

Spark, GraphX, and blockchains: Building a behavioral analytics platform for forensics, fraud, and finance



BLOCKCYPHER

Agenda

- Why the blockchain and analytics (business)
- Blockchain data complexity
- Existing challenges
- New approach
- What to look for
- Use cases



BlockCypher Background



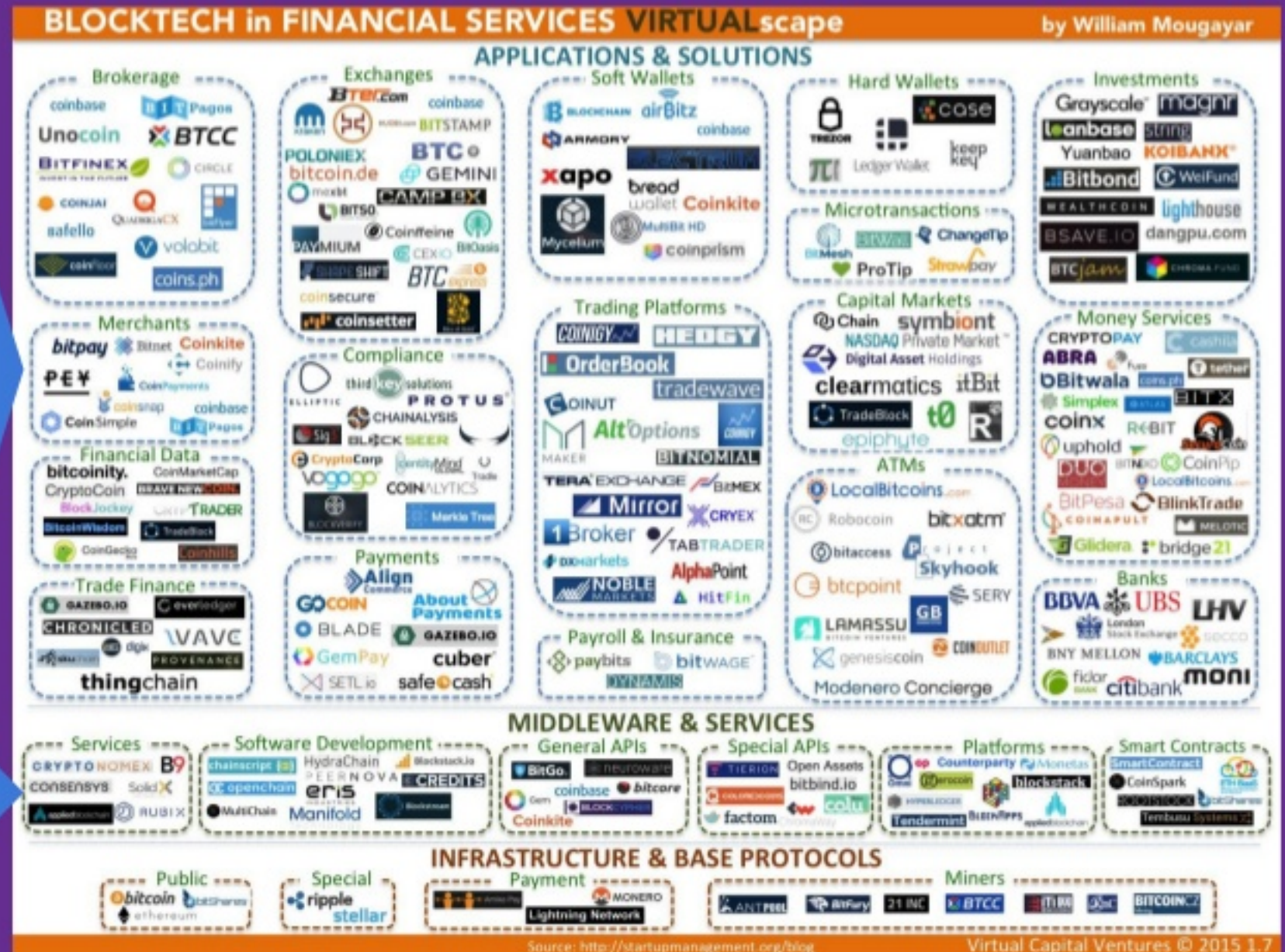
BLOCKCYPHER



BlockCypher: Blockchain Web Services

Our Customers

Who We Are



BlockCypher Blockchain Web Services

IDENTITY

Data Endpoint
Webhooks Websockets
New Transaction
Endpoint
Address Wallet API
Multisig Transaction

TRANSACTION

Payment Forwarding
API
Confidence Factor API
Microtransaction API
Asset API
Contract API
Address Balance
Endpoint

ANALYTICS


Analytics Engines and
Parameters
Create/Get Analytics
Job
Anomaly and Fraud
Detection



BLOCKCYPHER

BlockCypher Allows You to...

Outsource Infrastructure



6+
months
less time



35%+
less
costs



BLOCKCYPHER

Why the blockchain and blockchain analytics?



BLOCKCYPHER

Why the Blockchain?

Decentralized
Transactions



Transparency

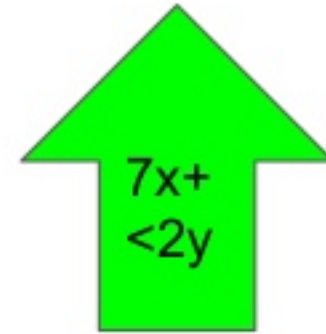
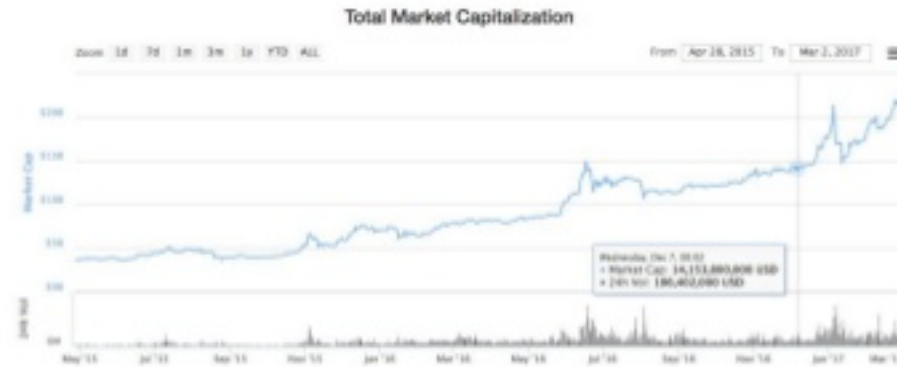


Security



Why Blockchain Analytics?

Market
Adoption



Regulation



Why Blockchain Analytics?

Technology Challenges

Blockchain Today

- data searching
- high-speed access to data for analytics

DTCC



What makes a blockchain?

(the 5 minute version)

The Problem Statement

How do you create a:

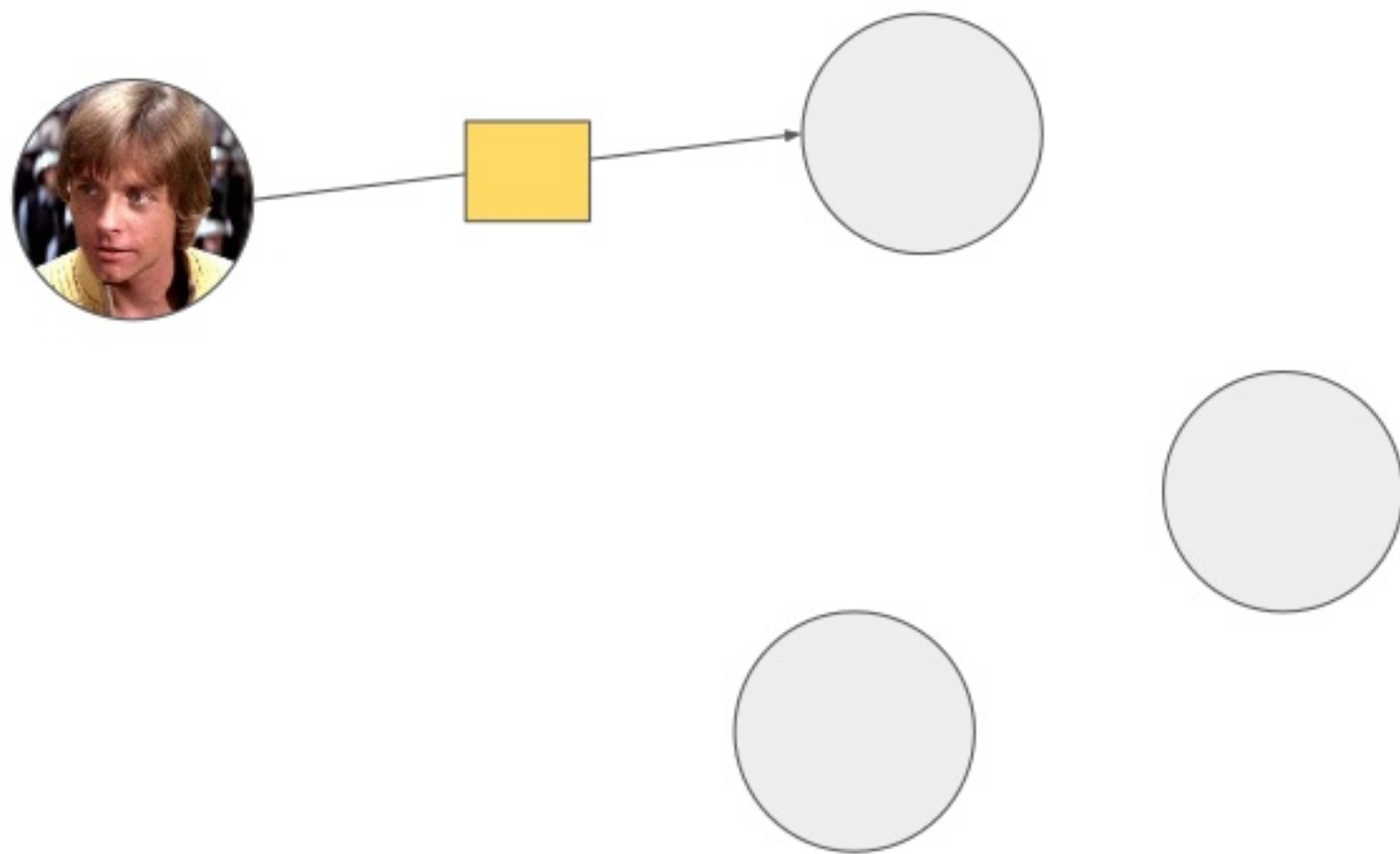
- Useful
- Decentralized
- Resilient

network?

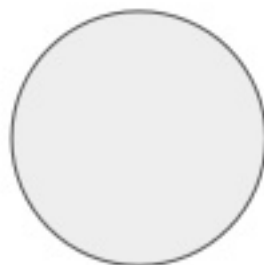
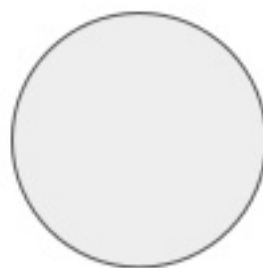
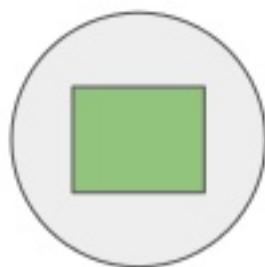
A Useful Network

- Founded on Cryptography to provide concrete, resilient **ownership** of tokens;
- Monetary policy built into the network
- A powerful scripting language to extend capabilities further:
 - **Multisignature**
 - **Smart Contracts**

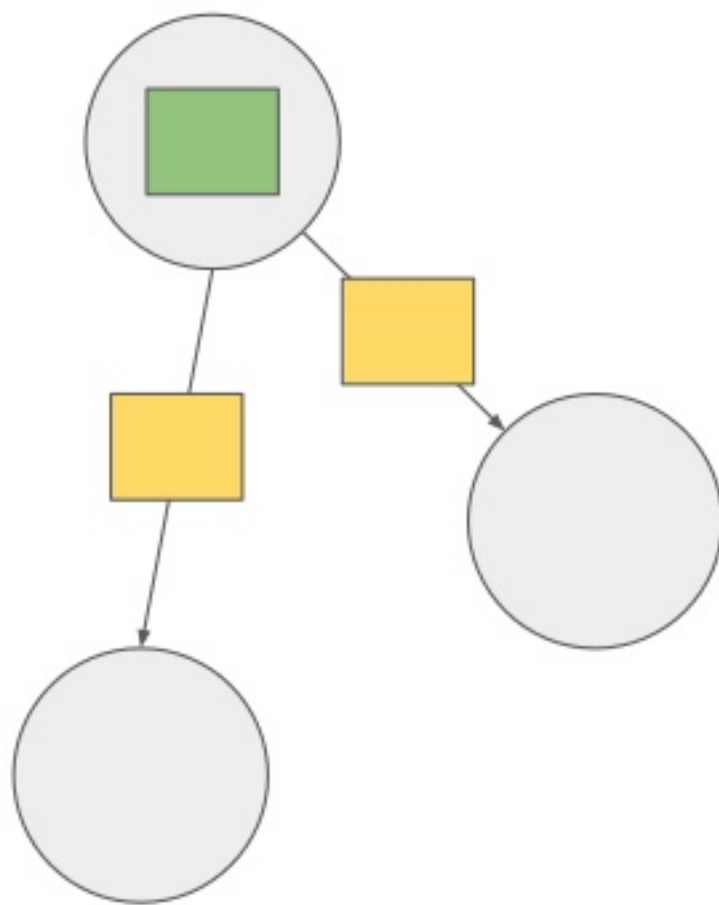
A Decentralized Network



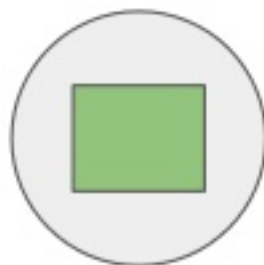
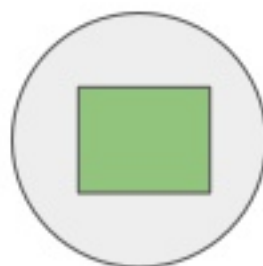
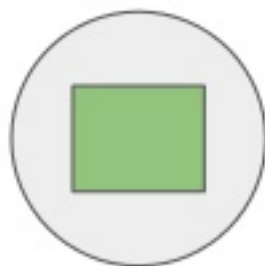
A Decentralized Network



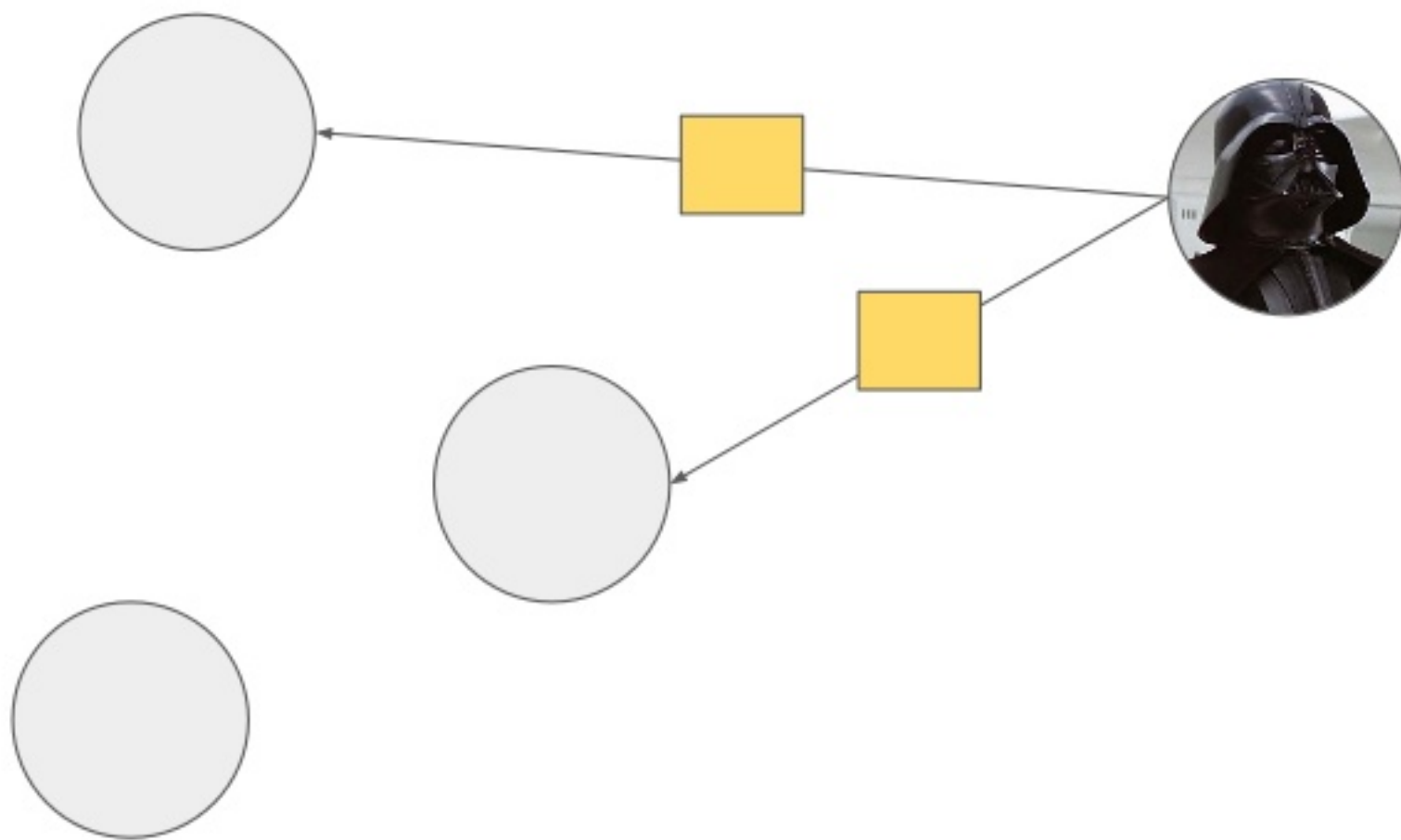
A Decentralized Network



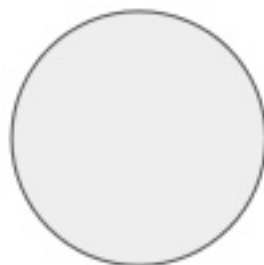
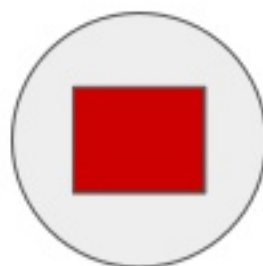
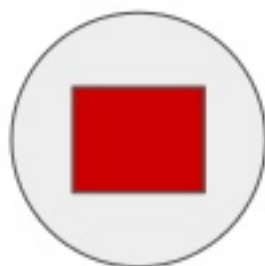
A Decentralized Network



A Decentralized Network

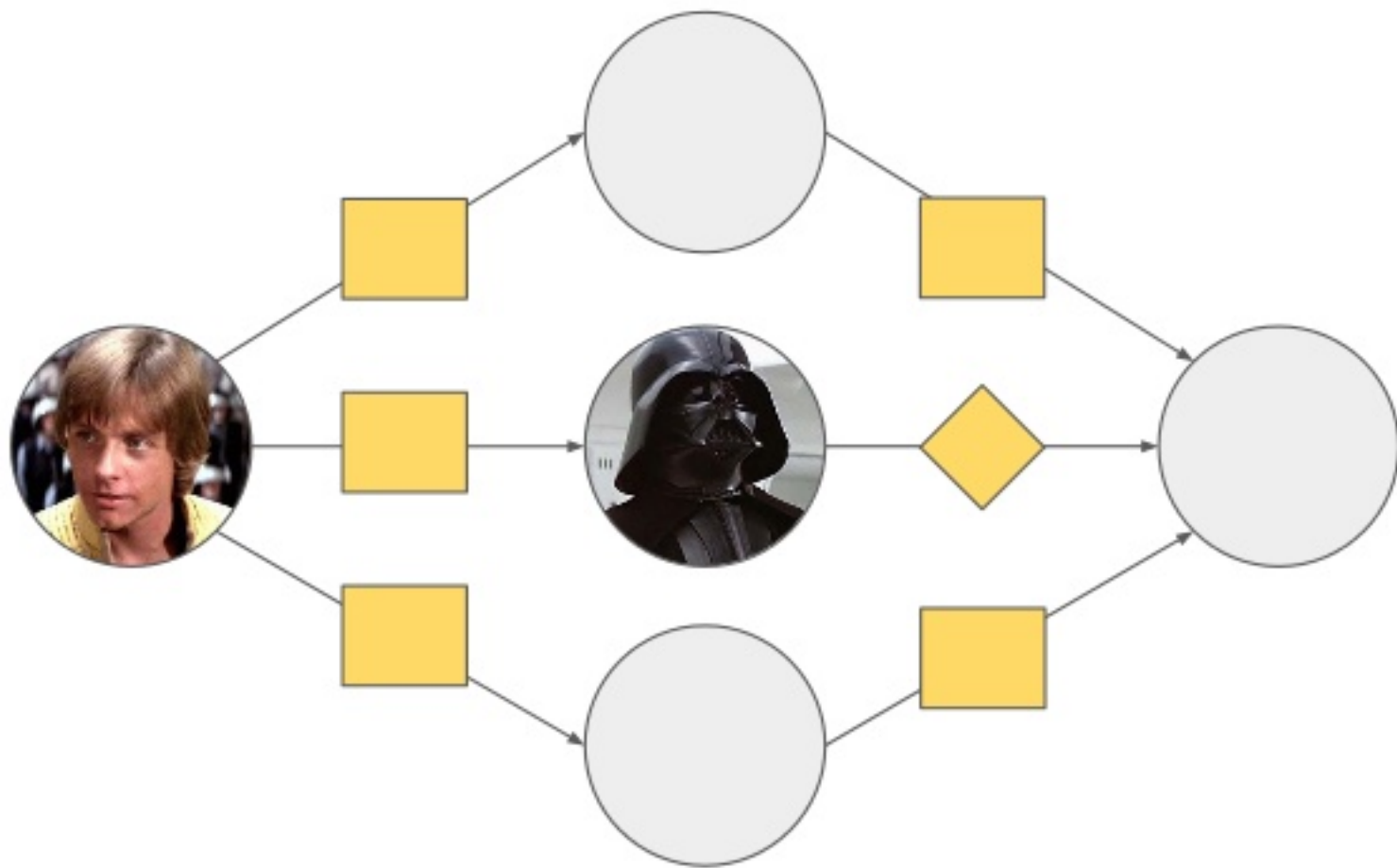


A Decentralized Network



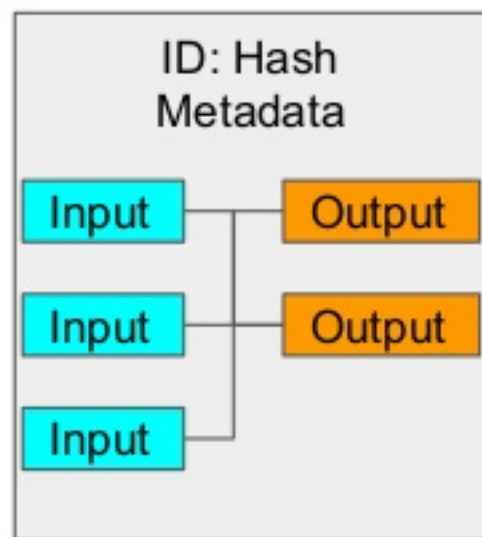
A Resilient Network

The Byzantine Generals Problem

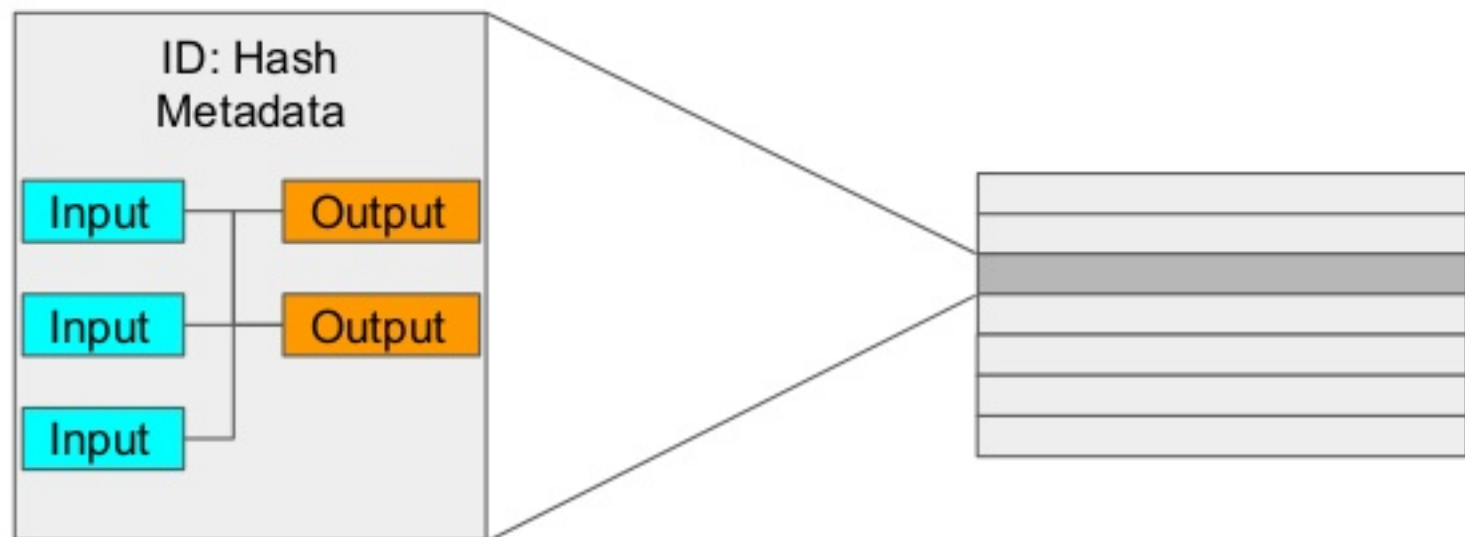


Consensus... But of What?

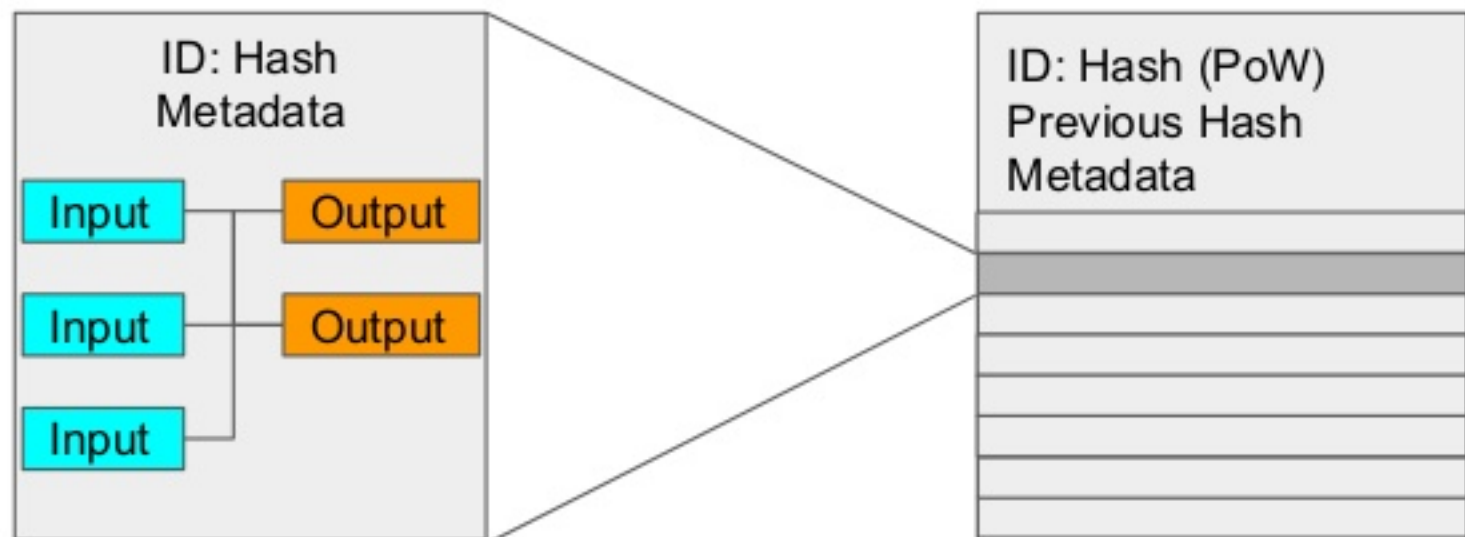
A Global Transaction Ledger



A Global Transaction Ledger

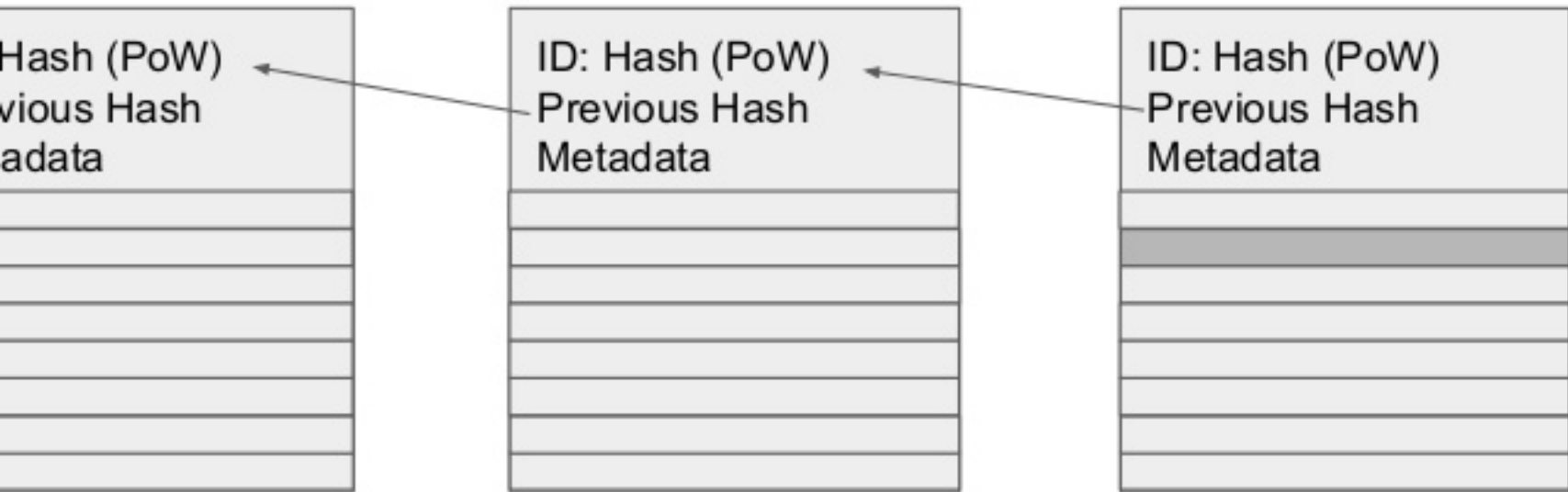


A Global Transaction Ledger

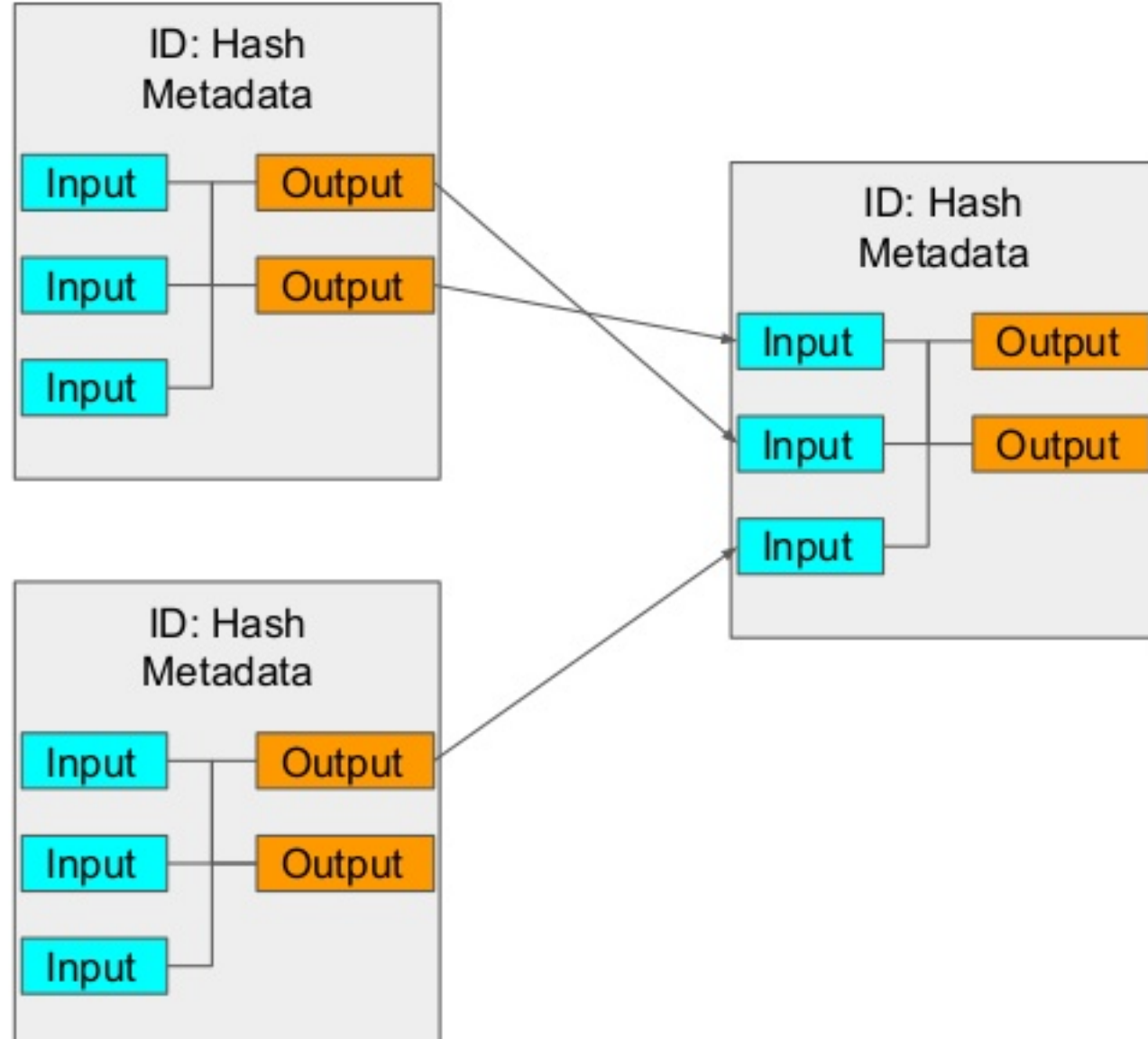


A “Block Chain”

Verifiable, including Proof-of-Work and all consensus rules, back to genesis

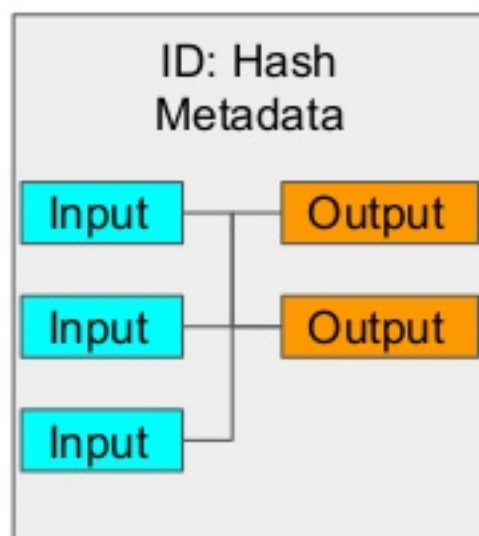


Inputs and Outputs



Inputs and Outputs

Outputs and inputs include **script**: code executed by all *fully validating* nodes to determine spending authority.



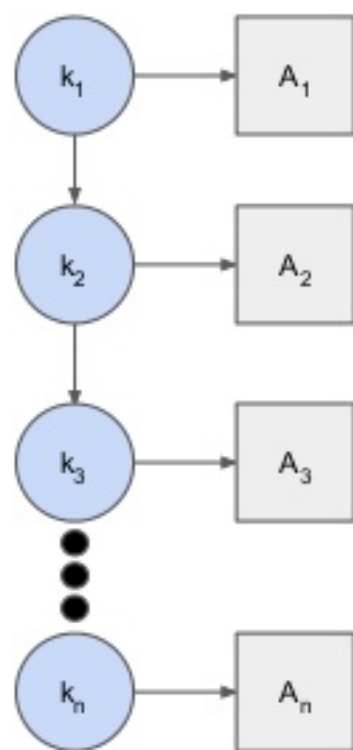
Pseudonymity in Bitcoin

Ownership identifiers are **addresses**:

- Most commonly owned by a single entity
- Sometimes a representation of not-yet-exposed code (script)

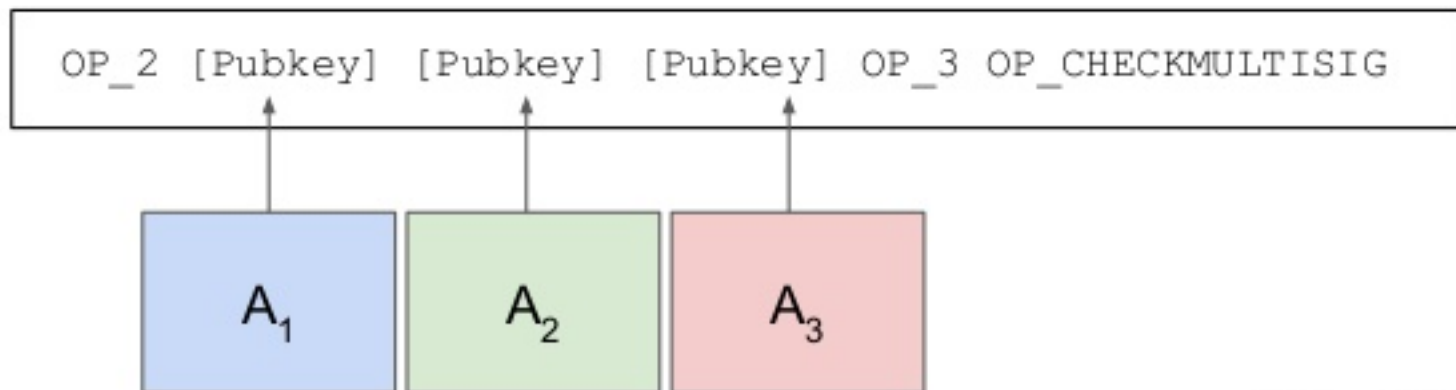
When are Identities not Identities?

1. HD Wallets



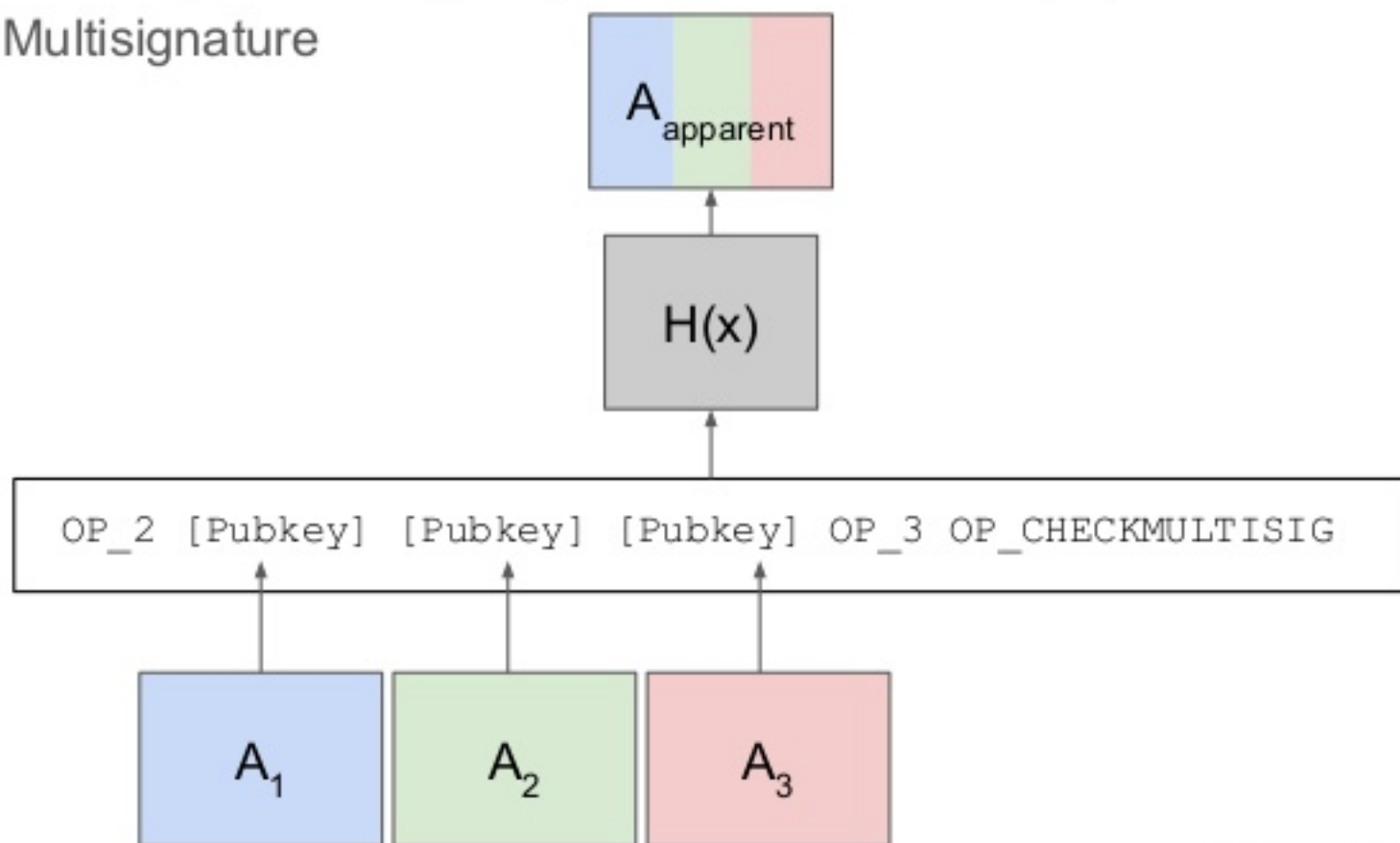
When are Identities not Identities?

1. HD Wallets
2. Pay to Script Hash (output bound to **script**):
 - a. Multisignature



When are Identities not Identities?

1. HD Wallets
2. Pay to Script Hash (output bound to **script**):
 - a. Multisignature



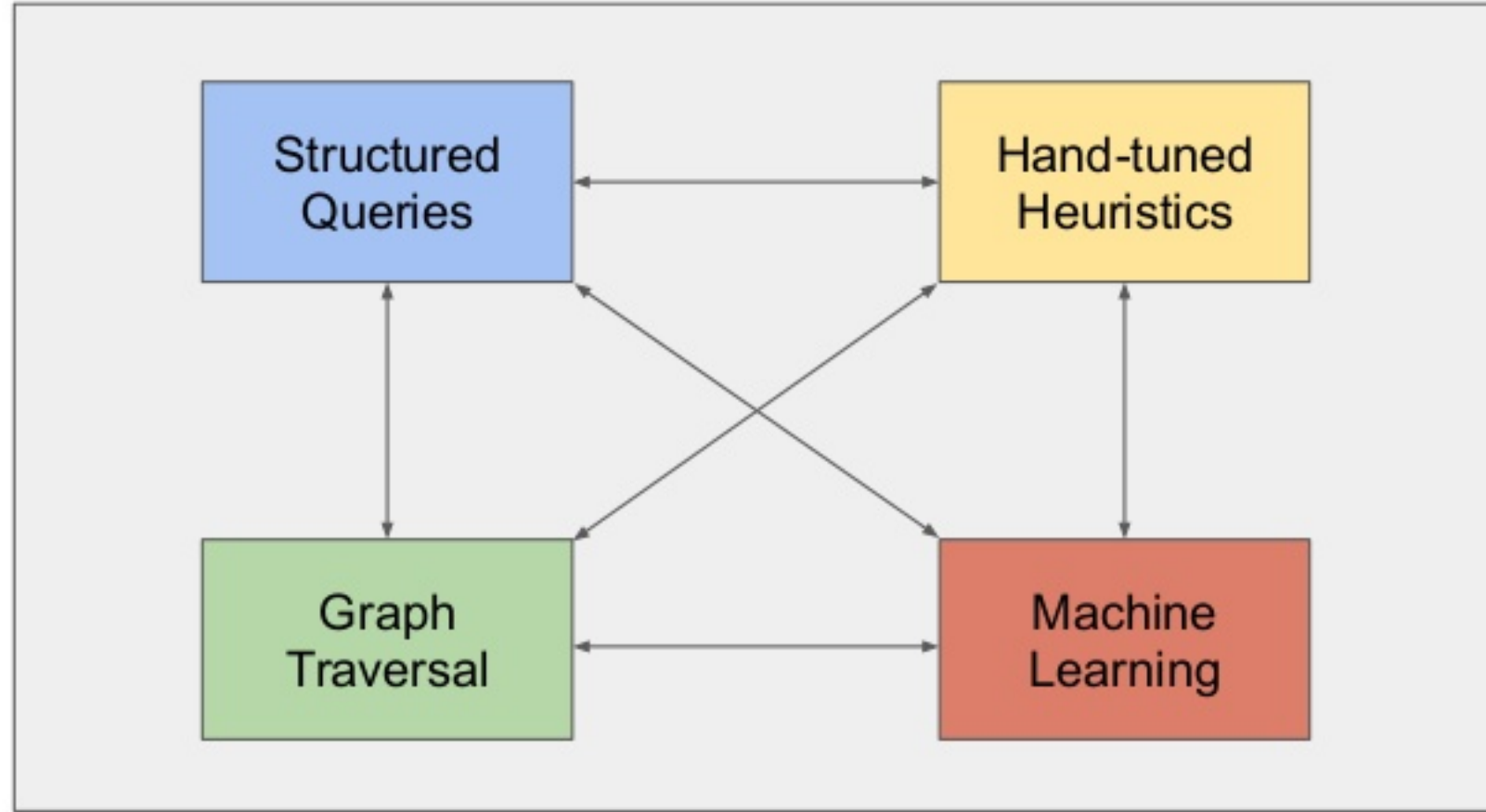
Current Tools are Lacking

Agents still use manual investigation and traversal

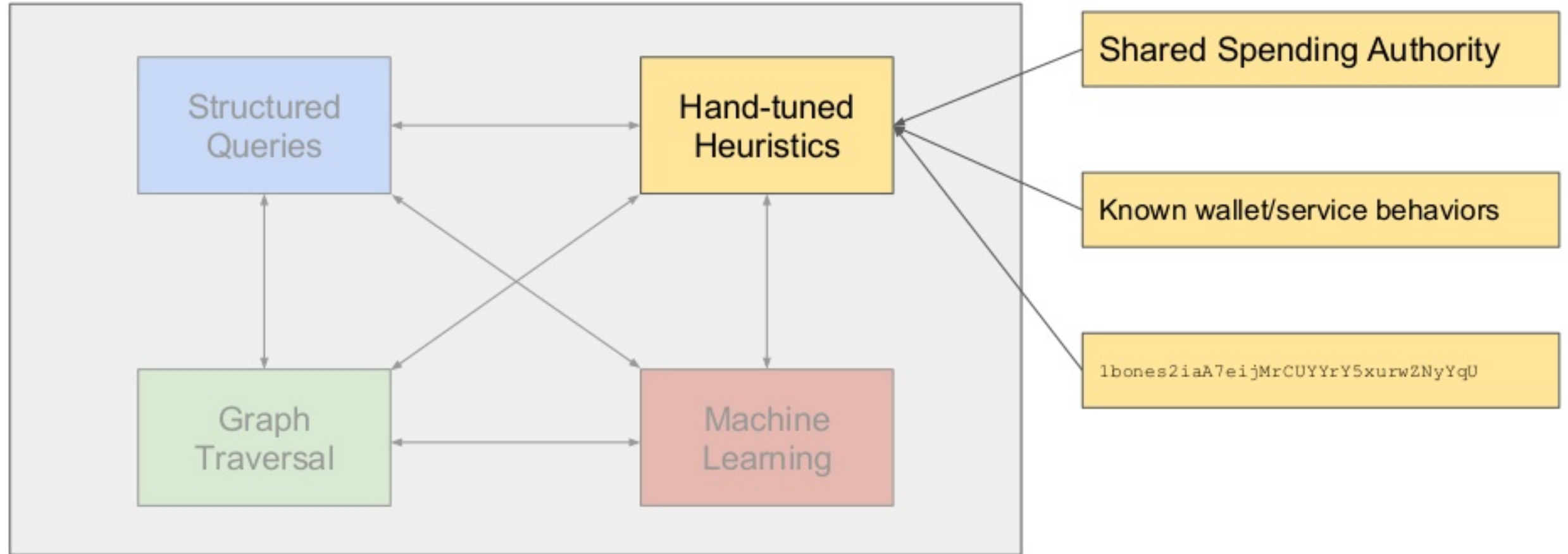
Traditional tools lack the flexibility to evolve alongside bitcoin's usage patterns and identity distributions

Behavioral Analytics bridges the gap: focus on patterns of movements, rather than easily-manipulated properties

You need a multi-layered solution

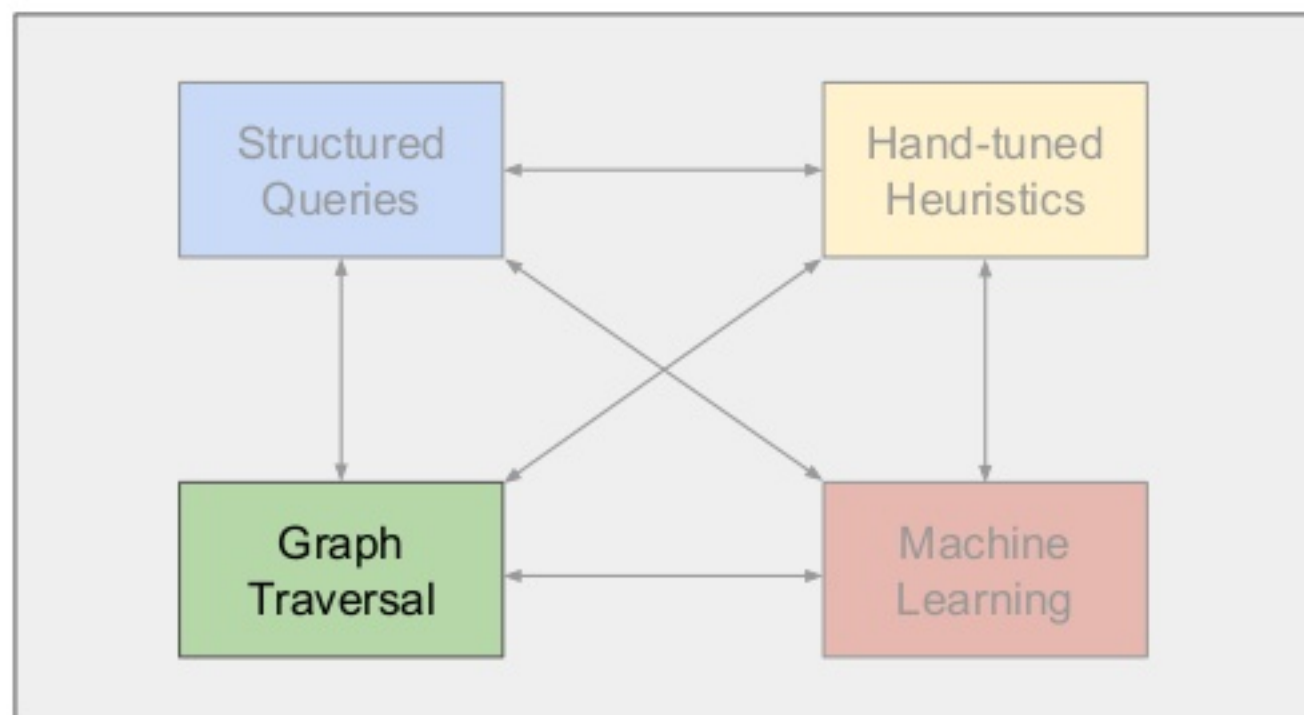


Don't discount the human element



Building a Practical Blockchain Analytics Platform for Forensics

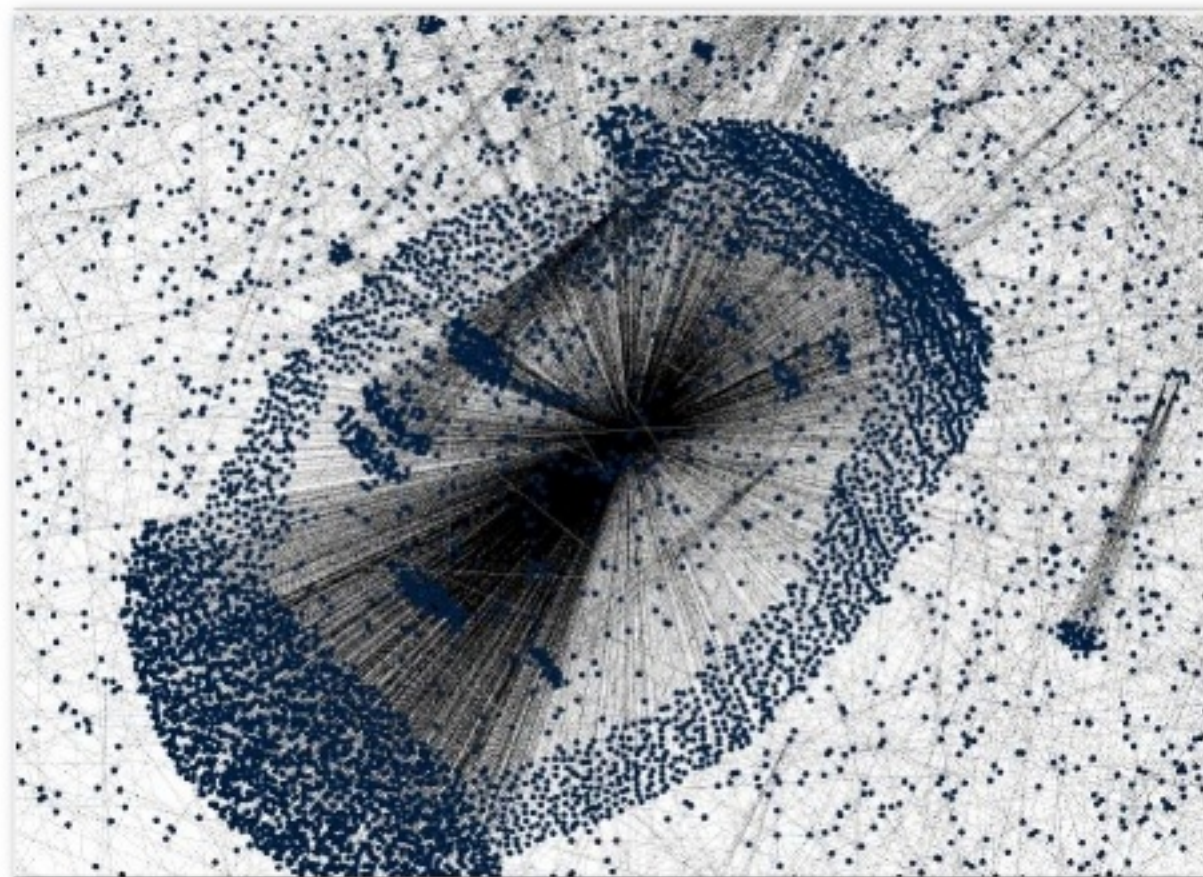
1. Representing the Graph



Graphs make natural blockchain representations...

Great for representing broader interactions, relationships, and trends

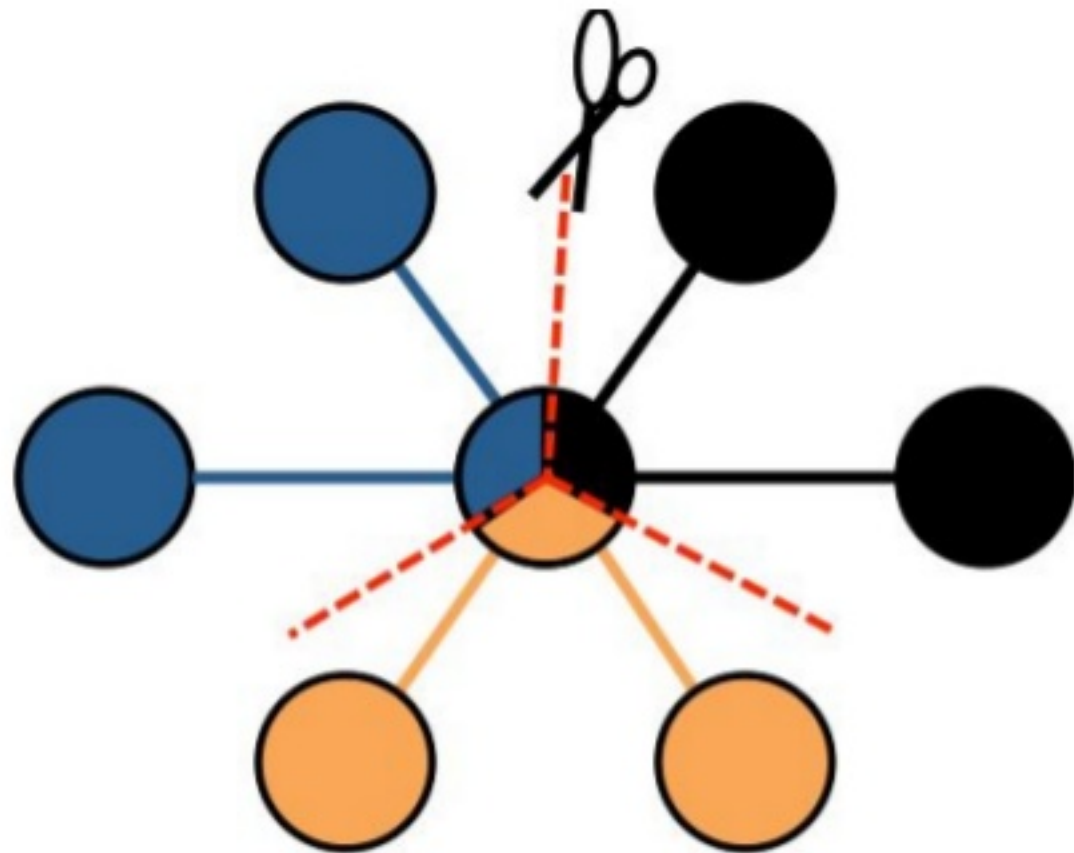
(plus they look cool)



But not all graphs are created equal.

Transformation to the Address-Address graph creates a loss of resolution and introduces assumptions

This is especially true with Spark



Implication: Operations will vary significantly in both parallelism and data locality based on edge versus vertex property distinctions.

Vertex Cut

Different graphs serve different purposes...

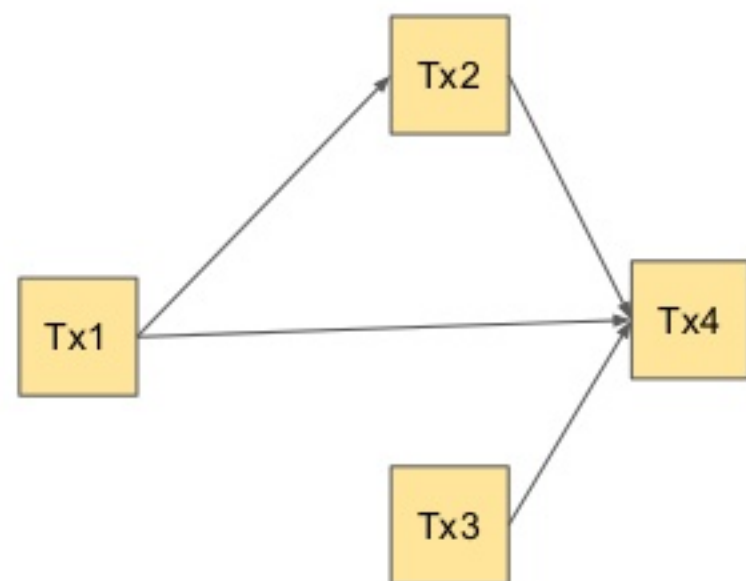
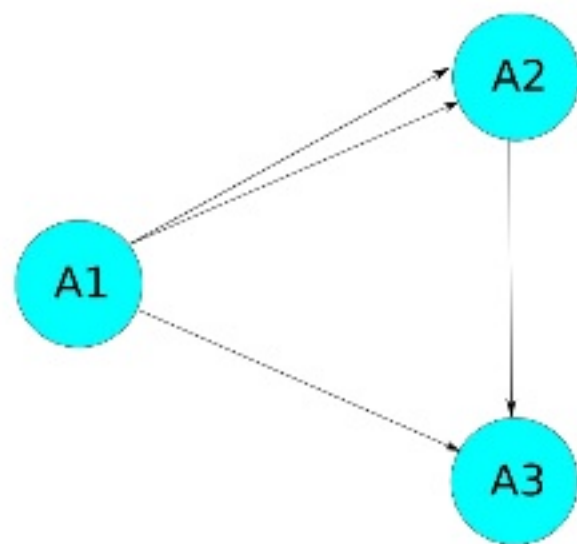
Address-Address

Address-Tx

Output-Output

Input-Output

Transaction-Transaction



... and have different properties

	Directed	Acyclic	Bipartite
Address/Address	Y	N	N
Address/Transaction	Y	N	Y
Input/Output	Y	Y	Y
Output/Output	Y	Y	N
Transaction/Transaction	Y	Y	N

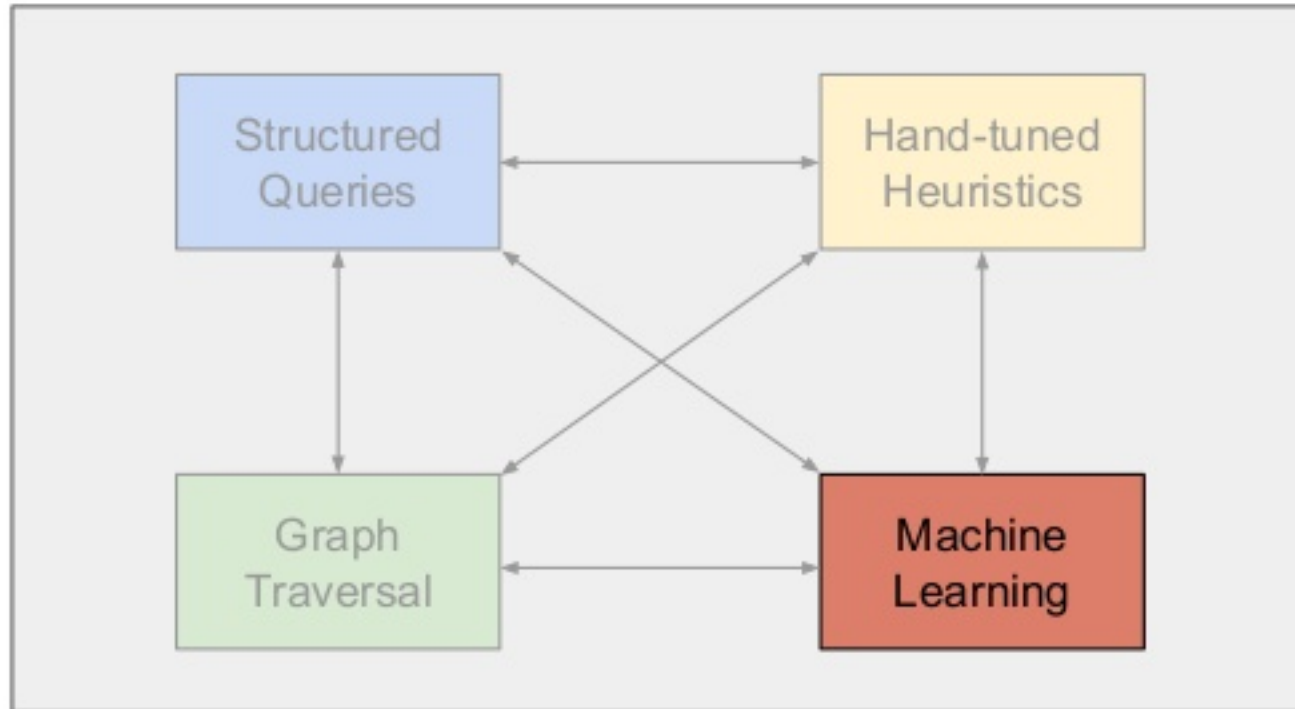
Why Spark + GraphX?

Single node graph databases great for performance re: traversal, etc.

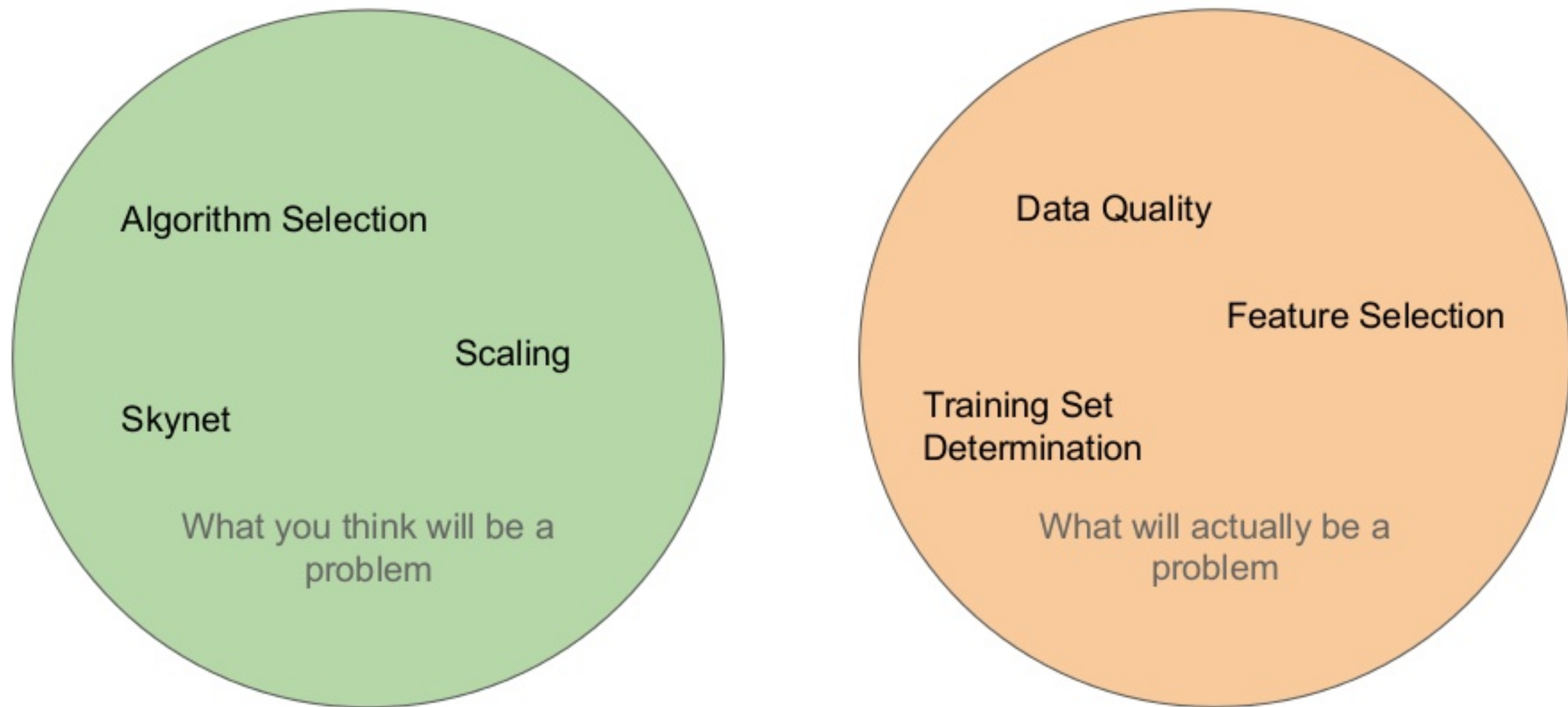
But Spark+GraphX gives:

- One infrastructure for graphs, ML models, computation
- Access to graph properties in training, eg. in-degree
- Compute + Storage scalability

2. Building the ML Pipeline



A Venn Diagram: Starting a ML Project

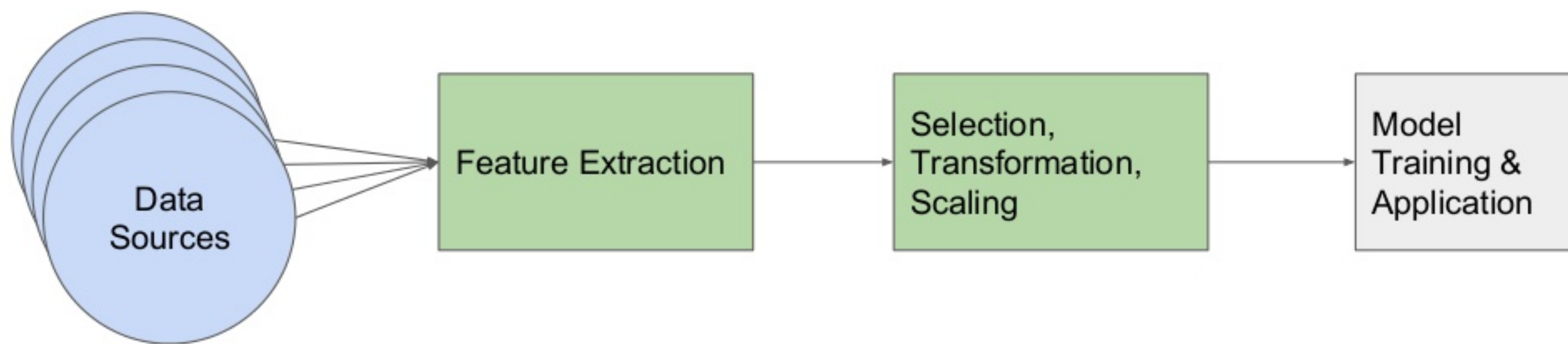


A simplifying recommendation

Start with unsupervised methods and iterate:

- Faster ramp-up
- Fewer parameters to tune
- Less overfitting risk

It begins and ends with data



Feature extraction is harder than it looks...

Which features to include?

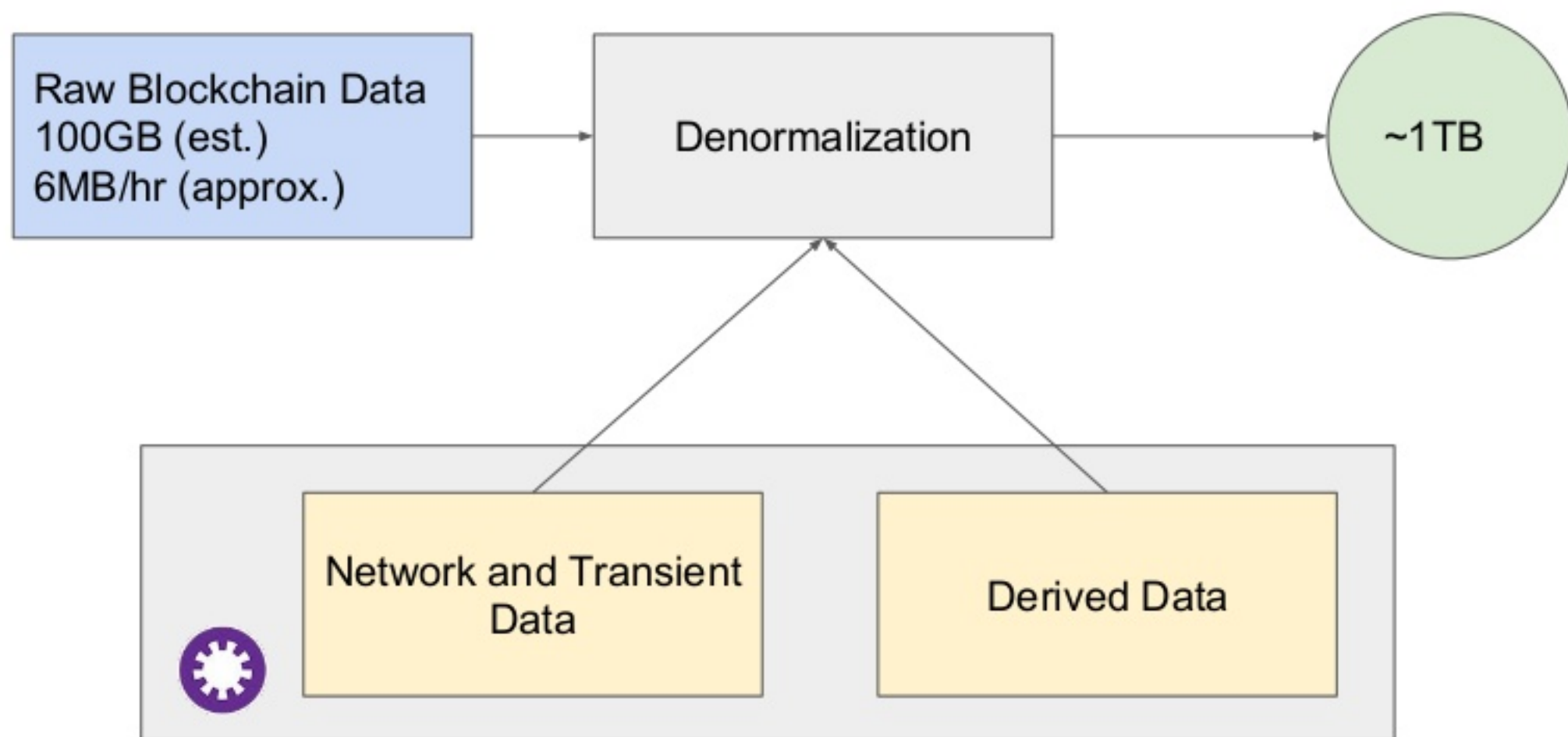
Balance risk of overfitting with need to capture relevant parameters.

... but it all comes down to $\text{Var}(x)$

When considering a variable, ask two questions:

- What variance does this variable introduce to the system?
- Is this variance correlated to the result I want?

Blockchain data is only one part of a larger picture

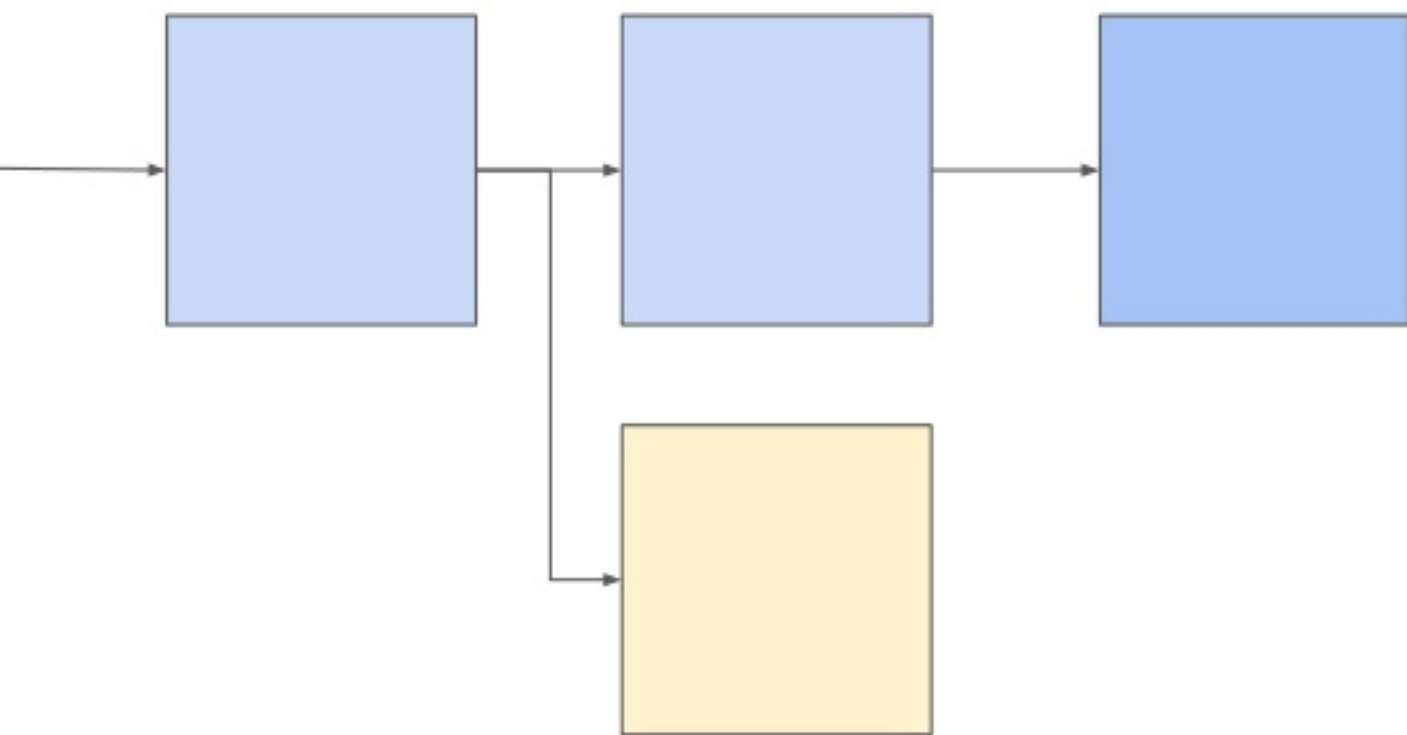


Big graphs need lots of tuning

- Edge/Vertex distinctions
 - Balance Locality, Parallelism
- Partition Strategies
- Data Denormalization
 - Eg. Reducing number, value of transactions between addresses to properties

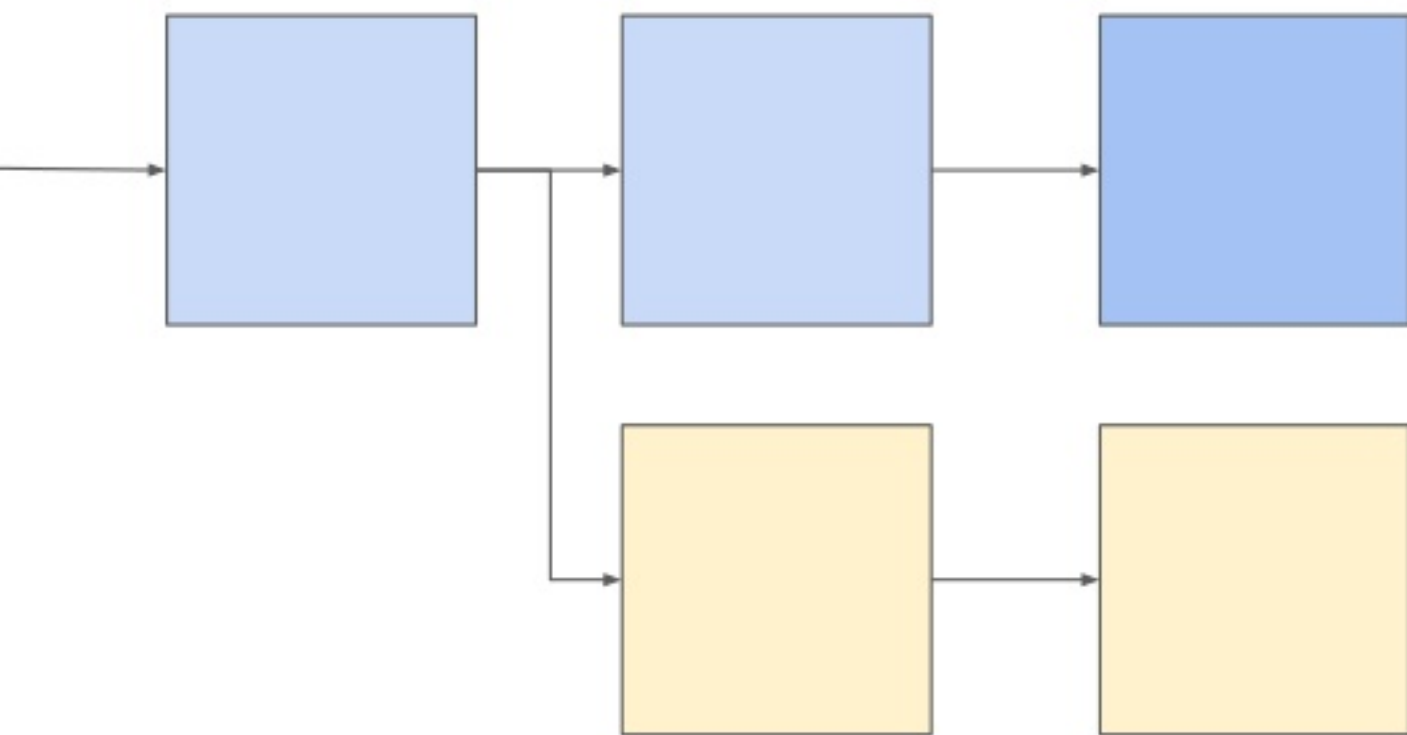
Blockchain data has some nuance...

Global consensus follows the chain with the most work:



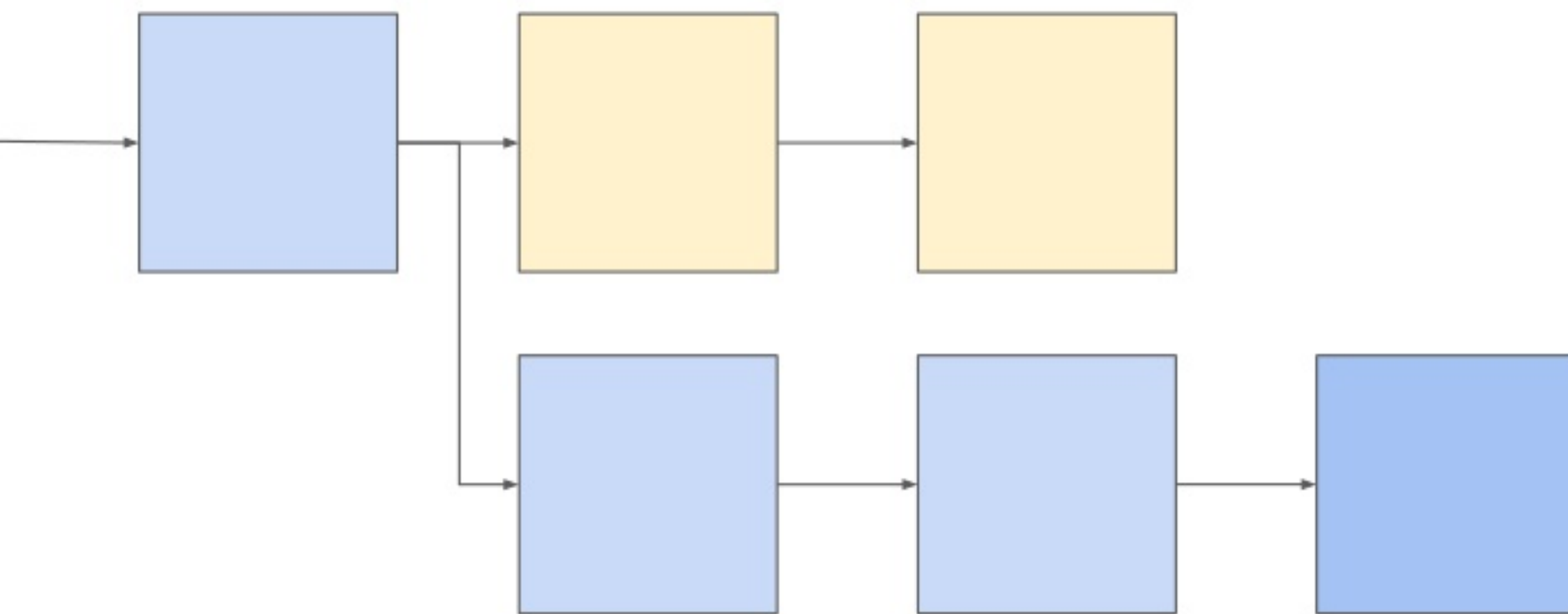
Blockchain data has some nuance...

Global consensus follows the chain with the most work:



Blockchain data has some nuance...

Global consensus follows the chain with the most work:

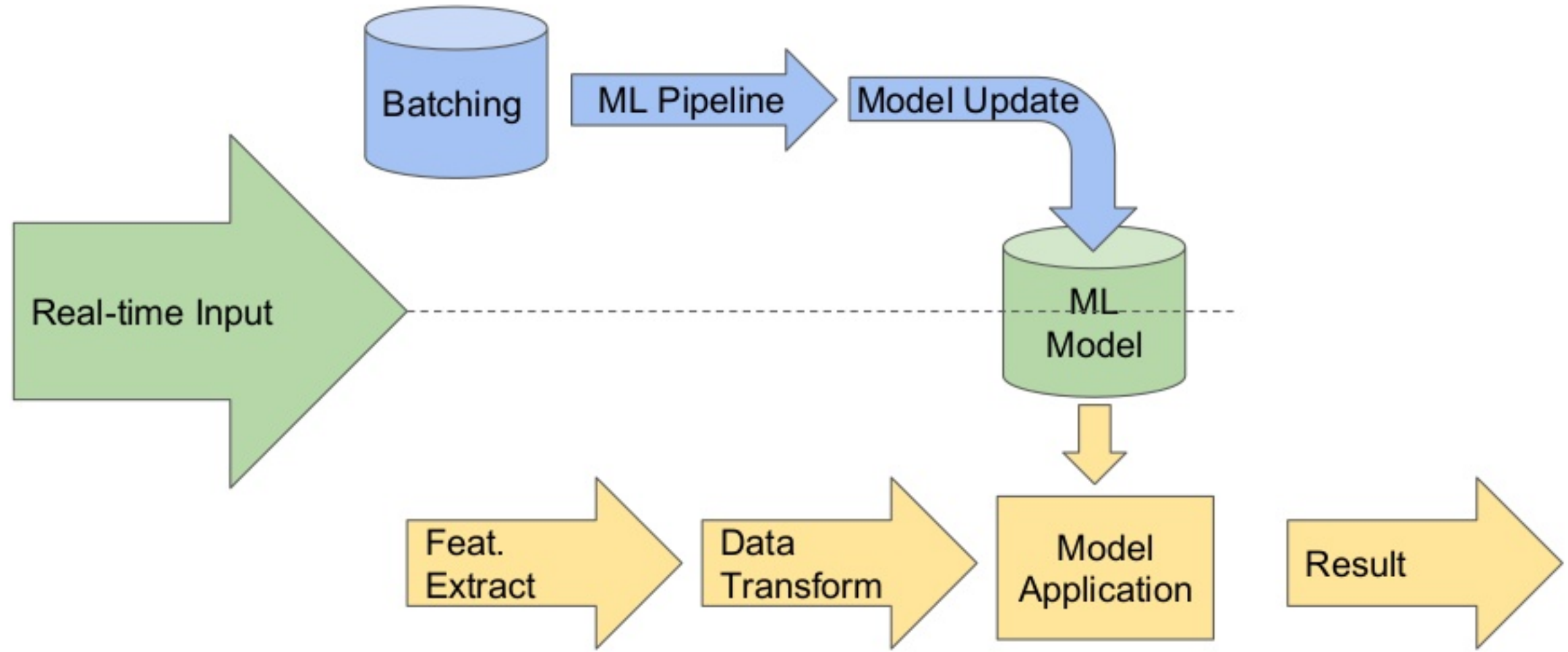


... but there are solutions.

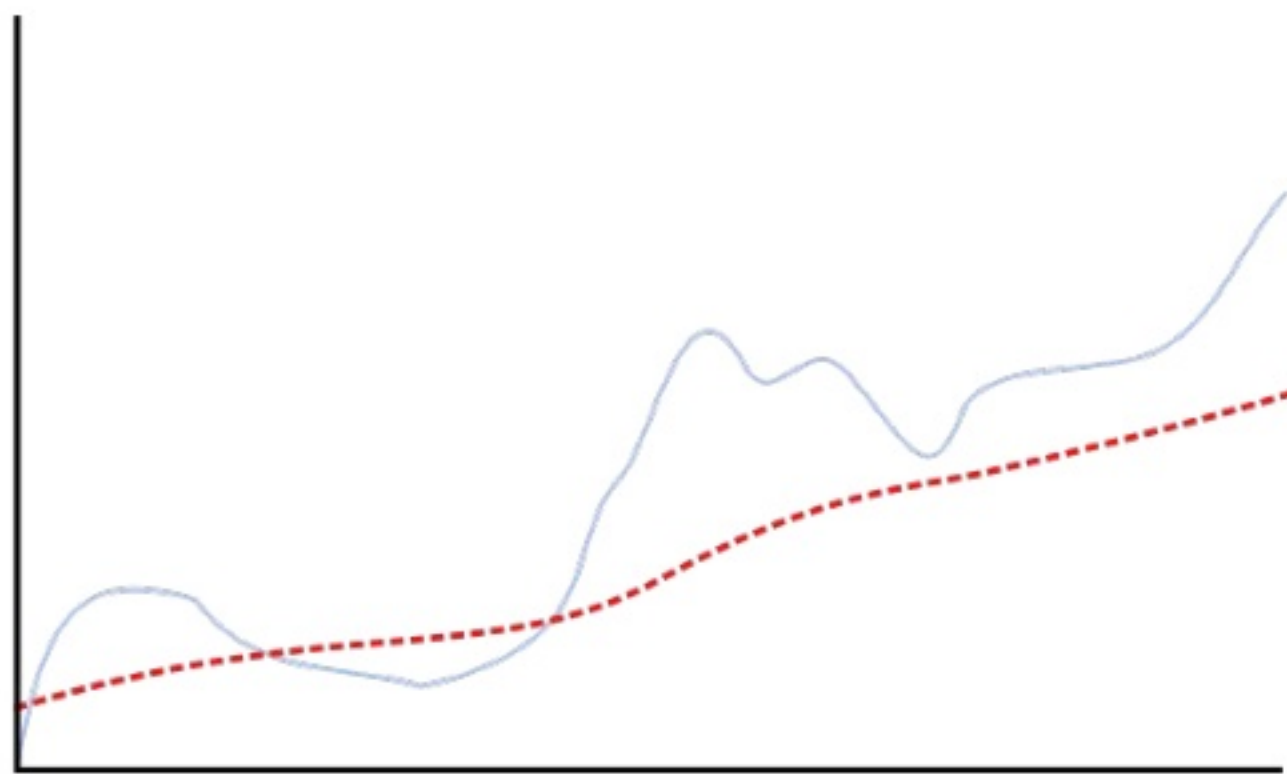
A processing delay can address virtually all rewrite cases, but hinders real-time analytics efforts.

The answer: Lambda Architecture

Lambda Architecture



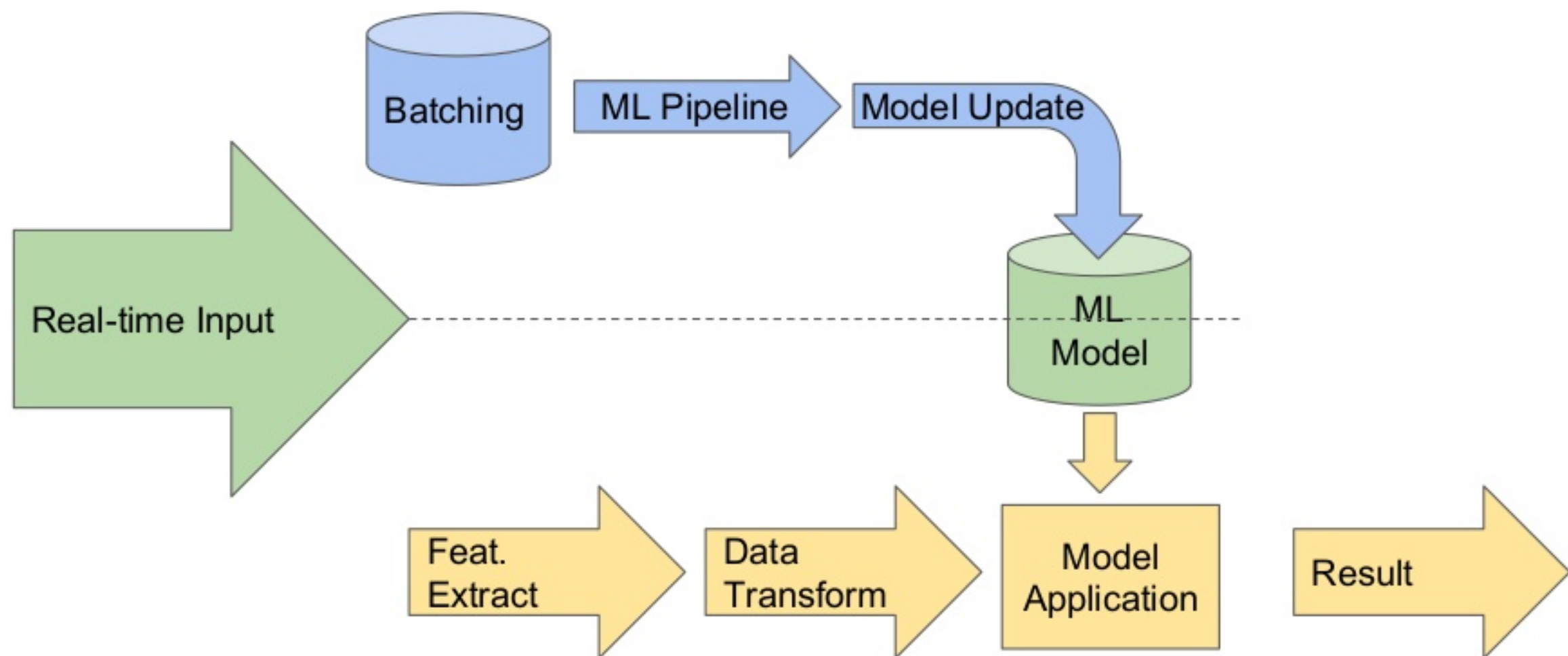
Why it works: Aggregates are slow to move...



... but what is true in the aggregate cannot predict individual measurements.

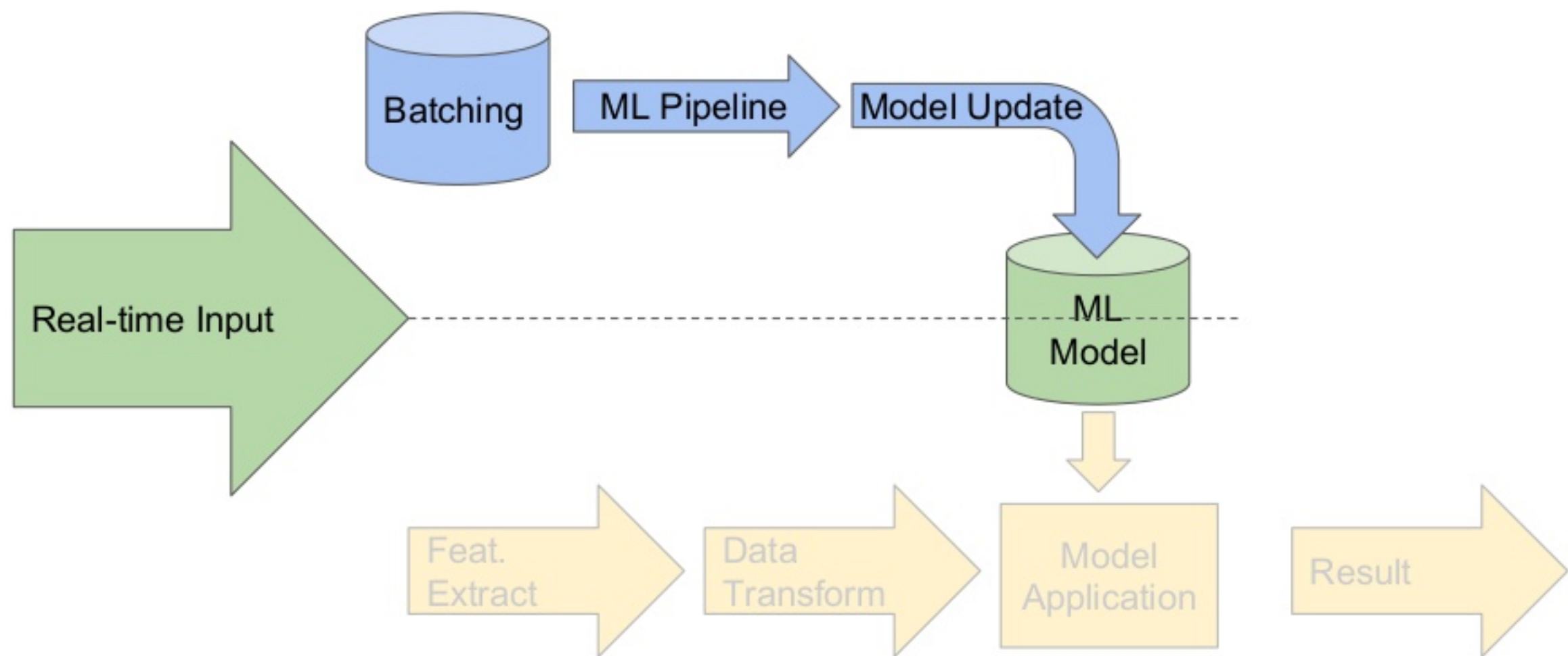
Example: Tx Clustering and Anomaly Detection

Lambda Architecture



05- Tx Clustering Example

Focus: Batching Process

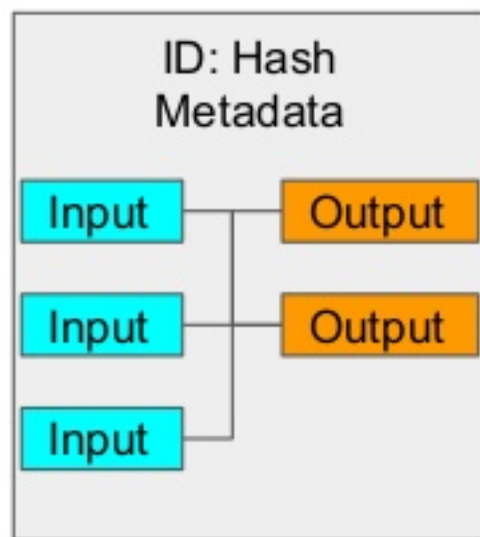


05- Tx Clustering Example

Delay in processing addresses mutability concerns



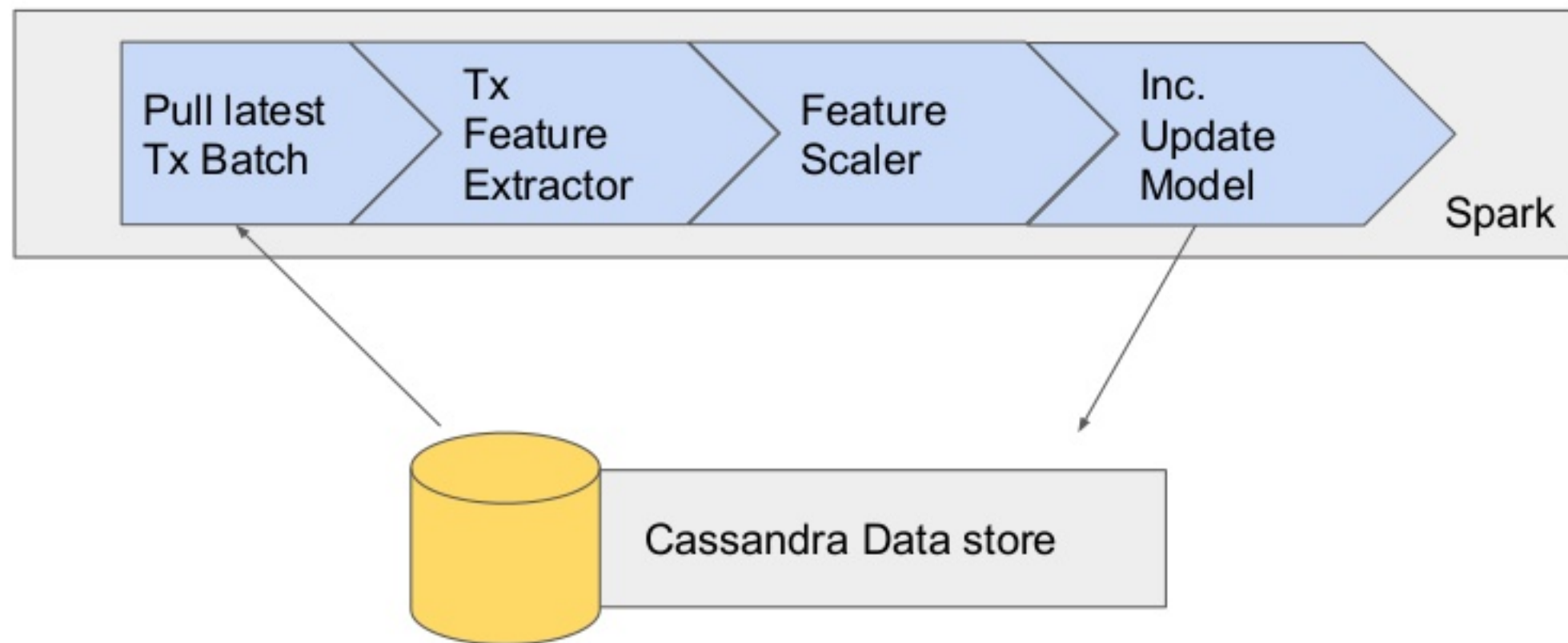
Feature extraction from live node data



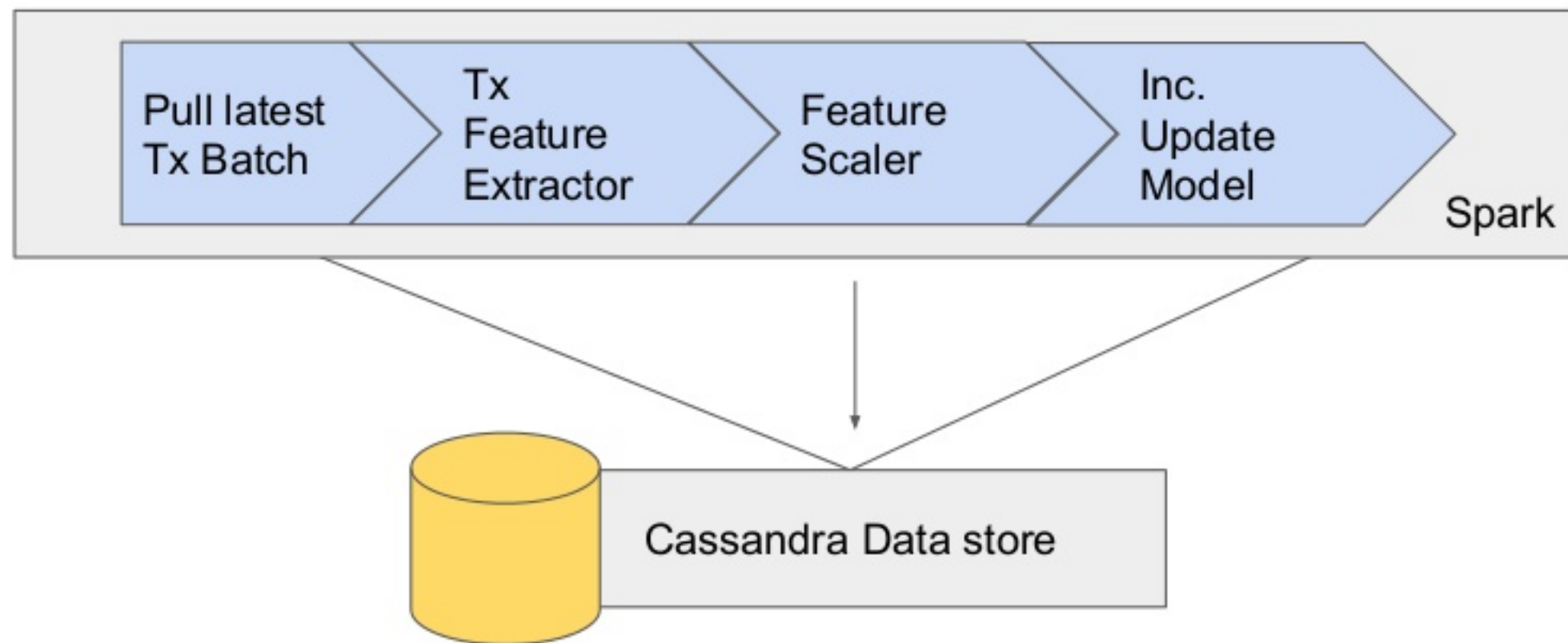
~20 features, including:

- Transaction shape (inputs, outputs)
- Value distribution
- Input types and ages

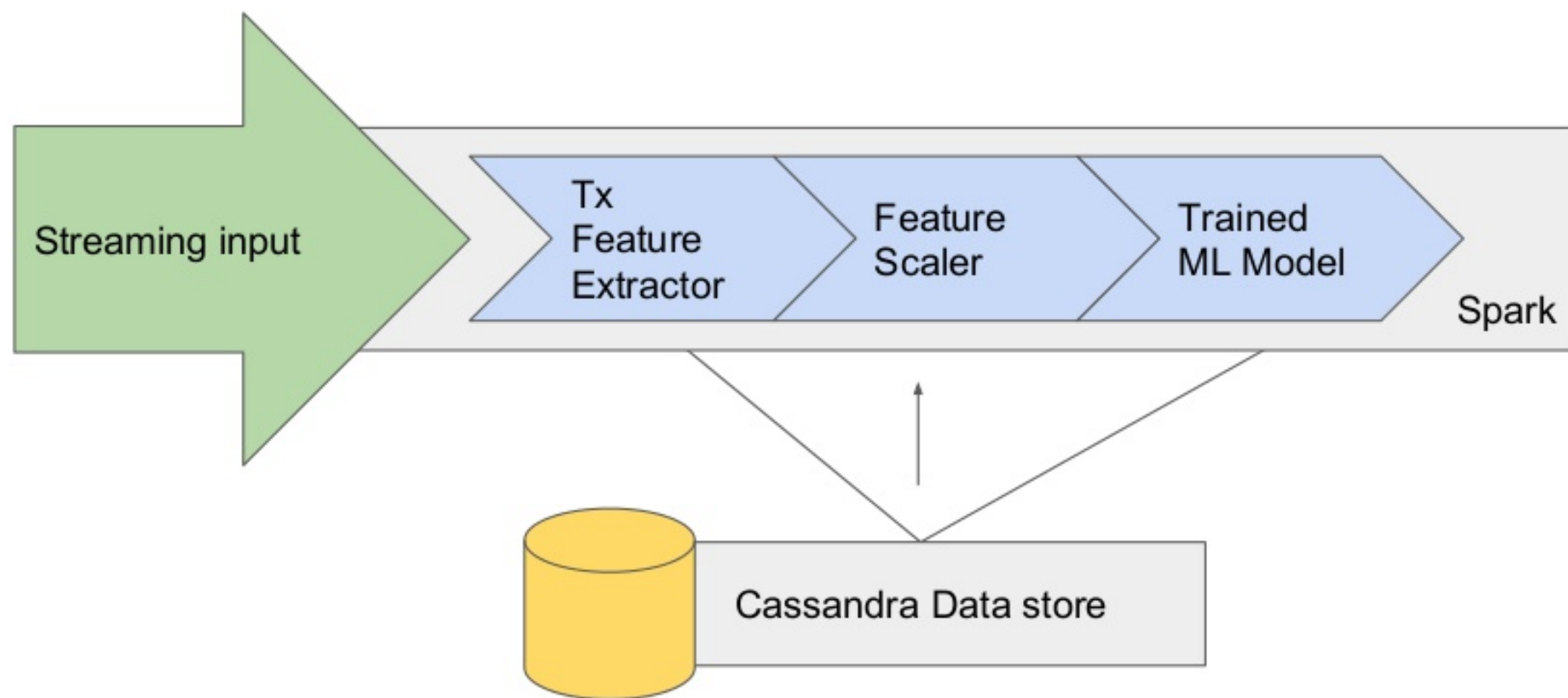
Data is fed into a kmeans pipeline



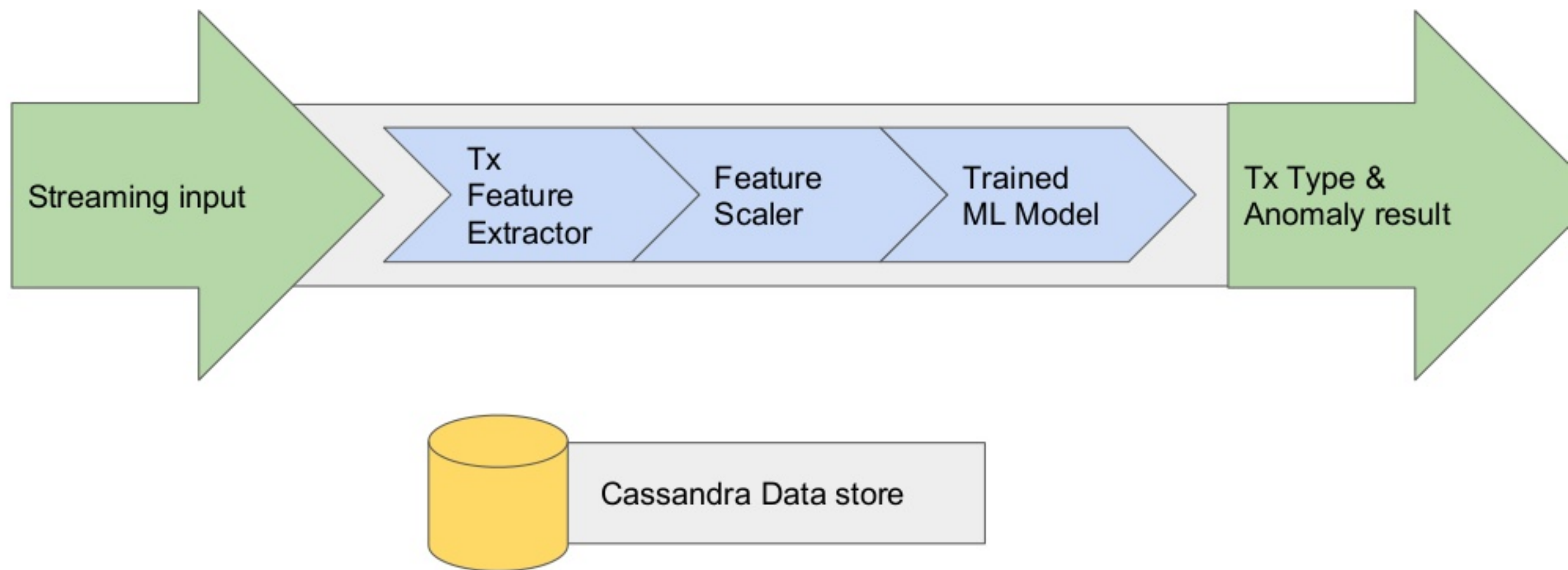
Spark 2.0 ML pipelines give you pipeline persistence



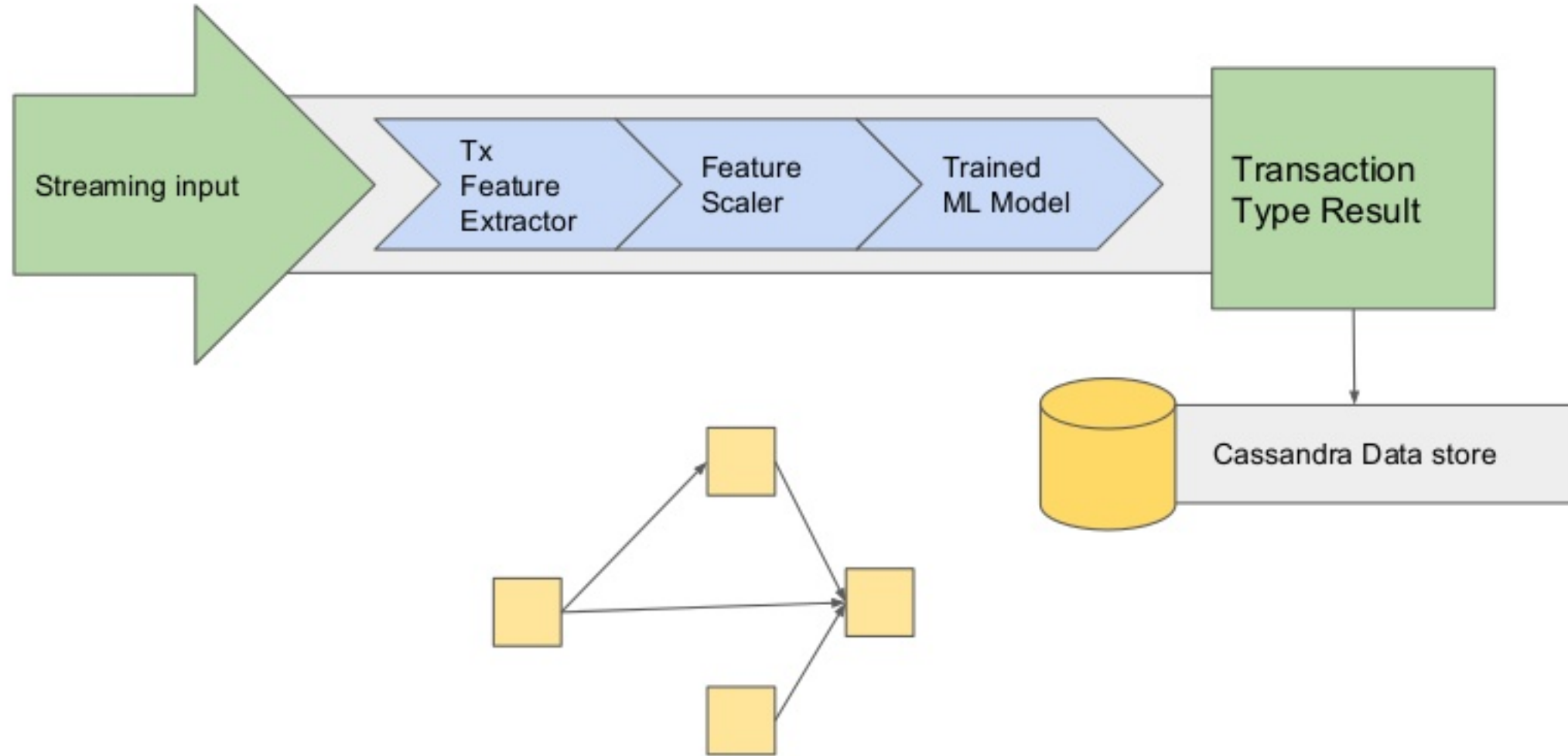
Streaming pipeline reuses data components



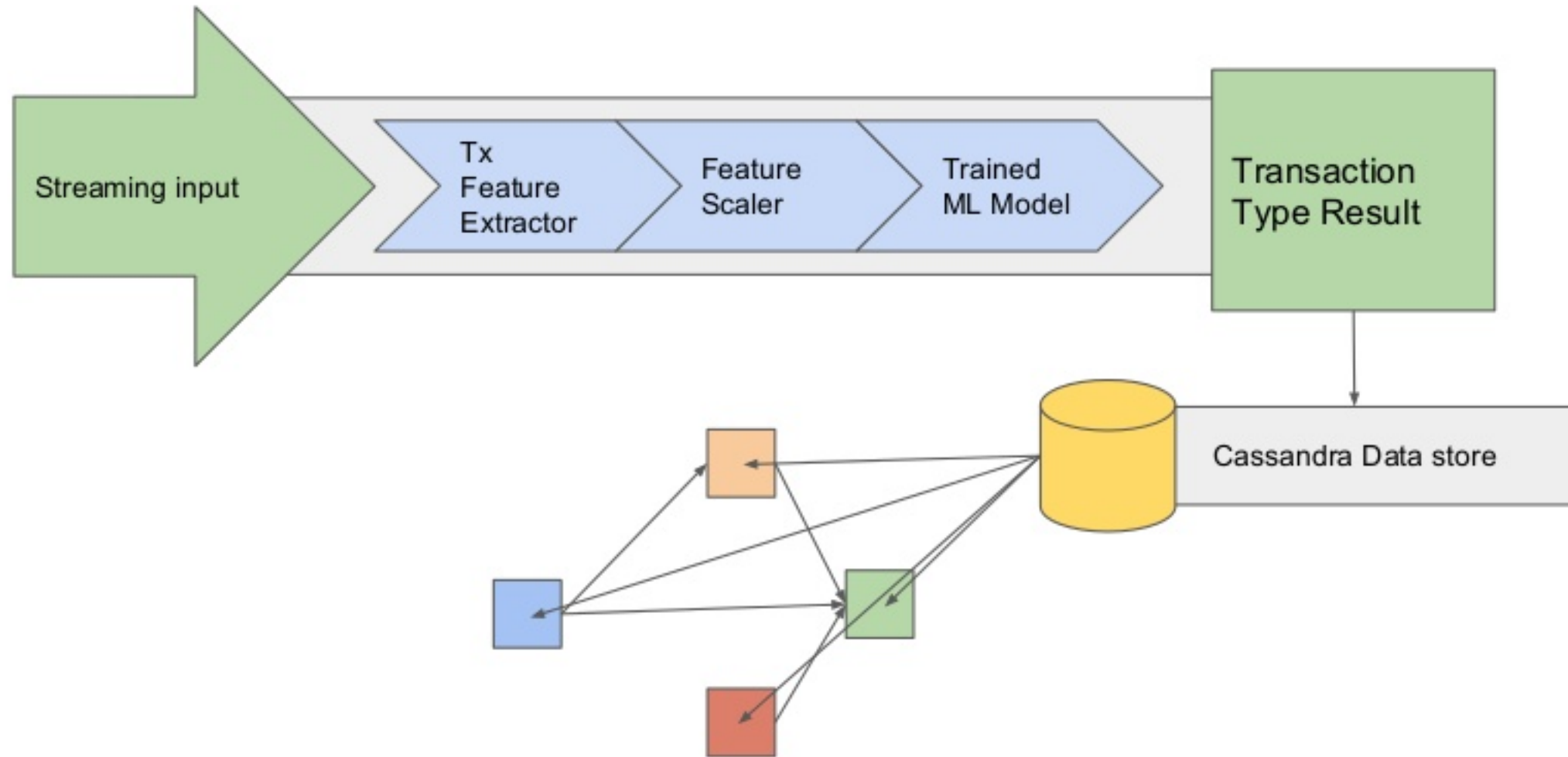
Trained model used for real-time anomaly detection



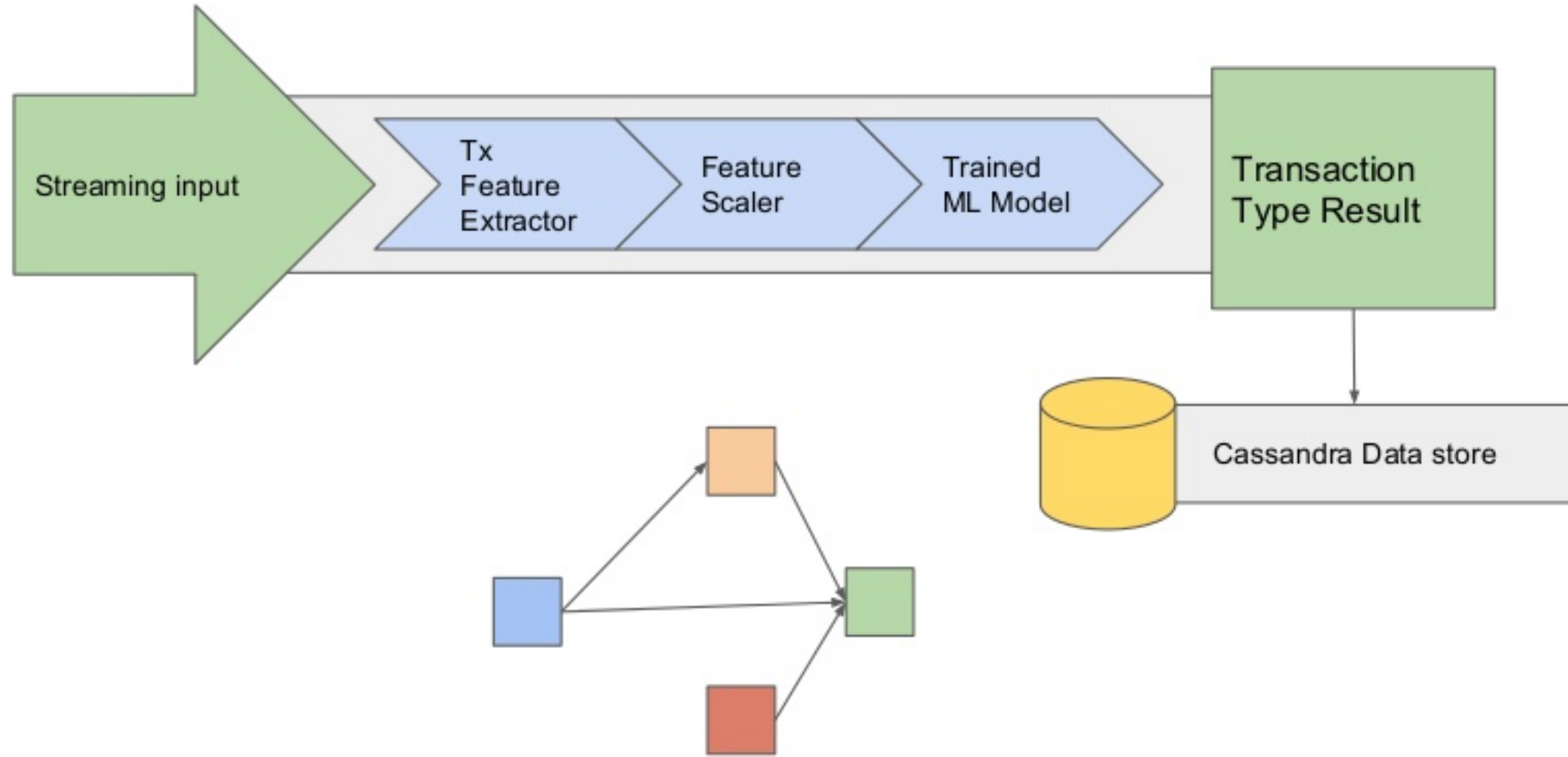
Also used to color graphs for efficient traversal decisionmaking



Also used to color graphs for efficient traversal decisionmaking



Also used to color graphs for efficient traversal decisionmaking

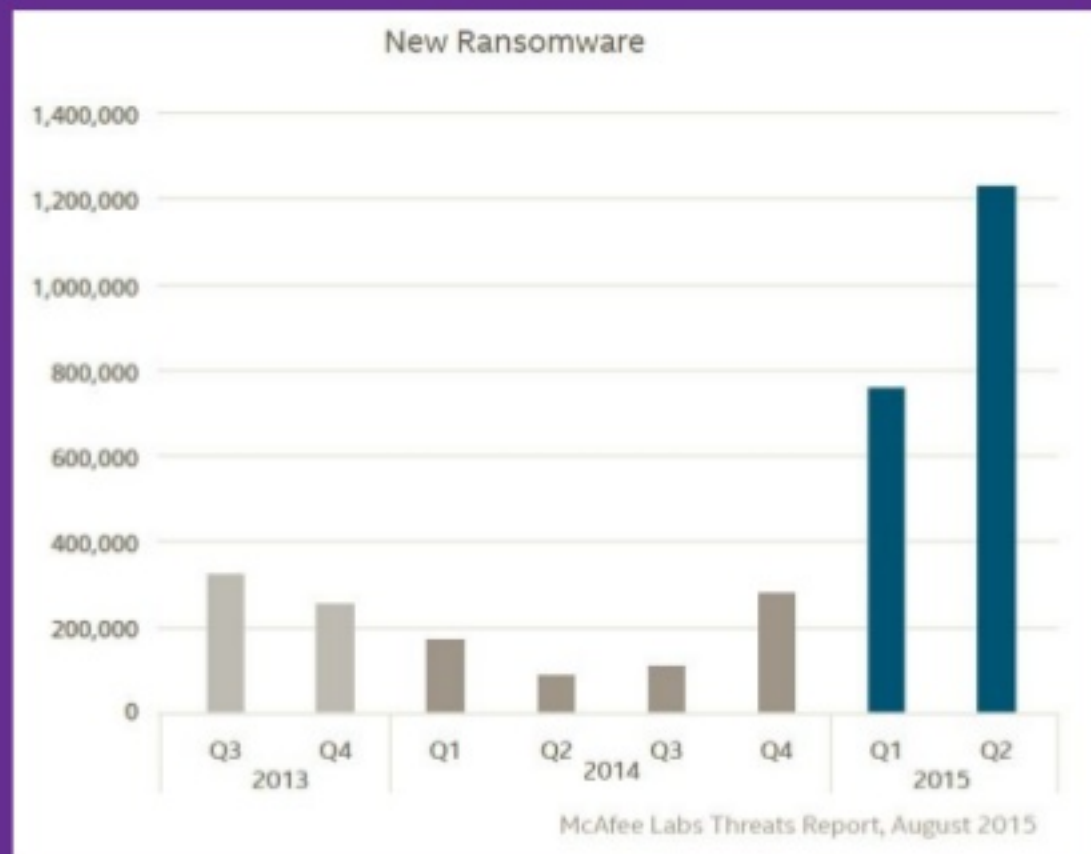


What to look for in Blockchain Analytics

- ❑ Infrastructure to ingest and transform metadata
- ❑ Multi-layered solution
- ❑ Graphs that address multiple purposes
- ❑ Resources for tuning graphs
- ❑ Real-time analytics (e.g. Lambda architecture)



Use Cases: Cybercrime



Source: FBI, SonicWall

Ransomware \$1B in 2016

Approaches

- Reactive - Follow the money
- Proactive - Anomaly detection



BLOCKCYPHER

Demo



BLOCKCYPHER

Use Cases

Use Case(s)	Industry
AML and financial crimes, regulatory compliance	Financial Services
Provider/patient demographic data analysis, Claims fraud	Healthcare
Mobile payments, Identity management	Telecommunications
Supply chain analytics	Manufacturing



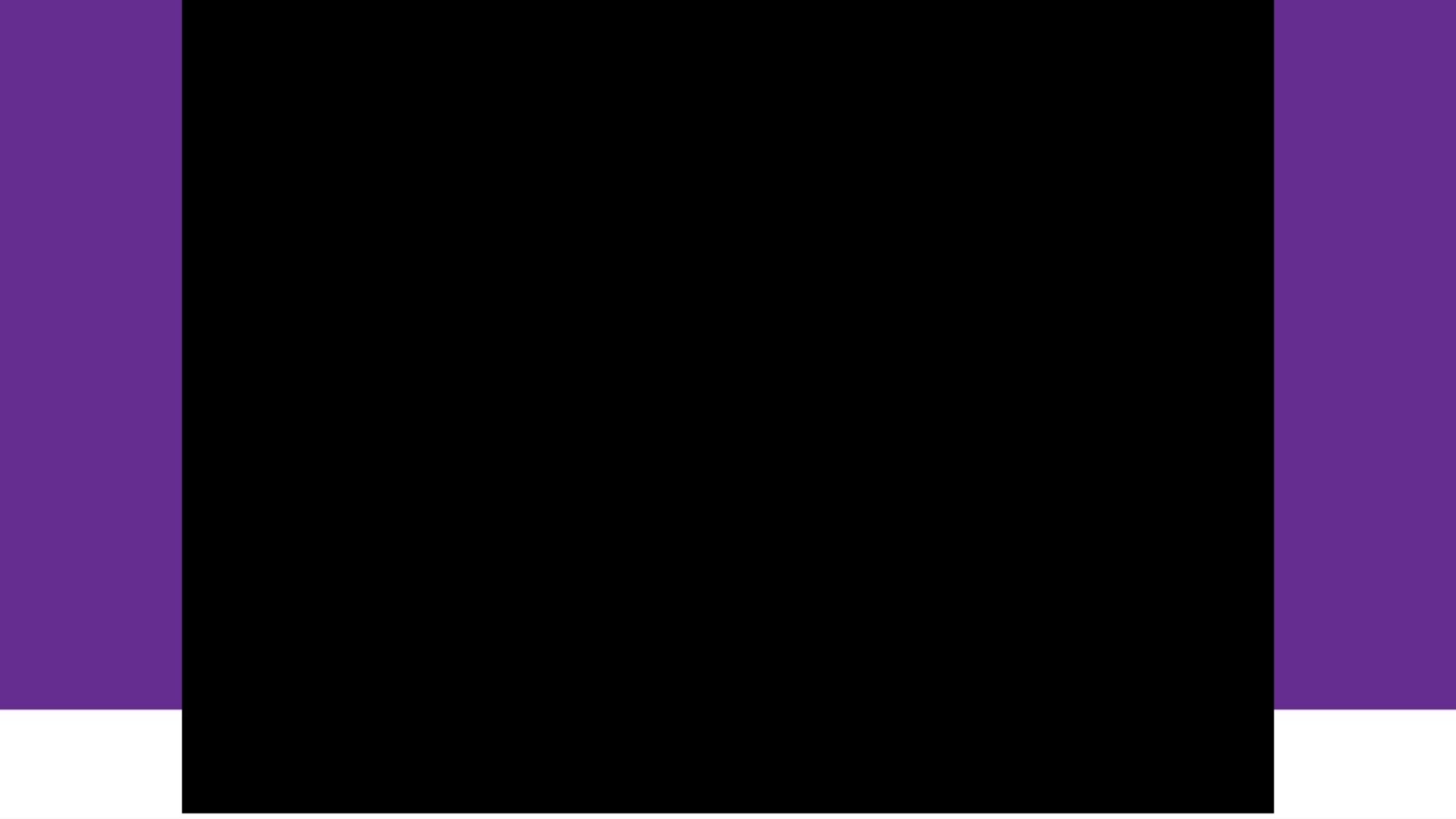
BLOCKCYPHER

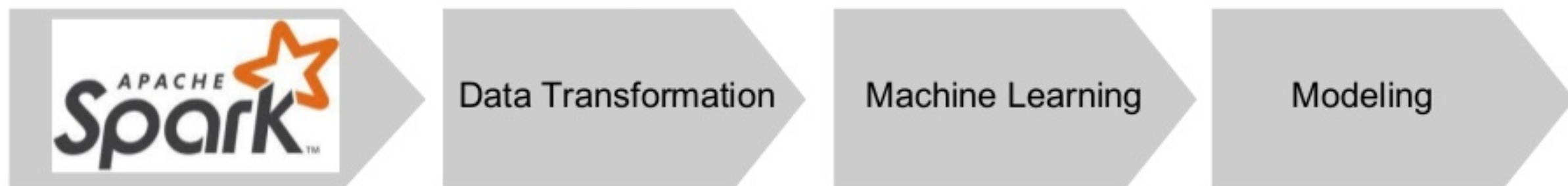
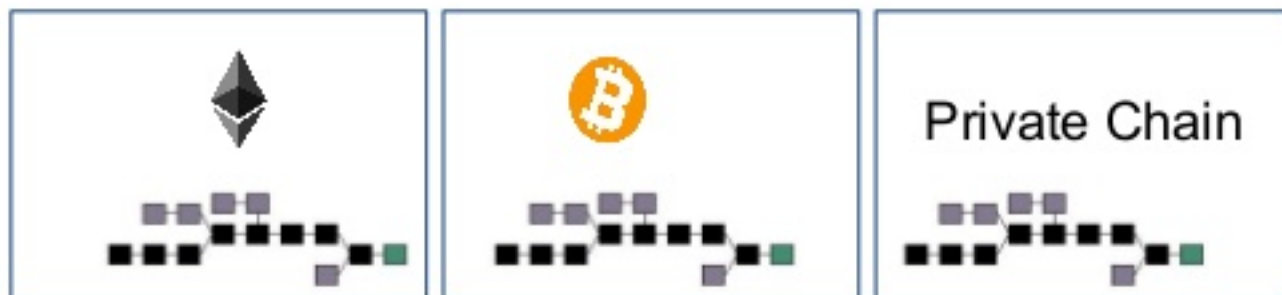
www.blockcypher.com

@blockcypher



BLOCKCYPHER





...