~~Вся правда~~ О специалистах по ~~данным~~ data science

Open Data Science

DM Labs

Алексей Натекин

Alex Natekin

SIEMENS fortiss DM Labs

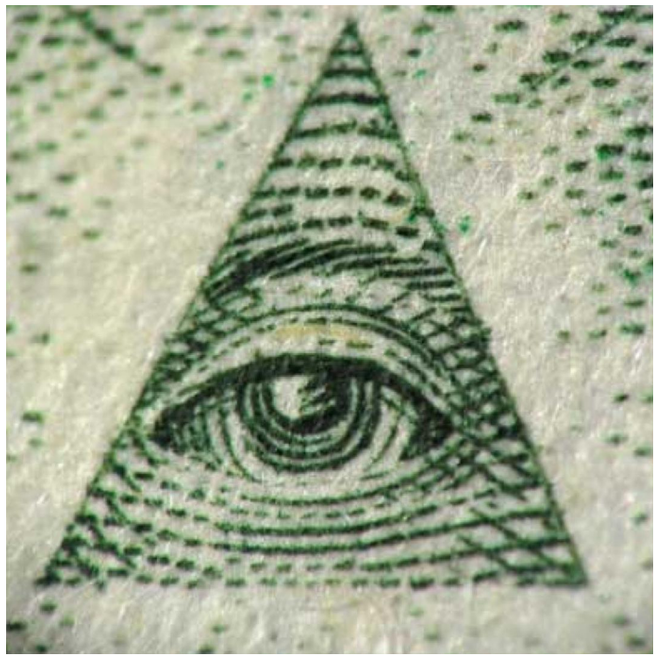
Deloitte. DIGINETICA ais group XOR

TECHNISCHE UNIVERSITÄT MÜNCHEN

Open Data Science

2008 2009 2010 2011 2012 2013 2014 2015 2016 2017

Finding ~~nemo~~ data scientists:



Who are all
these people

How and where
to find them

Brief manual:
dos and donts

Data [driven] Science:

- “The art of turning data into actions”
- “Interdisciplinary field about scientific methods, processes and systems to extract knowledge or insights from data in various forms”
- “A newly emerging field dedicated to analyzing and manipulating data to derive insights and build data products”

Data Science

Here be data visualization, decision theory, domain sciences, optimization, etc.

Statistics
Data Mining

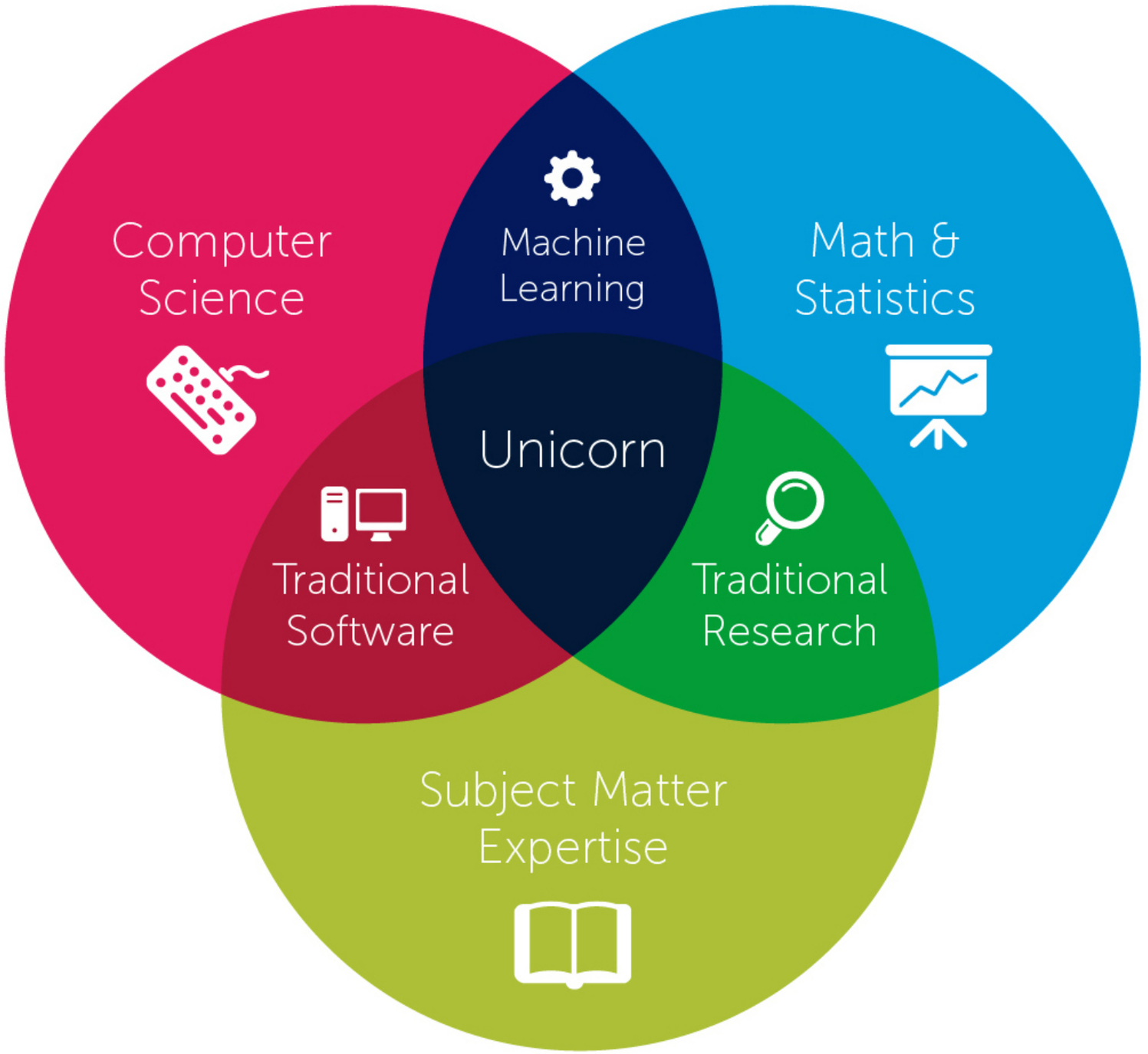
Machine Learning

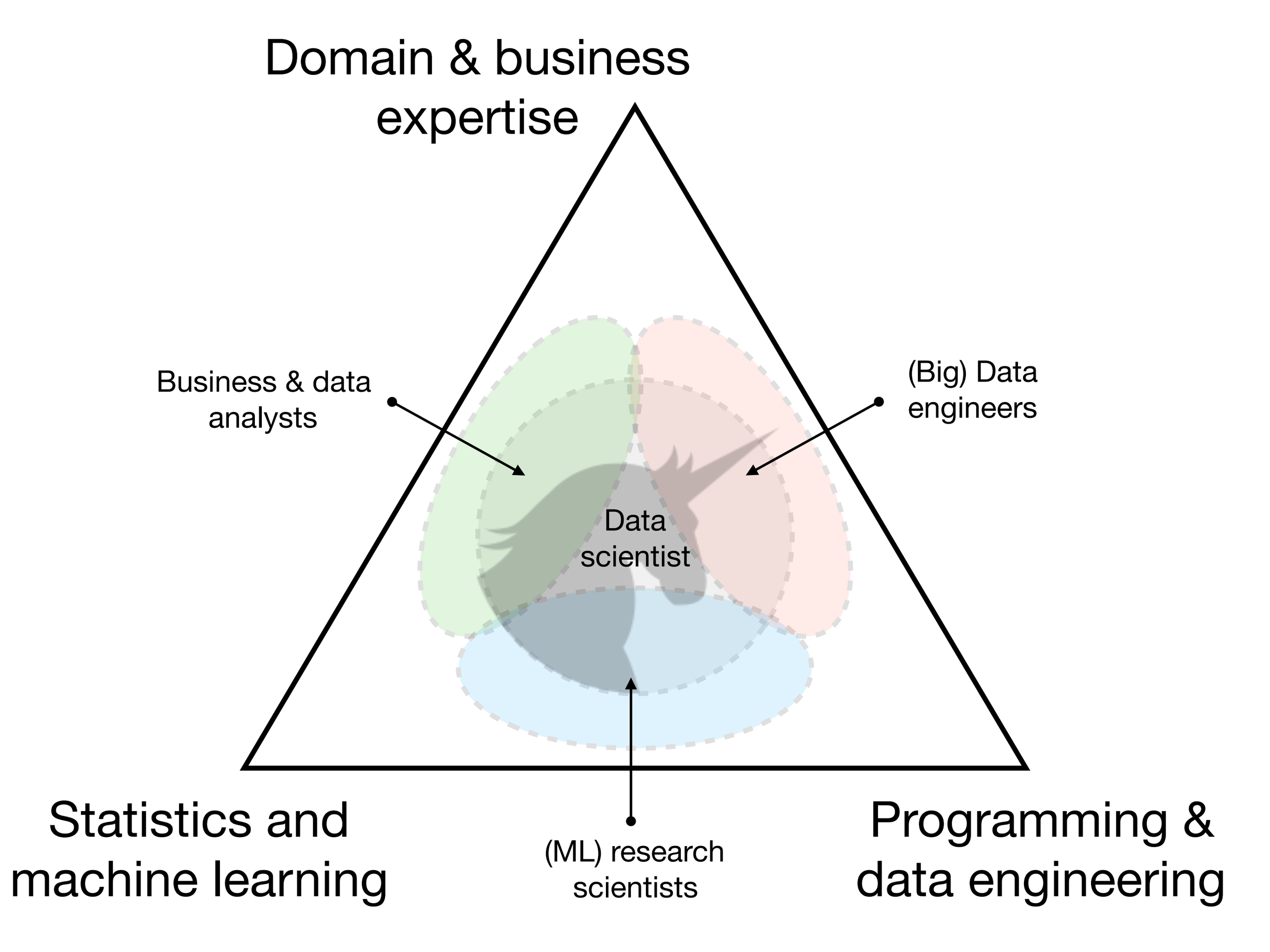
Big Data
HPC

AI

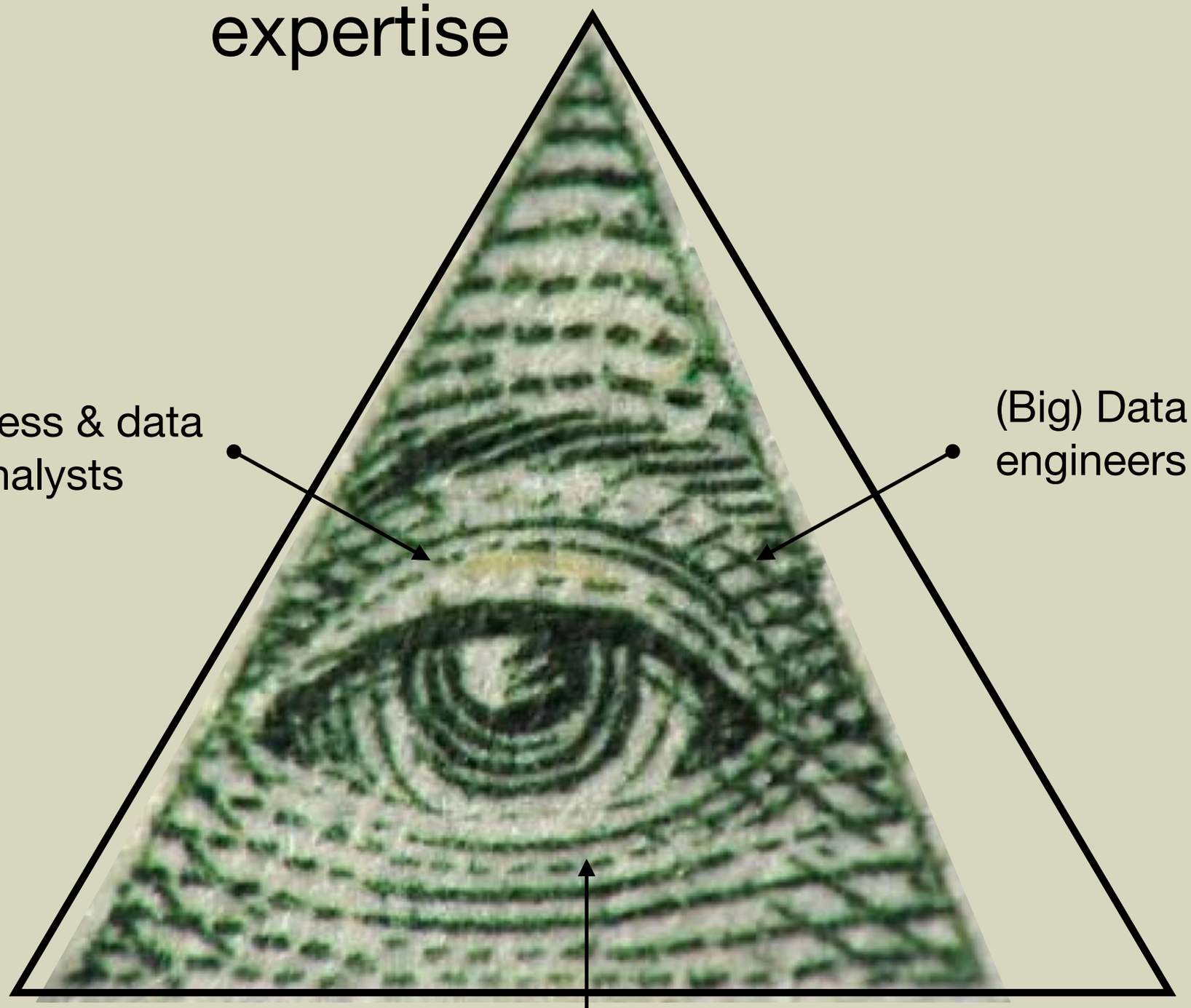
Here be dragons, AI safety, ethics, ASI philosophy, avatars, etc.

“A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.”




Domain & business expertise
Business & data analysts
(Big) Data engineers
Data scientist
(ML) research scientists
Statistics and machine learning
Programming & data engineering

Domain & business expertise



Business & data analysts

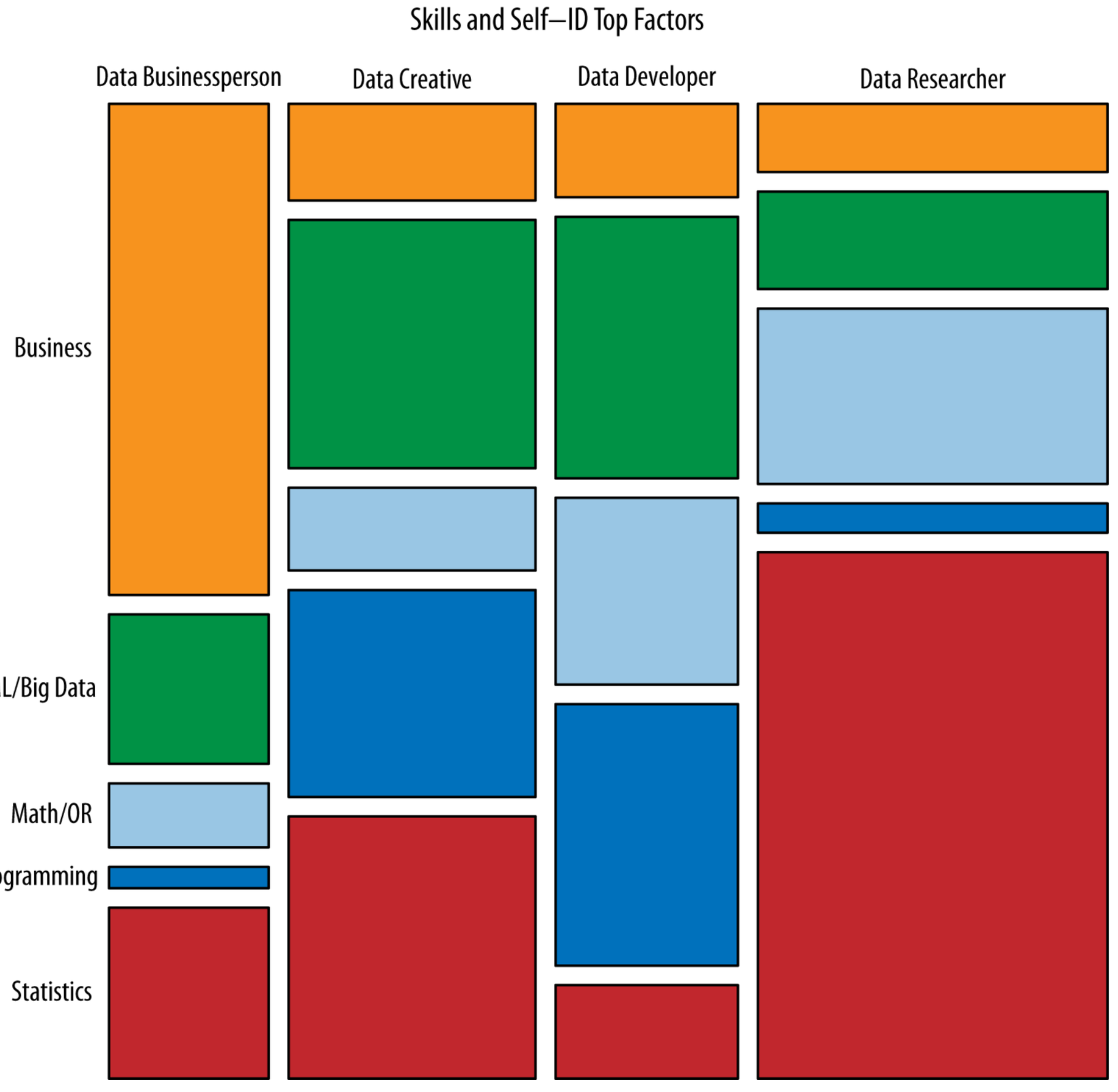
(Big) Data engineers

Statistics and machine learning

(ML) research scientists

Programming & data engineering

Analyzing the Analyzers



Data scientists maturity:

Senior

You give them a (business) problem, they (consistently) bring you a (working) solution

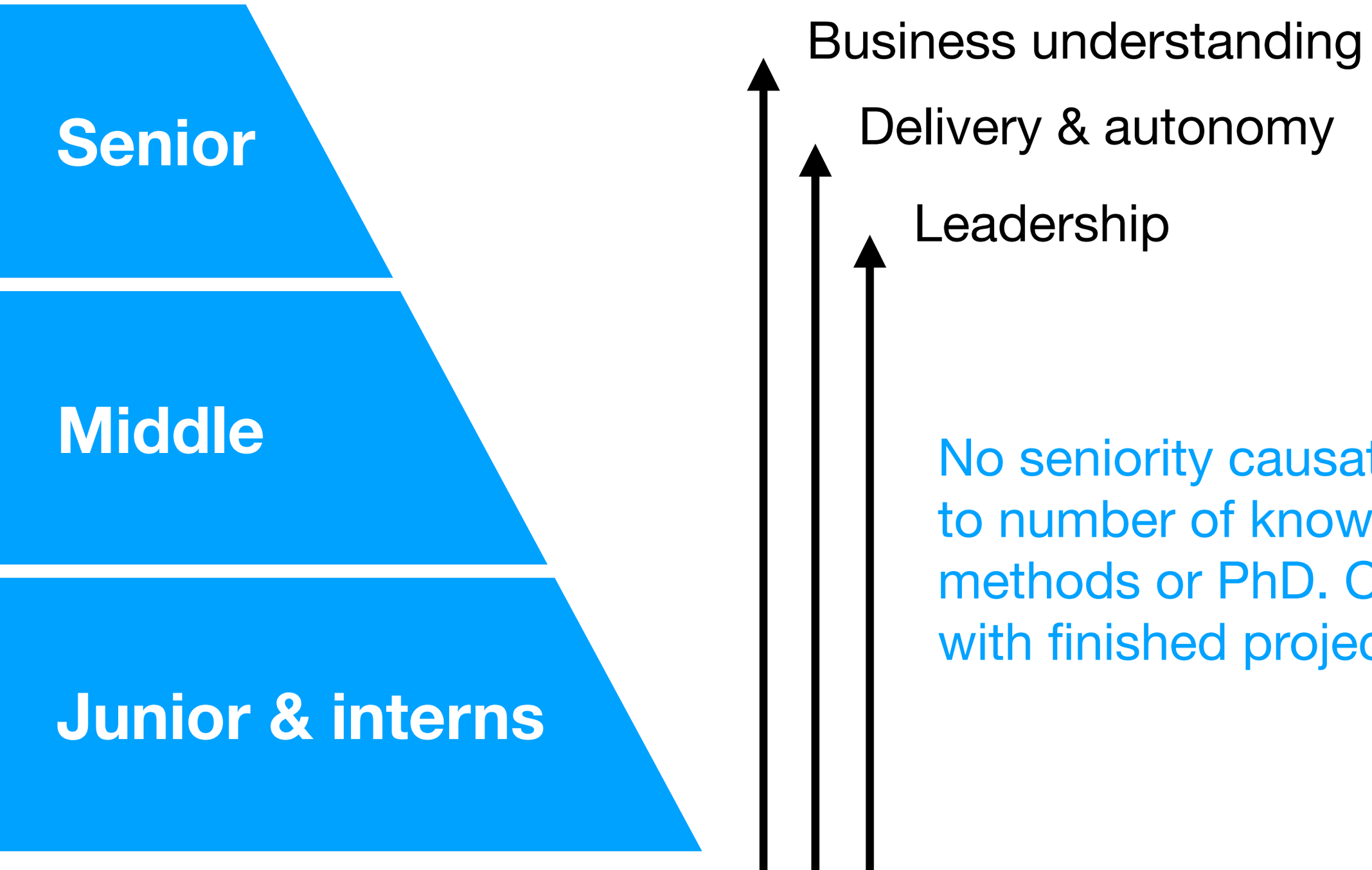
Middle

Given a prepared problem and paired with senior, you (weekly+) check on their progress

Junior & interns

Paired with mid and senior, they solve technical tasks and are checked daily

Data scientists progression:



Interview tips:

	can be hacked	hard to hack
Senior	~~Statistics, ML, tech stack,~~ processes, development skills **AND\OR** management, business understanding	Talk about finished projects: -business -pet projects -open source -hackathons -competitions -study All the projects, all the details, progression, why use A and not B…
Middle	Statistics, ML, tech stack, processes, development skills	
Junior & interns	Statistics, ML, development skills	

(1)

Senior	Large demand, very low supply	Found via recruiters Rare and few, easier to validate
Middle	Some demand, reasonable supply	Found via job boards ~~& recruiters~~. Can be a too early junior promo
Junior & interns	Many wannabe in supply	Found via job boards and academy. Many impostors, hard to pick

FANTASTIC BEASTS AND WHERE TO FIND THEM

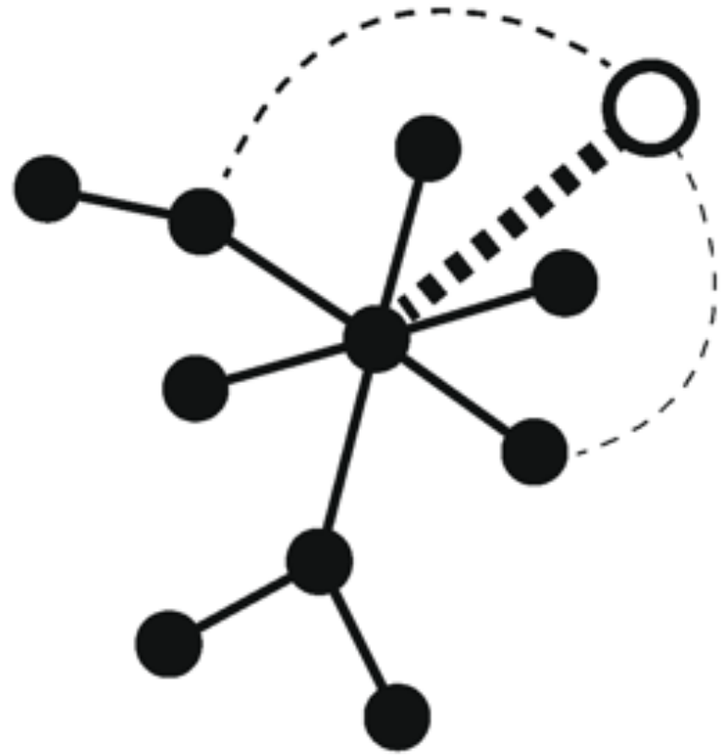
(2)

Senior

Middle

Junior & interns

Community and personal ties work best



Meetup

kaggle

Data scientists specifics:

Notes about DS feasibility:

- Data science is not a silver bullet
  Also, not all cases are reproducible and feasible
  Don’t overhype expectations

- Fail-safe ideology
  Avoid a costly failure from the larger team
  “If you fail, fail convincingly"

- Follow the “just do it” approach
  Prototype and Succeed or Fail Fast
  Help achieving more “80% solutions”

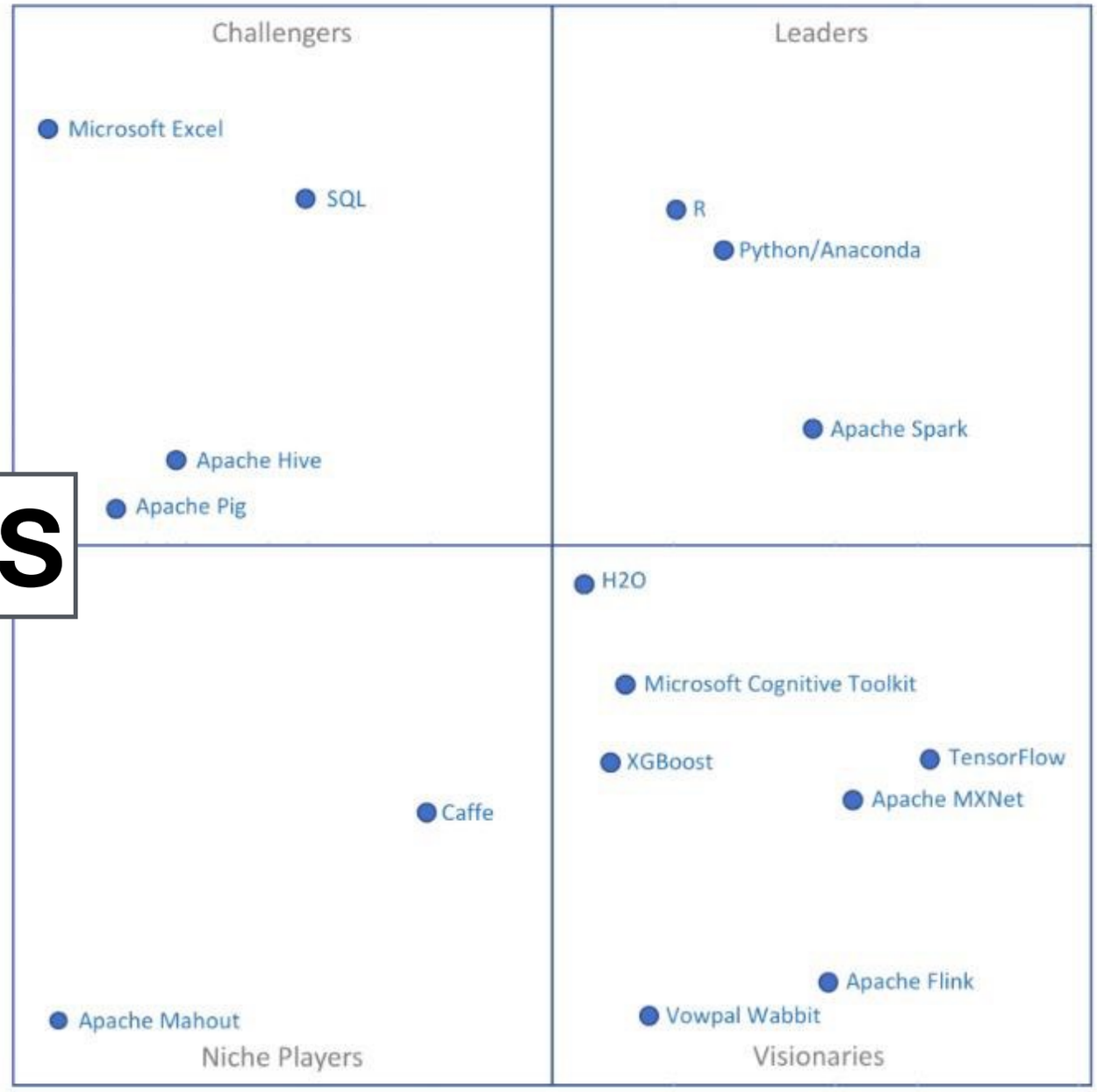
Notes about vacancies:

- Be precise with whom you are looking for
- If you want DS not to lie on an interview, don’t lie on the job description. Give true salary estimates*
- Painful truth about actual tasks (ETL, no ML) is better than unmatched promises

*use java developers with same amount of experience and “lead” functions for proxy

Software and open source:

VS



Common misconceptions:

- **Lack of vision in DS tasks**
  - Expect DS team to operate as engineering team
  - Confuse DS with pure data engineering role
  - Fall for the seduction of tons of data
- **Forcing typical business patterns**
  - Expect them to already know all the technologies, algorithms and approaches
  - Overgrow teams: the bigger the slower
  - Tell them what to do (like force them into agile*)

“We expect deep analytical talent shortage of between 140,000 and 190,000 by 2018 in the US only”

&

“We project a need for 1.5 million additional managers and analysts in the US who can ask the right questions and consume the results of the analysis of big data effectively.”

Summary:

- To prevent role mismatch, role of a data scientist requires more specification: data engineering, ml researcher, data analyst, domain expert, etc
- It’s hard to find seniors, but juniors are also quite tricky. Use community and references for both
- Offer you data scientists a fail-safe and experiment-friendly environment
- Help DS with the right questions from business

Thanks!

@natekin
alex.natekin@dmlabs.org