KO R EA
컨퍼런스

Begas
Best Expert Group for AnalytiS

데이터 옵스의 시작 데이터사이언스캔버스와 그 활용

베가스 박성우 연구위원, 김종대 연구위원 : 2021-11-19, 16:30 ~17:00

발표 개요

데이터 경제, 산업의 시대
기업 의사결정의 핵심요소로 데이터 & 데이터분석이 자리매김
데이터옵스 애자일/린 제조의 협업 개념과
10여년동안 제조 ·공공 ·금융 등 산업 현장에서 분석프로젝트 및
교육 경험을 체계화하여
시민과학자 양성과 분석 현장에서 활용 할 수 있는
데이터사이언스캔버스를 소개 합니다.
방법 뿐만 아니라 데이터과학도구 Jupyter와 연계를 통해
방법과 도구를 통한 분석업무의 어려움을 해소하고 효율성을
증진시키는 사례를 함께 공유 하고자 합니다.

1. DataOps ?
2. Data science canvas
3. Notebooks & Infrastructure Data science tool
4. Meet Data science canvas and Jupyter (Demo)

1. DataOps - 1.1 DataOps history

Start

1

IT 전통적인 개발 방법론

- 선형적(Linear) 접근
- 프로젝트 시작부터 최종결과물까지 **정확한 순서**로 실행 및 완료
- 요구사항분석 ~ 설계 ~ 테스팅 ~ 완료 ~ 운영

Water fall

2

소프트웨어 개발방법론

- Agile Manifesto
- **비즈니스 요구사항을 신속하게 반영** 할 수 있음
- 빠르게 테스트 하여 피드백을 얻고, 다시 제품에 반영함, Iteration

Agile

3

Agile의 보완

- **개발과 운영팀의 원활한 협업과 개발** 수명주기의 단축
- **빠르고 높은 품질의 코드 반영**을 위한 방법론 체계화

DevOps

4

Agile, DevOps, Lean Manufacturing 로 구성

- 데이터분석 개발 및 운영에 적용
- **데이터 과학, 데이터관리, 운영팀 간의 방향성 및 업무 조율**
- 비즈니스가 데이터를 최대한 활용하고 실행 가능한 통찰력을 가질 수 있도록 보장
- MLOps, ModelOps 등 으로 확대

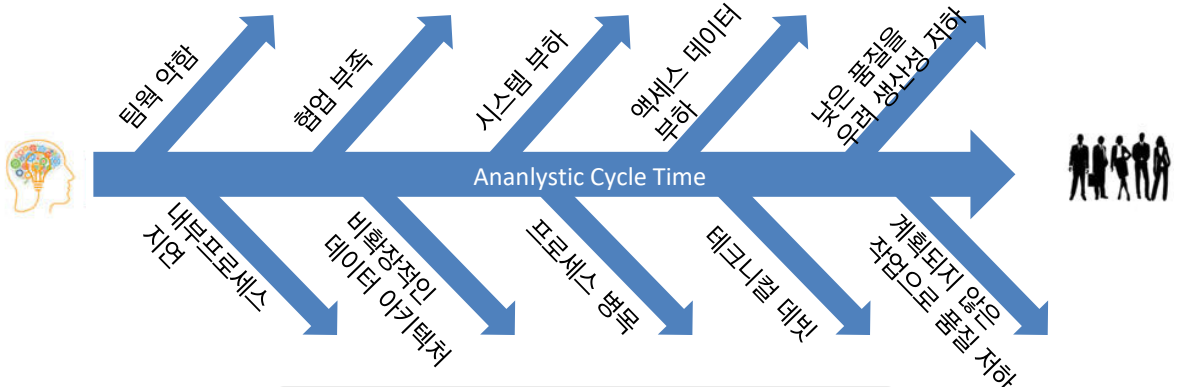
DataOPs

Present

출처 : Clover dx, datakitchen.io, Wikipedia

1. DataOps - 1.2 DataOps가 해결하려는 문제

데이터옵스는 워크플로우, 프로세스의 장애물을 제거 높은 생산성과 품질에 기여



분석 주기를 지연 시키는 장애물

- 팀웍/협업의 부재, 시스템 리소스/데이터 액세스/승인절차 대기 시간, 프로세스 병목현상, 버전으로 인한 Technical debt 등

- DataOps는 데이터 팀이 새로운 도구와 방법론을 활용할 때 품질과 주기 시간을 10배 정도 개선할 수 있음
- Celgene 제약회사는 주기 시간을 10배 개선했으며 데이터 엔지니어당 스키마 변경 수의 12배와 데이터 분석가 수의 24배를 지원할 수 있습니다

[예시] – 셀진(Celgene) 제약 데이터옵스 개선 사례

	Before DataOps	With DataOps
Schema changes per Data Engineer per week	1	12
Cycle time to publish new visualizations	weeks/months	next day
Data Analysts supported by ONE Data Engineer	0.5	12
Number of automated tests each build	0	1,000
Number of errors each build	frequent	0

DataOps 메트릭스 리포트

출처 : datakitchen.io

2. Data science canvas - 1.2 Stakeholder needs in data analytics pipeline

● 역할 ● 어려움

시티즌 데이터과학자[1], 학생	업무 전문가	데이터 과학자	데이터 엔지니어	데이터산업 분야 거래 전문가[2]
“산업과 기업 현장에서 필요로 하는 데이터분석 인력”	“데이터, 데이터분석의 업무 반영과 개선 및 활용”	“업무를 파악해서 구조, 비구조화 된 데이터를 통한 분석결과 제공 및 관리”	“데이터 파이프라인 구축 및 데이터 제공 및 관리”	“데이터경제·산업으로의 패러다임전환, 데이터 거래를 위한 신 직종”
“의욕만 가지고 데이터분석을 학습하여 현장 적용의 한계“ “현장·교육에 적절한 사례와 방법 및 도구 필요 ”	“업무 적용을 위한 분석 주제의 발굴과 활용에 시간이 걸림” “업무와 현장의 특성에 맞는 주제 발굴과 활용, 운영 방법 및 도구와 사례 필요“	“데이터수집·저장, 워크플루우 등 다양한 기술의 복잡성“ “이해관계자와 협업을 위한 방법 및 도구 필요 ”	“적절한 분석용 데이터 제공의 어려움“ “이해관계자와 협업을 위한 방법 및 도구 필요”	“데이터거래 상품 추천 및 평가를 위한 방법 및 도구 필요“

1) 주요업무가 통계, 분석 분야가 아니지만 고급 진단분석이나 예측(머신러닝)모델을 만들거나 생성하는 역할을 가짐

2) 데이터거래 상품을 발굴 추천/중개/사후관리 하는 역할과 책임을 가지며 데이터과학도구의 데이터거래 분야로 활용 할 수 있음

2. Data science canvas - 2.2 Overview

10여 년 동안 제조 · 공공 · 금융 등 기업 현장과 재직자 양성 교육 과정 체계화, 기업 현장의 업무 전문가, 엔지니어가 빠르고 쉽게 데이터 분석 프로젝트 실행 가이드

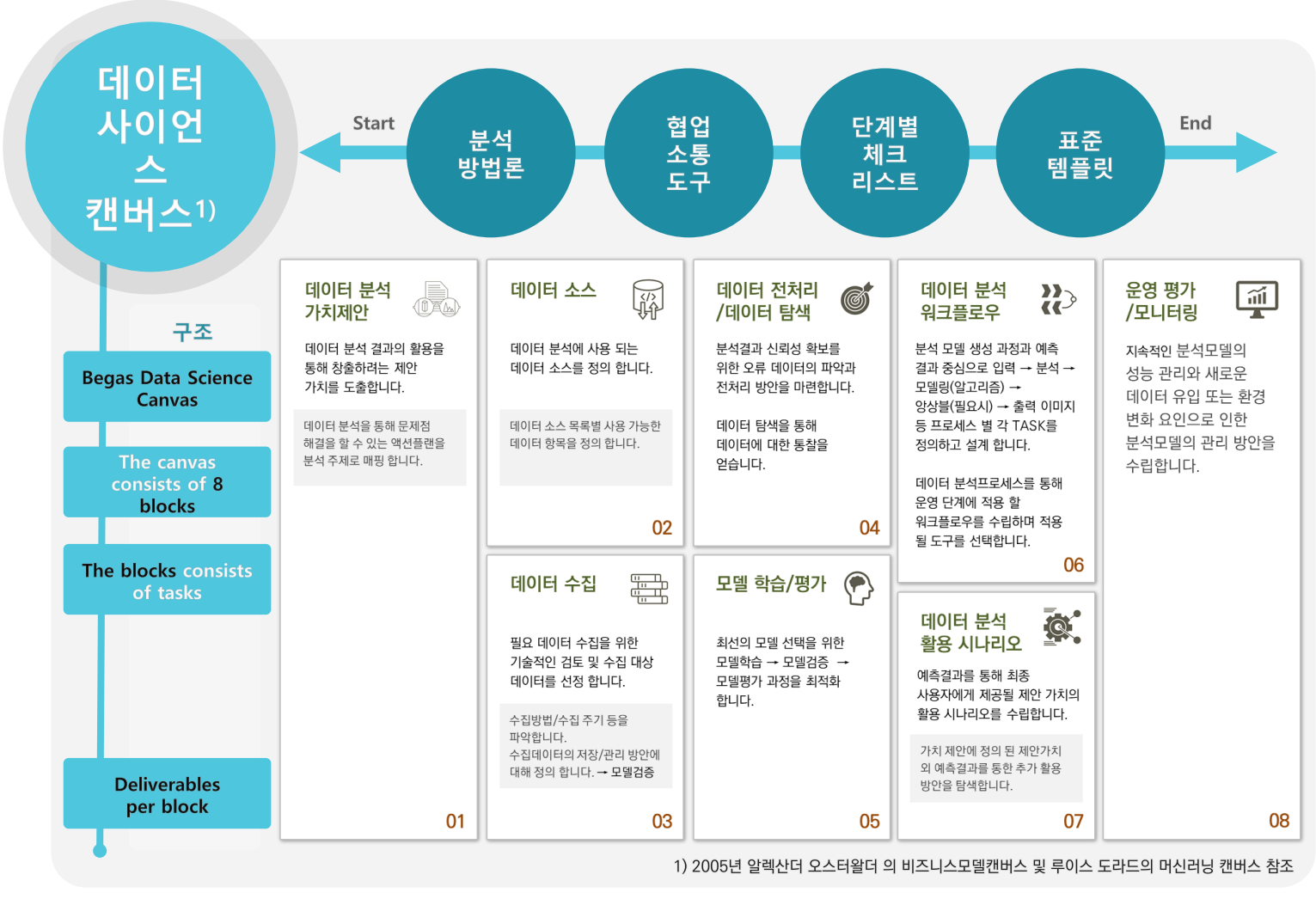
데이터 분석 방법

데이터 사이언스 캔버스란?

분석주제 ~ 운영평가/모니터링 까지 설계와 실행내역을 한 눈에 볼 수 있게 만든 데이터과학 도구 입니다.

!?

1) 시민데이터과학자, 학생
2) 업무전문가, 데이터 엔지니어
3) 데이터과학자
4) 데이터산업분야 종사자



1) 2005년 알렉산더 오스터왈더 의 비즈니스모델캔버스 및 루이스 도라드의 머신러닝 캔버스 참조

2. Data science canvas - 2.3 Purpose, User, Structure, Case Study

데이터 분석을 통해 유용한 정보를 창출, 비즈니스 문제를 해결하거나 혁신하기 위한 데이터과학자, 업무전무가, 운영자 등이 쉽고, 빠르게 이해하고 협업 할 수 있는 도구의 필요

데이터분석 가이던스 체계(방법론 및 표준산출물) 제공

관리자

『운영시』

- 빠른 인사이트 창출
- 다양한 활용 모색
- 데이터분석 쉬운 교육
- 협업·체계적인 관리

데이터분석 가이던스

데이터 전주기	분석 기법	협업도구	데이터유형·특성
모델안정성	히스토리,검증	데이터 통찰력	의사결정 인사이트 확장
체크리스트	데이터흐름중심	포인트,가이던스	운영·교육

분석참조모델

분석주제 별 참조모델 설계

가이던스 기준 사례 참조

- 데이터분석경험/노하우
- 다양한 분석기법, 사례 제공
- 산업·도메인표준모델

데이터 분석가

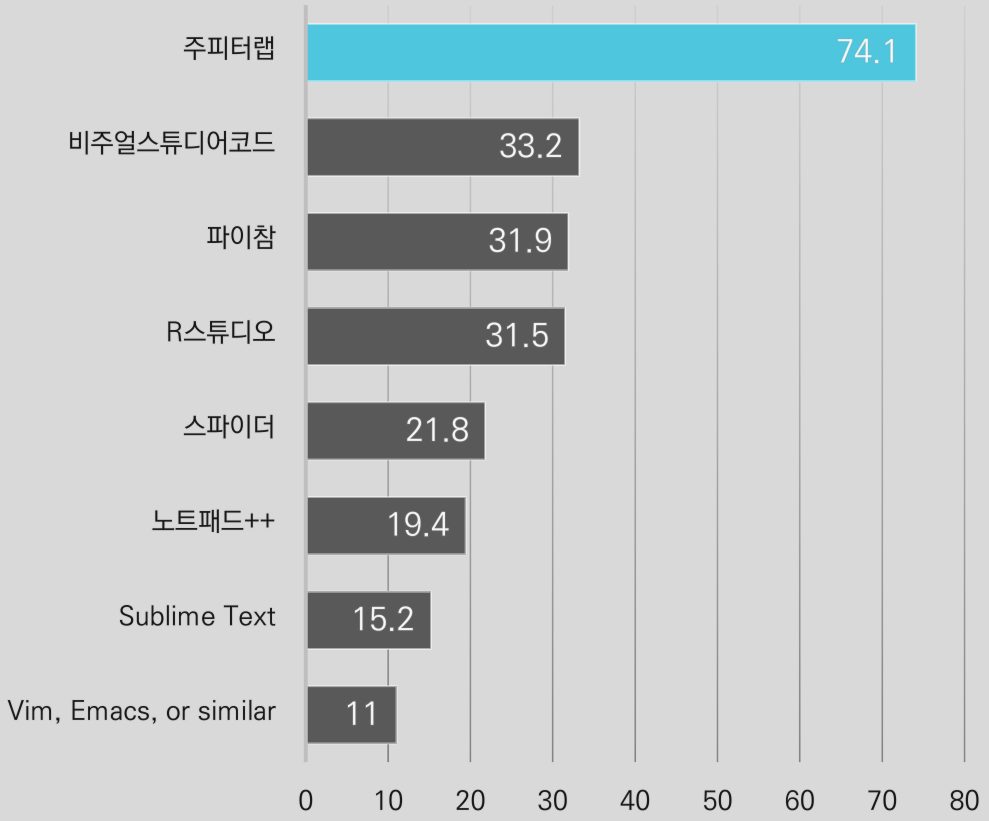
『프로젝트 수행 시』

- 효율적이고 신속한 데이터분석
- 분석모델링 품질보증
- 협업·체계적인 프로젝트 관리

3. Notebooks & Infrastructure Data science tool(1/2)

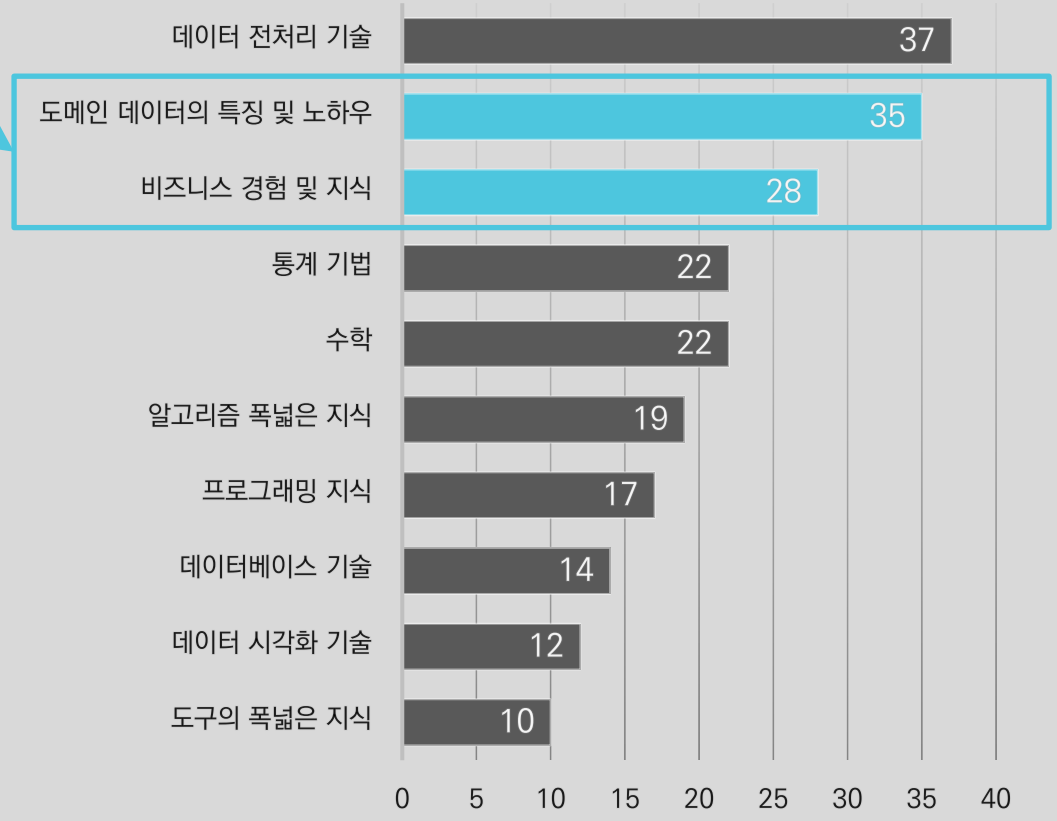
하나, 가장 선호하는 데이터과학 및 분석 도구 [주피터 노트북], 둘, 데이터과학자의 가장 중요한 스킬 [도메인 데이터의 특징 지식], [비즈니스 경험 및 지식]

데이터 과학 및 분석도구 & 인터랙티브 개발 환경



〈2020, Kaggle surveys, 응답자 다중 선택 방식 〉

데이터과학자를 위한 주요 스킬들



〈2017, kdnuggets surveys, 응답자 다중 선택 방식〉

3. Notebooks & Infrastructure Data science tool(1/2)

gradient by Paperspace

주피터 노트북 호스팅
프레임워크, 라이브러리, 드라이브 , 모델 학습 실시간 로그와 그래프 제공
초보자 무료 GPU 제공, 노트북 사용량에 따른 요금제

구글 코랩
주피터 노트북 호스팅
구글 플랫폼연동, 무료 GPUs 제공

kaggle

캐글 노트북
스크립트/주피터노트북 API 2가지 제공
트렌드, 인기, Auto 머신러닝 샘플 데이터/모델 제공

Deepnote

딥노트
주피터 노트북 지원 플랫폼, 팀/협업 지원
코드 탐색, 데이터 패턴 찾기, 코드 자동완성 등 제공
초보자에게 무료, 소규모 기업 등을 위한 노트북 요금제

플레이어들의 서비스 활성화를 위해 주피터 노트북을 채택
"Deepnote" 경우 노트북 중심의 서비스 진행

주피터 노트북, 오픈소스, R, 파이썬 등 40개 프로그래밍 언어 지원,
빅데이터 통합 및 인메모리 데이터 프로세싱을 위한 분석엔진 지원
머신러닝, pandas, scikit-learn, ggplot2, TensorFlow 등 분석, 시각화 패키지 제공

POINT 01
'21년 베스트 머신 러닝 데이터과학 도구들은 대부분 주피터 노트북 기반

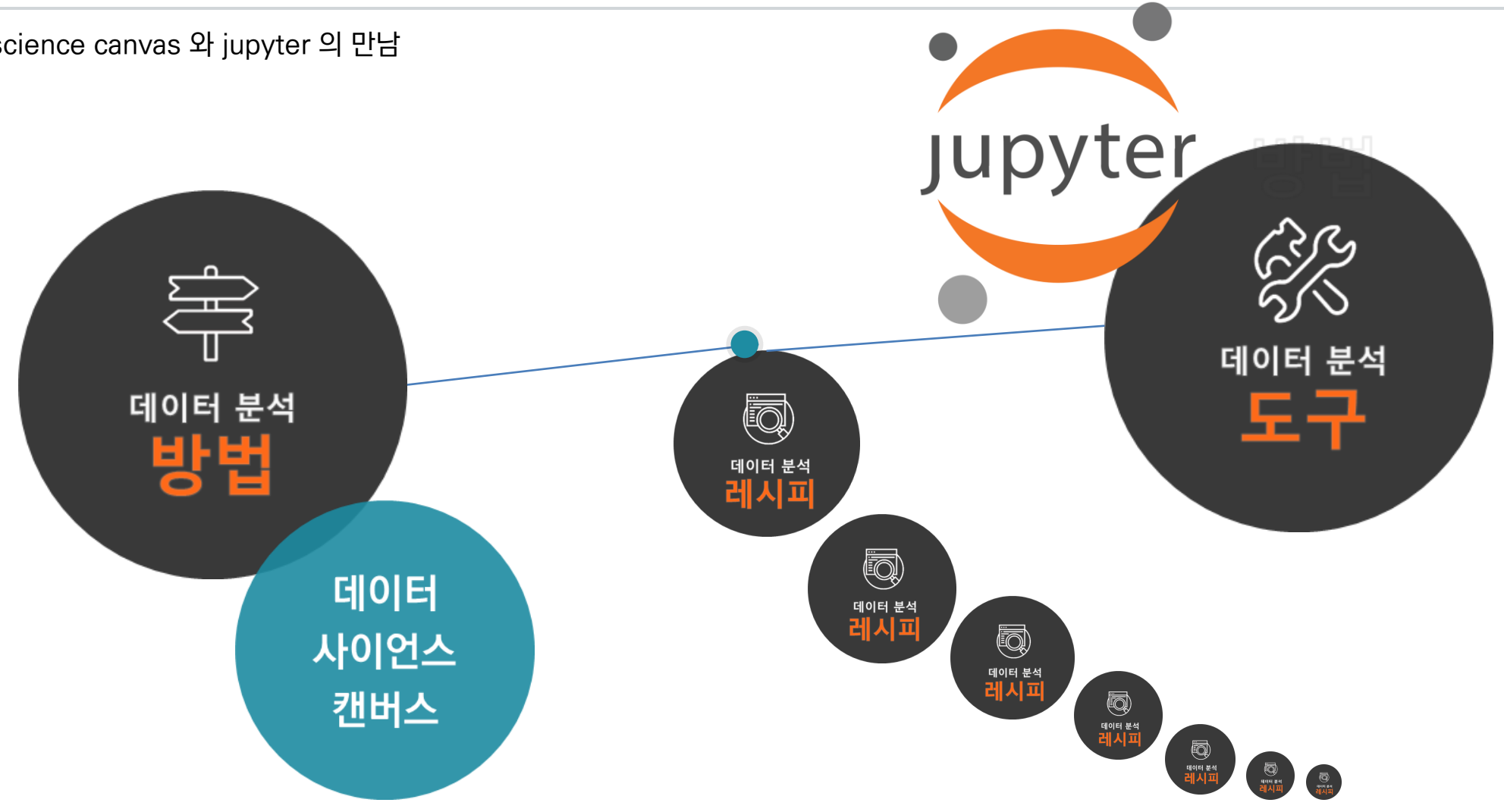
POINT 02
사업 목적에 부합하는 도구로 노트북 기능 활용(플랫폼, 클라우드, 서비스 등)

POINT 03
문서, 코드, 화면, 보고서 등 노트북으로 통합 제공

POINT 04
다양한 개발 및 교육 도구 중 접근성이 가장 뛰어남

4. 예시 - 방법과 도구를 통한 활용

Data science canvas 와 jupyter 의 만남



4. 예시 - 방법과 도구를 통한 활용

데이터 사이언스 캔버스를 주피터 노트북에 연계 사례 Screen Shot

The DataOps Manifesto

"조직, 도구 및 산업에 걸쳐 데이터로 작업하는 직접적 경험을 통해 우리는 우리가 DataOps라고 부르는 분석을 개발 및 전달 하기 위한 좋은 방법을 알아냈습니다." – 데이터옵스 선언문에서

The DataOps Manifesto

1. 지속적으로 고객을 만족시켜라
우리의 최고 우선 순위는 몇 분에서 몇 주에 이르는 **소중한 분석 통찰력의 조속하고 지속적인 전달**을 통해 고객을 만족시키는 것입니다.

2. 작동하는 분석을 소중하게 생각하라
우리는 데이터 분석 성능의 기본 척도가 강력한 프레임워크 및 시스템 위에 정확한 데이터를 통합하여 통찰력 있는 분석이 전달되는 정도라고 생각합니다.

3. 변화를 수용하라
우리는 **진화하는 고객 니즈를 환영하며 실제로 경쟁 우위를 만들어 내기 위해 이것을 수용**합니다. 우리는 고객과 소통하는 가장 효율적이고 효과적이며 기민한 방법은 마주보고 하는 대화라고 생각합니다.

4. 이것은 팀 스포츠다
분석 팀은 항상 다양한 역할, 기술, 선호하는 도구 및 제목을 갖습니다. 다양한 배경과 의견은 혁신과 생산성을 향상시킵니다.

5. 매일 일어나는 상호 작용
고객, 분석 팀 및 운영은 프로젝트 내내 일상적으로 협력 해야 합니다.

6. 스스로 조직화 하라
우리는 최고의 분석 통찰력, 알고리즘, 아키텍처, 요구 사항 및 설계는 자기 조직적 팀에서 나온다고 생각합니다.

7. 영웅주의를 줄여라
분석 통찰력 필요의 속도와 폭이 계속 증가함에 따라 우리는 **분석 팀이 영웅주의를 줄이고 지속 가능하고 확장 가능한 데이터 분석 팀 및 프로세스를 만들기 위해 노력** 해야 한다고 생각합니다.

8. 반성하라
분석 팀은 고객이 제공한 피드백, 자기 자신 및 운영 통계에 대한 **정기적 자기 반성을 통해 운영 성능을 미세 조정**해야 합니다.

9. 분석은 코드다
분석 팀은 데이터에 접근하고 데이터를 통합하고 모델링하고 시각화하기 위해 다양한 개별 도구를 사용합니다. 기본적으로, 이러한 도구 각각은 통찰력을 전달하기 위해 데이터에 대해 취해지는 조치를 설명하는 코드 및 구성을 생성합니다.

10. 결합하라
데이터, 도구, 코드, 환경 및 분석 팀 작업을 처음부터 끝까지 결합하는 것은 분석 성공의 핵심적인 요인 입니다.

11. 재현 가능하게 만들어라
재현 가능한 결과가 필요하며 따라서 우리는 데이터, 저 수준 하드웨어 및 소프트웨어 구성, 툴체인에 있는 각 툴에 **고유한 코드 및 구성 등 모든 것을 버전화** 합니다.

12. 일회용 환경
우리는 생산 환경을 반영하는 생성하기 쉽고 격리되며 안전한 일회용 기술 환경을 제공함으로써 **분석 팀원들이 실험하는 데 필요한 비용을 최소화** 하는 것이 중요하다고 생각합니다.

13. 단순성
우리는 기술적 우수성과 좋은 디자인에 대한 지속적인 관심이 민첩성을 강화한다고 생각합니다. **비슷하게 단순성–실행되지 않는 작업의 양을 극대화하는 기술–은 필수적** 입니다.

14. 분석은 제조다
분석 파이프라인은 린 제조 라인과 비슷합니다. 우리는 DataOps의 근본적인 개념이 **분석 통찰력의 생산에서 지속적인 효율성을 달성하는 것을 목표** 로 하는 과정적 사고에 대한 집중이라고 생각합니다.

15. 품질이 다른 무엇보다 중요하다
분석 파이프라인은 코드, 구성 및 데이터의 이상과 보안 문제에 대한 자동화된 탐지(지도카)가 가능한 기초를 가지고 구축되어야 하며 실수 방지(포카 요케)를 위해 운영자들에게 지속적인 피드백을 제공해야 합니다.

16. 품질 및 성능을 모니터링하라
우리의 목표는 예기치 않은 변동을 탐지하고 운영 통계를 생성하기 위해 지속적으로 모니터링되는 성능, 보안 및 품질 척도를 가지는 것입니다.

17. 재사용하라
우리는 분석 통찰력 생산성의 기초적 측면이 개인 또는 팀에 의한 이전 작업의 반복을 피하는 것이라고 생각합니다.

18. 사이클 타임을 개선하라
우리는 고객 니즈를 분석 아이디어로 바꾸고 그것을 개발에서 생성하고 그것을 반복 가능한 생산 프로세스로 배포하고 최종적으로 해당 제품을 리 팩터링 및 재사용하는 데 필요한 시간 및 노력을 최소화하기 위해 노력해야 합니다.

경청해 주셔서
감사합니다.

R 사용자회