Data Management Challenges in Machine Learning

Xu Chu

Assistant Professor in School of Computer Science

Georgia Tech

A Textbook View of ML

Machine learning is about learning patterns from data, often for making decisions or predictions.

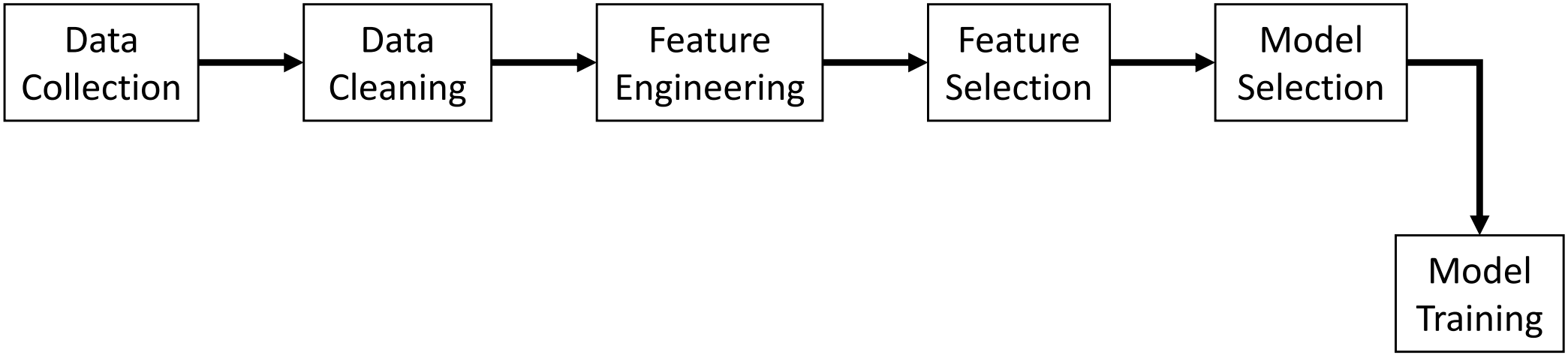
- Linear regression
- Logistic regression
- SVM
- Decision trees
- Neural nets
- …

Georgia Tech

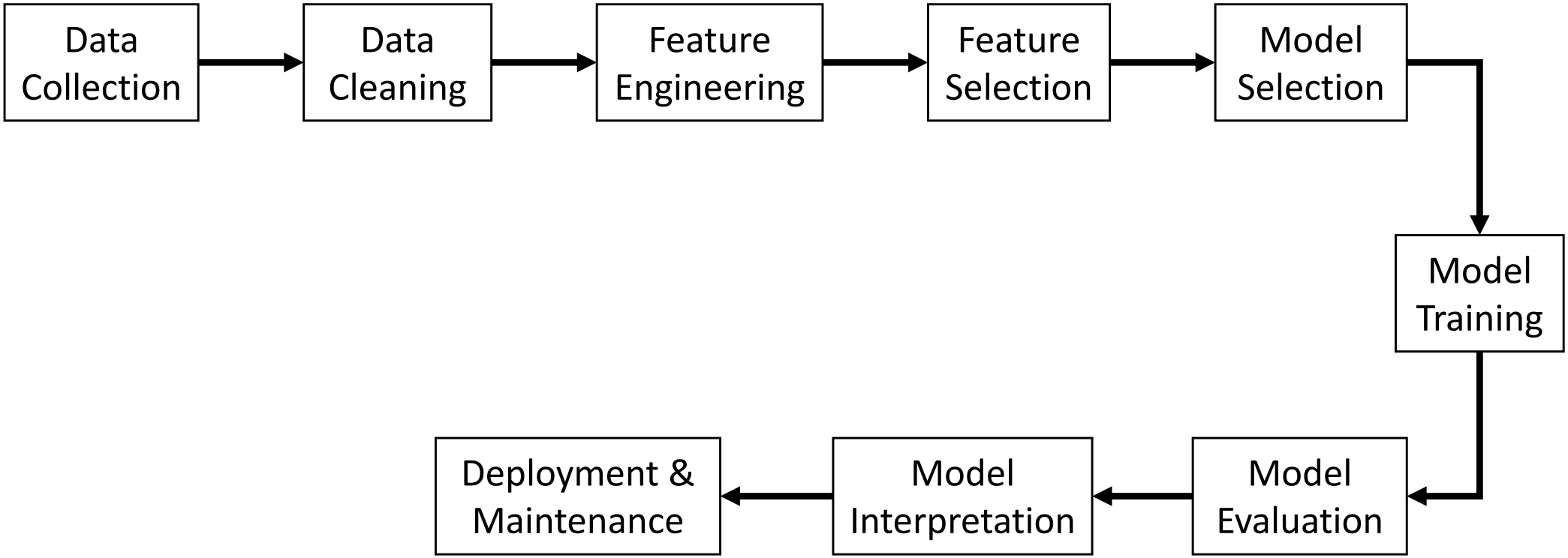
A Practitioner's View of ML

Model Training

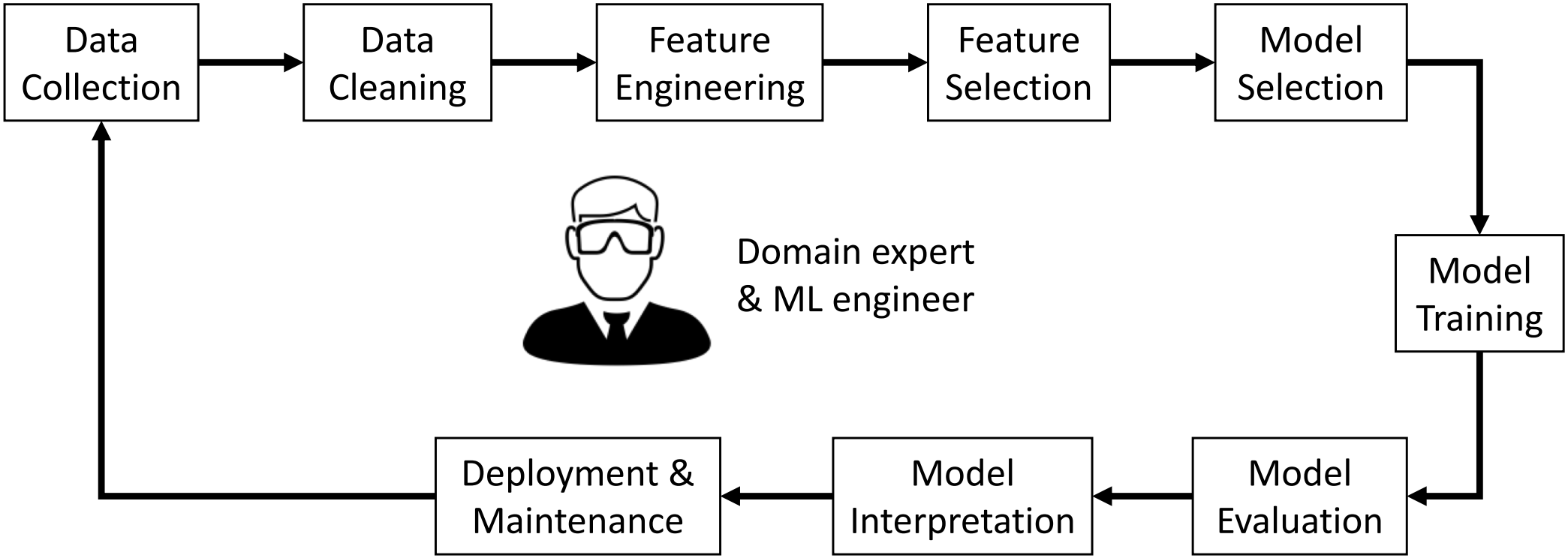
A Practitioner's View of ML



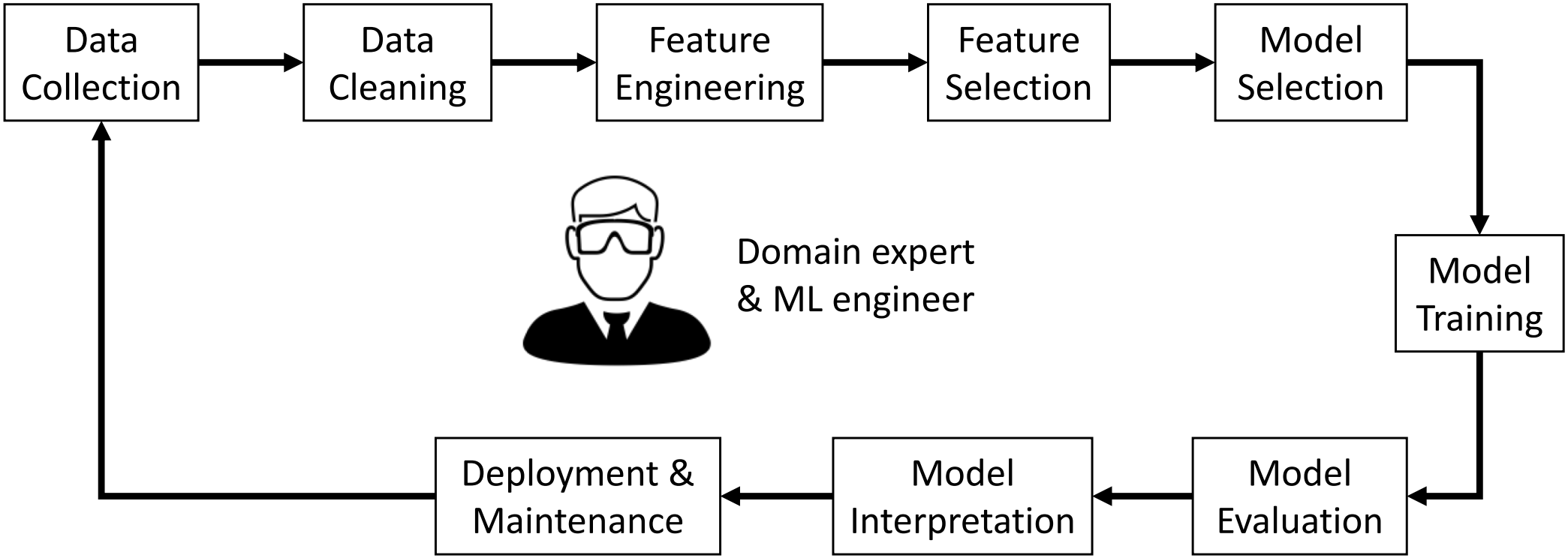
A Practitioner's View of ML



A Practitioner's View of ML

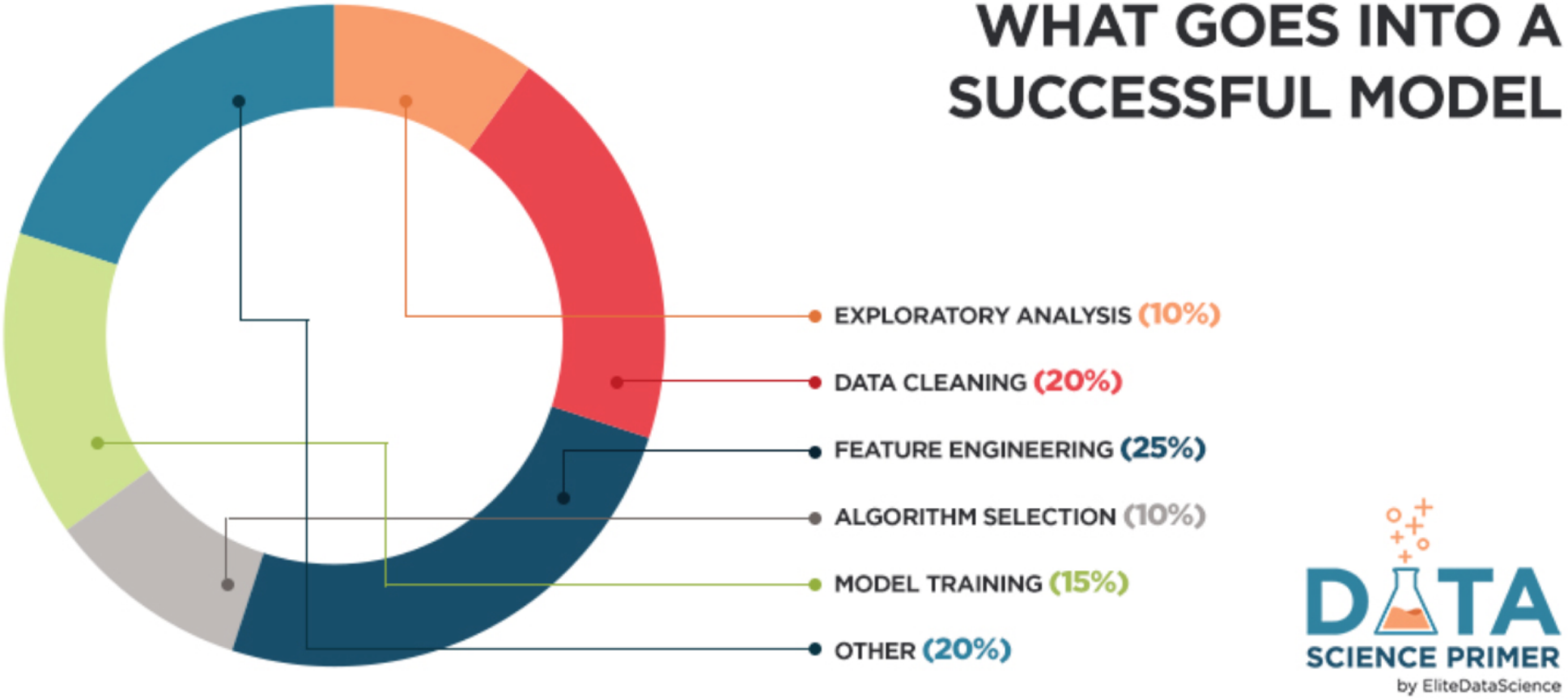


A Practitioner's View of ML



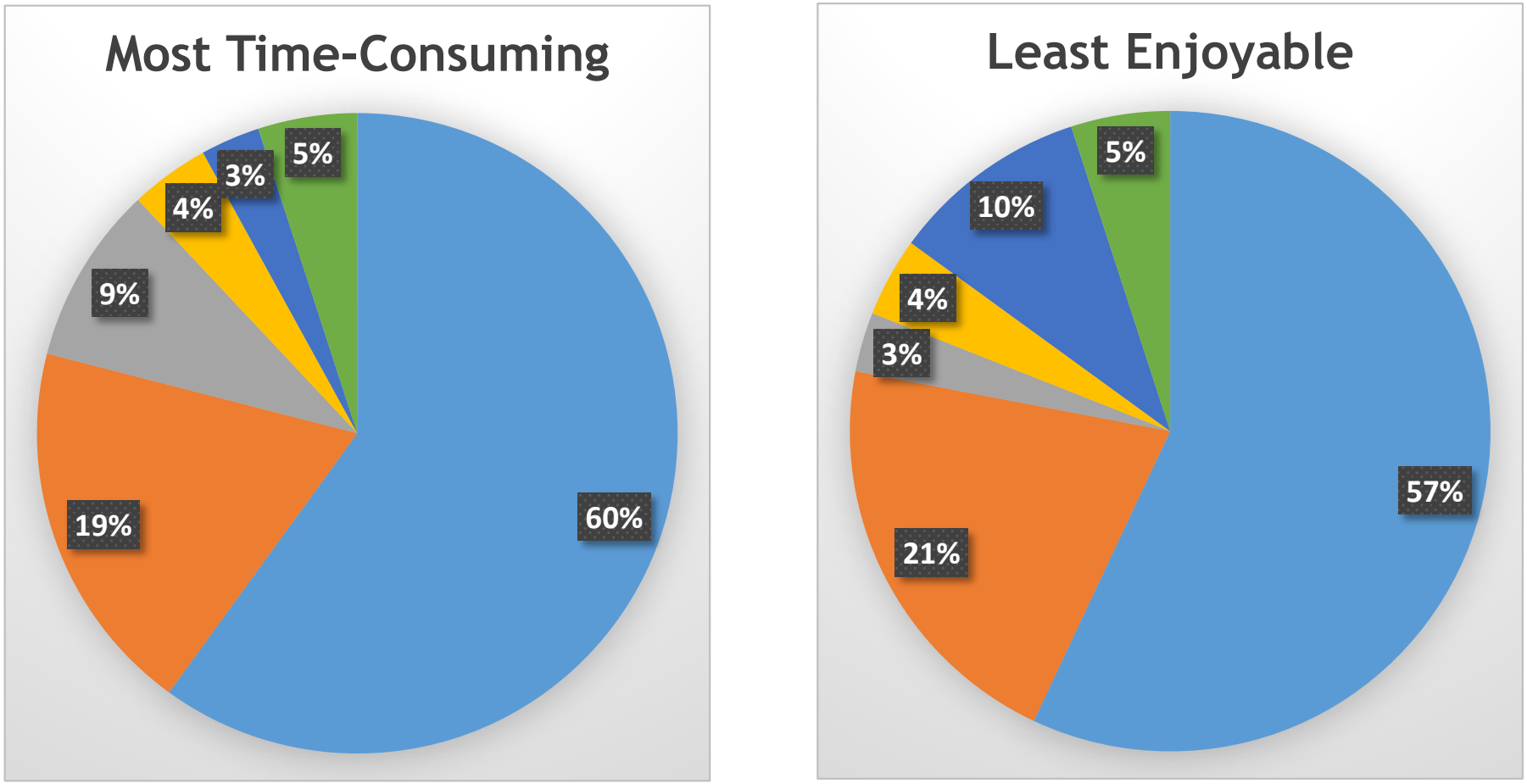
Task: Predict the selling price of a house in Atlanta

Statistics (1)



Georgia Tech

Statistics (2)



Forbes, 2016

Georgia Tech

Statistics (3)

Massive ongoing maintenance costs at the system level when applying ML

Machine Learning: The High-Interest Credit Card of Technical Debt

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov,
Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young
{dsculley,gholt,dgg,edavydov}@google.com
{toddphillips,ebner,vchaudhary,mwyoung}@google.com
Google, Inc

Abstract

Machine learning offers a fantastically powerful toolkit for building complex systems quickly. This paper argues that it is dangerous to think of these quick wins as coming for free. Using the framework of *technical debt*, we note that it is remarkably easy to incur massive ongoing maintenance costs at the system level when applying machine learning. The goal of this paper is highlight several machine learning specific risk factors and design patterns to be avoided or refactored where possible. These include boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, changes in the external world, and a variety of system-level anti-patterns.

Georgia Tech

How can database community help?

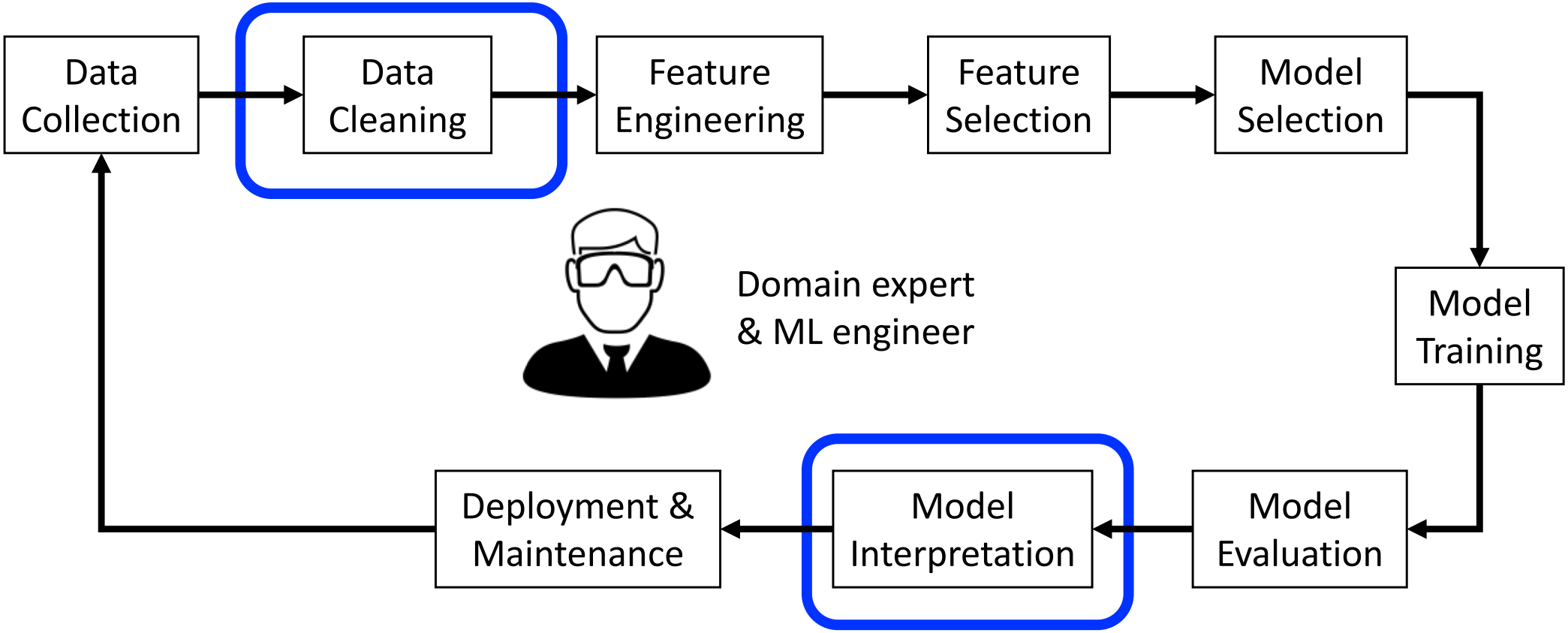
Database ≠ RDBMS

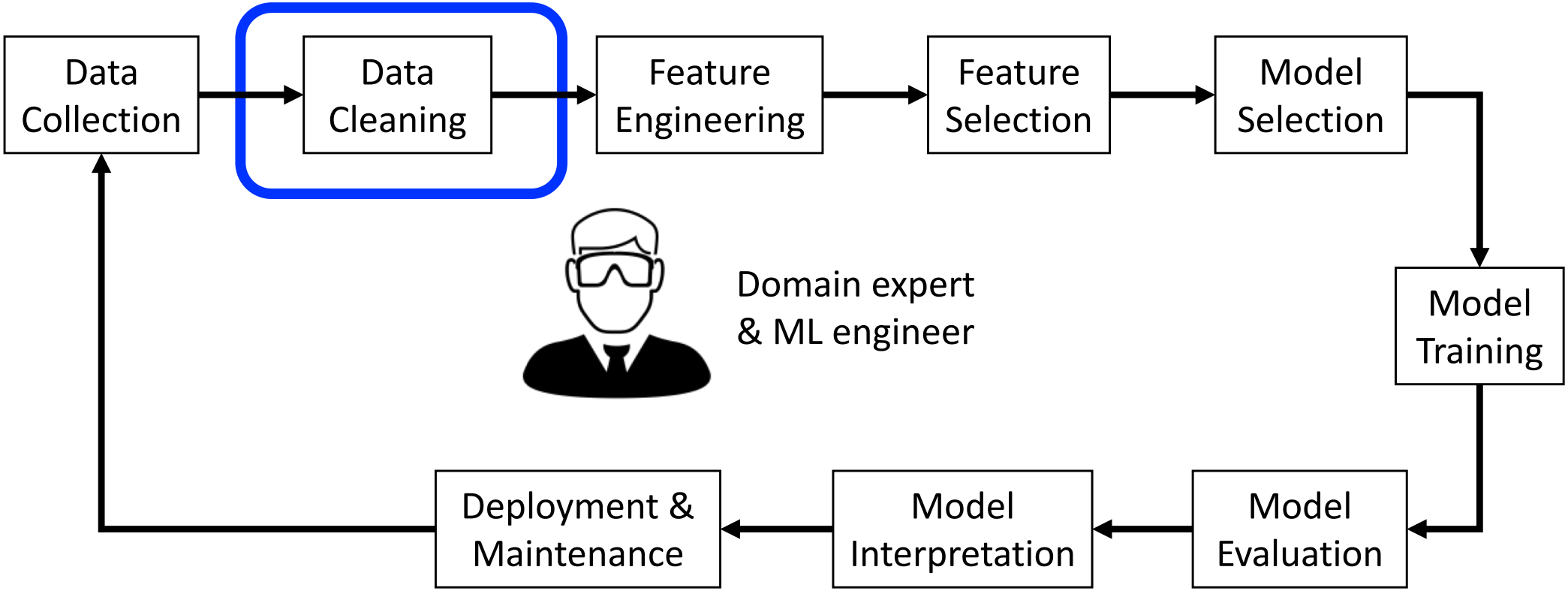
Data Management

- Information extraction
- Data profiling
- Data quality
- Data mining
- Data visualization
- Query optimization
- …

Georgia Tech

Today's Focus




Data Collection
Data Cleaning
Feature Engineering
Feature Selection
Model Selection
Model Training
Model Evaluation
Model Interpretation
Deployment & Maintenance
Domain expert & ML engineer

Common Data Errors

Incomplete

Country	UN R/P 10%[4]	UN R/P 20%[5]	World Bank Gini (%)[6]	WB Gini (year)	CIA R/P 10%[7]	Year	CIA Gini (%)[8]	CIA Gini (year)	GPI Gini (%)[9]
Seychelles			65.8	2007					
Comoros			64.3	2004					
Namibia	106.6	56.1	63.9	2004	129.0	2003	59.7	2010	
South Africa	33.1	17.9	63.1	2009	31.9	2000	65.0	2005	
Botswana	43.0	20.4	61.0	1994			63	1993	
Haiti	54.4	26.6	59.2	2001	68.1	2001	59.2	2001	
Angola			58.6	2000					62.0
Honduras	59.4	17.2	57.0	2009	35.2	2003	57.7	2007	

Georgia Tech

Common Data Errors

Inaccurate

Common Data Errors

Inconsistent

FlightView

American Airlines Flight Number 119 (AA119)

FLIGHT TRACKER

Departure

Airport:	
Scheduled Time:	6:15 PM, Dec 08
Takeoff Time:	6:53 PM, Dec 08
Terminal - Gate:	Terminal A - 32

ArrivalStatus: In Air

Airport:	
Scheduled Time:	9:40 PM, Dec 08
	9:42 PM, Dec 08
Estimated Time:	Track This Flight Live!
Time Remaining:	25 min
Terminal - Gate:	Terminal 4 - 42B
Baggage Claim:	4

FlightAware

AAL119 (Track inbound flight)

(web site) (all flights)

American Airlines "American"

Aircraft	Boeing 737-800 (twin-jet) (B738/Q - track or photos)
Origin	Terminal A / Gate 32 / Newark Liberty Intl (KEWR - tr
Destination	Terminal 4 / Gate 42B / Los Angeles Intl (KLAX - track
	Other flights between these airports
Route	ZIMMZ Q42 BTRIX Q480 AIR J80 VHP J80 MCI J24 SLN J102 ALS J44 RSK J (Decode)
Date	2011年 12月 08日 (Thursday)
Duration	5 hours 43 minutes
Progress	20 minutes left 5 hours 23 minutes
Status	En Route (2,284 sm down; 168 sm to go)
Distance	Direct: 2,451 sm Planned: 2,458
Fare	$51.99 to $3,561.11; average: $241.96 (airline insight)
Cabin	First: Dinner / Economy: Food for sale

	Scheduled	7-day Average	Actual/*Estimated*
Departure	06:15PM EST	07:08PM EST	06:53PM EST
Arrival	08:33PM PST	09:17PM PST	*09:36PM PST*

Orbitz

American Airlines # 119

Leg 1: In Transit

Departs: Newark (EWR) View real-time airpo

Gate: 32

Scheduled		Estimated	Actual
6:22p Dec 8	-		**6:32p** Dec 8

Arrives: Los Angeles (LAX) View real-time ai

Gate: 42B

Scheduled	Estimated	Actual
9:54p Dec 8		**9:47p** Dec 8

Georgia Tech

Common Data Errors

Duplicated

Merged citations

This "Cited by" count includes citations to the following articles in Scholar. The ones marked * may be different from the article in the profile.

Holistic data cleaning: Putting violations into context
X Chu, IF Ilyas, P Papotti
Data Engineering (ICDE), 2013 IEEE 29th International Conference on, 458-469, 2013

Holistic data cleaning: Put violations into context
X Chu, IF Ilyas, P Papotti
In ICDE, 2013

Georgia Tech

The Data Cleaning Landscape

Outlier detection

Data quality rules

Data deduplication

Data transformation

The Data Cleaning Landscape

Outlier detection

Data quality rules

Data deduplication

Data transformation

Outlier Detection

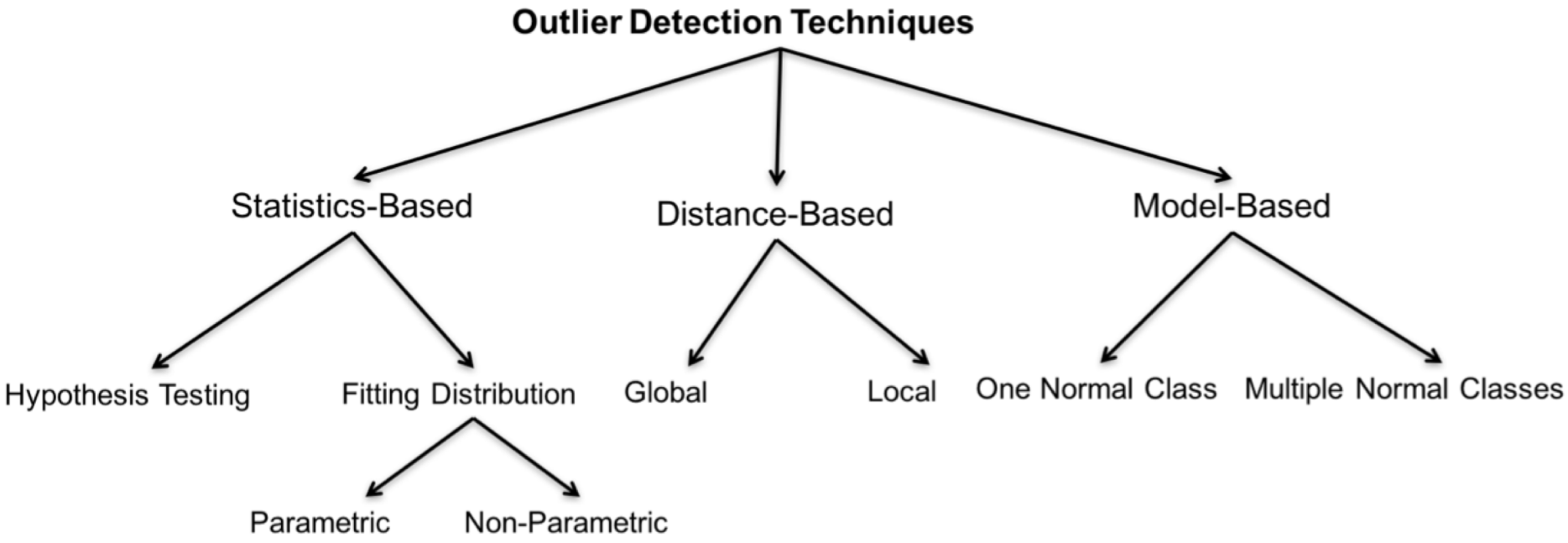
An outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism

--- Hawkins 1980

	Name	*age*	*income*	*tax*
t_1	Vivian Baskette	1	70	7
t_2	Jamison Marney	25	110	11
t_3	Marie Mulero	27	80	8
t_4	Trudi Kimmell	30	130	13
t_5	Stepanie Lindemann	32	120	7
t_6	Dia Werley	35	80	8
t_7	Abbie Lama	40	90	9
t_8	Misti Luce	41	100	10
t_9	Wilda Byerly	1000	120	12

Georgia Tech

How to Detect Outliers



The Data Cleaning Landscape

Outlier detection

Data quality rules

Data deduplication

Data transformation

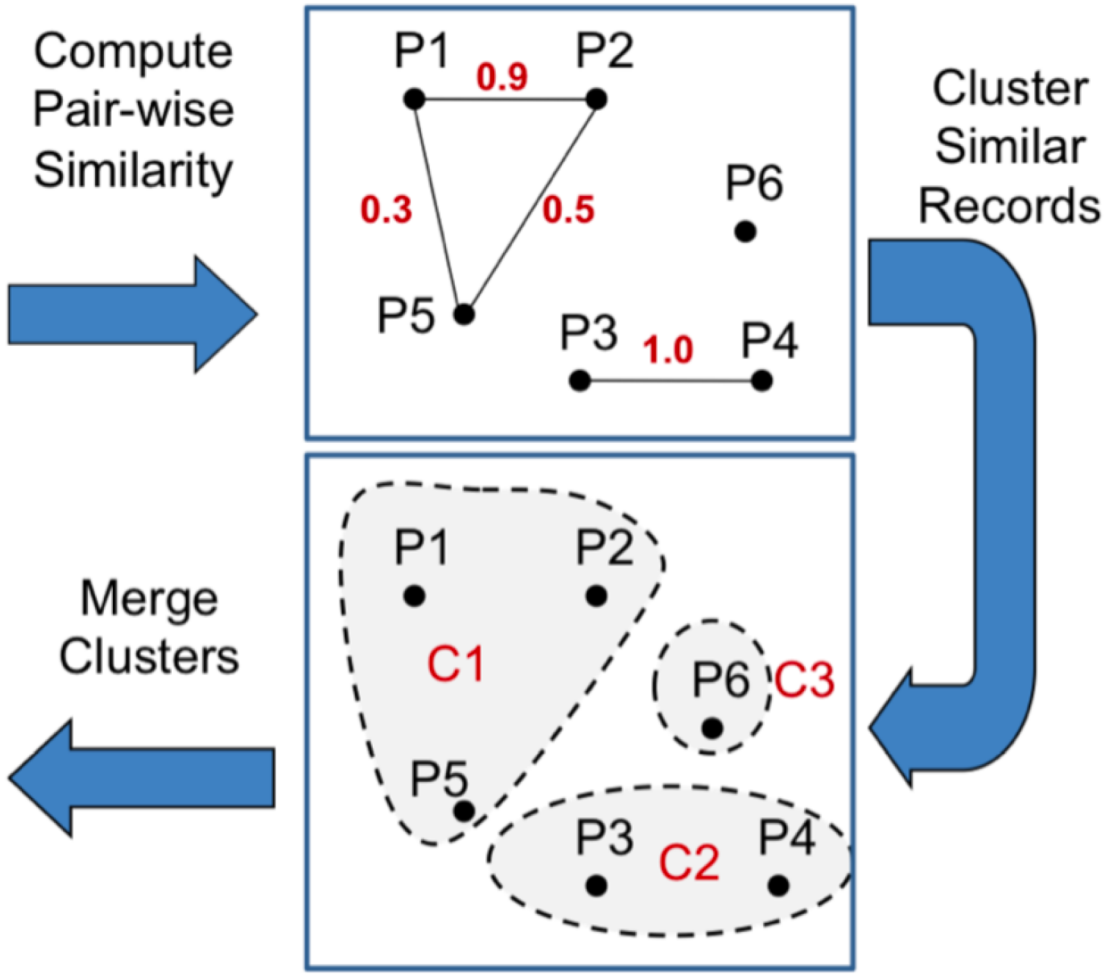
Data Deduplication

Unclean Relation

ID	name	ZIP	Income
P1	Green	51519	30k
P2	Green	51518	32k
P3	Peter	30528	40k
P4	Peter	30528	40k
P5	Gree	51519	55k
P6	Chuck	51519	30k

Clean Relation

ID	name	ZIP	Income
C1	Green	51519	39k
C2	Peter	30528	40k
C3	Chuck	51519	30k



The Data Cleaning Landscape

Outlier detection

Data quality rules

Data transformation

Data deduplication

Data Transformation

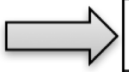
Syntactic

Phone
7188751243
7186359762
5198780763
5176543809

Phone
718-875-1243
718-635-9762
519-878-0763
517-654-3809

Semantic

Country Code
US
CA
CN
DE



Country
United States
Canada
China
Germany

Layout

Name	
John	Tel:7188751243
	Fax:7188751200
Mike	Tel:7186359762
	Fax:7186359700
Frank	Tel:5198780763
	Fax:5198780700
Julie	Tel:5176543809
	Fax:5176543800

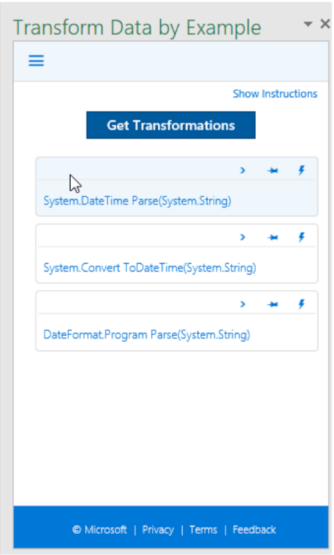
Name	Tel	Fax
John	7188751243	7188751200
Mike	7186359762	7186359700
Frank	5198780763	5198780700
Julie	5176543809	5176543800

Georgia Tech

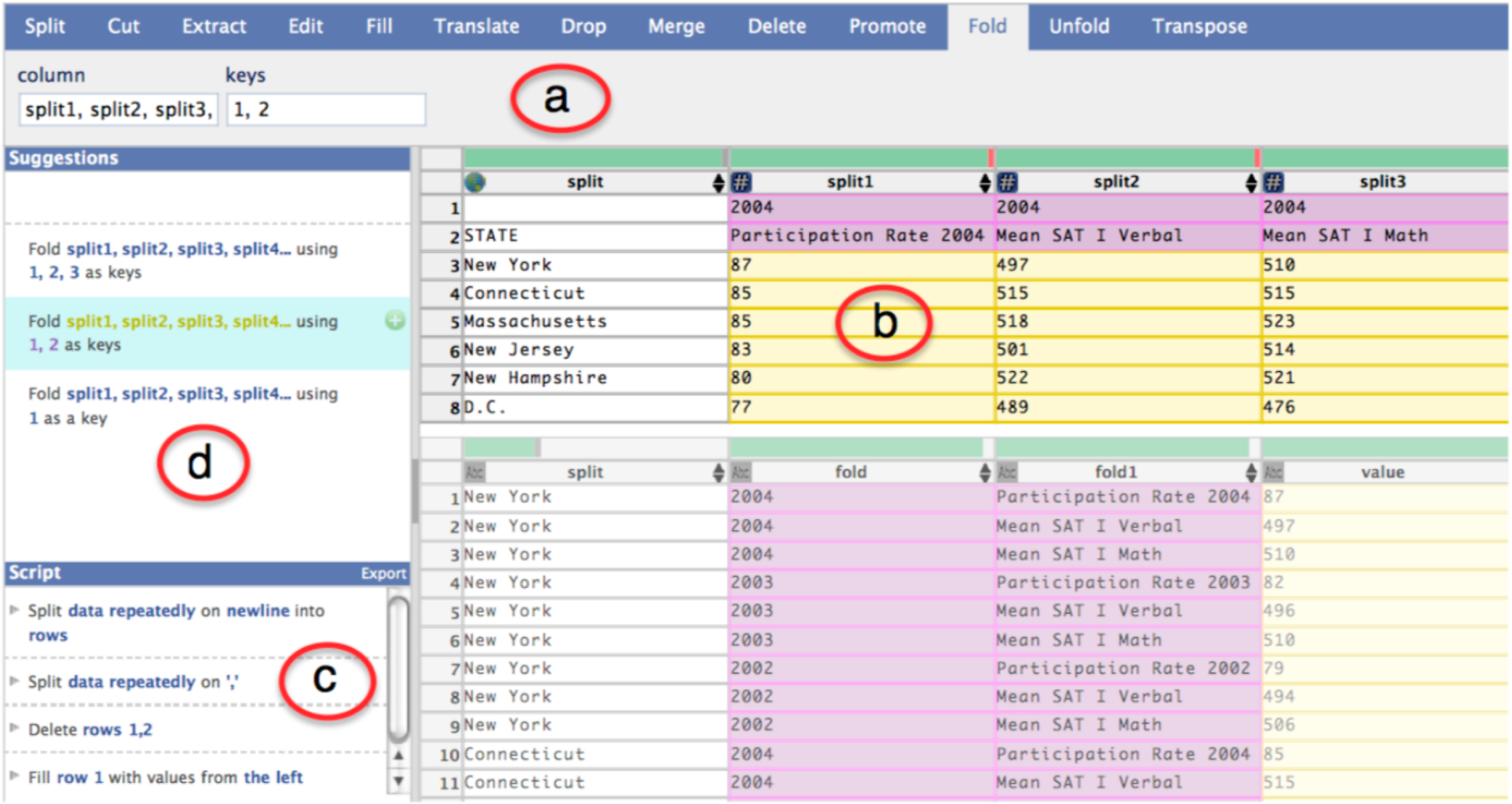
How to Transform: Example-Driven

C	D
Transaction Date	**output**
Wed, 12 Jan 2011	2011-01-12-Wednesday
Thu, 15 Sep 2011	2011-09-15-Thursday
Mon, 17 Sep 2012	
2010-Nov-30 11:10:41	
2011-Jan-11 02:27:21	
2011-Jan-12	
2010-Dec-24	
9/22/2011	
7/11/2012	
2/12/2012	

C	D
Transaction Date	**output**
Wed, 12 Jan 2011	2011-01-12-Wednesday
Thu, 15 Sep 2011	2011-09-15-Thursday
Mon, 17 Sep 2012	2012-09-17-Monday
2010-Nov-30 11:10:41	2010-11-30-Tuesday
2011-Jan-11 02:27:21	2011-01-11-Tuesday
2011-Jan-12	2011-01-12-Wednesday
2010-Dec-24	2010-12-24-Friday
9/22/2011	2011-09-22-Thursday
7/11/2012	2012-07-11-Wednesday
2/12/2012	2012-02-12-Sunday



How to Transform: Declarative



The Data Cleaning Landscape

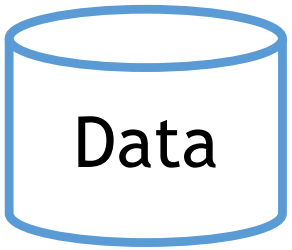
Outlier detection

Data quality rules

Data transformation

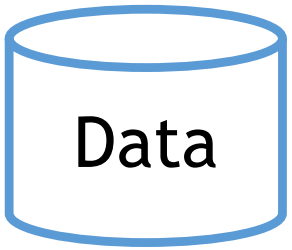
Data deduplication

Rule-based Data Cleaning



Name	ZIP	ST
Alice	10001	NM
Bob	87101	NM
Chris	10001	NY

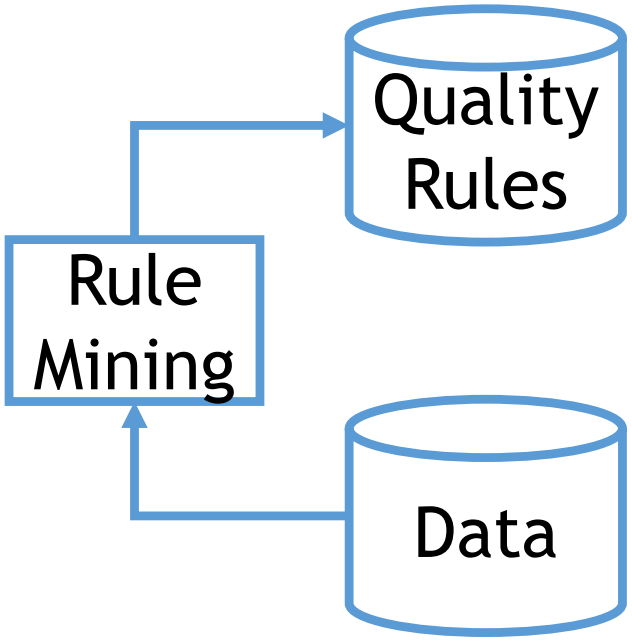
Rule-based Data Cleaning



Name	ZIP	ST
Alice	10001	NM
Bob	87101	NM
Chris	10001	NY

Two persons with the same ZIP live in the same ST

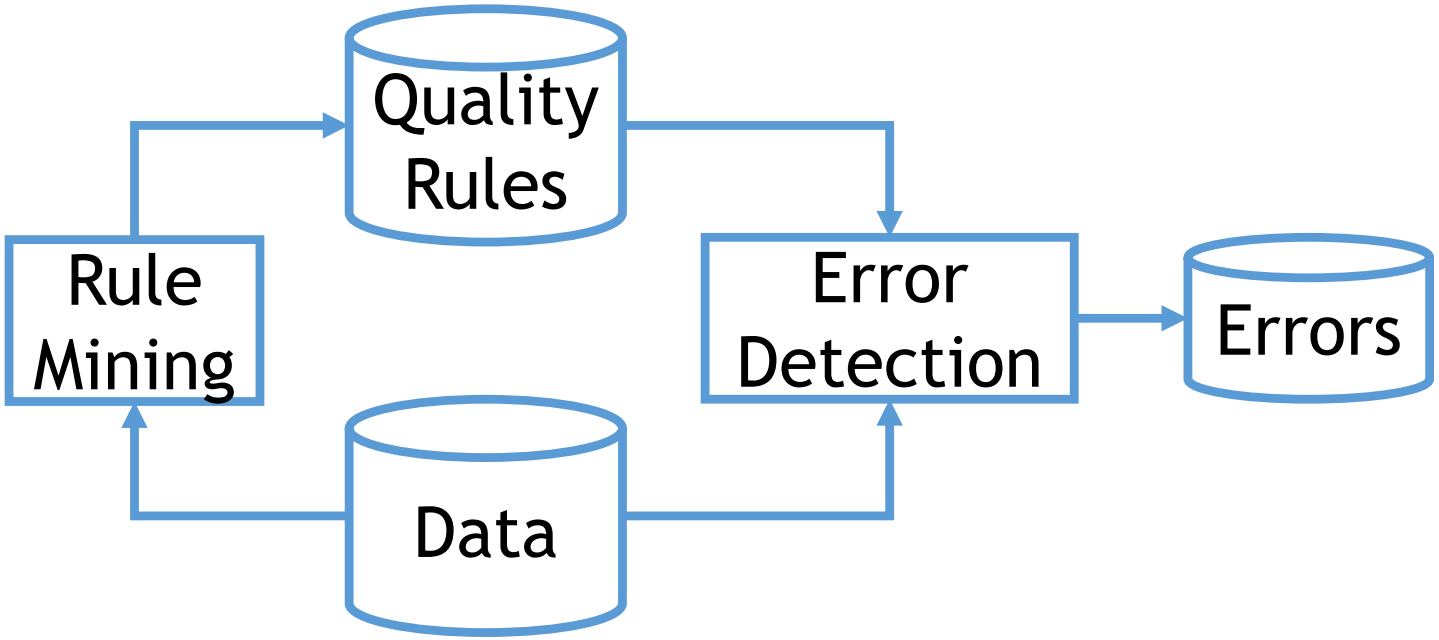
Rule-based Data Cleaning



Name	ZIP	ST
Alice	10001	NM
Bob	87101	NM
Chris	10001	NY

Two persons with the same ZIP live in the same ST

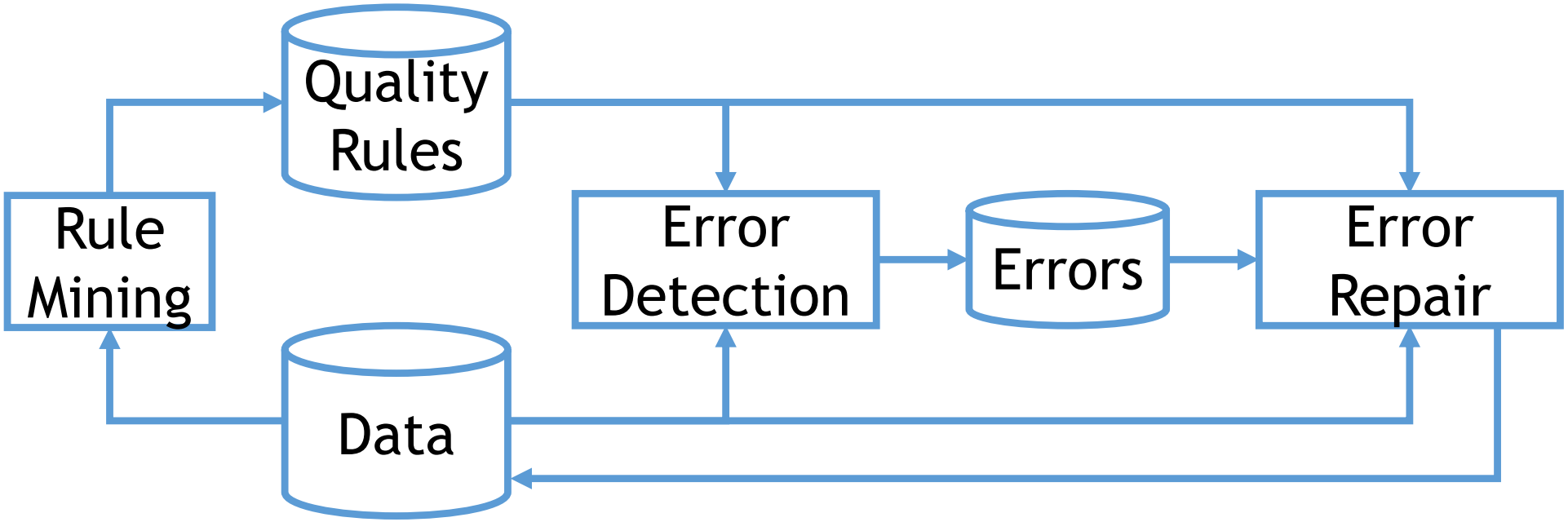
Rule-based Data Cleaning



Name	ZIP	ST
Alice	10001	NM
Bob	87101	NM
Chris	10001	NY

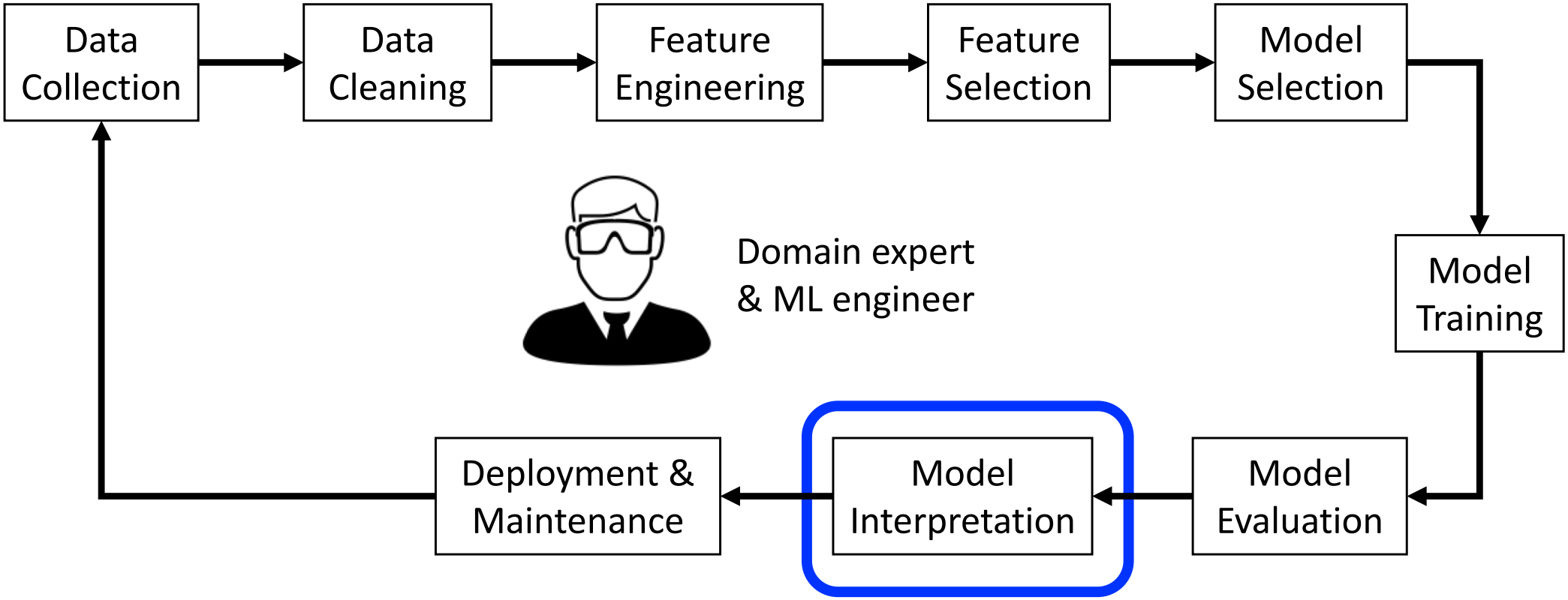
Two persons with the same ZIP live in the same ST

Rule-based Data Cleaning



Name	ZIP	ST
Alice	10001	NY
Bob	87101	NM
Chris	10001	NY

Two persons with the same ZIP live in the same ST


Data Collection
Data Cleaning
Feature Engineering
Feature Selection
Model Selection
Model Training
Model Evaluation
Model Interpretation
Deployment & Maintenance
Domain expert & ML engineer

Model Interpretation

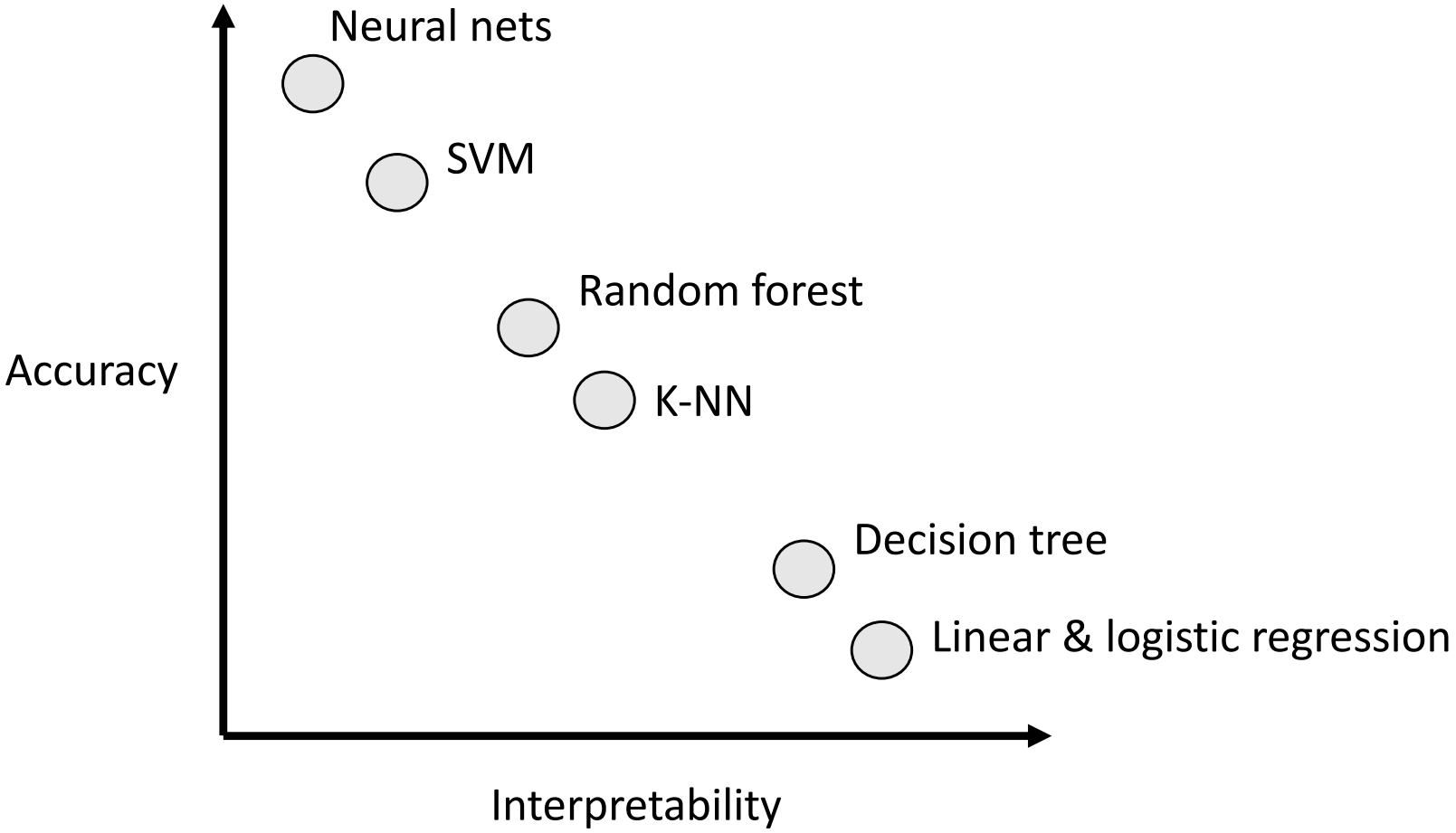
- What: Interpretability is the degree to which a human can understand the cause of a decision.

Miller, Tim. 2017. "Explanation in Artificial Intelligence: Insights from the Social Sciences." arXiv Preprint arXiv:1706.07269

- Why:
  - Improve ML users' understanding and trust
    - Model accuracy
    - How much users can understand the model's behavior
  - Enable ML System designers
    - Feature engineering
    - Parameter tuning
    - Model selection
    - …

Georgia Tech

Interpretability-Accuracy Trade-off



How to Achieve Interpretability

- Use interpretable models **(Model-specific)**
  - Decision trees
  - Linear regression
  - Logistic regression
  - …
- Interpreting complex, black-box models **(Model-agnostic)**
  - Neural nets
  - SVMs
  - …

Georgia Tech

Scope of Interpretation

- Global: the entire relationship the model encodes, i.e., the conditional distribution (hard, often approximate or based on average)

- Local: small regions of the model (more likely to be linear, monotonic)
  - Clusters of input examples
  - Quantiles of predictions
  - Often a single prediction

- Promote understanding by comparing global vs local
  - For examples with globally extreme predictions, determine if their local explanations justify their extreme predictions
  - For examples whose local and global explanations differ, determine if their local explanations are reasonable
  - For examples with globally median predictions or probabilities, analyze if their local explanations are similar to the global

Interpretable Models (Model-Specific)

- Linear regression
- Decision tree
- …

Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

Interpretation of a Numerical Feature

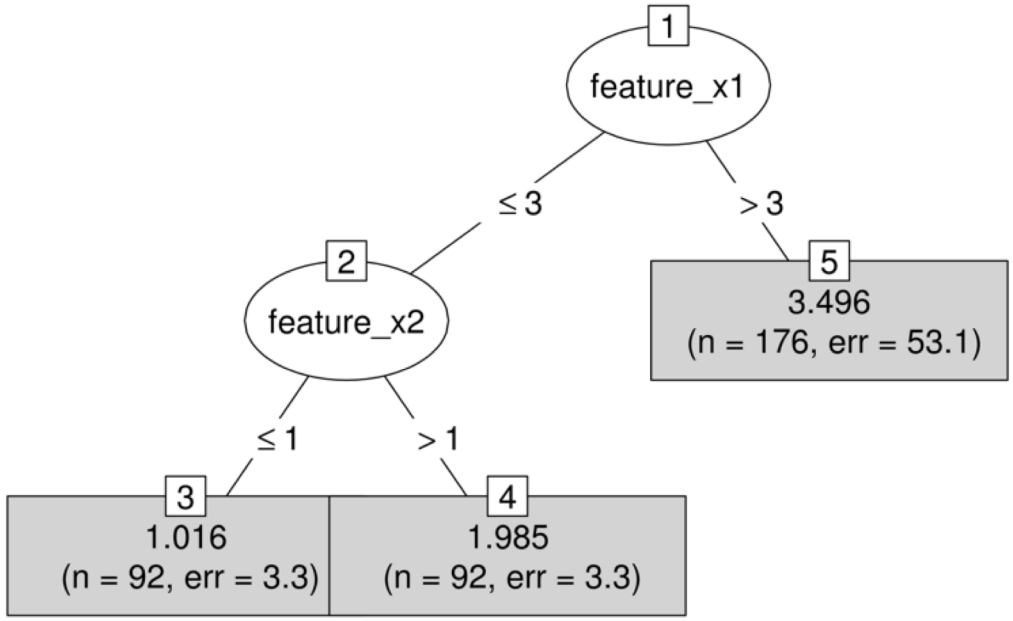
An increase of x_k by one unit increases the expectation for y by β_k units, given all other features stay the same.

Interpretation of a Categorical Feature

A change from x_k's reference level to the other category increases the expectation for y by β_k, given all other features stay the same.

Georgia Tech

Decision Trees



The interpretation is simple: Starting from the root node you go to the next nodes and the edges tell you which subsets you are looking at. Once you reach the leaf node, the node tells you the predicted outcome. All the edges are connected by 'AND'.

Template: If feature x is [smaller/bigger] than threshold c AND ..., then the predicted outcome is $\hat{y}_{\text{leafnode}}$.

Interpreting Complex Models (Model-Agnostic)

- Surrogate models and LIME
- PDP and ICE
- Feature importance
- Shapley value
- …

Surrogate Model (Global)

- A surrogate model h is trained on the predictions of the learned model g (not on the original dataset)
  - Use R^2 error to measure the fit of h to g

- Pro: Any interpretable models can be used as surrogates

- Con: No guarantee the surrogate h is representative of g

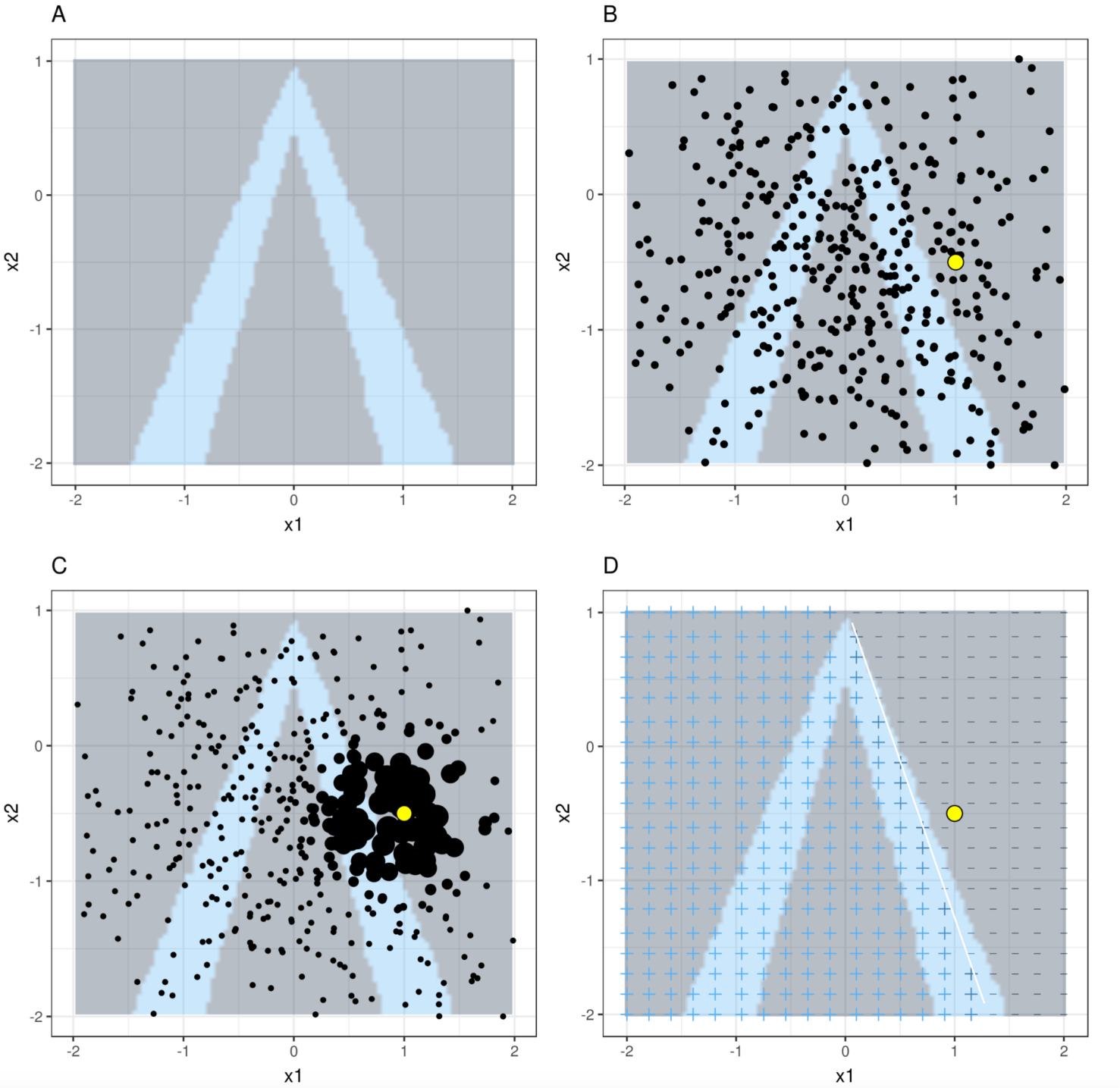
LIME (Local)

- Local interpretable model-agnostic explanations
  - Local, interpreting a single instance
  - Use surrogate models to fit a local region

Georgia Tech

Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). ACM

LIME



Source: https://christophm.github.io/interpretable-ml-book/lime.html

LIME (Local)

- Local interpretable model-agnostic explanations
  - Local, interpreting a single instance
  - Use surrogate models to fit a local region

- Pro:
  - Model-agnostic local interpretation
  - Can use any interpretable models

- Con:
  - Difficult to define/choose a local neighborhood

Georgia Tech

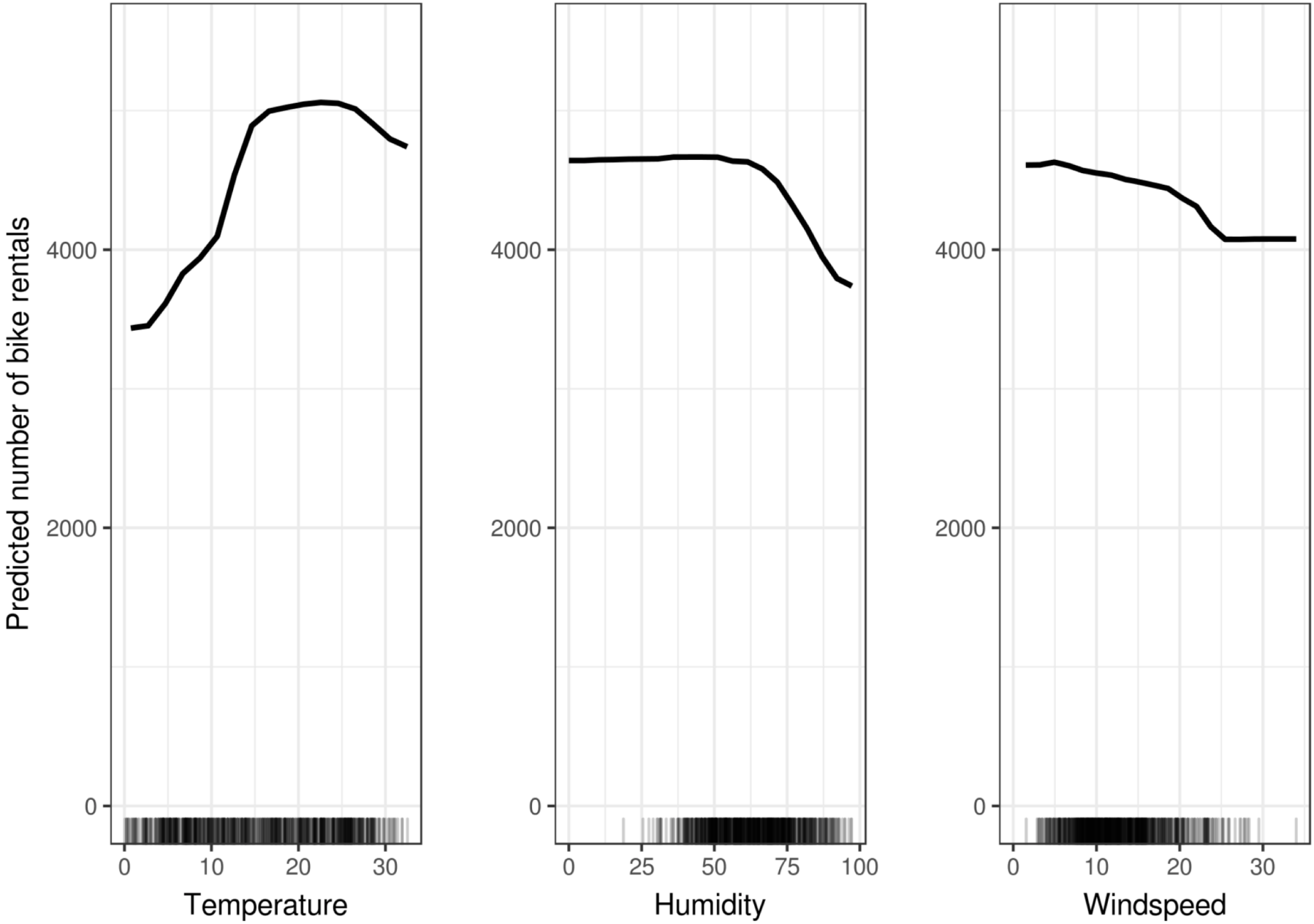
Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). ACM

PDP (Global)

- The partial dependence plot shows the marginal effect of a feature on prediction of a black-box model f

$$f(x[A_1]) = \frac{1}{n}\sum_{i=1}^{n} f(x[A_1], x_i[A_2, \ldots, A_n])$$

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." Annals of Statistics. JSTOR, 1189–1232



Source: https://christophm.github.io/interpretable-ml-book/pdp.html#fn25

PDP (Global)

- The partial dependence plot shows the marginal effect of a feature on prediction of a previously fit model

$$f(x[A_1]) = \frac{1}{n}\sum_{i=1}^{n} f(x[A_1], x_i[A_2, \ldots, A_n])$$

- Pro:
  - Model-agnostic
  - Intuitive

- Con:
  - Assumption of independence. May average over ”impossible” instances

Georgia Tech

Friedman, Jerome H. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” Annals of Statistics. JSTOR, 1189–1232

ICE (Local)

- Individual Conditional Expectation (ICE) plots draw one line per instance, representing how the instance's prediction changes when the feature changes.

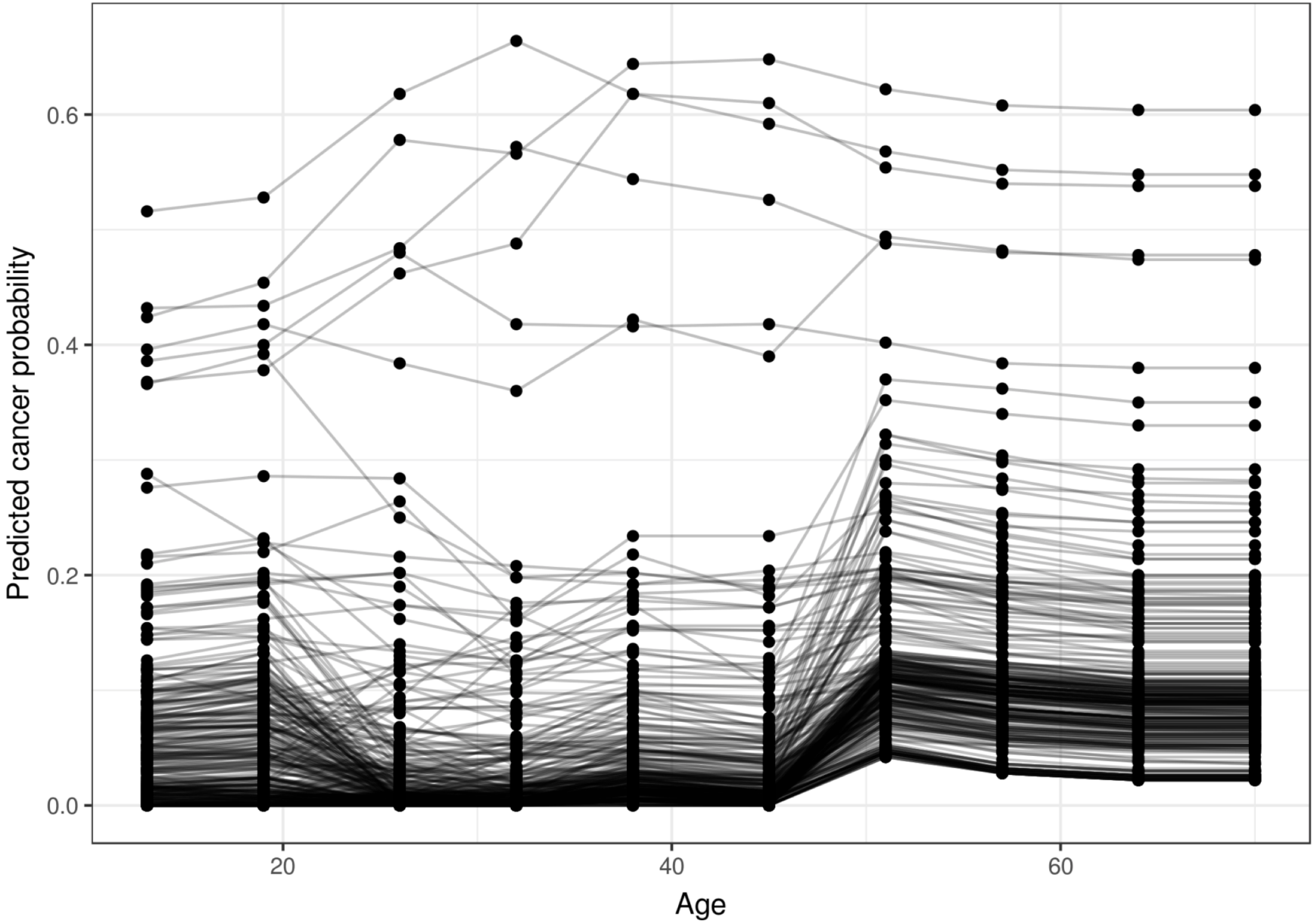
$$f(x[A_1]) = f(x[A_1], x_i[A_2, \ldots, A_n])$$

- Pro:
  - Model-agnostic
  - Intuitive

- Con:
  - Assumption of independence. May average over "impossible" instances

Georgia Tech

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." Annals of Statistics. JSTOR, 1189–1232



Georgia Tech

Source: https://christophm.github.io/interpretable-ml-book/ice.html#example-1

Feature Importance (Global)

- A feature's importance is the increase in the model's prediction error after we permuted the feature's values (breaks the relationship between the feature and the outcome).

- Pro:
  - Model-agnostic
  - Intuitive

- Con:
  - Tied to the error model
  - No access to actual predictions of the model

- Breiman, Leo. 2001. "Random Forests." Machine Learning 45 (1). Springer: 5–32
- Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. 2018. "Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the 'Rashomon' Perspective.

The algorithm:

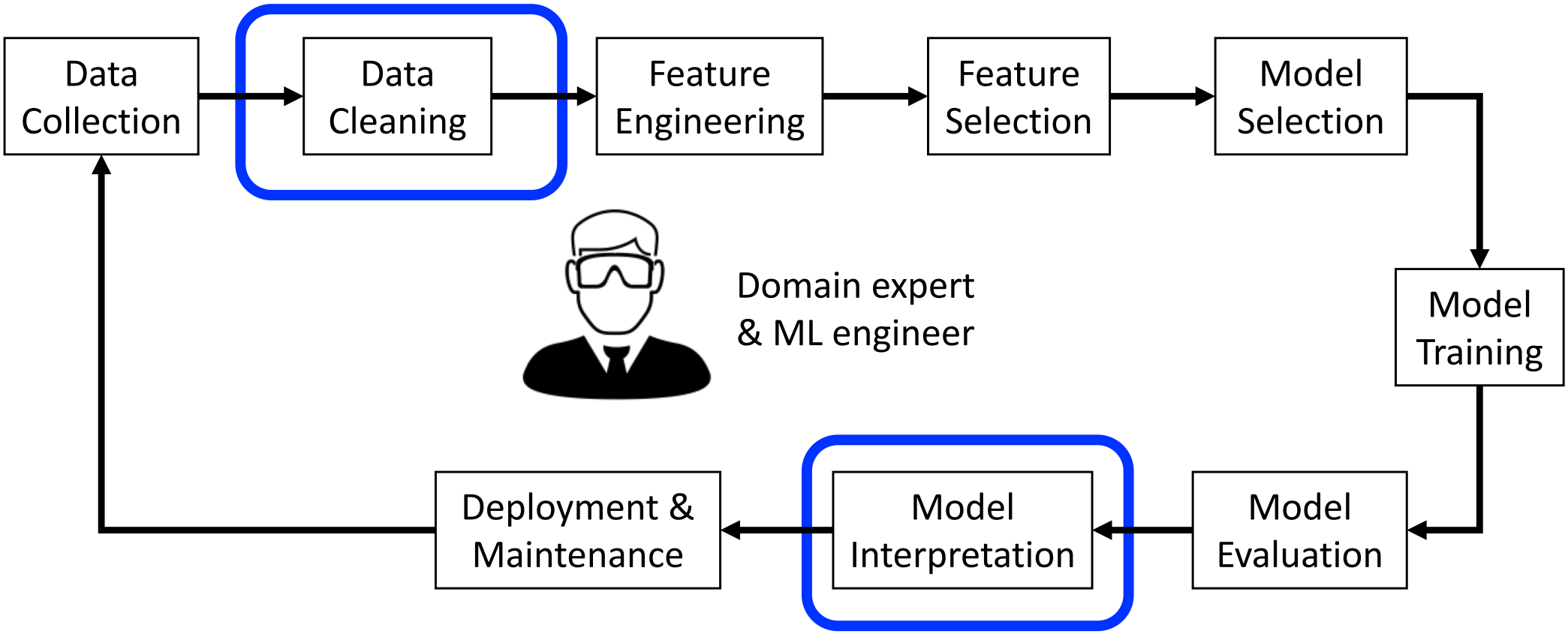
Input: Trained model $\hat{f}$, feature matrix X, target vector Y, error measure $L(Y, \hat{Y})$

1. Estimate the original model error $e_{orig}(\hat{f}) = L(Y, \hat{f}(X))$ (e.g. mean squared error)
2. For each feature

   $j \in 1, \ldots, p$ do
   - Generate feature matrix X_{perm_j} by permuting feature X_j in X. This breaks the association between X_j and Y.
   - Estimate error $e_{perm} = L(Y, \hat{f}(X_{perm_j}))$ based on the predictions of the permuted data.
   - Calculate permutation feature importance $FI_j = e_{perm}(\hat{f})/e_{orig}(\hat{f})$. Alternatively, the difference can be used: $FI_j = e_{perm}(\hat{f}) - e_{orig}(\hat{f})$
3. Sort variables by descending FI.

Source: https://christophm.github.io/interpretable-ml-book/feature-importance.html

Conclusion



CS 8803 (Fall 2018): Data Management Challenges in ML

xu.chu@cc.gatech.edu

Georgia Tech