Building Scalable Tools for Open edX Learning Analytics

Lauren Milechin*, Julie Mullen, Jeremy Kepner, Albert Reuther

2016 Open edX Conference

June 15, 2016

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering.

Outline

➡

- Introduction
- D4M and Analytics Pipeline
- Demo
- Ground Truth Data
- Results
- Conclusion

LLX Overview

LLx

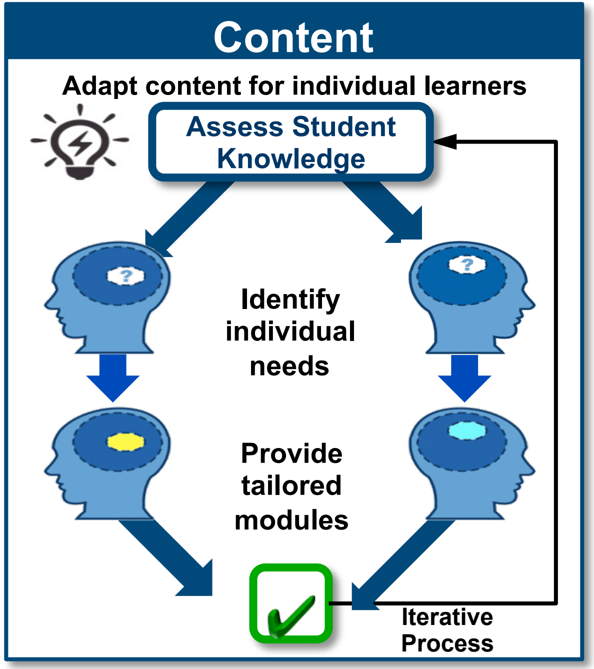
LLX provides online, self-paced and blended technical professional education as part of Lincoln Laboratory’s education portfolio

Goals

The LLx Team works with Laboratory staff to:

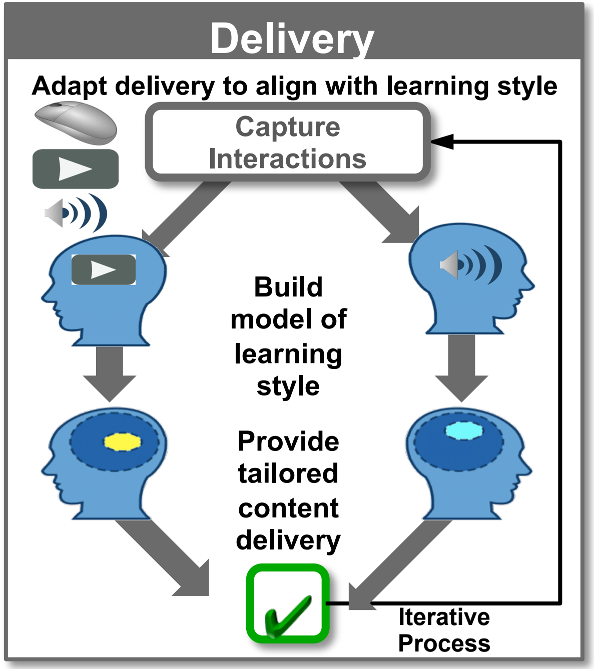
- Provide education at scale to assist the DoD in fulfilling its educational needs
- Transition Lincoln Laboratory technologies and expertise through course offerings
- Explore the pedagogy of new learning and teaching paradigms

Pedagogical Research
Empowering Adaptive Learning



Content

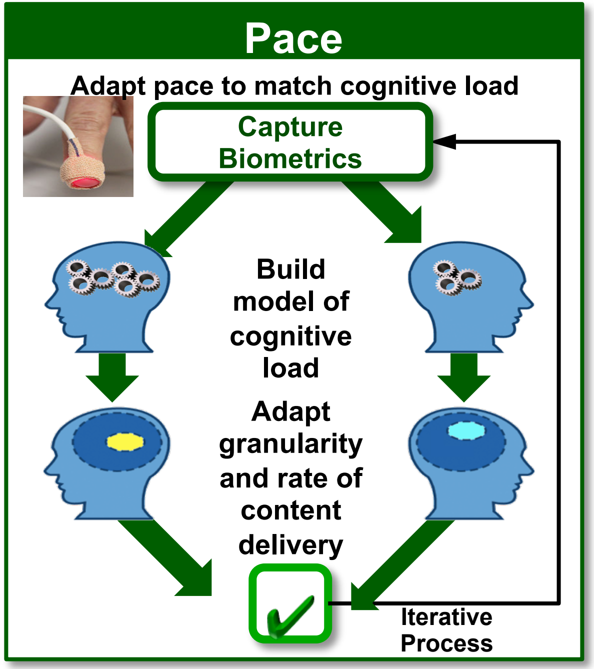
Adapt content for individual learners



Delivery

Adapt delivery to align with learning style

Capture Interactions
Build model of learning style
Provide tailored content delivery
Iterative Process



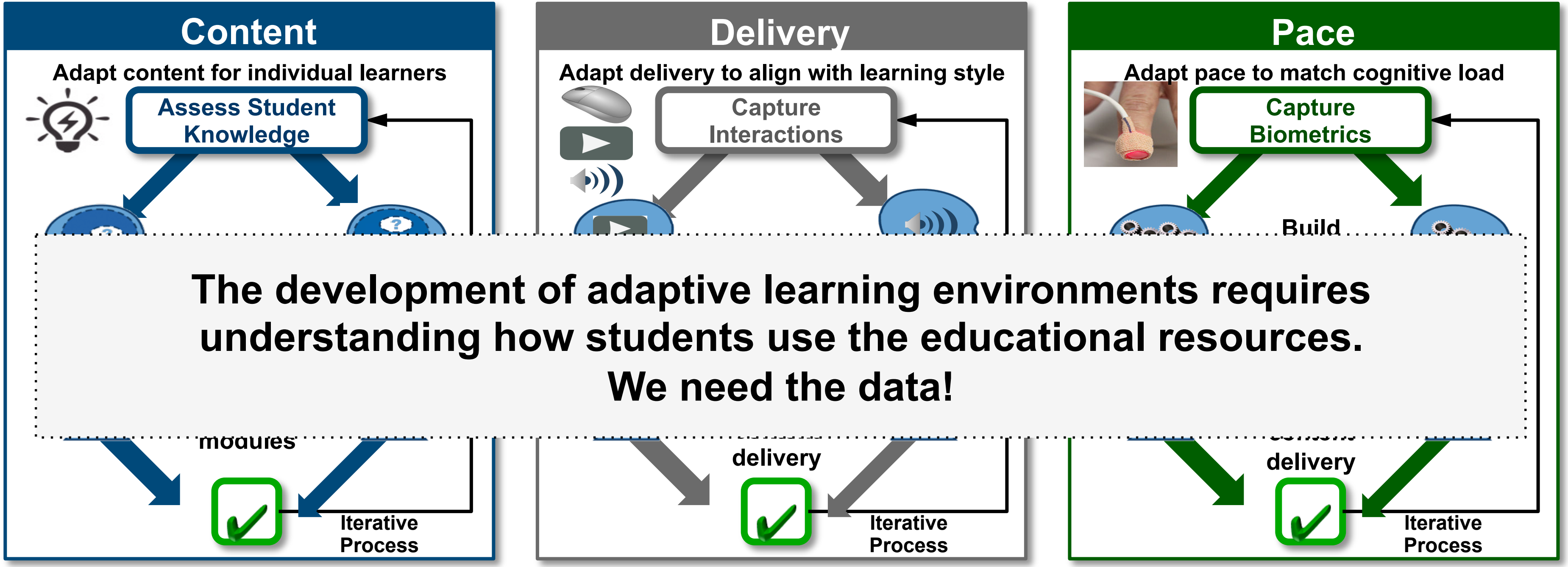
Pace

Adapt pace to match cognitive load

Capture Biometrics
Build model of cognitive load
Adapt granularity and rate of content delivery
Iterative Process

Adapting to student content, delivery, and pace needs yields deeper learning.

Pedagogical Research
Empowering Adaptive Learning



Content

Adapt content for individual learners

Assess Student Knowledge

modules

Iterative Process

Delivery

Adapt delivery to align with learning style

Capture Interactions

delivery

Iterative Process

Pace

Adapt pace to match cognitive load

Capture Biometrics

Build

delivery

Iterative Process

The development of adaptive learning environments requires understanding how students use the educational resources.

We need the data!

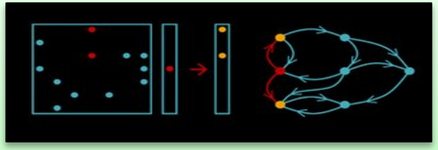
Adapting to student content, delivery, and pace needs yields deeper learning.

Pedagogical Research
Capturing Interactions

Data Collection

Capture Interactions

```
{"username": "", "event_type": "/notifier_api/v1/users/", "ip": "
7.3 Linux/3.13.0-32-generic", "host": "localhost", "referer": "",
\": {\"page\": [\"1\"], \"page_size\": [\"10\"]}}", "event_source
"course_id": "", "path": "/notifier_api/v1/users/"}, "time": "201
{"username": "", "event_type": "/notifier_api/v1/users/", "ip": "
7.3 Linux/3.13.0-32-generic", "host": "localhost", "referer": "",
\": {\"page\": [\"1\"], \"page_size\": [\"10\"]}}", "event_source
"course_id": "", "path": "/notifier_api/v1/users/"}, "time": "201
{"username": "", "event_type": "/notifier_api/v1/users/", "ip": "
7.3 Linux/3.13.0-32-generic", "host": "localhost", "referer": "",
\": {\"page\": [\"1\"], \"page_size\": [\"10\"]}}", "event_source
"course_id": "", "path": "/notifier_api/v1/users/"}, "time": "201
{"username": "JMullen", "event_type": "/course/course-v1:LLX+LLX0
```

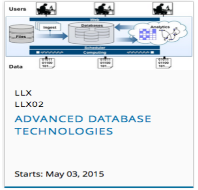


D4M: A Lincoln Laboratory Technology

Pedagogical Research

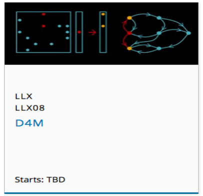
Assess student knowledge

- Module tailoring
- Module pacing / granularity
- New tools & techniques



LLX
LLX02
ADVANCED DATABASE TECHNOLOGIES

Starts: May 03, 2015



LLX
LLX08
D4M

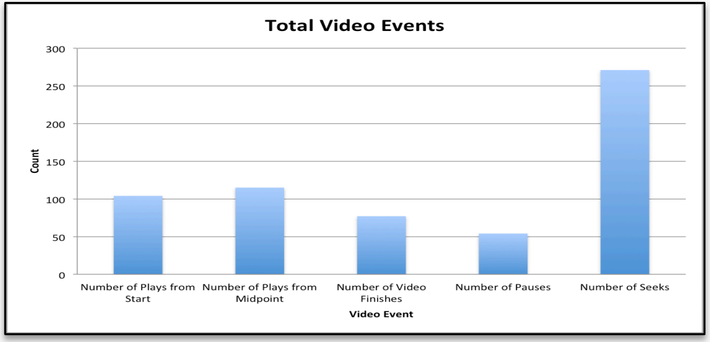
Starts: TBD

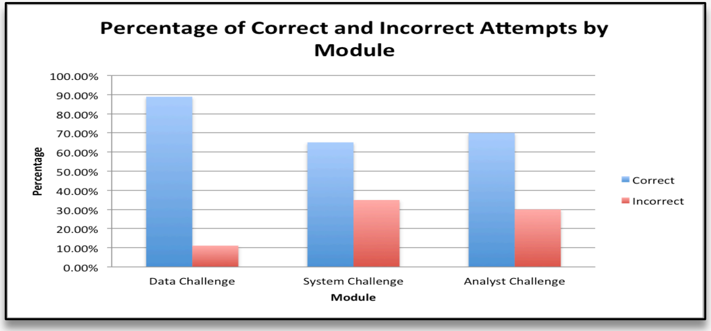
Courseware

Pedagogical Research
Preliminary Data Analytics

Analytics of Interest:

- Develop insight into completion of
  - Entire course
  - Individual sections and units
  - Individual videos
- Discover which
  - Sections or units receive most attention
  - Videos receive most attention
  - Questions are most troublesome
- Profile student paths
  - Linear
  - Topic by topic
  - Relation to student background
  - Responses to wrong questions



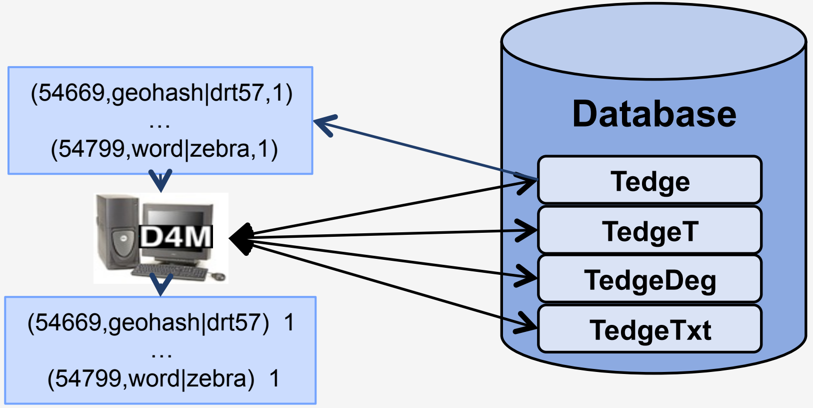


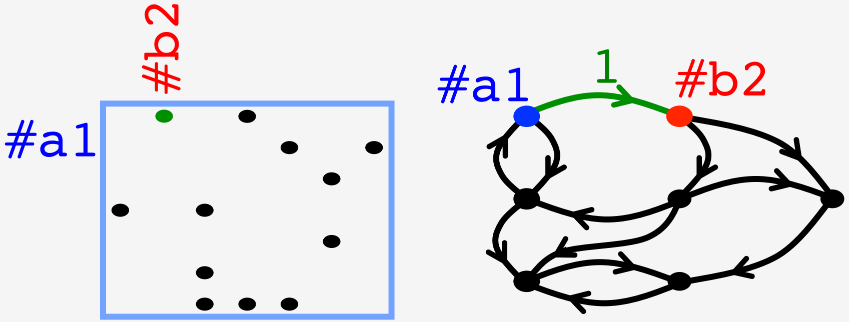
Outline

- Introduction
- D4M and Analytics Pipeline
- Demo
- Ground Truth Data
- Results
- Conclusion

D4M:
Dynamic Distributed Dimensional Data Model

- Library that allows you to
  - Represent data as Associative Arrays
  - Manipulate data using linear algebraic operations
  - Connect to and query high-performance databases
- Associative Arrays
  - Two keys mapped to one value
  - Similar to matrices with string indices
  - Easily represent graphs
  - Closed under algebraic and set operations
  - Composable array indexing
- Website: http://d4m.mit.edu



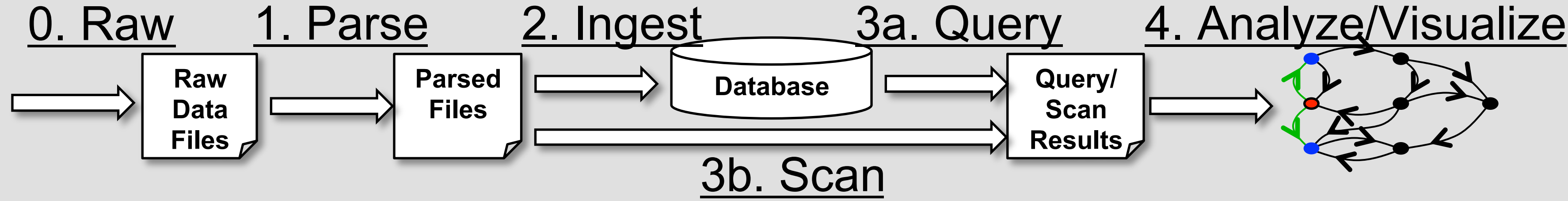


Advantages of D4M

- Associative Arrays can
  - Represent diverse types of data
  - Support a large variety of linear algebraic operations
- D4M is easy to set up and use
  - Download the library, then add the directory to your path in MATLAB® or Octave
  - Many native matrix functions and operations are overloaded to work seamlessly with Associative Arrays
- Great for
  - Rapid-prototyping analytics
  - Interactive data exploration
- With the right schema, easy to query for the data you need

Analytics Pipeline

Pipeline

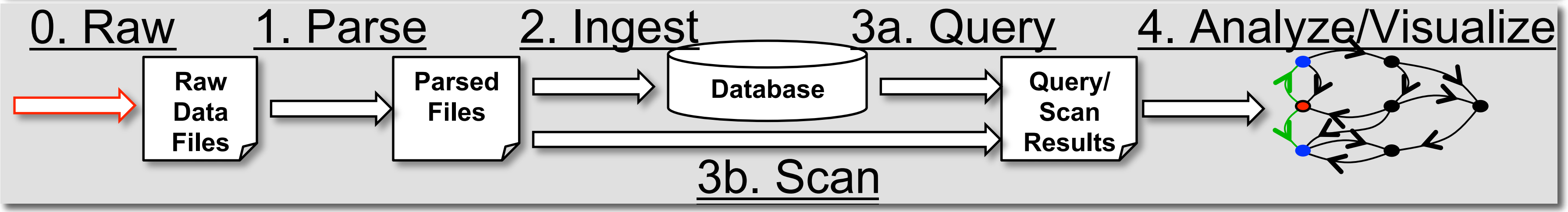


Steps

- Step 0: Retrieve raw data files
- Step 1: Parse raw data files
- Step 2: Ingest parsed data into a database (if needed)
- Step 3: Query database/Scan filesystem
- Step 4: Analyze and visualize the data

Analytics Pipeline

Pipeline



0. Raw

1. Parse

2. Ingest

3a. Query

4. Analyze/Visualize

Raw Data Files

Parsed Files

Database

Query/ Scan Results

3b. Scan

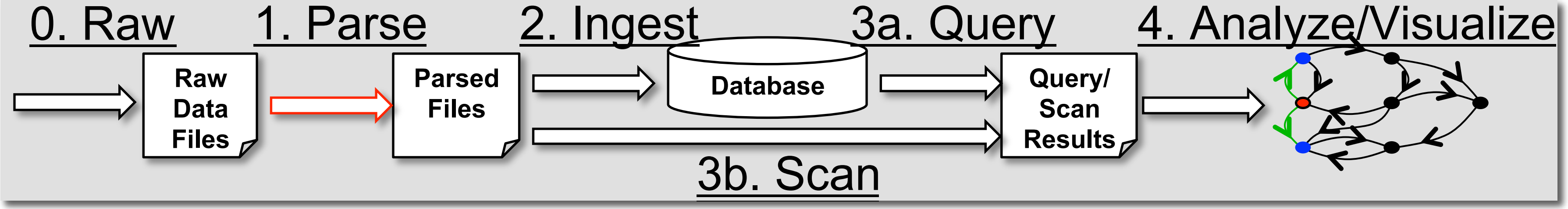
Step 0: Retrieve Raw Data

- Open edX Tracking Logs
- Located in course platform VM
- JSON format
- Logs transferred daily

```
{"username": "Lauren”,
 "event_type": “course.enrollment.activated”,
 "time": "2015-11-05T15:40:59.483662+00:00”,
 "session": “6435efe56dead9ad53438e662f0c14b”,
 "event": {
      "course_id": "course-v1:LLX02_ADT",
      "user_id": 9,
      "mode": "honor”}
}
```

Analytics Pipeline

Pipeline



Step 1: Parse Raw Data

- Parse JSON into D4M
- Written in MATLAB®
- Saved as .mat files
- Row keys: unique identifier for each event
- Column keys: concatenated attribute and corresponding value

```
"username": "Lauren"  ⟹  username|Lauren
```

Analytics Pipeline

Step 1: Parse Raw Data

```
{"username": "Lauren”,
 "event_type":
“course.enrollment.activated”,
 "time":
"2015-11-05T15:40:59.483662+00:00”,
 "session":
“6435efe56dead9ad53438e662f0c14b”,
 "event": {
     "course_id": "course-v1:LLX02_ADT",
     "user_id": 9,
     "
}
```

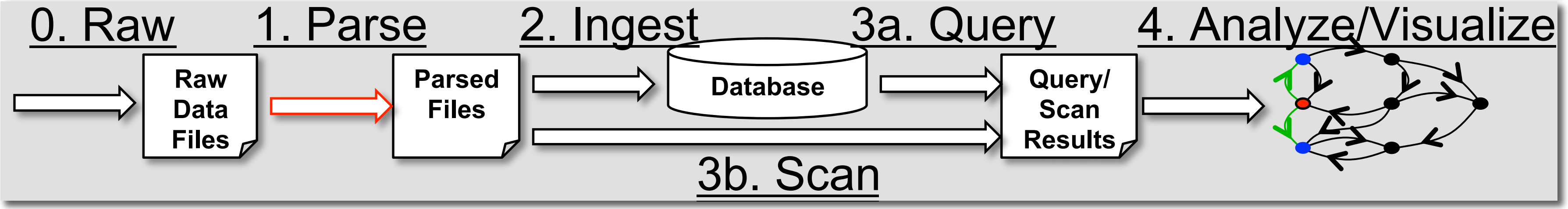
Parse

```
(201511050306,course_id|course-v1:LLX02_ADT) 1
(201511050306,enrollment_mode|honor)         1
(201511050306,event_type|
course.enrollment.activated)                 1
(201511105306,event|explicit)                1
(201511050306,time|
2015-11-05T15:40:59.483662+00:00)            1
(201511050306,user_id|9)                     1
(201511050306,username|Lauren)               1
```

	…	course_id| LLX01_D4M	course_id| LLX02_ADT	…	encrollment_mode| honor	…	event_type| enrollment.actiavated	…	user_id|8	user_id|9
…										
0306			1		1		1			1
…										
502		1			1		1		1	
…										

Analytics Pipeline

Pipeline

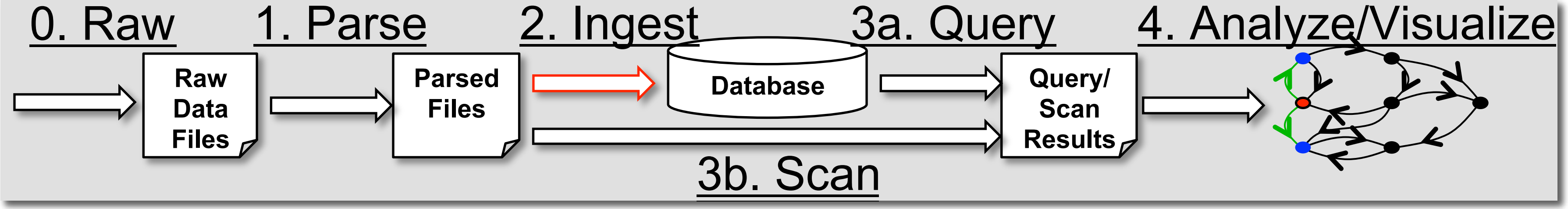


Step 1: Parse Raw Data

- Added:
  - Current unit, section, module names
  - Previous unit, section, module names
  - Whether an event is page navigation or explicit
- Removed:
  - Irrelevant server issued events

Analytics Pipeline

Pipeline

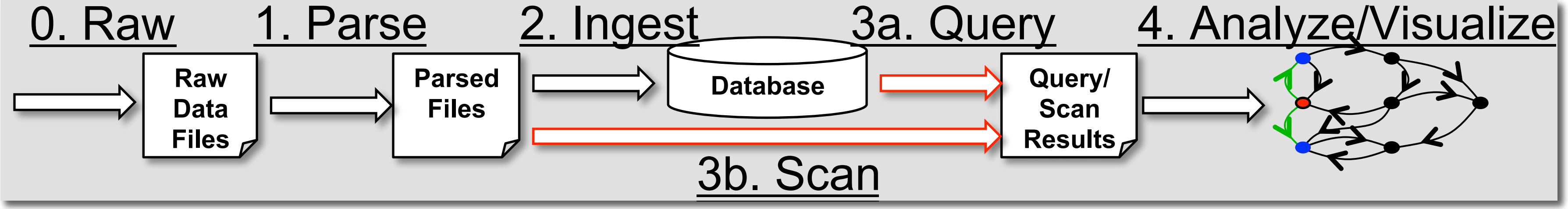


Step 2: Ingest

- D4M Associative Arrays can be easily ingested to a variety of databases
- File system is currently sufficient
- Will use Accumulo
    - NoSQL triple-store database for large, sparse data
    - Cell-level visibility labels ensure instructors see only their student’s data
- Parsed data files only need to be loaded and then inserted into Accumulo

Analytics Pipeline

Pipeline

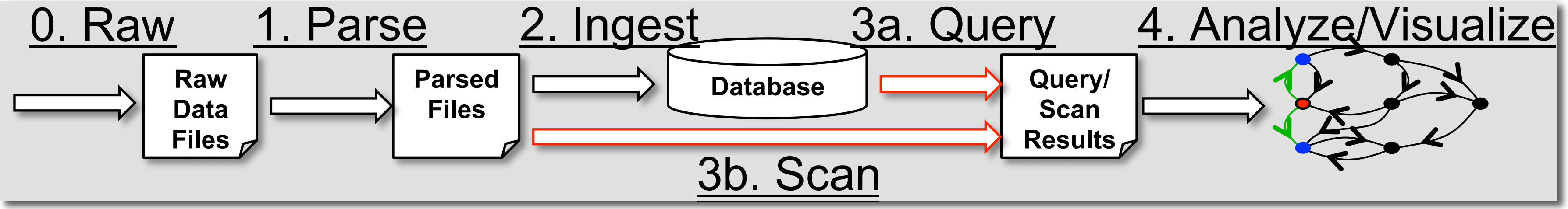


Step 3: Query/Scan

- Getting the data relevant to your analytic
    - Ex: Get all events triggered by students of a particular course
- Retrieve relevant rows by specifying columns of interest

Analytics Pipeline

Pipeline



0. Raw

1. Parse

2. Ingest

3a. Query

4. Analyze/Visualize

Raw Data Files

Parsed Files

Database

Query/ Scan Results

3b. Scan

Step 3: Query/Scan

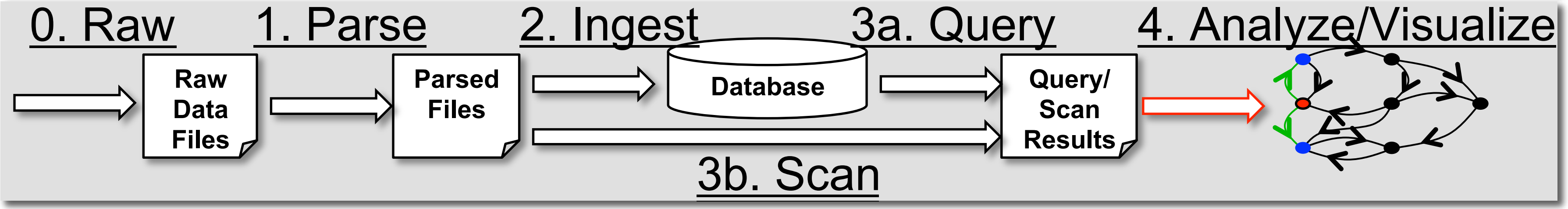
	…	course_id|LLX01_D4M	course_id|LLX02_ADT	…	user_id|8	user_id|9
…						
0306			1			1
…						
502		1			1	
…						

```
ids=Row(A(:,'course_id|LLX02_ADT'));
A_LLX02=A(ids,:);
```

	…	course_id|LLX02_ADT	…	user_id|9
…				
0306		1		1
…				

Analytics Pipeline

Pipeline

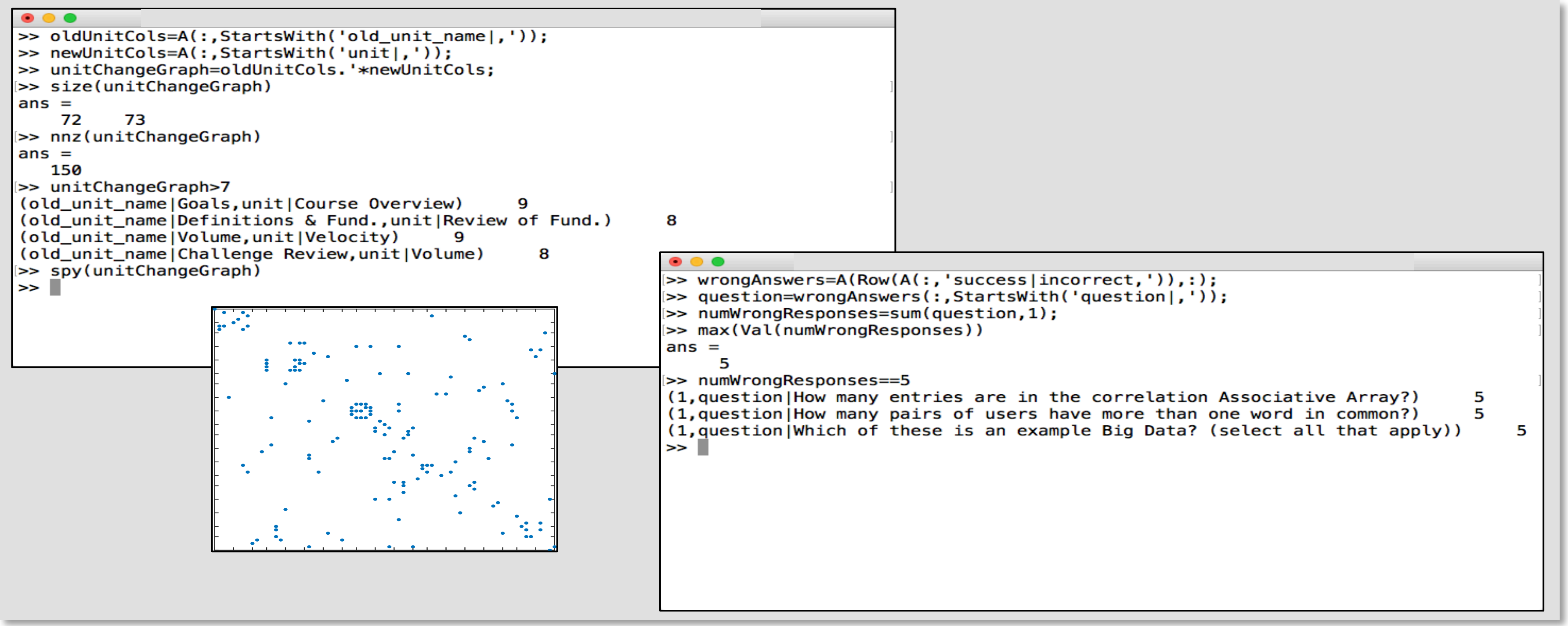


Step 4: Analyze/Visualize

- D4M Associative Arrays support addition, subtraction, matrix and element-wise multiplication, summing, etc
- One matrix multiplication on columns of interest yield adjacency matrix of a graph
  - Most graph algorithms can be expressed in terms of matrix operations on the adjacency matrix

Analytics Pipeline

Step 4: Analyze/Visualize



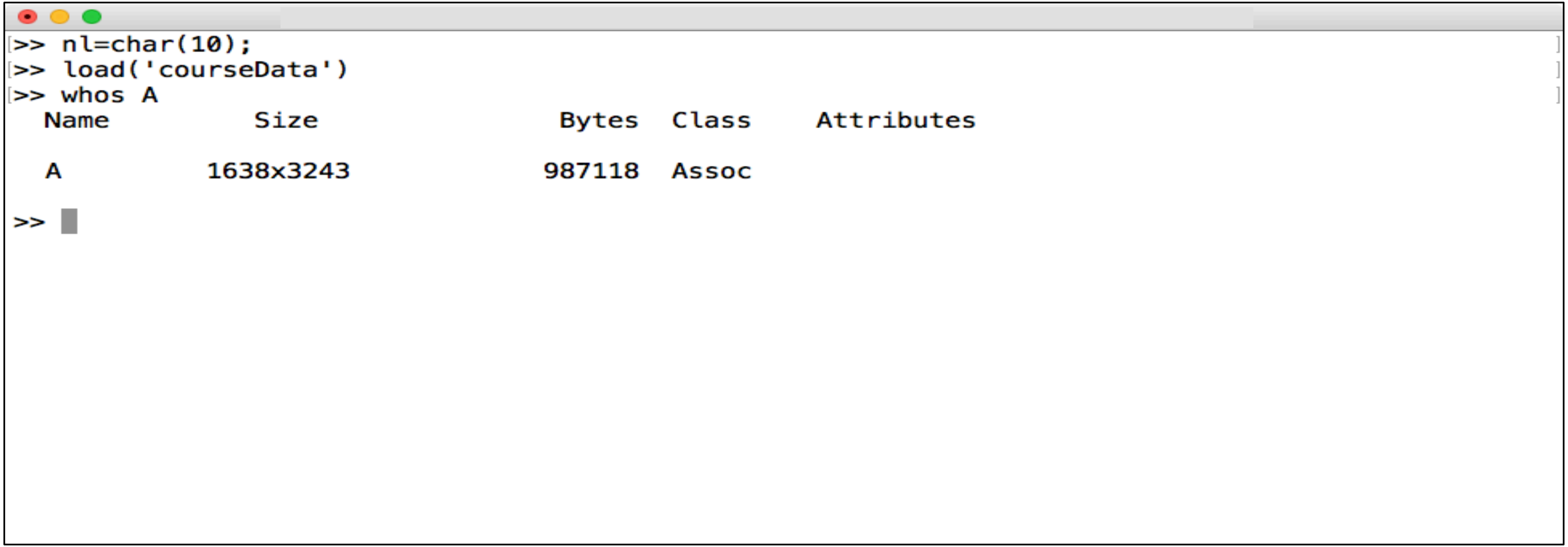
```
>> oldUnitCols=A(:,StartsWith('old_unit_name|,'));
>> newUnitCols=A(:,StartsWith('unit|,'));
>> unitChangeGraph=oldUnitCols.'*newUnitCols;
>> size(unitChangeGraph)
ans =
    72    73
>> nnz(unitChangeGraph)
ans =
   150
>> unitChangeGraph>7
(old_unit_name|Goals,unit|Course Overview)      9
(old_unit_name|Definitions & Fund.,unit|Review of Fund.)      8
(old_unit_name|Volume,unit|Velocity)      9
(old_unit_name|Challenge Review,unit|Volume)      8
>> spy(unitChangeGraph)
>> 
```

```
>> wrongAnswers=A(Row(A(:,'success|incorrect,')),:);
>> question=wrongAnswers(:,StartsWith('question|,'));
>> numWrongResponses=sum(question,1);
>> max(Val(numWrongResponses))
ans =
     5
>> numWrongResponses==5
(1,question|How many entries are in the correlation Associative Array?)      5
(1,question|How many pairs of users have more than one word in common?)      5
(1,question|Which of these is an example Big Data? (select all that apply))      5
>> 
```

Outline

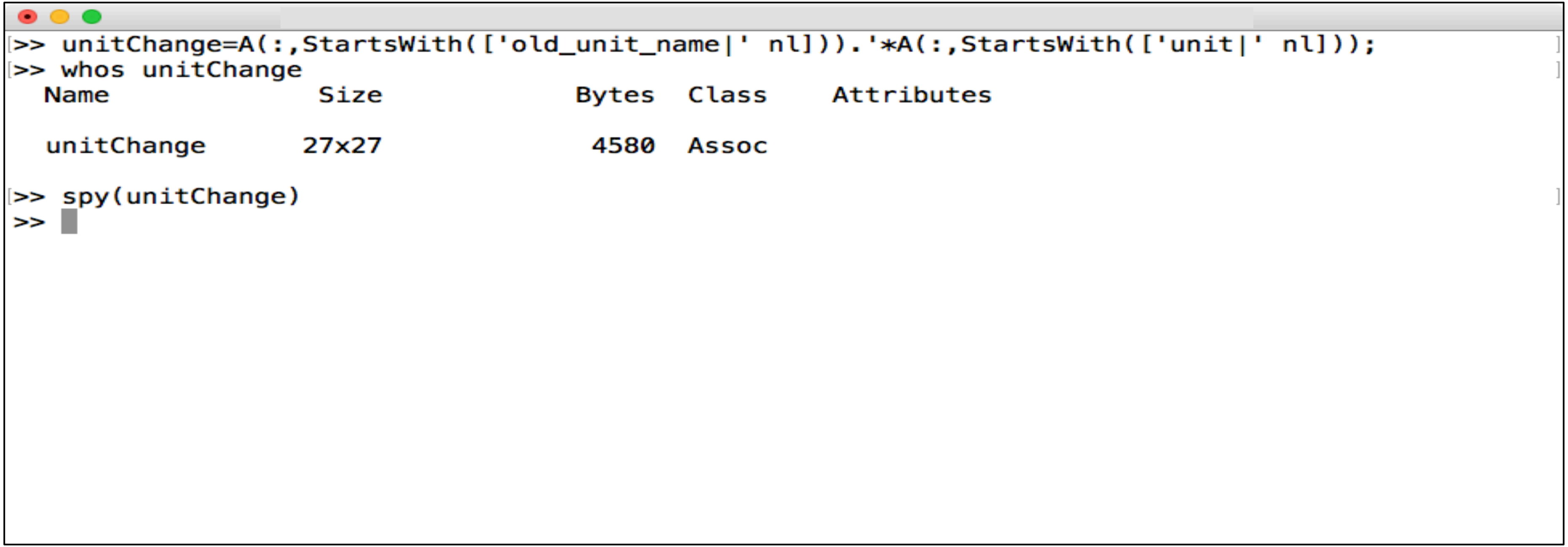
- Introduction
- D4M and Analytics Pipeline
- ➡ Demo
- Ground Truth Data
- Results
- Conclusion

Demo



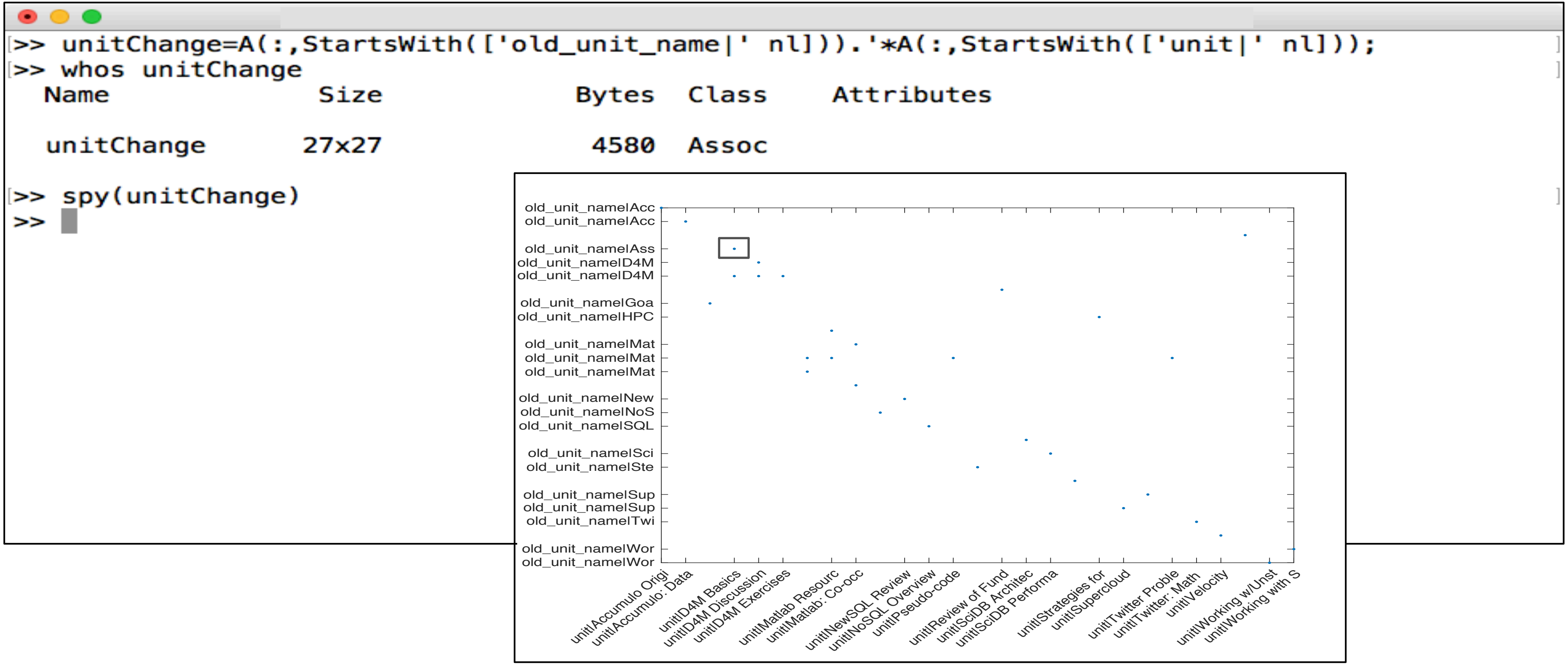
```
>> nl=char(10);
>> load('courseData')
>> whos A
  Name             Size                  Bytes  Class     Attributes

  A             1638x3243               987118  Assoc

>> 
```

Demo



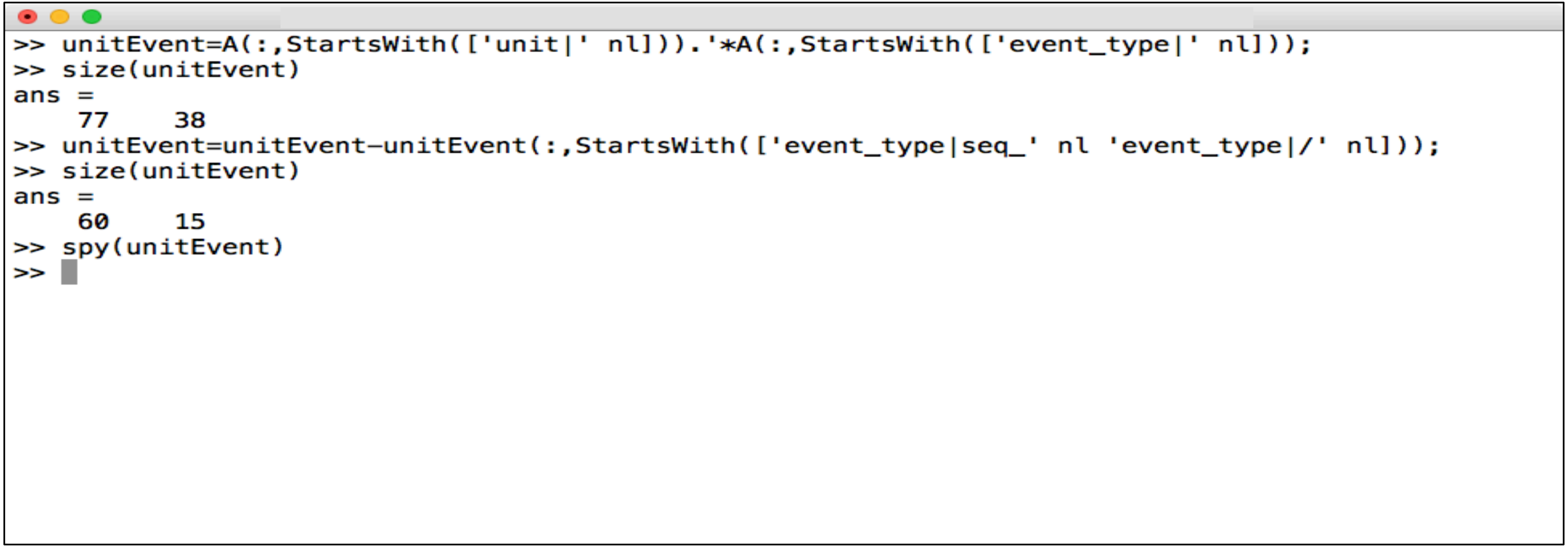
```
>> unitChange=A(:,StartsWith(['old_unit_name|' nl])).'*A(:,StartsWith(['unit|' nl]));
>> whos unitChange
  Name                 Size                Bytes  Class     Attributes

  unitChange          27x27                 4580  Assoc

>> spy(unitChange)
>> 
```

Demo



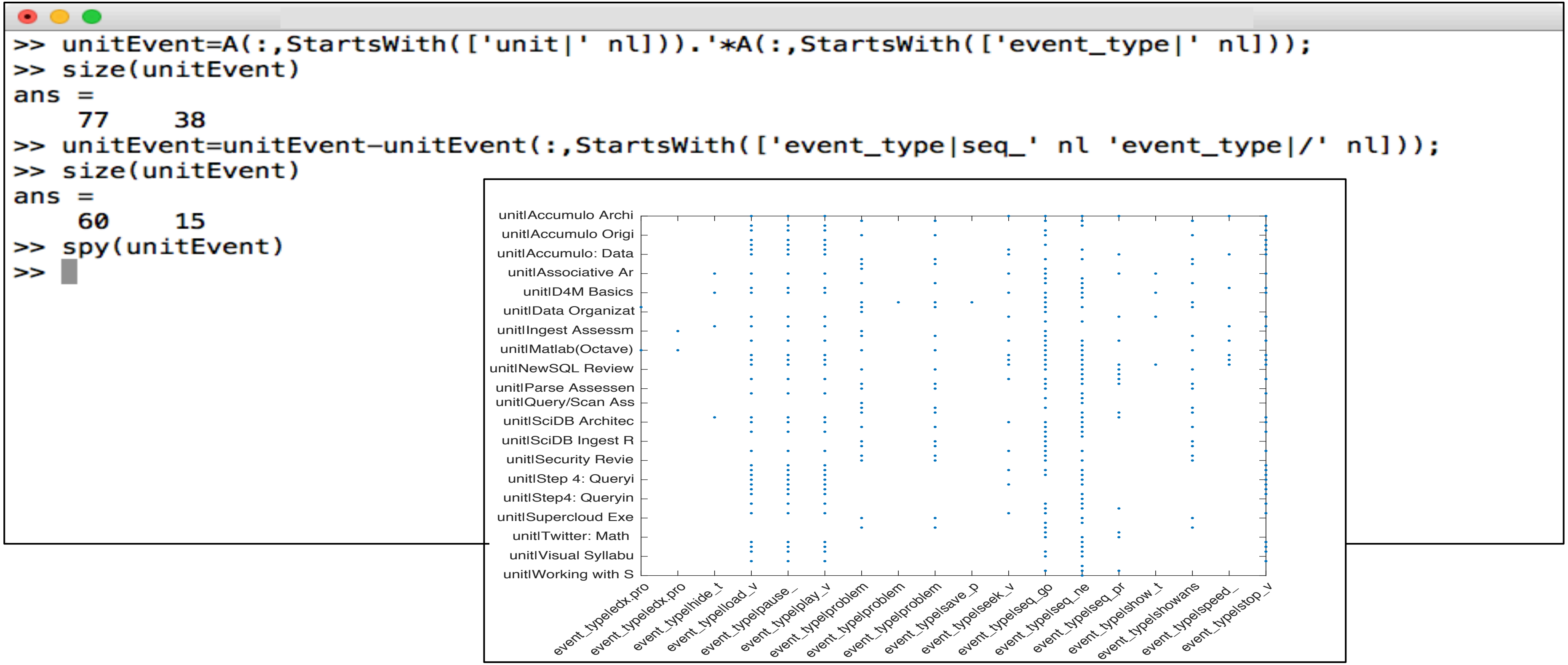
```
>> unitChange=A(:,StartsWith(['old_unit_name|' nl])).'*A(:,StartsWith(['unit|' nl]));
>> whos unitChange
  Name                 Size                 Bytes  Class    Attributes

  unitChange         27x27                   4580  Assoc

>> spy(unitChange)
>> 
```

Demo



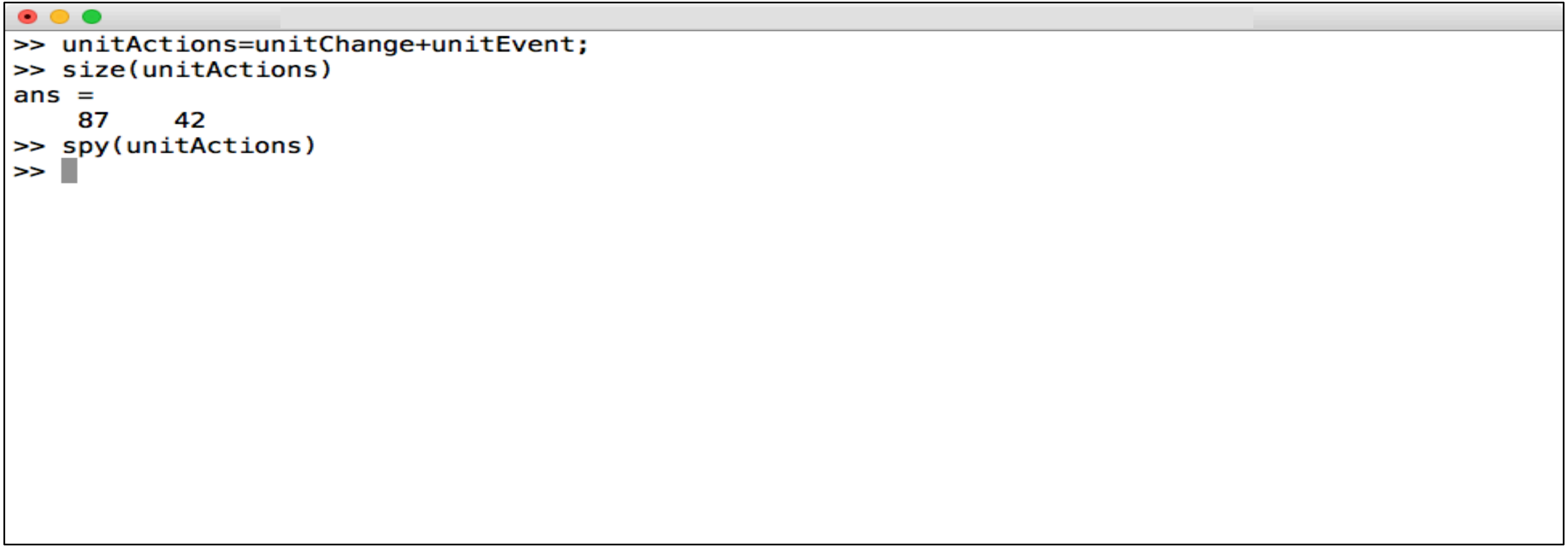
```
>> unitEvent=A(:,StartsWith(['unit|' nl])).'*A(:,StartsWith(['event_type|' nl]));
>> size(unitEvent)
ans =
    77    38
>> unitEvent=unitEvent-unitEvent(:,StartsWith(['event_type|seq_' nl 'event_type|/' nl]));
>> size(unitEvent)
ans =
    60    15
>> spy(unitEvent)
>> 
```

Demo



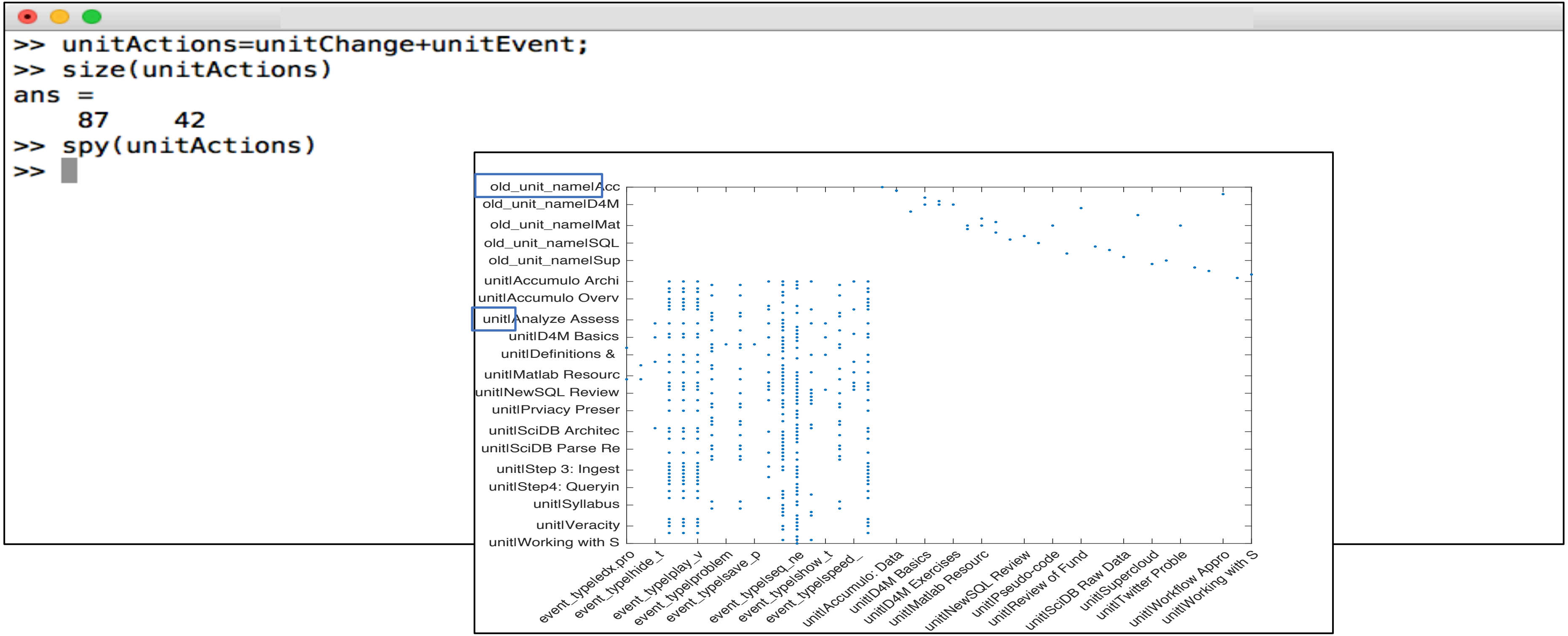
```
>> unitEvent=A(:,StartsWith(['unit|' nl])).'*A(:,StartsWith(['event_type|' nl]));
>> size(unitEvent)
ans =
    77    38
>> unitEvent=unitEvent-unitEvent(:,StartsWith(['event_type|seq_' nl 'event_type|/' nl]));
>> size(unitEvent)
ans =
    60    15
>> spy(unitEvent)
>>
```

Demo



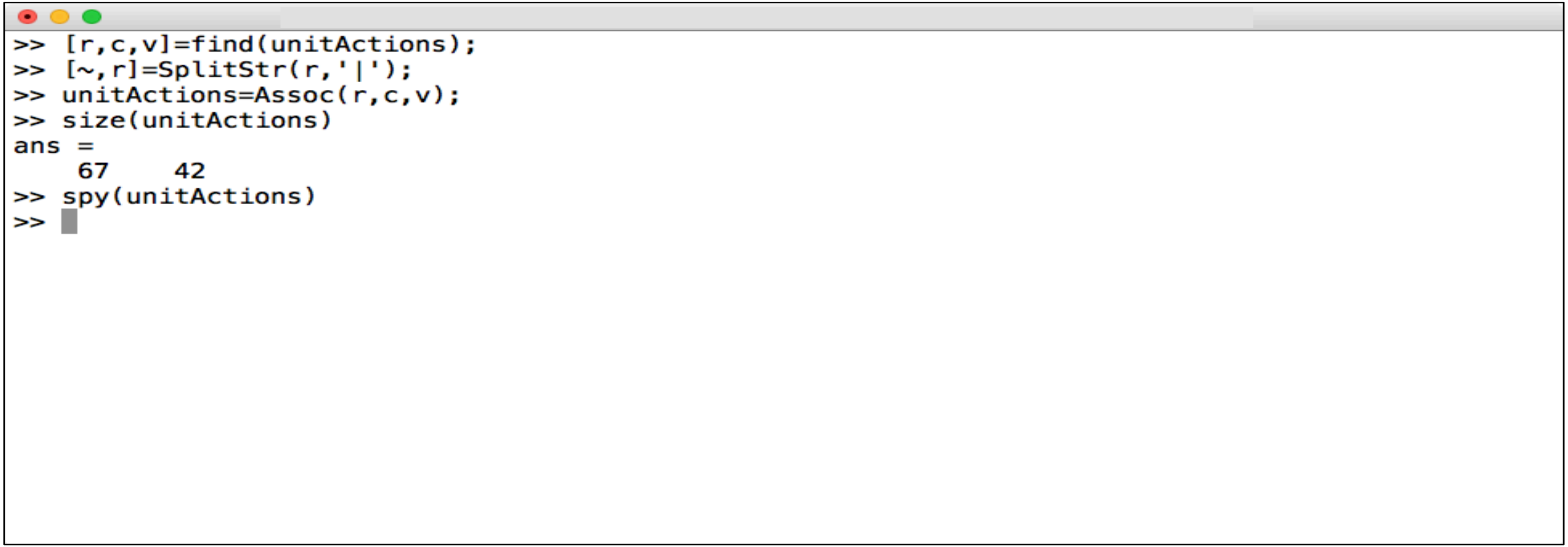
```
>> unitActions=unitChange+unitEvent;
>> size(unitActions)
ans =
    87    42
>> spy(unitActions)
>> 
```

Demo



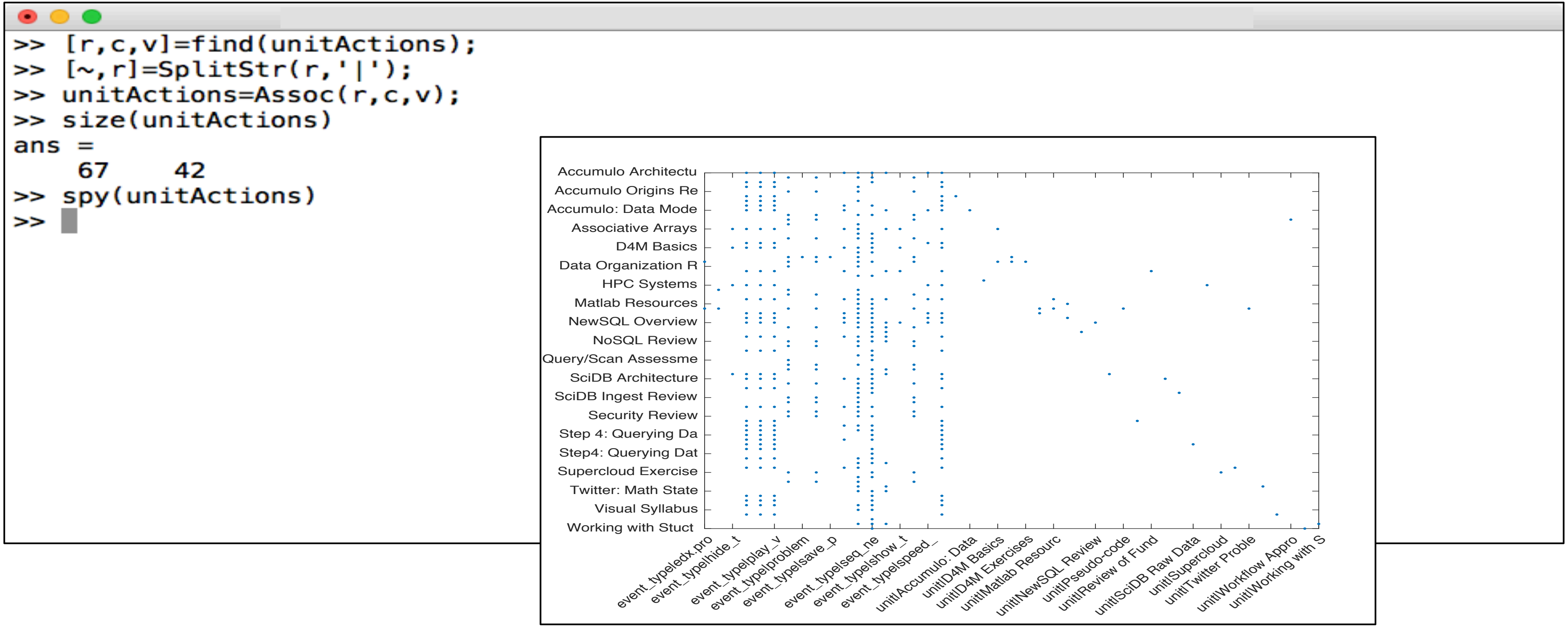
```
>> unitActions=unitChange+unitEvent;
>> size(unitActions)
ans =
    87    42
>> spy(unitActions)
>> 
```

Demo



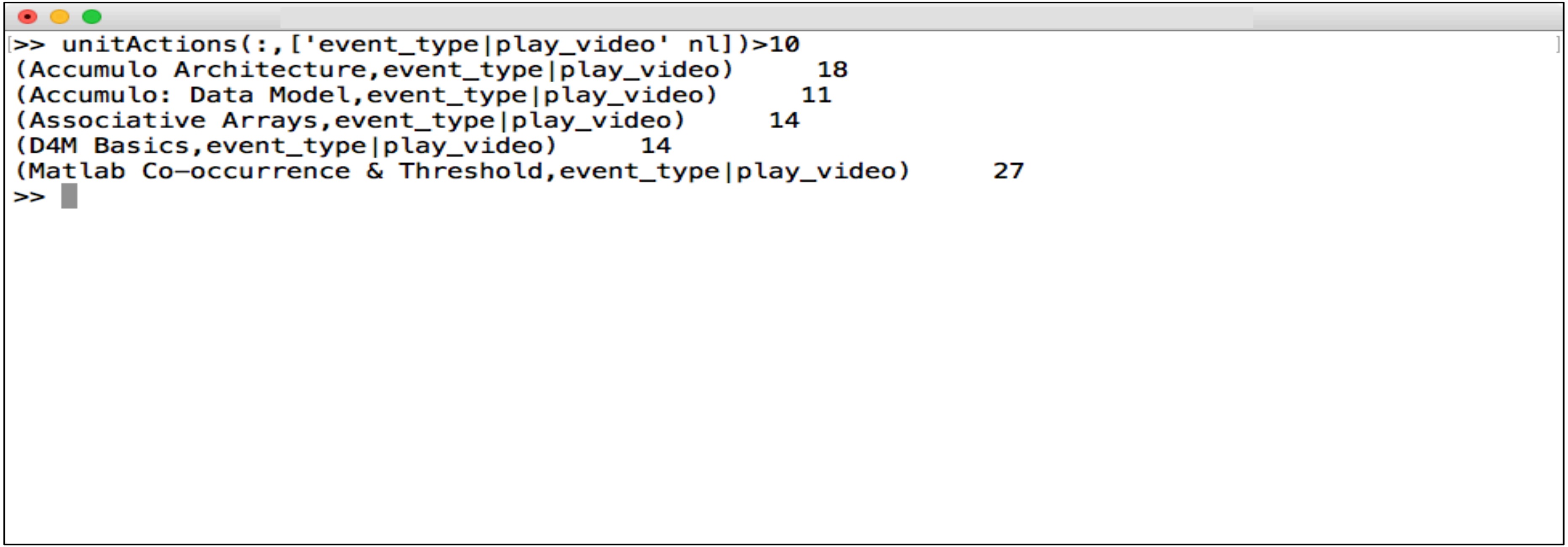
```
>> [r,c,v]=find(unitActions);
>> [~,r]=SplitStr(r,'|');
>> unitActions=Assoc(r,c,v);
>> size(unitActions)
ans =
    67    42
>> spy(unitActions)
>> 
```

Demo



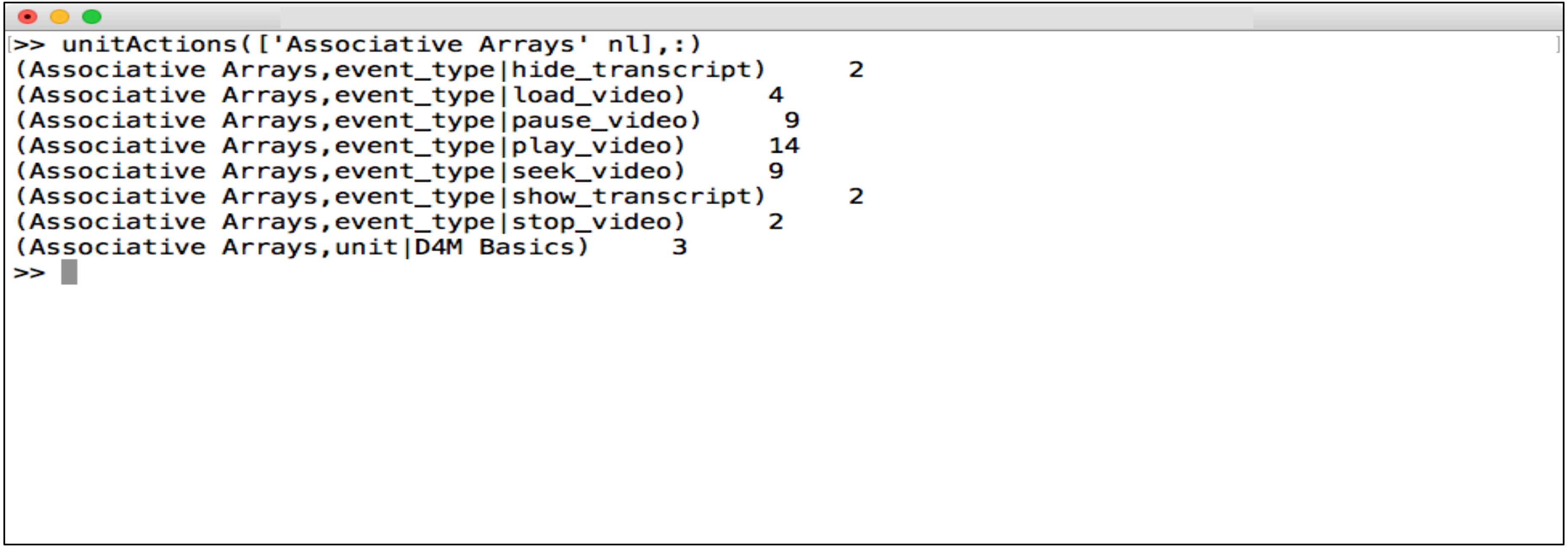
```
>> [r,c,v]=find(unitActions);
>> [~,r]=SplitStr(r,'|');
>> unitActions=Assoc(r,c,v);
>> size(unitActions)
ans =
     67     42
>> spy(unitActions)
>> 
```

Demo: Which units have the most video plays?



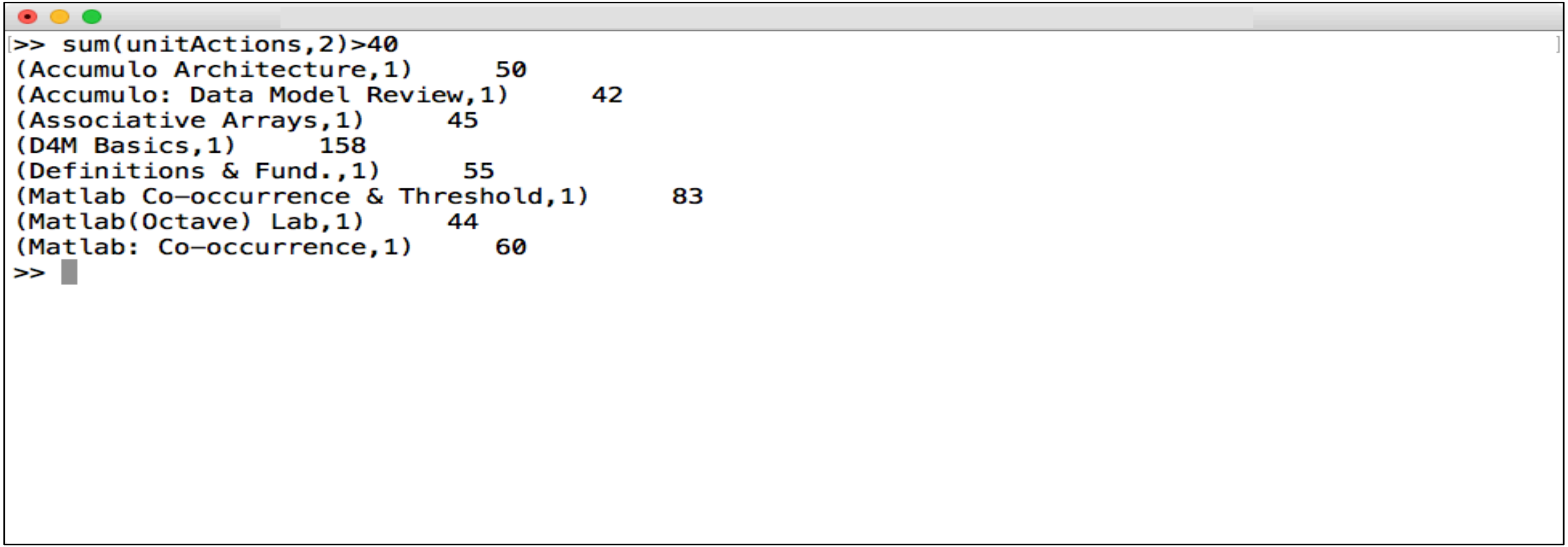
```
>> unitActions(:,['event_type|play_video' nl])>10
(Accumulo Architecture,event_type|play_video)      18
(Accumulo: Data Model,event_type|play_video)      11
(Associative Arrays,event_type|play_video)      14
(D4M Basics,event_type|play_video)      14
(Matlab Co-occurrence & Threshold,event_type|play_video)      27
>> 
```

Demo: What actions do students take from the “Associative Arrays” unit?



```
>> unitActions(['Associative Arrays' nl],:)
(Associative Arrays,event_type|hide_transcript)      2
(Associative Arrays,event_type|load_video)      4
(Associative Arrays,event_type|pause_video)      9
(Associative Arrays,event_type|play_video)      14
(Associative Arrays,event_type|seek_video)      9
(Associative Arrays,event_type|show_transcript)      2
(Associative Arrays,event_type|stop_video)      2
(Associative Arrays,unit|D4M Basics)      3
>>
```

Demo: Which units have the most activity?



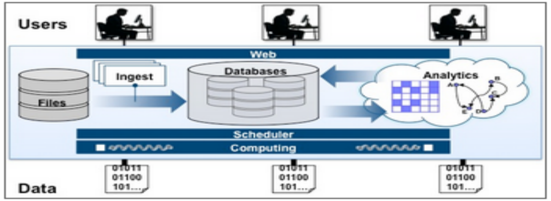
```
>> sum(unitActions,2)>40
(Accumulo Architecture,1)     50
(Accumulo: Data Model Review,1)     42
(Associative Arrays,1)     45
(D4M Basics,1)     158
(Definitions & Fund.,1)     55
(Matlab Co-occurrence & Threshold,1)     83
(Matlab(Octave) Lab,1)     44
(Matlab: Co-occurrence,1)     60
>> 
```

Outline

- Introduction
- D4M and Analytics Pipeline
- Demo
- ➡ Ground Truth Data
- Results
- Conclusion

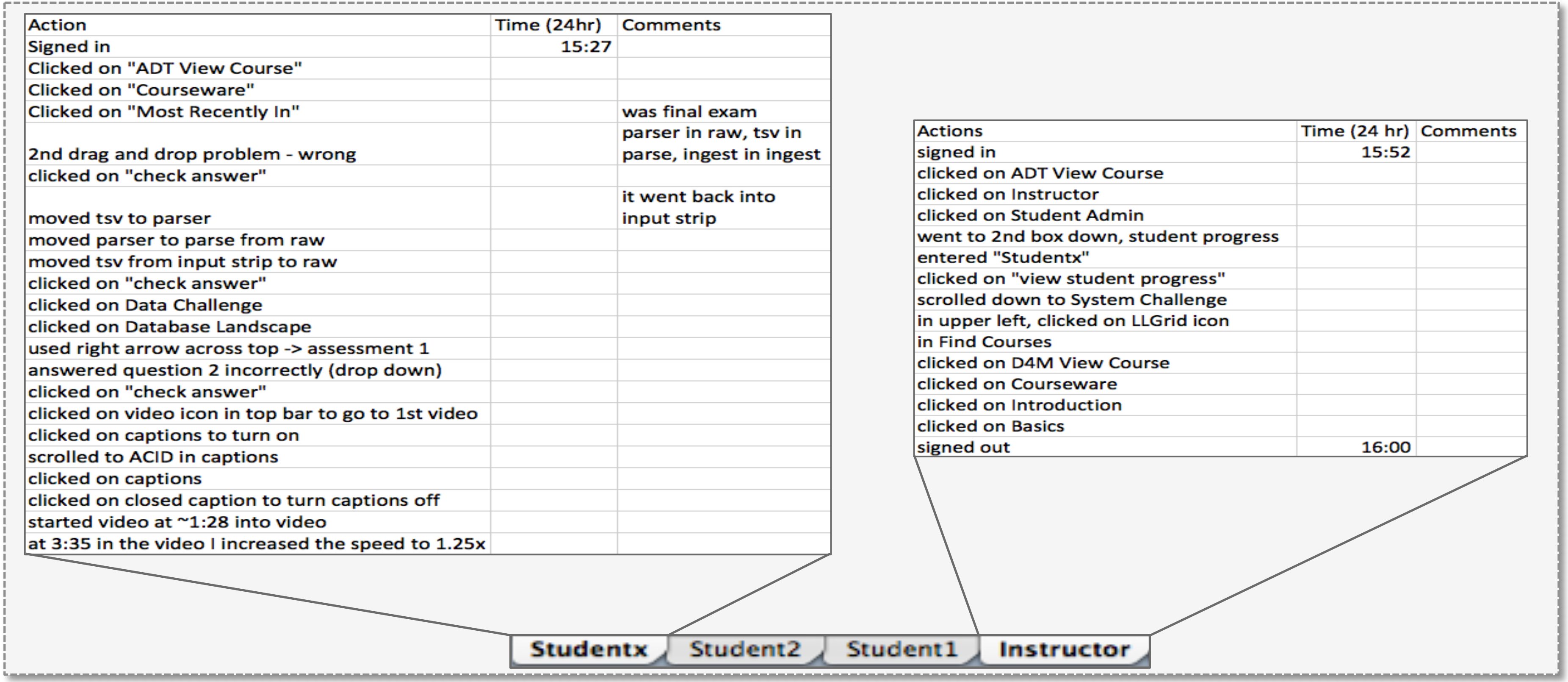
Ground Truth Data

- Can we accurately recreate a student’s actions?
  - What actions are captured by the tracking logs?
  - Which lines of the tracking logs contain useful information?
  - What should the parser add or remove to yield clear, informative events?
- Created using a script of actions for:
  - Three “students”
  - One “instructor”
- Actions executed, time and comments recorded



Advanced Database Technologies Course

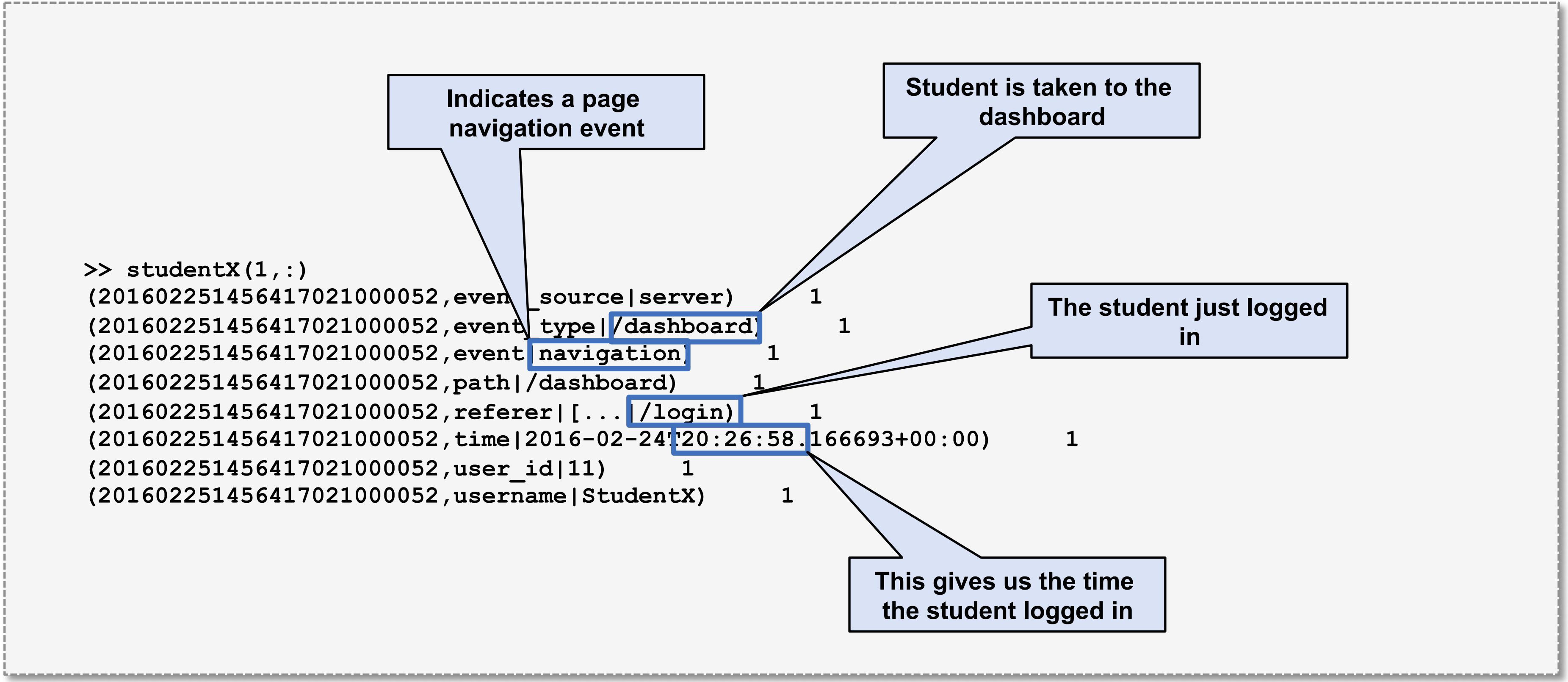
Ground Truth Data



Action	Time (24hr)	Comments
Signed in	15:27	
Clicked on "ADT View Course"		
Clicked on "Courseware"		
Clicked on "Most Recently In"		was final exam
2nd drag and drop problem - wrong		parser in raw, tsv in parse, ingest in ingest
clicked on "check answer"		
moved tsv to parser		it went back into input strip
moved parser to parse from raw		
moved tsv from input strip to raw		
clicked on "check answer"		
clicked on Data Challenge		
clicked on Database Landscape		
used right arrow across top -> assessment 1		
answered question 2 incorrectly (drop down)		
clicked on "check answer"		
clicked on video icon in top bar to go to 1st video		
clicked on captions to turn on		
scrolled to ACID in captions		
clicked on captions		
clicked on closed caption to turn captions off		
started video at ~1:28 into video		
at 3:35 in the video I increased the speed to 1.25x		

Actions	Time (24 hr)	Comments
signed in	15:52	
clicked on ADT View Course		
clicked on Instructor		
clicked on Student Admin		
went to 2nd box down, student progress		
entered "Studentx"		
clicked on "view student progress"		
scrolled down to System Challenge		
in upper left, clicked on LLGrid icon		
in Find Courses		
clicked on D4M View Course		
clicked on Courseware		
clicked on Introduction		
clicked on Basics		
signed out	16:00	

Interpreting Parsed Results



```
>> studentX(1,:)
(201602251456417021000052,event_source|server)       1
(201602251456417021000052,event_type|/dashboard)       1
(201602251456417021000052,event|navigation)       1
(201602251456417021000052,path|/dashboard)       1
(201602251456417021000052,referer|[...|/login)       1
(201602251456417021000052,time|2016-02-24T20:26:58.166693+00:00)      1
(201602251456417021000052,user_id|11)      1
(201602251456417021000052,username|StudentX)      1
```

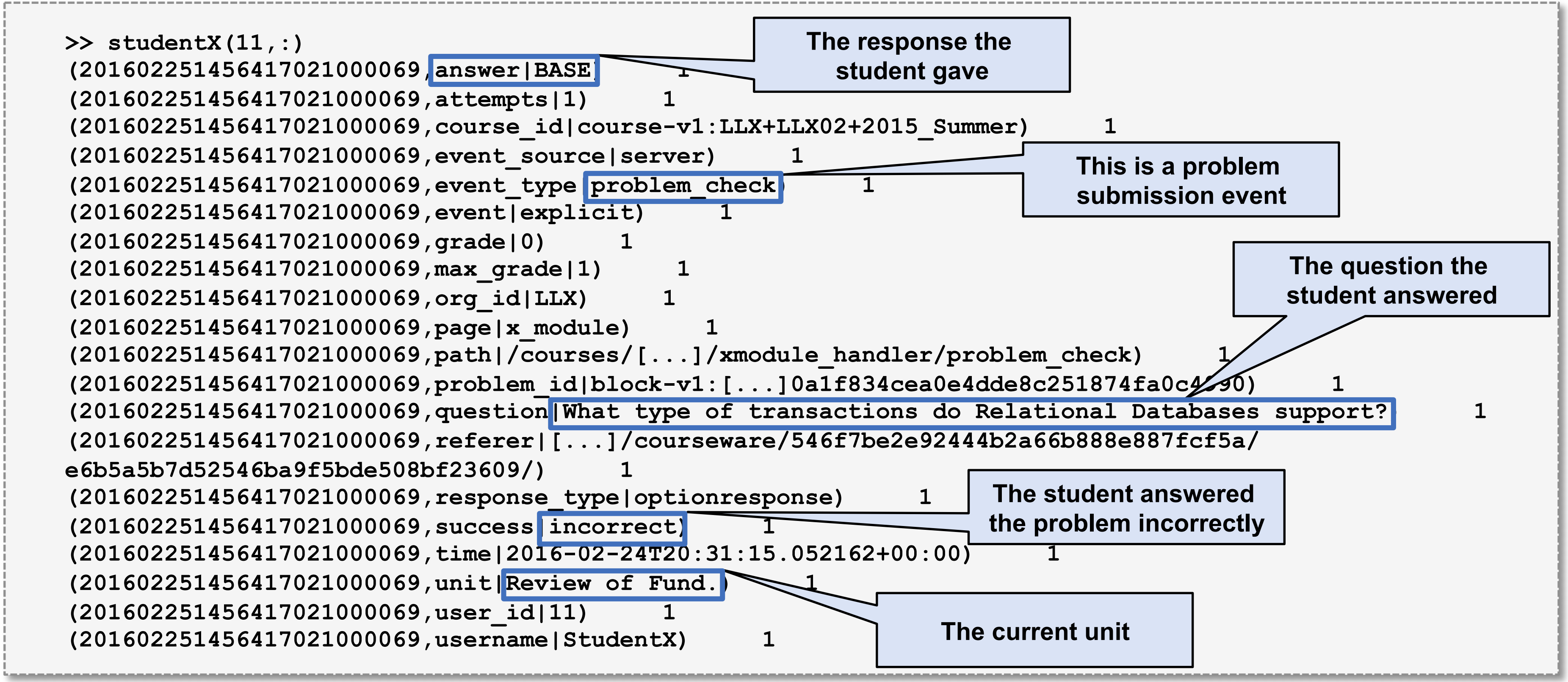
Indicates a page navigation event

Student is taken to the dashboard

The student just logged in

This gives us the time the student logged in

Interpreting Parsed Results



```
>> studentX(11,:)
(201602251456417021000069,answer|BASE)        1
(201602251456417021000069,attempts|1)      1
(201602251456417021000069,course_id|course-v1:LLX+LLX02+2015_Summer)      1
(201602251456417021000069,event_source|server)      1
(201602251456417021000069,event_type|problem_check)      1
(201602251456417021000069,event|explicit)      1
(201602251456417021000069,grade|0)      1
(201602251456417021000069,max_grade|1)      1
(201602251456417021000069,org_id|LLX)      1
(201602251456417021000069,page|x_module)      1
(201602251456417021000069,path|/courses/[...]/xmodule_handler/problem_check)      1
(201602251456417021000069,problem_id|block-v1:[...]0a1f834cea0e4dde8c251874fa0c4090)      1
(201602251456417021000069,question|What type of transactions do Relational Databases support?)      1
(201602251456417021000069,referer|[...]/courseware/546f7be2e92444b2a66b888e887fcf5a/
e6b5a5b7d52546ba9f5bde508bf23609/)      1
(201602251456417021000069,response_type|optionresponse)      1
(201602251456417021000069,success|incorrect)      1
(201602251456417021000069,time|2016-02-24T20:31:15.052162+00:00)      1
(201602251456417021000069,unit|Review of Fund.)      1
(201602251456417021000069,user_id|11)      1
(201602251456417021000069,username|StudentX)      1
```

The response the student gave

This is a problem submission event

The question the student answered

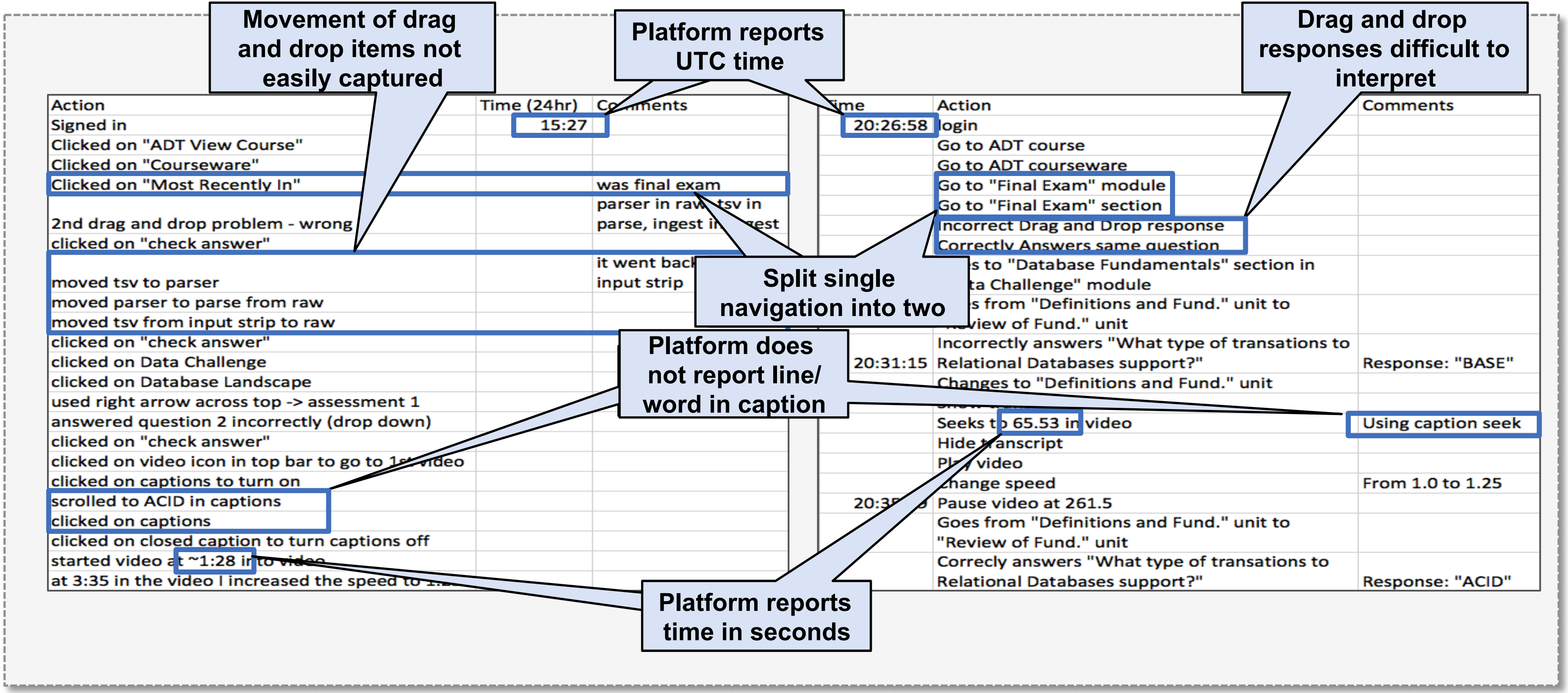
The student answered the problem incorrectly

The current unit

Outline

- Introduction
- D4M and Analytics Pipeline
- Ground Truth Data
- ➡ Results
- Conclusion

Parsed Results for StudentX



Movement of drag and drop items not easily captured

Platform reports UTC time

Drag and drop responses difficult to interpret

Split single navigation into two

Platform does not report line/ word in caption

Platform reports time in seconds

Action	Time (24hr)	Comments
Signed in	15:27	
Clicked on "ADT View Course"		
Clicked on "Courseware"		
Clicked on "Most Recently In"		was final exam
2nd drag and drop problem - wrong		parser in raw tsv in parse, ingest in ingest
clicked on "check answer"		
moved tsv to parser		it went back [illegible] input strip
moved parser to parse from raw		
moved tsv from input strip to raw		
clicked on "check answer"		
clicked on Data Challenge		
clicked on Database Landscape		
used right arrow across top -> assessment 1		
answered question 2 incorrectly (drop down)		
clicked on "check answer"		
clicked on video icon in top bar to go to 1st video		
clicked on captions to turn on		
scrolled to ACID in captions		
clicked on captions		
clicked on closed caption to turn captions off		
started video at ~1:28 into video		
at 3:35 in the video I increased the speed to 1.2		

Time	Action	Comments
20:26:58	login	
	Go to ADT course	
	Go to ADT courseware	
	Go to "Final Exam" module	
	Go to "Final Exam" section	
	Incorrect Drag and Drop response	
	Correctly Answers same question	
	Goes to "Database Fundamentals" section in "Data Challenge" module	
	Goes from "Definitions and Fund." unit to "Review of Fund." unit	
20:31:15	Incorrectly answers "What type of transations to Relational Databases support?"	Response: "BASE"
	Changes to "Definitions and Fund." unit	
	Show [illegible]	
	Seeks to 65.53 in video	Using caption seek
	Hide transcript	
	Play video	
	Change speed	From 1.0 to 1.25
20:3[illegible]	Pause video at 261.5	
	Goes from "Definitions and Fund." unit to "Review of Fund." unit	
	Correcly answers "What type of transations to Relational Databases support?"	Response: "ACID"

Results

- General path of students captured
  - General events (answering questions, playing videos)
  - Page navigation
  - Module/Section/Unit changes
- Some events are either not reported or not easily interpreted
  - Downloading files
  - Clicking on “send email” links
  - Actions associated with drag and drop problems
  - Answers for drag and drop problems
    - Exist but are hard to interpret
    - May report enough information to determine common incorrect answers
- Overall: can capture events tracked by the platform

Outline

- Introduction
- D4M and Analytics Pipeline
- Demo
- Ground Truth Data
- Results
- ➡ Conclusion

Conclusions and Future Work

- **Built up scalable tools for prototyping analytics for Open edX tracking log data**
- **Easy for instructors and researchers to**
  - **Query for data of interest**
  - **Form their own analytics**
- **Recreated student and instructor actions from known ground truth data**
- **Next steps:**
  - **Build up more learning analytics**
    - **Focus on student's paths through the material**
  - **Build models to enable adaptive learning**
    - **Assess and recommend additional actions to be captured**
  - **Ingest to Accumulo database using Accumulo's visibility labels**

Acknowledgements

- Vijay Gadepally
- Michael Houle
- Michael Jones
- Chansup Byun
- Anna Klein
- Matthew Hubbell
- Andrew Prout
- Siddharth Samsi
- Peter Michaleas
- William Arcand
- William Bergeron
- David Bestor
- Antonio Rosa
- Charles Yee

D4M:
http://d4m.mit.edu