avuxeni

sanibonani

dumelang

molweni

hallo

lotjhani

sanibonani

molweni

avuxeni

hello

hello

hallo

ri a vusa

ri a vusa

avuxeni

molweni

lotjhani

dumelang

hallo

dumelang

avuxeni

ri a vusa

Agenda and Webinar format

David Gouvias

Data Scientist

Data Science Hackathon 2021

- Welcome. PW Janse van Rensburg
- Graph Database Journey. Derick Schmidt
- Introduction to Graph Databases. Monika du Toit
- Introduction to AWS and SageMaker. Preshen Goobiah
- Neptune Graph Database and Gremlin (David Gouvias)
- Data Science Graph Algorithms (Ockert Janse Van Rensburg, Dalubuhle Mbune)
- Hackathon Challenge (David Gouvias)
- Data definition and reference Graph Database Design. (David Gouvias)
- Judges, Prizes and final logistics.

Welcome

PW Janse van Rensburg

Manager: Data Science - Client Insights

CAPITEC

Graph Database Journey

Derick Schmidt

Manager: Client Data Platform

CAPITEC

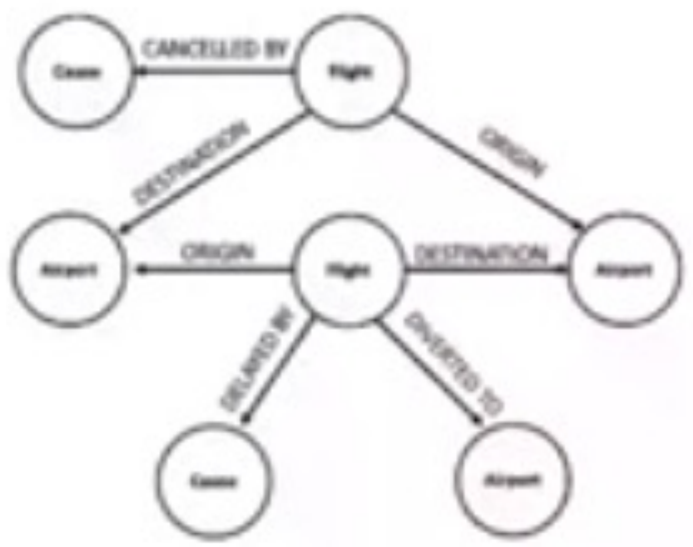
Introduction to GraphDB and Capitec Data Science

Monica Du Toit

Data Scientist

Graphs everywhere

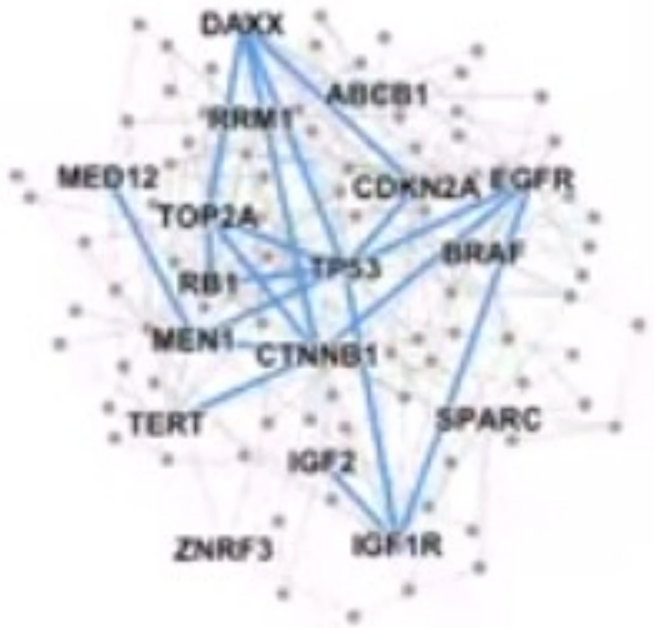
Relationships of highest priority



Event Graphs

Image credit: SalientNetworks

Computer Networks



Disease Pathways

Chesapeake Bay Waterbird Food Web

Image credit: Wikipedia

Food Webs

Image credit: Pinterest

Particle Networks

Image credit: visitlondon.com

Underground Networks

Graphs everywhere

Relationships of highest priority

Image credit: Medium

Social Networks

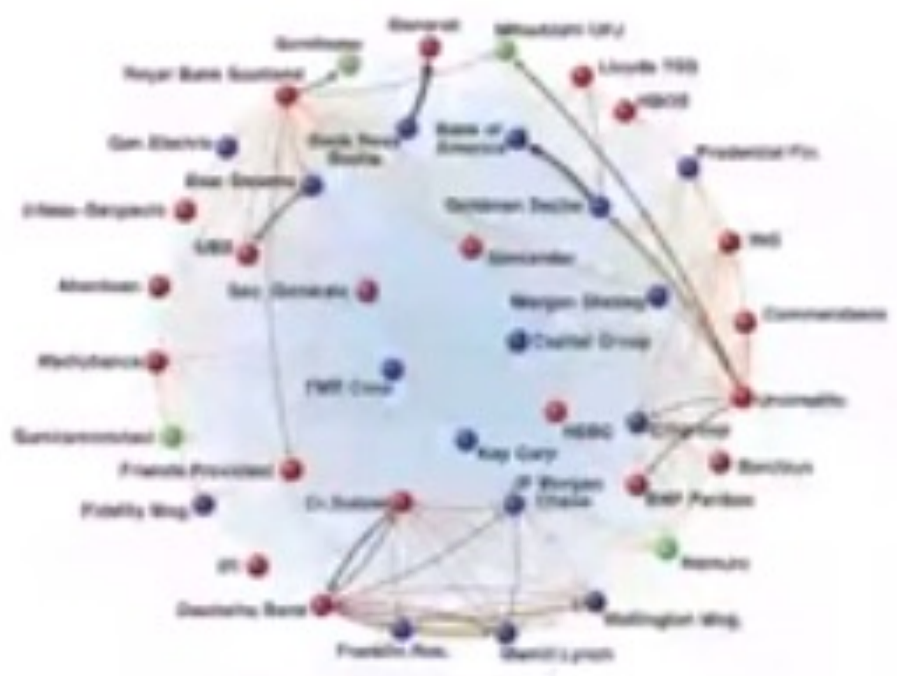


Image credit: Science

Economic Networks

Image credit: Lumen Learning

Communication Networks

Citation Networks

Image credit: Missoula Current News

Internet

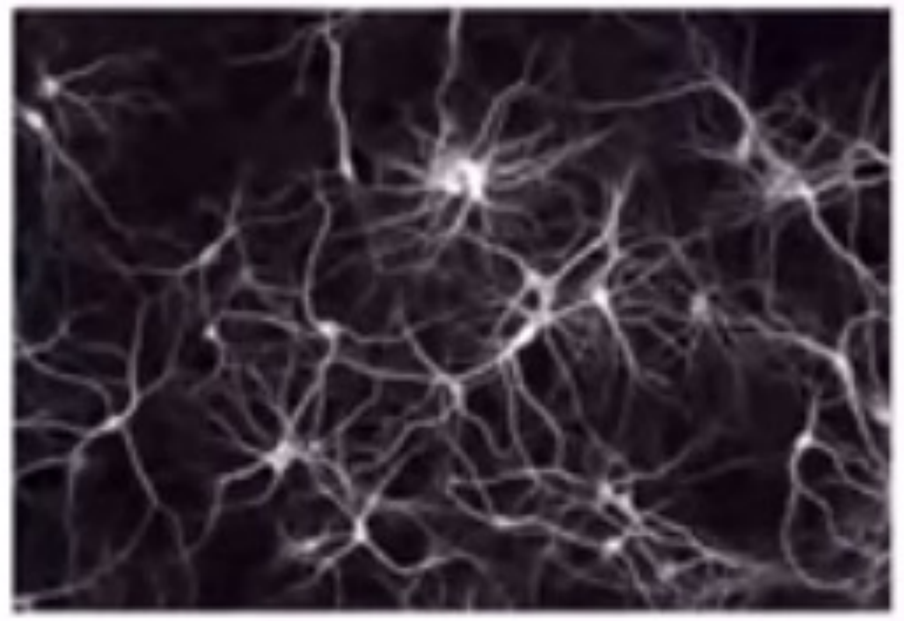
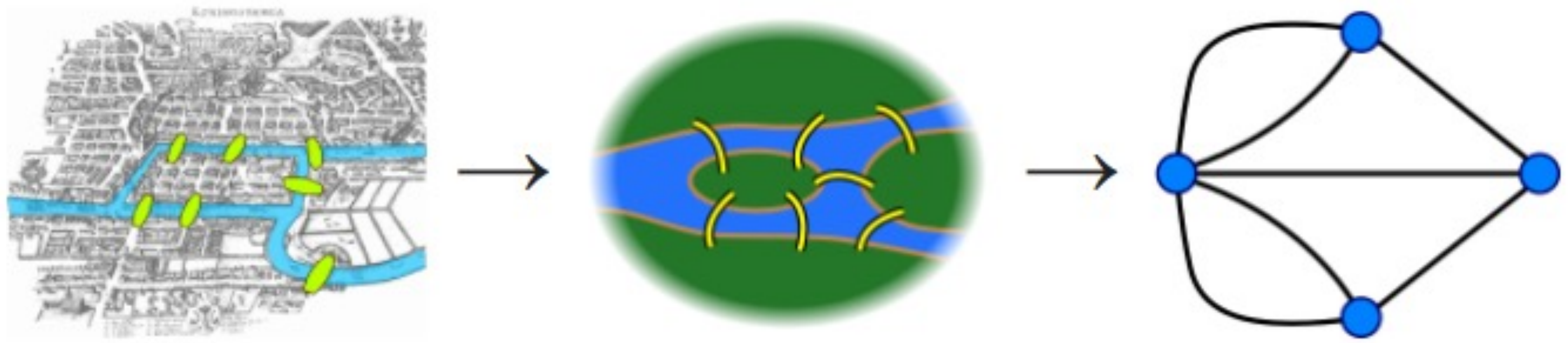


Image credit: The Conversation

Networks of Neurons

GraphDB

Relationships of highest priority



Leonard Euler

The Seven Bridges of Konigsberg (1736) laid the foundations of graph theory. Euler proved that the problem has no solution.

Eulerian Path = a walk through the city that would cross each bridge/edge only once.

GraphDB are a general language for describing entities with relationships.

Nodes represent entities or other domain components.

Edges connect two nodes and represent relationships between entities.

Nodes and edges can contain properties that hold name-value pairs of data.

https://fortelabs.co/blog/mapping-the-habit-graph/

https://blog.stratio.com/graph-database-clustering-solution/

GraphDB

Use cases across the world – node level

50 year old Protein Folding problem

Predict a protein’s 3D structure based solely on its amino acid sequence (DeepMind’s AlphaFold)

Represent underlying protein as a graph, using graph neural network, predicting new position of the amino acids.

Study living things in new ways, enable quicker and more advanced drug discovery.

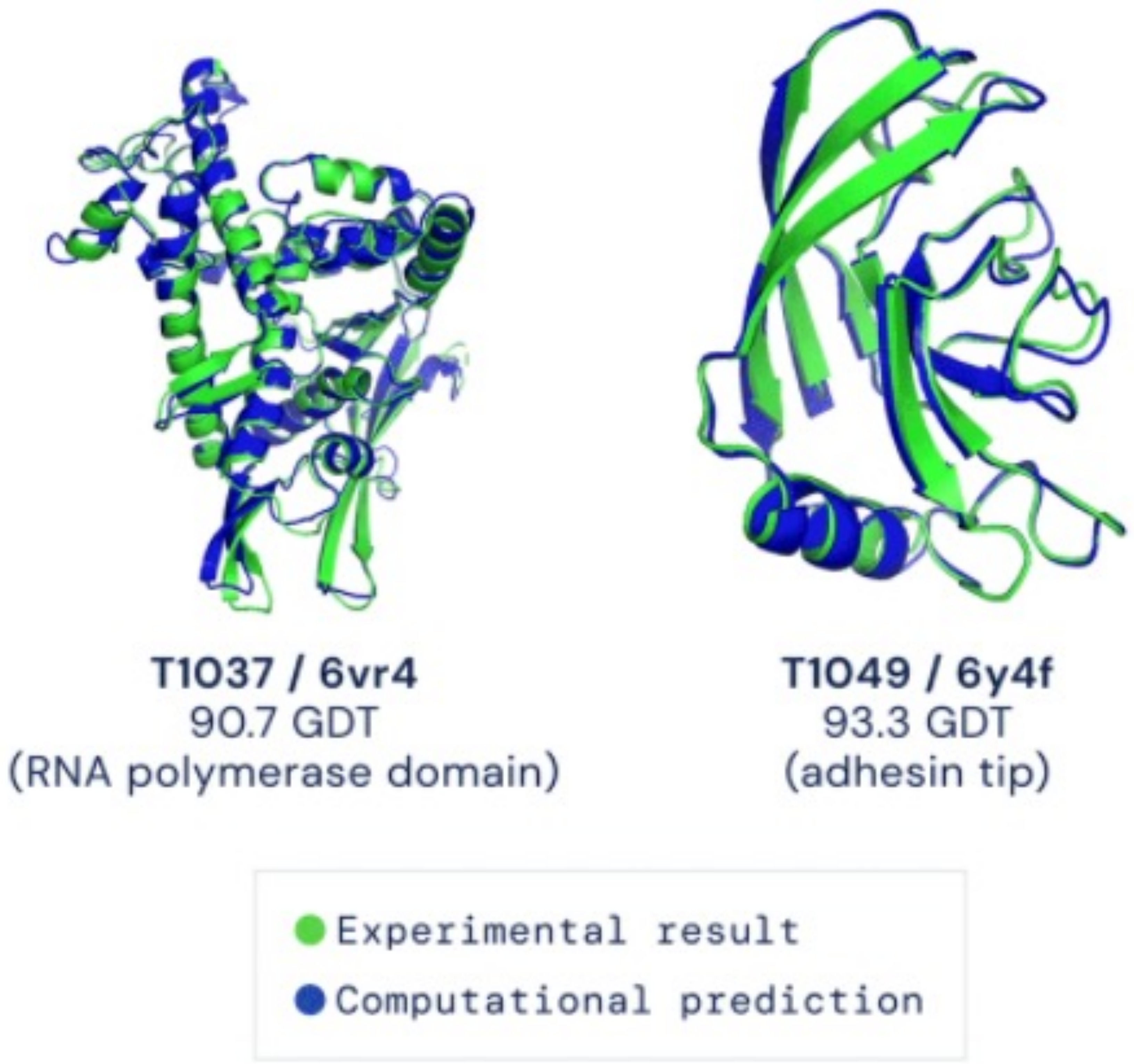
“Help to illuminate the function of the thousands of unsolved proteins in the human genome, and make sense of disease-causing gene variations that differ between people.”

https://www.nature.com/articles/d41586-020-03348-4

DeepMind

NEWS | 30 November 2020

‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures



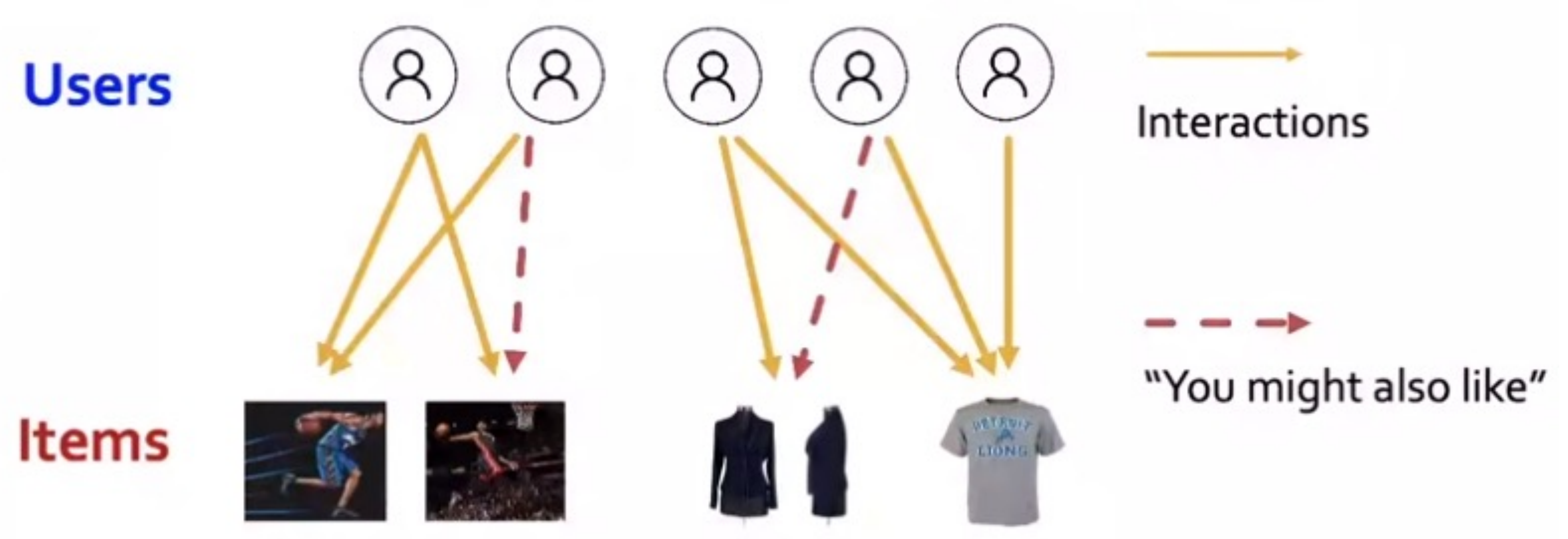
GraphDB

Use cases across the world – edge level

Recommender Systems

Nodes: users and items; Edges: user-item interactions

Recommend items users might like (watching movies, purchasing products, listening to music, etc)



Use graph neural network to predict clients' interests by considering relationships between clients and relationships between clients and their past interests.

"Existing research has shown the efficacy of graph learning methods for recommendation tasks."

Pinterest, LinkedIN, Facebook, Instagram, Alibaba, Netflix.

https://eng.uber.com/uber-eats-graph-learning/

Google Maps

DeepMind

GraphDB

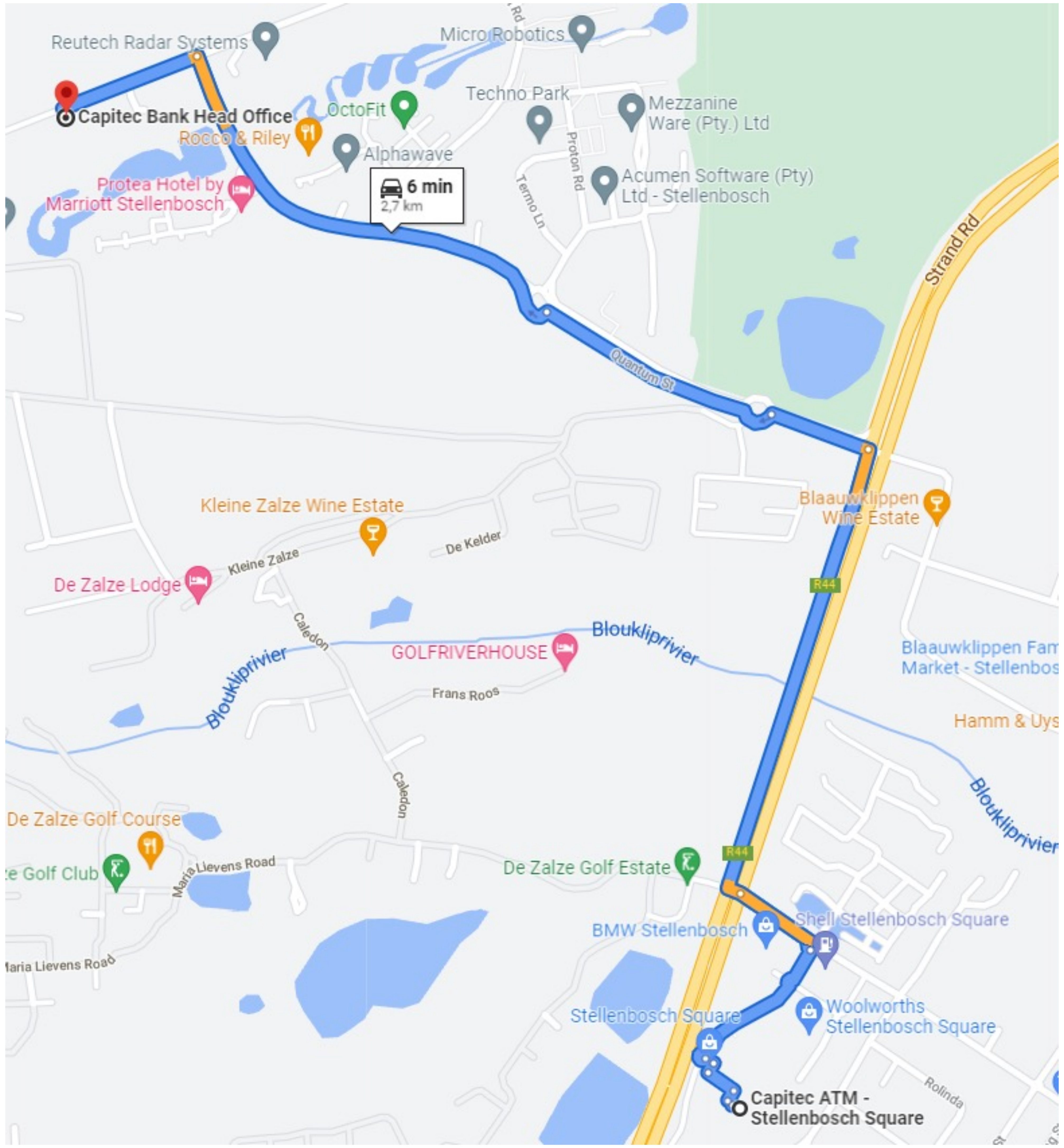
Use cases across the world – subgraph-level

Traffic prediction

Graph Neural Network approach based on collisions, traffic patterns trained and roads quality to find shortest path and predict travel time.

Nodes: road segments; Edges: connectivity between road segments

“Each day more than 1 billion km of road are driven with the app’s help. Google says using DeepMind’s AI tools have improved the accuracy of ETAs in Maps by up to 50 percent.”



https://www.theverge.com/2020/9/3/21419632/how-google-maps-predicts-traffic-eta-ai-machine-learning-deepmind

Capitec

We Believe That Banking is About People

20 years old

- 16.8mil clients
- 8.9mil active retail digital banking clients – biggest digital bank in SA
- 623 mil digital transactions last 6m
- Open GlobalOne account remotely

Our Fundamentals:

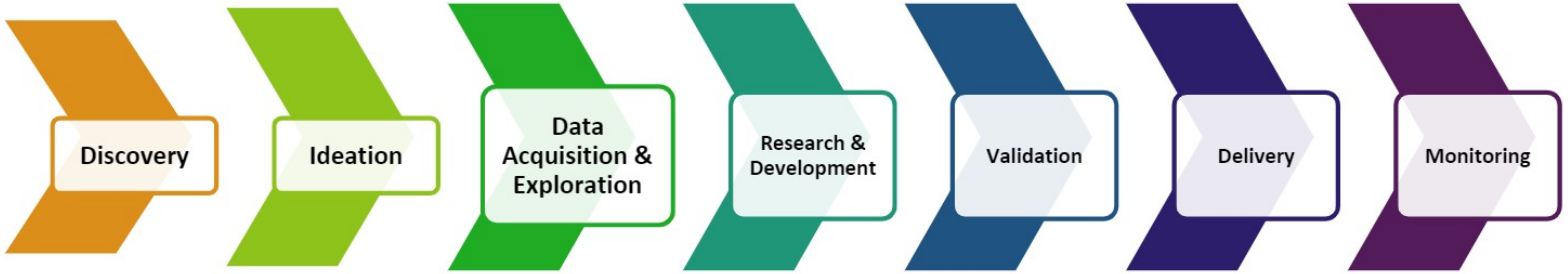
- Simplicity, Affordability, Accessibility, Personalized experience

New products:

Live Better Savings Account; Financial Education; Virtual Card; Scan to Pay; EasyEquities; Remote Credit; Business Bank

Capitec

Data Science team, est. 2017



Self-driving Credit

Client Engagement

ENGAGEMENT
YOU KEEP USING THAT WORD, I DON'T THINK IT MEANS WHAT YOU THINK IT MEANS

AML

Innovation / Optimisation

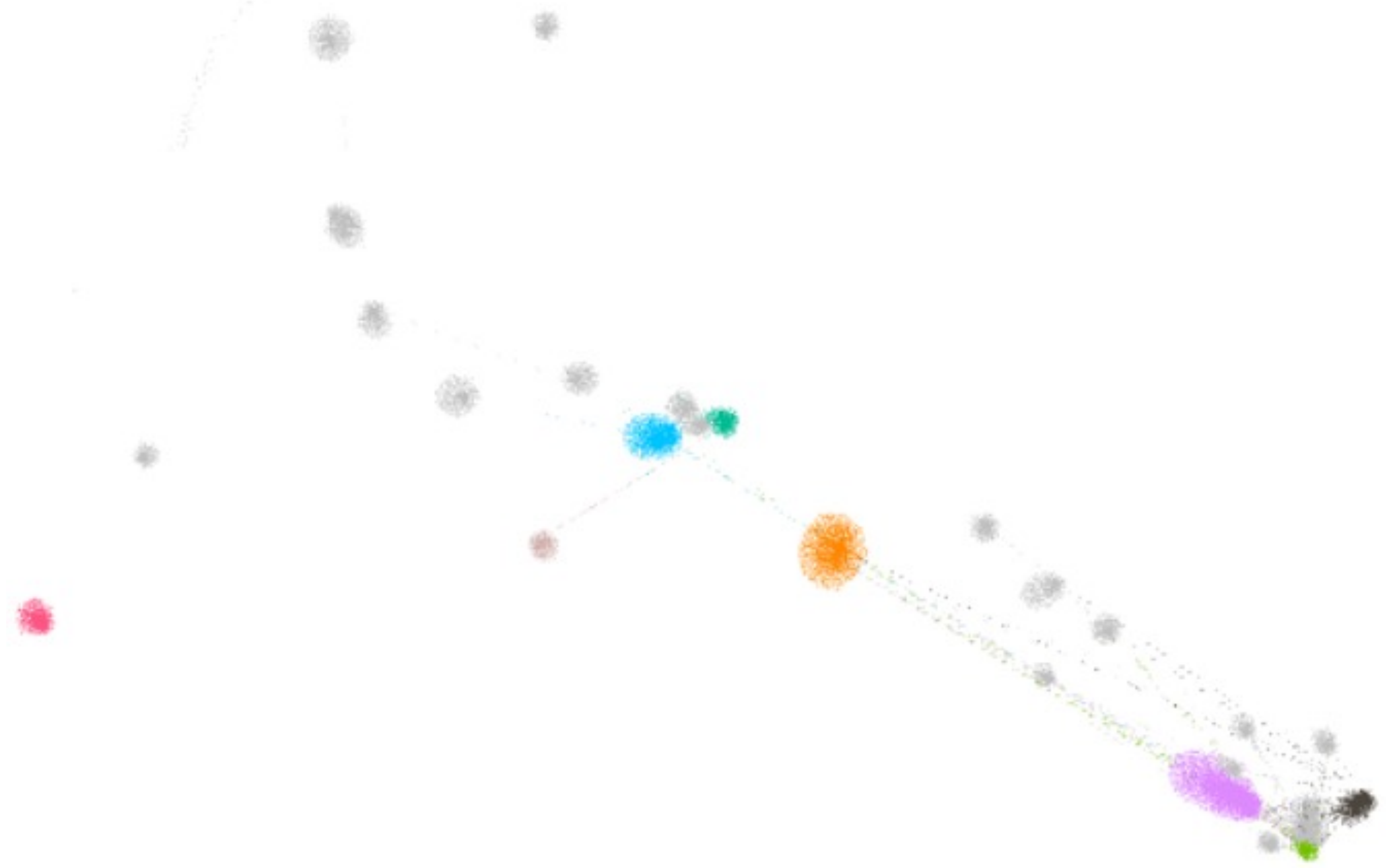
1. **People matter**: Data science will always impact people's lives
2. **Truth matters**: Keep searching
3. **Knowledge matters**: Keep learning
4. **Individual knowledge is limited**: Keep collaborating
5. **You matter**: Respect differences

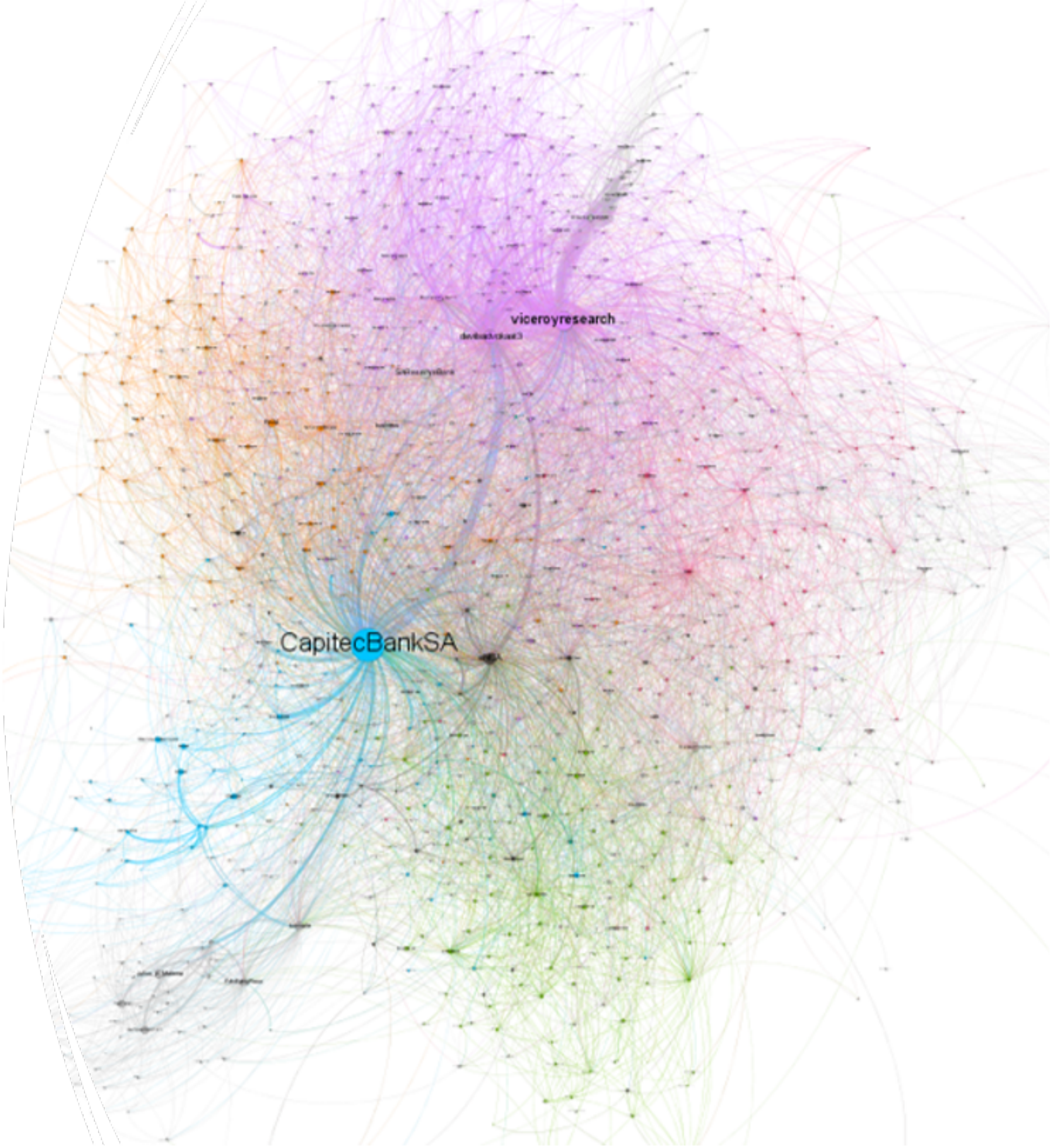
DATA SCIENCE

Capitec GraphML

Use cases

1. Identify likely merchant clients and convert to Business Bank clients.
2. Suggest potential new clients for existing business clients.
3. Identify fraudulent activity on client’s accounts.
4. Recommend most relevant product based on client’s need.
5. Discover Capitec client communities from social media data.





thank you

any questions?

#SimplifyBanking #LiveBetter

CAPITEC

capitecbank.co.za

Capitec Bank Limited is an authorised financial services provider (FSP 46669) and registered credit provider (NCRCP13) Capitec Bank Limited Reg. No.: 1980/003695/06

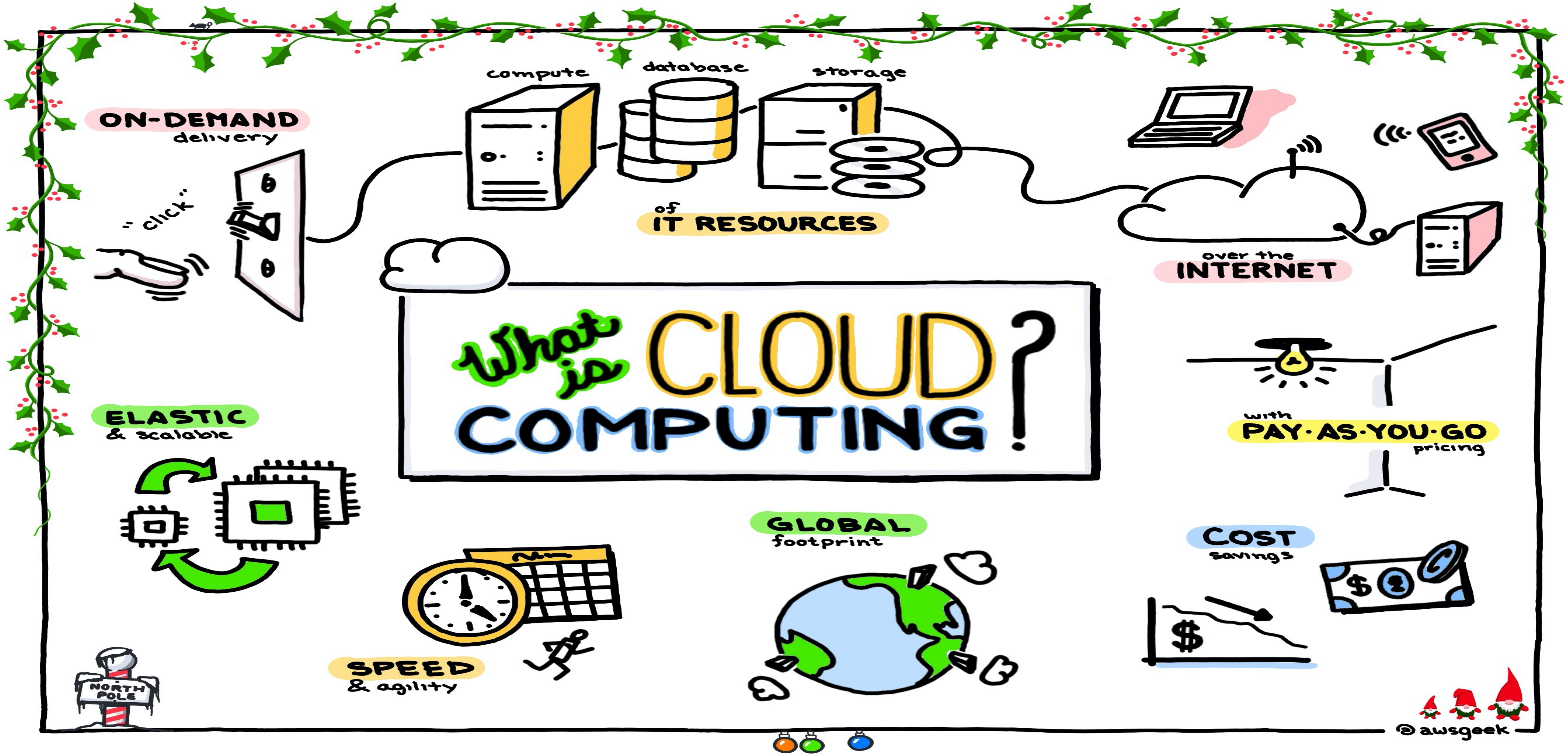
Cloud Computing and AWS

Preshen Goobiah

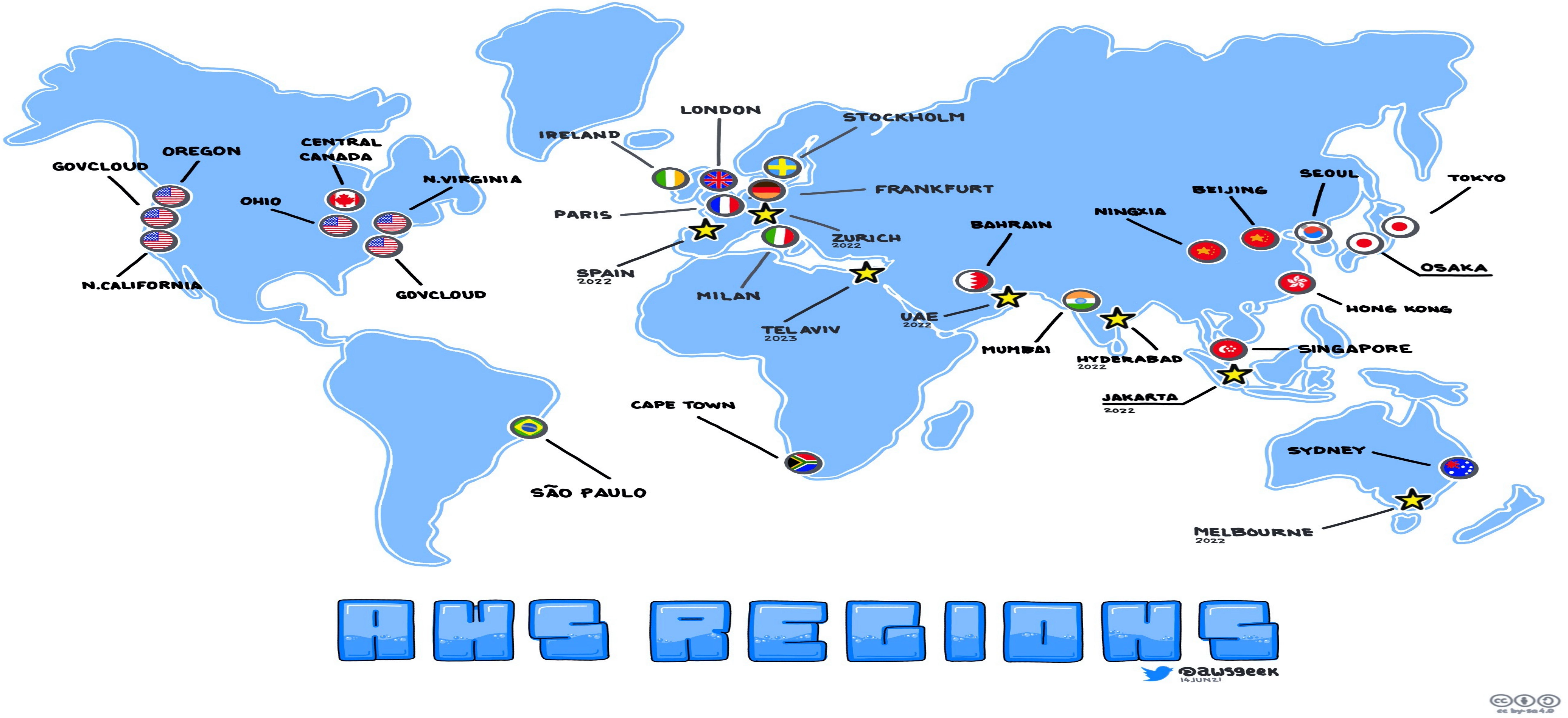
Machine Learning Engineer

CAPITEC

What is Cloud Computing and AWS



AWS Global Infrastructure

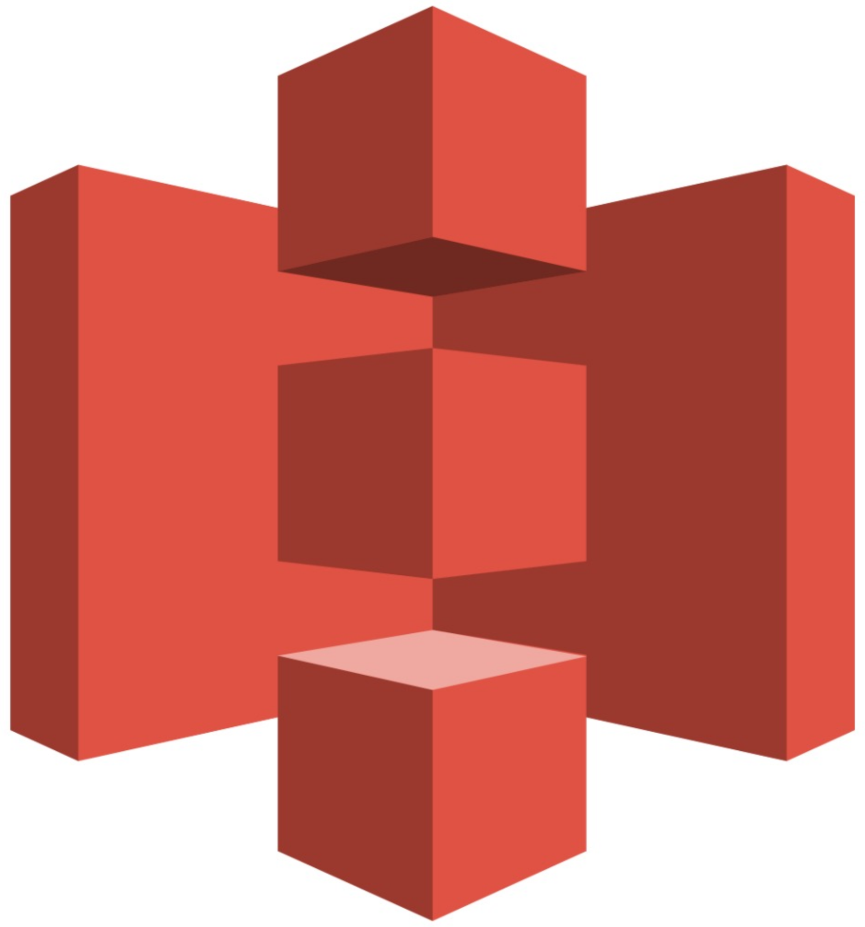


AWS Service Landscape

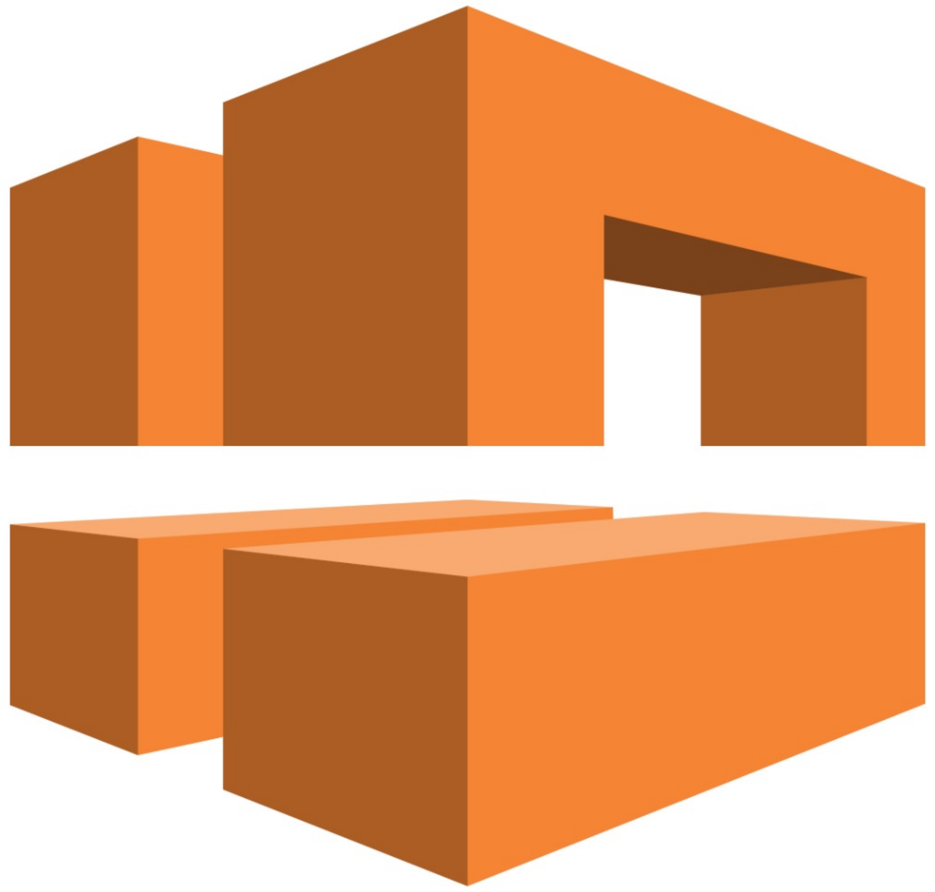
- IoT
- Game Development
- Customer Engagement
- Business Applications
- Security & Compliance
- AR & VR
- Application Integration
- Machine Learning
- Analytics
- Media Services
- Satellite
- Robotics
- Blockchain
- Mobile
- Migration & Transfer
- Network & Content Delivery
- Developer Tools
- Compute
- Storage
- Databases

AWS Core Services

EC2



S3



VPC

The AWS ML stack

Platform vs Application services

AI SERVICES

VISION	SPEECH		TEXT			SEARCH	CHATBOTS	PERSONALIZATION	FORECASTING	FRAUD	DEVELOPMENT	CONTACT CENTERS
Amazon Rekognition	Amazon Polly	Amazon Transcribe	Amazon Comprehend	Amazon Translate	Amazon Textract	Amazon Kendra	Amazon Lex	Amazon Personalize	Amazon Forecast	Amazon Fraud Detector	Amazon CodeGuru	Contact Lens
		+Medical	*+Medical*									*For Amazon Connect*

ML SERVICES

Amazon SageMaker	Ground Truth	AWS Marketplace for ML	SageMaker Studio IDE									Neo	Augmented AI
			Built-in algorithms	Notebooks	Experiments	Processing	Model training & tuning	Debugger	Autopilot	Model hosting	Model Monitor		

ML FRAMEWORKS & INFRASTRUCTURE

TensorFlow

PyTorch

mxnet

GLUON

HOROVOD

DeepGraphLibrary

Keras

scikit learn

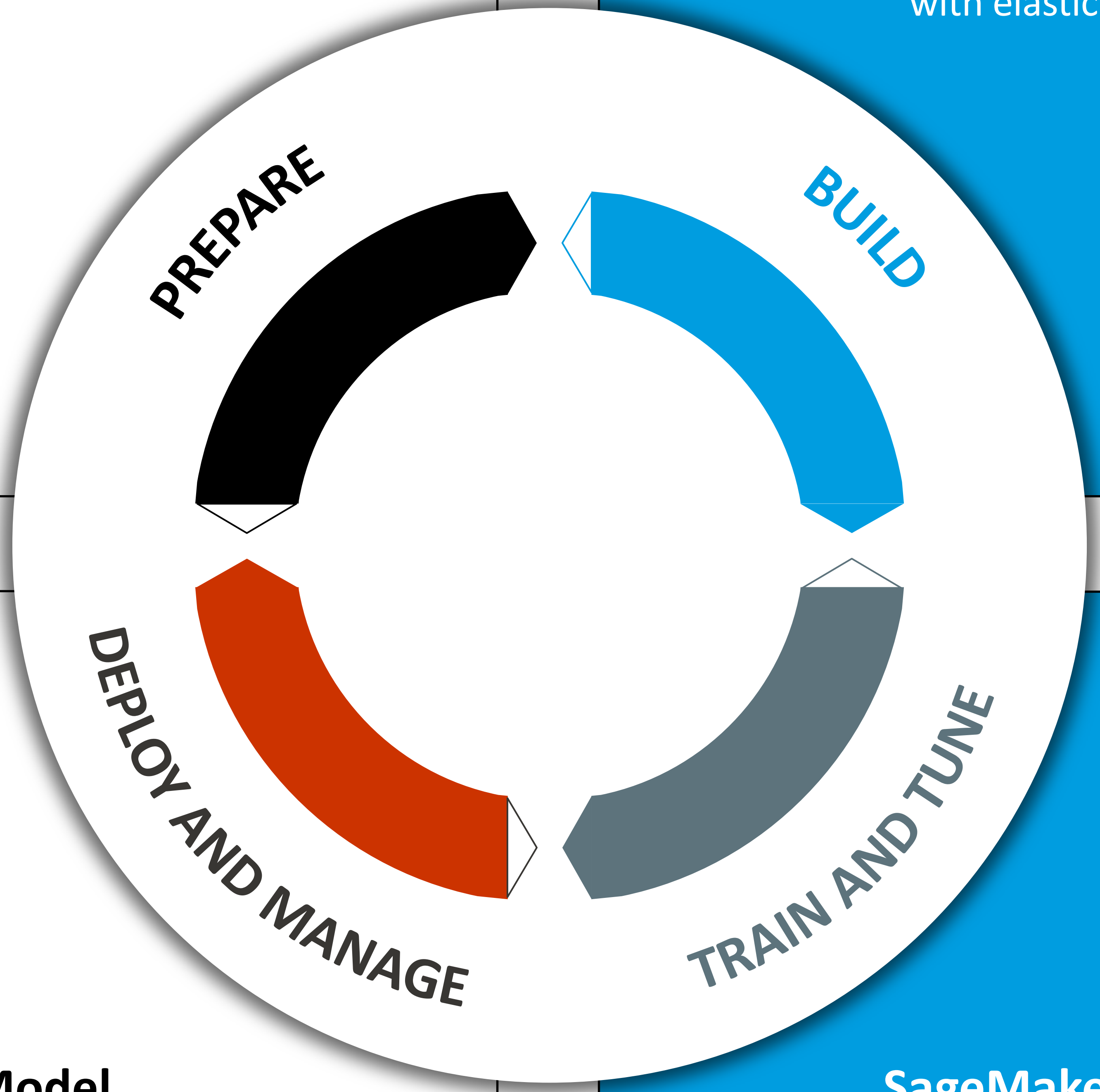
Deep Java Library

Deep Learning AMIs & Containers	GPUs & CPUs	Elastic Inference	Inferentia	FPGA

Amazon SageMaker features overview

SageMaker Ground Truth
Fully managed data labeling

SageMaker Processing
SKLearn, Spark, BYO



SageMaker Notebook Instances
One-click notebooks with elastic compute

SageMaker Studio Notebooks
One-click notebooks with elastic compute

Built-in and bring your-own algorithms
Supervised and unsupervised algorithms

AWS Marketplace
Pre-built algorithms and models

SageMaker Autopilot
Automatically build and train models

Inf1/Amazon Elastic Inference
High performance at lowest cost

Amazon Augmented AI
Add human review of model predictions

SageMaker Neo
Train once, deploy anywhere

Endpoint Deployment
Supports real-time, batch and multi-model

Model Monitor
Automatically detect concept drift

Training Jobs
Makes it easy to run training jobs

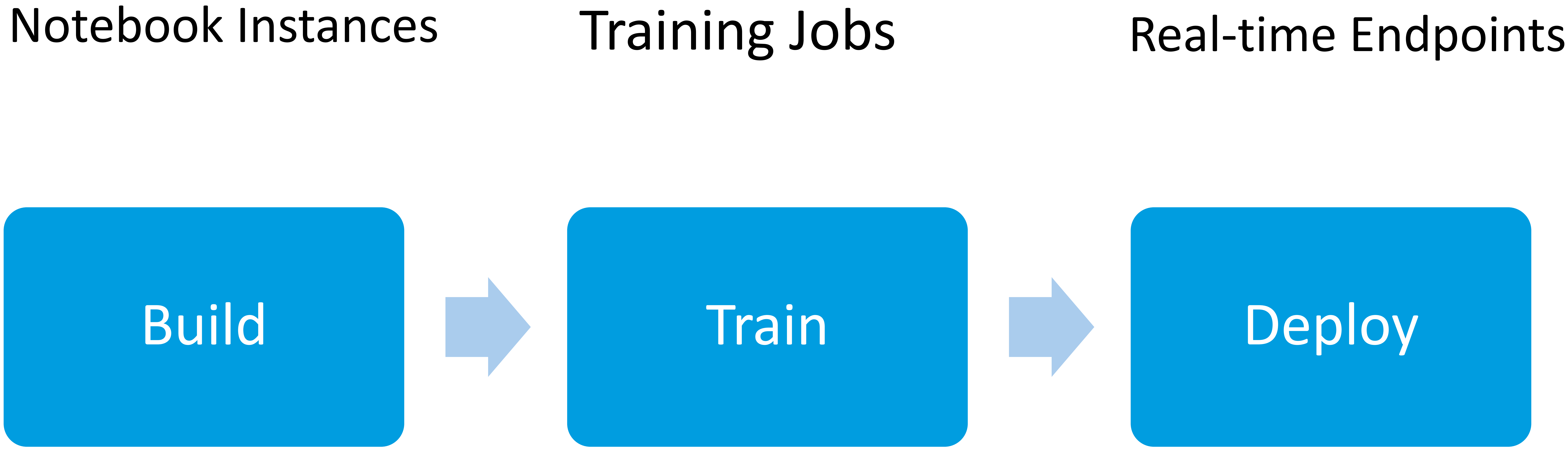
Automatic Model Tuning
One-click hyperparameter optimization

SageMaker Experiments
Capture, organize, and compare every step

SageMaker Debugger
Debug training runs

Managed Spot training
Reduce training cost by 90%

SageMaker Core Components



Notebook Instances

Training Jobs

Real-time Endpoints

Build

Train

Deploy

SageMaker Notebook Instances

Explore

Multi-level Donut Chart
Angular Gauge
Dot Plot
Pie Chart
Sociogram
Proportional Area Chart (Circle)
Waterfall Chart
Phase Diagram
Cycle Diagram
Population Pyramid
Boxplot
Three-dimensional Stream Graph
Semi Circle Donut Chart
Topographic Map
Radar Diagram

Built-in Kernels

Interact

Boto 3

Amazon SageMaker

SageMaker Python SDK

thank you

any questions?

#SimplifyBanking #LiveBetter

CAPITEC

capitecbank.co.za

Capitec Bank Limited is an authorised financial services provider (FSP 46669) and registered credit provider (NCRCP13) Capitec Bank Limited Reg. No.: 1980/003695/06

Amazon Neptune and Gremlin

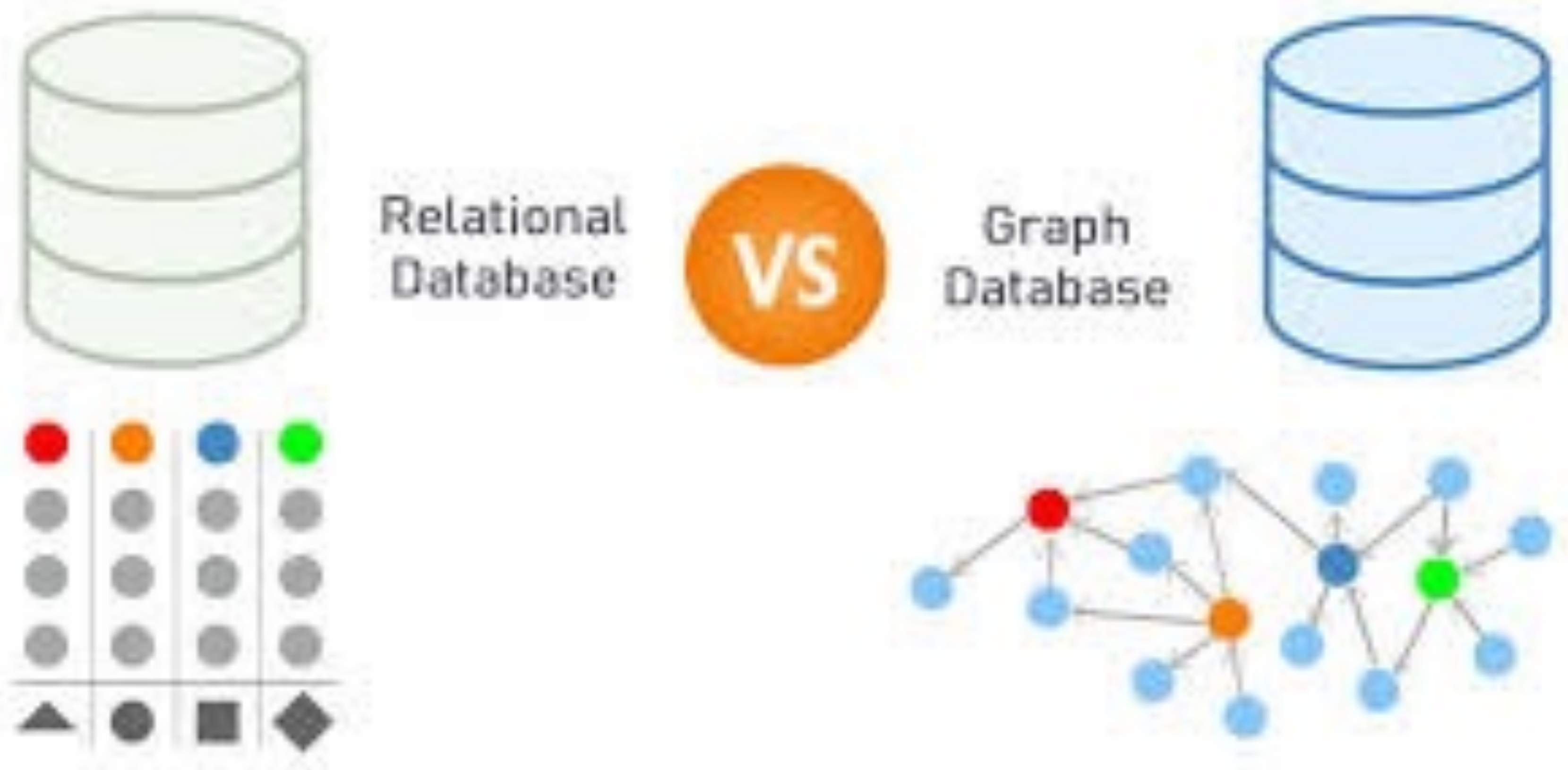
David Gouvias

Data Scientist

CAPITEC

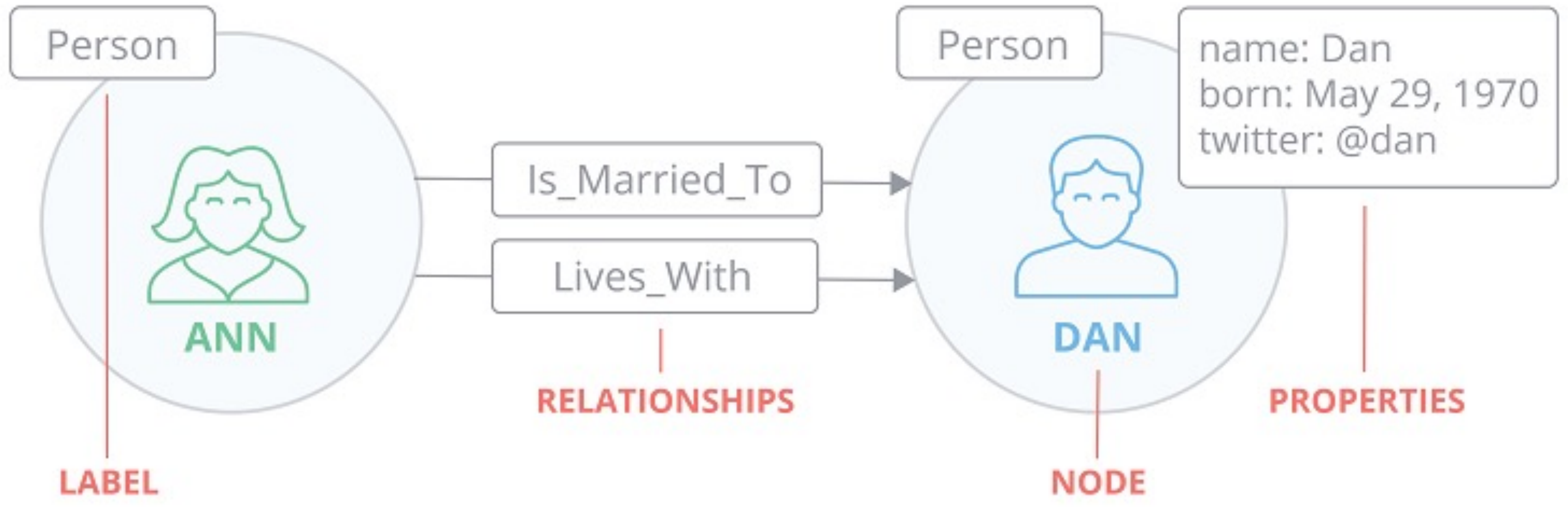
Amazon Neptune

- Purpose-build, high-performance graph database engine
- Optimized for storing billions of relationships
- Querying graphs with milliseconds latency
- Fully-managed (no hardware provisioning, software patching, setup)
- Supports graph model property graph and Resource Description Framework (RDF)
- Supports query languages Apache, TinkerPop, Gremlin and SPARQL

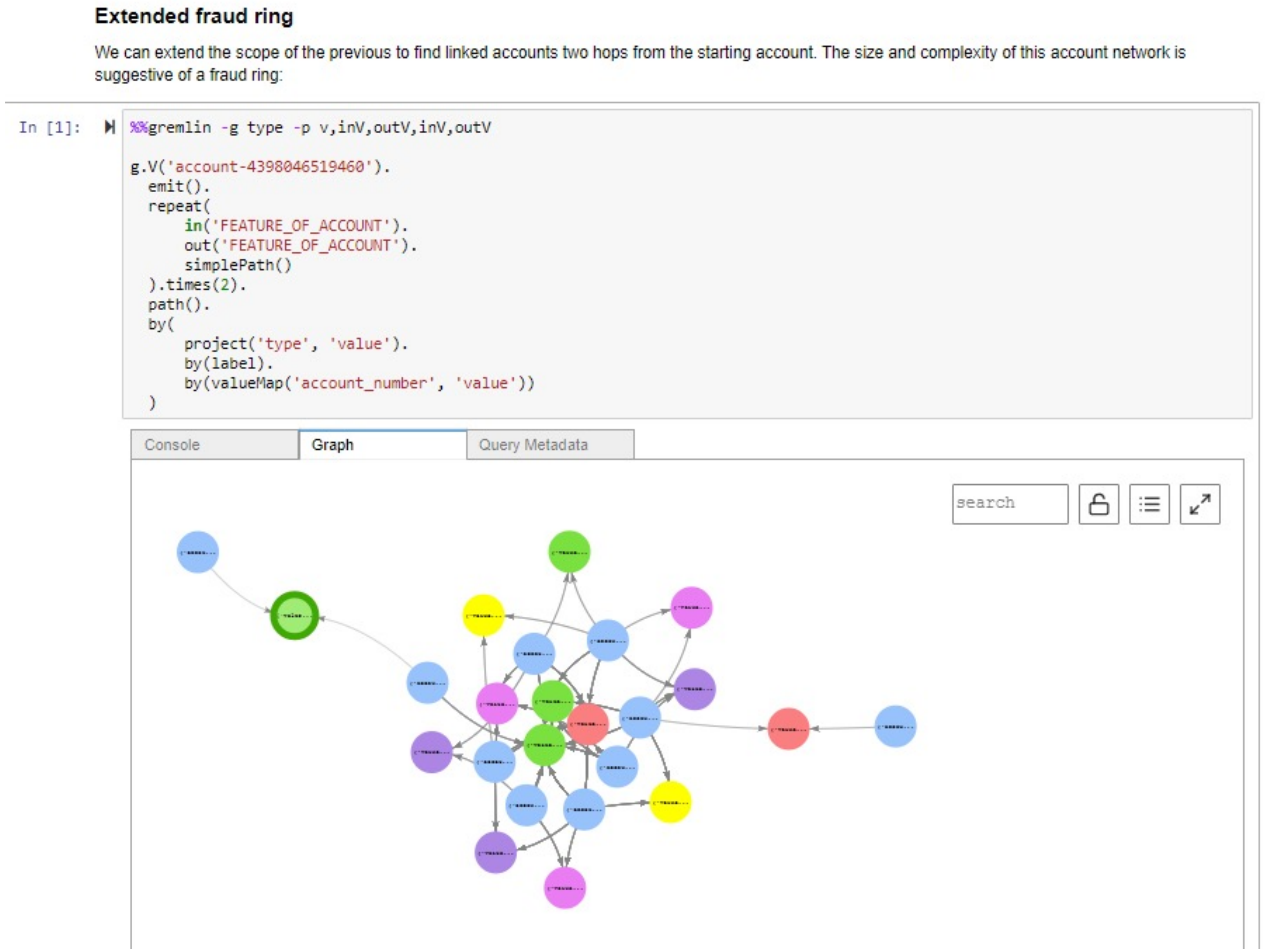


Gremlin - Graph Traversal Language

- Allows one to express complex queries that are not feasible and efficient in SQL.
- Many business problem solutions can be modeled as graph queries, including fraud typology detection.



Property Graph



Extended fraud ring

We can extend the scope of the previous to find linked accounts two hops from the starting account. The size and complexity of this account network is suggestive of a fraud ring:

```
In [1]:  %%gremlin -g type -p v,inV,outV,inV,outV

         g.V('account-4398046519460').
           emit().
           repeat(
               in('FEATURE_OF_ACCOUNT').
               out('FEATURE_OF_ACCOUNT').
               simplePath()
           ).times(2).
           path().
           by(
               project('type', 'value').
               by(label).
               by(valueMap('account_number', 'value'))
           )
```

thank you

any questions?

#SimplifyBanking #LiveBetter

CAPITEC

capitecbank.co.za

Capitec Bank Limited is an authorised financial services provider (FSP 46669) and registered credit provider (NCRCP13) Capitec Bank Limited Reg. No.: 1980/003695/06

Graph Algorithms

Ockert Janse Van Rensburg

Data Scientist

CAPITEC

Graph Algorithms

Extracting value from Graph Databases

Community Detection

Detects group clustering or partition options

Centrality (Importance)

Determines the importance of distinct nodes in the network

Similarity

Evaluates how alike nodes are

Heuristic Link Prediction

Estimates the likelihood of nodes forming a relationship

Pathfinding & Search

Finds optimal paths; evaluates route availability, quality

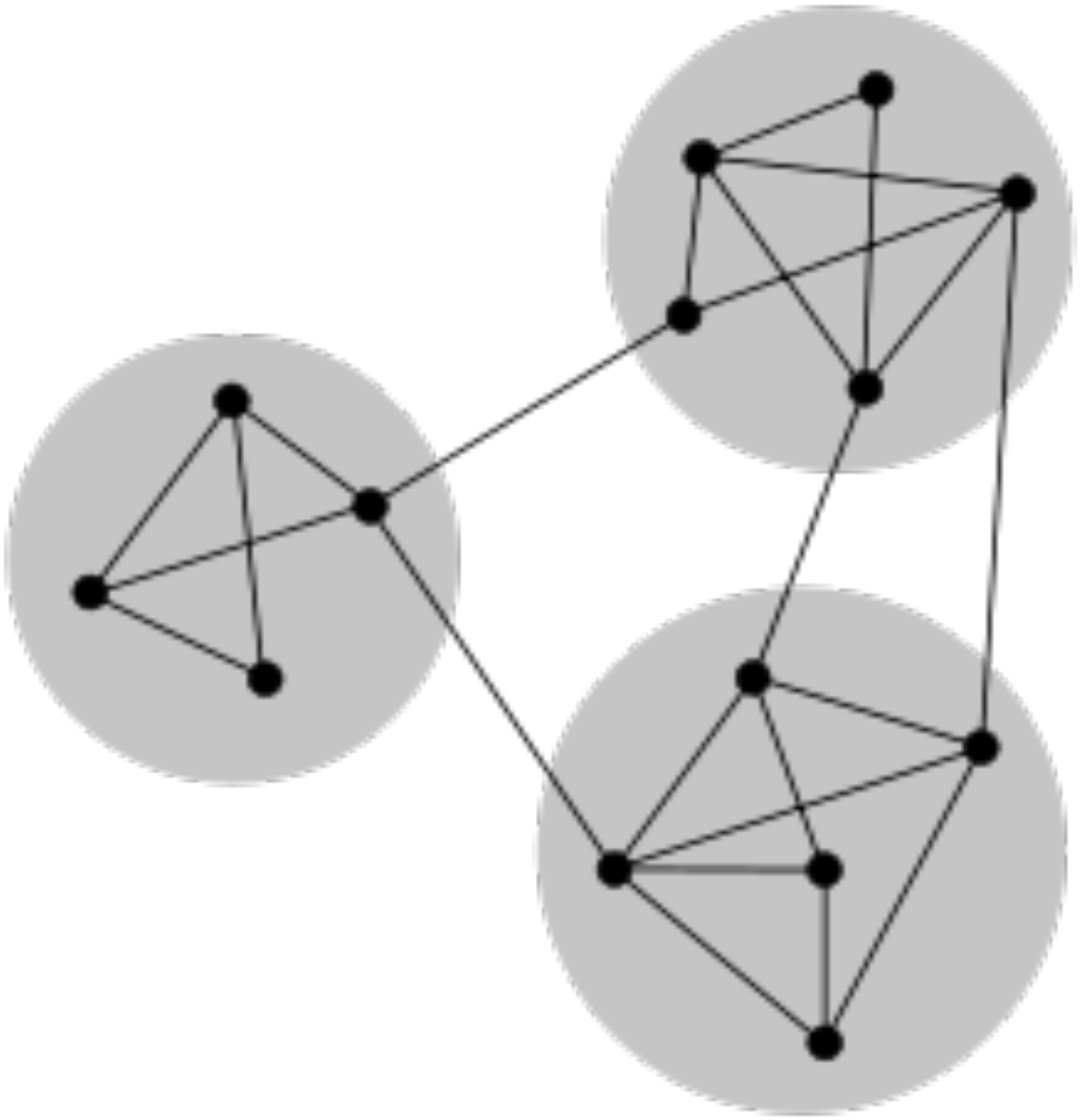
Node Embedding

Learns graph topology to reduce dimensionality for machine learning

Community Detection

Finding meaningful groups in complex phenomena

- What is a community?
  - **group, cluster, cohesive subgroup, module**
- Break up the network into **modular groups** where the edges within group are of higher density, than those of the other groups
- Multiple types of community detection algorithms (overlapping vs non-overlapping)
- The **Louvain method** commonly used due to its scaling properties
- More information on installing this method will be made available in the info pack to be distributed



Non-overlapping communities. Communities represented by the circles.

Graph Algorithms

Dalubuhle Mbune

Data Scientist

CAPITEC

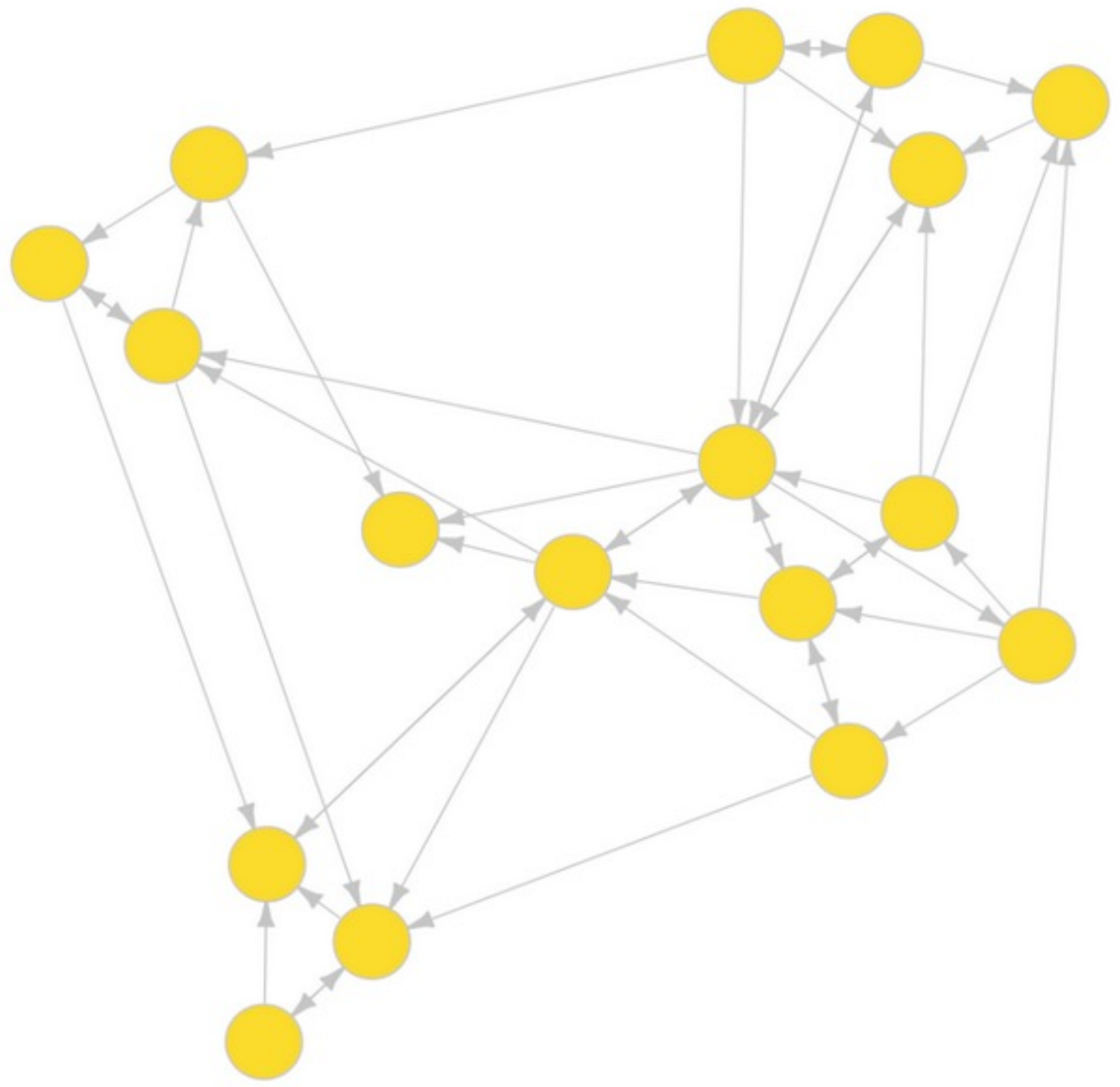
Graph Algorithms

Similarity Algorithms

- The similarity measure is a way of measuring how nodes are related or close to each other.
- Calculations are performed on vector representations of objects. Each object must first be converted to a numeric vector.

- Similarity/distance is calculated between a single pair of nodes at a time.

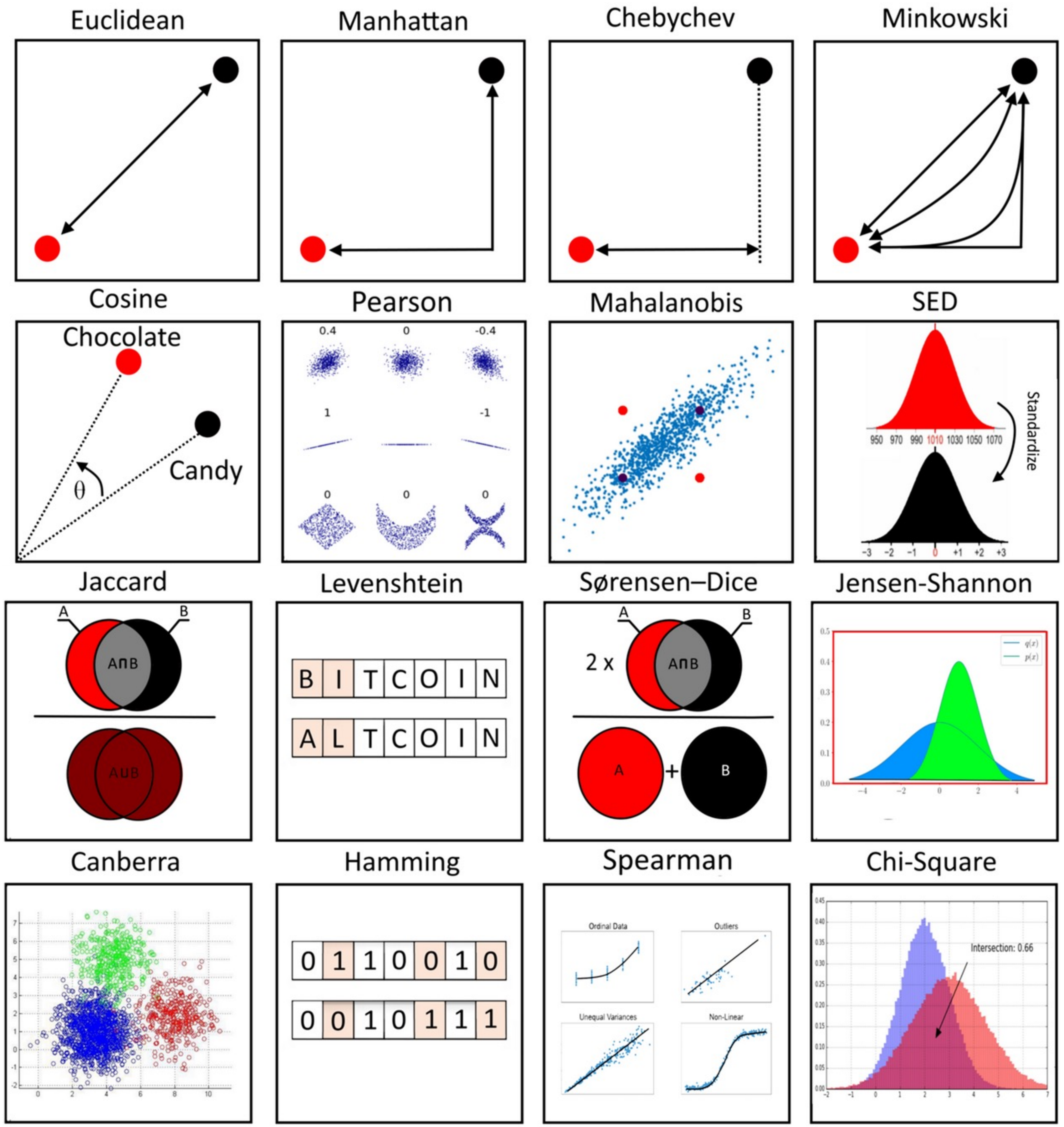
- There are numerous similarity algorithms
- Regardless of the algorithm, feature selection will have a huge impact on your results.



Similarity Algorithms

- Distance measures are the fundamental principle for classification

- The choice of distance measure plays a crucial role in the similarity algorithm's performance

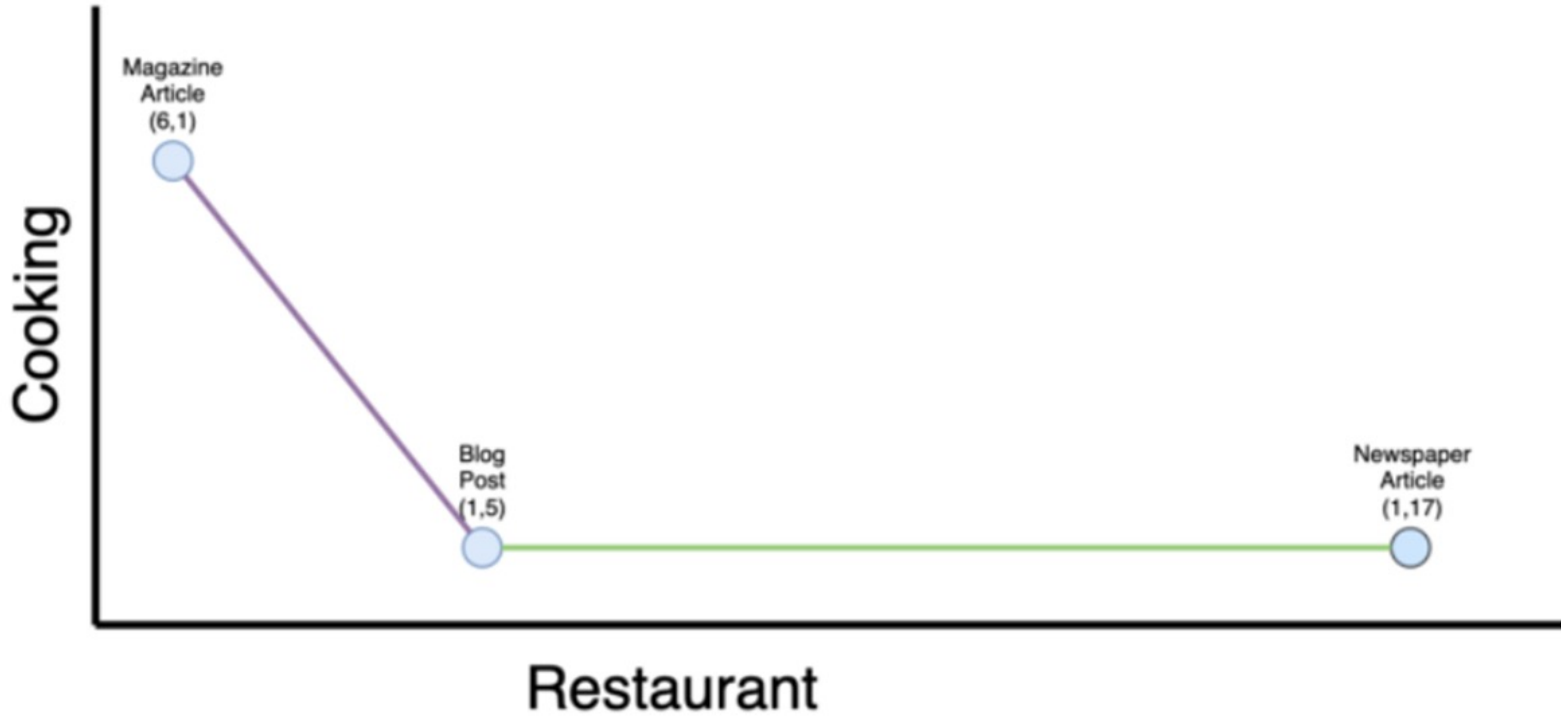


EXAMPLE

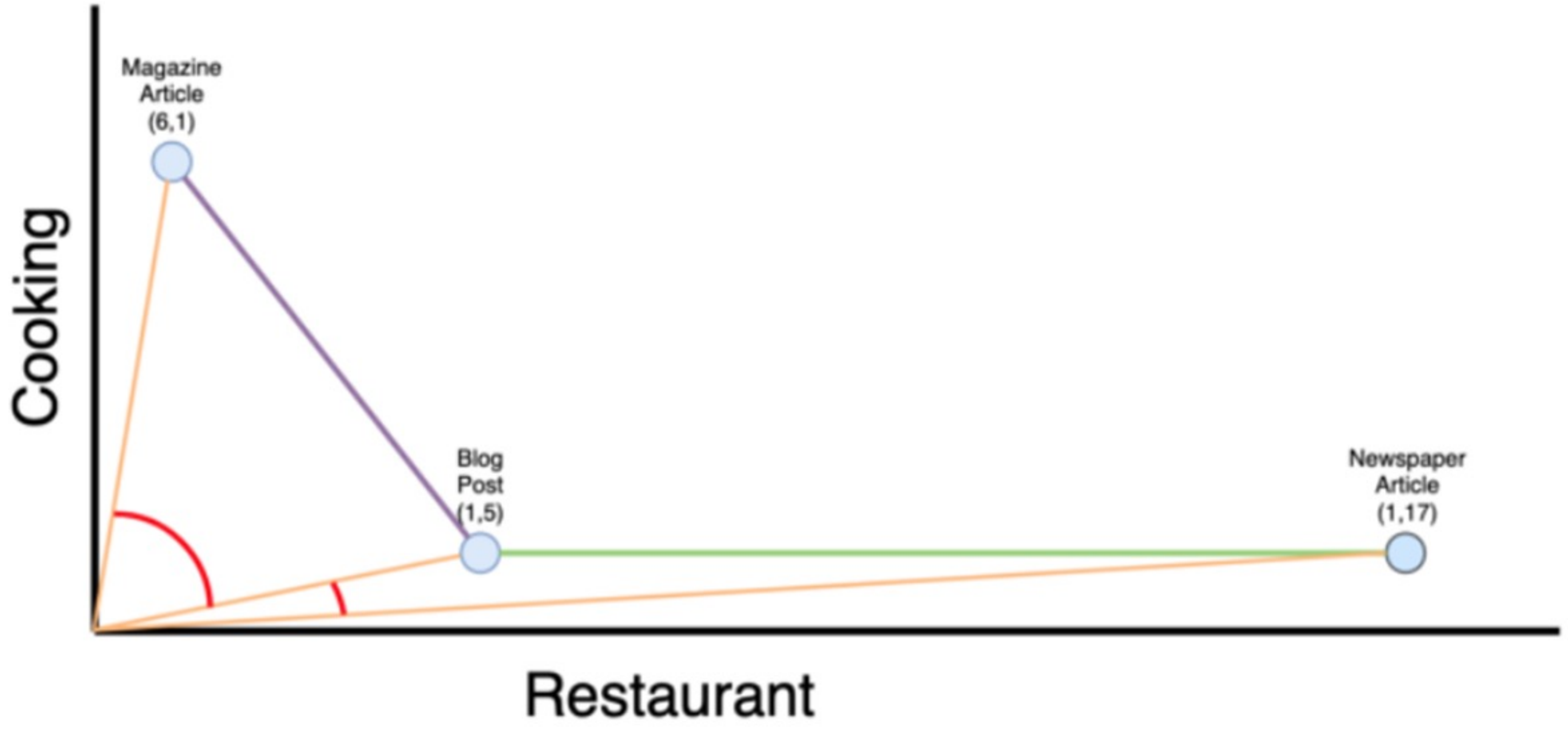
Euclidean and Cosine Similarity for Document Comparison

- Suppose we want to compare how frequent the words ‘Restaurant’ and ‘Cooking’ (Features) appear on a Blog Post, Newspaper Article, and Margazine Article.

Euclidean Similarity



Cosine Similarity



- In the above Example, we compare 3 documents based on how many times they contain the words “cooking” and “restaurant”.

- Euclidean distance tells us the blog and magazine are more similar than the blog and newspaper. But that’s misleading.

- The blog and newspaper could have similar content but are distant in a Euclidean sense because the newspaper is longer and contains more words.

- In reality, they both mention “restaurant” more than “cooking” and are probably more similar to each other than not. Cosine similarity doesn’t fall into this trap.

thank you

any questions?

#SimplifyBanking #LiveBetter

CAPITEC

capitecbank.co.za

Capitec Bank Limited is an authorised financial services provider (FSP 46669) and registered credit provider (NCRCP13) Capitec Bank Limited Reg. No.: 1980/003695/06

Hackathon Challenge

David Gouvias

Data Scientist

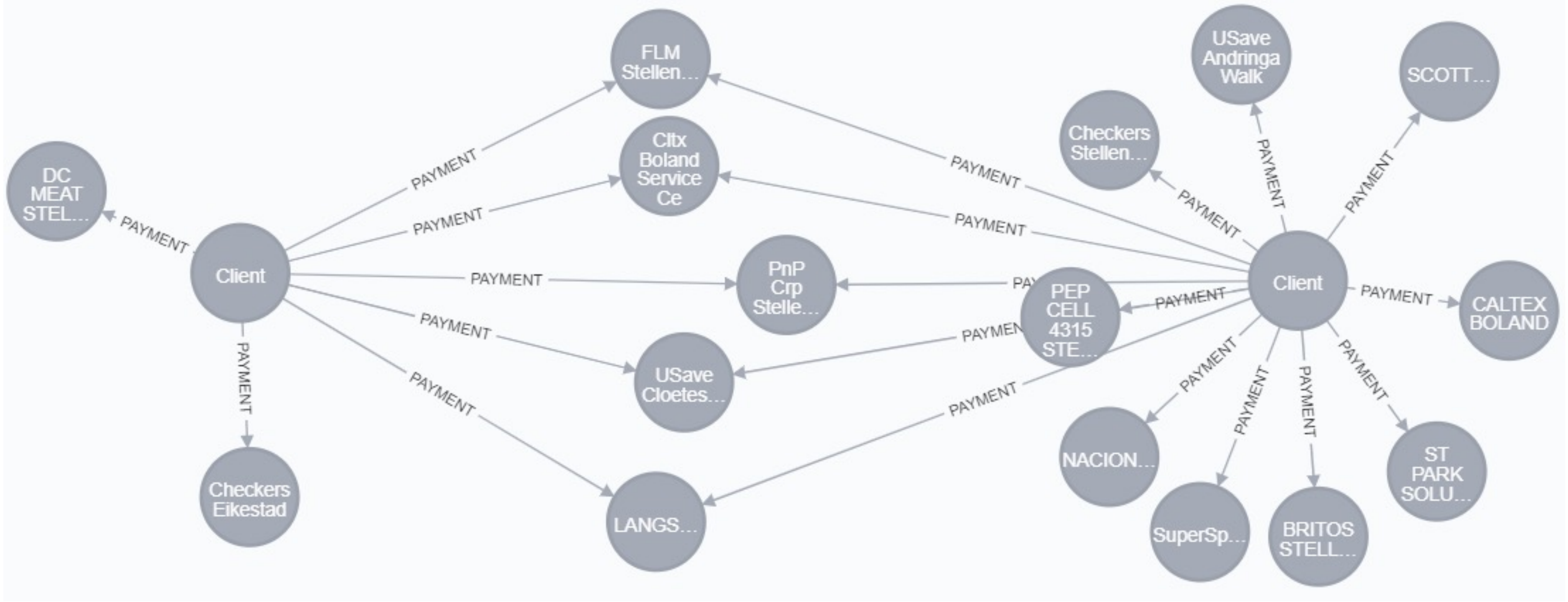
CAPITEC

Hackathon Challenge

Client – Merchant network

Your challenge is to use our AWS Neptune Graph Database and apply data science algorithms or graph queries to enrich the dataset through :

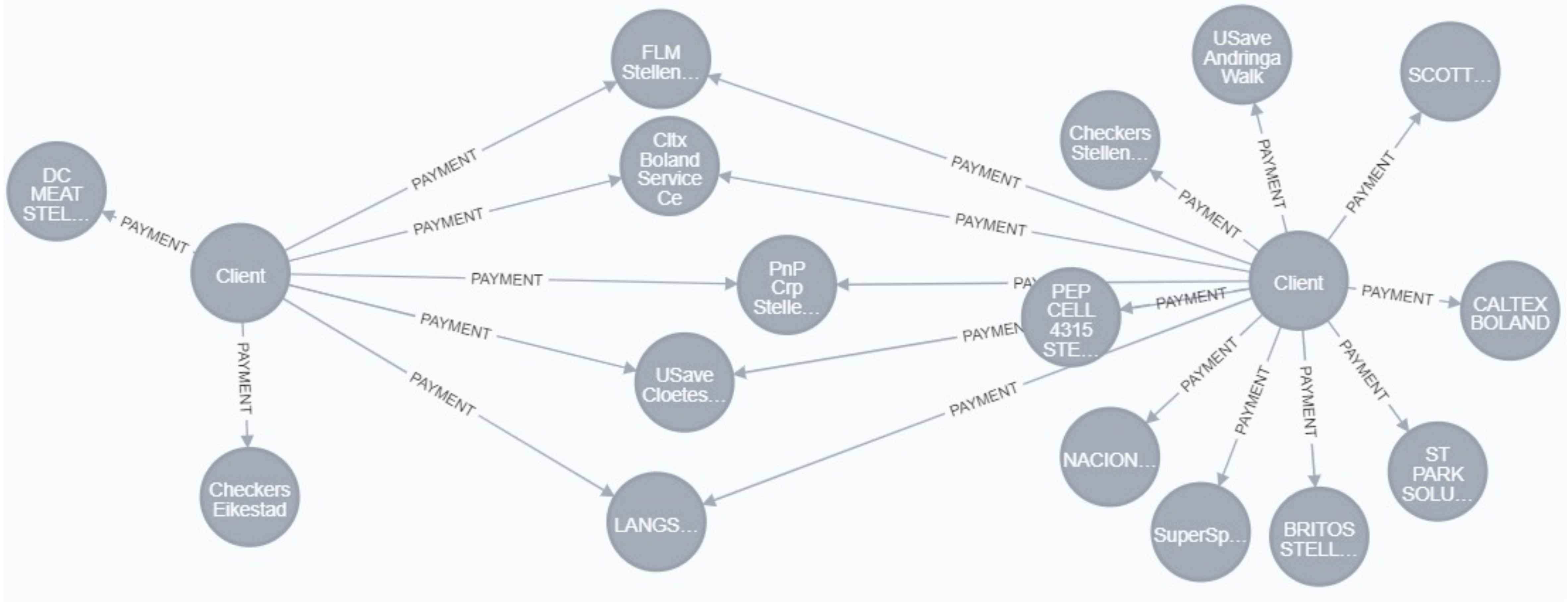
- Identifying community clusters (people with same shopping and movement patterns),
- Identifying Commuters, travellers, contract workers or traveling salesmen.
- Telling a story at scale of the client communities.
- Identifying fraudulent behaviour.
- Define your own problem you wish to solve.



Hackathon Challenge

Business Ideas

- Funeral cover recommender : Recommend which clients are likely to take out a funeral policy.
- Store Finder : Recommend a list of stores in a particular category for a customer need, e.g. Pharmacy.
- Merchant Assisted Marketing: Find a list of new customers that are likely to shop at a particular merchant.



thank you

any questions?

#SimplifyBanking #LiveBetter

CAPITEC

capitecbank.co.za

Data Model

Client Information

Client_UID	Age_Band
0xE3E097DC79D8161B2A2448F6C0930A8B081CD013	Band 10: 46 To 50
0xBE461A0CD1FDA052A69C3FD94F8CF5F6F86AFA34	Band 9: 41 To 45
0xE2154FEA5DA2DD0D1732FF30931723C2973003A0	Band 9: 41 To 45
0x4A0E88CF529FBBDC2C0A995BBE88A0A86212ED8D	Band 11: 51 To 55
0xCFA2ED2AAC6D61F44CA9CBA73E1E8946B7CD7D22	Band 11: 51 To 55

Merchant Information

Merchant_UID	Merchant_Name	Merchant_Type	Merchant_Type_Desc	Merchant_Category
0x8E6A682F75803D8F8090F99FBC303455489109D3	Clicks Somerset Mall CPT ZA	5912	Drug Stores and Pharmacies	All Other Merchants/U.S. Post Exchange OR Card...
0x93003523E22D055C5CC080807FCEB0FA7E0D67C5	Game Cape Gate Cape Town ZA	5311	Department Stores	All Other Merchants/U.S. Post Exchange OR Card...
0x2711A1A5DC6A4783B17D60E7444FA2AF2386305A	LINKS SERVICE STATION SOMERSET WEST ZA	5541	Service Stations (with or without Ancillary Se...	All Other Merchants/U.S. Post Exchange OR Card...
0xEEE999A1F8C76D2541ABFCB524709066D5B585AB	STEERS - CANAL WALK CENTURY CITY ZA	5814	Fast Food Restaurants	Restaurant
0xF40F5991406F281D684EEF2473692750573D1A95	SCOTTYS MIDAS STELLENBOSCH ZA	5511	Automobile and Truck Dealers: Sales, Service, ...	All Other Merchants/U.S. Post Exchange OR Card...

Data Model

Payments

Src_Client	Trg_Client	Tran_Date_Key	Amt_Trans	Num_Trans
0xA08170480197FFB2CCCA2671C63D7F9DD440DACD	0x0D80273C48EA052178805C8E0BAF5D99E2055A0F	11666	2650	1
0x97433A955B75A559C81E84E3BA9D1C3E75F6A1A7	0x0D01084F4C11AE10513480F1CF60271B8F1048CE	11580	145	1
0x5538DE60D60A00EC0A5CE8FC70D9431D3AB171D2	0x7BBAC91F5D41B0FDF9B3AE36FB417690C2024C63	11643	4400	2
0x0B4A6DC422CED9A7AF2B07867B91EE2B572CA451	0x867B6E1D45F7DCCE3B08AB67F85F298CB3F287E5	11638	500	2
0xC65CB7AD4C7F0C3560B1A1C953CB7664746DCC06	0x88A70DBF116D4DDF50BFB9962FEB2041C3A57BBA	11646	500	1

Purchases

Src_Client	Trg_Merchant	Tran_Date_Key	Amt_Trans	Num_Trans
0xDCD429E847183D910DBFBCB5A37214C2FAF4ACD5	0x5979712AC3DB16655C062AE7DEEB98A12106D4BB	11621	74	1
0x5D122FAFEDDCEFC8C4DBD9995EE058E0731BF712	0x1211AD3B70DC1FF4180AA6F46D3F72C0EBF9655E	11596	260	1
0x397A2F5AFE5F8A28D6F12F5B1757AC14E7367046	0x1211AD3B70DC1FF4180AA6F46D3F72C0EBF9655E	11637	342	1
0x39F8191CFA084AF00F9B530D900F9F34E3846904	0x1211AD3B70DC1FF4180AA6F46D3F72C0EBF9655E	11580	230	1
0x5A4001305F3A5A121A108A146AB96A67B5BC0D05	0x5979712AC3DB16655C062AE7DEEB98A12106D4BB	11627	696.5	1

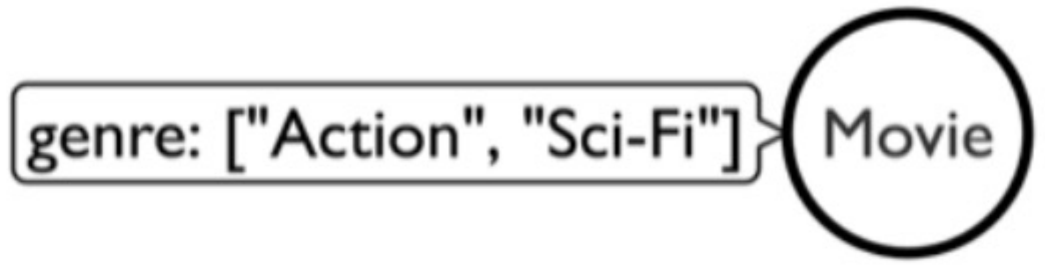
Data Model

Funeral Policy

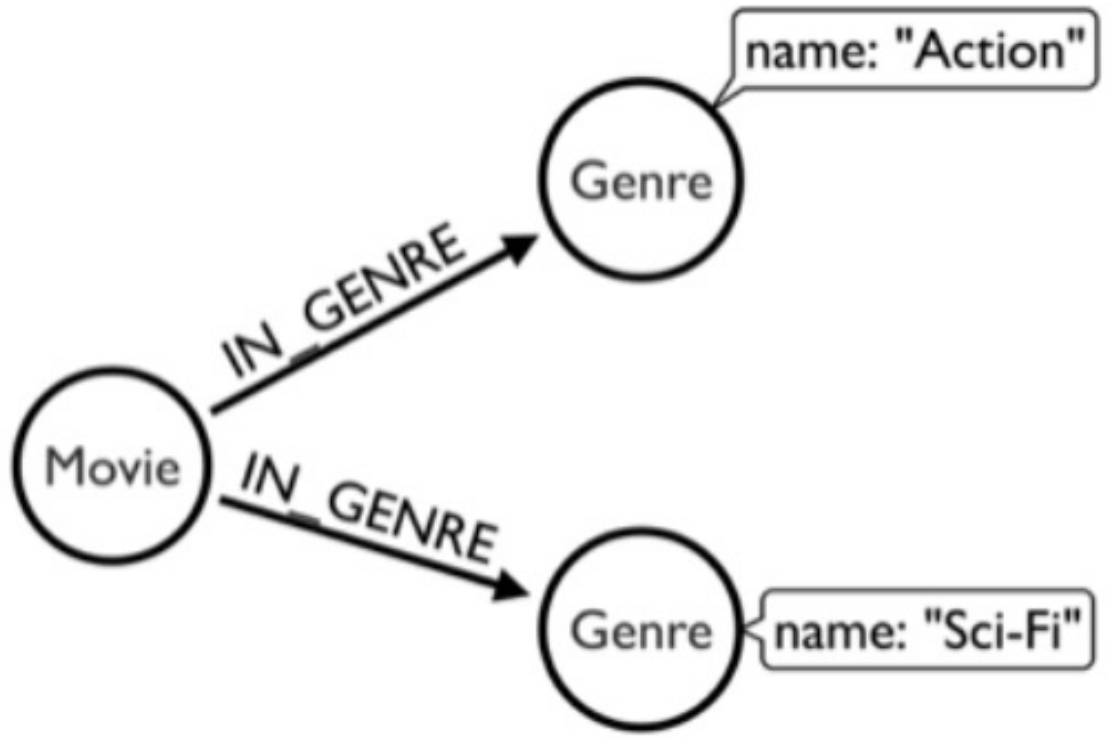
FC_Policy_UID	Policy_Holder_UID	Policy_Holder_Client_UID	Role_Type	Relationship	Role_Holder_UID	Role_Holder_Client_UID
0xBC2507F01F15E53D585CC349DDEBED803589E7A4	0xDE47A547194774E381F02999A3413EBEF771B84A6CDE...	NaN	Life Assured	Child	0x200ECFCEF62452CDFEEE703812F1CA1132CDBEFA5DBB...	0xD068C9D2BEC32C2C8D261211683C61F4843B8533
0x5CE150E313E4C52E250E199699E6D04CF84CE917	0x956390B3CD7558291775D4C32E0C820D07F11A0B1E23...	0x5641A37860F7B156FACEB5EE50A33D9538903F9C	Life Assured	Self	0x956390B3CD7558291775D4C32E0C820D07F11A0B1E23...	0x5641A37860F7B156FACEB5EE50A33D9538903F9C
0x68BDBBDBEFBE2744B9DE06E3C612C9C6FF8B2F5C	0xD67CB8DA6302E2B6695B06B2C5E23F7C0A7121B88797...	0x8D4F80DF0D37819CDE3E3D2BB9982D111EBAC97C	Life Assured	Self	0xD67CB8DA6302E2B6695B06B2C5E23F7C0A7121B88797...	0x8D4F80DF0D37819CDE3E3D2BB9982D111EBAC97C
0xAE0EFC73ADD762BF85AC79D3ADDC638F50EE87D8	0xB6F2555D8ED822AC1C2905940DFC21B611211C8A7365...	0xCBBE069D36EE6C3DA92B9E11C2AE6447FF6F359D	Life Assured	Self	0xB6F2555D8ED822AC1C2905940DFC21B611211C8A7365...	0xCBBE069D36EE6C3DA92B9E11C2AE6447FF6F359D
0x81E4AC72D5604FA9C06DEC4CE26BAA58F9FFB911	0xE754703894507606I3FF448403121227904771I1CE1E...	0xD5C6972618D4D3396A186726BE36049C39600298	Life Assured	Self	0xE7547038945076061 3FF4484031212279047711 1CE1E...	0xD5C6972618D4D3396A186726BE36049C39600298

Graph Database Design Guidelines

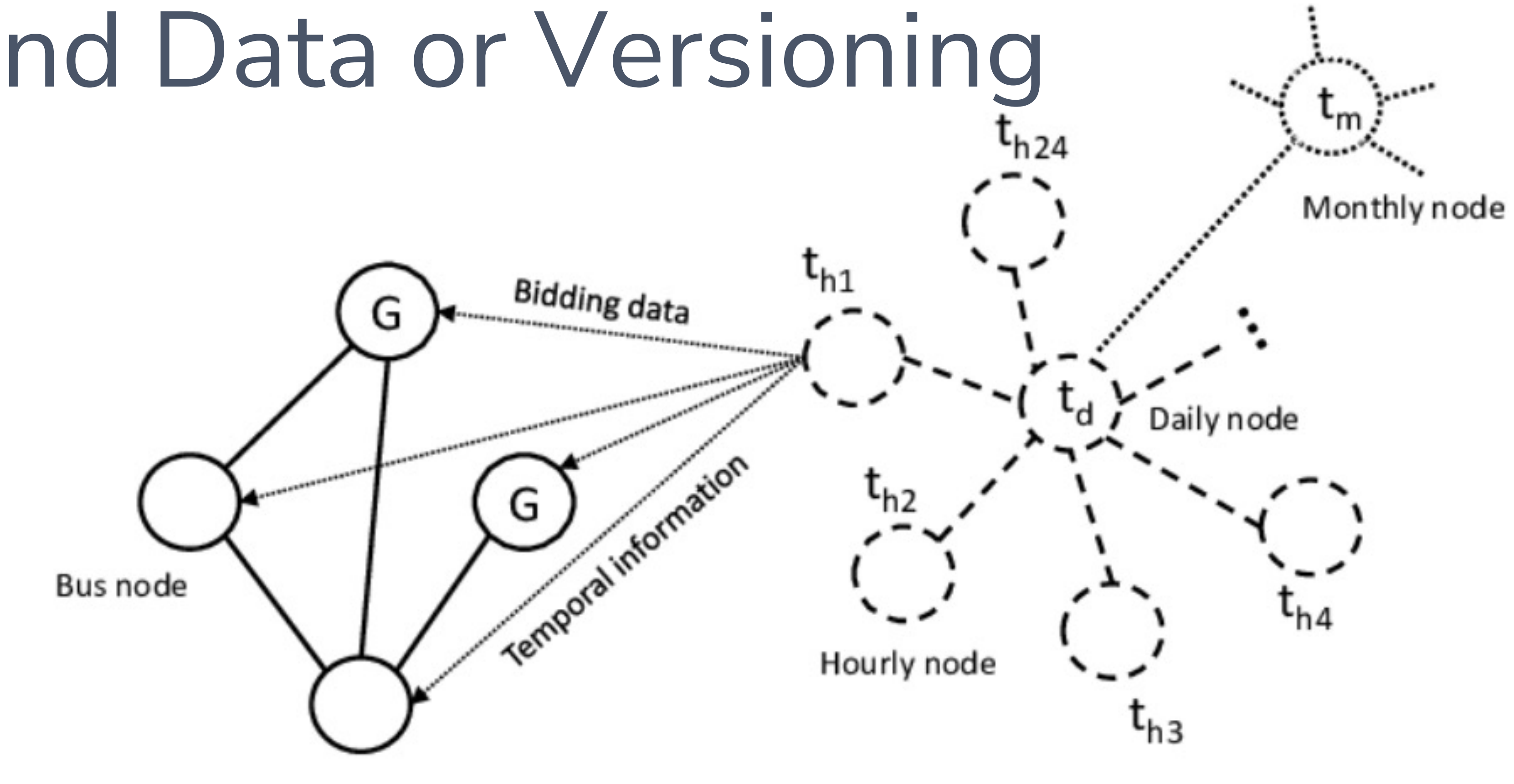
- Property vs Relationship



VS

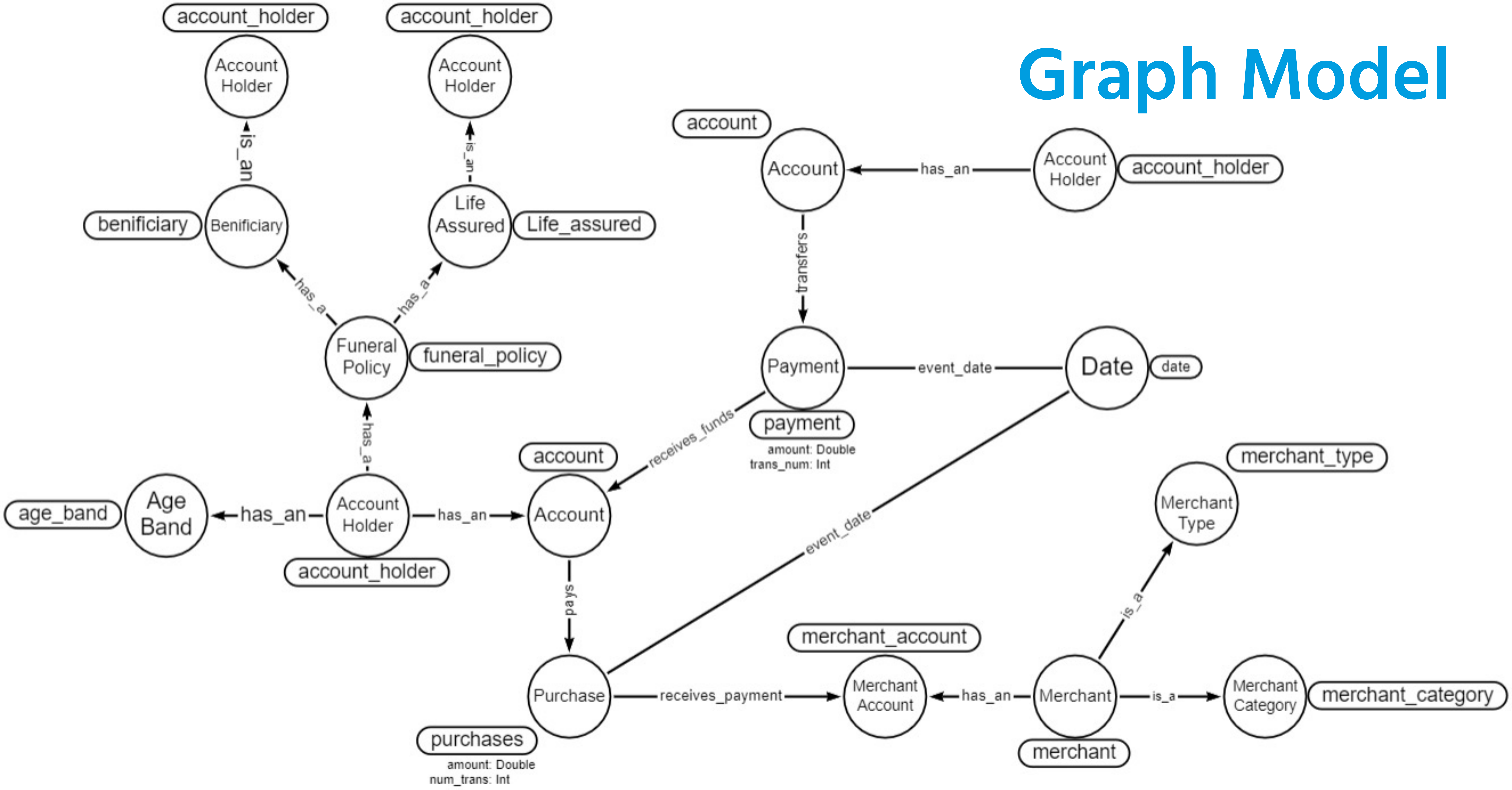


- Time-bound Data or Versioning



Time tree structure in graph database

Graph Model


account_holder
Account Holder
is_an
account_holder
Account Holder
is_an
benificiary
Benificiary
Life Assured
Life_assured
has_a
has_a
Funeral Policy
funeral_policy
has_a
age_band
Age Band
has_an
Account Holder
account_holder
has_an
account
Account
pays
Purchase
purchases
amount: Double
num_trans: Int
account
Account
has_an
Account Holder
account_holder
transfers
Payment
payment
amount: Double
trans_num: Int
receives_funds
event_date
Date
date
event_date
receives_payment
merchant_account
Merchant Account
has_an
Merchant
merchant
is_a
Merchant Category
merchant_category
is_a
Merchant Type
merchant_type

thank you

any questions?

#SimplifyBanking #LiveBetter

CAPITEC

capitecbank.co.za

Capitec Bank Limited is an authorised financial services provider (FSP 46669) and registered credit provider (NCRCP13) Capitec Bank Limited Reg. No.: 1980/003695/06

Next steps

What to expect next week.

- Info pack, including login details.
- Judges, PW Janse van Rensburg (Technical Value) and Chane Dewar (Business Value)
- Prizes

thank you

any questions?

#SimplifyBanking #LiveBetter

CAPITEC

capitecbank.co.za

Capitec Bank Limited is an authorised financial services provider (FSP 46669) and registered credit provider (NCRCP13) Capitec Bank Limited Reg. No.: 1980/003695/06

References

- https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg#/media/File:Present_state_of_the_Seven_Bridges_of_K%C3%B6nigsberg.png
- https://aws.amazon.com/nosql/graph/
- https://neo4j.com/
- Machine Learning with Graphs: https://www.youtube.com/watch?v=aBHC6xzx9YI
- https://eng.uber.com/uber-eats-graph-learning/
- https://www.nature.com/articles/d41586-020-03348-4
- https://www.theverge.com/2020/9/3/21419632/how-google-maps-predicts-traffic-eta-ai-machine-learning-deepmind
- https://www.capitecbank.co.za/