Executive Data Science

Foundations, Applications, Implementation, and Ethical Implications

Dr. Jennifer Priestley, Analytics and Data Science Institute
Professor of Statistics and Data Science

KENNESAW STATE UNIVERSITY

Concepts to be Covered

- The Evolution of Data Science
- Defining Data Science – and Data Scientists
- Demystifying Data Science
- Exercise: Developing an Analytical Plan
- Ethical Considerations in Data Science: Human Subjects and Algorithmic Bias
- Exercise: Ethics Case Study
- Summary and Wrap Up

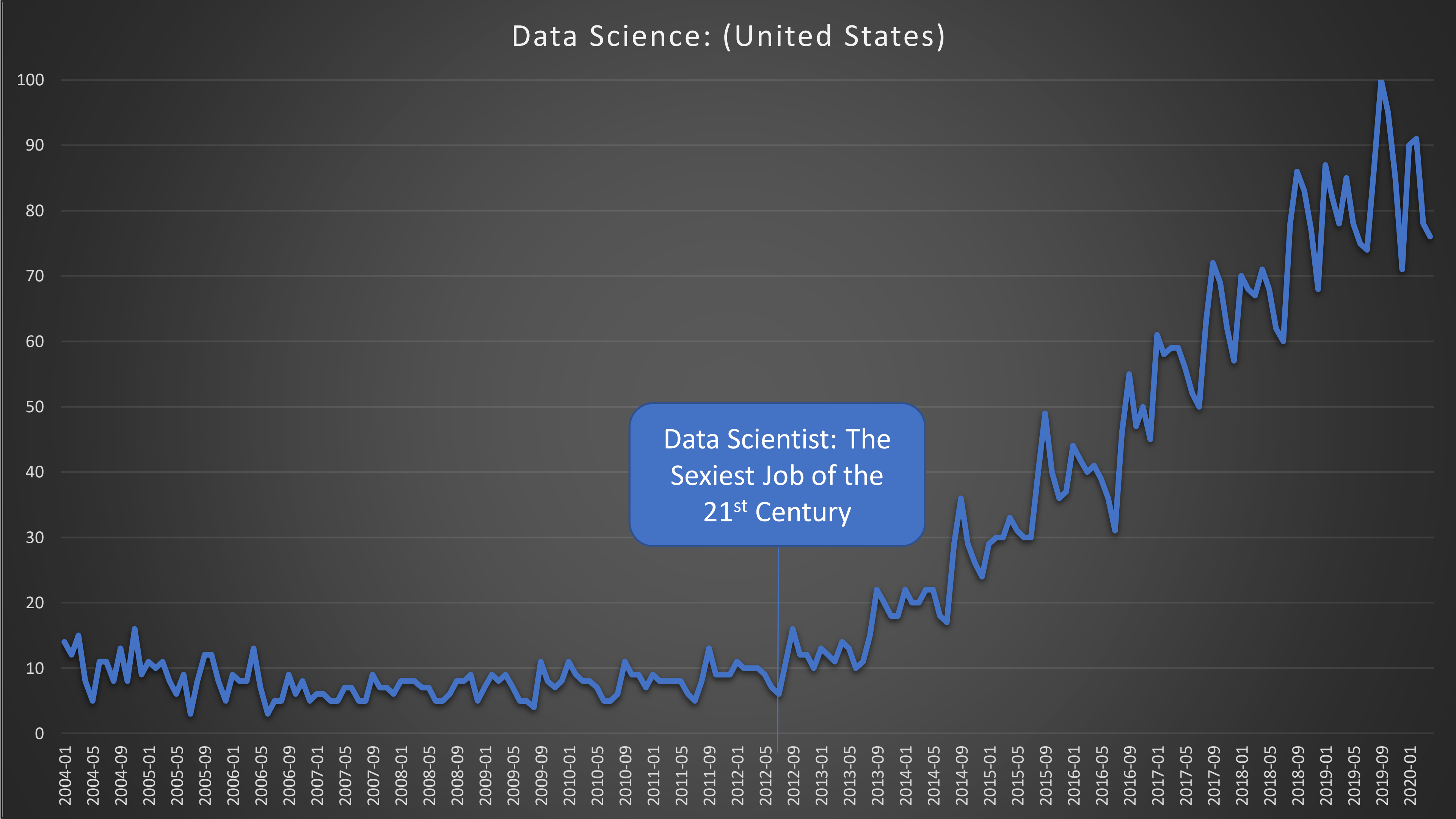
KENNESAW STATE UNIVERSITY

What do you think data science is?

Everyone Wants to Be a Data Scientist…

Finance
Retail
Political Science
Consulting
Healthcare
Economics
Manufacturing

KENNESAW STATE UNIVERSITY



Data Science: (United States)
100
90
80
70
60
50
40
30
20
10
0
Data Scientist: The Sexiest Job of the 21st Century
2004-01
2004-05
2004-09
2005-01
2005-05
2005-09
2006-01
2006-05
2006-09
2007-01
2007-05
2007-09
2008-01
2008-05
2008-09
2009-01
2009-05
2009-09
2010-01
2010-05
2010-09
2011-01
2011-05
2011-09
2012-01
2012-05
2012-09
2013-01
2013-05
2013-09
2014-01
2014-05
2014-09
2015-01
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05
2018-09
2019-01
2019-05
2019-09
2020-01

We have generated more data in the last two years than in the whole of human history.

How the Data Ecosystem is Evolving…

Massive, Integrated, and Dynamic

Artificial Intelligence

Deep Learning

Large, Unstructured, and In Motion

Recommendation Engines

Image Classification

Machine Learning

Small, Structured, and Static

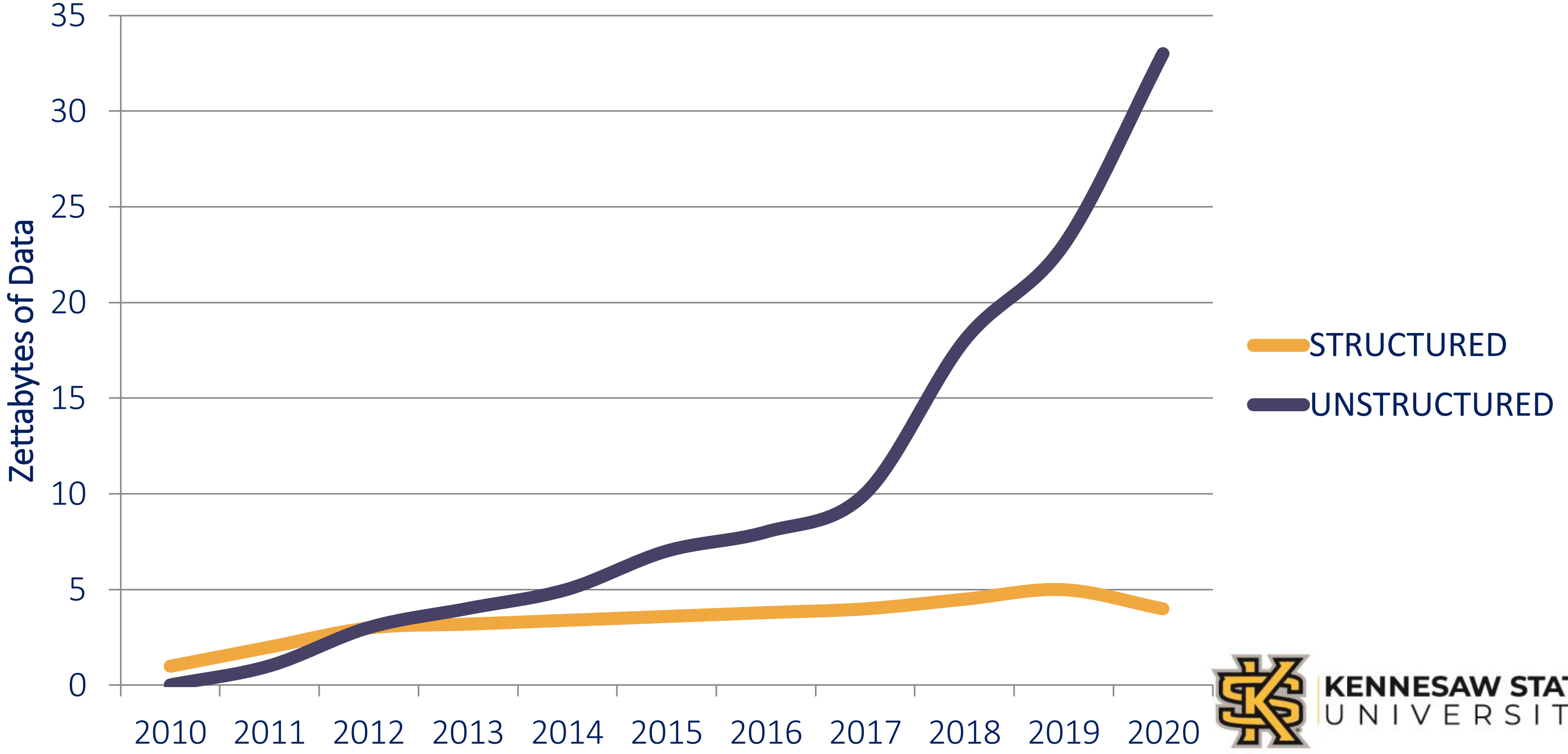
Classification

Predictive Modeling

Descriptive Statistics

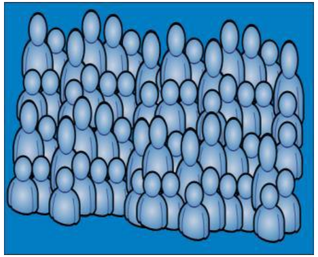
KENNESAW STATE UNIVERSITY

Production of Structured vs Unstructured Data

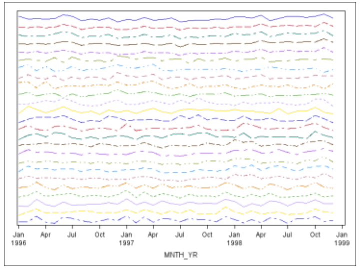


KENNESAW STATE UNIVERSITY

…and the data are different



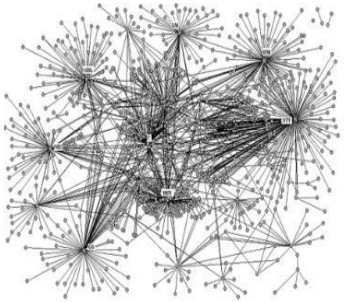
Cross Sectional



Streaming

Image/Video

Geo Spatial



Networks

Text

KENNESAW STATE UNIVERSITY

Who ARE these People?

"*Person who is better at statistics than any software engineer and better at software engineering than any statistician*" – Josh Wills, Director of Data Engineering at Slack

"*Person who is better at explaining the business implications of the results than any scientist and better at science than any business school graduate*"

Jennifer Priestley, Ph.D. Data Nerd

KENNESAW STATE UNIVERSITY

Percentage of Job Ads in 2018 for "Analytics", "Big Data", "Business Analyst" or "Data Scientist" that Contain the Following Key Words…

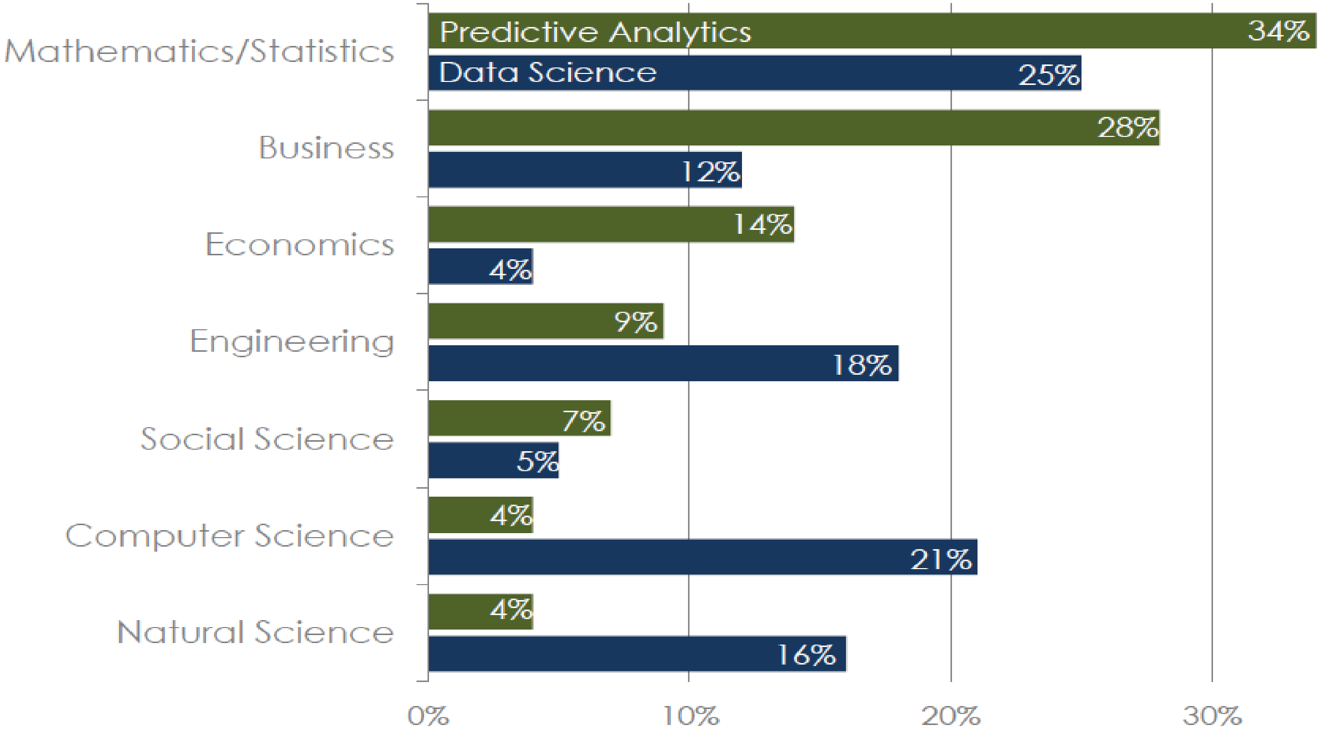
	Analytics, %	Big data, %	Business analyst, %	Data scientist, %
BI software	8.00	4.44	2.86	3.80
Big data	15.12	89.03	0.89	48.53
Business domain	23.04	4.54	36.90	7.22
Business intelligence	24.23	11.72	6.38	22.51
Cloud computing	1.88	6.42	0.19	2.12
Computer science	1.93	1.93	0.05	15.66
Data handling	17.90	34.92	6.73	16.55
Database	39.77	49.26	26.03	50.18
Managerial skills	36.96	13.14	36.63	14.98
Modeling and analysis	42.21	25.51	8.88	77.15
Communication and interpersonal skills	68.70	44.91	61.77	50.50
Programming	20.51	51.84	4.51	54.43
Scripting	15.92	47.10	2.61	62.84
System analysis and design	9.95	33.67	15.82	9.14
Tools	31.53	7.55	19.76	40.94
Web analytics	9.42	0.25	0.57	1.47
Count of job ads	147,525	44,348	365,183	46,368

Data source. Burning Glass Technologies (2018).
Note. Bold values are at least 25%.

KENNESAW STATE UNIVERSITY

What did practicing data scientists study in school (2018)?

Source: Burtchworks.com



Who ARE these People?

Business Disciplines ⟷ Computational Disciplines

Analysts

- ✓ Data Analyst
- ✓ Business/Marketing Analyst
- ✓ Operations/Systems Analyst

Researchers/Scientists

- ✓ Data Engineer
- ✓ Statistician/Modeler
- ✓ Data Architect
- ✓ Machine Learning Engineer

Who ARE these People?

P
o
s
i
t
i
o
n
s

1a – Data Scientist

Industry: Technology
Experience: 5+ Years
Qualifications:
History of applied data mining, ML, statistical modeling to solve business problem and delivered results
Self-driven individual, demonstrating continuous learning and creativity, and is naturally collaborative
Coding (Python or R)
Advanced SQL Skills
Experience in Hadoop and languages such as Pig and Hive. Experience with NoSQL Technologies (MongoDB, Neo4j)

1b – Data Scientist

Industry: Technology
Experience: 6+ Years
Qualifications:
Use machine learning, data mining and statistical techniques to create new, scalable solutions for business problems
Strong problem-solving ability
Strong communication and data presentation skills
Experience using Java or C/C++; Python and/or R
Experience using machine learning libraries, such as scikit-learn, caret, mlr, mllib

1c – Data Scientist

Industry: Technology
Experience: 3+ Years
Qualifications:
Excellent statistical, machine learning and data mining skills
A strong quantitative academic background
Proven oral and written communication skills
Superior data modeling and analysis in R/Python
Experience in data engineering using SQL, Tableau

2a – Data Analyst

Industry: Technology
Experience: 2+ Years
Qualifications:
In-depth knowledge of database structures/data warehousing principals
Revel in the complex, but seek to simplify and clarify it in your analysis and communication through strong writing, data visualization and communication skills.
SAS, R, Python, or Perl; Hadoop, Hive, NoSQL, Spark, Mahout, Impala, Pig, Cascading, Yarn, Theano Data Visualization with Tibco

2b – Data Analyst

Industry: Technology
Experience: 8+ Years
Qualifications:
Strong analytical skills with the ability to collect, organize, analyze and disseminate significant amounts of information with attention to detail and accuracy
Ability to document business processes
ETL; Microsoft Excel; Oracle; SQL

2c – Data Analyst

Industry: Technology
Experience: 3+ Years
Qualifications:
It is essential that the candidate has a good working knowledge of visualization, combining and analyzing large data sets, flat file structure (e.g. CSV) and proprietary data structures.
Tableau; SSAS/SPSS/SAS; Matlab, R, Python, and SQL

KENNESAW STATE UNIVERSITY

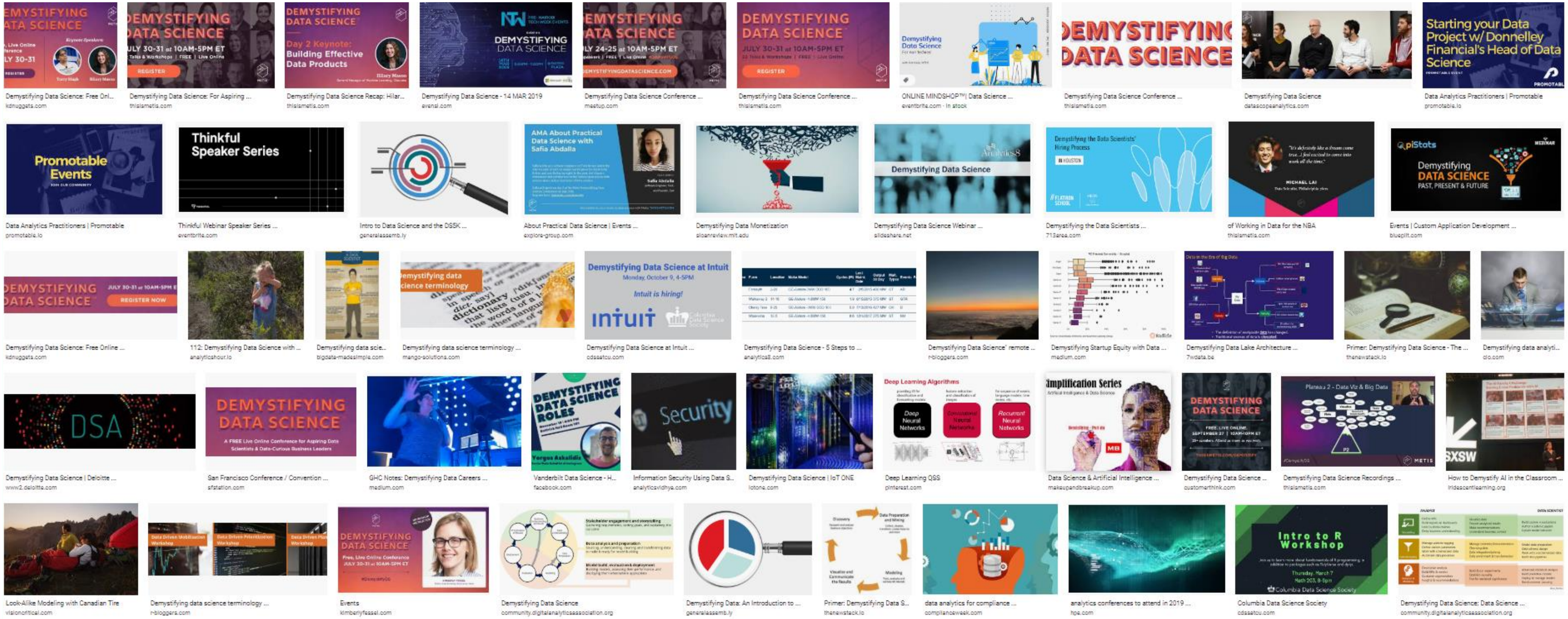
Concepts to be Covered

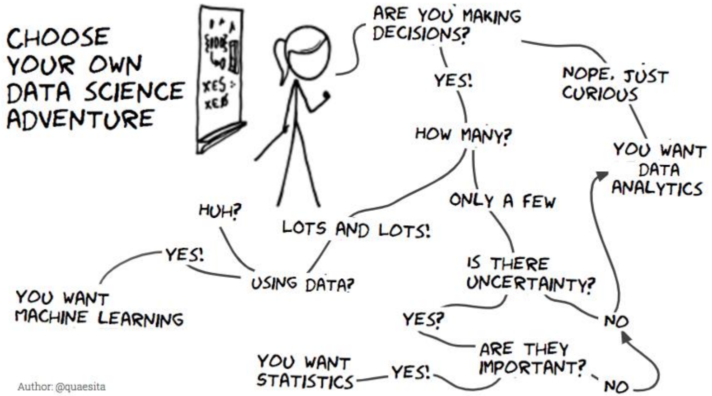
- ➢The Evolution of Data Science
- ➢Defining Data Science – and Data Scientists
- ➢Demystifying Data Science
- ➢Exercise: Developing an Analytical Plan
- ➢Ethical Considerations in Data Science: Human Subjects and Algorithmic Bias
- ➢Exercise: Ethics Case Study
- ➢Summary and Wrap Up

KENNESAW STATE UNIVERSITY

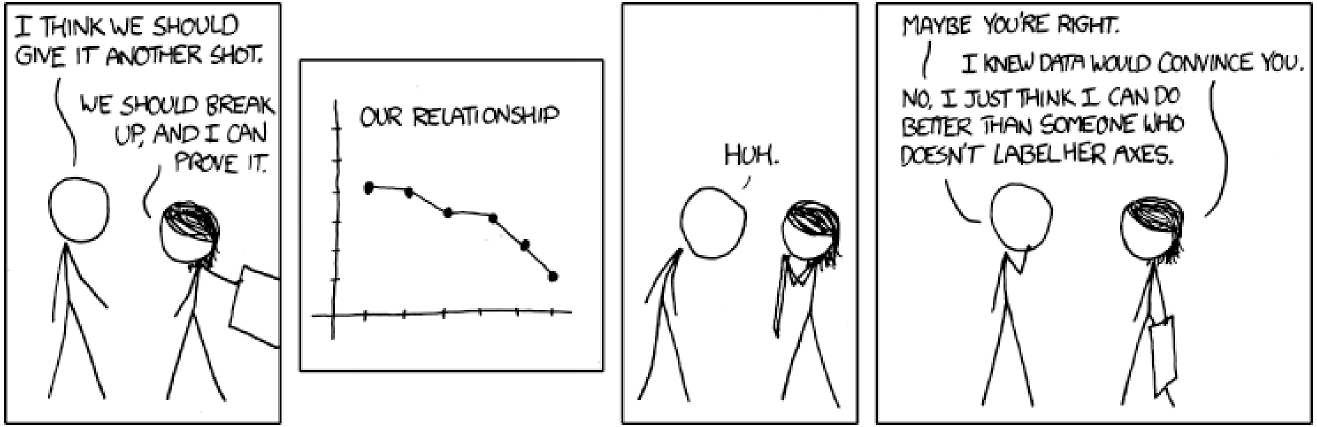
KENNESAW STATE UNIVERSITY

Demystifying Data Science…




CHOOSE YOUR OWN DATA SCIENCE ADVENTURE
ARE YOU MAKING DECISIONS?
YES!
NOPE, JUST CURIOUS
HOW MANY?
YOU WANT DATA ANALYTICS
ONLY A FEW
LOTS AND LOTS!
HUH?
YES!
USING DATA?
IS THERE UNCERTAINTY?
YOU WANT MACHINE LEARNING
YES?
NO
ARE THEY IMPORTANT?
YOU WANT STATISTICS
YES!
NO
Author: @quaesita


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS


I THINK WE SHOULD GIVE IT ANOTHER SHOT.
WE SHOULD BREAK UP, AND I CAN PROVE IT.
OUR RELATIONSHIP
HUH.
MAYBE YOU'RE RIGHT.
I KNEW DATA WOULD CONVINCE YOU.
NO, I JUST THINK I CAN DO BETTER THAN SOMEONE WHO DOESN'T LABEL HER AXES.

History of Data Science?

- The term “data science” has been traced back to computer scientist Peter Naur in 1960 (Naur, 1992).
- In 1962, the famed statistician John W. Tukey wrote:

“For a long time I thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have … come to feel that my central interest is in data analysis…data analysis is intrinsically an empirical science”

- Gregory Piatetsky-Shapiro organized and chaired the first Knowledge Discovery in Databases (KDD) workshop in 1989
- Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, published “From Data Mining to Knowledge Discovery in Databases” in 1996.
- A reference to the term “data science” as a discipline within statistics was made in the proceedings of the Fifth Conference of the International Federation of Classification Societies in 1996.
- In 1997, during his inaugural lecture as the H. C. Carver Chair in Statistics at the University of Michigan, Jeff Wu called for statistics to be renamed “data science” and statisticians to be renamed “data scientists”.

KENNESAW STATE UNIVERSITY

Analytics and Data Science Institute at KSU

- Ph.D. Program in Analytics and Data Science
- MS in Computer Science
- MS in Applied Statistics and Analytics
- MBA Program/PhD in BA
- Certificates
- Research Labs

KENNESAW STATE UNIVERSITY

Ph.D. in Data Science Curriculum

MATH CORE:
- ✓ THEORY OF LINEAR MODELS
- ✓ DISCRETE MATHEMATICS
- ✓ GRAPH THEORY
- ✓ MATHEMATICS FOR BIG DATA

CS CORE:
- ✓ ALGORITHM DESIGN
- ✓ WEB AND TEXT MINING
- ✓ MACHINE LEARNING

STAT CORE:
- ✓ DATA MINING I
- ✓ DATA MINING II
- ✓ BINARY CLASSIFICATION

OTHER:
- ✓ CSAR
- ✓ ETHICS/PRIVACY
- ✓ DISSERTATION
- ✓ TEACHING

Traditional Statistical Modeling

- Interpretability
- "White Box"
- Typically requires some theory/hypothesis
- Generalizable
- Results can be replicated more easily
- Works with smaller datasets

Machine Learning

- "User Friendly"
- Typically Better Predictions
- "Black Box" – Don't need to know what you are doing
- Problems with overfitting – specific to the data
- Generalizability can be an issue
- Can be completely "data driven"
- Needs large volumes of data

KENNESAW STATE UNIVERSITY

To clarify…these are data science "frameworks", not "methods":

- Data Mining
- Machine Learning
- Predictive Modeling
- Classification
- Neural Networks
- Artificial Intelligence
- Text Analytics
- Image Detection

DISCLAIMER…

Who ARE these People?

Business Disciplines ⟷ Computational Disciplines

Analysts

- ✓ Data Analyst
- ✓ Business/ Analyst
- ✓ Operations/Systems Analyst

Researchers/Scientists

- ✓ Data Engineer
- Modeler
- ✓ Data Architect
- ✓ Machine Learning Engineer

ANALYTICS TRANSLATOR

Demystifying Data Science…

"The Analytics Translator"

What do Analytics Translators actually do[1]?

1. Identify and prioritize problems that analytics is suited to solve
2. Help Identify the data that is needed to generate insights
3. Ensures the solution solves the business problem (i.e. the “right” answer to the “right” problem)
4. Synthesizes complex analytical results into easy-to-understand, actionable recommendations that end users can execute.
5. Drives adoption among business users.

1. hbr.org/2018/02/you-dont-have-to-be-a-data-scientist-to-fill-this-must-have-analytics-role

KENNESAW STATE UNIVERSITY

Methods (almost) every Data Scientist Uses:[1,2]

Supervised:

- Linear (and Multiple) Regression
- Logistic Regression
- Support Vector Machines
- Decision Trees and Random Forests

Unsupervised:

- Principal Components Analysis
- k-Means Clustering (versus hierarchical clustering)

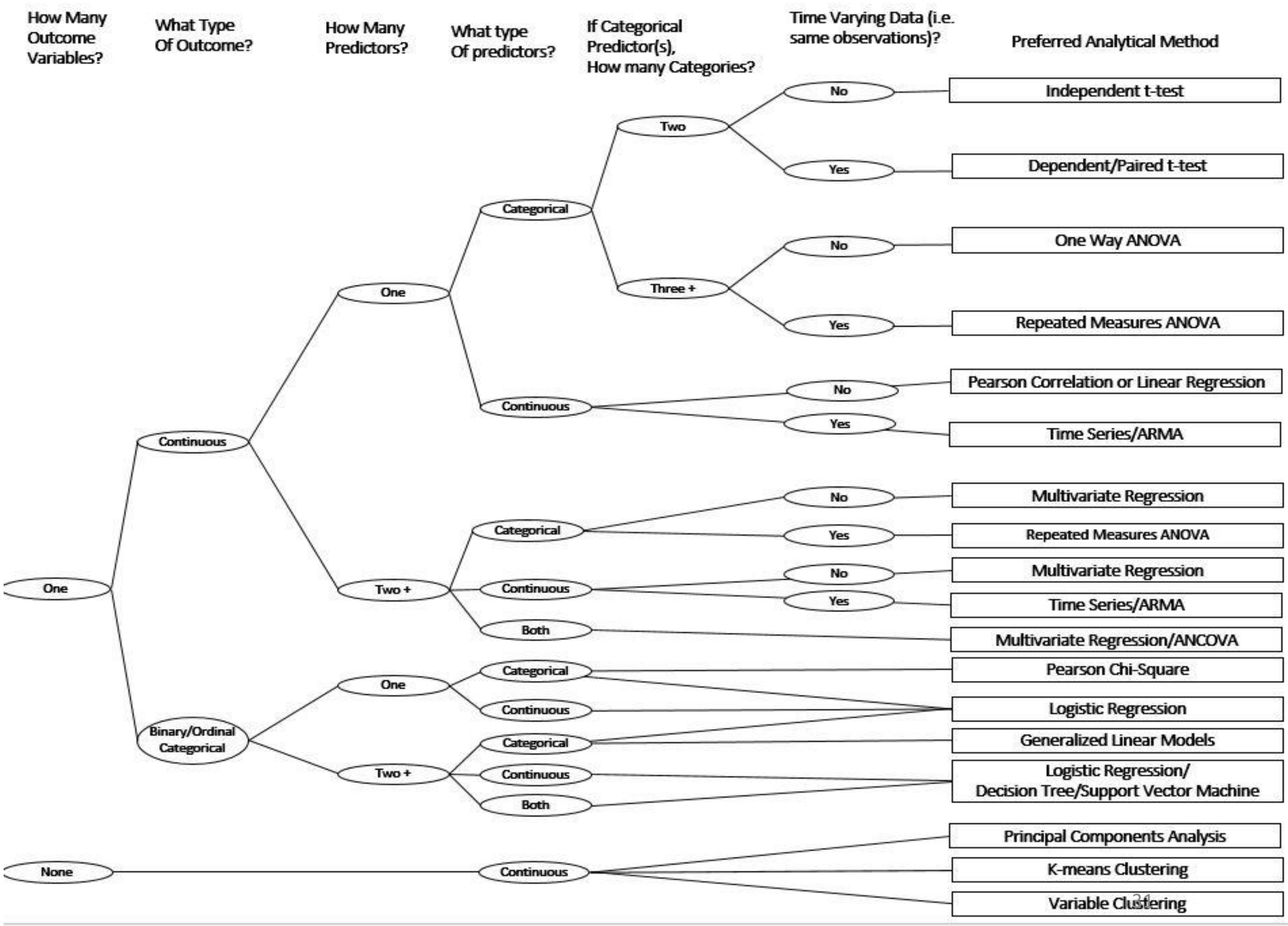
1. *Datasciencecentral.com*
2. *Towardsdatascience.com*

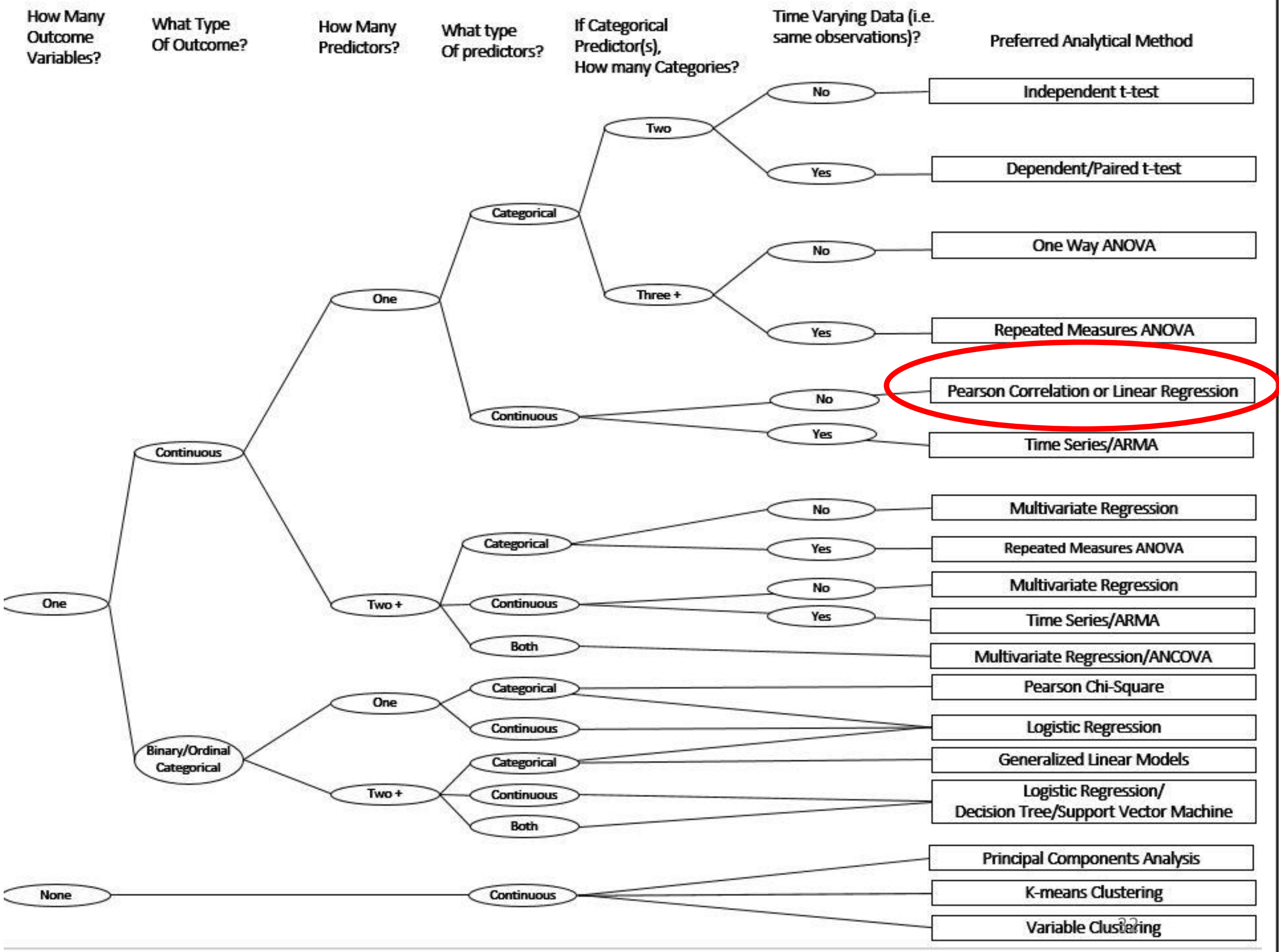
KENNESAW STATE UNIVERSITY

These are the most common data science methods ➔

Preferred Analytical Method

- Independent t-test
- Dependent/Paired t-test
- One Way ANOVA
- Repeated Measures ANOVA
- Pearson Correlation or Linear Regression
- Time Series/ARMA
- Multivariate Regression
- Repeated Measures ANOVA
- Multivariate Regression
- Time Series/ARMA
- Multivariate Regression/ANCOVA
- Pearson Chi-Square
- Logistic Regression
- Generalized Linear Models
- Logistic Regression/ Decision Tree/Support Vector Machine
- Principal Components Analysis
- K-means Clustering
- Variable Clustering


How Many Outcome Variables?
What Type Of Outcome?
How Many Predictors?
What type Of predictors?
If Categorical Predictor(s), How many Categories?
Time Varying Data (i.e. same observations)?
Preferred Analytical Method
One
Continuous
One
Categorical
Two
No
Independent t-test
Yes
Dependent/Paired t-test
Three +
No
One Way ANOVA
Yes
Repeated Measures ANOVA
Continuous
No
Pearson Correlation or Linear Regression
Yes
Time Series/ARMA
Two +
Categorical
No
Multivariate Regression
Yes
Repeated Measures ANOVA
Continuous
No
Multivariate Regression
Yes
Time Series/ARMA
Both
Multivariate Regression/ANCOVA
Binary/Ordinal Categorical
One
Categorical
Pearson Chi-Square
Continuous
Logistic Regression
Two +
Categorical
Generalized Linear Models
Continuous
Both
Logistic Regression/ Decision Tree/Support Vector Machine
None
Continuous
Principal Components Analysis
K-means Clustering
Variable Clustering


How Many Outcome Variables?
What Type Of Outcome?
How Many Predictors?
What type Of predictors?
If Categorical Predictor(s), How many Categories?
Time Varying Data (i.e. same observations)?
Preferred Analytical Method
One
Continuous
One
Categorical
Two
No
Independent t-test
Yes
Dependent/Paired t-test
Three +
No
One Way ANOVA
Yes
Repeated Measures ANOVA
Continuous
No
Pearson Correlation or Linear Regression
Yes
Time Series/ARMA
Two +
Categorical
No
Multivariate Regression
Yes
Repeated Measures ANOVA
Continuous
No
Multivariate Regression
Yes
Time Series/ARMA
Both
Multivariate Regression/ANCOVA
Binary/Ordinal Categorical
One
Categorical
Pearson Chi-Square
Continuous
Logistic Regression
Two +
Categorical
Generalized Linear Models
Continuous
Both
Logistic Regression/ Decision Tree/Support Vector Machine
None
Continuous
Principal Components Analysis
K-means Clustering
Variable Clustering

Linear (and Multiple) Regression – The Basics

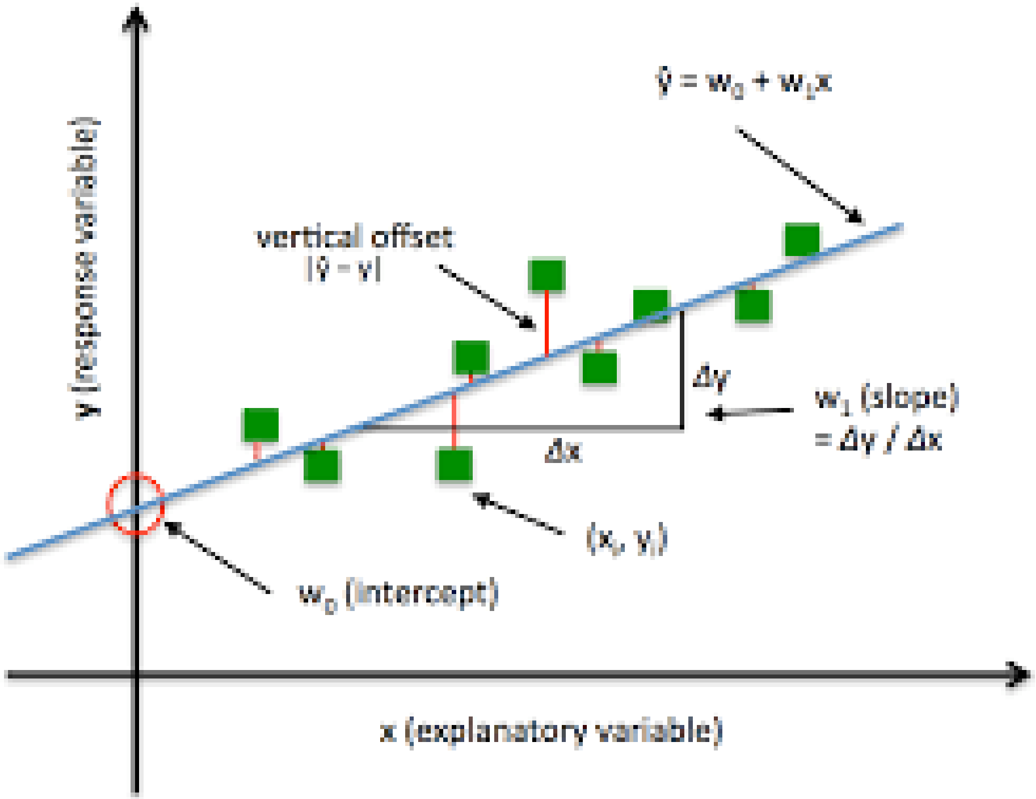
Application Example: You need to determine the fair price of a car, given the size of the engine.

Evaluating Model Performance:
- R^2 (percent of variance explained)

Pros:
- Simple – least complex
- Interpretability

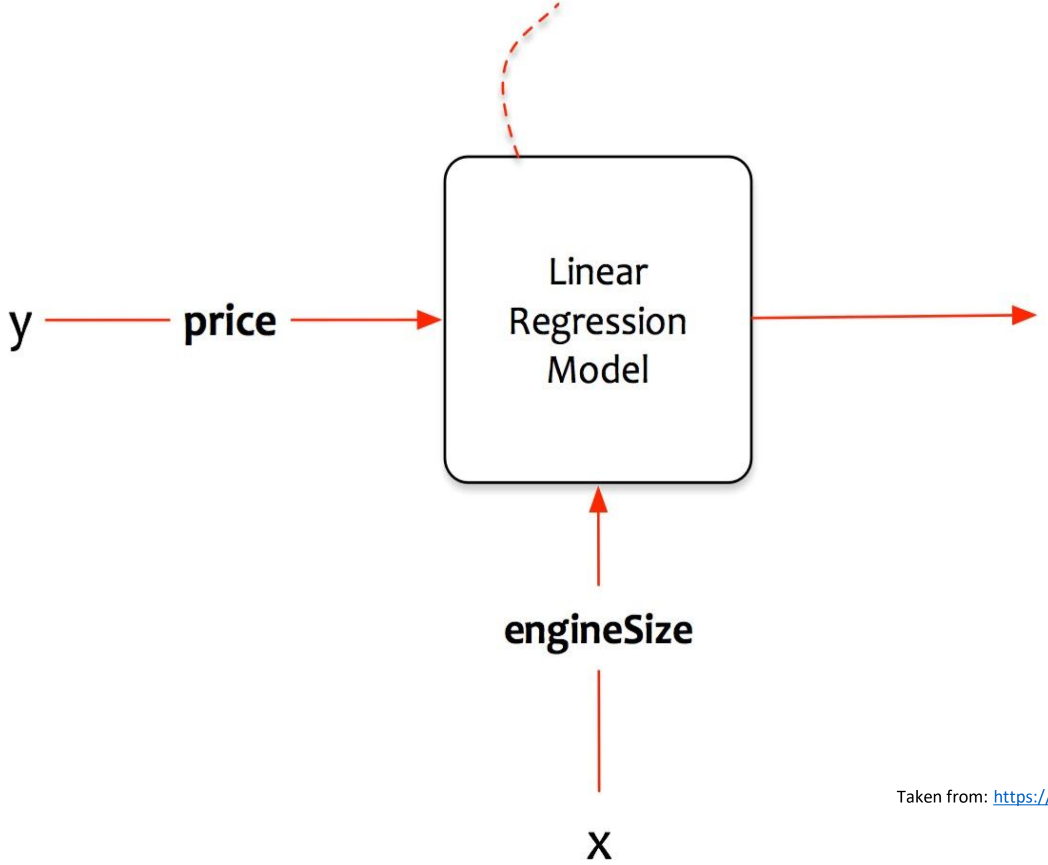
Cons:
- Requires a lot of parametric assumptions
- Simplistic
- Requires a lot of preprocessing

KENNESAW STATE UNIVERSITY



KENNESAW STATE UNIVERSITY

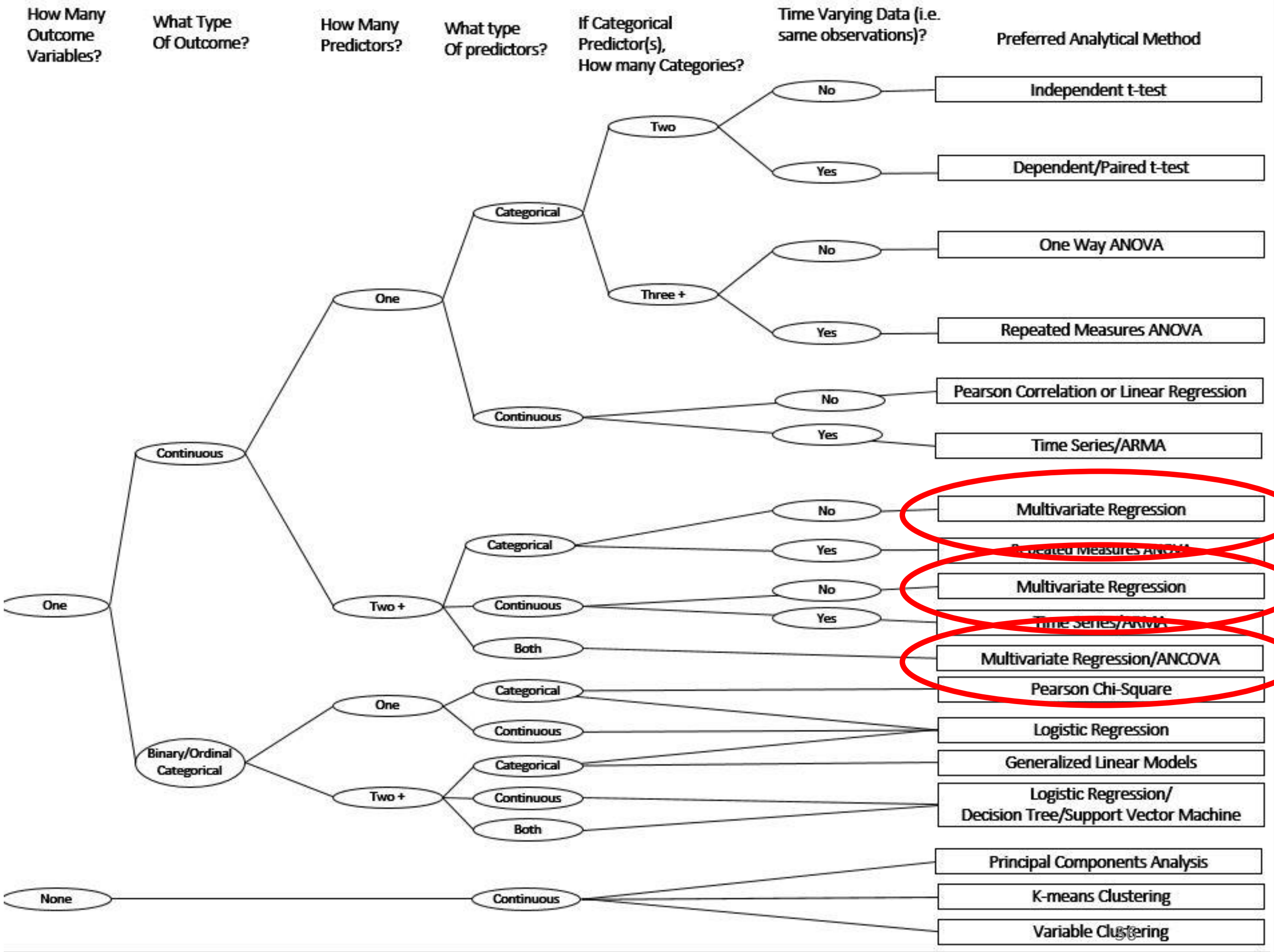
- estimates β0 and β1
- creates model performance metrics




```
Coefficients:
(Intercept)   engineSize
    -6870.1        156.9
```

price = -6870.1 + 156.9 * engineSize

Taken from: https://towardsdatascience.com/data-science-simplified-part-5-multivariate-regression-models-7684b0489015



How Many Outcome Variables?
What Type Of Outcome?
How Many Predictors?
What type Of predictors?
If Categorical Predictor(s), How many Categories?
Time Varying Data (i.e. same observations)?
Preferred Analytical Method
One
Continuous
One
Categorical
Two
No
Independent t-test
Yes
Dependent/Paired t-test
Three +
No
One Way ANOVA
Yes
Repeated Measures ANOVA
Continuous
No
Pearson Correlation or Linear Regression
Yes
Time Series/ARMA
Two +
Categorical
No
Multivariate Regression
Yes
Continuous
No
Multivariate Regression
Yes
Both
Multivariate Regression/ANCOVA
Binary/Ordinal Categorical
One
Categorical
Pearson Chi-Square
Continuous
Logistic Regression
Two +
Categorical
Generalized Linear Models
Continuous
Both
Logistic Regression/ Decision Tree/Support Vector Machine
None
Continuous
Principal Components Analysis
K-means Clustering
Variable Clustering

Multiple Regression – The Basics

make	fuelType	nDoors	driveWheels	engineSize	horsePower	peakRpm	cityMpg	price
alfa-romero	gas	two	rwd	130	111	5000	21	13495
alfa-romero	gas	two	rwd	130	111	5000	21	16500
alfa-romero	gas	two	rwd	152	154	5000	19	16500
audi	gas	four	fwd	109	102	5500	24	13950
audi	gas	four	4wd	136	115	5500	18	17450
audi	gas	two	fwd	136	110	5500	19	15250
audi	gas	four	fwd	136	110	5500	19	17710
audi	gas	four	fwd	136	110	5500	19	18920
audi	gas	four	fwd	131	140	5500	17	23875
bmw	gas	two	rwd	108	101	5800	23	16430
bmw	gas	four	rwd	108	101	5800	23	16925
bmw	gas	two	rwd	164	121	4250	21	20970
bmw	gas	four	rwd	164	121	4250	21	21105
bmw	gas	four	rwd	164	121	4250	20	24565
bmw	gas	four	rwd	209	182	5400	16	30760
bmw	gas	two	rwd	209	182	5400	16	41315
bmw	gas	four	rwd	209	182	5400	15	36880

Taken from: https://towardsdatascience.com/data-science-simplified-part-5-multivariate-regression-models-7684b0489015

KENNESAW STATE UNIVERSITY

$\beta_0 \ldots \beta_6$

t-stat

p-value

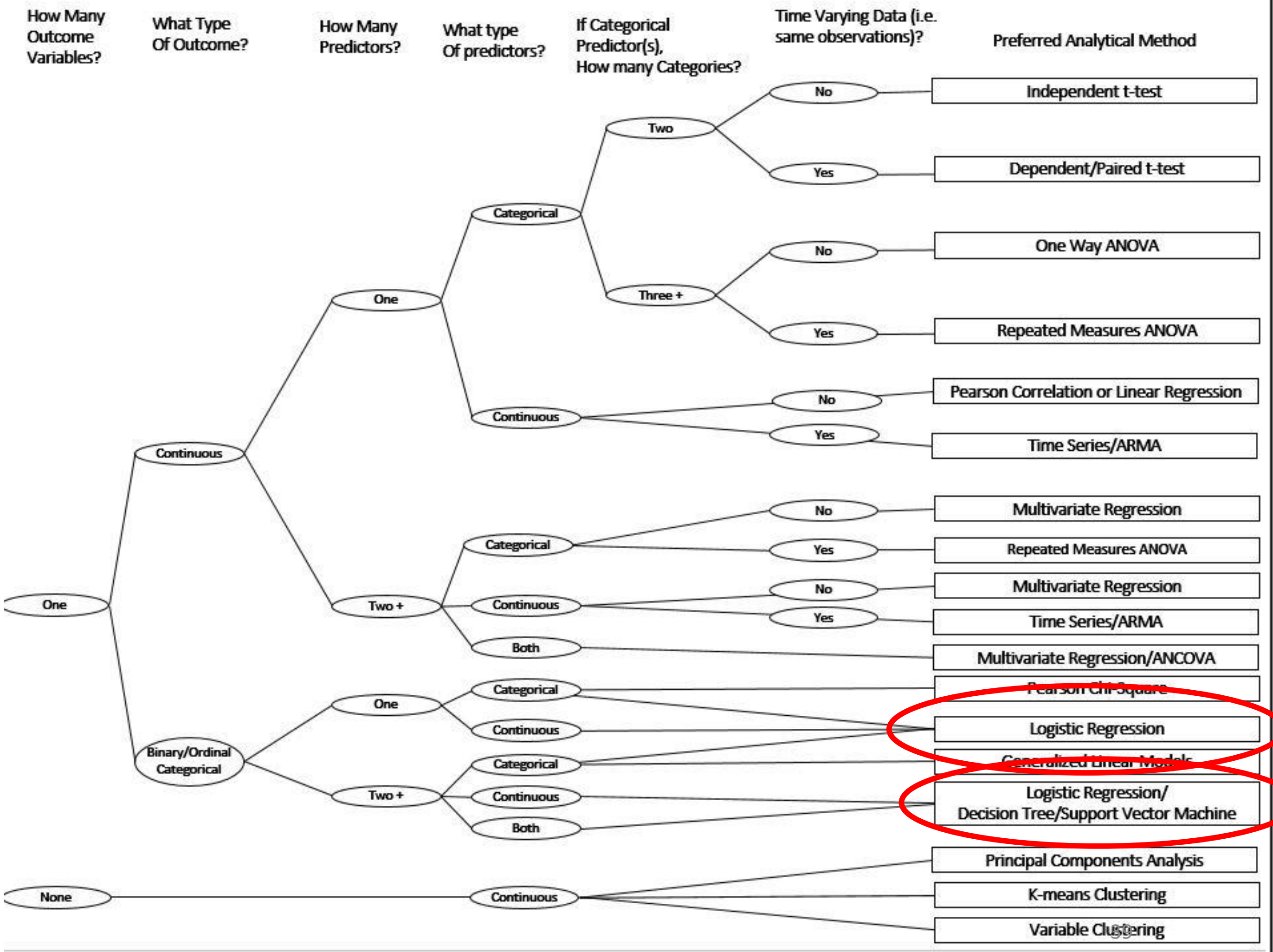
```
Coefficients:
              Estimate  Std. Error  t value       Pr(>|t|)
(Intercept)  -85086.39    15265.49    -5.57  0.00000012266 ***
engineSize      102.85       15.38     6.69  0.00000000049 ***
horsePower       43.79       16.41     2.67         0.0085 **
peakRpm           1.52        0.72     2.11         0.0367 *
length          -37.91       54.19    -0.70         0.4854
width           908.12      282.27     3.22         0.0016 **
height          364.33      153.36     2.38         0.0189 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3300 on 141 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.818,     Adjusted R-squared:  0.811
F-statistic: 106 on 6 and 141 DF,  p-value: <0.0000000000000002
```

coefficient of determiniation (R^2)

Adjusted coefficient of determiniation (R^2)

Taken from: https://towardsdatascience.com/data-science-simplified-part-5-multivariate-regression-models-7684b0489015


How Many Outcome Variables?
What Type Of Outcome?
How Many Predictors?
What type Of predictors?
If Categorical Predictor(s), How many Categories?
Time Varying Data (i.e. same observations)?
Preferred Analytical Method
One
Continuous
One
Categorical
Two
No
Independent t-test
Yes
Dependent/Paired t-test
Three +
No
One Way ANOVA
Yes
Repeated Measures ANOVA
Continuous
No
Pearson Correlation or Linear Regression
Yes
Time Series/ARMA
Two +
Categorical
No
Multivariate Regression
Yes
Repeated Measures ANOVA
Continuous
No
Multivariate Regression
Yes
Time Series/ARMA
Both
Multivariate Regression/ANCOVA
Binary/Ordinal Categorical
One
Categorical
Pearson Chi-Square
Continuous
Logistic Regression
Two +
Categorical
Generalized Linear Models
Continuous
Both
Logistic Regression/ Decision Tree/Support Vector Machine
None
Continuous
Principal Components Analysis
K-means Clustering
Variable Clustering

Logistic Regression – The Basics

Application Example: You need to classify borrowers into two groups - "loan/don't loan"

Application Example: Given an image, you need to determine if a tumor is malignant/not-malignant.

Performance Metrics:

- AUC
- Sensitivity/Specificity
- Precision/Recall
- Loss/Profit Function

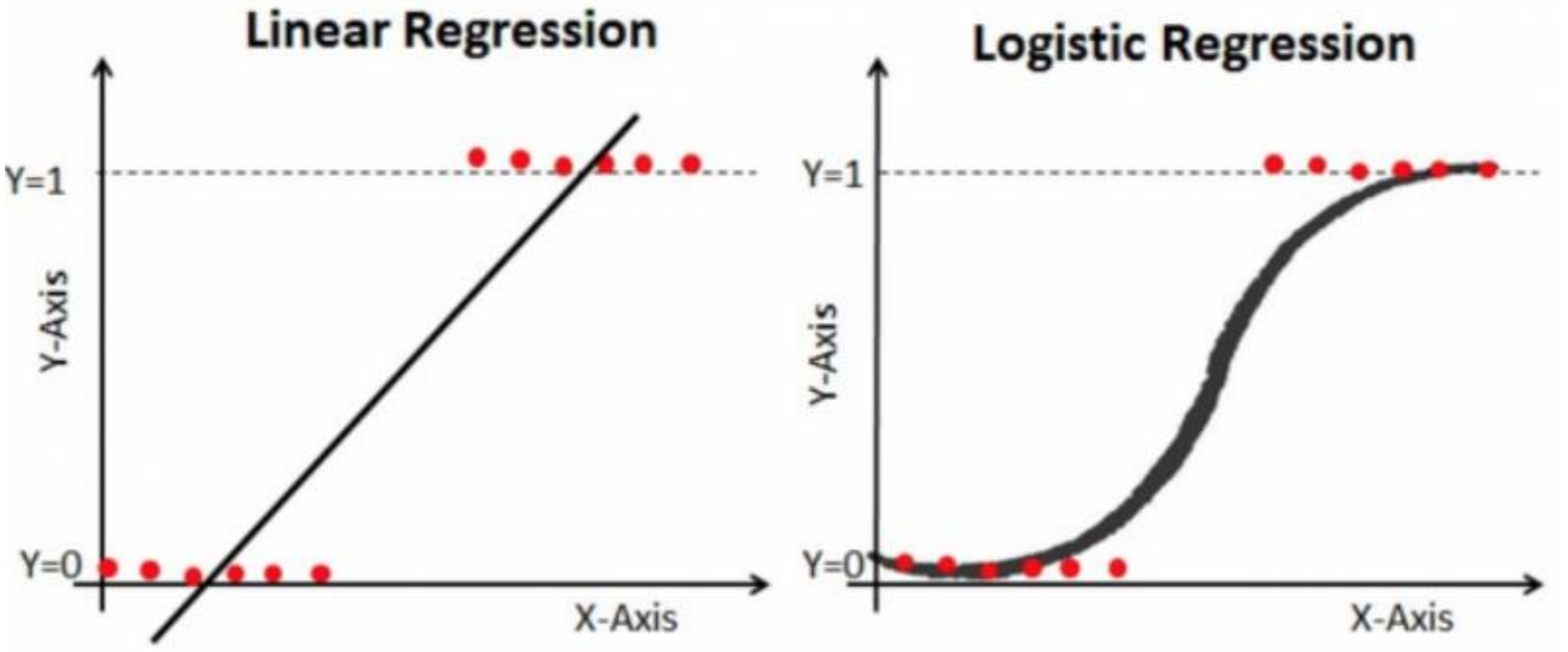
Pros:

- Commonly understood
- Scoring models are widely applicable

Cons:

- Loss of information with binary dependent variable
- Does not accommodate large number of predictors
- Subject to overfitting

KENNESAW STATE UNIVERSITY


Linear Regression
Y=1
Y=0
Y-Axis
X-Axis
Logistic Regression
Y=1
Y=0
Y-Axis
X-Axis

Notes on Binary Classification Assessment

Let's say, we are considering lending money to 100 people. There are four possible outcomes:

	Good Risk	Bad Risk
Predicted Good Risk	40	20
Predicted Bad Risk	15	25

True Positive = Good Predicted as Good

True Negative = Bad Predicted as Bad

False Positive = Bad Predicted as Good

False Negative = Good Predicted as Bad

Notes on Binary Classification Assessment

	Good Risk	Bad Risk
Predicted Good Risk	40	20
Predicted Bad Risk	15	25

Sensitivity (Recall) = 40/55 = 72.73%

Specificity = 25/40 = 62.50%

Precision = 40/(40+20) = 66%

Accuracy = (40+25)/(40+25+20+15) = 75%

F1 = 2*((Precision*Recall)/(Precision+Recall))

F1 = 2*(.4800/1.3873) = .6920

TP Rate = 40/55 = 72.73%

TN Rate = 25/45 = 55.55%

KENNESAW STATE UNIVERSITY

Notes on Binary Classification Assessment

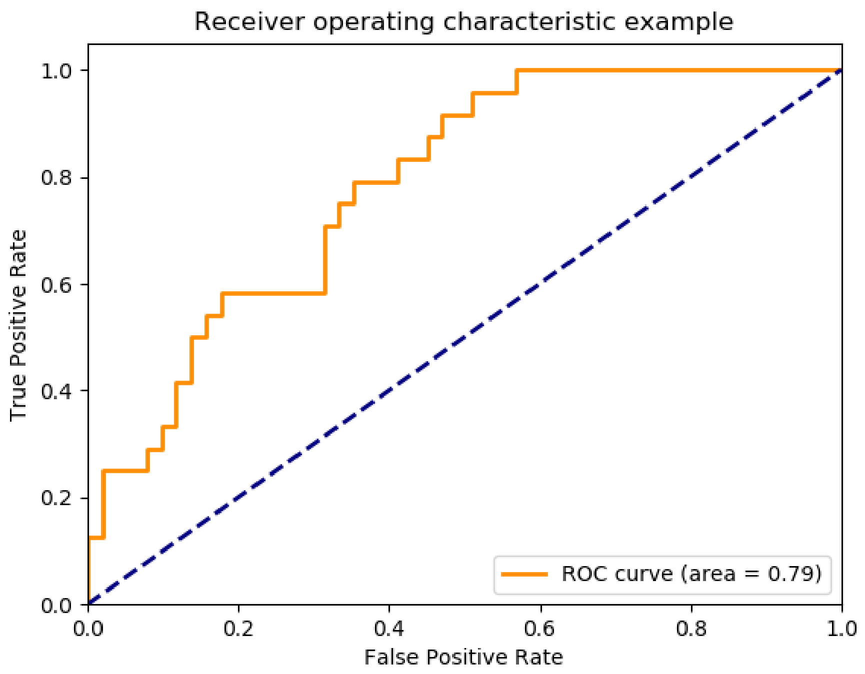
	Good Risk	Bad Risk
Predicted Good Risk	40	20
Predicted Bad Risk	15	25

Good Risk	Bad Risk
20	10
35	35

A more conservative specification for a "good" prediction, will generate fewer predicted "goods"…but may improve a metric like true negative rate (55% to 78%).

Its important to note that the analyst can "toggle" the definition of "good" and "bad" depending on the problem…50% cut off to define "good" or "bad" is just as "random" as 20% or 80%.

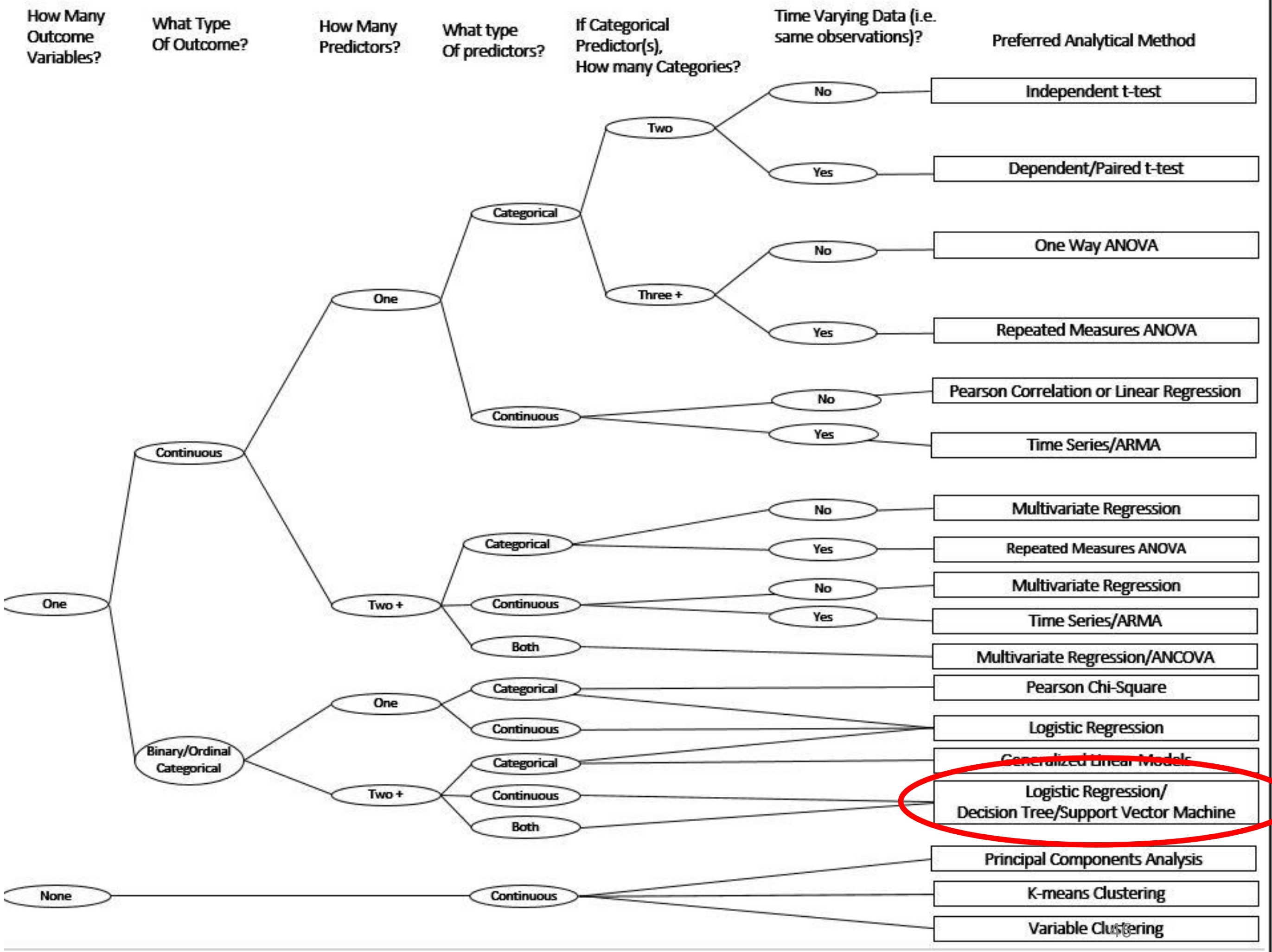
KENNESAW STATE UNIVERSITY



The AUC or “Area Under The Curve” is the area under the ROC curve that is created when the True Positive Rate (Sensitivity) is plotted against the False Positive Rate (1-Specificity) for all possible “cut points” for the probability of an event.

https://scikit-learn.org/stable/modules/model_evaluation.html

KENNESAW STATE UNIVERSITY



How Many Outcome Variables?
What Type Of Outcome?
How Many Predictors?
What type Of predictors?
If Categorical Predictor(s), How many Categories?
Time Varying Data (i.e. same observations)?
Preferred Analytical Method
One
Continuous
One
Categorical
Two
No
Independent t-test
Yes
Dependent/Paired t-test
Three +
No
One Way ANOVA
Yes
Repeated Measures ANOVA
Continuous
No
Pearson Correlation or Linear Regression
Yes
Time Series/ARMA
Two +
Categorical
No
Multivariate Regression
Yes
Repeated Measures ANOVA
Continuous
No
Multivariate Regression
Yes
Time Series/ARMA
Both
Multivariate Regression/ANCOVA
Binary/Ordinal Categorical
One
Categorical
Pearson Chi-Square
Continuous
Logistic Regression
Two +
Categorical
Generalized Linear Models
Continuous
Both
Logistic Regression/ Decision Tree/Support Vector Machine
None
Continuous
Principal Components Analysis
K-means Clustering
Variable Clustering

Support Vector Machines – The Basics

Example: You need to classify images into one of two groups (e.g. cat or dog).

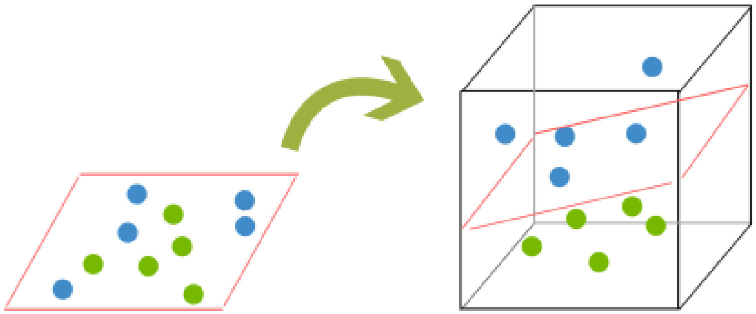
Common Performance Metrics:
- Accuracy
- Sensitivity/Specificity
- Precision/Recall
- F1

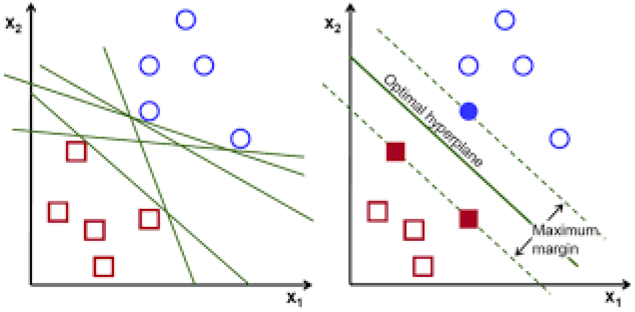
Pros:
- Particularly effective in higher dimensions.
- Effective when the number of features are more than training examples (k>n).
- The hyperplane is affected by only the support vectors thus outliers have less impact.

Cons:
- Interpretability
- Sensitive to selection of the Kernel Transformation
- Computationally intensive

KENNESAW STATE UNIVERSITY

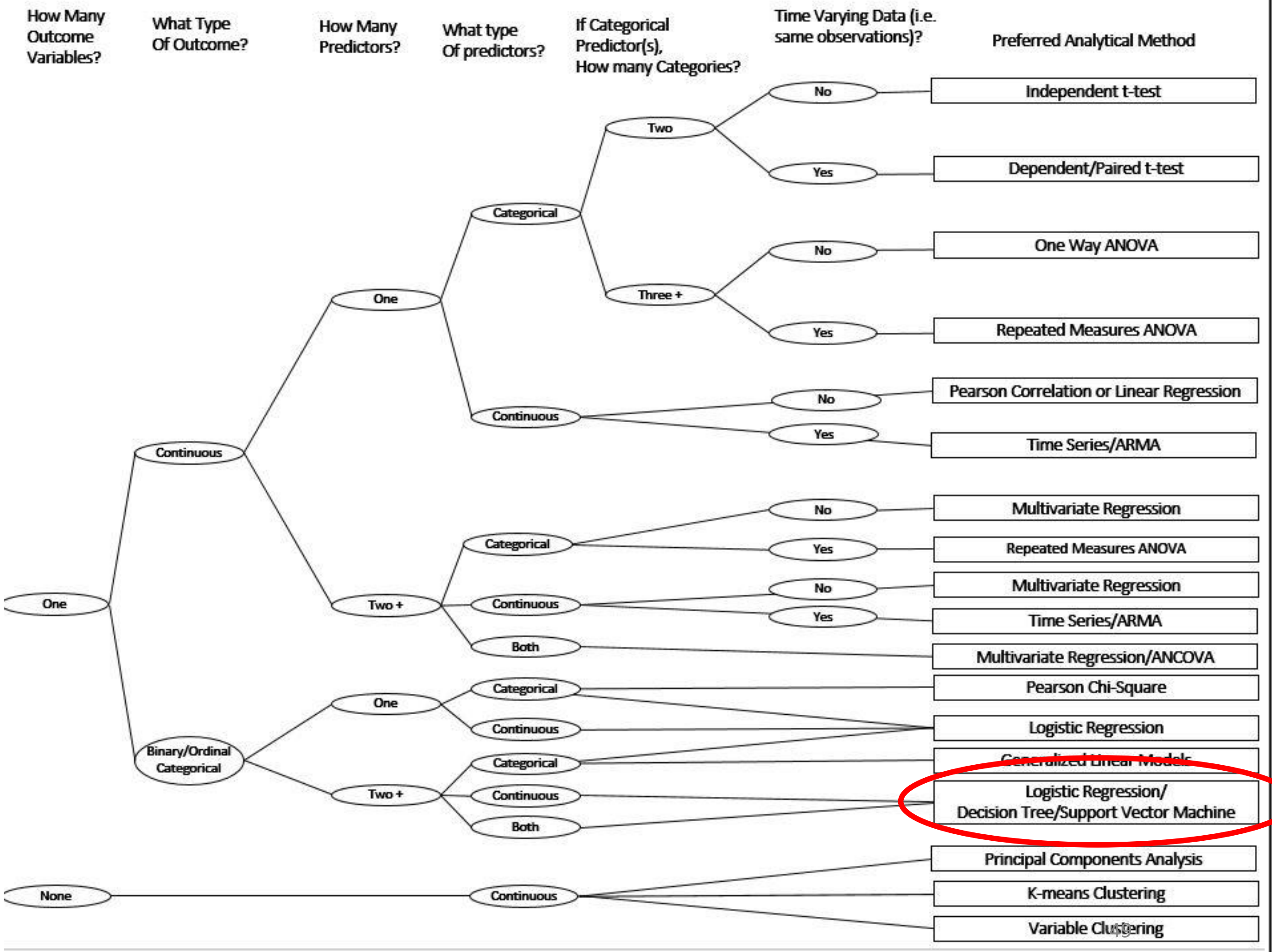
Kernel transformations ("tricks") commonly include linear, polymomial, radial basis function (nn) and sigmoid





The primary metric for evaluation is the F1 statistic – provided in the previous section

KENNESAW STATE UNIVERSITY


How Many Outcome Variables?
What Type Of Outcome?
How Many Predictors?
What type Of predictors?
If Categorical Predictor(s), How many Categories?
Time Varying Data (i.e. same observations)?
Preferred Analytical Method
One
None
Continuous
Binary/Ordinal Categorical
One
Two +
Categorical
Continuous
Both
Two
Three +
No
Yes
Independent t-test
Dependent/Paired t-test
One Way ANOVA
Repeated Measures ANOVA
Pearson Correlation or Linear Regression
Time Series/ARMA
Multivariate Regression
Repeated Measures ANOVA
Multivariate Regression
Time Series/ARMA
Multivariate Regression/ANCOVA
Pearson Chi-Square
Logistic Regression
Generalized Linear Models
Logistic Regression/ Decision Tree/Support Vector Machine
Principal Components Analysis
K-means Clustering
Variable Clustering

Decision Trees – The Basics

Example: Will this person default on their loan?

Common Performance Metrics:

- Gini coefficient
- Entropy
- Information Gain

Pros:

- Requires less effort for data preparation during pre-processing.
- Missing values in the data also does NOT affect the process of building decision tree to any considerable extent.
- Very intuitive and easy to explain to technical teams as well as stakeholders.
- Coding solutions can be simple in non-technical environments.

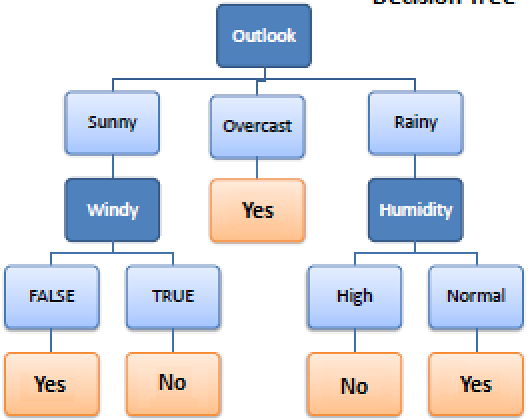
Cons:

- A small change in the data can cause a large change in the structure of the decision tree causing instability.
- Prone to overfitting.
- Decision tree training is sensitive to complexity.
- Does not adequately accommodate continuous data…therefore not great for regression and predicting continuous values.

KENNESAW STATE UNIVERSITY

Predictors				Target
Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

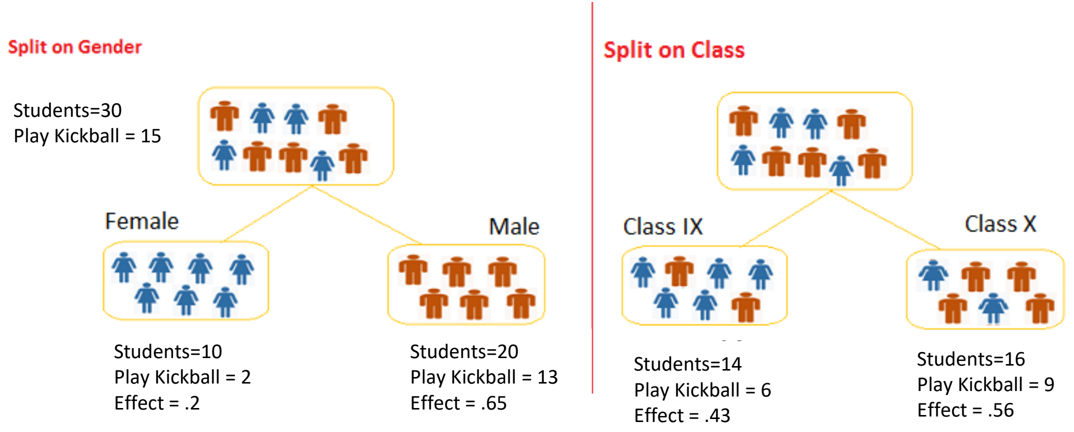
Decision Tree



Taken from: https://medium.com/@rishabhjain_22692/decision-trees-it-begins-here-93ff54ef134

KENNESAW STATE UNIVERSITY

We want to start a kickball team. Who is more likely to want to play?



Which piece of information is more informative?

Gini = p^2+q^2

Split on Gender:

1.Gini for sub-node Female = (0.2)*(0.2)+(0.8)*(0.8)=0.68

2.Gini for sub-node Male = (0.65)*(0.65)+(0.35)*(0.35)=0.55

3.Weighted Gini for Split Gender = (10/30)*0.68+(20/30)*0.55 = **0.59**

Similar for Split on Class:

1.Gini for sub-node Class IX = (0.43)*(0.43)+(0.57)*(0.57)=0.51

2.Gini for sub-node Class X = (0.56)*(0.56)+(0.44)*(0.44)=0.51

3.Weighted Gini for Split Class = (14/30)*0.51+(16/30)*0.51 = **0.51**

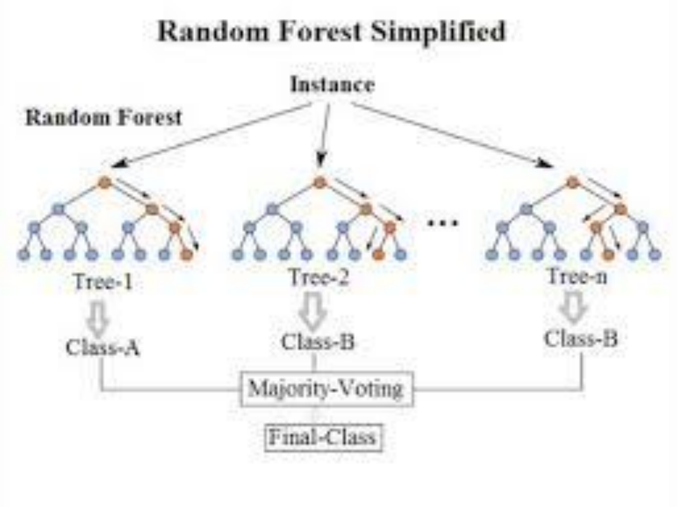
Taken from: https://medium.com/@rishabhjain_22692/decision-trees-it-begins-here-93ff54ef134

KENNESAW STATE UNIVERSITY

Note about Random Forests…

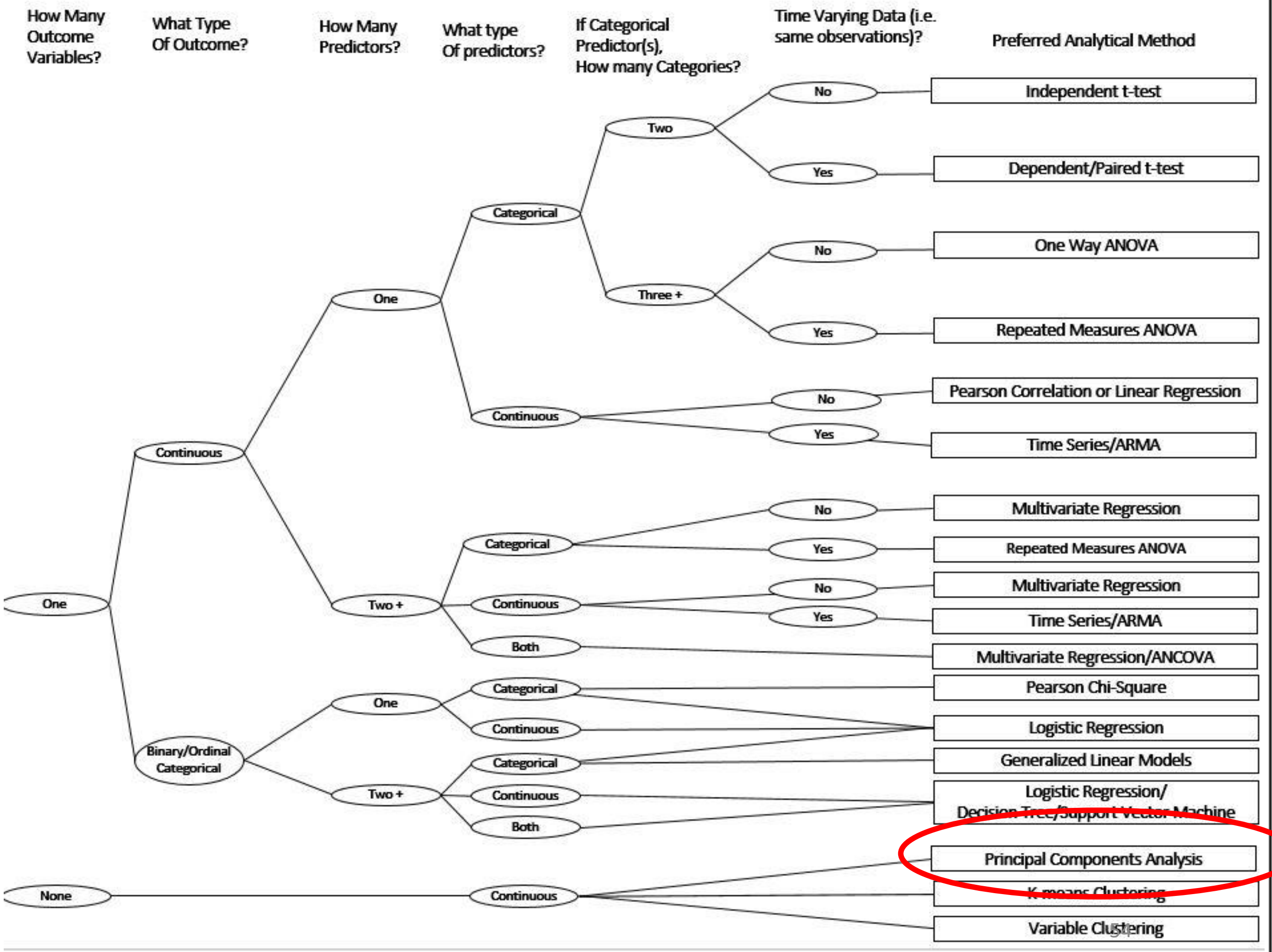
Simply put, RFs consist of randomly sampling subsets of training data, fitting decision trees, and aggregating the predictions. This approach introduces more randomness and diversity into the feature space. That is, instead of searching greedily for the best predictors to create branches, it randomly samples elements of the predictor space, thus adding more diversity and reducing the variance of the trees…

This typically leads to a more robust model, which is less subject to overfitting.



Taken from KDNUGGETS: https://www.kdnuggets.com/2017/10/random-forests-explained.html

KENNESAW STATE UNIVERSITY


How Many Outcome Variables?
What Type Of Outcome?
How Many Predictors?
What type Of predictors?
If Categorical Predictor(s), How many Categories?
Time Varying Data (i.e. same observations)?
Preferred Analytical Method
One
Continuous
One
Categorical
Two
No
Independent t-test
Yes
Dependent/Paired t-test
Three +
No
One Way ANOVA
Yes
Repeated Measures ANOVA
Continuous
No
Pearson Correlation or Linear Regression
Yes
Time Series/ARMA
Two +
Categorical
No
Multivariate Regression
Yes
Repeated Measures ANOVA
Continuous
No
Multivariate Regression
Yes
Time Series/ARMA
Both
Multivariate Regression/ANCOVA
Binary/Ordinal Categorical
One
Categorical
Pearson Chi-Square
Continuous
Logistic Regression
Two +
Categorical
Generalized Linear Models
Continuous
Both
Logistic Regression/ Decision Tree/Support Vector Machine
None
Continuous
Principal Components Analysis
K-means Clustering
Variable Clustering

Principle Components Analysis – The Basics

Example: A voter database has 100 pieces of information (variables) on each voter – which are highly correlated.

Example: A collection of images needs to be categorized (e.g., cats and dogs).

Performance Metrics

- Percent Variance explained

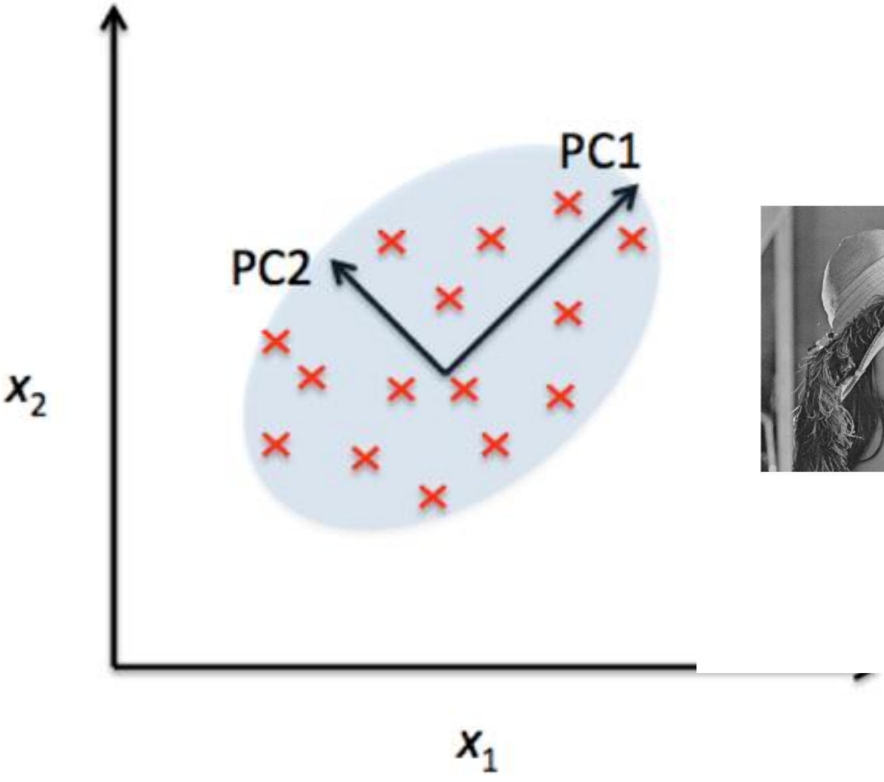
Pros:

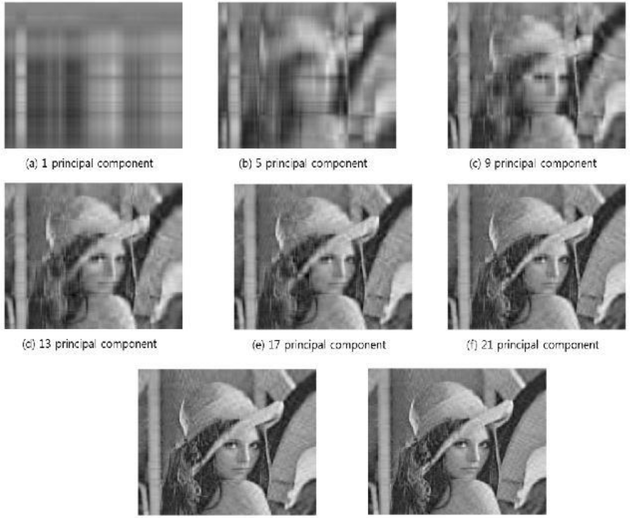
- Reduces Multicollinearity
- Parsimony
- Reduces Overfitting
- Reduces Complexity

Cons:

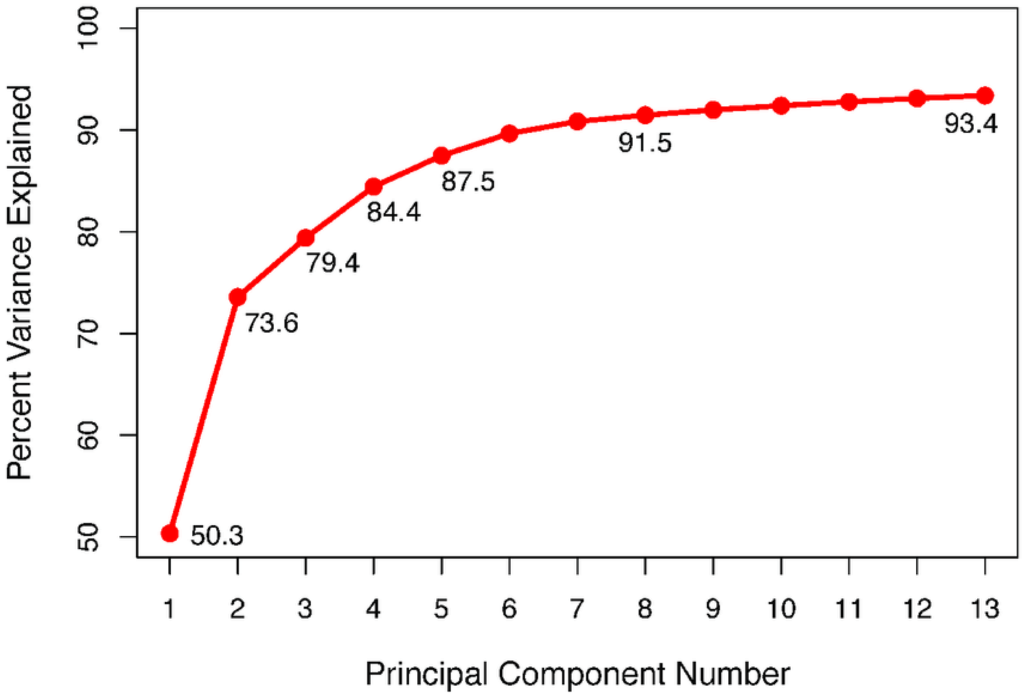
- Loss of Interpretability
- Loss of Information
- Required Standardization of Data

KENNESAW STATE UNIVERSITY





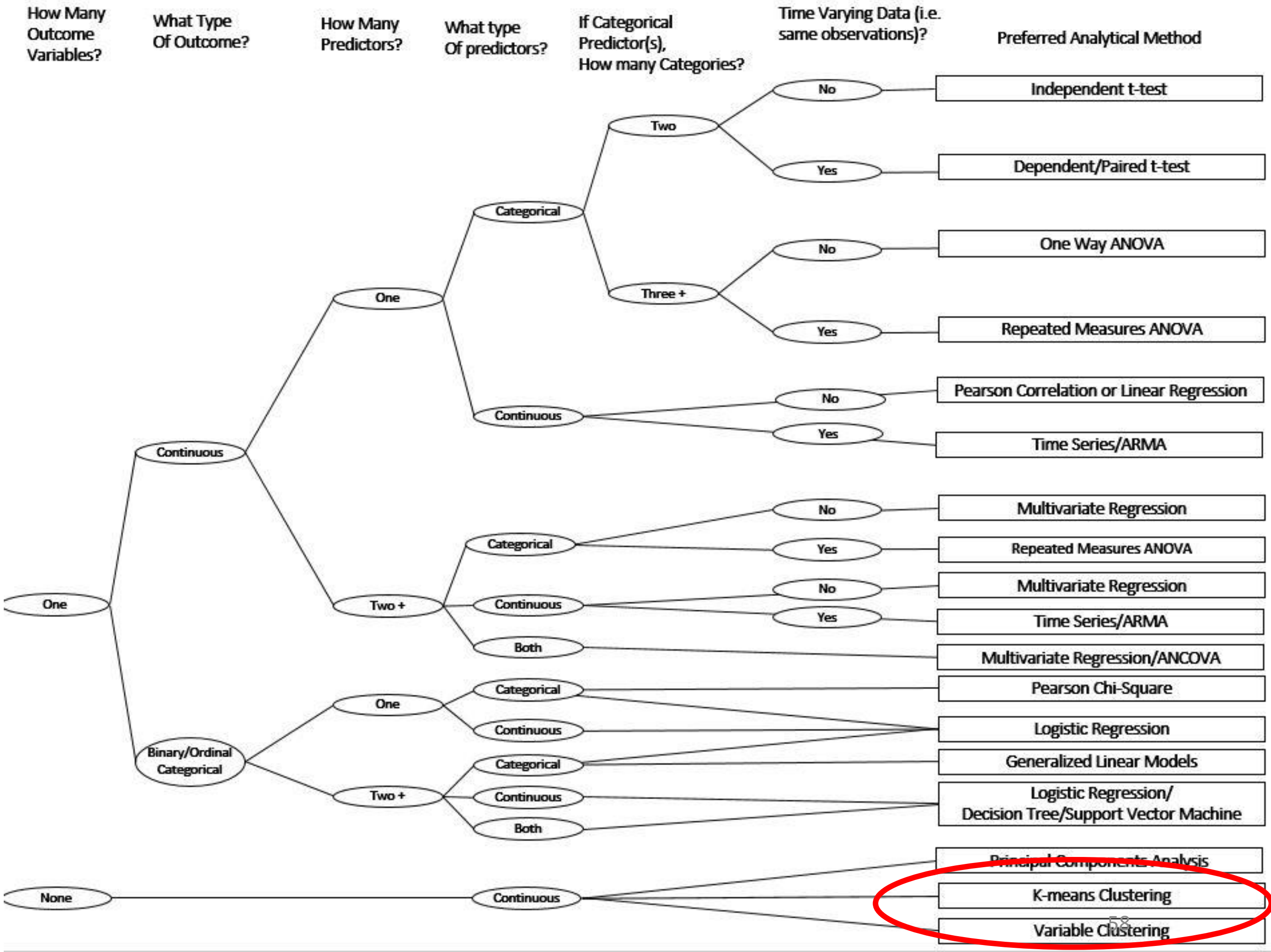
KENNESAW STATE UNIVERSITY



The objective of PCA is to explain as much variation with as few components as possible.

https://www.researchgate.net/publication/321789639_Genetic_relatedness_of_previously_Plant-Variety-Protected_commercial_maize_inbreds/figures?lo=1&utm_source=google&utm_medium=organic

KENNESAW STATE UNIVERSITY


How Many Outcome Variables?
What Type Of Outcome?
How Many Predictors?
What type Of predictors?
If Categorical Predictor(s), How many Categories?
Time Varying Data (i.e. same observations)?
Preferred Analytical Method
One
Continuous
One
Categorical
Two
No
Independent t-test
Yes
Dependent/Paired t-test
Three +
No
One Way ANOVA
Yes
Repeated Measures ANOVA
Continuous
No
Pearson Correlation or Linear Regression
Yes
Time Series/ARMA
Two +
Categorical
No
Multivariate Regression
Yes
Repeated Measures ANOVA
Continuous
No
Multivariate Regression
Yes
Time Series/ARMA
Both
Multivariate Regression/ANCOVA
Binary/Ordinal Categorical
One
Categorical
Pearson Chi-Square
Continuous
Logistic Regression
Two +
Categorical
Generalized Linear Models
Continuous
Logistic Regression/ Decision Tree/Support Vector Machine
Both
None
Continuous
Principal Components Analysis
K-means Clustering
Variable Clustering

K-Means Clustering – The Basics

Example: You have a large group of customers and you need to create segments.

Pros:

- Guaranteed convergence
- Cluster becomes an input to a model
- Relatively simple to explain

Performance Metrics:

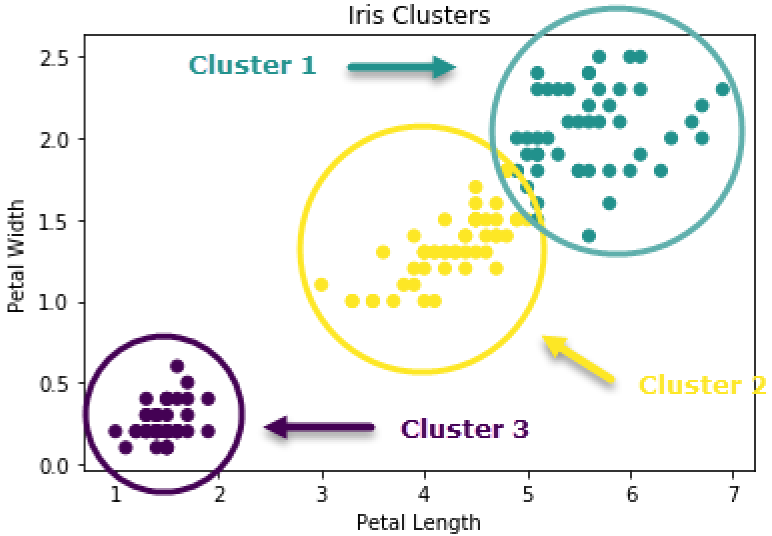
- Within Cluster Sum of Squares
- Between Clusters Sum of Squares
- Total variance explained

Cons:

- Requires (and is sensitive to) scaling
- May be sensitive to initial seeds
- Determination of k

KENNESAW STATE UNIVERSITY

K-Means Clustering – The Basics

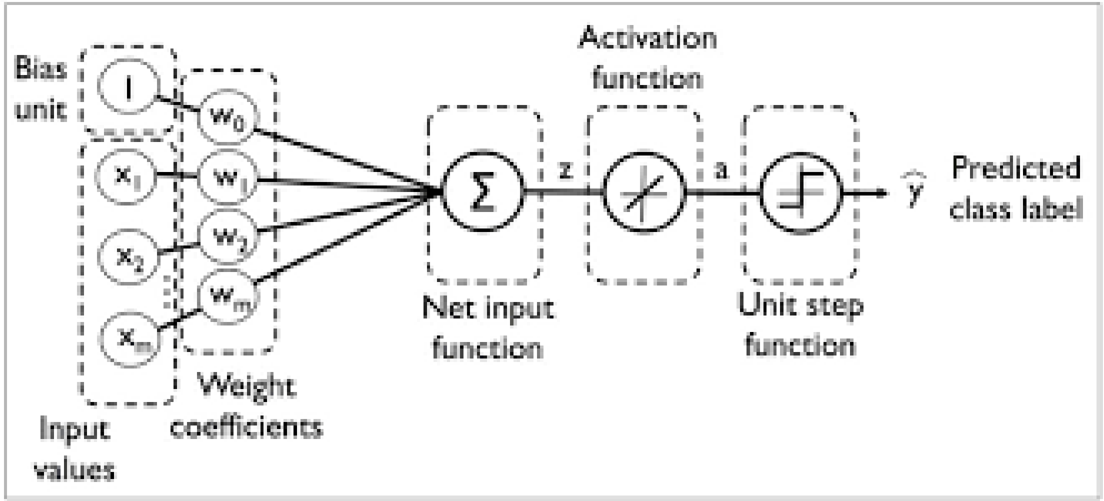


https://rpubs.com/vermaph/395036

https://en.wikipedia.org/wiki/Iris_flower_data_set

KENNESAW STATE UNIVERSITY

Note about Neural Networks…



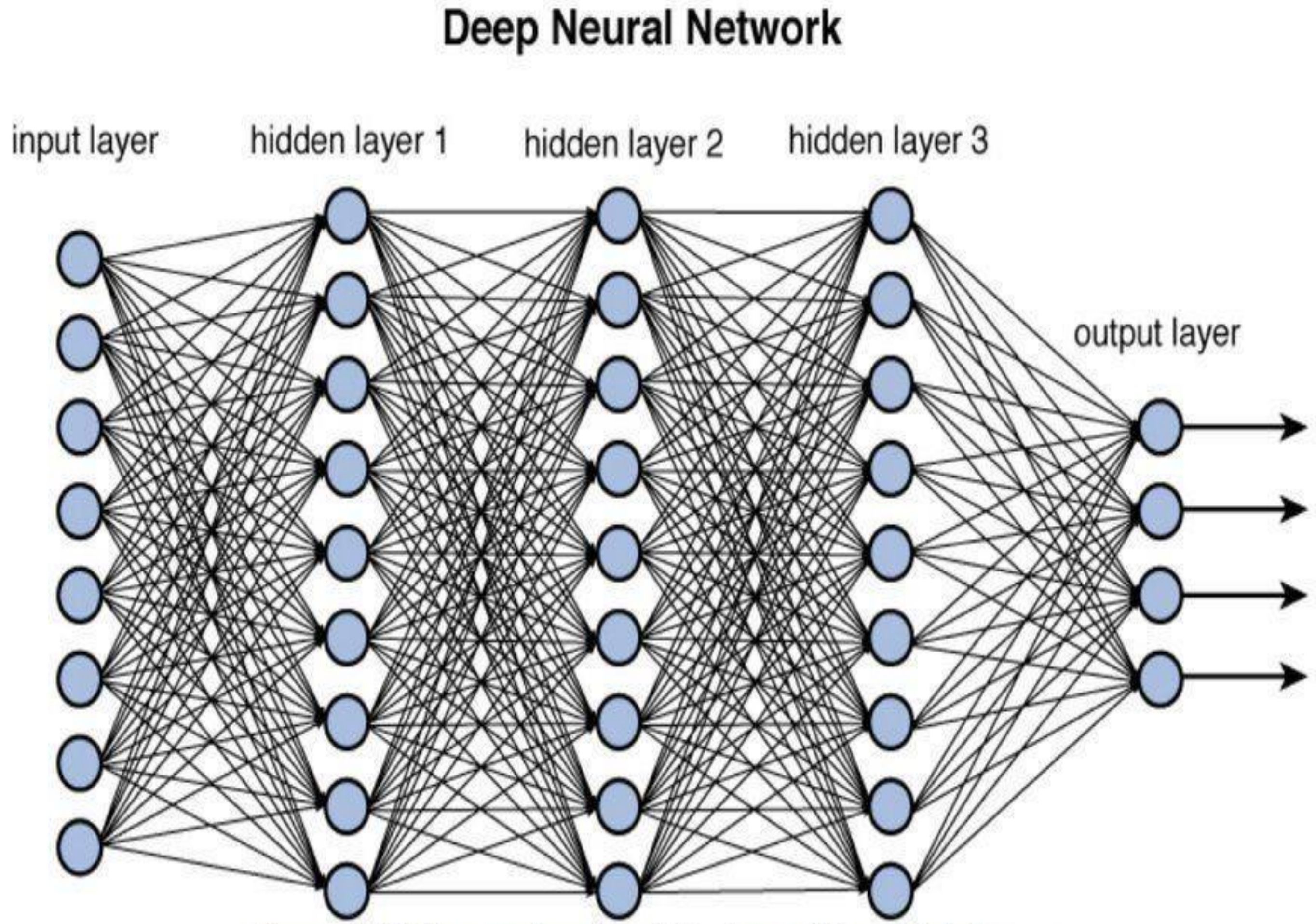


Figure 12.2 Deep network architecture with multiple layers.

Discussion:

You will be provided with a description of some data. The objective is to determine the expected lifetime value for customer segments.

Outline the methods that you would use to approach this task – including any pre-processing, transformations, etc.

Bonus – What data do you not have that you think you could use…and how would you get it?

KENNESAW STATE UNIVERSITY

Concepts to be Covered

- ➢The Evolution of Data Science
- ➢Defining Data Science – and Data Scientists
- ➢Demystifying Data Science
- ➢Exercise: Developing an Analytical Plan
- ➢Ethical Considerations in Data Science: Human Subjects and Algorithmic Bias
- ➢Exercise: Ethics Case Study
- ➢Summary and Wrap Up

KENNESAW STATE UNIVERSITY

Ethics in Data Science

The issues related to Ethics in Data Science are both broad and deep:

- Data Privacy
- Data Ownership
- "Algorithmic" Bias

Internet-based data collection

KENNESAW STATE UNIVERSITY

TIME

How Target Knew a High School Girl Was Pregnant Before Her Parents Did

By Keith Wagstaff @kwagstaff | Feb. 17, 2012

In Charles Duhigg's new piece for the New York *Times*, a father finds himself in the uncomfortable position of having to apologize to a Target employee. Earlier he had stormed into a store near Minneapolis and complained to the manager that his daughter was receiving coupons for cribs and baby clothes in the mail.

Turns out Target knew his daughter better than he did. She really was pregnant.

What is an IRB?

The Institutional Review Board (IRB) is an administrative body established to protect the rights and welfare of human research subjects recruited to participate in research activities conducted under the auspices of the institution with which it is affiliated.

The IRB is charged with the responsibility of reviewing, prior to its initiation, all research (whether funded or not) involving human participants. The IRB is concerned with protecting the welfare, rights, and privacy of human subjects.

KENNESAW STATE UNIVERSITY

The principles of ethical human subject studies:

- **Respect for persons** involves two ethical considerations: (1) individuals are and should be treated as autonomous agents and (2) individuals with diminished autonomy, due to youth, illness, mental disability, or restricted liberty (e.g., prisoners) should receive additional protections. The principle of respect for persons means recognizing the authority of an individual's preferences and choices about his or her life. In the context of research, the principle of respect for persons is expressed primarily in the use of informed consent, which requires that, as a general rule, individuals be afforded the opportunity to choose whether or not to be involved in research. It is incumbent upon investigators to disclose information about a study in language that is comprehensible to potential subjects so that they can provide meaningful and voluntary informed consent. These disclosures typically include the purpose of the research, the research procedures, risks, anticipated benefits (if any) to the subject, the opportunity to ask questions and receive satisfactory responses, and a statement that participation is voluntary and that the subject has the right to withdraw from the study at any time, for any reason.

- **Beneficence** involves two considerations: (1) the maximization of possible benefits for society and subjects; and (2) the minimization of possible harm to subjects. The principle of beneficence presents obligations that are woven throughout the research enterprise. Investigators, institutions, and sponsors must always endeavor to design and conduct research studies so that these obligations are met. Defining the optimum balance between the obligation to maximize benefit and minimize harm is often challenging. Notably, although the principle of beneficence refers to maximizing benefits for society, the Belmont Report does not expand upon this requirement.

- **Justice** is articulated in the Belmont Report as “fairness in distribution” of research benefits and burdens. Questions of justice and equal treatment in the research context are critical in the selection of subjects. The application of justice means that investigators must not offer potentially beneficial research only to some groups, nor select only some accessible, vulnerable, or disadvantaged groups for research that involves high risk or little prospect of direct benefit.

Which of the following on-line research strategies raises the most concerns regarding the ethical principle of respecting the autonomy of research subjects and the corresponding federal regulations requiring informed consent?

- ◯ A researcher proposes to join a moderated support group for cancer survivors posing as a survivor. She plans to insert comments to see how the members respond.
- ◯ A researcher observes the communications in an open support group without announcing her presence. She is interested in observing how long members participate and how the membership shifts over time.
- ◯ A linguist copies portions of postings on a political blog to document the use of expletives, abbreviations, and the use of irony in the postings.
- ◯ A researcher posts a notice on an open on-line support group for interracial adoptees asking anyone who would be interested in being interviewed for her study to contact her.

Taken from the quiz for the Social-Behavioral-Educational Internet Research Course in CITI

Facebook Manipulated User News Feeds To Create Emotional Responses

Gregory S. McNeal Contributor

Opinion

This article is more than 5 years old.

TWEET THIS

Facebook conducted a massive psychological experiment on 689,003 users, manipulating their news feeds to assess the effects on their emotions.

The short version is, Facebook has the ability to make you feel good or bad, just by tweaking what shows up in your news feed.

Facebook conducted a massive psychological experiment on 689,003 users, manipulating their news feeds to assess the effects on their emotions. The details of the experiment were published in an article entitled "Experimental Evidence Of Massive-Scale Emotional Contagion Through Social Networks" published in the journal *Proceedings of the National Academy of Sciences of the United States of America*.

KENNESAW STATE UNIVERSITY

Consider the following examples…is the collection of the data ethical? Why or why not?

Case Study: The Ethics of Using Hacked Data: Patreon's Data Hack and Academic Data Standards

April 4, 2016 / By Nathaniel Poor & Roei Davidson

Arguments in Favor of Use

1. Data is public, like a newspaper.
2. We hope to serve the public good via our work.
3. This is the data we want, but we can’t get it via other methods.

Arguments Against Use

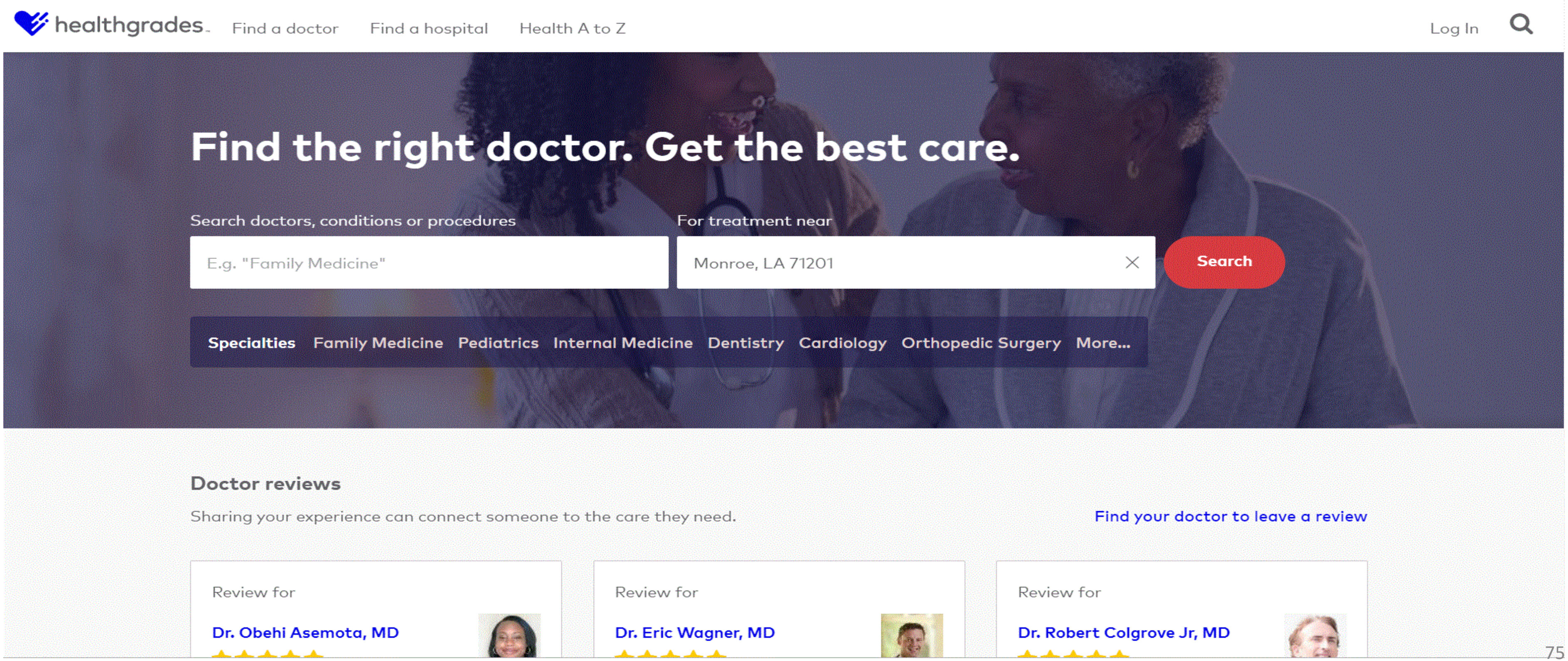
1. Researchers have a limited capability to distinguish between public and private information within the hacked data.
2. May see private data when cleaning the data.
3. Perhaps legitimizing criminal activity.
4. Violating users’ expectation of privacy.
5. Using people’s data without consent.
6. Other data can be ethically collected and used.

KENNESAW STATE UNIVERSITY

Fortnite-

An analyst with Epic creates an avatar to engage other avatars in Fortnite. They test different types of engagements to determine impacts on different types of behavior within the game – including purchase behavior, “success” in the game, interactions with other players. This information is then used to inform the next release of the game.

Healthgrades.com -



Healthgrades.com

What People Say About Dr. Eric Wagner, MD

Recent | Highest | Lowest | Most Helpful

★★★★★ Reply •••

I had a fracture and Dr. Wagner did my surgery. He did an amazing job and I am back to normal in just 4 months. He is very police, respectful, caring, and professional. I highly recommend him for the job he is doing.

Was this helpful?

Patel – Aug 23, 2019

★★★★★ Reply •••

Consider yourself very lucky if you live anywhere near Dr. Eric Wagner at Emory in Atlanta. I was in a car accident -- not my fault -- and the airbag broke my arm. Dr. Wagner handled the surgery so beautifully. I honestly felt like meeting this surgeon was one of the better aspects to the accident.... Read More

Was this helpful?

Wendy – Aug 22, 2019

In a 2016 study, the Pew Research Center found that 84% of all adults in the United States use online ratings sites to inform their product or service purchase decisions. The same is true for health care: patients increasingly access online ratings sites to inform their health care decisions, with online ratings emerging as the most influential factor for choosing a physician. In a 2017 study by the National Institutes of Health, 53% of physicians and 39% of patients reported visiting a health care rating website at least once. Overall, physicians indicated that the numerical results from these ratings websites were valid approximately 53% of the time, while patients indicated that they thought the ratings were valid 36% of the time.

This project will require you to “scrape” physician ratings data from the website healthgrades.com, and then find relevant patterns amongst numerical as well as text data (comments) that could inform medical practices.

KENNESAW STATE UNIVERSITY

…is the collection of the Fortnite and Healthgrades data ethical?

Does it stand the test of:

- ➢ Respect for Persons
- ➢ Beneficience
- ➢ Justice

Sources of “Algorithmic” Bias

Pre-Existing (Human)

Technical/ Sampling

Emergent

Pre-Existing Technical Emergent

"….a consequence of underlying and institutional ideologies…they may be explicit and conscious or implicit and unconscious…"

the NATIONAL BUREAU *of* ECONOMIC RESEARCH

Discrimination In The Age Of Algorithms

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, Cass R. Sunstein

NBER Working Paper No. 25548
Issued in February 2019
NBER Program(s): Children, Development Economics, Economics of Education, Health Care, Health Economics, Law and Economics, Labor Studies, Public Economics

"Algorithms are not only a threat to be regulated; with the right safeguards in place, they have the potential to be a positive force for equity."

KENNESAW STATE UNIVERSITY

Social Media and Credit Risk

Federal Deposit Insurance Corporation • Center for Financial Research

WORKING PAPER SERIES

On the Rise of the FinTechs—Credit Scoring using Digital Footprints

Pre-Existing | **Technical** | Emergent

"….emerges through limitations of a program, computational power, its design or other systemic constraint…

Uber crash shows 'catastrophic failure' of self-driving technology, experts say

Concerns raised about future testing as footage suggests fatal collision in Arizona was failing of system's most basic functions

- **Video released of fatal Uber self-driving crash**

MIT Researcher Exposing Bias in Facial Recognition Tech Triggers Amazon's Wrath

By Matt O'Brien | April 8, 2019

"…if AI systems are developed using images of mostly white men, the systems will work best in recognizing white men."

"Those disparities can sometimes be a matter of life or death…a study of the computer vision systems that enable self-driving cars to "see" the road shows they have a harder time detecting pedestrians with darker skin tones."

KENNESAW STATE UNIVERSITY

Pre-Existing Technical **Emergent**

"…result of the use and reliance on algorithms across new or unanticipated contexts…"

Unpredictable Correlations

Feedback Loops

KENNESAW STATE UNIVERSITY

HYPE

Microsoft's disastrous Tay experiment shows the hidden dangers of AI

By John West · April 2, 2016

- China plans to rank all its citizens based on their "social credit" by 2020.
- People can be rewarded or punished according to their scores.
- Like private financial credit scores, a person's social scores can move up and down according to their behavior.
- At the moment the system is piecemeal — some are run by city councils, while others are scored by private tech platforms that hold personal data.
- Scroll down to see how you can be punished or rewarded.

July 18, 2019

Algorithmic Accountability Act Introduced To Protect Against Bias In AI Systems

LinkedIn Facebook Twitter Send Embed

AI systems are being used for many applications, including facial recognition, determination of recidivism, and operation of autonomous vehicles. Some of the hardest problems with these systems are not in use of a neural network, but in the gathering the data that correlates with the outcomes to be predicted. In some cases, deficiencies in the datasets may result in outcomes that are biased against subgroups of people.

On April 10, US Senators Wyuden and Booker introduced the Algorithmic Accountability Act. This act intends to require companies to study automated decision systems to identify issues resulting in or contributing to inaccurate, unfair, biased, or discriminatory decisions impacting consumers. A copy of the Act is available here.

WRITTEN BY:

Brooks Kushmai

Contact + Foll

Isaac Slutsky

PUBLISHED IN:

Algorithms

Artificial Intelligence

Facial Recognition Technology

Science, Computers & Technology

What is the role of the academic community?

1) At a minimum, ensure students understand the law (e.g., GDPR)
2) Create awareness…ask questions at each stage of a project engagement – not just at the end.
3) Engage Institutional Research Boards
4) Don't shortcut the math/statistics

KENNESAW STATE UNIVERSITY

The PhD students in Analytics and Data took a course in ethical data science in Fall 2019. As a group of 21, they had to develop a series of guiding principles for ethical data science: https://datascience.kennesaw.edu/about/ethics.php

KENNESAW STATE UNIVERSITY

Concepts to be Covered

- ➢Overview of the Evolution of Data Science
- ➢Demystifying Data Science Methods
- ➢Table Talk: Machine Learning Case Study
- ➢Ethical Considerations in Data Science: Human Subjects
- ➢Table Talk: Ethics Case Study
- ➢Algorithmic Bias
- ➢Workshop Summary and Wrap Up

KENNESAW STATE UNIVERSITY

KENNESAW STATE UNIVERSITY

ANALYTICS AND DATA SCIENCE INSTITUTE

Jennifer Lewis Priestley, Ph.D. (jpriestl@kennesaw.edu)

datascience.kennesaw.edu

KENNESAW STATE UNIVERSITY