DaSSWeb
Data Science and Statistics Webinar

3rd December 2024

Designing Incentives for Collaborative Forecasting

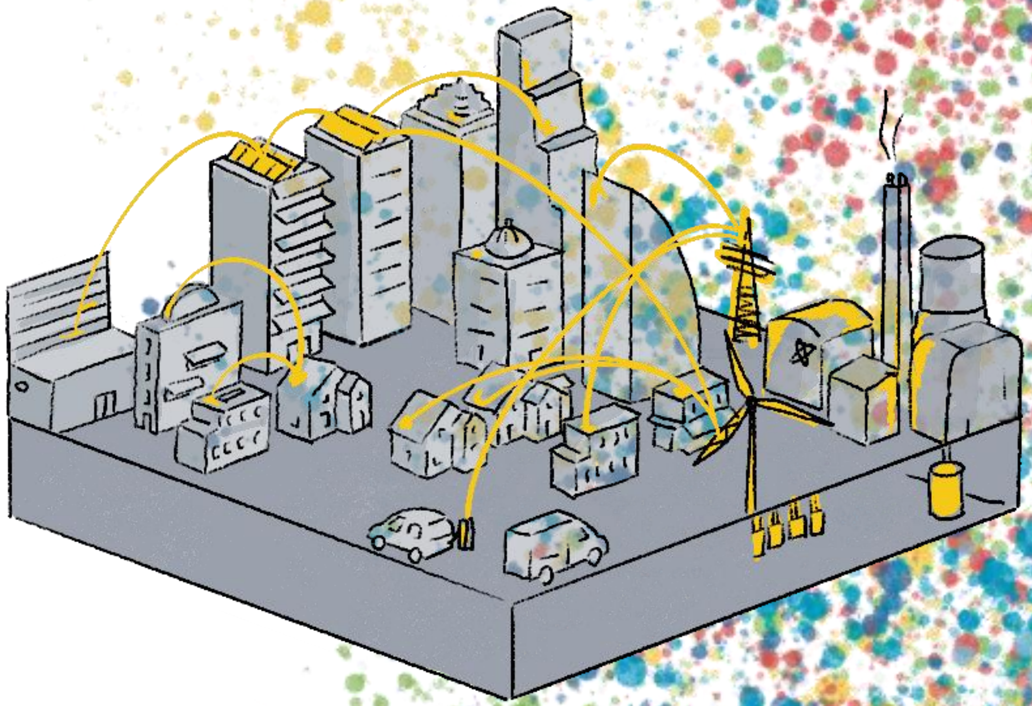
Carla Gonçalves, *carla.s.goncalves@inesctec.pt*

INESCTEC
INSTITUTE FOR SYSTEMS
AND COMPUTER ENGINEERING,
TECHNOLOGY AND SCIENCE

Context

Data Generation & Collection

Including in power systems …



Different types of data:

- Meteorological
- Network assets
- Generators
- Consumers
- …

Context

Data Generation & Collection

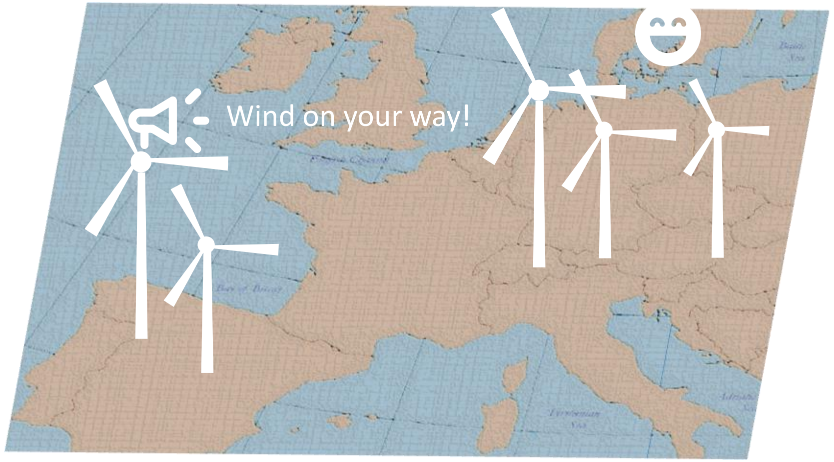
Collaborative Analytics

Combining data from multiple sources can increase the accuracy of forecasts

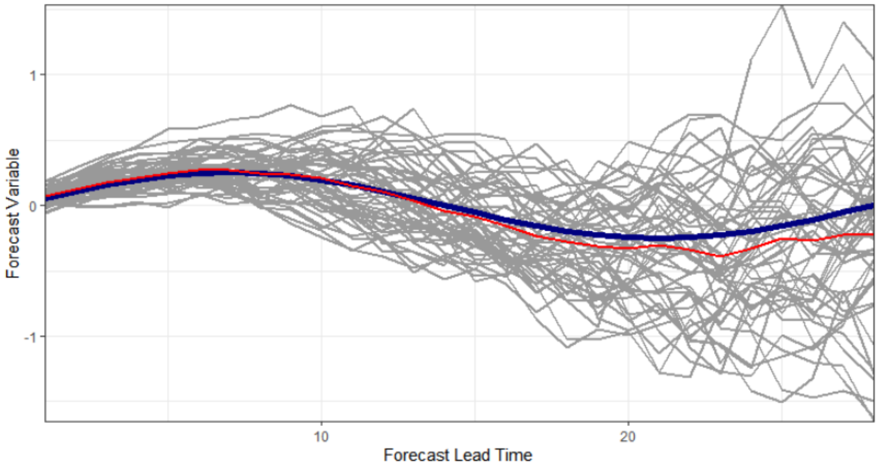
Open data sources

Proprietary data
(privacy, intellectual property, or regulatory reasons)

Spatio-temporal data
(helps models to capture e.g. weather patterns)

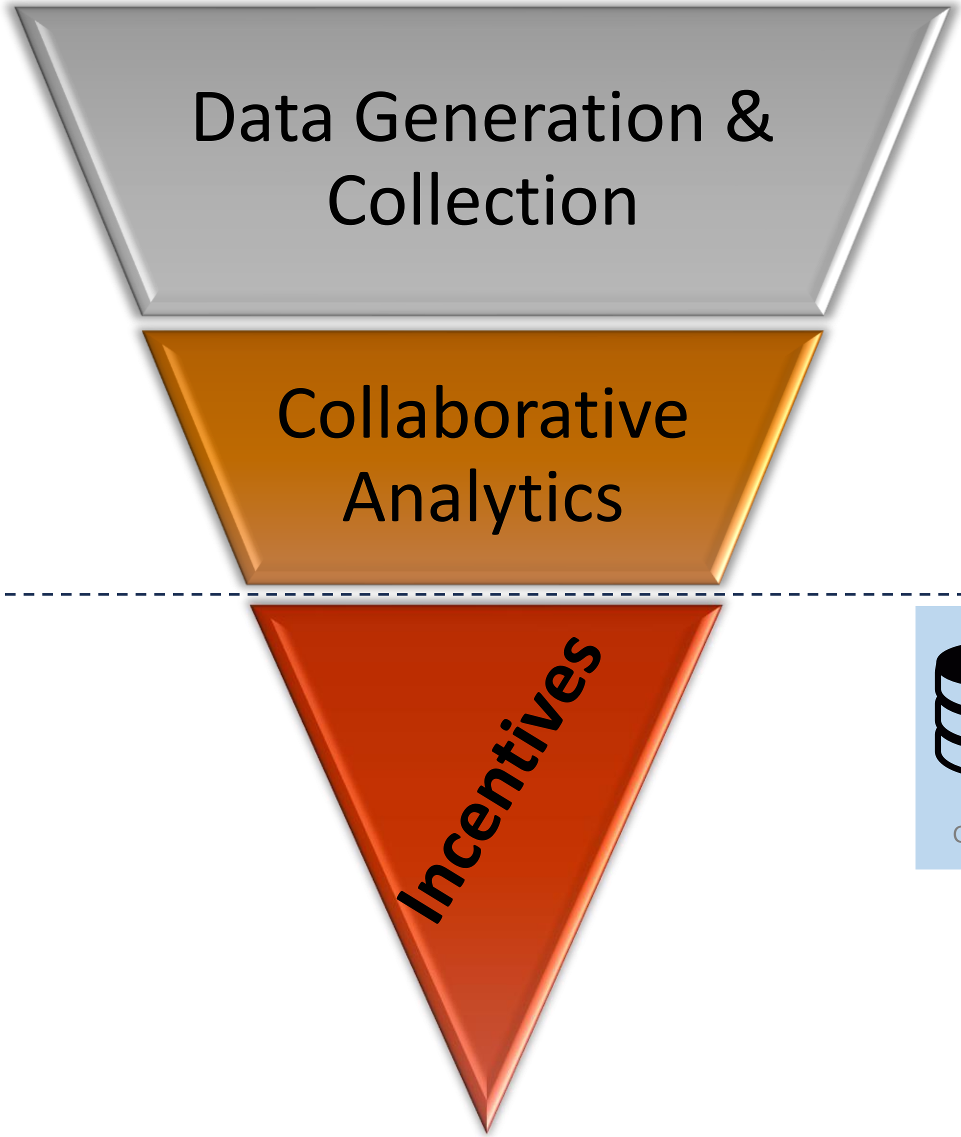


Multiple forecasts
(forecast ensemble are more stable)



Tastu et al. "Probabilistic forecasts of wind power generation accounting for geographically dispersed information." *IEEE Transactions on Smart Grid* 5.1 (2013): 480-489.

Context



Proprietary data
(privacy, intellectual property, or regulatory reasons)

Data privacy: collaborative analytics is performed without compromising data privacy

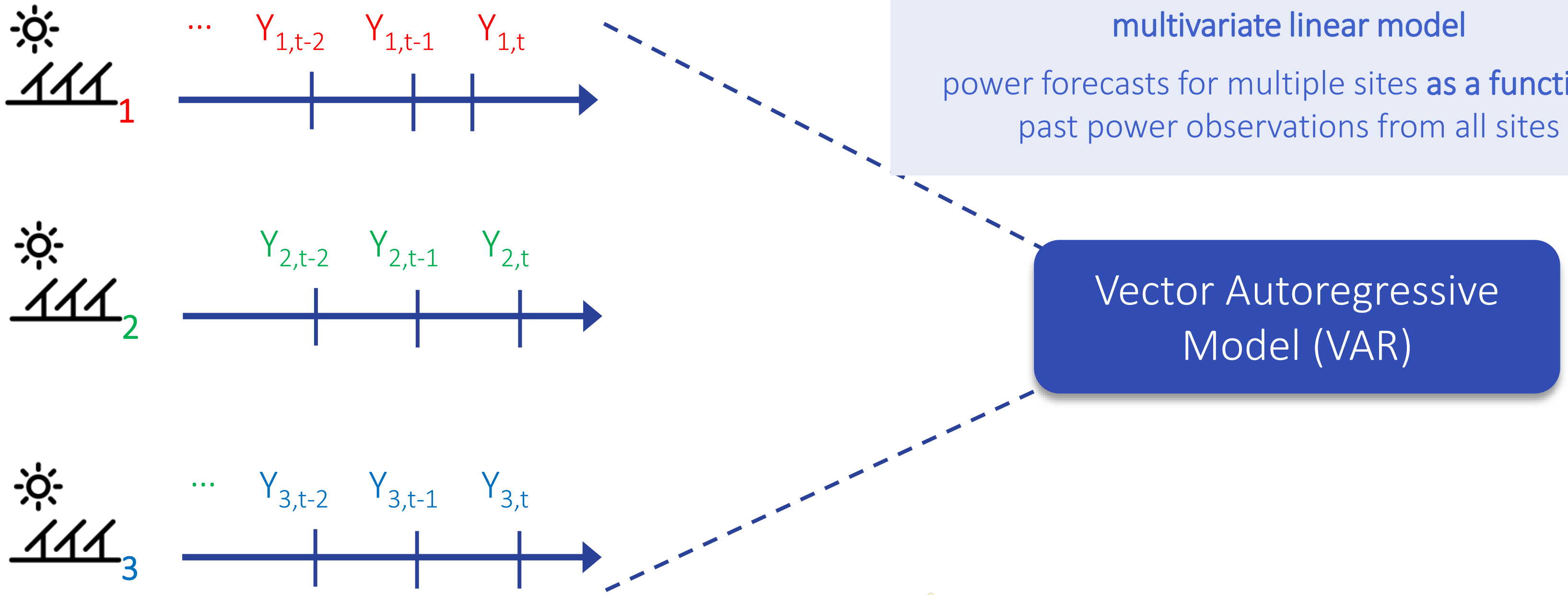
C. Gonçalves, R. J. Bessa, and P. Pinson. "Privacy-preserving distributed learning for renewable energy forecasting." *IEEE TSTE* 2021.

Data privacy: collaborative analytics is performed without compromising data privacy

C. Gonçalves, R. J. Bessa, and P. Pinson. "Privacy-preserving distributed learning for renewable energy forecasting." *IEEE TSTE* 2021.

multivariate linear model

power forecasts for multiple sites **as a function of** past power observations from all sites



Example: 3 PV sites, 1 lag

These connections are a problem!

$$Y_{1,t} = c_1 + B^1_{1,1} Y_{1,t-1} + \boxed{B^1_{2,1} Y_{2,t-1} + B^1_{3,1} Y_{3,t-1}} + E_{1,t}$$

$$Y_{2,t} = c_2 + \boxed{B^1_{1,2} Y_{1,t-1}} + B^1_{2,2} Y_{2,t-1} + \boxed{B^1_{3,2} Y_{3,t-1}} + E_{2,t}$$

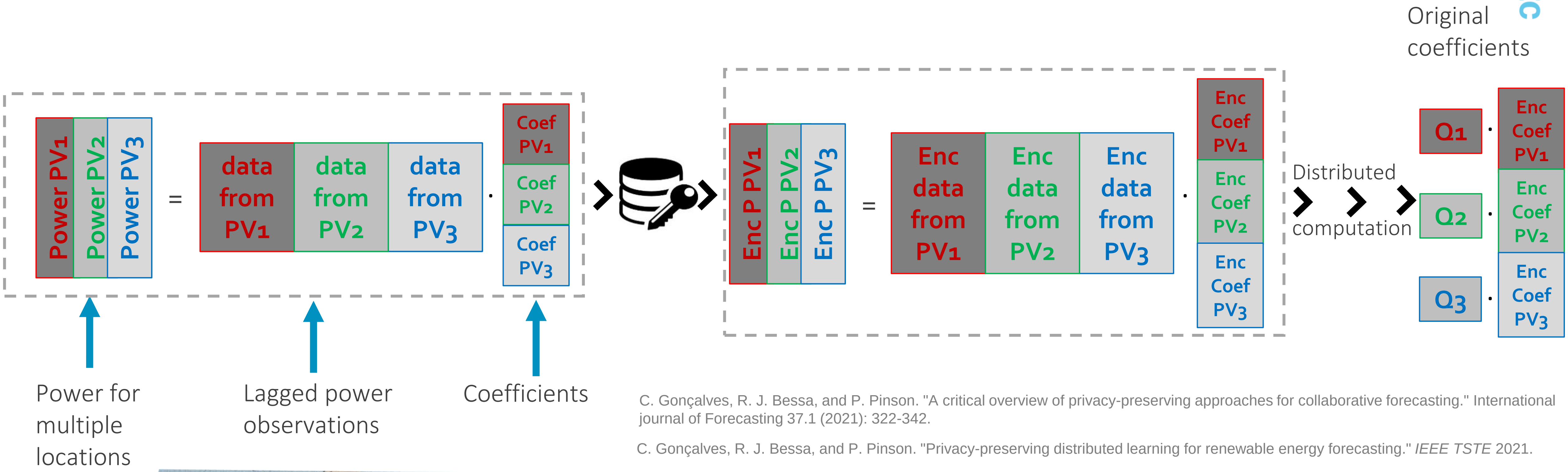
$$Y_{3,t} = c_1 + \boxed{B^1_{1,3} Y_{1,t-1} + B^1_{2,3} Y_{2,t-1}} + B^1_{3,3} Y_{3,t-1} + E_{3,t}$$

When compared against AR, VAR has reduced the average error by ~10% (h=1)

Data privacy: collaborative analytics is performed without compromising data privacy

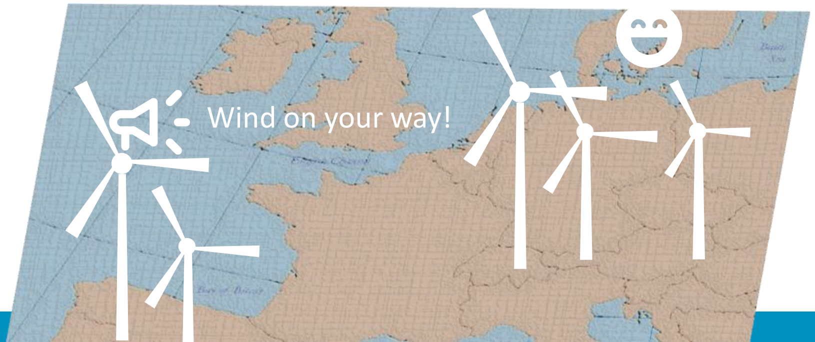
C. Gonçalves, R. J. Bessa, and P. Pinson. "Privacy-preserving distributed learning for renewable energy forecasting." *IEEE TSTE* 2021.

VAR MODEL

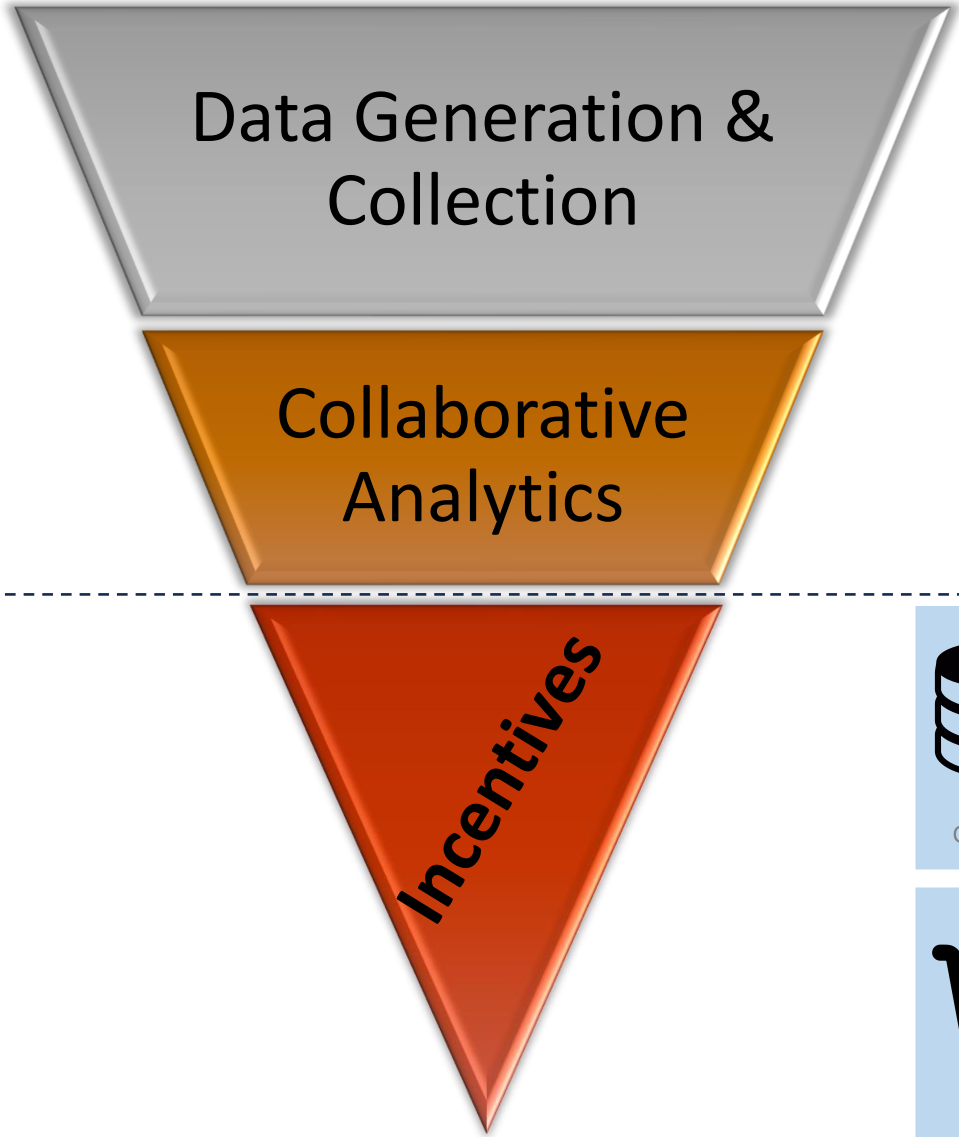


C. Gonçalves, R. J. Bessa, and P. Pinson. "A critical overview of privacy-preserving approaches for collaborative forecasting." International journal of Forecasting 37.1 (2021): 322-342.

C. Gonçalves, R. J. Bessa, and P. Pinson. "Privacy-preserving distributed learning for renewable energy forecasting." *IEEE TSTE* 2021.



Context



Proprietary data
(privacy, intellectual property, or regulatory reasons)

Data privacy: collaborative analytics is performed without compromising data privacy

C. Gonçalves, R. J. Bessa, and P. Pinson. "Privacy-preserving distributed learning for renewable energy forecasting." *IEEE TSTE* 2021.

Data markets: data privacy is a mandatory requirement but may not be enough

Data markets: data privacy is a mandatory requirement but may not be enough

Data owners want to improve their forecasts… or share data receive some compensation

data monetization

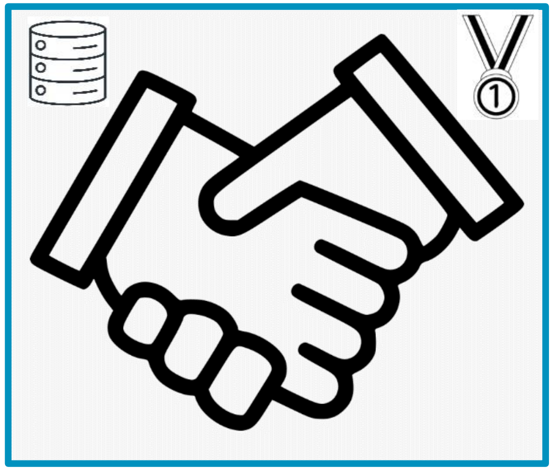
- ✓ **Data buyers: pay per accurate forecasts** with collaborative forecasting models
- ✓ **Data sellers: receive monetary compensation** proportional to the data importance when forecasting the others' data

barter trading (data by data)

- ✓ Data owners **provide and receive data** with approximately **the same value**
- ✓ **Value** is measured with metrics such as mutual information, correlation, etc.

data by service/recognition

- ✓ **Recognition** ,e.g., as a climate change mitigator
- ✓ Proportional to the data importance when forecasting the others' data



Simplest bidding strategy

Y, X4, P

I am willing to pay up to “P” for a prediction better than mine

Y = Const. + capturing nonlinearities

$$f1\_1(X1)*B1\_1 + \dots + f1\_M(X1)*B1\_M +$$
$$f2\_1(X2)*B2\_1 + \dots + f2\_M(X2)*B2\_M$$
$$f3\_1(X3)*B3\_1 + \dots + f3\_M(X3)*B3\_M$$
$$f4\_1(X4)*B4\_1 + \dots + f4\_M(X4)*B4\_M$$

subject to

$$p1*(1-I(B1\_1)*\dots*I(B1\_M)) +$$
$$p2*(1-I(B2\_1)*\dots*I(B2\_M)) +$$
$$p3\underbrace{*(1-I(B3\_1)*\dots*I(B3\_M)) \leq P}_{\text{check if at least one variable related to X3 is used}}$$

If my variable “X1” is used to generate forecasts, I want to receive “p1”

X1, p1

X2, p2

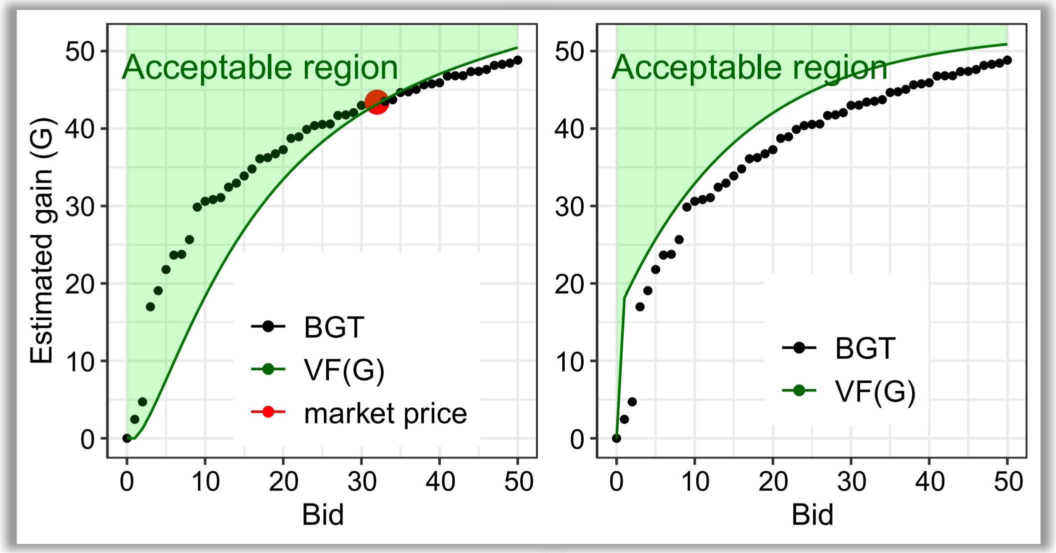
X3, p3

G. Yu, H. Fu, and Y. Liu. "High-dimensional cost-constrained regression via nonconvex optimization.“ *Technometrics* 64.1 (2022): 52-64.

Buyer bids based on accuracy improvement

I am willing to pay according to my accuracy improvement

Y, X4, P(gain)



If my variable "X1" is used to generate forecasts, I want to receive "p1"

X1, p1

X2, p2

X3, p3

C. Gonçalves, R. J. Bessa, T. Teixeira, J. Vinagre. "Budget-constrained Collaborative Renewable Energy Forecasting Market." under review IEEE TSTE (3rd round)

G. Yu, H. Fu, and Y. Liu. "High-dimensional cost-constrained regression via nonconvex optimization." *Technometrics* 64.1 (2022): 52-64.

INESCTEC

Case Study | Synthetic datasets

Y = f(X3, X7, X12, X21, X31, X48, X51, X63, X37, X90)

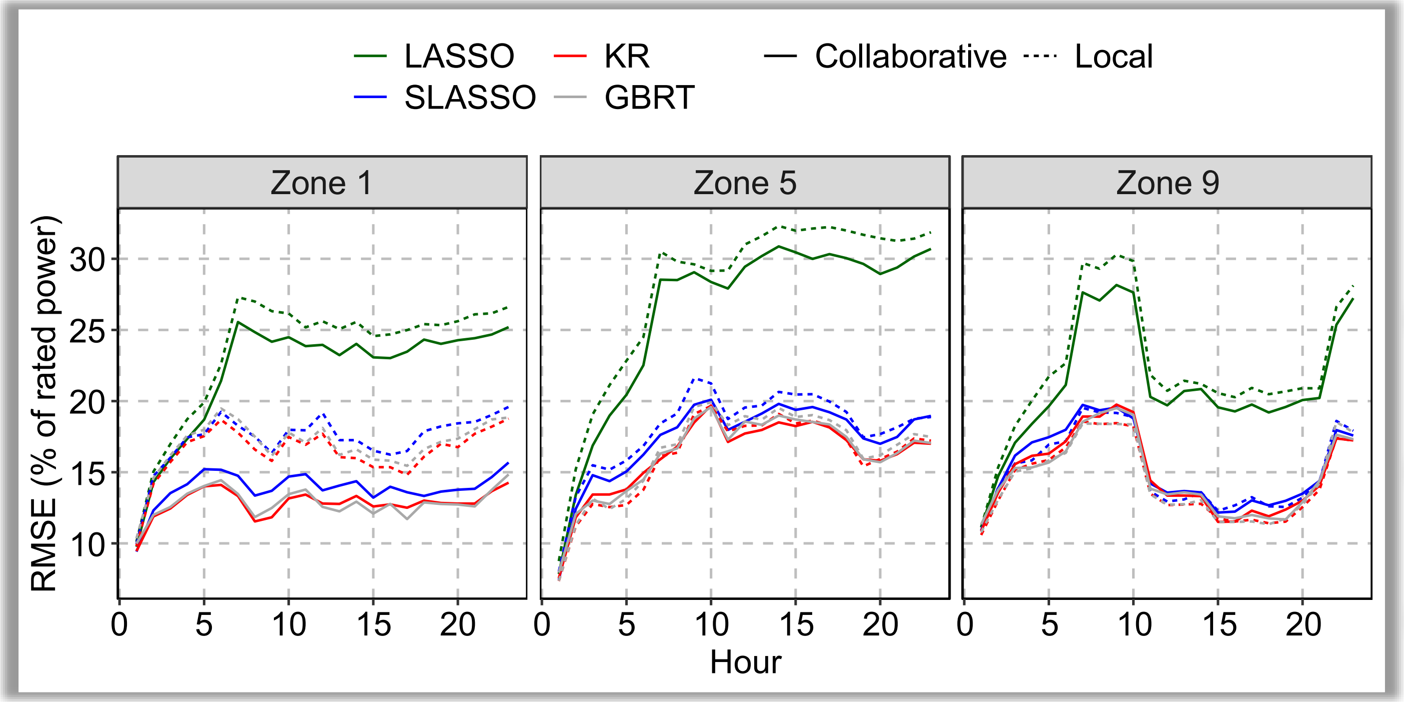
- **Data buyer** with **one target** and **ten covariates** (X1, X2, …, X10)
- **90 features available** on data market (X11, …, X100)
- X37 costs 11 cents, all the others cost 10
- X3 similar to X73 (will the market select X73?)
- X37 similar to X74 (will the market select the cheapest option?)

TABLE I: Data allocation and relevance (synthetic datasets).

Linear			**Non-linear**		
Optimal	**Case #1**	**Case #2**	**Optimal**	**Case #1**	**Case #2**
21 (22.73)	21 (26.33)	21 (22.86)	21 (15.45)	21 (28.50)	21 (25.30)
90 (20.45)	90 (23.88)	90 (20.69)	90 (14.70)	90 (22.60)	90 (20.01)
63 (18.18)	63 (20.86)	63 (18.07)	63 (13.98)	63 (20.52)	63 (17.92)
48 (15.91)	48 (18.25)	48 (15.81)	48 (13.30)	48 (18.67)	48 (17.18)
37 (9.09)	74 (10.68)	74 (8.95)	37 (11.45)	74 (9.72)	74 (8.39)
51 (6.81)	–	51 (6.78)	51 (10.89)	–	51 (7.32)
12 (4.55)	–	12 (4.51)	12 (10.36)	–	12 (3.88)
31 (2.27)	–	31 (2.33)	31 (9.85)	–	–
others (0)	–	–	others (0)	–	–

Case Study | Global Energy Forecasting Competition 2014 (GEFCom2014)

- **Normalized hourly wind power** measurements for 10 zones
- 24-hour-ahead **forecasts for zonal and meridional wind components** at 10m and 100m above ground level, issued daily at 00h00 to production locations.



Train + forecast times:

- KR 801 seconds
- GBRT 414 seconds
- Spline LASSO 93 seconds

! collaborative models outperfom local models

! improvement significantly varies between zones

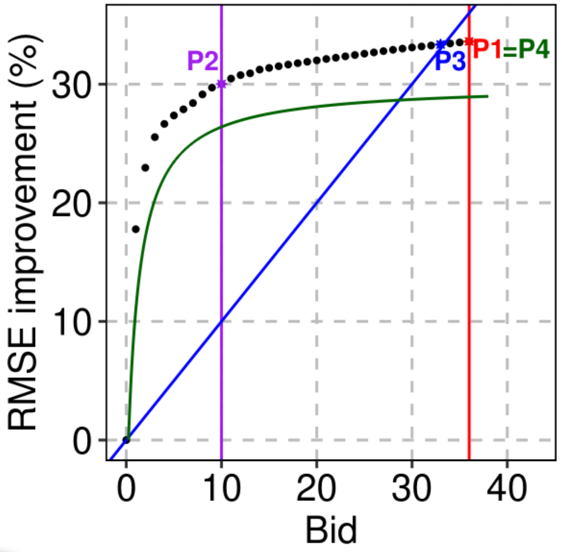
! these are mean values for the year 2013!

Case Study | Global Energy Forecasting Competition 2014 (GEFCom2014)

- **Normalized hourly wind power** measurements for 10 zones
- 24-hour-ahead **forecasts for zonal and meridional wind components** at 10m and 100m above ground level, issued daily at 00h00 to production locations.

! Higher buyer budgets lead to better performance

! Data owners improve accuracy and/or receive monetary compensations



! Sellers' bid are 1 per X

	SLCM											
	Payment (↑)				Revenue (↓)				Mean gain ($\mathcal{G}$(%))			
i	$\mathcal{VF}^1$	$\mathcal{VF}^2$	$\mathcal{VF}^3$	$\mathcal{VF}^4$	$\mathcal{VF}^1$	$\mathcal{VF}^2$	$\mathcal{VF}^3$	$\mathcal{VF}^4$	$\mathcal{VF}^1$	$\mathcal{VF}^2$	$\mathcal{VF}^3$	$\mathcal{VF}^4$
1	157941	22613	137907	60296	103211	17487	97387	44119	25.12	24.73	24.97	29.17
2	142742	29807	132772	59504	108121	22791	102774	47721	15.42	16.44	15.49	20.61
3	93827	14841	95458	33192	113268	21878	106349	48788	-1.87	-3.83	-1.86	-0.12
4	102468	22132	88939	47110	107611	10002	102073	43326	20.42	21.17	20.41	27.66
5	90894	14798	75095	44988	111547	18793	106193	46274	25.10	25.80	25.10	35.91
6	109366	18966	110560	42373	107990	18288	100552	44986	19.41	19.83	19.43	29.22
7	71848	12380	72478	38203	111038	21017	102800	44847	9.62	9.20	9.61	15.14
8	56730	10185	55863	36400	114681	28750	107278	47615	4.59	3.14	4.51	8.92
9	116838	25235	106607	48752	109046	19486	103044	46712	10.52	11.88	10.55	16.22
10	150819	32551	152278	52924	106960	25016	99507	46560	11.39	11.62	11.41	18.55

Next Generation Reservoir Computing

Daniel J. Gauthier,[1,2,*] Erik Bollt,[3,4] Aaron Griffith,[1] and Wendson A.S. Barbosa[1]

[1]The Ohio State University, Department of Physics, 191 West Woodruff Ave., Columbus, OH 43210, USA.

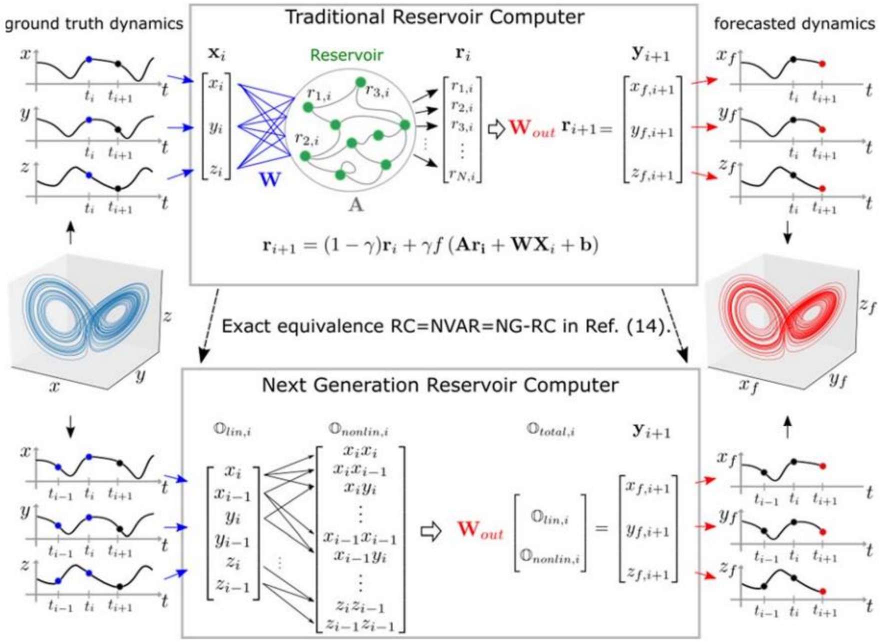
[2]ResCon Technologies, LLC, PO Box 21229, Columbus, OH 43221, USA

[3]Clarkson University, Department of Electrical and Computer Engineering, Potsdam, NY 13669

[4]Clarkson Center for Complex Systems Science (C^3S^2), Potsdam, NY 13699, USA

*Correspondence to: gauthier.51@osu.edu

July 21, 2021



- **Mixed effects:** product between lags

"a linear combination of mixed effects among (shifted) time series can serve as a powerful model(…), effectively capturing dynamic system data through observed time series."

(quadratic knapsack problems if we consider two times)

- **Classification:** logistic regression

Data markets: data privacy is a mandatory requirement but may not be enough

Data owners want to improve their forecasts… or share data receive some compensation

data monetization

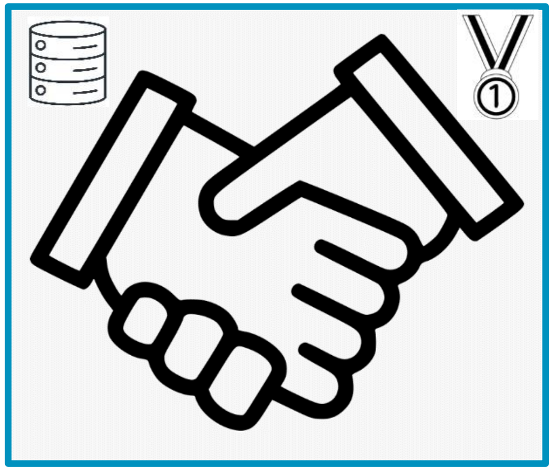
- ✓ **Data buyers: pay per accurate forecasts** with collaborative forecasting models
- ✓ **Data sellers: receive monetary compensation** proportional to the data importance when forecasting the others' data

barter trading (data by data)

- ✓ Data owners **provide and receive data** with approximately **the same value**
- ✓ **Value** is measured with metrics such as mutual information, correlation, etc.

data by service/recognition

- ✓ **Recognition** ,e.g., as a climate change mitigator
- ✓ Proportional to the data importance when forecasting the others' data



barter trading (data by data)

I am willing to share and receive data with similar value

Y4, X4_1, X4_2, X4_3

$$\arg\max_{z^j_{j'\to i}\in\{0,1\}} \underbrace{\sum_i \sum_{j\neq i} \sum_{j'\in\Omega_j} \mathrm{V}^j_{j'\to i} z^j_{j'\to i}}_{\text{total value exchange}}$$

subject to

$$\Big|\underbrace{\sum_{i'=1}^{n_i}\sum_j \mathrm{V}^i_{i'\to j} z^i_{i'\to j}}_{i\text{-th data owner shared value}} - \underbrace{\sum_k\sum_{k'=1}^{n_k} \mathrm{V}^k_{k'\to i} z^k_{k'\to i}}_{i\text{-th data owner received value}}\Big| \le \epsilon, \quad \forall i$$

$$\underbrace{\sum_{\{m',m\}\in\mathcal{C}_c} z^m_{m'\to i} \le 1, \forall i, c}_{\text{condition to capture redundancy}}$$

Linear Problem

$V^j_{j'\to i}$: value of Xj_j' when forecasting Y_i, conditional on Xi_1, Xi_2, Xi_3

maximize value allocation

subject to |shared-received value|<eps
no redundant data allocated

Y1, X1_1, X1_2, X1_3

Y2, X2_1, X2_2, X2_3

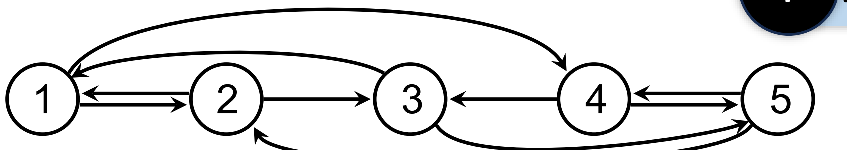
Y3, X3_1, X3_2, X3_3

barter trading (data by data)

Case Study | Synthetic datasets

… evaluating metrics

! Each agent has 30 covariates



$\Gamma_1 = \{2,3\}, \Gamma_2 = \{1,5\}, \Gamma_3 = \{2,4\}, \Gamma_4 = \{1,5\}, \Gamma_5 = \{3,4\}$

$$60X_i^3 + 40X_i^7 + \sum_{j\in\Gamma_i}\left[16.1X_j^{11} + 6.5X_j^{13} + 19.4X_j^{19} + 25.8X_j^{24} + 32.3X_j^{27}\right] + \varepsilon$$

	linear			exponential			quadratic (no interactions)		
	recall	precision	time (s)	recall	precision	time (s)	recall	precision	time (s)
Pearson	1.00(±0.00)	0.64(±0.06)	5.13	1.00(±0.00)	0.59(±0.09)	5.12	0.12(±0.14)	0.16(±0.18)	5.06
Spearman	1.00(±0.00)	0.63(±0.08)	16.59	1.00(±0.00)	0.64(±0.10)	16.19	0.07(±0.07)	0.15(±0.17)	16.54
Kendall	1.00(±0.00)	0.63(±0.08)	9.59	1.00(±0.00)	0.64(±0.09)	9.39	0.05(±0.08)	0.11(±0.17)	9.65
MI	0.95(±0.04)	0.69(±0.11)	3.29	0.80(±0.09)	0.69(±0.07)	3.48	0.98(±0.04)	0.64(±0.08)	3.54
ϕ_k	0.90(±0.07)	0.68(±0.08)	1665.08	0.58(±0.14)	0.48(±0.10)	1658.37	0.90(±0.04)	0.58(±0.05)	1667.59
dcor	1.00(±0.00)	0.65(±0.03)	64.95	1.00(±0.00)	0.65(±0.08)	64.02	0.98(±0.04)	0.66(±0.13)	65.24
Partial Pearson	1.00(±0.00)	0.57(±0.05)	83.49	1.00(±0.00)	0.58(±0.09)	83.64	0.07(±0.07)	0.13(±0.13)	83.18
Partial Spearman	1.00(±0.00)	0.58(±0.06)	422.48	1.00(±0.00)	0.60(±0.05)	422.05	0.03(±0.08)	0.07(±0.15)	421.78
CMI	0.90(±0.04)	0.81(±0.08)	35.96	0.75(±0.06)	0.96(±0.08)	39.36	0.98(±0.04)	0.66(±0.11)	37.60
MRMR Pearson	1.00(±0.00)	0.62(±0.07)	17.38	1.00(±0.00)	0.71(±0.18)	17.24	0.10(±0.11)	0.14(±0.16)	17.80
MRMR Spearman	1.00(±0.00)	0.60(±0.09)	141.44	1.00(±0.00)	0.65(±0.06)	141.35	0.05(±0.08)	0.10(±0.17)	141.66
MRMR Kendall	1.00(±0.00)	0.60(±0.09)	839.84	1.00(±0.00)	0.65(±0.06)	842.25	0.05(±0.08)	0.10(±0.17)	841.30
MRMR MI	1.00(±0.00)	0.69(±0.16)	104.99	0.80(±0.04)	0.66(±0.11)	105.72	1.00(±0.00)	0.65(±0.06)	105.55
MRMR ϕ_k	0.95(±0.04)	0.67(±0.15)	2550.40	0.48(±0.15)	0.50(±0.18)	2571.69	0.93(±0.07)	0.61(±0.05)	2562.87
MRMR dcor	1.00(±0.00)	0.58(±0.01)	1606.65	1.00(±0.00)	0.62(±0.11)	1601.20	0.98(±0.04)	0.67(±0.07)	1604.79
PermImp (GBR)	1.00(±0.00)	0.72(±0.04)	3288.71	0.90(±0.07)	1.00(±0.00)	1812.93	1.00(±0.00)	0.63(±0.12)	2122.35
Impurity (GBR)	0.83(±0.00)	1.00(±0.00)	1956.00	0.70(±0.04)	0.98(±0.05)	1149.63	0.83(±0.00)	1.00(±0.00)	1404.10
SHAP (GBR)	1.00(±0.00)	0.10(±0.00)	2020.25	1.00(±0.00)	0.12(±0.02)	1152.52	1.00(±0.00)	0.10(±0.00)	1409.82

Non-conditional

Conditional to owned data

! Permutation hypothesis test (if p-value <5% feature is relevant)

! Recall = TP / (TP + FN)

! Precision = TP / (TP + FP)

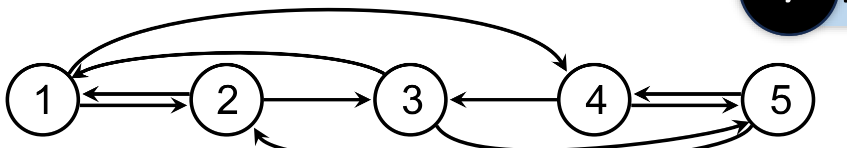
! MRMR MI has the best trade-off between the 3 criteria

barter trading (data by data)

Case Study | Synthetic datasets

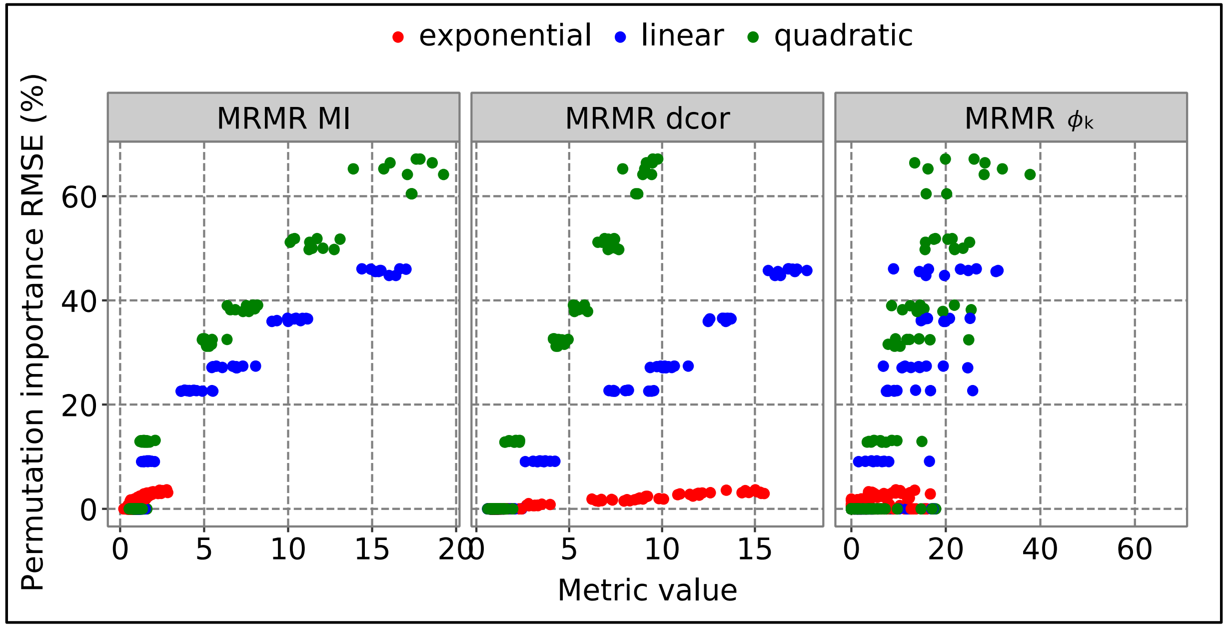
… relation with data value?

! Each agent has 30 covariates



$\Gamma_1 = \{2,3\}, \Gamma_2 = \{1,5\}, \Gamma_3 = \{2,4\}, \Gamma_4 = \{1,5\}, \Gamma_5 = \{3,4\}$

$$60X_i^3 + 40X_i^7 + \sum_{j\in\Gamma_i}\left[16.1X_j^{11} + 6.5X_j^{13} + 19.4X_j^{19} + 25.8X_j^{24} + 32.3X_j^{27}\right] + \varepsilon$$



		Owner 1	Owner 2	Owner 3	Owner 4	Owner 5
Linear	Correctly exchanged	9	10	9	10	9
	Wrongly exchanged	2	4	3	2	8
Exponential	Correctly exchanged	8	6	9	7	6
	Wrongly exchanged	4	3	2	4	6
Quadratic	Correctly exchanged	10	10	9	7	10
	Wrongly exchanged	2	5	7	5	1

! MRMR MI has the best trade-off between the 3 criteria

Case Study | Global Energy Forecasting Competition 2014 (GEFCom2014)

RMSE for January 2013

Zone	Local	Proposal	Collaborative
1	19.72	14.58	13.15
2	27.57	26.53	25.97
3	34.37	35.06	34.96
4	27.15	27.09	27.00
5	31.26	31.06	31.07
6	32.38	32.46	31.44
7	14.22	14.11	14.4
8	14.49	14.66	14.77
9	20.37	19.94	20.03
10	43.48	42.92	42.09

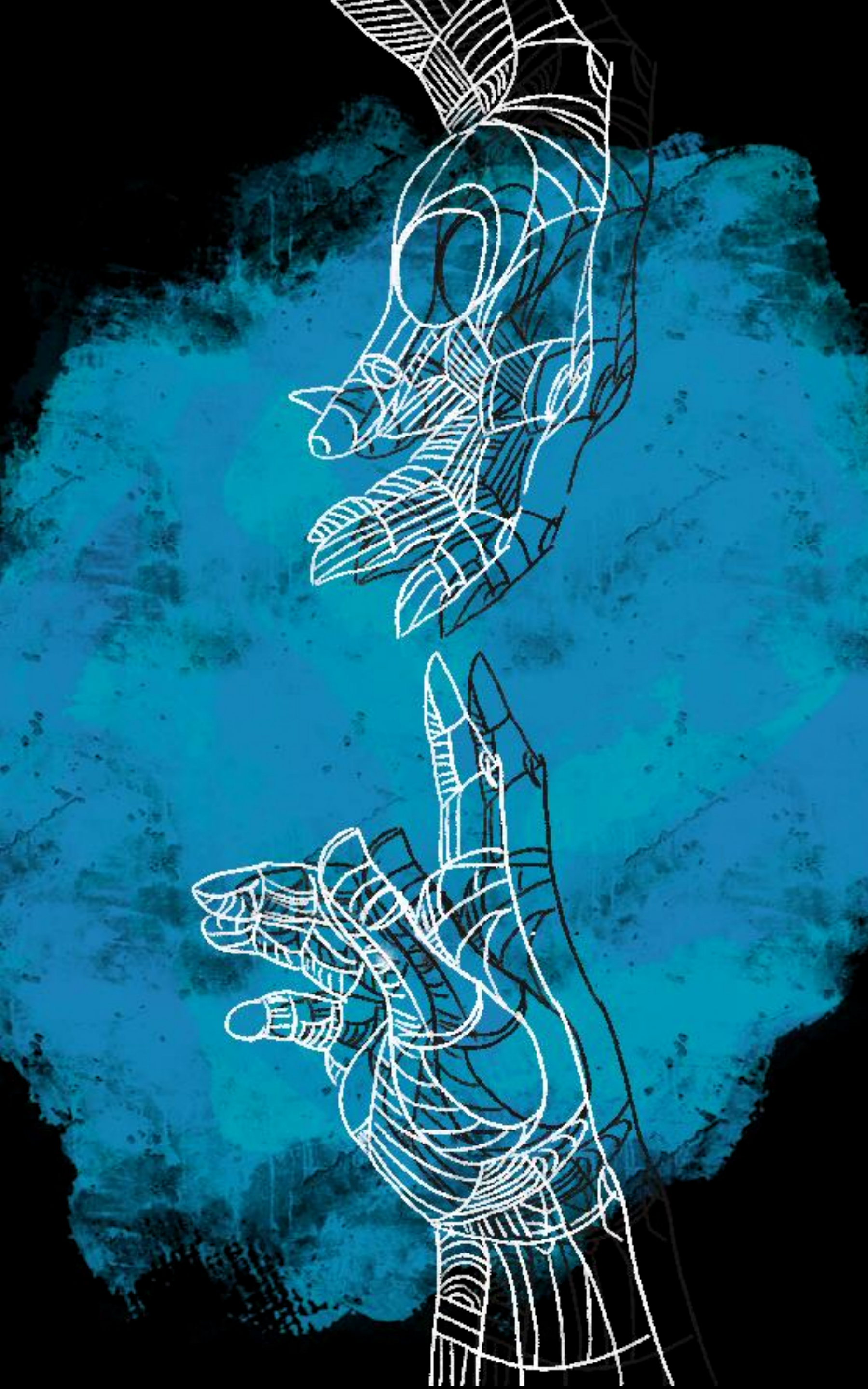
! Zone 1 has higher rmse improvement → shared information with all other zones!

! Zone 2 has the second higher rmse improvement and the second higher shared value

! Remaining zones shared lower value

Key messages

- **Collaborative forecasting** has potential to improve forecasting accuracy in many use cases
- **Data sharing incentives** are needed to promote such collaboration
- Incentives explored by our team:
    - Data monetization
    - Barter trading
- Projects



José Andrade
Researcher
CPES @ INESC TEC

André Garcia
Researcher
CPES @ INESC TEC

Giovanni Buroni
Researcher
CPES @ INESC TEC

Carla Gonçalves
Researcher
CPES @ INESC TEC

Ricardo Bessa
Coordinator & Researcher
CPES @ INESC TEC

Fernando Paula
Researcher
CPES @ INESC TEC

Thank you!

Rua Dr. Roberto Frias

4200-465 Porto

Portugal

INESCTEC

T +351 222 094 000
info@inesctec.pt
www.inesctec.pt