

Data Visualization for Educational Research

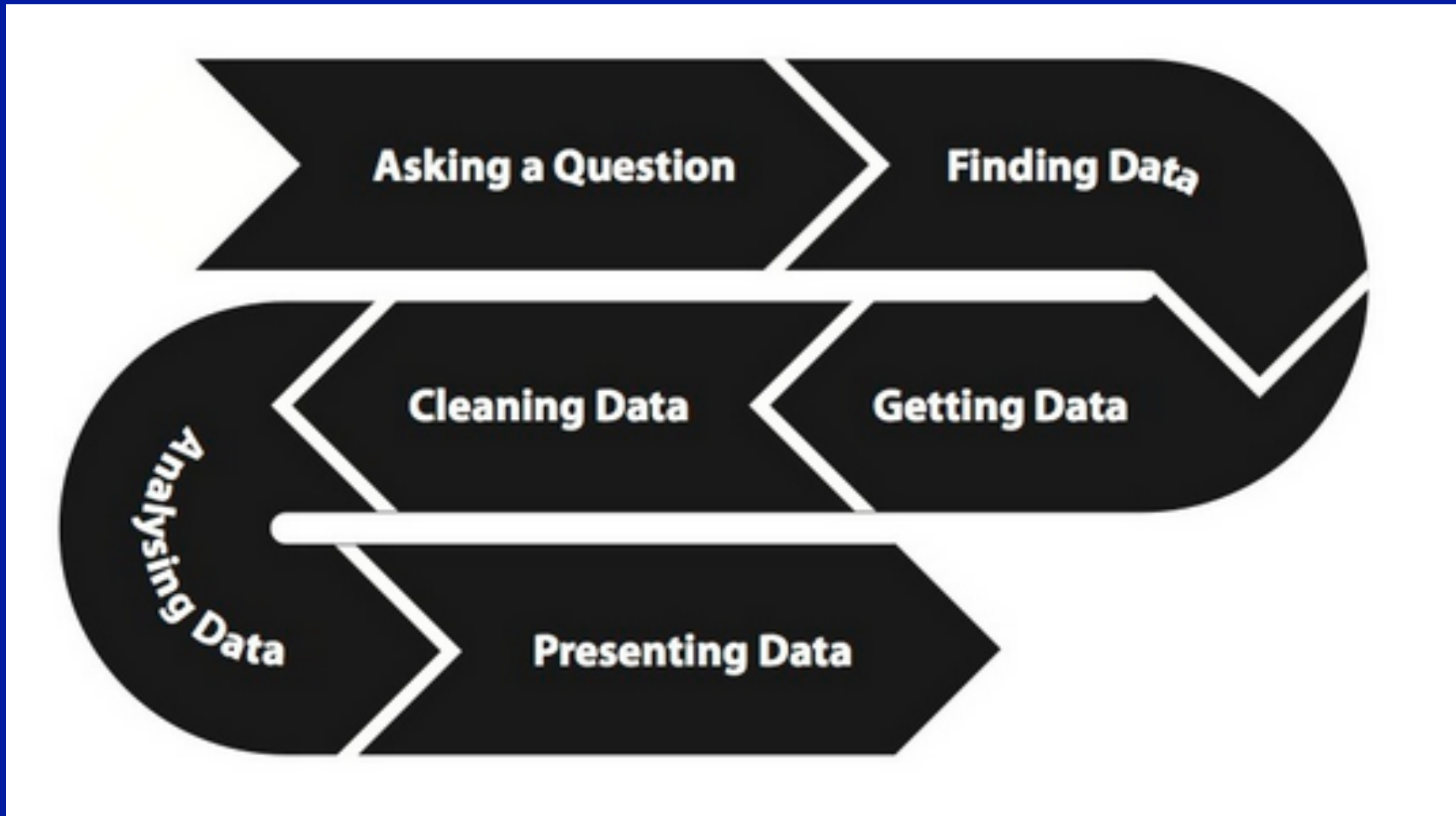


Taylor Martin
Active Learning Lab
Utah State University

The Opportunity

- I had a bunch here, but you fabulous folks have already covered it!
- The new microscope
- Rich and growing streams of digital learning data
- Better measures of learning and teaching

Data Viz across the pipeline



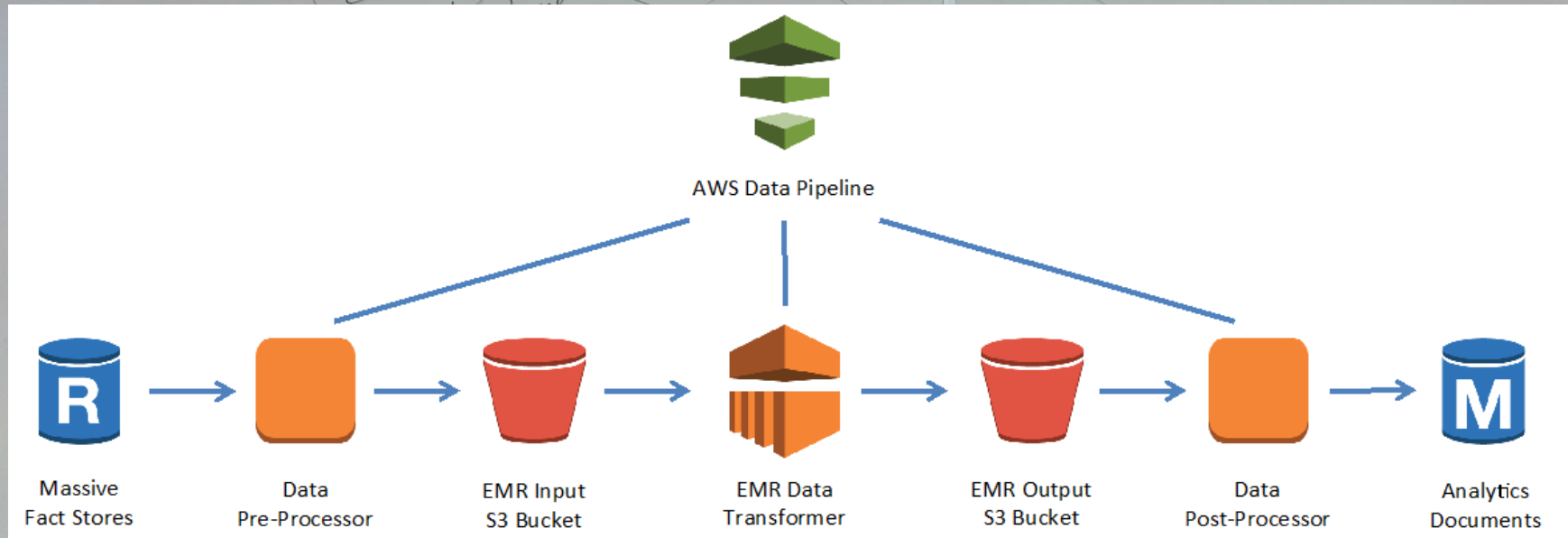
The Data Pipeline?



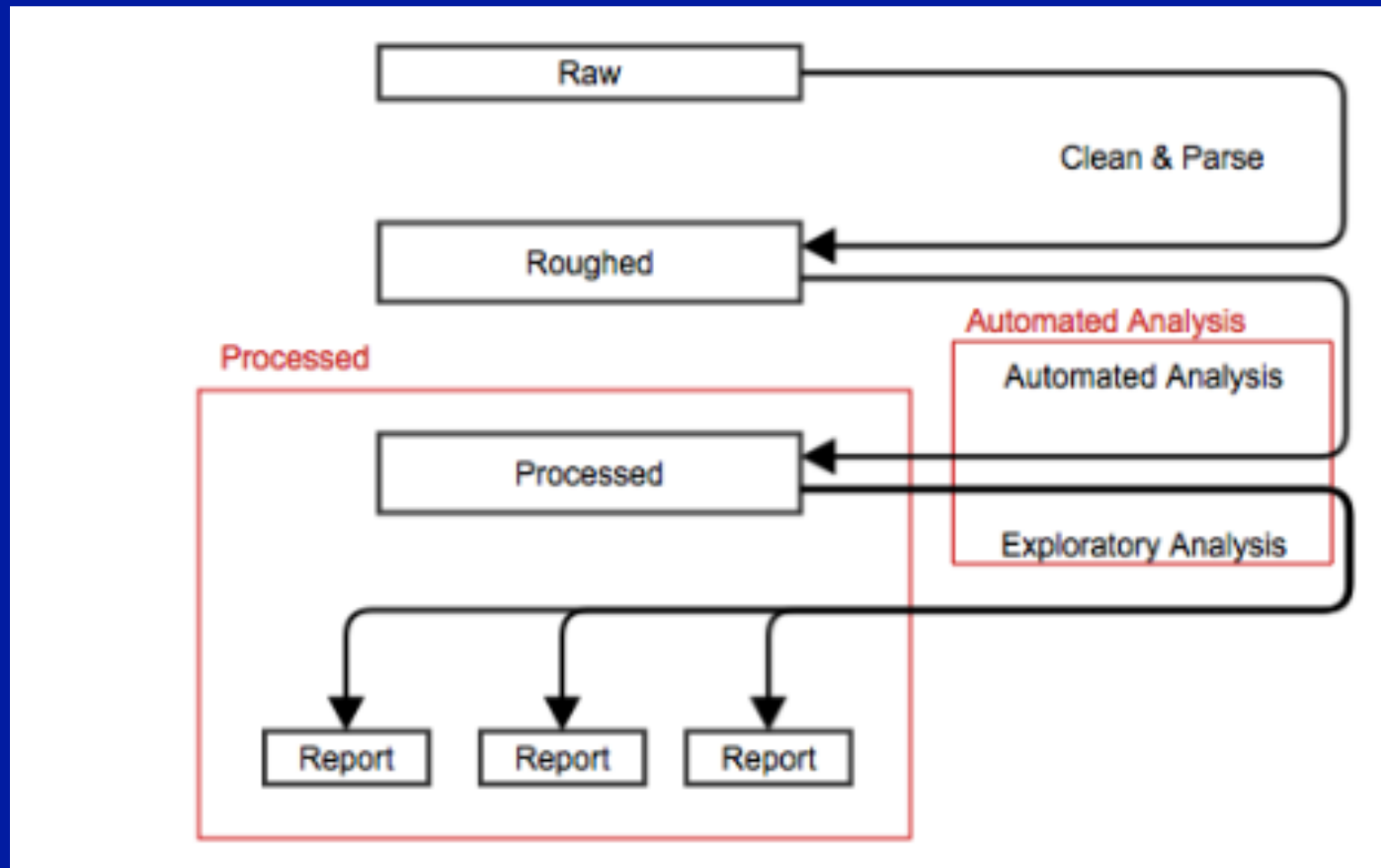
Collection

everything
xml
json

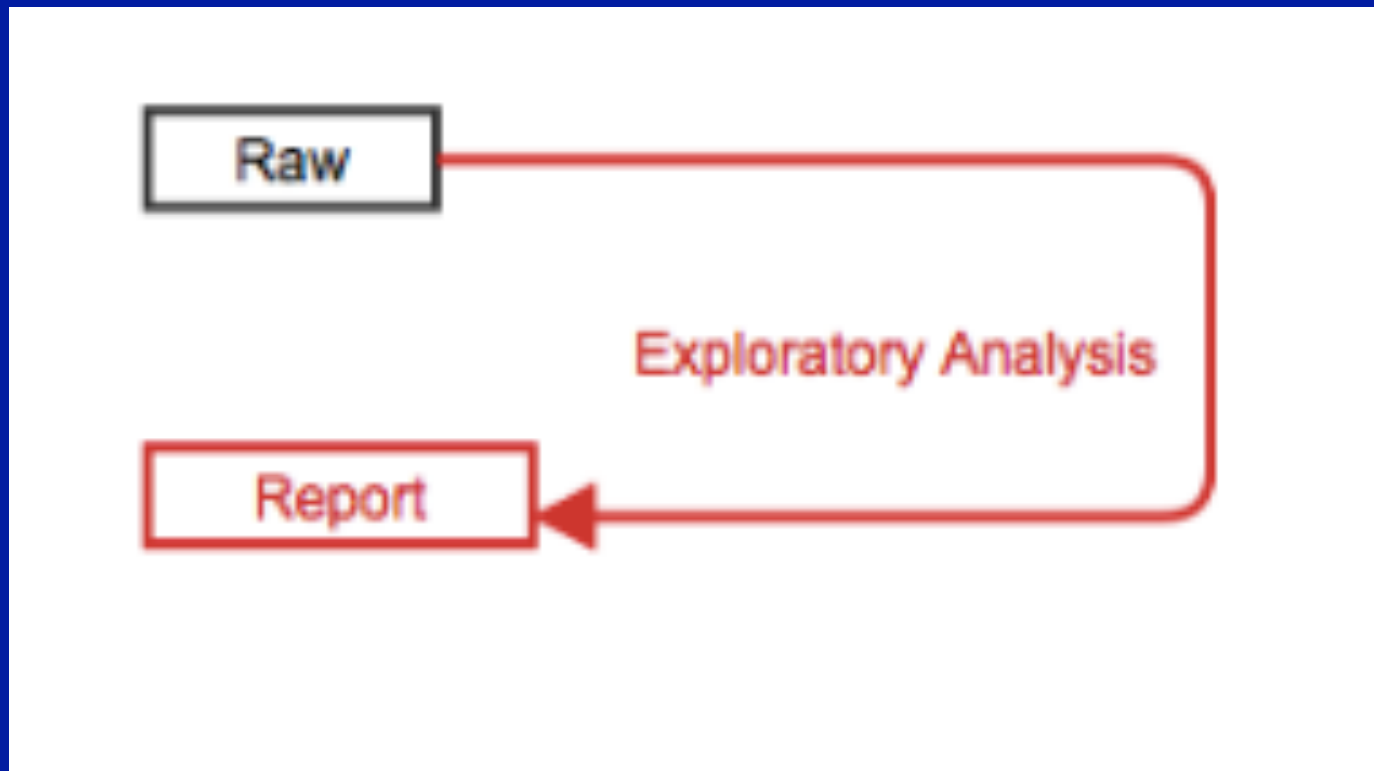
rawish data



For our group...



So at Stage 1...



Cleaning

Cluster & Edit column "Type of Contract"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method: Keying Function: 51 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
6	18	<ul style="list-style-type: none">Firm Fixed Price IDIQ (7 rows)FIRM FIXED PRICE IDIQ (5 rows)Firm Fixed Price (IDIQ) (2 rows)IDIQ Firm Fixed Price (2 rows)Firm Fixed Price - IDIQ (1 rows)Firm Fixed Price ID/ID (1 rows)	<input type="checkbox"/>	Firm Fixed Price IDIQ
6	868	<ul style="list-style-type: none">Firm Fixed Price (836 rows)firm fixed price (22 rows)FIRM FIXED PRICE (5 rows)Firm fixed price (2 rows)Firm Fixed price (2 rows)Firm Fixed Price (1 rows)	<input type="checkbox"/>	Firm Fixed Price
5	7	<ul style="list-style-type: none">FFP & T/M (3 rows)FFP T&M (1 rows)FFP & T&M (1 rows)T&M FFP (1 rows)T&M & FFP (1 rows)	<input type="checkbox"/>	FFP & T/M
5	9	<ul style="list-style-type: none">Fixed price labor hour (5 rows)Fixed Price - Labor Hour (1 rows)Fixed Price - Labor hour (1 rows)Fixed Price / Labor Hour (1 rows)	<input type="checkbox"/>	Fixed price labor hour

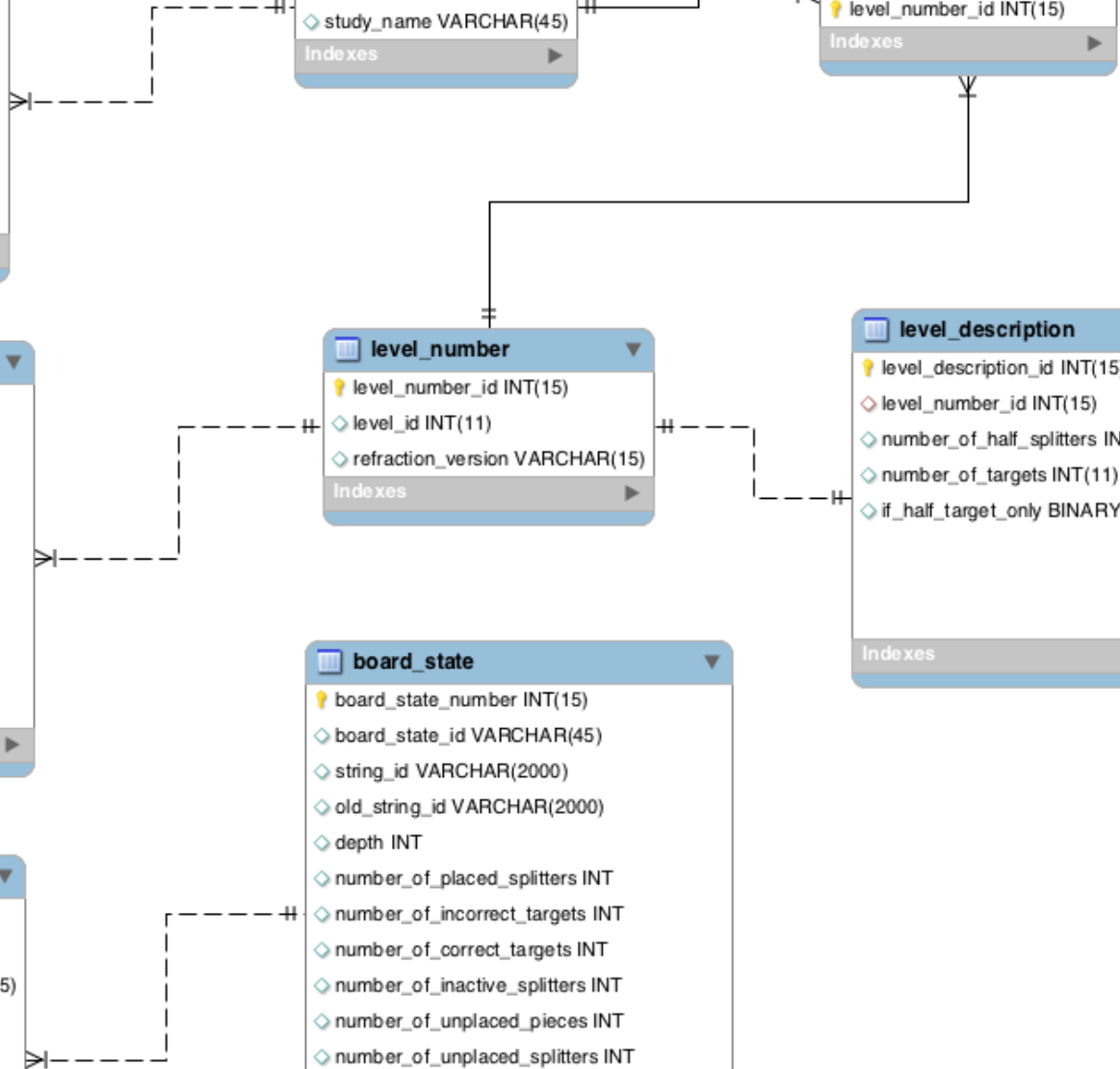
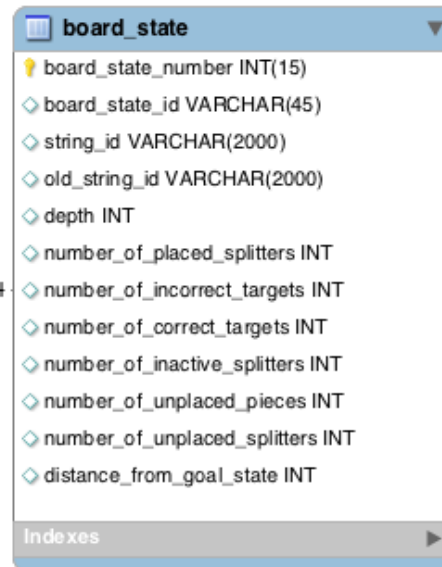
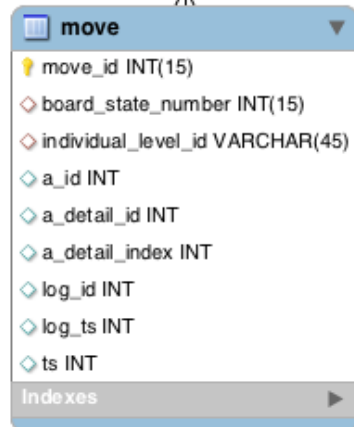
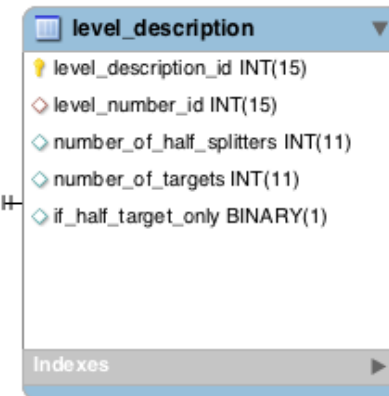
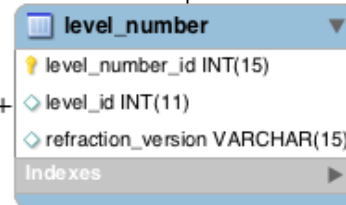
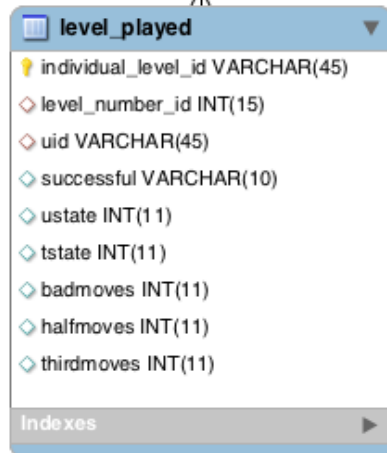
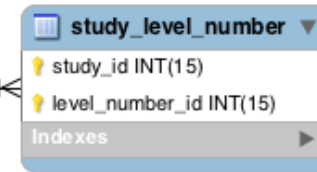
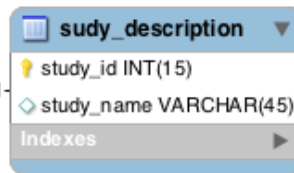
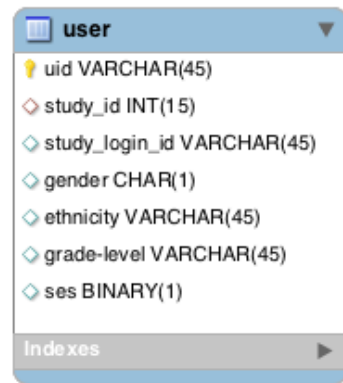
Choices in Cluster

Rows in Cluster

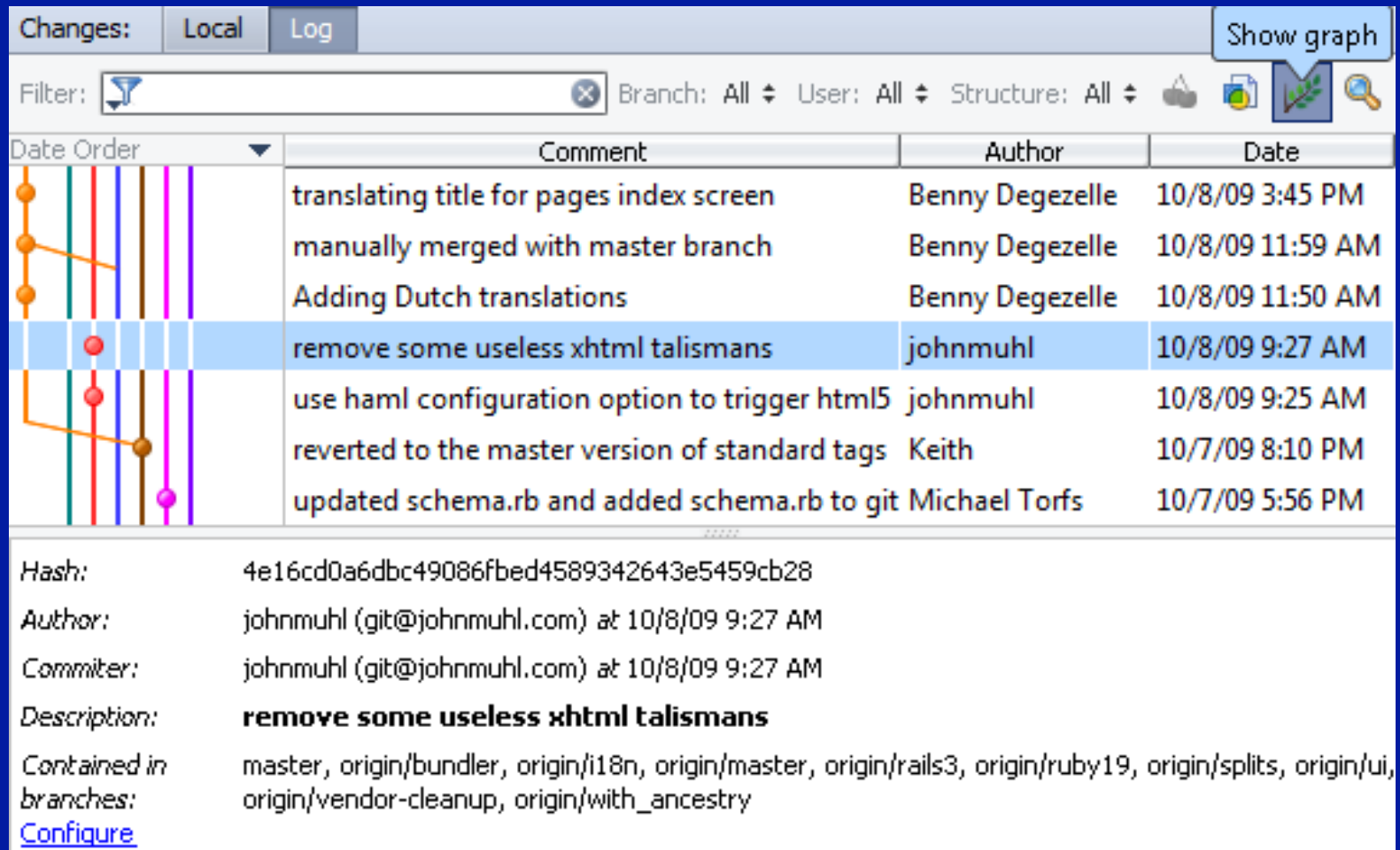
Average Length of Choices

Length Variance of Choices

Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close



Knowing what you did later



The screenshot shows a Git log window with the following details:

- Buttons: Changes, Local, Log, Show graph
- Filter: Branch: All User: All Structure: All
- Table columns: Date Order, Comment, Author, Date
- Selected commit: remove some useless xhtml talismans (johnmuhl, 10/8/09 9:27 AM)
- Other visible commits: translating title for pages index screen, manually merged with master branch, Adding Dutch translations, use haml configuration option to trigger html5, reverted to the master version of standard tags, updated schema.rb and added schema.rb to git
- Commit details for the selected commit:
 - Hash: 4e16cd0a6dbc49086fbed4589342643e5459cb28
 - Author: johnmuhl (git@johnmuhl.com) at 10/8/09 9:27 AM
 - Committer: johnmuhl (git@johnmuhl.com) at 10/8/09 9:27 AM
 - Description: **remove some useless xhtml talismans**
 - Contained in branches: master, origin/bundler, origin/i18n, origin/master, origin/rails3, origin/ruby19, origin/splits, origin/ui, origin/vendor-cleanup, origin/with_ancestry
 - Configure

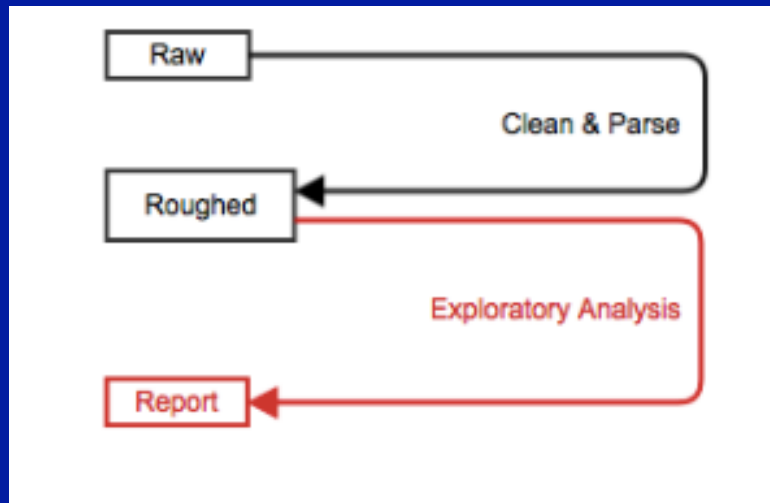
Challenges

- Capacity
 - Particularly for real Big Data
 - Quickly changing teams
- Disparate Data sources and shapes
 - xml combined with json to make sense of game data
 - accessing data through a variety of protocols
 - web services through SOAP or REST
 - Hadoop data through Hive
 - Other types of NOSQL data through proprietary APIs.
 - Tabular or relational still there but changing
- Keeping the pipeline as your guiding framework

But

- New and developing tools to help at this point, e.g.:
 - OpenRefine
 - Trifacta
 - DataTamer
 - Hunk (for Big Data)
- Capacity building efforts within the field
 - At this level, this is just starting. Probably most behind here.

Stage 2

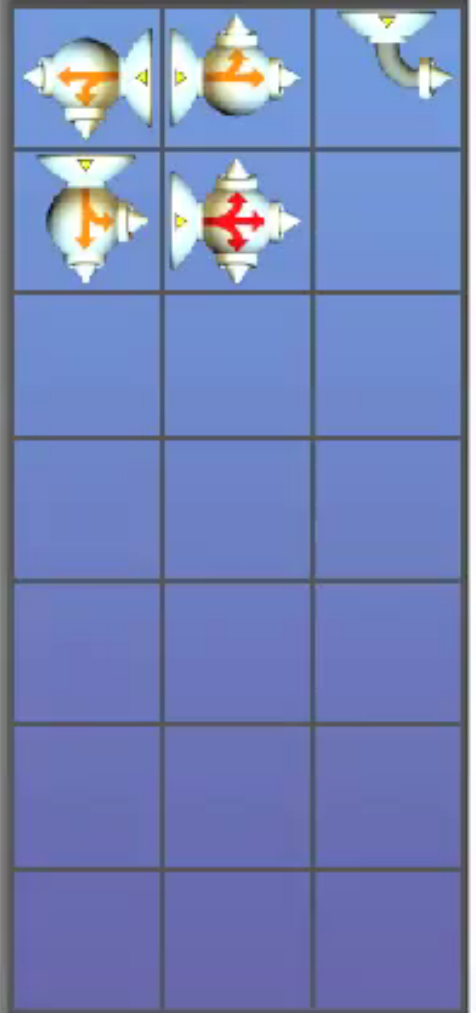


Example tasks:

Find/define/refine
variables of interest
(predicted or emergent)

Visualizations of
preliminary results






Level 2:1
The Big E



MENU

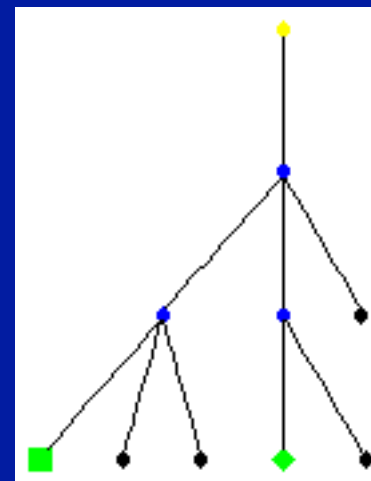
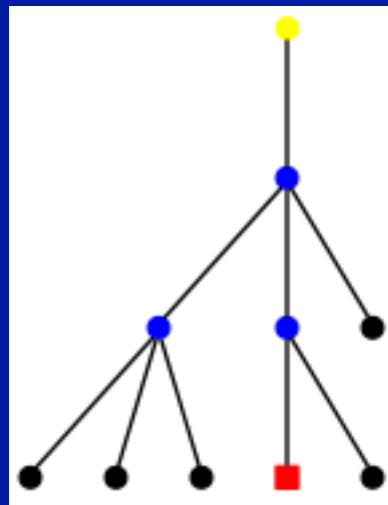
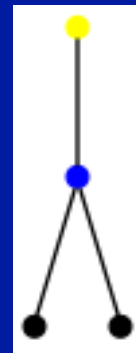
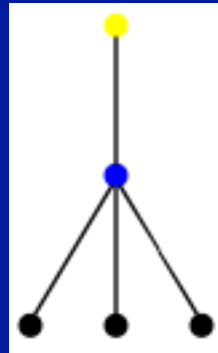
OPTIONS

Level 7:5

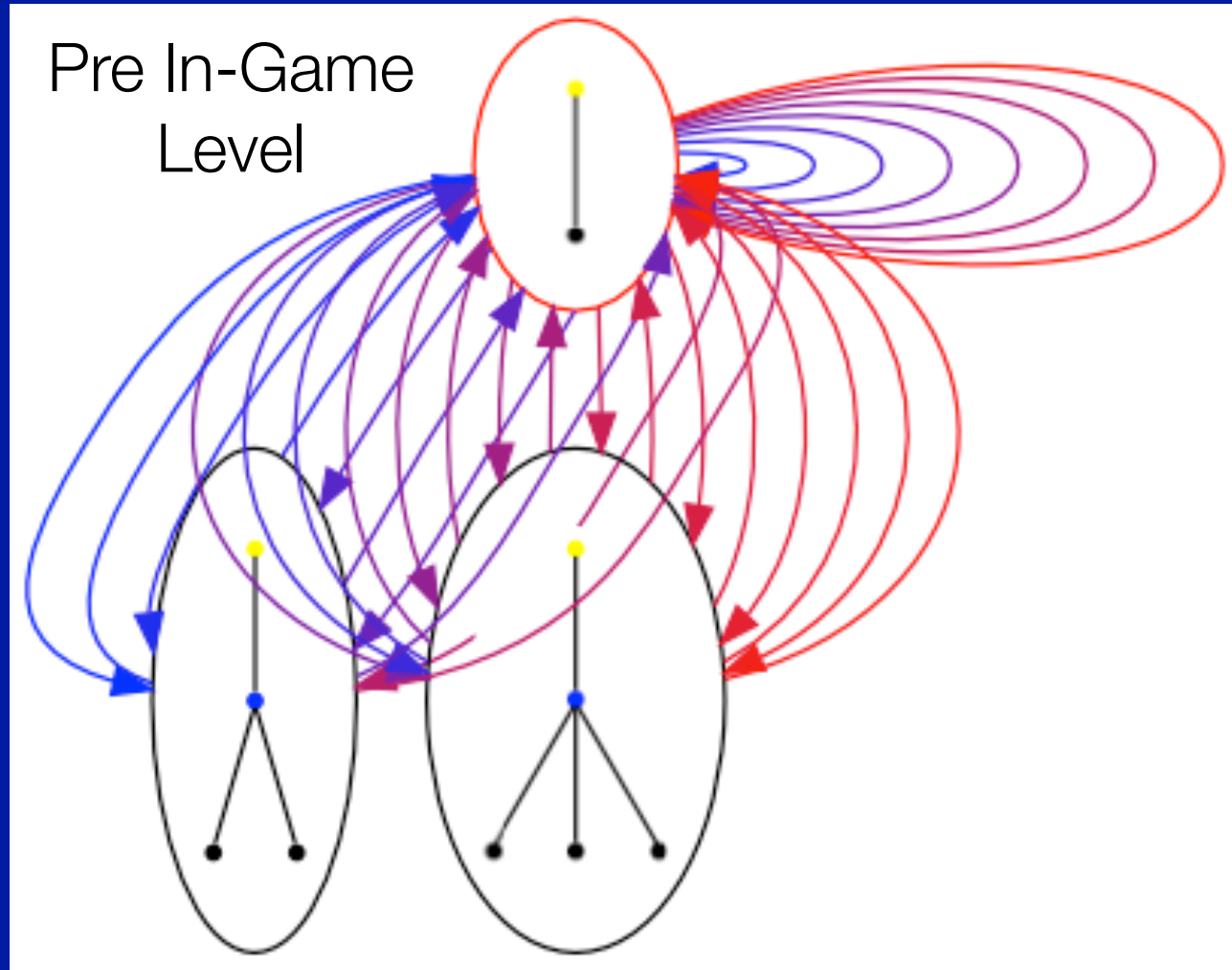
		
		



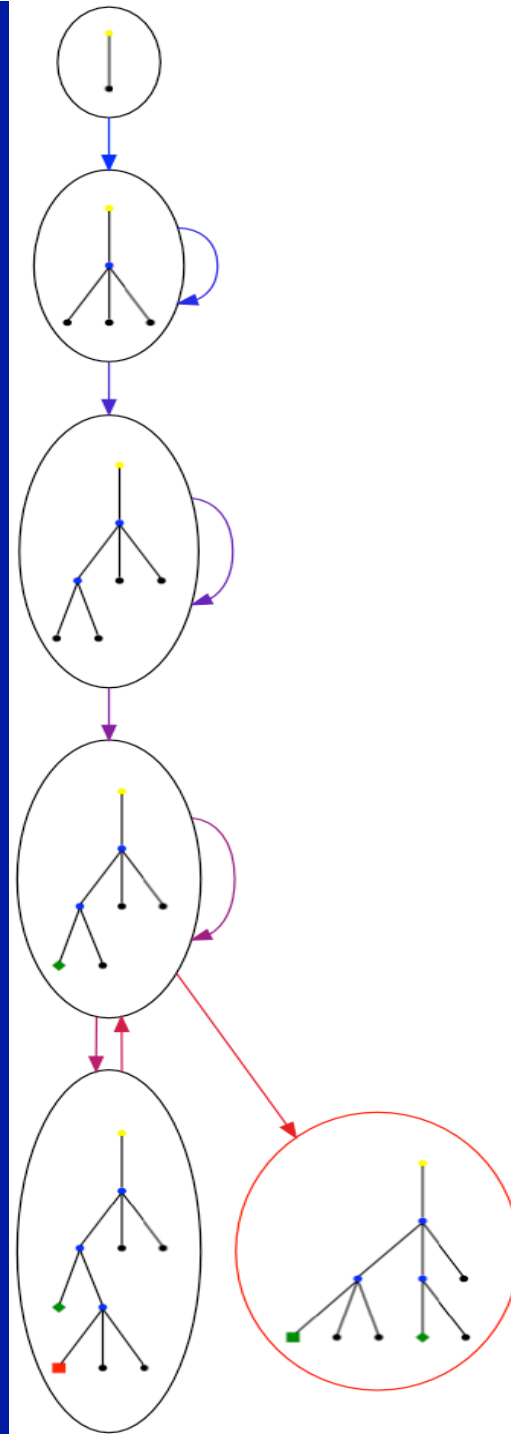
Mathematical States



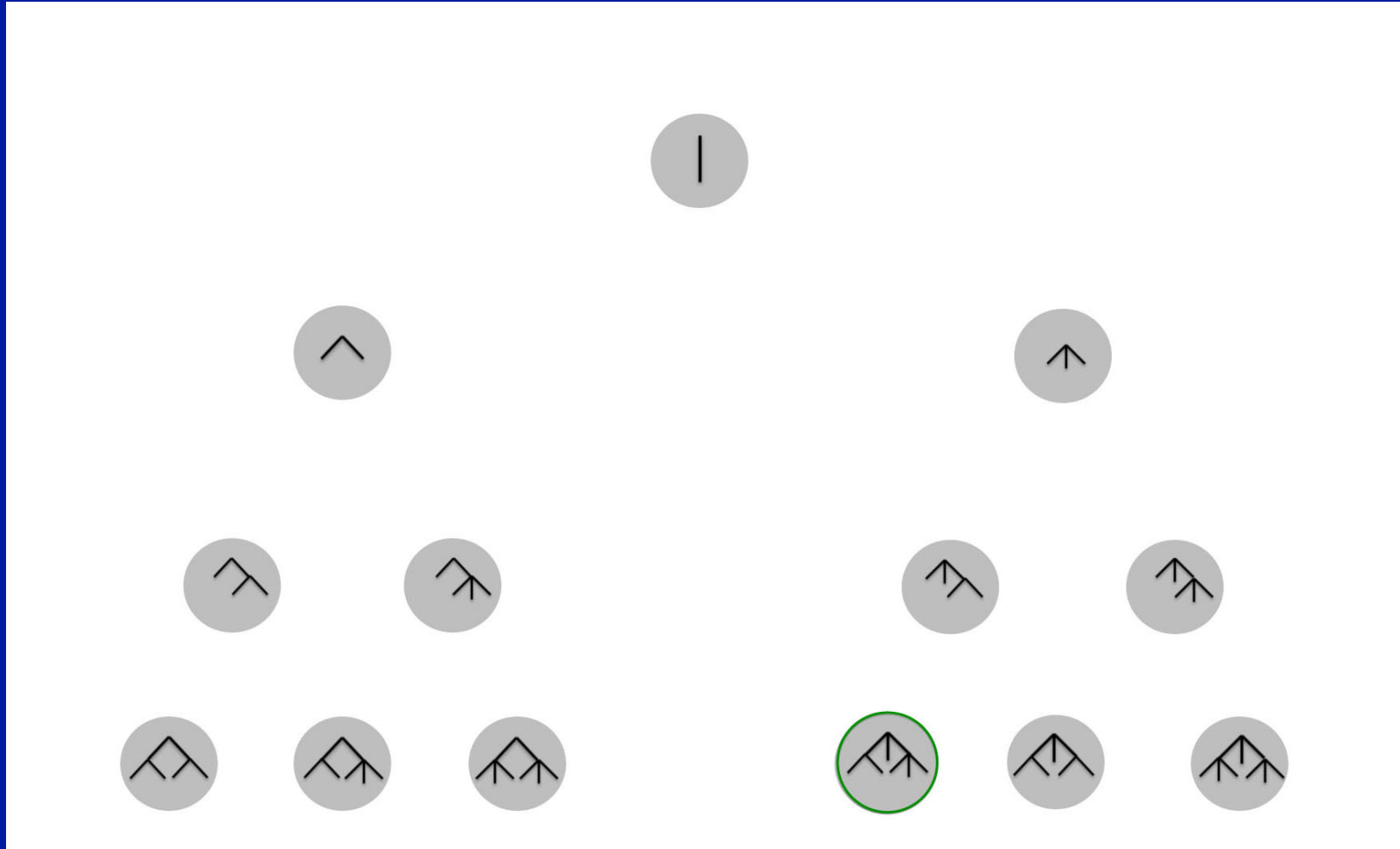
Visualizing Trajectories

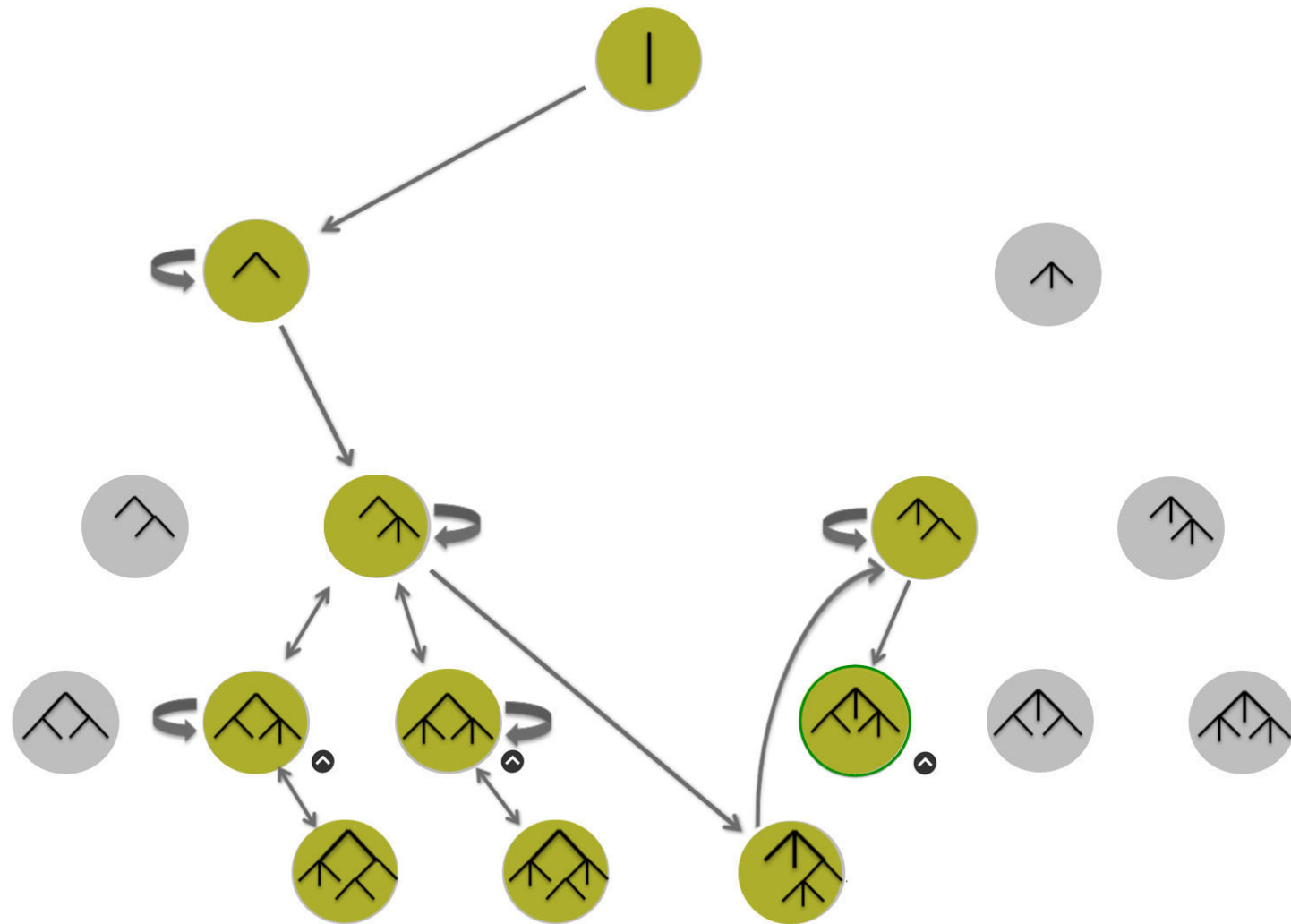


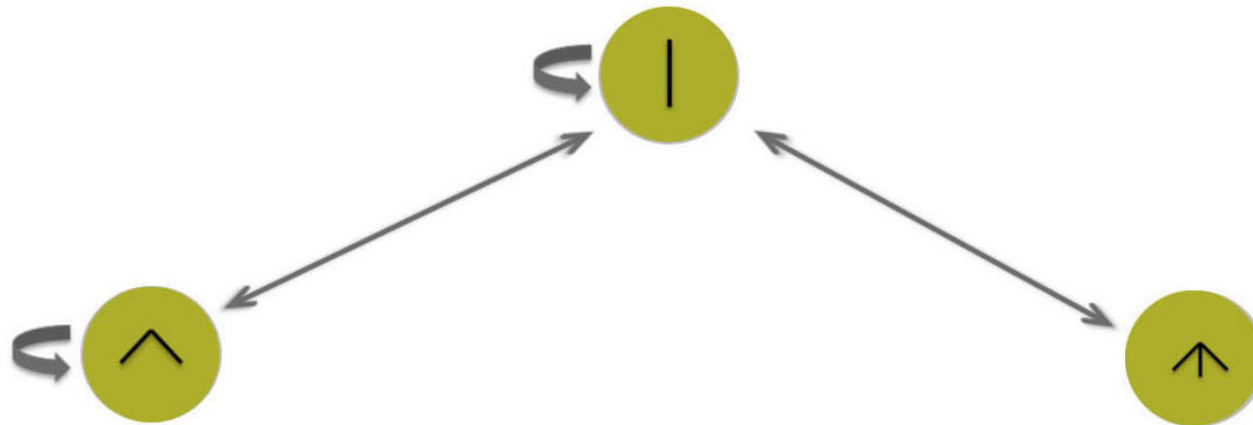
Post In-Game Level



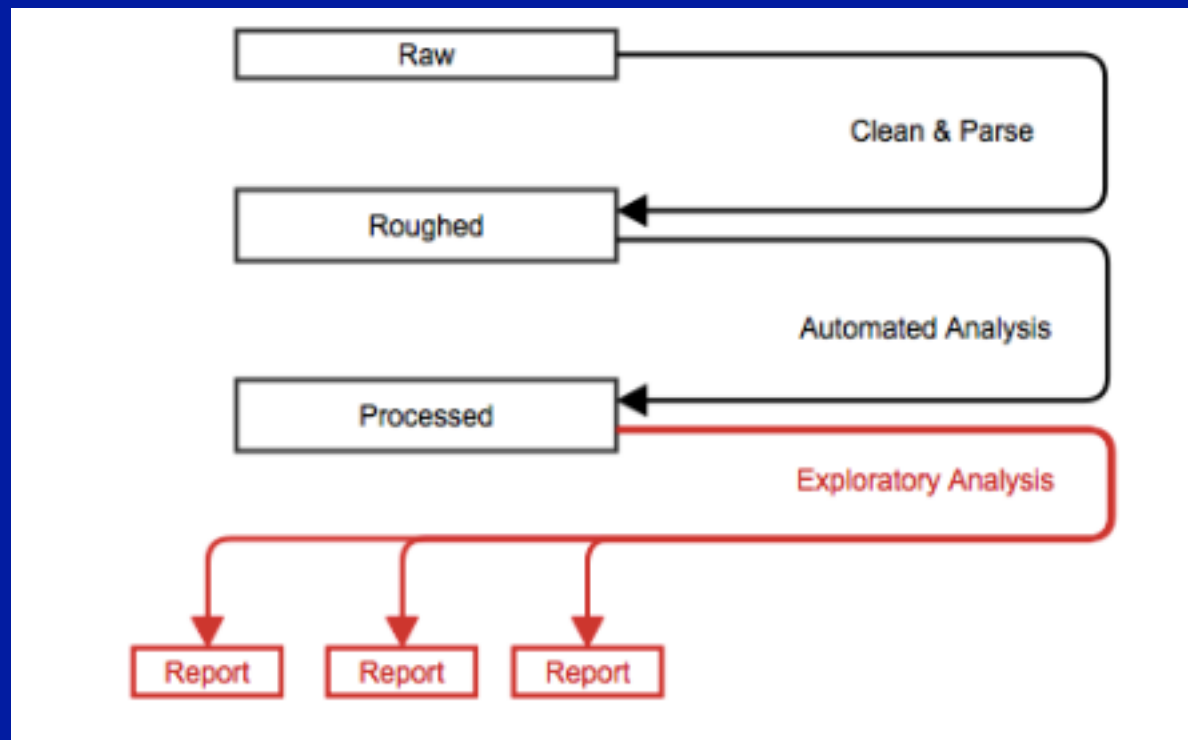
Iterate on Visualizations







Stage 3a



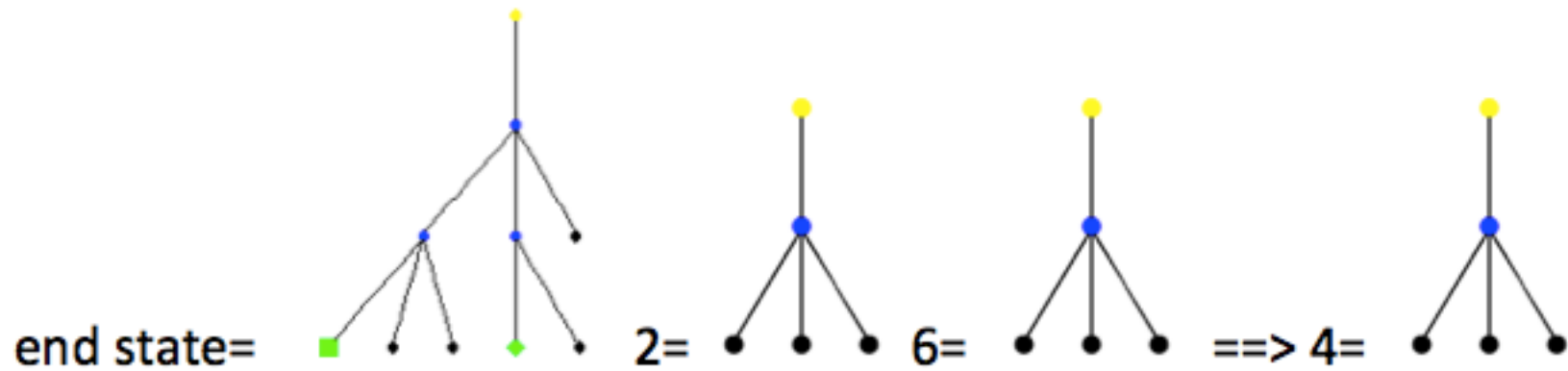
Theory Building

- Splitting is a theory of how kids learn fractions
- Look splitting does matter

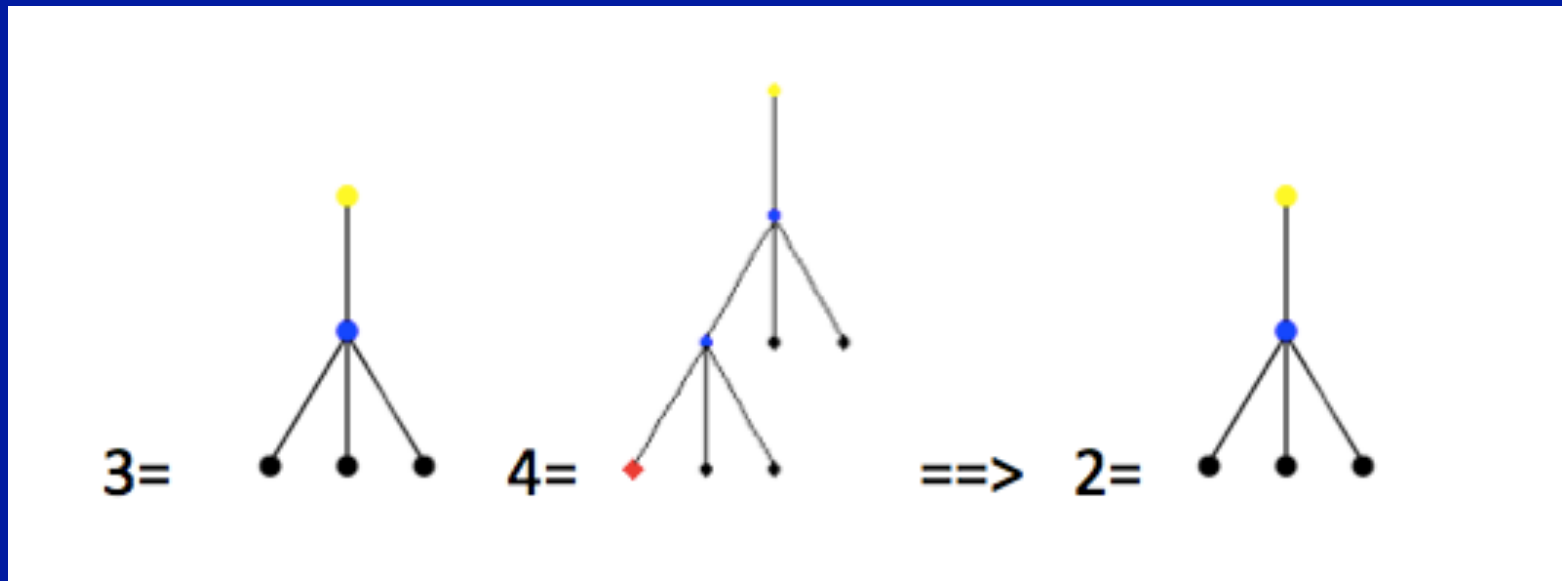
Prelevel

- Used ARM to determine if a learner bought onions and potatoes (i.e., a $1/3$ and a $1/6$ state), what else did they buy (e.g., hamburger or a $1/9$ state)

Pre Level: Association with Success



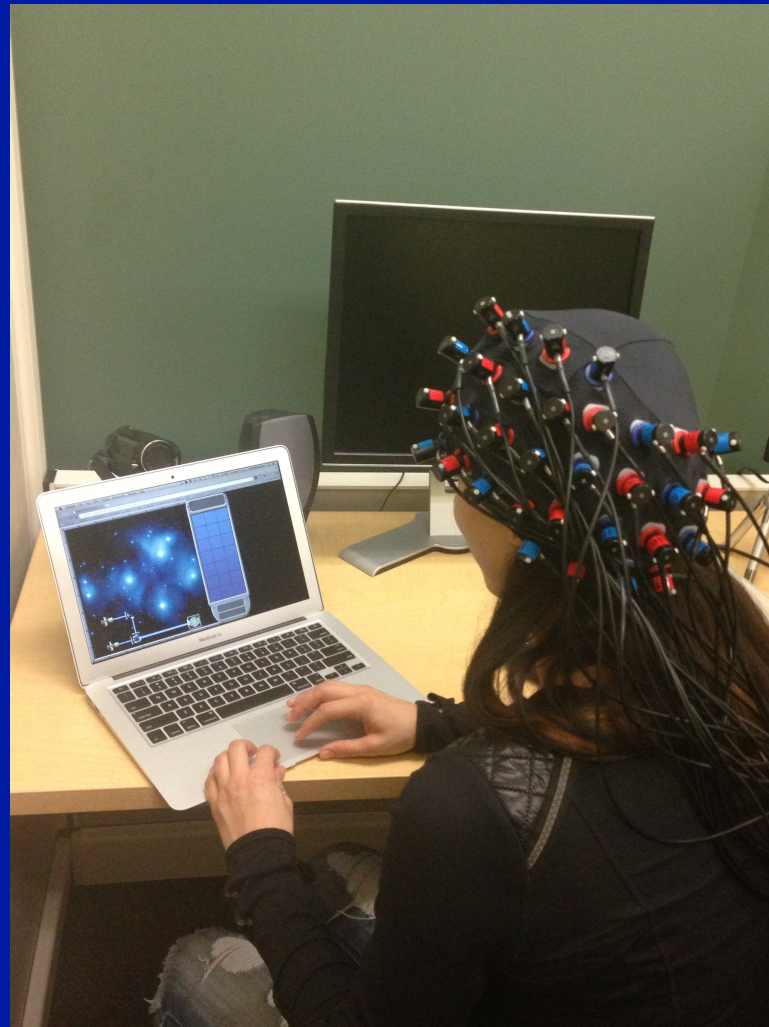
Post Level: Association with Success



Present Results to a Variety of Audiences

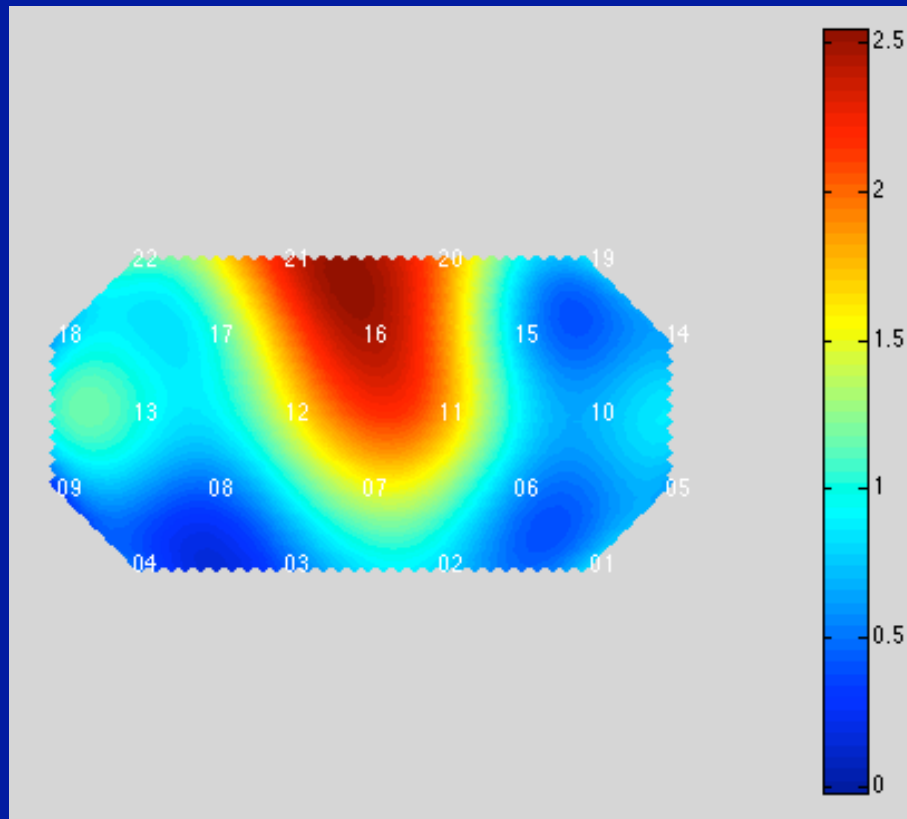
- Neuroscience for AERA

Refraction NIRS Study

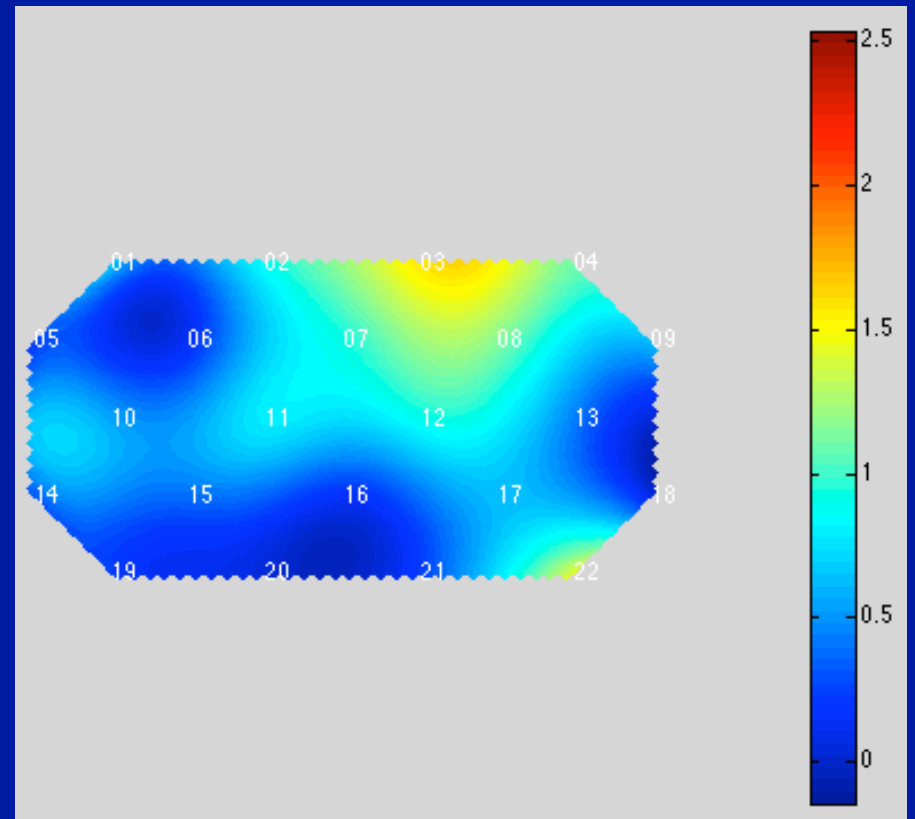


Math vs. Refraction Condition

Prefrontal Patch (viewing head-on)



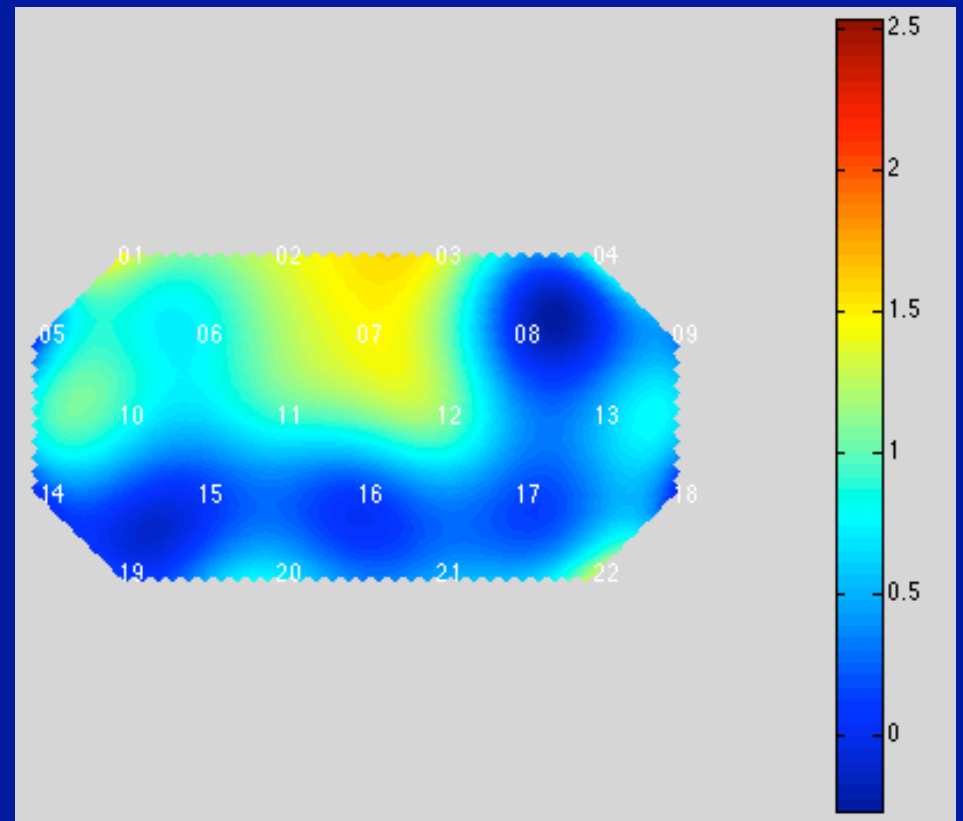
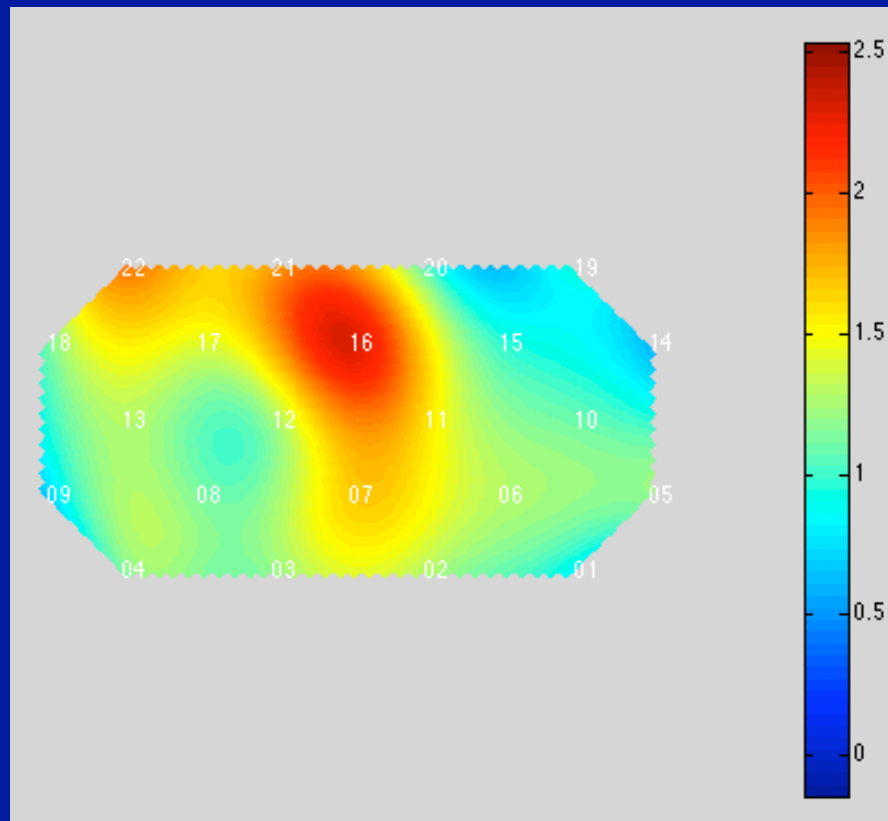
Parietal Patch (viewing from behind)



Math vs. Spatial Condition

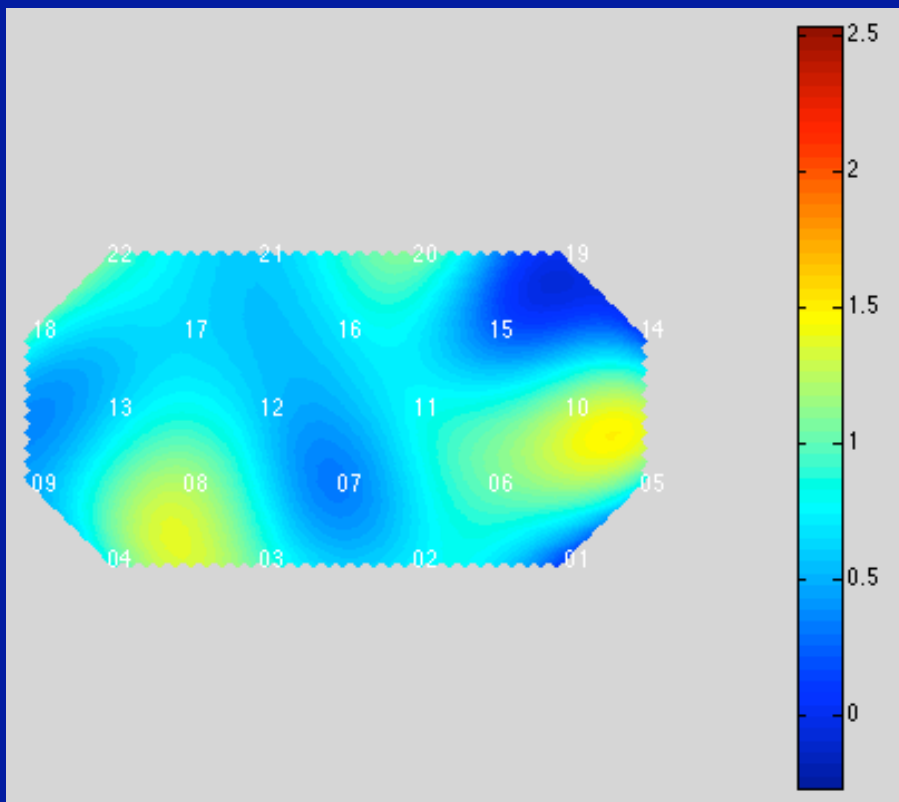
Prefrontal Patch (viewing head-on)

Parietal Patch (viewing from behind)

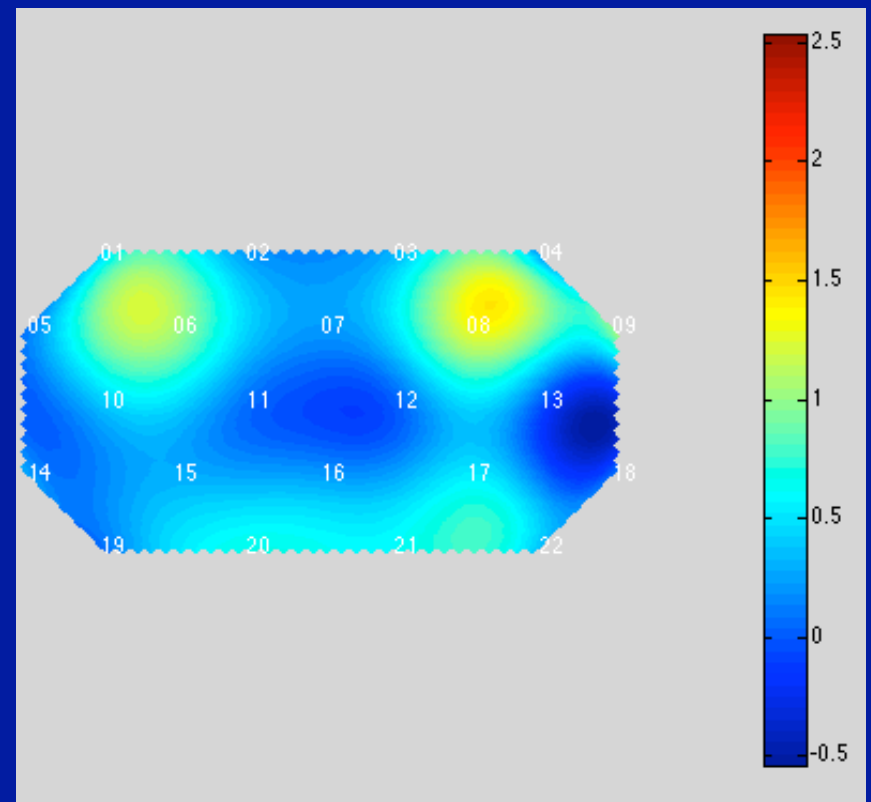


Refraction vs. Spatial Condition

Prefrontal Patch (viewing head-on)



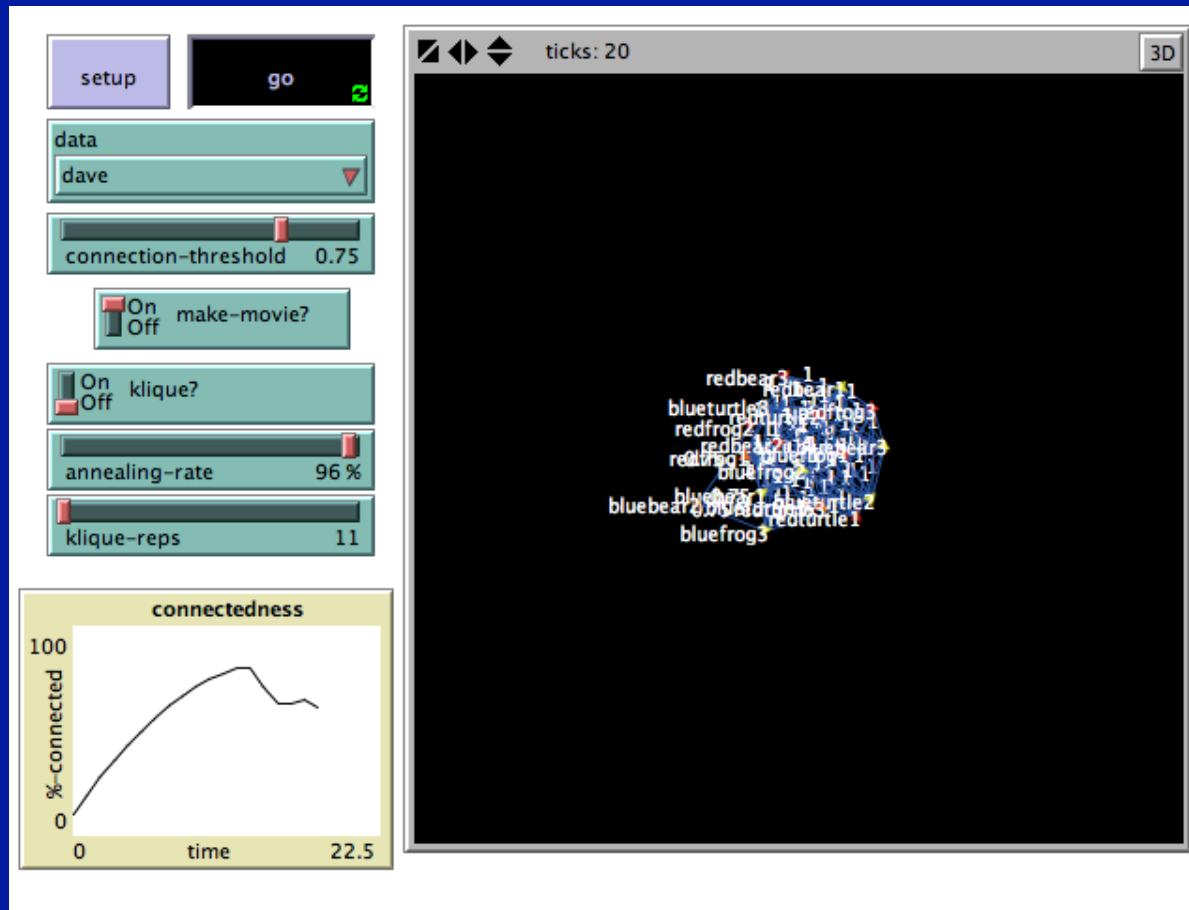
Parietal Patch (viewing from behind)



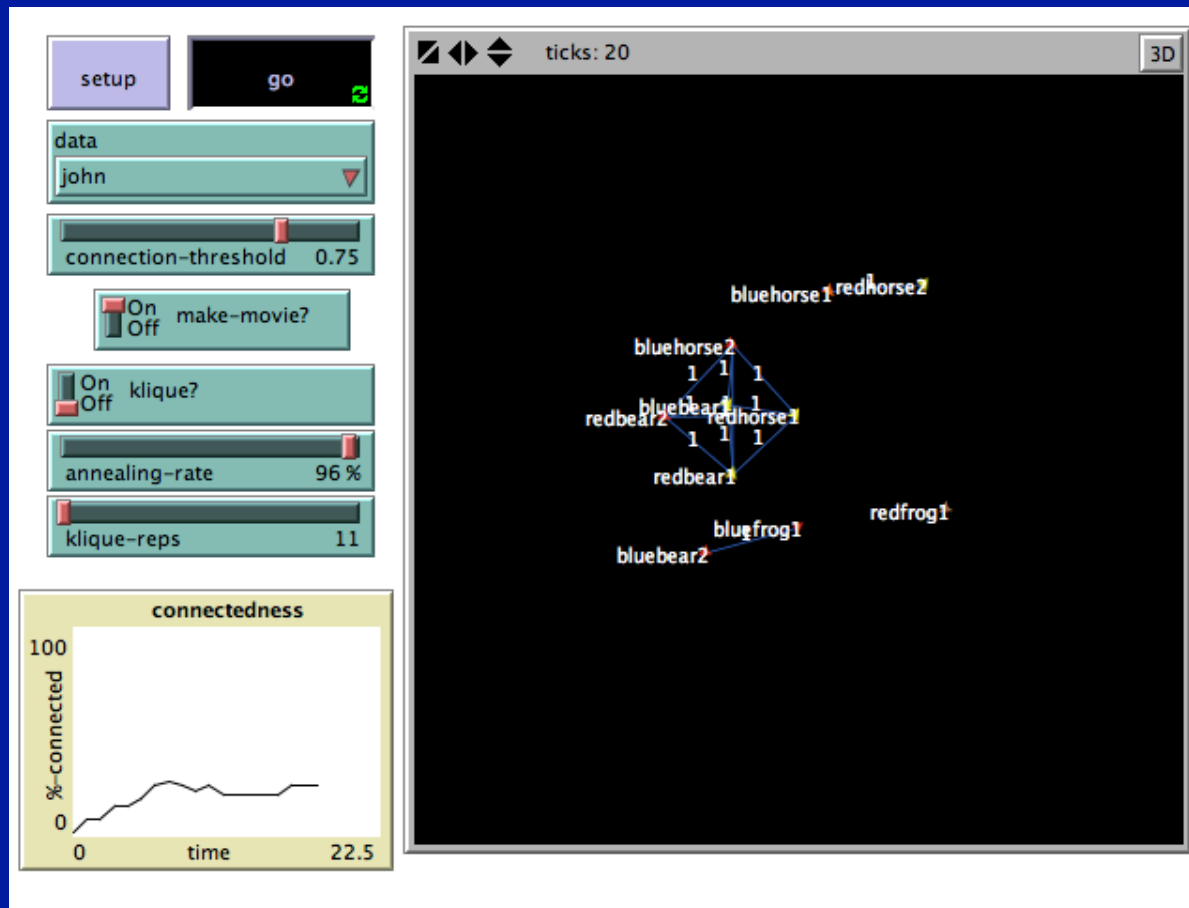
Representing Change over Time

- Changes in students' programming
 - Comparing novices and experts
 - Development of programming in Scratch

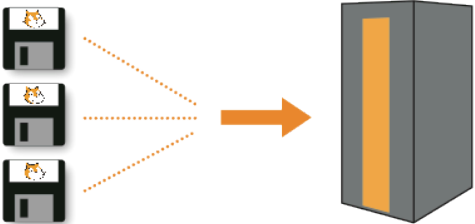
Similarity in Novices' programs



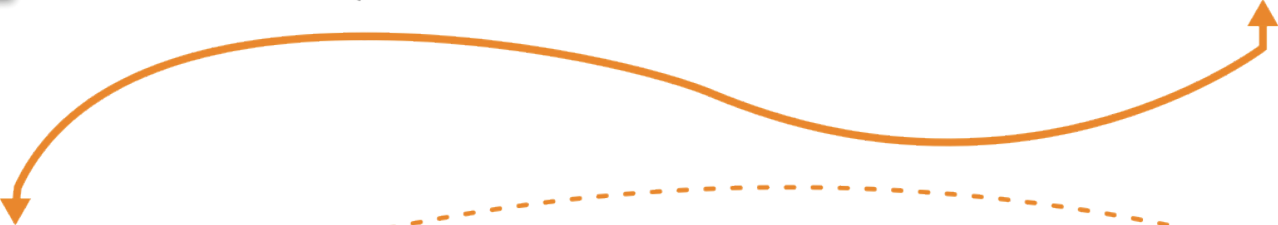
Similarity in (relative) experts' programs



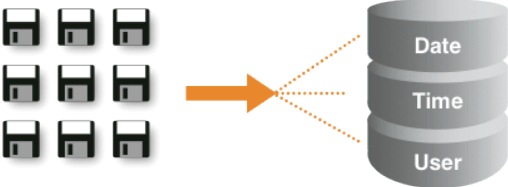
Collect



Parse



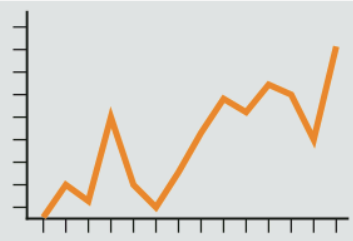
Manage



Analyze



Report



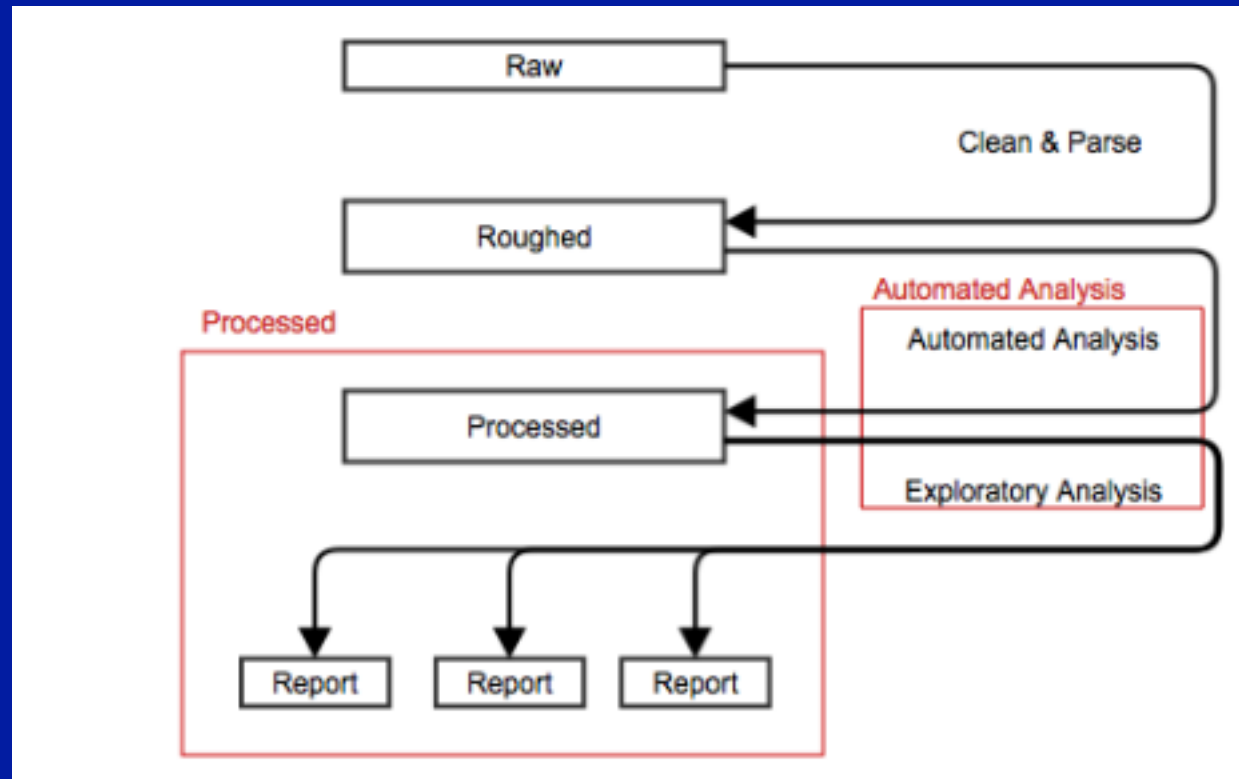
Challenges (Stages 2 & 3a)

- Capacity
 - Particularly for real Big Data
 - Quickly changing teams
- Keeping the pipeline as your guiding framework

But

- New and developing tools to help at this point, e.g.:
 - RStudio
 - Rapid Miner
 - Weka
 - MySQLWorkbench
- Capacity building efforts within the field
 - LAMP, programs at CMU, TC, etc
 - LASI and events at LAK & EDM
 - LearnLab at CMU

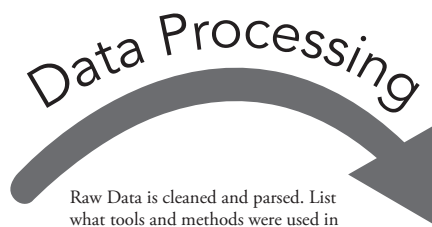
Stage 3b



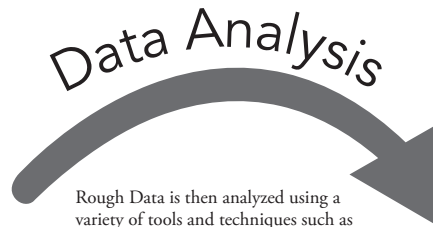
Auto generated representations of student learning, progress, engagement, etc for teachers, parents, students...

We all know I mean dashboards, but hey, I don't have a picture I like

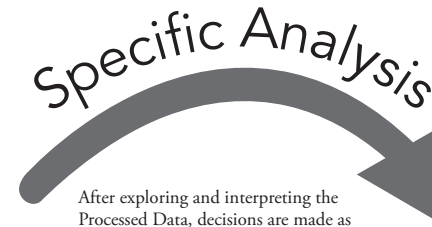
Standard Data Pipeline



Raw Data is cleaned and parsed. List what tools and methods were used in the process. After the data has been cleaned and parsed it then becomes Rough Data.



Rough Data is then analyzed using a variety of tools and techniques such as R, Python, Tableau, or SPSS. The results of the analysis of the Rough Data become Processed Data.



After exploring and interpreting the Processed Data, decisions are made as to what specific calculations and analysis should be done for a specific report or paper. The results become Report Specific Data.

Raw Data

Format

Here is where the exact format or file for all data is specified. So for example raw data could be in the form of a .csv, json, .mp4, .doc, etc. It is also important to document version history and dates.

Location

Usually this Raw Data should be stored on a secure server, or even in a database. This should be made explicit with instructions on how to access the data.

Data Cleaning

Before the data can be analyzed it has to be cleaned and processed in to a usable format and stored in a database.

Rough Data

Format

What format is the data in after it has been cleaned and parsed? What are the files named? What version number is this? All these questions should be answered here, and include meta-data.

Location

Typically, the data at this point should be stored in a database of some kind. Where this data lives and how it is structured should be very explicit.

Analysis

Data is analyzed using tools such as R, Python, Tableau, or SPSS. Often the analysis at this stage is exploratory.

Processed Data

Format

Measures, visualizations, exploratory data files, graphs, charts, etc.

Location

Some processed data might live in a database, but most will reside either on a server or Dropbox. The exact location should be documented.

Analysis

After the data has been explored and processed it is analyzed using calculations and methods with a specific paper or report in mind.

Report Specific Data

Format

Visualizations, Charts, Graphs, Descriptive Statistics, Inferential Statistics.

Location

Data to be used in a report or paper should be located in the same location as the report. This could either be in Dropbox, or in a Google doc. These should also be archived for future reference.

Challenges

- Teachers are overwhelmed with 900 dashboards, LMSs, games, etc.
- Tools for creating and maintaining your data pipeline limited
- Managing teams

But

- Not a lot here now, guess that's why I thought I might talk about this
- Ideas?

Thank You!

- activelearninglab.org
- taylor.martin@usu.edu

BILL & MELINDA
GATES *foundation*



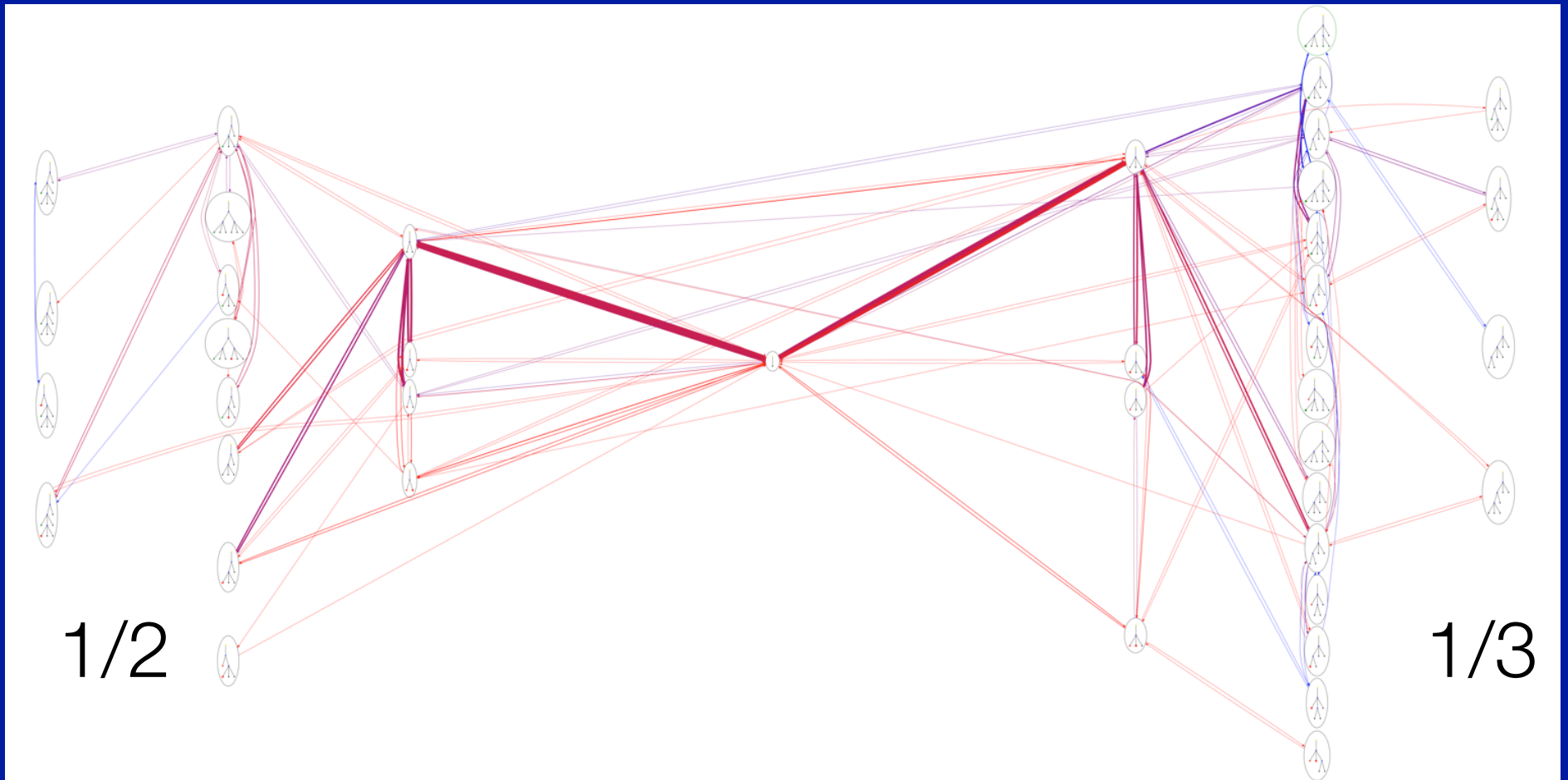
ies INSTITUTE OF
EDUCATION SCIENCES

Extras

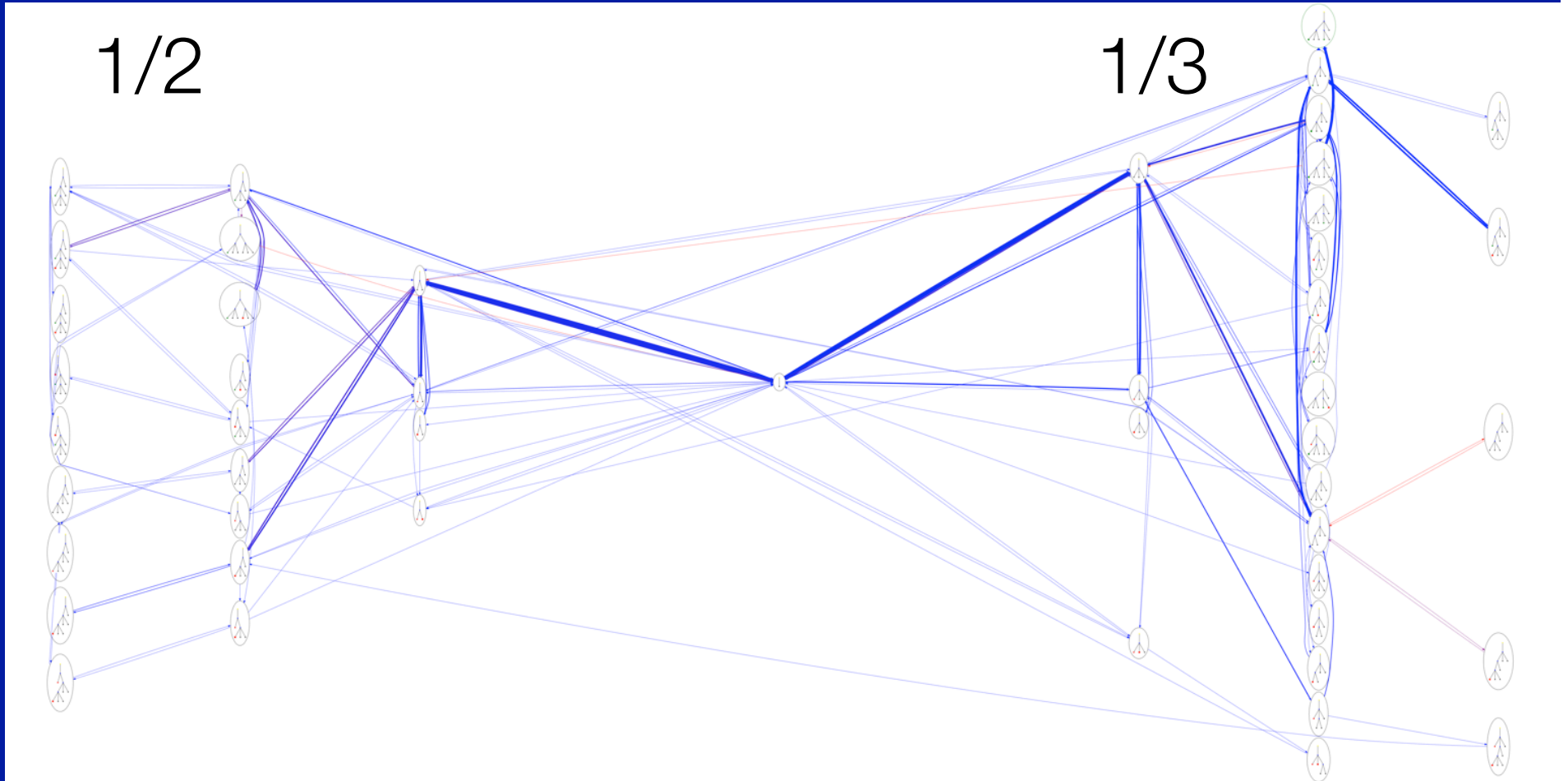
Refraction



Pre



Post



Association Rule Learning

- Discover regularities between variables in large datasets
- e.g., Large-scale transaction data recorded by POS systems in supermarkets:

{Onions, Potato Chips} → {Burger}

*Agrawal et al

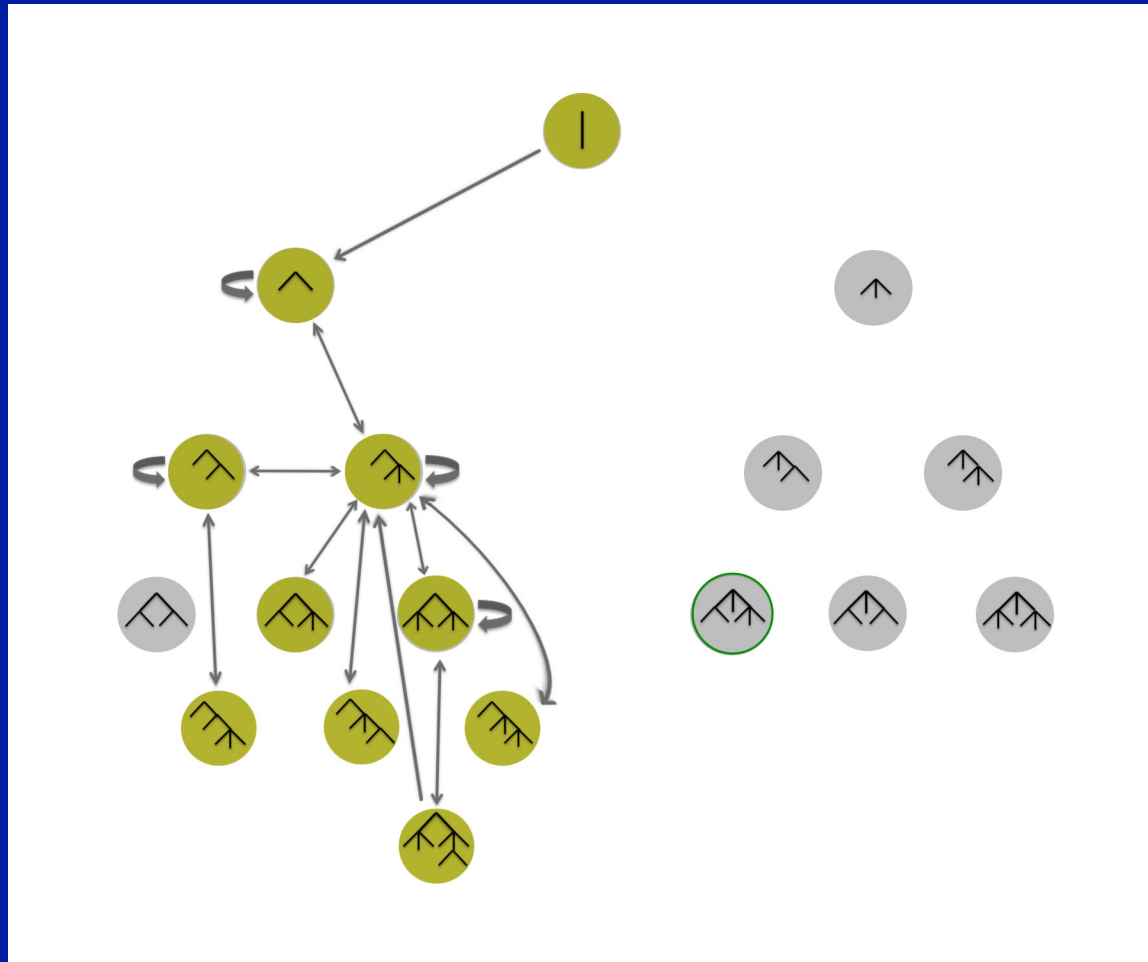
Initial Conclusions

- Fussing with $1/3$ (central conceptual hurdle)
 - Productive even if not achieving obvious goal
- Fussing with $1/2$
 - Unproductive unless used to correctly hit target

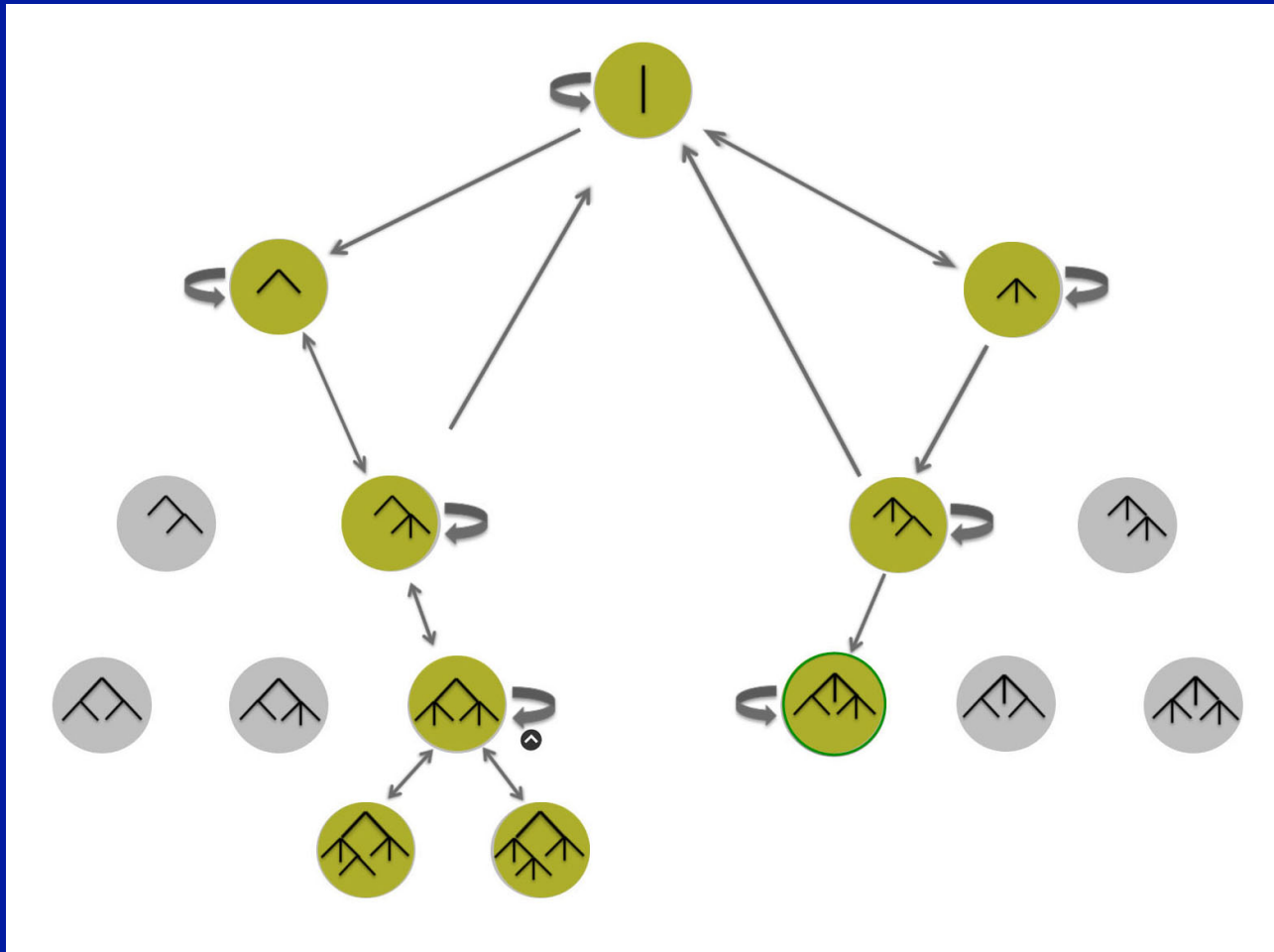
Cluster Analysis

- Explore Fussing in more depth
- Variables (Generated in Stage 2 with visualizations)
 - Number of unique board states
 - Total number of board states
 - Average time on board state
 - Number of moves till hit 1/3 board state
 - Time on level

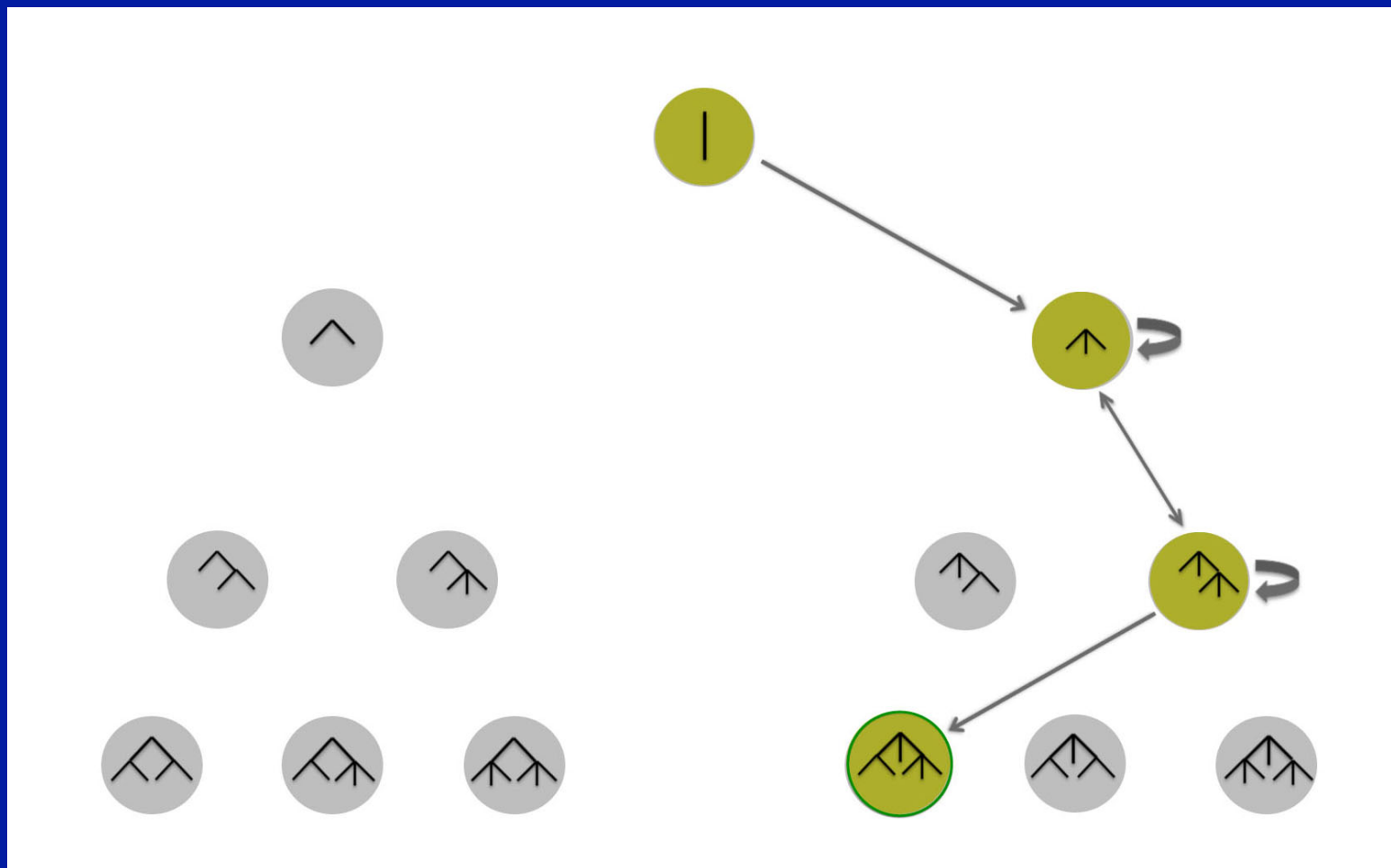
Haphazard



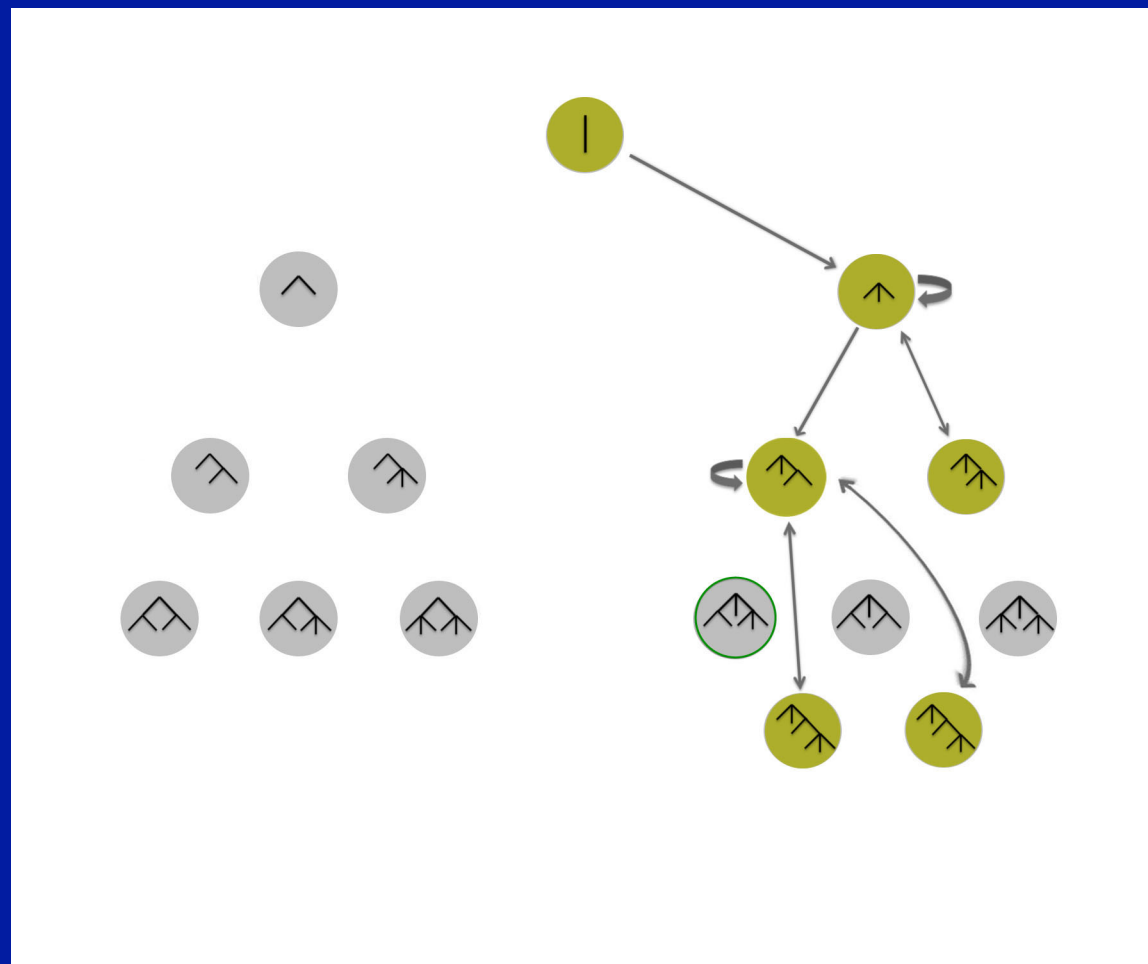
Exploration



Careful



Minimal



Relate Clusters to Transfer

- Unproductive
 - Haphazard
 - Minimal
- Productive
 - Exploration
 - Careful