RSAConference2022
San Francisco & Digital | June 6 – 9

TRANSFORM

SESSION ID: **MLAI-M01**

Assessing Vendor AI Claims Like a Data Scientist, Even if You Aren't One

Joshua Saxe

Chief Scientist
Sophos
Twitter: @joshua_saxe

Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

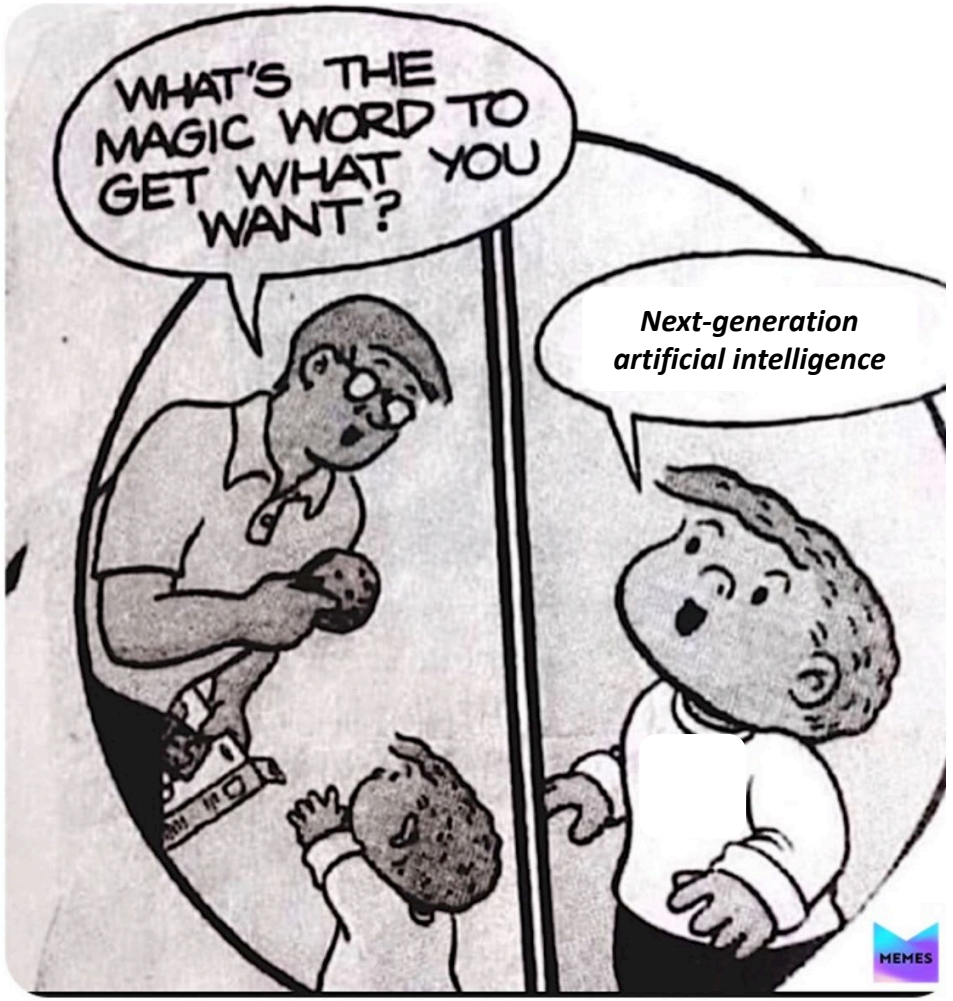
Presentation structure

- Why do we need this talk?
- How machine learning works in theory
- How security machine learning works in practice
- Questions to ask a vendor and what good answers look like
- Where to go from here

RSAConference2022

Why do we need this talk?

Why do we need this talk?

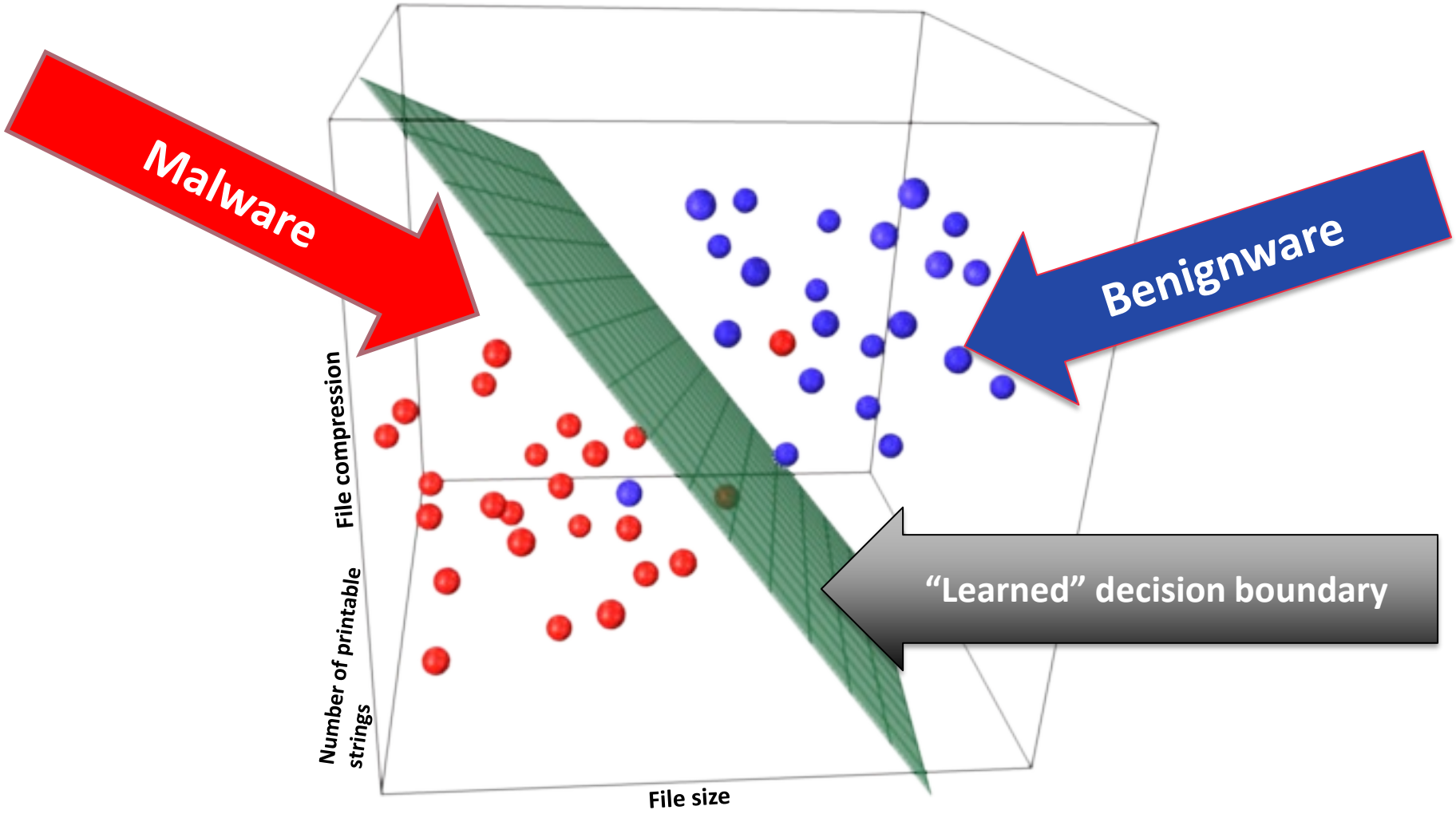


Source: https://memes.com/

RSAConference2022

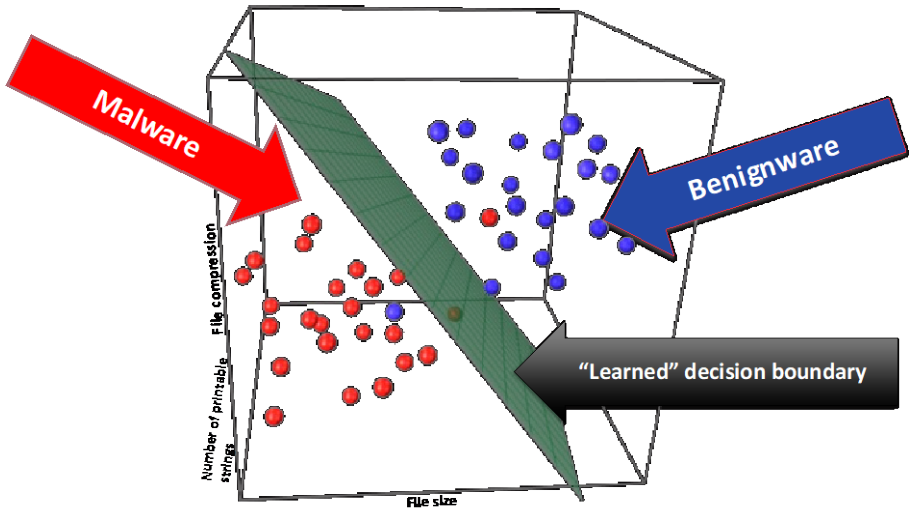
How machine learning works in theory

How does machine learning-based detection work?



Machine learning geometry operates in high dimensional spaces

- Every pixel in images could get a “dimension”
- Every field in file headers could get a dimension
- Every word in a language could get a dimension



Simple machine learning ideas operating in high dimensions can yield impressive results

https://pjreddie.com/darknet/yolo/ (Credit:Joseph Redmon, Ali Farhadi)

RSAConference2022

Security machine learning in the real world

What matters most about building effective AI systems

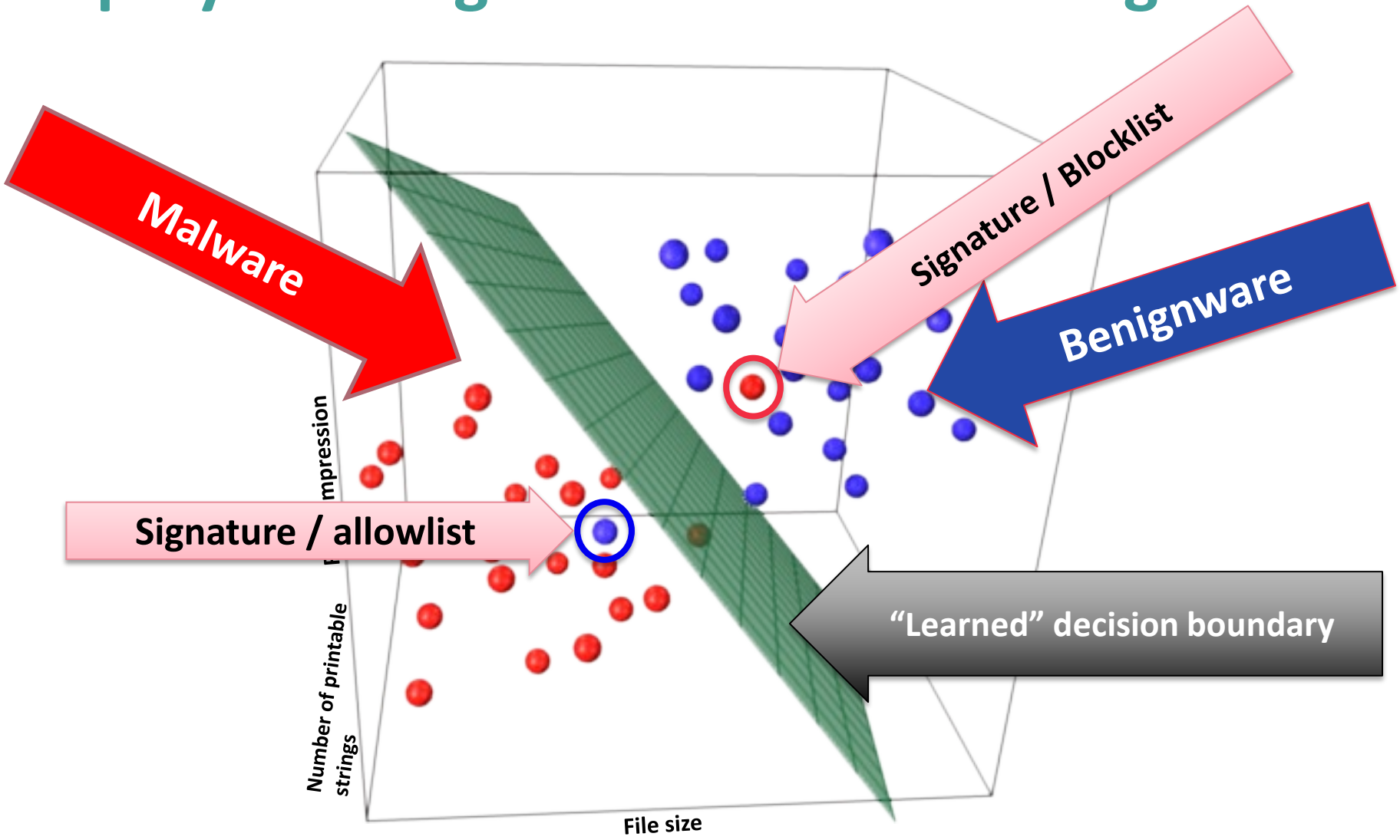
Common myths about what matters in security machine learning

- **Myth:** Mathematical sophistication is what drives accuracy
- **Myth:** The learning algorithm used is what drives accuracy
- **Myth:** Original ideas, beyond what's already public knowledge, are what drive accuracy
- **Myth:** Machine learning supersedes signatures, allowlists and blocklists

What actually matters in security machine learning

- **Reality:** Timely real-world data and data volumes drive accuracy
- **Reality:** Correct data labels drive accuracy
- **Reality:** An evaluation that accurately estimates how well a given approach will actually work drives accuracy
- **Reality:** You can't deploy machine learning well without including signatures, allowlists, and blocklists

Why signatures, blocklists and allowlists need to be deployed alongside machine learning



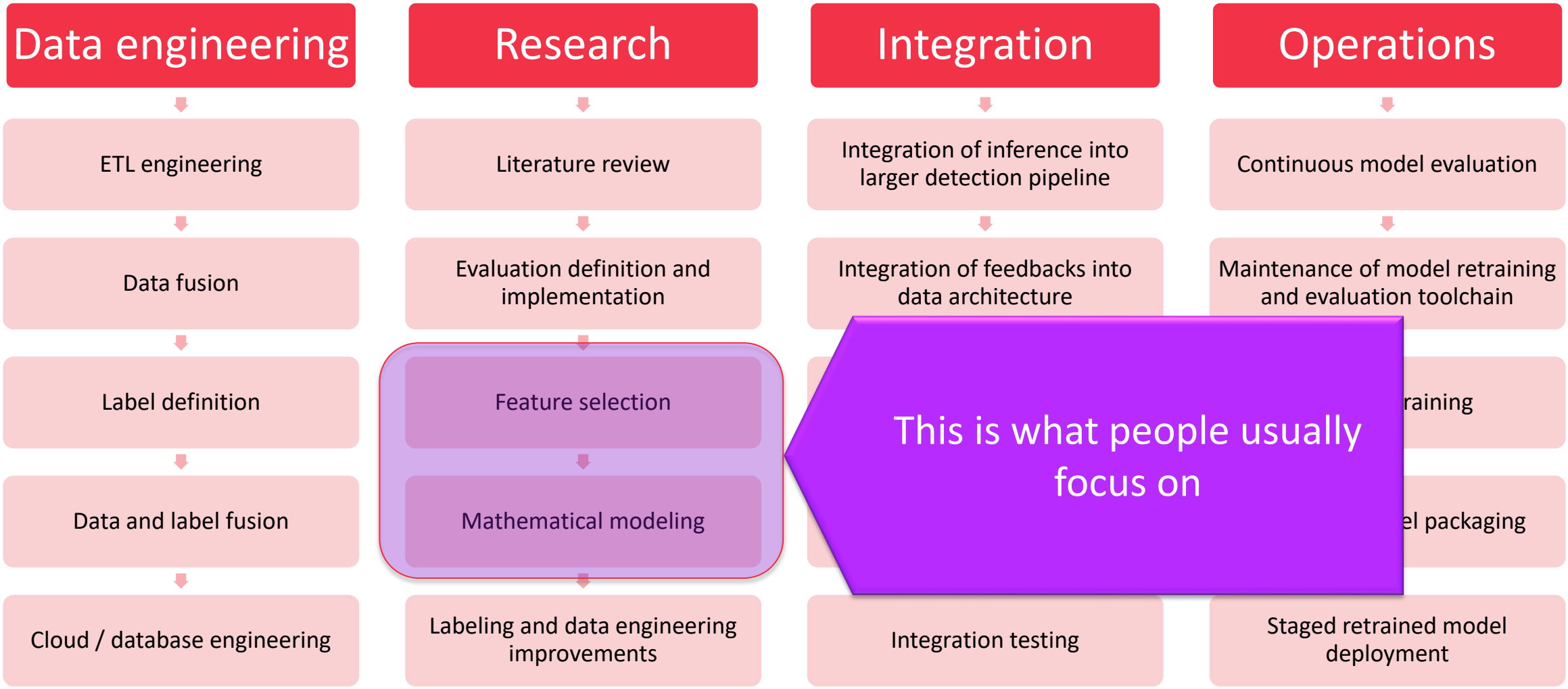
RSAConference2022

How to (politely) interrogate a security machine learning team

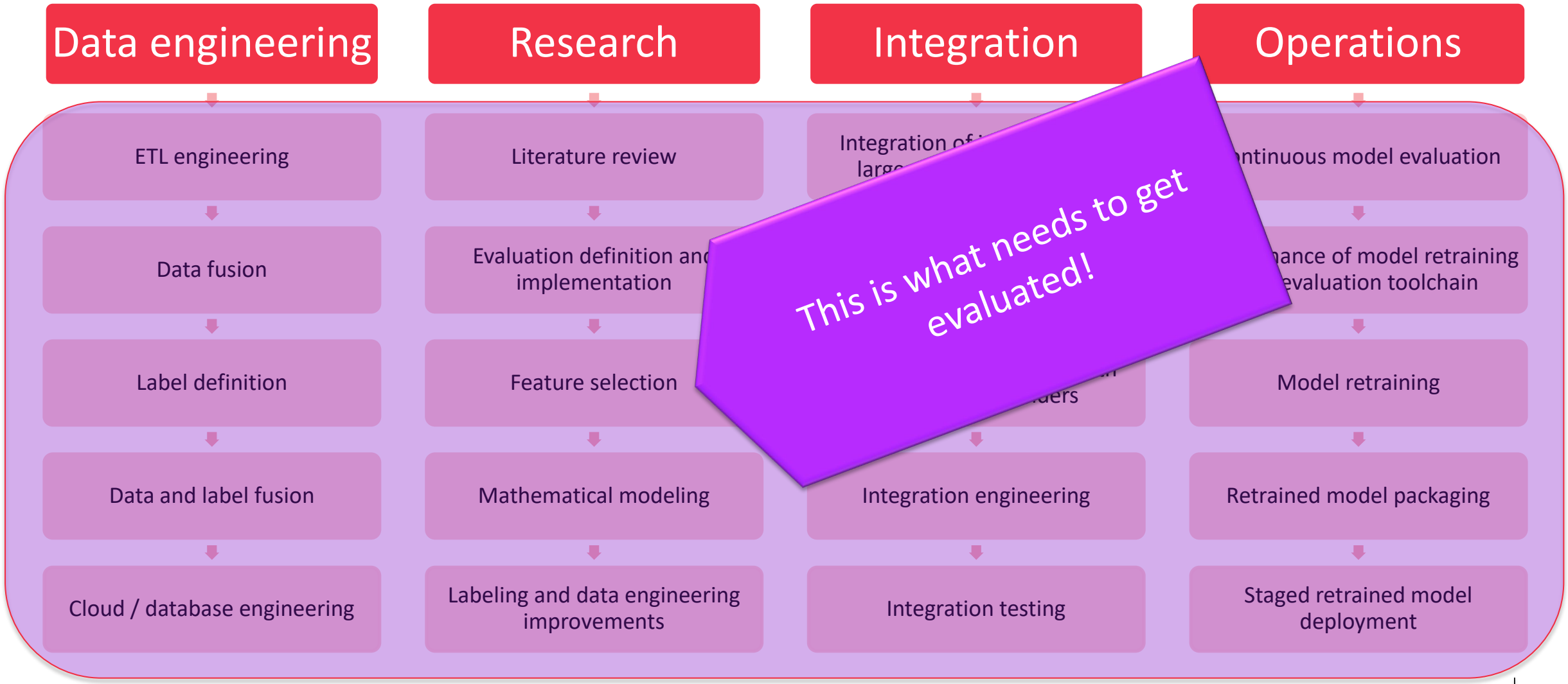
The security machine learning workflow

Data engineering	Research	Integration	Operations
ETL engineering	Literature review	Integration of inference into larger detection pipeline	Continuous model evaluation
Data fusion	Evaluation definition and implementation	Integration of feedback into data architecture	Maintenance of model retraining and evaluation toolchain
Label definition	Feature selection	Discussion and planning with product stakeholders	Model retraining
Data and label fusion	Mathematical modeling	Integration engineering	Retrained model packaging
Cloud / database engineering	Labeling and data engineering improvements	Integration testing	Staged retrained model deployment

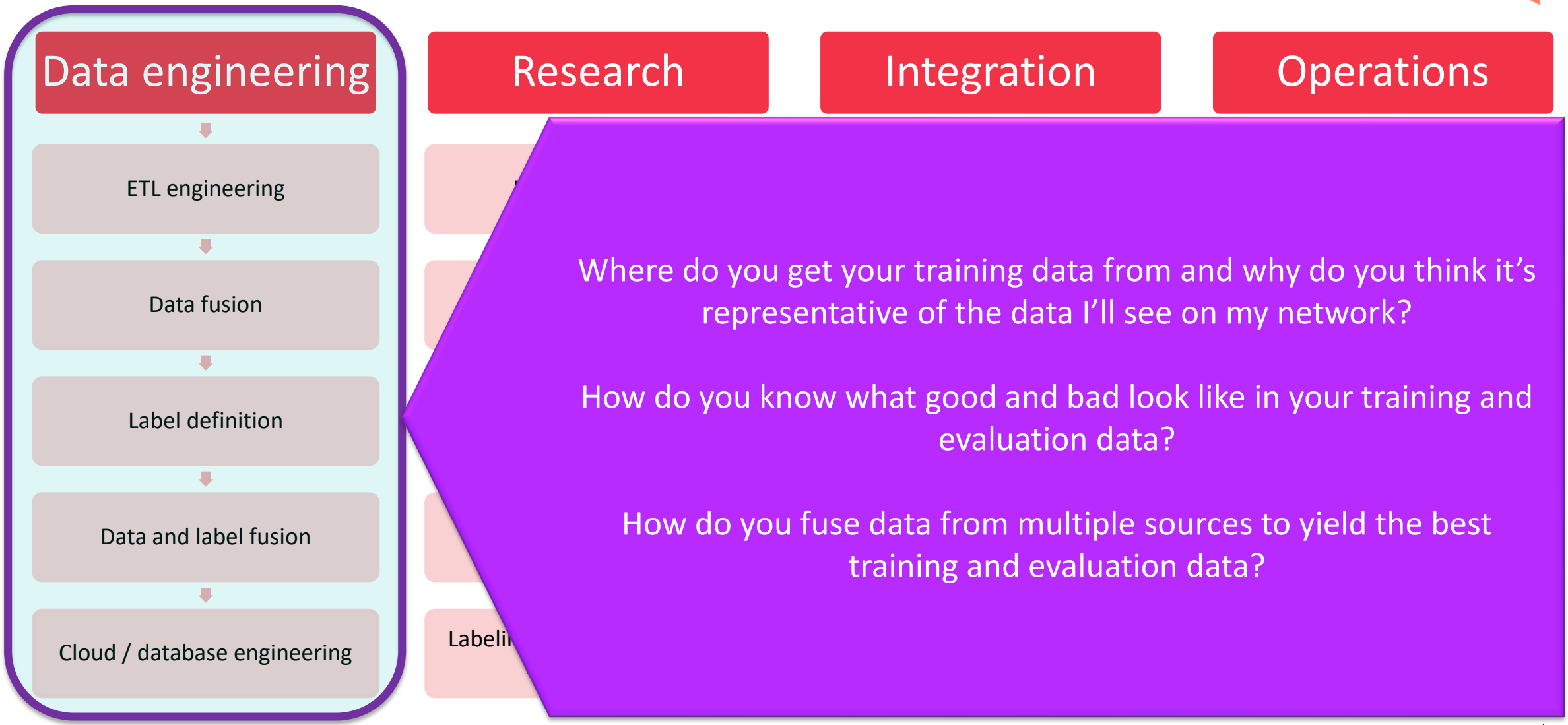
The security machine learning workflow



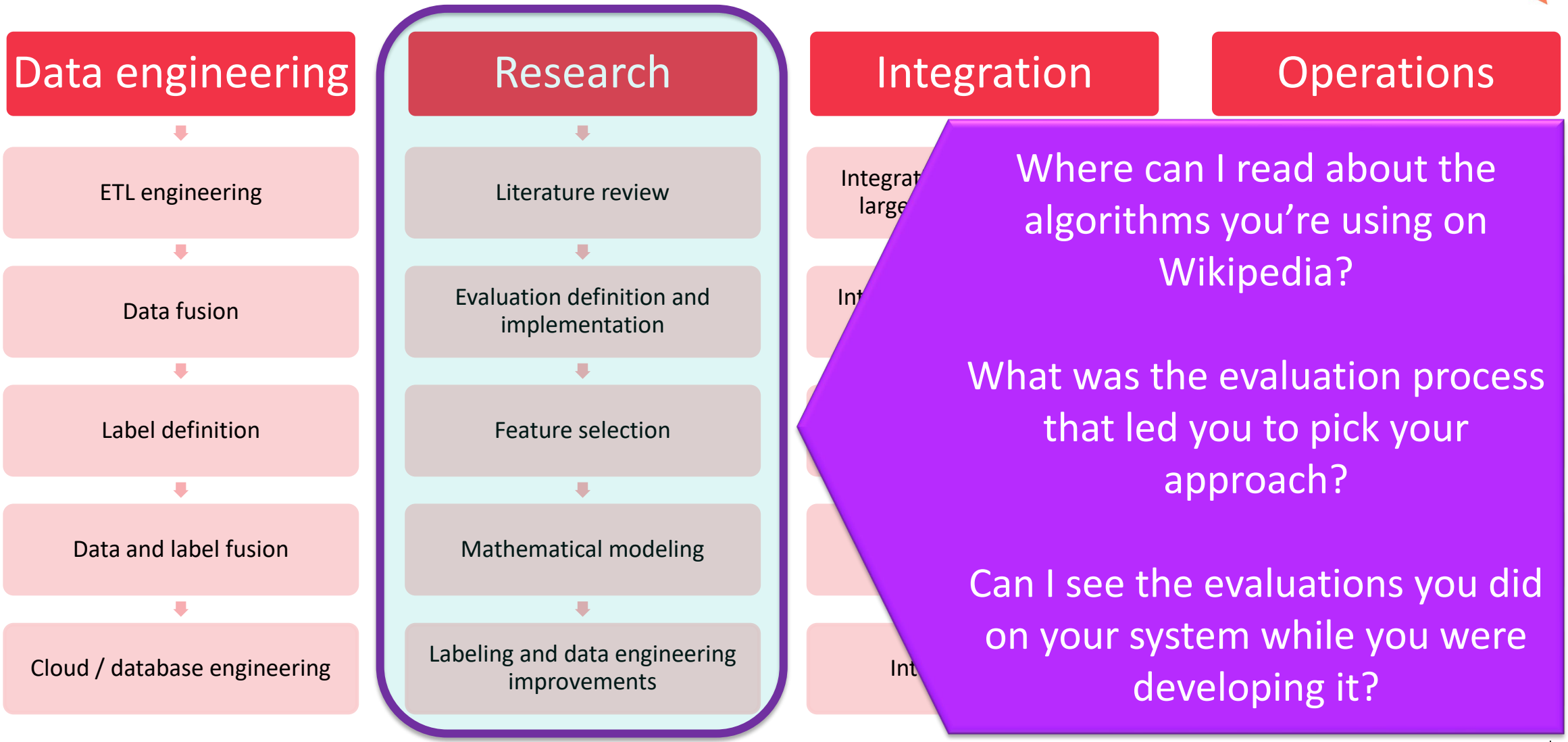
The security machine learning workflow



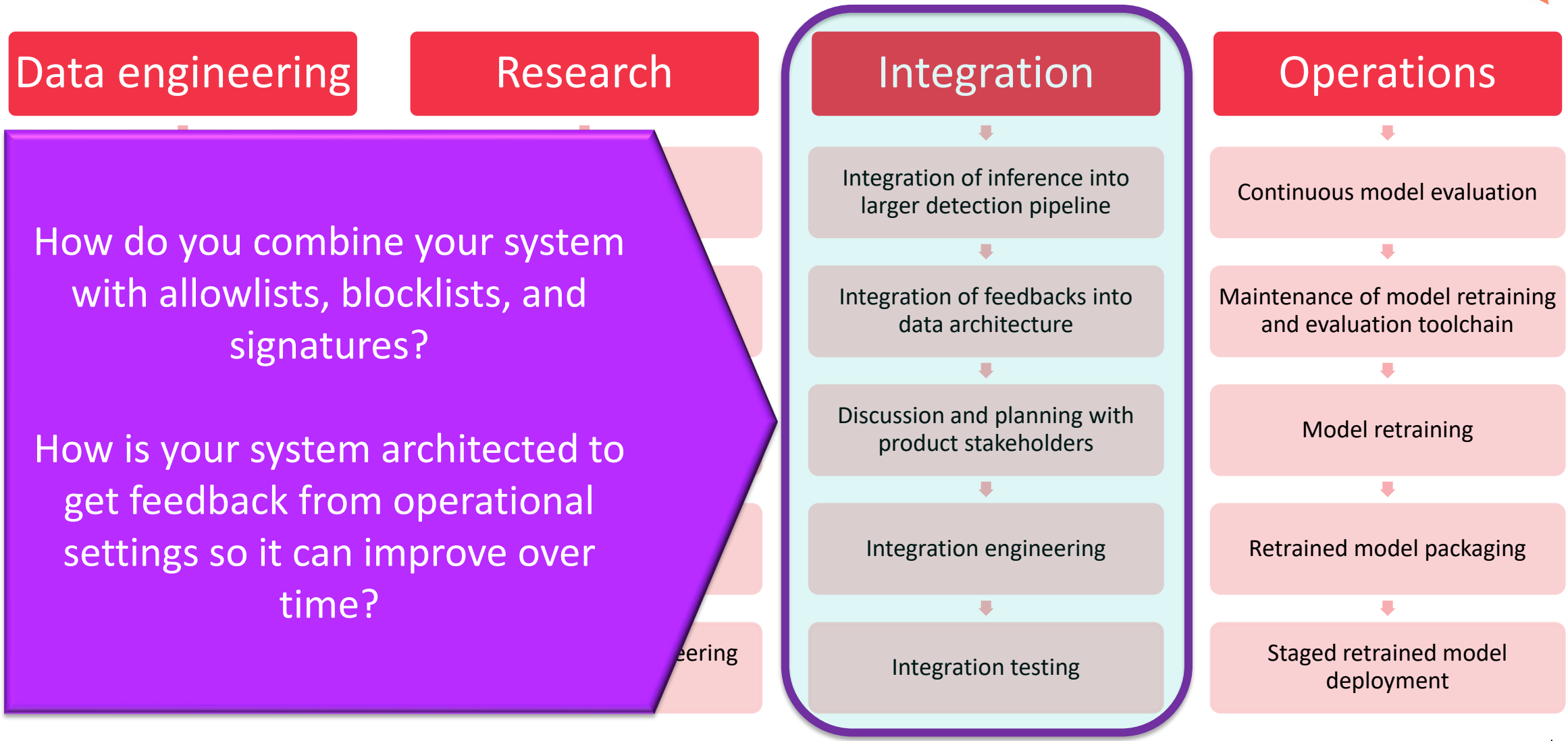
The security machine learning workflow



The security machine learning workflow



The security machine learning workflow



The security machine learning workflow

Data engineering | Research | Integration

How do you monitor the accuracy of your system in deployment?

How have you dealt with accuracy crises operationally?

How often do you retrain and redeploy your system? What triggers training and redeployment?

Operations

- Continuous model evaluation
- Maintenance of model retraining and evaluation toolchain
- Model retraining
- Retrained model packaging
- Staged retrained model deployment

Takeaways

- **Effective security ML isn't mostly about math!** Timely real-world data and data volumes drive accuracy
- The key to effective security ML is an evaluation that accurately estimates how well a given approach will actually work
- You can't deploy machine learning well without including signatures, allowlists, and blocklists
- ***Grill security vendors about all this!***

Applying the lessons of this presentation

- Immediate actions you can take
  - Use notes from this presentation to grill AI-focused security vendors
  - Use notes from this presentation to think about your own detection-oriented work
- Actions you can take in the next three months
  - Learn more about real-world machine learning; read Malware Data Science (Saxe & Sanders)
  - If you operate machine learning technology, refocus on your main levers: data and data label quality + operational hygiene