

About Data



1. Natural Language Processing

2. Data Types

3. Resources



The image shows a dark, futuristic digital interface. In the center, there is a glowing blue circular element with the letters 'NLP' in white. The background is filled with intricate, glowing blue circuitry and data lines, creating a high-tech, cybernetic feel. The overall color palette is dominated by dark blues and blacks, with bright blue highlights from the digital elements.

NLP

Natural language processing is the ability to take a body of text and **extract meaning** from it using a computer.

Natural Language Challenge: Human vs Computer

HUMAN

“I drove my friend Mary to the park in my Tesla while listening to music on my iPhone.”



Humans understand that Mary is a friend and that a Tesla is likely a car. Additionally, after many years of popularity and cultural references, we all know that an iPhone is a smartphone.

MACHINE

Structured data

{
<friend>Mary</friend>
<car>Tesla</car>
<phone>iPhone</phone>

None of this is understood by a computer without assistance.



NLP Components

Let's analyze the phrase ...

"I drove my friend Mary to the park in my Tesla while listening to music on my iPhone"

Entities

The people, places, organizations, and things in your text.

Example: friend, car, and phone

Relations

How entities are related.

A "createdBy" relation might connect the entities "iPhone" and "Apple."

Concepts

Extracting reference to topics that do not explicitly appear in the text.

An article about Tesla may refer to concepts "electric cars" or "Elon Musk," even if those terms are not explicitly mentioned.

Keywords

Identify the important and relevant keywords in your content.

Semantic Roles

Subjects, actions, and objects in the text.

"IBM bought a company." The subject is "IBM," the action is "bought," and the object is "company."

Categories

Describing what a piece of content is about at a high level.

Categories could be sports, finance, travel, computing, and so on.

Emotion

Understanding the emotion or tone conveyed.

Is the content conveying anger, disgust, fear, joy, or sadness?

Sentiment

Is the feeling/attitude positive, neutral, or negative?

The level of positive or negative sentiment can be scored.

Natural Language Processing

The semantic
behind the syntactic

Syntactic messages

Subject-verb-object

Semantic messages

Agent and Patient

What is the message sentiment?



* Search for 'moving' at <http://www.shoecomics.com/>

Natural language processing: a classification problem

- Difficulty of language: Subtleties, idiosyncrasies, idioms, ambiguities, nuances and gaps

Noses run	but	feet smell
Slim chance	=	Fat chance
Wise man	not	Wise guy
Alarm goes off	while	it goes on
Fill in a form	by	filling it out
House burns up	while	it burns down
Ship by truck	but	send cargo by ship

- It is highly contextual, imprecise and has gaps (context known outside the conversation)

Syntactic versus the Semantic



Syntax is easy: parse, annotate and tokenize the sentence.

Semantics is harder:

- recognize the situation, type and roles of the two agents,
- Relate the word 'thing' to the picture and the concept car,
- Relate the words 'take' and 'move' to the situation.
- Pragmatics is the hardest: explain the irony and the humor.

* Search for 'moving' at <http://www.shoecomics.com/>

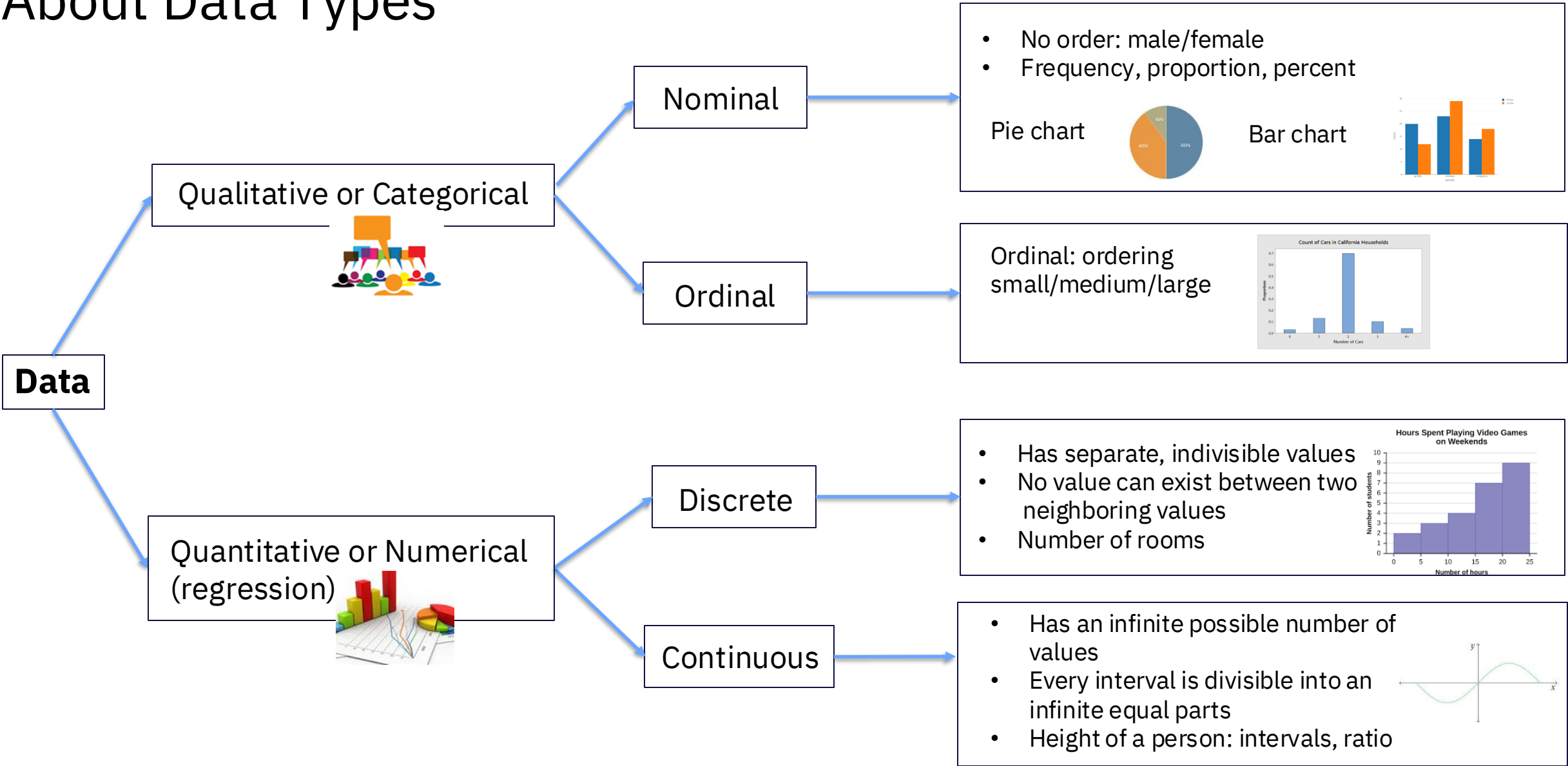
LECTURE 2

About Data

1. Natural Language Processing
- ▶ **2. Data Types**
3. Resources



About Data Types



Characteristics of data

The 5 Vs

Volume

Refers to the vast amounts of data generated every second.

Variety

Refers to the different types of data we can now use.

Velocity

Refers to the speed at which new data is generated and the speed at which data moves around.

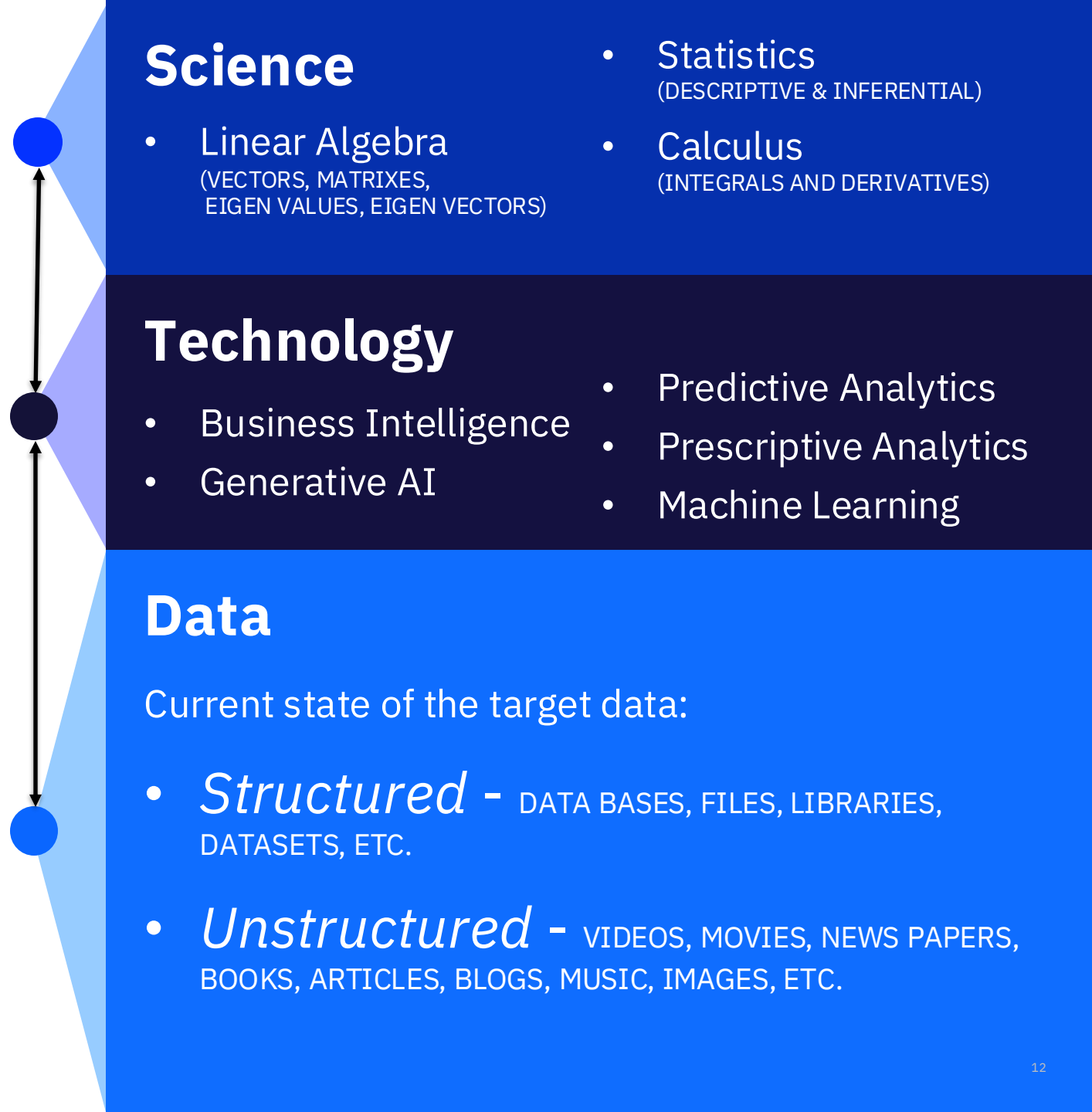
Veracity

Refers to the messiness or trustworthiness of the data.

Value

Refers to having access to big data is no good unless we can turn it into value.

Science, Technology and **Data** are intrinsically connected



Data cleansing

Data **Analysts** can spend up to **80% of the time cleaning data.**

- Importing data
- Joining multiple datasets
- Detecting missing values
- Detecting anomalies
- Imputing for missing values
- Data quality assurance

Tidy data lends itself to efficient data analysis and processing

“Tidy datasets are all alike, but every messy dataset is messy in its own way.”

– Hadley Wickham

- **Transforming your data into standard format**, or tidy data, makes analysis and storage easier down the road
- Additionally, one must make sure the data are in its **appropriate type**

What is tidy data?

Tidy data satisfies 3 components

1. Each variable forms a column.

Brand	Size	Price
McDonald's	Small	1.00
Dunkin Donuts	Small	1.99
Starbucks	Small	2.99
McDonald's	Medium	1.00
Dunkin Donuts	Small	2.49
Starbucks	Medium	3.49

2. Each observation forms a row.

Brand	Size	Price
McDonald's	Small	1.00
Dunkin Donuts	Small	1.99
Starbucks	Small	2.99
McDonald's	Medium	1.00
Dunkin Donuts	Small	2.49
Starbucks	Medium	3.49

3. Each type of observational unit forms a table.

Brand	Size	Price
McDonald's	Small	1.00
Dunkin Donuts	Small	1.99
Starbucks	Small	2.99
McDonald's	Medium	1.00
Dunkin Donuts	Small	2.49
Starbucks	Medium	3.49

Example of tidy data

Not so tidy.....

Brand	Size	Price
McDonald's	Small	1.00
Dunkin Donuts	Small	1.99
Starbucks	Small	2.99
McDonald's	Medium	1.00
Dunkin Donuts	Medium	2.49
Starbucks	Medium	3.49

Tidy data....also known as feature engineering

Size	McDonald's	Dunkin Donuts	Starbucks
Small	1.00	1.99	2.99
Medium	1.00	2.49	3.99

From outlier checking to feature engineering to missing value imputation

Column headers are values, not variable names.

Multiple variables are stored in one column.

Variables are stored in both rows and columns.

Multiple types of observational units are stored in the same table.

A single observational unit is stored in multiple tables.

Missing Data

Explicit

- Marked with NULL or NA
- Use summary functions, eg, `summary()`, as its output will list each variable's count of missing values

Implicit

- Not there at all
- You must explore and visualize your data to notice things that appear 'off'

Methods for handling missing data

Remove the observation completely

Imputation methods

- Replace with summary statistic such as mean, median, or mode
- Create a new variable that flags a missing column
- Replace NA with an outlier (tree-based models will implicitly understand that these outliers are associated with missing values)

Statistical analysis

The data representation phase should leverage mathematical tools such as:

- **Statistics**
- **Correlations**
- **Chi-squared tests**

Descriptive statistics allow you to describe a vast, complex data set using just a few key numbers.

You can also create a table that displays summarized statistics for cases grouped by categorical data based on a single measure.

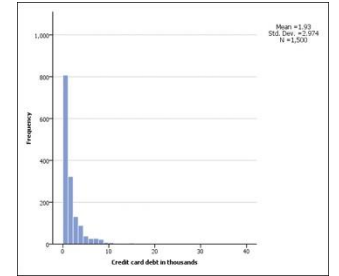
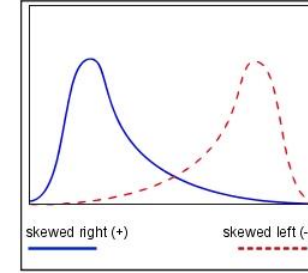
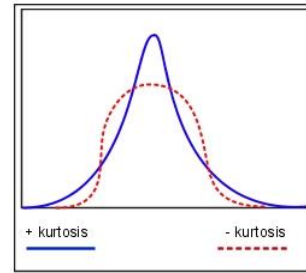
The table below shows the mean household income for customers grouped by education level.

Descriptive Statistics					
	High school degree	Post-undergraduate degree	Did not complete high school	Some college	College degree
Mean	52.00	99.71	51.48	56.90	70.94
Std. Deviation	56.370	147.769	51.855	53.836	67.940
N	527	84	246	333	310
Median	35.00	59.50	36.00	39.00	49.00
Minimum	12	16	15	13	15
Maximum	533	1,079	497	403	512

Descriptive Statistics Quantitatively Summarize a Data Set

You can use descriptive statistics to:

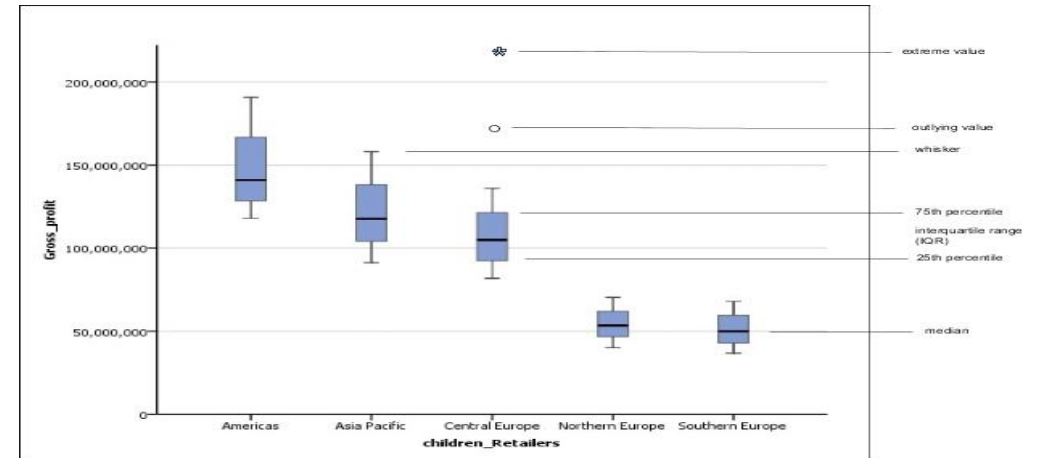
- Look at averages, such as the **mean** or **median**.
- Obtain information, such as the **mean for groups of interest**, that you might need to interpret other statistical tests.
- Provide graphical representations of data, such as **histograms** and **boxplots**.



Descriptive Tables

- Measures of Central Tendency
- Measures of Dispersion
- Measures of Distribution

Histograms



Boxplots

Variance vs Standard Deviation

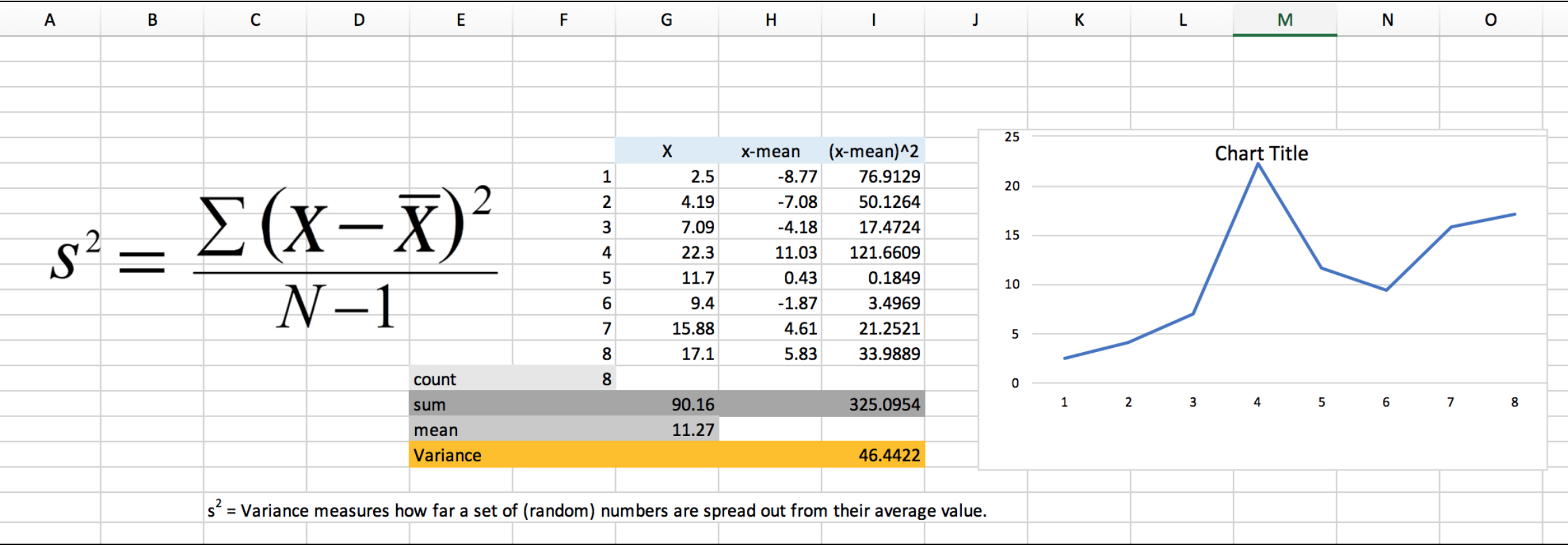
Variance measures the average degree to which each point differs from the mean.

The greater the variance, the larger the overall data range. It is a good way to spot the outliers and gives you an idea of the **overall spread**.

Standard deviation is the square root of the variance.

The calculation of variance uses squares because it weights outliers more heavily than data very near the mean.

It starts with calculating the variance



Risk of using summary statistics

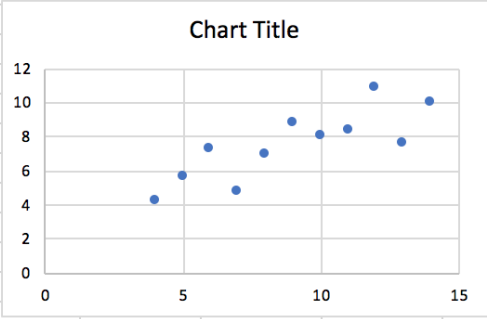
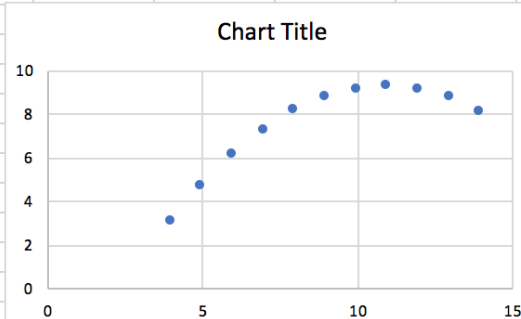
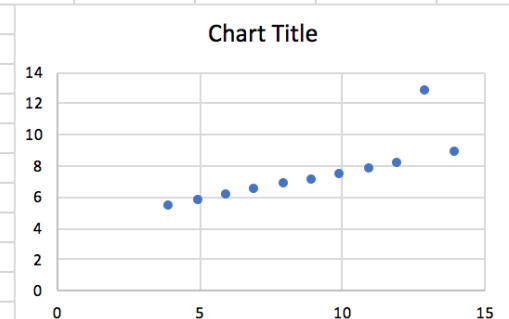
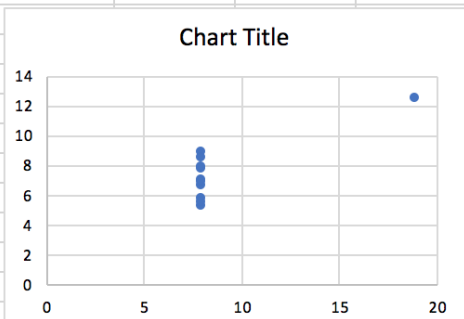
There's a **danger** in relying only on summary statistics and **ignoring the overall distribution**.

Anscombe's quartet offers a classic example of this risk.

It comprises four data sets of 11 x, y points that have nearly identical simple statistical properties, yet appear very different when graphed.

Anscombe's Quartet

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Sources:	Edward R. Tufte, <i>The Visual Display of Quantitative Information</i> (Cheshire, Connecticut: Graphics Press, 1983), pp. 14-15. F.J. Anscombe, "Graphs in Statistical Analysis," <i>American Statistician</i> , vol. 27 (Feb 1973), pp. 17-21.																
Observation	x1	y1				x2	y2				x3	y3			x4	y4	
1	10	8.04				10	9.14				10	7.46			8	6.58	
2	8	6.95				8	8.14				8	6.77			8	5.76	
3	13	7.58				13	8.74				13	12.74			8	7.71	
4	9	8.81				9	8.77				9	7.11			8	8.84	
5	11	8.33				11	9.26				11	7.81			8	8.47	
6	14	9.96				14	8.1				14	8.84			8	7.04	
7	6	7.24				6	6.13				6	6.08			8	5.25	
8	4	4.26				4	3.1				4	5.39			19	12.5	
9	12	10.84				12	9.13				12	8.15			8	5.56	
10	7	4.82				7	7.26				7	6.42			8	7.91	
11	5	5.68				5	4.74				5	5.73			8	6.89	
Sum		99	82.51			99	82.51				99	82.5			99	82.51	
count(n)	11																
avgerage (mean)		9	7.50090909			9	7.50090909				9	7.5			9	7.50090909	
variance s ²		11	4.12726909			11	4.12726909				11	4.12262			11	4.12324909	

Using representation to help us detect bad data

Univariate analysis:

- Histograms allow one to understand the distribution of a variable

Bivariate analysis:

- Bar charts and Box Plots can help an analyst compare groups
- Line charts help an analyst understand trends over time
- Scatterplots are best used to understand relationships of values

Categorical variables must be mapped to a number in order to be used by a machine learning model

Nominal variables:

- colors
- animal species
- countries

Ordinal:

- rankings
- socioeconomic status

How do you represent categories as numbers?

Naïve approach:

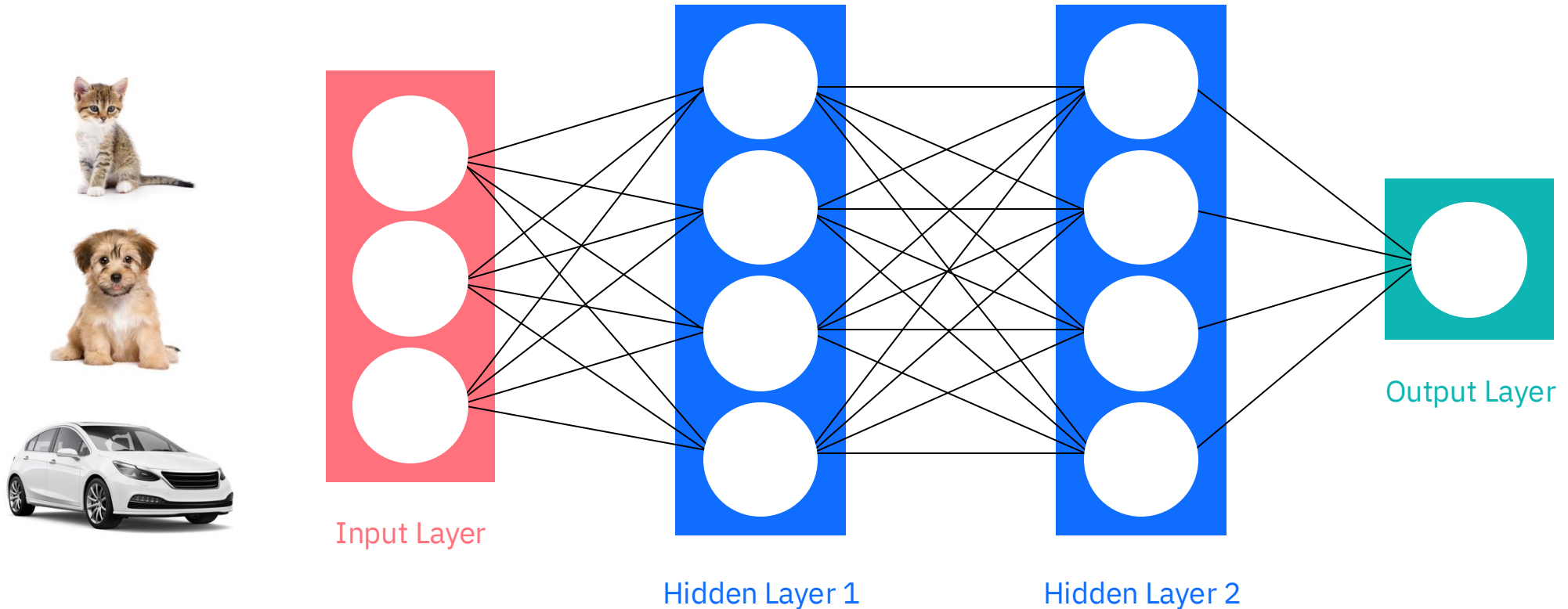
Map a category to a number: [red, blue, green] = [1, 2, 3]

This is misleading!

By mapping the data like this, you are implying green is 3x greater than red.

How to make use of 1-hot encoding in machine learning

Labeled Input → Supervised Learning



What if your data comprise images, tweets, videos?



$[x, x, x]$

$[1, 0, 0]$



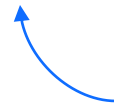
$[x, x, x]$

$[0, 1, 0]$



$[x, x, x]$

$[0, 0, 1]$



Index or element in a vector

1-Hot Encoding allows you to encode categorical data into numbers

1-Hot Encoding:

- Create a matrix of 0's and 1's
- Make each category a column in a table
- **WARNING:** *you must encode (n-1) categories you have in the variable*
 - Otherwise, you will have *perfect multicollinearity* and encounter mathematical issues

Cons:

- This makes the data very large and sparse
- Better methods for text data
- Does not scale well to big data

Sample	Species
1	Human
2	Human
3	Penguin
4	Octopus

Solutions:

- Bag of words / TF-IDF for text data
- Advanced algorithms such as neural networks with embedding

Sample	Human	Penguin	Octopus
1	1	0	0
2	1	0	0
3	0	1	0
4	0	0	1

Python: data science language

Python remains the leader for data science because of the massive scope of libraries that have been developed for it:

SciPy: A set of scientific computing tools for Python that provide many performant numerical routines covering optimization, integration, interpolation, and linear algebra

NumPy: Adds support for large, multidimensional arrays and matrices into Python and includes several mathematical functions for use on these structures

scikit-learn: a Python library that includes a variety of machine learning algorithms, such as support vector machines, random forests, and principal component analysis

Natural Language Toolkit: A Python module that allows you to build Python programs that work with human language as data

Pandas: A Python module that offers data structures and operations for manipulating numerical tables and time series

Roles integrated enviroment

[7]
You need framework for **Machine Learning** and likely framework for **Deep Learning**
Scikit, TensorFlow, Keras, Torc, Theano, etc.



[6]
Choose your scientific computing and statistic packages:
SciPy, NumPy are widely utilized



[5]
Choose your visualization and plotting tools:
Matplotlib, PixieDust are top market trends



[4]
Choose your data munging libraries and tools:
Pandas is a very flexible library under the python framework



[3]
Choose your programming language:
python is most versatile, R and Scala are mostly specialized statistical packages



[2]
Now you need a development environment.
• Get **Jupyter Notebooks** (julia+python+R)
• Get it from **Anaconda** (www.continuum.io)



[1]
So you have your collected data:
Is it structured, semi-structured, unstructured, mix?

Text editor	MS Excel	DB2, MS SQL	Hadoop	MongoDB	Watson Studio	Sentiment
CSV	xls	RDBMS	DFS	JSON	images	text

LECTURE 2

About Data

1. Natural Language Processing
2. Data Types
- 3. Resources**



Resources

- [1] Beyond the hype: A guide to understanding and successfully implementing artificial intelligence within your business
<https://www.ibm.com/downloads/cas/8ZDXNKQ4>
- [2] A Practical Guide to Building Enterprise Applications: by Tom Markiewicz and Josh Zheng – Feb 2018 O'Reilly
<https://tmarkiewicz.com/getting-started-with-artificial-intelligence/>
- [3] The New York Times - Nils Nilsson
<https://www.nytimes.com/2019/04/25/obituaries/nils-nilssen-dead.html>
- [4] Why artificial intelligence is enjoying a renaissance
<http://www.economist.com/blogs/economist-explains/2016/07/economist-explains-11>
- [5] How Cognitive Systems Could Redefine The Way Governments Work
<http://www.forbes.com/sites/ibm/2016/09/20/how-cognitive-systems-could-redefine-the-way-governments-work/#1e1ed4f52ff1>
- [6] March of the Machines
<http://www.economist.com/news/leaders/21701119-what-history-tells-us-about-future-artificial-intelligenceand-how-society-should>
- [7] We have to upgrade our skills to match intelligent machines
<http://www.businessinsider.com/how-labor-can-keep-up-with-artificial-intelligence-2016-10?IR=T>
- [8] Forbes & IBM - Intelligent Automation: How AI and Automation are Changing the Way Work Gets Done
<https://www.ibm.com/downloads/cas/RE2XMOLR>