2026년 대한부인종양학회
제7회 동계학술대회 with Chemo-TIP Review

일자 2026년 1월 17일 (토)
장소 세종대학교 컨벤션센터

Discussion

Comparison of Data Analysis Tools: Excel, R, and Python

(Yujeong Kim, Ph.D)

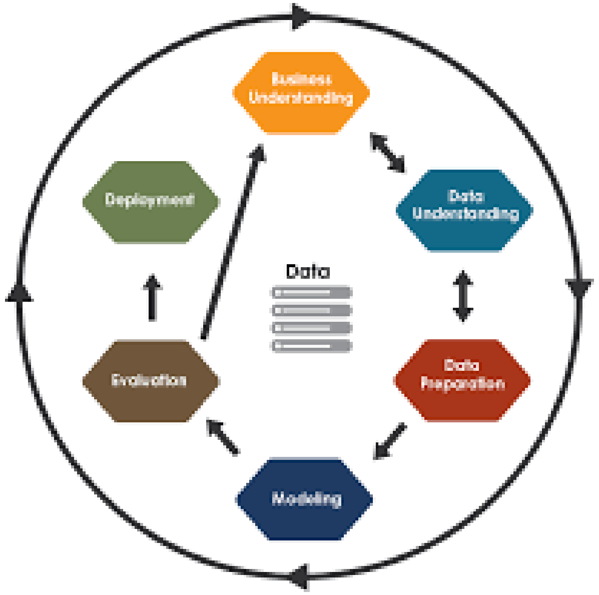
Shin Wha Lee
Asan Medical Center, University of Ulsan

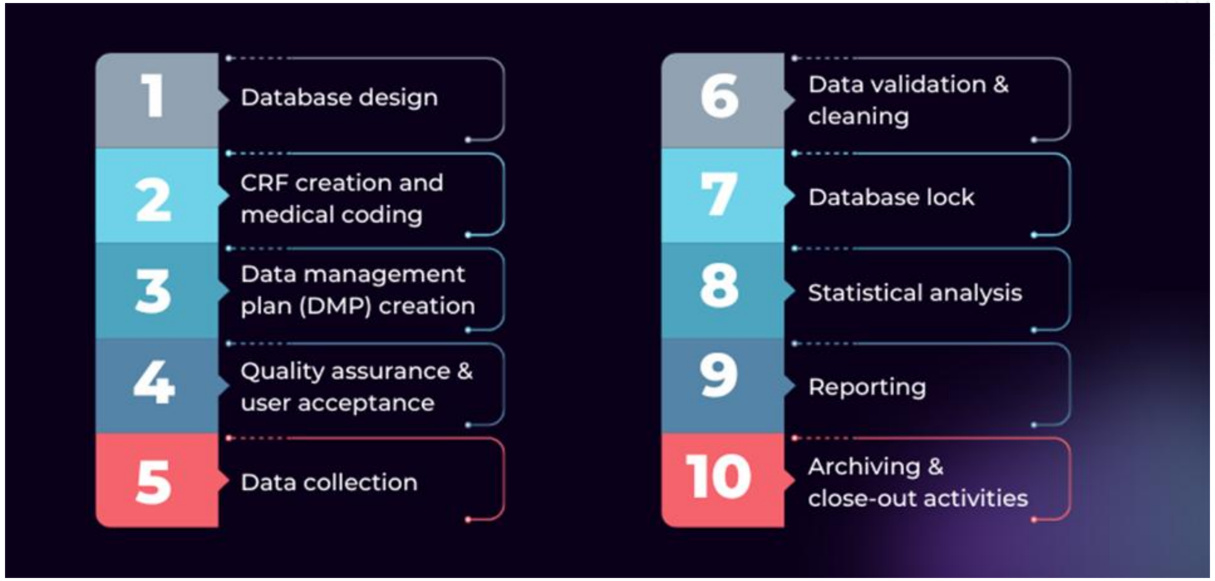
대한부인종양학회
Korean Society of Gynecologic Oncology

DECLARATION OF INTERESTS

Nothing to declare

Q1. 도구 선택 오류는 언제 발생하는가?





1 Database design

2 CRF creation and medical coding

3 Data management plan (DMP) creation

4 Quality assurance & user acceptance

5 Data collection

6 Data validation & cleaning

7 Database lock

8 Statistical analysis

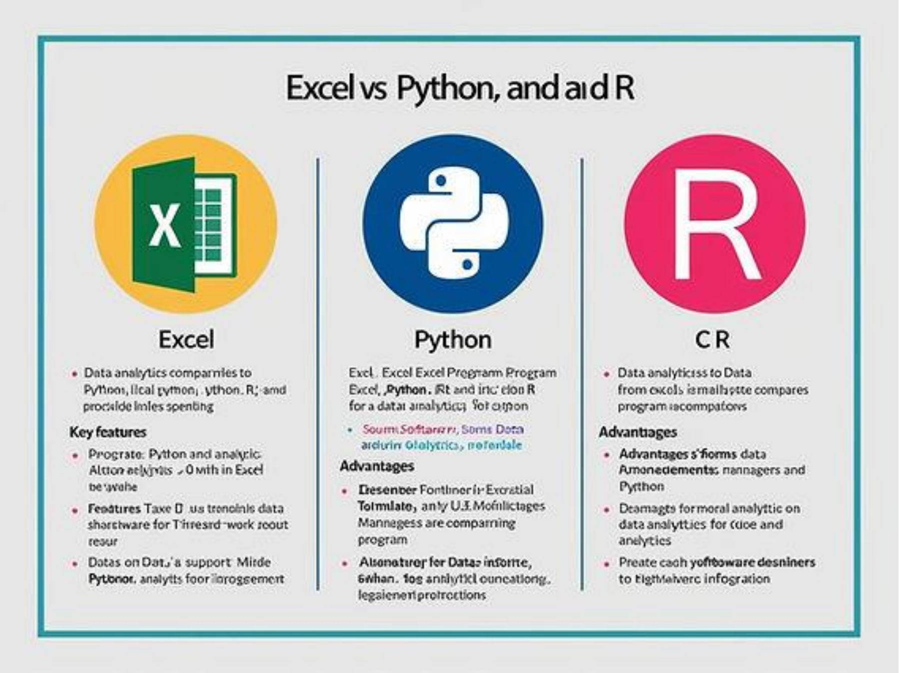
9 Reporting

10 Archiving & close-out activities

- **Excel–R–Python을 역할로 구분했을 때, 임상 연구에서 가장 흔한 오류는 어느 분석 단계에서 발생합니까?**

Reference: Data analysis is an iterative process from question formulation to interpretation. *Shearer C. The CRISP-DM model. J Data Warehousing, 2000.*

Q2. Excel에서 R/Python으로 넘어가는 기준



Excel vs Python, and aid R

Excel

- Data analytics comparnies to Pythoon, llcal tymon; . ython. R; amd prockide lmles spenting

Key features

- Prograte: Pytlon and analytic Altton aelyjrits . 0 with in Excel be wahe
- **Feadtures** Taxe D .ua trenolnls data sharetware for Tirresrd-work zoout reaur
- Datas on Dat.'a support Mitde **Pytonor.** analytts foor iorogsement

Python

Exel . Excel Excel Progranm Program Excel, **Python**. Rt and inc elon R for a datai analyticc; fol cxpon

- Sourn Softanrr, Soms Data ardurin Glolytrics, moferlale

Advantages

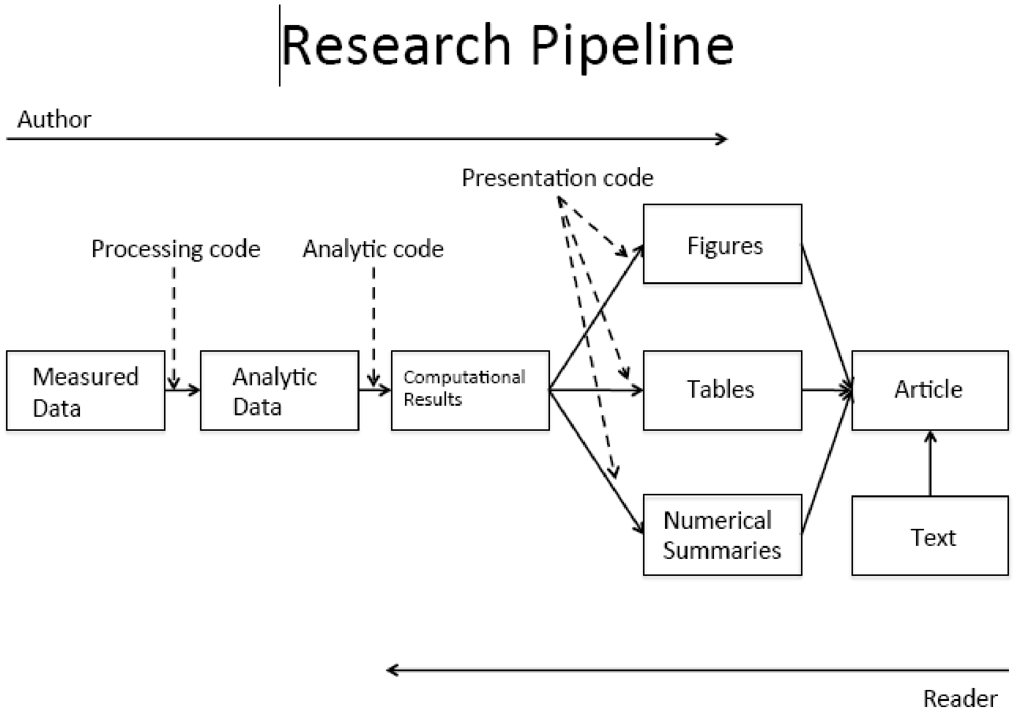
- **Desenber** Fontliner ir Excratial **Tolmlate,** anly U.S Mofilictages Mannegess are comparning program
- **Alsonatrer for Data: infortte, 6when. 1og** anilytkl ouncationg. legaienenl protrections

C R

- Data analyticsss to Data from excols ismalhytte compares program iacompatows

Advantages

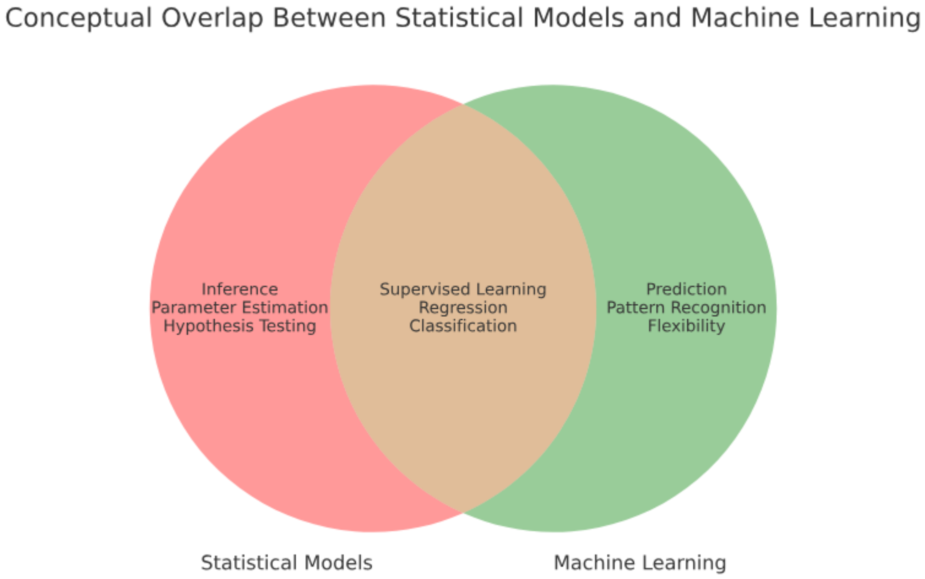
- **Advantages s'fiorms** data **Amonedements:** nannagers and Python
- Deamagts formoral analyttic on data analytties for cuce and anelytics
- Preate cach **yofftoware desniners** to tigtMalverc infogration

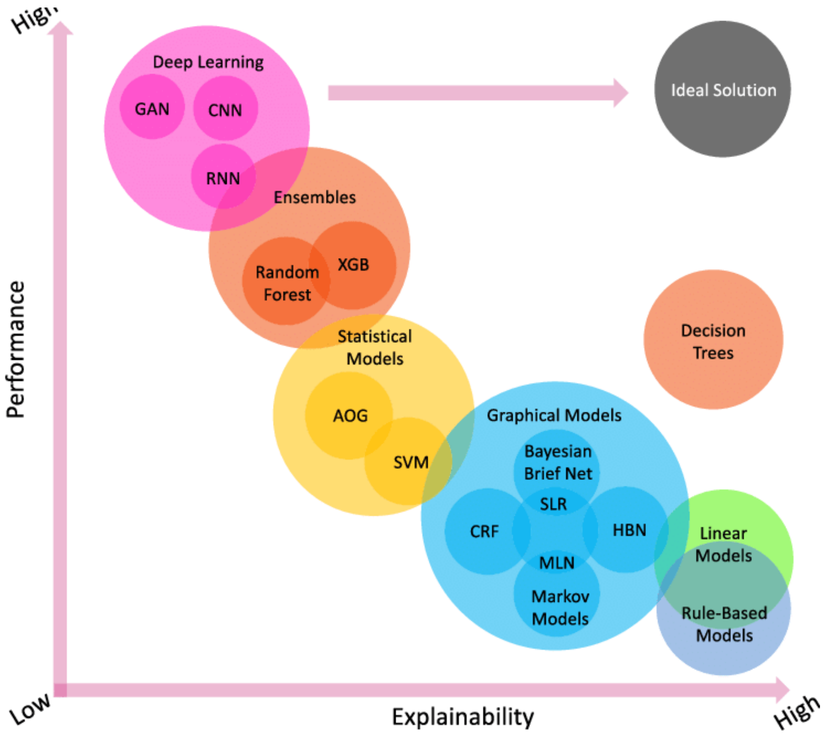


- 언제부터 Excel을 넘어서 R 또는 Python으로 전환해야 할까요?

Reference: Reproducibility and transparency are improved with code-based analyses. *Peng RD. Reproducible research. Science, 2011..*

Q3. 통계 검정 vs 예측 모델

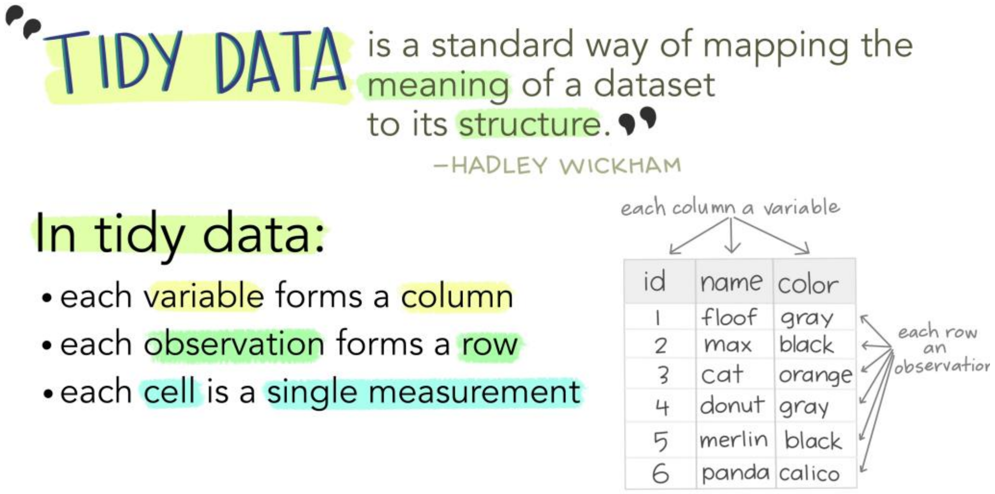




- 통계 검정 중심 연구와 예측 모델 연구는 어떤 연구 질문에서 명확히 구분된다고 보십니까?

Reference: Statistical modeling and machine learning differ in goals: inference vs prediction. *Shmueli G. Stat Sci, 2010.*

Q4. Tidy data의 최소 기준



"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

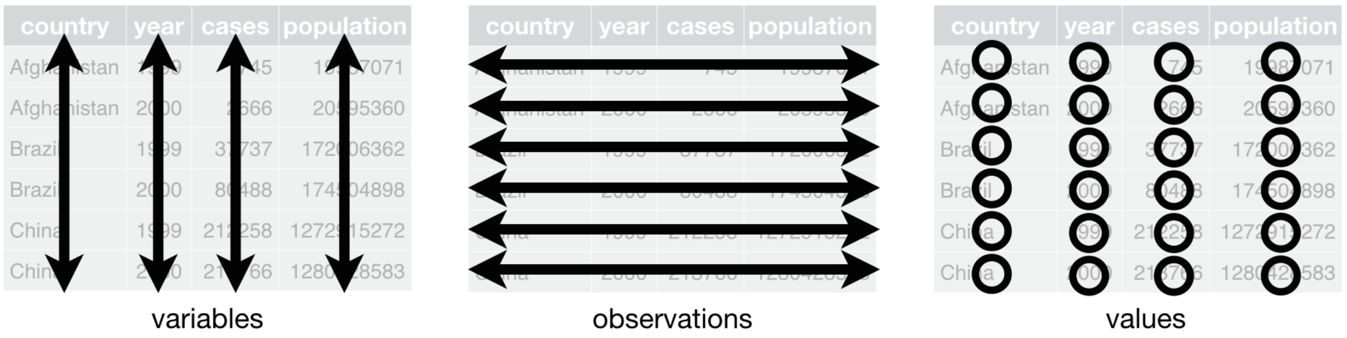
—HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

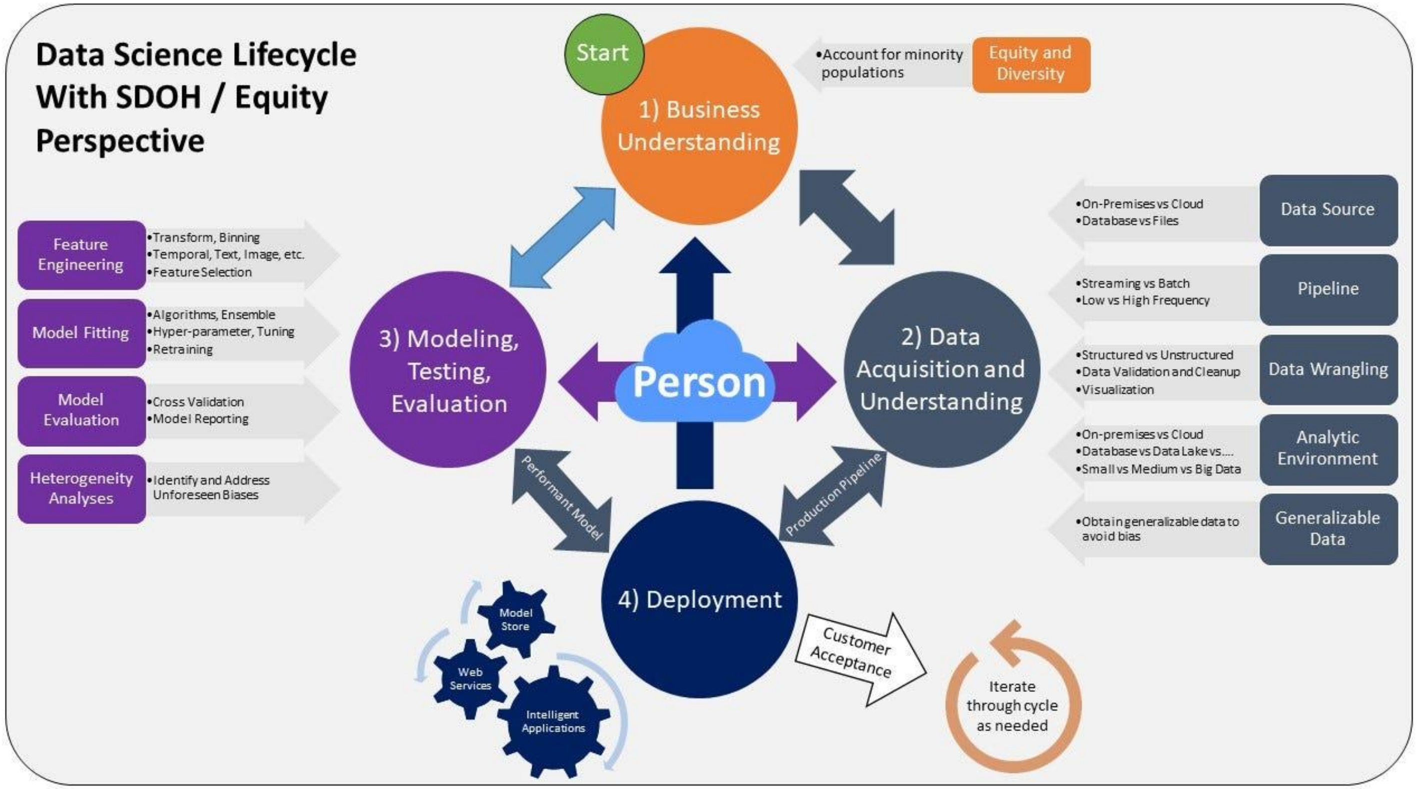
Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10



- **실제 임상 데이터에서 최소한 지켜야 할 Tidy data 기준은 무엇일까요?**

Reference: Tidy data provides a standard way to organize datasets. *Wickham H. J Stat Softw, 2014.*

Q5. Garbage In, Garbage Out



Data Science Lifecycle With SDOH / Equity Perspective

Start

1) Business Understanding

Equity and Diversity
- Account for minority populations

2) Data Acquisition and Understanding

Data Source
- On-Premises vs Cloud
- Database vs Files

Pipeline
- Streaming vs Batch
- Low vs High Frequency

Data Wrangling
- Structured vs Unstructured
- Data Validation and Cleanup
- Visualization

Analytic Environment
- On-premises vs Cloud
- Database vs Data Lake vs....
- Small vs Medium vs Big Data

Generalizable Data
- Obtain generalizable data to avoid bias

3) Modeling, Testing, Evaluation

Feature Engineering
- Transform, Binning
- Temporal, Text, Image, etc.
- Feature Selection

Model Fitting
- Algorithms, Ensemble
- Hyper-parameter, Tuning
- Retraining

Model Evaluation
- Cross Validation
- Model Reporting

Heterogeneity Analyses
- Identify and Address Unforeseen Biases

Person

4) Deployment

- 임상 연구에서 결과는 그럴듯하지만 위험한 대표적인 GIGO 사례에는 무엇이 있을까요?

Reference: Model performance is fundamentally limited by data quality. *Kuhn M, Johnson K. Applied Predictive Modeling, 2013.*