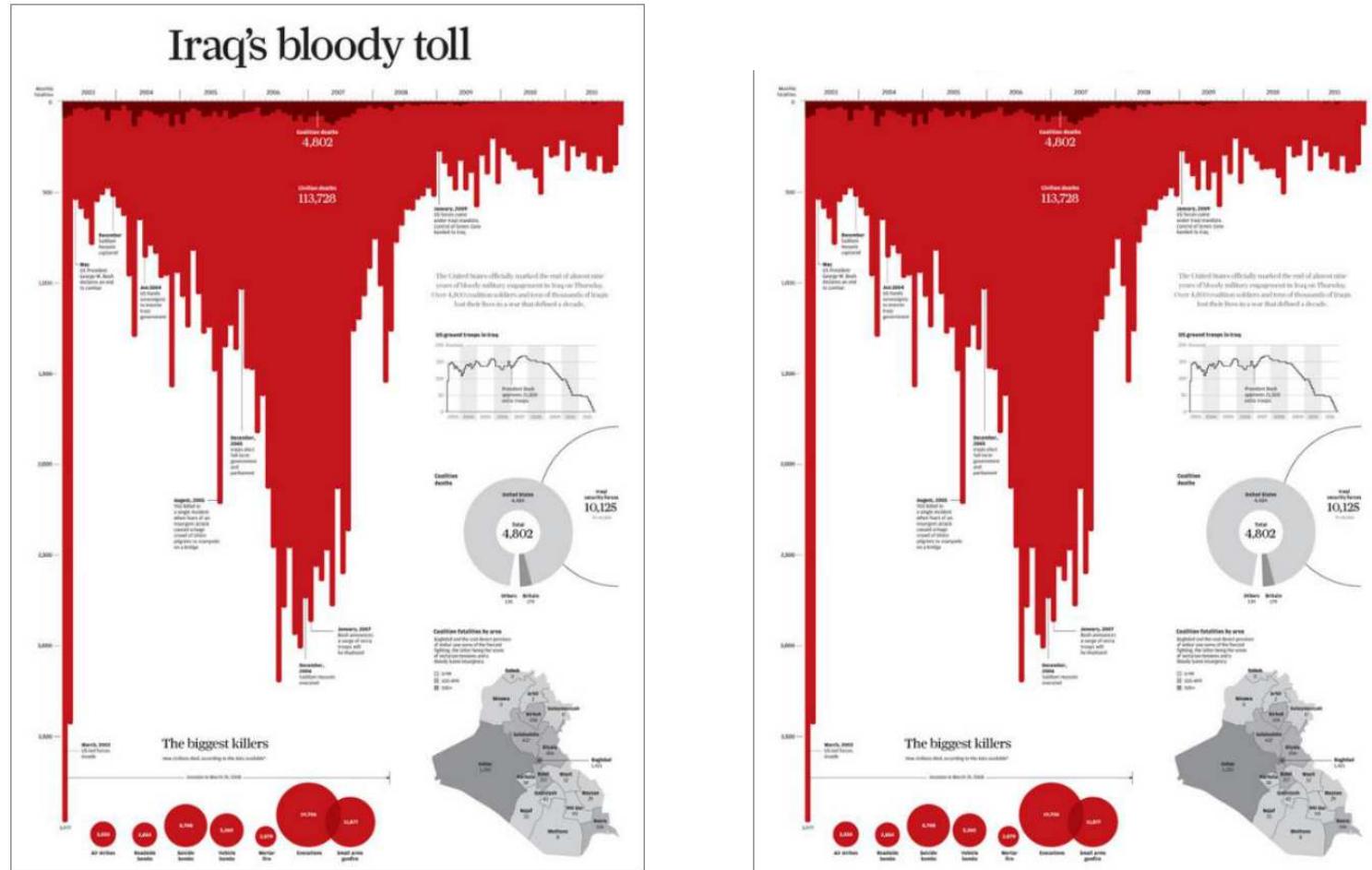
Visual cognition is fastest **and** much of it is subconscious

**“The best way to capture imagination is to speak to the eyes”**  
William Playfair [18<sup>th</sup> century Economist, early “data visualizer”]

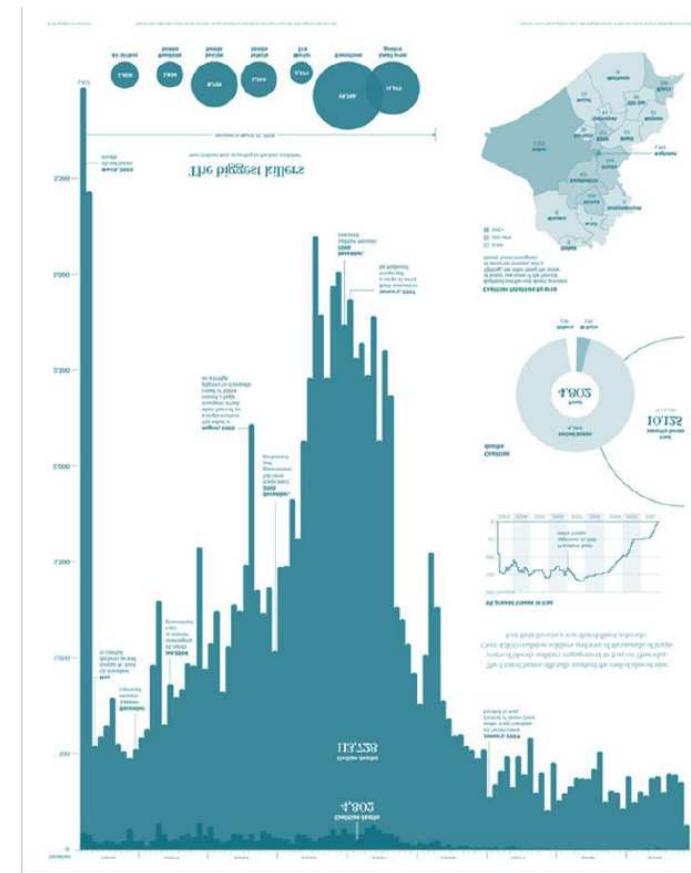
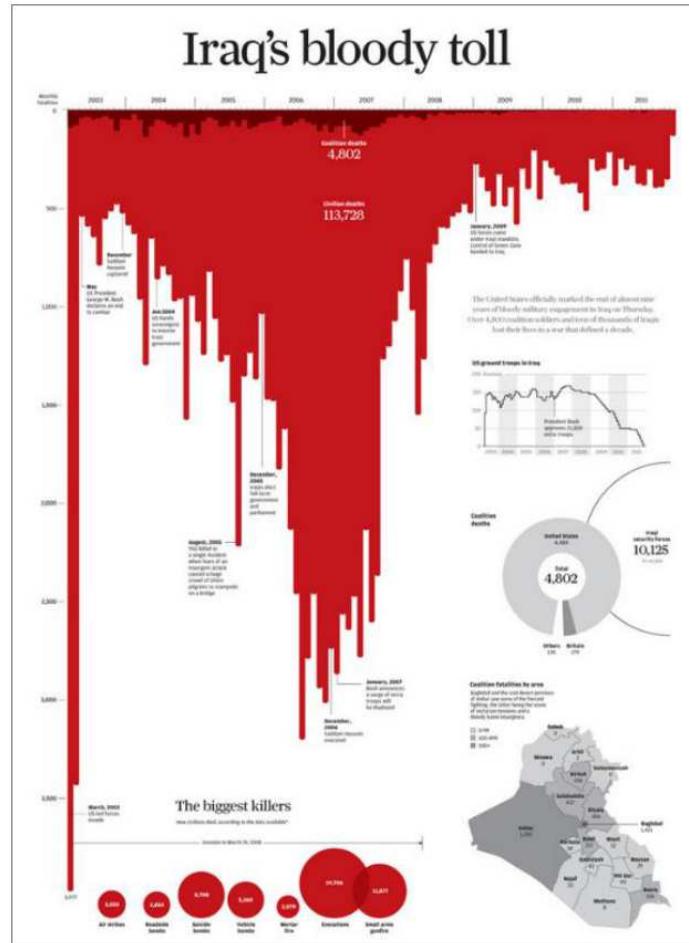
# Why data visualization?



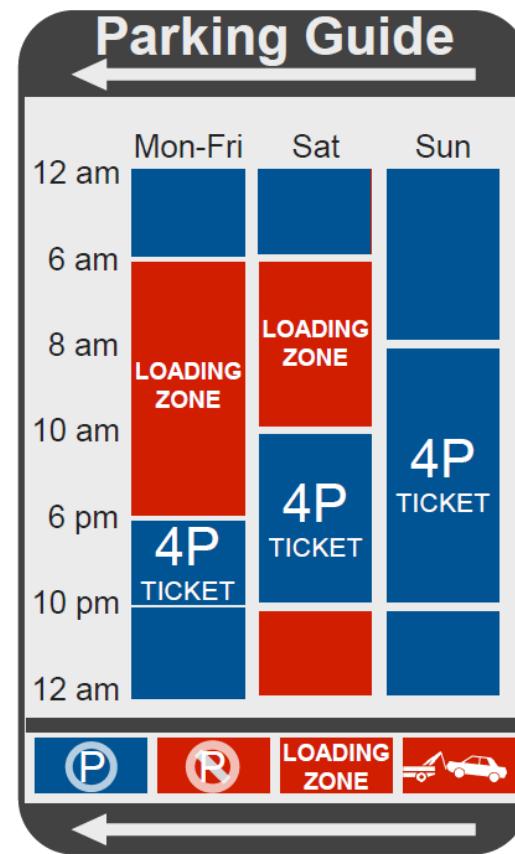
# Why data visualization?



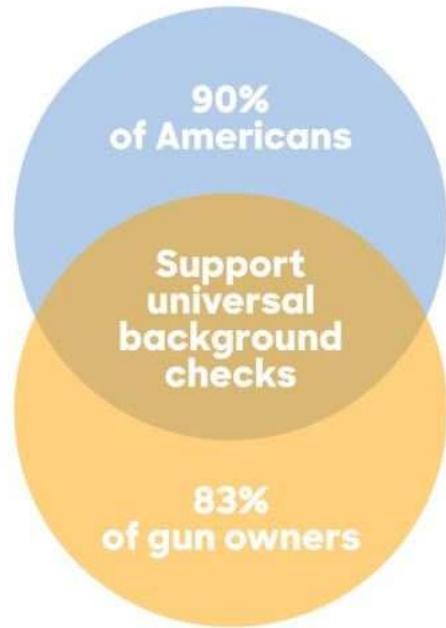
# Why data visualization?



# Why read it when you can **see** it?



# Beware of fails



 Hillary Clinton   
@HillaryClinton

 Follow

Dear Congress,

Let's get this done.

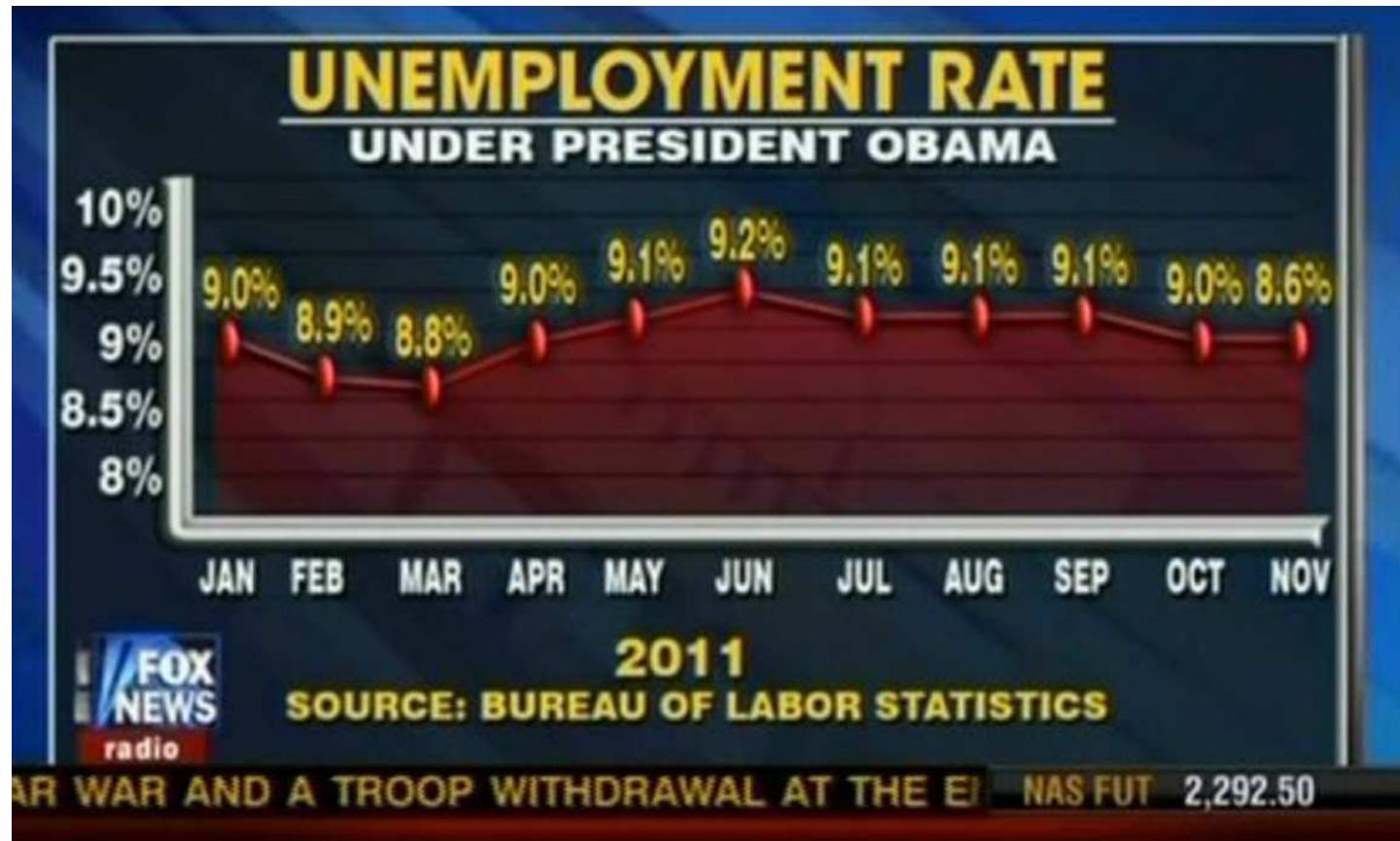
Thanks,

The vast majority of Americans

12:21 PM - 20 May 2016

 2,251  5,191

# Beware of manipulations



# Beware of manipulations



Orange F 7:28 AM 61%

Starting today, we're introducing new ways to earn more Points and new perks. Now, every night you stay counts for more.

New Silver, Gold & Diamond earn rates:

Points earned with every \$1 spent on stays:

Member	Silver	Gold	Diamond
5 Pts.	6 Pts.	9 Pts.	10 Pts.
<b>10 Pts.</b>	<b>12 Pts. (20% Bonus)</b>	<b>18 Pts. (80% Bonus)</b>	<b>20 Pts. (100% Bonus)</b>

Points earned at Tru & Home2:

here's how to earn elite status

4 stays or 10 nights	20 stays, 40 nights or 75K Pts.	30 stays, 60 nights or 120K Pts.
----------------------	---------------------------------	----------------------------------



**“It’s a common mistake to think that charts are just a fancy way of showing numbers. They’re not. They’re tools for understanding”**

Robert Kosara  
Senior Research Scientist,  
Tableau

## How many 9s?



4	7	7	5	5	2	7	4	7	1
4	9	2	5	7	7	2	6	1	7
1	7	6	9	3	4	7	5	1	2
5	1	6	3	3	8	4	8	6	6
6	5	6	4	9	3	8	9	1	9
3	8	1	5	2	2	3	6	3	9
4	6	4	5	6	3	7	7	9	1
9	1	3	3	6	1	3	3	1	8
8	1	1	8	7	5	8	1	7	4
3	6	9	2	8	9	3	7	5	7
4	4	4	2	8	2	2	9	2	8

# How many 9s?

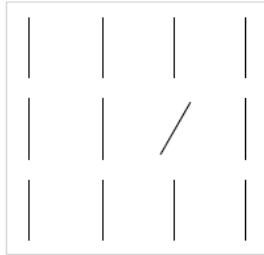


4	7	7	5	5	2	7	4	7	1
4	9	2	5	7	7	2	6	1	7
1	7	6	9	3	4	7	5	1	2
5	1	6	3	3	8	4	8	6	6
6	5	6	4	9	3	8	9	1	9
3	8	1	5	2	2	3	6	3	9
4	6	4	5	6	3	7	7	9	1
9	1	3	3	6	1	3	3	1	8
8	1	1	8	7	5	8	1	7	4
3	6	9	2	8	9	3	7	5	7
4	4	4	2	8	2	2	9	2	8

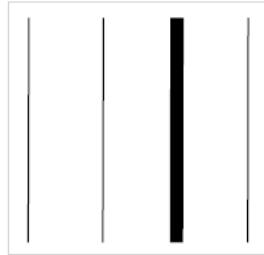
# Sub-conscious (“pre-attentive”) visual attributes



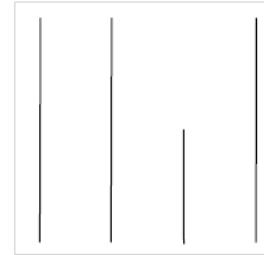
Orientation



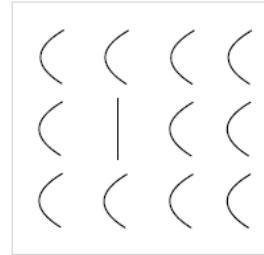
Line Width



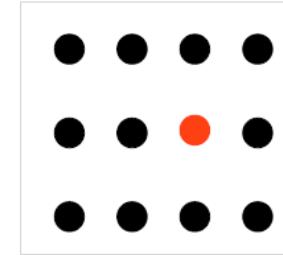
Line Length



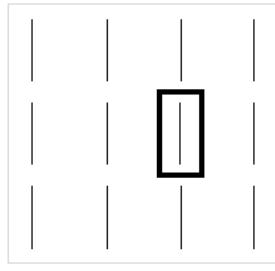
Curved/Straight



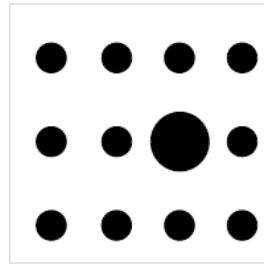
Colour/Hue



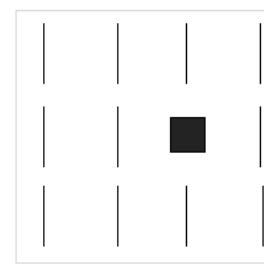
Enclosure



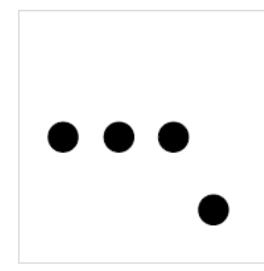
Size



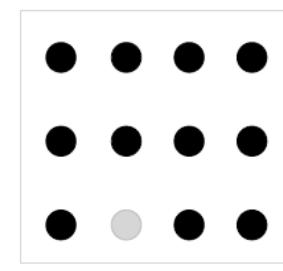
Shape



2D Position



Intensity



# Tableau: leader in data visualization



<https://www.tableau.com/>

Use your @insead.edu email to request a student product key

# Sarah “color” visual in Tableau



The screenshot shows the Tableau software interface. On the left, the 'Connect' pane is open, with 'Microsoft Excel' highlighted and circled in red. Below it are other connection options like 'Text file', 'JSON file', 'Microsoft Access', etc. The 'Open' pane in the center displays two workbooks: '0102 Sarah Gets a Diam...' and 'Tableau Workshop,\_INSE...'. The 'Discover' pane on the right provides links to training, sharing, resources, and forums. At the bottom, there's a 'VIZ OF THE WEEK' section featuring an influenza season visualization.

Tableau - Book1

File Data Server Help

Connect

Microsoft Excel

Text file

JSON file

Microsoft Access

PDF file

Spatial file

Statistical file

More...

To a Server

Tableau Server

Microsoft SQL Server

MySQL

Oracle

Amazon Redshift

More... >

Saved Data Sources

Sample - EU Superstore

Sample - Superstore

World Indicators

Open

0102 Sarah Gets a Diam...

Tableau Workshop,\_INSE...

Discover

Open a Workbook

Training

Getting Started

Connecting to Data

Visual Analytics

Understanding Tableau

More training videos...

Sharing

Learn more about ways to share

Resources

Get Tableau Prep

Blog - 4 ways the Tableau Community visualizes World Cup data

Tableau Conference

Register by 7/13 to save \$200

Forums

VIZ OF THE WEEK

Influenza Season →

More Samples

Sample Workbooks

Superstore

Regional

World Indicators

# Sarah “color” visual in Tableau



Tableau - Book1

File Data Server Window Help

Connections Add  
0102 Sarah G...Diamond data Microsoft Excel

Sheets  
 Use Data Interpreter  
Data Interpreter might be able to clean your Microsoft Excel workbook.  
Raw Data  
ST\_CaratWeight  
ST\_Clarity  
ST\_Color  
ST\_Cut  
ST\_ID  
ST\_Polish  
ST\_Preport  
ST\_Report  
ST\_Symmetry  
New Union

Raw Data (0102 Sarah Gets a Diamond data)

Raw Data

Sort fields Data source order ▾  
 Show aliases  Show hidden fields 1000 rows

Data types, Names/roles, Calculated fields

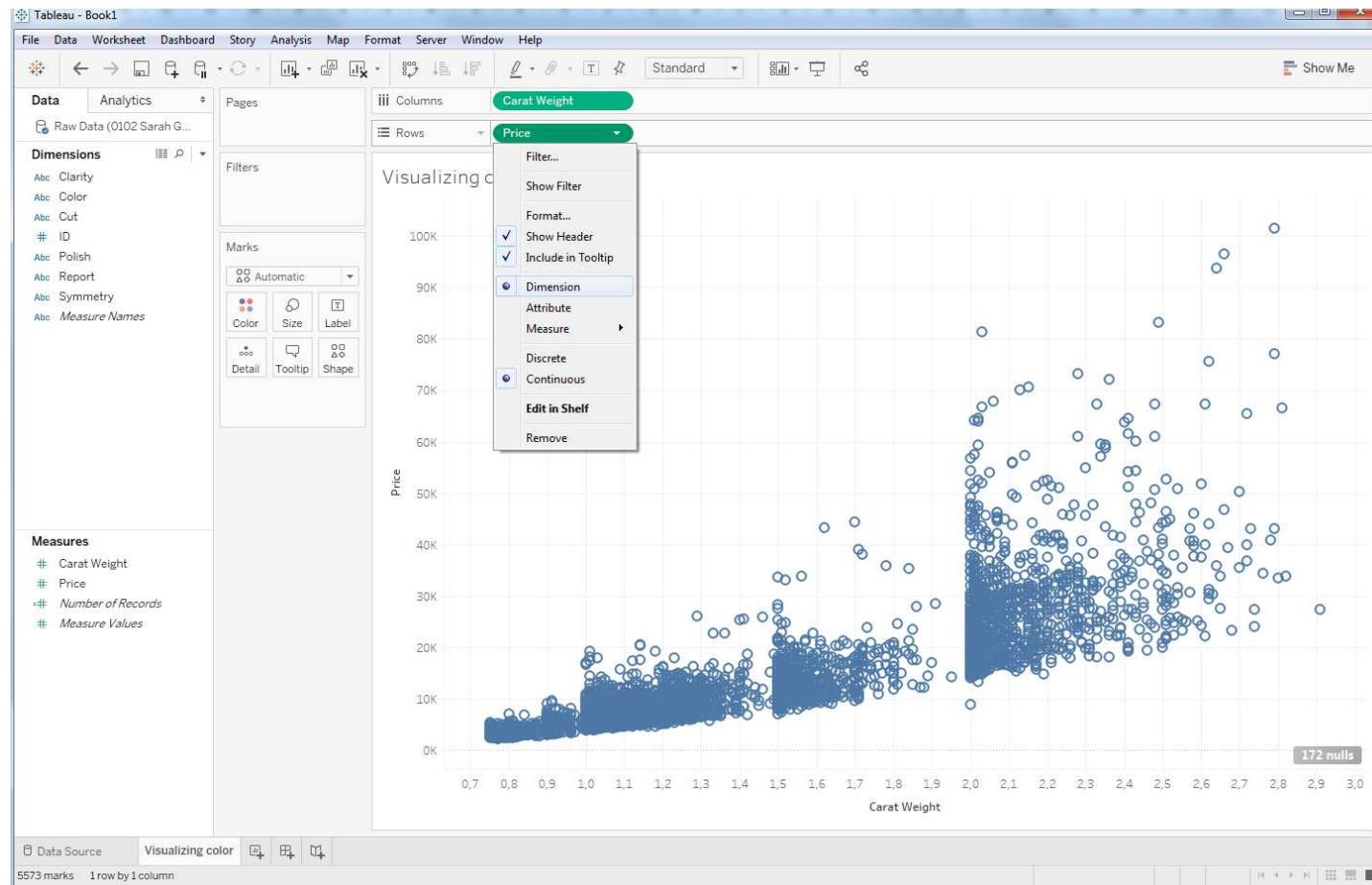
Carat Weight

	Carat Weight	Cut	Color	Clarity	Polish	Symmetry	Report	Price
1	1.10000	Ideal	H	SI1	VG	EX	GIA	5169
2	0.83000	Ideal	H	VS1	ID	ID	AGSL	3470
3	0.85000	Ideal	H	SI1	EX	EX	GIA	3183
4	0.91000	Ideal	E	SI1	VG	VG	GIA	4370
5	0.83000	Ideal	G	SI1	EX	EX	GIA	3171
6	1.53000	Ideal	E	SI1	ID	ID	AGSL	12791
7	1.00000	Very Good	D	SI1	VG	G	GIA	5747
8	1.50000	Fair	F	SI1	VG	VG	GIA	10450
9	2.11000	Ideal	H	SI1	VG	VG	GIA	18609
10	1.05000	Very Good	E	VS1	VG	G	GIA	7666
11	0.91000	Ideal	D	VS2	VG	VG	GIA	6224
12	1.01000	Good	E	SI1	G	G	GIA	5161
13	0.92000	Good	I	VS2	VG	VG	GIA	3679

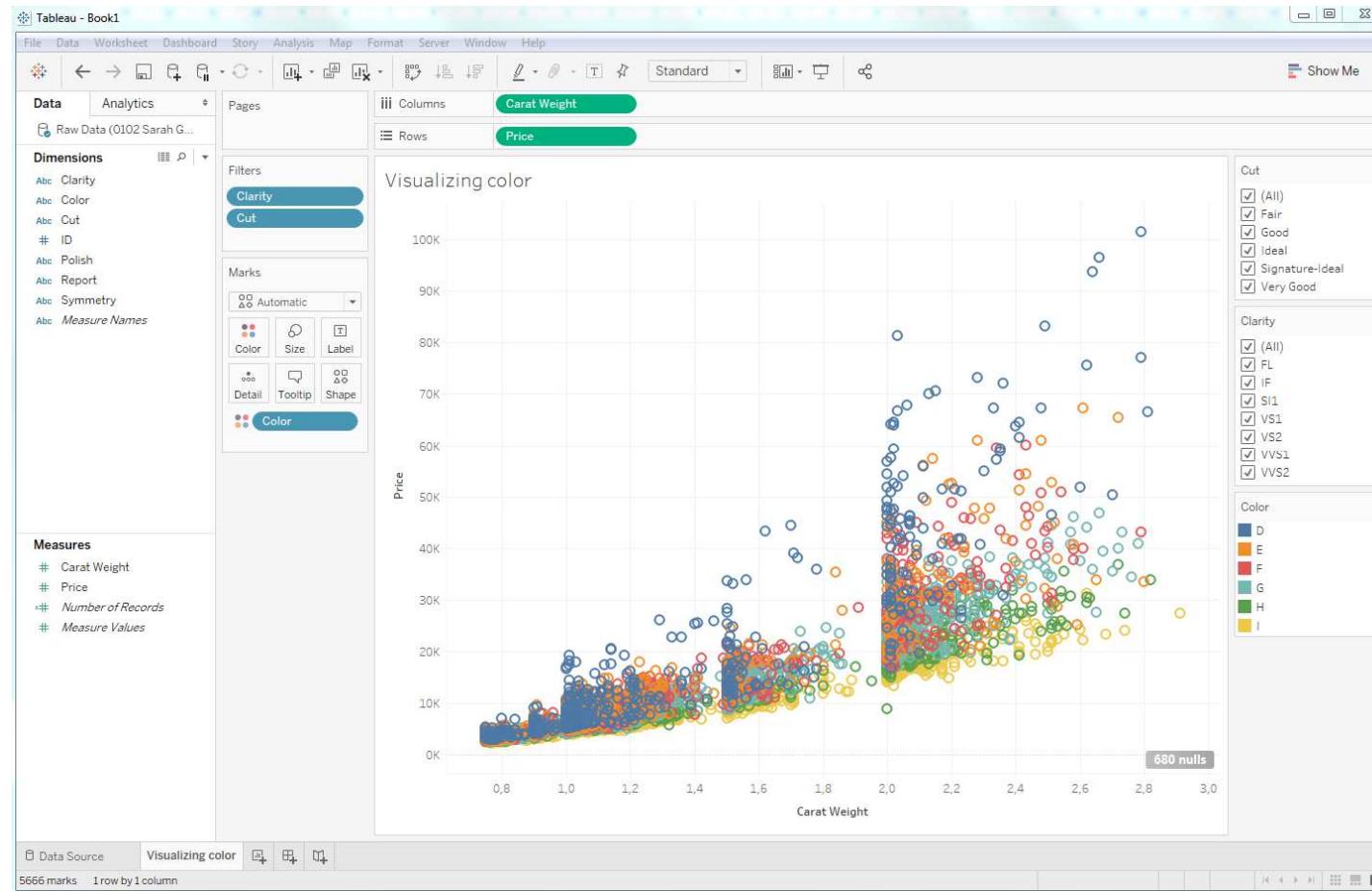
Go to Worksheet

Data Source Sheet1

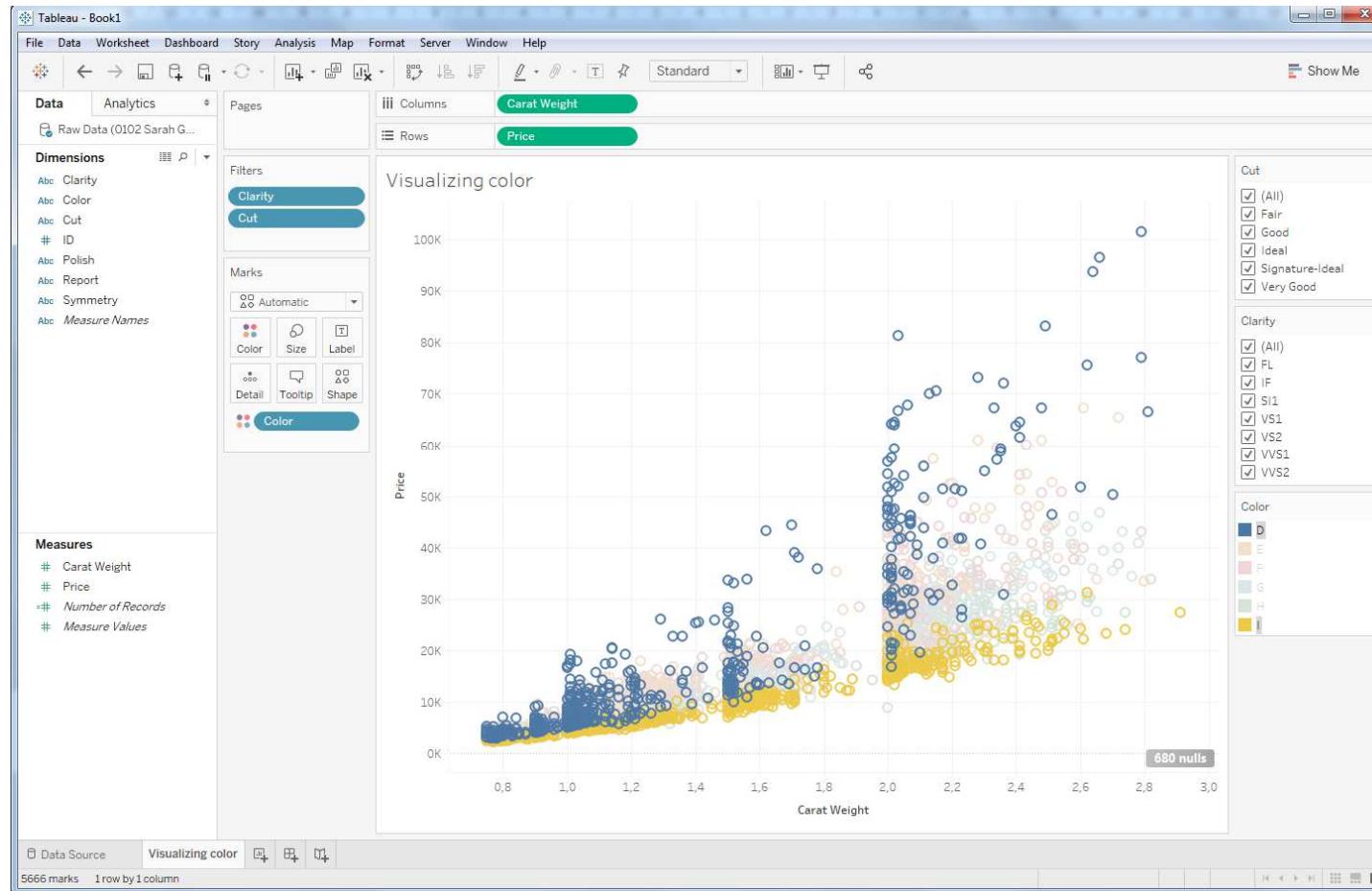
# Sarah “color” visual in Tableau



# Sarah “color” visual in Tableau



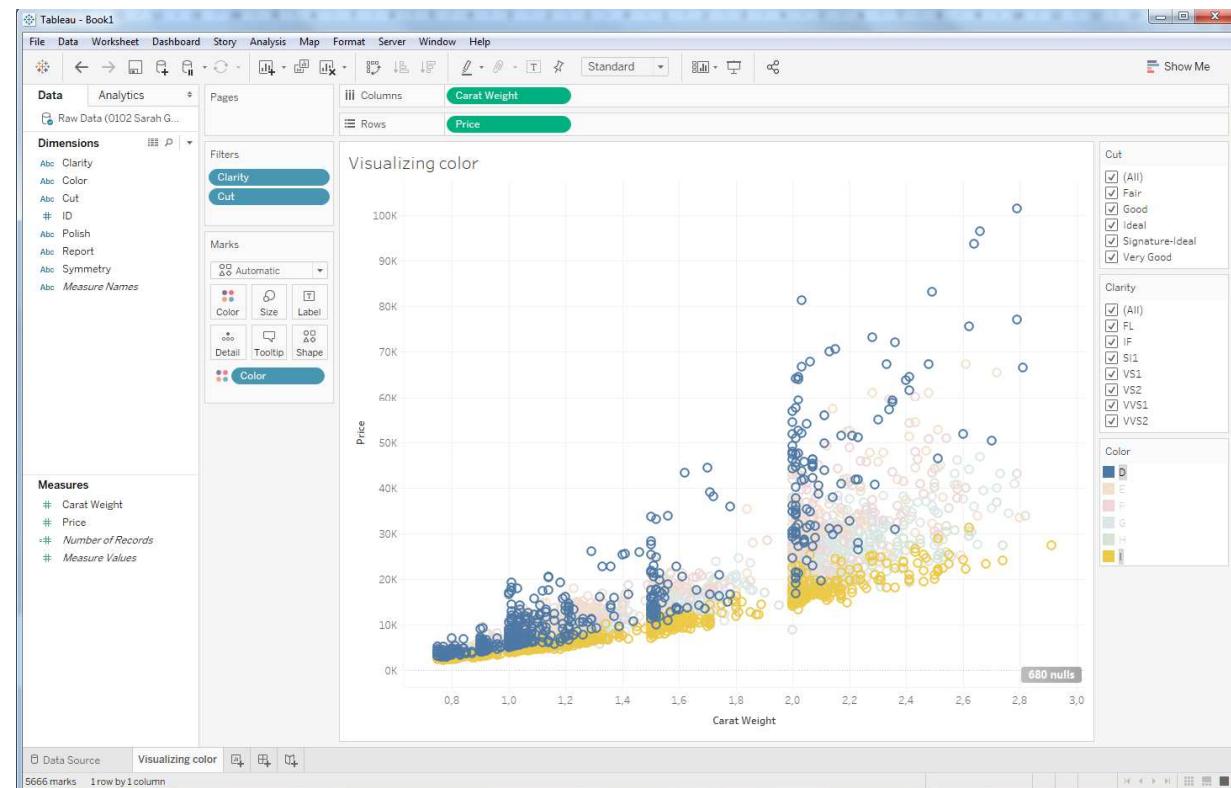
# Sarah “color” visual in Tableau



# Sarah “color” visual in Tableau What do we see? What should the “machine” “learn”?



- 1) Price increases with Carat Weight
- 2) Price for “better”-Color diamonds is higher, on average
- 3) Price increases non-linearly (faster for high Carat Weights)  
↔ There is more price dispersion for high Carat Weight diamonds
- 4) Price increases faster for “better”-Color diamonds



# Tableau on your own!



- Guide for students: <https://www.tableau.com/university-students>
- Numerous inspirational examples (with data and workbooks) at Tableau Public
  - E.g. FIFA Worldcup visuals <https://www.tableau.com/about/blog/2018/6/4-ways-tableau-community-visualizes-world-cup-data-90686>
- Dashboards – live and interactive tools to “play” / explore your data, Can be hosted online via Tableau Server
  - E.g., McKinsey Uses and Potential of AI and Analytics:  
<https://www.mckinsey.com/featured-insights/artificial-intelligence/visualizing-the-uses-and-potential-impact-of-ai-and-other-analytics>

# Next: going beyond Excel



- Intro to data analyses in R
  - What is R?
  - Why R?
  - “How to” R?: By following a simple code for the Sarah’s case you just did in Excel
- Session 0304: time series models in R
  - How? Again, by following a simple code that I will provide
- Assignment 1, Yahoo/Tumblr case: combining Excel and R for startup valuation.
  - How? By modifying the code from sessions 01-02 and 03-04
- **Before that, however:** data visualization in Tableau

# What is R?



- R is an open-source (free) programming language for statistical computing, [just one letter, “R”]
  - CRAN: R archive network that hosts “packages” (more next)
- Convenient user interface: R-Studio



[Home]

Download

CRAN

R Project

About R  
Contributors  
What's New?  
Mailing Lists  
Bug Tracking  
Conferences  
Search

## The R Project for Statistical Computing

### Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

### News

- [R version 3.2.2 \(Fire Safety\)](#) has been released on 2015-08-14.
- [The R Journal Volume 7/1](#) is available.
- [R version 3.1.3 \(Smooth Sidewalk\)](#) has been released on 2015-03-09.

The homepage of the RStudio website. It features the RStudio logo at the top left. A navigation bar includes links for rstudio::conf, Products, Resources, Pricing, About Us, Blogs, and a search icon. Below the navigation is a large banner with the text "RStudio" and "Open source and enterprise-ready professional software for R". To the right of the banner are four buttons: "Download RStudio", "Discover Shiny", "shinyapps.io Login", and "Discover RStudio Connect".



RStudio



Shiny



R Packages

RStudio makes R easier to use. It includes a code editor, debugging & visualization tools.

Shiny helps you make interactive web applications for visualizing data. Bring R data analysis to life.

Our developers create popular packages to expand the features of R. Includes ggplot2, dplyr, R Markdown & more.

- Main “competitor”: language called “python”
  - R is more for data and learning, python is more for development and coding. The two are mostly identical in their analytical capabilities (can run each-others code via interpreters)
  - <https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis>

# Why R?



- Free, open source software
- Over 10,000 open-source packages
  - Forefront of developing new methods and tools (most stats/data science research papers come with R implementations of the proposed methods)
- Engaged community of data-scientists
- Numerous excellent learning and help resources, free online “books” and inexpensive courses (datacamp, udemy, coursera – see course website)
- Learning Data Science is effectively inseparable from learning R (or python)
- “Why R” vs “Why coding?”
  - **Replicability:** anyone who has your code can replicate exactly what you’ve done. Fundamentally different from point-and-click Excel
  - **Reusability:** copy-paste code written for one project into another
  - **Algorithmic thinking:** becomes very important in man-machine world

# “How to” R?



- Data:
  - CSV format (“comma separated values”) [note: regional settings]
  - variables, vectors/matrices and dataframes
  - `data.frame` is the “mother” of all datatypes in R
- commenting out: `# your comment`
- packages and libraries
  - The standard installation comes with only basic commands; most functionality is in packages/libraries
  - Installing packages (need to do that only once) + calling libraries (need to do that every time you relaunch R); will see in Session 0304
- getting help:
  - Question mark and command opens help: e.g., `?plot`
  - Search the issue online
  - Ask TA ☺

# “How to” R? – Sarah’s case



Here is the code to create a model with MAPE ~ 7.6%

Yes, only six lines of code! Main function: `lm` (“linear model”)

```
diamond.data<-read.csv(file.choose(), header=TRUE, sep=",") #load the data into the
diamond.data dataframe
diamond.data.training<-subset(diamond.data, ID<=6000) #separate ID 1...6000 into "training"
diamond.data.prediction<-subset(diamond.data, ID>=6001) #separate ID 6001...9142 into
"prediction"
fit<-lm(Price~Carat+Weight+Cut+Color+Clarity+Polish+Symmetry+Report,
data=diamond.data.training) #run a multiple linear regression model (lm) on the training
data, call it "fit"
predicted.prices<-predict(fit, diamond.data.prediction) #use the "fit" model to predict
prices for the prediction data
write.csv2(predicted.prices, file = "Predicted Diamond Prices.csv") #export the predicted
prices into a CSV file
```

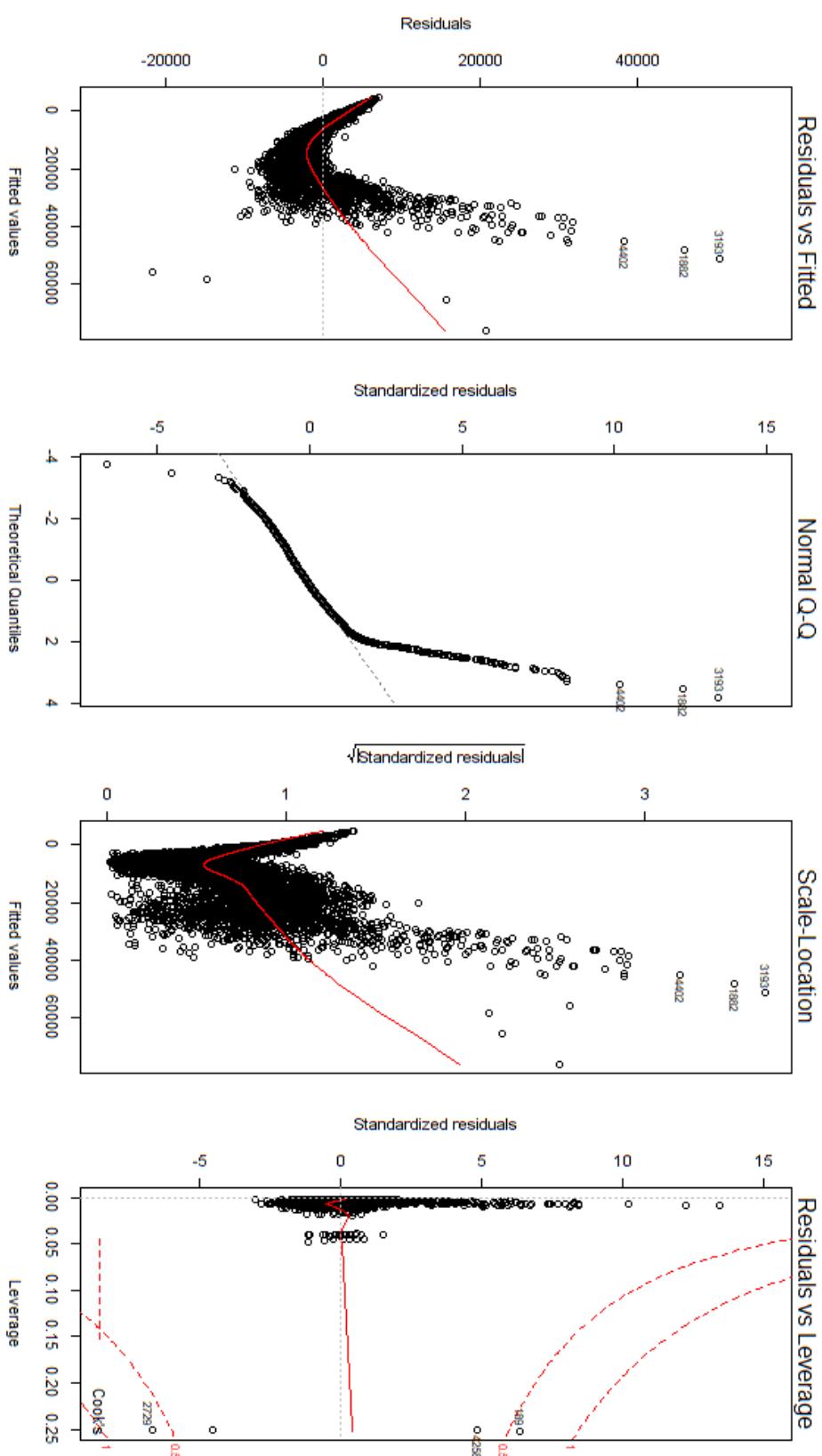
Download two files from course portal:

- CSV Data: [0102 CSV data -- Sarah Gets a Diamond.csv](#)
- R code: [0102 R Code -- Sarah Gets a Diamond -- linear regression.R](#)

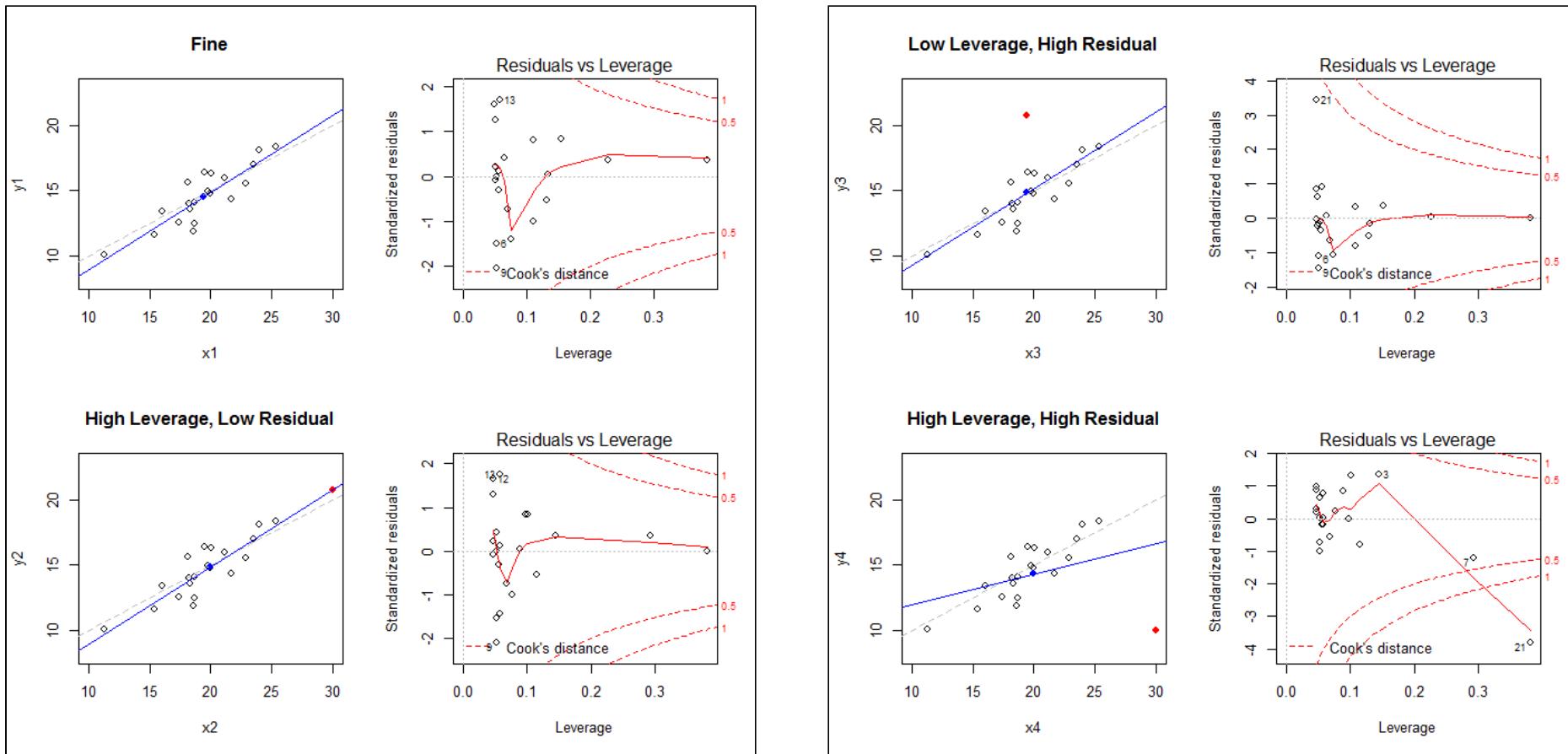
Open code in RStudio, and lets proceed from there step-by-step

# Regression Diagnostic Plots

## Homoscedasticity or Heteroscedasticity?



# Outliers: Residuals versus Leverage, “Bill Gates plot”



# Additional R commands and “ideas”



- Continuous vs categorical variables – recognized automatically (nothing special you need to do!)
- Exploring the data: `str`, `head`, `tail`, `summary`
- Basic plotting: `plot`, `hist`, `abline`, `par`
- Log-transforms:
  - `fit.log<-lm(log(Price)~log(Carat.Weight)+...`
- Assessing the quality of the model: holdout / train-test / cross-fold validation
- Interactions:
  - `fit.log.i<-lm(log(Price)~log(Carat.Weight)*Color+...`
  - [time-permitting] Advanced plotting
  - [time-permitting] Stepwise variable selection

# Cross-fold validation / train-test / holdout



- Recall: What was the task? To predict the prices of new diamonds, for which we know the features (Xs) but not the prices (Y). How can we recreate this task using the current data on 6000 diamonds?
- **Main idea:** set a portion of the data (e.g., 1000) as a “holdout” (“testing”) sample. Build a model on the remaining 5000 (“training data”) and use it to predict the prices of the holdout data
  - But we already know the prices of those 1000 diamonds(!)
  - We can thus measure the errors in predictions:
    - Estimate how well our model will perform, select the best model
    - Automatic resampling/”rotation” of holdout (`rms` package)

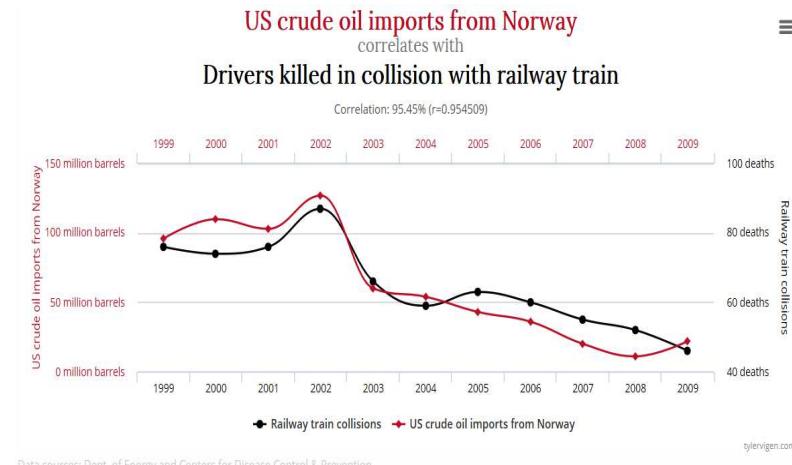
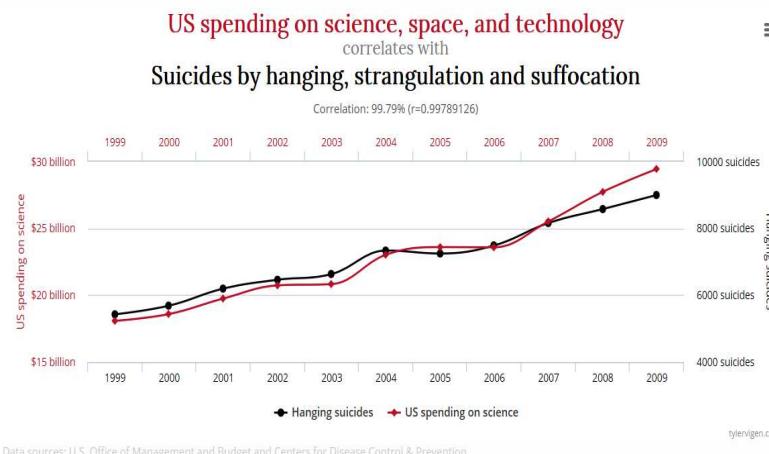
# Cross-fold validation / holdout



```
diamond.data.testing<-subset(diamond.data, (ID>=5001 & ID<=6000)) #withhold 1000  
datapoints into a "testing" data  
diamond.data.training<-subset(diamond.data, ID<=5000) #redefine the training data  
  
fit<-lm(Price~Carat.Weight+Cut+Color+Clarity+Polish+Symmetry+Report,  
data=diamond.data.training) #build a model on training data  
predicted.prices.testing<-predict(fit, diamond.data.testing) #predict the prices of  
the 1000 diamonds left for testing the model  
percent.errors <- abs((diamond.data.testing$Price-  
predicted.prices.testing)/diamond.data.testing$Price)*100 #calculate absolute  
percentage errors  
mean(percent.errors) #display Mean Absolute Percentage Error (MAPE)  
  
# repeat the same for the log model  
fit.log<-lm(log(Price)~log(Carat.Weight)+Cut+Color+Clarity+Polish+Symmetry+Report,  
data=diamond.data.training)  
predicted.prices.testing.log<-exp(predict(fit.log, diamond.data.testing))  
percent.errors.log <- abs((diamond.data.testing$Price-  
predicted.prices.testing.log)/diamond.data.testing$Price)*100  
mean(percent.errors.log)
```

# Why “train-test” is so important? Spurious correlations/overfitting

- In the world of “Big Data” it is not difficult to find variables that would correlate almost perfectly with any data (“overfitting”), without reflecting any underlying relationship - “spurious correlations”



- The best model is thus **not** one that does best on training data (has high  $r^2$  ~ overfitting), but one that does best on the new data

# Additional R commands and “ideas”

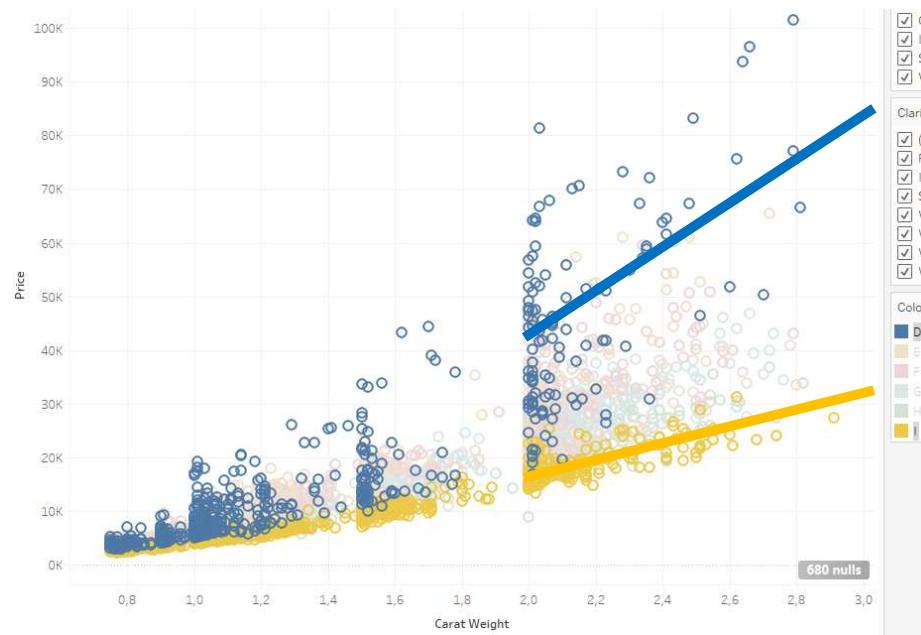


- Continuous vs categorical variables – recognized automatically (nothing special you need to do!)
- Exploring the data: `str`, `head`, `tail`, `summary`
- Basic plotting: `plot`, `hist`, `abline`, `par`
- Log-transforms:
  - `fit.log<-lm(log(Price)~log(Carat.Weight)+...`
- Assessing the quality of the model: holdout / train-test / cross-fold validation
- Interactions:
  - `fit.log.i<-lm(log(Price)~log(Carat.Weight)*Color+...`
  - [time-permitting] Advanced plotting
  - [time-permitting] Stepwise variable selection

# Interactions



- Main idea: how to capture simultaneous influence of two variables that is not additive?
  - Dummy variable: changes intercept, but not slope
  - Interaction: changes slope
  - `fit.log.i<-lm(log(Price)~log(Carat.Weight)*Color+...`



# Understanding Interactions vs Dummies

## Dummies only:



### Data:

log(Carat.Weight)	ColorE	ColorF	ColorG
0.095310180	0	0	0
-0.186329578	0	0	0
-0.162518929	0	0	0
-0.094310679	1	0	0
-0.186329578	0	0	1

### Model:

```
> summary(fit.log)
```

```
Call:
lm(formula = log(Price) ~ log(Carat.weight) + cut + Color + clarity +
    Polish + Symmetry + Report, data = diamond.data.training)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.65518 -0.06189 -0.00326  0.06122  0.55758 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 9.446e+00 5.227e-02 180.732 < 2e-16 ***
log(Carat.weight) 1.987e+00 4.053e-03 490.292 < 2e-16 ***
CutGood      3.018e-02 9.732e-03  5.156 2.60e-07 ***
CutIdeal     1.061e-01 9.540e-03 11.125 < 2e-16 ***
CutSignature-Ideal 2.518e-01 1.180e-02 21.332 < 2e-16 ***
CutVery Good 7.017e-02 9.275e-03  8.428 < 2e-16 ***
ColorE       -8.583e-02 5.381e-03 -15.951 < 2e-16 ***
ColorF       -1.000e-01 5.096e-03 -24.181 < 2e-16 ***
ColorG       -2.192e-01 4.787e-03 -45.787 < 2e-16 ***
ColorH       -3.495e-01 5.050e-03 -69.199 < 2e-16 ***
colorI       -4.944e-01 5.180e-03 -95.443 < 2e-16 ***
```

- **What did the model with dummies only learn?**

- If Color = E, then Model is:  $\log(\text{Price}) = 9.446 + 1.987 * \log(\text{C.W}) - 0.0858 * 1 = 9.3602 + 1.987 * \log(\text{C.W})$
- If Color = G, then Model is:  $\log(\text{Price}) = 9.446 + 1.987 * \log(\text{C.W}) - 0.2192 * 1 = 9.2268 + 1.987 * \log(\text{C.W})$
- **In English:** “E-color diamonds are on average more expensive than G (9.36 vs 9.22) but their price increases the same with C.W (1.987)”

# Understanding Interactions vs Dummies

## Dummies + Interactions:



**Data:** `log(Carat.Weight) * Color`

log(Carat.Weight)	ColorE	ColorF	ColorG	log(Carat.Weight):ColorE	log(Carat.Weight):ColorG
0.095310180	0	0	0	0.000000000	0.000000000
-0.186329578	0	0	0	0.000000000	0.000000000
-0.162518929	0	0	0	0.000000000	0.000000000
-0.094310679	1	0	0	-0.094310679	0.000000000
-0.186329578	0	0	1	0.000000000	-0.186329578

**Model:**

```
> summary(fit.log.i)
```

Call:

```
lm(formula = log(Price) ~ log(carat.weight) * color + cut + color +
clarity + Polish + Symmetry + Report, data = diamond.data.training)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	9.353630	0.050617	184.793	< 2e-16 ***
log(carat.weight)	2.147211	0.012553	171.045	< 2e-16 ***
colorE	-0.066750	0.006208	-10.752	< 2e-16 ***
colorF	-0.109461	0.006057	-18.104	< 2e-16 ***
colorG	-0.196335	0.005770	-34.026	< 2e-16 ***
colorH	-0.295292	0.006135	-48.130	< 2e-16 ***

log(carat.weight):colorE	-0.107736	0.017128	-6.290	3.44e-10 ***
log(carat.weight):colorF	-0.144098	0.016185	-6.389	1.83e-10 ***
log(carat.weight):colorG	-0.153511	0.015125	-10.150	< 2e-16 ***
log(carat.weight):colorH	-0.260390	0.015675	-16.612	< 2e-16 ***
log(carat.weight):colorI	-0.292555	0.016244	-18.010	< 2e-16 ***

- **What did the model with dummies + interactions learn?**

- Color = E, Model:  $\log(\text{Price}) = 9.353 + 2.14 * \log(\text{C.W}) - 0.0667 * 1 - 0.1077 * \log(\text{C.W}) = 9.286 + 2.032 * \log(\text{C.W})$
- Color = G, Model:  $\log(\text{Price}) = 9.353 + 2.14 * \log(\text{C.W}) - 0.1963 * 1 - 0.1535 * \log(\text{C.W}) = 9.1594 + 1.987 * \log(\text{C.W})$
- **In English:** “E-color diamonds are on average more expensive than G (9.28 vs 9.15) and their price increases faster with C.W (2.032 vs 1.987)”

Dummies change only the intercept. Interactions change both intercept and slope → testing error drops to 7.5%

# Summary of Sessions 01-02



- This course is about Data Science: most-demanded, highest-paid (“sexiest”) skill in the MBA job market today
- You will learn the basics of several analytical (stats and machine learning) techniques, learn the basics of coding (in R, leading open-source soft), and most importantly, learn how to use this knowledge in business applications
- Key learnings:
  - Visualizations: “speak directly to their eyes,” Tableau
  - Feature Engineering: (Dummies, Log-transform, Interactions)
  - Cross-validation (train-test) approach of machine learning [so-far, based on regression]
  - We predicted prices  $\pm \sim 7.5\%$ ! More interactions can improve to  $\sim 5.5\%$  (still with regression)
- “Excel to R”: codified solutions to analytical problems are often easier and more powerful
  - But coding is new to many of you. It’ll take practice. We’ll move slowly. Tutorials will help
  - **... and if nothing else, you’ve learned my name and office # ;)**

# Next...



- Finish group formation
- Tutorial 1:
  - Getting comfortable with R [replicating today's concepts on a new dataset]
  - Basic data manipulation in R [`dplyr` package, time-permitting]
- Sessions 3-4: **Time-series analyses**
  - Same as today, R code will be provided and we will work through it learning the underlying methods, how to code them, and how to use them



The Business School  
for the World®

Europe

|

Asia

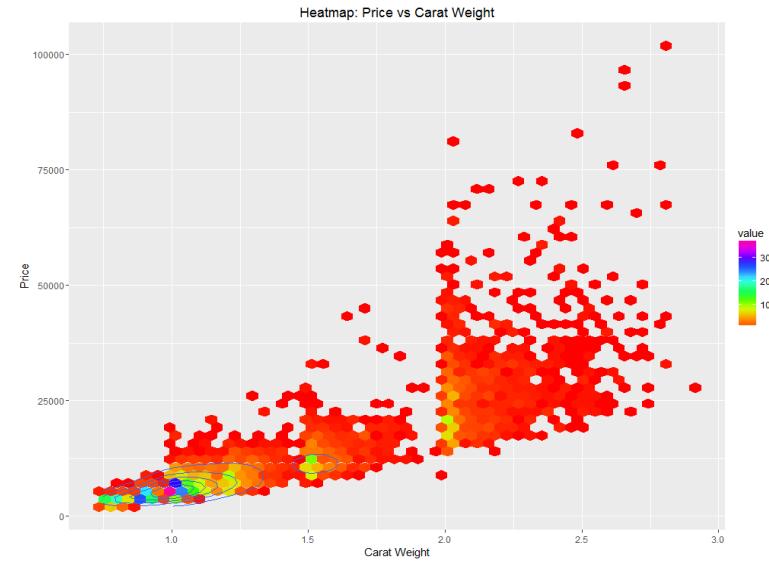
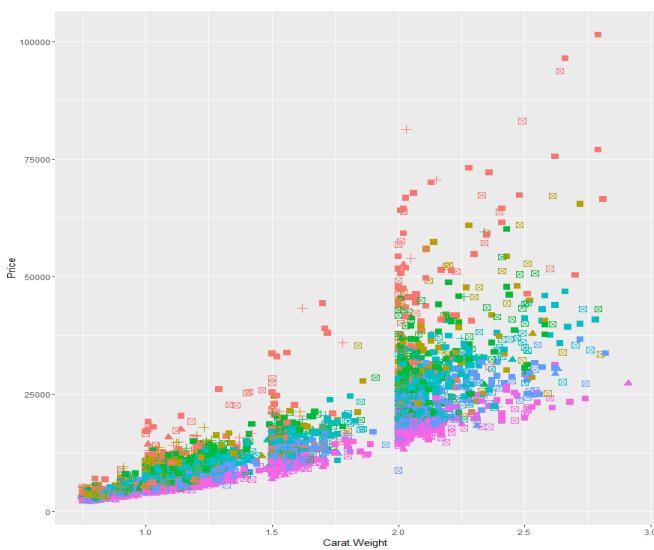
|

Middle East

# [Optional / Time Permitting]: Advanced plotting



```
#install.packages("ggplot2") # installing a package -- do this only once, the first time you plan to use it
library(ggplot2) # calling a library from the package -- do this every time you plan on using it
ggplot(diamond.data.training, aes(x=Carat.Weight, y=Price, shape=Cut, color=Color)) + geom_point(size=3)
#create a Tableau-like plot of price vs carat with color representing "color" and point shapes for cut
heatmap<-ggplot(diamond.data.training, aes(Carat.Weight, Price)) + geom_hex(bins=50)
+scale_fill_gradientn(colours = rainbow(10) )+ggtitle("Heatmap: Price vs Carat Weight") + theme(aspect.ratio = 0.8)+ labs(x = "Carat Weight", y = "Price") +geom_density2d()
print(heatmap)
```



## [Optional / Time Permitting]: Variable selection



- In the world of “Big Data” 000s of variables can be easily mined/created (e.g., cut\*color\*clarity interactions result in  $5*6*7=350$  new variables).

- How to select which variables should be in the model?

- Forward/Backward/Stepwise regressions:

```
fit.log.step<-step(lm(log(Price)~log(Carat.Weight)*Color*Cut + ... ,direction="backward")
summary(fit.log.step)
fit.log.step<-step(lm(log(Price)~log(Carat.Weight)*Color*Cut + ...,direction="both")
summary(fit.log.step)
```

- Main idea (forward example):

- Find which single X is most correlated with Y, add X1 to the model.
- Given that X1 is already in the model, find which other variable adds most explanatory power. Add X2 and re-estimate the model.
- Repeat until no variable can be added
- step function uses significance. More advanced: stepAIC ["Akaike information criterion", sessions 5-6], LASSO/Ridge – penalties for many variables [sessions 7-8]