Scalable Machine Learning to Exploit Big Data for Knowledge Discovery

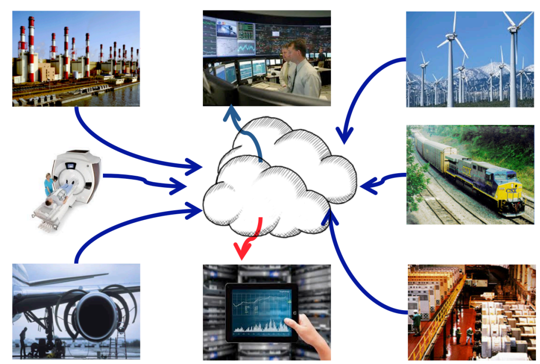
Una-May O'Reilly
MIT

MIT ILP-EPOCH Taiwan Symposium
Big Data: Technologies and Applications

ALFA
ANYSCALE LEARNING FOR ALL

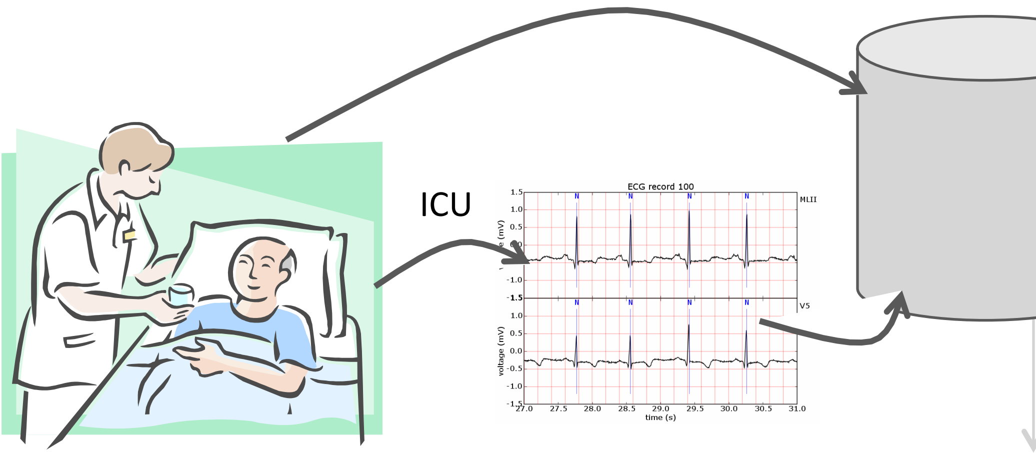
CSAIL

Lots of Data Everywhere



The GigaBeats Project

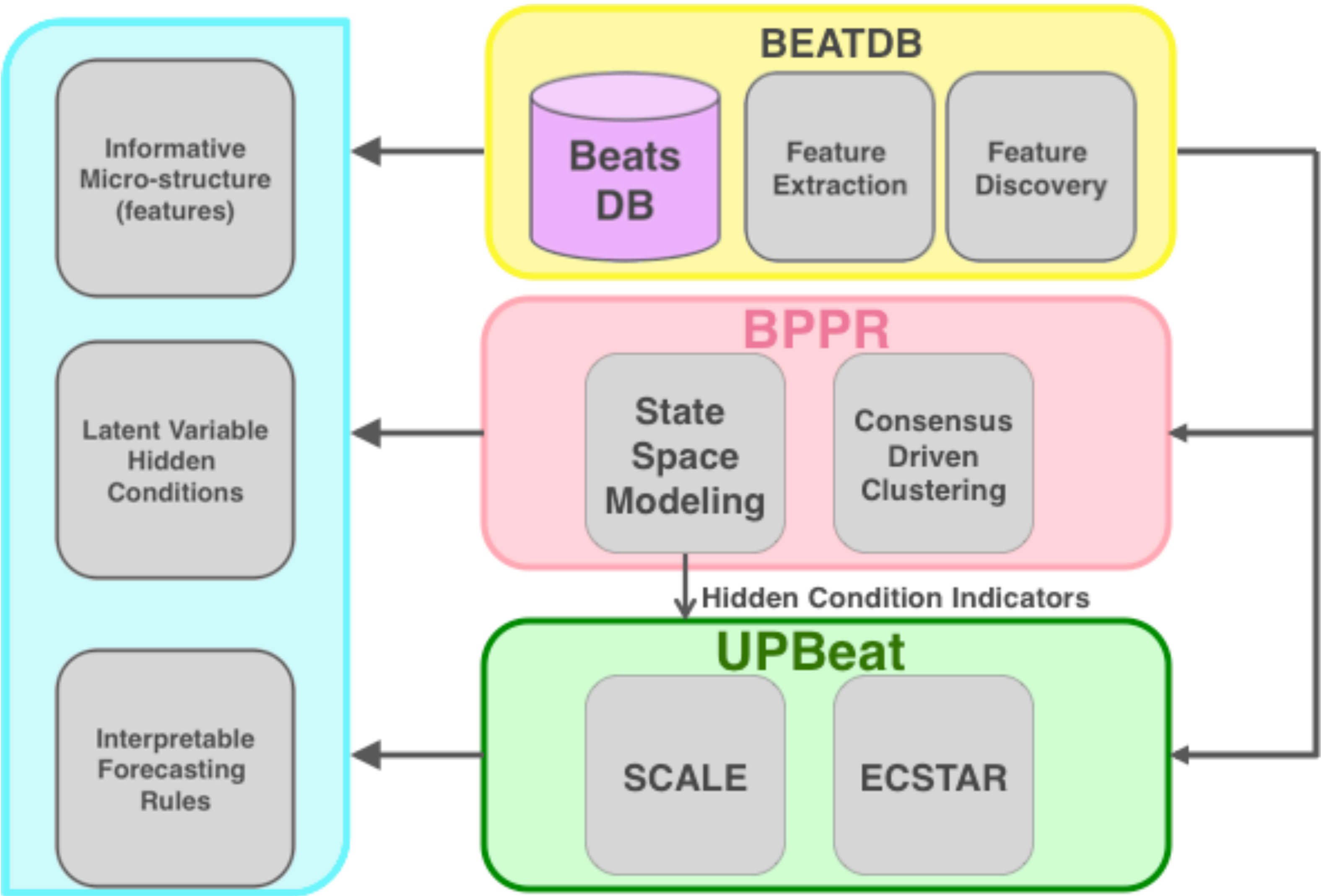
In an ICU environment, physiologic data is collected at high frequency but is ignored because of the need for immediate focus of attention.



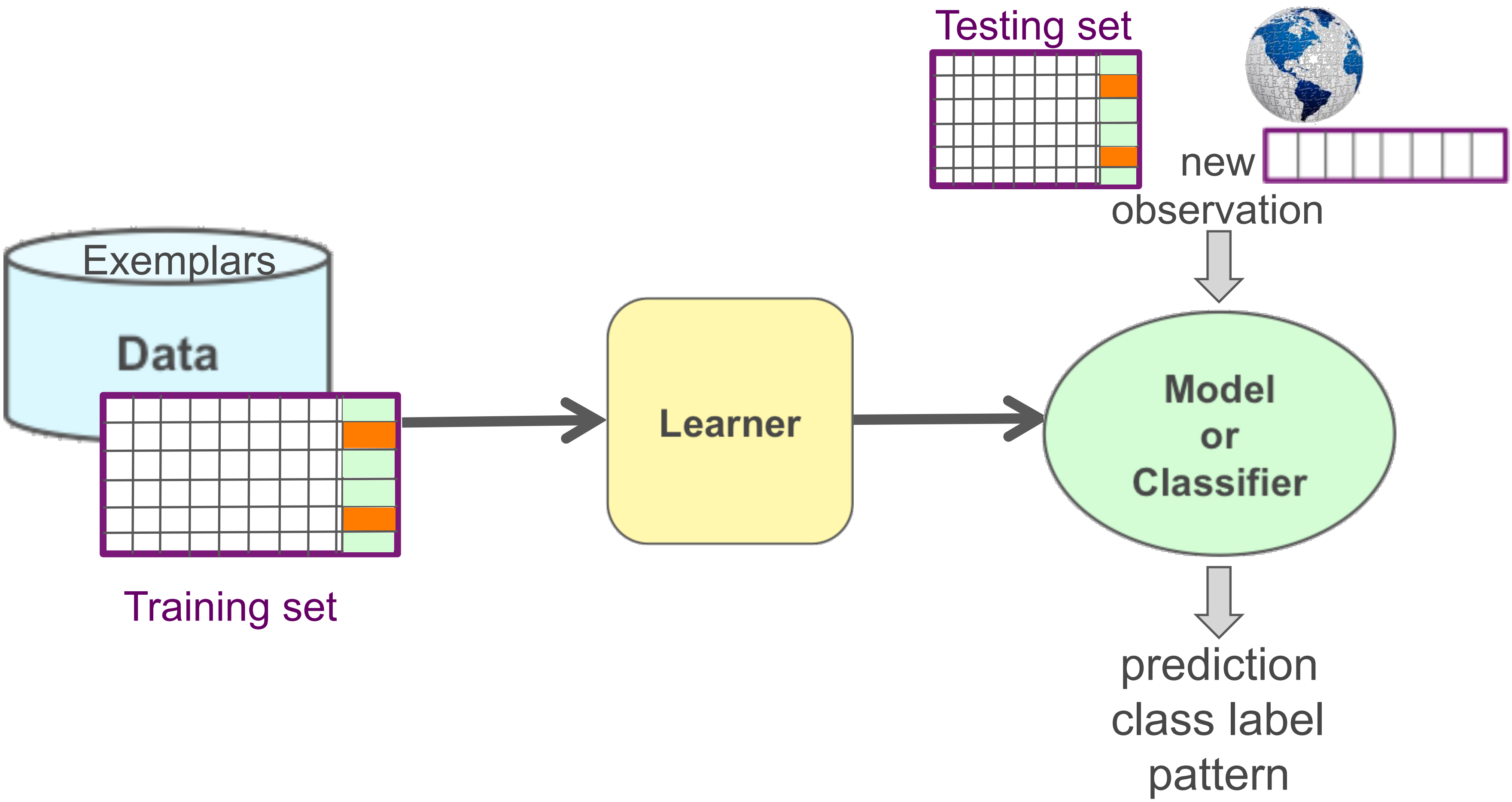
MIMIC Waveform database

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

GigaBeats Project



Machine Learning Primer



ANYSCALE LEARNING FOR ALL

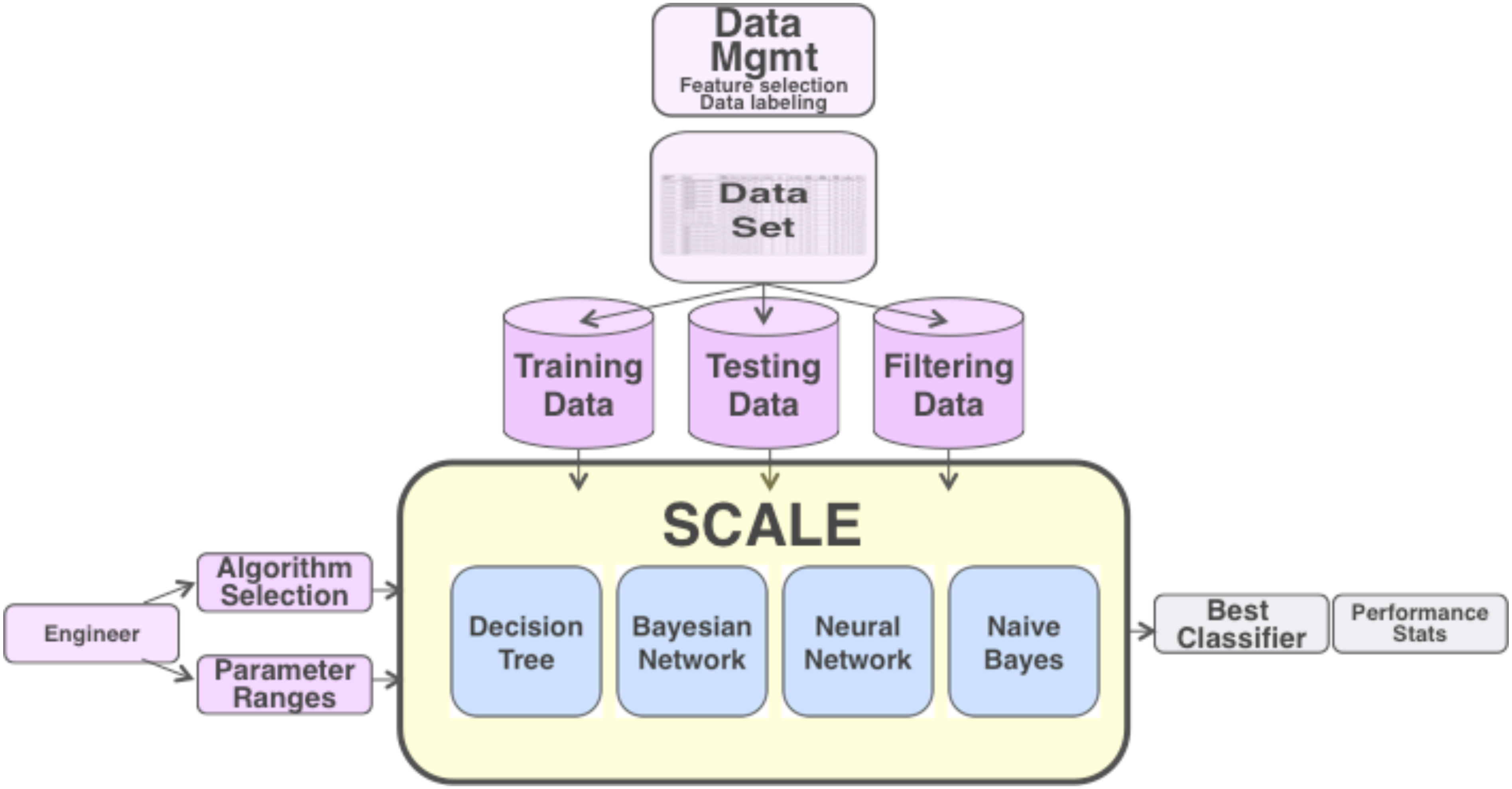
CSAIL

Agenda

Distributed Computation
+
Scalable Machine Learning

SCALE
FlexGP
EC-Star

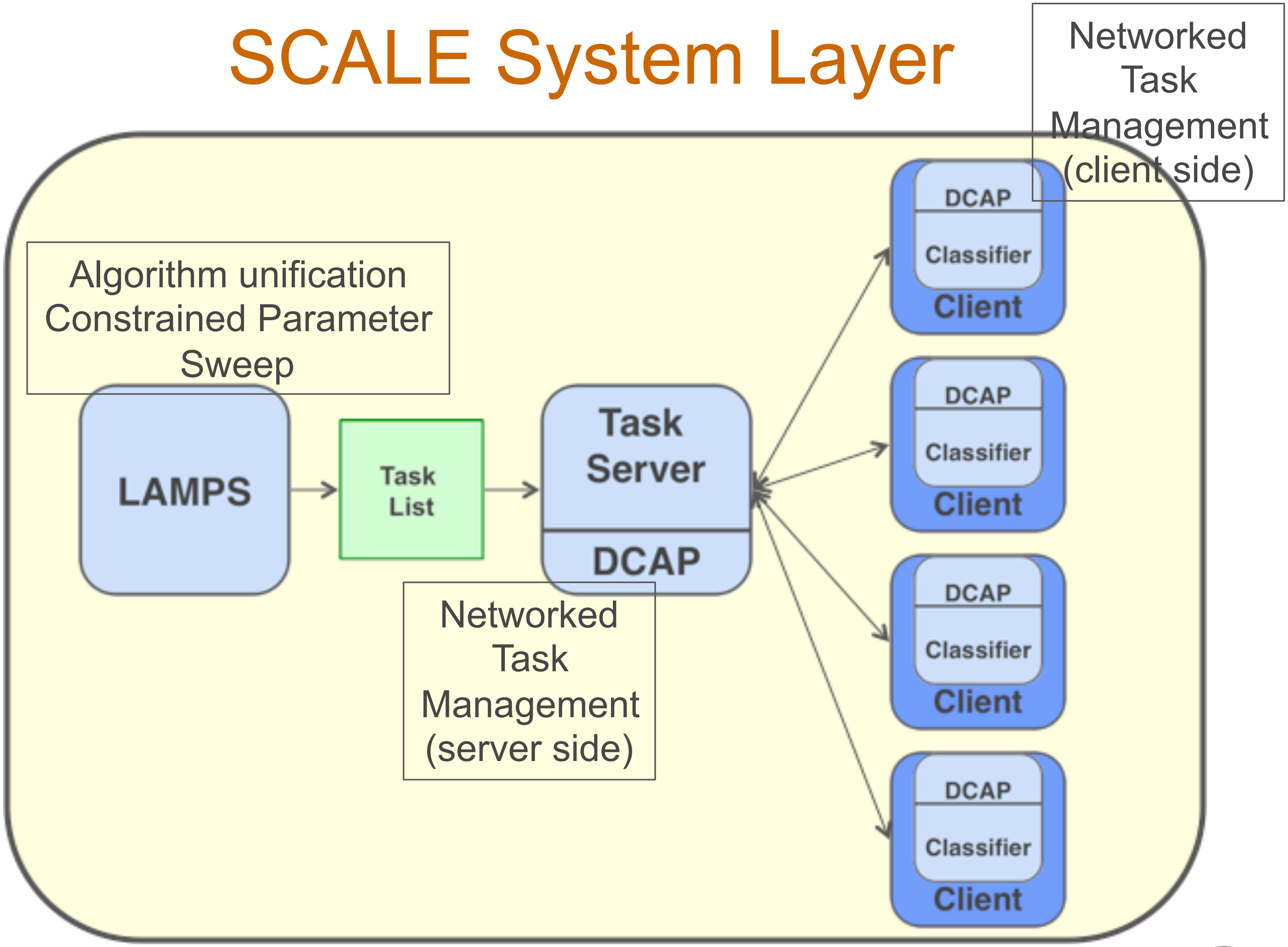
SCALE



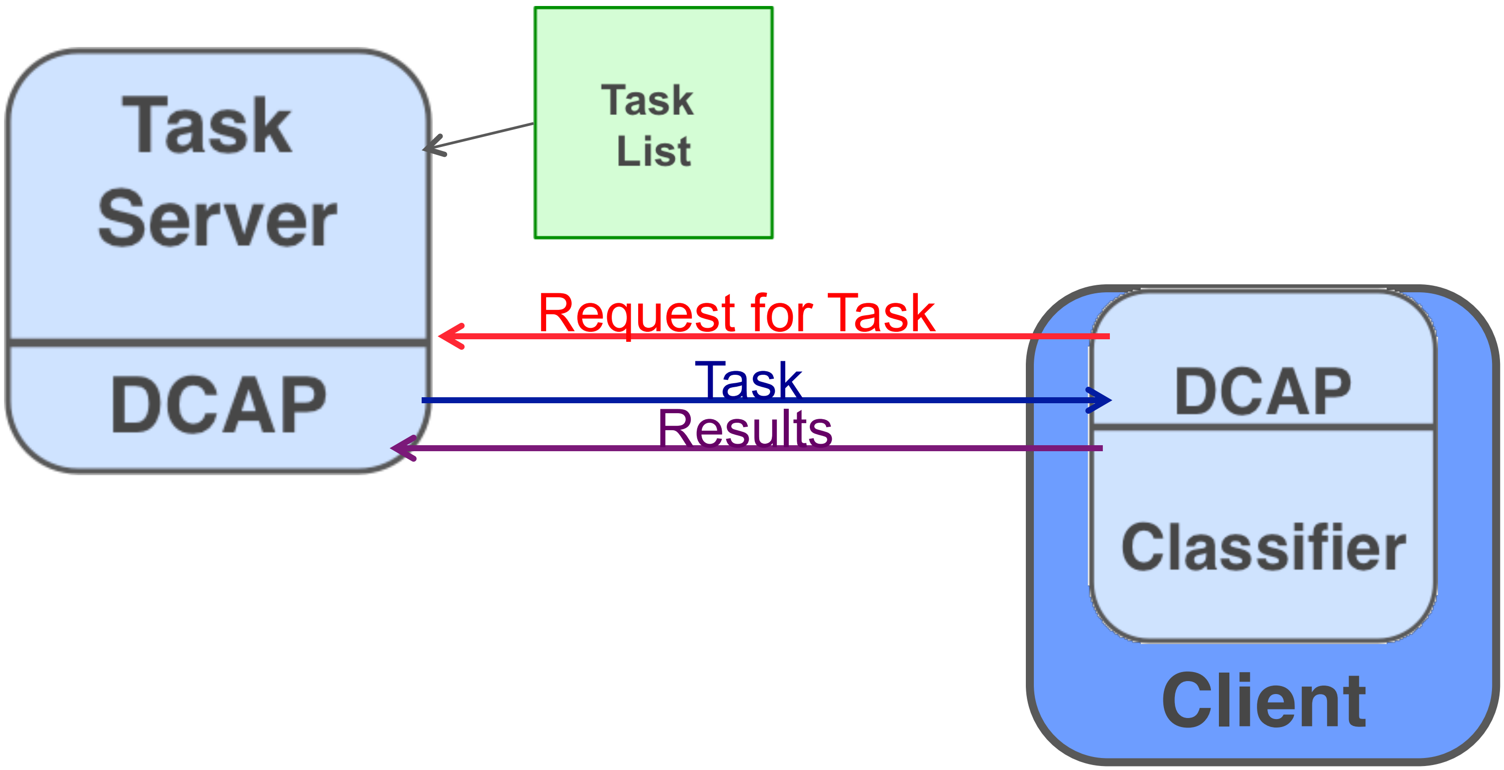
ALFA
ANYSCALE LEARNING FOR ALL

MIT CSAIL

SCALE System Layer



DCAP Protocol



SCALE Demonstration: WD-1

Forecasting ABP

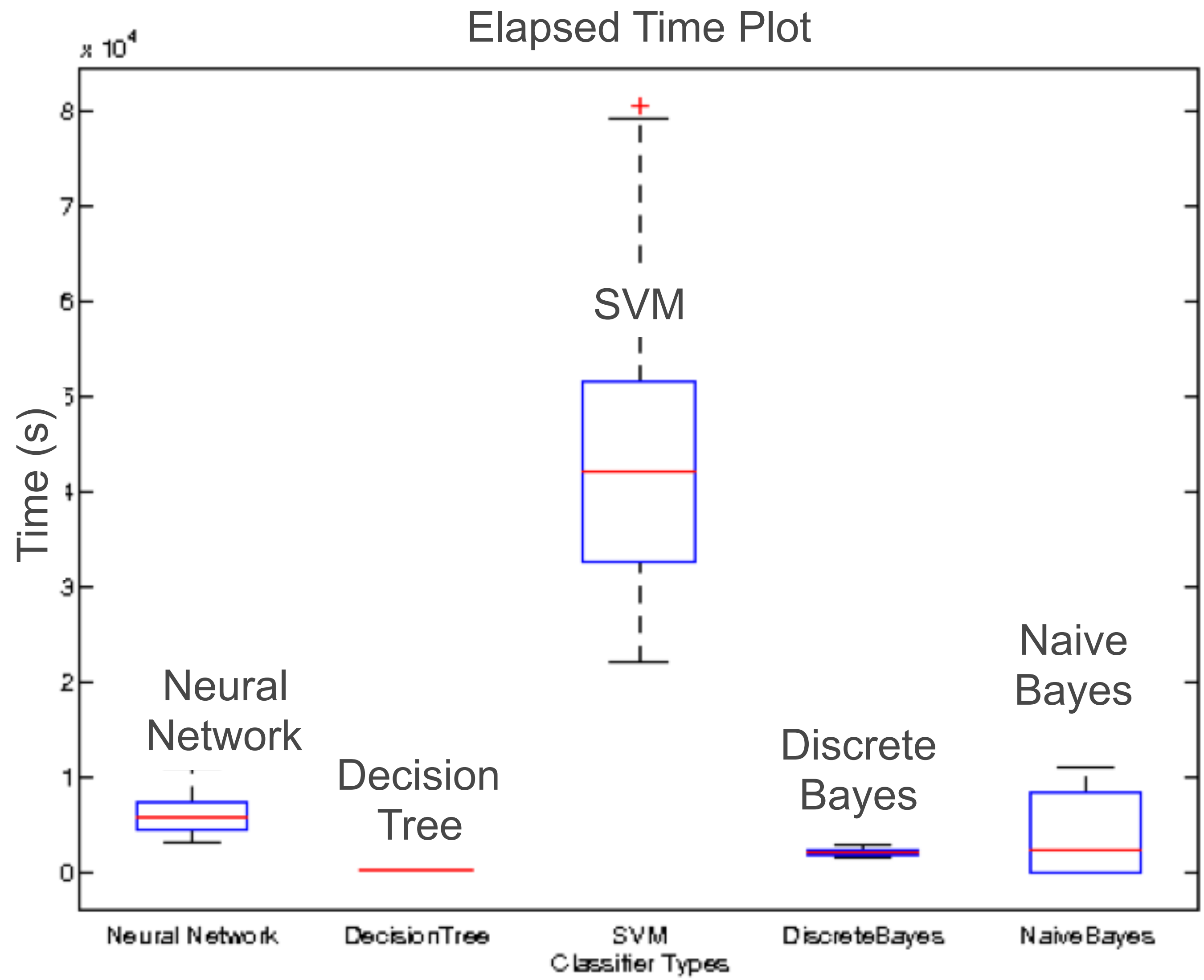
- **10 levels**
- **95K exemplars**
- **67K/28K training/test split**
- **7 dimensions**
  - **Stats, trends on MAP**
- **1225 learner tasks (10 fold cross validation)**
- **80 nodes, ~2 days, ~4000 node hours**

Classifier	Number of Instances	Errors
Neural Networks	720	0
Discrete Bayes	18	0
Naive Bayes	64	0
Support Vector Machines (SVM)	66	34
Decision Trees	324	0

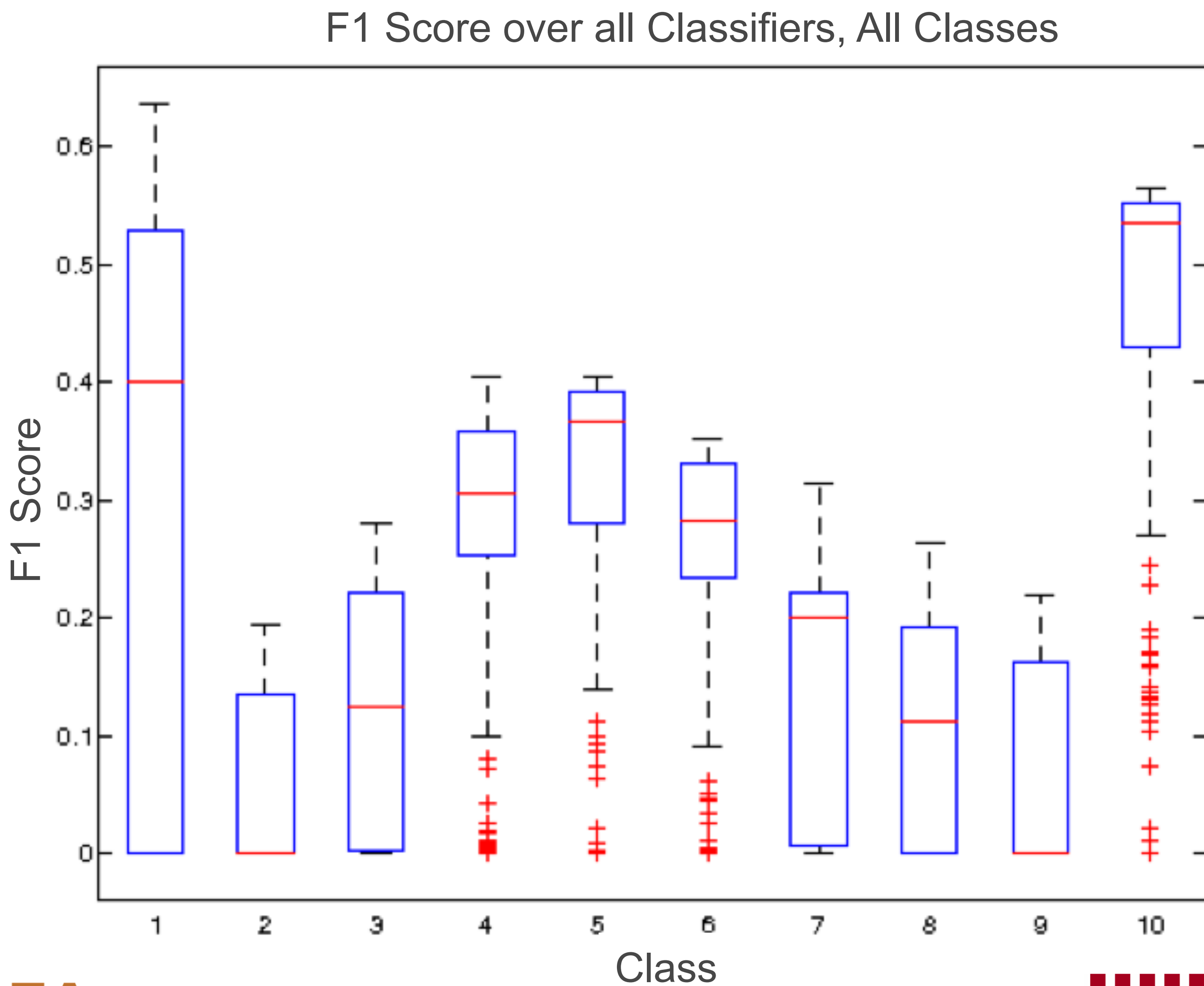
ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

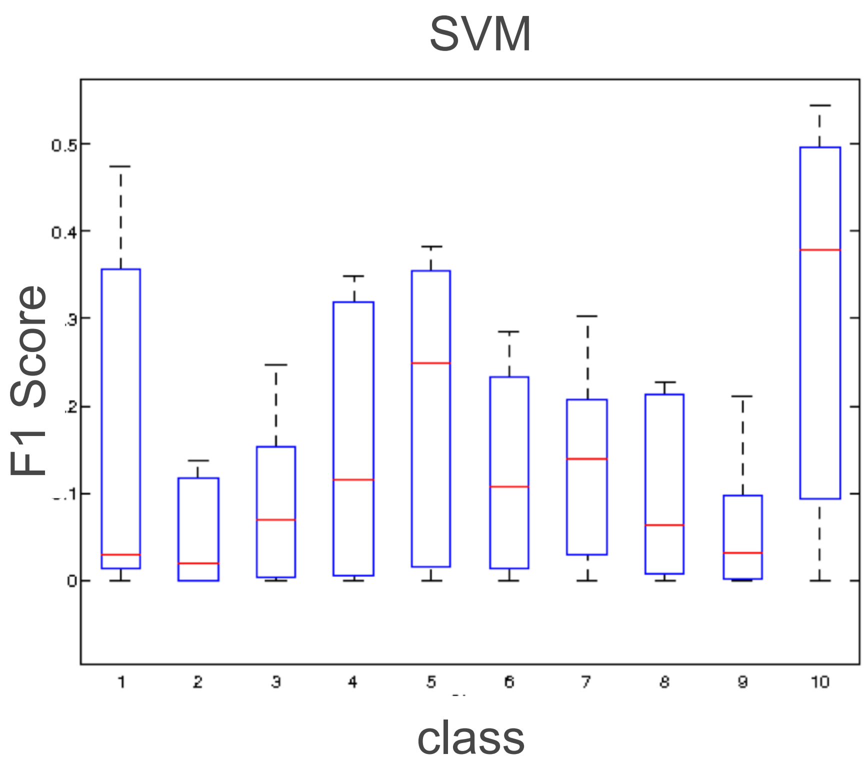
SCALE: WD-1 Running time

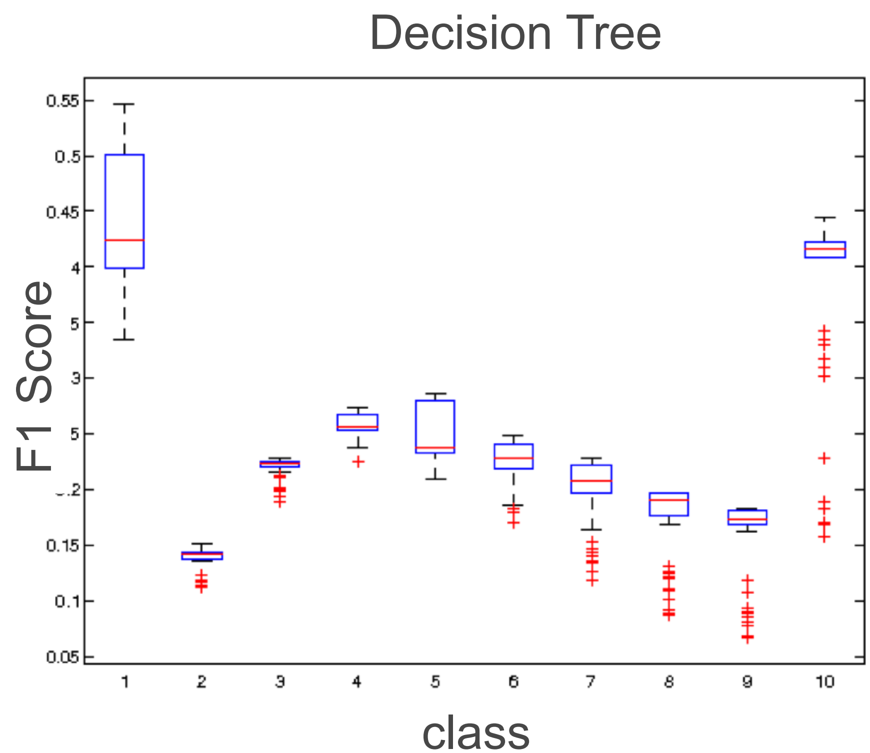


SCALE: WD-1 Accuracy results



SCALE: WD-1 Comparison SVM v DT





Scaling up: from SCALE to FlexGP

SCALE

- Modest 10's of features
- Assumes all training data fits into RAM

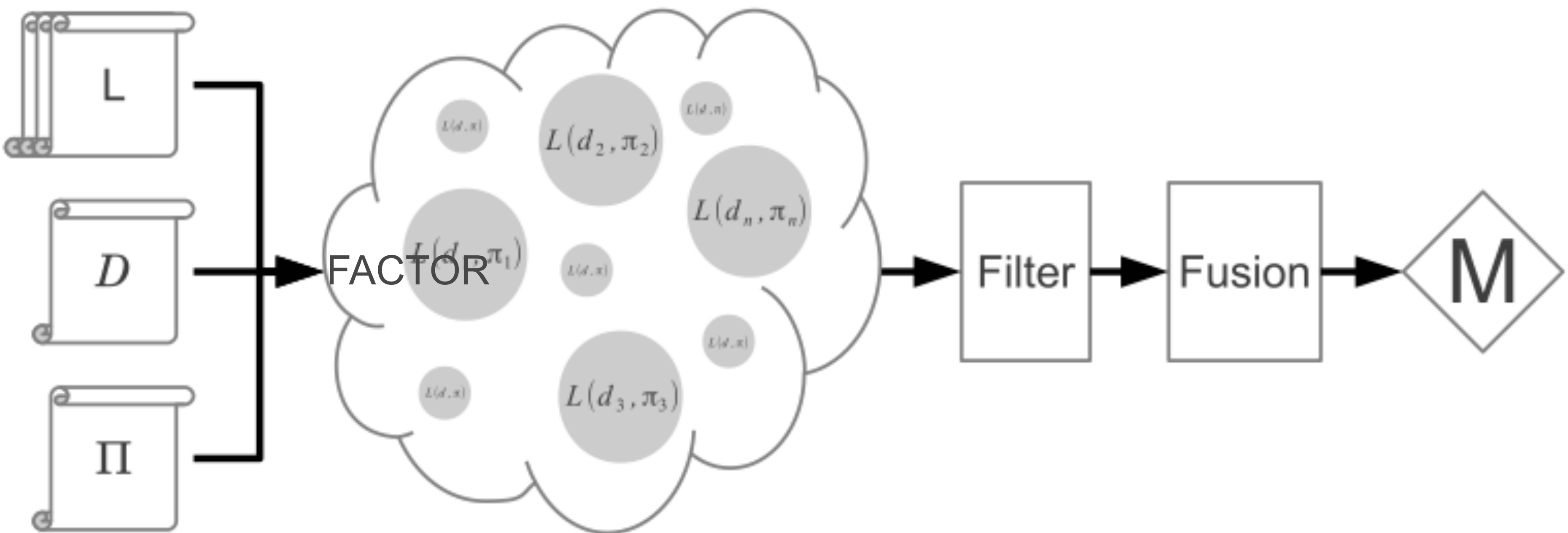
FlexGP

- 100's of features
- Big Data
- Big Data requires multidimensional factoring, filtering then fusion

ANYSCALE LEARNING FOR ALL

CSAIL

FlexGP ML Layer

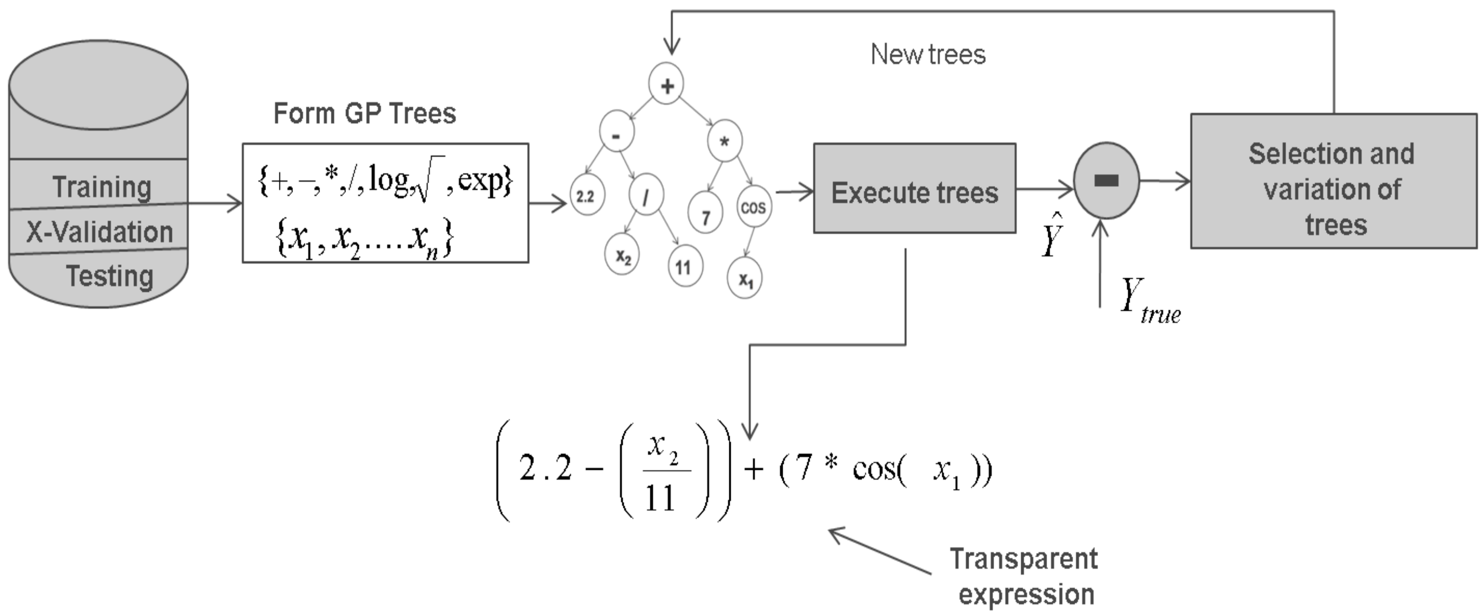


FACTORING

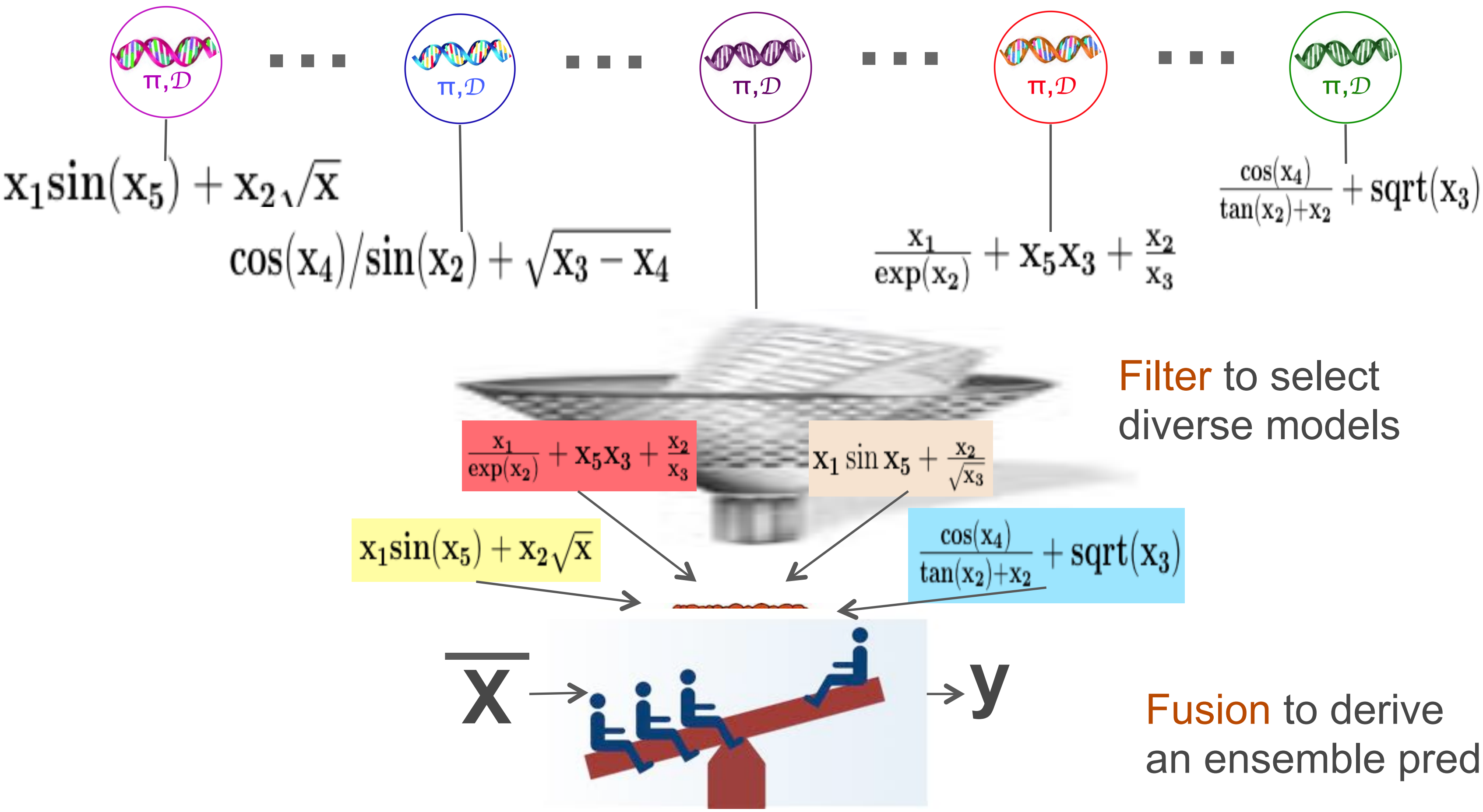
Π: Probability of feature, objective function, operator
$\mathcal{D}$: factoring of the data

FlexGP Learner

Goal: Model $y = f(x_1, x_2, \ldots x_n)$



FlexGP Filter and Fusion

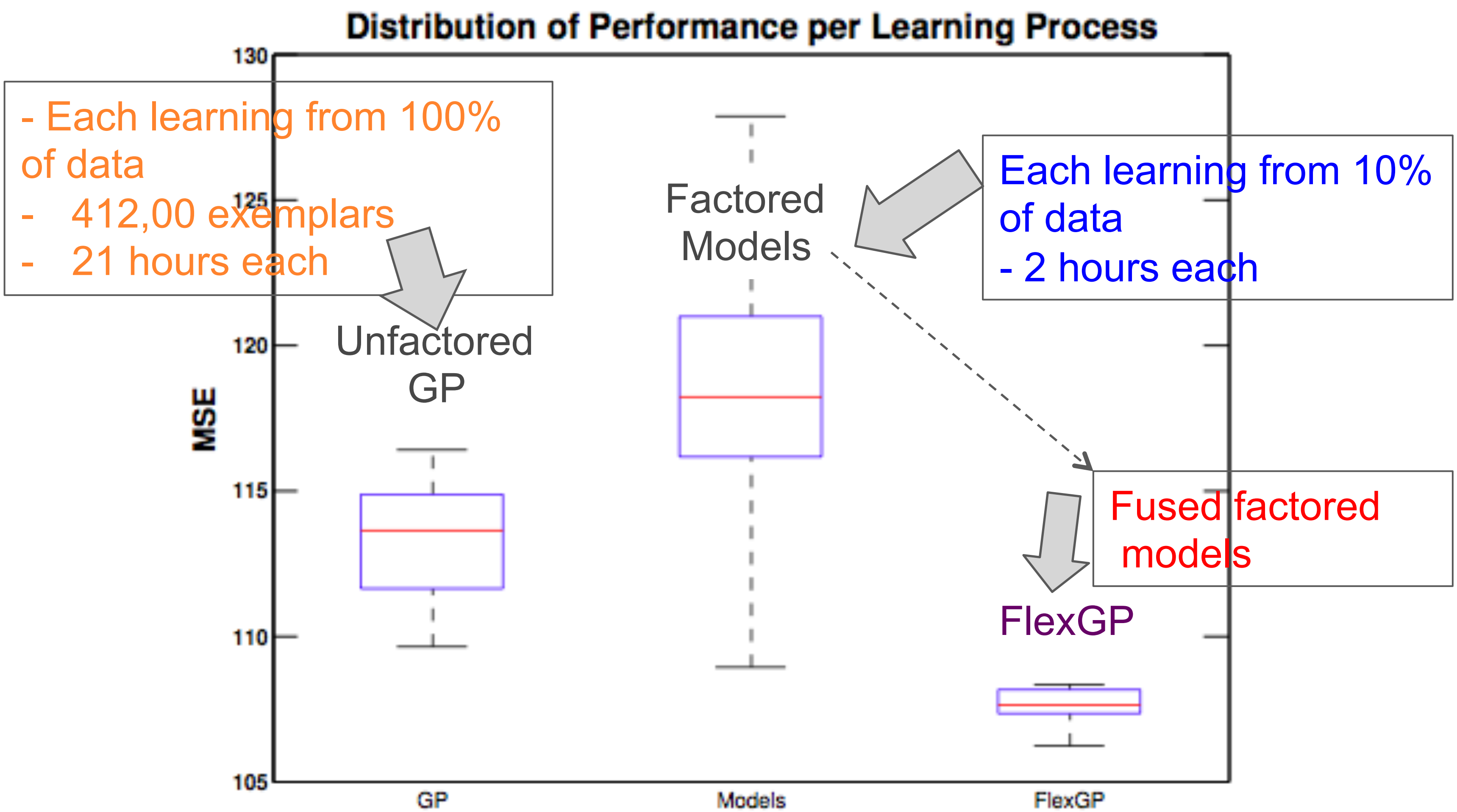


Filter to select diverse models

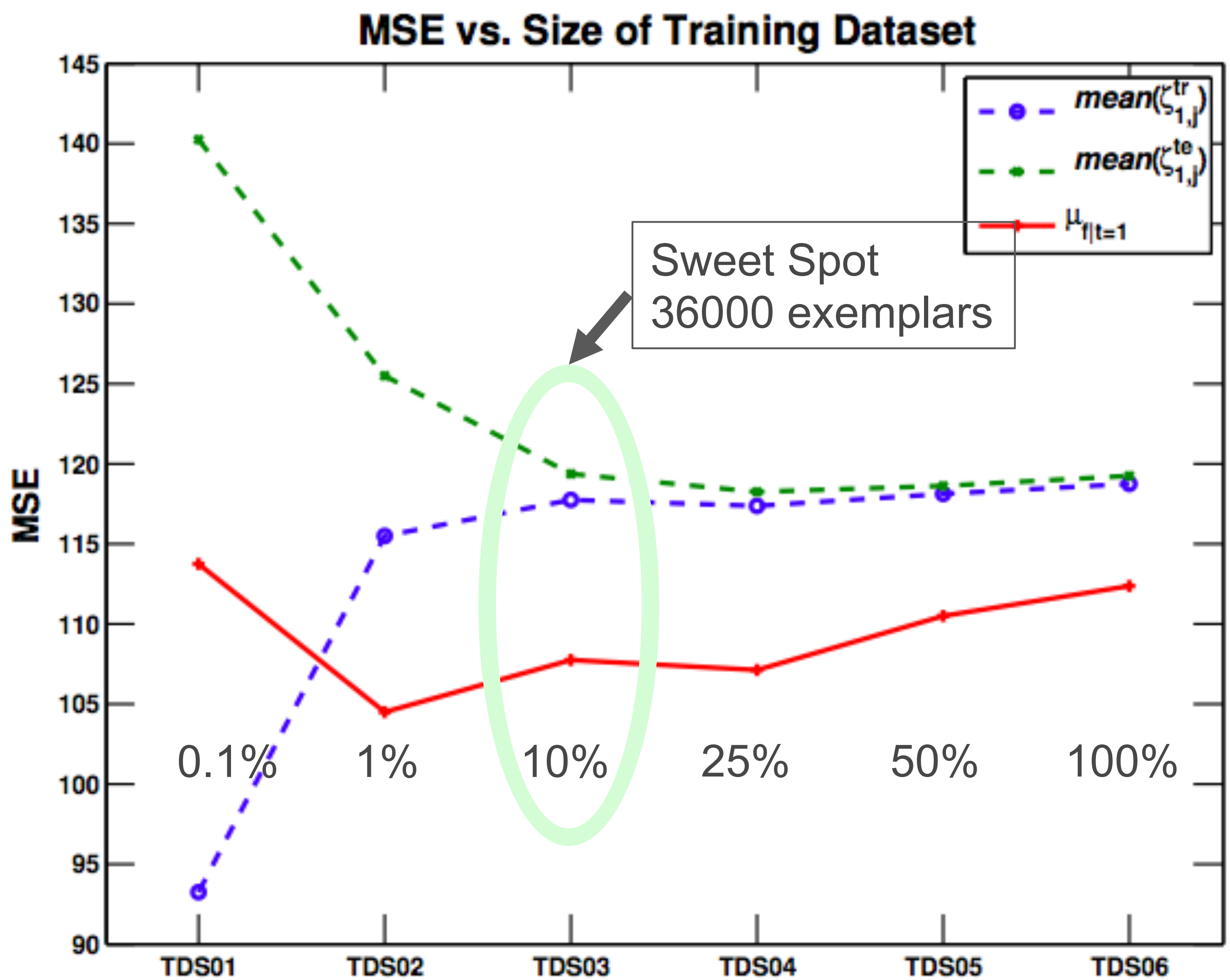
Fusion to derive an ensemble prediction

Adaptive Regression Mixing

Factoring is Better



FlexGP: Data Factoring Size Study



Dataset	$\lvert D_{gp} \rvert$	$\lvert D_f \rvert$	$\lvert \tilde{D}_t \rvert$	Total	$\lvert \bar{\mathbf{x}}_j \rvert$	Range of z
MSD	398,924	67204	49436	515564	90	[1922 2011]

MIT CSAIL

Resource and System Management Layer

Completely decentralized launch

- Start up
  - Cascading launch is decentralized and allows elastic retraction and expansion

Completely decentralized network support

- IP discovery and launch of gossip protocol added to start up cascade
  - Launch Ip discovery and ongoing gossip protocol enables communication network among nodes at algorithm level
  - Support for Monitoring/reporting/harvesting current best

Remarks

- Essential design for resilience to node failure
  - Launch or running node loss will not halt computation,
  - lost launch branch or node can be integrated seamlessly
  - Node loss in a communication network won't break the network

Statistical oversight of learning algorithm's execution parms and data

- distribution for parms and data added to start up cascade, negotiated between parent and child at launch of child

System Layer: SCALE vs FlexGP

SCALE:

- LAMPs pre-defines the tasks
- Every learner has to know IP of task handler
- Task handler is a bottleneck and central point of failure

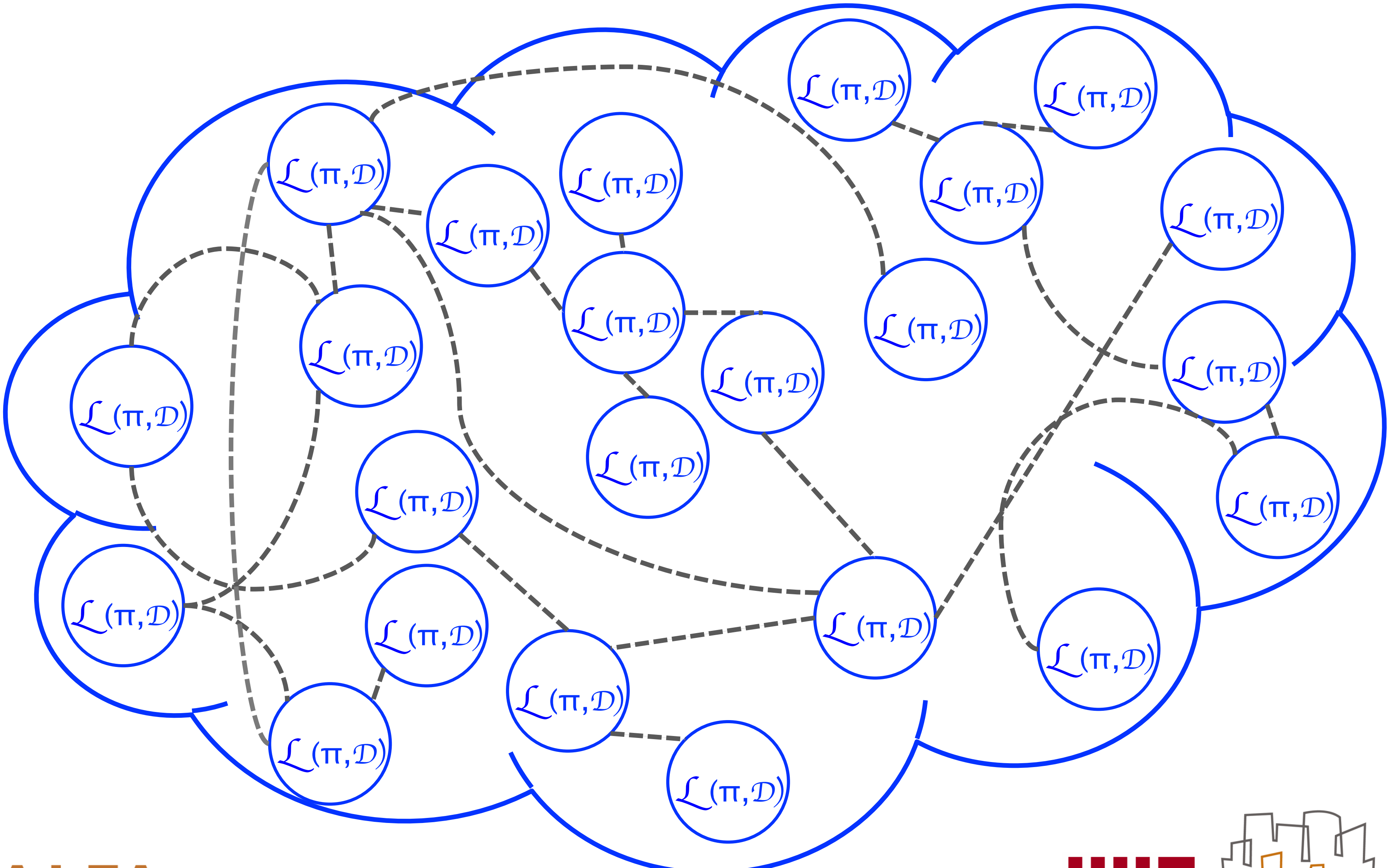
FlexGP

- Autonomous task specification
- Learners gossip to learn each others' IP
- No central task handler or point of failure

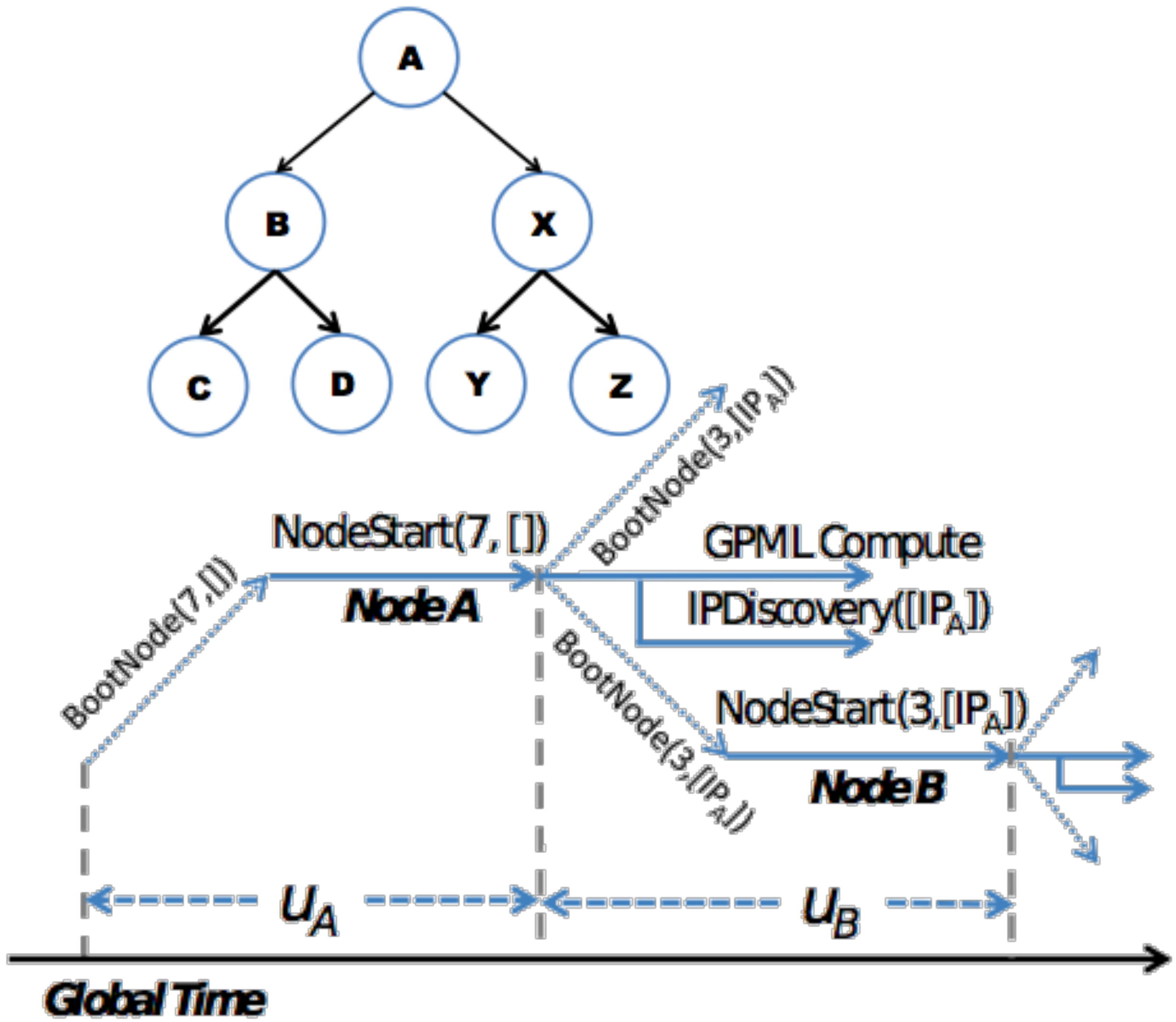
ANYSCALE LEARNING FOR ALL

CSAIL

FlexGP System Layer



FlexGP Launch

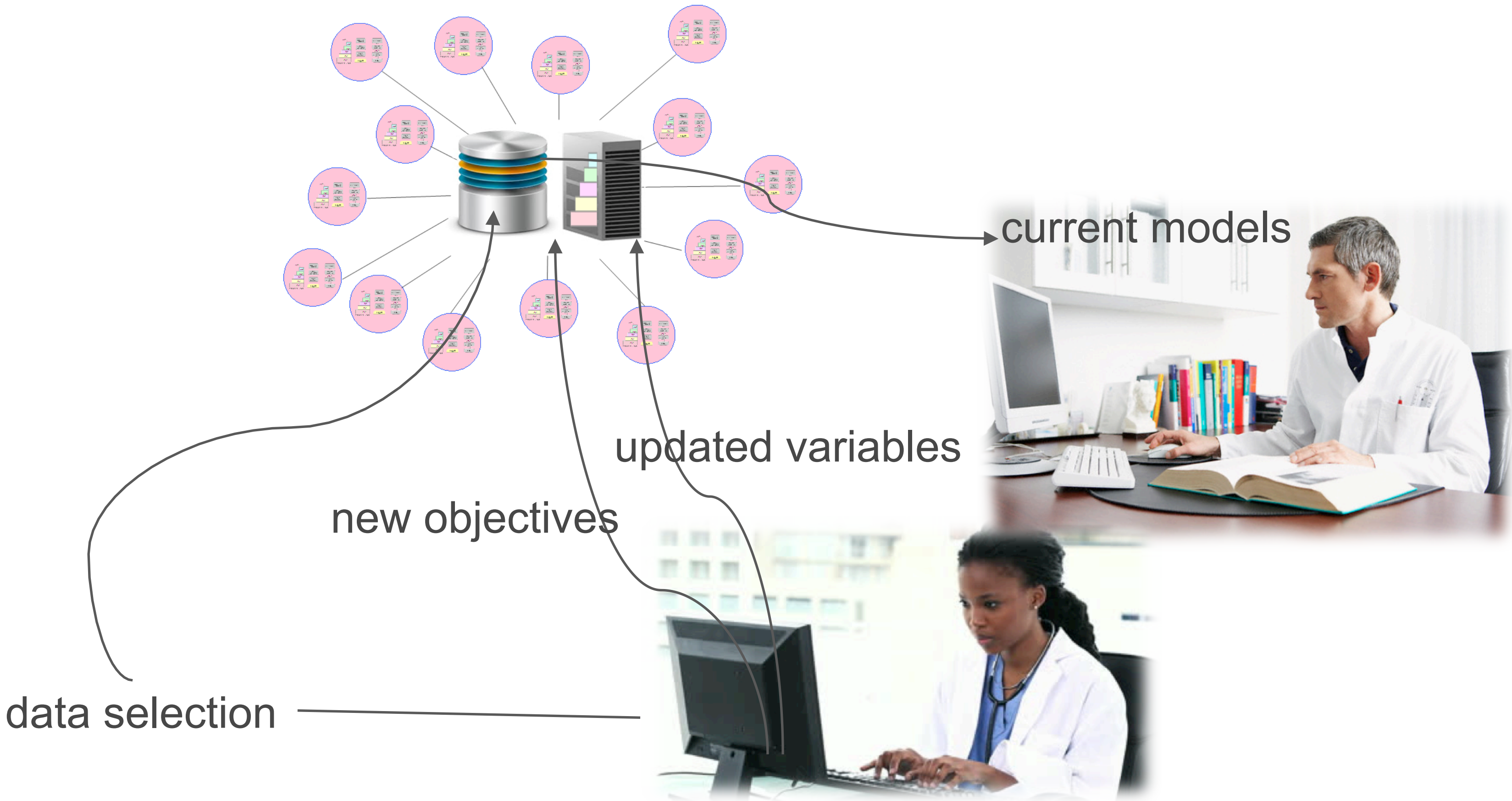


ANYSCALE LEARNING FOR ALL

ECStar

- Goal: compute very cost effectively on *VAST* number of nodes…with a lot of training data
  - Runs on thousand to 10’Ks 100K’s million nodes
  - Vast requires cost effective -> volunteer
- Domain: learn from time series
  - Finance, medical signals domain
- Solution is strategy or classifier expressed as rule sets

EC-Star Paradigm
Digital Directed Evolution of Models

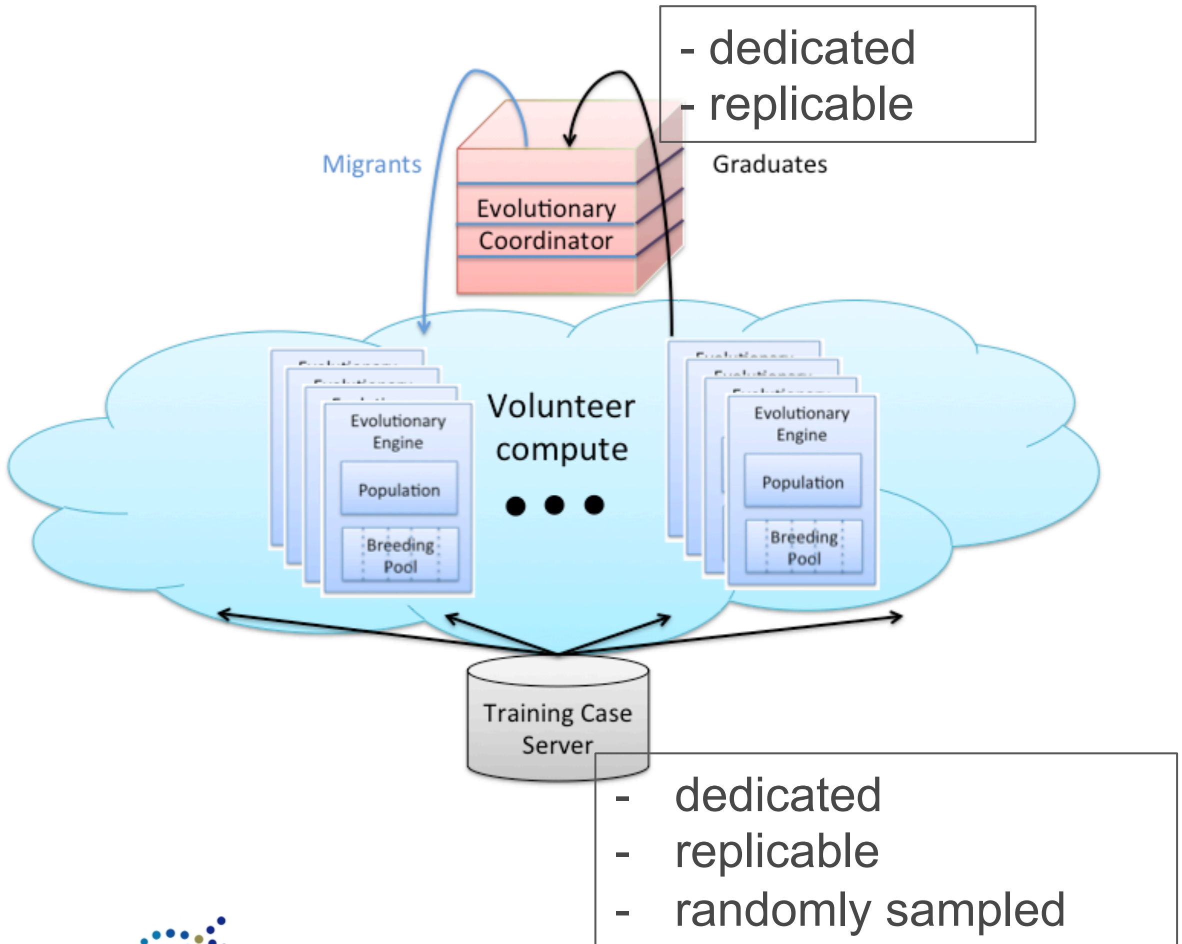


Use Case #1

ALFA
ANYSCALE LEARNING FOR ALL

MIT CSAIL

EC-Star



BigData factoring strategy

- **Use fewer training samples and cull poor models**
- **Increase training samples on better models**
- **Local and distributed randomization**
- **Oversampling on superior models**

System Layer Comparison

	Scale	FlexGP	EC-Star
ML domain	Classification	Regression Classification	Rule Learning
Resource Scale	10's to 100	100's to 1000	10^3 to 10^6
Resource Type	Cloud	Cloud	Volunteer and Dedicated
Fusion	External	External	Integrated
Local Algorithm	Different	Same	Same
Server:Client ratio	1: many	Decentralized	Few: many

ALFA ANYSCALE LEARNING FOR ALL

CSAIL

Scalable Machine Learning Comparison

	SCALE	FlexGP	EC-Star
Factor	Algorithm	Algorithm and Data	Data: Under to oversampling
Filter		Correlation Accuracy	Layered competition
Fuse		Non-parametric output space approaches	Migration and ancestral properties

ALFA
ANYSCALE LEARNING FOR ALL

MIT CSAIL

Automation

- **"In the end, the biggest bottleneck is not data or CPU cycles, but human cycles."**

ML requires a lot of Human Effort

Domain Knowledge Analysis and Transfer

Problem Definition

Data Preconditioning

Feature Identification

Feature Extraction

Algorithm Selection

Algorithm Customization

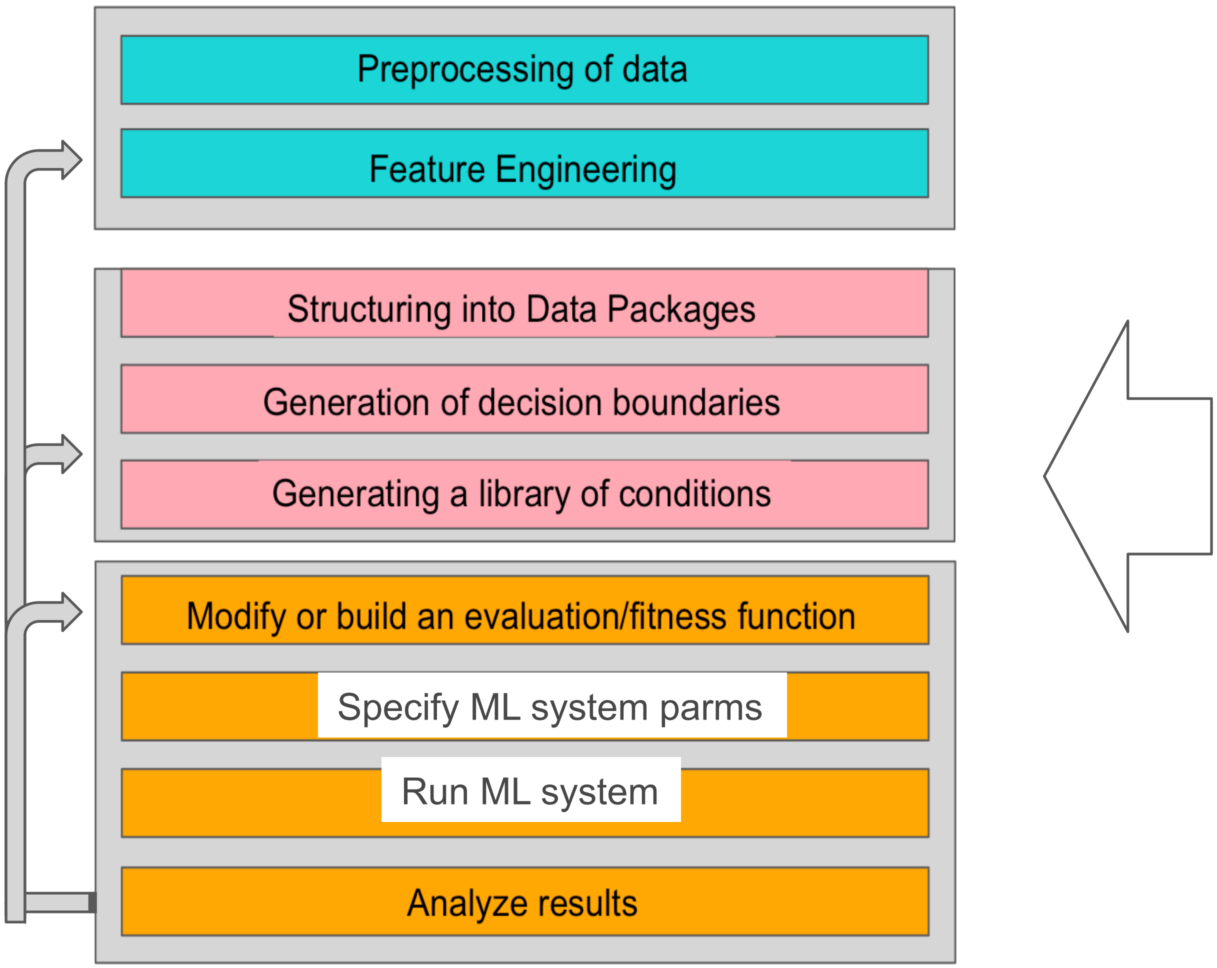
Parameter Selection

Training and Test Data Selection

Results Evaluation

Solution Deployment

Compressing the ML Endeavor



- **From desktop**
  - **Multi-core**
- **To GPU**
- **To Cloud**

Seamlessly!

Rapidly!

Flexibly!

Scalably!

Wrap Up

DCAP is open source: https://github.com/byterial/dcap

FlexGP is documented in publications and thesis

Owen Derby, MEng, 2013

Thanks to…

- **ALFA group members**
  - **Large team of students**
  - **Postdoc: Dr. Erik Hemberg**
  - **Research Scientist: Dr. Kalyan Veeramachaneni**
- **CSAIL Quanta cloud**
- **Our collaborators and sponsors**

genetic FINANCE

李嘉誠基金會
LI KA SHING FOUNDATION

imagination at work

We bring good things to life.

CSAIL

ALFA
ANYSCALE LEARNING FOR ALL