Machine Learning

Optimizing the Commodity Flow Survey

Andrew Cukierwar

Washington University in St. Louis

Mentor: James Hinckley, Economic Reimbursable Surveys Division

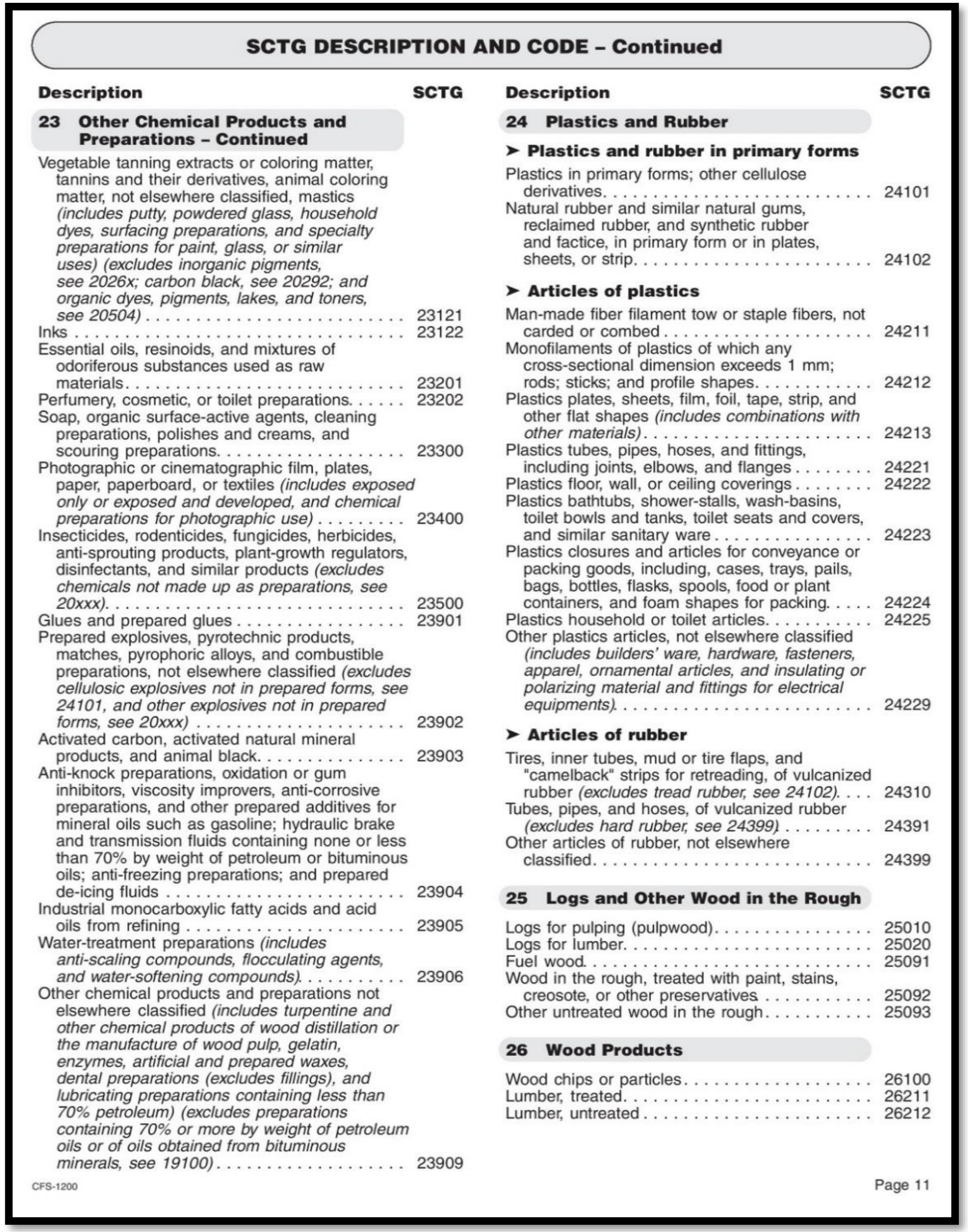
Commodity Flow Survey

- Conducted **every 5 years**
- Co-sponsored by Census and BTS
- Each quarter, select businesses report data on up to 80 of their shipments

SCTG Codes

Standard Classification of Transported Goods

- SCTG code for **each shipment**
- Finding code is **time-intensive**
  - Over 500 SCTG codes
- **Incorrect** SCTG codes
- <u>Ex:</u> 01003 – Live Poultry



SCTG DESCRIPTION AND CODE – Continued

Description	SCTG
23 Other Chemical Products and Preparations – Continued	
Vegetable tanning extracts or coloring matter, tannins and their derivatives, animal coloring matter, not elsewhere classified, mastics *(includes putty, powdered glass, household dyes, surfacing preparations, and specialty preparations for paint, glass, or similar uses) (excludes inorganic pigments, see 2026x; carbon black, see 20292; and organic dyes, pigments, lakes, and toners, see 20504)*	23121
Inks	23122
Essential oils, resinoids, and mixtures of odoriferous substances used as raw materials	23201
Perfumery, cosmetic, or toilet preparations	23202
Soap, organic surface-active agents, cleaning preparations, polishes and creams, and scouring preparations	23300
Photographic or cinematographic film, plates, paper, paperboard, or textiles *(includes exposed only or exposed and developed, and chemical preparations for photographic use)*	23400
Insecticides, rodenticides, fungicides, herbicides, anti-sprouting products, plant-growth regulators, disinfectants, and similar products *(excludes chemicals not made up as preparations, see 20xxx)*	23500
Glues and prepared glues	23901
Prepared explosives, pyrotechnic products, matches, pyrophoric alloys, and combustible preparations, not elsewhere classified *(excludes cellulosic explosives not in prepared forms, see 24101, and other explosives not in prepared forms, see 20xxx)*	23902
Activated carbon, activated natural mineral products, and animal black	23903
Anti-knock preparations, oxidation or gum inhibitors, viscosity improvers, anti-corrosive preparations, and other prepared additives for mineral oils such as gasoline; hydraulic brake and transmission fluids containing none or less than 70% by weight of petroleum or bituminous oils; anti-freezing preparations; and prepared de-icing fluids	23904
Industrial monocarboxylic fatty acids and acid oils from refining	23905
Water-treatment preparations *(includes anti-scaling compounds, flocculating agents, and water-softening compounds)*	23906
Other chemical products and preparations not elsewhere classified *(includes turpentine and other chemical products of wood distillation or the manufacture of wood pulp, gelatin, enzymes, artificial and prepared waxes, dental preparations (excludes fillings), and lubricating preparations containing less than 70% petroleum) (excludes preparations containing 70% or more by weight of petroleum oils or of oils obtained from bituminous minerals, see 19100)*	23909

Description	SCTG
24 Plastics and Rubber	
➤ Plastics and rubber in primary forms	
Plastics in primary forms; other cellulose derivatives	24101
Natural rubber and similar natural gums, reclaimed rubber, and synthetic rubber and factice, in primary form or in plates, sheets, or strip	24102
➤ Articles of plastics	
Man-made fiber filament tow or staple fibers, not carded or combed	24211
Monofilaments of plastics of which any cross-sectional dimension exceeds 1 mm; rods; sticks; and profile shapes	24212
Plastics plates, sheets, film, foil, tape, strip, and other flat shapes *(includes combinations with other materials)*	24213
Plastics tubes, pipes, hoses, and fittings, including joints, elbows, and flanges	24221
Plastics floor, wall, or ceiling coverings	24222
Plastics bathtubs, shower-stalls, wash-basins, toilet bowls and tanks, toilet seats and covers, and similar sanitary ware	24223
Plastics closures and articles for conveyance or packing goods, including, cases, trays, pails, bags, bottles, flasks, spools, food or plant containers, and foam shapes for packing	24224
Plastics household or toilet articles	24225
Other plastics articles, not elsewhere classified *(includes builders' ware, hardware, fasteners, apparel, ornamental articles, and insulating or polarizing material and fittings for electrical equipments)*	24229
➤ Articles of rubber	
Tires, inner tubes, mud or tire flaps, and "camelback" strips for retreading, of vulcanized rubber *(excludes tread rubber, see 24102)*	24310
Tubes, pipes, and hoses, of vulcanized rubber *(excludes hard rubber, see 24399)*	24391
Other articles of rubber, not elsewhere classified	24399
25 Logs and Other Wood in the Rough	
Logs for pulping (pulpwood)	25010
Logs for lumber	25020
Fuel wood	25091
Wood in the rough, treated with paint, stains, creosote, or other preservatives	25092
Other untreated wood in the rough	25093
26 Wood Products	
Wood chips or particles	26100
Lumber, treated	26211
Lumber, untreated	26212

CFS-1200

Page 11

One of 15 Pages in Booklet

Solution

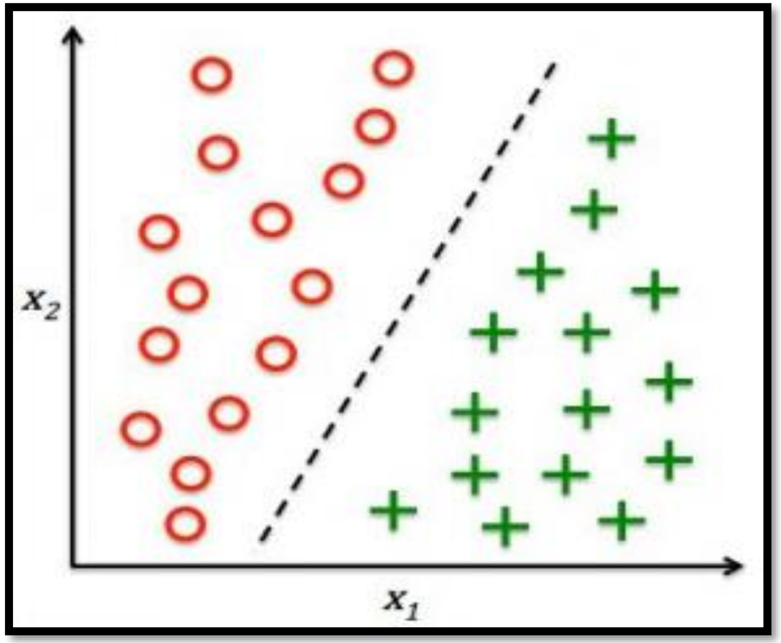
Machine Learning

- Use **machine learning** techniques to **automate** assignment of SCTG codes to shipments
  - Classify shipments into SCTG codes
  - Allows survey responders to **save great amount of time**
- Implemented in Python using the Scikit-Learn library

Classification

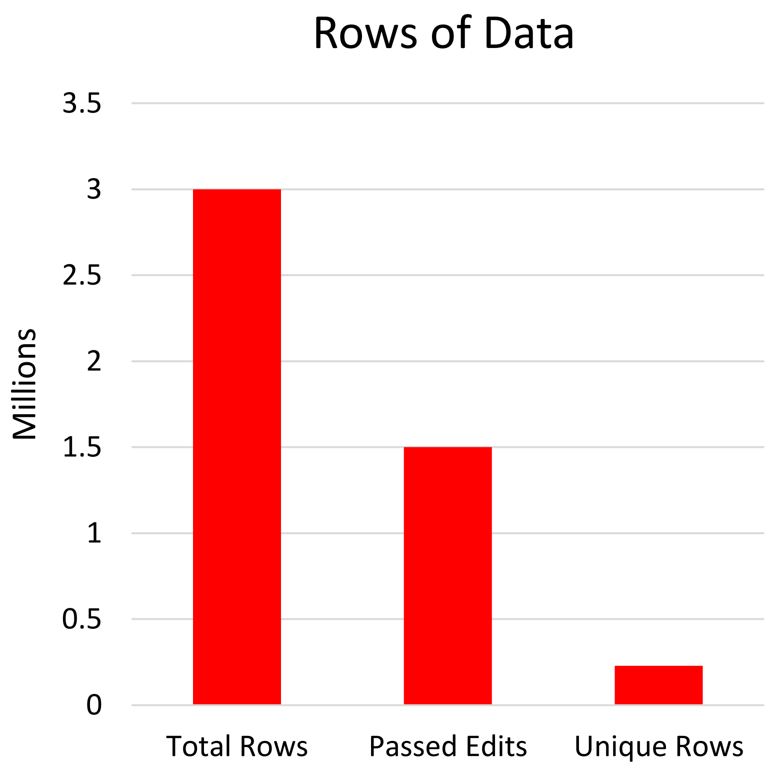
- Use X-variables to identify what class (Y-variable) an item belongs to

	X1	X2	X3	Y
Item 1	5	1	3	O
Item 2	4	5	4	+
Item 3	2	7	6	+
Item 4	2	9	8	O
Item 5	3	7	6	+



Initial Data Preparation

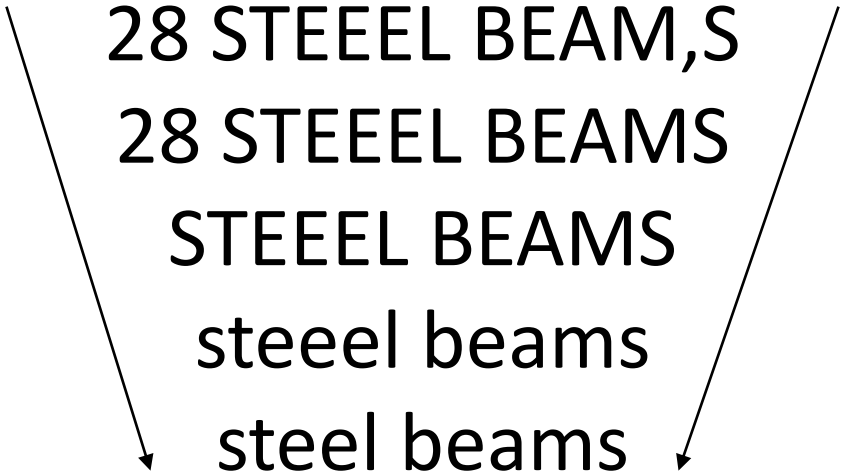
- Initially **3 million** rows of data
- Only **1.5 million** rows pass edits
  - Ex: Shipment is missing SCTG code
- Remove Duplicate Rows
  - Left with **230,000 unique** rows



Text Preprocessing

- Data includes short **text description of shipment**

- Manipulate text to make it **uniform**



28 STEEEL BEAM,S

28 STEEEL BEAMS

STEEEL BEAMS

steeel beams

steel beams

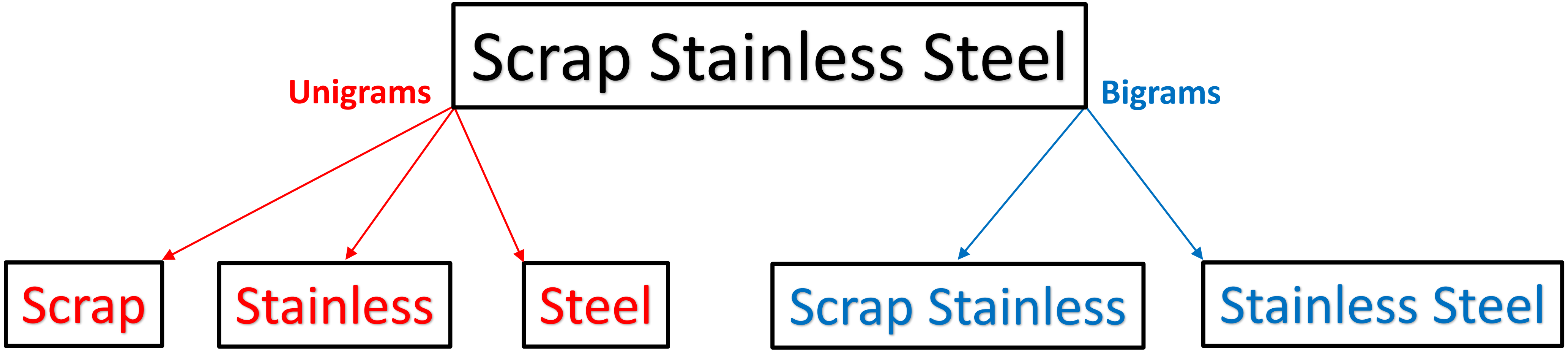
Text to Matrix

- **Must convert text to matrix form**
    - Algorithms only work on **numerical data**

- **TF-IDF** (Term Frequency – Inverse Document Frequency)
    - Bag of Bigrams

Bag of Bigrams

Each description is split into a ‘bag’ of individual words (**unigrams**) and pairs of consecutive words (**bigrams**)



Scrap Stainless Steel

Unigrams

Scrap

Stainless

Steel

Bigrams

Scrap Stainless

Stainless Steel

TF-IDF

- Create matrix of counts of unigrams and bigrams

	Description	steel	beams	pipe	steel beams	steel pipe	steel steel
1	**steel beams**	1	1	0	1	0	0
2	**steel pipe**	1	0	1	0	1	0
3	**steel steel**	2	0	0	0	0	1

- Multiply matrix values by inverse of frequency in full body of text

- Resulting matrix is **large but sparse**
  - 230,000 Rows and 170,000 Cols = **39.1 billion values**
  - Nearly all values are **0's**

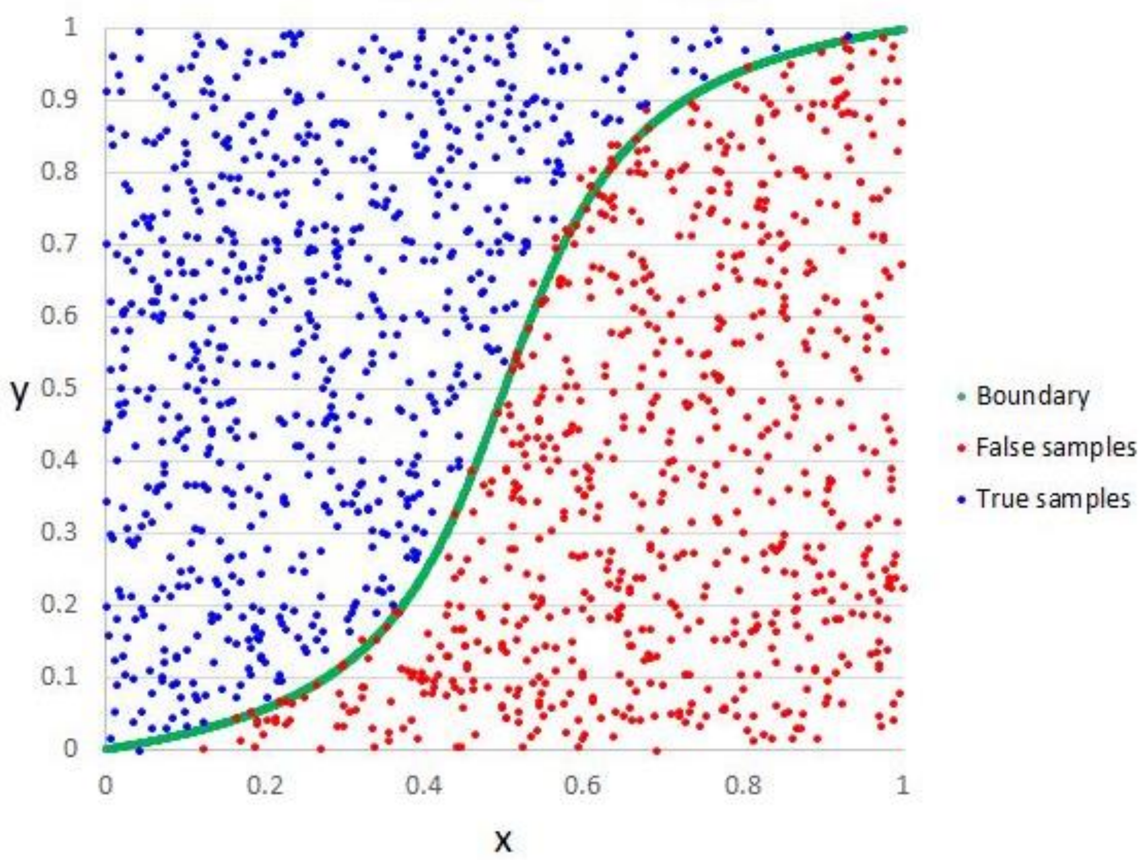
Matrix Processing

- Add two columns of data to the matrix
    - Price per weight
    - Average word length

- **Normalize** all data to be between 0 and 1

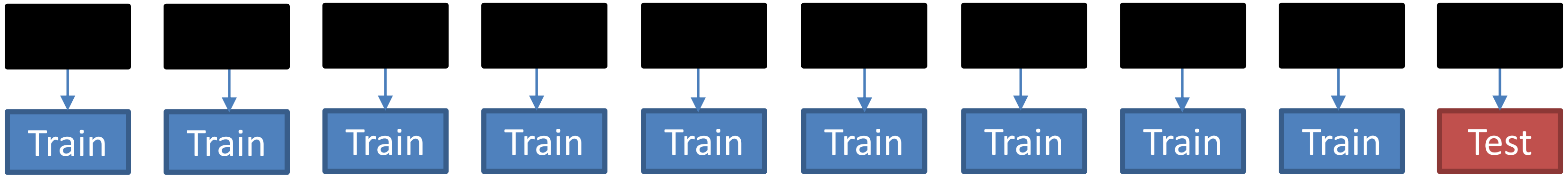
Logistic Regression Classifier

- **Fast** for large and sparse data
- **Highest accuracy** for this data
- Returns **probability** scores
  - Probability of X shipment having SCTG code Y is Z%



Training and Testing

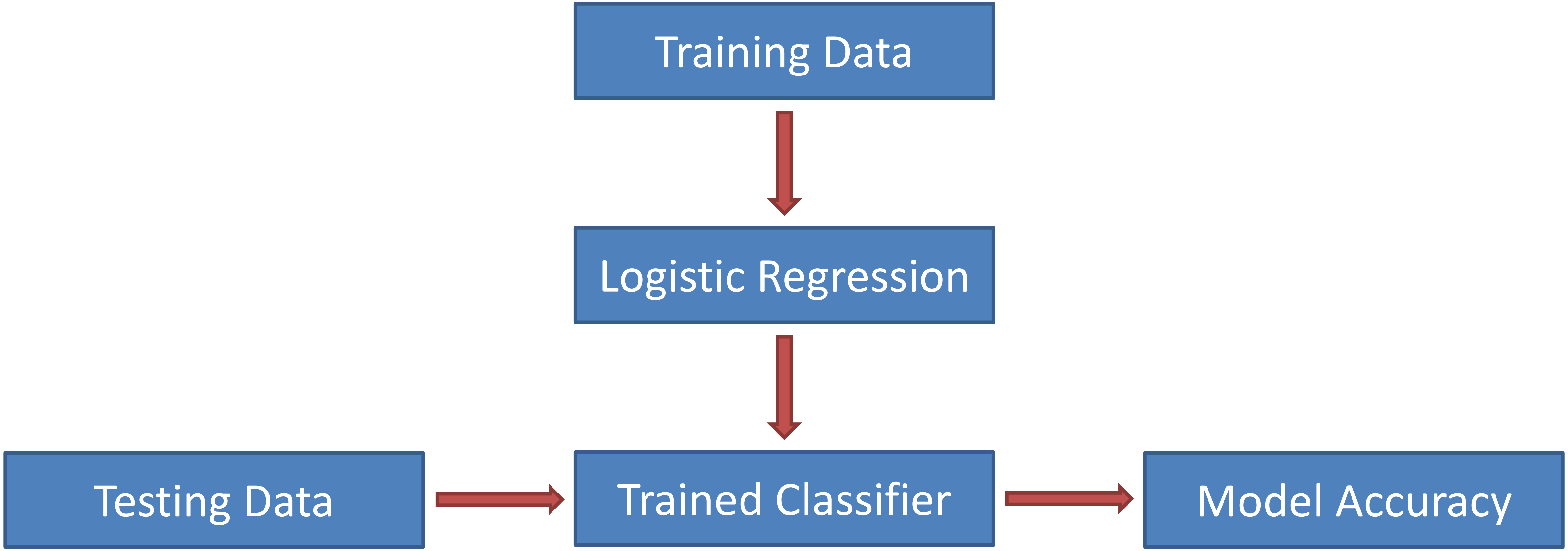
- Split data into training and testing data (90-10 split)



- Use training data to “create” model

- Use testing data to determine accuracy of model

Model Process



Training Data

Logistic Regression

Trained Classifier

Testing Data

Model Accuracy

Predict classes of testing data

Compare predicted classes against actual classes

Results

- Predict **first 2 digits** of SCTG code with **65% accuracy**

- Predict **full SCTG code** with **48% accuracy**

- **Given first 2 digits** of SCTG code, can predict **full SCTG code** with **70% accuracy**

Limitations

- Model is only as good as the data used
  - **Using survey respondent data to train**
  - Short descriptions of shipments (100 char limit)
  - Many classes
  - 43XXX – Mixed Freight
  - Codes for miscellaneous products

Future Improvements

- Improve quality of data
  - Train model only using label-verified shipments
  - Ensure spelling of descriptions
  - Increase shipment description length
  - Train model with more data

- Use more advanced algorithms (deep learning)