Foundations of Data Science

DS 3001

Data Science Program
Department of Computer Science
Worcester Polytechnic Institute
Instructor: Prof. Kyumin Lee

Project Teams

1. Clay Oshiro-Leavitt, Hunter Caouette, Nick Alescio
2. Danielle Angelini, Elijah Ellis, Ryan Candy, Rob Wondolowski
3. Eva (Yingbing) Lu, Manasi Danke, Erica Lee, Jonathan Dang
4. Arianna Kan, Yihan Lin, Margaret Goodwin, Ken Snoddy
5. Yang Gao, Jose Li, Sarah Burns, Daniel McDonough
6. Noah Puchovsky, Katherine Handy, Alex Tavares, Angelica Puchovsky
7. Armando Zubillaga, Gabriel Rodrigues, Humberto Leon, Joao Omena de Lucena
8. Edward Carlson, Samuel Goldman, Nick Krichevsky, Christopher Myers
9. Jessie White, Lindsay MacInnis, Bao Huynh, Ziqian Zeng
10. Suverino Frith, Nicholas Odell, Fay Whittall, Johvanni Perez
11. Alp Piskin, Robert Scalfani, Jake Barefoot, Mark Bernardo
12. Amanda Chan, Nugzar Chkhaidze, Luke Gebler
13. Daniel Pelaez, Nathan Savard, Kate Sincaglia
14. Maan Alneami

HW2

- Implement linear regression
  - https://canvas.wpi.edu/courses/18106/assignments/131989
  - Due date is April 17

Upcoming Schedule

- Exam 1 on April 17 at 2pm

- Project Proposal
  - https://canvas.wpi.edu/courses/18106/assignments/132329
  - Due date: April 21

Analyzing Board Health Log Data by Applying Machine Learning Techniques

- MQP opportunity for CS students who are looking for MQP opportunity

- Sponsored by Dell EMC

- https://eprojects.wpi.edu/group/11761

- Email me if you are interested in this project

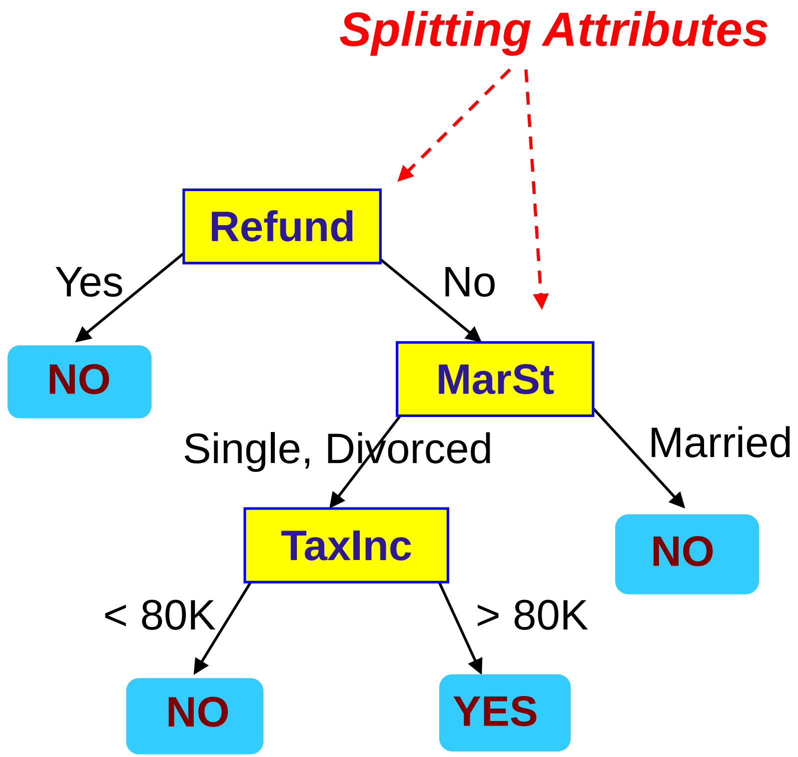
Mining and Analytics: Classification +

Decision Trees

Example of a Decision Tree

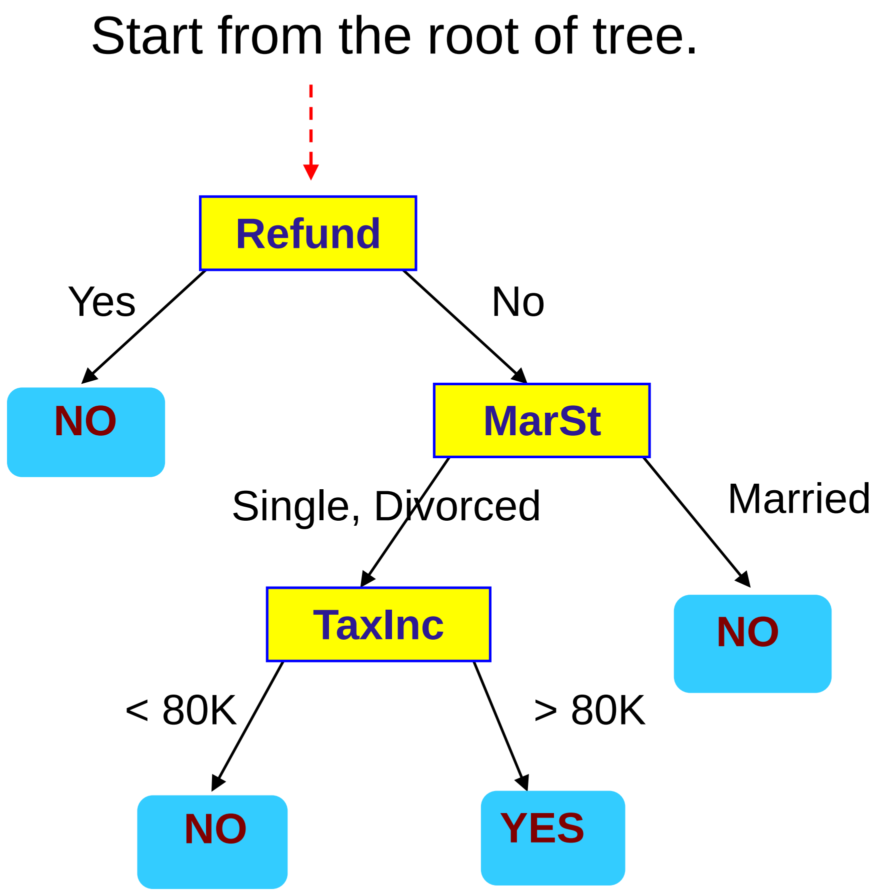
	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Training Data



Model: Decision Tree

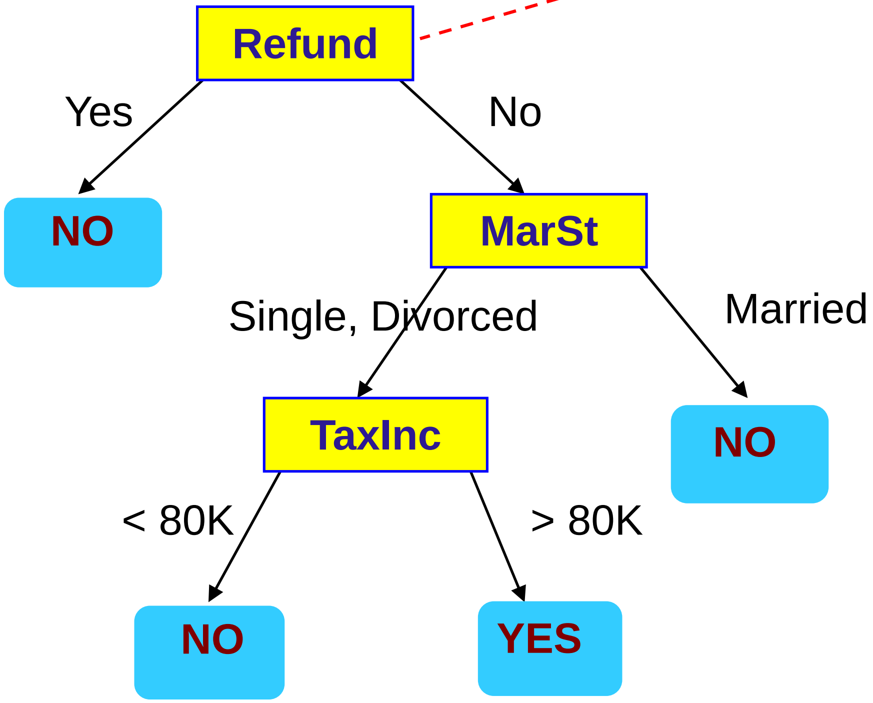
Apply Model to Test Data



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Apply Model to Test Data



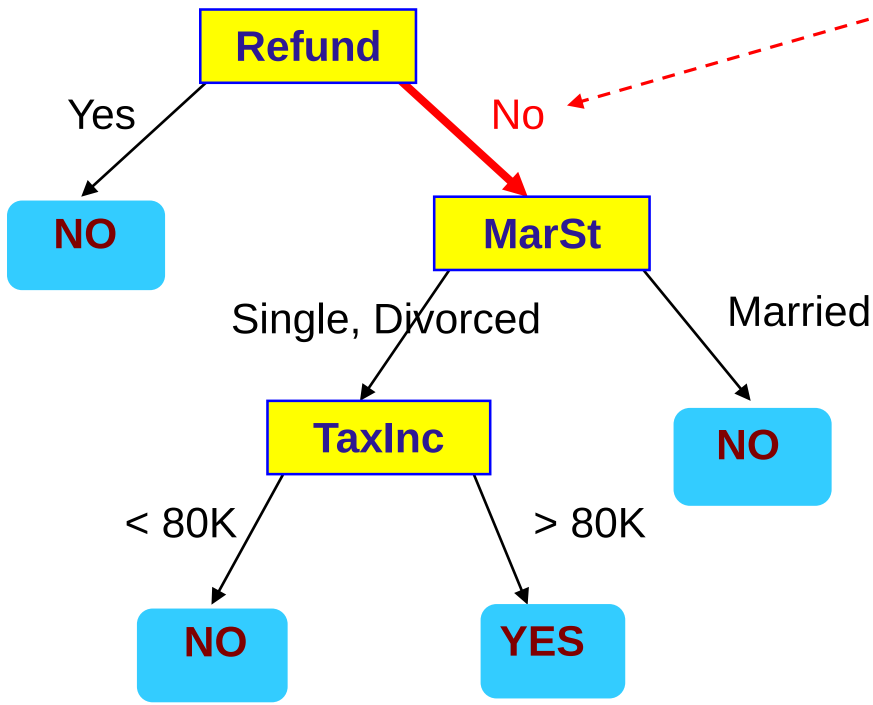
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

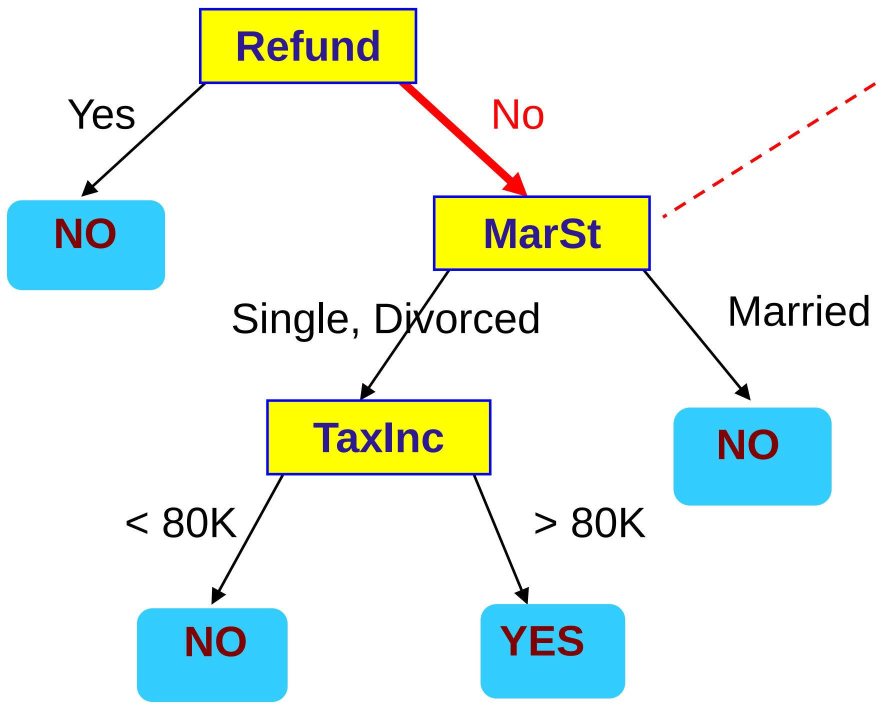
Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data



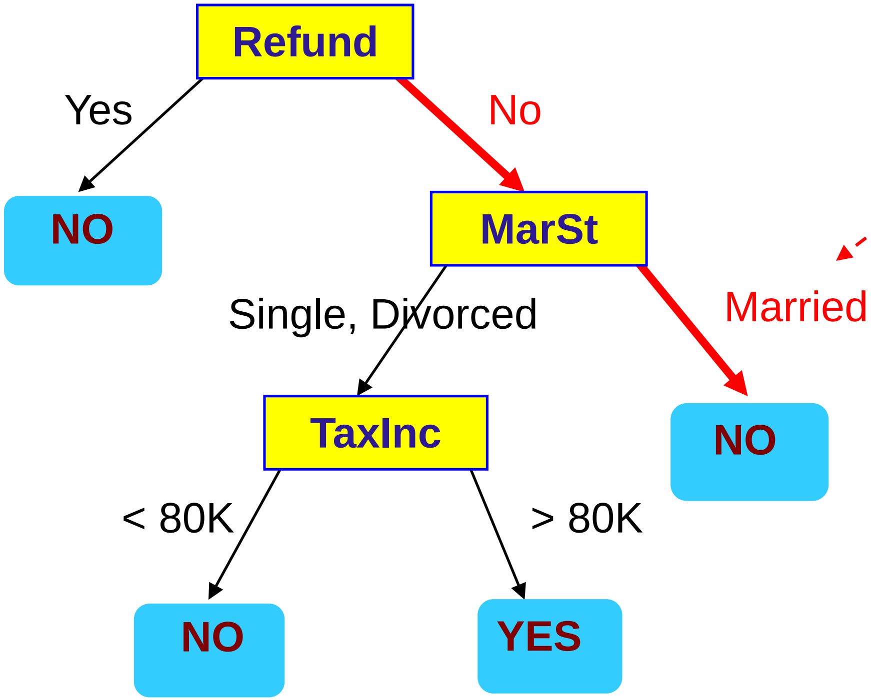
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

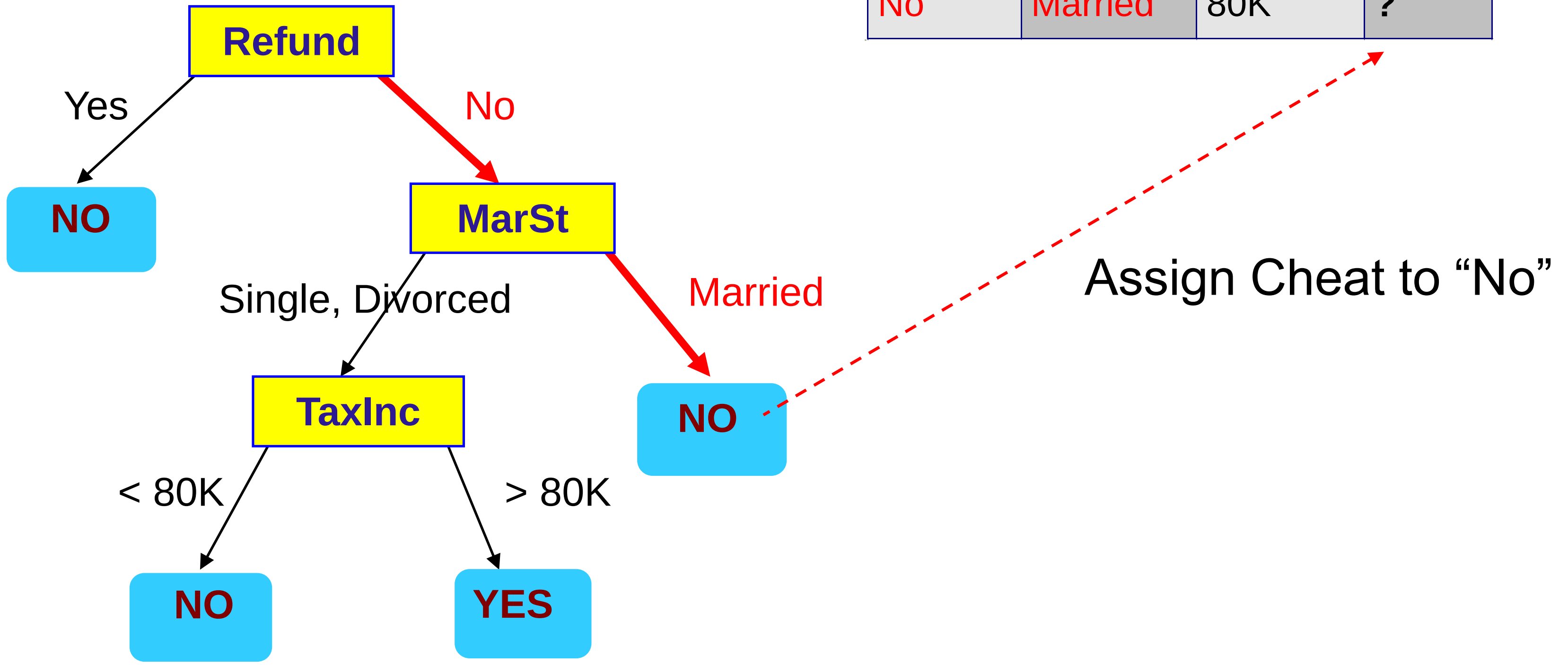
Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Splitting Criterion

- Ideas?
- Intuition: Prefer nodes with *homogeneous* class distribution

C0: 5	C0: 9
C1: 5	C1: 1
Non-homogeneous,	**Homogeneous,**
High degree of impurity	**Low degree of impurity**

- Typical methods (i.e., measuring impurity)
  - Gini Index
  - Entropy / Information Gain
  - Classification error

Splitting Criterion: GINI

- Gini Index for a given node t :

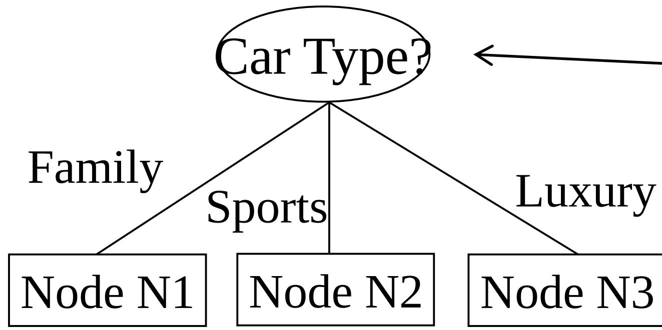
$$\boxed{GINI(t) = 1 - \sum_j [p(j \mid t)]^2}$$

(NOTE: $p(j \mid t)$ is the relative frequency of class j at node t).

- Measure the impurity of a node
  - Maximum $(1 - 1/n_c)$ when records are equally distributed among all classes, implying least interesting information
  - Minimum (0.0) when all records belong to one class, implying most interesting information

GINI Example

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$



	Parent
C1	6
C2	6

Node N1

C1	0
C2	6

P(C1) = 0/6 = 0　　P(C2) = 6/6 = 1

Gini = 1 – $P(C1)^2$ – $P(C2)^2$ = 1 – 0 – 1 = 0

Node N1

C1	1
C2	5

P(C1) = 1/6　　P(C2) = 5/6

Gini = 1 – $(1/6)^2$ – $(5/6)^2$ = 0.278

Node N1

C1	2
C2	4

P(C1) = 2/6　　P(C2) = 4/6

Gini = 1 – $(2/6)^2$ – $(4/6)^2$ = 0.444

Splitting Based on GINI

- Used in CART, SLIQ, SPRINT.
- When a node p is split into k partitions (children), the quality of split is computed as,

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

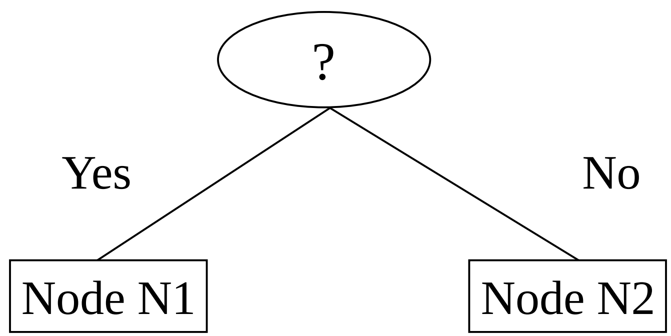
where, n_i = number of records at child i,

n = number of records at node p.

Also called collective impurity of child nodes

GINI for Binary Attributes

- Splits into two partitions
- Effect of Weighing partitions:
  - Larger and Purer Partitions are sought for.



	Parent
C1	**6**
C2	**6**
Gini = 0.500	

Gini(N1)
$= 1 - (5/7)^2 - (2/7)^2$
$= 0.409$

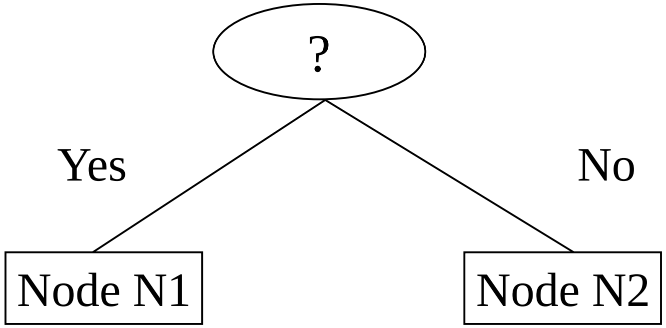
Gini(N2)
$= 1 - (1/5)^2 - (4/5)^2$
$= 0.32$

	N1	N2
C1	**5**	**1**
C2	**2**	**4**
Gini=0.371		

Gini(Children)
$= 7/12 * 0.409 +$
$\quad 5/12 * 0.32$
$= 0.371$

GINI for Binary Attributes

- Splits into two partitions
- Effect of Weighing partitions:
  - Larger and Purer Partitions are sought for.



	Parent
C1	**6**
C2	**6**
Gini = 0.500	

Attribute A

	N1	N2
C1	**0**	**6**
C2	**6**	**0**
Gini=0.000		

Attribute B

	N1	N2
C1	**5**	**1**
C2	**1**	**5**
Gini=0.278		

Attribute C

	N1	N2
C1	**4**	**2**
C2	**3**	**3**
Gini=0.486		

Attribute D

	N1	N2
C1	**3**	**3**
C2	**3**	**3**
Gini=0.500		

GINI for Nominal Attributes

- For each distinct value, gather counts for each class in the dataset
- Use the count matrix to make decisions

Multi-way split

	CarType		
	Family	Sports	Luxury
C1	1	2	1
C2	4	1	1
Gini	0.393		

Two-way split
(find best partition of values)

	CarType	
	{Sports, Luxury}	{Family}
C1	3	1
C2	2	4
Gini	0.400	

	CarType	
	{Sports}	{Family, Luxury}
C1	2	2
C2	1	5
Gini	0.419	

GINI for Quantitative Attributes

- Use Binary Decisions based on one value
- Several Choices for the splitting value
  - Number of possible splitting values = Number of distinct values
- Each splitting value has a count matrix associated with it
  - Class counts in each of the partitions, $A < v$ and $A \geq v$
- Simple method to choose best v
  - For each v, scan the database/dataset to gather count matrix and compute its Gini index
  - Computationally Inefficient! Repetition of work. $O(n)^2$

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

GINI for Quantitative Attributes

- For efficient computation: for each attribute,
  - Sort the attribute on values
  - Linearly scan these values, each time updating the count matrix and computing gini index
  - Choose the split position that has the least gini index
  - O(nlogn)

Sorted Values →

Split Positions →

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
Taxable Income	60	70	75	85	90	95	100	120	125	220

	55		65		72		80		87		92		97		110		122		172		230	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
No	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0.420		0.400		0.375		0.343		0.417		0.400		*0.300*		0.343		0.375		0.400		0.420	

Other splitting criteria

- Information Gain
- Classification Error

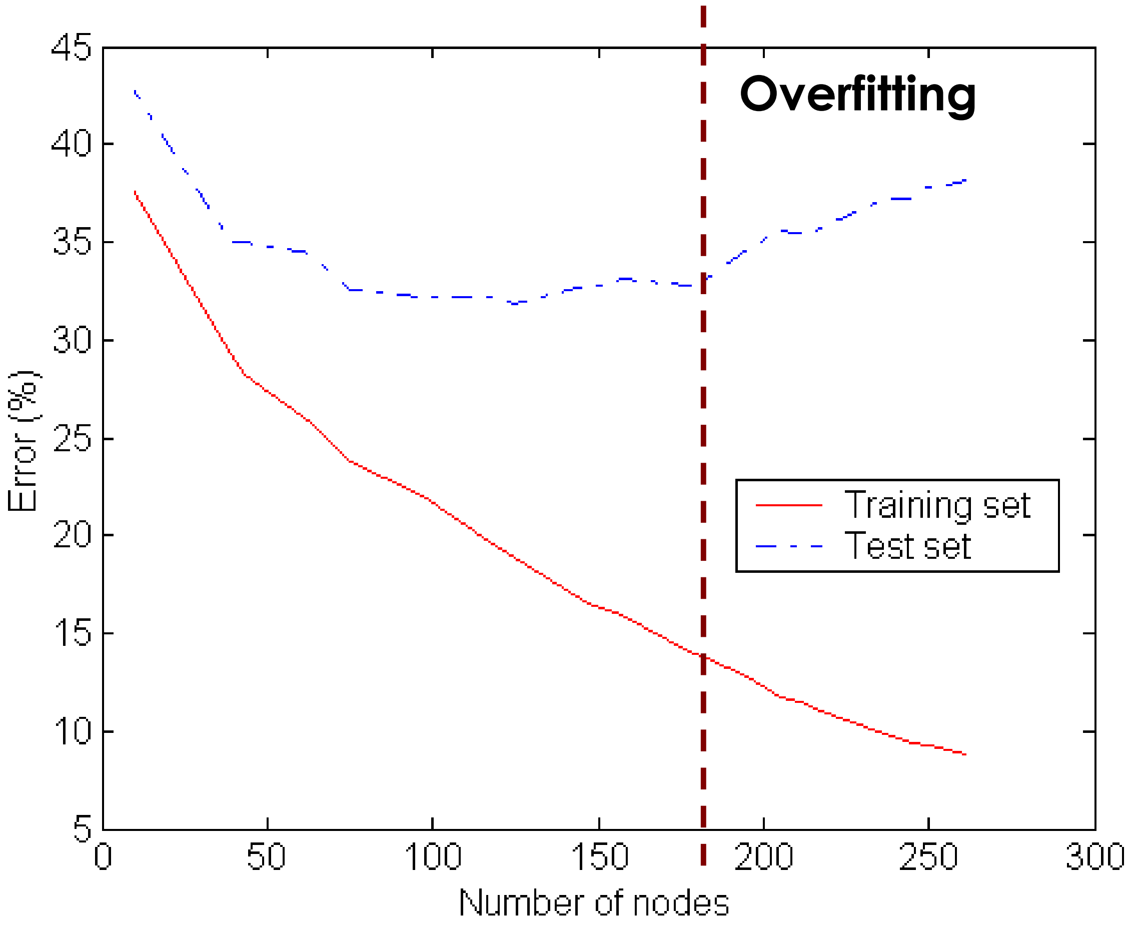
(See readings)

When do we stop splitting?

- Stop expanding a node when all the records belong to the same class
- Stop expanding a node when all the records have similar attribute values
- Early termination

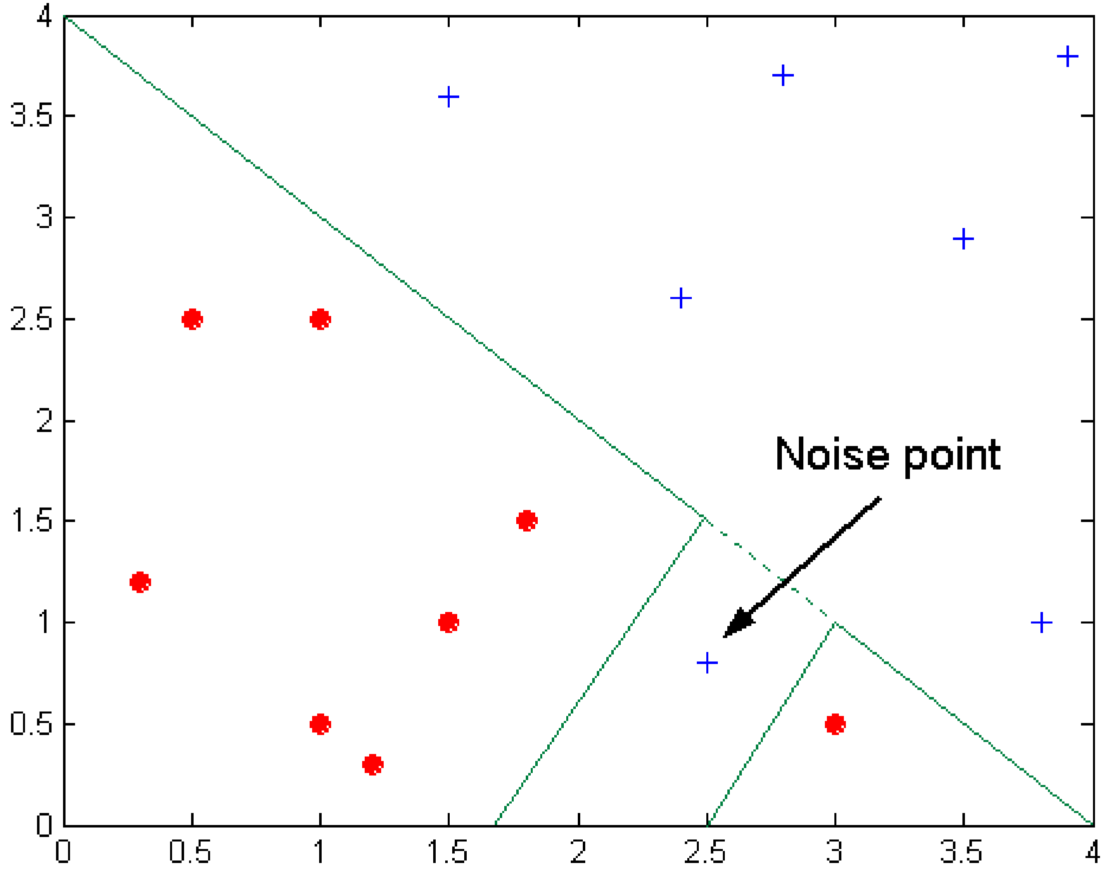
Issues with Classification

Underfitting and Overfitting



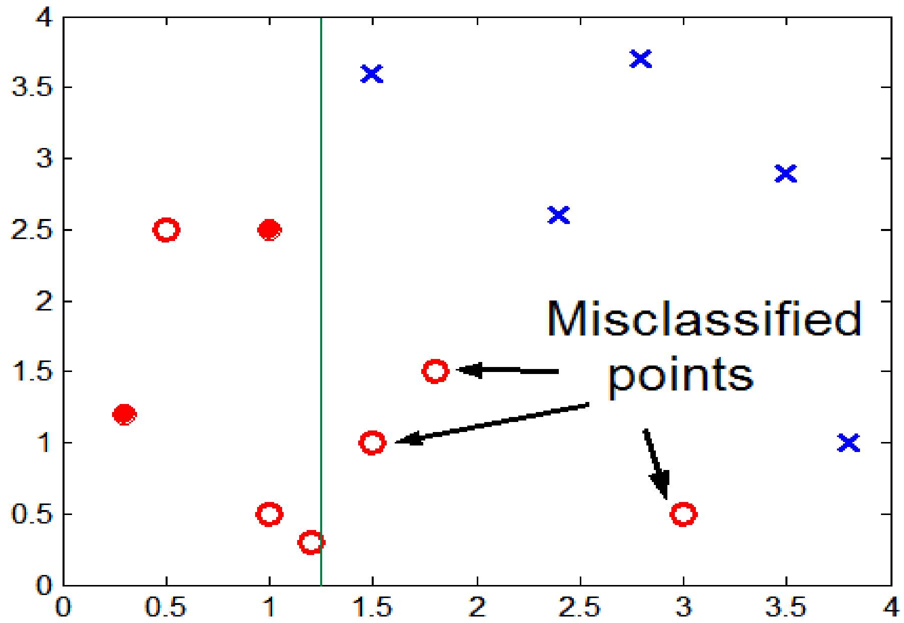
Underfitting: when model is too simple, both training and test errors are large

Overfitting due to Noise



Decision boundary is distorted by noise point

Overfitting due to Insufficient Examples



Lack of data points in the lower half of the diagram makes it difficult to predict correctly the class labels of that region

- Insufficient number of training records in the region causes the decision tree to predict the test examples using other training records that are irrelevant to the classification task

Notes on Overfitting

- Overfitting results in decision trees that are more complex than necessary

- Training error no longer provides a good estimate of how well the tree will perform on previously unseen records

- Need new ways for estimating errors

Evaluating a Classifier

Accuracy

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Cost Matrix

	PREDICTED CLASS	
ACTUAL CLASS C(i|j)	**Class=Yes**	**Class=No**
Class=Yes	C(Yes|Yes)	C(No|Yes)
Class=No	C(Yes|No)	C(No|No)

C(i | j): Cost of misclassifying class j example as class i

Computing Cost of Classification

Cost Matrix	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	**+**	**-**
	+	-1	100
	-	1	0

Model M_1	PREDICTED CLASS		
ACTUAL CLASS		**+**	**-**
	+	150	40
	-	60	250

Accuracy = 80%

Cost = 3910

Model M_2	PREDICTED CLASS		
ACTUAL CLASS		**+**	**-**
	+	250	45
	-	5	200

Accuracy = 90%

Cost = 4255

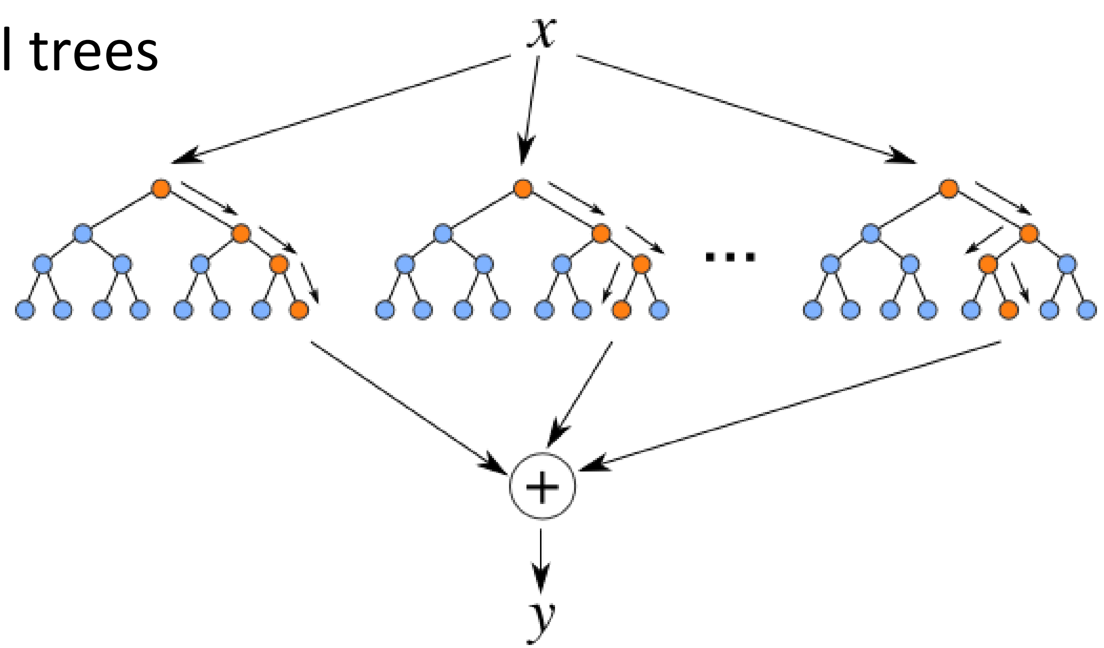
How to Estimate True “Accuracy” (or whatever we’re measuring)

- Holdout
  - Reserve 2/3 for training and 1/3 for testing
- Cross validation
  - Participation data into k disjoint subsets
  - K-fold: training on k-1 partitions, test the remaining one
- Bootstrap
  - Sampling with replacement
  - https://machinelearningmastery.com/statistical-sampling-and-resampling/

Random Forests

- Construct decision trees on bootstrap replicas
  - **Bootstrap replication:** Given *n* training examples, construct a new training set by sampling *n* instances with replacement
  - Restrict the node decisions to a small subset of features picked randomly for each node

- Average the output of all trees



Exam1

- The exam will be held at 2pm this Friday in the class time. The exam sheet will be released at 2pm on Canvas.
- You may print out the exam sheet or use blank papers to write your answers. Then, **scan/take a picture and submit it to Canvas**. Make sure your scanned doc/picture have high resolution so that we can clearly see and grade it.

- The exam is closed book.

- You may prepare and use one standard 8.5" by 11" piece of paper with any notes you think appropriate or significant (use only *one-side*).

- You may use a calculator if it make you feel comfortable. But no other electronic devices are allowed (e.g., cell phone, tablet and computer).

Previous Class…

Attribute and Data Object

Previous Class…

Attribute and Data Object

Types of Attributes
→ Nominal, Ordinal, and Quantitative

Previous Class…

Attribute and Data Object

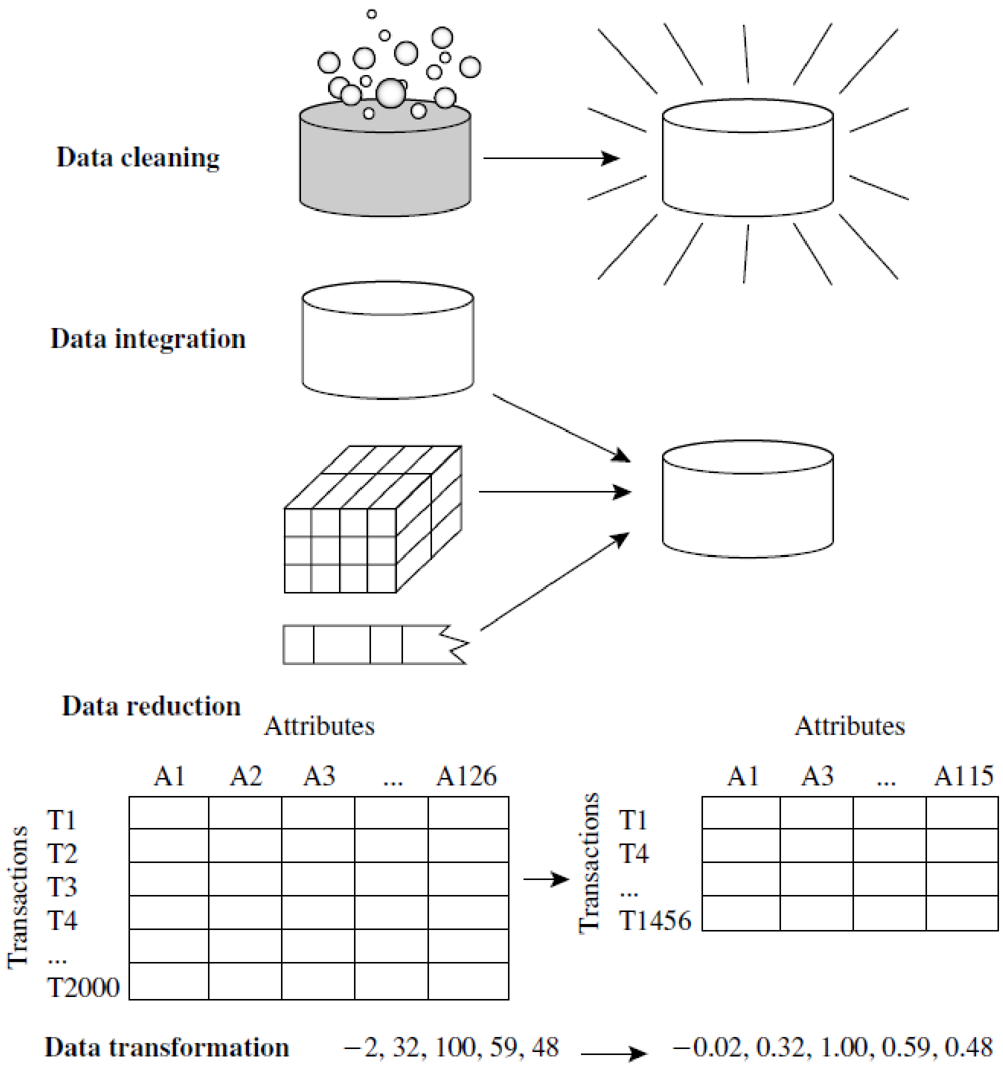
Types of Attributes
→ Nominal, Ordinal, and Quantitative

Measuring the Central Tendency
→ Mean, Median, Mode

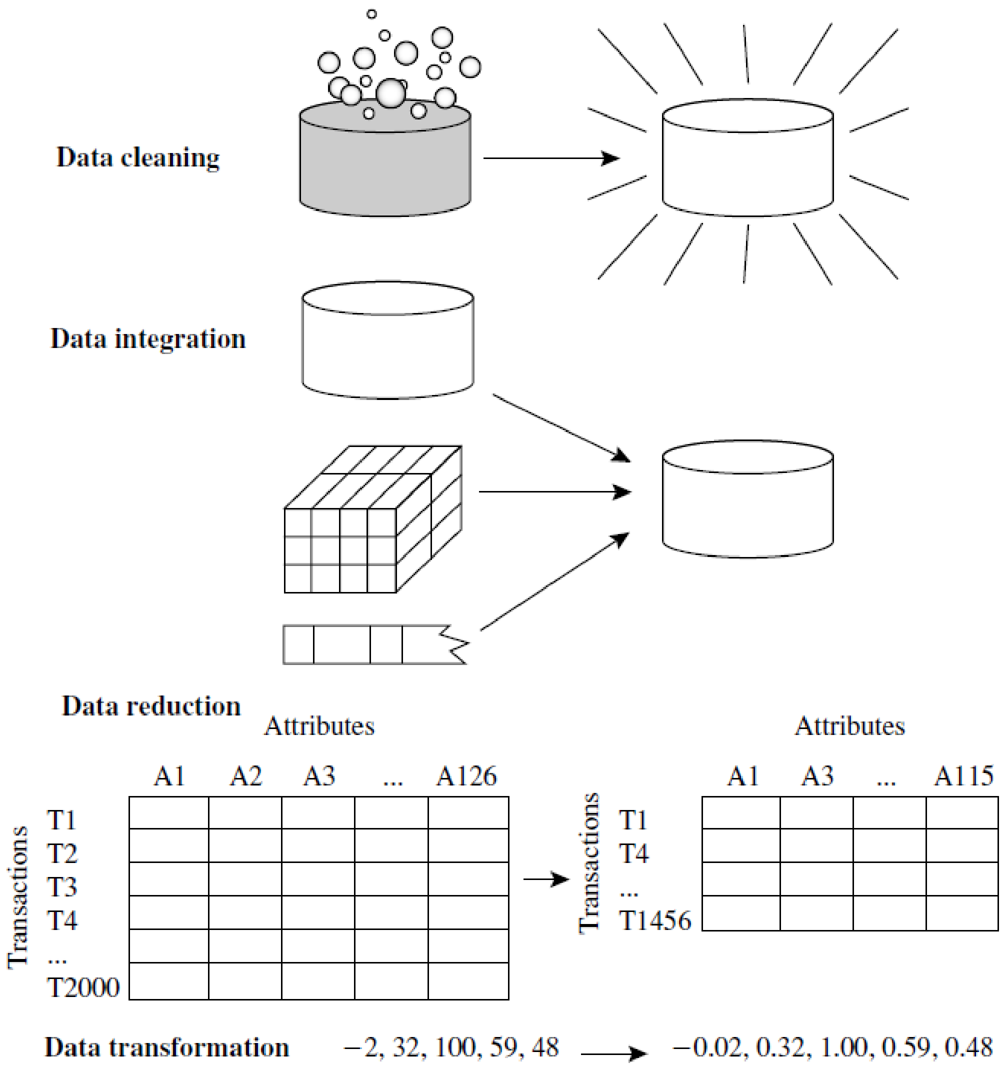
Previous Class…

Measuring the Dispersion of Data
➔ Quartiles, outliers and boxplots

Previous Class…



Previous Class…



Previous Class…

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases/data sources, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization

Previous Class…

Correlation Analysis
→ Correlation Coefficient

Previous Class…

Correlation Analysis
→Correlation Coefficient

Classification: Decision Tree

Previous Class…

Correlation Analysis
→Correlation Coefficient

Classification: Decision Tree

GINI Index

Previous Class…

Linear Regression

Data Science: The Context

Ask question: What data needs to be recorded? or collected?

Why? What research question am I going to answer?

What do I want it to look like?

Real World

Humans behaving
Biology
Finance
Internet
Medicine
Sociology
Olympics

Raw Data is Collected / Recorded

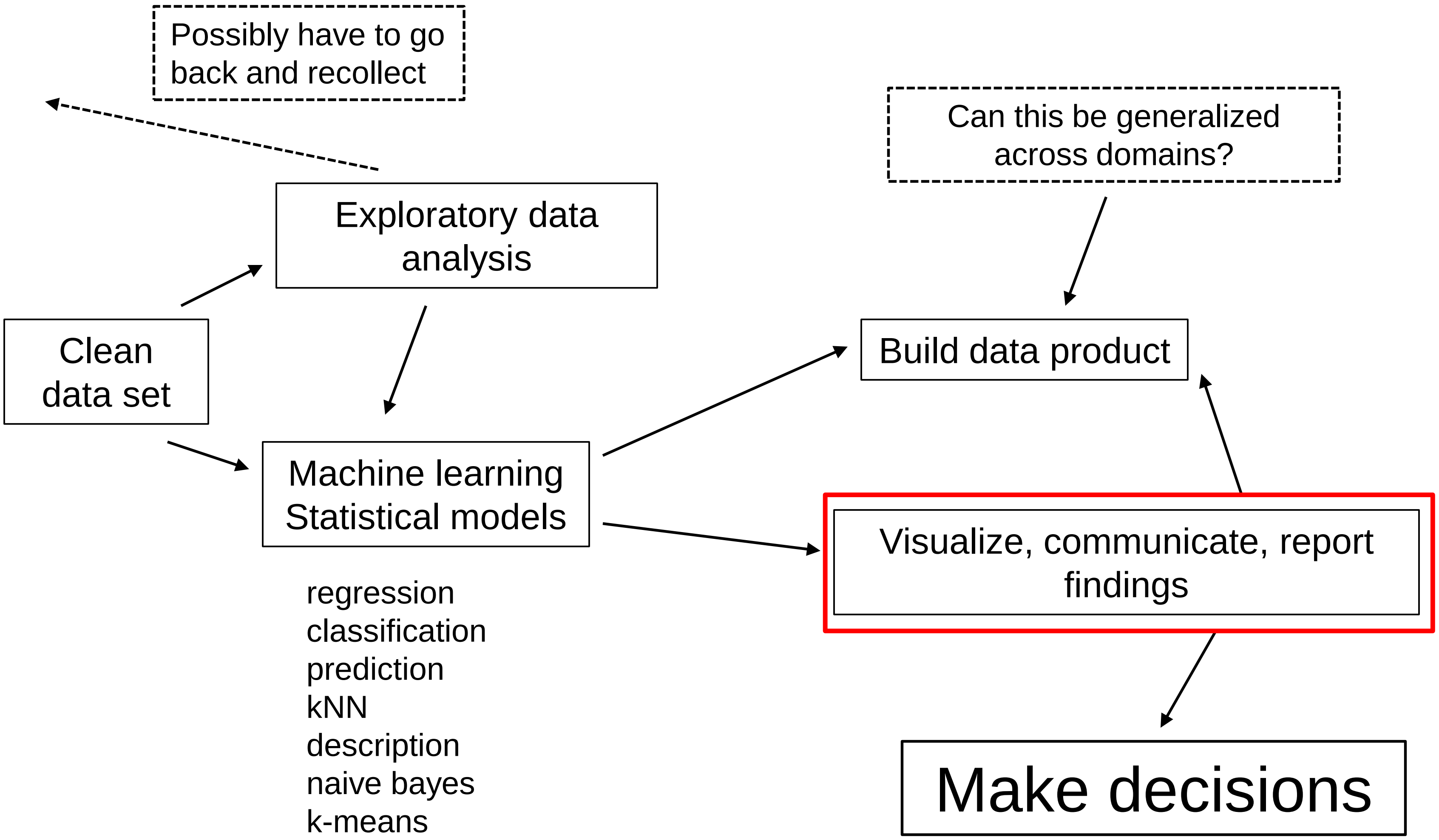
email
logs
medical records
surveys
blood drawn
(microarray)
olympic records
NYT web pages

Data is Processed

pipelines
web scraping
cleaning
munging
joining
wrangling

Data Set

“clean” table



Possibly have to go back and recollect
Can this be generalized across domains?
Exploratory data analysis
Clean data set
Build data product
Machine learning
Statistical models
Visualize, communicate, report findings
regression
classification
prediction
kNN
description
naive bayes
k-means
Make decisions

Data Visualization

What is visualization?

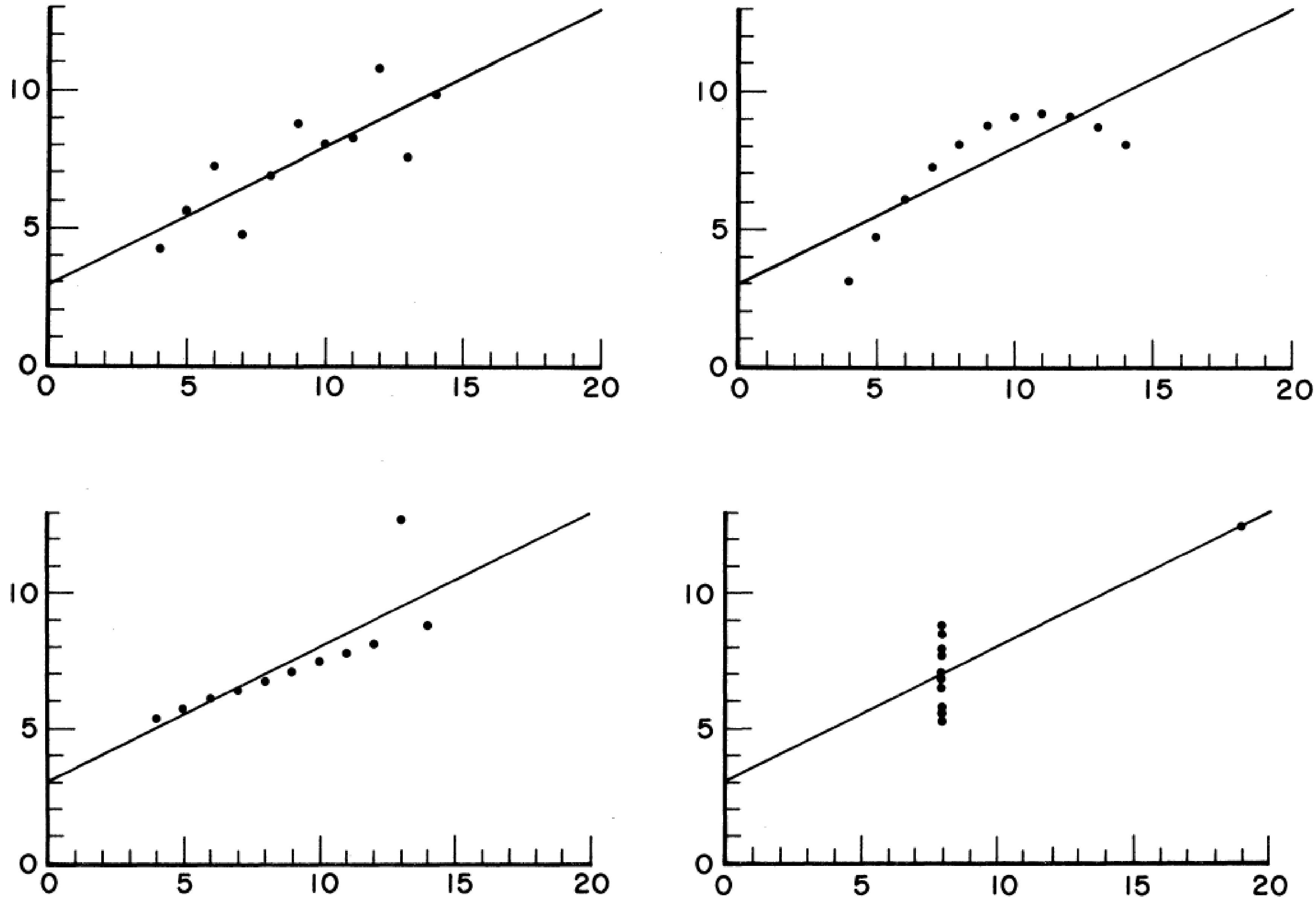
- “Transformation of the symbolic into the geometric” [McCormick et al. 1987]
- “... finding the artificial memory that best supports our natural means of perception.” [Bertin 1967]
- “The use of computer-generated, interactive, visual representations of data to amplify cognition.” [Card, Mackinlay, & Shneiderman 1999]

Four Datasets

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Anscombe “Graphs in Statistical Analysis” 1973

Number of observations (n) $=$ 11
Mean of the x's ($\bar{x}$) $=$ 9.0
Mean of the y's ($\bar{y}$) $=$ 7.5
Regression coefficient (b_1) of y on x $=$ 0.5
Equation of regression line: $y = 3 + 0.5\,x$
Sum of squares of $x - \bar{x} = 110.0$
Regression sum of squares $=$ 27.50 (1 d.f.)
Residual sum of squares of y $=$ 13.75 (9 d.f.)
Estimated standard error of $b_1 = 0.118$
Multiple $R^2 = 0.667$


10
5
0
0 5 10 15 20
10
5
0
0 5 10 15 20
10
5
0
0 5 10 15 20
10
5
0
0 5 10 15 20

Why Do We Create Visualizations?

- Answer questions (or discover them)
- Make decisions
- See data in context
- Expand memory
- Support graphical calculation
- Find patterns
- Present argument or tell a story
- Inspire

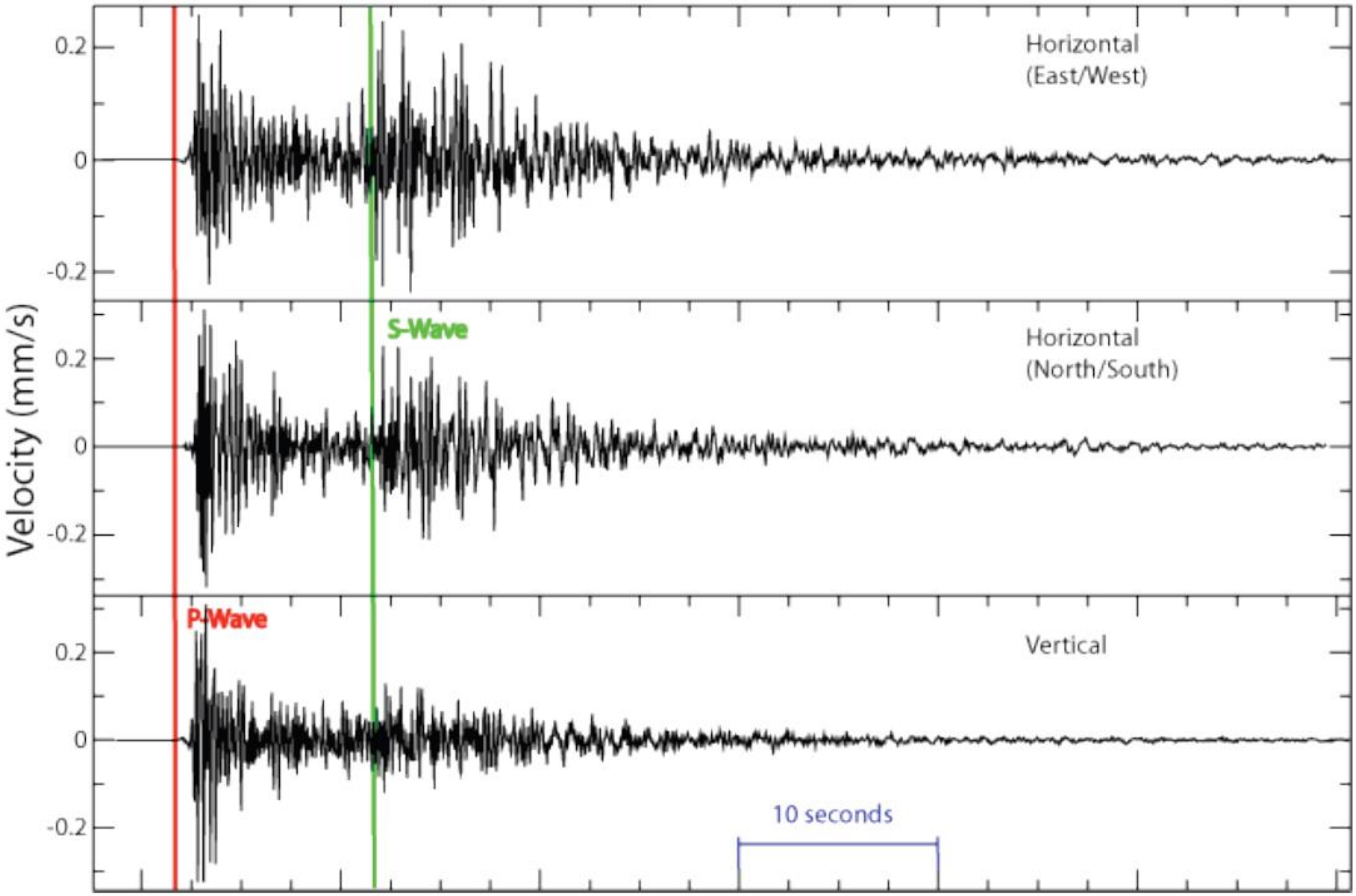
Jeff Heer

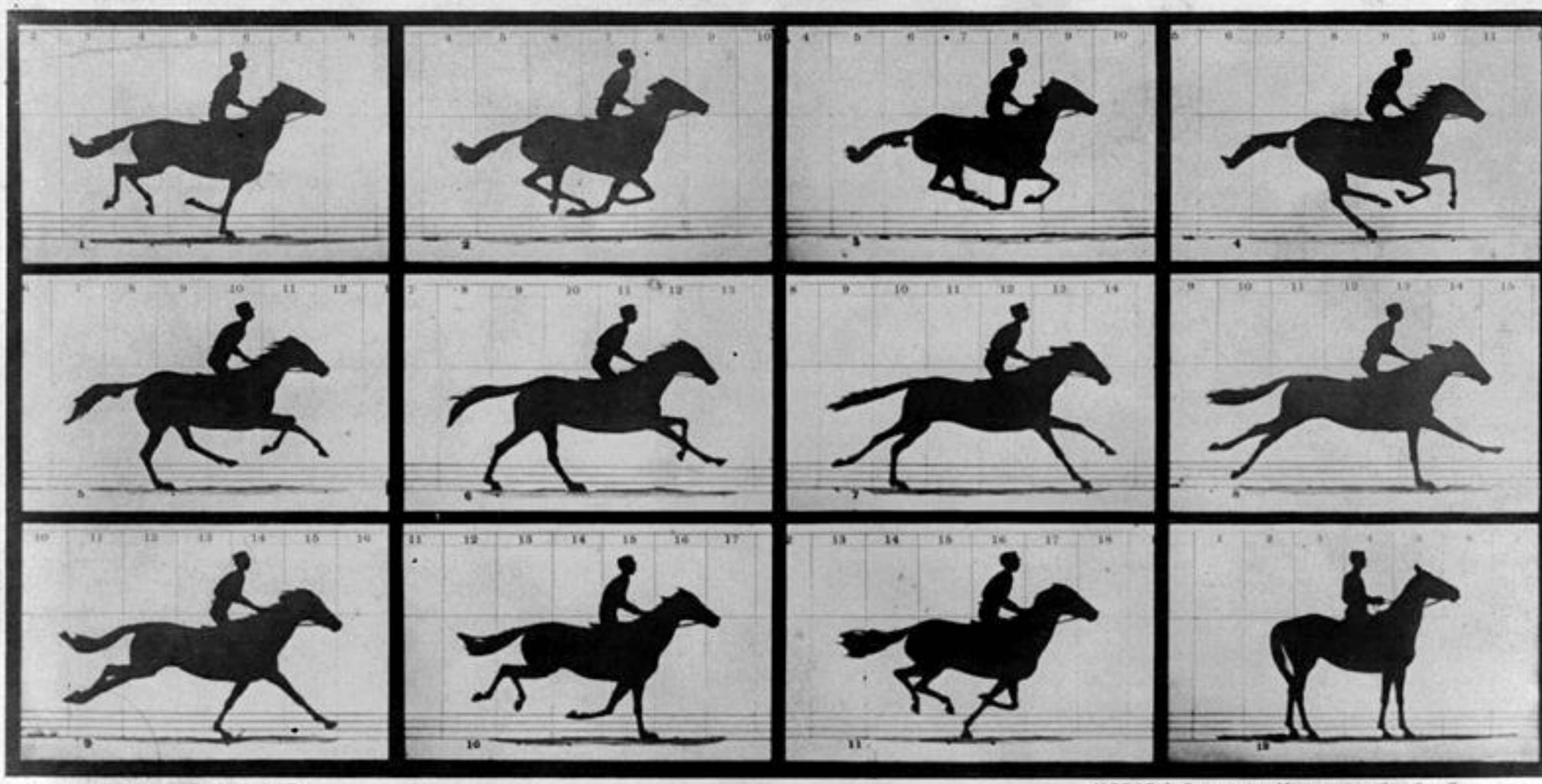
The Value of Visualization

- **Record** information
  - Blueprints, photographs, seismographs, …
- **Analyze** data to support reasoning
  - Develop and assess hypotheses
  - Discover errors in data
  - Expand memory
  - Find patterns
- **Communicate** information to others
  - Share and persuade
  - Collaborate and revise

Record Information

Seismic waves





MORSE'S Gallery, 417 Montgomery St., San Francisco.

THE HORSE IN MOTION.

Illustrated by

MUYBRIDGE.

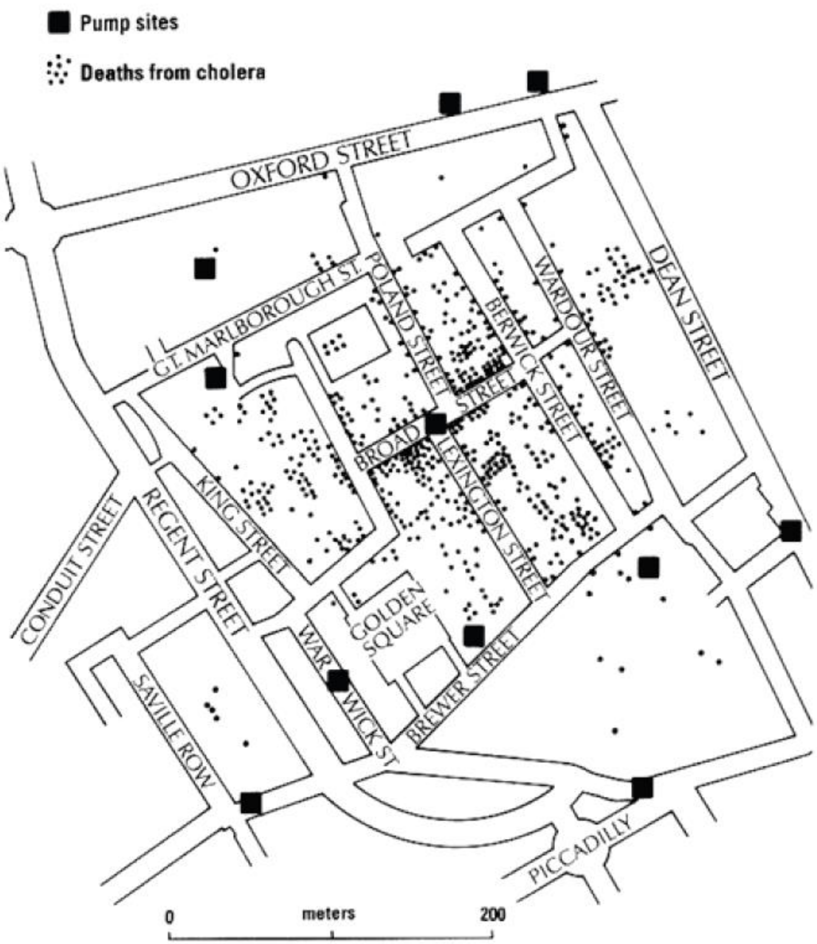
AUTOMATIC ELECTRO-PHOTOGRAPH.

"SALLIE GARDNER," owned by LELAND STANFORD; running at a 1.40 gait over the Palo Alto track, 19th June, 1878.

The negatives of these photographs were made at intervals of twenty-seven inches of distance, and about the twenty-fifth part of a second of time; they illustrate consecutive positions assumed in each twenty-seven inches of progress during a single stride of the mare. The vertical lines were twenty-seven inches apart; the horizontal lines represent elevations of four inches each. The exposure of each negative was less than the two-thousandth part of a second.

Analyze

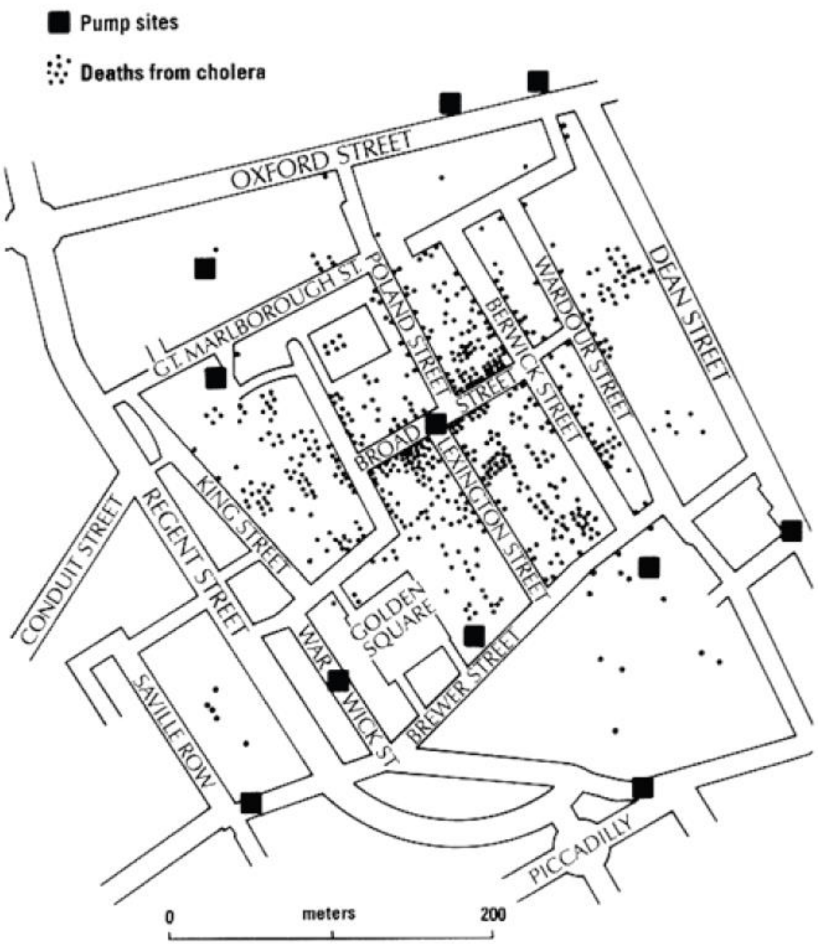
Data in context: Cholera outbreak



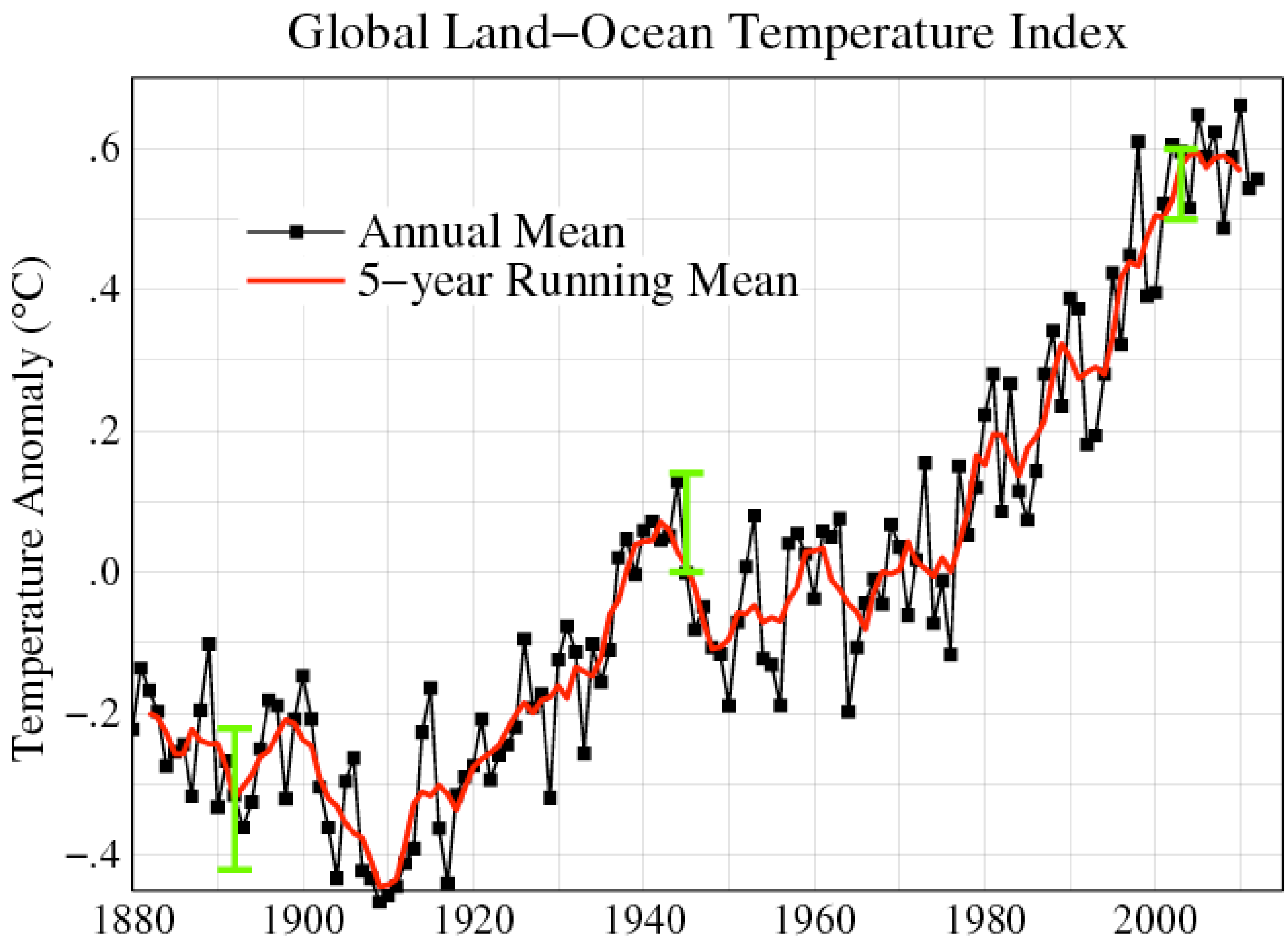
In 1854 John Snow plotted the position of each cholera case on a map. [from Tufte 83]

https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

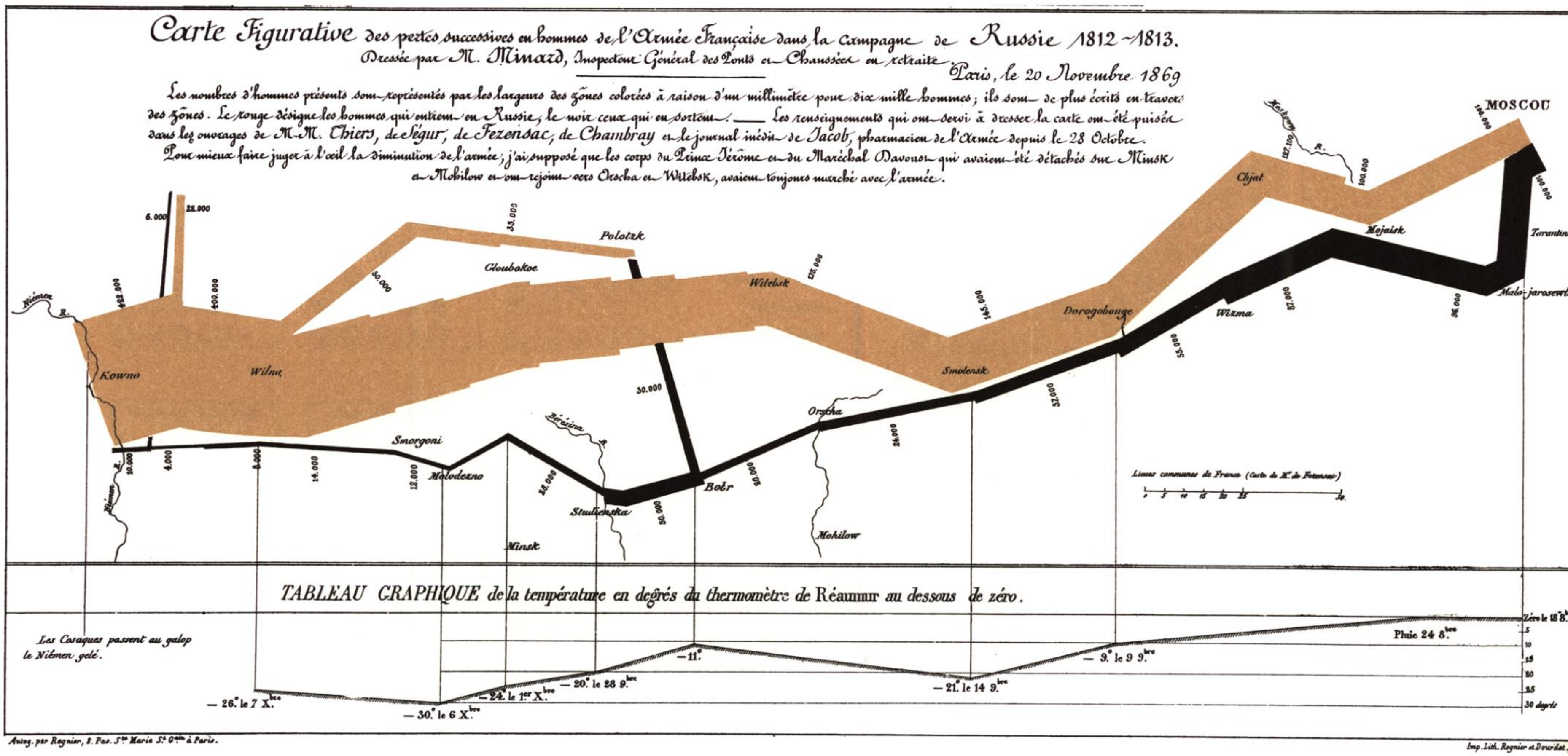
Data in context: Cholera outbreak



Used map to hypothesize that pump on Broad St. was the cause. [from Tufte 83]

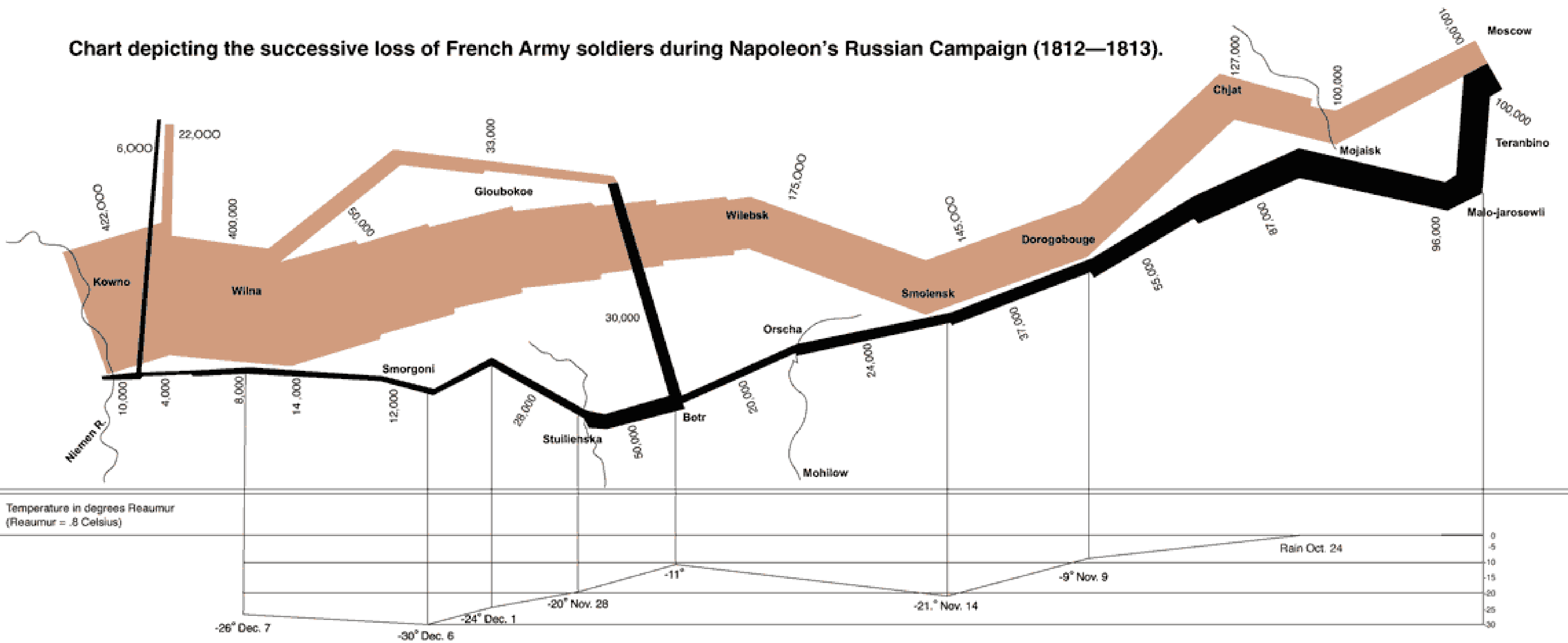

Global Land−Ocean Temperature Index
Annual Mean
5−year Running Mean
Temperature Anomaly (°C)
.6
.4
.2
.0
−.2
−.4
1880
1900
1920
1940
1960
1980
2000

Communicate Information to Others

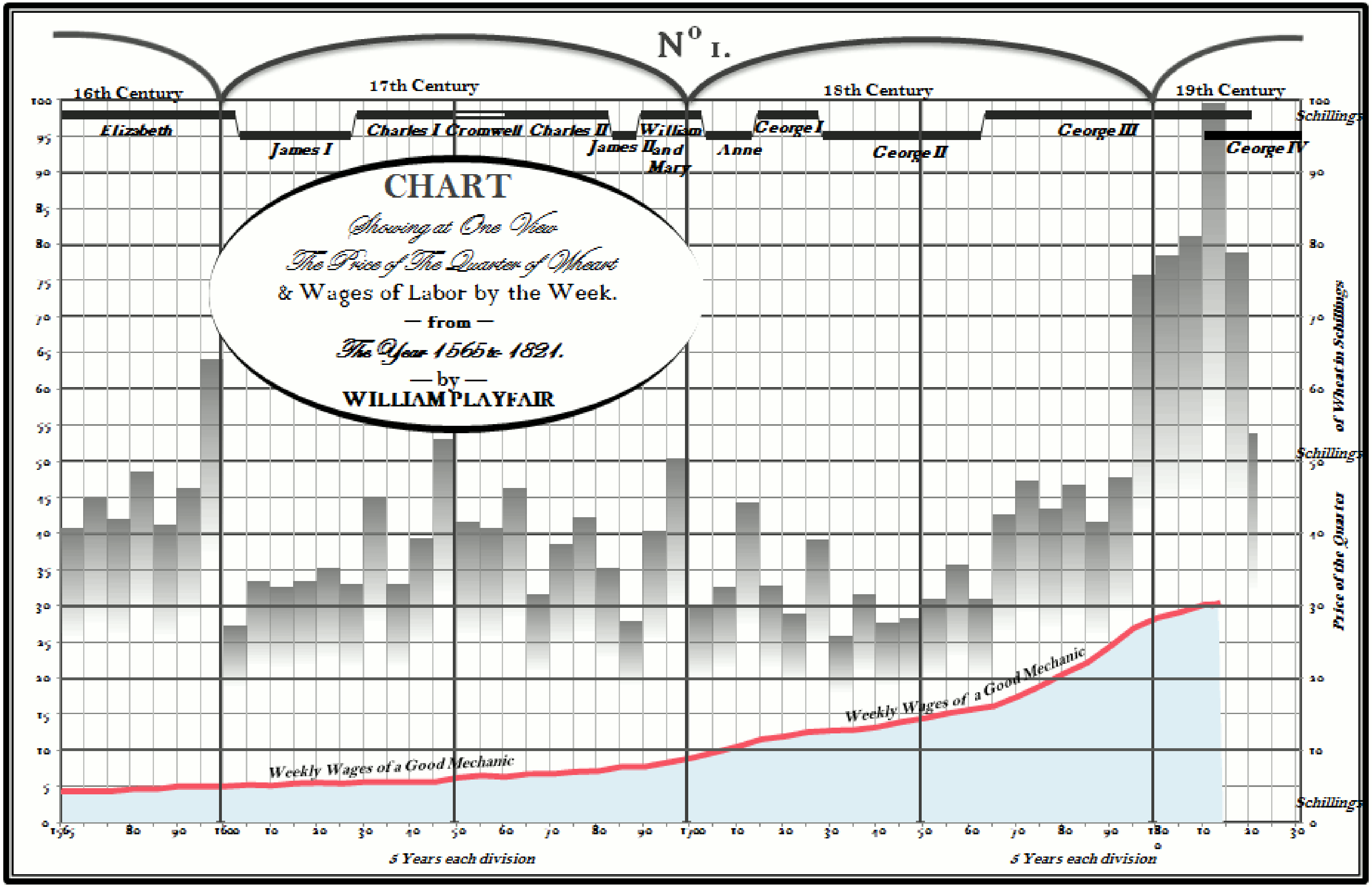


Napolean's invasion of Russia, as drawn by Charles Joseph Minard (1781-1870)

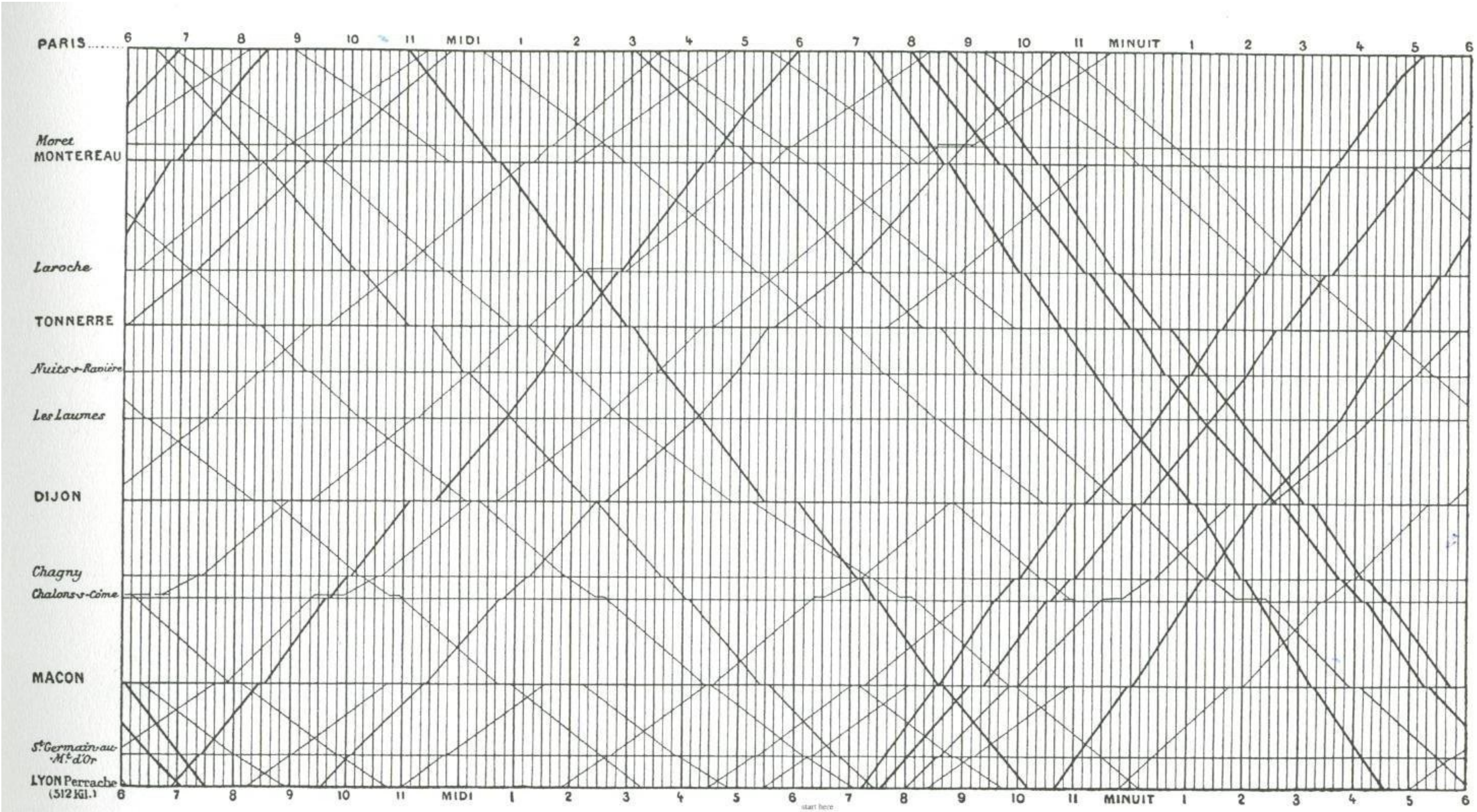
Show Space, Time and Temperature



Display the size of the army, its location on a two-dimensional surface, direction of the army's movement and temperature



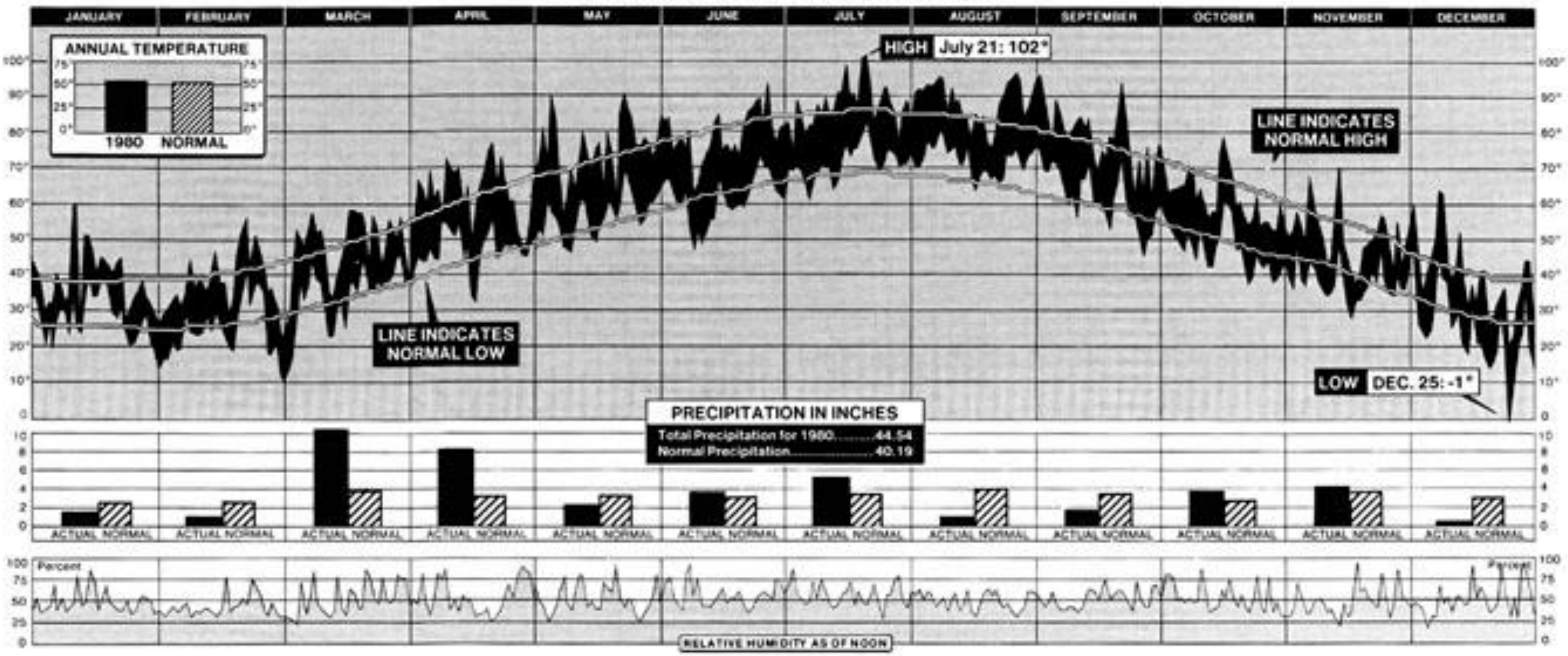
The price of wheat compared to labor wages, William Playfair (1759-1823)



E. J. Marey, *La méthode graphique* (Paris, 1885), 20. The method is attributed to the French engineer, Ibry.

French train schedule, as drawn by E.J. Marey (1830-1904)

NEW YORK CITY'S WEATHER FOR 1980



Reasons?

- Lots of data -- compact representation
- Identify what is being represented
  - Data clarity
- Choice of presentation matters (pie chart vs. time series vs. map ...)
- Easy to compare / contrast (ANALYZE)
- Multi-data types

Tufte: Principles of Graphical Excellence

- Graphical excellence is the well-designed presentation of interesting data – a matter of *substance*, *statistics*, and *design*

- Graphical excellence consists of complex ideas communicated with *clarity*, *precision*, and *efficiency*