Is prediction STILL difficult?

Data Workshop Warsaw 2019

Marta Markiewicz

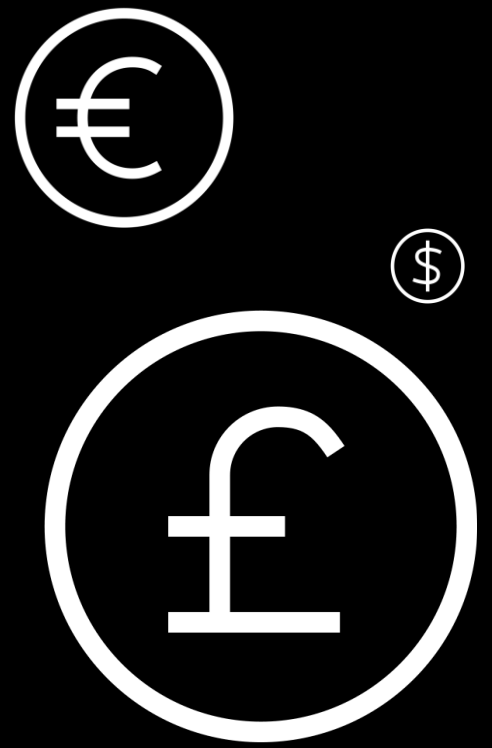
Objectivity

Agenda

1. Business challenge description
2. Evolution
   I. Classical approach
   II. Monte Carlo v1
   III. Monte Carlo v2
   IV. MCML
3. Summary

Business Challenge

- Objectivity specialises in bespoke software solutions and digital transformation
- We run multiple projects, of various team sizes and duration, for many customers at the same time
- Billed work for customers is our source of revenue
- Important tactical and strategical decisions regarding the future have to be made based on assumed future revenue levels

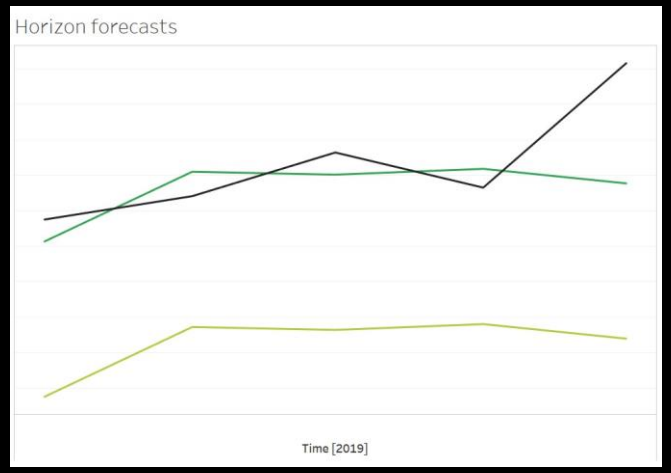


How to accurately forecast the revenue ?

2019
2018
2017
2016
2015

Classical approach

- Naive
- Exponential Smoothing
- SARIMAX
- Linear Regression



✓ fast
✓ familiar
✓ easy to explain

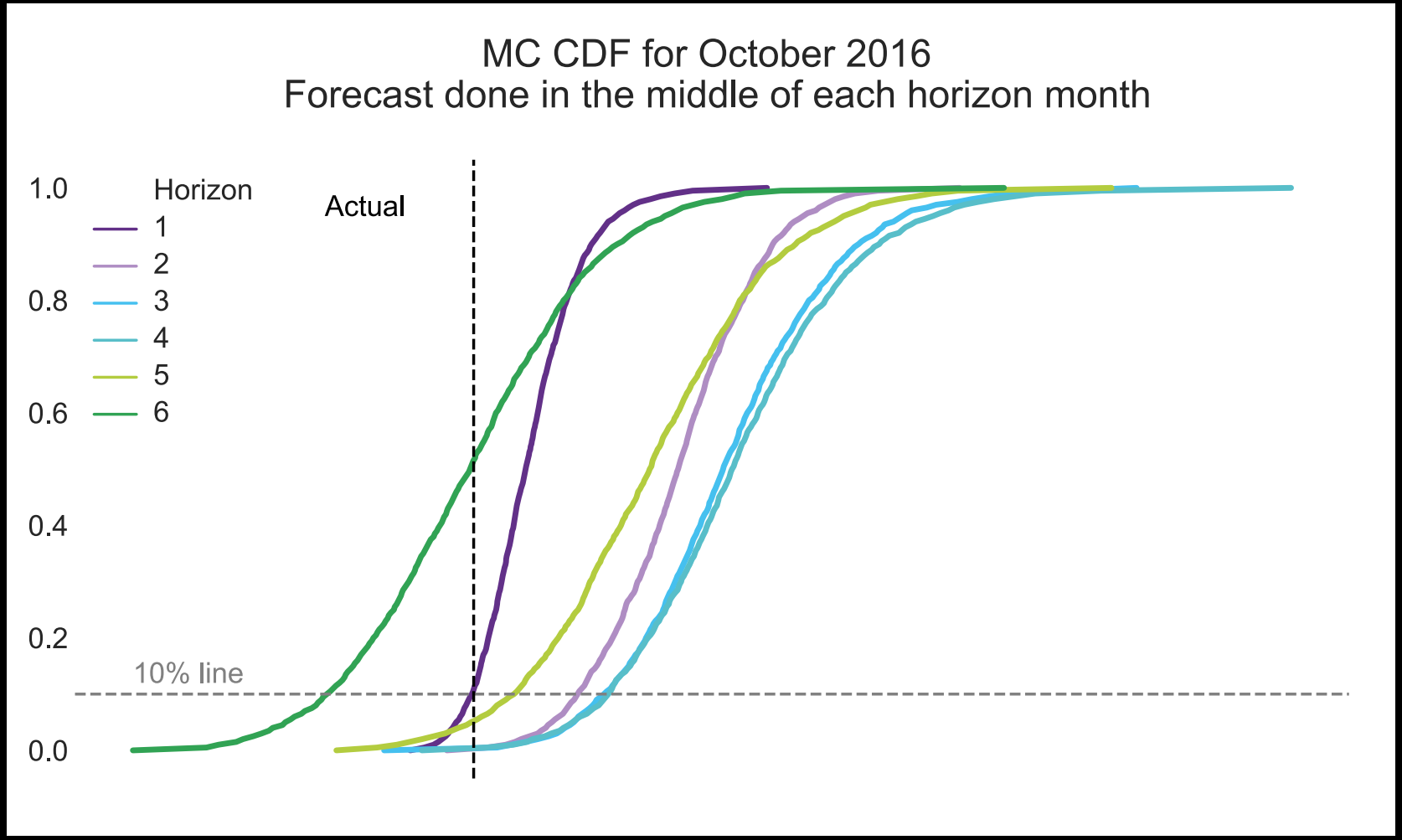
- failed when trend changed
- was not convincing for a bank

MC v1 Example data input

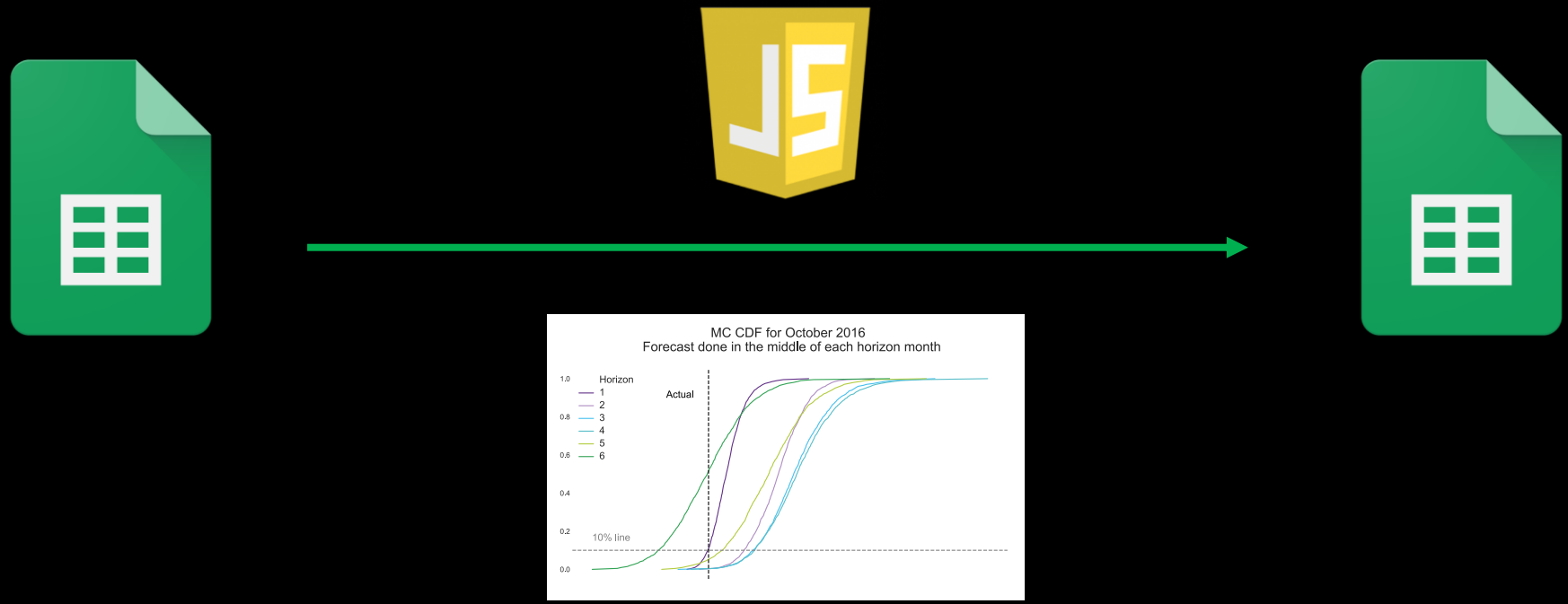
	A	B	C	D	E	F	G	H	I	J
1					Months					
2	**Opportunity Name**	**Group**	**Probability**	**Day Rate**	**1**	**2**	**3**	**4**	**5**	**6**
3	Cool Project 1	1	0.8	[illegible]	5	5	5	10	10	10
4	Cool Project 2	2	0.2	[illegible]	0	0	0	0	4	4
5										
6	Total estimated expected revenue				[illegible]000	[illegible]000	[illegible]000	[illegible]000	[illegible]200	[illegible]200

My Customer

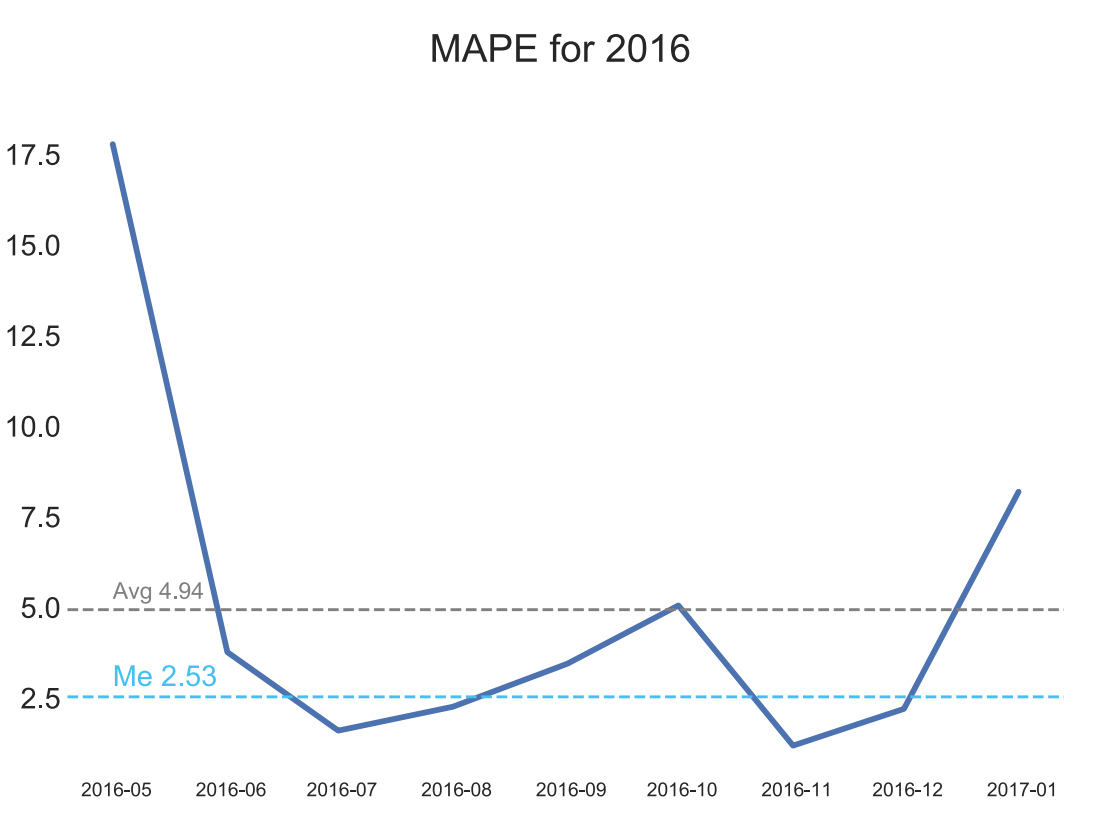
MC v1 Horizon changes



MC v1 Technology stack



MC v1 P&C



- ✓ simple understandable model
- ✓ fast productionisation
- ✓ driven by CEO
- ✓ frequent users input
- ✓ easy access

- ○ too simple to reflect reality
- ○ prone to users' errors
- ○ poorer with increase of horizon
- ○ no changes log
- ○ PoC style

MC v2 Methodology

Let's denote multipoint distribution by $\mathcal{M}(\mathbf{p}, \mathbf{v})$. $X \sim \mathcal{M}(\mathbf{p}, \mathbf{v})$ satisfies the following conditions:

$$
\begin{aligned}
&\mathbf{p} = (p_1, p_2, ..., p_N), \mathbf{p} \in [0,1]^N, N \in \mathbb{N} \\
&\mathbf{v} = (v_1, v_2, ..., v_N), \mathbf{v} \in \mathbb{R}^N \\
&\sum_{j=1}^{N} p_j = 1 \\
&P(X = v_j) = p_j \\
&E(X) = \sum_{j=1}^{N} p_j v_j
\end{aligned}
\tag{3}
$$

$$
\begin{aligned}
&O \sim \mathcal{D}(p, \mathbf{a}(S) * R * \mathbf{E}(S)) \\
&\mathbf{a} \in \mathbb{R}^{H+max(S)} \\
&N = |\mathbf{p}^{(rate)}| = |\mathbf{p}^{(effort)}| = |\mathbf{p}^{(start)}| \\
&R \sim \mathcal{M}(\mathbf{p}^{(rate)}, \mathbf{r}), \mathbf{r} \in \mathbb{R}_+^N \\
&\mathbf{E} \sim \mathcal{M}(\mathbf{p}^{(effort)}, \mathbf{E}), \mathbf{E} \in \mathbb{R}^{N \times H} \\
&S \sim \mathcal{M}(\mathbf{p}^{(start)}, \mathbf{s}), \mathbf{s} \in \mathbb{N}_+^N
\end{aligned}
\tag{4}
$$

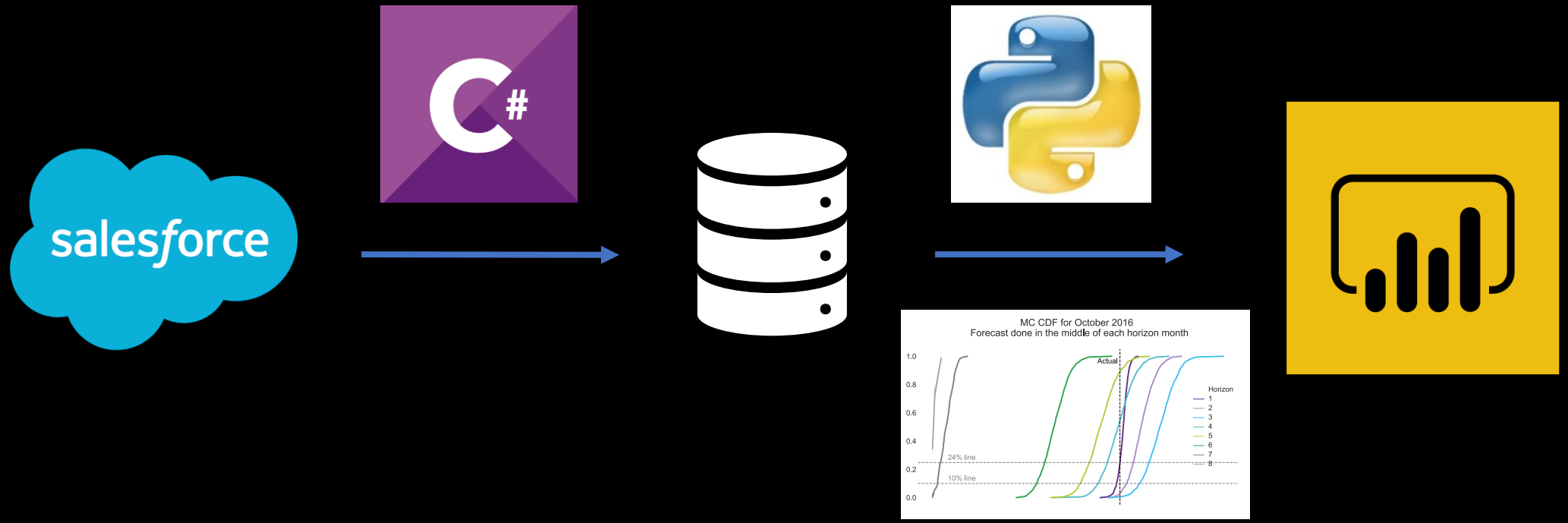
and the expected value of it:

$$
E(O^{(h)}) = p \sum_{j,k,l} p_j^{(rate)} p_k^{(effort)} a_{h+s_l} r_j E_{k,h+s_l}
$$

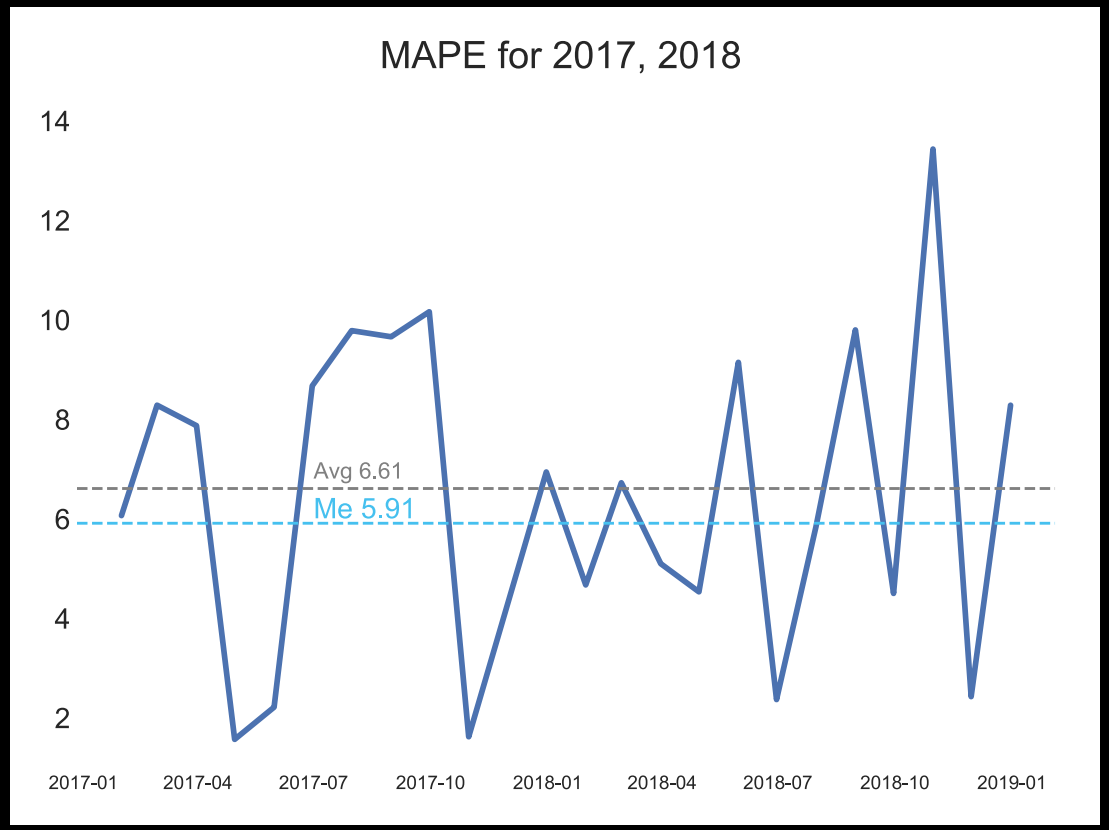
MC v2 Example data input

Opportunity	Card
Probability	0.8
Day Rate	P(Day Rate = [illegible]) = 0.4, P(Day Rate = [illegible]) = 0.6
Team Size	P(Team Size = [5,5,5,10,10,10]) = 1
Delay	P(Delay = 0) = 0.8, P(Delay = 1) = 0.1, P(Delay = 3) = 0.1
Skills Distribution	Devs = 60%, QE = 20%, BA = 10%, PM = 10%

MC v2 Technology stack



MC v2 P&C



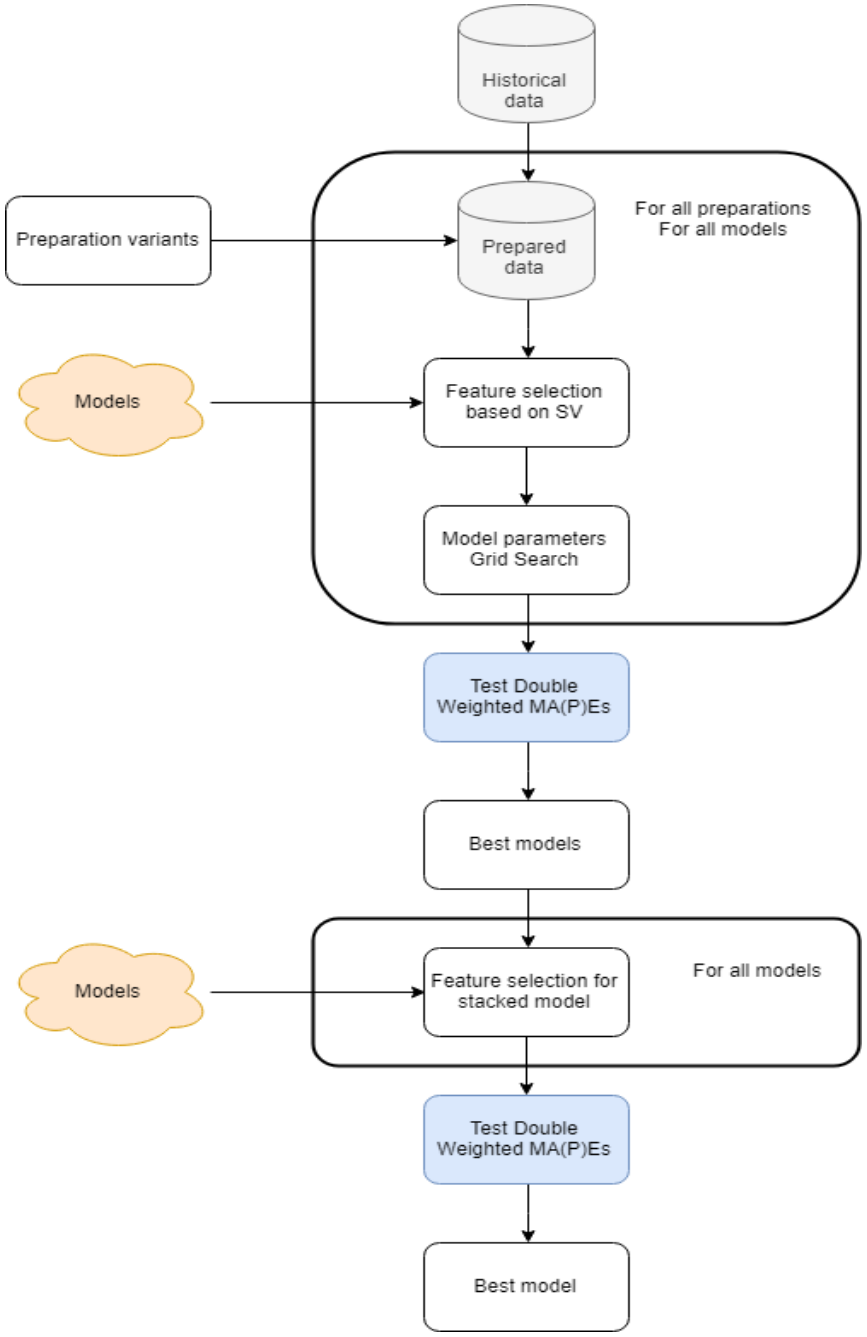
- ✓ close to real digital twin
- ✓ changes log
- ✓ immune to users' mistakes
- ✓ paired with dashboard

- painful tooling switch
- too complicated to understand
- data quality drop
- unfavorable effort to gain ratio
- high variance

2019
2018
2017
2016
2015

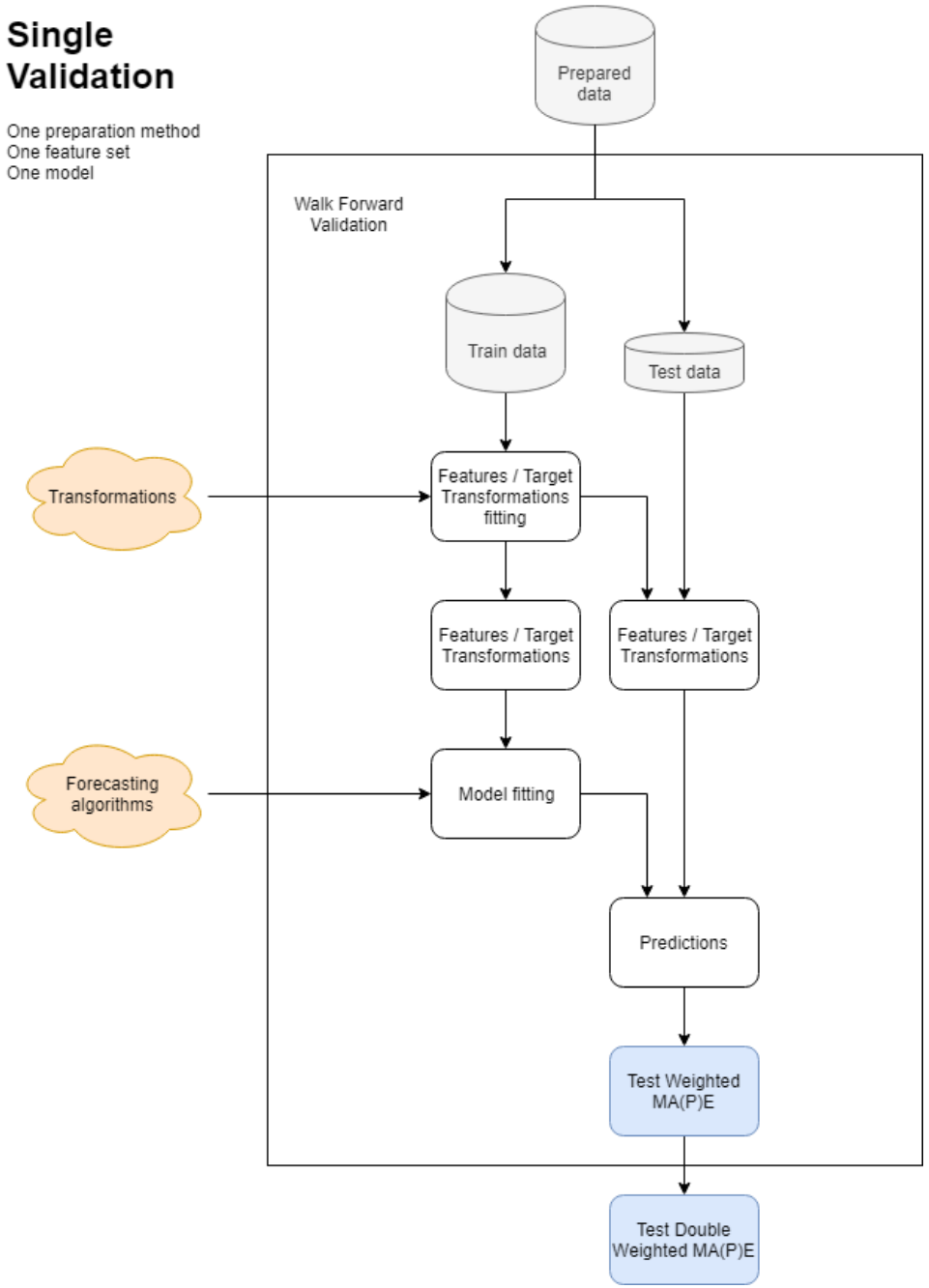
MCML Methodology

- Profits from users' input, Monte Carlo and Machine Learning algorithms
- MC, time and ARIMA like features
- Takes into consideration daily time aggregation, monthly time aggregation and the combined version of both

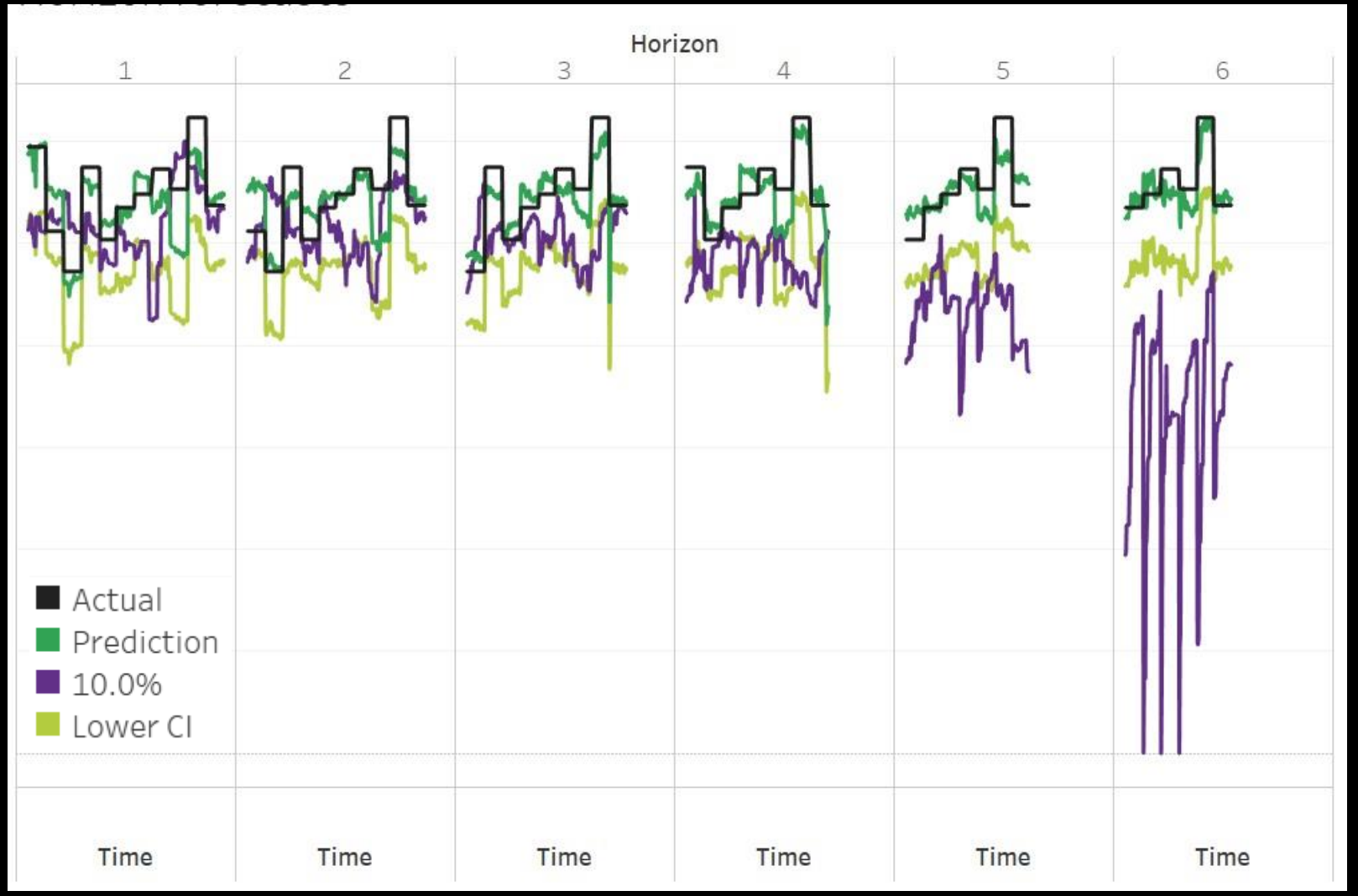


MCML Validation

- Every day new data are introduced to the CRM system
- Every day there is a 6 months ahead forecast produced
- Actual revenue values are known with a lag of about 2 weeks in relation to month's end



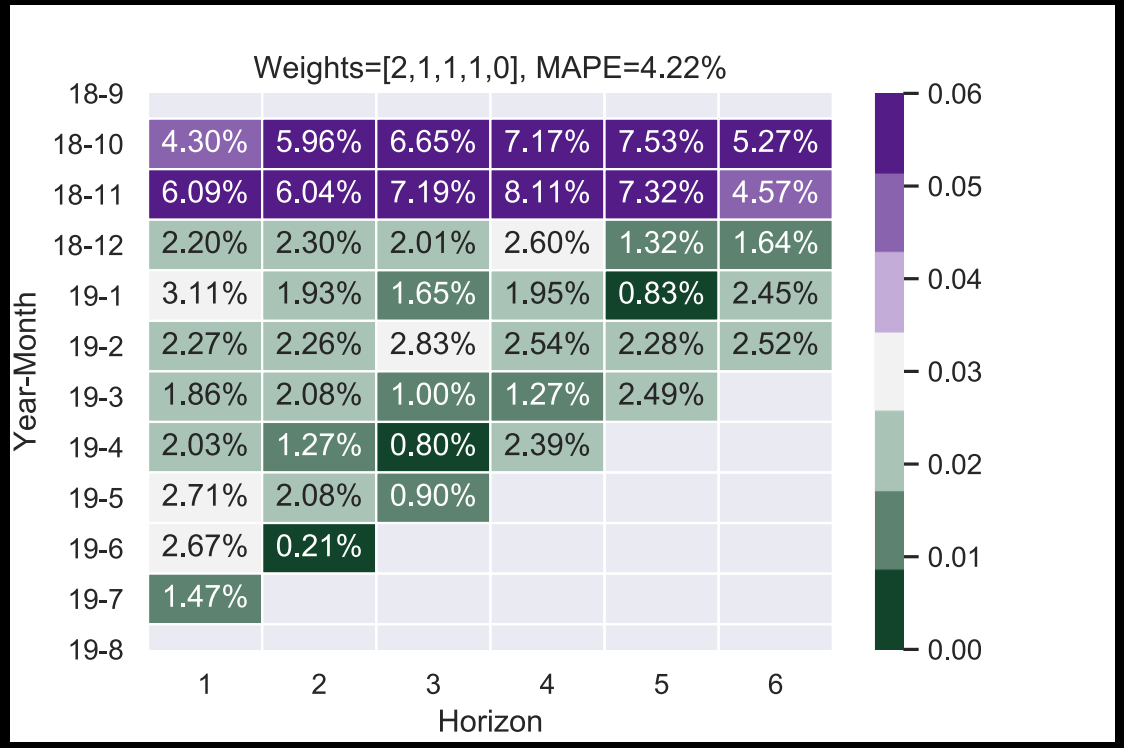
MCML Exemplary results

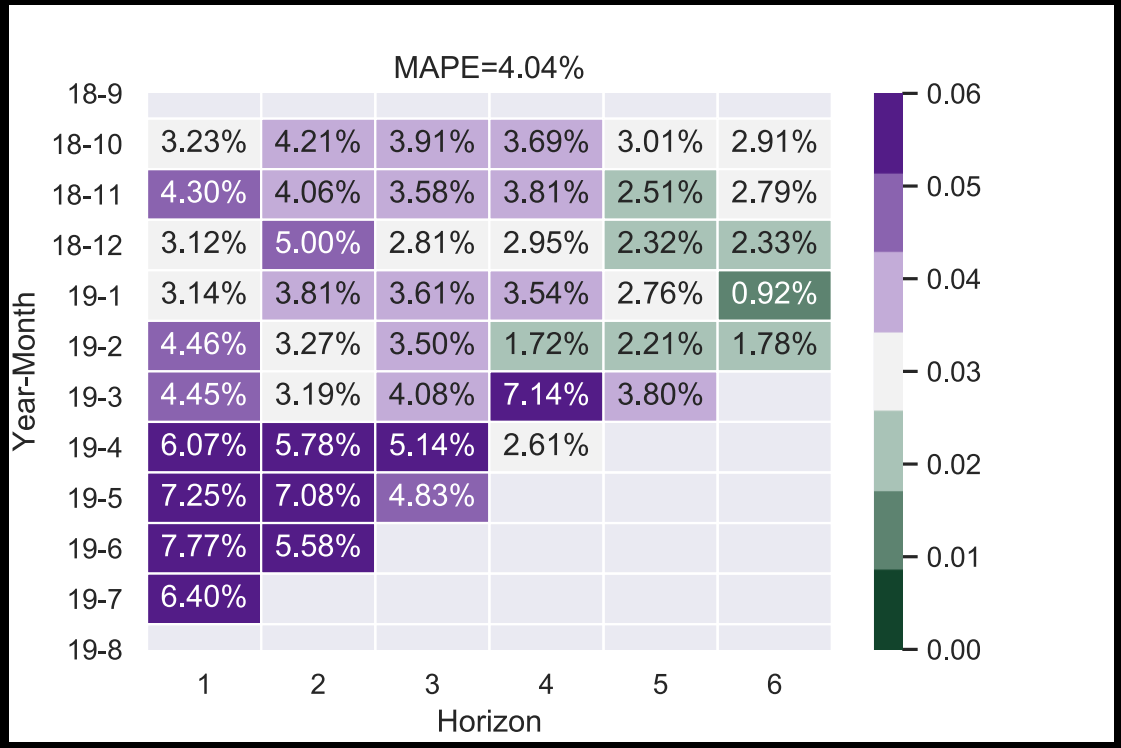


MCML Metric

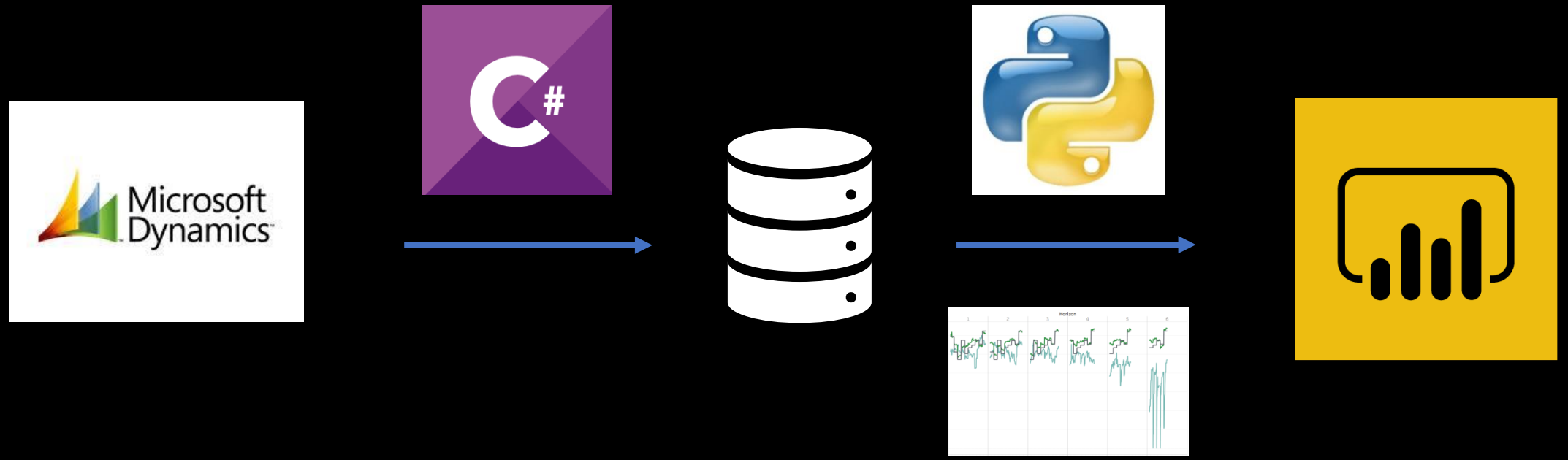
- Short term perfectness
  accuracy should increase with the decrease of h
- Continuous improvement
  accuracy should increase with every month that model runs on production
- Simplicity
  model should be as simple and explainable as possible (as it may happen that we will need to explain it in front of bank guys)
- Reliable worse scenario
  lower prediction intervals should have nice coverage (Financial Director needs to be prepared for the worse)

MCML Metrics comparison

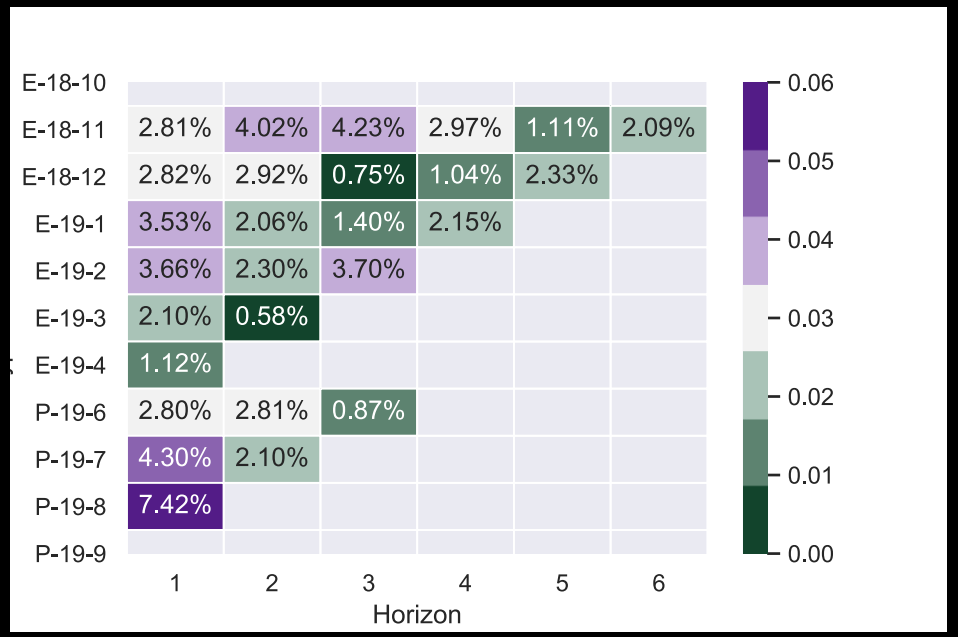


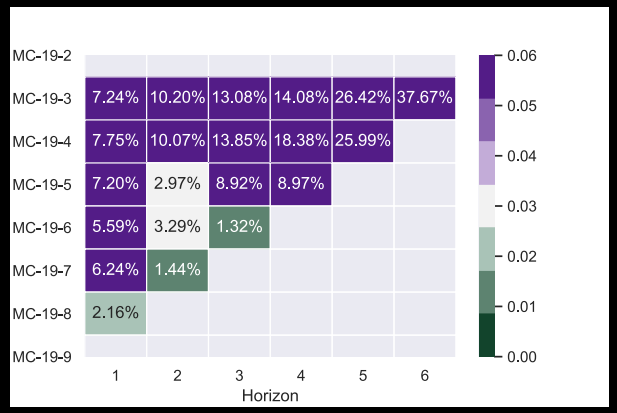


MCML Technology stack



MCML P&C

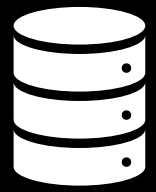


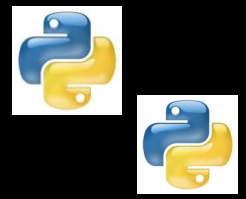


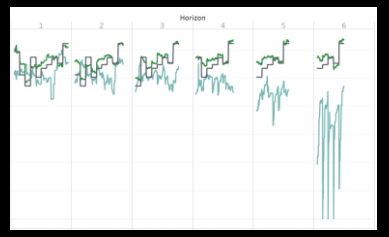
- ✓ promising results
- ✓ understands users optimism / pessimism
- ✓ improves over time

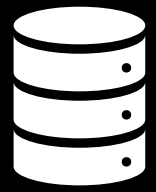
- ○ very complicated to explain
- ○ decreases the motivation for users to keep the data clean
- ○ too short on production to be sure there are no other cons

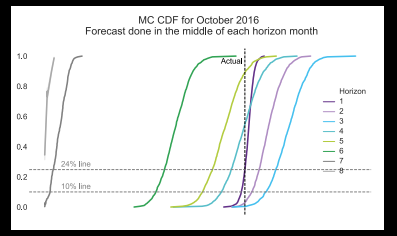
Evolution Summary

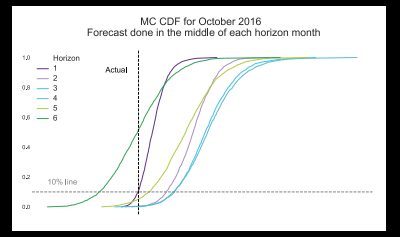












Pieces of advice

Start small

Don’t be afraid to start small. If it’s not harmful, it’s better to have simple solution but on production, that have only pure research

Challenge complexity

Challenge any additional complexity in manual data collection. Gain may not be worth it

Define success metrics wisely

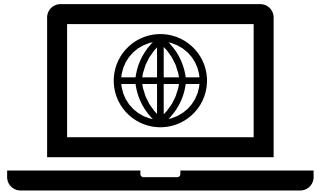
Of-the-self classics may not be suited for your needs

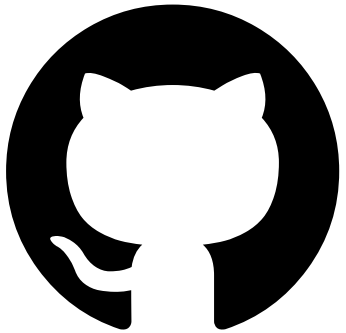
Leverage expert knowledge

Domain knowledge in people heads runs the world for years, use it to your advantage

Don’t be ashamed to use simple models

Check simple models along with state-of-the-art. Worse case scenario, you will have benchmark values. Best - they may surprise you





Thank you!