CustID (Varchar)	Zipcode (Integer)	Income (Varchar)	Age (INT)	XYZ (Varchar)
101	92092	12000	25	005
102	92093	USD 100	56	001
103	92093	50000	34	002
104	92093	1000	56	002

Raw CSV file (DB schema)

Base Featurization

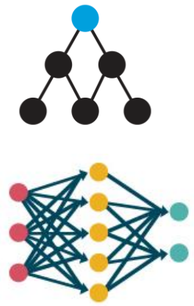
#	Attribute Name	Descriptive Statistics			Sample Values			Label
		Mean	% Distinct Vals	...	Sample1	Sample2	...	
1	Age	42.75	75		34	56		Usable Numeric
2	CustID	102.5	100		102	104		Unusable
3	XYZ	2.5	75		002	001		Context-Specific
4	ZipCode	92092.75	50		92093	92092		Usable Categorical
5	Income		100		USD 100	1000		Usable-with-Extraction

Offline Phase

Model-specific secondary featurization

#	3-gram on Attribute Name	Descriptive Statistics	3-grams on Sample Values
1	age	...	34
2	cus,ust,sti,tid	...	102
3	xyz	...	002
4	zip,ipc,pco,cod,ode	...	920,209,093
5	inc,nco, com,ome	...	usd,sd ,d 1, 10,100

Training



Trained ML model

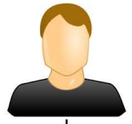
new CSV file

Gender	CompName
M	AMAZ
F	MSFT
M	GOOGL
M	MSFT

Online Phase

Label	Confidence	
	Gender	CompName
Usable Numeric	0.0	0.0
Usable Categorical	0.99	0.45
Usable-with-Extraction	0.0	0.0
Context-Specific	0.01	0.50
Unusable	0	0.05

Model Predictions



Data Scientist only inspects the less confident columns