AVESTERRA

# Big Data for Big Problems

*34th International Conference on Massive Storage Systems and Technology (MSST 2018)*
*Monday, May 14, 2018, 2-6 p.m.*
*Santa Clara University*


A tutorial presented by:

- Norman R. Kraft (Georgetown)
- Helen Karn (Georgetown)
- Stephen Baird (AdaCore)


GEORGETOWN UNIVERSITY

# *AvesTerra Tutorial Schedule*

| | |
|---|---|
| 2:00 to 2:30 p.m. | Ada 2012, Spark, and AdaCore |
| 2:30 to 3:45 p.m. | AvesTerra architecture |
| 3:45 to 4:15 p.m. | Break |
| 4:15 to 5:30 p.m. | AvesTerra adapters |
| | AvesTerra toolkit and API |
| 5:30 to 6:00 p.m. | AvesTerra roadmap |
| | Questions and Discussion |

# The AvesTerra Team

## Georgetown University

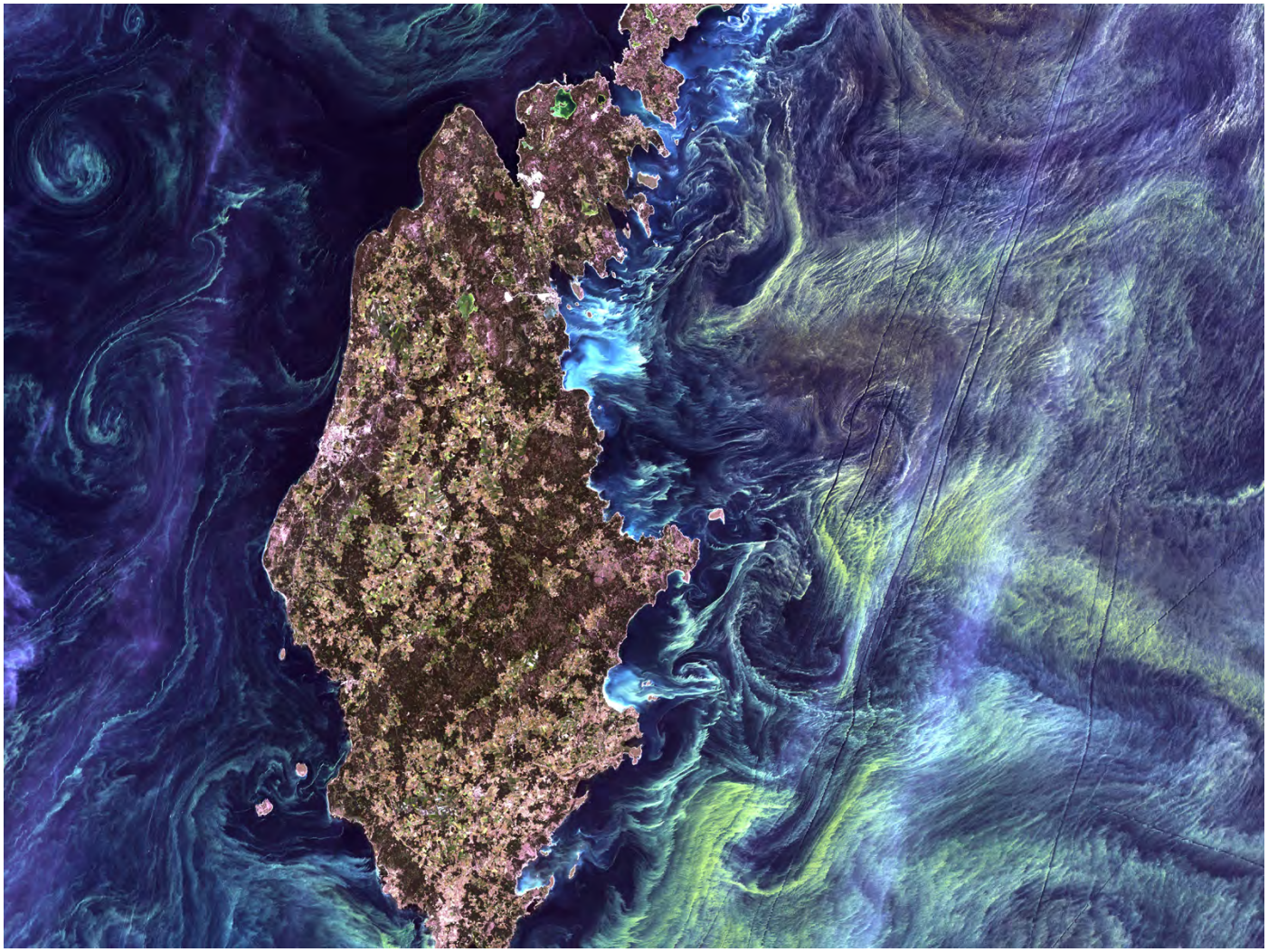
- J. C. Smart, *AvesTerra Chief Scientist*
- David Bridgeland, *AvesTerra adapters, toolkit, API*
- Norman Kraft, *AvesTerra adapters, toolkit, API*
- Helen Karn, *AvesTerra ontology and taxonomies*
- John Cederholm, *AvesTerra Visualization Utility (AVU)*
- Jianan Su, *AvesTerra testing*

## Collaborating institutions

- American University
- Lawrence Livermore National Laboratory (LLNL)
- LEDR Technologies Inc.
- Oak Ridge National Laboratory (ORNL)
- Pacific Northwest National Laboratory (PNNL)

GEORGETOWN UNIVERSITY

# AvesTerra architecture



*Van Gogh from Space*, USGS/NASA/Landsat 7, July 13, 2005

# *AvesTerra's Big Data focus*

1. Extreme scale
2. Highly distributed
3. Highly complex organizations
4. Multi-disciplinary use cases
5. Unique organizational cultures
6. Differing privacy policies
7. Data sharing but not sharing data

*It's not necessarily* the amount of data...
but where it is and how it's shared

# Why is "connecting-the-dots" so hard?

- **Plumbing:** Massive logistics problem to integrate thousands of government/non-government data systems at scale
  - *Different standards, models, security, infrastructure, procedures, policies, networks, access, compartments, applications, tools, protocols, etc. … all at immense scale!*
- **Protection:** Large-scale integration of data resources increases cyber security risks
  - *Prevention of adversary exploitation of strategic national assets.*
- **Patterns:** Lack of analytic algorithm techniques to automatically detect data patterns and alert
  - *Transition from "analytic dumpster diving" to early-warning indication and real-time notification*
- **Privacy:** Significant tension between security and liberty
  - *Who trusts the "watchers"?*
  - *Who watches the watchers?*
- **Politics:** What's in it for me?

GEORGETOWN UNIVERSITY

# *Use case: U.S. intelligence community*

17 different organizations with different missions

Sharing without sharing


Department of the Treasury

U.S. Marine Corps

U.S. Coast Guard

Central Intelligence Agency

National Security Agency/ Central Security Service

Drug Enforcement Administration

U.S. Air Force

Department of State

Office of the Director of National Intelligence

National Geospatial-Intelligence Agency

Department of Energy

Defense Intelligence Agency

Federal Bureau of Investigation

Department of Homeland Security

National Reconnaissance Office

U.S. Navy

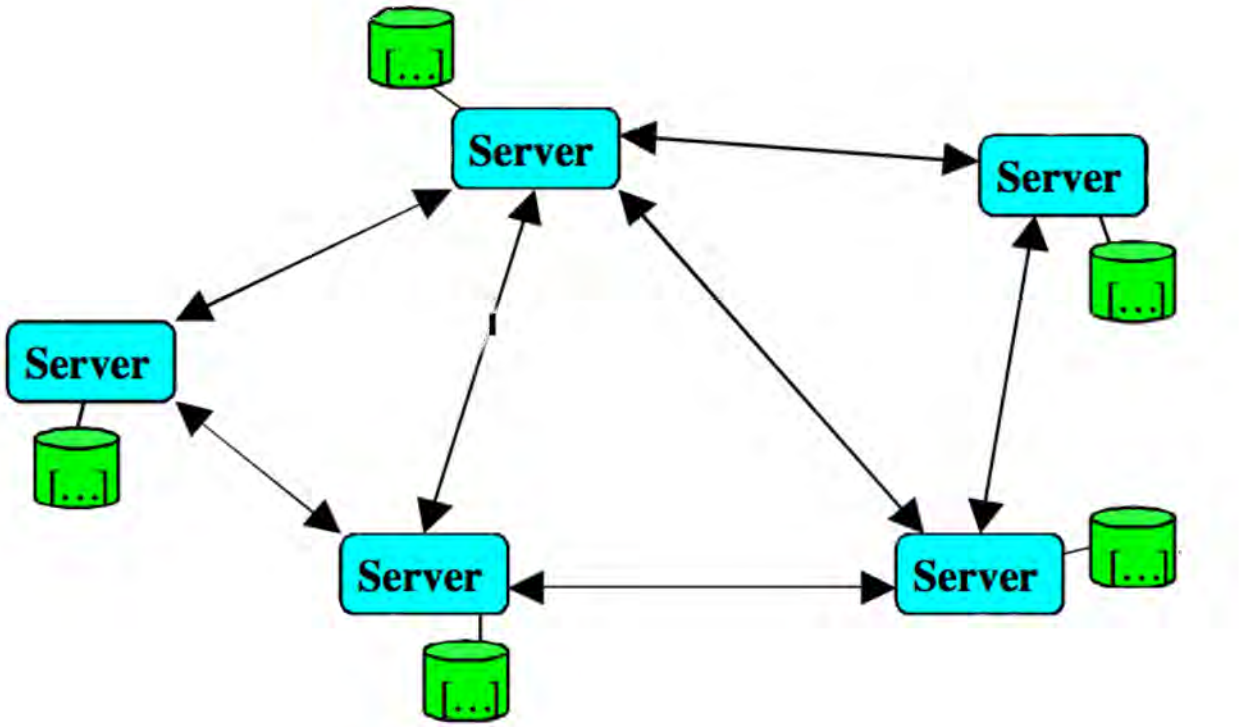
U.S. Army



GEORGETOWN UNIVERSITY

# *Solution: A large graph*

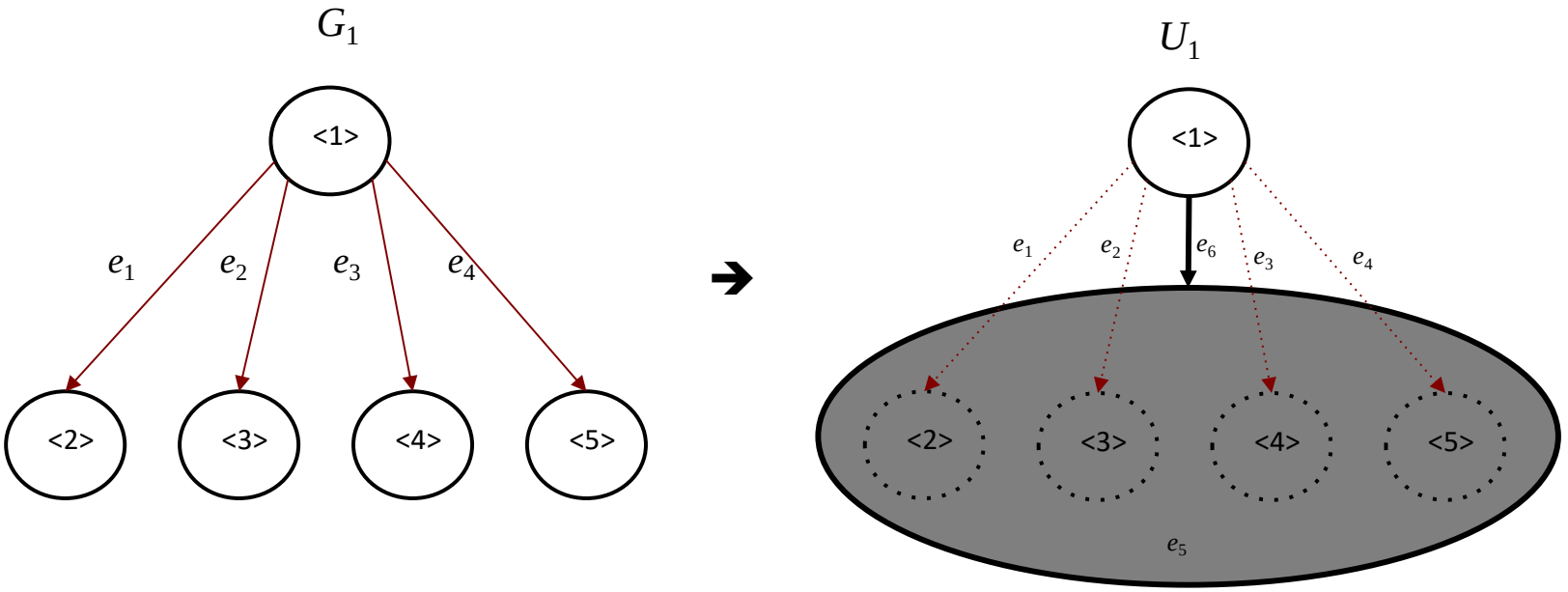
- AvesTerra manifests the appearance of a big graph
- The graph is a knowledge hypergraph






GEORGETOWN UNIVERSITY

## *Ultragraph example*

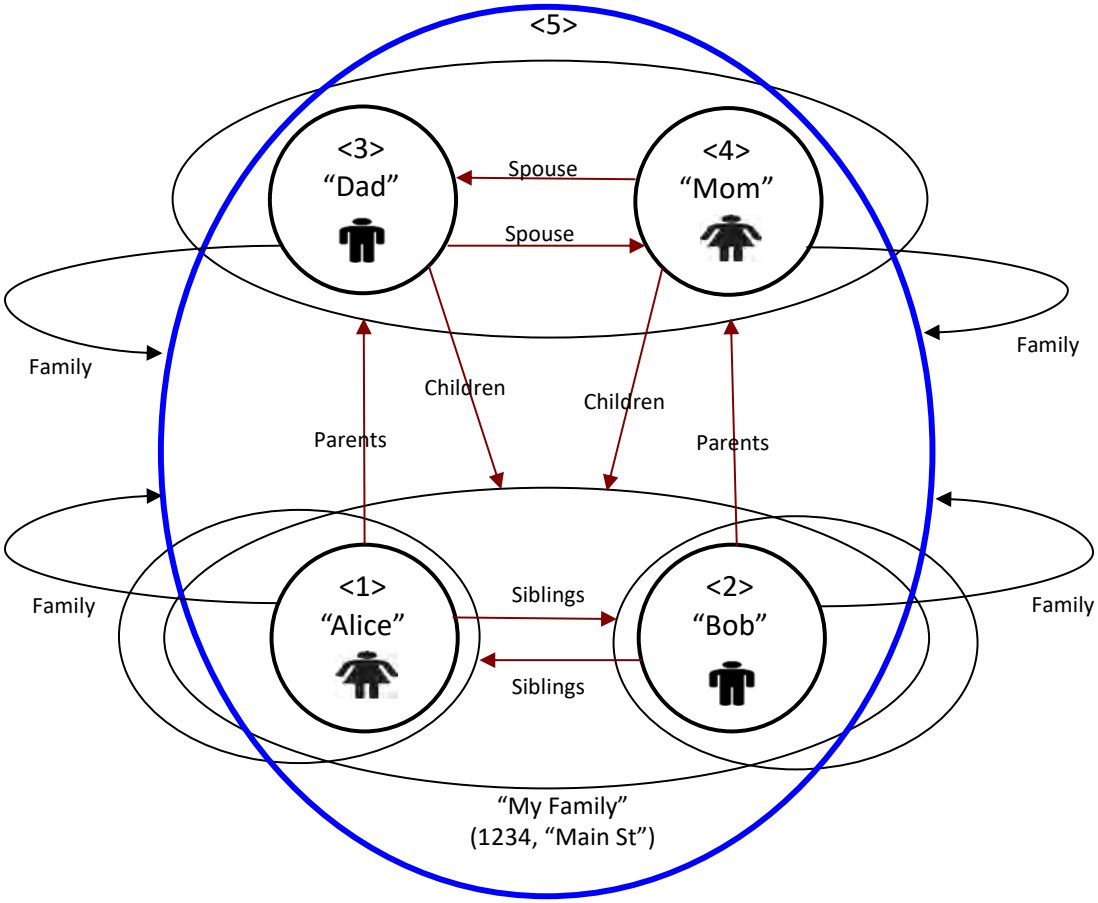



$$R \subseteq E \times E$$

$$R \subseteq \mathcal{P}(\mathcal{P}(\ldots \mathcal{P}(E)))$$

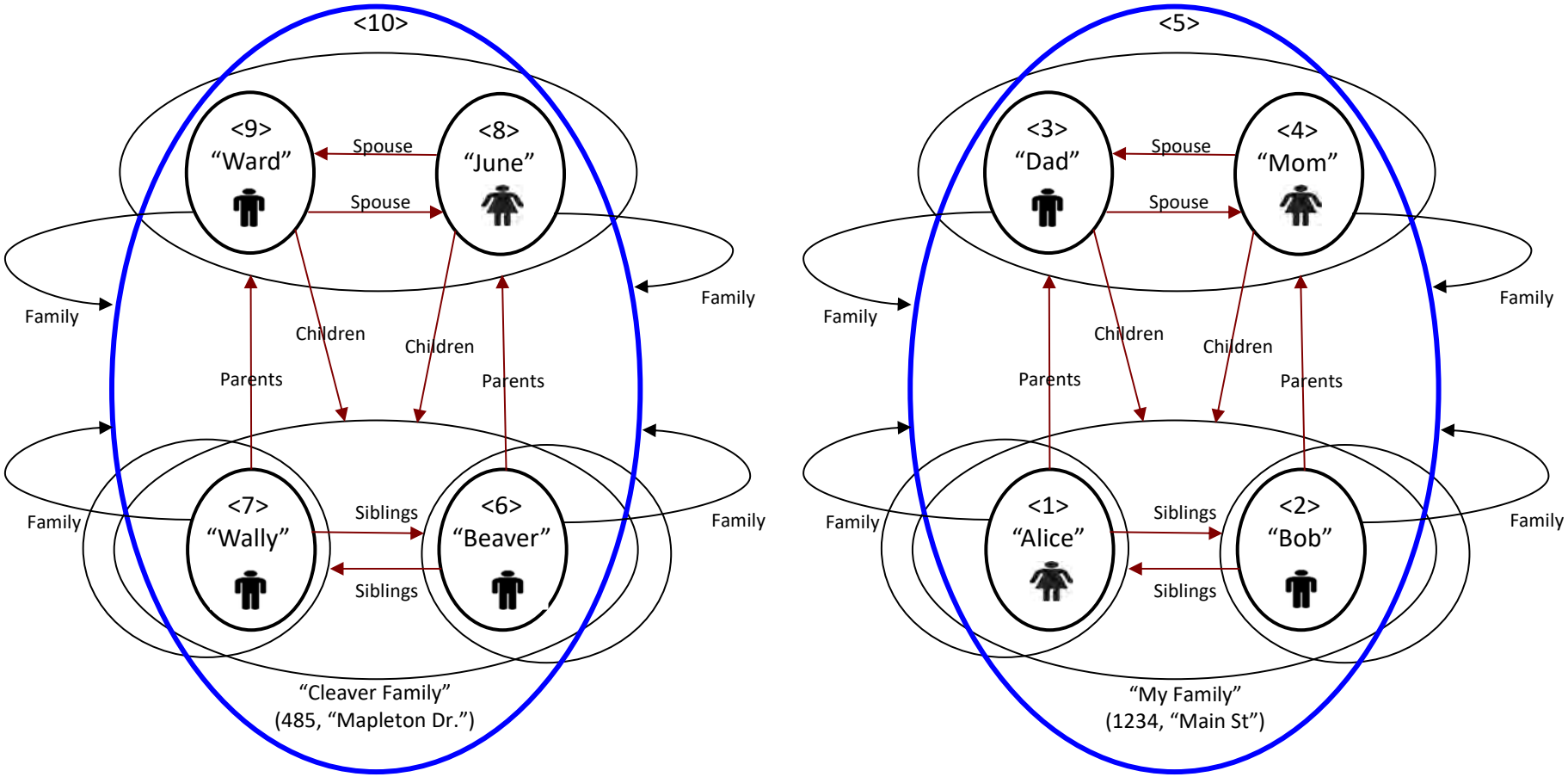
GEORGETOWN UNIVERSITY

# Ubergraph example




D= 411 I= 42 K= 20 DD= 82.2 ID= 8.4 KD= 0.8000 J= 25.0

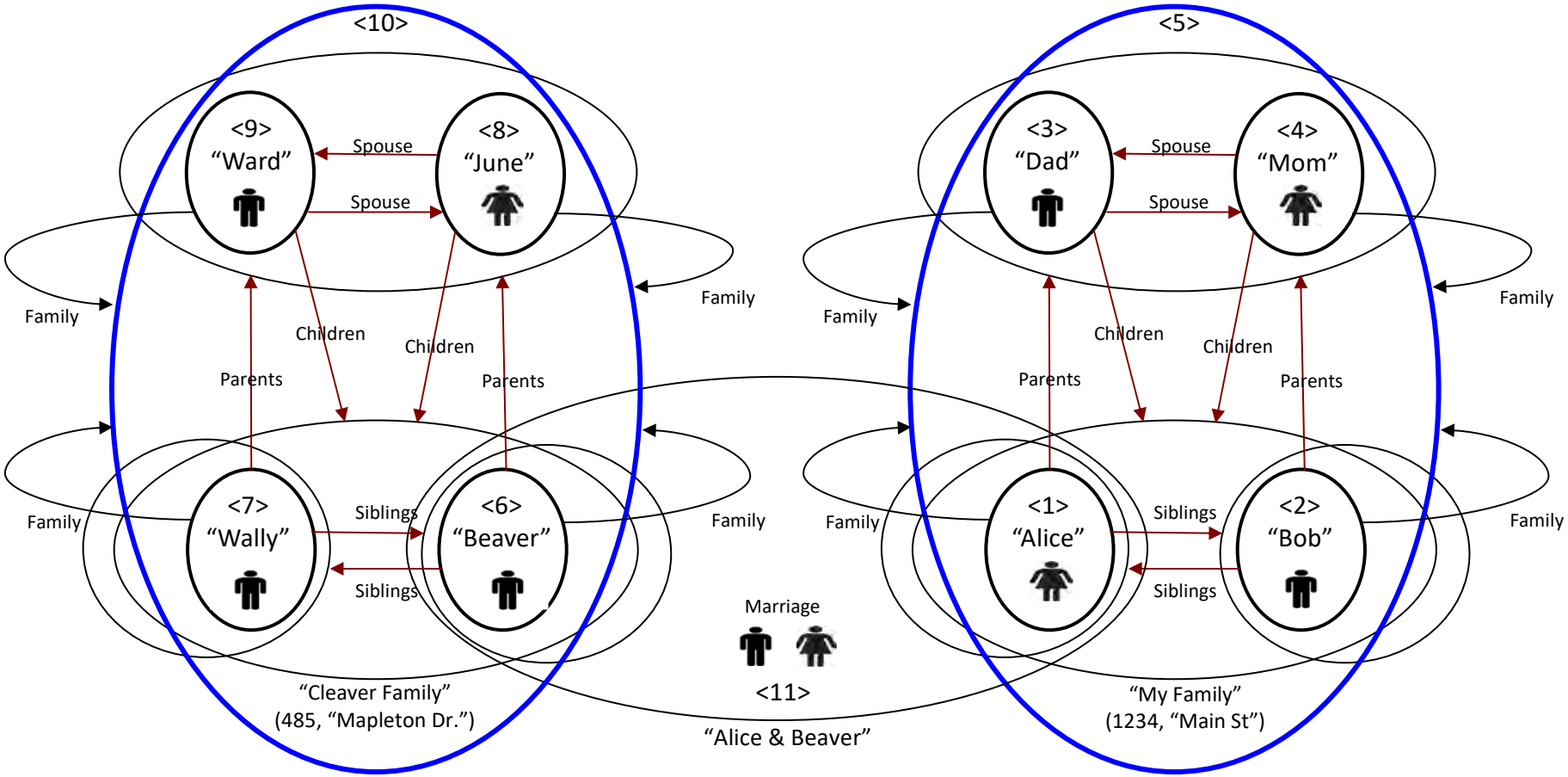
# Ubergraph example (cont)




D= 814 I= 84 K= 40 DD= 81.4 ID= 8.4 KD= 0.4000 J= 50.0

GEORGETOWN UNIVERSITY

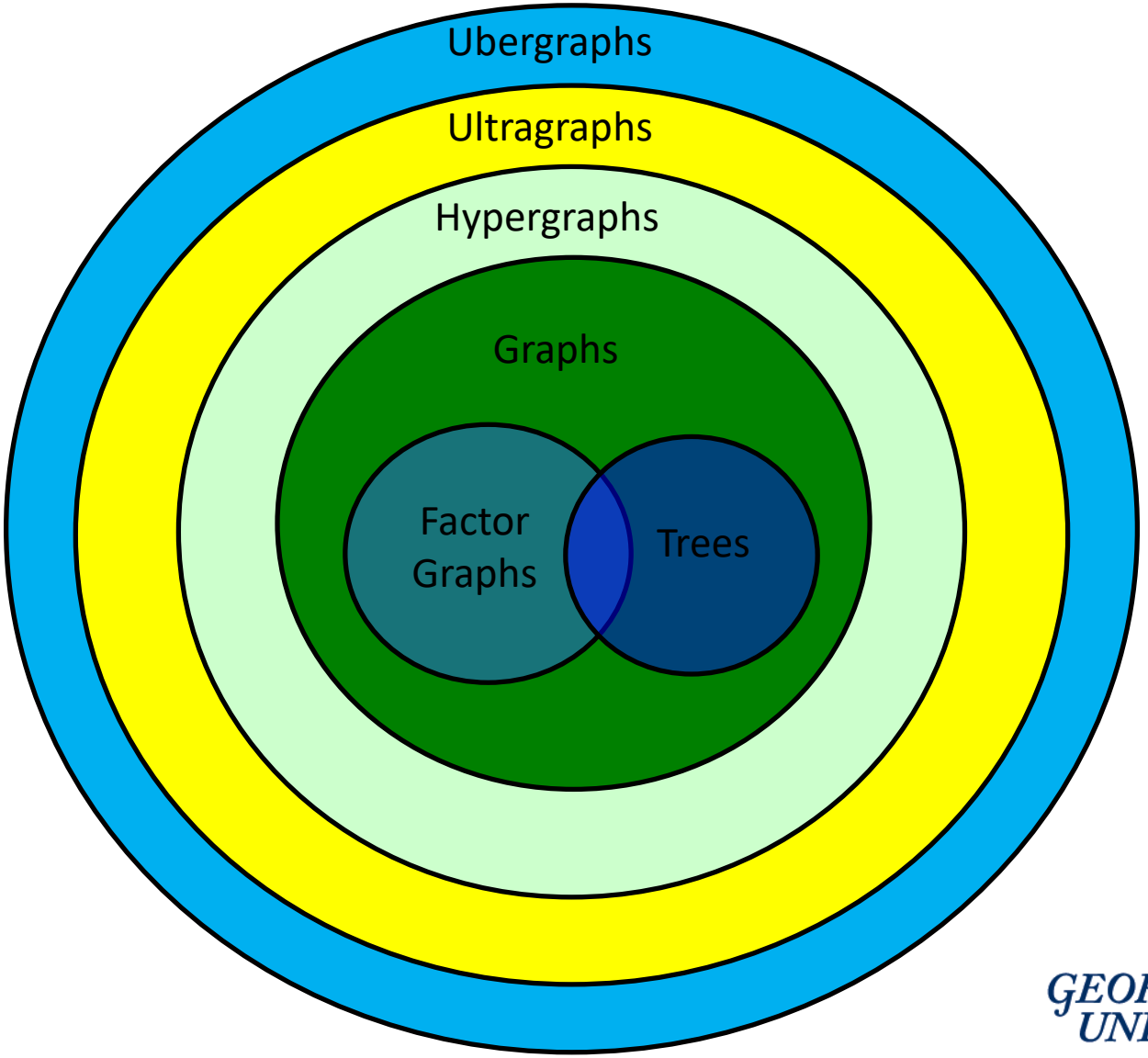
# Ubergraph example (cont)




D= 868 I= 89 K= 42 DD= 78.9 ID= 8.1 KD= 0.3471 J= 121.0

GEORGETOWN UNIVERSITY

## *AvesTerra Graph Families*




GEORGETOWN UNIVERSITY

## *The nodes of the big graph = entities*

An entity is any THING that can be observed, measured, or described:

- MSST 2018, MSST 2017
- Adacore, Apple, Facebook, Google, Qualcomm
- Santa Clara, Silicon Valley, Washington, DC, California, U.S.A., Basque Country, Spain
- Santa Clara University, Georgetown University
- Norm Kraft, Helen Karn, Steve Baird
- the Society of Jesus ("the Jesuits")
- Association of Jesuit Colleges & Universities

# *Every entity belongs to a class*

PERSON_CLASS

Steve Baird, Norm Kraft, Helen Karn, Ignatius of Loyola

ORGANIZATION_CLASS

AdaCore, Georgetown University, Santa Clara University

Society of Jesus ("the Jesuits"), Association of Jesuit Colleges & Universities (AJCU)

LOCATION_CLASS

Santa Clara, District of Columbia, Pamplona, California, Basque Country, United States of America, Spain

# *Classes and subclasses form a taxonomy*

ORGANIZATION_CLASS

- BUSINESS_SUBCLASS
  - Adacore
- COLLEGE_SUBCLASS
  - Santa Clara University
  - Georgetown University
- COMMUNITY_SUBCLASS
  - Society of Jesus [the Jesuits]
- ORGANIZATION_SUBCLASS
  - Association of Jesuit Colleges & Universities (AJCU)

# *Entities have attributes*

PERSON_CLASS

Entity: Ignatius of Loyola

Attributes:

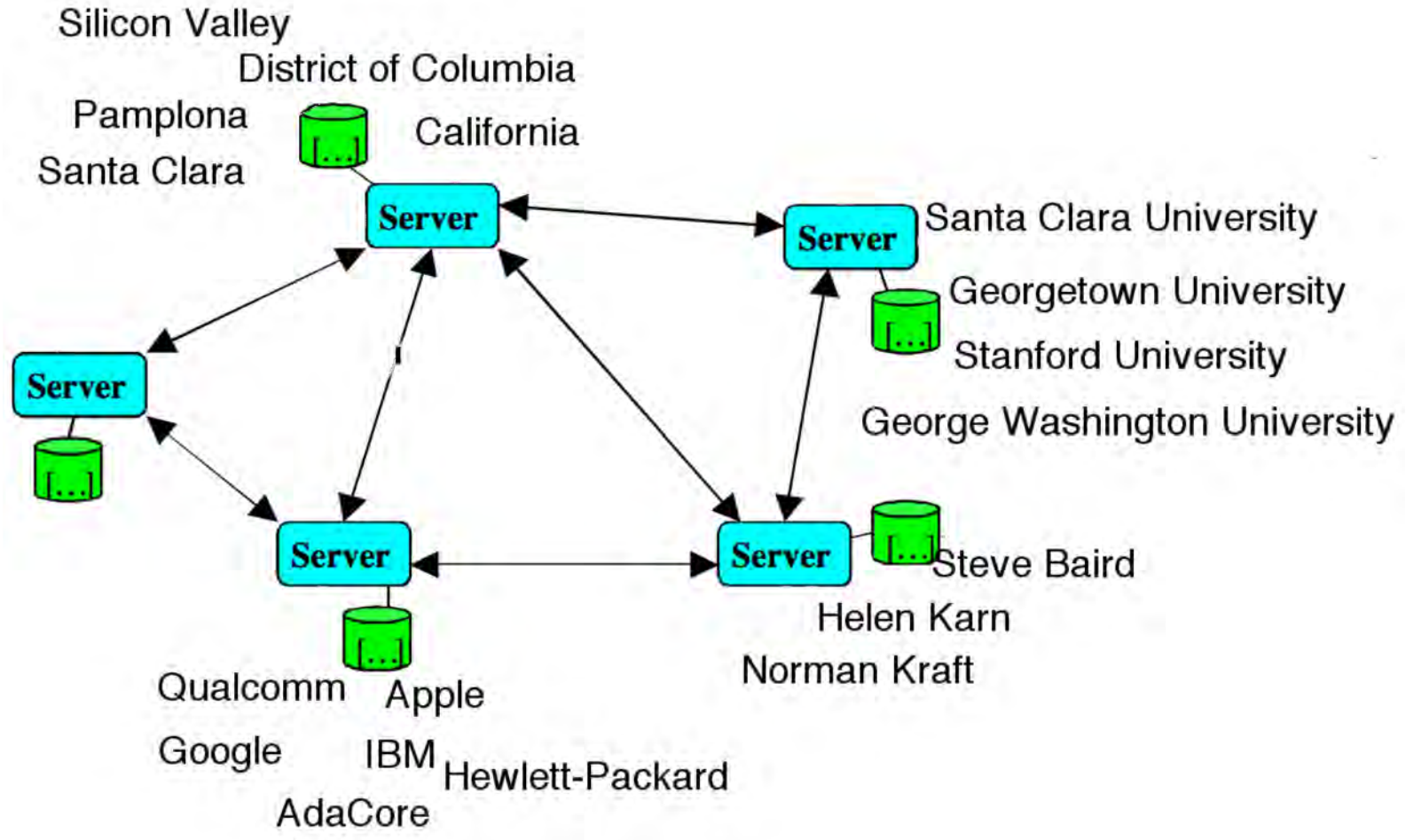
SEX_ATTRIBUTE: Male
LANGUAGE_ATTRIBUTE: Basque
NAME_ATTRIBUTE: Iñigo, Igacio, Ignatius
FAMILY_ATTRIBUTE: López de Loyola y Onaz
HEALTH_ATTRIBUTE: Poor
RELIGION_ATTRIBUTE: Roman Catholic

# *An attribute value can be another entity*

Entity: **Ignatius of Loyola**
Attributes: OCCUPATION_ATTRIBUTE: page, soldier, priest, saint
LOCATION_ATTRIBUTE: Azpeitia, Pamplona, Montserrat, Manresa, Rome
**ASSOCIATION_ATTRIBUTE**: **Society of Jesus**

**Entity:** **Society of Jesus**
Attributes: NAME_ATTRIBUTE: Compañía de Jesús, the Jesuits
MANAGER_ATTRIBUTE: Saint Ignatius of Loyola, [+29 more], Arturo Sosa
ASSOCIATION_ATTRIBUTE: Georgetown University, Santa Clara University, Loyola University Chicago [+90 more]
PURPOSE_ATTRIBUTE: education, retreats

# Attribute values cross server boundaries






GEORGETOWN UNIVERSITY

# *Entities can have properties*

Entity: Ignatius of Loyola

Properties:

- hobbies: riding, dueling, gambling, billiards, dancing, womanizing
- injured-year: 1521
- injured-by: cannonball
- injured-place: Pamplona
- friends-of: Francis Xavier, Peter Faber
- canonized-date: March 12, 1622

# *AvesTerra properties vs. attributes*

| Properties | Attributes |
|---|---|
| Situation-specific | Universal semantics |
| Unlimited number | Limited number |
| User-defined strings | AvesTerra ecosystem-defined |

## *AvesTerra – Unique Aspects*

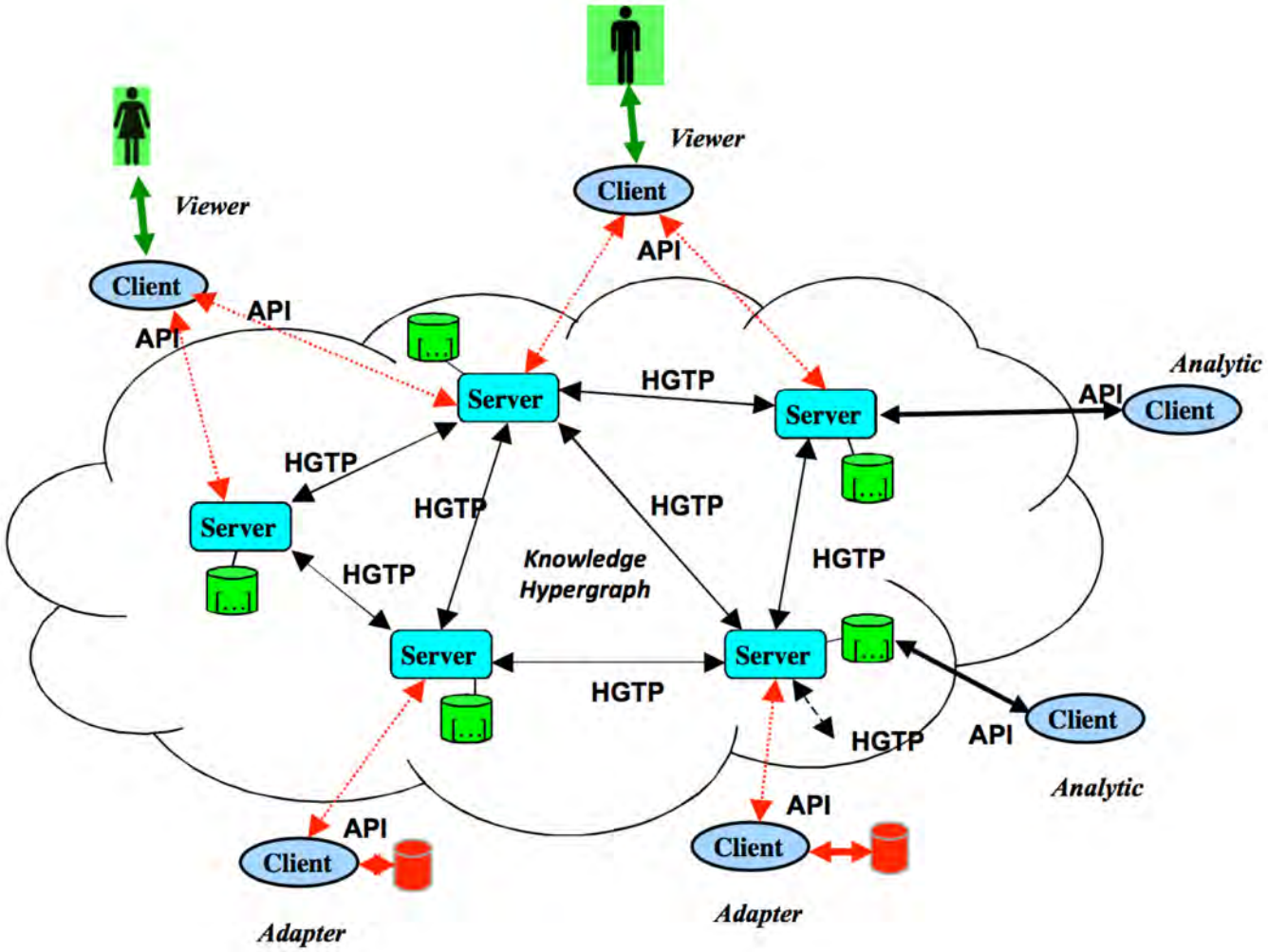
**Knowledge Overlay** is a technique used to create a shared semantic representation of a complex system that spans *many* diverse contributing organizations, data sources, and analytic components.

**AvesTerra Design Criteria**:

- **Global-scale (trillions of entities/quadrillions of relationships)**
- **Collaborative/Distributed (thousands/millions of participants)**
- **Semantic expressivity (complex physical and virtual systems)**
- **Multi-domain (hundreds/thousands)**
- **Multi-modal (hundreds/thousands)**
- **Multi-fidelity (microscopic to macroscopic)**
- **Dynamic (real-time, changing information flows)**
- **Analytic/semantic interoperability**
- **Minimized data movement/replication**

GEORGETOWN UNIVERSITY

# *AvesTerra:*
# *An* overlay *on existing data sources*



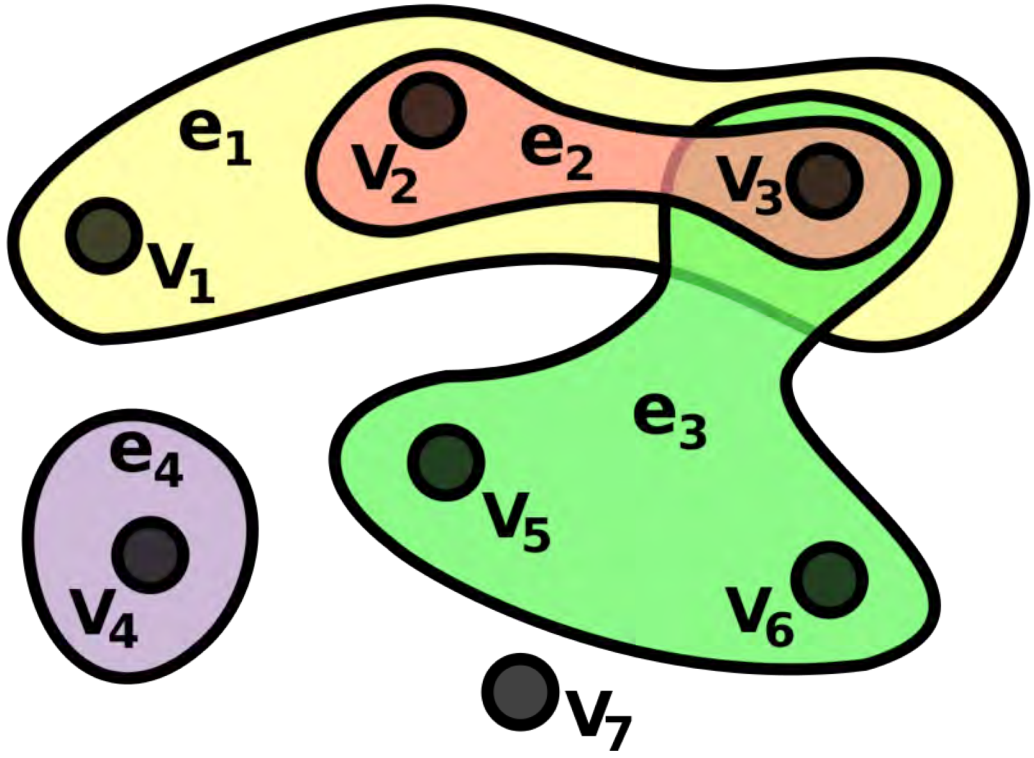



GEORGETOWN UNIVERSITY

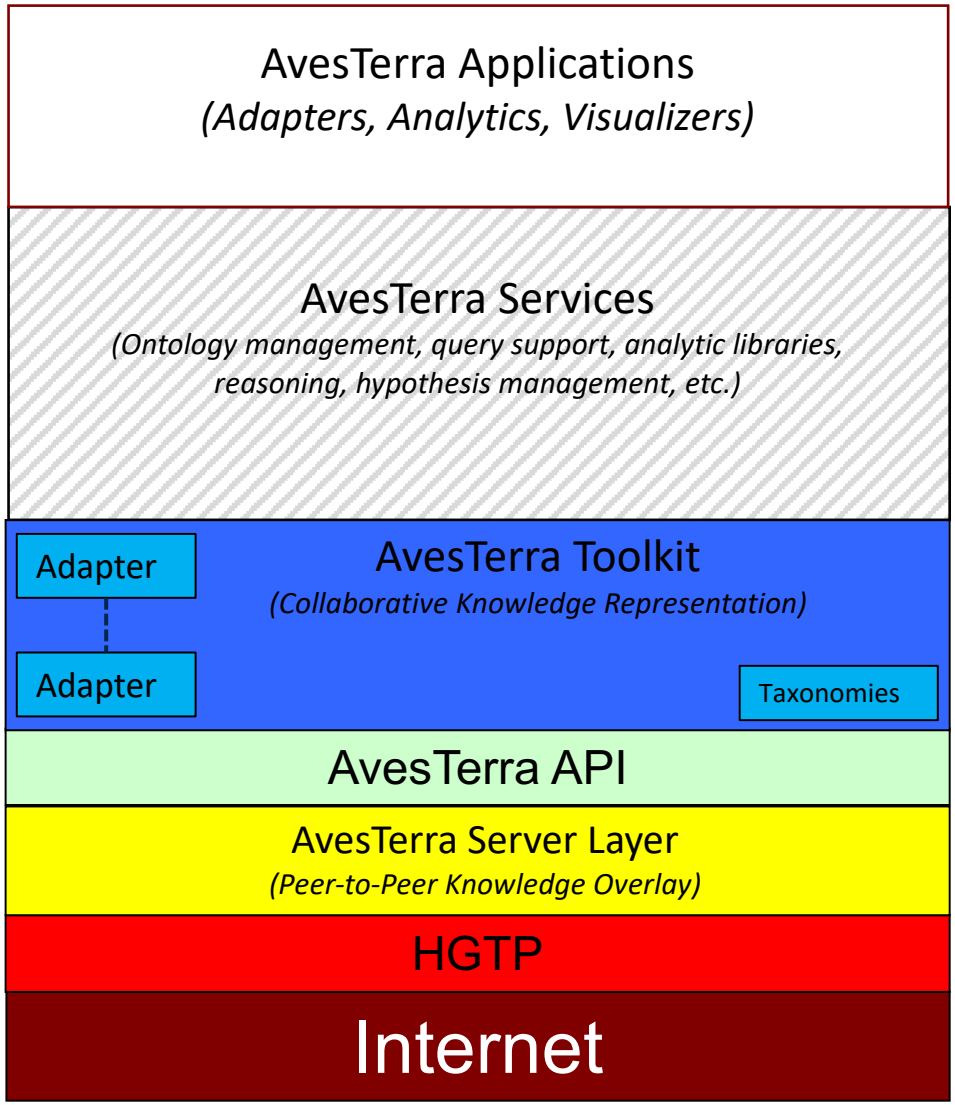
# AvesTerra Adapters

The AvesTerra global knowledge network is an ambitious project to connect the world's knowledge.

However the conceptual graph by itself isn't useful.



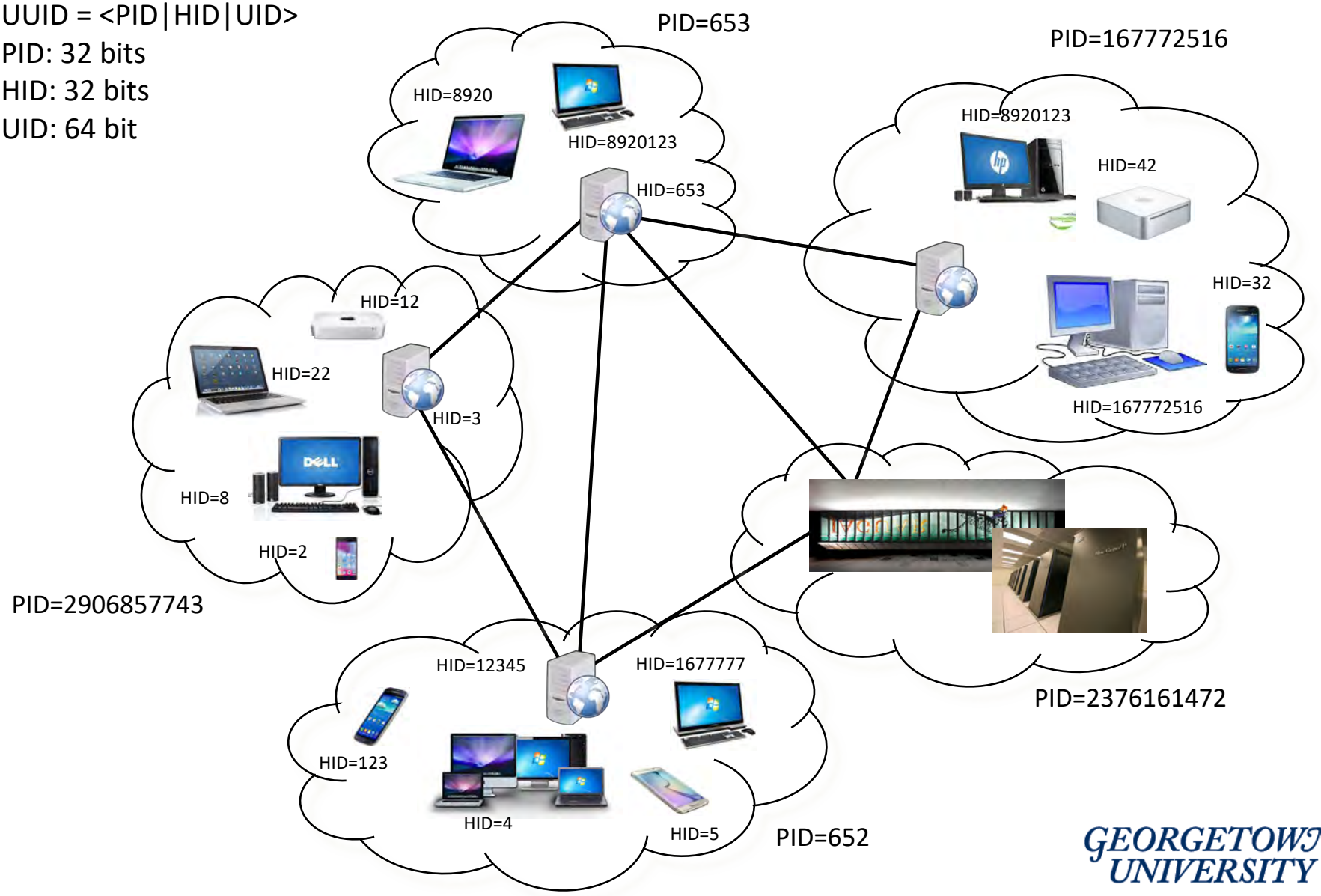
# *AvesTerra Layered Architecture*




GEORGETOWN UNIVERSITY

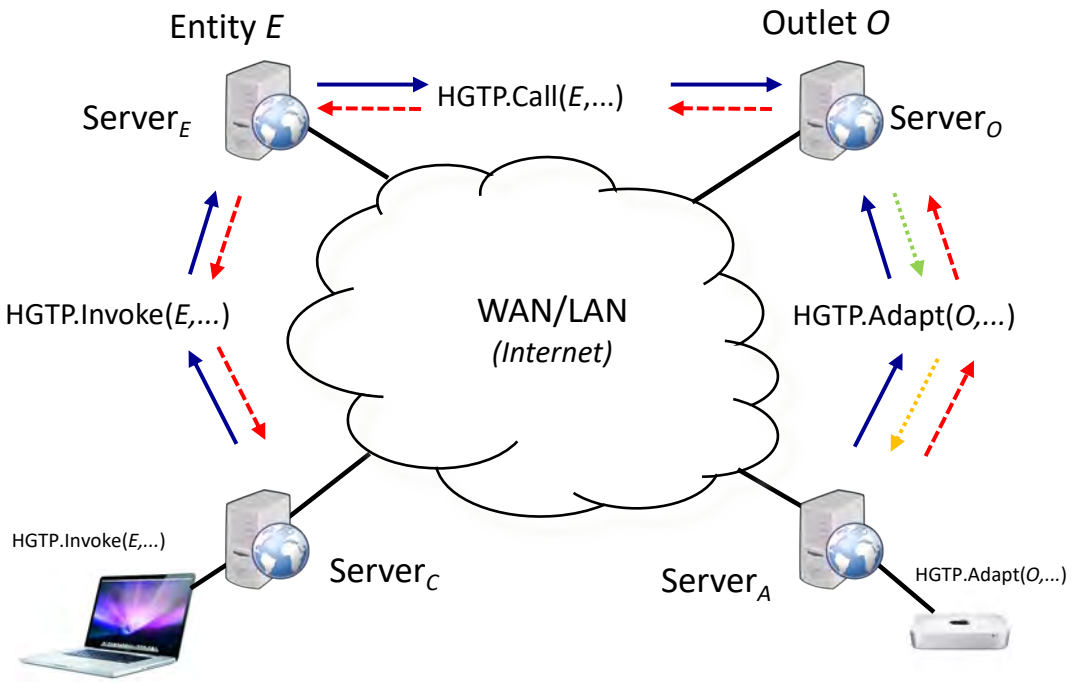
# *AvesTerra Entity Logical Addressing*

UUID = <PID|HID|UID>
PID: 32 bits
HID: 32 bits
UID: 64 bit




GEORGETOWN UNIVERSITY

# *AvesTerra Distributed Remote Rendezvous*




Client *C*

```
API.Create(E)
API.Connect(E,O)
....

API.Invoke(E,...)
API.Invoke(E,...)
API.Invoke(E,...)
```

Adapter *A*

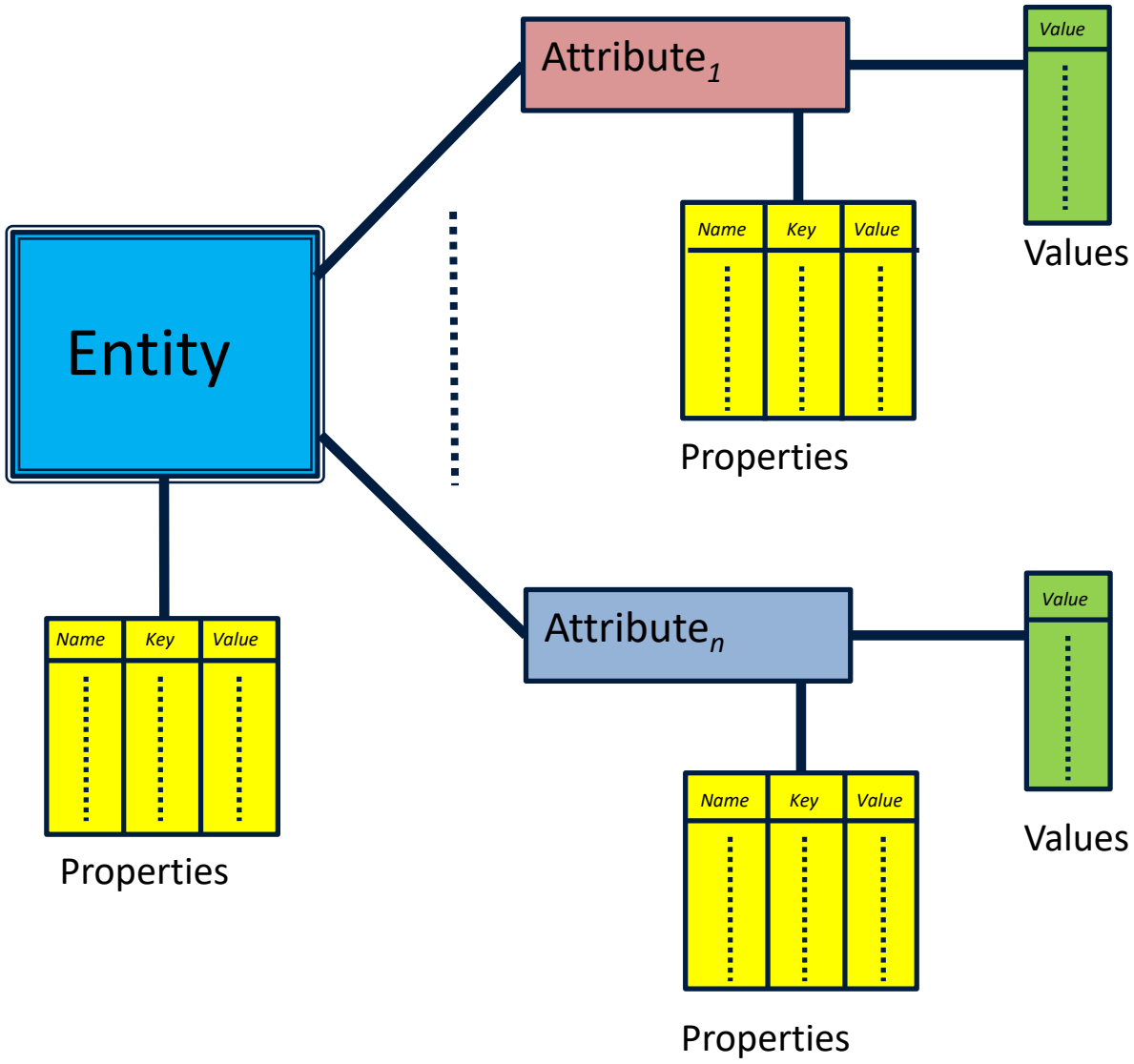
```
API.Create(O)
API.Activate(O)
....

loop
    API.Adapt(O,...)
end loop
```



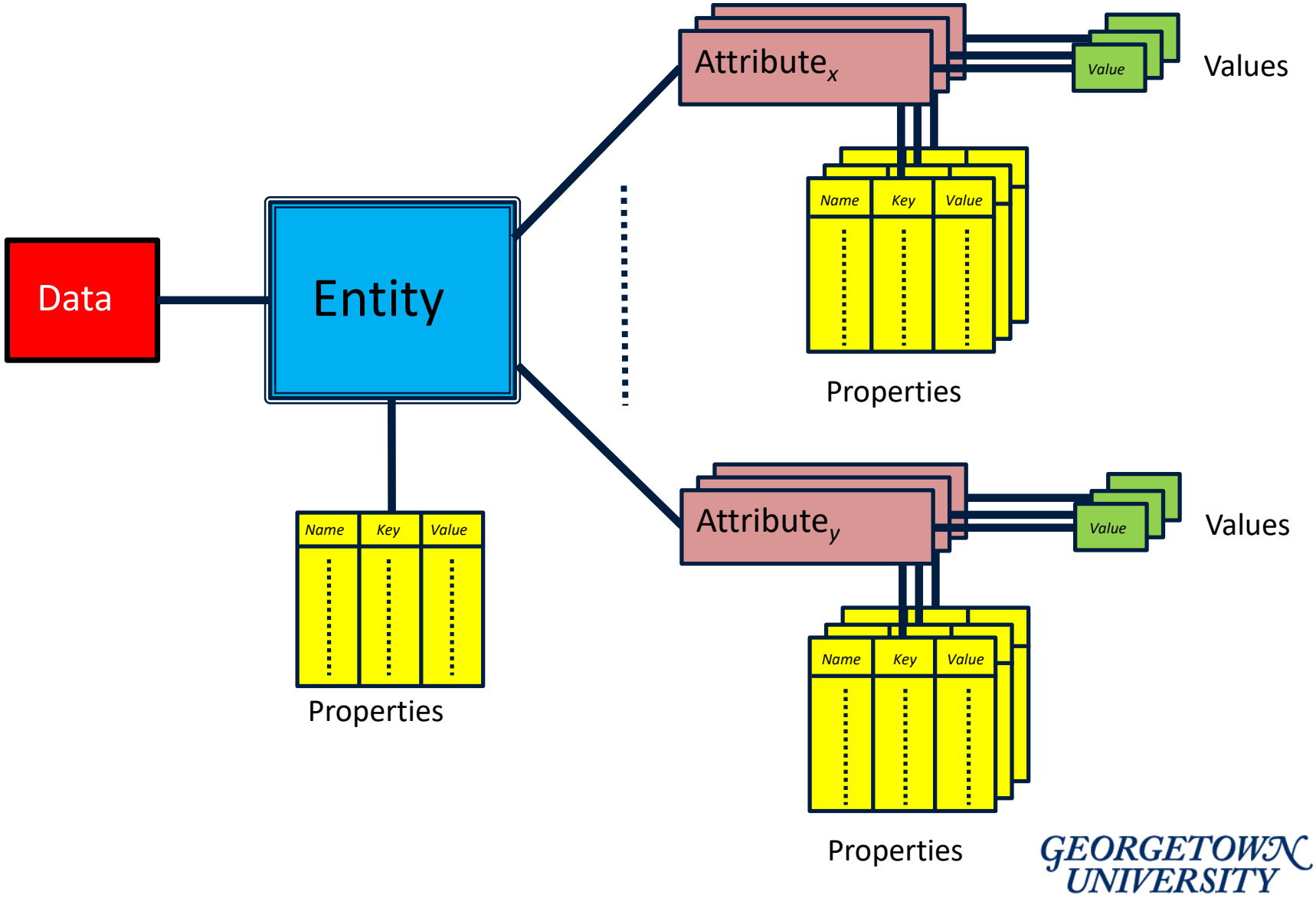
GEORGETOWN UNIVERSITY

# *AvesTerra Common Model*



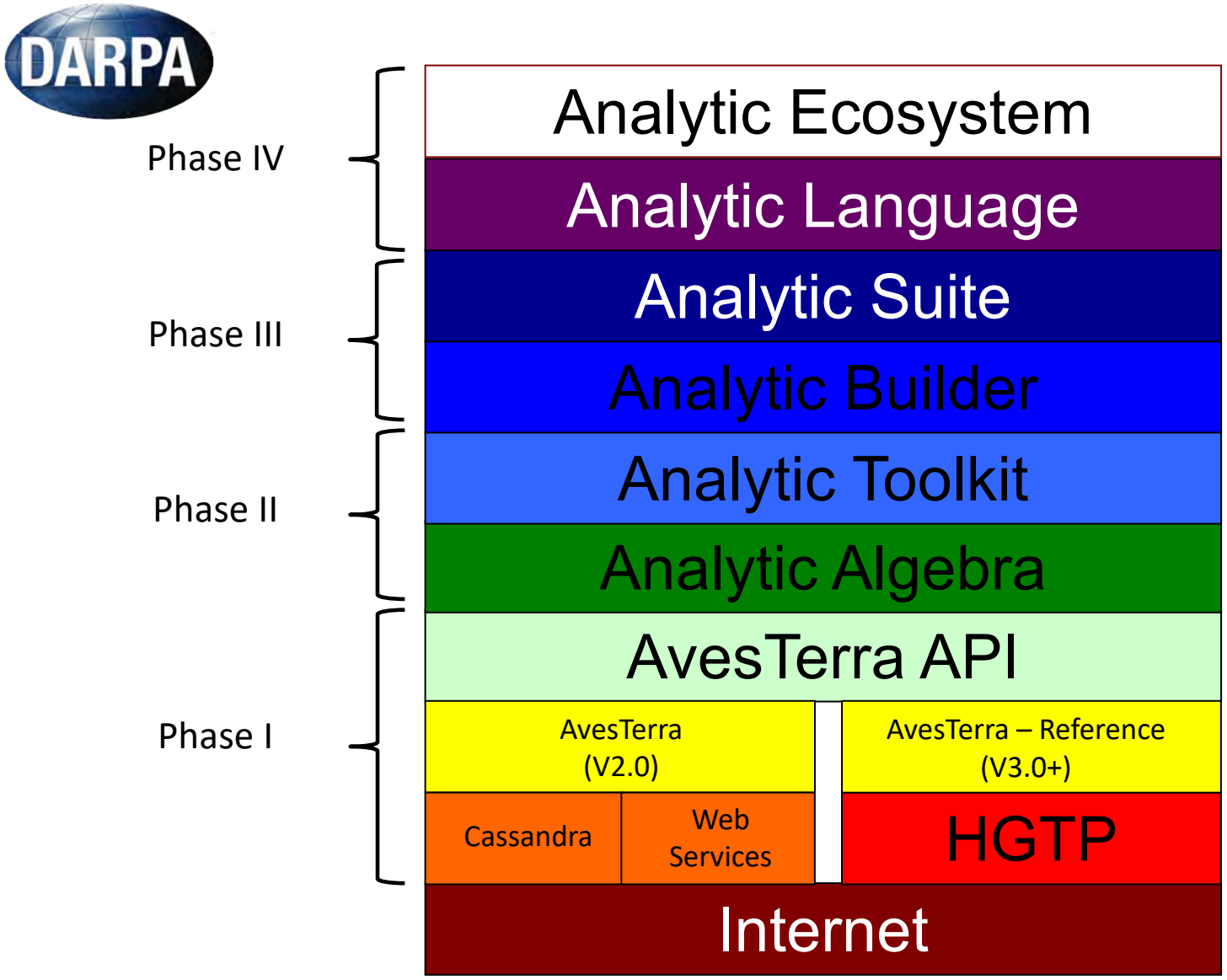

GEORGETOWN UNIVERSITY

# *AvesTerra Common Model (Enhanced)*




GEORGETOWN UNIVERSITY

## *AvesTerra Layered Architecture*




GEORGETOWN UNIVERSITY

## *AvesTerra API*

### *Summary*

**Primitives:**

- CREATE/DELETE (entities)
- CONNECT/DISCONNECT (methods)
- ATTACH/DETACH (attributes)
- INVOKE/INQUIRE (method/attribute access)
- REFERENCE/DEREFERENCE (garbage collection)
- ACTIVATE/DEACTIVATE (rendevous/queues)
- ADAPT/RESET (Adapters)
- PUBLISH/SUBSCRIBE/CANCEL (Events)
- WAIT/CLEAR (Subscribers)
- CALL/NOTIFY (RPC & subscriber notification)
- AUTHORIZE/DEAUTHORIZE (access control)
- REPORT (auxiliary functions)

**Current Bindings:** **Python, Swift, Ada, Java, C++**

**In Progress:** **Clojure, R**

**Available:** **avesterra@georgetown.edu**

GEORGETOWN UNIVERSITY

# *What is an Adapter?*

The AvesTerra graph needs information, and most of that will come from traditional data sources.

Adapters connect and translate data sources like databases and spreadsheets for the AvesTerra network and nodes communicate with each other via the HGTP protocol.

Adapters can also apply analytics to entities, on demand.

Adapters are gateways between your data and the knowledge network.

# *There Are Three Kinds of Adapters*

1. Read-only adapters (Web, Secured Data, REST, etc)

2. Read-Write adapters
   - Difference between knowledge network enrichment and updating a data store

3. Information adapters

# *Adapter Methods*

Eight Basic Operations on Entities:

- CREATE
- DELETE
- INVOKE
- INQUIRE
- REFERENCE
- DEREFERENCE
- SEARCH
- INDEX

Two Bulk Entity Operations:

- STORE
- RETRIEVE

# *Things an Adapter Must Do*

- Receive AvesTerra entities with methods and attributes
- Translate the request into a local datastore query
- Apply analytics, if needed or requested
- Convert the query response into an AvesTerra entity
- If an entity is created, keep track of entity ID and local data ID
- Optional but recommended: cache frequent requests

# *Adapters are Customized to Your Data Source*

While you may have a large PostgreSQL database, the adapter would present data in your tables as AvesTerra entity attributes and properties.

In the construction of the adapter, an organization has an opportunity to limit what data is shared.

# *AvesTerra Toolkit and API*

Since AvesTerra uses a custom protocol (HGTP) for communication, an API has been defined for the protocol.

From this API, low level bindings have been developed in several programming languages for the purpose of adapter and application development.

Because the low level bindings are not fun to use, we have also created the AvesTerra toolkits for each language binding. These toolkits simplify development.

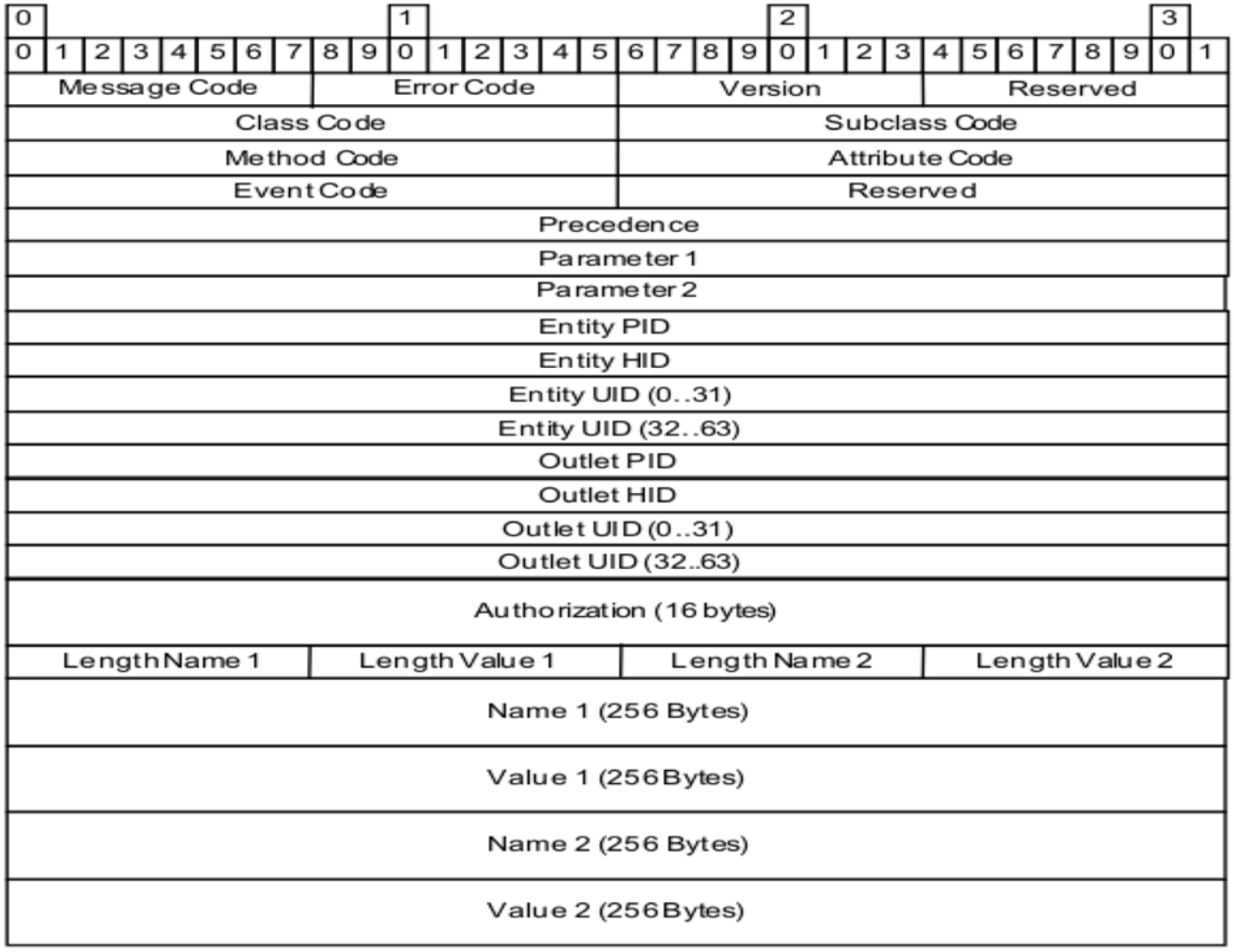
# *What is the HGTP Communications Protocol?*

HGTP is a simple all-text protocol with positional content. The communications protocol is based on a ACK/NAK system.

The simplicity of the HGTP protocol has two goals:

- To make application and binding development available across a wide variety of programming languages.
- A simple protocol  allows for enhanced security, as scanning the positional data in a message does not require a complex parser.

# *Message Packets Contain Entities*



# An Entity

```
Entity:         <2906857743|167772516|471952>
Name:           "winona"
Class:          PERSON_CLASS
Subclass:       NULL_SUBCLASS
Server:         <2906857743|167772516|0>
Timestamp:      2018-01-30 19:04:27
Activated:      FALSE
References:     0
Attachments:    0
Connections:    1
    <0|0|11>  NULL_METHOD  1
Subscriptions: 0
Authorizations: 0
Attributes:
   HEIGHT_ATTRIBUTE                1.61
   LOCATION_ATTRIBUTE              <2906857743|167772516|471953>
   AGE_ATTRIBUTE                   46
Properties:
   acting credits                                  []              Edward Scissorhands
   acting credits                                  []              The Age of Innocence
   acting credits                                  []              A Scanner Darkly
   acting credits                                  []              Stranger Things
   producing credits                               []              Girl, Interrupted
   acting credits                                  []              Girl, Interrupted
```

# *The Avesterra Toolkits*

The AvesTerra bindings are available in several languages.

The AvesTerra Toolkits have been created to make working with those bindings easier and at a higher level.

What would have taken five or six method calls at the binding level is often only one or two method calls in the toolkit.

Toolkit-based code is shorter, more functional and easier to comprehend.

# *AvesTerra Roadmap (2017-2018)*

| Version | Release Date | Features |
|---|---|---|
| 3.0 | October 2017 | •AvesTerra API<br>•AvesTerra Toolkit<br>•AvesTerra Visualization Utility (AVU) |
| 3.1 | January 2018 | •Weather adapter<br>•Enhancements to API exception handling<br>•API bindings refresh<br>•Enhancements to AVU (interim) |
| 3.2 | April 2018 | •Climate adapter<br>•GMS adapter (Phase I)<br>•EOS adapter (Phase I)<br>•Ontology manager (Phase I)<br>•Uncertainy Model (Design)<br>•Provenance Model (Design)<br>•Python callbacks binding<br>•Enhancements to exception reporting<br>•Enhancements to AVU (Phase I) |

# *AvesTerra Roadmap (2018-2020)*

| Version | Release Date | Features |
|---|---|---|
| 3.3 | July 2018 | •GMS adapter (Phase II)<br>•EOS adapter (Phase II)<br>•Ontology manager (Phase II)<br>•Uncertainty Model (Implementation)<br>•Provenance Model (Implementation)<br>•Enhancements to AVU (Phase II) |
| 3.4 | October 2018 | •GMS adapter (Phase III)<br>•Ontology manager (Phase III)<br>•Enhancements to AVU (Phase III) |
| 4.0 | 2019 | •Server protocol plug-ins<br>•Server routing plug-ins<br>•Containerization |
| 5.0 | 2020 | •AvesTerra Analytic Language<br>•AvesTerra Analytic Environment<br>•ATra integration |

# *Questions and Discussion*

- Norman Kraft, Senior Software Engineer, Georgetown University (Norman.Kraft@georgetown.edu)
- Helen Karn, Research Specialist – Computational Sciences, Georgetown University (karnh@georgetown.edu)
- Steve Baird, Senior Software Engineer, AdaCore (baird@adacore.com)

# *AvesTerra Resources*

Web: https://avesterra.georgetown.edu/
Email: avesterra_admin@georgetown.edu
Sponsor: Office of the Senior Vice President for Research, https://osvpr.georgetown.edu

Browse the AvesTerra web site for:

- The Four-Color Framework *(April 2016)*
- Theoretical Framework *(May 2017)*
- AvesTerra Application Programming Interface (API) *(v. 3.2, May 2018)*
- AvesTerra Toolkit *(as of March 2018)*
- AvesTerra JSON Schemas *(as of February 13, 2018)*
- Hypergraph Transfer Protocol (HGTP) *(v. 1.4, May 2018)*
- AvesTerra Taxonomies and Ontology *(as of May 2018)*
- AvesTerra Roadmap *(last updated December 8, 2017)*

# *Acknowledgements*

This work discussed in this tutorial presentation was supported in part by the following funding to Georgetown University. The tutorial contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention (CDC), the Department of Health and Human Services, or the National Institutes of Health (NIH).

"Towards an Infectious Disease Insight Center: Privacy and Efficacy Pilot Study". Supplement to NIH Award 5U01A1034994.

"Privacy Data Sharing Tool to Support De-duplication of Cases in the National HIV Surveillance System (NHSS)", CDC Contract 211-2016-M-92074.

"De-duplication of Case Pairs in the National HIV Surveillance System Using the Black Box", CDC Project Grant NU62PS924580-01-00.

# *Additional references*

Ocampo JMF, et al. (2016) Improving HIV Surveillance Data for Public Health Action in Washington, DC: A Novel Multiorganizational Data-Sharing Method. *JMIR Public Health Surveill* 2(1):e3. doi: 10.2196/publichealth.5317

Smart J.C. (2016) Technology for Privacy Assurance. In: Collmann J., Matei S. (eds) *Ethical Reasoning in Big Data*. Computational Social Sciences. Springer.


GEORGETOWN UNIVERSITY