Pacific Northwest
NATIONAL LABORATORY

Classification without Representation: Interactive Machine Leaning at Scale with CHISSL

November 1, 2018

Dustin Arendt

Chesapeake Large-Scale Analytics Conference

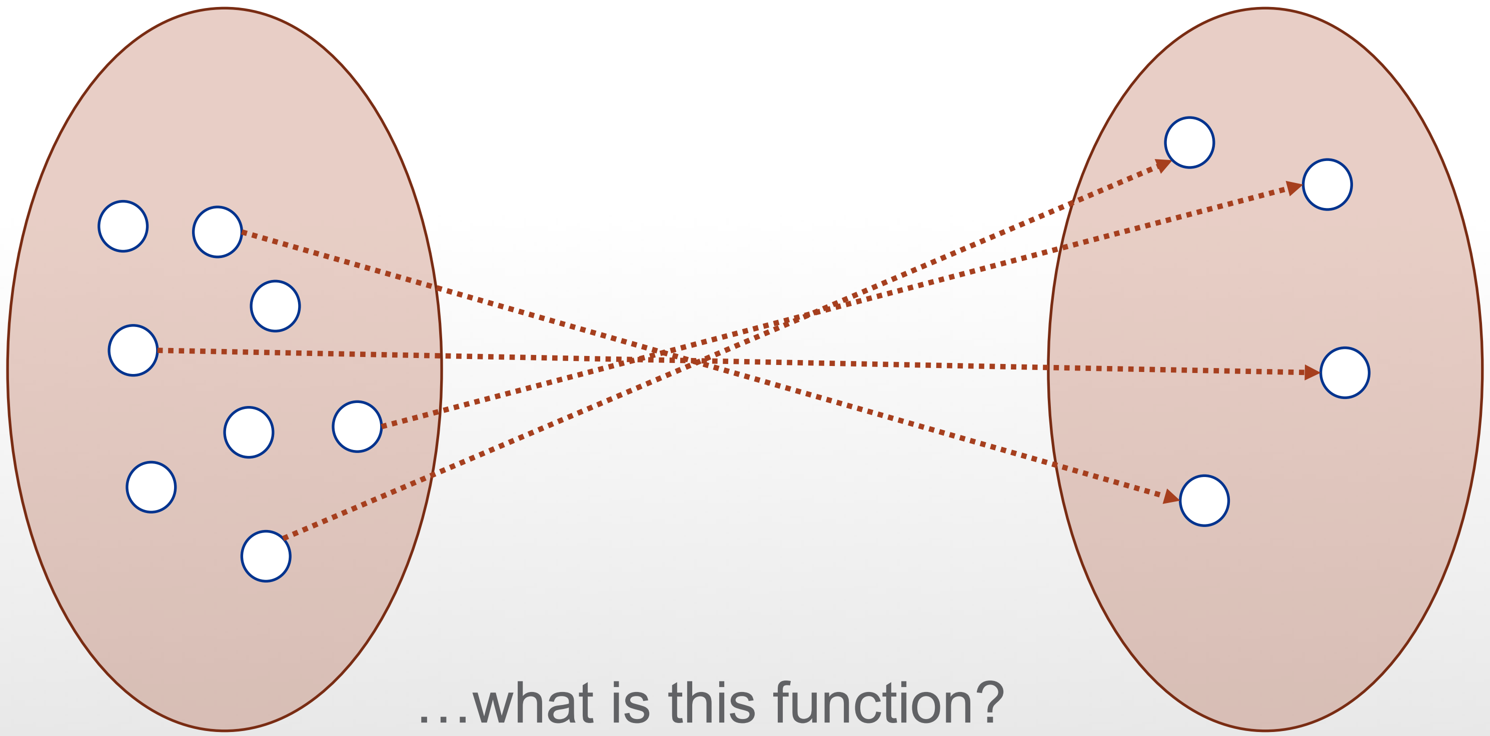
U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

Machine Learning in 30 Seconds

Given some input…

…and some output…



…what is this function?

ML Challenge: Eliciting Domain Knowledge

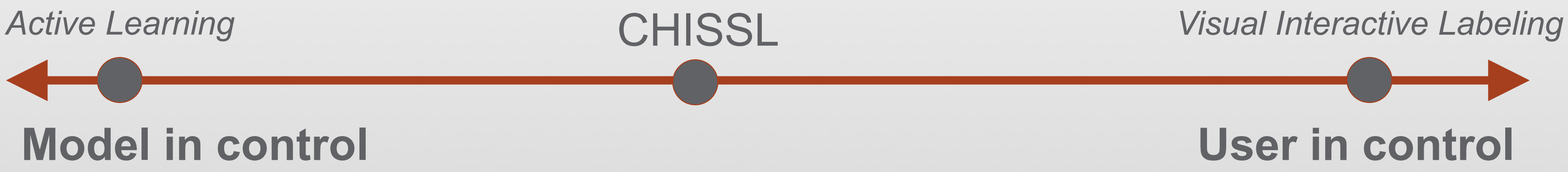
Data + **Model** = win

- Labels

Domain Knowledge

- Task Design
- Feature Engineering
- Model Selection

Label Elicitation Strategies



Active Learning

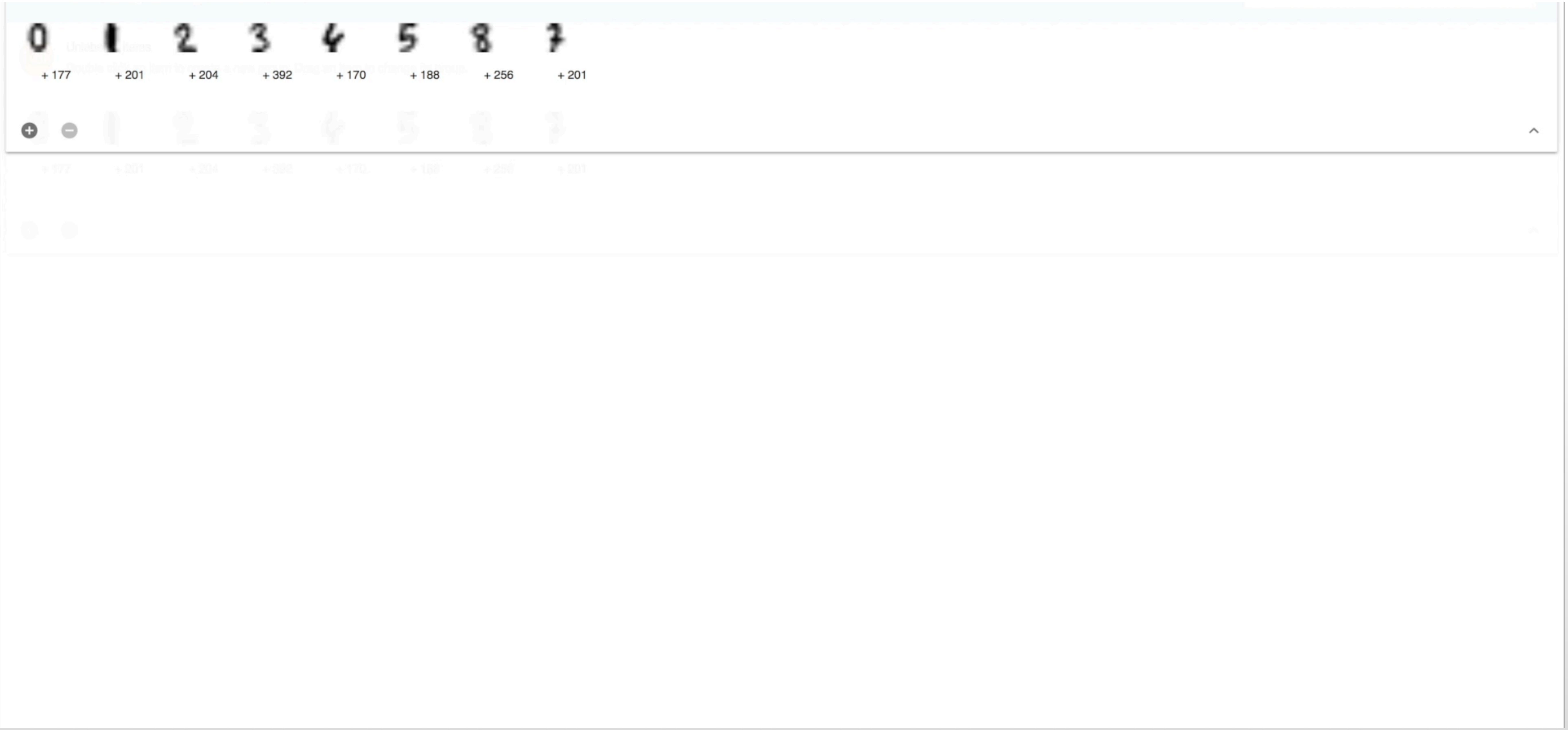
CHISSL

Visual Interactive Labeling

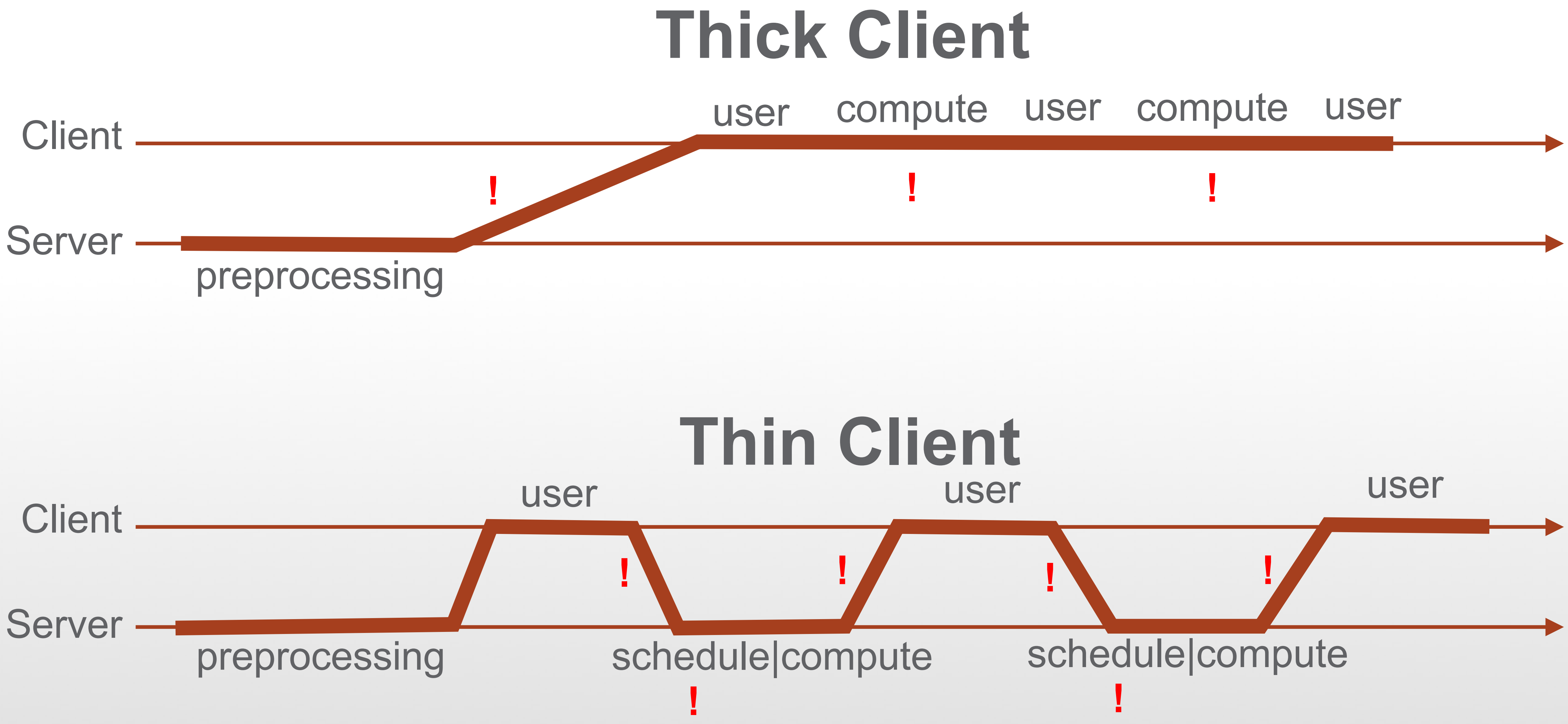
Model in control

User in control

CHISSL Demonstration

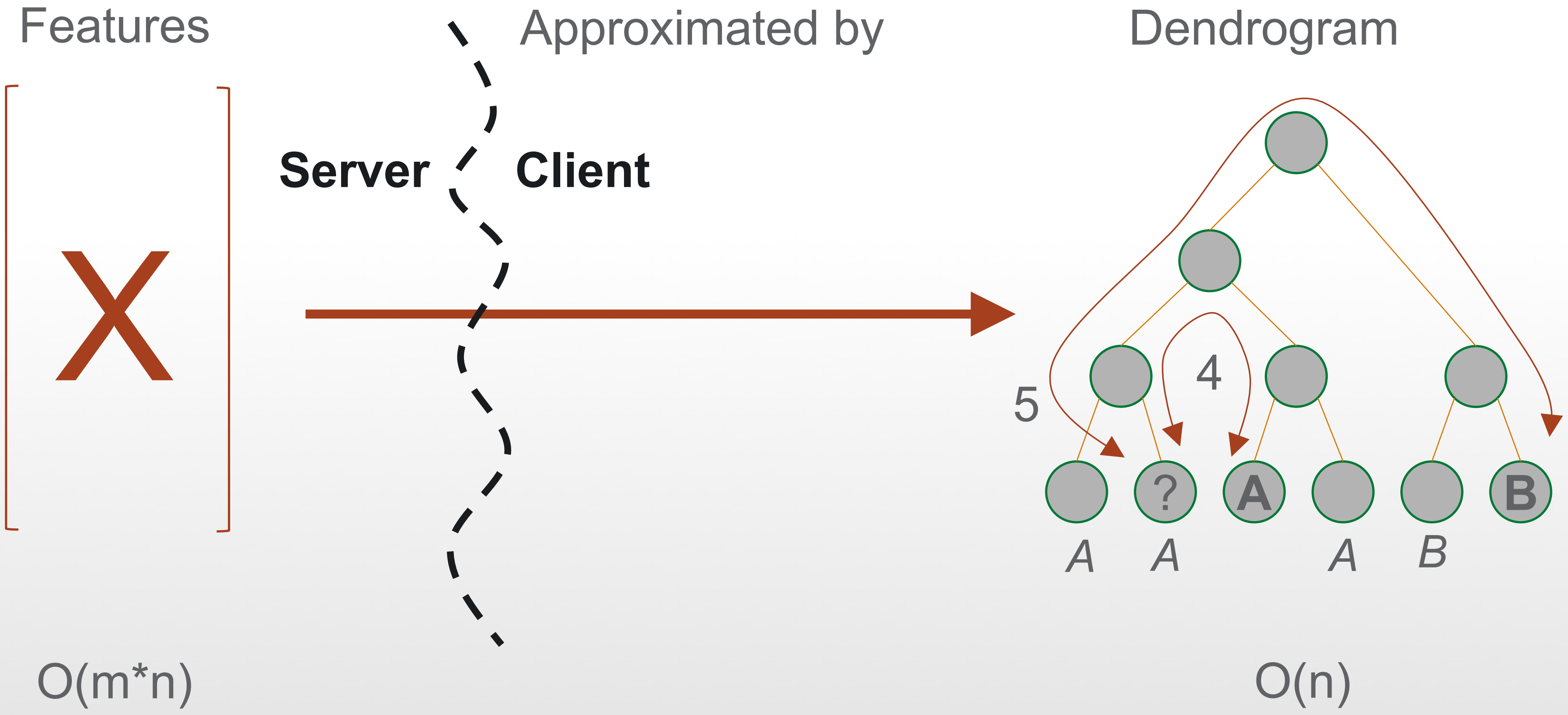


Design Considerations



! denotes a threat to scalability

Approach: Representation-free Classification


Features

Approximated by

Dendrogram

Server

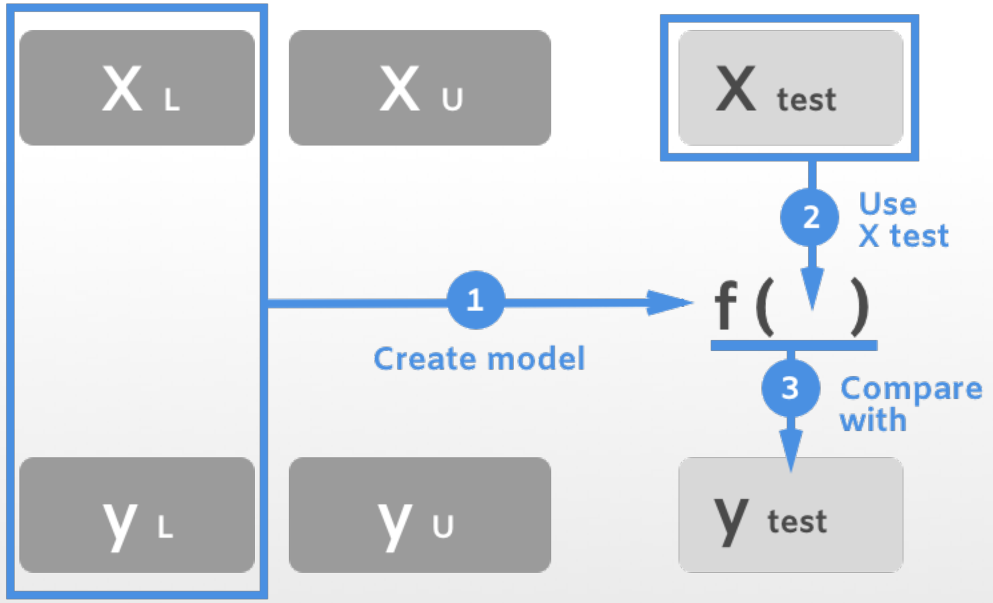
Client

O(m*n)

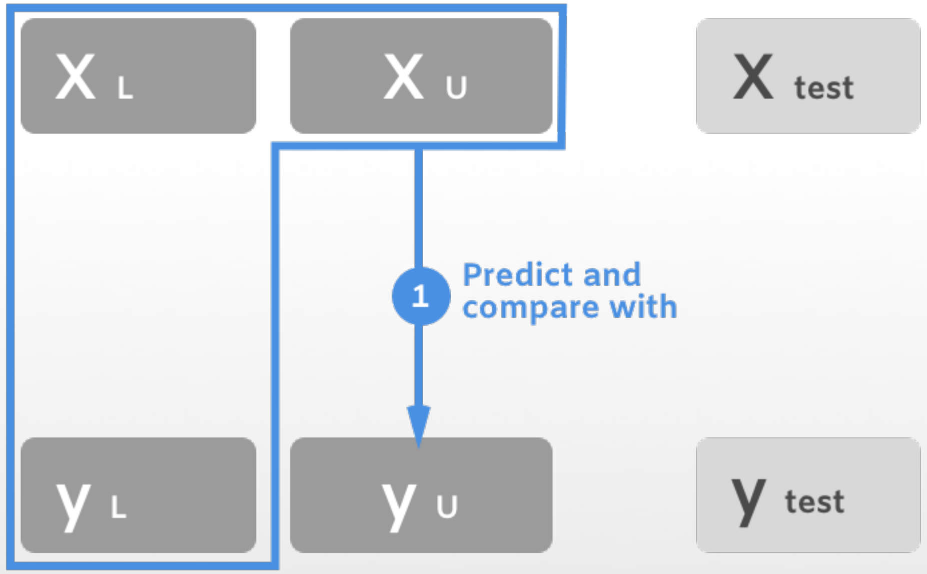
O(n)

Computational Evaluation

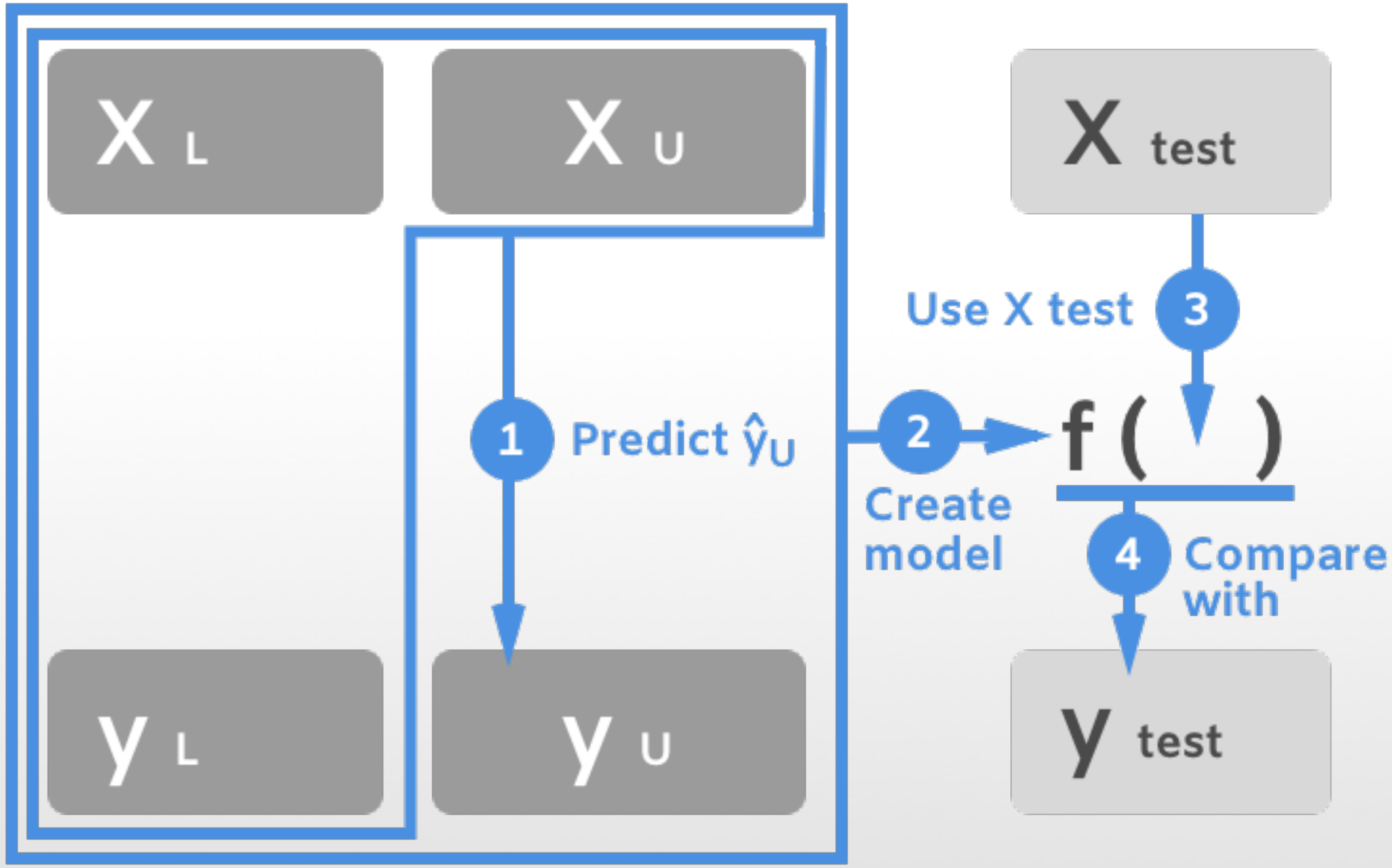
Induction



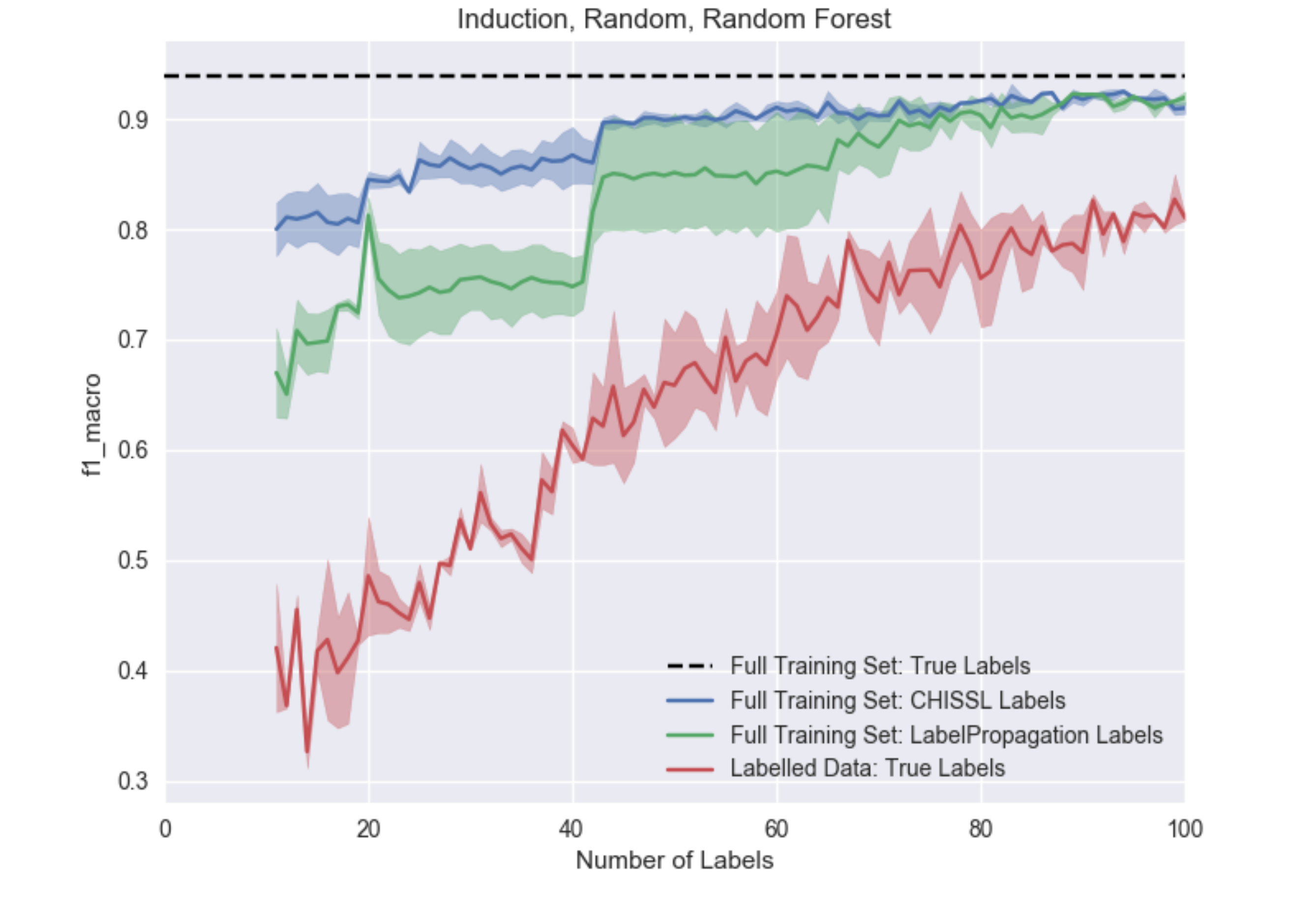
Transduction

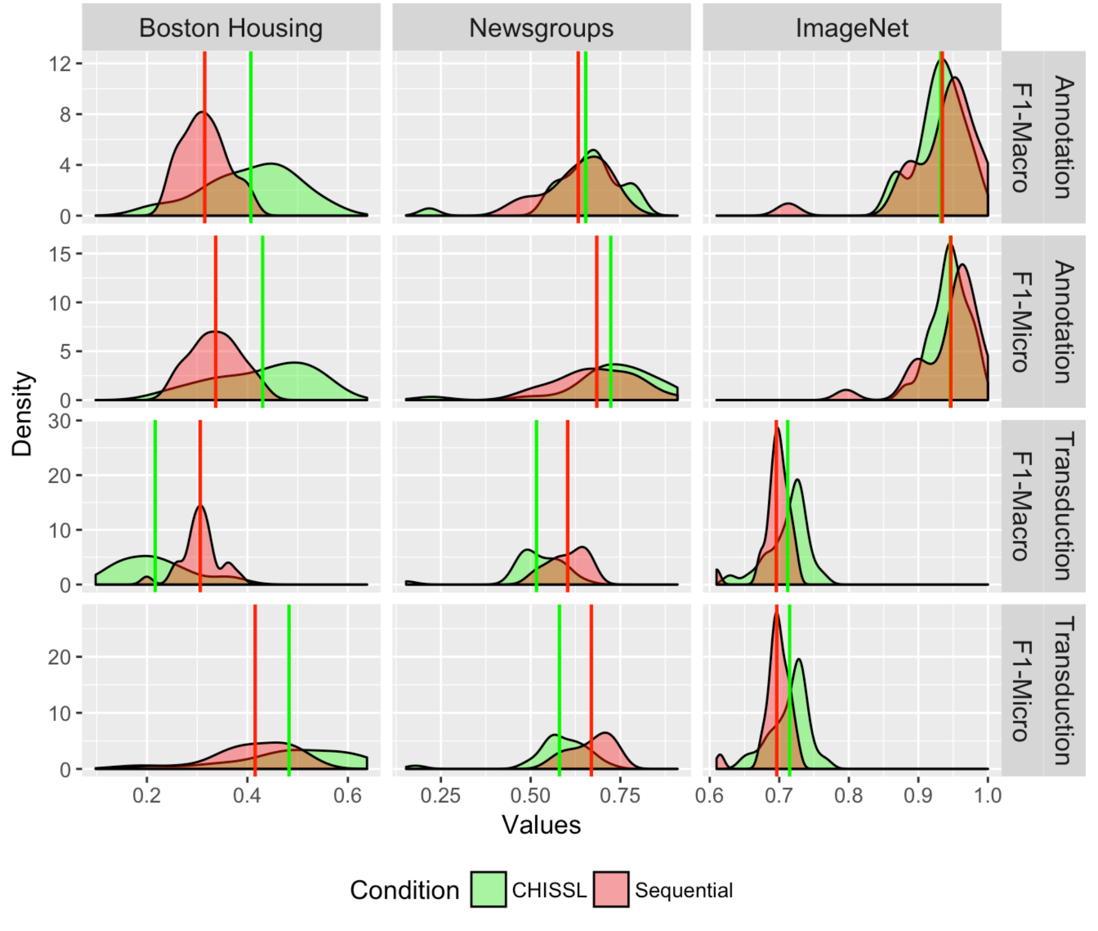


Bootstrapped Induction

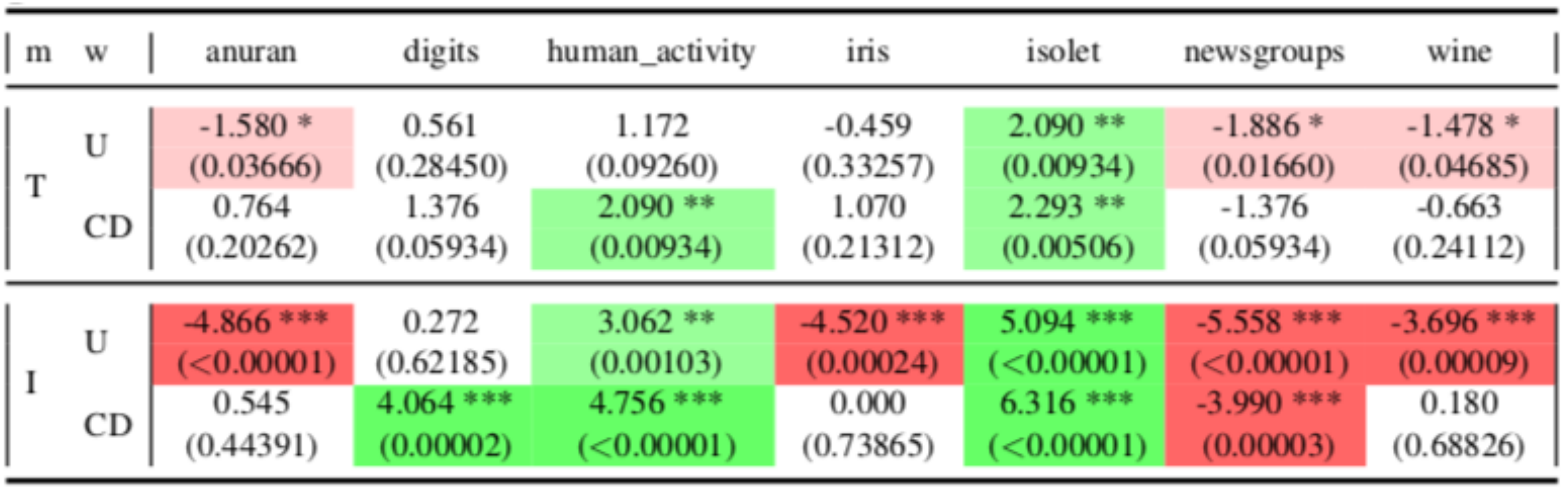


Results: Faster, Smaller, Better



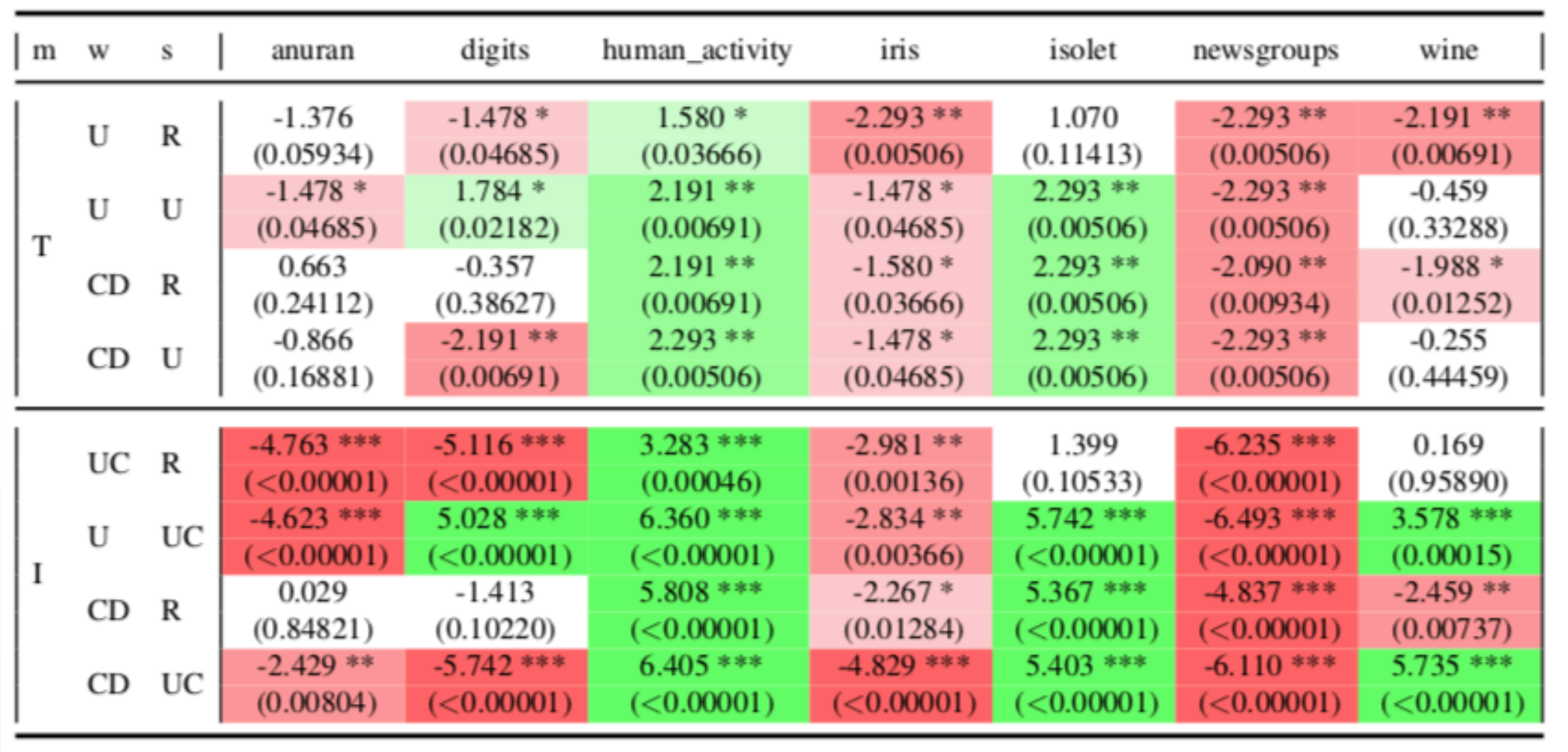


Results: Statistical Tests



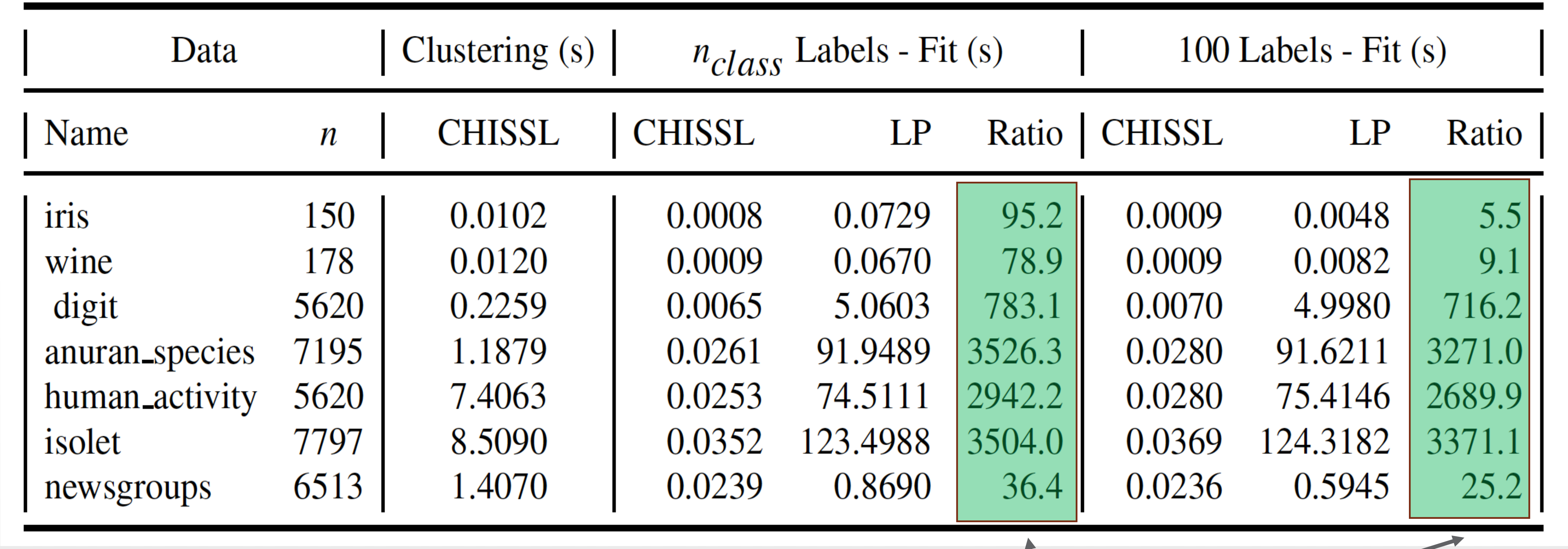
m	w	anuran	digits	human_activity	iris	isolet	newsgroups	wine
T	U	-1.580 * (0.03666)	0.561 (0.28450)	1.172 (0.09260)	-0.459 (0.33257)	2.090 ** (0.00934)	-1.886 * (0.01660)	-1.478 * (0.04685)
	CD	0.764 (0.20262)	1.376 (0.05934)	2.090 ** (0.00934)	1.070 (0.21312)	2.293 ** (0.00506)	-1.376 (0.05934)	-0.663 (0.24112)
I	U	-4.866 *** (<0.00001)	0.272 (0.62185)	3.062 ** (0.00103)	-4.520 *** (0.00024)	5.094 *** (<0.00001)	-5.558 *** (<0.00001)	-3.696 *** (0.00009)
	CD	0.545 (0.44391)	4.064 *** (0.00002)	4.756 *** (<0.00001)	0.000 (0.73865)	6.316 *** (<0.00001)	-3.990 *** (0.00003)	0.180 (0.68826)

Results: Statistical Tests



m	w	s	anuran	digits	human_activity	iris	isolet	newsgroups	wine
T	U	R	-1.376 (0.05934)	-1.478 * (0.04685)	1.580 * (0.03666)	-2.293 ** (0.00506)	1.070 (0.11413)	-2.293 ** (0.00506)	-2.191 ** (0.00691)
	U	U	-1.478 * (0.04685)	1.784 * (0.02182)	2.191 ** (0.00691)	-1.478 * (0.04685)	2.293 ** (0.00506)	-2.293 ** (0.00506)	-0.459 (0.33288)
	CD	R	0.663 (0.24112)	-0.357 (0.38627)	2.191 ** (0.00691)	-1.580 * (0.03666)	2.293 ** (0.00506)	-2.090 ** (0.00934)	-1.988 * (0.01252)
	CD	U	-0.866 (0.16881)	-2.191 ** (0.00691)	2.293 ** (0.00506)	-1.478 * (0.04685)	2.293 ** (0.00506)	-2.293 ** (0.00506)	-0.255 (0.44459)
I	UC	R	-4.763 *** (<0.00001)	-5.116 *** (<0.00001)	3.283 *** (0.00046)	-2.981 ** (0.00136)	1.399 (0.10533)	-6.235 *** (<0.00001)	0.169 (0.95890)
	U	UC	-4.623 *** (<0.00001)	5.028 *** (<0.00001)	6.360 *** (<0.00001)	-2.834 ** (0.00366)	5.742 *** (<0.00001)	-6.493 *** (<0.00001)	3.578 *** (0.00015)
	CD	R	0.029 (0.84821)	-1.413 (0.10220)	5.808 *** (<0.00001)	-2.267 * (0.01284)	5.367 *** (<0.00001)	-4.837 *** (<0.00001)	-2.459 ** (0.00737)
	CD	UC	-2.429 ** (0.00804)	-5.742 *** (<0.00001)	6.405 *** (<0.00001)	-4.829 *** (<0.00001)	5.403 *** (<0.00001)	-6.110 *** (<0.00001)	5.735 *** (<0.00001)

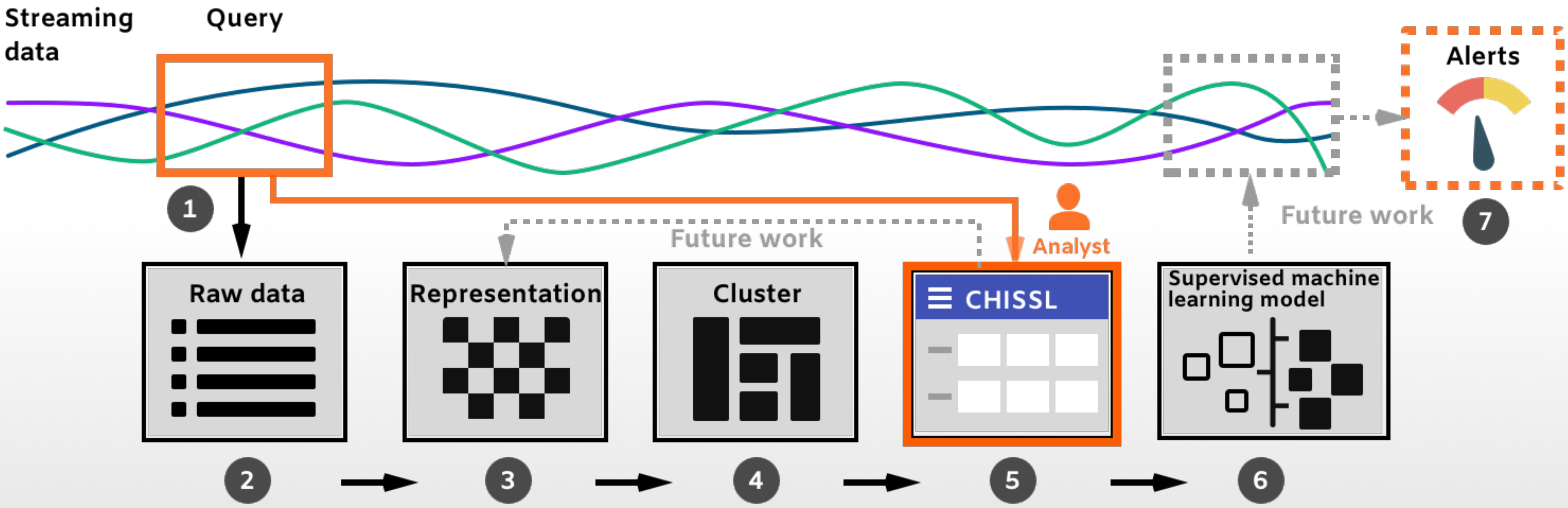
Results: Speed Benchmarks



Data		Clustering (s)	n_{class} Labels - Fit (s)			100 Labels - Fit (s)		
Name	n	CHISSL	CHISSL	LP	Ratio	CHISSL	LP	Ratio
iris	150	0.0102	0.0008	0.0729	95.2	0.0009	0.0048	5.5
wine	178	0.0120	0.0009	0.0670	78.9	0.0009	0.0082	9.1
digit	5620	0.2259	0.0065	5.0603	783.1	0.0070	4.9980	716.2
anuran_species	7195	1.1879	0.0261	91.9489	3526.3	0.0280	91.6211	3271.0
human_activity	5620	7.4063	0.0253	74.5111	2942.2	0.0280	75.4146	2689.9
isolet	7797	8.5090	0.0352	123.4988	3504.0	0.0369	124.3182	3371.1
newsgroups	6513	1.4070	0.0239	0.8690	36.4	0.0236	0.5945	25.2

Speed improvement factor relative
to Label Propagation

Future Work: the "Big Picture"



Conclusions

- Rapid—much faster than baselines
- Accurate—better than supervised and competitive with semi-supervised baselines
- Helpful—users gave more accurate labels and built more accurate models
- Application Domains
  - Geo-temporal analysis
  - Insider threat detection

- Questions? Contact me.
  - Dustin.Arendt@pnnl.gov

Available on GitHub:
https://github.com/pnnl/chissl