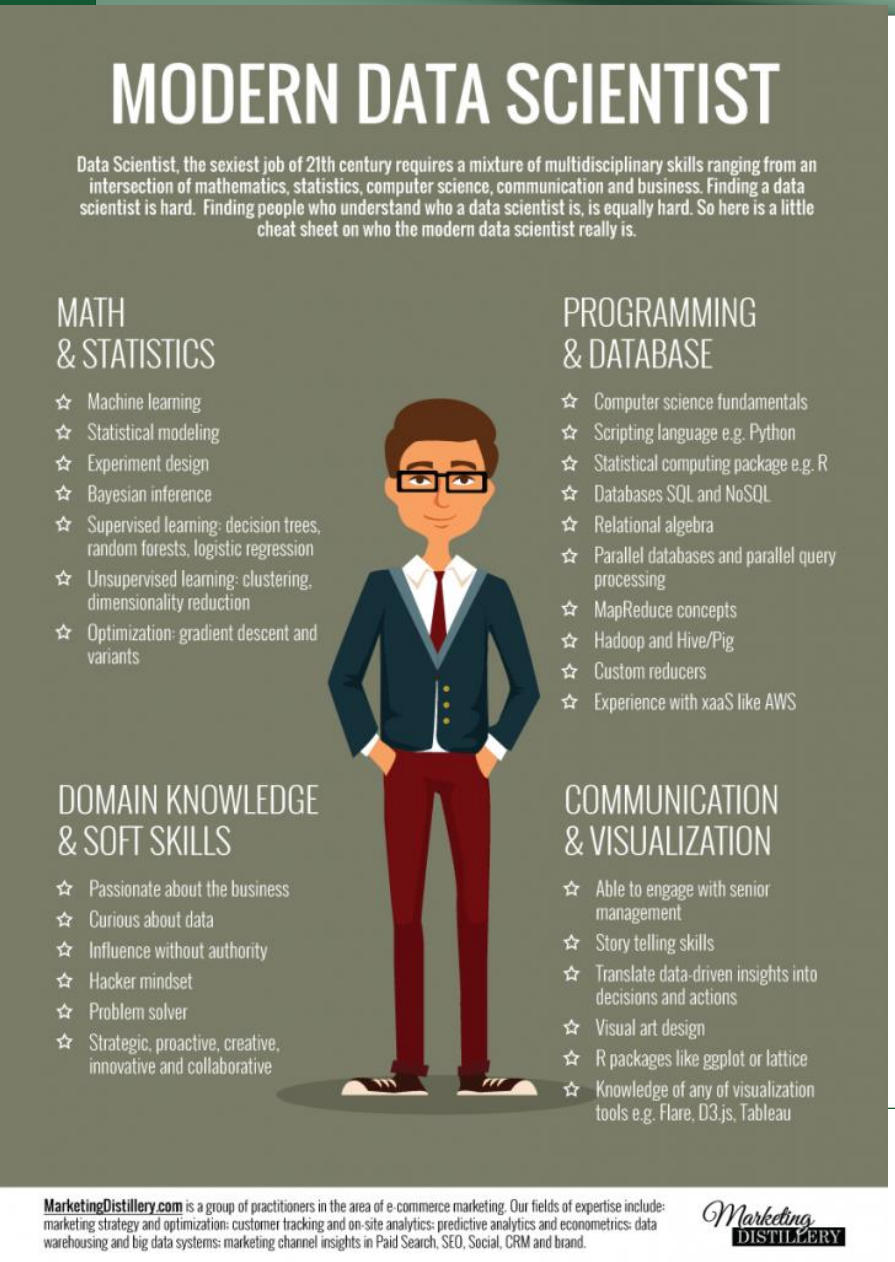
Jak kopać w danych - sieci społeczne i data mining

Łukasz Ryniewicz

Bank Zachodni WBK

Grupa Santander

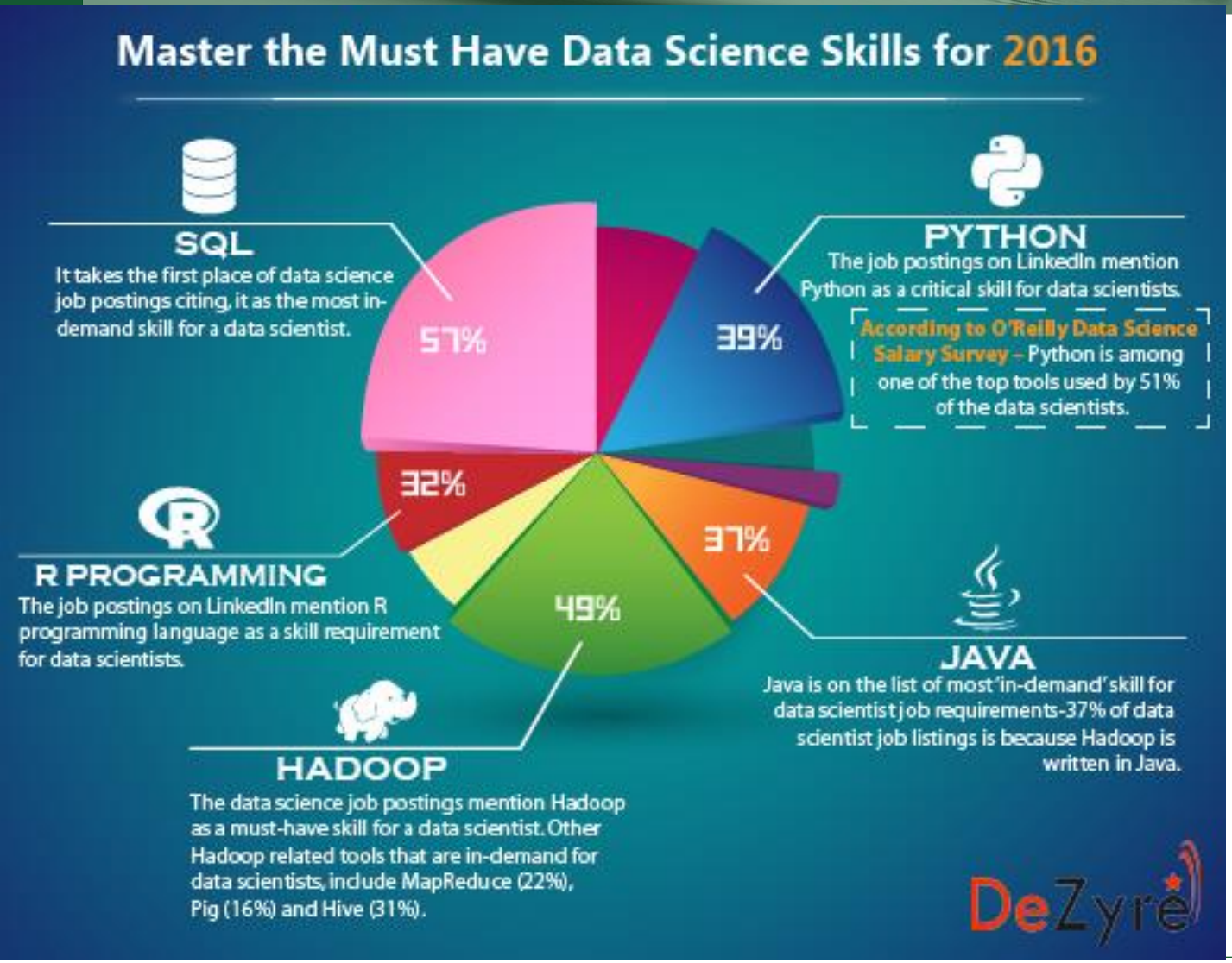
Data Scientist



A w Banku…

- Wiedza o produktach bankowych
- Wiedza o regulacjach
- Umiejętność poszukiwania dobrych danych
- Umiejętność wyjaśnienia analiz
- Znalezienie przełożenia na biznes

Data Scientist



- SAS 4GL
- SAS VA
- Unix (shell)
- Control-M
- Spark
- AbIninito

Data Scientist

- Ciągłe poszukiwanie rozwiązań, łączenie faktów, analiza nowych źródeł
- Śledzenie rynku, co robi konkurencja, czego nie robi, co można zrobić lepiej
- Śledzenie technologii, co można z nią robić, nowe funkcjonalności

BigData

Obecnie zbieramy dużo danych

- Volume (duża ilość, objętość danych)
- Variety (duża różnorodność danych)
- Velocity (duża szybkość zmian)
- Veracity (duża wiarygodność)
- Value (duża wartość)

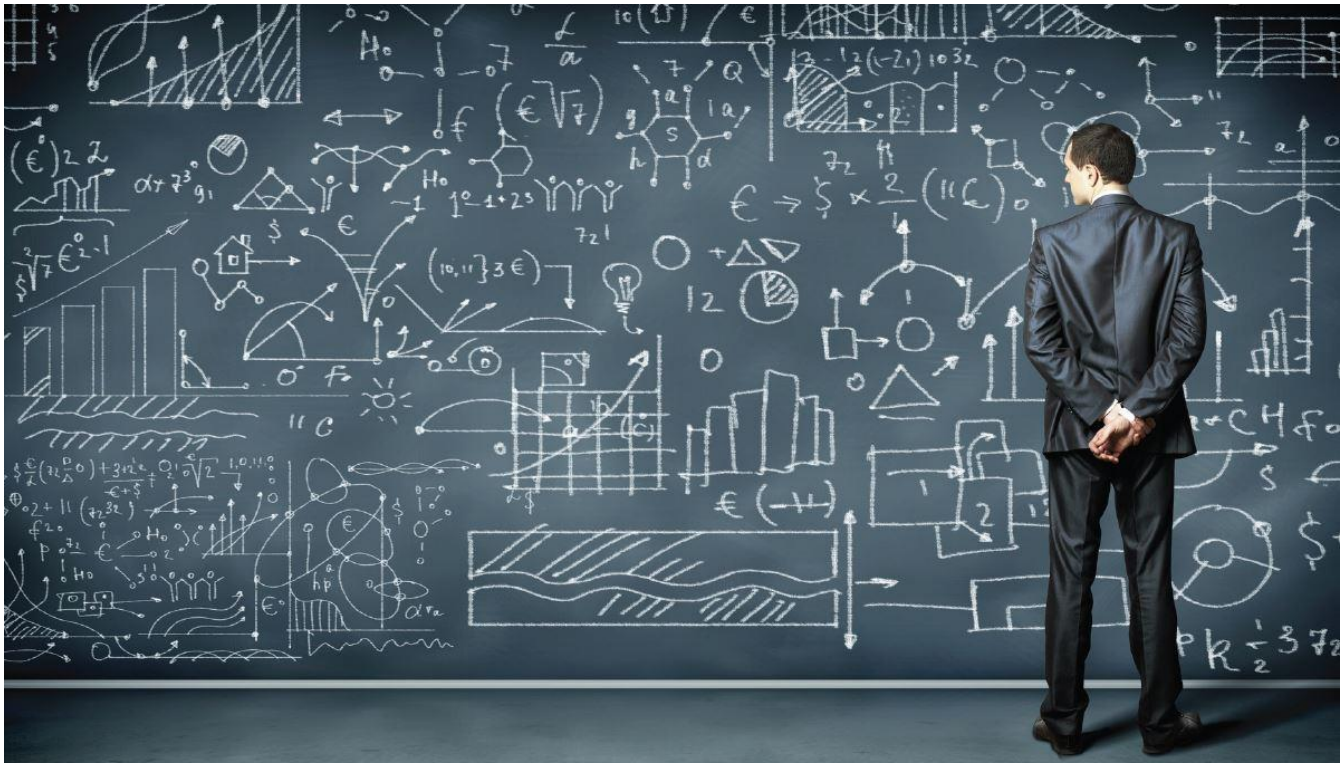
Często są to:

- Dane nieuporządkowane
- Dane nieoczyszczone
- Dane niespójne

Dane z różnych systemów, formatów

- Unix, Windows, Android, etc.
- Txt, csv, dat, sas, etc.
- Teradata, Oracle, MSSQL, etc.

BigData → Smart Data

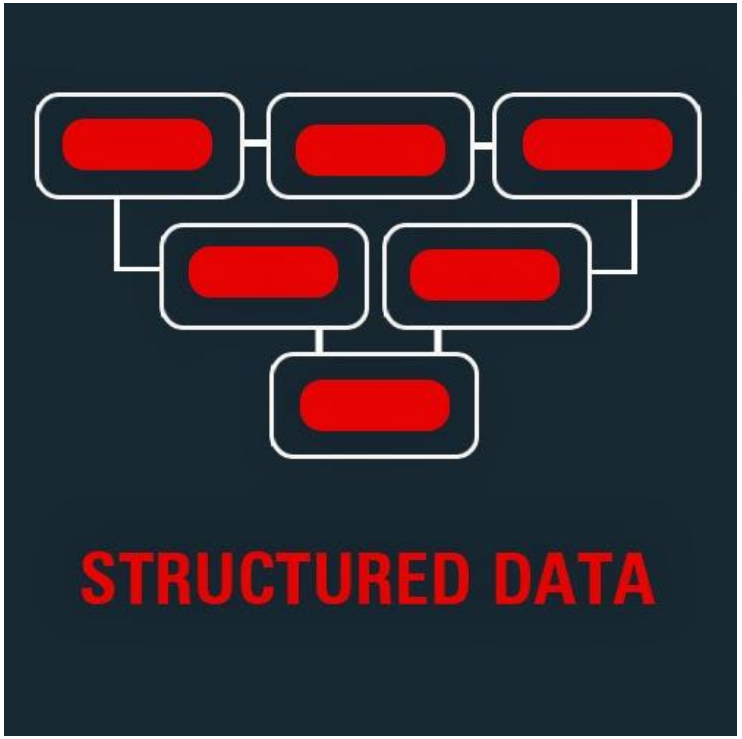


- Dane dla większości osób nieczytelne, nie możliwe do pracy na nich
- Nie można wyciągać wprost wniosków z danych
- Małe możliwości analityczne

→ Trzeba coś z nimi zrobić

Smart Data

- Uporządkowane
- Pogrupowane
- Zinterpretowane
- Przekładają się na podejmowanie właściwych decyzji
- Rozwiązują problem
- Łatwy i zrozumiały dostęp
- Możliwe użycie wielu narzędzi do tych danych



Sprowadzenie danych do poziomu klienta

- Jak? → Agregacje, łączenia, słowniki, rozproszenia
- Ułatwia analizy, pracę wielu działów (CRM, Ryzyko, Fraudy, etc)
- Czasem nie takie łatwe:
  - ID klienta, WH_ID, NIK, CUSTOMER_ID, IBAN, PESEL
  - Tworzenie map do poziomu klienta CUST_ID
  - Potrzebne przejście przez wiele relacji (np. karta → konto → typ właścicielstwa)

Źródła danych

- Hurtownia danych
- Systemy transakcyjne
- Systemy kartowe
- Logi strony www
- Logi aplikacji mobilnej
- Dane z wniosków kredytowych
- Dane o ubezpieczeniach kredytów

- Różne systemy
- Różne formaty
- Różne definicje (czasem dla tego samego bytu)

Jak zmapować dane

Dla każdej zmiennej

- Nazwa, opis
- Pochodzenie: z jakich systemów, jakie formaty
- Jakie reguły słownikowe
- Jakie reguły mapujące
- Jakie reguły agregacji
- Itp..

Ile tych danych

- Ponad 12 000 finalnych zmiennych (per klient)
- Grupowane w ponad 20 obszarach
- Dane źródłowe pochodzące z parunastu systemów
- Przetwarzane ponad 100 000 000 000 rekordów…

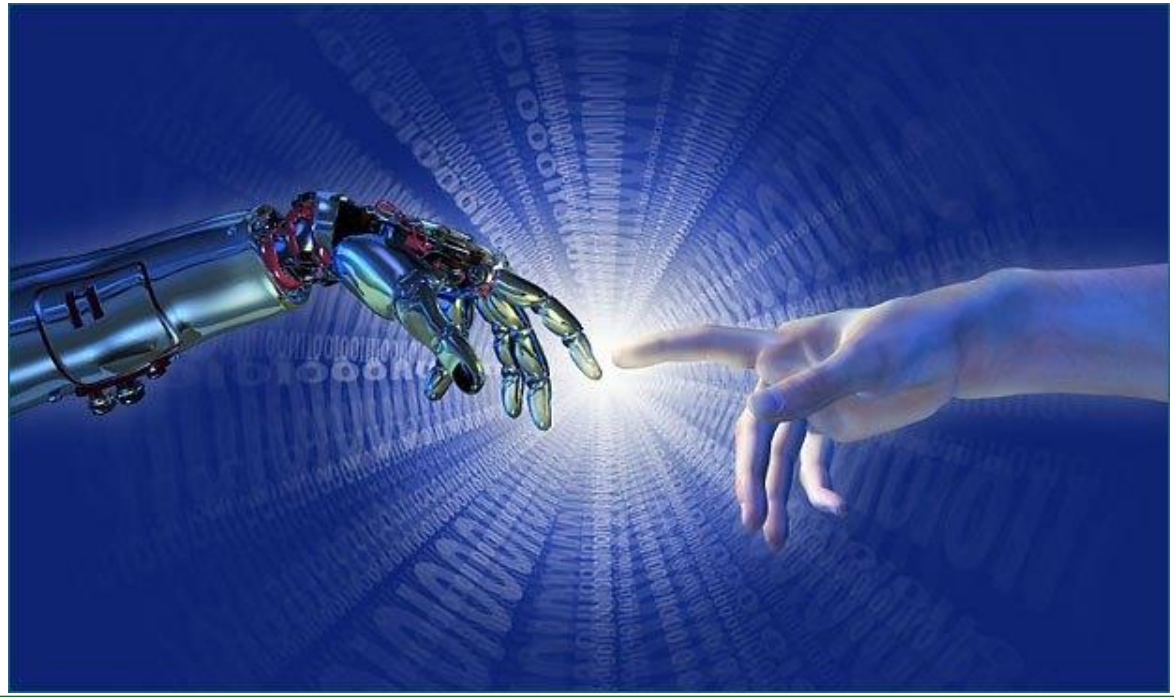
Jak te dane przetworzyć…

1. Wnikliwa analiza wszystkich zmiennych
2. Wyłapanie jak największej liczby zależności
3. Podział na etapy
   a. Dane źródłowe
   b. Operacje na danych źródłowych
   c. Agregacje, wyliczenia
4. Znalezienie jak największej liczby wzorców
5. Określenie jakie struktury danych będą mogły te wzorce obsłużyć
6. Znalezienie jakie potrzebne są transformacje by te struktury zasialć
7. Określenie reguł atomowych na poszczególnych transformacjach
8. Jak łatwo wynieść reguły atomowe do parametryzacji
9. Stworzenie parametryzacji
10. Stworzenie procesu
11. Testy
12. Wdrożenie LIVE
13. Parametryzacja

Jak powstają nowe zmienne

Proces powstawania nowych zmiennych można sprowadzić do:

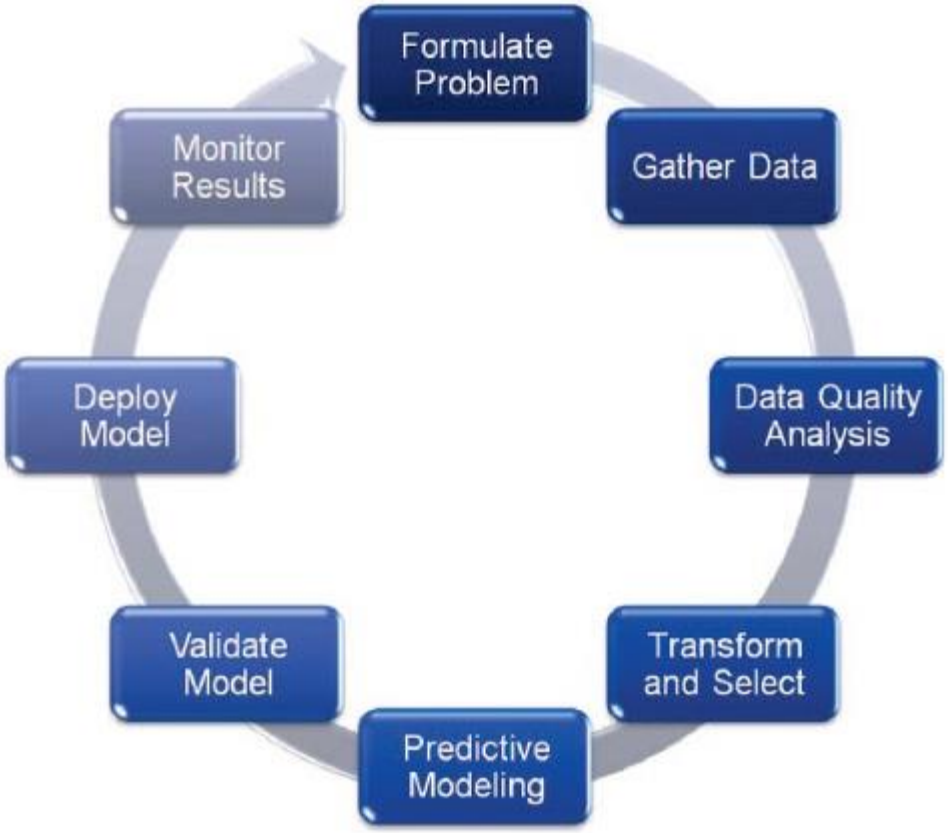
1. Identyfikacja potrzeby biznesowej
2. Analiza dostępności źródeł do zaspokojenia danej potrzeby biznesowej
3. Określenia listy charakterystyk i wykonania wstępnej specyfikacji na źródłach (np. SQL)
4. Stworzenie finalnej specyfikacji w narzędziach Środowiska
5. Wykonanie testów
6. Wdrożenie



Data Mining

Co chcemy modelować

- Kto będzie chciał Kartę Kredytową
- Kto będzie chciał wziąć kredyt
- Kiedy te zdarzenia nastąpią
- Jaki kolejny produkt sprzedać klientowi
- Jakie jest ryzyko udzielenia kredytu danej osobie
- Czy klient chce odejść
- Jaki kanał jest najlepszy dla klienta



Populacja, target, wybór zmiennych

Dla każdego modelu musimy zdecydować co jest jego targetem

- Wzięcie kredytu, odejście z banku, etc.
- Wtedy tworzymy nową zmienna i dla tych warunków wstawiamy 1

Potrzebujemy określić populacje docelową

- Zawiera wykluczenia biznesowe (np. klienci starsi niż 40 lat, z co najmniej 3 miesięczną historią)
- Zawiera wykluczenia techniczne (wyrażone odpowiednie zgody, wypełnione odpowiednie dane, klienci poniżej 18 lat, klienci z flagą zgonu)

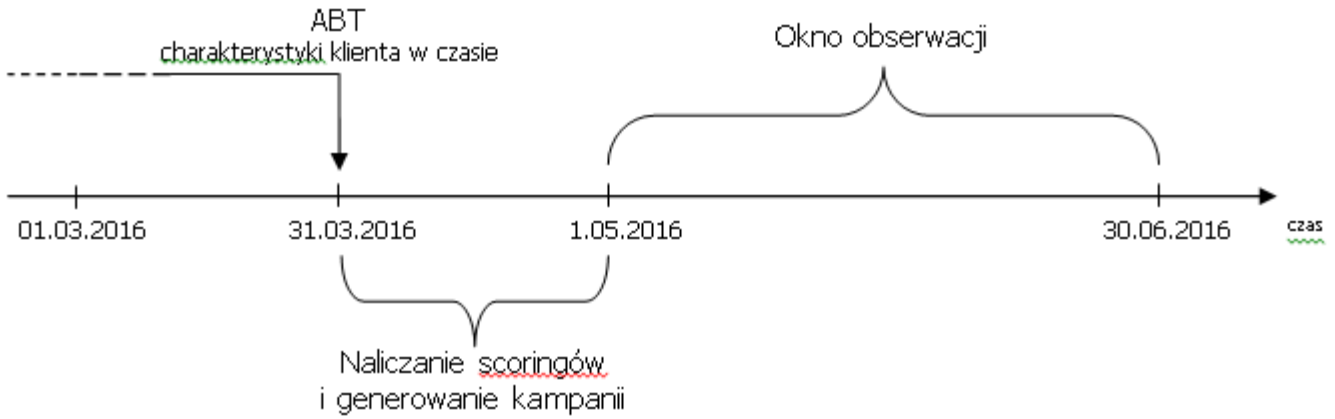
Wybór zmiennych (do modelowania)

- Zmienne stabilne w czasie
  - testy KS i Tstudenta dla ciągłych
  - Chi-kwadrat i Population stability index (PSI) dla kategorycznych
- Zmienne niepuste (określa się próg wypełnienia)
- Ocinanie wartości odstających
- Często doprowadza się zmienne do rozkładów symetrycznych, standaryzacja zmiennych
- Ograniczenie liczby zmiennych < 500

Dane do modelowania

Okno obserwacji

Okno obserwacji obejmuje 2 miesiące (maj-czerwiec 2016), następujące miesiąc po okresie, z którego pochodzi ABT (marzec 2016).

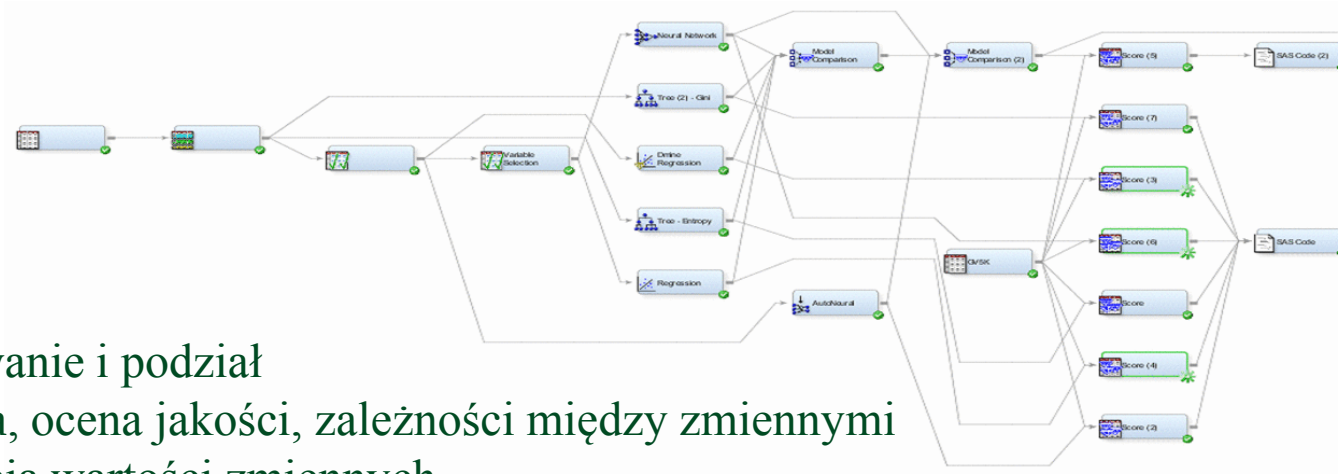


- Próba jest losowana, tak by stosunek sukcesów (target) był zadowalający (min 30%)
- Stosuje się przepróbkowania
- Gdy mało targetów w jednym okresie, dodaje się kolejne okresy

Modelowanie

W zależności kto jest odbiorcą modelu są dodatkowe narzuty

- Ryzyko: nie można używać blackbox (sieci neuronowe, lasy losowe, itp.)
  - Wymóg prawny, każdą zmienną należy przed regulatorem umieć wyjaśnić
  - Modele muszą być budowane zgodnie z obwiązującymi regulacjami
- CRM: tylko dla osób z odpowiednią zgodą marketingową
  - Brak restrykcyjnych regulacji



Metodyka SEMMA

- Sample – przygotowanie i podział
- Explore – exploracja, ocena jakości, zależności między zmiennymi
- Modify – modyfikacja wartości zmiennych
- Model – modelowanie
- Assess - ocena jakości modeli, wybór najlepszego

Best Practices

Zgodnie z dobrymi praktykami spośród zmiennych wybierane są następujące, spełniające wstępne kryteria:

- Poziom wypełnienia zmiennej – co najmniej 60%
- Udział obserwacji niezerowych – powyżej 30%
- Statystyka Kolmogorova-Smirnova (KS) – co najmniej 0.07 (dla zmiennych ciągłych)
- Statystyka t-Studenta (p-value) – maksymalnie 0.05 (dla zmiennych ciągłych)
- Population Stability Index (PSI) – co najmniej 0.2 (dla zmiennych znakowych)
- Współczynnik V Cramera - co najmniej 0.2 (dla zmiennych znakowych)

Dla tak wybranych zmiennych wykonuje się następujące kroki:

- kategoryzacja zmiennych i przekodowanie na wartości WOE
- analiza czynnikowa i wybór reprezentantów klastrów
- usunięcie zmiennych skorelowanych

WOE (Weight Of Evidence) – transformacja zmiennych ciągłych do zmiennych kategorycznych przy równoczesnej maksymalizacji statystyki Gini (mierzącej moc zależności pomiędzy zmienną objaśniającą i zmienną zależną).

Gini (Somers'D) – określa moc i kierunek zależności pomiędzy zmienną objaśniającą i zmienną celu. Przyjmuje wartości od -1 do 1.

$$Gini = (n_c - n_d)/t$$

gdzie n_c - liczba par zgodnych, n_c - liczba par niezgodnych
oraz t – liczba wszystkich par złożonych z obserwacji obserwowanej i przewidywanej.

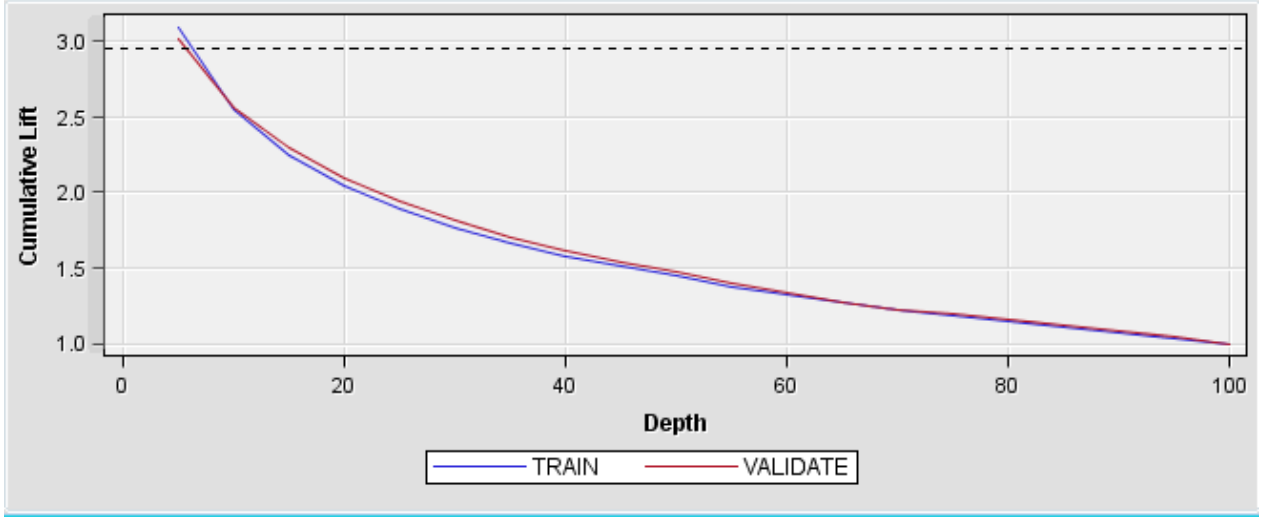
$$WOE_i = \ln\left(\frac{\%sukcesów_i}{\%nonsukcesów_i}\right)$$

gdzie i - i-ta kategoria zmiennej objaśniającej.

Ocena modelu

Do finalnej oceny modelu wykorzystujemy:

- Współczynniki korelacji rang Spearmana
- Gini
- KS
- Skumulowany lift 10%
- Skumulowany lift 20%



Scoring

Każdy model produkuje wynik który nas interesuje: prawdopodobieństwo 1

Scoring zawiera zestaw reguł które określają to prawdopodobieństwo

Często (zwłaszcza w Ryzyku) tworzy się też karty scornigowe

Po naliczeniu danych za dany okres nalicza się scoringi dla wszystkich modeli na tych danych

Scoringi są dystrybuowane do odbiorców

Monitoring

Monitorowanie zmiennych

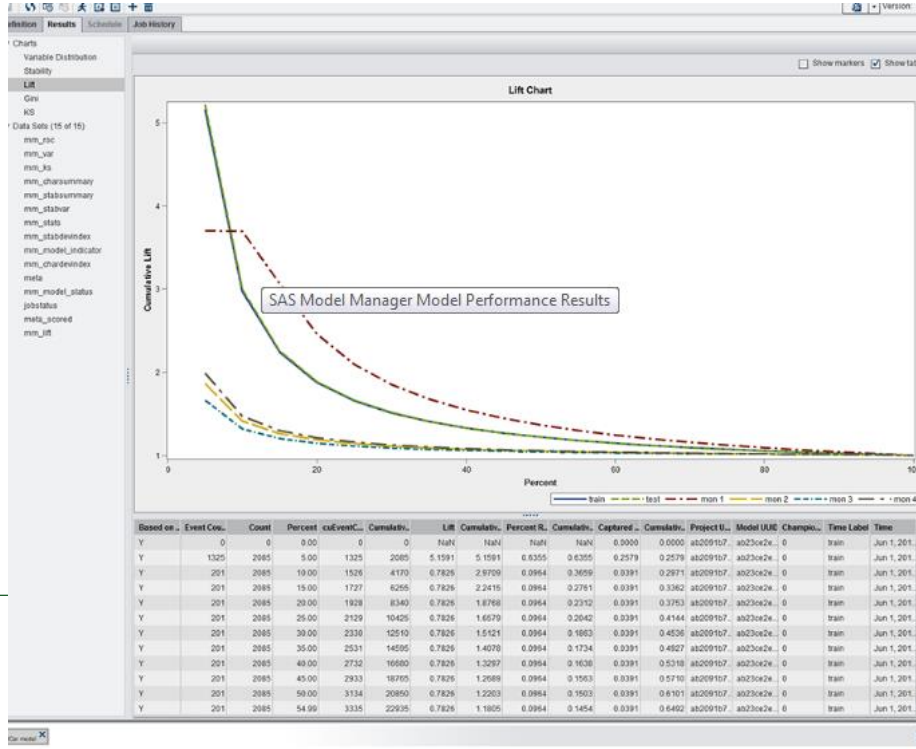
- Wszystkich zmiennych wyliczanych cyklicznie – podstawowe statystyki
- Zmienne użyte w modelach – rozszerzone statystyki

Monitoring jakości i stabilności modeli

- Jakość i wielkość prognozowanych targetów
- Lift
- Gini
- KS

Cykliczne odświeżanie modelu

- W wyniku dryfu w danych
- Zmiany w sytuacji na rynku

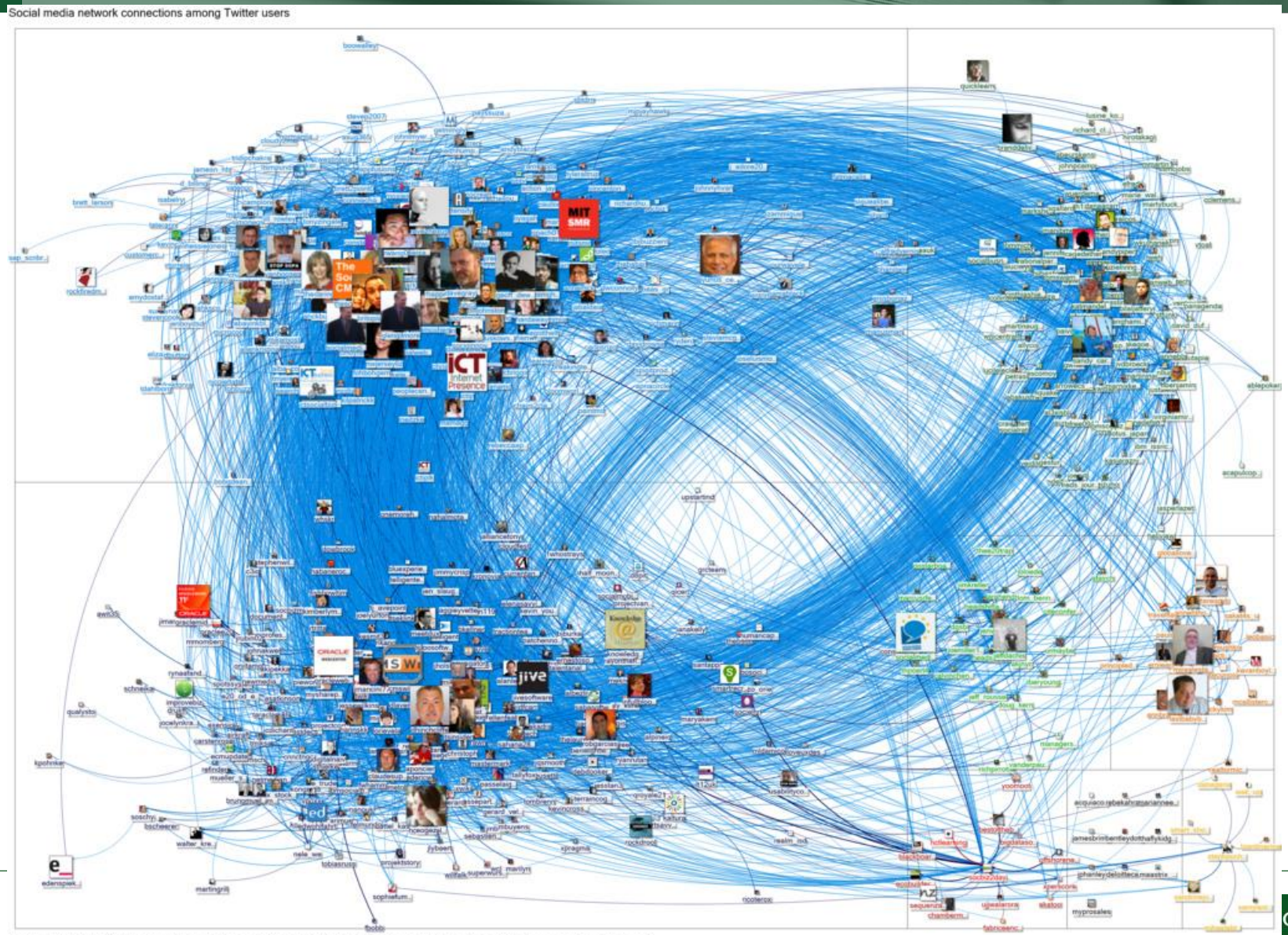


I co z tych modeli…

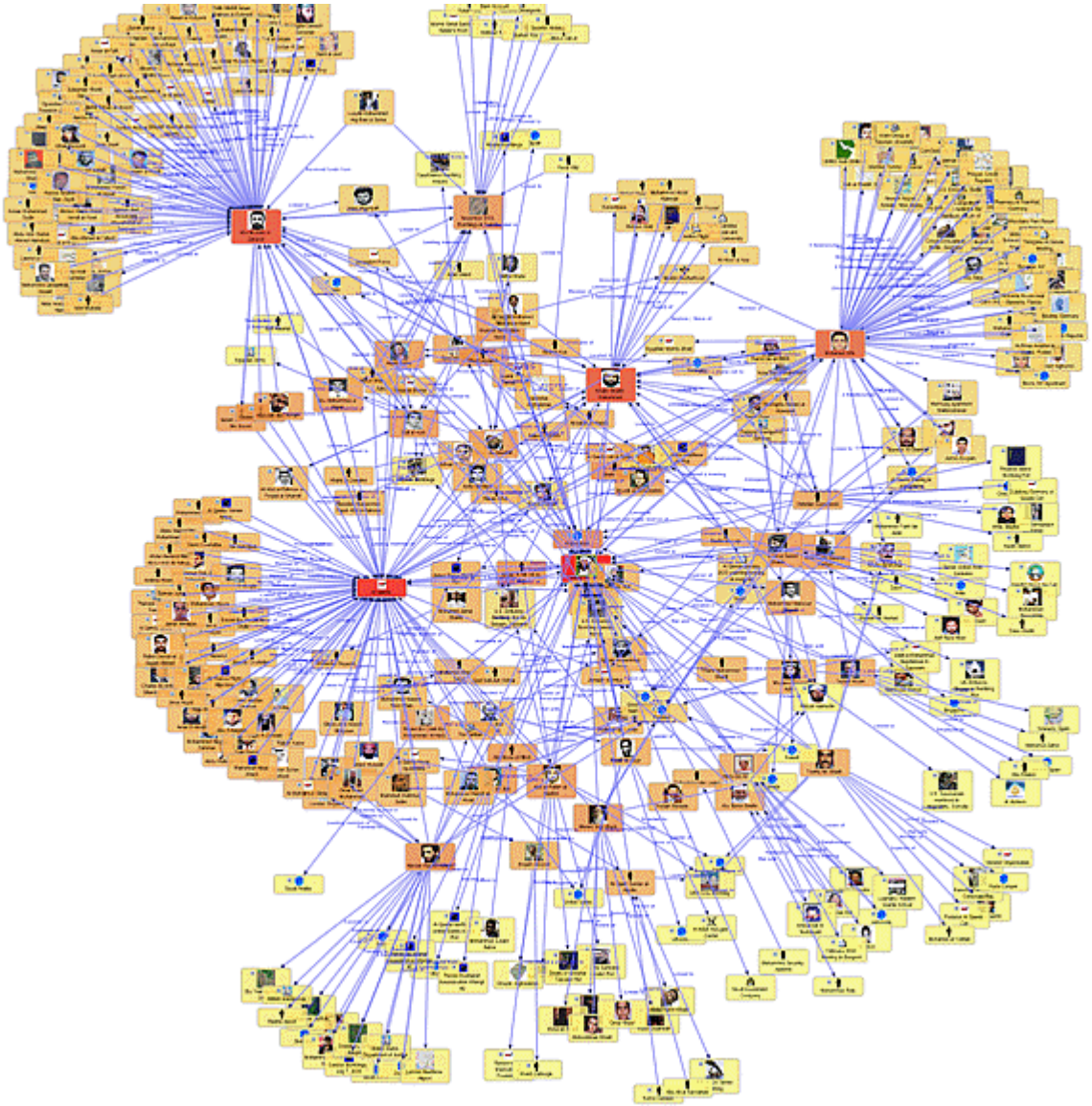
Jakość modeli przekłada się na zysk dla Banku w postaci

- Mniejszego ryzyka udzielenia złych kredytów
- Lepszego zarządzania i prognozowania ryzyka
- Lepszego doboru prób do kampanii marketingowych

Sieci społeczne



Sieci Społeczne



Sieci społeczne

Zupełnie inny rodzaj przedstawiania danych

- Obrazujemy zależności społeczne zachodzące między ludźmi
- Możemy w różny sposób określać siłę relacji
  - Na postawie liczby kontaktów
  - Na podstawie ich czasów
  - Na podstawie ich jakości
- Przedstawiamy dane w zupełnie nowy sposób
- Znajdujemy grupy społeczne
- Znajdujemy grupy znajomych
- Możemy określić jak ludzie faktycznie żyją i jak się zachowują

SNA gdzie stosowane

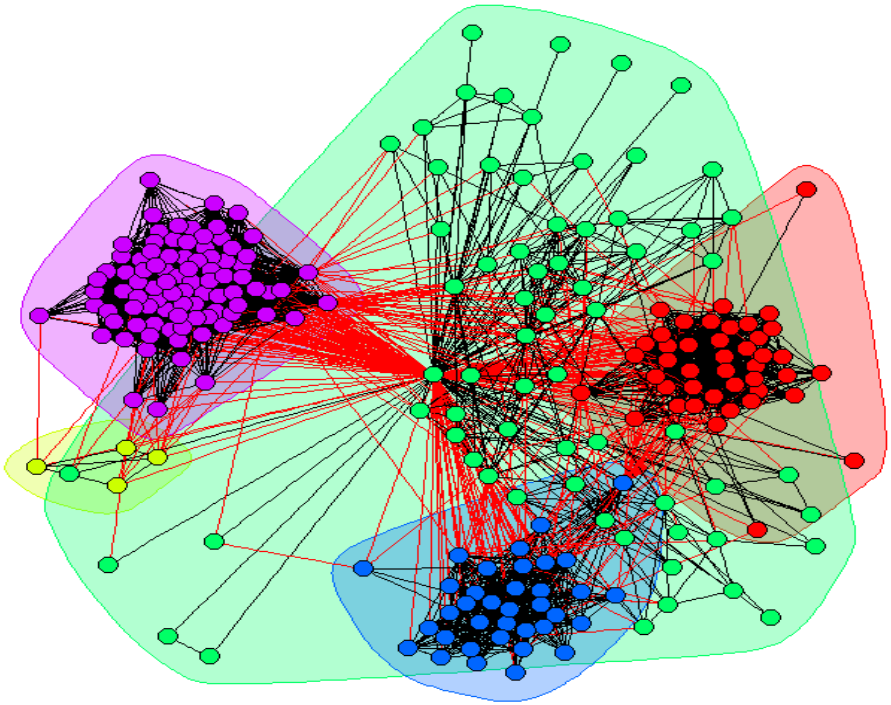
Zastosowania Social Network Analysis (SNA)

- Rozprzestrzenianie się wiadomości, polaryzacja polityczna
- Rozprzestrzenianie się wirusów i chorób
- Rozprzestrzenianie się wiadomości (marketing wirusowy)
- TELCO:
  - Analiza budowy nowych nadajników
  - Analiza powiązań miedzy klientami → lepsze kampanie, oferty
- Ubezpieczenia;
  - Analiza zdarzeń do wykrywania fraudów
- Analiza cytowań → kto jest „najlepszym" naukowcem
- Analiza stron WWW → Google PageRank ☺
- Analiza powiązań terrorystów → odnajdywanie kluczowych osób (Saddam)

SNA w Banku

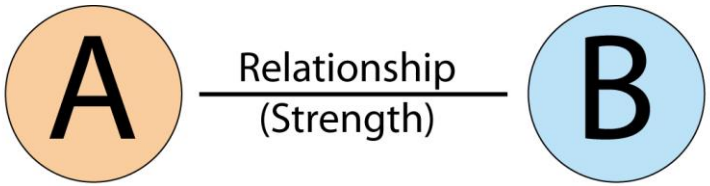
W banku trochę trudniej o dobre dane, ale możliwe są również zastosowania SNA

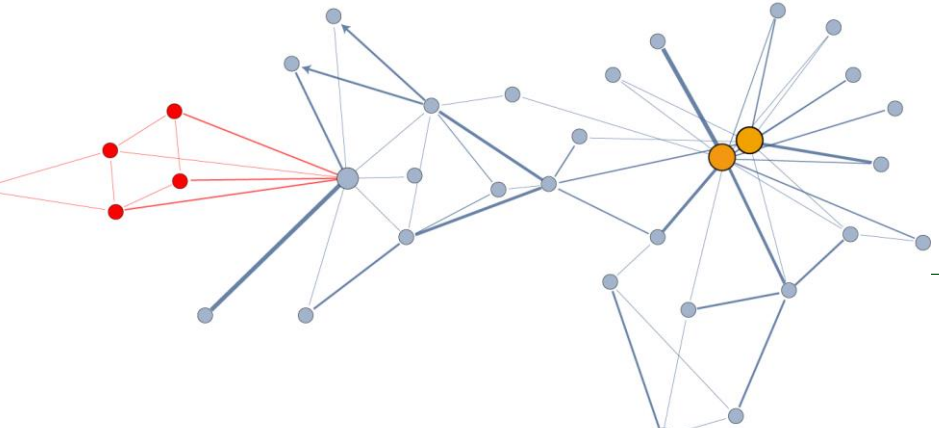
- Budowanie grafów powiązań między klientami
- Śledzenie przepływów finansowych → wykrywanie fraudów, terroryzm
- Analiza rentowności firm SME i ich wpływu na pozostałe
- Znajdowanie społeczności i liderów → lepsze kampanie CRM
- Lepsza ocena ryzyka dla nowych klientów
- Wzbogacenie danych o kliencie



SNA jak się to robi

1. Należy znaleźć relacje między ludźmi
2. Należy zagregować te relacje
3. Dla każdej relacji wyznaczyć funkcję wagi
4. Znaleźć społeczności
5. Wyliczyć miary wierzchołkowe
6. Znaleźć liderów





SNA jak się to robi

Dane agregowane do unikalnych par (FROM, TO)

Na każdej parze wyliczane wielkości, np.:

- Liczba wspólnych przelewów
- Liczba wspólnych płatności w POS po godzinie 18 w weekendy
- Liczba wspólnych wniosków o kredyt hipoteczny itd.

Na podstawie tych miar wyliczana funkcja wagi

- Kombinacja liniowa powyższych miar

Odfiltrowanie niepotrzebnych wierzchołków

- Płatności
- Taxi, pizza
- PayU itp..

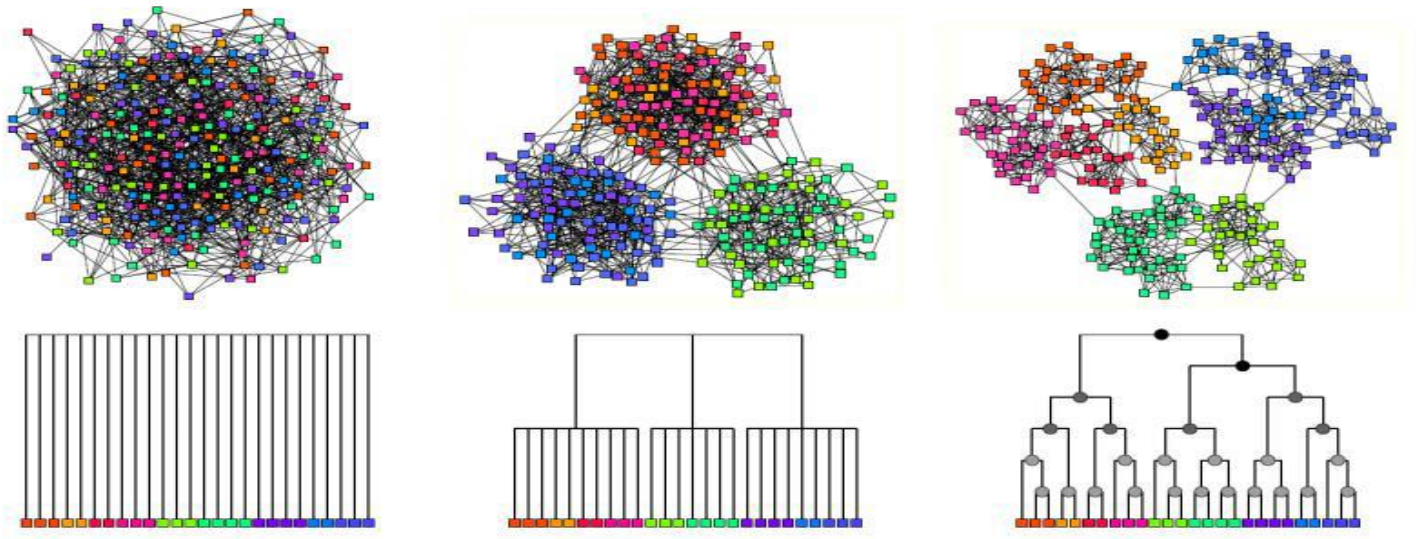
Budowa finalnego grafu

- Krawędzie (FROM, TO)
- Wagi (funkcja wagi)

SNA jak się to robi

Mając graf szukamy społeczności – grup wierzchołków silniej powiązanych miedzy sobą niż pomiędzy takimi grupami

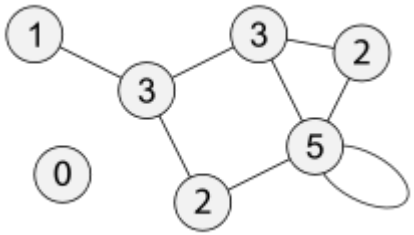
- Optymalizacja modularności grafu
- Algorytm Louvain
- Algorytm Label propagation
- Algorytmy hierarchiczne

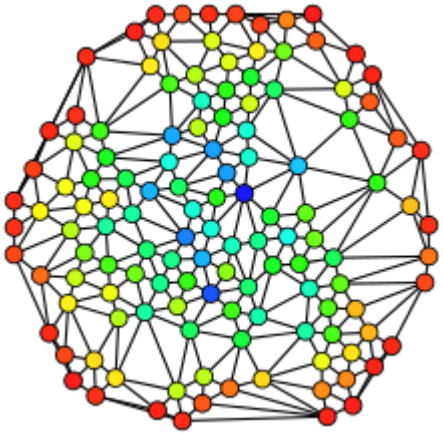


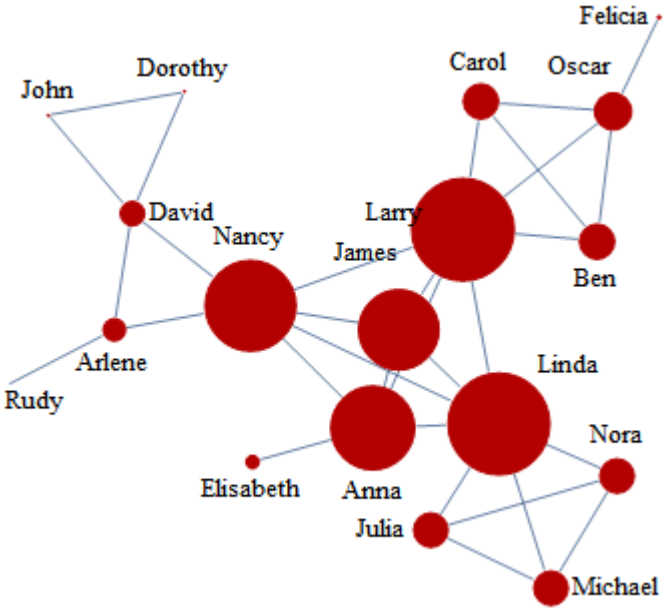
SNA jak się to robi

Mając społeczności można określić role ich członków na podstawie miar

- Stopień wierzchołka
- Pośrednictwo
- Zasięg
- Centralność
- Ważność







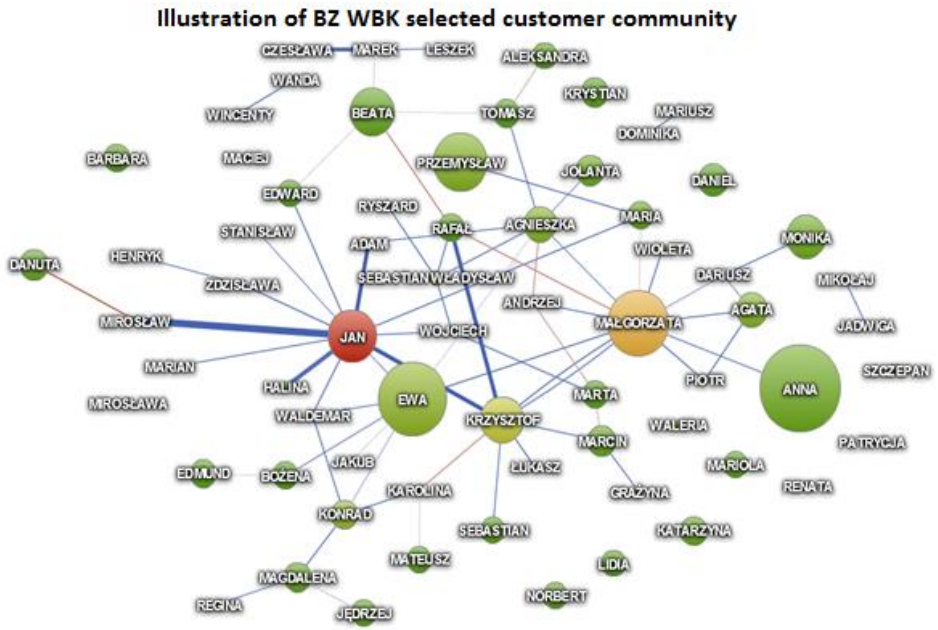
SNA technologie

Do przetwarzania danych, tworzenia grafów

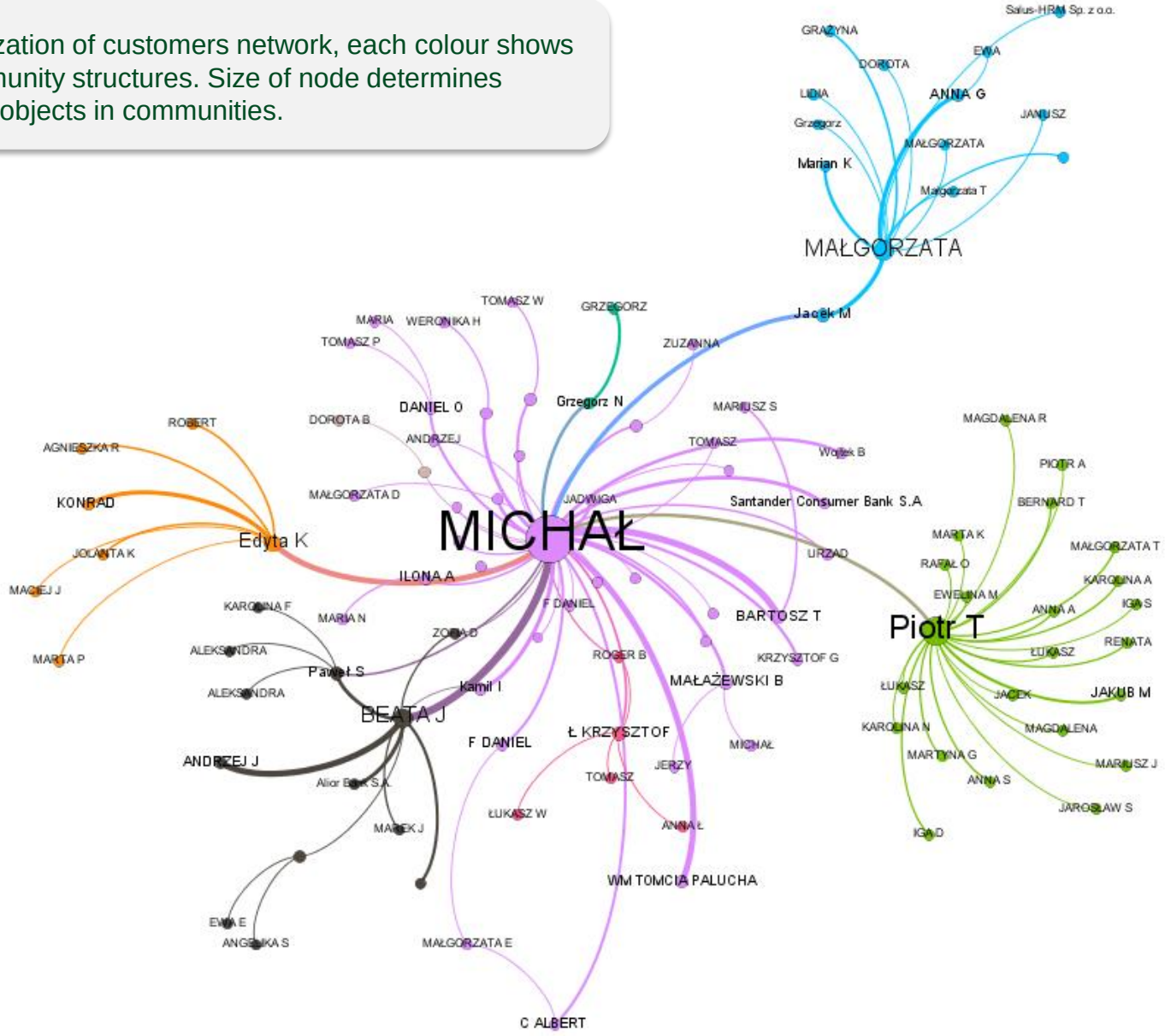
- SAS implementacje własne (jest też OPTGRAPH)
- R (sna, igraph)
- Python (SNAP, NetworX, igraph)
- Scala (głównie implementacje własne)

Wizualizacje

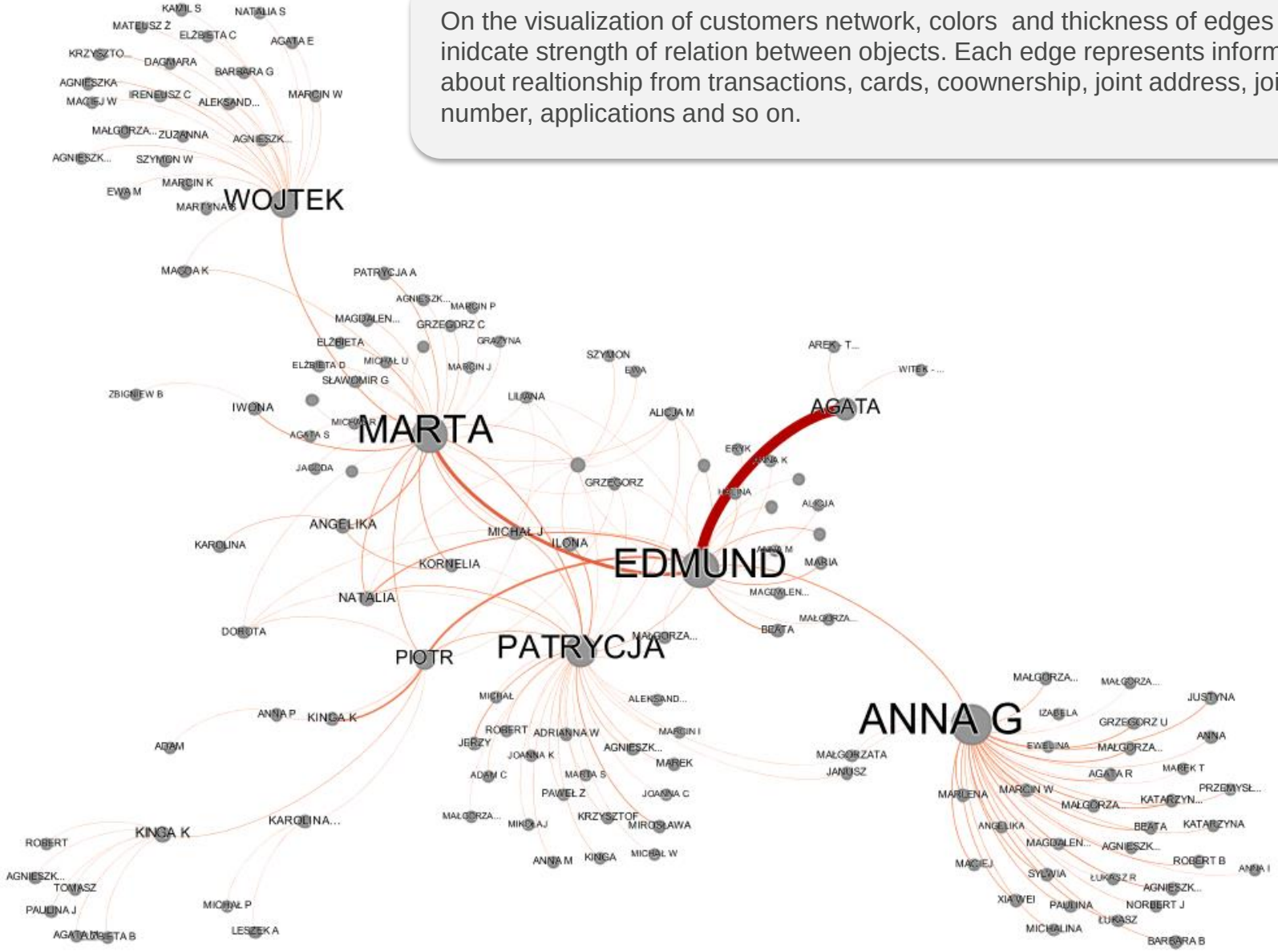
- SAS VA
- Gephi



On the visualization of customers network, each colour shows different community structures. Size of node determines importance of objects in communities.



On the visualization of customers network, colors and thickness of edges (links) inidcate strength of relation between objects. Each edge represents information about realtionship from transactions, cards, coownership, joint address, joint phone number, applications and so on.



SNA zastosowania

- **Kampanie marketingowe**
- **Przewidywanie połączeń**
- **Upadłości firm**
- **Wzbogacanie bieżących danych**

Odkryj Santander Universidades

Globalny program, który wspiera rozwój uczelni i inicjatyw studenckich na całym świecie.

W Polsce program koordynuje Bank Zachodni WBK.

- Bogata oferta staży i praktyk
- Programy stypendialne
- Wymiana międzynarodowa
- Wsparcie projektów badawczych

santanderuniversidades.pl

Santander UNIVERSIDADES

Bank Zachodni WBK

Grupa Santander