Eksploracja Danych Geoprzestrzenych i Uczenie maszynowe

Geo-Data Scicence

Jarosław Jasiewicz

Eksploracja danych i Uczenie maszynowe

Geoinformacja program magisterski
Specjalność Geoinformatyka

Data Science

- Nowa dyscyplina
- Termin zaproponowany w 2012 roku, polskie określenia to **Inżynieria danych**, blisko pokrewny z terminem analiza danych.
- Zasadnicze różnice:
  - Rozumie **kontekst** swoich analiz – na przykład w zakresie **nauk o Ziemi**
  - Umiejętności komunikacyjne w zakresie dziedziny w której pracuje – na przykład **nauk o Ziemi**
  - Wszechstronna wiedza, umiejętność poruszania się w **różnych, nieuporządkowanych** zbiorach danych
  -
- Analiza danych == statystyka

Plan zajęć

1) Wstęp i podstawowe pojęcia, obszary zastosowań,
2) Struktury danych, porządkowanie danych, uzupełnianie braków, typy danych i atrybutów, analiza danych geoprzestrzennych
3) Wizualizacja i eksploracyjna analiza danych
4) Regresja, wykrywanie zależności między atrybutami, różne typy regresji, analiza wielowymiarowa
5) Klasyfikacje bez znanej zmiennej zależnej (nienadzorowane,grupowanie), ocena jakości klasyfikacji
6) Analizy asocjacyjne, wykrywanie wzorców
7) Klasyfikacje ze znaną zmienną zależną (nadzorowane, uczenie z danych), Ocena jakości klasyfikacji

Podstawowa literatura

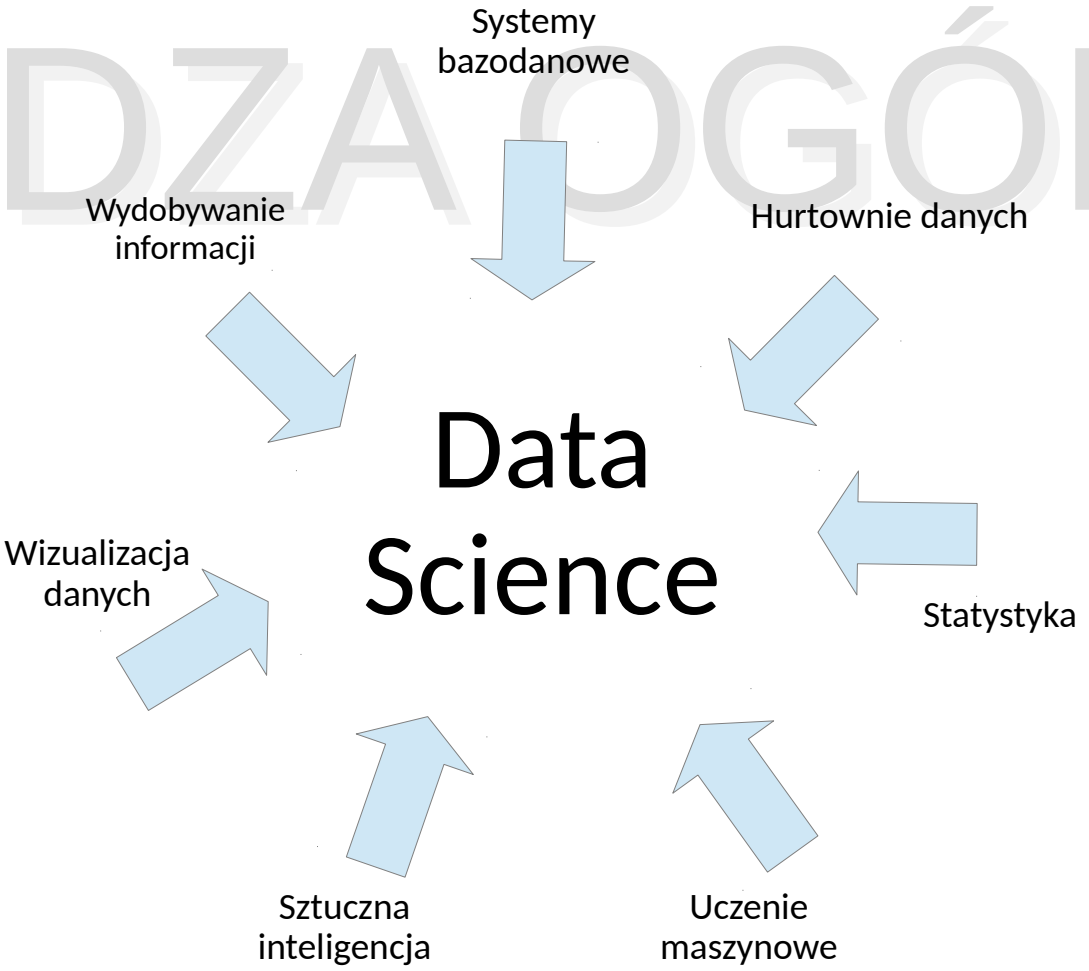
- Larose D., Odkrywanie wiedzy z danych. Wprowadzanie do eskploracji danych, PWN, 2006.
- Larose D., Metody i modele eksploracji danych, PWN 2008
- P. Biecek – Na przełaj przez Data Mining (internet)
- Krawiec K, Stefanowski J., Uczenie maszynowe i sieci neuronowe, Wyd. PP, 2003.
- Podręczniki do data mining w środowisku R...

Dobry podręcznik do Geo-Data Science nie istnieje...

Dlaczego Data Science

- Współczesne techniki zbierania danych (GPS, wysoko--rozdzielcze obrazy satelitarne, narzędzia lokalizacji w czasie rzeczywistym, wolontariacki GIS) dostarczają ilości danych niemożliwych do przetwarzania **metodami obserwacyjnymi**.
- Dostępność ogromnej ilości danych geoprzestrzennych i czasoprzestrzennych umożliwia zdobywanie **nowej wiedzy** oraz lepszego zrozumienia procesów geograficznych (np. interakcje człowiek-środowisko, procesy społeczno-ekonomiczne, globalne zmiany klimatu, itp.)
- Dotychczas poznane metody statystyczne posiadają wiele ograniczeń i wymagają spełnienia wielu założeń:
  - Ciągłość danych
  - Liniowość relacji pomiędzy zmiennymi
  - Wymóg kompletności danych
  - Odrębne metody dla danych ilościowych i jakościowych

Interdyscyplinarność



Ewolucja zagadnienia

- Lata 60: tworzenie baz danych, bazy sieciowe;
- Lata 70: model relacyjny;
- Lata 80 zaawansowane modele RDBMS, modele obiektowe i inne, SQL;
- Lata 90 hurtownie danych, web, multimedia;
- Lata 2000 uczenie maszynowe, big Data;
- Lata 2010+, Data Science, Data-driven AI...

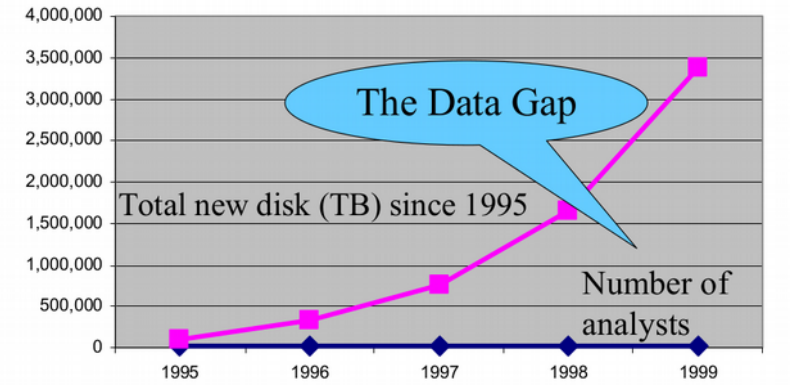
Definicje

- **Eksploracja danych** (ang. data mining) odkrywanie niejawnych, wcześniej nieznanych i potencjalnie użytecznych informacji, zależności i związków w zbiorze danych.
- Wykorzystuje zaawansowane metody i algorytmy,pozwala na zbadanie charakteru zgromadzonych danych oraz pozwala na wyciągnięcie z nich konkretnych informacji i wiedzy.
- Bardzo często pojęcia eksploracji danych i odkrywania wiedzy a dokładniej odkrywania wiedzy w bazach danych (ang. k**nowledge discovery in databases**) przeplatają się ze sobą a często używane są wymiennie w zależności od podejścia.
- **Odkrywanie wiedzy** odnosi się całościowo do procesu odkrywania przydatnych i pożytecznych informacji i wiedzy poprzez eksplorowanie **baz danych**, podczas gdy eksploracja danych ma węższe znaczenie, gdyż dotyczy samego wyboru i wykorzystania algorytmów oraz aplikacji służących do wydobycia z baz reguł, zależności, schematów.
- Narzędzia eksploracji danych są wykorzystywane do automatyzacji procesu poszukiwania związków, zależności, relacji czy schematów i generują rezultaty, które mogą zostać użyte zarówno bezpośrednio w procesie podejmowania decyzji przez określone osoby jak również zaawansowane systemy wspomagania decyzji.

Inne definicje: knowledge extraction,data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Wiedza a dane

- „Topimy się w morzu danych szukając wiedzy”
- Nie wszystkie dane są użyteczne – większość nie jest (szum)
- Problemem jest wydobycie użytecznej wiedzy
- Dane przyrastają szybciej niż możliwości ich przetwarzania

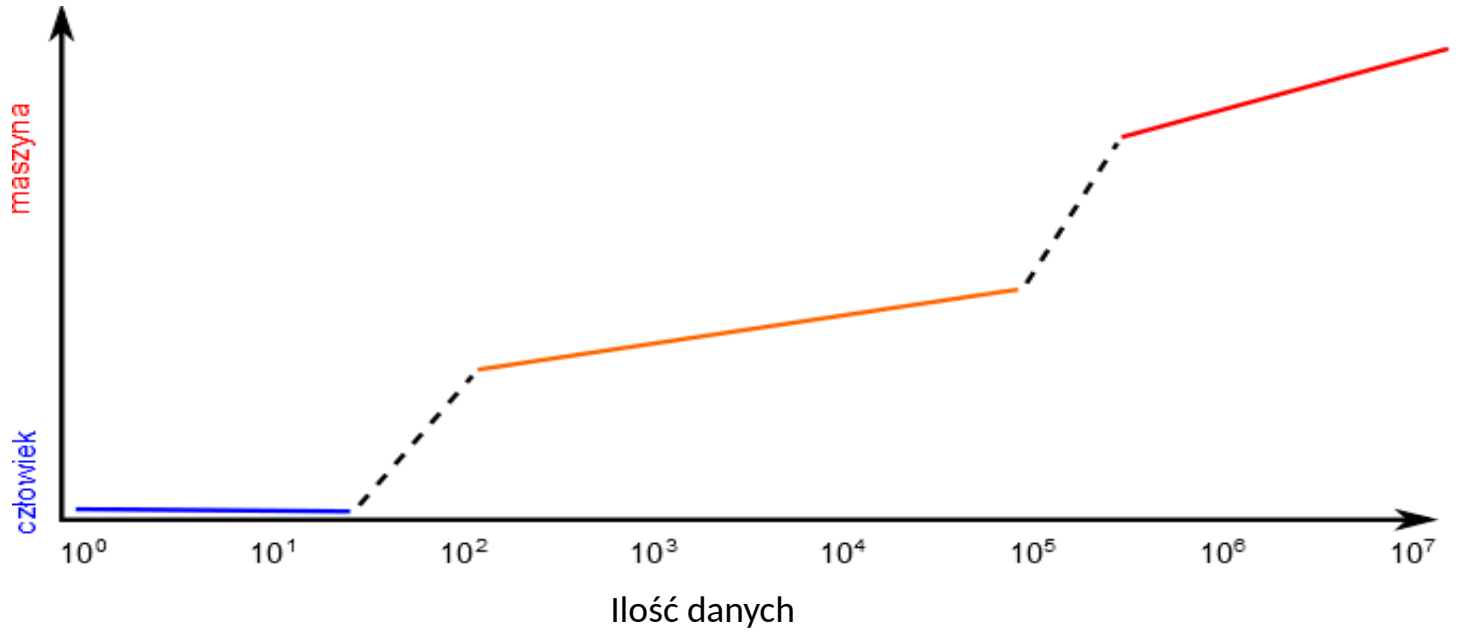


Przykład: mamy TB zdjęć satelitarnych **dziennie**, na nich dane na temat pożarów lasów. Ale nie wiemy gdzie są te pożary? -

odpowiedź: katalogowanie obrazów na podstawie zawartości

Wydobywanie wiedzy

- Odkrywanie ukrytych, wcześniej nieznanych a potencjalnie przydatnych informacji z dużych i złożonych zasobów danych
- proces analityczny (oparty o wnioskowanie z wykorzystaniem określonych algorytmów), przeznaczony do **badania dużych zasobów danych** w poszukiwaniu regularnych wzorców oraz systematycznych współzależności pomiędzy zmiennymi, a następnie do oceny wyników poprzez zastosowanie wykrytych wzorców do nowych podzbiorów danych. Końcowym celem data mining jest najczęściej WIEDZA.



Analiza danych a Data Science

	Analiza danych	Data Science
Typ danych	• Uporządkowanie strukturalizowane	• Uporządkowane niestrukturalizowane • Nieuporządkowane
Ilość obiektów	• Mała/średnia	• Duża (dowolna)
Ilość cech	• Mała	• Duża
Cel	• Estymacja • Weryfikacja hipotez • Badanie rozkładu	• Poszukiwanie prawidłowości, wzorców, związków i anomalii
Wynik	• Interpretacja danych na podstawie zadanych kryteriów	• Odkrywanie relacji między cechami/obiektami nie zawsze oczywistych

Data science to nie jest...

- Rozszerzona „zaawansowana” statystyka
- Tylko uczenie maszynowe
- Zapytania do baz danych (SQL)
- Systemy ekspertowe

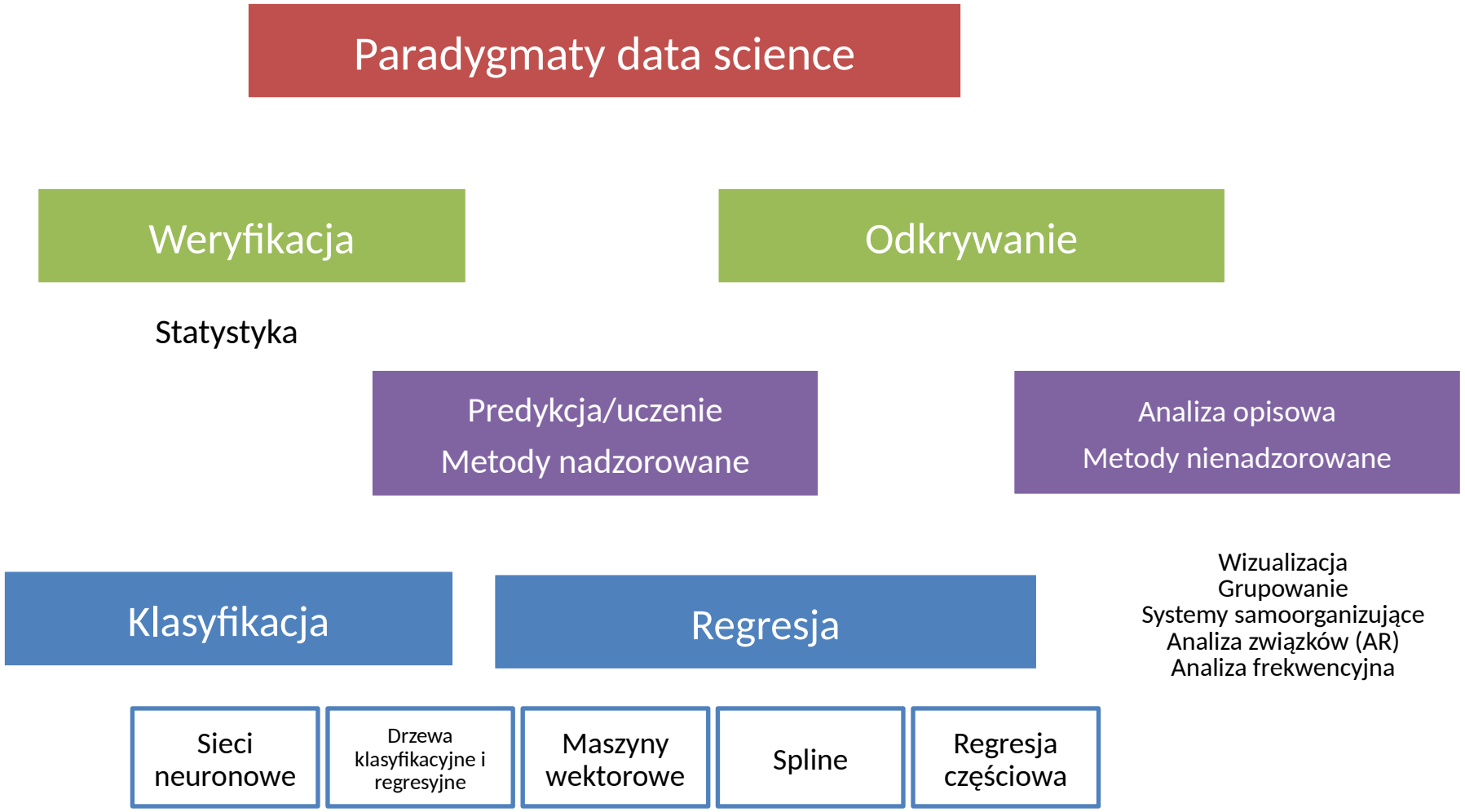
Jakie typy danych

- Tabele atrybutowe
- Bazy relacyjne (wymóg I postaci normalnej)
- Sekwencyjne bazy danych
- Bazy geoprzestrzenne
- Serie czasowe danych
- Bazy tekstów
- Zasoby WWW
- Grafy
- Dane nieuporządkowane

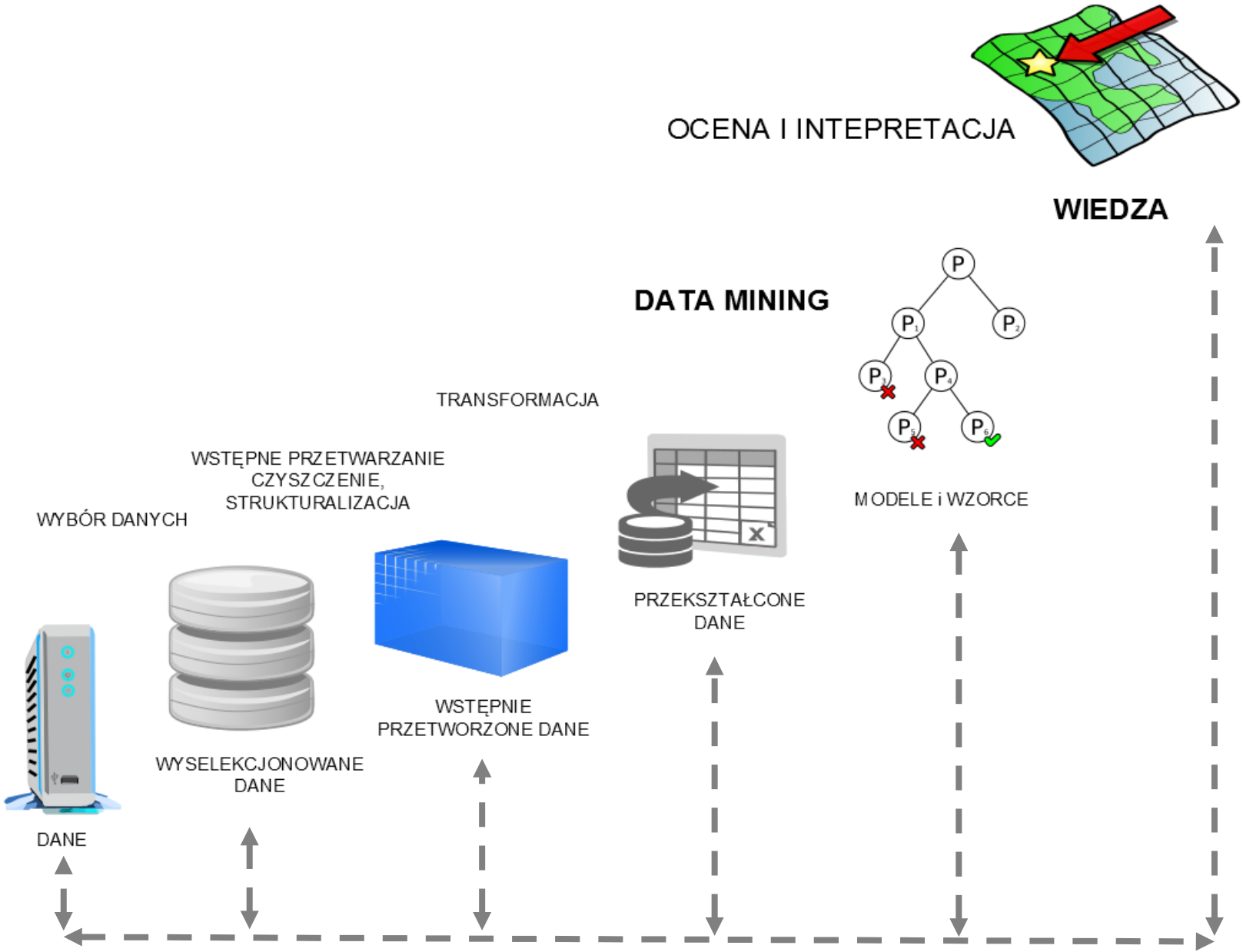
Rodzaje analiz - jak odkrywamy

Wizualizacja	Analiza, której celem jest odbiór abstrakcyjnych danych czy działania algorytmów w formie obrazów, map, animacji w sposób zrozumiały dla człowieka
Statystyka matematyczna i teoria prawdopodobieństwa	Reprezentowanie wiedzy jako prawdopodobieństwa w danych warunkach i określonym stopniu prawdziwości hipotezy.
Statystyka przestrzenna i geostatystyka	Odkrywanie przestrzennych trendów w danych przy użyciu struktury kowariancji i ew. zmiennych pomocniczych
Analiza rozmyta i przybliżona	Analiza wykonywana przy założeniu częściowej prawdziwości wyników, gdzie pojęcie przynależności znajduje się w przedziale od 0 do 1
Uczenie	Budowanie systemów klasyfikacyjnych na podstawie kolekcji wzorców
Predykcja	Stosowanie klasyfikatorów do nowych danych, czyli takich dla których zmienna zależna nie jest znana

Paradygmaty Data Science



Proces wydobywania wiedzy



Procedury analityczne

1) Zrozumienie problemu z danej dziedziny wiedzy i główne cele zbadanej dziedziny
2) Utworzenie docelową kolekcję danych - selekcja danych
3) Czyszczenie i wstępne przetwarzanie danych (często główna część pracy)
4) Redukcja i transformacja danych: Znaleźć użyteczne atrybuty, redukcja wymiarów, inna reprezentacja
5) Wybór odpowiednich narzędzi data mining: klasyfikacja, regrecja, asocjacja, grupowanie, ...
6) Wybór algorytmów: drzewa regresyjne, maszyny wektorowe, SOM...
7) Szukanie wzorców, modeli...
8) Ocena wzorców i prezentacja wyników: Wizualizacja, transformacja, usuwanie zbędnych wzorców ...
9) Zastosowanie odkrywanej wiedzy

Krótka definicja wzorca...

Wzorzec to reprezentacja wiedzy.

- Typy wzorców:
  - **subiektywne**: oparte o ufność (belief) użytkownika w dane, nowość, coś nieoczekiwanego
  - **obiektywne**: wynik analizy, oparte o wiedzę
- Cechy wartościowych wzorców:
  - musi być rozpoznawalny dla człowieka
  - regularny
  - oryginalny
  - użyteczny
  - prawdziwy dla nowych danych

Reguły i wzorce – co odkrywamy

Reguły	Opis	Przykład
Asocjacyjne *Associacion rules*	to logiczne połączania pomiędzy zjawiskami i obiektami (entity). W analizie przestrznnej pozwalają badać częstość obiektów występujących wspólnie w ramach jednego obszaru	Jeżeli DUZA KUCHNIA => Minimum 3 pokoje
Charakteryzujące *Characteristics rules*	Wspólne cechy zjawiska lub grupy zjawisk. Znajdowanie wyróżniających cech w tej samej klasie obiektów lub obszarów w analizie przestrzennej	Domy na przedmieściach są niskie
Odróżniające *Dyscrimianant rules*	Cechy, które pozwalają na odróżnienie jednego zjawiska lub grupy zjawisk (w analizie przestrzennej obszarów) od innych	Cena lokali na przedmieściach i w centrum
Porządkowe *Serial rules*	Reguły ograniczone czasowo-przestrzennie, które odnoszą się do związków zjawisk z czasem: powtarzalnością, trendem, występowaniem wzorców i podobieństw między sekwencjami	Wahania cen mieszkań są przesunięte względem wahań zatrudnienia
Grupujące *Clustering rules*	Reguły, która grupują zjawiska, obiekty lub obszary poprzez ich wzajemne podobieństwo,**bez wstępnej wiedzy** na temat liczby i charakteru docelowych klas (unsupervised). Klasy są interpretowane po zakończeniu procesu grupowania (a posteriori)	Grupowanie typów zabudowy
Klasyfikujące *Classification rules*	Reguły decydujące, czy dany obiekt (entity) należy czy nie należy do **uprzednio określonego typu klasy** (supervised). Liczba i charakterystyka klas jest znana przed (a priori) rozpoczęciem procesu klasyfikacji	Klasyfikacja obrazów satelitarnych
Prognozujące *Predictive rules*	Reguły pozwalające na przewidywanie występowania lub oszacowanie wartości zjawiska, jeżeli zmienią się wartości lub rozmieszczenie innych zjawisk. Przewidywanie nieznanych lub brakujących atrybutów	Ceny mieszkań spadną jesienią..
Wyjątki *Exceptions*	Obiekty lub zjawiska, które odbiegają znacząco od pozostałych.	Wyjątkowo drogie mieszkanie

Cele Data science – po co odkrywamy?

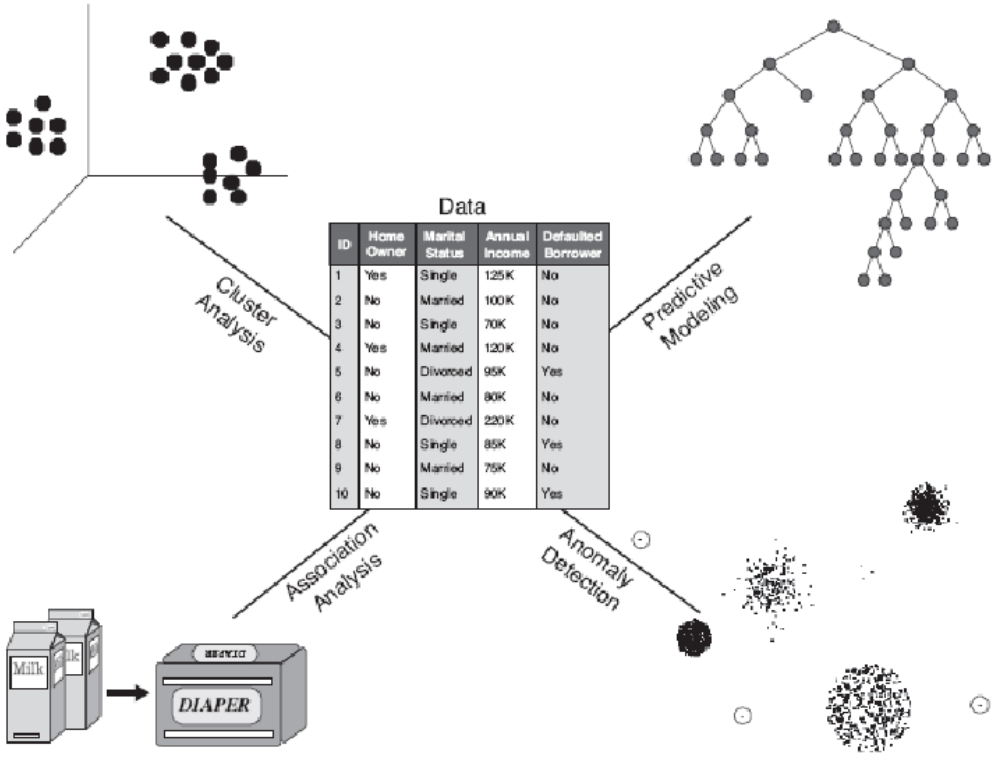
- **Przewidywanie:** określenie możliwej przyszłej sytuacji na podstawie danych historycznych (np. ocena ryzyka)
- **Opisywanie:** przyczyna dla której zachodzi jakieś zjawisko (np. dlaczego dochodzi do wypadku, dlaczego spada sprzedaż)
- **Weryfikacja hipotez:** czy jedzenie pomidorów zmniejsza zachorowanie na raka?, czy lokalny wzrost opadów powoduje zwiększenie ryzyka powodzi
- **Wykrywanie sytuacji nietypowych:** są przypadki wskazujące na nietypowe zachowania (wykrywanie oszustw)

Cele analizy danych i Data Science

Modelowanie predykcyjne

Opisywanie (streszczanie) danych	Wstępne informacje na temat danych, cechy, typy i zakresy danych, statystyki podsumowujące, tworzenie metadanych
Analiza jakościowa i ilościowa	Opisywanie zjawisk bez (jakościowa) lub z (ilościowa) użyciem parametrów liczbowych. Celem analizy jakościowej jest określenie klasy obiektu (zjawiska) lub stwierdzenie jego występowania a ilościowej podanie charakterystyk obiektów i zjawisk
Klasyfikacja	Nadawanie etykiet (klas, przynależności) obiektom, uprzednio niezaklasyfikowanym, na podstawie klasyfikatora wyuczonego na podstawie pozbioru, gdzie klasa jest już określona
Regresja/predykcja	Budowanie modeli pozwalających przewidzieć wartość (ilość) cechy lub zjawiska na podstawie innych wartości lub cech. Predykcyjne modele regresejne stosują tę samą klasę algorytmów co klasyfikacja
Grupowanie (uczenie nienadzorowane i częściowo nadzorowane)	Wykrywanie skupień w zestawie danych nie mających uprzednio określonej przynależności, dzielenie zbioru obiektów na naturalnych grupy (skupienia)
Wykrywanie asocjacji	Znajdowanie związków pomiędzy obiektami występującymi na tyle często, że nie można ich intrpretować jako przypadkowy
Znajdowanie wyjątków i wykrywanie nowości	Celem analizy jest znajdowanie obiektów i zjawisk odbiegających znacząco od większości (wyjątków) lub obiektów, które nie pasują do żadnej klasy stosowanego systemu klasyfikacyjnego (nowości)

Najważniejsze zadania



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Klasyfikacje nadzorowane ws nienadzorowane

Klasyfikacje:	Nadzorowane	Nienadzorowane
Zmienna zależna (znana)	jest	Nie ma
Zbiór uczący	jest	Nie musi być
Zbór testowy	jest	Nie ma
Ocena jakości	Wydajność względem zbioru testowego	Odrębność skupień
Typ zmiennych	dowolne	numeryczne
Klasy	A priori	A posteriori

Big data

- termin odnoszący się do **dużych, zmiennych i różnorodnych** zbiorów danych, których przetwarzanie i analiza jest trudna, ale jednocześnie wartościowa, ponieważ może prowadzić do zdobycia nowej wiedzy.
- W praktyce pojęcie dużego zbioru danych **jest względne** i oznacza sytuację, gdy zbioru nie da się przetwarzać przy użyciu trywialnych, powszechnie dostępnych metod
- Dziedziny:
  - Astronomia i Teledetekcja
  - Fizyka
  - Bankowość i finanse
  - Diagnostyka medyczna
  - Telekomunikacja
  - Szpiegostwo

Terrorbytes

Problemy związane z wydobywaniem wiedzy

- Różne typy wiedzy, różne typy abstrakcji
- Użycie wiedzy zastanej
- Szumy i dane niekompletne
- Złożone typy danych
- Ważność odkrytych wzorców i ich ocena
- Wydajność i skalowalność algorytmów
- Strategie przeszukiwania (podejście heurystyczne)
- Zastosowanie wydobytej wiedzy i jej integracja z istniejącą wiedzą
- Bezpieczeństwo i prywatność

Zastosowania data science

- Biznes i ekonomia,
- marketing, reklamy kierowane
- Ubezpieczenia, ocena ryzyka
- Bankowość i wykrywanie nadużyć (fraud detection)
- Diagnostyka medyczna, badania genetyczne
- Rozpoznawanie obrazów, pisma
- Zarządzanie ryzykiem, wykrywanie oszustw, NSA
- Indeksowanie tekstów (AI), inteligentne systemy wyszukiwania informacji

Spatial Data Science

- Celem klasycznej eksploracji danych jest poszukiwanie nowych, nieoczywistych, ukrytych wzorców/wiedzy
- Celem przestrzennej eksploracji danych jest znajdowanie interesujących wzorców/układów obejmującą zarówno cechy przestrzenne jak i nieprzestrzenne
- Główna różnica to założenie że na obiekt mają wpływ nie tylko jego cechy (jawnie - *explicit*) ale też obiekty i cechy obiektów sąsiednich (niejawnie - *implicit*)

Statystyczne podstawy

Metoda	Opis	Przykład
Analiza układów punktowych	Analiza rozkładu przestrzennego obiektów punktowych i liniowych, skupień przestrzennych, losowości przestrzennej	Analizy kernelowe zbiorów punktowych/linijnych
Analiza siatek	Analiza regularnych i nieregularnych obiektów przestrzennych znajdujących się w relacjach topologicznych poprzez wspólne sąsiedztwo	Regresja i autokorelacja przestrzenna, Moran I,
Geostatystyka	Analiza ciągłości zjawisk i zróżnicowania przestrzennego, badanie trendów i stacjonarności	Predykcja nieznanej wartości na podstawie lokalizacji

Ogólne i przestrzenne strategie

Zadanie	Ogólne	Przestrzenne
Optymalizacja	„Dziel i zwyciężaj”	Podział na jednostki przestrzenne
Filtrowanie	Filtrowanie na podstawie atrybutów	Filtrowanie na podstawie zasięgu
Sortowanie	Sortowanie	Generalizacja przestrzenna
Indeksowanie	Standardowe indeksy bazodanowe, B-Trees	Przestrzenne struktury hierarchiczne (R-Trees)
Brakujące dane	Estymacja na podstawie pozostałych atrybutów	Estymacja na podstawie autokorelacji przestrzennej

Zastosowania Geo-data Science

- Klasyfikacja obrazów satelitarnych
- Wykrywanie cech (np. stanowisk archeologicznych)
- Analiza trendów czasowych (klimatologia, hydrologia)
- Modelowanie ekologiczne i paleogeograficzne
- Analiza zagrożeń (powodzie, pożary, osuwiska, erozja itp.)
- Wykrywanie zależności geoprzestrzennych (związki przestrzenne między obiektami)


data
influenced
spatial
now
accuracy
need
concepts
areas
discipline
traditional
perspective
Crampton
methodological
level
humanity
geospatial
approach
geographic
make
changing
interrelations
domain
improvement
background
increasingly
different
research
making
side
technology
cartograms
thesis
base
sensed
social
part
work
geography
define
within
regard
aims
human
main
processes
still
help
map
particular
results
assessed
using
additional
focus
chapter
natural
cartographic
discussed
see
development
needs
statistical
showing
complex
means
new
science
always
practice
dissertation
cartography
looks
visualising
maps
developed
worldmapper
initial
understanding
many
Chapter
links
element
integrated
Hennig
scope
remotely
issues
dimensions
spaces
underlying
study
understand
mapping
academic
visualisation
world
environments
piece
often
view
patterns
findings
methods
information
analysis
globalised
conventional
contemporary
investigated
investigation
follows
time
digital
contribute
technical
complexity
knowledge
physical
geographical
general
fields
questions
explain
population
way
environment
also
used
sense
visual
focuses
ways
project
interdependencies
visualise
determine
special
understood
implications
Cartographic
contribution
provide
role
living
expectations
aesthetic
wider