# Delivering Cutting-Edge Data Science across Industries

Moorissa Tjokro

DATA SCIENCE INDONESIA

Block71, Jakarta Kuningan | March 17, 2018

# About

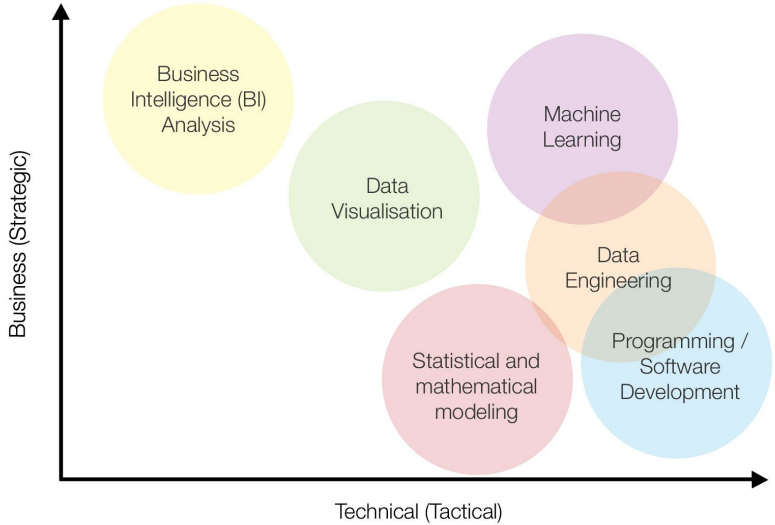
- Born and raised in Malang, East Java
- B.S. in Industrial Engineering & Statistics ‘14 from Georgia Institute of Technology
- M.S. in Data Science ‘17 from Columbia University
- Data science industries: from marketing and entertainment to space and automotive
- Hobbies: painting, running, and traveling

# Overview

1. What is Data Science?
2. Data Science across Industries (use cases & techniques)
3. How to be a Data Scientist?

# What is Data Science?

*An interdisciplinary field of statistics, computer science, algorithms and related methods to extract knowledge from data.*



# Foundations of Data Science

- Probability and Inferential Statistics / Modeling
- Algorithms and data structures
- Computer systems and database / data wrangling
- Exploratory data analysis, visualizations and descriptive statistics
- Machine Learning and related methods

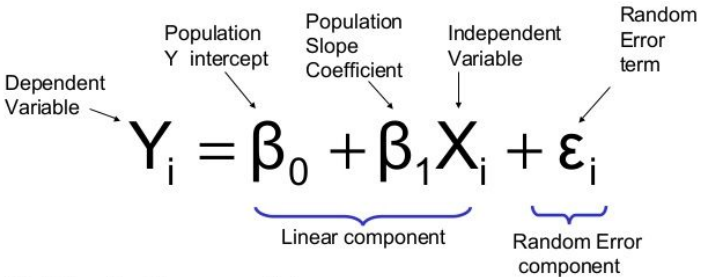
# Common Techniques

| Supervised Learning | Unsupervised Learning |
| --- | --- |
| Linear Models / Regressions | Clustering |
| Naive Bayes Classifiers | Principal Component Analysis (PCA) |
| Decision Trees & Ensembles | Non-Neg Matrix Factorization (NMF) |
| Support Vector Machines | Manifold Learning with t-SNE |
| Neural Networks | |

# Data Science across Industries

# Nonprofits & Public Sector

- Use case: **Linear regression** models to select top 10k people who will donate the most




- **A/B testing** to test if a new feature works well

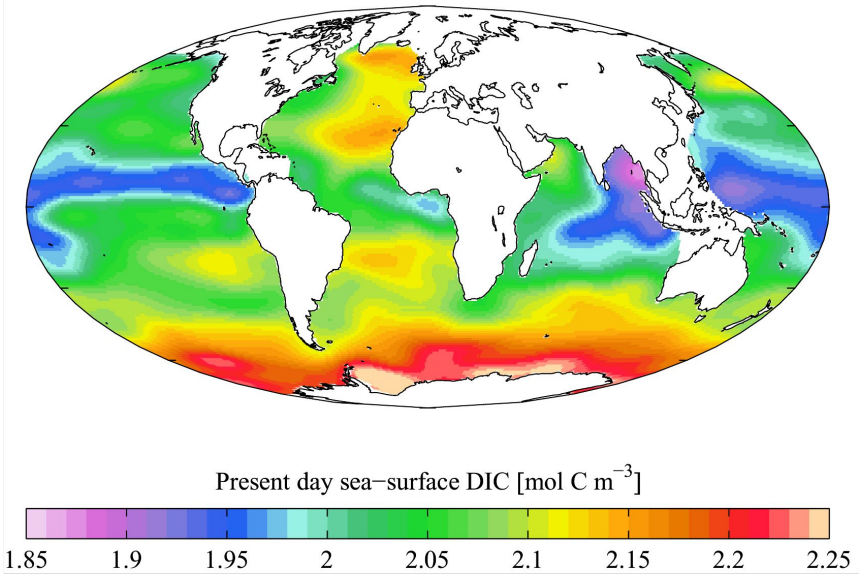
# Finance

- Use case: **Logistic regression** for developing fraud detection algorithms
- More advanced technique is *meta-learning approach*

| | Actual Fraud $y_i = 1$ | Actual Legitimate $y_i = 0$ |
|---|---|---|
| Predicted Fraud $c_i = 1$ | $C_{TP_i}$ | $C_{FP_i}$ |
| Predicted Legitimate $c_i = 0$ | $C_{FN_i}$ | $C_{TN_i}$ |

# Goddard Institute for Space Studies

- Use case: **Clustering algorithms** to classify ocean carbon states
- Goal: To gain insight into the physical and biogeochemical processes controlling the ocean carbon cycle in nature



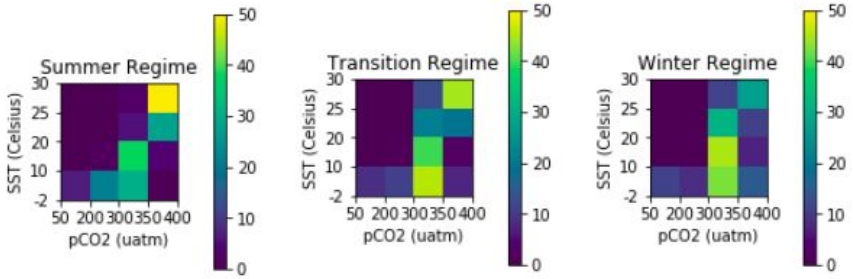
# Goddard Institute for Space Studies

**Average months are classified into 3 clusters:**

winter summer transitional

| Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|

**2-D Histograms with SST and pCO2 variables:**

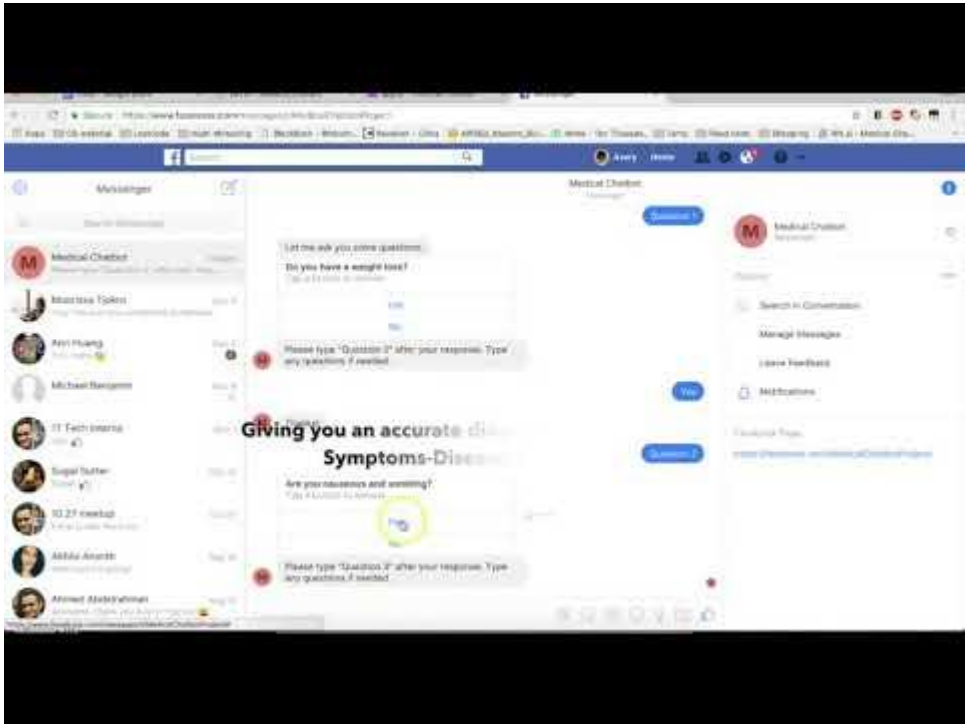


# Self-Driving Cars

- Use case: **Reinforcement learning** that is inspired by biological neural networks that learns environment of the real world.

# Healthcare

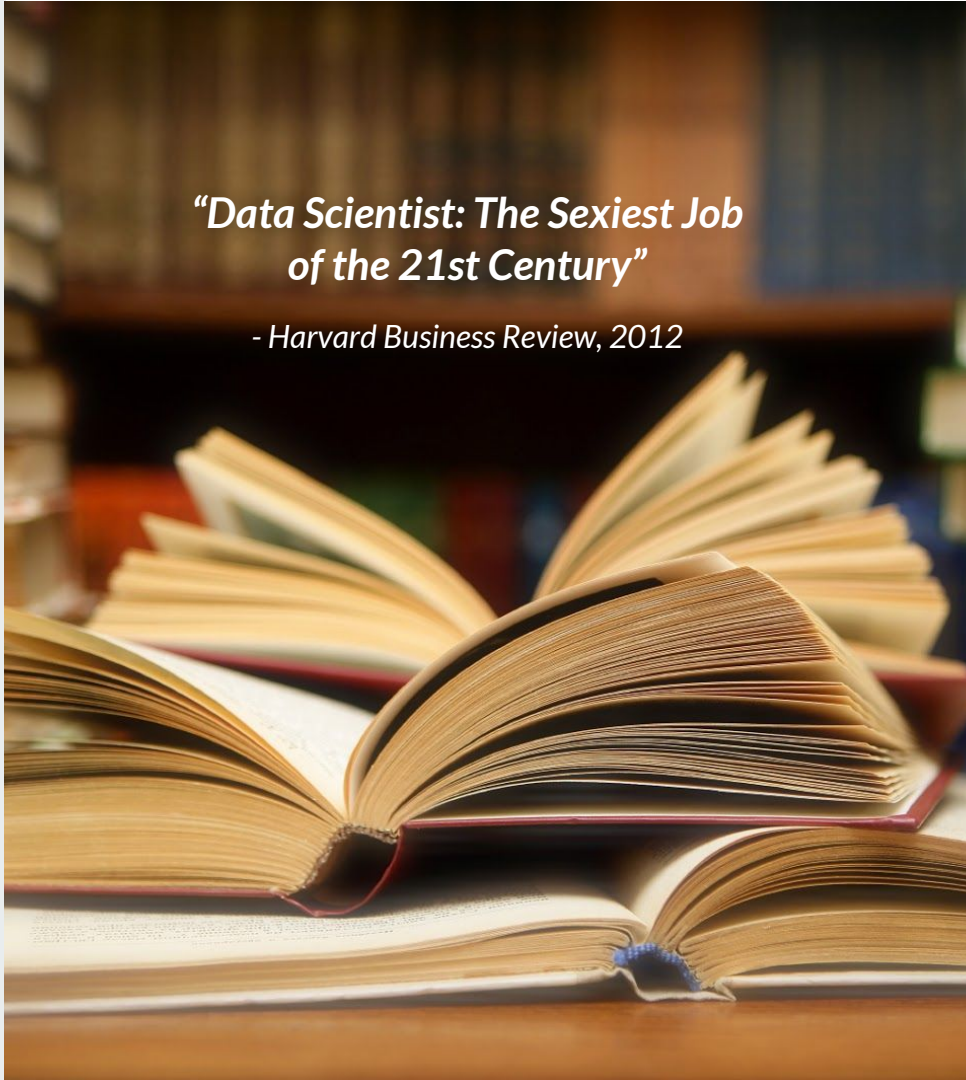
- Use case: **Natural Language Processing** to learn scripts and extract keywords and important information.
- Combined with matching algorithms



# How Data Science Flourish in Great Companies

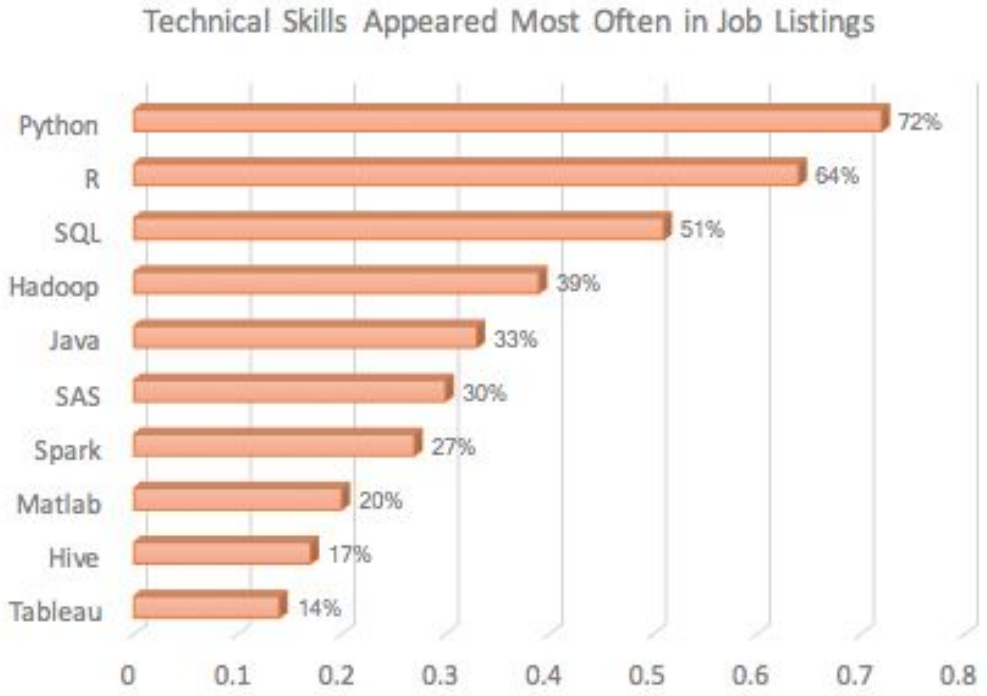
- Invention / innovation (competitiveness in the industry to bring more efficiency for customers)
- Creativity to be different because employee's voice matters
- Empowering leadership in work setting
- Bringing the smartest, most-skilled people by offering great compensations

# How to become a Data Scientist?



# Hard Skills

- Be fluent in Python, R, SQL
- Know the use of computing clusters, e.g. Google Clouds or Amazon Web Services (AWS)
- Learn data science techniques using different libraries




*Glassdoor, Jan - Jul 2017*

# Soft Skills

- Be an excellent team player, “Smart people aren’t smart if they can’t work with others.”
- Curiosity puts you in a sea of opportunities
- Critical thinking skills & analytical acumen
- In industries, efficiency > quality

# Resources

- Elements of Statistical Learning (Friedman, Hastie, and Tibshirani, 2001)
- Introduction to Machine Learning with Python (Mueller and Guido, 2017)
- Coursera, Udacity, Udemy, and other online learning platforms (basic/advanced)
- Subscribe to machinelearningmastery.com for regular learning of theories/applications
- Kaggle.com for data challenge and projects
- Leetcode.com for coding practices at all levels


- Thank you -
Questions?
DATA
SCIENCE
INDONESIA