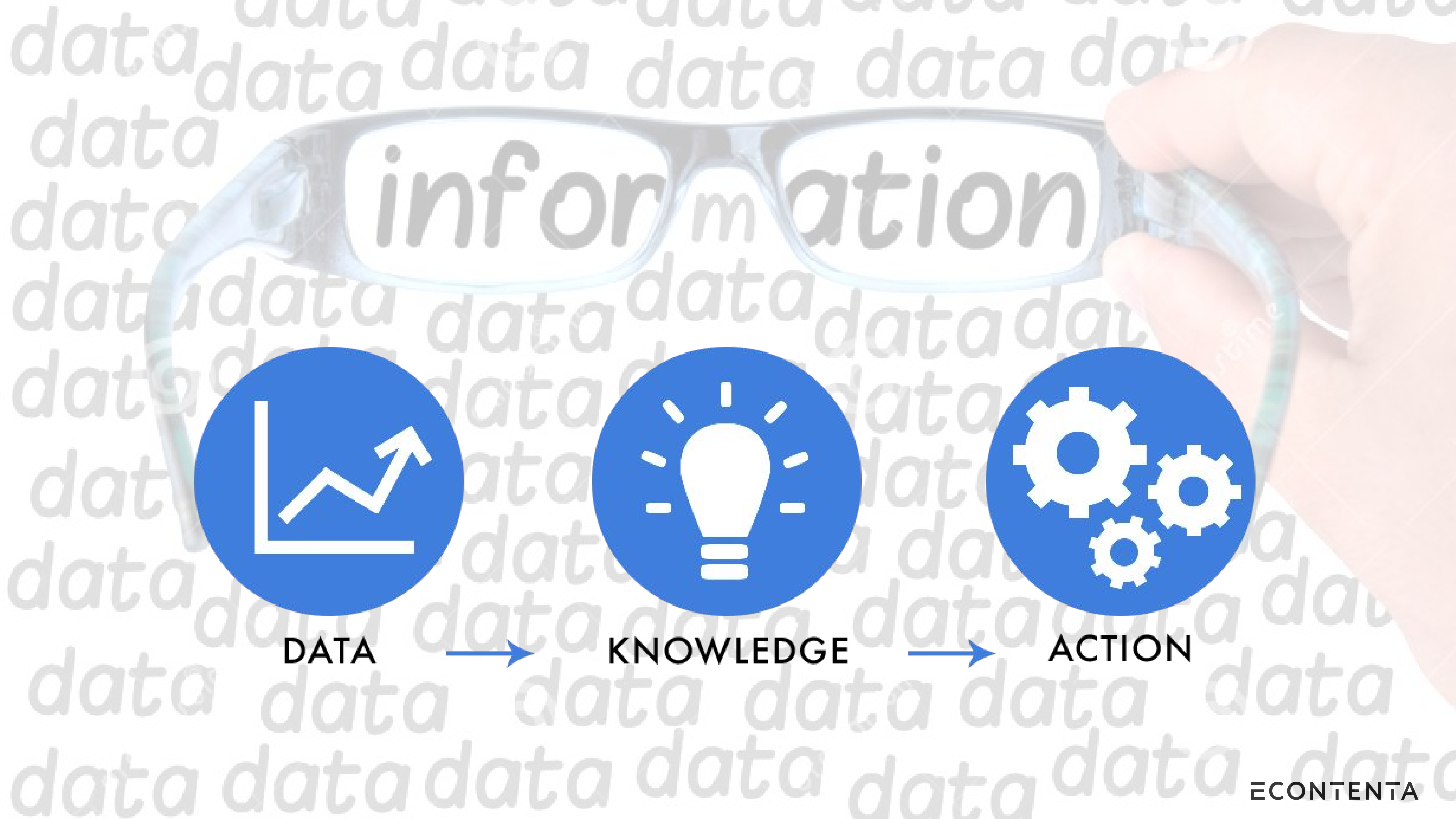
Введение в Анализ Данных

Александр Сенов

Data Sceintist @ E-Contenta

СУТЬ



DATA → KNOWLEDGE → ACTION

ПРИЛОЖЕНИЯ

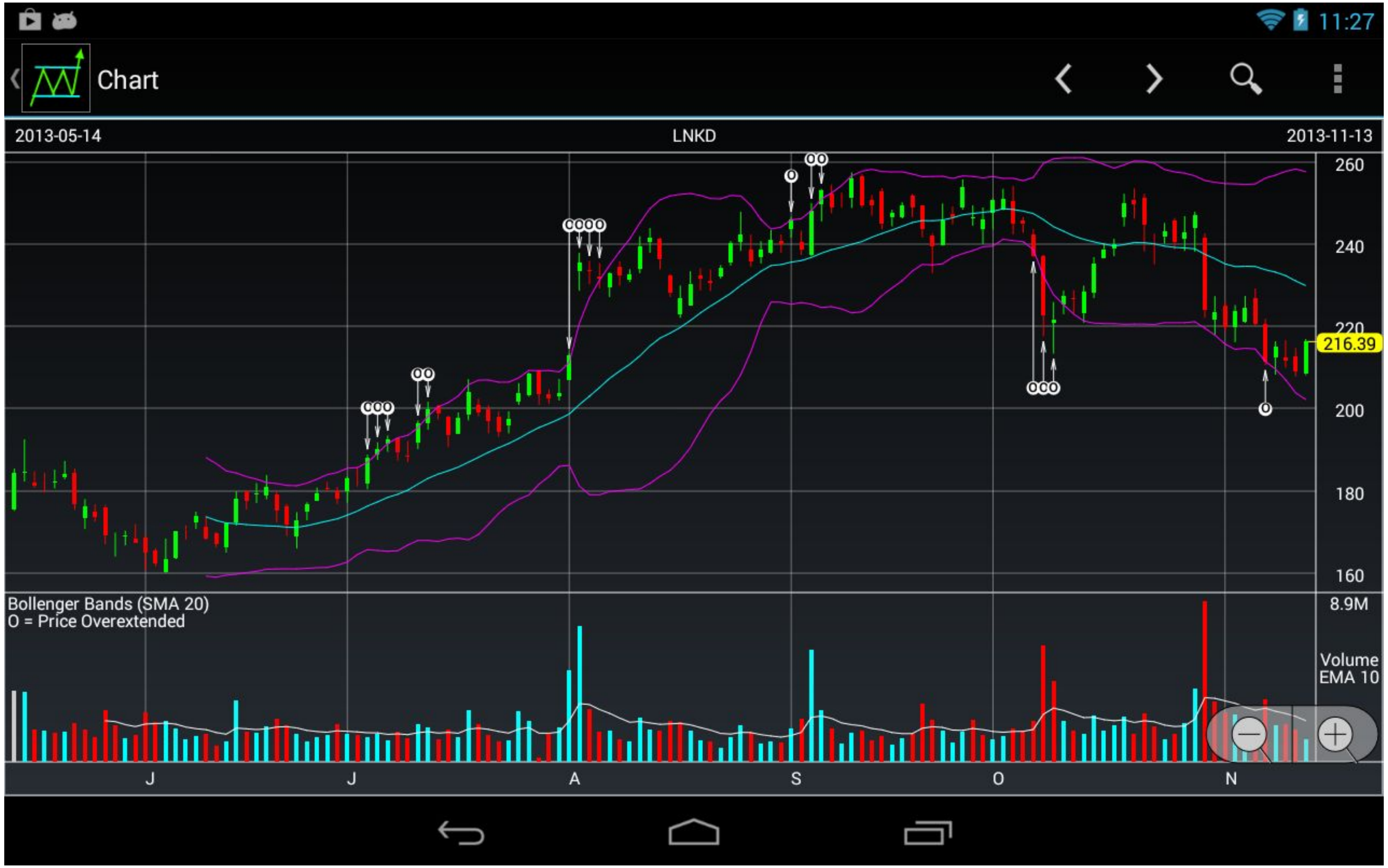
WIN %
RECORD
EXPECTED
ACTUAL
EXPECTED
ACTUAL
OAK 2001
.6525
.6296
106-56
102-60
OAK
.6193
99-63
0-0
Sea 2001
.6861
.7160
111-51
.6097
99-63
.5597
.5864
91-71
.6178
100-62

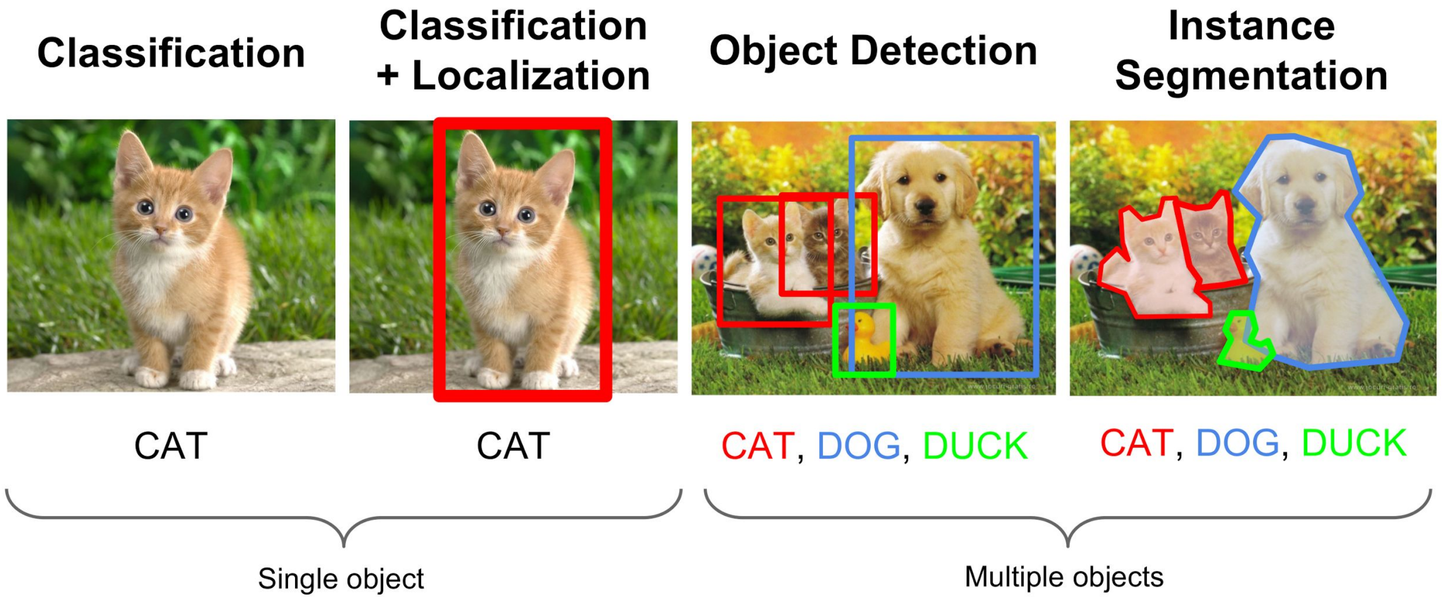
SPAM

Search

Credit Score
Excellent
Good
Fair
Uncertain

7017
1234
7000
VALID THRU
7000
CARDHOLDER NAME
COMPANY


11:27
Chart
2013-05-14
LNKD
2013-11-13
260
240
220
216.39
200
180
160
Bollenger Bands (SMA 20)
O = Price Overextended
8.9M
Volume
EMA 10
J
J
A
S
O
N



Classification

Classification + Localization

Object Detection

Instance Segmentation

CAT

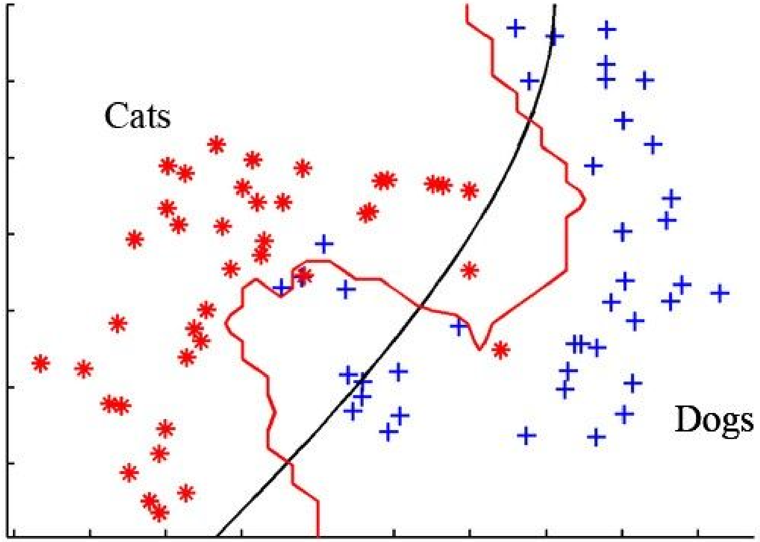
CAT

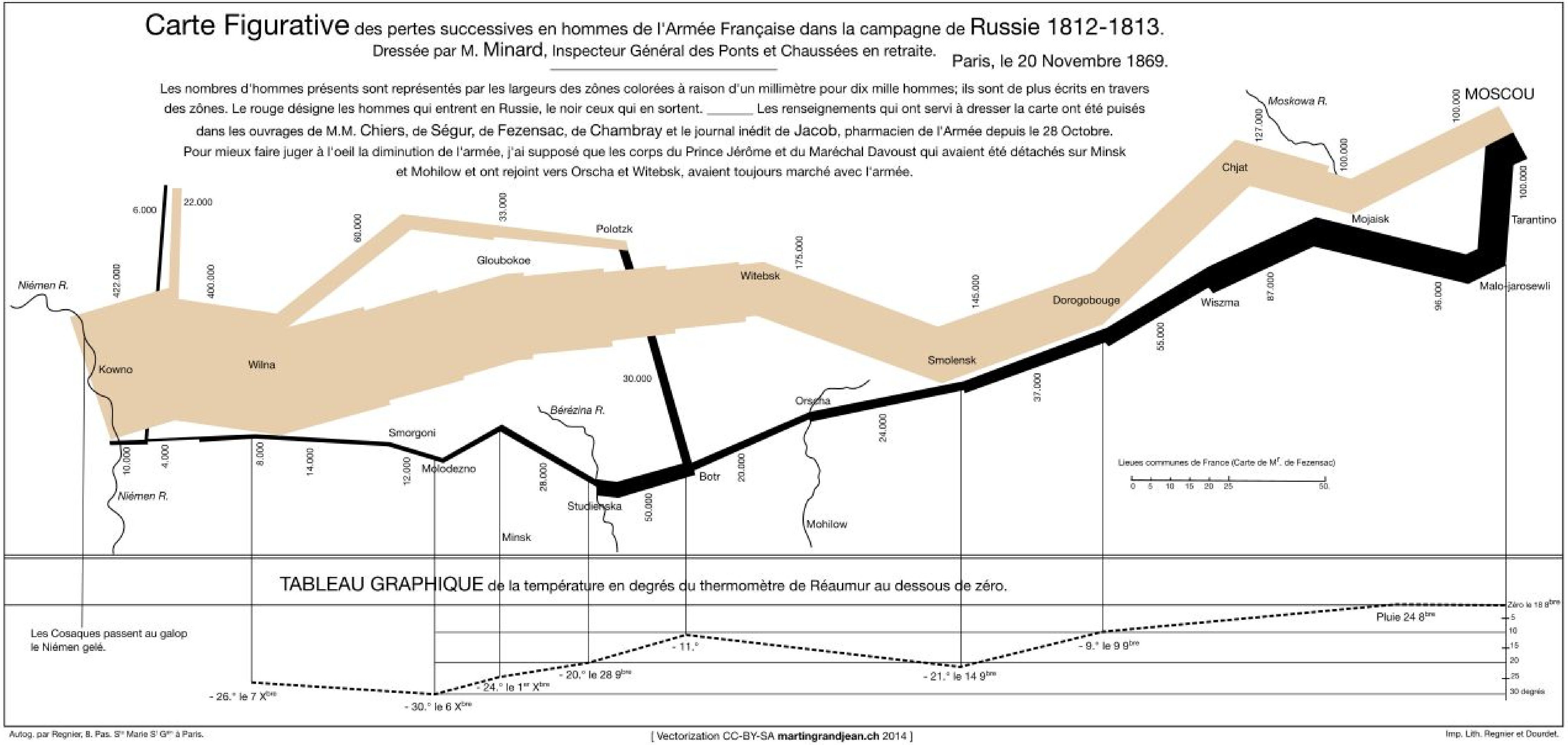
CAT, DOG, DUCK

CAT, DOG, DUCK

Single object

Multiple objects





Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

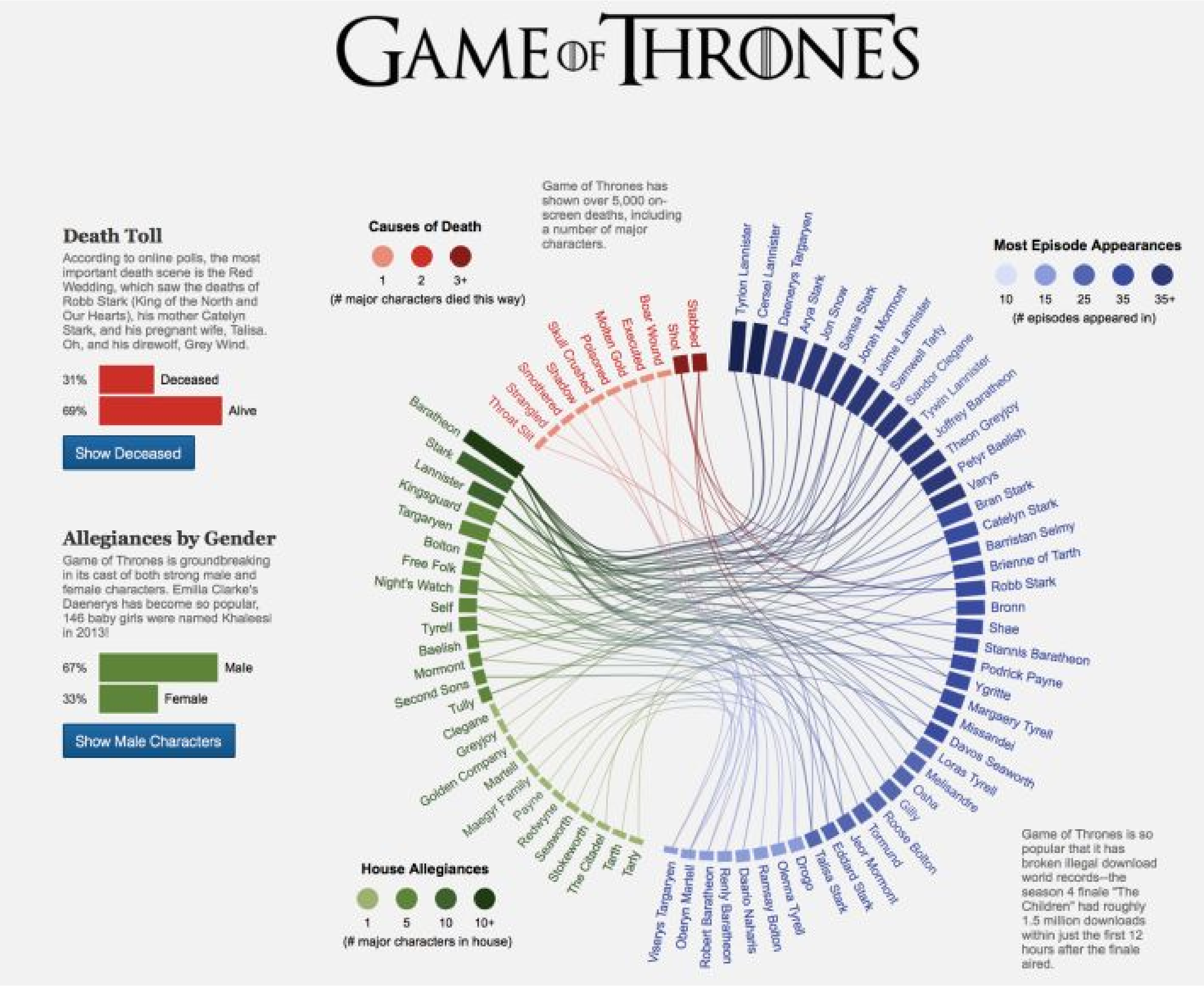
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. ______ Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Chiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre. Pour mieux faire juger à l'oeil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Cosaques passent au galop le Niémen gelé.



GAME OF THRONES

Death Toll

According to online polls, the most important death scene is the Red Wedding, which saw the deaths of Robb Stark (King of the North and Our Hearts), his mother Catelyn Stark, and his pregnant wife, Talisa. Oh, and his direwolf, Grey Wind.

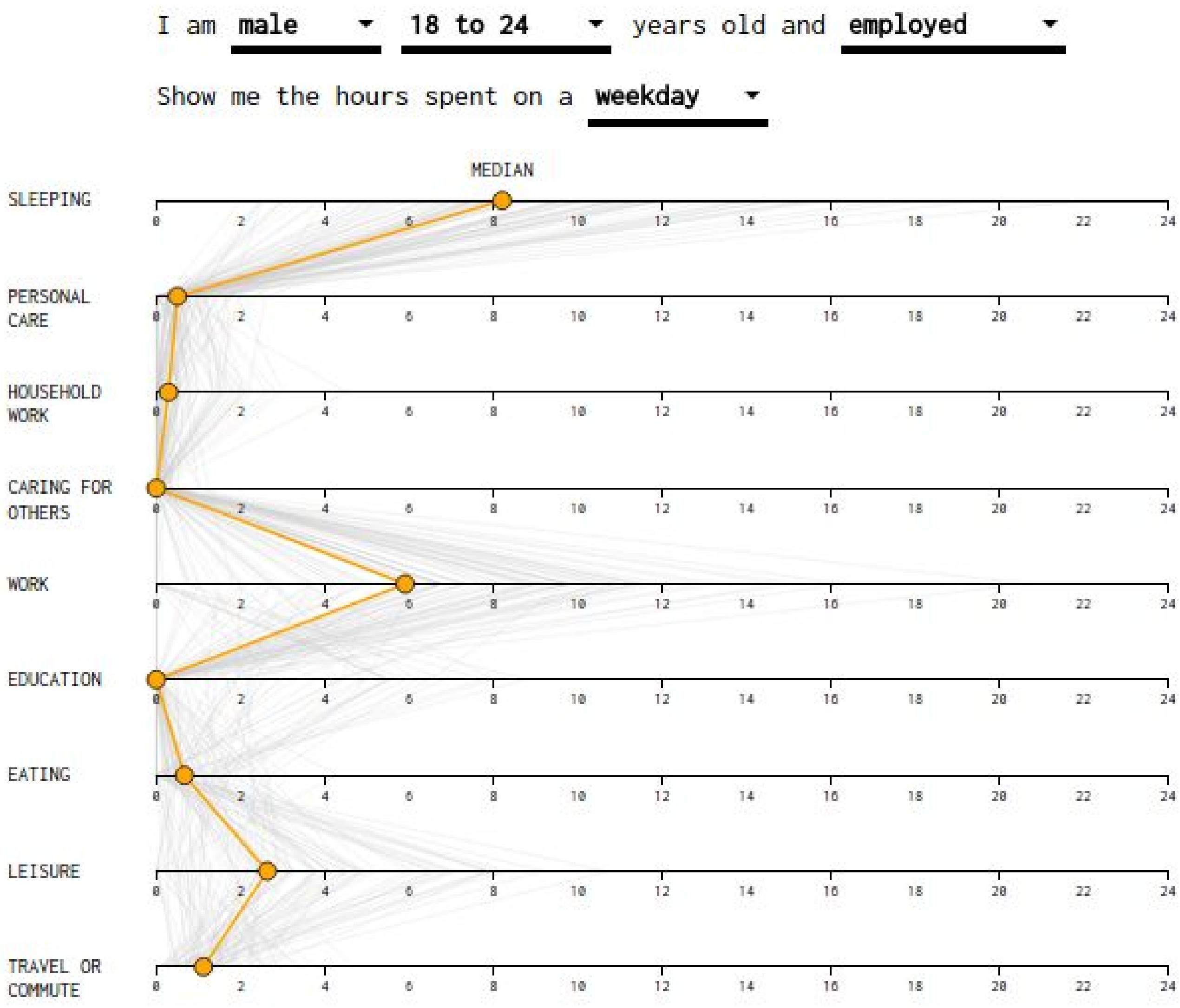
Allegiances by Gender

Game of Thrones is groundbreaking in its cast of both strong male and female characters. Emilia Clarke's Daenerys has become so popular, 146 baby girls were named Khaleesi in 2013!

Game of Thrones has shown over 5,000 on-screen deaths, including a number of major characters.

Game of Thrones is so popular that it has broken illegal download world records—the season 4 finale "The Children" had roughly 1.5 million downloads within just the first 12 hours after the finale aired.

HTTP://THRONESVIZ.GITHUB.IO/

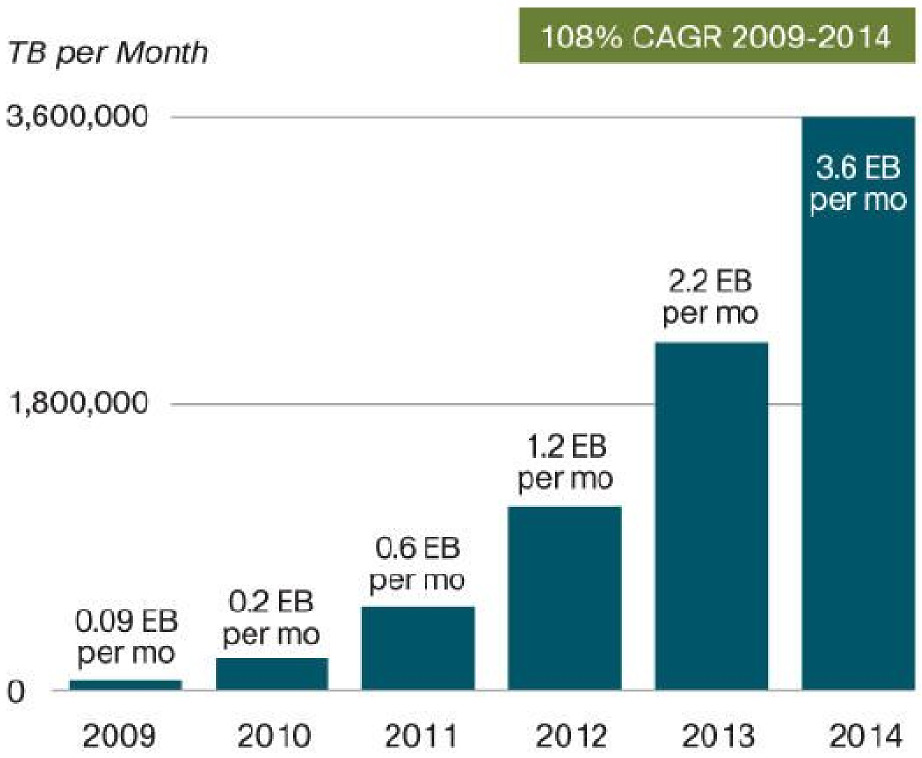


I am male ▾ 18 to 24 ▾ years old and employed ▾

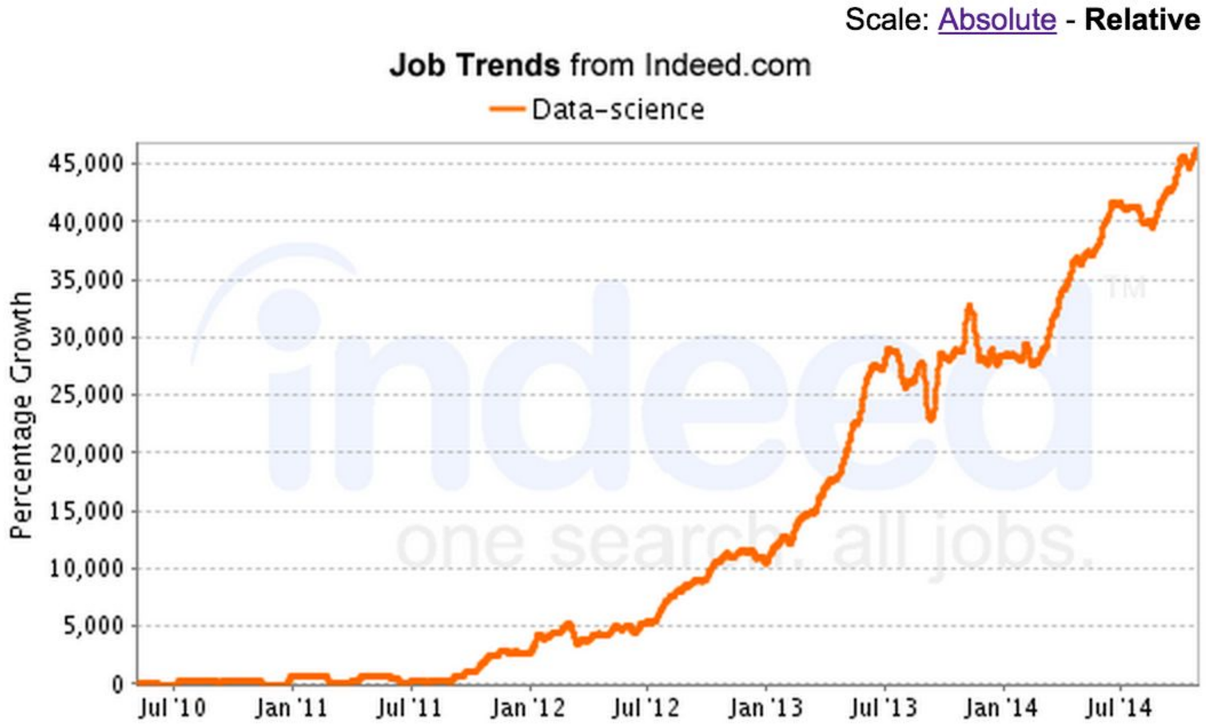
Show me the hours spent on a weekday ▾

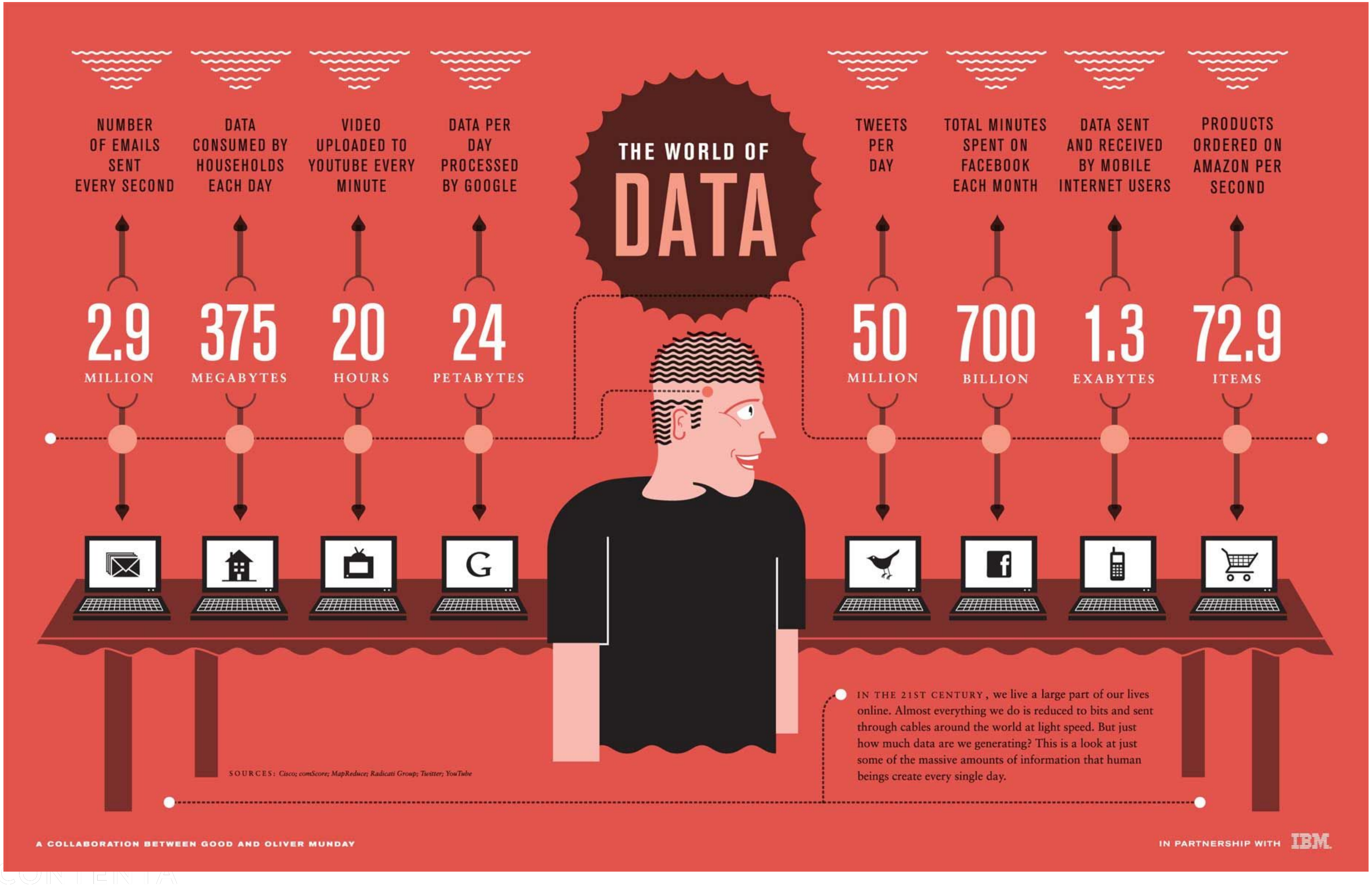
http://blog.visme.co/best-information-graphics-2016/

ЗАЧЕМ



For more details, see Appendix B: Forecast and Methodology.
Source: Cisco VNI Mobile, 2010





THE WORLD OF DATA

NUMBER OF EMAILS SENT EVERY SECOND
2.9 MILLION

DATA CONSUMED BY HOUSEHOLDS EACH DAY
375 MEGABYTES

VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20 HOURS

DATA PER DAY PROCESSED BY GOOGLE
24 PETABYTES

TWEETS PER DAY
50 MILLION

TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700 BILLION

DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3 EXABYTES

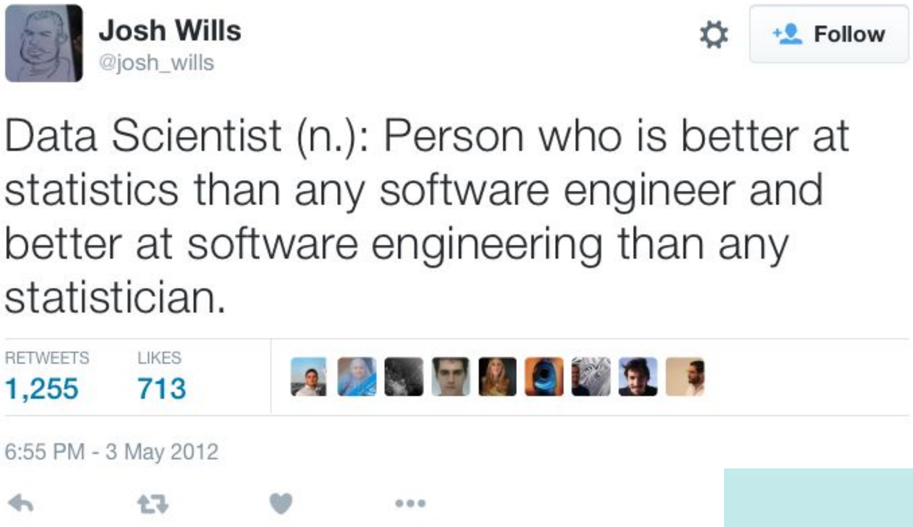
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9 ITEMS

IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

SOURCES: *Cisco; comScore; MapReduce; Radicati Group; Twitter; YouTube*

A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY

IN PARTNERSHIP WITH IBM



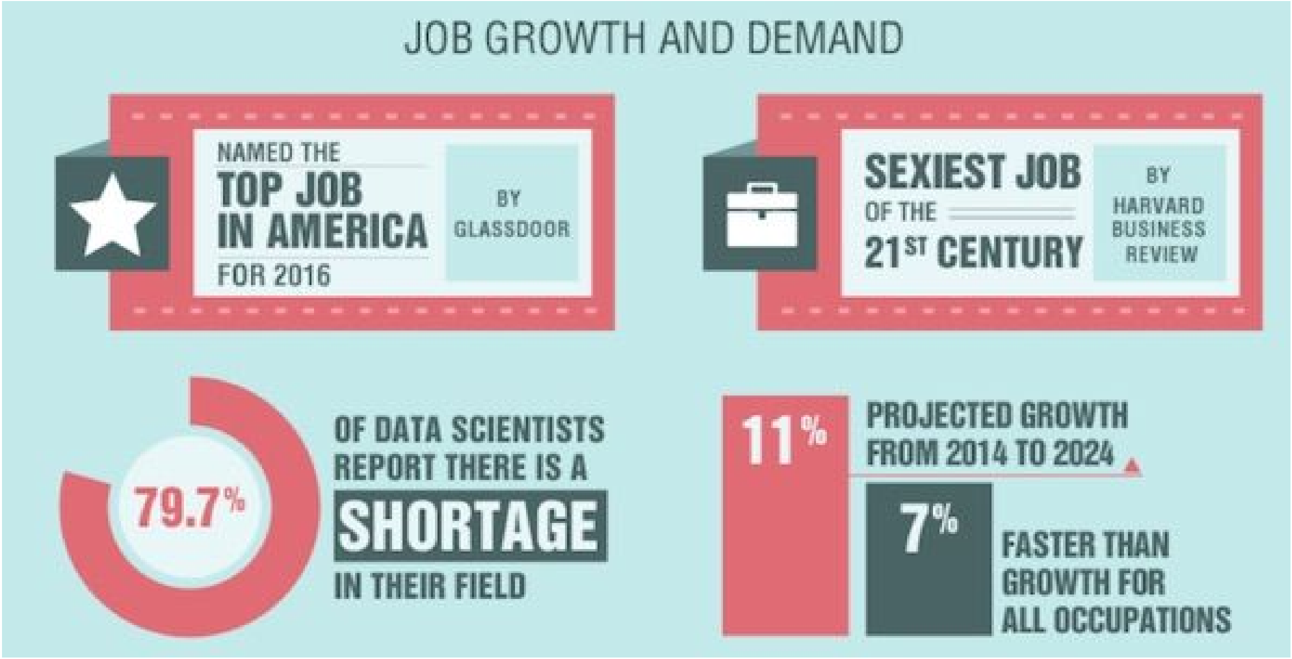
Josh Wills
@josh_wills

Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

RETWEETS 1,255 LIKES 713

6:55 PM - 3 May 2012



JOB GROWTH AND DEMAND

NAMED THE TOP JOB IN AMERICA FOR 2016 BY GLASSDOOR

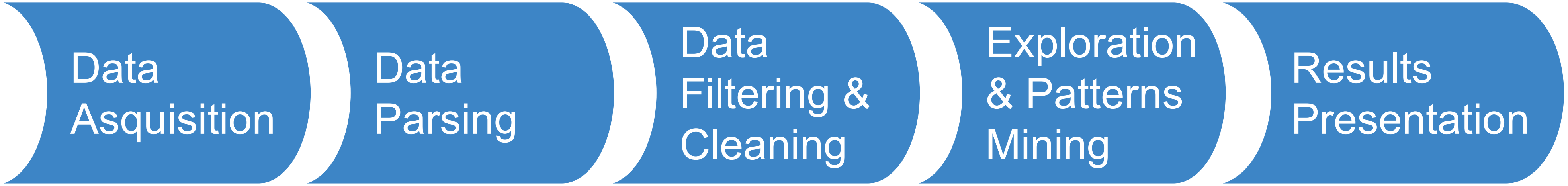
SEXIEST JOB OF THE 21ST CENTURY BY HARVARD BUSINESS REVIEW

79.7% OF DATA SCIENTISTS REPORT THERE IS A SHORTAGE IN THEIR FIELD

11% PROJECTED GROWTH FROM 2014 TO 2024

7% FASTER THAN GROWTH FOR ALL OCCUPATIONS

ЧТО



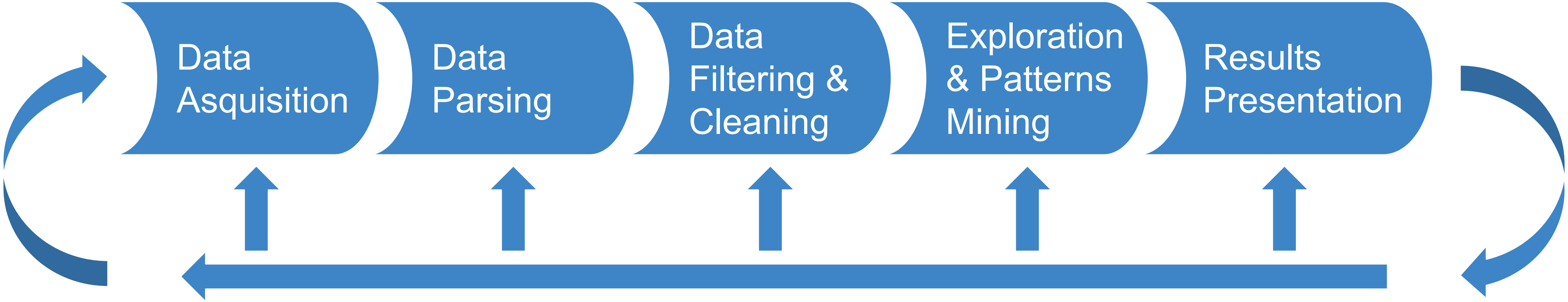
Data Asquisition

Data Parsing

Data Filtering & Cleaning

Exploration & Patterns Mining

Results Presentation



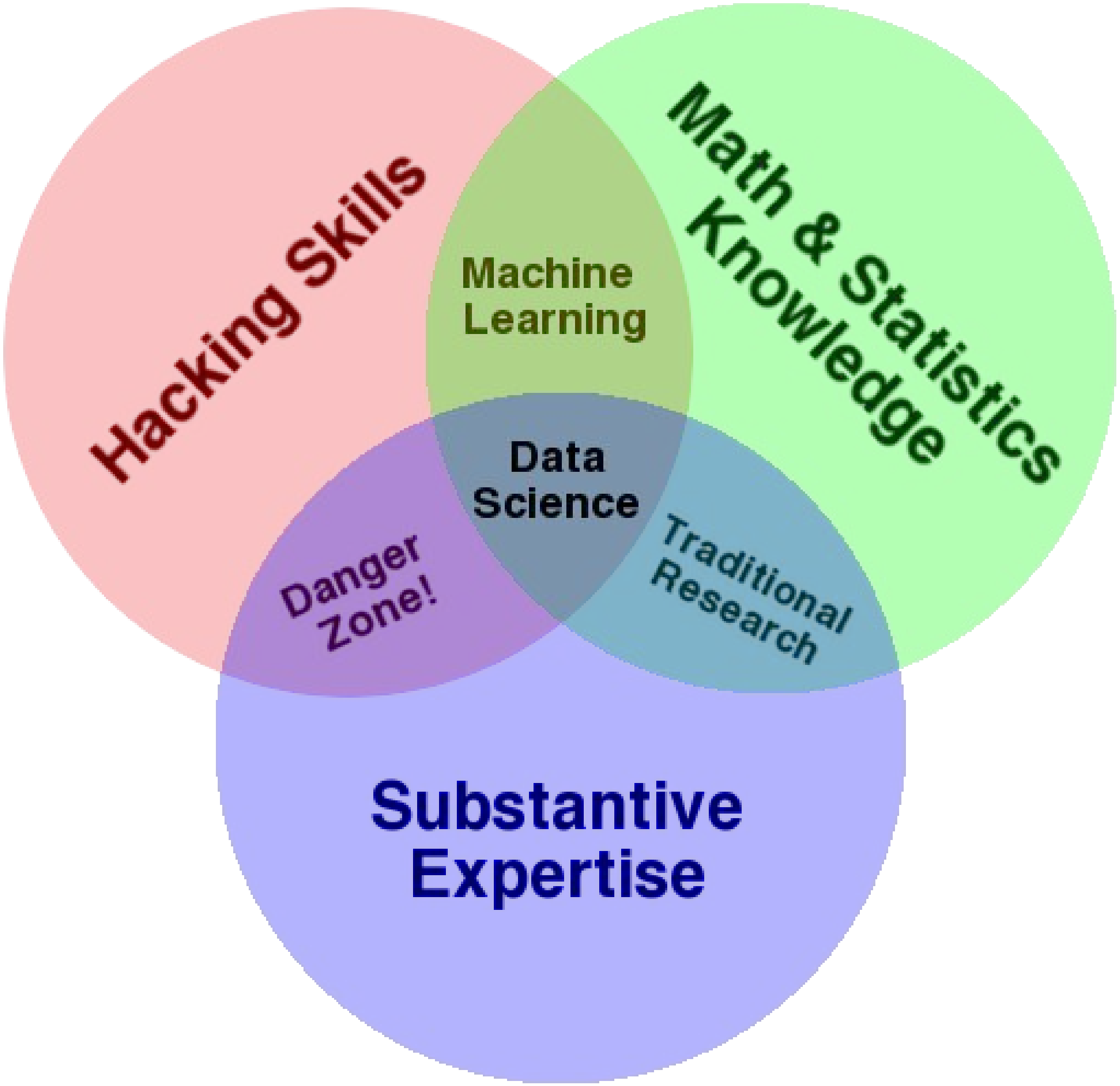
Data Asquisition

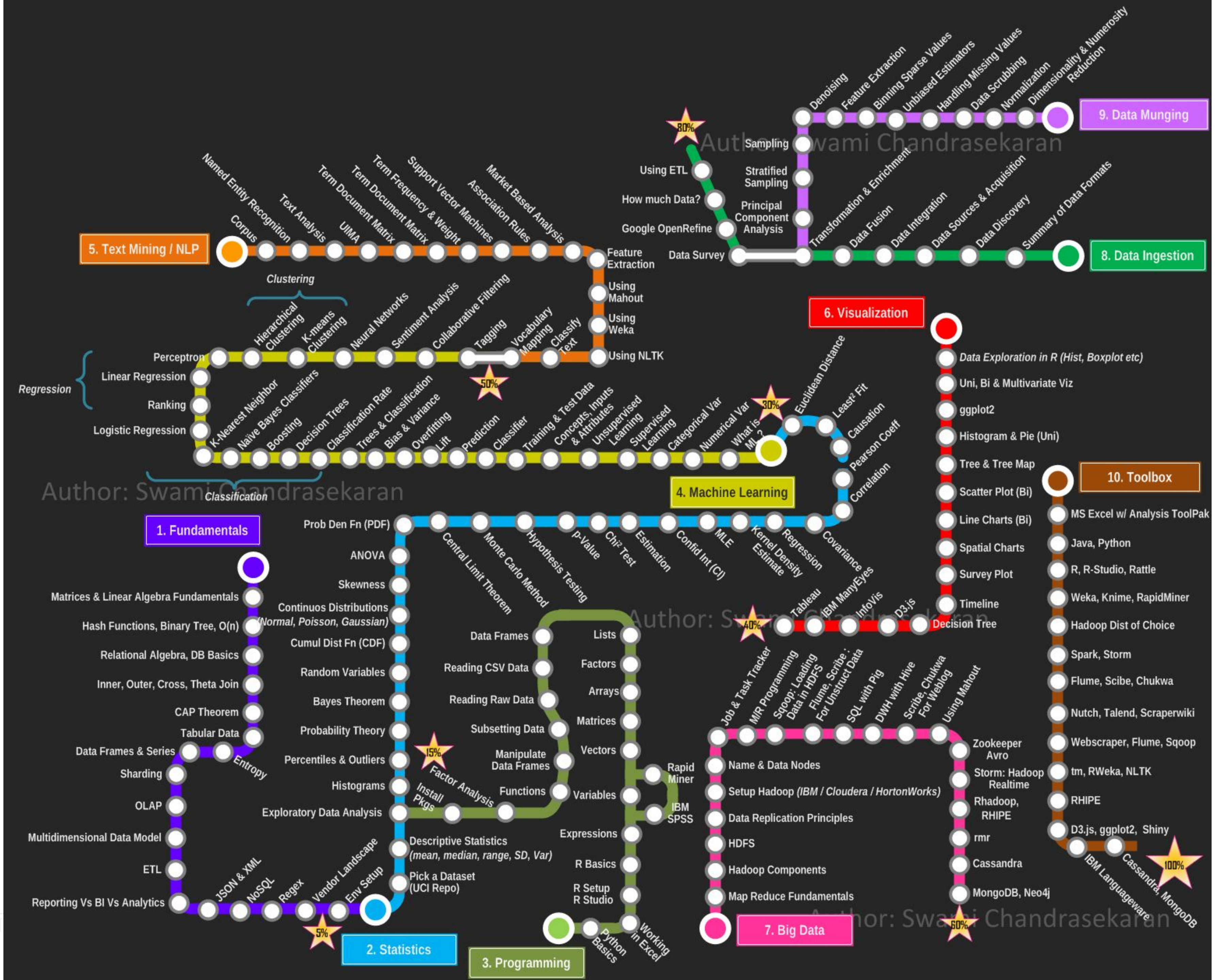
Data Parsing

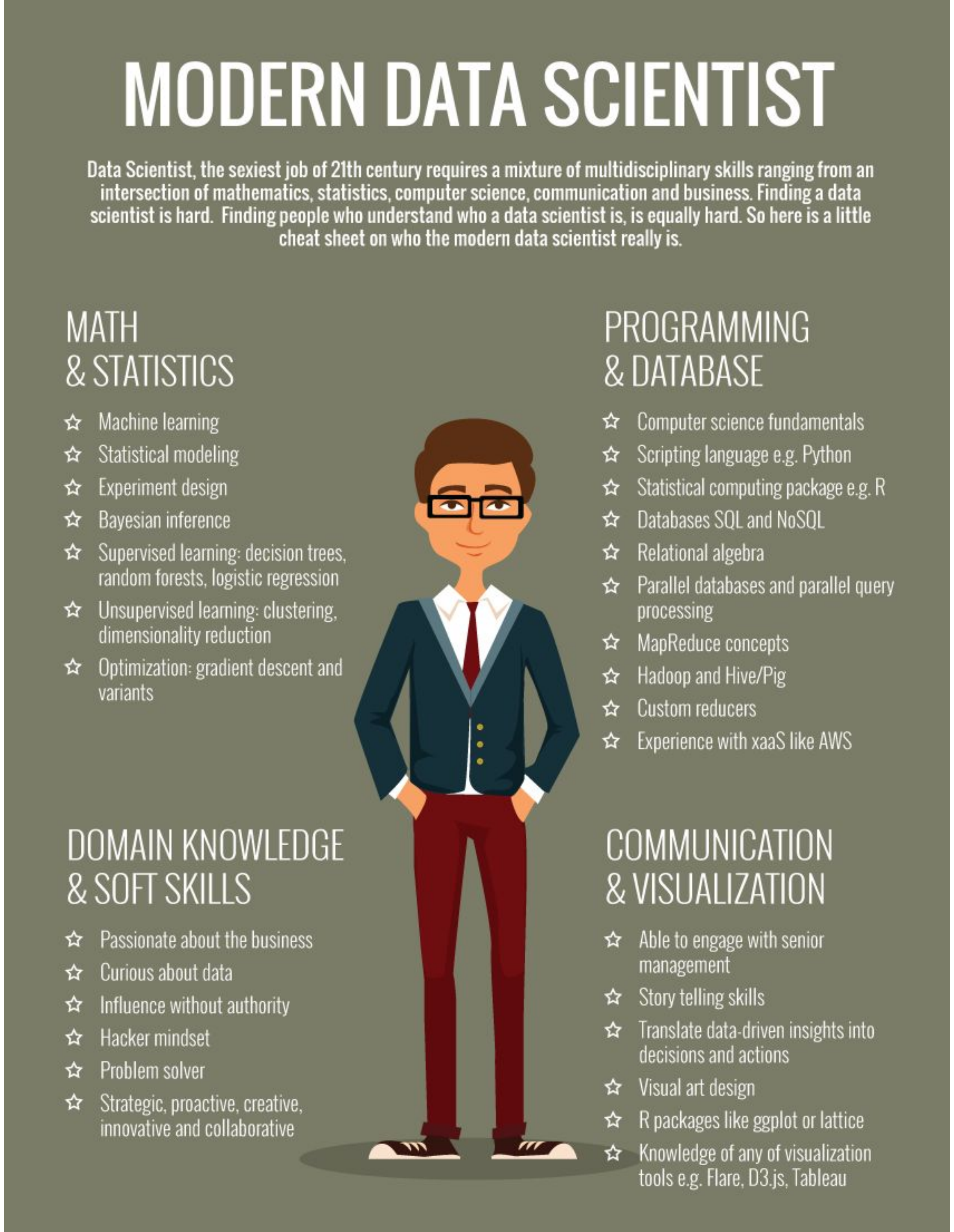
Data Filtering & Cleaning

Exploration & Patterns Mining

Results Presentation


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise


1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Entropy
Data Frames & Series
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Pick a Dataset (UCI Repo)
Descriptive Statistics (mean, median, range, SD, Var)
Exploratory Data Analysis
Histograms
Percentiles & Outliers
Probability Theory
Bayes Theorem
Random Variables
Cumul Dist Fn (CDF)
Continuos Distributions (Normal, Poisson, Gaussian)
Skewness
ANOVA
Prob Den Fn (PDF)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
Correlation
Pearson Coeff
Causation
Least² Fit
Euclidean Distance
30%
3. Programming
Python Basics
Working in Excel
R Setup R Studio
R Basics
Expressions
Variables
IBM SPSS
Rapid Miner
Vectors
Matrices
Arrays
Factors
Lists
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Install Pkgs
Factor Analysis
15%
4. Machine Learning
What is ML?
Numerical Var
Categorical Var
Supervised Learning
Unsupervised Learning
Concepts, Inputs & Attributes
Training & Test Data
Classifier
Prediction
Lift
Overfitting
Bias & Variance
Trees & Classification
Classification Rate
Decision Trees
Boosting
Naive Bayes Classifiers
K-Nearest Neighbor
Classification
Logistic Regression
Ranking
Linear Regression
Regression
Perceptron
Hierarchical Clustering
K-means Clustering
Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
50%
5. Text Mining / NLP
Vocabulary Mapping
Classify Text
Using NLTK
Using Weka
Using Mahout
Feature Extraction
Market Based Analysis
Association Rules
Support Vector Machines
Term Frequency & Weight
Term Document Matrix
UIMA
Text Analysis
Named Entity Recognition
Corpus
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
D3.js
InfoVis
IBM ManyEyes
Tableau
40%
7. Big Data
Map Reduce Fundamentals
Hadoop Components
HDFS
Data Replication Principles
Setup Hadoop (IBM / Cloudera / HortonWorks)
Name & Data Nodes
Job & Task Tracker
M/R Programming
Sqoop: Loading Data in HDFS
Flume, Scribe: For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%
8. Data Ingestion
Summary of Data Formats
Data Discovery
Data Sources & Acquisition
Data Integration
Data Fusion
Transformation & Enrichment
Data Survey
Google OpenRefine
How much Data?
Using ETL
80%
9. Data Munging
Dimensionality & Numerosity Reduction
Normalization
Data Scrubbing
Handling Missing Values
Unbiased Estimators
Binning Sparse Values
Feature Extraction
Denoising
Sampling
Stratified Sampling
Principal Component Analysis
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%



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

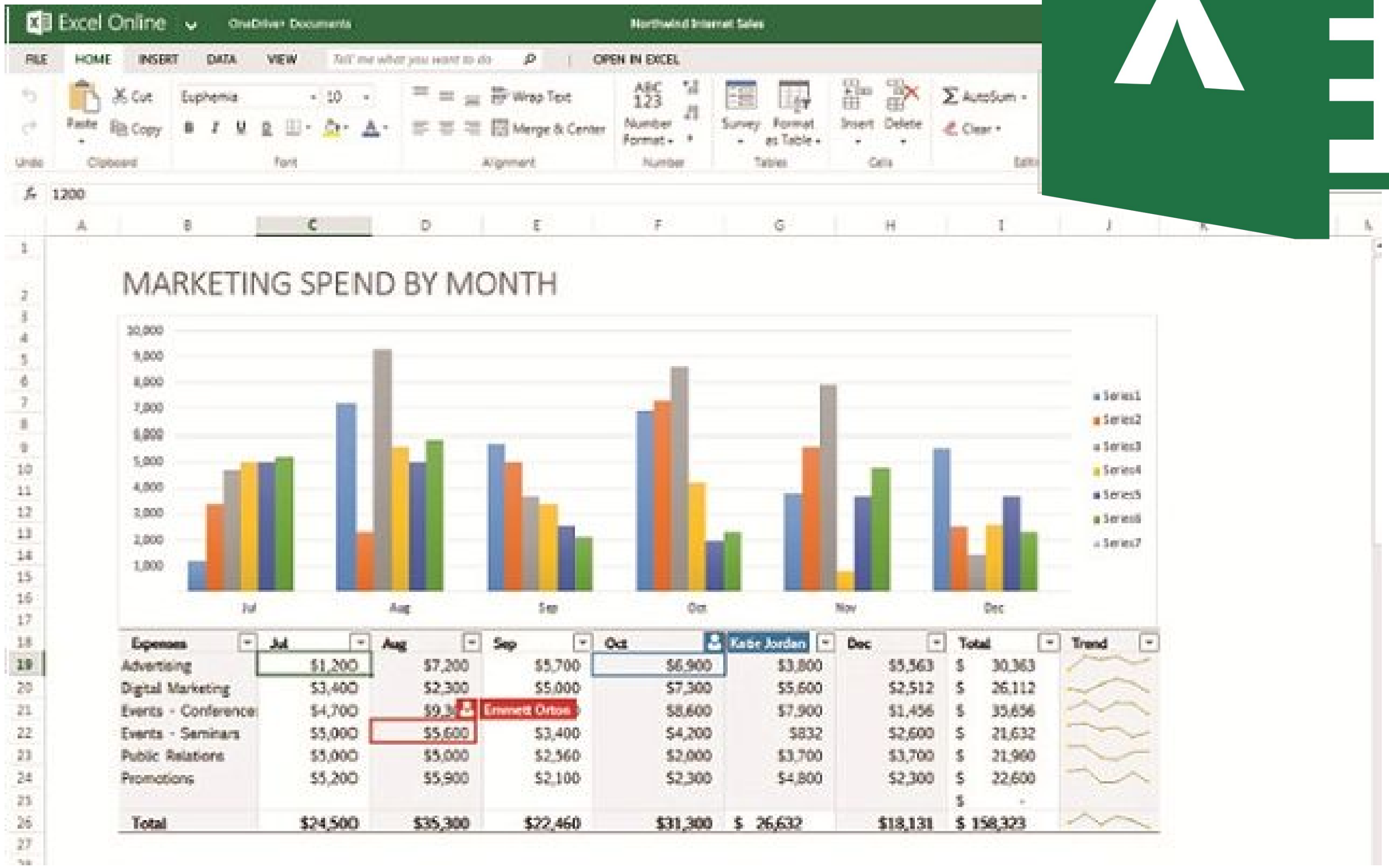
DOMAIN KNOWLEDGE & SOFT SKILLS

- 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

ЧЕМ

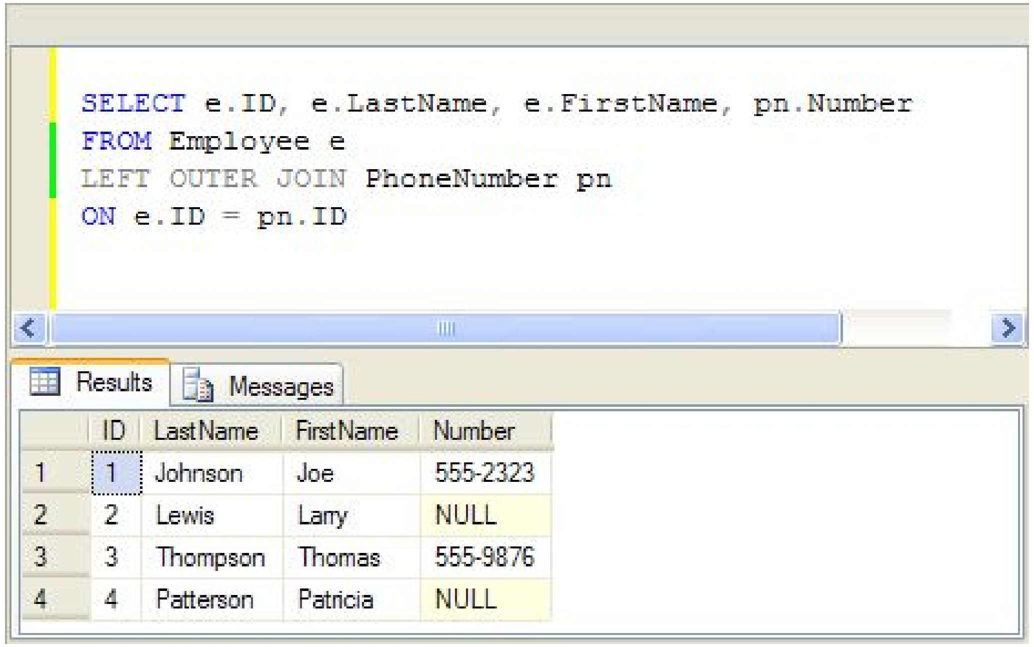


1200

MARKETING SPEND BY MONTH

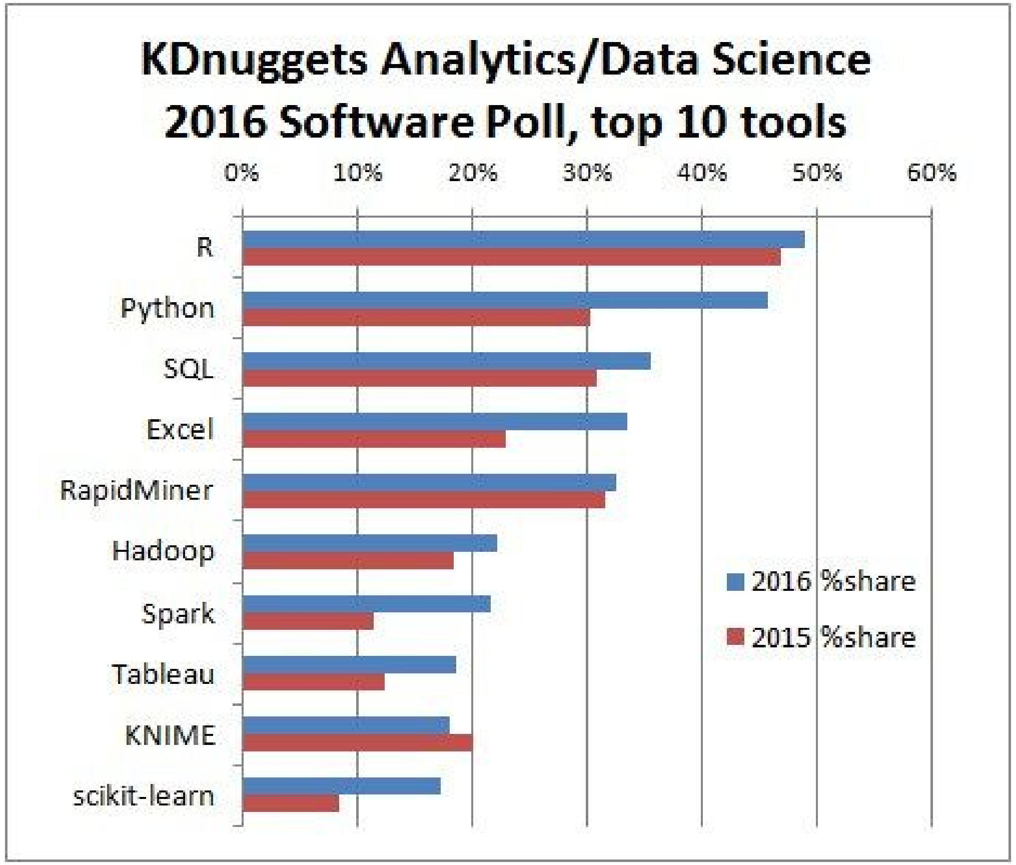
Expenses	Jul	Aug	Sep	Oct	Katie Jordan	Dec	Total	Trend
Advertising	$1,200	$7,200	$5,700	$6,900	$3,800	$5,563	$ 30,363	
Digital Marketing	$3,400	$2,300	$5,000	$7,300	$5,600	$2,512	$ 26,112	
Events - Conference	$4,700	$9,3		$8,600	$7,900	$1,456	$ 35,656	
Events - Seminars	$5,000	$5,600	$3,400	$4,200	$832	$2,600	$ 21,632	
Public Relations	$5,000	$5,000	$2,560	$2,000	$3,700	$3,700	$ 21,960	
Promotions	$5,200	$5,900	$2,100	$2,300	$4,800	$2,300	$ 22,600	
							$ -	
Total	$24,500	$35,300	$22,460	$31,300	$ 26,632	$18,131	$ 158,323	

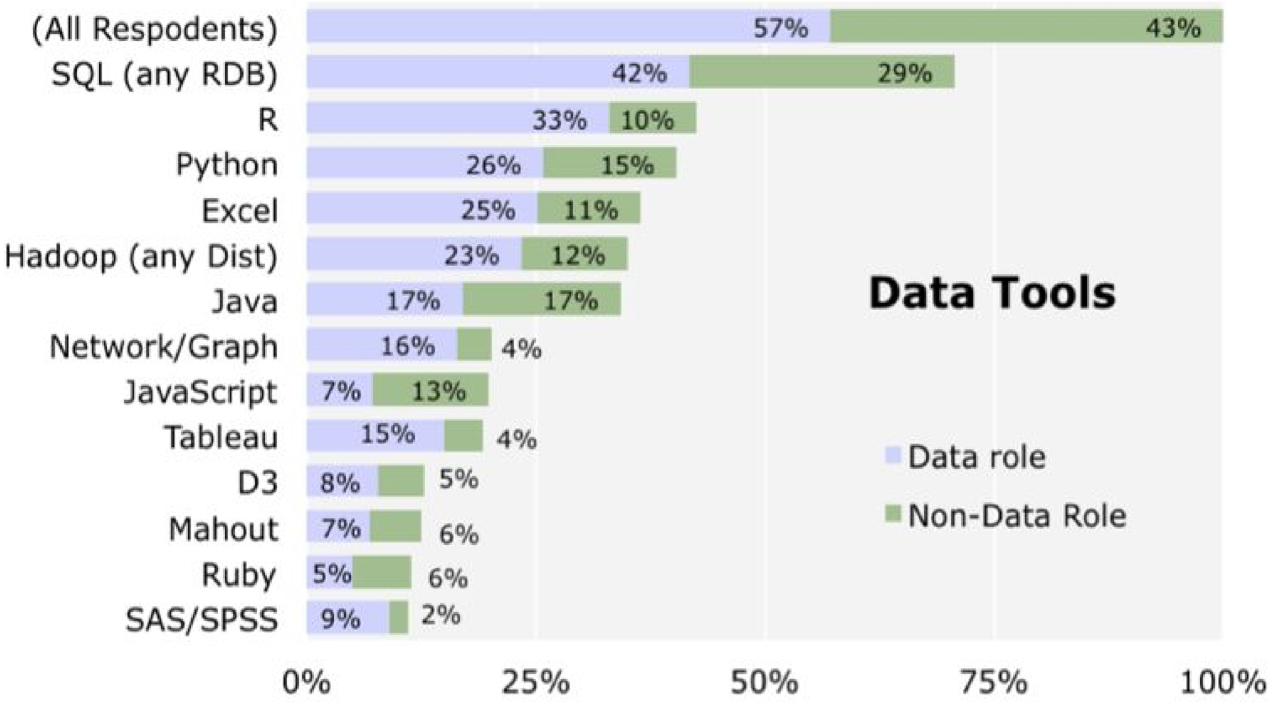
ECONTENTA



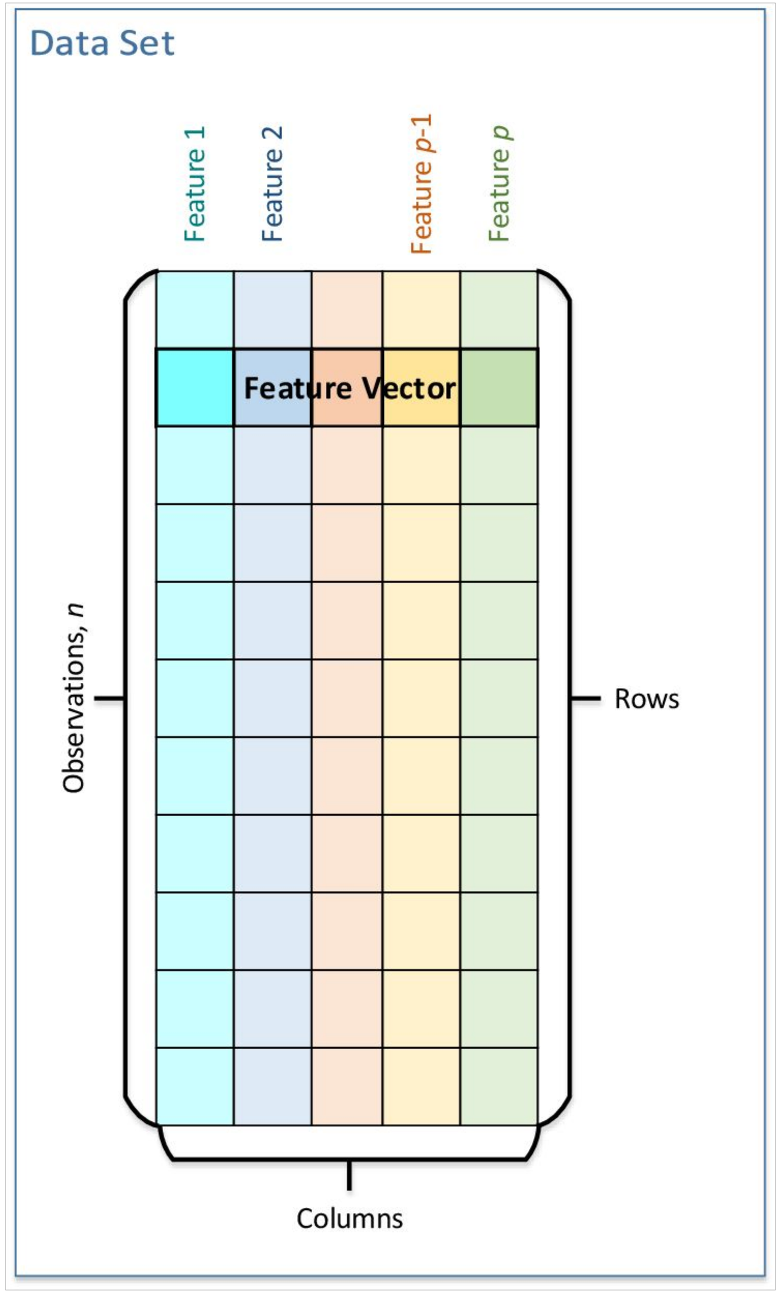
```sql
SELECT e.ID, e.LastName, e.FirstName, pn.Number
FROM Employee e
LEFT OUTER JOIN PhoneNumber pn
ON e.ID = pn.ID
```

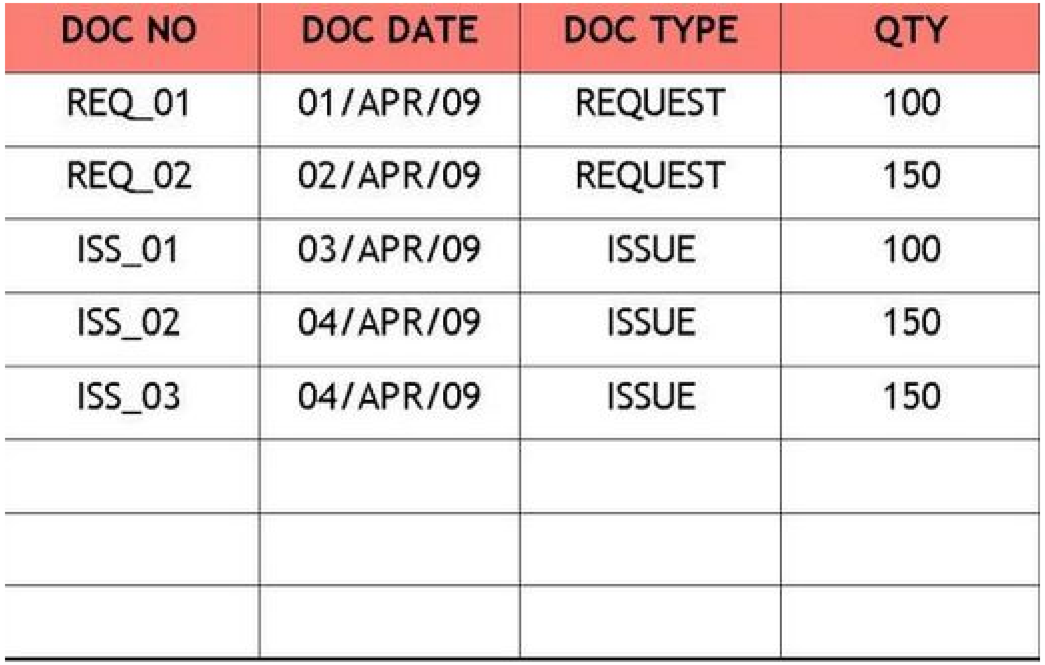
	ID	LastName	FirstName	Number
1	1	Johnson	Joe	555-2323
2	2	Lewis	Larry	NULL
3	3	Thompson	Thomas	555-9876
4	4	Patterson	Patricia	NULL


KDnuggets Analytics/Data Science
2016 Software Poll, top 10 tools
0%
10%
20%
30%
40%
50%
60%
R
Python
SQL
Excel
RapidMiner
Hadoop
Spark
Tableau
KNIME
scikit-learn
2016 %share
2015 %share


(All Respodents)
57%
43%
SQL (any RDB)
42%
29%
R
33%
10%
Python
26%
15%
Excel
25%
11%
Hadoop (any Dist)
23%
12%
Java
17%
17%
Network/Graph
16%
4%
JavaScript
7%
13%
Tableau
15%
4%
D3
8%
5%
Mahout
7%
6%
Ruby
5%
6%
SAS/SPSS
9%
2%
Data Tools
Data role
Non-Data Role
0%
25%
50%
75%
100%

ДАННЫЕ



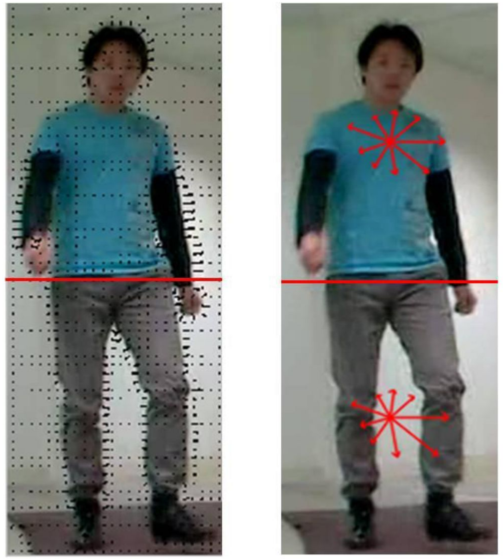


DOC NO	DOC DATE	DOC TYPE	QTY
REQ_01	01/APR/09	REQUEST	100
REQ_02	02/APR/09	REQUEST	150
ISS_01	03/APR/09	ISSUE	100
ISS_02	04/APR/09	ISSUE	150
ISS_03	04/APR/09	ISSUE	150

Features types

- Quantitative
  - Continious
    - Real
    - Natural
    - Ratio, Percentage
    - Bounded
    - Interval
  - Discrete
    - Binary
    - Nominal
    - Ordinal

- Qualitative
  - Text
  - Images
  - Sound



(a) (b)

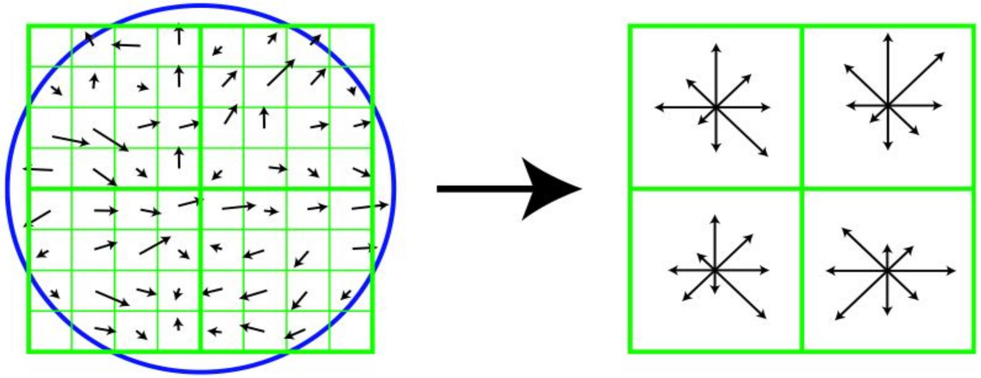
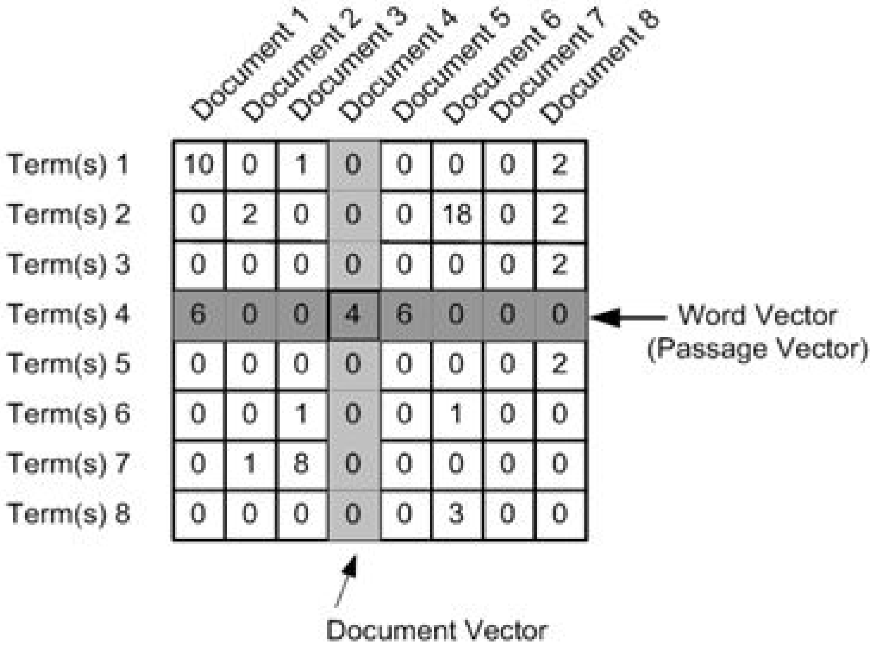


Image gradients

Keypoint descriptor



	Document 1	Document 2	Document 3	Document 4	Document 5	Document 6	Document 7	Document 8
Term(s) 1	10	0	1	0	0	0	0	2
Term(s) 2	0	2	0	0	0	18	0	2
Term(s) 3	0	0	0	0	0	0	0	2
Term(s) 4	6	0	0	4	6	0	0	0
Term(s) 5	0	0	0	0	0	0	0	2
Term(s) 6	0	0	1	0	0	1	0	0
Term(s) 7	0	1	8	0	0	0	0	0
Term(s) 8	0	0	0	0	0	3	0	0

Word Vector (Passage Vector)

Document Vector

Спасибо!

…теперь немного практики

ECONTENTA