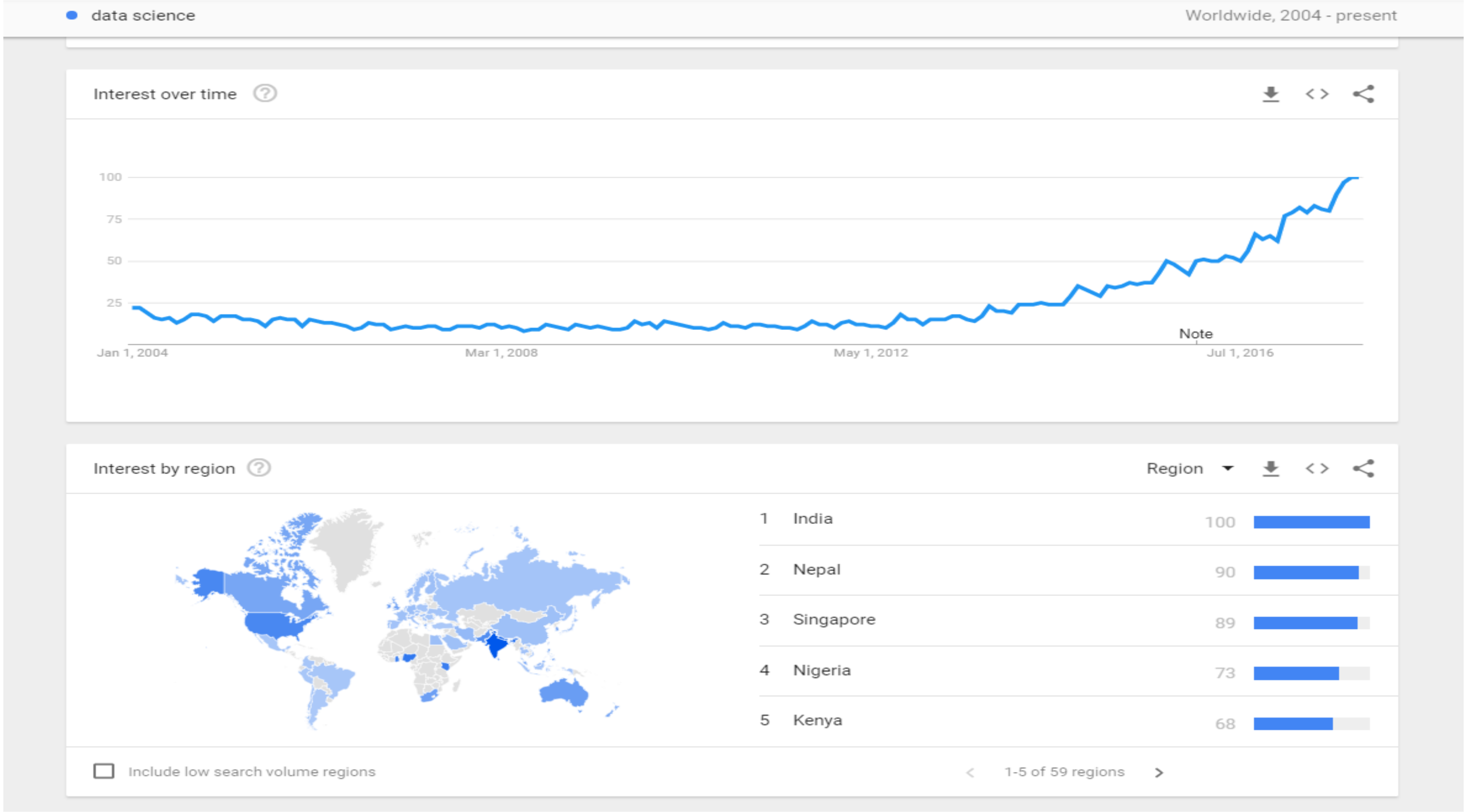
Introduction to Data Science

Hui Lin, DowDuPont

2017/11/29 @ ASA

Outline

- Slides: http://scientistcafe.com/IDS/slides/IntroToDataScience.html
- What is data science? Some definitions
- Brief history
- What questions can data science answer?
- Types of Learning
- Types of Algorithm
- Data Scientist Skill Set
- Data Science Pipeline



● data science

Worldwide, 2004 - present

Interest over time

Interest by region

Region

1	India	100
2	Nepal	90
3	Singapore	89
4	Nigeria	73
5	Kenya	68

□ Include low search volume regions

< 1-5 of 59 regions >

What is data science?

What is data science?

- Does big data matter?
- Back to 1962, John Tukey wrote in “The Future of Data Analysis”:

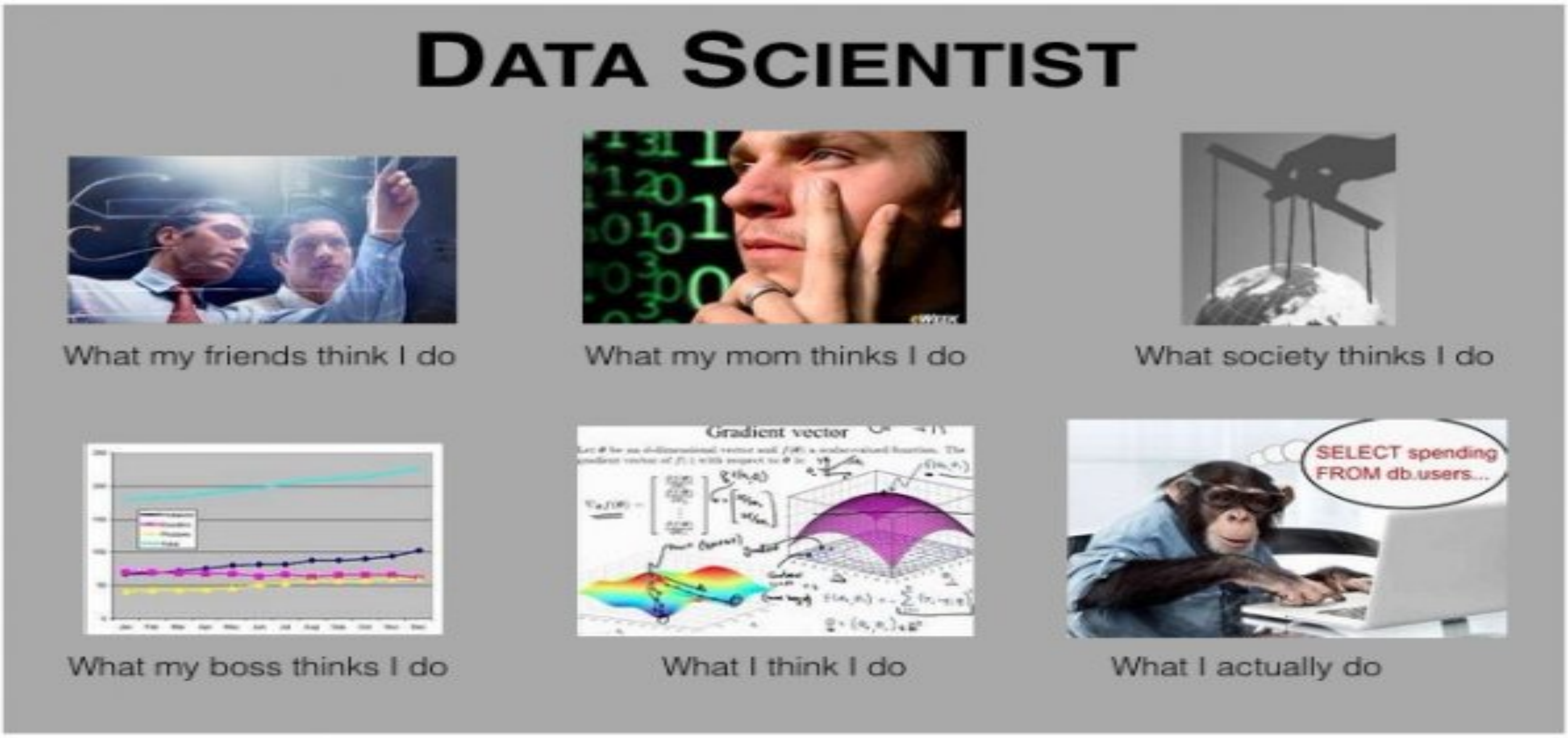
> *For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. … All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.*

What is data science?

- The website for DSI gives us an idea what Data Science is:

  "This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and interdisciplinary applications."

What is a data scientist?



What is a data scientist?

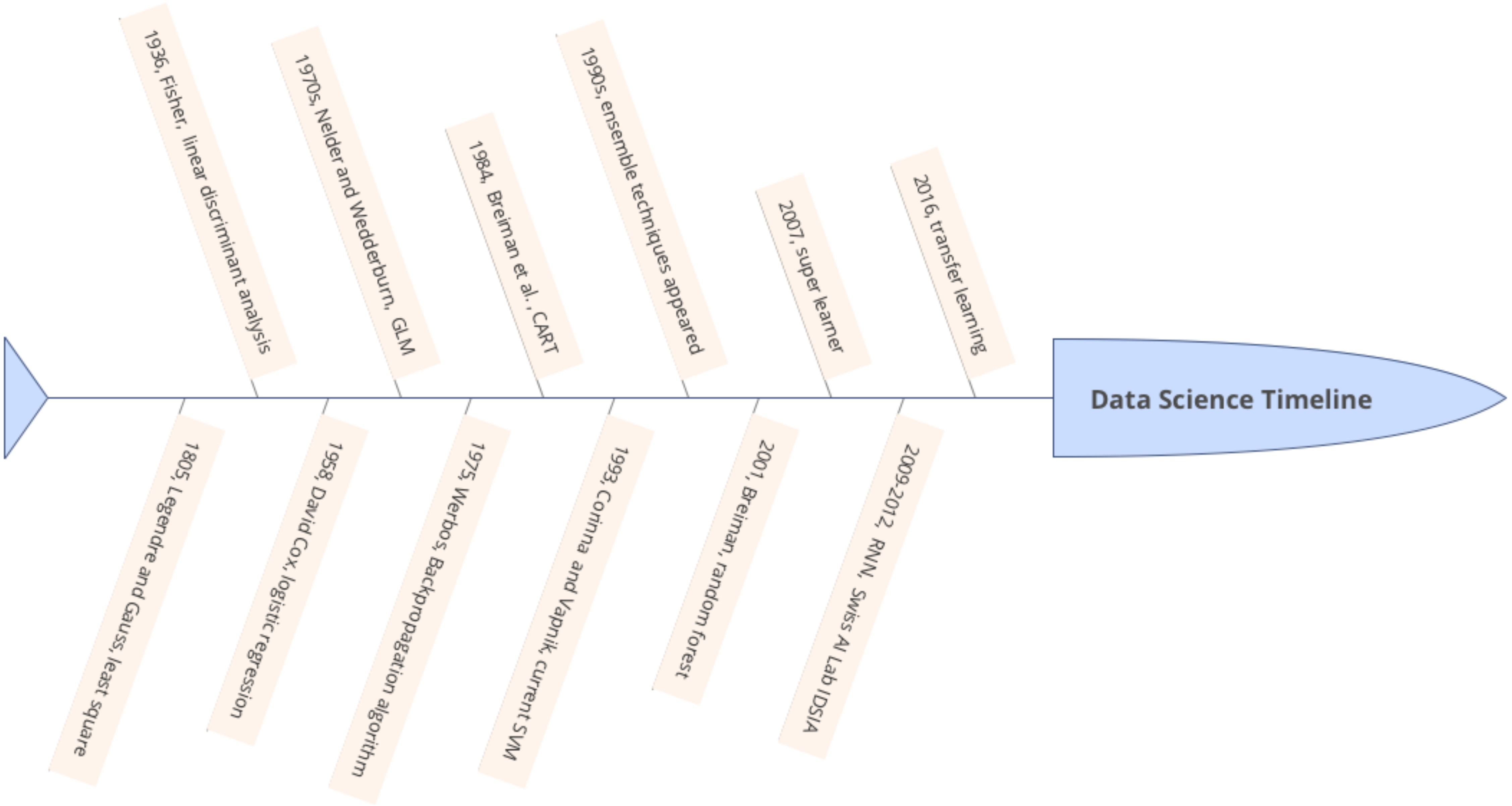
Here is a list of definitions for a “data scientist”:

- “A data scientist is a data analyst who lives in California”
- “A data scientist is someone who is better at statistics than any software engineer and better - at software engineering than any statistician.”
- “A data scientist is a statistician who lives in San Francisco.”
- “Data Science is statistics on a Mac.”

%&^%$*(^)…..

- You know it when you see it.

Brief History



Data Science Timeline

1805, Legendre and Gauss, least square

1936, Fisher, linear discriminant analysis

1958, David Cox, logistic regression

1970s, Nelder and Wedderburn, GLM

1975, Werbos, Backpropagation algorithm

1984, Breiman et al. , CART

1993, Corinna and Vapnik, current SVM

1990s, ensemble techniques appeared

2001, Breiman, random forest

2007, super learner

2009-2012, RNN, Swiss AI Lab IDSIA

2016, transfer learning

Driving Forces

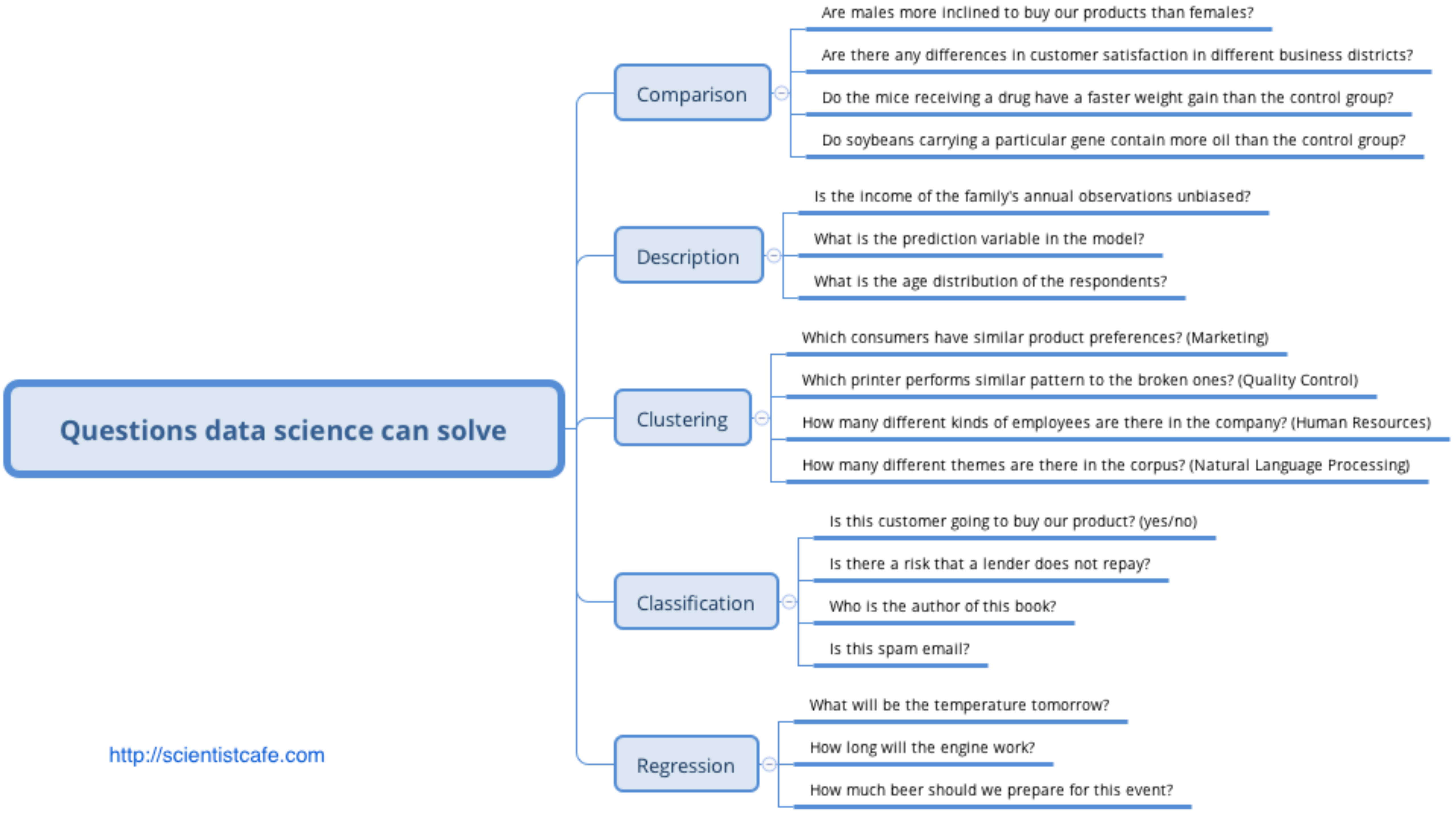
- John Tukey identified 4 forces driving data analysis (there was no "data science" then):

    1. The formal theories of math/stat
    2. Acceleration developments in computers and display devices
    3. The challenge, in many fields, of more and ever larger bodies of data
    4. The emphasis on quantification in an ever wider variety of disciplines

What questions can data science answer?

- Specific
  1. *How can we increase sales?*
  2. *Does the January campaign on product X increase the amount of purchase from our 2017 retained customers?*
- Data
  1. *Representative*
  2. *Relevant*
  3. *Quality*

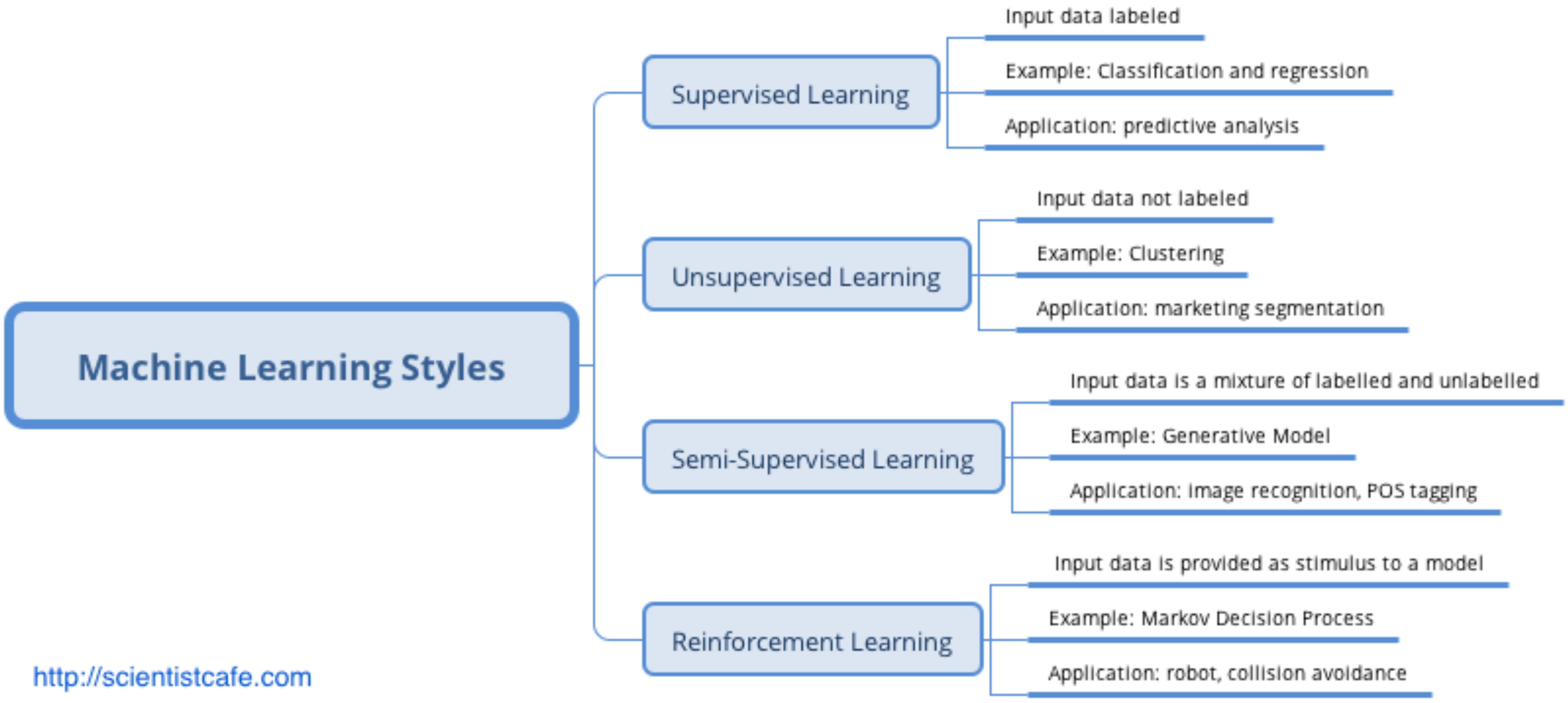
Types of Questions



- Questions data science can solve
  - Comparison
    - Are males more inclined to buy our products than females?
    - Are there any differences in customer satisfaction in different business districts?
    - Do the mice receiving a drug have a faster weight gain than the control group?
    - Do soybeans carrying a particular gene contain more oil than the control group?
  - Description
    - Is the income of the family's annual observations unbiased?
    - What is the prediction variable in the model?
    - What is the age distribution of the respondents?
  - Clustering
    - Which consumers have similar product preferences? (Marketing)
    - Which printer performs similar pattern to the broken ones? (Quality Control)
    - How many different kinds of employees are there in the company? (Human Resources)
    - How many different themes are there in the corpus? (Natural Language Processing)
  - Classification
    - Is this customer going to buy our product? (yes/no)
    - Is there a risk that a lender does not repay?
    - Who is the author of this book?
    - Is this spam email?
  - Regression
    - What will be the temperature tomorrow?
    - How long will the engine work?
    - How much beer should we prepare for this event?

http://scientistcafe.com

Types of Learning



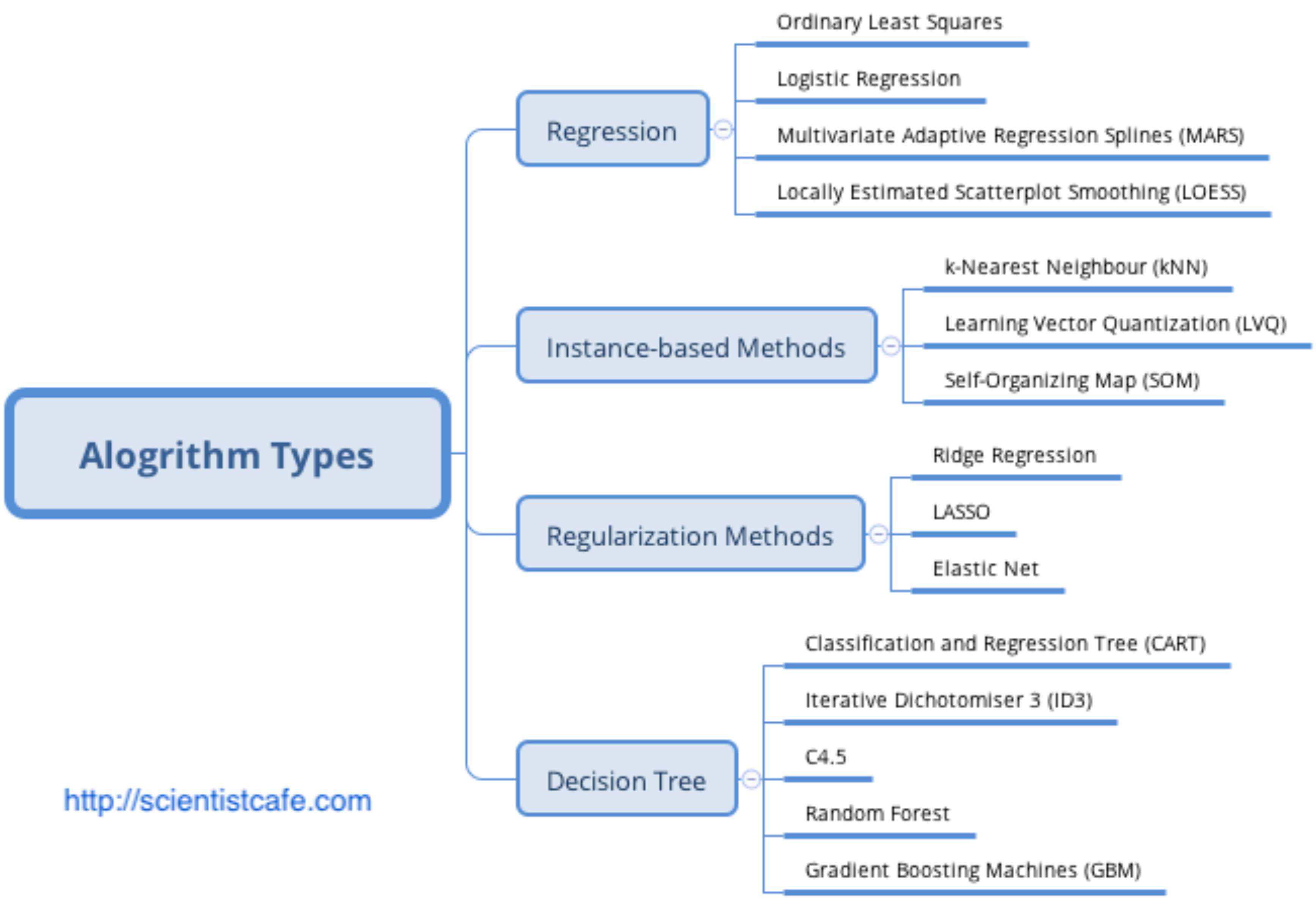
Machine Learning Styles

- Supervised Learning
  - Input data labeled
  - Example: Classification and regression
  - Application: predictive analysis
- Unsupervised Learning
  - Input data not labeled
  - Example: Clustering
  - Application: marketing segmentation
- Semi-Supervised Learning
  - Input data is a mixture of labelled and unlabelled
  - Example: Generative Model
  - Application: image recognition, POS tagging
- Reinforcement Learning
  - Input data is provided as stimulus to a model
  - Example: Markov Decision Process
  - Application: robot, collision avoidance

http://scientistcafe.com

Types of Algorithm (1)

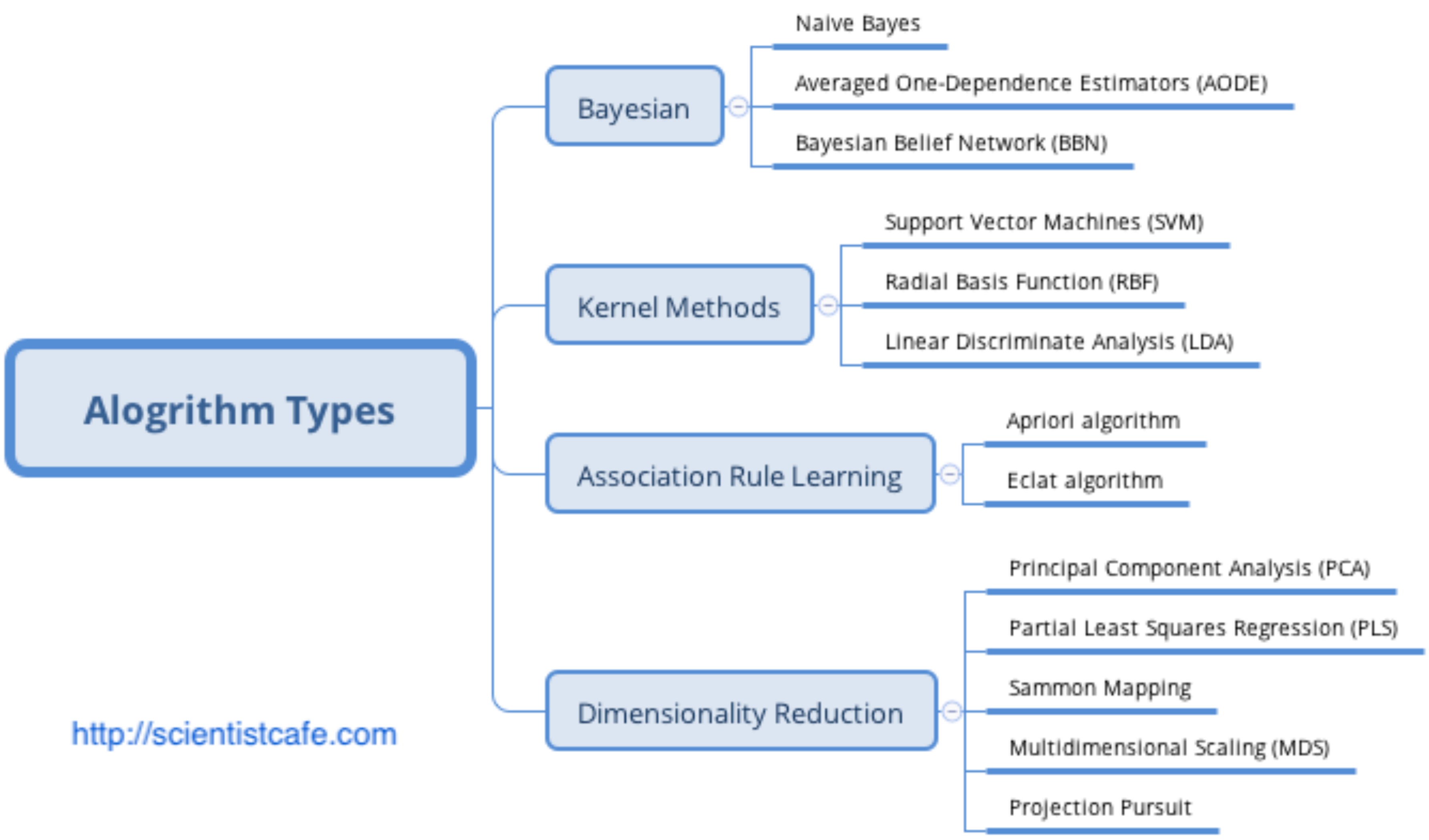
- http://scientistcafe.com/2017/07/08/MachineLearningAI.html



Alogrithm Types
Regression
Ordinary Least Squares
Logistic Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Instance-based Methods
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Regularization Methods
Ridge Regression
LASSO
Elastic Net
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
Random Forest
Gradient Boosting Machines (GBM)
http://scientistcafe.com

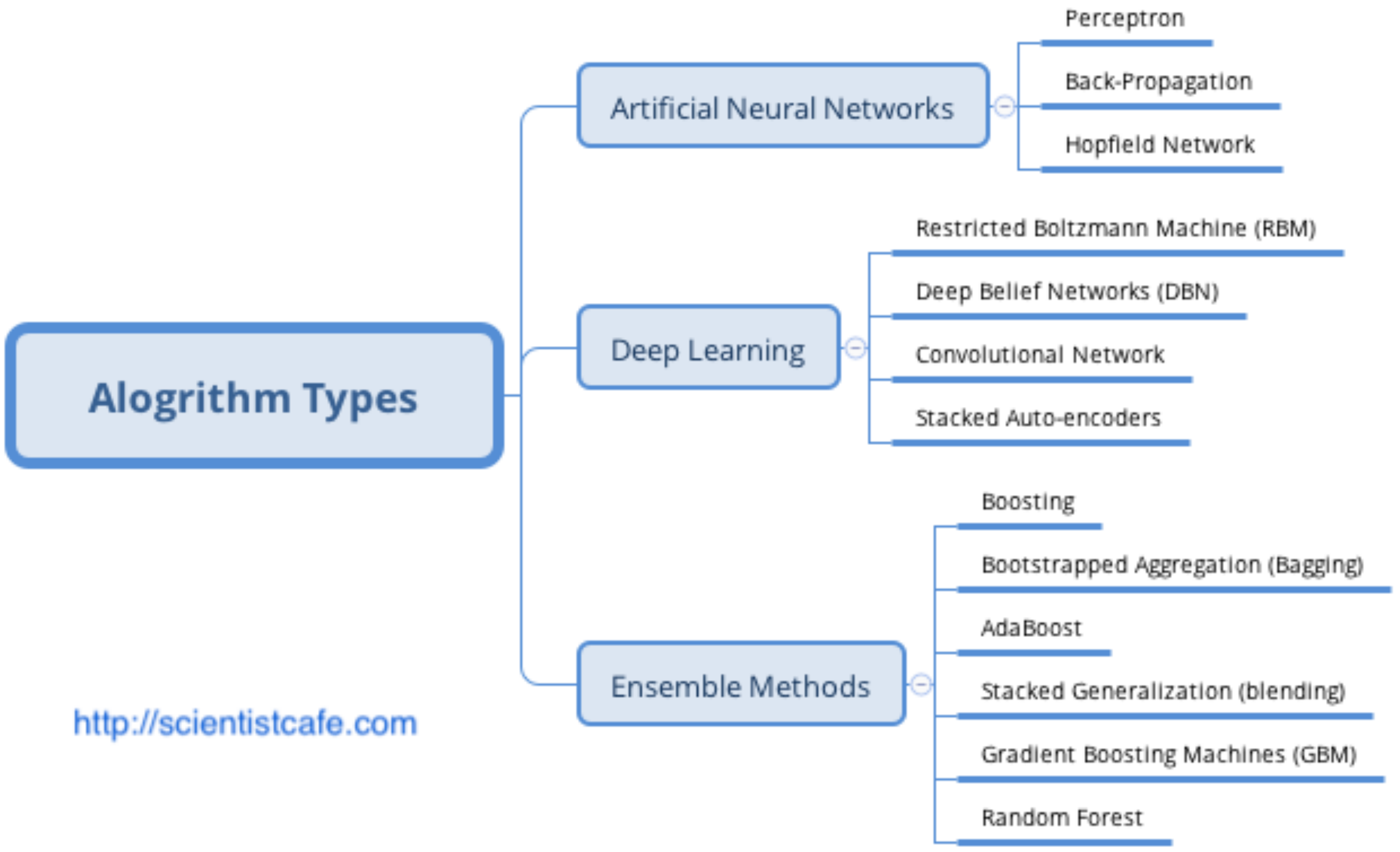
Types of Algorithm (2)

- http://scientistcafe.com/2017/07/08/MachineLearningAI.html

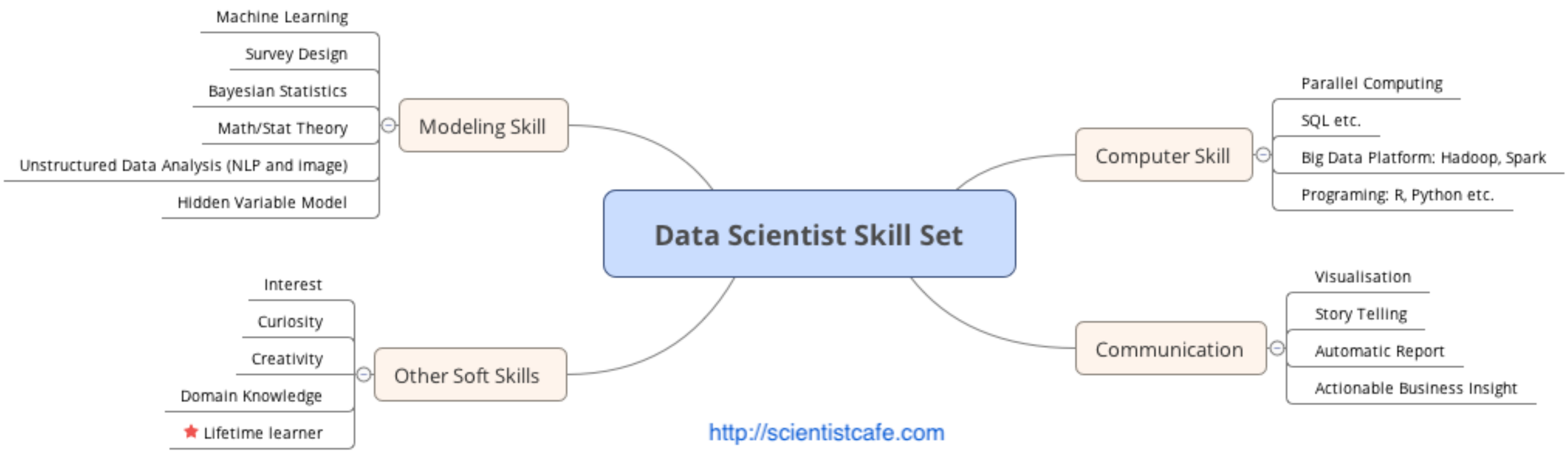

Alogrithm Types
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Kernel Methods
Support Vector Machines (SVM)
Radial Basis Function (RBF)
Linear Discriminate Analysis (LDA)
Association Rule Learning
Apriori algorithm
Eclat algorithm
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLS)
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
http://scientistcafe.com

Types of Algorithm (3)

- http://scientistcafe.com/2017/07/08/MachineLearningAI.html


Alogrithm Types
Artificial Neural Networks
Perceptron
Back-Propagation
Hopfield Network
Deep Learning
Restricted Boltzmann Machine (RBM)
Deep Belief Networks (DBN)
Convolutional Network
Stacked Auto-encoders
Ensemble Methods
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (blending)
Gradient Boosting Machines (GBM)
Random Forest
http://scientistcafe.com

Data Scientist Skill Set

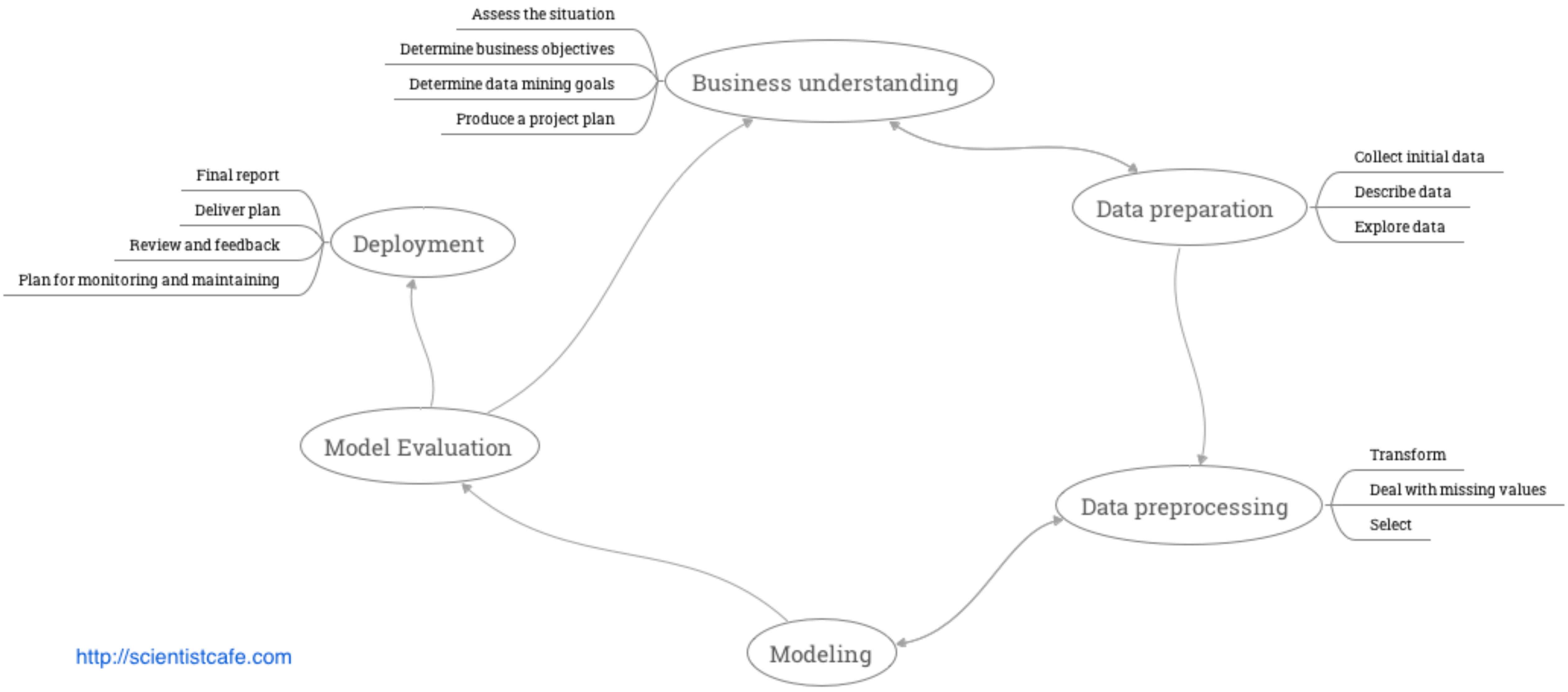


Data Scientist Skill Set

- Modeling Skill
  - Machine Learning
  - Survey Design
  - Bayesian Statistics
  - Math/Stat Theory
  - Unstructured Data Analysis (NLP and image)
  - Hidden Variable Model
- Other Soft Skills
  - Interest
  - Curiosity
  - Creativity
  - Domain Knowledge
  - ★ Lifetime learner
- Computer Skill
  - Parallel Computing
  - SQL etc.
  - Big Data Platform: Hadoop, Spark
  - Programing: R, Python etc.
- Communication
  - Visualisation
  - Story Telling
  - Automatic Report
  - Actionable Business Insight

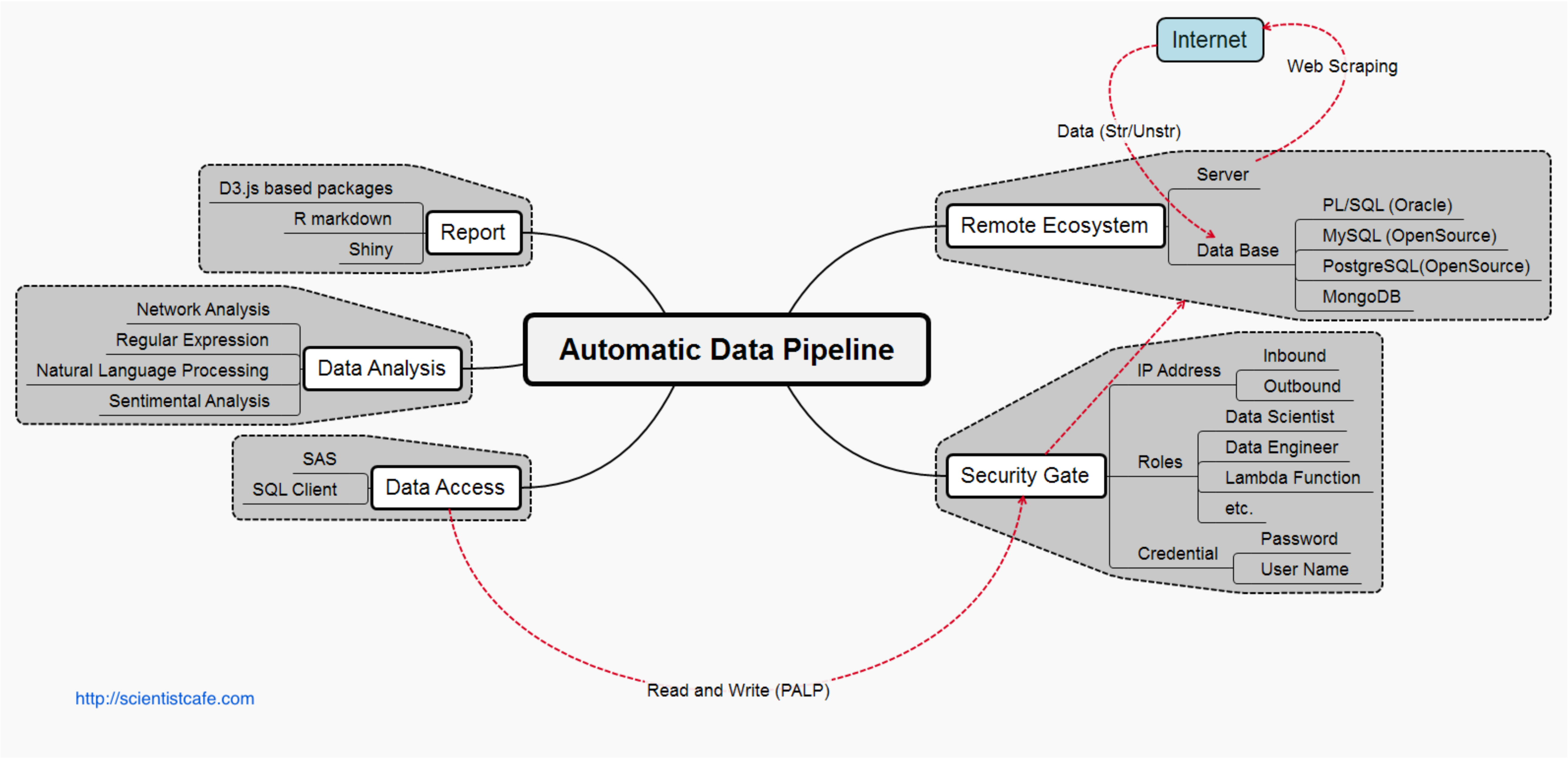
http://scientistcafe.com

General Process



http://scientistcafe.com

Automatic Data Science Pipeline



- Automatic Data Pipeline
  - Report
    - D3.js based packages
    - R markdown
    - Shiny
  - Data Analysis
    - Network Analysis
    - Regular Expression
    - Natural Language Processing
    - Sentimental Analysis
  - Data Access
    - SAS
    - SQL Client
  - Remote Ecosystem
    - Server
    - Data Base
      - PL/SQL (Oracle)
      - MySQL (OpenSource)
      - PostgreSQL(OpenSource)
      - MongoDB
  - Security Gate
    - IP Address
      - Inbound
      - Outbound
    - Roles
      - Data Scientist
      - Data Engineer
      - Lambda Function
      - etc.
    - Credential
      - Password
      - User Name

Internet

Web Scraping

Data (Str/Unstr)

Read and Write (PALP)

http://scientistcafe.com

Some links

- Types of Machine Learning Algorithm

- Online books:
    - *The Elements of Statistical Learning*
    - *An Introduction to Statistical Learning*
    - *Introduction to Data Science(still writing)*

- Hard copy books:
    - *Applied Predictive Modeling*
    - *R for Marketing Research and Analytics*
    - *套路！机器学习*

- Awesome-Data-Science-Materials