Graphs & Systems Thinking Workshop

Using Systems Thinking to Guide Enterprise Knowledge Graph Adoption

Dan McCreary

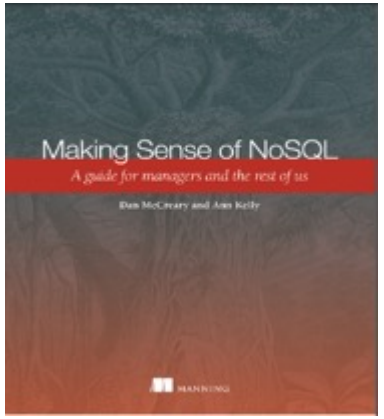
Welcome!


KGC
Dan McCreary
Distinguished Engineer
Optum
How Graphs Improve Systems Thinking
The Knowledge Graph Conference
May 3-6, 2021

Hello, my name is *Dan*

- Distinguished Engineer in AI and Graph Technologies at Optum's Advanced Technology Collaborative
- Co-founder of "NoSQL Now!" conference (now part of Dataversity)
- Author of "Making Sense of NoSQL" (w. Ann Kelly)
- Background in solution architecture, metadata management, NLP, semantics, text analytics and knowledge representation for AI
- **Personal Mission:** Use storytelling to unleash the power of connected data

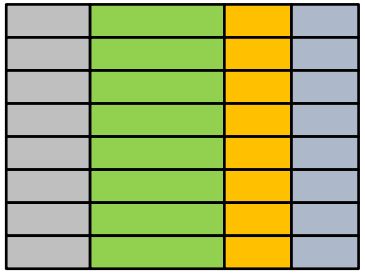
Disclaimer: All opinions are my own and may not reflect the views of my employer



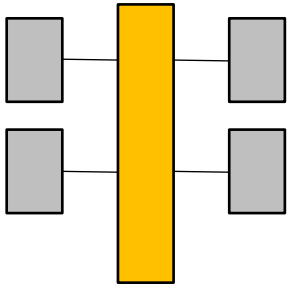
dan.mccreary@gmail.com

Graph is a “NoSQL” Data Architecture

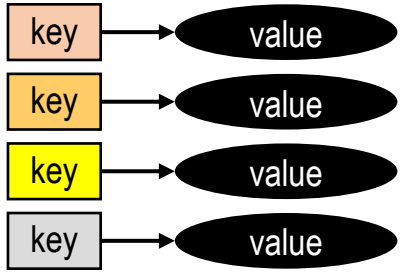
Relational



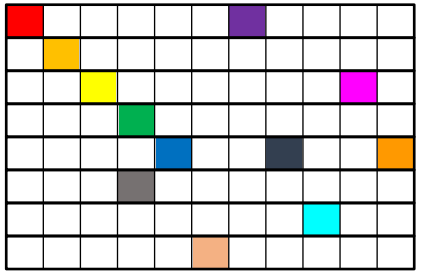
Analytical (OLAP)



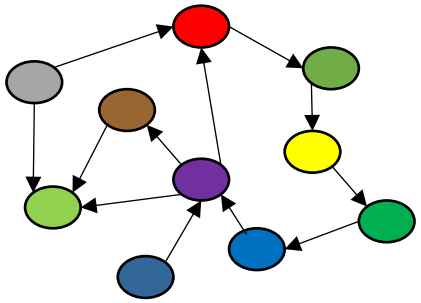
Key-Value



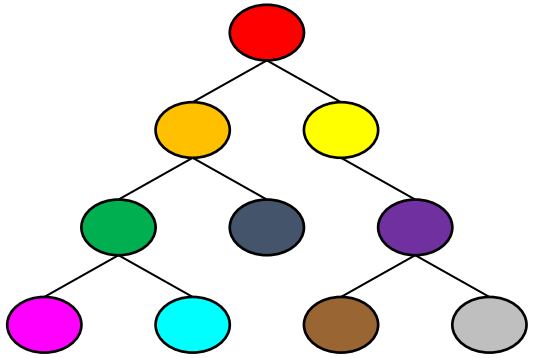
Column-Family



Graph



Document



See Chapter 1: https://www.manning.com/books/making-sense-of-nosql

Graph Timeline

Euler Solves 7 Bridges of Königsberg 1736

Sir Tim's Vision RDF 2001

On Intelligence 2005

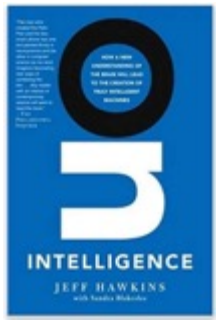
W3C SPARQL 2008

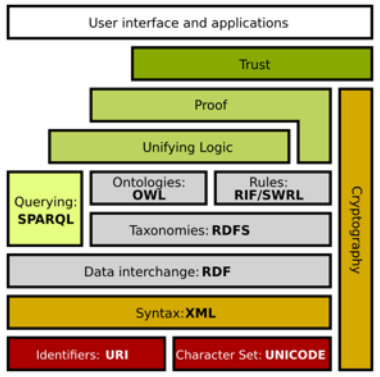
Labeled Property Graph Neo4j 1.0 2010

Google's Knowledge Graph "Things Not Strings" May 2012

AlexNet Sept 2012

LPG in AI 2018

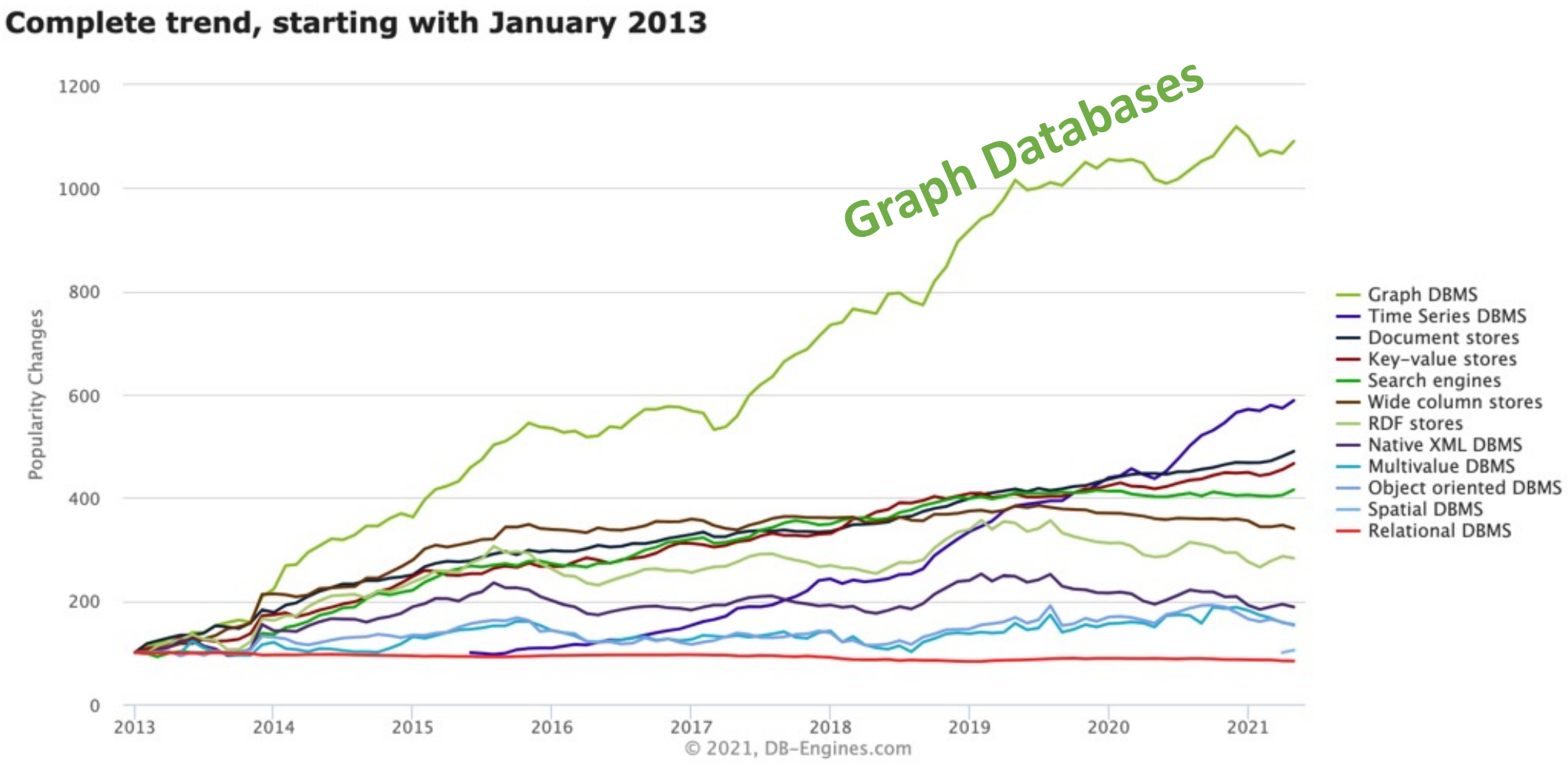




Semantic Web Stack

Graphs Rise

Graph Databases are Hot!



Workshop Description

Graph databases are characterized by their ability to model relationships between entities. Graph models allow for relationship traversals up to 3-5 orders of magnitude faster than traditional relational models.

This performance advantage allows graphs to include more data and different types of data than were possible in the past. This workshop will introduce the fundamental concepts around **Enterprise Knowledge Graphs** (EKGs) and **Systems Thinking** and how they fit together. We then will walk the group through a series of exercises to demonstrate how the two concepts are related with examples.

EKG Architects Questions

- What other data should we include in our graph?
- What benefits could we gain if we began to think of our organization as a more holistic and integrated system?
- How do the diverse datasets interact to give us deeper insights into how to optimize our operations?

Objectives for the Workshop

- Define the characteristics of an Enterprise Knowledge Graph (EKG)
- Allow participants to understand how EKG data modeling processes determine what is stored in an enterprise knowledge graph
- Learn the fundamentals of Systems Thinking and how large graph models help us with Systems Thinking
- See the role of time in data models (temporal modeling)
- How to align the EKG data model with enterprise strategy (lower costs, increase revenue, increase agility)
- Learn how to predict the value of insights as you connect more systems together

Outline for the Workshop

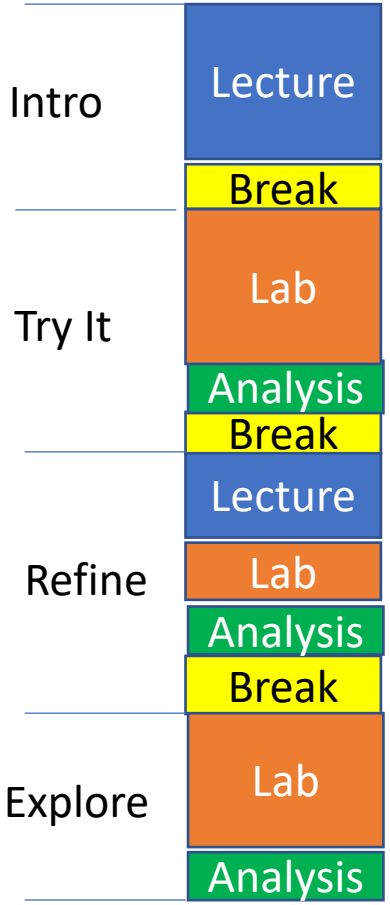
- What is an Enterprise Knowledge Graph?
- What is Systems Thinking?
- What are predictive graph models?
- What are feedback loops?
- What are externalities?
- How do we look for unintended consequences?
- How do we decide what new data to model?
- How do large, diverse connected datasets change the way we think?
- What new insights can we discover?
- How do we decide what datasets can provide the most value to our system?

Workshop Philosophy

- 1/3 lecture – light introduction
- 1/3 trying out systems thinking on your challenges
- 1/3 shared analysis of what worked/what did not work

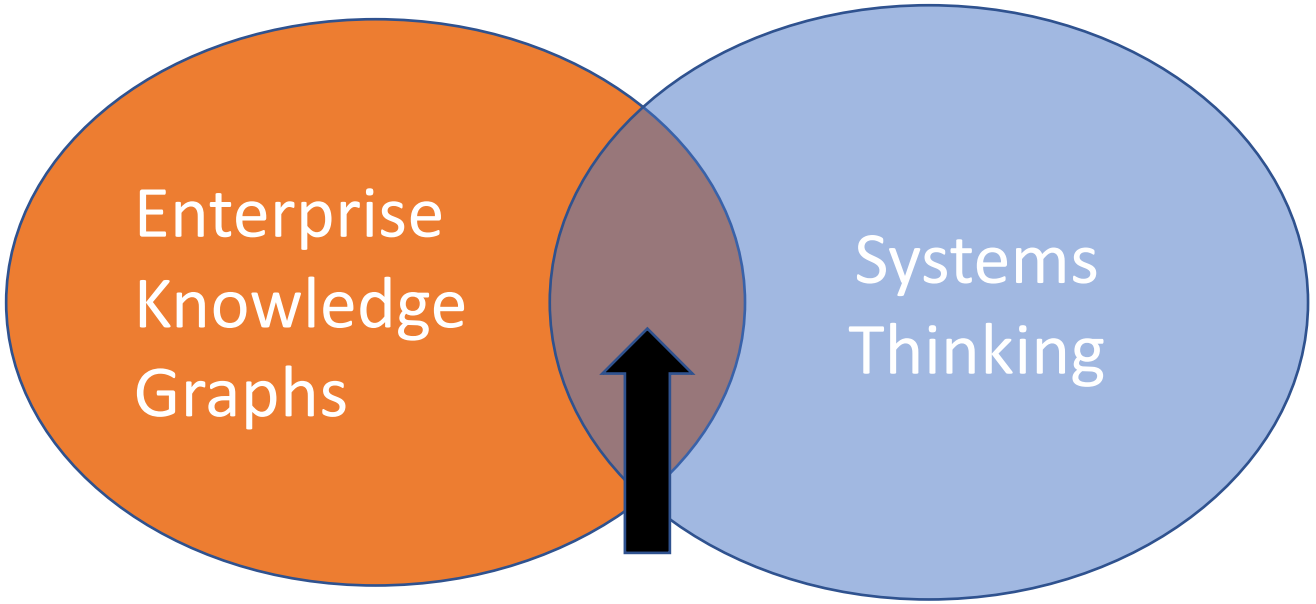
There are no right answers, only trade offs

Class Structure



- 4 hours wall-clock time – 4 50-minute sessions with 10-minute breaks between sessions
- Part 1: **Introduction**
  - 50 minutes overview of graph systems thinking
  - EKG assumptions, causal loop diagrams
- Part 2: **Try It**
  - 35 minutes, 15 minutes - analysis
  - Draw causal loop diagrams for things that might impact your EKG evolution
- Part 3: Refine Connected Emergence
  - How does connecting new data sources trigger insight?
  - Exercise: Predict new insights
- Part 4: Working Session
  - Break up into groups
  - 35 minutes: Focus on key problems that you are interested in
  - Final 15 minutes: share results

Agenda



How to use systems thinking to guide the adoption of Enterprise Knowledge Graphs (EKGs)

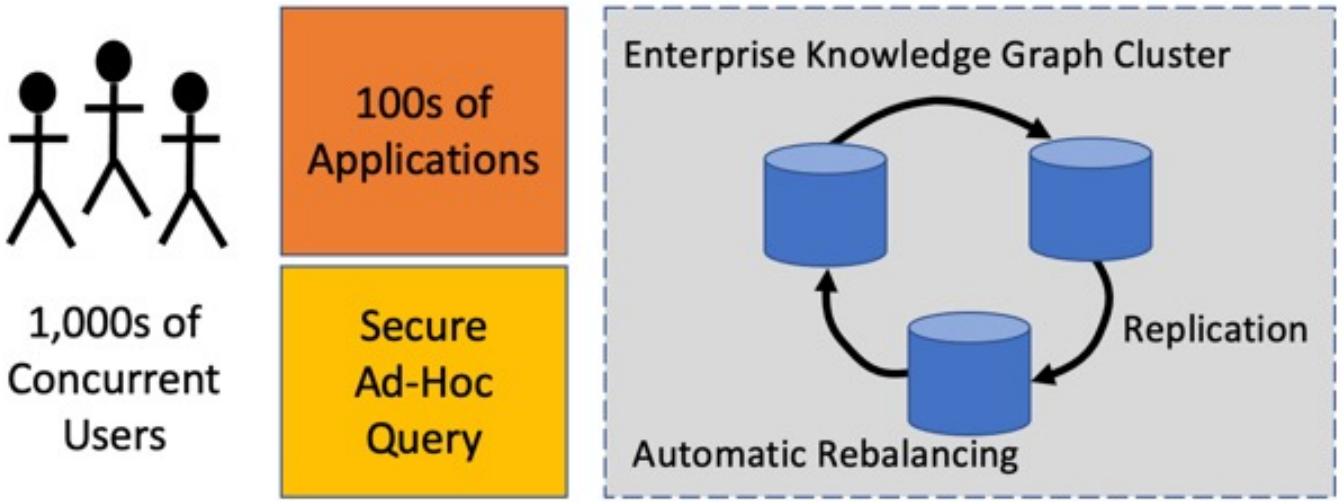
Terminology

https://dmccreary.github.io/graph-systems-thinking/glossary/

Assumptions

- Enterprise Knowledge Graphs (EKGs) are going to become a central force in organizational dynamics
- They are becoming the Central Nervous System (CNS) of organizations
- We need tools to manage the adoption and growth of EKGs
- *Systems Thinking* is an appropriate tool to help us guide EKG Growth

Definition of an Enterprise Knowledge Graph



- High Availability (HA): over 99.99% uptime
- Rolling upgrades: never shut down services for upgrades
- Vertex-level Role-based Access Control (RBAC)
- Resource quotas
- Large library of graph algorithms

An Enterprise Knowledge Graph (EKG) is a type of graph database designed to **scale-out** to meet large organizations' demanding requirements to store diverse forms of connected knowledge.

Seven Measure of Scale-out Graphs

1. **Scale-out data size** — adding more RAM, SSD, and spinning disk should not interrupt services
2. **Scale-out compute** — adding additional CPUs should be possible without service interruption
3. **Scale-out security** — adding more projects with more roles and more users should not impact system performance
4. **Scale-out manageability** — monitoring the continual performance of 100s of applications executing thousands of graph queries is a complex process.
5. **Scale-out data quality** —EKG software must make it easy to perform data validation as it enters the EKG and as it evolves within the EKG as new relationships are inferred
6. **Scale-out algorithms** — EKGs need to run an extensive library of standard graph algorithms and a new generation of machine-learning algorithms that create graph embedding
7. **Scale-out query** — EKGs need query software that allows developers to express distributed queries in high-level query languages

https://dmccreary.medium.com/a-definition-of-enterprise-in-ekgs-561283d37deb

EKGs: Today vs Future

Today

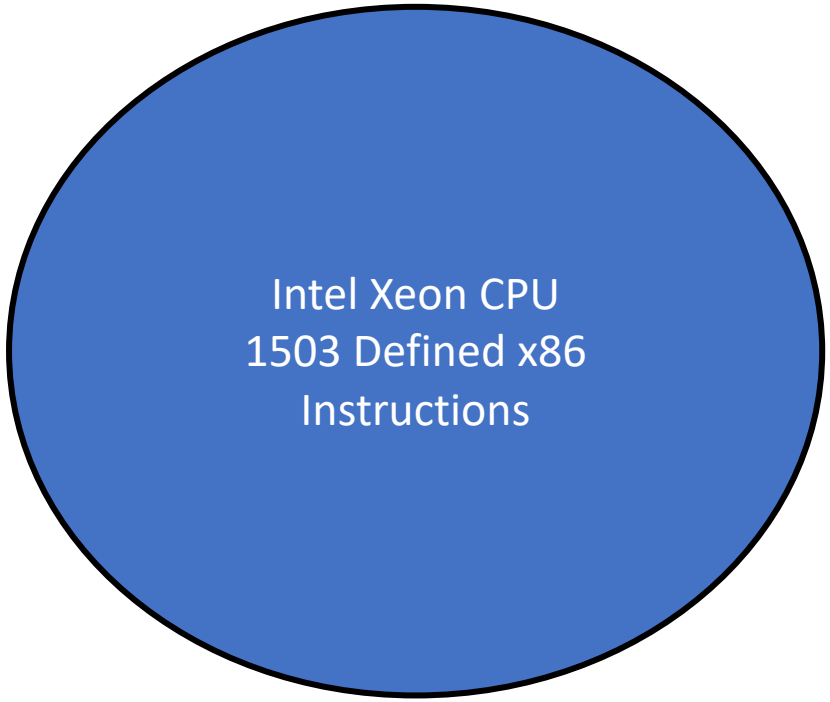
- Many graph database don't scale well over 100s of servers
- Distributed graph database licenses are prohibitively expensive ($1M/TB)
- Only the largest companies can afford them
- No specialized graph chips
- No built-in machine learning

Future

- Most software vendors will have scale-out graph solutions
- Open source distributed graphs will be common
- Even small-medium companies will have robust EKGs
- Specialized graph hardware
- Out-of-the box machine learning

General CPU Hardware vs. Graph Hardware

Complex Instruction Set Computer (CISC)

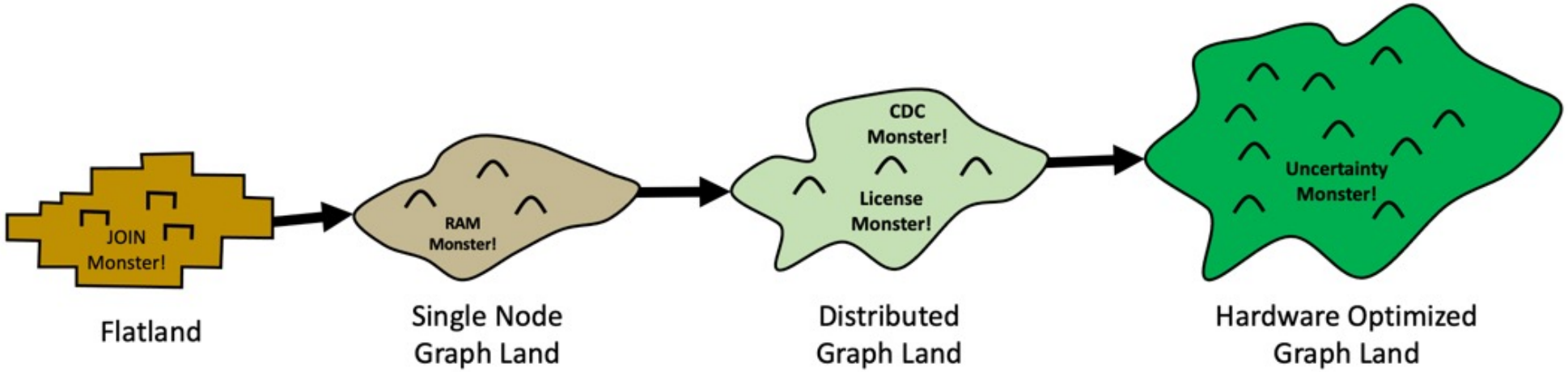


Reduced Instruction Set Computer (RISC)

Fewer Instructions -> More Cores

- Most graph traversal algorithms only need simple pointer hopping
- How efficient are CPU and GPUs at running graph algorithms?
  - No need for floating point
  - No need for matrix multiplication

Four Stages of EKG Adoption



1. Flatland
2. Single node graph
3. Distributed graph
4. HOG Heaven

From Flatland to HOG Heaven
https://towardsdatascience.com/from-flatland-to-hog-heaven-the-four-lands-of-ekg-adoption-945571c09b67

HOG Heaven

Photo by Bruno van der Kraan on Unsplash

- Hardware optimized graph solutions
- 100B vertices
- Vertex-level role-based access control
- Complete views of customers in under 100 milliseconds
- Embeddings for every vertex to enable fast similarity at scale

Key Question

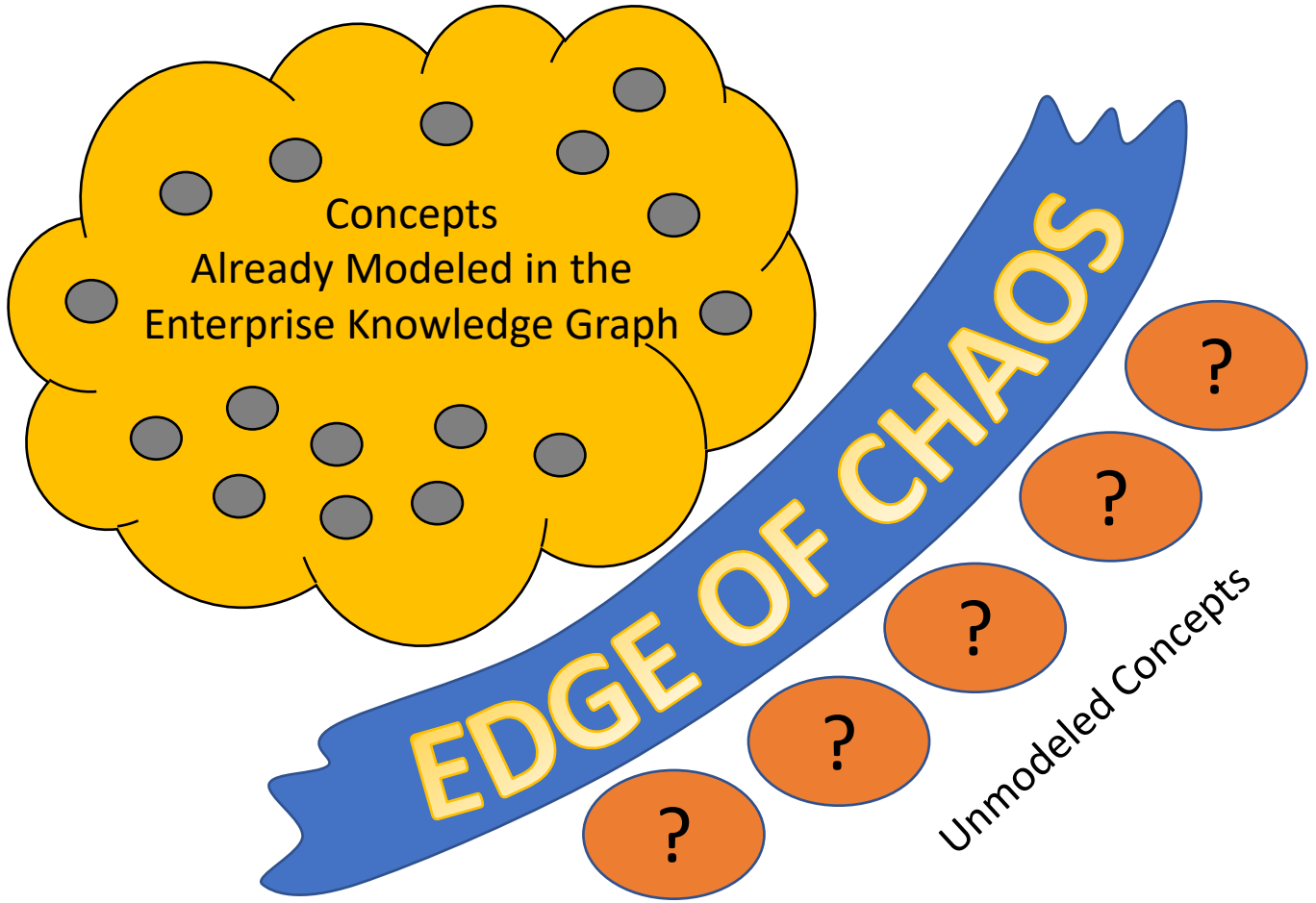
*How do we choose **what** entities in our organization should be in our EKG?*

Answer: Use Systems Thinking!

Exercise

- How do you currently decide what items to put in your EKG?

Edge of Chaos



Consider two regions of your data model:

1) The part of the world that you have modeled with precision. We call this region the EKG region

2) The part of the world that you have not modeled yet. We call this the “region of chaos”

The border between the EKG and the region of chaos we will call “The Edge of Chaos”

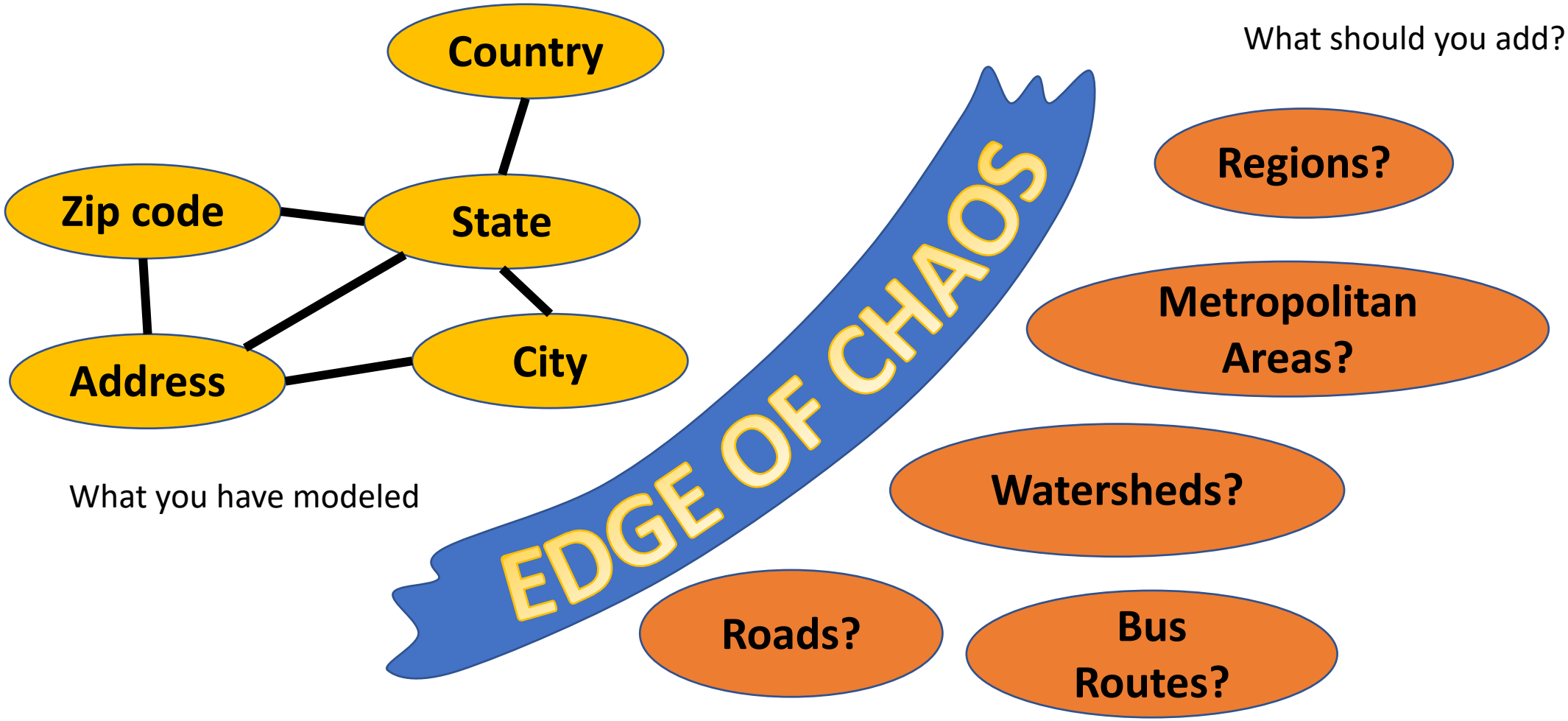
What is a System?

- A collection of components that interact together to produce some sort of behavior of the whole
- Systems can have subsystems

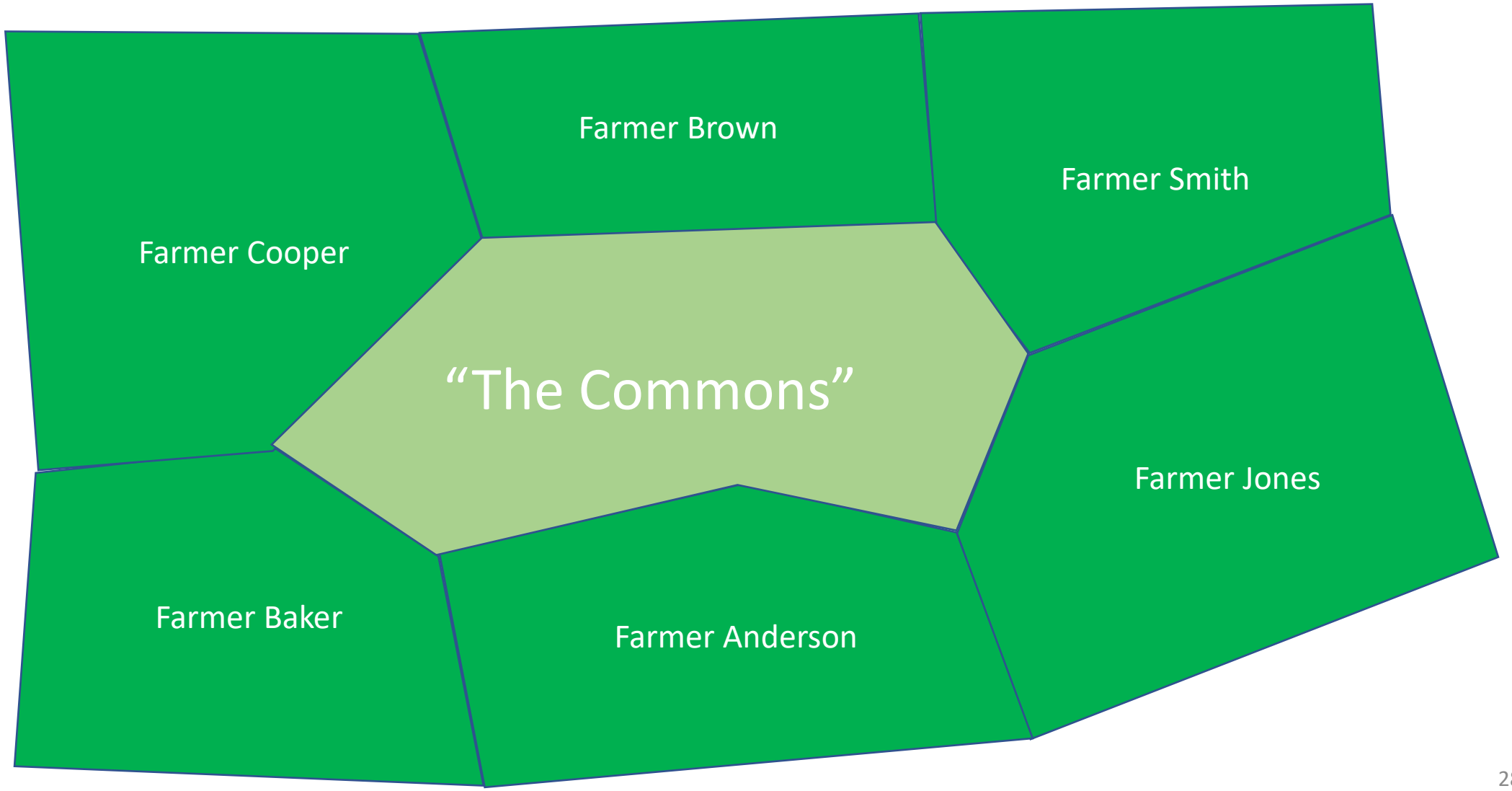
Example: Provider Recommendation

*When our senior members call into our call centers to find a healthcare provider in their area, many of them **don't drive**. They only want providers that are on **bus routes**. However, our current system does not store this information.*

Example: Geospatial Models



Tragedy of The Commons



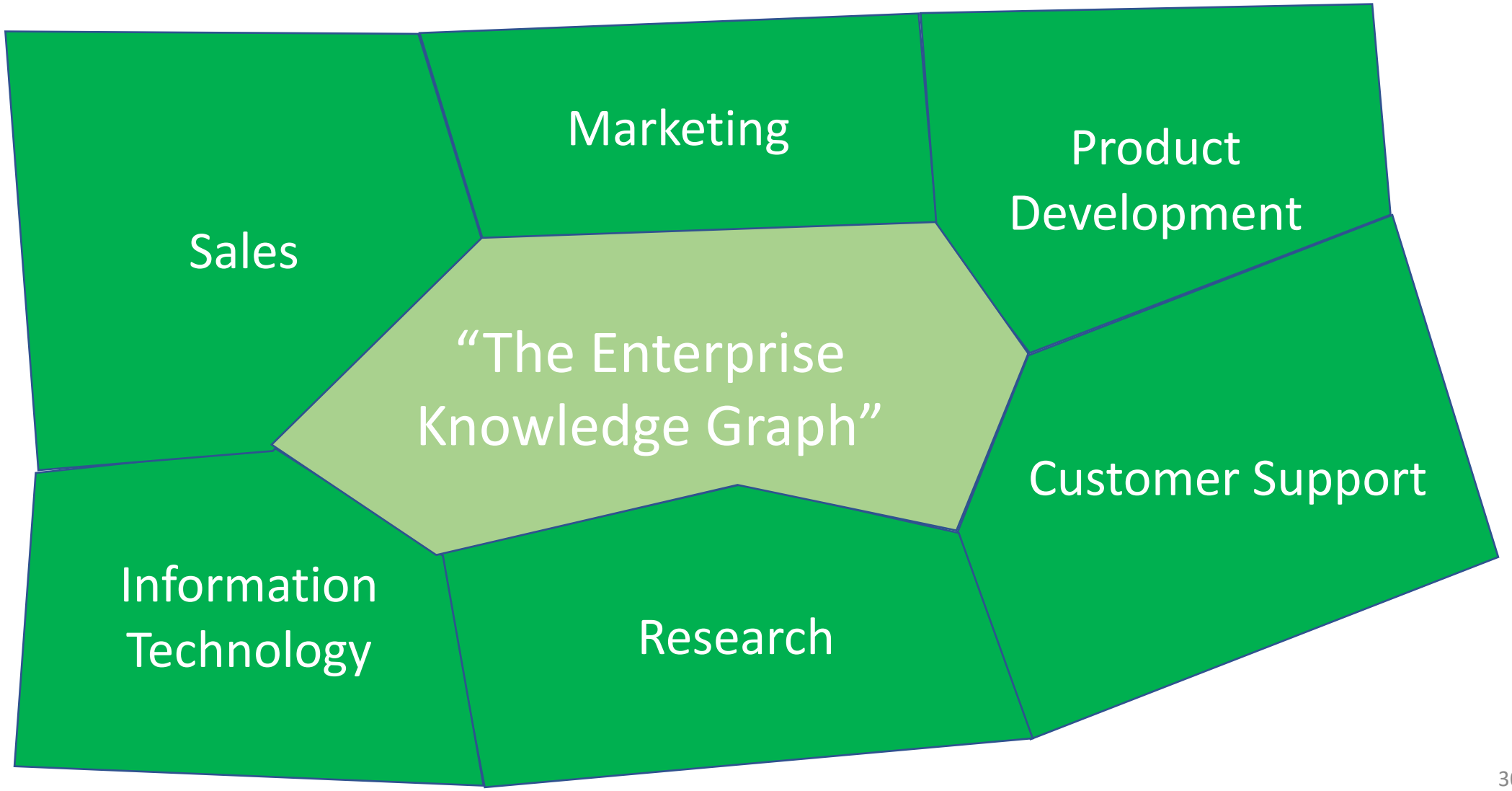
Tragedy of The Commons

A situation in **economic science** when individual users, who have **open access to a resource** unhampered by shared social structures or **formal rules** that govern access and use, act independently according to their own **self-interest** and, contrary to the common good of all users, cause depletion of the resource through their **uncoordinated** action.

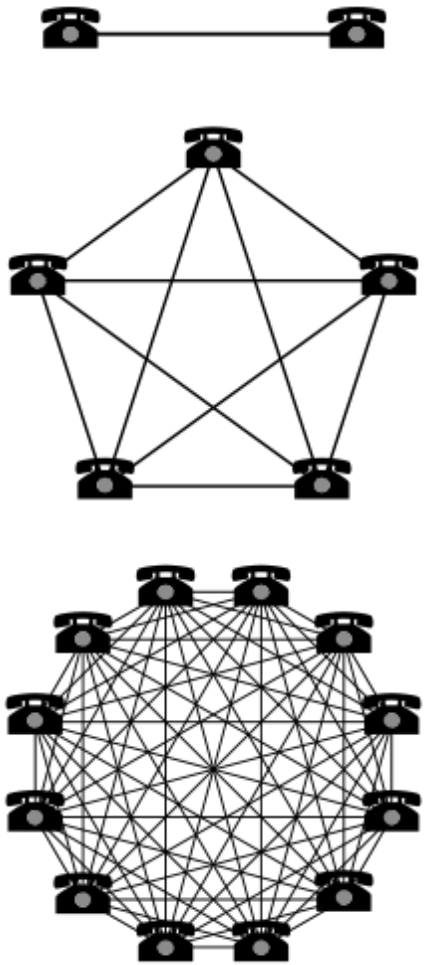
The more everyone uses a resource, the less valuable it becomes.

https://en.wikipedia.org/wiki/Tragedy_of_the_commons

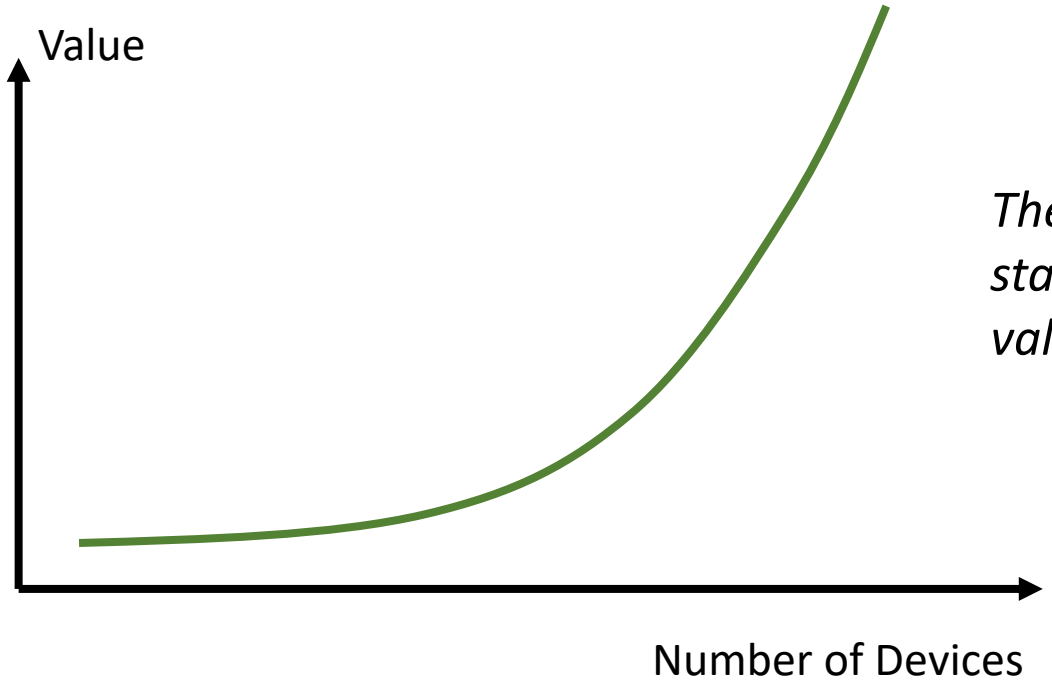
EKGs are Also Shared Resources



Metcalf's Law



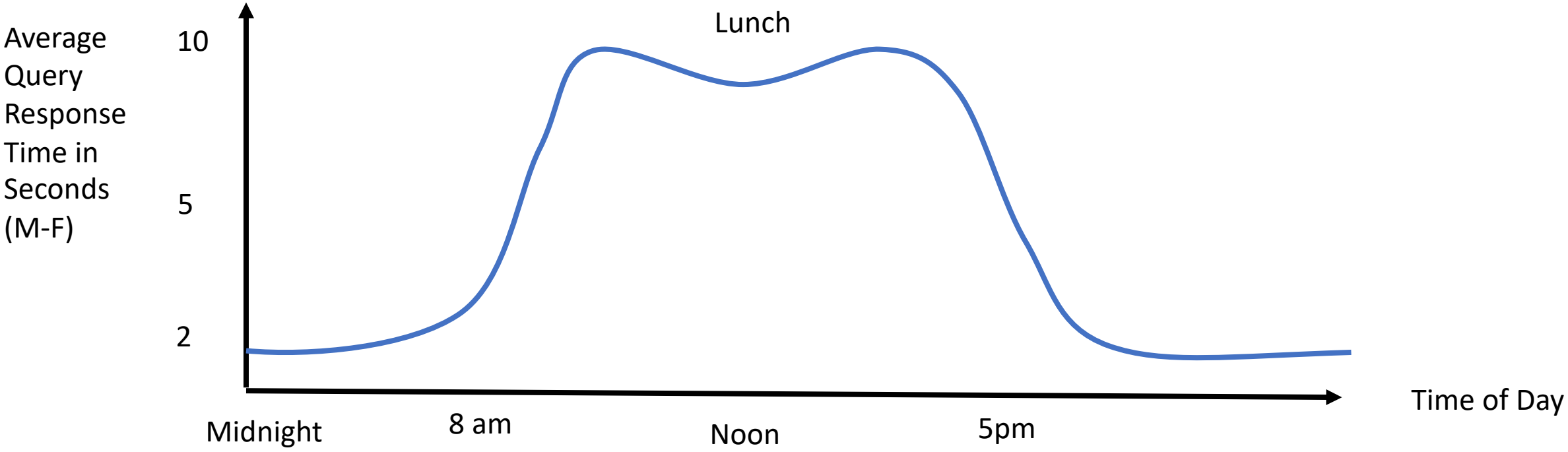
The value of a telecommunications network is proportional to the square of the number of connected users of the system (n^2)



The more everyone uses a standard, the more valuable it becomes.

https://en.wikipedia.org/wiki/Metcalfe%27s_law

Query Response vs. Time of Day



- Many users fighting for a shared resource with limited capability
- The more you use it the lower the value to others

Systems Thinking Definition

Systems thinking is a holistic approach to analysis that focuses on the way that a **system's** constituent parts interrelate and how **systems** work over **time** and within the context of larger **systems**.

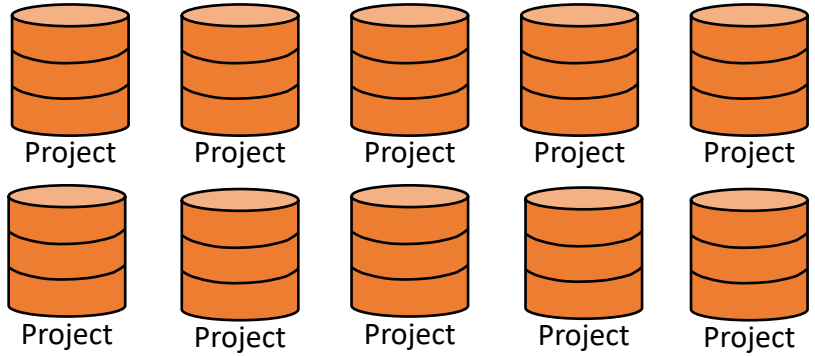
A Definition of Systems Thinking: A Systems Approach

Ross D. Arnold and Jon P. Wade

https://www.sciencedirect.com/science/article/pii/S1877050915002860

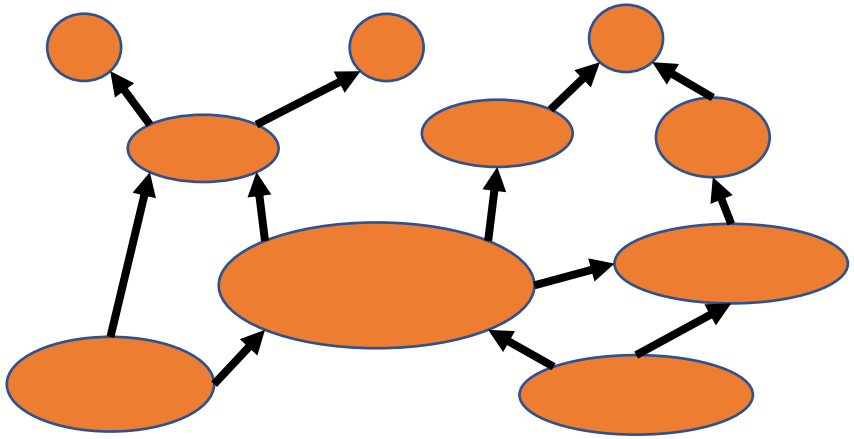
Project Silos vs Systems Thinking

Isolated Silo View



- Each project is an **independent** silo of effort
- The success of any project will not impact the success of **other** projects
- Project **order** is not relevant and project value is static in time
- Project costs and benefits are easy to represent in a simple spreadsheet
- The spreadsheet may not reflect the complexities of the real world

Systems Thinking View



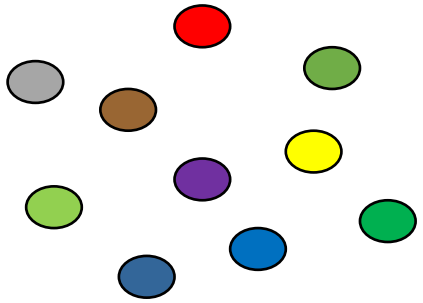
- Projects are **dependent** on another project success
- The success of **foundational** projects may have a dramatic impact on other projects (x10 faster)
- Project **order is relevant** and deferring customer benefit is needed until foundational projects are complete
- Requires a deep understanding of how resources created by one project can be leveraged by other projects
- Reflects the tacit knowledge gained over years of working in research projects and observing different teams' ability to build reusable artifacts

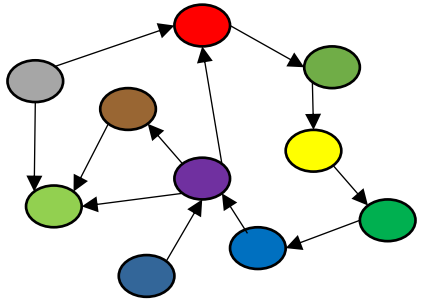
Exercise

- How do you determine what projects to start?
- How would Systems Thinking change project priorities?

Connected Data Strategy

Connected Data Strategy: *An **Enterprise Data Strategy** pattern that brings **focus** to the business **value** of connecting disparate silos of data.*



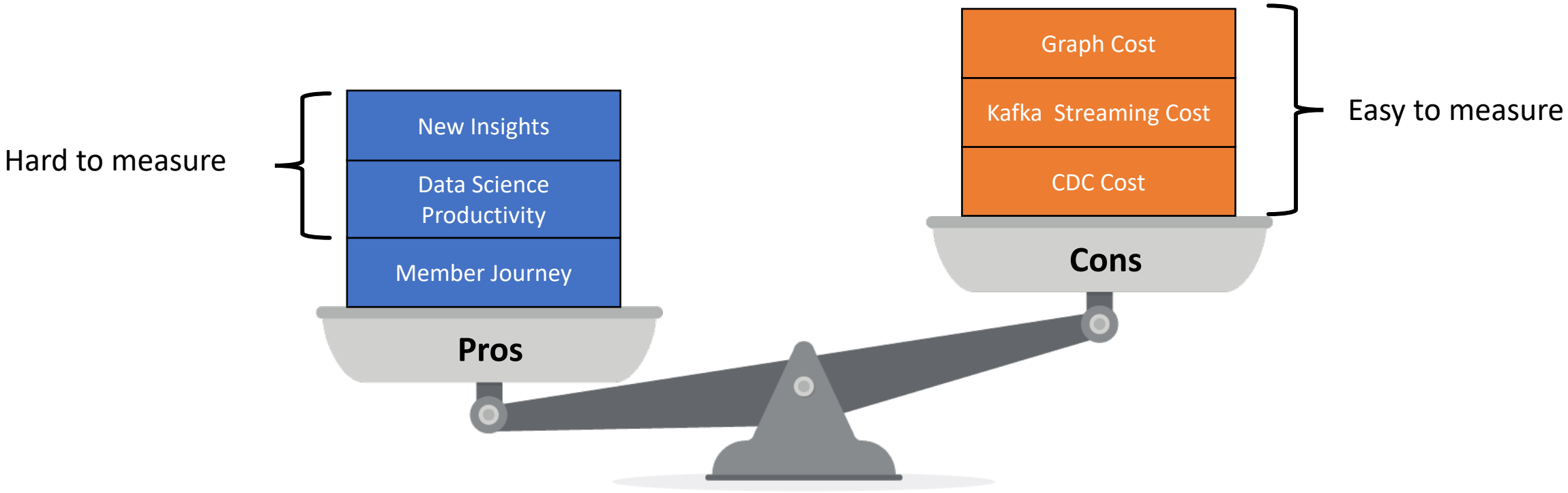


How much value?
At what cost?

$ — Lower Value

$$$$ — Higher Value

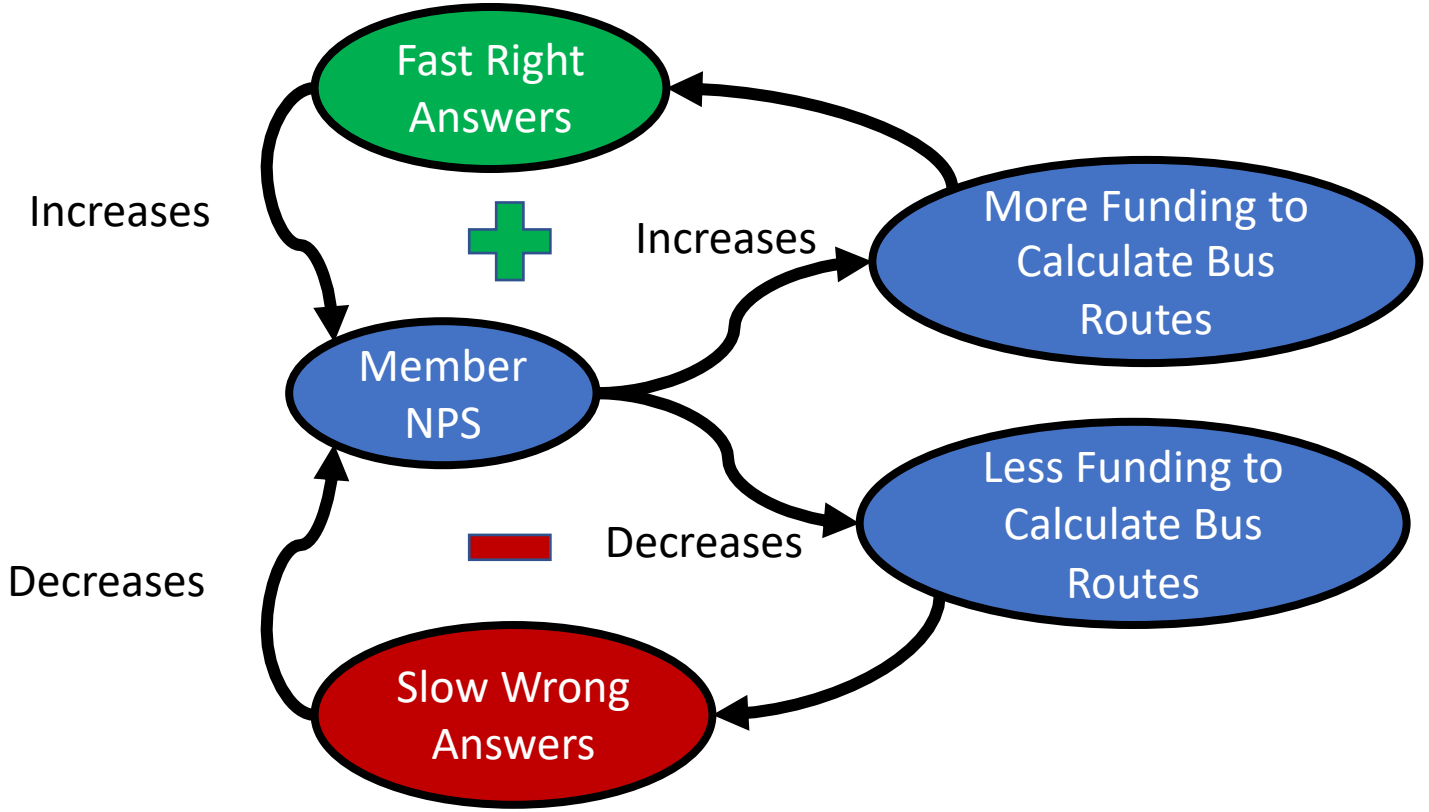
Goal: Objective Weighing of Pros and Cons



- We are starting to learn the benefits of connected data: Customer 360
- It takes time and effort to create high-quality connected data
- Can we objectively measure each of the pros and cons?
- **What is the value of "new insights?"**
- How is our decision making driven by what is easy to measure?

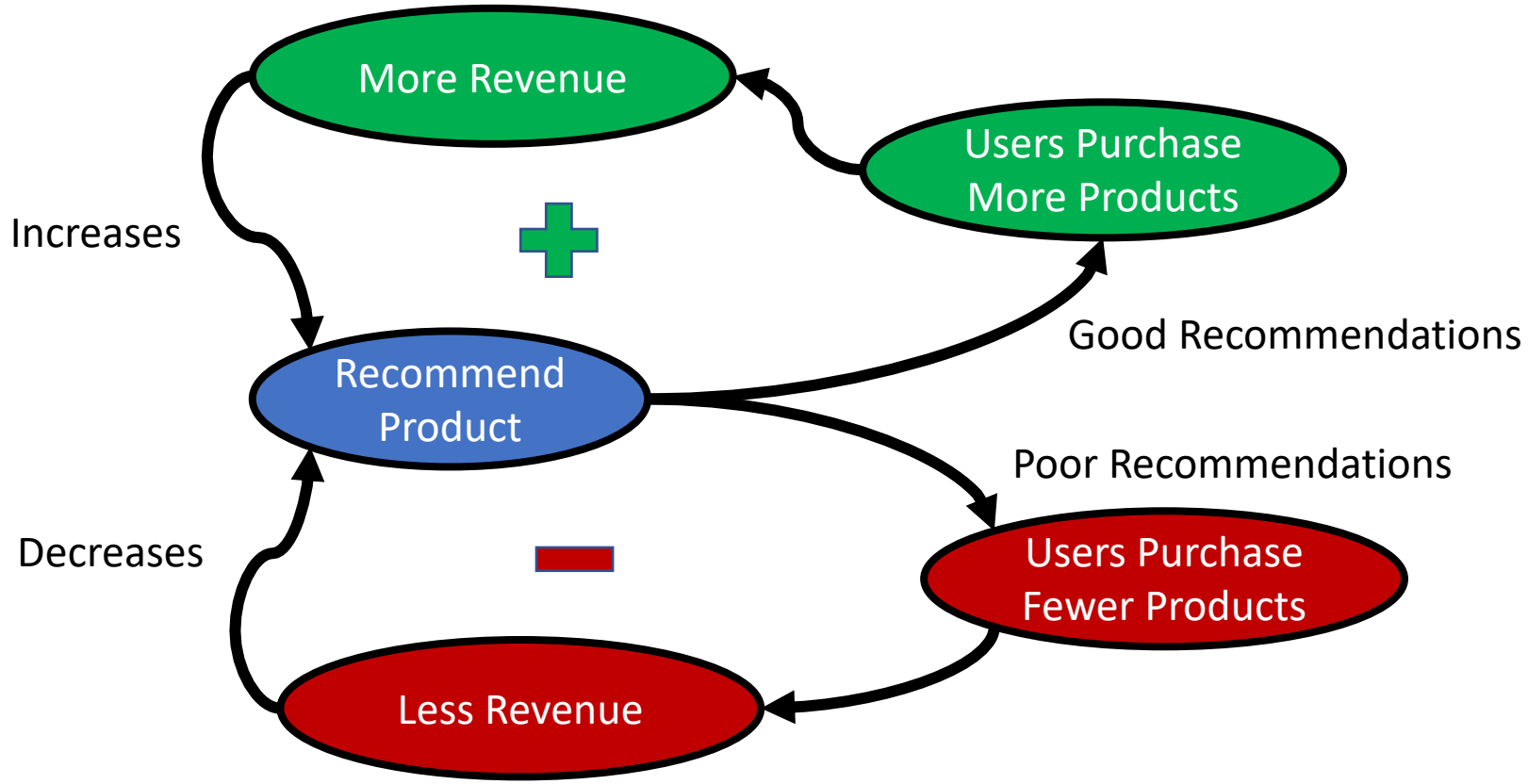
Causal Loop Diagram

NPS = Net Promoter Score – how happy are our customers

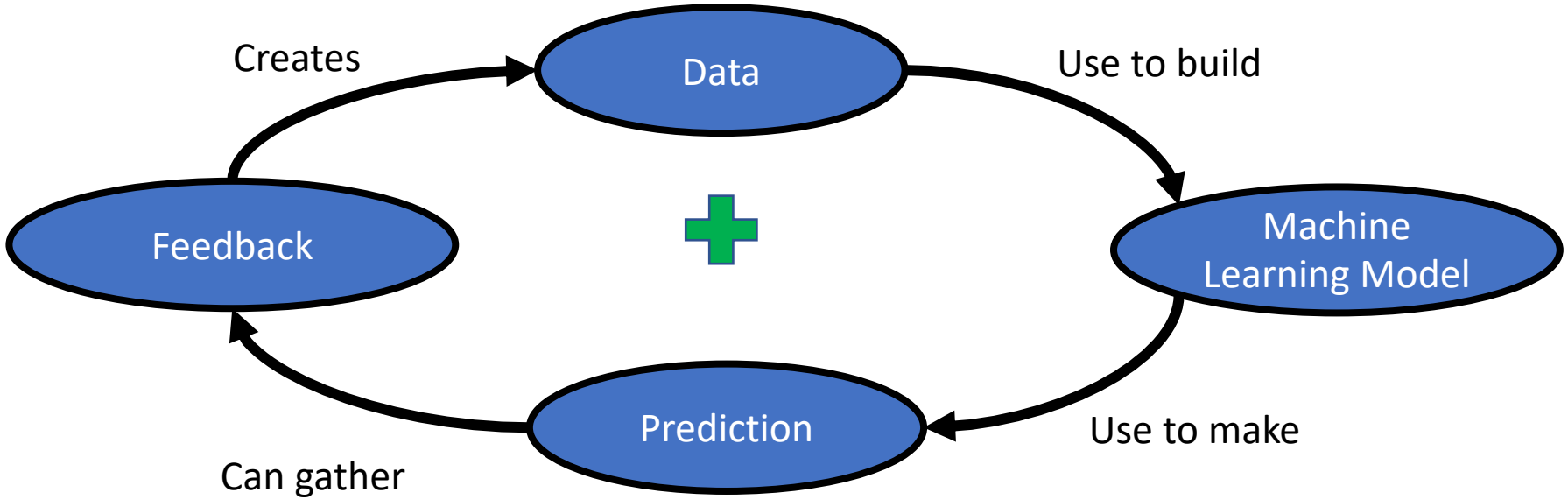


- Assumption: Higher NPS promotes higher market share

Predictive Feedback Cycle

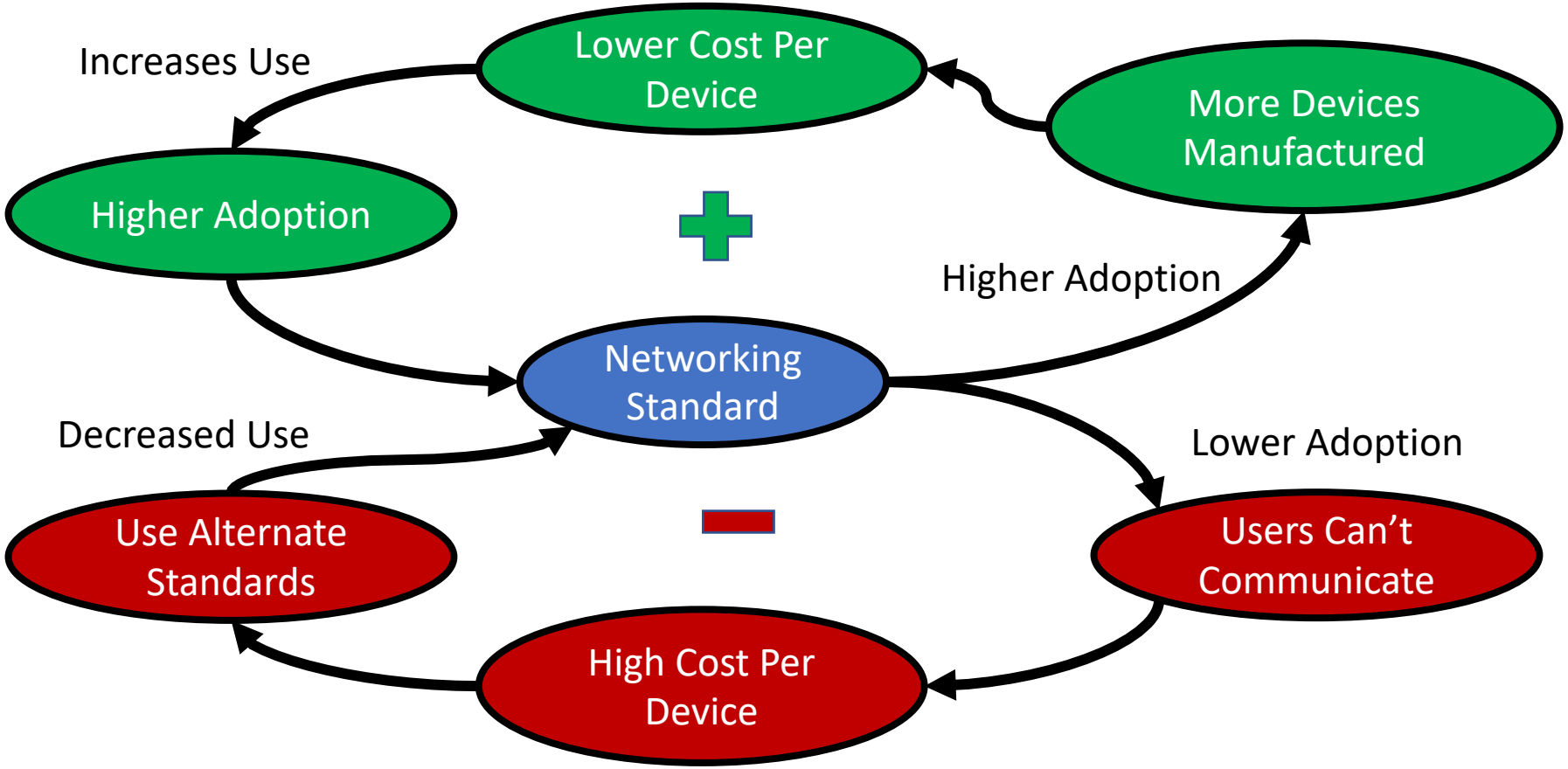


The AI Flywheel



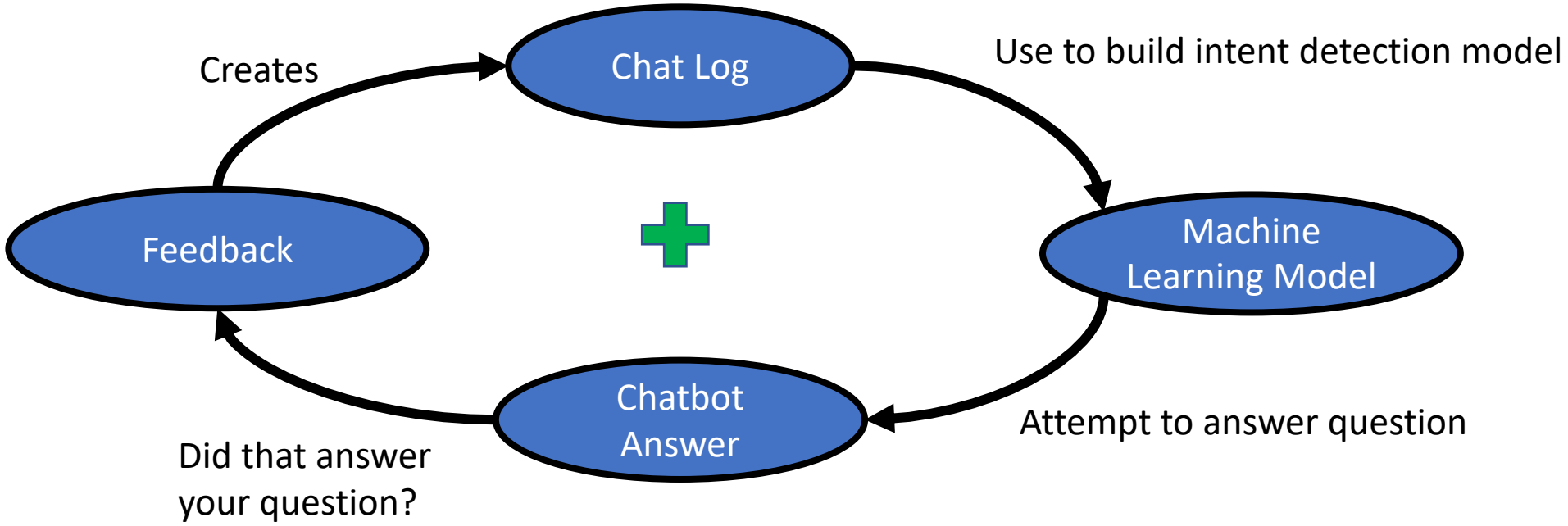
- More data creates more precise machine learning models

Network Effects (Metcalfs's Law)



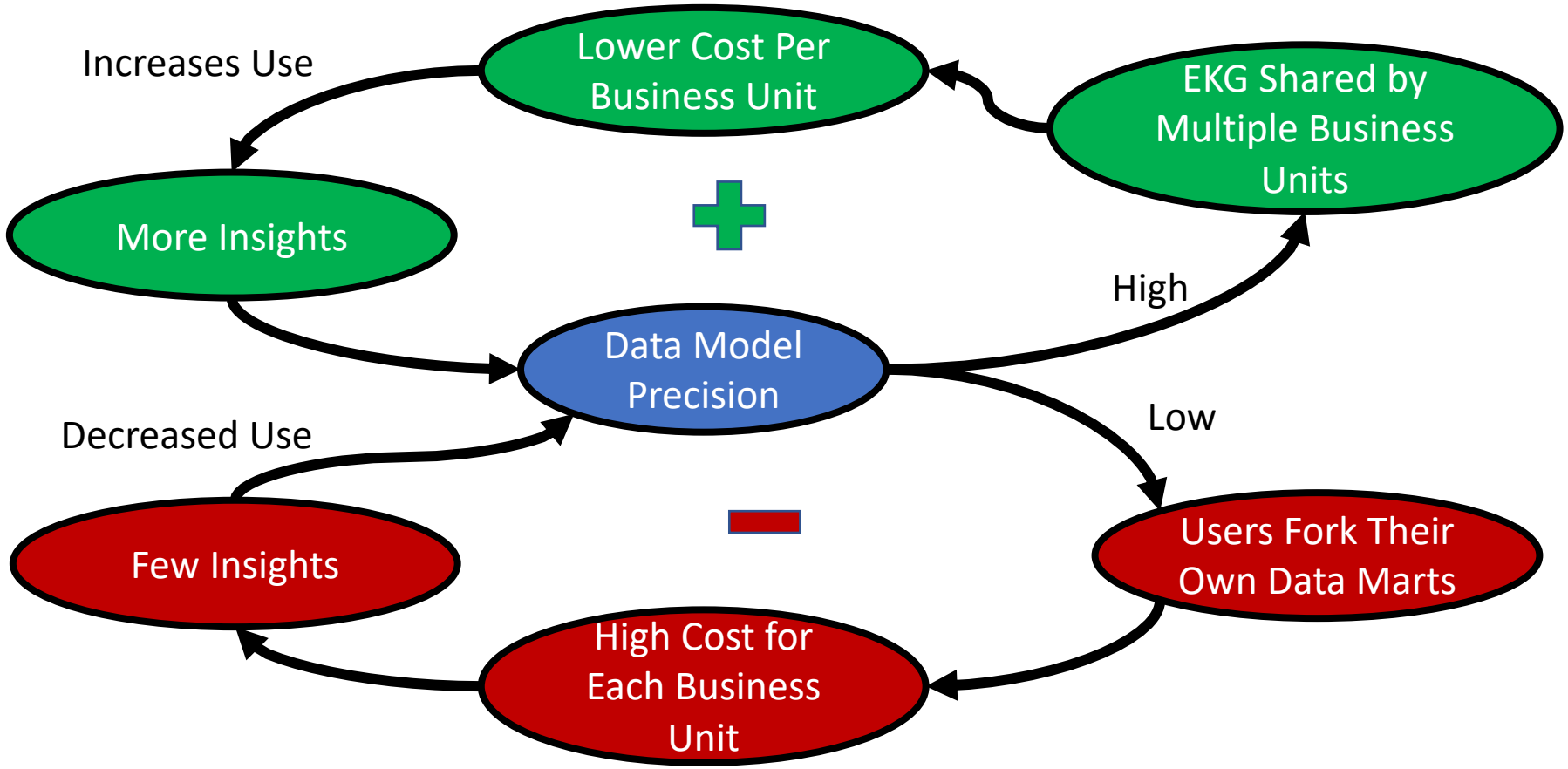
The value of any device on a network standard grows exponentially as the number of connections increase

Customer Support Chatbot



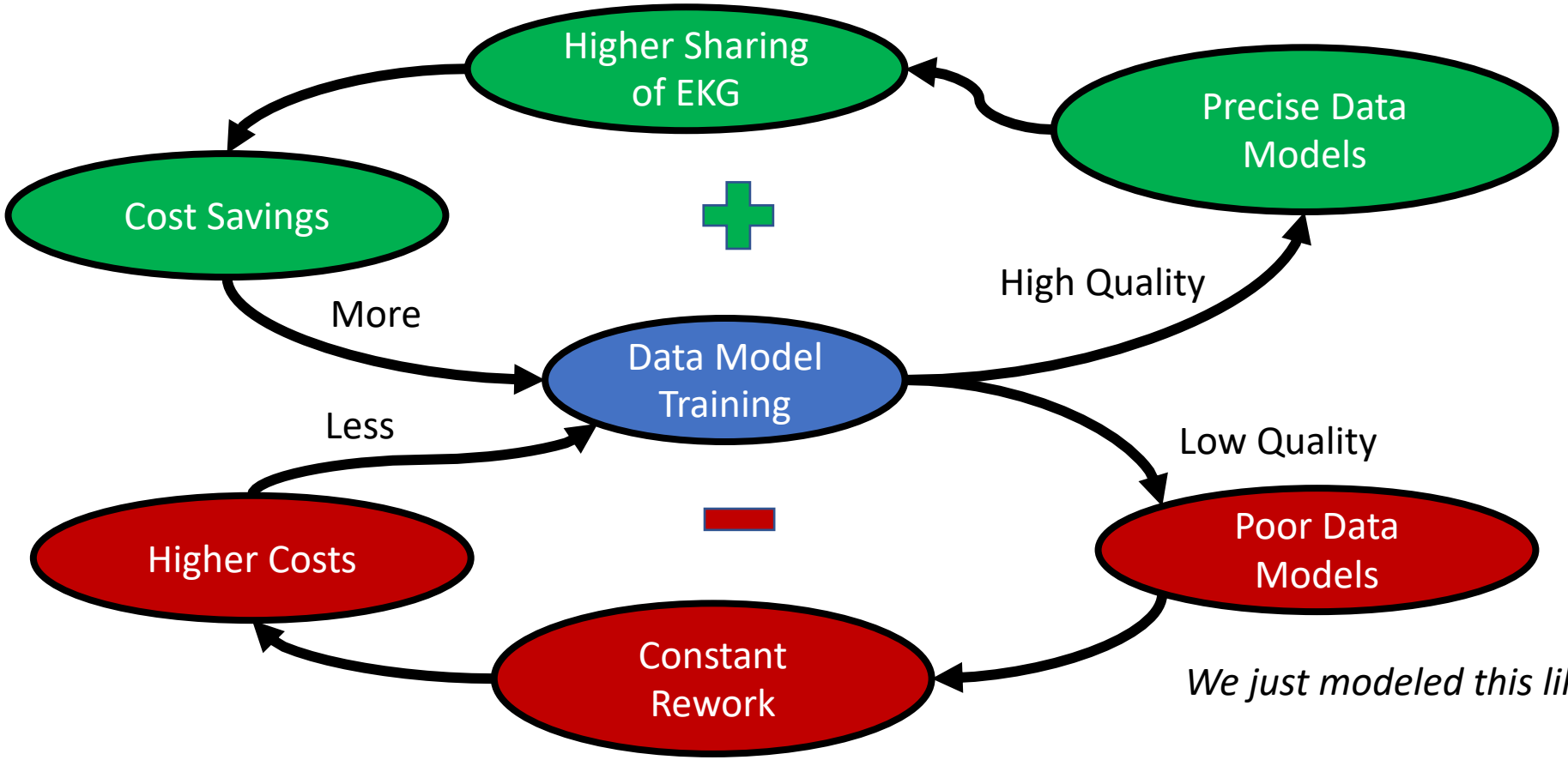
- More feedback is used to build better intent detection models

Data Model Precision and Cost Sharing



- Precise data models allow multiple business units to share EKG infrastructure costs

Data Model Precision and Cost Sharing



We just modeled this like our RDBMS

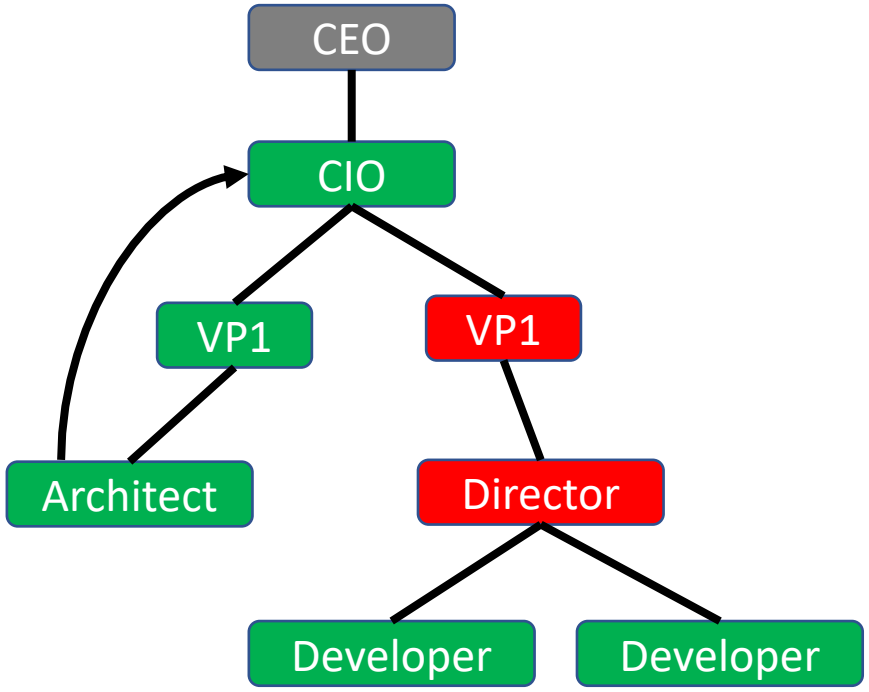
- Precise data models allow multiple business units to share EKG infrastructure costs

Org Chart vs Influence Diagram

Promoters

Neutral

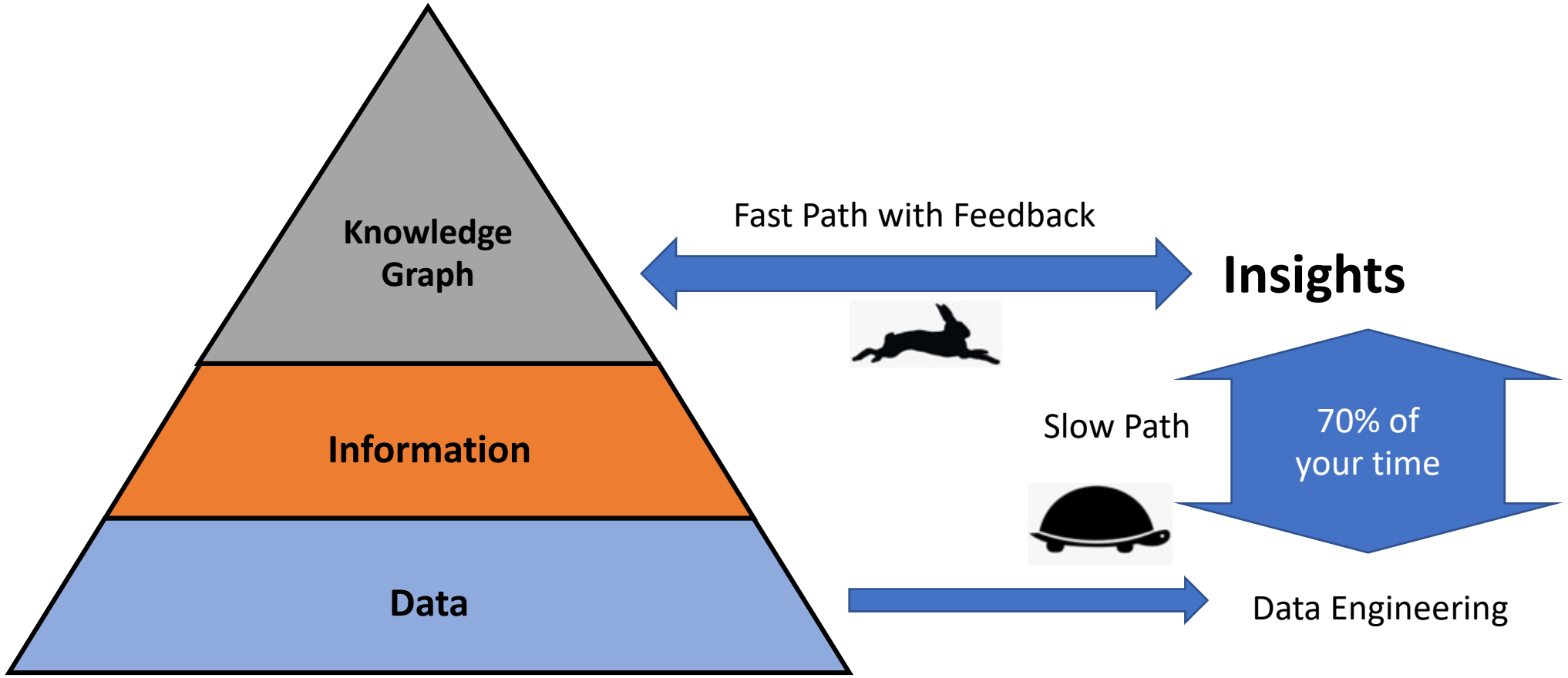
Detractors



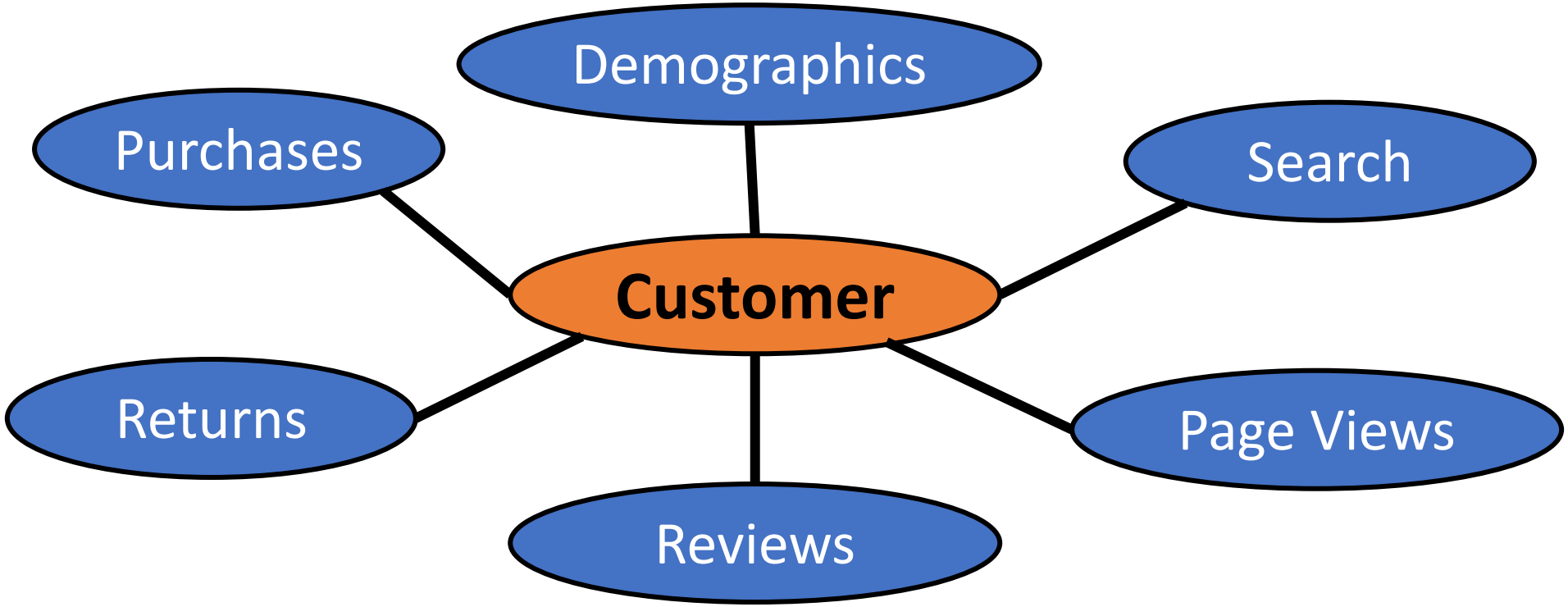
Systems Thinking Terminology

- Archetypes
- Balancing Loop
- Digital Twin
- Dynamics
- Feedback
- Influence Diagram
- Unintended consequences
- Local optimization
- Emergence
- Flow
- Leverage Points
- Limiting Factor
- Nonlinear Relationships
- Resilience
- Reinforcing Loop
- Self Organization
- Sustainability
- Unintended consequences

From Data Scientist to Knowledge Scientist



Customer-Centric EKG Strategy



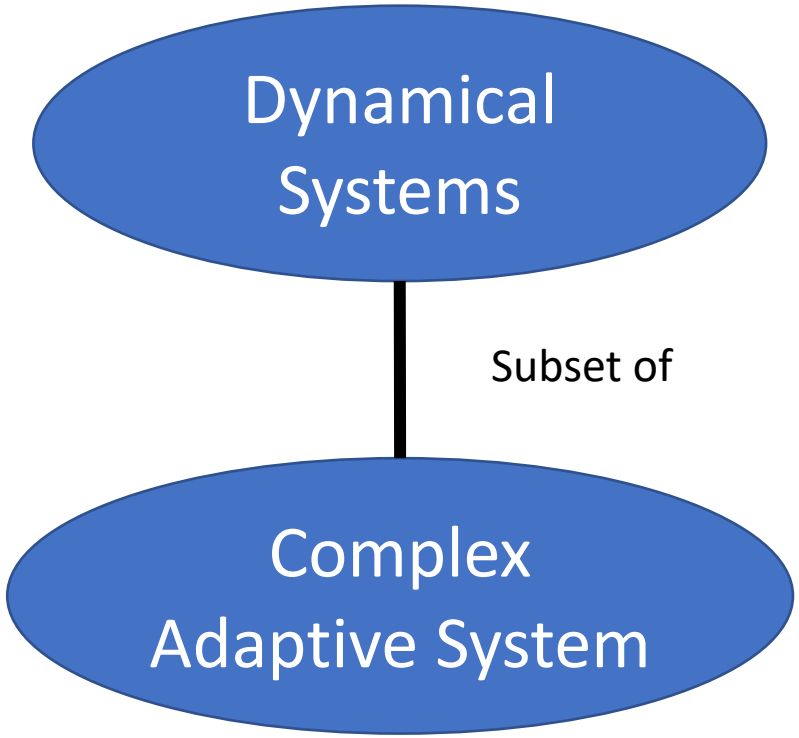
- Put your customer at the center of your EKG

Modeling Precision

All models are wrong. Some models are useful. - George Box

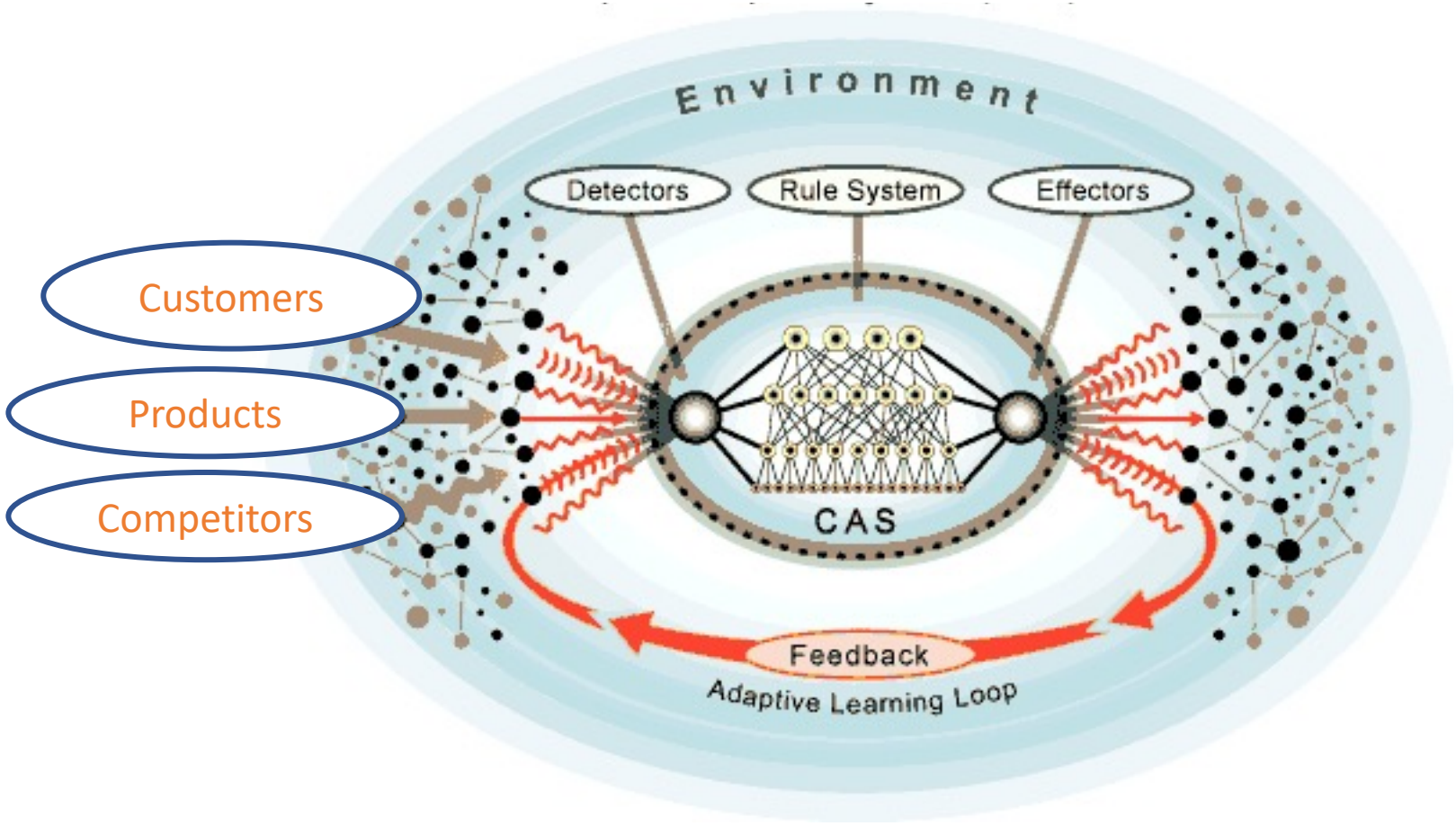
- Graph data models are better at modeling the real world
- Models that are correct are easily shared across business units

Advanced Systems Theory



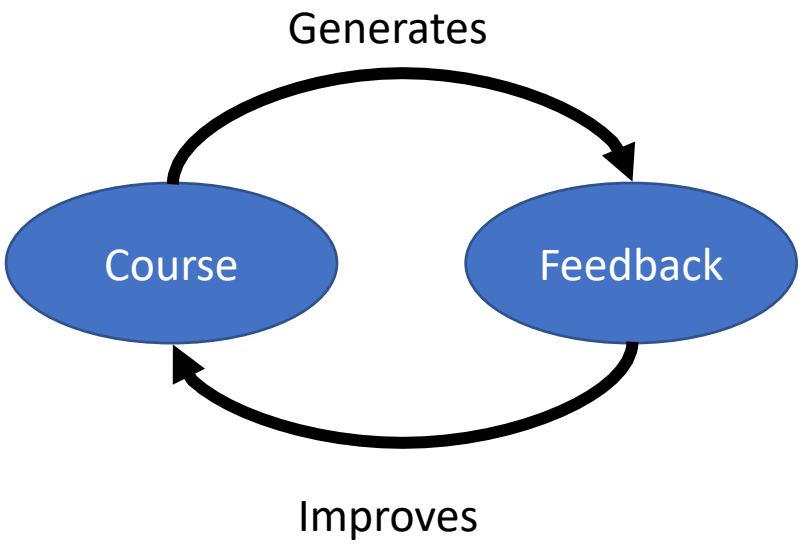
- Fields of Study

Complex Adaptive System As Rules



Inspired by http://www.calresco.org/lucas/cas.htm

Thank You!



Dan.McCreary@gmail.com

https://www.linkedin.com/in/danmccreary

https://dmccreary.medium.com/

@dmccreary