WALKTHROUGH & STATUS

2013 08 28

S. Yoakum-Stover, Ph.D.
M. Andrew Eick

Institute for Modern Intelligence
& Mission Focus

Move it forward. Make it happen.

CONTENTS

Move it forward. Make it happen.

WHAT IS THING 5?

Move it forward. Make it happen.

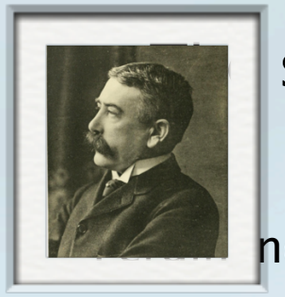
THING 5

A **semiotic** compute & storage apparatus,
built on an open-source **cloud** technology stack,
for the **storage** and **processing** of diverse data at scale,

including geospatial, temporal, human, social, cultural, behavioral,
as well as traditional Intel data types and
all modalities from documents to streaming video.

SEMIOTICS

The science which studies how meanings are made and how reality is **represented**

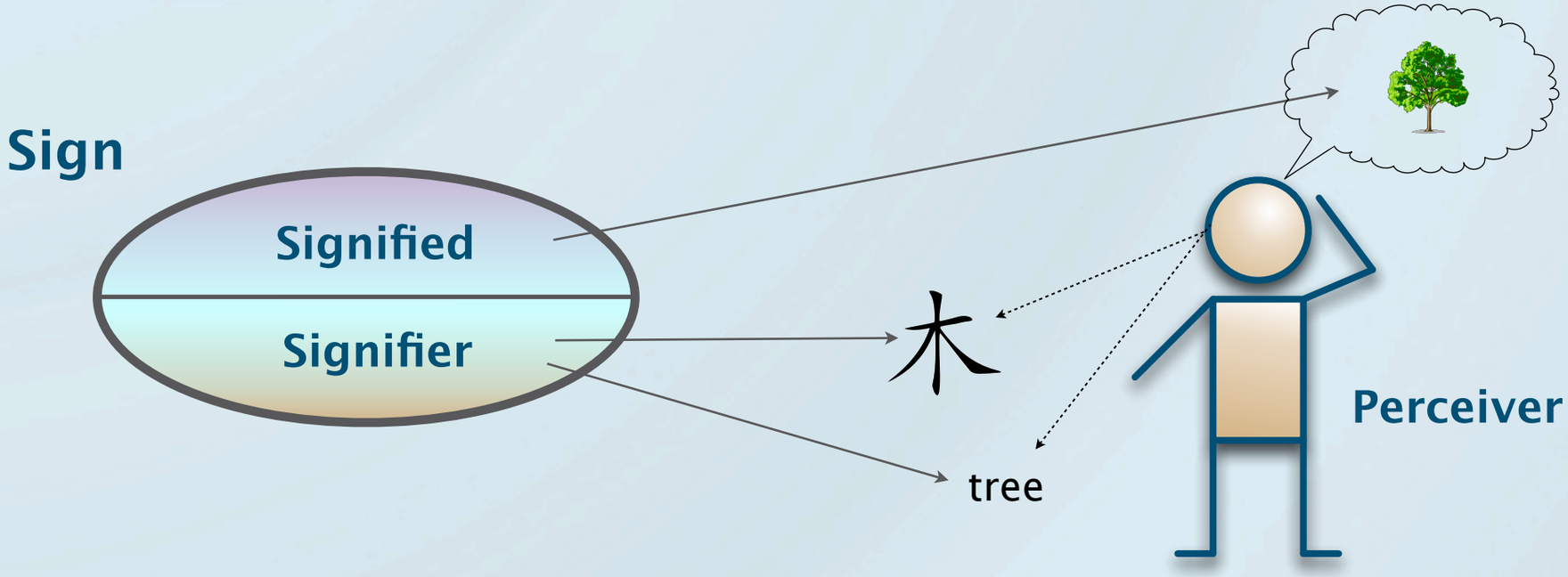


study of **signs** - anything which stands for something

Developed by

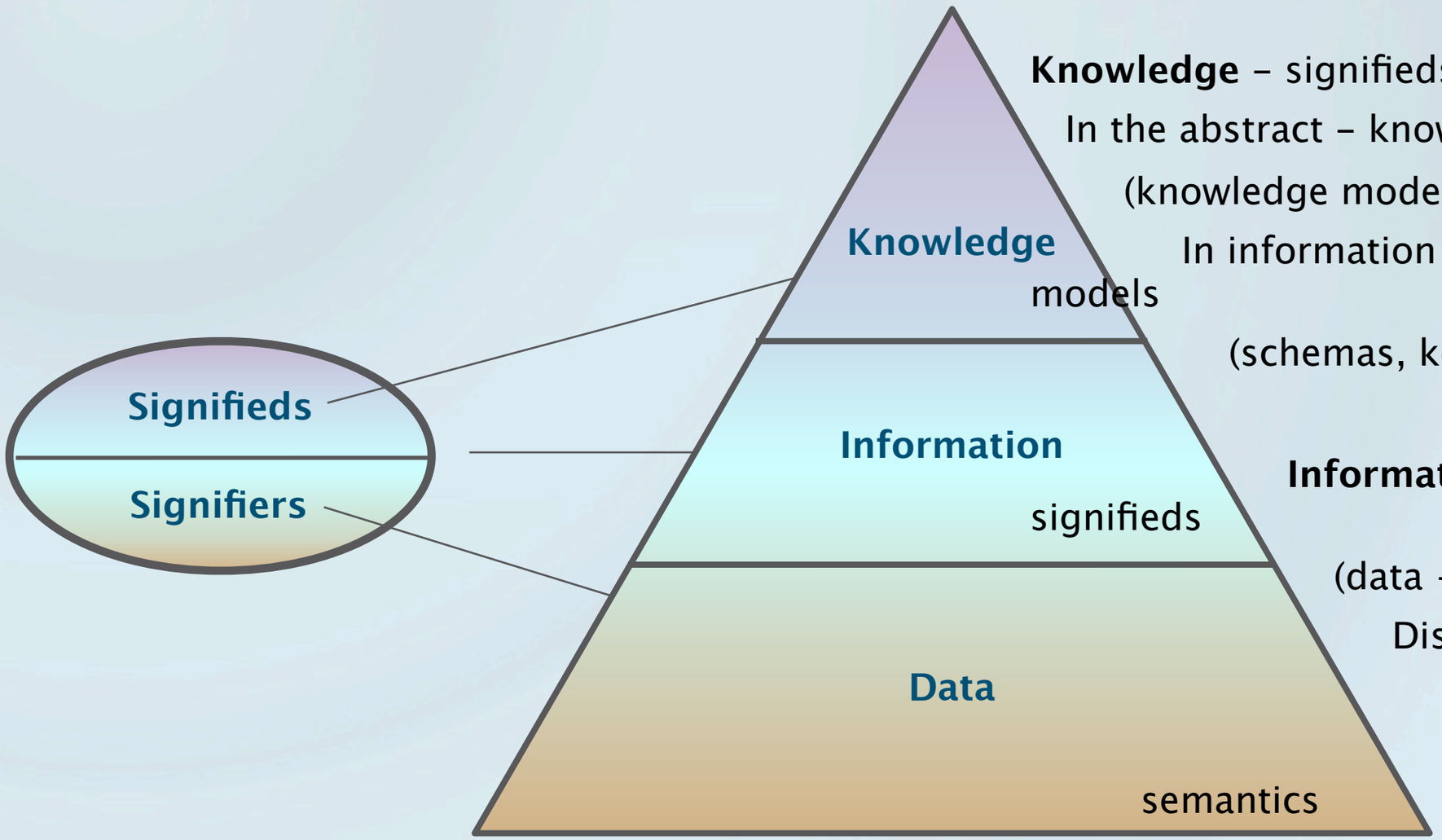
nd de Saussure (1857-1913) Charles Sanders Peirce (1839

Logic and Linguistics may be classified as branches of semiotics



To represent a thing, an information system must capture both signifier and signified

SIGN STORE



Knowledge – signifieds in relationship

In the abstract – knowledge representation (knowledge models)

In information systems – domain data-models

(schemas, keys, tags, ...)

Information – signifiers bound to signifieds

(data + semantics)

Disambiguation, Association

Data – signifiers lacking semantics

In typical information system, signifieds are presented as the storage-model (schema)

This approach restricts the scope to 1 perspective (set of meanings)

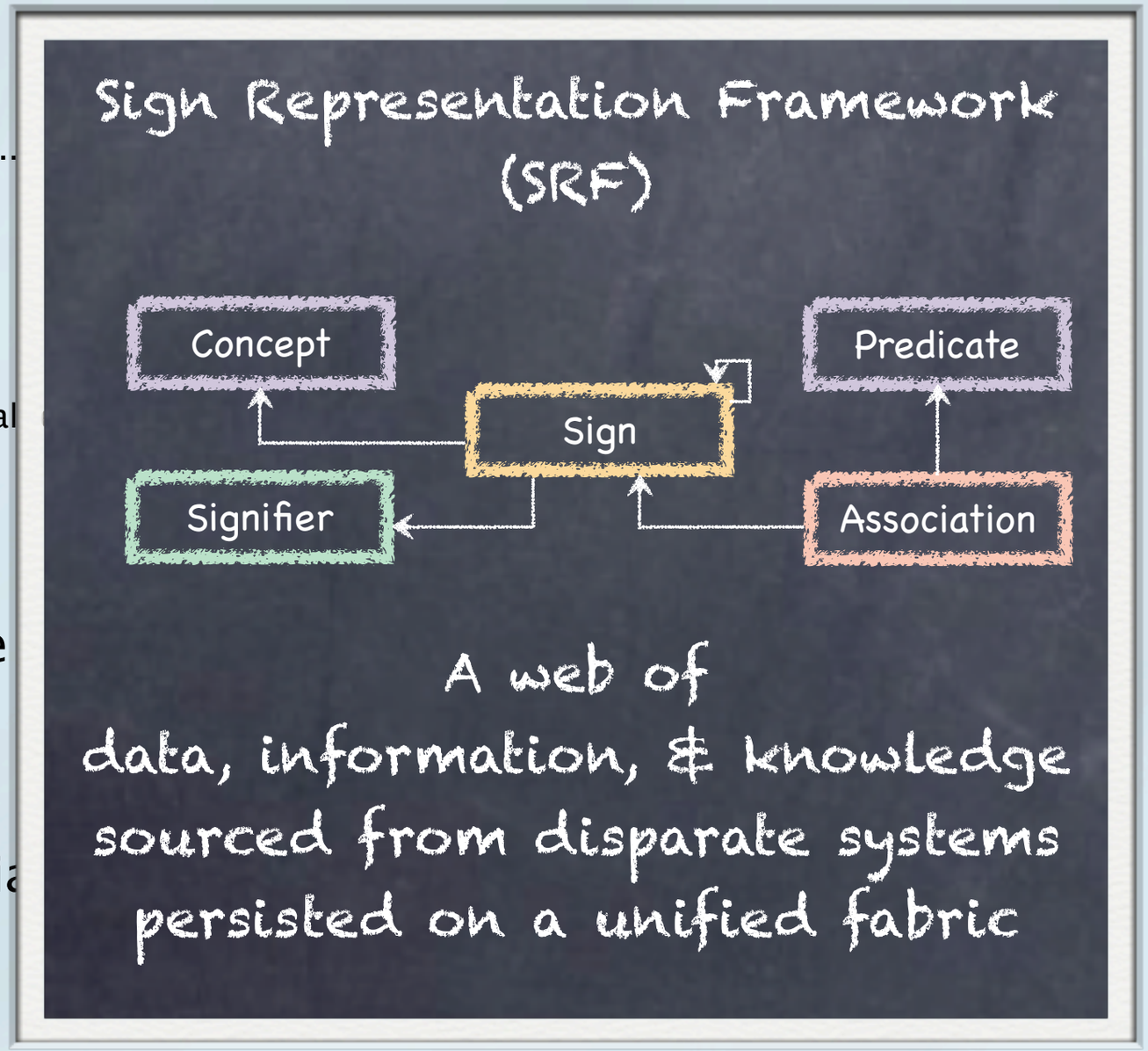
Thing 5 represents signifiers and signifieds as equal citizens

The storage-model is decoupled from the perspective

SIGN REPRESENTATION FRAMEWORK

SIGN STORAGE MODEL FOR ULTRA-LARGE SCALE

- Accommodate all, specify none
  - Any encoding, any modality
    - Video, images, audio, text, numeric series, ...
  - File standards of any kind
    - NITF, MP2TS, XML, IRC, email, HTML, ...
  - Structured data & data-models of all kinds
    - ShapeFiles, spreadsheets, metadata,...
    - Relational, object, hierarchical, graph, key-val
    - Taxonomies, schemas, ontologies, ...
  - Anything with a geometry / time
    - GIS features, events, ...
- Access all content - unified interface
- Ingest anything
- Preserve domain semantics
- Disambiguate semantically, geospatia temporally, contextually
- Harmonize data-models, or not
- Integrate information
- Build entirely new kinds of tools, analytics, applications



COMPUTE & STORAGE CLOUD

CHARACTERISTICS

- Internet-scale data store
- Massively parallel computation engine
- Tremendous aggregate bandwidth
- Commodity HW
- Open source SW

ACHIEVING SCALE

- Cloud achieves scale
  - Data size & processing performance
  - Marrying economy with technology
- SRF achieves another level of scale
  - Production and content management
  - Faithfully representing meaning and organizing diversity

SEMIOTIC APPARATUS (SEMAPP)

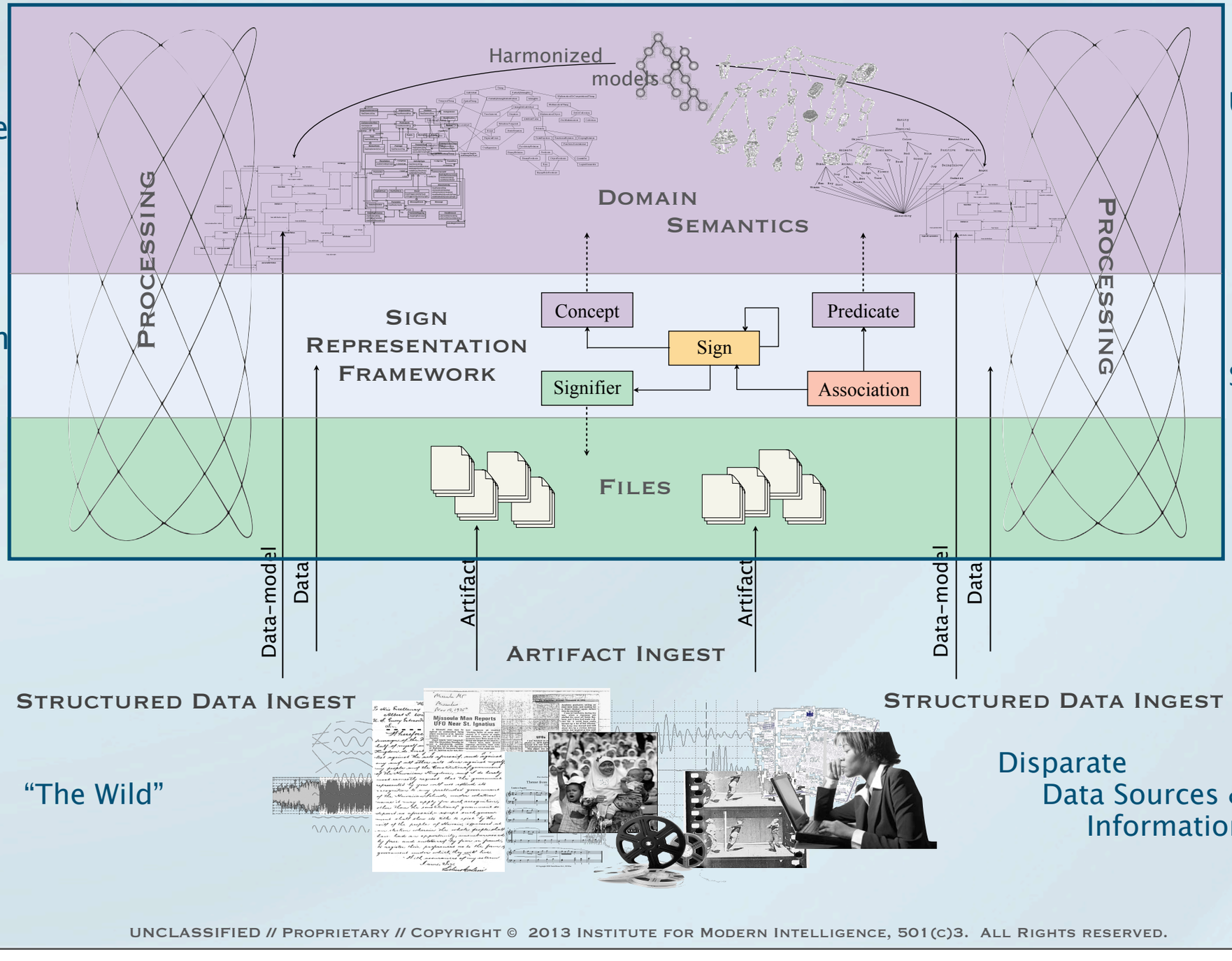
Move it forward. Make it happen.

SIGN-SPACE = SEMIOTIC APPARATUS + DATA + PROCESSING

Knowledge

Information

Data



Universal Store for Data-models

Universal Store for Structured Data

Universal Store for Unstructured Data

"The Wild"

Disparate Data Sources & Information Systems

BENEFITS & CAPABILITIES

BENEFITS

- No barriers to data ingest, its fast, simple, and universal
- All users & all processes can access all the data
- Disparate data are unified and operationalized without information loss or distortion
- Richness and meaning of information from the source is preserved
- Data-model harmonization is more powerfully supported, but not required
- True data integration across domains – connecting the dots – is enabled
- Entirely new kinds of analytics, tools, and applications become possible

CAPABILITIES

- Assert, retrieve, delete, tally, query data, information, knowledge
- Search – Keyword, semantic, geospatial, temporal
- Extract, characterize, expose – Make information more discoverable
- Connect – Assert new associations (e.g. social network analysis, registration)
- Surf – Follow associations within and across semantic domains
- Mine – Discover and expose new information (e.g. identification, tracking)

WHY'S IT CALLED THAT?

Move it forward. Make it happen.

In the beginning

CDR Kraft, Sir... So where do we fit in?

I got a handful of things I worry about

Thing 1 is NVS

Thing 2 is ...

You're Thing 5

Yes, Sir :-)

WHAT PROBLEM DOES IT SOLVE?

THE "DATA" PROBLEM

Move it forward. Make it happen.

OUR INTELLIGENCE ASSETS

DATA, INFORMATION, KNOWLEDGE, ANALYTICS, TOOLS, APPLICATIONS...

ARE FRACTURED

IN PHYSICALLY AND SEMANTICALLY DISPARATE SYSTEMS

OPERATIONS ARE IMPEDED

BY OUR INABILITY TO FULLY

SEARCH, EXPLORE, ENRICH, MANAGE, AND EXPLOIT OUR ASSETS

TO CRACK THE DATA / PROCESSING / GLOBAL SENSE-MAKING PROBLEM

THING 5

UNIFIES ALL DATA & SUPPORTS ALL PROCESSING

RELEVANCE TO NGA

Move it forward. Make it happen.

BOLD VISION WICKED BARRIERS

As long as we have the courage to
"**let the data surprise us**,"
I am convinced our analysts will lead to new discoveries—and not only by NGA.
We will **enrich the analysis of *others***,
especially the analysis performed by the **all source analyst.**
Letitia Long, Director, NGA

NATIONAL GEOSPATIAL-INTELLIGENCE AGENCY

Know the Earth… Show the Way… Understand the World

NGA Has a *Bold* Vision

Putting the Power of GEOINT in Your Hands

Provide online, on-demand access to our GEOINT knowledge

Broaden and deepen our analytic expertise to produce new value

"We will continue to deliver to our varied customer set what they need, when they need it, how they need it. But we have to be thinking about the future. We have to be continually pushing ourselves so that we do remain at the forefront."

– Letitia A. Long, Director, NGA

Approved for Public Release – NGA Case #13-153

1

BARRIERS

- The Data Problem
- The Processing Problem
- The Network Effect Problem

THING 5

BREAKS THE BARRIERS

Move it forward. Make it happen.

KNOW THE EARTH

MEANS KNOWING WHAT IS - The Physical & the Human terrain

FOUNDATION DATA - From imagery to custom / thematic maps

- Representing our human reality
- Telling the story of the earth and the dynamic things, natural and man-made, that operate upon it
- Foundation for analytic discourse

MAKING & MANAGING FOUNDATION DATA - Core to NGA's core domain

- Ravenous and impatient data consumer
- Starving and languishing amidst troves of data, information, and knowledge
- Something is wrong

THE DATA PROBLEM - Diversity, ambiguity, scale

- How meaning is represented – knowledge representation and semiotics
- Domain models and schemas built into every information system
- Not just more data, profoundly different – really, Really Big Data
- Inability to access, understand, manage and use our data assets as a coherent whole

THING 5

- Accommodates diversity, reduces ambiguity, and handles scale
- Access, understand, manage and use our data assets as a coherent whole
- Feed the production chain with rich, diverse information
- Custom / thematic maps as linked data using information from any source

SHOW THE WAY

MEANS NAVIGATING THAT TERRAIN SAFELY

- Tactical level of capability that a local understanding enables

ANALYTICS

- Modeling, simulation, what-if analyses, …
- Operationalizing the data
- Conduct of analytic discourse, a man-machine partnership

MAKING & USING ANALYTICS - Essential to NGA's core domain

- Scattered across programs and organizations
- Competing for resources and saddled with overhead
- Something is wrong

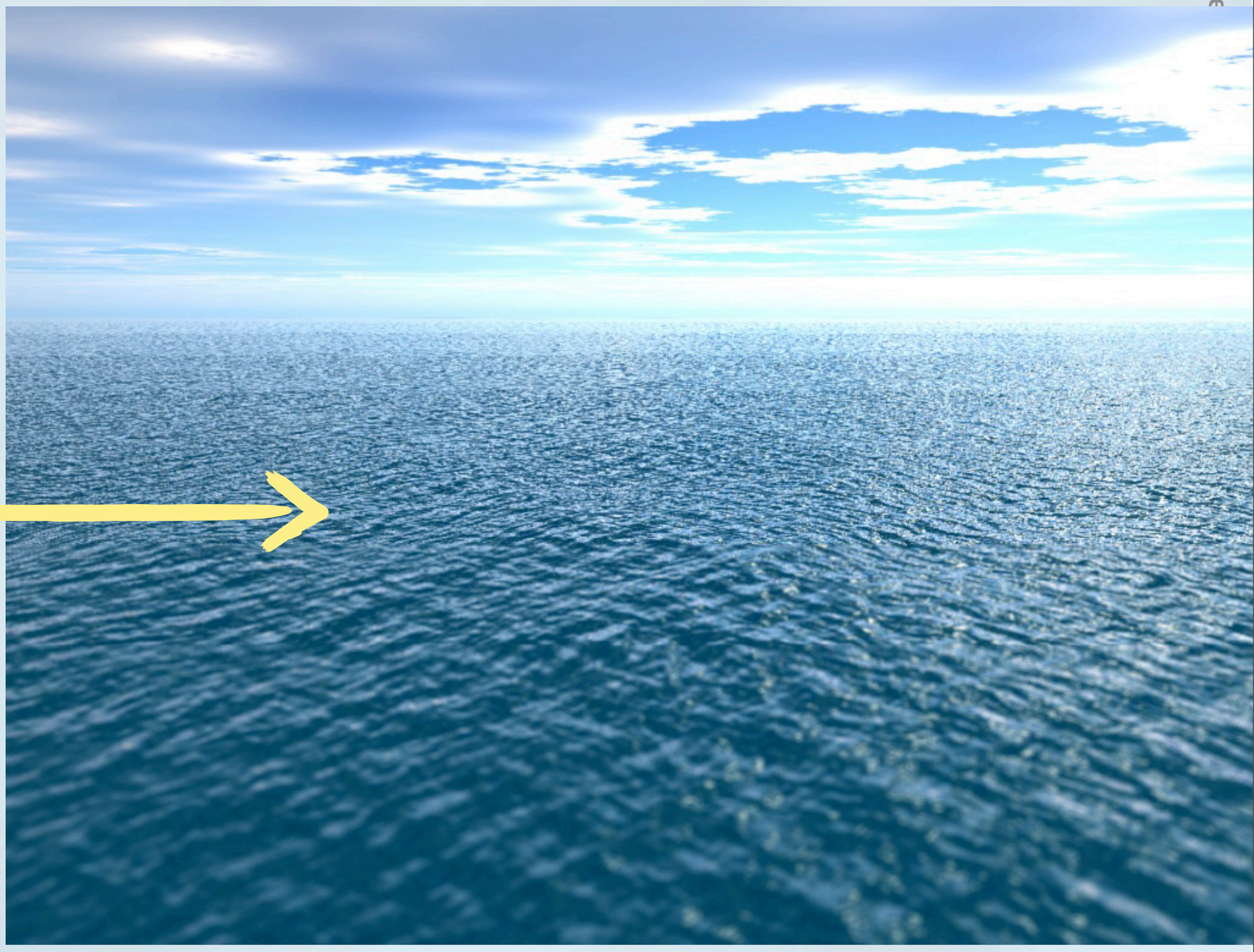
THE PROCESSING PROBLEM - The pipeline model

- Data flowing between processes
- Data and processing assets are different at every point in the enterprise
- Bandwidth-limited & complex
- Inability to leverage, understand, manage and apply our analytic assets as a coherent whole

THING 5

- Replaces the pipeline model with a wave model – analytic waves over data at rest
- Leverage, understand, manage and apply our analytic assets as a coherent whole
- All analytics access and enrich a shared ocean of data, information, knowledge
- Simple, robust, efficient, powerful, creative

PIPELINES TO WAVES



UNDERSTAND THE WORLD

MEANS KNOWING WHAT IT ALL MEANS

- A strategic level of capability that a global understanding enables

INTERCONNECTED DATA - Interconnectedness at all scales

- Putting the pieces together – a living web of information
- Sharpening the saw – adjusting data and analytic assets in anticipation
- Product of analytic discourse

MAKING & USING INTERCONNECTED DATA - Essential to NGA's core domain

- Shared objective across the IC
- Struggling despite troves of data, information, knowledge, and analytics
- Something is wrong

THE NETWORK EFFECT PROBLEM

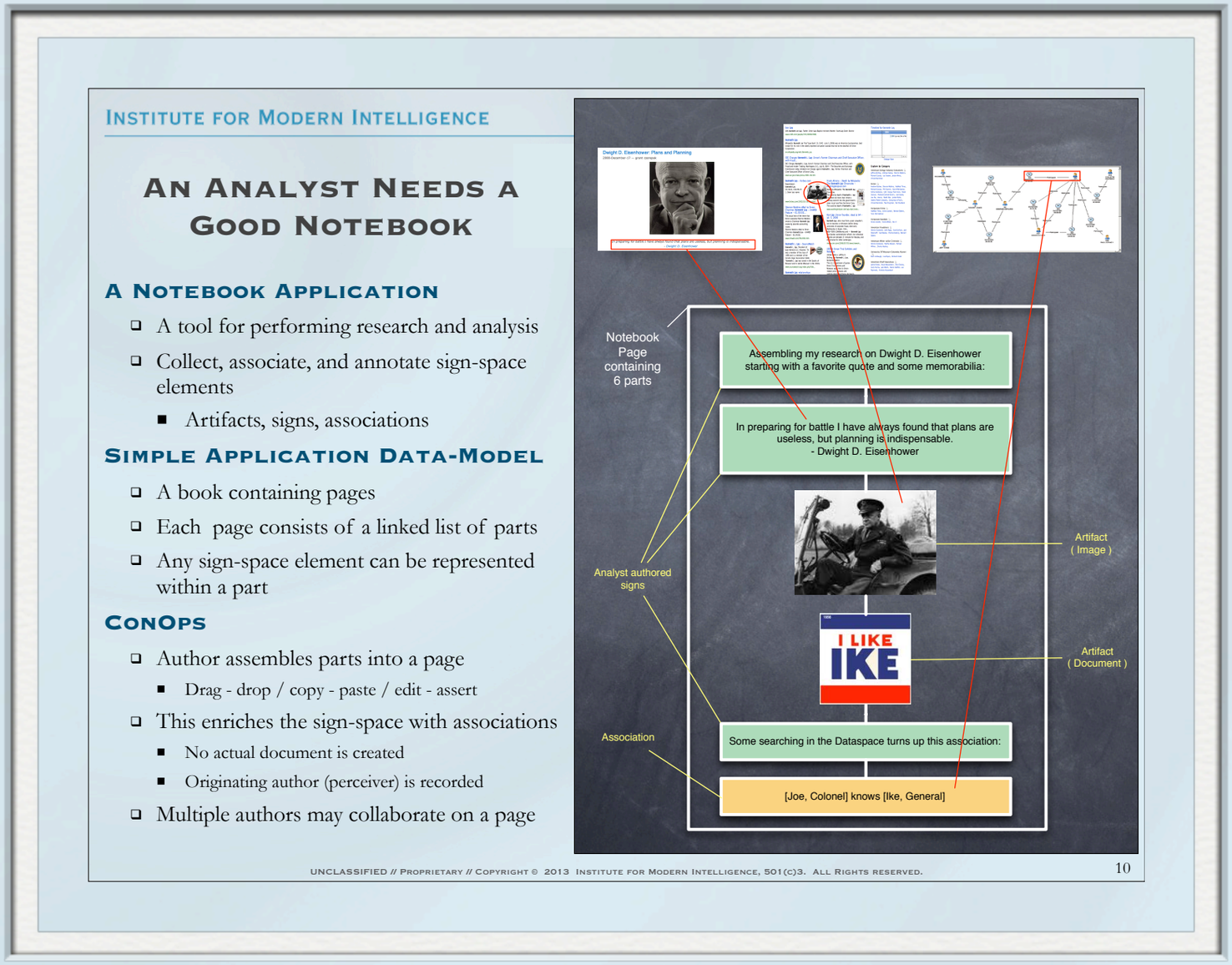
- Not compounding (each contribution increases the value of the whole and vice versa)
- No feedback mechanism whereby output (product) enriches the input
- Inability to fully benefit from achievements and function strategically

THING 5

- A rich, living web of information naturally emerges
- Each contribution increases the value of the whole and vice versa
- Products are an enrichment of the web
- Fully implement Structured Observation Management
- Achieve and apply new levels of understanding to achieve mission objectives and make NGA better

PRODUCTS AS INTERCONNECTED DATA

- A product consists of
  - Selected elements of information (signs and associations)
  - Linked together (associations)
  - Displayed by an application to suit a particular purpose
- Examples
  - Notebook
  - Thematic map
  - Social network
  - Geospatial track
- Information elements can appear in any number of different products
- Product creation enriches the web of interconnected data
- Nothing to “re-ingest”
- Every product makes the information more valuable



INFORMATION WEB GROWS AS A SIDE-EFFECT OF ANALYTIC WORK

SPACE & TIME

SPACE & TIME

- The physics that makes NGA and Thing 5 special
- The bedrock of the IC
- The unifying threads throughout our data

THING 5 CORE

- Semantic disambiguation
  - Binding of a element of **data** (the signifier) to a **concept** (the signified)
  - Forms a Sign – the atom of semantics in Thing 5
  - e.g. [systover@imintel.org, Sender]
- Geospatial disambiguation
  - A Sign that also has a **geometry** (representing geospatial coordinates) e.g. [Nile, River, 30°10'N 031°06'E]
  - Isomorphic to a GIS feature
  - Applies to associations as well
- Temporal disambiguation
  - A sign that also has a **period** (representing temporal coordinates) e.g. [Wall Street bombing, Terrorist Act, 19200916]
  - Isomorphic to an event
  - Applies to associations as well

POWER

- Native GEOINT support
- Ideal platform for ABI
- Search, explore, discover, enrich, manage, exploit these unifying threads throughout our data
- IC solution for GEOINT storage and processing

DEVELOPING THE SCIENCE, PRACTICE, AND GOVERNANCE OF MODERN INTELLIGENCE

SEMAPP GEMS

UNIQUE SOLUTIONS & PERSPECTIVES

MISSION FOCUS

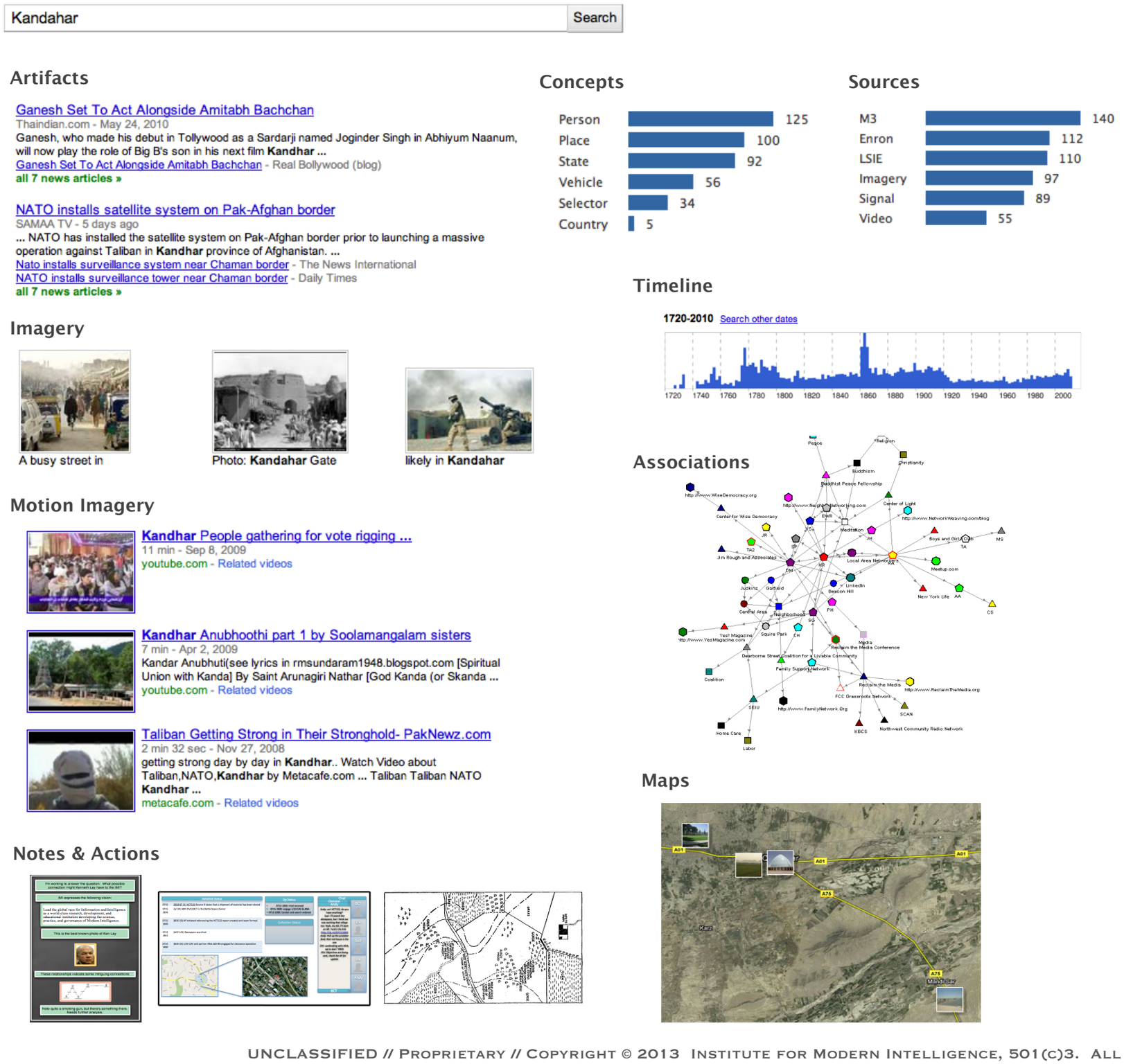
Move it forward. Make it happen.

Unique solutions & perspectives

APPLICATIONS

Move it forward. Make it happen.

SEARCH-SURF WEB APP



A SIGN-SPACE NEEDS A GOOD SEARCH ENGINE

- Familiar Google-like interface
- Ask questions of the data
- Boolean and semantic search over the entire Sign-Space
- Navigate hyperlinks to surf associations
- Drill down to originating artifacts and sources
- Quickly locate rich analysis products
  - Notebook pages
  - HorseBlankets
  - Enhanced audio, imagery, and video
- Mouse-over for

Move it forward. Make it happen.

An Analyst Needs a Good Notebook

A Notebook Application

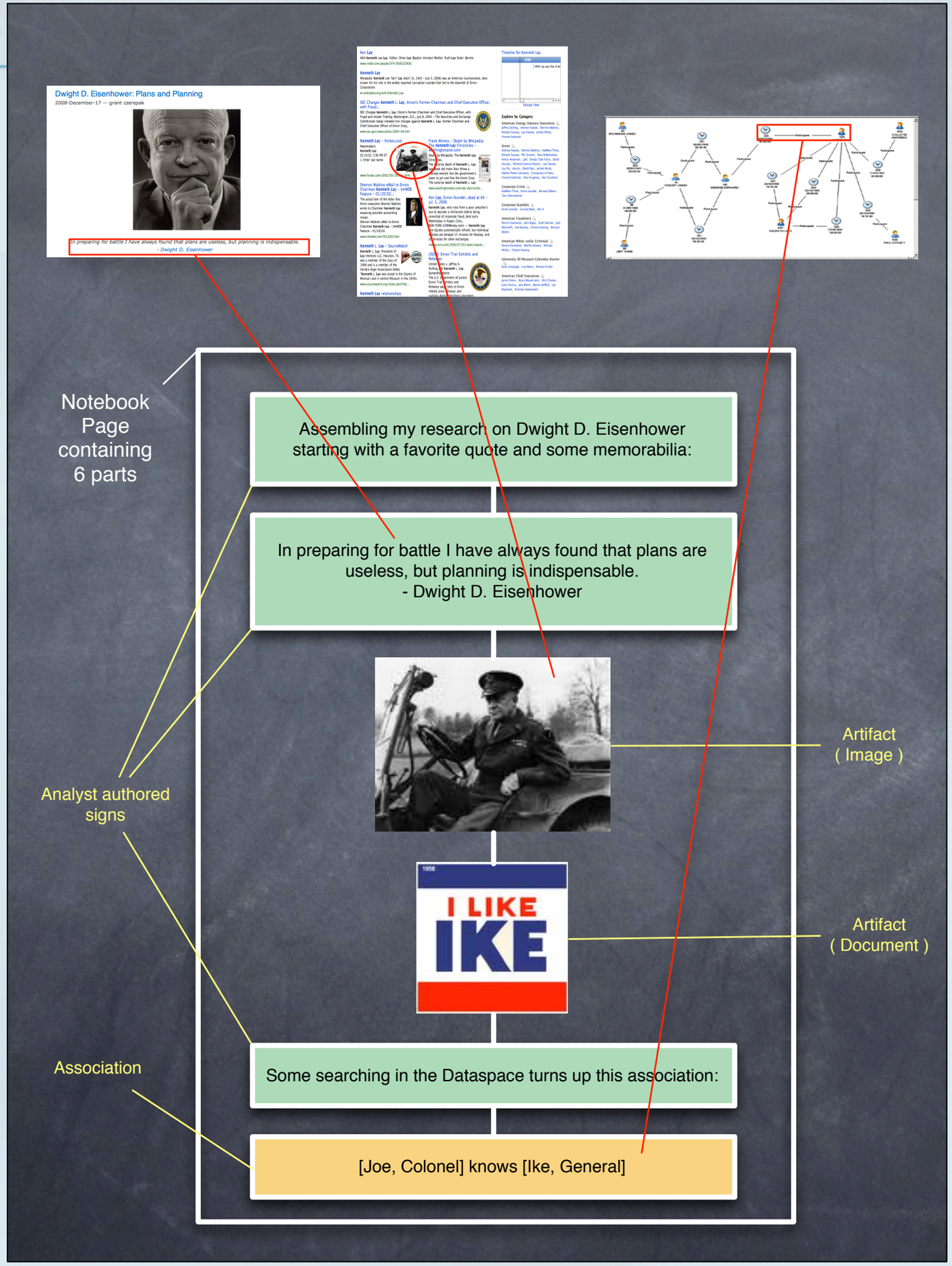
- A tool for performing research and analysis
- Collect, associate, and annotate sign-space elements
  - Artifacts, signs, associations

Simple Application Data-Model

- A book containing pages
- Each page consists of a linked list of parts
- Any sign-space element can be represented within a part

ConOps

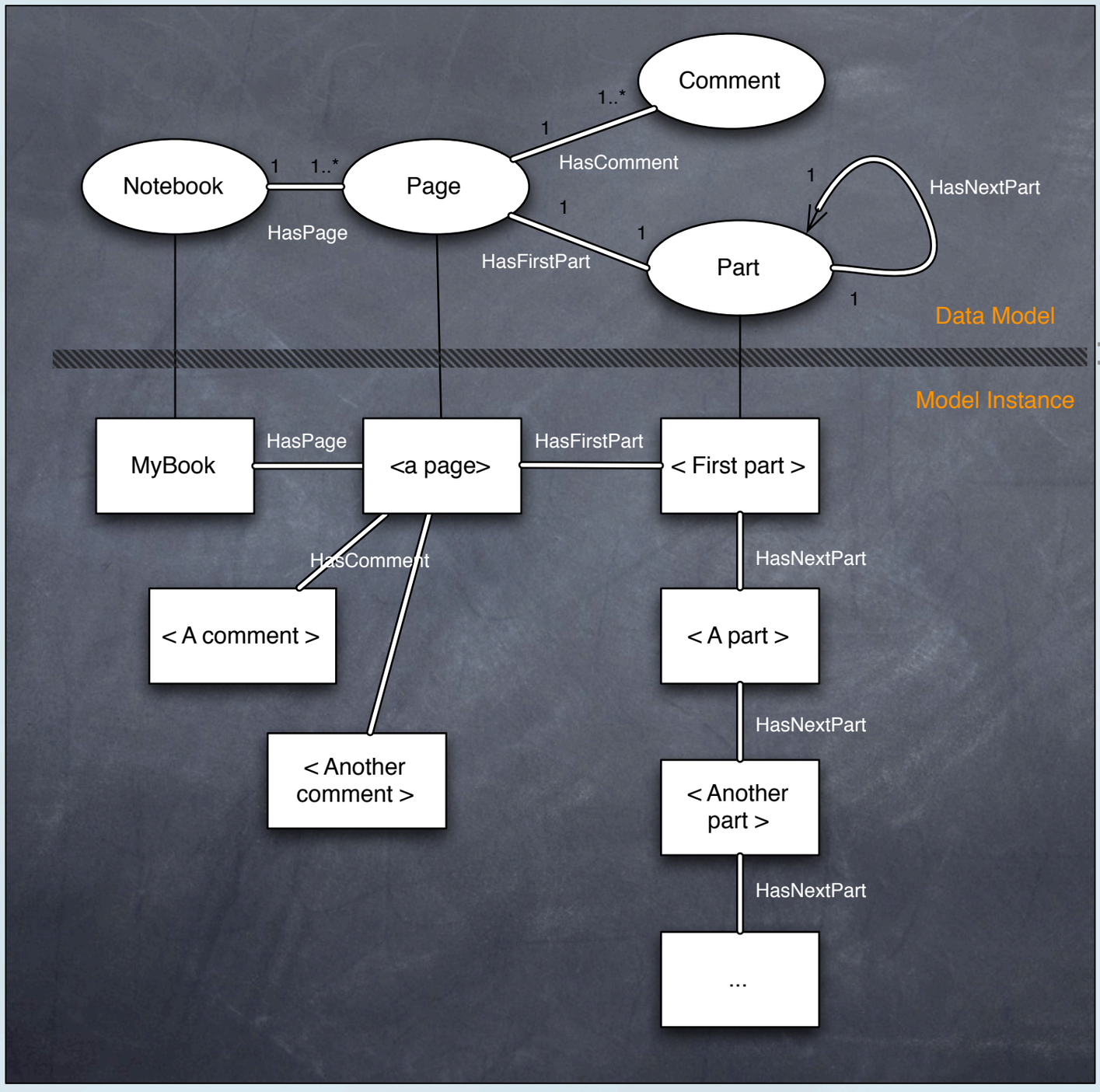
- Author assembles parts into a page
  - Drag – drop / copy – paste / edit – assert
- This enriches the sign-space with associations
  - No actual document is created
  - Originating author (perceiver) is recorded
- Multiple authors may collaborate on a page



NEW SPACE TO WORK IN

APPLICATION DATA-MODEL

- A "Container" Model
  - Imposes minimal semantics
  - Parts serve as containers for sign-space elements
  - Elements in containers carry their original domain-semantics
- A Mash-Up of Semantic Elements
  - Not data-model harmonization
  - Not exactly data integration
  - More than a visual mash-up
  - A persistent, information mash-up
- New Species of Application
  - Enabled by the unified interface
  - Operating on information absent any data-model harmonization
  - A new space in which to work with data!



SIGN-SPACE EVOLVED VIDEO APPLICATION

Move it forward. Make it happen.

ENABLES AN ANALYST TO ENRICH VIDEO WITH

- Other contextual data from the Sign-Space
- Annotations including text, images, voice, and telestration

ARCHITECTURE VERY SIMILAR TO SIGN-SPACE NOTEBOOK

- A container model with parts associated by **time**
  - SSEVA Stream has a MainStream and PartStreams
  - PartSteams have start, length, and placement relative to MainStream
  - PartStreams may contain any Sign-Space element
- Dropping parts into a SSEVA flow automatically creates PartStreams relative to the MainStream
- The SSEVA engine knows how to displays SSEVA streams

CONTEXT MATTERS

Move it forward. Make it happen.

CONTEXT

THE INFORMATION ENVIRONMENT IN WHICH A THING MAKES SENSE

- Context is a kind of Thirdness (Aristotle, Kant, Hegel, Pierce, Husserl, Whitehead, ...)
  - Triad of primary categories of the kinds of things that are
    - Firstness – Existence independent of anything else
      - Characterized by intrinsic qualities (e.g. systover, Mission Focus, D&B)
    - Secondness – Being relative to something else
      - Characterized by dyadic, directed relations between entities (e.g. mother, attorney, wife, employee)
    - Thirdness – A mediation whereby a multiplicity of entities are brought into relation, creating new relationships
      - Characterized by a nexus of related entities (e.g. motherhood, legal system, marriage, business enterprise)
- Examples: Organizational, cultural, economic, religious, technical, ...

CONTEXT IN INFORMATION SYSTEMS

- A data-model implicitly defines a domain context
- Beyond the data-model, context is almost never represented
- Using information outside it's originating domain without preserving the context is perilous

CONTEXT IN THE SIGN-SPACE

- SRF captures firstness, secondness, and thirdness
- Context is represented at several levels
  - Structured data retain their native domain-model associations
  - Data retain their source context (source metadata)
  - Information extracted from artifacts retains the artifact context (mentions)
  - Context may be explicitly represented as signs in the SRF
  - Contextual models may be explicitly represented in SRF

Move it forward. Make it happen.

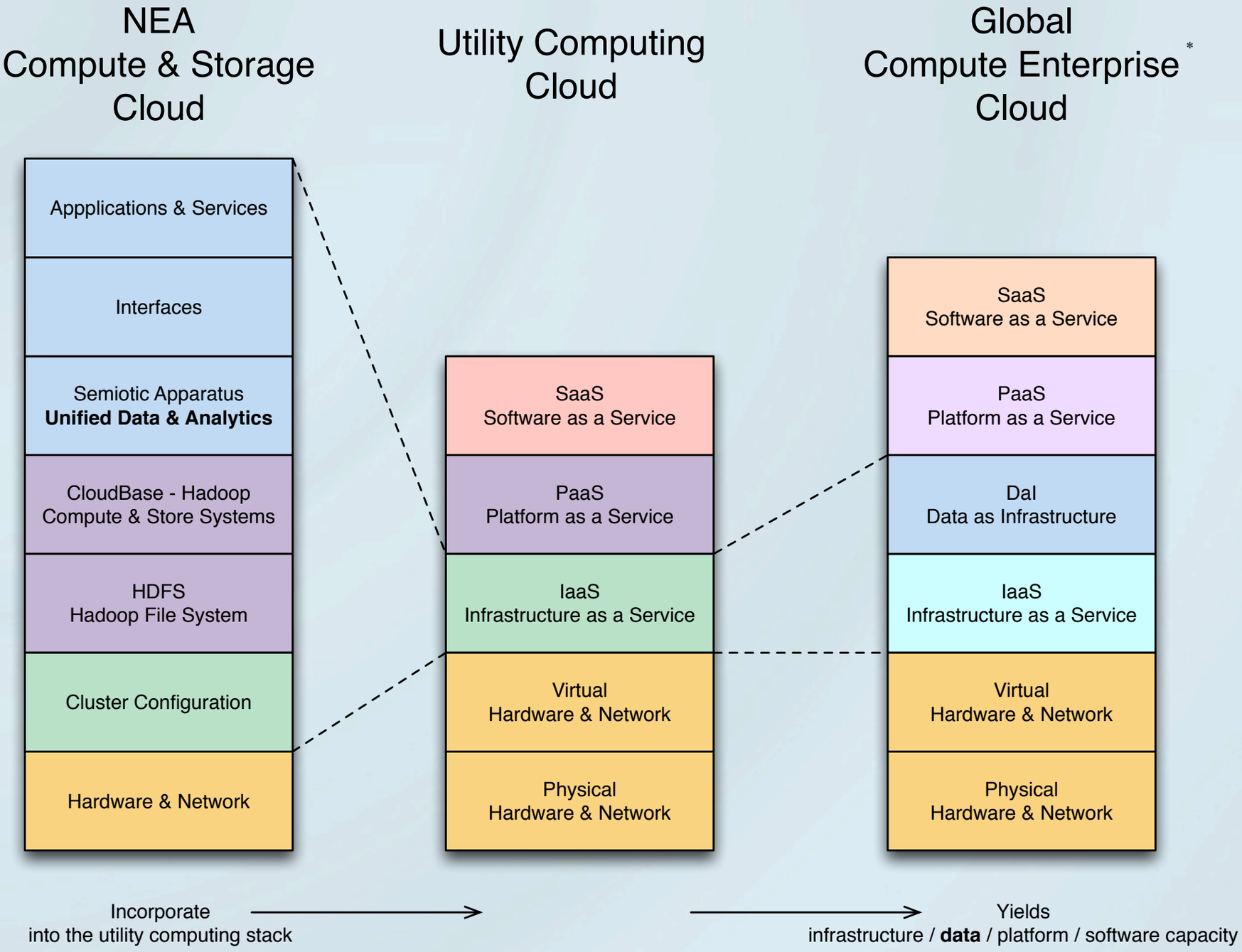
DEVELOPING THE SCIENCE, PRACTICE, AND GOVERNANCE OF MODERN INTELLIGENCE

DATA-AS-INFRASTRUCTURE

Move it forward. Make it happen.

NEW CAPABILITY → NEW CAPACITY

Move it forward. Make it happen.



* **Global Compute Enterprise** white paper, T. Myers et al.

NEW CAPACITY FROM NEW CAPABILITY

BROADEN NGA'S VISION OF UTILITY COMPUTING BY PROVIDING

"DATA-AS-INFRASTRUCTURE"

(DATA & PROCESSING AS A UBIQUITOUS RESOURCE)

- How
  - Incorporate the SemApp into the infrastructure layer of the GCE
    - Hook platforms to the Thing 5 cloud at the infrastructure layer
    - Expose cloud data and processing via APIs that hide the underlying transport and storage mechanisms
- Result
  - SemApp becomes an open commodity accessible by default
  - Using and contributing to the unified data assets becomes easy by default

As part of the Data-as-Infrastructure strategy,
Thing 5 transforms the "data problem" into a National Asset
&
Achieves the IC's goals for data sharing
as a matter of course

DATA-MODELS & THE TROUBLE WITH DATA

Move it forward. Make it happen.

DATA-MODELS

SPECIFY

- Vocabulary, Structure, Semantics, and Constraints

BUILT FOR A PURPOSE

- Make a particular analytic, tool, application work efficiently
- Entail a particular perspective on the data
- They're all different for a reason

PROVIDE THE INTERFACE TO AND PERSPECTIVE ON THE DATA

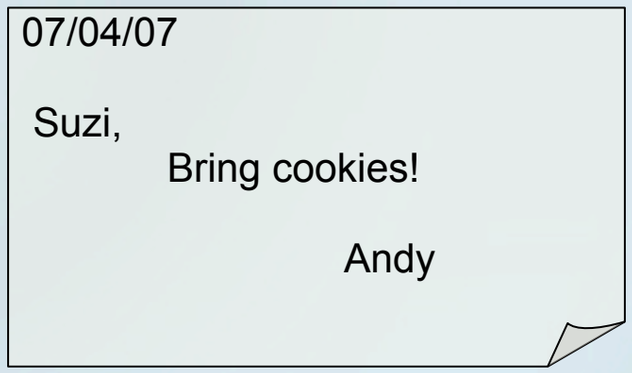
- All interactions with data are mediated by a data-model
  - We go thru the data-model to ask questions about the data
    - What is the **name** of the **Student** with **sid** 53666?
    - How many **Students** have **age** = 18?
    - What is average **gpa** of **Students**?
  - We can not directly question the dat
    - What is "Jones"?
    - Is there anything on "Guldu" ?
    - What is "53831" related to?
  - Indexes are no exception
    - Every index reflects a hidden data-m
    - The "things" being indexed
    - e.g. Index on **Students name**

Students

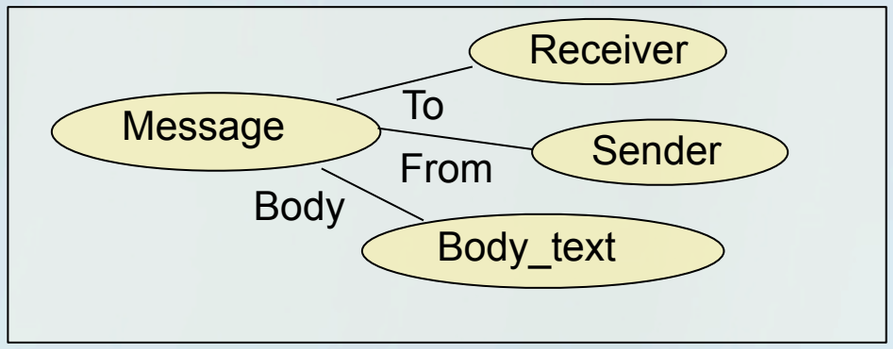
sid	*name*	*login*	*age*	*gpa*
53666	Jones	jones@cs	18	3.4
53688	Smith	smith@ee	18	3.2
53650	Smith	smith@math	19	3.8
53831	Madayan	madayan@music	11	1.8
53832	Guldu	guldu@music	12	2.0

Move it forward. Make it happen.

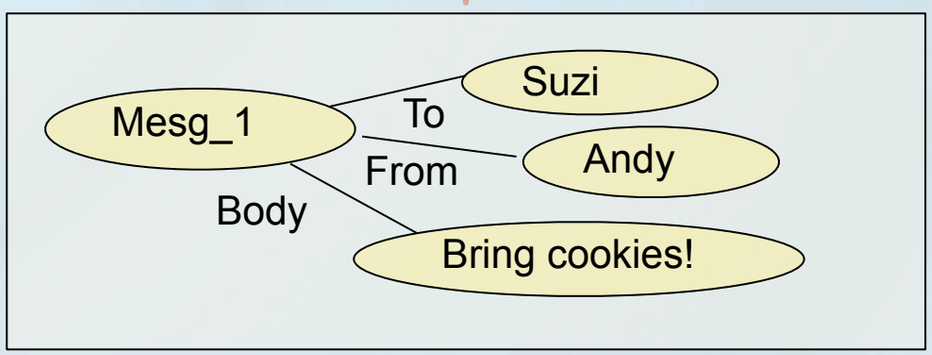
DATA-MODEL IN ACTION (RDB-STYLE)



(a) Unstructured Data



(b) Data-model



(c) Structured Data

Message	Receiver	Sender	Body_text
Msg_1	Suzi	Andy	Bring cookies!
…			

(d) Typical database structure

The data-model is imposed on the data store (schema)

The data values are frozen beneath it (e.g. We can't say new things about Andy)

Once loaded, it's painful to modify

Introducing data from other sources is hard as data-models must conform to the storage structure

Once you do that however, the data is "integrated" (but according to only 1 perspective)

Information is typically lost /

THE PROBLEM WITH DATA

is…

It Has to Go Somewhere and often goes Lots of Places

Then it Tends to

Stick where it Lands and take the Shape of its Container

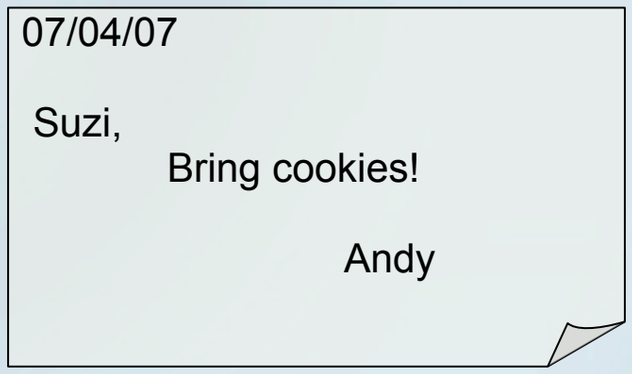
In other words,

The Data-Model gets

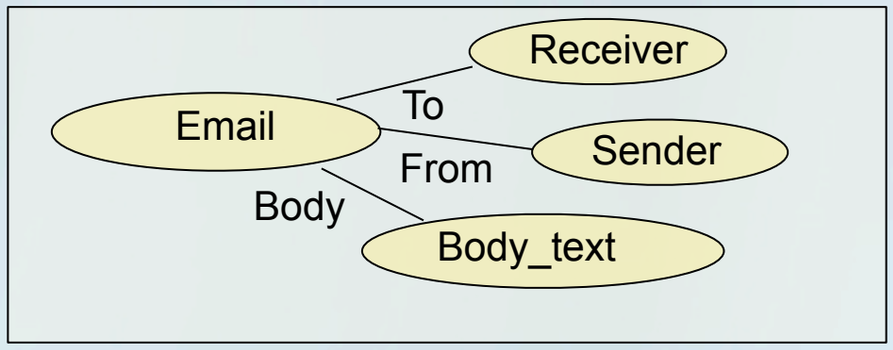
Imposed on the Data Store and the Data is then Frozen Into it

So Don't do That

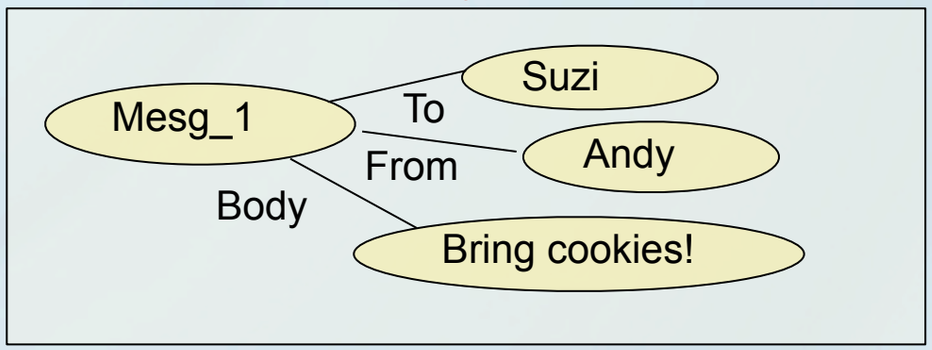
DATA-MODEL IN ACTION (CLOUD-STYLE)



(a) Unstructured Data



(b) Data-model



(c) Structured Data

The data-model is still there, but it's mingled with the data

We still need to know what it is to access our data

We also need to know how data

Row	Family	Qualifier	TimeStamp	Value
Mesg_1	Email	Sender	20090224123	Andy
Mesg_1	Email	Body_text	20090224125	Suzi, Bring cookies!

(d) Typical CloudDB structure (e.g. BigTable)

Introducing data from other sources is easy

Keeping track their data-models (interfaces) is hard

The data is in not integrated

THE PROBLEM WITH CLOUD

is…

The Data Structure has to GO SOMEWHERE

AND CAN GO

almost **Anywhere**

So

Data flows Easily INTO its Container

BUT LOSES ITS SHAPE

In other words,

The Data-Model gets

Mixed with the Data in a Specific but Arbitrary way

So Don't do That

DATA INTEGRATION CONUNDRUM

SITUATION WE HAVE:

A data-model Provides

The Interface to the data AND the Semantics of the data

For data to be integrated it must have a Unified interface

Unified Interface ⇒

A single data-model

A single meaning / perspective

But there is no single right way to represent all Knowledge ⇒

There can be no single data-model for all intel

SITUATION WE WANT:

A unified interface to All the data

that

THE SOLUTION

Move it forward. Make it happen.

CONSIDER **DATA-MODELS** FROM A

Higher Level of Abstraction

Distill from there a

Minimal set of Elements Sufficient to Capture

Any data-Model

and then

Build Storage Model based On that That

GEOSPATIAL INDEXING

Move it forward. Make it happen.

GeoHash Overview

Binary Decomposition Scheme Representing lat-longs by recursively dividing the associated angular ranges ([-90, 90] , [0, 180]) in half

Bit value of 0 indicates the lower/left half. Bit value of 1 indicates the upper/right half

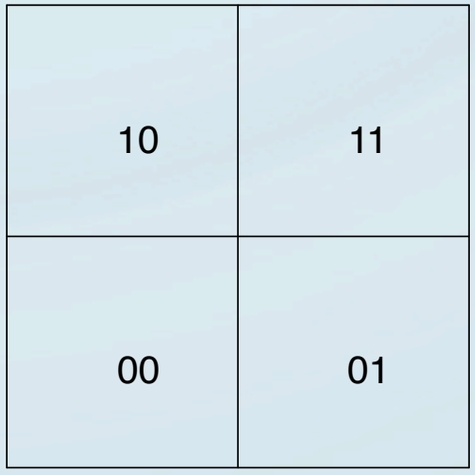
0111110000000 (long), 1**01111**001001 (lat)

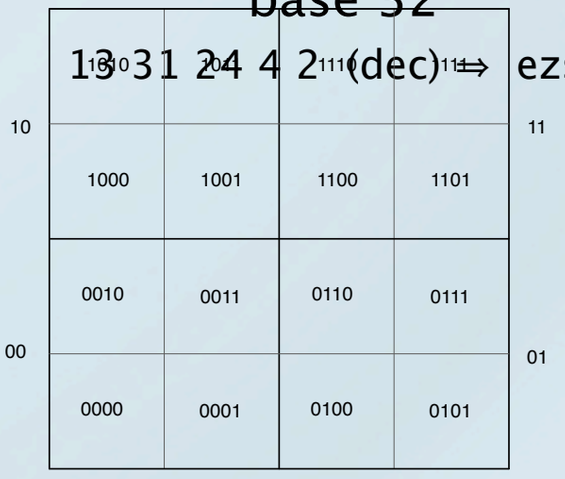
The two lat-long bit sequences are interleaved from left to right with even bits taken for the longitude

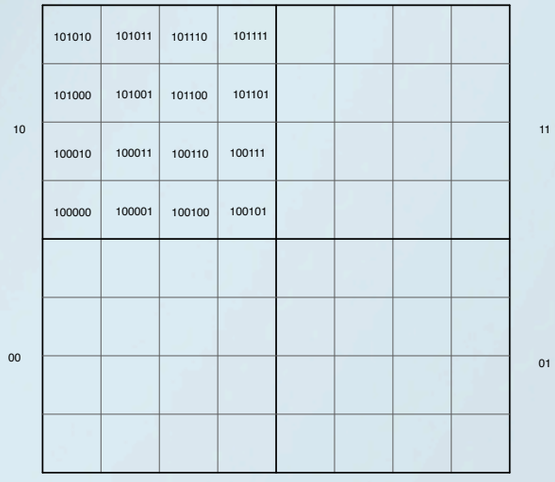
01101 11111 11000 00100 00010 (bin)

GeoHash results by grouping the sequence of bits by 5 and representing each group in base 32

13 31 24 4 2 (dec) ⇒ ezs42



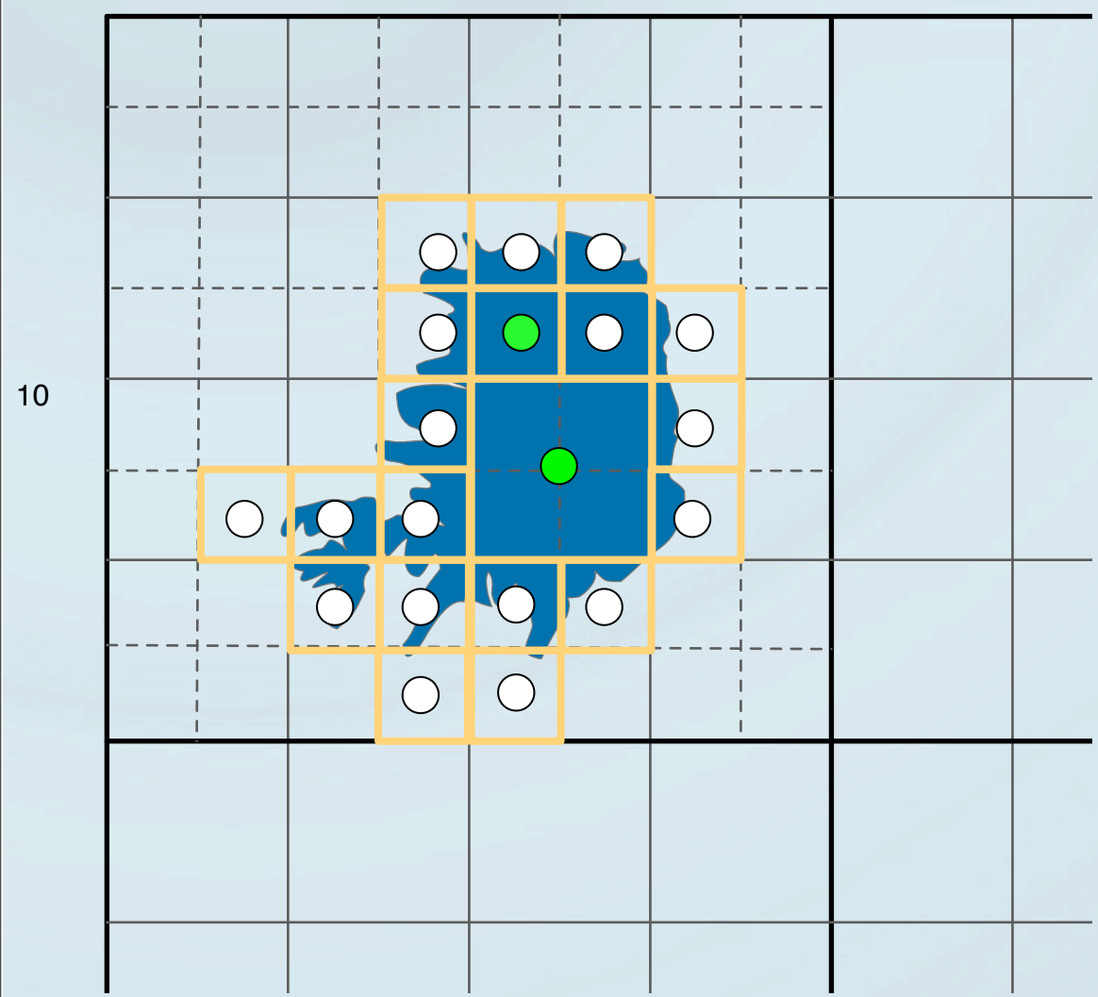




Characteristics

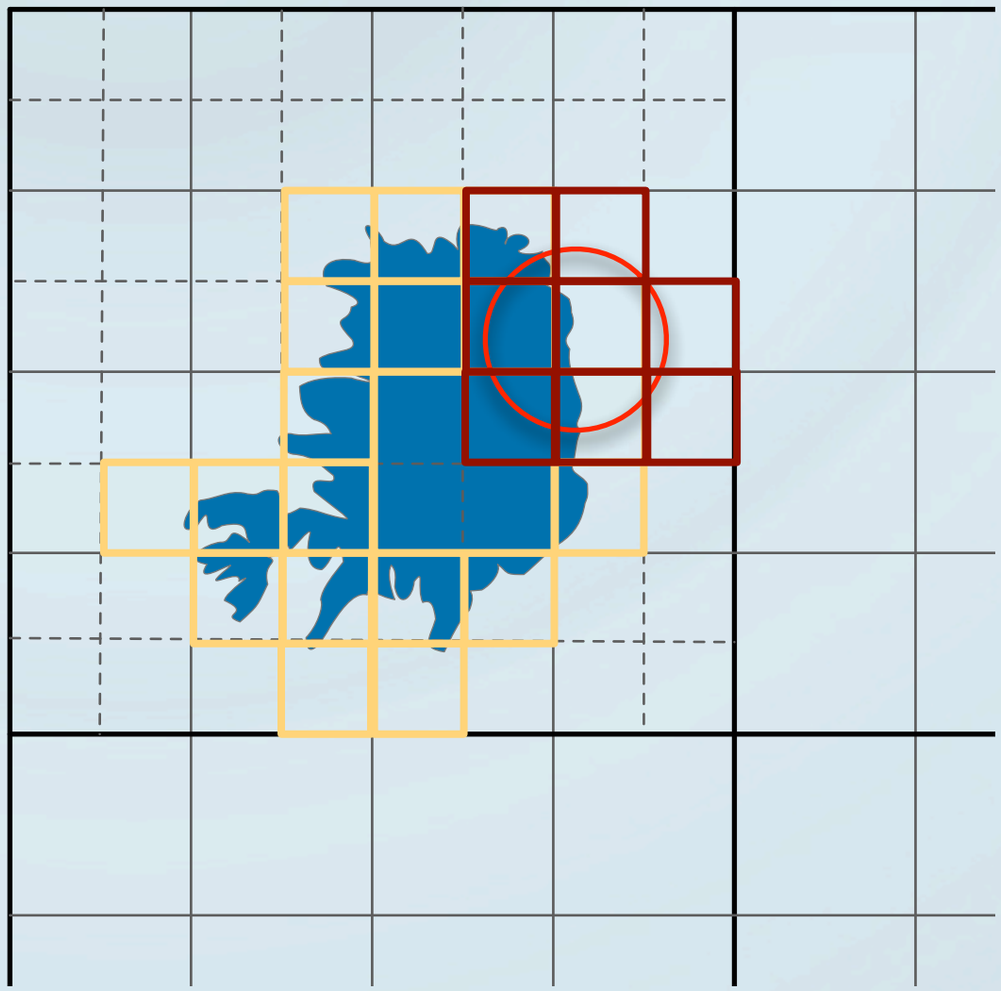
Any level of precision can be obtained by increasing the number of bits
Resolution can be dialed down or up by ignoring or not ignoring additional bits

INDEXING A FEATURE



- If the feature is a point, simply index it at the highest resolution GeoHash (GH) that contains it
- Otherwise, find the GHs fully covering the feature (10)
- For each such GH, evaluate it's children's relationship with the feature geometry recursively
  - If disjoint, return
  - If fully containing, index the feature at this GH, noted as "inside"
  - If at the max depth below the start GH, index the feature at this GH, noted as "related"

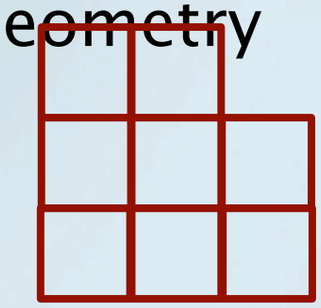
QUERYING GEOMETRY



10

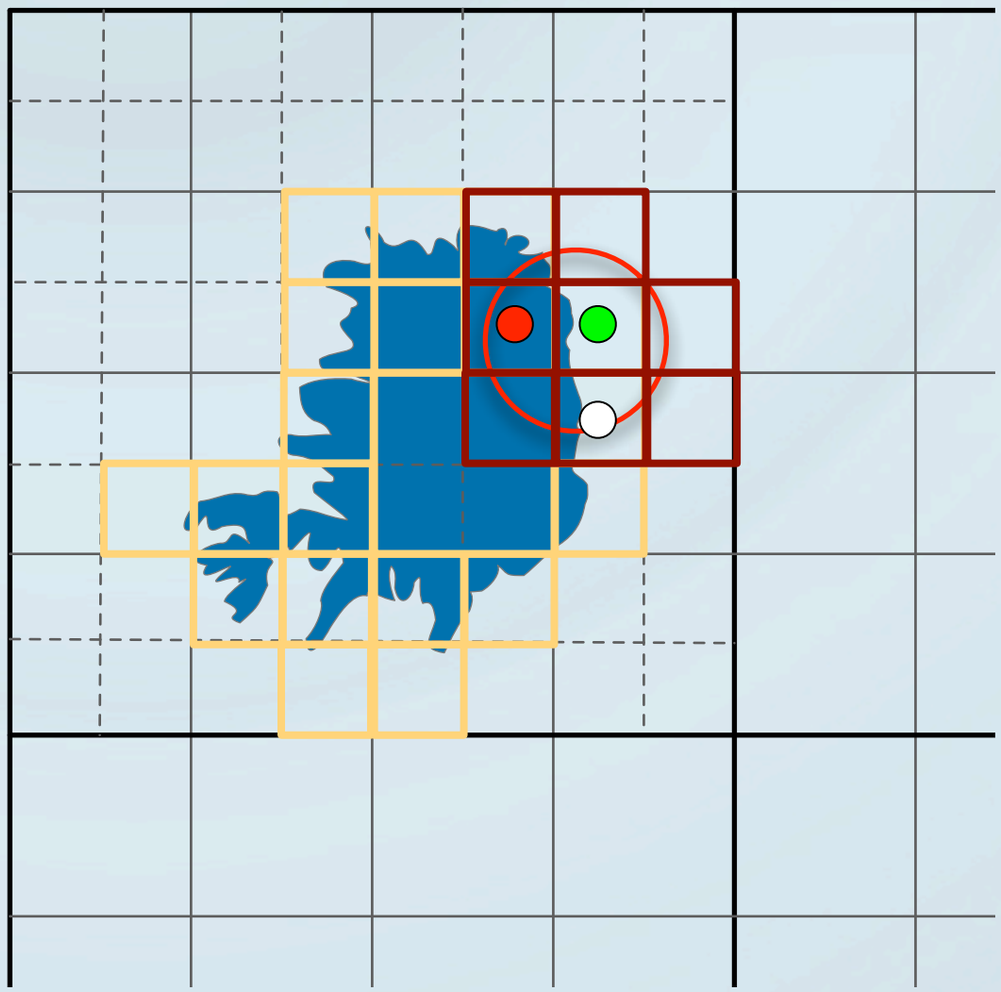
- Specify the query geometry

- Find GeoHashes completely inside or intersecting the query geometry



- Result is the set of query relevant GHs

QUERYING GEOMETRY



- For each query-relevant GH
  - If the query GH is inside the query geometry, return all features in this GH and all those of all its children
  - Otherwise
    - Return features for which the GH is inside or is related and overlaps the query geometry
      - Until the highest resolution is reached, repeat
- Until lowest resolution is reached, consider the parent GH*
  - Return features for which the GH is inside, or is related and overlaps the query geometry

KEY POINTS

- While a given geometry may be indexed at many levels of resolution, each portion of it is indexed at only 1 level of resolution
- Indexing scheme naturally floats geometries into smaller GHs (i.e. higher resolution)
  - Biggest GH for a geometry is on the scale of the geometry itself (the GH that fits inside the geometry)
  - Only extremely large geometries are indexed under very big GeoHashes
  - As you walk up in GH size (down in resolution) fewer geometries are returned, thus fewer overlap calculations need to be made
- Although many query-relevant GHs may intersect a given geometry (e.g. along its edges), only 1 overlap calculation is ever made for that geometry
  - Once we determined the geometry is relevant to the query, we don't have to check it for every query-relevant GH
- *For each query-relevant GH we explore all the child hashes, but only about 1/4 of them must explore all the parent hashes
- Indexes are sharded across the cluster and all index operations (including overlap calculation) are performed server side on all shards in parallel

INSTITUTE FOR MODERN INTELLIGENCE

DEVELOPING THE SCIENCE, PRACTICE, AND GOVERNANCE OF MODERN INTELLIGENCE

GEOSPATIAL-TEMPORAL DATA & THE ATOM OF SEMAPP SEMANTICS

Move it forward. Make it happen.

GEOSPATIAL-TEMPORAL DATA

"EVERYTHING HAS A SOMEWHERE AND A SOME-WHEN"

- CDR J. SMITH, USN RETIRED

- Geometry represents geospatial information a graphic system
  - Three elementary figures of (planar) geometry: point, line, area
  - Represented by a delimited set of geographic coordinates (i.e. points)
- Feature
  - In a GIS, a feature is a geometry with a label and a type (e.g. 30°10′N 031°06′E, Nile, River)
  - In the SRF, a feature is a sign that also has a geometry (e.g. [Nile, River, 30°10′N 031°06′E])
    - Geospatial disambiguation
    - GIS features fit naturally into the sign-space
    - Sign-Space can easily emit features to GIS systems
    - Associations may be asserted between GIS features and other signs / associations in the sign-space
- Interval is the elementary figure representing temporal information
- Event
  - In the SRF, is a sign that has an interval (e.g. [Wall Street bombing, Terrorist Act, 19200916])
    - Temporal disambiguation
    - Events fit naturally into the sign-space
    - Sign-Space can easily emit events to ABI systems
    - Associations may be asserted between events and other signs in the sign-space
- Layers
  - In a GIS, a layer created by displaying a set of features by type (e.g. Rivers, Roads)
  - In SRF, a layer can include signs of any type and types from any domain
  - Imagine temporal layers!

GIS SYSTEM'S SET OF TYPES COMPRISE A DATA-MODEL

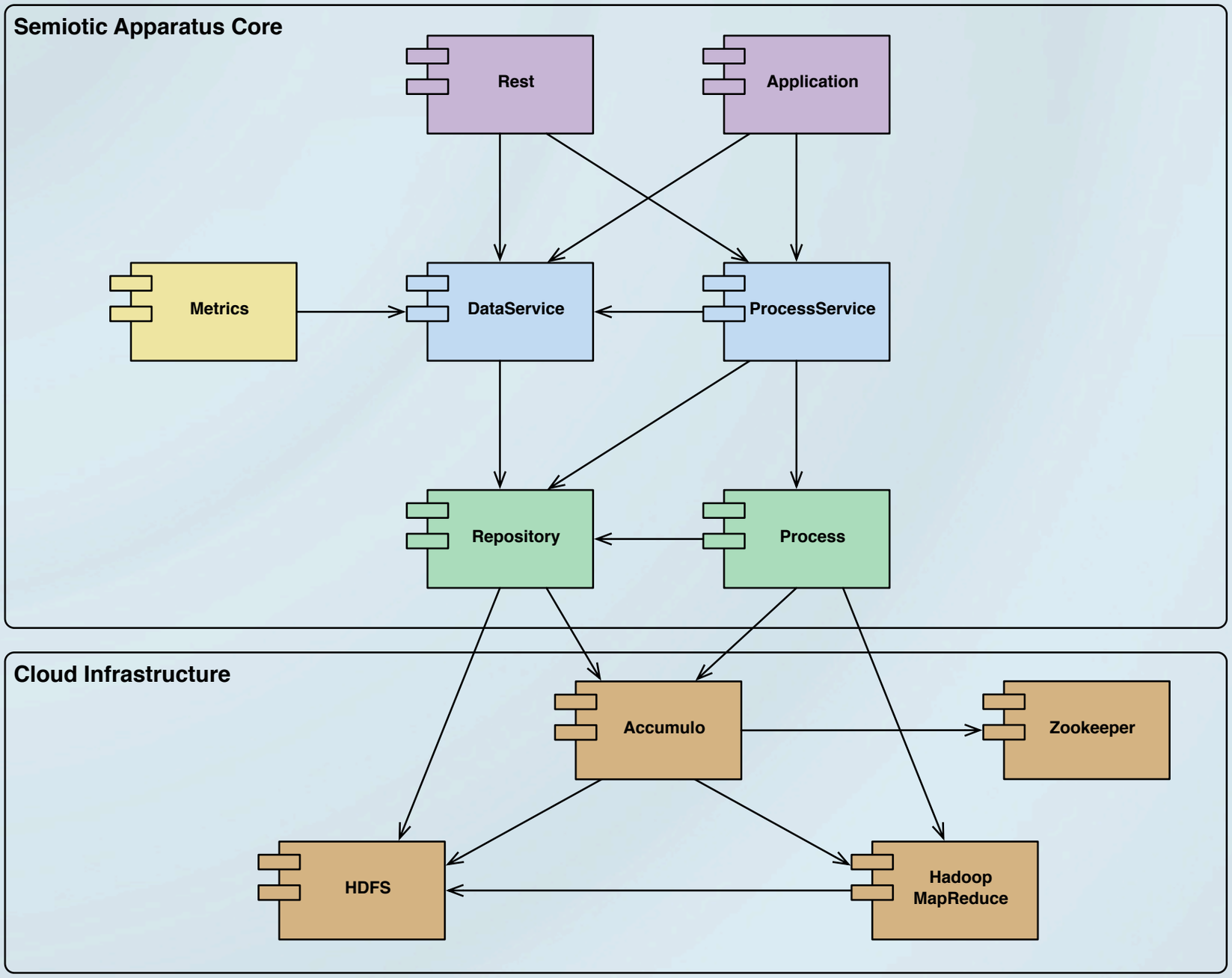
Move it forward. Make it happen.

THING 5 IS A UNIVERSAL STORE FOR GEOSPATIAL TEMPORAL INFORMATION

INFRASTRUCTURE

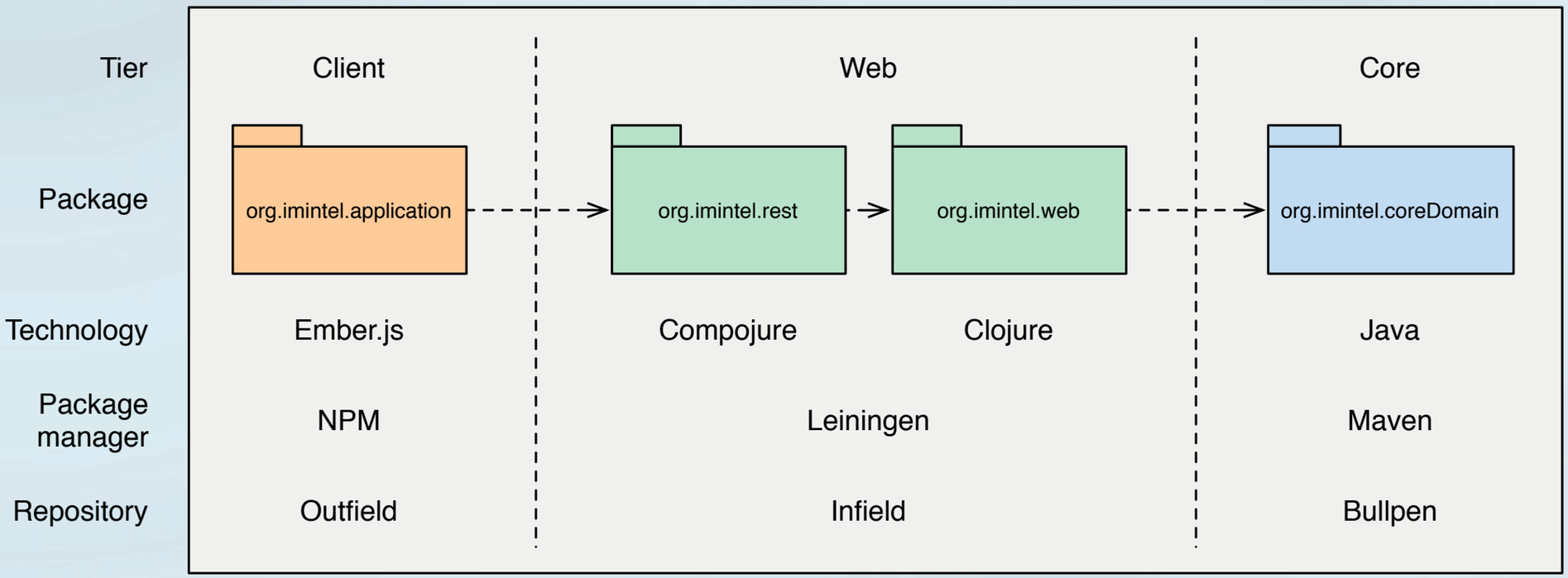
Move it forward. Make it happen.

LOGICAL ARCHITECTURE CORE



SemApp Logical Architecture
R 0.14 2011 01 23
S. Yoakum-Stover

Logical Architecture Web



SemApp Web Architecture
R 0.17
S. Yoakum-Stover
2013 05 10

CENTRAL SWITCH

2 Cisco 6504 in VSS mode
connected by 2 or 4 Gbps peer links
(performing as a single device)
Each switch has:
ACE module (application control engine) and
FWSM (firewall services module)

At least one 10 Gbps link per switch

Nexus 7k with
2 Sup modules
2M1 (routing) modules
12 F1 (layer 2) line cards
Full set (5?) of fabric modules

10 Gbps uplinks to site network

6504 Border Security

N7k Central Switch
net: 10.254.0.254
data:10.0.0.254
bmc:10.20.0.254

Links to only 1 rack shown

Compute & Store Rack
net:10.254.0.2
data:10.0.1.0/24
bmc:10.20.1.0/24
see detail

Links to only 1 server are shown

Server Rack
net:10.254.0.1
data:10.0.0.0/24
bmc:10.20.0.0/24
see detail

NETWORK

There are 2 subnets:
data and
management (denoted bmc represented in blue).

10 F1 modules should be sufficient to accommodate all the Cisco switches in the racks, which leaves 2 F1 modules to host the server nodes. Each server should have a 10Gbps uplink to each of the two F1 modules. The links should be evenly distributed between the upper and lower halves of the modules.

There should be at least one 10Gbps link from the N7k to each of the 6504's. Recommend putting one in each of the M1 modules, taking care not to use more than one port per 4-port group - the link from the 6504's to the N7k will be on a different LAN from the cluster-internal networks, so communication between it and the cluster internals will require routing.

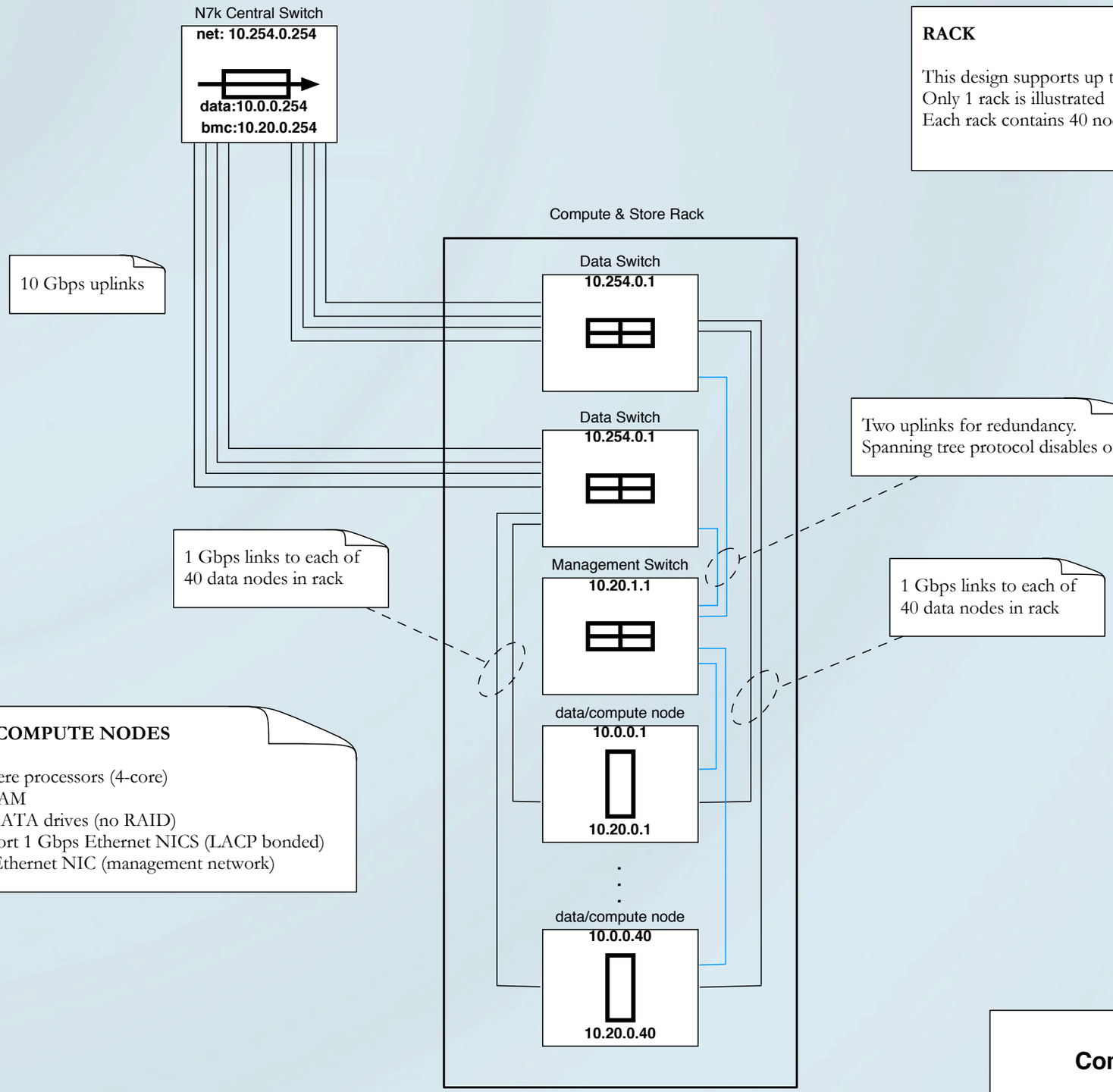
IP NUMBERING SCHEME

Data Networks: 10.0.<rack>.<node>/24
BMC Networks: 10.0.<rack>.<node>/24
Network Equipment: 10.254.0.0/24

Design follows NSA R6 (Not a Ghost Machine)

Data Cloud
Border Security and Central Switch
R7
S. Yoakum-Stover
2011 10 26

COMPUTE & STORAGE RACK



RACK

This design supports up to 10 racks
Only 1 rack is illustrated
Each rack contains 40 nodes

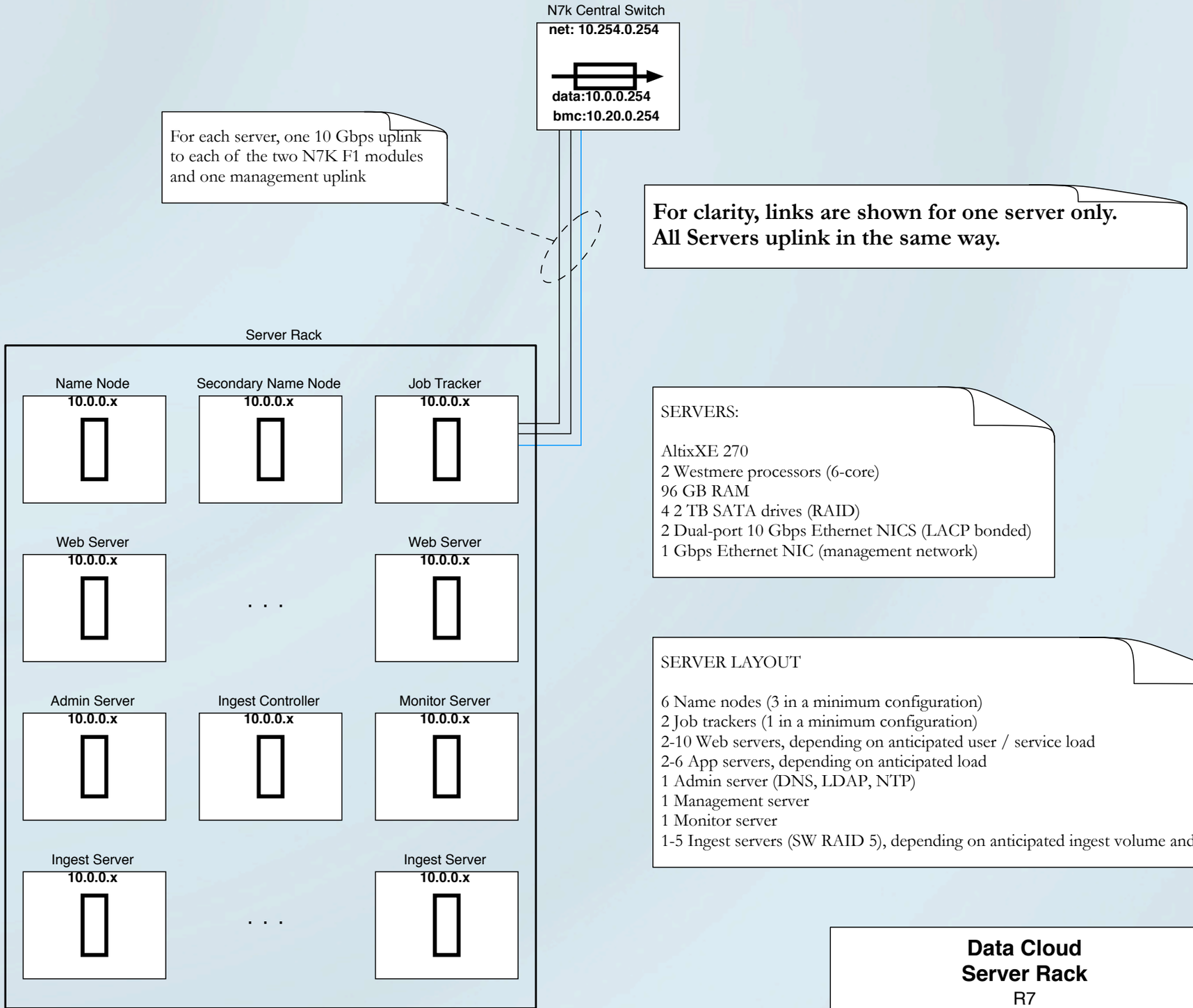
DATA/COMPUTE NODES

2 Westmere processors (4-core)
48 GB RAM
4 2 TB SATA drives (no RAID)
2 Dual-port 1 Gbps Ethernet NICS (LACP bonded)
1 Gbps Ethernet NIC (management network)

Data Cloud
Compute & Store Rack
R7
S. Yoakum-Stover
2011 10 26

Move it forward. Make it happen.

SERVER RACK



For clarity, links are shown for one server only.
All Servers uplink in the same way.

SERVERS:

AltixXE 270
2 Westmere processors (6-core)
96 GB RAM
4 2 TB SATA drives (RAID)
2 Dual-port 10 Gbps Ethernet NICS (LACP bonded)
1 Gbps Ethernet NIC (management network)

SERVER LAYOUT

6 Name nodes (3 in a minimum configuration)
2 Job trackers (1 in a minimum configuration)
2-10 Web servers, depending on anticipated user / service load
2-6 App servers, depending on anticipated load
1 Admin server (DNS, LDAP, NTP)
1 Management server
1 Monitor server
1-5 Ingest servers (SW RAID 5), depending on anticipated ingest volume and rate

Data Cloud
Server Rack
R7
S. Yoakum-Stover
2011 10 26

INGEST

Move it forward. Make it happen.

Ingest Process

Core Tenants

- Never drop data – ingest is transactional
- Make it extremely robust – ingest it, then fix it
- As simple as possible and no simpler
  - Take the data as it lies – no transformation, no normalization
  - Move the bytes
    - Read bytes from endpoint (filesystem, JMS, (S)FTP, multi-cast...)
    - Persist raw bytes artifact
    - Persist a sign representing the artifact
  - Attach out-of-band metadata (ingest metadata)
    - Defined by the ingest manager and applied to all artifacts from a given endpoint
    - Fields
      - Required: authoritative source, ingest source, modality, classification and releasability
      - Optional: from-file-path, custom fields, detailed class info (e.g. system, declass, downgrade)
    - Persisted as signs associated to the artifact sign
- Everything else is a map-reduce process

SOME TYPES OF INGEST

ARTIFACT INGEST

- Unstructured / semi-structured files of any type / modality
- NITF, IRC, email, MP2TS, ...
- Persist the artifact raw bytes
- Persist a sign representing the artifact
- Associate out-of-band metadata

RELATED ARTIFACT INGEST (E.G. SHAPE FILES)

- Shapes, index, shape attributes, and a decoder
  - Index relates shapes to attributes
  - Decoder relates attribute key-value pairs to signifier-concept pairs
- Ingest tarball of shapes, index, attributes and decoder as an artifact
- A map-reduce job persists shapes and attributes as signs and associations
- Spreadsheet ingest is similar

COLLECTION INGEST (E.G. WAMI)

- Maintains relationships among ingested artifacts

WEB SERVICE INGEST

- Collect data and model (service spider)
- Ingest the model, ingest the data as an artifact
- Process with map-reduce to reveal semantics
- Can be more sophisticated, incremental, dynamic

RDB INGEST

- Operates at the level of relational semantics (not domain semantics)
- Ingest the model
- Ingest the data and relations (i.e. signs and associations)

MODEL INGEST

- Represent model as OWL ontology
- Ingest the model
- Model, concepts, predicates are persisted with expressivity equal to OWL

ORGANIZED TESTING

Move it forward. Make it happen.

TESTING

ORGANIZATION OF TESTS

- Core service boundary
  - Where the system data architecture is fully enforced
  - Dividing line between managed & non-managed tests
- Non- Managed tests
  - Unit tests
  - Result from TDD
  - Not included in test plans
- Managed tests
  - Included in test plans
  - Functional
  - Configuration
  - UI
  - Performance
  - Acceptance
- Regression testing
  - All automated tests are part of continuous integration process
    - Check-in – build – deploy – test – manual test

Type	Visibility	Mode	Level	Subject
Unit	white box	automated	repository process	CRUD SemApp elements
Functional	black box	mostly automated	service	data services,
Configuration	black box	various	n/a	infrastructure
UI	black box	various	application	UI
Performance	black box	various	various	user interactions,
Acceptance	black box	manual	application	applications

PARALLEL PROCESS VIDEO

Move it forward. Make it happen.

PARALLEL PROCESS VIDEO?

WHY YES!

Ruminate

Segment out the metadata groups

Segment out the frames

Exposes the structural elements so they may be more easily exploited

Digest

Extract signs and associations representing the static metadata

The first group is all you need

Now it's discoverable

Process

Extract the collection footprint

Now it's discoverable from a geospatial query and we can do collection management

Exploit the pixels (e.g. object extraction)

etc...

PIPELINES VS WAVES

Move it forward. Make it happen.

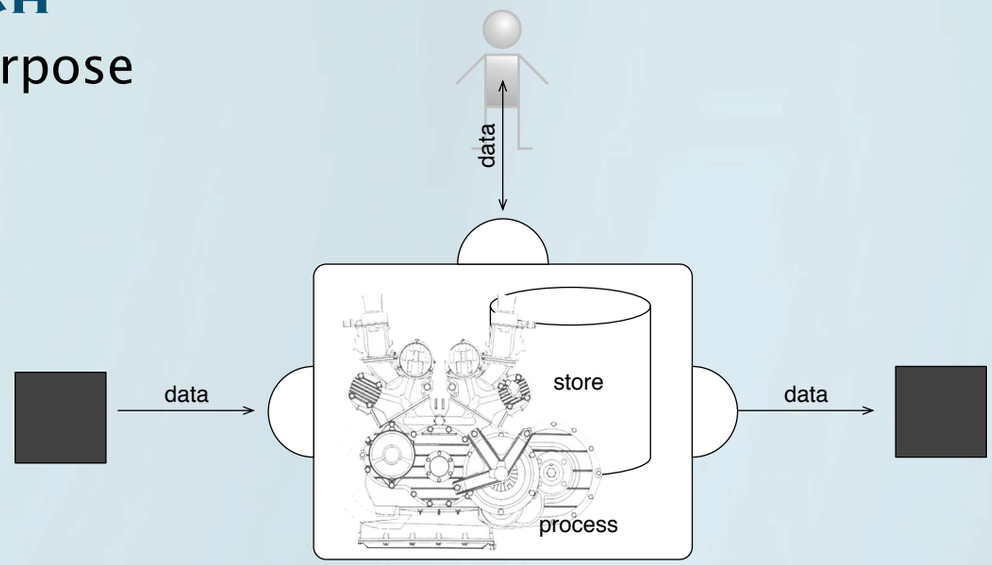
PROCESSING PIPELINES

DIVERSITY OF INFORMATION SYSTEMS, EACH

- Has its own processing, data representation, purpose
- Gets data from somewhere
- Transforms / creates data
- Outputs data

PROCESSING PIPELINE

- Common pattern used within systems
  - Used to chain modular process components
  - Examples: NLP, ETL-style ingest system
  - Data flows down the pipe
  - At every point in the pipe, the data is different
- Common pattern used in Enterprise integration
  - Services hooked up to fulfill business processes (SOA, BPEL, ESB)
  - Data flows between systems



ISSUES

- If processing on a piece of data fails
  - Subsequent processing can not occur (bad data)
  - Dropped on the floor / handled on dead letter channel
- If a component fails
  - Processing can not continue, data stops flowing (broken pipe)
- If a new component is added, downstream data is not affected
- The data assets are different at every point in the Enterprise
  - Collecting into separate pools of information
- Bandwidth & Complexity (IT, data management)

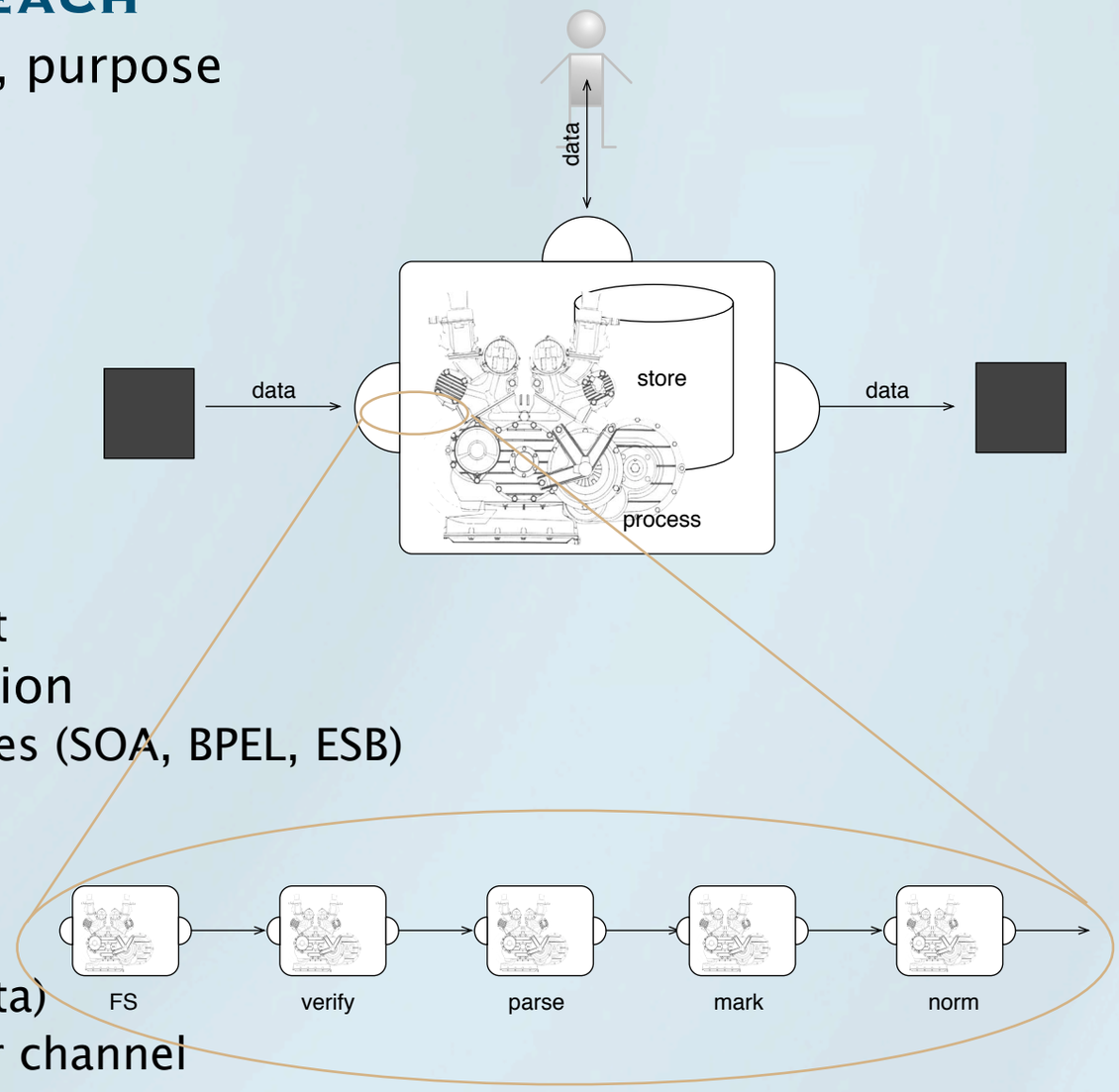
PROCESSING PIPELINES

DIVERSITY OF INFORMATION SYSTEMS, EACH

- Has its own processing, data representation, purpose
- Gets data from somewhere
- Transforms / creates data
- Outputs data

PROCESSING PIPELINE

- Common pattern used within systems
  - Used to chain modular process components
  - Examples: NLP, ETL-style ingest system
  - Data flows down the pipe
  - At every point in the pipe, the data is different
- Common pattern used in Enterprise integration
  - Services hooked up to fulfill business processes (SOA, BPEL, ESB)
  - Data flows between systems



ISSUES

- If processing on a piece of data fails
  - Subsequent processing can not occur (bad data)
  - Dropped on the floor / handled on dead letter channel
- If a component fails
  - Processing can not continue, data stops flowing (broken pipe)
- If a new component is added, downstream data is not affected
- The data assets are different at every point in the Enterprise
  - Collecting into separate pools of information
- Bandwidth & Complexity (IT, data management)

PROCESSING PIPELINES

DIVERSITY OF INFORMATION SYSTEMS, EACH

- Has its own processing, data representation, purpos
- Gets data from somewhere
- Transforms / creates data
- Outputs data

PROCESSING PIPELINE

- Common pattern used within systems
  - Used to chain modular process components
  - Examples: NLP, ETL-style ingest system
  - Data flows down the pipe
  - At every point in the pipe, the data is different
- Common pattern used in Enterprise integration
  - Services hooked up to fulfill business processes (SOA, I
  - Data flows between systems

ISSUES

- If processing on a piece of data fails
  - Subsequent processing can not occur (bad data)
  - Dropped on the floor / handled on dead letter channel
- If a component fails
  - Processing can not continue, data stops flowing (broke
- If a new component is added, downstream data is nc
- The data assets are different at every point in the En
  - Collecting into separate pools of information
- Bandwidth & Complexity (IT, data management)

Move it forward. Make it happen.

THE PROBLEM WITH PROCESSING

is…

We got Lots of Different Processes and Lots of Data to process

and Everyone Wants everyOneElse's processed Data

So we we Hook Up the processes and Send the Data

In other words,

The Data FlOws around the Enterprise

(or NOT)

and is Different at Every Place

So Don't do That

PROCESSING WAVES

SEMAPP DECOUPLES PROCESSING FROM DATA FLOW

- Ingest is Not a pipeline
  - A single process: Move file/stream bytes into the sign-space (and persist OOB metadata)
  - Works the same regardless of what sort of thing the bytes are
  - Everything else is a process
- Sign-Space is an ocean of data
  - Data stays in place
  - Processing enriches the ocean
  - All processing sees the same ocean

PROCESSING OCCURS IN WAVES

- Each wave processes all the data available to it
  - Available means: Is processable & Not already processed
- Processes leave their fingerprints on the things processed
  - Mechanism for determining availability during the next wave / for another process
- Processes execute independent of other processes

NO FLOW, NO LEAKS, NO BREAKS

- All processing that can occur on an element will
- The sign-space is the floor, data is never dropped (i.e. lost)
- If a component fails, dependent processing finds no available input
  - Once fixed, dependent processing again finds available input on the next wave
- If processing on a piece of data fails, dependent processing will not find it to be available
  - Enrichment produced by all previous processing still present
  - If the issue can be corrected, no special handling is required, dependent processing find it on the next wave
- If a new process is added, all the data that can be processed will
- Bandwidth reduced, complexity simplified, and robust

TEMPORAL INDEXING

Move it forward. Make it happen.

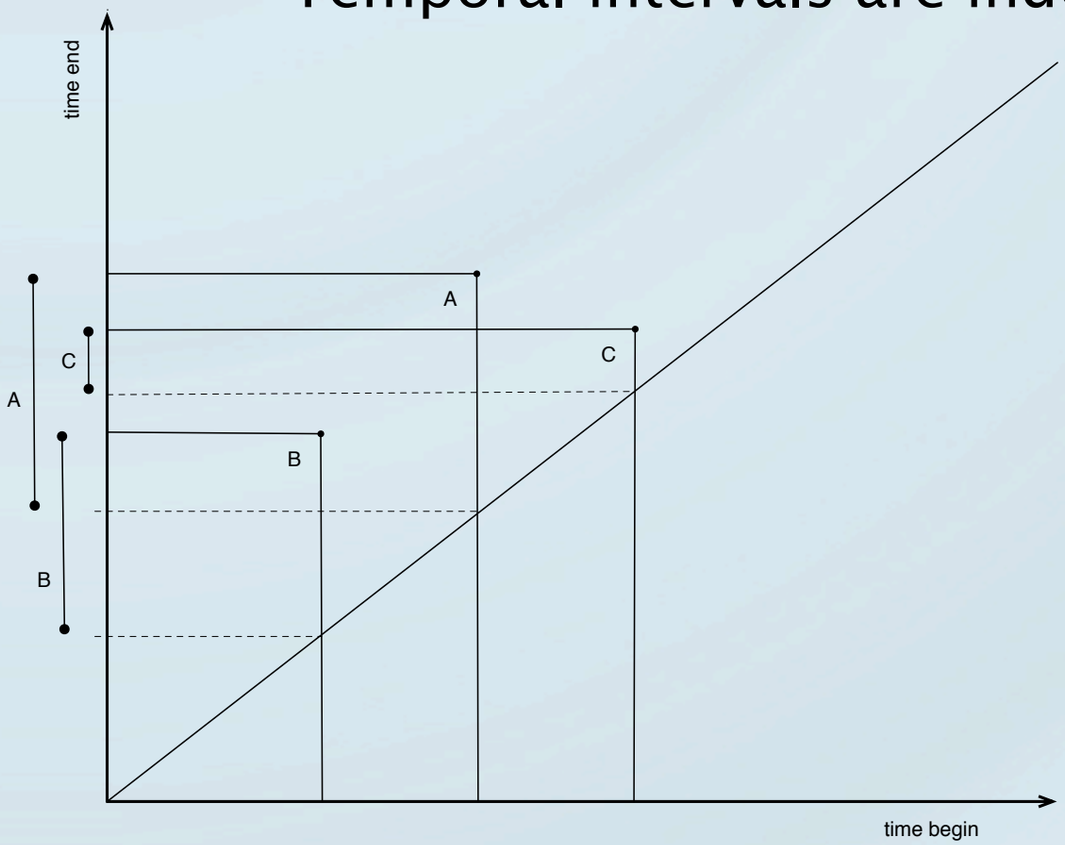
TD-Tree Overview

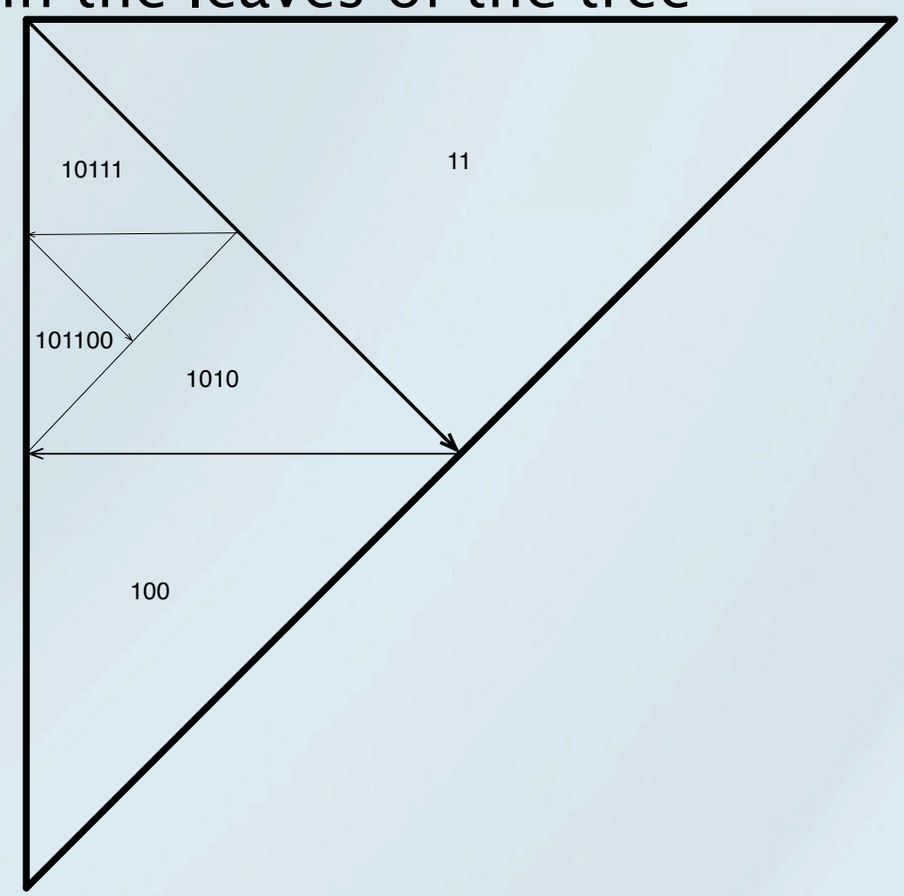
Temporal Decomposition Tree* (TDT) represents temporal intervals as points in 2D

Region of interest is a triangle, which is is decomposed into a tree of ever smaller triangles

Each represented with a binary identifier

Temporal intervals are indexed in the leaves of the tree





*Stantic, B., Topor, R., Terry, J., Sattar, A.: Advanced Indexing Technique for Temporal Data. Computer Science and Information Systems, Vol. 7, No. 4, 679-703. (2010)

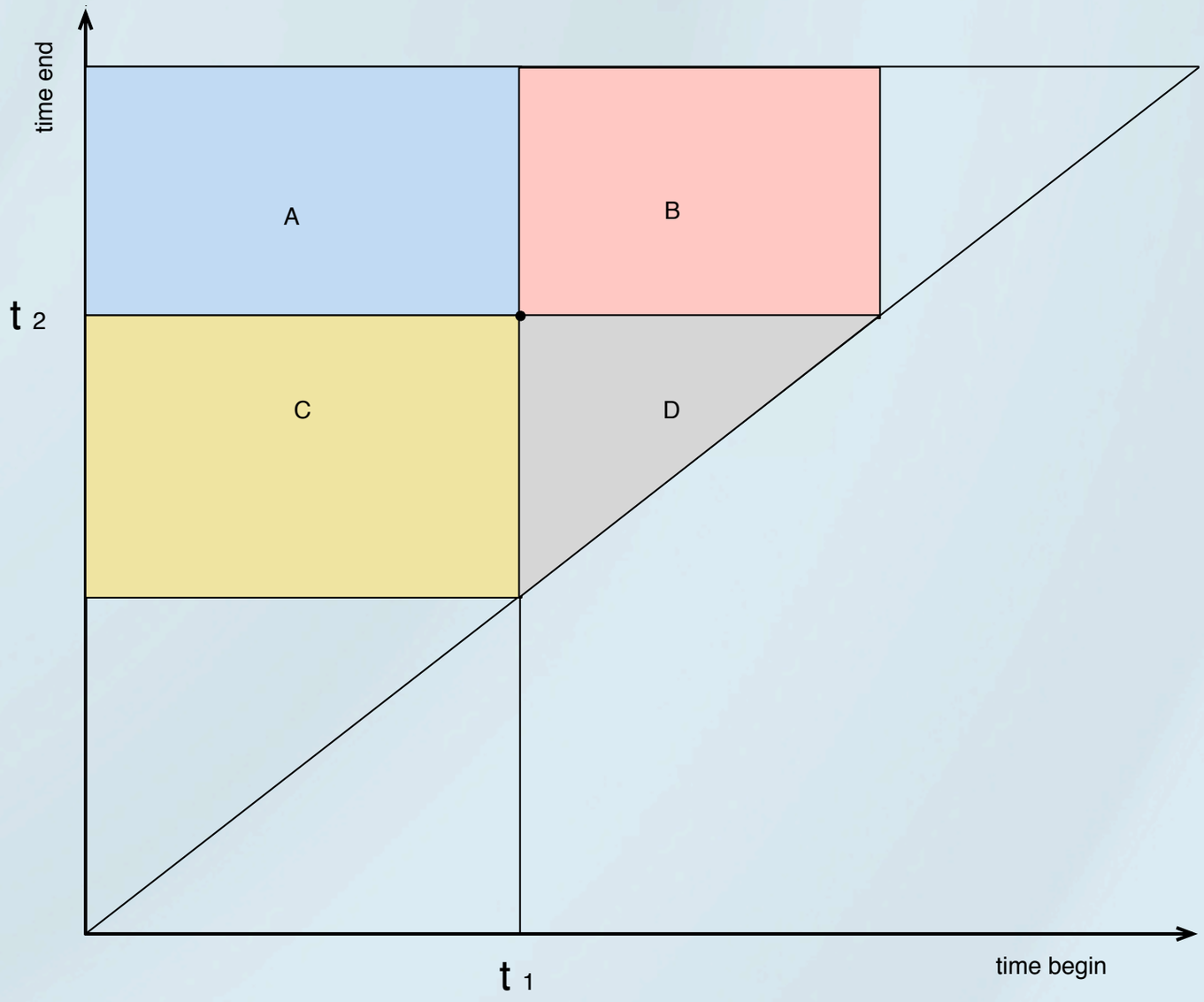
TEMPORAL INDEXING & QUERYING

INDEXING

- Index the interval at the TDT leaf which contains it (our approach differs slightly from the TDT paper)

QUERYING

- Standard queries expressed as rectangles, for example:
  - A: Intervals starting before t_1 and ending after t_2
  - B: Intervals starting between t_1 and t_2 and ending after t_2
  - C: Intervals starting before t_1 and ending between t_1 and t_2
  - D: Intervals starting between t_1 and t_2 and ending between t_1 and t_2

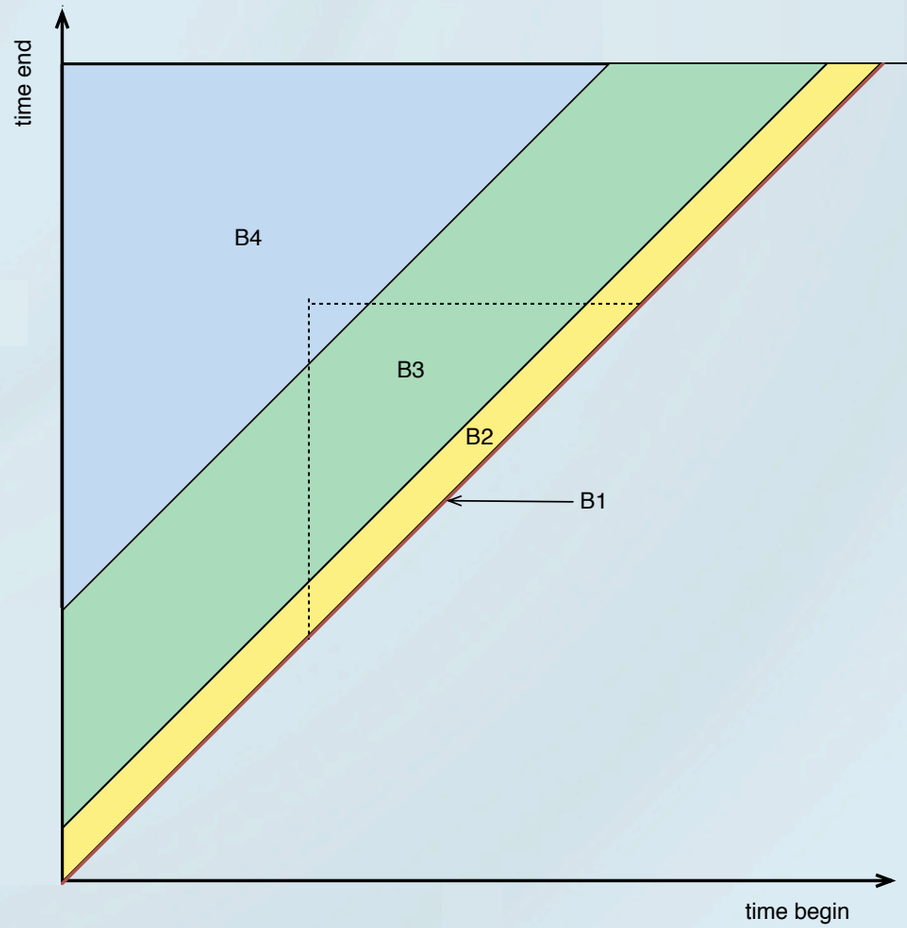


KEY POINTS

- Earliest time, latest time, and max resolution are specified prior to constructing the TDT, however
  - A large offset is used in case we need to go back farther in time
  - Max resolution can be increased later on
- Indexes are sharded across the cluster and all index operations are performed server side on all shards in parallel
  - Keeps the index balanced across the cluster
- TDT approach works best for intervals that are widely distributed in size
  - Points in time are more efficiently indexed using straight-forward approach
  - Intervals relevant to Intel will likely cluster just above the $t_1 = t_2$ line
  - Few intervals are likely to appear in the future

TEMPORAL BUCKET INDEX

- Addresses the shortcomings of the TDT index
- Separately index intervals with different lengths
  - Instants (intervals with 0 length) should just be lookups (B1)
  - Longer intervals may require some small amount of analysis (B2, B3, B4, ...)



- Index
  - Assert element in the smallest bucket size that fully contains its interval, ordered by start time
- Query
  - For each bucket, starting with the earliest start time that could satisfy our query and ending with the

There is No Metadata

Move it forward. Make it happen.

METADATA

- Metadata is structured data
  - Generally distinguished by its target
    - Not the X, but information about (or extracted from) the X, where X = artifact, data, model ...
    - Subjective distinction – One person's metadata is another's data
  - Variety of metadata standards/models reflect a variety of uses and perspectives
- SemApp accommodates all, but sanctifies none
  - Metadata is represented just like any other kind of structured data
  - Metadata and metadata models are persisted in SRF
  - SemApp imposes no special metadata model
- Pedigree, logs, and system actors are part of the SemApp machinery
  - History of sign-space changes (who, what, when) are persisted
  - System logs are persisted
  - System actor (person/process known to the SemApp) information is also persisted
    - Supports history, logging, and security

METADATA IS STRUCTURED DATA

TREATING SEMI-STRUCTURED DATA RIGHT

SEMI-STRUCTURED DATA

Semi-structured data is data that conforms to a standard

(e.g. HTML, RSS, MIME, NITF, MP2 TS, ...)

Standard is a less glamorous cousin to data-model

Mix of

Semantic elements (e.g. NITF originating station Id, MIME in-reply-to)

Structural elements (e.g. NITF file header length)

How to handle it?

Preserve & expose both the structure & the semantics

Beyond Ingest

Ingest

Stream raw bytes into the Sign-Space

Ruminate

Segment according to the standard

Typically only 1 way to do this for a given standard, but there are many standards

Exposes the structural elements so they may be more easily exploited

Digest

Extract signs and associations according to a semantic model based on the standard

Expose the semantic elements expressed by the standard

(create signs with mentions referring to appropriate segments)

May be done in slightly different ways

(different semantic models interpreting the standard)

Process

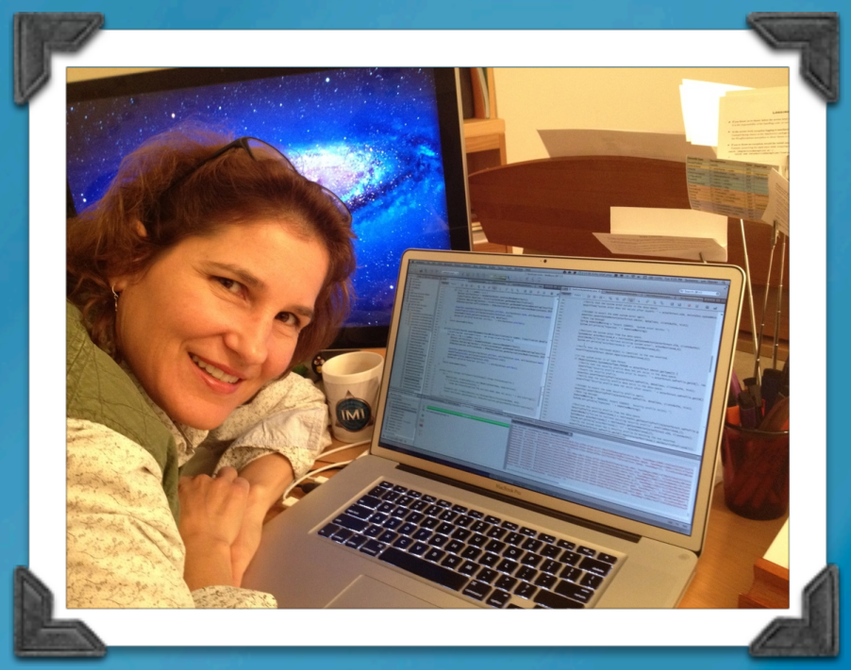
Extract semantics from unstructured elements based on a domain model (e.g. NLP)

Expose and enrich semantic elements extracted from unstructured elements

May be done in many different ways according to many different models

DEVELOPING THE SCIENCE, PRACTICE, AND GOVERNANCE OF MODERN INTELLIGENCE

WHO ARE WE?



MISSION FOCUS

Move it forward. Make it happen.

WHO WE ARE

Institute for Modern Intelligence

IMI aims to change the world!
by developing
an Ultra-Large Scale systems infrastructure
for data-intensive computing & Data-intensive operations

Our mission:

Develop the Science, Practice, and Governance of Modern Intelligence

We are a non-profit, 501(c)3 organization
whose purpose is to fulfill our mission.

S. Yoakum-Stover & M. A. Eick are the IMI founders

Mission Focus

Mission Focus aims to build clean code that works for its customers.
We are an agile software development shop that takes
domain design and development as seriously as
system design and development.

Our tag line is:

Move it Forward, Make it Happen

We are a small, for-profit organization
whose purpose is to fulfill its customers' missions.
Our profits help to support the IMI.

M. A. Eick & S. Yoakum-Stover are CEO and Chief Scientist respectively

Move it forward. Make it happen.

Core Domain

Mission Focus and IMI work mostly in the intelligence domain with DoD and IC customers / partners.
Our core domain addresses the storage, processing, and utilization of data in the context of immense volume and diversity.

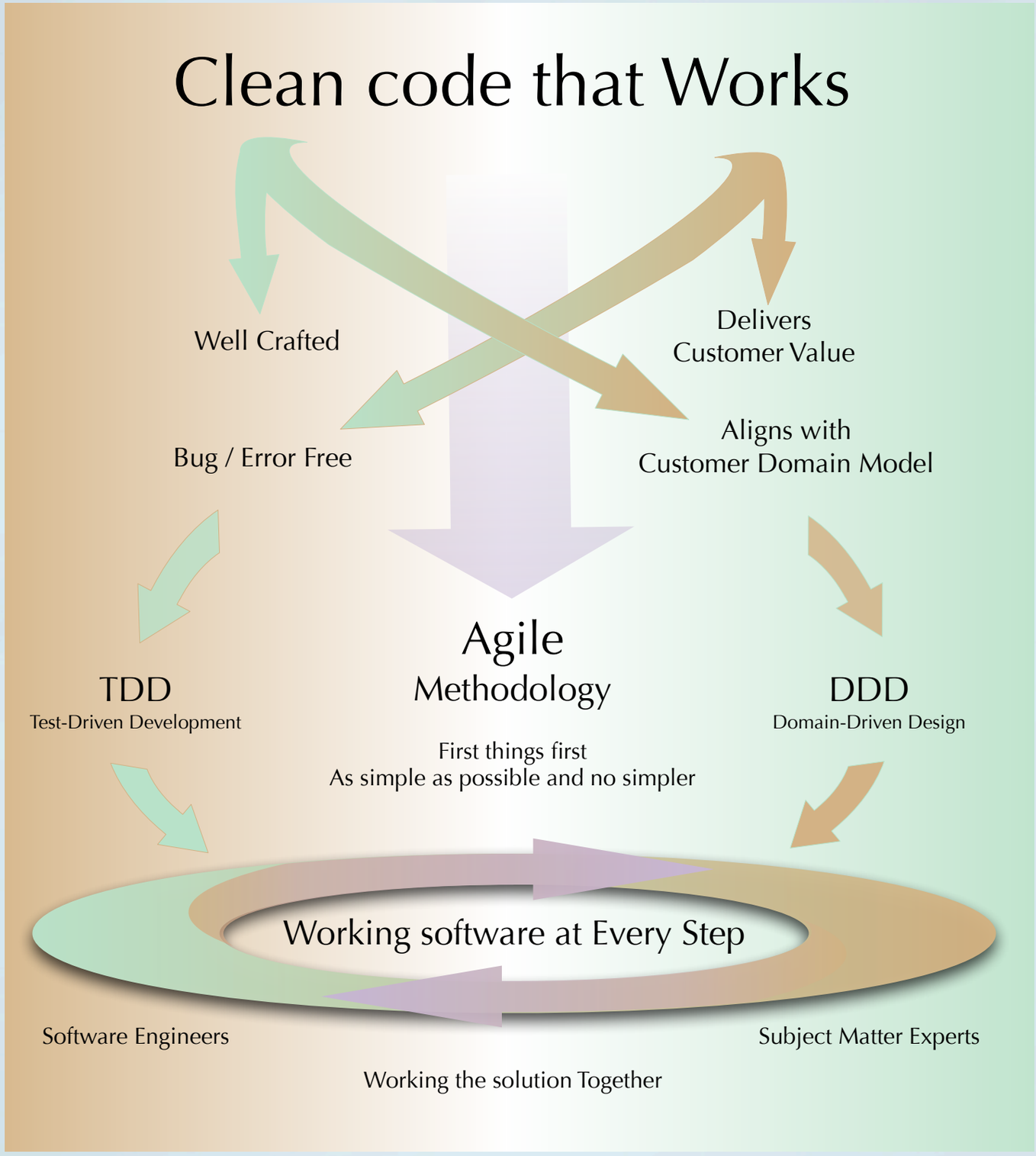
We are experts in cloud computing and storage technology.
We invented the Sign Representation Framework which underpins a game-changing approach to data unification.

We pride ourselves on our disciplined engineering practices and distinguish ourselves by our ability to continually learn and innovate.

The work we do is meaningful and intentional and is wrapped with our integrity.

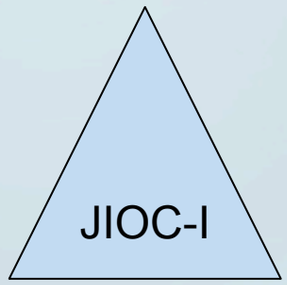
Put simply,
We just think harder and work better than the rest.

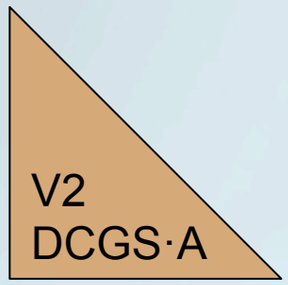
How We Work



Move it forward. Make it happen.

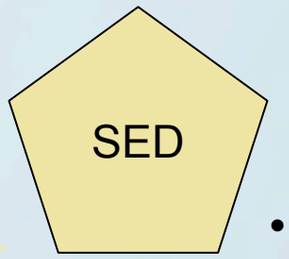
THING 5 HERITAGE





- Dr. Suzi Object Sciences @ I2WD
- JIOC-I V2 integration
- Universal store for documents meets UFEIEO
- We need a universal store for structured data

- Andy SSS / TexelTek under Dr. Heath @ NSA
- GeoBoost – thin client Geo visualization
- Ozone widget framework
- Cloudbase analytics under Dr. Dowd



- Dr. Suzi PIPS @ I2WD Tech Dir
- Andy SSS / TexelTek
- Cross-domain MLS Oracle
- Data Description Framework
- The data integration problem

- Dr. Suzi IMI @ I2WD SETA Tech Dir
- Andy MissionFocus
- Compute & storage cloud aspiration
- Legacy integration implementation

- Dr. Suzi IMI & MissionFocus
- Andy MissionFocus & IMI
- GEOINT Compute & storage cloud implementation
- Sign Representation Framework
- PL4 design
- Marry cloud technology with semiotic science

BATTLE SCARRED AND A WHOLE LOT SMARTER

Move it forward. Make it happen.

Current Capabilities

Move it forward. Make it happen.

Capabilities - Data

- collect, Ingest, Export
  - All unstructured / semi-structured file types
  - From endpoint types: File system, JMS topic / queue, (S)FTP, multi-cast / uni-cast stream
  - Related files (e.g. shape)
  - Collections (e.g. WAMI)
  - Web service collectors
  - Domain data-models as OWL
- Assert, Retrieve, Delete, Exists, Tally
  - Artifacts of any type and modality
  - Signs & associations
  - Domain Models
  - System actors
  - Process elements
- Index & QUERY
  - Full Text
  - Semantic (Signifier, concept, predicate)
  - Geospatial (including compound geometries)
  - Temporal
- Preserve History
  - All sign-space changes including deletes
- Semiotic refinement

Move it forward. Make it happen.

Capabilities - Process

- Process waves infrastructure
- SemApp map-reduce process framework
- Parallel processes
  - Assess
    - Gather artifact, SRF, and model statistics
  - Detect
    - Artifact standard
    - Artifact language
  - enrich
    - Extract & expose place names from artifact text
    - Extract & expose Semantics from Email, NITF, IRC, TIFF, MP2TS, CSV
    - Extract & expose NITF images as artifacts
    - Extract & expose Email attachments as artifacts
    - Extract & apply security markings from NITF & MP2TS
    - Extract & expose video collection footprint from MP2TS, KML fly-over display
    - Find shortest path between 2 semantic elements
    - Extract & expose shape-file features and attributes
    - Extract and expose faces within images
    - Generate standard image representations (thumbnail, small, medium, full)
    - Model reasoners - Functional, PredicateChain, Transitive
    - Relatedness within archives
  - index
    - artifacts, signs, associations, mentions, models
  - load
    - SystemActors, OWL models, process specifications, gazetteers
  - mend
    - Maintain integrity of index and material signs
  - Util
    - Parse > 50 document, audio, video, & image file types

Move it forward. Make it happen.

Image & Video Processing

Move it forward. Make it happen.

Image Tag-Value Standard	Executable Process								
	Artifact Ingest	Artifact Standard Detector	Classification Markings Extractor	Artifact Ruminate	Artifact Digest	Artifact Text Extractor	Place Name Extractor	Image Proofer	Image Face Extractor
KODAK_DCR	✓	✓		✓	✓	✓	✓		
KODAK_KDC	✓	✓		✓	✓	✓	✓	✓	✓
EPSON	✓	✓		✓	✓	✓	✓	✓	✓
HASSELBLAD	✓	✓		✓	✓	✓	✓	✓	✓
NIKON	✓	✓		✓	✓	✓	✓	✓	✓
OTHER_RAW	✓	✓		✓	✓	✓	✓		
BMP	✓	✓		✓	✓	✓	✓	✓	✓
GIF	✓	✓		✓	✓	✓	✓	✓	✓
JPEG	✓	✓		✓	✓	✓	✓	✓	✓
PNG	✓	✓		✓	✓	✓	✓	✓	✓
PSD	✓	✓		✓	✓	✓	✓		
TIFF	✓	✓		✓	✓	✓	✓	✓	✓
NITF2_0	✓	✓	✓	✓	✓	✓	✓	✓	
NITF2_1	✓	✓	✓	✓	✓	✓	✓	✓	

Standard	Executable Process								
	Artifact Ingest	Artifact Standard Detector	Classification Markings Extractor	Artifact Ruminate	Artifact Digest	Artifact Text Extractor	Place Name Extractor	Footprint Creator	Video Proofer
MP2TS	✓	✓	✓	✓	✓	✓	✓	✓	✓

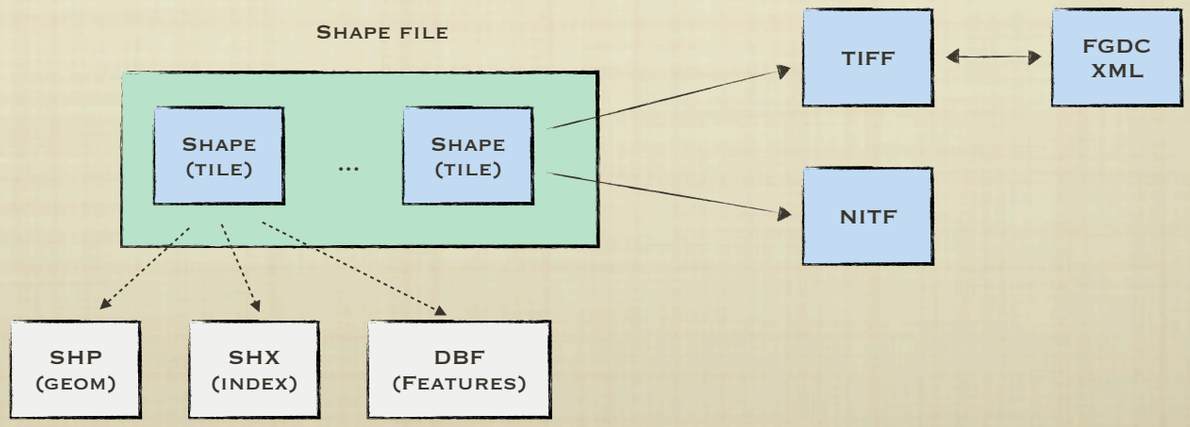
Text Processing

Text Standard	Executable Process							
	Artifact Ingest	Artifact Standard Detector	Artifact Ruminate	Artifact Digest	Artifact Text Extractor	Place Name Extractor	Artifact Language Detector	FGDC Metadata Digest
CSV	✓	✓	✓	✓	✓	✓	✓	
XLS	✓	✓	✓	✓	✓	✓	✓	
ODS	✓	✓	✓	✓	✓	✓	✓	
PLAIN_TEXT	✓	✓			✓	✓	✓	
XML	✓	✓			✓	✓	✓	
RSS	✓	✓			✓	✓	✓	
ODT	✓	✓			✓	✓	✓	
OTT	✓	✓			✓	✓	✓	
STW	✓	✓			✓	✓	✓	
SXW	✓	✓			✓	✓	✓	
FGDC_METADATA	✓	✓			✓	✓	✓	✓
HTML	✓	✓			✓	✓	✓	
DOC	✓	✓			✓	✓	✓	
DOCX	✓	✓			✓	✓	✓	
CHAT	✓	✓	✓	✓	✓	✓	✓	
EMAIL	✓	✓	✓	✓	✓	✓	✓	
RTF	✓	✓			✓	✓	✓	
PPT	✓	✓			✓	✓	✓	
PDF	✓	✓			✓	✓	✓	
PAGES	✓	✓			✓	✓	✓	
EPUB	✓	✓			✓	✓	✓	

About FGDC (Federal Geographic Data Committee)

LiDAR collection (SAIC) includes shape files, tifs, xml, and NITFs.

Each shape files references a tiff which references an FGDC XML file containing metadata about the shape whose associated image is in a NITF



Move it forward. Make it happen.

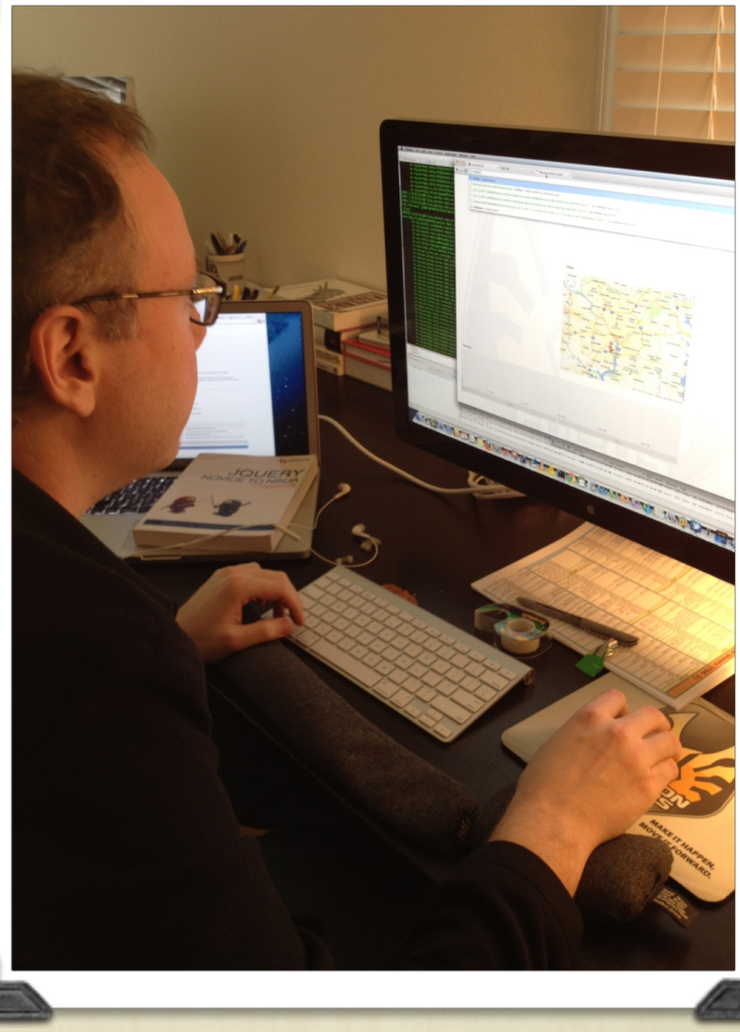
Processing (misc.)

Move it forward. Make it happen.

Standard	Executable Process								
	Artifact Ingest	Artifact Standard Detector	Classification Markings Extractor	Artifact Ruminate	Artifact Digest	Artifact Text Extractor	Place Name Extractor	Footprint Creator	Archive Exploder
SHAPE		✓			✓				
SHP	✓	✓							
SHX	✓	✓							
DBF	✓	✓							
MODEL	✓	✓							
MP3	✓	✓							
MIDI	✓	✓							
MP4	✓	✓							
RIFF	✓	✓							
UNKNOWN	✓	✓							
DECODER	✓	✓							
ZIP	✓	✓							✓
GZIP	✓	✓							✓
BZIP2	✓	✓							✓
TAR	✓	✓							✓

Capabilities - Application

- User Applications
  - The Eye (search application)
  - Data SA (survey application)
- Privileged User Applications
  - **Ingest Manager**
    - Configure and execute artifact ingest
    - Specify out-of-band metadata
  - **System actor manager**
    - browse & manage users / organizations / applications known t
      the system
  - **Process manager**
    - schedule and execute processes
    - Browse process issues
  - **Model Manager**
  - **Gazetteer Manager**
  - **Launcher**
- API
  - SemApp Map/Reduce framework
  - Restful Data & Query Web Services
  - KML export
- Documentation
  - Deployment guide, Management guide, REST API



Move it forward. Make it happen.

Search UI

Text Search:

Search Examples

Text:

Will Return any data space element or artifact entailing this text.
Example: `Suzi, Andy, Car`

Semantic Search:

Search Examples

Text:

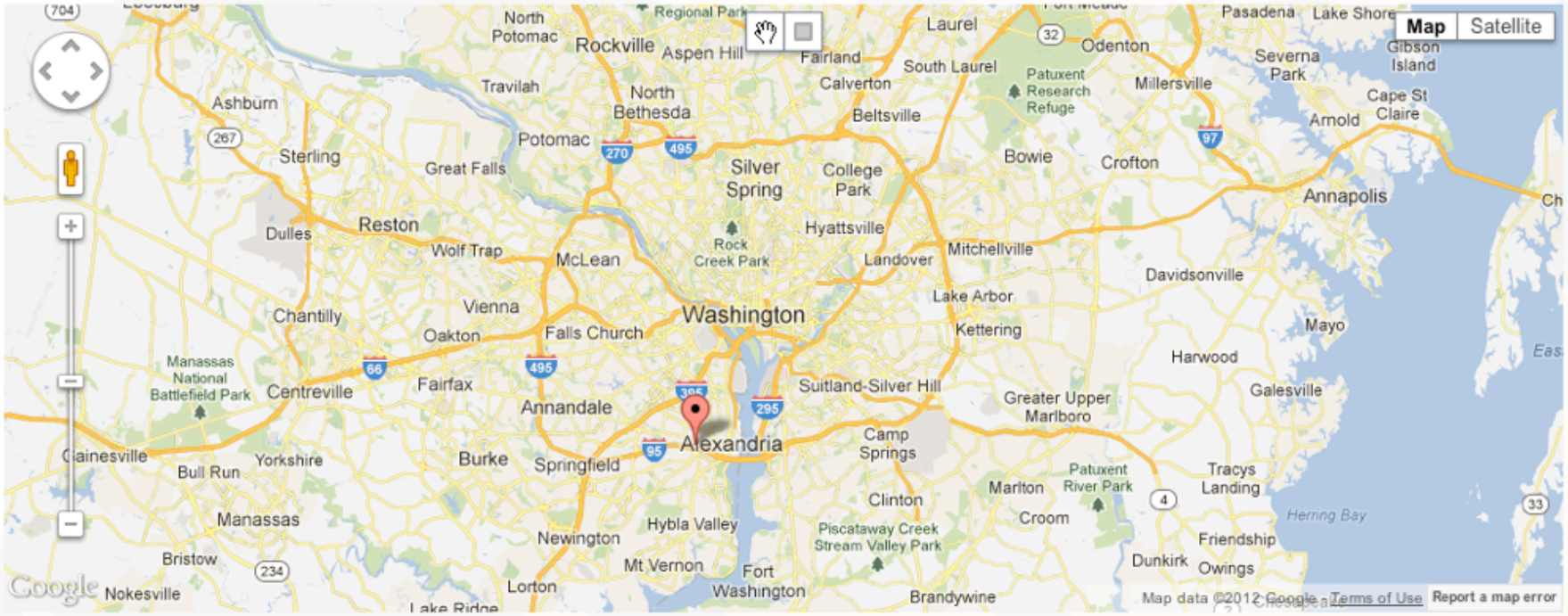
Will Return symantic elements that contain the following text in the sign. Example: `Suzi, Andy, Artifact`

Concept:

Will Return symantic elements that contain the following concept.
Example: `Email, Sender, Receiver`

Visual Search:

Select an area to Search



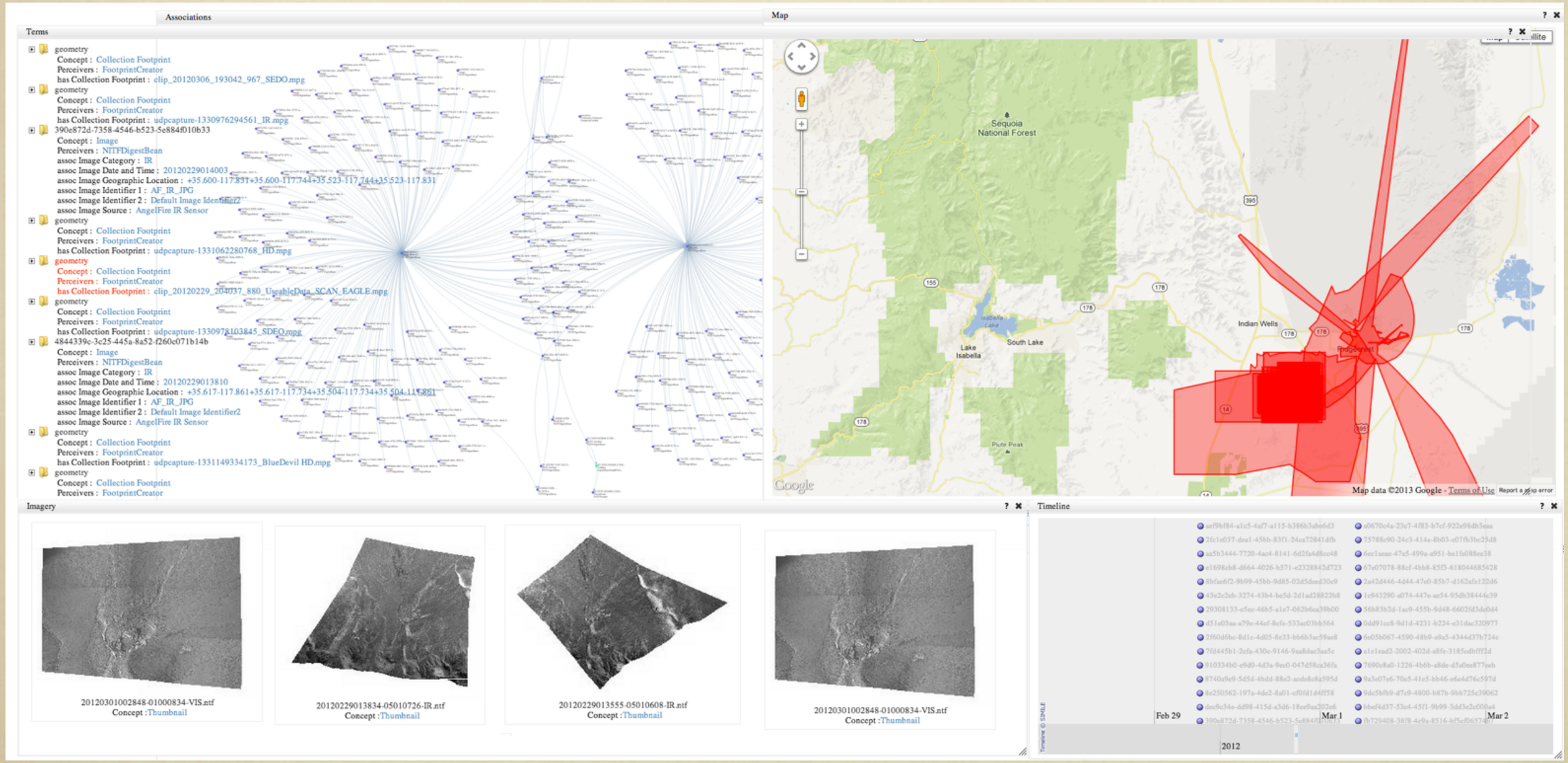
Temporal Search:

From to

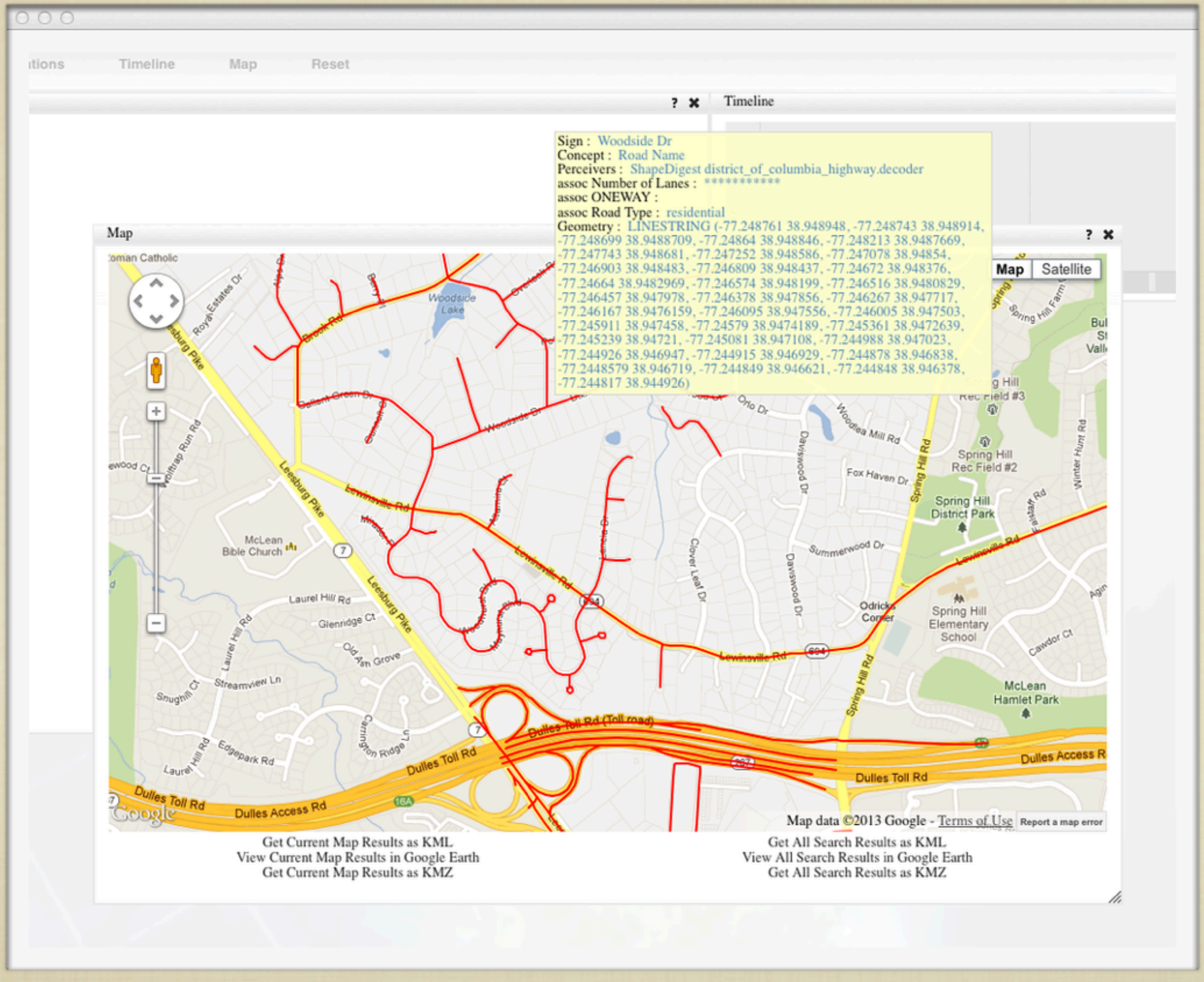
Search

Results Page

- List of results
- Web of information
- Image thumbnails
- Map display
- Collection footprints
- Timeline display



ShapeFile Features



Move it forward. Make it happen.

Capabilities - Other

- Monitoring & Metering
  - Persist log messages in the sign-space
  - Monitor & meter
    - Disk & CPU utilization across the cluster
    - Hadoop & Accumulo SW infrastructure
    - SemApp core methods and services
    - Ingest message transfers
- Security
  - Data classification persisted at the cell-level
  - Releasability determined by user credentials & data classification
  - Applications user authentication & authorization
  - Engineered for PL4
- Continuous Integration
  - Unit tests (as results from Test Driven Development)
  - Test framework - set sign-space state, manage test suites
  - Maven build & dependency management
  - Leiningen & NodeJS web test & dependency management
  - GIT code & artifact repositories
  - Jenkins automated build, quality checks, & regression test
  - Ambient visualization of build & test status
- Deployment - Install, config, test on production hardware
  - SemApp Kickstart
    - RHEL + packages install (STIG'd)
    - CentOS
    - Disk partitioning & NIC bonding
  - Procedures for NTP config, cloud infrastructure and SemApp software deploy
  - Test Procedures for configuration and core functionality
  - Passed Area 59 FAT

Move it forward. Make it happen.

Test Data

- Collected from these sources
  - ALOE 1 & 2
    - Imagery, video, WAMI, chat
  - Blue Devil online archive
    - Imagery, video
  - Michelson Lab archives, China Lake
    - Various
  - Open source
    - GovTrack
    - Twitter
    - Wikipedia
    - Federal Contract Spending
    - Audio Metadata
    - US Patents
    - Project Gutenberg
    - Open secrets
    - OpenStreeMap
  - SAIC GSTI
    - LIDAR
    - HSI

Development Backlog

Move it forward. Make it happen.

Ongoing Work

- Assert, Retrieve, Delete, Exists, Tally
  - Geometry extensions
  - Signs with multiple geometries
- API for video playback
- Applications
  - Data Management
  - Foundation data builder
  - Executable process loader
  - MI player
  - Operators console
  - SemApp notebook
  - SemApp suite
  - Tracking
  - User Asserts
  - User SA
  - User Self-registration
  - ...
- Capacity & Stability
  - Caching
  - Cucumber
  - Load balancing
- Deployment Activities
  - Deployment readiness
  - IA
  - Site integration
  - Theater support
- Documentation
  - Application user guides
  - ConOps
  - Data architecture guide
  - Developer's guide
  - Management guide
  - Operator's guide
  - Process documentation
  - Training materials
- Ingest
  - direct from RDB
- Integration
  - Bulk import / export
  - DIB
  - JPIP
  - NVS-ABI
  - SOM
  - Virgo
  - WorldWind
- M&M query performance
- Model services
  - Query broaden / narrow
  - SameAs handling
- Process
  - Advanced reasoning engines
  - Image processing
  - Message chaining
  - Microsoft products
  - NLP analytics
  - RDF
  - User-defined m/r
  - ...
- Mixed modality compound queries

Move it forward. Make it happen.

Recent Release Record

Move it forward. Make it happen.

Release 0.13 Focus

How do we know it's working?

How do we keep it working as it grows?

Testing

- Test case development
  - Functional
    - Data & index services
    - Process services
    - Processes
    - Query
  - Performance
    - Data & index services
    - Ingest
    - Processing
    - Query
  - Configuration
    - Hadoop, Zookeeper, Accumulo
    - SemApp core
    - M&M
    - Enterprise
  - Acceptance
    - Search - Surf application
    - Command-line applications
    - Web services

Performance

Quality

- Test management
  - Organization - test cases, sets, & plans
  - Documentation & CM
  - Continuous integration
  - Manual Testing procedures
- Code quality
  - Coding standards compliant throughout
  - Consistent use of patterns
  - Scalable
  - Thread safe
  - Useful log / error messages
- Monitoring & Metering
  - Operator monitoring console
  - Core health & status indicators
  - Aggregated performance metrics
- Practice
  - Deploy it
  - Stress it
  - Treat it like a production system

Move it forward. Make it happen.

Release 0.14 Focus

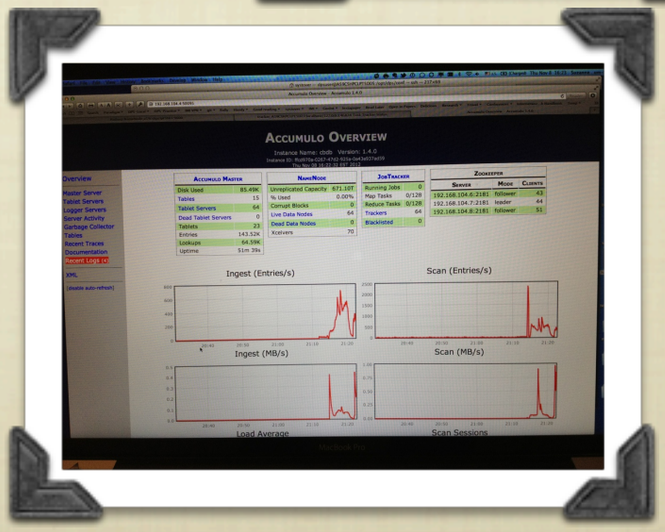
Deploy onto Production Hardware

Move it forward. Make it happen.

The "Paradigm Cluster"

On all 64 data nodes and 15 servers

- Fix initial RHEL install
- Develop SemApp Kickstart
- Partition the drives
- Bond the NICS
- Configure NTP
- Resolve hardware failures
- Work through UCSM issues
- IA scan
- Install, configure, and test cloud infrastructure
- Install, configure, and test SemApp software
- Document / update deployment procedures
- Organize and refine test Procedures for configuration and core functionality
- Participate in Area 59 FAT



Release 0.15 Focus

Administrator, Operator, User Support

- Theater ConOps
- theater work force
- NVS integration
- Applications
  - Search-Surf
  - Data management
  - Ingest
  - process management
- Documentation
  - Applications
  - Operation
  - Management
  - System services / APIs
  - Training

Usability

Release 0.16 Focus

Domain Integration

- Activity Based Intelligence
- Broaden scope of data
- A framework for web applications

Getting together with domain SMEs

Release 0.17 Focus

Front End

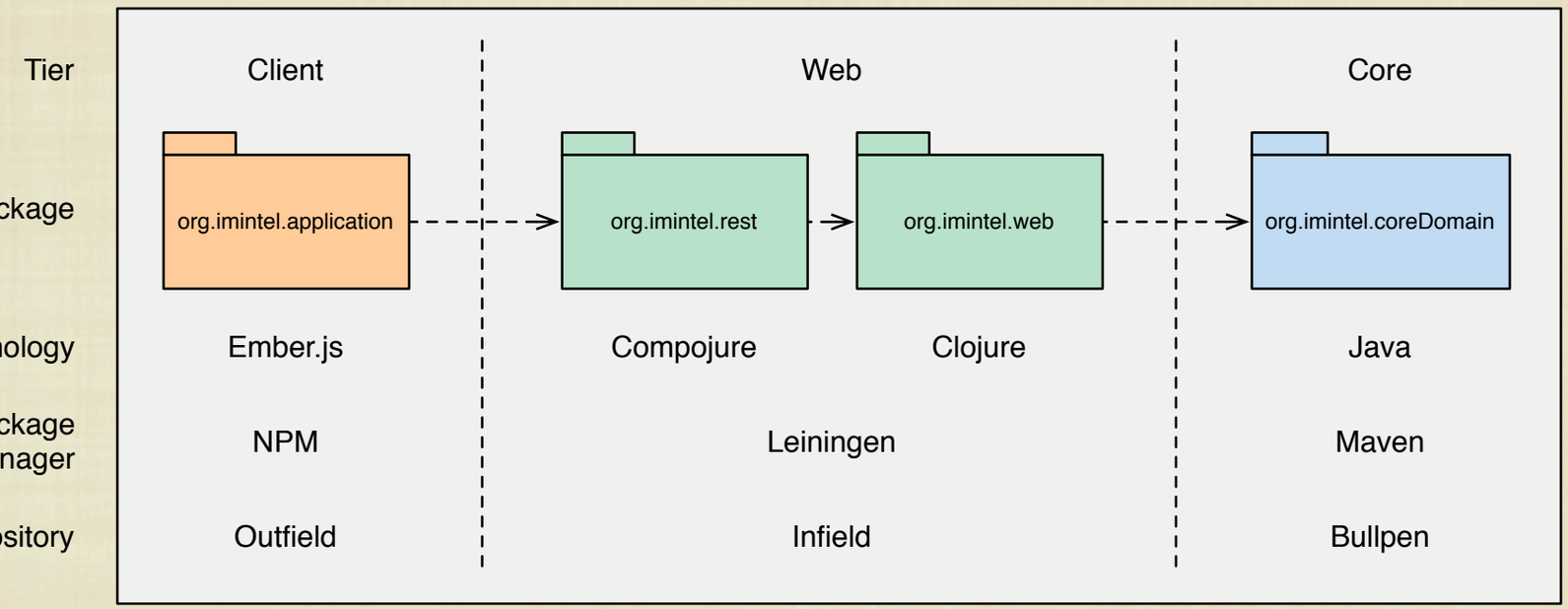
- Look & feel
- Application prototype(s)
- Information visualization
- More GeoINT data types

Move it forward. Make it happen.

Release 0.18 Focus

New Apps

- The Eye
- Data SA



Move it forward. Make it happen.

Looking Forward

Release It!

- Deploy to laboratory / operational (LabOp) environment(s)
- Pursue certification and accreditation to operate
- Integrate with LabOp services (e.g. monitoring)
- Ingest LabOp data
- Process data
- Work with users (e.g. analysts) and LabOp operators
- Interoperate with Enterprise services (e.g. GeoBus, DIB)
- Refine / extend / develop applications, interfaces, services, processes, and domain-models as needed to support LabOp objectives
- Support and maintain system operation

Move it forward. Make it happen.

Release Plans

Move it forward. Make it happen.

R0.13 - R0.19 Schedule

-- Obsolete --

	Sept-Oct 2012	Nov-Dec 2012	Jan-Feb 2013	Mar-Apr 2013	May-Jun 2013	Jul-Aug 2013	Sept-Oct 2013
	R 0.13	R 0.14	R 0.15	R 0.16	R 0.17	R 0.18	R 0.19
Core Services & Processes	Gazetteer JPIP Shape File M&M Performance	Virgo Integration NVS Integration SensrWeb Integration	Query Broaden Query Narrow RDB Ingest Named Entity	SameAs Handling Walk Back Model Export Query Bool Ext	Bulk Export SDA Improvements	Search Engine NVS ABI Integration	User Messaging Automated Reasoning
Applications	Unif Geo Extr File Std Det Monitor	System Mgmt	Process Reg User Self-Reg	User Assert User Ingest	SemApp Notebook Data SA	Search Engine ABI	User SA Mission Scrum
Test & IA	Func, Perf, Conf Tests Test Plans Test Mgmt Acceptance Tests	IV&V	Theater Smoke Test Regression Testing	Regression Testing Func, Perf, Conf Tests Test Plans Acceptance Tests	Regression Testing Func, Perf, Conf Tests Test Plans Acceptance Tests	Regression Testing Func, Perf, Conf Tests Test Plans Acceptance Tests	Regression Testing Func, Perf, Conf Tests Test Plans Acceptance Tests
Deployment & O&M	ConUS Install&Config	IA Scan&Remediate	Theater Deploy Site Integration	Theater Support Config & SW Update New Feature Deploy	Theater Support Config & SW Update New Feature Deploy	Theater Support Config & SW Update New Feature Deploy	Theater Support Config & SW Update New Feature Deploy
Training	Process Docs	Operator's Guide Priv Usr App Guides	Process Reg Guide User Self-Reg Guide Operator Training	User Assert Guide User Ingest Guide User Training	Developers Guide Data Arch Guide Developer Training	Operator Training User Training Developer Training	Operator Training User Training Developer Training

Completed
In progress
Defined

R0.15 - R0.21 Schedule
-- Obsolete --

	Jan-Feb 2013	Mar-Apr 2013	May-Jun 2013	Jul-Aug 2013	Sept-Oct 2013	Nov-Dec 2013	Jan-Feb 2014
	R 0.15	R 0.16	R 0.17	R 0.18	R 0.19	R 0.20	R 0.21
Core Services & Processes	Gazetteer NVS Integration Shape File Index decoupling	ABI Integration Virgo Integration Index associations OWL expressiveness	Bulk Export Model Export RDB Ingest	SameAs Handling Query Bool Ext	Query Broaden Query Narrow Search Engine	Named Entity JPIP	User Messaging Automated Reasoning
Applications	Web app security Data Mgmt Process Mgmt	Operator Console	User Assert User Ingest	SemApp Notebook Data SA	ABI	Search Engine User SA	Process Reg Mission Scrum
Test & IA	Performance Testing	Interface Testing	Theater Smoke Test Regression Testing IA Scan	Regression Testing Theater ATO	Regression Testing	Regression Testing	Regression Testing
Deployment & O&M	Deployment prep	Deployment prep	China Lake Update	Theater Deploy Site Integration	Theater Support Config & SW Update New Feature Deploy	Theater Support Config & SW Update New Feature Deploy	Theater Support Config & SW Update New Feature Deploy
Documentation & Training	Operator's Guide Management Guide Process Docs REST API Doc	ConOps	Data Arch Guide Priv User App Guides Search-Surf Guide	Operator Training Manager Training User Assert Guide User Ingest Guide	User Training Developers Guide Notebook Guide Data SA Guide	Developer Training Training (as needed) ABI Guide	Training (as needed) User SA Guide Mission Scrum Guide

Completed
In progress
Slow / Blocked

R0.15 - R0.21 Schedule

	May-Jun 2013	Jul-Aug 2013	Sept-Oct 2013	Nov-Dec 2013	Jan-Feb 2014	Mar-Apr 2014
	R 0.17	R 0.18	R 0.19	R 0.21	R 0.22	R 0.23
Core Services & Processes	Index p-wave CAPCPO ext FS Detection p-wave TIFF digest LiDAR digest Index predicate	Off Contract				
Applications	Data SA					
Other	Office move Dev ref refactor					

Now

Completed
In progress
Slow / Blocked

Move it forward. Make it happen.

Scale changes everything

Walk with your head in the clouds
and your feet in the dirt

Everything changes when you
deploy to a production
environment

The ground state of
data is not a pure
crystalline form

Everything changes when
you start writing the code

There's More
In_all_of_it_Together
Than we are getting or
could possibly get out of
All of the Parts

Our universe of data & processing is an
Ultra-Large Scale system

New science waits to be discovered

Simple is hard

Diversity gives Intel its richness

Every system should
have the ability to
export its data

Our data & processing assets
belong to the nation
(not the system or the contractor)

What you don't have
in your head, you gotta
have in your feet

The sink is not a
repository for dirty dishes

We are

Be happy, write code!

Godspeed

Don't sell yourself short (the rest of the world will do it for you soon enough)