DECISION PATTERNS

01-Introduction

Practical Machine Learning (with R)

UC Berkeley

Fall 2016

Topics

- Introduction
- Tools and Environment
- Exercise
- Introductions (continued)
- Data Science, Machine Learning and Opportunities
- Machine Learning

INTRODUCTIONS

Me (Personally)

⚠ Shameless Plug

My Skills

- R /Python Programmer (>15 years)
- Machine Learning (>15 years)
- DevOps
- Researcher and Writer : Machine Learning, Clinical Medicine, Chemistry, Finance

Education

- UC Berkeley → (UT Austin)→ UC Santa Barbara → UC Berkeley
- Post-graduate: UC Berkeley, Stanford

Professional Experience

- Lawrence Berkeley National Lab, Allianz, Open Data
- Sept. 2010 Founded Decision Patterns

Professional Interests

- Machine Learning / Statistics
- High Performance Computing
- Applied Statistics and Visualization
- Management of Data Organizations

(Decision Patterns)

⚠ Shameless Plug

Decision Patterns

- Founded 2010
- Bring together complementary skills for data strategy:

Acquisition → Organization → Storage Access → Utilization

- Our Model
  - Service Consulting
  - Not a start-up -- no VC funding
  - Use consulting margins from to niche products

- Our Customers
  - Financial Services, Retail, Entertainment, Food, Communications, Defense, Environmental Sciences

What do I like *most* about what I do?

We get to work on a

- variety of problems,
- with a variety of technologies
- in a variety of fields

What do I like *least* about what I do?

We have to work on a

- variety of problems,
- with a variety of technologies
- in a variety of fields

TOOLS AND ENVIRONMENT

Exercise: Set-Up Tools And Environment

- Install **R** → **CRAN**
- Install **R Studio Desktop™** (IDE)
- Install **git**

- Create **github** account
  - Send name, student id, github id to:
    christopher.brown@berkeley.edu

GIT

Git / GITHUB / Source Tree Workflow

What is it?

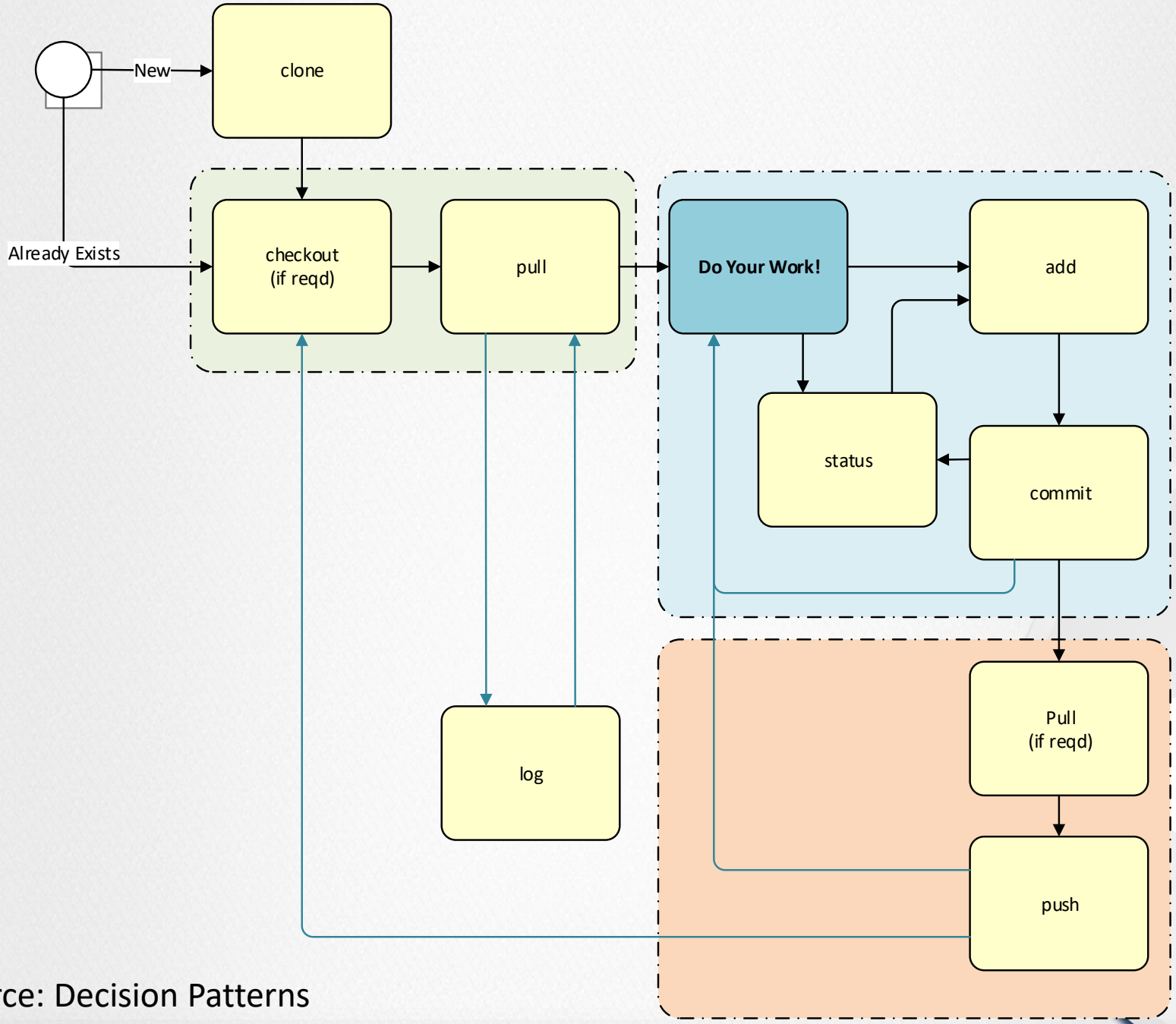
A source control tool (and ***process***) to promote collaborative development

Interfaces

- Github
- Source Tree™
- *R Studio*
- *command-line*

Features

- Distributed
- Each clone contains complete history
- Ability to return to
- Branch in and merging



source: Decision Patterns

Git Commands

- **Repo(sitory)**: location where files are stored. If different from original source: "fork"
- **Branch**: Copy of code that can be independently worked on.

- checkout: Change to specific branch/commit.

- **add**: Tell which files to "stage" (accept) commit. Done on a per-file basis.
- **commit:** accept changes.

- **pull**: Retrieve changes from remote repository
- **push**: Send committed change to remote repository

- **log**: review history of commits
- **status**: review "staged" status

EXERCISES

- Create github account
- Send account log-in, student id to christopher.brown@berkeley.edu

- Clone class repository

```
git clone https://github.com/csx460
```

You?
DISCUSSION OF INDIVIDUAL GOALS ?

You

- Occupation (student/professional)? Employer?
- Experience Using R?
  - None
  - < 1 year
  - < 3 years
  - > 5 years
- How many use R as your principal data.science tool?
- How many use/have used?
  - Python
  - SAS or SPSS
  - Matlab
  - Statistica
  - Spark/Scala
  - Java
  - C/C++
- Ever spend too much time debating which technology fits?

Class / Objectives

Theory	Practice
Distinguish fundamental aspects of machine learning algorithms	Frame problems to make the suitable for solution via machine learning
Build (train) machine learning models	Collaborate in a group using tools for collaborative/social programming
Evaluate (score) machine learning models	Deploy machine learning models to operations
	Generate high quality, graphical and textual results

Kuhn, Max and Johnson, Kjell
Springer Science+Business
2013

Text

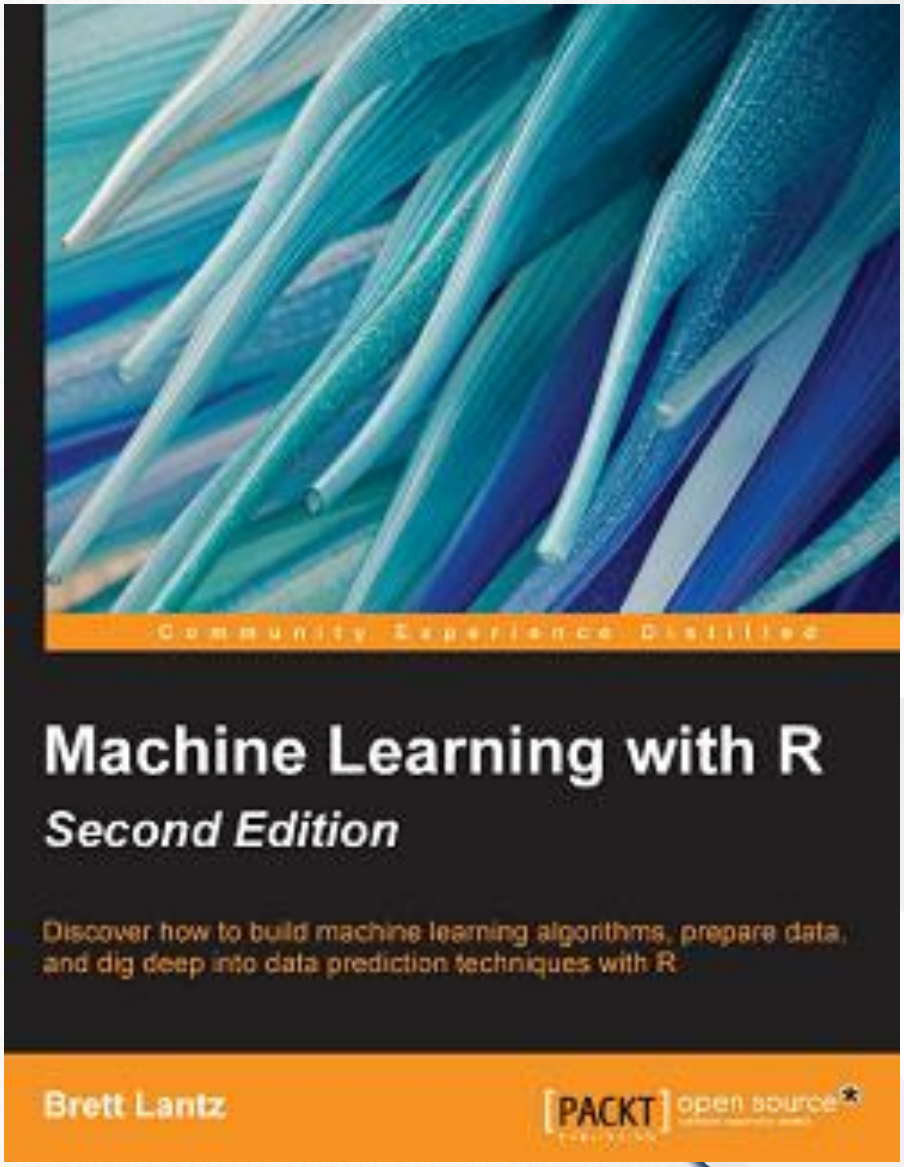
Machine Learning with R, 2nd Edition

ISBN: 978-1-78439-390-8

Lantz, Brett

Packt Publishing

2015



Recommended Text

Applied Predictive Modeling

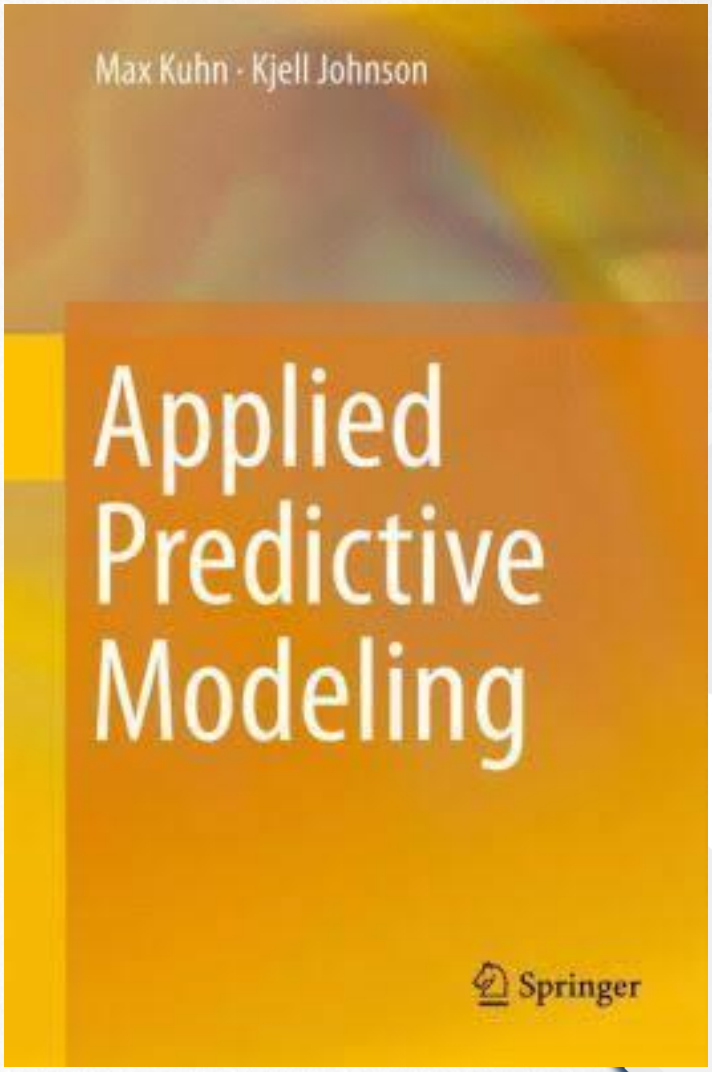
ISBN-13: 978-1461468486

ISBN-10: 1461468485

Kuhn, Max and Johnson, Kjell

Springer Science+Business

2013



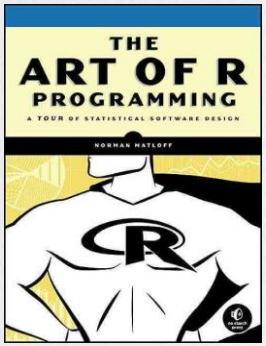
Additional Resources

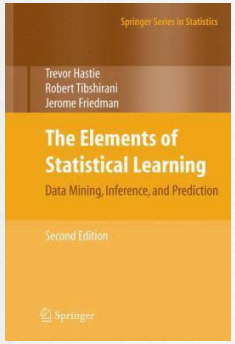
Texts
(not used in this class)

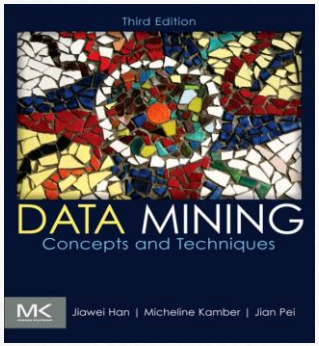
The Art of R Programming
by Norm Matloff

Elements of Statistical Learning
by Hastie, Friedman, Tibshirani

Data Mining Concepts and Techniques
by Han, Kamber, Pei







Online

- CRAN
  - Packages
  - Task Views
- Metacran (r-pkg.org)
- Stackoverflow.com
- r-bloggers.com
- H. Wickham Online Resources:
  - Advanced R Programming
  - R for Data Science
- Github

Contacts / Coordinates

- Christopher Brown
  christopher.brown@berkeley.edu
  checked once / day (mornings)
  phone # (provided in class)

- Class Website
  - https://github.com/CSX460

- Class group (invitation only)
  - https://groups.google.com/forum/#!forum/csx460

Grading

- **~10 Weekly Exercises (80%)**
  - Exercises are **Rmarkdown** in the github
  - Due at the beginning of class each week
  - Submitted via **github** commits
    - Please email me your **github** login
    - ! Github commits are timestamped
  - Answers reviewed in class
  - Work on them in class, time-permitting
- **Class Participation (20%)**
- **Attendance is Mandatory**
  - **no unexcused absences.**

** PARTICIPATE **

RMARKDOWN/KNITR

RMARKDOWN

What is it?

- Simple text mark-up syntax
- that supports the markdown standard
- And allows incorporation of R analysis and graphical output

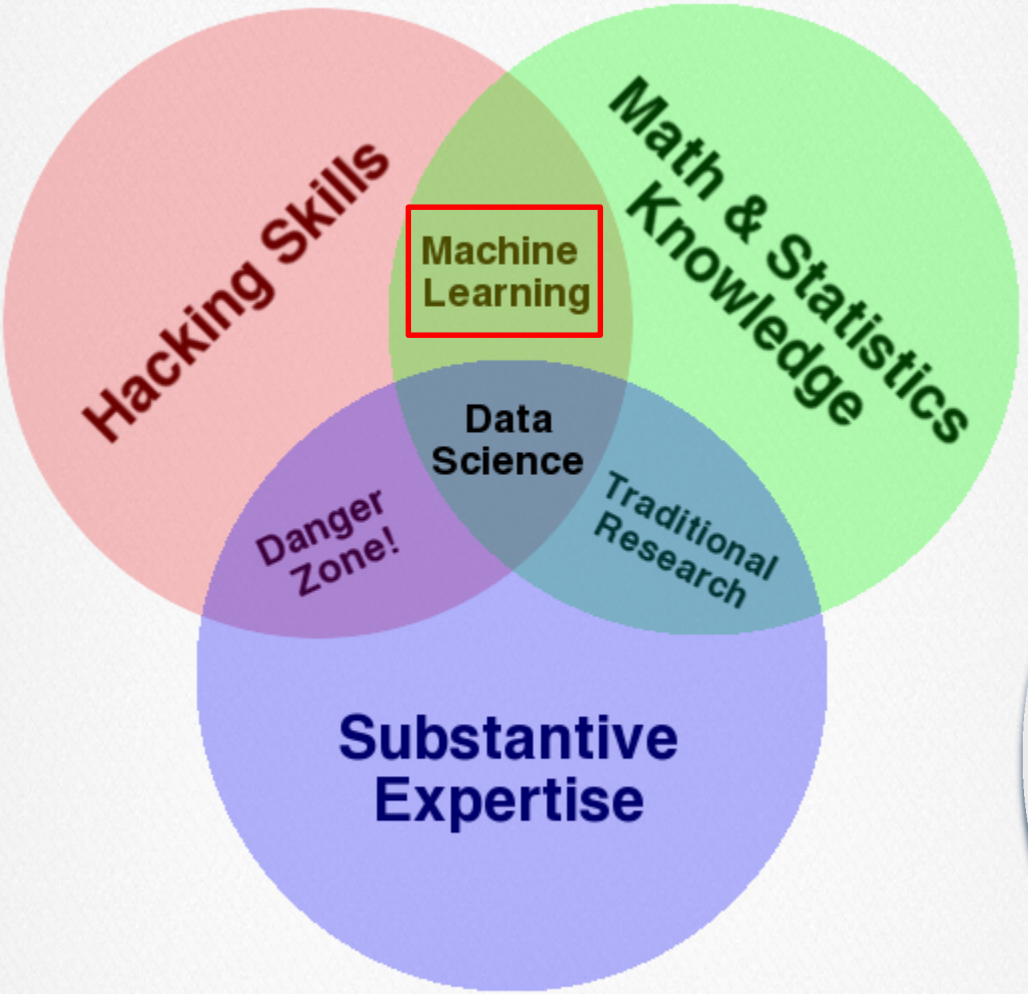
- Are assignments will be done in Markdown …
- Simply put your answers in the space provided
- → Demonstration

ML OVERVIEW

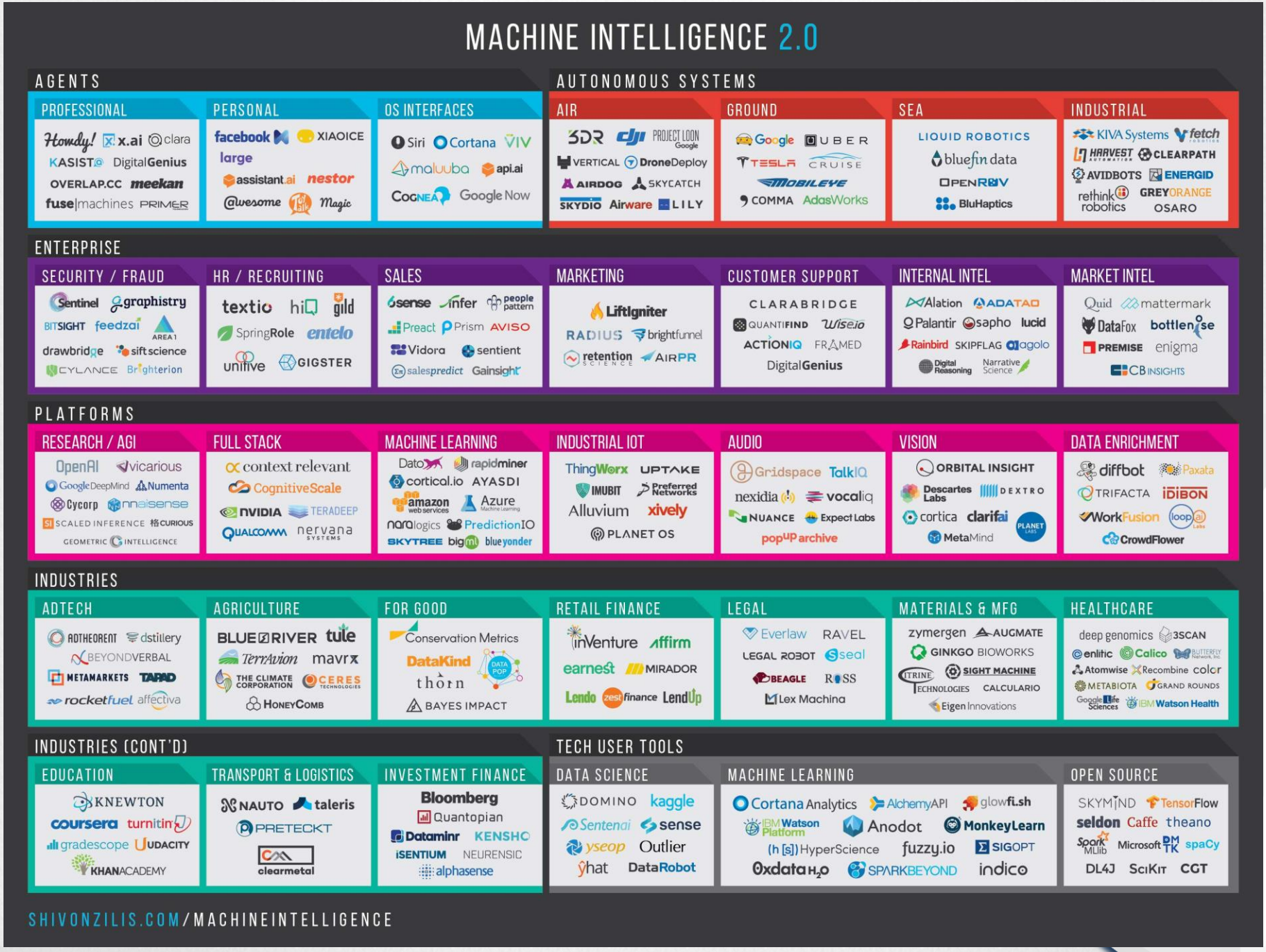
Example of ML algorithm(s)

- Spam Filter
- handwriting recognition (svm)
- Traffic engineering (lights)
- Weather prediction
- Sentiment analysis (social media)
- Netflix Recommender
- Fraud detection (Visa)
- Imaging processing
- Intrution detection
- Self-driving cars

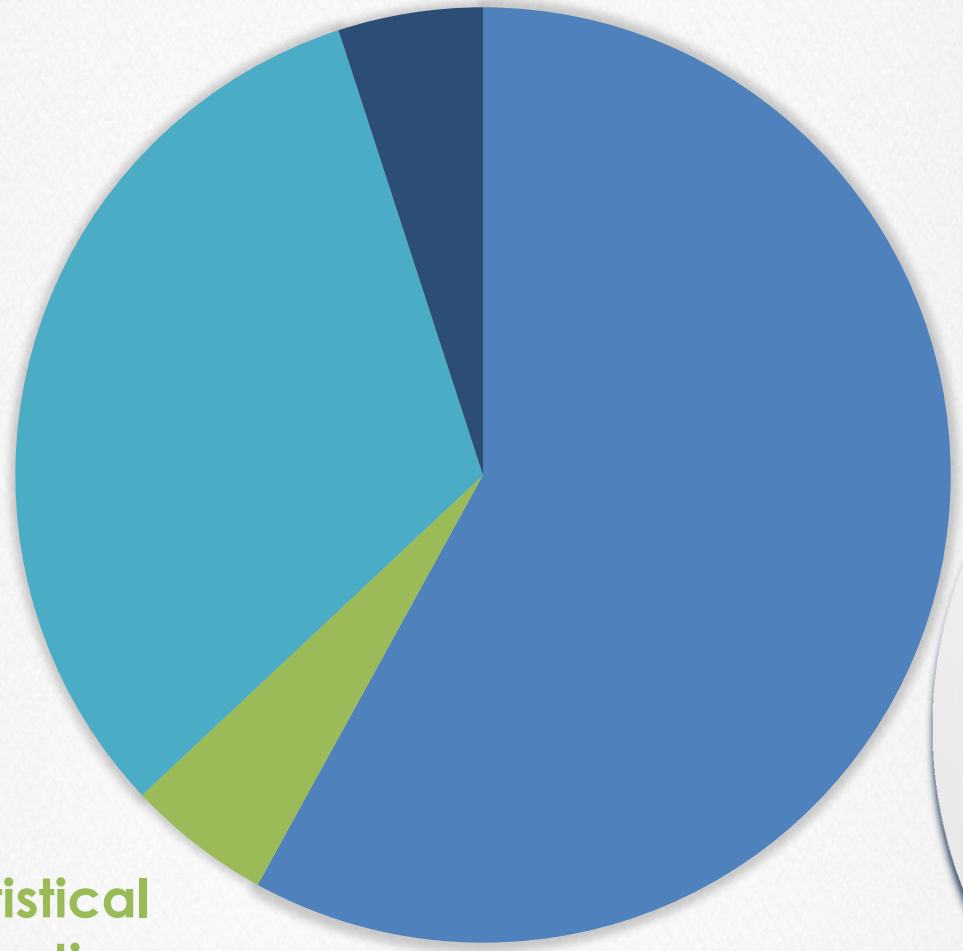
Data Science Venn Diagram



Ref. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram


MACHINE INTELLIGENCE 2.0
AGENTS
PROFESSIONAL
Howdy! x.ai clara
KASIST DigitalGenius
OVERLAP.CC meekan
fuse|machines PRIMER
PERSONAL
facebook XIAOICE
large
assistant.ai nestor
@wesome Magic
OS INTERFACES
Siri Cortana VIV
maluuba api.ai
CogNEA Google Now
AUTONOMOUS SYSTEMS
AIR
3DR dji PROJECT LOON Google
VERTICAL DroneDeploy
AIRDOG SKYCATCH
SKYDIO Airware LILY
GROUND
Google UBER
TESLA CRUISE
MOBILEYE
COMMA AdasWorks
SEA
LIQUID ROBOTICS
bluefin data
OPENROV
BluHaptics
INDUSTRIAL
KIVA Systems fetch
HARVEST CLEARPATH
AVIDBOTS ENERGID
rethink robotics GREYORANGE
OSARO
ENTERPRISE
SECURITY / FRAUD
Sentinel graphistry
BITSIGHT feedzai AREA 1
drawbridge sift science
CYLANCE Brighterion
HR / RECRUITING
textio hiQ gild
SpringRole entelo
unitive GIGSTER
SALES
6sense infer people pattern
Preact Prism AVISO
Vidora sentient
salespredict Gainsight
MARKETING
LiftIgniter
RADIUS brightfunnel
retention SCIENCE AIRPR
CUSTOMER SUPPORT
CLARABRIDGE
QUANTIFIND Wise.io
ACTIONIQ FRAMED
DigitalGenius
INTERNAL INTEL
Alation ADATAO
Palantir sapho lucid
Rainbird SKIPFLAG algolo
Digital Reasoning Narrative Science
MARKET INTEL
Quid mattermark
DataFox bottlenose
PREMISE enigma
CB INSIGHTS
PLATFORMS
RESEARCH / AGI
OpenAI vicarious
Google DeepMind Numenta
Cycorp nnaisense
SCALED INFERENCE CURIOUS
GEOMETRIC INTELLIGENCE
FULL STACK
context relevant
CognitiveScale
NVIDIA TERADEEP
QUALCOMM nervana SYSTEMS
MACHINE LEARNING
Dato rapidminer
cortical.io AYASDI
amazon web services Azure Machine Learning
narrologics PredictionIO
SKYTREE bigml blueyonder
INDUSTRIAL IOT
ThingWorx UPTAKE
IMUBIT Preferred Networks
Alluvium xively
PLANET OS
AUDIO
Gridspace TalkIQ
nexidia vocaliq
NUANCE Expect Labs
popup archive
VISION
ORBITAL INSIGHT
Descartes Labs DEXTRO
cortica clarifai PLANET LABS
MetaMind
DATA ENRICHMENT
diffbot Paxata
TRIFACTA iDIBON
WorkFusion loop.ai Labs
CrowdFlower
INDUSTRIES
ADTECH
ADTHEORENT dstillery
BEYONDVERBAL
METAMARKETS TAPAD
rocketfuel affectiva
AGRICULTURE
BLUE RIVER tule
TerrAvion mavrx
THE CLIMATE CORPORATION CERES TECHNOLOGIES
HONEYCOMB
FOR GOOD
Conservation Metrics
DataKind DATA POP
thorn
BAYES IMPACT
RETAIL FINANCE
inVenture affirm
earnest MIRADOR
Lendo zestfinance LendUp
LEGAL
Everlaw RAVEL
LEGAL ROBOT seal
BEAGLE ROSS
Lex Machina
MATERIALS & MFG
zymergen AUGMATE
GINKGO BIOWORKS
ITRINE TECHNOLOGIES SIGHT MACHINE CALCULARIO
Eigen Innovations
HEALTHCARE
deep genomics 3SCAN
enlitic Calico
Atomwise Recombine color
METABIOTA GRAND ROUNDS
Google Life Sciences IBM Watson Health
INDUSTRIES (CONT'D)
EDUCATION
KNEWTON
coursera turnitin
gradescope UDACITY
KHANACADEMY
TRANSPORT & LOGISTICS
NAUTO taleris
PRETECKT
clearmetal
INVESTMENT FINANCE
Bloomberg
Quantopian
Dataminr KENSHO
iSENTIUM NEURENSIC
alphasense
TECH USER TOOLS
DATA SCIENCE
DOMINO kaggle
Sentenai sense
yseop Outlier
ŷhat DataRobot
MACHINE LEARNING
Cortana Analytics AlchemyAPI glowfi.sh
IBM Watson Platform Anodot MonkeyLearn
(h [s]) HyperScience fuzzy.io SIGOPT
0xdata H2O SPARKBEYOND indico
OPEN SOURCE
SKYMIND TensorFlow
seldon Caffe theano
Spark MLlib Microsoft DMTK spaCy
DL4J SciKit CGT
SHIVONZILIS.COM/MACHINEINTELLIGENCE

Breakdown of Code Tasks



ELITE CODING

USEFUL R PACKAGES

```
> install.packages("package-name")
```

- **Tables**: *data.tables, tibble*
- **Data Manipulation** (tidyverse)
  - **Pipe operators**: *magrittr (pipeR, backpipe)* (shiny)
  - **Tables:** dplyr, tidyr
  - **Read/Writing**: readr, readxl, foreign
- **Visualization**: *ggplot2, ggvis*
- **Reporting**: *rmarkdown/knitr, shiny*
- ***ML Framework:*** *caret* (Classification and Regression Training)

EXERCISES IN CLASS

QUESTION 1

What is machine learning?

A formal ***process*** for building a ***model***

QUESTION 2

What is a model?

a ***function*** that ***estimates*** a ***response*** associated with (a set of) known ***predictors***

$$\hat{y} = f(\vec{x})$$

Question 3: What are the Properties of f

- Should be easy* to evaluate
- Takes a one or more values of inputs
- Yields a single output value for each input

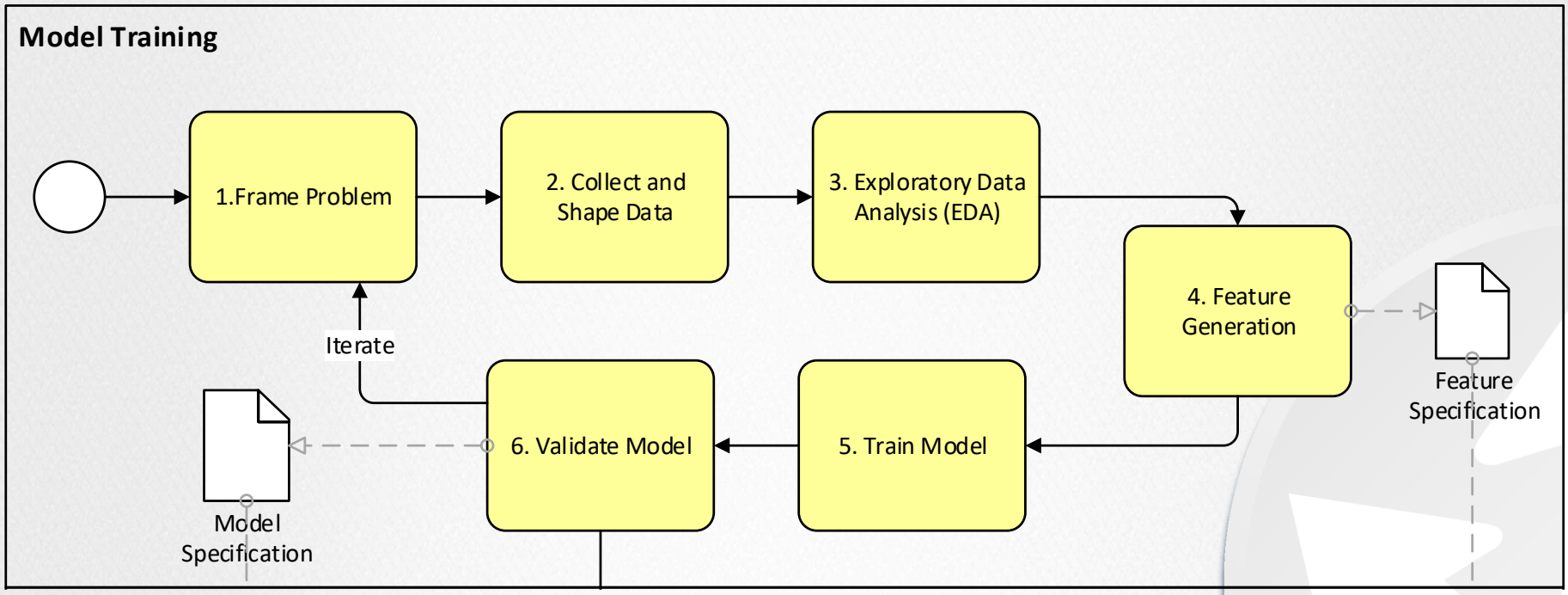
- Output, $\hat{\mathbf{y}}$, should be "close to" observed values, $\mathbf{y}$:

$$\hat{\mathbf{y}} \sim \mathbf{y}$$

* Computational cheap/efficient

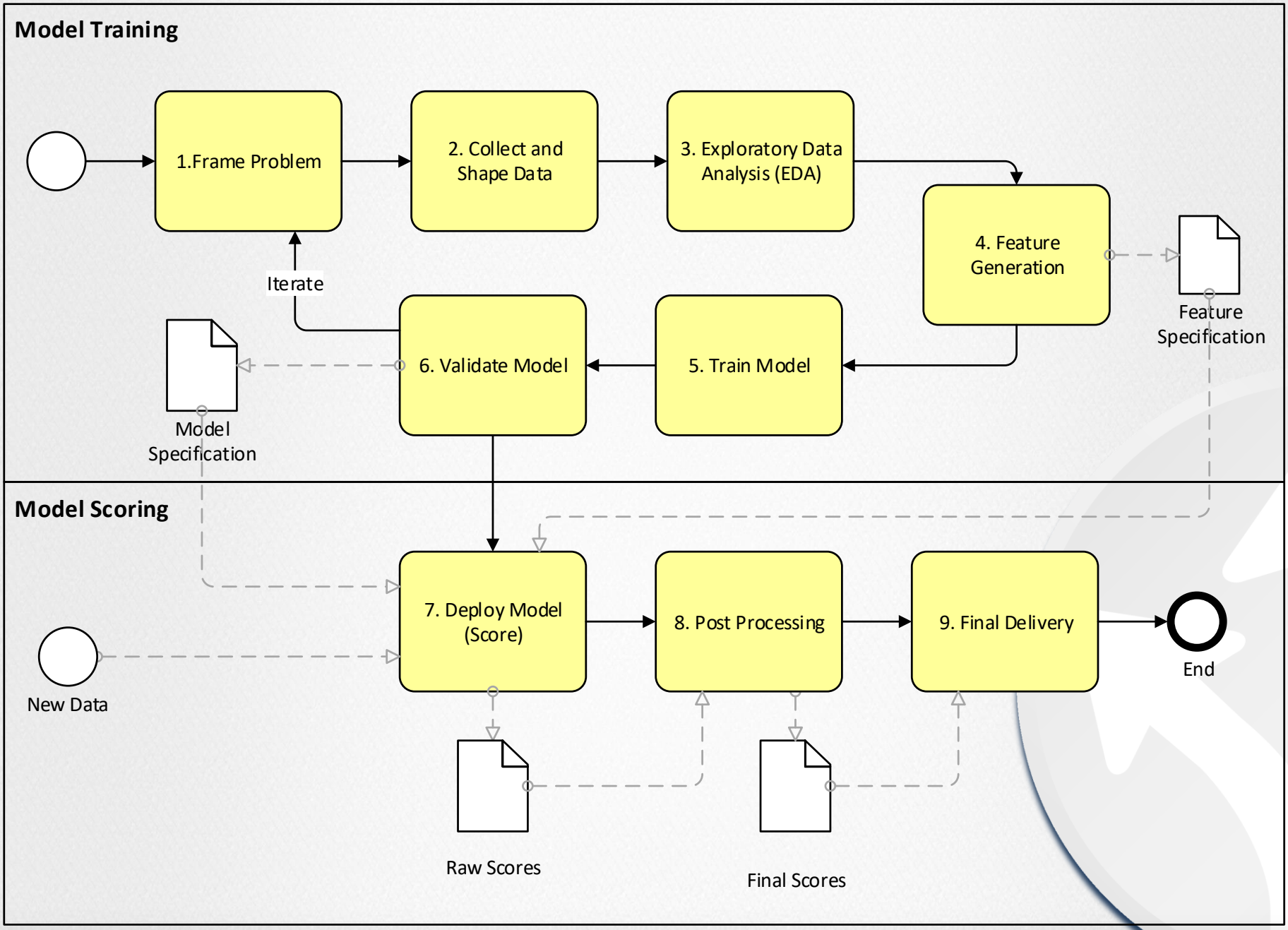
QUESTION 4

How do we find f?


Model Training
1.Frame Problem
2. Collect and Shape Data
3. Exploratory Data Analysis (EDA)
4. Feature Generation
Feature Specification
5. Train Model
6. Validate Model
Iterate
Model Specification

QUESTION 4

How do we use f?


Model Training
1.Frame Problem
2. Collect and Shape Data
3. Exploratory Data Analysis (EDA)
4. Feature Generation
Feature Specification
Iterate
6. Validate Model
5. Train Model
Model Specification
Model Scoring
7. Deploy Model (Score)
8. Post Processing
9. Final Delivery
End
New Data
Raw Scores
Final Scores