



Rensselaer

why not change the world?®

Unstructured Information, Information Management & Discovery, Information Workflows

Ahmed Eleish

Xinformatics 4400/6400

April 02, 2024

Tetherless World Constellation
Rensselaer Polytechnic Institute



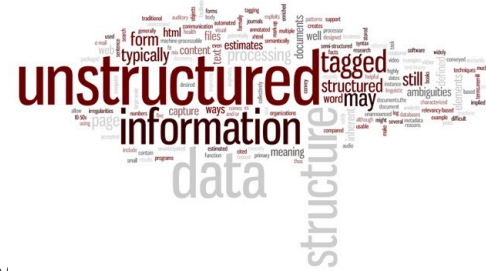
Class Agenda

- Unstructured Information
- Information Management & Discovery
- Information Workflows
- Group Project Check in!



“Many organizations are becoming overwhelmed with the volumes of unstructured information -- audio, video, graphics, social media message -- that falls outside the purview of their "traditional" databases.”

“Organizations that do get their arms around this data will gain significant competitive edge”

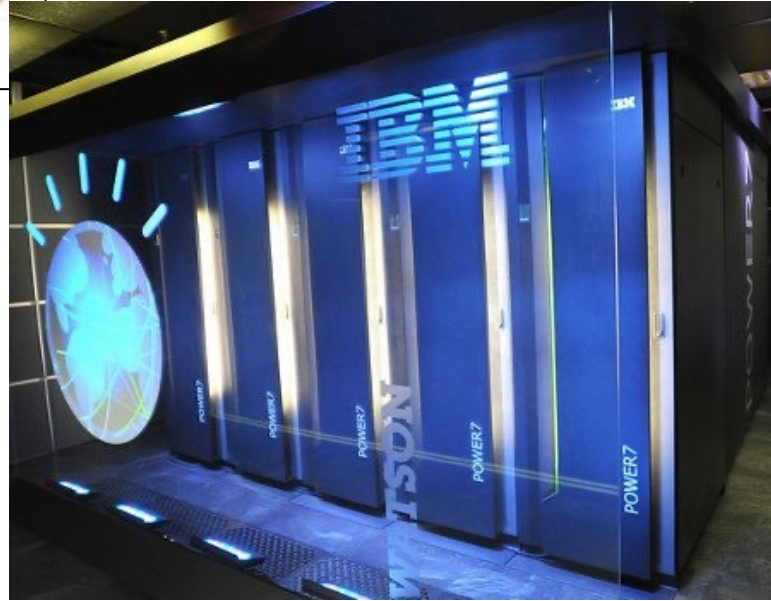


Resource: <https://www.zdnet.com/article/unstructured-data-the-elephant-in-the-big-data-room/>



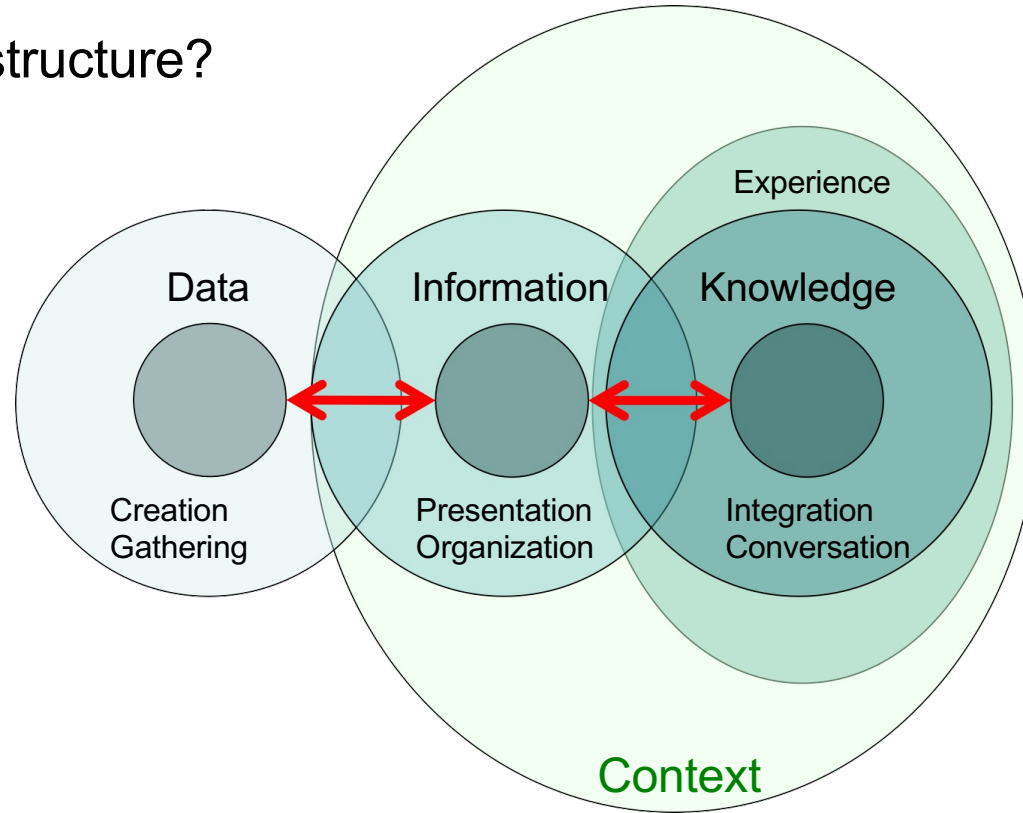
Sources and uses of unstructured information

- audio, video, graphics, social media messages, etc. – that which fall outside the purview of traditional databases



Data<->Information<->Knowledge

- Where is the structure?



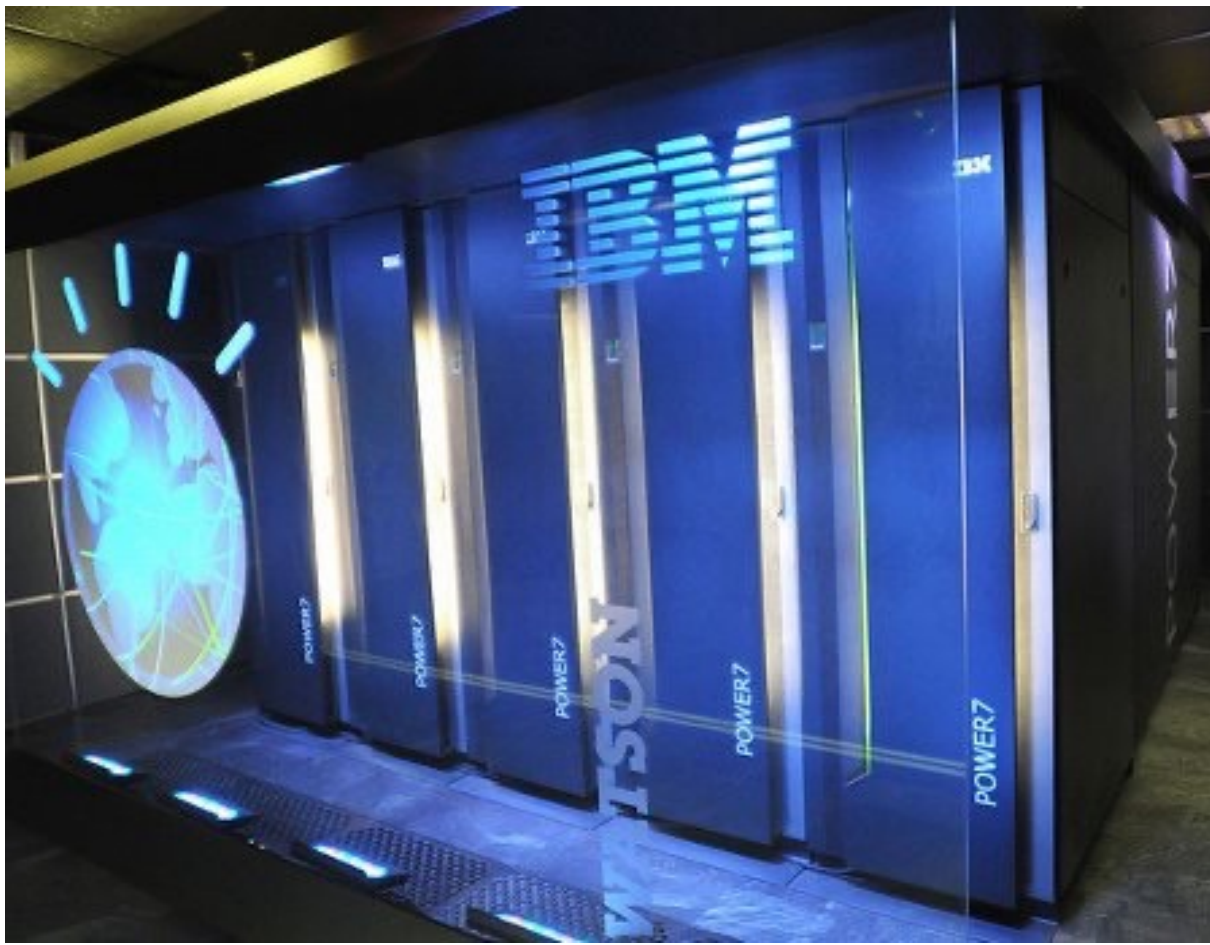
Informatics

- Oh, wait – people *structure* information!
- Cognitive processes
 - Semiotics
 - Mental representation
 - Intuition
 - Expertise
- But not in the same way computers can!



More than 10 years ago...

- Unstructured Information Management Architecture (UIMA) from IBM
 - “Unstructured information management (UIM) applications are software systems that analyze unstructured information (text, audio, video, images, and so on) to discover, organize, and deliver relevant knowledge to the user. In analyzing unstructured information, UIM applications make use of a variety of analysis technologies, including statistical and rule-based Natural Language Processing (NLP), Information Retrieval (IR), machine learning, and ontologies.
 - IBM's Unstructured Information Management Architecture (UIMA) is an architectural and software framework that supports creation, discovery, composition, and deployment of a broad range of analysis capabilities and the linking of them to structured information services, such as databases or search engines.
 - The UIMA framework provides a run-time environment in which developers can plug in and run their UIMA component implementations, along with other independently-developed components, and with which they can build and deploy UIM applications.”



Unstructured Information Management Architecture (UIMA)

- The **bridge from the unstructured world to the structured world** is built through the composition and deployment of these analysis capabilities.

Figure 1. UIMA helps you to build the bridge between unstructured and the structured world

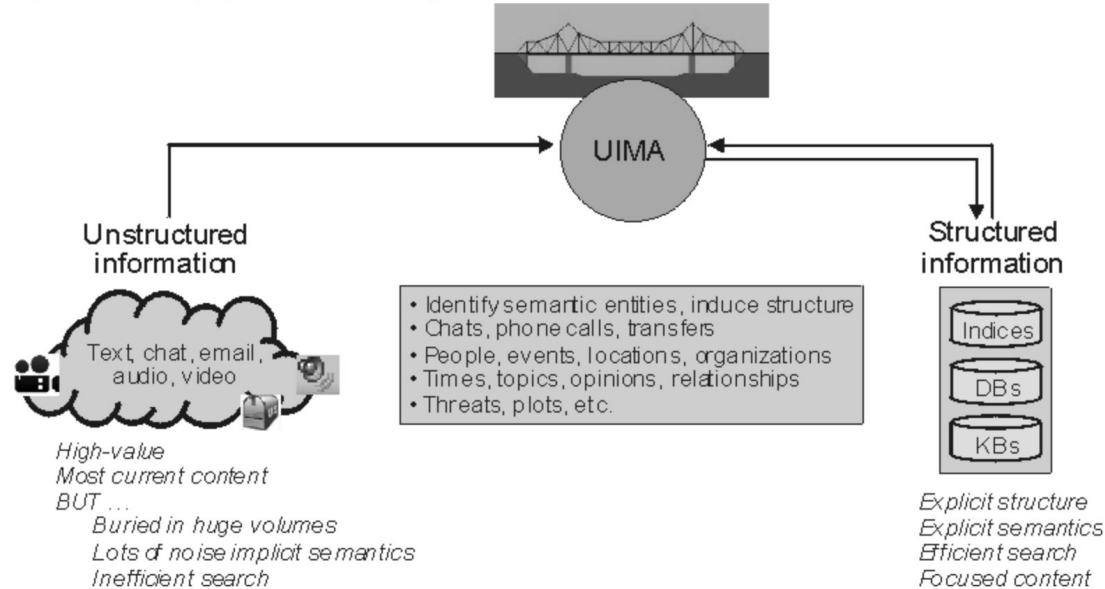


Image Credit: IBM

<https://www.ibm.com/docs/en/db2/9.7?topic=examples-uima-concepts>

Unstructured Information Management Architecture (UIMA)

- **The bridge from the unstructured world to the structured world is built through the composition and deployment of these analysis capabilities.**
- The Unstructured Information Management Architecture (UIMA) is an architecture and software framework that helps you build that bridge.
- It supports creating, discovering, composing, and deploying a broad range of analysis capabilities and linking them to structured information services.

<https://www.ibm.com/docs/en/db2/9.7?topic=examples-uima-concepts>

Information Management & Discovery



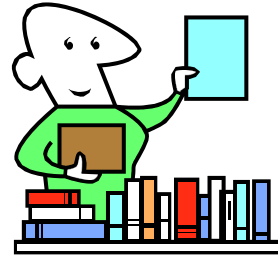
Information Management Activities

- Creation of logical collections
- Physical handling
- Interoperability support
- Security support
- Ownership
- Metadata collection, management and access.
- Persistence
- Dissemination and publication
- Knowledge and information discovery



Logical Collections

- The **primary goal of a Management system is to abstract the physical collection into logical collections**. The resulting view is a *uniform homogeneous* collection.
- Note the analogy with logical models and information integration: so EARLY ON
 - Identifying naming conventions and organization
 - Aligning cataloguing and naming to facilitate search, access, use (who uses?)
 - Provision of ****contextual**** information



Physical Handling

- Map between physical and logical.
- Where and who does it come from?
 - Is there a transfer into a physical form?
 - Is it backed-up, archived, cached? ...
 - What formats?
 - Naming conventions – do they change?



Interoperability Support

- Bit/byte and platform/wire neutral encodings
 - Programming or application interface access
 - Data structure and vocabulary (metadata) conventions and standards
-
- Definition of interoperability?
 - Smallest number of things to agree on so that you do not need to agree on anything else



Security

- What mechanisms exist for securing data?
- Who performs this task?
- Change and versioning (yes, the data may change), who does this, how?
- Who has access?
- How are access methods controlled, audited?
- Who and what – authentication and authorization?
- Encryption and data integrity



Data Ownership

- Rights and policies – definition and enforcement
- Limitations on access and use
- Requirements for acknowledgement and use
- Who and how is quality defined and ensured?
- Who may ownership migrate too?
- How to address replication?
- How to address revised/ derivative products?



Metadata (Data About Data)

- Know what conventions, standards, best practices exist
- Use them – can be hard, use tools
- Understand costs of incomplete and inconsistent metadata
- Understand the line between metadata and data and when it is blurred
- Know where and how to manage metadata and where to store it (and where not to)
- Metadata CAN be added later in many cases

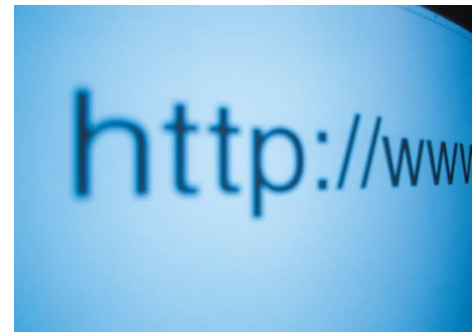
Persistence

- Where will you put your data so that someone else (e.g. one of your class members) can access it?
- What happens after the class, the semester, after you graduate?
- What other factors are there to consider?



Dissemination

- Mechanisms to make interested parties aware of changes and additions to the collections.
- Do you rely on information retrieval? The Web?
- How and what needs to be put in place?
- How to advertise?
- How to inform about updates?
- How to track use, significance?



Discovery

- If you choose so (see ownership and security), how does someone find your data?
- How would you provide discovery of collections, versus files, data points?
- How would you find? =>



Information Management Activities

- Creation of logical collections
- Physical handling
- Interoperability support
- Security support
- Ownership
- Metadata collection, management and access.
- Persistence
- Dissemination and publication
- Knowledge and information discovery

Information discovery

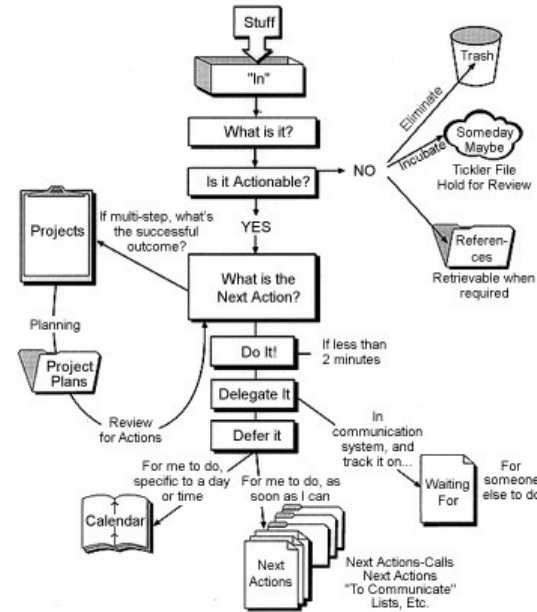
- What makes discovery work?
 - Metadata
 - Logical organization
 - Attention to the fact that someone would want to discover it
 - It turns out that file types are a key enabler or inhibitor to discovery
 - Result ranking using *tuned* algorithm
- What does not work?
 - Dumping project files (excel workbooks) into the cloud!

Information Workflows



Information Workflow

- What is a workflow?
- Why would you use it?
- Key considerations for information, cf. data
- Some pointers to workflow systems



Copyright © 1996-2004, David Allen & Co. All rights reserved.

www.davidco.com

Image Credit: David Allen

What is a workflow?

- General definition: “**series of tasks performed to produce a final outcome**” (taxes?)
- Information workflow – involves people but potentially want to
 - Automate jobs that a person traditionally performed manually
 - Process large volumes of information faster than one could do by hand

Adapted from Slides by Bill Howe UW



Background: Business Workflows

- Example: planning a trip
- Need to perform a series of tasks: book a flight, reserve a hotel room, arrange for a rental car, etc.
- Each task may depend on outcome of previous task
 - Days you reserve the hotel depend on days of the flight
 - If hotel has shuttle service, may not need to rent a car
- Prior information, experience, preferences...

Adapted from Slides by Bill Howe UW



Tripit.com?

Everything looks good, but TripIt Pro will keep monitoring this trip.

San Francisco, CA

Jan 11 - 14, 2013 / San Francisco, CA

Create an itinerary like this by forwarding your travel confirmation emails to plans@tripit.com.

Travelers: Madalynn Priester · Robin Andersen · Richard Adams · Lisa Silveria

Map

Sharing

Options

Print

Fri, Jan 11

San Francisco, CA: AVG: HI 58°F / LO 45°F

Add Plans

Scheduled

4:30
EST PM

8:20
PST PM

Edit
 Copy
 Move
 Delete

New York (JFK) to San Francisco (SFO)
Virgin America 27, Terminal 3, Gate 38

Arrive San Francisco (SFO)
Terminal 2, Gate 53

Confirmation # JD58439
[Get Seat Advice](#)

[More Info](#) ▶

TripIt Pro: Alternate Flights Not Tracking Price

9:00
PST PM

Edit
 Copy
 Move
 Delete

Avis Car Rental (pick-up)
SFO, 780 McDonnell Road San Francisco CA 94128
650-877-6780 , Sun - Sat Open 24 hrs
Confirmation # 7548329

[More Info](#) ▶

9:30
PST PM

Directions from SFO to San Francisco Marriott Marquis ▼

What about information workflows?

- Perform a set of transformations/
operations on information source(s)
- Examples
 - Generating images from raw data
 - Identifying areas of interest from a large
information source (e.g. word cloud)
 - Classifying a set of objects
 - Querying a web service for more information
on a set of objects
 - Many others...

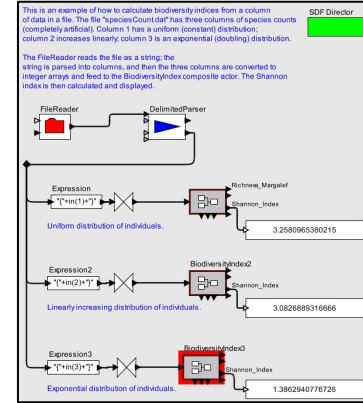
Adapted from Slides by Bill Howe UW



More on Workflows

- Can process many information types:

- Archives
- Web pages
- Streaming/ real time
- Images
- Semiotic systems



- Robust workflows depending on formal (concept and logical) models of the flow of information among components
- May be simple and linear or very complex

Adapted from Slides by Bill Howe UW

Challenges

- Mastering a programming language
- Visualizing workflow
- Sharing/exchanging workflow
- Formatting issues
- Locating datasets, services, or functions



Workflow Management

What is Workflow Management?

<https://www.youtube.com/watch?v=3KJjKY8k9Lk>



Workflow Management Systems

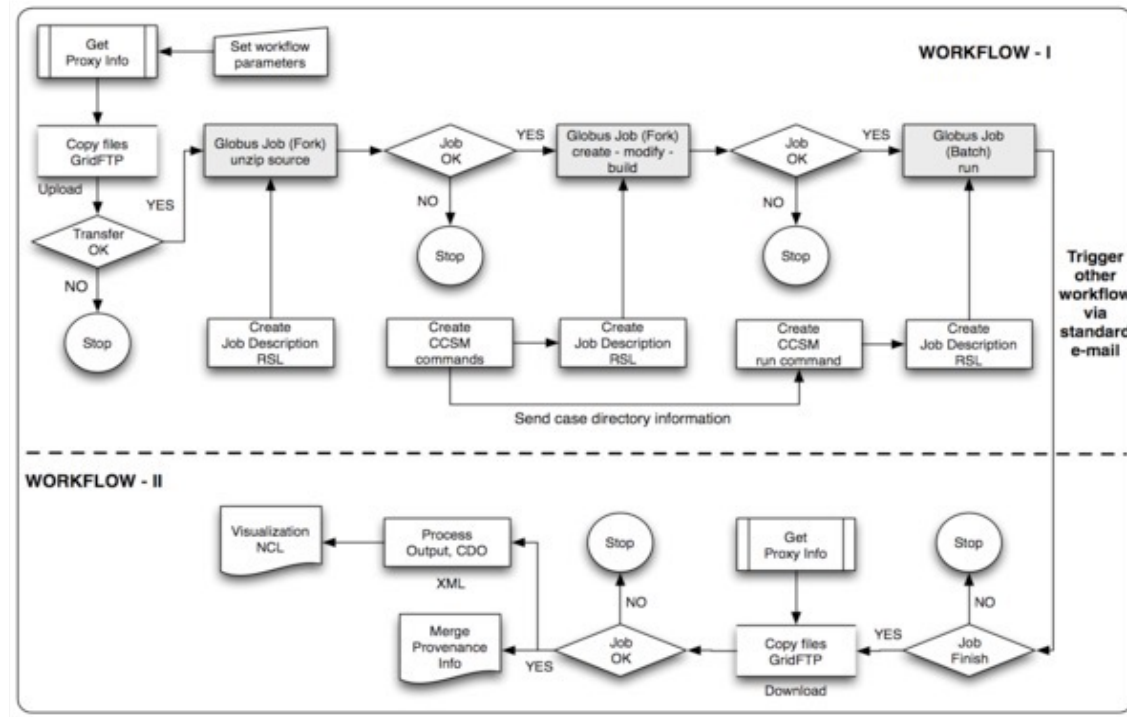


Image Credit: <http://www.earthsystemcurator.org/projects/workflow.shtml>

Benefits of Workflows

- Documentation of aspects of analysis
- **Visual communication of analytical steps**
- **Ease of testing/debugging**
- **Reproducibility**
- Reuse of part or all of workflow in a different project

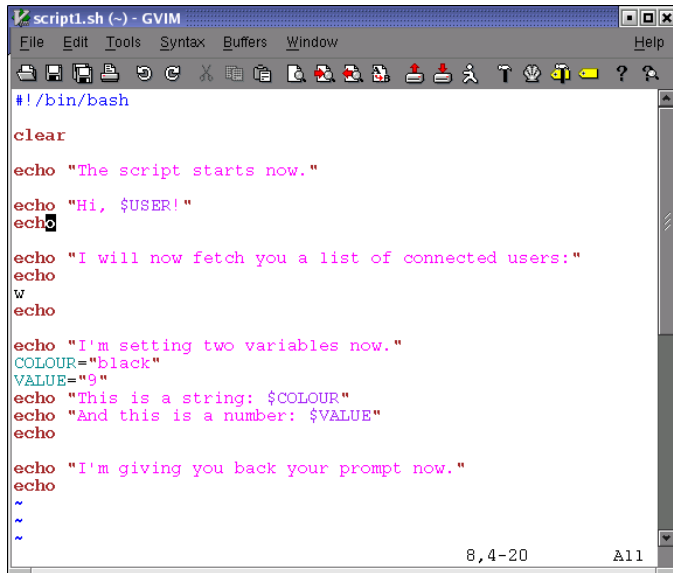


Adapted from Slides by Bill Howe UW

Image Credit: http://www.prestek.com/Collateral/Images/English-US/WorkflowConcept_vf.jpg

Why not just use a script?

- Script does not specify low-level task scheduling and communication
- May be platform-dependent
- Can't be easily reused
- May not have sufficient documentation to be adapted for another purpose



```
script1.sh (-) - GVIM
File Edit Tools Syntax Buffers Window Help
#!/bin/bash

clear

echo "The script starts now."

echo "Hi, $USER!"
echo

echo "I will now fetch you a list of connected users:"
echo
echo
echo

echo "I'm setting two variables now."
COLOUR="black"
VALUE="9"
echo "This is a string: $COLOUR"
echo "And this is a number: $VALUE"
echo

echo "I'm giving you back your prompt now."
echo
~
~
~

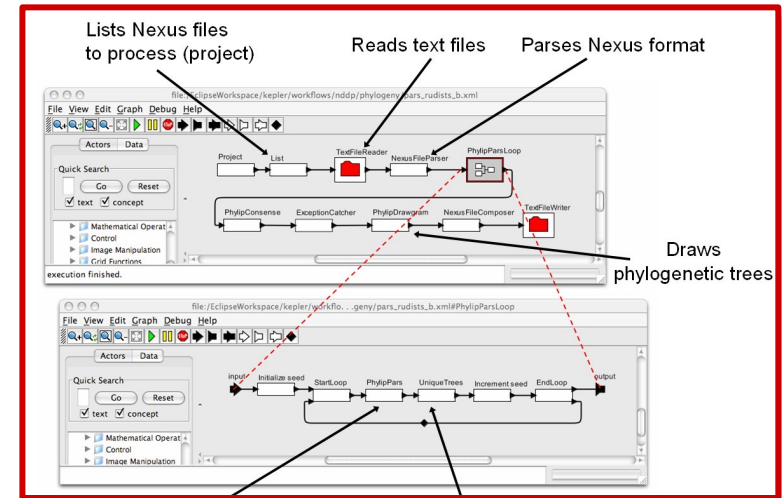
8, 4-20 All
```

Based on Bill Howe

Image Credit: <http://tldp.org/LDP/Bash-Beginners-Guide/html/images/script1.sh.png>

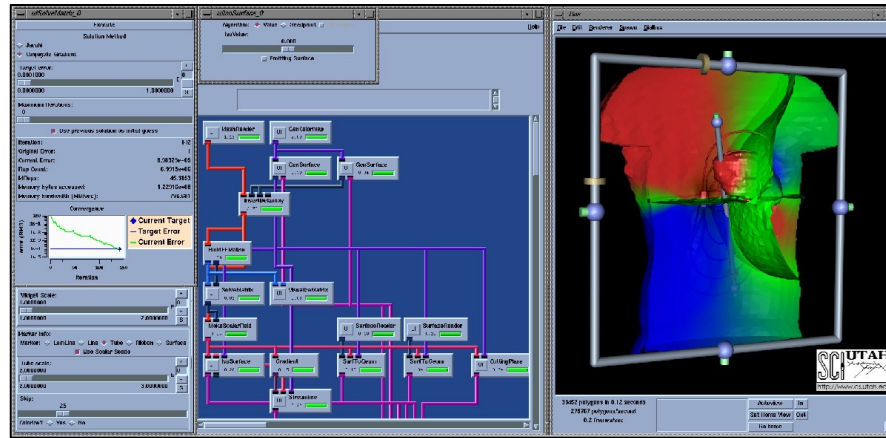
Why can a GUI be useful?

- No need to learn a programming language
- **Visual representation of what workflow does**
- Allows you to monitor workflow execution
- Enables user interaction (though not necessarily collaboration)
- Facilitates sharing of workflows



Some workflow systems

- **Kepler**
- **SCIRun**
- **Sciflo**
- **Triana**
- **Taverna**
- **Pegasus**



- Some commercial tools:
 - **Windows Workflow Foundation**
 - **Mac OS X Automator**
- <http://www.isi.edu/~gil/AAAI08TutorialSlides/5-Survey.pdf>
- <http://www.isi.edu/~gil/AAAI08TutorialSlides/>
- <https://kepler-project.org/>

Check-in for Project Assignment



Thanks!

