MoocDB
Taming MOOC Big Data while Fostering Collaboration in Online Education Research

Una-May O'Reilly

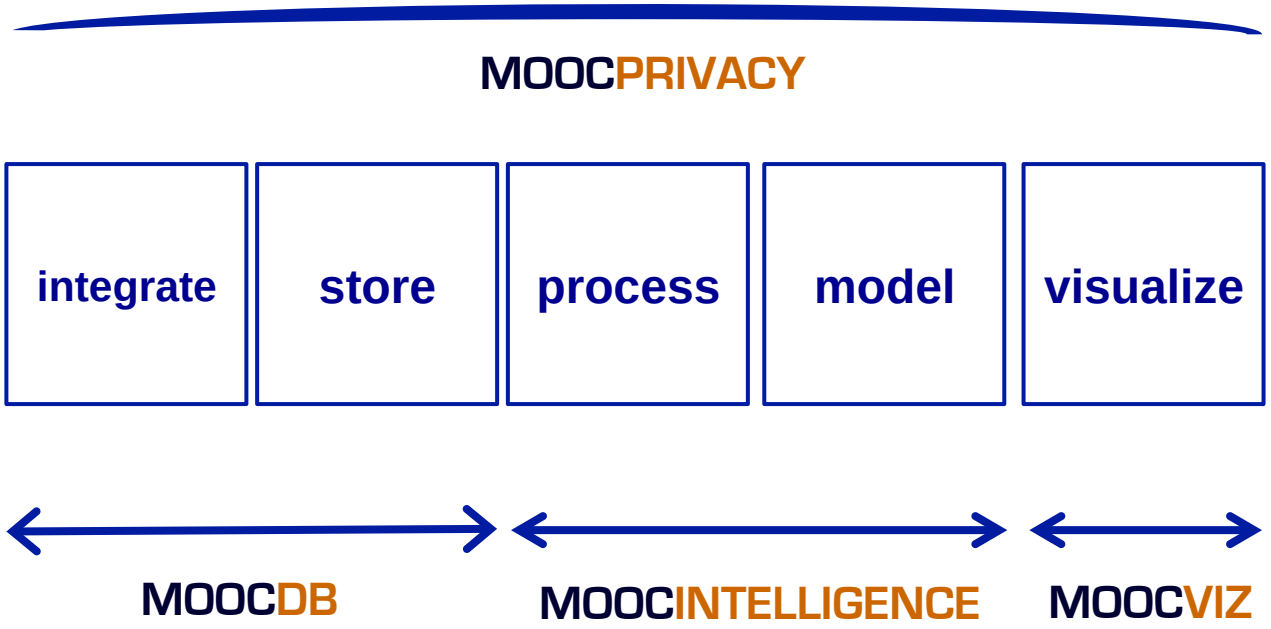
AnyScale Learning for All Group: ALFA

Computer Science and Artificial Intelligence Lab

MIT

http://groups.csail.mit.ed/ALFA/groupWebSite/index.php?n=Site.AlfaX

ALFA MOOC Data Science

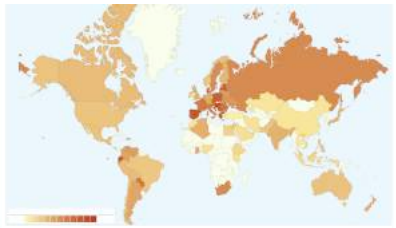


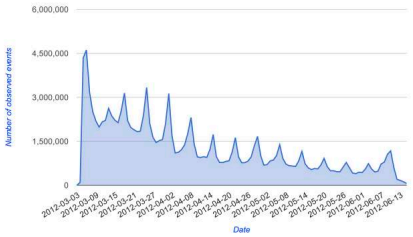
integration… to… insight

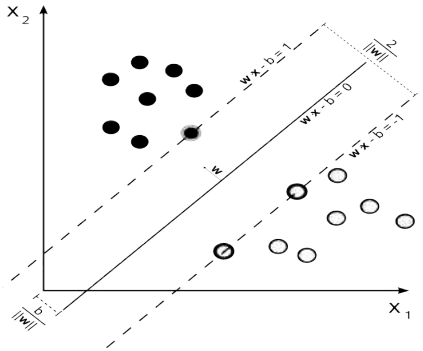
ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

ALFA MOOC Research







Who is likely to stopout?
Community detection
Weekly topic analysis

Crowd Sourcing

Access
Privacy Protection
Policy
Differential privacy

Shared data model

Shared analytics

Machine Learning

Privacy as a service

MOOCDB

MOOCVIZ

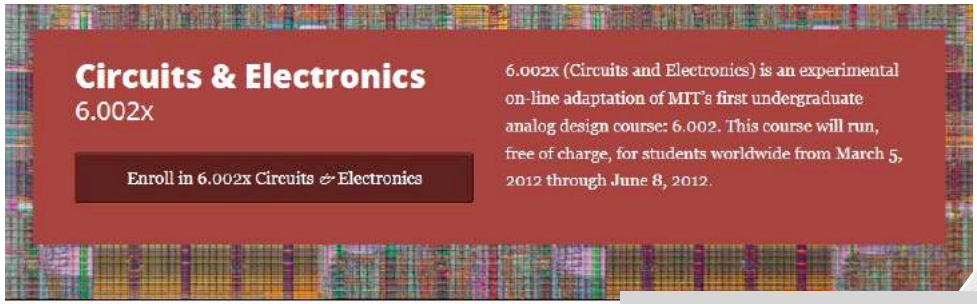
MOOCINTELLIGENCE

MOOCPRIVACY

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Massive Online Open Courses



6.002x on *MITx*

If you successfully complete the course, you will receive an electronic certificate of accomplishment from *MITx*. This certificate will indicate that you earned it from *MITx*'s pilot course. In this prototype version, *MITx* will not require that you be tested in a testing center or otherwise have your identity certified in order to receive this certificate.

ABOUT THE COURSE STAFF

Anant Agarwal

Director of MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) and a professor of the Electrical Engineering and Computer Science department at MIT. His research focus is in parallel computer architectures

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

MOOC Stakeholders

- Instructors
- Students
- Data custodians/guardians
- Course designers
  - Education technology specialists
- Education/Learning researchers

MOOC Research Questions

- Descriptive information
  - Who? When?
    - Demographics and grades, statistical correlations
- MOOC specific
  - Trajectory related
  - Resource related
  - Using the crowd
  - Response related
- General questions about learning and education
  - Learning styles?
  - Knowledge acquisition
  - Flipped classrooms, blended learning

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

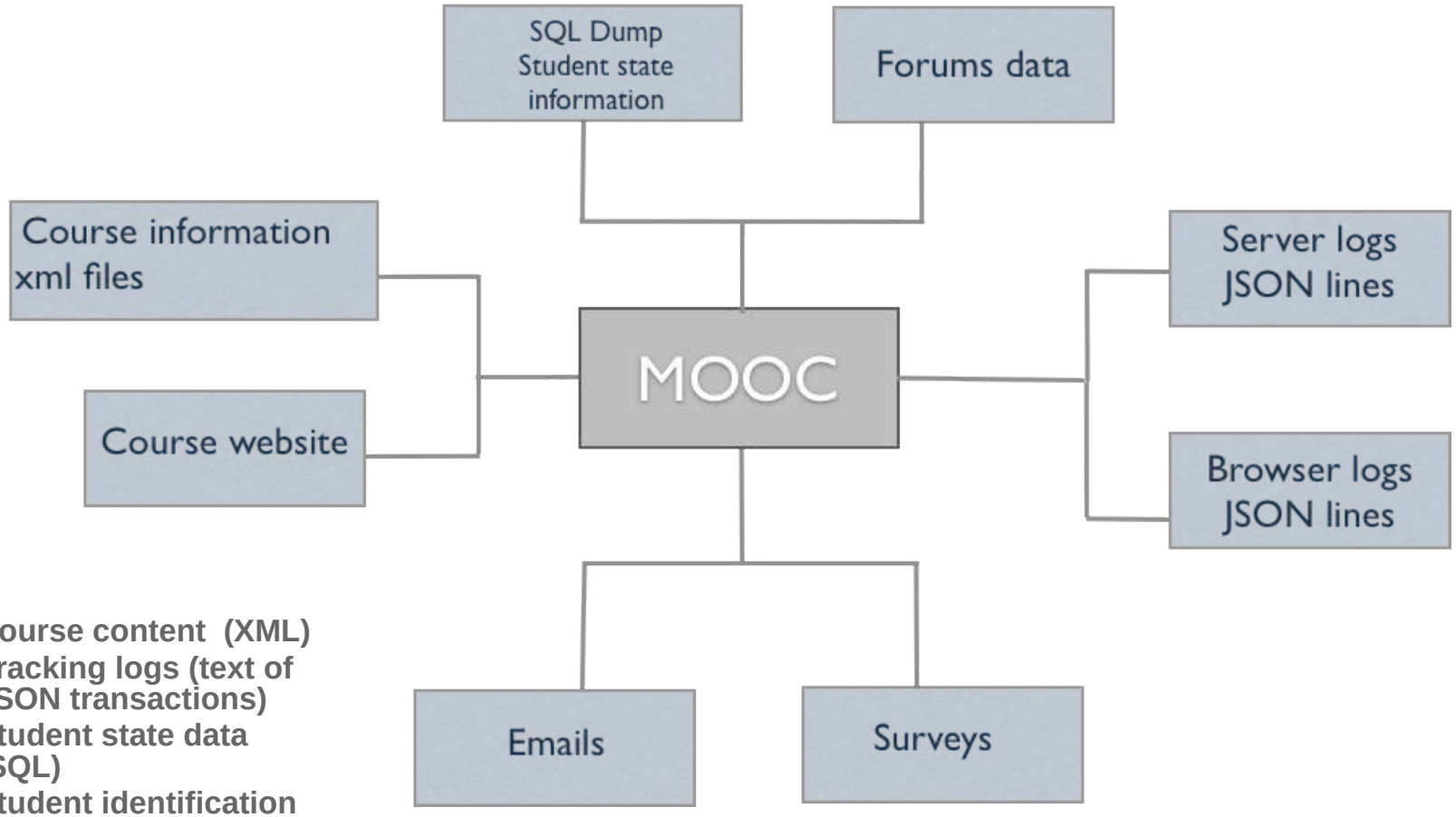
Behavioral Analysis

- Hypothesis
- Assemble data and features
- Statistical model
  - Validate, inspect, interpret
  - visualize

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

A MOOC Data Management Problem



- … **Course content (XML)**
- … **Tracking logs (text of JSON transactions)**
- … **Student state data (SQL)**
- … **Student identification data (SQL)**
- … **Forum data (NOSQL)**
- … **Wiki data (SQL)**

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Pain Points and Bottlenecks

- …Heterogeneous data formats
- …Bloated raw data storage
- …Lack of a comprehensive view of the data
  - …Needs to be organized according to use!
- …Un-identified cross-platform compatibilities
- …Wasted effort replicating efforts of others

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

What about …

Multiple courses?

Multiple platforms?

How can we bring many eyes to the data?

Enable and encourage community reflection and intellectual engagement around it

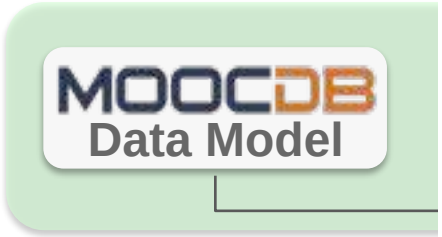
http://www.flickr.com/photos/hagdorned/7434861784/

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

MoocDB

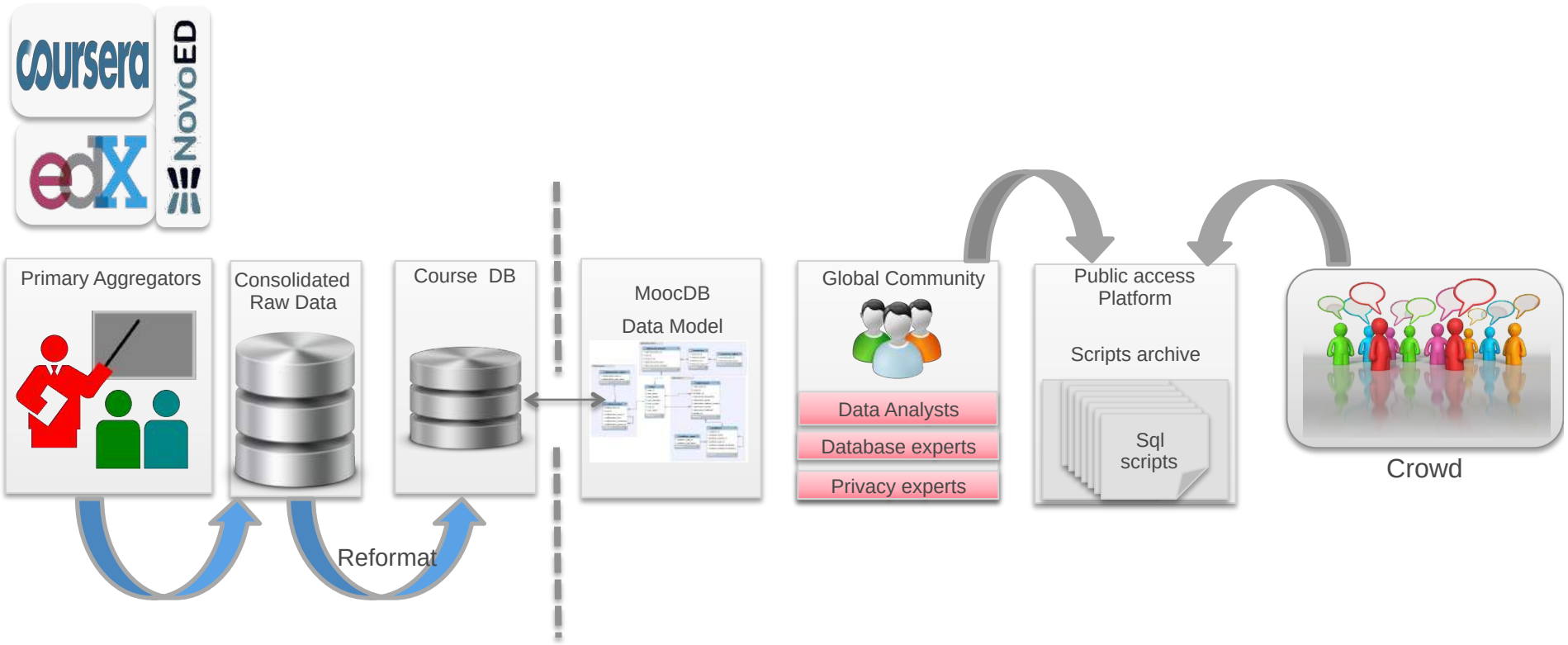
Shared data model



ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

MoocDB



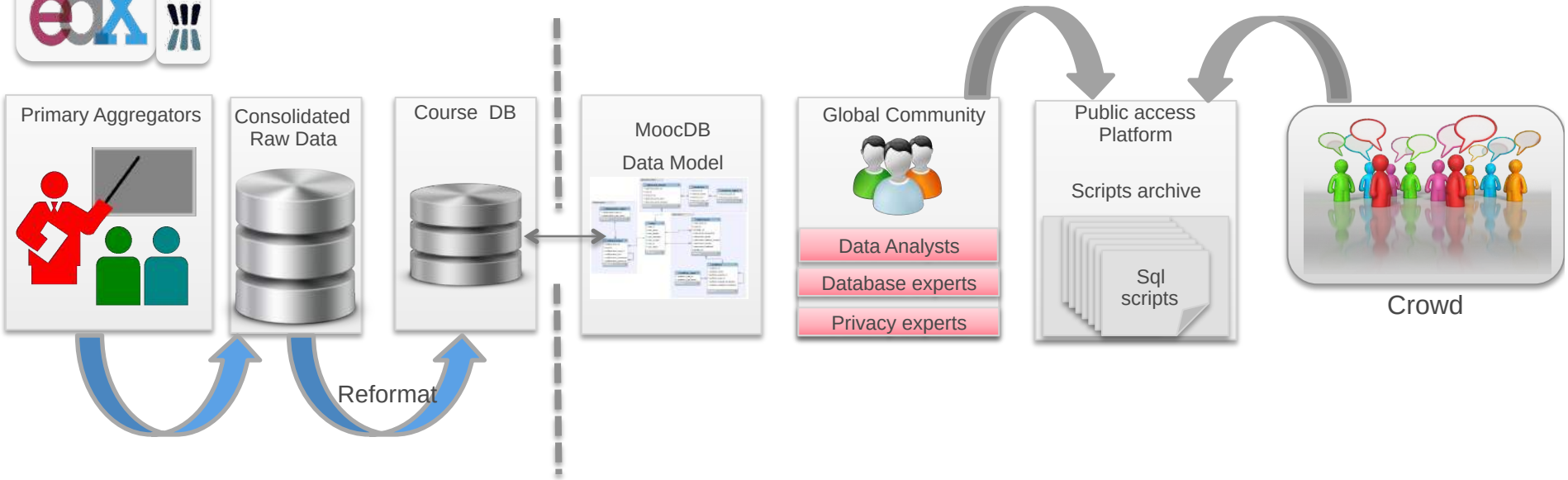
MoocDB: Data organization to support many eyes on the data

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

MoocDB

There is an transparent but protective interface between course DB and the researchers which is facilitated by the data model

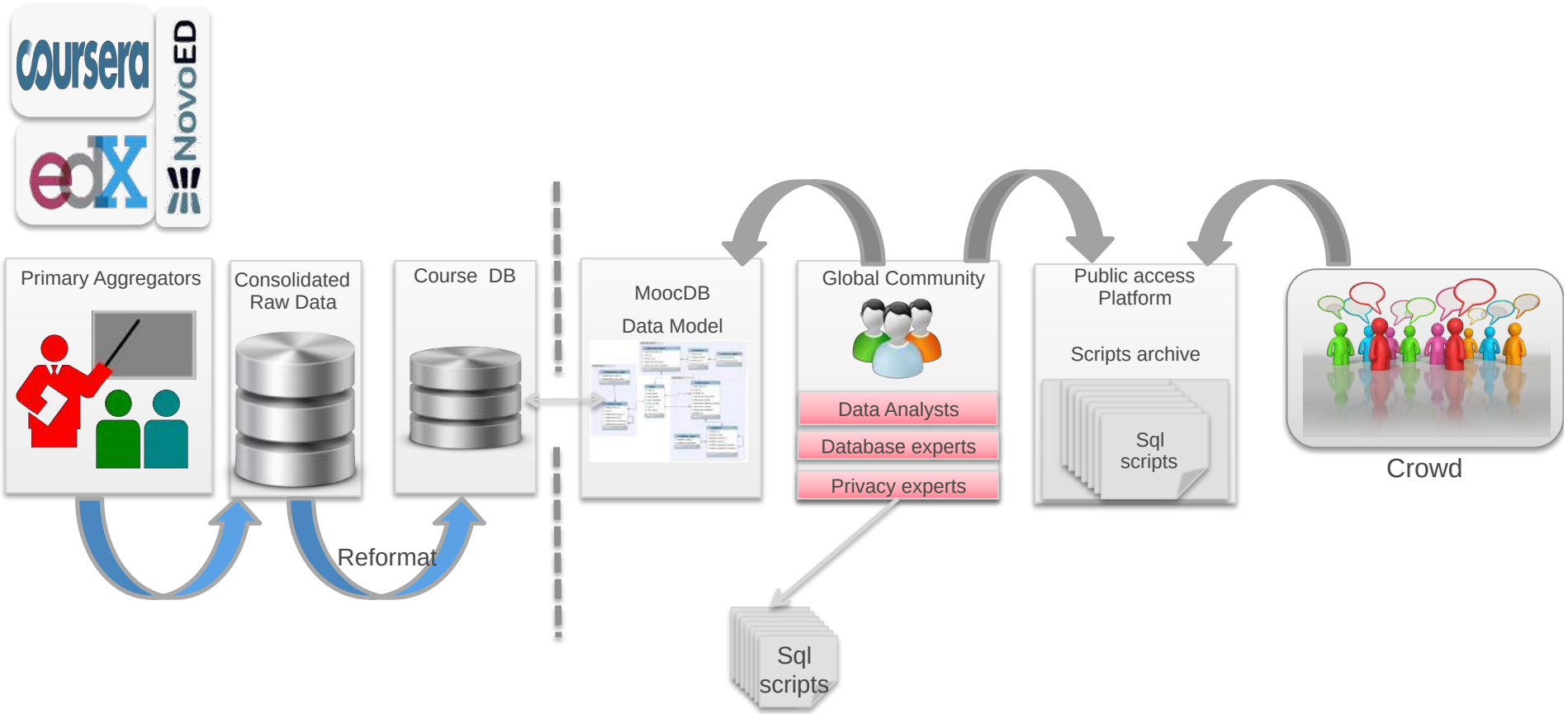


MoocDB: Data organization to support many eyes on the data

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

MoocDB

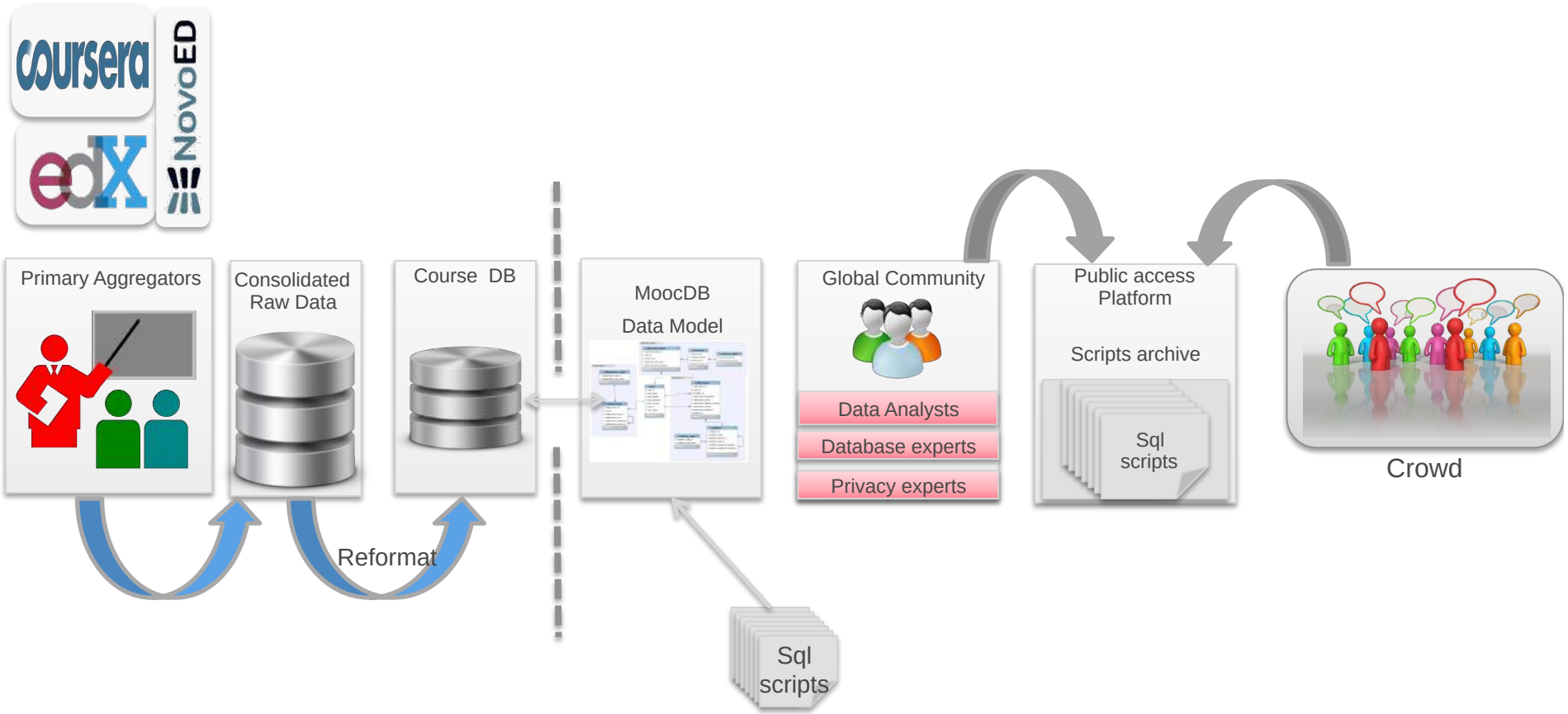


Community Visualization and Data Analysis:
Step 1: analysts write scripts by consulting the data model

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

MoocDB

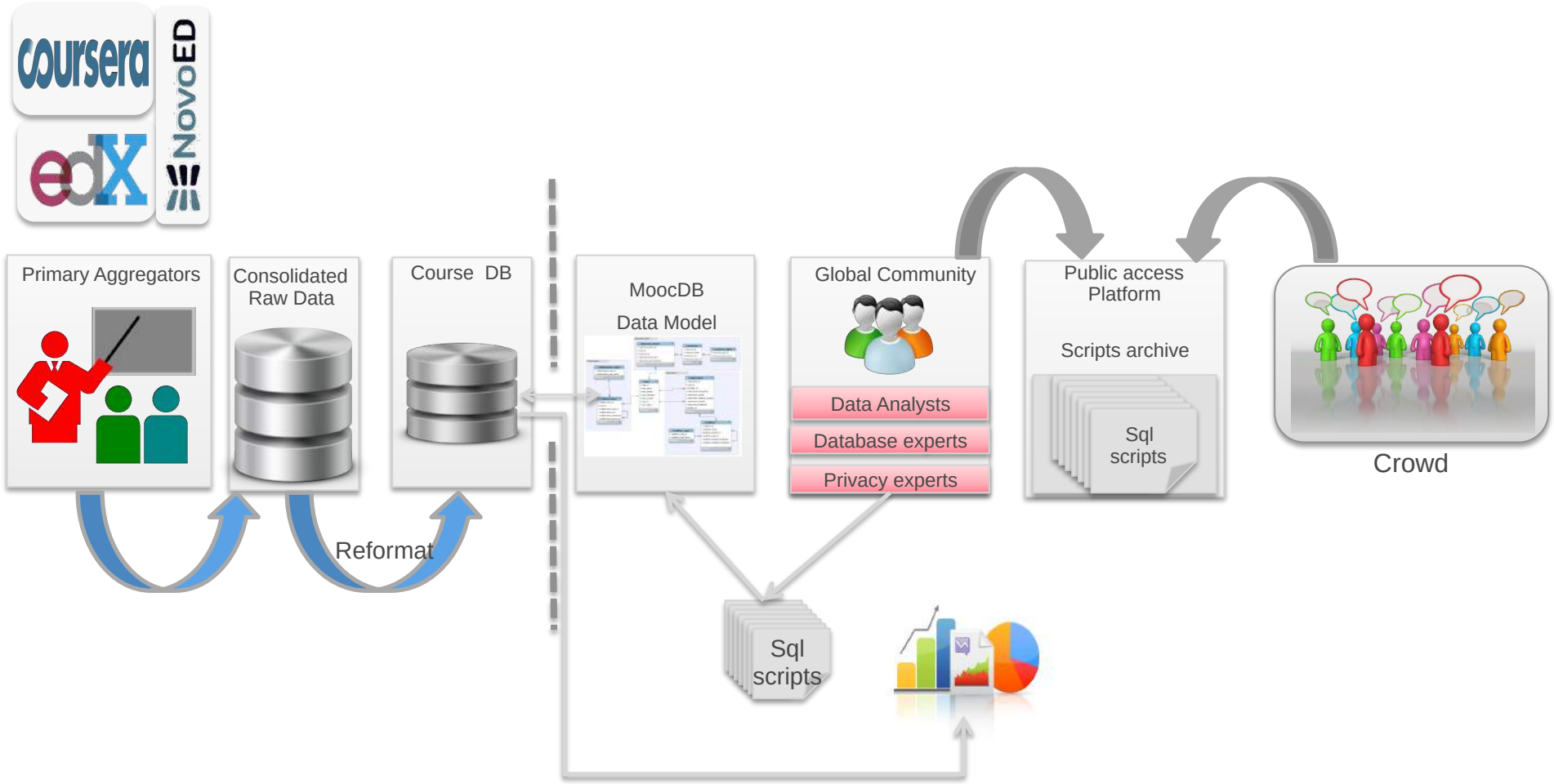


Step 2: their scripts use the schema to reference the data in the course DB

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

MoocDB



Step 3: The script executing over the data, referencing the data model, allows the insights from the course DB to be returned

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

MOOCDB supports multiple frameworks

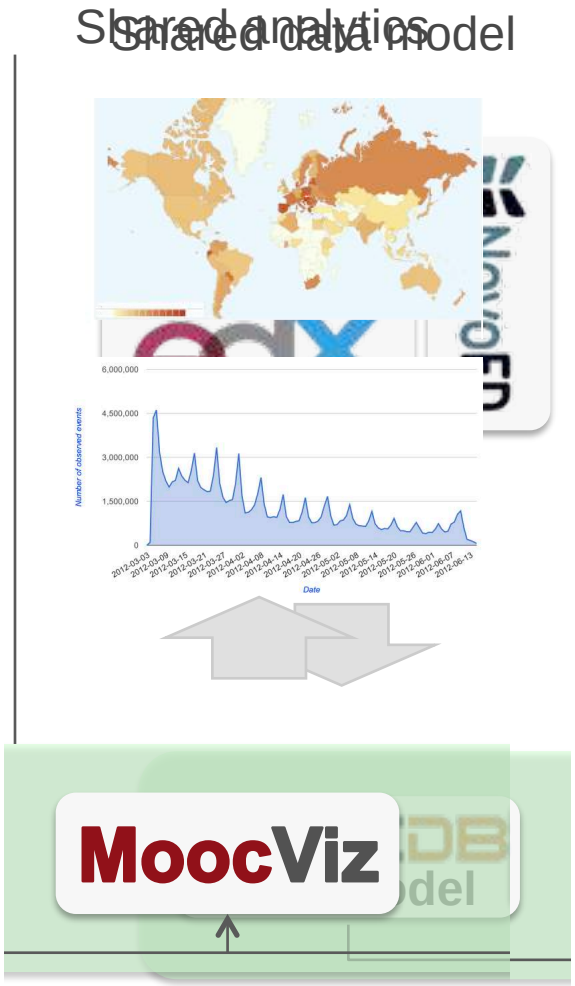
- Our 6.002x DB using MOOCDB model
  - 17 million submission mode events
  - 150M observing mode events
  - 96K collaborative events
  - Collapsed from 60GB to 6 GB

- Multiple Frameworks based on MOOCDB
  1. Export of data from course db
  2. Interoperability with programming languages
  3. Privacy protection via differential privacy
  4. Visualization and analytics -> MOOCVIZ

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

MoocVIZ



ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

MOOCDB and MOOCVIZ

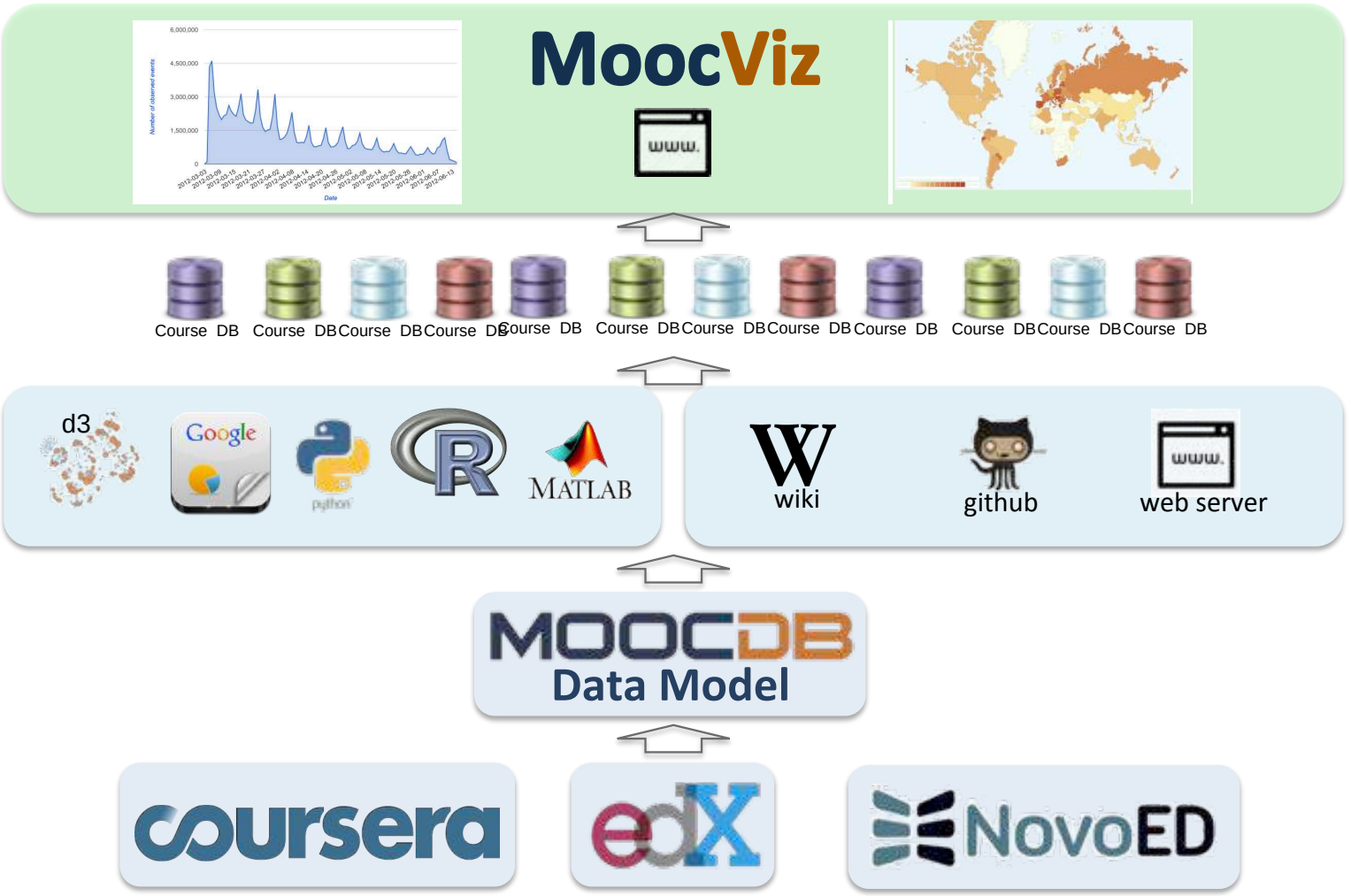
Use MOOCVIZ to demonstrate

- **MOOCDB's collaboration support**
  - **On Stanford and MIT courses**
  - **On 2 different platforms EDX and COURSERA**
- **New analytics around resource usage**
  - **Visualization and statistical support**

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

MoocViz Analytics Platform



ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

MoocVIZ User Types

- Arms-length observers
  - Checking out the website
- Technology-savvy crowd
  - Vote on utility of a visualization
  - Contribute s/w from other domains
- Course instructors
- MOOC providers
- Education researchers

"arm-chair quarterbacks"

Script developers

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

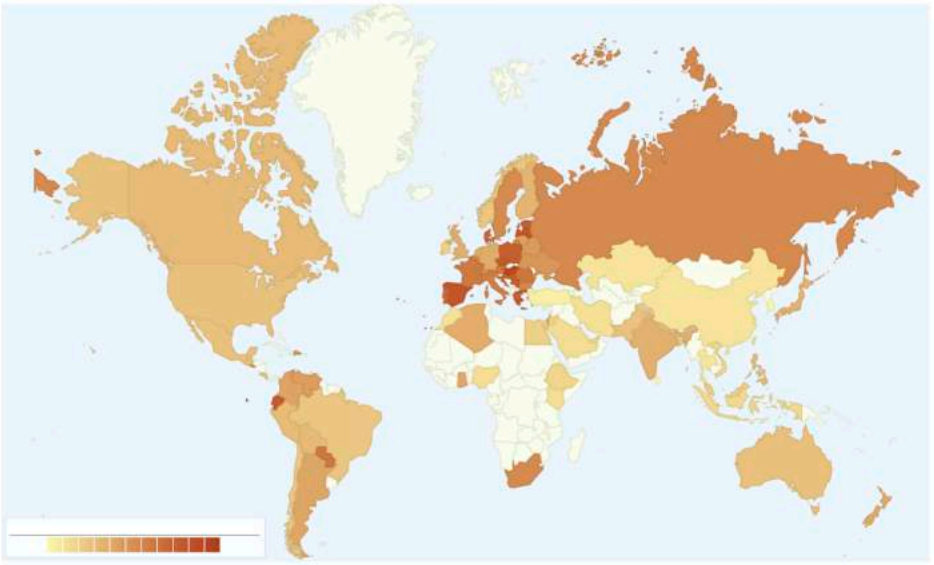
	Coursera course	edX Course
Title	Cryptography I	Circuits and Electronics (6.002x)
Instructors	Dan Boneh	Anant Agarwal, Gerald Sussman, Piotr Mitros
University	Stanford University	MIT
Length	6 weeks	14 weeks
Platform	Coursera	edX
Start date	Jan 13th, 2013	March 5, 2012
Registrants	21,744	154,763

ALFA
ANYSCALE LEARNING FOR ALL

Stanford and MIT visualiations side by side

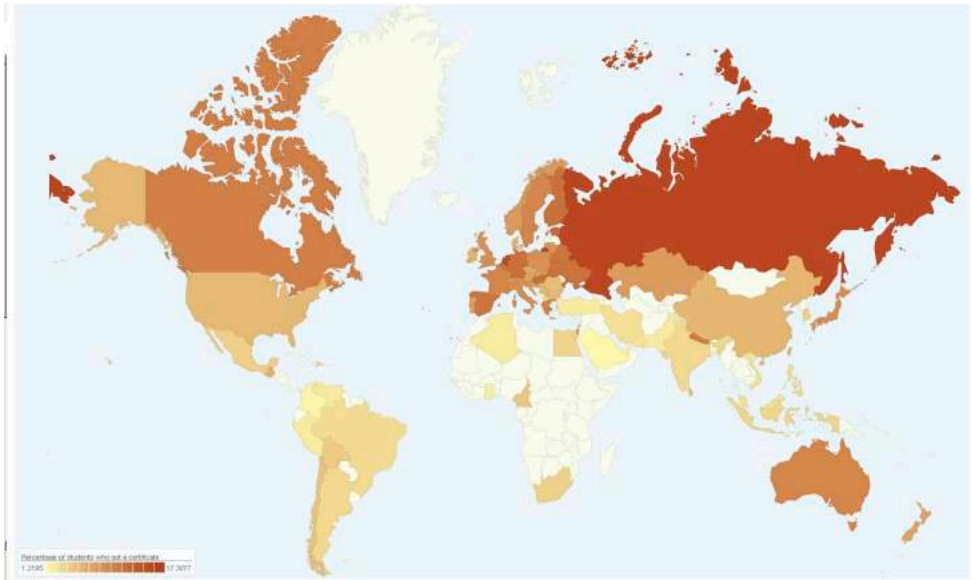
CSAIL

MoocViz



6.002x
user certificates per country,
normalized, cutoff 100

Hungary 16.2%
Spain 14.55%
Latvia 14.40%



Crypto 1-Stanford,
user certificates per country,
normalized, 2 columns 100

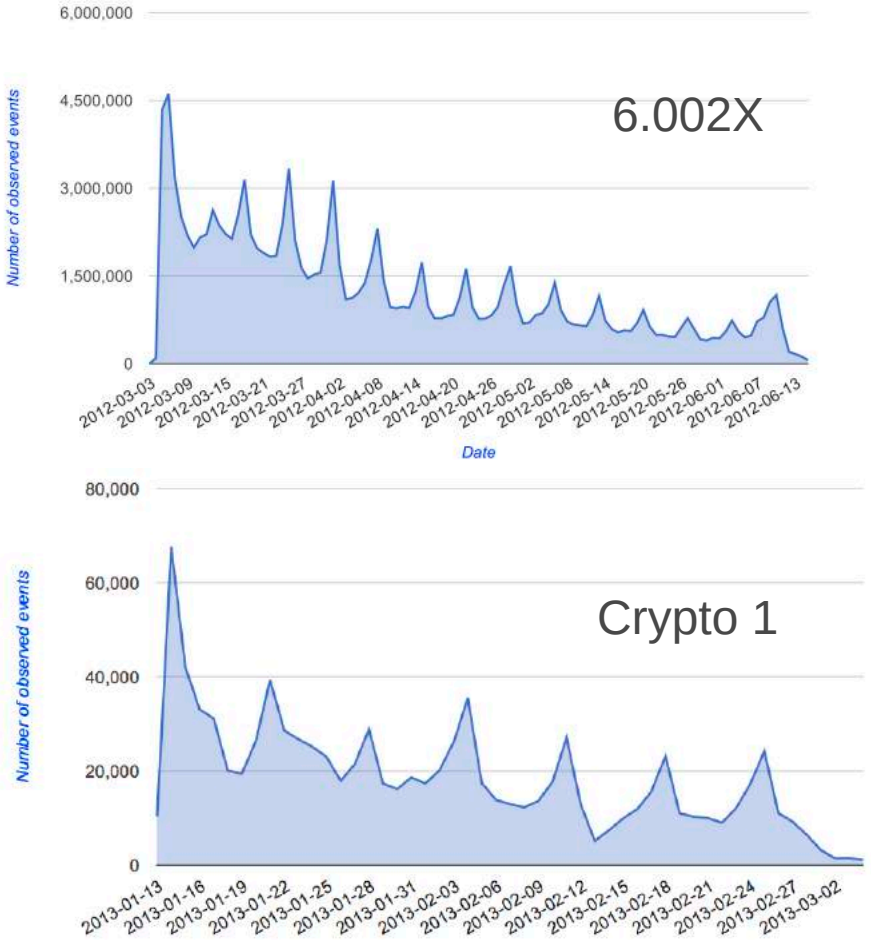
Russia 17.4%
Netherlands 16.43%
Germany 12.95%

A comparison using the classic world map graphic

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

MoocViz



6..002x

ANYSCALE LEARNING FOR ALL

The “Classic” Events cross Time Graphic

CSAIL

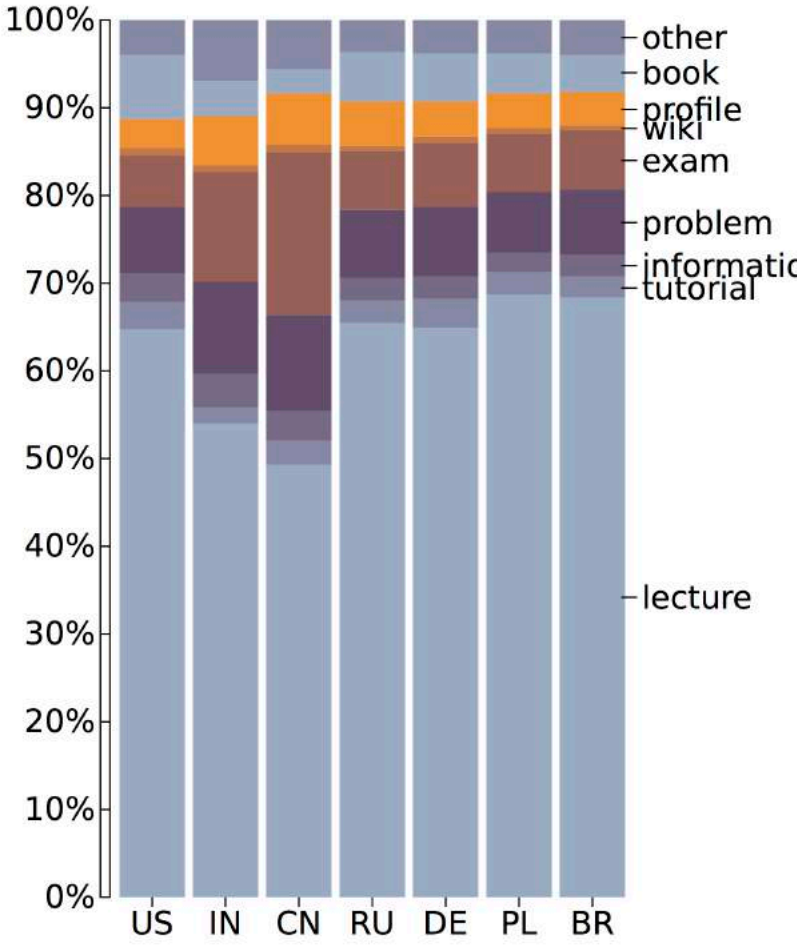
Resource Types

Resource id	Content	Medium
1	Lecture	Text
2	Lecture	Video
3	Tutorial	Text
4	Tutorial	Video
5	Informational	Any
6	Problems	Any
7	Wiki	Any
8	Forum	Any
9	Profile	Any
10	Index	Any
11	Book	Any
12	Survey	Any
13	Home	Any
14	Other	Any

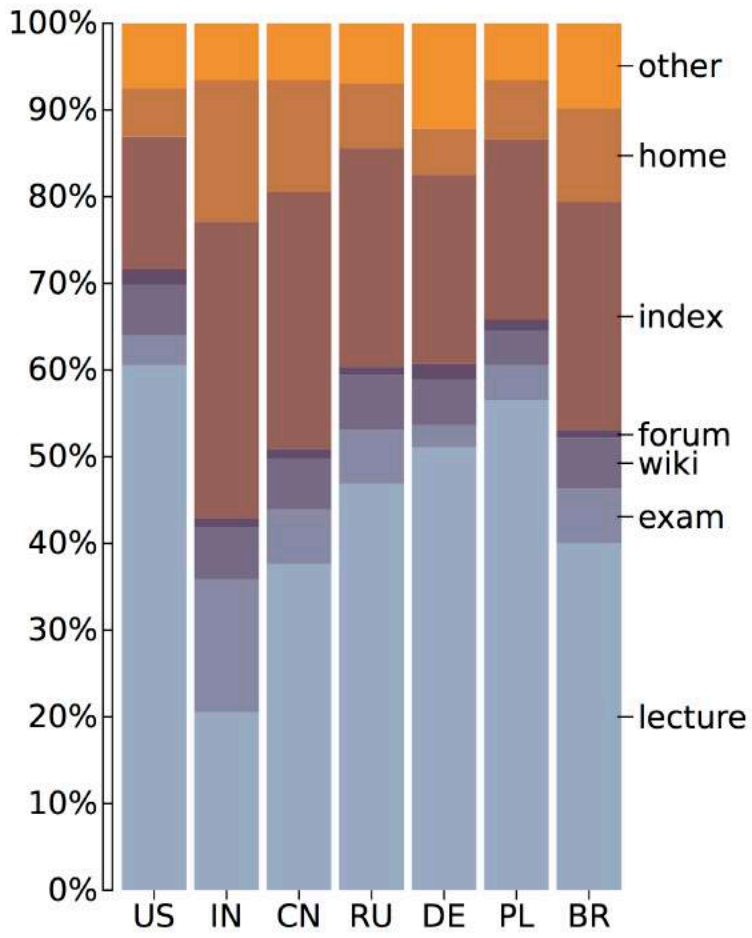
ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Studying Resource Use



6.002X



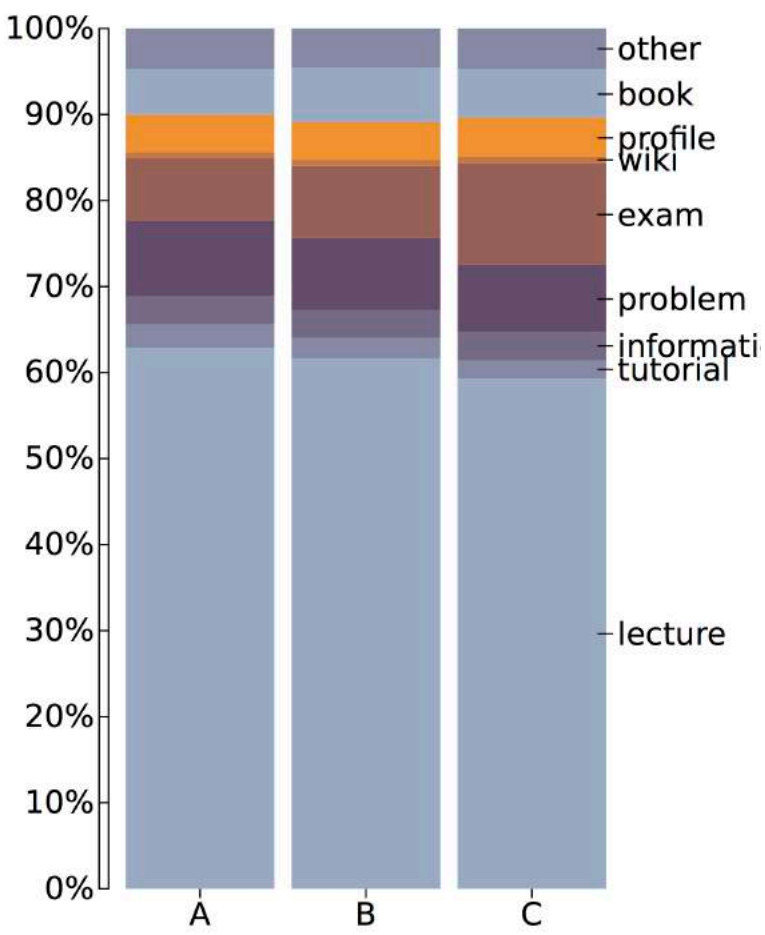
Crypto 1

ALFA
ANYSCALE LEARNING FOR ALL

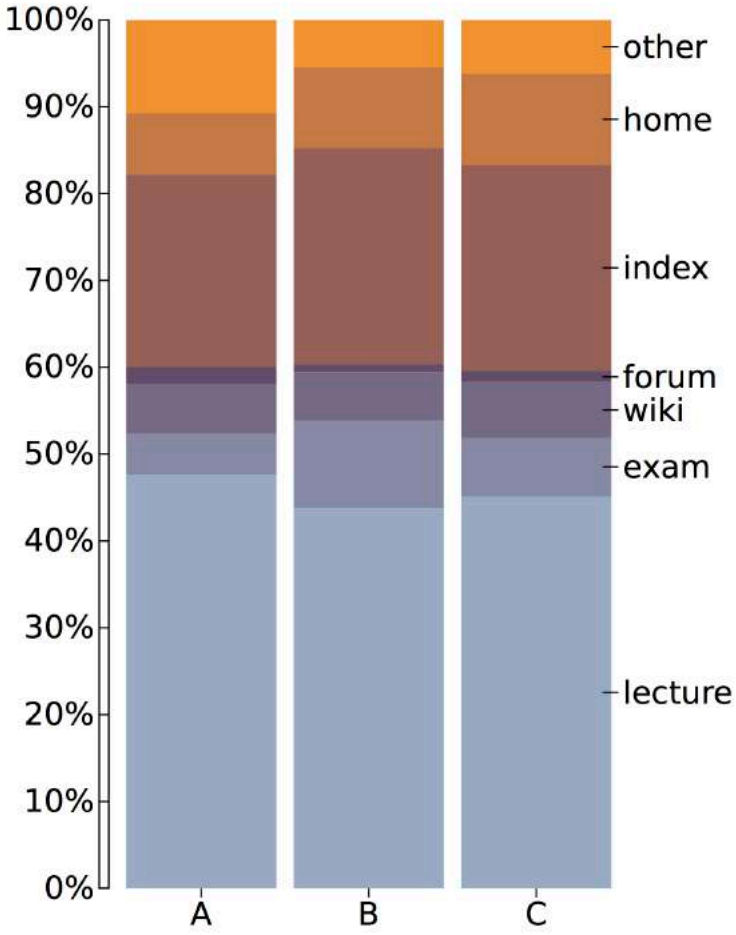
Resource Use Compared by Country

CSAIL

Studying Resource Use



6.002x



Crypto 1

Resource Use Compared by Grade

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Analytics: Statistical Comparisons

	Lecture		*Exam*		*Problems*		*Book*	
	l	u	l	u	l	u	l	u
BR *vs.*DE	-35405	138970	-2563.1	9993.1	-7198.3	12247	-37926	13293
BR *vs.*IN	165720	287580	-1668.5	7106.2	2319.2	15908	-8910.5	26882
BR *vs.*RU	-60298	94243	-1072.3	10056	-5549.5	11684	-29058	16336
BR *vs.*US	-113120	4533	-5271.2	3201	-16114	-2993.5	-53044	-18485
DE *vs.*IN	102530	24721	-6205	4212.7	-1477.3	14656	55.123	42550
DE *vs.*RU	-121360	51742	-5455	7009.4	-9108.7	10195	-19468	31378
DE *vs.*US	-176650	-35499	-9832.1	332.03	-19948	-4207.7	-44179	-2717.7
IN *vs.*RU	-269690	-149660	-2548.6	6094.8	-12739	646.57	-32976	2281.4
IN *vs.*US	-313990	-247900	-6133.3	-1374.5	-22352	-14982	-54456	-35045
RU *vs.*US	-129150	-13383	-9695.1	-1358.9	-19076	-6166.2	-46405	-12401

Table 2: Analysis of the duration spent on resource type by country-based student cohorts for the edX course. The value of each cell is the 95% confidence interval given by the Tukey-Kramer method for the true difference of the means of the duration for the two cohorts indicated in the first column, e.g. students BR and DE in the first row. Cells with colored background indicate that the two cohorts have a significant difference in terms of mean of duration, which corresponds to the interval encompassing 0.

Analytics – Statistical Comparisons

	Lecture		*Exam*	
	l	u	l	u
BR *vs.*DE	-28253	-1117.2	-1024.1	592.15
BR *vs.*IN	-461.58	21961	-1335.2	0.2361
BR *vs.*RU	-20329	3236.1	-758.69	644.88
BR *vs.*US	-24271	-2475.1	-685.78	612.37]
DE *vs.*IN	15939	34930	-1017.1	114
DE *vs.*RU	-4025.8	16302	-446.33	764.43
DE *vs.*US	-7811.2	10435	-364.14	722.64
IN *vs.*RU	-25998	-12595	211.44	1009.8
IN *vs.*US	-29106	-19138	333.96	927.65
RU *vs.*US	-10989	1337	-346.88	387.29

Crypto 1: Comparison of country based cohorts

Analystics – Statistical Comparisons

	Lecture		Exam		Problems		Book	
	l	u	l	u	l	u	l	u
A *vs.* B	-16144	22245	-1346	1117.5	-3689.6	315.26	-13797	-3978.9
A *vs.* C	64383	119560	3935.6	7476.6	8216	13972	-2728.8	11383
B *vs.* C	59184	118660	3912	7728.6	9679	15884	5609.6	20820

Table 4: Analysis of the duration spent on resource type by grade-based student cohorts for the edX course. See Table 2's caption for the explanation on how to read the table.

MoocDB Resources

Publications: groups.csail.mit.edu/EVO-DesignOpt/groupWebSite/index.php?n=Site.Publications

- MOOCdb: Developing Data Standards for MOOC Data Science

  Kalyan Veeramachaneni, Franck Dernoncourt, Colin Taylor, Zachary A. Pardos, Una-May O'Reilly,

  MOOCShop at Artificial Intelligence in Education, 2013

Other Resources

- Wiki site documenting data model
  - will be perpetually updated
  - **http://moocdb.csail.mit.edu/wiki.**
- Web-based software repository (not yet public)
  - https://github.com/ organizations/MOOCdb

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

MoocViz Resources

- MoocVIZ: A Large Scale, Open Access, Collaborative Analytics Platform for MOOCs
  - Dernoncourt, O'Reilly, Veeramachaneni, S. Wu, C. Do, S. Halawa
  - NIPS 2013 Workshop on Data Directed Education
- Wiki site documenting MoocDB data model
  - will be perpetually updated
  - http://moocdb.csail.mit.edu/wiki.
- Web-based software repository (not yet public)
  - https://github.com/ organizations/MOOCdb
  - R, Python, Matlab
- Web server
  - Local and community versions
  - Visualizations are described in html

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Future MoocDB and MOOCViz work

- MoocDB Scaling Up
  - Move from grass roots, bottom one step up to institutions
    - Legacy course data
    - EDX, Coursera and Kahn Academy joining
- MoocVIZ: Visualization building
- Leveraging MoocDB for ALFA research
  - Crowd sourcing
  - Tiger team research with fielding
  - Problem response behavior
  - Understanding MOOC attrition
  - Studying social interactions

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Acknowledgments

ALFA Mooc Data Science Team

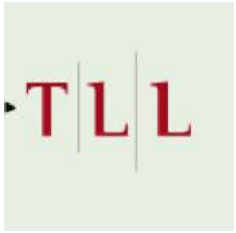
- Kalyan Veeramachaneni (Lead)
- Franck Dernoncourt
- Elaine Han
- Colin Taylor
- Sherwin Wu
- Kristin Asmus
- John O'Sullivan
- Will Grathwohl
- Josep Mingot

ALFA
ANYSCALE LEARNING FOR ALL

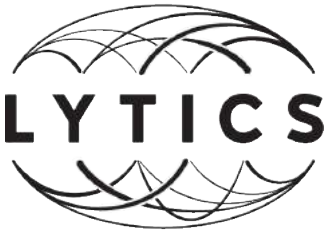
CSAIL

Partners

Lori Breslow
Jennifer Deboer
Glenda Stump



Sherif Halawa
Andreas Paepcke
Rene Kizilcec
Emily Schneider



Piotr Mitros
James Tauber

Chuong Do

Sponsors

- …Mooc Research Initiative
- …Quanta Research