The Business School for the World®

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

Profs. Anton Ovchinnikov, Theos Evgeniou, Spyros Zoumpoulis

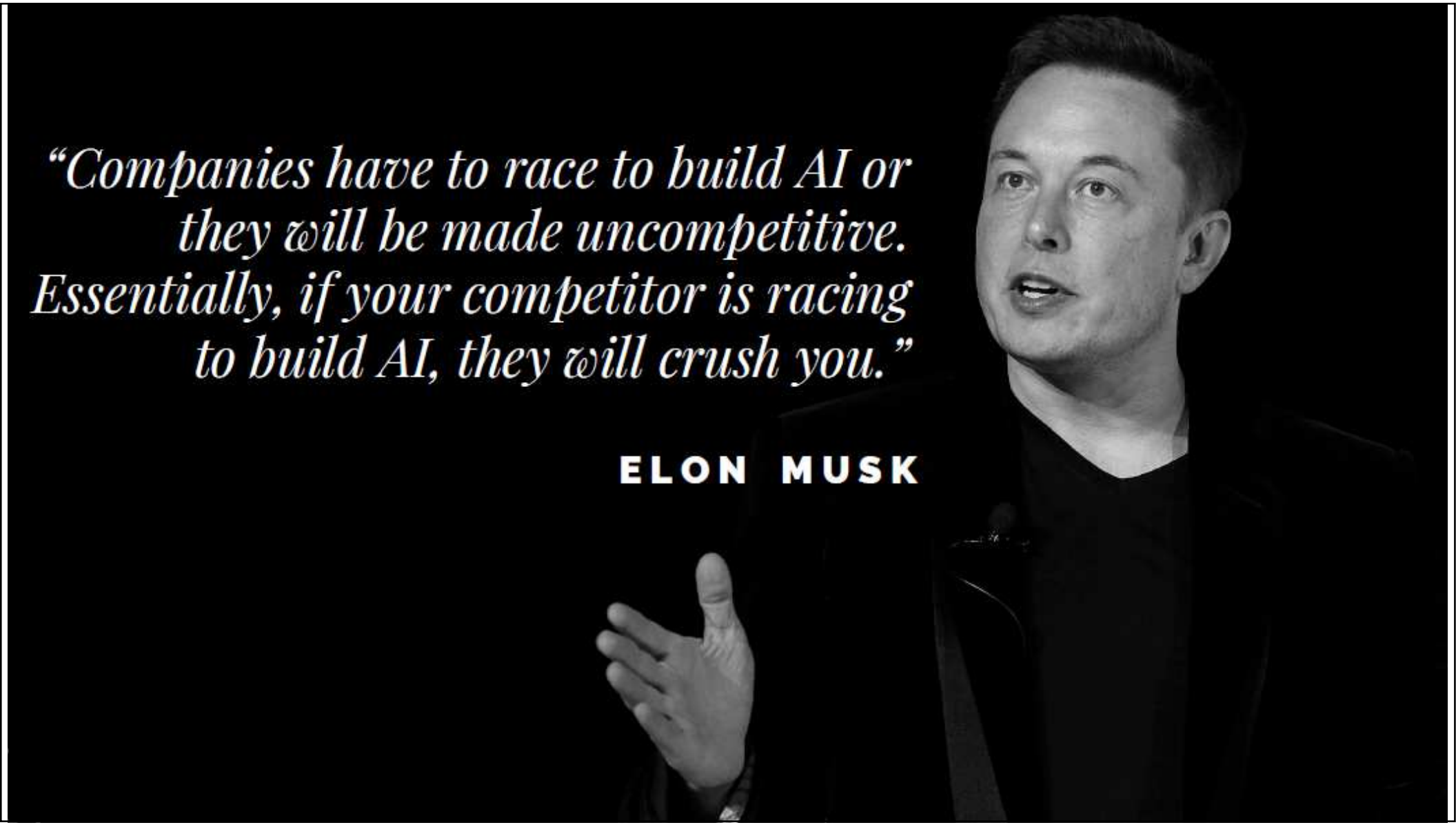
Sessions 01-02

- 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/#66bebaf0701a
- 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/1st 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/louiscolumbus/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

“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

Spotlight

ARTWORK Tamar Cohen, *Andrew J. Bubolz*
2011, silk screen on a page from a high school yearbook, 8.5" x 11"

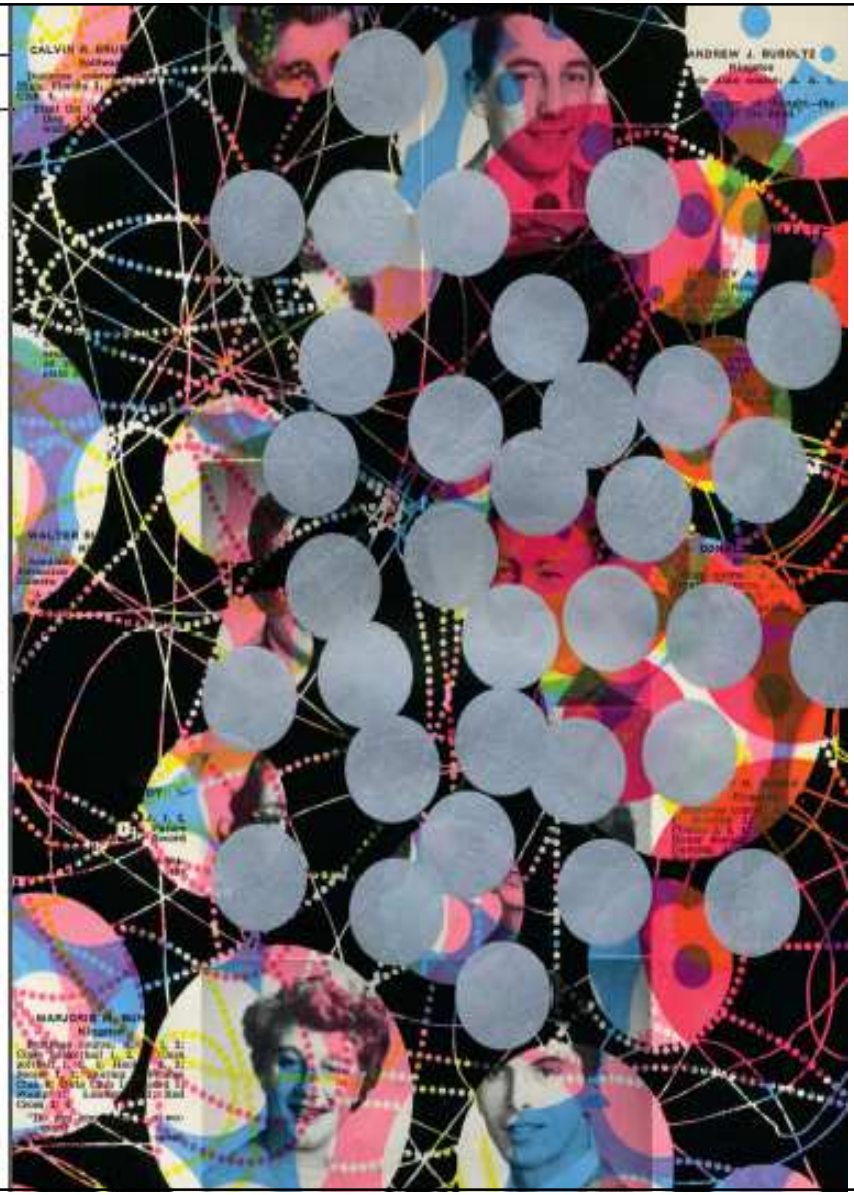
Data Scientist:

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

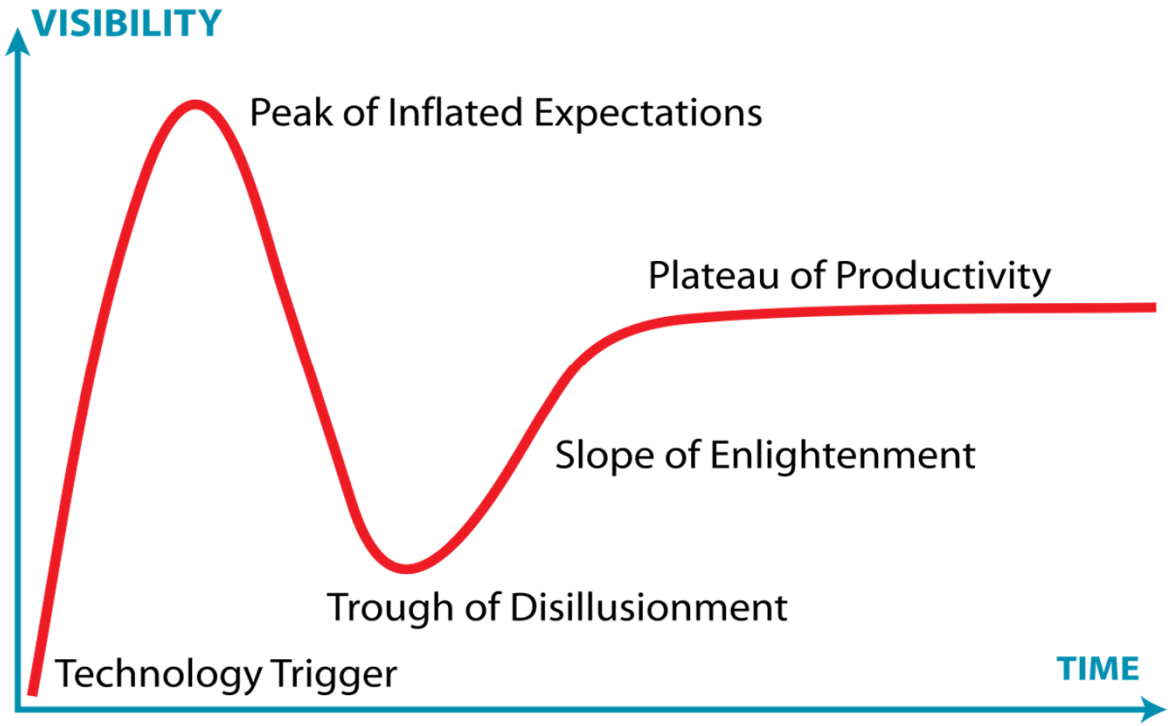
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking 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 leave early."



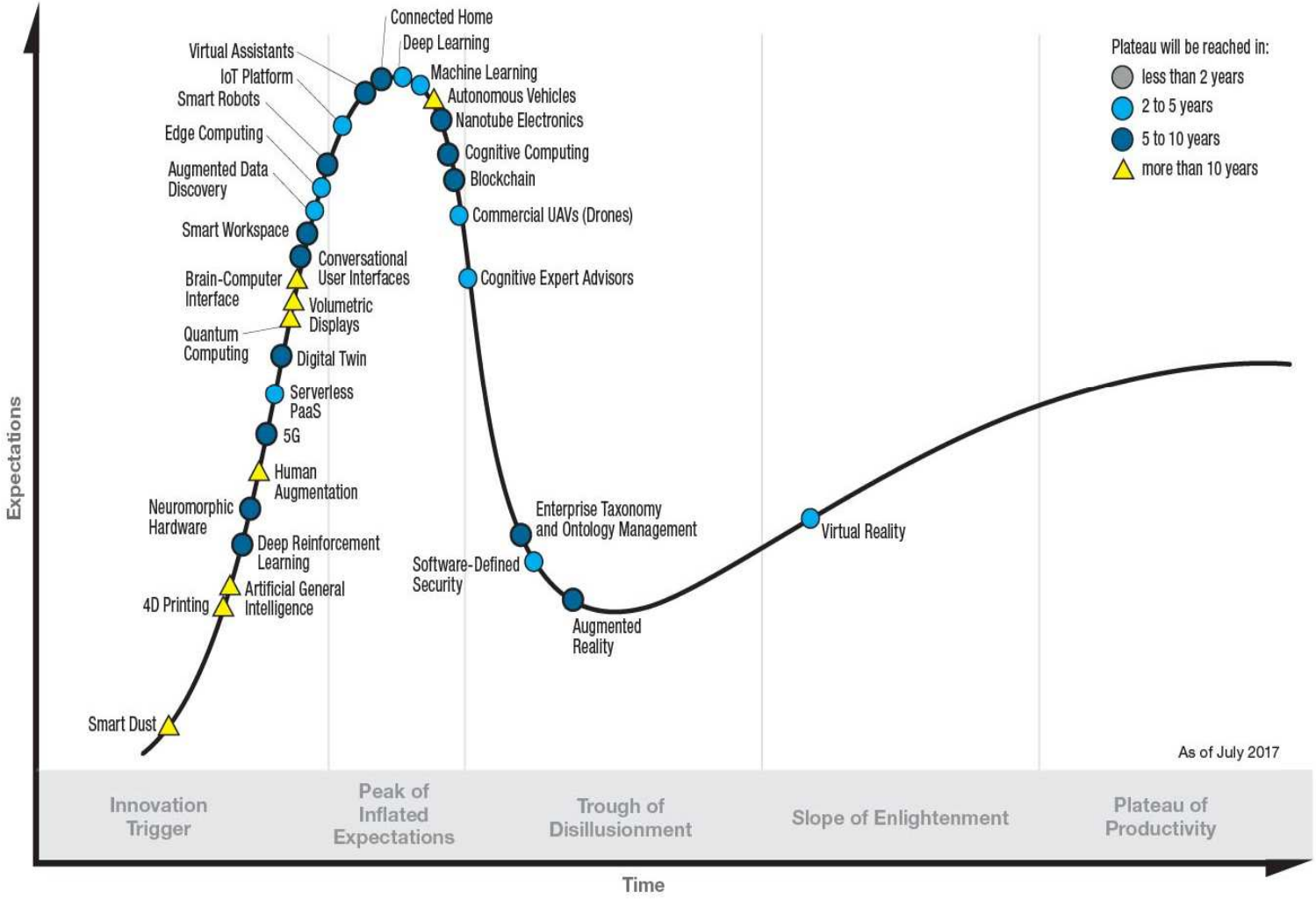
INSEAD

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



Gartner Hype Cycle for Emerging Technologies, 2017



Source: https://www.gartner.com/doc/3783465?ref=SiteSearch&sthkw=Hype%20Cycle%20for%20Analytics%20and%20Business%20Intelligence&fnl=search&srcId=1-3478922254#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

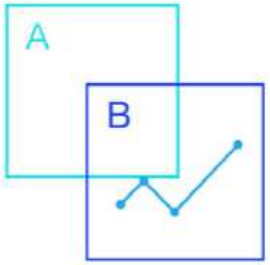
Descriptive

- 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

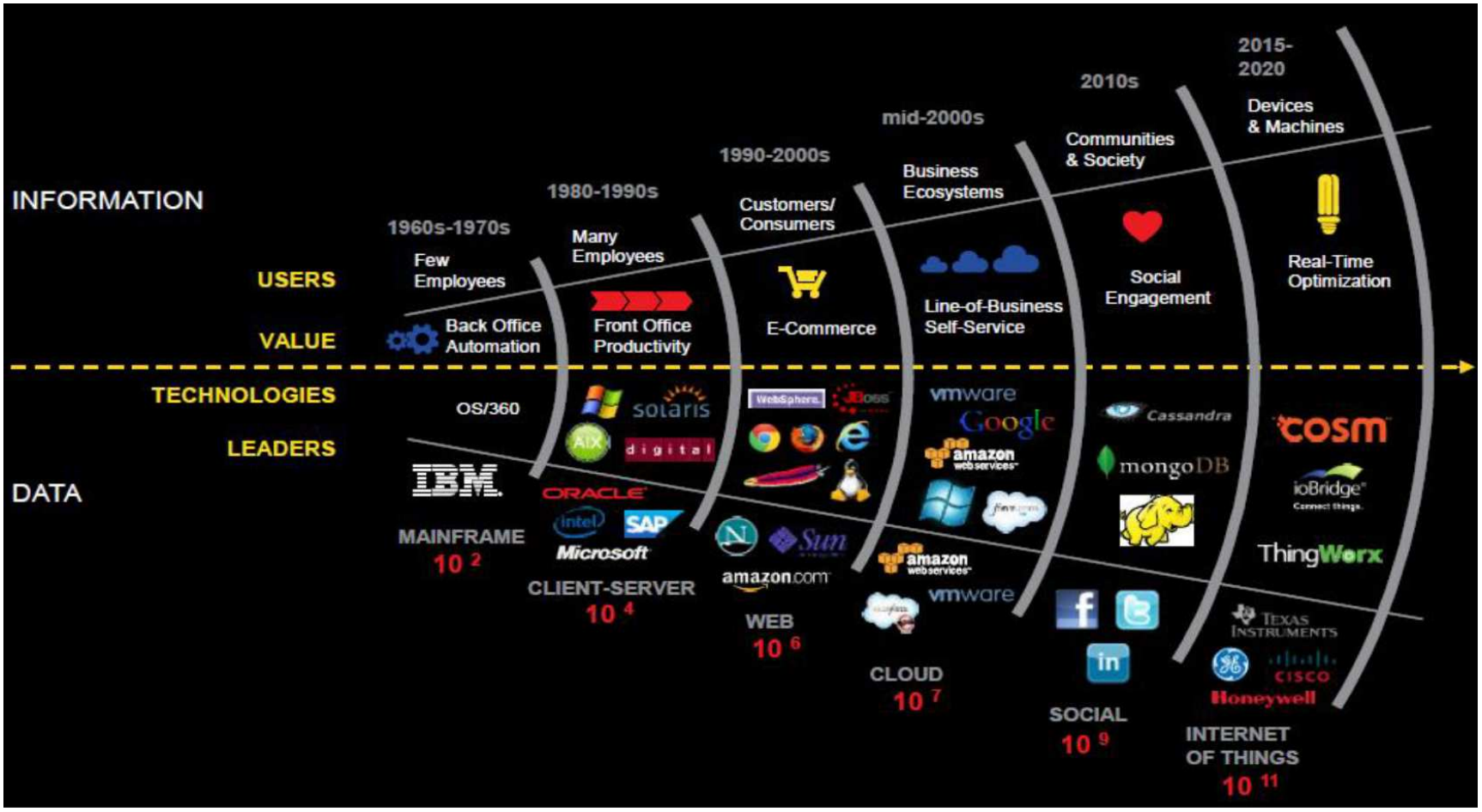


- Provide recommendations on **what to do** to achieve goals
- Employed heavily by leading data and Internet companies

Focus of machine learning

Source: McKinsey "An Executive's Guide to AI" https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/an-executives-guide-to-ai

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

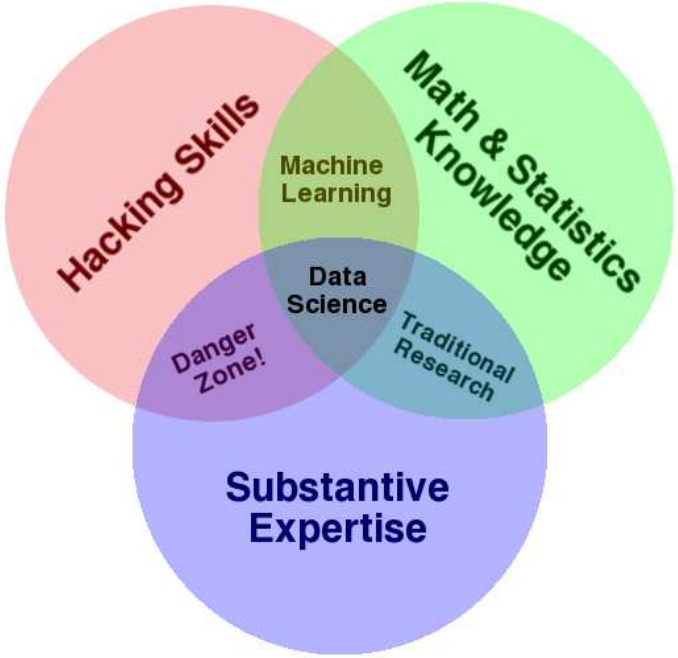
Data Engineering

Data Analytics

Business Expertise

The driver of “Smart Data”: Data Science

Data Engineering



Analytics

Business Expertise

Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

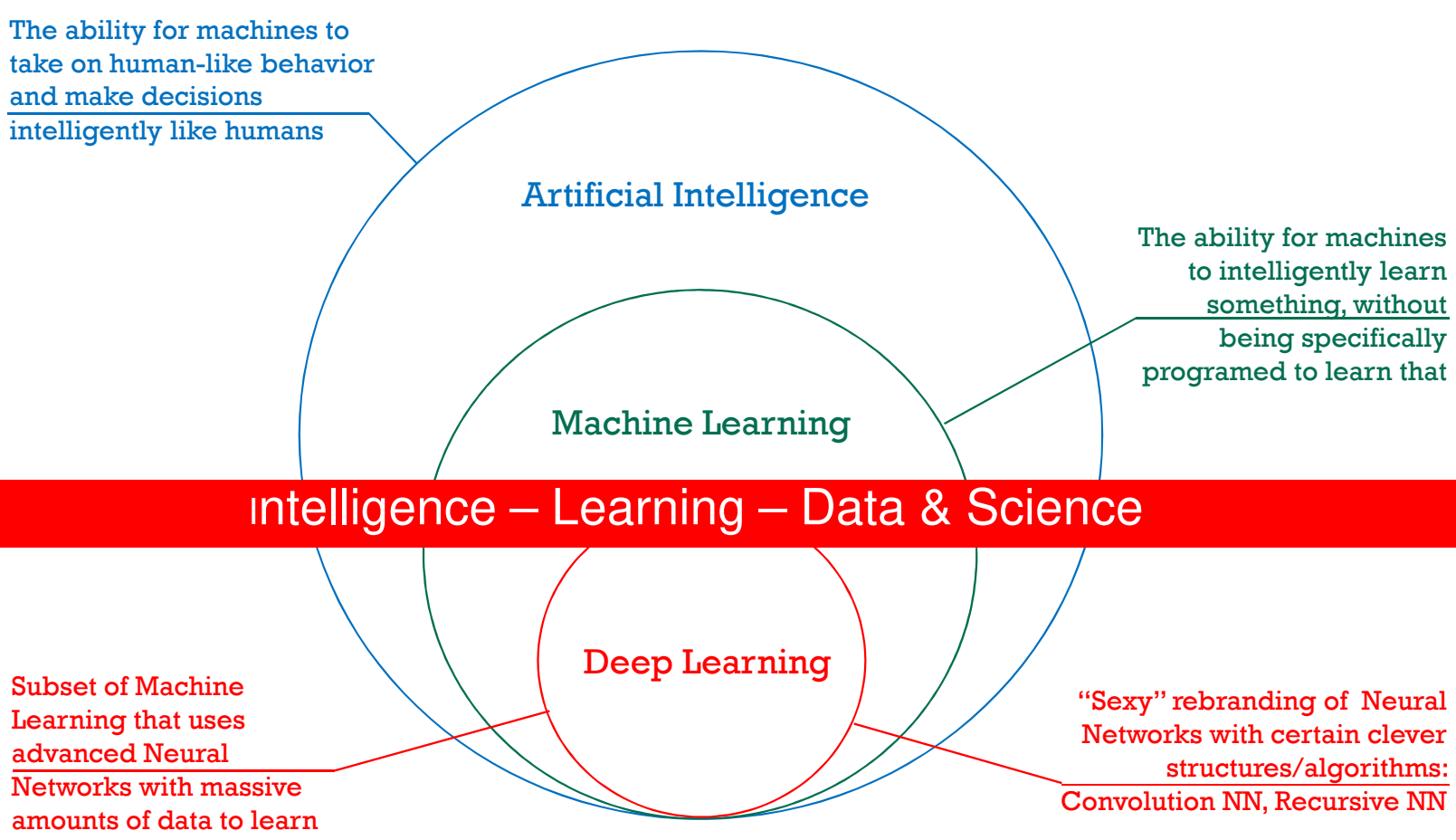
Making sense of AI and ML

ENCYCLOPÆDIA BRITANNICA

Artificial intelligence
/ˌɑː.tɪ.fɪʃ.əl ɪnˈtel.ɪ.dʒəns /

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

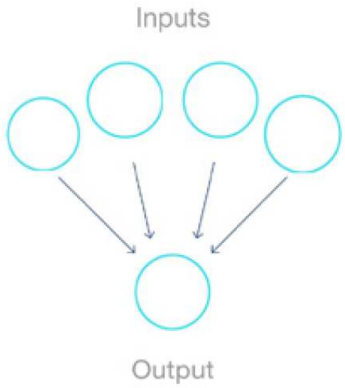
Making sense of AI and ML



Learning: Supervised vs Unsupervised vs Reinforcement

INSEAD

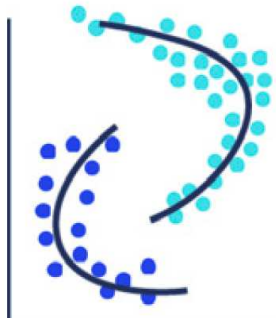
Supervised learning



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

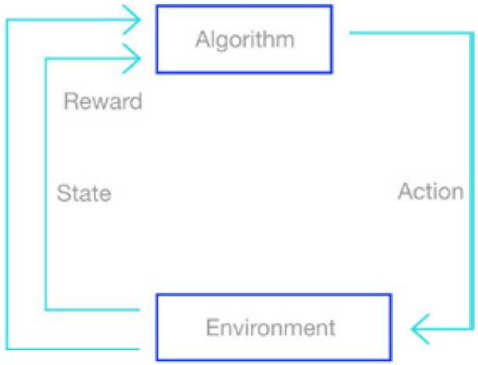
Unsupervised learning



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

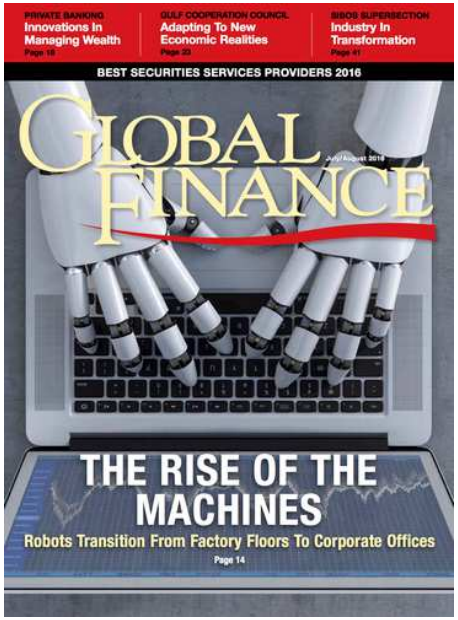
Reinforcement learning



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

8 PROCEEDINGS OF THE IRE January

Steps Toward Artificial Intelligence*

MARVIN MINSKY†, MEMBER, IRE

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 work together into a coherent summary, supplement it with appropriate explanatory or theoretical noncomputer information, and introduce his assessment of the state-of-the-art. This paper emphasizes the class of activities in which a general purpose computer, complete with a library of basic programs, is further programmed to perform operations leading to ever higher-level information processing functions such as learning and problem solving. This informative article will be of real interest to both the general PROCEEDINGS reader and the computer specialist.—*The Guest Editor*

***Summary*—The problems of heuristic programming—of making computers solve really difficult problems—are divided into five main areas: Search, Pattern-Recognition, Learning, Planning, and Induction.**

A computer can do, in a sense, only what it is told to do. But even when we do not know how to solve a certain problem, we may program a machine (computer) to *Search* through some large space of solution attempts. Unfortunately, this usually leads to an enormously inefficient process. With *Pattern-Recognition* techniques, efficiency can often be improved, by restricting the application of the machine's methods to appropriate problems. Pattern-Recognition, together with *Learning*, can be used to exploit generalizations based on accumulated experience, further reducing search. By analyzing the situation, using *Planning* methods, we may obtain a fundamental improvement by replacing the given search with a much smaller, more appropriate exploration. To manage broad classes of problems, machines will need to construct models of their environments, using some scheme for *Induction*.

Wherever appropriate, the discussion is supported by extensive citation of the literature and by descriptions of a few of the most successful heuristic (problem-solving) programs constructed to date.

INTRODUCTION

A VISITOR to our planet might be puzzled about the role of computers in our technology. On the one hand, he would read and hear all about wonderful "mechanical brains" baffling their creators with prodigious intellectual performance. And he (or it) would be warned that these machines must be restrained, lest they overwhelm us by might, persuasion, or even by the revelation of truths too terrible to be borne. On the other hand, our visitor would find the machines being denounced, on all sides, for their slavish obedience, unimaginative literal interpretations, and incapacity for innovation or initiative; in short, for their inhuman dullness.

Our visitor might remain puzzled if he set out to find, and judge for himself, these monsters. For he would find only a few machines (mostly "general-purpose" computers, programmed for the moment to behave according to some specification) doing things that might claim any real intellectual status. Some would be proving mathematical theorems of rather undistinguished character. A few machines might be playing certain games, occasionally defeating their designers. Some might be distinguishing between hand-printed letters. Is this enough to justify so much interest, let alone deep concern? I believe that it is; that we are on the threshold of an era that will be strongly influenced, and quite possibly dominated, by intelligent problem-solving machines. But our purpose is not to guess about what the future may bring; it is only to try to describe and explain what seem now to be our first steps toward the construction of "artificial intelligence."

Along with the development of general-purpose computers, the past few years have seen an increase in effort toward the discovery and mechanization of problem-solving processes. Quite a number of papers have appeared describing theories or actual computer programs concerned with game-playing, theorem-proving, pattern-recognition, and other domains which would seem to require some intelligence. The literature does not include any general discussion of the outstanding problems of this field.

In this article, an attempt will be made to separate out, analyze, and find the relations between some of these problems. Analysis will be supported with enough examples from the literature to serve the introductory function of a review article, but there remains much relevant work not described here. This paper is highly compressed, and therefore, cannot begin to discuss all these matters in the available space.

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 acknowledgments here—suffice it to say that we have discussed these matters with almost every one of the cited authors.

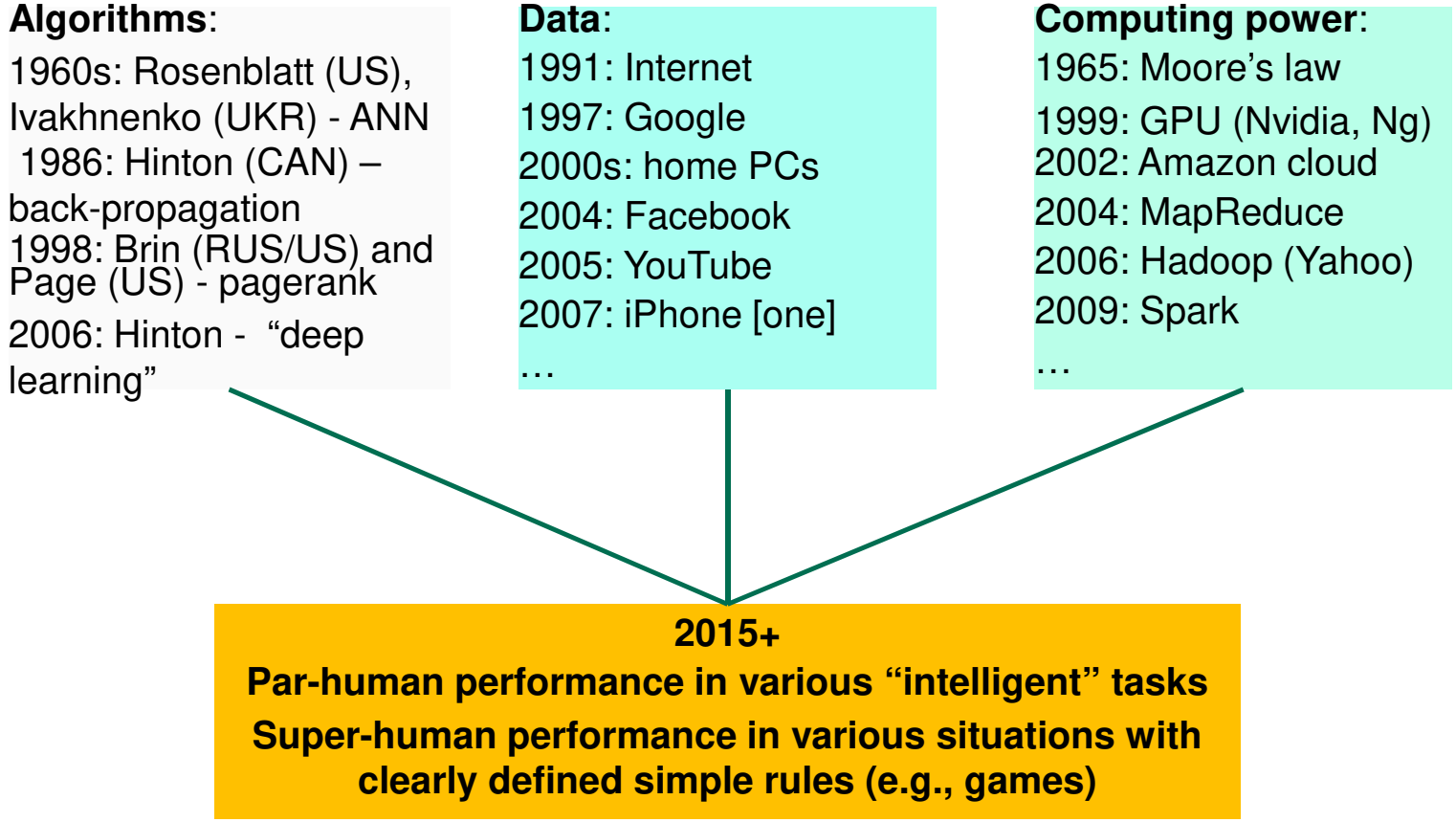
It is convenient to divide the problems into five main areas: Search, Pattern-Recognition, Learning, Planning, and Induction; these comprise the main divisions

* Received by the IRE, October 24, 1960. The author's work summarized here—which was done at Lincoln Lab., a center for research operated by M.I.T. at Lexington, Mass., with the joint support of the U. S. Army, Navy, and Air Force under Air Force Contract AF 19(604)-5200; and at the Res. Lab. of Electronics, M.I.T., Cambridge, Mass., in part under support given jointly by Bell Telephone Labs., the Ford Foundation, and the ONR—is based on earlier work done by the author as a Junior Fellow of the Society of Fellows, Harvard Univ., Cambridge.

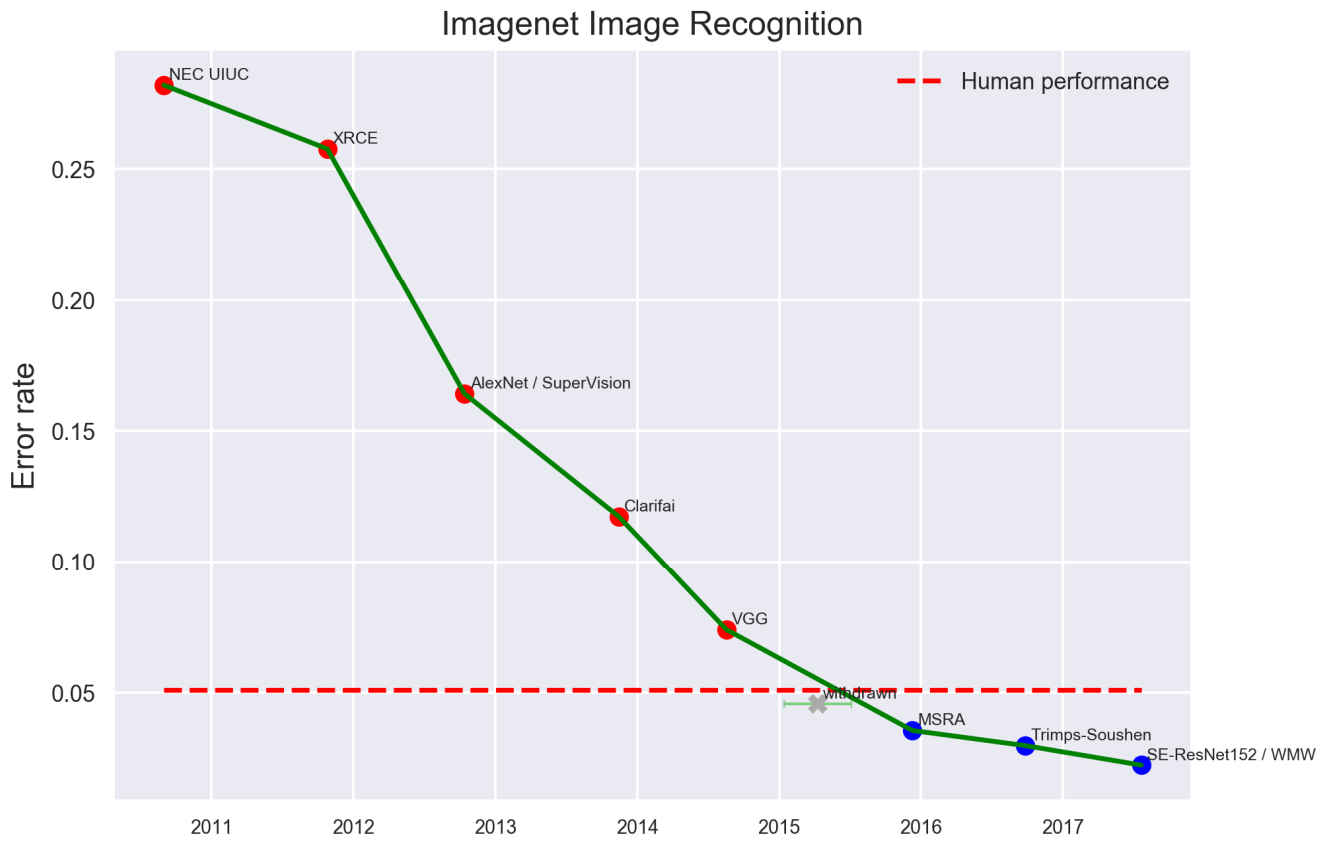
† Dept. of Mathematics and Computation Center, Res. Lab. of Electronics, M.I.T., Cambridge, Mass.

1961

Why this became so important NOW?

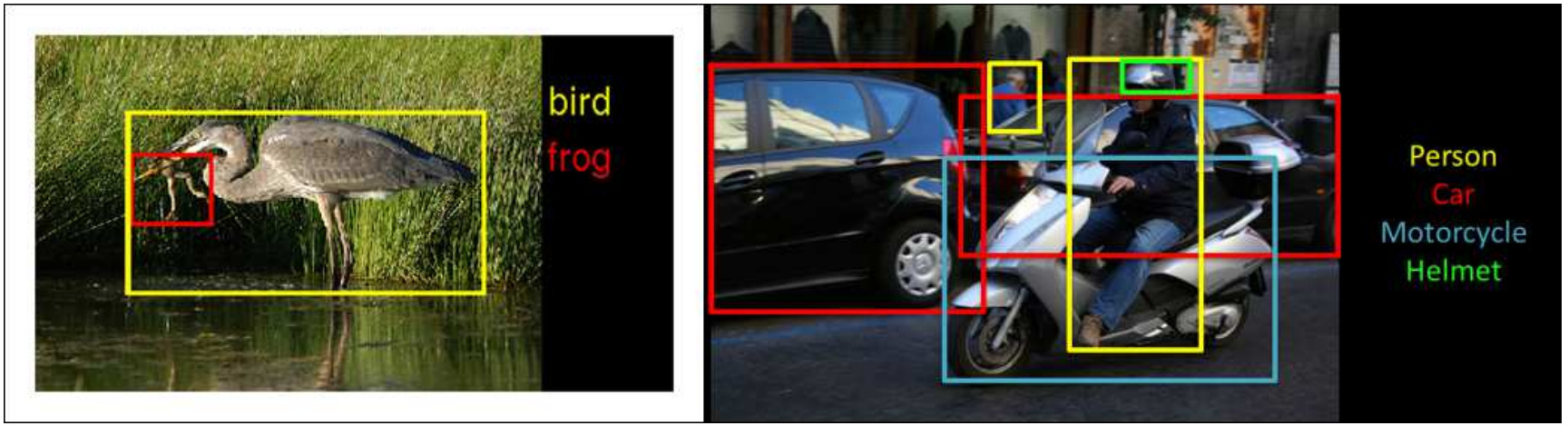


Why this became so important NOW?
Image recognition

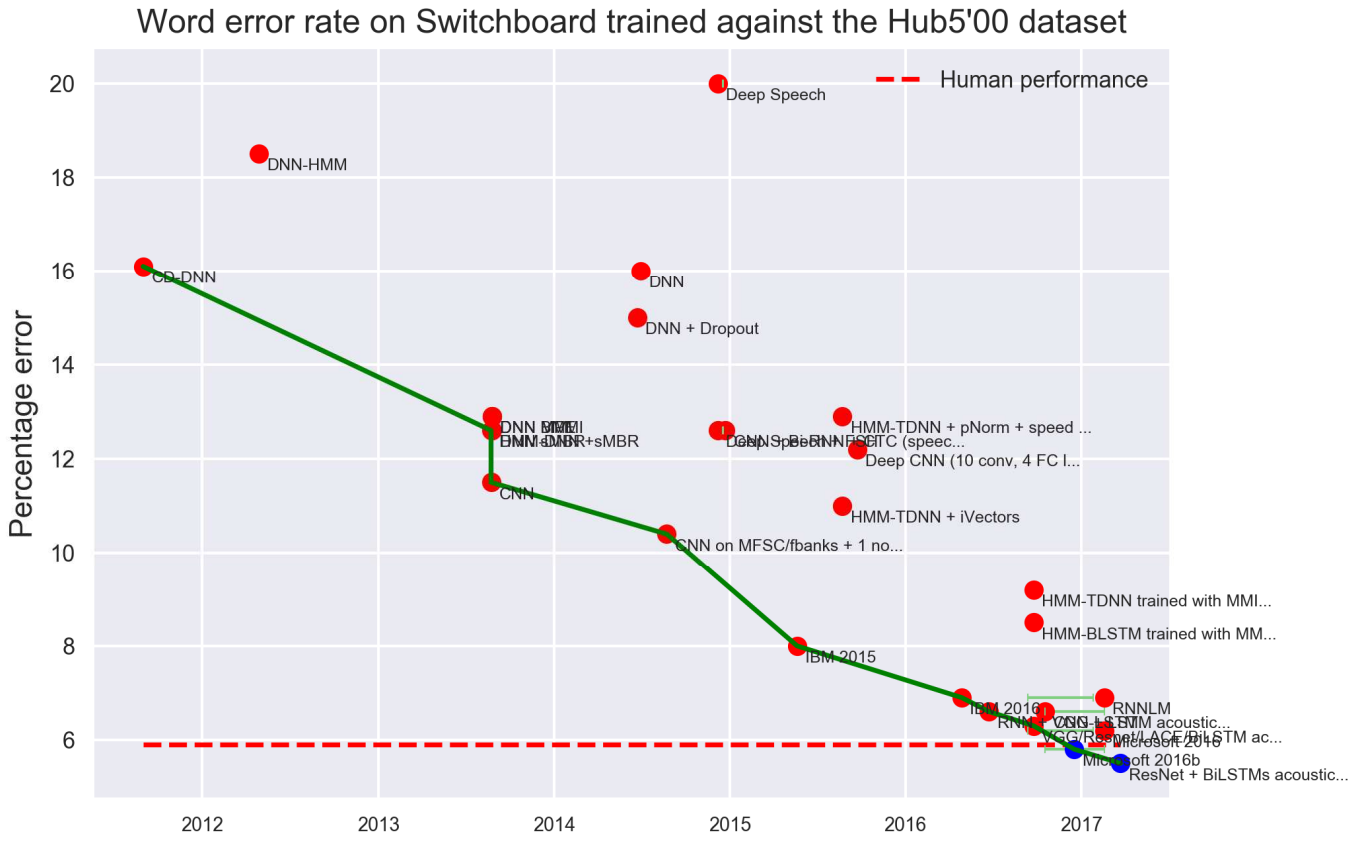


https://www.eff.org/ai/metrics

Why this became so important NOW?
Image recognition

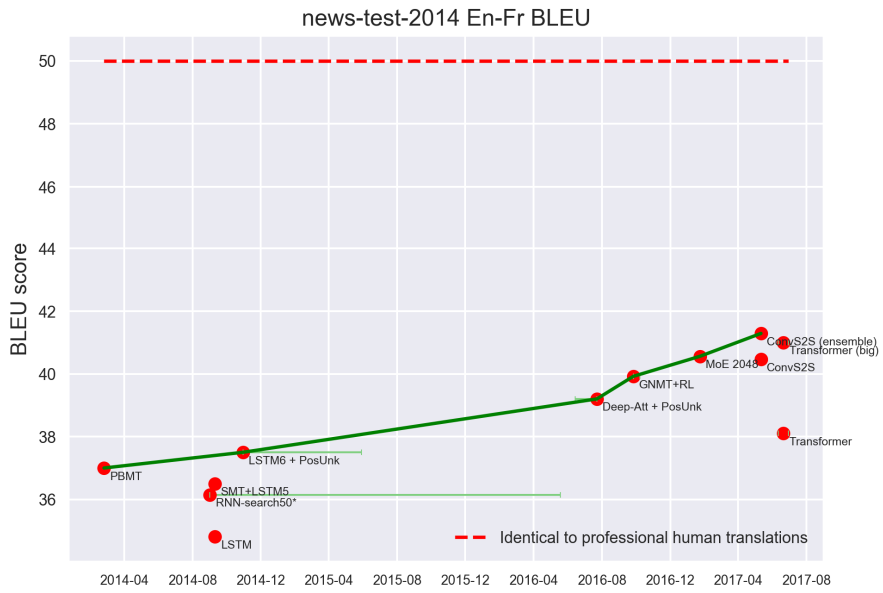


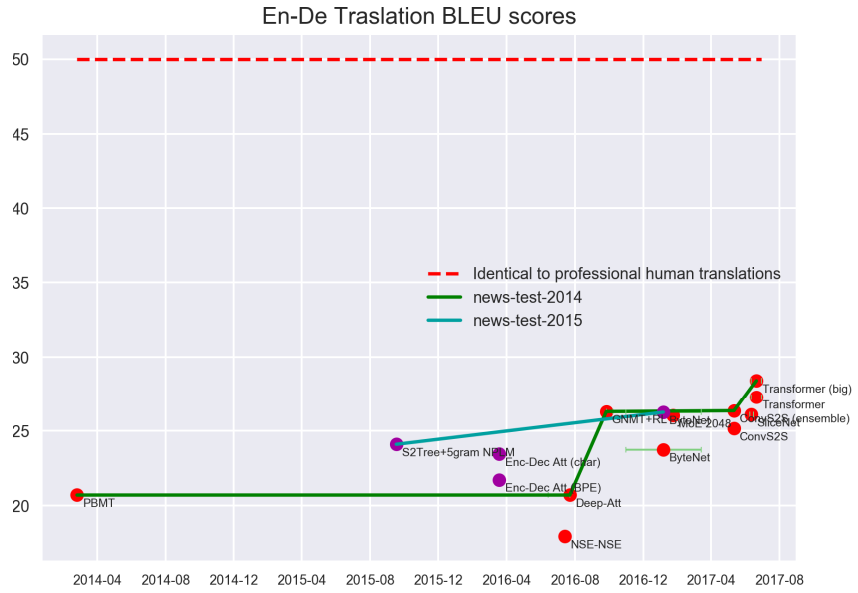
Why this became so important NOW?
Voice-to-text



https://www.eff.org/ai/metrics

Why this became so important NOW?
Translations





Why worse than images or text?

Why performance in French is better than in German?

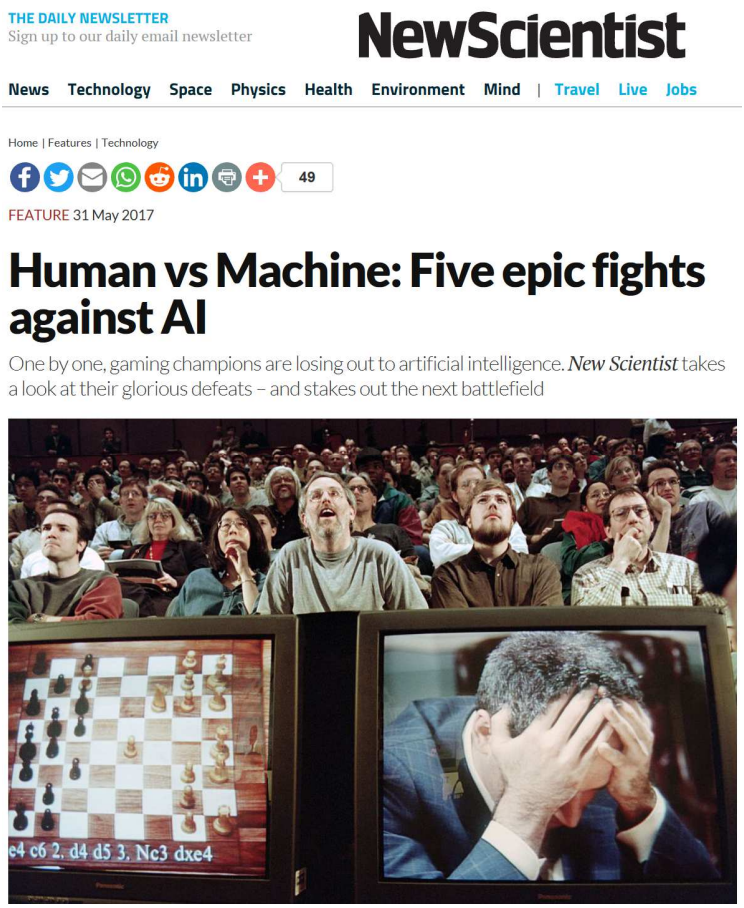
More complex task

More data (Canada's Parliament Records)

https://www.eff.org/ai/metrics

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 bet 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

49

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

https://www.newscientist.com/article/2133146-human-vs-machine-five-epic-fights-against-ai/

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's_paradox

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's_paradox

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

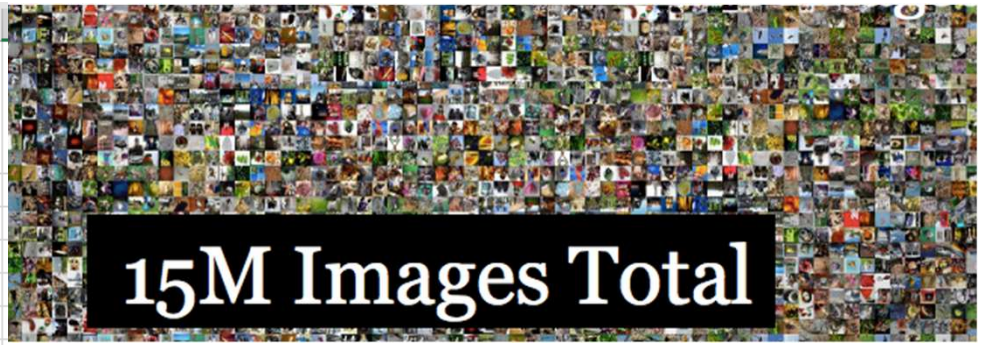
Do they learn from / like humans?

- 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

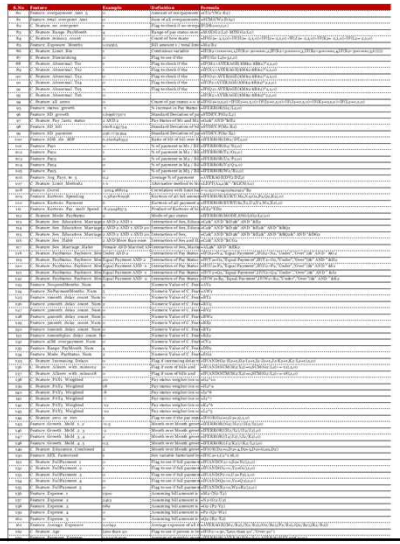


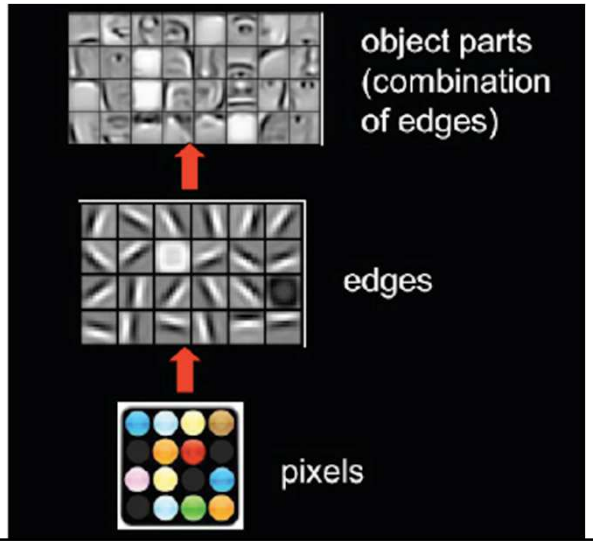
Why this became so important NOW?
How do these algorithms learn

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” (age>65), “single and male”, etc. → VERY MANY





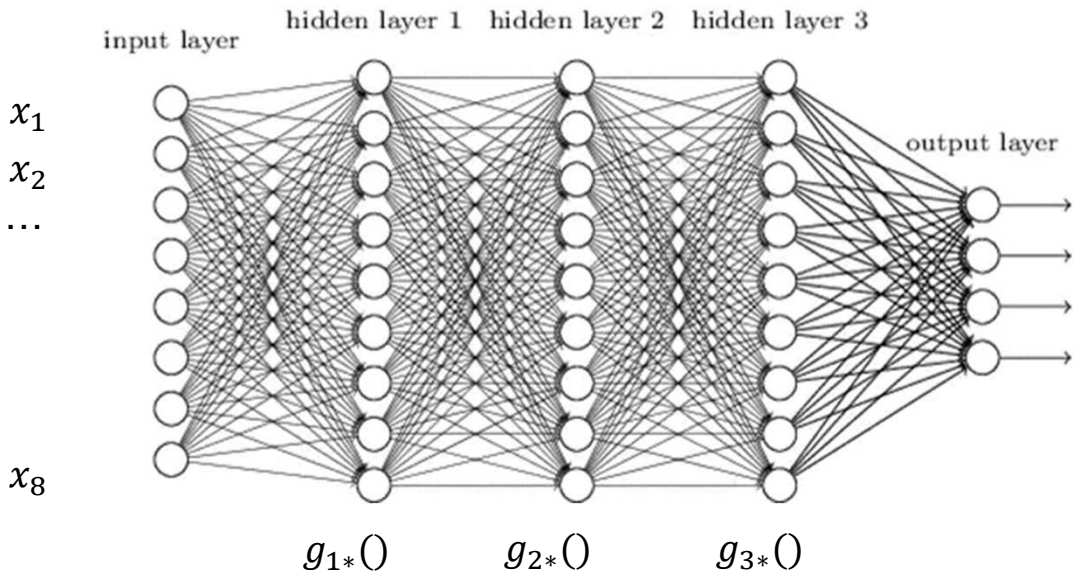
Why this became so important NOW?
How do these algorithms learn

Do they learn from / like humans?

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

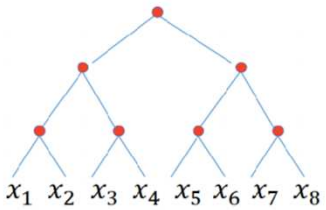
1. Data: “features” / variables that describe the situation
2. Feature Engineering: creating many new variables
3. Indirect (non-linear) functions / relationships / “representations”
   - Regression: Y = f(X) = a + b * X
   - Modern ML: Y = crazy complicated function (of functions (of other functions (of many-many Xs)))

Deep neural network

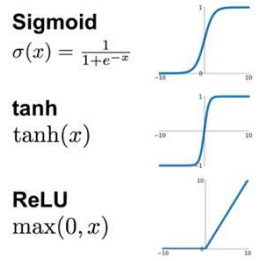


$$Y \equiv f(x_1, x_2, \ldots)$$

$$f(x_1,x_2,\ldots,x_8) = g_3(g_{21}(g_{11}(x_1,x_2),g_{12}(x_3,x_4))g_{22}(g_{11}(x_5,x_6),g_{12}(x_7,x_8)))$$



Activation Functions



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 = 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

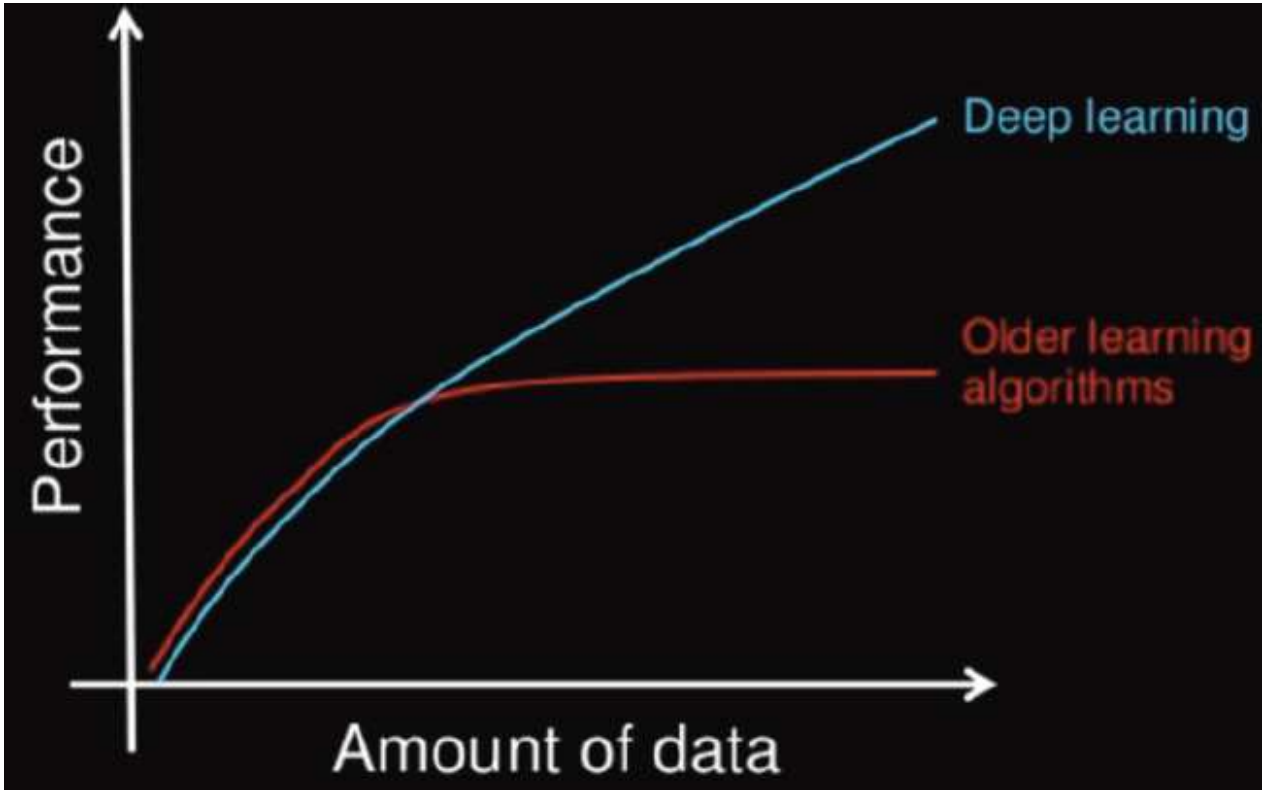
ImageNet classification with deep convolutional neural networks

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 overriding 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?

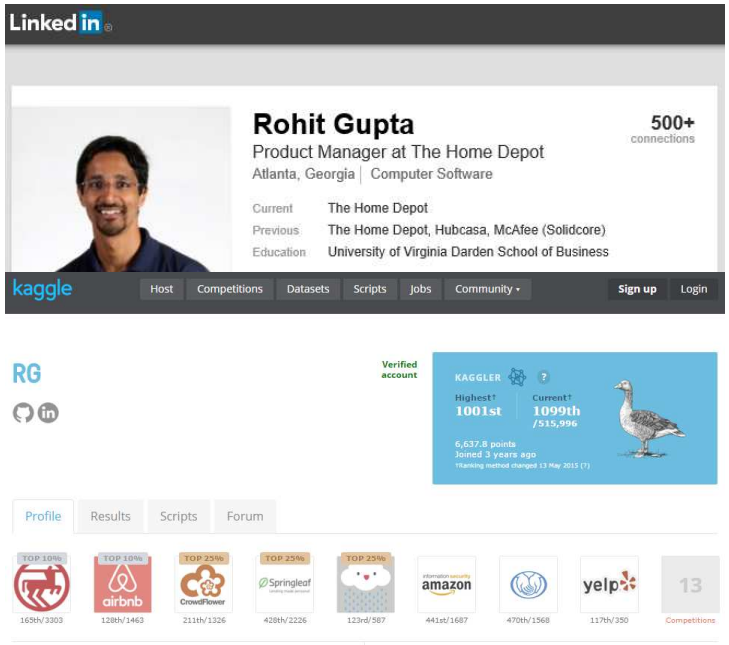
https://www.kaggle.com/competitions

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"]

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



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
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

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

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) 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:
    - **M**ean **A**bsolute **P**ercentage **E**rror (MAPE)
    - **(R**oot) **M**ean **S**quared **E**rror (MSE, RMSE)

- **Does any team know how well they did?**
  - This 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 mem	MAPE
1	Team Jetson	Nick Czuzo	4.72%
2	ERI	JE&his tear	5.36%
3	Team HERTz Analytics	Hasan Too	5.52%
4	Team MDP2	Adam Liu	5.67%
5	Team GOAT (Greatest Team of All Time)	Kasturi Kur	5.70%
6	Default H2O	SP&AO	6.07%
7	Team YAY!	Steven Che	6.51%
8	Team Rita	Rita Wu	6.79%
9	The One Hit Wonders		7.18%
10	Team ABB		7.25%
11	Team "Gonna-Forward-This-To-My-Boyfriend-NOW"	Susan Wan	7.43%
12	Team Sarah gets a Diamond makes Spencer Nervous		7.53%
13	The Dummy Variables		7.54%
14	We've Regressed		7.54%
15	The Nostradamii of Diamonds		7.60%
16	Fly Like A G6		7.60%
17	Yes, I do.		7.60%
18	Two girls, one guy, and a lot of rocks		7.60%
19	Juri dreams of Vodka	Jan Hendri	7.60%
20	Sarah's Answer		7.60%
21	Real Gems		7.60%
22	Team Night Owl		7.60%
23	The Blood Diamonds		7.60%
24	Every Kiss Begins With Jay		7.60%
25	This is why you get cubic zirconia		7.60%
26	Team: Diamonds in the Rough		7.61%
27	Sorry for Partying		7.61%
28	World's Tiniest Handcuffs		7.61%
29	Team Members	Rain Huang	7.62%
30	The Blood Diamonds		7.62%
31	Team Splendites		7.62%
32	The HPs		7.62%
33	She'll never notice it's zirconia?		7.62%
34	Team KissMyDiamond	Poorya Fer	7.64%
35	DLS		7.64%
36	ShinyShinyShiny		7.65%
37	Team Victory	Parikshit Ch	7.77%
38	The Stoners		7.77%
39	Vivek is a Slacker		7.77%
40	M-Cubed		7.77%
41	Team Eglinton	Jaime Lern	7.78%
42	DS Wiz	Rachit Aga	7.78%
43	TeamDataMaffia	Maxime	7.78%
44	Gods of Numbers	Nadezhda	7.78%
45	Dummy	Patricia Du	7.79%
46	GUS JOHNSON: ALL-AMERICAN HERO		7.79%
47	Darden Dummies		7.85%
48	Team Tornado Watch		7.94%
49	"Where is Nik?"	Gregory Fa	8.13%
50	Diamonds Are Overpriced (DAO)		8.25%
51	Cubic Zirconia		8.26%
52	Team Individual	Gary Dhaliv	8.28%
53	The Foxfielders		8.30%
54	Bling, Bling		8.41%
55	PRICING FOR THE WIN	FINJA BOC	8.46%
56	Data Geeks	Kila Walser	8.47%
57	Kohinoor	Sachin Aga	8.59%
58	The Antwerp 4	Dikaios, Jo	8.59%
59	Drink Milk & Kick Ass	Lin Lin	8.59%
60	Data Sharks	Nancy Sriv	8.59%
61	Prestige Worldwide		8.59%
62	SteakSauce	Shi Xian	8.61%
63	Regress My Heart		8.65%
64	The Doug	Paras Gupt	8.70%
65	Team Winning		8.91%
66	TheSeven	Jan Focked	9.61%
67	Diamond District		9.80%
68	abracaDATA	Andrew Ko	10.02%
69	Diamond Data Miners		10.41%

	TEAM	First mem	MAPE
70	Team Aspiring-Minds	Ashish Gup	10.68%
71	LMM		10.83%
72	I heard it was a CZ		10.84%
73	Breakfast at Tiffany's		10.91%
74	AubergineAnalytica	Alexander A	10.92%
75	Team Waitlist	Fernando F	10.93%
76	Rohan's Five	Sehrish Sa	11.06%
77	The Dummy Variables	Ishan Kawl	11.44%
78	Gradient		11.62%
79	fn(input_your_group_name)	Ivan Gn	11.78%
80	Just the Two of Us, We can Make it if we Try	Tovey, Jos	11.85%
81	UNPREDICTABLÉZ	Hill Pruksar	11.87%
82	Team Name: Black Hole	Jay Bhatna	11.87%
83	Team Y-Acasting	Yeganeh G	12.82%
84	VWZ		13.55%
85	Team Chloe	Chloe Chi	14.83%
86	MakeISPRedundantAgain	Bery Prana	15.38%
87	Wildcats	Nambaya C	17.53%
88	Midnight lamps		20.07%
89	Group 8 - AA	Giuditta Sa	20.07%
90	Naive single variable Log-Log	AO&CL	20.07%
91	BloodDiamond Regressors		20.07%
92	Luso-Mandarin	Xiaoyu	22.40%
93	Diamonsters	Bernardo A	22.56%
94	Russ-ians	Darya Ber	24.11%
95	Crack R	Han Songyi	25.02%
96	Okay-Lah	Chris Leun	26.23%
97	Error 505	Shapour Sa	26.37%
98	Pink Panther	TOREN Roy	26.50%
99	Shallow Un-Learning	Stefano Ta	27.44%
100	Group R	Eszter Sant	27.94%
101	Team Starrk	Sheng Jian	27.95%
102	DataHustlers		27.95%
103	Group 9	Antoine	27.95%
104	TeamOfOne	Simeen Ala	27.96%
105	Diamonds are forever	Dimitar To	28.00%
106	FiveGuys+1	Edwar	28.00%
107	Lucy in the Sky	Ishika Saha	28.05%
108	Rinsead	Florian	28.41%
109	BrazilianBus	Leonardo C	28.96%
110	Geeky Durians	Camille An	29.00%
111	G7	Gokalp	29.08%
112	Outliers	Tabriz Aliy	29.12%
113	G10	Archit	29.15%
114	The Undatables	Dianne Gor	29.23%
115	Data Dumpstfiers	John Davie	29.42%
116	Uh-Oh Oreos	Felipe Arav	29.43%
117	Elon's Angels		29.62%
118	like a g6	paul	29.99%
119	CP3		40.14%
120	Team Leonard	Sean O'Ma	45.66%
121	Team Superneura	Kenneth Br	48.13%
122	3L's Diamond Trade		55.88%
123	DAOMG		68.75%
124	Group 1	Clement Va	86.80%
125	Machine Unlearning	Chi Sheng	90.28%
126	Team Guess	Joy Wang	91.67%
127	Team No Regressts	Mitchell Po	96.90%
128	Swagatha Christie	Ross MacD	103.62%
129	Fantastic 5	Isabelle Pe	106.18%
130	Group 8	Mi Zhou	116.37%
131	3 Girls, 3 Guys, No Diamonds	Eva Perrett	118.19%
132	Hypothetically Null	Flora Xu	120.56%
133	Team Lone Solider	Mohib Rab	228.65%
134	randomer	Cody Li	313.90%
135	Can Log La	David Gala	830.47%
136	DATA DORKS	Noemi	######
137	Diamond Hunters	Mariana Le	#VALUE!
138	G03 - Cube Datos	Nathalie Ko	#VALUE!

Blue -- FBL 20J (AA in white, AB in black)
Green -- SGP 20J (AA in white, AB in black)
Black / Gray -- 19D
Yellow -- prior years INSEAD MBAs

Brown -- other MBAs
White -- MMAs
Red -- special guests

INSEAD

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?

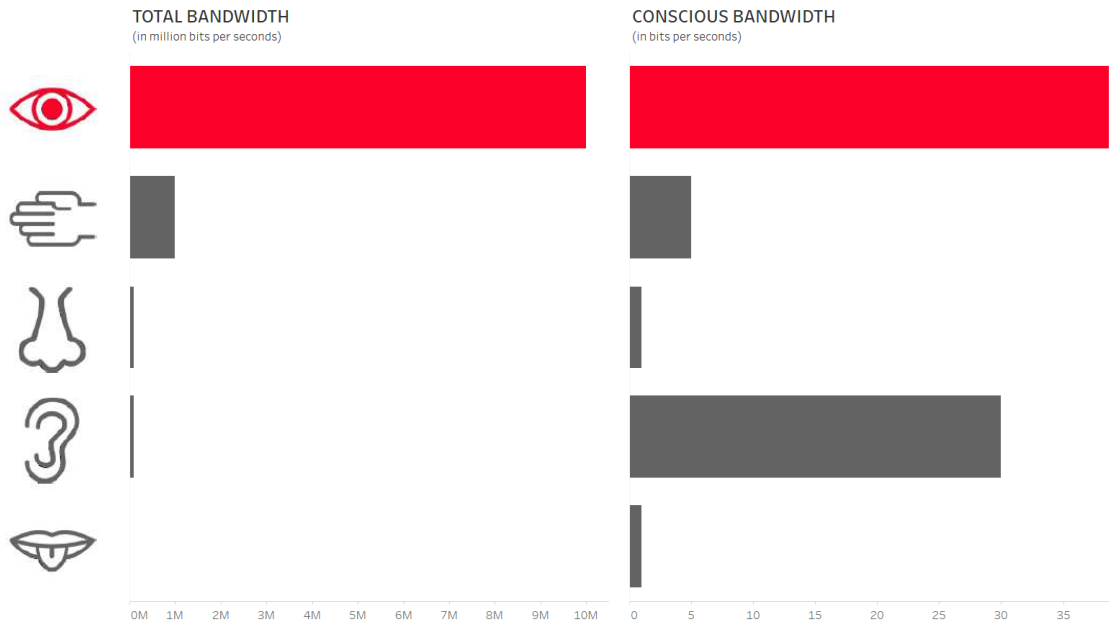
- **Q**: “What is really scarce in the Internet economy?”
- **A**: “Attention”

  2009, Hal Varian [chief economist, Google; Professor, UCBerkeley]

- “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”]

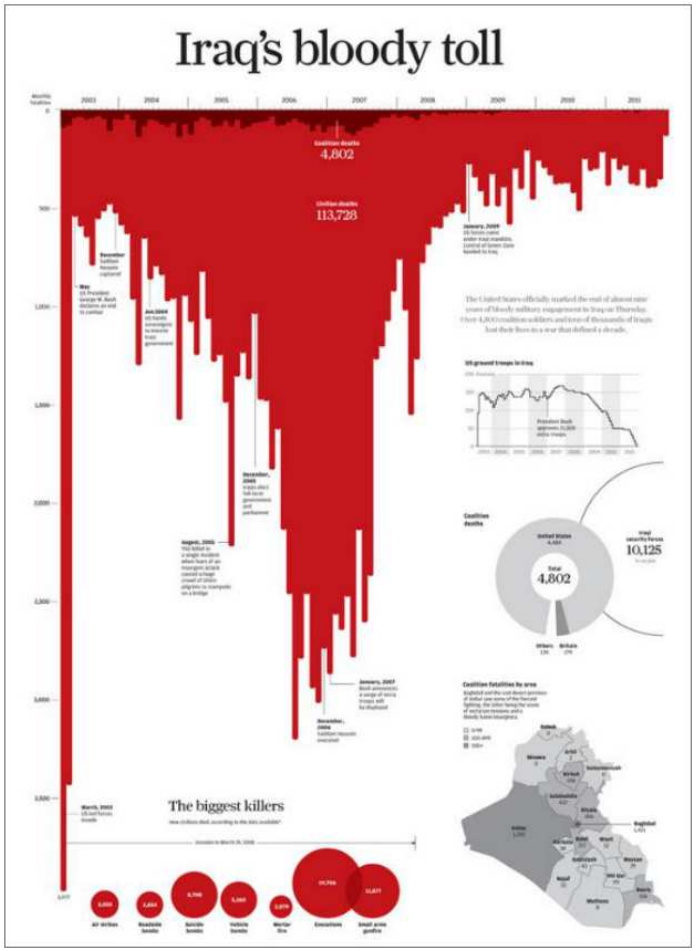
Why data visualization?



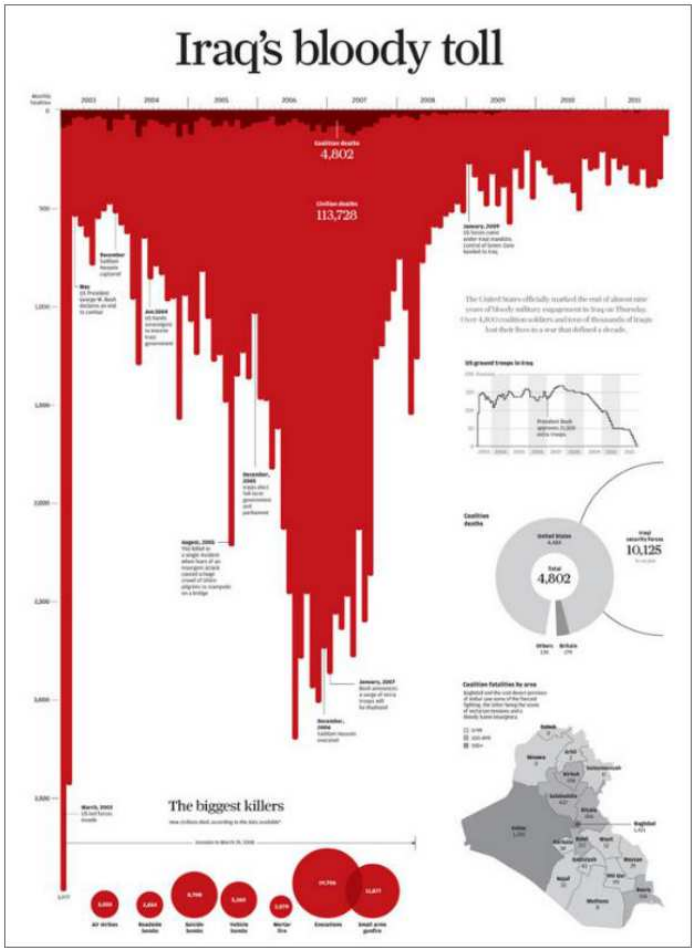
Visual cognition is fastest **and** much of it is subconscious

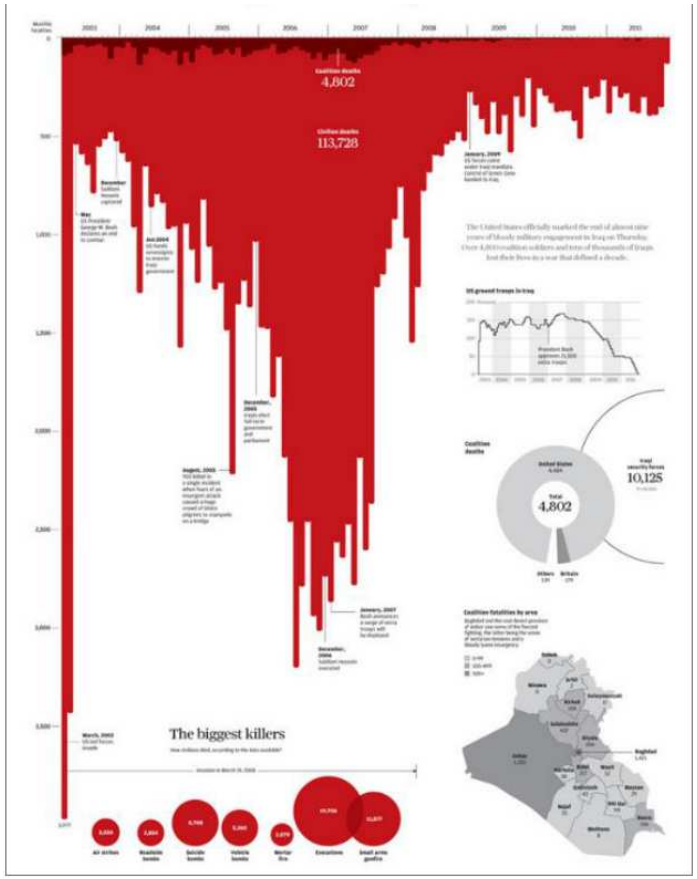
"The best way to capture imagination is to speak to the eyes"
William Playfair [18th century Economist, early "data visualizer"]

Why data visualization?

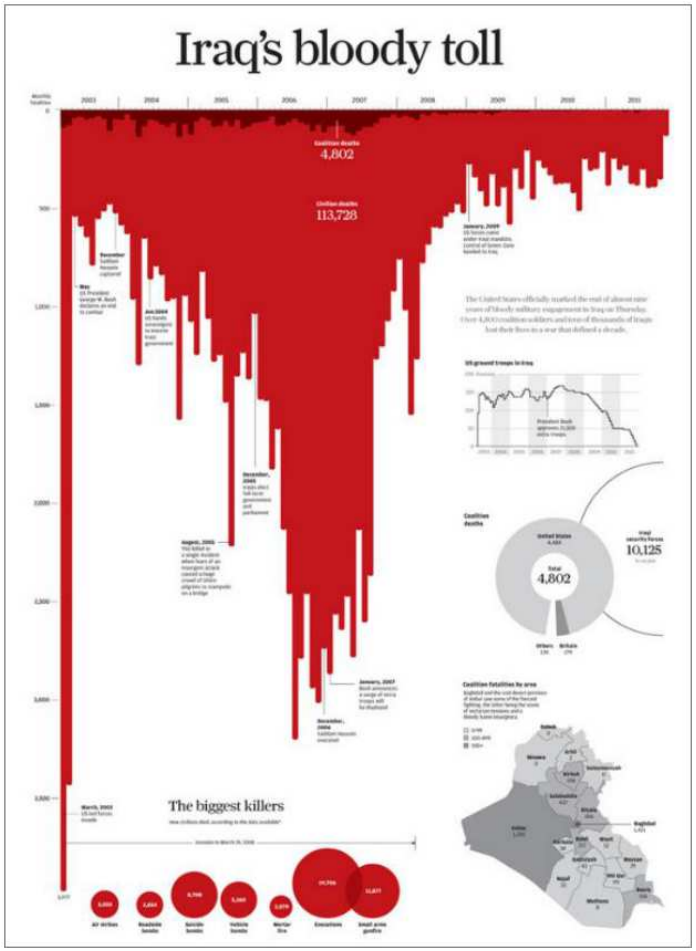


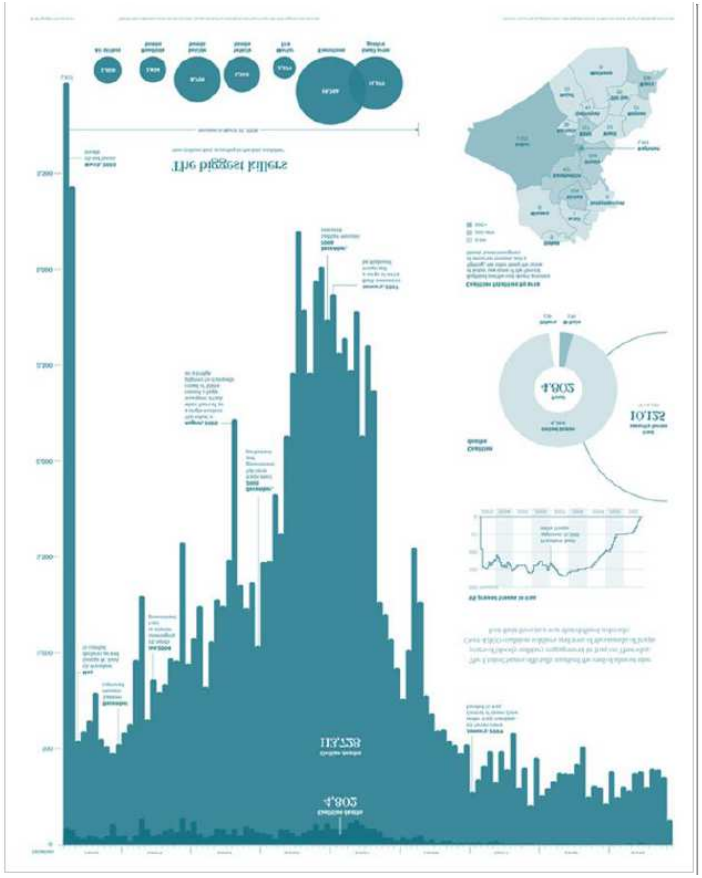
Why data visualization?



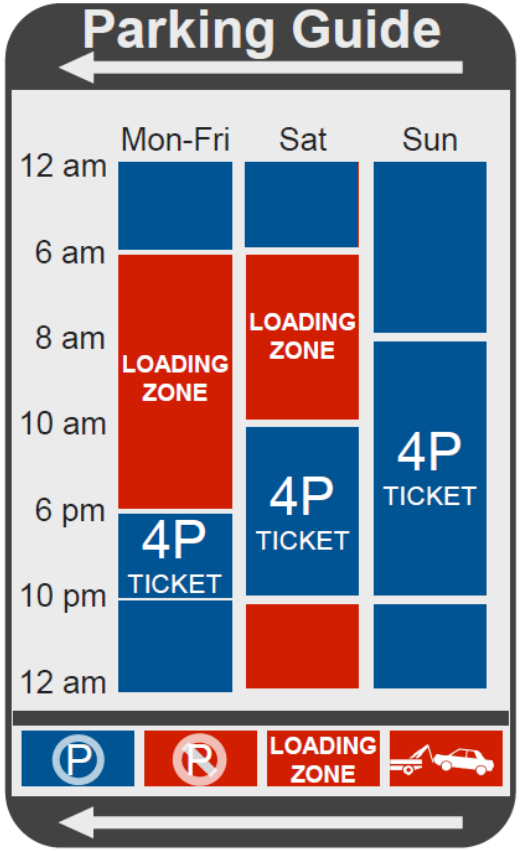


Why data visualization?

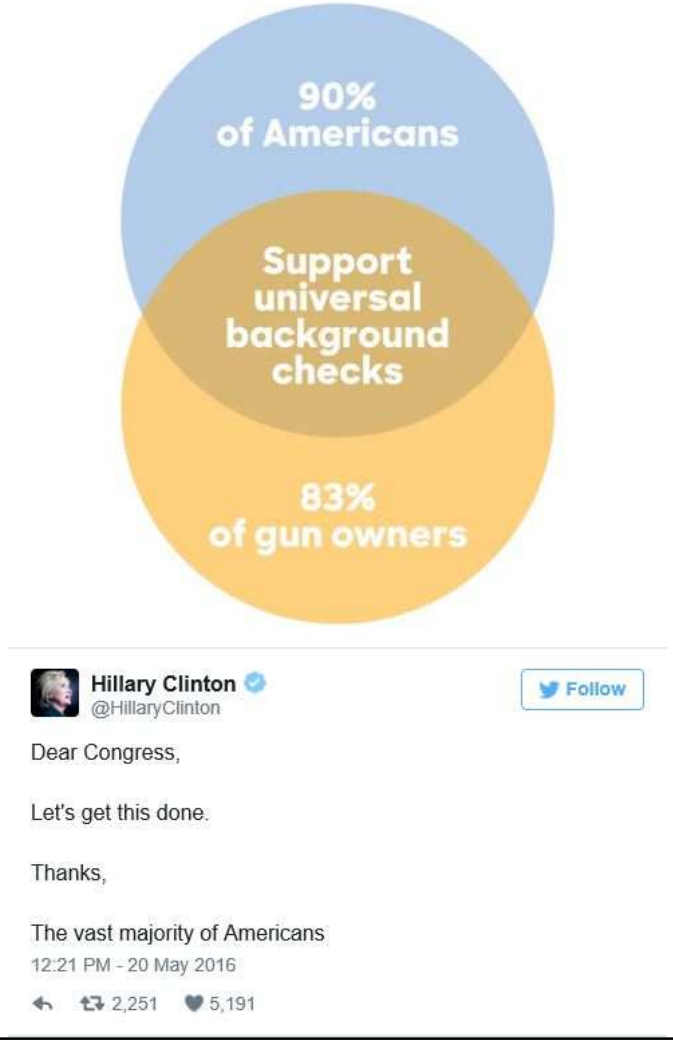




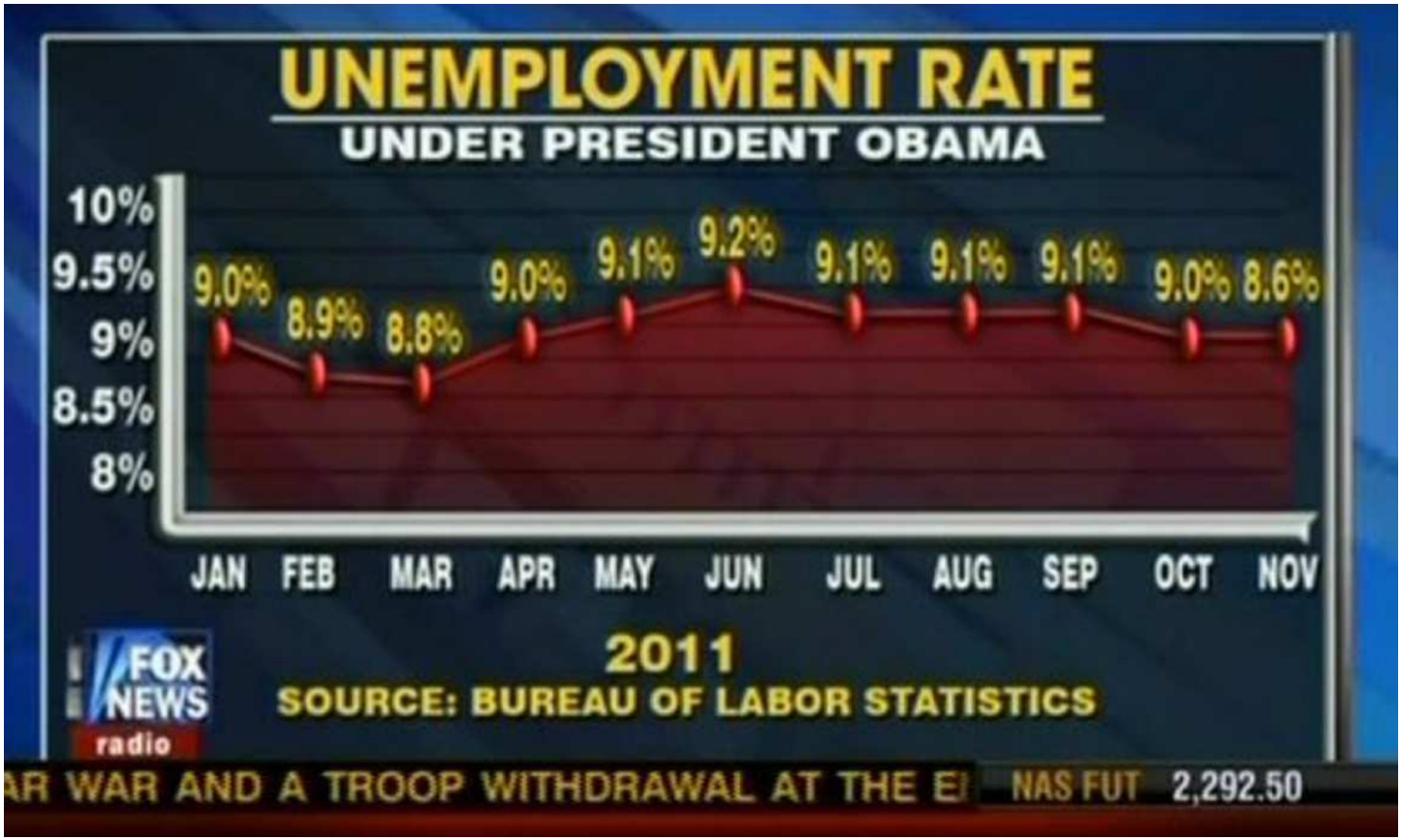
Why read it when you can **see** it?



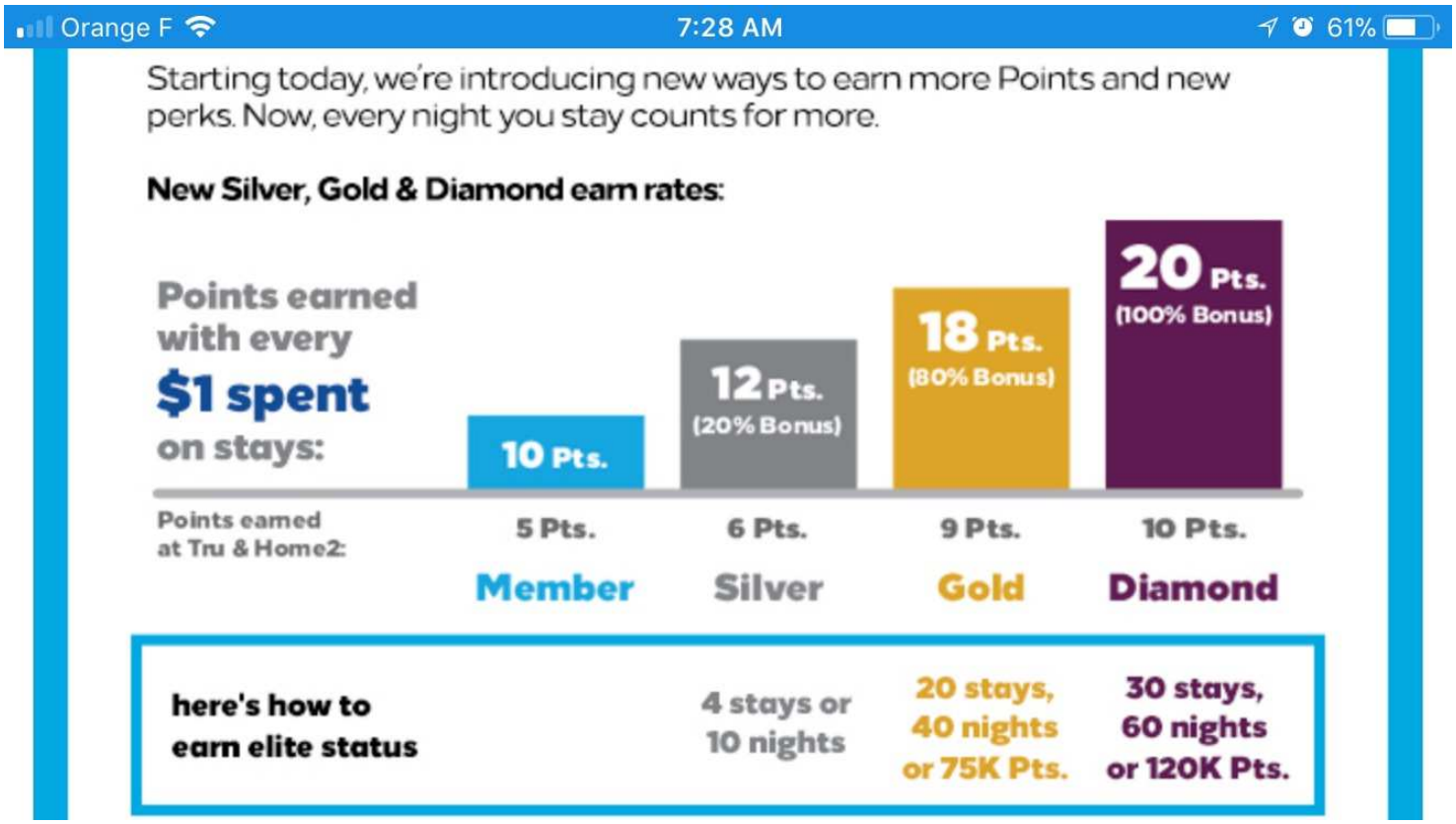
Beware of fails



Beware of manipulations



Beware of manipulations

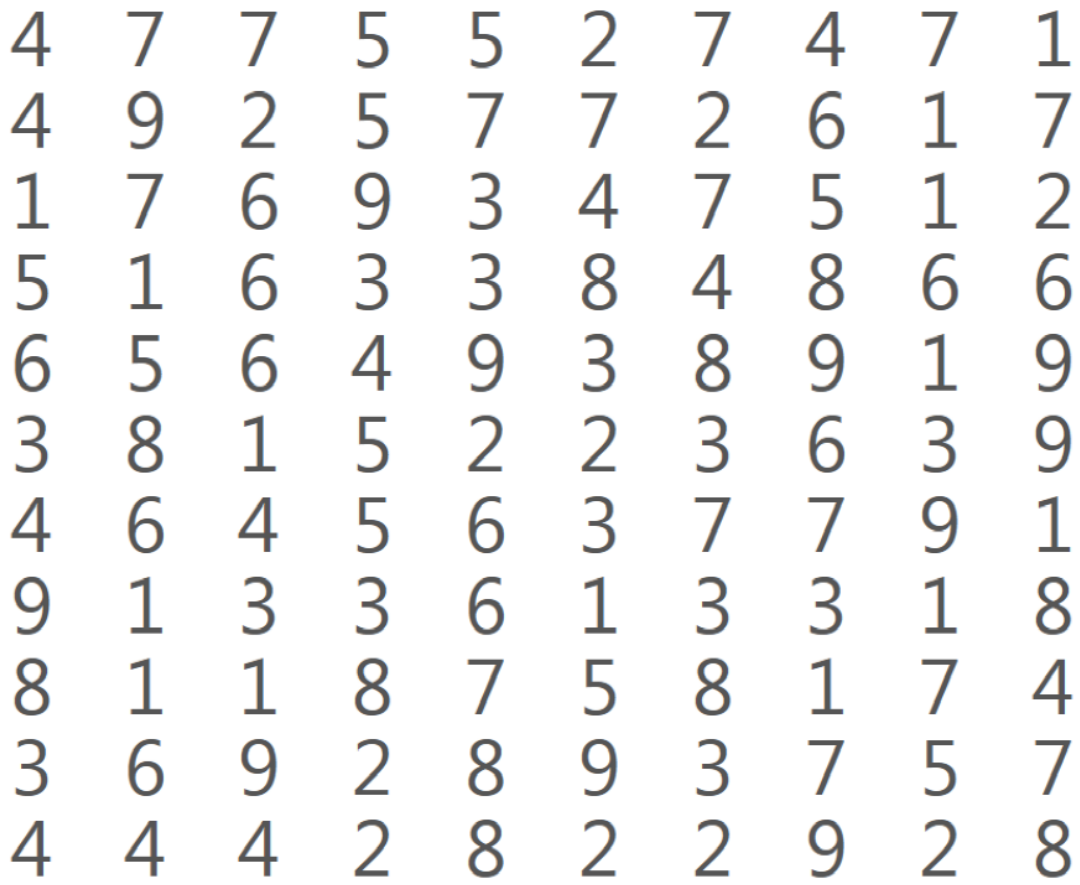


INSEAD

"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?

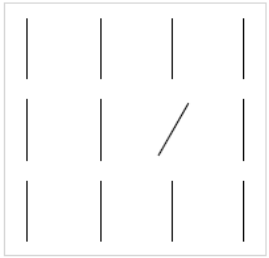

4 7 7 5 5 2 7 4 7 1
4 9 2 5 7 7 2 6 1 7
1 7 6 9 3 4 7 5 1 2
5 1 6 3 3 8 4 8 6 6
6 5 6 4 9 3 8 9 1 9
3 8 1 5 2 2 3 6 3 9
4 6 4 5 6 3 7 7 9 1
9 1 3 3 6 1 3 3 1 8
8 1 1 8 7 5 8 1 7 4
3 6 9 2 8 9 3 7 5 7
4 4 4 2 8 2 2 9 2 8

How many 9s?

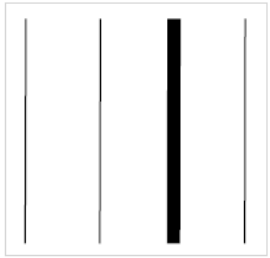

4 7 7 5 5 2 7 4 7 1
4 9 2 5 7 7 2 6 1 7
1 7 6 9 3 4 7 5 1 2
5 1 6 3 3 8 4 8 6 6
6 5 6 4 9 3 8 9 1 9
3 8 1 5 2 2 3 6 3 9
4 6 4 5 6 3 7 7 9 1
9 1 3 3 6 1 3 3 1 8
8 1 1 8 7 5 8 1 7 4
3 6 9 2 8 9 3 7 5 7
4 4 4 2 8 2 2 9 2 8

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

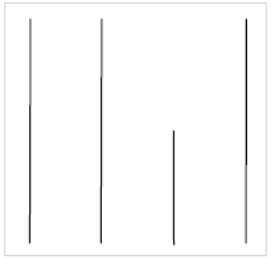
Orientation



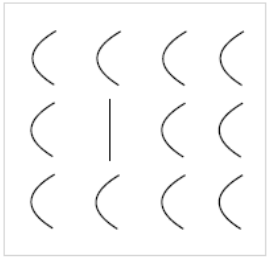
Line Width



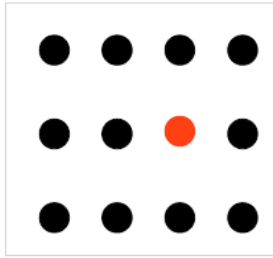
Line Length



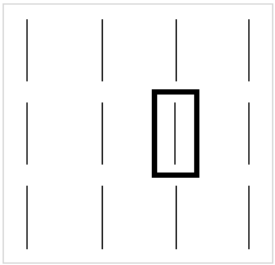
Curved/Straight



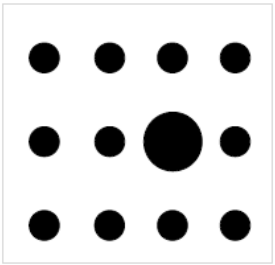
Colour/Hue



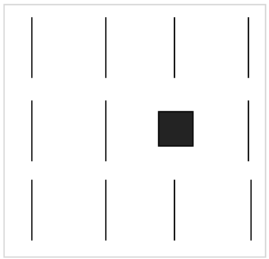
Enclosure



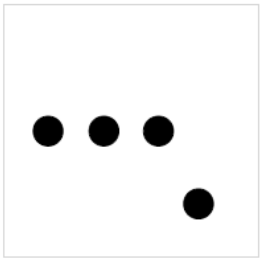
Size



Shape



2D Position



Intensity

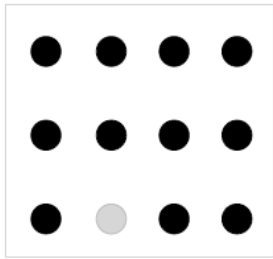


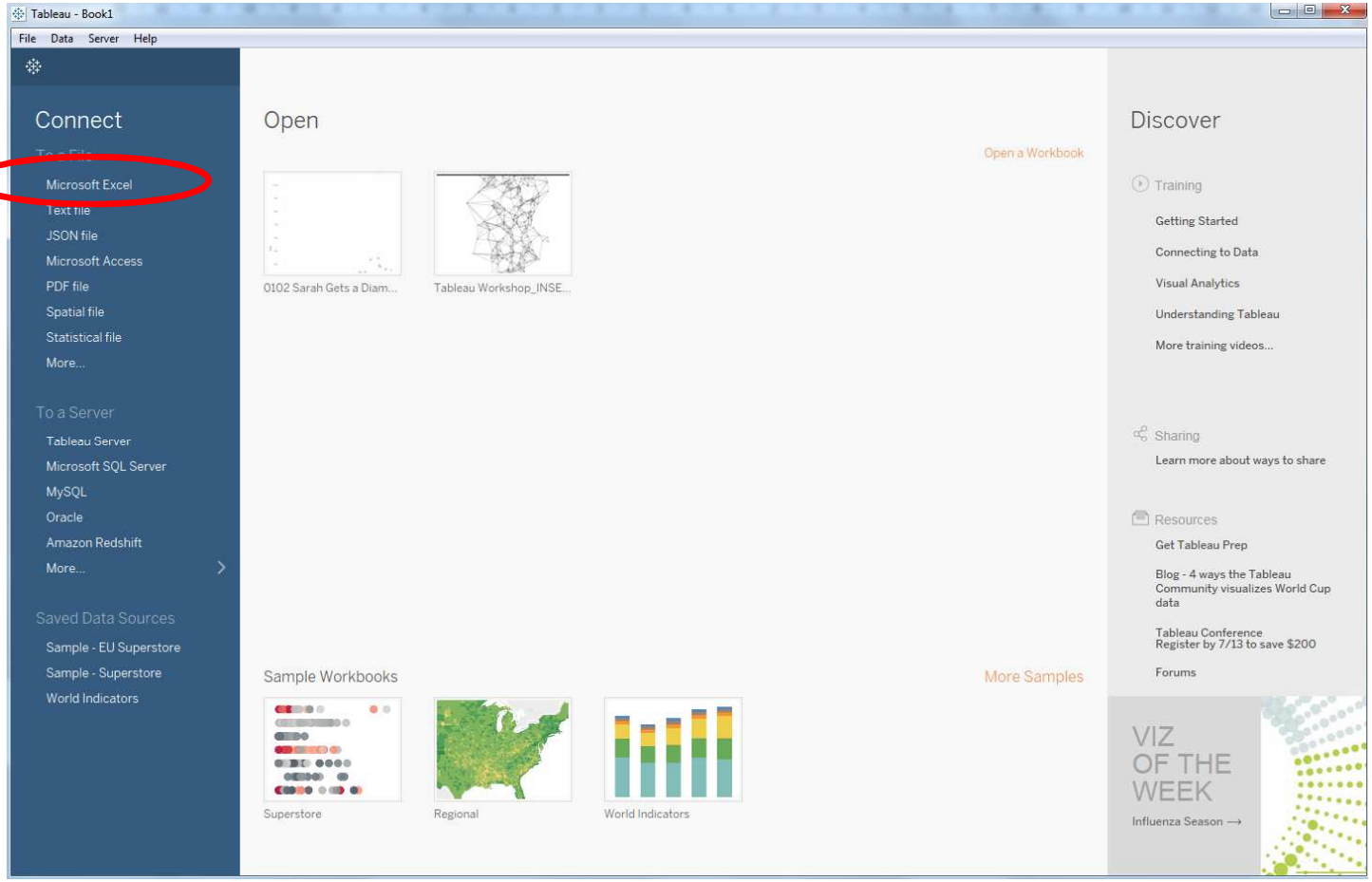
Tableau: leader in data visualization

https://www.tableau.com/

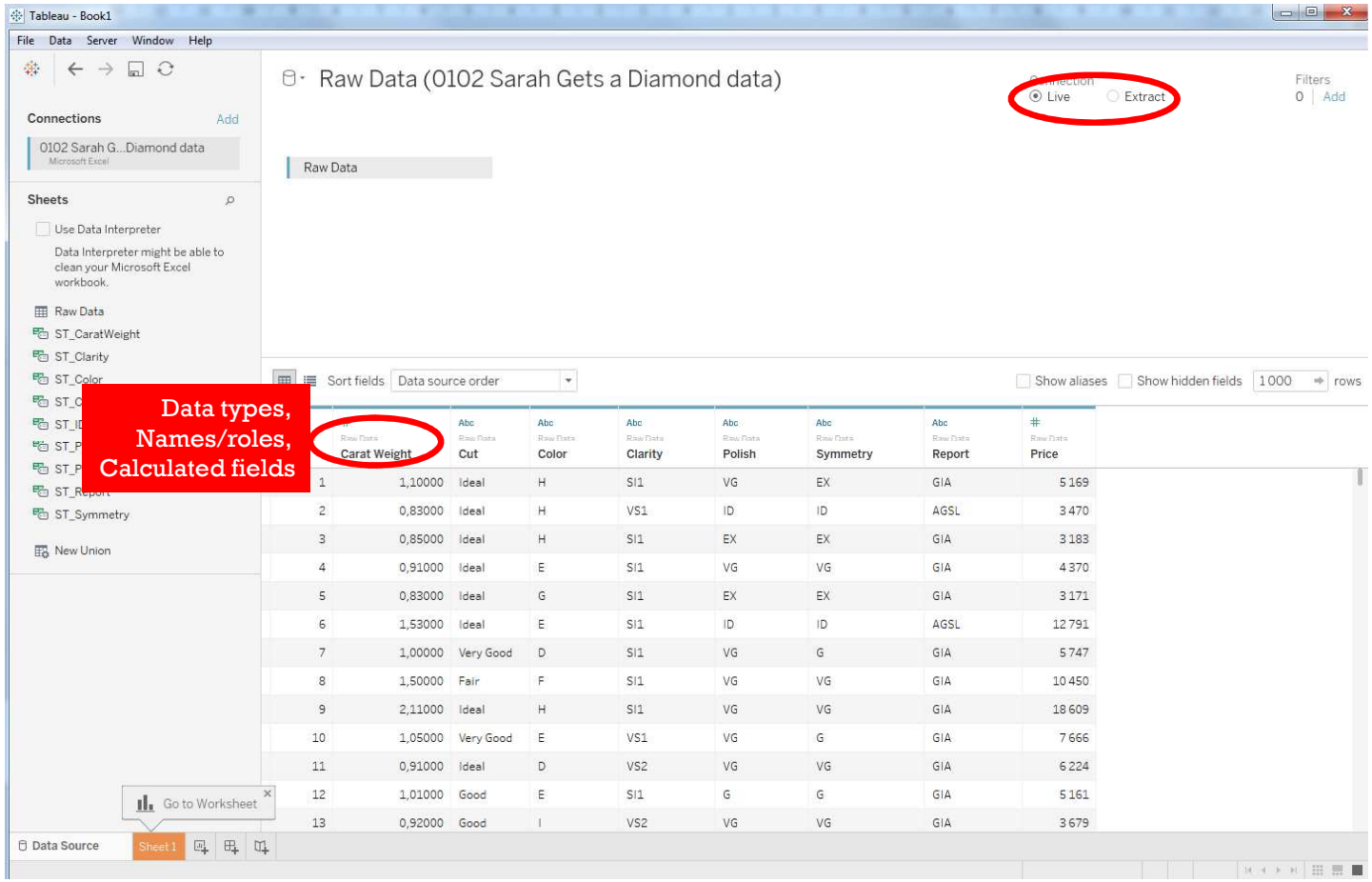
Use your @insead.edu email to request a student product key

https://www.tableau.com/about/blog/2018/2/tableau-named-leader-gartner-magic-quadrant-six-years-row-82534

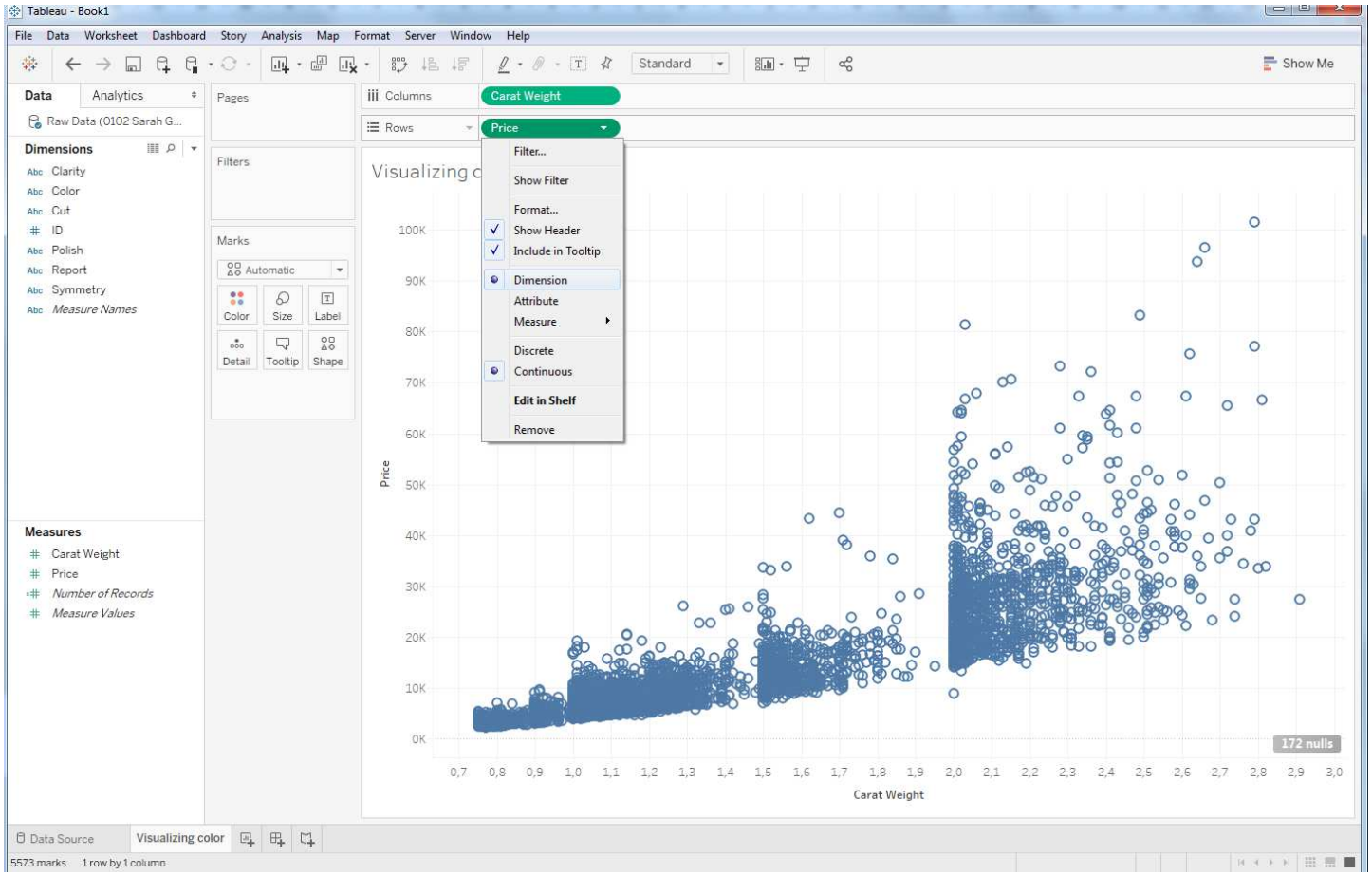
Sarah “color” visual in Tableau



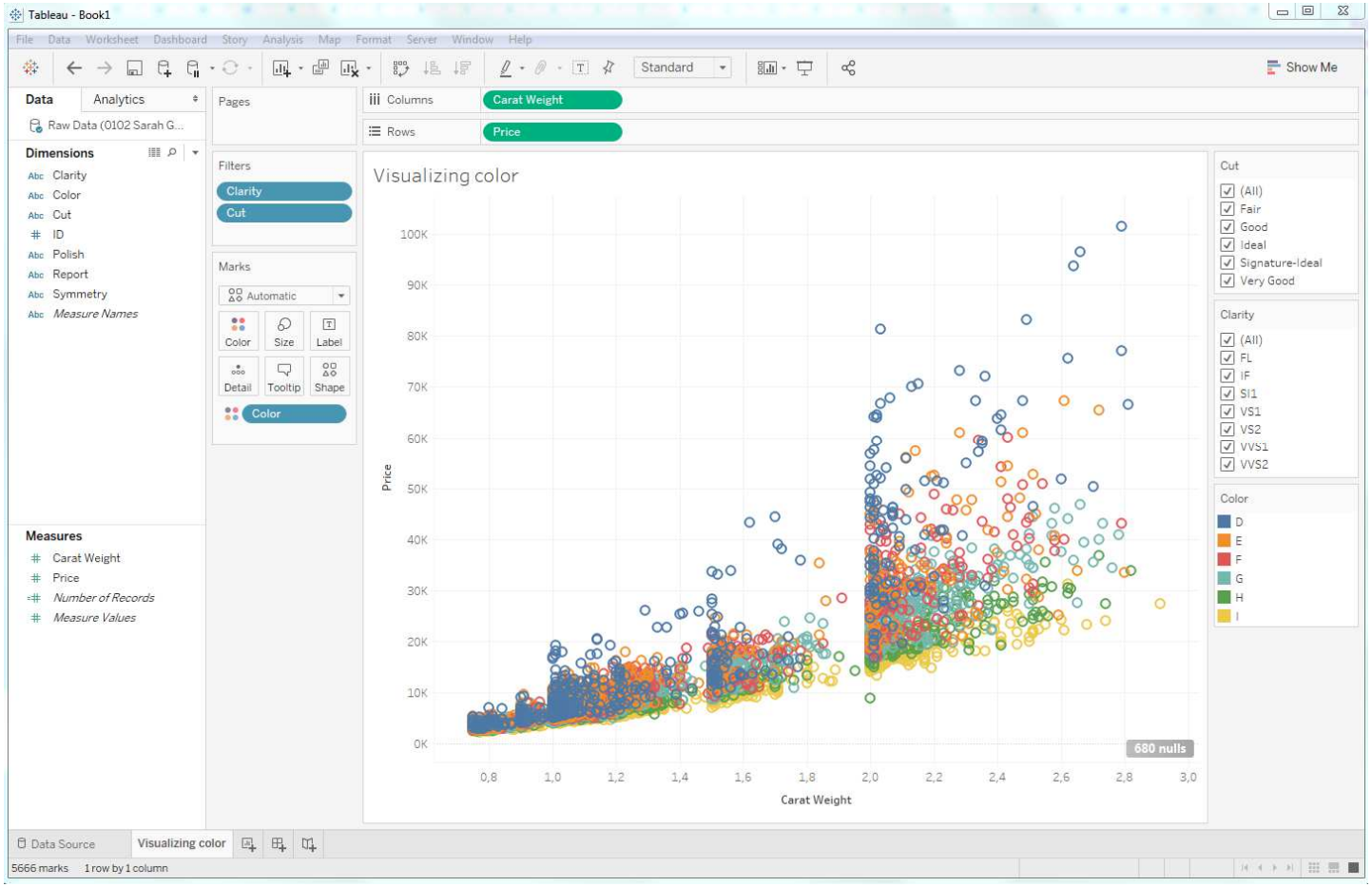
Sarah “color” visual in Tableau



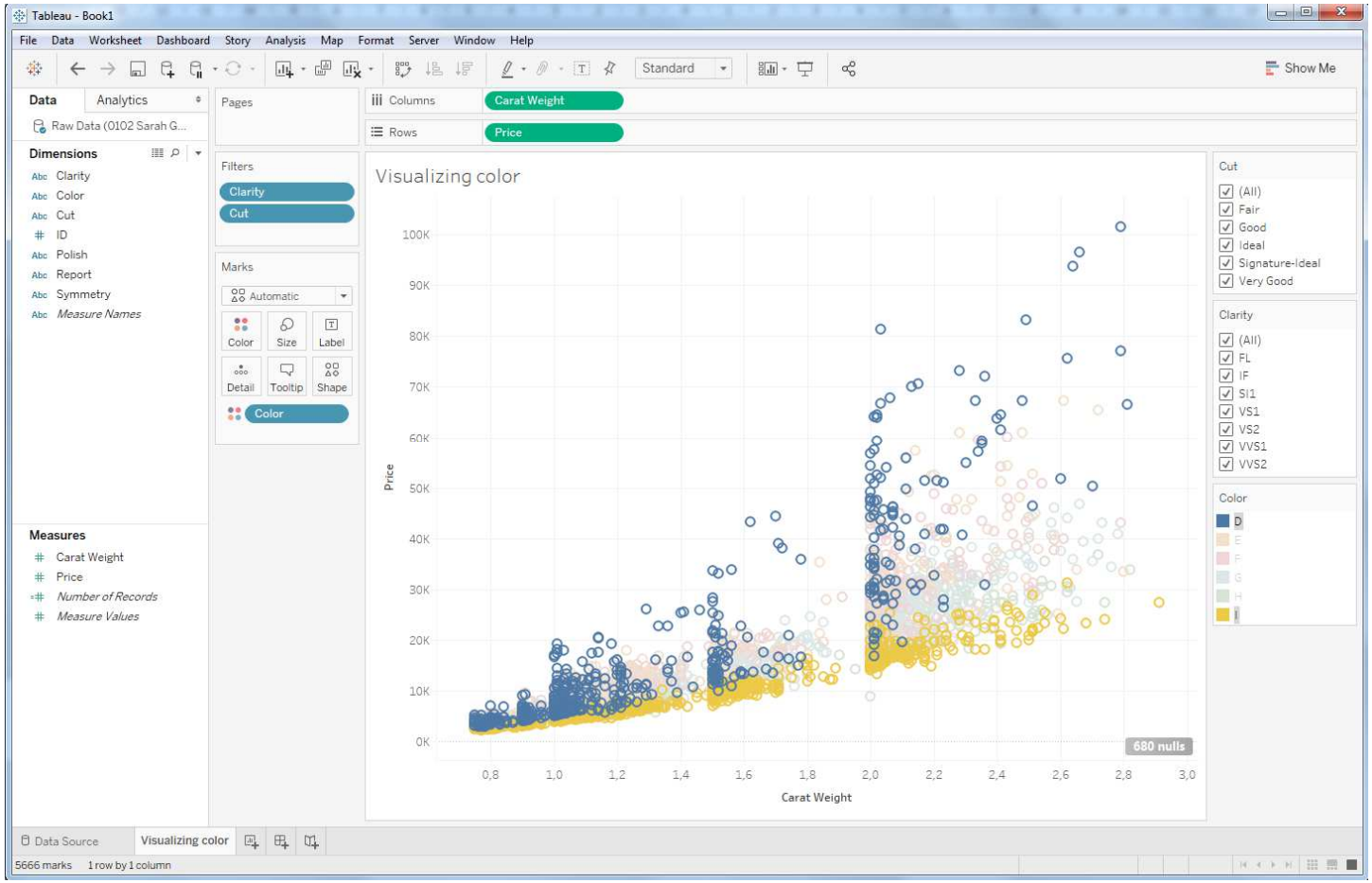
Sarah “color” visual in Tableau



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**

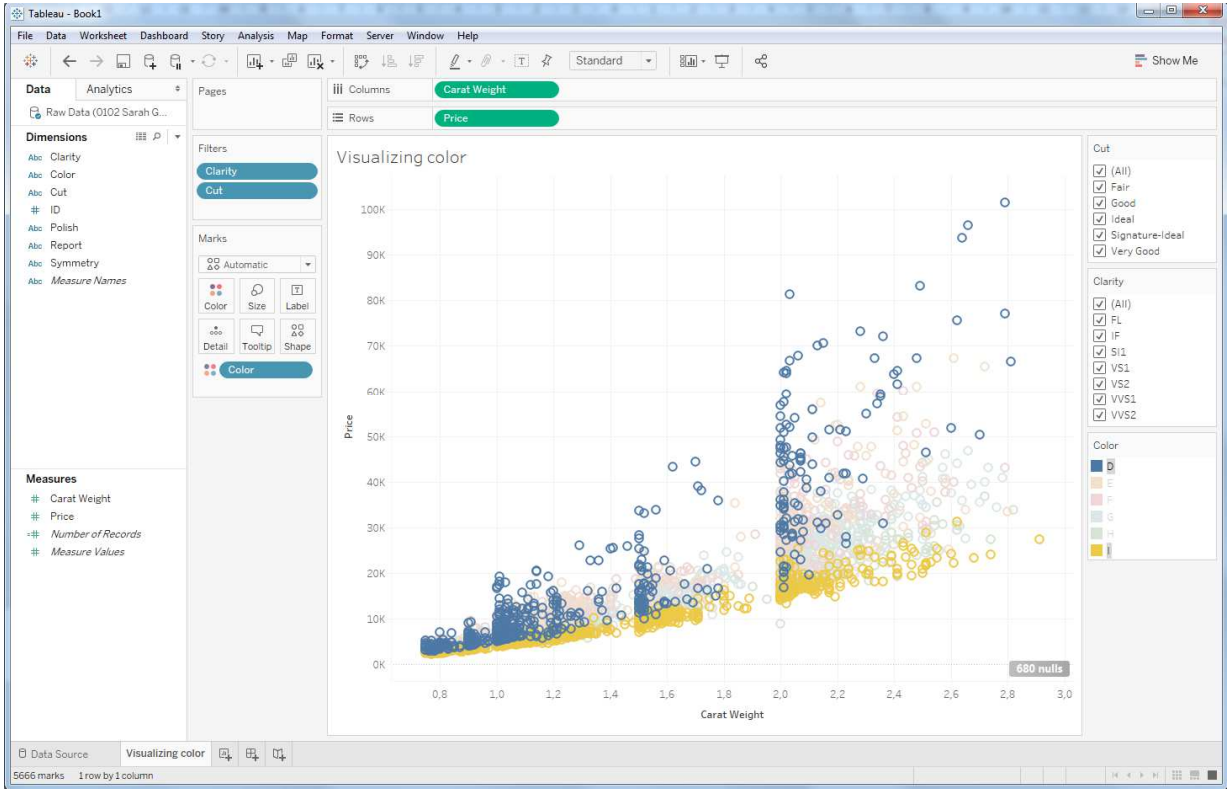


Tableau on your own!

- 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

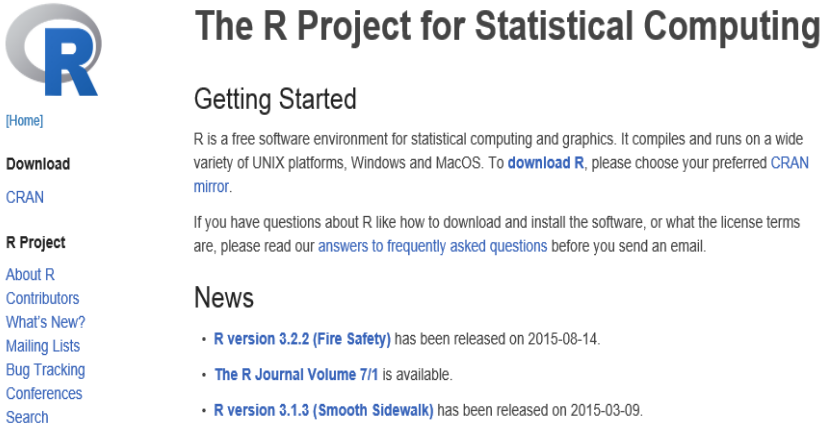
Next: going beyond Excel

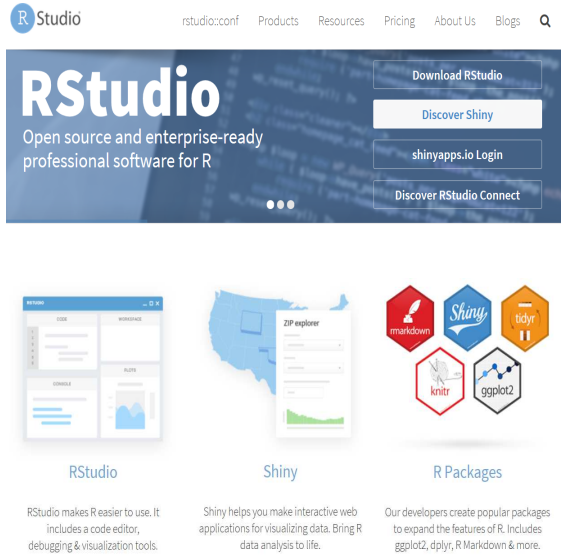
- Intro to data analyses in R
  - What is R?
  - Why R?
  - "How to" R?: By following a simple code for the Sarah's case you just did in Excel
- Session 0304: time series models in R
  - How? Again, by following a simple code that I will provide
- Assignment 1, Yahoo/Tumblr case: combining Excel and 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





- 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
```

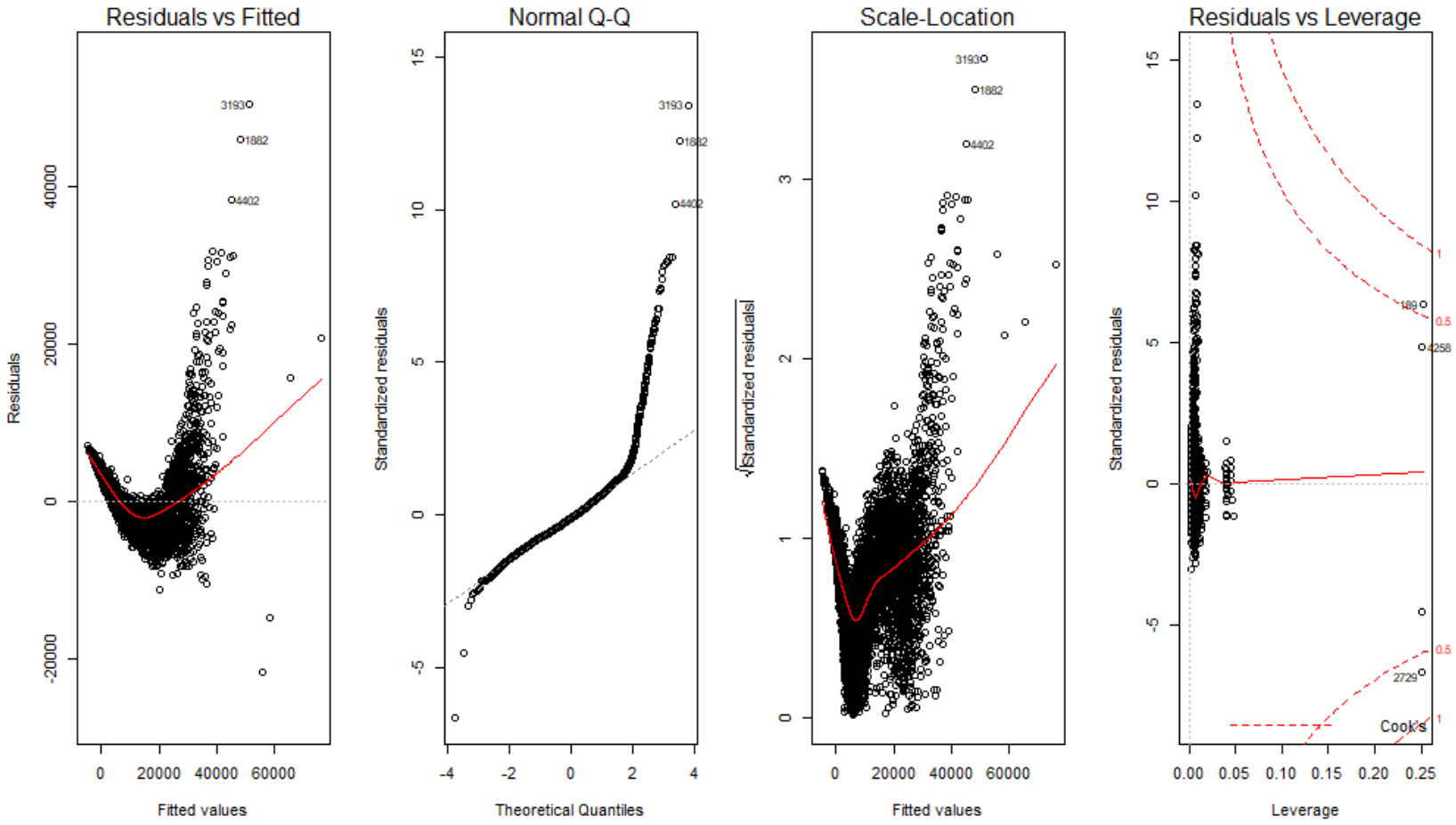
Download two files from course portal:

- CSV Data: 0102 CSV data -- Sarah Gets a Diamond.csv
- R code: 0102 R Code -- Sarah Gets a Diamond -- linear regression.R

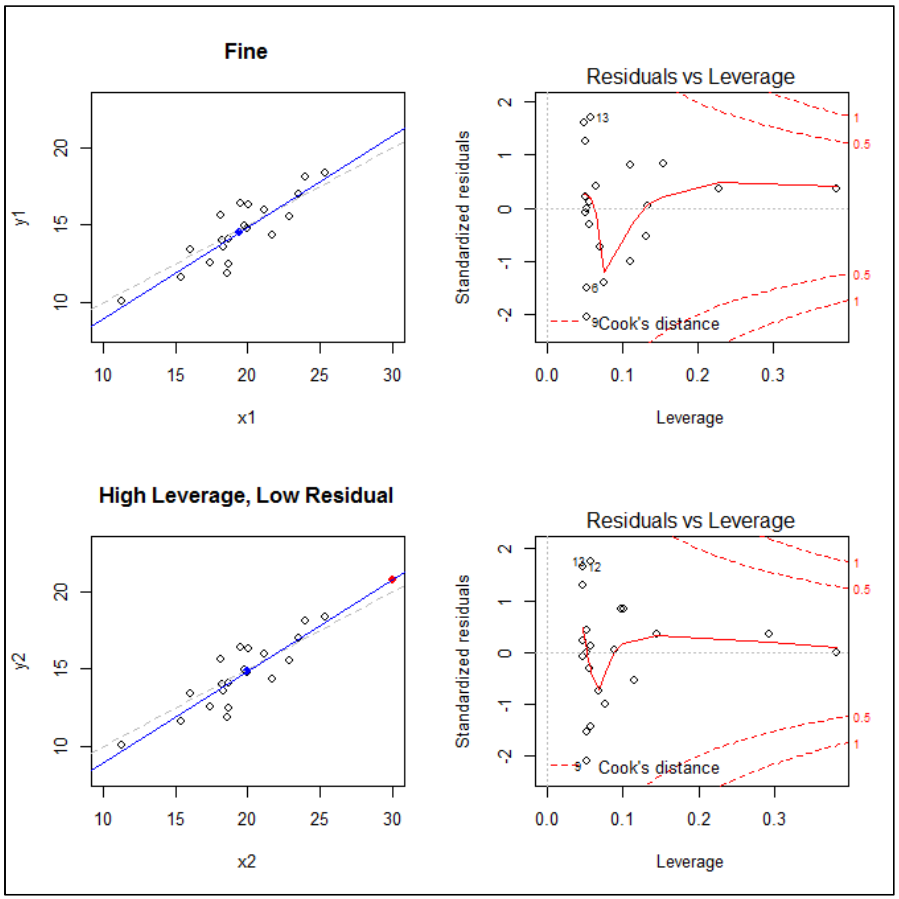
Open code in RStudio, and lets proceed from there step-by-step

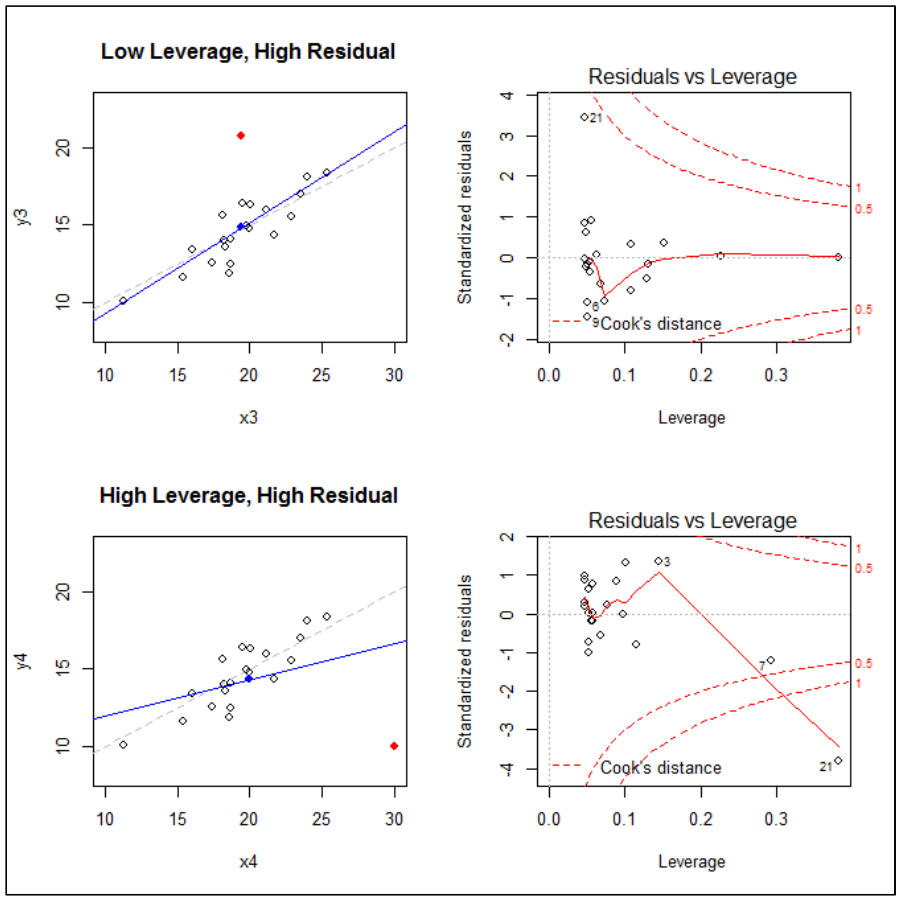
Regression Diagnostic Plots

Homoscedasticity or Heteroscedasticity?



Outliers: Residuals *versus* Leverage, "Bill Gates plot"





https://stats.stackexchange.com/questions/58141/interpreting-plot-lm

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

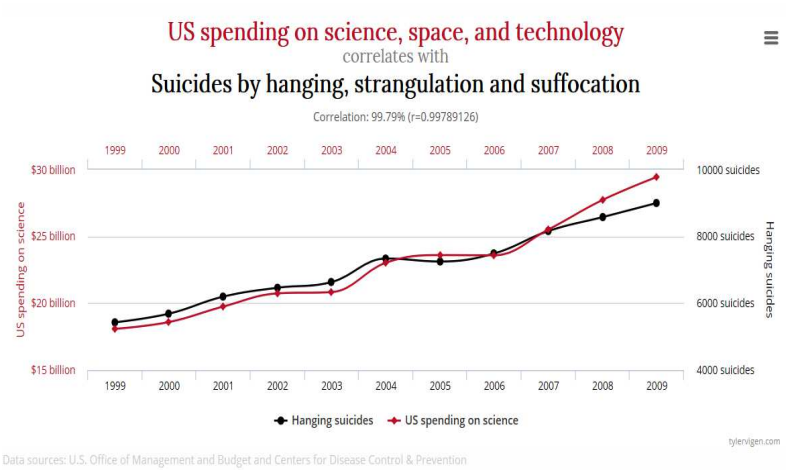
- Recall: What was the task? To predict the prices of new diamonds, for which we know the features (Xs) but not the prices (Y). How can we recreate this task using the current data on 6000 diamonds?
- Main idea: set a portion of the data (e.g., 1000) as a "holdout" ("testing") sample. Build a model on 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)

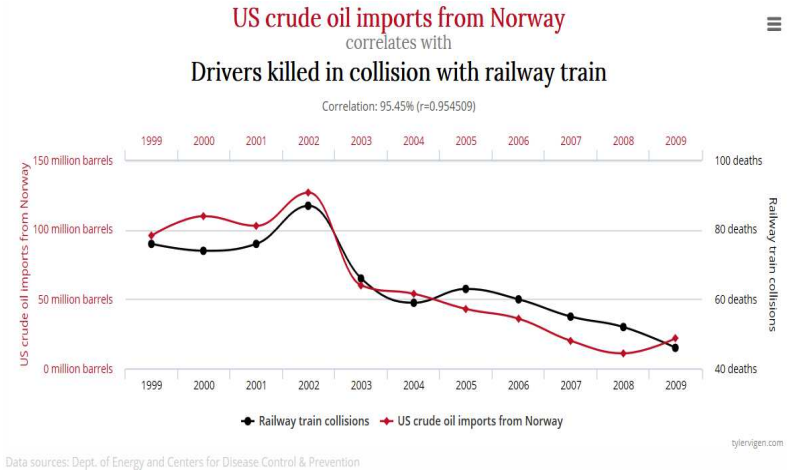
Cross-fold validation / holdout


```
diamond.data.testing<-subset(diamond.data, (ID>=5001 & ID<=6000)) #withold 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

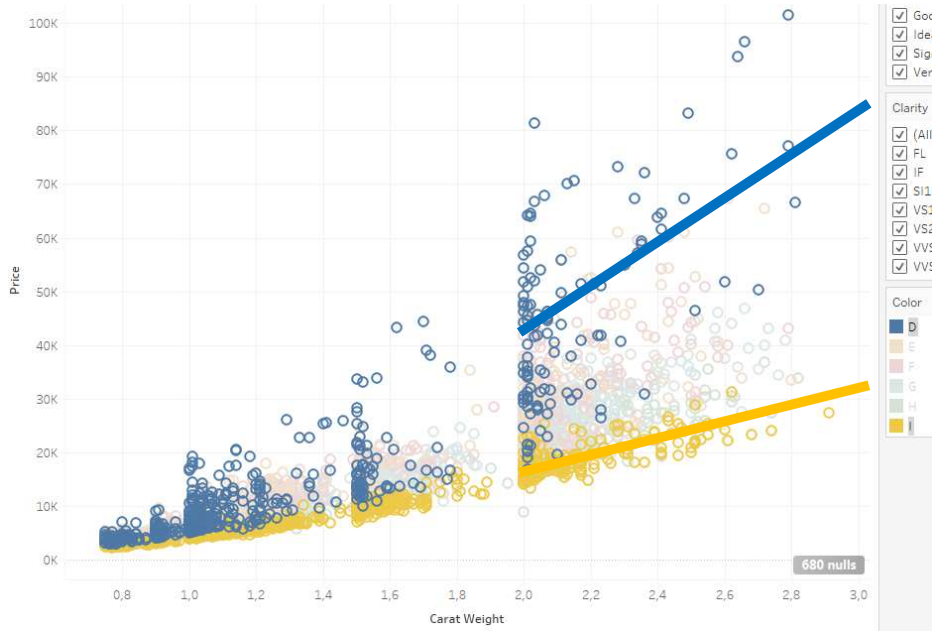
Source: http://tylervigen.com/spurious-correlations

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            5.018e-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       7.817e-02  9.275e-03   8.428  < 2e-16 ***
ColorE            -8.583e-02  5.381e-03 -15.951  < 2e-16 ***
ColorF            -1.301e-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(Price) = 9,446 + 1.987 * log(C.W) – 0,0858*1 = 9.3602 +1.987* log(C.W)
  - If Color = G, then Model is: log(Price) = 9,446 + 1.987 * log(C.W) – 0,2192*1 = 9.2268 +1.987* log(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                       -0.109661   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.[illegible]   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(Price) = 9,353 + 2.14 * log(C.W) – 0,0667*1 – 0,1077 * log(C.W) = 9.286 + 2.032 * log(C.W)
  - Color = G, Model: log(Price) = 9,353 + 2.14 * log(C.W) – 0,1963*1 – 0,1535 * log(C.W) = 9.1594 + 1.987 * log(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 ± ~7.5%! More interactions can improve to ~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]

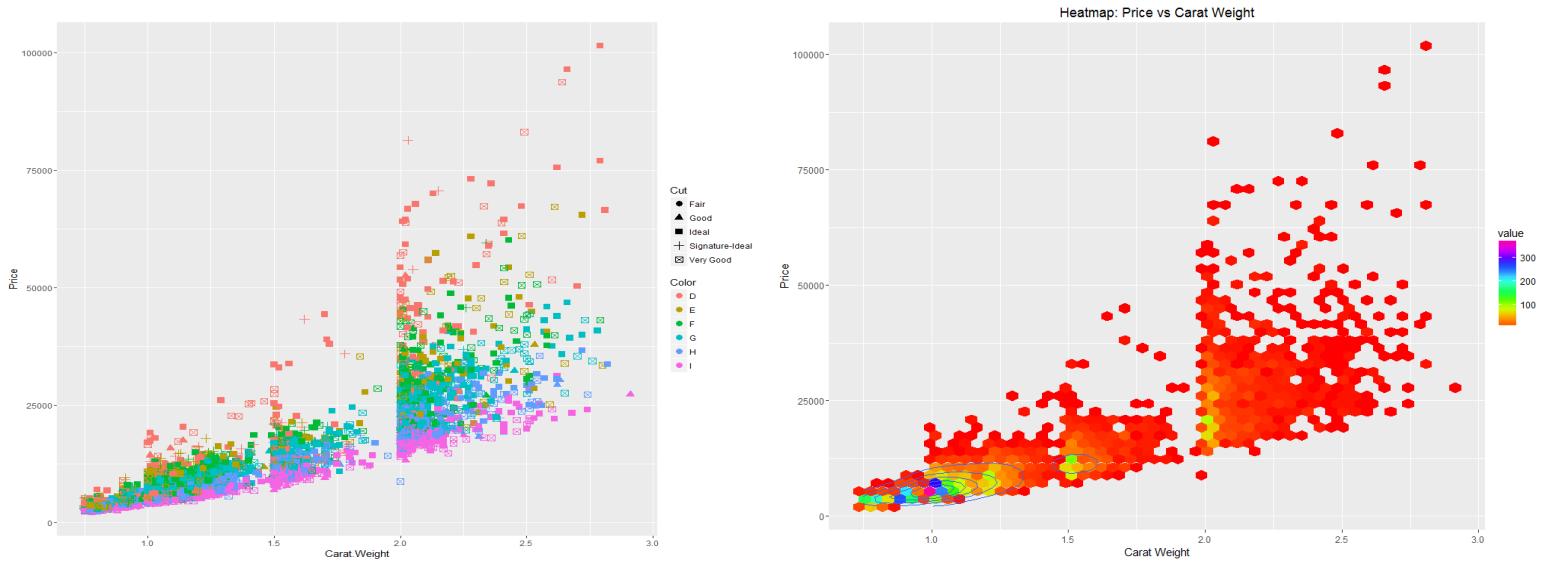
- Sessions 3-4: Time-series analyses
  - Same as today, R code will be provided and we will work through it learning the underlying methods, how to code them, and how to use them

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 i
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.rati
= 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]