PRODUCED BY:

DR. BRENDA MULLALLY BMULLALLY@WIT.IE

DEPARTMENT COMPUTING MATHS AND PHYSICS

WATERFORD INSTITUTE OF TECHNOLOGY

WWW.WIT.IE

MOODLE.WIT.IE

BUSINESS ANALYTICS

DATA MINING MODEL

- THE DM TECHNIQUES USED FOR THE STUDY/PROJECT CAN ORIGINATE FROM THE FIELDS OF STATISTICS OR COMPUTER SCIENCE.
- THE FIRST STEP IN ORDER TO DEFINE WHICH DM TECHNIQUE TO USE FOR THE APPLICATION IS TO IDENTIFY THE GOALS OF THE STUDY, DESCRIBING WHAT HAS HAPPENED IN THE PAST **OR** PREDICTING WHAT WILL HAPPEN IN THE FUTURE:

DATA MINING MODEL

- DESCRIBING WHAT HAPPENED:
  - CHARACTERISATION - IS A SUMMARISATION OF GENERAL FEATURES OF OBJECTS IN A TARGET CLASS (E.G. CHARACTERISE THE OURVIDEOSTORE CUSTOMERS WHO REGULARLY RENT MORE THAN 30 MOVIES A YEAR)
  - PATTERNS/ASSOCIATIONS/CORRELATIONS
    - FREQUENT ITEMSET, (E.G. SET OF ITEMS THAT APPEAR TOGETHER – MILK AND BREAD)
    - FREQUENT SUBSEQUENCES (SEQUENTIAL PATTERN – CUSTOMERS BUY A LAPTOP AND THEN A DIGITAL CAMERA)
  - CLUSTERING ANALYSIS – IDENTIFY NATURAL GROUPS BASED ON THEIR KNOWN CHARACTERISTICS
    - OUTLIER DETECTION – DATASET MAY CONTAIN OBJECTS THAT DON'T COMPLY

- PREDICTING THE FUTURE:
  - CLASSIFICATION AND REGRESSION

THE DATA MINING MODEL: DESCRIBING WHAT HAPPENED

- **DESCRIBE** WHAT HAPPENED
- *DESCRIPTIVE TECHNIQUES* ARE USED TO LOOK FOR PATTERNS
- DESCRIPTIVE TECHNIQUES CAN BE OF TWO TYPES:
  - *ASSOCIATION – ESTABLISHES RELATIONSHIPS ABOUT ITEMS THAT OCCUR TOGETHER IN A GIVEN RECORD*
  - *CLUSTERING – PARTITION DATA INTO SEGMENTS IN WHICH THE MEMBERS OF DATA SEGMENT SHARE SIMILAR QUALITIES*

ASSOCIATION/CLUSTERING TECHNIQUES

Association Techniques

Goal	Input Variables (Predictor)	Output Variables (Outcome)	Statistical Technique	Examples
Find large groups of cases in large data files that are similar on a small set of input characteristics,	Continuous or Discrete	No outcome variable	K-means Cluster Analysis	• Customer segments for marketing • Groups of similar insurance claims
To create large cluster memberships			Kohonen Neural Networks	• Cluster customers into segments based on demographics and buying patterns
Create small set associations and look for patterns between many categories	Logical	No outcome variable	Market Basket or Association Analysis with Apriori	• Identify which products are likely to be purchased together • Identify which courses students are likely to take together
Create small set associations and look for patterns between many categories	Logical or numeric	No outcome variable	Market Basket or Association Analysis with GRI	• Identify which courses students are likely to take together
To create linkages between sets of items to display complex relationships	Continuous or Discrete	No outcome variable	Link Analysis	• To identify a relationship between a network of physicians and their prescriptions

CLUSTER ANALYSIS FOR DATA MINING

- USED FOR AUTOMATIC IDENTIFICATION OF NATURAL GROUPINGS OF THINGS (E.G. CUSTOMERS)
- PART OF THE MACHINE-LEARNING FAMILY
- EMPLOY UNSUPERVISED LEARNING
- LEARNS THE CLUSTERS OF THINGS FROM PAST DATA
- THERE IS NO OUTPUT VARIABLE
- ALSO KNOWN AS SEGMENTATION
- MOST COMMON CLUSTERING ALGORITHM – K-MEANS

ASSOCIATION RULE MINING

- A VERY POPULAR DM METHOD IN BUSINESS
- FINDS INTERESTING RELATIONSHIPS (AFFINITIES) BETWEEN VARIABLES (ITEMS OR EVENTS)
- PART OF MACHINE LEARNING FAMILY
- EMPLOYS UNSUPERVISED LEARNING
- THERE IS NO OUTPUT VARIABLE
- ALSO KNOWN AS MARKET BASKET ANALYSIS
- OFTEN USED AS AN EXAMPLE TO DESCRIBE DM TO ORDINARY PEOPLE, SUCH AS THE FAMOUS "RELATIONSHIP BETWEEN DIAPERS AND BEERS!"
- MOST COMMON ALGORITHM - APRIORI

ASSOCIATION RULE MINING

- ARE ALL ASSOCIATION RULES INTERESTING AND USEFUL?

  A GENERIC RULE: **X** ⇒ **Y** [**S%, C%**]

  X, Y: PRODUCTS AND/OR SERVICES

  X: LEFT-HAND-SIDE (LHS)

  Y: RIGHT-HAND-SIDE (RHS)

  S: SUPPORT: HOW OFTEN **X** AND **Y** GO TOGETHER

  C: CONFIDENCE: HOW OFTEN **Y** GO TOGETHER WITH THE **X**

  EXAMPLE: {LAPTOP COMPUTER, ANTIVIRUS SOFTWARE} ⇒ {EXTENDED SERVICE PLAN} [30%, 70%]

ASSOCIATION RULE MINING

- INPUT: THE SIMPLE POINT-OF-SALE TRANSACTION DATA
- OUTPUT: MOST FREQUENT RELATIONSHIPS AMONG ITEMS
- EXAMPLE: ACCORDING TO THE TRANSACTION DATA…

  "CUSTOMER WHO BOUGHT A LAPTOP COMPUTER AND A VIRUS PROTECTION SOFTWARE, ALSO BOUGHT EXTENDED SERVICE PLAN 30 PERCENT OF THE TIME" "70% OF THE TRANSACTIONS FOR THE SALE OF EXTENDED SERVICE PLAN ALSO PURCHASED A LAPTOP AND VIRUS PROTECTION"
- HOW DO YOU USE SUCH A PATTERN/KNOWLEDGE?
  - PUT THE ITEMS NEXT TO EACH OTHER FOR EASE OF FINDING
  - PROMOTE THE ITEMS AS A PACKAGE (DO NOT PUT ONE ON SALE IF THE OTHER(S) ARE ON SALE)
  - PLACE ITEMS FAR APART FROM EACH OTHER SO THAT THE CUSTOMER HAS TO WALK THE AISLES TO SEARCH FOR IT, AND BY DOING SO POTENTIALLY SEE AND BUY OTHER ITEMS

ASSOCIATION RULE MINING

- REPRESENTATIVE APPLICATIONS OF ASSOCIATION RULE MINING INCLUDE
  - IN BUSINESS: CROSS-MARKETING, CROSS-SELLING, STORE DESIGN, CATALOG DESIGN, E-COMMERCE SITE DESIGN, OPTIMIZATION OF ONLINE ADVERTISING, PRODUCT PRICING, AND SALES/PROMOTION CONFIGURATION
  - IN MEDICINE: RELATIONSHIPS BETWEEN SYMPTOMS AND ILLNESSES; DIAGNOSIS AND PATIENT CHARACTERISTICS AND TREATMENTS (TO BE USED IN MEDICAL DSS); AND GENES AND THEIR FUNCTIONS (TO BE USED IN GENOMICS PROJECTS)

ARE ALL PATTERNS INTERESTING?

- WHAT MAKES A PATTERN INTERESTING?
- CAN A SYSTEM GENERATE ONLY THE INTERESTING ONES?

  - IS IT EASILY UNDERSTOOD BY HUMANS?
  - IS IT VALID ON NEW OR TEST DATA WITH SOME DEGREE OF CERTAINTY?
  - IS IT POTENTIALLY USEFUL?
  - IS IT NOVEL?

  - IT IS ALSO INTERESTING IF A PATTERN VALIDATES A HYPOTHESIS THAT THE USER SOUGHT TO CONFIRM.
  - AN INTERESTING PATTERN REPRESENTS KNOWLEDGE.

DATA MINING MODEL: PREDICTING WHAT WILL HAPPEN

- TO PREDICT WHAT WILL HAPPEN MEANS TO DEVELOP A MODEL THAT USES HISTORICAL DATA TO PREDICT AN OUTCOME BASED ON A SET OF INPUT CHARACTERISTICS.
- PREDICTIVE TECHNIQUES REQUIRE THE USE OF PAST HISTORY WITH THE INTENT TO PREDICT FUTURE BEHAVIOUR.
- DM TECHNIQUES IN THIS AREA SERVE TO CLASSIFY THE OUTCOME VARIABLE INTO PREDEFINED CATEGORIES.

DATA MODEL: PREDICTING WHAT WILL HAPPEN

- OBJECTIVE OF STATISTICAL DM TECHNIQUES IS TO FIND HOW TWO OR MORE VARIABLES ARE RELATED TO EACH OTHER.
  - **PREDICTION**/FORECASTING ESTIMATES FUTURE VALUES BASED ON PATTERNS WITHIN LARGE SETS OF DATA, THIS PREDICTION CAN BE LABELED FOR DETERMINING WEATHER FORECAST AS 'SUNNY' OR 'RAINY' . USE INPUT TO PRODUCE SOME CLASSIFICATION OF OUTPUT, E.G.
    - A PATTERN HAS BEEN FOUND ALREADY IN A SET OF DATA.
    - GIVEN A NEW SET OF DATA, YOU CAN PREDICT WHICH OF THESE CLASSES IT BELONG TOO.
- **REGRESSION** IS A WELL-KNOWN STATISTICAL TECHNIQUE THAT IS USED TO MAP DATA TO A PREDICTION VALUE E.G. A REAL NUMBER 65°F
  - HOW ACCURATE AM I WITH THIS? USE CLASSIFICATION MODEL TO GIVE A ACTUAL MEASURE WITH HOW CLOSE YOU ARE TO THE TARGET – 95% CORRECT

DATA MINING MODEL - PREDICTION

- CLASSIFICATION

  SUPERVISED INDUCTION USED TO ANALYZE THE HISTORICAL DATA STORED IN A DATABASE AND TO GENERATE A MODEL THAT CAN PREDICT FUTURE BEHAVIOR

- PART OF THE MACHINE-LEARNING FAMILY
- EMPLOY SUPERVISED LEARNING
- LEARN FROM PAST DATA, CLASSIFY NEW DATA
- THE OUTPUT VARIABLE IS CATEGORICAL (NOMINAL OR ORDINAL) IN NATURE
- MOST COMMON ALGORITHM/TECHNIQUE: DECISION TREES

DATA MINING CONCEPTS AND APPLICATIONS

- **SEQUENCE DISCOVERY**

  THE IDENTIFICATION OF ASSOCIATIONS OVER TIME

- **VISUALIZATION** CAN BE USED IN CONJUNCTION WITH DATA MINING TO GAIN A CLEARER UNDERSTANDING OF MANY UNDERLYING RELATIONSHIPS

DATA MINING CONCEPTS AND APPLICATIONS

- **HYPOTHESIS-DRIVEN DATA MINING**

  BEGINS WITH A PROPOSITION BY THE USER, WHO THEN SEEKS TO VALIDATE THE TRUTHFULNESS OF THE PROPOSITION

- **DISCOVERY-DRIVEN DATA MINING**

  FINDS PATTERNS, ASSOCIATIONS, AND RELATIONSHIPS AMONG THE DATA IN ORDER TO UNCOVER FACTS THAT WERE PREVIOUSLY UNKNOWN OR NOT EVEN CONTEMPLATED BY AN ORGANIZATION

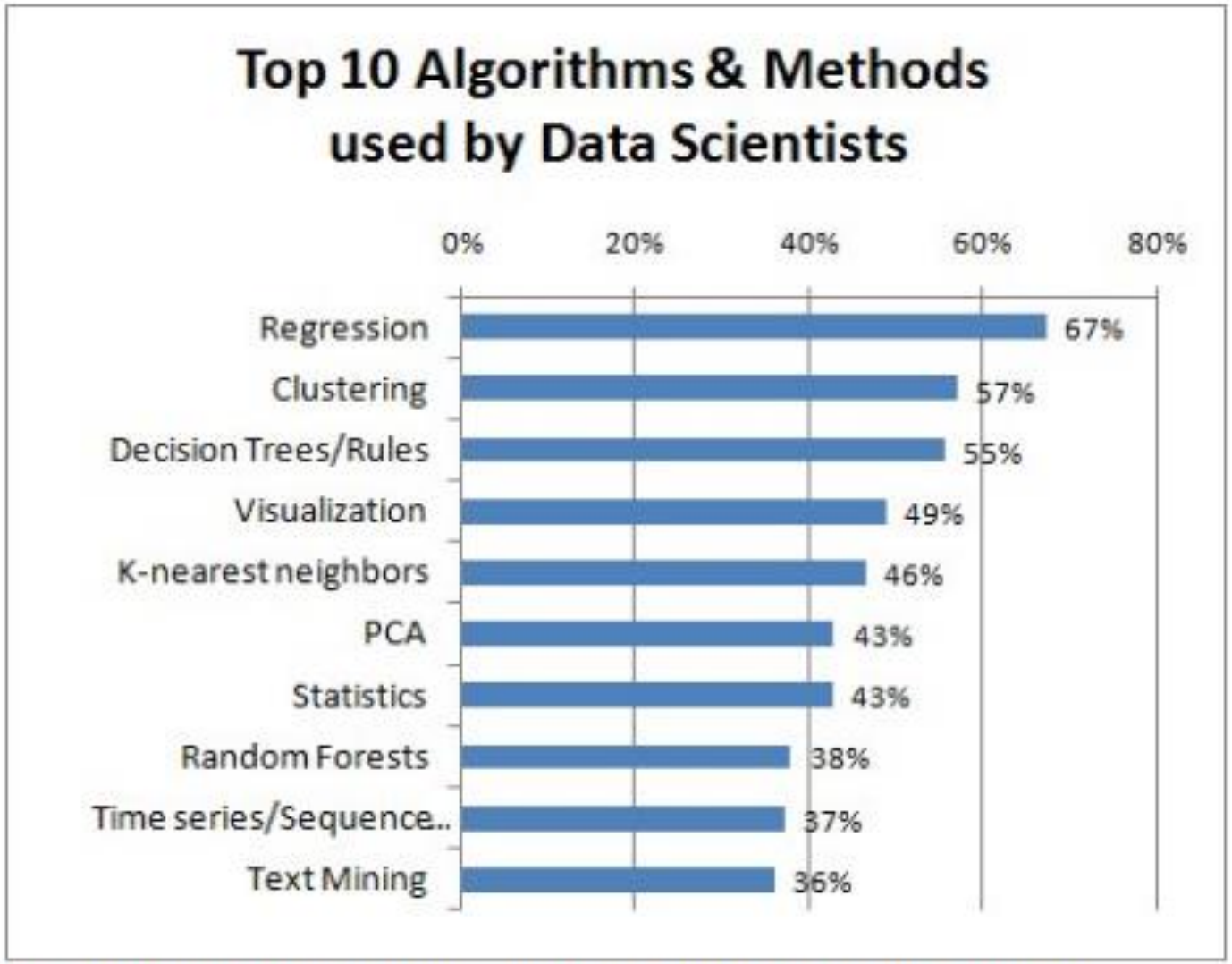

Fig. 1: Top 10 algorithms & methods used by Data Scientists.
See full table of all algorithms and methods at the end of the post.

- WWW.KDNUGGEST.COM

TOOLS

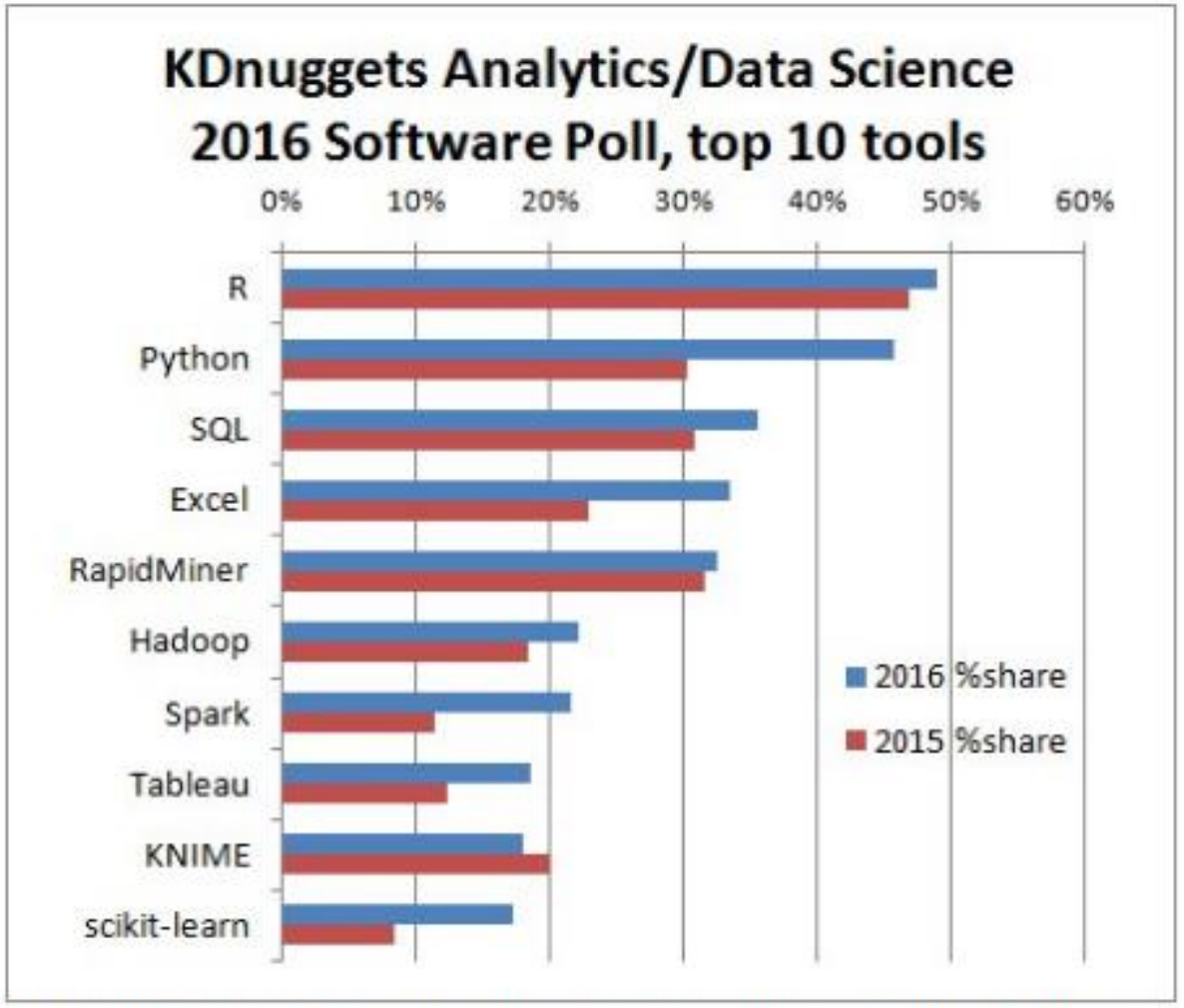

Fig 1: KDnuggets Analytics/Data Science 2016 Software Poll: top 10 most popular tools in 2016

www.kdnuggets.com

3 year comparison

CAREERCAST.COM

- TOP 10 CAREERS 2016
  - DATA SCIENTIST MADE IT TO THE LIST IN 2015 AND NOW TOPS NUMBER 1 FOR 2016.
  - STATISTICIAN IS SECOND.
  - MATHEMATICIAN IS SIXTH
  - ACTUARY IS TENTH.
- BOOMING MARKET FOR THOSE THAT DEAL WITH NUMBERS.

- U.S STUDY CONDUCTED EACH YEAR USING BUREAU OF LABOUR STATISTICS(BLS) USING ENVIRONMENTAL (WORK WEEK, EMOTIONAL, PHYSICAL), INCOME (START, MIDLEVEL, TOP) , OUTLOOK (GROWTH, UNEMPLOYMENT) AND STRESS (TRAVEL, DEADLINES, COMPETITIVENESS ETC) FACTORS.

DATA MINING MYTHS

- DATA MINING …
  - PROVIDES INSTANT SOLUTIONS/PREDICTIONS.
  - IS NOT YET VIABLE FOR BUSINESS APPLICATIONS.
  - REQUIRES A SEPARATE, DEDICATED DATABASE.
  - CAN ONLY BE DONE BY THOSE WITH ADVANCED DEGREES.
  - IS ONLY FOR LARGE FIRMS THAT HAVE LOTS OF CUSTOMER DATA.
  - IS ANOTHER NAME FOR GOOD-OLD STATISTICS.

DATA MINING?

"NOT EVERYTHING THAT CAN BE COUNTED COUNTS, AND NOT EVERYTHING THAT COUNTS CAN BE COUNTED" WILLIAM BRUCE CAMERON 1963

"PREDICTION IS VERY DIFFICULT, ESPECIALLY ABOUT THE FUTURE" NEIL BOHR 1918

"IF WE HAVE DATA, LET'S LOOK AT DATA. IF ALL WE HAVE ARE OPINIONS, LET'S GO WITH MINE." – JIM BARKSDALE, FORMER CEO OF NETSCAPE COMMUNICATIONS CORPORATION.

"TORTURE THE DATA, AND IT WILL CONFESS TO ANYTHING." – RONALD COASE, ECONOMICS, NOBEL PRIZE LAUREATE

- TED TALKS WHAT DO WE DO WITH ALL THIS DATA?

- TED TALKS HOW TO USE DATA TO MAKE A HIT TV SHOW