Lecture Notes for

Machine Learning in Python

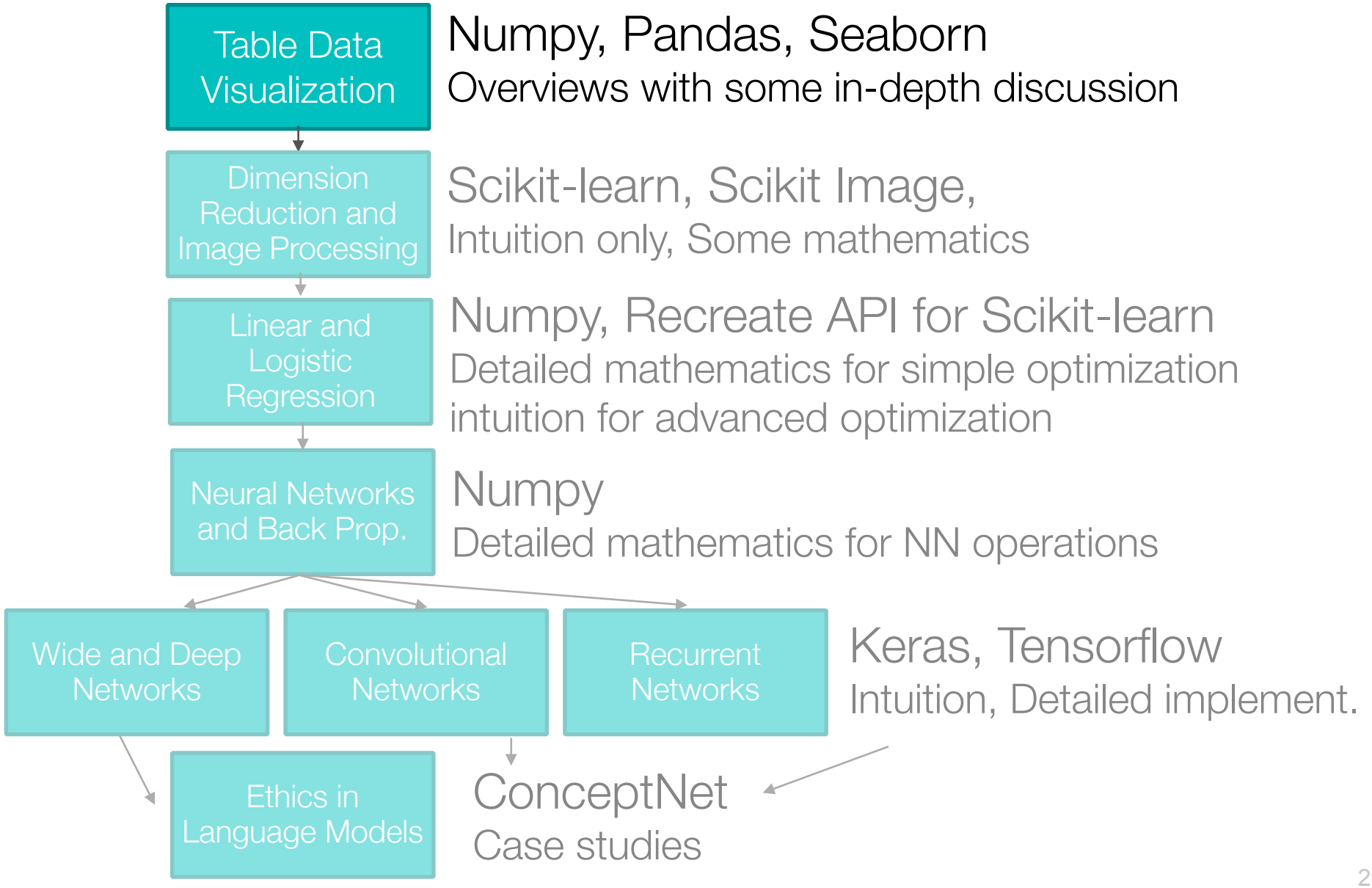
Professor Eric Larson

Numpy, Pandas, Document Features

Class Logistics and Agenda

- Canvas? Anaconda Installs?
- In-person versus Zoom and other classes
- Agenda:
  - Finish Table Data, Numpy
  - Data Quality
  - Attributes Representation
    - documents
  - The Pandas eco-system
    - loading and manipulating attributes
- Needing some more help?
  - **fast.ai** has great, free resources

Class Overview, by topic



Review: Example Classification: Malware

- Classify files as malware based on size and naming.
- Approach:
  - Use already classified malware files
  - Must translate name to set of features
  - ***{malware, not malware}*** decision forms the ***class attribute***
  - Collect various malware examples and a number of safe files, providing labels for each and a set of features

Training Set

TID	Name	Size	Class
1	erte.dll	916 b	not
2	fufu.bin	1M	yes
3	exe.exe	1G	not
4	ex.py	113 b	not

Unknown

TID	*Name*	*Size*
1	asdf.dll	11b

Review: Example Regression: Housing Price

- Predict a value of a given *continuous valued* variable based on the values of other variables
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Predicting House Sales

Training Set

TI	# Rms	Sq Ft	Zip	Price
1	2	1125	74012	150K
2	2	2525	75155	200k
3	10	4678	90210	3M
4	4	2678	75154	350k

Unknown

TI	# Rms	Sq Ft	Zip
1	2	2200	75115

Example Classifying: Objects in Images

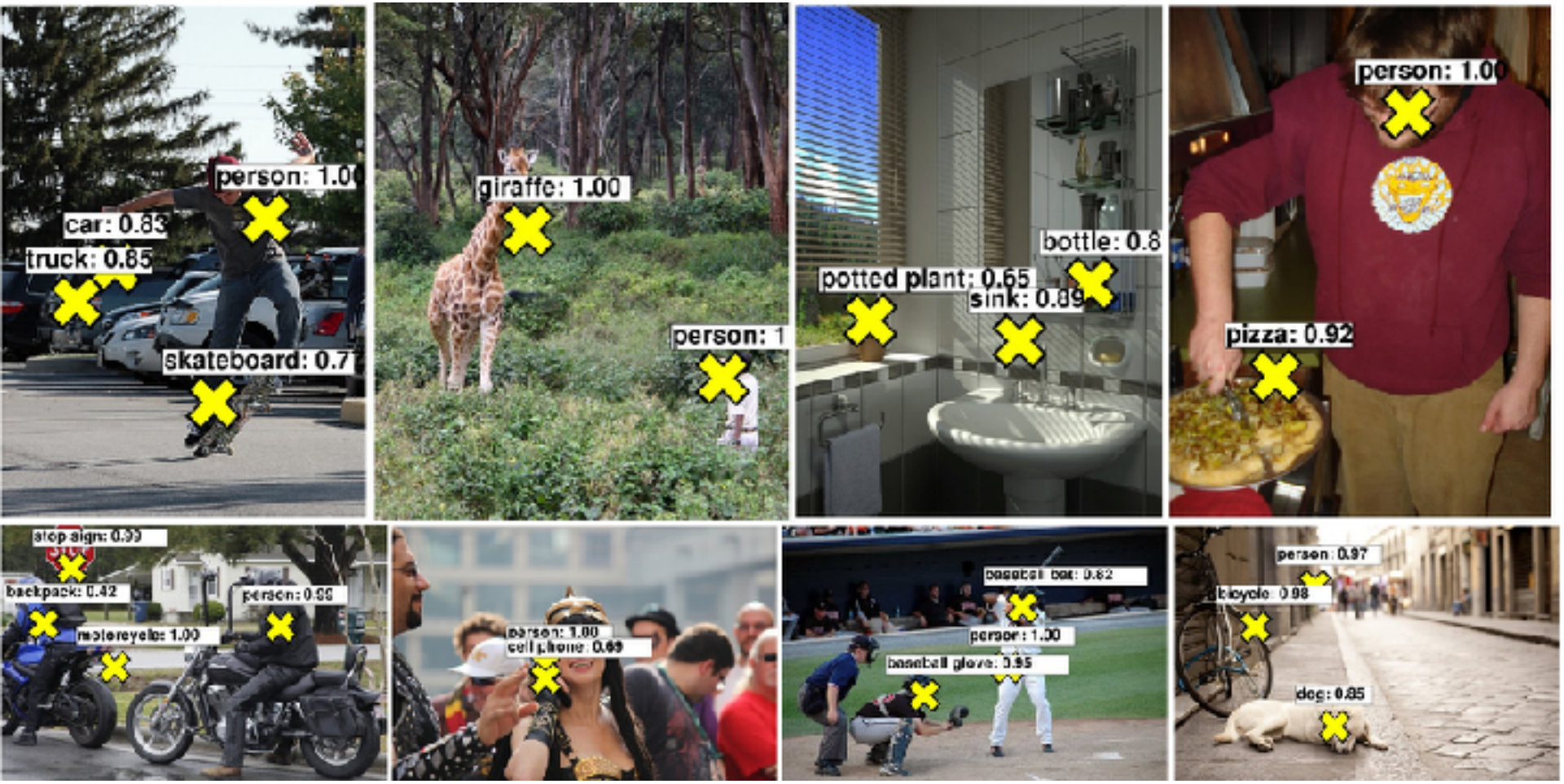


Image Net:

- 14 million images
- 200 Labeled Categories
- 1000 Location Labels

Attributes:

- Images

Self Test

- **A. Classification**
  B. Regression
  C. Not Machine Learning

- Dividing up customers by potential profitability?

- Extracting frequency of sound?

Types of Data and Categorization

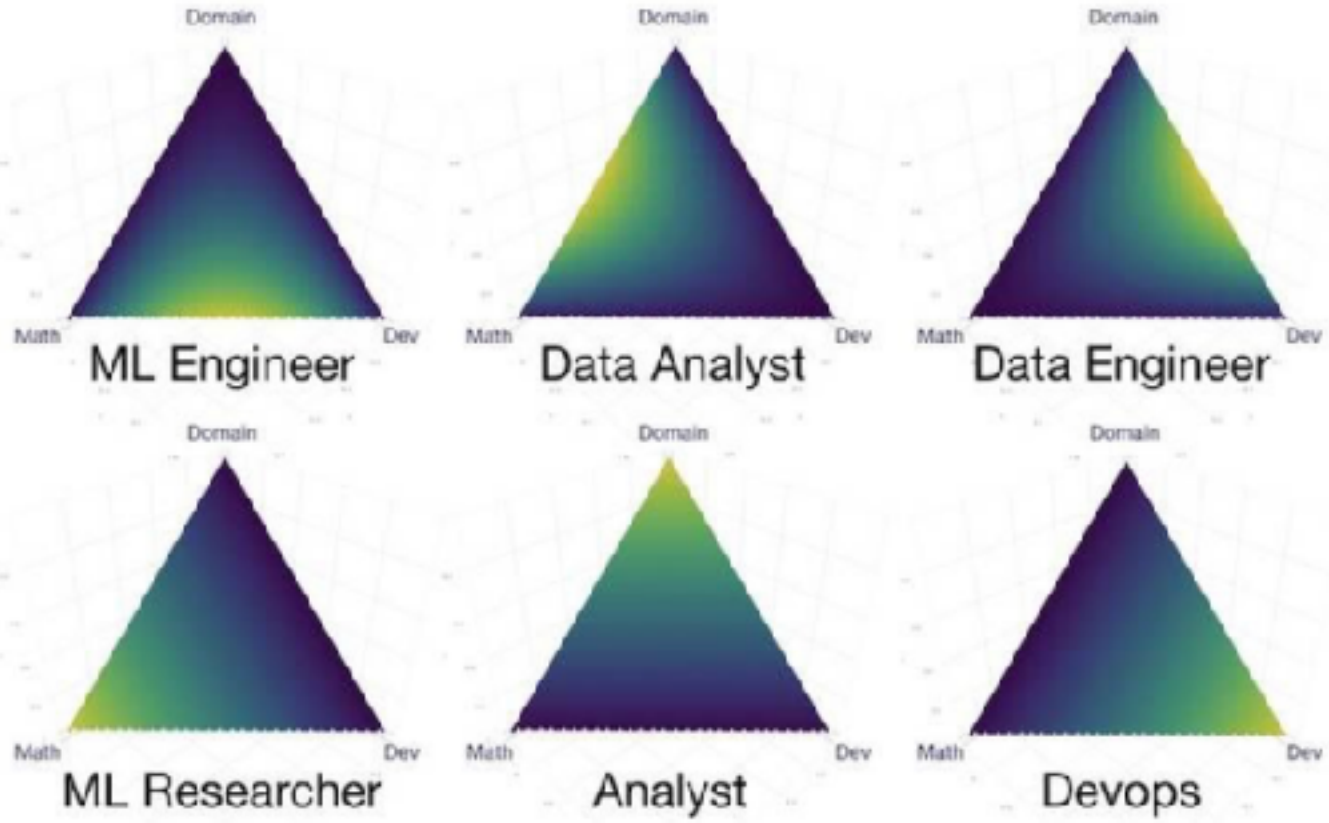


Table Data

- **Table Data**: Collection of data **instances** and their **features**
  - **Python:** Pandas Dataframe
  - **R:** Data.frame
  - **Matlab:** Table Class
  - **C++**: Trick Question

Attributes, columns, variables, fields, characteristics, **Features**

Objects, records, rows, points, **samples**, cases, entities, **instances**

TID	Pregnant	BMI	Age	Diabetes
1	Y	33.6	41-50	positive
2	N	26.6	31-40	negative
3	Y	23.3	31-40	positive
4	N	28.1	21-30	negative
5	N	43.1	31-40	positive
6	Y	25.6	21-30	negative
7	Y	31.0	21-30	positive
8	Y	35.3	21-30	negative
9	N	30.5	51-60	positive
10	Y	37.6	51-60	positive

Feature Type Representation

	Attribute	Representation Transformation	Comments
Discrete	**Nominal**	Permutation of values only. **one hot encoding or hash function**	If all **employee ID** numbers were reassigned, would it make any difference?
Discrete	**Ordinal**	Order must be preserved `new_value = f(old_value)` where *f* is a monotonic function. **integer**	An attribute encompassing the notion of **good, better best** can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Continuous	**Interval**	`new_value =f(old_value) + b` *f* is monotonic through origin **float**	Thus, the **Fahrenheit** and **Celsius** temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Continuous	**Ratio**	`new_value = f(old_value)` *f* is monotonic through origin **float**	**Length** can be measured in meters or feet, but **zero is zero**

from Tan et al. Introduction to Data Mining

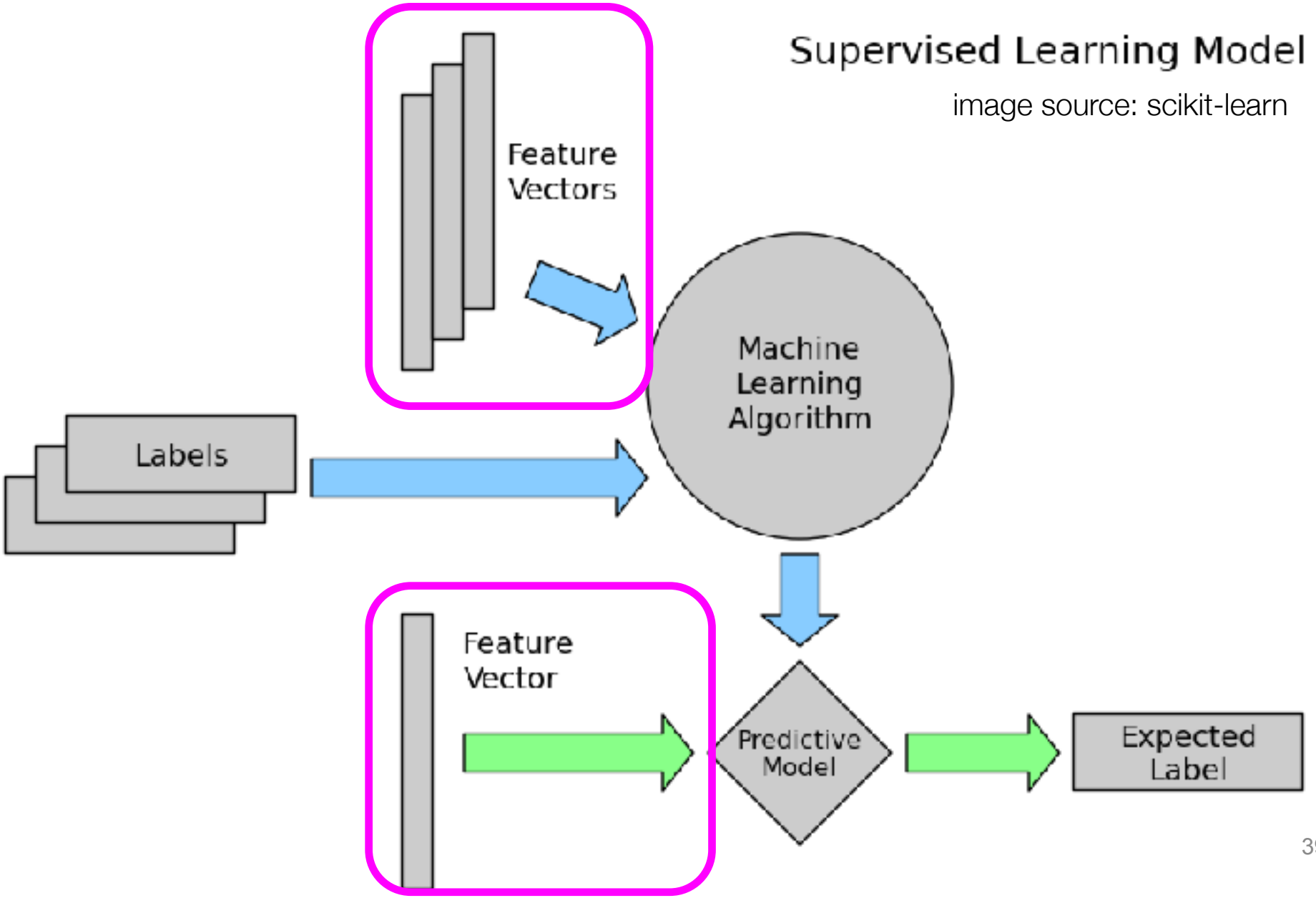
"Finish"
Jupyter Notebooks

`01_Numpy and Pandas Intro.ipynb`

Data Quality

programmers
commenting their code

Review of Feature Data



Data Quality Problems

- Missing
  - Easy to find, NaNs
- Duplicated
  - Easy to find, hard to verify
- Noise or Outlier
  - Hard to define
  - Hard to catch

Information is not collected (e.g., people decline to give their age and weight)

Features **not applicable** (e.g., annual income for children)

UCI ML Repository: 90% of repositories have missing data

TID	*Hair Color*	*Height*	*Age*	*Arrested*
1	Brown	5'2"	23	no
2	Hazel	1.5m	12	no
3	Bl	5	999	no
4	Brown	5'2"	23	no