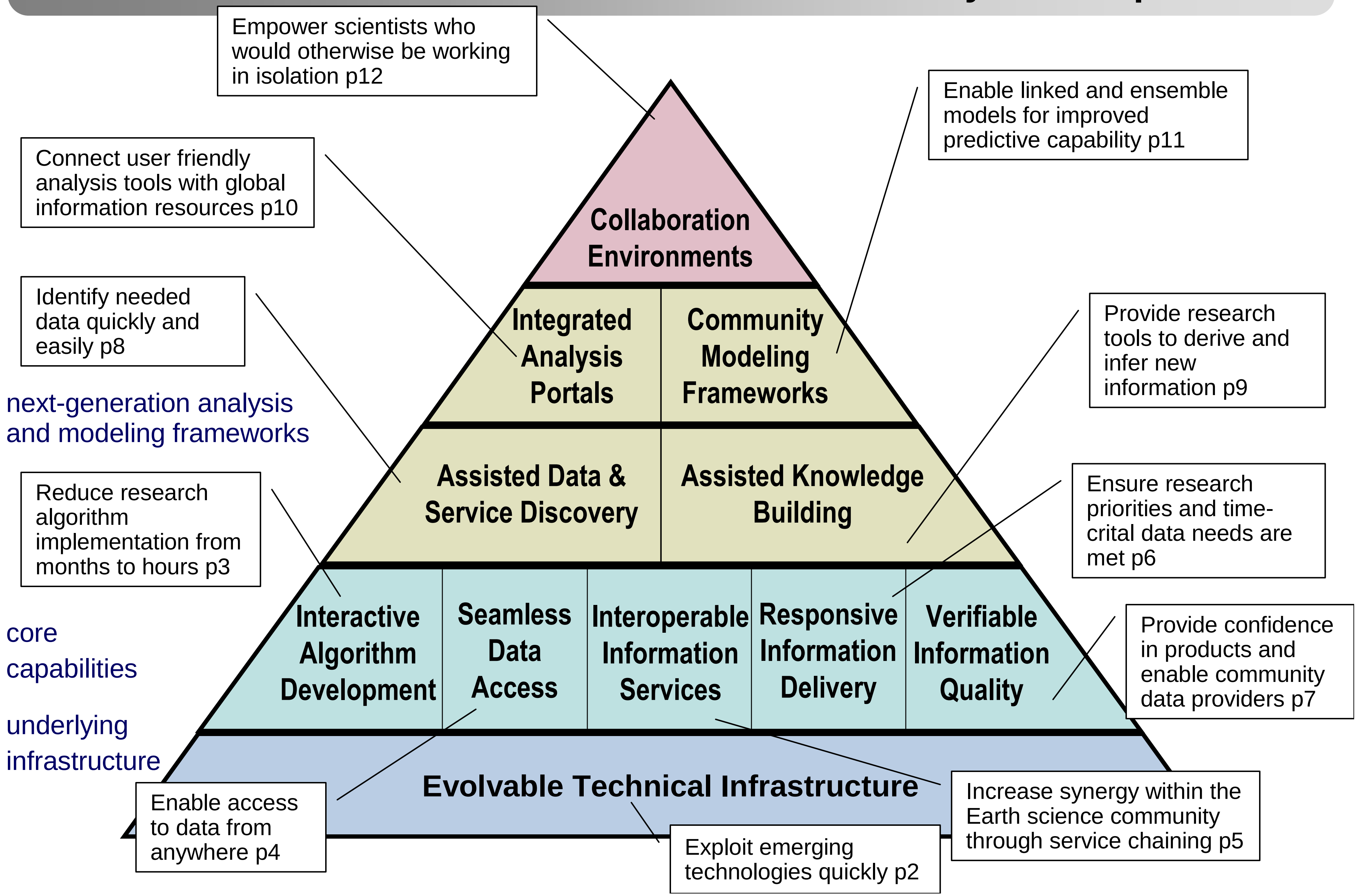
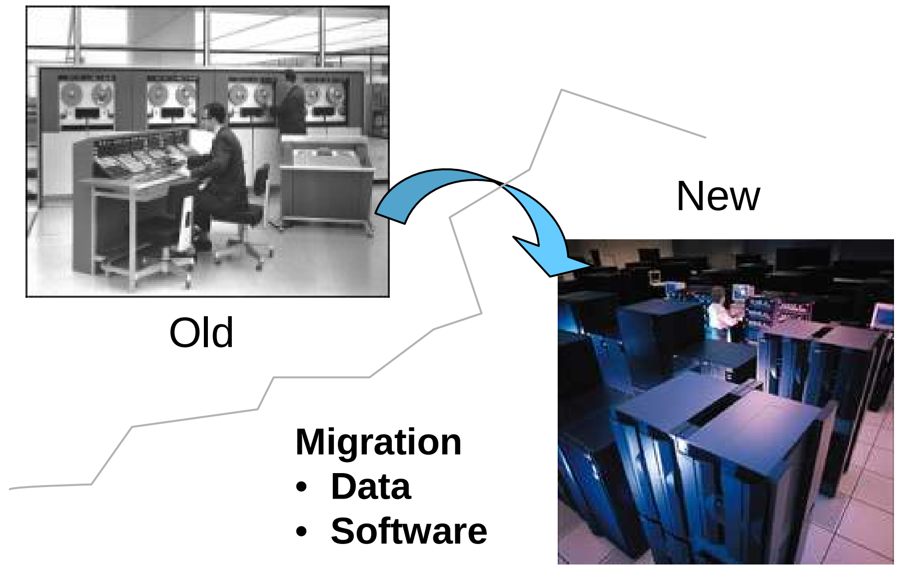
TIWG Earth Science Information Systems Capability Vision
New Information System Capabilities



Evolvable Technical Infrastructure

- Need
  - Researcher needs to take advantage of new processing, storage, and communications technologies to improve performance and reduce costs
- Vision
  - Exploit emerging technologies quickly and maximize utilization of resources
- Enabling technologies
  - Cloud computing, cloud storage, and infrastructure as a service (Amazon EC2 & S3, NASA Nebula)
  - Processor and storage virtualization (VMware, Xen, logical volume management)
  - Scalable architectures (Beowolf, Grid)
  - Sensor webs (SWE)
  - Bandwidth-on-demand
  - Resource allocation control engines (RACE)

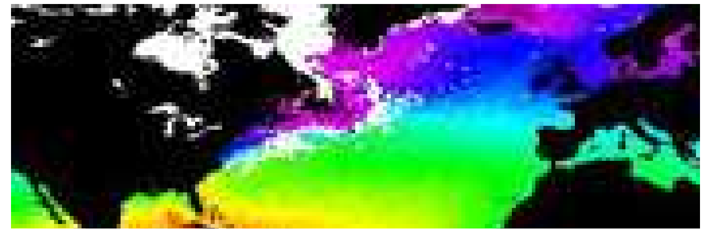
- Current State
  - Network capacity established early in mission and difficult to change
  - Processing, storage, and communications upgrades are difficult and disruptive
    - Manual migration of data
    - Cutover is risky, and parallel operations are costly
    - Communication outages common during upgrades
  - Non-standard interfaces impede introduction of new technologies



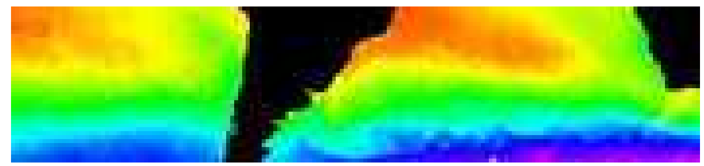
- Future Vision
  - Researcher simply plugs in new equipment to meet storm track model demands
  - Researcher places on-line order for additional processing, storage, and communications capacity based on requirements and budget
  - Additional capacity is obtained dynamically as load dictates
  - Data and processes automatically migrate to take advantage of new equipment or capacity retaining semantic mappings
  - Sensor webs allow temporary tasking of additional sensor resources for high priority observations

Interactive Algorithm Development

- Need
  - Researcher needs to implement a new algorithm in software to calculate ocean heat flux
- Vision
  - Reduce research algorithm implementation from months to hours
- Enabling technologies
  - Visual grammars
  - Visual programming environments (Cantata, Triana, Grist/Viper, Wit)
  - High-level analysis tools (IDL, Matlab, Mathematica)



$$\rho C_{p_g} u \frac{\partial T}{\partial x} = \lambda \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} \right) + G$$

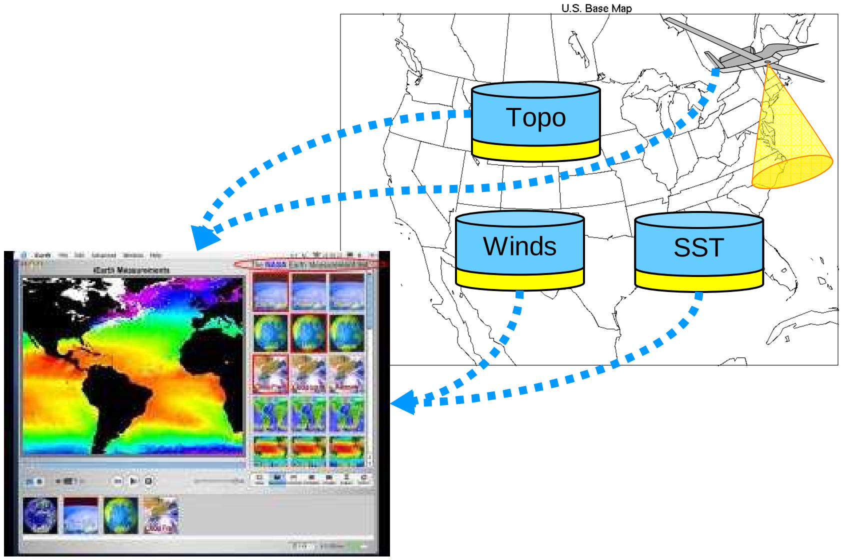


- Current State
  - Coding, debugging, and deploying algorithms takes months of work
  - Algorithms must be implemented by software engineers, not scientists, using custom procedural code
  - Algorithm developers must learn complex application program interfaces for data manipulation and production control
  - Monolithic programming & production environments do not support algorithm sharing

- Future Vision
  - Researcher uses a visual programming environment to create a new heat flux product in hours rather than months
  - Researcher plugs useful transforms created by others into the visual programming environment as needed
  - Researcher analyzes data with interactive tool to identify and quantify relationships between sea surface winds, temperature, topography, and heat transfer
  - Researcher publishes analysis results and relevant metadata as a data product for use in hurricane models

Seamless Data Access

- Need
  - Researcher needs to incorporate a variety of data such as sea winds, sea surface temperature, and ocean topography into the heat flux analysis
- Vision
  - Users can access current data from authoritative sources from any programming environment or analysis tool regardless of the data's physical location
- Enabling technologies
  - Network data access protocols (OpenDAP, WMS/WCS, WebDAV, GridFTP)
  - Sensor web protocols (SWE: SOS, SAS, SensorML, WNS)
  - Established data server tools (MapServer, DODS/LAS, ArcWeb)
  - Semantic metadata (OWL, OWL-S)
  - Single sign-on / shared authentication

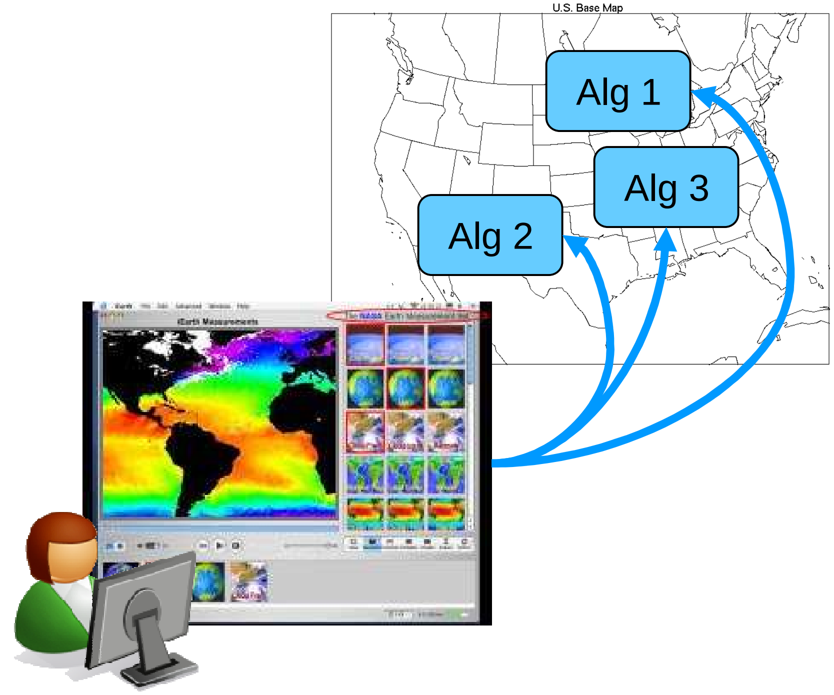


- Current State
  - Data access is broken into separate search, order, and ingest processes
  - Remote data products must first be imported into local storage systems before they can be accessed by analysis tools
  - Different logins are required to access each data product
  - Information on file format and data semantics is not bound to the data and must be manually interpreted

- Future Vision
  - Researcher simply opens remote data and sensing resources from within any analysis tool as if they were local
  - Researcher obtains access to all datasets using single sign-on
  - Data are correctly interpreted and automatically combined by the analysis tool using the associated semantic metadata
  - Sea winds, sea surface temperature, ocean topography, and other data are quickly incorporated into the heat flux analysis

Interoperable Information Services

- Need
  - Researcher needs to incorporate algorithms available at remote locations into the local heat flux analysis
- Vision
  - Increase synergy in the Earth science community by leveraging in-place resources and expertise to provide information services on demand
- Enabling technologies
  - Web service frameworks (SOAP/ WSDL, REST/ WADL)
  - OpenGIS specifications
  - Service orchestration (BPEL, SciFlo, Taverna, Kepler)
  - Sensor web protocols (SWE: SOS)
  - Semantic metadata (RDF, OWL, OWL-S), ESIP data-type and service type ontologies

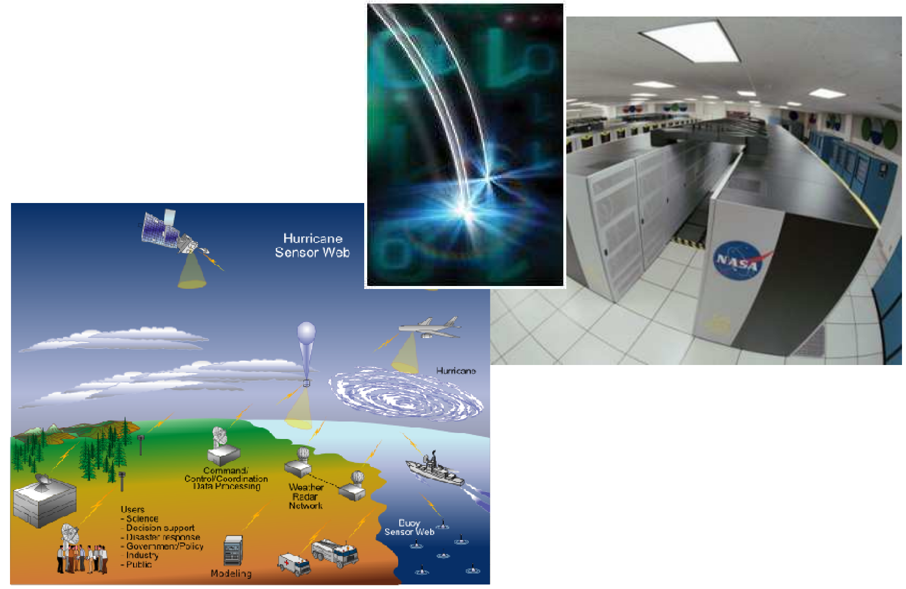


- Current State
  - Remote algorithms must first be ported to the local environment before they can be run
  - Software release policies necessitate obtaining release approvals before algorithms can be transferred
  - Incompatibilities and dependencies sometimes result in recoding of the entire algorithm

- Future Vision
  - Researcher simply invokes remote services and service chains from within the local analysis tool
  - New workflows are dynamically composed and orchestrated to meet the requirements of the researcher’s request
  - Ocean topography data is sent to proven services for sea roughness calculation and reprojection to enhance heat transfer calculation

Responsive Information Delivery

- Need
  - Researcher needs current storm data and real-time observations to update the storm track prediction
- Vision
  - Ensure ready availability of time-critical data
- Enabling technologies
  - Sensor webs (SWE: SPS, SOS, WNS)
  - High speed optical networks (National LambdaRail)
  - Peer-to-peer networks (BitTorrent)
  - Direct downlink (MODIS/AIRS DDL)

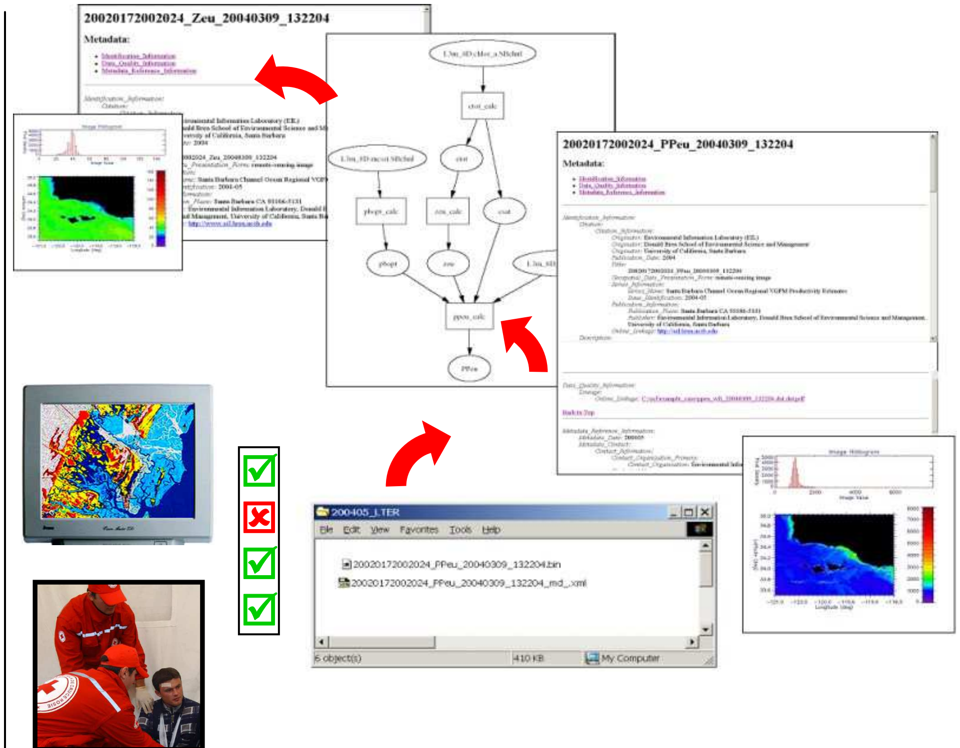


- Current State
  - Static products delivered days after collection
  - Data is stored, cataloged, and delivered in granules that reflect processing and storage constraints more than end user needs
  - First-come first-served data dissemination regardless of intended use

- Future Vision
  - Automated data quality assurance and autonomous operations are used to expedite time-critical data
  - Sensors in the storm path are tasked to acquire new observations
  - Researcher obtains storm data within minutes of sensor overpass based on the application's assigned priority
  - Data are delivered in the preferred format specified in the researcher's profile
  - Data are delivered with the extents and parameter subsets specifically needed by the storm track model

Verifiable Information Quality

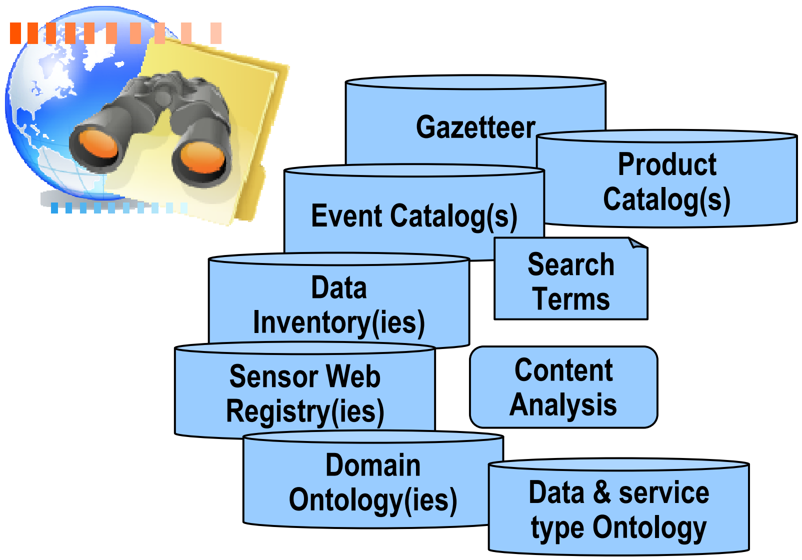
- Need
  - Relief and evacuation planners need to assess the quality of the coastal inundation prediction, which has been based on a long chain of calculations
- Vision
  - Provide confidence in information products and enable the community information provider marketplace
- Enabling technologies
  - Machine-readable formats (XML)
  - Machine readable semantics (information quality ontology)
  - Data pedigree algorithms (e.g., Ellis)



- Current State
  - End user has little insight into the quality of the analysis
  - Data quality is sometimes implicit or assumed based on provider or dataset reputation
  - Non-standard quality indicators cannot be automatically interpreted by COTS analysis software and are sometimes overlooked
  - No machine-readable, standard representation of data lineage
- Future Vision
  - Users can easily explore data pedigree determine its reliability
  - Commercial tools understand data quality flags and automatically handle issues such as missing data
  - Researcher and end user can quantify the quality of the inundation prediction and use the results appropriately

Assisted Data & Service Discovery

- Need
  - Researcher needs to discover datasets, sensor web resources, and other information services required for heat flux calculations
- Vision
  - Identify needed information quickly and easily
  - Automated discovery of resources to support the researcher's goal
- Enabling technologies
  - Data and service description (XML, WSDL, SWE, SensorML, RDF, RDFa, OWL, OWL-S)
  - Service directories (UDDI)
  - Syndication services (RSS, Atom)
  - Domain vocabularies, topic maps, ontologies (SWEET, MMI, VSTO)
  - Established directory services (GCMD, ECHO, THREDDS)
  - Federated search (OpenSearch)
  - Semantic search: facets, synonyms, data integration



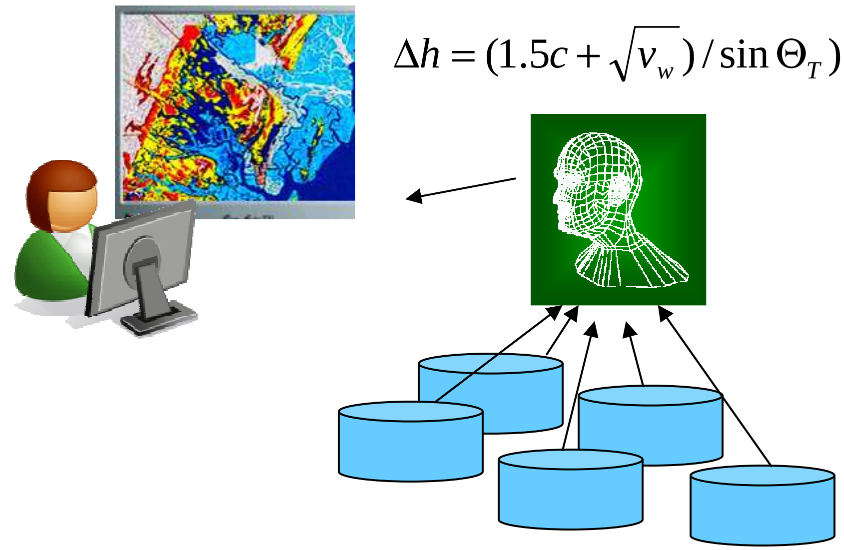
- Current State
  - Manual catalog searches result in dozens of similar datasets, many of which are unsuited to the intended use
  - Inventory searches must be carefully constrained and user must know the exact data product needed, otherwise too much or too little data is returned
  - Disparate catalog approaches impeded cross-catalog searches
- Future Vision
  - Researcher uses semantic and content-based search to search for data and services using physical quantity names, domain-specific jargon, and high-level specifications
  - Federated search enables the simultaneous search of multiple online resources

DRAFT

Assisted Knowledge Building

- Need
  - Researcher needs to determine how the storm track and other storm parameters affect storm surge
- Vision
  - Provide research and operations assistance using intelligent systems to discover granularity in data/information
- Enabling technologies
  - Data mining algorithms and ontologies
  - Data mining toolkits (Adam, D2K, Darwin) and plug-ins (IMAGINE, ENVI, ArcGIS) with semantic annotations
  - Data and service description standards, web service directories, syndication services, topic maps
  - Cross-domain data mining and rule-based smart data mining
  - Environments for ingesting and creating lightly and heavily annotated data and information
  - Semantic languages and tools: Web Ontology Language (OWL) and services ontology (OWL-S).

- Current State
  - Manual generation and testing of hypotheses regarding data interrelationships is time consuming and misses unexpected relationships.
  - Manual analysis misses infrequent events and results in lost opportunities to collect additional data related to the event

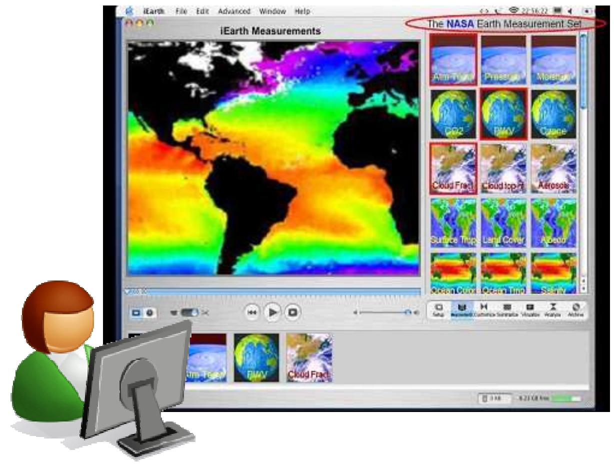


$$\Delta h = (1.5c + \sqrt{v_w}) / \sin \Theta_T)$$

- Future Vision
  - Data mining algorithms automatically infer a statistical model of storm surge based on storm size, angle of track, speed along track, wind speed, lunar phase, coastal shelf depth, and other parameters
  - Researcher combines the inferred model and physical models to create a precision storm surge model
  - Toolkit environments help naïve as well as sophisticated users semi-automatically and automatically create marked up data and knowledge for smart applications with persistent provenance-aware encodings.

Integrated Analysis Portals

- Need
  - Researcher needs to combine a variety of local and remote data products and services to produce a new data product of estimated heat flux at ocean surface boundary
- Vision
  - Connect user friendly analysis tools with global information resources using common semantics.
- Supporting capabilities
  - Assisted data & service discovery
  - Interactive data analysis
  - Seamless data access
  - Interoperable information services
  - Responsive information delivery
  - Verifiable information quality

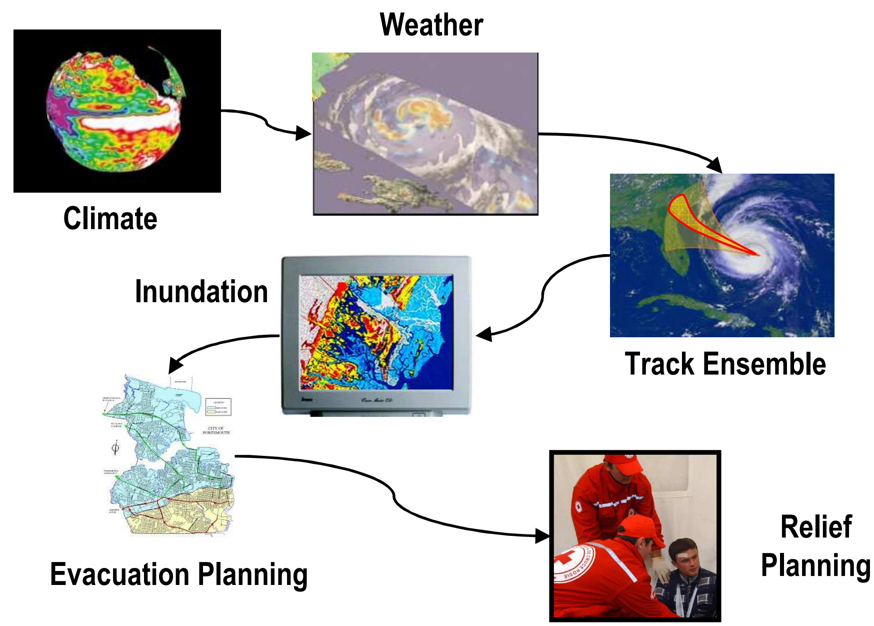


- Current State
  - Researcher combines local and remote data products and services manually with significant time and effort and detailed knowledge of underlying data and services
  - User friendly analysis tools are not common
  - Poor supporting capabilities for discovery, interaction, controlling/ specifying responses verifying information quality

- Future Vision
  - Portal modal functions respond to user selection
  - Easy to plug-in domain terminology
  - Capabilities to propagate provenance and other key metadata
  - Ability to describe analysis functions and results

Community Modeling Frameworks

- Need
  - Researcher needs to couple hurricane forecast model to storm surge model to create more accurate predictions of coastal inundation
- Vision
  - Enable linked and ensemble models for improved predictive capability
- Enabling technologies
  - Multi-model frameworks (ESMF, Tarsier, MCT, COCOLIB)
  - Model data exchange standards (BUFR, GRIB, CF-1)
  - Semantic metadata (domain ontologies, ontology matching)

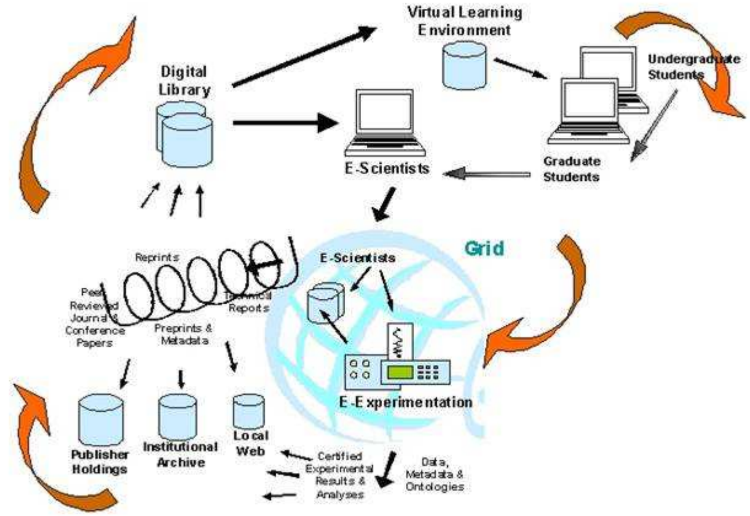


- Current State
  - Disparate and non-interoperable modeling environments with language and OS dependencies
  - Scientific models and remote sensing observations rarely connected directly to decision support systems
  - Evacuation and relief planning based largely on historical averages and seat-of-the-pants estimates
- Future Vision
  - Researcher combines multiple models into an ensemble model to forecast the hurricane's track
  - Researcher couples the storm track model to the storm surge model
  - Analyst assesses property and transportation impact in decision support system fed by storm surge/inundation model

DRAFT

Collaborative Frameworks / Environments

- Need
  - Organic collaboration to shorten the cycle of data to knowledge, share and disseminate the knowledge of the science process/ workflow, and allow reproducibility
- Vision
  - Empower scientists who would otherwise be working in isolation and without full access to modern resources
- Enabling technologies
  - Wiki, Content Management System, blog, IM, micro-blog, web telecon
  - New forms of publication
  - Full provenance



- Current State
  - Most often the only artifact shared is the end product: the publication
  - Workflow sharing only at the end of the cycle; little or no publication, perhaps conference presentation, maybe on a web site (but not integrated, scalable, reproducible, etc.)
  - Collaborations are mostly initiated in person, and there are barriers to entry (e.g. a student may find it impossible to collaborate with a senior scientist)
  - Publishing companies own copyright and dictate the process

- Future Vision
  - Shared environment that scales from two people to workgroup to community to dissemination; fine-grain, versioned access control to all artifacts that are part of the process
  - Full traceability from data to results and back
  - All the current and new tools work seamlessly in support of collaboration and new science; shared and annotated science workflows
  - Science fully democratized
  - Gathering collective knowledge
  - Rapid advancement of science through collaboration
  - Science benefits extend easily to society