

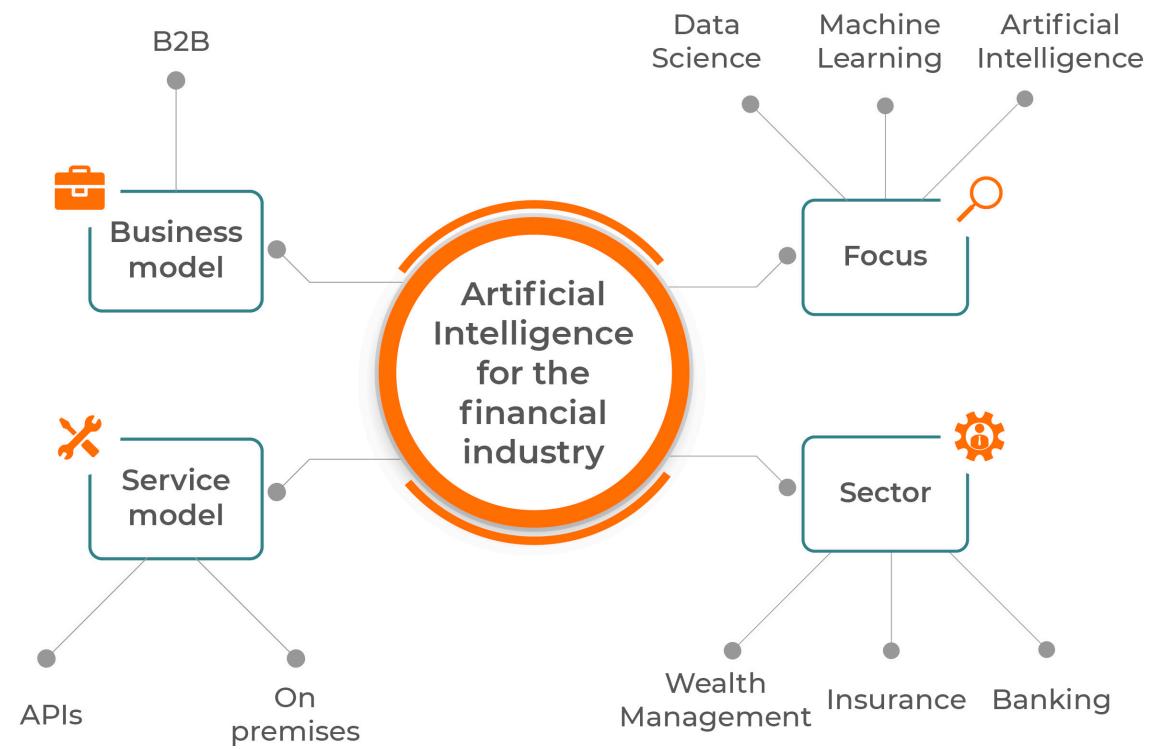
Module #1

Data-driven customer profiling for financial institutions/Fintechs

Raffaele Zenti (raffaele.zenti@virtualb.it)

Co-Founder, Head of Financial Data Science, Virtual B SpA

Why here?
Because of
Virtual B SpA



Our clients



ING DIRECT[®]



CheBanca!
Gruppo Mediobanca



Allianz Bank
Financial Advisors

BLACKROCK[®]



Schroders



Etica SGR
GRUPPO BANCA POPOLARE ETICA



BancoPostaFondi SGR



ZENIT Online

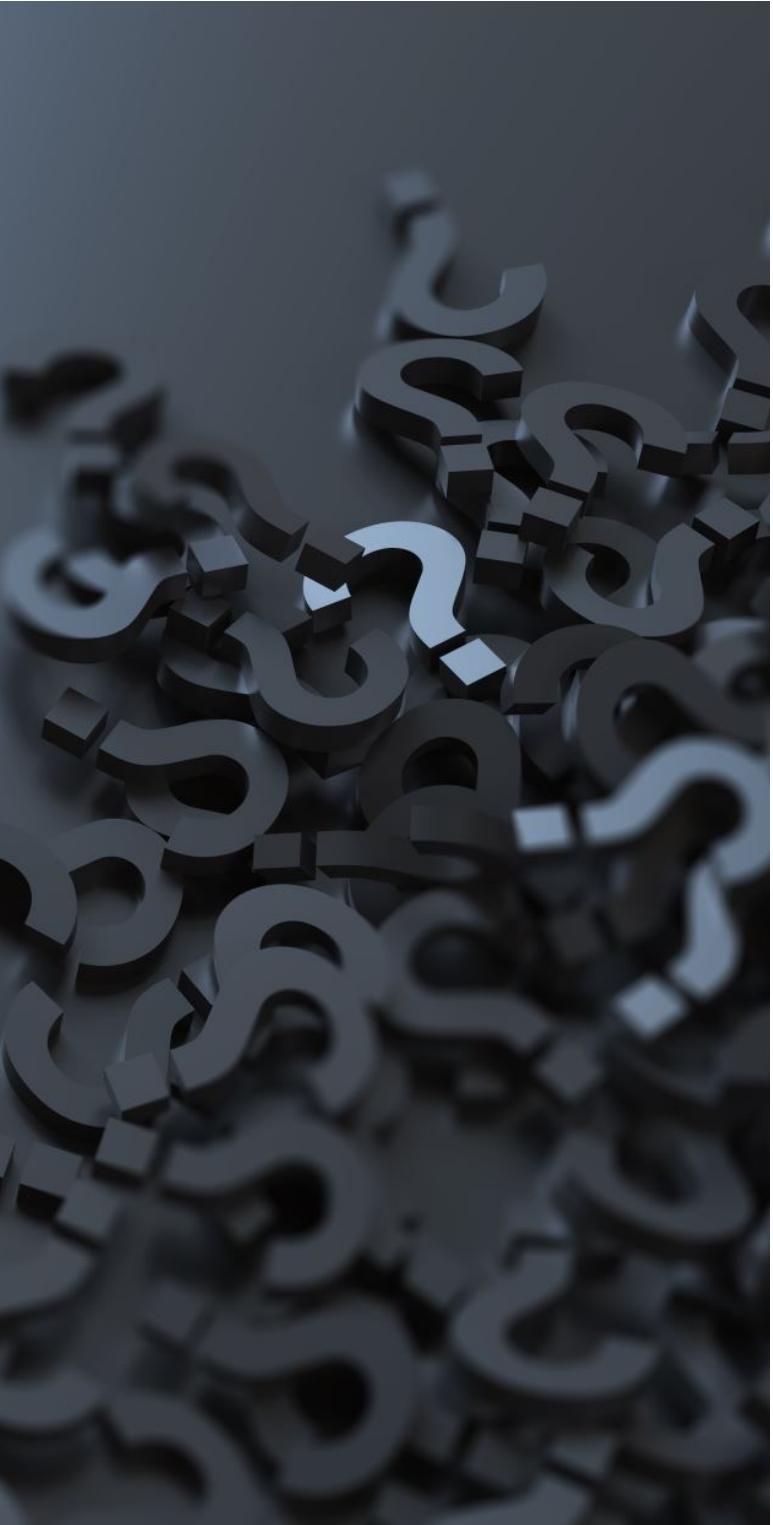


N26

Overall goals

- You will savor the practice of Machine Learning in the Fintech sector with a series of business cases
- We will cover various aspects of the business (e.g., customer intelligence, digital marketing, investing, risk analysis) and various Data Science techniques
- I would like to stimulate your ability to solve real problems with Data Science - having modeling vision – not just writing a piece of code, or using that new fancy model...





Organizational aspects

- Computing language: Matlab
- We will use a notebook – the Live Editor (because it's easier to show the analysis, but notebooks are great just for prototyping and this kind of situations - to get an idea of how to code neatly, take a look at:
<https://drivendata.github.io/cookiecutter-data-science/>)
- Each business case is as follow:
 - I present the case from a "business" point of view
 - I present an associated dataset
 - We will do some brainstorming: I want to see you come up with ideas on how to solve the problem
 - I show you some code that solves the problem (there is rarely only one way)
 - I leave you to develop your own code, playing at home
 - Next time we discuss how you solved it, issues, etc



Business case #1:

Segmenting Clients

Why segment? To efficiently customize the service



A black and white photograph of three men from the early 20th century. They are all wearing dark suits, white shirts, and dark top hats. The man on the left has a mustache and is looking towards the camera. The man in the center is looking directly at the camera with a neutral expression. The man on the right is partially visible, also looking towards the camera.

Client segmentation: the old school in finance

- Money, e.g
 - AuM
 - insurance premiums
 - number of trades p.a.
 - ...
- Age
- A combination of 1. and 2.
- (Yeah. Boring. There is room for some improvement - unless you think that money and/or age fully define a human being)

Personas: general idea

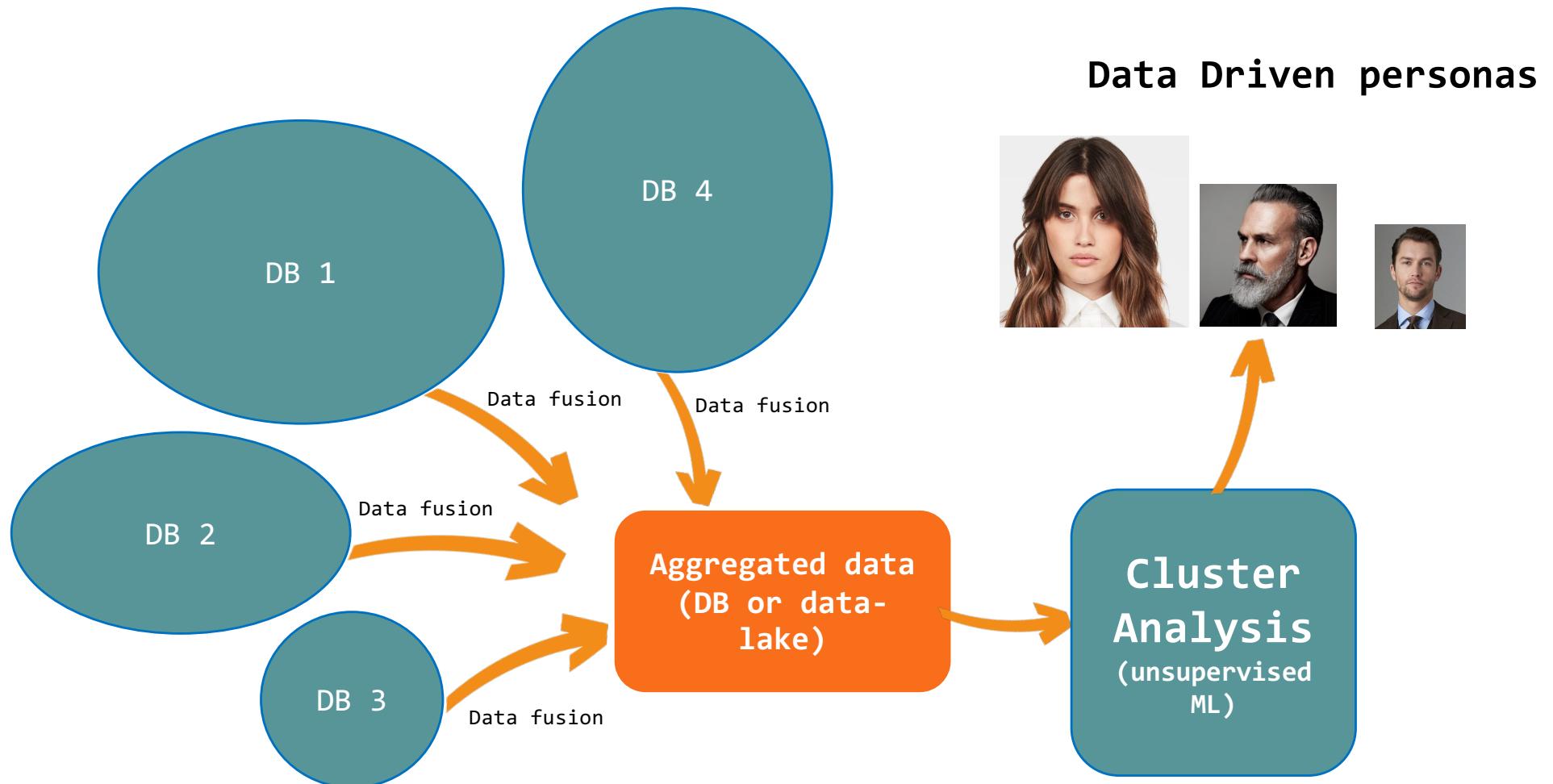
- Premise: we are all different, but with common traits
- Personas = **human prototype with similar traits**
- Depending on the perspective, human prototypes change – we take the financial perspective
- Key to summarize information ↔ **dimensionality reduction**
- That is, instead of reasoning on, say, 400k customers, you can reason on 6 TYPES of customers – which is easier



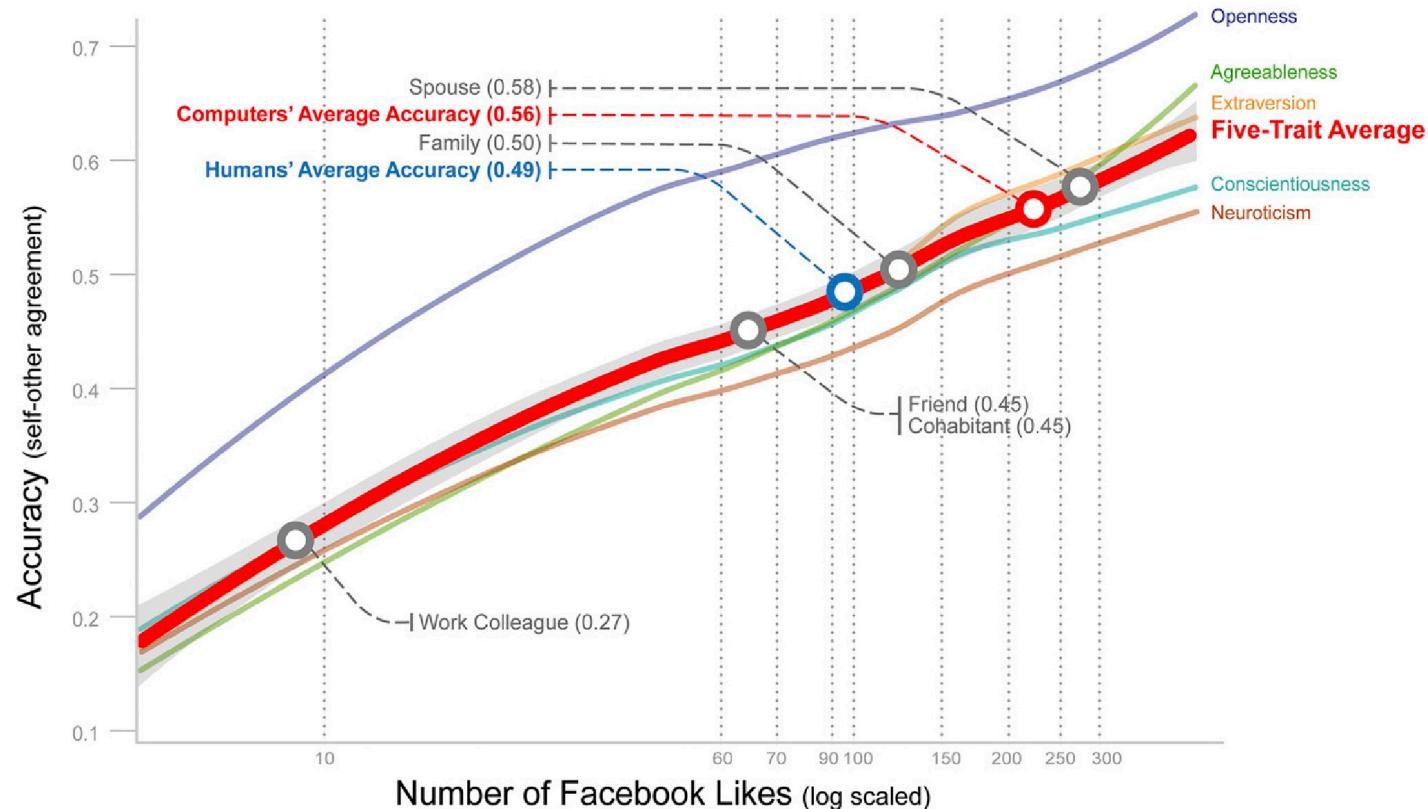
Personas according to traditional marketing

- Put some experts in a room
- Super-brainstorming
- Based on their experience, they will identify «customer prototypes», i.e., personas – they will likely include their biased personal ideas
- (Better. But there is still room for some improvement)

Data-driven personas: (simplified) overview



Motivation: AI/ML and people



A Machine Learning algorithm needs 70, 150 and 300 Facebook Likes, respectively, to outperform an average friend, family member, and spouse in assessing your personality*

*Five-Factor Model model

Source: Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. PNAS Proceedings of the National Academy of Sciences of the United States of America, 112(4), 1036-1040

The problem

- **Goal: segment clients based on the information contained in their corresponding vectors of features**
- Spoiler: in real world client data contain heterogeneous data (categorical nominal/ordinal, numerical ordinal/continuous)
- Introductory courses to unsupervised learning quite often discuss ideal use cases, such as tutorials using k-means, which works great, but only apply to numerical features – so we are out of the comfort zone...
- Now: let's inspect data, and discuss!



Let's look at the data



Brainstorming

(don't be shy)

- What would you do?
- What are the key steps to start from the data and arrive at customer segmentation?
- What kind of algorithms would you use? Why?
- ...



Matlab session
starts

At the end of the process...

- Qualitative overlay might change the number of clusters, or might change centroids/medoids

- Each client is a vector:

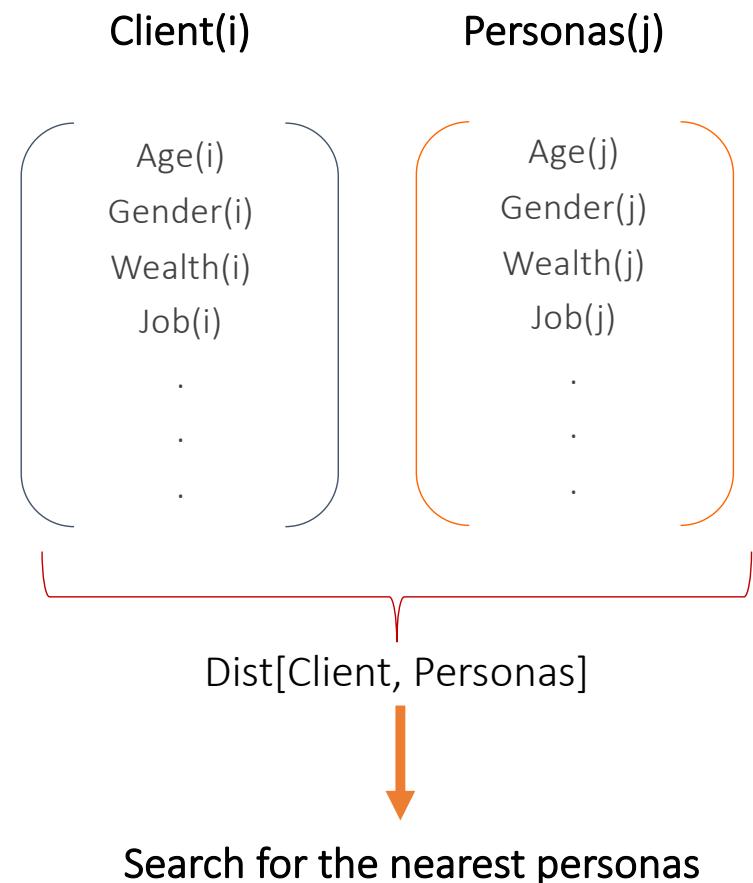
$z(i), i = 1, 2, 3, \dots n_{Client}$

- Each personas (centroid/medoid) is a vector:

$C(j), j = 1, 2, 3, \dots n_{Personas}$

- Client(i) belongs to the closest personas(j), i.e. the rule is:

$\text{ArgMin}_j \{ dist[z(i), C(j)] \}$



Real example – inside clusters (it's not your dataset)

Features	Cluster (2) = 21% “Wealthy widow”
Age	55-70
Gender	F
Job	Housewife, retired
Marital status	Widow, separated, divorced
Family	1
Financial education	Below average
Geographical area	Italy
Size of the municipality	20k÷50k
Income	Above average
Mortgage	N
Short term loans	N
Real estate wealth	Huge
Financial assets	Above average
Socio-demographic risk	Average
Geo-seismic risk	Above average
Digital propensity	Low



Service model:

- Physical
- Physical+Call center



Main needs:

- Long Term Care
- Inheritance
- Investments (low risk, capital protection)
- Premium credit cards

Real example – inside clusters

Features	Cluster (3) = 2% “Top jobs”
Age	50-65
Gender	M
Job	Manger, professional, business owner
Marital status	Married
Family	2-3
Financial education	High
Geographical area	North Italy, large cities
Size of the municipality	Above 200k
Income	High
Mortgage	Y
Short term loans	N
Real estate wealth	Above average
Financial assets	High
Socio-demographic risk	Above average
Geo-seismic risk	Low
Digital propensity	High



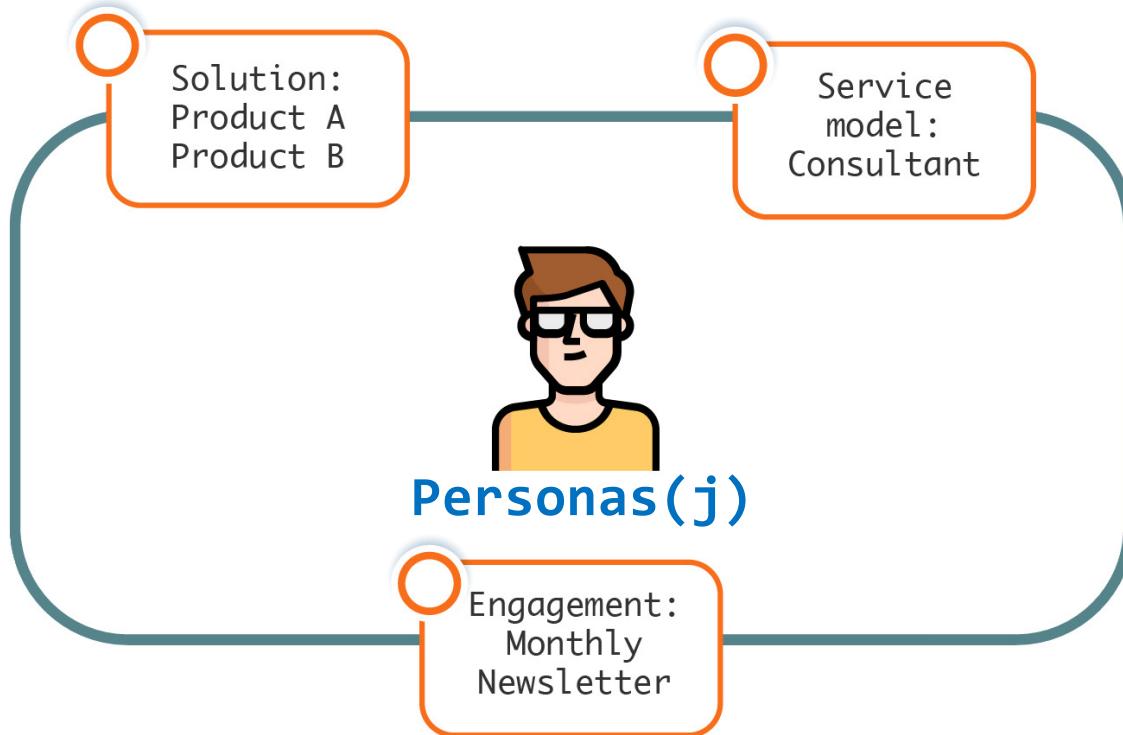
Service model:

- Digital
- Physical+call center

Main needs:

- Long Term Care
- Family/Home protection
- Death insurance
- Investments
- Premium credit cards

Using personas to customize financial services in an industrial way



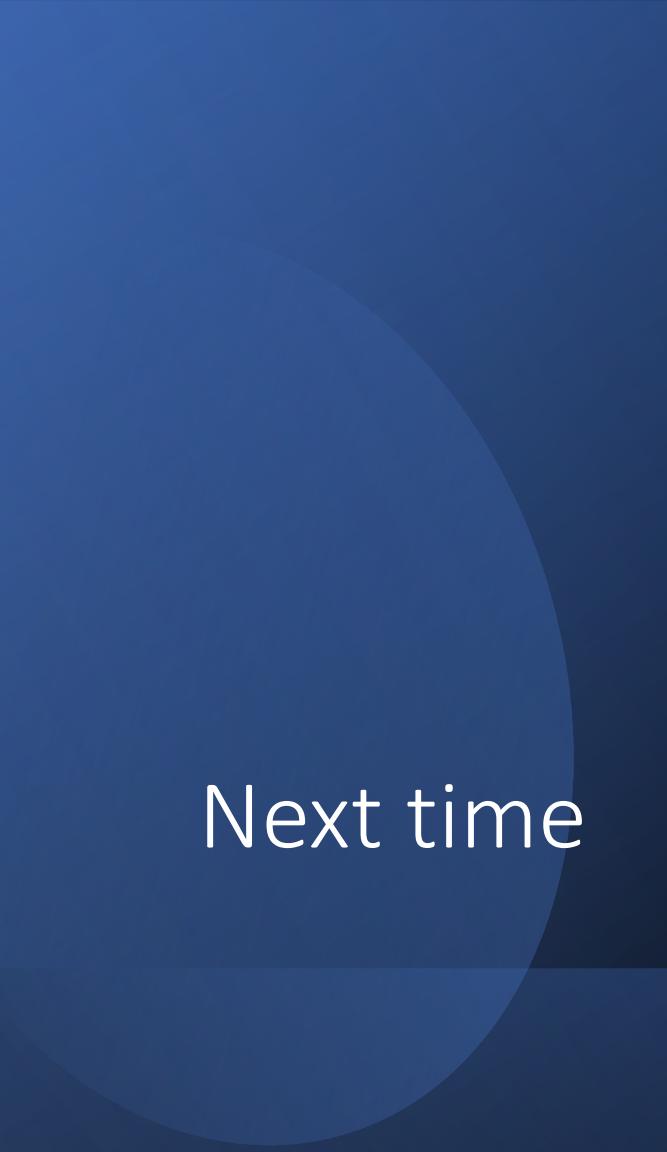
- Each personas has her service/communication model, products, etc
- But personas are typically 8÷15 → you reduce business complexity

Take home on data-driven personas

- Application of unsupervised ML + qualitative overlay (strongly suggested)
- Much better segmentation, based on empirical evidence
- Several business applications:
 - customized products, services, channels
 - dedicated communication tools for each personas
 - precision marketing online/offline
 - data enrichment if we have a limited amount of information (a handful of features) → then you can start engaging customers
 - generating synthetic data, for ML training and simulations
 - ...

Now **YOU**

- **It's your turn: use your favorite techniques to segment customers, write the code (use my code, or start from scratch, or whatever...as you like), and we'll talk about it next time**
- **Work in small groups: use collective intelligence (Data Science = teamwork)**
- **Get your hands on that data...**



Next time

- Each group will present ideas, results, doubts, code snippet, etc
- Be short and concise (you are... many)
- Prepare plots, charts, tables, commented code snippets
- Get ready to share your screen and your work - be it little or a lot – don't be shy