FUNDAMENTALS OF DATA SCIENCE

...

MODULE 1

IN PROGRESS

AGENDA

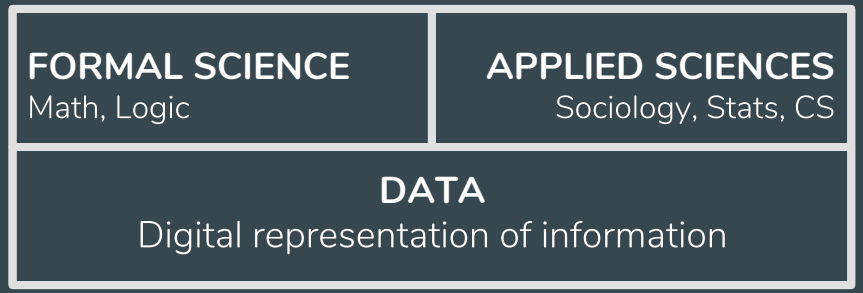
1. WHAT IS DATA SCIENCE
2. SKILLS OF DATA SCIENTISTS
3. EVOLUTION & TRENDS

WHAT IS DATA SCIENCE | SKILLS | EVOLUTION AND TRENDS

A. ...as understood through **FORMAL DEFINITIONS**
B. ...as understood through **INDUSTRY LEADERS**
C. ...as misunderstood through **MYTHS**
D. ...as understood through **DATA PRODUCTS**

WHAT IS DATA SCIENCE | SKILLS | EVOLUTION AND TRENDS

definitions - industry leaders - myths - products



DATA SCIENCE is an **interdisciplinary** and **applied** science, applying scientific knowledge from the **formal sciences** with that from the **social or natural sciences** onto the **data**, the digital representation of information.
The goals in this sub-discipline include providing actionable intelligence in the form of testable explanations, predictions, interactive intelligence, and intelligent machines.

WHAT IS DATA SCIENCE | SKILLS | EVOLUTION AND TRENDS

definitions - industry leaders - myths - products

Machine Learning is the process of 'learning' underlying patterns in a set of observations that are represented through data. These patterns are extracted using algorithms and can then be represented in a mathematical model used to predict the outcomes of new observations.

If we want to predict weight from height, we would need a formula we could run when we got the weight of new people, such as the common y = mx + b. In this case, y = weight (because it is what we are predicting) and x = height (because it is what we know). Our current observations include Joe, height = 66" and weight = 130 lbs, another is Maria, height = 65" and weight = 125, etc. A regression algorithm would identify what values 'm' and 'b' should be set to in our formula. Let's say m = 2 and b = 10. Our new mathematical model will be able to predict future observations using y = 2x + 10.
The goal of the algorithm is to pick parameters (m & b in this case) that minimize the error (how far off we are) of our predictions of height from weight.

WHAT IS DATA SCIENCE | SKILLS | EVOLUTION AND TRENDS

definitions - **industry leaders** - myths - products

"...a hybrid skill set that combines analytical, statistical, development and engineering skills that enable a team to provide value, insights, and direction to people."
Ann-Jinette Hess – Data Scientist/Manager @ Rackspace

"...take huge complicated databases, decipher business needs and come back with intelligence that quantifies spending, profits, and trends."
Carla Gentry – Data Scientist @ Analytical-Solution

"...equal parts hacker, stats geek, and entrepreneur."
Chris Chapo – Data Scientist/VP @ Gap

"...detecting patterns that can then be used to help people make better decisions"
Alice Zhen – Data Scientist/Manager @ Amazon

"...being able to leverage all the data we have to inform our decision making in order to make better Products."
Nathan Anderson – Product Engineering Director @ ClearDATA

"As a reader, data science is an easy way to reduce people to numbers, but a possible if difficult way to augment people's understanding of each other."
Jason Gignac

"As a policy advocate (political hack), data science helps me mobilize support for a policy position."
Ana Unruh Cohen

"As an Innovation Director, to me data science is critical for us to understand if we should pursue an experiment or not."
Marina Alderete Gavito

WHAT IS DATA SCIENCE | SKILLS | EVOLUTION AND TRENDS

definitions - industry leaders - **myths** - products

Myth #1: Data Science == Statistics

Statistics forces us to make assumptions about the nature of the relationship between variables, distribution of the data, hypothesis, e.g.

Statistics is used in data science, but it is only a small part of it. Machine learning turns this process around. Given a large trove of data, the computer taunts us by saying "If only you knew what question to ask me, I would give you some very interesting answers based on the data!"[1]

Myth #2: Data Science ==Business Analyst

Structured data, incoming requests => dashboards with KPI's or regular reporting. LIght on the 'decision science' and heavy on the KPI reporting. Often a data or business analyst role will progress into a data science role.

Myth #3: Data Science == Data Science

Do not assume a common understanding to every hiring manager, recruiter and applicant

Myth #4: Data science curriculum is well defined and consistent.

Myth #5: If I want to be a data scientist, I just need to learn how to use R or Python

Just as knowing how to use autoCAD does not make me an architect...those are tools that enable someone to do much of the work required in data science.

WHAT IS DATA SCIENCE | SKILLS | EVOLUTION AND TRENDS

definitions - industry leaders - myths - **products**

NOTES for use on this topic (WIP):

Insert images of example products, or links to products, such as netflix, amazon, facebook, google

Data driven: creating a data culture – pg 5 UPS example: if drivers took only right turns, it would see a large improvement in fuel savings and safety, while reducing wasted time…shaved 20.4 million miles off routes in one year. (https://money.cnn.com/video/news/2010/12/13/n_cs_ups_no_left_turn.fortune/)

"the best minds of my generation are thinking about how to make people click ads." – Jeff Hammerbacher, early facebook employee

"work on stuff that matters." – Tim O'Reilly

Collaborative filtering or recommendation engines, Apple genius, apps you may like, people you may know, news (danger of that)

Try google's incognito searching compared to searching as you usually do. Differences?

From *Women in Data:* Pg 21 Laurie Skelly: researching the neural processes behind empathy and psychopathy, connecting the dots between behavior and emotions using ML algorithms"

From *Building Data Science Teams:* "People who viewed this item bought…", "People You May Know", Netflix

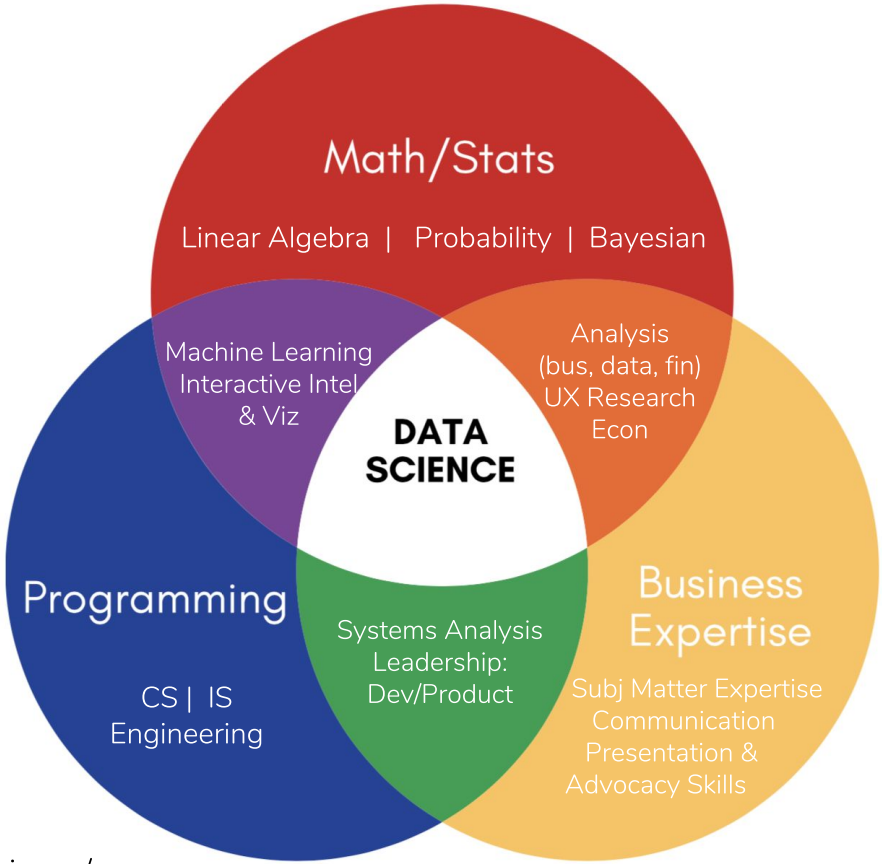
target marketing/personalized content

Products that drive a company's value proposition

Products that facilitate the introduction into other products (linkedin groups)

Products that prevent dead ends (amazon)

Products that stand alone



Check out our recent blog post:
https://codeup.com/what-is-data-science/

MODERN DATA SCIENTIST

Data Scientist, the sexiest job the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MarketingDistillery.com

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Advocacy | Strategic | Technical expertise | Curiosity | Storytelling | Cleverness | Diverse | Connectedness | Detailed & Focused

WHEN INTERVIEWING:

1. Would we be willing to do a start up with you? We can be locked in a room with you for long periods of time; we can trust you; we can communicate with each other quickly and efficiently
2. Can you “knock the socks off” of the company in 90 days
3. In 4-6 years, will you be doing something amazing?
4. Skills and curiosity to ask big questions
5. Diverse backgrounds who have histories of playing with data to create something novel
6. Incredibly bright and creative people right out of college and put them through a very robust internship program.

D.J. Patil in *Building Data Science Teams*
http://www.datascienceassn.org/sites/default/files/Building%20Data%20Science%20Teams.pdf

Advocacy | Strategic | Technical expertise | Curiosity | Storytelling | Cleverness | Diverse | Connectedness | Detailed & Focused

WAYS TO SHARPEN SKILLS:

1. Finding rich data sources
2. Working with large volumes of data despite hardware, software and bandwidth constraints
3. Cleaning the data and making sure that data is consistent
4. Melding multiple data sets together
5. Visualizing the data
6. Building the rich tooling that enables others to work with data effectively

D.J. Patil in *Building Data Science Teams*

D.J. Patil in *Building Data Science Teams*

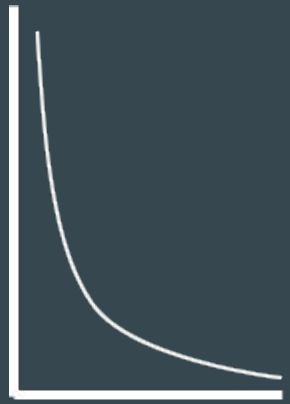
Advocacy – “data scientists have an influence out of proportion to their numbers.”
Strategic - ability to recognize which data needs to be collected
Technical expertise - deep expertise in some scientific discipline
Curiosity – “desire to go beneath the surface and discover and distill a problem down into a very clear set of hypotheses that can be tested.”
Storytelling - “ability to use data to tell a story and to be able to communicate it effectively.”
Cleverness - “ability to look at a problem in different, creative ways”
Diverse - “My best data scientists have come from very different backgrounds”
Connectedness - ability bring disparate areas together in a novel way…”I’ve seen data scientists apply novel DNA sequencing techniques to find patterns of fraud”
Detailed & Focused - “Good data scientists understand, in a deep way, that the heavy lifting of cleanup and preparation isn’t something that gets in the way of solving the problem: it *is* the problem.”

WHAT IS DATA SCIENCE | SKILLS | **EVOLUTION & TRENDS**

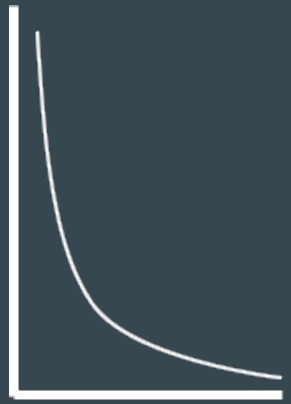
yesterday - today - tomorrow

Capability has been driven by the decreasing costs of storage, CPU, and bandwidth and increase in amount of data generated

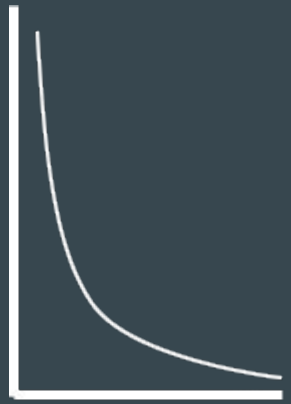
"**Demand** has been driven by the impact of an organization using data effectively." *Data Driven: Creating a Data Culture*



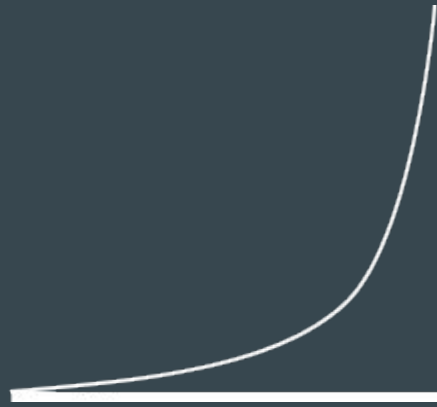
STORAGE COST



CPU COST



BANDWIDTH COST



DATA GENERATED

WHAT IS DATA SCIENCE | SKILLS | **EVOLUTION & TRENDS**

yesterday - **today** - tomorrow

- Educational Opportunities
- Buzzword
- Options now exist for consumers to tools abstracted from dependencies on deep technical skills, but this space needs more refining still to make it ‘commonplace’
- Listed in companies’ ‘top 5 priorities’
- Subdisciplines of data science are emerging: Machine Learning Engineer, Data Visualization Developer, Data Journalist, Big Data Engineer, etc.
- “Google needs to move beyond the current search format of you entering a query and getting 10 results. The ideal would be us knowing what you want before you search for it…” *The Evolution of Data Products*: Eric Schmidt June 25, 2011
    - Google search are giving ”human time”…<score of spurs game if I google spurs while a game is going on>

WHAT IS DATA SCIENCE | SKILLS | **EVOLUTION & TRENDS**

yesterday - today - **tomorrow**

- Educational opportunities at the bachelor's level
- Ethics in data science, especially around "Black Box" Models
- More products for reducing the time spent in wrangling the data
- Automation around machine learning (development, deployment, management)
- More presence in Cyber Security
- Continue expansion into all industries
- More presence in Healthcare
- Field will follow similar path Web Development has taken
- As with all developments in technology...used for bad as well as good. (https://www.wired.com/story/machine-learning-backdoors/)

ASSIGNMENT #2: THE DATA SCIENCE COMMUNITY

GOAL: Explore and integrate yourselves into the data science community so that you have that to use for networking, learning, sharing once this course is over and you are off doing amazing things.

In each of the following, you will share what is asked via google classroom, where I have posted specific questions for each item. Your sources must be unique to what has already been posted for that question (e.g. blog, leaders, podcasts)

NETWORKS AND PORTFOLIOS	Create an account (if you do not already have one), share your username/handle with the class , and connect or follow everyone in this class in each of the following apps: • Twitter | LinkedIn | Kaggle.com | Data.world | Github.com | public.tableau.com
BLOGS	Find one interesting blog related to data science • Subscribe to it, find a post that is interesting to you from it, summarize in a sentence or two something you learned from it
DISCIPLINE LEADERS	Find 3 leaders in the field of data science • Follow them, share with the class so they can follow them, also, and share one useful post in their feed
PODCASTS	Find one podcast that discusses data science topics • Subscribe to it, and share the name so others can also subscribe
USING R	Find & share one online resource related to R that you can reference often when working in R

For next week

1. Data Science Mindset, Methodologies and Misconceptions: Intro through Chapter 4
2. Weapons of Math Destruction: Get book and begin reading
3. Exploring Data Science Assignment #1: Example in everyday life

   Using the app you are assigned, write up what the 'data product' is, why it is useful for you as the end user, and what data you think it is using to produce output. Explore it, experiment with it, and try to manipulate the output. Then explain how you experimented with it and what resulted from it. What did you learn from your experiments related to the way the model behind the scenes might work?

Enrique	Netflix
Kit	Google Search
Maria	App Store, Google Play, SIRI, Alexa, Google Voice, or gmail
Paola	Hulu
Adrienne	Pinterest
Josephina	LinkedIn

Brad	Twitter
Evan	YouTube
Tam	Instagram
Samantha	Spotify
Nishant	Google Maps
Spencer	Facebook

EXAMPLES OF DS | PIPELINE | METHODOLOGIES

A. WHAT IS YOUR DATA PRODUCT?
B. WHY IS IT USEFUL TO THE CONSUMER?
C. WHAT DATA DO YOU THINK IT IS USING TO PRODUCE OUTPUT?
D. HOW DID YOU EXPERIMENT WITH IT?
E. WHAT WAS THE RESULT OF YOUR EXPERIMENT?
F. WHAT DID YOU LEARN RELATED TO HOW THE MODEL BEHIND THE SCENES MIGHT WORK?

Enrique	Netflix
Kit	Google Search
Spencer	Facebook
Paola	Hulu
Adrienne	Pinterest
Antonio	LinkedIn

Brad	Twitter
Evan	YouTube
Tam	Instagram
Samantha	Spotify
Nishant	Google Maps
Maria	App S

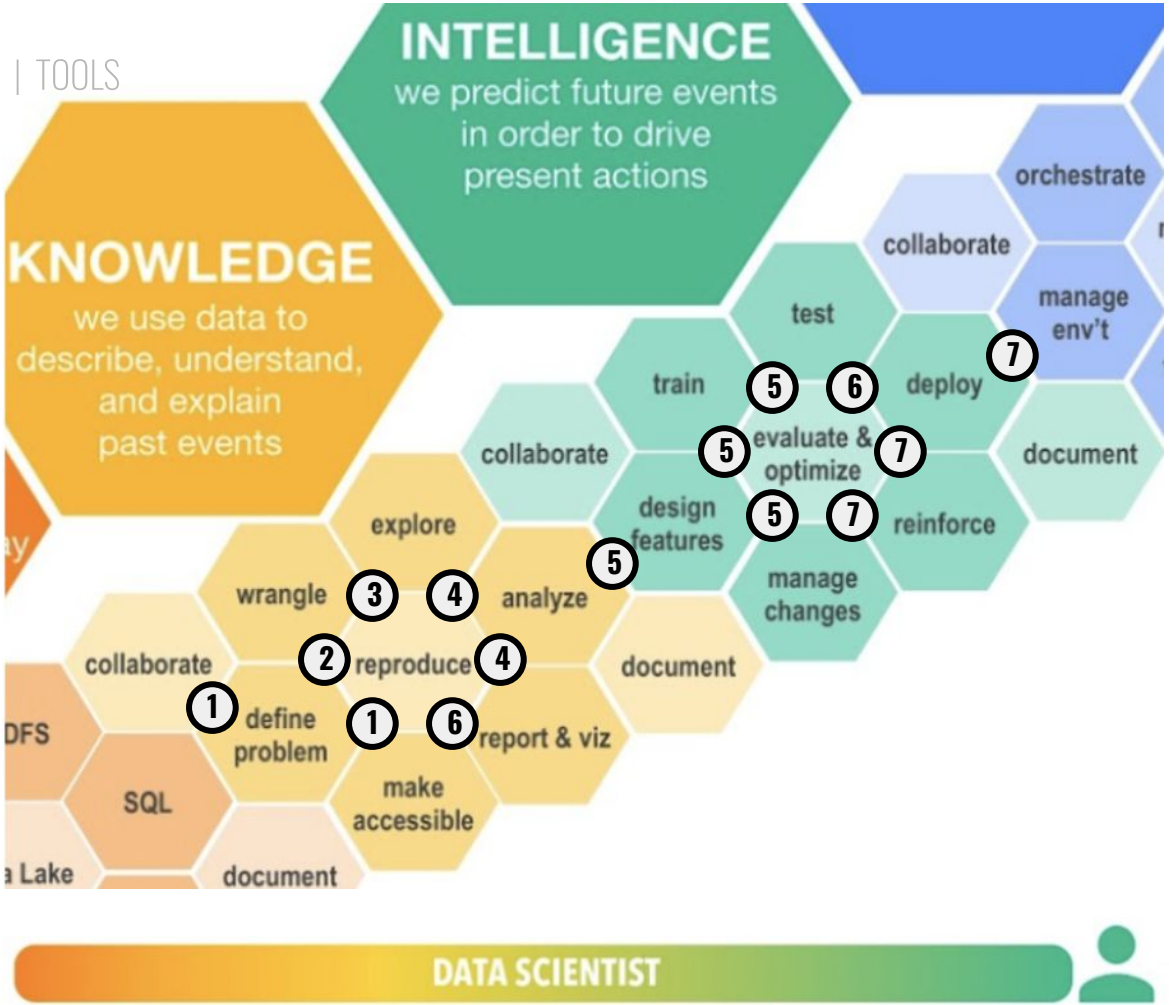
AGENDA

1. THE DATA SCIENCE PIPELINE
2. TOOLS
3. METHODOLOGIES

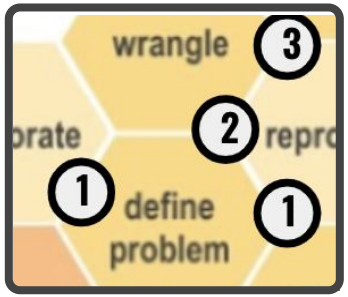
PIPELINE | TOOLS | METHODOLOGIES

PIPELINE | METHODOLOGIES | TOOLS

1. Planning
2. Acquisition
3. Preparation
4. Exploration
5. Modeling
6. Delivery
7. Maintenance



PIPELINE: planning



1. **Planning**
2. Acquisition
3. Preparation
4. Exploration
5. Modeling
6. Delivery
7. Maintenance

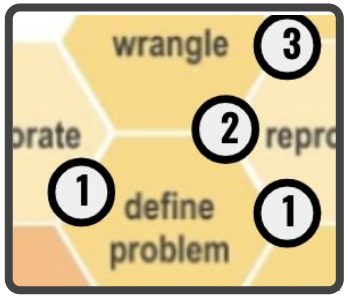
Ask questions. Ask the 'right' questions: Ask your own questions!

1. Business stakeholders and end users ask either a) more general questions that are very hard to answer directly or b) extremely specific questions that are not going to achieve the ultimate goal they hope the project will achieve. Both of these could lead to miscommunication, time spent on work that will not end up being useful, or inadequate understanding of the underlying problem being investigated.
2. As you work with the various data streams at your disposal, you gain a better understanding of problems and can ask more informative & specialized

Hypothesizing:

1. Is feature X1 related to feature X2?
2. How similar are features X1 and X2?
3. Is subset A of target variable Y significantly different from subset B in their measures in feature X1?
4. Do features X1 and X2 collaborate well with each other for predicting target variable Y?
5. Should we remove X1 from the feature set?
6. Does feature X1 cause feature X2 or target variable Y?
7. Code to acquire is documented, cleaned and reproducible
8. Access to desired data is available from a single environment.

PIPELINE: planning for answering questions and testing hypotheses



1. **Planning**
2. Acquisition
3. Preparation
4. Exploration
5. Modeling
6. Delivery
7. Maintenance

1. Is feature X1 related to feature X2? e.g. Is a person's age related to their work experience?

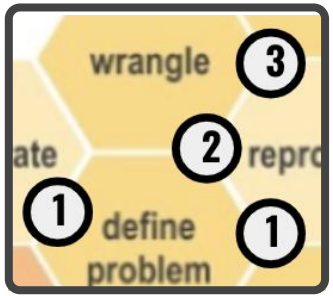
- H0 (Null Hypothesis): Similarity between values of X and Y is 0, e.g. there is no measurable relationship between age and work experience.
- Ha (Alternative Hypothesis): There is measurable similarity between X and Y, e.g. age is related to work experience.
- If X and Y are continuous variables, normalize the data first.

2. How similar are variables X and Y? Similarity is generally measured through correlation statistics, we will usually want to use metrics that don't rely on the distribution of the data (such as Pearson's R) in order to measure similarity. Similarity of 2 features helps to reduce the feature set by omitting one of them or through merging them into a single feature. A t-test (for continuous variables) or a chi-squared test (for discrete variables) can often be used to test.

3. Is subset A significantly different from subset B? A and B are subsets of the target variable, usually, and the goal is to compare the groups' measures in feature X to each other to see if they are statistically different. Outliers have been removed or changed to adapt to the variables distribution. This can be used to test if the observations with target class Y1 (subset A) has significantly different measures of variable X than the observations of target class Y2 (subset B). If there is difference, then variable X is a good choice to keep as a feature.

- H0: the difference between subset A and subset B is insubstantial (basically 0)
- Ha: the difference between subset A and subset B is substantial

PIPELINE: data acquisition



DELIVERABLES:

1. Code to acquire is documented, cleaned and reproducible
2. Access to desired data is available from a single environment.

GOALS: Gather data from its sources in order to prepare and clean it in the next step. Essentially, you are creating a plath from original sources to the environment in which you will work with the data.

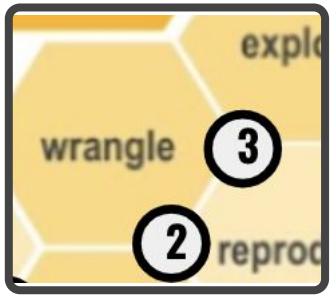
TOOLS: SQL | R | Spark | Python | Hive | Command line

1 Planning
2 Acquisition
3 Preparation
4 Exploration
5 Modeling
6 Delivery
7 Maintenance

SOURCE TYPES: RDBMS | NoSQL | HDFS | cloud files (S3, google drive) | static local flat files (csv, xlsx, txt) | Command line

A.K.A: Data Gathering | Data Import | Data Wrangling (Acquisition + Prep)

PIPELINE: data preparation



DELIVERABLES:

1. Code to prepare data is documented, cleaned and reproducible
2. Single data set that has integrated all the original datasets into one, where each row is an observation and each column is an attribute/single variable, missing values have been addressed and erroneous outliers have been resolved.
3. Two samples from the dataset above, the training set making up about 60% of the total and the test set being the remaining 40%.

GOALS: GHave a training dataset in a format that can be easily explored, analyzed and visualized. The test set should be in exactly the same state, but set aside for the final evaluation at the end of modeling.

TOOLS: R | Spark | Python

A.K.A: Data Tidying | Data Cleansing | Data Wrangling (Acquisition + Prep)

1. Planning
2. Acquisition
3. **Preparation**
4. Exploration
5. Modeling
6. Delivery
7. Maintenance

PIPELINE: data exploration

DELIVERABLES:

1. Reproducible analysis demonstrating findings though verbiage, visualizations, and tables.
2. Clean script fo the transformation the training dataset went through in this step, in order to apply the same transformations to the test dataset.
3. Refined training dataset with no missing values and ready to be modeled.
4. Documented code so someone else could pick it up and understand what is going on without you there to explain.

GOALS:

1. Signals in the data are understood, i.e. variables that help drive the outcome are identified and we are able to narrow in on those variables
2. Remove those variables that are noisy, provide no valuable information, are redundant, or provide no new information (i.e. no information gain)
3. Acquire knowledge that will help guide construction of new features

1 Planning
2 Acquisition
3 Preparation
4 **Exploration**
5 Modeling
6 Delivery
7 Maintenance

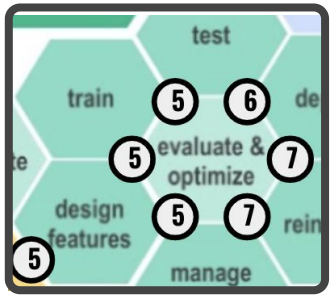
TOOLS: R | Spark | Python | Tableau | Plot.ly

METHODOLOGIES:

1. Statistics: descriptive, inferential
2. Correlation Analysis
3. Association Rules
4. Data Visualizations

A.K.A: Exploratory Analysis | Analytics | Exploratory Visualization

PIPELINE: data modeling



DELIVERABLES:

1. Program that includes the code for data acquisition, data transformations (prep, feature engineering, pre-processing), predictions that can be run on new data in the same form as the sample data was in at step 0.
2. Actual predictions made on test data using the newly trained model.
3. Evaluation metrics from predictions made on test data

GOALS:

1. **Feature Engineering:** Engineer features to enable an optimal model of the underlying structure of the observations that are represented by data.
2. **Train a model:** Abstract from the observations the patterns that drive the outcomes observed.
3. **Test the model:** Effectively scale the model performance from the training data to the test, unseen dataset.

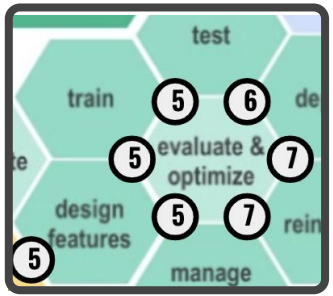
1 Planning
2 Acquisition
3 Preparation
4 Exploration
5 Modeling
6 Delivery
7 Maintenance

TOOLS: R | Spark | Python | Tableau | Plot.ly

METHODOLOGIES:

1. Statistics: descriptive, inferential
2. Correlation Analysis
3. Association Rules
4. Data Visualizations

PIPELINE: data modeling - feature engineering



DELIVERABLES:

1. A dataset with features that you know drive the behavior of the outcome your are targeting.
2. A program with code to reproduce those features in a new dataset.
3. Documentation

GOAL: "What's the best representation of the sample data to learn a solution to your problem?"

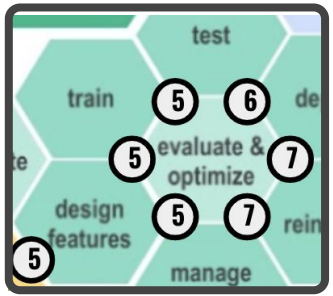
TOOLS: R | Spark | Python

METHODOLOGIES:

Brainstorm: Manually identify features that are of greater value. Think about the problem and assess qualitatively.
Feature Extraction: Automated construction of new, synthetic features from raw data, solving for high dimensionality. Common methods include principal components analysis (PCA), clustering.
Feature Construction: Spend time with the data thinking about the underlying form of the structures and how to best expose them to the algorithm. Features could be constructed through aggregating or combining attributes, or decomposing or splitting attributes.
Feature importance: Scoring features based on their importance prior to the modeling algorithm using association rules, correlation coefficients, or as part of the modeling algorithms of MARS, Random Forest, Gradient Boosted Machines, as examples.
Feature Selection: Filter out any remaining redundant variables, select those with most value, highest correlation. This by creating subsets of feature groups and evaluate performance. Stepwise regression is an algorithm that incorporates selection as part of it's model.
Feature Transformation: Dummy variables, normalizing, standardizing, center & scale, data types

1 Planning
2 Acquisition
3 Preparation
4 Exploration
5 Modeling
6 Delivery
7 Maintenance

PIPELINE: data modeling - train



DELIVERABLES:

1. A fit model with evaluation metrics

GOALS: Create a robust and generalizable model that is a mapping between features and a target based on discoveries and transformations made in the previous steps.

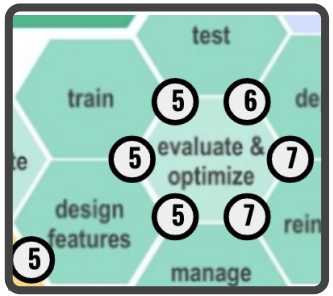
1. Test and identify hyper-parameters
2. Sample the data
3. Train and fit algorithms
4. Select evaluation metric and evaluate performance
5. Select optimal model (algorithm + hyperparameters + features)

TOOLS: R | Spark | Python

1 Planning
2 Acquisition
3 Preparation
4 Exploration
5 **Modeling**
6 Delivery
7 Maintenance

METHODOLOGIES:
Cross Validation Methods: Bagging, Boosting
Predictive Modeling: Classification, Regression, Time Series Analysis, Text Predictions, Anomaly Detection
Recommender Systems: Content-Based Filtering, Collaborative Filtering, Non-Negative Matrix Factorization
Graph Analysis: Modeling relationships among networks, e.g. social networks.
NLP: Analysis of text to derive insights related to its content. Use cases include topic modeling, sentiment analysis, text summarization

PIPELINE: data modeling - test



DELIVERABLES:
1. Predictions on new data (where applicable)
2. Evaluation metrics
3. Code to deploy model to be run on new data

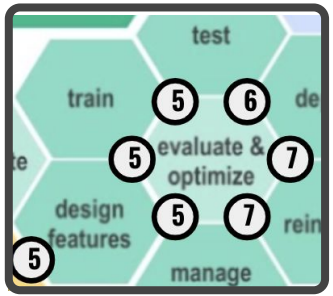
GOALS:
1. Preprocess new data using training parameters
2. Predict using the trained model on test/unseen dataset
3. Evaluate results and look out for overfitting.

TOOLS: R | Spark | Python

METHODOLOGIES:
Python ‘Pipeline’
Spark ‘Transform’ and ‘Pipeline’

1 Planning
2 Acquisition
3 Preparation
4 Exploration
5 Modeling
6 Delivery
7 Maintenance

PIPELINE: data modeling



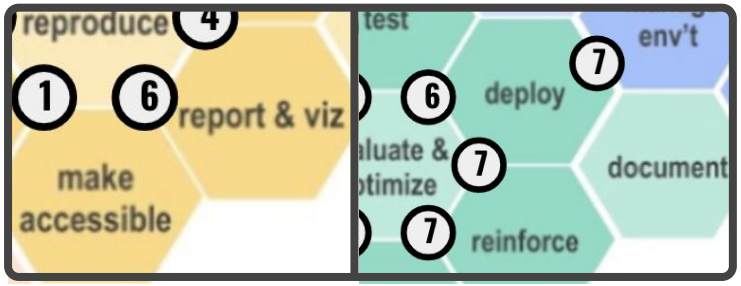
1 Planning
2 Acquisition
3 Preparation
4 Exploration
5 Modeling
6 Delivery
7 Maintenance

Let's start with data and what is a feature. Tabular data is described in terms of observations or instances (rows) that are made up of variables or attributes (columns). An attribute could be a feature. The idea of a feature, separate from an attribute, makes more sense in the context of a problem. A feature is an attribute that is useful or meaningful to your problem. It is an important part of an observation for learning about the structure of the problem that is being modeled.

Machine learning algorithms learn a solution to a problem from sample data. In this context, feature engineering asks: what is the best representation of the sample data to learn a solution to your problem? It's deep. Doing well in machine learning, even in artificial intelligence in general comes back to representation problems. It's hard stuff, perhaps unknowable (or at best intractable) to know the best representation to use, *a priori*.

"you have to turn your inputs into things the algorithm can understand" — Shayne Miel, answer to "What is the intuitive explanation of feature engineering in machine learning?"

PIPELINE: delivery of data product



GOAL: Enable a machine learning model to be consumed through an interface by an end user. This is about deploying a model to production, or productionizing the model, so that it is live and consumable in the desired manner. This is usually done through an API. The web application will take data input by the user, and the model will take that data, extract/compute the features, feed them to the model which will generate results. These results will then be fed back to the web application for it to respond accordingly, such as display a particular product for the end user.

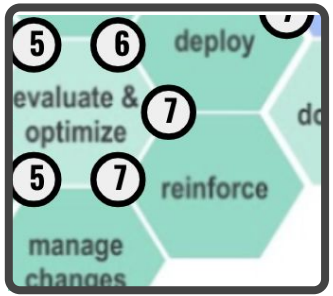
1. Planning
2. Acquisition
3. Preparation
4. Exploration
5. Modeling
6. **Delivery**
7. Maintenance

DELIVERABLES:

- **Clear Communication and documentation** on how to take action on your deliverable
- **Report:** Summarize findings in words and visualizations and make recommendations of next steps
- **Dashboard/interactive visualizations** for future exploration and insights by others
- **Predictive model** to be run at a regular cadence 'deployed to production'
- **Predictions** from a predictive model that is run ad-hoc
- **Sentiment Modeling/Topic Models/Text summarization models** deployed to be run across new documents

TOOLS: Jupyter, R-Markdown, ShinyR, Tableau, AWS Sagemaker, Spark, Python

PIPELINE: maintenance



1. Planning
2. Acquisition
3. Preparation
4. Exploration
5. Modeling
6. Delivery
7. **Maintenance**

BIG DATA & AI LANDSCAPE 2018

TOOLS

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

DATABASE PLATFORMS

1. SQL (Structured Query Language)
   - Relational Database Management System (RDBMS)
   - Structured data only
   - Primary data source for business intelligence.
   - Original use case is for managing data for software applications
   - Examples: Microsoft SQL Server, MySQL (open-source), Oracle PostgresQL
   - https://en.wikipedia.org/wiki/SQL_Server_Management_Studio
2. NoSQL (Not-Only SQL)
   - Accommodates both structured and unstructured data
   - Used for big data and data science
   - Benefits: elastic scaling & flexibility, low cost, many can be integrated with hadoop ecosystem
   - Examples: Cassandra, HBase, Hive, MongoDB
3. Graph-based
   - Focus on creating, storing, querying and processing graphs
   - Data science use cases include social network analysis and security threat detection
   - Examples: Neo4j, Graphbase

https://www.tutorialspoint.com/database_tutorials.htm

PROGRAMMING LANGUAGES

1. Julia
    a. Functional programming language, developed at MIT
    b. Pros: fast, easy for prototyping, can live in production in Julia
    c. Cons: still early in its development
2. Python
    a. Object-oriented programming language
    b. Top data science libraries: Numpy, SciPy, Matplotlib, Pandas, Statsmodels
    c. Pros: easy to learn and debug, large number of libraries for data science and growing, plenty of resources for learning and using
    d. Cons: slower in performance speed
3. R
    a. Open-source version of the proprietary statistical software ‘S’
    b. Pros: optimal for statistics, intuitive, easy to learn, great IDE (RStudio), great plotting libraries
    c. Cons: slow in performance speed, not good when dealing with large amounts of data
4. Scala
    a. Functional programming language, stemming from Java
    b. Pros: seamless integration with Apache Spark, used when dealing with distributed, big data
    c. Cons: not easy to learn

PIPELINE | **TOOLS** | METHODOLOGIES

DATA ANALYTICS SOFTWARE: MATLAB (open-source version: Octave), **Analytica, Mathematica**

VISUALIZATION SOFTWARE

Plot.ly: work with plots in different platforms, easy to use, great quality
D3.js: javascript library, fast, supports a variety of datasets, and allows for animations and interactive plots. (d3js.org)
Tableau: great for BI and data science visualizations, especially for creating dashboards for business users.

DATA GOVERNANCE: Storing, managing & processing data in a distributed environment

Spark: Pros are integration with R and Python, generality (used for different data science applications), versatility (run on different environments), performance speed, ease of use, growing community of users. Through Spark, you can use SQL-like language for querying (Spark SQL), do machine learning analytics (MLlib), graph analytics (GraphX), and stream analysis (Spark Streaming)
Hadoop: original big data platform, comprised of an ecosystem included HDFS (hadoop distributed file system), MapReduce, HBase (NoSQL DB), Hive (query platform)
Storm: used for streaming data...data coming in at high speeds or velocity.

VERSION CONTROL SYSTEMS (VCS): Important when working in a team that handles the same scripts and project files in general. Good for tracking changes in a program and reverting to previous versions.

Git: open source, cross platform, command line interface
Github: git with an available GUI, most commonly used

PIPELINE | TOOLS | METHODOLOGIES

STATISTICS
FEATURE ENGINEERING
- PCA
- regularization

PREDICTIVE ANALYTICS
- Classification
- Regression
- Time series analysis
- Text prediction

CLUSTERING
- K-means
- DBSCAN
- hierarchical

RECOMMENDER SYSTEMS
- Content based filtering
- Collaborative filtering

NLP
- Topic modeling
- Sentiment analysis
- Text summarization

CROSS VALIDATION
- Bootstrapping
- Bagging
- Boosting

PIPELINE | TOOLS | METHODOLOGIES

Algorithm cheat sheet

As a data scientist

Projects you might work on: types, examples
Teams
Industries
Business Units
A ‘typical’ day
Challenges