Plan

3) identify variables

Prof. Dr. Jan Kirenz

Mainly relevant for classical ML projects (not Gen AI)

4 Major **types** of data science projects

Features

	Structured data: Variables	**Unstructured data:** Labels
Small data (n <= 10,000)	Type A	Type C
Big data (n > 10,000)	Type B	Type D
	Feature engineering	Humans label data / Data augmentation

A "**feature**" refers to the entire column in the dataset

Feature = **Variable**

Transaction_id	in_foreign_country	size_compared_to_avg_transaction	fraud?
7485	False	0.8x	False
46854	True	21.2x	True
3521	True	1.1x	False

A "**feature value**" refers to a single value of a feature column

If we have **unstructured data**, we need to identify relevant **labels**

Computer vision

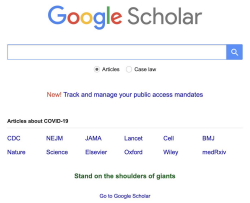
Natural language processing

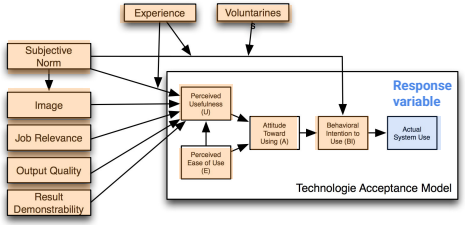
Audio processing

Source: AWS (2021)

For **structured data** problems, we need to identify potentially relevant **variables**

- **Goal**: show the primary variable of interest (response variable) and possible factors (explanatory variables)
    - explanatory variable →might affect → response variable
- Speak to **domain experts**
- Do **literature research** (e.g., using google scholar) to identify possible relationships between variables
- (In business use cases you can also use a **strategy map - see appendix)**





Venkatesh & Davis (2000)