1. Wprowadzenie do kursu. Business Analytics (BI), Data Analytics i Data Science. Proces odkrywania wiedzy. Etapy: definicja problemu, pozyskiwanie danych, czyszczenie danych, modelowanie danych, ewaluacja, komunikowanie wyników. Studium przypadku demonstrujące kolejne etapy procesu.

Przetwarzanie danych i odkrywanie wiedzy

Tomasz Kajdanowicz, Kamil Tagowski

Plan wykładu

1. Analityka biznesowa
2. Analityka danych
3. Data Science - nauka o danych
4. Proces odkrywania wiedzy
    a. Crisp DM-2
    b. Team Data Science Process
    c. Semma
5. Studium przypadku demonstrujące kolejne etapy procesu odkrywania wiedzy

Literatura

LITERATURA PODSTAWOWA:

1. Provost F., Fawcett T., “Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking”, O'Reilly Media, 2013.
2. VanderPlas J., “Python Data Science Handbook: Essential Tools for Working with Data”, O'Reilly Media, 2016.
3. William McKinney, “Python for Data Analysis”, O’Reilly Media, 2012.
4. Emmanuel Ameisen, “Building Machine Learning Powered Applications - Going from Idea to Product”, O’Reilly Media, 2020.

Analityka Biznesowa

- Analityka Biznesowa termin z połowy lat 90 XX w. - Gartner Group
- korzenie już w latach 70
- koncepcja rozszerzająca wsparcie managerów przez zastosowanie wielowymiarowego raportowania na żądanie
- obecnie systemy BI integrują narzędzia i metody sztucznej inteligencji

Analityka Biznesowa

Analityka Biznesowa to używanie:

- danych
- technik informacyjnych
- analizy statystycznej
- metod ilościowych
- modeli matematycznych i sztucznej inteligencji

aby pomóc menedżerom uzyskać lepszy wgląd w działania biznesowe i podejmować lepsze, oparte na faktach decyzje

Przykładowe zastosowania

Cennik: ustalanie cen towarów konsumpcyjnych i przemysłowych

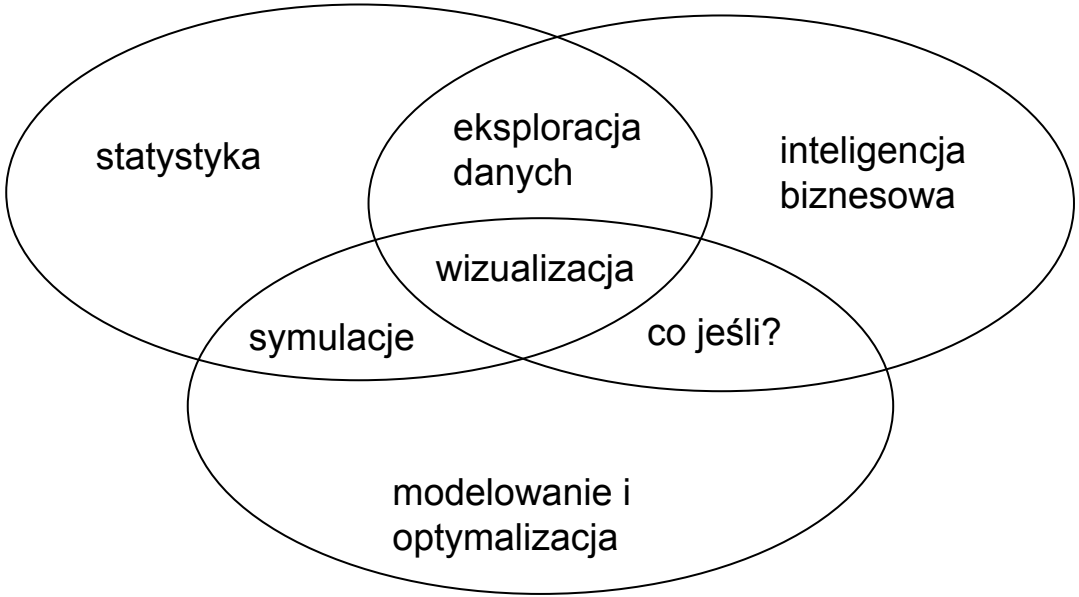
Segmentacja klientów: identyfikacja i dotarcie do kluczowych grup klientów w handlu detalicznym, ubezpieczeniach i branży kart kredytowych

Sprzedaż: określanie marek do zakupu, ilości i przydziałów

Lokalizacja: znalezienie najlepszej lokalizacji dla oddziałów banków i bankomatów lub dokąd się udać po serwis urządzeń przemysłowych

Media społecznościowe: zrozumienie trendów i postrzeganie klientów; wspomaganie marketingu

Narzędzia Analityki Biznesowej



Benefity i wyzwania

Benefity

- niższe koszty
- lepsze zarządzanie ryzykiem
- szybsze decyzje
- lepsza produktywność
- lepsze wyniki finansowe

Wyzwania

- brak zrozumienia, jak korzystać z analityki,
- konkurencyjne priorytety biznesowe, niewystarczające umiejętności analityczne,
- trudności w uzyskaniu dobrych danych i udostępnianiu informacji,
- brak zrozumienia dla korzyści
- koszty badań analitycznych

Zakres Analityki Biznesowej

Analiza opisowa

wykorzystanie danych do zrozumienia dotychczasowych i obecnych wyników biznesowych do podejmowania świadomych decyzji

Analiza predykcyjna

przewidywanie przyszłości poprzez badanie danych historycznych, wykrywanie wzorców lub relacji w danych, a następnie ekstrapolacja i predykcja w przyszłość

Analiza zaleceniowa

zidentyfikuj najlepsze alternatywy aby zminimalizować lub zmaksymalizować jakiś cel

Przykład

- U większości sprzedawców funkcjonuje mechanizm obniżek cen w zależności od pór roku, zapasów, wydarzeń.
- Kluczowe pytania:
  - od kiedy i do kiedy obniżyć cenę?
  - o ile aby zmaksymalizować przychody?

Analiza opisowa	Analiza predykcyjna	Analiza zaleceniowa
sprawdź dane historyczne pod kątem podobnych produktów (ceny, sprzedane jednostki, reklama,…)	przewiduj sprzedaż na podstawie historii sprzedaży	znajdź najlepsze zestawy cen i reklamy w celu maksymalizacji przychodów ze sprzedaży

Narzędzia

dostęp do składowanych danych

arkusze kalkulacyjne

wizualizacja danych

pulpity do raportowania kluczowych mierników wydajności

metody statystyczne

symulacje

analizy scenariuszy i „co by było, gdyby”

klasyfikacja i Predykcja

grupowanie

uczenie reprezentacji

optymalizacja

analiza danych tekstowych, audio, obrazów, wideo

analityka sieci i mediów społecznościowych

Narzędzia software’owe

- SQL, NoSQL - dostęp do różnych baz danych (relacyjne, dokumentowe, grafowe, strumieniowe, ...)
- Excel, Tableau (proste narzędzia typu „przeciągnij i upuść” do wizualizacji danych)
- IBM Cognos (zintegrowana inteligencja biznesowa i rozwiązanie planistyczne zaprojektowane z myślą o potrzebach firm średniej wielkości)
- SAS / SPSS / Rapid Miner (modelowanie predykcyjne, wizualizacja, prognozowanie, optymalizacja i zarządzanie modelami, statystyka)
- R / Python (rozwiązania otwarte oparte na dostępnym oprogramowaniu)

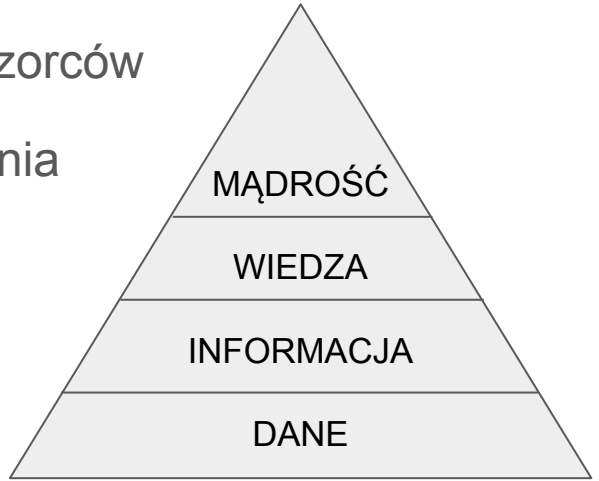
Analityka Danych

Dane - liczbowe lub tekstowe fakty i liczby, które są zbierane za pomocą pewnego rodzaju pomiarów procesu

Informacje - podsumowanie danych, kombinacje i przecięcia danych, relacje

Wiedza - stwierdzenia, agregacje, reguły, zrozumienie wzorców

Mądrość - zrozumienie zjawiska, możliwość abstrahowania



Źródła danych - przykłady

- Dane transakcji sprzedażowych
- Raporty roczne
- Trendy gospodarcze
- Badania marketingowe
- Zachowanie użytkowników w sieci, w sklepach internetowych
- Media społecznościowe
- Telefon komórkowy
- Urządzenia IOT

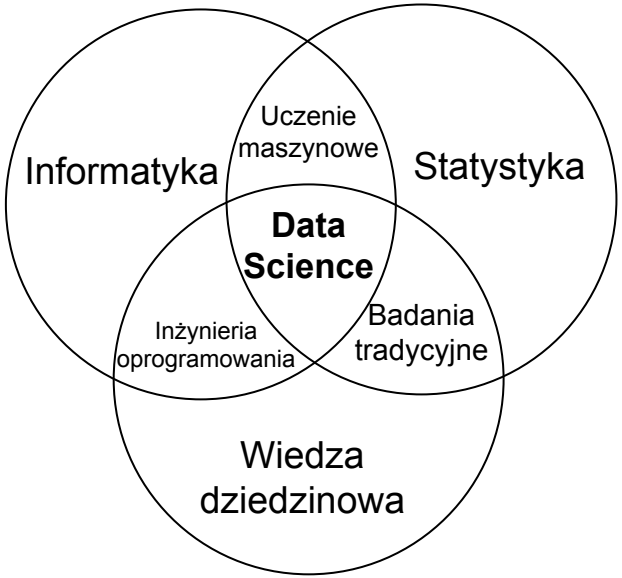
Typy danych

- Dane odzwierciedlają konkretne obiekty
- Obiekty mogą być najróżniejsze: dane tabelaryczne, obrazy, muzyka, video, grafy, strumienie danych, serie czasowe, ...
- Dane mogą być ustrukturyzowane i nieustrukturyzowane
- Dane co do zasady są wielowymiarowe
- Typy danych
    - Dyskretne - wywodzący się z liczenia czegoś, np. dostawa odbywa się na czas lub nie; zamówienie jest kompletne lub niekompletne; faktura ma jeden, dwa, trzy lub dowolną liczbę błędów
    - Ciągłe oparte na ciągłej skali pomiarów, np. wszelkie wskaźniki obejmujące dolary, długość, czas, objętość, wagę, ...

Typy pomiarów

- kategoryczne (nominalne) - według kategorii lub określonych cech
- porządkowe - można je uporządkować lub uszeregować według pewnych relacji między sobą
- przedziałowe - jak porządkowe, ale mają stałą różnice między obserwacjami
- dane typu współczynnik - ciągłe i mają naturalne zero

Nauka o danych - Data Science



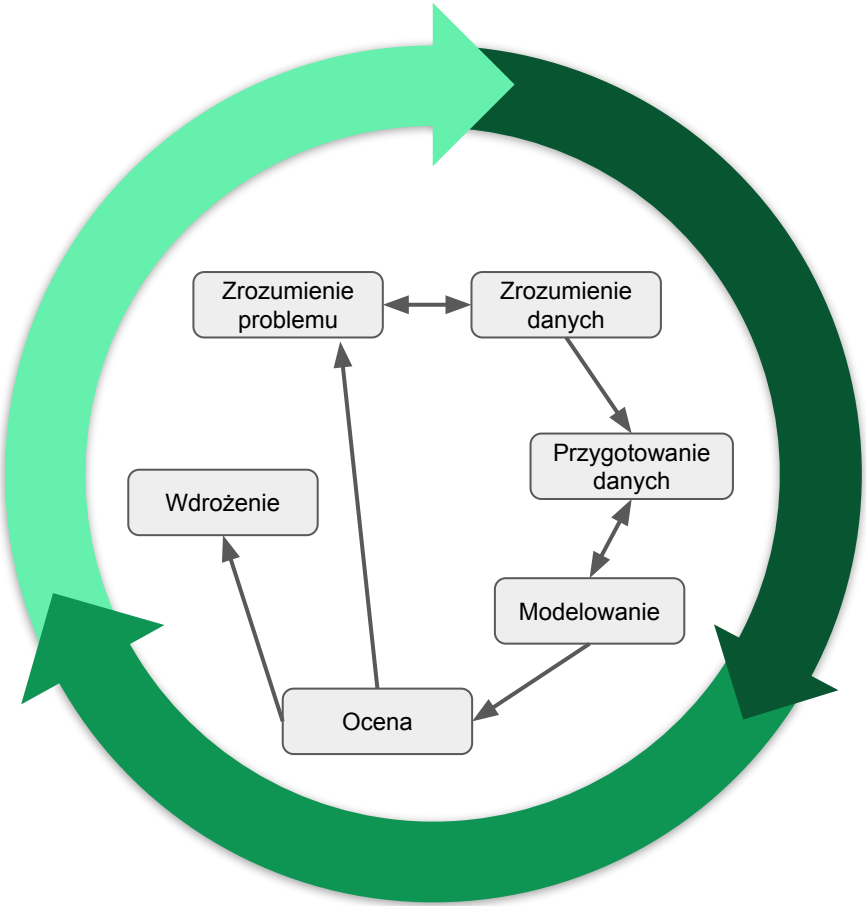
- Rozwój usług i produktów w oparciu o dane
  - wykorzystanie danych jako wejście
  - przetwarzanie danych w celu zwrócenia wygenerowanych algorytmicznie wyników

Data Science to przede wszystkim dociekliwość, zadawanie nowych pytań, dokonywanie nowych odkryć i uczenie się nowych rzeczy.

Źródło:
https://www.youtube.com/watch?v=4g-a5cs7tQ0

Crisp DM-2

- CRoss Industry Standard Process for Data Mining (CRISP-DM)
- model procesu przetwarzania danych i odkrywania wiedzy
- sześć faz, które opisują cykl życia projektu
  - Zrozumienie biznesu/problemu - czego potrzebuje firma? jaki problem właściwie chcemy rozwiązać?
  - Zrozumienie danych - jakie dane mamy / potrzebujemy? Czy jakość danych jest odpowiednia?
  - Przygotowanie danych - Jak organizujemy dane do modelowania?
  - Modelowanie - jakie techniki modelowania powinniśmy zastosować?
  - Ocena - który model najlepiej spełnia cele biznesowe?
  - Wdrożenie - w jaki sposób interesariusze uzyskują dostęp do wyników?


Zrozumienie problemu
Zrozumienie danych
Przygotowanie danych
Modelowanie
Ocena
Wdrożenie

Zrozumienie biznesu/problemu

- Określenie celów biznesowych/problemów
  - „dokładnie zrozumieć, co naprawdę chce osiągnąć”
  - jakie są kryteria sukcesu ?
- Ocena sytuacji
  - dostępne zasoby, wymagania, ryzyko, analiza kosztów i korzyści
- Cele analizowania danych
- Plan projektu
  - wstępny wybór technologii i narzędzi
  - WBS, backlog, specyfikacja produktu/usługi, ...

Zrozumienie danych

- identyfikacja posiadanych zasobów
- gromadzenie brakujących danych
- analizowanie danych pod kątem osiągnięcia celów projektu

Zagadnienia:

- zbieranie danych
- opisanie danych: właściwości, formaty, kodowania, liczby rekordów, rozkłady, znaczenie
- eksploracyjna analiza danych: wizualizuj i identyfikuj relacje między danymi
- jakość danych

Przygotowanie danych

- przygotowanie ostatecznych zbiorów danych do modelowania

Zagadnienia

- zwężenie i wybranie docelowych danych (odpowiedź dlaczego akurat tak)
- czyszczenie danych: poprawianie, przypisywanie lub usuwanie błędnych wartości
- inżynieria cech (jeśli potrzebna), uczenie reprezentacji
- integracja danych: łączenie danych z wielu źródeł
- formatowanie danych: odpowiednie typy dla dat, liczb, danych złożonych

Modelowanie

- tworzenie i ocena modeli w oparciu o wiele różnych technik modelowania

Zagadnienia

- wybór techniki modelowania: które algorytmy wypróbować?
- zaplanowanie testowania: zbiory uczące, testowe i walidacyjne, walidacje krzyżowe, boostrapy, ...
- uczenie modelu: UWAGA! duże modele długo się mogą liczyć
- ewaluacja modelu: interpretacja wyniku, sprawdzenie jak się ma do kryteriów sukcesu

krok wielokrotnie iterowany

Ocena

- ocena modelu ≠ ocena biznesowa

Zagadnienia

- czy modele spełniają kryteria sukcesu biznesowego?
- przeglądu prac: Czy aby wszystkie kroki zostały wykonane prawidłowo?
- kolejne kroki: czy przystąpić do wdrażania?

Wdrożenie

- model jest nieprzydatny, póki nie dostarcza wartości dla odbiorcy

Zagadnienia

- opracowanie i udokumentowanie planu wdrożenia modelu
- plan monitorowania i konserwacji, aby uniknąć problemów w fazie operacyjnej (lub fazie poprojektowej) modelu.
- raport końcowy, przegląd projektu (retrospekcja)
- utrzymanie w ‘produkcji’

CRISP-DM - podsumowanie

- **KLUCZOWE**: Zrozumienie biznesu/problemu
- **TO TEŻ KLUCZOWE**: Zrozumienie danych
- **ZAJMUJE 80% CZASU**: Przygotowanie danych
- **NAJBARDZIEJ EKSCYTUJĄCE**: Modelowanie
- **MOMENT PRAWDY**: Ocena
- **PROJEKT SAM W SOBIE**: Wdrożenie

Team Data Science Process

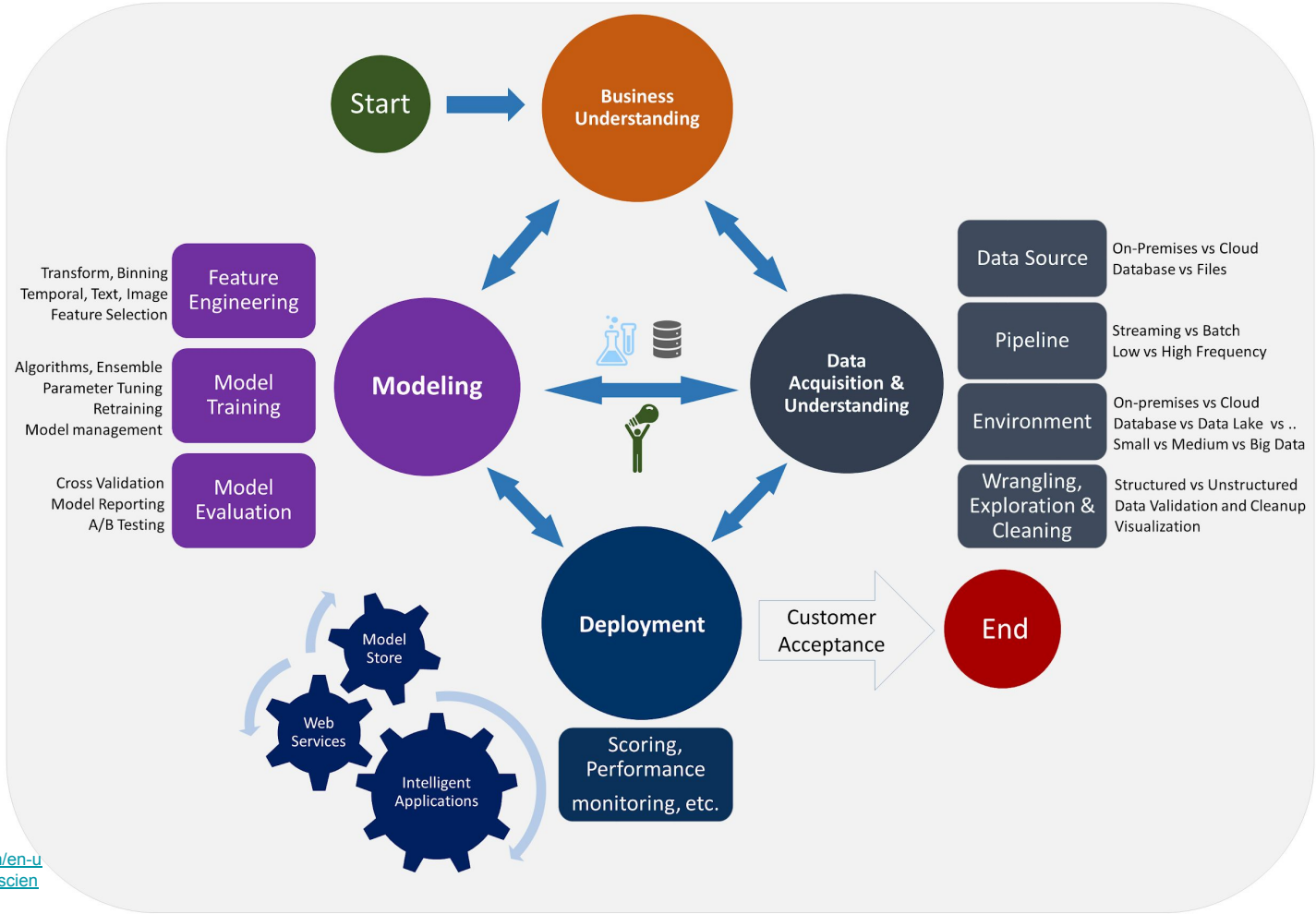
- zwinna, iteracyjna metodologia nauki o danych
- wydajne dostarczanie rozwiązań do analizy predykcyjnej i inteligentnych aplikacji
- pomaga usprawnić współpracę zespołową i uczenie się, sugerując, jak najlepiej współpracują ze sobą role w zespole
- zawiera najlepsze praktyki i struktury firmy Microsoft i innych liderów w branży

Komponenty TDSP

- Definicja cyklu życia
- Ujednolicona struktura projektu
- Infrastruktura i zasoby zalecane dla projektów data science
- Narzędzia i programy narzędziowe zalecane do realizacji projektów

Cykl życia projektu:

- Zrozumienie biznesu
- Gromadzenie i zrozumienie danych
- Modelowanie
- Wdrażanie



Źródło:
https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview

Role w projekcie

Architekt rozwiązań

Menadżer projektu (project lead)

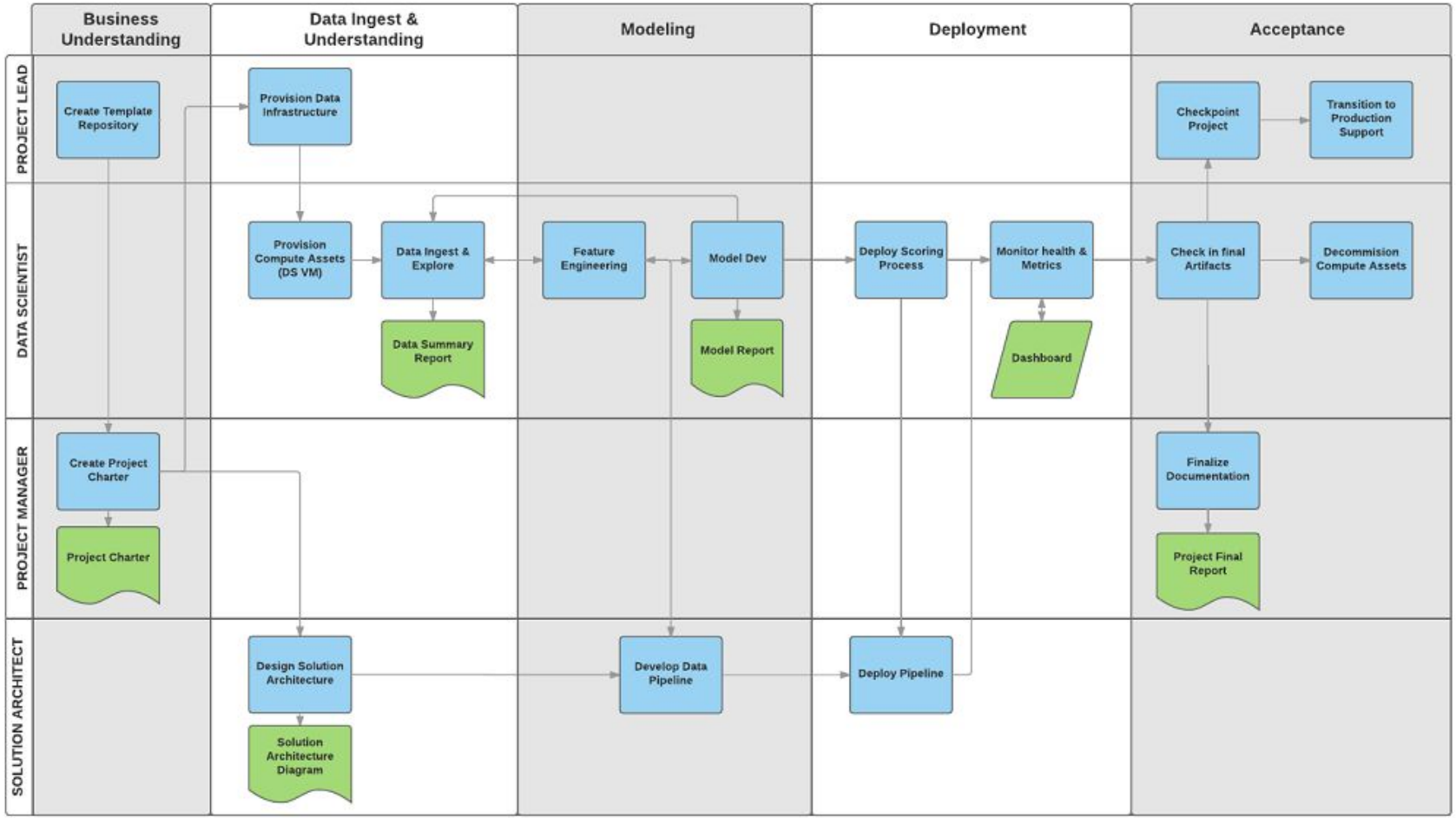
Inżynier danych

Danolog

Twórca aplikacji (programista)

ML/DS DevOps

Kierownik projektu



Źródło:
https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview
Licencja otwarta.

Semma

- Sample, Explore, Modify, Model, and Assess
- SAS

Zagadnienia

- próbkowanie danych: jedno lub więcej źródeł
- eksploracja danych: relacje, trendy, anomalie, ...
- modyfikacja danych: tworzenie, wybieranie, przekształcanie zmiennych
- modelowanie
- ocena: użyteczność i wiarygodność ustaleń z procesu

Podsumowanie

Analityka biznesowa

Analityka danych

Data Science - nauka o danych

Proces odkrywania wiedzy

Metodyki

- Crisp DM-2
- Team Data Science Process
- Semma