September 2017

Artificial Intelligence in Online Education

With an emphasis on personalization

Curtis G. Northcutt
Massachusetts Institute of Technology
EECS, Office of Digital Learning

Disclaimer

Many of these slides contain ideas which are unsolved or unpublished. Interested? Please reach out :) Collaboration is encouraged!

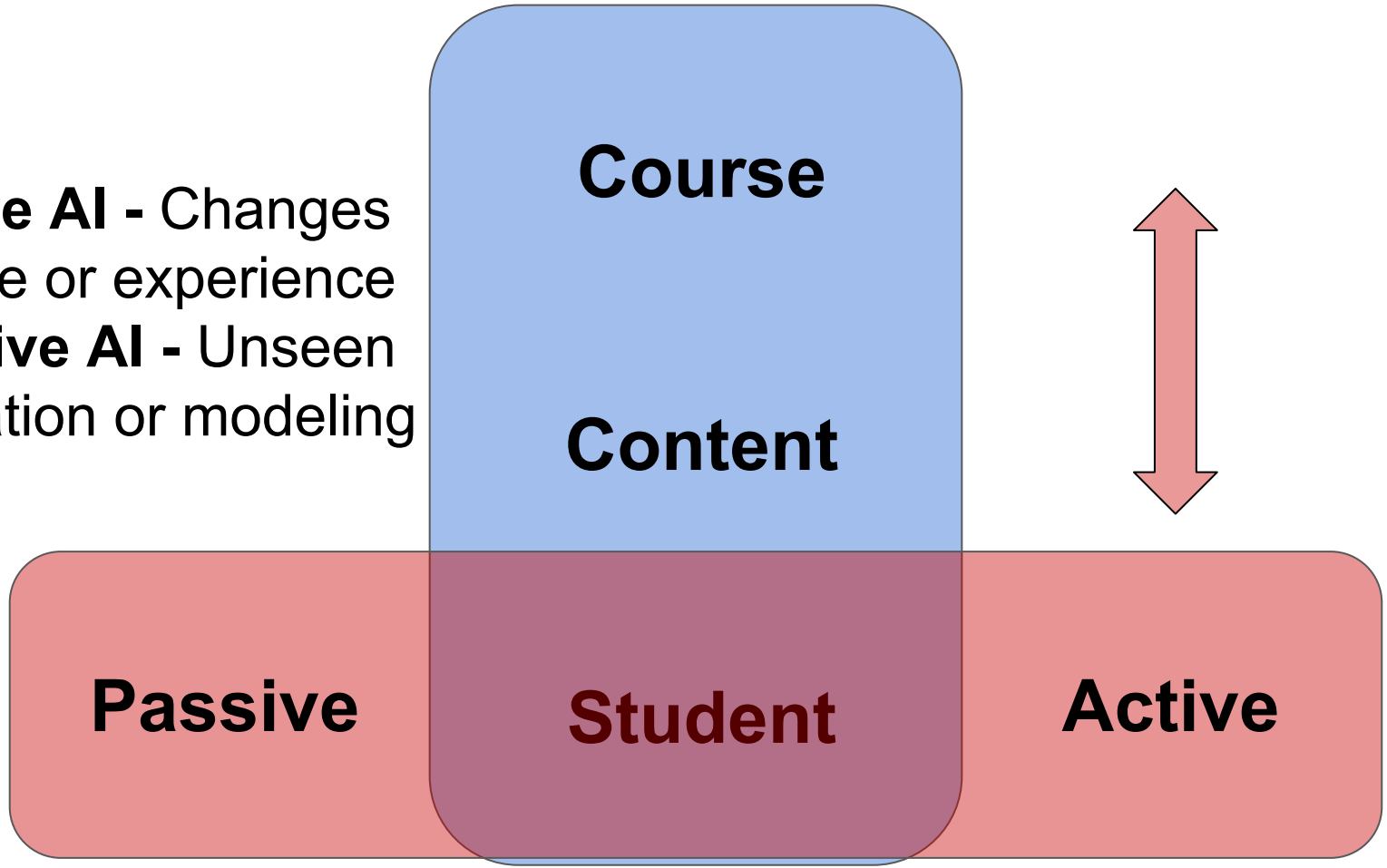
These slides are intended for **researchers**, **practitioners**, and **course staff** who want to improve their **online courses** with **AI**.

I use the term **student**, but these concepts apply to **any learner**.

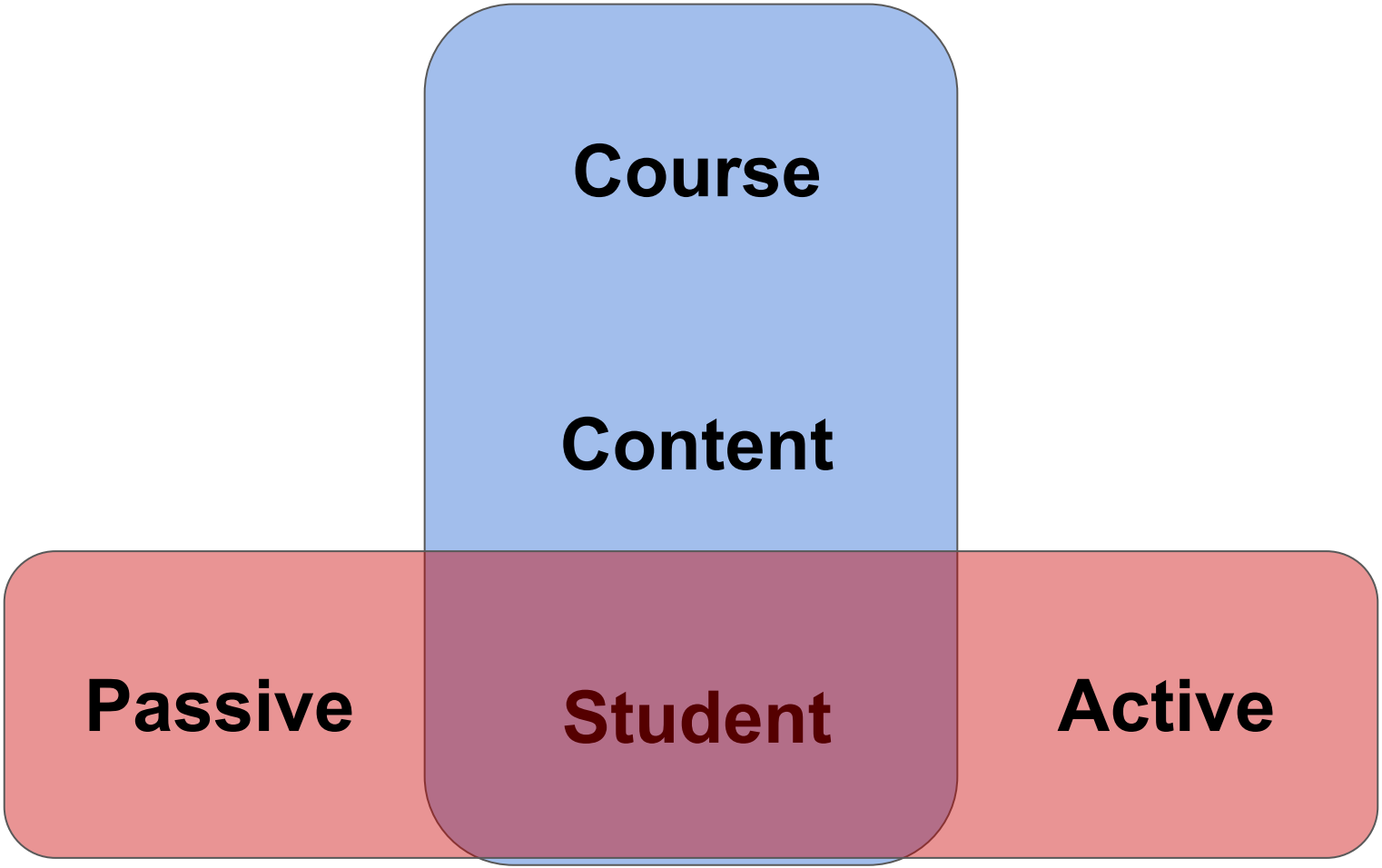
A **framework** within which we can **organize topics** in the intersection of AI & OE

Framework for AI in Online Education

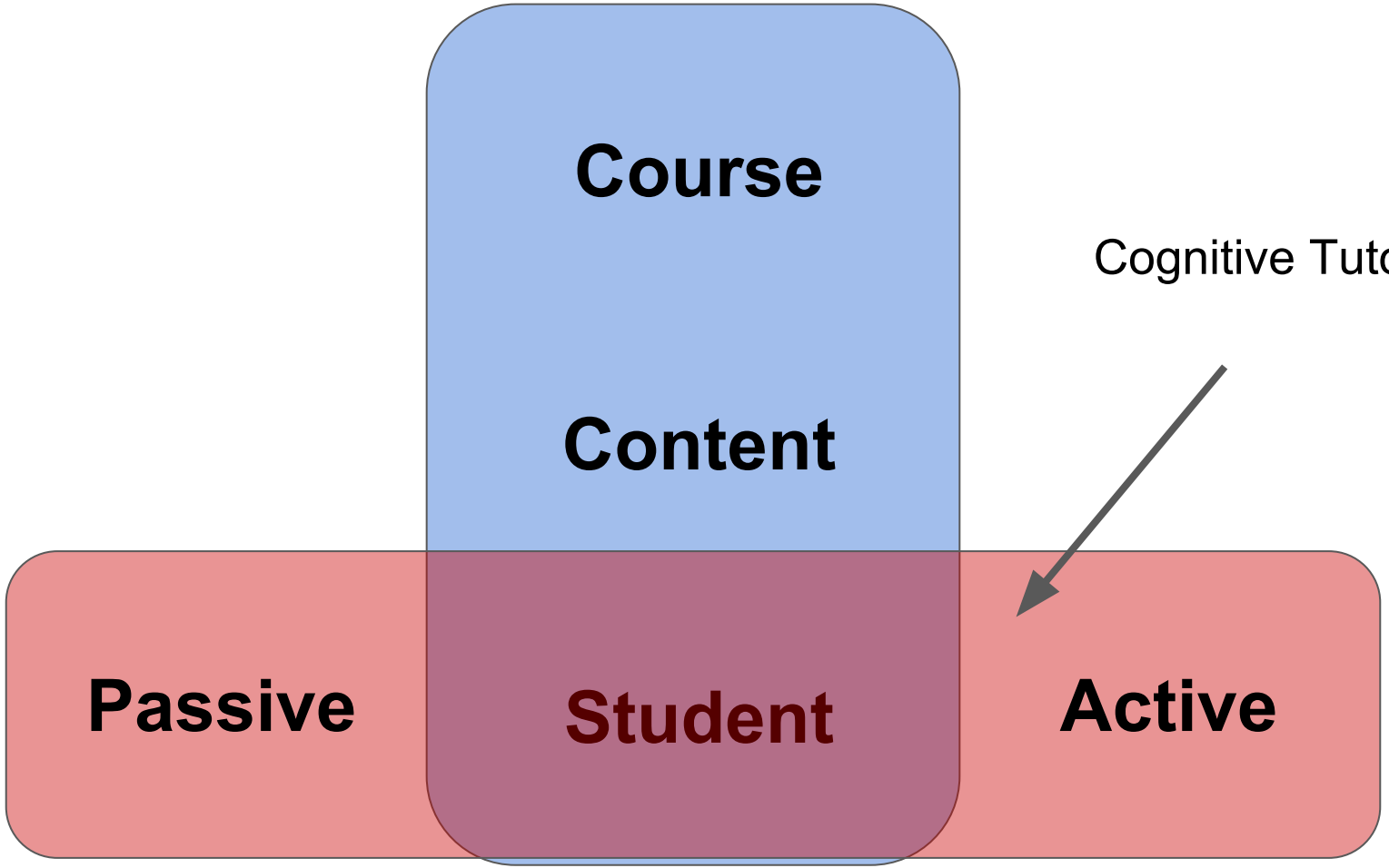
Active AI - Changes course or experience
Passive AI - Unseen estimation or modeling



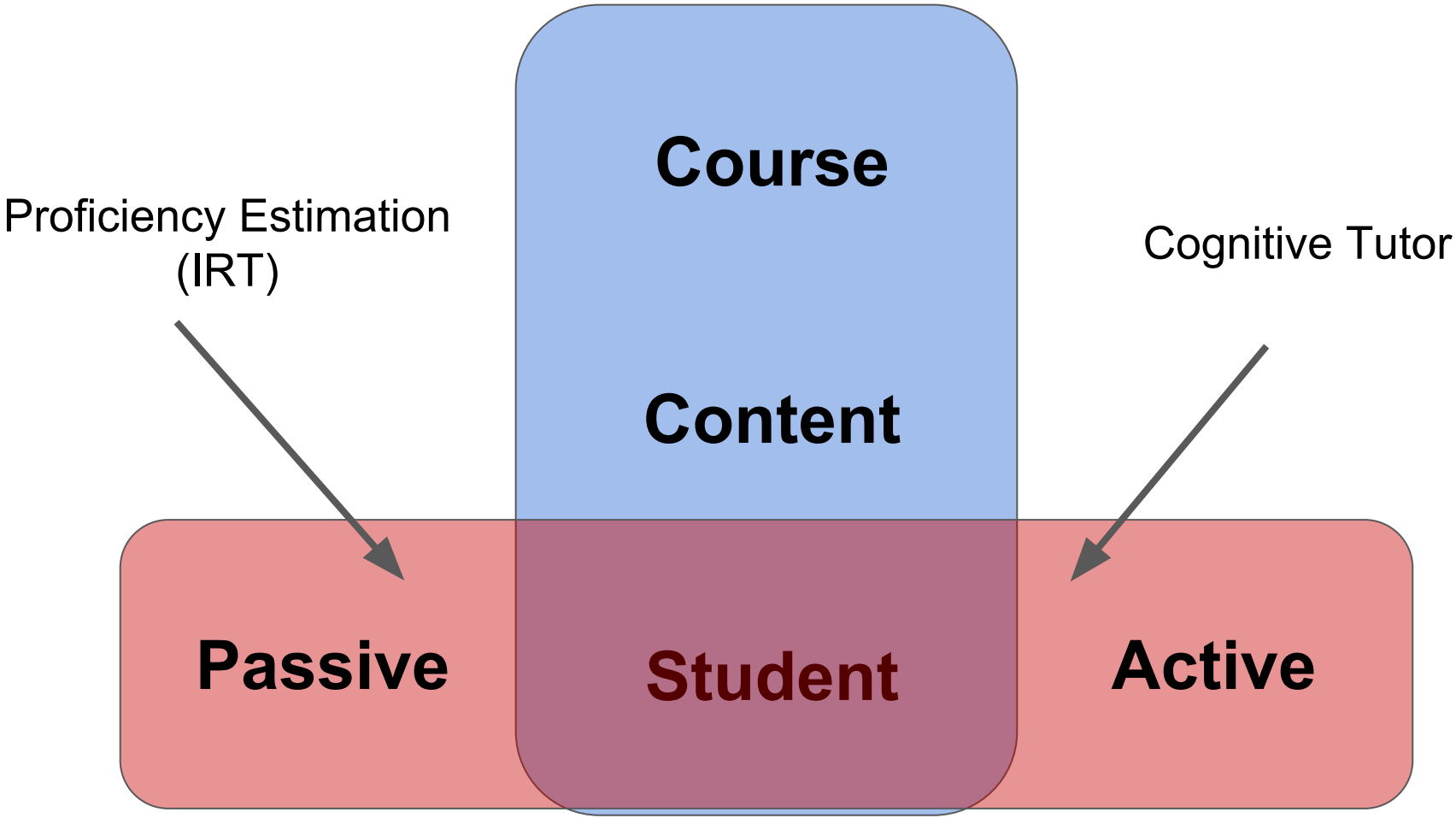
Framework for AI in Online Education



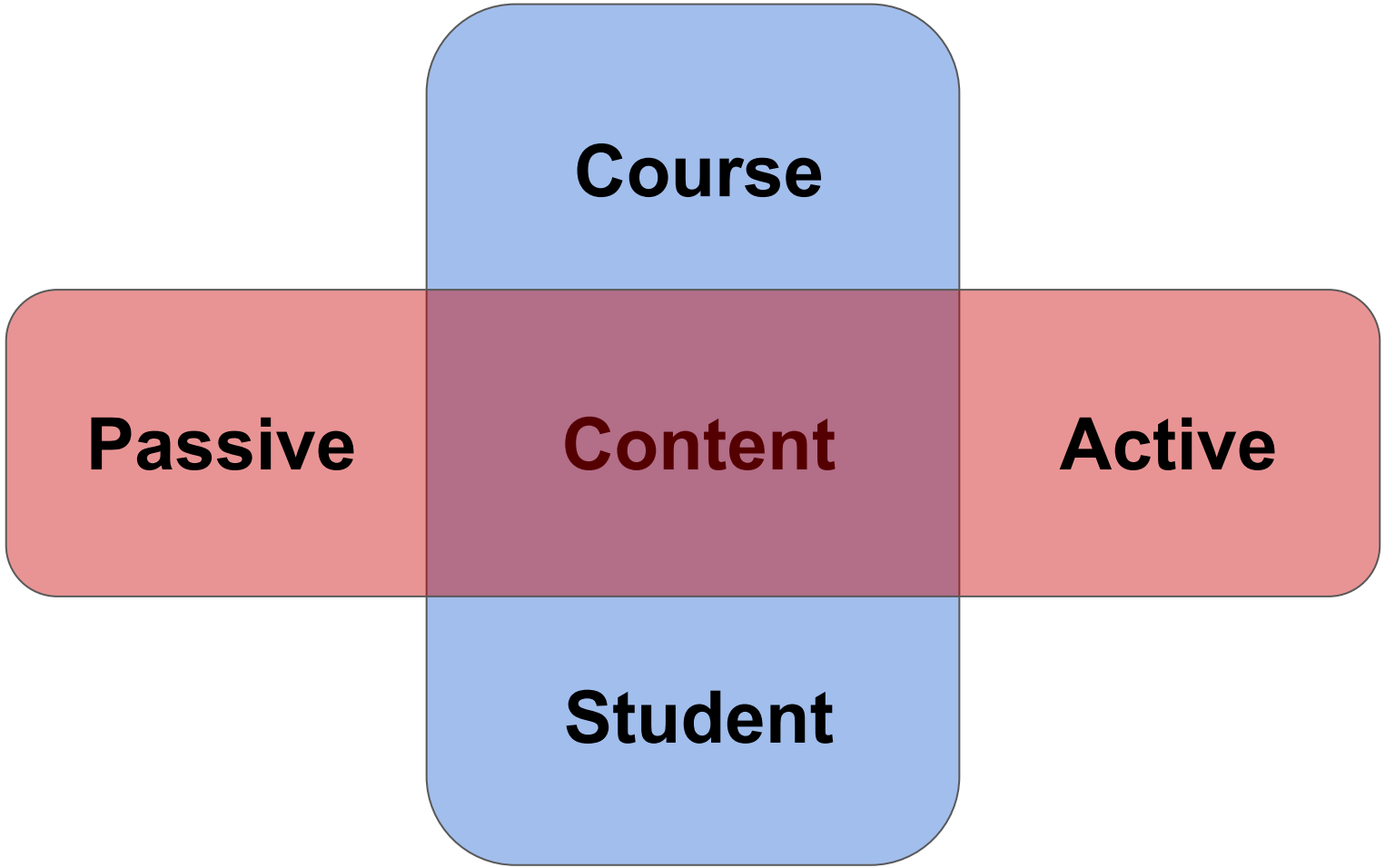
Framework for AI in Online Education



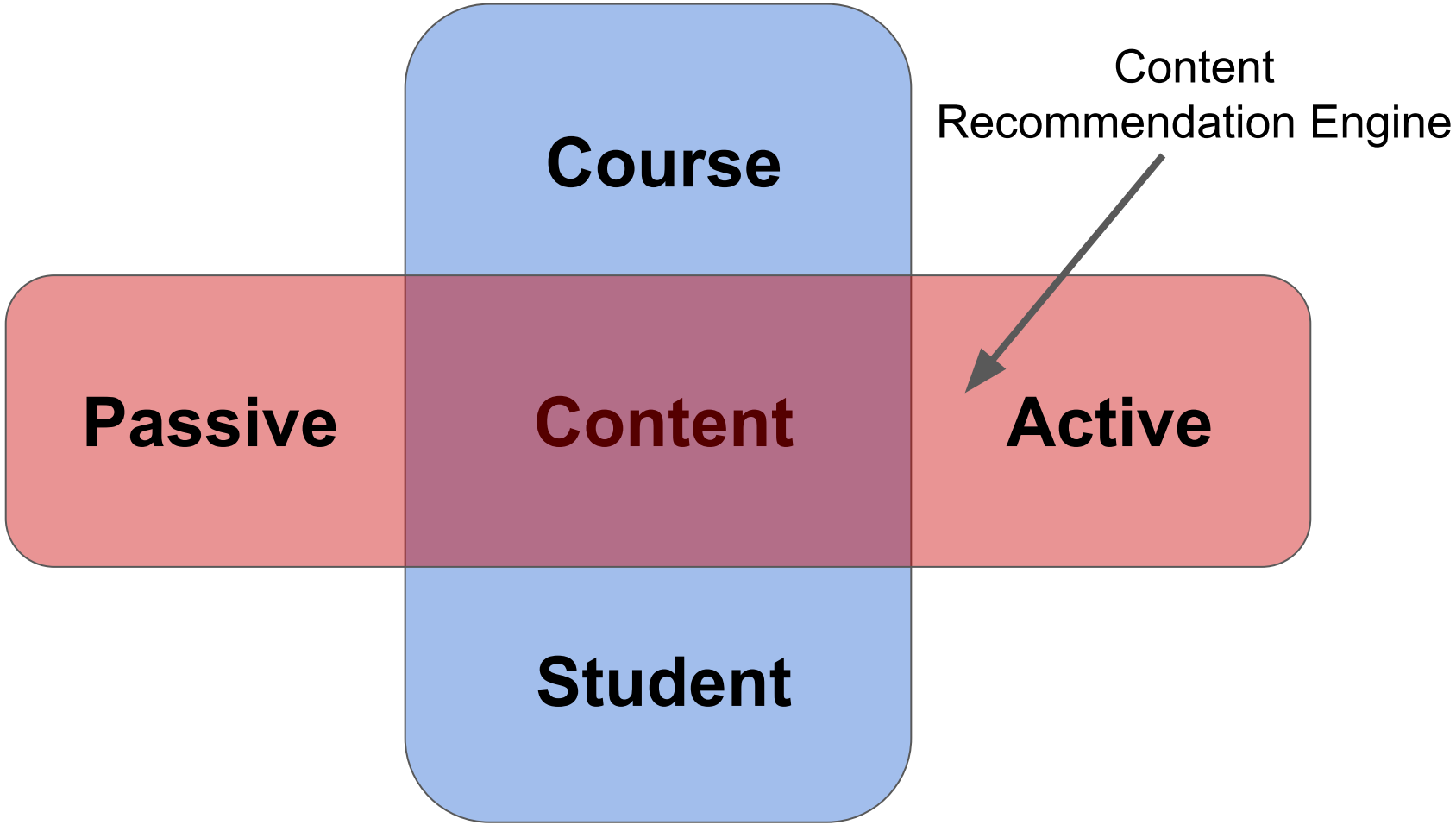
Framework for AI in Online Education



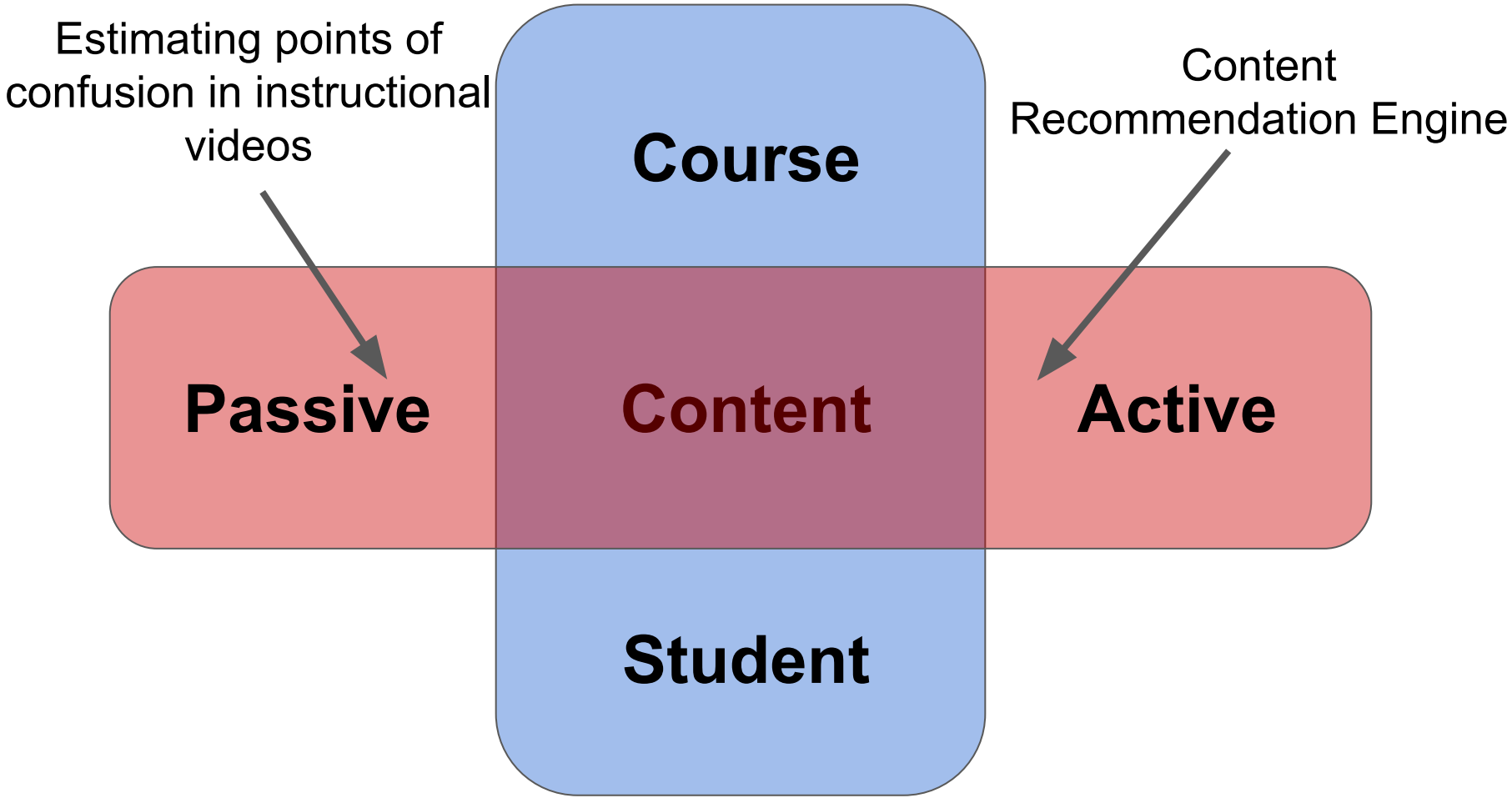
Framework for AI in Online Education



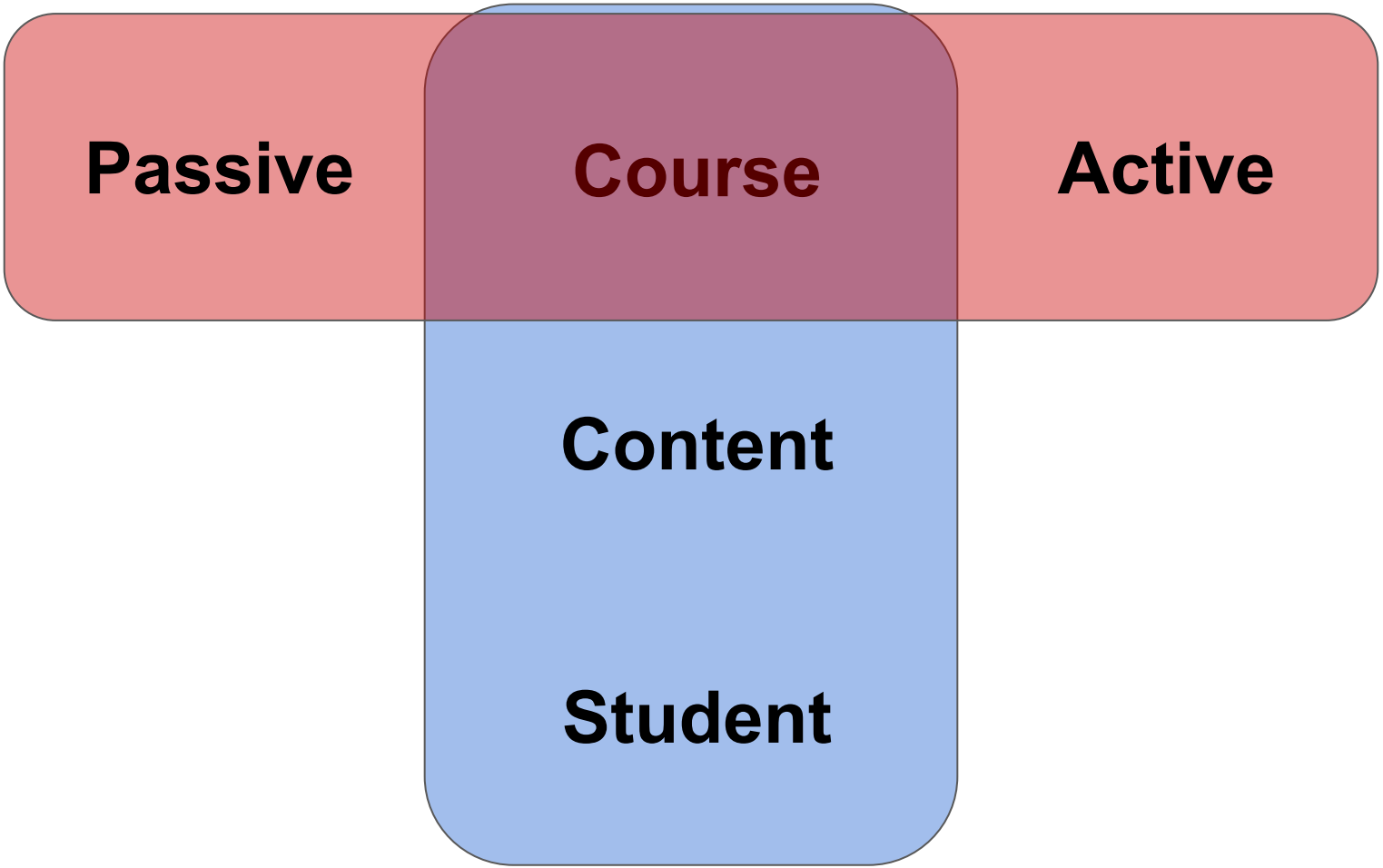
Framework for AI in Online Education



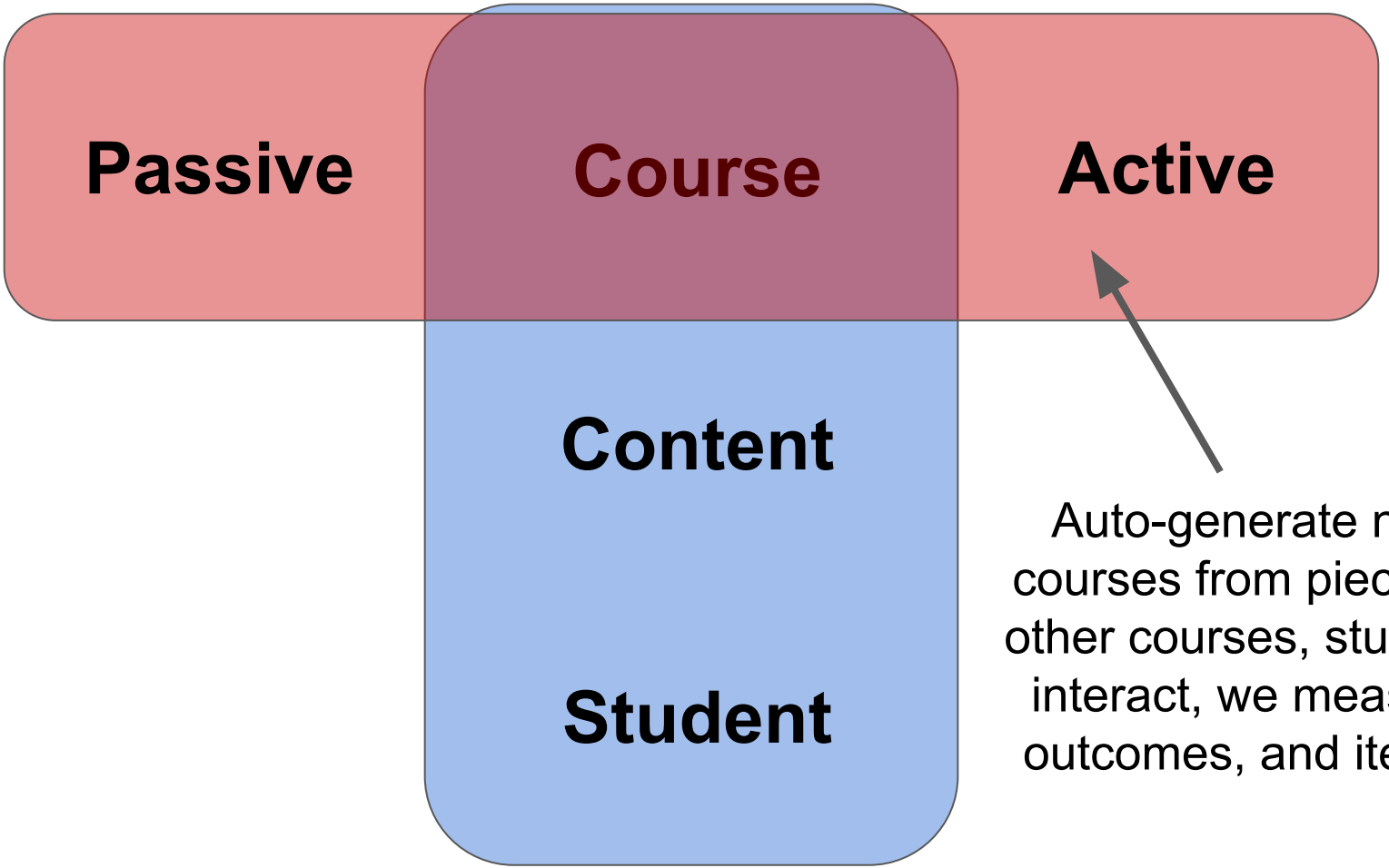
Framework for AI in Online Education



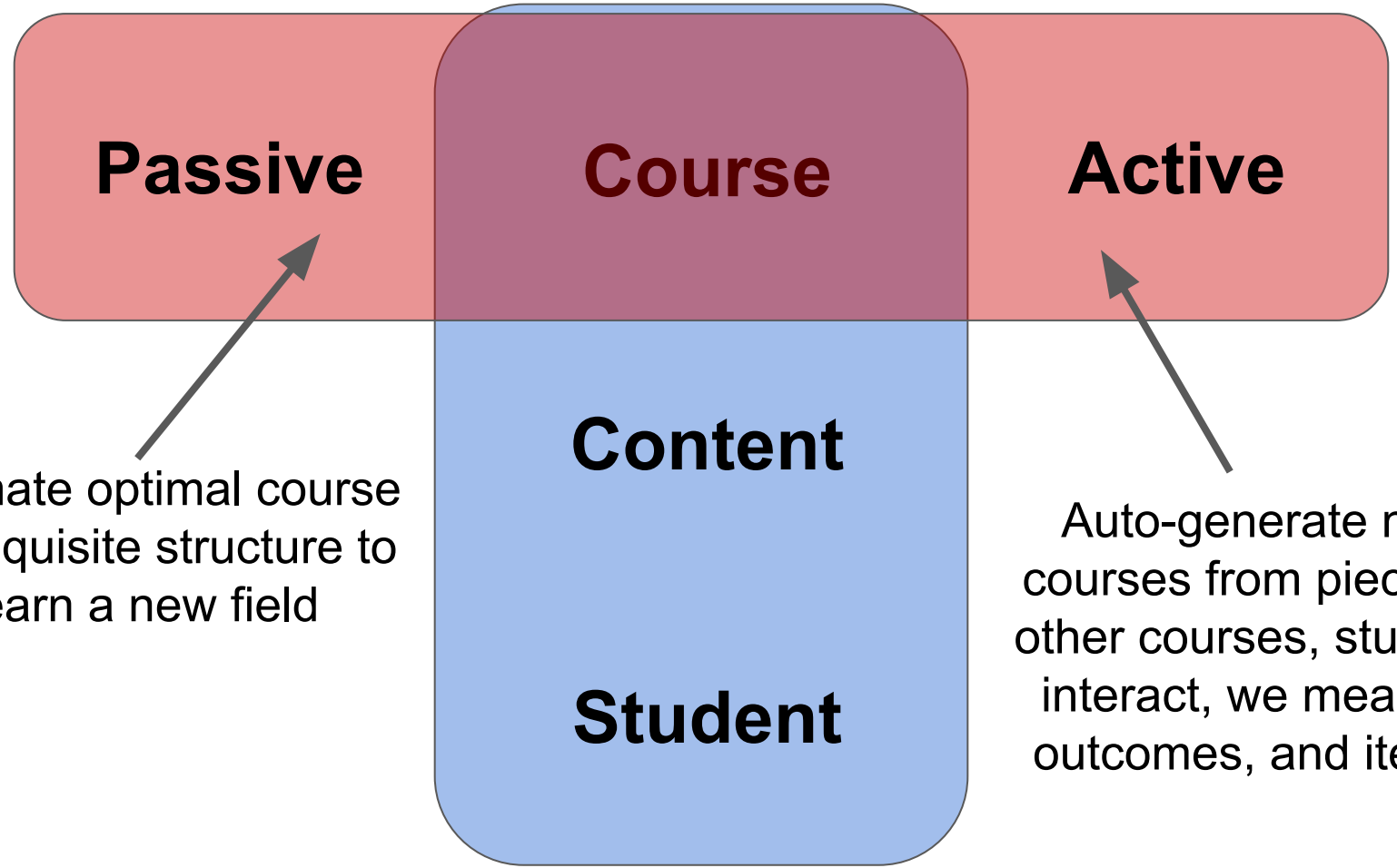
Framework for AI in Online Education



Framework for AI in Online Education



Framework for AI in Online Education



Example: active → passive

Students can ask for info about their progress (active). Based on students who do well, we can learn which information to provide students that is likely to increase learning (passive).

For the rest of these slides, we'll consider 18 examples of how to improve

online learning with
machine learning

⟷

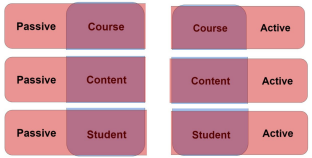
human intelligence with
artificial intelligence

1

Problem: Representation

How do we…

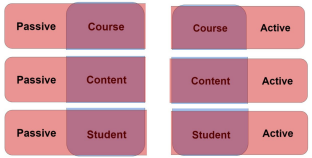
- represent courses as vectored-data?
- know if two courses/videos/students are similar/different?
- recommend content to students?
- match students with other students?



1

Solution: Representation (Embeddings)

1. Course-level embeddings
2. Content-level embeddings
3. Student-embeddings

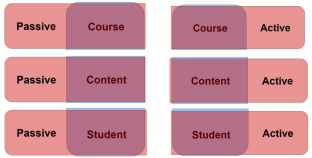


1

Solution: Representation

(How to generate embeddings)

Capture courses/content/students interactions as **high-dimensional** feature matrices. Reduce dimensionality using **PCA, SVD**, etc or use **hidden weights** of a **neural network**.



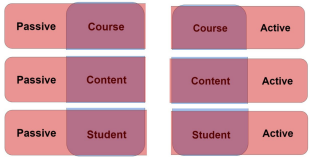
1

Solution: Representation

(How to generate embeddings)

"Embedded" dense low-dim representations enable:

1. Generation of new courses from existing courses
2. Student pairing and inference
3. Content structuring

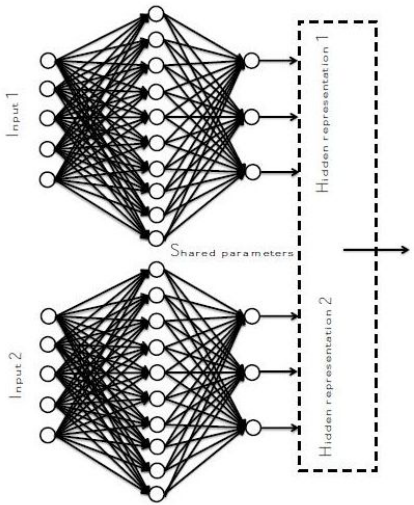


1

Recommendation Engine for MOOCs using Embeddings

Siamese neural network architecture.

Paper → http://www.mit.edu/~jonasm/info/MuellerThyagarajan_AAAI16.pdf

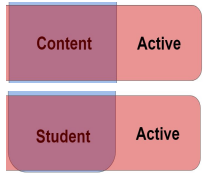


One network - **students**

Other network - **content**

Produces embeddings s.t.

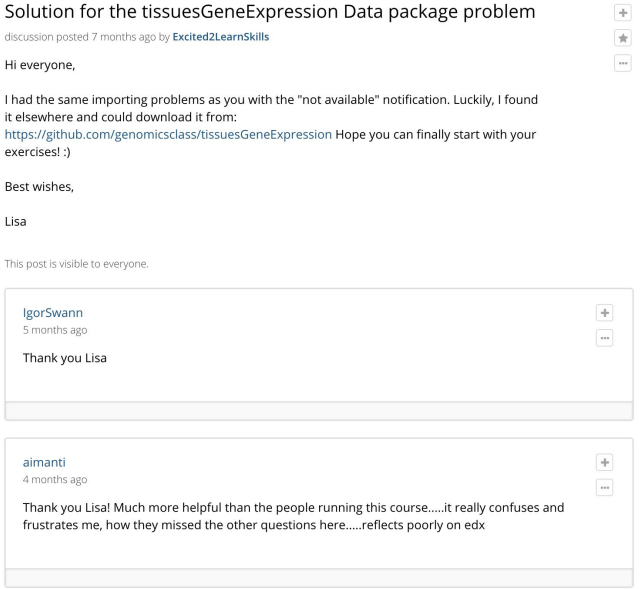
cosine(student,content) = how good content is for student

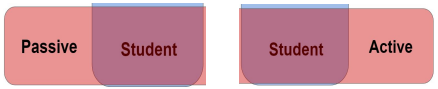


2

Problem: Detect struggling

Alert course teams of learners who are struggling with the material.

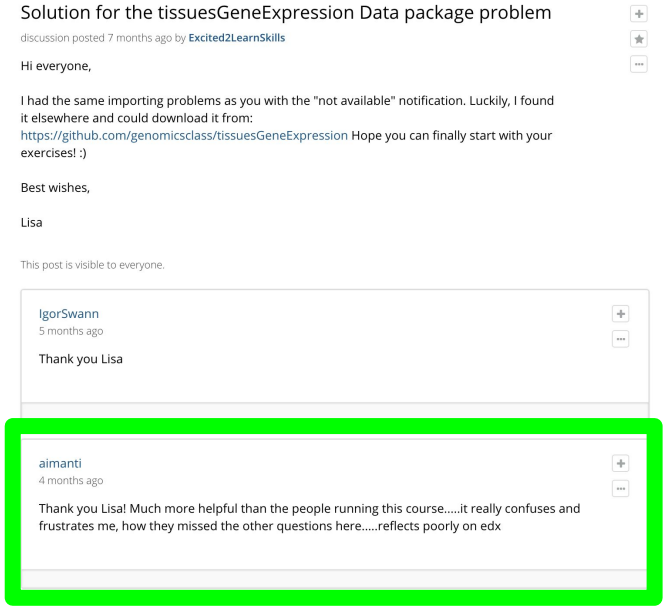


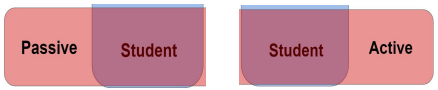


2

Problem: Detect struggling

Alert course teams of learners who are struggling with the material.

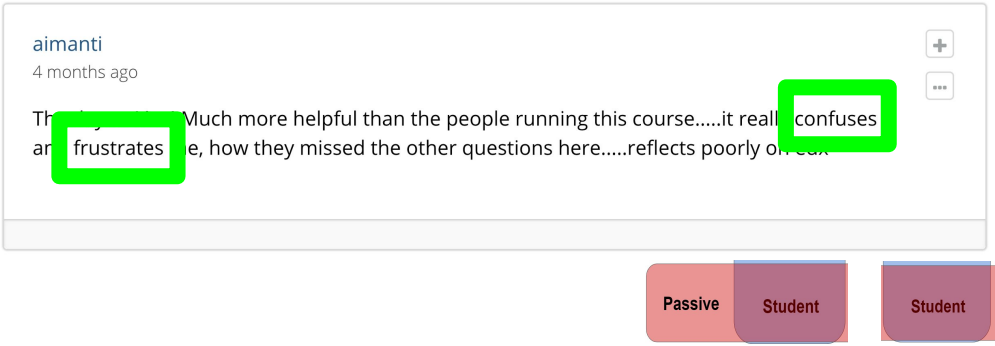




2

Problem: Detect struggling

Alert course teams of learners who are struggling with the material.

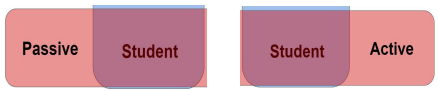


2

Solution: Detect struggling
(Noisy labels + Rank Pruning)

Paper -> Learning with Confident Examples: Rank Pruning for Robust Classification with Noisy Labels | Northcutt, Wu, Chuang (2017)

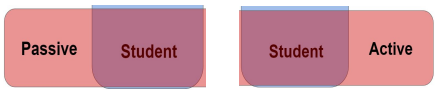
1. Train classifier
   a. student interaction features *X*
   b. forum posts containing “confused” and “frustrated” as labels *y*



2

Solution: Detect struggling
(Noisy labels + Rank Pruning)

Account for noisy labels using Rank Pruning (next slide).



The Rank Pruning Algorithm

Northcutt, Wu, Chuang (UAI 2017)
http://auai.org/uai2017/proceedings/papers/35.pdf

① Confident Examples ② Estimate Noise Rates ③④ Denoise(Prune) & Refit

$$\tilde{P}_{y=1} = \{x \in \tilde{P} \mid g(x) \geq LB_{y=1}\} \rightarrow s=1, y=1$$
$$\tilde{N}_{y=1} = \{x \in \tilde{N} \mid g(x) \geq LB_{y=1}\} \rightarrow s=0, y=1$$
$$\tilde{P}_{y=0} = \{x \in \tilde{P} \mid g(x) \leq UB_{y=0}\} \rightarrow s=1, y=0$$
$$\tilde{N}_{y=0} = \{x \in \tilde{N} \mid g(x) \leq UB_{y=0}\} \rightarrow s=0, y=0$$

Confident examples → agreement of predicted probability and training label.

$$LB_{y=1} := P(\hat{s}=1 \mid s=1) = E_{x \in \tilde{P}}[g(x)]$$
$$UB_{y=0} := P(\hat{s}=1 \mid s=0) = E_{x \in \tilde{N}}[g(x)]$$
$$g(x) = P(\hat{s}=1|x)$$

$$\rho_1 = P(s=0|y=1),\ \rho_0 = P(s=1|y=0)$$

$$\hat{\rho}_1^{conf} := \frac{|\tilde{N}_{y=1}|}{|\tilde{N}_{y=1}| + |\tilde{P}_{y=1}|},\ \hat{\rho}_0^{conf} := \frac{|\tilde{P}_{y=0}|}{|\tilde{P}_{y=0}| + |\tilde{N}_{y=0}|}$$

$$\rho_1 = \frac{\#(s=0, y=1)}{\#(s=0, y=1) + \#(s=1, y=1)}$$

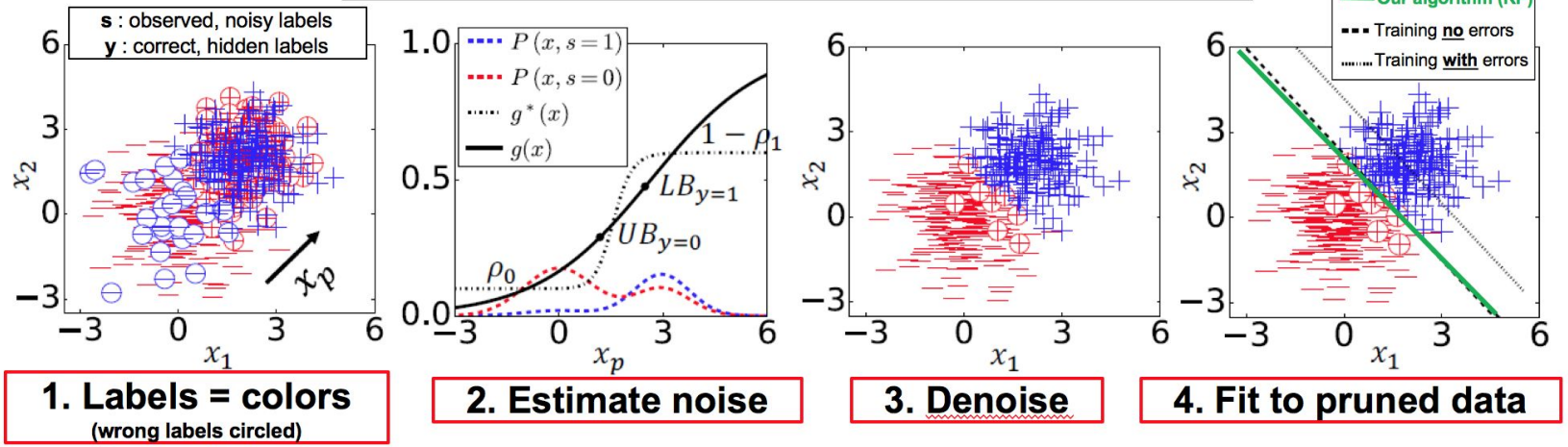
$\pi_1 = P(y=0|s=1),\ \pi_0 = P(y=1|s=0)$

$k_1 :=$ $(\hat{\pi}_1|\tilde{P}|)^{th}$ smallest $g(x)$ for $x \in \tilde{P}$.
$k_0 :=$ $(\hat{\pi}_0|\tilde{N}|)^{th}$ largest $g(x)$ for $x \in \tilde{N}$.
BFPRT ($\mathcal{O}(n)$) to compute k_1 and k_0.

$$\tilde{P}_{conf} := \{x \in \tilde{P} \mid g(x) \geq k_1\}$$
$$\tilde{N}_{conf} := \{x \in \tilde{N} \mid g(x) \leq k_0\}$$

Train with $\tilde{P}_{\text{conf}} \cup \tilde{N}_{\text{conf}}$

$s=0, y=0$ $s=0, y=1$ $s=1, y=0$ $s=1, y=1$



1. Labels = colors (wrong labels circled)

2. Estimate noise

3. Denoise

4. Fit to pruned data

3 Problem: Rampant cheating

The rampant multiple-account cheating (around 10% of completers) problem poses a **serious threat** to the promise of democratization of education and to accredited MicroMasters programs.

3 Solution: Rampant cheating

(CAMEO on steroids + Rank Pruning + HITT)

- Generate 100+ statistical matching features based on timing, ip address, etc. and combine five filter-based algorithms to produce noisy labels.
- Fix with Rank Pruning using a Neural Network.
- Validate and combine results with HITT labels (unpublished).

Source: MIT S.M. thesis, Curtis G. Northcutt (2017)

4 Problem: Stats Collaboration

In general, we'd like to know which two accounts are most similar (cheating, working together, same person, etc.). To answer this, we can answer the question - **which pairs of accounts are drawn from the same distributions?**

4

Solution: Stats Collaboration

Measure how different distributions P and Q are using Maximum Mean Discrepancy or Wasserstein Distance based on work by **Aruther Gretton** (Univ College London) - Two-sample testing

Paper (sec. 8): http://www.gatsby.ucl.ac.uk/~gretton/papers/GreBorRasSchetal12.pdf

5

Problem: How to personalize

Most machine learning models focus on class-level prediction, and do not consider that each student has their own predictive model. What are some tricks to deal with this?

5

Solution: How to personalize

- Active Learning
- Domain Adaptation
- Gaussian Processes with class-level priors and student-level training
- DNN with class-level initial weights and student-level additional training
- 100s of sources. Check out Oggi Rudovic (MIT)

6

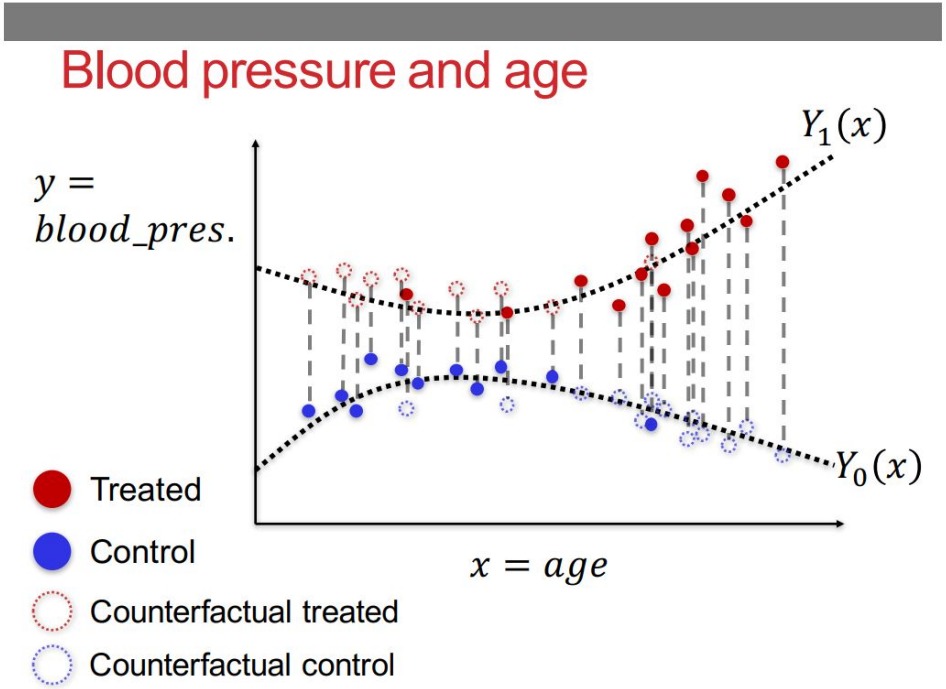
Problem: ITE/Counterfactuals

The standard for A/B testing is we apply a treatment/condition to students and measure class-level difference between the two conditions on some outcome of interest (grade, IRT proficiency, etc.).

We can do better than this… What we really want is the **ITE = independent treatment effect**, but we need the counterfactuals.

6 Solution: ITE/Counterfactuals

(iterative ite estimation)



Slide by Uri Shalit & David Sontag, taken from: http://www.cs.nyu.edu/~shalit/slides.pdf

6 Solution: ITE/Counterfactuals
(iterative ite estimation)

For education, we measure grade/proficiency/time as the outcome and the treatment is an A/B condition, learning condition, etc. Sources:

Learning Representations for Counterfactual Inference, Johansson, Shalit, Sontag (2016)
Estimating individual treatment effect: generalization bounds and algorithms, Shalit, Johansson, Sontag (2017)

6

Solution: ITE/Counterfactuals
(iterative ite estimation)

A version of algorithm for Online Education:

1. Train a classifier on treatment group A and group B separately.
2. Predict B outcome for group A, and vice versa.
3. Retrain with new info until convergence

7

Problem: Trajectory prediction

Will a student dropout? Perform exceedingly well? Any outcome of your choice...

7

Solution: Trajectory prediction

- Trajectory outcome prediction
  - Reliable Decision Support using Counterfactual Models Schulam and Saria (2017)
- Learning Models from Observational Traces
  - Takes into account treatments - tell students they might fail (interventions)
  - Dropout prediction is trajectory based
- Counterfactual GPs for inference from traces
  - Marked point processes (MPP)
  - Gaussian Process (GP)

8

Problem: How to order content

Course staff orders problems, videos, and topics using their expertise, which often works well. What would be better? Ordering content in a way to **maximize learning outcomes** (increase in proficiency, decrease in time spent)

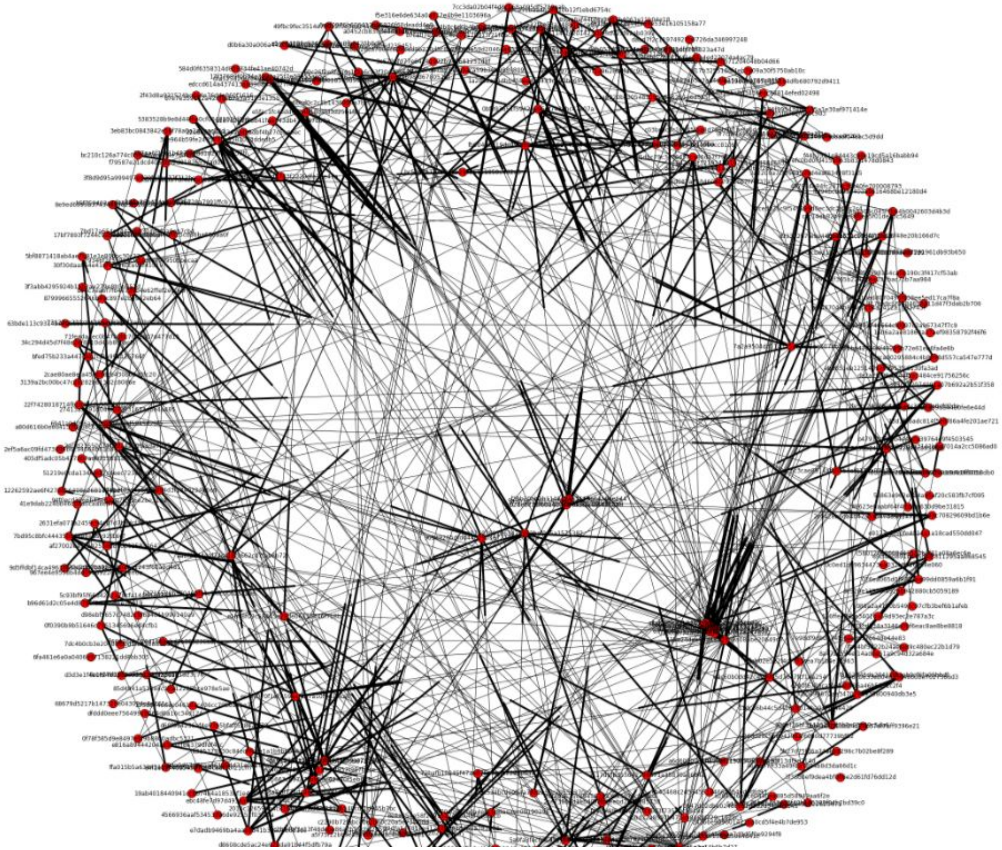
8 Solution: How to order content (mutual information)

For all pairs of problems, determine

- Prob(answer A correctly | understand B)
- Prob(answer A incorrectly | understand B)
- Prob(answer A correctly | not understand B)
- Prob(answer A incorrectly | not understand B)

From these we can derive directional mutual information: **the prerequisite nature of A → B**

8 Solution: How to order content



Passive | Content

9 Problem: Adaptive Learning

Although embeddings, siamese networks, sentiment analysis LSTMs, and other solutions presented in this deck can be used for adaptive learning, another approach is to use a knowledge graph and proficiency model, but how?

Content Active

9

Solution: Adaptive Learning (Knewton and IRT)

Knewton:

- Proficiency Model via IRT
- Knowledge Graph (propagate knowledge based on a **given** prerequisite graph of concepts.

Content Active

10

Problem: Google Scholar

Student's try to increase their knowledge using Google Scholar, but can we develop an intelligent platform that understands the types of scholarly articles and what we really want when we search?

Content Active

10

Solution: Semantic Scholar (maybe)

AI2 (Allen Institute for Artificial Intelligence) and others are actively working on more intelligent information transfer.

Content Active

11

Problem: Majority bias in forums

Its 1600s and we are in a MOOC. The MOOC instructor asks 50k students, “What shape is the earth?” to which the majority of students respond, “It is flat”, +1, thumbs up, liking each others’ posts. The minority arguing for a round model.. Their answers are never seen, down-ranked far below where anyone will read.

Content Active

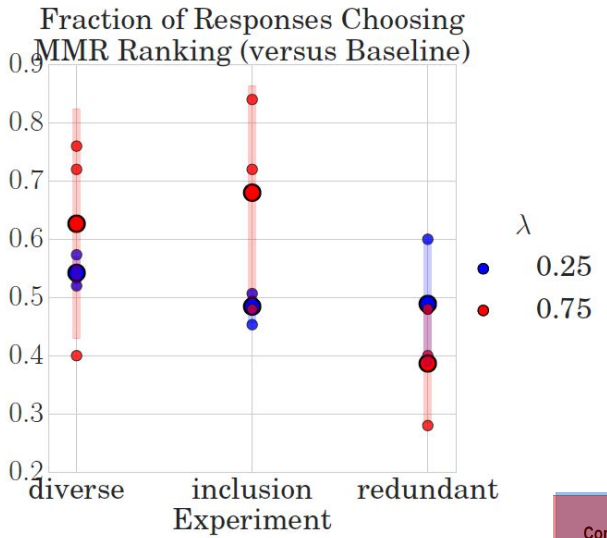
11

olution: Majority bias in forums

(diversification of comment rankings)

Paper -> Comment Ranking Diversification in Discussion Forums, Northcutt et al. (2017)

```
Trial 11

|------------|
| Question Q |
|------------|
'Whether you are deeply familiar with Christianity or new to the tradition, please share 1-3 things...

|--------|
| List A |
|--------|

[1] 'Christianity was explained in just 4 minutes. Very good video.'

[2] "Interesting that so many listed as making up part of Christianity don't recognize each other as such!"

[3] "I appreciated the point about how Eastern Christianity emphasizes the Incarnation...  "

[4] 'Christianity is based not on the writings of Jesus Christ, but the writings of others...'

[5] 'I have always found it interesting that Christianity has both condemned and enabled oppression. '


|--------|
| List B |
|--------|

[1] 'Christianity was explained in just 4 minutes. Very good video.'

[2] 'I found it interesting that some missions working in other cultures recognized the active presence of God...

[3] "Interesting that so many listed as making up part of Christianity don't recognize each other as such!"

[4] 'Christianity is based not on the writings of Jesus Christ, but the writings of others...'

[5] 'Orthodox is newer to me.   I was raised in a Pentecostal church and converted...  '
```



Content | Active

12

Problem: Feature Extraction

How do we get good features from student data?

12

Solution: Feature Extraction

- Examples (Instructor's toolkit / student weathermap/instrument panel)
  1. Item Response Theory graphs
  2. Student Pathway Fingerprint (2-D voronoi)
  3. Sorted Order Most Common Responses
  4. Student Pathway State Space Diagram
  5. IFACAs
  6. Max Correct Sliding Window Curve - MEPTI
  7. Average Correct Sliding Window Curve
  8. Histogram of time-between-correct-responses
  9. Histogram of time-from-viewing-to-submitting
  10. Time maps - https://districtdatalabs.silvrback.com/time-maps-visualizing-discrete-events-across-many-timescales
  11. Scatter plot of total time spent versus percent correct, red X for user, blue dots for everyone elseZ-scores
- Then train using known labels and find out which features are most predictive of that outcome. Use those features in the future.

13

Problem: Cognitive State

If students lose interest, they don't learn.

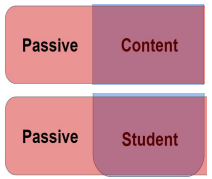
Can we detect student affective and cognitive states?

13

Solution: Cognitive States

(AutoTutor: Affective Computing)

- Roz Picard in Media Lab.
- Uses webcam and facial tracking
  - AutoTutor Detects and Responds to Learners Affective and Cognitive States, Mello et al. (2008)
- AI part: We can use these labels to learn which content leads to boredom and which types of students tend to get bored.



14

Problem: Content Likeability

Non-AI solution - students can rate content. But the **problem** is that how they rate it **depends** on their **knowledge** and **what content they saw previously**.

14

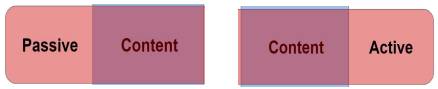
Solution: Content Likeability:
(Sentiment Analysis LSTM)

- Train a Sentiment Analysis LSTM Using Noisy Crowd Labels
  - https://github.com/HazyResearch/snorkel/blob/master/tutorials/crowdsourcing/Crowdsourced_Sentiment_Analysis.ipynb
- If you have user sentiment labels (active AI) → easy.
- No labels? Use a proxy like sentiment = sum of time spent by learners who got the right answer / time by those who got it wrong.

15

Problem: Points of Confusion

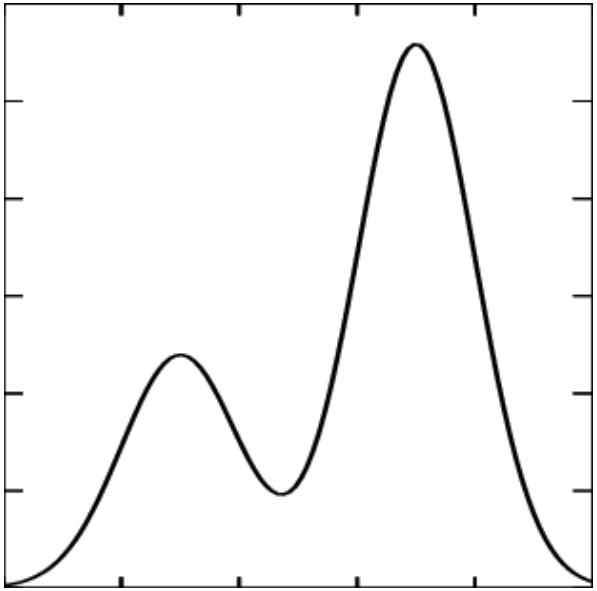
In which locations do students become confused? Can we identify them and notify instructors?



15

Solution: Points of Confusion

(Video interaction Density estimation)



Passive Content

Content Active

16

Problem: Cognitive Modeling

We'd like to understand what the student knows to better serve them content and optimize for learning constraints.

16

Solutions: Cognitive Modeling

- ITS (Intelligent Tutoring Systems)
  - Gautam Biswas, Ken Koedinger, Vincent Aleven (many papers)
- Bayesian Knowledge Tracing
  - Albert Corbett (Baker, Koedinger)
  - More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing, Baker, Corbett, Aleven (2009)
- Cognitive Tools / Agents / Tutors
  - Hundreds of papers and authors

17

Problem: Human Intelligence vs. Artificial Intelligence

We don't yet fundamentally understand how humans learn, but we can see when human learning matches up with artificial learning. Given a paragraph of text and student-specific features, can we predict a numerical score for how "much" the student will learn from the text?

17

Solution: HI vs. AI

- Generate a million versions of an instructional paragraph by toggling a few words each time. Choose the paragraph that "maximizes learning."
- With better models of human learning, we can have better AI models for education.
- For other work in this context (but different scenarios)
  - Patrick Winston (MIT) - Human Intelligence Enterprise
    i. http://groups.csail.mit.edu/genesis/HIE/white.html

18

Problem: The next edX

How do we reach the lofty goal of democratization of education affordably, realistically, achieving both openness and value?

Course Active

18

Part of a possible solution

Virtual Reality with four components:

1. ML
2. ASR
3. VR
4. Pedagogy

Course Active

Machine Learning Data Expressed in OE Terminology

ML Topic	Data Type	Online Education Concept
Support Vector Machines	Support vectors	Borderline student/content/course between two conditions
Active Learning	High variance examples	Student/content/course which greatly differs from other students/content/courses
Image Recognition	Hard Negatives	Student/content/course that is likely to be misclassified
Rank Pruning	Confident Examples	Student/content/course we can confidently estimate belongs to a group of interest

Other Resources

Papers

Two different ways to estimate time-dependent individual treatment effect via counterfactual prediction:
1. **What-if reasoning with Counterfactual Gaussian Processes** *(Peter Schulam*, Suchi Saria, 2017)
2. **A deep-learning approach - Structured Inference Networks for Nonlinear State Space Models** *(Rahul G. Krishnan, Uri Shalit, David Sontag, 2017)*

One way to "*embed"* content / students / courses (into low dimensional vector representations) from textual (i.e. NLP) representation as input:
1. Starspace - Embed all the things! (FAIR, 9/12/17)

Videos/Slides

1. 2 hours - Suchi Saria - Modern healthcare methods all super applicable to education
2. 350 slides - Uri Shalit and David Sontag - Counterfactual Methods for Observational Studies