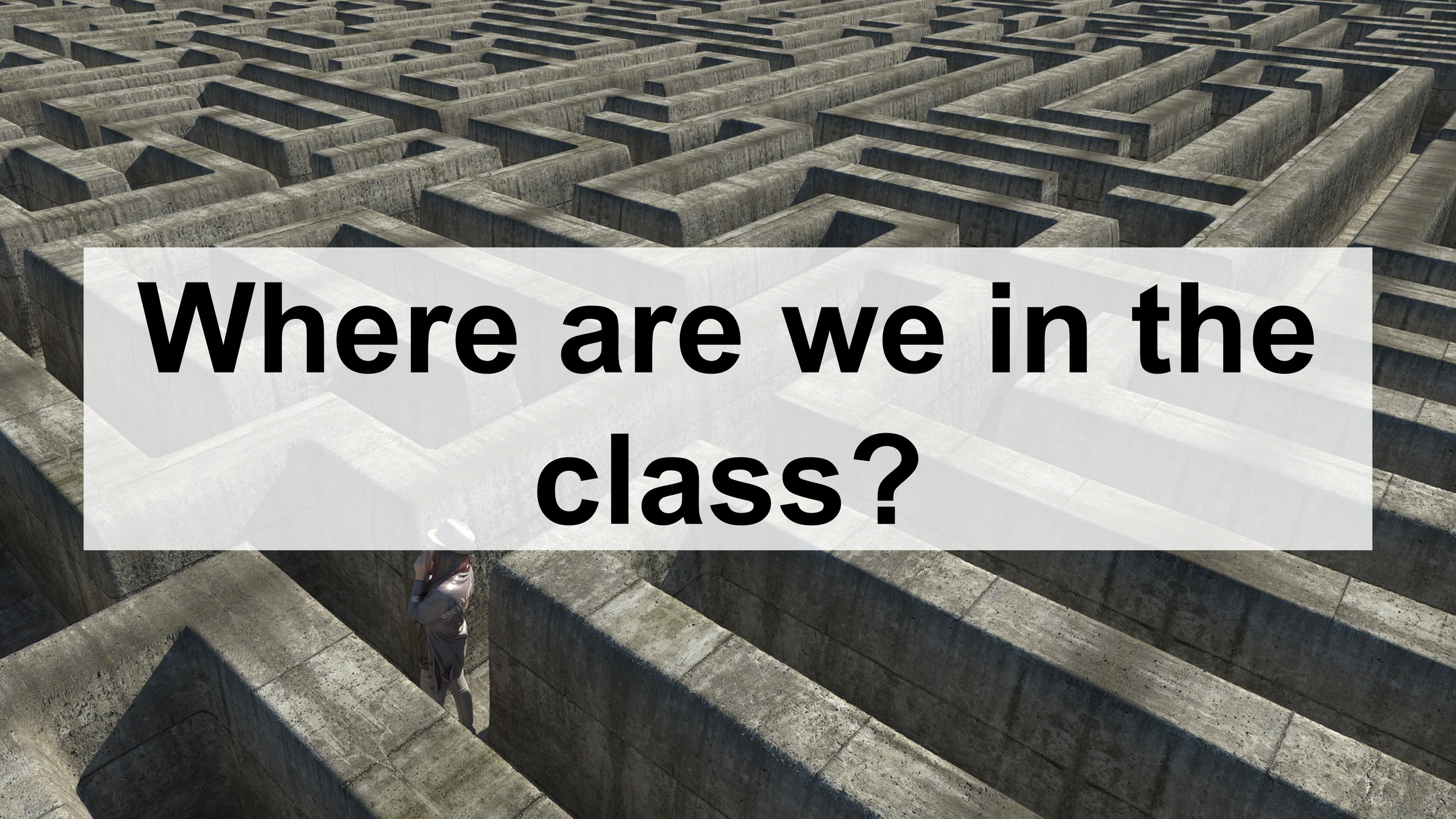
Gies Business

ETL & EDA in R

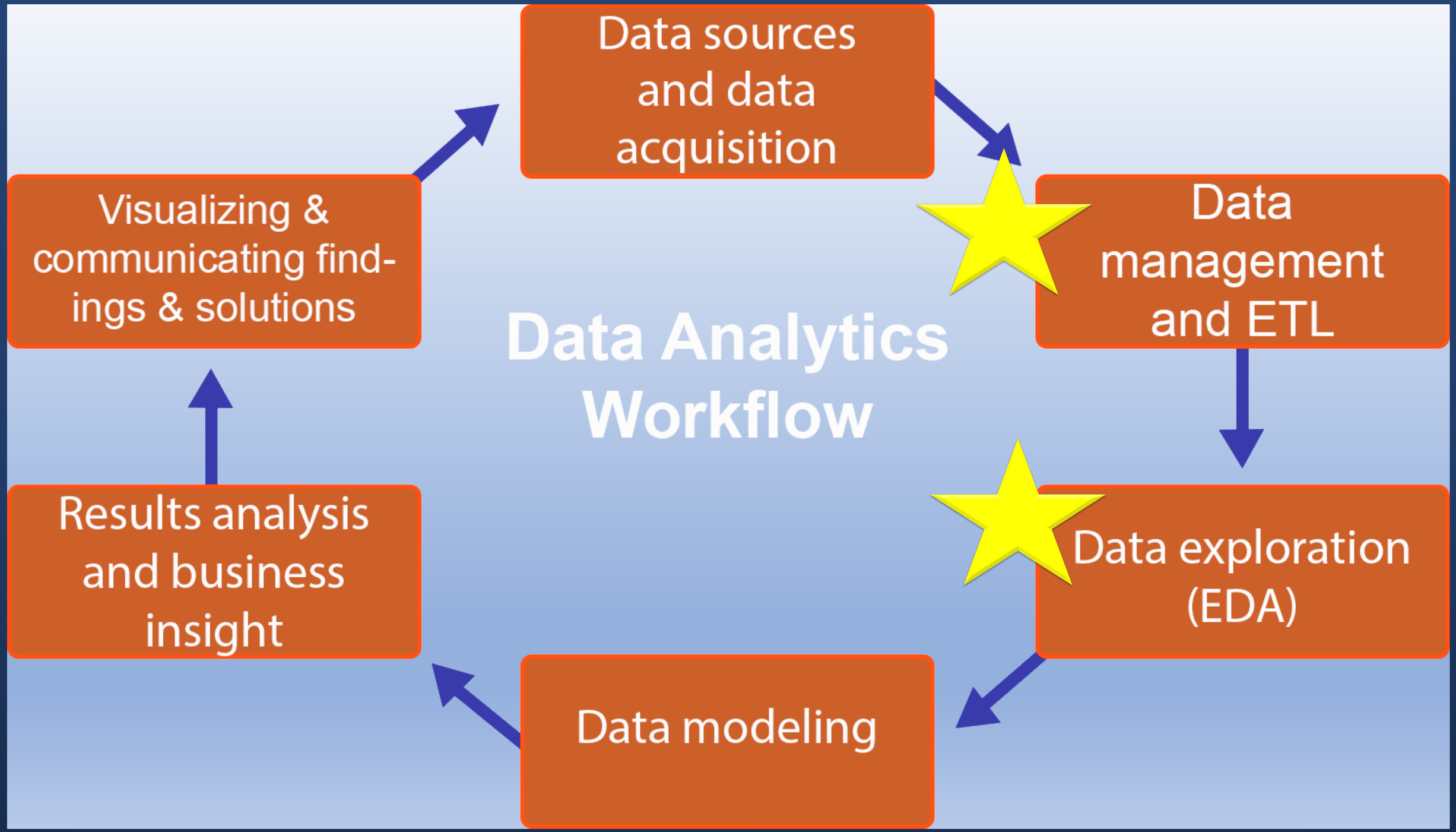
Module 3

Ron Guymon and Jessen Hobson

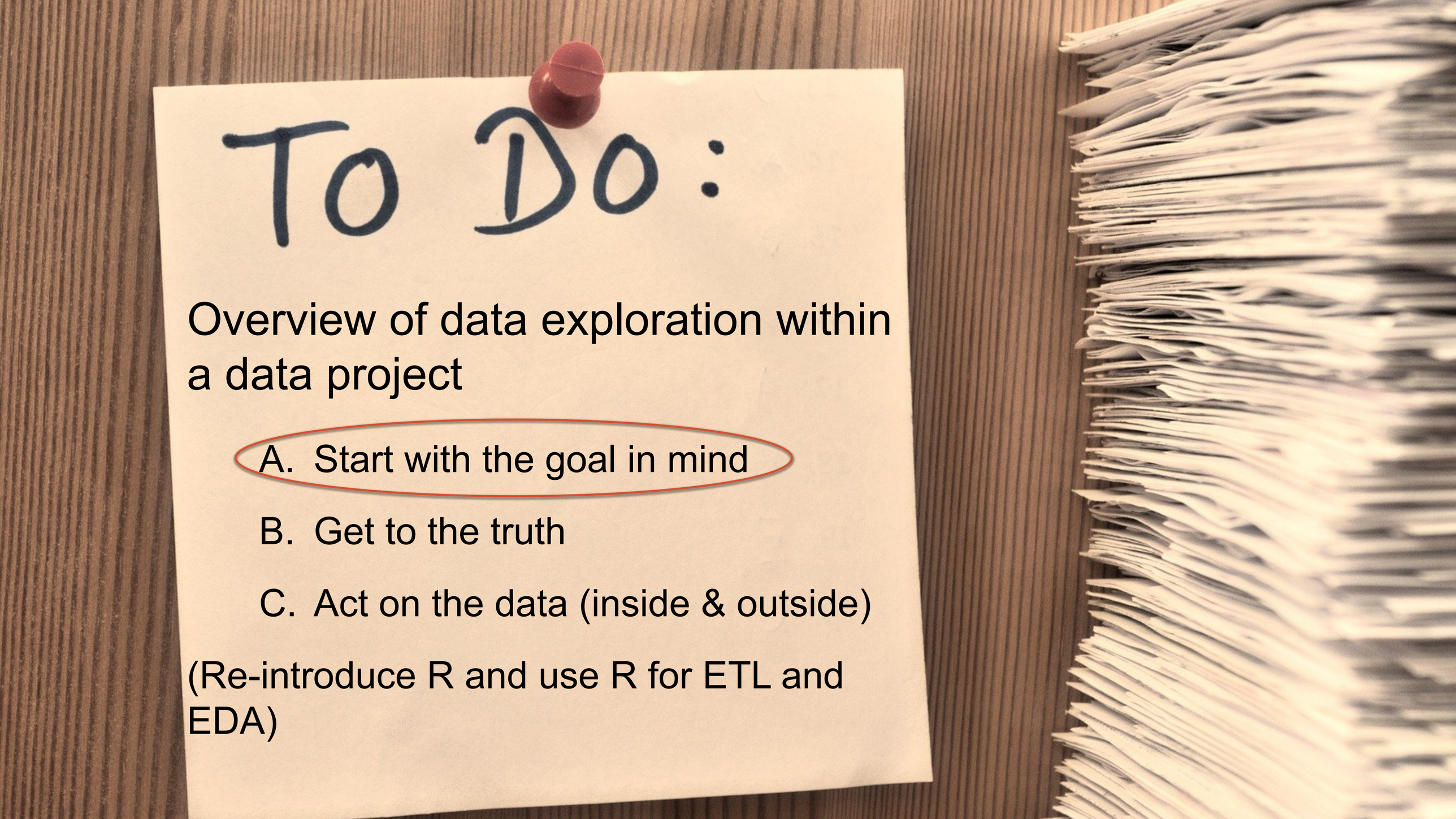
Last week’s assignments



Where are we in the class?


Data Analytics Workflow
Data sources and data acquisition
Data management and ETL
Data exploration (EDA)
Data modeling
Results analysis and business insight
Visualizing & communicating find-ings & solutions

Review Coursera



Overview of data exploration within a data project

A. Start with the goal in mind

B. Get to the truth

C. Act on the data (inside & outside)

(Re-introduce R and use R for ETL and EDA)

Start with the goal in mind

Two ways data enters the business problem

1. Start with a question or problem
2. Start with data

Breakout! – 3 minutes

Breakouts

1. Start with a question or problem
2. Start with data

What data do you have in your organization that has helped you solve a problem or has had a significant impact? What was the problem it solved or the impact it had?

What came first, the problem or the data?

Mini Assignment 1 – your own data

Find data you are inherently interested in

Learn by doing

Use the tools you are learning

Two ways data enters the business problem

1. Start with a question or problem
2. Start with data
   - Still need an ultimate goal / use for the data

- Avoid at all costs playing with data with no goal in mind

Don’t throw money at data/technology without a goal

“As corporate spending on artificial intelligence systems is set to pass **$50 billion** this year, the **vast majority of companies** may not be seeing much return on that record investment. In a survey of more than 3,000 company managers about their AI spend, **only 10%** reported significant financial benefits from their investment so far.”

https://www.forbes.com/sites/davidjeans/2020/10/20/bcg-mit-report-shows-companies-will-spend-50-billion-on-artificial-intelligence-with-few-results/

Don’t throw money at data/technology/AI without a goal

“Thinking about [the production environment for AI models] **<u>from the beginning</u>** will help you manage your project, …”
(https://bg.hbr.org/2020/06/the-dumb-reason-your-ai-project-will-fail)

“To achieve significant financial benefits, companies must **<u>look beyond the initial</u>**, albeit fundamental, steps of AI adoption—of having the right data, technology, and talent in place, and organizing these elements **<u>around a corporate strategy</u>**.”
(https://www.bcg.com/publications/2020/is-your-company-embracing-full-potential-of-artificial-intelligence)

Research Question: auditors detecting fraud

Can auditor detect deception from voice?

They are not as good as you would hope

Fraud Detectors within External Governa	
Analyst	24 (16.9%)
Auditor	16 (11.3%)
Client or competitor	9 (6.3%)
Employee	26 (18.3%)
Equity holder	5 (3.5%)
Industry regulator, gvt. agency or self-regulatory organization	20 (14.1%)
Law firm	5 (3.5%)
Media (incl. academic publications)	22 (15.5%)
SEC	10 (7.0%)
Short-seller	5 (3.5%)
External governance total cases	142 (100%)

Dyck, Morse, Zingale`s (2010)

They should be good – why aren't they?

They talk to managers all day

They train to see mistakes

They have hundreds of clients

They have many DISincentives to find fraud

Let's test it (experimental mindset)

Can auditors detect deception from client narratives, and can we help them improve detection?

Next. provide an overall judgment of whether it is more likely than not that fraud was being committed at this company during this quarter. That is, did this company later restate this quarter's financial results due to one or more of the following: fraudulent financial results, a regulatory investigation, or a class action lawsuit?

- Yes, fraud was likely being committed during this quarter.
- No, fraud was not likely being committed during this quarter.

Experiment with experienced auditors

Judge fraud & list red flags from 4 sequential companies

First 5 minutes of CEO Q&A from conference call

We learn by doing! Let’s get to some data and get our hands dirty!

1.1

	A	B	C	D	E	G	H	I	J	K	L	M	N	O
1	V1	V2	V3	V4	V5	V7	V8	V9	V10	type	set	Q1.1	Q1.2_1_TE	Q1.2_2_
2	ResponseID	ResponseS	Name	ExternalDa	EmailAddre	Status	StartDate	EndDate	Finished	type	set	Intro	Browser Me	Browser
3	R_780CFm	Default Res	Anonymous			0	3/7/2012 9:32	3/7/2012 9:34	0	TCD	73	1		
4	R_6y5PfEC	Default Res	Anonymous			0	3/7/2012 10:00	3/7/2012 11:13	0	ACD	10	1	Safari	5.1.3
5	R_cA6HKV	Default Res	Anonymous			0	3/7/2012 13:06	3/7/2012 13:26	0	TN	20	1	Chrome	12.0.74
6	R_0CEHwV	Default Res	Anonymous			0	3/7/2012 14:54	3/7/2012 16:09	1	AN	17	1	MSIE	
7	R_bKQzKl4	Default Res	Anonymous			0	3/7/2012 16:42	3/7/2012 18:45	1	AN	70	1	Chrome	17.0.96
8	R_00AATal	Default Res	Anonymous			0	3/7/2012 20:49	3/7/2012 22:26	1	ACD	2	1	Chrome	17.0.96
9	R_djnAZWi	Default Res	Anonymous			0	3/7/2012 15:51	3/8/2012 7:13	1	TN	47	1	Chrome	17.0.96
10	R_3HMlhtV	Default Res	Anonymous			0	3/8/2012 14:26	3/8/2012 15:59	1	TCD	18	1	Chrome	17.0.96
11	R_9zalVj5o	Default Res	Anonymous			0	3/8/2012 17:44	3/8/2012 18:59	1	AN	26	1	Chrome	17.0.96
12	R_51G7pC	Default Res	Anonymous			0	3/8/2012 19:00	3/8/2012 19:39	1	ACD	60	1	Chrome	17.0.96
13	R_6FD20W	Default Res	Anonymous			0	3/8/2012 18:41	3/8/2012 19:43	0	TN	57	1	Chrome	17.0.96
14	R_3rxl3NY	Default Res	Anonymous			0	3/9/2012 20:56	3/9/2012 22:59	1	TCD	71	1	Safari	5.1.2
15	R_eqHDIP9	Default Res	Anonymous			0	3/9/2012 21:13	3/10/2012 9:33	1	TN	3	1	Chrome	17.0.96
16	R_eqzvCW	Default Res	Anonymous			0	3/9/2012 15:45	3/10/2012 12:10	1	ACD	19	1	Safari	5.0.6
17	R_a3Ra3Q	Default Res	Anonymous			0	3/10/2012 11:48	3/10/2012 13:45	1	TCD	24	1	Firefox	10.0.2
18	R_0wEntbn	Default Res	Anonymous			0	3/10/2012 16:36	3/10/2012 19:55	1	AN	75	1	Chrome	17.0.96
19	R_42z0CcE	Default Res	Anonymous			0	3/10/2012 13:20	3/11/2012 21:59	0			1	Firefox	10.0.2
20														

	BJJ	BJK	BJL	BJM	BJN	BJO	BJP	BJQ	BJR	BJS	BJT	BJU	BJV	BJW	BJX
1	Q2.4	Q2.5	Q2.6	Q2.7	Q2.8	Q2.9	Q2.10	Q2.11	Q2.12	Q2.13_1	Q2.13_2	Q2.13_3	Q2.13_4	Q2.14	Q2.15
2	infointro	name	backgrnd	bs	IS	CF	audio	audiotrans	todvs	Timing-Firs	Timing-Last	Timing-Pag	Timing-Clic	Unable to d	fraudyn
3															
4	1	1	1	1	1	1	0.24 - 0.26	1	1	53.435	518.926	520.063	3	1	
5															
6															
7															
8	1	1	1	1	1	1	0.24 - 0.26	1	1	34.424	517.228	520.344	13	1	
9															
10															
11															
12	1	1	1	1	1	1	0.24 - 0.26	1	1	5.094	5.094	6.332	1	1	
13															
14															
15															
16	1	1	1	1	1	1	0.24 - 0.26	1	1	162.342	375.176	376.514	5	1	
17															

	CEW	CEX	CEY	CEZ	CFA	CFB	CFC	CFD	CFE	CFF	CFG	CFH	CFI	CFJ
1	Q14.27_3	Q14.27_4	Q6.1	Q6.2_1_TE	Q6.2_2_TE	Q6.2_3_TE	Q6.2_4_TE	Q6.2_5_TE	Q6.2_6_TE	Q6.2_7_TE	Q6.3	Q6.4_1	Q6.5_1	Q6.6
2	Timing-Pag	Timing-Clic	Exit Survey	Questions /	Questions /	Questions /	Questions /	Questions /	Questions /	Questions /	8. Gender	9. How mu	10. Compa	11. Plea
3														
4														
5														
6			1	0	0	Student	0	Bachelor			2	1	-1	Only dis
7			1	0	0	Student	0	High schoo	Nope	None	1	0.82	0.49	I have ve
8			1	0	0	Student	0	GED	none	none	1	0.93	0	None
9			1	0	0	student	0	Bachelor	none	none	2	0	-2.09	No spec
10			1	0	0	0	0	Some College			1	0		None, th
11			1	0	0	student	0	BA			2	0.73		I have n
12			1	0	0	Student	0	Bachelors	N/A	N/A	2	0.27	2.06	None
13														
14			1	0	0	Student	0	Bachelors	None	None	1	0	-2	1) I have
15			1	0	0	student	0	University			1	0	-5	The train
16			1	1	0	student	1 summer	College Jur	Logistics		1	2.24	0.22	I have n
17			1	0	0	student	0	BA	Accountanc	n/a	2	0	1.05	none
18			1	0	0	Intern	0.3	3 years of college			2	0	-2.09	I have re
19														

1.2 – ETL Tasks keeping the goal in mind

What features do we actually care about?

What should guide our analyses?

Focus on the data that matters

Any good project has a lot of data.

Let your problem/question by your guide!

What about the “NA”s?

804 “NA”s in `auditor_fraud`

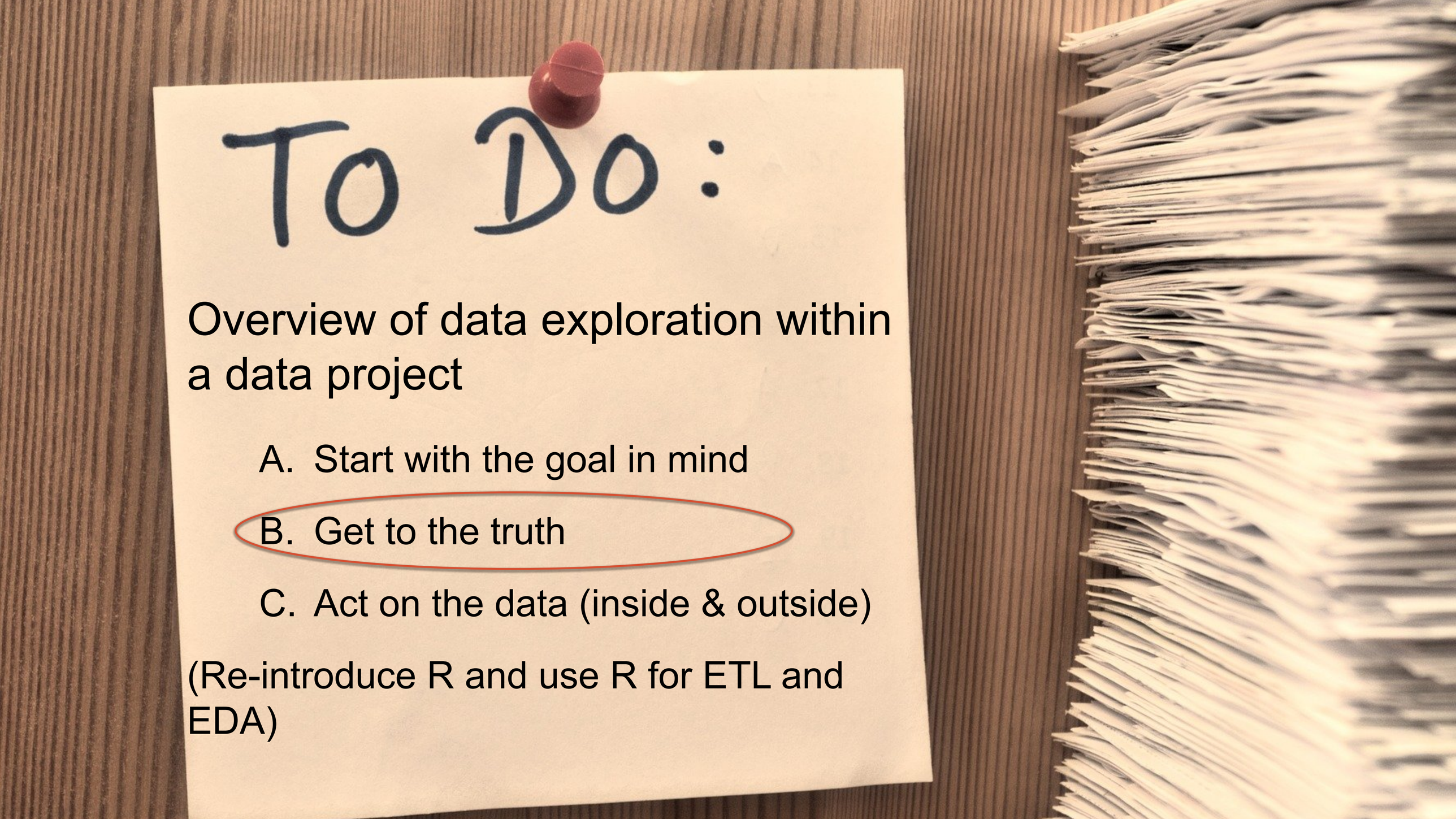
So, let’s delete them all, right?

Create Variable for Professional

How do we know whether someone is an auditor or not from these data?

What variables should we look at?

```
 [1] "subject"       "company"       "fraud"         "judgment"      "aud_exp"
 [6] "prompt"        "overall_time"  "jobtitle"      "auditor_fraud" "confidence"
[11] "flagchars"     "accurate"
```



Overview of data exploration within a data project

A. Start with the goal in mind

B. Get to the truth

C. Act on the data (inside & outside)

(Re-introduce R and use R for ETL and EDA)

Get to the underlying truth

Get to the underlying truth

Every data project is ultimately concerned with sifting through the detritus of the data to get to truth

- Causal analysis
- Prediction

What is the detritus?

Noise in the data

- Measurement error
- Outliers
- Bias in the data
- Etc.

Insignificant results

Unimportant results (spurious correlations)

Omitted, correlated variables

What went wrong here? 3 min

Breakouts

"Take the case of Harry Rosen, a 17-store luxury menswear chain based in Toronto. In 2018, it did a customer survey, asking if buyers would use PayPal if the company offered it on the site—and a significant number said yes. So, the company hired a systems integrator to add PayPal into its e-commerce platform, at a cost of $80,000. But then only a small portion of the company's online payments came through PayPal—just 10% by late 2018. The trouble seemed to be that PayPal wasn't a one-step process. Customers had to click "buy," fill out their payment information on the PayPal platform, and then go back to the Harry Rosen site to complete the transaction."

Harry gathered data to solve a problem and let the data speak, but the solution failed. What went wrong? What should he have done instead?

https://www.wsj.com/articles/the-many-ways-companies-waste-money-with-technology-spending-11615298419

Detritus – measuring the wrong thing

“In retrospect, says Ian Rosen, Harry Rosen’s executive vice president, digital and strategy, the company should also have asked customers about a broader question: What kind of payment experience best suits your needs? The company might have realized that customers wanted a simpler payment experience overall, not necessarily PayPal itself. Mr. Rosen says the company is rolling out a new e-commerce platform and payment technology that allows it to connect to multiple payment methods with one click. PayPal is “a fantastic solution” if you optimize your systems to match PayPal’s strengths, says Mr. Rosen, but “when you try to force it into another process, you may lose some of the great benefits.”

Bias in AI results

Garbage in => garbage out

“AI-driven systems and models will stop functioning when being fed wrong and malformed data.” (https://hbr.org/2020/06/the-dumb-reason-your-ai-project-will-fail)

IBM Watson Oncology

Amazon hiring

2.1 - Outliers

Why do we get rid of outliers?

Look at the summary. Do any features have potential outliers?

Which ones should we take out?

Visualize the data – 10 min.

Breakouts

In groups, do a simple visualization of the data. Pick one or two variables to look at and create a plot with `ggplot2()`.

Use data “Mod3HE_HMPV_eda.rds”

Share what you find

2.2 – Main effects

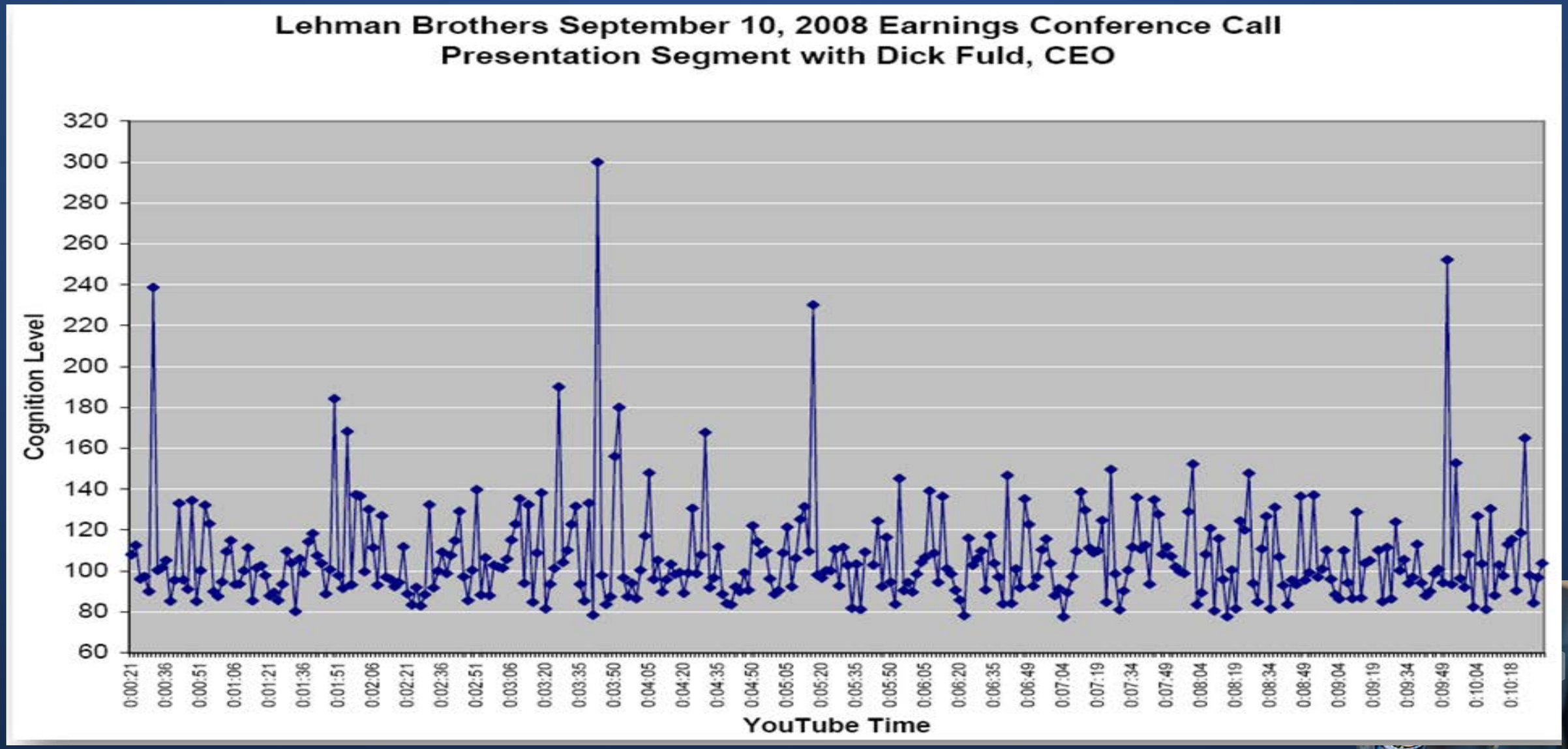
2.3 - Interactions

The effect of some features depends upon other features

Think of one feature turning another on or off

1. Weather + sale of umbrellas
2. Elon Musk's latest statement and price of crypto currency
3. Product placement (cross selling)

Can we get them into system 2?



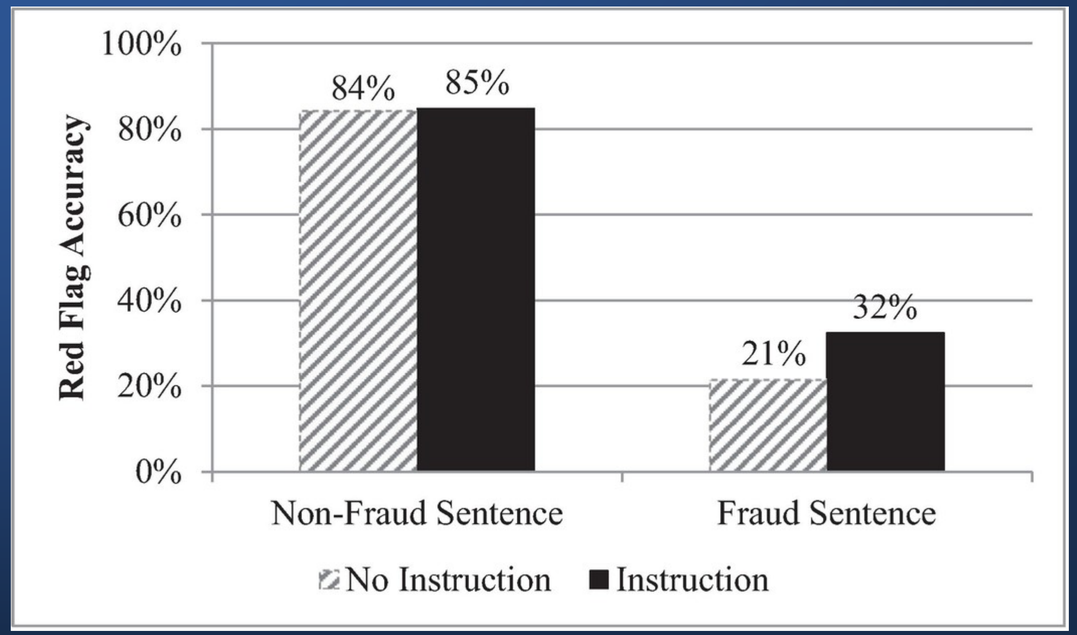
2.2 – Interactions

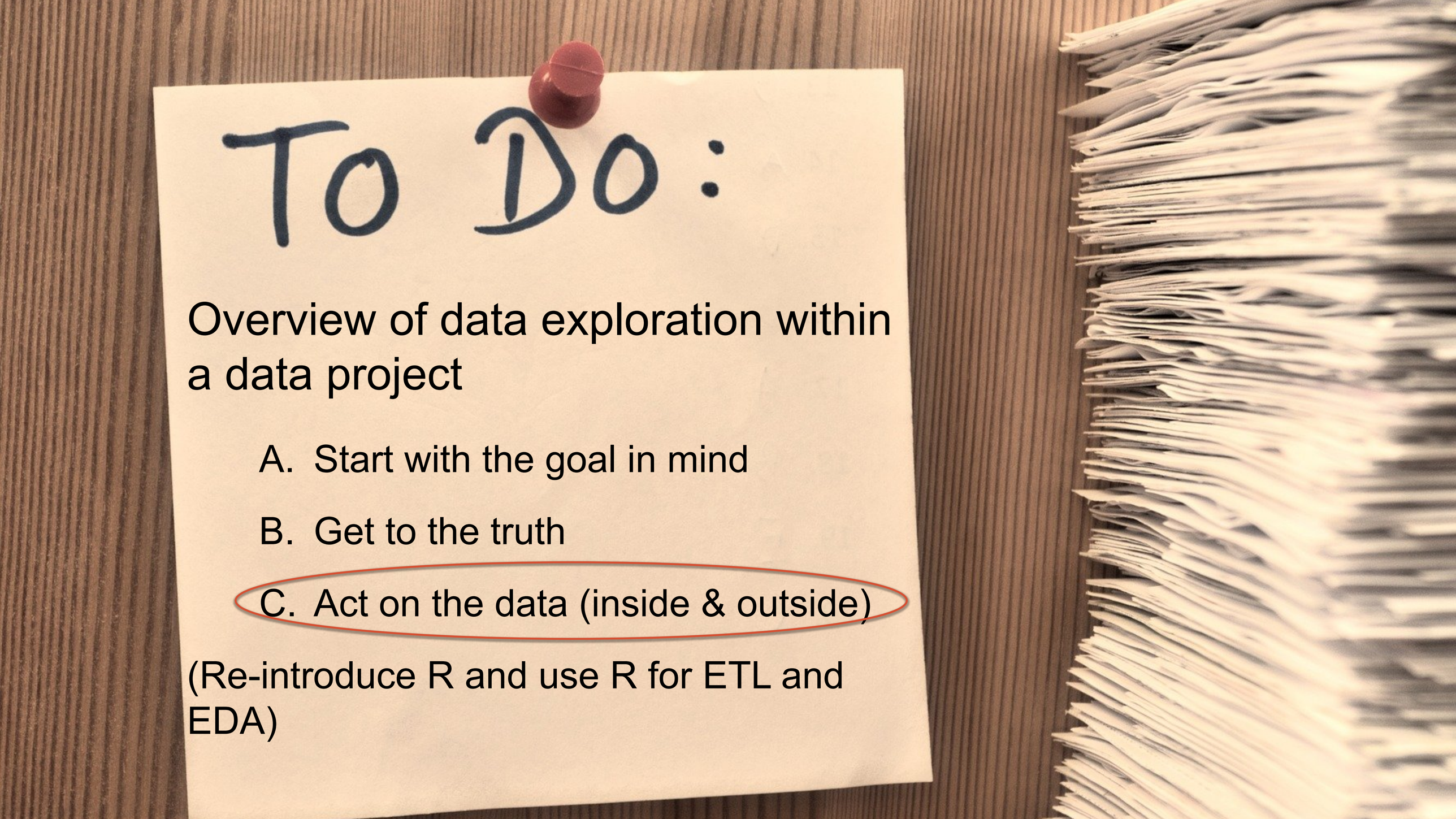
Missing Data – omitted correlated variables

E.g., PayPal example from before

E.g., any prediction model with low R^2

E.g., bringing in fraud sentence data in our study





Overview of data exploration within a data project

A. Start with the goal in mind

B. Get to the truth

C. Act on the data (inside & outside)

(Re-introduce R and use R for ETL and EDA)

Act on what the data tells you (inside & outside the project)

Act on what the data tells you (inside & outside the project)

Inside the project

- Let the data speak!
- Continuous feedback, revision, and adaptation

Outside the project

- Put your results to use
- Last mile problem

Continuous feedback, revision, and adaptation in our project

Started as pros versus joes

Changed to the value of detecting cognitive dissonance and auditor's inability to sense fraud

This change altered the analysis, how we looked at the data, as well as what we did with it

Put your results to use

The whole point of a data project is to help the company!

Visualizing & communicating find-ings & solutions

Results analysis and business insight

Last mile problem

“In 2016, Gartner estimated that 60% of data science projects were failing. In 2017, Gartner analyst Nick Heudecker said it was likely closer to 85%....In my experience, data science projects fail not because of poor technology or lack of available data. Nor is it because of a lack of talent or technical skills. It’s because of the last mile problem. Data scientists are struggling to deploy their models into business processes and applications where they are used by the business.”
(https://www.forbes.com/sites/forbestechcouncil/2019/07/23/solving-the-last-mile-problem-for-data-science-project-success/)

This week's assignments

Assessments this week:

- Individual Coursera Quiz
- Individual Assignment
- Mini-Assignment – your own data

Two individual assignments due this week

1. Individual analysis using R and NANSE data
2. Mini-Assignment (your own data)

NANSE Data

Aggregation level – store (average weekly results per store)

2015

Goal: maximize profitability of weekly store sales

NANSE Data - features

“store”

“week”

“high_med_gp”

“size”

“promo_units_per

“velocityA_units_per” (B, C, D)

“velocityNEW_units_per”

Percentage of total units sold for the top 20% of categories (e.g., gum, potato chips, energy drinks, 20oz. soda, etc.)

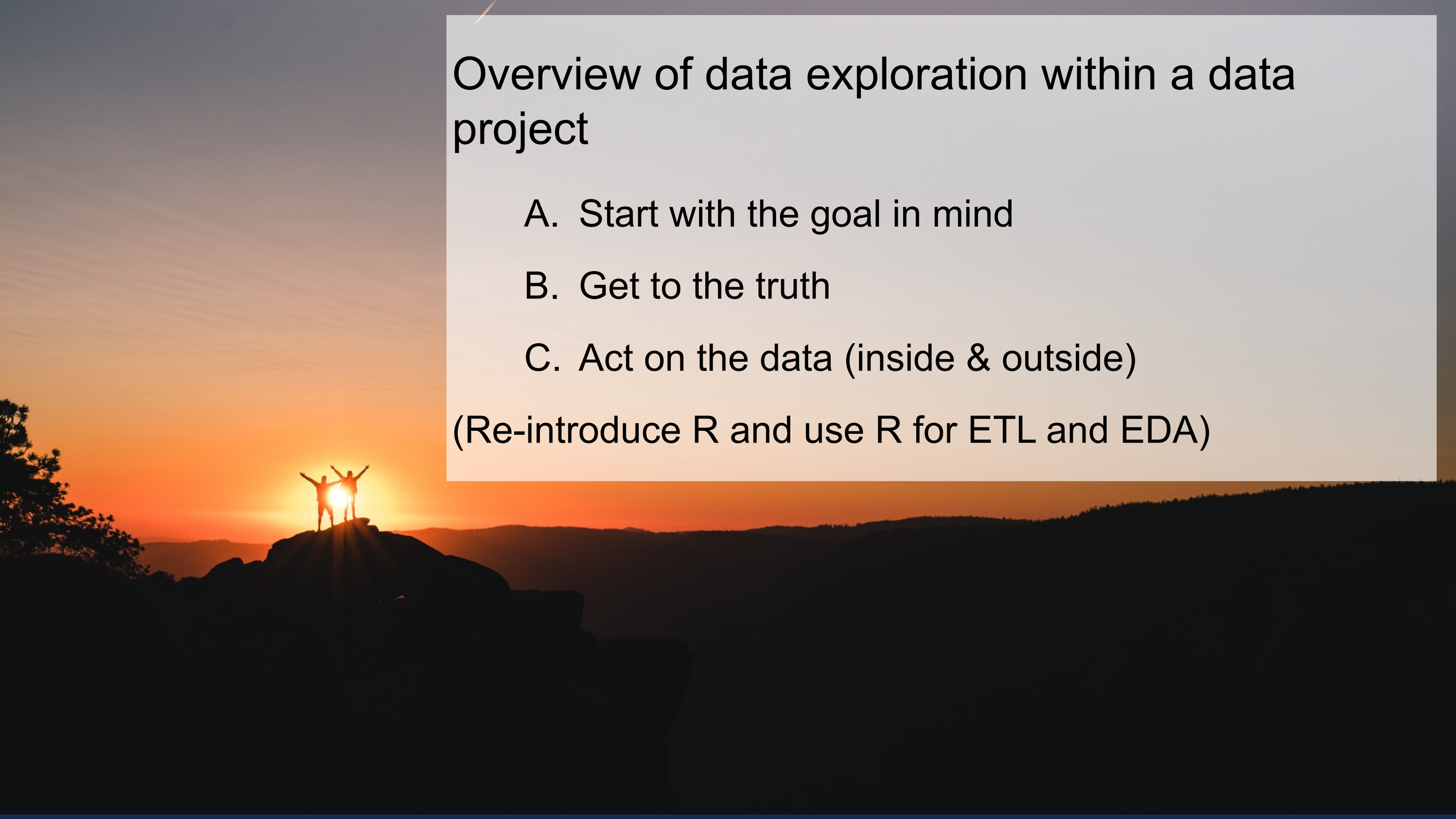
NANSE – Business problem

Mini Assignment 1 – your own data

Find data you are inherently interested in

Learn by doing

Use the tools you are learning



Overview of data exploration within a data project

A. Start with the goal in mind

B. Get to the truth

C. Act on the data (inside & outside)

(Re-introduce R and use R for ETL and EDA)