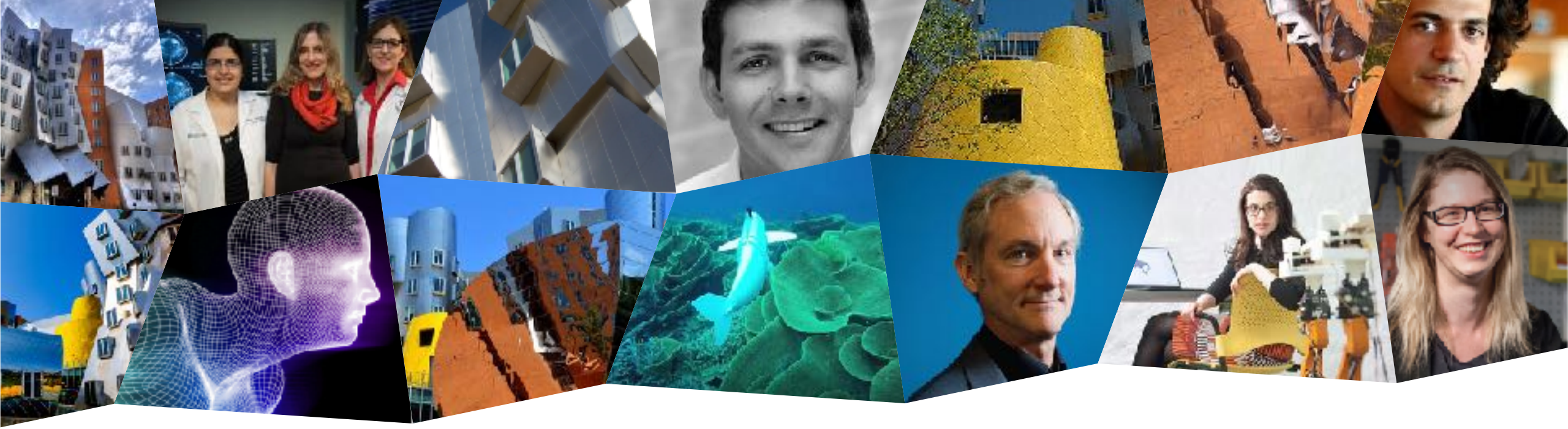


Causal Concept Explanations for Deep Neural Models

Josh Rountree[1] · Pulkit Verma[2] · Oswin So[3] · Chuchu Fan[3] · Julie Shah[1]

MIT CSAIL[1] · IIT Madras[2] · MIT LIDS[3]

CCW-Net: Causal Concept Wrapper Network

- Do a model's explanations cause its outputs?
- Uses human-meaningful concepts as explanations
- Trains for causal faithfulness and measures it

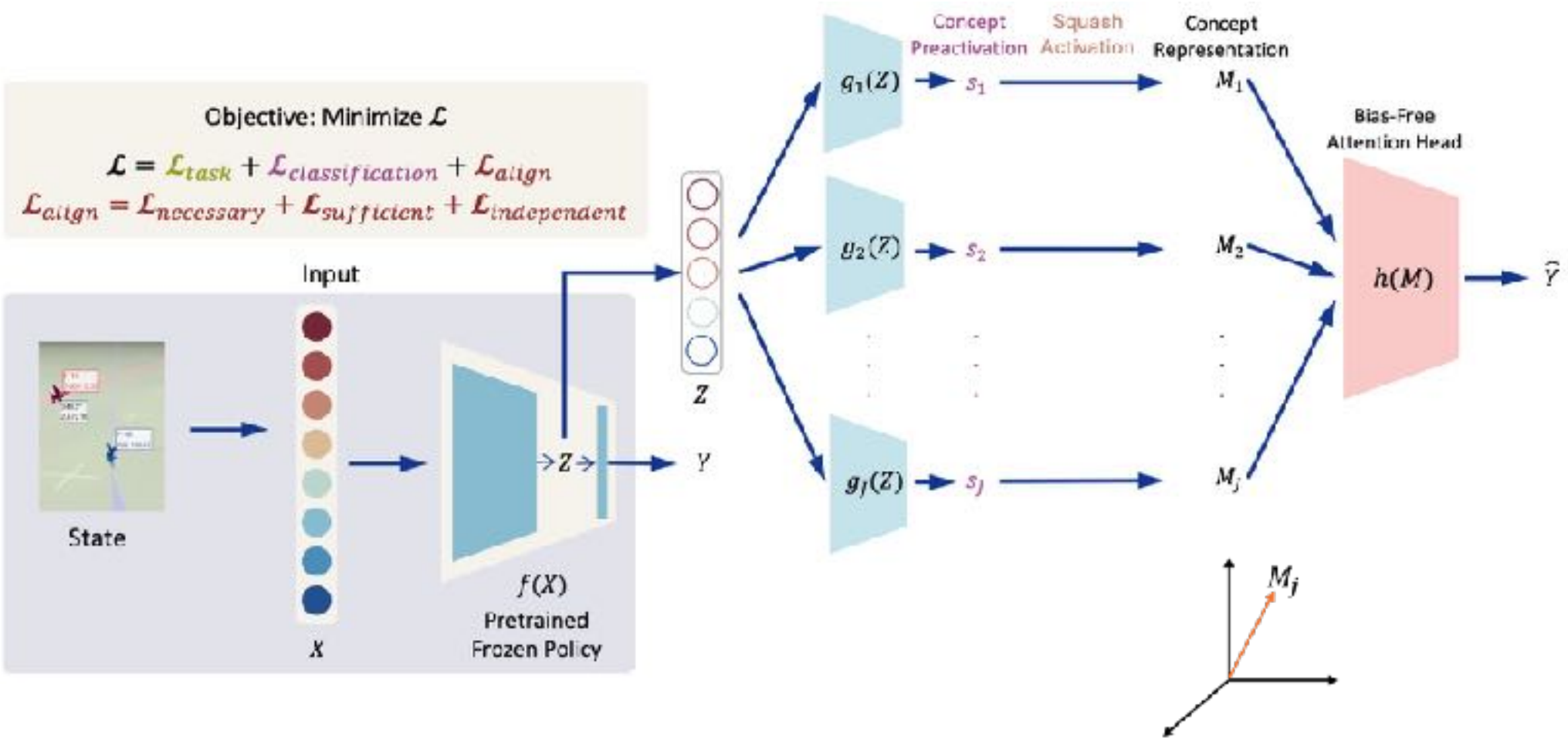
Correct looking explanations can be causally wrong.
Autonomous vehicles: Does a "passenger pickup" concept activation simply correlate with the actions?

Correlational Explanations
Concepts co-occur with scenes and model outputs
$p(y \mid c)$ $p(c \mid y)$

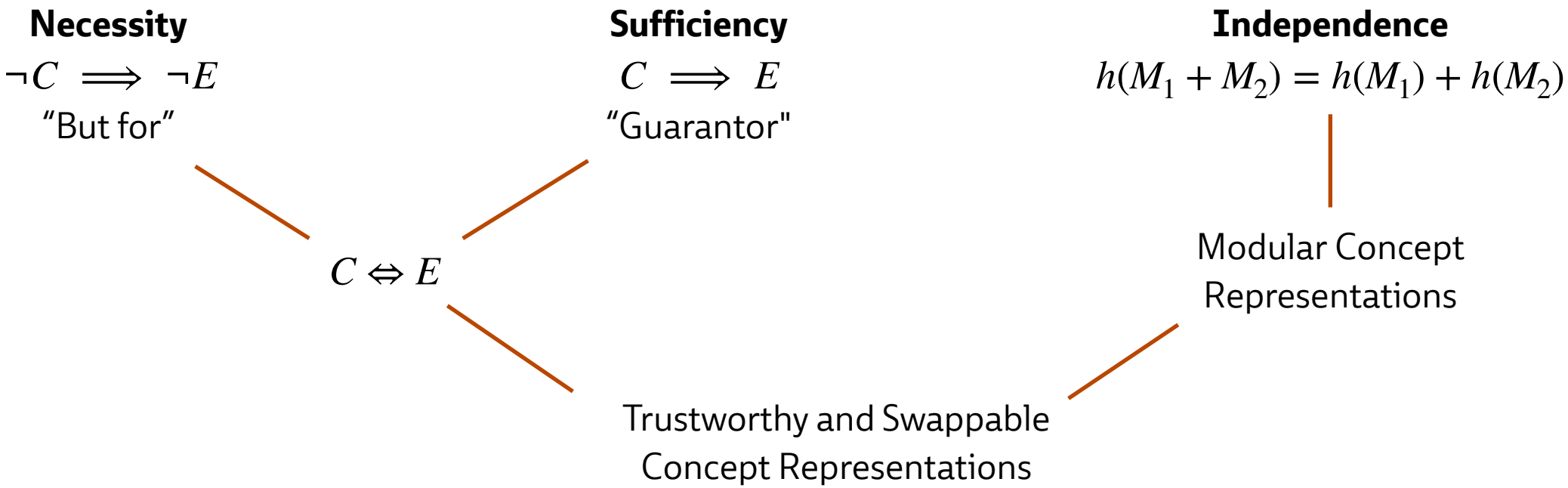


Causal Explanations
Concepts causally drive model outputs
$p(y \mid do(c))$

Method - Architecture



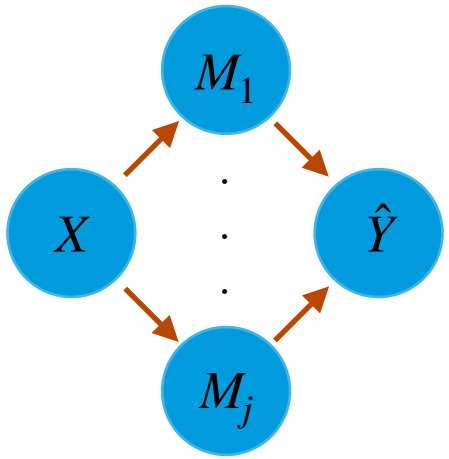
Method - Training



Method - Training

Primary causal measurement: Change in model output, $\hat{Y}$, caused by changes in concept representation, M_j:

$$\Delta^0 = \hat{Y} - h(\mathbf{0}) \qquad \Delta_j^0 = h(\mathbf{0}_{j \leftarrow M_j}) - h(\mathbf{0}) \qquad \Delta_j^{cf} = h(M_{j \leftarrow \tilde{M}_j}) - \hat{Y}$$



Convert necessity, sufficiency, and independence into loss terms:

$$\mathscr{L}_{\text{nec}} = \mathbb{E}\Big[\sum_{j:c_j=0} \|\Delta_j^{\text{cf}}\|^2\Big]$$

Inactive concepts should not change output
$\neg C \implies \neg E$

$$\mathscr{L}_{\text{suf}} = \mathbb{E}\Big[\|h(\mathbf{0}) + \sum_{j:c_j=1} \Delta_j^{\mathbf{0}} - Y\|^2\Big]$$

Active concepts should produce expert output
$C \implies E$

$$\mathscr{L}_{\text{ind}} = \mathbb{E}\Big[\|\Delta^{\mathbf{0}} - \sum_{j} \Delta_j^{\mathbf{0}}\|^2\Big]$$

Concept representations should contribute additively
$h(M_1 + M_2) = h(M_1) + h(M_2)$

Method - Evaluation

Interventional Indirect Effect (IE) of concept j:

$$\mathrm{IE}_j = \mathbb{E}\left[h\left(M_j^{\mathscr{B}}, M_{-j}\right) - h\left(M_j^{\mathscr{A}}, M_{-j}\right)\right]$$

Fraction of total effect caused by concept j

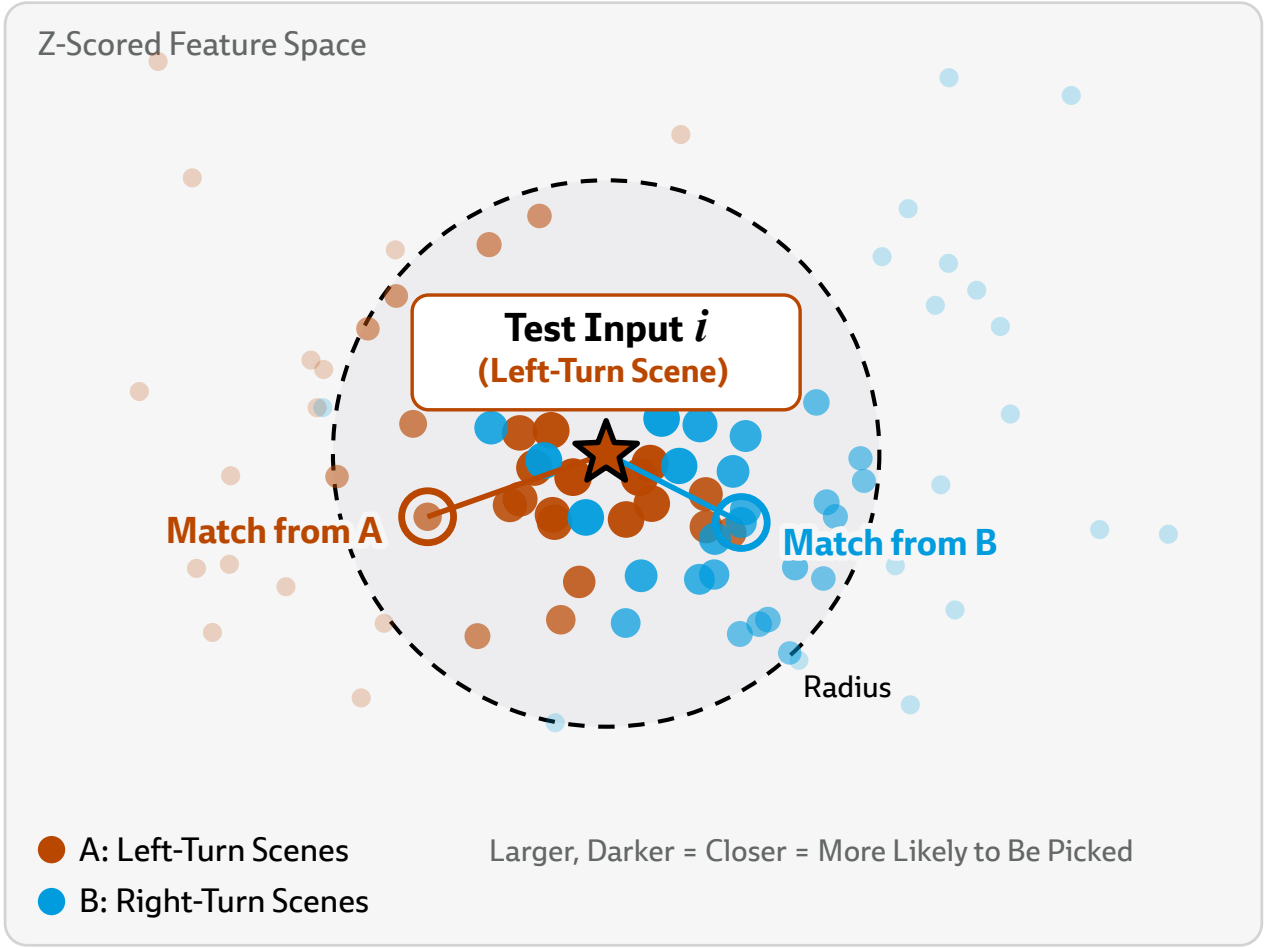
$$\rho_j = \frac{\|\mathrm{IE}_j\|}{\sum_k \|\mathrm{IE}_k\|}$$

Fraction of total effect caused by active concepts

$$\rho_{\text{active}} = \sum_{j:c_j=1} \rho_j$$

Fraction of total effect caused by inactive concepts

$$\rho_{\text{inactive}} = \sum_{j:c_j=0} \rho_j$$



Selected Results

How much of the effect flows through active concepts?

Aircraft Control

94.5%

CCW-Net (ours)

50.0% Prior concept wrapper approach

CarRacing

89.8%

CCW-Net (ours)

33.1% Prior concept wrapper approach

CUB-200

91.3%

CCW-Net (ours)

41.6% Prior concept bottleneck approach

CCW-Net's explanations are more faithful than prior art.

Selected Results

Accurate or disentangled concepts are insufficient for causal faithfulness.

Prior Wrapper on Aircraft Control

Concept Accuracy

91.2%/98.2%

Effect via active concepts:

50.0%

Concept Bottleneck on CUB-200

Concept Accuracy

97.7%

Effect via active concepts:

41.6%

Prior Wrapper on Aircraft Control

DCI Disentanglement

0.998

Effect via active concepts:

50.0%

Causal faithfulness is its own property.
We provide a method to design for and measure it.

Causal Faithfulness can be Trained for and Measured

1. Mediation analysis becomes a training objective for deep networks.
2. Vector concepts with zero-valued absence make interventions well defined.
3. Evaluation method quantifies each concept's causal share of the output



Thank you!


INDIAN INSTITUTE OF TECHNOLOGY MADRAS


MIT
CSAIL
Computer Science &
Artificial Intelligence
Laboratory


MIT
LIDS