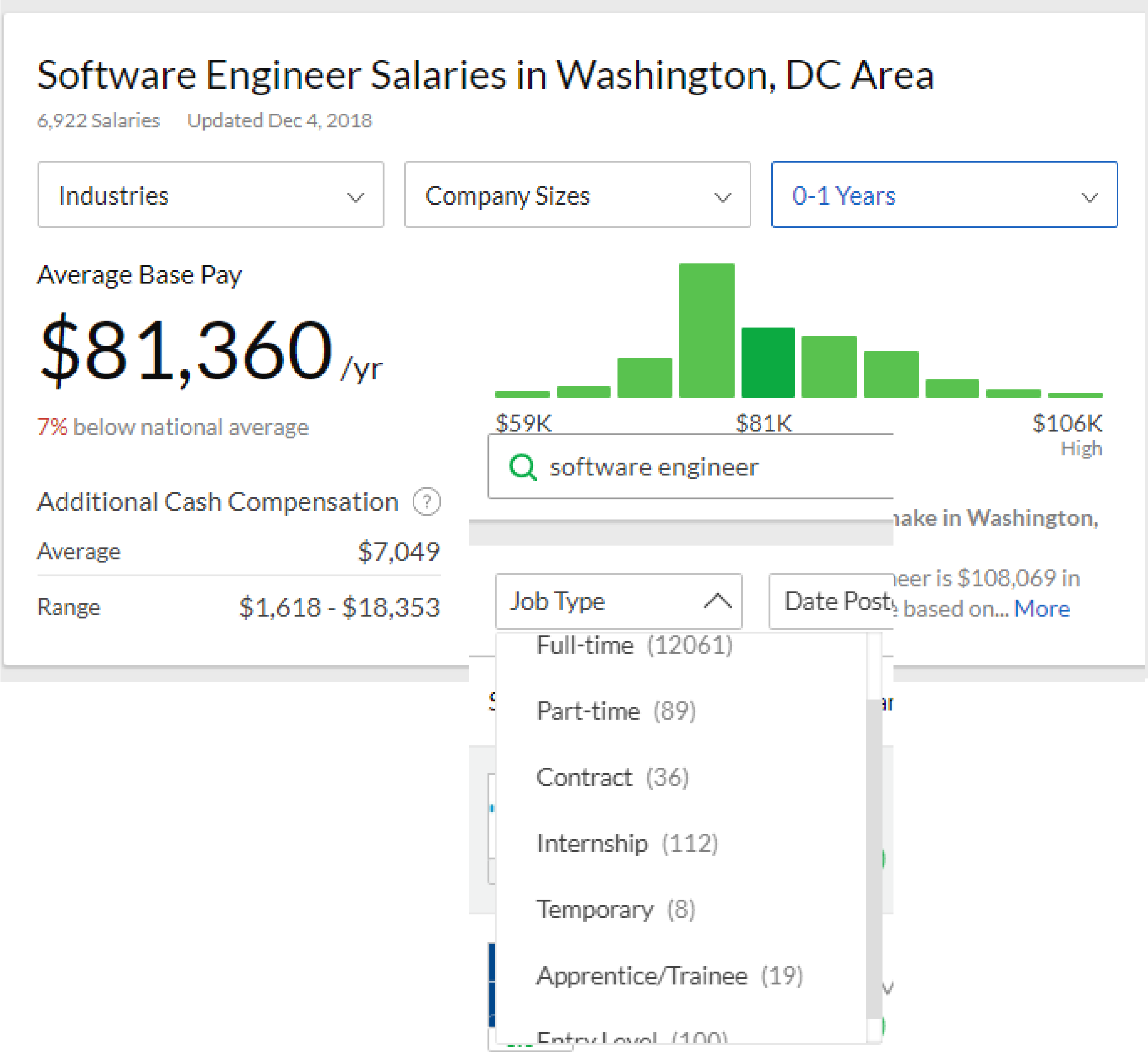
Data Science in Industry



Software Engineer Salaries in Washington, DC Area

6,922 Salaries   Updated Dec 4, 2018

Industries | Company Sizes | 0-1 Years

Average Base Pay

$81,360 /yr

7% below national average

software engineer

ıake in Washington,

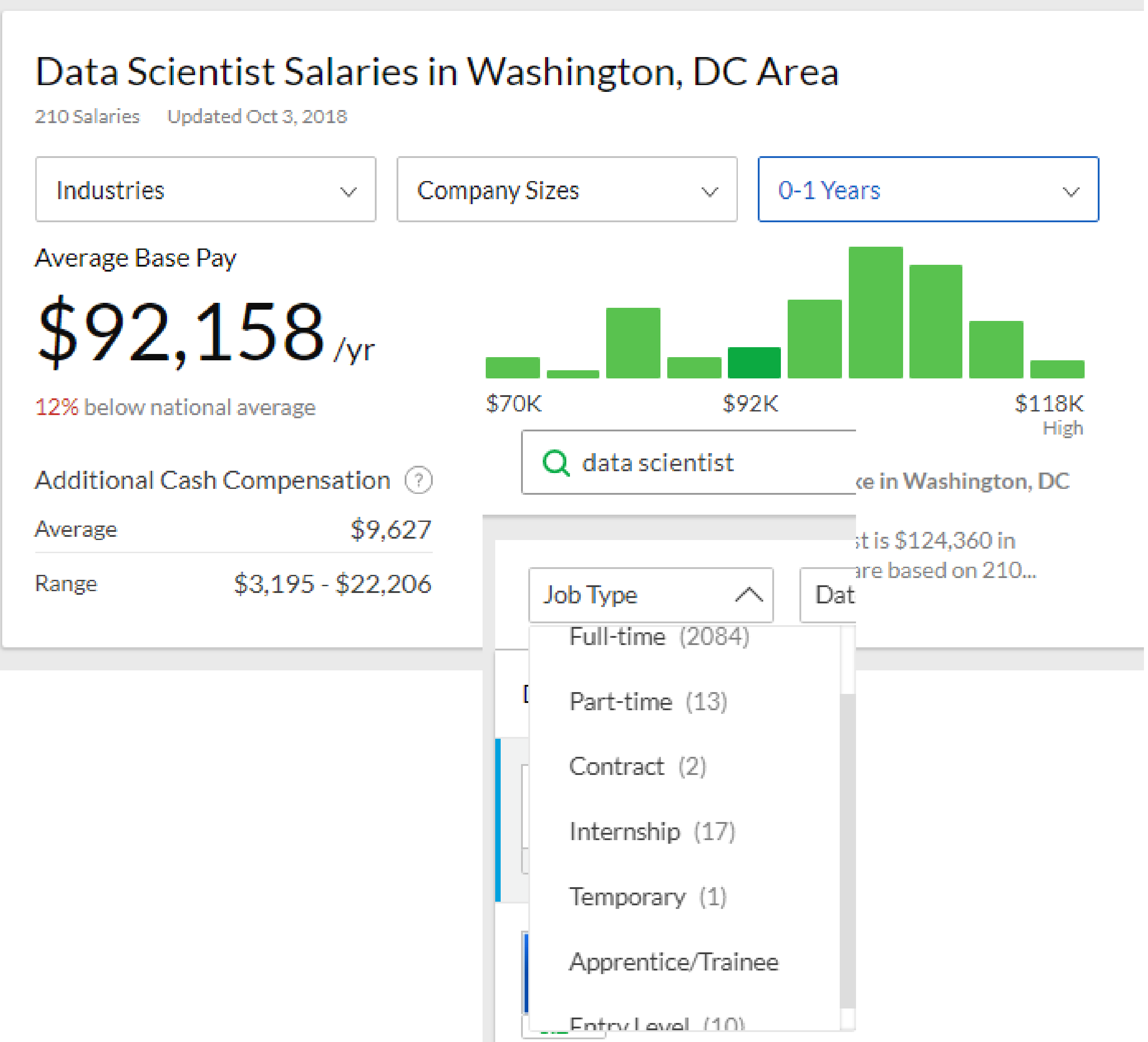
ıeer is $108,069 in
: based on... More

Additional Cash Compensation

Average	$7,049
Range	$1,618 - $18,353

Job Type | Date Posti

- Full-time (12061)
- Part-time (89)
- Contract (36)
- Internship (112)
- Temporary (8)
- Apprentice/Trainee (19)
- Entry Level (100)



Data Scientist Salaries in Washington, DC Area

210 Salaries   Updated Oct 3, 2018

Industries | Company Sizes | 0-1 Years

Average Base Pay

$92,158 /yr

12% below national average

data scientist

ːe in Washington, DC

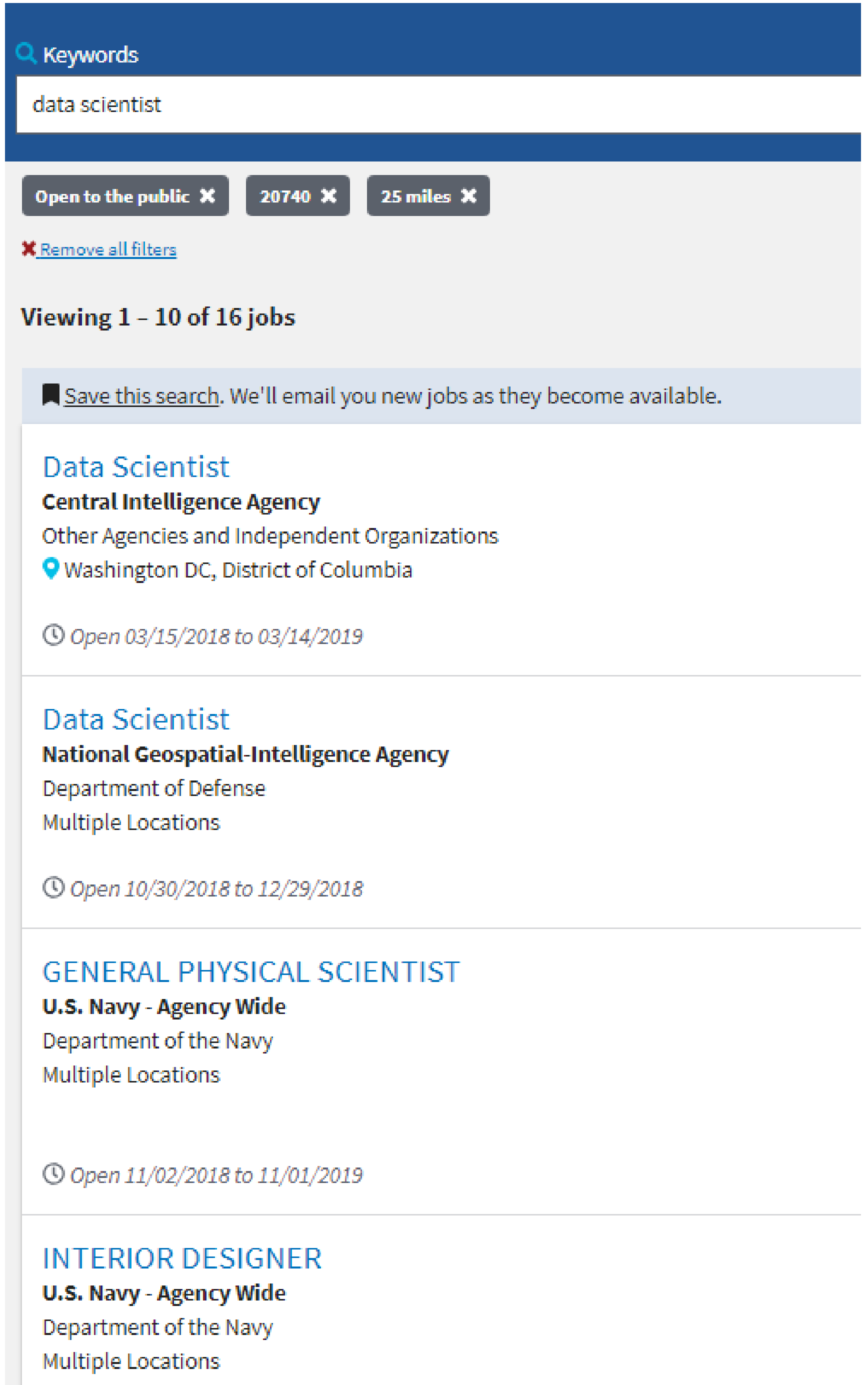
ːt is $124,360 in
ːre based on 210...

Additional Cash Compensation

Average	$9,627
Range	$3,195 - $22,206

Job Type | Dat

- Full-time (2084)
- Part-time (13)
- Contract (2)
- Internship (17)
- Temporary (1)
- Apprentice/Trainee
- Entry Level (10)



Keywords

data scientist

Open to the public ✖ 20740 ✖ 25 miles ✖

✖ Remove all filters

Viewing 1 – 10 of 16 jobs

Save this search. We'll email you new jobs as they become available.

Data Scientist
Central Intelligence Agency
Other Agencies and Independent Organizations
Washington DC, District of Columbia

Open 03/15/2018 to 03/14/2019

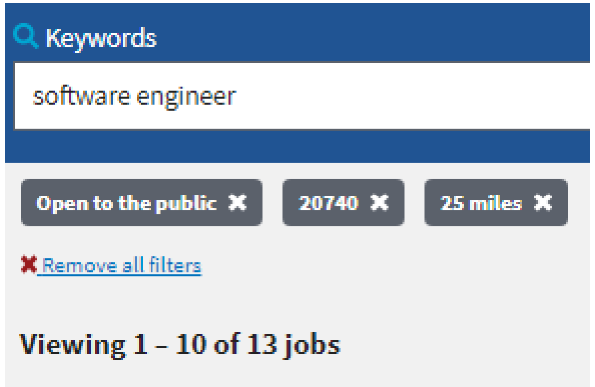
Data Scientist
National Geospatial-Intelligence Agency
Department of Defense
Multiple Locations

Open 10/30/2018 to 12/29/2018

GENERAL PHYSICAL SCIENTIST
U.S. Navy - Agency Wide
Department of the Navy
Multiple Locations

Open 11/02/2018 to 11/01/2019

INTERIOR DESIGNER
U.S. Navy - Agency Wide
Department of the Navy
Multiple Locations



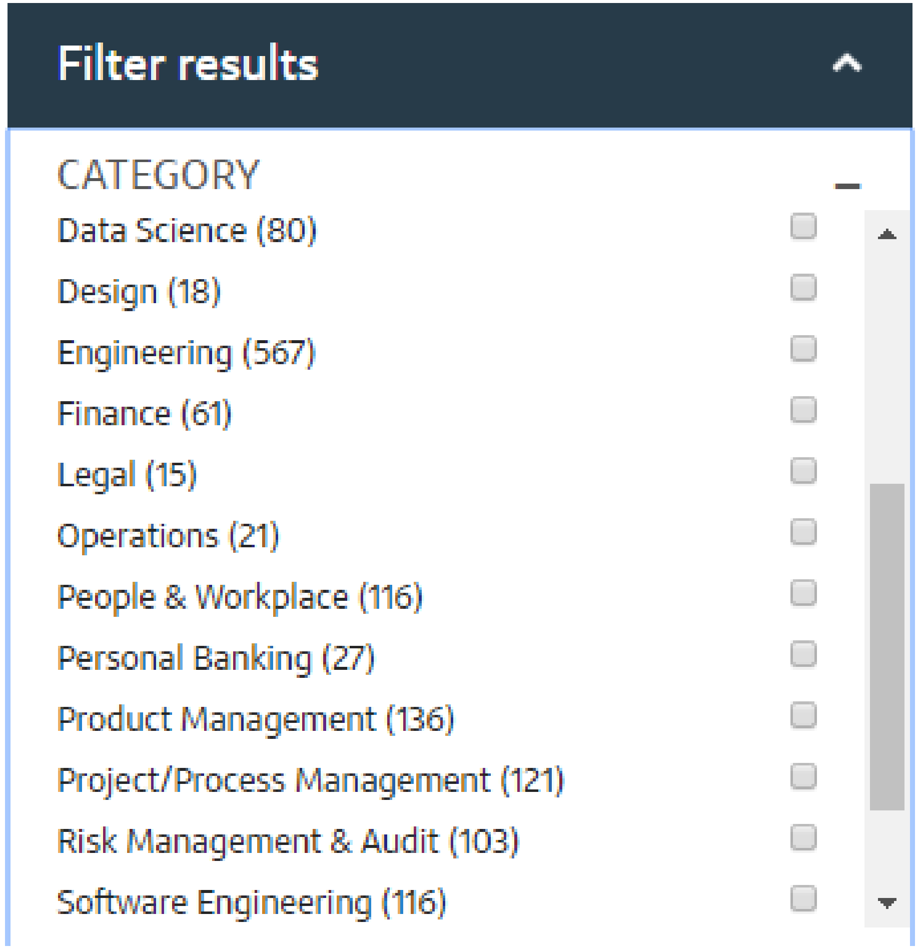
Keywords

software engineer

Open to the public ✖ 20740 ✖ 25 miles ✖

✖ Remove all filters

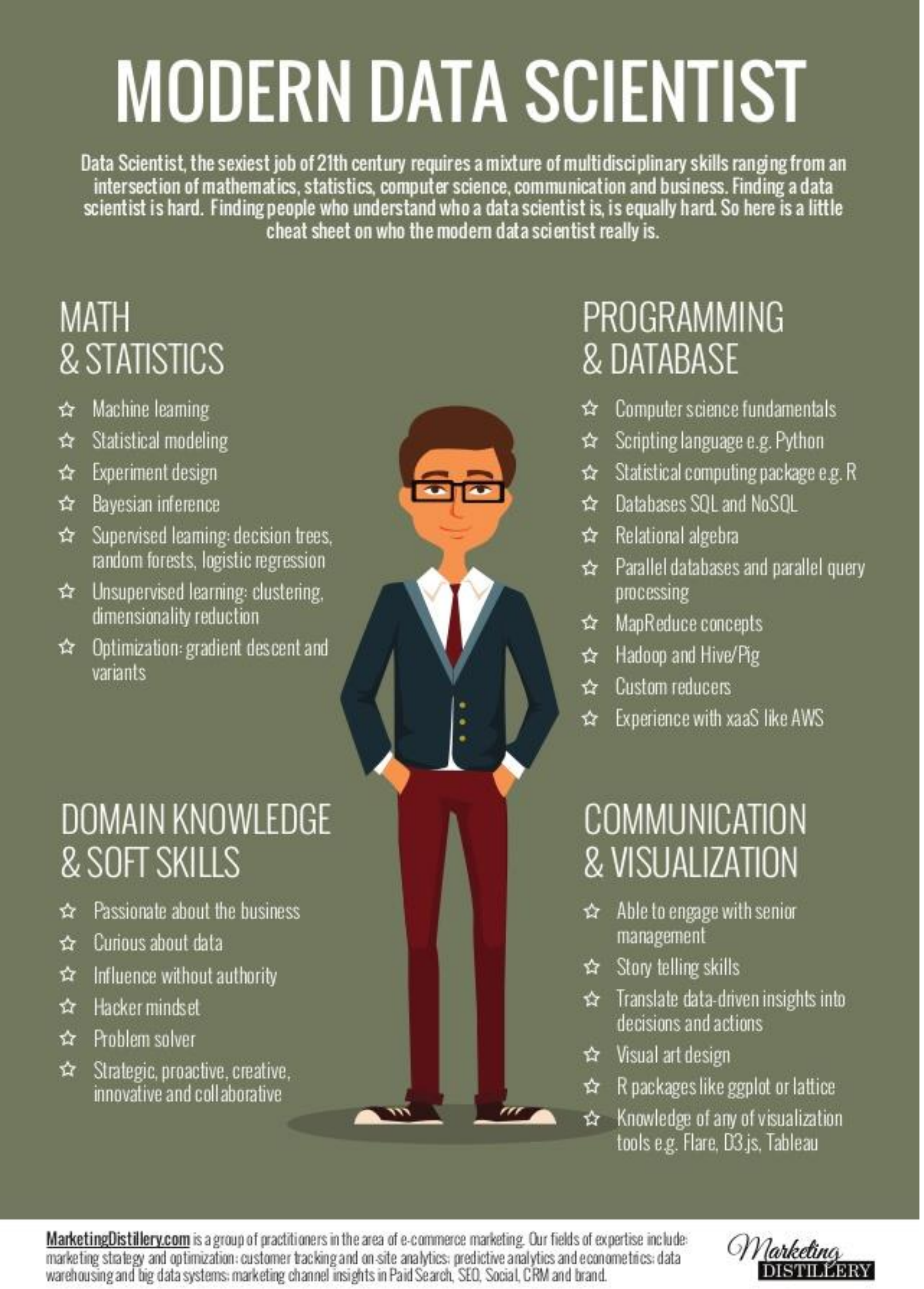
Viewing 1 – 10 of 13 jobs



Filter results

CATEGORY

Data Science (80) □
Design (18) □
Engineering (567) □
Finance (61) □
Legal (15) □
Operations (21) □
People & Workplace (116) □
Personal Banking (27) □
Product Management (136) □
Project/Process Management (121) □
Risk Management & Audit (103) □
Software Engineering (116) □



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th 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.

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 package 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

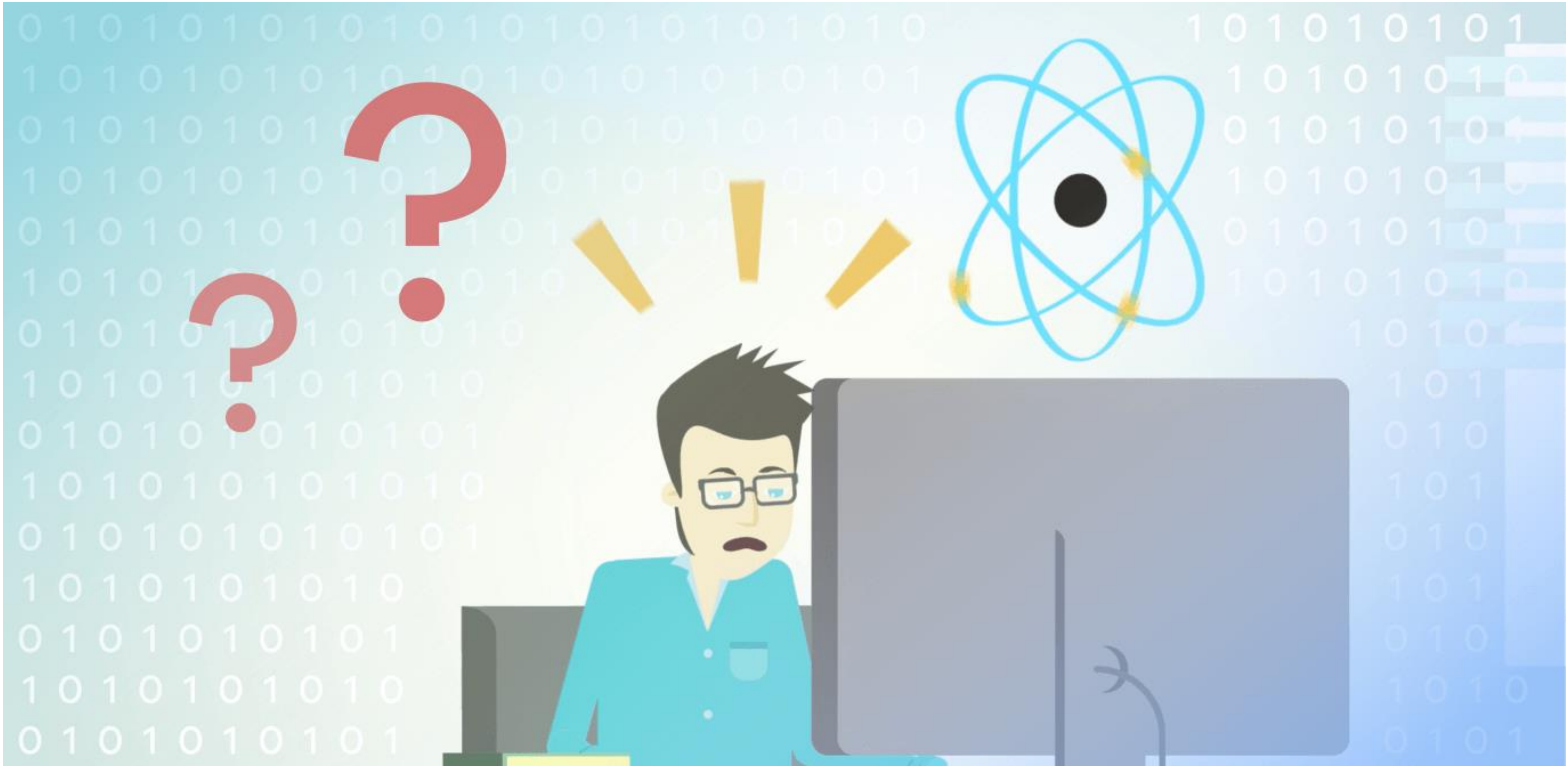
COMMUNICATION & VISUALIZATION

- ☆ 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

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

What is the hardest part of data science?



What is the hardest part of data science?

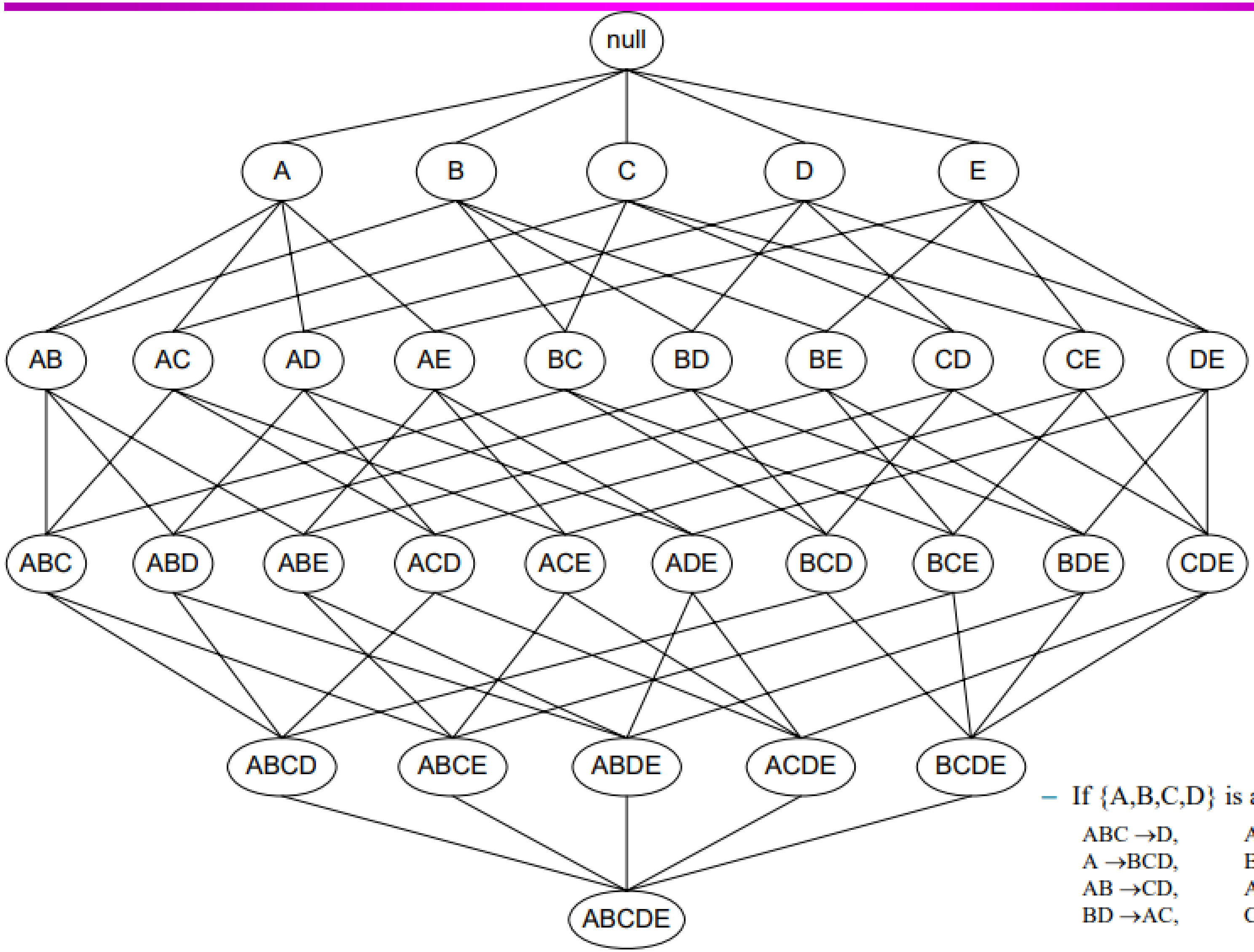
1. **Dealing with people ☺**
2. Figuring out what questions to ask (domain knowledge)
3. Getting the data for those questions
4. Organizing the data
5. Dealing with missing data
6. Training supervised Machine Learning

How to become successful?

Association Rules

1. Gather all frequent item sets (specified by support % of occurrences)
2. From those frequent items consider each possible combination and calculate the supports.
3. The most frequent rules that match our support level are presented.

Basic idea behind rule generation



– If {A,B,C,D} is a frequent itemset, candidate rules:

ABC →D, ABD →C, ACD →B, BCD →A,
A →BCD, B →ACD, C →ABD, D →ABC
AB →CD, AC → BD, AD → BC, BC →AD,
BD →AC, CD →AB,

Association rule terminology

- Confidence: Confidence (x→y) = support(xUy)/support(x)
- Lift: Lift (x→y) = support(xUy)/(support(x)*support(y))
    - If the rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events.
- Conviction:

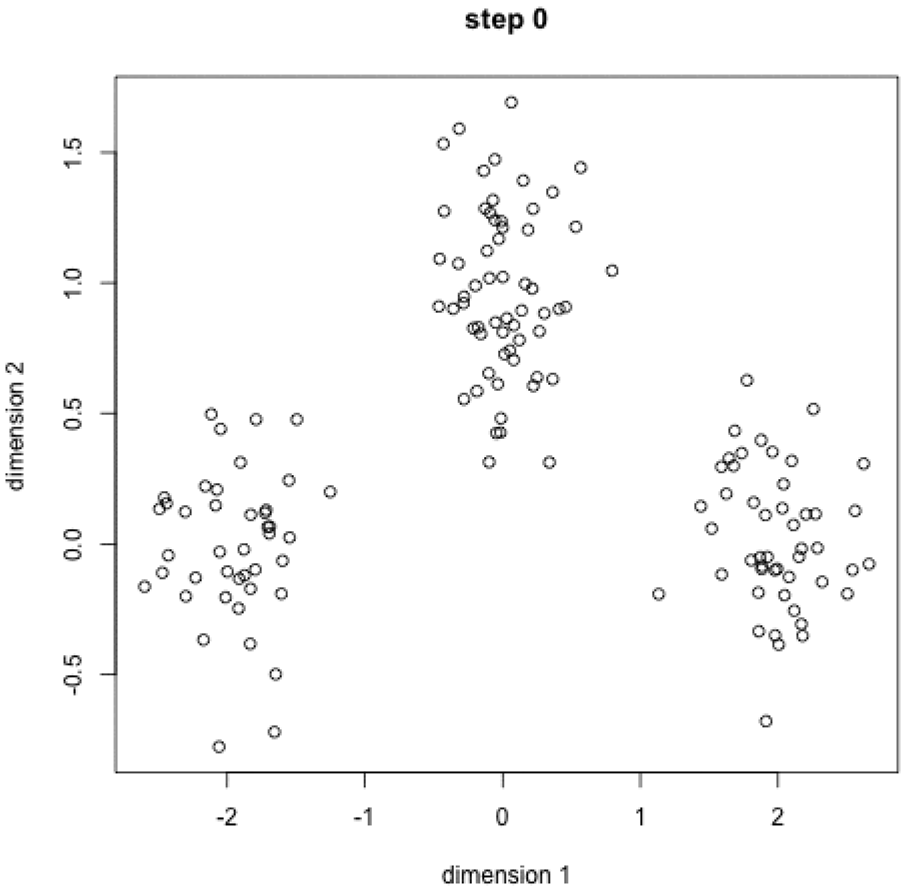
$$\operatorname{conv}(X \Rightarrow Y) = \frac{1 - \operatorname{supp}(Y)}{1 - \operatorname{conf}(X \Rightarrow Y)}.$$

    - interpreted as the ratio of the expected frequency that X occurs
    - without Y (that is to say, the frequency that the rule makes an incorrect prediction)
    - if X and Y were independent divided by the observed frequency of incorrect
    - predictions.

Jupyter Example

Clustering

K-Means will give you clusters that you can label!



Putting it together

1. From clusters you can label them which allows you to engineer statistically the association rule role ups.
2. From there you can see all the rules and target a dependent variable
3. Set your decision tree, random forests, or neural network to target this dependent variable and independent variables
4. Update with feedback from the field and you don’t have to worry about changing any code. Plus what code/variables would you change? The time savings of ML ☺.

Questions ☺