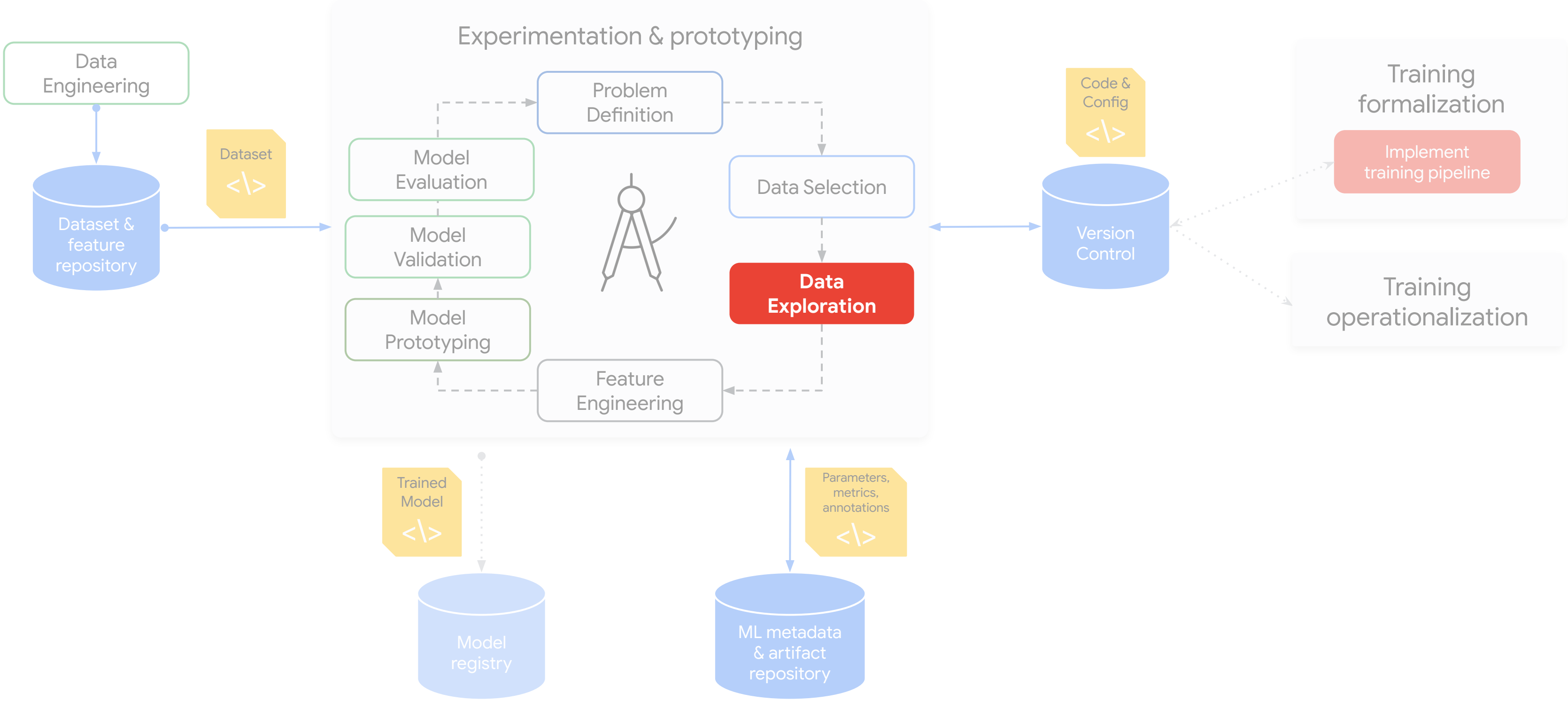
ML Development: Data Exploration

MLOps: ML Development



The MLOps **Personas**

ML Engineer

ML Researcher

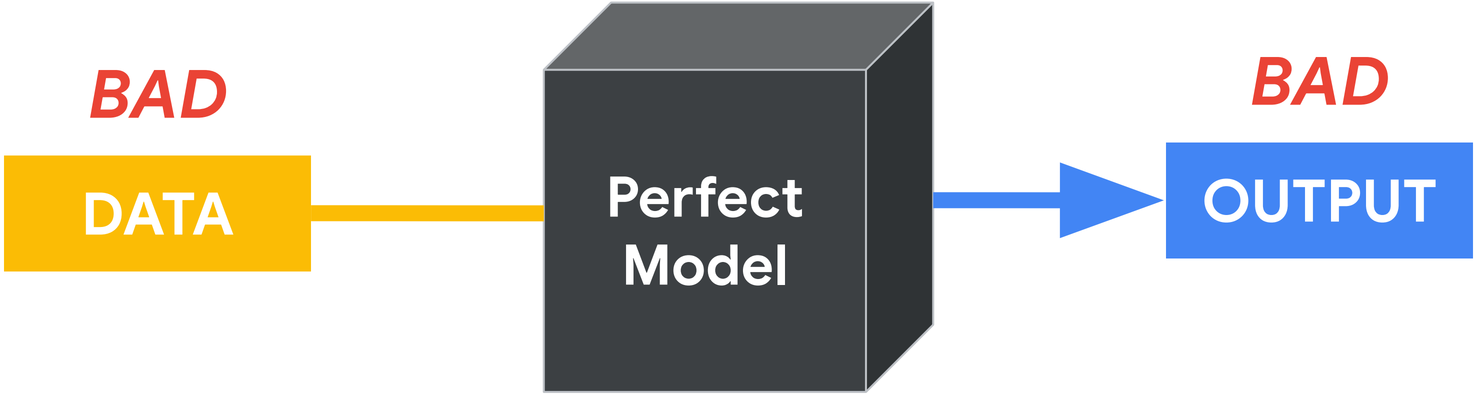
Data Scientist

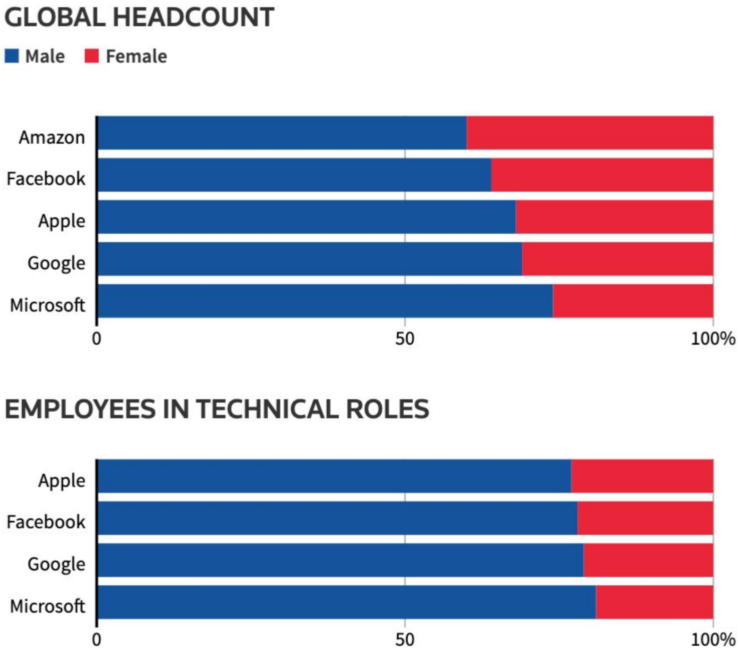
Data Engineer

Software Engineer

DevOps

Business Analyst


BAD
DATA
Perfect
Model
BAD
OUTPUT



GLOBAL HEADCOUNT

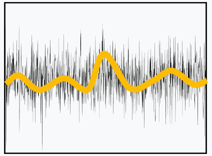
EMPLOYEES IN TECHNICAL ROLES

Note: Amazon does not disclose the gender breakdown of its technical workforce.
Source: Latest data available from the companies, since 2017.
By Han Huang | REUTERS GRAPHICS

AI **Hiring Bias**

- Caused by dataset bias
- Experiment by Amazon to **automate hiring** of developers
- Use the **past 10 years of** Amazon applicant **data to train** the model

Clean or **Noisy** Data



Label ***Noise***

Reservoir ***Sampling***

Human *Test*

Listen

Bounding boxes

Clean or **Noisy** Data

Label ***Noise***

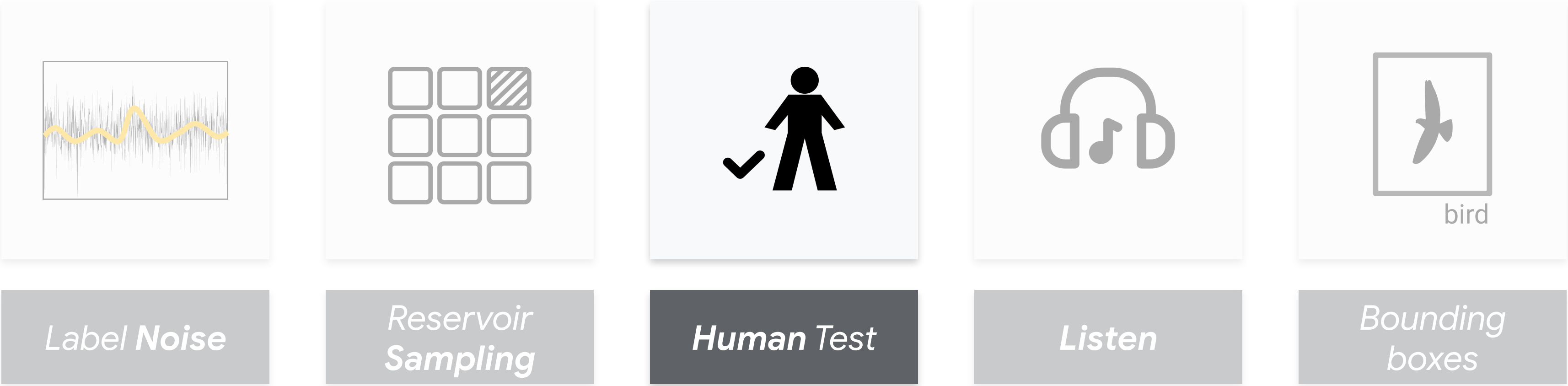
Reservoir ***Sampling***

Human *Test*

Listen

Bounding boxes

Clean or **Noisy** Data



*Label **Noise***

*Reservoir **Sampling***

***Human** Test*

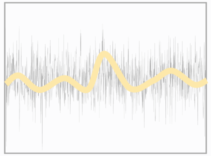
Listen

Bounding boxes

Clean or **Noisy** Data

- *Label **Noise***
- *Reservoir **Sampling***
- ***Human** Test*
- ***Listen***
- *Bounding boxes*

Clean or **Noisy** Data



*Label **Noise***

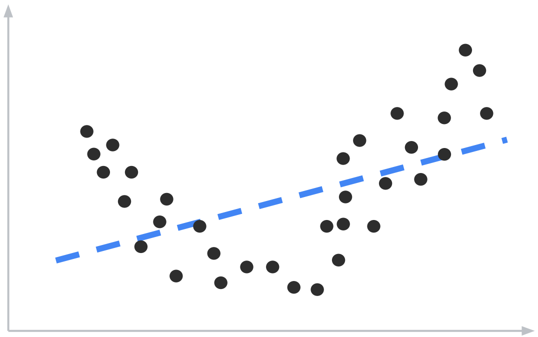
*Reservoir **Sampling***

***Human** Test*

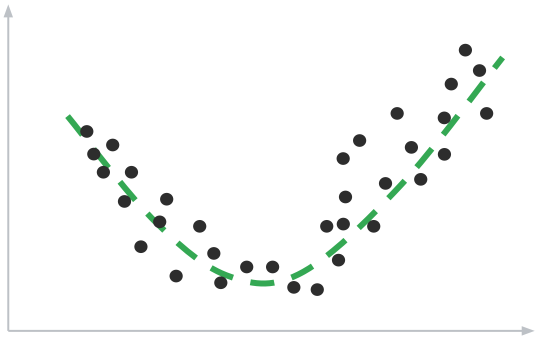
Listen

Bounding boxes

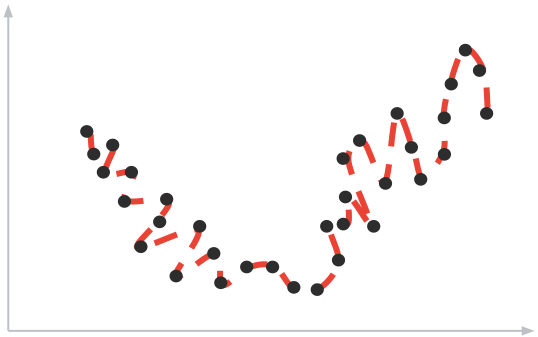
Small experiments



Underfitting



Good fit



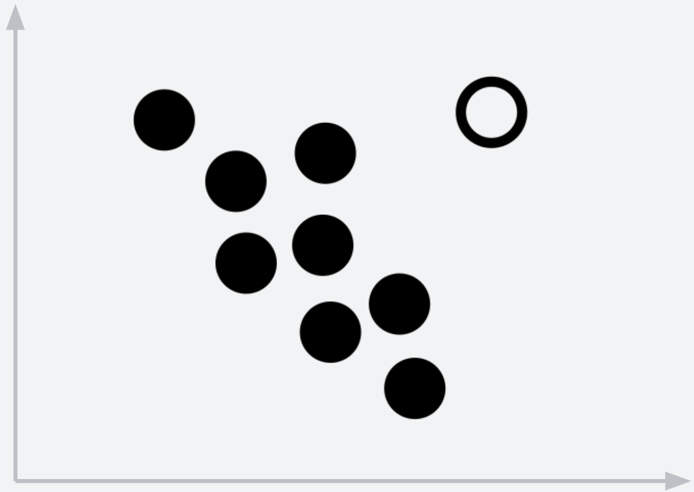
Overfitting

Data **Exploration** Protocols

1. Outliers in the data
2. Homogeneity in variance
3. Normally distributed data
4. Missing values in the data
5. Collinearity in covariates
6. Interaction between variables
7. Independence in the dataset
8. …

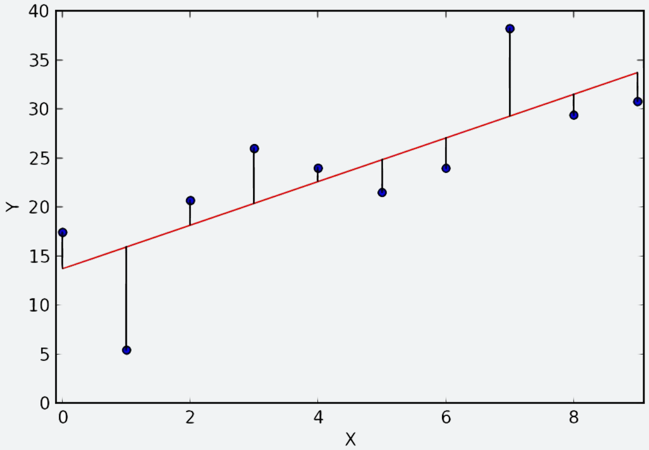
Data **Exploration** Protocols

1. **Outliers in the data**
2. Homogeneity in variance
3. Normally distributed data
4. Missing values in the data
5. Collinearity in covariates
6. Interaction between variables
7. Independence in the dataset
8. ...



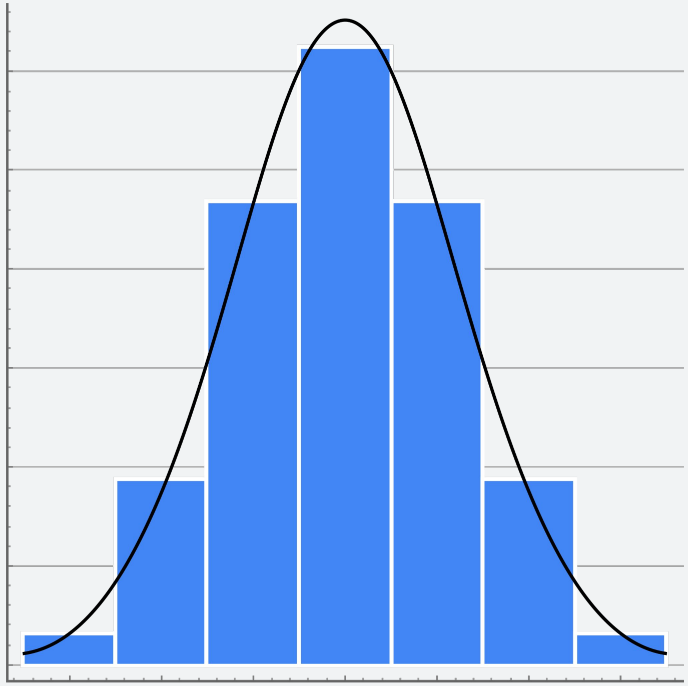
Data **Exploration** Protocols

1. Outliers in the data
2. **Homogeneity in variance**
3. Normally distributed data
4. Missing values in the data
5. Collinearity in covariates
6. Interaction between variables
7. Independence in the dataset
8. ...



Data **Exploration** Protocols

1. Outliers in the data
2. Homogeneity in variance
3. **Normally distributed data**
4. Missing values in the data
5. Collinearity in covariates
6. Interaction between variables
7. Independence in the dataset
8. ...



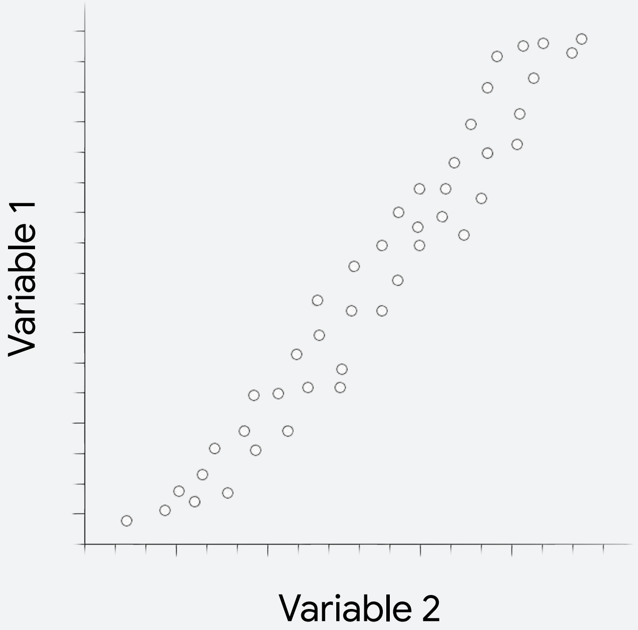
Data **Exploration** Protocols

1. Outliers in the data
2. Homogeneity in variance
3. Normally distributed data
4. **Missing values in the data**
5. Collinearity in covariates
6. Interaction between variables
7. Independence in the dataset
8. ...

Name	Trial A	Trial B	Trial C
VJ	15	32	09
Colby	11	42	
Lara	16	77	35

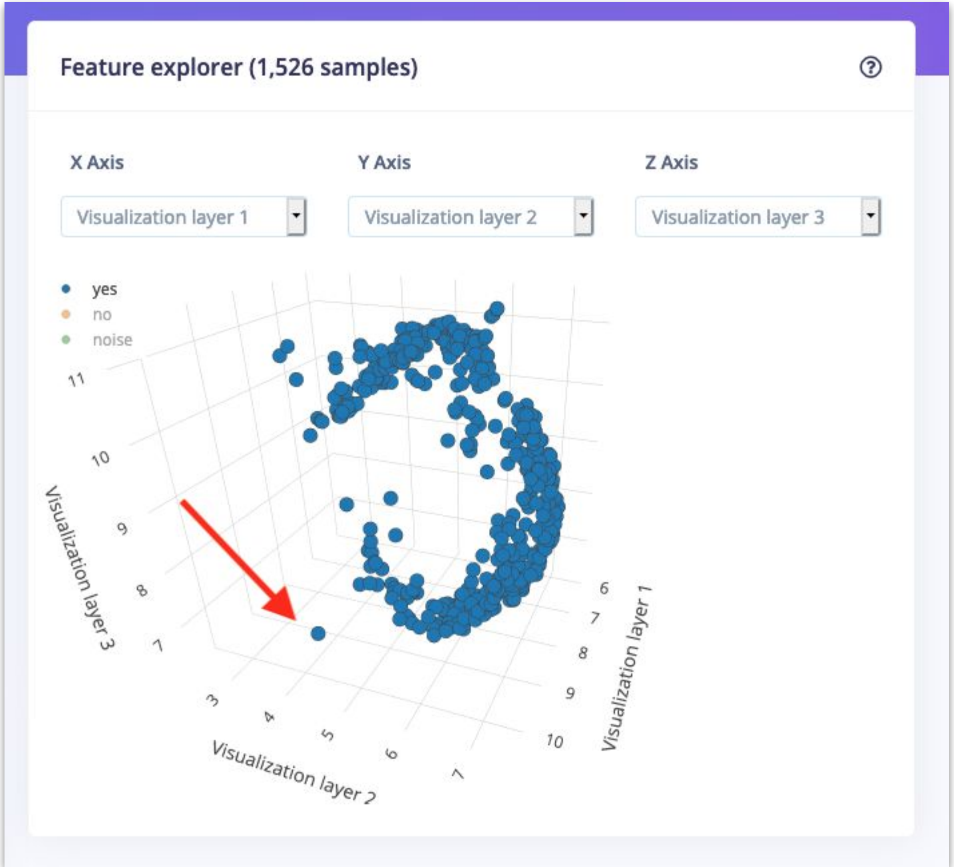
Data **Exploration** Protocols

1. Outliers in the data
2. Homogeneity in variance
3. Normally distributed data
4. Missing values in the data
5. **Collinearity in covariates**
6. Interaction between variables
7. Independence in the dataset
8. ...



Data **Exploration** Protocols

1. Outliers in the data
2. Homogeneity in variance
3. Normally distributed data
4. Missing values in the data
5. Collinearity in covariates
6. Interaction between variables
7. Independence in the dataset
8. …



Feature explorer (1,526 samples)

X Axis: Visualization layer 1

Y Axis: Visualization layer 2

Z Axis: Visualization layer 3