



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Why Use Tables and Graphs for Knowledge Discovery

JOHN FEO

*NORTHWEST INSTITUTE FOR ADVANCED COMPUTING
PACIFIC NORTHWEST NATIONAL LABORATORY*

HPC USER FORUM
SEPTEMBER, 2016

Why graphs? Good for representing unstructured data

Mary called her sister Sally to discuss buying her 6-year daughter a pony for Christmas.

- 1) Mary called Sally
- 2) Mary has a sister named Sally
- 3) Sally has a sister named Mary
- 4) Either Mary or Sally has a daughter
- 5) The daughter is 6 years old
- 6) Mary wants to buy a pony



Sally rented Joe's condo in Hawaii for a two week vacation. She paid \$1200 rent.

- 1) Sally traveled to Hawaii
- 2) Sally vacationed in Hawaii
- 3) Joe owns a condo
- 4) Joe's condo is in Hawaii
- 5) Sally rented Joe's condo
- 6) Joe rented his condo for \$600 per week



NAME	SIBLING	CHILD	AGE	CALLED	FUTURE PURCHASES
Mary	Sally	?		Sally	Pony
Sally	Mary	?			
?			6		

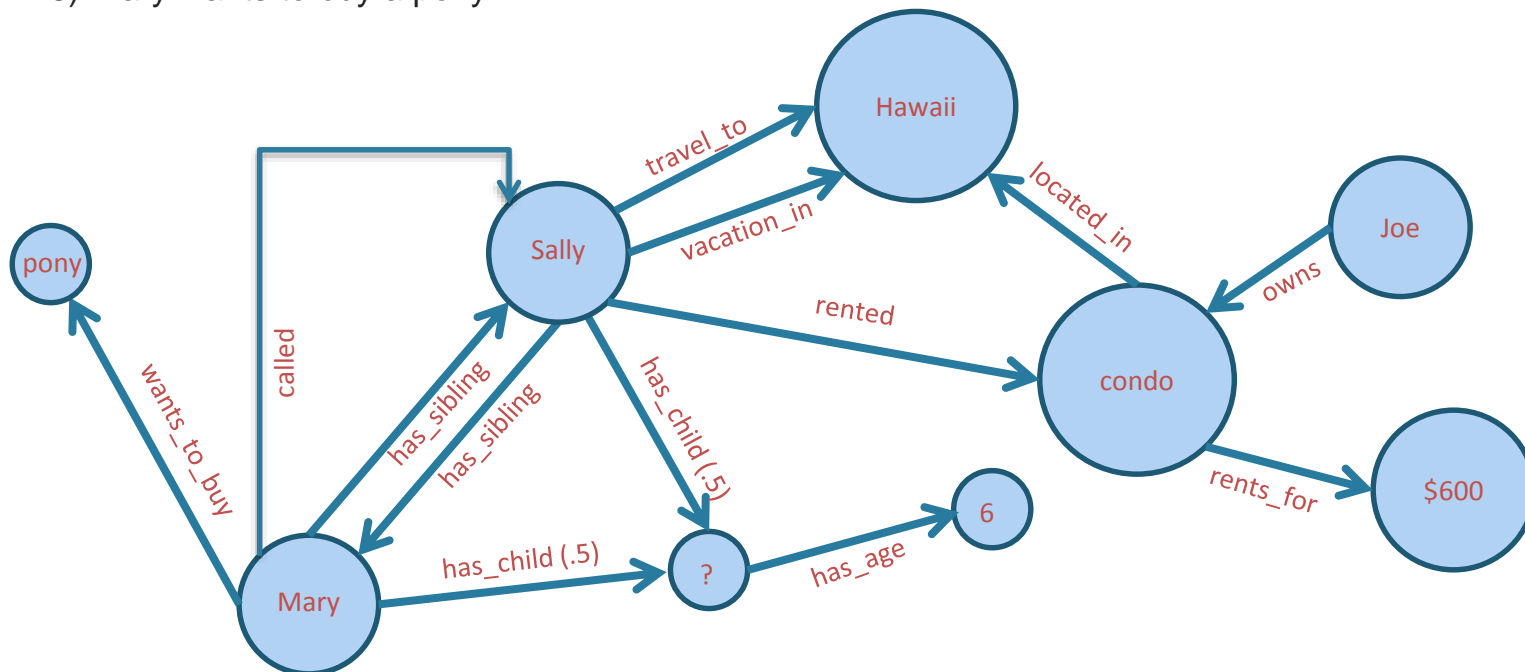
Why graphs? Good for representing unstructured data

Mary called her sister Sally to discuss buying her 6-year daughter a pony for Christmas.

- 1) Mary called Sally
- 2) Mary has a sister named Sally
- 3) Sally has a sister named Mary
- 4) Either Mary or Sally has a daughter
- 5) The daughter is 6 years old
- 6) Mary wants to buy a pony

Sally rented Joe's condo in Hawaii for a two week vacation. She paid \$1200 rent.

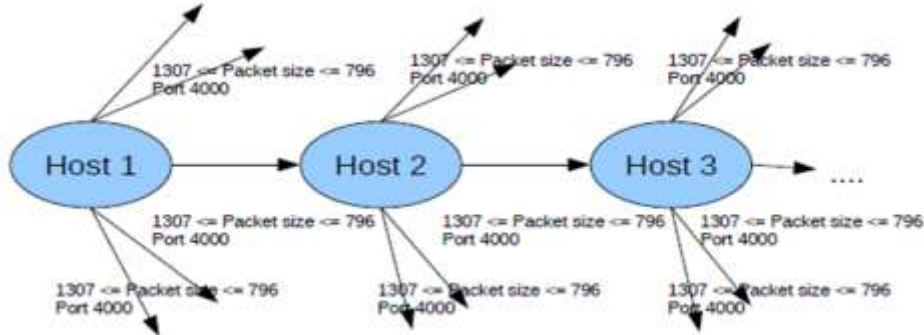
- 1) Sally traveled to Hawaii
- 2) Sally vacationed in Hawaii
- 3) Joe owns a condo
- 4) Joe's condo is in Hawaii
- 5) Sally rented Joe's condo
- 6) Joe rented his condo for \$600 per week



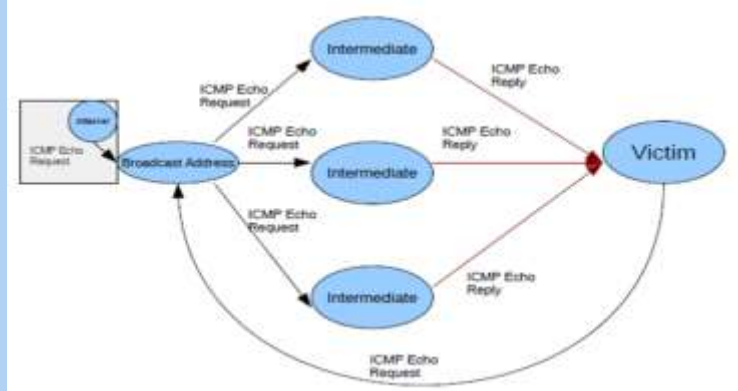
Why graphs? Good for finding patterns



Witty Worm

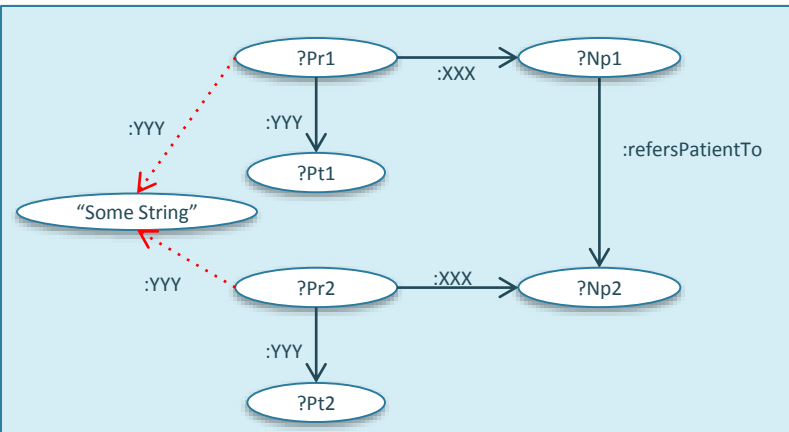
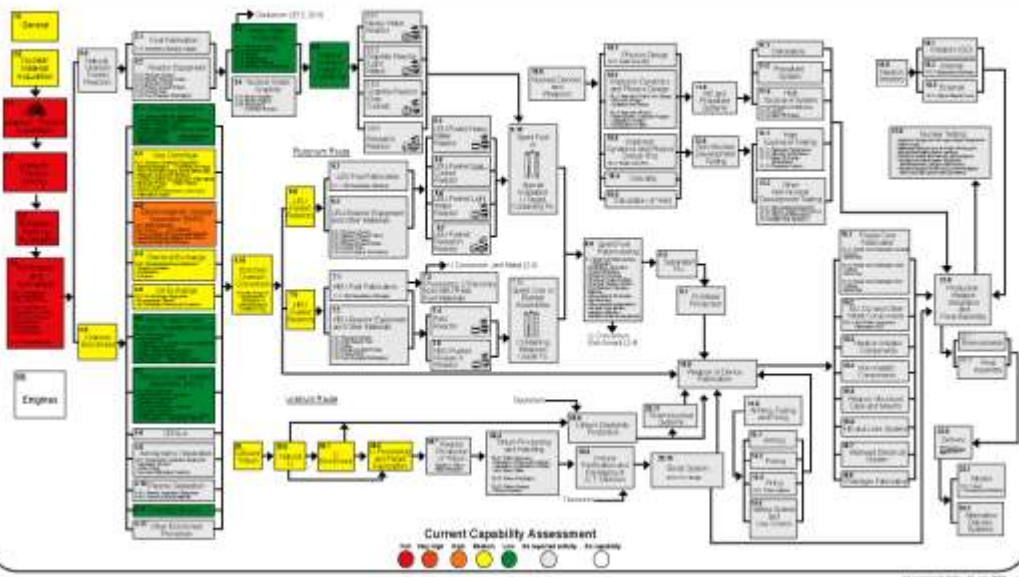


Distributed DoS Smurf Attack



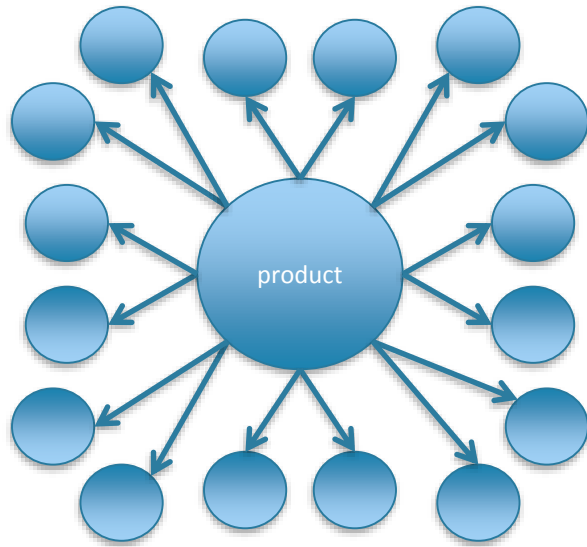
HYPOTHETICAL PROLIFERANT COUNTRY

Nuclear Fuel Cycle and Weapons Development Process



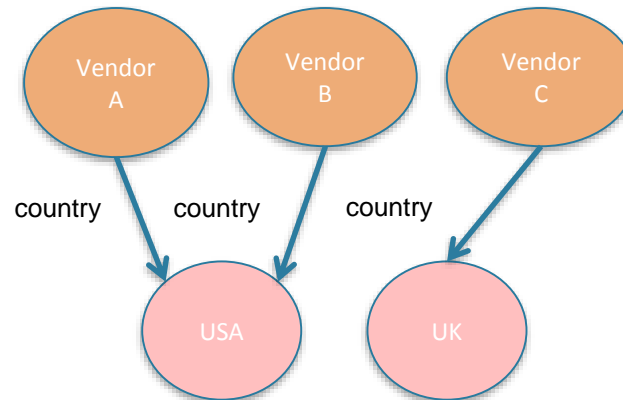
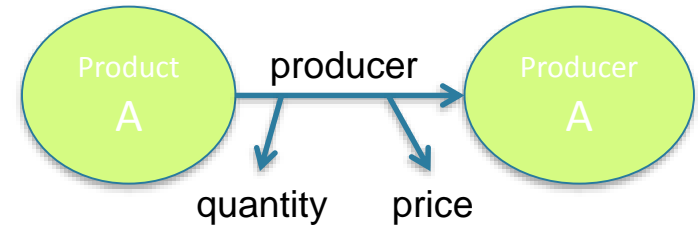
Why not graphs?

id	srcaddr	dstaddr	protocol	srcport	dstport	start	end
1	62.175.55.106	131.215.122.25	6	25	5641	5682	9084
2	67.72.170.48	131.215.122.25	6	57947	3724	17364	24602
3	67.72.170.48	131.215.122.25	6	60373	3724	28494	31684
4	67.72.170.48	179.137.146.48	6	80	64845	41239	43766
5	67.72.170.48	131.215.122.25	17	58020	9050	45516	46136



Starification

Edge attributes

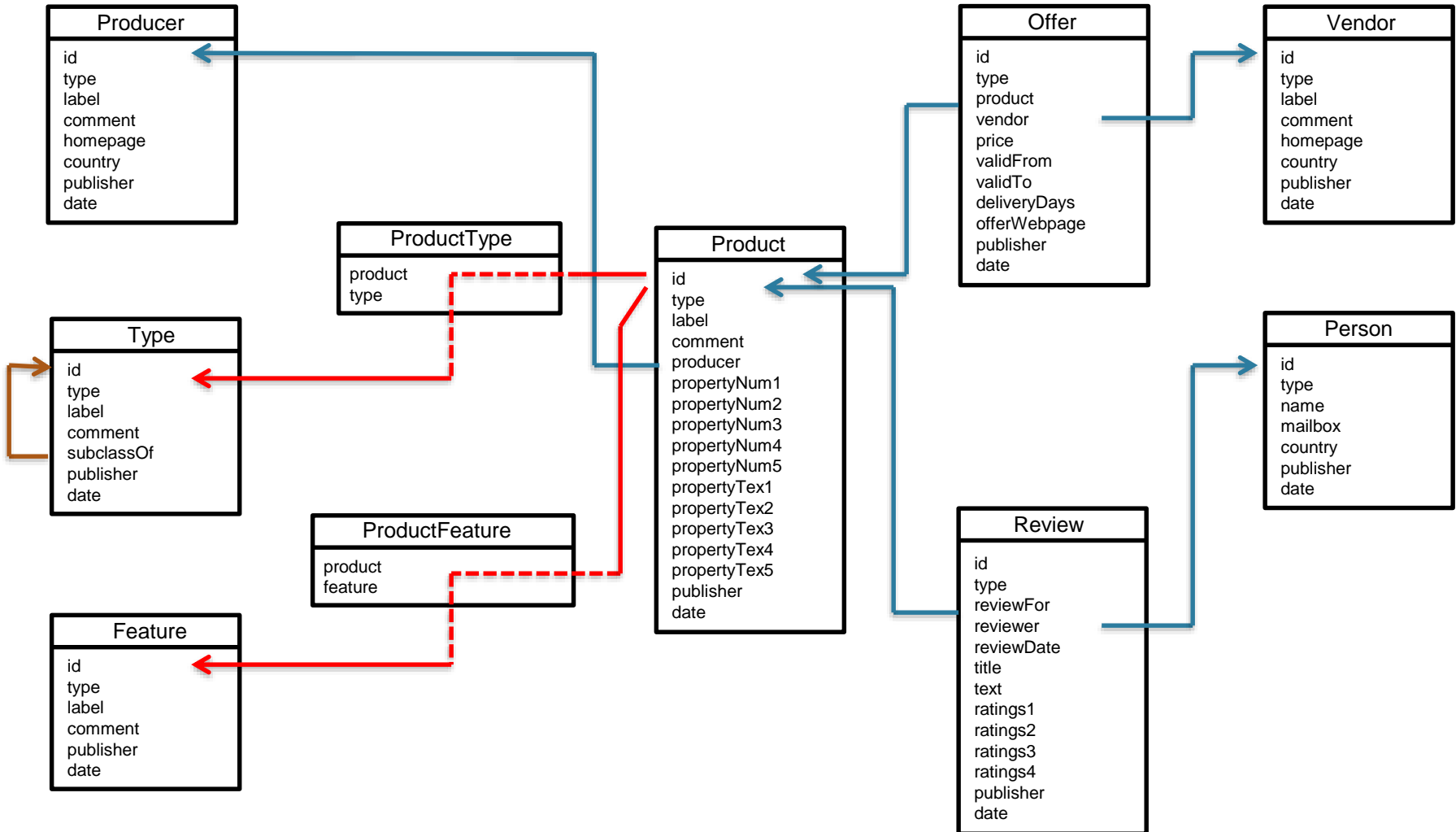


Inefficient selects

So, may be we need ...

- ▶ Table and graph data structures
- ▶ Table operations and graph methods
- ▶ Relational and graph/knowledge databases
- ▶ New query languages? **No**, just a few extensions
- ▶ Disruptive platforms? **No**, just integrate new features
- ▶ New workflows? **No**, just new options

Some attributes are edges



GraQL – adding a graph view to SQL

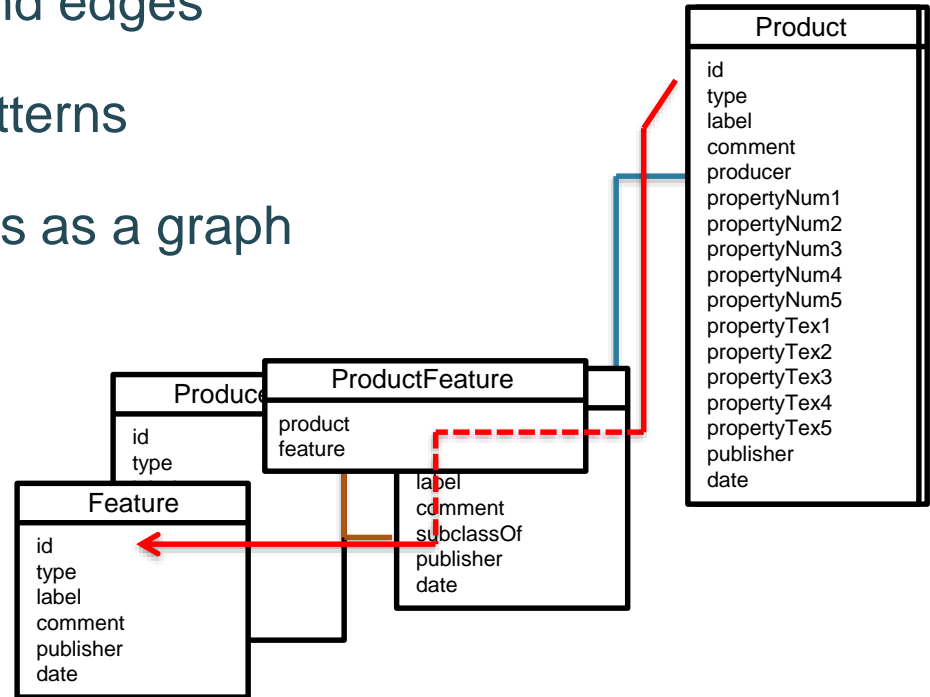
- ▶ Extensions to declare vertices and edges
- ▶ Extensions to describe graph patterns
- ▶ Extensions to return query results as a graph

```
create vertex ProductVtx(id)
from table Product

create vertex ProducerVtx(id)
from table Producer

create vertex TypeVtx(id)
from table Type

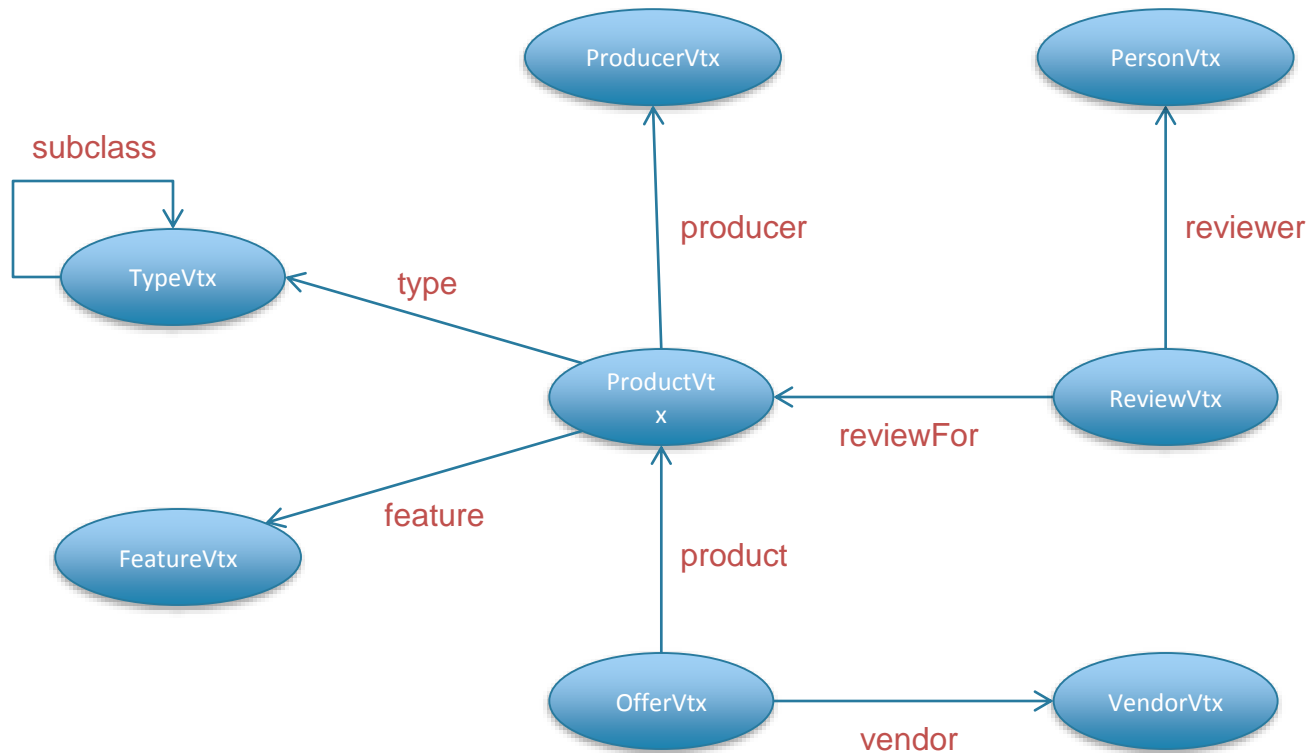
create vertex FeatureVtx(id)
from table Feature
```



CREATE EDGE

```
create edge feature with vertices (ProductVtx, FeatureVtx)
from table ProductFeature
where ProductFeature.product = ProductVtx.id and ProductFeature.feature =
FeatureVtx.id
```


A graph view



A graph makes it easy to find patterns

SELECT THE TOP 10 PRODUCTS MOST SIMILAR TO A SPECIFIC PRODUCT,
RATED BY THE COUNT OF FEATURES THEY HAVE IN COMMON



```
select y.id
from graph ProductVtx(id = %Product%) -feature-> FeatureVtx()
      <-feature- foreach y: ProductVtx(id != %Product%)
into table T1
```

```
select top 10
  id, count(*) as sameFeature
from table T1
  group by id
  order by count(*) desc
```

```
SELECT ?otherProduct ?sameFeatures {
  ?otherProduct type Product .
  FILTER(?otherProduct != %Product%)
  {SELECT ?otherProduct (count(%otherFeature) as ?sameFeatures)
    {%Product% productFeature ?feature .
     ?otherProduct productFeature ?otherFeature .
    FILTER(?feature = ?otherFeature)
  }
  Group By ?otherProduct
} }
Order By desc(?sameFeatures) ?otherProduct
Limit 10
```

More complex definitions and queries

```
create vertex server(ipAddr)
from table Netflow
with ip = srcaddr or ip = dstaddr
```

```
create edge comm with vertices (server as src, server as dst)
from table Netflow
with attributes (id, protocol, srcport, dstport, size, start, end)
where src.ipAddr = srcaddr and dst.ipAddr = dstaddr
```

```
from graph
  ( server -comm(protocol = 1)-> foreach ip1: server -comm(protocol = 1)->
    server -comm(protocol = 1)-> foreach ip3: server -comm(protocol = 1)-> ip1
  )
and
  (ip1 -comm(protocol = 1)-> server -comm(protocol = 1)-> ip3
  )
and
  (ip1 -comm(protocol = 1)-> server -comm(protocol = 1)-> ip3
  )
```

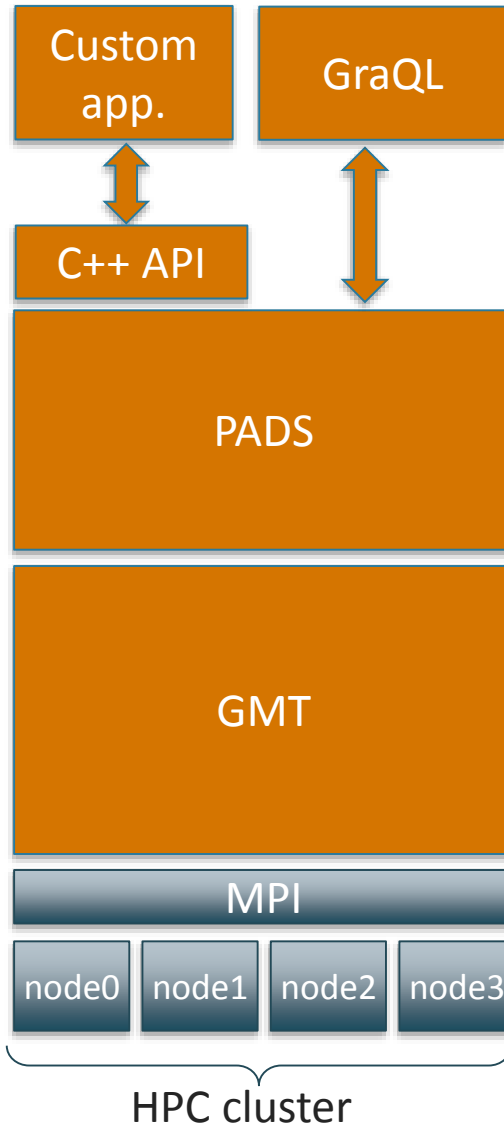
```
-x->      out-edge
<-x-      in-edge
~x->      conditional edge
|-x->      no edge
```

Graph Engine for Multithreaded Systems (GEMS)

Provide an in-memory property-graph database capable of doing **deep analytic** query processing over **multi-terabyte datasets** with human-interactive response times on **commodity computing parts**

► Focus on

- Blending *relational* and *graph* data representations, methodologies, algorithms, and queries
- Providing an application programming interface (API) for custom applications using embedded graph databases
- Supporting data feeds that supply data in batches (e.g., hourly)
- Serving a range of user profiles: application programmer, subject matter expert analyst



Scalable, in-memory hybrid search engine for pattern discovery

Castellana, et. al, In-Memory Graph Databases for Web-Scale Data.
IEEE Computer 48(3): 24-35 (2015)

Parallel and Distributed Algorithms and data Structures:

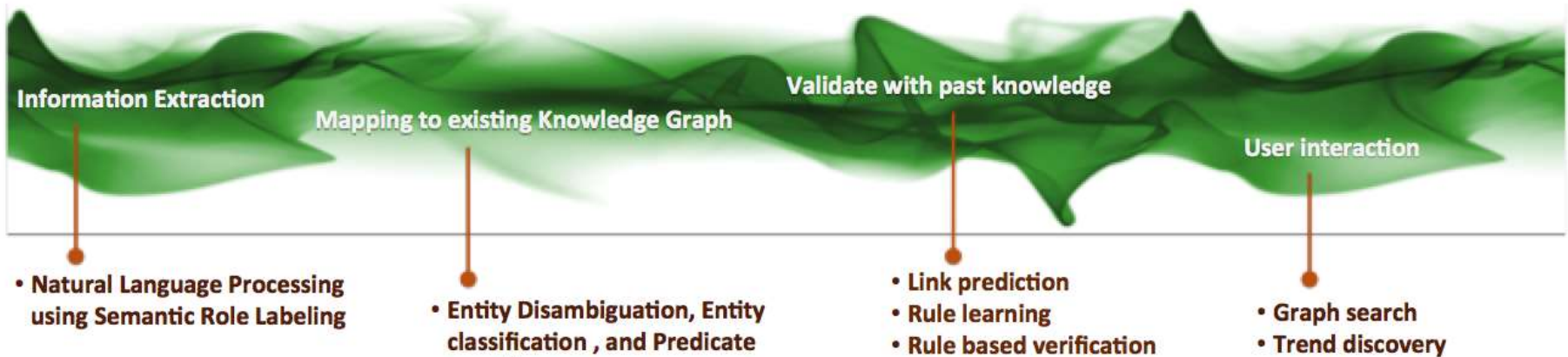
- ▶ Table to store tuples
- ▶ Index (Graph) for fast neighbours access
- ▶ Dictionary for encoding
- ▶ CSV and RDF parser

Global Memory and Threading runtime system:

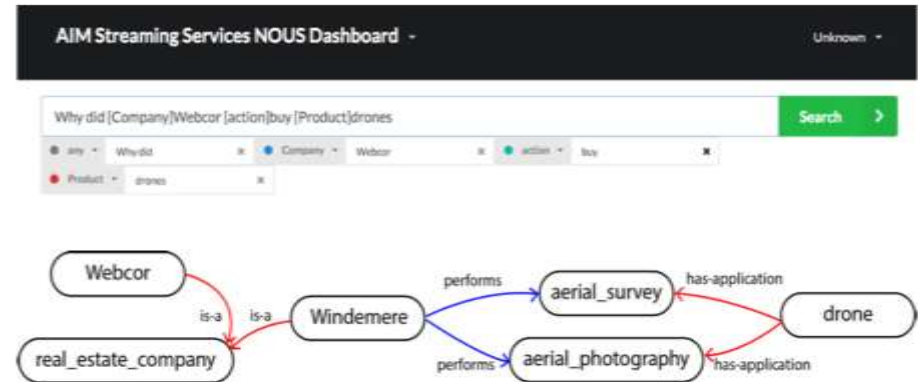
- ▶ Runtime library implemented in C
- ▶ Requirements: MPI, Linux, x86
 - ▶ Single node uses Intel TBB
- ▶ Partitioned Global Address Space (PGAS)
- ▶ Parallel loops program structure (*parFor*)
- ▶ Massive user-level asynchronous tasks
- ▶ Software multi-threading to hide latency
- ▶ Message aggregation

Morari, et.al. Scaling Irregular Applications through Data Aggregation and Software Multithreading. IPDPS 2014: 1126-1135

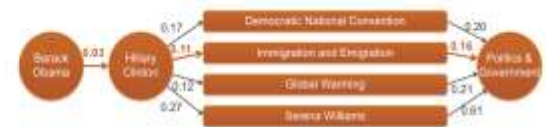
NOUS: Incremental Knowledge Graphs



- ▶ Incrementally construct KGs from streaming data and multiple sources
- ▶ Support advanced trending and explanatory questions
- ▶ Automatically generate hypotheses to guide users in information discovery
- ▶ Web-based, interactive UI

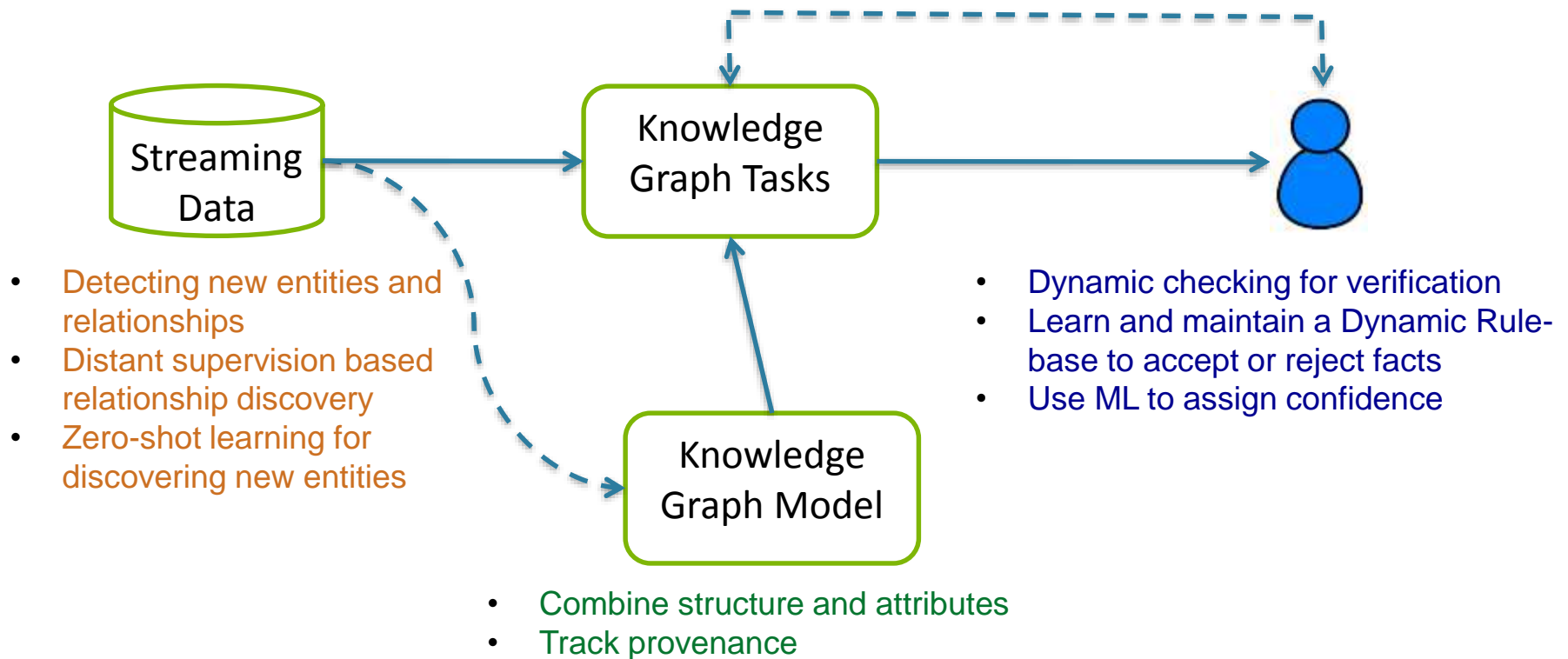


Toyota announced	Elron loss	2010-01-01T00:01:00.000
Microsoft developing	all- microsoft case	2010-01-01T00:01:00.000
Klasse developing	i use gmail	2010-01-01T00:01:00.000
Chrysler halt	frighting sales pions	2010-01-01T00:01:00.000
Chrysler halt	29 k	2010-01-01T00:01:00.000
Toyota recalled	Last part:	2010-01-01T00:01:00.000
Toyota recalled	a million cars	2010-01-01T00:01:00.000
Toyota recalled	see pedals	2010-01-01T00:01:00.000
Toyota recalled	floor mats	2010-01-01T00:01:00.000
Toyota recalled	warn	2010-01-01T00:01:00.000
Toyota recalled	universal	2010-01-01T00:01:00.000
Lawyers begin	settlements	2010-01-01T00:01:00.000
Lawyers begin	C.S.	2010-01-01T00:01:00.000
Obama began	speech	2010-01-01T00:01:00.000



We use topic distribution of entities to find coherent paths in Knowledge Graph

- Continuous pattern (or rule) discovery, search, and reasoning
- Algorithms that monitor emerging “solution sketches” or “pathways”
- Integrates machine-learning models (e.g., LDA, RNN) to steer the search





United States government	crack	an iPhone that belonged to a gunman in the San Bernardino
Apple	repair	the particular iPhone hole that the government hacked.
Federal officials	specify	the procedure used to open the iPhone
Federal officials	deny	to specify the procedure used to open the iPhone.
Jay Kaplan, chief executive of Synack and a former NSA analyst.	says	Apple has to earn the trust of Apples customers,"
F.B.I.	cracks	Mr. Farook's
LegbaCore, which previously found and fixed flaws for Apple.	found and fixed	flaws for Apple.

- ▶ Built on top of Stanford CoreNLP and OpenIE
- ▶ Adds a layer with heuristics to minimize the noise in triples

Navigating data

Find news about autonomous drones



What are their capabilities?



DJI Phantom can now perform autonomous flight

San Diego, CA | July 2, 2014



Thanks to an app update, the DJI Phantom 2 Vision and Vision+ are now able to follow pre-programmed flight paths. [View gallery \(8 images\)](#)

What are their software components?

Hardware - The embedded systems and peripheral sensors that act as the vehicle's brain, eyes, ears, etc. Almost any mobile machine can be transformed into a robot, by simply integrating a small hardware package into it.

Firmware - The "skill set" code running on the hardware, which configures it for the kind of vehicle you've put it in. You choose the firmware and vehicle that match your mission: Plane, Copter, Rover...

Software - Your interface to the hardware. Initial set-up, configuration, and testing; mission-planning/operation, and post-mission analysis. Point-and-click intuitive interaction with your hardware, or advanced custom scripting for niche mission profiles. Options are everything with ArduPilot.

Ardupilot: An Open Source award winning platform

Navigating data (cont.)

Which vendors are mentioned with Ardupilot?



Arduino

This little board is truly disruptive because it breaks down several barriers that existed in the hardware world. It illustrates this with an example: in order to develop a hardware product (a drone, a smart thermostat, a phone...) one could buy an ARM processor on a board and program it

Where is Arduino?



Who sells Arduino?



YKS Upgraded Arduino APM 2.8 Flight Controller Board built-in Compass w/ Shock Absorber for RC Quadcopter

by YYS
★★★★★ 3 customer reviews

See: \$48.00 A **FREE Shipping** on orders over \$40. Details
+ \$0.00 estimated tax

In Stock

Want it tomorrow, May 13? Order within 5 hrs 18 mins and choose **One-Day Shipping** at checkout. Details

Sold by **EWKStore** and Fulfilled by Amazon.

- APM2.8 Flight Controller, APM2.8 2.8 Upgraded version, revised APM version
- Includes 3-axis gyros, accelerometer and magnetometer, along with a high-performance barometer, barometric pressure sensor upgraded to MS5611-01BA03, from Measurement Specialties
- Onboard 4 MegaByte Dataflash chip for automatic datalogging. Optional off-board GPS, uBlox LEA-6H module with Compass.
- One of the first open source autopilot systems to use InvenSense's 6 DoF: Accelerometer/Gyro MPU-6



Summary

- ▶ Big data is changing our world
- ▶ Winning platforms will implement a variety of methods and structures
- ▶ PNNL has developed *an in-memory, scalable table-graph engine* capable of supporting a trillion attributed vertices and edges
- ▶ ... and *extended SQL* to define graph views and queries
- ▶ ... and built *a knowledge graph* capable of processing two million web crawls and analyzing trends, connecting dots, and presenting hypotheses



Gartner Hype Cycle, https://en.wikipedia.org/wiki/Hype_cycle

Contacts

- ▶ GraQL – Daniel Chavaria, daniel.chavaria@pnnl.gov
- ▶ GEMS – David Haglin, david.haglin@pnnl.gov
- ▶ NOUS – Sutanay Choudhury, sutanay.choudhury@pnnl.gov