

# **Environmental Health Big Data Analysis**

**Il-Youp Kwak**

Chung-Ang University

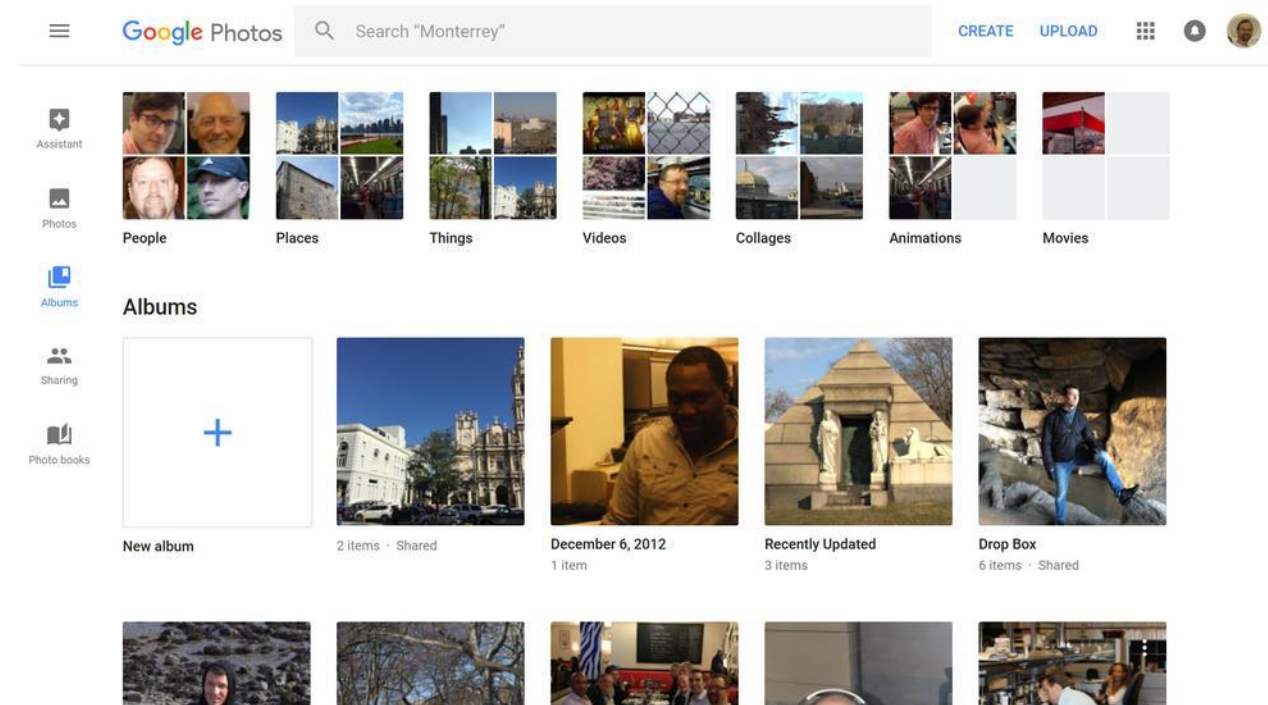
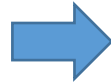
# About Me



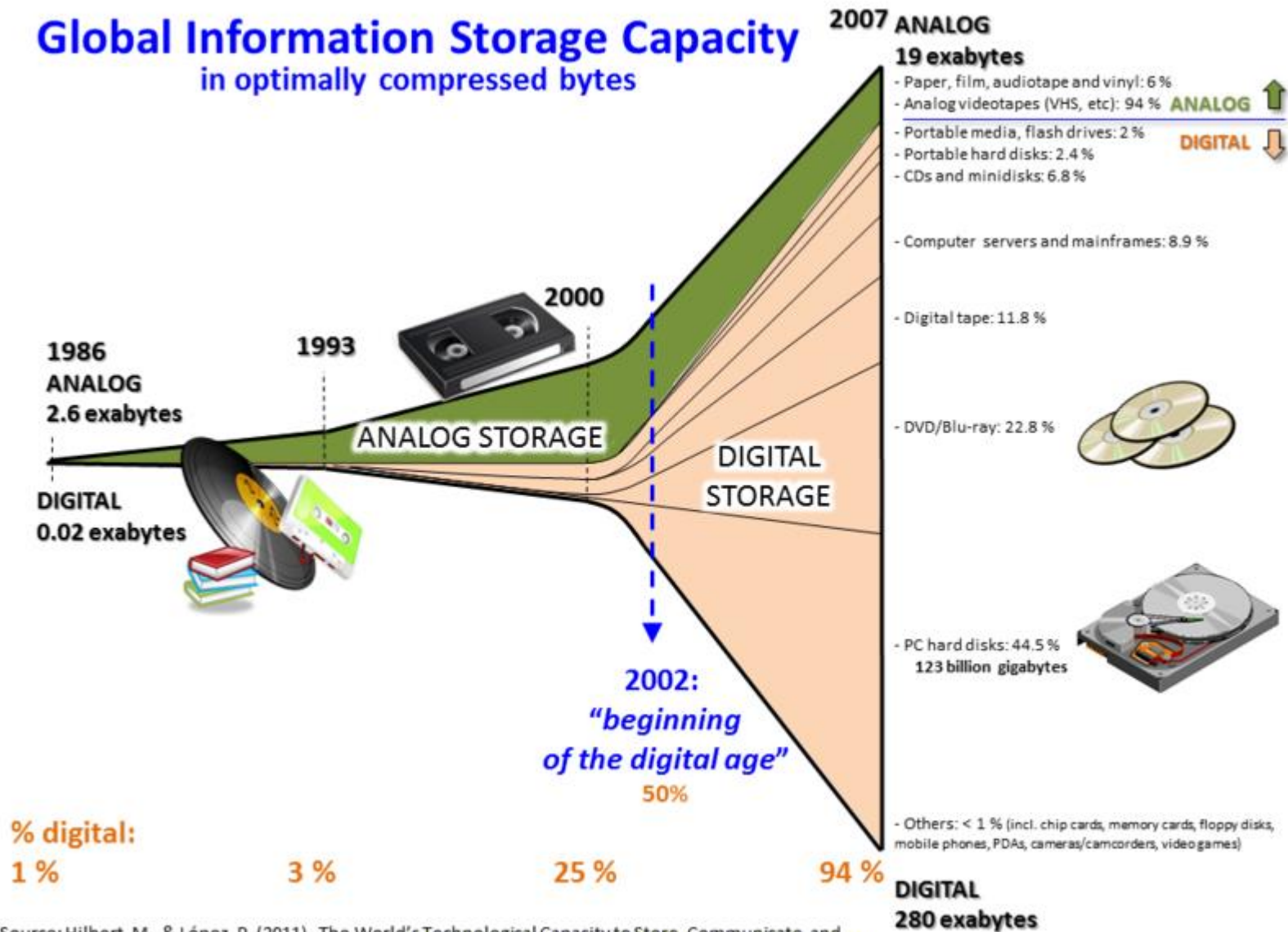
I am an Assistant Professor in [the Department of Applied Statistics](#) at [Chung-Ang University](#); research in Statistical Genetics and Deep Learning.

I am eager to design new methods for big and complicated data. I am also excited to develop new useable tools for the method so that those methods can be widely adopted and used by other researchers

# Old Photo Books to Google Photos



# Global Information Storage Capacity is Growing

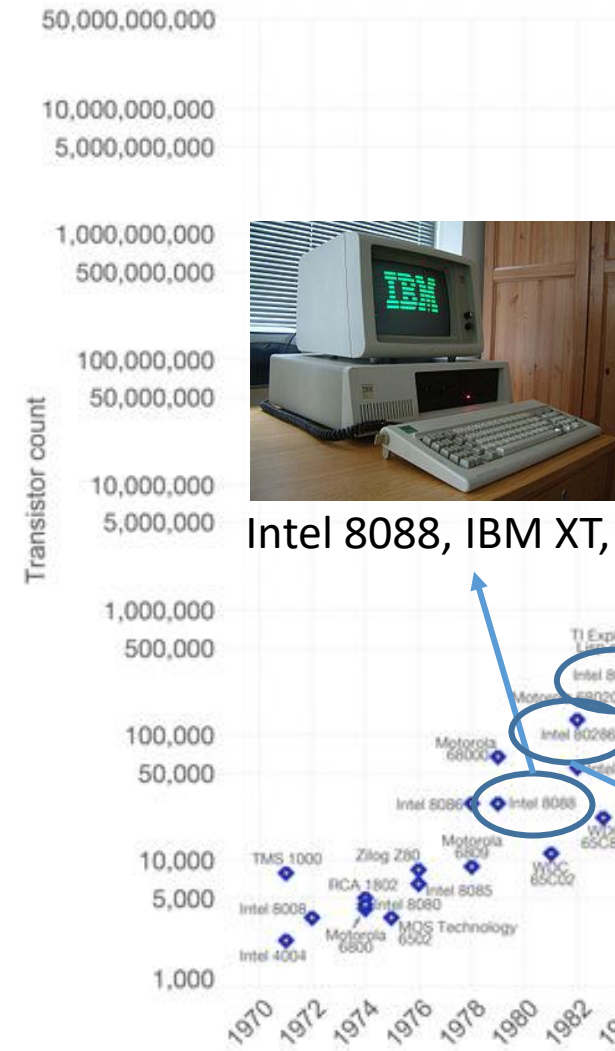


Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60–65. <http://www.martinhilbert.net/WorldInfoCapacity.html>



# Exponential Laws of Computing Growth

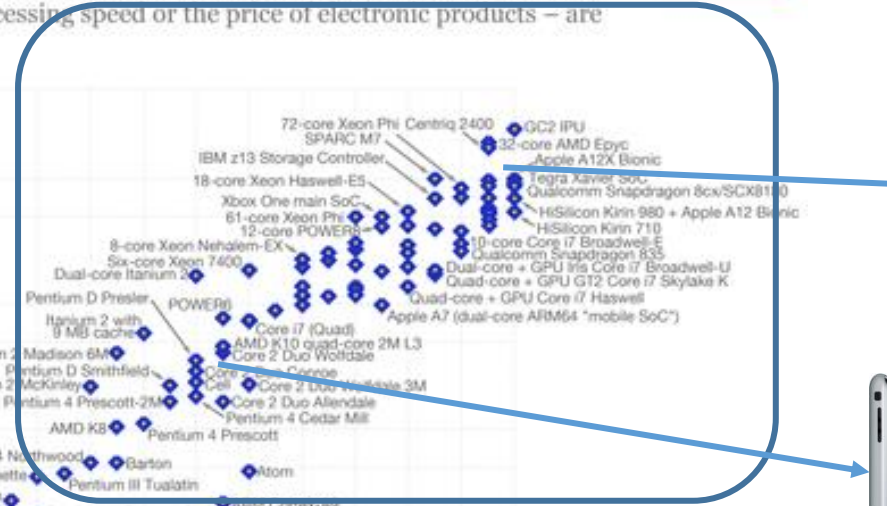
**Moore's Law** – The number of transistors on a microchip doubles approximately every two years. This advancement is important as other aspects of technology linked to Moore's law.



Intel Pentium, i486

Integrated circuit chips (1971-2018)

Integrated circuits doubles approximately every two years. Processing speed or the price of electronic products – are





Getting Multiple cores



**Computers are getting faster,**

**Information storage growing bigger**

# Opens **Big Data** Era





A good eco for **AI**





# Growing need for **Data Scientists**

## BEST JOB IN AMERICA: DATA SCIENTIST!

Glassdoor recently ranked the 50 Best Jobs in America  
3 of the top 5 are in the Data Analytics Field



### #1 Data Scientist

4000+ openings on Glassdoor  
\$110K median base salary  
4.4/5.0 level of job satisfaction



### #3 Data Engineer

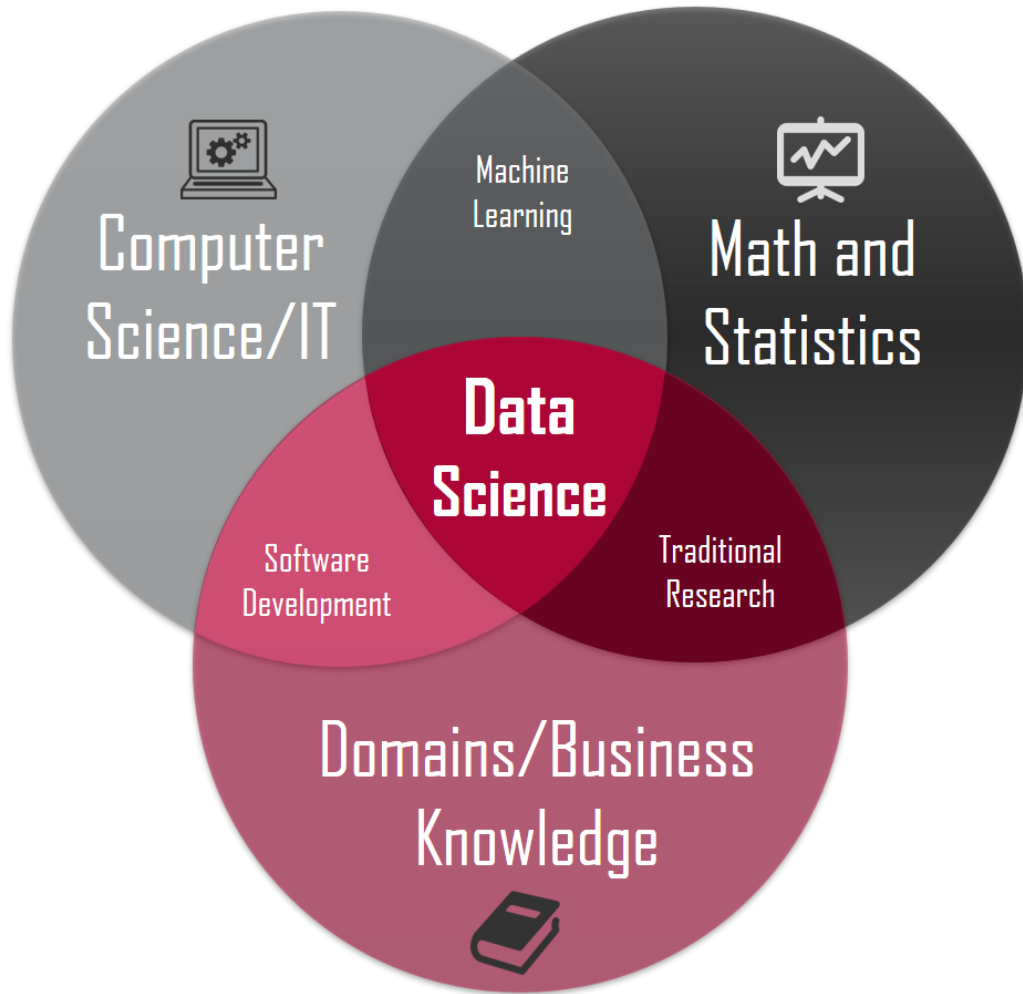
2500+ openings on Glassdoor  
\$106K median base salary  
4.3/5.0 level of job satisfaction



### #5 Analytics Manager

2000 openings  
\$112K median base salary  
4.1/5.0 level of job satisfaction

# Core Skills for Data Scientist



**Intellectual curiosity**

**Openness to learning new things**

**Ability to solve problems in unique way**

**Passion for innovation**

# **About this course?**

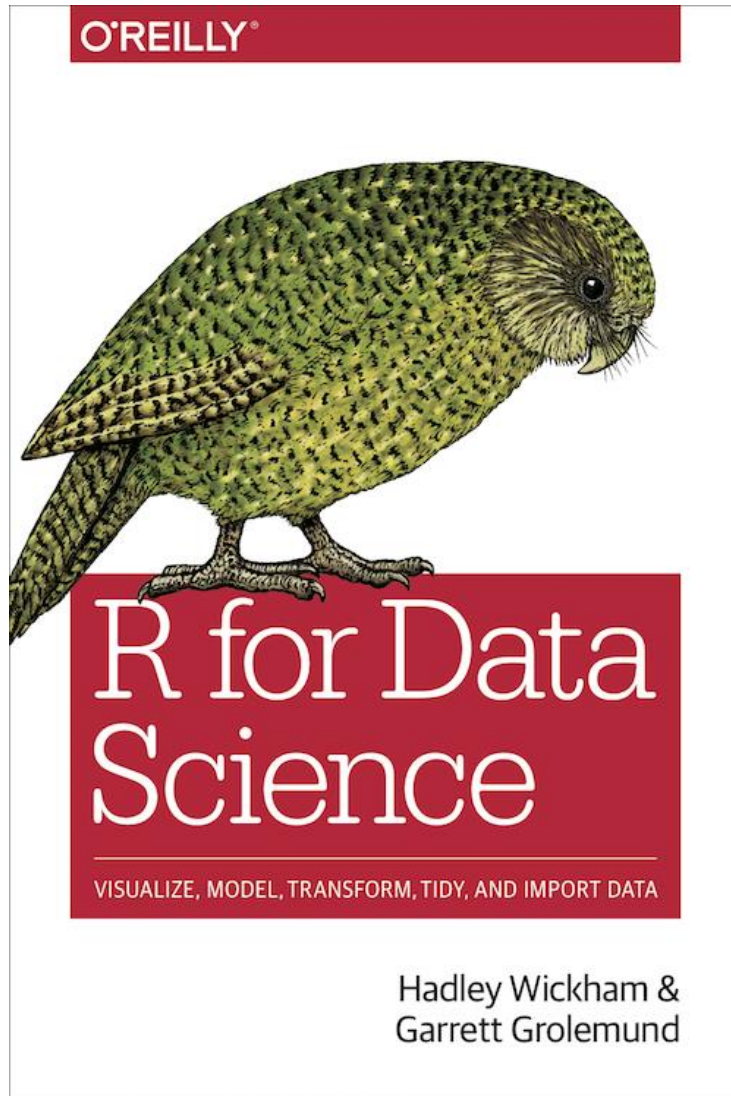
**Improve Practical Computational Skills (R and Python)**

**Learn reproducible research techniques  
(Github, Rmd, Jupyter Notebook)**

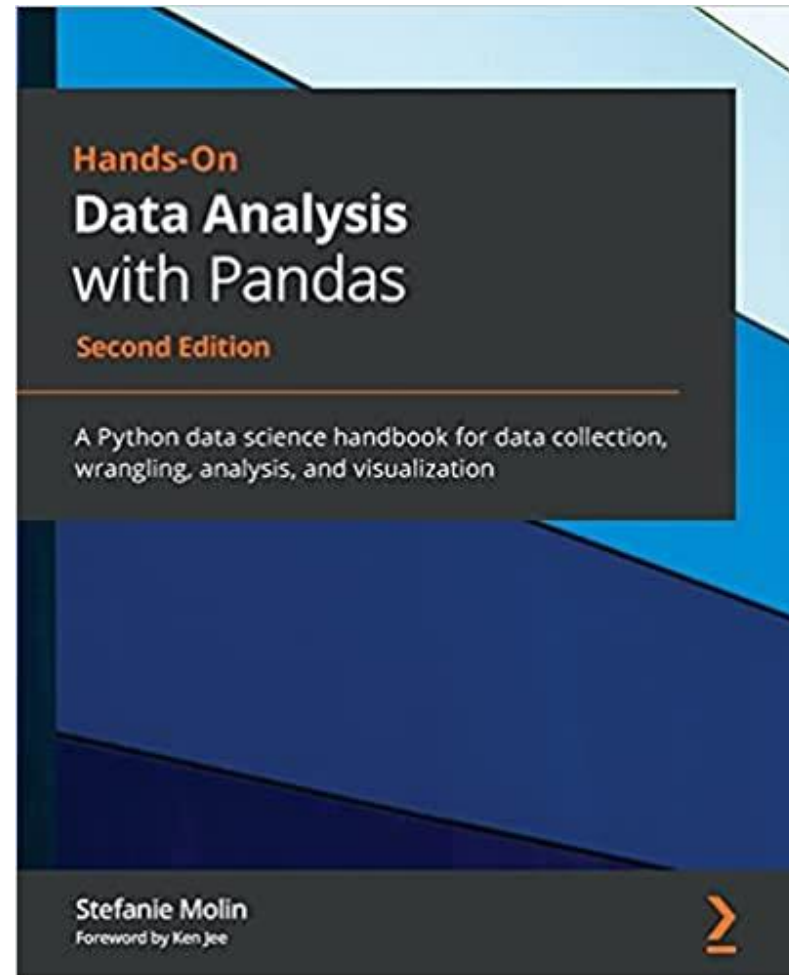
**Team project using big data (Final Project)**



# Textbooks



















<https://r4ds.had.co.nz/>





















<https://github.com/stefmolin/Hands-On-Data-Analysis-with-Pandas-2nd-edition>

# Plans

1	과일업	Course Introduction. Colab 실습, Jupyter notebook 실습, Learn Github			
2	과일업	Markdown, Rmarkdown, Jupyter notebook			
3	과일업	R for Data Science 1 (dplyr)	HW1		
4	과일업	R for Data Science 2 (tidyr)			
5	과일업	R for Data Science 3 (ggplot)			
6	과일업	Working with Pandas dataframe	HW2		
7	과일업	Data Wrangling with Pandas			
8	과일업	중간고사			

# Plans

9	과일업	Aggregating Pandas DataFrames	HW3		 
10	과일업	Visualizing Data with Pandas and Matplotlib, Visualizing multi-dimensional data (t-SNE, UMAP)			 
11	과일업	Plotting with Seaborn and Customization Techniques			 
12	과일업	Financial Analysis – Bitcoin and the Stock Market ;		Zoom, 5 min plan presentation	 
13	과일업	Rule-Based Anomaly Detection			 
14	과일업	Getting Started with Machine Learning in Python			 
15	과일업	Talk from the industry (talk from data scientist?)			 
16	과일업	팀별 데이터 분석 결과 발표	기말프로젝트 레포트 완성본, 기말 ppt, 개인 제출자료 제출	기말고사는 팀별 발표로 대체. 각 팀별로 데이터 분석 및 결과를 약 15분간 발표 진행	 
					 



# Discussion (Team Project)

- Each team will have 3~4 members (4~5?)
- Pick any data
  - <https://datahub.io/search>
  - <https://datasetsearch.research.google.com/>
  - <https://registry.opendata.aws/>
  - Etc
- Zoom session for team selection (when?)

**Thank you! 😊**