

# 데이터 역량 개발을 위한 자생적 연구모임

## 스터디그룹 2회차

이 현 창  
디지털혁신실 디지털신기술반

2022.11.9

# Outline

---

1. 데이터 전문가
2. RAP와 효율적 데이터 분석
3. 분석 단계별 RAP
4. RAP 도구

# **데이터 전문가**

# 데이터 전문가

## 데이터 전문가는

- 다양한 데이터를 효율적으로 분석함으로써 데이터기반 의사결정을 지원

- [데이터] 여러 데이터를 빠르게 조회하고 입수하여 분석에 활용
- [효율적 분석] 데이터 입수부터 전처리, 분석, 시각화, 보고서 작성 등 전 과정을 자동화
- [의사결정 지원] 데이터 분석 결과가 새로운 데이터에 대해 얼마나 유효할 것인지 평가
- 현재, 효율적인 분석 프로세스 공유 및 축적을 위한 여건이 마련되어 있지 않고, 데이터 분석은 주로 표본내 설명력을 높이는데 초점

# 데이터 기반 의사결정

## 의사결정

- 데이터 분석의 표본외 예측력(out-of-sample predictability)을 제고함으로써 데이터 기반 의사결정을 지원
  - [과적합] 기존 분석은 대체로 관측된 변수의 상호관계에 대한 추론, 즉 표본내 설명(in-sample description)에 초점, 분석 결과가 표본외 데이터에서 얼마나 유효한지에 대한 평가없이는 표본외 예측력을 개선하기 어렵고 의사결정에도 활용하기 어려움
  - [비선형, 상호의존성] 대부분 계량경제모형은 변수 간 선형 관계를 가정함에 따라, 결과를 설명하기는 좋지만 예측력이 낮아질 가능성

# 효율적 데이터 분석

## 효율적 분석

- 분석 프로세스의 모듈화, 자동화, 문서화를 통해 데이터 분석을 효율화
  - 현재 데이터 분석 업무는 개인 역량과 스타일, 업무 관행에 크게 의존
  - [모듈화] 전체 프로세스를 입력과 출력이 구분되는 단계들로 구분하고, 각 단계를 수행하는 코드(모듈)를 작성하거나 라이브러리에서 호출
  - [자동화] 데이터 입수부터 보고서 작성까지의 각 모듈을 독립적으로 수정하고, 전체 프로세스를 일괄 실행할 수 있는 분석 환경 활용
  - [문서화] 각 모듈에 대한 설명과 함께 작성된 분석 파일(코드, 데이터)은 손쉽게 공유되고 업데이트되며 한국은행 지적 자산으로 축적

# 데이터 전문가를 위한 디지털 신기술

---

## 데이터 입수    효율적 분석    의사결정

---

1. 데이터 플랫폼(BReiT)	●	○	○
2. RAP(파이썬/노트북)	○	●	○
3. 데이터 분석 라이브러리	○	●	○
4. AI/ML	○	○	●
5. 모형 검증	○	○	●

---

Note: ●는 큰 도움, ○는 작은 도움

# RAP와 효율적 데이터 분석

# 데이터 분석 과정

---

1. 주제(질문) 설정
2. 기존 연구 조사(이론, 가설, 모형, 데이터)
3. 데이터 수집 및 탐색적 자료분석(전처리, 기초통계, 시각화)
4. 모형 분석(계량경제, ML, 모형 검증)
5. 발표, 보고서 작성 및 리뷰
  - [1.] - [5.] 반복
6. 대외 발간, 분석 코드 및 결과 공개

# 기존 데이터 분석xlsx, csv

Excel Tutorial 1 - Excel

	A	B	C	D	E	F	G	H	I	J	K	L
1		증정	제품	금액								
2	1	A	1000		공장별	I	6000					
3	1	C	2000			Z						
4	1	A	1000			3						
5	1	B	2000		제품별	A	6000					
6	2	D	1000			B						
7	2	A	2000			C						
8	2	C	1000		공장별 제품별	I-A	2000					
9				3000								

Excel Tutorial 3 - Excel

	A	B	C	D	E	F	G	H	I	J	K	L
1	0	0.004321	0.0086	0.012837	0.017033	0.021189	0.025306	0.029384	0.033424	0.037426		
2	1.1	0.041393	0.045323	0.049218	0.053078	0.056905	0.060698	0.064458	0.068180	0.071882	0.075547	
3	1.2	0.079181	0.082785	0.0863	0.089900	0.093422	0.096911	0.100371	0.103800	0.10721	0.11059	
4	1.3	0.113943	0.117271	0.120574	0.123852	0.127105	0.130334	0.133539	0.136721	0.139879	0.143015	
5	1.4	0.146128	0.149219	0.152288	0.155336	0.158362	0.161368	0.164353	0.167317	0.170262	0.173186	
6	1.5	0.176091	0.178977	0.181844	0.184691	0.187521	0.190332	0.193125	0.1959	0.198657	0.201397	
7	1.6	0.20412	0.206826	0.209515	0.21218	0.214844	0.217484	0.22010	0.222716	0.225309	0.227887	
8	1.7	0.230449	0.232996	0.23552	0.23804	0.240549	0.24303	0.245513	0.247973	0.25042	0.252853	
9	1.8	0.255273	0.257679	0.260071	0.262451	0.264818	0.267172	0.269513	0.271842	0.274158	0.276462	
10	1.9	0.287875	0.281033	0.283301	0.285557	0.287802	0.290003	0.292256	0.294466	0.296665	0.298853	
11												

Excel Tutorial 4 - Excel

	A	B	C	D	E	F	G	H	I	J	K	L
1			A	B	C	D						
2	단가	100	200	300	400							
3	수량	1	2	3	4							
4	합계	100	400	900	1600							
5			단가	수량	합계							
6			100	200	300	400						
7												

Excel Tutorial 2 - Excel

	A	B	C	D	E	F	G	H	I	J	K	L
1	구분 1	구분 2	값									
2	A	D	100		전체 평균	550						
3	A	D	200		400 이상 값의 평균	700						
4	A	E	300		A의 평균	200						
5	B	E	400		A & D의 평균	150						
6	B	F	500									
7	B	F	600									
8	C	G	700									
9	C	G	800									
10	C	H	900									

# 기존 데이터 분석xlsx, csv, .. handicraft

기존 데이터 분석 xlsx, csv, .. handicraft

기존 데이터 분석 xlsx, csv, .. handicraft

기존 데이터 분석 xlsx, csv, .. handicraft

글자	제작	설비
A	100	300
B	2	3
C	1000	2000
D	1000	2000
E	1000	2000
F	1000	2000
G	1000	2000
H	1000	2000
I	1000	2000
J	1000	2000
K	1000	2000
L	1000	2000
M	1000	2000
N	1000	2000
O	1000	2000
P	1000	2000
Q	1000	2000
R	1000	2000
S	1000	2000
T	1000	2000
U	1000	2000
V	1000	2000
W	1000	2000
X	1000	2000
Y	1000	2000
Z	1000	2000

	5	6	7	8	9
1	0.025306	0.029384	0.033424	0.037426	
2	0.064458	0.068186	0.071882	0.075547	
3	0.106371	0.103880	0.10721	0.11059	
4	0.133539	0.136721	0.139879	0.143015	
5	0.164353	0.167317	0.170262	0.173186	
6	0.191125	0.1959	0.198657	0.201397	
7	0.220510	0.222716	0.225309	0.227787	
8	0.245513	0.247973	0.25042	0.252853	
9	0.269513	0.271842	0.274158	0.276462	
10	0.292256	0.294466	0.296665	0.298853	

	A	B	C	D
1	단가	100	200	300
2	수량	1	2	3
3	합계	100	400	900
4				
5	단가	수량	합계	
6				
7				
8				
9				
10				

# 기존 데이터 분석 NOT scalable!

---

데이터 분석의 확장이 어려움

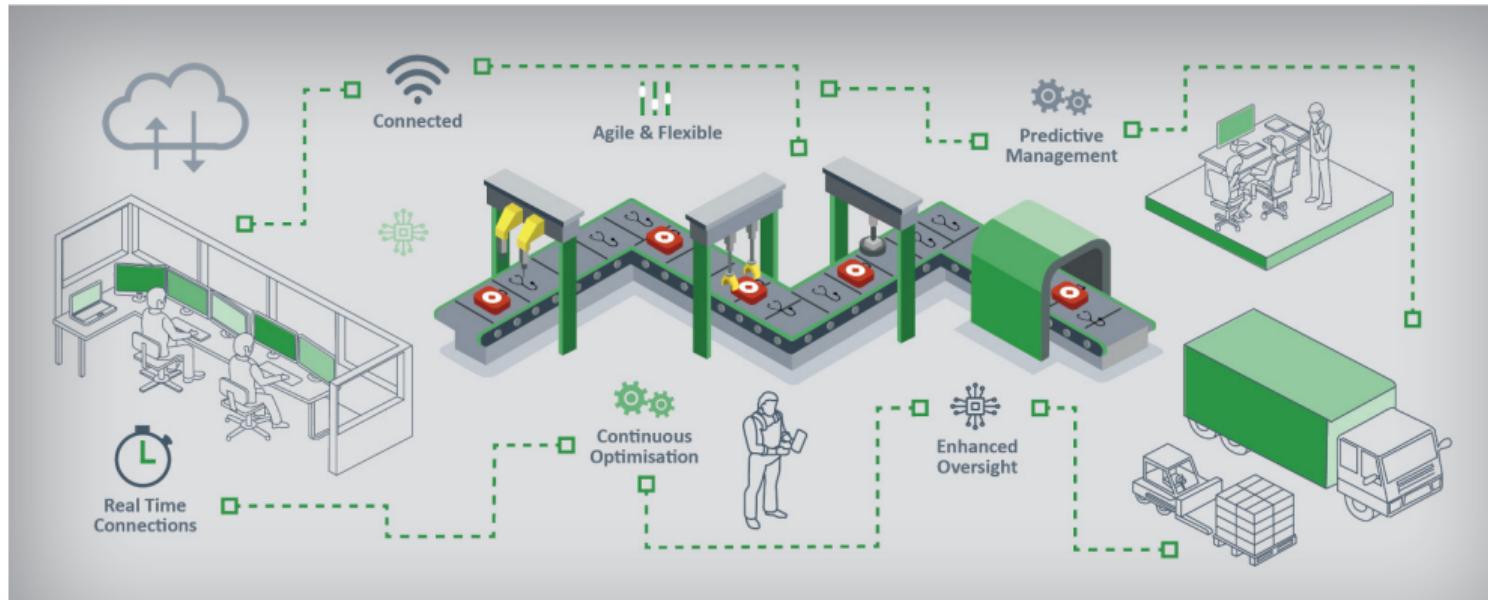
- 변수 추가 및 변환, 표본기간 변경시 동일한 시간과 수고가 필요
- 분석 프로세스를 다른 연구 과제에 적용할 수 없음

데이터 분석 결과를 업데이트하거나 공유하기 어려움

- 데이터 분석 절차에 대해 따로 문서화하지 않는 경우, 분석 결과를 재현하는 것이 현실적으로 불가능

# 스마트 팩토리 Smart Factory

## 제조업의 애자일 생산 공정



# 디지털 혁신과 중앙은행 한은소식 2019년 1월호

## 중앙은행의 애자일 생산(데이터 분석) 공정은?

금융연구 트렌드 · 이현창 금융안정연구팀 과장

디지털 혁신과  
중앙은행



# RAP 기반 애자일 분석

---

데이터 분석의 강건성과 생산성 제고를 위한 방법론

- replication crisis and reproducible research
- RAP champions

데이터 입수부터 전처리, 분석, 시각화 등 전 과정을 모듈화.  
자동화함으로써 RAP 구현(스마트 팩토리의 컨베이터 벨트)

- 모든 모듈이 서로 연계되어 원클릭으로 전 과정을 실행
- 데이터에 이상이 있는 경우 애자일하게 이전 단계 모듈을 점검하고 재실행

# RAP 기반 애자일 분석 - WIOT 사례

In [95]:

```
import pandas as pd
import numpy as np
from IPython.display import IFrame, Image
idx = pd.IndexSlice
```

레온티에프 역행렬 함수

투입산출표

중간투입행렬  $Z$ , 최종수요행렬  $F$ , 부가가치벡터  $V$ , 충산출ベ터  $X$

$$\begin{bmatrix} Z \\ V' \\ X' \end{bmatrix}, \text{where } Z = \begin{bmatrix} Z_{c-m,c-m} & Z_{c-m,c-s} & Z_{c-m,k-m} & Z_{c-m,k-s} \\ Z_{c-s,c-m} & Z_{c-s,c-s} & Z_{c-s,k-m} & Z_{c-s,k-s} \\ Z_{k-m,c-m} & Z_{k-m,c-s} & Z_{k-m,k-m} & Z_{k-m,k-s} \\ Z_{k-s,c-m} & Z_{k-s,c-s} & Z_{k-s,k-m} & Z_{k-s,k-s} \end{bmatrix} \text{ and } F = \begin{bmatrix} F_{c-m,c-c} & F_{c-m,c-i} & F_{c-m,k-c} & F_{c-m,k-i} \\ F_{c-s,c-c} & F_{c-s,c-i} & F_{c-s,k-c} & F_{c-s,k-i} \\ F_{k-m,c-c} & F_{k-m,c-i} & F_{k-m,k-c} & F_{k-m,k-i} \\ F_{k-s,c-c} & F_{k-s,c-i} & F_{k-s,k-c} & F_{k-s,k-i} \end{bmatrix}$$

레온티에프 역행렬:  $(1 - A)^{-1}$

$X = AX + f$  with  $A_{i,j} = Z_{i,j}X_j$  and  $f_i = \sum_j F_{i,j}$

$X = (I - A)^{-1}f$

In [96]:

```
def Leon(A):
    return pd.DataFrame(np.linalg.inv(np.identity(A.shape[0]) - A),
                        index=A.index, columns=A.columns)
```

주피터노트북에서 수식 편집

In [97]:

```
url = ('http://jupyter-notebook.readthedocs.io/en/latest/examples/Notebook/Typesetting%20Equations.html')
IFrame(url, width=900, height=450)
```

Out[97]:

- Motivating Examples
- The Lorenz Equations

Docs » Notebook Examples » Motivating Examples [Edit on GitHub](#)

# RAP 기반 애자일 분석 Scalable!

---

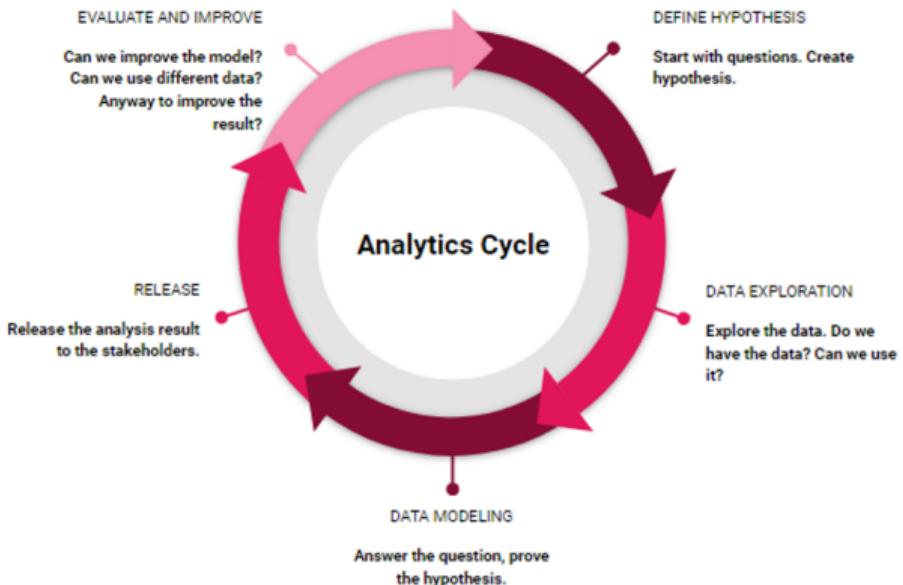
데이터 분석의 확장이 쉬움

- 변수 추가 및 변환, 표본기간 변경시 추가적인 시간과 수고가 없음
- 분석 프로세스를 다른 연구 과제에 간단히 적용

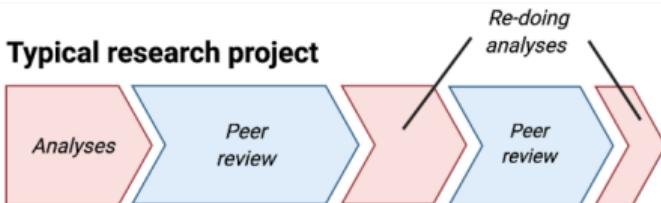
데이터 분석 결과를 업데이트하거나 공유하기 쉬움

- 추가적인 작업 없이 원클릭으로 분석 결과를 업데이트하거나 공유

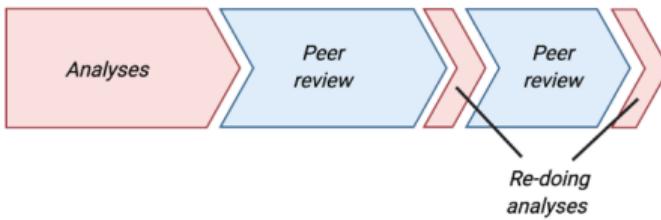
# [참고] 애자일 분석 Agile Analytics



## Typical research project



## Research project using reproducible practices



# **분석 단계별 RAP**

# 전처리·정제

---

입수한 데이터를 분석에 용이한 구조로 정형화하고 정제

- 테이블 만들기(행과 열 라벨 지정)
- 데이터 정제(결측치 처리, 변수타입 변환, 국가코드 통일 등)
- 데이터 구조화(시계열, 멀티인덱스 등)
- 데이터 결합(BIS + IMF + Bloomberg + ...)
- 그룹 리스트 정의(국가, 지역, 업권 등)
- 빈도 변환(일 ↔ 월 ↔ 분기 ↔ 연)

# 입수 데이터 형태 keb, ecos vs bis

## 데이터의 각 행과 열에 의미를 쉽게 파악할 수 있는 라벨을 지정

	A	B	C	D	E	F	G	H	I	J
1	* 통계명 :	월간 매매가격지수_종합								
2	* 수록기간 :	2003년 11월 ~ 2022년 09월								
3	* 조회기간 :	2003년 11월 ~ 2022년 09월								
4	* 종목 :	한국부동산원								
5	* 자료다운일자 :	2022.10.17 10:49:04								
6	* 단위 :	지수								
7										
8										
9										
10										
11	지 역	2003년 11월	2003년 12월	2004년 01월	2004년 02월	2004년 03월	2004년 04월			
12	전국	61.4518237	60.9631845	60.6520091	60.7438288	60.8423757	60.9356494			
13	수도권	55.765538	55.3591949	55.1164374	55.2289581	55.3764484	55.5114429			
14	지방권	68.358855	67.7607968	67.3565003	67.416319	67.441927	67.47331			
15	6대광역시									
16	5대광역시									
17	9개도									
18	8개도									
19	서울									
20	강북지역									
21	도심권									
22	종로									
23	중구									
24	용산									
25	동북권									
26	성동구									
27	광진구									
28	동대문구									
29	중랑구									
	Sheet1	Sheet2	Sheet3							

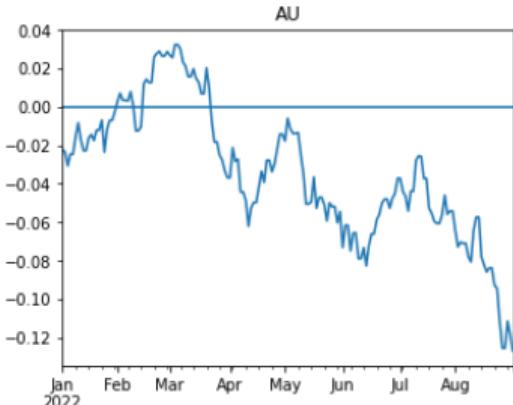
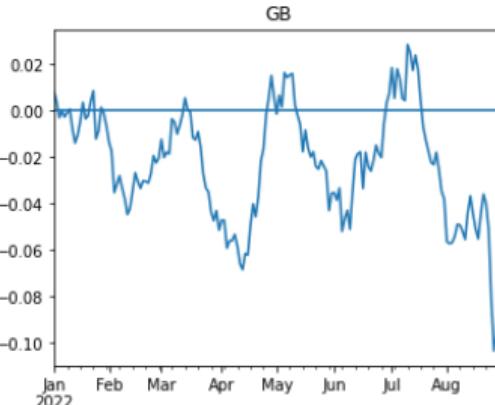
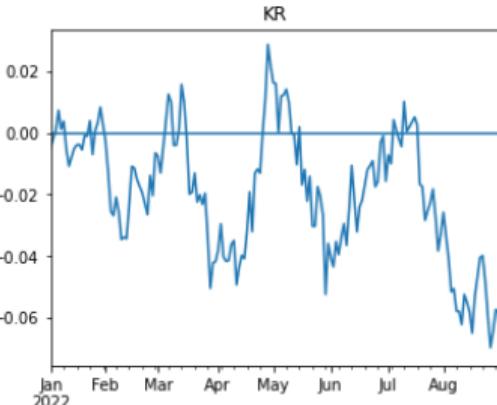
	A	B	C	D	E
	Residential property prices selected - Nominal - Index, 2010 = 100	Residential property prices selected - Nominal - Year-on-year changes, in per cent	Residential property prices selected - Real - Index, 2010 = 100	Residential property prices selected - Real - Year-on-year changes, in per cent	
	<a href="#">Back to menu</a>				
1					
2	Index, 2010 = 100 (-)	Year-on-year changes, in per cent (-)	Index, 2010 = 100 (-)	Year-on-year changes, in per cent (-)	
3	Emerging market economies (aggregate)	Emerging market economies (aggregate)	Emerging market economies (aggregate)	Emerging market economies (aggregate)	
4 Period	Q:4:T:N:628	Q:4:T:N:771	Q:4:T:R:628	Q:4:T:R:771	
374	30.06.2019	160.6871	5.3621	114.812	1.8126
375	30.09.2019	162.5165	4.863	114.9866	1.4986
376	31.12.2019	164.1597	4.4712	114.4569	0.4736
377	31.03.2020	165.925	4.6423	114.3476	0.0907
378	30.06.2020	167.7486	4.3946	116.1724	1.1849
379	30.09.2020	169.3663	4.2148	116.2018	1.0568
380	31.12.2020	171.8334	4.6745	117.011	2.2315
381	31.03.2021	174.2555	5.0206	117.2152	2.5078
382	30.06.2021	177.2083	5.6392	118.378	1.8986
383	30.09.2021	179.6238	6.0564	118.8758	2.3012
384	31.12.2021	183.3819	6.7208	119.0632	1.7539
385	31.03.2022	188.6016	8.2328	119.1865	1.6818
386	30.06.2022				

# 시각화·기초통계

빠르게 데이터를 개관하고 오류·이상치를 식별하기 시각화 및 기초통계를 활용

```
fig, axs = plt.subplots(1, 4, figsize=(24, 4))

for ax, co in zip(axs, ['KR', 'GB', 'AU', 'CA']):
    er.loc['2022', co].plot(ax=ax)
    ax.set_title(co)
    ax.axhline(y=0)
```



# [참고] 파일 연수 자료

github.com/hyunchangyi/python101

Product Solutions Open Source Pricing Search Sign in Sign up

hyunchangyi / python101 Public Notifications Fork 0 Star 3

Code Issues Pull requests Actions Projects Security Insights

main 1 branch 0 tags Go to file Code

hyunchangyi Delete GDP nowcasting_202107.pdf	e26056f on 22 Jul 116 commits
input Delete open_in_colab.png	11 months ago
script Update plots.py	11 months ago
EWS_slides_20220718.pdf Add files via upload	3 months ago
Ensemble.ipynb Colaboratory를 통해 생성됨	3 months ago
Python_for_BigData.pdf Add files via upload	3 months ago
README.md Update README.md	3 months ago
RF.ipynb Colaboratory를 통해 생성됨	3 months ago
intro.ipynb Rename python_intro.ipynb to intro.ipynb	3 months ago
intro_new.ipynb Colaboratory를 통해 생성됨	3 months ago
lecture_note.pdf Add files via upload	3 months ago
pandas.ipynb Colaboratory를 통해 update	11 months ago
preprocess.ipynb Colaboratory를 통해 생성됨	3 months ago
test.ipynb Update test.ipynb	11 months ago
wiot.ipynb Colaboratory를 통해 update	11 months ago
파이썬 신입직원연수(2022).pdf Add files via upload	3 months ago

About  
No description, website, or topics provided.

Readme  
3 stars  
2 watching  
0 forks

Releases  
No releases published

Packages  
No packages published

Languages  
Jupyter Notebook 100.0%

README.md

23 / 35

# 모형 분석

---

계령경제모형, ML 등 분석 모형의 입력 데이터 생성을 자동화하여 모형 분석 및 검증을 빠르게 수행

## 모형 분석

- 데이터 패턴을 잘 설명하는 모형 탐색
- 표본기간 및 빈도, 변수 생성, 모형 설정(입력변수 구성, 하이퍼파라미터)

## 모형 검증

- 분석 결과의 강건성을 검증
- 표본기간 및 빈도, 변수 생성, 모형 설정(입력변수 구성, 하이퍼파라미터)

# 모래밭에서 바늘찾기... 검증

---



# RAP 도구

# RAP 도구

---

여러 데이터 분석 단계를 효율적으로 연계할 수 있는 범용 프로그래밍 언어와 분석 환경

## 파이썬과 주피터 노트북/랩

- 파이썬은 배우기 쉽고 데이터 입수부터 분석까지 모든 기능을 지원
- 주피터 랩은 단계별 점검·수정·실행 및 일괄 실행을 위한 환경

## 플랫폼

- PC(anaconda), BReiT(모델허브), Colab(구글)

# 파이썬

---

단 하나의 프로그래밍 언어를 배운다면 파이썬!

- 범용 프로그래밍 언어(glue language)

데이터 분석, ML, NLP, 계량경제모형, ABM, 웹개발 등을 위한 라이브러리

- 가독성이 좋고 배우기 쉬움
- 느린 속도가 가끔 문제가 되나 극복할 수 있음

Cython, Numba

# 주피터 노트북/랩 .ipynb

---

## 효율적인 RAP 구축을 위한 환경

- 데이터 분석의 모듈화, 자동화, 문서화가 용이
- 모듈 단위 코드 편집을 위한 다양한 편의 기능

Table of Contents, 셀 접기, 셀 편집, 대시보드 등

- 스크립트 파일(.py)을 이용하여 일부 모듈을 라이브러리처럼 활용
- R, Julia 등 프로그래밍 언어 지원
- console, spyder(R studio)

# 플랫폼

---

플랫폼에 따라 장단점이 있으므로 필요에 맞게 선택

- PC - [Anaconda python](#) 배포판 설치
- BReiT(모델허브) - 고급데이터분석환경(R, Python, MATLAB) - 내부망  
RAP 공유 및 협업을 위한 기능 개발 중
- Colab - 라이브러리 설치 및 협업이 용이  
분석 환경이 매번 초기화, 주피터 단축키 설정이 다름

# 분석 결과 공유

---

RAP(파이썬과 주피터) 프로젝트 폴더 형태로 손쉽게 공유

USD forecast

  └ data

    └ xlsx, csv, txt, ...

  └ graphs

    └ 1.png, 2.png, ...

  └ 0.ipynb

  └ 1.py

  └ 2.py

# RAP 공유 사례 - GDP nowcasting

BANK OF KOREA Model Hub

고급 데이터 분석환경  
GDP nowcasting  
주택시장 실거래가격지수  
AI Transcribe  
데이터 분석 참고자료 (18)  
R io  
데이터 입수 및 전처리  
데이터 분석 시각화  
데이터 모델분석  
데이터 분석 참고자료  
Python io  
데이터 입수 및 전처리  
데이터 분석 시각화  
데이터 모델분석  
데이터 분석 참고자료  
Stata  
EVViews

File Edit View Insert Cell Kernel Widgets Help

Not Trusted Python 3.0

### 실시간 당분기 경제전망(GDP nowcasting) 시스템

- 연구자: 디지털혁신실 이정장, 최동규, 경제연구원 김윤진, 주철  
• BOK 이슈노트[제2022-7호] 디지털 신기술을 이용한 실시간 경제전망(GDP nowcasting) 시스템 개발” [이슈노트 다운로드]
- 데이터 기반 경기호흡 판단 지표 제공  
• 수시 입수된 경기지표를 이용하여 당분기 GDP 성장을 예측하는 지표
- 동적요인모형(DFM)과 딥러닝 알고리즘(LSTM) 적용  
• DFM은 경제지표의 구조적 관계를 고려하여 일주기를 보간, LSTM은 변수간 비선형, 상호의존적 관계 표착  
• ENSEMBLE 기법(DFM과 LSTM 결합)의 풍금을 이용하여 전망성과 개선
- 파이썬/주피터노트북 및 컨테이너 기술 활용  
• 마지막 시스템 개발 및 업데이트, 연구결과 공유 및 출판에 활용  
• BReit 데이터 등수, 변수 명수, 모델 추정회수, 실시간 전망, 예측력 평가 등 각 모듈 단위 업데이트 및 실행

### 실시간 당분기 및 다음분기 경제전망

실시간 전망 결과는 매주 금요일 업데이트됩니다.  
ENSEMBLE 전망(DFM과 LSTM 전망치 평균) 주기(2022.10.24)  
전망치 및 캐드가 대외 유출되지 않도록 유의하여 주시기 바랍니다.

2022년 3분기 last updated: 2022-11-04

Jun Jun 24 Jul 08 Jul 22 Aug 05 Aug 19 Sep 02 Sep 16 Sep 30 Oct 14 Oct 28

DFM LSTM Prices Manufacturing Labor Surveys Retail and Consumption

2022년 4분기 last updated: 2022-11-04

Sep 09 Sep 23 Oct 07 Oct 21 Nov 04 Nov 18 Dec 02 Dec 16 Jan 13 Jan 27

DFM LSTM Prices International Trade Labor Surveys Retail and Consumption

32 / 35

# RAP 공유 사례 - 실시간 주택가격지수

BANK OF KOREA  
BReiT

Model Hub

File Edit View Insert Cell Kernel Widgets Help

Not Trusted Python 3

## 실시간 주택시장 실거래가격지수(BReiT/CS index)

디지털혁신실(디지털신기술)

주택 실거래가 데이터를 이용하여 주택시장 동향을 빠르게 파악할 수 있는 새로운 모니터링 지수(BReiT/CS index) 개발

- 기존 주택 가격지수는 공표시가가 길거나 실제 거래가에 기반하지 않아 현저 시장상황 파악에 한계
- BReiT/CS index는 기존 실거래가격 지수(한국부동산원, REI)와 유사한 움직임을 보이면서도 실시간으로 산출되어 속보성이 높음
- 참고자료 [\[업무정보\]](#)

```
# 주요 출연기관
# regions : 전국대상지역
# b_month : 조정기준월 2022년 2월 => pd.Timestamp('2022-2')
# e_month : 조정기준월 2022년 7월 => pd.Timestamp('2022-7')
# v_from : 반기지 시작월
# v_to : 반기지 종료월
# v_interval: 그래프에 표시되는 반기지별과 간격

run house.py
plot_by_vintage(plot_rnames=['N', 'n', 'NH'], b_month=pd.Timestamp('2022-2'), e_month=pd.Timestamp('2022-7'),
v_from='2022-10-13', v_to='2022-12-31', v_interval=3);
```

(Figure size 432x228 with 0 axes)

National      Metropolitan Area      Non-Metropolitan

10.13 10.16 10.19 10.22 10.25 10.28 10.31 11.3 11.6 REI index

지표 조회

# 2022년 향후 일정

---

- 3회차(11.23), 라이브러리 개발 계획
- 5회차(12.7), AI/ML과 모형검증
- 4회차(12.14), 통합 데이터 플랫폼 구축 현황 및 방향
- 6회차(12.21), 데이터 분석 사례

# **Q & A**