

Socjolog wśród Data Scientistów

Opis plemienia

Remigiusz Żulicki







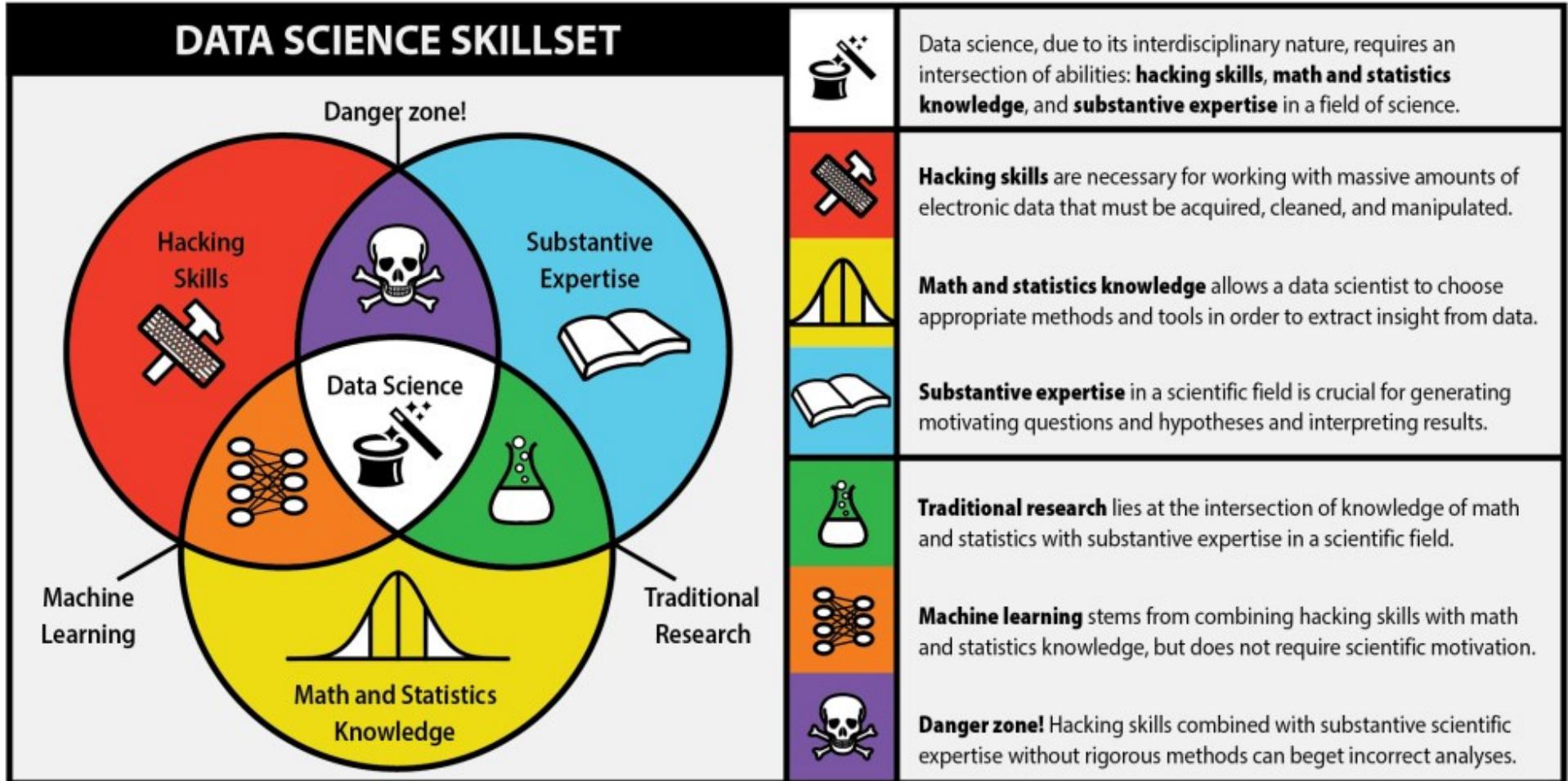
ja

data scientist

Plan:

- Data scientist, czyli kto? - charakterystyka
- Podziały wewnętrzne
- Jak to jest być data scientistem?

Data scientist, czyli kto? 1/3



„Tutaj są ci, co potrafią programować, tutaj są ci, którzy znają statystykę, tutaj są ci co dobrze mówią i tłumaczą, a tutaj na środku nie ma nikogo i tam są data scientisti. Musimy sobie uprościć i sobie podzielić na tych data engineer, data steward, jakieś takie stanowiska (...). Data scientist nie koniecznie musi umieć wszystko wszystko, tylko musi umieć tę esencję, rozwiązać problem biznesowy korzystając z analityki a teraz czy on to zaprogramuje, czy on to zaprezentuje, jakich narzędzi użyje, czy zrobi to superoptymalnie czy mniej no to jest sprawa drugorzędna”

Data scientist, czyli kto? 2.1/3

- Data scientist to osoba, która **rozwiązuje praktyczne problemy za pomocą ilościowej analizy danych cyfrowych:**
 - Przy użyciu programowania skryptowego
 - Przy użyciu szerokiej gamy metod analitycznych

**Me: This doesn't seem to be all that complicated. I
might even be able to solve it using logistic regression.**

Me to Me: Use deep learning!



Data scientist, czyli kto? 2.2/3

- Data scientist to osoba, która **rozwiązuje praktyczne problemy za pomocą ilościowej analizy danych cyfrowych:**
 - Przy użyciu samodzielnie pisanych skryptów
 - Przy użyciu szerokiej gamy metod analitycznych
 - **Przy konieczności dookreślenia problemu, sposobu jego rozwiązania i miar powodzenia**
 - **Przy braku wiedzy eksperckiej co do merytoryki problemu**

„My mamy postawione zadanie, na przykład zwiększyć sprzedaż, albo zmniejszyć ilość wypadków (...). I my **nie wiemy jak to zrobić**. Nasza robota polega na tym, że tworzymy takie pole, taką przestrzeń, w której modelujemy różne zjawiska (...). Nasza praca polega na pomaganiu modelowi żeby dobrze działał. Aczkolwiek ta różnica między nami a normalnymi programistami czy w ogóle normalnymi inżynierami jest taka, że my w zasadzie nie wiemy jak rozwiązać problem”

„Programiści bardzo by chcieli i przyzwyczajeni są do takiego trybu, że dostają gotowe wymagania, od klienta, że ta aplikacja to ma robić to to to to to, i tamto. (...) Natomiast w naszych projektach problem jest często bardziej abstrakcyjny, i z góry nie ma gotowej definicji tego problemu. I często widzę że też programiści borykają się z tym”

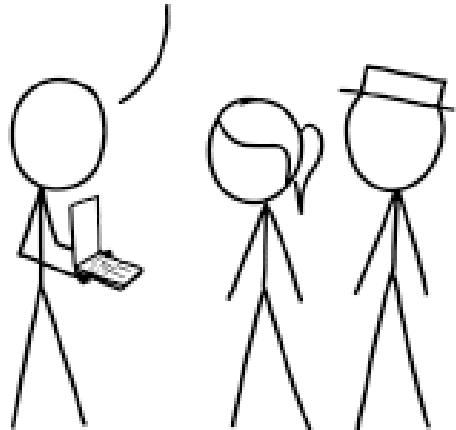
„Trzeba być człowiekiem w data science, robienie modeli jest najprostsze, a najważniejsze jest komunikować się i być kreatywnym i wytłumaczyć to biznesowi. Nikt nie potrzebuje perfekcyjnego modelarza, a potrzebuje człowieka, którego kreatywności i dociekliwości nie da się zautomatyzować.”

„Jestem ciekawską osobą bardzo z natury, więc dobrze się czuję w sprawach, które nie są już wyjaśnione. Czyli bardzo lubię, jestem rybą mętnej wody. Się śmieją tutaj koledzy ze mnie w pracy że ja najczęściej biorę projekty które albo się nie udało, przynajmniej z dwa razy, albo nikt nie wie jak się za nie zabrać, albo w ogóle problem jest tak rozmyty, że pfffff (...) Więc akurat te rzeczy które są u nas w zawodzie dla wielu osób wyzwaniem, to mnie cieszą, bo mam taki charakter”

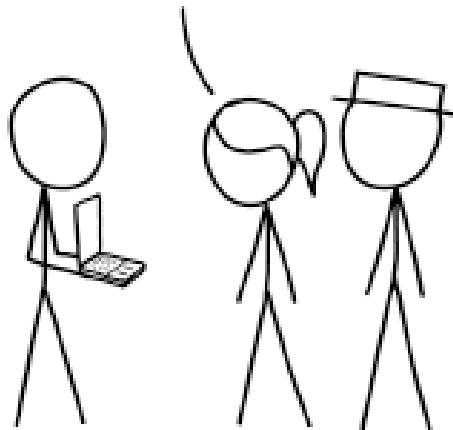
Data scientist, czyli kto? 3/3

- **Magik**
 - Robi rzeczy niemożliwe
 - Robi rzeczy niezrozumiałe
 - Robi rzeczy pożąданie
- **Majsterkowicz**
 - Nie wie, jak rozwiązać problem
 - Nie ma jasnych wytycznych
 - Szuka, kombinuje, eksperymentuje, skleja, docina

CHECK IT OUT—I MADE A
FULLY AUTOMATED DATA
PIPELINE THAT COLLECTS
AND PROCESSES ALL THE
INFORMATION WE NEED.

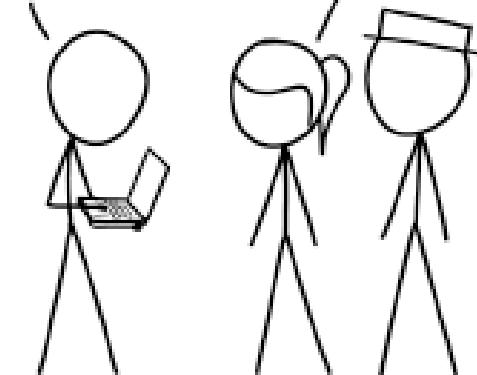


IS IT A GIANT HOUSE OF CARDS
BUILT FROM RANDOM SCRIPTS
THAT WILL ALL COMPLETELY
COLLAPSE THE MOMENT ANY
INPUT DOES ANYTHING WEIRD?



IT... MIGHT NOT BE.

| I GUESS THAT'S SOMETH-
WHOOPS, JUST
COLLAPSED. HANG
ON, I CAN PATCH IT.



Podziały wewnętrzne

- **Rodzaj firmy:** klient swój vs klient obcy
- **Background:** matematycy vs informatycy vs inni
- **Specjalizacje:** data engineer, ML engineer, ML researcher, analytics translator...

Podziały wewnętrzne

- Rodzaj firmy
- Backgroud
- Specjalizacje
- **Narzędzia: Python vs R**
 - R: dplyr vs data.table
 - Python: TensorFlow vs Theano vs Torch



HATING EXCEL

R USERS

PYTHON USERS

„Nie widzę sensu w ogóle dzielenia tego, natomiast oczywiście wszyscy w tym zawodzie, trzeba być, są ambitni, każdy uważa się za specjalistę w czymś bardzo często jest, więc mamy te swoje małe wewnętrzne zabawy w kłócenie się o wyższości Pythona nad R lub przeciwnie, co moim zdaniem jest fajne bo też powoduje że obie strony się uczą tak? (...) Więc summa summarum no czymś musimy się zajmować oprócz tej analizy danych. O czymś musimy się przy piwie tam dyskutować”

PYT: To jest wojna między Pythonem a R?

ODP: Wiesz co, niby trochę jest, ale wydaje mi się że to bardziej dla beki

„Data scientist, to nie jest to osoba która używa eRa, czy Pythona, tylko ona zadaje właściwe pytania, tak? Ogólnie, dla mnie, nie ma czegoś takiego jak wojna, jeśli ktoś się bawi w wojnę między językami, no to dla mnie jest takie trochę, taka trochę dzieciada, wiesz? Ze względu, jeśli ktoś by chciał robić model w VBA, no to niech robi, tylko pod warunkiem że on wie, że to będzie bardziej opłacalne jakby to zrobił w Pythonie. Więc ogólnie wiem że jest taka wojna, są osoby które się jakoś oburzają i jakoś starają się nie wiem, wmaćwać że to co one potrafią jest lepsze od tego co inni potrafią. (...) Oba narzędzia trzeba byłoby poznać, żeby znać ich plusy i minusy”

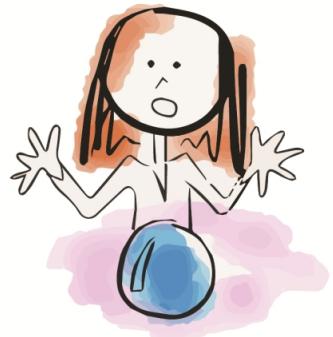
Jak to jest być data scientistem?

Jak to jest być data scientistem?

- Dobrze, ale...

DATA SCIENTIST

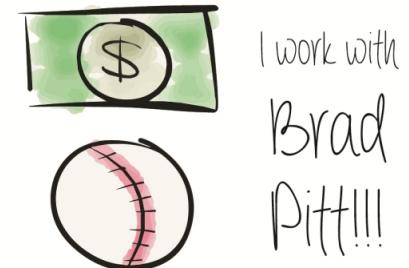
What my **CUSTOMERS** think I do



What my **MUM** thinks I do



What my **FRIENDS** think I do



What my **HUSBAND** thinks I do



What **I** think I do



What **I** **ACTUALLY** do



„Najbardziej podoba mi się to, że jest różnorodna. Że wymaga myślenia. Że projekt za projektem to nie są powtarzalne rzeczy. Jakieś tam części się powtarzają jasne, ale jakby za każdym razem jest to trochę inny problem do którego trzeba wymyślić jakieś podejście. (...) Właściwie cały czas śledząc rozwój tej dziedziny, czytając różne artykuły patrząc jakie się pakiety pojawiają chociażby tam właśnie w R, można się dowiadywać zupełnie nowych rzeczy i robić w zupełnie inny sposób to co się jeszcze robiło jaką konkretną metodą, nie wiem pół roku temu chociażby. Więc to jest tak że można dużo, trzeba dużo myśleć, dużo szukać różnych źródeł i być cały czas na bieżąco”

„Wcale nie jest tak że data scientist to spiła śmietankę i cały czas sobie buduje te modele i w ogóle to ma tak świetnie. (...) To co ja widziałem w Polsce to jednak jest ścieranie się ze zmieniającymi się wymaganiami biznesu, walka z tym że dane w firmie są w zlej kondycji i tak dalej i tak dalej. (...) Projekty są żmudne, tak jak w tym związanym z wykrywaniem wyłudzeń, po prostu musieliszy siedzieć i sprawdzać cecha po cesze. To jest żadna przyjemność i spora część projektów taka jest. Jak projekt jest źle poprowadzony to wtedy te rzeczy przygotowawcze zżerają prawie cały czas i to co jest najfajniejsze o czym się mówi: modelowanie, właśnie ta zabawa, z dużymi komputerami, zabawa z GPU, wiele procesorów, to na to zostaje 10% czasu”

„Myślę że przede wszystkim nie jest dla ludzi którzy, nie lubią siedzieć dłucho nad jedną rzeczą, nie lubią dłucho siedzieć sami, z komputerem, no bo jednak to bardzo wymaga żebyś nie potrzebował więcej bodźców, prawda. Więc myślę że wszystkie osoby które lubią pracować z klientami no to są raczej spisane na straty tutaj, bo po prostu się wymęczą w tym. No tak jak na studiach ten kto rozumiał statystkę i lubił matematykę to tu się nada, a ten kto nie rozumiał i nie lubił myślę że tu się, też nie nada”

„Ona brzmi super jak się opowiada znajomym, ale jak już siadasz do tego i to robisz, to nie jest seksowne, to statystyka. Więc to jest fajne ale dla kogoś kto właśnie to czuje, nie?”

Dziękuję!

remigiuszzulicki@gmail.com

Remigiusz Żulicki

