



빅데이터 모델링과 플랫폼 = 빅모플

Big Data Modeling and Platforms

AIX30011 Spring 2025

한동대학교
HANDONG GLOBAL UNIVERSITY

School of Applied Artificial Intelligence | 전재영

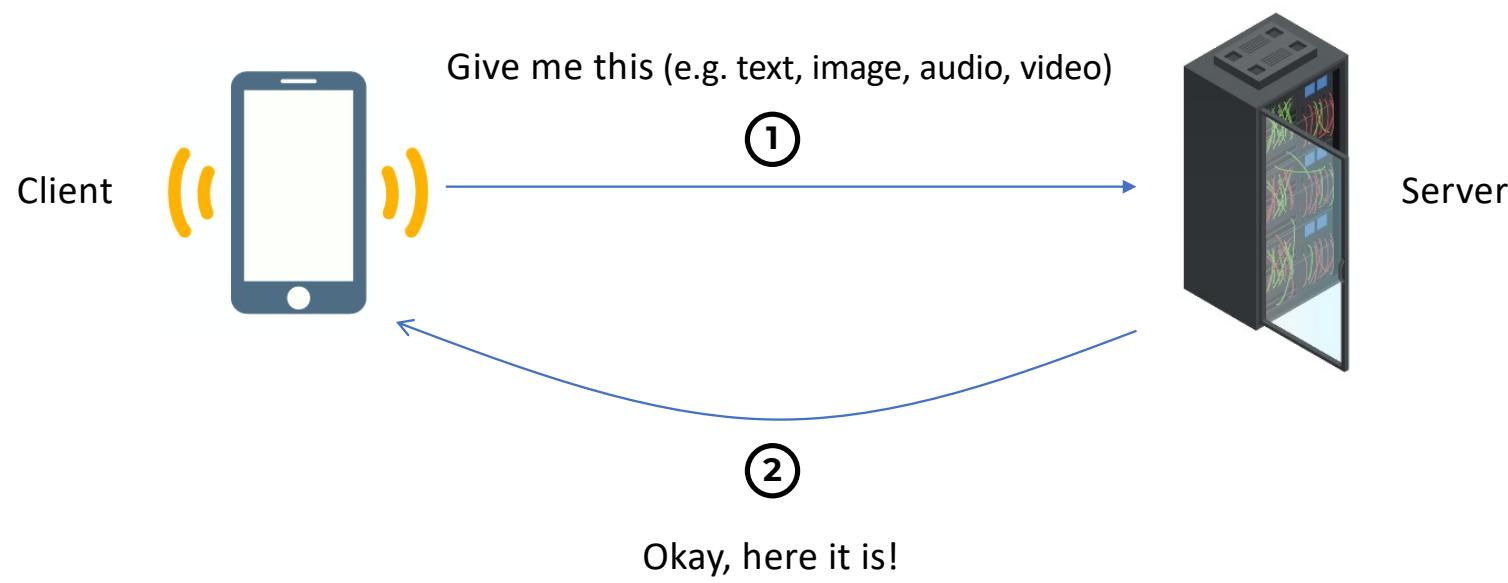
Copyright Notice

These slides are available for educational purposes. You may not use or distribute these slides for commercial purposes. You may make copies of these slides and use or distribute them for educational purposes as long as you cite the author as the source of the slides.



Jaeyoung Chun

Browsing



CNN World

Hear why some Russian troops are refusing to fight and say they're in a 'desperate position' 02:32

LIVE UPDATES

Russia's war in Ukraine

By Tara Subramaniam, Jack Guy, Aditi Sangal, Mike Hayes, Leinz Vales and Maureen Chowdhury, CNN

Updated 8:08 p.m. ET, February 27, 2023

What we're covering

- The situation in the eastern city of Bakhmut is "getting more and more challenging," President Volodymyr Zelensky said Monday. Separately, a Ukrainian commander said the situation in Bakhmut was "extremely difficult" due to non-stop Russian assaults.
- US Treasury Secretary Janet Yellen met with President Volodymyr Zelensky in Kyiv on Monday to reaffirm US economic support for

Network Performance Monitor (NPM) Screenshot:

The screenshot shows a Network Performance Monitor tool with the following details:

- Filter:** All, Fetch/XHR, JS, CSS, Img, Media, Font, Doc, WS, Wasm, Manifest, Other, Has blocked cookies.
- Blocked Requests:** Blocked Requests, 3rd-party requests.
- Timeline:** 20000 ms, 40000 ms, 60000 ms, 80000 ms, 100000 ms.
- Table Headers:** Name, Status, Type, Initiator, Size, Ti..., Waterfall.
- Table Data:** A list of network requests including:
 - get?pub=13060000-6761-40... 200 script controltag.js...
 - f3b6d00d-676f-48d8-80ef-2... 200 script controltag.js...
 - pixel.gif?source=smarttag&... 204 text/p... controltag.js...
 - eligible?wkz=M4Uw5gtiB2... 200 gif main_a5903...
 - eligible?wkz=M4Uw5gtiB2... 200 gif main_a5903...
 - eligible?wkz=M4Uw5gtiB2... 200 gif main_a5903...
 - ppub_config?ippd=edition.c... (blocker) xhr gpt.js:10 0 B 2...
 - pubads_impl_2023022301.js 200 script gpt.js:10 133 kB 9...
 - integrator.js?domain=editio... 200 script pubads_im...
 - integrator.js?domain=editio... 200 script pubads_im...
 - ads?psid=7228592286732... (blocker) xhr pubads_im...
 - container.html (blocker) docu... pubads_im...
 - data:image/png;base... 200 png chrome-erro...
 - data:image/png;base... 200 png chrome-erro...
 - audience_id (blocker) xhr cnn-ais.min...
 - pubsync?redir=https://umto... (blocker) xhr cnn-ais.min...
 - zion-telemetry.api.cnn.io 200 ping metricRepor...
 - sodar?sv=200&tid=gpt&tv=... 200 xhr pubads_im...
 - zion-telemetry.api.cnn.io 200 ping metricRepor...
 - favicon.ico 200 x-icon Other 5.7 kB 5...
 - zion-telemetry.api.cnn.io 200 ping metricRepor...
 - ping?h=edition.cnn.com&p... 200 gif chartbeat_v...
 - sodar2.js 200 script pubads_im...
 - runner.html 200 docu... sodar2.js:32 5.5 kB 4...
 - aframe 200 docu... aframe:1 1.2 kB 4...
 - bP143D2MfrYa-8L1g1kZr... 200 script runner.html:25 14.8 kB 4...
 - fabric-icons.css 200 styles... content.js:2 242 B 1...
 - sodar?id=sodar2&v=225&li... 204 text/h... aframe:1 0 B 4...
 - fabric-icons.css 200 styles... content.js:2 242 B 5...

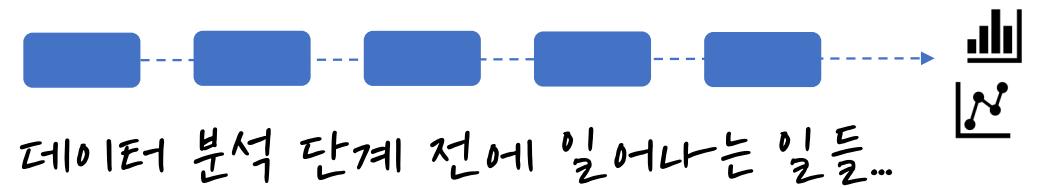
Live



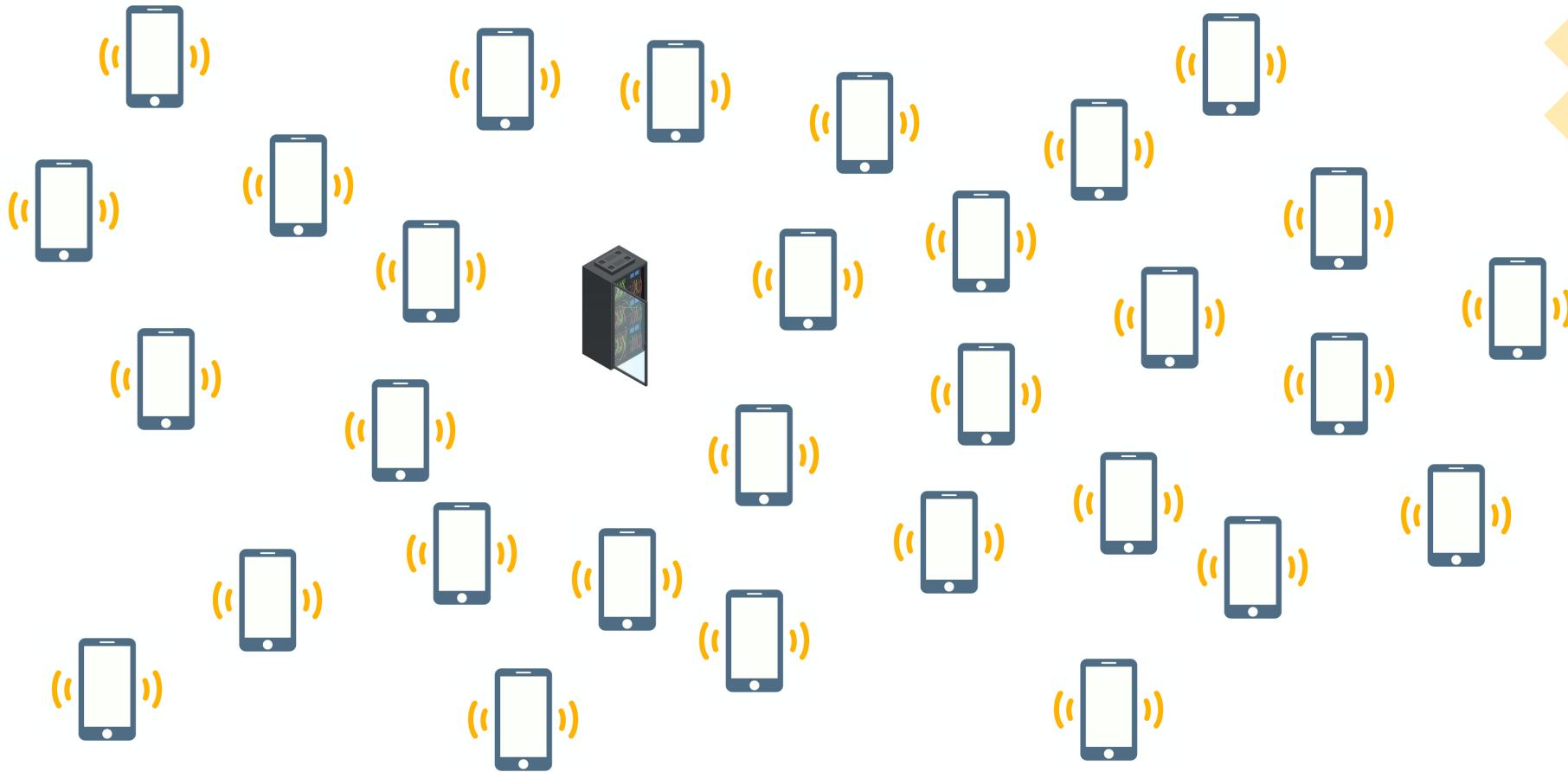
Jaeyoung Chun

Jaeyoung Chun

Big Data

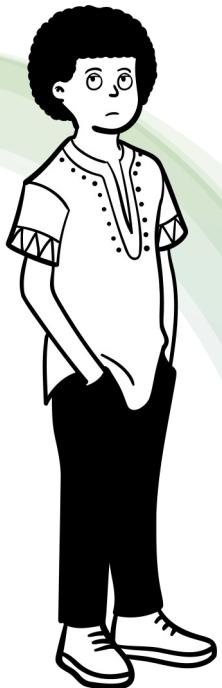


데이터 분석 단계 전에 일어나는 일들...



WE

- 경영+데이터사이언스
- 생명+데이터사이언스
- ACE+데이터사이언스
- ICT융합+데이터사이언스
- 상담심리+데이터사이언스
- 도시환경+데이터사이언스
- 건설+데이터사이언스
- 경영+AI융합
- 경제+AI융합
- 생명+AI융합
- AI-컴퓨터공학심화
- ACE+수학통계
- 경제+ICT융합



'데이터' 다루는 것이 재미 없는 사람?

수강 여부를 파악하기 전에 먼저 ○○○

우리가 해보려는 것?

1. 데이터를 잘 다룬다는 것은 단순히 분석을 잘 하는 것인가?
2. 나도 정답을 맞출 수 있고 다른 사람도 정답을 맞출 수 있다면 이기는 자는 누구인가?
3. 실패했을 때 다시 일어날 수 있는가? 실패를 예방할 수 있는가?

80%



Big Data
Real-time

4. 기술적(technical) 역량 갖추면 충분한가?
5. 도메인 지식 없이 분석을 잘 할 수 있나?

Jaeyoung Chun

"Big Data Modeling and Platforms"

①

②

③



Big Data Modeling and Platforms

빅모풀

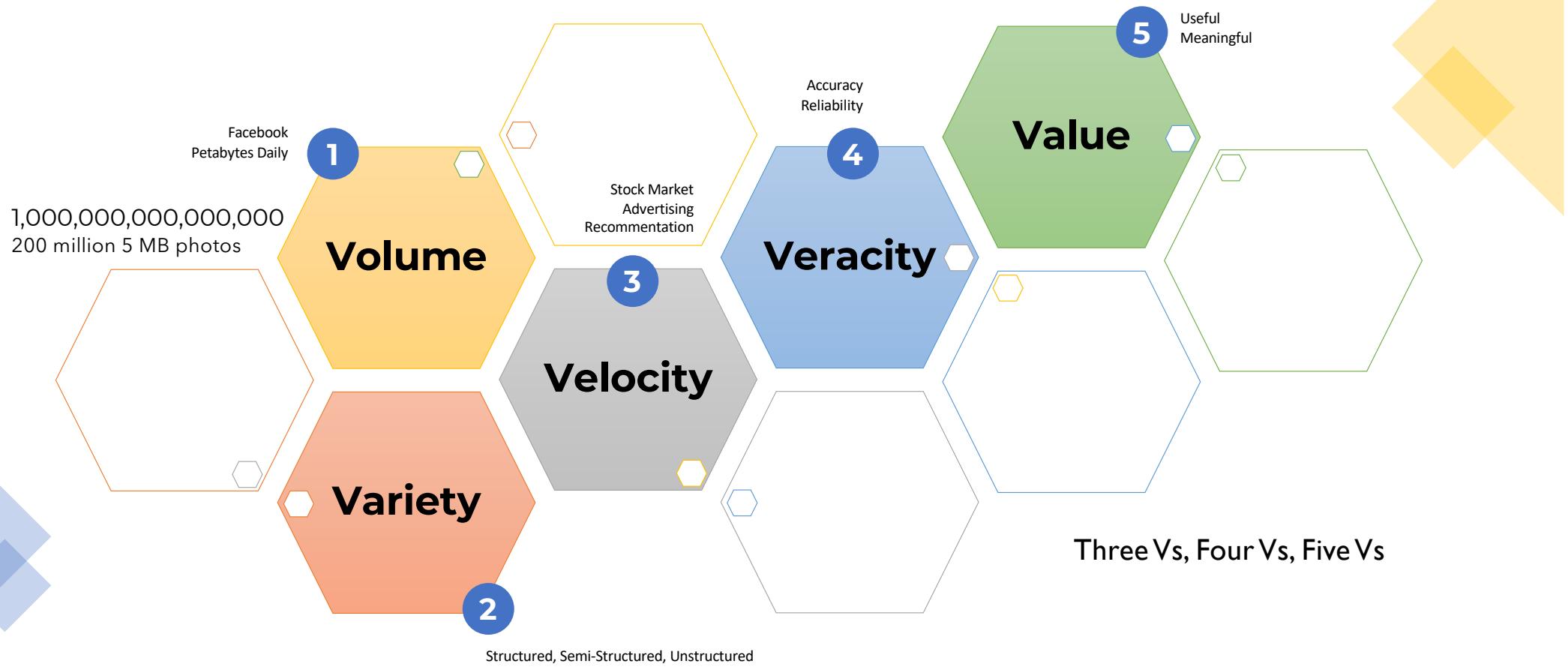
1

Big Data

Jaeyoung Chun

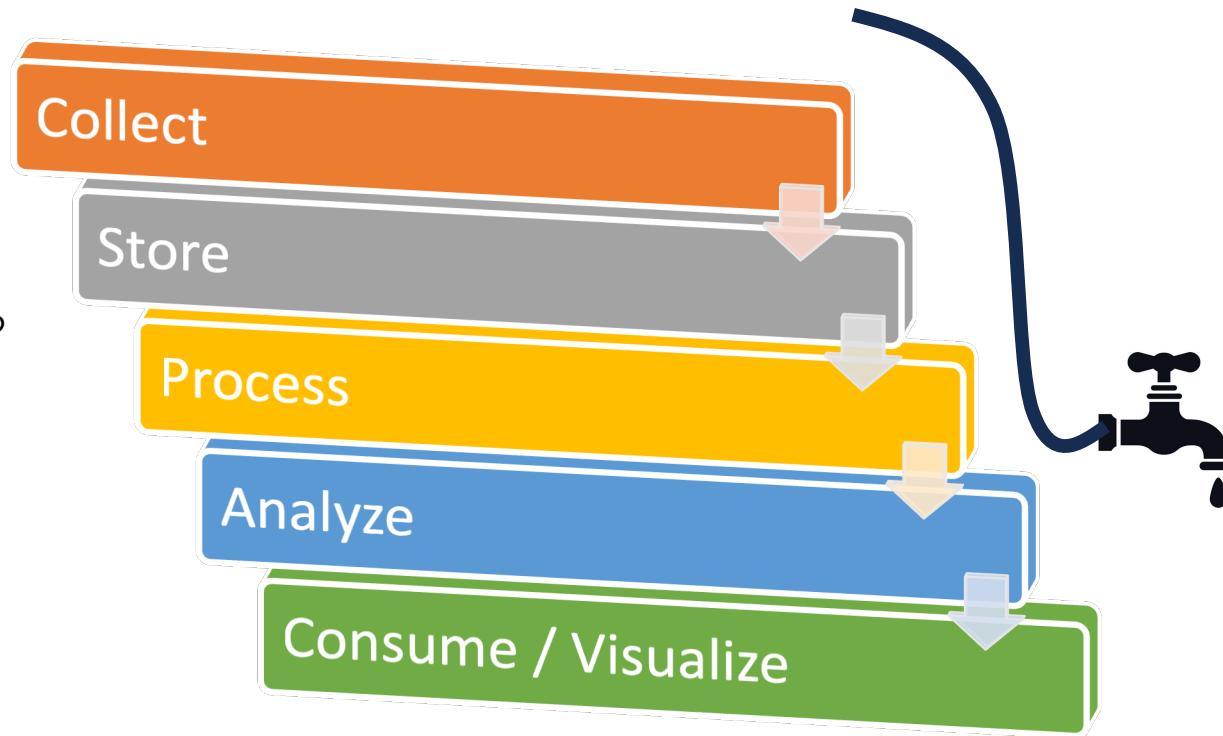
Big Data

단순히 크기만을 말하는 것이 아니다.



How Does Big Data Work?

- 어떻게 저장?
- 무엇을 어떻게 검색?



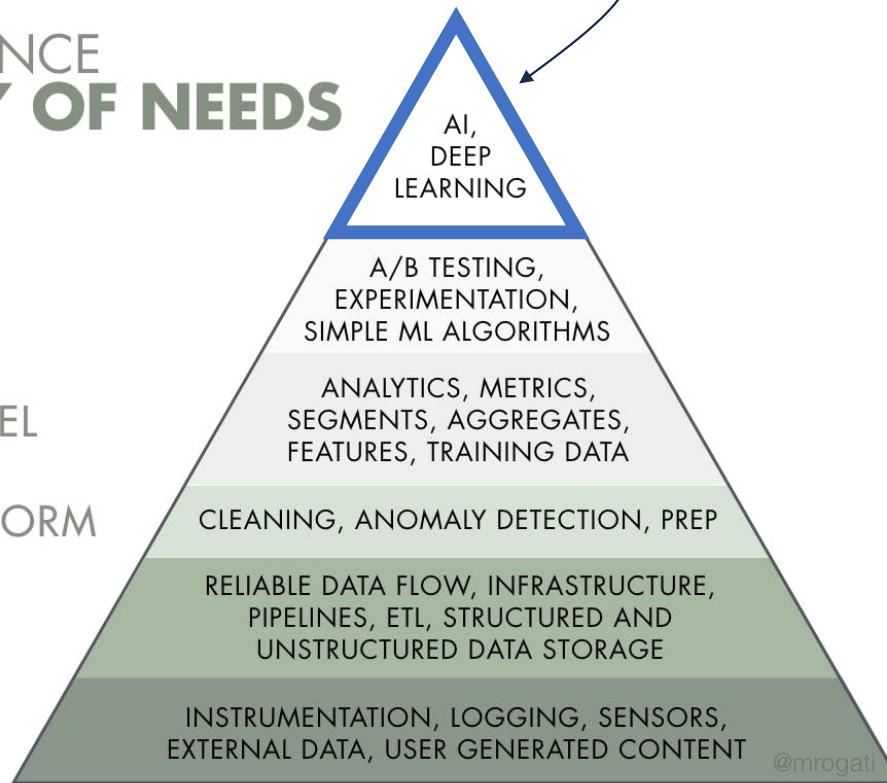
AI 만들어줘요!?!

!?!
!



- LEARN/OPTIMIZE
- AGGREGATE/LABEL
- EXPLORE/TRANSFORM
- MOVE/STORE
- COLLECT

THE DATA SCIENCE HIERARCHY OF NEEDS



Monica Rogati
@mrogati

Source: <https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007>

Jaeyoung Chun

속해 있는 조직에 따라 달라지는 직무

THE DATA SCIENCE HIERARCHY OF NEEDS

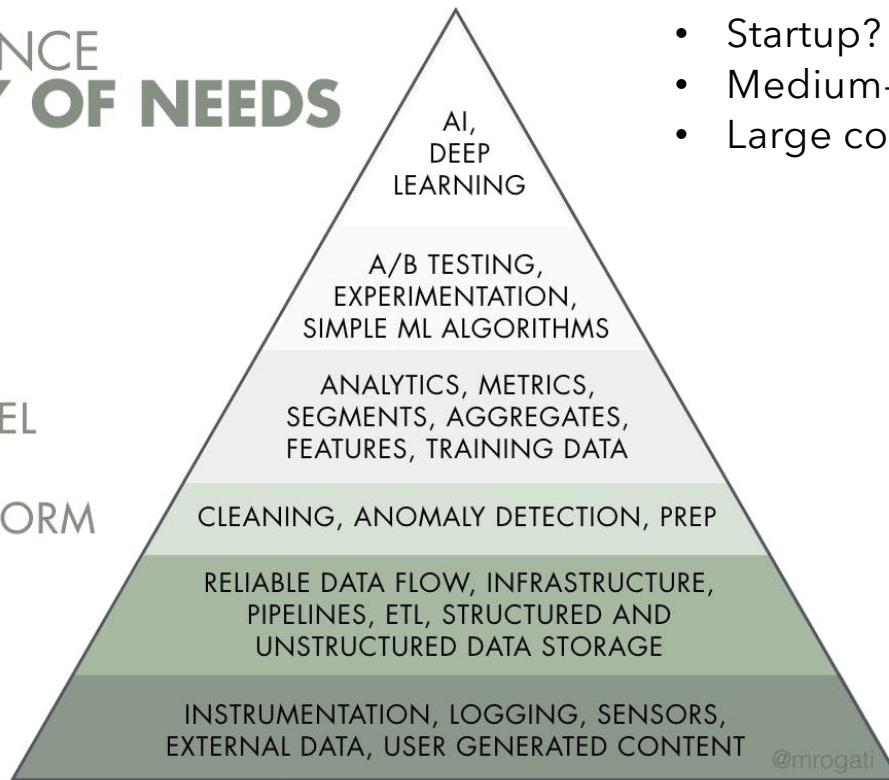
LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT



- Academia?
- Startup?
- Medium-sized company?
- Large corporation?



Monica Rogati
@mrogati



Big Data **Modeling** and Platforms

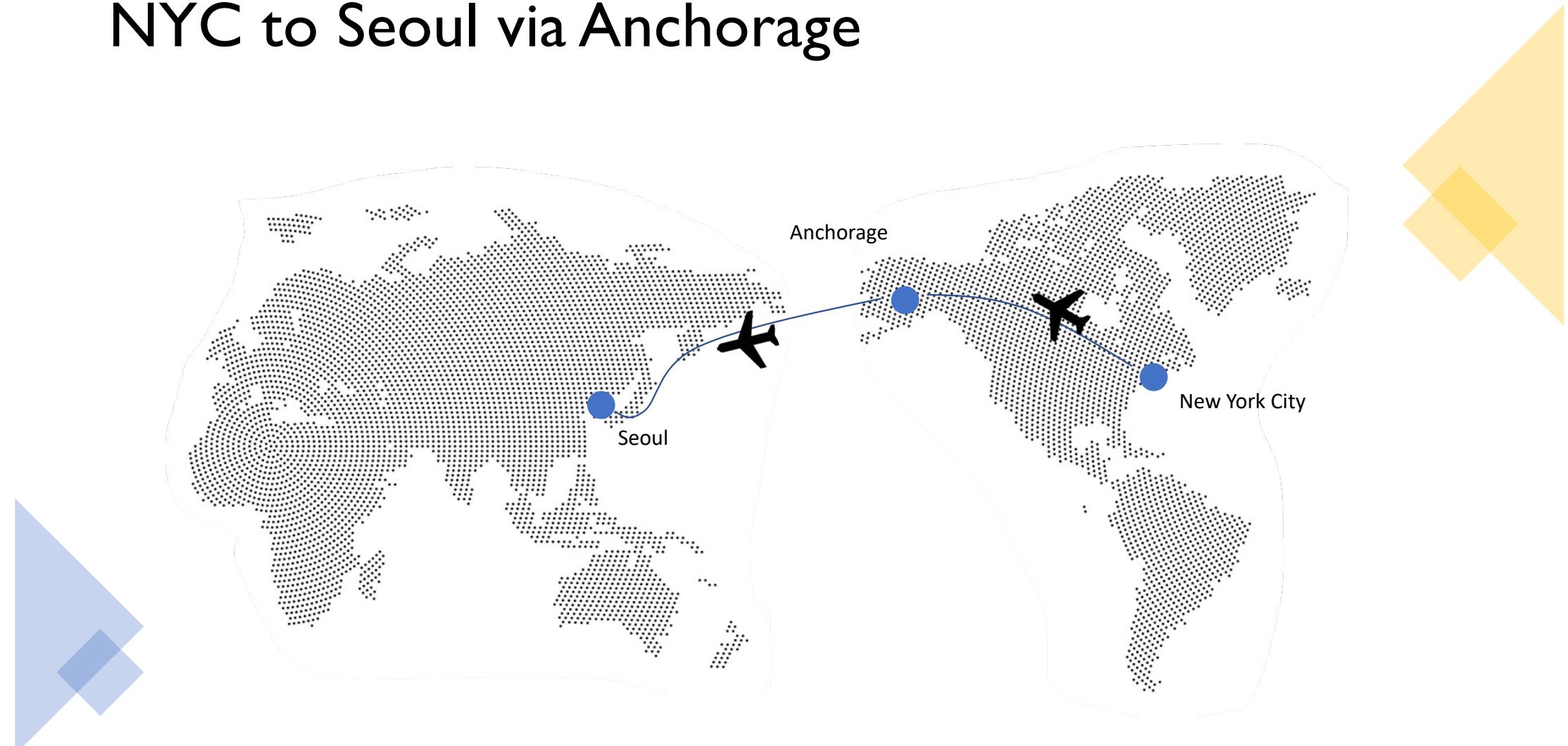
빅모풀

2

Modeling

Jaeyoung Chun

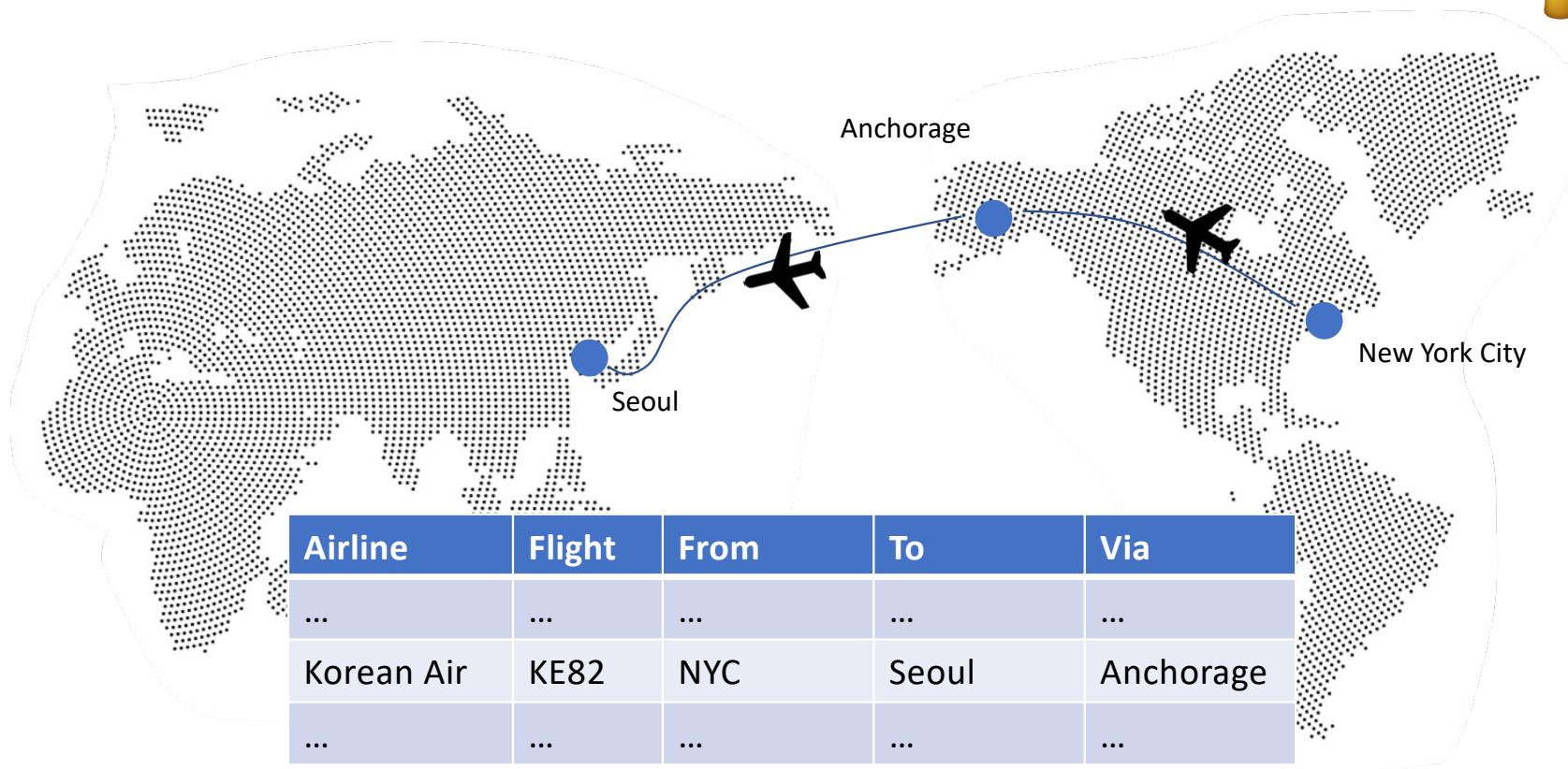
NYC to Seoul via Anchorage



Jaeyoung Chun

Any Issue?

이 항공 노선에서 비행기가 데칼로 바뀐다면?



Any Issue?

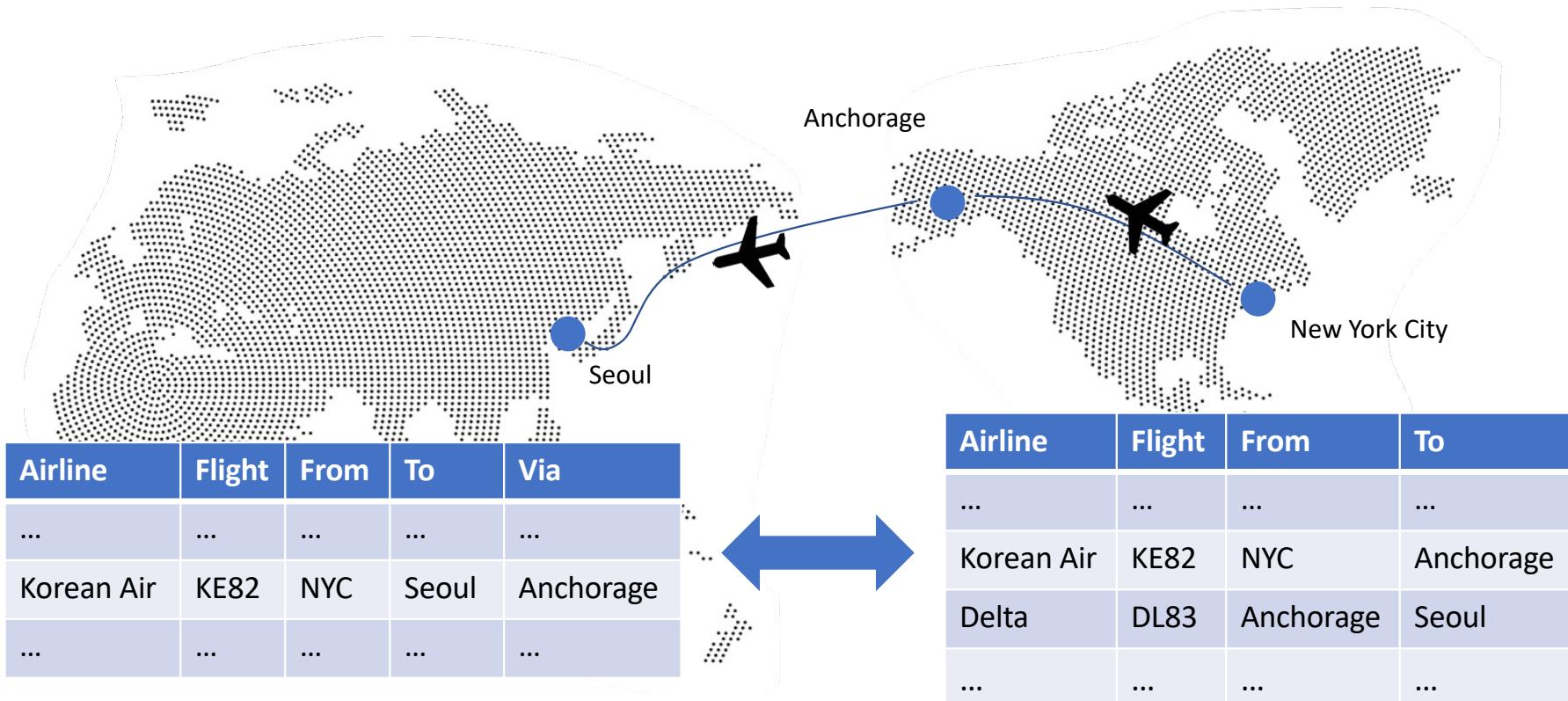
\$2K-3K per year

알래스카 유입 인구는?

여행 횟수는?



Transformation



Jaeyoung Chun

Non-Tabular?



Jaeyoung Chun



3

Big Data Modeling and **Platforms**

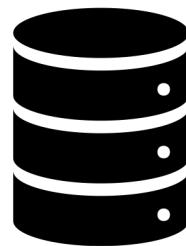
빅모풀

Platforms

Jaeyoung Chun

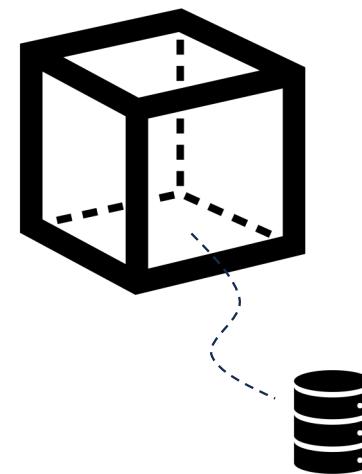
Database vs. Big Data Platform

Database



- Store
- Retrieve
- Update
- Manage

Platform



- Data Ingestion
- Real-time processing
- ...



데이터 저장

어떤 그릇에 물을 담느냐에 따라
물의 모양도 변하지만



데이터 검색

물을 부을 때
떨어지는 모습이나 속도도 바뀐다



Jaeyoung Chun

가격도 변한다;;;;



Type	Price per GB	Price per 100 TB	Retrieval Speed
S3 Standard	\$0.023 per GB	\$2,300	Instant
S3 Glacier Deep Archive	\$0.00099 per GB	\$99	12 hours

(*) per month, as of 2025-03-03

Platforms

- MySQL
- Redis
- Neo4j
- MongoDB
- Big Query
- Spark
- Kafka
- ...

“새 포도주를 낡은 가죽 부대에 넣지 아니하나니
그렇게 하면 부대가 터져 포도주가 쓸어지고
부대도 버리게 됨이라
새 포도주는 새 부대에 넣어야
둘이다 보전되느니라.”
(마태복음 9:17)





Top Data Scientist Technical Skills

Jaeyoung Chun

Top Data Scientist Technical Skills

- Programming
 - Python, R, SAS, **SQL**
- Statistics and Probability
- Data Wrangling and **Database Management**
- Machine Learning and Deep Learning
- Data Visualization
- Cloud Computing
- Data Visualization
- Python
- **SQL/NoSQL**
- Social Media Mining
- Fundamental Statistics
- Natural Language Processing/Machine Learning
- Microsoft Excel
- High-Level Math

<https://www.coursera.org/articles/data-scientist-skills>

<https://bootcamp.berkeley.edu/blog/data-scientist-skills/>

Jaeyoung Chun

Top Data Scientist Technical Skills

- Python
- R
- Statistics and Math
- SQL
- NoSQL
- Data Visualization
- Machine Learning
- Deep Learning
- Natural Language Processing
- Big Data
- Cloud Computing
- Cloud Computing
- Statistics and Probability
- Advanced Mathematics
- Machine Learning
- Data Visualization Tools
- Query Languages
- Database Management
- Python coding
- Microsoft Excel
- R programming
- Data wrangling

<https://www.datacamp.com/blog/top-15-data-scientist-skills>

<https://www.indeed.com/career-advice/resumes-cover-letters/skills-for-a-data-scientist>

Jaeyoung Chun



Course Logistics

Jaeyoung Chun

Instructor

이름	전재영
소속	AI 융합교육원 (전산전자공학부 & 생명과학부 겸임)
이메일	hisplan@handong.edu
웹사이트	hisplan.github.io

Jaeyoung Chun

TA

-
- 이찬규
 - 생명+AI융합 학부생
 - 22000584@handong.ac.kr



-
- 구나영
 - 생명+AI융합 학부생
 - nayoungku1@gmail.com

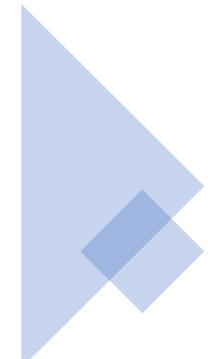
-
- There will be TA sessions.

Misc.

- Announcements
 - Check LMS frequently.
- Office Hours
 - Starting from March 7 (Friday)
 - Make appointments at <https://calendly.com/hisplan-handong/casual-chat>.
 - Ebenezer Hall 310

Prerequisites?

- Linux (connect, transfer, execute via terminal)
- Docker
- Python



Terminologies

- Will use English terminologies most of the time.
 - Relational Database (관계형 데이터베이스)
 - Replication (복제)
 - Disaster Recovery (재해복구)
 - Virtualization (가상화)
 - ...



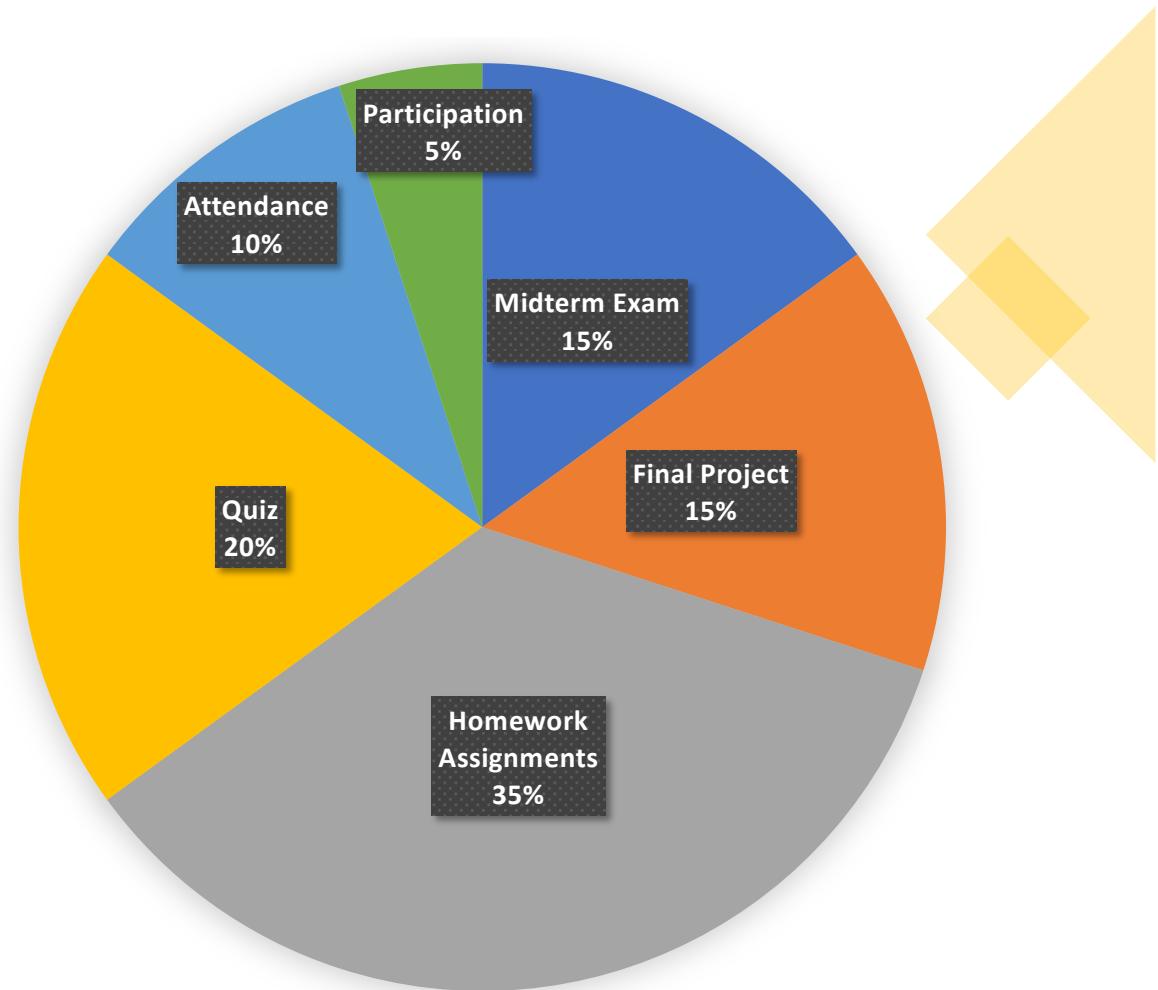
Honor Code

- You are responsible for understanding and complying with Handong Honor Code.
- If you copy someone else's coding or homework assignments, or use someone else's creative work without any indication, you will get an F grade. Anyone who share the code with you will also get an F grade.

Evaluation

- All the assignments must be done **individually**.
- All the assignments are expected to be submitted on the day they are due.
- **Any late assignment will be penalized 50%.**
- Exams will be based on lectures and assignments. If you don't do the assignments by yourself, you will be troubled.

50%



Attendance

1. Quite many things are packed into a single course, and I use multiple sources when I create teaching materials, which means if you miss a class and try to learn things on your own, it's going to be a very tough course.
2. All lectures will be held in person. Joining via Zoom is only allowed if you're in a special circumstance (e.g. tested COVID positive)
3. You are expected to attend every lecture, to arrive on time, and to participate in all class activities.
4. **If you miss more than three lectures without a legitimate excuse, you will get an F grade.**
5. If you arrive more than 15 minutes late, you will be counted absent.
6. If you have a legitimate reason to miss lectures, please notify the instructor in advance, if possible. In any event, you are responsible for any work that you miss, and missing class is no excuse for not turning in an assignment.

3

Live Demo

- Whenever you see this icon, that means I'm going to do live coding or show video clips.

