Better Machine Learning Through Data

Saleema Amershi
Machine Teaching Group
Microsoft Research

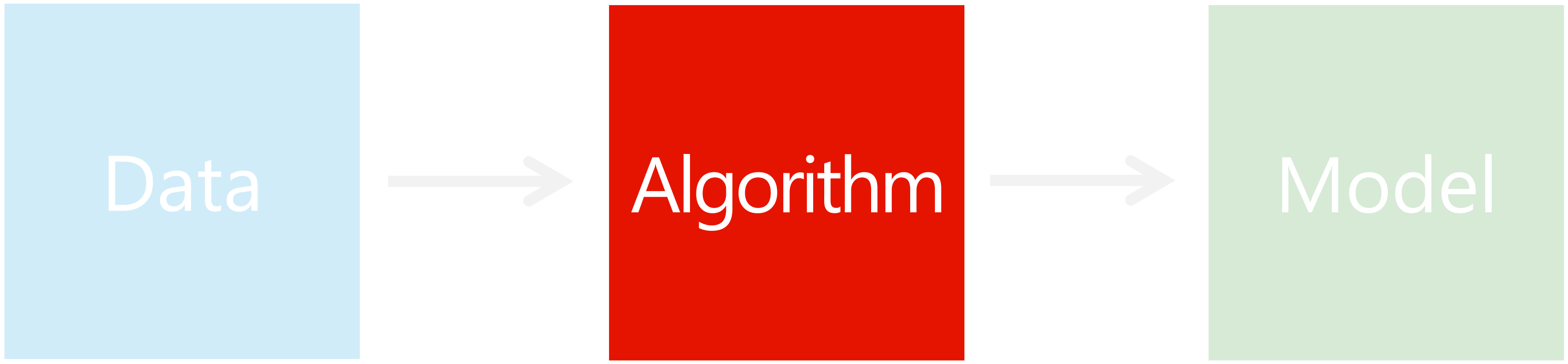
August 14, 2016

Making better sense of data.

Better data makes better machine learning.

Data + Algorithm = Model

Data + Algorithm = Model



Machine learning research often
takes the data as given.

When Algorithms Discriminate *– The New York Times, 2015*

Big Data's all-too-human failings *– Reuters, 2016*

Artificial Intelligence's White Guy Problem
– The New York Times, 2016

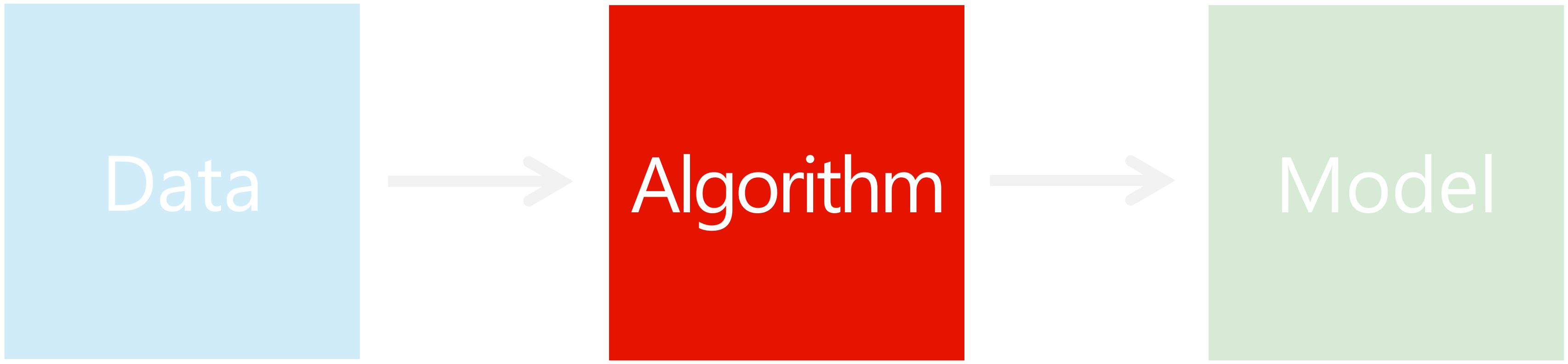
Mapping Crime – Or Stirring Hate?*– Financial Times, 2014*

Making better sense of data.

Better data makes better machine learning.

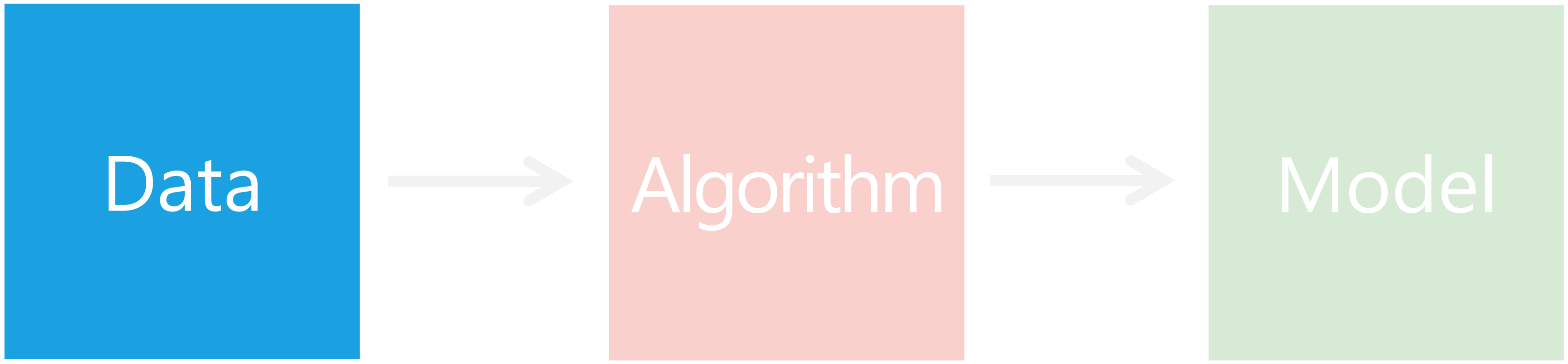
Most influence practitioners have on machine learning is through data.

Data + Algorithm = Model



In research,
data is often
taken as given.

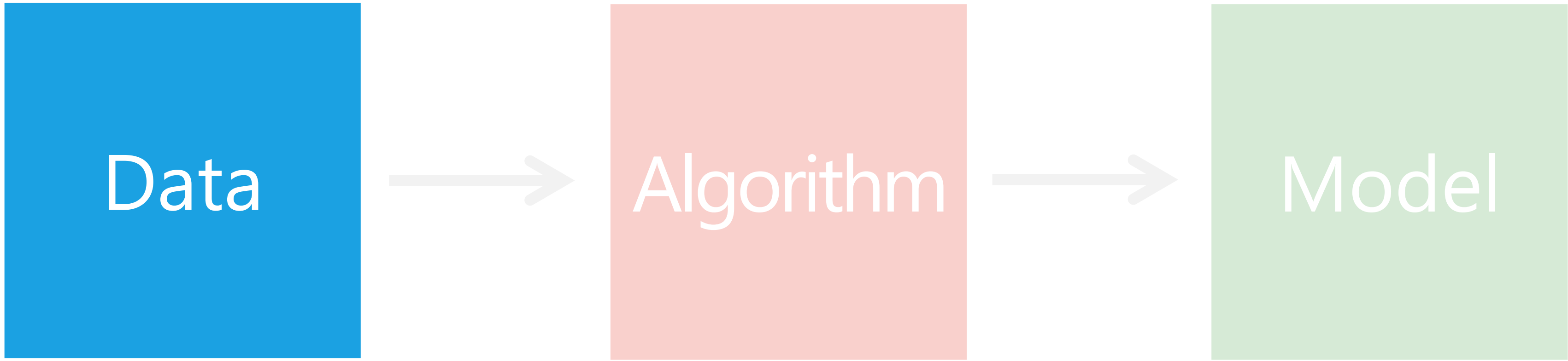
Data + Algorithm = Model



In practice, the algorithm is often taken as given.

In research, data is often taken as given.

Data + Algorithm = Model



In practice, the algorithm is often taken as given.

"Data scientists, according to interviews and expert estimates, spend ***50 percent to 80 percent of their time*** *mired in this more mundane labor of collecting and preparing unruly digital data."*

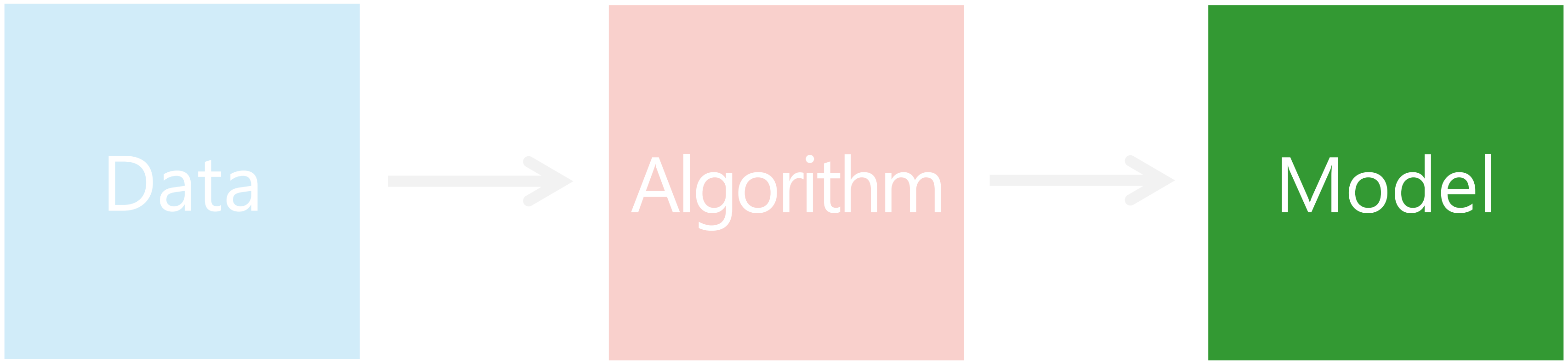
- New York Times, 2014

Data + Algorithm = Model

[Patel et al., CHI 2008]

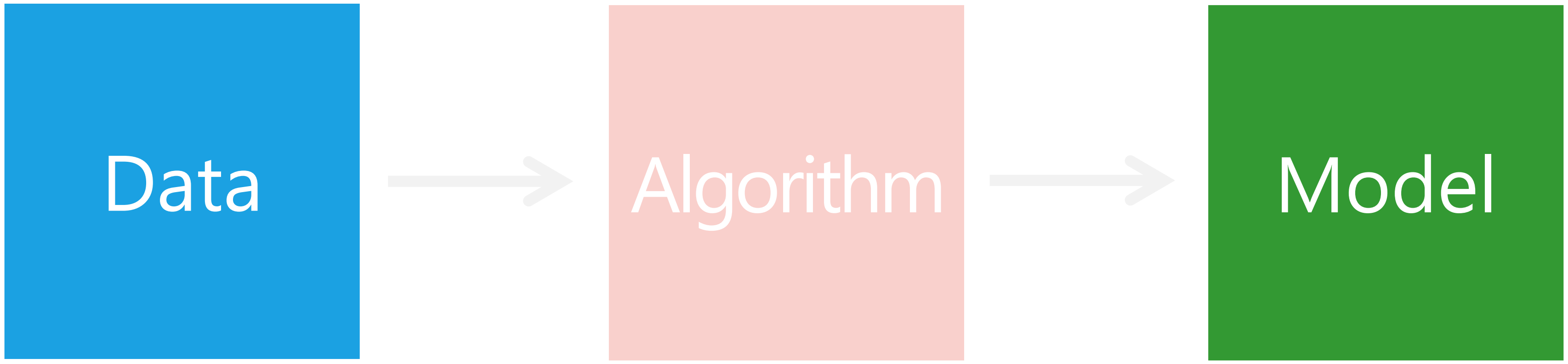
Data + Algorithm = Model

Data + Algorithm = Model



Iterations are driven by evaluating models on data.

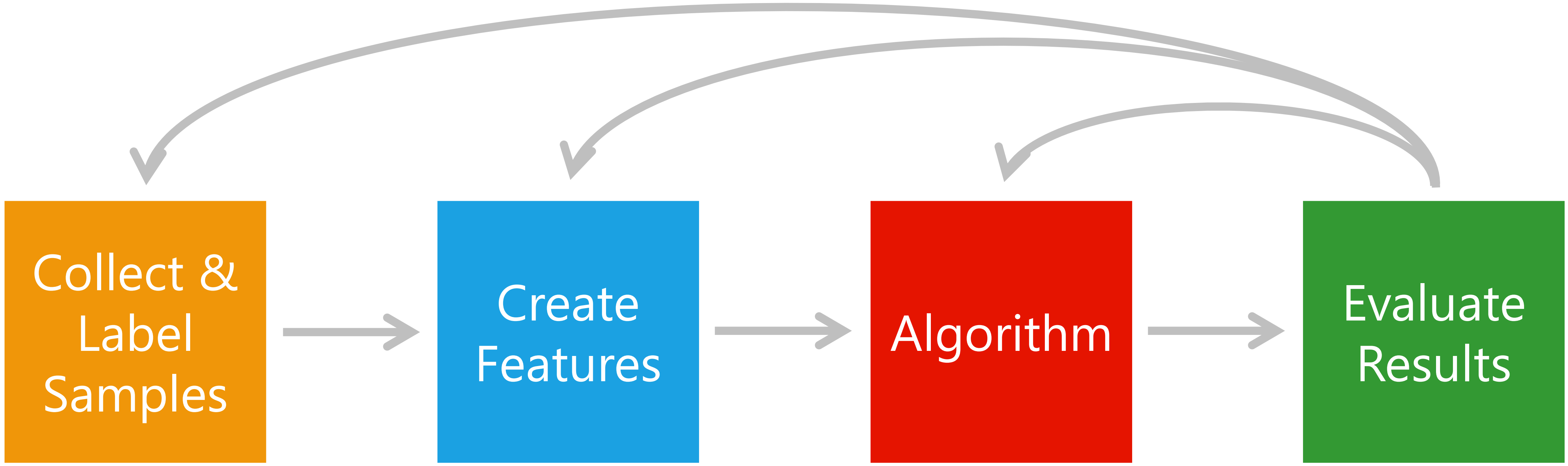
Data + Algorithm = Model



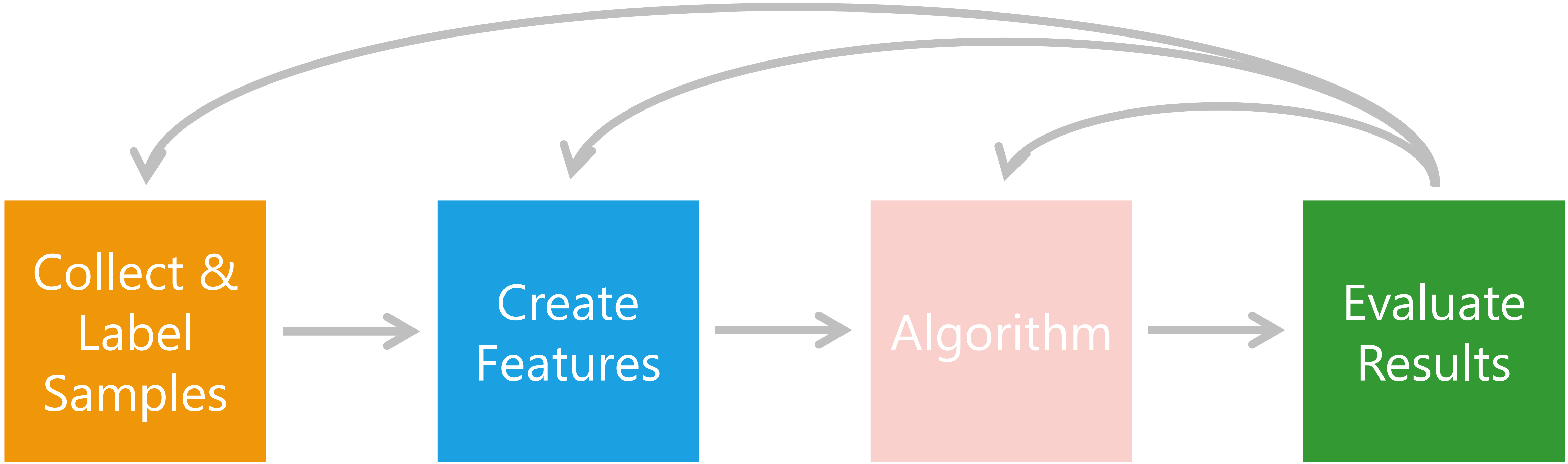
In practice, most effort is spent crafting input data.

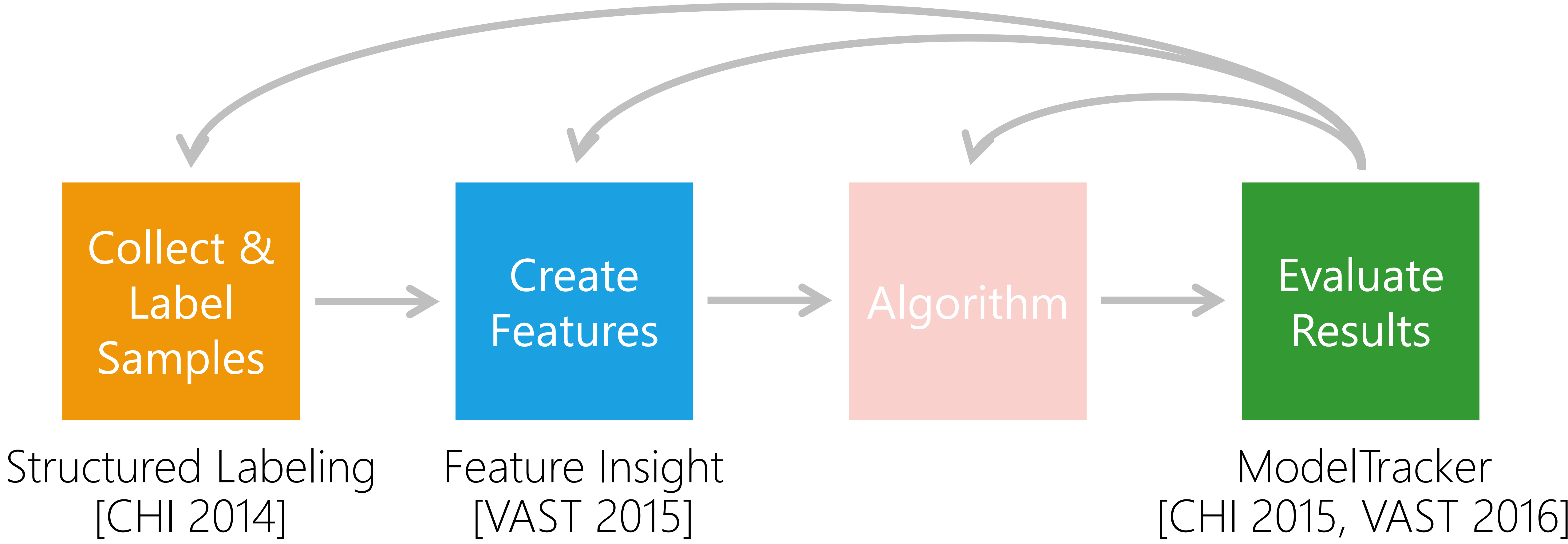
Iterations are driven by evaluating models on data.

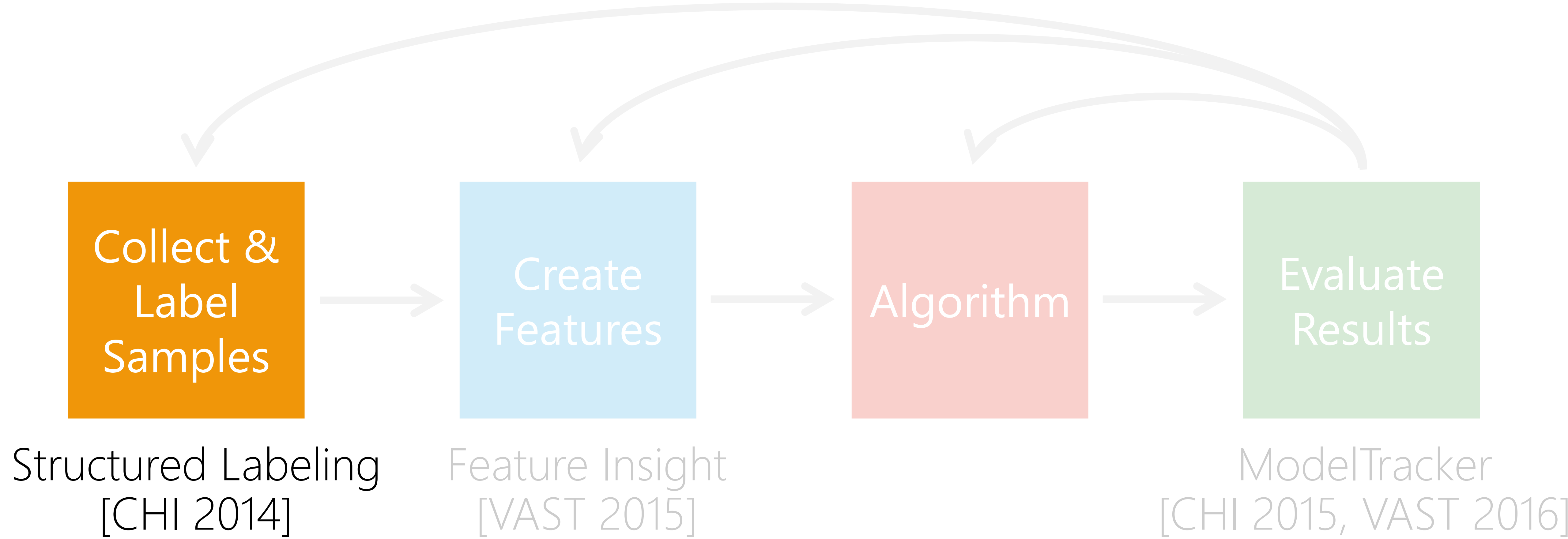
Machine learning in theory



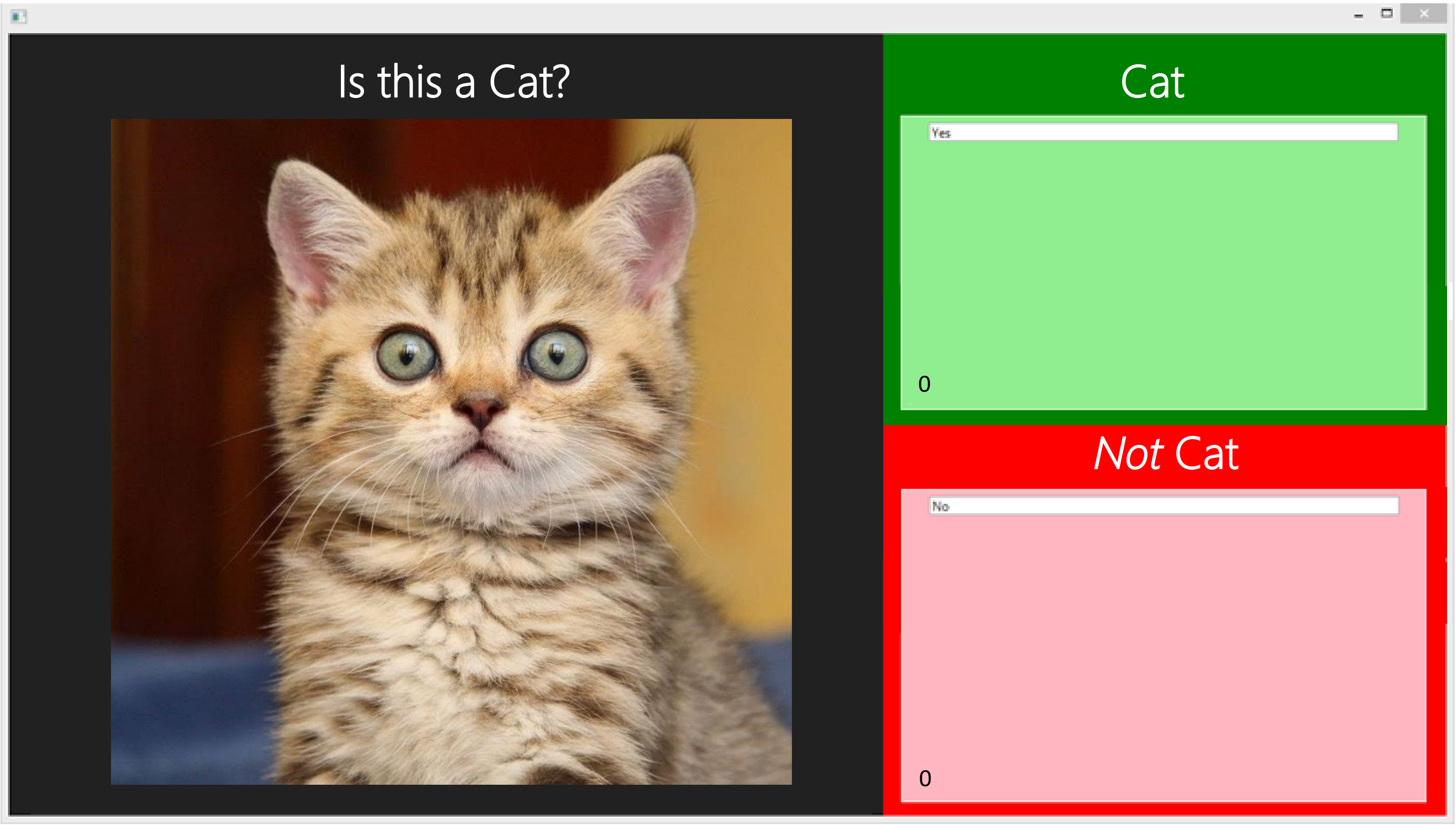
Machine learning in practice


Collect & Label Samples
Create Features
Algorithm
Evaluate Results


Collect & Label Samples
Create Features
Algorithm
Evaluate Results
Structured Labeling
[CHI 2014]
Feature Insight
[VAST 2015]
ModelTracker
[CHI 2015, VAST 2016]

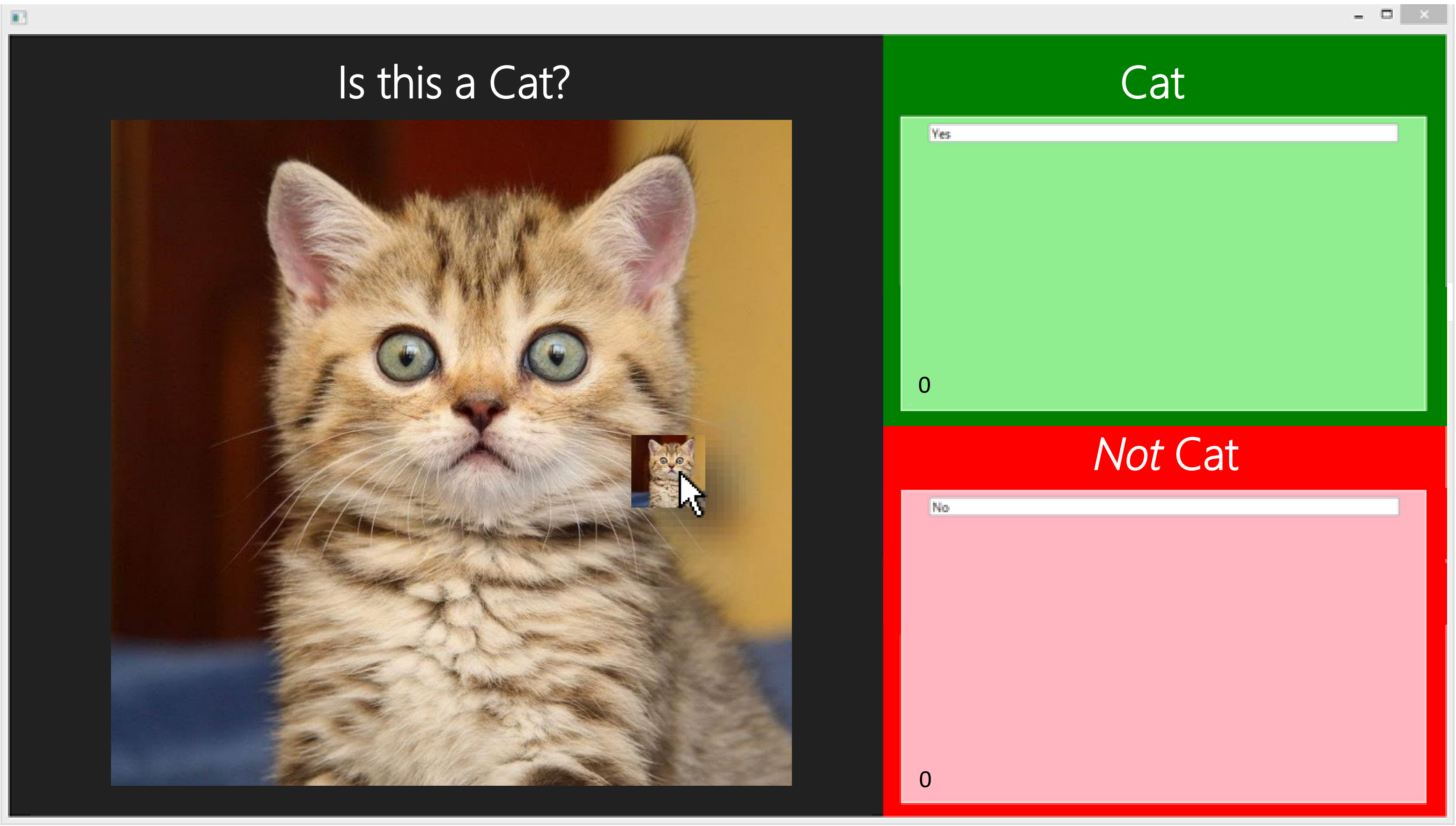

Collect & Label Samples
Create Features
Algorithm
Evaluate Results
Structured Labeling
[CHI 2014]
Feature Insight
[VAST 2015]
ModelTracker
[CHI 2015, VAST 2016]

Traditional Labeling



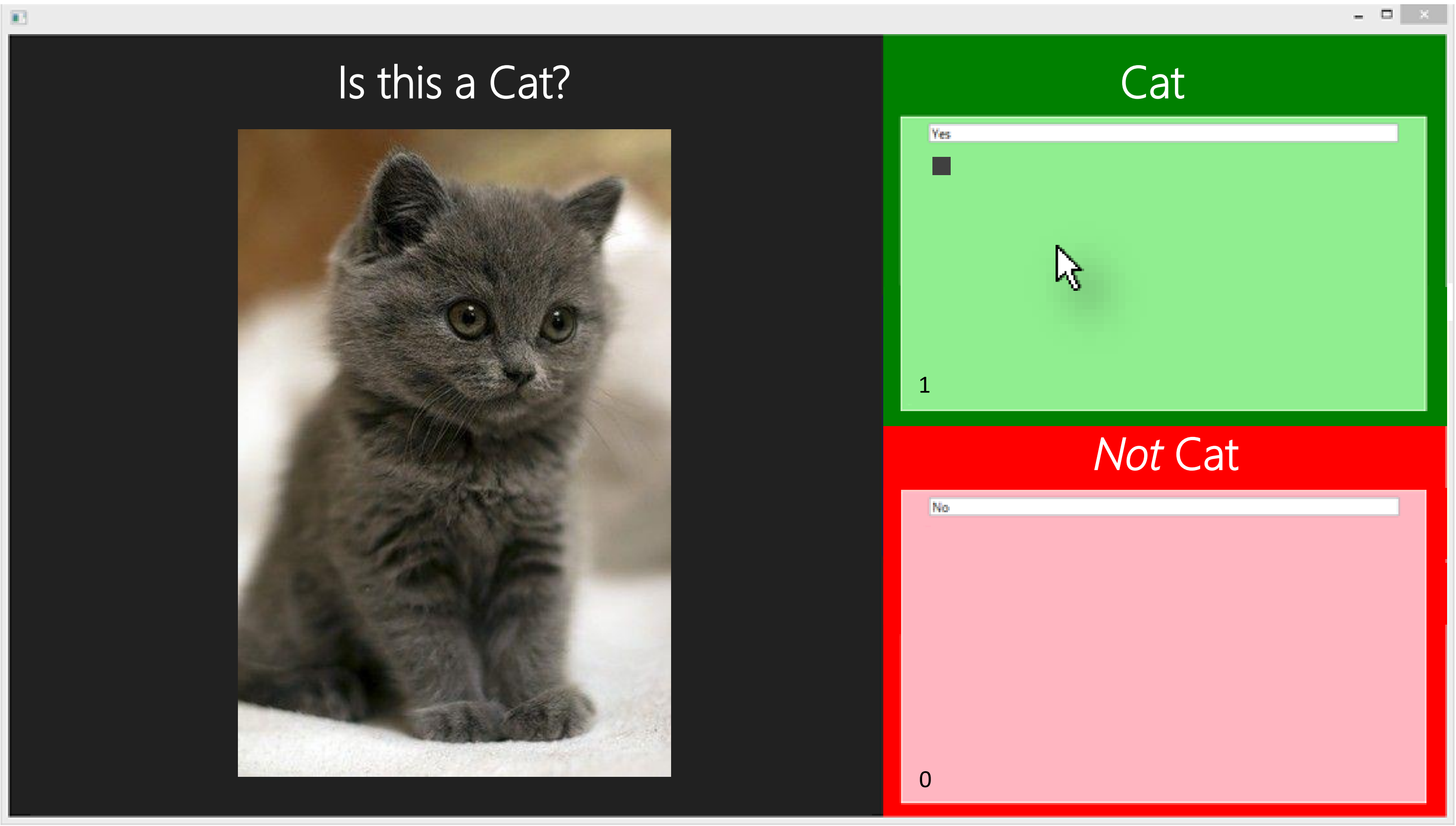
Pre-defined high-level categories.

Traditional Labeling



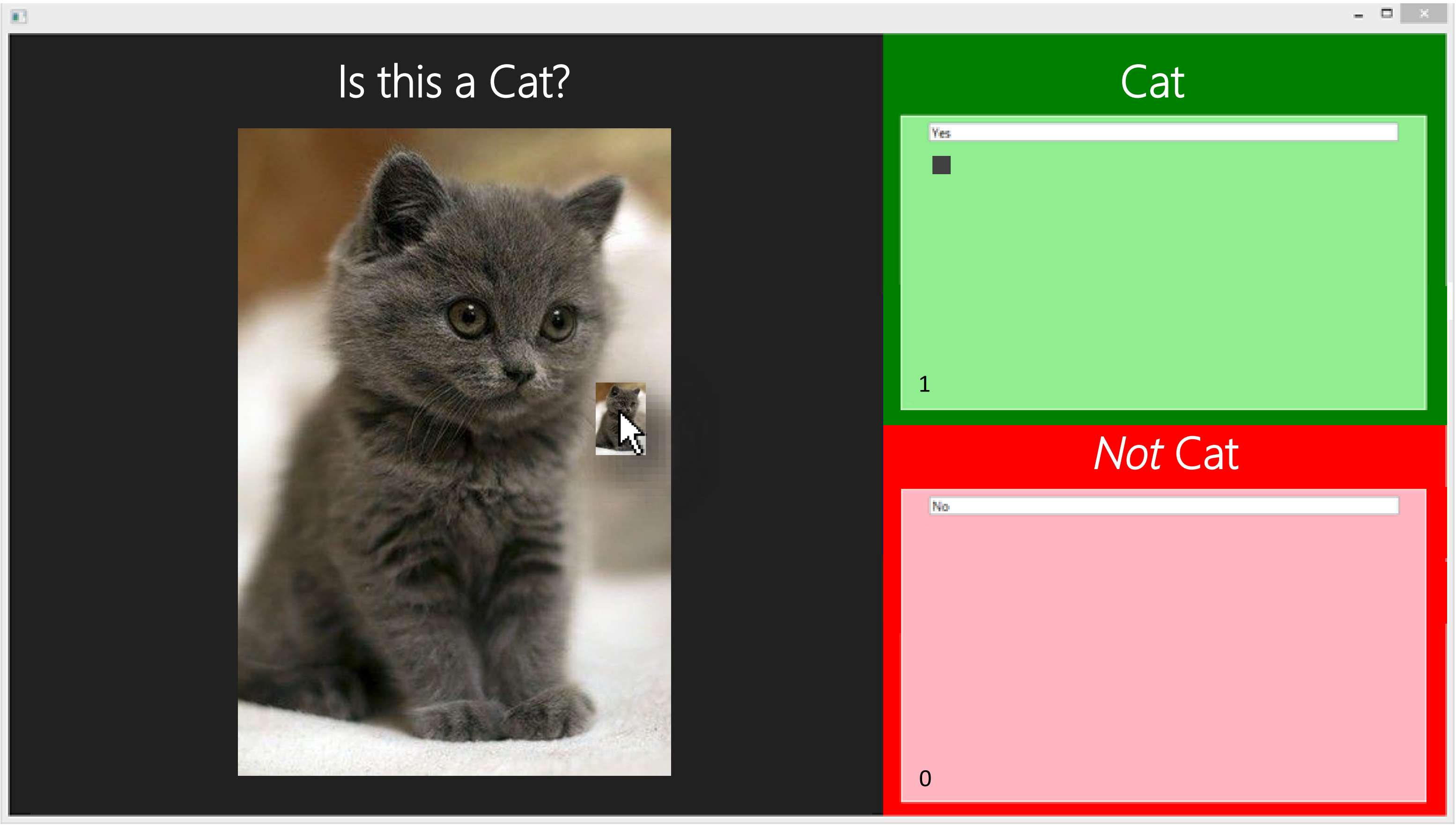
Pre-defined high-level categories.

Traditional Labeling



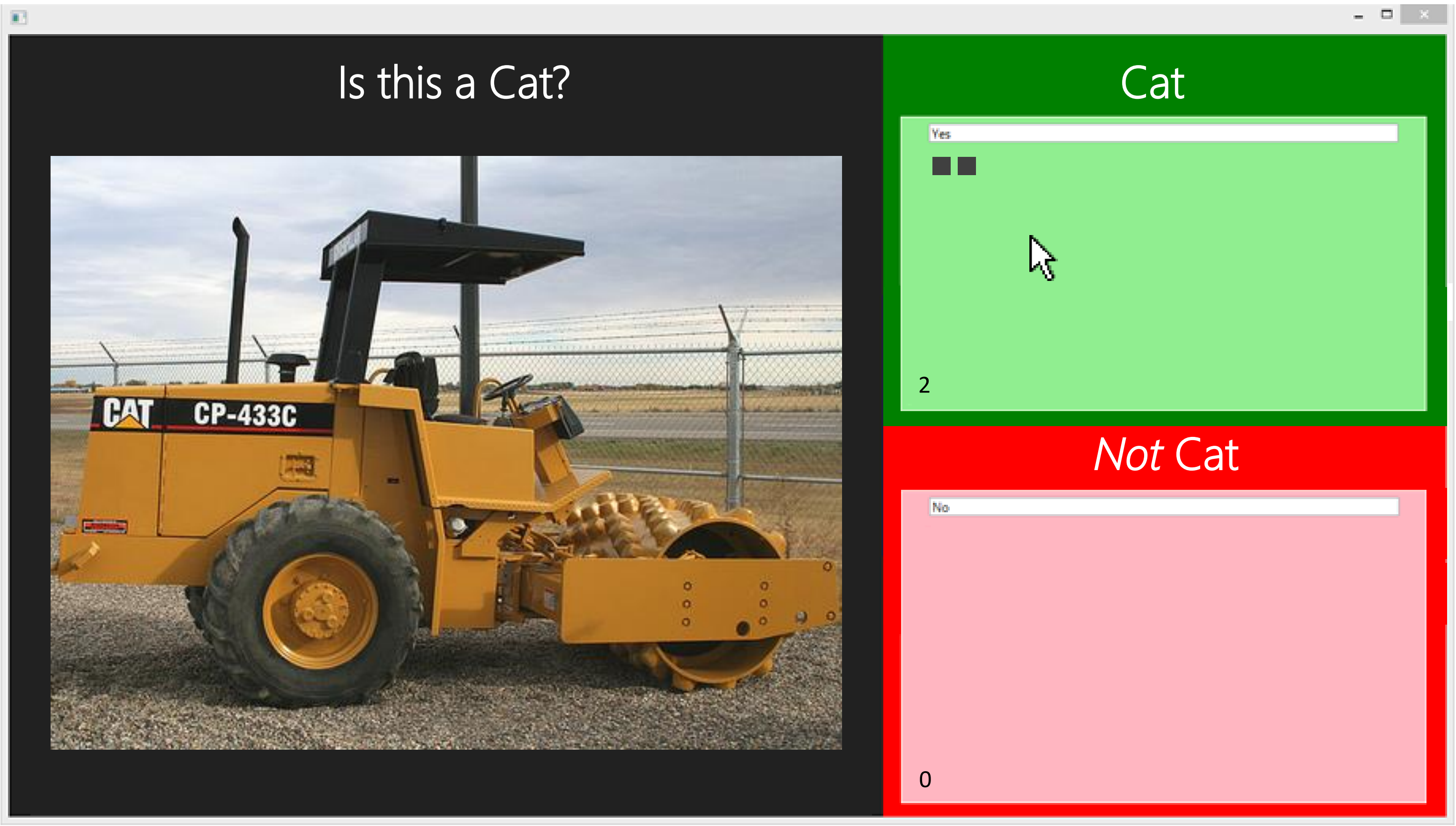
Pre-defined high-level categories.

Traditional Labeling



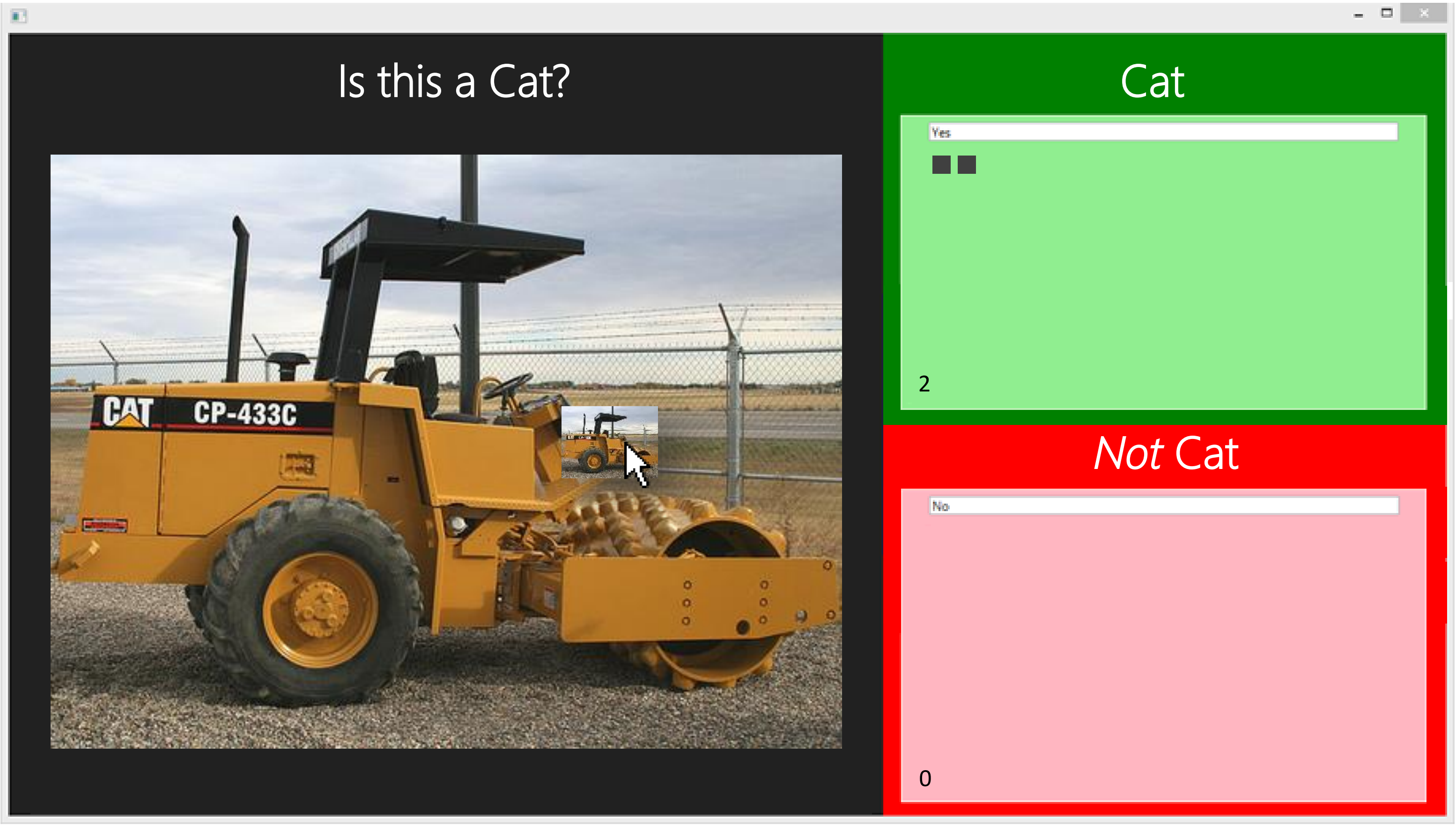
Pre-defined high-level categories.

Traditional Labeling



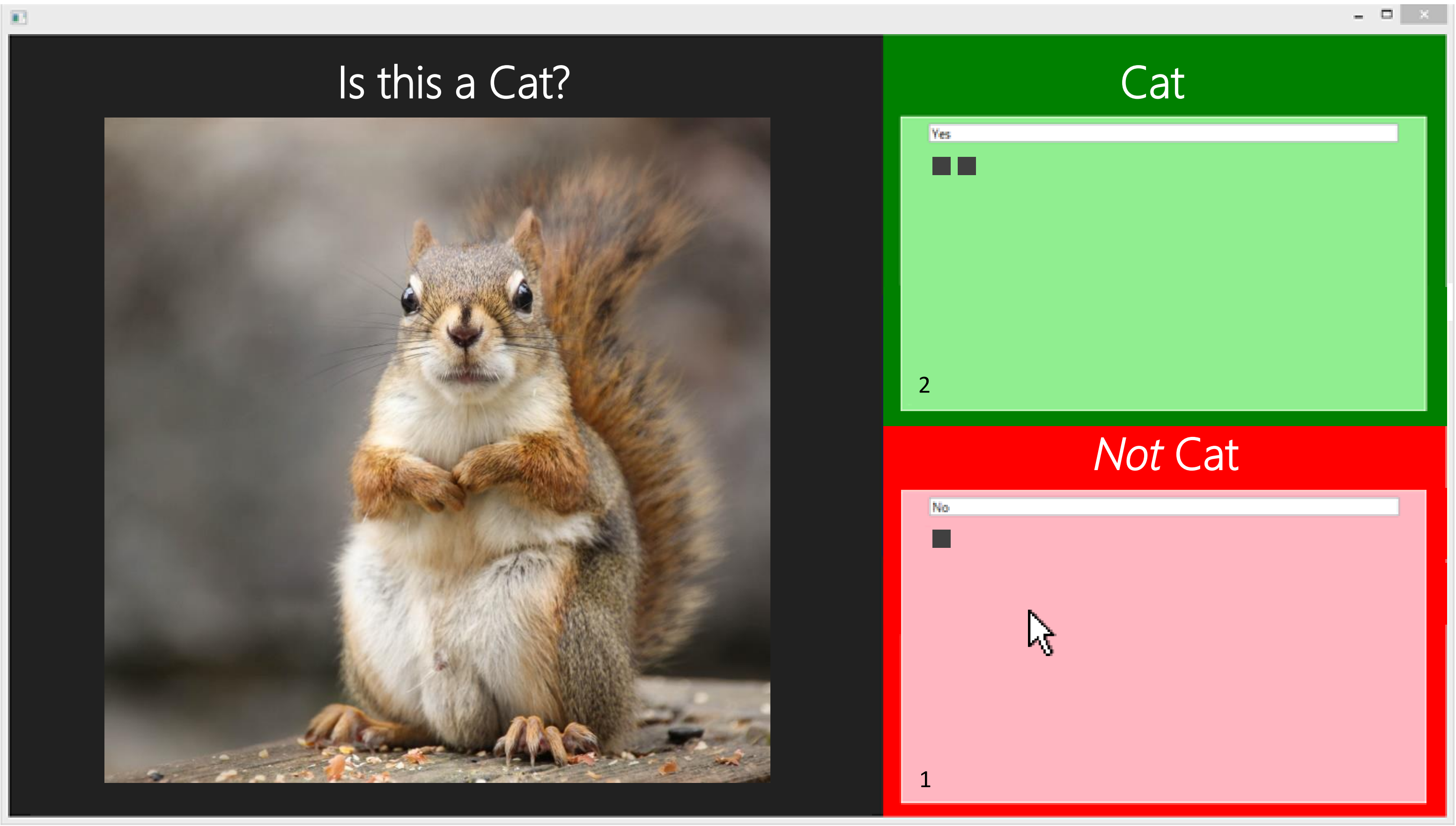
Pre-defined high-level categories.

Traditional Labeling



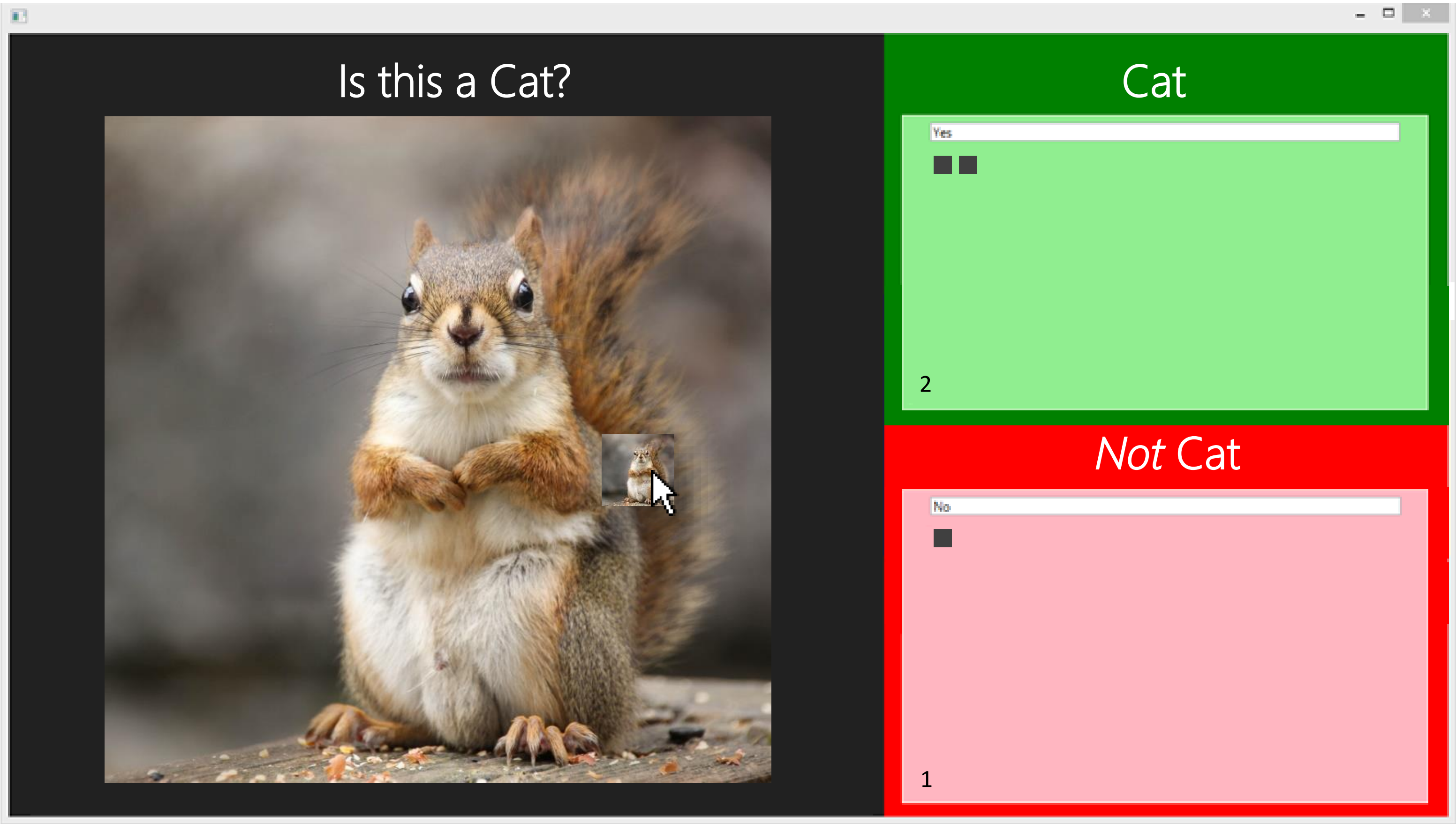
Pre-defined high-level categories.

Traditional Labeling



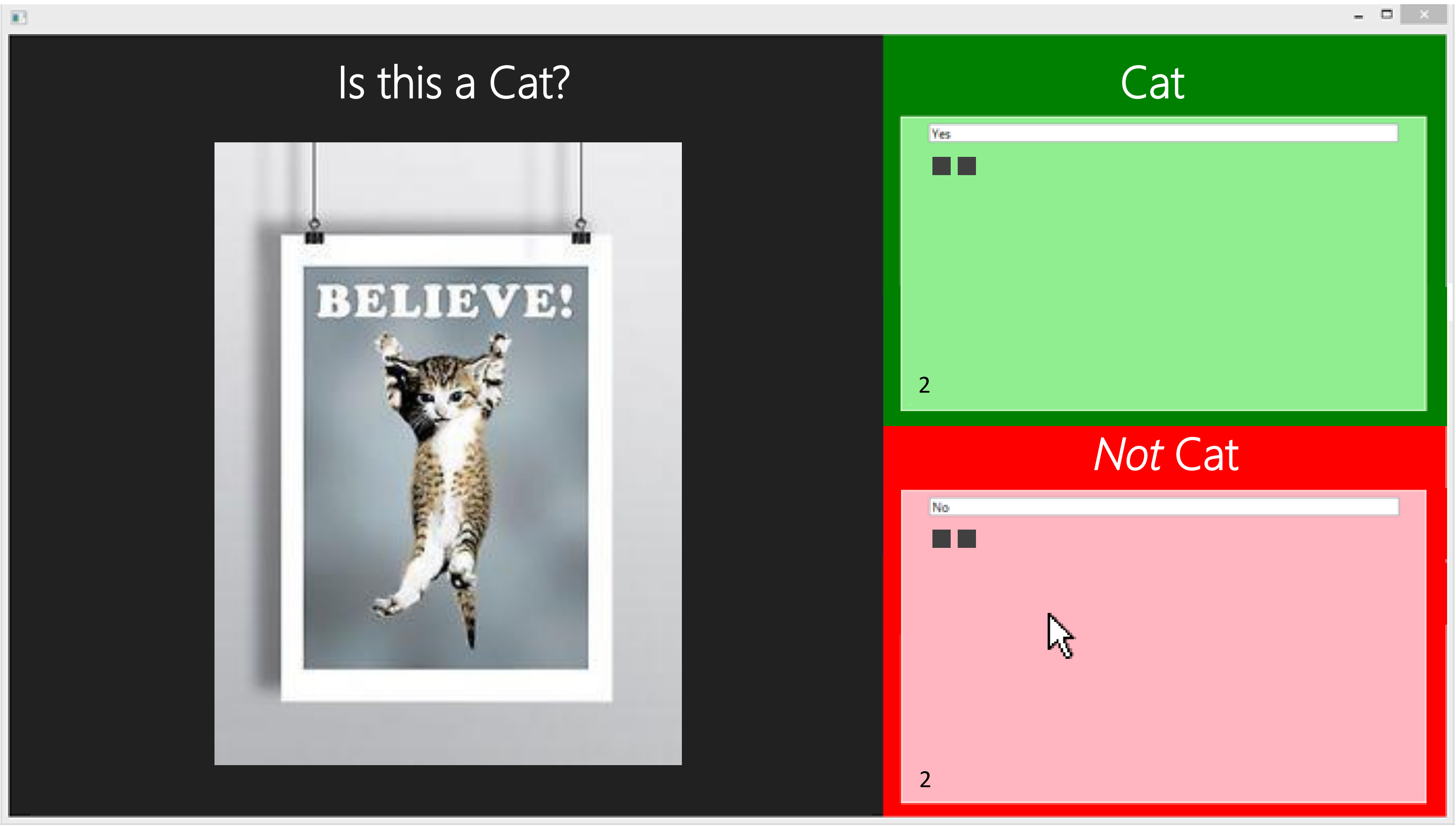
Pre-defined high-level categories.

Traditional Labeling



Pre-defined high-level categories.

Traditional Labeling

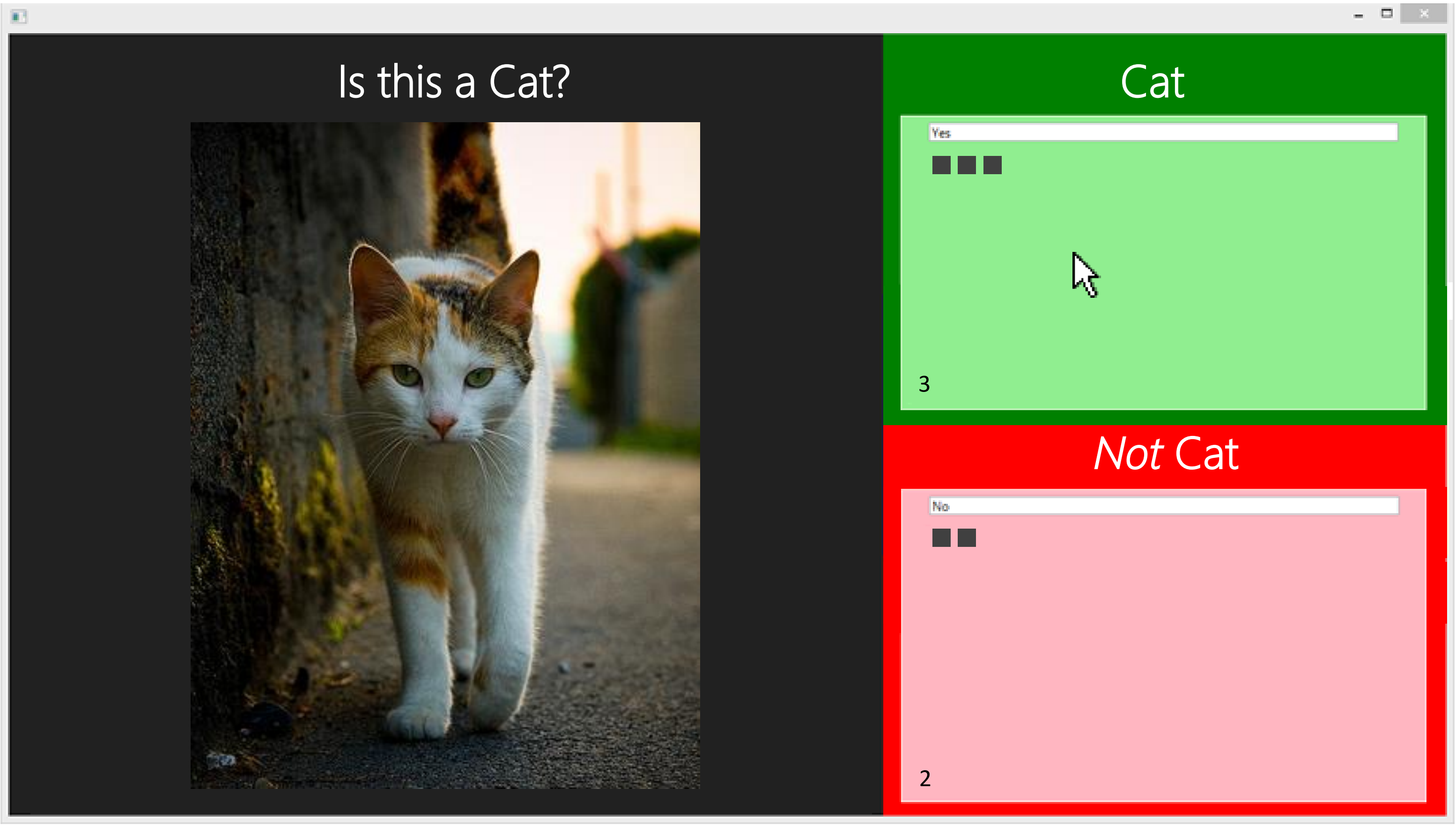


Pre-defined high-level categories.

Traditional Labeling

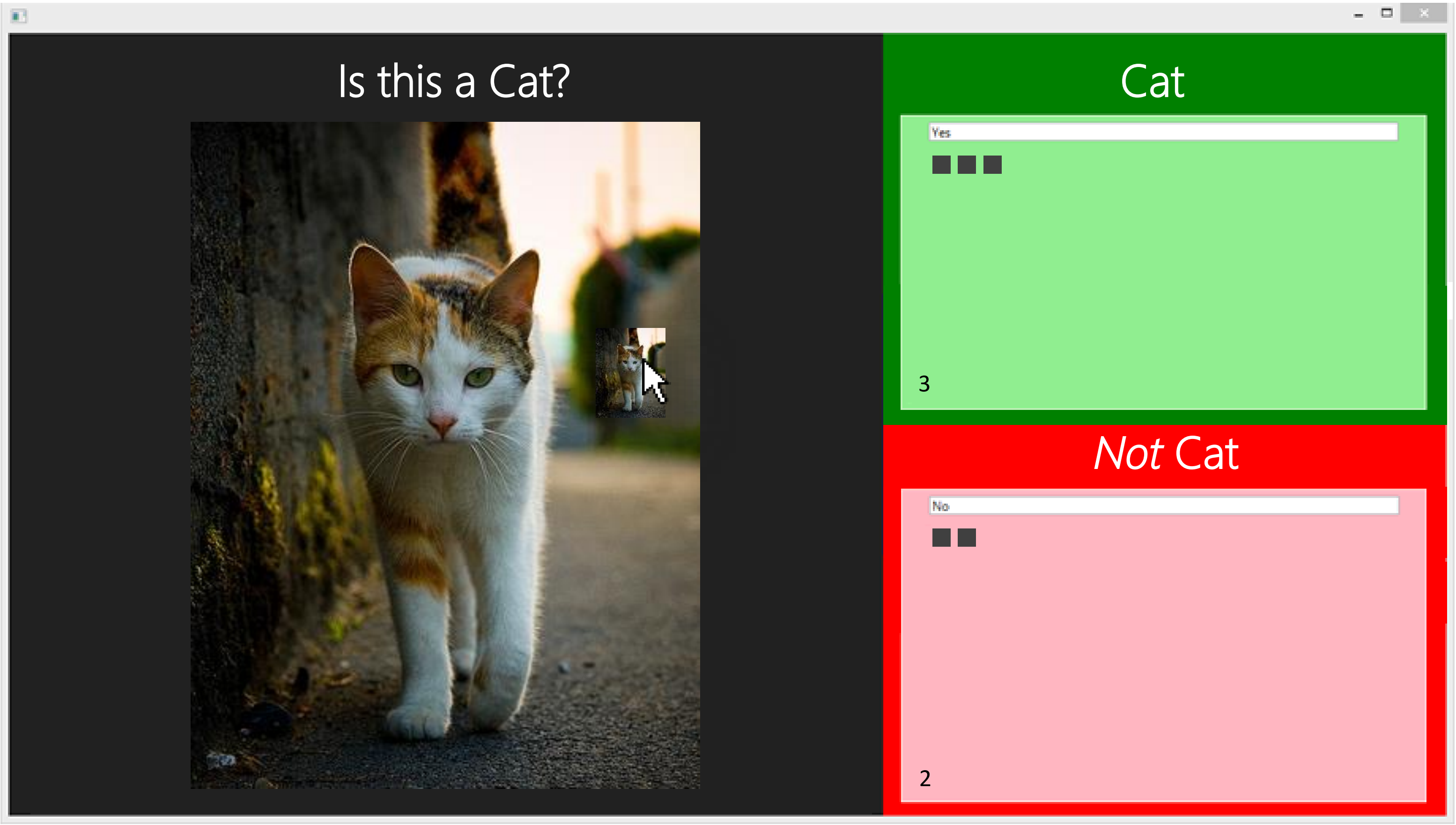
Pre-defined high-level categories.

Traditional Labeling



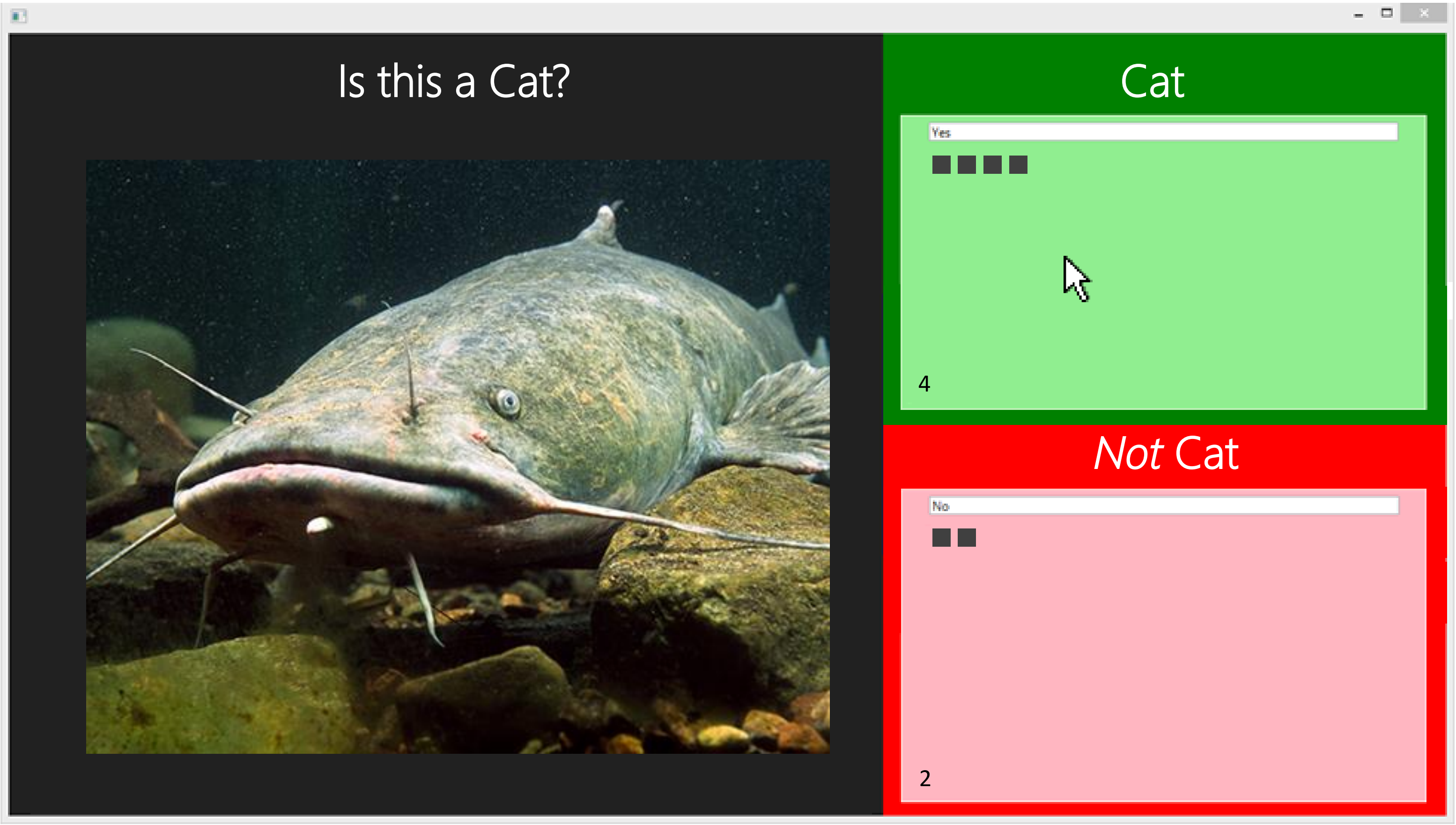
Pre-defined high-level categories.

Traditional Labeling



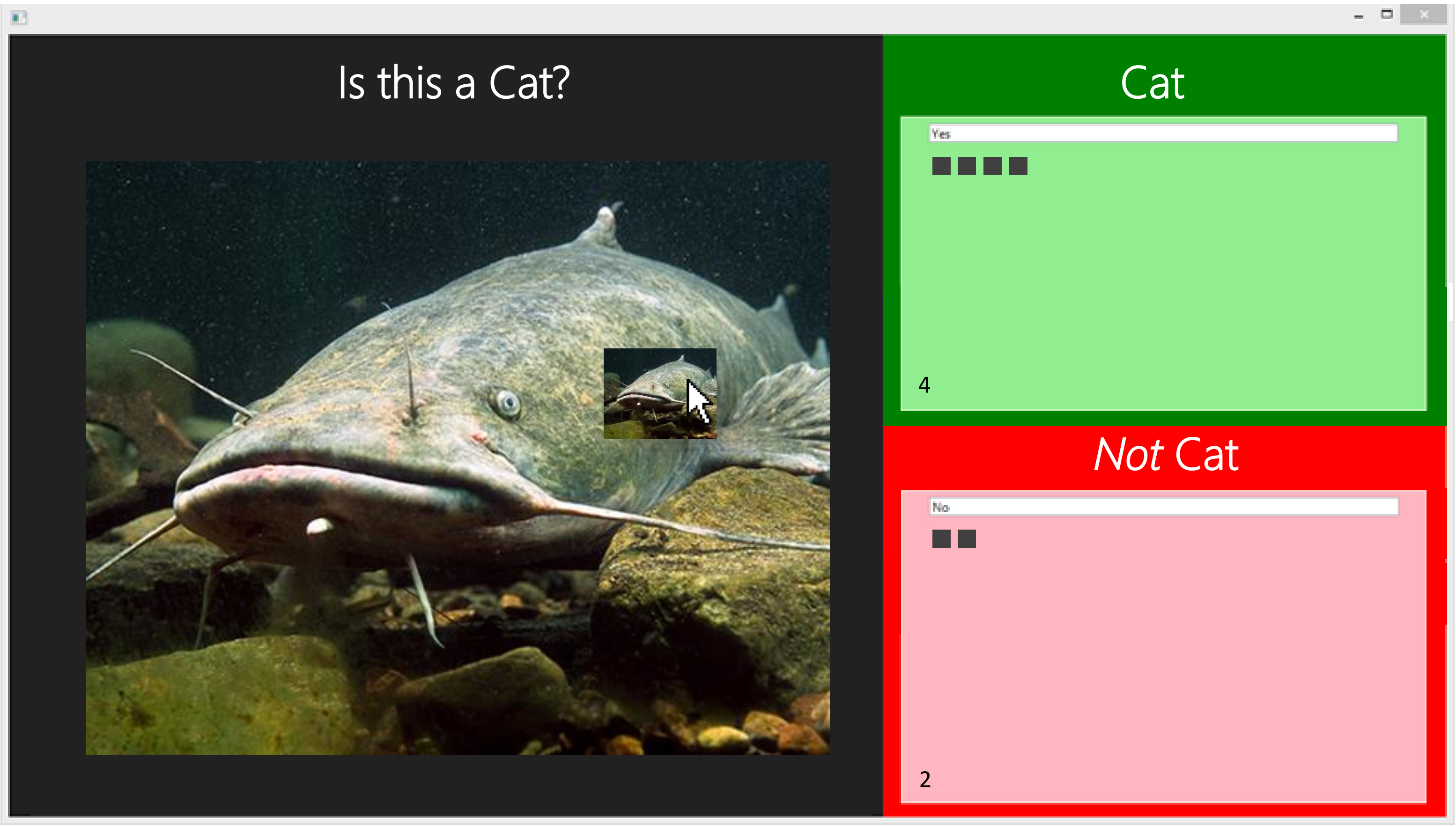
Pre-defined high-level categories.

Traditional Labeling



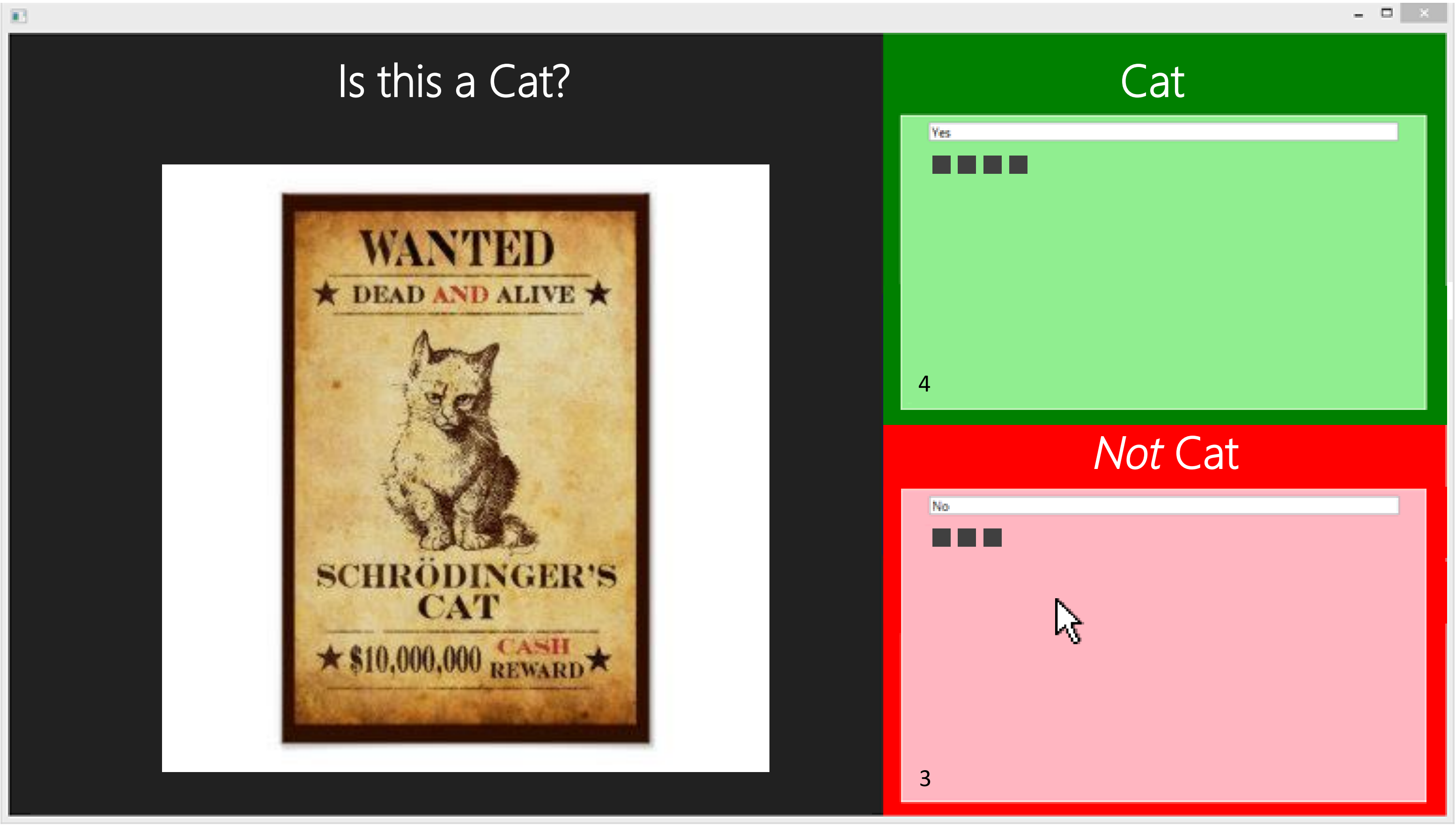
Pre-defined high-level categories.

Traditional Labeling



Pre-defined high-level categories.

Traditional Labeling

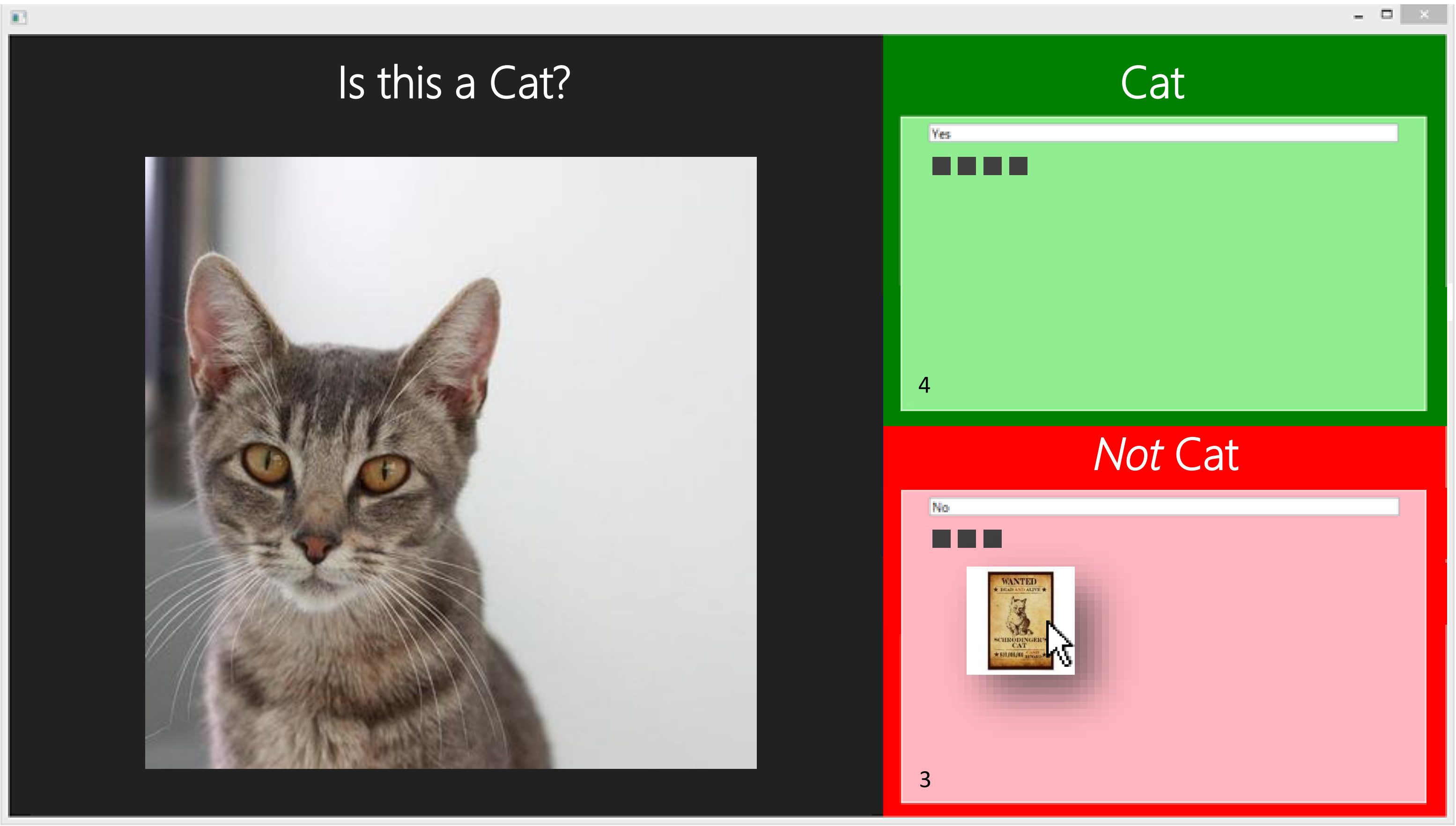


Pre-defined high-level categories.

Traditional Labeling

Pre-defined high-level categories.

Traditional Labeling

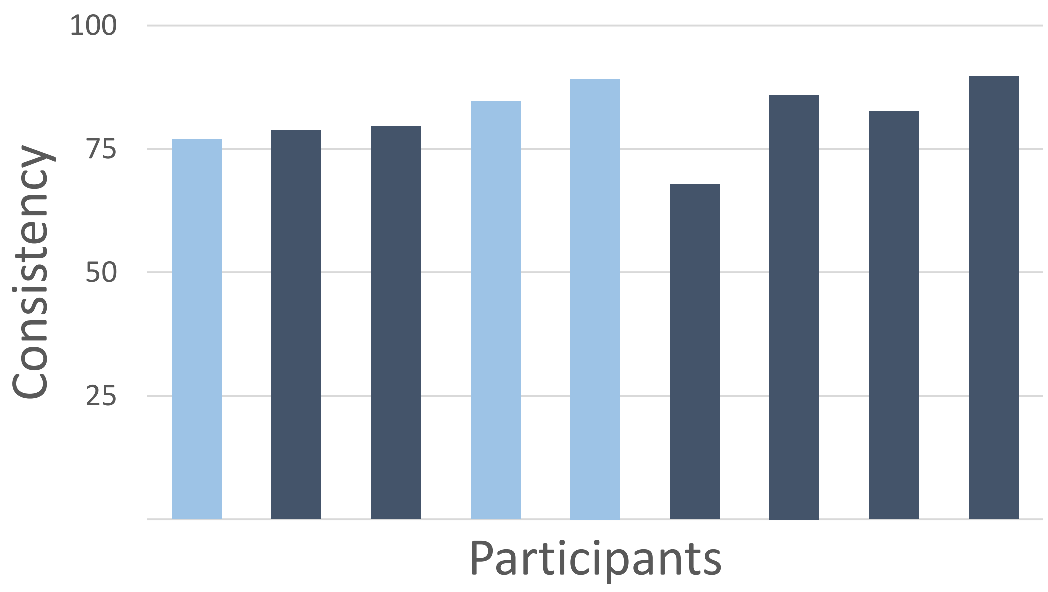


Pre-defined high-level categories.

Does not support **concept evolution** (refining the target concept as data is observed).

How common is concept evolution?

Nine machine learning experts labeled the same 200 pages in two sessions (4 weeks apart).



Average consistency 81.7% (SD=6.8%)

6 out of 9 people's labels changed significantly (via Chi Square test of symmetry)

Proposed Solution – Structured Labeling

Enable people to **explicitly organize their concept** via *grouping* and *tagging* within a traditional labeling scheme.

Traditional Labeling

Pre-defined high-level categories.

Structured Labeling

Structured Labeling

Structured Labeling

Grouping within high-level categories.

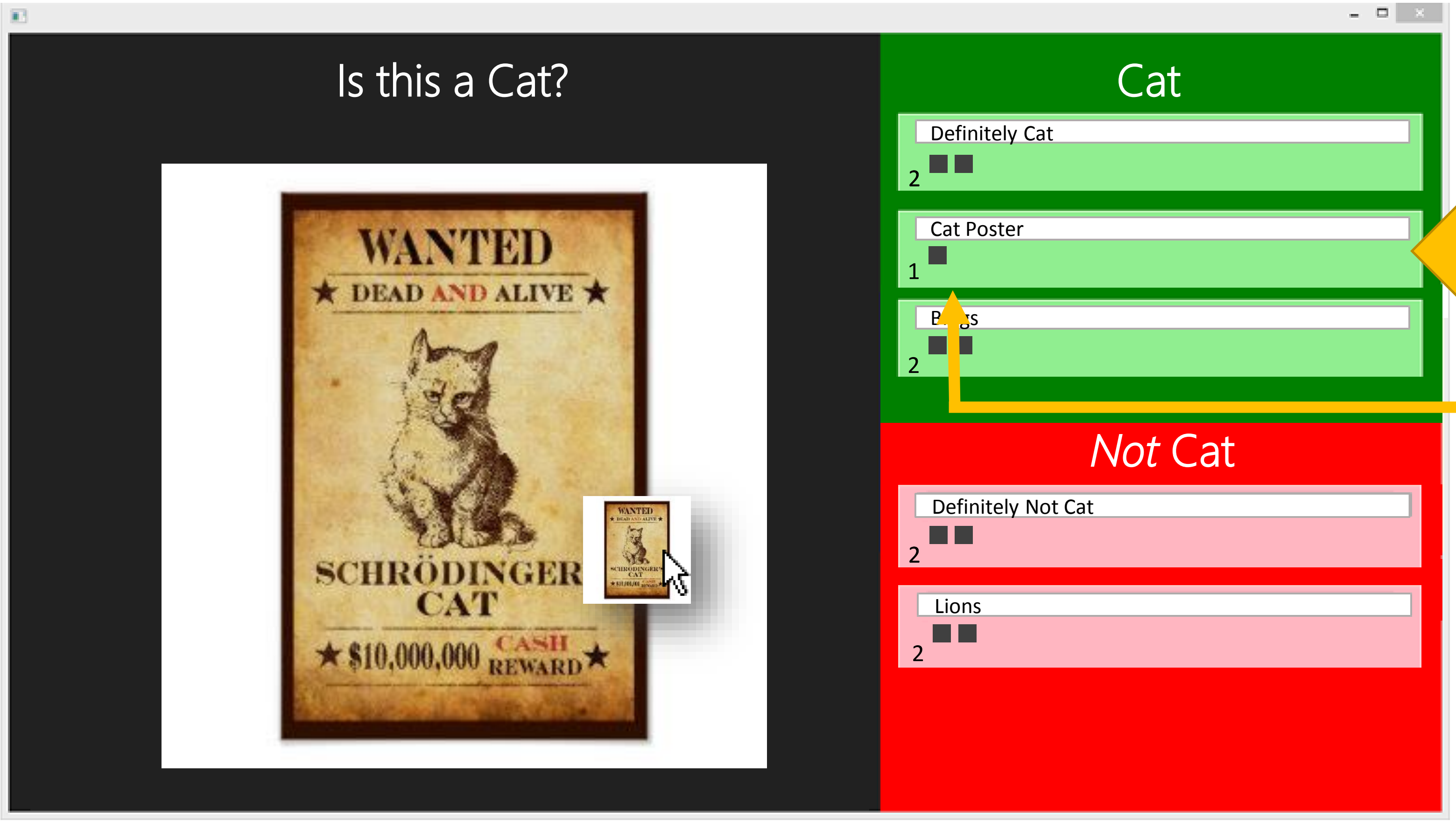
User provided tags on groups aid recall.

Structured Labeling

Grouping within high-level categories.

User provided tags on groups aid recall.

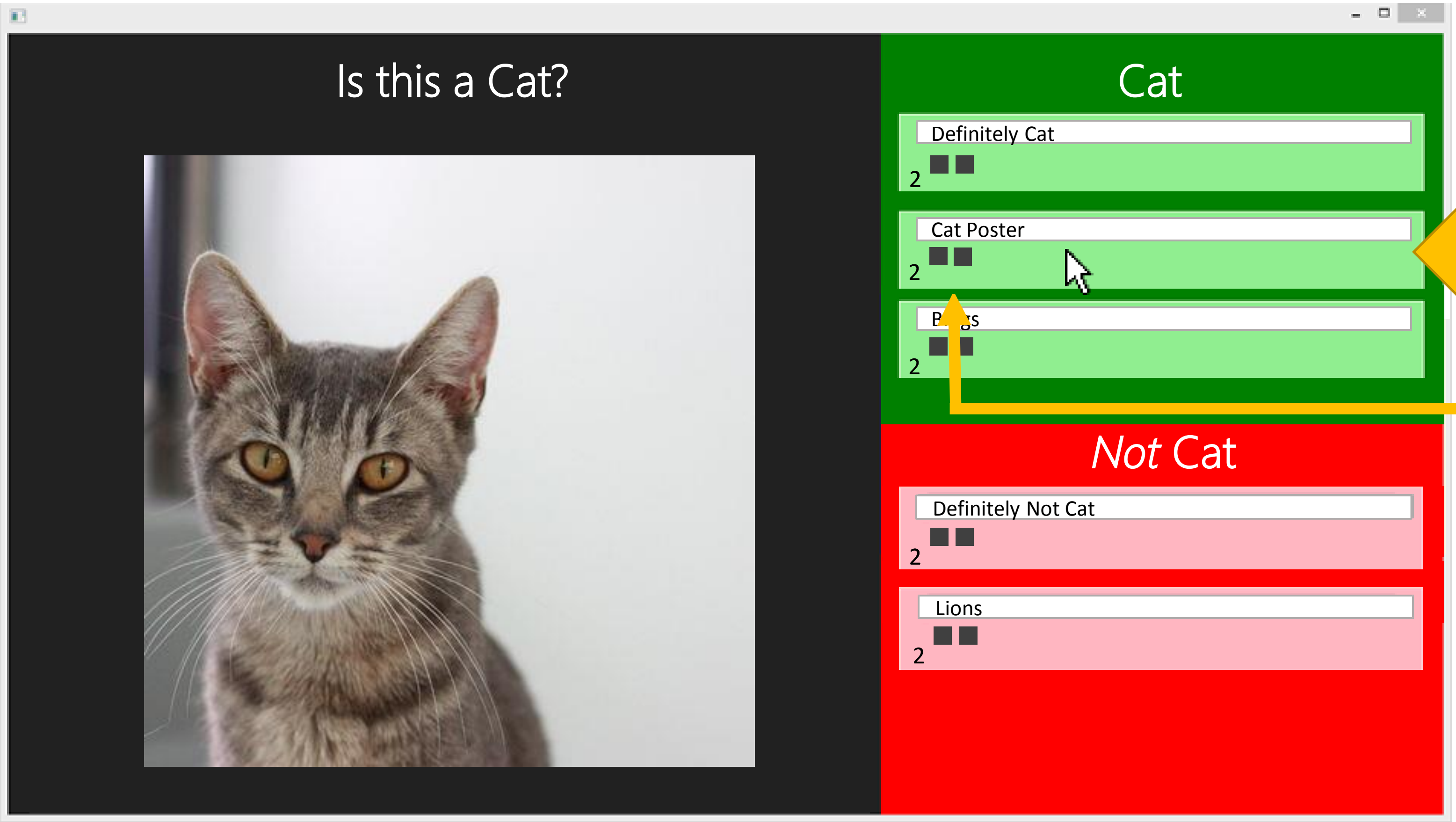
Structured Labeling



Grouping within high-level categories.

User provided tags on groups aid recall.

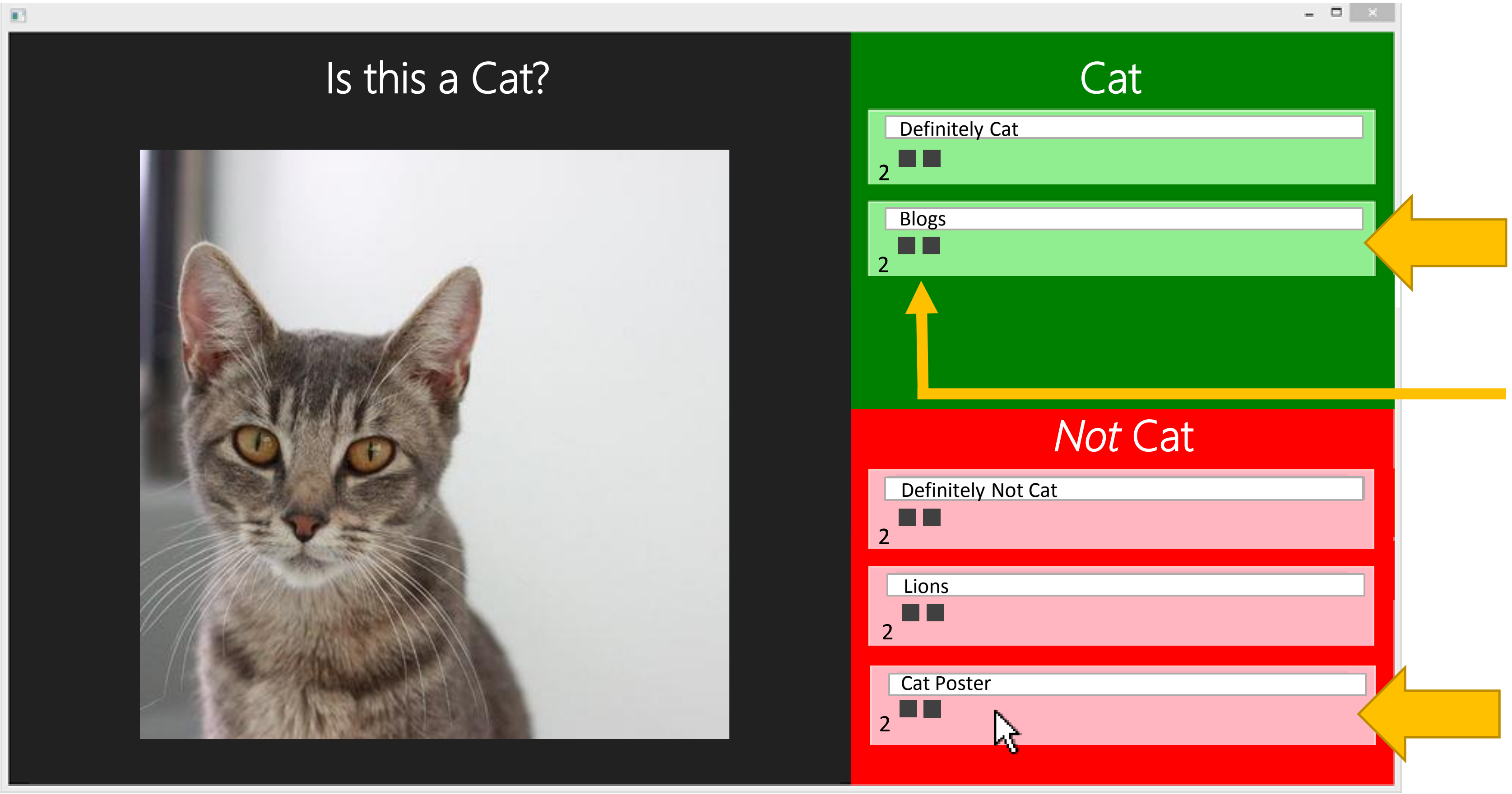
Structured Labeling



Grouping within high-level categories.

User provided tags on groups aid recall.

Structured Labeling

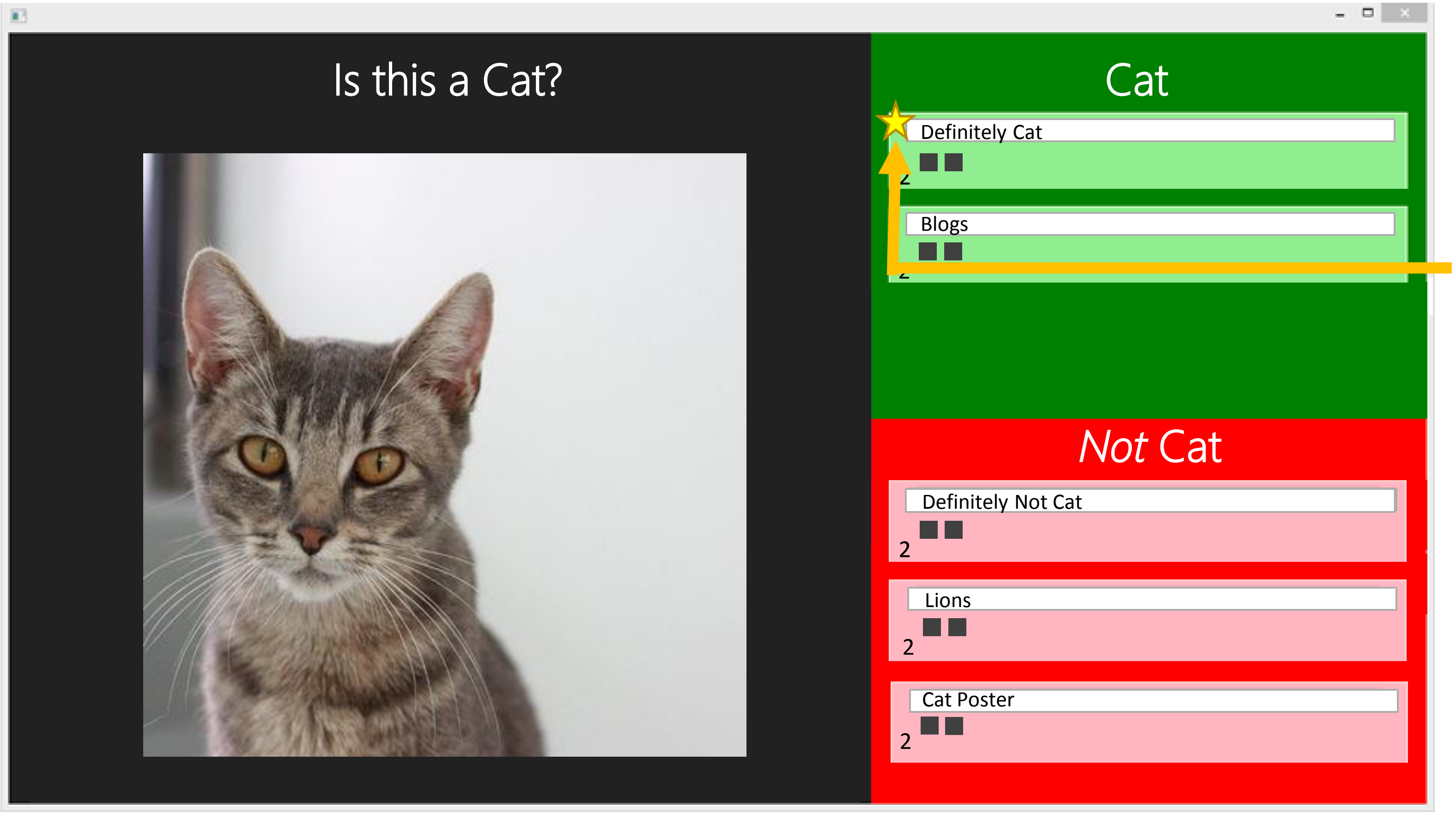


Grouping within high-level categories.

User provided tags on groups aid recall.

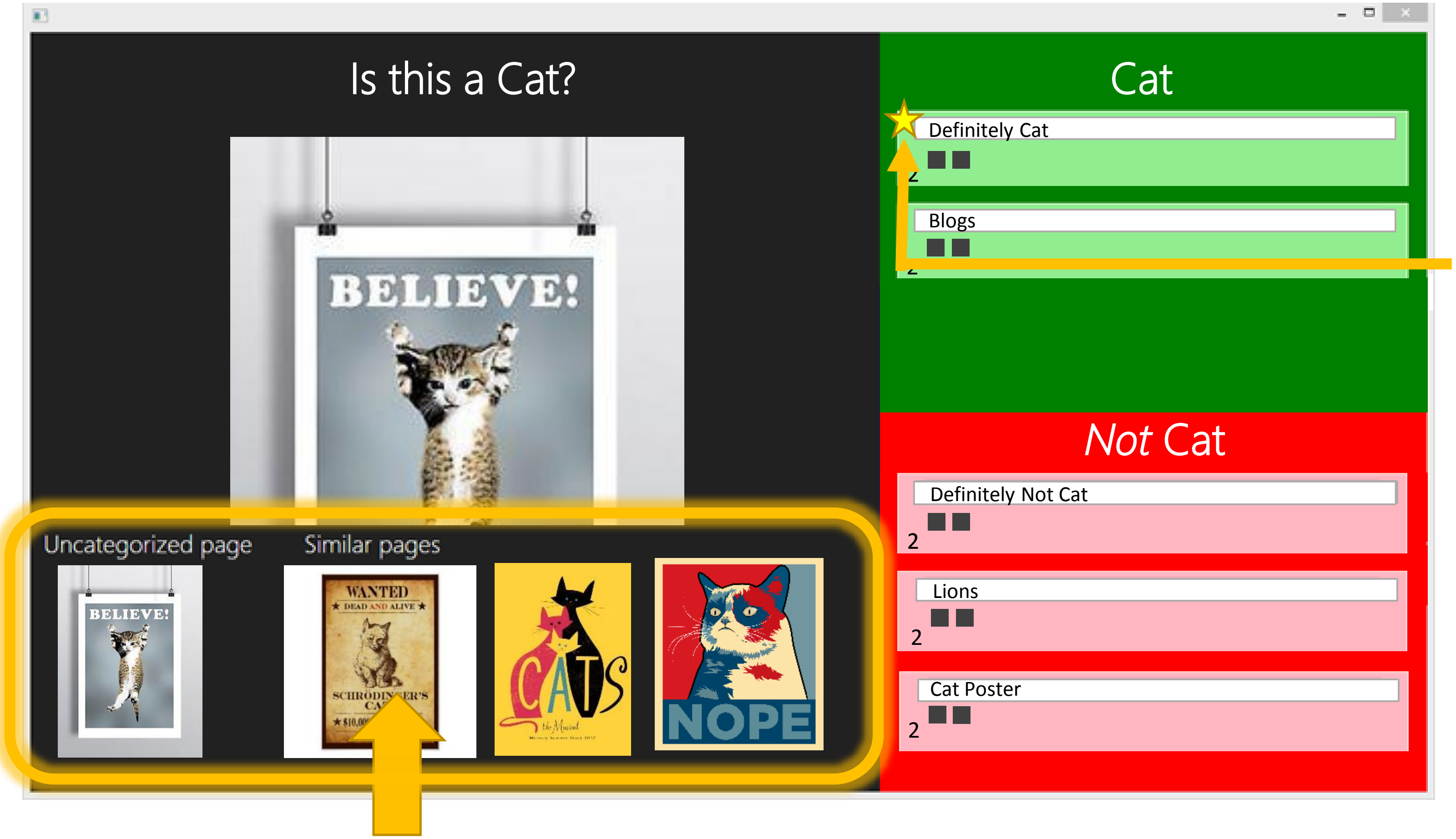
Can move, merge and split groups as desired.

Assisted Structured Labeling



Grouping recommendations to improve label consistency.

Assisted Structured Labeling



Grouping recommendations to improve label consistency.

Similar items to help users make decisions.

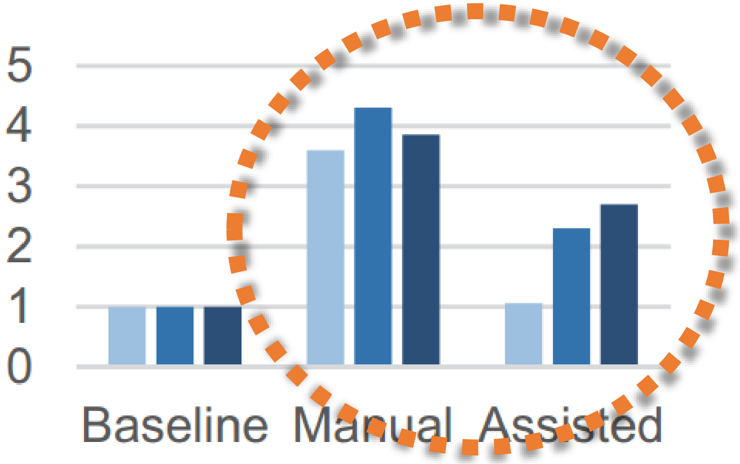
Findings

People revised labels significantly more with structured labeling

People labeled more consistently

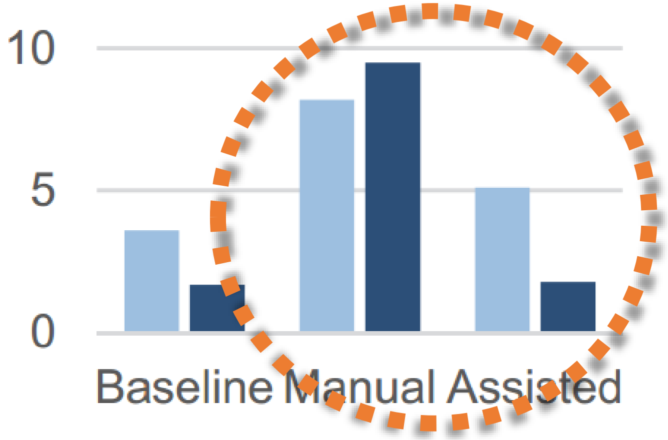
People preferred it over traditional labeling

Mean # Groups



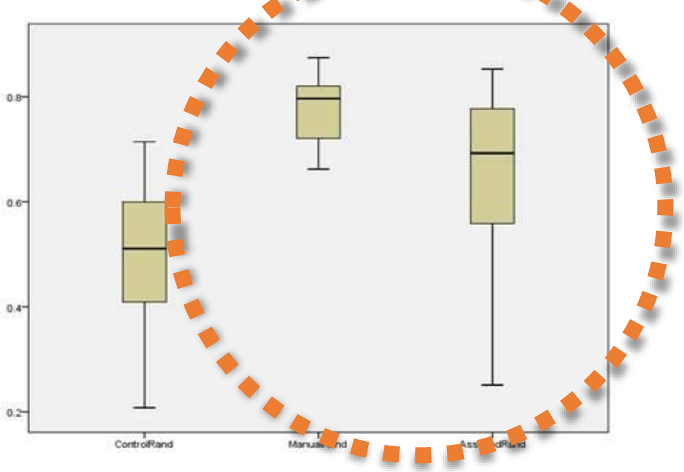
(X^2=20.19, df=2, p < .001)

Revisions



(X^2=12, df=2, p < .002)

Label Consistency



(X^2=6.53, df=2, p < .038)

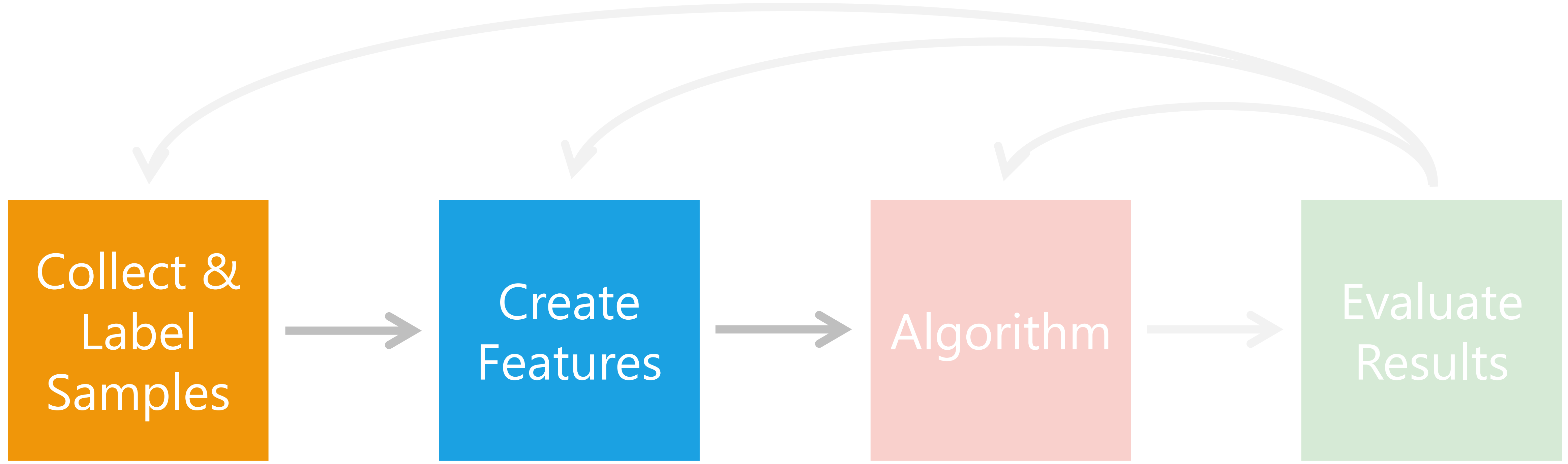
Structured Labeling Summary

Current tools do not support **concept evolution**.

Structured labeling helps people refine their concepts by surfacing labeling decisions and aiding recall.

People used structured labeling when it was available and **labeled more consistently**.

Structure contains additional information (e.g., group related features, group related accuracy, decisions made…)



Structured labeling improves consistency [CHI 2014]

Feature Insight [VAST 2015]

ModelTracker [CHI 2015, VAST 2016]

*"At the end of the day, some machine learning projects succeed and some fail. What makes the difference? **Easily the most important factor is the features used.**"* [Domingos, CACM 2012]

...yet, little guidance or best practices exist.

How do people come up with features?

Look for features used in related domains.

Use intuition or domain knowledge.

Apply automated techniques

Feature ideation – Think of and experiment with custom features (a “black art”).

Proposed Solution – Feature Insight

Support **compare and contrast** of data.

What makes a cat a cat?

What makes a cat a cat?

Proposed Solution – Feature Insight

Support **compare and contrast** of data.

Comparing pairs vs sets?

Comparing Pairs vs Sets

Sets may help people think of generalizable features.

Positive

Negative

VS

Positives

Negatives

Proposed Solution – Feature Insight

Support **compare and contrast** of data.

Comparing pairs vs sets?

Raw data vs visual summaries?

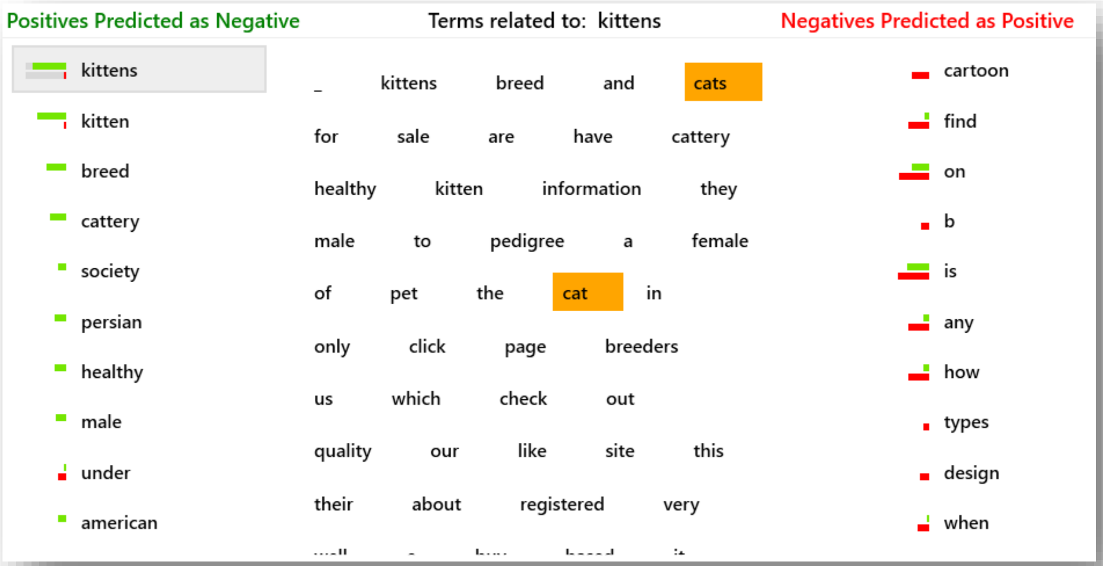
Looking at Raw Data vs. Visual Summaries

Visual summaries may reveal relevant characteristics and hide irrelevant noise.

Raw Data

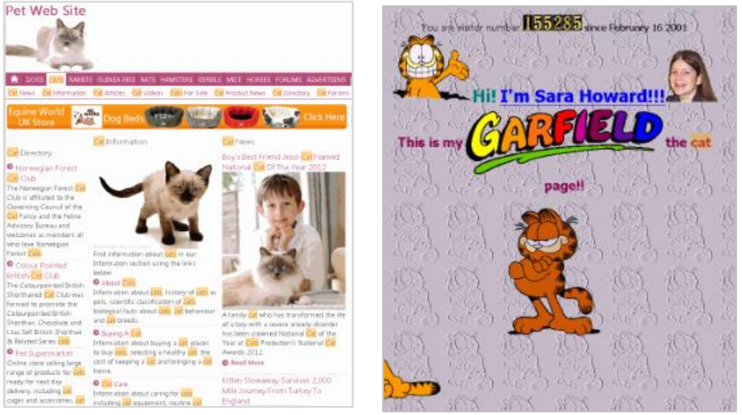
VS

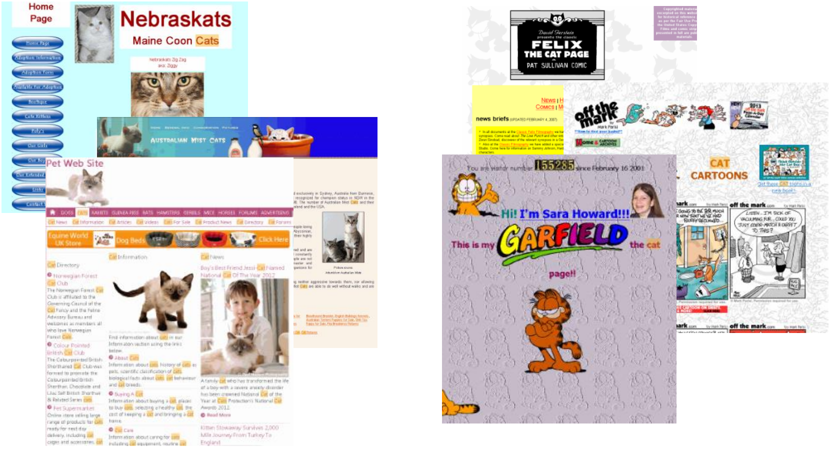
Visual Summary



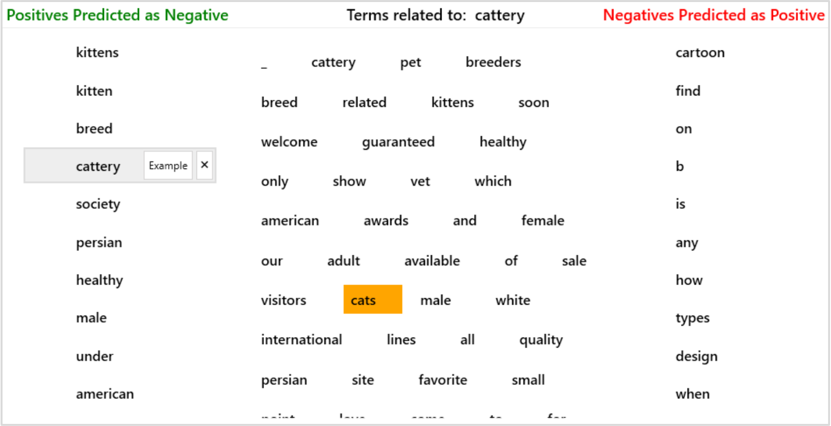
	Individual Comparison	Set Comparison

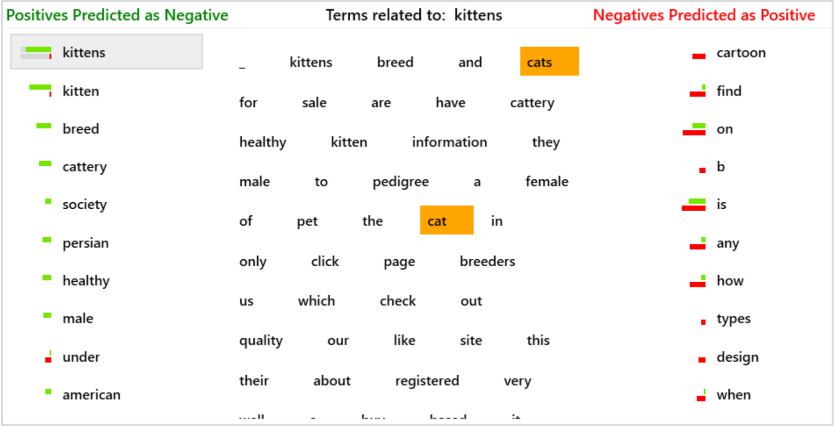
Raw Data



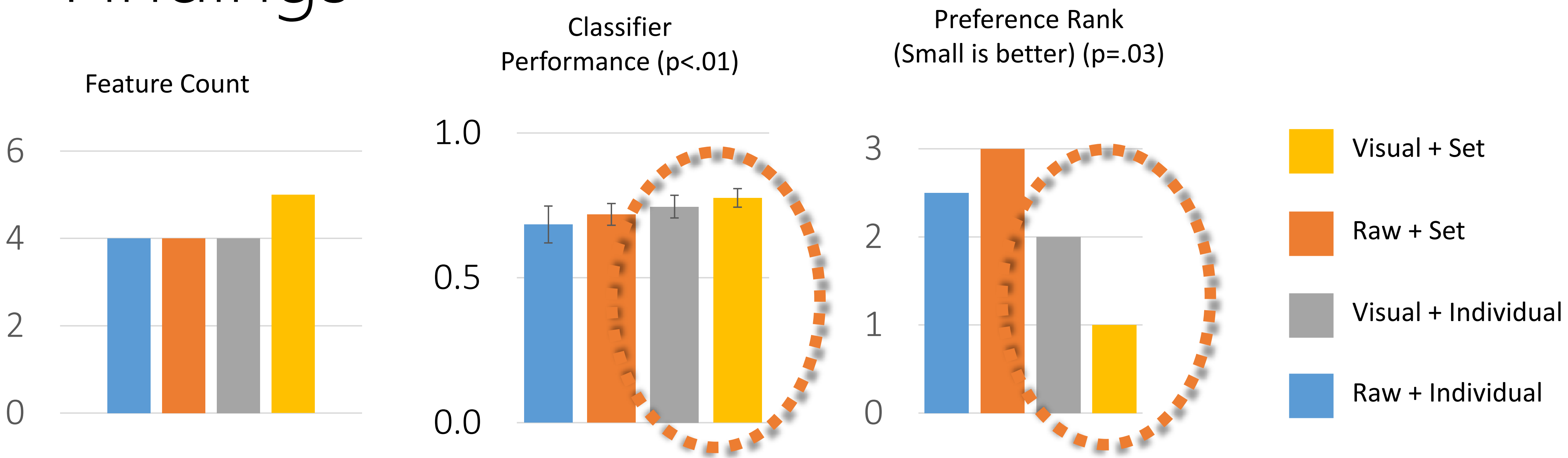


Visual Summaries





Findings



Visual summaries led to *better* features

Visual summaries preferred over looking at raw data

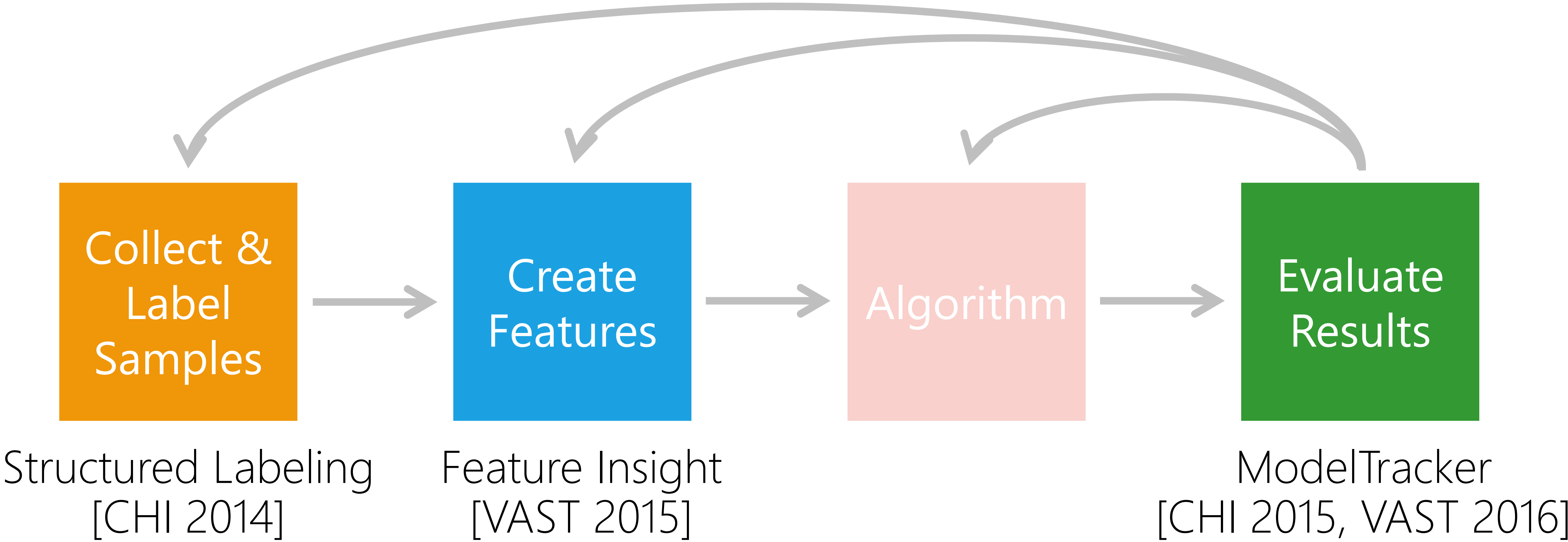
Sets useful only in combination with visuals

Feature Insight Summary

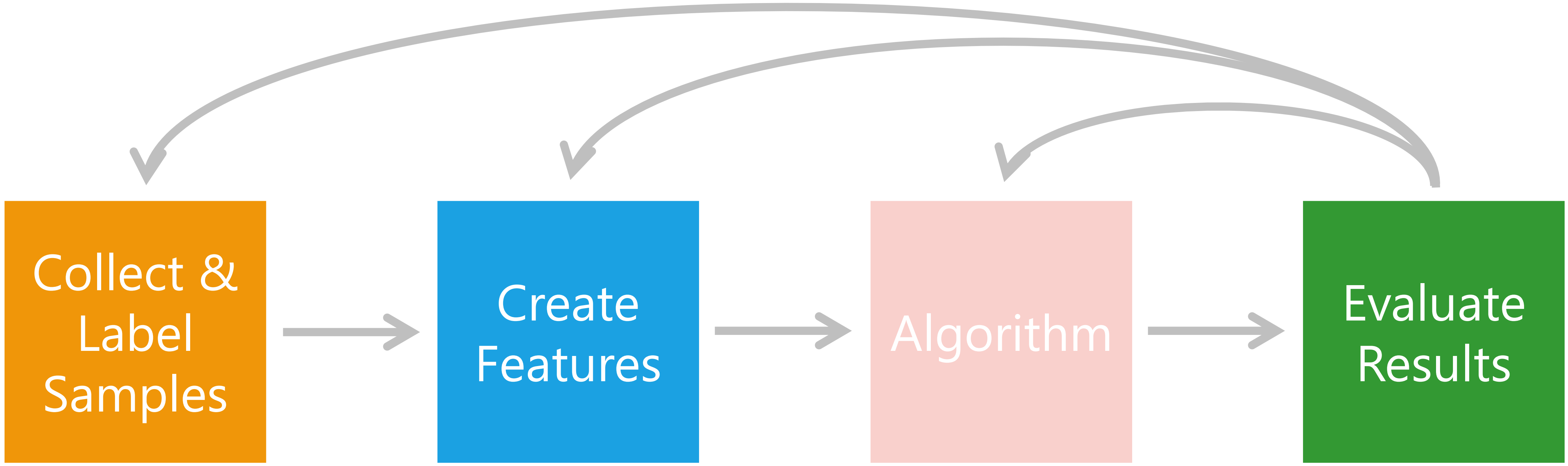
Featuring is arguably the most important step in machine learning, but there is little guidance on **feature ideation**.

Feature Insight supports error comparison, examination of sets, and visual summaries.

Visual summaries help people create **better quality features**.


Collect & Label Samples
Create Features
Algorithm
Evaluate Results
Structured Labeling [CHI 2014]
Feature Insight [VAST 2015]
ModelTracker [CHI 2015, VAST 2016]

How do people evaluate performance?

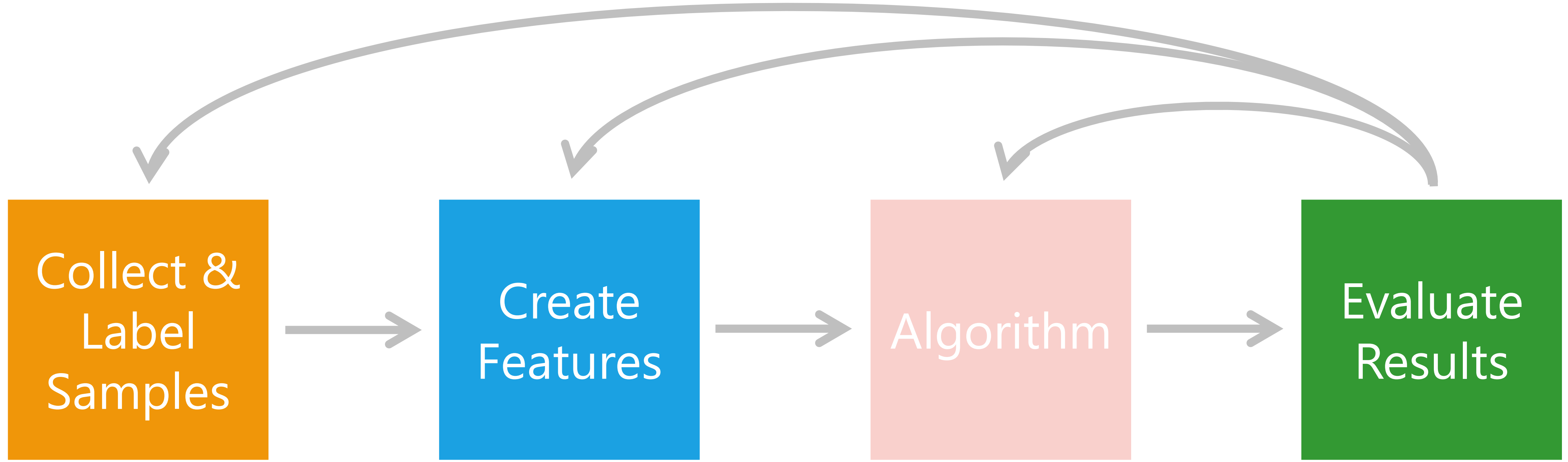


Structured Labeling
[CHI 2014]

Feature Insight
[VAST 2015]

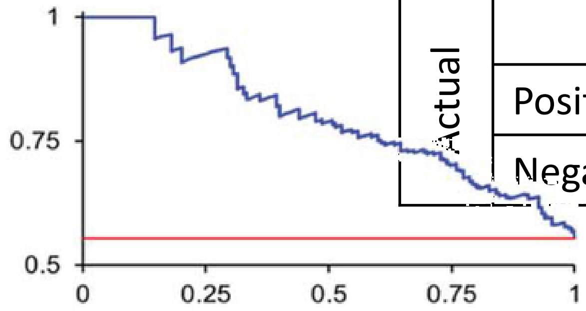
ModelTracker
[CHI 2015, VAST 2016]

How do people evaluate performance?



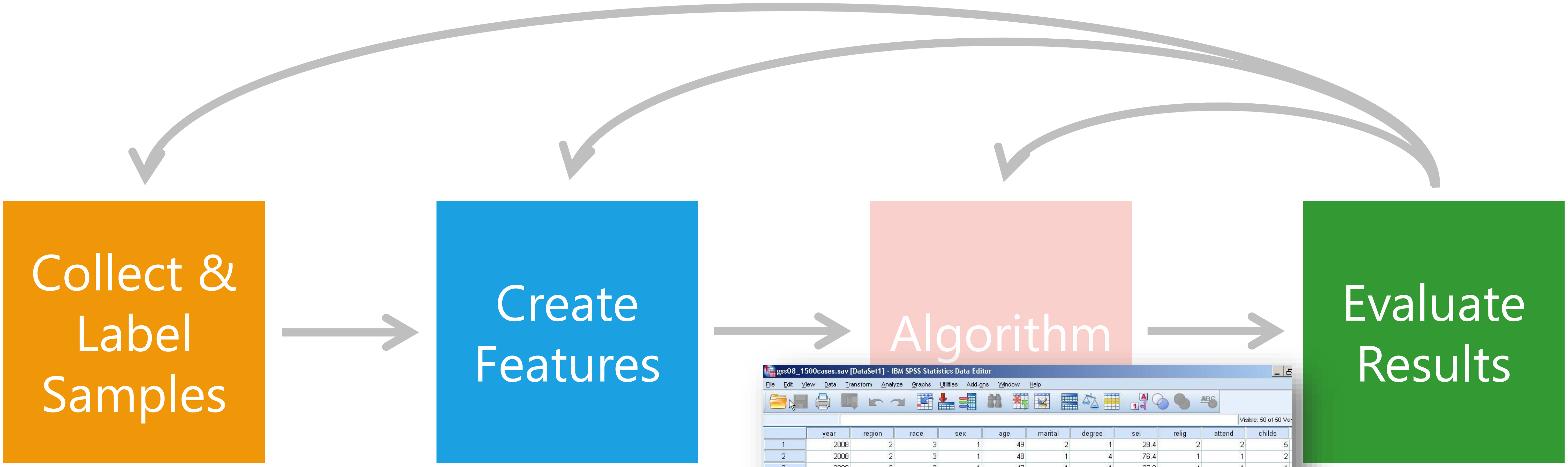
Summary statistics **hide important information** about model behavior.

0.67 0.70 0.71 ???

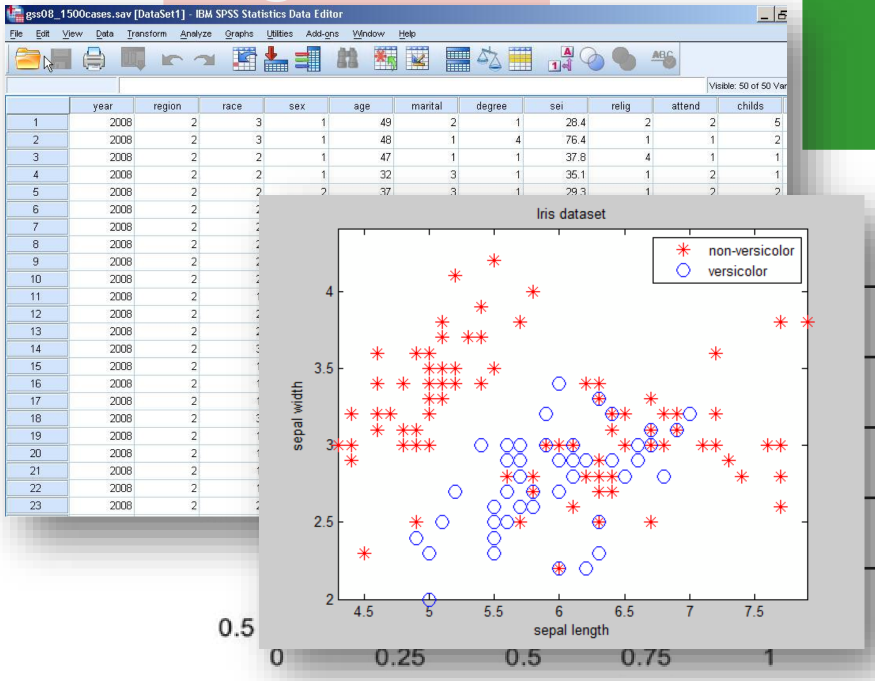


		Predicted	
		Positive	Negative
Actual	Positive	143	72
	Negative	35	190

How do people evaluate performance?



Summary statistics **hide important information** about model behavior.

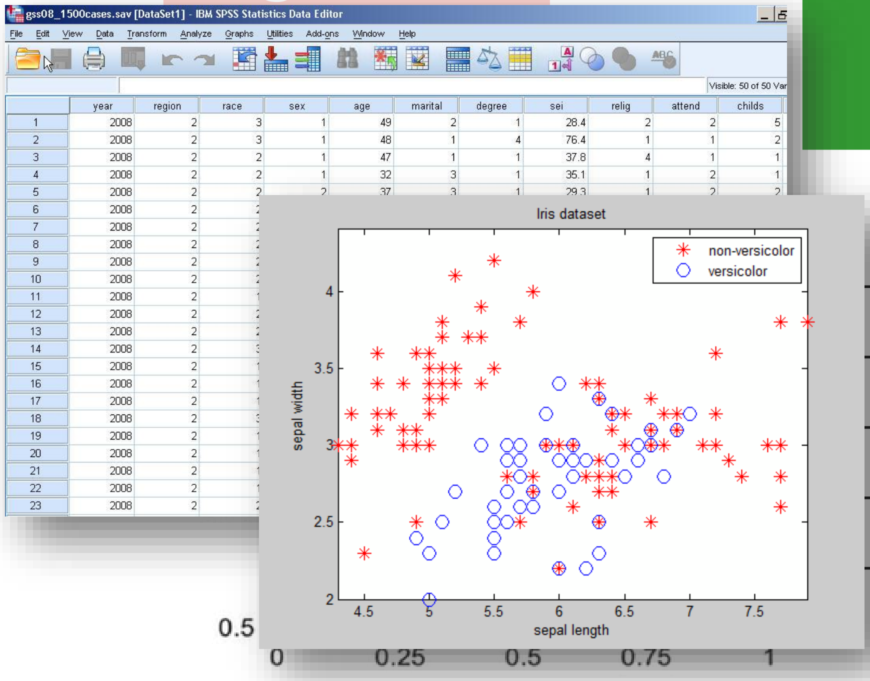


0.70 0.71 ???

Predicted	
Positive	Negative
143	72
35	190

How do people evaluate performance?

Collect & Label Samples → **Create Features** → **Algorithm** → **Evaluate Results**



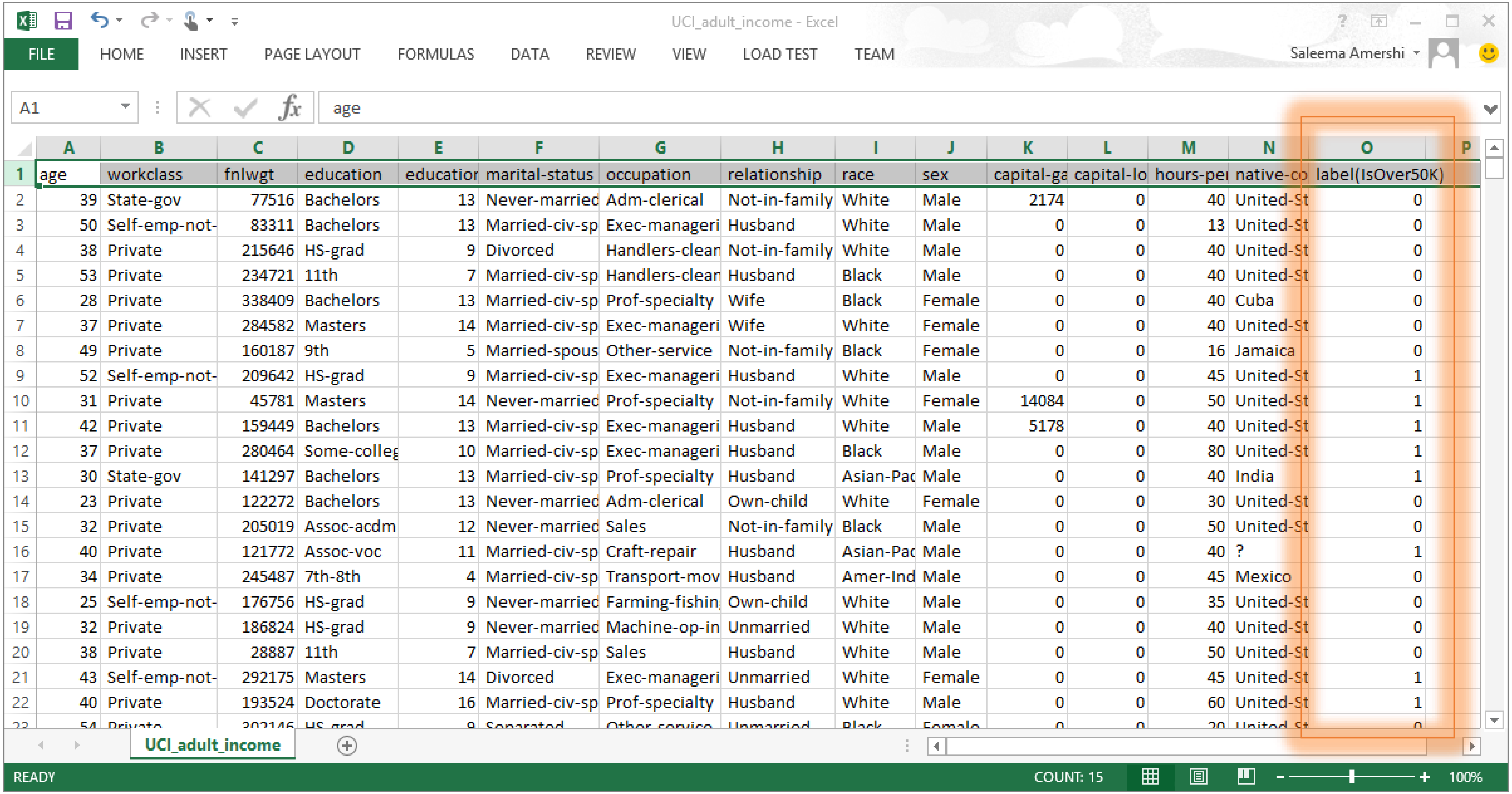
0.70 0.71 ???

Predicted	
Positive	Negative
143	72
35	190

Summary statistics **hide important information** about model behavior.

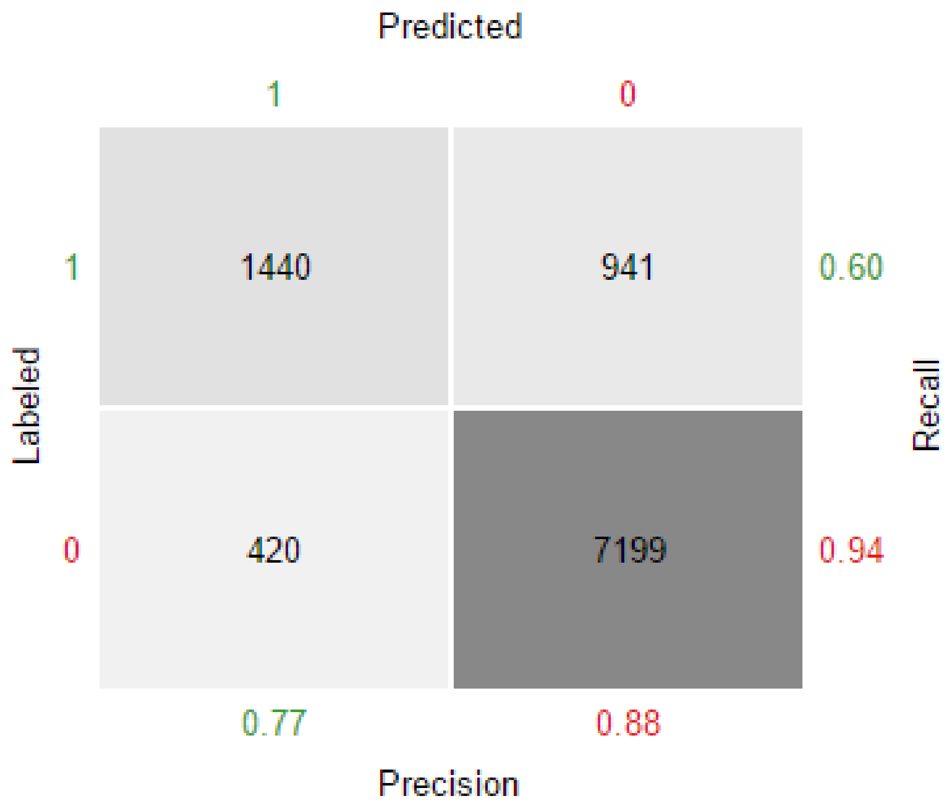
Switching tools to examine data is **disruptive** and leads to a trial-and-error approach [Patel *et al.*, AAAI 2008].

Example: Predicting Income Levels

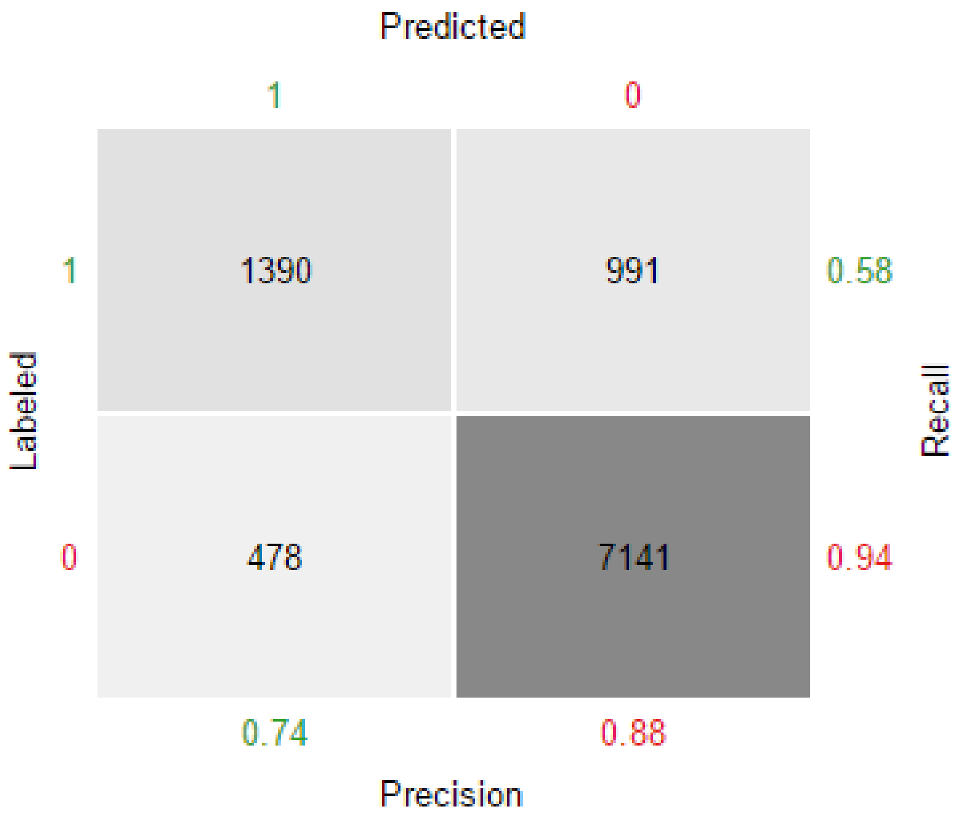


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	age	workclass	fnlwgt	education	educatior	marital-status	occupation	relationship	race	sex	capital-ga	capital-lo	hours-pei	native-co	label(IsOver50K)
2	39	State-gov	77516	Bachelors	13	Never-marriec	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-St	0
3	50	Self-emp-not-	83311	Bachelors	13	Married-civ-sp	Exec-manageri	Husband	White	Male	0	0	13	United-St	0
4	38	Private	215646	HS-grad	9	Divorced	Handlers-clear	Not-in-family	White	Male	0	0	40	United-St	0
5	53	Private	234721	11th	7	Married-civ-sp	Handlers-clear	Husband	Black	Male	0	0	40	United-St	0
6	28	Private	338409	Bachelors	13	Married-civ-sp	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	0
7	37	Private	284582	Masters	14	Married-civ-sp	Exec-manageri	Wife	White	Female	0	0	40	United-St	0
8	49	Private	160187	9th	5	Married-spous	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	0
9	52	Self-emp-not-	209642	HS-grad	9	Married-civ-sp	Exec-manageri	Husband	White	Male	0	0	45	United-St	1
10	31	Private	45781	Masters	14	Never-marriec	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-St	1
11	42	Private	159449	Bachelors	13	Married-civ-sp	Exec-manageri	Husband	White	Male	5178	0	40	United-St	1
12	37	Private	280464	Some-colleg	10	Married-civ-sp	Exec-manageri	Husband	Black	Male	0	0	80	United-St	1
13	30	State-gov	141297	Bachelors	13	Married-civ-sp	Prof-specialty	Husband	Asian-Pac	Male	0	0	40	India	1
14	23	Private	122272	Bachelors	13	Never-marriec	Adm-clerical	Own-child	White	Female	0	0	30	United-St	0
15	32	Private	205019	Assoc-acdm	12	Never-marriec	Sales	Not-in-family	Black	Male	0	0	50	United-St	0
16	40	Private	121772	Assoc-voc	11	Married-civ-sp	Craft-repair	Husband	Asian-Pac	Male	0	0	40	?	1
17	34	Private	245487	7th-8th	4	Married-civ-sp	Transport-mov	Husband	Amer-Ind	Male	0	0	45	Mexico	0
18	25	Self-emp-not-	176756	HS-grad	9	Never-marriec	Farming-fishin	Own-child	White	Male	0	0	35	United-St	0
19	32	Private	186824	HS-grad	9	Never-marriec	Machine-op-in	Unmarried	White	Male	0	0	40	United-St	0
20	38	Private	28887	11th	7	Married-civ-sp	Sales	Husband	White	Male	0	0	50	United-St	0
21	43	Self-emp-not-	292175	Masters	14	Divorced	Exec-manageri	Unmarried	White	Female	0	0	45	United-St	1
22	40	Private	193524	Doctorate	16	Married-civ-sp	Prof-specialty	Husband	White	Male	0	0	60	United-St	1
23	54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-St	0

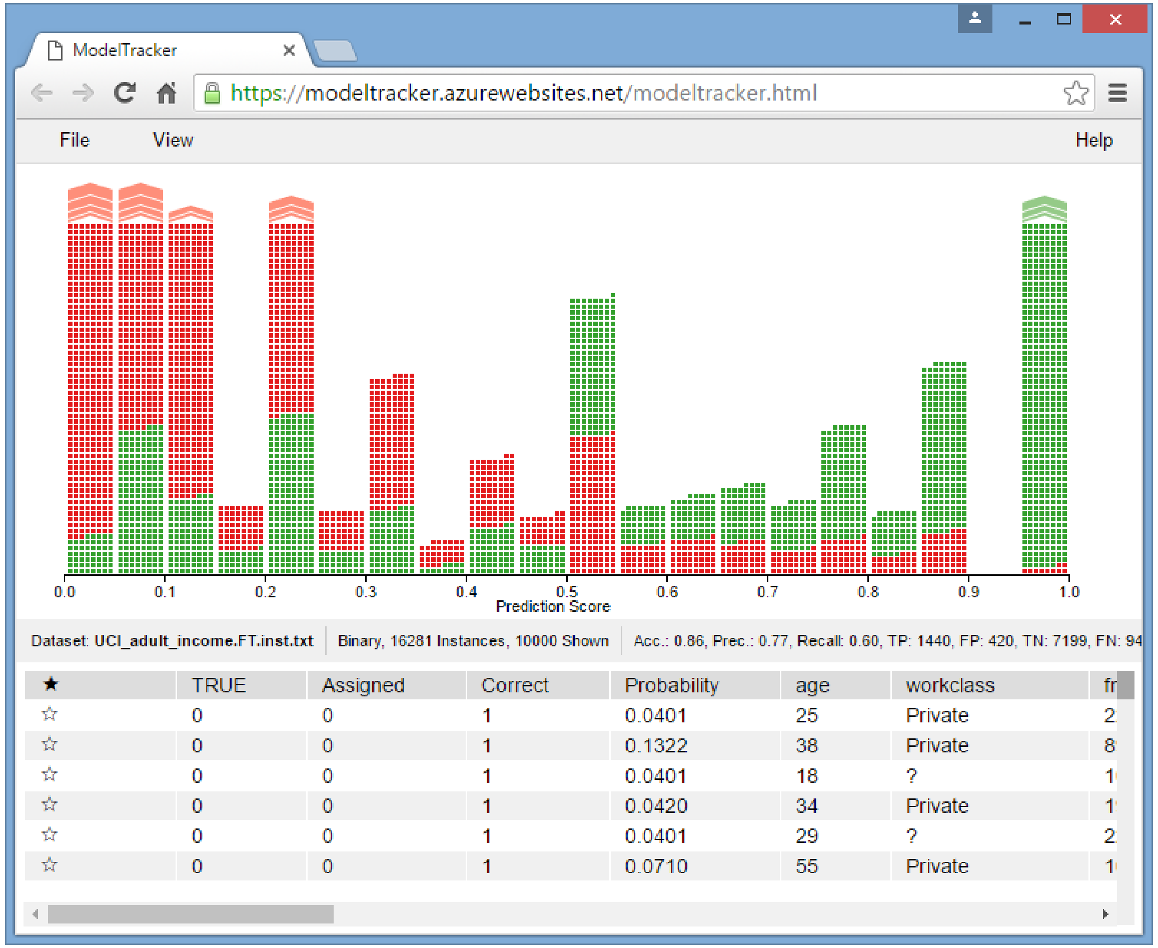
Decision Tree
86% Accuracy



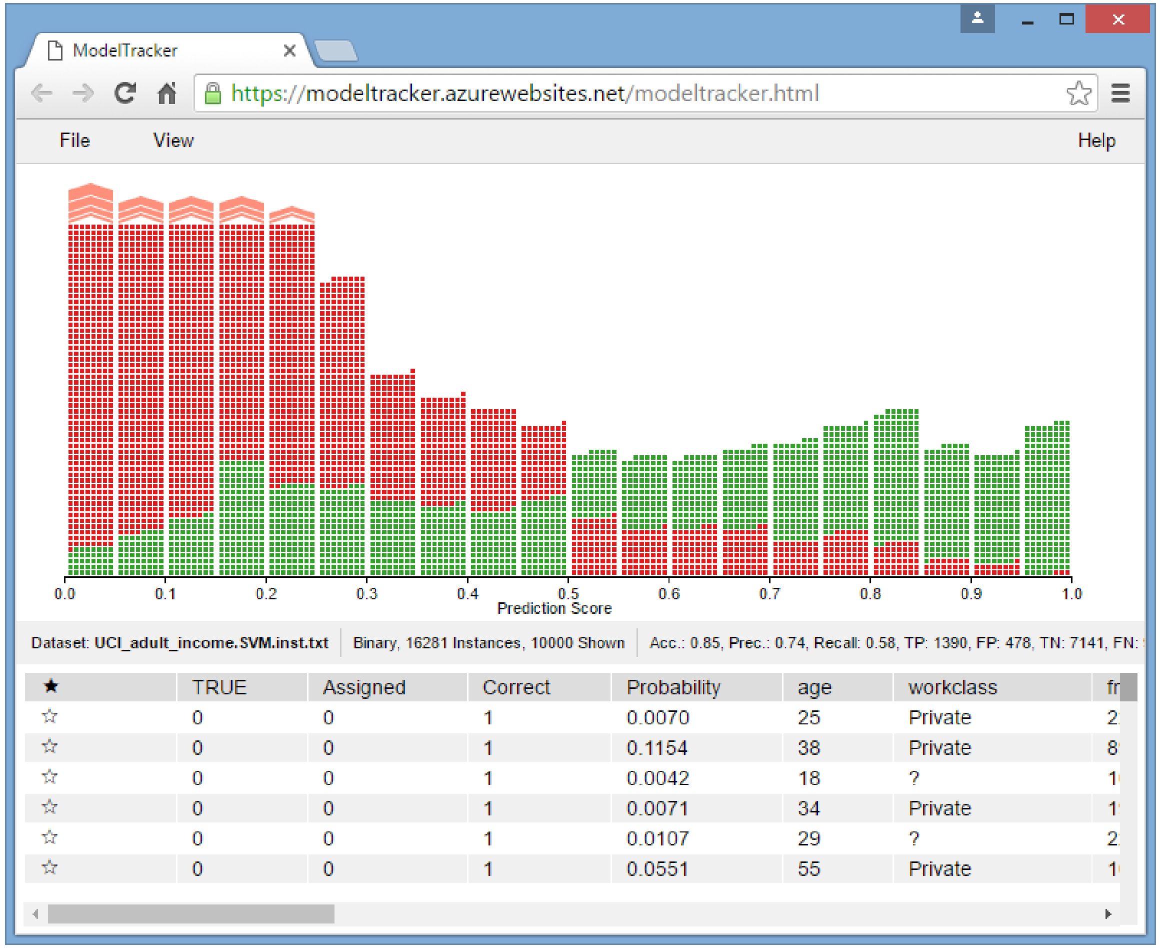
Support Vector Machine
85% Accuracy



Decision Tree
86% Accuracy

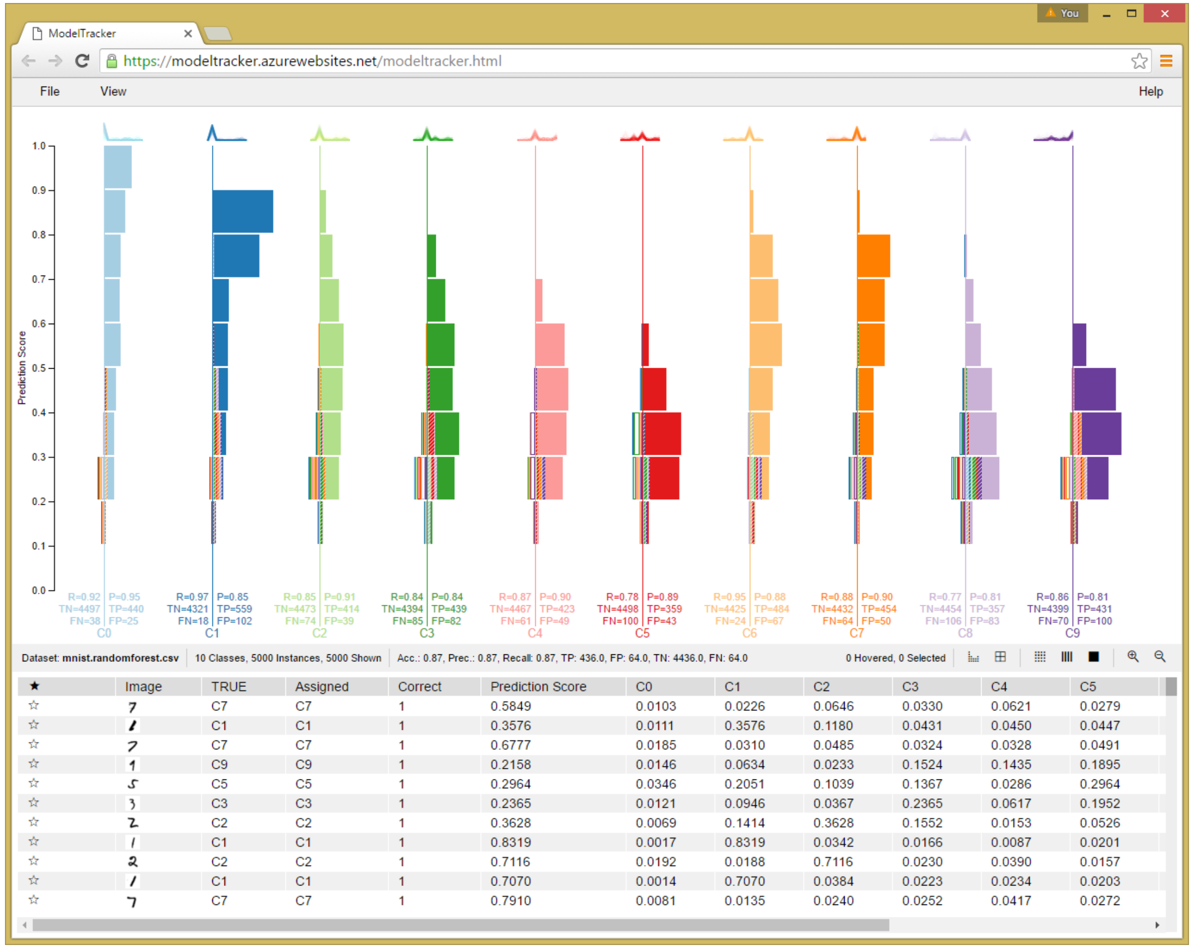


Support Vector Machine
85% Accuracy

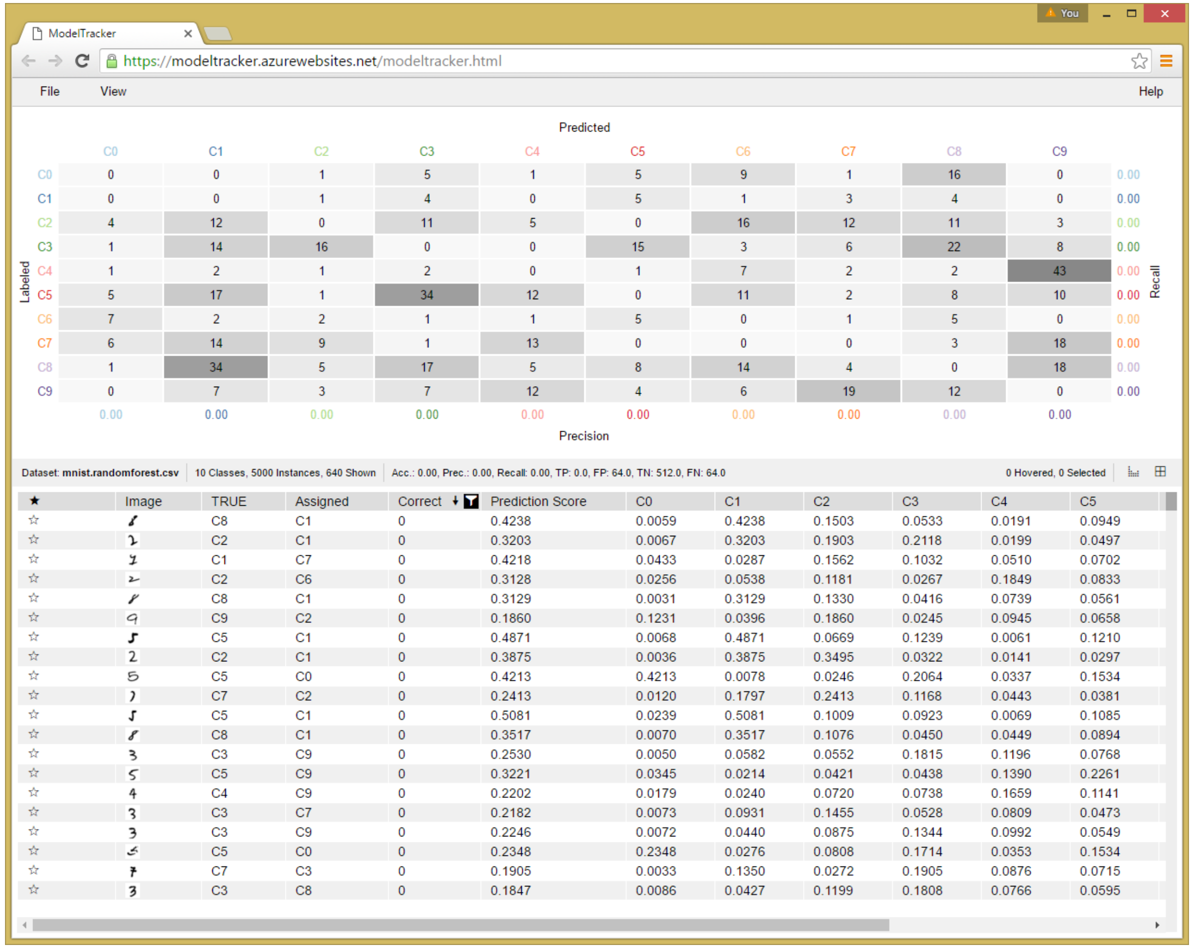


ModelTracker Demo

Significantly faster and more accurate performance analysis



ModelTracker



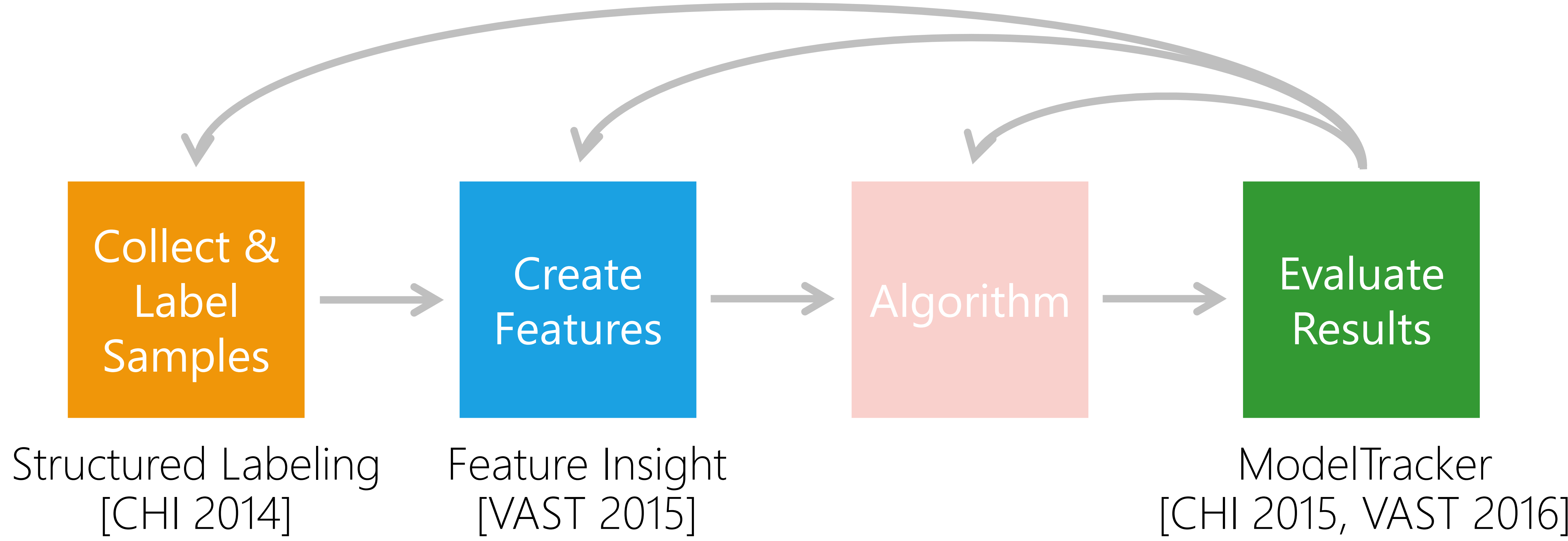
Common Confusion Matrix

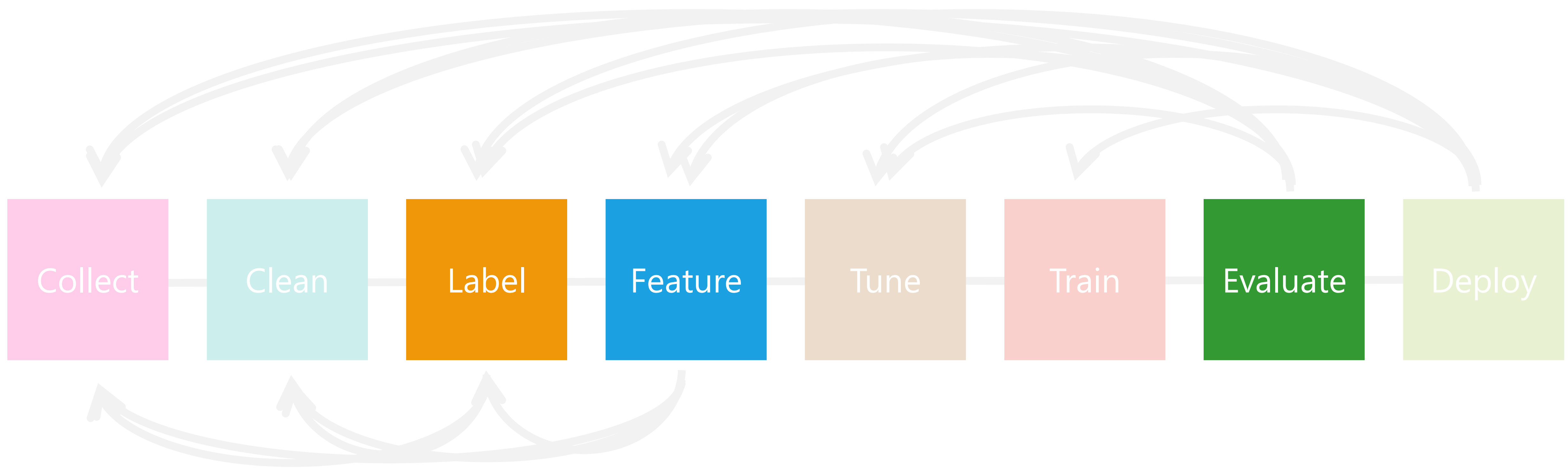
ModelTracker Summary

Current tools for performance analysis and debugging hide a lot of important information about model behavior.

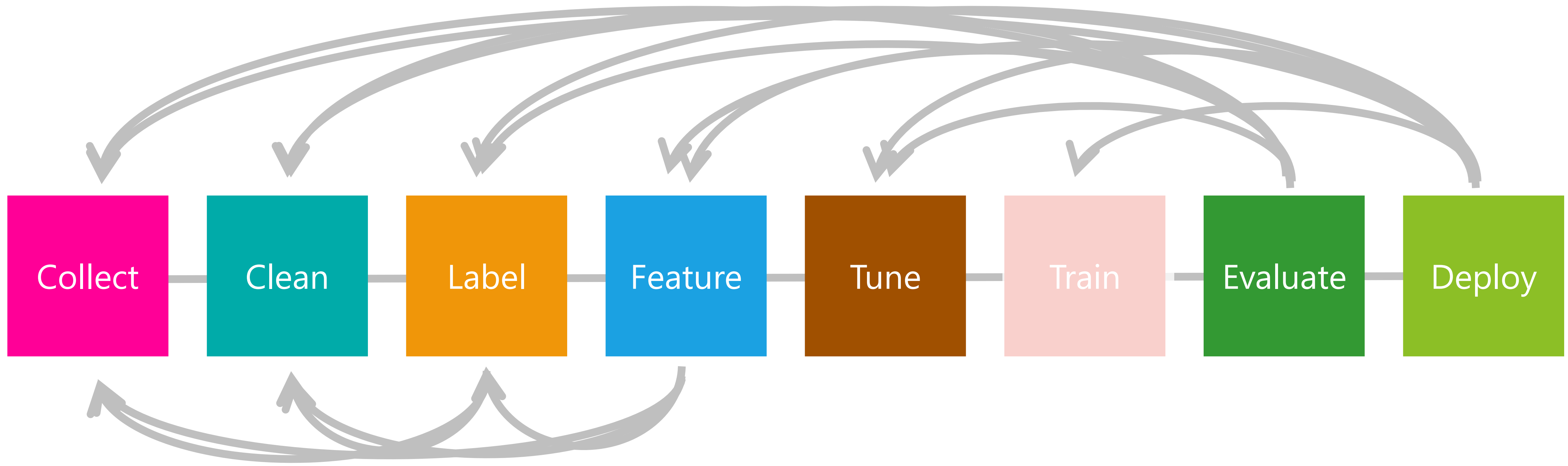
ModelTracker **supports estimating performance at multiple levels of granularity** while enabling **direct access to data**.

People are significantly **faster and more accurate** at performance analysis with ModelTracker.

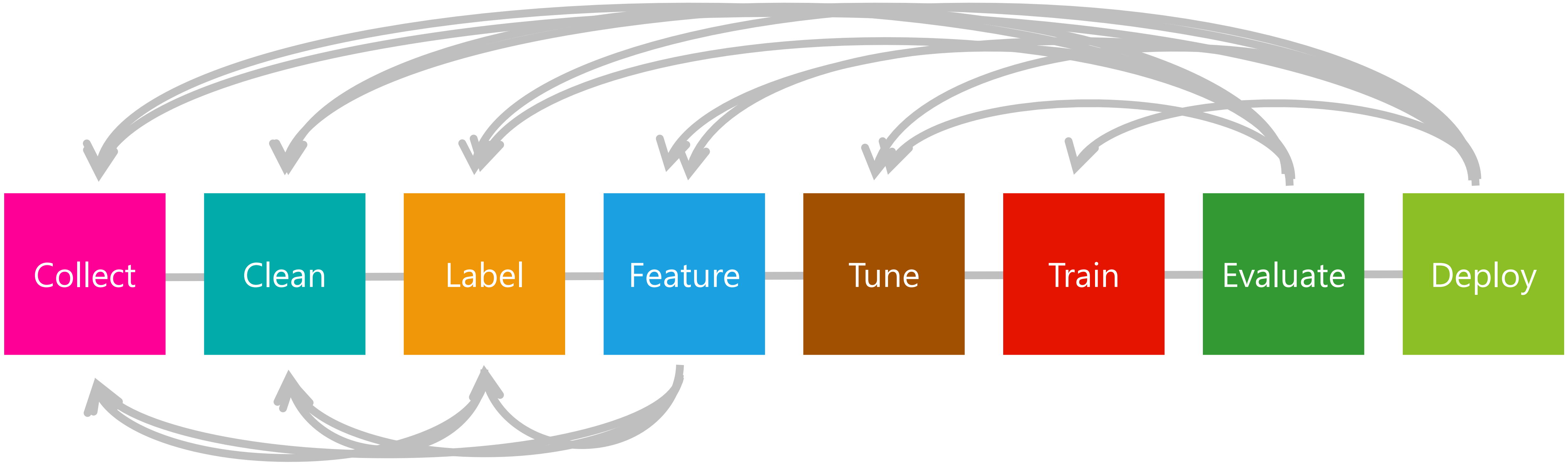

Collect & Label Samples
Create Features
Algorithm
Evaluate Results
Structured Labeling [CHI 2014]
Feature Insight [VAST 2015]
ModelTracker [CHI 2015, VAST 2016]



Many more opportunities to better support machine learning in practice.



Many more opportunities to better support machine learning in practice.



Many more opportunities to better support machine learning in practice and theory.

Making better sense of data.

Better data means better machine learning.

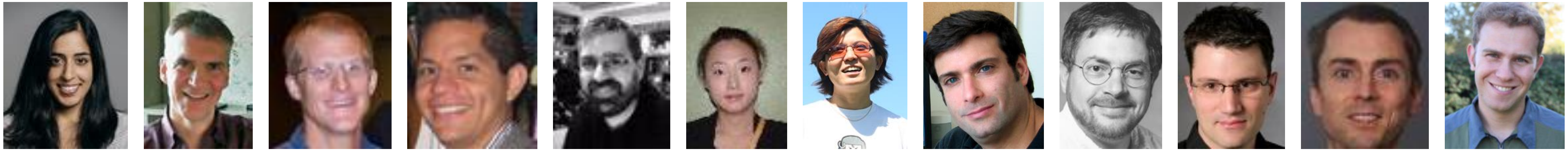
Most influence practitioners have on machine learning is through data.

Many more opportunities!

Thanks! Questions?

Better Machine Learning Through Data

Saleema Amershi, samershi@microsoft.com
Machine Teaching Group
Microsoft Research



August 14, 2016