NYU Center for Data Science

DS-GA 1003
Machine Learning

Week 1: Lecture 1
Model Selection – Clustering and Regression

NYU Center for Data Science

How can machine learning help us make guesses or detect patterns?

DS-GA 1003 Machine Learning

Week 1: Lecture 1

Model Selection – Clustering and Regression

Adapted from Rosenberg, Rudin, Rangan, Steinhardt

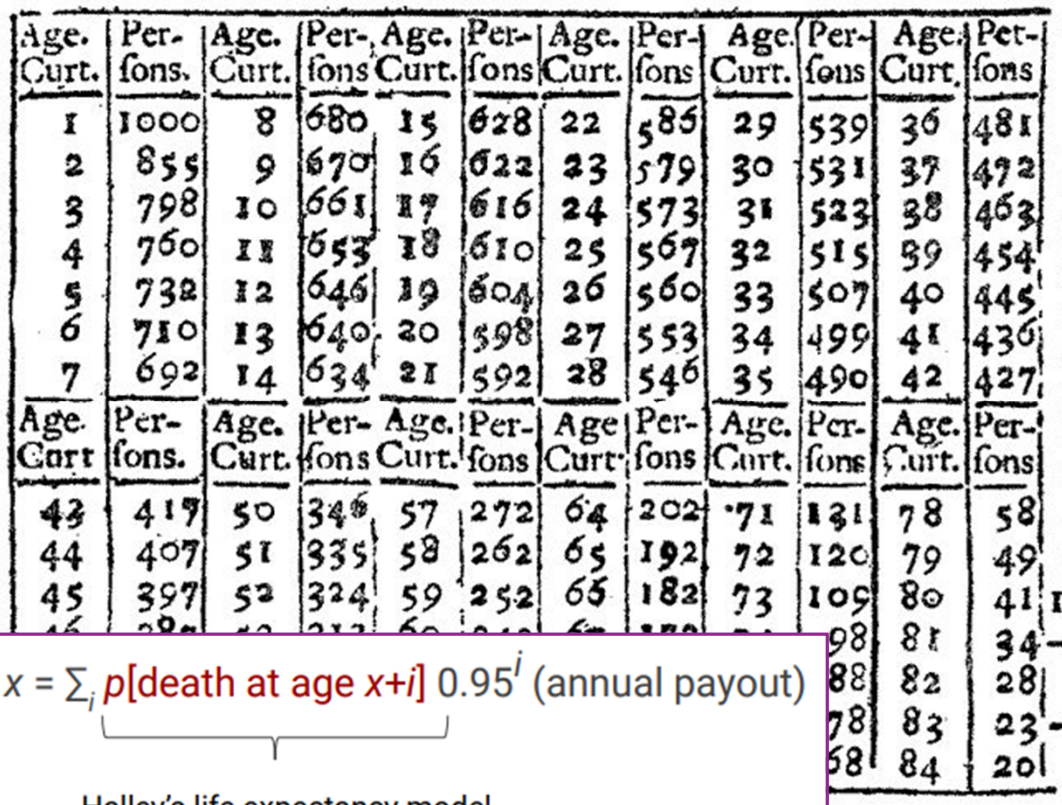
What is machine learning?

Age. Curt.	Per- sons.	Age. Curt.	Per- sons	Age. Curt.	Per- sons	Age. Curt.	Per- sons	Age. Curt.	Per- sons	Age. Curt.	Pet- sons
1	1000	8	680	15	628	22	585	29	539	36	481
2	855	9	670	16	622	23	579	30	531	37	472
3	798	10	661	17	616	24	573	31	523	38	463
4	760	11	653	18	610	25	567	32	515	39	454
5	732	12	646	19	604	26	560	33	507	40	445
6	710	13	640	20	598	27	553	34	499	41	436
7	692	14	634	21	592	28	546	35	490	42	427
Age. Curt	Per- sons.	Age. Curt.	Per- sons	Age. Curt.	Per- sons	Age Curt	Per- sons	Age. Curt.	Per- sons	Age. Curt.	Per- sons
43	417	50	346	57	272	64	202	71	131	78	58
44	407	51	335	58	262	65	192	72	120	79	49
45	397	52	324	59	252	66	182	73	109	80	41
46	387	53	313	60	242	67	172	74	98	81	34
47	377	54	302	61	232	68	162	75	88	82	28
48	367	55	292	62	222	69	152	76	78	83	23
49	357	56	282	63	212	70	142	77	68	84	20

Age.	Persons.
7	
14	
21	
28	
35	
42	3178
49	2709
56	2194
63	1694
70	1204
77	692
84	253
100	107
	34000
	Sum Total.

Based on data collected between 1687 and 1691 in Wroclaw, Poland

What is machine learning?



Age. Curt.	Per- sons.	Age. Curt.	Per- sons	Age. Curt.	Per- sons	Age. Curt.	Per- sons	Age. Curt.	Per- sons	Age. Curt.	Pet- sons
1	1000	8	680	15	628	22	585	29	539	36	481
2	855	9	670	16	622	23	579	30	531	37	472
3	798	10	661	17	616	24	573	31	523	38	463
4	760	11	653	18	610	25	567	32	515	39	454
5	732	12	646	19	604	26	560	33	507	40	445
6	710	13	640	20	598	27	553	34	499	41	436
7	692	14	634	21	592	28	546	35	490	42	427
Age. Curt	Per- sons.	Age. Curt.	Per- sons	Age. Curt.	Per- sons	Age Curt	Per- sons	Age. Curt.	Per- sons	Age. Curt.	Per- sons
43	417	50	346	57	272	64	202	71	131	78	58
44	407	51	335	58	262	65	192	72	120	79	49
45	397	52	324	59	252	65	182	73	109	80	41
									98	81	34
									88	82	28
									78	83	23
									68	84	20

Age.	Persons.
7	
14	
21	
28	
35	
42	3178
49	2709
56	2194
63	1694
70	1204
77	692
84	253
100	107
	34000
	Sum Total.

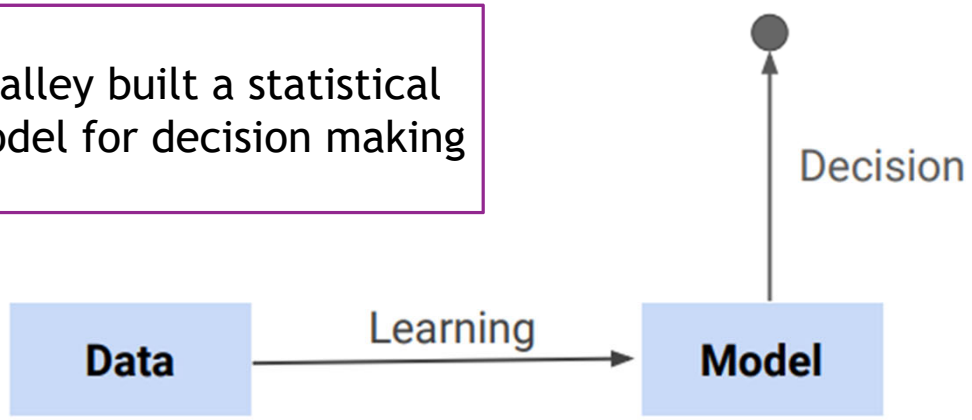
Based on data collected between 1687 and 1691 in Wroclaw, Poland

Price at age $x = \sum_i p[\text{death at age } x+i]\ 0.95^i$ (annual payout)

Halley's life expectancy model

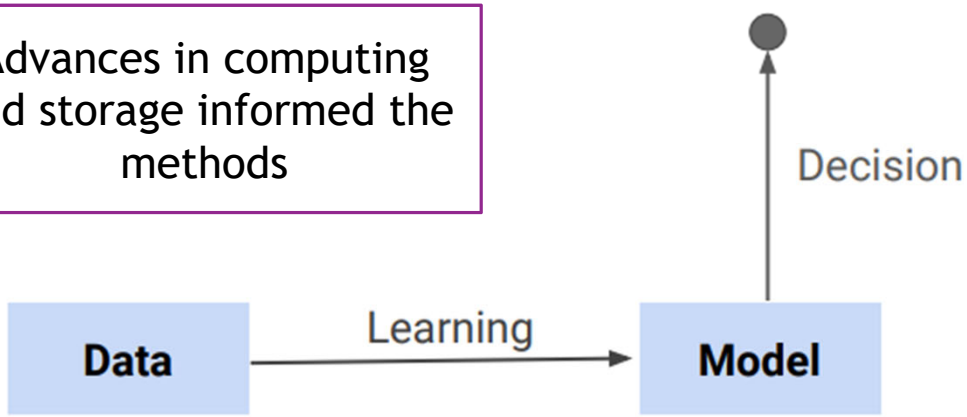
What is machine learning?

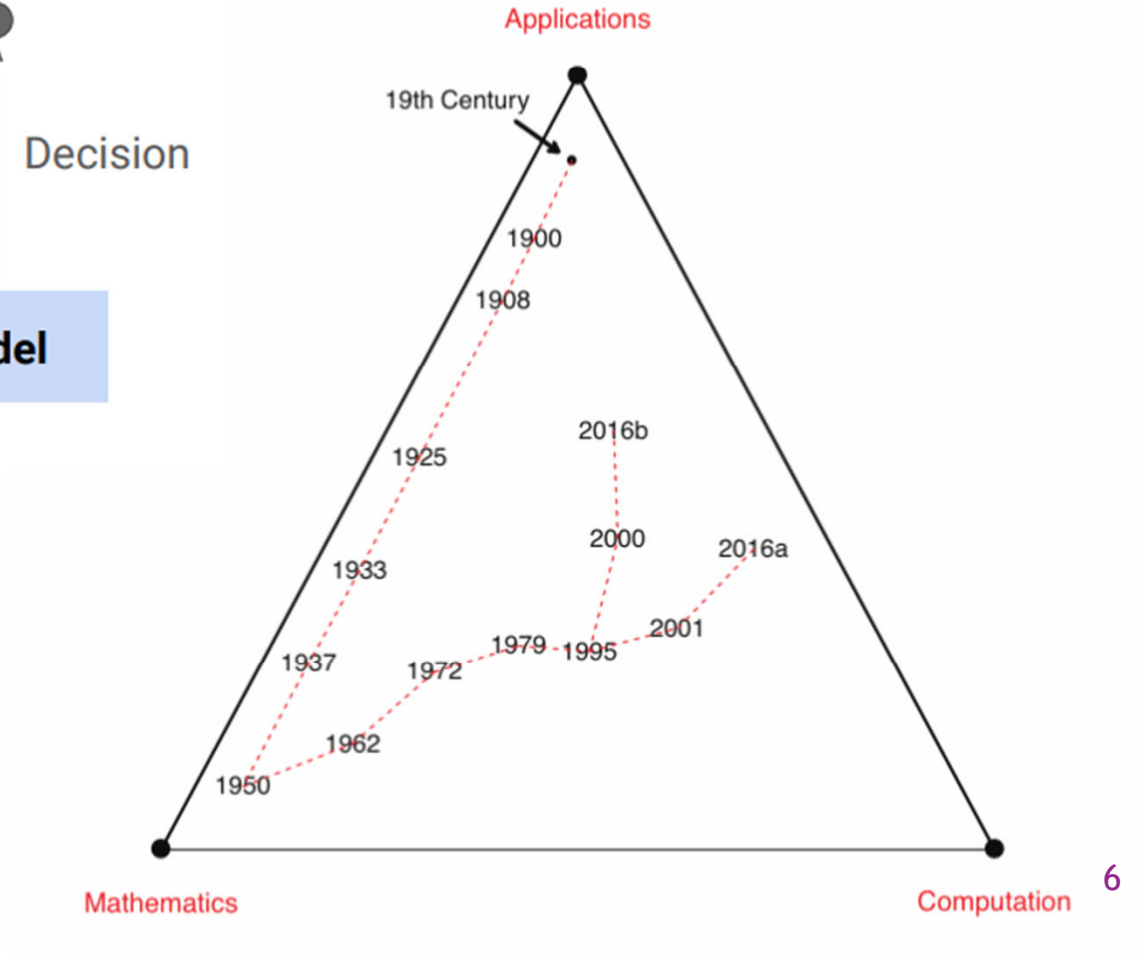
Halley built a statistical model for decision making



What is machine learning?

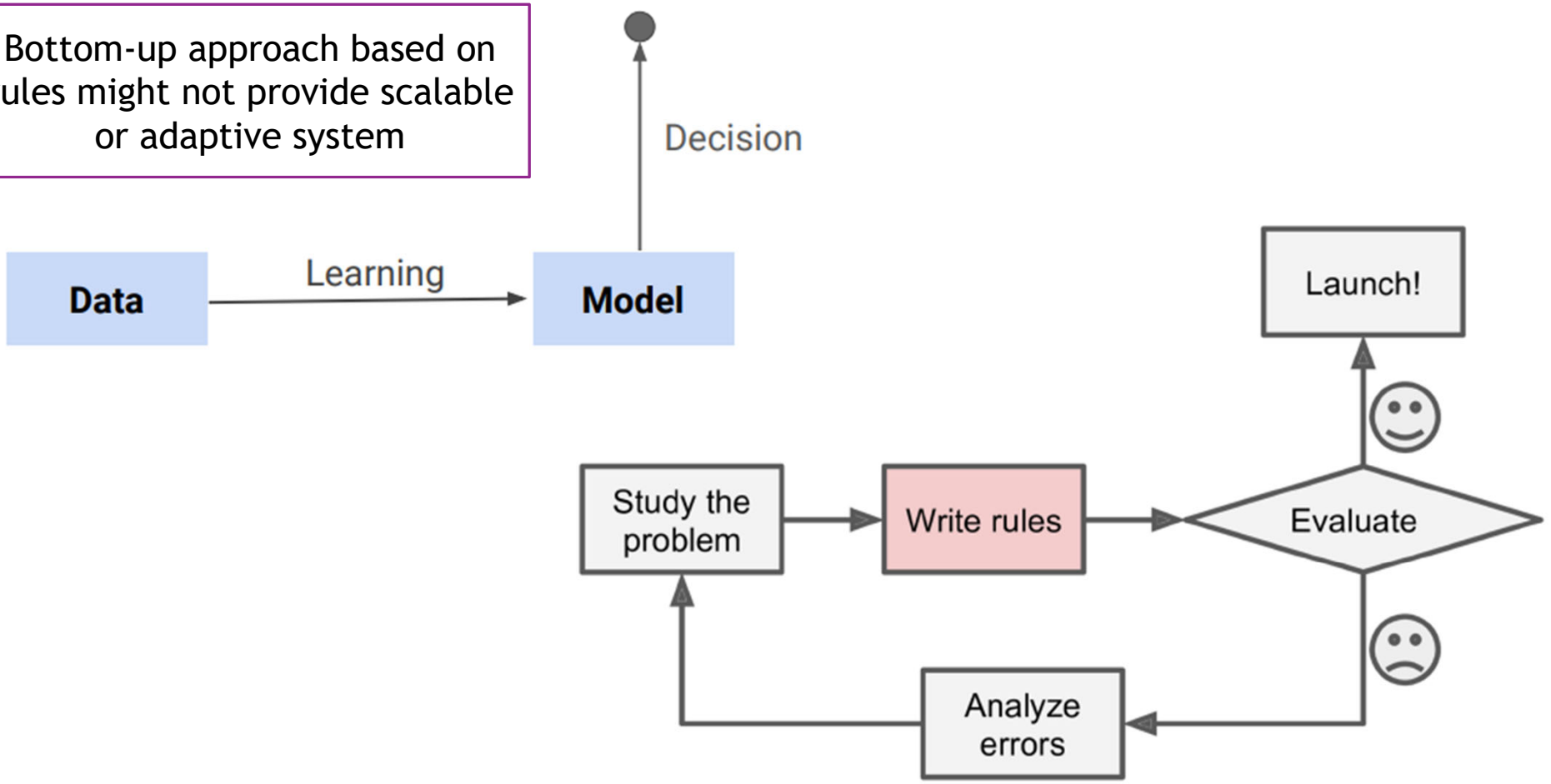
Advances in computing and storage informed the methods



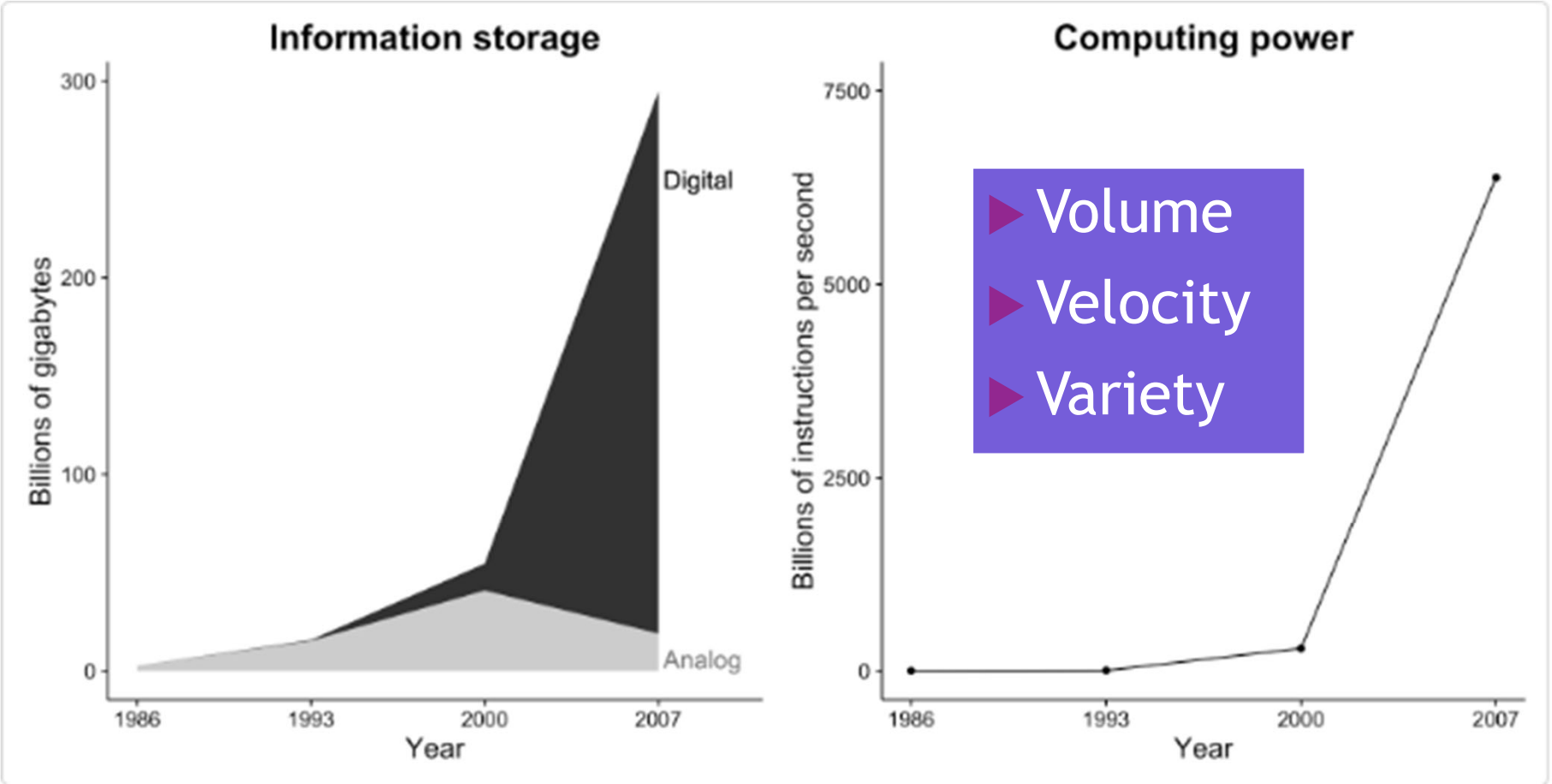


What is machine learning?

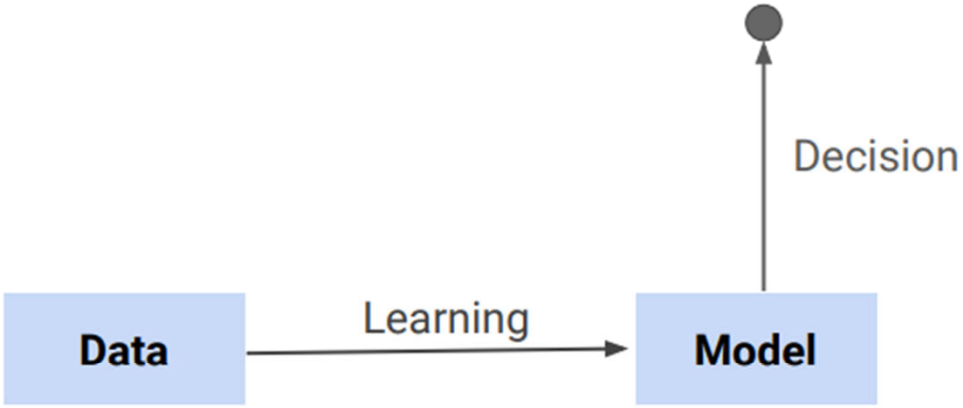
Bottom-up approach based on rules might not provide scalable or adaptive system

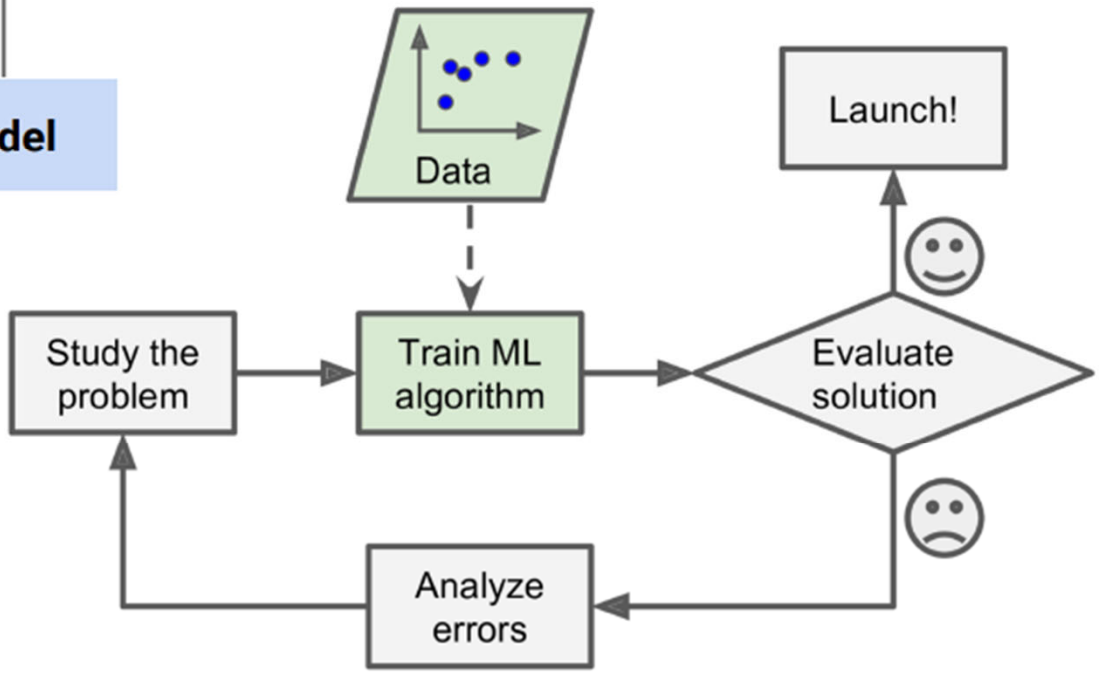


What is machine learning?



What is machine learning?

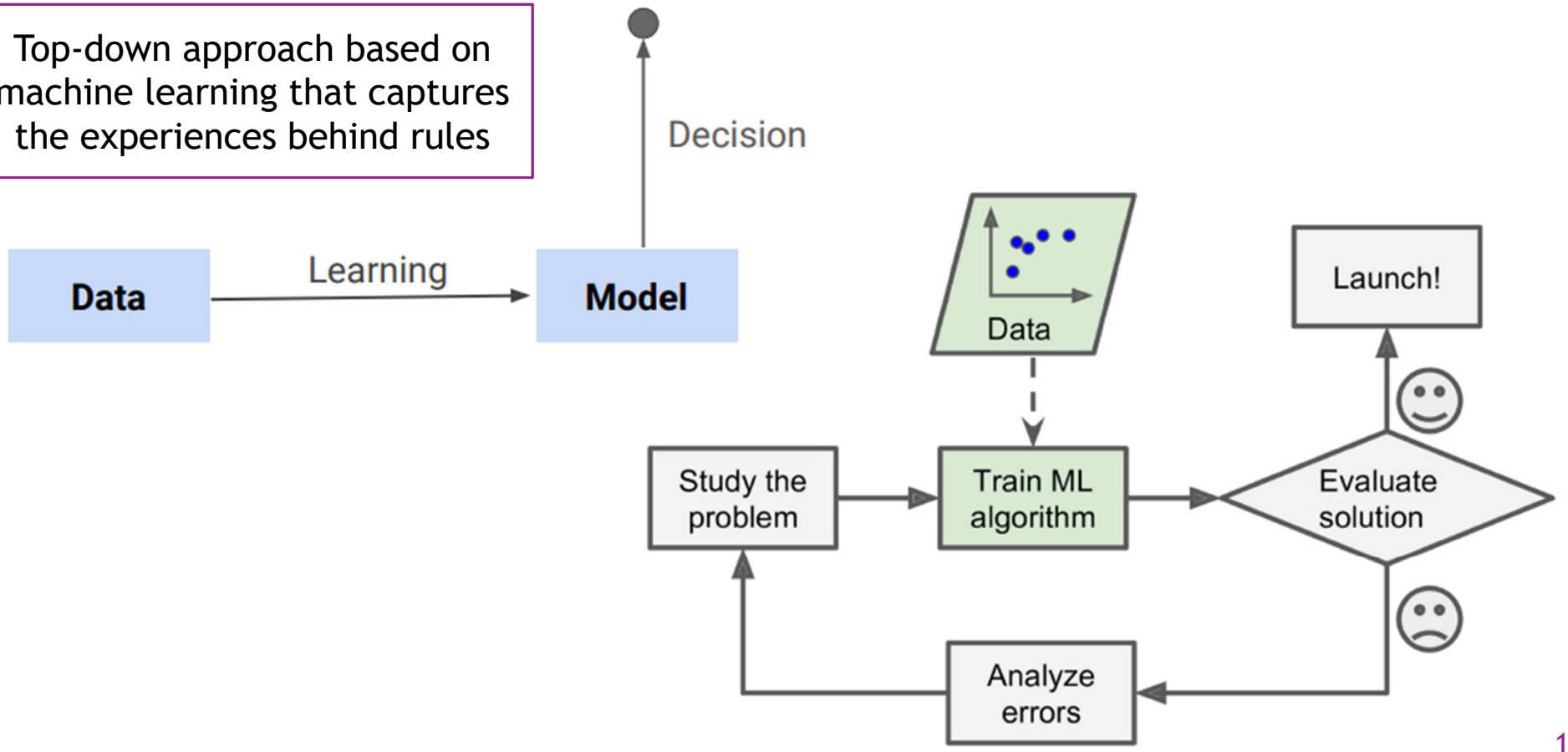




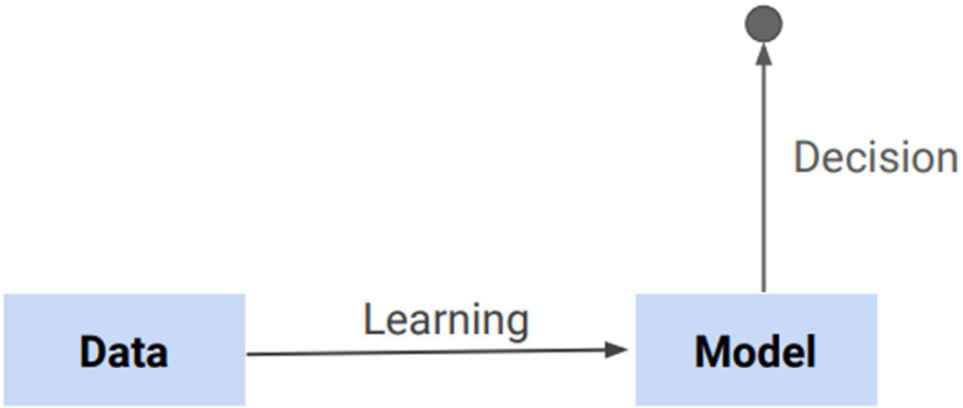
A computer program learns from experience ***E*** with respect to some task ***T*** and some performance measure ***P***, if its performance on task ***T***, as measured by ***P***, improves with experience ***E***.

What is machine learning?

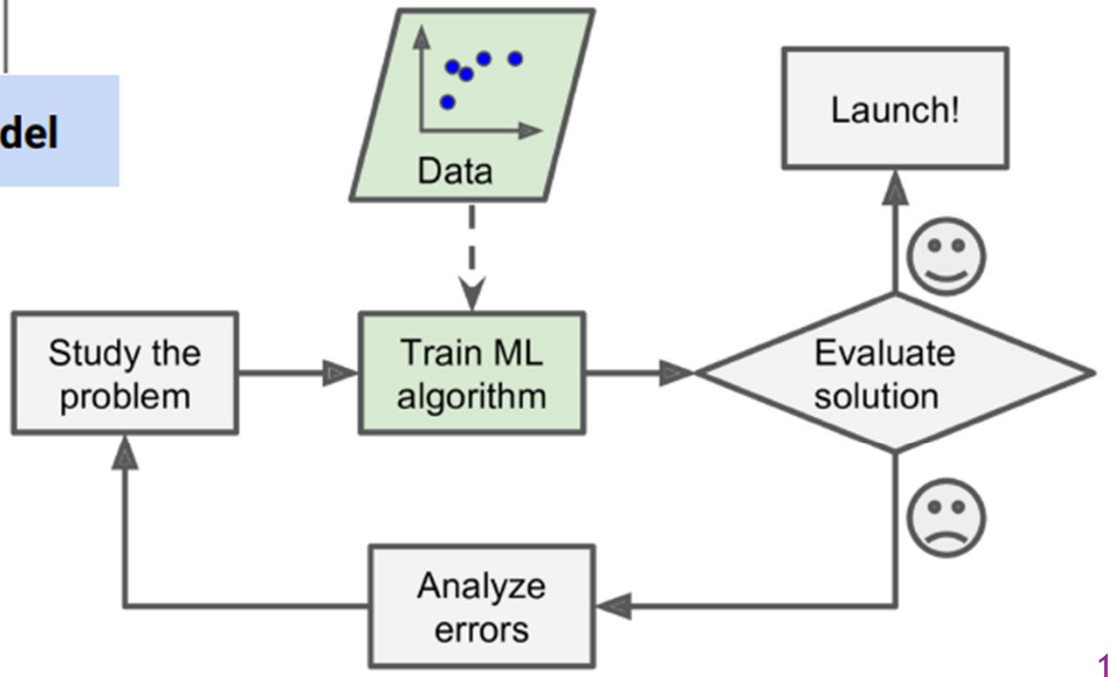
Top-down approach based on machine learning that captures the experiences behind rules



What is machine learning?



Connecting questions and answers through computational and statistical methods that extract information from data



What is machine learning?

	id	subject	email	spam
0	0	Subject: A&L Daily to be auctioned in bankrupt...	url: http://boingboing.net/#85534171\n date: n...	0
1	1	Subject: Wired: "Stronger ties between ISPs an...	url: http://scriptingnews.userland.com/backiss...	0
2	2	Subject: It's just too small ...	<html>\n <head>\n </head>\n <body>\n <font siz...	1
3	3	Subject: liberal defnitions\n	depends on how much over spending vs. how much...	0
4	4	Subject: RE: [ILUG] Newbie seeks advice - Suse...	hehe sorry but if you hit caps lock twice the ...	0

Homework 1

```
url: http://boingboing.net/#85534171
date: not supplied

arts and letters daily, a wonderful and dense blog, has folded up its tent due
to the bankruptcy of its parent company. a&l daily will be auctioned off by the
receivers. link[1] discuss[2] (_thanks, misha!_)

[1] http://www.aldaily.com/
[2] http://www.quicktopic.com/boing/h/zlfterjnd6jf
```

What is machine learning?

Homework 1

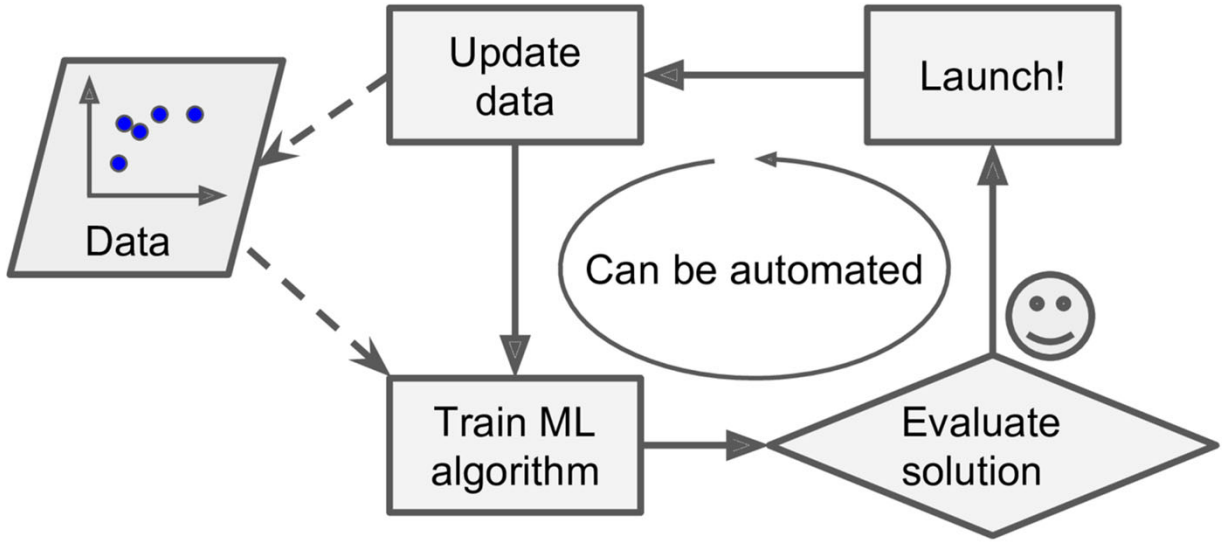
	id	subject	email	spam
0	0	Subject: A&L Daily to be auctioned in bankrupt...	url: http://boingboing.net/#85534171\n date: n...	0
1	1	Subject: Wired: "Stronger ties between ISPs an...	url: http://scriptingnews.userland.com/backiss...	0
2	2	Subject: It's just too small ...	<html>\n <head>\n </head>\n <body>\n <font siz...	1
3	3	Subject: liberal defnitions\n	depends on how much over spending vs. how much...	0
4	4	Subject: RE: [ILUG] Newbie seeks advice - Suse...	hehe sorry but if you hit caps lock twice the ...	0

```
dear ricardo1 ,

<html>
<body>
<center>
<b><font color = "red" size = "+2.5">cost effective direct email advertising</font><br>
<font color = "blue" size = "+2">promote your business for as low as </font><br>
<font color = "red" size = "+2">$50</font> <font color = "blue" size = "+2">per
<font color = "red" size = "+2">1 million</font>
<font color = "blue" size = "+2"> email addresses</font></font><p>
<b><font color = "#44c300" size ="+2">maximize your marketing dollars!<p></font></b>
<font size = "+2">complete and fax this information form to 309-407-7378.<br>
a consultant will contact you to discuss your marketing needs.<br>
```

What is machine learning?

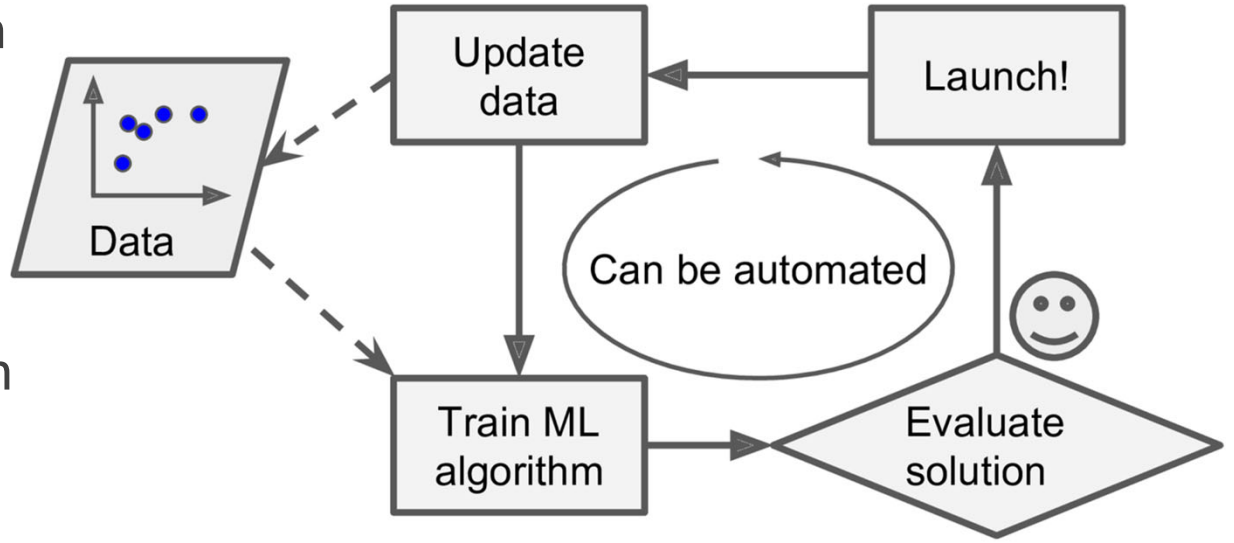
- Rules not Appropriate
- Possess Data
- Pattern in the Data



What is machine learning?

- Natural Language Processing
  - Text classification
  - Part of speech recognition
- Computer Vision
  - Character recognition
  - Image retrieval
- Speech Recognition
  - Source separation
  - Speaker Identification

- Rules not Appropriate
- Possess Data
- Pattern in the Data

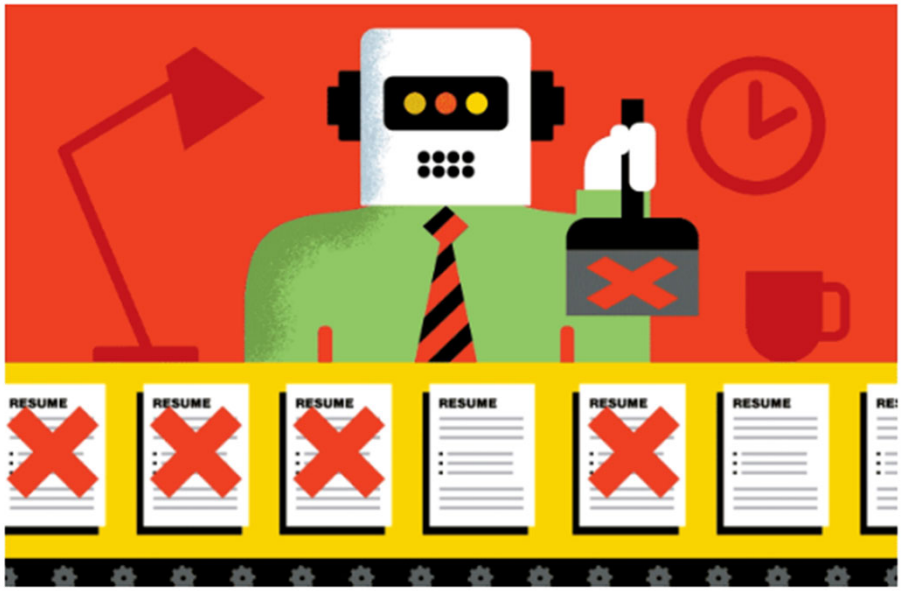


What is machine learning?

Make Your Job Application Robot-Proof

It takes planning to make sure AI gatekeepers don't bounce your résumé—sometimes for arbitrary reasons—before a human can make a call

"Often a job candidate doesn't even know a system is in use," and employers aren't required to disclose it, says Sarah Myers West, a researcher at the AI Now Institute, a New York University research group.



What is machine learning?

ProPublica

Machine Bias

Machine Bias ... In 2014, then U.S. Attorney General Eric Holder warned that the risk scores might be injecting bias into the courts. ... So ProPublica did, as part of a larger examination of the powerful, largely hidden effect of

May 23, 2016

A commercial tool COMPAS automatically predicts some categories of future crime to assist in bail and sentencing decisions. It is used in courts in the US.

What is machine learning?

IEEE Spectrum

Are Your Students Bored? This AI Could Tell You

Qu and his colleagues tested their AI system in two classrooms ... chief architect at Squirrel AI Learning, who was not involved in the paper.

2 weeks ago

The system "provides teachers with a quick and convenient measure of the students' engagement level in a class," says Huamin Qu, a computer scientist at the Hong Kong University of Science and Technology, who co-authored the paper. "Knowing whether the lectures are too hard and when students get bored can help improve teaching."

What is machine learning?

NYU

Memorandum

Date: January 24, 2020

To: THE NYU COMMUNITY

From: Carlo Ciotoli, MD, Associate Vice President for Student Health and Executive Director of the Student Health Center

Re: The Emergence of the Novel Coronavirus

An AI Epidemiologist Sent the First Warnings of the Wuhan Virus

WIRED · 2 days ago

Logistics

For more information please refer to the syllabus on NYU Classes.

Please complete **Homework 0** and **Survey 1** linked to the Weekly Agenda on NYU Classes

Logistics

- Instructors: He and Chris
  - Office Hours: Thursdays 9:30-10:30AM at 60 Fifth Avenue, Room 650

Logistics

- Section Leaders: Shubham, Joshua, Yiqiu
- Graders: Raghav, Sarthak, Peeyush, Aniket, Ieshan, Chinmay

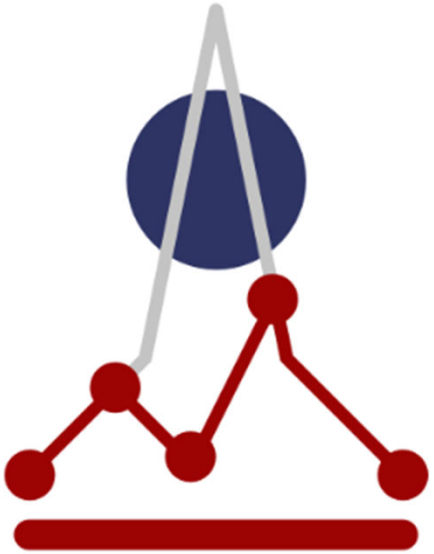
Logistics

- NYU Classes
  - Weekly Agenda, Zoom Conference, Syllabus
- JupyterHub
  - Class Materials, Submission Labs/Homework/Projects
- Piazza
  - Announcements, Discussion
- Gradescope
  - Submission Homework/Projects, Retrieve Exams

Logistics

- Homework 40%
- Project 15%
- Exams
  - Midterm 20%
  - Final 25%
- Extra Credit
  - Optional Problems
  - Office Hours
  - Piazza

Logistics



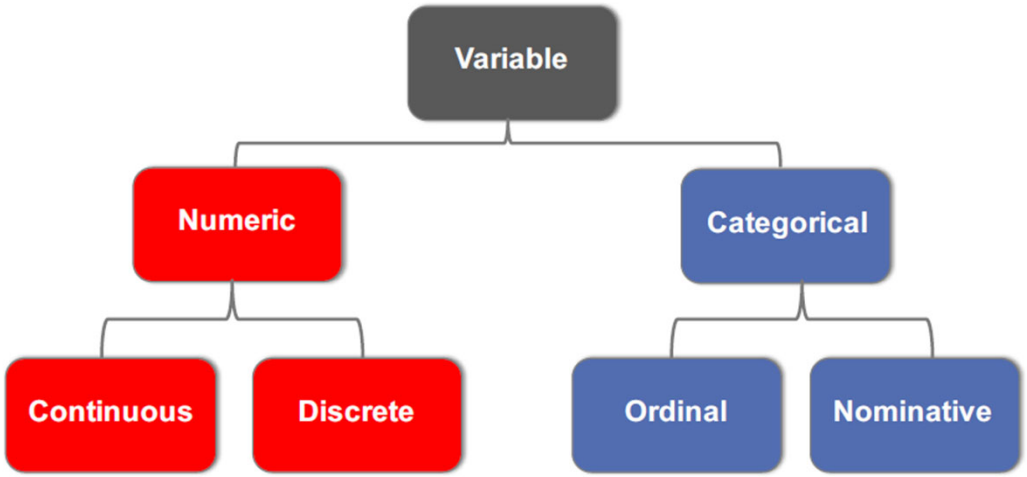
GOVLAB

CEAP
Center for Environmental & Animal Protection

- Project Goals
  - Formulate
  - Process
  - Access
  - Visualize
  - Fit
  - Evaluate

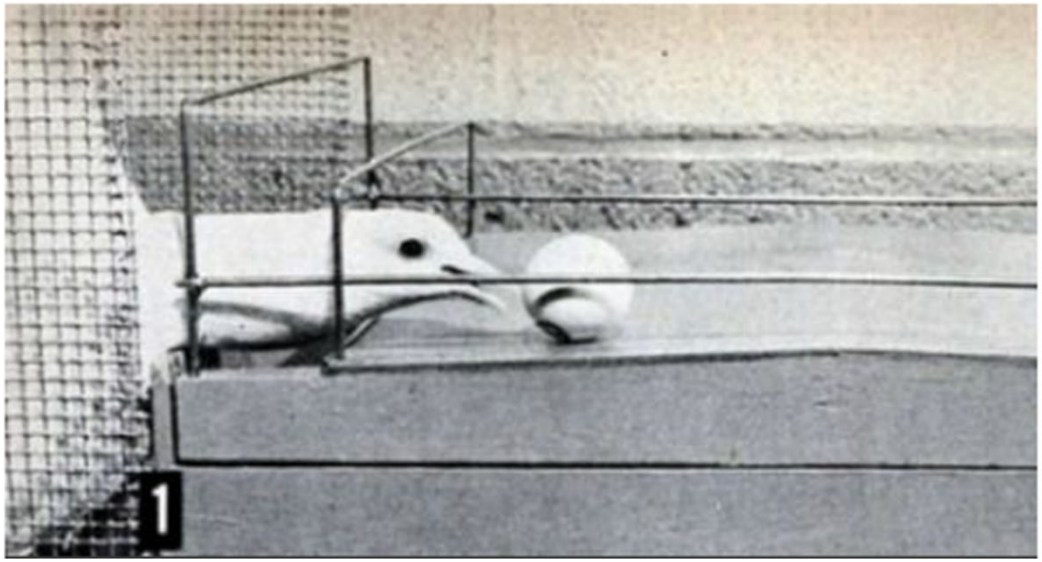
Types of machine learning

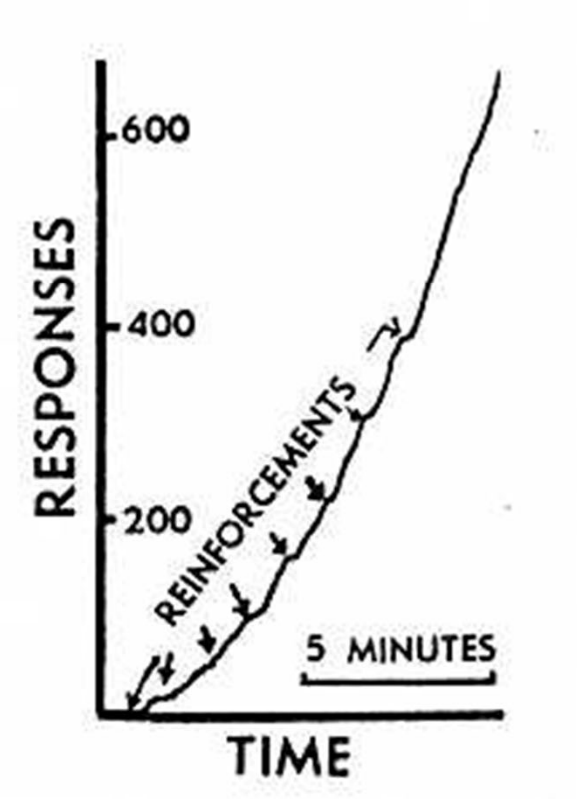
- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
  - Clustering
  - Dimension Reduction
- Reinforcement Learning



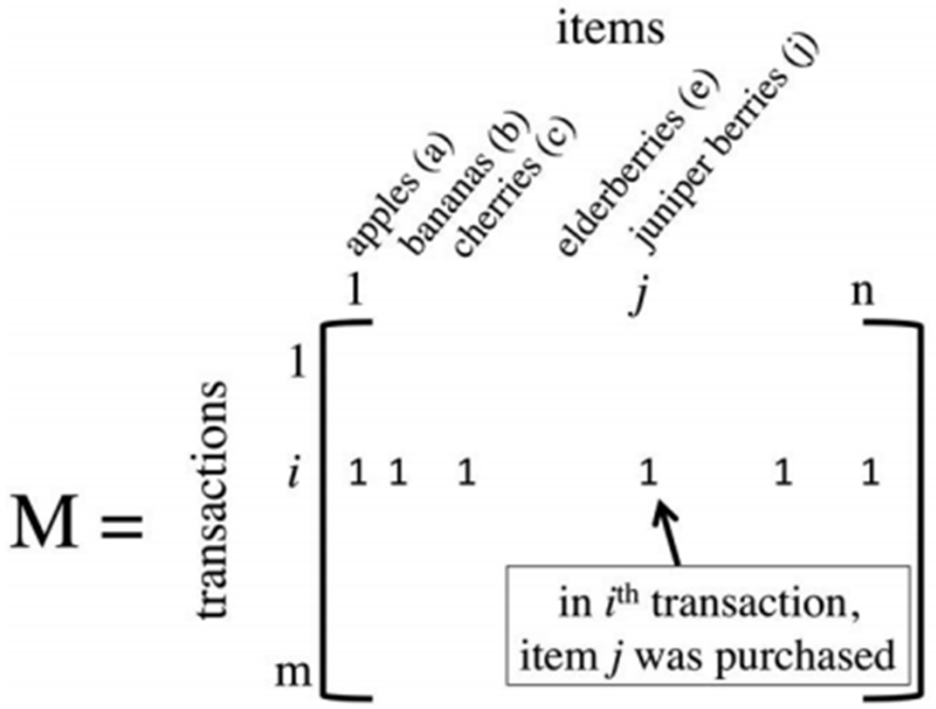
Types of machine learning

- B.F. Skinner (1950)
- Garcia and Koelling (1996)





Clustering


items
apples (a)
bananas (b)
cherries (c)
elderberries (e)
juniper berries (j)
1
j
n
M =
transactions
1
i
m
1 1 1 1 1 1
in i^{th} transaction,
item j was purchased

Clustering

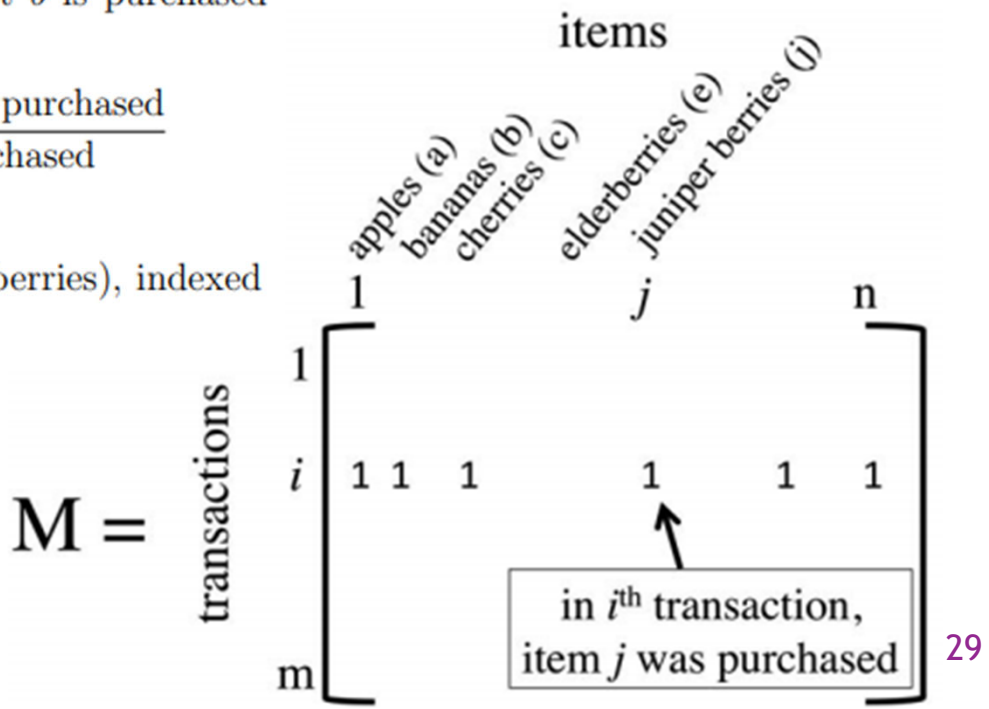
- **Support** of an itemset: number of transactions containing it,

$$\text{Supp(bananas, cherries, elderberries)} = \sum_{i=1}^{m} M_{i,2} \cdot M_{i,3} \cdot M_{i,5}.$$

- **Confidence** of rule $a \to b$: the fraction of times itemset b is purchased when itemset a is purchased.

$$\begin{aligned} \text{Conf}(a \to b) &= \frac{\text{Supp}(a \cup b)}{\text{Supp}(a)} = \frac{\#\text{times } a \text{ and } b \text{ are purchased}}{\#\text{times } a \text{ is purchased}} \\ &= \hat{P}(b|a). \end{aligned}$$

- **Itemset**: a subset of items, e.g., (bananas, cherries, elderberries), indexed by $\{2, 3, 5\}$.



Clustering

- **Support** of an itemset: number of transactions containing it,

$$\text{Supp(bananas, cherries, elderberries)} = \sum_{i=1}^{m} M_{i,2} \cdot M_{i,3} \cdot M_{i,5}.$$

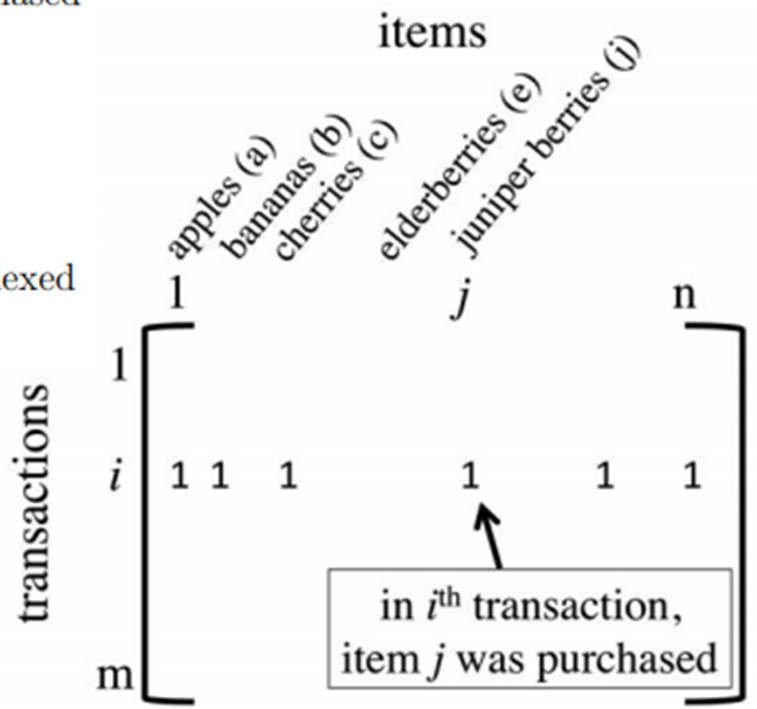
- **Confidence** of rule $a \to b$: the fraction of times itemset b is purchased when itemset a is purchased.

$$\begin{aligned} \text{Conf}(a \to b) &= \frac{\text{Supp}(a \cup b)}{\text{Supp}(a)} = \frac{\#\text{times } a \text{ and } b \text{ are purchased}}{\#\text{times } a \text{ is purchased}} \\ &= \hat{P}(b|a). \end{aligned}$$

- **Itemset**: a subset of items, e.g., (bananas, cherries, elderberries), indexed by $\{2, 3, 5\}$.

$$\text{Supp}(a \cup b) \geq \theta, \text{ and } \text{Conf}(a \to b) \geq \text{minconf}.$$

$$\mathbf{M} =$$

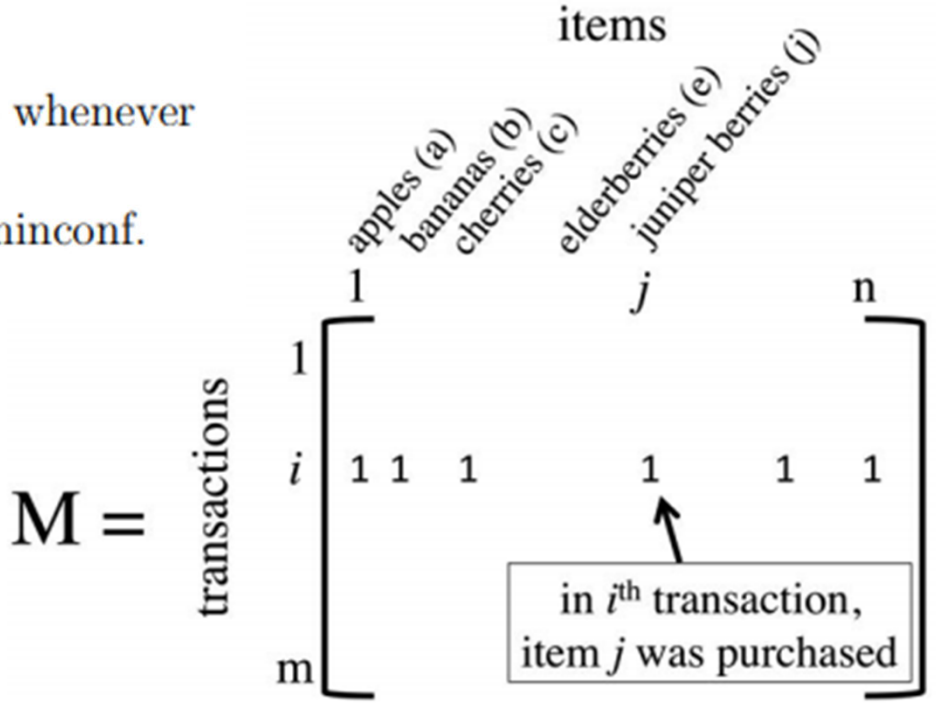


Clustering

Confidence

- For each frequent itemset ℓ:
  - Find all nonempty subsets of ℓ
  - For each subset a, output $a \rightarrow \{\ell \setminus a\}$ whenever

$$\frac{\text{Supp}(\ell)}{\text{Supp}(a)} \geq \text{minconf.}$$



Clustering

If Supp($a \cup b$) $\geq \theta$ then Supp(a) $\geq \theta$ and Supp(b) $\geq \theta$.

Support

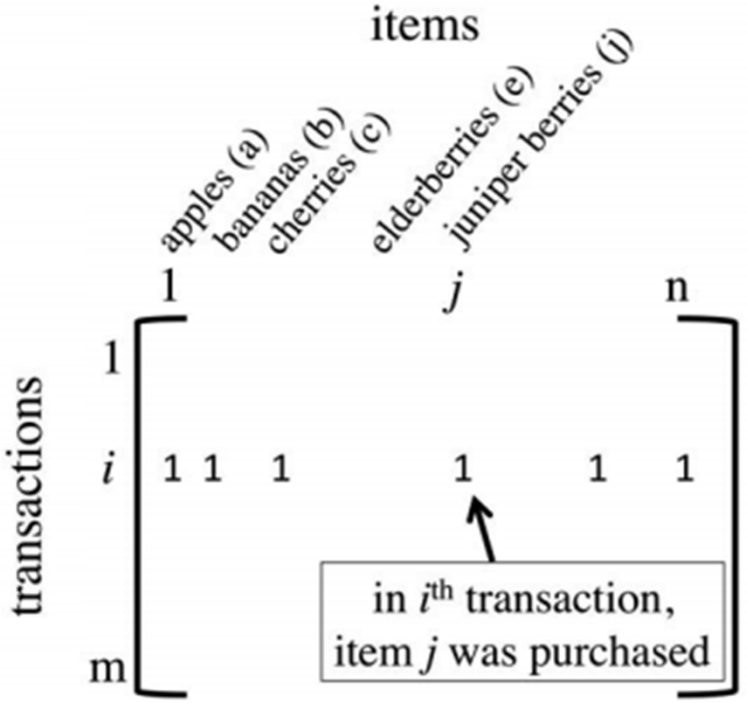
	apples	bananas	cherries		elderberries		grapes
1-itemsets:	a	b	c	d	e	~~f~~	g
supp:	25	20	30	45	29	5	17

2-itemsets:	~~{a,b}~~	{a,c}	{a,d}	{a,e}	...	~~{e,g}~~
supp:	7	25	15	23		3

3-itemsets:	{a,c,d}	{a,c,e}	{b,d,g}	...
supp:	15	22	15	

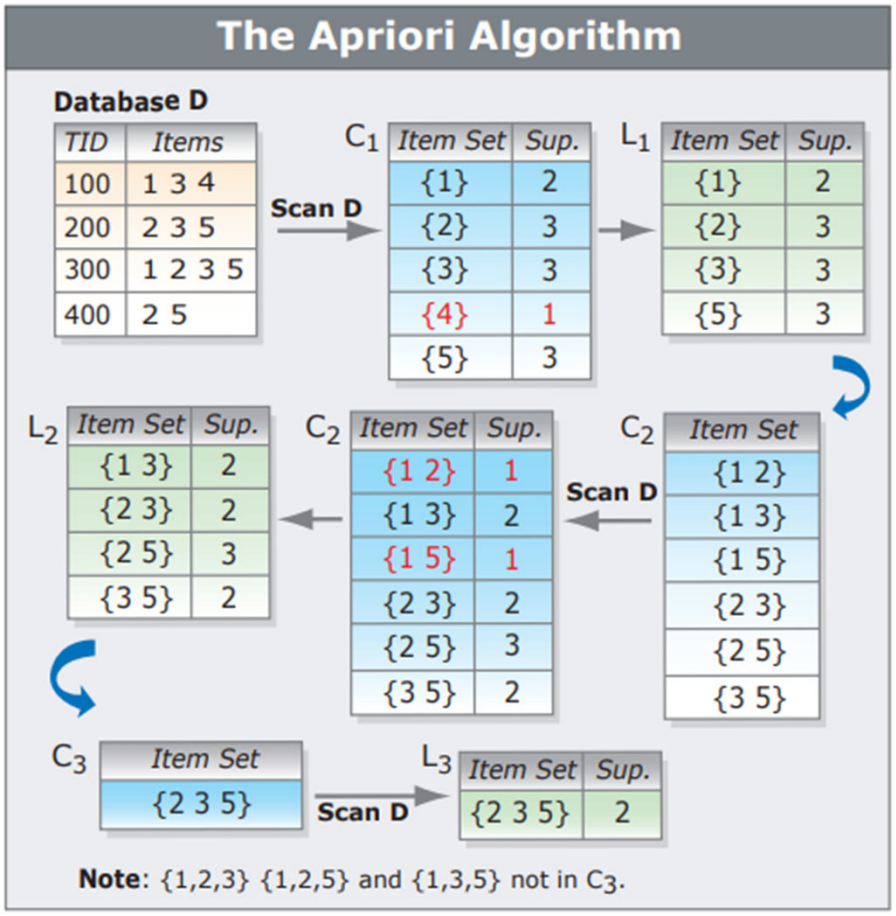
4-itemsets:	{a,c,d,e}
supp:	12

=



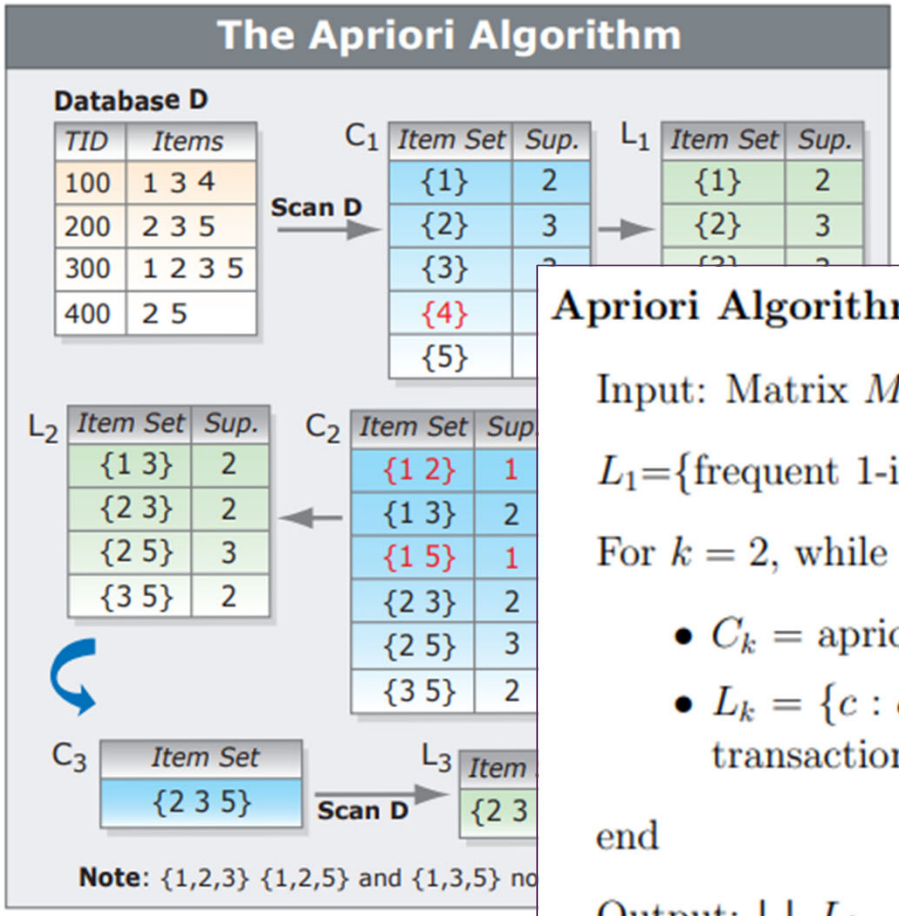
Clustering

Support



Clustering

Support



Apriori Algorithm:

Input: Matrix M

L_1={frequent 1-itemsets; i such that Supp$(i) \geq \theta$}.

For $k = 2$, while $L_{k-1} \neq \emptyset$ (while there are large $k-1$-itemsets), $k++$

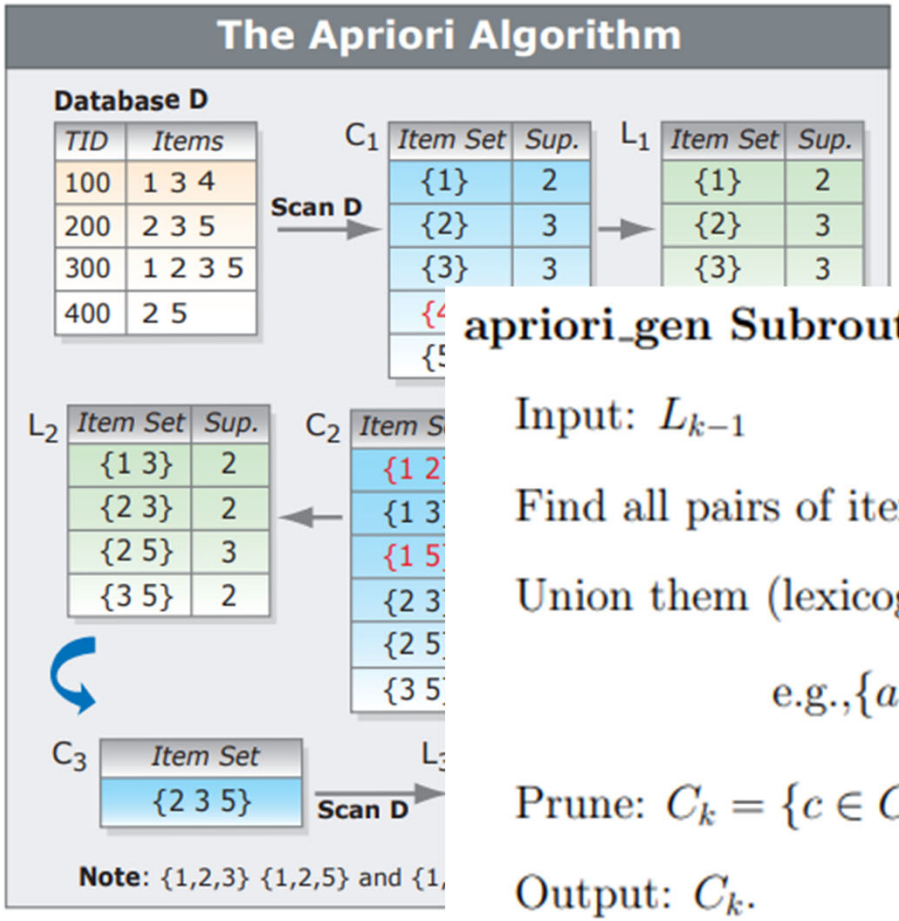
- C_k = apriori_gen(L_{k-1}) generate candidate itemsets of size k
- $L_k = \{c : c \in C_k,$ Supp$(c) \geq \theta\}$ frequent itemsets of size k (loop over transactions, scan the database)

end

Output: $\bigcup_k L_k$.

Clustering

Support



apriori_gen Subroutine:

Input: L_{k-1}

Find all pairs of itemsets in L_{k-1} where the first $k-2$ items are identical.

Union them (lexicographically) to get $C_k^{\text{too big}}$,

$$\text{e.g.,} \{a,b,c,d,e,f\}, \{a,b,c,d,e,g\} \rightarrow \{a,b,c,d,e,f,g\}$$

Prune: $C_k = \{c \in C_k^{\text{too big}}, \text{all } (k-1)\text{-subsets } c_s \text{ of } c \text{ obey } c_s \in L_{k-1}\}$.

Output: C_k.

Regression

Assume a linear relation

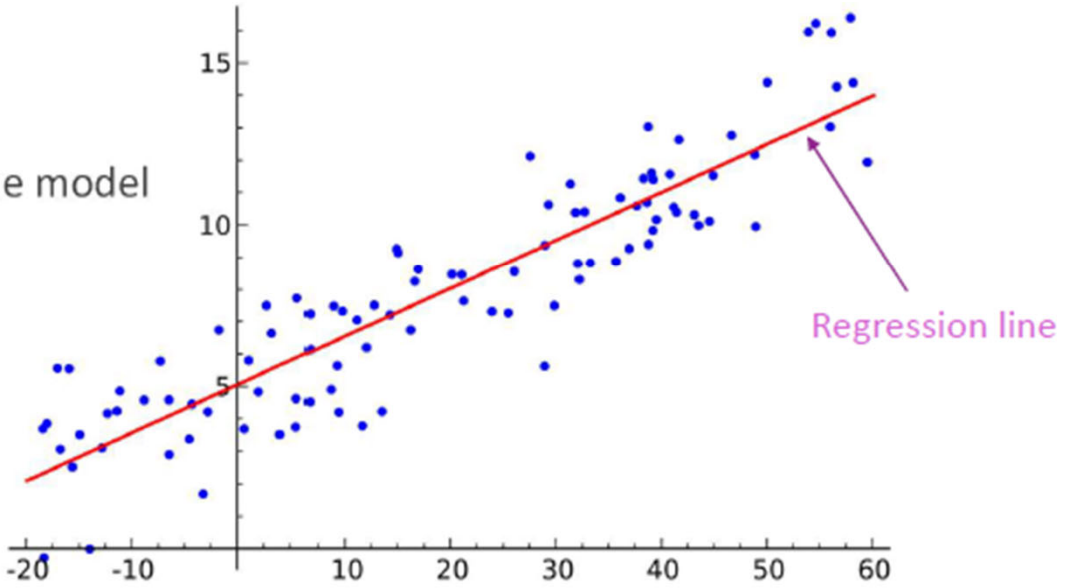
$$y \approx \beta_0 + \beta_1 x$$

- β_0 = intercept
- β_1 = slope

$\beta = (\beta_0, \beta_1)$ are the parameters of the model

What are the units of β_0, β_1?

When is this model good?



Regression

Knowing x does not exactly predict y

- Variation in y due to factors other than x

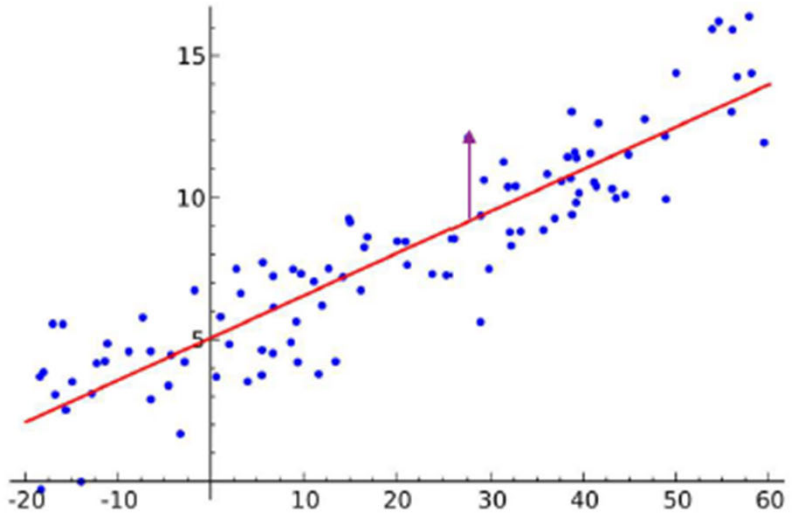
Add a residual term

$$y = \beta_0 + \beta_1 x + \epsilon$$

Residual = component the model does not explain

- Predicted value: $\hat{y}_i = \beta_1 x_i + \beta_0$
- Residual: $\epsilon_i = y_i - \hat{y}_i$

Vertical deviation from the regression line



Regression

How do we select parameters $\beta = (\beta_0, \beta_1)$?

Define $\hat{y}_i = \beta_1 x_i + \beta_0$

- Predicted value on sample i for parameters $\beta = (\beta_0, \beta_1)$

Define average residual sum of squares:

$$\text{RSS}(\beta_0, \beta_1) := \sum_{I=1}^{n} (y_i - \hat{y}_i)^2$$

- Note that $\hat{y}_i$ is implicitly a function of $\beta = (\beta_0, \beta_1)$
- Also called the sum of squared residuals (SSR) and sum of squared errors (SSE)

Least squares solution: Find (β_0, β_1) to minimize RSS.

- Geometrically, minimizes squared distances of samples to regression line

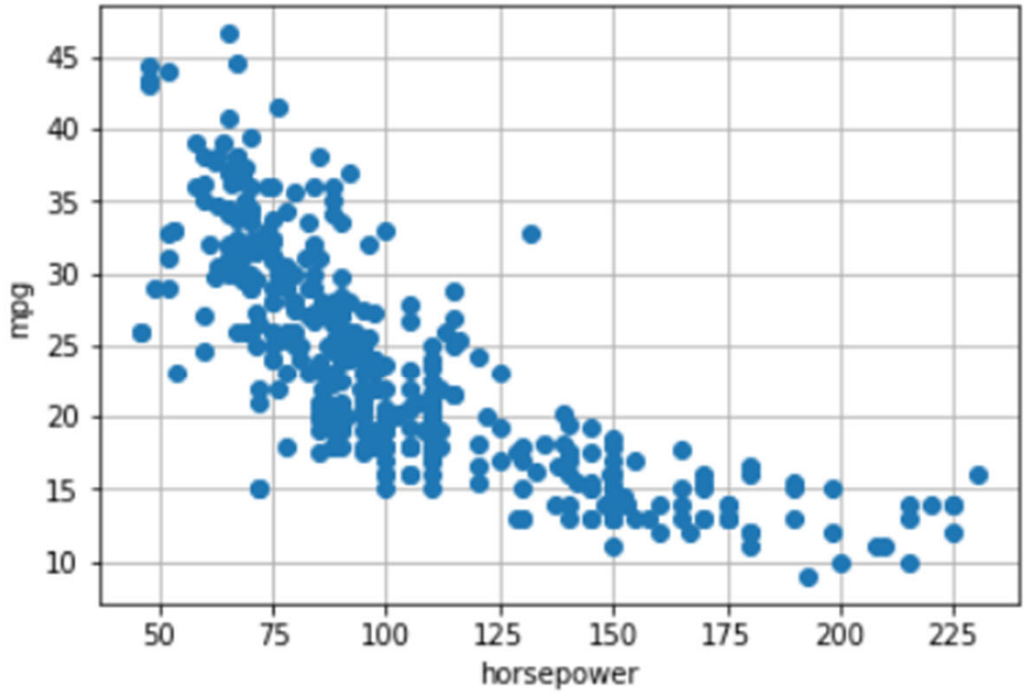
Exercise

SA Scientific American

VW Emissions Cheating Scandal Increased Children's ...

The Volkswagen emissions cheating scandal—also known as ... September 2015, when a team of scientists at West Virginia University discovered that the ... The working paper from the Chicago Fed relied

Jul 17, 2019



Regression

General ML problem		Simple linear regression
IFind a model with parameters	→	Linear model: $\hat{y} = \beta_0 + \beta_1 x$
IGet data	→	Data: $(x_i, y_i), i = 1,2, \dots, N$
IPick a loss function ◦ Measures goodness of fit model to data ◦ Function of the parameters	→	Loss function: $RSS(\beta_0, \beta_1) \coloneqq \sum(y_i - \beta_0 + \beta_1 x_i)^2$
IFind parameters that minimizes loss	→	Select β_0, β_1 to minimize $RSS(\beta_0, \beta_1)$

Questions

- Questions on Piazza?
- Question for You!
  - Lift is Weighted Confidence?

$$\frac{\hat{P}(b|a)}{\hat{P}(b)}$$

- What is the interpretation of lift?
- Why would we use lift instead of confidence?
- Compute lift in terms of the supports of a,b and aUb