| Cancer Data Science 101 |

Presented by

NIH
NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics
& Information Technology

&

Frederick National Laboratory
for Cancer Research
sponsored by the National Cancer Institute

April 23, 2019

NCI Shady Grove, Terrace East, Seminar 408/410

Contents

Past presentations and additional resources are available at the following links:

- Presentations https://nciphub.org/groups/cancerdatascience/presentations
- Additional Resources https://docs.google.com/document/d/1fBBBBaN6-yXJdXM8qpilfGwcmgcjfaNY9Qr4M8NQqDk/edit

Organized by NCI Center for Bioinformatics and Information Technology (**CBIIT**) and

Frederick National Laboratory for Cancer Research (**FNLCR**)

AGENDA

Welcome and Introduction of speakers: NCI and FNL 1:00 pm – 1:10 pm

Session I: *What is Data Science?* 1:10 pm – 1:45 pm

Mark Jensen, PhD, **FNLCR BIDS**
- Definition
- History

Session II: *Data Science Methodology* 1:45 pm – 2:20 pm

Randall Johnson, PhD, **FNLCR ABCS**
- Methodology for cancer data science that includes knowledge of the problem, understanding the data, data preparation, etc.

PANEL Discussion: *Q & A for Sessions I and II* 2:20 pm – 2:45 pm

BREAK 2:45 pm – 2:55 pm

Session III: *Formulating the Question* 2:55 pm – 3:30 pm

Martin Skarzynski, PhD, **NCI DCEG**
- Understanding the research question
- Analytic Approach: use cases on
  - Regression
  - Classification
  - Clustering

Session IV: Data Gathering 3:30 pm – 4: 15 pm

Simina Boca, PhD, **Georgetown University Medical Center**
- Data requirements
  - Content, format, representation
- Data collection
- Data understanding
  - Descriptive statistics and visualization
  - Additional data collection
- Data preparation

PANEL Discussion: *Q & A for Sessions III and IV* 4:15 pm – 4:50 pm

Wrap-Up: *Feedback and Next Workshop* 4:50 pm – 5:00 pm

Session I: What is Data Science?

Instructor: **Mark Jensen, PhD, FNLCR BIDS**

History

Definition

Notes:____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

Session II: Data Science Methodology

Instructor: **Randall Johnson, PhD, FNLCR ABCS**

Methodology for cancer data science that includes knowledge of the problem, understanding the data, data preparation, etc.

Notes:______________________________________

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

Session III: Formulating the Research Question

Instructor: **Martin Skarzynski, PhD, NCI DCEG Fellow**

Understanding the research question

Analytic Approach: Use Cases on regression, classification, clustering

Notes:____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

Session IV: Data Gathering

Instructor: **Simina Boca, PhD, Georgetown University Medical Center**

Data requirements: Content, format, representation
Data collection
Data understanding: Descriptive statistics and visualization, additional data collection
Data preparation

Notes:_______________________________________

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

For More Information

NCI Scientific Computing Hub Site:
https://nciphub.org/groups/cancerdatascience

- **Presentations**: https://nciphub.org/groups/cancerdatascience/presentations
- **Bios**: https://nciphub.org/groups/cancerdatascience/members
- **Additional Resources for Training:** https://docs.google.com/document/d/1fBBBBaN6-yXJdXM8qpilfGwcmgcjfaNY9Qr4M8NQqDk/edit?usp=sharing

For information or assistance

- **Contact: Miles Kimbrough (Miles.Kimbrough@nih.gov)**

NCI/NIH Data Science Resources[*]

Bioinformatics

- **Bioinformatics User Forum** (https://abcsfrederick.info/BUF): The Bioinformatics User Forum is a place to build the bioinformatics community within the NCI and Frederick National Laboratory as well as among our collaborators at NIH, Ft Detrick and beyond. Anyone interested in the bioinformatics challenges in these communities is welcome to join us (search for “Bioinformatics-User-Forum” at list.nih.gov to sign up for our list serve).

- **Statistics for Lunch** (https://abcsfrederick.info/Stats4Lunch): Statistics for Lunch focuses on exploring topics in statistics at an accessible, intuitive level. We usually provide a high level, easy to understand seminar followed by a practical, hands-on session to practice what is covered in the seminar.

- **Programmer's Corner** (https://abcsfrederick.info/ProgrammersCorner): Programmer’s Corner is a quarterly meeting of programmers to discuss topics of interest to the group. We usually have 2 or 3 short talks centered around a common topic, followed by discussion.

Data Science

- **Data Science at NIH:** https://datascience.nih.gov/

- **Data Science Bowl -** a collection of tutorials on image segmentation https://www.kaggle.com/c/data-science-bowl-2017/kernels

- **CBIIT Computational Genomics and Bioinformatics Group:** https://datascience.cancer.gov/about/staff-directory/daoud-meerzaman#person-projects
  - Contact Dr. Daoud Meerzaman

[*] Partial list. See https://bit.ly/CancerDataScienceResources for the full list which includes advanced concepts.

- **Big Data to Knowledge (BD2K) Training Coordination Center** https://bigdatau.ini.usc.edu
    - Provides an educational resource discovery index to explore, aggregate and organize biomedical and data science educational content. The database contains more than 10,000 videos, documents, assignments, books and courses for personal and professional use.
- **NLM's University-based Biomedical Informatics and Data Science Research Training Programs** https://www.nlm.nih.gov/ep/GrantTrainInstitute.html
    - The U.S. National Library of Medicine supports research training in biomedical and data science at sixteen educational institutions in the United States. This website provides information related to Informatics Training and a list of academic institutions participating in the NLM training program.
- **HHS Data Science CoLab.** Brings together learners, practitioners, and collaborators to participate in ***Reimagine Data***, building a collaborative community around data science, and learning cohorts. The curriculum includes: GitHub, R programming, data visualization, foundational statistics, linear regression, and Machine Learning. https://www.hhs.gov/cto/initiatives/data-science-colab/index.html
- **NIH Data Science Distinguished Seminar Series.** A dynamic lecture series that brings high-profile researchers and experts to the NIH main campus to exemplify the roles that integration of the computational and the quantitative sciences play in the today's most innovative biomedical research, health research policy development, and sustainable research practices.https://datascience.nih.gov/community/datascience-at-nih/seminar
- **Cold Spring Harbor Laboratories Courses.** A list of courses, and workshops sponsored by the Cold Spring Harbor Laboratories in cancer research, cell biology, omics, neuroscience, genetics, immune/infect, neural data science, imaging structure, programing for biology, and computational genomics among other topics. https://meetings.cshl.edu/courseshome.aspx

Machine Learning

- **The Exascale CANcer Distributed Learning Environment (CANDLE)** https://cbiit.cancer.gov/ncip/hpc/candle
  - Using the challenges within each JDACS4C pilot to shape priorities for the CANDLE project, the DOE laboratories are drawing upon their strengths in high-performance computing (HPC), machine learning and data analytics, coupled with domain strengths at NCI and Frederick National Laboratory for Cancer Research to establish the foundations for CANDLE. Exploiting exascale technologies and capabilities anticipated for deep and machine learning, the project is scheduled to deliver critical technologies to the community that will be used to advance precision oncology.
  - CANDLE Workshops at NIH - https://wiki.nci.nih.gov/display/HPC/CANDLE+Workshops
- **Deep Learning in Oncology** A list of initiatives using Machine Learning to fight cancer. https://emerj.com/ai-sector-overviews/deep-learning-in-oncology/

Links to Training Resources

- **NIH Library training** - https://www.nihlibrary.nih.gov/training/calendar
- **Machine Learning Glossary -** https://developers.google.com/machine-learning/glossary/
- **NIH HPC Training** - https://hpc.nih.gov/training
- **Foundation for Advanced Education in the Sciences** (FAES). Graduate school for the NIH (https://faes.org)
  - Serves NIH sites in Bethesda, Frederick, Rockledge, Baltimore, Arizona, North Carolina, and Montana. Offers 150 academic for-credit courses per year, including over 60 biotechnology training workshops (Advanced Studies in Bioinformatics and Data Science)
  - **"Introduction to Text Mining," course, Fall 2019** (Instructed by Dr. Yifan Peng, NLM/NCBI and Dr. Qingyu Chen, NLM/NCBI) https://faes.org/sites/default/files/FAES-2018-19_Catalog_FINAL_1.pdf
  - **Course Catalog of all FAES Courses** https://faes.org/content/explore-courses

- **GitHub Resources for Continued Learning**
  https://github.com/justmarkham/DAT4/blob/master/resources.md

- **HHS Data Science Co-Lab** - Offers occasional NIH-wide training
  https://www.hhs.gov/idealab/dscolab/

- **NCI's Cancer Imaging Archive (TCIA)** http://www.cancerimagingarchive.net/

- NCI@Frederick Data Science Initiative - a collaboration between NCI and the University of Maryland. Offers a series of four workshops to provide a meeting ground for informal exchange of knowledge, ideas and networking to facilitate collaborative projects that will be carried out by UMD data science graduate and undergraduate students under the joint supervision of UMD and NCI researchers.

Glossary of Key Terms in Data Science[*]

A

algorithm

A series of repeatable steps for carrying out a certain type of task with data. As with data structures, people studying computer science learn about different algorithms and their suitability for various tasks. Specific data structures often play a role in how certain algorithms get implemented. See also data structure

AngularJS

An open-source JavaScript library maintained by Google and the AngularJS community that lets developers create what are known as Single [web] Page Applications. AngularJS is popular with data scientists as a way to show the results of their analysis. See also JavaScript, D3

artificial intelligence

Also, *AI*. The ability to have machines act with apparent intelligence, although varying definitions of "intelligence" lead to a range of meanings for the artificial variety. In AI's early days in the 1960s, researchers sought general principles of intelligence to implement, often using symbolic logic to automate reasoning. As the cost of computing resources dropped, the focus moved more toward statistical analysis of large amounts of data to drive decision making that gives the appearance of intelligence. See also machine learning, data mining

B

backpropagation

Also, *backprop*. An algorithm for iteratively adjusting the weights used in a neural network system. Backpropagation is often used to implement gradient descent. See also neural network, gradient descent

[*] Available online at http://www.datascienceglossary.org.

Bayes' Theorem

Also, *Bayes' Rule*. An equation for calculating the probability that something is true if something potentially related to it is true. If $P(A)$ means “the probability that A is true” and $P(A|B)$ means “the probability that A is true if B is true,” then Bayes' Theorem tells us that $P(A|B) = (P(B|A)P(A)) / P(B)$. This is useful for working with false positives—for example, if x% of people have a disease, the test for it is correct y% of the time, and you test positive, Bayes' Theorem helps calculate the odds that you have the disease.

The theorem also makes it easier to update a probability based on new data, which makes it valuable in the many applications where data continues to accumulate. Named for eighteenth-century English statistician and Presbyterian minister Thomas Bayes. See also Bayesian network, prior distribution

Bayesian network

Also, *Bayes net*. “Bayesian networks are graphs that compactly represent the relationship between random variables for a given problem. These graphs aid in performing reasoning or decision making in the face of uncertainty. Such reasoning relies heavily on Bayes’ rule.”[bourg] These networks are usually represented as graphs in which the link between any two nodes is assigned a value representing the probabilistic relationship between those nodes. See also Bayes' Theorem, Markov Chain

bias

In machine learning, “bias is a learner’s tendency to consistently learn the same wrong thing. Variance is the tendency to learn random things irrespective of the real signal.... It’s easy to avoid overfitting (variance) by falling into the opposite error of underfitting (bias). Simultaneously avoiding both requires learning a perfect classifier, and short of knowing it in advance there is no single technique that will always do best (no free lunch).”[domingos] See also variance, overfitting, classification

Big Data

As this has become a popular marketing buzz phrase, definitions have proliferated, but in general, it refers to the ability to work with collections of data that had been impractical before because of their volume, velocity, and variety (“the three Vs”). A key driver of this new ability has been easier distribution of storage and processing across networks of inexpensive commodity hardware using technology such as Hadoop instead of requiring larger, more powerful individual computers. The work done with these large amounts of data often draws on data science skills.

binomial distribution

A distribution of outcomes of independent events with two mutually exclusive possible outcomes, a fixed number of trials, and a constant probability of success. This is a discrete probability distribution, as opposed to continuous—for example, instead of graphing it with a line, you would use a histogram, because the potential outcomes are a discrete set of values. As the number of trials represented by a binomial distribution goes up, if the probability of success remains constant, the histogram bars will get thinner, and it will look more and more like a graph of normal distribution. See also probability distribution, discrete variable, histogram, normal distribution

C

chi-square test

Chi (pronounced like "pie" but beginning with a "k") is a Greek letter, and chi-square is "a statistical method used to test whether the classification of data can be ascribed to chance or to some underlying law."[websters] The chi-square test "is an analysis technique used to estimate whether two variables in a cross tabulation are correlated."[shin] A chi-square distribution varies from normal distribution based on the "degrees of freedom" used to calculate it. See also normal distribution and Wikipedia on the chi-squared test and on chi-squared distribution.

classification

The identification of which of two or more categories an item falls under; a classic machine learning task. Deciding whether an email message is spam or not classifies it among two categories, and analysis of data about movies might lead to classification of them among several genres. See also supervised learning, clustering

clustering

Any unsupervised algorithm for dividing up data instances into groups—not a predetermined set of groups, which would make this classification, but groups identified by the execution of the algorithm because of similarities that it found among the instances. The center of each cluster is known by the excellent name "centroid." See also classification, supervised learning, unsupervised learning, k-means clustering

coefficient

"A number or algebraic symbol prefixed as a multiplier to a variable or unknown quantity (Ex.: *x* in *x(y + z)*, *6* in *6ab*"[websters] When graphing an equation such as *y = 3x + 4*, the coefficient of *x* determines the line's slope. Discussions of statistics often mention specific coefficients for specific tasks such as the correlation coefficient, Cramer's coefficient, and the Gini coefficient. See also correlation

computational linguistics

Also, *natural language processing*, *NLP*. A branch of computer science for parsing text of spoken languages (for example, English or Mandarin) to convert it to structured data that you can use to drive program logic. Early efforts focused on translating one language to another or accepting complete sentences as queries to databases; modern efforts often analyze documents and other data (for example, tweets) to extract potentially valuable information. See also GATE, UIMA

confidence interval

A range specified around an estimate to indicate margin of error, combined with a probability that a value will fall in that range. The field of statistics offers specific mathematical formulas to calculate confidence intervals.

continuous variable

A variable whose value can be any of an infinite number of values, typically within a particular range. For example, if you can express age or size with a decimal number, then they are continuous variables. In a graph, the value of a continuous variable is usually expressed as a line plotted by a function. Compare discrete variable

correlation

"The degree of relative correspondence, as between two sets of data."[websters] If sales go up when the advertising budget goes up, they correlate. The *correlation coefficient* is a measure of how closely the two data sets correlate. A correlation coefficient of 1 is a perfect correlation, .9 is a strong correlation, and .2 is a weak correlation. This value can also be negative, as when the incidence of a disease goes down when vaccinations go up. A correlation coefficient of -1 is a perfect negative correlation. Always remember, though, that correlation does not imply causation. See also coefficient

covariance

“A measure of the relationship between two variables whose values are observed at the same time; specifically, the average value of the two variables diminished by the product of their average values.”[websters] “Whereas variance measures how a single variable deviates from its mean, covariance measures how two variables vary in tandem from their means.”[grus] See also variance, mean

cross-validation

When using data with an algorithm, “the name given to a set of techniques that divide up data into training sets and test sets. The training set is given to the algorithm, along with the correct answers... and becomes the set used to make predictions. The algorithm is then asked to make predictions for each item in the test set. The answers it gives are compared to the correct answers, and an overall score for how well the algorithm did is calculated.”[segaran] See also machine learning

D

D3

“Data-Driven Documents.” A JavaScript library that eases the creation of interactive visualizations embedded in web pages. D3 is popular with data scientists as a way to present the results of their analysis. See also AngularJS, JavaScript

data engineer

A specialist in data wrangling. “Data engineers are the ones that take the messy data... and build the infrastructure for real, tangible analysis. They run ETL software, marry data sets, enrich and clean all that data that companies have been storing for years.”[biewald] See also data wrangling. (A Wikipedia search for “data engineering” redirects to “information engineering,” an older term that describes a more enterprise-oriented job with greater system architecture responsibility and less hands-on work with the data.)

data mining

Generally, the use of computers to analyze large data sets to look for patterns that let people make business decisions. While this sounds like much of what data science is about, popular use of the term is much older, dating back at least to the 1990s. See also data science

data science

“The ability to extract knowledge and insights from large and complex data sets.”[patil] Data science work often requires knowledge of both statistics and software engineering. See also data engineer, machine learning

data structure

A particular arrangement of units of data such as an array or a tree. People studying computer science learn about different data structures and their suitability for various tasks. See also algorithm

data wrangling

Also, *data munging*. The conversion of data, often through the use of scripting languages, to make it easier to work with. If you have 900,000 *birthYear* values of the format yyyy-mm-dd and 100,000 of the format mm/dd/yyyy and you write a Perl script to convert the latter to look like the former so that you can use them all together, you're doing data wrangling. Discussions of data science often bemoan the high percentage of time that practitioners must spend doing data wrangling; the discussions then recommend the hiring of data engineers to address this. See also Perl, Python, shell, data engineer

decision trees

“A decision tree uses a tree structure to represent a number of possible decision paths and an outcome for each path. If you have ever played the game Twenty Questions, then it turns out you are familiar with decision trees.”[grus] See also random forest

deep learning

Typically, a multi-level algorithm that gradually identifies things at higher levels of abstraction. For example, the first level may identify certain lines, then the next level identifies combinations of lines as shapes, and then the next level identifies combinations of shapes as specific objects. As you might guess from this example, deep learning is popular for image classification. See also neural network

dependent variable

The value of a dependent value “depends” on the value of the independent variable. If you're measuring the effect of different sizes of an advertising budget on total sales, then the advertising budget figure is the independent variable and total sales is the dependent variable.

dimension reduction

Also, *dimensionality reduction*. “We can use a technique called *principal component analysis* to extract one or more dimensions that capture as much of the variation in the data as possible... Dimensionality reduction is mostly useful when your data set has a large number of dimensions and you want to find a small subset that captures most of the variation.”[grus] Linear algebra can be involved; “broadly speaking, linear algebra is about translating something residing in an *m*-dimensional space into a corresponding shape in an *n*-dimensional space.”[shin] See also linear algebra

discrete variable

A variable whose potential values must be one of a specific number of values. If someone rates a movie with between one and five stars, with no partial stars allowed, the rating is a discrete variable. In a graph, the distribution of values for a discrete variable is usually expressed as a histogram. See also continuous variable, histogram

E

econometrics

“The use of mathematical and statistical methods in the field of economics to verify and develop economic theories”[websters]

F

feature

The machine learning expression for a piece of measurable information about something. If you store the age, annual income, and weight of a set of people, you're storing three features about them. In other areas of the IT world, people may use the terms property, attribute, or field instead of “feature.” See also feature engineering

feature engineering

“To obtain a good model, however, often requires more effort and iteration and a process called feature engineering. Features are the model’s inputs. They can involve basic raw data that you have collected, such as order amount, simple derived variables, such as ‘Is order date on a weekend? Yes/No,’ as well as more complex abstract features, such as the ‘similarity score’ between two movies. Thinking up features is as much an art as a science and can rely on domain knowledge.”[anderson] See also feature

G

GATE

"General Architecture for Text Engineering," an open source, Java-based framework for natural language processing tasks. The framework lets you pipeline other tools designed to be plugged into it. The project is based at the UK's University of Sheffield. See also computational linguistics, UIMA

gradient boosting

"Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function."[wikipediagb]

gradient descent

An optimization algorithm for finding the input to a function "that produces the largest (or smallest) possible value... one approach to maximizing a function is to pick a random starting point, compute the gradient, take a small step in the direction of the gradient (i.e., the direction that causes the function to increase the most), and repeat with the new starting point. Similarly, you can try to minimize a function by taking small steps in the opposite direction."[grus] See also backpropagation

H

histogram

A graphical representation of the distribution of a set of numeric data, usually a vertical bar graph. See also probability distribution, binomial distribution, discrete variable

I

independent variable

See dependent variable

J

JavaScript

A scripting language (no relation to Java) originally designed in the mid-1990s for embedding logic in web pages, but which later evolved into a more general-purpose development language. JavaScript continues to be very popular for embedding logic in web pages, with many libraries available to enhance the operation and visual presentation of these pages. See also AngularJS, D3

K

k-means clustering

"A data mining algorithm to cluster, classify, or group your *N* objects based on their attributes or features into *K* number of groups (so-called clusters)."[parsian] See also clustering

k-nearest neighbors

Also, *kNN*. A machine learning algorithm that classifies things based on their similarity to nearby neighbors. You tune the algorithm's execution by picking how many neighbors to examine (*k*) as well as some notion of "distance" to indicate how near the neighbors are. For example, in a social network, a friend of your friend could be considered twice the distance away from you as your friend. "Similarity" would be comparison of feature values in the neighbors being compared. See also classification, feature

L

latent variable

"In statistics, latent variables (from Latin: present participle of *lateo* ('lie hidden'),as opposed to observable variables), are variables that are not directly observed but are rather inferred (through a mathematical model) from other variables that are observed (directly measured). Mathematical models that aim to explain observed variables in terms of latent variables are called latent variable models."[wikipedialv]

lift

"Lift compares the frequency of an observed pattern with how often you'd expect to see that pattern just by chance... If the lift is near 1, then there's a good chance that the pattern you observed is occurring just by chance. The larger the lift, the more likely that the pattern is 'real.'"[zumel]

linear algebra

A branch of mathematics dealing with vector spaces and operations on them such as addition and multiplication. “Linear algebra is designed to represent systems of linear equations. Linear equations are designed to represent linear relationships, where one entity is written to be a sum of multiples of other entities. In the shorthand of linear algebra, a linear relationship is represented as a linear operator—a matrix.”[zheng] See also vector, vector space, matrix, coefficient

linear regression

A technique to look for a linear relationship (that is, one where the relationship between two varying amounts, such as price and sales, can be expressed with an equation that you can represent as a straight line on a graph) by starting with a set of data points that don't necessarily line up nicely. This is done by computing the “least squares” line: the one that has, on an x-y graph, the smallest possible sum of squared distances to the actual data point *y* values. Statistical software packages and even typical spreadsheet packages offer automated ways to calculate this. People who get excited about machine learning often apply it to problems that would have been much simpler by using linear regression in an Excel spreadsheet. See also regression, logistic regression, machine learning

logarithm

If $y = 10^x$, then $log(y) = x$. Working with the log of one or more of a model's variables, instead of their original values, can make it easier to model relationships with linear functions instead of non-linear ones. Linear functions are typically easier to use in data analysis. (The $log(y) = x$ example shown is for log base 10. Natural logarithms, or log base *e*—where *e* is a specific irrational number a little over 2.7—are a bit more complicated but also very useful for related tasks.) See also dependent variable, linear regression

logistic regression

A model similar to linear regression but where the potential results are a specific set of categories instead of being continuous. See continuous variable, regression, linear regression

M

machine learning

The use of data-driven algorithms that perform better as they have more data to work with, “learning” (that is, refining their models) from this additional data. This often involves cross-validation with training and test data sets. “The fundamental goal of machine learning is to generalize beyond the examples in the training set.”[domingos] Studying the practical application of machine learning usually means researching which machine learning algorithms are best for which situations. See also algorithm, cross-validation, artificial intelligence

machine learning model

"The process of training an ML model involves providing an ML algorithm (that is, the learning algorithm) with training data to learn from. The term ML model refers to the model artifact that is created by the training process." [amazonml] See also algorithm, machine learning, model

Markov Chain

An algorithm for working with a series of events (for example, a system being in particular states) to predict the possibility of a certain event based on which other events have happened. The identification of probabilistic relationships between the different events means that Markov Chains and Bayesian networks often come up in the same discussions. See also Bayesian network, Monte Carlo method

MATLAB

A commercial computer language and environment popular for visualization and algorithm development.

matrix

(Plural: *matrices*) An older Webster’s dictionary with a heavier emphasis on typographical representation gives the mathematical definition as “a set of numbers or terms arranged in rows and columns between parentheses or double lines”[websters].

For purposes of manipulating a matrix with software, think of it as a two-dimensional array. As with its one-dimensional equivalent, a vector, this mathematical representation of the two-dimensional array makes it easier to take advantage of software libraries that apply advanced mathematical operations to the data—including libraries that can distribute the processing across multiple processors for scalability. See also vector, linear algebra

mean

The average value, although technically that is known as the "arithmetic mean." (Other means include the geometric and harmonic means.) See also median, mode

Mean Absolute Error

Also, *MAE*. The average error of all predicted values when compared with observed values. See also Mean Squared Error, Root Mean Squared Error

Mean Squared Error

Also, *MSE*. The average of the squares of all the errors found when comparing predicted values with observed values. Squaring them makes the bigger errors count for more, making Mean Squared Error more popular than Mean Absolute Error when quantifying the success of a set of predictions. See also Mean Absolute Error, Root Mean Squared Error

median

When values are sorted, the value in the middle, or the average of the two in the middle if there are an even number of values. See also mean, mode

mode

"The value that occurs most often in a sample of data. Like the median, the mode cannot be directly calculated"[stanton] although it's easy enough to find with a little scripting. For people who work with statistics, "mode" can also mean "data type"—for example, whether a value is an integer, a real number, or a date. See also mean, median, scripting

model

"A specification of a mathematical (or probabilistic) relationship that exists between different variables."[grus] Because "modeling" can mean so many things, the term "statistical modeling" is often used to more accurately describe the kind of modeling that data scientists do.

Monte Carlo method

In general, the use of randomly generated numbers as part of an algorithm. Its use with Markov Chains is so popular that people usually refer to the combination with the acronym MCMC. See also Markov Chain

moving average

"The mean (or average) of time series data (observations equally spaced in time, such as per hour or per day) from several consecutive periods is called the *moving average*. It is called *moving* because the average is continually recomputed as new time series data becomes available, and it progresses by dropping the earliest value and adding the most recent."[parsian] See also mean, time series data

N

n-gram

The analysis of sequences of *n* items (typically, words in natural language) to look for patterns. For example, trigram analysis examines three-word phrases in the input to look for patterns such as which pairs of words appear most often in the groups of three. The value of *n* can be something other than three, depending on your needs. This helps to construct statistical models of documents (for example, when automatically classifying them) and to find positive or negative terms associated with a product name. See also computational linguistics, classification

naive Bayes classifier

"A collection of classification algorithms based on Bayes Theorem. It is not a single algorithm but a family of algorithms that all share a common principle, that every feature being classified is independent of the value of any other feature.

So, for example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A Naive Bayes classifier considers each of these 'features' (red, round, 3" in diameter) to contribute independently to the probability that the fruit is an apple, regardless of any correlations between features. Features, however, aren't always independent which is often seen as a shortcoming of the Naive Bayes algorithm and this is why it's labeled 'naive'."[aylien] This naiveté makes it much easier to develop implementations of these algorithms that scale way up. See also Bayes' Theorem, classification

neural network

Also, *neural net* or *artificial neural network* to distinguish it from the brain, upon which this algorithm is modeled. “A robust function that takes an arbitrary set of inputs and fits it to an arbitrary set of outputs that are binary... In practice, Neural Networks are used in deep learning research to match images to features and much more. What makes Neural Networks special is their use of a hidden layer of weighted functions called neurons, with which you can effectively build a network that maps a lot of other functions. Without a hidden layer of functions, Neural Networks would be just a set of simple weighted functions.”[kirk] See also deep learning, backpropagation, perceptron

normal distribution

Also, *Gaussian distribution*. (Carl Friedrich Gauss was an early nineteenth-century German mathematician.) A probability distribution which, when graphed, is a symmetrical bell curve with the mean value at the center. The standard deviation value affects the height and width of the graph. See also mean, probability distribution, standard deviation, binomial distribution, standard normal distribution

NoSQL

A database management system that uses any of several alternatives to the relational, table-oriented model used by SQL databases. While this term originally meant “not SQL,” it has come to mean something closer to “not only SQL” because the specialized nature of NoSQL database management systems often have them playing specific roles in a larger system that may also include SQL and additional NoSQL systems. See also SQL

null hypothesis

If your proposed model for a data set says that the value of *x* is affecting the value of *y*, then the null hypothesis—the model you're comparing your proposed model with to check whether *x* really is affecting *y*—says that the observations are all based on chance and that there is no effect. “The smaller the P-value computed from the sample data, the stronger the evidence is against the null hypothesis.”[shin] See also P value

O

objective function

“When you want to get as much (or as little) of something as possible, and the way you’ll get it is by changing the values of other quantities, you have an optimization problem...To solve an optimization problem, you need to combine your decision variables, constraints, and the thing you want to maximize together into an objective function. The objective is the thing you want to maximize or minimize, and you use the objective function to find the optimum result.”[milton] See also gradient descent

outlier

“Extreme values that might be errors in measurement and recording, or might be accurate reports of rare events.”[downey] See also overfitting

overfitting

A model of training data that, by taking too many of the data's quirks and outliers into account, is overly complicated and will not be as useful as it could be to find patterns in test data. See also outlier, cross-validation

P

P value

Also, *p-value*. “The probability, under the assumption of no effect or no difference (the null hypothesis), of obtaining a result equal to or more extreme than what was actually observed.”[goodman] “It’s a measure of *how surprised you should be* if there is no actual difference between the groups, but you got data suggesting there is. A bigger difference, or one backed up by more data, suggests more surprise and a smaller *p* value...The *p* value is a measure of surprise, not a measure of the size of the effect.”[reinhart] A lower *p* value means that your results are more statistically significant. See also null hypothesis

PageRank

An algorithm that determines the importance of something, typically to rank it in a list of search results. “PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.”[googlearchive] PageRank is not named for the pages that it ranks but for its inventor, Google co-founder and CEO Larry Page.

Pandas

A Python library for data manipulation popular with data scientists. See also Python

perceptron

"Pretty much the simplest neural network is the perceptron, which approximates a single neuron with *n* binary inputs. It computes a weighted sum of its inputs and 'fires' if that weighted sum is zero or greater."[grus] See also neural network

Perl

An older scripting language with roots in pre-Linux UNIX systems. Perl has always been popular for text processing, especially data cleanup and enhancement tasks. See also scripting, data wrangling

pivot table

"Pivot tables quickly summarize long lists of data, without requiring you to write a single formula or copy a single cell. But the most notable feature of pivot tables is that you can arrange them *dynamically*. Say you create a pivot table summary using raw census data. With the drag of a mouse, you can easily rearrange the pivot table so that it summarizes the data based on gender *or* age groupings *or* geographic location. The process of rearranging your table is known as *pivoting* your data: you're turning the same information around to examine it from different angles. [macdonald]

Poisson distribution

A distribution of independent events, usually over a period of time or space, used to help predict the probability of an event. Like the binomial distribution, this is a discrete distribution. Named for early 19th century French mathematician Siméon Denis Poisson. See also spatiotemporal data, discrete variable, binomial distribution

posterior distribution

See prior distribution

predictive analytics

The analysis of data to predict future events, typically to aid in business planning. This incorporates predictive modeling and other techniques. Machine learning might be considered a set of algorithms to help implement predictive analytics. The more business-oriented spin of "predictive analytics" makes it a popular buzz phrase in marketing literature. See also predictive modeling, machine learning, SPSS

predictive modeling

The development of statistical models to predict future events. See also predictive analytics, model

principal component analysis

“This algorithm simply looks at the direction with the most variance and then determines that as the first principal component. This is very similar to how regression works in that it determines the best direction to map data to.”[kirk] See also regression

prior distribution

“In Bayesian inference, we assume that the unknown quantity to be estimated has many plausible values modeled by what's called a prior distribution. Bayesian inference is then using data (that is considered as unchanging) to build a tighter posterior distribution for the unknown quantity.”[zumel] See also Bayes' Theorem

probability distribution

“A probability distribution for a discrete random variable is a listing of all possible distinct outcomes and their probabilities of occurring. Because all possible outcomes are listed, the sum of the probabilities must add to 1.0.”[levine] See also discrete variable

Python

A programming language available since 1994 that is popular with people doing data science. Python is noted for ease of use among beginners and great power when used by advanced users, especially when taking advantage of specialized libraries such as those designed for machine learning and graph generation. See also scripting, Pandas

Q

quantile, quartile

When you divide a set of sorted values into groups that each have the same number of values (for example, if you divide the values into two groups at the median), each group is known as a quantile. If there are four groups, we call them quartiles, which is a common way to divide values for discussion and analysis purposes; if there are five, we call them quintiles, and so forth. See also median

R

R (Programming Language)

An open-source programming language and environment for statistical computing and graph generation available for Linux, Windows, and Mac.

random forest

An algorithm used for regression or classification that uses a collection of tree data structures. "To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree 'votes' for that class. The forest chooses the classification having the most votes (over all the trees in the forest)."[breiman] The term "random forest" is actually trademarked by its authors. See also classification, vector, decision trees

regression

"...the more general problem of fitting any kind of model to any kind of data. This use of the term 'regression' is a historical accident; it is only indirectly related to the original meaning of the word."[downey] See also linear regression, logistic regression, principal component analysis

reinforcement learning

A class of machine learning algorithms in which the process is not given specific goals to meet but, as it makes decisions, is instead given indications of whether it's doing well or not. For example, an algorithm for learning to play a video game knows that if its score just went up, it must have done something right. See also supervised learning, unsupervised learning

Root Mean Squared Error

Also, *RMSE*. The square root of the Mean Squared Error. This is more popular than Mean Squared Error because taking the square root of a figure built from the squares of the observation value errors gives a number that's easier to understand in the units used to measure the original observations. See also Mean Absolute Error, Mean Squared Error,

Ruby

A scripting language that first appeared in 1996. Ruby is popular in the data science community, but not as popular as Python, which has more specialized libraries available for data science tasks. See also scripting, Python

S

S curve

Imagine a graph showing, for each month since smartphones originally became available, how many people in the US bought their first one. The line would rise slowly at first, when only the early adopters got them, then quickly as these phones became more popular, and then level off again once nearly everyone had one. This graph's line would form a stretched-out “S” shape. The “S curve” applies to many other phenomena and is often mentioned when someone predicts that a rising value will eventually level off.

SAS

A commercial statistical software suite that includes a programming language also known as SAS.

scalar

“Designating or of a quantity that has magnitude but no direction in space, as volume or temperature — ***n.*** a scalar quantity: distinguished from vector”[websters] See also vector

scripting

Generally, the use of a computer language where your program, or script, can be run directly with no need to first compile it to binary code as with with languages such as Java and C. Scripting languages often have simpler syntax than compiled languages, so the process of writing, running, and tweaking scripts can go faster. See also Python, Perl, Ruby, shell

serial correlation

“As prices vary from day to day, you might expect to see patterns. If the price is high on Monday, you might expect it to be high for a few more days; and if it’s low, you might expect it to stay low. A pattern like this is called serial correlation, because each value is correlated with the next one in the series.

To compute serial correlation, we can shift the time series by an interval called a lag, and then compute the correlation of the shifted series with the original... 'Autocorrelation' is another name for serial correlation, used more often when the lag is not 1.”[downey] See also correlation

shell

When you use a computer's operating system from the command line, you're using its shell. Along with scripting languages such as Perl and Python, Linux-based shell tools (which are either included with or easily available for Mac and Windows machines) such as grep, diff, split, comm, head, and tail are popular for data wrangling. A series of shell commands stored in a file that lets you execute the series by entering the file's name is known as a shell script. See also data wrangling, scripting, Perl, Python

spatiotemporal data

Time series data that also includes geographic identifiers such as latitude-longitude pairs. See also time series data

SPSS

A commercial statistical software package, or according to the product home page, "predictive analytics software."[spss] The product has always been popular in the social sciences. The company, founded in 1968, was acquired by IBM in 2009. See also predictive analytics

SQL

The ISO standard query language for relational databases. Variations of this extremely popular language are often available for data storage systems that aren't strictly relational; watch for the phrase "SQL-like."

standard deviation

The square root of the variance, and a common way to indicate just how different a particular measurement is from the mean. "An observation more than three standard deviations away from the mean can be considered quite rare, in most applications."[zumel] Statistical software packages offer automated ways to calculate the standard deviation. See also variance

standard normal distribution

A normal distribution with a mean of 0 and a standard deviation of 1. When graphed, it's a bell-shaped curve centered around the *y* axis, where *x*=0. See also normal distribution, mean, standard deviation

standardized score

Also, *standard score*, *normal score*, *z-score*. “Transforms a raw score into units of standard deviation above or below the mean. This translates the scores so they can be evaluated in reference to the standard normal distribution.”[boslaugh] Translating two different test sets to use standardized scores makes them easier to compare. See also standard deviation, mean, standard normal distribution

Stata

A commercial statistical software package, not to be confused with strata. See also strata, stratified sampling

strata, stratified sampling

“Divide the population units into homogeneous groups (strata) and draw a simple random sample from each group.”[gonick] Strata also refers to an O'Reilly conference on big data, data science, and related technologies. See also Stata

supervised learning

A type of machine learning algorithm in which a system is taught to classify input into specific, known classes. The classic example is sorting email into spam versus ham. See also unsupervised learning, reinforcement learning, machine learning

support vector machine

Also, *SVM*. Imagine that you want to write a function that draws a line on a two-dimensional *x*-*y* graph that separates two different kinds of points—that is, it classifies them into two categories—but you can't, because on that graph they're too mixed together.

Now imagine that the points are in three dimensions, and you can classify them by writing a function that describes a plane that can be positioned at any angle and position in those three dimensions, giving you more opportunities to find a working mathematical classifier. This plane that is one dimension less than the space around it, such as a two-dimensional plane in a three-dimensional space or a one-dimensional line on a two-dimensional space, is known as a hyperplane.

A support vector machine is a supervised learning classification tool that seeks a dividing hyperplane for any number of dimensions. (Keep in mind that “dimensions” don't have to be *x*, *y*, and *z* position coordinates, but any features you choose to drive the categorization.) SVMs have also been used for regression tasks as well as categorization tasks. See also supervised learning, feature

T

t-distribution

Also, *student's t distribution*. A variation on normal distribution that accounts for the fact that you're only using a sampling of all the possible values instead of all of them. Invented by Guiness Brewery statistician William Gossett (publishing under the pseudonym "student") in the early 20th century for his quality assurance work there. See also normal distribution

Tableau

A commercial data visualization package often used in data science projects.

time series data

"Strictly speaking, a time series is a sequence of measurements of some quantity taken at different times, often but not necessarily at equally spaced intervals."[boslaugh] So, time series data will have measurements of observations (for example, air pressure or stock prices) accompanied by date-time stamps. See also spatiotemporal data, moving average

U

UIMA

The "Unstructured Information Management Architecture" was developed at IBM as a framework to analyze unstructured information, especially natural language. OASIS UIMA is a specification that standardizes this framework and Apache UIMA is an open-source implementation of it. The framework lets you pipeline other tools designed to be plugged into it. See also computational linguistics, GATE

unsupervised learning

A class of machine learning algorithms designed to identify groupings of data without knowing in advance what the groups will be. See also supervised learning, reinforcement learning, clustering

V

variance

"How much a list of numbers varies from the mean (average) value. It is frequently used in statistics to measure how large the differences are in a set of numbers. It is calculated by averaging the squared difference of every number from the mean."[segaran] Any statistical package will offer an automated way to calculate this. See also mean, bias, standard deviation

vector

Webster's first mathematical definition is "a mathematical expression denoting a combination of magnitude and direction," which you may remember from geometry class, but their third definition is closer to how data scientists use the term: "an ordered set of real numbers, each denoting a distance on a coordinate axis"[websters]. These numbers may represent a series of details about a single person, movie, product, or whatever entity is being modeled. This mathematical representation of the set of values makes it easier to take advantage of software libraries that apply advanced mathematical operations to the data. See also matrix, linear algebra

vector space

A collection of vectors—for example, a matrix. See also vector, matrix, linear algebra

W

Weka

An open source set of command line and graphical user interface data analysis tools developed at the University of Waikato in New Zealand.

Glossary References

Amazon Machine Learning Developer Guide, Training ML Models, accessed 2018-04-15.

Carl Anderson, *Creating a Data-Driven Organization* (Sebastopol: O'Reilly Media, 2015).

Naive Bayes for Dummies; A Simple Explanation, accessed 2015-08-21.

Sarah Boslaugh, *Statistics in a Nutshell*, 2nd Edition (Sebastopol: O'Reilly Media, 2012).

David M. Bourg and Glenn Seeman *AI for Game Developers* (Sebastopol: O'Reilly Media, 2004).

Pedro Domingos, A Few Useful Things to Know about Machine Learning, accessed 2015-09-05.

Leo Breiman and Adele Cutler, Random Forests, accessed 2015-08-22.

Allen B. Downey *Think Stats*, 2nd Edition (Sebastopol: O'Reilly Media, 2014).

Larry Gonick and Woolcott Smith, The Cartoon Guide to Statistics (New York: HarperCollins, 1993)

S. N. Goodman, *Toward evidence-based medical statistics. 1: The P value fallacy*. Annals of Internal Medicine, 130:995–1004, 1999. (quoted in Reinhart)

Matthew Kirk, *Thoughtful Machine Learning* (Sebastopol: O'Reilly Media, 2014).

Matthew MacDonald, *Excel 2013: The Missing Manual* (Sebastopol: O'Reilly Media, 2013).

Michael Milton, *Head First Data Analysis* (Sebastopol: O'Reilly Media, 2009).

DJ Patil, A Memo to the American People from U.S. Chief Data Scientist Dr. DJ Patil, 2015-02-15

Alex Reinhart, "An introduction to data analysis" in Statistics Done Wrong: The Woefully Complete Guide (San Francisco: No Starch Press, 2015)

Facts about Google and Competition, archive.org version accessed 2015-09-05.

Joel Grus, *Data Science from Scratch: First Principles with Python*, (Sebastopol: O'Reilly Media, 2015).

David M. Levine, *Statistics for Six Sigma Green Belts with Minitab and JMP* (Upper Saddle River: Pearson, 2006).

Mahmoud Parsian, *Data Algorithms*, (Sebastopol: O'Reilly Media, 2015).

Toby Segaran, *Programming Collective Intelligence*, (Sebastopol: O'Reilly Media, 2015).

Shin Takahashi, *The Manga Guide to Statistics*, (Sebastopol: O'Reilly Media, 2008).

SPSS Software, accessed 2015-08-22.

82. Stanton, J.M. (2012). *Introduction to Data Science*, Third Edition. iTunes Open Source eBook. Available: https://itunes.apple.com/us/book/introduction-to-data-science/id529088127?mt=11

Victoria Neufeldt, Editor in Chief, *Webster's New World College Dictionary* , Third Edition (New York: Macmillan, 1997).

Wikipedia: Gradient boosting, accessed 2016-02-28.

Wikipedia: Latent variable, accessed 2016-02-28.

Alice Zheng, *Striking parallels between mathematics and software engineering*, accessed 2015-09-11.

Nina Zumel and John Mount, *Practical Data Science with R* (Shelter Island: Manning Publications, 2014).

NIH NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics
& Information Technology

Frederick National Laboratory
for Cancer Research
sponsored by the National Cancer Institute

Please Share Your Feedback on Today's Workshop!

This Feedback form is also available online at bit.ly/CDS101feedback

Your feedback is very important to us. It helps us understand your needs and improve the quality of our workshops.

On a scale of 1 through 5, 1 = strongly disagree, 2 = disagree, 3=somewhat agree, 4=Agree, 5=Strongly agree. (circle one for each question below)

1) The workshop was a good use of my time.

1 2 3 4 5

2) I would like to know more about:

3) In future workshops, I would like to see:

4) How would you compare your overall experience in this workshop to others you have attended?

1	2	3	4	5
Won't attend another one	Not very beneficial	Average	Surprisingly good	GREAT! Can't wait for the next one!

5) This workshop inspired me to further explore Cancer Data Science.

1 2 3 4 5

6) Did this workshop help you create/strengthen relationships with potential future collaborators?

No, none Met one or two Yes, a few Too many to count

Other:____

7) What else would you like to add about your experience?

8) Would you be interested in contributing to the workshop series as an instructor/presenter in future workshops?

Yes, here is my contact information (*please include your division, office or center*):

Name/DOC____

Expertise: ____

Email: ____

Phone: ____

No, but I suggest you contact the following as a potential instructor(s):

Name/DOC: ____

Contact Info: ____

Name/DOC: ____

Contact Info:____

Thank you very much for your participation and feedback!