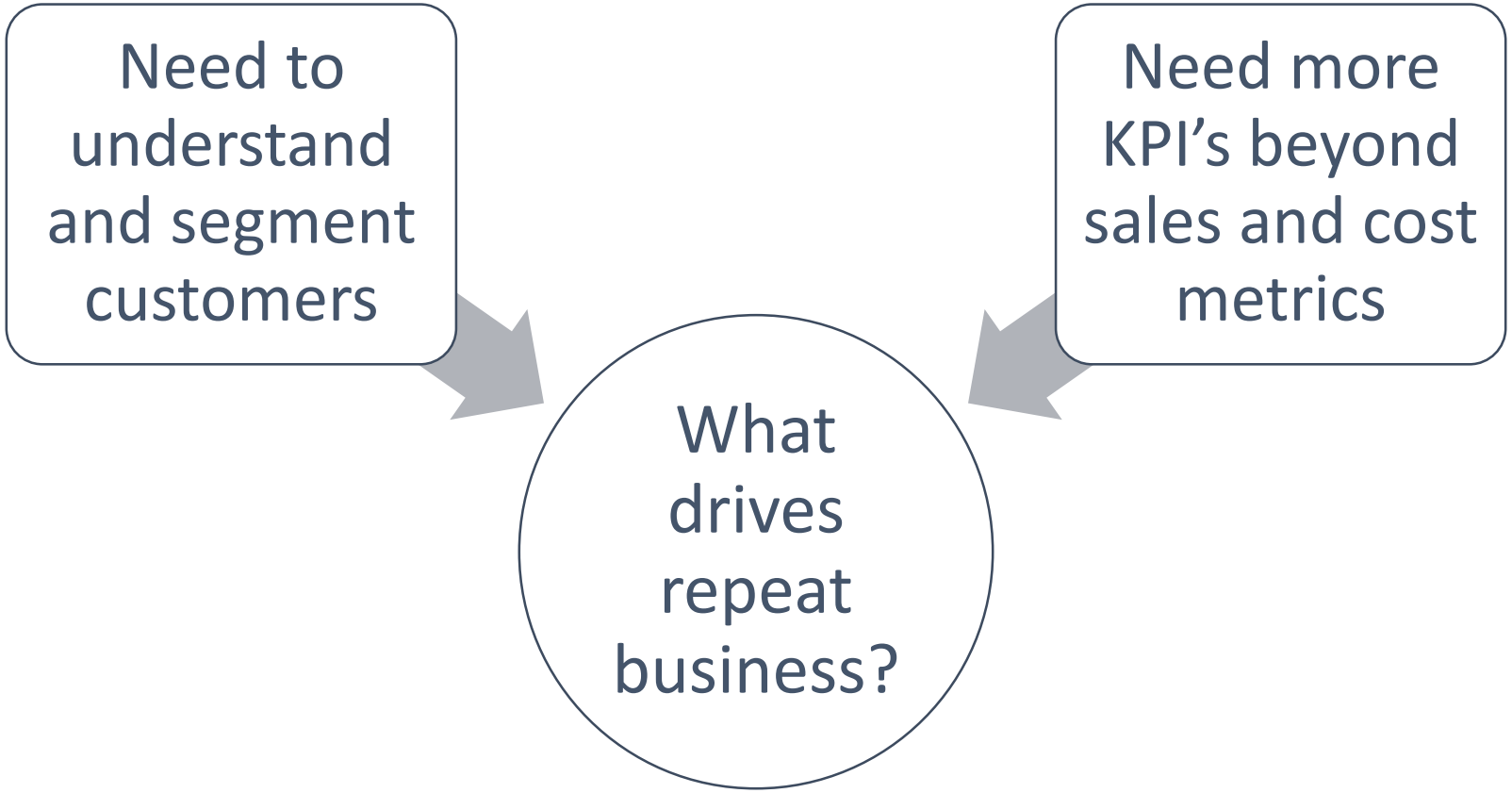
repeataly

predicting repeat customers with sales data

Eataly needs to better analyze its customers



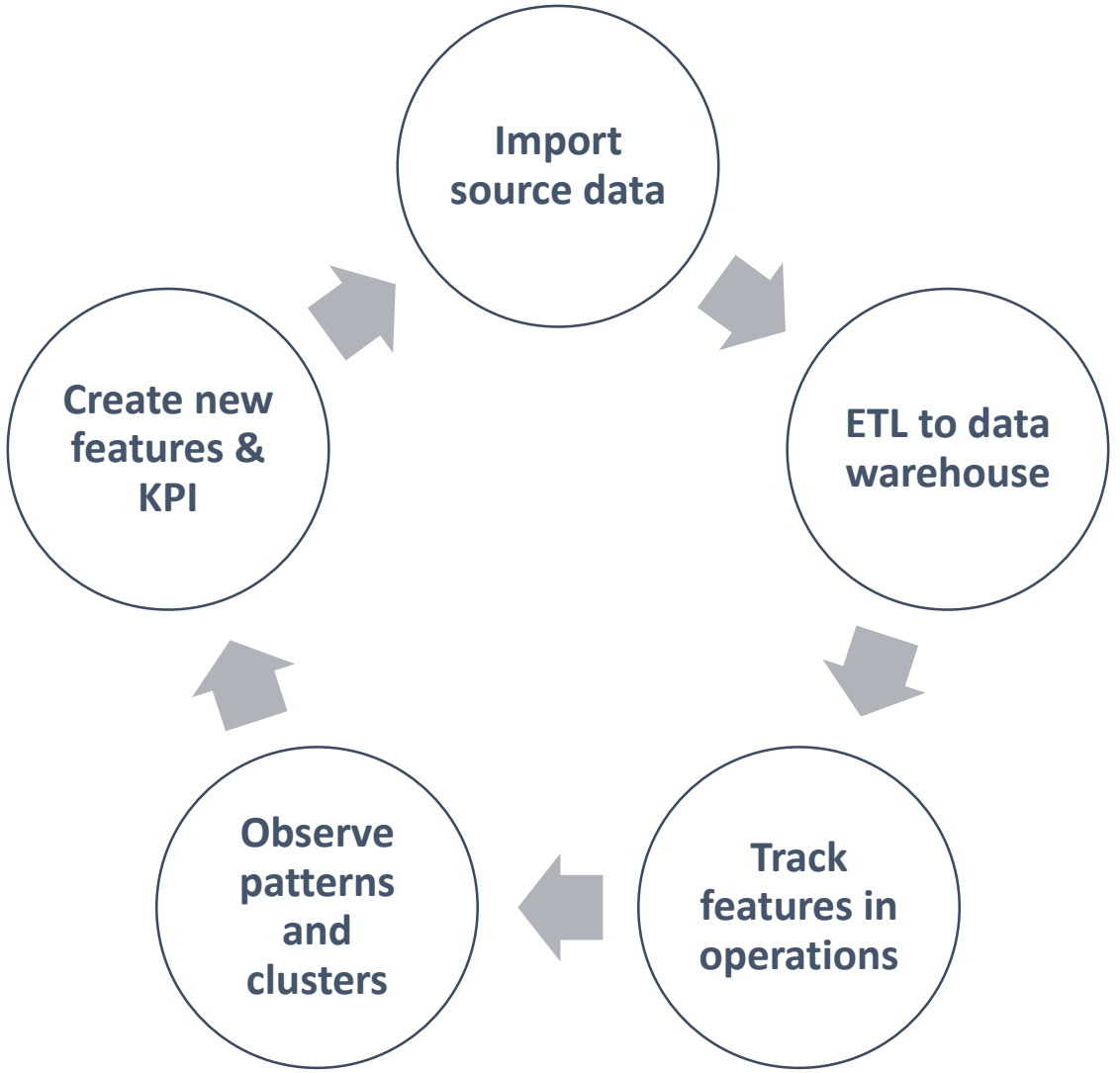
problem | dataset | progress

repeataly

Scope of question must be limited

What features and values from Eataly's POS transaction data will best predict the likelihood of a repeat visit within 90 days?

The goal is to track new features



problem → dataset → progress

repeataly

POS application provides a sea of clean data

TicketDate	TicketTime	PriorVisits	SaleLines	ReturnLines	NetAmount	NetRetail Amount	NetQSR Amount	Station Group	Unique Items	UniqueCategories	Has Retail Products	Has QSR Products	Returned Bags	WillReturn
2015-08-31	14:07:30	0	4	0	13.20	13.20	0	Other	2	1	True	False	False	True
2015-08-31	14:15:12	30	2	0	6.60	6.60	0	Other	2	1	True	False	False	True
2015-08-31	14:17:04	1	3	0	6.16	6.16	0	Other	2	2	True	False	False	False
2015-08-31	14:25:18	0	1	0	2.80	2.80	0	Other	1	1	True	False	False	True
2015-08-31	14:25:49	0	1	0	3.40	3.40	0	Other	1	1	True	False	False	True

828K observations | **15** features | **NYC** market | **2015** Jan - Dec

k-means clustering shows potential already

	Lunch Junkie	Sampler	Regular	Pro
PriorVisits	2.170	3.711	3.169	6.016
SaleLines	1.687	12.728	3.111	5.437
ReturnLines	0.001	0.001	0.003	1.034
NetAmount	14.349	105.468	22.869	39.195
NetRetailAmount	1.706	102.893	22.474	38.506
NetQSRAmount	12.604	2.137	0.323	0.722
UniqueItems	1.638	11.534	2.895	6.075
UniqueCategories	1.263	5.849	2.041	3.991
HasRetailProducts	0.186	1.000	0.993	0.994
HasQSRProducts	0.998	0.161	0.032	0.062
ReturnedBags	0.000	0.001	0.000	0.987
IsFrontEnd	0.012	0.981	0.716	0.995

Next steps are feature selection and creation

- Random forests and logistic regression will help me select features more effectively
- Domain knowledge suggests that certain best-selling products may deserve their own feature
- I will definitely add more features, although my computer can hardly handle the current dataset.

Questions?