Data Visualization

CentraleDigitalLab
@LaPlateforme_

Nicolás Benjamin Ocampo

Ph.D. candidate in Computer Science at the
Wimmics team - INRIA/UCA

nicolas.ocampo@inria.fr

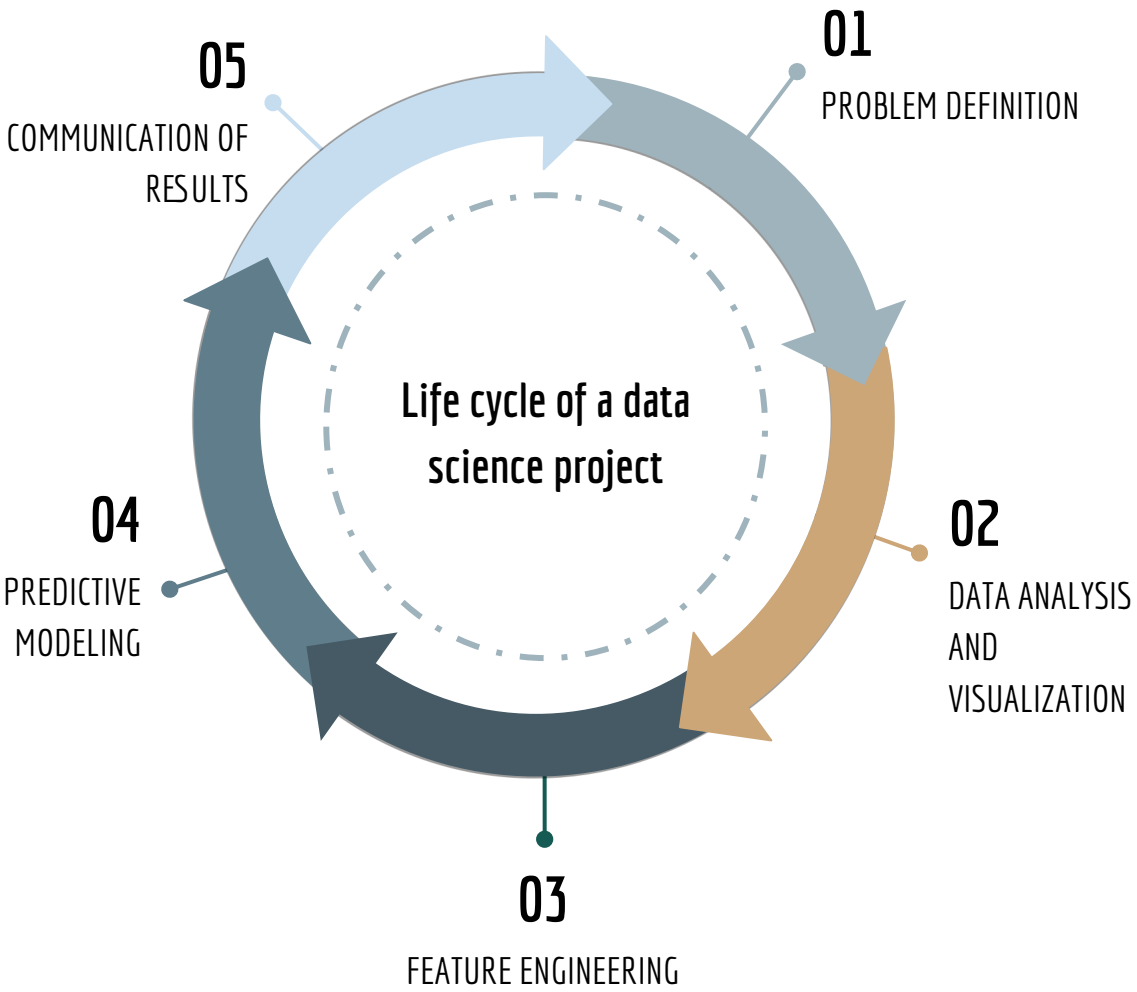
What is Data Science?

Data Analysis	Data Science	Machine Learning
Needs of **concrete questions**	Needs of a **problematic** in a domain	Needs of a **task** and a **dataset**.
Explains data to take a **future decision**	Aims to develop a **product based on data**	**Optimizes** a metric that measures **performance**
Guided by the data **analyst**	Guided by the **interpretation** of the data	Guided by the **model theory**
Detects **superficial patterns**	Highlights **deep patterns**	Detects **deep patterns**

Endless lifecycle

In this course we'll see concepts related to:

1. Statistical and visualization tools for the **step 02**.
2. Statistical methods to understand results for the **step 04**.
3. Visualization and effective communication for the **step 05**.



What if we learn Machine Learning without data analysis and visualization?

- We **don't know what to model** unless we are told to.
- We **can't understand** the **impact** of our results
  - Long term impact usually given by **bias**, **unfairness**, or **information filtering**.
  - Impact given by **business metrics**.
- We waste so much time and effort developing **models that don't answer our questions**.

What if we learn Data Science without Machine Learning?

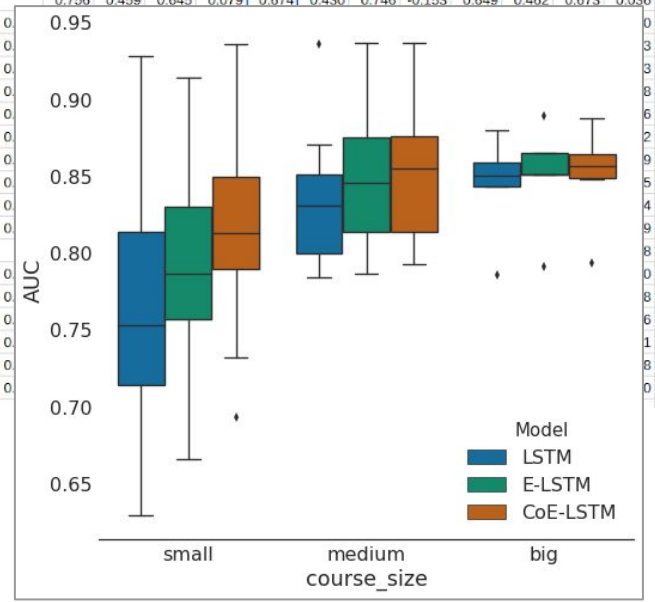
- We are **limited** to **simple analysis**. Or we use models without understanding them.
- Use of machine learning **models that are inappropriate for our dataset**. For example models that don't work well on categorical data.
- We **spend so much time optimizing a model** since we don't know how to properly do it.
- We can **only detect superficial patterns**.

Visualization in Data Science

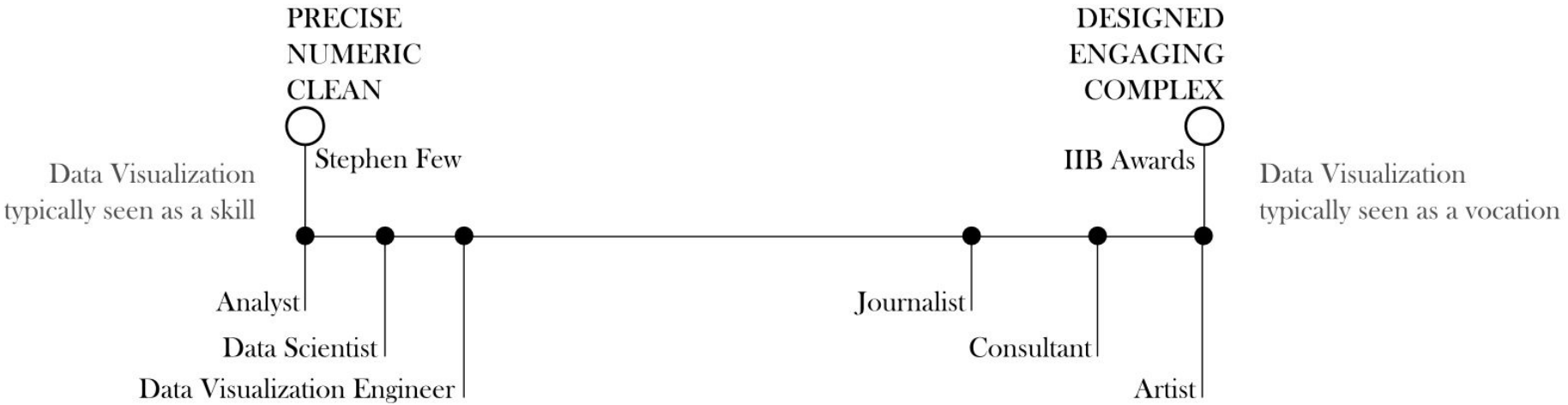
- Data visualization is for communication!
- There’s a sender, receiver, a message, and a communication channel. All these factors affect the communication process.
- Data visualization helps to define our message and determine how the receiver will interpret it.
- Depending on the visualization it might improve or complicate the communication.

	A	B	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL
1					Optimizer			Config						Dataset						Metrics COURSE 1				Metrics COURSE 6				Metrics COURSE 16				Metrics COURSE 21				
2	Results	Log	Trucar	Diff	Optim	LR	Cell	batch_	dropc	epoch	embedc	hidder	max_s	Filter	Merge	Pretra	Fineti	AUC R	rmse	AUC R	rmse	Accura	R2	AUC R	rmse	Accura	R2	AUC R	rmse	Accura	R2	AUC R	rmse	Accura	R2	Eval all
3																																				
4	LSTM																																			
5	../results/kddcup/lstm/pre	14648	N					100	0.3	500	-	100	50	> 5	Y	-	-			0.787	0.388	0.795	0.278													
6	/home/mteruel/edm/resul	14662	N					100	0.3	500	-	100	50	> 5	Y	-	-			0.879	0.359	0.831	0.439	0.794	0.345	0.850	0.275	0.605	0.533	0.672	-0.308	0.498	0.456	0.770	-0.431	
7	/home/mteruel/edm/resul	14663	N					100	0.3	500	-	50	50	> 5	Y	-	-			0.880	0.359	0.827	0.438	0.802	0.339	0.858	0.296	0.656	0.511	0.685	-0.183	0.597	0.414	0.800	-0.199	
8	/home/mteruel/edm/resul	14664	N					100	0.3	500	-	50	100	> 5	Y	-	-	0.814	0.374	0.881	0.361	0.822	0.431	0.804	0.336	0.864	0.315	0.666	0.490	0.713	-0.108	0.626	0.405	0.795	-0.094	
9	../results/kddcup/lstm/pre	14735	Y					50	0.3	500	-	50	50	> 5	Y	-	-			0.759	0.467	0.617	0.050	0.657	0.415	0.784	-0.053	0.611	0.500	0.613	-0.164	0.634	0.398	0.783	-0.074	
10	../results/kddcup/lstm/pre	14885	N					50	0.3	500	-	50	50	N	Y	-	-	0.837	0.335	0.871	0.341	0.855	0.450	0.842	0.290	0.895	0.369	0.741	0.386	0.796	0.153	0.662	0.418	0.771	-0.195	14939
11			N					100	0.3	500	-	50	100	N	Y	-	-			0.871	0.339	0.851	0.458	0.837	0.289	0.896	0.375	0.748	0.390	0.800	0.138	0.633	0.416	0.797	-0.181	14937
12																																				
13	Embeddings																																			
19	/home/mteruel/edm/resul	14661	N					100	0.3	500	50	50	100	> 5	N	N				0.885	0.352	0.836	0.457	0.814	0.335	0.856	0.326	0.703	0.466	0.723	0.006	0.724	0.393	0.797	-0.047	
20	/home/mteruel/edm/resul	14667	N					100	0.3	500	50	50	50	> 5	N	N				0.880	0.359	0.829	0.439	0.813	0.332	0.867	0.330	0.687	0.478	0.705	-0.053	0.668	0.405	0.810	-0.142	
21	../results/kddcup/embedc	14706	Y					50	0.2	500	50	50	20	> 5	N	Y	Y			0.728	0.481	0.614	-0.010	0.683	0.390	0.798	0.069	0.641	0.467	0.687	0.004	0.599	0.404	0.773	-0.113	
22	../results/kddcup/embedc	14716	Y					100	0.3	500	50	50	50	> 5	Y	N				0.756	0.459	0.645	0.079	0.674	0.430	0.746	-0.153	0.649	0.462	0.673	0.036	0.534	0.422	0.770	-0.207	
23	../results/kddcup/embedc	14821	N					100	0.3	500	50	50	100	> 5	N	N		0.830	0.363	0.884	0.357	0.830	0.443	0.810	0.334	0.868	0.319	0.740	0.455	0.727	0.050	0.685	0.379	0.815	0.029	
24	../results/kddcup/embedc	14823	N					100	0.3	500	50	50	100	> 5	Y	N		0.834	0.361	0.884	0.358	0.825	0.442	0.813	0.333	0.865	0.322	0.731	0.456	0.715	0.043	0.688	0.375	0.831	0.051	
25	../results/kddcup/embedc	14828	N					50	0.2	500	50	100	20	> 5	Y	Y	Y	0.808	0.378	0.871	0.365	0.816	0.417	0.801	0.344	0.846	0.279	0.701	0.479	0.711	-0.053	0.575	0.407	0.813	-0.122	
26	../results/kddcup/embedc	14832	N					50	0.3	500	50	100	200	> 5	Y	Y	Y	0.818	0.370	0.879	0.359	0.825	0.436	0.788	0.340	0.861	0.296	0.707	0.471	0.711	-0.018	0.672	0.388	0.811	-0.021	
27	../results/kddcup/embedc	14858	N					50	0.3	500	50	100	200	> 5	Y	Y	N	0.825	0.365	0.875	0.365	0.827	0.420	0.806	0.338	0.854	0.302	0.714	0.449	0.739	0.076	0.650	0.383	0.824	0.007	
28	../results/kddcup/embedc	14873	N					50	0.3	500	20	100	200	> 5	Y	Y	Y	0.831	0.363	0.879	0.361	0.827	0.433	0.815	0.335	0.857	0.316	0.733	0.438	0.742	0.122	0.675	0.381	0.826	0.020	
29	../results/kddcup/embedc	14875	N					50	0.3	500	20	100	200	> 5	Y	Y	N	0.835	0.362	0.880	0.361	0.822	0.432	0.815	0.340	0.846	0.293	0.722	0.445	0.736	0.089	0.712	0.371	0.823	0.069	
30	../results/kddcup/embedc	14877	N					50	0.3	500	20	50	200	> 5	Y	Y	N	0.841	0.360	0.880	0.364	0.818	0.423	0.819	0.334	0.859	0.320	0.753	0.432	0.735	0.145	0.715	0.372	0.826	0.065	14941
31	../results/kddcup/embedc	14886	N					100	0.3	500	50	50	100	N	Y	N		0.850	0.330	0.887	0.338	0.853	0.461	0.850	0.291	0.895	0.366	0.783	0.379	0.807	0.184	0.699	0.396	0.813	-0.069	14922 +14
32	../results/kddcup/embedc	14887	N					50	0.3	500	20	100	200	N	Y	Y	Y	0.846	0.328	0.879	0.339	0.852	0.458	0.843	0.289	0.896	0.373	0.788	0.371	0.820	0.219	0.652	0.399	0.804	-0.089	x
33			N					50	0.3	500	20	50	200	N	Y	Y	N			0.881	0.339	0.851	0.456	0.843	0.290	0.892	0.366	0.804	0.362	0.831	0.258	0.740	0.379	0.814	0.020	14940
41	../results/kddcup/embedc	15167	N		adam	0.01	gru	100	0.3	500	20	50	200	N	Y	Y	Y	0.818	0.343	0.890	0.334	0.854	0.472	0.839	0.294	0.896	0.349	0.732	0.391	0.807	0.130	0.576	0.437	0.787	-0.305	
42	../results/kddcup/embedc	15177	N		adam	??	gru	100	0.3	500	20	50	200	N	Y	Y	Y	0.811	0.345	0.886	0.336	0.858	0.466	0.823	0.304	0.890	0.307	0.683	0.438	0.773	-0.088	0.627	0.419	0.801	-0.199	
43	../results/kddcup/embedc	15235	N		adam	0.01	lstm	100	0.3	500	50	50	100	N	Y	N		0.811	0.347	0.882	0.345	0.841	0.439	0.814	0.304	0.892	0.305	0.684	0.425	0.799	-0.026	0.621	0.411	0.810	-0.153	
44	../results/kddcup/embedc	15236	N		adam	0.01	lstm	50	0.3	500	20	100	200	N	Y	Y	Y	0.812	0.345	0.884	0.340	0.850	0.452	0.826	0.300	0.892	0.325	0.700	0.405	0.800	0.071	0.589	0.431	0.786	-0.265	
45	../results/kddcup/embedc	26285	N					100	0.3	500	20	20	300	N	Y	N		0.853	0.325	0.881	0.335	0.854	0.469	0.841	0.291	0.895	0.362	0.790	0.369	0.814	0.228	0.675	0.397	0.796	-0.076	
46	../results/kddcup/embedc	26286	N					100	0.3	500	20	20	300	N	Y	Y	Y	0.857	0.322	0.883	0.334	0.857	0.474	0.845	0.288	0.895	0.375	0.783	0.364	0.826	0.250	0.757	0.371	0.820	0.062	

	A	B	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL
1					Optimizer			Config						Dataset						Metrics COURSE 1				Metrics COURSE 6				Metrics COURSE 16				Metrics COURSE 21				
2	Results	Log	Trucar	Diff	Optim	LR	Cell	batch_	dropc	epoch	embedc	hidder	max_s	Filter	Merge	Pretra	Fineti	AUC R(	rmse	AUC R(	rmse	Accurac	R2	AUC R(	rmse	Accurac	R2	AUC R(	rmse	Accurac	R2	AUC R(	rmse	Accurac	R2	Eval all	
3																																				
4	LSTM																																			
5	../results/kddcup/lstm/pre	14648	N					100	0.3	500	-	100	50	> 5	Y	-	-			0.787	0.388	0.795	0.278														
6	/home/mteruel/edm/resul	14662	N					100	0.3	500	-	100	50	> 5	Y	-	-			0.879	0.359	0.831	0.439	0.794	0.345	0.850	0.275	0.605	0.533	0.672	-0.308	0.498	0.456	0.770	-0.431		
7	/home/mteruel/edm/resul	14663	N					100	0.3	500	-	50	50	> 5	Y	-	-			0.880	0.359	0.827	0.438	0.802	0.339	0.858	0.296	0.656	0.511	0.685	-0.183	0.597	0.414	0.800	-0.199		
8	/home/mteruel/edm/resul	14664	N					100	0.3	500	-	50	100	> 5	Y	-	-	0.814	0.374	0.881	0.361	0.822	0.431	0.804	0.336	0.864	0.315	0.666	0.490	0.713	-0.108	0.626	0.405	0.795	-0.094		
9	../results/kddcup/lstm/pre	14735	Y					50	0.3	500	-	50	50	> 5	Y	-	-			0.759	0.467	0.617	0.050	0.657	0.415	0.784	-0.053	0.611	0.500	0.613	-0.164	0.634	0.398	0.783	-0.074		
10	../results/kddcup/lstm/pre	14885	N					50	0.3	500	-	50	50	N	Y	-	-	0.837	0.335	0.871	0.341	0.855	0.450	0.842	0.290	0.895	0.369	0.741	0.386	0.796	0.153	0.662	0.418	0.771	-0.195	14939	../
11			N					100	0.3	500	-	50	100	N	Y	-	-			0.871	0.339	0.851	0.458	0.837	0.289	0.896	0.375	0.748	0.390	0.800	0.138	0.633	0.416	0.797	-0.181	14937	../
12																																				
13	Embeddings																																			
19	/home/mteruel/edm/resul	14661	N					100	0.3	500	50	50	100	> 5	N	N				0.885	0.352	0.836	0.457	0.814	0.335	0.856	0.326	0.703	0.466	0.723	0.006	0.724	0.393	0.797	-0.047		
20	/home/mteruel/edm/resul	14667	N					100	0.3	500	50	50	50	> 5	N	N				0.880	0.359	0.829	0.439	0.813	0.332	0.867	0.330	0.687	0.478	0.705	-0.053	0.668	0.405	0.810	-0.142		
21	../results/kddcup/embedc	14706	Y					50	0.2	500	50	50	20	> 5	N	Y	Y			0.728	0.481	0.614	-0.010	0.683	0.390	0.798	0.069	0.641	0.467	0.687	0.004	0.599	0.404	0.773	-0.113		
22	../results/kddcup/embedc	14716	Y					100	0.3	500	50	50	50	> 5	Y	N				0.756	0.459	0.645	0.079	0.674	0.430	0.746	-0.153	0.649	0.462	0.673	0.036	0.534	0.422	0.770	-0.207		
23	../results/kddcup/embedc	14821	N																													0.685	0.379	0.815	0.029		
24	../results/kddcup/embedc	14823	N																													0.688	0.375	0.831	0.051		
25	../results/kddcup/embedc	14828	N																													0.575	0.407	0.813	-0.122		
26	../results/kddcup/embedc	14832	N																													0.672	0.388	0.811	-0.021		
27	../results/kddcup/embedc	14858	N																													0.650	0.383	0.824	0.007		
28	../results/kddcup/embedc	14873	N																													0.675	0.381	0.826	0.020		
29	../results/kddcup/embedc	14875	N																													0.712	0.371	0.823	0.069		
30	../results/kddcup/embedc	14877	N																													0.715	0.372	0.826	0.065	14941	../
31	../results/kddcup/embedc	14886	N																													0.699	0.396	0.813	-0.069	14922 +14	../
32	../results/kddcup/embedc	14887	N																													0.652	0.399	0.804	-0.089	x	
33			N																													0.740	0.379	0.814	0.020	14940	../
41	../results/kddcup/embedc	15167	N																													0.576	0.437	0.787	-0.305		
42	../results/kddcup/embedc	15177	N																													0.627	0.419	0.801	-0.199		
43	../results/kddcup/embedc	15235	N																													0.621	0.411	0.810	-0.153		
44	../results/kddcup/embedc	15236	N																													0.589	0.431	0.786	-0.265		
45	../results/kddcup/embedc	26285	N																													0.675	0.397	0.796	-0.076		
46	../results/kddcup/embedc	26286	N																													0.757	0.371	0.820	0.062		



```
enrollment_id,username,course_id
1,9Uee7oEuuMmgPx2IzPfFkWgkHZyPbWr0,DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
3,1qXC7Fjbwp66GPQc6pHLfEuO8WKozxG4,7GRhBDsirIGkRZBtSMEzNTyDr2JQm4xx
4,FIHlppZyoq8muPbdVxS44gfvceX9zvU7,DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
5,p1Mp7WkVfzUijX0peVQKSHbgd5pXyl4c,7GRhBDsirIGkRZBtSMEzNTyDr2JQm4xx
6,dpK33RH9yepUAnyoywRwBt1AJzxGlaja,AXUJZGmZ0xaYSWazu8RQ1G5c76ECT1Kd
7,I1KwJ6EdCZnEPLfC8Q7yWpIkLOHn7h02,7GRhBDsirIGkRZBtSMEzNTyDr2JQm4xx
9,J1oRHoSJOInehnrxVdh32dK7QnDuCJWo,DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
12,9tsGjrRgtMZ6V7yrAOyfOQPZHaltDHAp,DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
13,hDbSkVrFRj9Ryk3c5E1JYJQLyxm4jLRb,5X6FeZozNMgE2VRi3MJYjkkFK8SETtu2
14,XOhIczT5nEeO52jMq1vN7QziDk8L2jnI,DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
16,mPSPvu82Gr17tV9GJ95bDC7exvsVnwDE,DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
18,b0Hk5D3sJulvyuC4JEm5kvAvOLAxswgQ,DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
20,BoK7CAUaCFqnLgmWLxeOHg8YkXUSeCtc,DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
22,dPBUV0FPFjTZZK079rPAeq0WXhW4DUkF,7GRhBDsirIGkRZBtSMEzNTyDr2JQm4xx
23,BoK7CAUaCFqnLgmWLxeOHg8YkXUSeCtc,AXUJZGmZ0xaYSWazu8RQ1G5c76ECT1Kd
26,vcAiZWU2sfUKO0mnfjDwm0iTzACrKr78,DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
28,BoK7CAUaCFqnLgmWLxeOHg8YkXUSeCtc,TAYxxh39I2LZnftBpL0LfF2NxzrCKpkx
30,JPkczYOxyoDZBjwZAAQHmjpSvnPQzwV0,DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila
```



Credit: Medium Article

Can you identify your roll in this line? Where would you like to be?

Bias in perception

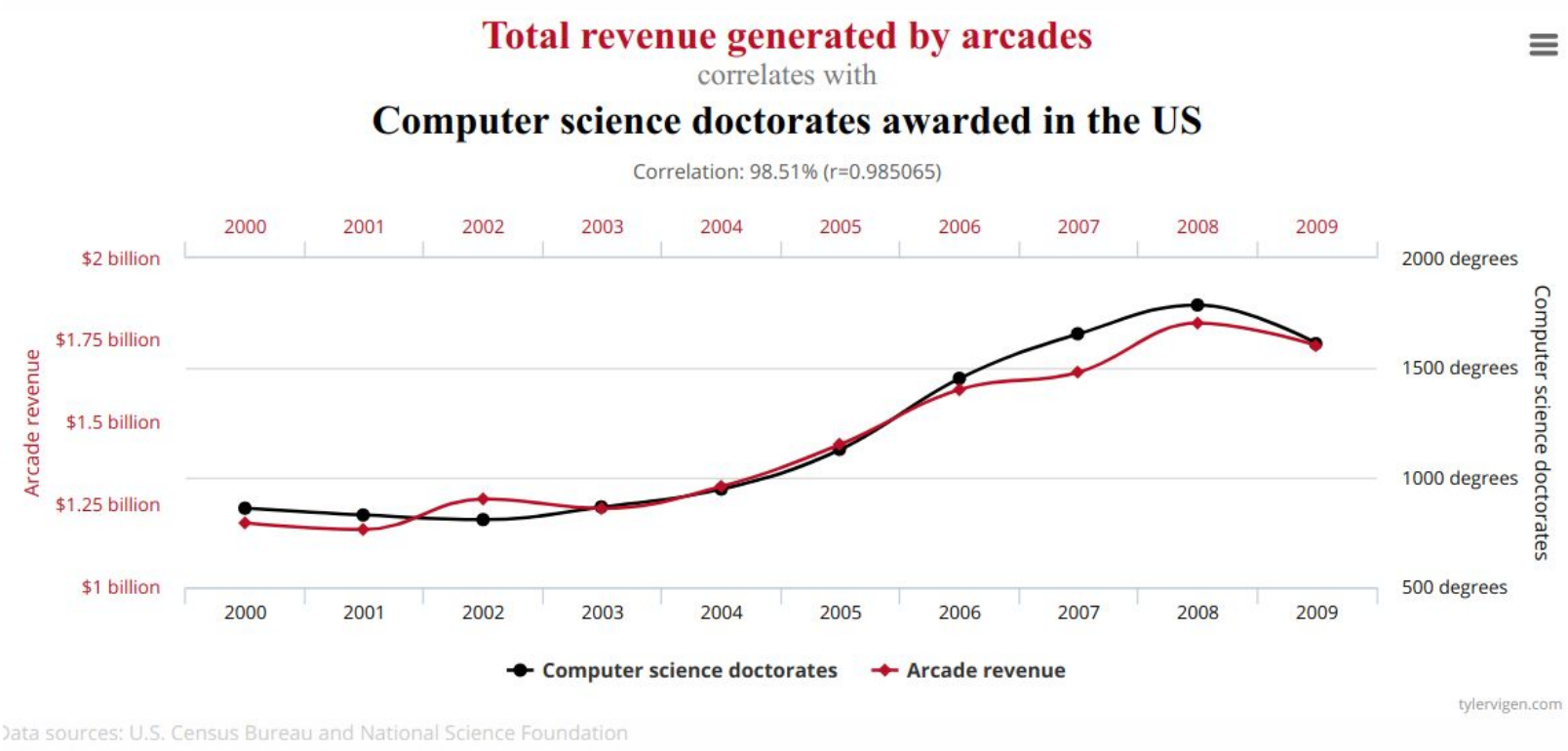
What do we think when we don't stop to analyze?

Patternicity bug!

Tendency to find patterns in objects and to perceive the whole as more than the sum of its parts.

The face of mars

Correlation is not causation



https://www.tylervigen.com/spurious-correlations

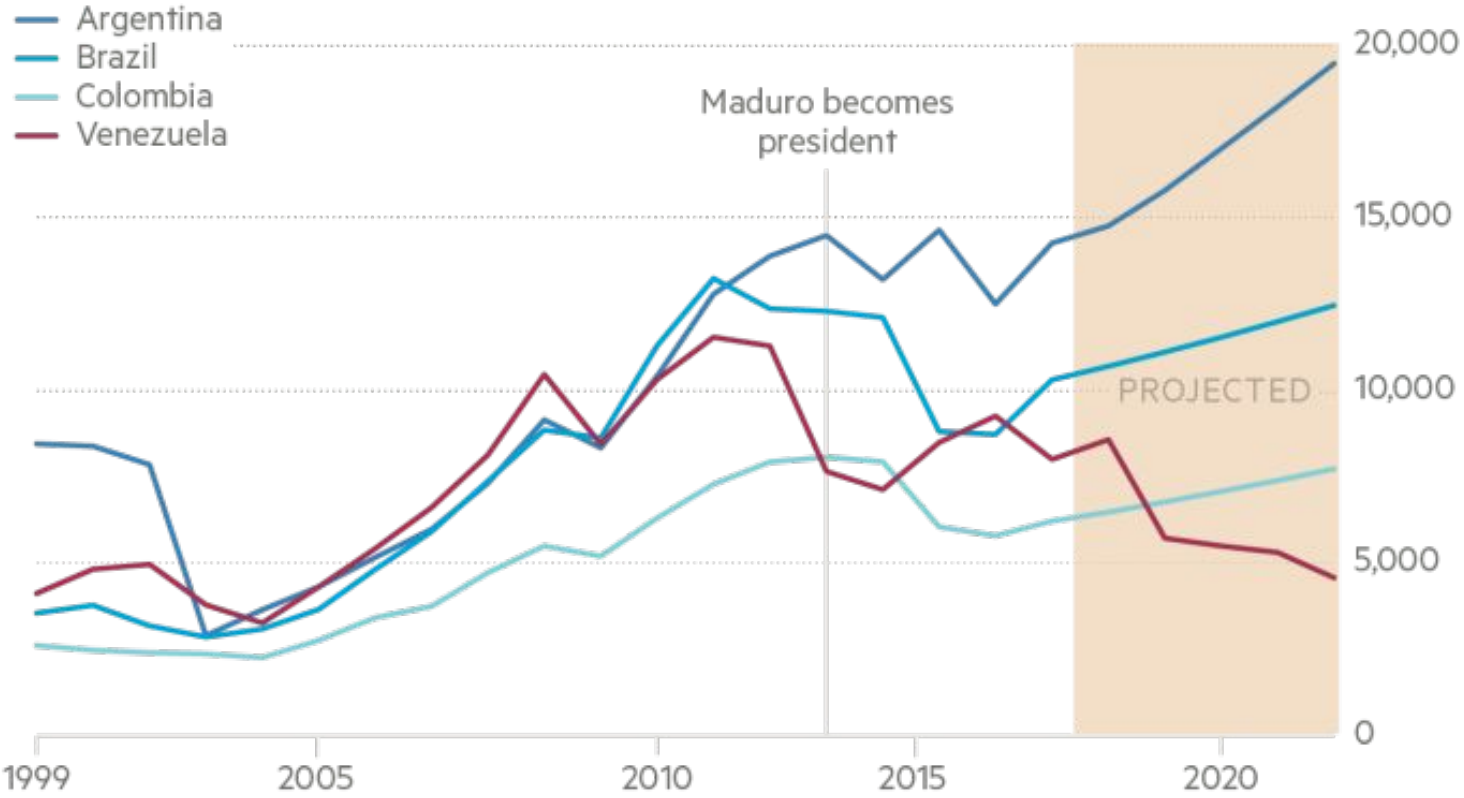
Storytelling bug!

Tendency to find reasons that explain the

presence of patterns in objects

Oil-rich Venezuela will have a lower per-capita GDP than its peers

GDP per capita in current US dollars

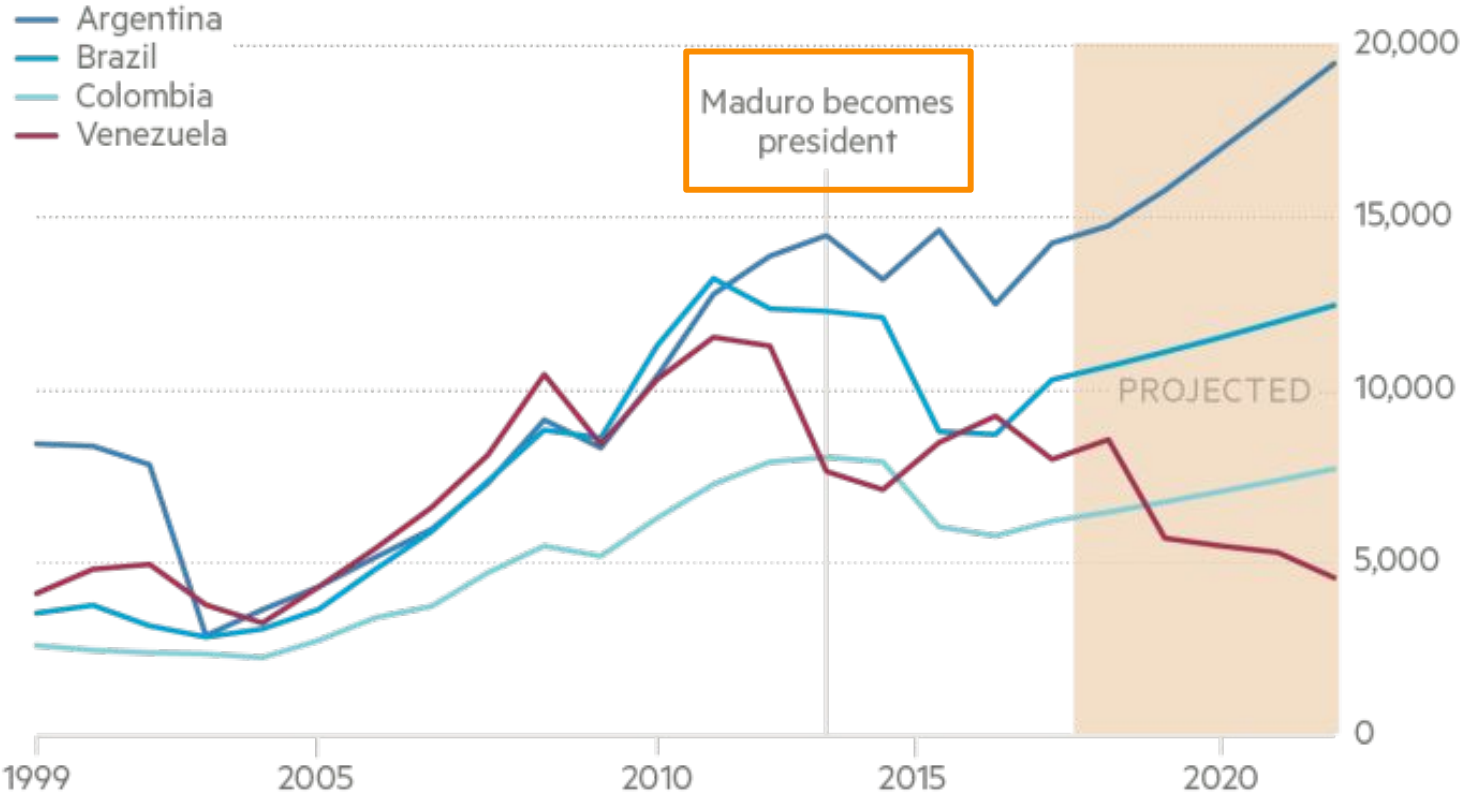


Source: IMF World Economic Outlook Database

FT

Oil-rich Venezuela will have a lower per-capita GDP than its peers

GDP per capita in current US dollars



Source: IMF World Economic Outlook Database

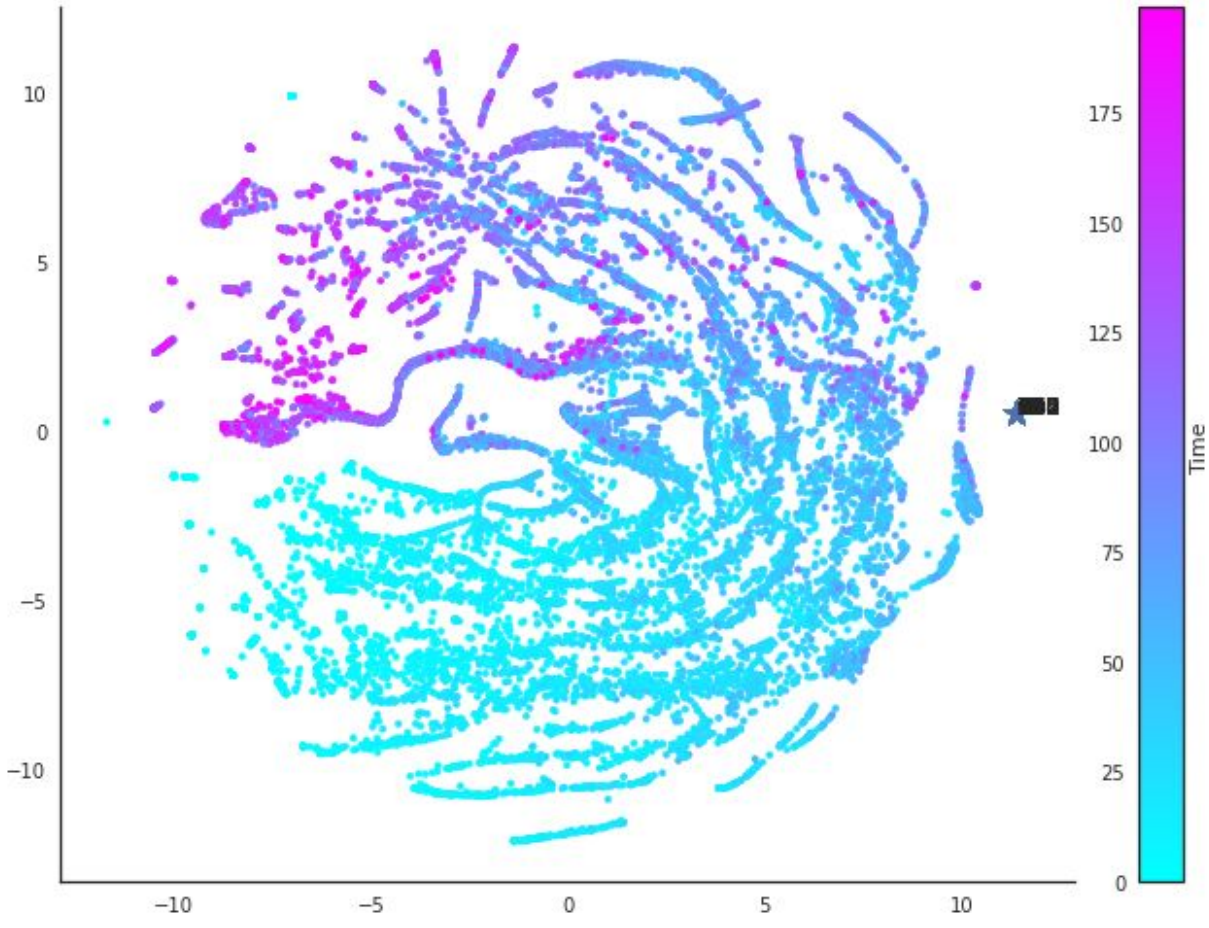
FT

Confirmation bug!

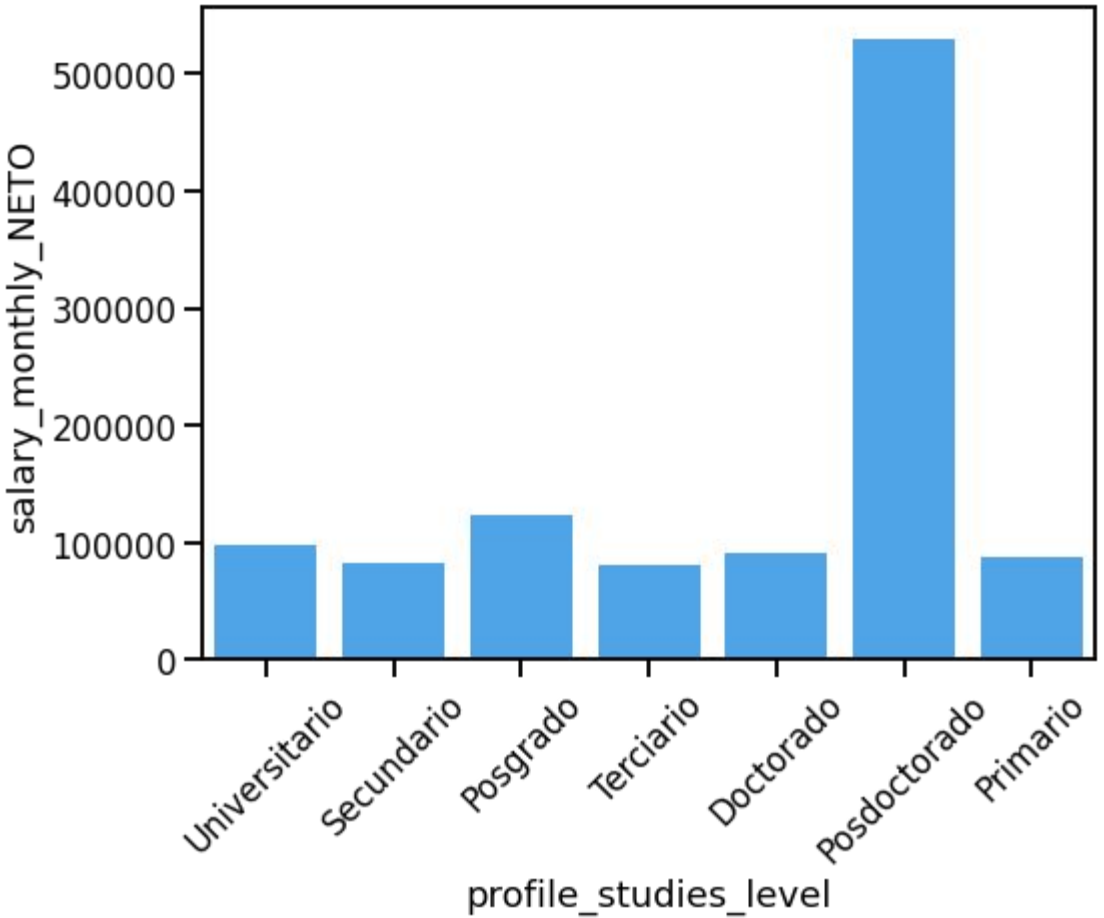
Tendency to believe (more) true the information that supports our existing beliefs.

Are the patterns real?

- Color patterns
- "Worm" patterns



Do we doubt the data or not?



Presentation

Visualization for others

Optimize the communication process

Generate a message that is quickly and faithfully decoded

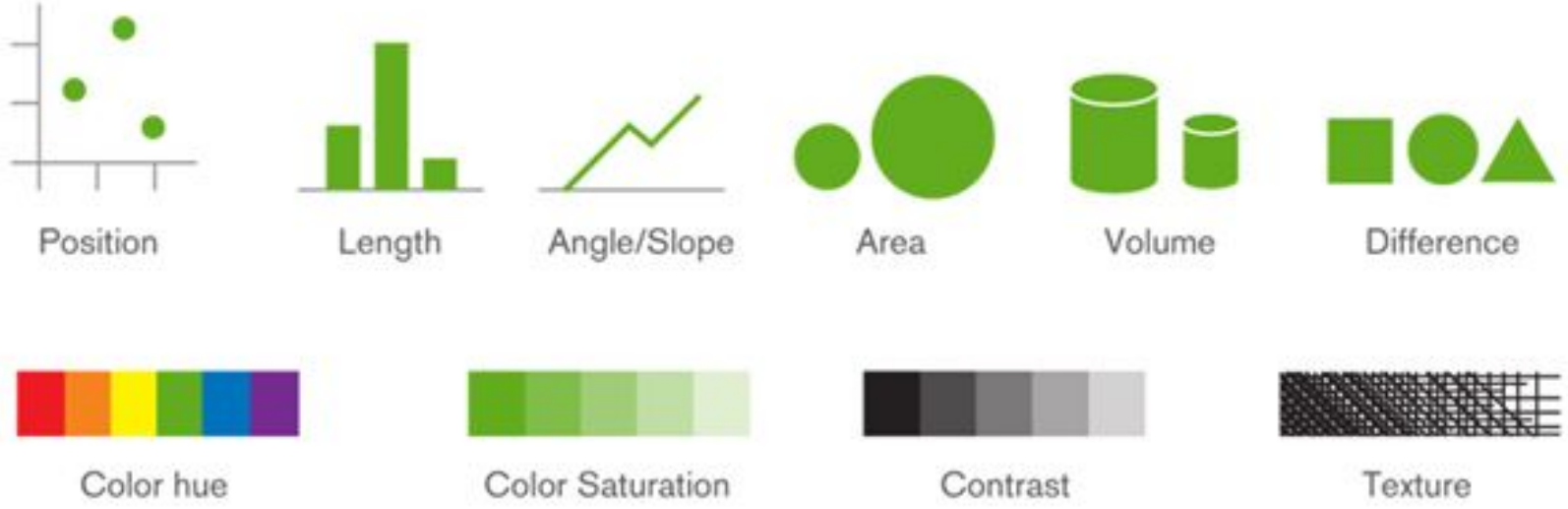
Characteristics of a good visualization

- Honest: represents data that is correct.
- Functional: represents data so that they can be properly interpreted.
- Enlightening: It must show patterns that would not be easily perceived using other media.
- Esthetic
- Informative

Visual Encodings

Data mapping  visual elements

Visual elements



Second Pillar of Mapping Data to Visualizations: Visual Encoding

How to choose the visual elements?

Principle of **consistency**: the properties of the image must correspond to the properties of the data.

- The lie factor

Principle of **ordering by importance**: the most important information must be coded in the most efficient way possible.

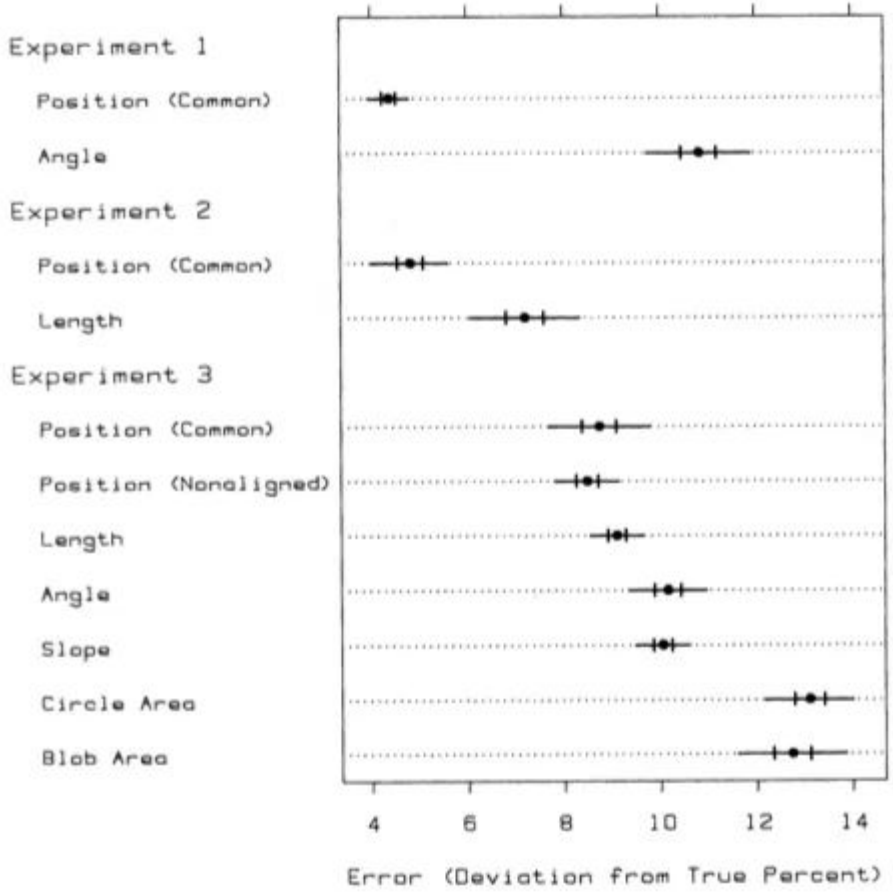
- What is the most important information?
- What are the most effective encodings?

How do we measure the error?

Different encodings allow us to better

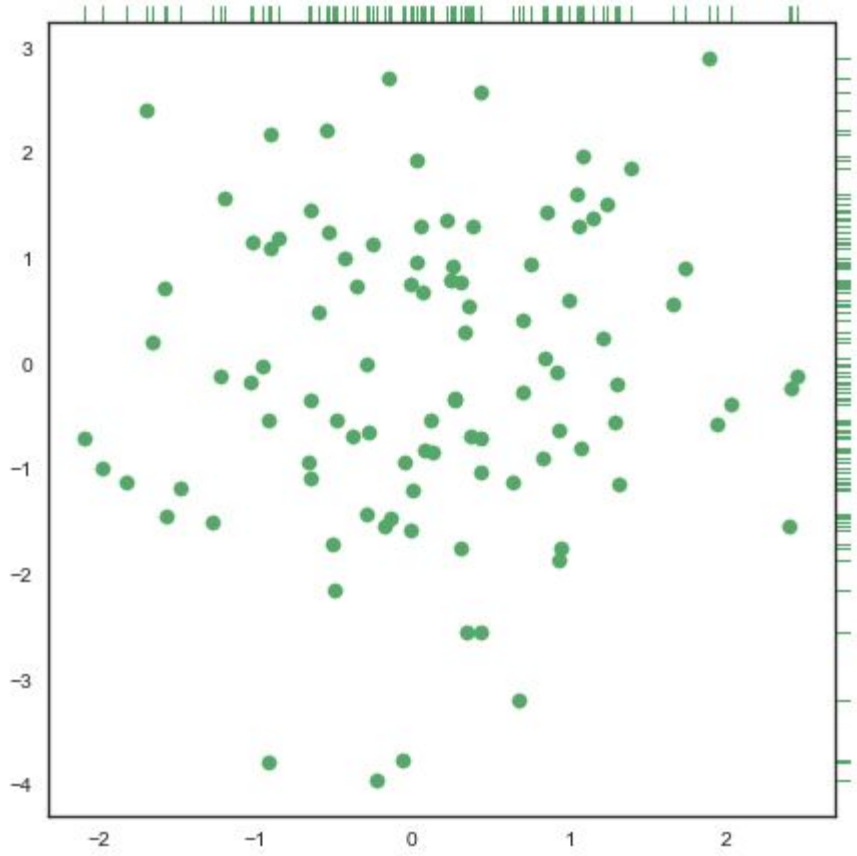
or worse estimate the difference

between two quantities



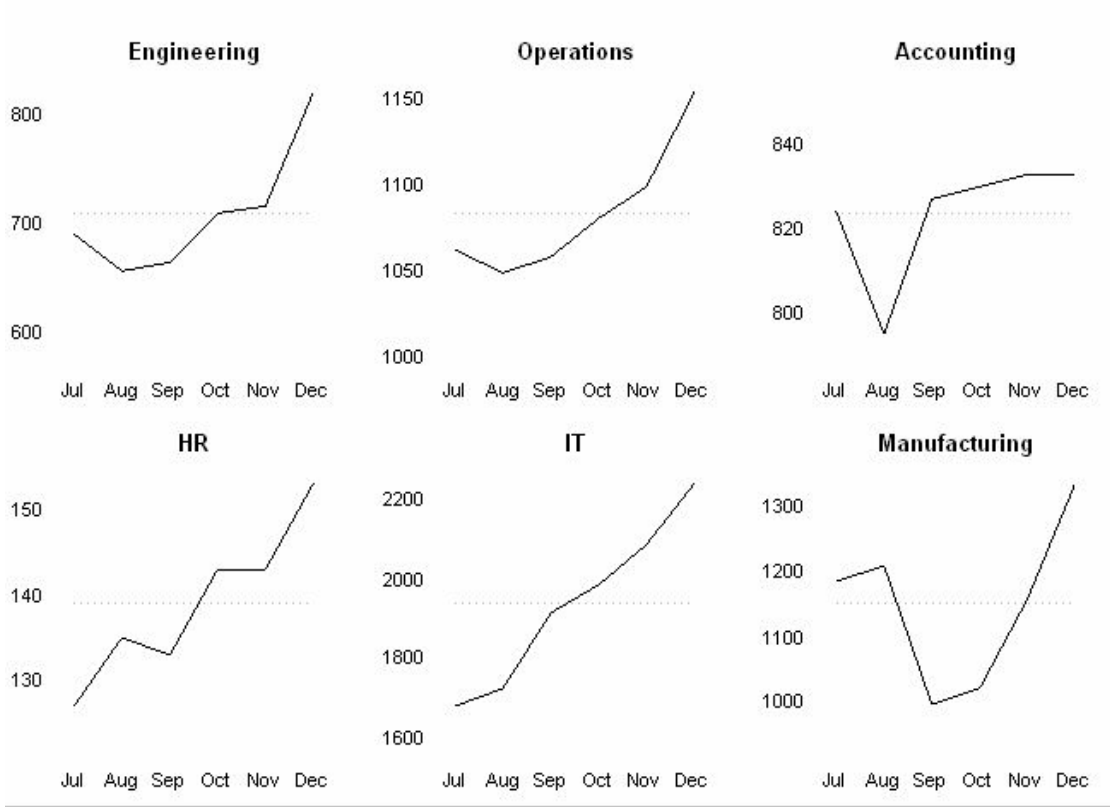
How do we measure the error?

1. Position on a common scale
2. Position on unaligned scales
3. Length
4. Angle
5. Area
6. Volume, density and color saturation
7. chromatic hue



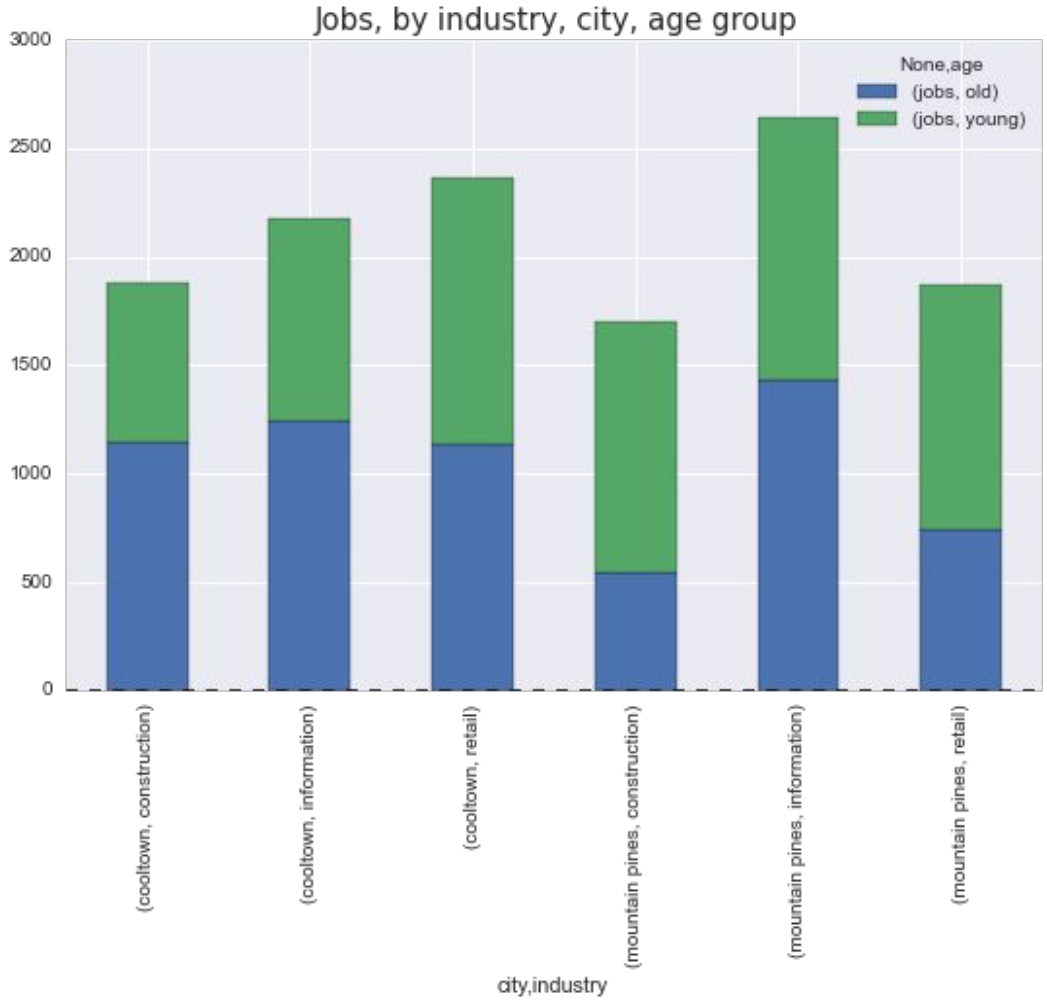
How do we measure the error?

1. Position on a common scale
2. Position on unaligned scales
3. Length
4. Angle
5. Area
6. Volume, density and color saturation
7. chromatic hue



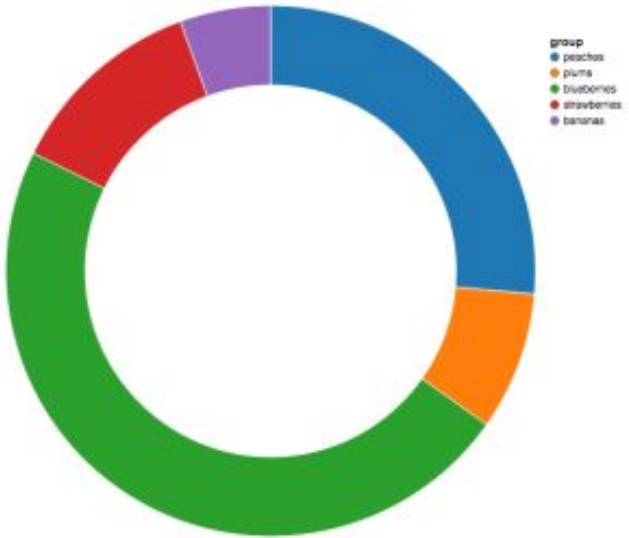
How do we measure the error?

1. Position on a common scale
2. Position on unaligned scales
3. Length
4. Angle
5. Area
6. Volume, density and color saturation
7. chromatic hue



How do we measure the error?

1. Position on a common scale
2. Position on unaligned scales
3. Length
4. Angle
5. Area
6. Volume, density and color saturation
7. chromatic hue

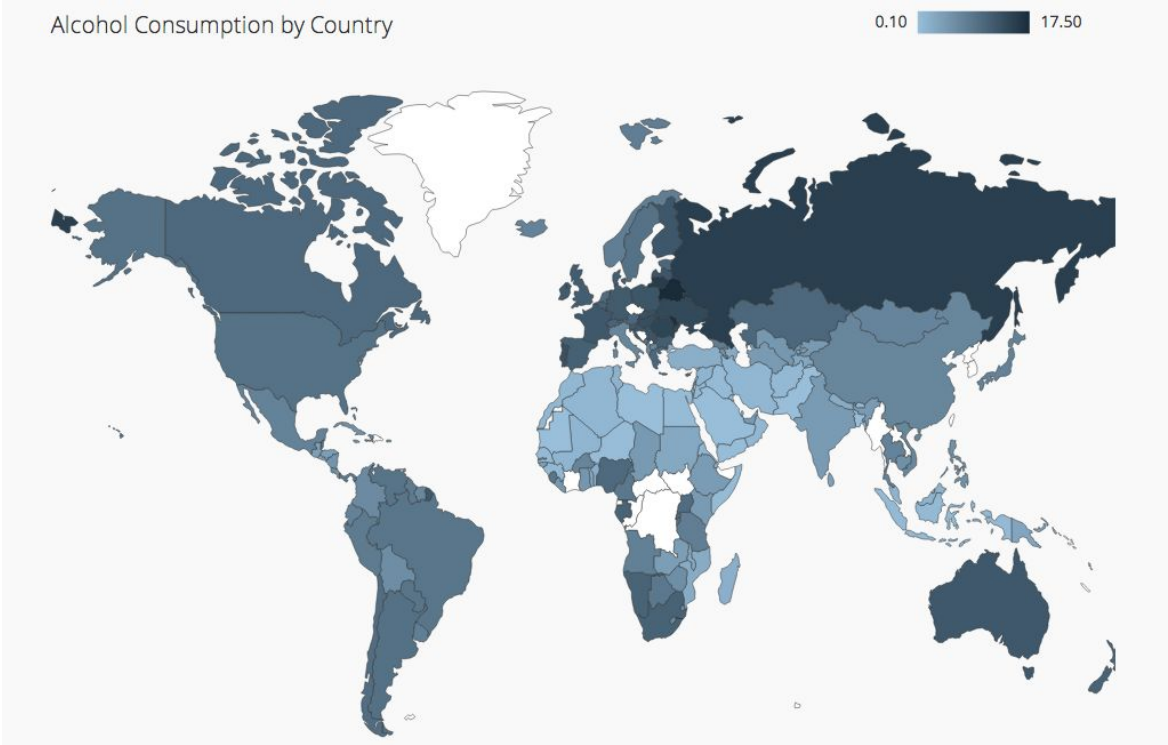


How do we measure the error?

1. Position on a common scale
2. Position on unaligned scales
3. Length
4. Angle
5. Area
6. Volume, density and color saturation
7. chromatic hue

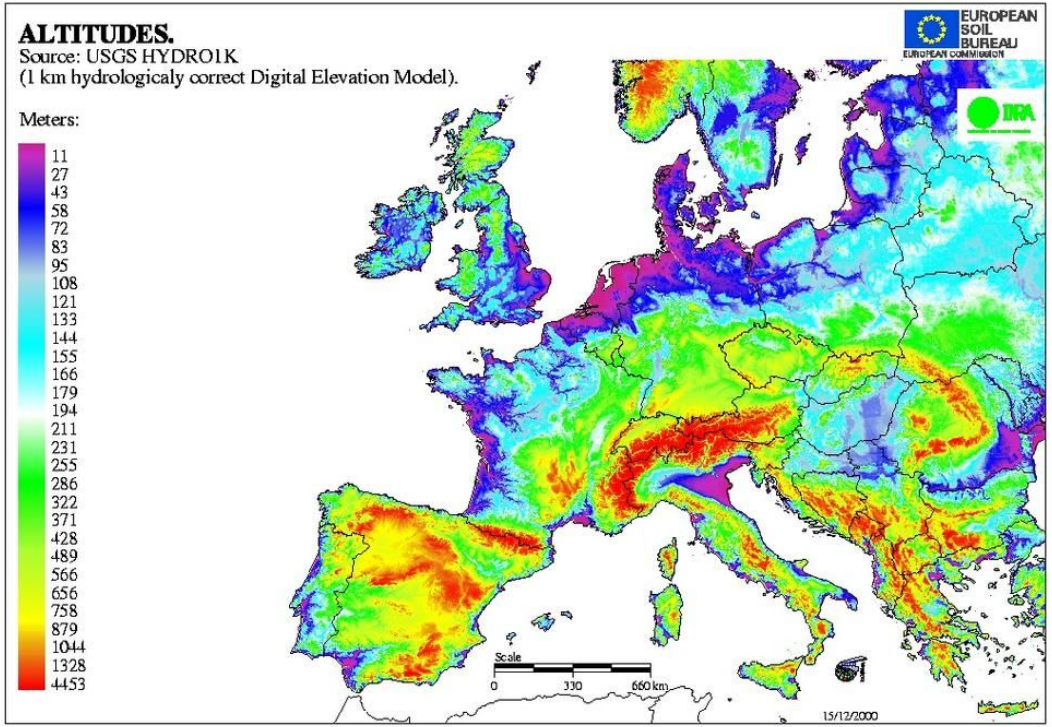
How do we measure the error?

1. Position on a common scale
2. Position on unaligned scales
3. Length
4. Angle
5. Area
6. Volume, density and color saturation
7. chromatic hue



How do we measure the error?

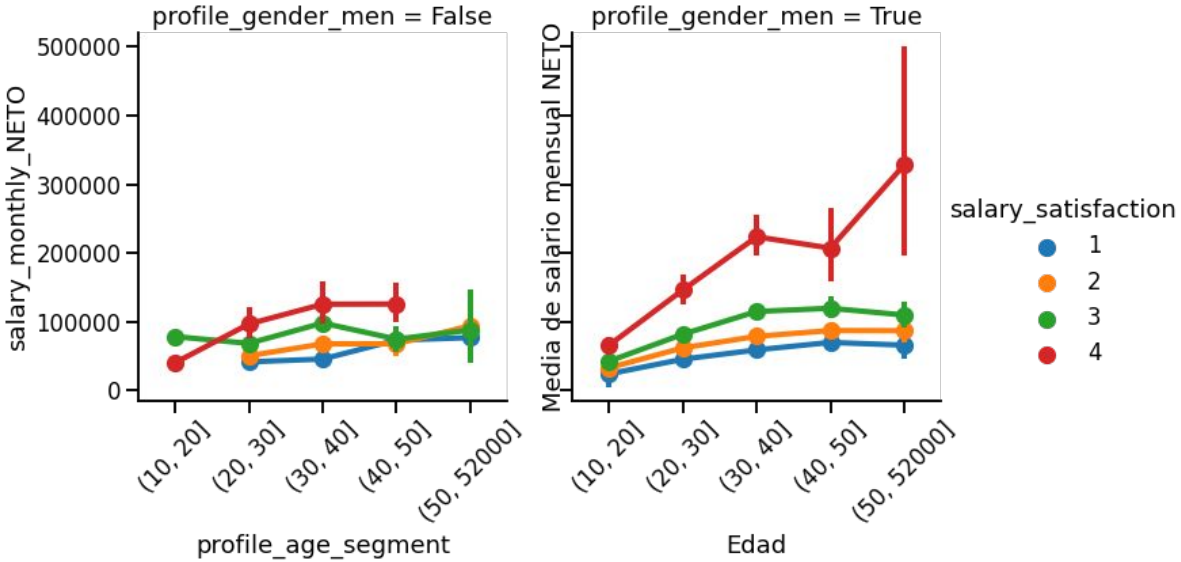
1. Position on a common scale
2. Position on unaligned scales
3. Length
4. Angle
5. Area
6. Volume, density and color saturation
7. chromatic hue



Complex Graphics

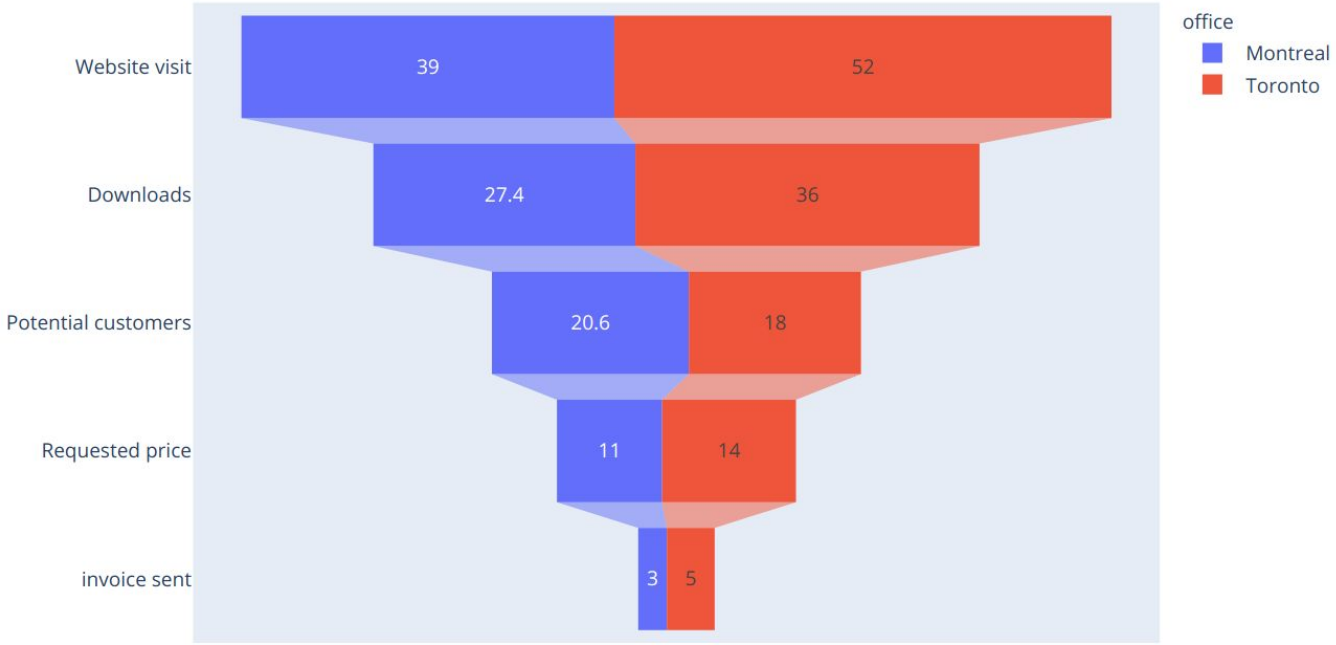
Adding more variables

- Every new variable we add needs a new encoding.
- In seaborn the structure is slightly different
- Other libraries like Plotly allow more complex and interactive graphs



Harder to read graphics

- Less intuitive but more informative visualizations
- Complex transformations to data
- Uncommon Chart Types

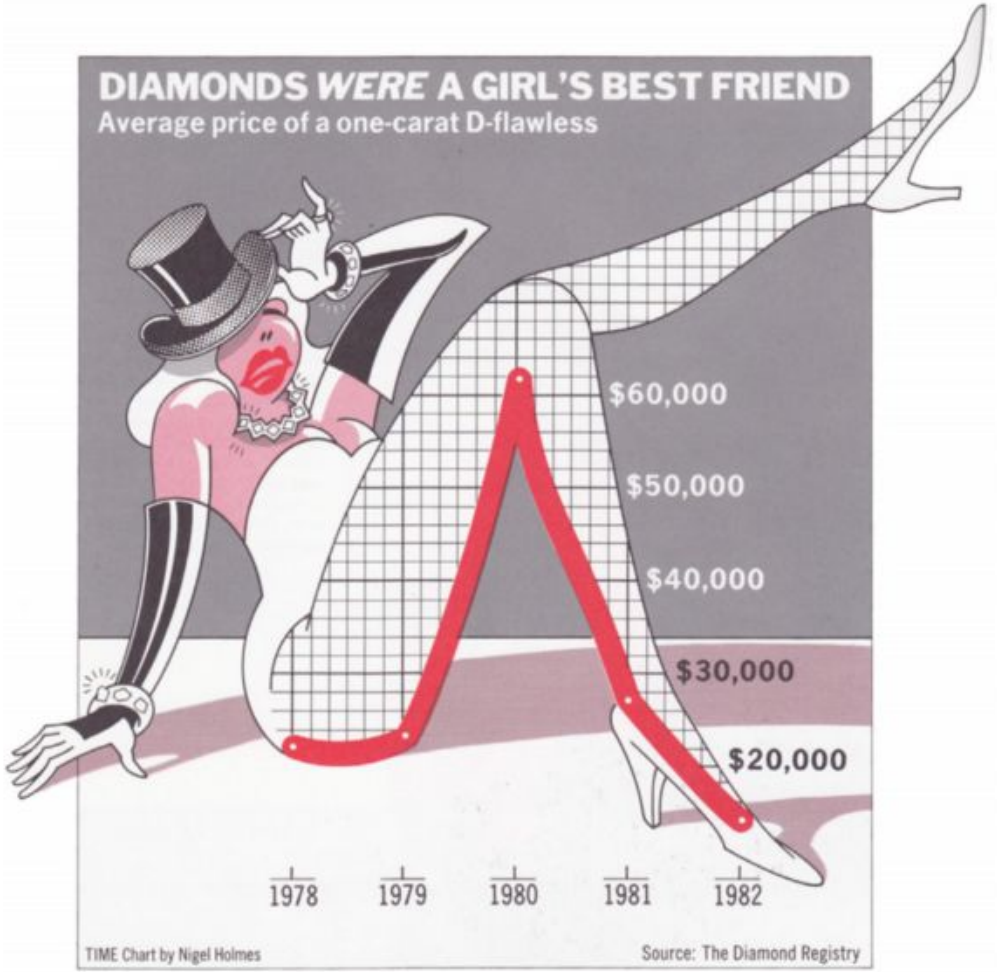


https://plotly.com/python/funnel-charts/

Example

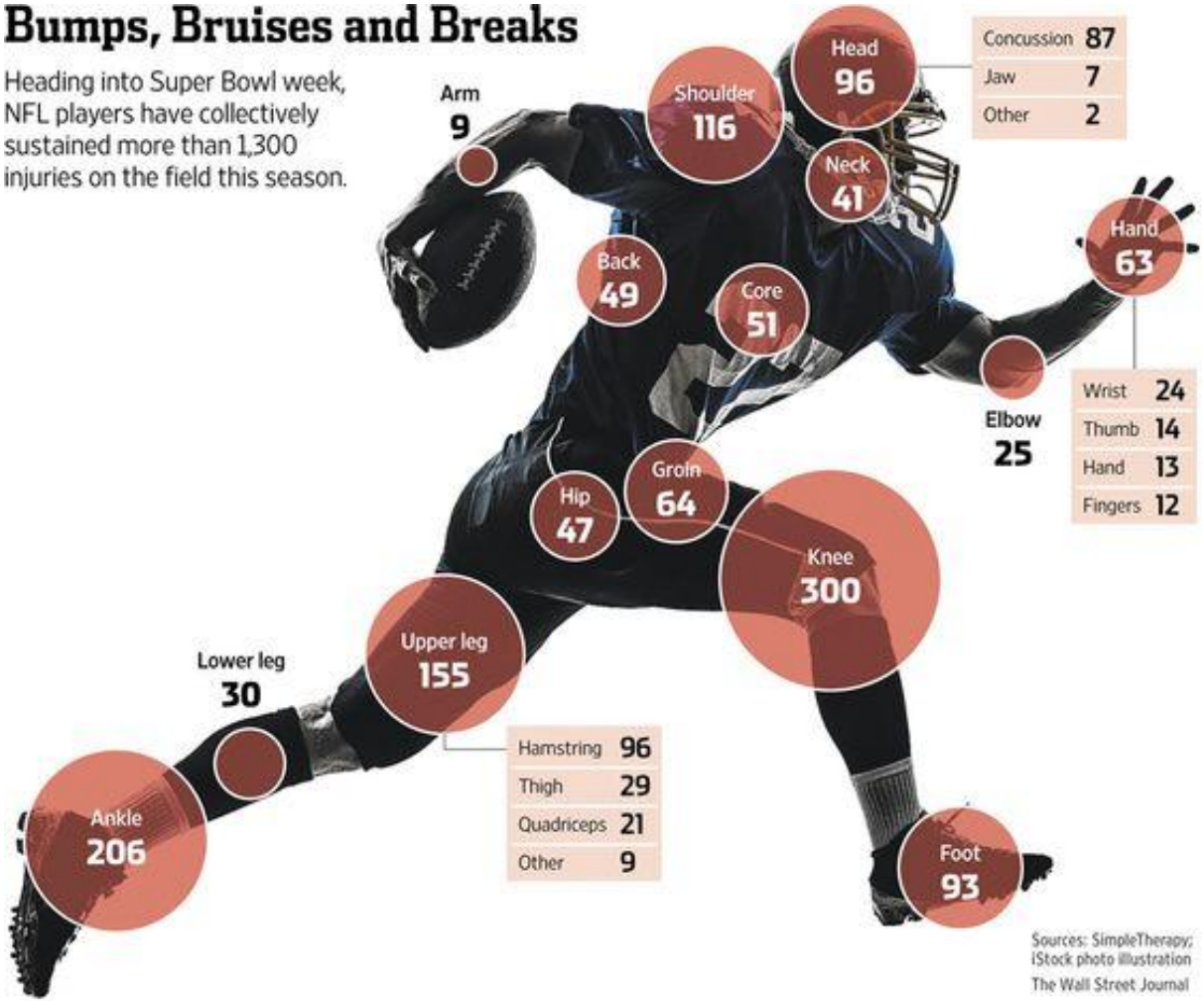
Only a small number of visual

elements are relevant!



Example

All visual elements are relevant to the information conveyed



Questions?

https://github.com/benjaminocampo/dl_data_visualization_2023