1rst) Applied Data Science - Overview

by Kumaran Ponnambalam F on Skillshare

This course focuses on the overview of Data Science. It explains how Data Science works from data elements through relationships and predictions. It then walks through the stages of a Data Science Project

9 Lessons (1h 44m)

URL: https://www.skillshare.com/classes/Applied-Data-Science-1-Overview/259054106

1. About Applied Data Science Series 8:12

Course goal

- Train students to be full-fledged data science **practitioners** who could execute **end-to-end** data science projects to achieve **business results**
- This course is focused on practice (i.e. applied data science)
  - Technologies
  - Processes
  - Tools and Techniques
  - Focus on purpose, usage, advantages with adequate understanding of concepts
  - Available tools and libraries

Achievements

- Understand the concepts and life cycle of Data Science
- Develop proficiency to use R for all stages of analytics
- Learn Data Engineering tools and techniques
- Acquire knowledge of different machine learning techniques and know when and how to use them.
- Become a full-fledged Data Science Practitioner who can immediately contribute to real-life Data Science projects

Three technical areas

- Math and Statistics
- Machine Learning foundations
- Programming

2. What is Data Science - One 11:51

Data Science - understanding the domain

What is Data Science? (definitions, aims)

- Skill of extracting of knowledge from data
- Using knowledge to predict the unknown
- Improve business outcomes with the power of data
- Employ techniques and theories drawn from broad areas of mathematics, statistics and information technology

What is a Data Scientist? (practioner)

- A practitioner of data science
- Expertise in data engineering, analytics, statistics and business domain
- Investigate complex business problems and use data to provide solutions

What are the elements of Data Science? (#6)

1. Entity

- A thing that exists about which we research and predict in data science.
- Entity has a business context.
- Customer of a business
- Patient at a hospital. The same person can be a patient and a customer, but the business context is different.
- Car. Entities can be non living things

2. Characteristics (= properties = attributes)

- Every entity has a set of characteristics. These are unique properties
- Properties too have a business context
- Customer : Age, income group, gender, education
- Patient: Age, Blood Pressure, Weight, Family history.
- Car: Make, Model, Year, Engine, VIN

3. Environments

- Environment points to the eco-system in which the entity exists or functions.
- Environment is shared among entities. Multiple entities belong to the same environment
- Environment affects an entity's behavior
- Customer : Country, City, Work Place
- Patient: City, Climate .
- Car: Use (City/highway), Climate

4. Events

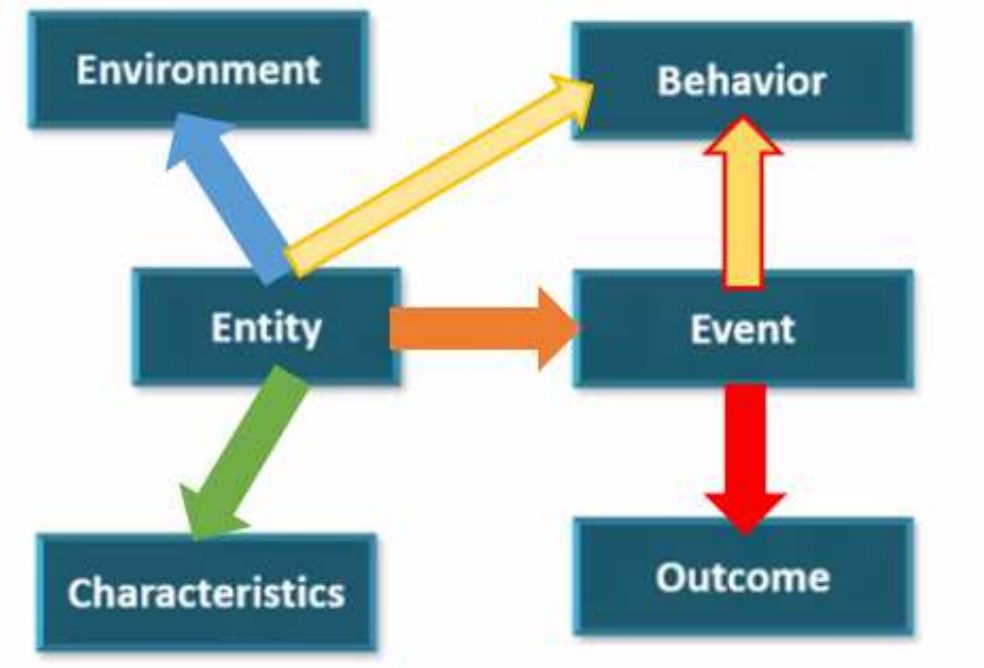
- A significant business activity in which an entity participates.
- Events happen in a said environment.
- Customer : Browsing, store visit, sales call
- Patient: Doctor visit, blood test
- Car: Smog test, comparison test

5. Behavior

- What an entity does during an event.
- Entities may have different behaviors in different environments
- Customer : Phone Call vs email, Clickstream, response to offers
- Patient: Nausea, light-headed, cramps
- Car: Skid, acceleration, stopping distances

6. Outcome

- The result of an activity deemed significant by the business.
- Outcome values can be
  - Boolean (Yes/No, Pass/Fail)
  - Continuous (a numeric value)
  - Class (identification of type)
- Customer : Sale (Boolean), sale value (continuous)
- Patient: Blood Pressure value (continuous). Diabetes type (class)
- Car: Smog levels (class), stopping distances (continuous), smog passed (Boolean), car type (class)



3. What is Data Science - Two 10:44

What is Data? (introduction)

Observation

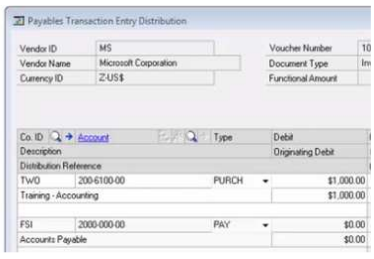
- A measurement of an event deemed significant by the business.
- Captures information about
  - Entities involved
  - Characteristics of the entities
  - Behavior
  - Environment in which the behavior happens
  - outcomes
- An observation is also called a system of record
- Customer : A phone call record, a buying transaction, an email offer
- Patient: A doctor visit record, a test result, a data capture from a monitoring device
- Car: Service record, smog test result

Dataset

- A collection of observations
- Each observation is typically called a record
- Each record has a set of attributes that point to characteristics, behavior or outcomes.
- A dataset can be
  - Structured (database records, spreadsheet)
  - Unstructured (twitter feeds, news paper articles)
  - Semi-structured (email)
- Data scientists collect and work on datasets to learn about entities and predict their future behavior/ outcomes.

Structured Data

- Attributes are labeled and distinctly visible.
- Easily searchable and query able.
- Stored easily in tables



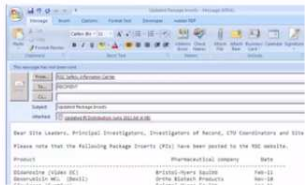
Unstructured Data

- Data is continuous text
- Attributes are not distinctly labeled. They are present within the data.
- Querying is not easy.

The Mazda3 is on a very short list of compact cars that are available as a hatchback or a sedan. It also comes with two 6-speed transmissions -- manual or automatic -- and choice of two 4-cylinder engines -- a 155-horsepower 2.0-liter or a 184-horsepower 2.5-liter -- and all of those variations are available with either body style. Its best fuel economy is an EPA-rated 41 mpg on the highway, which is

Semi-structured Data

- Mix of structured and unstructured.
- Some attributes are distinctly labeled. Others are hidden within free text



Summary : the #8 fundamental keys of Data Science

- Entity
- Characteristics
- Environment
- Event
- Behavior
- Outcomes
- Observation
- Dataset

4. What is Data Science - Three 12:55

What is Learning? i.e. discovering knowledge from data

Relationships

- Attributes in a dataset exhibit relationships
- Relationships "model" the real world and have a logical "explanation"
- For attributes A and B the relationships can be
  - When A occurs, B also occurs
  - When A occurs B does not occur
  - When A increases B also increases
  - When A increases B decreases
- Relationships can involve multiple attributes too
  - When A is present and B increases, C will decrease

Examples

- Customer
  - As age goes up, spending capacity goes up. (AGE and REVENUE)
  - Urban customers buy more internet bandwidth (LOCATION and BANDWIDTH)
- Patient
  - Older patients have more prevalence of Diabetes (AGE and DISEASE LEVEL)
  - Overweight patients typically have higher cholesterol levels (WEIGHT and HDL)
- Car
  - The more cylinders a car has, the mileage tends to be lower (CYLINDERS and MILEAGE)
  - Sports Cars have more insurance rates (TYPE and RATES)

Types of relationship (#3)

- Consistent vs Incidental Patterns in Data
- Correlations
- Signals and noise

Learning is …

- Learning implies learning about relationships.
- It involves
  - Taking a domain
  - Understanding the attributes that represent the domain
  - Collecting data
  - Understanding relationships between the attributes
- Model is the outcome of learning

Model

- A simplified, approximated representation of a real world phenomenon
- Captures key attributes and their relationships
- Mathematical model – represents relationships as an equation
- Blood Pressure

  BP = 56 + (AGE * .8) + (WEIGHT * .14) + (LDL * .009)
- Decision Tree model – represents the outcome as a decision tree
- Buying a music CD

  If AGE < 25 and GENDER=MALE, buy BEYONCE-CD = YES
- Accuracy of models depends on strength of relationships between attributes

5. What is Data Science - Four 9:31

Prediction

- A model can be used to predict unknown attributes

  BP = 56 + (AGE * .8) + (WEIGHT * .14) + (LDL * .009)
- The above model represents the relationships between BP, AGE, WEIGHT and LDL.
- If 3 of the 4 attributes are known, the model can be used to predict the 4th.
- The above equation can be considered the prediction algorithm
- Relationships can be a lot more complex, leading to complex models and prediction algorithms.

Predictors and outcomes

- Outcomes are attributes that you want to predict
- Predictors are attributes that are used to predict outcomes.
- Learning is all about building models that can be used to predict outcomes (outputs) using the predictors (inputs)

Example	Predictors	Outcomes
Customer	Age, Income Range, Location	Buy? Yes/No
Patient	Age, Blood Pressure, Weight	Diabetic?
Car	Cylinders, acceleration	Sports vs family

Humans vs. Machines

- Humans understand relationships and predict all the time.
- Build humans can only handle finite amount of data
  - One shop keeper can know preferences of 100 customers, not 10 million of them
- Machines (computers) come into play when the number of entities and data about them are large
- There in comes machine learning, predictive analytics and data science

So, Data Science consists in 6 major steps :

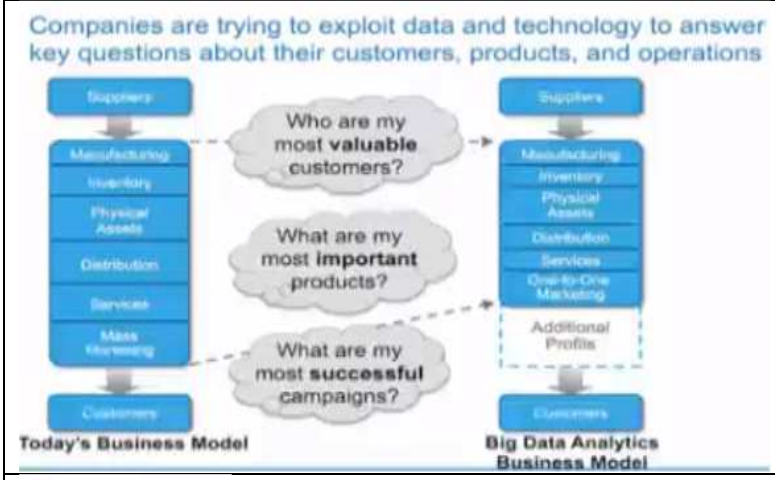
- Picking a problem in a specified domain
- Understanding the problem domain (entities and attributes)
- Collect datasets that represent the entities
- Discover relationships (Learning)
  - When computers are used for this purpose, its called machine learning.
- Build models that represent relationships
  - Uses past data where all predictors and outcomes are known
- Use models for predicting outcomes
  - Current/ future data – predictors known, outcomes unknown

Ex. Website shopper

- Problem : Predict if the shopper will buy a smartphone
- Data: Past purchase history of shoppers
  - Shopper characteristics (age, gender, income etc.)
  - Seasonal information
  - Others..
- Build Model
  - Decision model based on shopper and seasonal entities
  - Built every week
- Prediction
  - When a new shopper is browsing, predict if the shopper will buy
- Action
  - Offer Chat help

6. Data Science Use Cases 7:47

Data Science Use Cases (benefits in various Business domains)



Finance

Making money and saving money => **Fraud detection** (fraud score is calculated for each transaction based on models. High score is suspected to be a fraud…)

Retailing

Sell more => **Recommendations** **(when an item is bought by a customer, the system offers related items with a high affinity score ... Affinity score is built based on patterns)**

Contact centers

Improving efficiency => **Scoring of callers and agents** **(Caller/Agent score is built based on ability to solve specific type of problem or to sell… the system will improve the matching. Call recordings are also analyzed by ML algorithm to grade quality and outcome)**

Health Care

Preventive Care => **Predicting disease outbreaks** **(Dataset and location are collected from public domains. Model uses to track potential outbreak and take preventive action)**

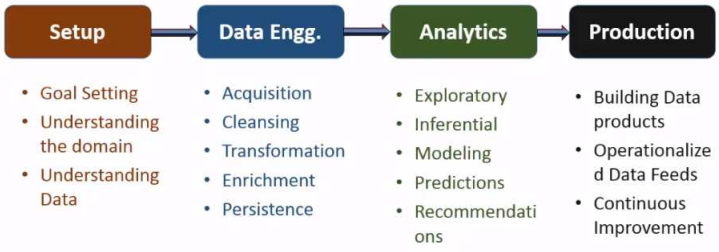
7. Data Science Life Cycle - Setup 11:46

Data Science Life Cycle (activities and sequencing)

Introduction

- Data Science efforts are typically executed as projects
- They are typically research projects, not build-and-operate projects
- Projects typically have a start and end state
- Projects have phases and activities
- Transitions happen between phases / activities

Data Science Project Phases and list of activities (#4 phases)



Setup

Setup -> 1. Goal setting

- Every Data Science Project will / should have a goal.
- The team will focus their activities based on this goal
- Projects without goals are cars without a driver.

Setup –> 2. Understandng Problem Domain

- A data science team should have solid understanding of the problem domain
  - Business basics (finance, CRM, medical)
  - Business processes and workflows
  - Key Performance metrics
- Machines only know numbers and strings, they needs humans to associate meaning to them.
- Knowledge of the domain helps the team to understand entities, relationships and patterns.
- It helps validate assumptions, identify errors and analyze if predictions will work.

Setup –> 3. Understandng Data

- Business Processes and workflows generate Data
  - Application Data Entries
  - Reports and Visualizations
  - Sensor Data feeds
  - Web clicks in a browser
  - Point-of-Sale Transactions
  - Social media feeds
- Data can be structured, unstructured or semi-structured
- Data have different origins, stored in different silos and have logical relationships

- Source of the Data
- Processing /transformation steps performed
- Storage (enterprise databases, cloud, news feeds)
- Synchronization
- Relationships
- Ordering
- Understanding data helps the team identify possible sources of predictive patterns.

8. Data Science Life Cycle - Data Engg 11:57

Data Engi neering (get the data to the form you need it)

Data Engg.

1. Data Acquisition

- Acquire Data from different data sources
  - Enterprise Databases (Oracle, MySQL)
  - Cloud APIs (Sales Force)
  - Scanner / sensor feeds (Bar code scanners)
  - Social media downloads (Twitter, Facebook)
- Real time/ interval and bulk
- Sanity checking
- Most cumbersome and time-consuming to setup.
- Establishing connections to machines and humans involved can be frustrating ☹

2. Data Cleansing

- Data have different degrees on cleanliness and completeness
- Structured data from corporate applications are usually clean and complete
- Data from internet, social media or voice transcripts need significant cleansing.
- Handling missing data is a key decision

3. Data Transformation

- Data is transformed to extract required information while discarding un-necessary baggage
- Processing and summarization to logical activity levels
- Transformation helps cutting down data size and minimizes further processing needs

4. Data Enrichment

- Add additional attributes to data records that improves the quality of information

5. Data Persistence

- Processed data is stored in a reliable, retrievable data sink.
- All relevant information captured in a single local record as much as possible

- Scaling and query performance are important factors will choosing a data sink
  - Flat files
  - Traditional SQL databases
  - Big Data technologies.

9. Data Science Life Cycle - Analysis ... 19:16

Analytics (learn and predict)

Analytics

1. Exploratory Data Analysis

- Understand individual attribute patterns (range, minimum, maximum, frequency, mean etc.)
- Understand relationships between attributes (how does change in one affect another)
- Reasoning (is the behavior explainable?)
- Outliers (odd values)
- Possible errors in processing
- Validate findings with domain experts.

2. Inferential Analysis

- Look for signals in the data
  - Patterns
  - Correlations
  - Reasoning
- Check if patterns are consistent and reproducible
  - Month after month
  - Different use cases
- Statistical Tests
  - Can results be extrapolated for the entire population?

3. Modeling

- Use machine learning algorithms to build models
- Build multiple models based on different algorithms and different datasets
- Test models for accuracy
- Identify best performing models
  - Accuracy
  - Response Time
  - Resources
- As simple as an equation or a decision tree. As complex as a neural network.

4. Predicting

- Use models built to predict outcomes for new data
- Keep validating model accuracy to make sure accuracy levels are consistent for different variations in data
- Response time and resource usage are critical when predictions need to happen in real time
- Measure improvements made to outcome predictions using the model
- Simulations might be performed to validate prediction benefits

5. Recommendations

- At the end of the project, recommendations need to be provided to the project owners on the algorithms to use and expected benefits
- A Data science project might have no recommendations to make if the dataset does not exhibit any exploitable patterns
  - Does not mean it's a failure
- Sometimes unexpected patterns are discovered that might lead to other benefits
- A final presentation is made to stake holders.

6. Iterations

- Based on intermediate or at-the-end analysis and feedback, the analysis phase might be repeated with different objectives
- The project team "responds" to findings in the data, which might lead to multiple analysis paths.

Production (Implement Continuous Process)

Production

1. Builing Data Product

- Once the modeling and prediction algorithms are "firmed up", data products are built that would use the algorithms for production level modeling and predictions
- Have quality software rigor in development and testing
- Deployed in enterprise or cloud models.

2. Operationalized Data Feeds

- Continuous data feeds into data products
  - Instantaneous
  - Every day
  - Periodic
- Data products perform cleansing, transformation and error reporting
- Pruning of old data might be necessary

3. Continous Improvement

- Changes in business environment might affect learning and prediction
- The learning and prediction steps need to be re-validated at appropriate intervals to make sure they continue to work as desired.
- Revalidation needs to happen when business processes change.
- Efforts to generate better models should be ongoing.

Summary

- Data Science projects follow a life cycle
- Data Science projects are research type projects – there is a lot of experimentation and sometimes no end result
- Signals in data drives results, not the algorithms
- Multiple iterations might be necessary before reasonable results are achieved.