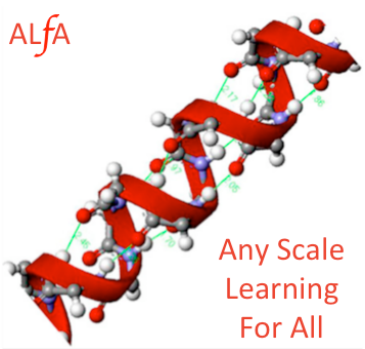
Engaging Big Data and Cloud Computing for Large Scale Machine Learning and Modeling



The EC-Star Platform

Una-May O'Reilly
Computer Science and Artificial Intelligence Lab
MIT

EvoDesignOpt

genetic FINANCE

李嘉誠基金會 LI KA SHING FOUNDATION

CSAIL

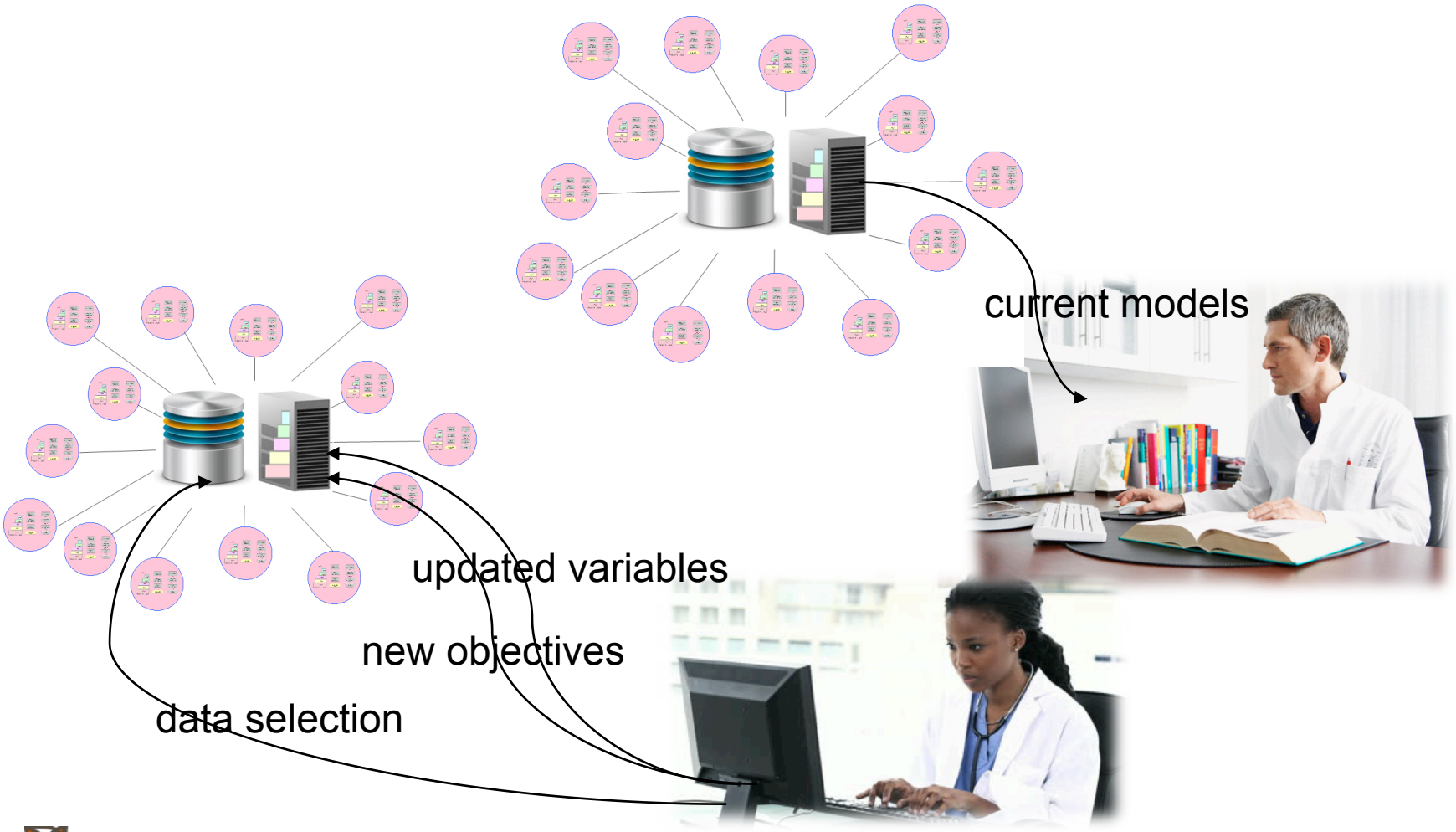
EC-Star

- is very well suited to knowledge mine "bigData"
- generates interpretable models
- performs scalable ML/quantitative modeling
- recasts ML/quantitative modeling

EvoDesignOpt

CSAIL

Digital Directed Evolution of Models



EvoDesignOpt

Use Case #1

CSAIL

EC-Star

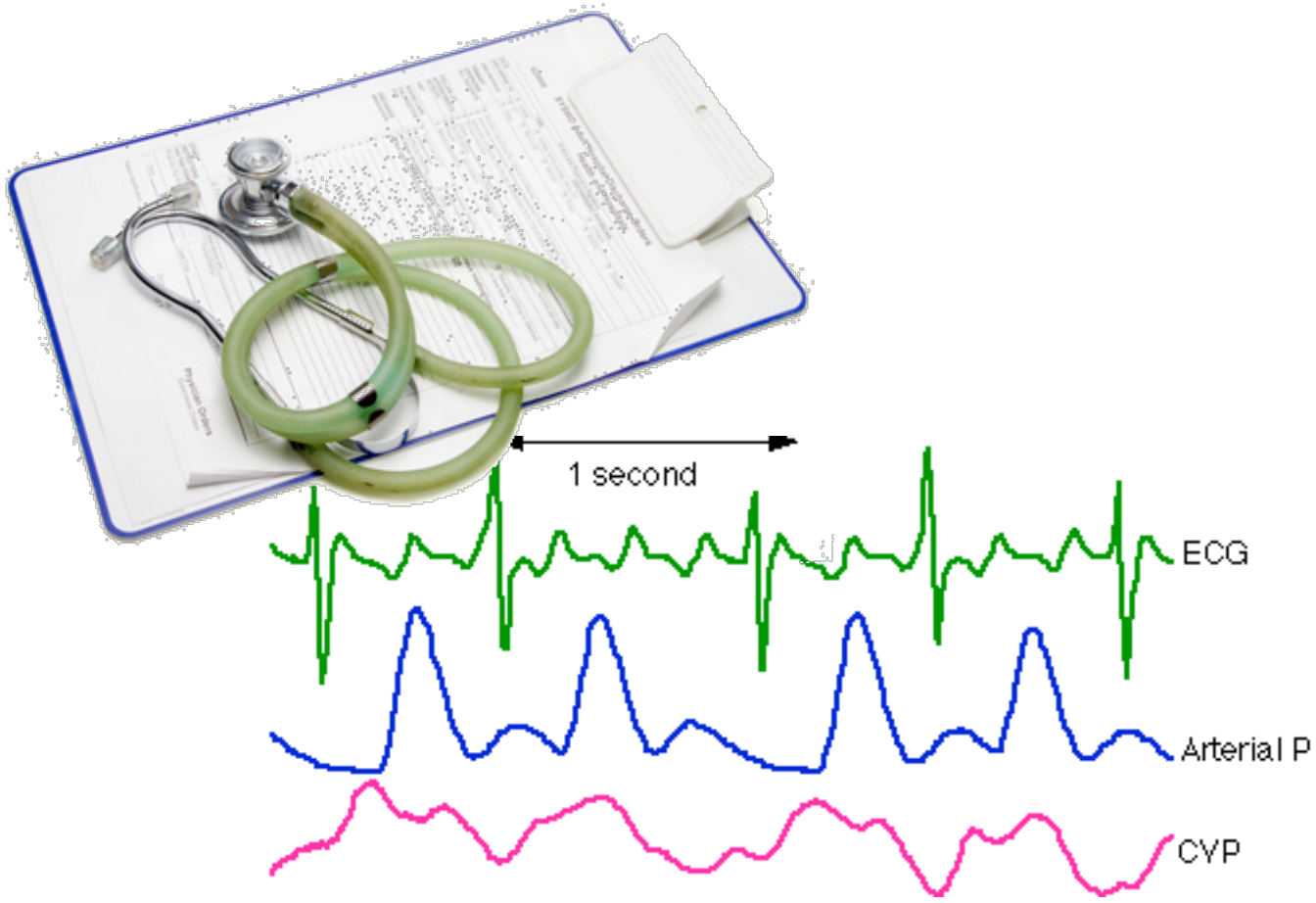
- **recasts ML/quantitative modeling**
  - **it's a sustained or continuously learning system that is capable of executing for extended durations**
  - **as it is running, we can manipulate its data, features and objectives it presently considers**
  - **As it is running, we can collect its current best models and externally test and interpret them**
  - **This allows us to derive more than one explanation for your data. ''Philosophy of pluralism''**
    - **To mine for multiple correlations between the explanatory variables and their response.**

EvoDesignOpt

CSAIL

EvoDesignOpt

CSAIL



EvoDesignOpt

CSAIL

EC-Star

- Is a platform we can run
  - On a laptop, workstation, cluster, or cloud
  - On dedicated or idle resources
  - Elastically: shrink and expand its available resources

- Current status:
  - Genetic Finance deployment on 100's thousands of computer resources for financial trading
  - Ongoing development for predictive modeling, in general:
    - » domain: physiological times series data
    - » Future: add clinical intervention and records data
  - Ready for another pilot project?

EvoDesignOpt

CSAIL

Agenda

- How EC-Star represents its strategies, rules, models
  - They are INTERPRETABLE
- EC-Star's algorithmic approach is
  - Local and global evolutionary computation
    - » SCALABLE
- Envisioned Use Cases
  - digital "directed evolution" of models
  - Many Minds – collaborative, scientific knowledge discovery
- Examples of EC-Star learning
  - Blood pressure forecasting

EvoDesignOpt

CSAIL

What's in an ML Algorithm's Representation?

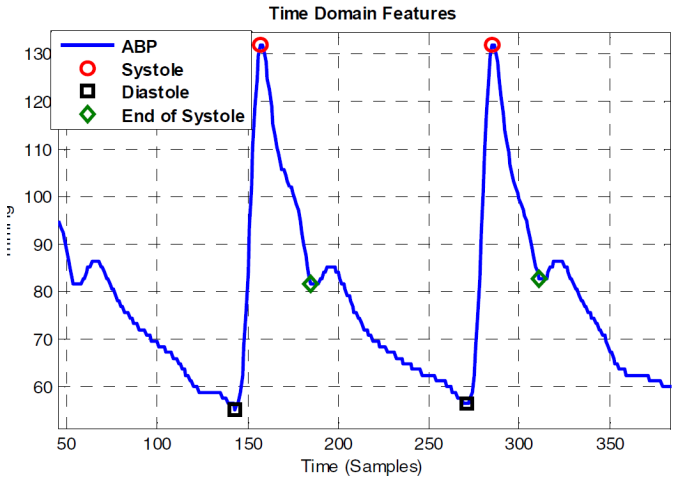
- **Models! ..quantitative, data-driven models**
  - **Graphical model**
  - **Neural network**
  - **Support vectors**
  - **Decision tree -> rule**
- **EC-Star uses rule sets**
  - **provides freedom to express some relationships completely flexibly, constrain others,**
  - **Linear or non-linear**
  - **after the fact:**
    - » **readability,**
    - » **sensitivity analysis,**
    - » **confidence measures**

EvoDesignOpt

CSAIL

EC-Star's Rule Set Representation

- Rules in a rule set
- Rule has LHS, RHS
- LHS is a list of condition tests
  - All must be true for rule to be fired
- RHS is the prediction, label, strategy



SBP: Systolic blood pressure – Max
DBP: Diastolic blood pressure - Min
PD: Pulse Duration
MAP: Mean arterial pressure

PD@t-12 ≥ 68.0 & MAP@t ≤ 89.0 → normal

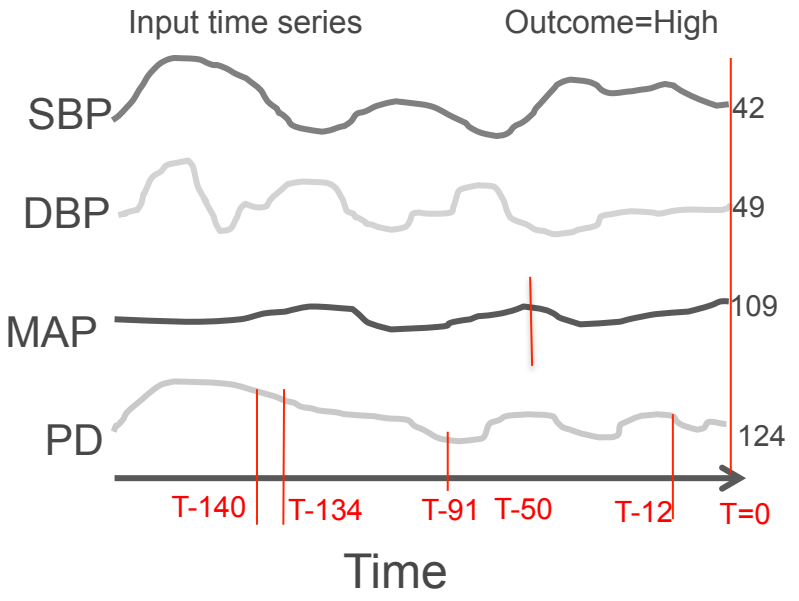
SBP@t ≥ 32.0 & DBP@t ≥ 50.0 & PD@t ≥ 124.0 → normal

- LHS: a variable at lag X is greater than another value
  - » EC-Star can discover this X
- Additional conditions
  - RHS: Introspection: self- reference to a previous prediction, while predicting over a duration

EvoDesignOpt

CSAIL

Rule Set Interpretation



SBP: Systolic blood pressure – Max
DBP: Diastolic blood pressure - Min
PD: Pulse Duration
MAP: Mean arterial pressure

Rule set:

MAP@t ≥50.0 & ! MAP@t ≥ 89.0	→ low
PD @t-12 ≥68.0 & !MAP@t≤89.0	→ normal
PD@t-91 ≥68.0 & !PD@t-140≥120.0	→ low
MAP@t-50≥ 90.0 & !MAP@t≥112.0	→ high
PD@t-134 ≥ 68.0 & PD@t≥118.0	→ high
SBP@t≥ 32.0 & DBP@t ≥50.0 & !PD@t≥ 124.0	→ normal
DBP@t ≥50.0 & !PD@t≥106.0	→ low
!DBP@t≥ 40.0& PD@t ≥ 68.0	→ low

Prediction bids

High	2
Normal	1
Low	0

$Outcome_p$=High

EvoDesignOpt

CSAIL

What's in a Machine Learning Algorithm's Method

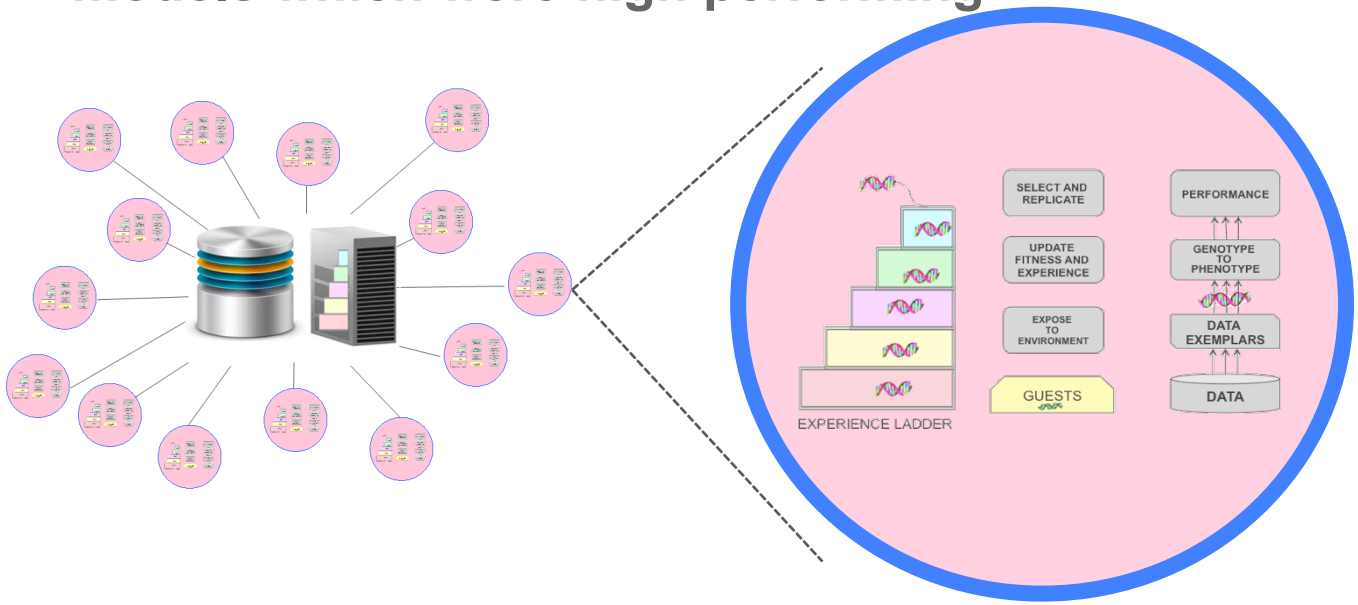
- **Needs to be scalable**
  - **Handle bigData, exploit big Compute**
  - **EC-Star uses evolutionary computation and distributed data handling**
- **Needs to be effective**
  - **Evolutionary algorithms are well used in Engineering, Finance…Science?**
- **Needs to support a pluralistic modeling approach**
  - **There is more than one explanation**
  - **any explanation can be refined, qualified**

EvoDesignOpt

CSAIL

Overview

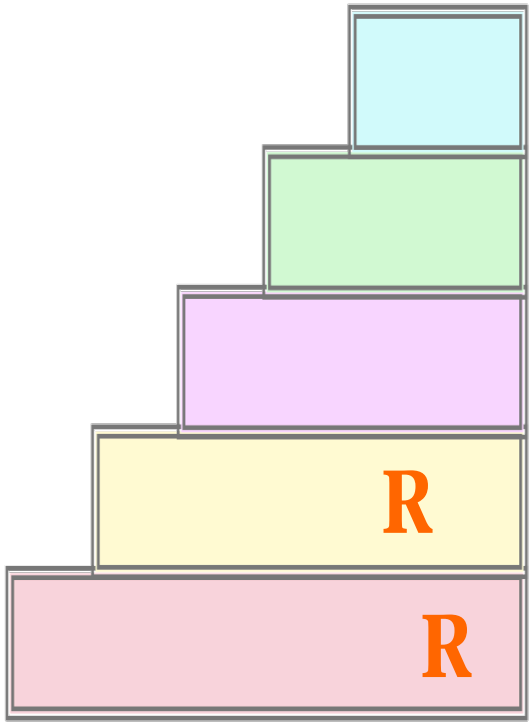
- Evolutionary computation is ARTIFICIAL
- Local learners
  - Replicated stochastic algorithm, uniquely randomized data
  - They learn independently by sampling a portion of the data
  - They “evolve” models that progressively use ancestral models which were high performing



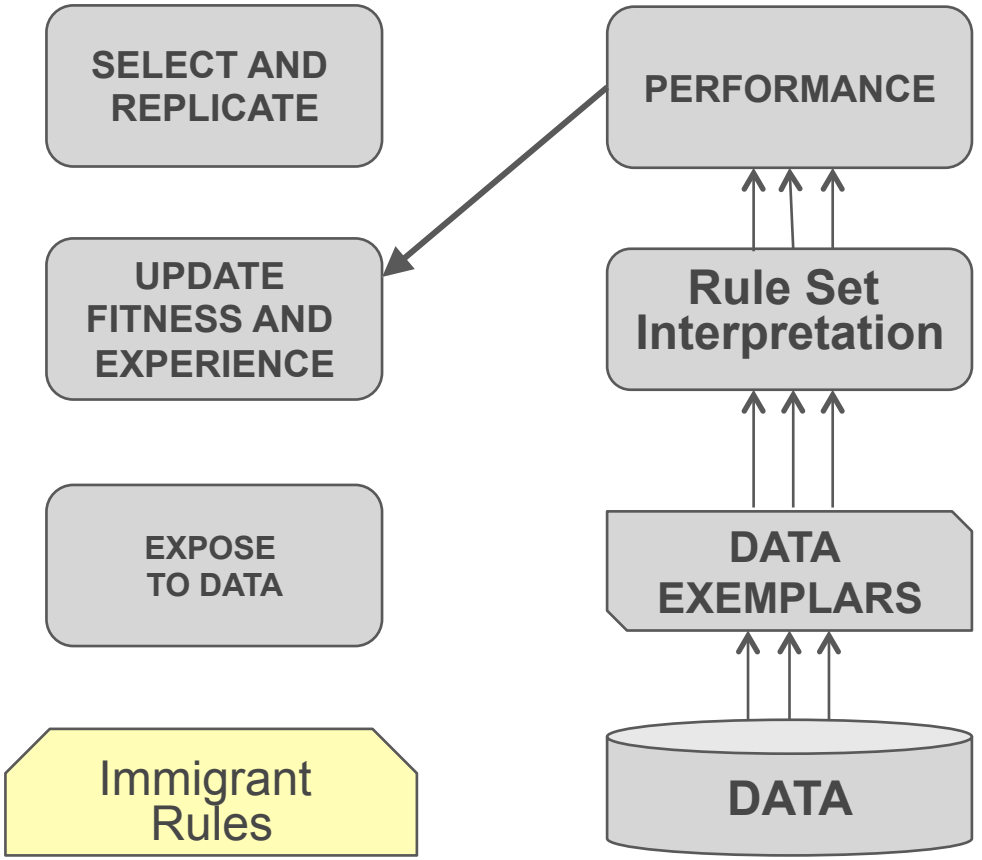
EvoDesignOpt

CSAIL

Interpret New RuleSet



EXPERIENCE LADDER

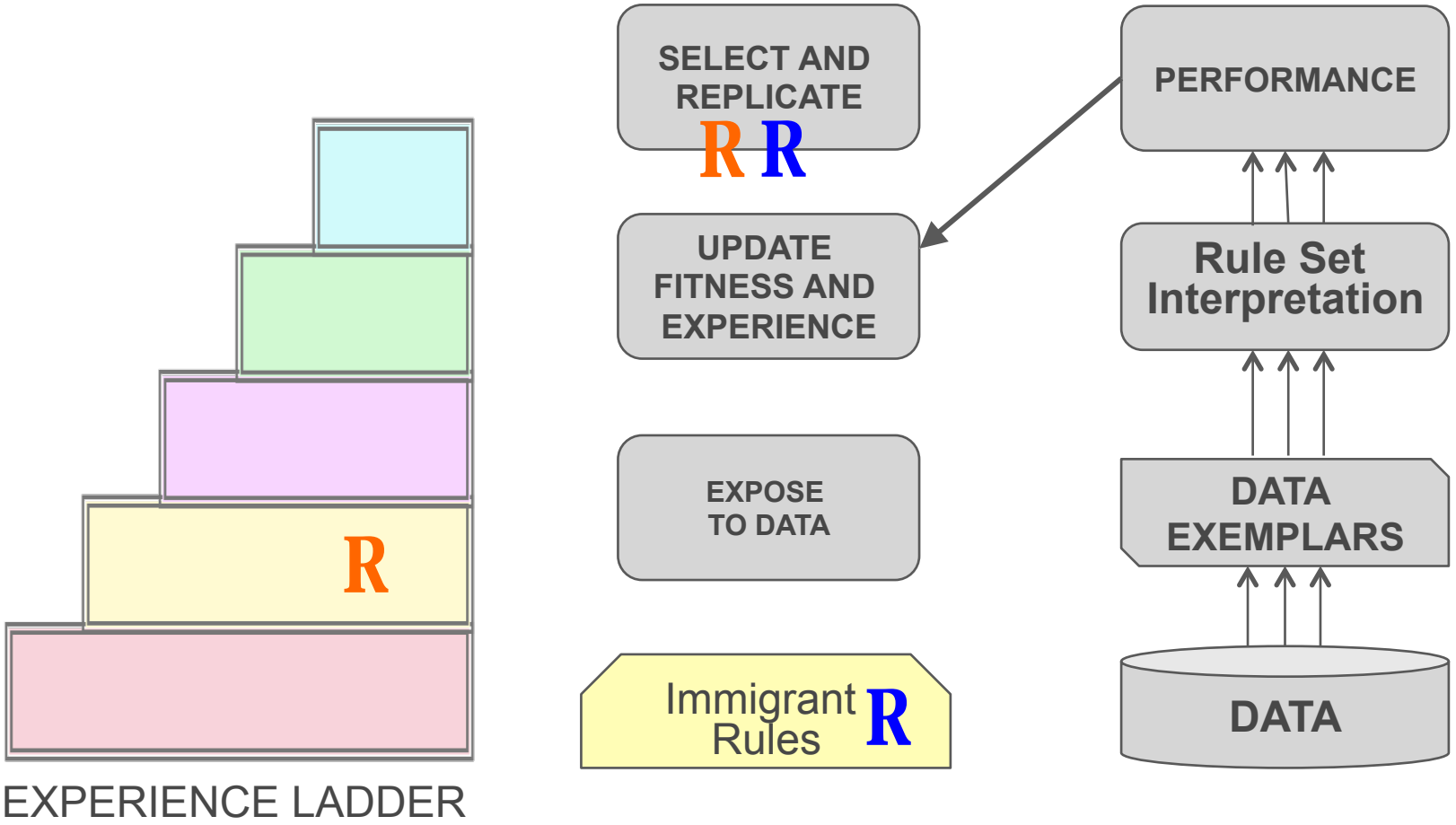


EvoDesignOpt

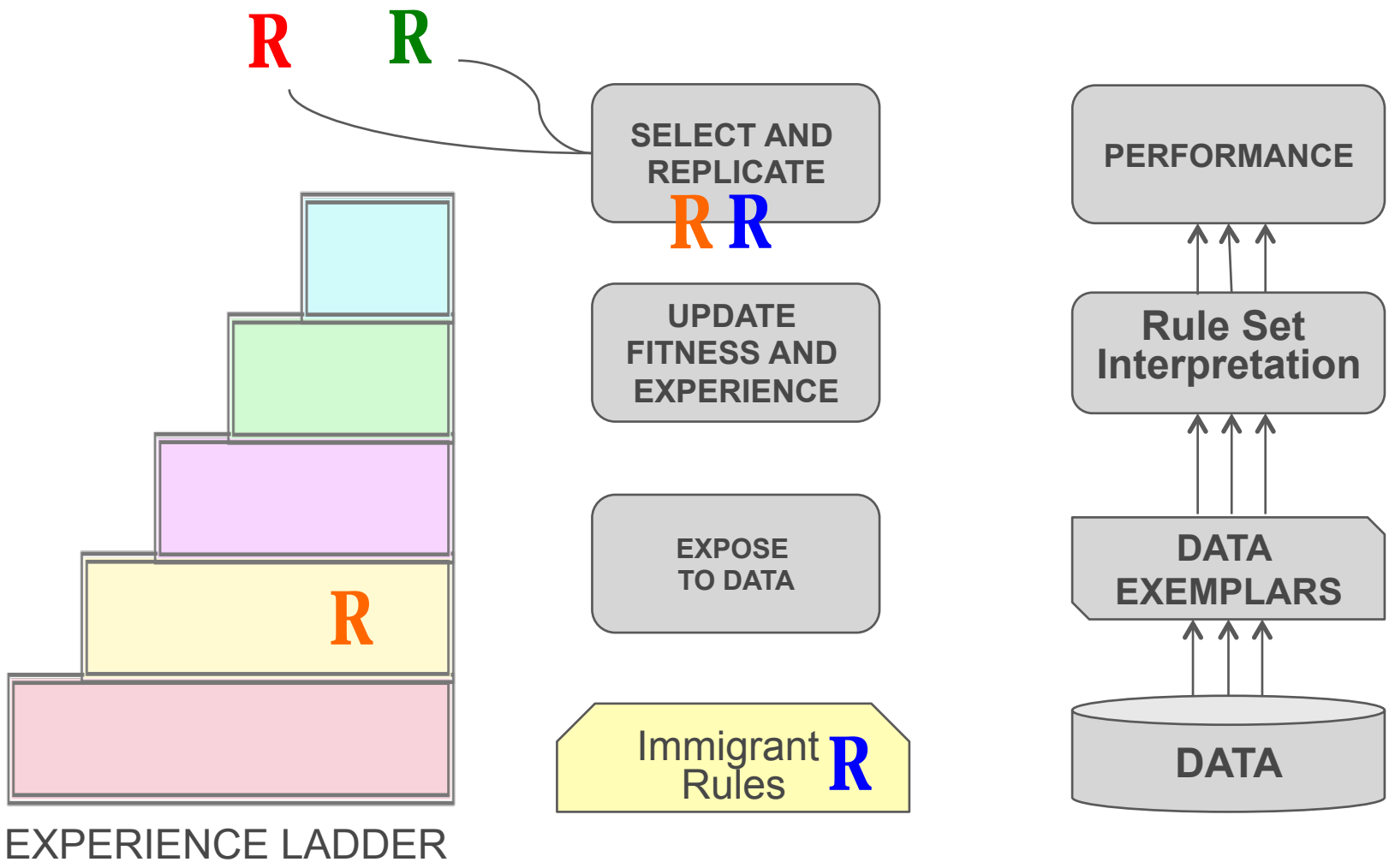
Local Learning Process

CSAIL

Interpret Migrant RuleSet



“Evolve” new RuleSet

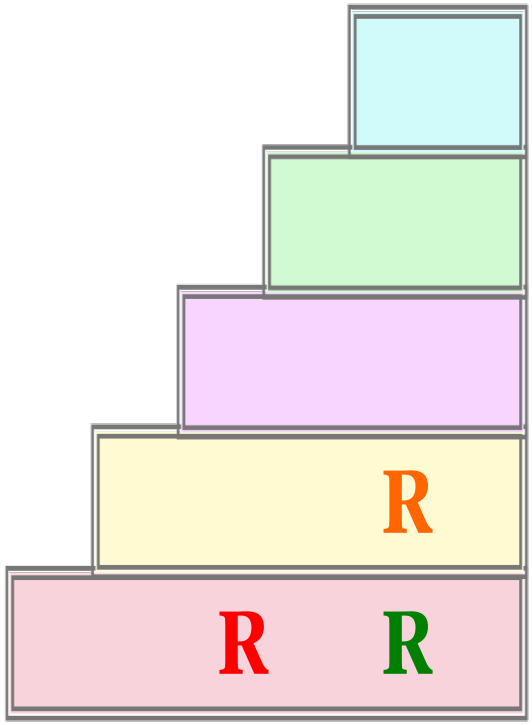


EvoDesignOpt

Local Learning Process

CSAIL

Climbing Experience Ladder



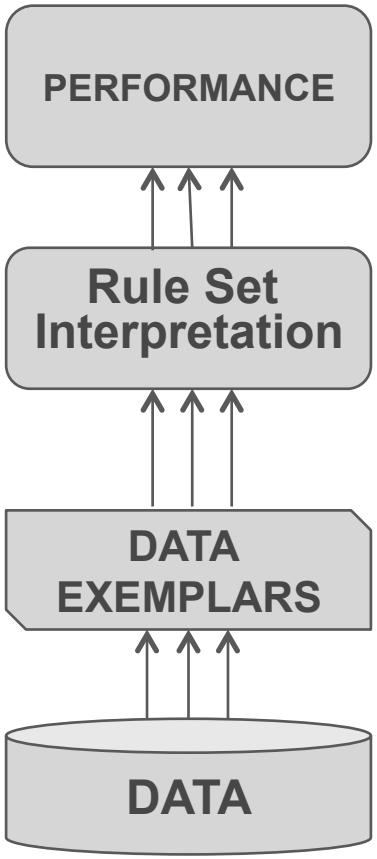
EXPERIENCE LADDER

SELECT AND REPLICATE

UPDATE FITNESS AND EXPERIENCE

EXPOSE TO DATA

Immigrant Rules R

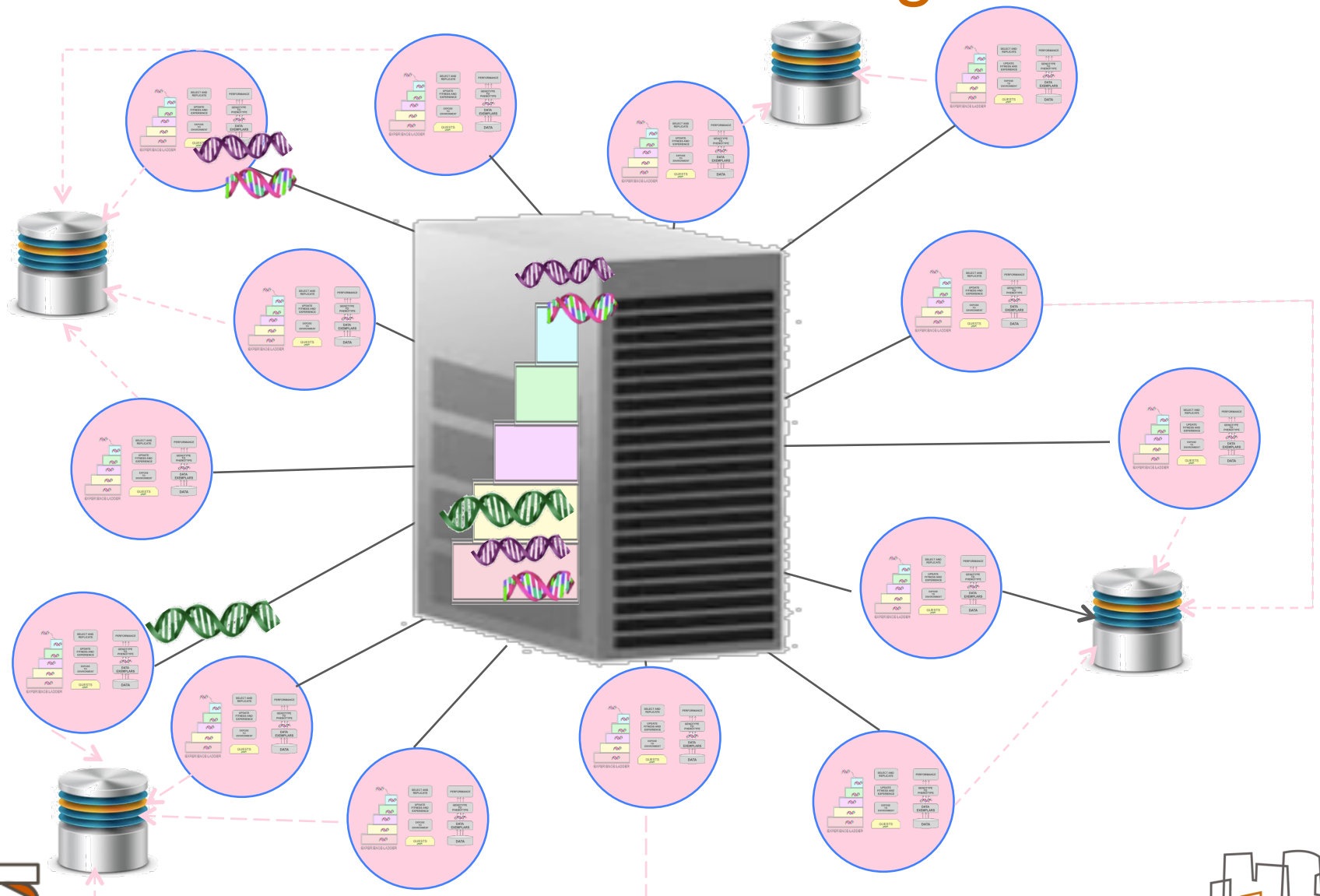


EvoDesignOpt

Local Learning Process

CSAIL

Global Learning



EvoDesignOpt

CSAIL

Platform Dynamics

- **Temporal and spatial artificial evolutionary dynamics**
  - **Population selection, replication with inheritance and variation**
  - **Asynchronous**
  - **Spatial mixing**
  - **Long duration**
  - **No single optimal solution**
  - **Large scale**
  - **Continuous adaptation**

- **Dynamics support continuous model adaptation**

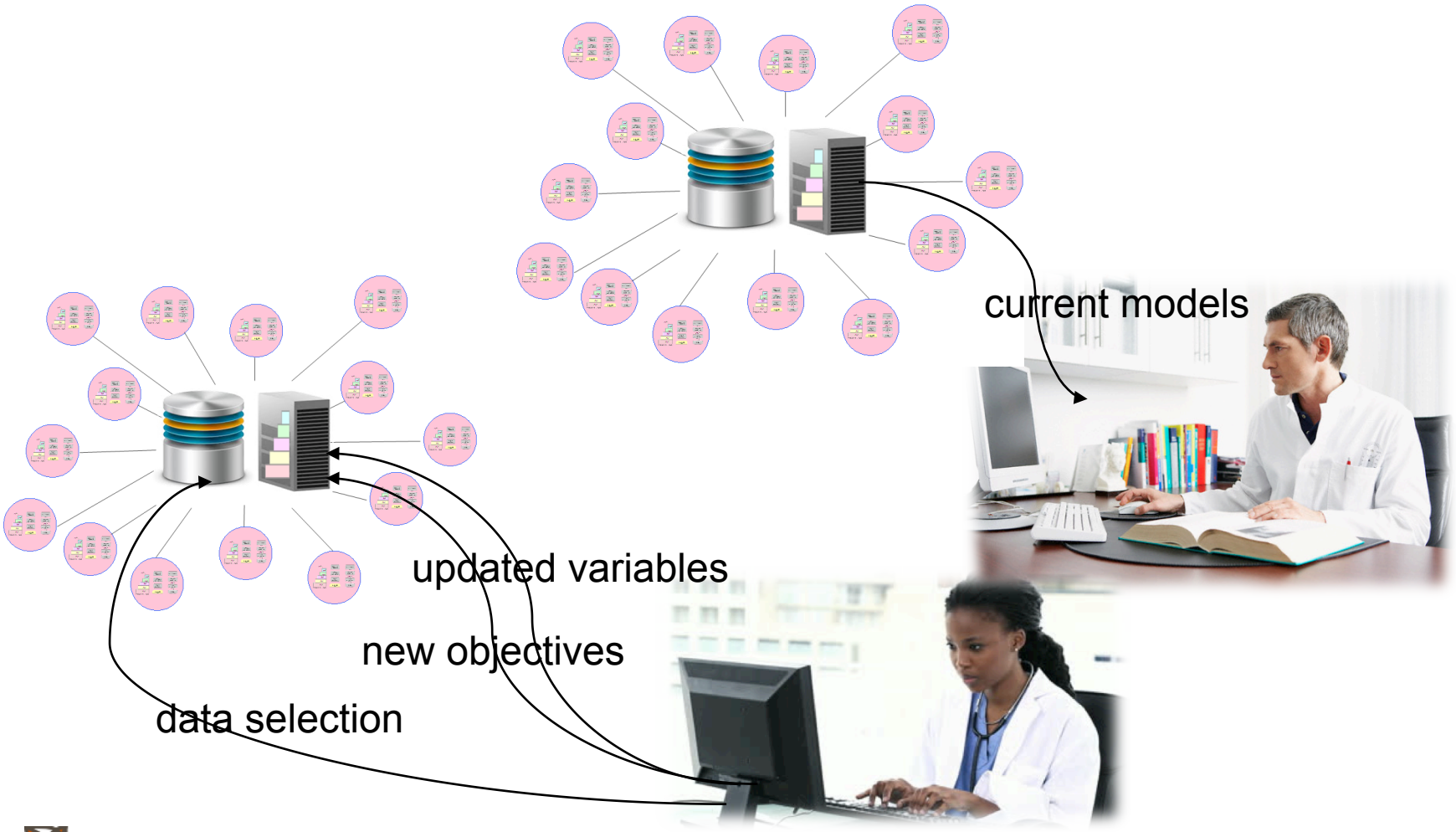
Scalability

- The data is distributed
- Each local learner is mapped to one of our computers
  - The /global learner/ hub is mapped to a dedicated computer
- Learners don't depend on each other so they can be added, removed, pause without stopping the system
- Sustained, non-blocked, fault tolerant computation

EvoDesignOpt

CSAIL

Digital Directed Evolution of Models

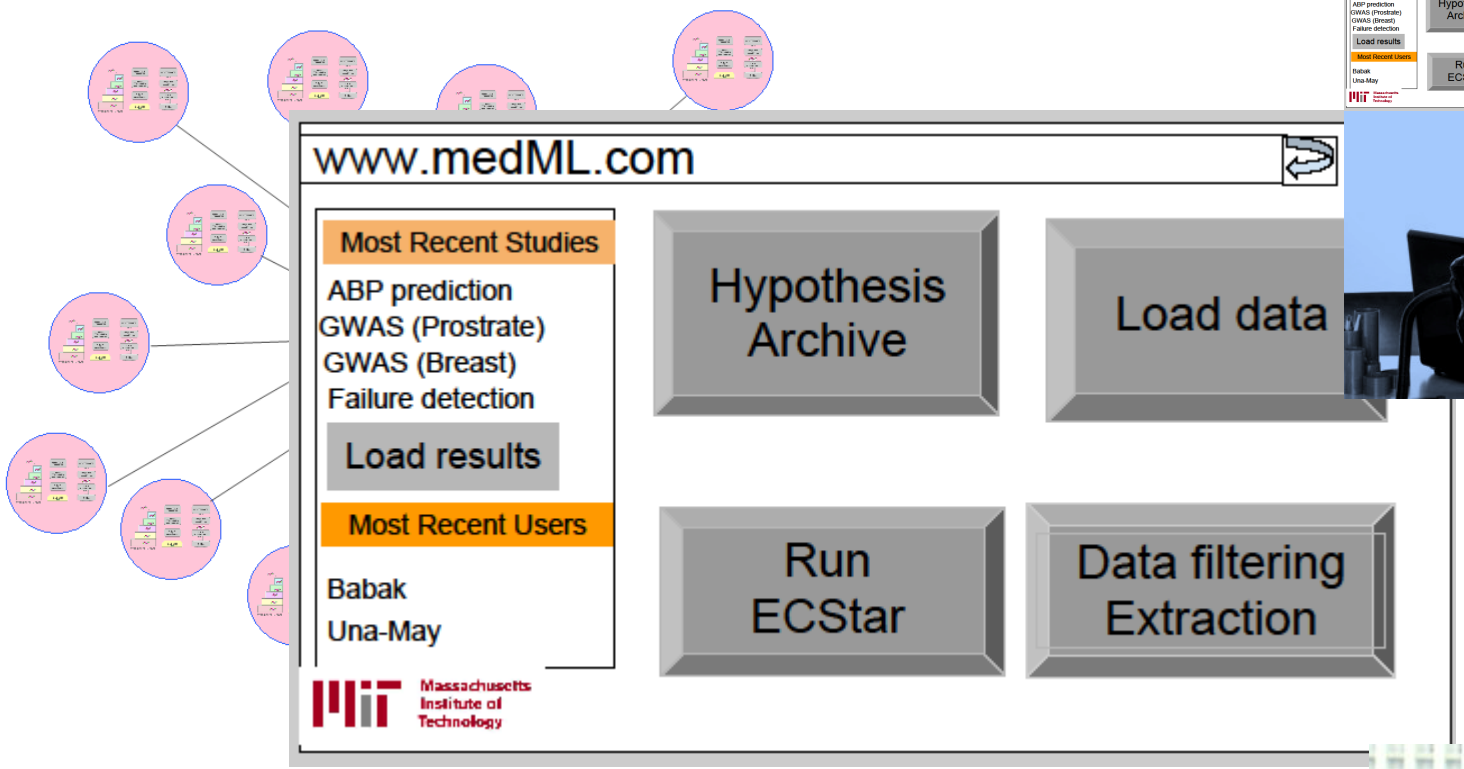


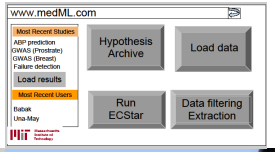
EvoDesignOpt

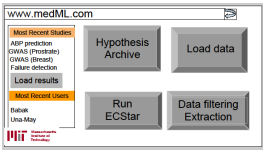
Use Case #1

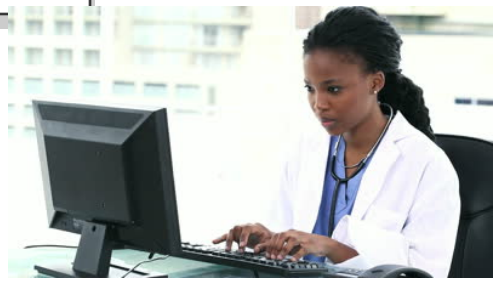
CSAIL

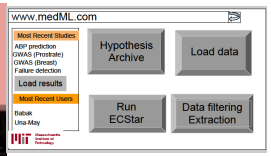
An Elastic Analysis Component of a “Many Minds” Platform











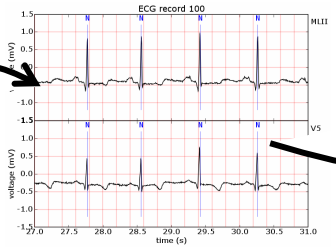
EvoDesignOpt

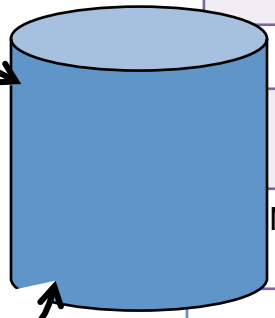
CSAIL

Arterial Blood Pressure Prediction Project

In an ICU environment, physiologic data is collected at high frequency but is ignored because of the need for immediate focus of attention.

ICU





MIMIC Waveform database

Feature	Value
Total size	4 Terra Bytes
Waveform types	22
Signal sampling frequency	125 samples/sec
Number of samples	500m

U of T, St Michael's Hospital, Canada
ICU specialists:
Dr. Jan Friedrich,
Dr. David Klein
Dr. M Mamdami

- Pressure (mmHg)
  - PAP(pulmonary), CVP(central venous), ART(arterial), RAP(right arterial), LAP(left arterial), ABP(arterial)
  - AOBP, UAP, ICP, P1
- ECG (mV)
  - Leads- ECG V+, I , ECG III, V, ECG V2, II , ECG I, ECG II, ECG V, ECG II+, III
  - Other – MCL1, ECG AVF, ECG MCL, ECG AVL, ECG AVR, AVR, AVF , AVL
- Other Signals
  - RESP(pm)- respiration
  - PLETH - plethysmogram, …

EvoDesignOpt

CSAIL

"Universal" Blood Pressure Prediction System

Parameterizations :

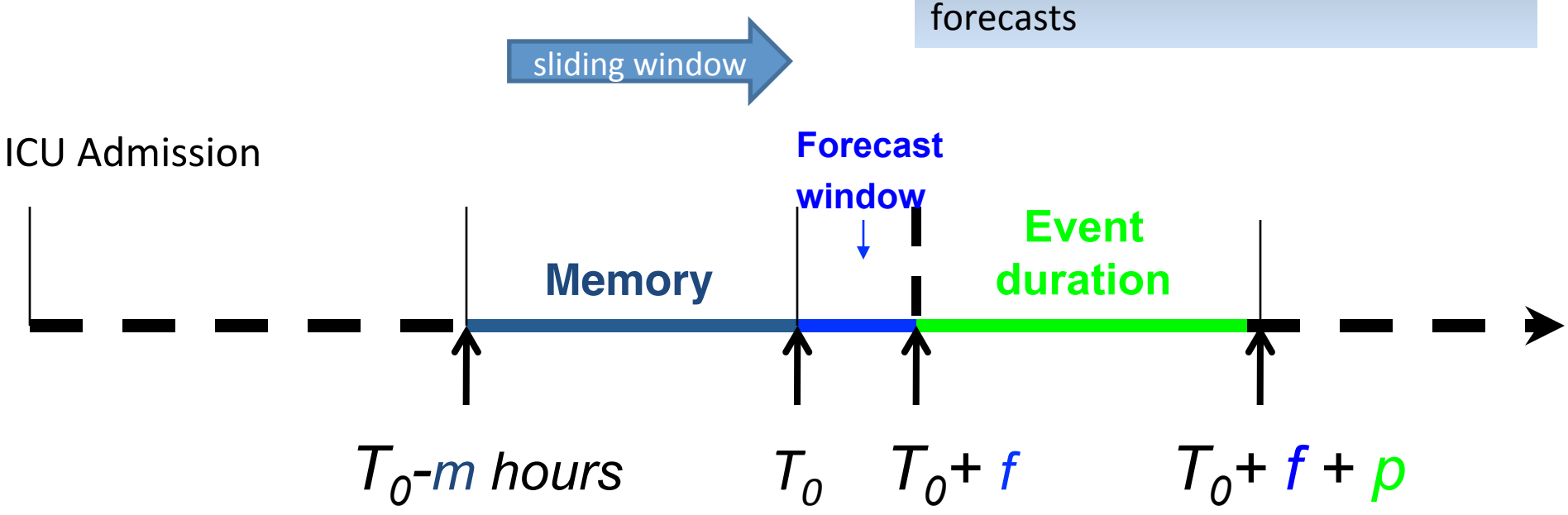
***m*- hours of past data used to forecast**

***f* – forecast window, lead**

***p*- duration of event**

In Training: System sequentially processes data over the *m* hours making predictions, with forecast window data is blacked out.

In Deployment: System will move over data and make real-time forecasts



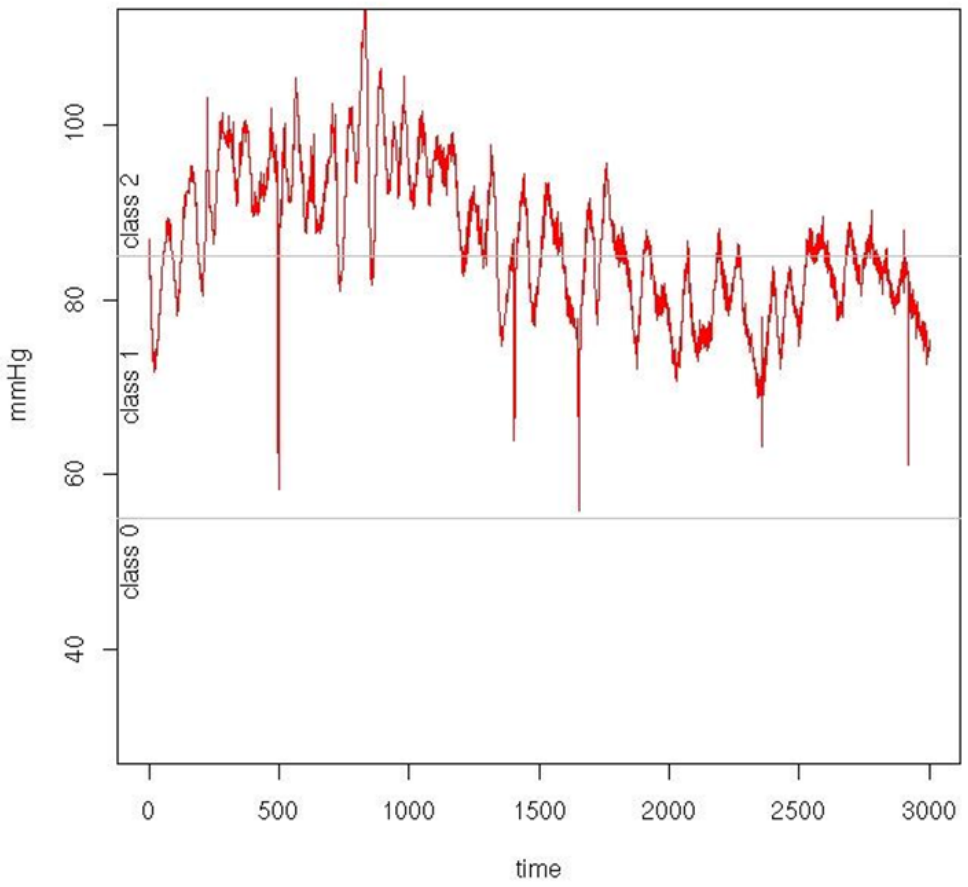
EvoDesignOpt

CSAIL

Forecast Classes: 3 label assignment

Data is categorized into a minimal three classes as prediction targets (also used as indicators).

The goal is to predict the class, based on current and past information, over a forecast period starting after a blackout forecast window.



Class 0: Low

ABP ≤ 55mmHg

Class 1: Normal

55mmHg < ABP ≤ 85mmHg

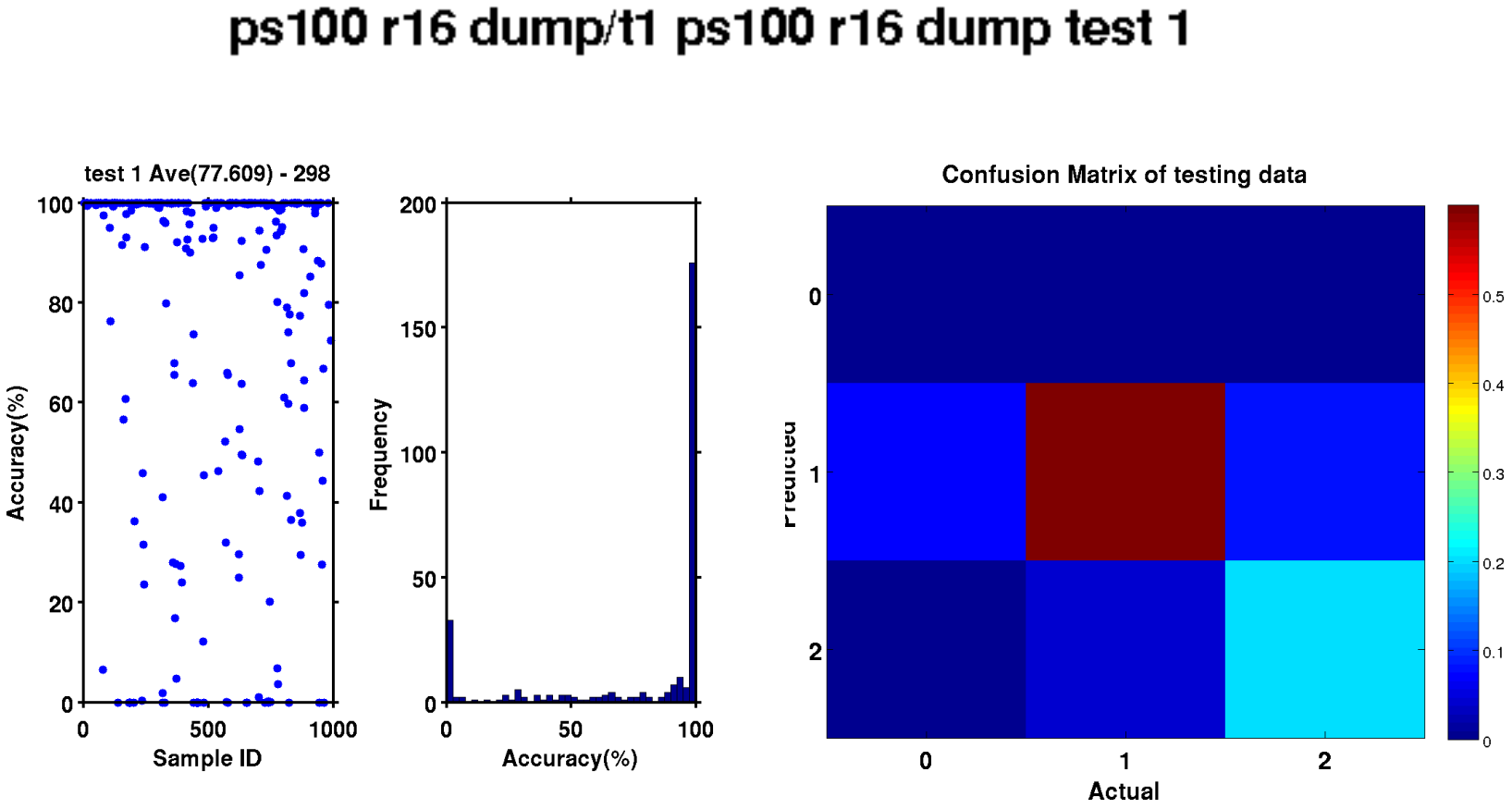
Class 2: High

85 < ABP

Class 0: Low

ABP < 55mmHg

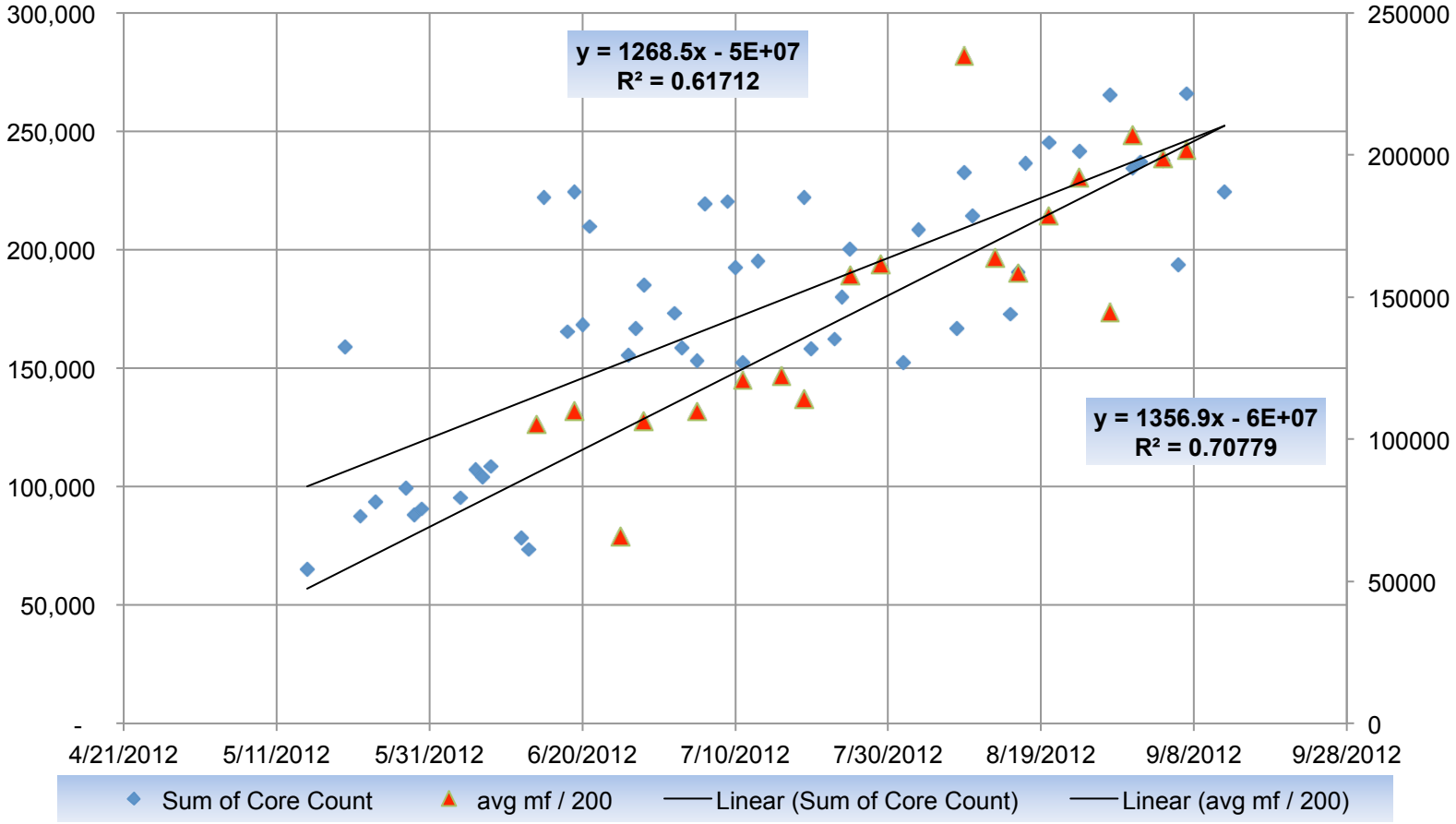
Result Interpretation



EvoDesignOpt

CSAIL

Scalability: improvement with more time and resources



EvoDesignOpt

CSAIL

EC-Star

- **A sustained, fault tolerant, scalable learning system**
  - **Integrating local and global computational evolutionary intelligence**
    - » **Spatial and temporal dynamics**
    - » **Large scale**
    - » **Stochastic**
- **Deployable with any size computational resources and data sets**
- **Emphasizes interpretable representation**
- **Please talk to me if you want to try a pilot project**
- **Generously supported by Li Ka Shing Foundation**