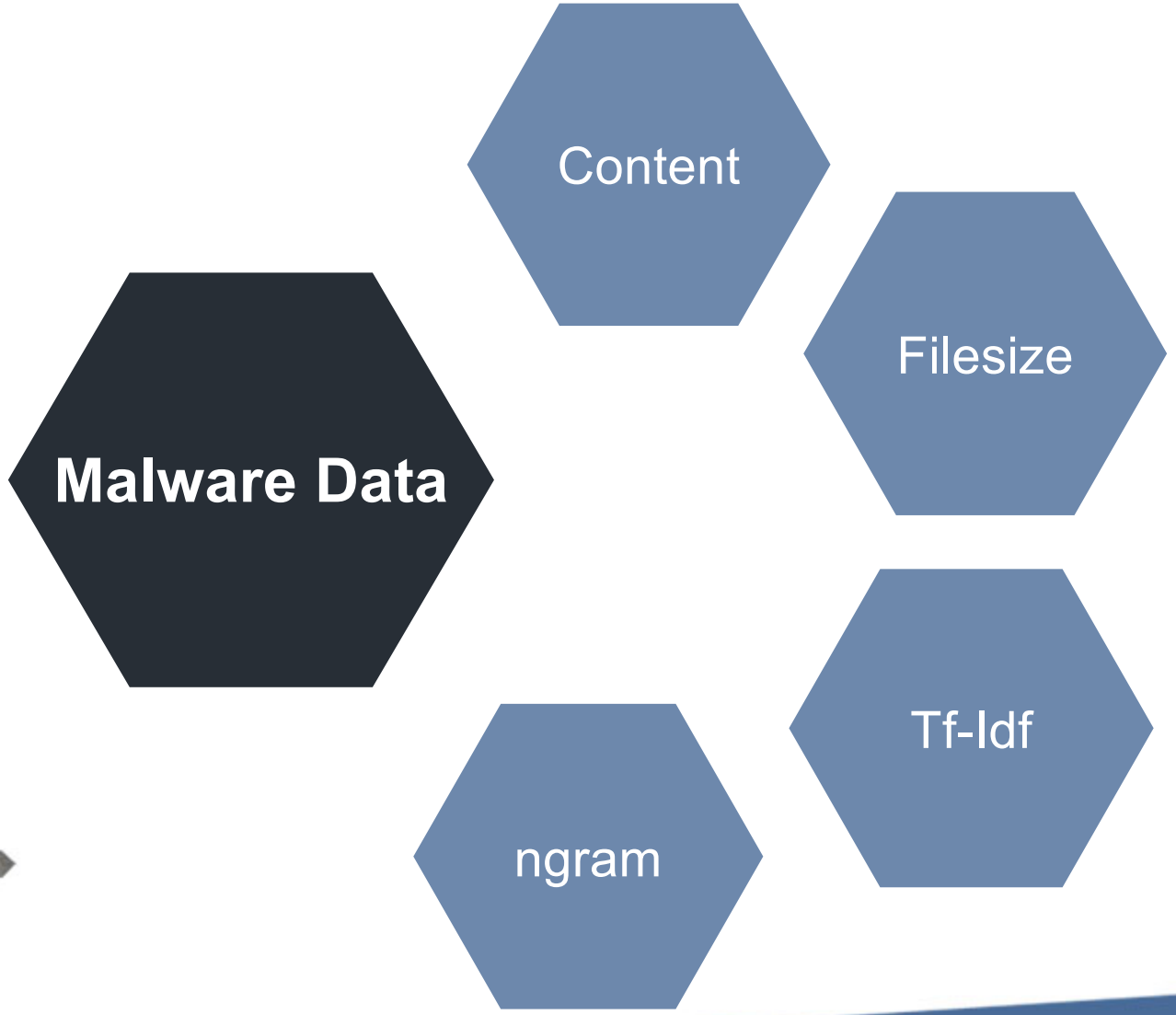
Malware Classification

Team Pardus
Haixing Dai
Xian Lei
Caleb

Feature Extraction

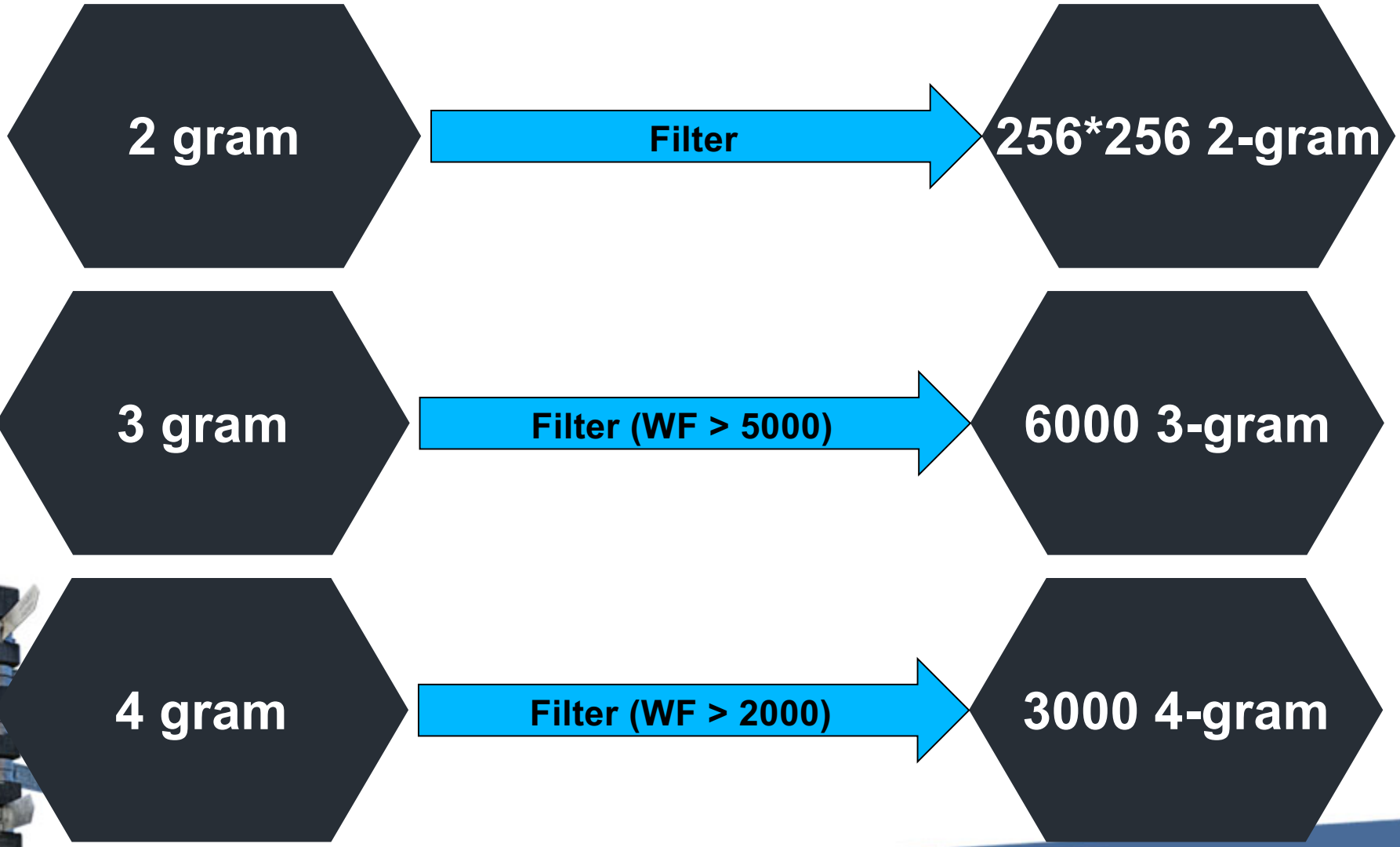


N-gram

N-gram Feature

- Using small sample to gather useful n-gram
- Extract ngram list in data
- Predict

N-gram


2 gram
Filter
256*256 2-gram
3 gram
Filter (WF > 5000)
6000 3-gram
4 gram
Filter (WF > 2000)
3000 4-gram

Models

1. Decision Tree Classifier
2. Random Forest Classifier
3. Extra Trees Classifier

Accuracy

	Content	File-size	2-gram	3-gram	4-gram	All-features	Best Solution
Accuracy	30%	80%	92%	97%	99%	96%	99.2%

Future Discussion

Find more golden features

- Asm Data
- Word2vec
- xg-boost to make a combined model

Using more models

- LSTM model, RNN deal with the sequence data.
- Convert the file to a graph.


BEIJING
12 933 KM
BERLIN
9 575 KM
TOKYO
4 724 KM
SYDNEY
11 642 KM
RIO DE JANEIRO
6 055 KM
JERUSALEM
7 468 KM
SOUTH POLE
6248 KM
EW DELHI
296 KM
INHALA

Thank you !