Data Science opportunities outside academia

Data Science in the private sector and transitioning from research in fundamental physics.

Pizza Seminar IFAE

Mateo García Pepin

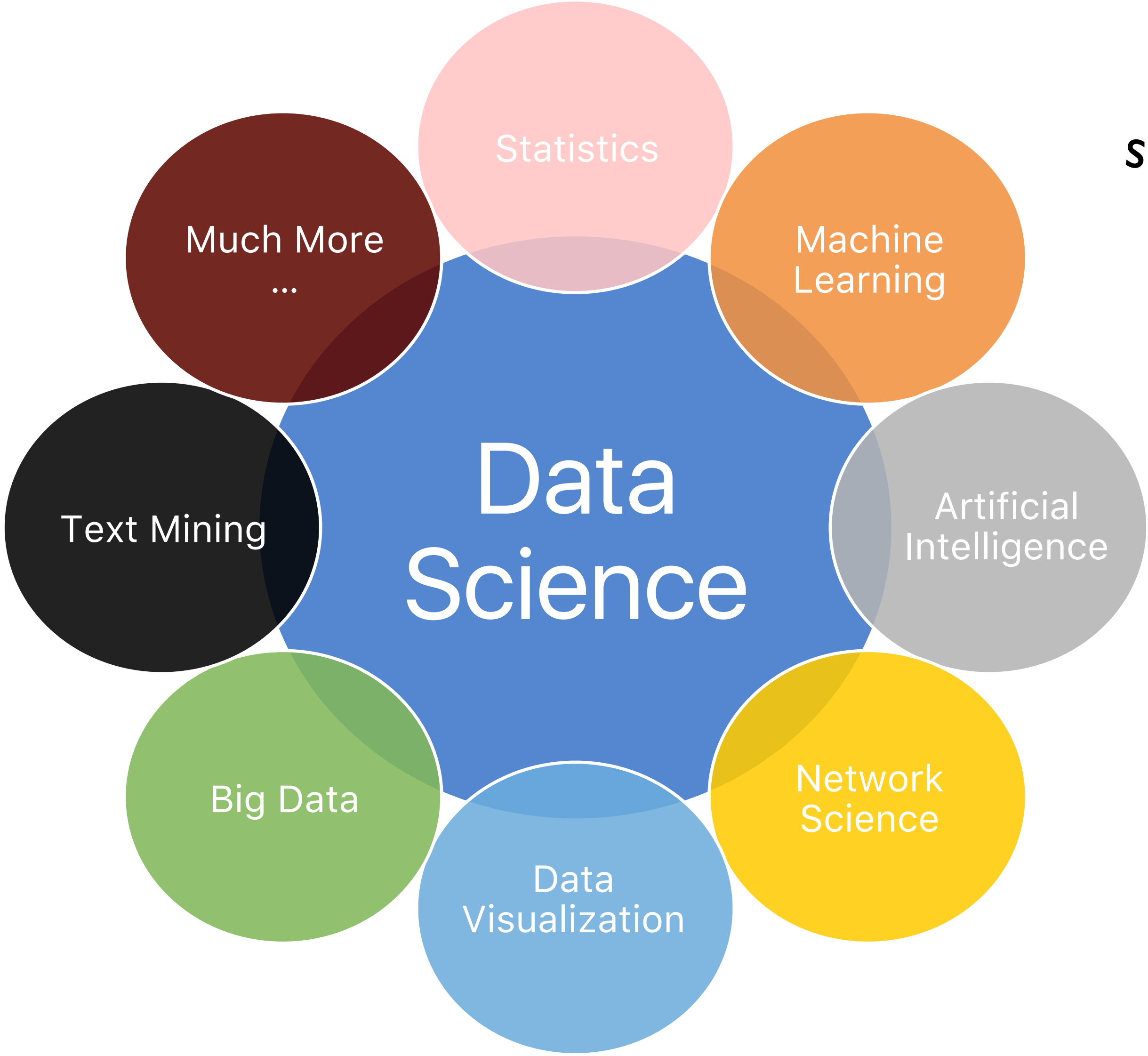
Disclaimer

I’m quite new to this world, however, I still feel that my recent experience might be interesting or useful for a few people.

What is Data Science?

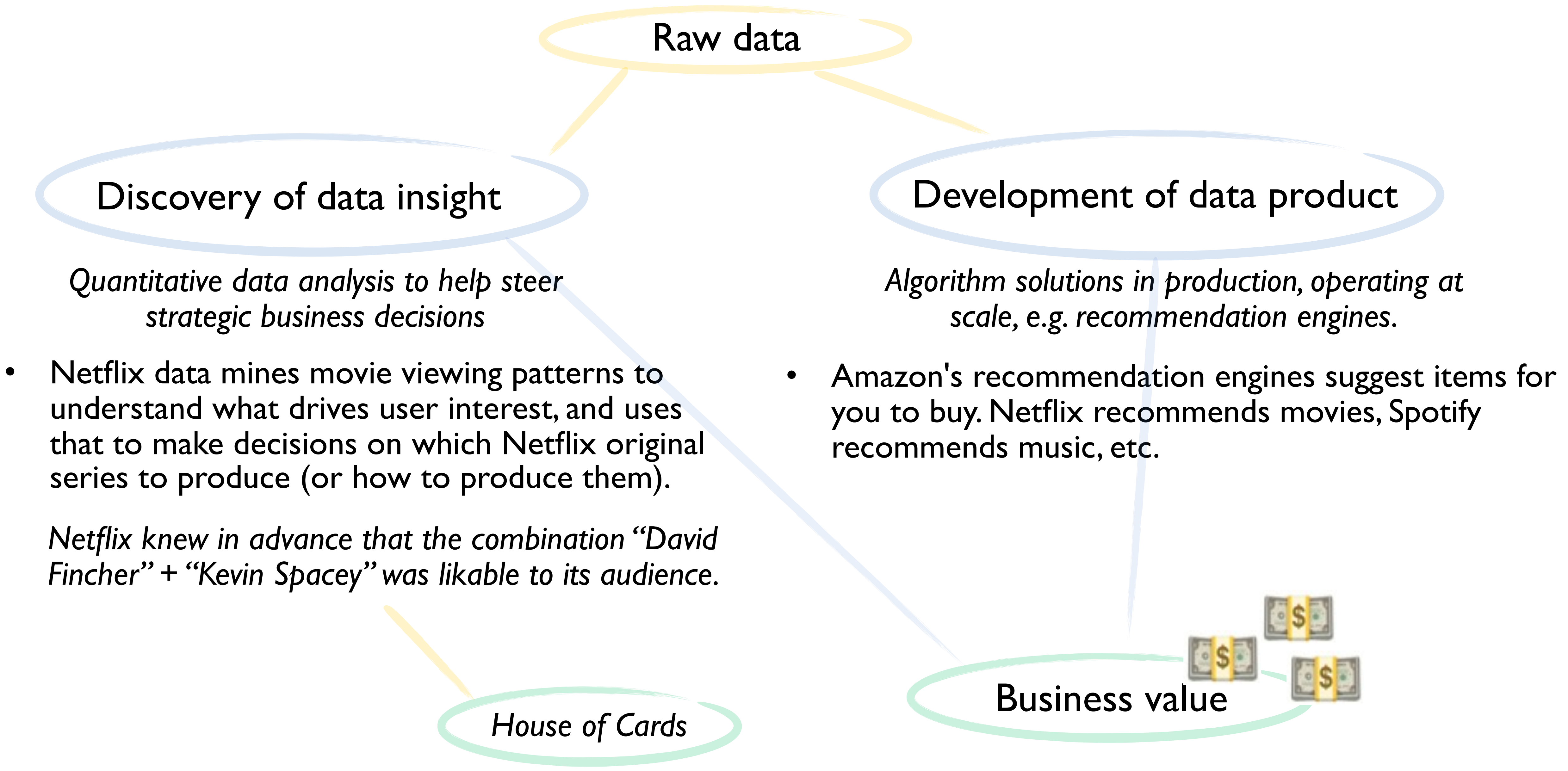
In particular, what is Data Science in the private sector? How is it useful for companies?

Data science is a multidisciplinary blend of data inference, algorithm development, and technology in order to solve analytically complex problems.



A very wide field which uses several tools depending on the problem to be solved

How can companies make a profit of it?



Raw data

Discovery of data insight

Quantitative data analysis to help steer strategic business decisions

- Netflix data mines movie viewing patterns to understand what drives user interest, and uses that to make decisions on which Netflix original series to produce (or how to produce them).

Netflix knew in advance that the combination "David Fincher" + "Kevin Spacey" was likable to its audience.

House of Cards

Development of data product

Algorithm solutions in production, operating at scale, e.g. recommendation engines.

- Amazon's recommendation engines suggest items for you to buy. Netflix recommends movies, Spotify recommends music, etc.

Business value

Particular examples

Customer value management

- Client segmentation
- Client behavior forecasting
- Target marketing optimization

Demand Forecasting

- Sales estimation by product or point of sale
- Replenishment Optimization
- Route planning

Geomarketing

- Quantification of potential demand and performance
- Point of sale network optimization

Personalized recommendation

- Identification of the best product or offer for each client in a given moment and channel

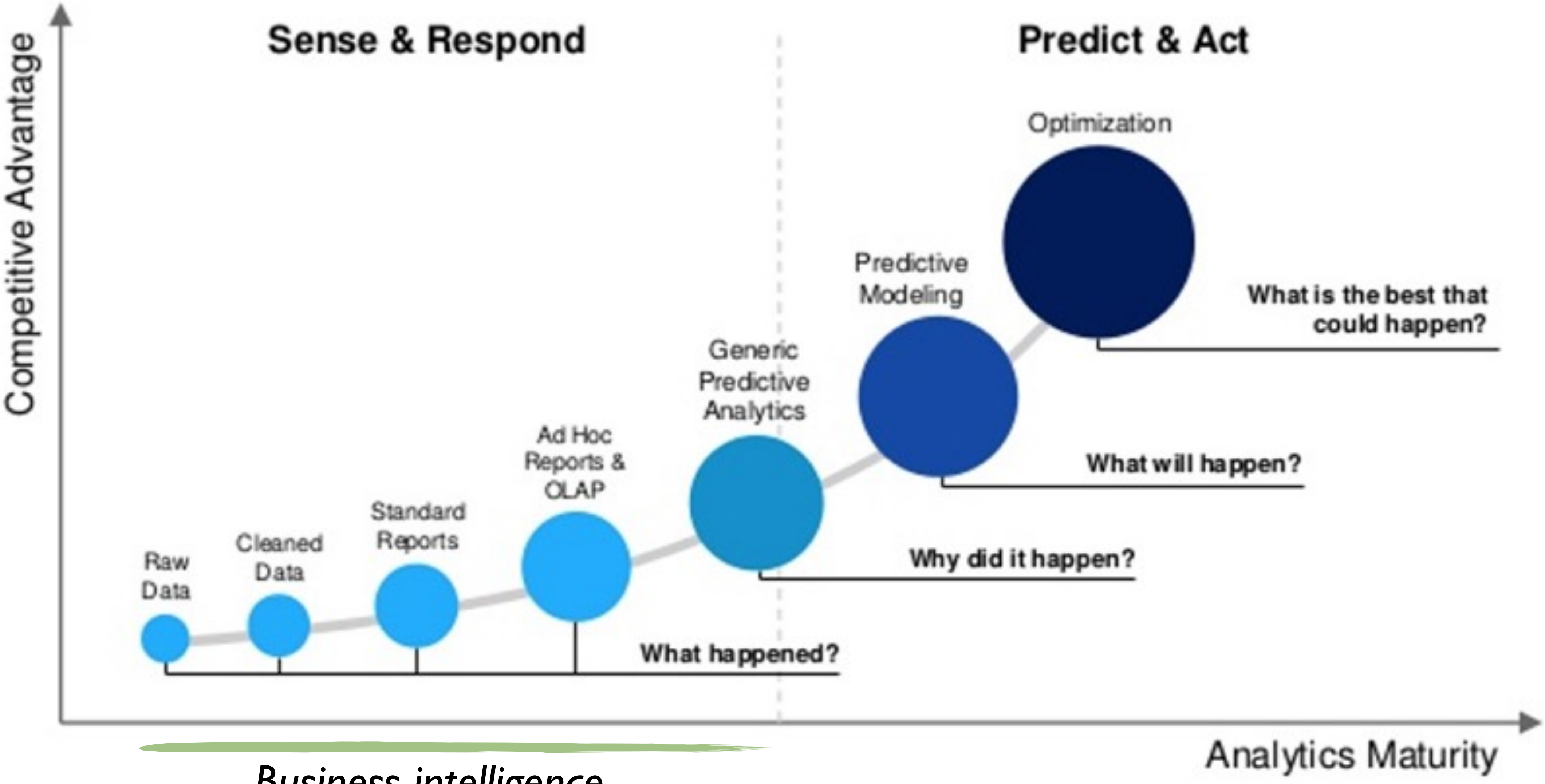
Revenue management

- Assortment optimization
- Setting prices
- Discount optimization and dynamic pricing

Marketing

- Measurement of advertisement impact both online and offline
- Marketing mix optimization

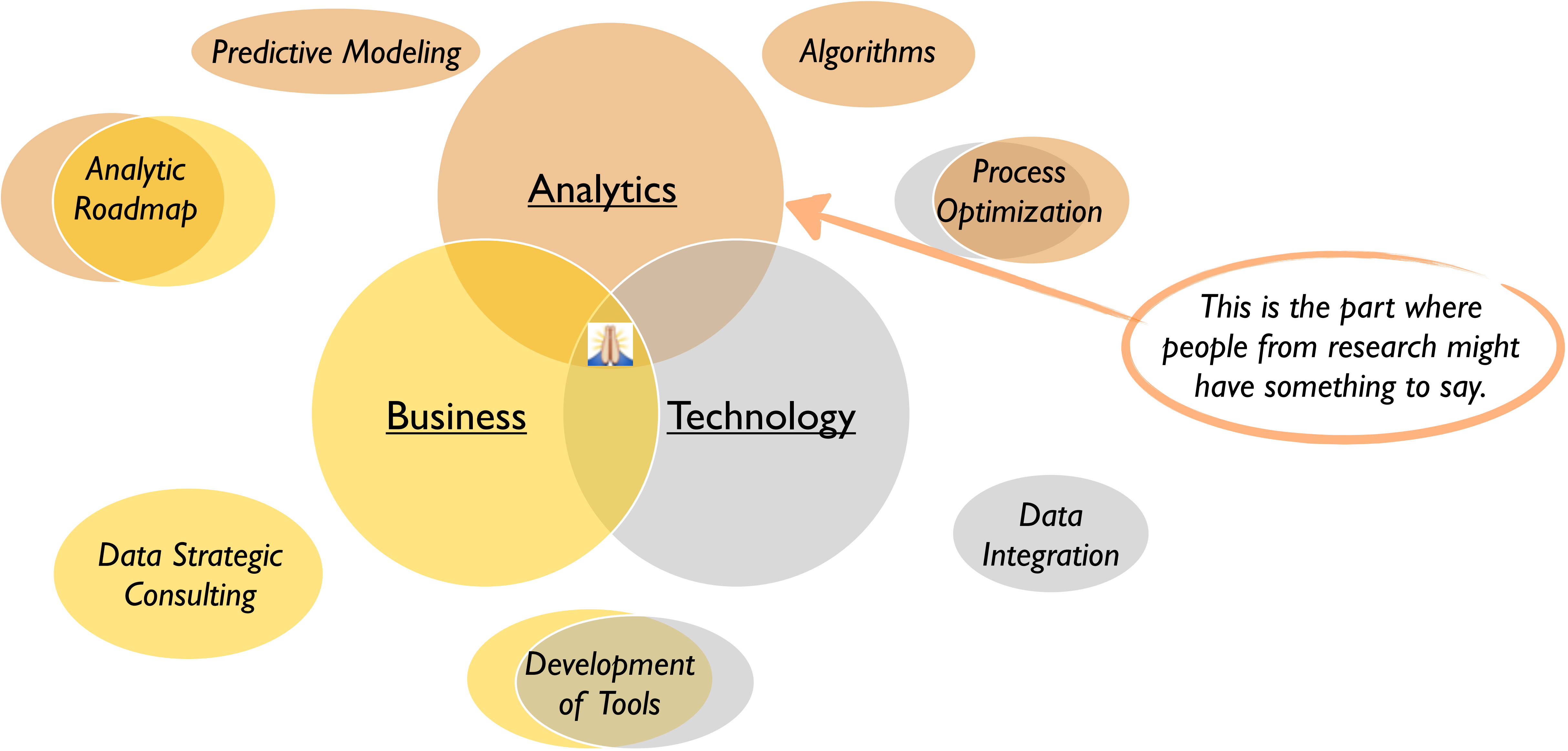
Evolution of Data Science



Business intelligence

Data Science

The set of required abilities for Data Science in the industry is threefold



Predictive Modeling

Algorithms

Analytic Roadmap

Analytics

Process Optimization

🙏

Business

Technology

This is the part where people from research might have something to say.

Data Integration

Data Strategic Consulting

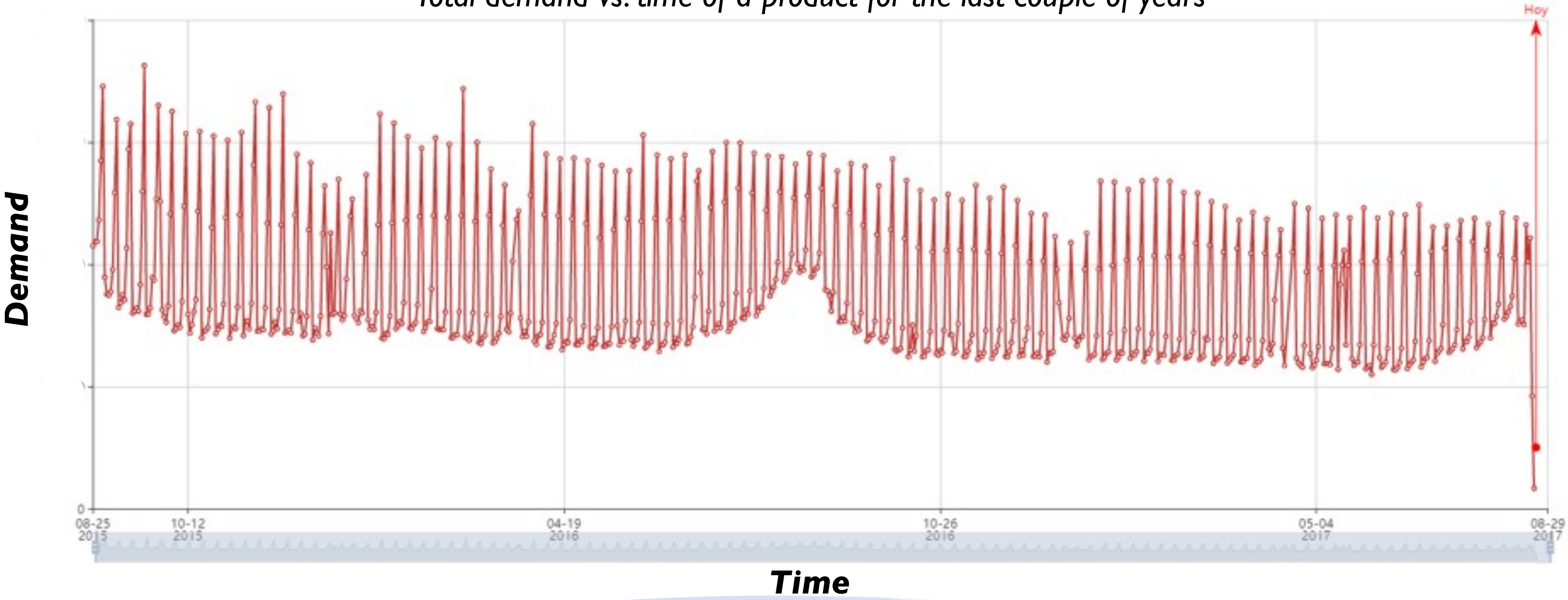
Development of Tools

What do we work on?

Example a project I have been involved in:
Development and maintenance of a web tool that uses time series, ARIMA models and other quantitative tools to forecast demand for a newspaper / magazine distributor.

The client needs to know how many units of its different products should be assigned to each sale point each day.

Total demand vs. time of a product for the last couple of years



How do we get future points in this plot?

In time series analysis, *autoregressive integrated moving average* (ARIMA) models are used to better understand data or predict future points of the series (forecasting).

Autoregressive (AR)

$$X_t = c + \sum_{i=1}^{p} \alpha_i X_{t-1} + \epsilon_t$$

+

Moving-average (MA)

$$X_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

ARMA

$$X_t = c + \epsilon_t + \sum_{i=1}^{q} \alpha_i X_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

Coefficients fitted when training with historic demand

The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values,

$$X'_t = X_t - X_{t-1}$$

Differencing removes the changes in the level of a time series, eliminating trend and seasonality and consequently stabilizing the mean of the time series. The differenced data is then used for the estimation of an ARMA model.

It makes sense to think that the demand of a Sunday is more related to the last Sunday than to the Saturday that came immediately before.

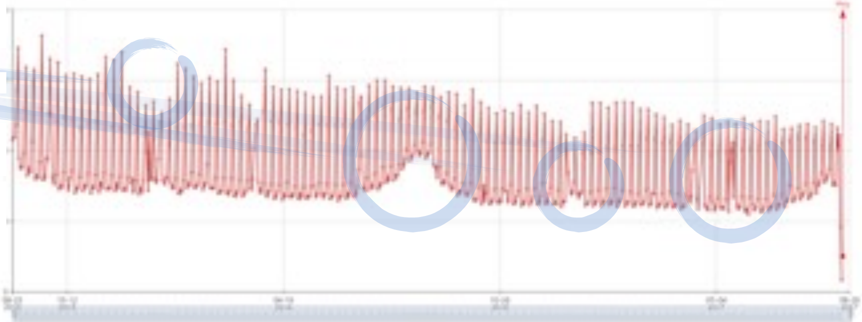
For that we use seasonal ARIMA models,

$$\mathrm{ARIMA}\underbrace{(p,d,q)}_{(X_t,\,X_{t-1})}\underbrace{(P,D,Q)}_{(X_t,\,X_{t-m})}{}_m$$

$$\mathrm{ARIMA}(0,0,1)(1,1,1)_7$$

Our parameter choice (the one that worked best without taking to long to train)

What about these bumps and dips?



<u>Annual seasonality</u>

Summer holiday season, Christmas, etc.

+ <u>Special events that modify demand</u>

News, results in sports, non-working days, etc.

To model these behaviors, we introduce extra variables,

$$X_t = \beta Y_t + \cdots$$

Can be a discrete (rare events, news, etc.) or continuous (Fourier coefficients to model yearly seasonality).

Some extra technicalities to take into account:

- **Non-observed demand**
  The informed demand is not the real one (it does not take into account clients who cannot buy because the sale point is soldout) and we have to add extra ghost demand.

- **Distribute demand**
  We aggregate sale points by zones of different characteristics (region, exact situation, etc.). For each of these we a have a separate ARIMA model that predicts demand which then has to be redistributed by sale points.

We do so by assigning a weight to each sale point that will determine how much of the overall zone demand will be assigned to it.

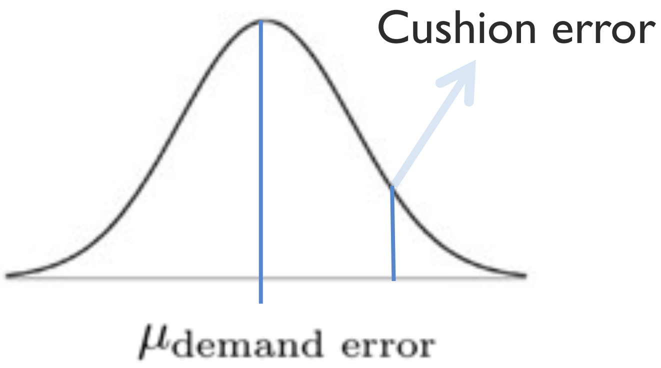
$$\omega_i = \frac{x_i}{\sum_j x_j}$$

For each sale point, sum of demand for similar days.

Sum over all sale points in each zone

- **Error handling**
  The client wants a limited number of soldout sale points. We add an extra cushion to our predictions to ensure that statistically the percentage of sold out sale points is where the client has set the objective.

For each sale point we have a distribution for prediction error in days of similar demand.



The cushion error is set at a percentile given by

$$\%_{\text{Cushion}} = 1 - \%_{\text{Soldouts}}$$

then

$$\text{Demand} + \text{Cushion} = \text{Demand}\,(1 + \text{Cushion error})$$

In general, I think that these are the main differences between industry projects and research:

- Here, you have to take care of many small details rather than one or two complicated problems.
- Timing is quite different. Less flexible time constrains, more urgencies to deal with, etc.
- The code needs to be easy to read and modify by other people. Comment code, avoid hard coding as much as possible, etc.
- You are building something that will be used by people other than you and your collaborators.
- The clients are not your peers and you cannot communicate with them as you would with people in your field. It is important to know what to say and how to say it.

Transition from research?

What did I do to get a job?

As a theorist I had very little programming skills and not so much knowledge of probability and statistics.

+

In Data Science, you have to compete with engineers and mathematicians who actually know how to code.

However, with a bit of preparation finding a job coming form research is pretty easy, even for a theorist. Mainly because:

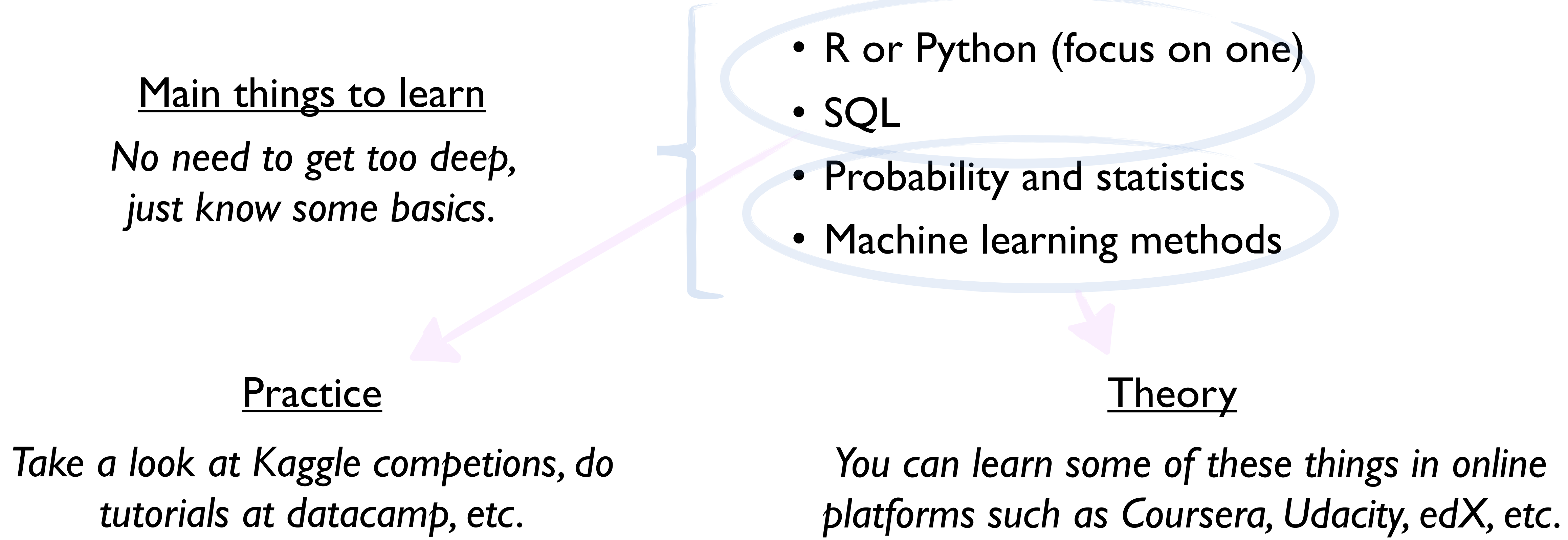
There is a lot of demand

+

Recruiters value very much these research-like abilities:

- Being able to explain the motivations and reasoning behind one’s work.
- Think critically about hard problems.
- Learn and adapt quickly.

An experimentalist can make a smoother transition to industry Data Science as their coding skills and statistical intuition are more developed.



Main things to learn

No need to get too deep, just know some basics.

- R or Python (focus on one)
- SQL
- Probability and statistics
- Machine learning methods

Practice

Take a look at Kaggle competions, do tutorials at datacamp, etc.

Theory

You can learn some of these things in online platforms such as Coursera, Udacity, edX, etc.

A couple of basics which are pretty useful (at least they where to me):

- The *Coursera* introduction to Machine Learning (the Andrew Ng course) is considered as a standard in the industry.
- Kaggle "Titanic: Machine learning from disaster" competition. Plenty of tutorials and scripts available.

Final comments

In my opinion, it is not as fun as physics, however, if you are decided to leave academia for whatever reasons you might also find it interesting / entertaining.

Things you probably never heard about: Git, database architecture, etc.

I spent 8 months working on a consultancy firm and I just started working on a US based start-up company.

- There are plenty of things to learn.

- Very wide field with possibilities of working in very different subjects.

- Interesting to work with people with different sets of skills, web developers, marketing experts, business managers, etc.

- Large job market (specially now, specially in Barcelona).

A bit of advertising

I started working at Kernel Analytics, a consultancy firm founded in Barcelona in 2013 specialized in analytics.

I now left for other reasons, still, they are nice people who value academic knowledge and they deserve the ad.

Kernel Analytics is a consultancy firm founded in Barcelona in 2013 specialized in analytics.

- It currently has offices in Barcelona and Madrid.
- About 70 employees with different backgrounds, some of them hold PhD degrees.
  - Engineering (Computer Science, Telecommunications, etc.).
  - Physics
  - Mathematics
  - Economy

Kernel
analytics

If you are interested send an email to the HR manager:
laura.moraleda@kernel-analytics.com

Thank you!