НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Профессия Data Scientist

Леонид Жуков

Отделение Прикладной Математики

Director Data Science Ancestry.com

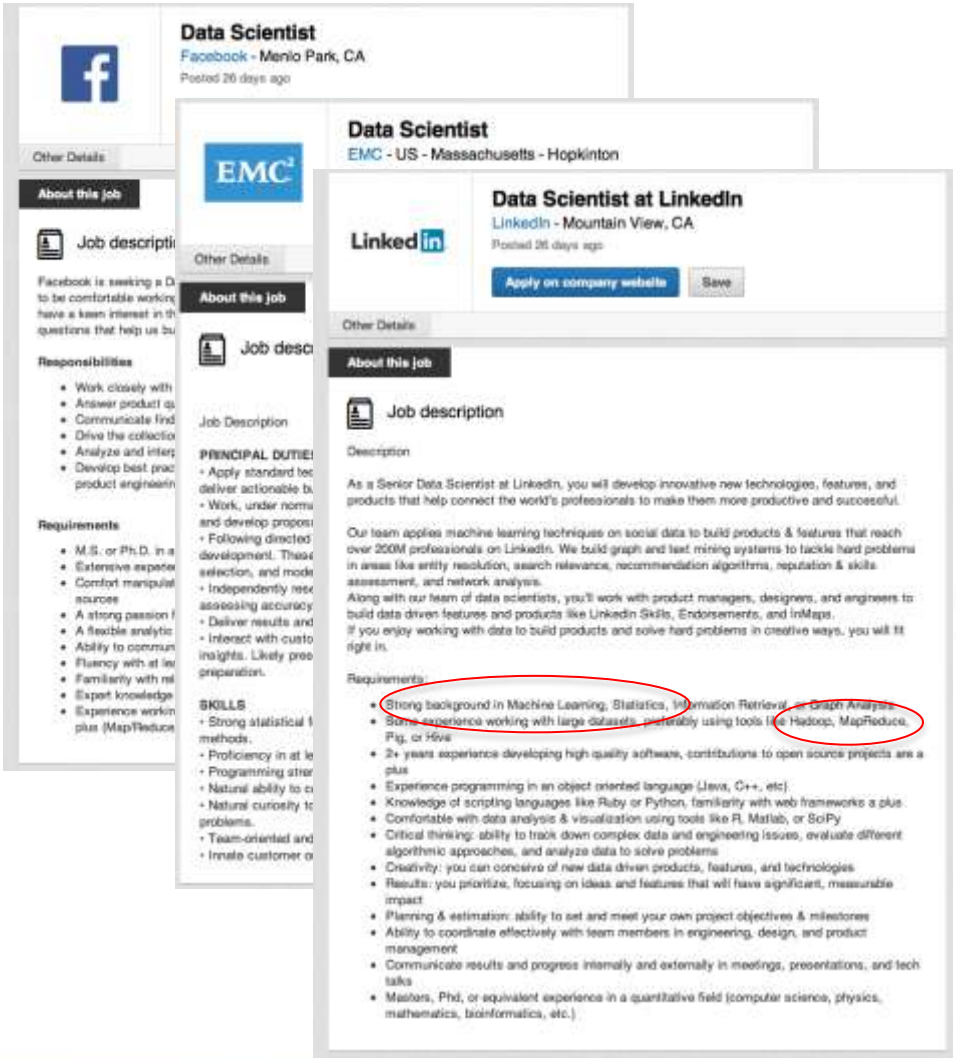
lzhukov@hse.ru

Конференция «Большие Данные в национальной экономике»
Москва 2013

Sexiest job of the 21scentury

McKinsey оценивает нехватку в 140,000-190,000 специалистов к 2018г

Требуются Data Scientists!



Data Scientist
Facebook - Menlo Park, CA
Posted 26 days ago

Other Details
About this job
Job descripti
Facebook is seeking a D
to be comfortable workin
have a keen interest in th
questions that help us bu
Responsibilities
- Work closely with
- Answer product qu
- Communicate find
- Drive the collectio
- Analyze and interp
- Develop best prac
product engineerin
Requirements
- M.S. or Ph.D. in a
- Extensive experie
- Comfort manipulat
sources
- A strong passion f
- A flexible analytic
- Ability to commun
- Fluency with at le
- Familiarity with re
- Expert knowledge
- Experience workin
plus (MapReduce

Data Scientist
EMC - US - Massachusetts - Hopkinton

Other Details
About this job
Job desc
Job Description
PRINCIPAL DUTIES
- Apply standard tec
deliver actionable bu
- Work, under norma
and develop propos
- Following directed
development. These
selection, and mode
- Independently rese
assessing accuracy
- Deliver results and
- Interact with custo
insights. Likely pres
preparation.
SKILLS
- Strong statistical
methods.
- Proficiency in at le
- Programming strer
- Natural ability to c
- Natural curiosity to
problems.
- Team-oriented and
- Innate customer o

Data Scientist at LinkedIn
LinkedIn - Mountain View, CA
Posted 26 days ago
Apply on company website | Save

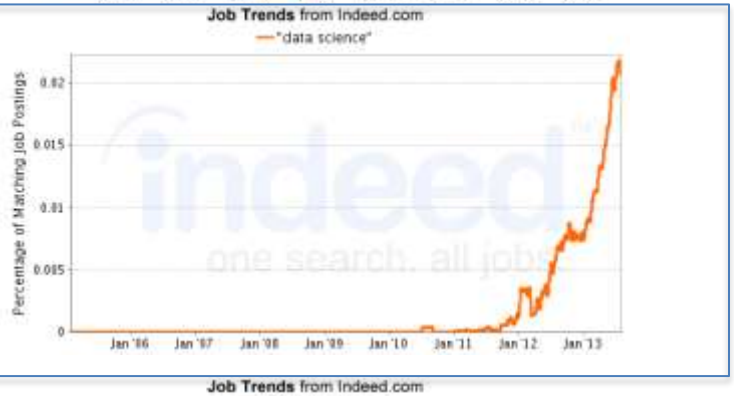
Other Details
About this job
Job description
Description
As a Senior Data Scientist at LinkedIn, you will develop innovative new technologies, features, and products that help connect the world's professionals to make them more productive and successful.
Our team applies machine learning techniques on social data to build products & features that reach over 200M professionals on LinkedIn. We build graph and text mining systems to tackle hard problems in areas like entity resolution, search relevance, recommendation algorithms, reputation & skills assessment, and network analysis.
Along with our team of data scientists, you'll work with product managers, designers, and engineers to build data driven features and products like LinkedIn Skills, Endorsements, and InMaps.
If you enjoy working with data to build products and solve hard problems in creative ways, you will fit right in.
Requirements
- Strong background in Machine Learning, Statistics, Information Retrieval, or Graph Analysis
- Some experience working with large datasets, preferably using tools like Hadoop, MapReduce, Pig, or Hive
- 2+ years experience developing high quality software, contributions to open source projects are a plus
- Experience programming in an object oriented language (Java, C++, etc)
- Knowledge of scripting languages like Ruby or Python, familiarity with web frameworks a plus
- Comfortable with data analysis & visualization using tools like R, Matlab, or SciPy
- Critical thinking: ability to track down complex data and engineering issues, evaluate different algorithmic approaches, and analyze data to solve problems
- Creativity: you can conceive of new data driven products, features, and technologies
- Results: you prioritize, focusing on ideas and features that will have significant, measurable impact
- Planning & estimation: ability to set and meet your own project objectives & milestones
- Ability to coordinate effectively with team members in engineering, design, and product management
- Communicate results and progress internally and externally in meetings, presentations, and tech talks
- Masters, Phd, or equivalent experience in a quantitative field (computer science, physics, mathematics, bioinformatics, etc.)

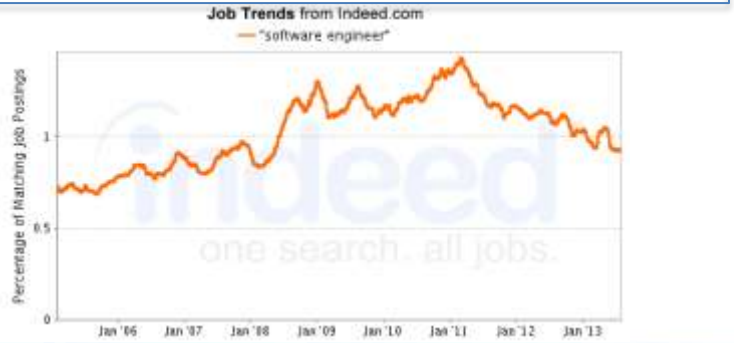
Data Scientist
Apple - Santa Clara Valley - California -US
Posted 19 days ago
Apply on company website | Save

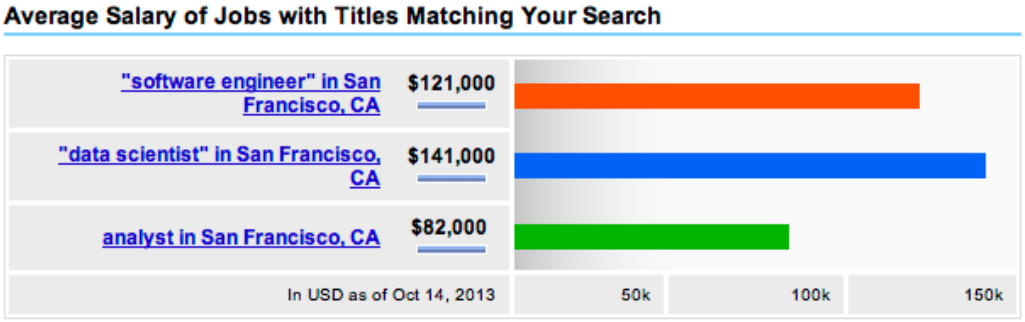
Other Details
About this job
Job description
Apple has a tremendous amount of data, and we have just scratched the surface in pattern detection, anomaly detection, predictive modeling, and optimization. There are many exciting problems to be discovered and solved. We encourage scientists to stay abreast of data mining research by attending conferences and working with academic faculty and students. We foster a collaborative work environment, but allow solution autonomy on projects.
The iTunes Engineering team has a proud tradition of delivering cutting-edge products in a competitive marketplace. We seek to maintain a challenging and rewarding environment where the best engineers and scientists can collaborate and produce real-world improvements in customers' online experience. Successful candidates will solve problems unique in scale and concept in the pursuit of new and original features.
Key Qualifications
- Strong working knowledge of data mining algorithms including decision trees, probability networks, association rules, clustering, regression, and neural networks.
- Familiarity with database modeling and data warehousing principles with a working knowledge of SQL.
- Familiarity with Big Data tools and techniques, including MapReduce, NoSQL stores, and unbounded stream processing.
- Creativity to go beyond current tools to deliver best solution to the problem
- Strong programming skills in Java, Python, or similar language
- Excellent interpersonal, written, and verbal communication skills
- Ability and comfort working independently and making key decisions on projects

Description
We are seeking an outstanding data mining scientist who is interested in designing, developing, and fielding data mining solutions that have direct and measurable impact to Apple. This person will work within and across teams to help identify viable data mining opportunities and then implement end to end analytical solutions. The role requires both a broad knowledge of existing data mining algorithms and creativity to invent and customize when necessary.
Education
Ph.D. in Data Mining, Machine Learning, Statistics, Operations Research or related field
M.S. in related field with 5 years experience applying data mining techniques to real business problems.

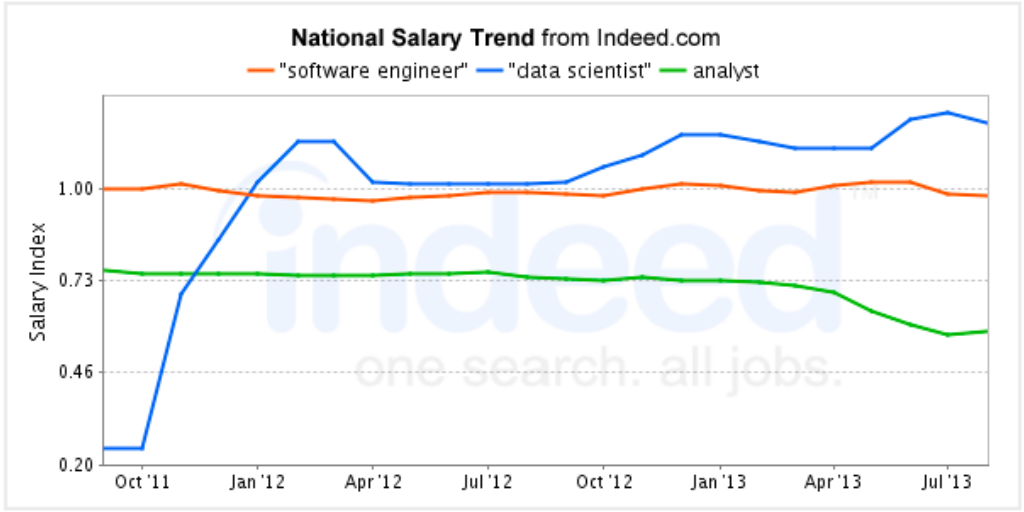
Спрос и предложение







Average "data scientist" salaries for job postings in San Francisco, CA are 71% higher than average analyst salaries for job postings in San Francisco, CA.

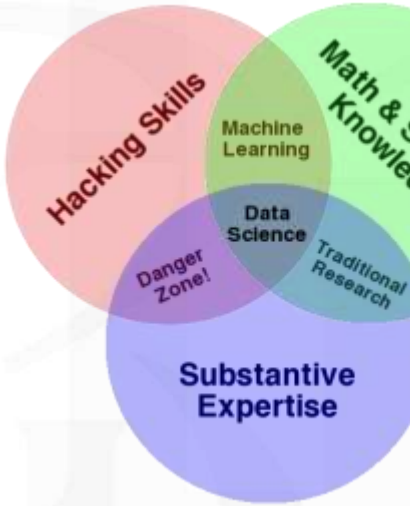


Кто такие Data Scientists?

A practitioner of data science is called a data scientist (Wikipedia)

Data Scientist:

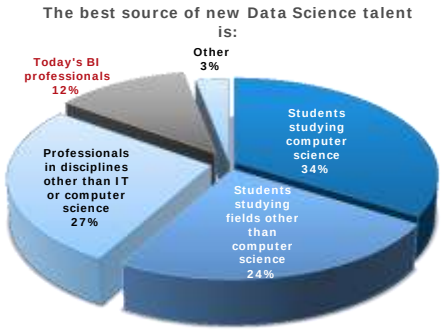
- Любит данные
- Исследовательский склад ума
- Цель работы – нахождение закономерностей в данных
- Практик, не теоретик
- Умеет и любит работать руками
- Эксперт в прикладной области (*)
- Работает в команде



Drew Conway, 2010

Предпочтительное образование:

- Computer Science
- Статистика, математика
- Точные науки: Физика, Инженерия, итд
- Магистры и кандидаты наук

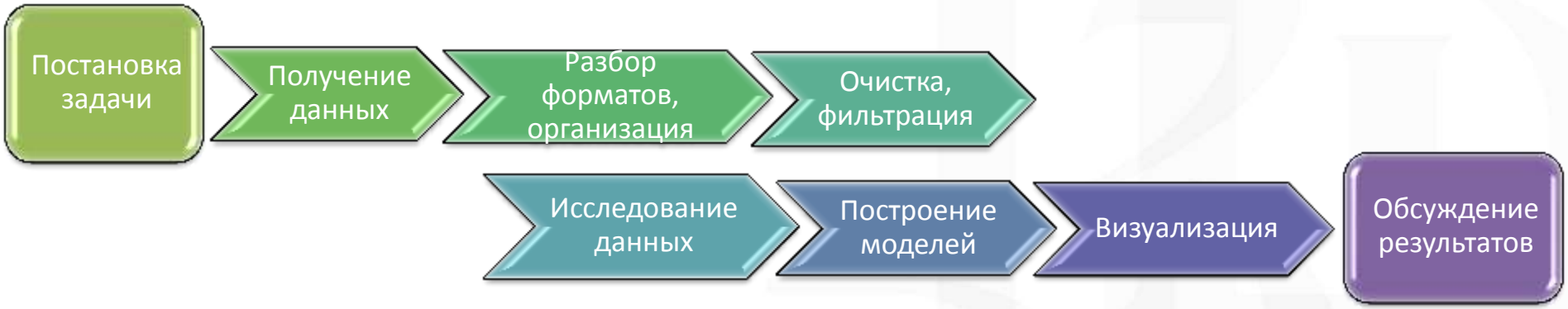


EMC Data Science Community Survey, 2011

Рабочие инструменты

- Operating systems:
  - Linux + shell tools
- Big data instruments:
  - Hadoop (MapReduce) + hadoop tools
  - Hive, Pig
  - NoSQL (Hbase, MongoDB, Cassandra, Neo4J)
- Database:
  - SQL
- Programming:
  - Python
  - Java
  - Scala
- Machine Learning:
  - R
  - Matlab
  - Python libraries (NumPy, SciPy, Nltk,…)
  - Java libraries (Mahaut)

День из жизни Data Scientist

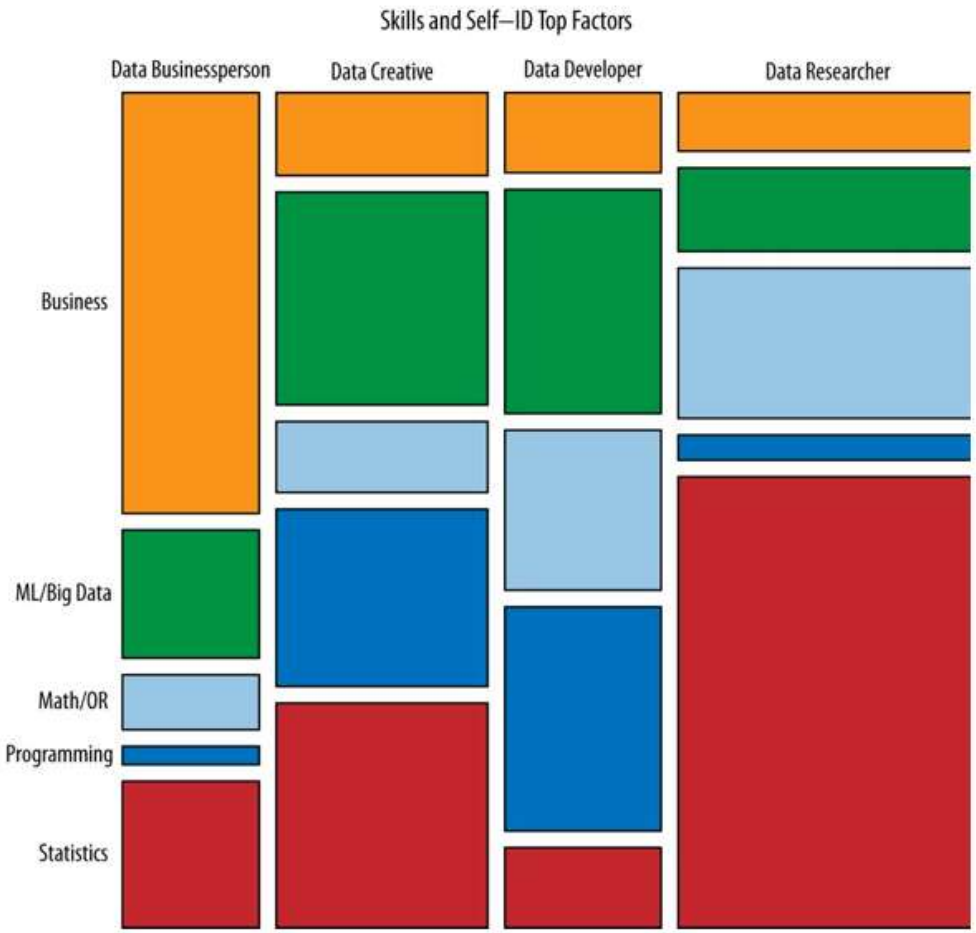


Data Scientist или Аналитик

- Data Scientist:
  - Используют Hadoop, MapReduce, Hive, R
  - Создают специализированные системы и инструменты
  - Работают со структурированными и не структурированными данными
  - Рабочие данные измеряются в TB, PB
  - Опыт научной работы, экспертиза в статистке, машинном обучении, программировании
  - Магистры и кандидаты наук (PhDs)
  - Разрабатывают предсказательными модели
  - Создают data products
- Analysts:
  - Используют Excel, SQL
  - Используют существующие инструменты и системы
  - Работают с табличными данными
  - Данные измеряются MB,GB
  - Профессиональное образование, нет формального научного
  - Бакалавры etc (BS, BA, MS, MBA)
  - Работают тесно с BI и маркетингом
  - Создают отчеты и описывают данные
  - Чаще всего данные о показателях работы бизнеса

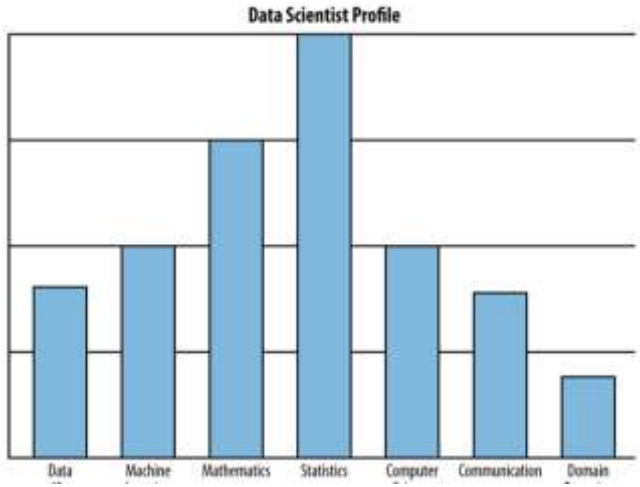
Если Вы программируете, то скорее всего Вы - Data Scientist, если используете Excel, то - аналитик

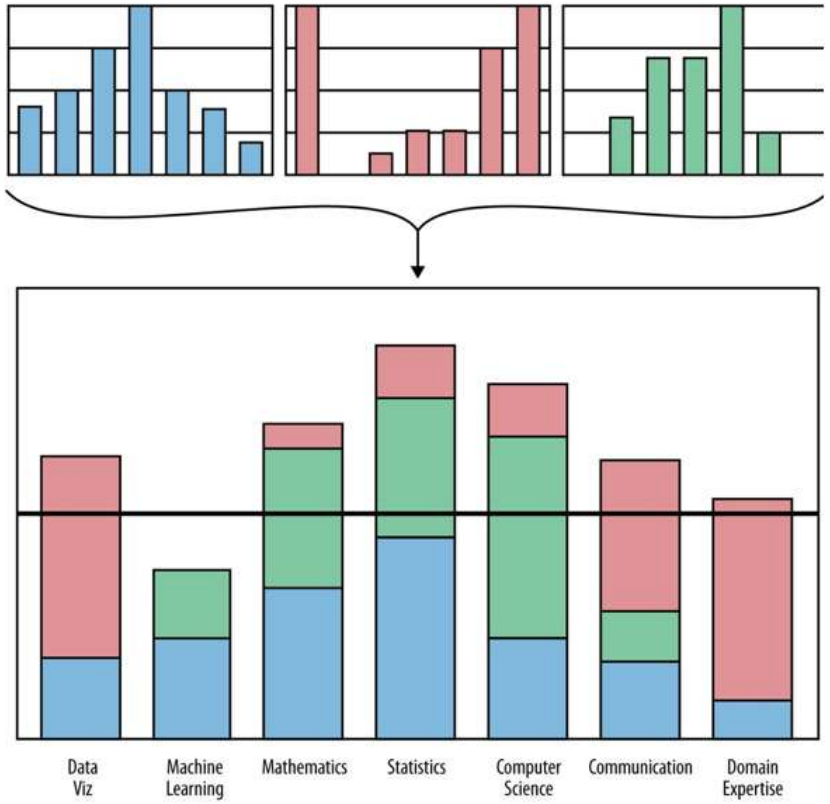
Опрос: роли и навыки Data Scientist



From: “Analyzing the Analyzers” by Harlan Harris, Sean Murphy, and Marck Vaisman , O’Reilly Strata 2012

Data Science команда - ”the dream team”





From: “Doing Data Science: Straight Talk from the Frontline”, Rachel Schutt, Cathy O'Neil, O'Reilly Media, 2013

Прикладные задачи

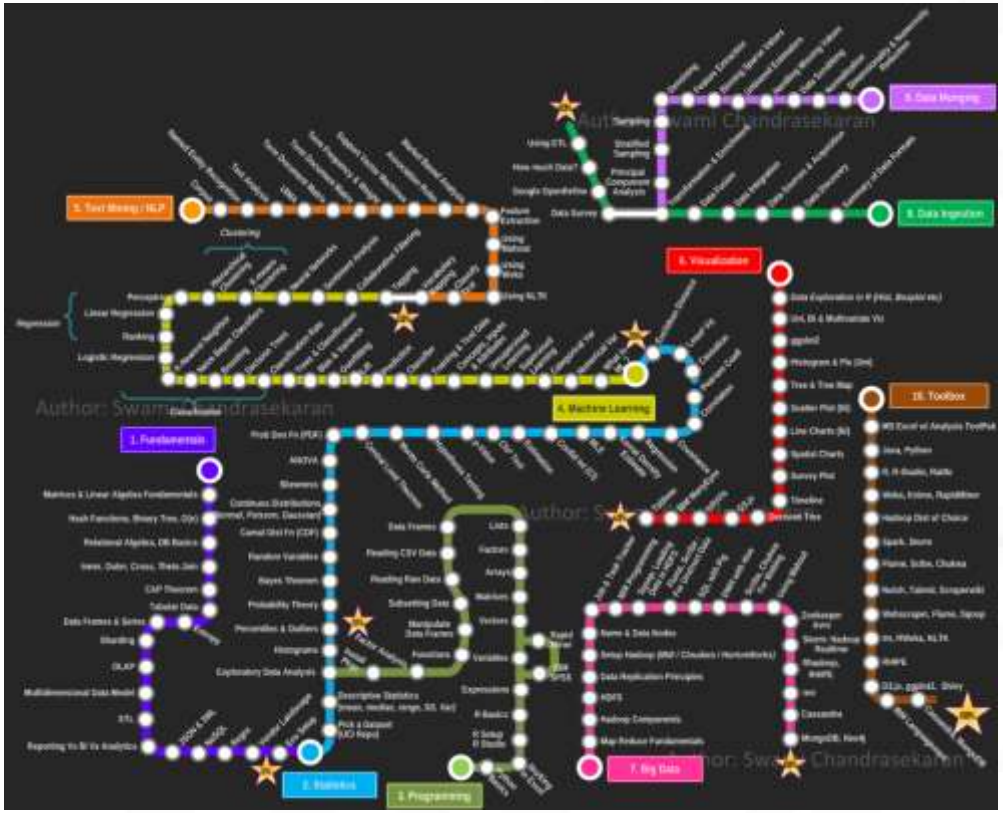
- Маркетинг:
  - Сегментация рынка
  - Моделирование приобретения и оттока клиентов
  - Рекомендательные системы
  - Анализ социальных медиа

- Финансовые и страховые компании:
  - Предотвращение fraud
  - Детектирование аномального поведения
  - Анализ кредитных рисков
  - Страховые моделирование
  - Оптимизация портфолио

- Здравоохранение и Фармакология:
  - Генетический анализ
  - Анализ клинических испытаний
  - Клинические системы принятия решени

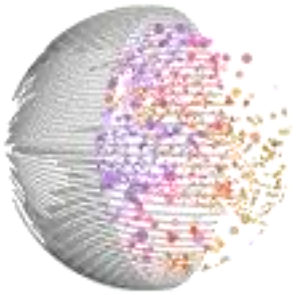
Дорога дальняя...

- Программирование
- Алгоритмы и структуры данных
- Базы данных
- Статистика
- Анализ данных
- Машинное обучение
- Компьютерная обработка текста
- Распределенные системы
- Инструменты Big Data
- Визуализация данных



From: Swami Chandrasekaran,Executive Architect, IBM, Watson Solutions

Подготовительные программы в индустрии



cloudera
Ask Bigger Questions

Course Outline: Cloudera Introduction to Data Science

Introduction

Data Science Overview
- What Is Data Science?
- The Growing Need for Data Science
- The Role of a Data Scientist

Use Cases
- Finance
- Retail
- Advertising
- Defense and Intelligence
- Telecommunications and Utilities
- Healthcare and Pharmaceuticals

Project Lifecycle
- Steps in the Project Lifecycle
- Lab Scenario Explanation

Data Acquisition
- Where to Source Data
- Acquisition Techniques

Evaluating Input Data
- Data Formats
- Data Quantity
- Data Quality

Data Transformation
- Anonymization
- File Format Conversion
- Joining Datasets

Data Analysis and Statistical Methods
- Relationship Between Statistics and Probability
- Descriptive Statistics
- Inferential Statistics

Fundamentals of Machine Learning
- Overview
- The Three Cs of Machine Learning
- Spotlight: Naïve Bayes Classifiers
- Importance of Data and Algorithms

Recommender Overview
- What Is a Recommender System?
- Types of Collaborative Filtering
- Limitations of Recommender Systems
- Fundamental Concepts

Introduction to Apache Mahout
- What Apache Mahout Is (and Is Not)
- A Brief History of Mahout
- Availability and Installation
- Demonstration: Using Mahout's Item-Based Recommender

Implementing Recommenders with Apache Mahout
- Overview
- Similarity Metrics for Binary Preferences
- Similarity Metrics for Numeric Preferences
- Scoring

Experimentation and Evaluation
- Measuring Recommender Effectiveness
- Designing Effective Experiments
- Conducting an Effective Experiment
- User Interfaces for Recommenders

Production Deployment and Beyond
- Deploying to Production
- Tips and Techniques for Working at Scale
- Summarizing and Visualizing Results
- Considerations for Improvement
- Next Steps for Recommenders

Conclusion

Appendix A : Hadoop Overview

Appendix B: Mathematical Formulas

Appendix C : Language and Tool Reference

Подготовительные программы в индустрии

EMC² PROVEN PROFESSIONAL

EMC²

DATA SCIENCE AND BIG DATA ANALYTICS COURSE OUTLINE

Applying a hands-on practitioner's approach to the techniques and tools required for Big Data Analytics.

Introduction	Big Data Overview	State of the practice in analytics	The role of the Data Scientist	Big Data Analytics in industry verticals			
	Introduction to Big Data Analytics						
Analytics Lifecycle	Key roles for a successful analytics project	Main phases of the lifecycle	Developing core deliverables for stakeholders				
	End-to-end data analytics lifecycle						
Basic Methods	Introduction to R	Analyzing and exploring data with R	Statistics for model building and evaluation				
	Using R to execute basic analytics methods						
Adv. Methods	K-Means Clustering	Association Rules	Linear and Logistic Regression	Naïve Bayesian Classifier	Decision Trees	Time Series Analysis	Text Analysis
	Advanced analytics and statistical modeling for Big Data – Theory and Methods						
Tools	Using MapReduce/Hadoop for analyzing unstructured data	Hadoop ecosystem of tools	In-database Analytics	MADlib and Advanced SQL Techniques			
	Advanced analytics and statistical modeling for Big Data – Technology and Tools						
Lab	How to operationalize an analytics project	Creating the Final Deliverables	Data Visualization Techniques	Hands-on Application of Analytics Lifecycle to a Big Data Analytics Problem			
	Endgame, or Putting it all together						

Образовательные программы

Университетские программы:

- University of Washington: Certificate in Data Science
- UC Berkeley: Master of information and data science program
- New York University: Data Science at NYU
- Columbia University: Institute for Data Sciences and Engineering
- University of Southern California (UCS) : Master of Science in Data Science

Онлайн курсы обучения (MOOC):

- Coursera
- edX
- Udacity

Ускоренные образовательные программы (компании):

- Zipfian Academy (12 weeks intensive program)
- Insight Data Science Fellows program (6 weeks post doc training)

Конференции

Индустрийные конференции и выставки:

- O'Reilly Strata Conference Making Data Work
- Hadoop world
- Big Data Techcon
- Big Data Innovation summits

Научные и академические конференции (peer reviewed):

- IEEE & ACM Supercomputing
- IEEE Big Data
- ACM KDD Knowledge Discovery and Data Mining
- ACM SIGIR Information Retrieval
- ICML International Conference on Machine Learning
- NIPS Neural Information Processing
- WWW World Wide Web Conference
- VLDB Very Large Data Bases
- IEEE Visualization

Meetups («кружки по интересам»)

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Книги

Открытые вопросы

- Насколько важно быть экспертом в предметной области решаемой задачи (domain expertise) ?
- Что более важно в профессии Data Scientist : образование или практический опыт?
- Перспективы профессии Data Scientist, будут ли она замещена программными решениями?

ВШЭ Отделение Прикладной Математики

Курсы, читаемые на отделении:

- Программирование (Python, Java, Matlab)
- Методы разработки данных
- Машинное обучение
- Статистика
- Компьютерная лингвистика
- Анализ социальных сетей
- Распределенные системы
- Основы визуализации

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Спасибо за внимание!

101000, Россия, Москва, Мясницкая ул., д. 20
Тел.: (495) 621-7983, факс: (495) 628-7931
www.hse.ru