Data Science at MDCS

IGOR ILIC

Microsoft Development Center Serbia

Established 2005

220 employees:

- Software Engineers
- Data Scientists
- Labelers
- Designers
- PR, HR, Finance

Teams:

- SQL
- Office
- Bing
- Handwriting recognition
- etc.

Known for hiring local “super-stars”, but focused on discovering new talents

Start-up within a corporation

Data Science

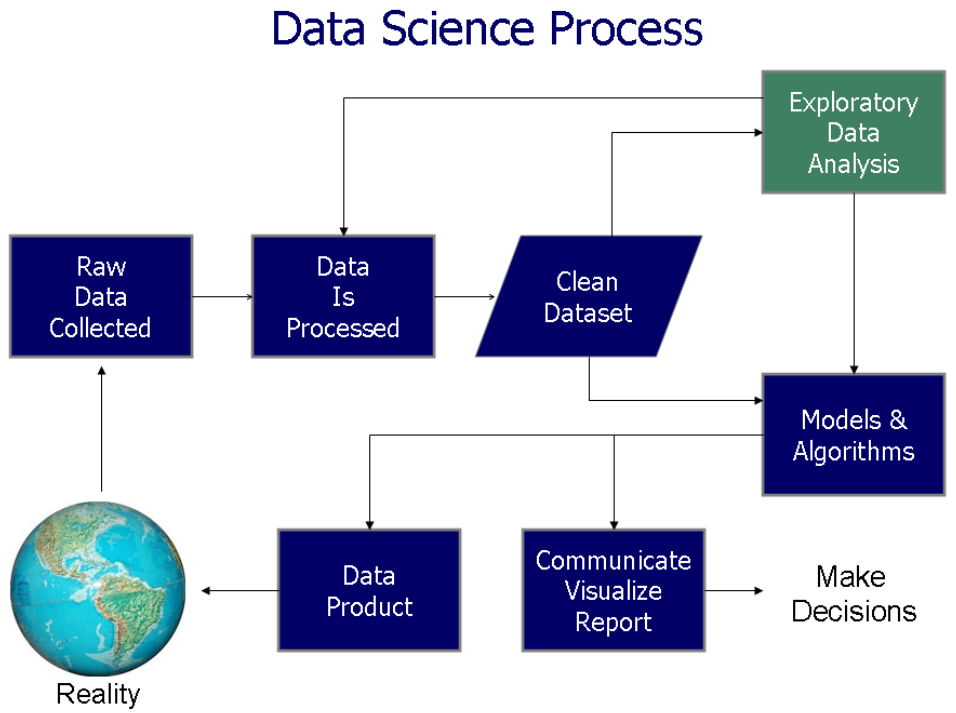
Interdisciplinary field:

- Mathematics
- Statistics
- Computer Science
- Etc.

Extracting knowledge or insights from data

Used in many areas:

- IT
- Genetics
- Finance
- Business Intelligence
- etc.



Global State of Data Science

Growth – Number of data scientists doubles over the last 4 years

IT industry hires largest number of data scientists

55% of all data scientists globally in USA, followed by Western Europe, Canada, India and Israel

Huge demand – 33.000 vacancies in USA on one job website (Glassdoor) alone

Salaries higher 30% on average than the rest of IT industry

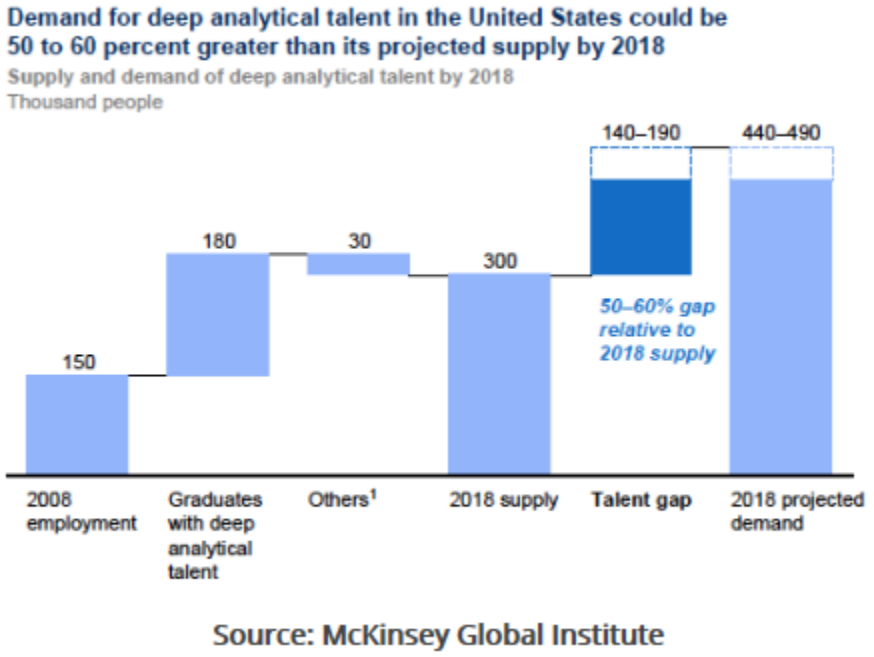
Prospects of Data Science

Over 100 data science programs in US universities

- MIT's $75.000 MSc

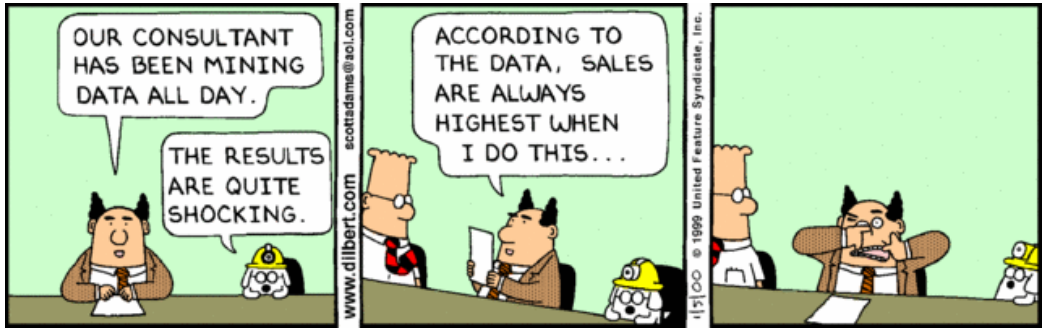
Demand still much greater than prospective supply

Serbia – with growing start-up scene and big investments in IT, growth expected to catch-up in the following years



Data Science Skills

Intuition for numbers



Data Analysis, Machine Learning, R, SAS, Python, SQL, etc.

Hiring from academia and other majors:

- Statistics
- Physics
- Computer Science
- Mathematics

38% with a PhD

Soft skills – Communication, Presentation, Management

Data Science at MDCS

Microsoft (corporation) invests heavily in data-driven culture

MDCS: follow -> lead the way

Several areas being worked on in several teams:

- Machine Learning
- Time Series Analysis
- Hypothesis testing
- etc.

Expanding role – **proving our value increases the team**

Data Science at SQL Team

Azure – Microsoft's cloud service

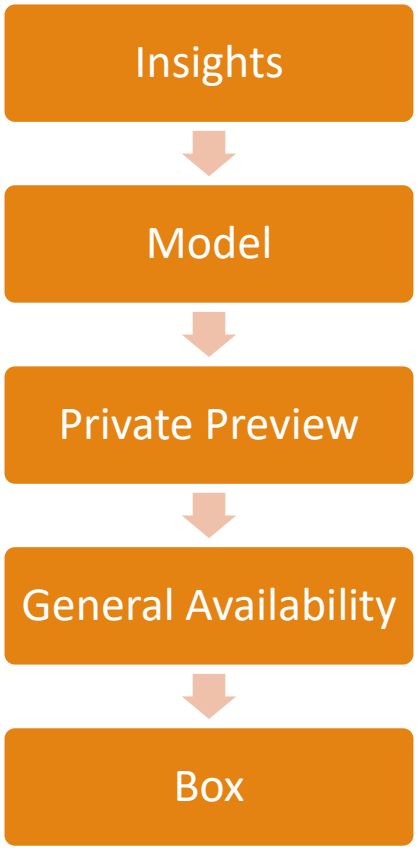
Azure SQL DB – service for storing SQL DBs on Azure

Box products – based on assumptions and communication with limited amount of customers

Cloud – possibilities of customer insight like never before

Insights transformed into models in order to fit SQL features to most customers

- 99% approach



Data Science Challenges

Recognizing a gap:

- Data Science for marketing (IT) or regulations (banks) purposes
- Apart from the trivial ones, insights/models not used on a fundamental level

Why is there a gap?

- Independence/Separation of Data Science teams
- Working on complex models
- Not scalable
- Needs internal marketing to push through
- Policies don’t work

Data Science Projects Participation

Our solution – **mixed teams**:

- Building expert systems
- No data science specific projects
- Working to the smallest details with SEs
- Complete understanding between each other's work and models

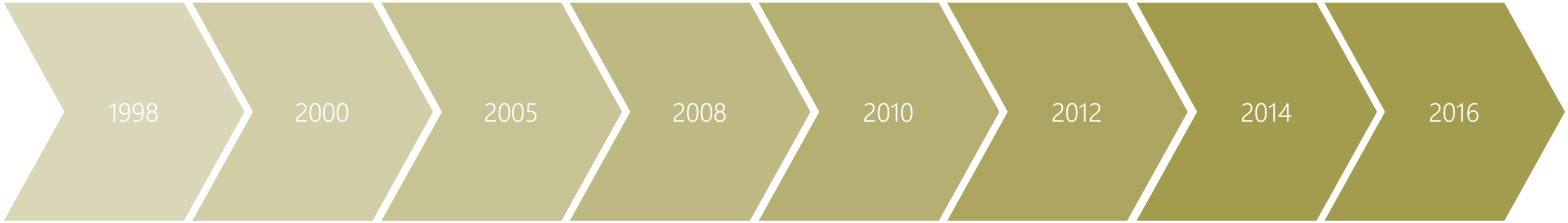
This leads to products with data scientists' signatures all over them

Absolutely never do DS specific stuff? Of course not!

- ML for targeted marketing
- AR + FFT for predictability
- Expert systems – root which brings freedom to experiment

DEVELOPMENT IN CLOUD

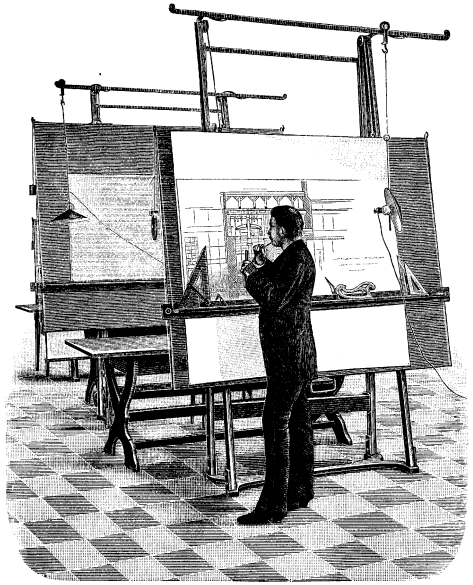
Traditional development cycle



Microsoft® SQL Server®

Traditional development cycle

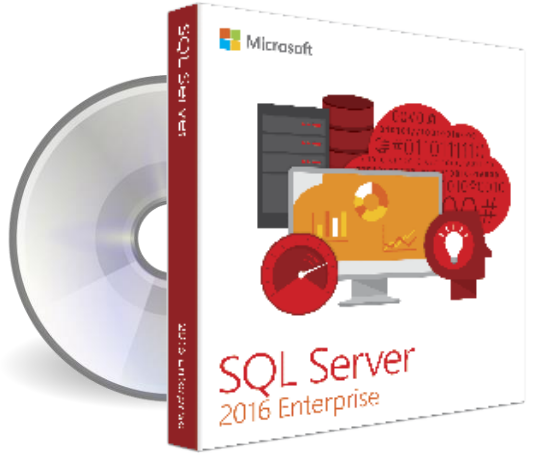
Design



Build

Test

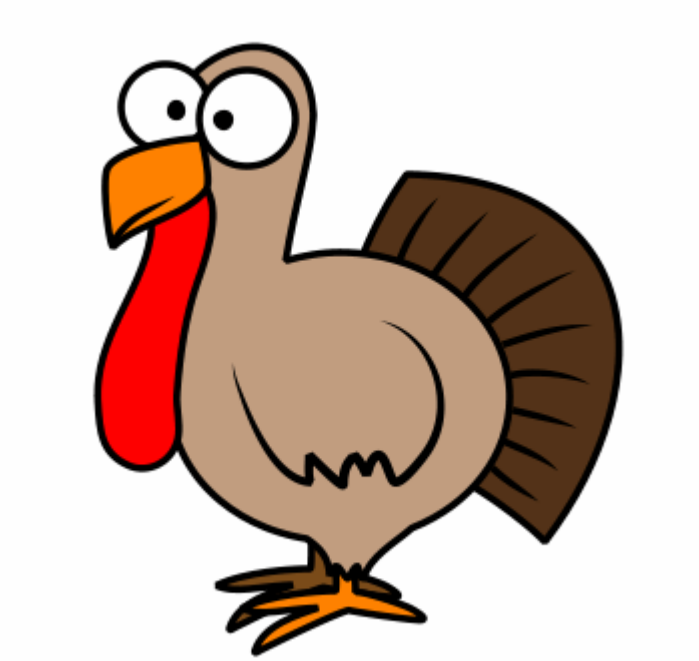
Release



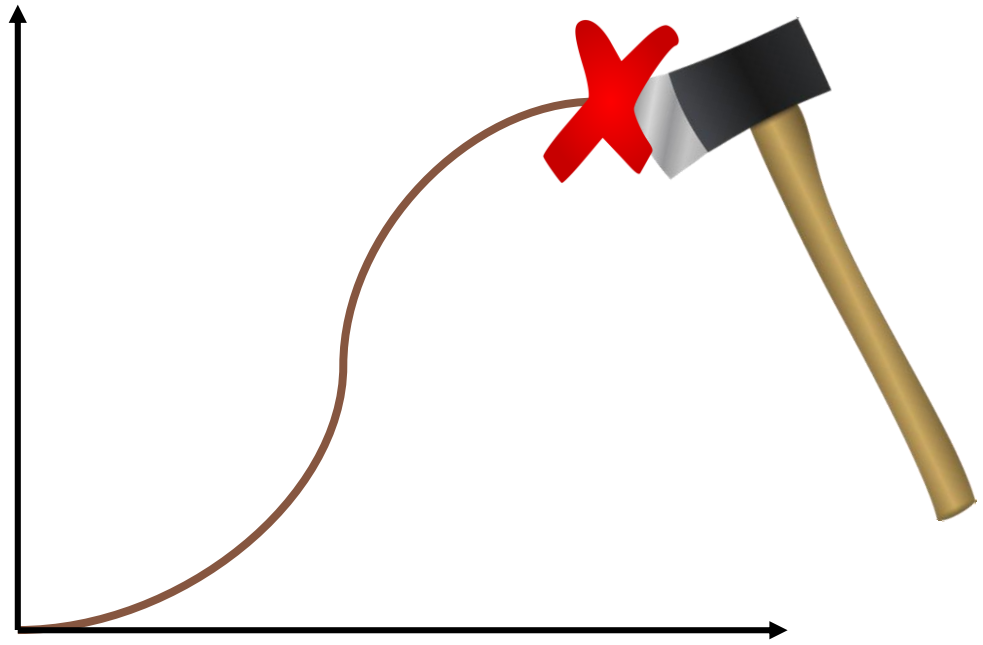
Main issue – fixed date:
Predicting future state of the market

Unknown unknowns

LIFE OF A TURKEY



LIFE OF A BOX FEATURE



Service development

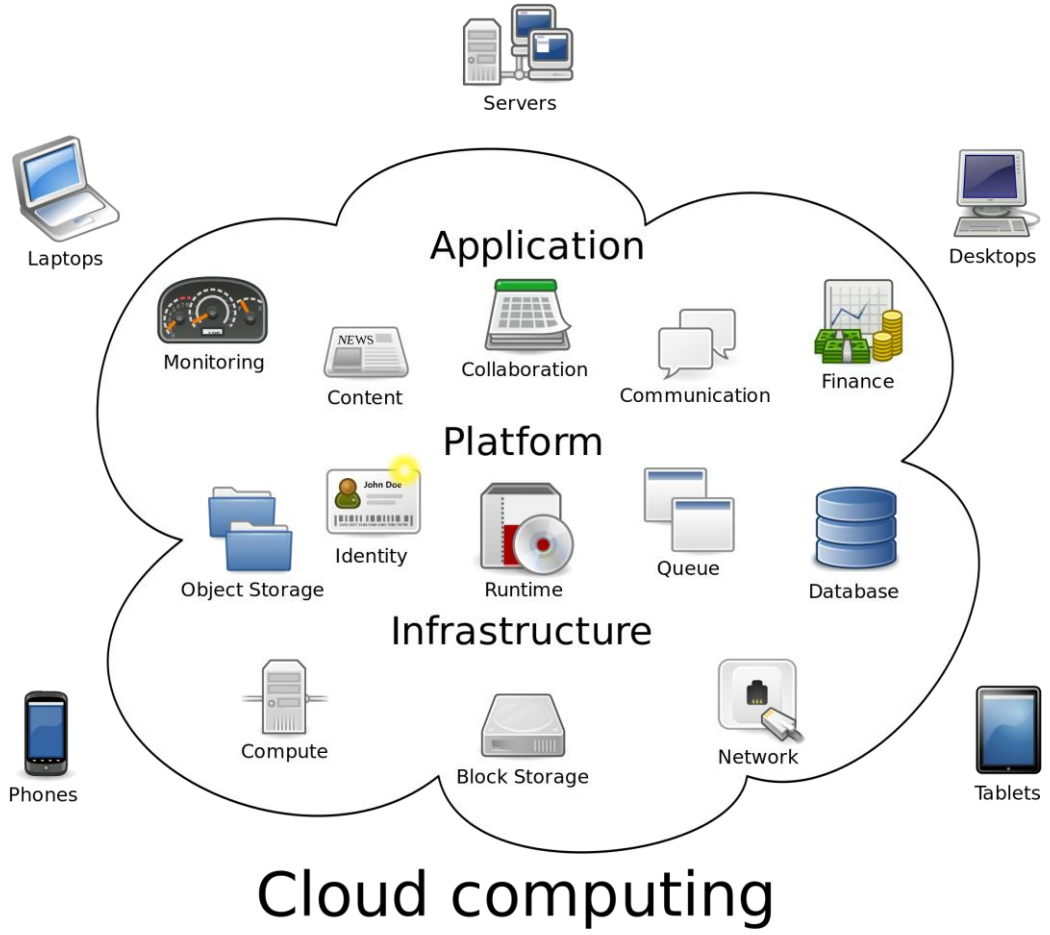
Service as opposed to product

Common framework for everyone

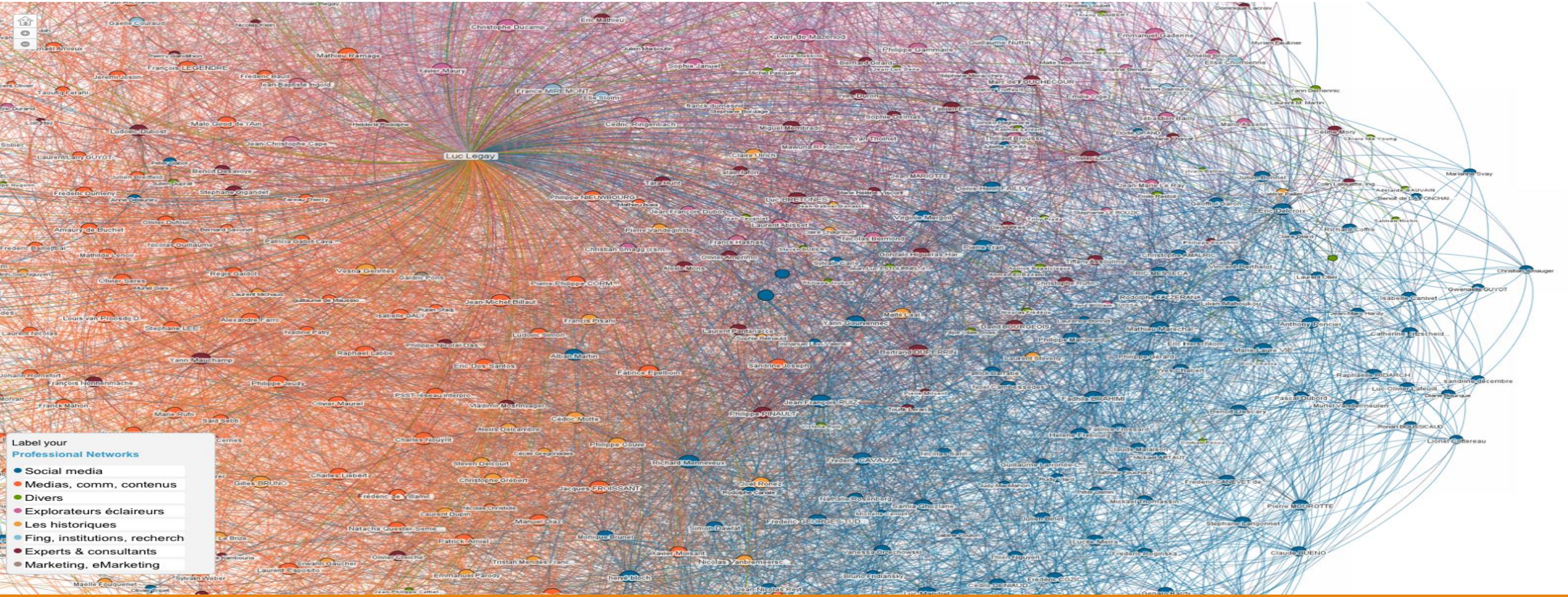
Allows for constant development and changes – development in real-time

Provides huge possibilities in collecting data about usage

Customization options



Now we have the data

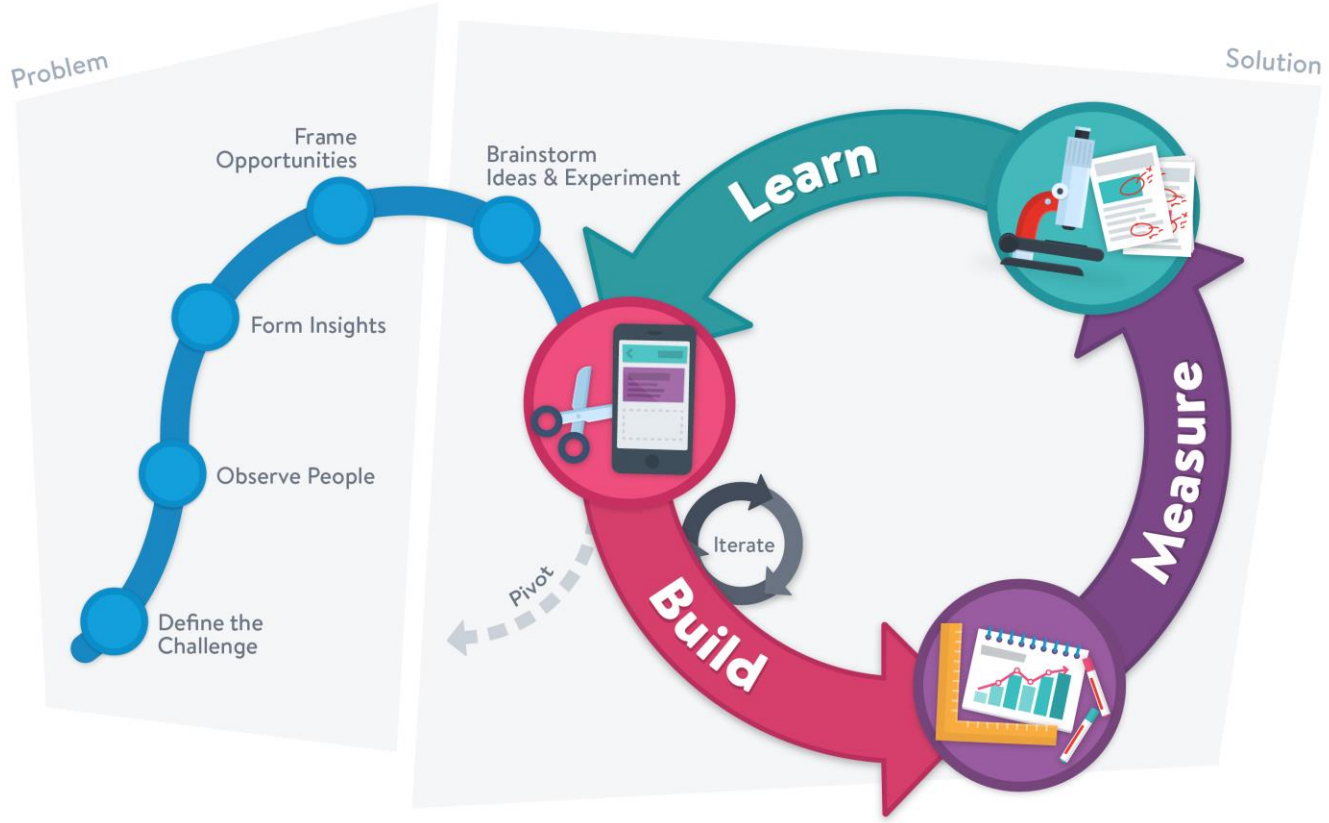


Service development cycle

Increasing amount of data -> increasing analysis possibilities

Data Science – part of each development component

Data Science in Build – automatic customization based on customer usage



Example - APM

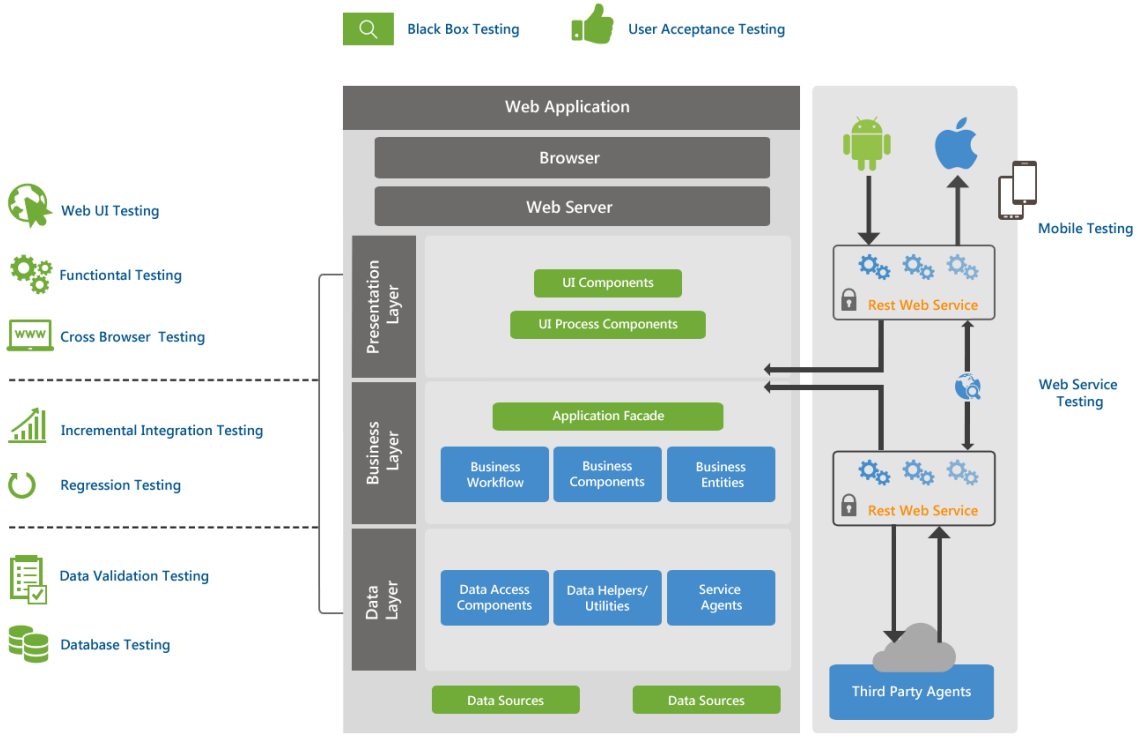
Most stressful and time consuming part of any development project – debugging the issues

Application – enormous amount of relations between codes/components unknown to even the best experts in the team

Cloud – constant deployments/changes increase and less testing increase the risk of failures in production

Incidents, lack of sleep

These need to be tracked proactively



Example - APM

Application Performance Management

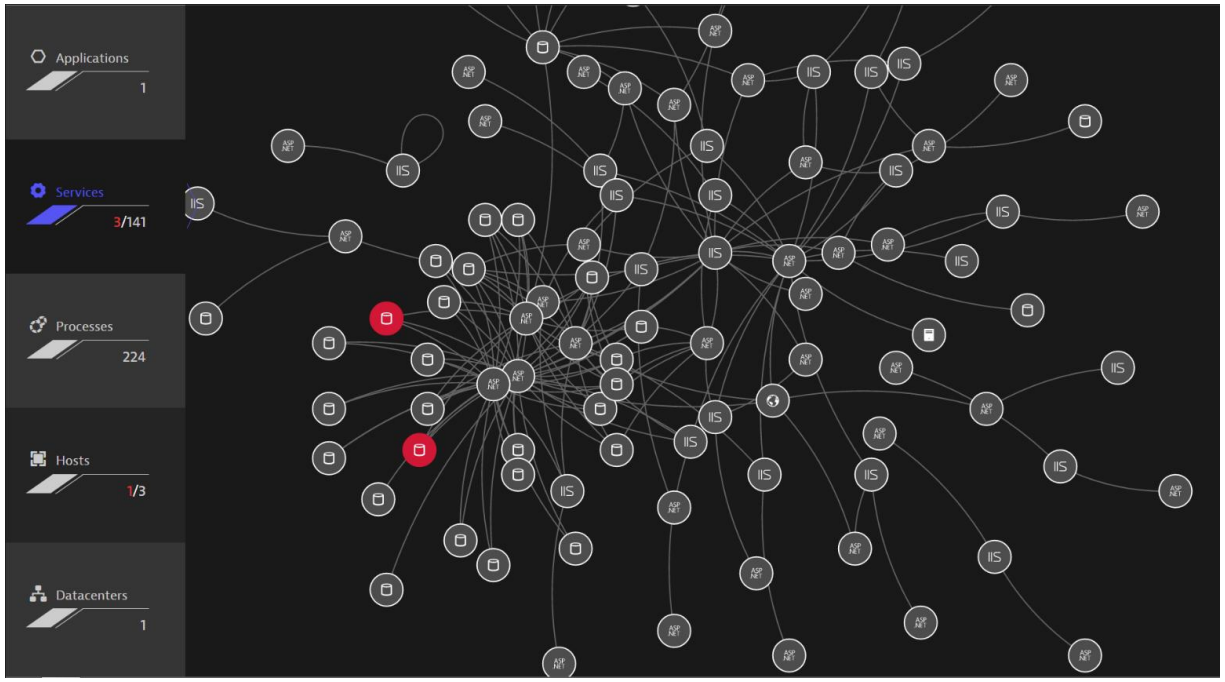
Monitoring/Management of performance and availability of software applications

Starting point of debugging

Current level (market leaders) – point to a component which underperforms

Mostly machine learning

No deep level root-causing



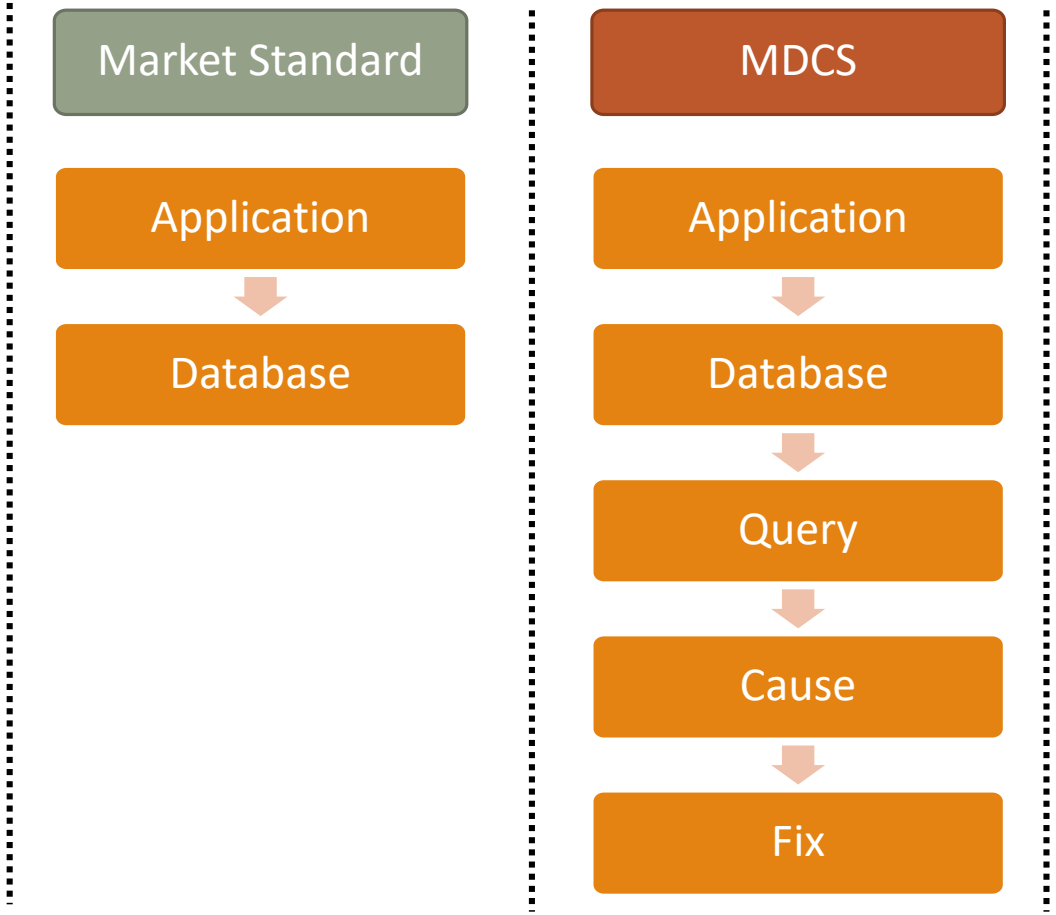
Example – APM

Niche – automatic RCA of application performance issues

Azure SQL DB at the moment

Three levels deeper than the current market leader

Huge value for the customers



Example - APM

Implement expert's (DBA's) mind into the code

How?

Combination of engineering and data science – Expert System:

- Data Science models – expert's eyes
- Engineering models – expert's knowledge


CENSURED


CENSURA


CENSORED


CENSURED


CENSORED

Instead of conclusion

If you want to become a data scientist:

- You need to be interested in getting insights from numbers
- Work hard
- Practice, practice, practice -> get intuition
- Be the best programmer among statisticians
- Be the best statistician among programmers
- Don’t just do it for the money – you ***will*** fail

Thank you for your attention


Q
&
A