From MOOCdb
to
MOOC Data Science Commons

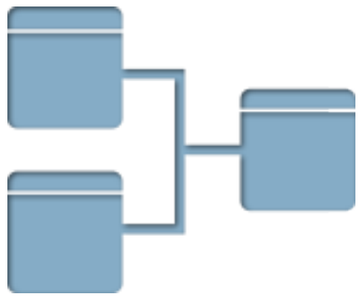
Goal: To build the largest shared repository of MOOC data science apps

Kalyan Veeramachaneni
Una-May O'Reilly
CSAIL, MIT

NSF DIBBS KickOff
Carnegie Mellon University
January 15, 2015

Where were we?

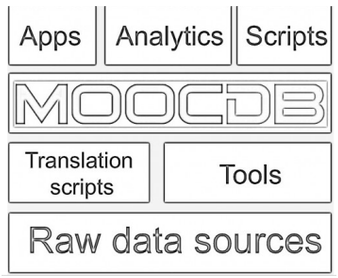
Circa. April 2014 (when we started to think about this proposal)



Shared data model

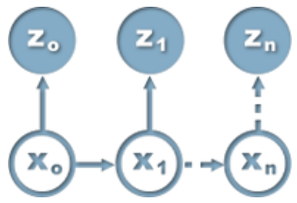
Discussions with community

Community input



Software stack

Feature foundry

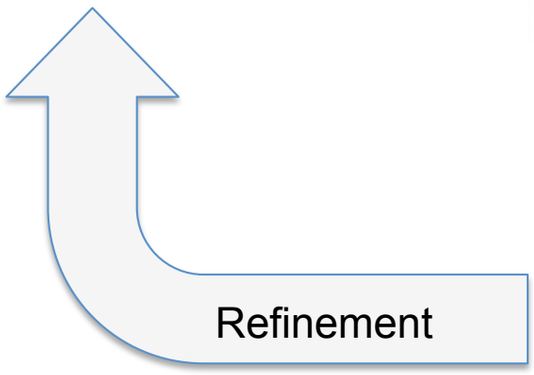


Machine learning

MOOCviz

Privacy preserving
Data access

Collaborative, open source frameworks

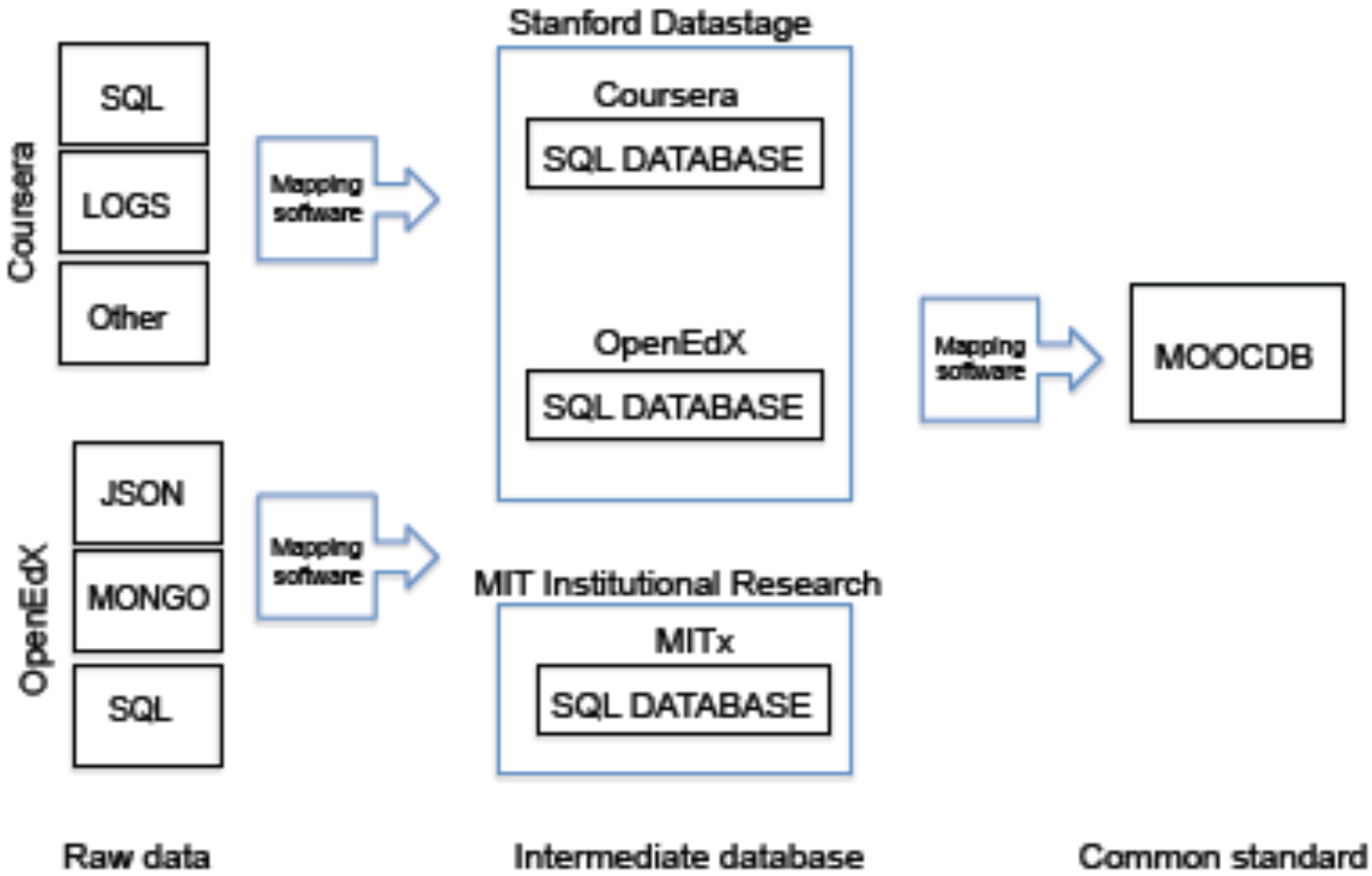


Phase 1

Phase 2

Phase 3

Focus was on translating raw data to MOOCdb



Current state of our resources?

Circa. January 2015

coursera

Translate

edX

Translate

Software

Github

MOOCdb Docs

Wiki

Where are we today?

Circa. January 2015

> 100 course offerings
67 Coursera courses
30 edX courses
5 OpenedX courses

Total number of events processed
> 1 Billion

Total amount of data

~2TB

Finished

THE UNIVERSITY OF TEXAS AT AUSTIN

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Carnegie Mellon University

In progress

What is an app?

- An end-to-end packaged software chain with several value additions:
  - Integrated curation and processing tools
  - Appropriate and easy-to-use interfaces
  - Documentation (web based)
  - Use cases
  - Demonstration

Types of apps

- Machine learning (ML) apps
- A machine learning app is designed to build a model of the data that solves a predictive, correlative and/or analysis problem.
    - end-to-end packaged software includes:
        - data preprocessing and curation
        - data assembly to make it ready for the machine learning approach
        - machine learning algorithm implementation
        - structured model output and storage
        - model analysis
        - visualization

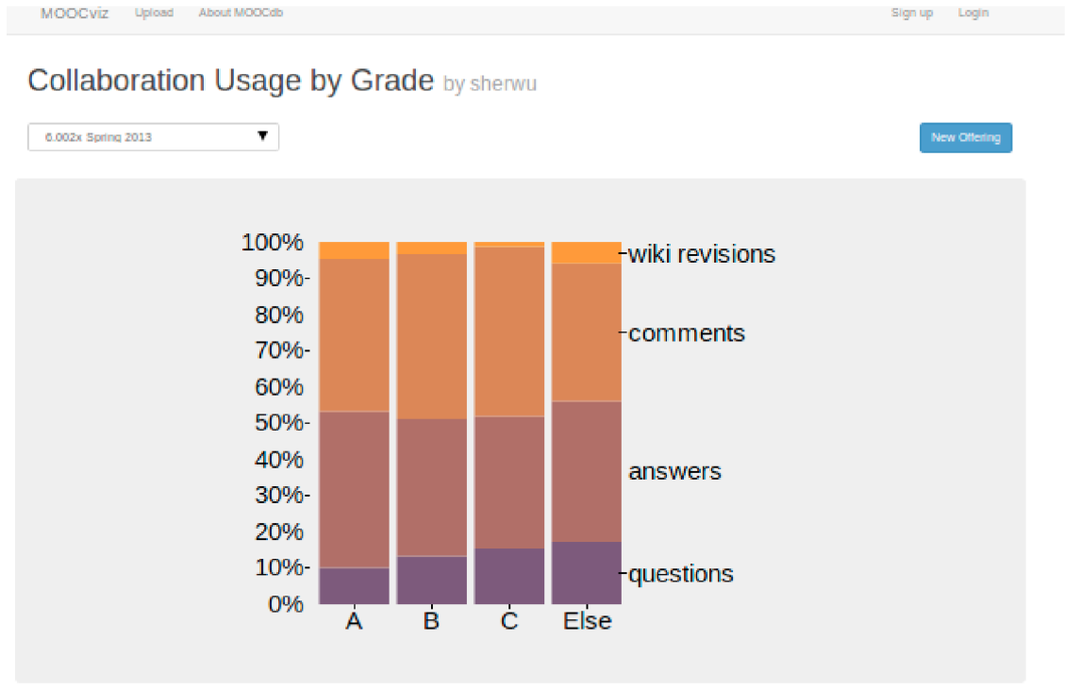
Types of apps

- Apps that are “Platforms”
- A “Platform” according to our definition is a web based system with which humans (analysts, instructors, data scientists, researchers) interact and it collects “human input and stores it in a persistent database” *.
- Platforms enable us to solve endeavors that are human driven and are complimentary to automatic data processing, machine learning and analytic tools.
- They allow us to collaboratively work together and share our work through web based interfaces.
- Examples are: MOOCviz, Feature Factory, LabelMe-Text
    - An end-to-end packaged software includes:
        - data preprocessing and curation
        - data assembly to make it ready for the platform
        - user interfaces
        - backend database and storage design
        - analytics on the backend database
        - modules for platform administration

* Ken’s suggested definition.

To read about our current design, please read here and here about MOOCviz development and here about our feature factory.

An example platform app- MOOCviz



This visualization shows what different types of collaborations were used by different types of students. We looked at students based on their final grade in the class - 'A', 'B', 'C', or 'else', where else means the student didn't obtain a passing certificate for the class.

Download ZIP of all files

collaborations_percentage_by_user_grade.csv

```
State,questions,answers,comments,wiki revisions
A,0.0001574493356,0.433542424467,0.429347412061,0.0460527132358
B,0.13322248779,0.379091750327,0.454380130936,0.0333955219474
C,0.152642322522,0.365842504693,0.469040508775,0.01247448401
Else,0.173850241749,0.387092705487,0.38891233971,0.0580447139544
```

Download File

© 2013 MIT CSAIL ALFA Lab.

Current apps

as listed on our website (some are in progress)

Contributed by

Github and docs location

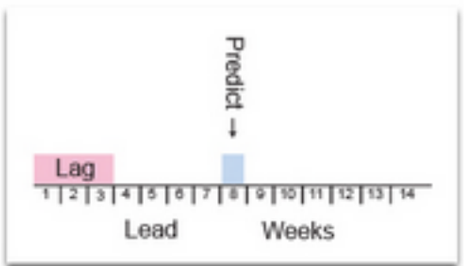
Coursera
Stanford . github . docs

edx
MIT . github . docs

My Feature Factory
MIT . github . docs

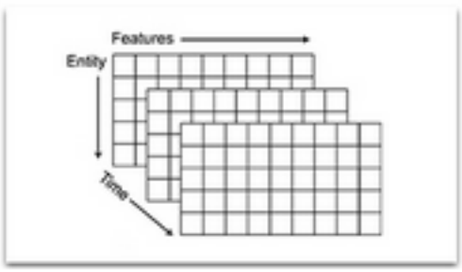
MY MOOCViz
MIT . github . docs

Whats happening in forums?
MIT . github . docs

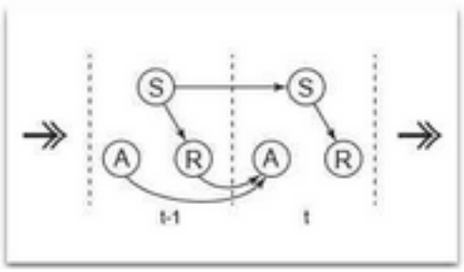


Predict who is likely to stop?
MIT . github . docs

Community detection
MIT . github . docs



Digital Learner Quantified
MIT . github . docs



Problem Analytics
MIT . github . docs

LabelMe
MIT . github . docs

Translation and curation apps

coursera

Translate

edX

Translate

MY MOOCViz

LabelMe

MY FEATURE FACTORY

Apps that are platforms

coursera

Translate

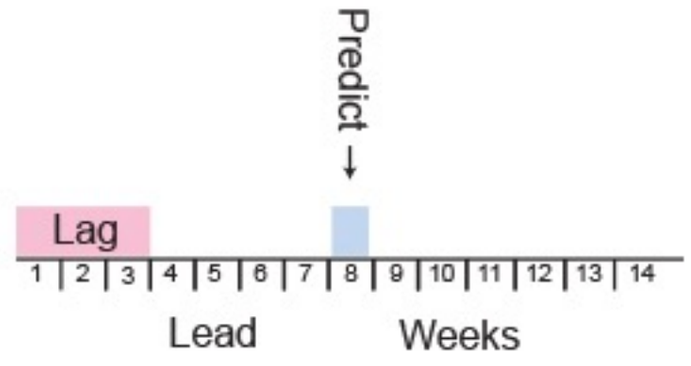
edX

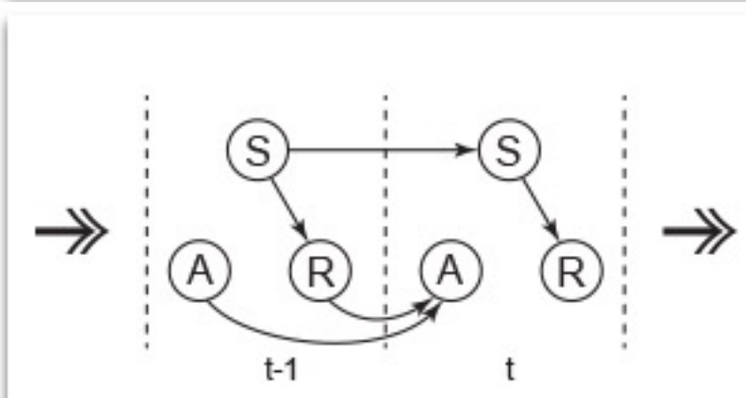
Translate

MY MOOCViz

LabelMe

MY FEATURE FACTORY





Machine learning apps

coursera

Translate

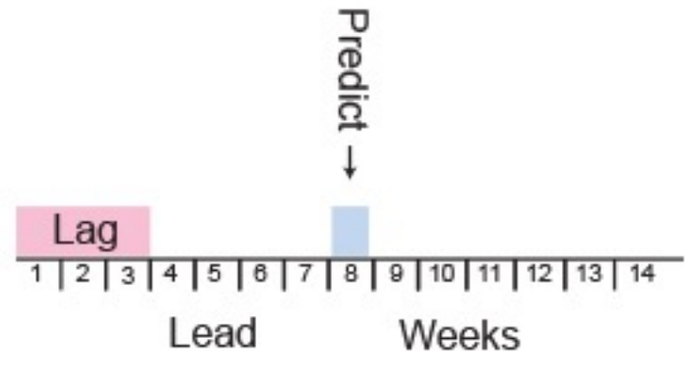
edX

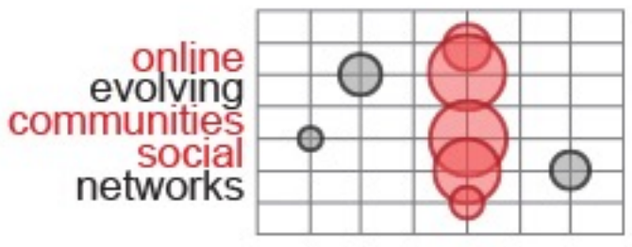
Translate

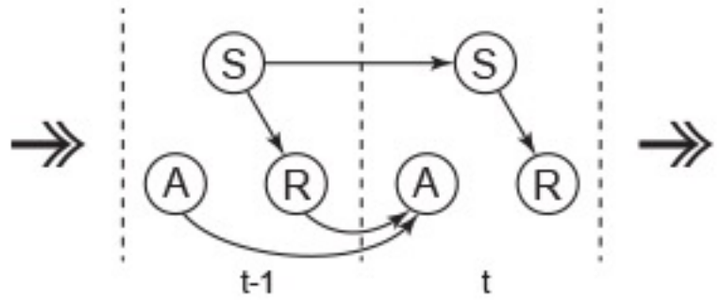
MY MOOCViz

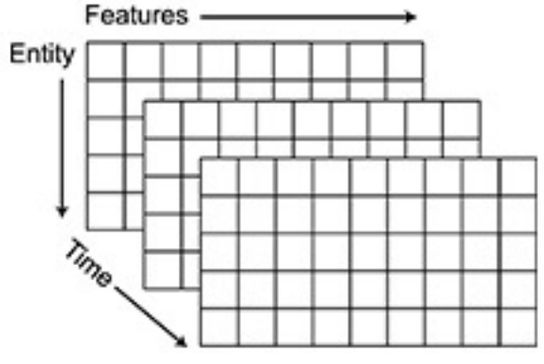
LabelMe

MY FEATURE FACTORY









Analytics apps

coursera

Translate

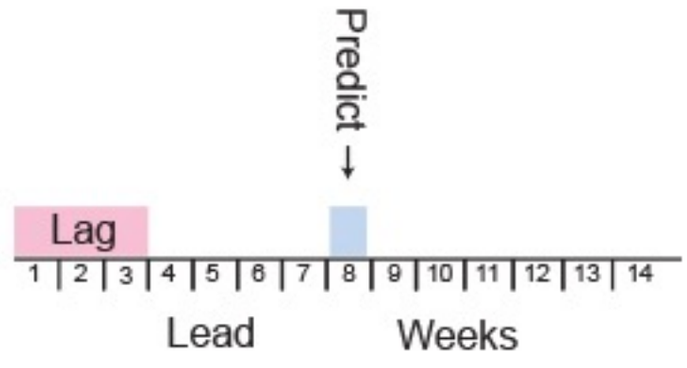
edX

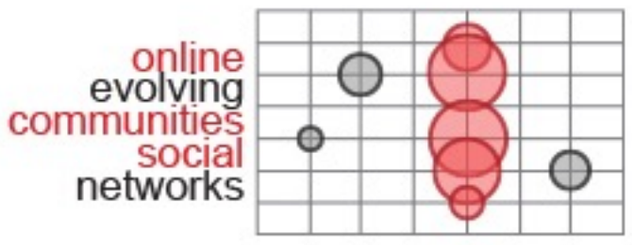
Translate

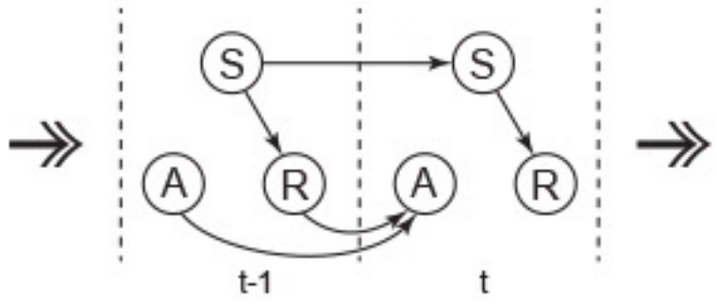
MY MOOCViz

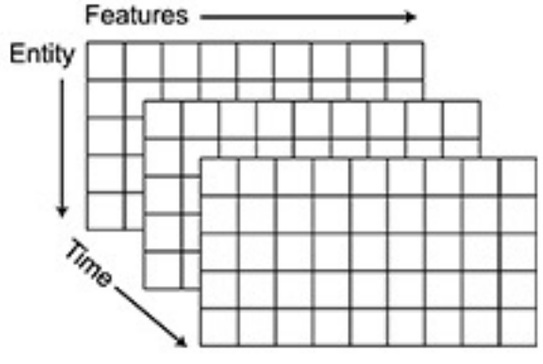
LabelMe

MY FEATURE FACTORY









Current apps
circa 2015

How does an app come about?

- Starts with an idea and then:

    1. Refine the idea with input from multiple stakeholders
    2. Test it at MIT (use an MIT course data, if it is a data/machine learning app)
    3. Test it with external partners
    4. Generalize and refine
    5. Release (github source code, documentation)

- For a full release (finishing these steps 1-5), and development of an app it takes us 2 - 2.5 years

- Note that we build multiple apps simultaneously

- Some of the apps we showed today have been through Steps 1, 2 and 3 and some are at 2 and some are at 5.