STAT3612 Lecture 1

Introduction to Statistical Machine Learning

Dr. Aijun Zhang

1 September 2020

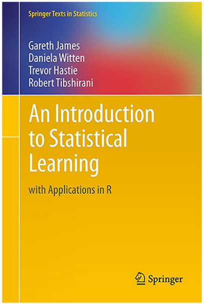
Department of 統計及精算學系
Statistics & Actuarial Science

Table of Contents

STAT3612 Course Outline

- Course website: `https://github.com/ajzhanghku/Stat3612`
- Click to view the syllabus (PDF) ...
- Check out the tentative class schedule ...
- We need to fix the tutorial hours with Yuyang and Yifeng ...

Reference Books

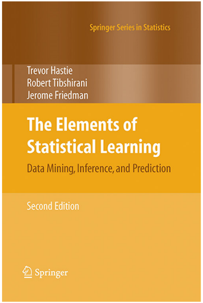

Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer


Springer Series in Statistics

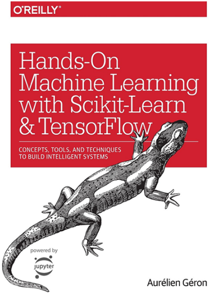
Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

Springer

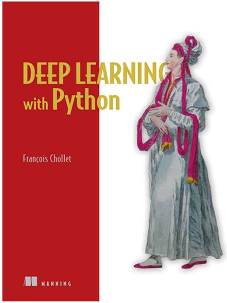

O'REILLY

Hands-On Machine Learning with Scikit-Learn & TensorFlow

CONCEPTS, TOOLS, AND TECHNIQUES TO BUILD INTELLIGENT SYSTEMS

powered by jupyter

Aurélien Géron

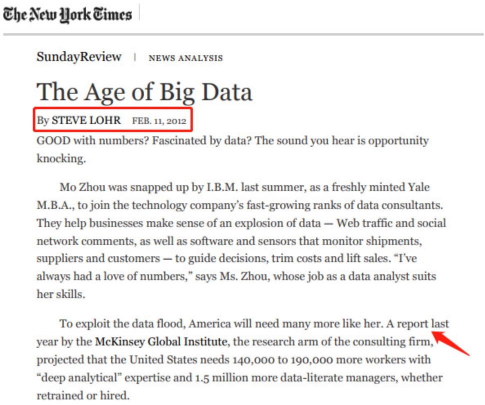

DEEP LEARNING with Python

François Chollet

MANNING

Table of Contents

The Age of Big Data



The New York Times

SundayReview | NEWS ANALYSIS

The Age of Big Data

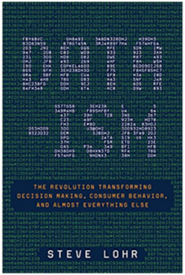
By STEVE LOHR FEB. 11, 2012

GOOD with numbers? Fascinated by data? The sound you hear is opportunity knocking.

Mo Zhou was snapped up by I.B.M. last summer, as a freshly minted Yale M.B.A., to join the technology company's fast-growing ranks of data consultants. They help businesses make sense of an explosion of data — Web traffic and social network comments, as well as software and sensors that monitor shipments, suppliers and customers — to guide decisions, trim costs and lift sales. "I've always had a love of numbers," says Ms. Zhou, whose job as a data analyst suits her skills.

To exploit the data flood, America will need many more like her. A report last year by the McKinsey Global Institute, the research arm of the consulting firm, projected that the United States needs 140,000 to 190,000 more workers with "deep analytical" expertise and 1.5 million more data-literate managers, whether retrained or hired.

Steve Lohr, New York Times Reporter
in 2013 Pulitzer Prize Winning Team


DATA-ISM
THE REVOLUTION TRANSFORMING DECISION MAKING, CONSUMER BEHAVIOR, AND ALMOST EVERYTHING ELSE
STEVE LOHR

HarperCollins, 2015

Read the complete article at nytimes.com

McKinsey 2011 Report

McKinsey&Company

McKinsey Global Institute

May 2011

Big data: The next frontier for innovation, competition, and productivity

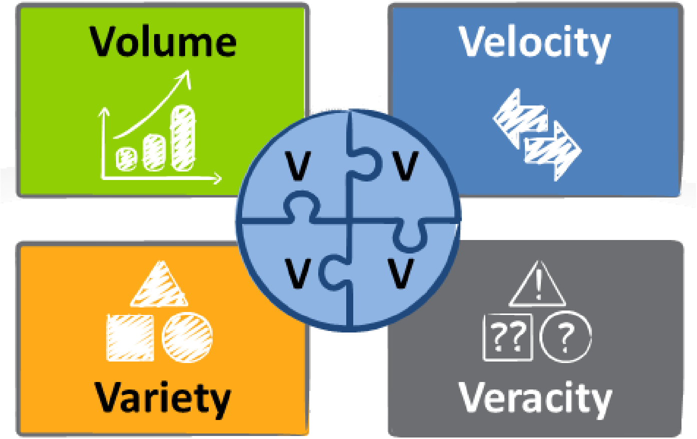
- In 2011, McKinsey Global Institute claimed that:

  > *"Big data" refers to datasets whose size is beyond the ability of typical database software to caputure, store, manage, and analyze.*
  >
  > *By 2018, the United States needs 140,000 to 190,000 more workers with "deep analytical" expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data.*

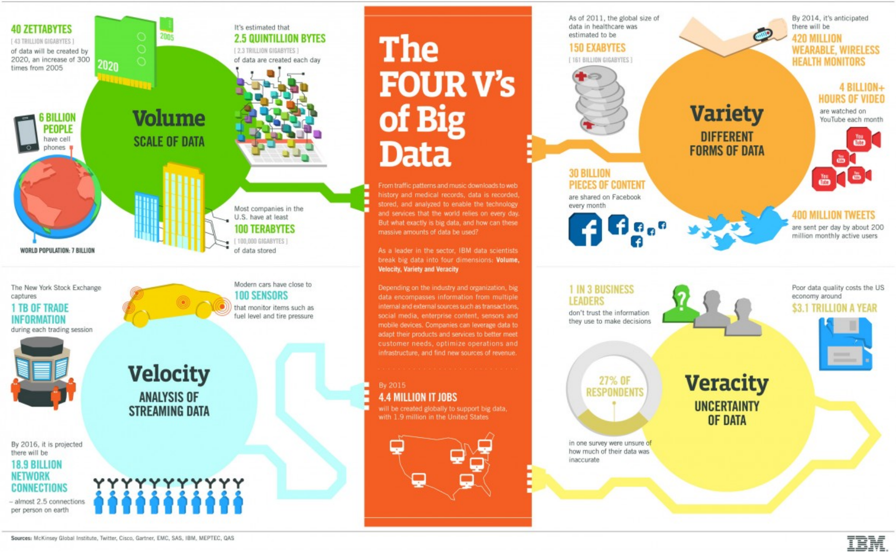
- Download the full report at mckinsey.com

Four V's of Big Data

Source: haritbigdata.wordpress.com



Four V's of Big Data (IBM version)



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE have cell phones

It's estimated that **2.5 QUINTILLION BYTES** [2.3 TRILLION GIGABYTES] of data are created each day

Most companies in the U.S. have at least **100 TERABYTES** [100,000 GIGABYTES] of data stored

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** [161 BILLION GIGABYTES]

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

30 BILLION PIECES OF CONTENT are shared on Facebook every month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

Veracity

UNCERTAINTY OF DATA

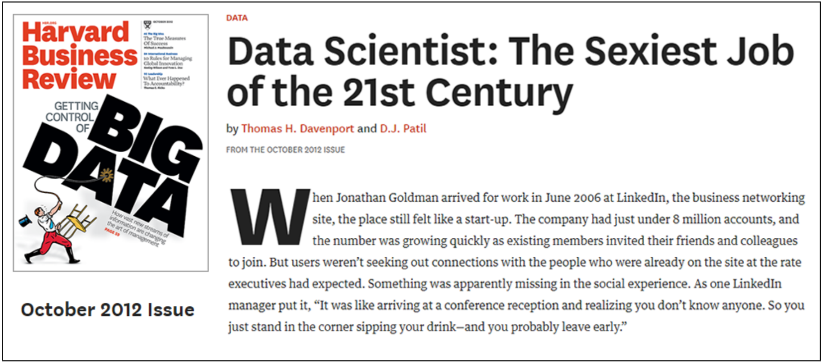
1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around **$3.1 TRILLION A YEAR**

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Data Scientist


Harvard Business Review
GETTING CONTROL OF BIG DATA
How vast new streams of information are changing the art of management

October 2012 Issue

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

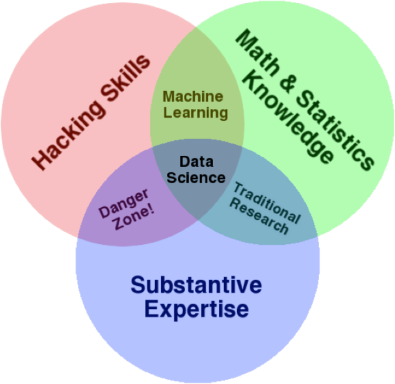
FROM THE OCTOBER 2012 ISSUE

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

The trending job markets; search LinkedIn

Three specialties are required for a data scientist: math & statistics, computer science, and domain expertise.

Data Science Venn Diagram



Created by Drew Conway (2010), Click Here

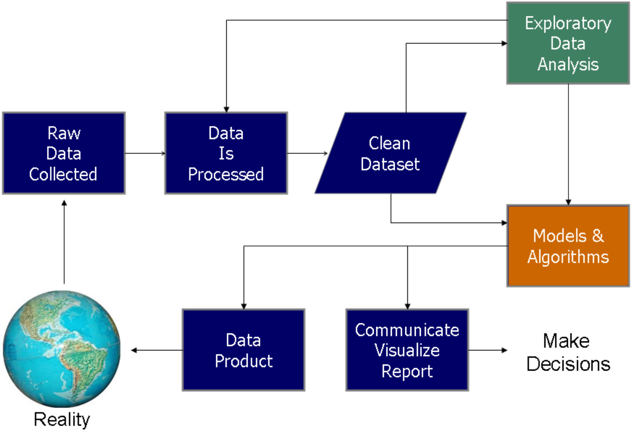
Data Science vs. Statistics

© morganimation - Fotolia.com

Data Science vs. Statistics

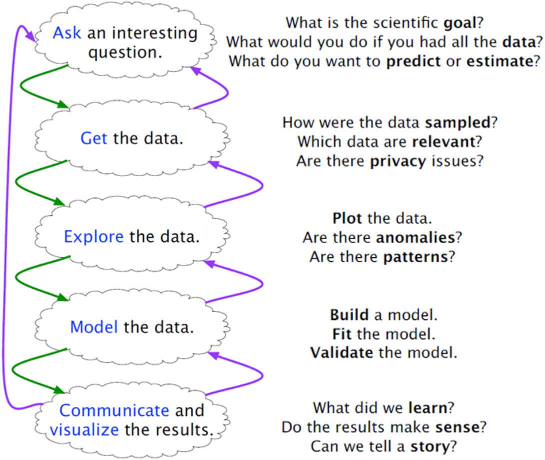
Linear
Approximation
Selection
Matrices
Prediction
Algorithm
Statistics
Distribution
Regression
Nonparametric
Rate
MonteCarlo
Empirical
Function
Estimation
Scale
poisson
Correlation
Data
Ratio
Nearest
Risk
Estimator
Condition
Matrix
Covariance
Simulation
Partial
Point
Moment
Multivariate
Operator
semiparametric
Parametric
Random
Differential
Hazard
Mixture
Solution
Local
Population
Parameter
Value
Density
General
Bayes
Models
Spectral
Order
Study
High
Design
Effects
Mixed
Numerical
Square
Series
Least
Structure
Spatial
Gaussian
Latent
Markov
Nonlinear
Problem
Principal
Probability
Asymptotic
Functional
Large
Kernel
Test
Convergence
Finite
Maximum
Clustering
Learning
Space
Process
Model
Time
Sampling
Gene
Optimal
Theory
Error
Tail
Variable
Equation
Conditional
bayesian
Inference
Normal
Mining
Information
Stochastic
Analysis
Classification
Likelihood
Method

Data Science Workflow v1



Statistical modeling & machine learning lies in “Models & Algorithms”.

Data Science Workflow v2



Ask an interesting question.

What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

Get the data.

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Explore the data.

Plot the data.
Are there **anomalies**?
Are there **patterns**?

Model the data.

Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

See also “What is the work flow or process of a data scientist?” on the Quora

Table of Contents

1 STAT3612 Course Outline

2 Introduction to Data Science
- The Age of Big Data
- Data Science Venn Diagram
- Data Science Workflow

3 Statistical Machine Learning
- Interpretable Machine Learning
- Automated Machine Learning

4 Python and Jupyter Notebook

Machine Learning

- **Machine Learning** is an integral part of data science. It refers to the study of computer algorithms that build models of observed data in order to make predictions or decisions.
- Machine learning can find patterns and discover knowledge from data. It is also called **Data Mining** or **KDD**.
- Machine learning refers to a whole set of algorithms, including **unsupervised**, **supervised**, and **reinforcement** learning.
- **Statistical machine learning** emphasizes statistical models, inferences and interpretations.

Machine Learning: Stat3612 Landscape

Supervised Learning: (both features X and response y)

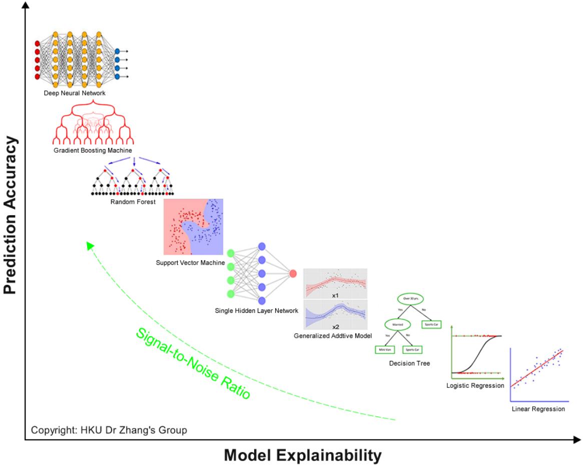
- Parametric regression: GLM, basis expansion, sparse modeling
- Nonparametric regression: splines, piecewise smooth modeling
- Kernel methods: support vector machines, Gaussian processes
- Tree-based methods: decision tree, random forest, gradient boosting
- Neural networks: Single-Hidden Layer Network, DNN/CNN/RNN

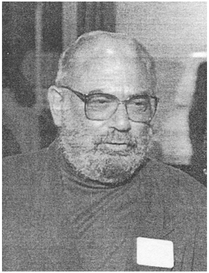
Unsupervised Learning (only features X)

- Dimension reduction: PCA, matrix factorization, auto-encoder
- Others: hierarchical clustering, K-means, t-SNE, outlier peeling

Supervised Machine Learning

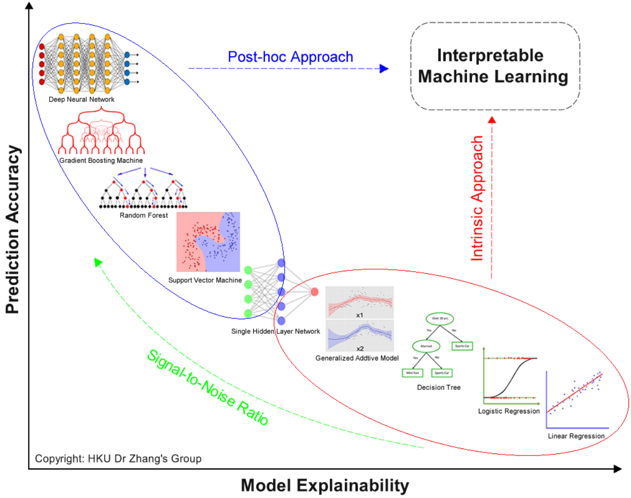
“Statistical Modeling: The Two Cultures” (Breiman 2001): Occam dilemma





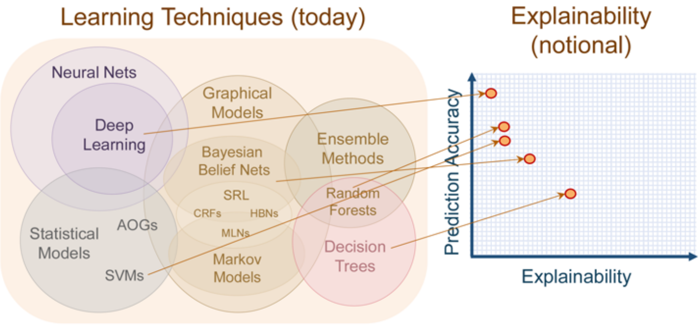
Leo Breiman
(1928–2005)

Interpretable Machine Learning (IML)



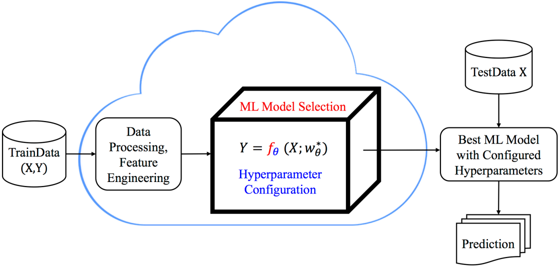
Copyright: HKU Dr Zhang's Group

IML a.k.a. XAI (Explainable Artificial Intelligence)



Gunning (2017). Explainable Artificial Intelligence (XAI). *US Defense Advanced Research Projects Agency (DARPA) Report.*

Automated Machine Learning (AutoML)


TrainData (X,Y)
Data Processing, Feature Engineering
ML Model Selection
$Y = f_\theta\,(X; w_\theta^*)$
Hyperparameter Configuration
TestData X
Best ML Model with Configured Hyperparameters
Prediction

- AutoML is to perform model selection and hyperparameter configuration automatically for maximizing prediction accuracy.
- Also: progressive automation of data preprocessing, feature engineering and postprocessing.
- AutoML alone is a lame duck. We actually need the AutoIML.

Table of Contents

1. STAT3612 Course Outline
2. Introduction to Data Science
   - The Age of Big Data
   - Data Science Venn Diagram
   - Data Science Workflow
3. Statistical Machine Learning
   - Interpretable Machine Learning
   - Automated Machine Learning
4. Python and Jupyter Notebook

Python and Jupyter Notebook

- Download and install Python from https://www.anaconda.com/
- Download and install Jupyter Notebbook from https://jupyter.org/
- Note the new release of JupyterLab IDE ......
- Free Google Colaboratory; Click here
- You will learn about Python/Notebook coding through the tutorials.
- **Important Note:** Jupyter Notebook is the recommended format for the assignments and the final project report.

Preview of Statistical Machine Learning

Demo of Google Colab (Python)

https://colab.research.google.com/

Thank You!

Q&A or Email ajzhang@umich.edu