

Data Markets for Energy

Pierre Pinson

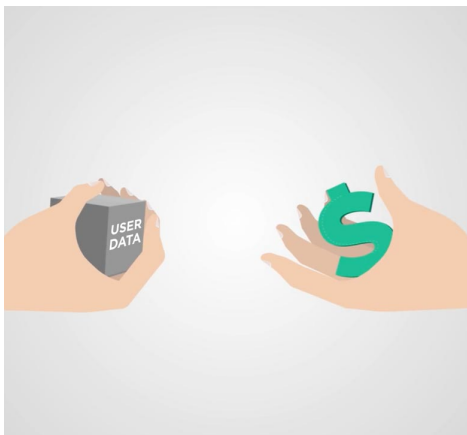
Technical University of Denmark
Department of Technology, Management and Economics

DTU Summer School 2022 – 22 June 2022

(acknowledgements to all collaborators, funders and data providers)

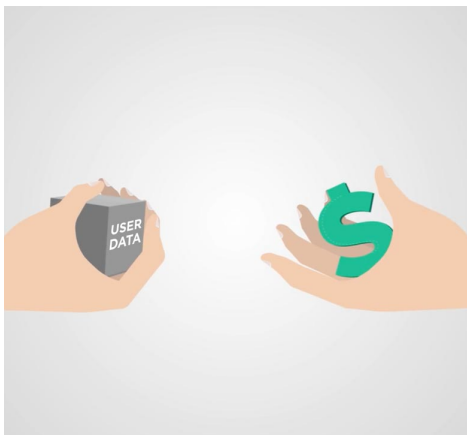
- 1 A few introductory experiments
- 2 The context
- 3 The how's and the why's of collaborative and market-based analytics
- 4 Data markets for regression and forecasting problems
- 5 Simulations and case-study application
- 6 Concluding thoughts and discussion

1 A few introductory experiments



Experiment 1:

Do you have a data point you may like to sell?



Experiment 1:

Do you have a data point you may like to sell?

The underlying “issues”:

- What is a data point?
- Why would I sell it?
- How do I determine the price I would want to sell it for?



Experiment 2:

Do you want to buy a data point from that seller?



Experiment 2:

Do you want to buy a data point from that seller?

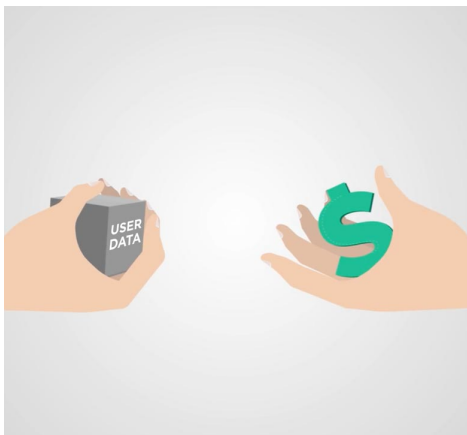
The underlying “issues”:

- How would you use it?
- What would you use it for?
- Is that even useful?
- How much you would be ready to pay for it?



Experiment 3:

Do you see ways to be malicious (and to prevent it)?



Experiment 3:

Do you see ways to be malicious (and to prevent it)?

Example "issues":

- What if copy my data (with slight modifications) and sell it several times?
- What if I lie and sell rubbish data?

other more general issues?

2 The context

Today, everything has to be smart!

Smart Energy



Smart Cities



Industry 4.0



Smart Transport



Smart Agriculture



etc.

Today, everything has to be smart!

Smart Energy

Smart Cities

Industry 4.0

Smart Transport

Smart Agriculture

“Smart”
=
Data
+
Analytics

Data is more valuable than oil

The Economist, May 2017:

"The world's most valuable resource is no longer oil, but data"



For many application areas, we are reaching a tipping point where data may become the most valuable commodity... through analytics!

1995/2005 – ...



- Data collection is increasing at an astounding rate (order of billions of GB per day!)
- This motivated research efforts towards **big data** analytics

- Data collection and storage is decentralized
- This led to a focus on **edge**, **cloud** and **fog computing**

1995/2005 – ...



- Data collection is increasing at an astounding rate (order of billions of GB per day!)
- This motivated research efforts towards **big data** analytics



- Data collection and storage is decentralized
- This led to a focus on **edge**, **cloud** and **fog computing**



Remaining
gap:

Data ownership is also **distributed**, with agents having **heterogeneous preferences** (privacy, competition, willingness to share, etc.)

1995/2005 – ...



- Data collection is increasing at an astounding rate (order of billions of GB per day!)
- This motivated research efforts towards **big data** analytics

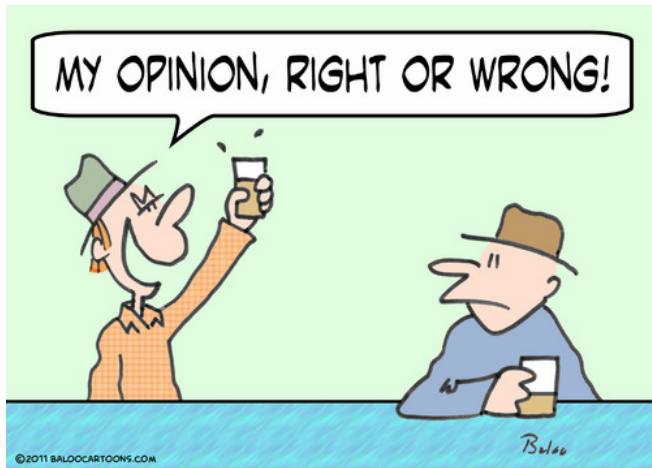


- Data collection and storage is decentralized
- This led to a focus on **edge**, **cloud** and **fog computing**



Proposal
solution:

Collaborative and market-based analytics!



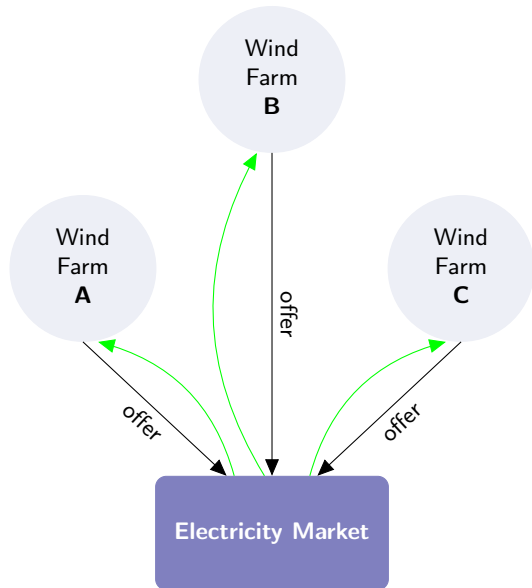
Are we not doing it already?

- ③ **The how's and the why's of collaborative and market-based analytics**

A motivating real-world example

Context:

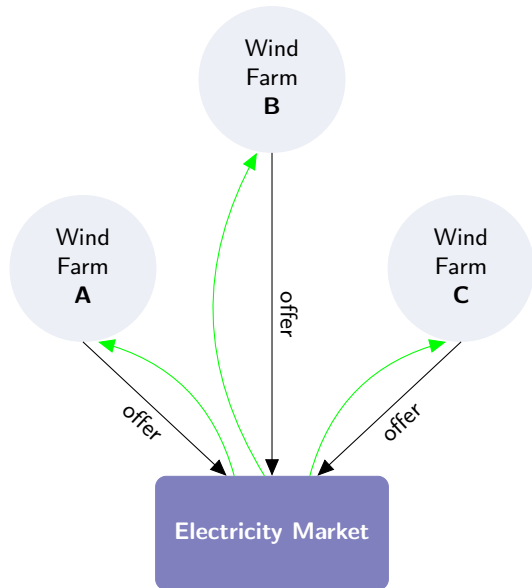
- Wind farms offer in electricity markets based on their individual forecasts and private information
- Their revenue is affected by their (lack of) forecast accuracy



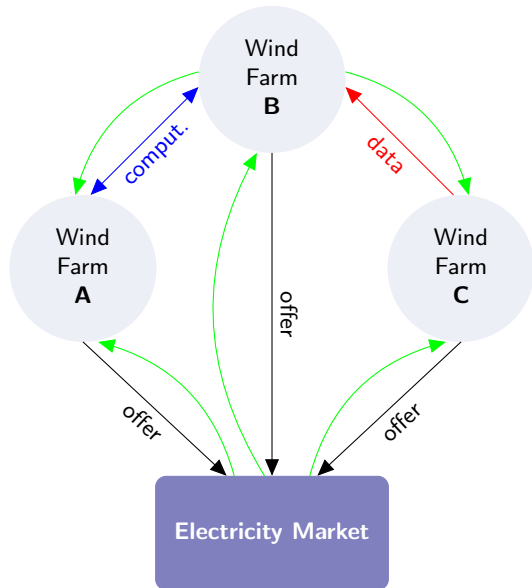
A motivating real-world example

Opportunity: All *could* benefit from some form of collaboration (e.g., information sharing)

Challenge: They have no interest in doing so



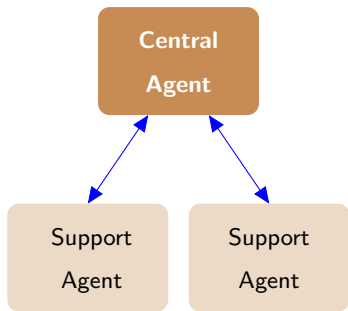
A motivating real-world example



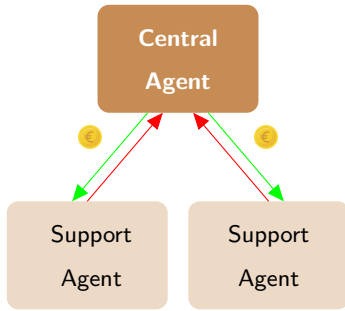
Proposal: Design a regression market framework allowing for all agents to collaborate and benefit from it

Agents meet through **analytics platforms** supporting collaborative and market-based analytics

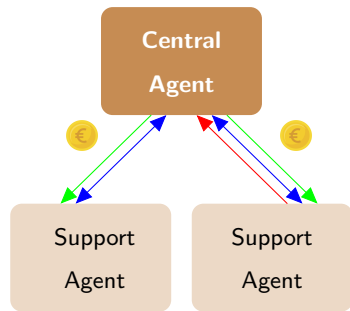
Collaborative Analytics:



Data Markets:



Analytics Markets:



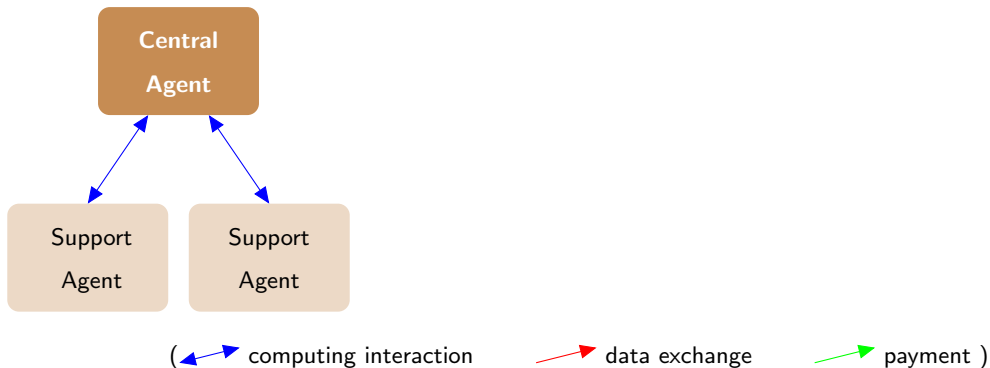
(computing interaction

data exchange

payment)

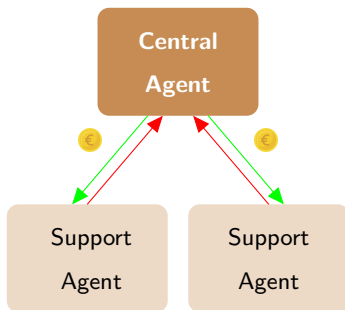
Agents meet through **analytics platforms** supporting collaborative and market-based analytics

Collaborative Analytics:



Agents meet through **analytics platforms** supporting collaborative and market-based analytics

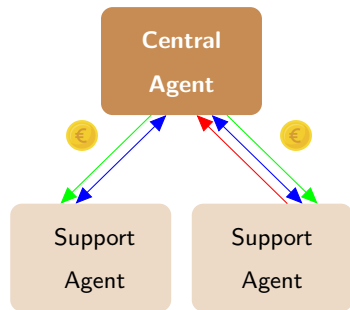
Data Markets:




( computing interaction  data exchange  payment)


Agents meet through **analytics platforms** supporting collaborative and market-based analytics

Analytics Markets:



( computing interaction

 data exchange

 payment)

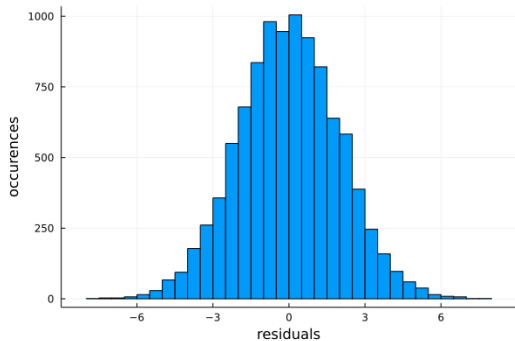
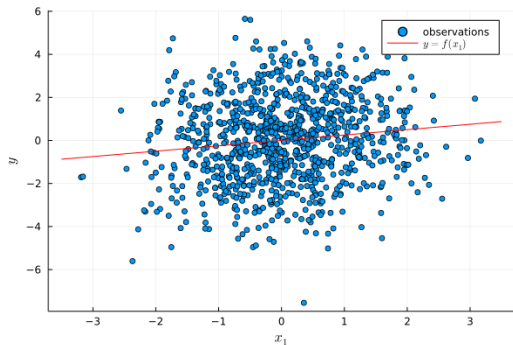
- **Data markets for regression and forecasting problems**

A toy model example to illustrate data markets for regression

- A *central agent* has a response variable y of interest, and one explanatory variable x_1 ,
- Two other agents have explanatory variables x_2 and x_3 ,
- It happens that the true data generation process (over $T = 10000$ time steps) is

$$y_t = \theta_1 x_{1,t} + \theta_2 x_{2,t} + \theta_3 x_{3,t} + \epsilon_t, \quad t = 1, \dots, T$$

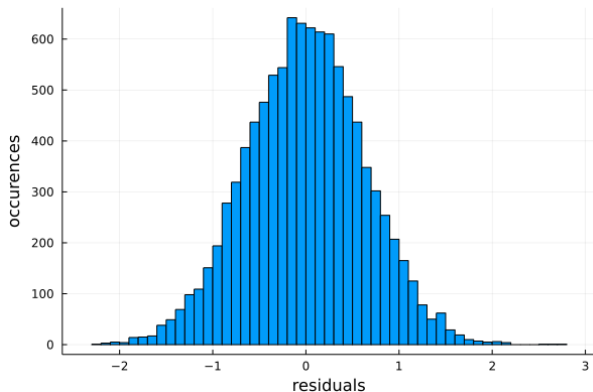
In the first stage, the central agent may only use her own data, and model y using x_1 only (with linear regression):



This yields an MSE (Mean Square Error) of 3.96...

Using data of others may help!

Then, the central agent gets access to the data of the other 2 agents, and model y using x_1 , x_2 and x_3 (with linear regression):



- The MSE is now of 0.4
- This means a decrease of MSE of $(3.96-0.4) = 3.56$ (or a 89.9% improvement)
- This is a substantial improvement for the central agent!

From a market point of view,

- how to define revenues and payments?
- what market properties do we need?
- what underlying techniques can be employed?

Formally: the central agent and the regression problem

- Consider a *central agent* (“**Forecaster**”) with a regression problem, e.g., as a basis to forecast renewable power generation for a given site (y_{t+k})
- **Forecaster** owns a set ω of m features, $\omega = \{x_1, \dots, x_m\}$

The following regression problem could be used as basis for eventual prediction,

$$Y_{t+k} = \beta_0 + \sum_{i=1}^m \beta_i x_{i,t} + \varepsilon_t, \quad t = 1, \dots, T$$

The vector of parameters $\beta = [\beta_0 \dots \beta_m]^\top$ can easily be learned by minimizing an appropriate loss function

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} S_\omega(\beta), \quad S_\omega(\beta) = \frac{1}{T} \sum_{t=1}^T \rho \left(y_{t+k} - \left(\beta_0 + \sum_{i=1}^m \beta_i x_{i,t} \right) \right)$$

where ρ may be any convex loss function (e.g., quadratic, pinball loss, etc.)

Based on the data available, the minimum loss function value is $S_\omega^* = S_\omega(\hat{\beta})$

- **Forecaster** could post the regression task on an analytics platform, to improve model fit
- **Forecaster** declares a willingness to pay of $\phi = 1\text{€}$ per percent-point improvement in S and per data point provided.

- Two *support agents* **Good Data** and **Useful Features** may bring in additional features z_1 and z_2 , to be remunerated
- The overall set of features now is $\Omega = \omega \cup \{z_1, z_2\}$

The regression problem can then be augmented, as

$$Y_{t+k} = \underbrace{\beta_0 + \sum_{i=1}^m \beta_i x_{i,t}}_{\text{Forecaster}} + \underbrace{\gamma_1 z_{1,t}}_{\text{Good Data}} + \underbrace{\gamma_2 z_{2,t}}_{\text{Useful Features}} + \varepsilon_t, \quad t = 1, \dots, T$$

where the augmented vector of coefficients $\beta^+ = [\beta_0 \dots \beta_m \gamma_1 \gamma_2]^\top$ can be learned similarly, by minimizing an appropriately chosen convex loss function ρ , i.e.,

$$\hat{\beta}^+ = \underset{\beta^+}{\operatorname{argmin}} S_\Omega(\beta^+), \quad S_\Omega(\beta^+) = \frac{1}{T} \sum_{t=1}^T \rho \left(y_{t+k} - \left(\beta_0 + \sum_{i=1}^m \beta_i x_{i,t} + \gamma_1 z_{1,t} + \gamma_2 z_{2,t} \right) \right)$$

We eventually write $S_\Omega^* = S_\Omega(\hat{\beta}^+)$

- If z_1 and/or z_2 are informative features, one expects $S_\Omega^* < S_\omega^*$

- How to define **revenues** and **payments** in such a regression market?

For each support agent j ($j = 1, 2$), the revenue is given by

$$\pi_j = (S_\omega^* - S_\Omega^*) T \phi \psi_j, \quad j = 1, 2$$

where ψ_j is an allocation policy based on feature valuation (can be obtained with, e.g., leave-one-out or Shapley-based allocation), such that $\sum_j \psi_j = 1$

For **Forecaster**, the payment is

$$\pi_c = \phi(S_\omega^* - S_\Omega^*) T$$

Such a simple approach actually yields a market with a wealth of good properties, i.e.,

- budget balance
- symmetry (or anonymity)
- zero element
- incentive compatibility
- individual rationality (truthfulness)

- Let us assume that the central agent is ready to pay 1€ per data point and per unit improvement in MSE
- We use a leave-one-out allocation policy

To apply a leave-one-out policy, we need to calculate the loss function (MSE), with and without the various explanatory variables, i.e.

features	MSE
$\{x_1\}$	3.96
$\{x_1, x_2\}$	3.66
$\{x_1, x_3\}$	0.71
$\{x_1, x_2, x_3\}$	0.4

$$\bullet \psi_2 = \frac{\text{MSE}_{\{x_1, x_2, x_3\}} - \text{MSE}_{\{x_1, x_3\}}}{\text{MSE}_{\{x_1, x_2, x_3\}} - \text{MSE}_{\{x_1\}}} = \frac{0.4 - 0.71}{0.4 - 3.96} = 0.09$$

$$\bullet \psi_3 = \frac{\text{MSE}_{\{x_1, x_2, x_3\}} - \text{MSE}_{\{x_1, x_2\}}}{\text{MSE}_{\{x_1, x_2, x_3\}} - \text{MSE}_{\{x_1\}}} = \frac{0.4 - 0.3.66}{0.4 - 3.96} = 0.91$$

We can then deduce the payment and revenues for all agents:

- The central agent should pay: $1 \times 10000 \times 3.56 = 35600\text{€}$
- The first support agent (for x_2) should receive: $1 \times 10000 \times 3.56 \times 0.09 = 3204\text{€}$
- The second support agent (for x_3) should receive: $1 \times 10000 \times 3.56 \times 0.91 = 32396\text{€}$

Typically,

- we learn in-sample (batch or online)
- we predict out-of-sample...

Can we use the above concepts more generally?

In-sample and out-of-sample

Typically,

- we learn in-sample (batch or online)
- we predict out-of-sample...

Can we use the above concepts more generally?

In **online regression** markets, the payments can be generalized with

$$\pi_{j,t} = (S_{\omega,t}^* - S_{\Omega,t}^*) \phi \psi_{j,t}, \quad j = 1, 2$$

where $S_{\omega,t}^*$ and $S_{\Omega,t}^*$ are time-varying estimator of the loss function, and $\psi_{j,t}$ is a time-varying estimate of allocation policies (profiting of their linearity property)

And, in **out-of-sample regression** markets (i.e., for genuine forecasting),

$$\pi_{j,t} = (s_{\omega,t}^* - s_{\Omega,t}^*) \phi \psi_{j,t}, \quad j = 1, 2$$

where $s_{\omega,t}^*$ and $s_{\Omega,t}^*$ are time-varying estimator of the loss function, and $\psi_{j,t}$ is the instantaneous allocation policies (i.e., readily Shapley additive explanation)

Batch, online, and out-of-sample regression markets all enjoy the same properties.

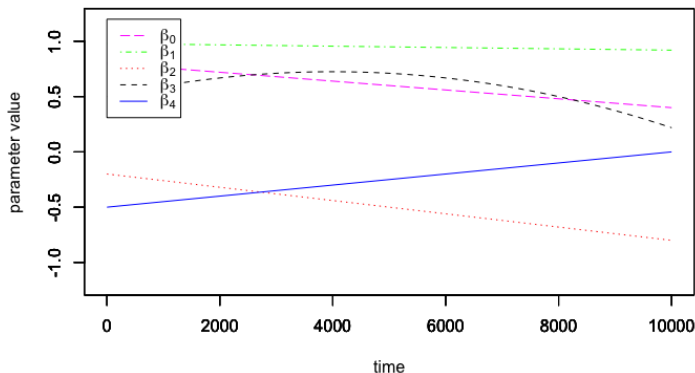
(given convex loss functions and models that are linear in their parameters)

5 Simulations and case-study application

Simulation-based example: RLS with an ARX model

True **data generation process** is $Y_t = \beta_0 + \beta_{1,t}y_{t-1} + \beta_{2,t}x_{2,t-1} + \beta_{3,t}x_{3,t-1} + \beta_{4,t}x_{4,t-1} + \varepsilon_t$

Time varying parameters:

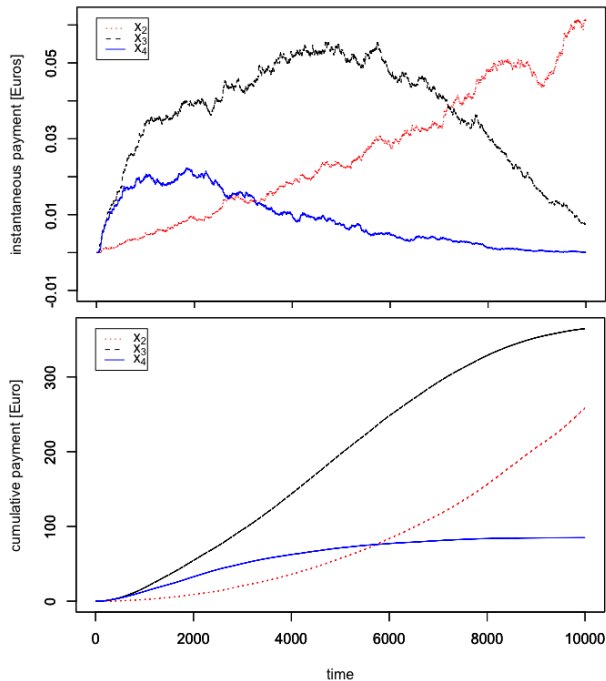


Central agent a_1 :

- focuses on target variable y and owns feature y_{t-1}
- willingness-to-pay $\phi = 0.1\text{€}$ per time instant and per unit improvement in a quadratic loss function

Support agents:

- a_2 owns feature x_2
- a_3 owns features x_3 and x_4

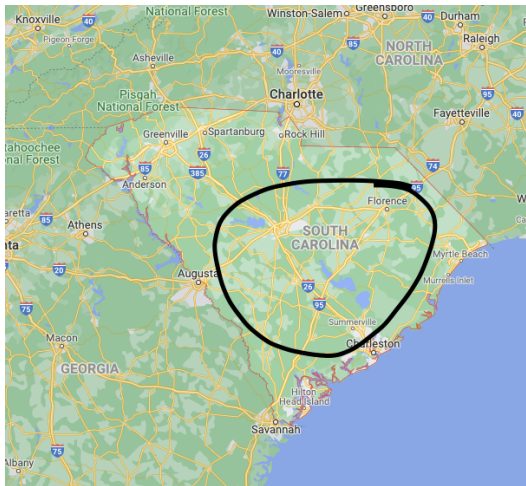


Focusing on both **instantaneous** (top) and **cumulative** (bottom) payments:

- payments are always positive (!?!)
- they vary as a function of the relative importance of the features
- that importance change with time
- monetary compensation piles up nicely in time...

A real world case study in South Carolina (USA)

Wind power generation for 9 locations in South Carolina (US) – 7 years of data with hourly resolution



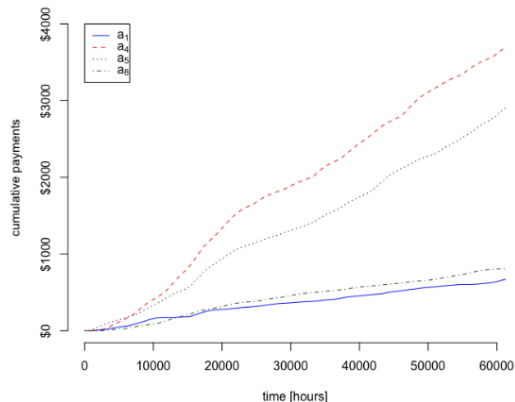
Agent	P_n [MW]	Lat./Long.	County
a_1	1.75	34.248/-79.75	Florence
a_2	2.96	34.02/-79.537	Florence
a_3	3.38	33.925/-79.958	Florence
a_4	16.11	34.732/-82.122	Laurens
a_5	37.98	34.556/-81.889	Laurens
a_6	30.06	34.334/-82.133	Laurens
a_7	2.53	33.136/-80.857	Colleton
a_7	2.6	33.112/-80.665	Colleton
a_9	1.24	32.641/-80.504	Colleton

1-hour ahead forecasting based on $Y_{i,t} = \beta_0 + \sum_{\delta=1}^{\Delta} y_{i,t-\delta} + \sum_{j \neq i} \sum_{\delta=1}^{\Delta} y_{j,t-\delta} + \varepsilon_{i,t}$
 where a_i is the central agent and a_j ($j \neq i$) are the support agents

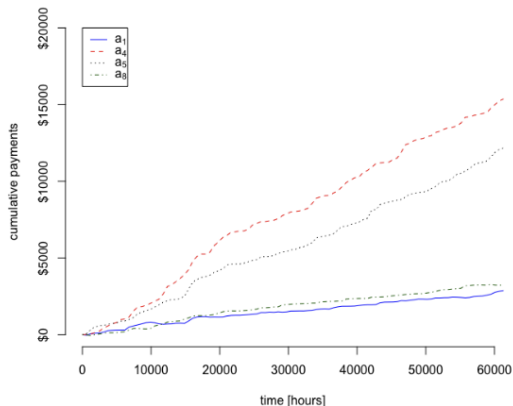
Online and out-of sample regression markets

- Online quantile regression ($\tau = 0.55$) in models with 2 lags for a_i and 1 lag for a_j ($j \neq i$)
- $\phi = 0.2$ in-sample, and $\phi = 0.8$ out-of-sample (per unit loss, per data point)

Cumulative payments of a_6 towards others:



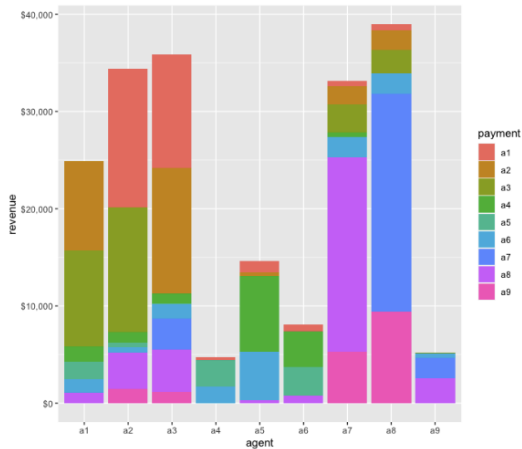
(a) Online regression market.



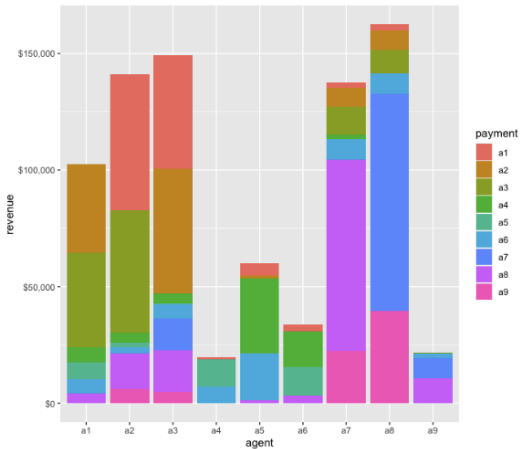
(b) Out-of-sample regression market.

Overall payments

Let's see what happens if they all pay each other for data to improve forecasts...



(a) Online regression market.



(b) Out-of-sample regression market.

On a per-data-point basis:

- a_4 only gets 0.39\$ per data point
- a_8 gets 3.26\$ per data point

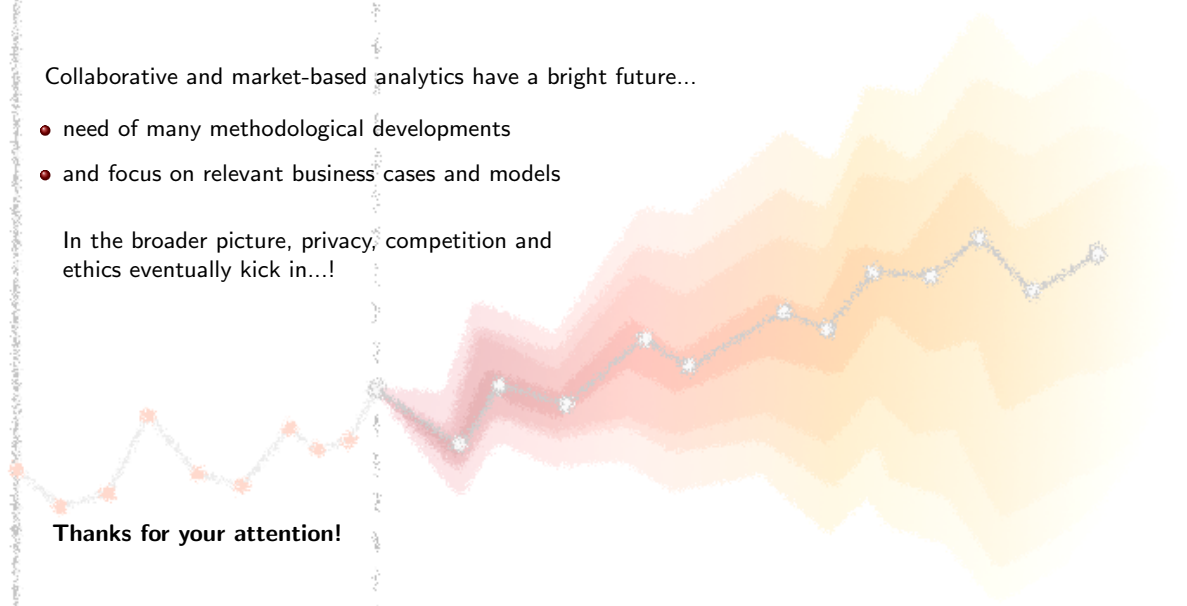
6 Concluding thoughts and discussion

Collaborative and market-based analytics have a bright future...

- need of many methodological developments
- and focus on relevant business cases and models

In the broader picture, privacy, competition and ethics eventually kick in...!

Thanks for your attention!



- 1 P. Pinson, L. Han, J. Kazempour (2022) Regression markets and application to energy forecasting. *TOP*, available online ([pdf](#))
- 2 L. Han, P. Pinson, J. Kazempour (2022) Trading data for wind power forecasting: A regression market with Lasso regularization. Proc. of the Power System Computation (PSCC) conference 2022, Porto, Portugal ([arxiv.org/pdf/2110.07432](#))
- 3 C. Goncalves, P. Pinson, R. Bessa (2022) Towards data markets in renewable energy forecasting. *IEEE Transactions on Sustainable Energy* **12**(1): 533-542 ([pdf](#))
- 4 A. M. Kharman, C. Jursitzky, Q. Zhao, P. Ferraro, J. Marecek, P. Pinson, R. Shorten (2022) On the design of decentralised data markets. Preprint, under review ([arxiv.org/abs/2206.06299](#))
- 5 S. R. Pandey, P. Pinson, P. Popovski (2022) Participation and data valuation in IoT data markets through distributed coalitions. Preprint, under review ([arxiv.org/abs/2206.07785](#))

... among many other papers appearing lately about data markets!

Hands-on session available at: [pdf link](#)

Solution available at: [pdf link](#) (in Julia)