MLBlocks

Towards building machine learning blocks and predictive modeling for MOOC learner data

Kalyan Veeramachaneni

Joint work with

Una-May O'Reilly, Colin Taylor, Elaine Han, Quentin Agren, Franck Dernoncourt, Sherif Halawa, Sebastien Boyer, Max Kanter

Any Scale Learning for All Group
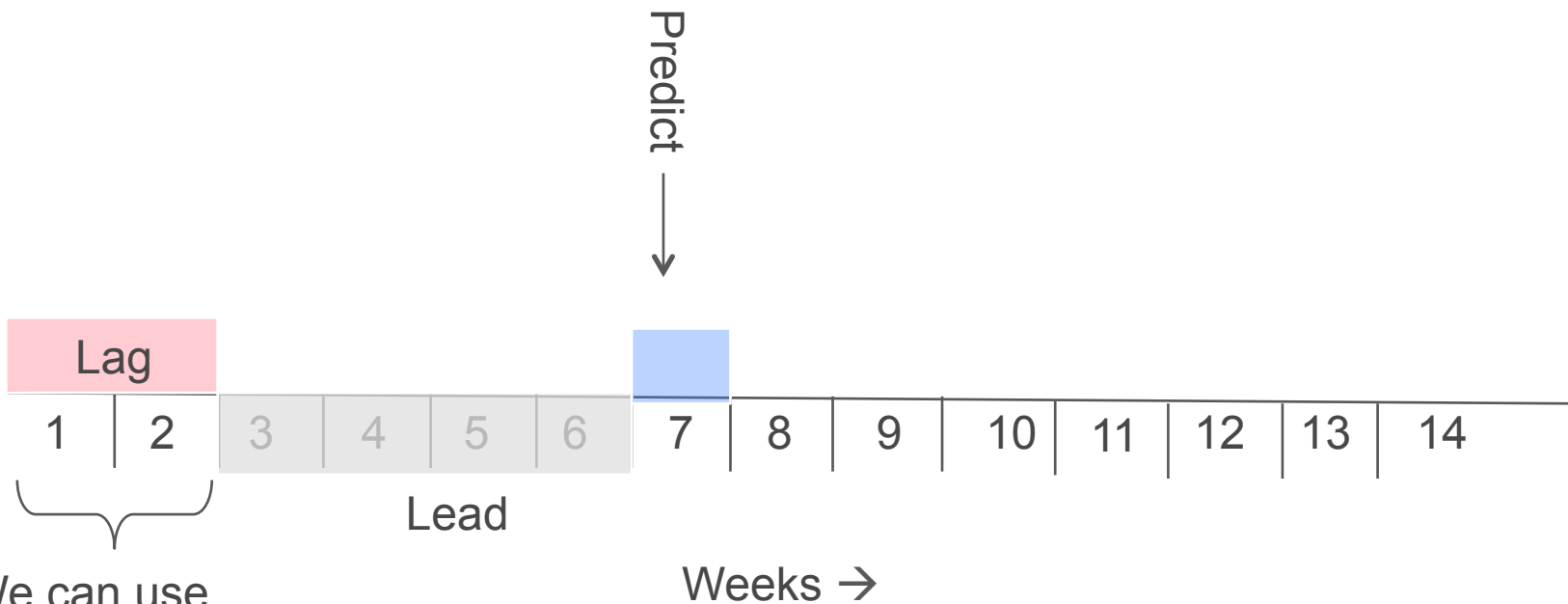**CSAIL, MIT**

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Suppose…

Given learners interactions up until a time point, we want to predict if s/he will dropout/stopout in the future?

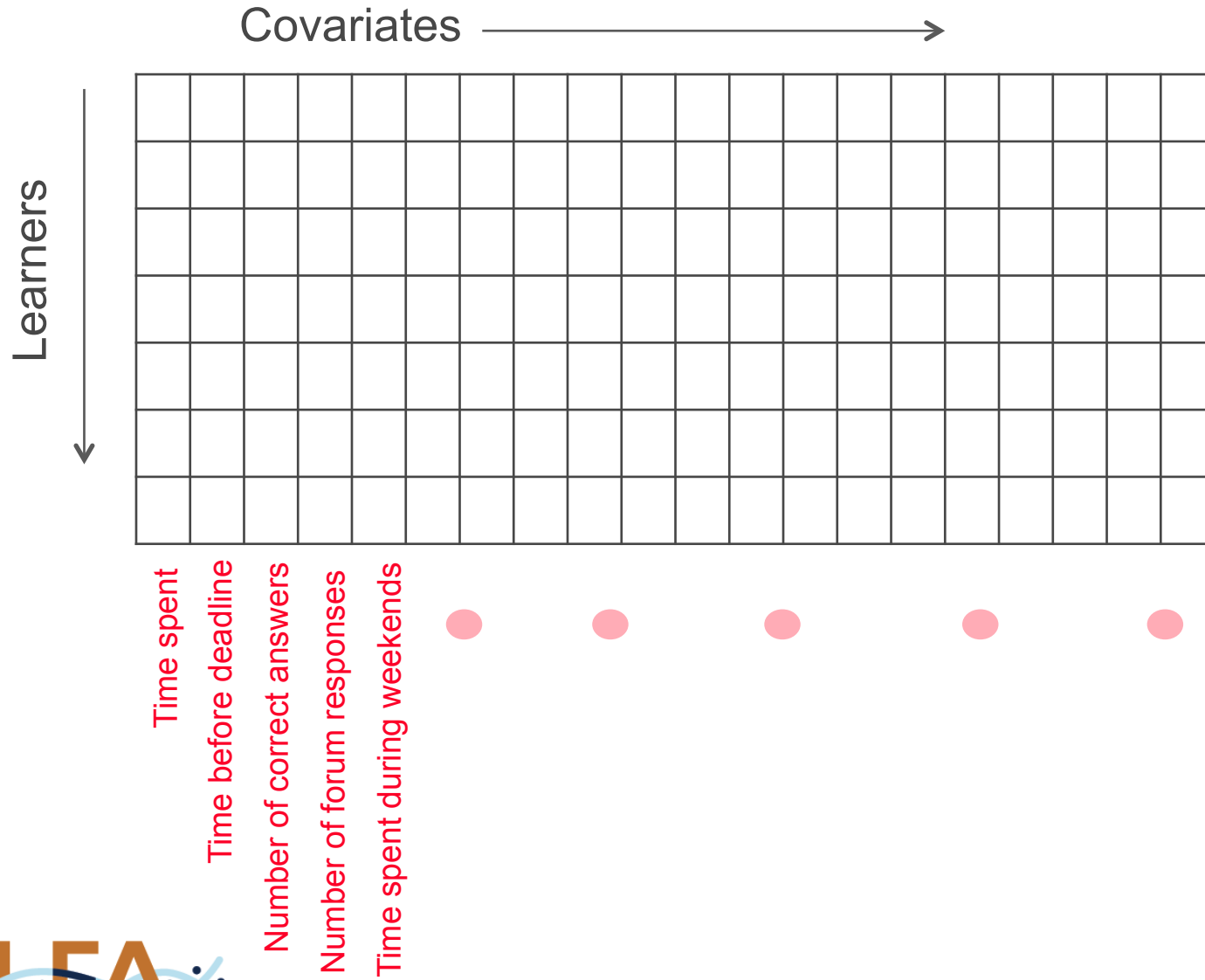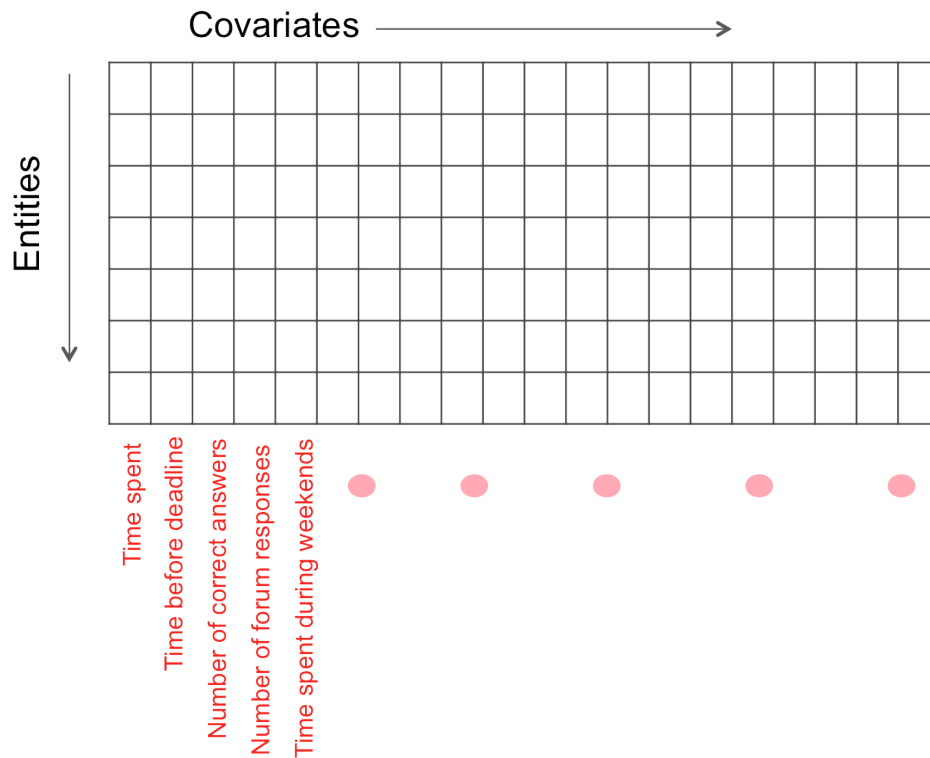- We must use *click stream*, *forums* as well *assessments*

Predict
↓

Lag													
1	2	3	4	5	6	7	8	9	10	11	12	13	14

Lead

We can use students data during these weeks

Weeks →

Note: By varying lead and lag we get 91 prediction problems

The Quintessential Matrix

Covariates →

Learners ↓

Time spent
Time before deadline
Number of correct answers
Number of forum responses
Time spent during weekends

What can we do with that matrix ?

Covariates →

Entities ↓

Time spent
Time before deadline
Number of correct answers
Number of forum responses
Time spent during weekends

Cluster/segment

Lurkers,
high achievers,
interactive

Predict an outcome

Who is likely to dropout?

Analytics

Did this video help?
Correlation with
performance

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

What can we do with that matrix ?

Covariates →

Entities ↓

Time spent
Time before deadline
Number of correct answers
Number of forum responses
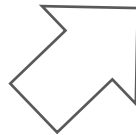Time spent during weekends
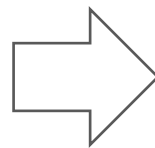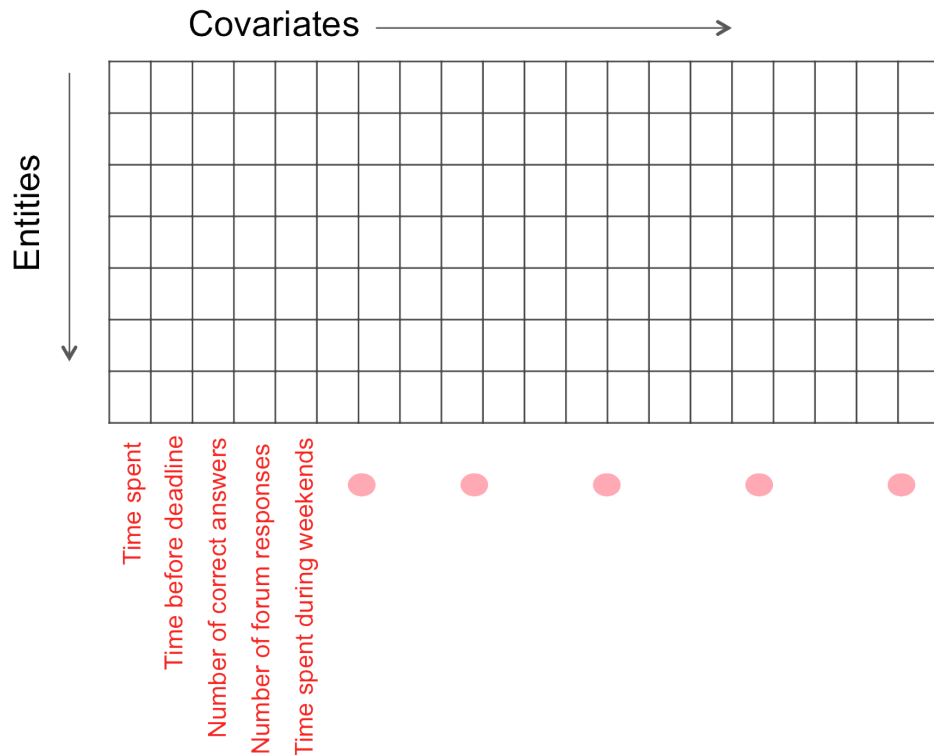
Cluster/segment

Lurkers,
high achievers,
interactive

Predict an outcome

Who is likely to dropout?

Analytics

Did this video help?
Correlation with
performance

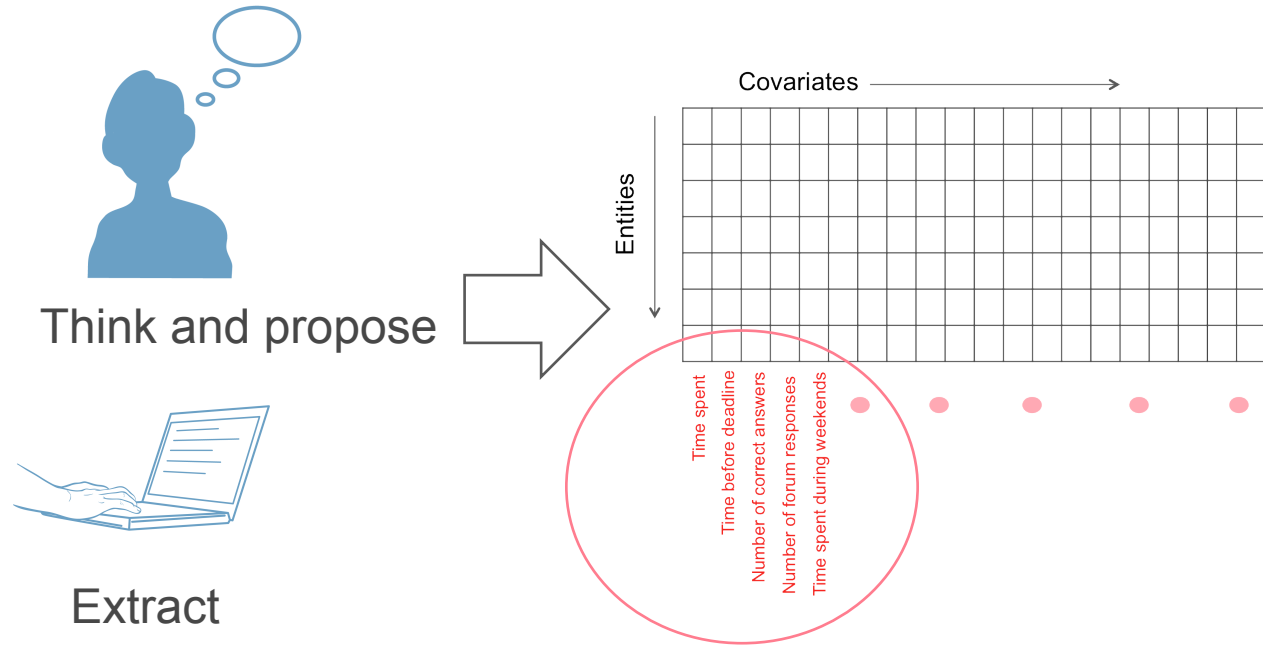Supervised learning machinery
Neural networks, SVMs, Random Forests
Unsupervised learning machinery
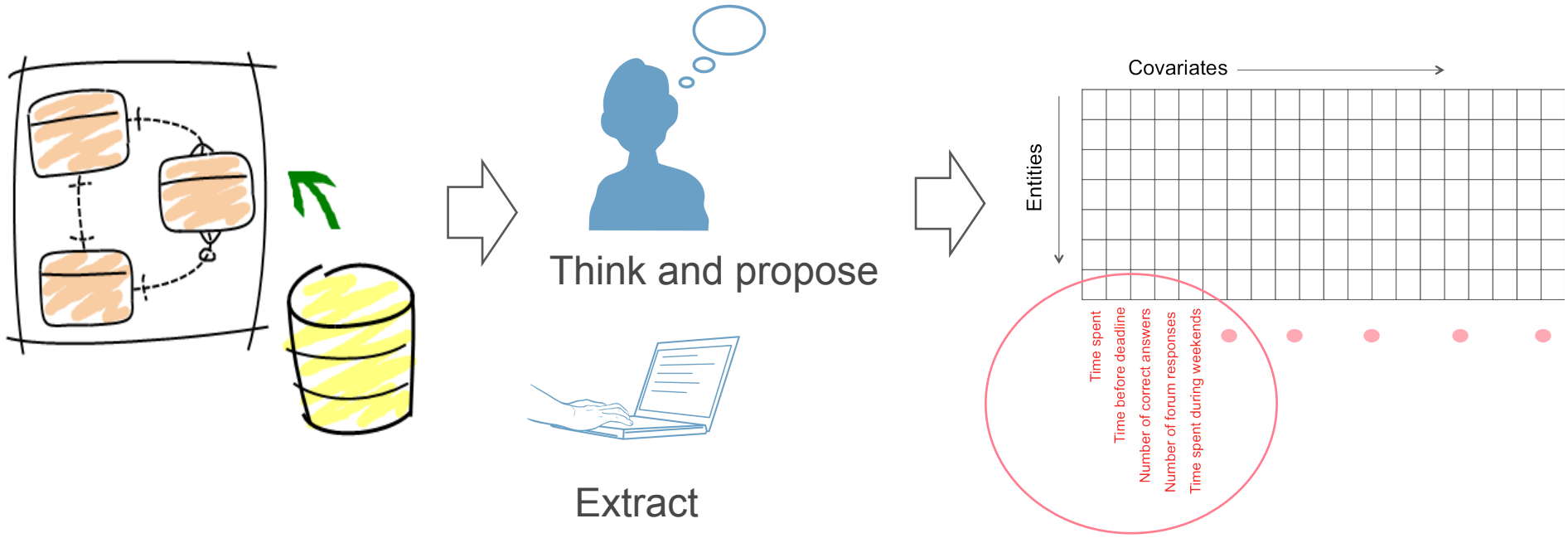Gaussian mixture models, Bayesian clustering
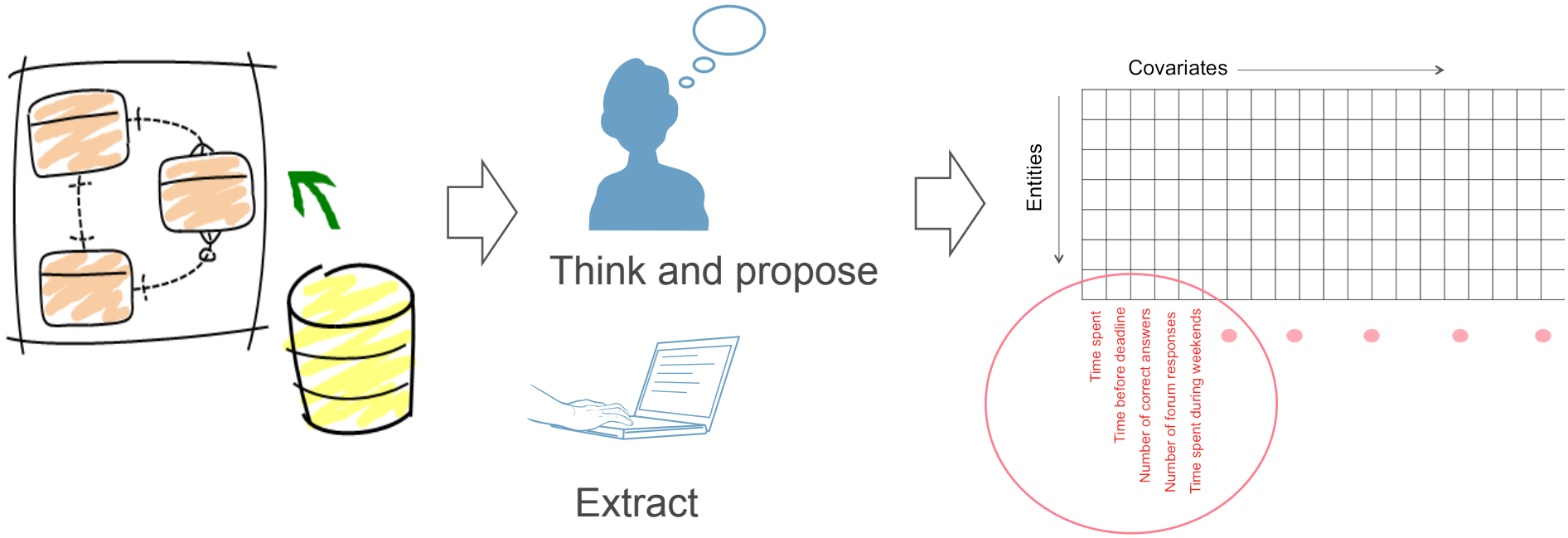Probabilistic modeling
Graphical models, HMMs

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

But…. How did the matrix come about?

But….How did the matrix come about?

Think and propose

Extract

Curation of
raw data

Variable engineering

Covariates

Entities

Time spent
Time before deadline
Number of correct answers
Number of forum responses
Time spent during weekends

But….How did the matrix come about?

Think and propose

Extract

Covariates

Entities

Time spent
Time before deadline
Number of correct answers
Number of forum responses
Time spent during weekends

Curation

Variable engineering

Machine learning

How do we shrink this?

Think and propose

Extract

Curation

Variable engineering

>6 months

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

How did the matrix come about?

Think and propose

Extract

Covariates

Entities

Time spent
Time before deadline
Number of correct answers
Number of forum responses
Time spent during weekends
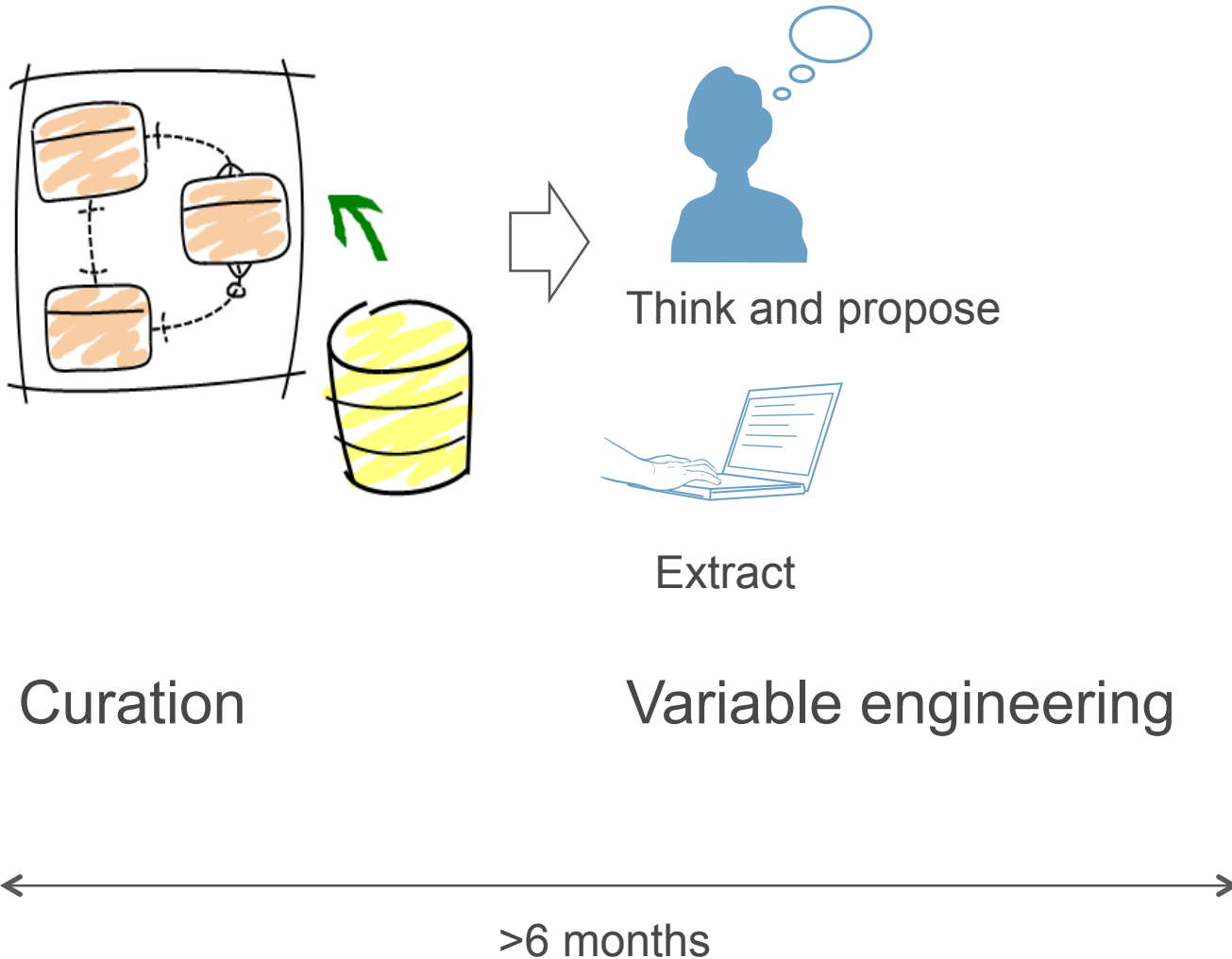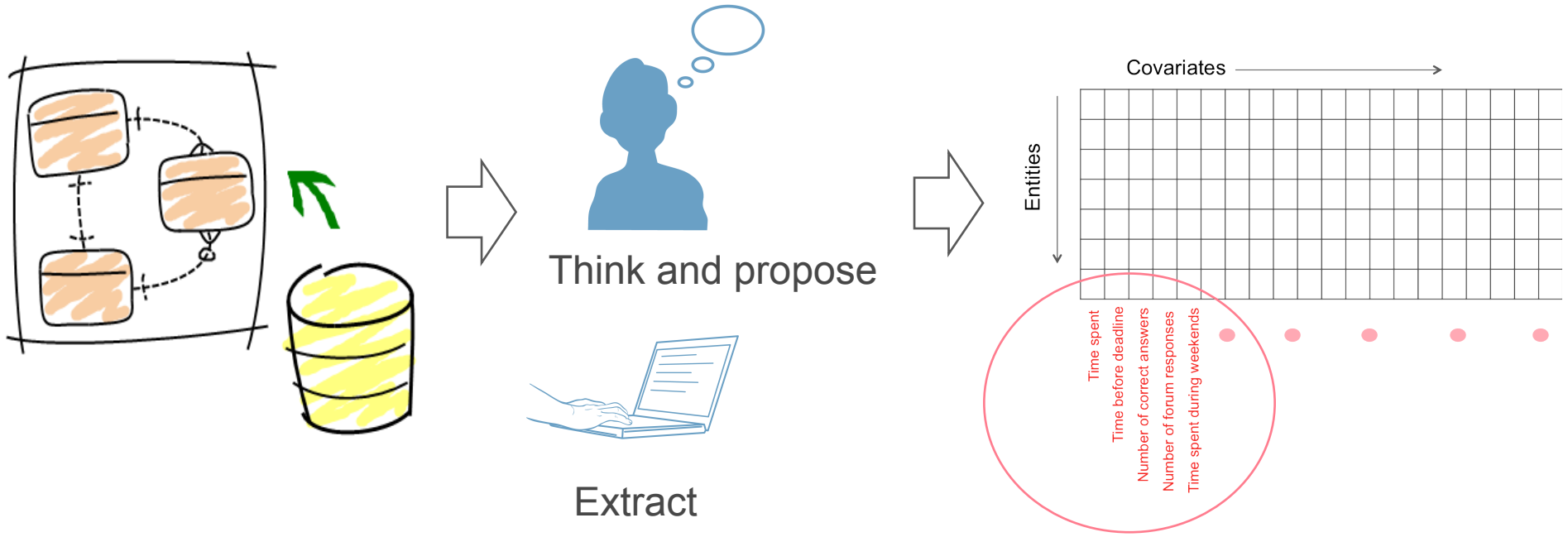
Curation

Variable engineering

Machine learning

> 6 months
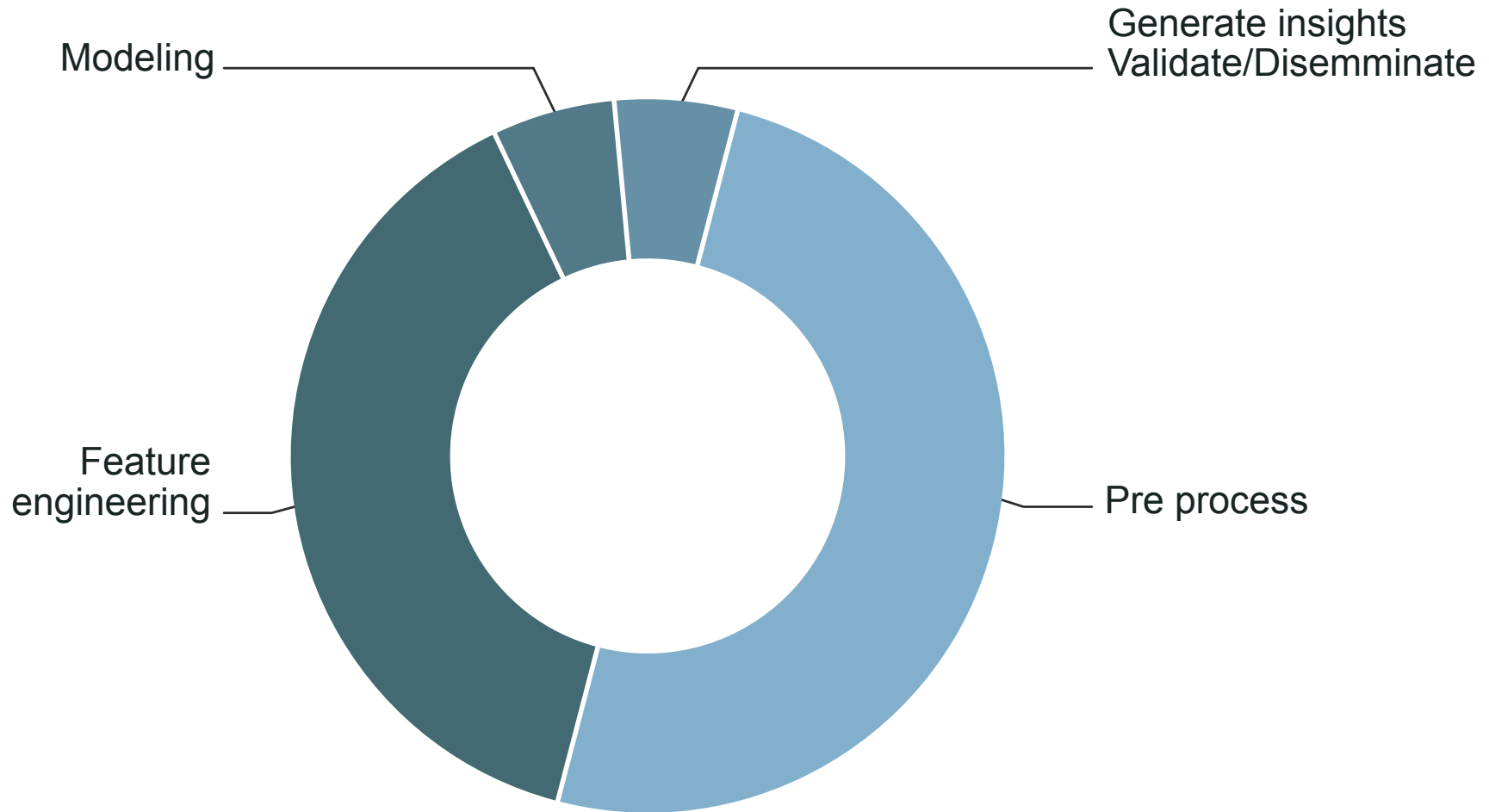
a week

The Overarching theme of my research

- **How can we reduce time to process, analyze, and derive insights from the data?**

How to shrink this time?

- Build fundamental building blocks for reuse
- Understand how folks in a certain domain interact with the data
    - make this interaction more efficient
- Increase the pool of folks who can work with the data
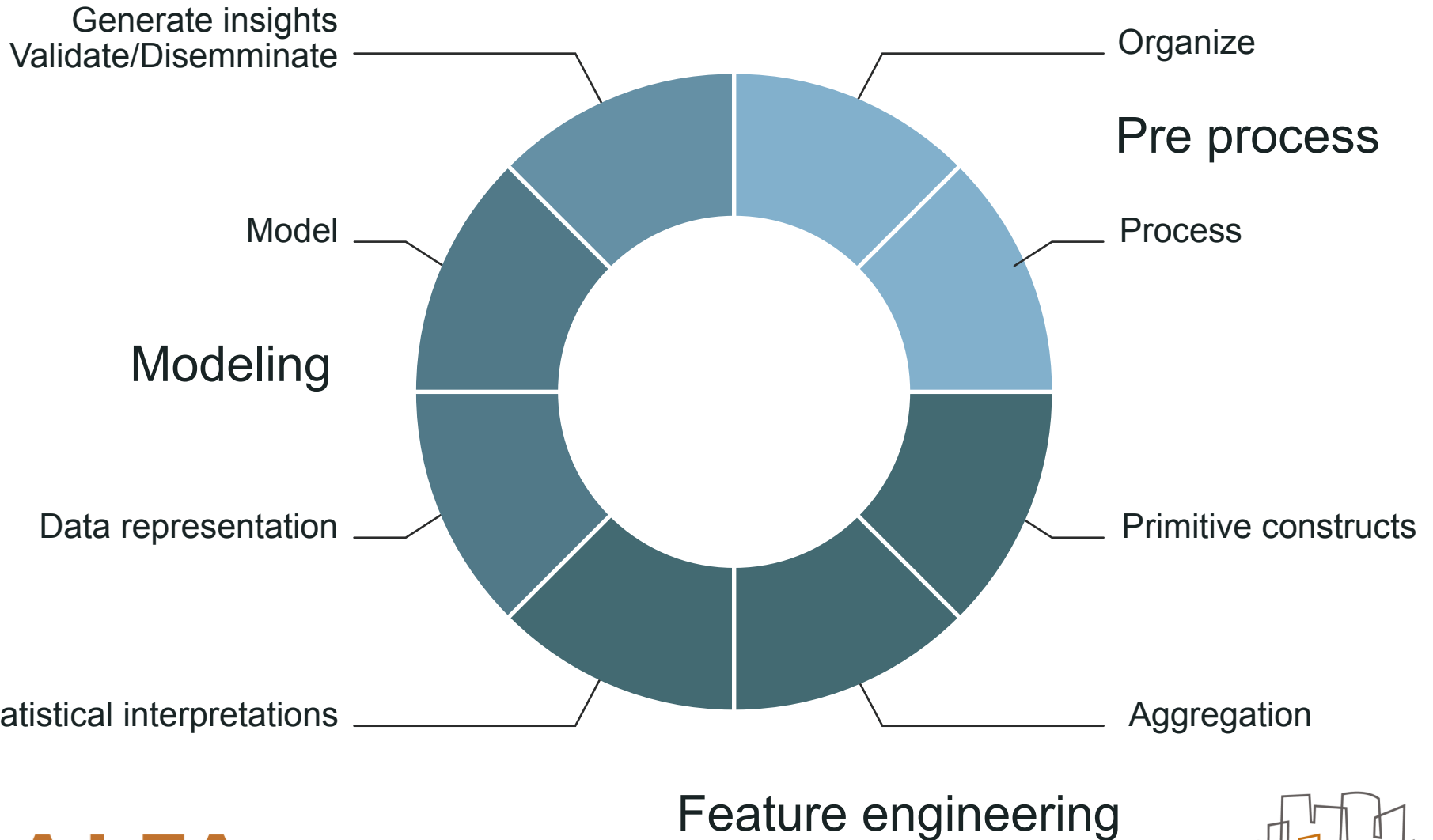
So what are MLBlocks?
Size of the arc corresponds to time spent

Generate insights
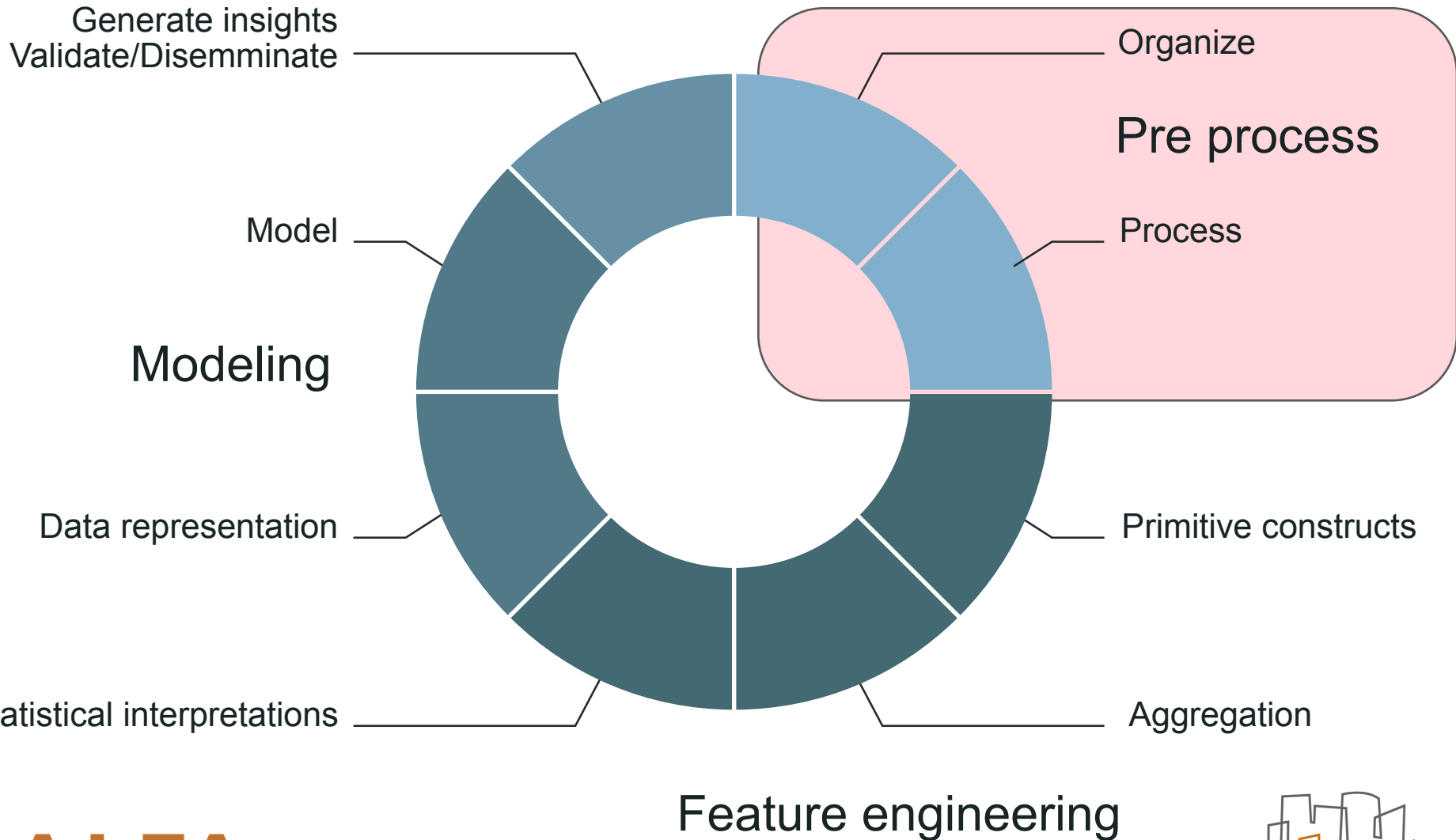Validate/Disemminate

Modeling

Feature
engineering

Pre process

A typical ML process

So what are MLBlocks?
Detailed breakdown

Generate insights
Validate/Disemminate

Organize

Pre process

Model

Process

Modeling

Data representation

Primitive constructs

Statistical interpretations

Aggregation

Feature engineering

So what are MLBlocks?
Detailed breakdown

Generate insights
Validate/Disemminate

Model

Modeling

Data representation

Statistical interpretations

Organize

Pre process

Process

Primitive constructs
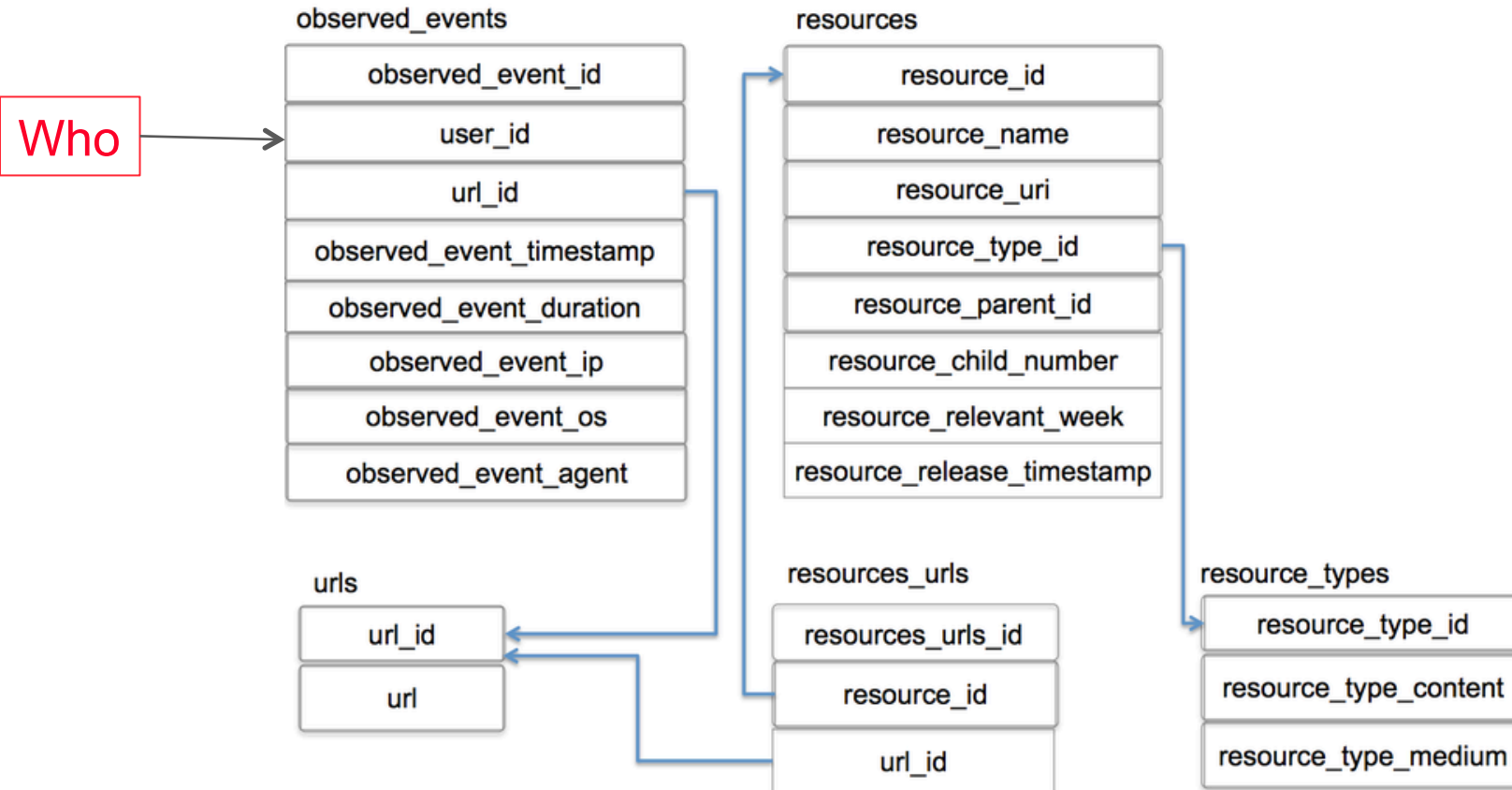
Aggregation

Feature engineering

What we would like to capture and store?

- **Who, When, What Where ?**   Organize

What we would like to capture and store?

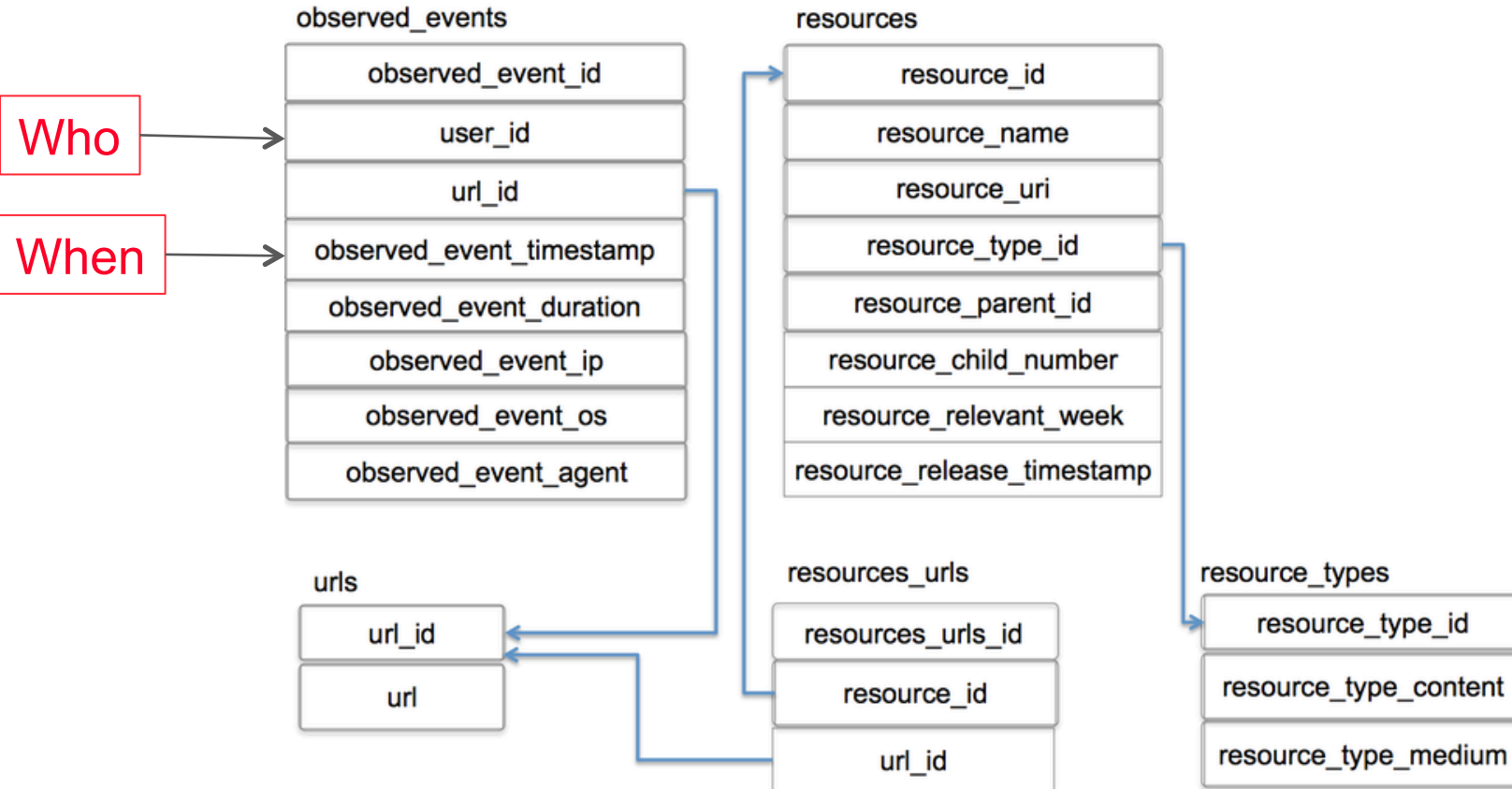- **Who**, When, What Where ?   Organize

What we would like to capture and store?
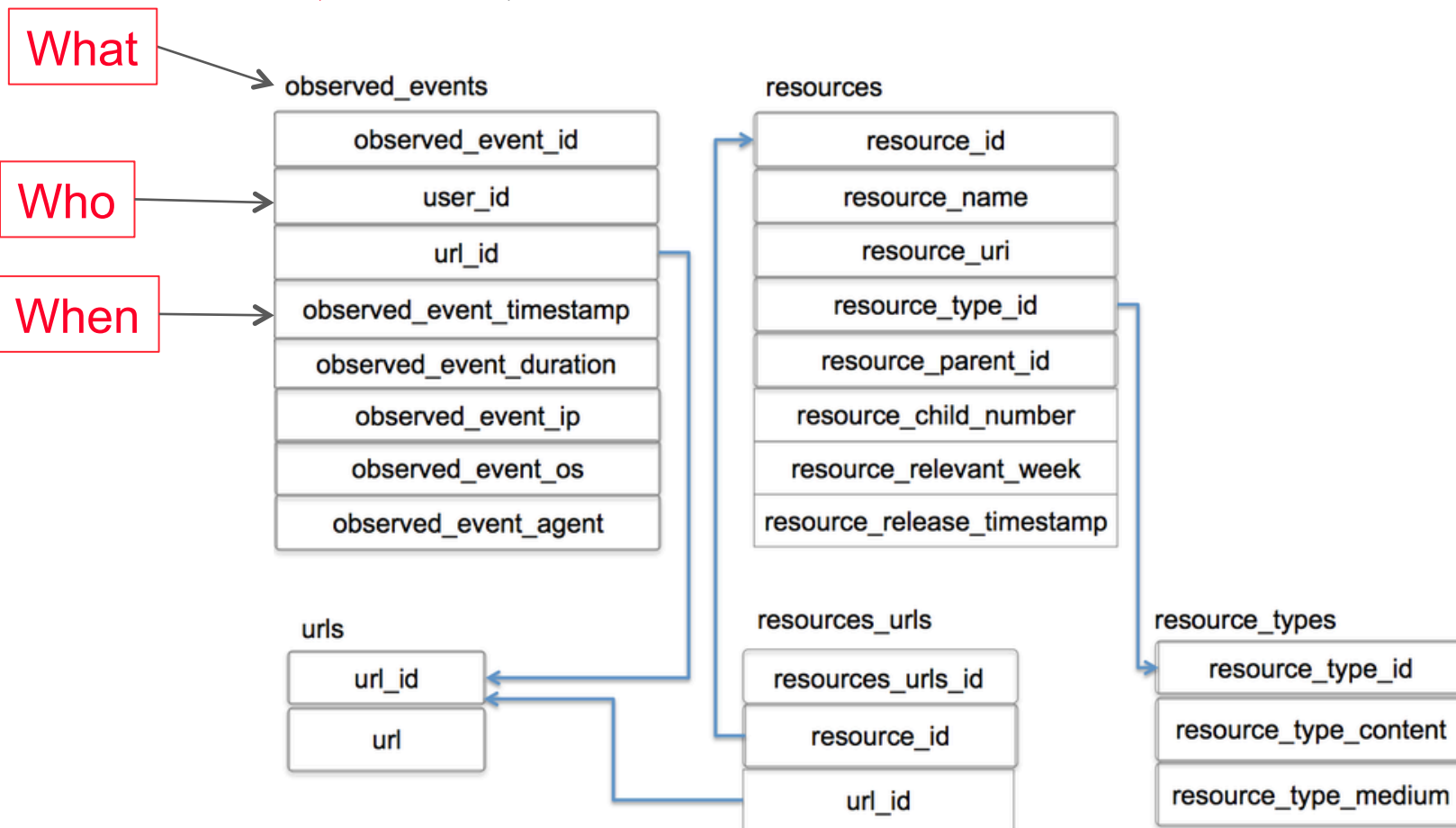
- **Who, When**, What Where ?   Organize

What we would like to capture and store?
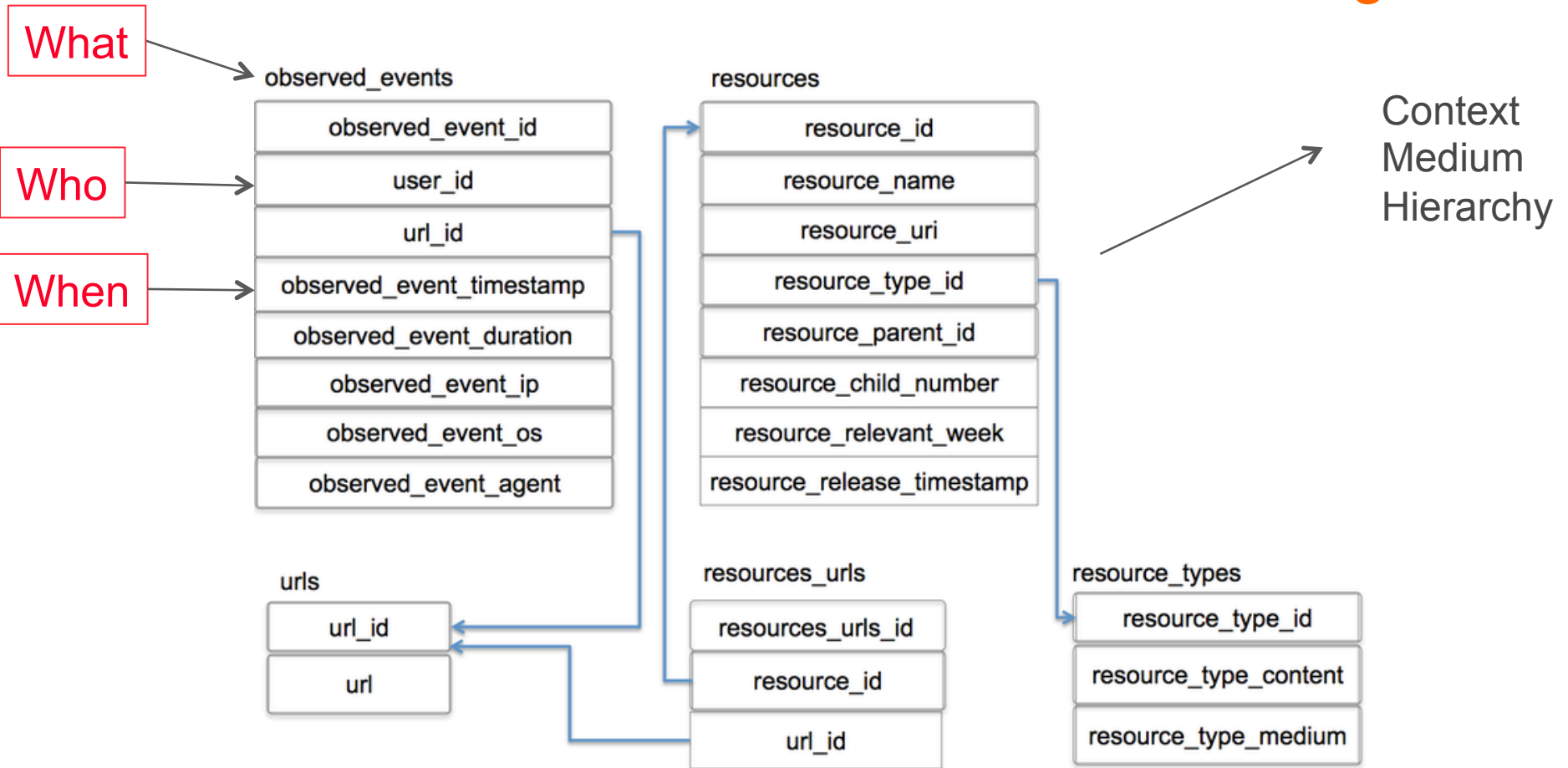
- **Who, When, What** Where ?   Organize

What we would like to capture and store?
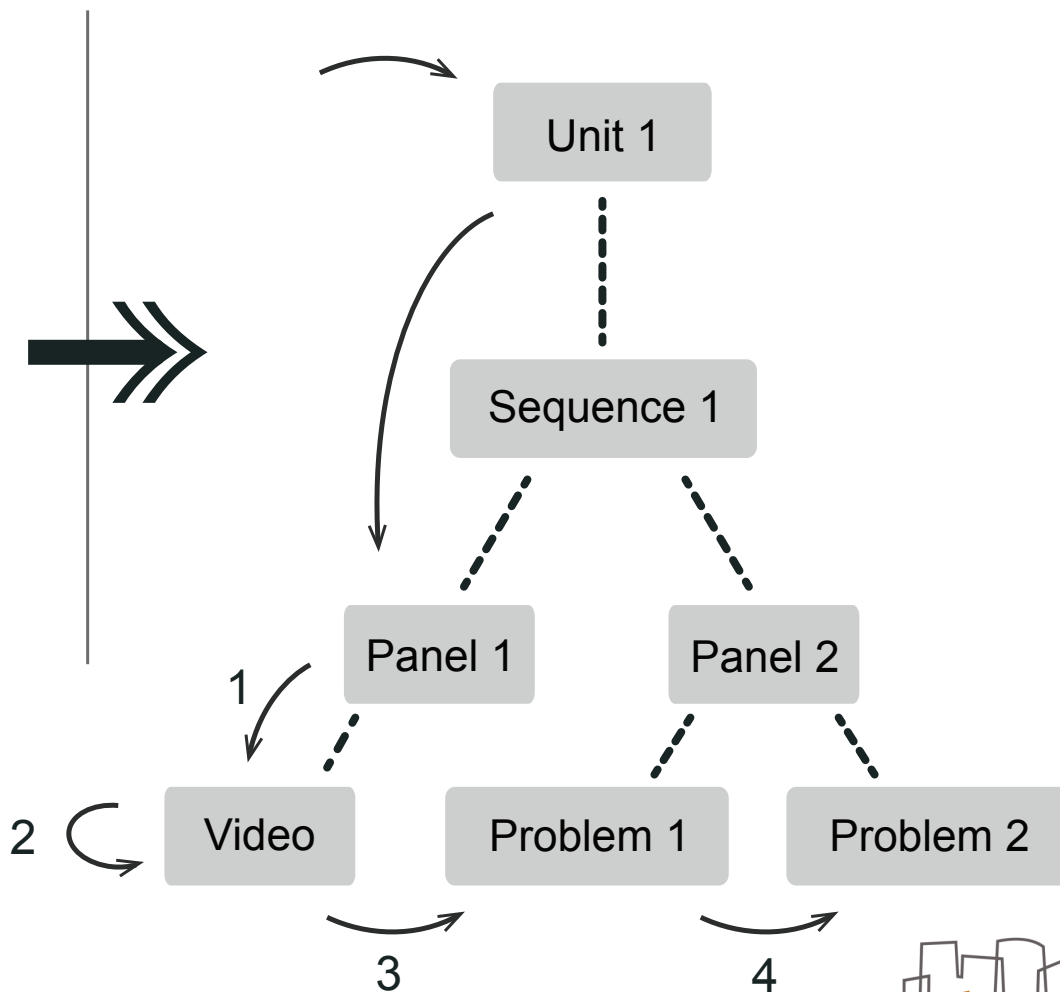
- **Who, When, What, Where ?**   Organize

What

Who

When

Context
Medium
Hierarchy

observed_events
- observed_event_id
- user_id
- url_id
- observed_event_timestamp
- observed_event_duration
- observed_event_ip
- observed_event_os
- observed_event_agent

resources
- resource_id
- resource_name
- resource_uri
- resource_type_id
- resource_parent_id
- resource_child_number
- resource_relevant_week
- resource_release_timestamp

urls
- url_id
- url

resources_urls
- resources_urls_id
- resource_id
- url_id

resource_types
- resource_type_id
- resource_type_content
- resource_type_medium

Organize: Constructing deeper hierarchies
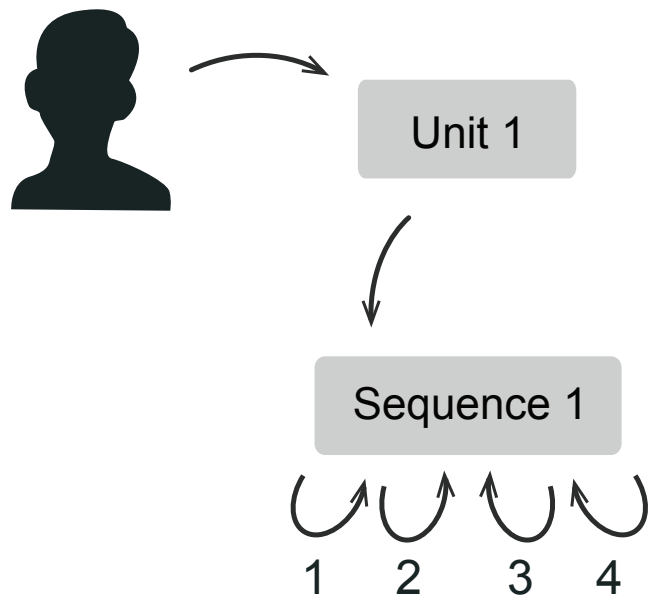
Organize: Contextualizing an event

Interation event

problem_check

URL

http://a/b/c/

Module URI

i4x://problem/123

http://a/b/c/problem/123

Constructe URI
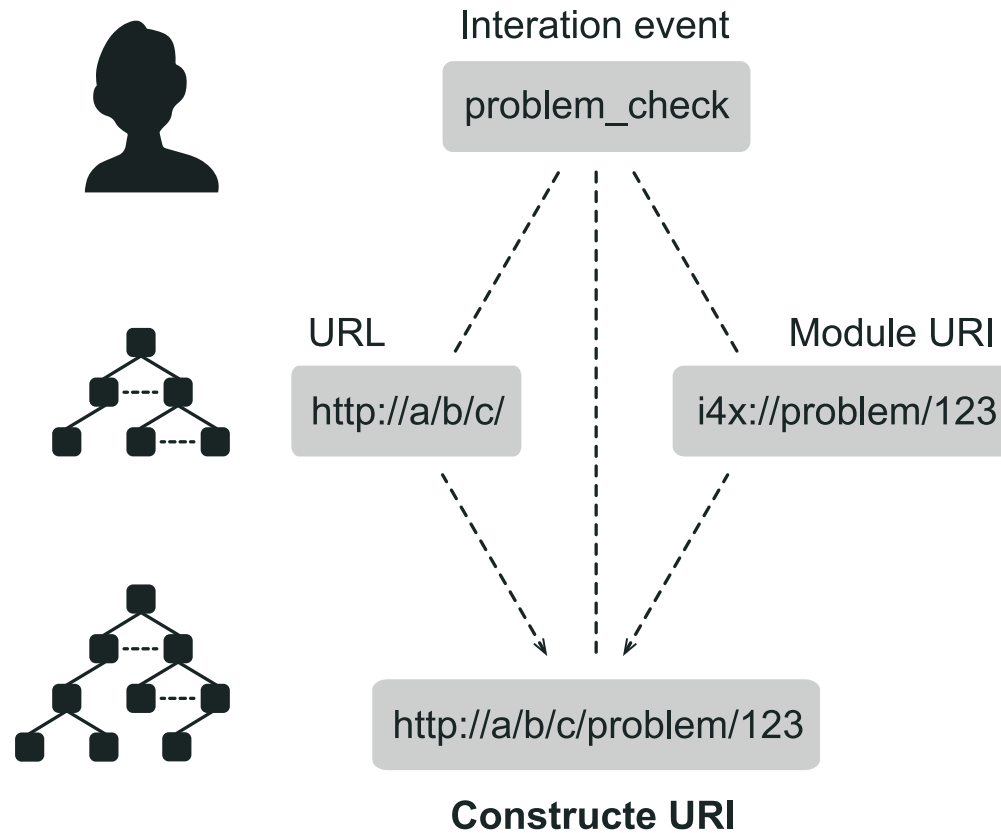
event_type	page	problem_id	resource
problem_check	http://a/b/c/	i4x://problem/123	http://a/b/c/problem/123

Organize: Inheritance
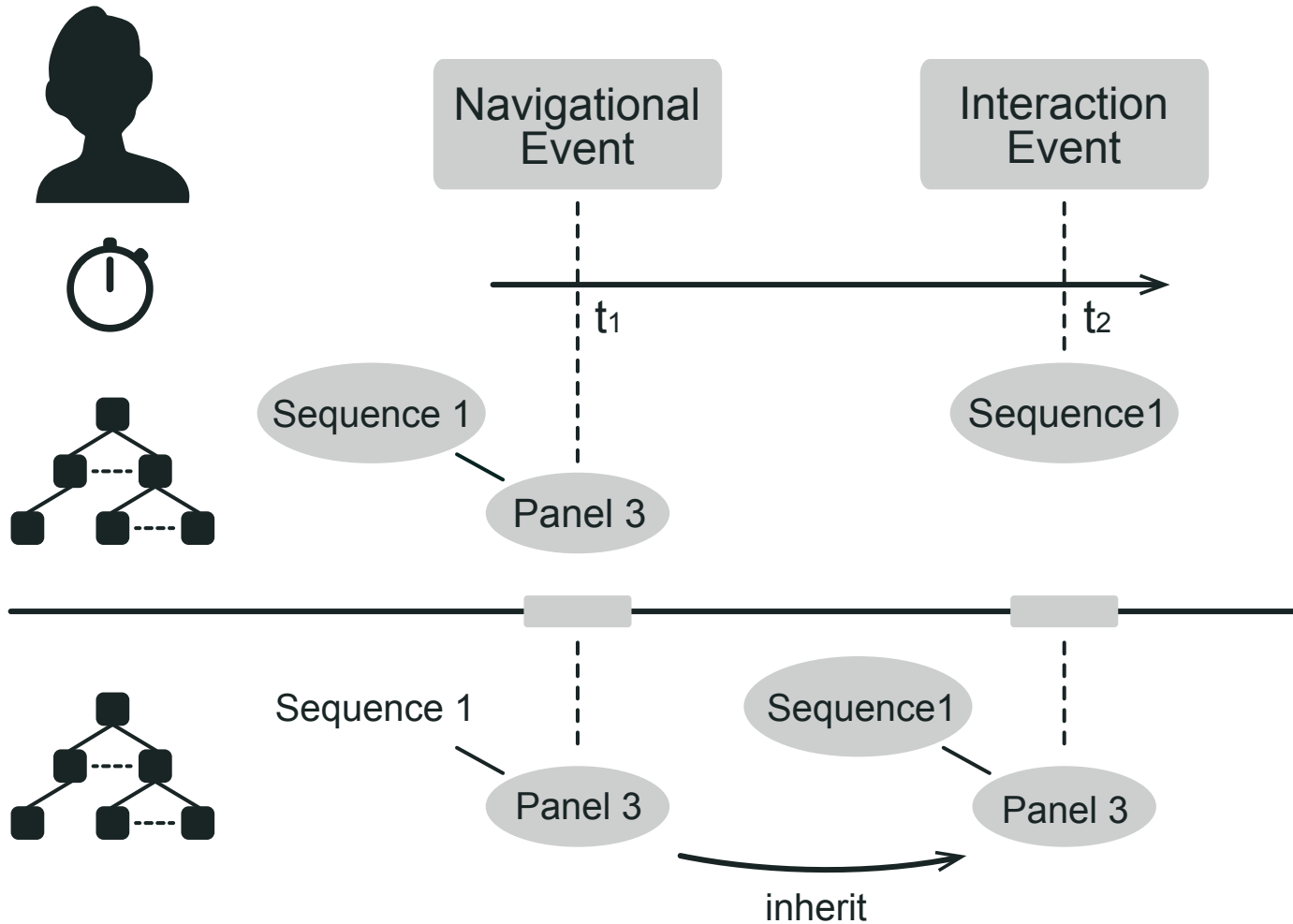
Organize: Inheritance
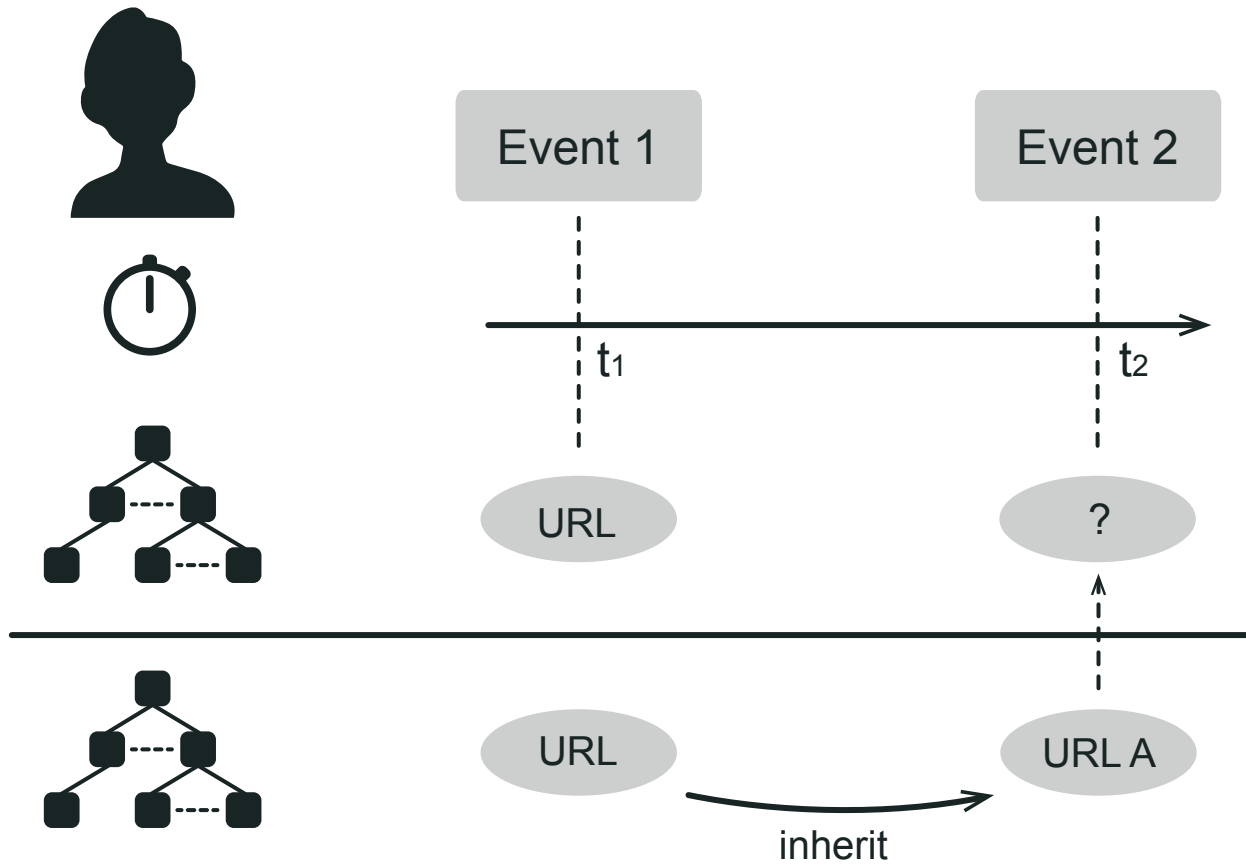
Organize: preprocess
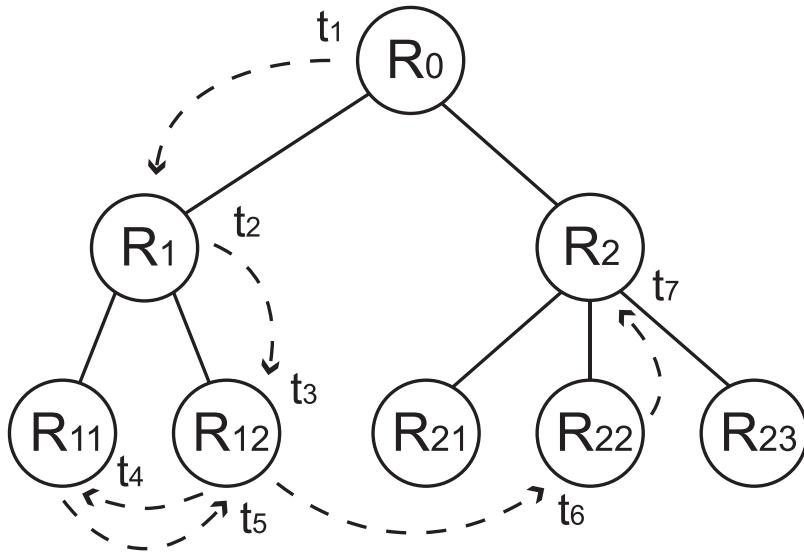
	Resource	Time spent
a	R_0	$t_2 - t_1$
b	R_1	$t_3 - t_2$
c	R_{12}	$t_4 - t_3 + t_6 - t_5$
d	R_{11}	$t_5 - t_4$
e	R_{22}	$t_7 - t_6$

So what are MLBlocks?
Detailed breakdown

Generate insights
Validate/Disemminate

Organize

Pre process

Model

Process

Modeling

Data representation

Primitive constructs

Statistical interpretations

Aggregation

Feature engineering

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL
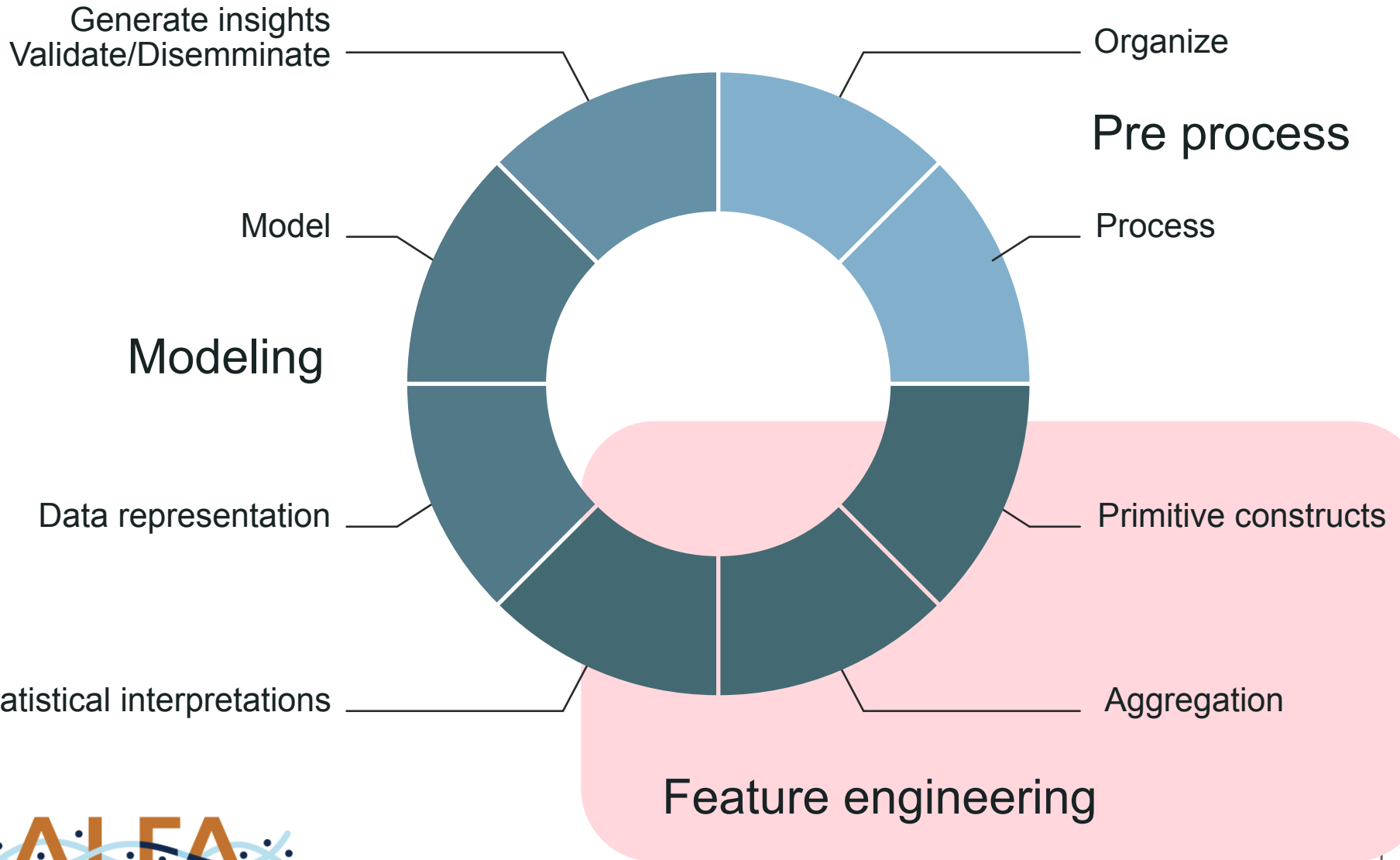
Feature engineering
Primitive constructs

- **Students activity falls into either of three**
  - **Spending time on resources**
  - **Submitting solutions to problems**
  - **Interacting with each other**
  - **Other (peer grading, content creation etc)**

- **Basic constructs**
  - **Number of events**
  - **Amount of time spent**
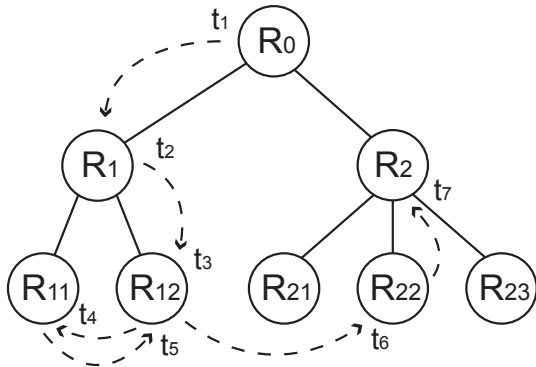 - **Number of submissions, attempts**

Feature engineering
Primitive constructs
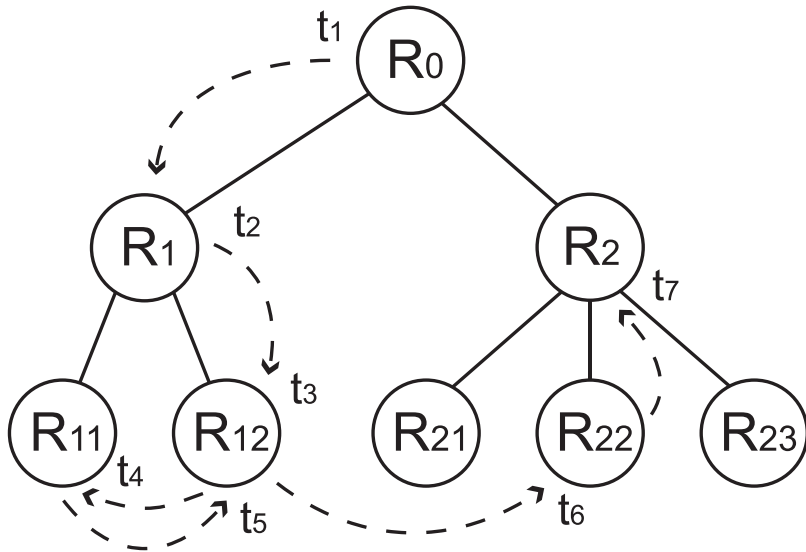
	Resource	Time spent
a	R_0	$t_2 - t_1$
b	R_1	$t_3 - t_2$
c	R_{12}	$t_4 - t_3 + t_6 - t_5$
d	R_{11}	$t_5 - t_4$
e	R_{22}	$t_7 - t_6$

Feature engineering
Aggregates

	Resource	Time spent
a	R_0	$t_2 - t_1$
b	R_1	$t_3 - t_2$
c	R_{12}	$t_4 - t_3 + t_6 - t_5$
d	R_{11}	$t_5 - t_4$
e	R_{22}	$t_7 - t_6$

Resource	Aggregate
R_0	$a + b + c + d + e$
R_1	$b + c + d$
R_2	e
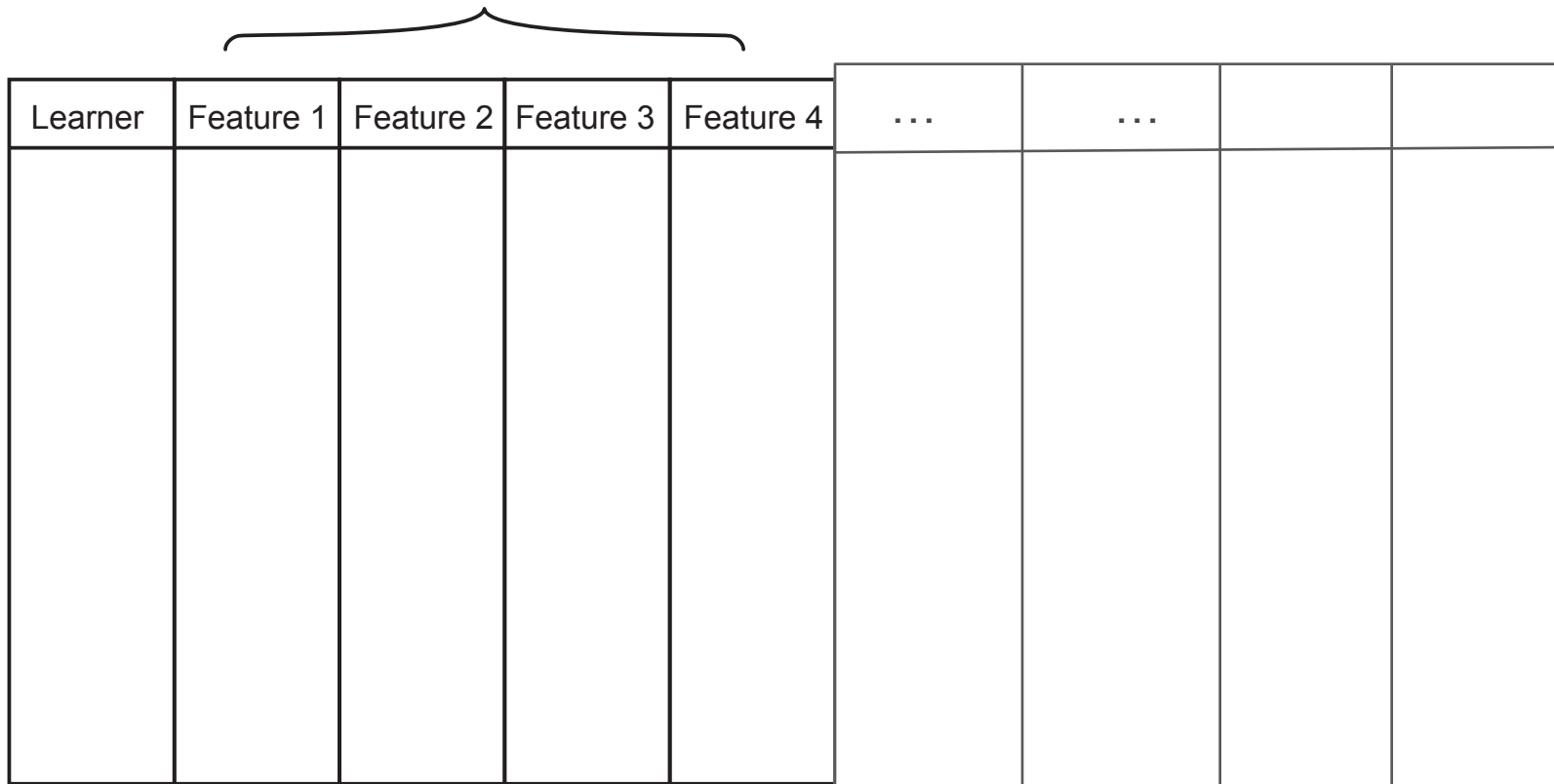
- Aggregate by resource hierarchy
- Aggregate by resource type
  - Book, lecture, forums

Feature Engineering: Primitive aggregates

Total time spent on the course
number of forum posts
number of wiki edits
number of distinct problems attempted
number of submissions (includes all attempts)
number of collaborations
number of correct submissions
total time spent on lecture
total time spent on book
total time spent on wiki
Number of forum responses
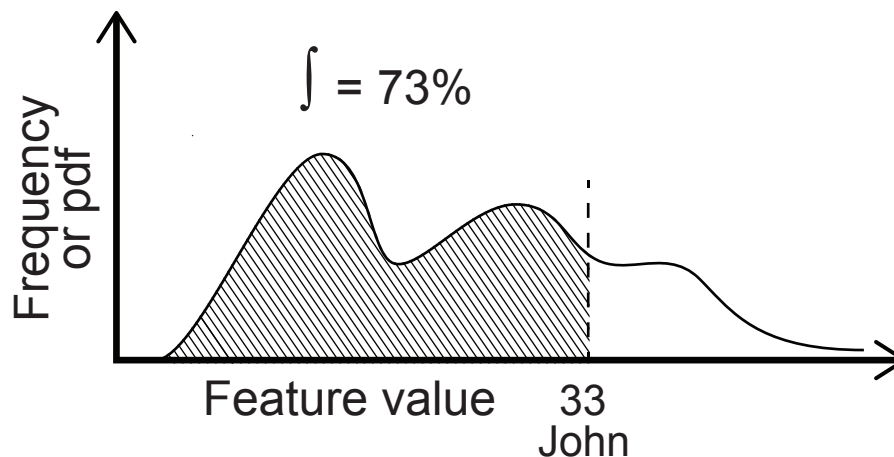
Feature Engineering : Primitive constructs

Primitive

Learner	Feature 1	Feature 2	Feature 3	Feature 4	. . .	. . .		

Feature Engineering - Statistical interpretations

Percentiles, relative standing of a learner amongst his peers
Uni-variate explanation

Learner	Feature value
Verena	32
Dominique	61
Sabina	21
Kalyan	12
Fabian	32
John	33
⋮	⋮
Sheila	88

Feature Engineering : Statistical interpretations

Percentiles, relative standing of a learner amongst his peers
Multivariate explanation

Learner	Feature value 1	Feature value 2
Verena	32	12.4
Dominique	61	2.3
Sabina	21	6.1
Kalyan	12	7.8
Fabian	32	12.4
John	33	12
⋮	⋮	⋮
Sheila	88	12.4
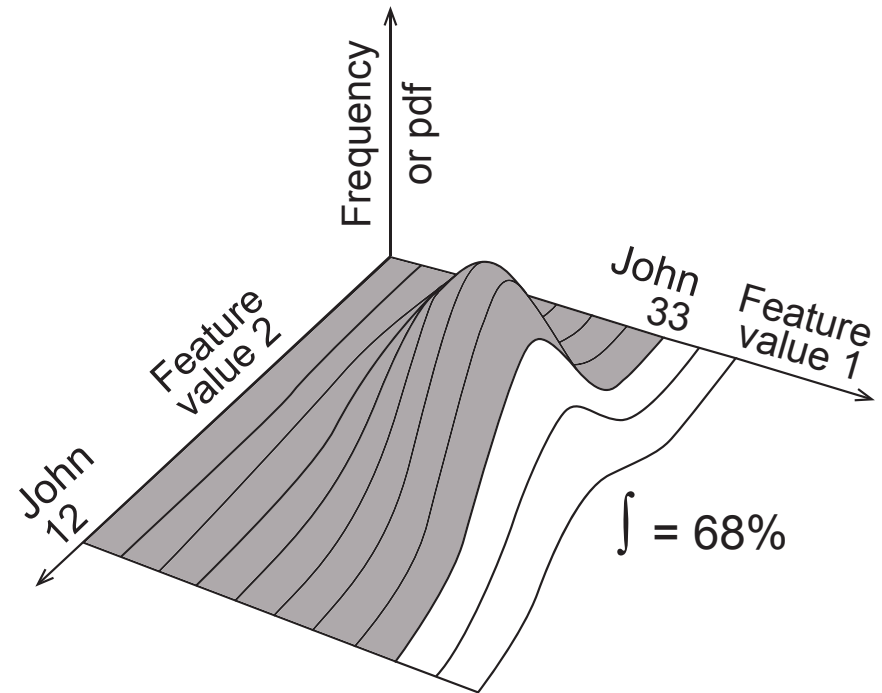
Feature Engineering : Statistical interpretations

Trend of a particular variable over time
Rate of change of the variable

John	
t	Feature value
8	38
9	33
10	44
. . . .	. . . .
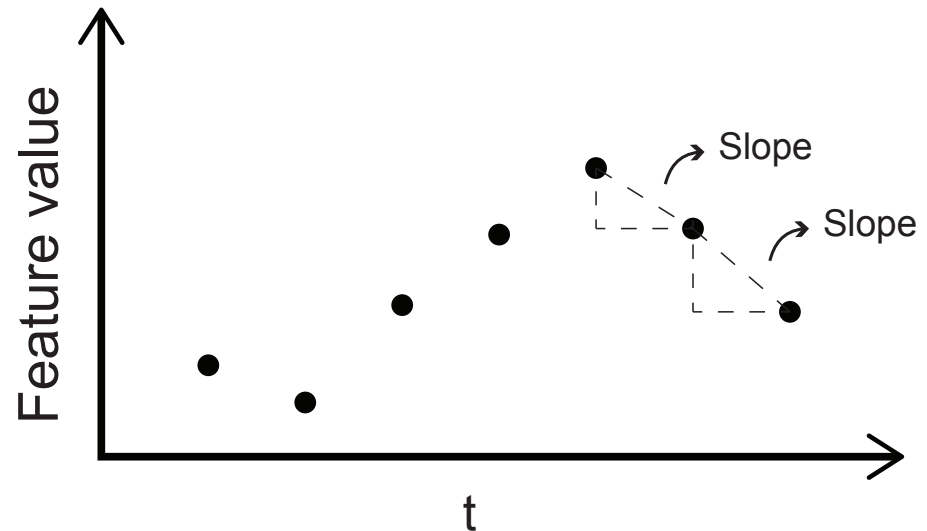
More complex
Learner's topic distribution on a weekly basis
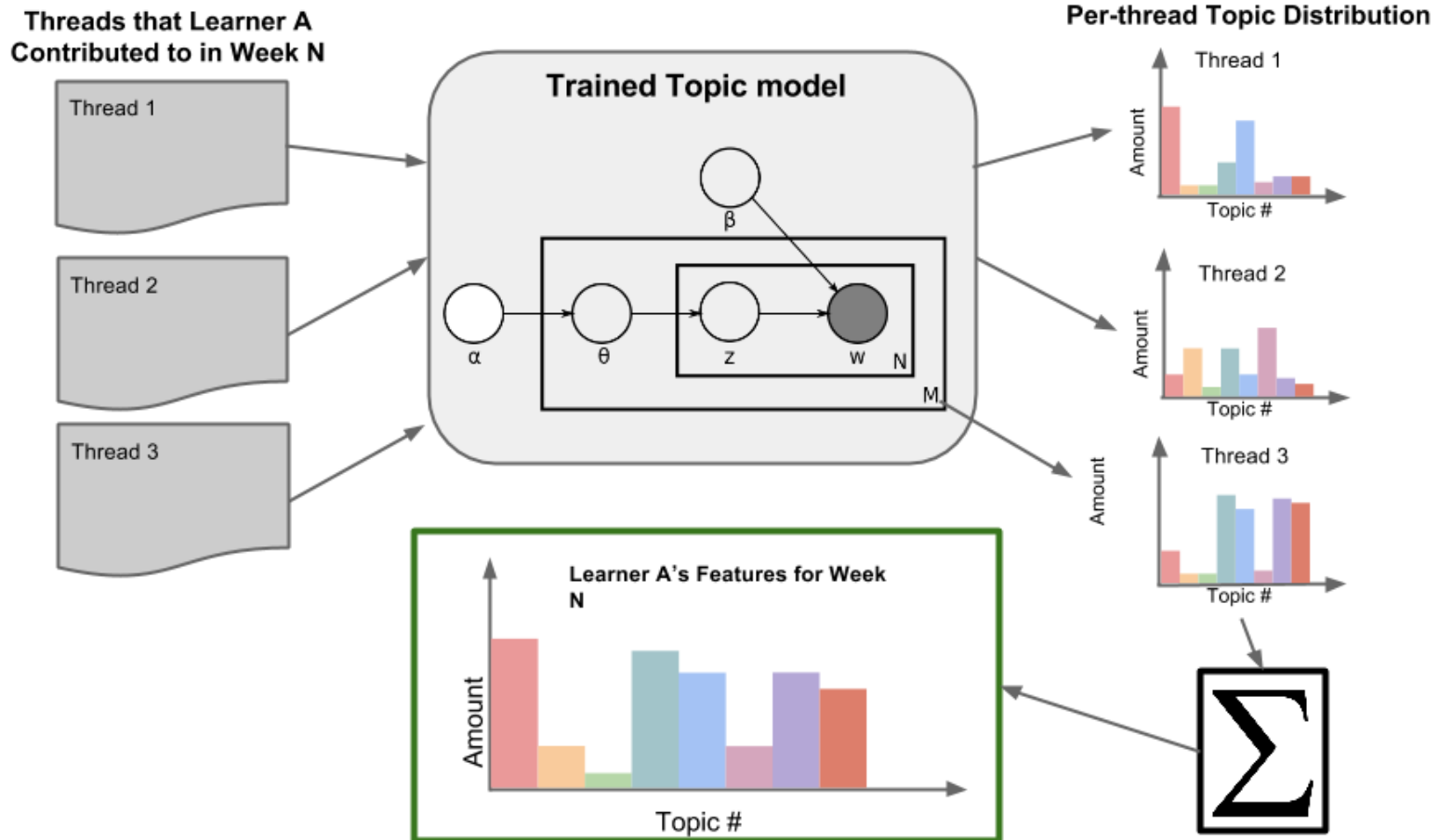
Modeling the Learners time series using HMM
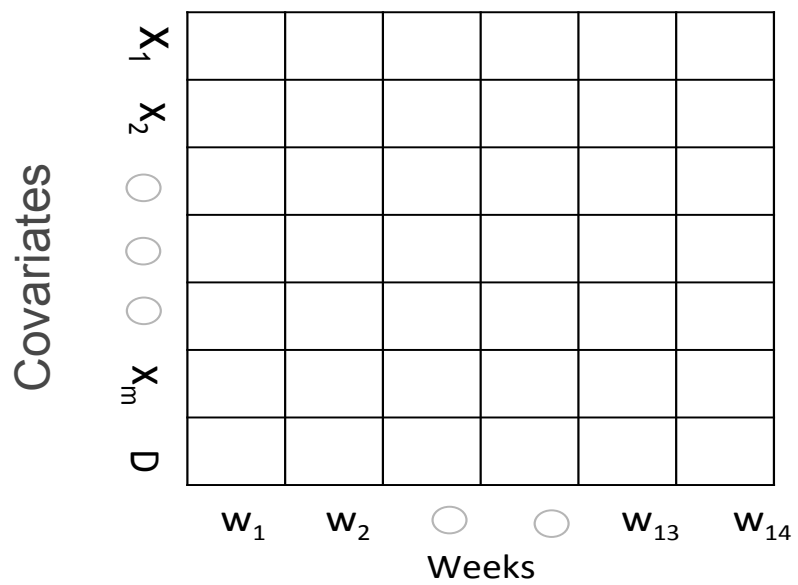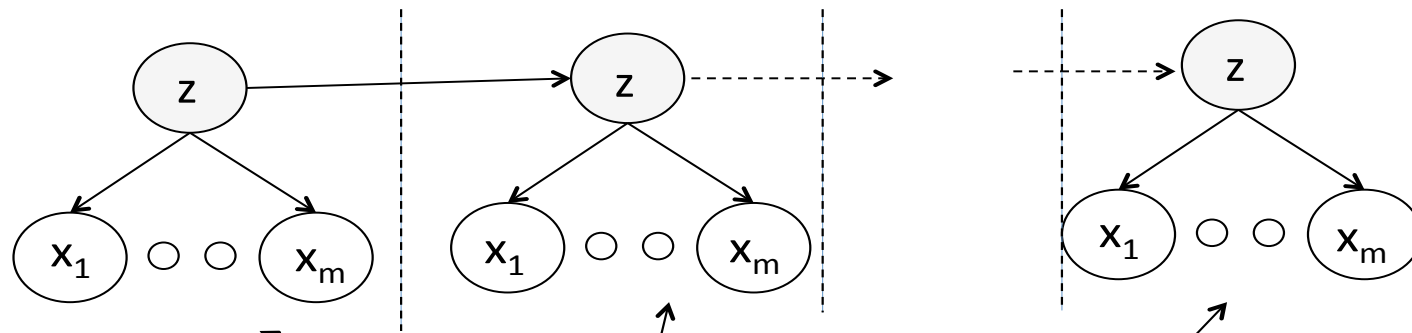
One learners matrix

HMM state probabilities as features

$p(z_1)$

$p(z_2)$

t=1

t=2

$p(z_1)$ $p(z_2)$

covariates

Z

Z

X_1 ○ ○ X_{m+1}

X_1 ○ ○ X_{m+1}
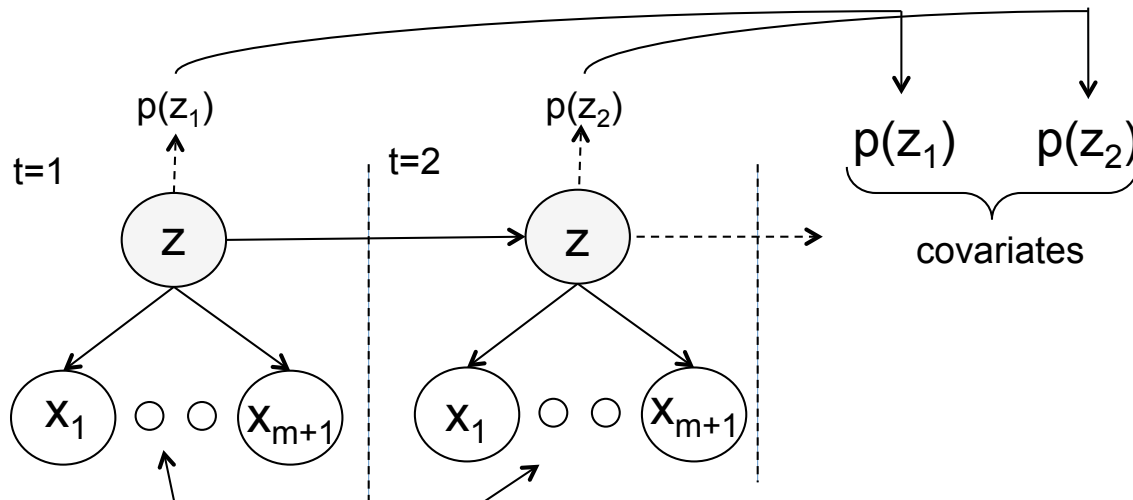
Features for a learner
at the end of second week

X_1 X_2 ○   S

w_1 w_2 ○   ○ w_{13} w_{14}

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

More specifically

State	Probability
1	0.02
2	0.001
3	0.84
4	0.03
5	0.109

State	Probability
1	0.23
2	0.001
3	0.112
4	0.12
5	0.5370

Features

H   H   H   H

3   4   4   5   V₁   Vₙ   V₁   Vₙ

Week 1   Week 2   t=3   Week 14
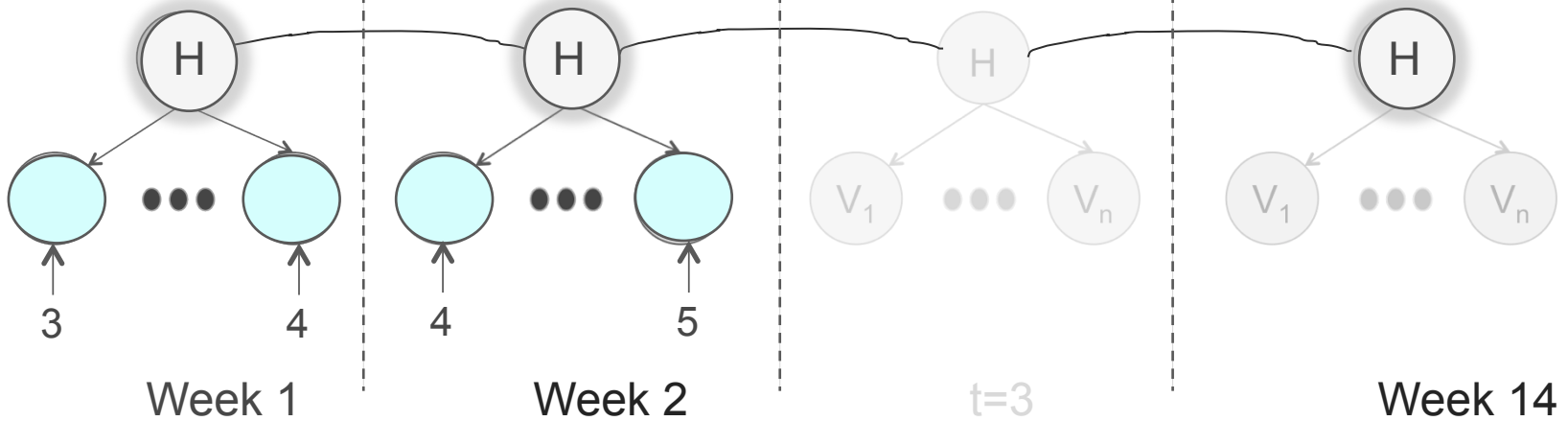
Feature Engineering
Digital learner quantified!

Primitive   Statistical   time series based (including hmm)

Learner	Feature 1	Feature 2	Feature 3	Feature 4	...	...	...	...	...	...	Feature n-1	Feature n

Feature Engineering

Features/ Variables	
Statistical interpetations	
Aggregation	
Primitive constructs	

What we can't automate ?

- **Constructs that are based on our intuition**
  - **average time to solve problem**
  - **observed event variance (regularity)**
  - predeadline submission time (average)
  - Time spent on the course during weekend
- **Constructs that are contextual**
  - pset grade (approximate)
  - lab grade
  - Number of times the student goes to forums while attempting problems
- **Ratios**
  - **time spent on the course per-correct-problem**
  - **attempts per correct problems**
- **Constructs that are course related**
  - **Performance on a specific problem/quiz**
  - **Time spent on a specific resource**

Feature Factory
Crowd source variable discovery

Feature Factory MIT CSAIL ALFA Lab

Feature discovery is a challenging aspect of the data science and knowledge discovery. Creating an online interactive space where data scientists can benefit from each other's ideas on various features can significantly simplify and expedite the process. Feature Factory is an online platform where ALFA@CSAIL will present a prediction problem for which features are sought. For the prediction problem, the group will provide downloadable mock data so users can test their scripts and submit. Feature Factory seeks three kinds of contributions: ideas of new features, feature extraction code and comments on existing ones.

Upon the submission of the feature extraction code , it will be validated on our online mock dataset and you will be notified of the result immediately. Upon validation, our team will execute the code on the real dataset to generate the features and insert the new feature into a number of machine learning models using discriminative (Decision trees, Neural networks, support vector Machines), generative (logistic regression, Gaussian process) and time series models. As a result, your features will be ranked against one another.

Current Focus Problem: Predict Student Stopouts on Massive Open Online Courses

Data model

In this problem, our goal is to predict when a student will stop engaging with the course. A student is assumed to have stopped out from a course when s/he stops to attempt problems/homeworks. We have data captured from students online behavior, which includes click stream data, their online forum interactions and their submissions for problems. We have a comprehensive data schema, called MOOCdb which captures the student activity data on a MOOC platform. The data schema is documented here. A small mock dataset that is in the form of the data schema can be downloaded in two formats: sql or csv.

Featurefactory.csail.mit.edu

Feature Factory

Featurefactory.csail.mit.edu

How does one participate?

featurefactory.csail.mit.edu

① Think and propose

② Comment

③ Help us extract by writing scripts

Extract
Supplying us a script

Pause and exercise

- **Based on your experience**
- **Propose a variable or a feature that we can form for a student on a weekly or per module basis**

- **Current list of extracted variables and proposals made by others are at:**
  - **http://featurefactory.csail.mit.edu**
- **You can add your idea there**
  - **http://featurefactory.csail.mit.edu**
- **Or you can add your idea and more detail with this google form**
  - **http://shoutkey.com/attractive**

That URL again is

http://shoutkey.com/attractive

What did we assemble as variables so far?

Simple

Total time spent on the course
number of forum posts
number of wiki edits
average length of forum posts (words)
number of distinct problems attempted
number of submissions (includes all attempts)
number of distinct problems correct
average number of attempts
number of collaborations
max observed event duration
number of correct submissions

Complex

average time to solve problem
observed event variance (regularity)
total time spent on lecture
total time spent on book
total time spent on wiki
Number of forum responses
predeadline submission time (average)

Derived

attempts percentile
pset grade (approximate)
pset grade over time
lab grade
lab grade over time
time spent on the course per-correct-problem
attempts per correct problems
percent submissions correct

What did we assemble as variables so far?

Simple

Total time spent on the course
number of forum posts
number of wiki edits
average length of forum posts (words)
number of distinct problems attempted
number of submissions (includes all attempts)
number of distinct problems correct
average number of attempts
number of collaborations
max observed event duration
number of correct submissions

Note:
- Red were proposed by crowd
- For definitions of *simple*, *complex* and *derived*
  Please check out http://arxiv.org/abs/1407.5238

Complex

average time to solve problem
observed event variance (regularity)
total time spent on lecture
total time spent on book
total time spent on wiki
Number of forum responses
predeadline submission time (average)

Derived

attempts percentile
pset grade (approximate)
pset grade over time
lab grade
lab grade over time
time spent on the course per-correct-problem
attempts per correct problems
percent submissions correct

So what are MLBlocks?
Detailed breakdown

Generate insights
Validate/Disemminate

Organize

Pre process

Model

Process

Modeling

Data representation

Primitive constructs

Statistical interpretations

Aggregation

Feature engineering

Dropout prediction problem

Given current student behavior if s/he will dropout in the future?

Predict

| Lag | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |

Lead

We can use students data during these weeks

Weeks →

Note: By varying lead and lag we get 91 prediction problems

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL
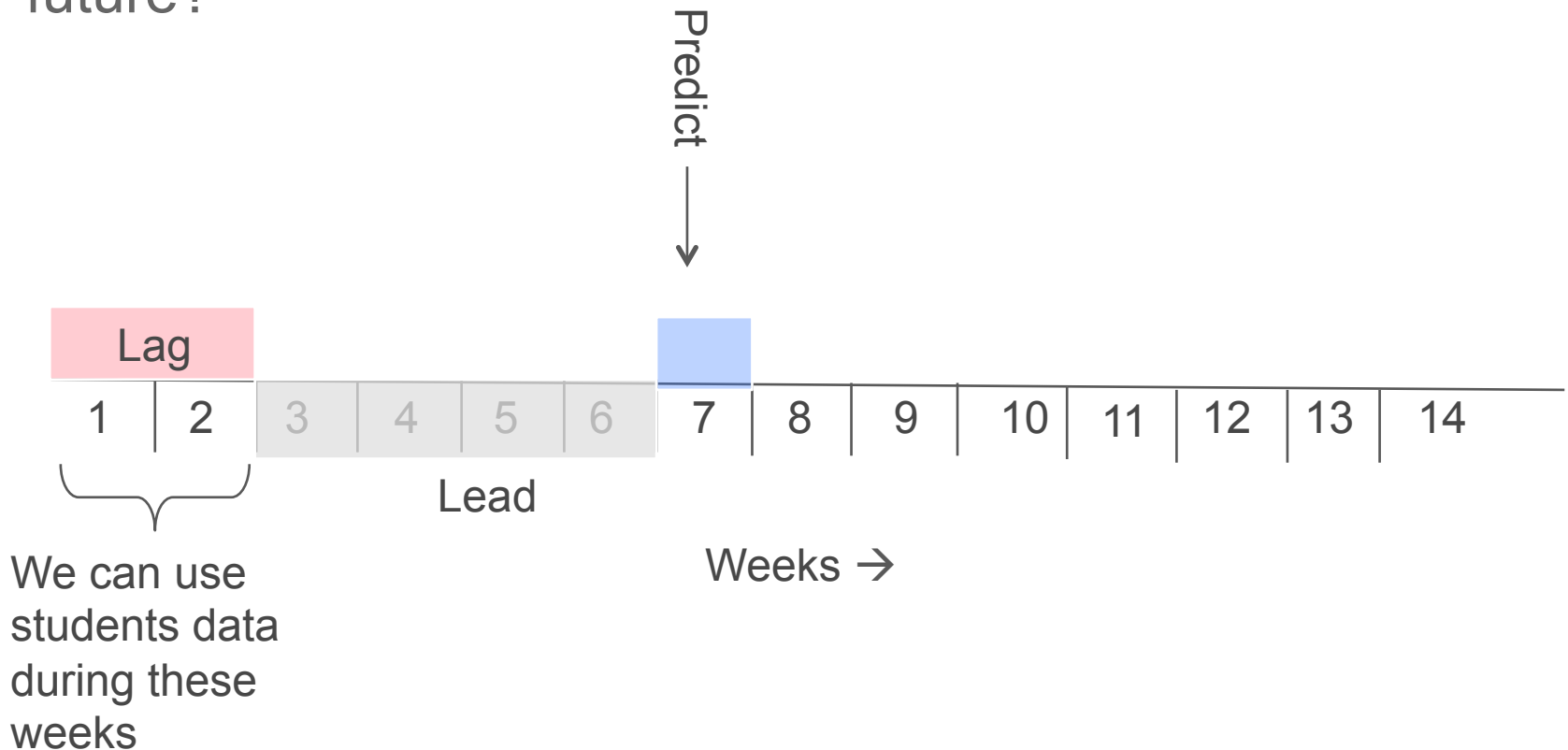
The Numbers

- **154,763 students registered in 6.002x Spring 2012**
- **200+ Million events**
  - 60 GB of raw click stream data
- **52000+ students in our study**
  - 130 Million events
- **44,526 never used forum or wiki**
- **Models use 27 predictors with weekly values**
  - 351 dimensions at max
- **Predictors reference clickstream to consider**
  - Time, performance on assessment components
    - » homeworks, quizzes, lecture exercises
  - Time, use of resources
    - » videos, tutorials, labs, etexts, …
- **5000+ models learned and tested**
  - 91 prediction problems for each of 4 cohorts
  - 10 fold cross validation and once on entire training -> 11 models per problem
  - Extra modeling to examine influential features
  - Multi-algorithm modeling on problems with less accurate models
  - HMM modeling and 2-level HMM-LR modeling

Splitting into cohorts
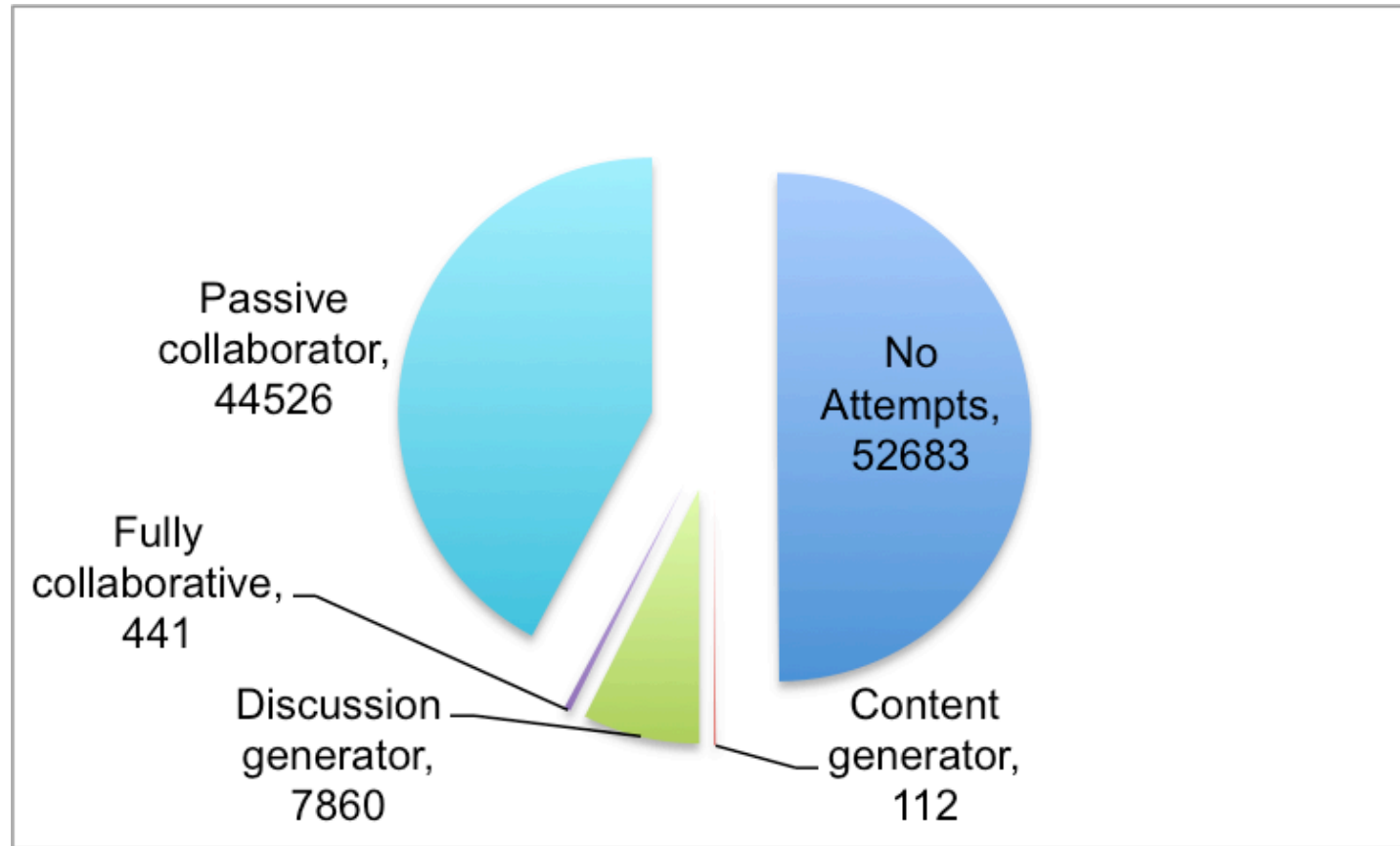
Models

- **Logistic regression**
- **Hidden markov models**
- **Hidden markov models + LR**
- **Randomized logistic regression**
  - **For variable importance**

Learner per-week variable matrix

	x_1	x_2	$\circ$	$\circ$	x_m	S
w_1						
w_2						
w_{13}						
w_{14}						

Weeks

Data Representation
Flattening it out for Discriminatory Models
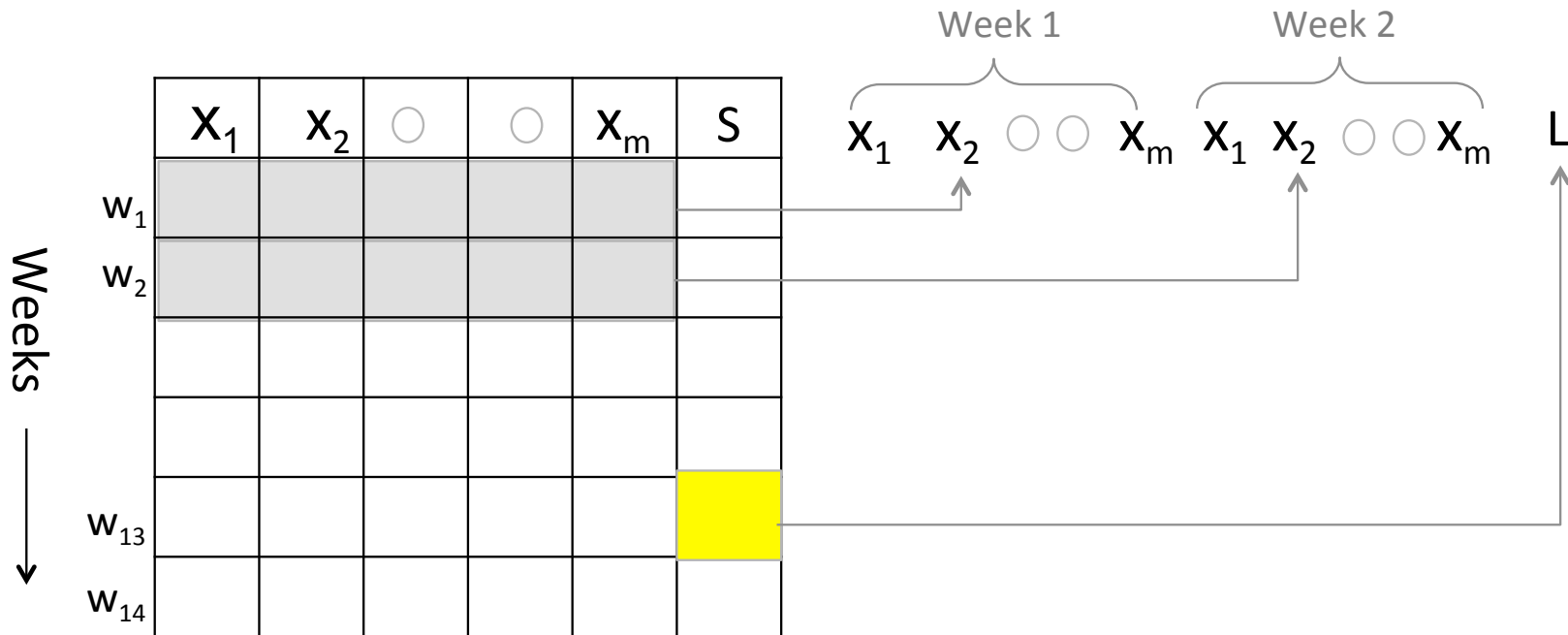
Lag 2 – Lead 11 prediction problem

Logistic Regression

Hidden Markov Model as a Prediction Engine
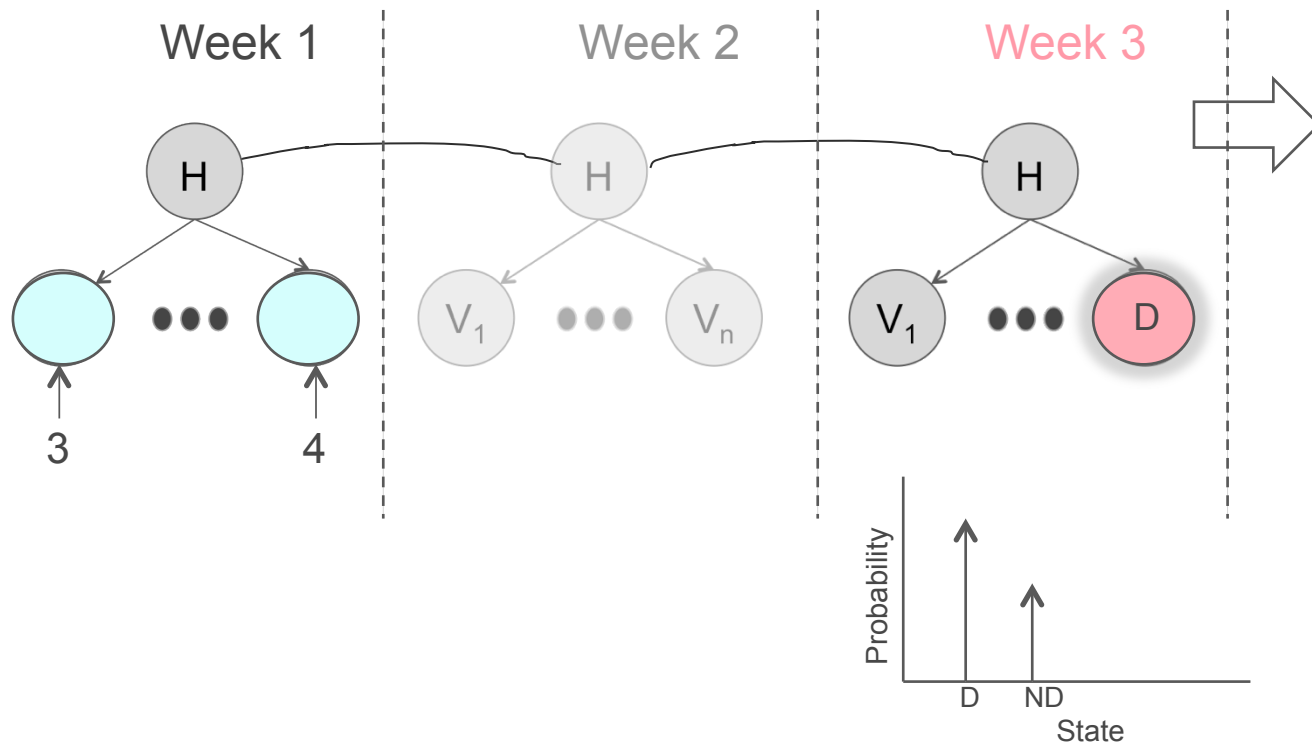
Week 1 data, predict 2 weeks ahead

Hidden Markov Model as a Prediction Engine

Week 1   Week 2   Week 3   Week 4
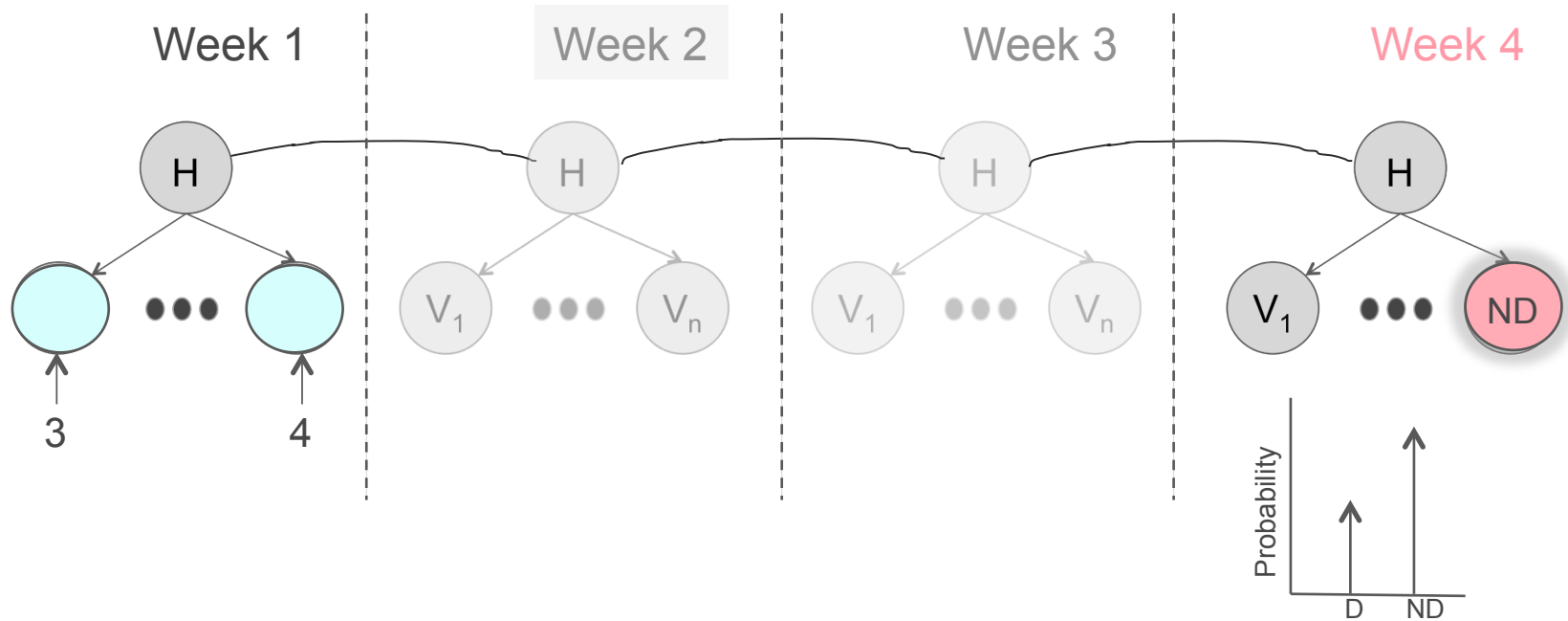
Week 1 data, predict 3 weeks ahead

HMM performance
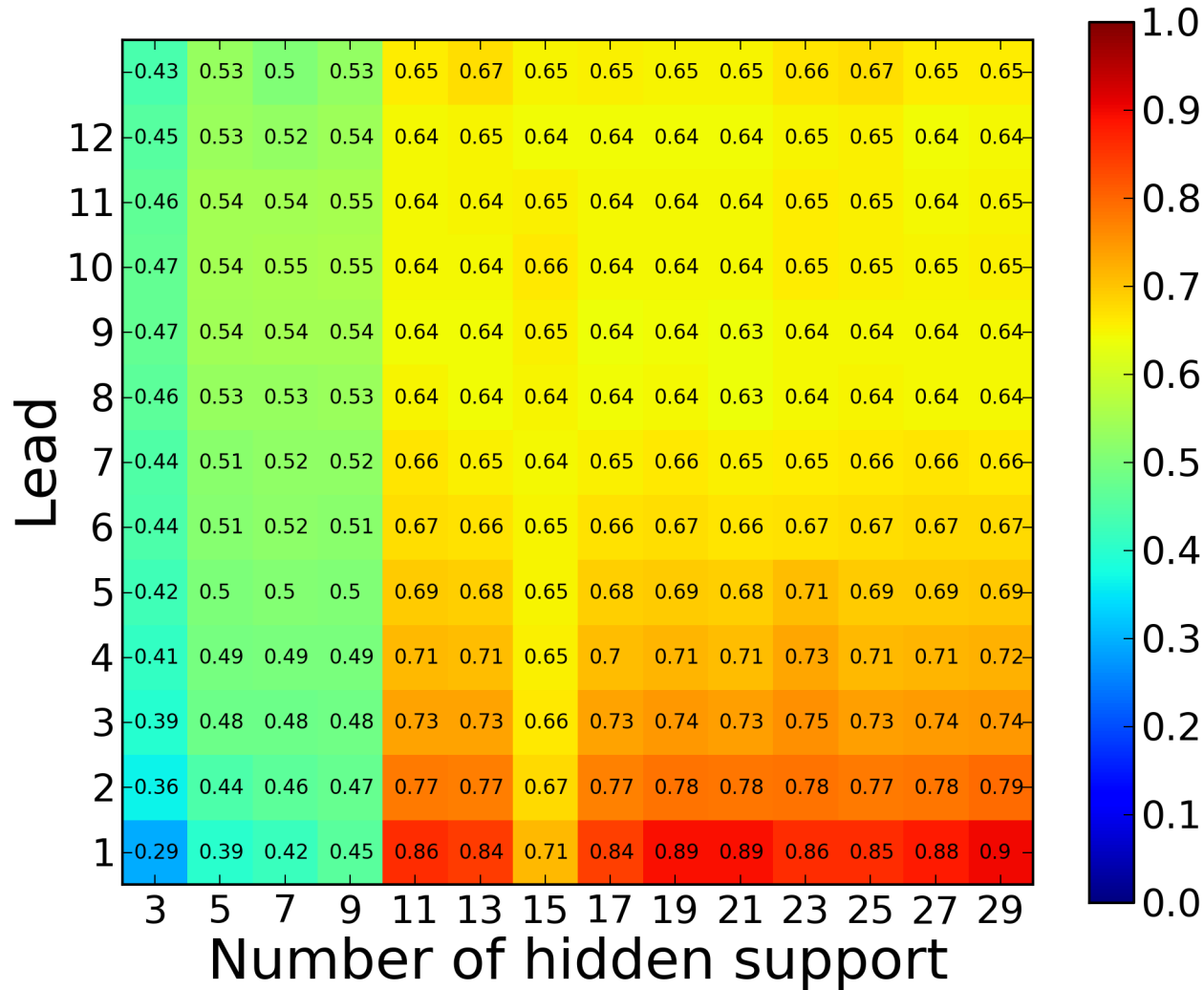
Hidden state probabilities as variables

State	Probability
1	0.02
2	0.001
3	0.84
4	0.03
5	0.109

State	Probability
1	0.23
2	0.001
3	0.112
4	0.12
5	0.5370

Variables

0.23, 0.001, 0.112, 0.12, 0.5370

Week 1   Week 2   Week 3   Week 4   Week 5

Class label → 5

Lag=2 weeks   Lead=2 weeks
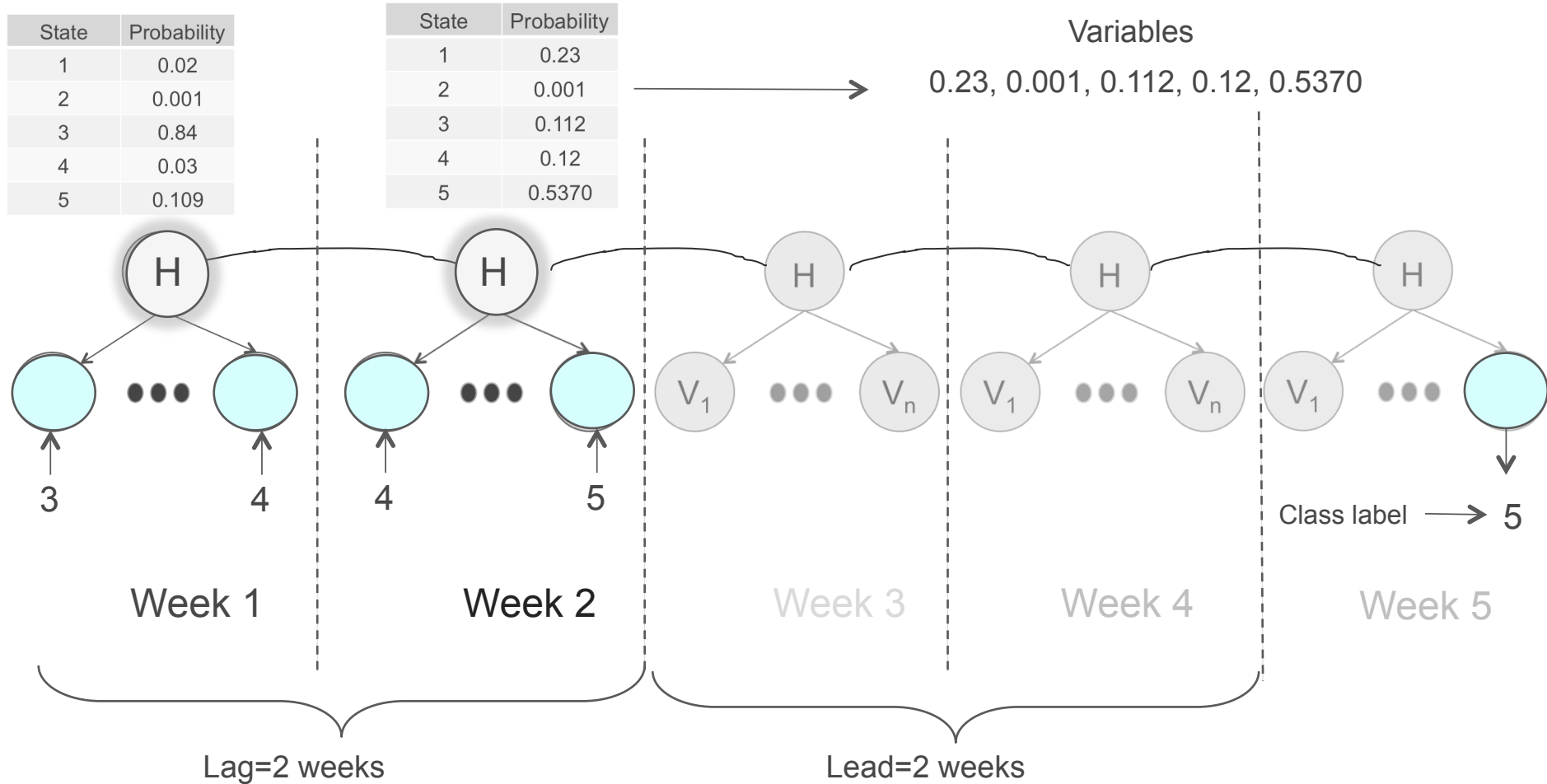
Use 2 weeks data, predict 3 weeks ahead

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Hidden state probabilities
→ Logistic Regression
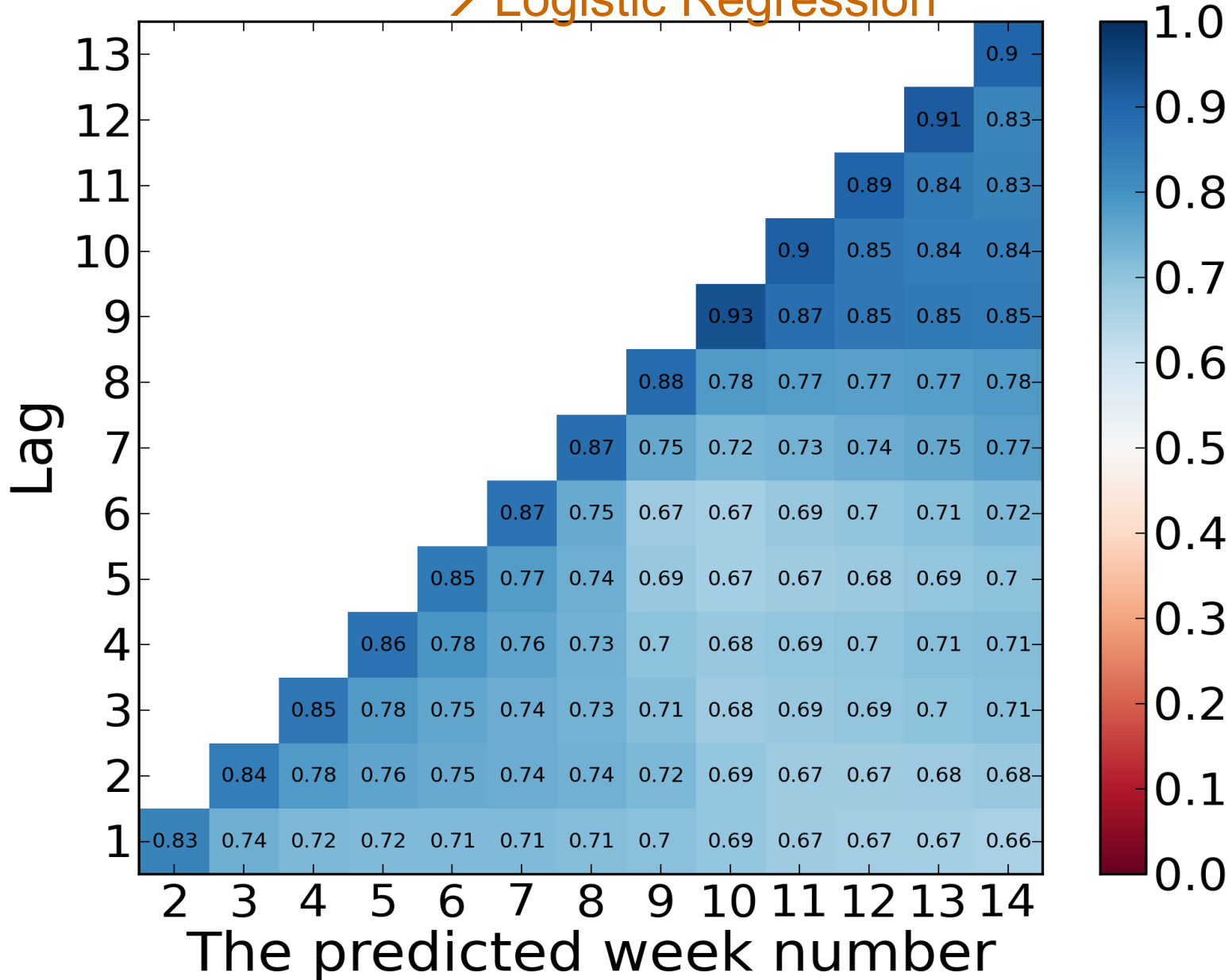
Number of hidden states - 27

So what are MLBlocks?
Detailed breakdown

Randomized Logisitic Regression

- Iterate over data multiple times
  - Each time resampling from data
  - Identify feature weights
  - Aggregate over multiple trials

Randomized Logisitic Regression

Influential Predictors

Q. What predicts a student successfully staying in the course through the final week?

Answer: A student's average number of weekly "submissions" (attempts on all problems include self-tests and homeworks for grade) *relative* to other students', e.g. a percentile variable, is highly predictive.

Relative and trending predictors drive accurate predictions. E.G. a student's lab grade in current week relative to average in prior weeks is more predictive than the grade alone.

Influential Predictors

Q. Across different cohorts of students what is the single most important predictor of dropout?

Answer: A predictor that appears among the most influential 5 in all 4 cohorts is the average "pre-deadline submission time". It is the average duration between when the student submits a homework solution and its deadline.

Interesting Predictors

Human: how regularly the student studies
- X13 "observed event variance"
    - Variance of a students observed event timestamp

Human: Getting started early on pset
- X210: average time between problem submission and pset deadline

Human: how rewarding the student's progress feels
- "I'm spending all this time, how many concepts am I acquiring?"
- X10: Observed events duration / correct problems

Student: it's a lot of work to master the concepts
- Number of problems attempted vs number of correct answers
- X11: submissions per correct problem

Instructor: how is this student faring vs others?
- tally the average number of submission of each student,
- student variable is his/her percentile (x202) or percentage of maximum of all students (X203)

Instructor: how is the student faring this week?
- X204: pset grade
- X205: pset grade trend: difference in pset grade in curent week to student's average pset grade in past weeks

Top 10 features/variables that mattered

- For an extremely hard prediction problem
- Week 1
  - Number of distinct problems correct
  - Predeadline submission time
  - number of submissions correct
- Week 2
  - Lab grade
  - Attempts per correct problem
  - Predeadline submission time
  - Attempts percentile
  - Number of distinct problems correct
  - Number of submissions correct
  - Total time spent on lectures

Predict

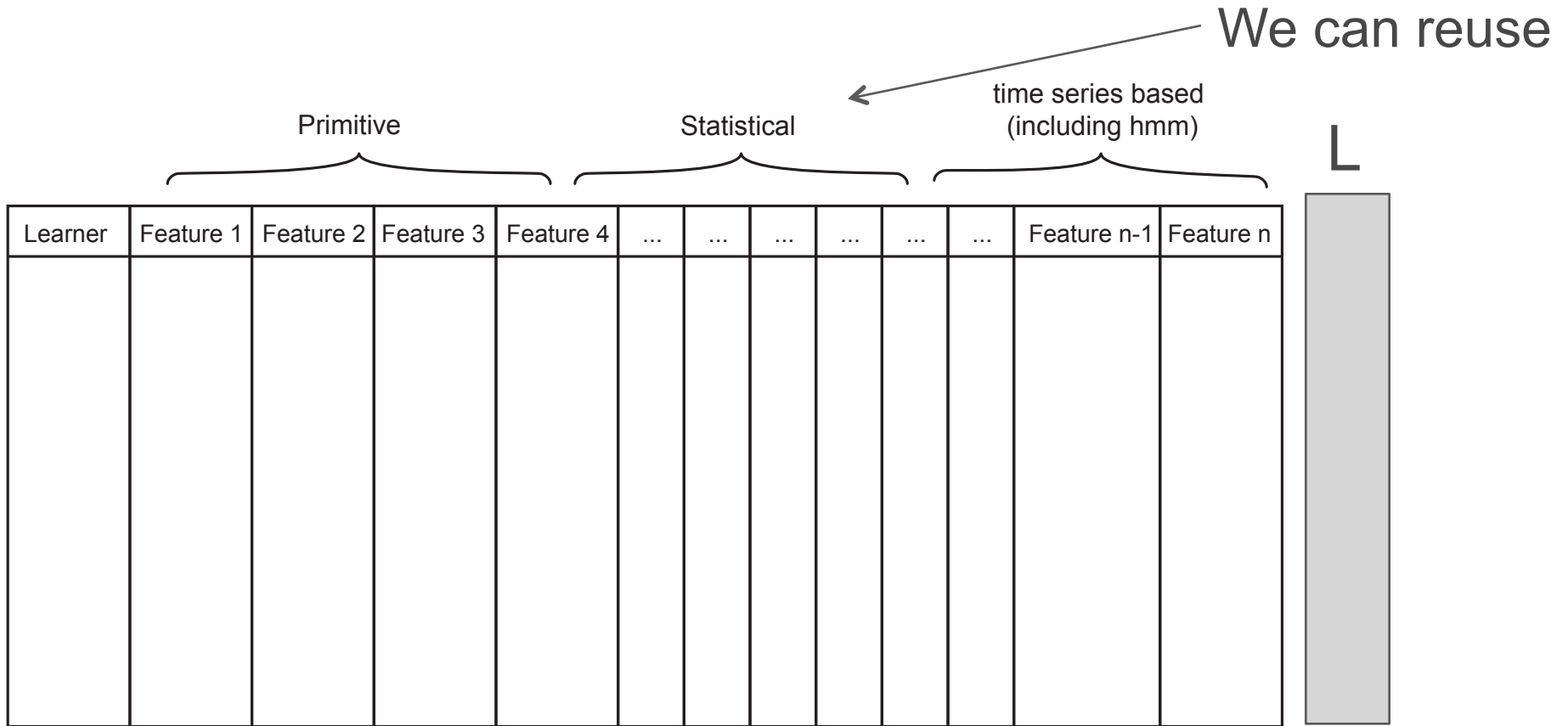| Lag													
1	2	3	4	5	6	7	8	9	10	11	12	13	14

Lead

We can use students data during these weeks

Weeks →

Parameters throughout this process

- Choices we make during the calculations of primitive constructs
  - Cut-offs for duration calculation
- Aggregation parameters

- Parameters for models
  - Number of hidden states
  - Number of topics
- We would next like tune these parameters against a prediction goal

What else can we predict?

We can reuse

time series based (including hmm)

Primitive

Statistical

Learner	Feature 1	Feature 2	Feature 3	Feature 4	...	...	...	...	...	...	Feature n-1	Feature n	L

We can change this

What else should we predict?

- **We want your thoughts/ideas as to what we should next predict using the same matrix**

- **The prediction problem has to be something in future:**
  - **Like whether the student will stopout (we already did that)**
  - **Whether the student will return after stopping out**
  - **Success in next homework**

- **We created a google form and is available at:**

  - **http://shoutkey.com/dissociate**

That URL is

http://shoutkey.com/dissociate

Acknowledgements- Students

- **Roy Wedge**
- **Kiarash Adl**
- **Kristin Asmus**
- **Sebastian Leon**
- **Franck Dernoncourt**
- **Elaine Han aka Han Fang**
- **Colin Taylor**
- **Sherwin Wu**
- **Kristin Asmus**
- **John O'Sullivan**
- **Will Grathwohl**
- **Josep Mingot**
- **Fernando Torija**
- **Max Kanter**
- **Jason Wu**

Acknowledgments

Sponsor: Project QMULUS

PROJECT Q MULUS

PARTNERS

- **Lori Breslow**
- **Jennifer Deboer**
- **Glenda Stump**

T|L|L
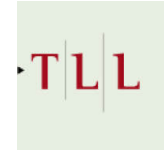
- **Sherif Halawa**
- **Andreas Paepcke**
- **Rene Kizilcec**
- **Emily Schneider**

LYTICS

- **Piotr Mitros**
- **James Tauber**

edX

- **Chuong Do**

coursera

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL