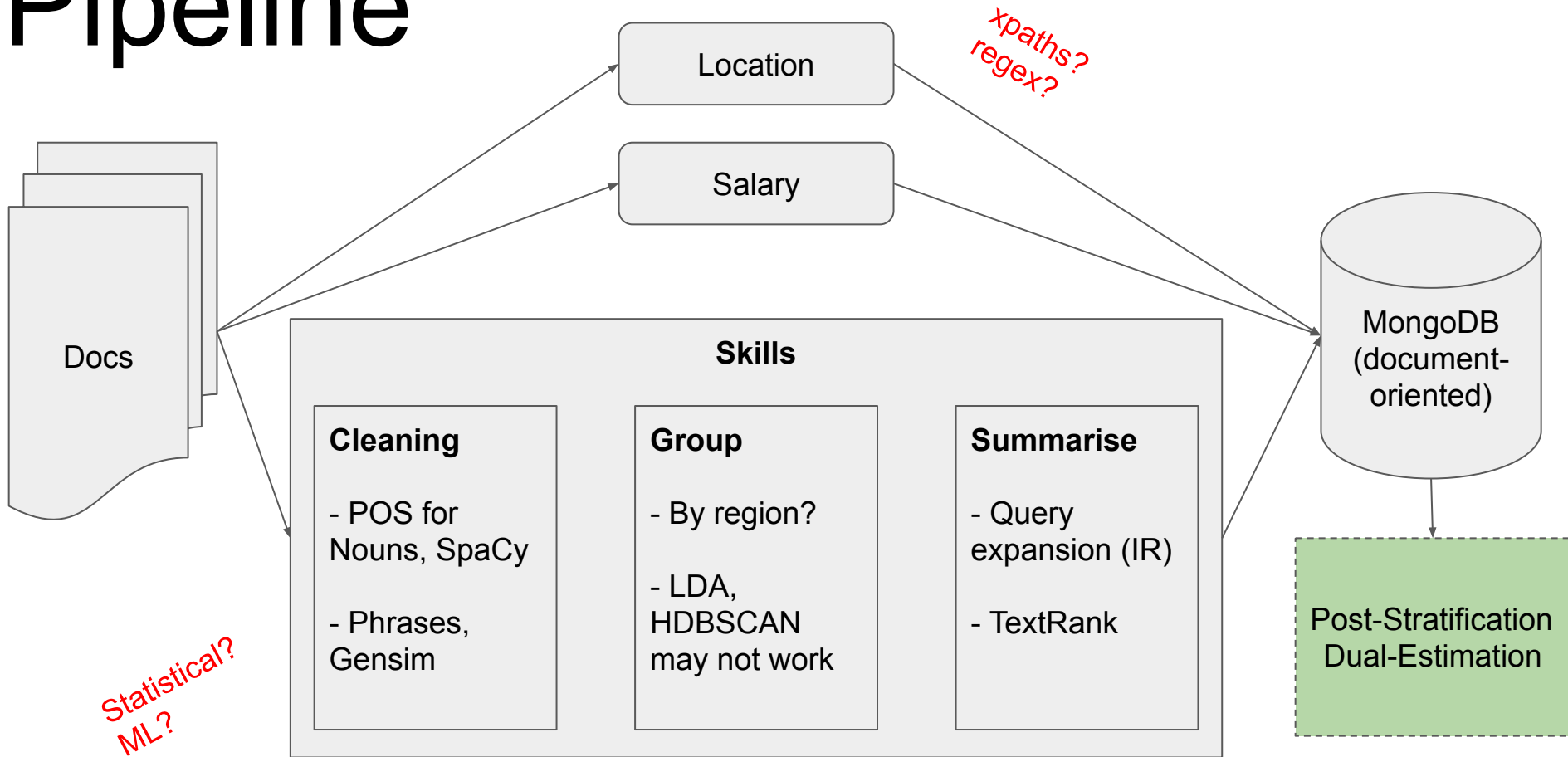# The problem

- Don't know what we're looking for = Text Mining
- Also info on Location & Salary
- Webscraping! Demo data - CV's + Job descriptions

- Outcome = top skills, faceted by region & salary
- Info on career stage?

Pipeline

Location

Salary

xpaths?
regex?

Docs

MongoDB
(document-
oriented)

Skills

Cleaning

- POS for
Nouns, SpaCy

- Phrases,
Gensim

Group

- By region?

- LDA,
HDBSCAN
may not work

Summarise

- Query
expansion (IR)

- TextRank

Post-Stratification
Dual-Estimation

Statistical?
ML?

https://github.com/OzwaldCavendish/mining_cv_data

Plan & Deliver

- Before beginning - Read! Consult!
- Build "online", AGILE can be done even without formal ceremonies
- If possible/secure, introduce customer to GitHub, show them progress & outputs folder, get a graph early
- Handle expectations/uncertainty
- Consider scale right at the beginning, build appropriately

https://github.com/OzwaldCavendish/mining_cv_data