빅데이터기획전문가
회사 내 기능조직, 비즈니스 분석 또는 BI조직에 소속되어 있으면서 빅데이터 분석 전문 조직과 협력을 통하여 업무에 필요한 분석 모델이나 예측 모델을 Self Service Analytics도구를 활용하여 구현하는 전문가

데이터베이스
문자, 기호, 음성, 화상, 영상 등 상호 연관된 다수의 콘텐츠를 정보 처리 및 정보통신 기기에 의하여 체계적으로 수집·축적하여 다양한 용도의 방법으로 이용할 수 있도록 정리한 정보의 집합체

데이터사이언스
데이터로부터 의미있는 정보를 추출해 내는 학문으로, 통계학과는 달리 정형 또는 비정형을 막론하고 다양한 유형의 데이터를 분석 대상으로 한다. 또한 분석에 초점을 두는 데이터마이닝과는 달리 분석 뿐 아니라 이를 효과적으로 구현하고 전달하는 과정까지 포함하는 포괄적인 개념

데이터웨어하우스
기업의 의사결정 과정을 지원하기 위한 주제 중심적이고 통합적이며 시간성을 가지는 비휘발성 데이터의 집합

데이터마트
데이터 웨어하우스 환경에서 정의된 접근 계층으로 데이터 웨어하우스에서 데이터를 꺼내 사용자에게 제공하는 역할을 한다. 보통 특정한 조직 혹은 팀에서 사용하는 것을 목적으로 한다.

데이터거버넌스
전사차원의 모든 데이터에 대하여 정책 및 지침, 표준화, 운용조직 및 책임등의 표준화된 관리체계를 수립하고 운영을 위한 프레임워크 및 저장소를 구축하는 것을 말한다. 특히 (마스터데이터, 메타데이터, 데이터 사전) 은 중요한 관리대상이다.

데이터사이언티스트가 갖춰야 할 역량은 빅데이터 처리 및 분석에 필요한 이론적 지식과 기술적 숙련과 관련된 능력인 (1)와 데이터 속에 숨겨진 가치를 발견하고 새로운 발전 기회를 만들어 내기 위한 능력인 (2)로 나누어진다.
(1)=**하드스킬(hard)**
(2)=**소프트스킬(soft)**

비즈니스 모델 캔버스는 9가지 블록을 단순화하여 (1),(2), 고객단위로 문제를 발굴하고 이를 관리하는 규제와 감사, (3)영역으로 나눠 분석 기회를 도출한다.
(1)=**업무**
(2)=**제품**
(3)=**지원인프라**

능력성숙도통합모델(CMMI)
소프트웨어와 시스템공학의 역량 성숙도를 측정하기 위한 모델로 소프트웨어 품질보증과 시스템 엔지니어링 분야의 품질보증 기술을 통합하여 개발된 평가모델로 1~5단계로 구성된 성숙도 모델

정보
데이터의 가공 및 상관관계간 이해를 통해 패턴을 인식 그 의미를 부여한 데이터로 지식을 추출하기 위한 것
-데이터의 가공 및 상관관계 간 이해를 통해 패턴을 인식하고 의미를 부여한 데이터
-지식을 도출할 때 사용하는 데이터

플랫폼
이것은 비즈니스 측면에서 일반적으로 '공동 활용의 목적으로 구축된 유무형의 구조물'을 의미
각종사용자 데이터나 M2M센서 등에서 수집된 데이터를 가공·처리·저장해두고 이 데이터에 접근할 수 있도록 API(Application Program Interface)를 공개한다. 그러면 다양한 서드파티사업자들이 비즈니스에 필요한 정보를 추출해 활용하게 되고 빅데이터는 그 자체로 이 역할 수행하게 된다.
-페이스북은 2006년 F8행사를 기점으로 소셜 그래프 자산을 외부 개발자들에게 공개하고 서드파티 개발자들이 페이스북 위에서 작동하는 앱을 만들기 시작하면서
( )역할을 하기 시작했다.
-하둡은 대규모 분산 병렬 처리의 업계 표준으로 맵리듀스 시스템과 분산파일시스템인 HDFS로 구성된 ( )기술이며, 선형적인 성능과 용량 확장성, 고장 감내성을 가지고 있다. 아마존은 S3와 BC2환경을 제공함으로써
( )을(를) 위한 클라우드 서비스를 최초로 실현하였다.

맵리듀스
하둡분산파일시스템(HDFS)에 저장된 대용량의 데이터들을 대상으로 SQL을 이용하여 사용자의 질의를 실시간으로 처리하는 기술
대표적인 예 Apache Hive, Apache Tajo, Cloudera의 Impala, Facebook의 Presto, Pivotal HD의 HAWQ, Apache Drill 등이 있다.

메타데이터
데이터 표준화는 데이터표준용어설정, 명명규칙수립 ( )구축, 데이터사전구축 등의 업무로 구성
데이터 표준 용어는 표준 단어 사전, 표준 도메인사전, 표준 코드 등으로 구성되며 사전 간 상호 검증이 가능하도록 점검 프로세스를 포함해야 한다.

IoT(사물인터넷)
인터넷을 기반으로 모든사물을 연결해 사람과 사물, 사물과 사물 간의 정보를 상호 소통하는 지능형기술 및 서비스
사물에서 생성되는 Data를 활용한 분석을 통해 마케팅 등에 활용할 수 있다.
인터넷에 연결된 기기가 사람의 개입 없이 상호간에 알아서 정보를 주고 받아 처리
EX)구글의 Google Glass, 나이키의 Fuel band

블록체인
거래정보를 하나의 덩어리로 보고 이를 차례로 연결한 거래장부
기존 금융회사의 경우 중앙 집중형 서버에 거래 기록을 보관하는 반면 이것은 거래에 참여하는 모든 사용자에게 거래 내역을 보여주며 거래 때마다 이를 대조해 데이터 위조를 막는 방식을 사용한다.

BI
데이터 기반 의사결정을 지원하기 위한 리포트 중심의 도구

ERP(Enterprise Resource Planning)
기업내부 데이터베이스 중 기업 전체가 경영자원을 효과적으로 이용하기 위해 통합적으로 관리하고 경영의 효율화를 기하기 위한 수단으로 정보의 통합을 위해 기업의 모든 자원을 최적으로 관리하기 위한 기업 경영 정보시스템

KMS(지식관리시스템)
조직 내 구성원들이 축적하고 있는 노하우 등 암묵적 지식을 형식지로 표출화 될 수 있도록 지원하는 등 조직의 경쟁력 향상을 위해 지식자원을 체계화하고 원활하게 공유가 될 수 있도록 지원하는 시스템

SCM(공급망 관리)
기업이 외부 공급업체 또는 제휴업체와 통합된 정보시스템으로 연계하여 시간과 비용을 최적화시키기 위한 것으로 자재구매, 생산·재고, 유통·판매, 고객 데이터로 구성된다.

ISP(정보전략계획)
기업 및 공공기관에서의 시스템 중장기 로드맵 정의를 위해 수행, 정보기술 또는 정보시스템을 전략적으로 활용하기 위하여 조직 내·외부 환경을 분석하여 기회나 문제점을 도출하고 사용자의 요구사항을 분석하여 시스템 구축 우선순위를 결정하는 중장기 마스터 플랜을 수립하는 절차

머신러닝(기계학습)
-데이터의 패턴을 발견하고 데이터 모델의 매개 변수를 자동으로 학습한다.
-자체 알고리즘을 사용하여 시간이 경과함에 따라서 경험을 축적하면서 작업 성능이 향상된다.
-인공지능의 한 분야로 컴퓨터가 스스로 많은 데이터를 분석해서 패턴과 규칙을 찾아내고, 학습된 패턴과 규칙을 활용하여 분류나 예측을 하는 것

분석인프라
데이터 분석 도입의 수준을 파악하기 위한 분석 준비도의 6가지 구성요소 중 하나로서 운영시스템 데이터 통합, 빅데이터 분석환경, 통계분석환경 등을 진단하는 구성요소

분석 유즈 케이스
현재의 비즈니스 모델 및 유사/동종사례 탐색을 통해서 빠짐없이 도출한 분석 기회들을 구체적인 과제로 만들기 전에 ( )로 표기하는 것이 필요하다.
풀어야 할 문제에 대한 상세설명 및 해당 문제 해결했을 때 발생하는 효과를 명시함으로써 향후 데이터 분석 문제로의 전환 및 적합성 평가에 ( )를 활용

분석 과제 관리 프로세스는 크게 과제 발굴과 (1)으로 나누어진다. 조직이나 개인이 도출한 분석아이디어를 발굴하고 이를 과제화하여 분석 과제 풀(Pool)로 관리하면서 분석과제가 확정되면 (2), (3), (4)분석과제 결과 공유/개선의 분석과제 관리 프로세스를 수행하게 된다.
(1)=**과제수행**
(2)=**팀구성**
(3)=**분석과제실행**
(4)=**분석과제진행관리**

분석적 기업으로 도약을 위해서는 가장 먼저 조직의 분석도입여부 및 활용 수준에 대한 명확한 진단이 요구된다. 특히 분석수준진단방법 중 조직의 분석 및 활용을 위한 역량수준을 파악하기 위해 '도입->(1)->확산->최적화'의 분석 성숙도단계 포지셔닝을 파악한다.
(1)=**활용**

상향식 접근법
문제의 정의 자체가 어려운 경우 데이터를 기반으로 문제의 재정의 및 해결방안을 탐색하고 이를 지속적으로 개선하는 분석과제발굴방식

하향식 접근법
문제가 주어지고 이에 대한 해법을 찾기 위하여 각 과정이 체계적으로 단계화되어 수행하는 분석과제발굴방식

디자인사고(Design Thinking)
상향식 접근 방식의 발산단계와 도출된 옵션을 분석하고 검증하는 하향식 접근 방식의 수렴단계를 반복하여 과제를 발굴하는 방법

다차원척도법
-여러 대상 간의 거리가 주워져 있을 때, 대상들을 동일한 상대적 거리를 가진 실수 공간의 점들로 배치시키는 방법
-여러 대상 간의 관계에 대한 수치적자료를 이용해 유사성에 대한 측정치를 상대적 거리로 시각화

최단연결법
계층적 군집분석 방법 중 하나로 군집과 군집, 또는 데이터와의 거리계산 시, 최단거리를 계산하여 거리가 가까운 데이터, 또는 군집을 새로운 군집으로 형성하는 방법
사슬 구조의 군집이 생길 수 있다.

와드연결법 (계층적군집수행 시 두 군집간의 거리측정방법)
군집 내의 오차제곱합에 기초하여 군집을 수행

정상성
-회귀모형의 가정 중 잔차항이 정규분포를 이루어야 하는 가정을 의미하는 용어
-시계열의 수준과 분산에 체계적인 변화가 없고 엄밀하게 주기적 변동이 없다는 것으로 미래는 확률적으로 과거과 동일하다는 것을 의미

차분
시계열 분석을 위해서는 정상성을 만족해야한다. 따라서 주어진 자료가 정상성을 만족하는지 판단하는 과정이 필요하다. 자료가 추세를 보이는 경우 현 시점의 자료값에서 전 시점의 자료를 빼는 방법을 통해 비정상시계열을 정상시계열로 바꾸어 주는 방법

AR모형(자기회귀모형)
-시계열 모델 중 자기 자신의 과거 값을 사용하여 설명하는 모형
-백색잡음의 현재값과 자기 자신의 과거값의 선형가중합으로 이루어진 정상확률 모형
-모형에 사용하는 시계열 자료의 시점에 따라 1차,2차, ..., p차 등을 사용하나 정상시계열모형에서는 주로 1,2차사용

특이도
오분류표를 활용하여 모형을 평가하는 지표 중 실제값이 FALSE인 관측치 중 예측치가 적중한 정도

향상도
연관규칙의 측정 지표 중 도출된 규칙의 우수성을 평가하는 기준으로 두 품목의 상관관계를 기준으로 도출된 규칙의 예측력을 평가하는 지표

유전자 알고리즘(유전알고리즘)
-생명의 진화를 모방하여 최적화를 구하는 알고리즘으로 존 홀랜드가 1975에 개발
-'최대의 시청률을 얻으려면 어떤 프로그램을 어떤 시간대에 방송해야 하는가?' 와 같은 문제를 해결할 때 사용
-어떤 미지의 함수 Y=f(x)를 최적화하는 해 x를 찾기 위해, 진화를 모방한 탐색알고리즘

앙상블
다수 모델의 예측을 관리하고 조합하는 기술을 메타 학습이라 한다. 여러 분류기들의 예측을 조합함으로써 분류 정확성을 향상시키는 기법

배깅
원 데이터 집합으로부터 크기가 같은 표본을 여러 단순임의복원추출하여 각 표본에 대해 분류기를 생성한 후 그 결과를 앙상블하는 기법

부스팅
앙상블 기법 중 붓스트랩 표본을 구성하는 재표본 과정에서 분류가 잘못된 데이터에 더 큰 가중치를 주어 표본을 추출하는 기법

랜덤포레스트
배깅에 랜덤과정을 추가한 방법
원 자료로부터 붓스트랩 샘플을 추출하고,
각 붓스트랩 샘플에 대해 트리를 형성해 나가는 과정은 배깅과 유사하나,
각 노드마다 모두 예측변수 안에서 최적의 분할을 선택하는 방법 대신 예측변수를 임의로 추출하고 추출된 변수 내에서 최적의 분할을 만들어 나가는 방법을 사용

홀드아웃
모형평가방법 중 주어진 원천 데이터를 랜덤하게 두 분류로 분리하여 교차검정을 실시하는 방법으로 하나는 모형의 학습 및 구축을 위한 훈련용자료로 다른하나는 성과 평가를 위한 검증용자료로 사용

역전파알고리즘
인공신경망에서 동일 입력층에 대해 원하는 값이 출력되도록 개개의 가중치를 조정하는 방법

기울기소실문제(Gradient Vanishing)
-신경망의 모형에는 Visible Hidden Layer로 구성되어 Layer가 많은 다층 퍼셉트론에서 Hidden Layer를 많이 거칠수록 전달되는 오차가 크게 줄어들어 학습이 되지 않는 현상
-신경망 모형의 학습을 위한 역전파 과정에서 오차를 더 줄일 수 있는 가중치가 존재함에도 기울기가 0이 되어버려 더 이상 학습이 진행되지 않는 문제

과대적합(현상자체를 의미)
-의사결정나무 모형에서 가지를 끝까지 모두 사용해 순도 100%상태로 만들면 실제 데이터에 적용할 수 없게 되는 문제점 발생, 분기가 너무 많아 발생

계절요인
시계열 분석에서 일정한 주기를 가지고 반복적으로 같은 패턴을 보이는 변화를 나타내는 요인

과적합(이러한 상태를 의미)
-의사결정나무에서 끝마디가 너무 많으면 모형에 ( )인 상태로 현실문제에 적용될 수 있는 적절한 규칙이 나오지 않게 된다. 따라서 분류된 관측치의 비율 또는 MES(Mean Square Error) 등을 고려하여 적절한 수준의 가지치기 규칙을 제공해야 한다.

정지규칙
의사결정 나무에서 더 이상 분기가 되지 않고 현재의 마디가 끝마디가 되도록 하는 규칙

ESD
이상값 탐색기법 중 하나로 평균으로부터 k*표준편차만큼 떨어져 있는 값들을 이상값으로 판단하는 방법

포아송분포
이산형 확률분포 중 주어진 시간 또는 영역에서 어떤 사건의 발생 횟수를 나타내는 확률분포

분류분석
은행에서 대출 심사를 할 때 소득, 카드 사용액, 나이 등 해당 고객의 개인정보 정보를 바탕으로 그 고객이 대출 상환을 잘 하는 집단에 속할지 그렇지 않은 집단에 속할지를 예측할 수 있다.

실루엣계수
군집분석의 품질을 정량적으로 평가하는 대표적인 지표
군집 내의 데이터 응집도와 군집 간 분리도를 계산하여
군집 내의 데이터의 거리가 짧을수록,
군집 간 거리가 멀수록 데이터의 값이 커지며
완벽한 분리일 경우 1의 값을 가짐

프레이밍 효과
합리적 의사결정을 방해하는 요소로써
표현 방식 및 발표자에 따라
동일한 사실에도 판단을 달리하는 현상

오즈
-로지스틱 회귀모형에서 한 단위 증가할 때마다 성공(y=1)의 ( )이 증가하는지 나타낸다
-로지스틱 회귀분석에서는 이산형 종속변수가 1일 확률을 모형화한다.
설명변수가 한 단위 증가할 때 종속변수가 1인 확률과 0인 확률 비의 증가율을 나타내는 것

SOM(자기조직화지도)
코호넨에 의해 제시되었으며, 비지도 신경망으로 고차원의 데이터를 저차원의 뉴런으로 정렬하여 지도의 형태로 형상화라는 클러스터링방법

인과관계
어떤 현상에 대하여 현상을 발생시킨 원인과 그 결과 사이의 관계

상관관계
어떤 두 현상이 관계가 있음을 말하지만 어느 쪽이 원인인지 알 수 없는 관계

문제정의
-문제 탐색을 통해서 식별된 비즈니스 문제를 변환하는 단계로써, 문제 탐색 단계가 무엇을 어떤 목적으로 수행해야 하는가에 대한 관점이었다면, ( )단계는 이를 달성하기위해서 필요한 데이터 및 기법(HOW)을 도출하기 위한 데이터 분석의 문제로의 변화을 수행하게 된다.
-식별된 비즈니스 문제를 데이터의 문제로 변환하여 정의하는 단계

모델링
분석용 데이터를 이용한 가설 설정을 통하여 통계모델을 만들거나 기계학습을 이용한 데이터의 분류,예측,군집 등의 기능을 수행하는 모델을 만드는 과정

소프트맥스함수(softmax함수)
신경망 모형에서 표준화지수함수로 불리며, 출력값 z가 여러 개로 주어지고, 목표치가 다범주인 경우 각 범주에 속할 사후확률을 제공하여 출력노드에 주로 사용되는 함수

최소제곱법
회귀모형의 계수를 추정하는 방법
잔차제곱합을 최소화하는 계수를 찾는 방법

후진제거법
최적방정식을 선택하기 위한 방법 중 모든 독립변수 후보를 포함한 모형에서 시작하여 가장 적은 영향을 주는 변수를 하나씩 제거하면서 더 이상 유의하지 않은 변수가 없을 때까지 설명변수를 제거하는 방법

층화추출법
상당히 이질적인 원소들로 구성된 모집단에서 각 계층을 고루 대표할 수 있도록 표본을 추출하는 방법이다. 이질적인 모집단의 원소들로 서로 유사한 것끼리 몇 개의 층을 나눈 후, 각 계층에서 표본을 랜덤하게 추출한다.

내용 기반 필터링(Content-based filtering)
아이템에 대한 설명과 사용자 선호를 기반으로 하여 과거에 사용자가 좋아했던 것과 비슷한 아이템을 추천하는 알고리즘

로짓변환
로지스틱 회귀분석에서 어떠한 일이 일어날 확률을 일어나지 않을 확률을 나누어 log를 취하고 이를 0~1의 값이 아닌 (-∞, +∞) 범위에서 선형함수를 시그모이드 함수로 변환하는 방법

시급성
전략적 중요도가 핵심이며, 이는 현재의 관점에서 전략적 가치를 둘 것인지, 미래의 중장기적 관점에 전략적인 가치를 둘 것인지를 고려하고, 분석 과제의 목표가치(KPI)를 함께 고려하여 ( )의 여부를 판단할 수 있다.

나이브베이즈분류
베이즈 정리와 특징에 대한 조건부 독립을 가설로 하는 알고리즘으로 클래스에 대한 사전 정보와 데이터로부터 추출된 정보를 결합하고 베이즈 정리를 이용하여 특정 데이터가 특정 클래스에 속하는지를 분류하는 알고리즘

스테밍(어간추출)
텍스트 마이닝에서 어근에 차이가 있더라도 관련이 있는 단어들을 동일한 어간으로 매핑이 될 수 있도록 정해진 규칙에 따라 단어에서 어간을 분리하여 공통 어간을 가지는 단어를 묶는 작업

분해시계열
시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법

인공신경망모형
데이터 마이닝 기법 중 동물의 뇌신경계를 모방하여 분류(또는 예측)을 위해 만들어진 모형

ROC Curve
레이더 이미지 분석의 성과를 측정하기 위해 개발된
이 그래프는 두 분류분석모형을 비교 분석 결과를 가시화 할 수 있다는 점에서 유용한 평가도구
X축에는 FP Ratio(1-특이도)를 나타내며
Y축에는 민감도를 나타내 두 평가 값의 관계로 모형을 평가 모형의 성과를 평가하는 기준은 그래프의 밑 부분 면적이 넓을수록 좋은 모형으로 평가

향상도곡선(lift curve)
분류분석의 모형을 평가하는 방법으로 랜덤모델과 비교하여 해당 모델의 성과가 얼마나 향상되었는지를 각 등급별로 파악하는 그래프

이익도표
분류분석모형을 사용하여 분류된 관측치가
각 등급별로 얼마나 포함되는지를 나타내는 도표

나선형모델
반복을 통하여 점진적으로 개발하는 방법
처음 시도하는 프로젝트에 적용이 용이하지만
관리체계를 효과적으로 갖추지 못한 경우 복잡도가 상승하여 프로젝트 진행이 어려울 수 있다.

제1종 오류
가설검정 결과에서
귀무가설이 옳은데도 귀무가설을 기각하게 되는 오류

점추정
통계분석 개념 중 모집단의 특성을 단일한 값으로 추정하는 방법

지니지수
불순도를 측정하는 지표로 노드의 불순도를 의미
클수록 이질적이며 순수도가 낮다
CART에서 목적변수가 범주형일 경우 사용

집중구조
-전사 분석업무를 별도의 분석전담조직에서 담당
-전략적 중요도에 따라 분석조직이 우선순위를 정해서 진행가능
-현업 업무부서의 분석업무와 이중화/이원화 가능성 높음

비모수모형
의사결정나무와 같이 선형성, 정규성, 등분산성 등의 가정을 필요로 하지 않는 모형

검증용데이터
데이터 마이닝 적용한 후 그 결과의 신빙성을 검증하기 위해 데이터를 분할하는데 구축된 모델의 과잉 또는 과소 맞춤 등에 대한 미세조정 절차를 위해 사용되는 데이터