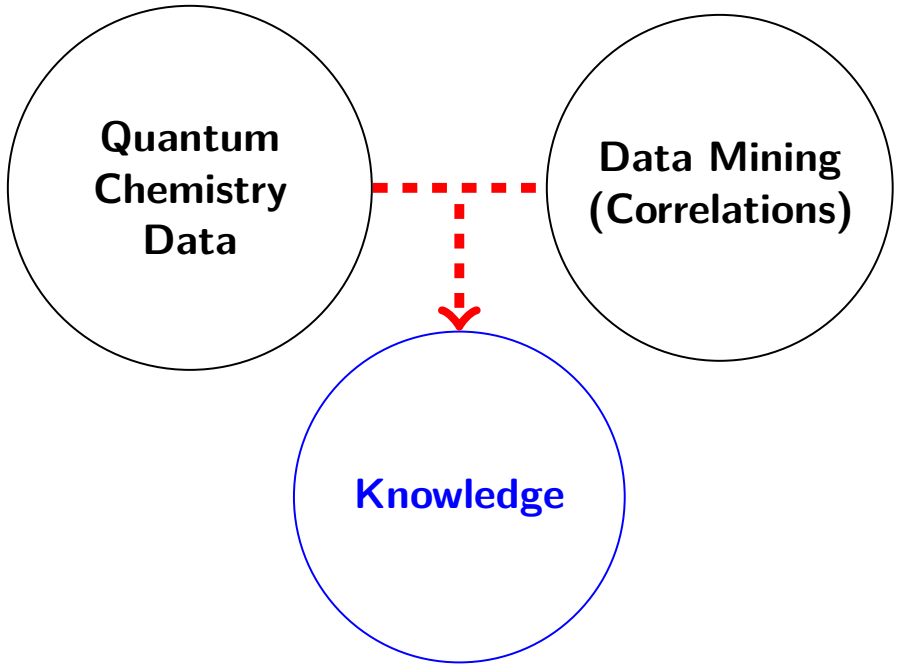
IV CINE-CMSC Workshop

A Correlation and Feature Engineering Framework to Obtain Insights from Quantum Chemistry Datasets

Johnatan Mucelini

Advisor:
Juarez L. F. Da Silva

Collaborators:
Paulo de Carvalho Dias Mendes,
Priscilla Felício-Sousa,
Karla F. Andriani,
Marcos G. Quiles,
Ronaldo C. Prati.



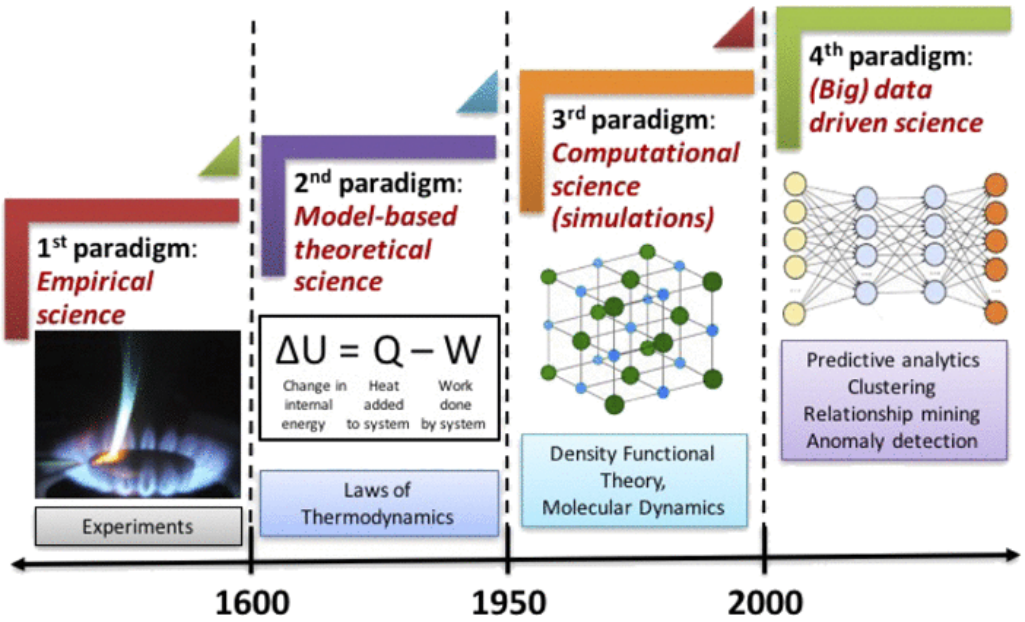
February, 2020

Outline

- Motivation
- Correlation Analysis
  - Pearson, Spearman, and Kendal Correlations
  - Trustfull
  - Applications in QC data
- Feature Engineering
  - Featurization - Mining Features Strategy
  - Results Visualization
  - Practical Tips
  - Implementation - Quandarium
- Applications
  - Materials
  - Workflow
  - Results
- Conclusions

Motivation

Scenario:



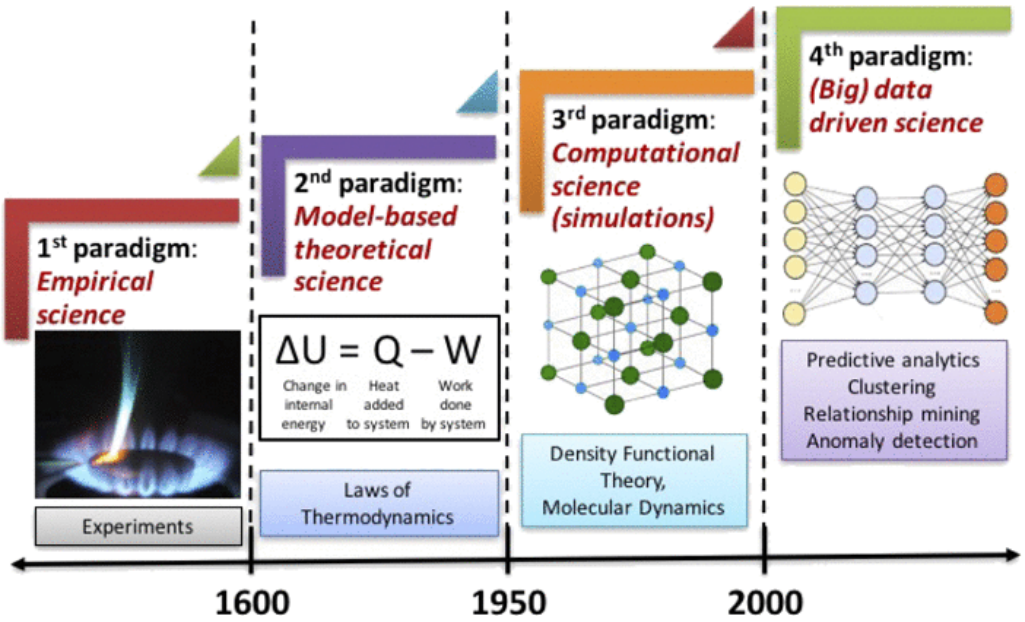
Font: *APL Materials*, **2016**, 4, 053208

Problems in QC Studies:

- Insights are obtained using few calculations;
- Visual description of the atomic systems;
- Trends are visually identified.

Motivation

Scenario:



Font: *APL Materials*, **2016**, 4, 053208

Problems in QC Studies:

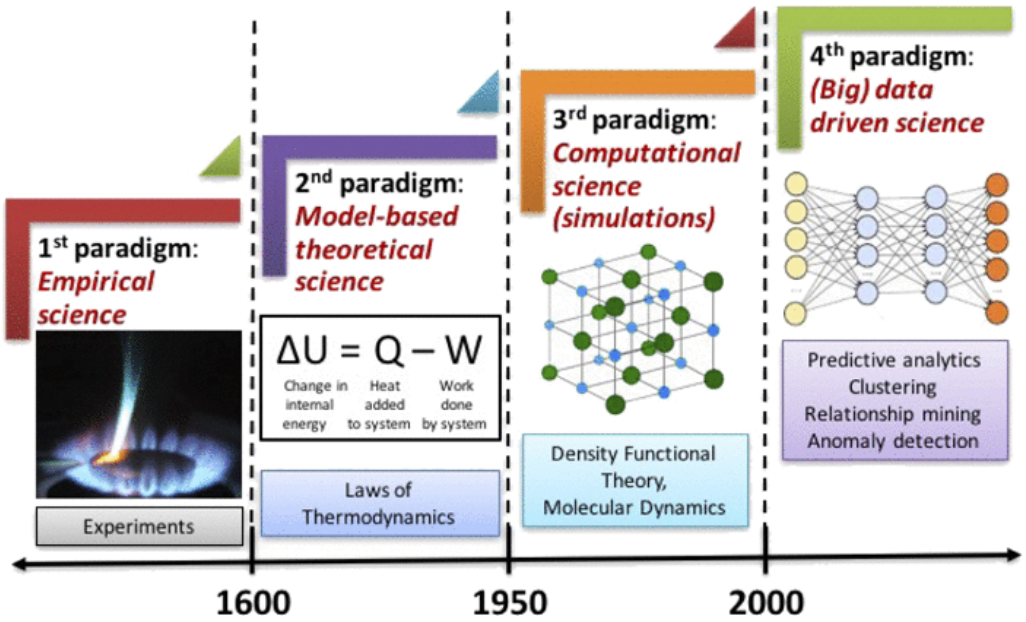
- Insights are obtained using few calculations;
- Visual description of the atomic systems;
- Trends are visually identified.

Proposed **Solution**: to empoloy Correlation Analysis...

- Large amount of data can be used;
- Methodological analysis of trends;
- Easy to calculate and implement.

Motivation

Scenario:



Font: *APL Materials*, **2016**, 4, 053208

Problems in QC Studies:

- Insights are obtained using few calculations;
- Visual description of the atomic systems;
- Trends are visually identified.

Proposed **Solution**: to empoloy Correlation Analysis...

- Large amount of data can be used;
- Methodological analysis of trends;
- Easy to calculate and implement.

Chalange: A good featurization processes in mandatory...

Correlation Analysis

Pearson Correlation Coefficient:

$$r = \frac{cov(x, y)}{\sigma(x)\sigma(y)}$$

Correlation Interpretation:

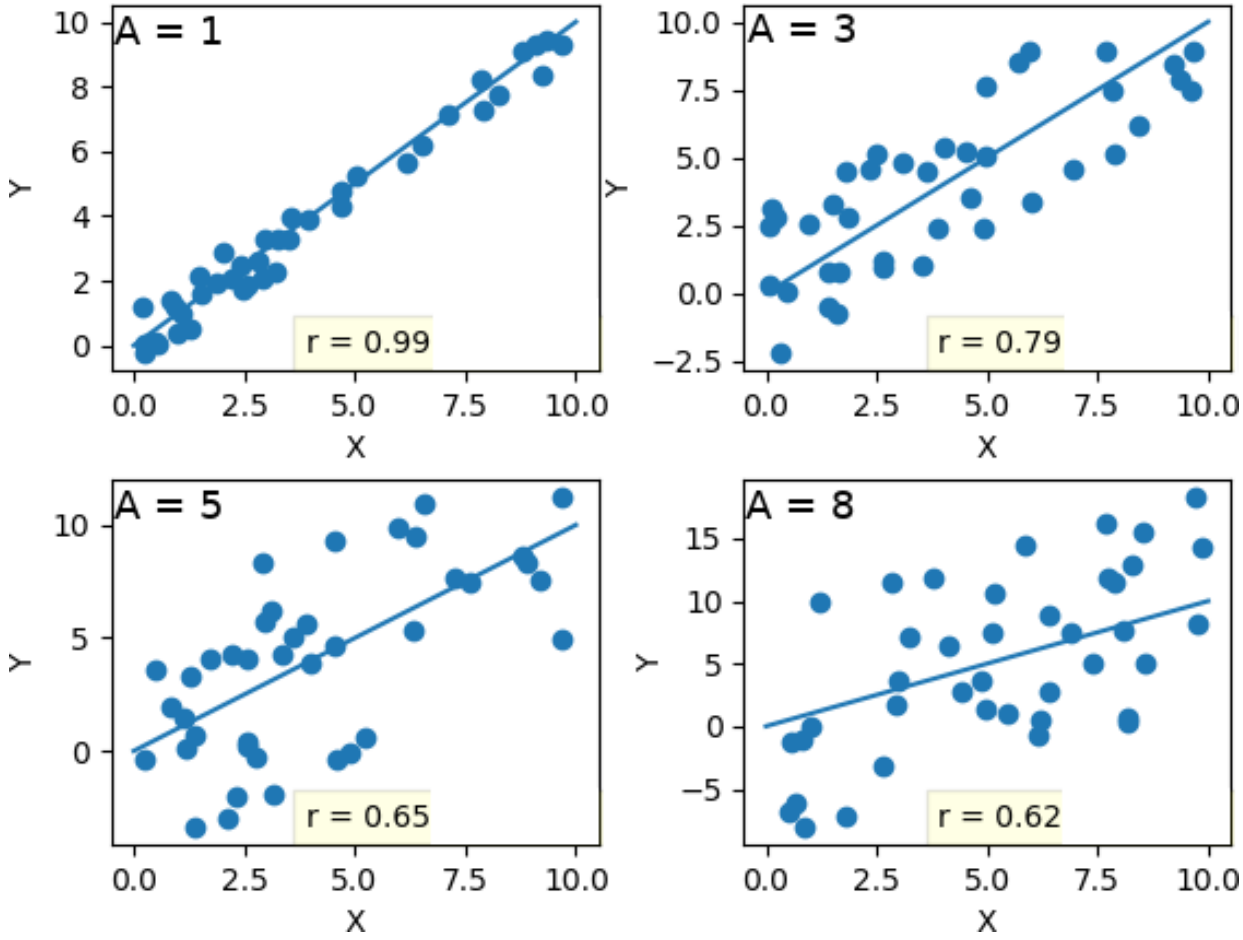
- $-1 \geq r \geq 1$
- If X increase, Y is **expected** to go:

  $\overbrace{r <}^{\text{downward}} \underbrace{0}_{\text{upward}} < r$

- How much expected?
  "How strongly correlated?"

  "As large as was $|r|$."

$$X = \{\text{random points in } (0,1)\} * 10,$$
$$Y = X * 10 + \{\text{random noise in } (-5,5)\}$$



Pearson, Spearman, and Kendall Correlations

Pearson

$$r = \frac{cov(x, y)}{\sigma(x)\sigma(y)}$$

- Non-ranked data;
- Sensitive to outliers.

Spearman

$$r_s = \frac{cov(r_x, r_y)}{\sigma(r_x)\sigma(r_y)}$$

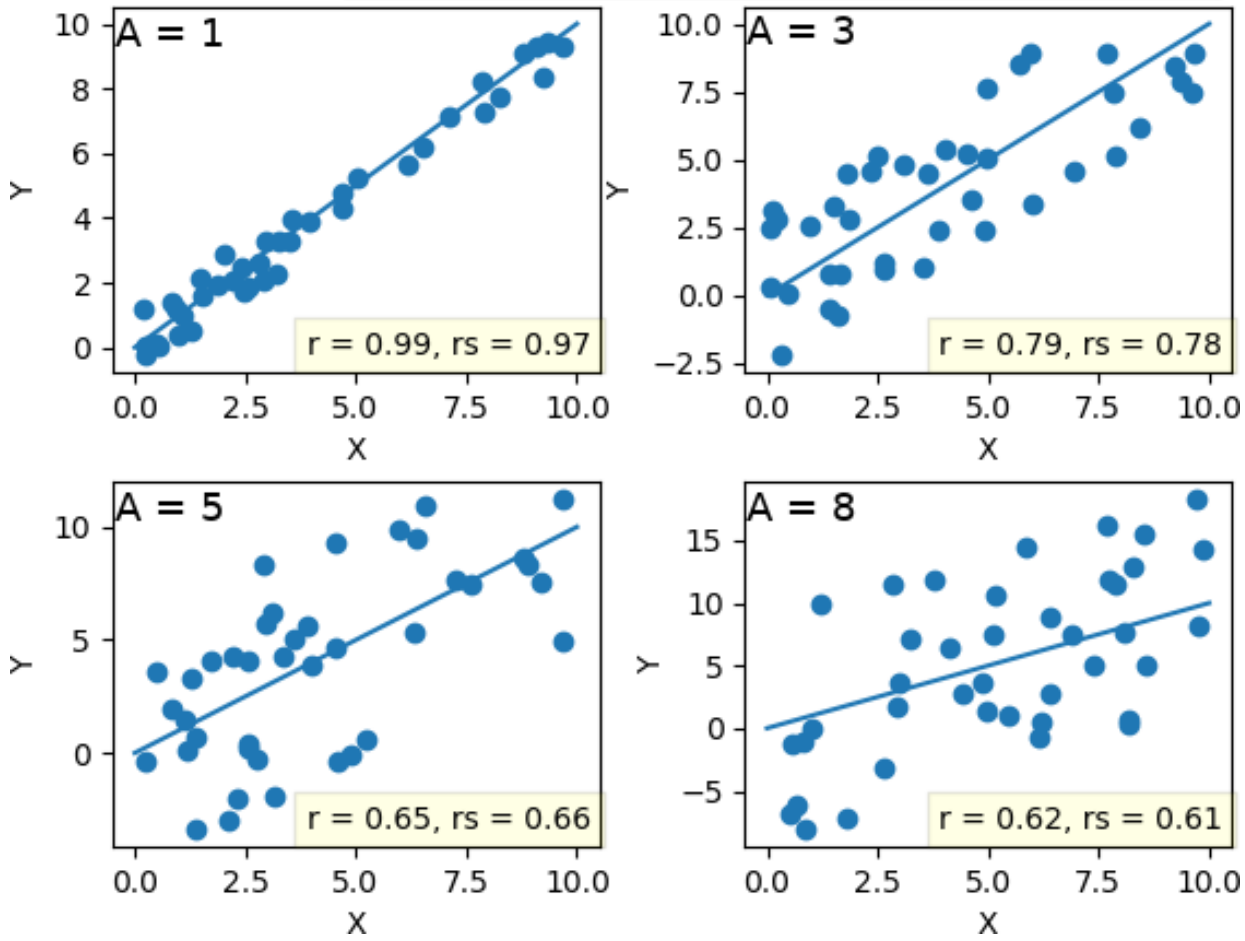
- Ranked data;
- Robust to outliers.

Spearman and Pearson are very similar (for data without outliers):

- r_s normaly is smaller then r.

For data without outlier:

- r reduced 0.89 (sensitive);
- r_s reduced 0.12 (robust).



Pearson, Spearman, and Kendall Correlations

Pearson

$$r = \frac{cov(x, y)}{\sigma(x)\sigma(y)}$$

- Non-ranked data;
- Sensitive to outliers.

Spearman

$$r_s = \frac{cov(r_x, r_y)}{\sigma(r_x)\sigma(r_y)}$$

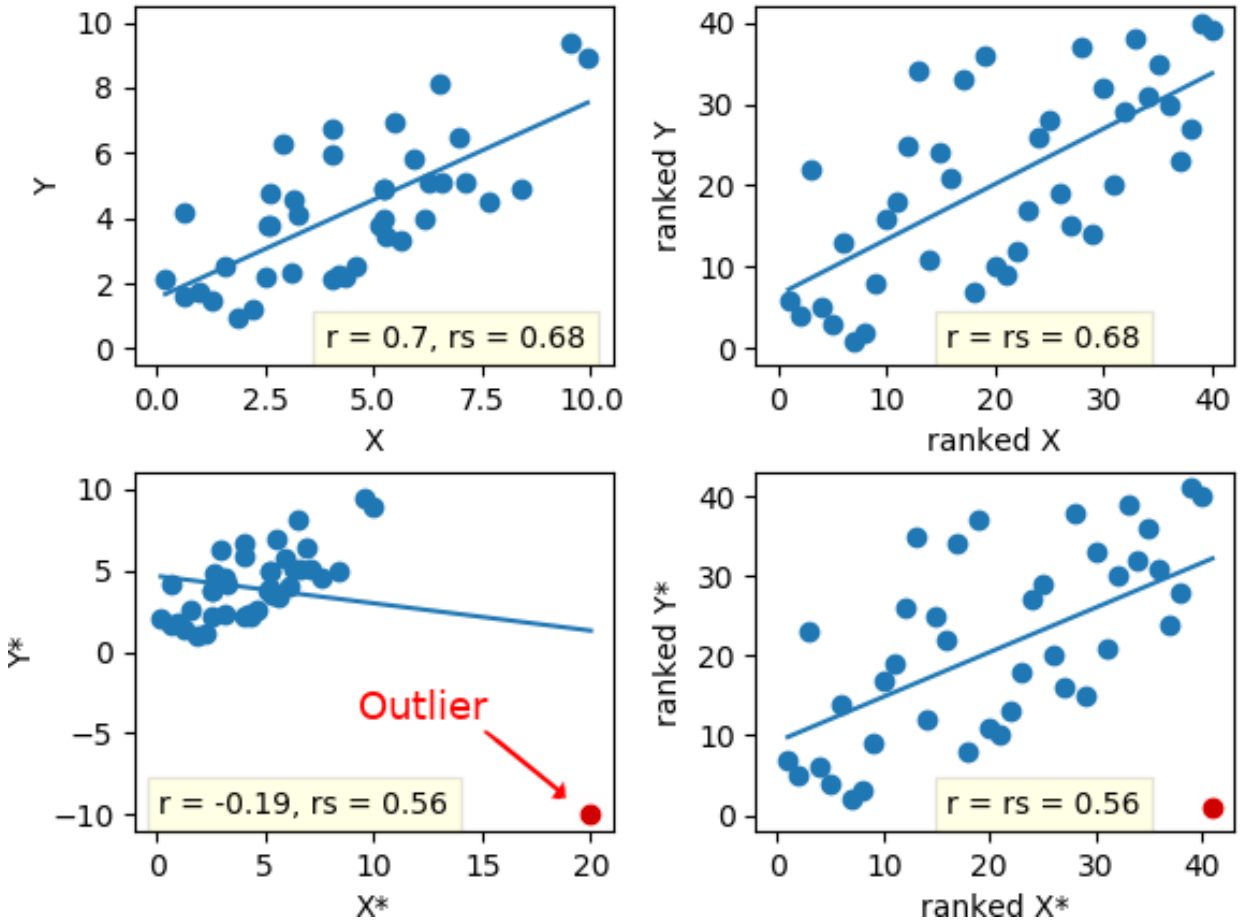
- Ranked data;
- Robust to outliers.

Spearman and Pearson are very similar (for data without outliers):

- r_s normaly is smaller then r.

For data without outlier:

- r reduced 0.89 (sensitive);
- r_s reduced 0.12 (robust).



Pearson, Spearman, and Kendall Correlations

Pearson

$$r = \frac{cov(x, y)}{\sigma(x)\sigma(y)}$$

- Non-ranked data;
- Sensitive to outliers.

Spearman

$$r_s = \frac{cov(r_x, r_y)}{\sigma(r_x)\sigma(r_y)}$$

- Ranked data;
- Robust to outliers.

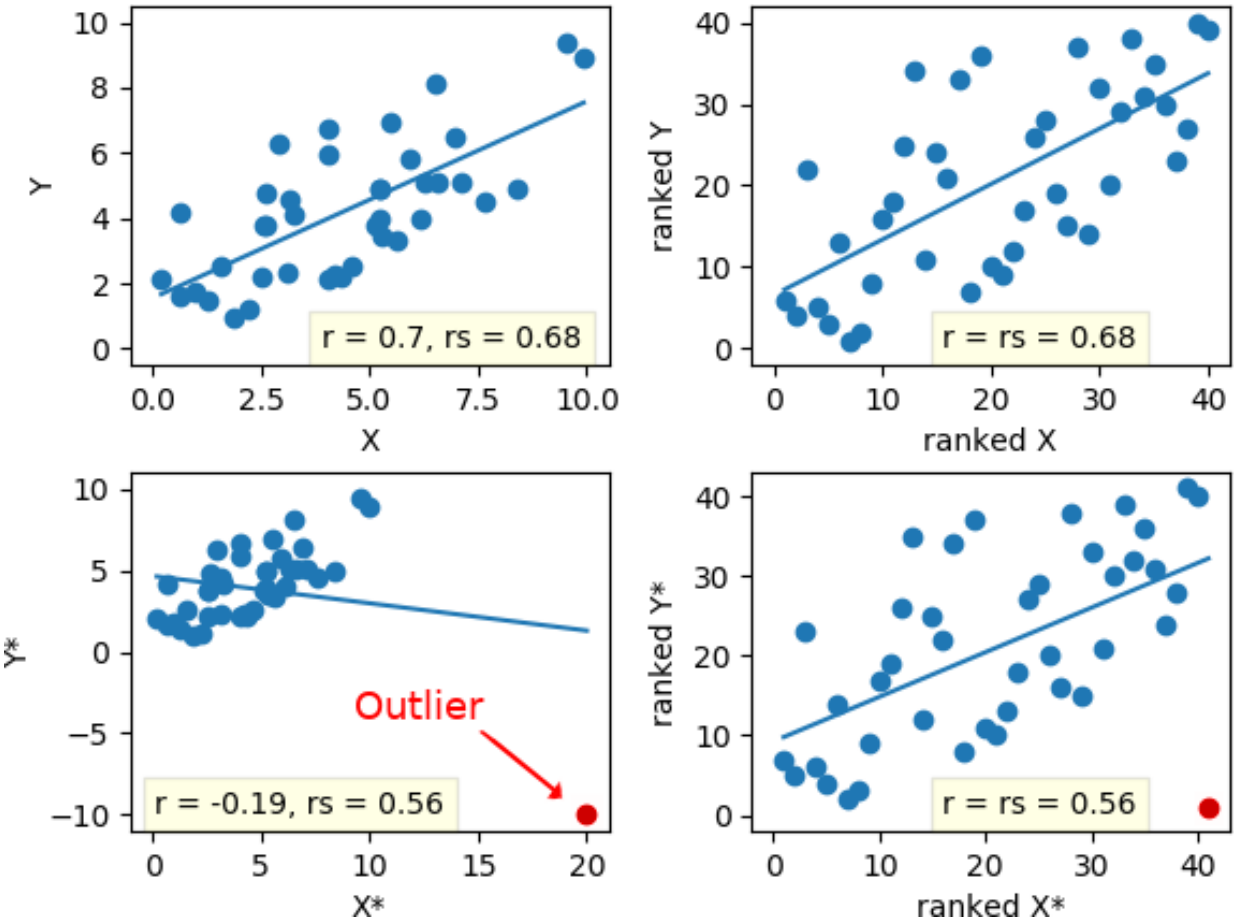
Kendall

$$\tau = \frac{\#\, cp - \#\, dp}{n(n-1)/2};$$

- Very similar to Spearman.

Differences:

- Values magnitude;
- P values;
- Probability interpretation.

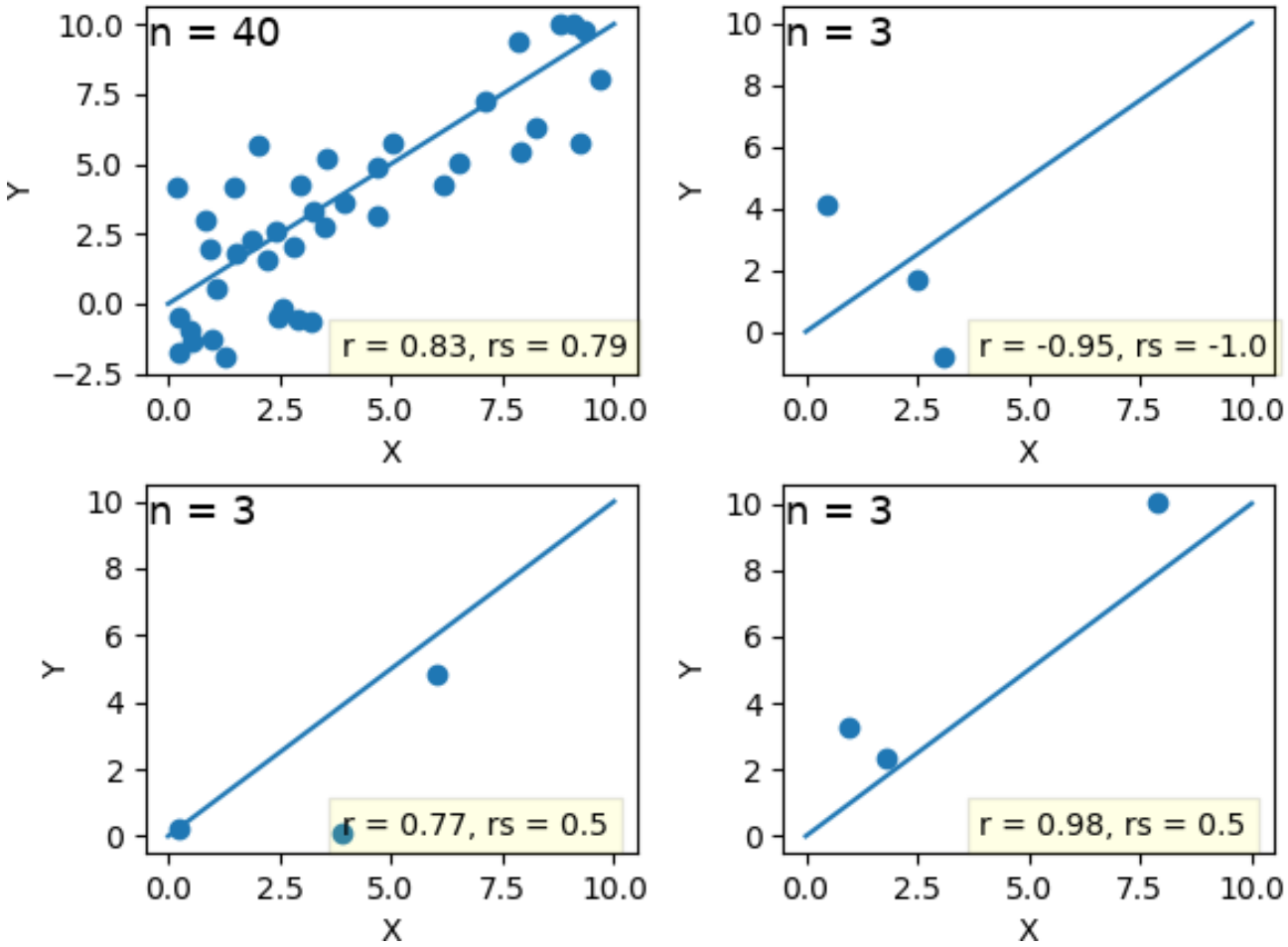


Significance of the Correlations

Should I trust large correlations?

It depend on data **size** and its **distribution**.

$$X = \{\text{n random points in } (0,10)\},$$
$$Y = X * 10 + \{\text{random noise in } (-5,5)\}$$

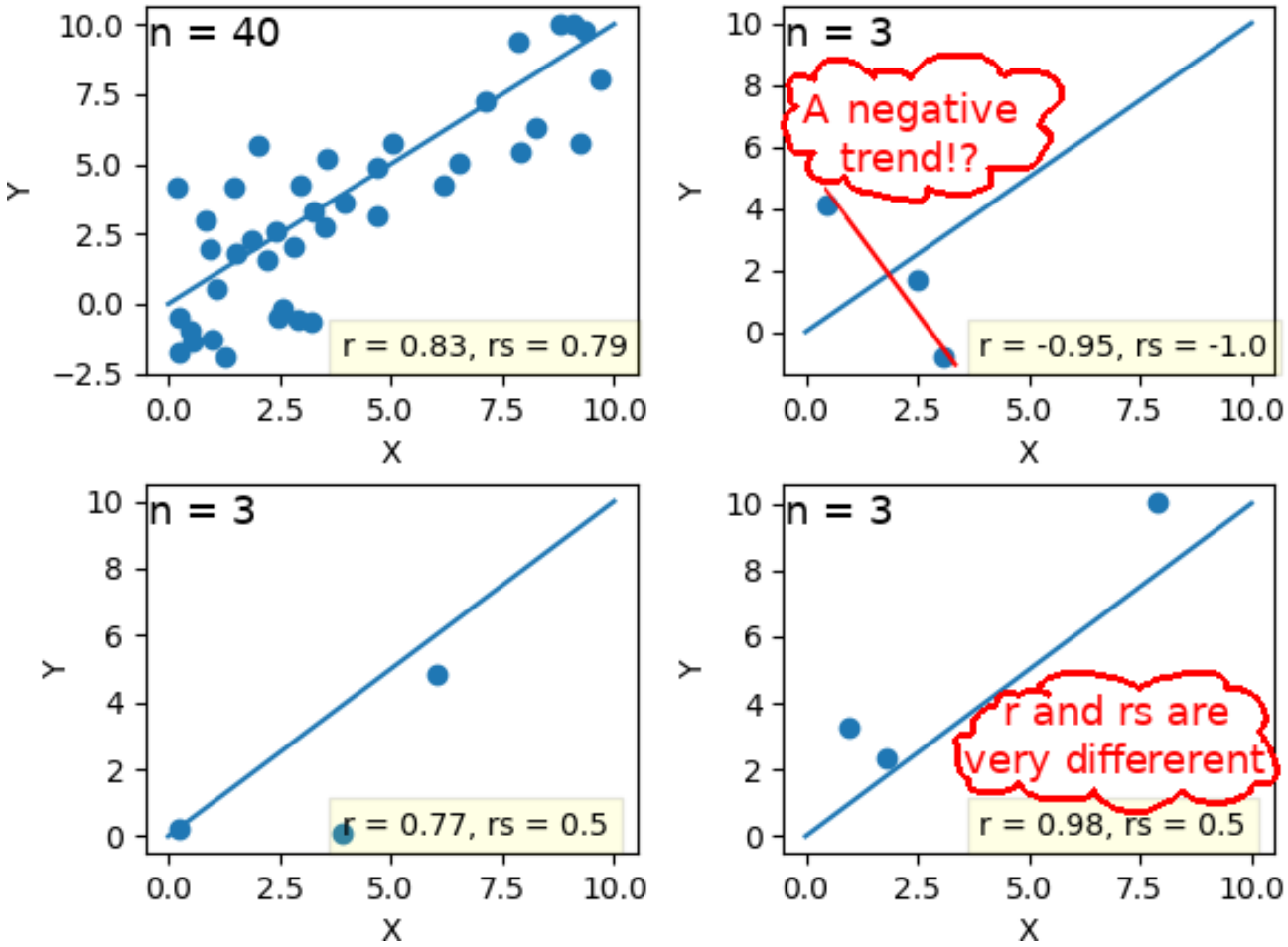


Significance of the Correlations

Should I trust large correlations?

It depend on data **size** and its **distribution**.

$$X = \{\text{n random points in } (0,10)\},$$
$$Y = X * 10 + \{\text{random noise in } (-5,5)\}$$



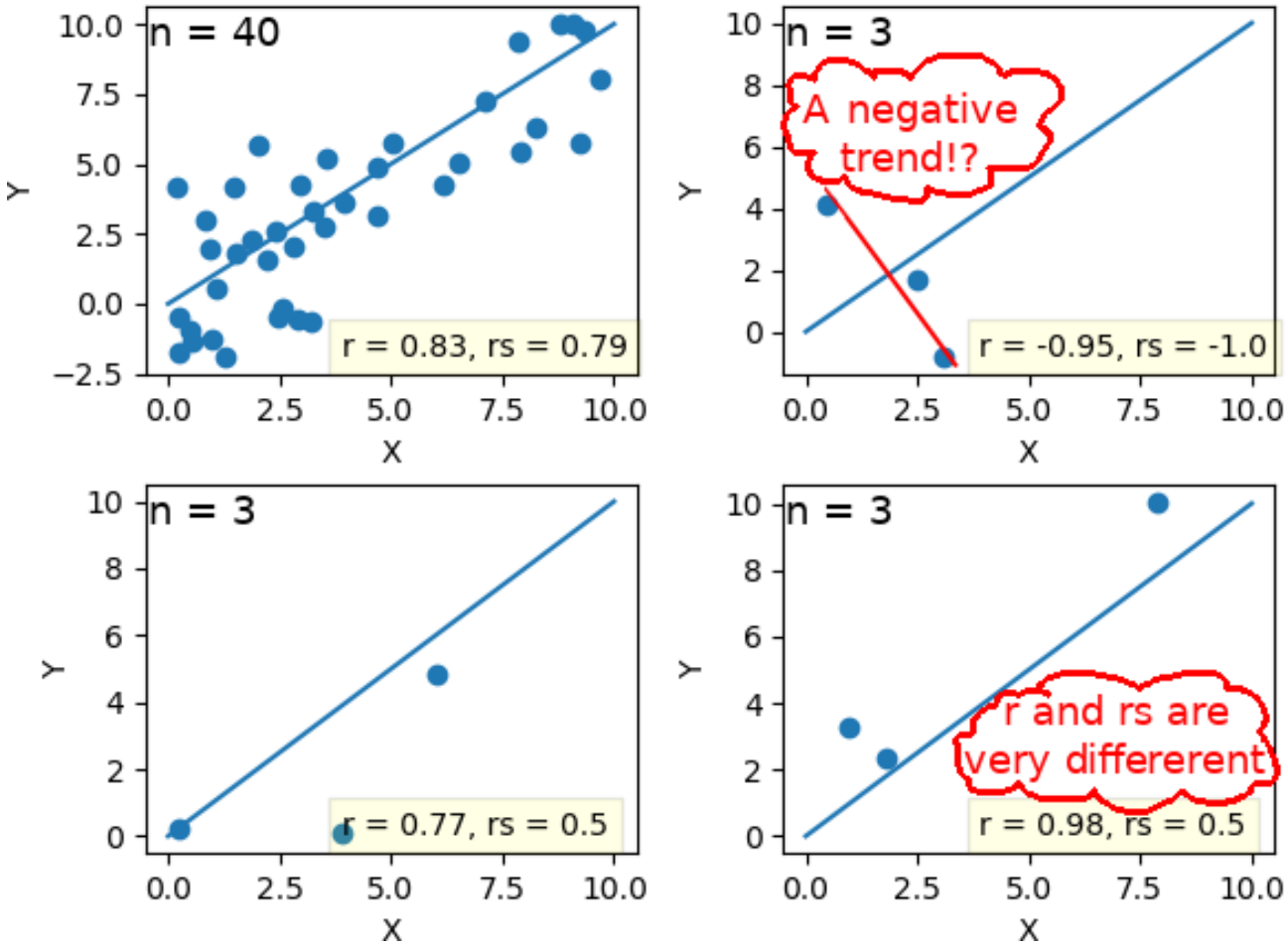
Significance of the Correlations

Should I trust large correlations?

It depend on data **size** and its **distribution**.

Good practice: Hypothesis test and P-values.

$$X = \{\text{n random points in } (0,10)\},$$
$$Y = X * 10 + \{\text{random noise in } (-5,5)\}$$



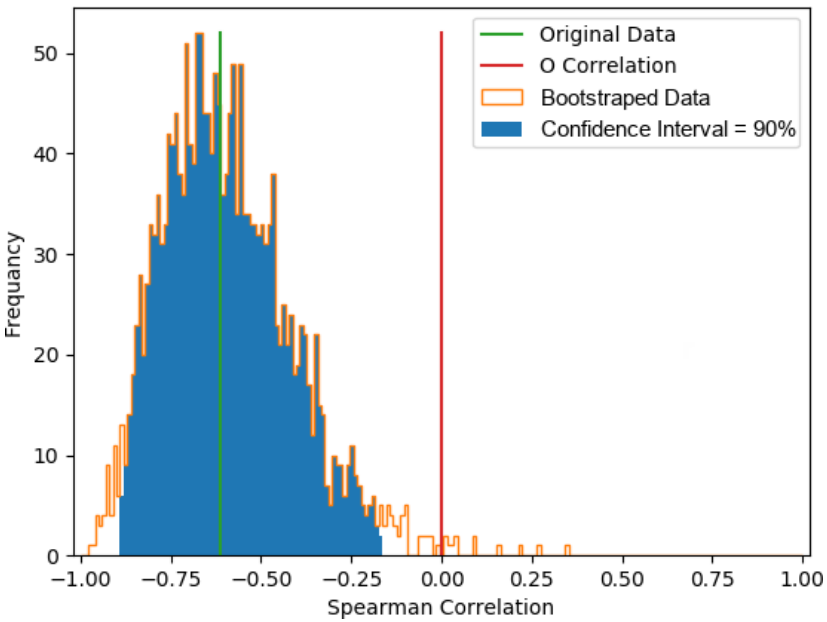
Bootstrap

Data Resampling Methods with Replacement (Ex.: $[1, 3, 4, 5] \rightarrow [1, 1, 3, 5]$).

Original samples $\rightarrow$ Bootstrap samples $\rightarrow$ **Statistical Information**

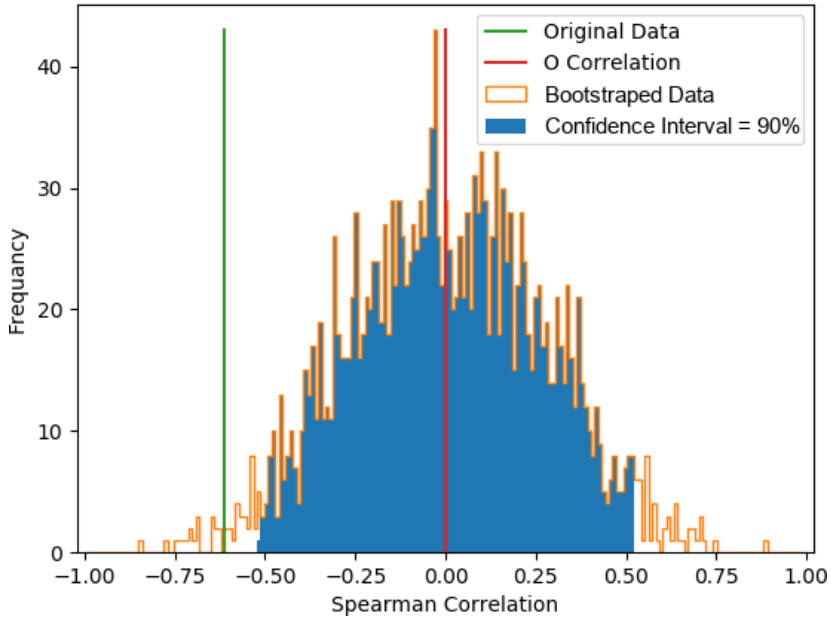
The **alternative hypothesis**: X and Y are correlated ($r_s \neq 0$).

$[(1,4),(2,5),(3,6)] \rightarrow [(1,4),(2,5),(2,6)]$
$\rightarrow [(3,6),(1,4),(2,5)]$
$\rightarrow [(3,6),(1,4),(3,6)]$
$\rightarrow \ldots$



The **null hypothesis**: X and Y are **not** correlated, ($r_s = 0$).

$[(1,4),(2,5),(3,6)] \rightarrow [(1,6),(3,5),(1,4)]$
$\rightarrow [(3,5),(1,6),(2,4)]$
$\rightarrow [(2,4),(2,4),(3,6)]$
$\rightarrow \ldots$



Featurization - Challenge

DM/ML Input:

attribute-value table	feature0	feature1	feature2...	
sample0	data00	data01	data02	...
sample1	data10	data11	data12	...
...	...	...	...	...

Initial Data:

Molecular info (structured data):

- Energy [−42.0 eV];
- HOMO [−4.123 eV];

Atomic info (attribute-vector):

- Exposed to vacuum [True, False, ...];
- Atomic Charges [0.7, -0.8, ...].

Initial Data:

All data table	Energy	HOMO	Charges	...
molecule0	-42.0	-4.123	[0.7, -0.8, ...]	...
molecule1	-42.1	-4.042	[0.6, -0.3, ...]	...
...	...	...	...	...

Featurization - Challenge

DM/ML Input:

attribute-value table	feature0	feature1	feature2...	
sample0	data00	data01	data02	...
sample1	data10	data11	data12	...
...	...	...	...	...

Initial Data:

Molecular info (structured data):

- Energy [−42.0 eV];
- HOMO [−4.123 eV];

Atomic info (attribute-vector):

- Exposed to vacuum [True, False, ...];
- Atomic Charges [0.7, -0.8, ...].

Initial Data:

All data table	Energy	HOMO	Charges	...
molecule0	-42.0	-4.123	[0.7, -0.8, ...]	...
molecule1	-42.1	-4.042	[0.6, -0.3, ...]	...
...	...	...	...	...

How to get molecular data from atomic data?

Average per atoms???
Average per atomic species???

Featurization - Mining Features Strategy

How to get molecular data from atomic data?

Take **molecular data** from **operator** over a **bag** (properties), for all or a **class** of atoms.

$$\underbrace{OPERATOR(OP)}_{\text{average}}\left[\underbrace{BAG}_{\text{ECN}}\left[\underbrace{CLASS}_{\text{Pt}}\right]\right] = OP\left[\underbrace{Selected\_Data}_{\text{Pt atoms ECN}}\right] = Molecular\_Data$$

$$\text{Av.}\left[\begin{bmatrix}1.1\\3.3\\2.2\\3.1\end{bmatrix}\begin{bmatrix}\checkmark\\\times\\\checkmark\\\times\end{bmatrix}\right] = \text{Av.}\left[\begin{bmatrix}1.1\\2.2\end{bmatrix}\right] = 1.65$$

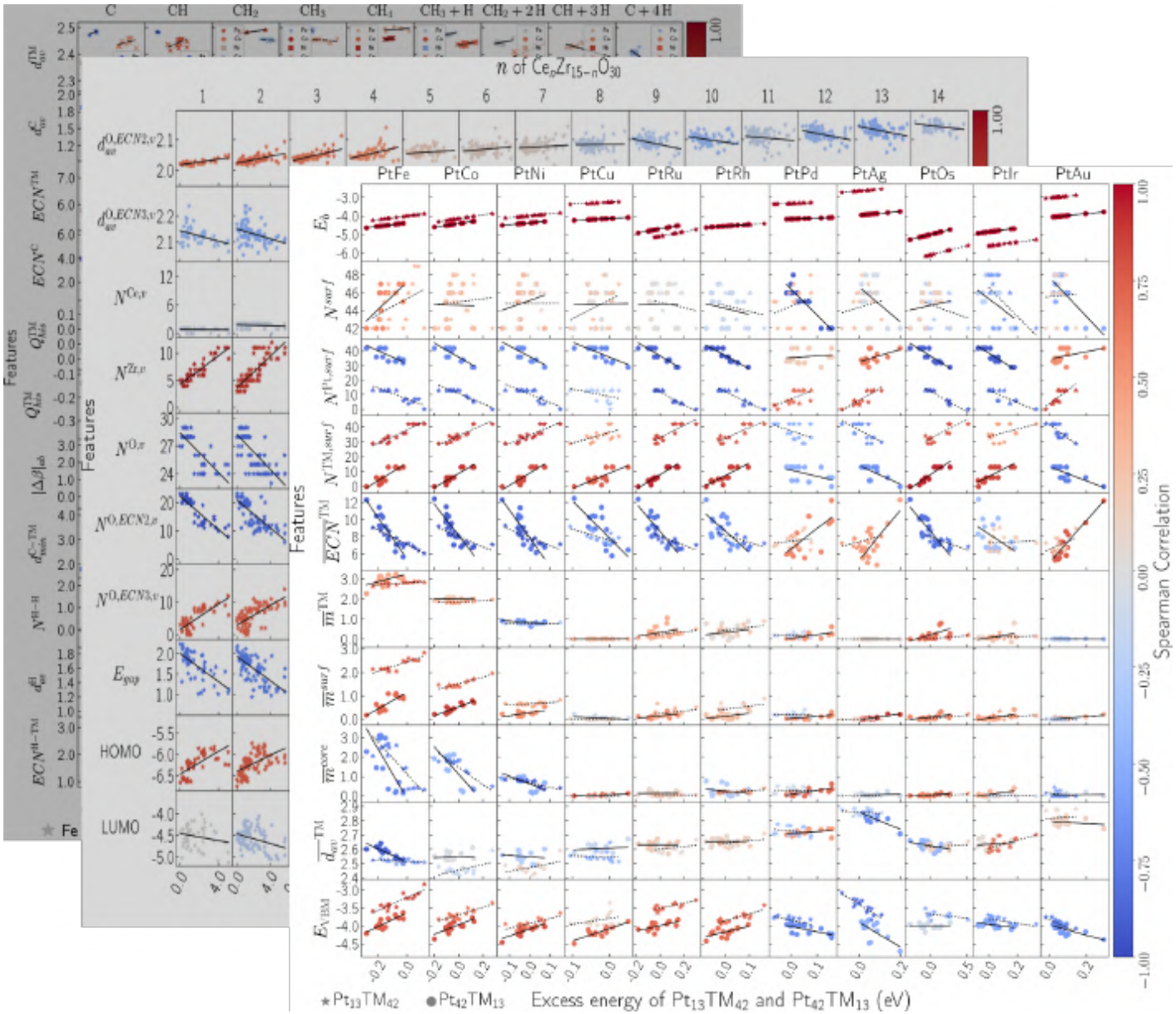
Operator: Operates over an array argument, and return a number (Ex.: sum).

Bag: Array with atomic data (Ex.: d_{av})...

Classes: Set of atoms that meet a condition! Ex.:

- O;
- O exposed to the vacuum;
- O exposed to the vacuum with $1 < ECN < 2$;
- (be criative)...

Results Visualization



Scatter-plot Matrix:

- Rows: Features;
- Columns: Groups of data. (to avoid Simpson's paradox)

Cell:

- Scatter-plot;
- Y axis: Feature;
- X axis: Energy;
- Correlation: Colors;
- Linear Model.

Practical Tips

Tipical Problem	Tips
Small significace, small correlations.	1) If especific classes? Generalize your classes. 2) If general classes? Split more your classes. 3) Else, probable the features are not suitable. Then, be creative and develop new features for your study case.
Small significace, large correlations.	Probable a few samples (ex.: $n < 15$) case. Then, generalize you classes to have more samples or calculate more samples.
Unexpected correlation A vs B	A may be correlated with C that is correlated with B. Verify other features searching for a C...
Expected correlation A vs B is too small	If B is based in a atomic feature, Generalize or especify the classes.

Good Scenarios	Tips
Large significace, small correlations.	It is not a problem. But you can try to specify more your classes to get more crucial features.
Large significace, Large correlations.	It is not a problem. Already get crusial features.

Implementation - Quandarium

Create your data mining procedure. Usage:

- Find Calculation → Extract Info → Molecular Analysis → Featurization → Plots
- Find Calculation → Extract Info → Molecular Analysis → Save Data
- Read Data → Featurization → Save Data
- Read Data → Plots

Find Calculation:

- Search calculations folders.

Extract Info:

•Energy; •State Energies;
•Positions; •Chemical Species;
•Charges, •....

Molecular Analysis:

•ECN; •d_{av}; •Connectivy;
•Surf_or_core; •Site_geometry;
•Number_of_conections.

Featurization:

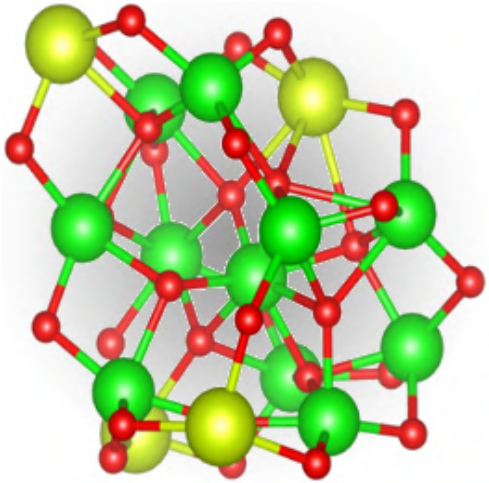
- bag ↔ class
- class1, class2 → class3
- bag1 → bag2
- bag1, bag2, ... → bag3
- bag[class] → molecular_data
- bag → molecular_data
- class → molecular_data

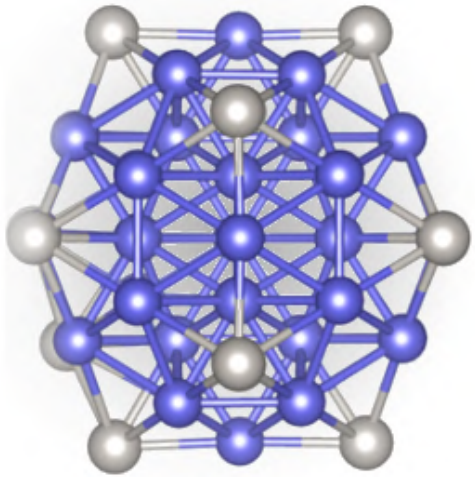
Plots:

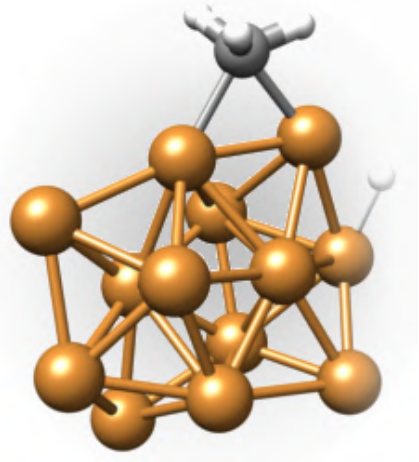
- Scatternplot with correlations;
- Bag histograms.

It operate recursively!

Introduction - Materials







$Zr_qCe_{15-q}O_{30}$

- Employed in TWC and candidate for others process.
- Explore material in nanoparticles structures.

Pt_lTM_{55-l}

- Pt-based Catalysts are wildly employed.
- Explore this alloys structural preferences.

$CH_m/xH/TM_{13}$

- CH_4 dehydrogenation candidates.
- Reaction intermediates study.

Motivation:

- Develop the chem/comp overlap;
- Develop material science area.

Objectives:

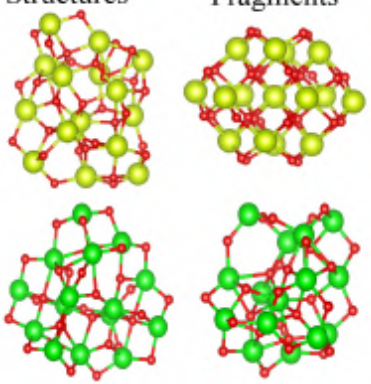
- Find new patterns;
- Initial Studies;
- Tools Development.

Workflow

Zibordi-Besse *et al.* work:

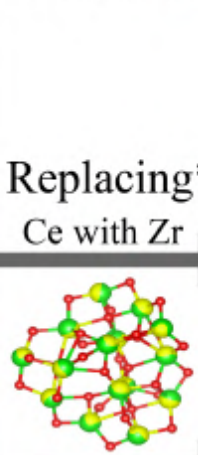
174 $Ce_{15}O_{30}$ structures
184 $Zr_{15}O_{30}$ structures

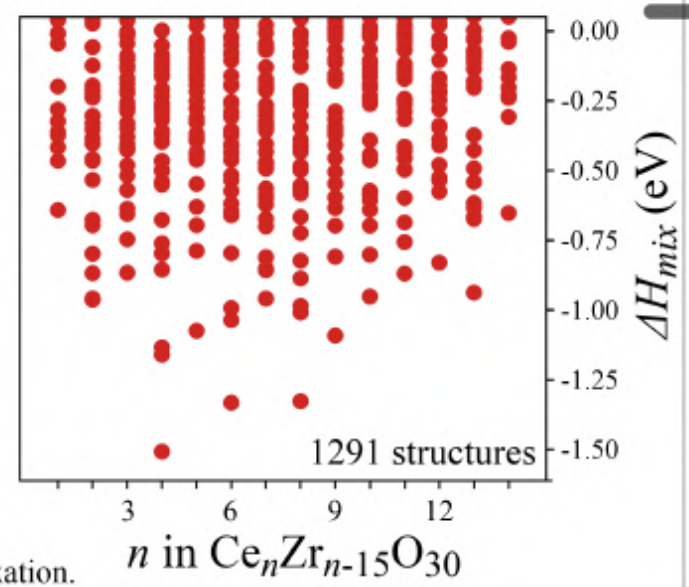
Stablest Structures | Bulk Fragments



Felício-Sousa *et al.* work:

Replacing*:
Ce with Zr
Zr with Ce



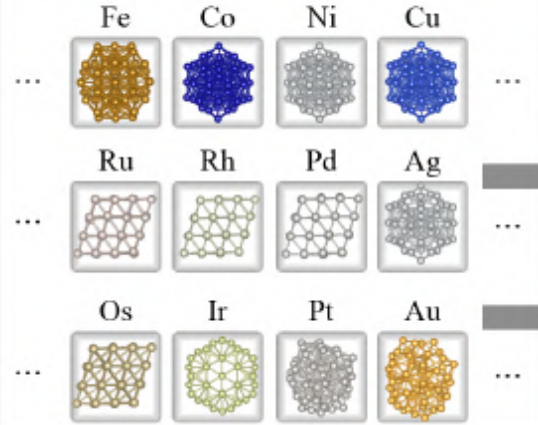


* Followed by relaxation.

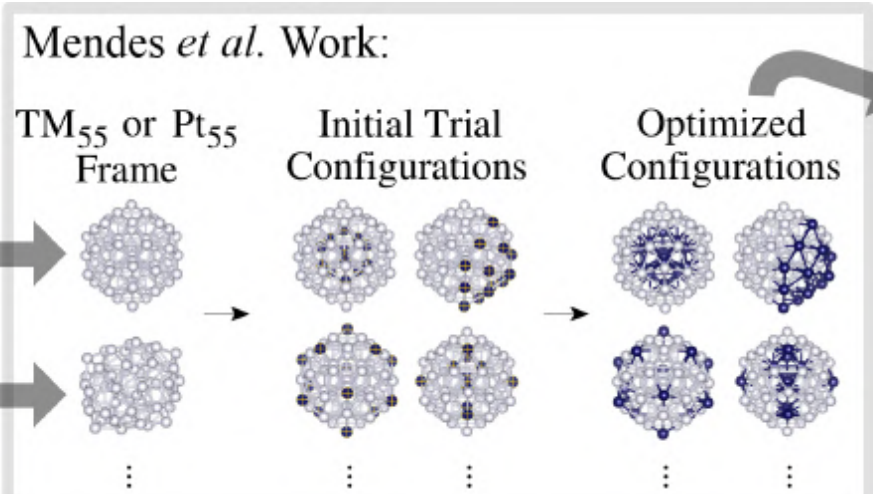
Johnatan's work:

- **Data Extraction:**
  Molecular Properties.
  Atomic Properties (bags).
- **Data Mining:**
  Obtain Classes.
  Obtain Mined Features.
  Correlation Analysis.
- **Data Presentation:**
  Select Important features.
  Build up Scatter-plot Matrix.

Piotrowski *et al.* Work:



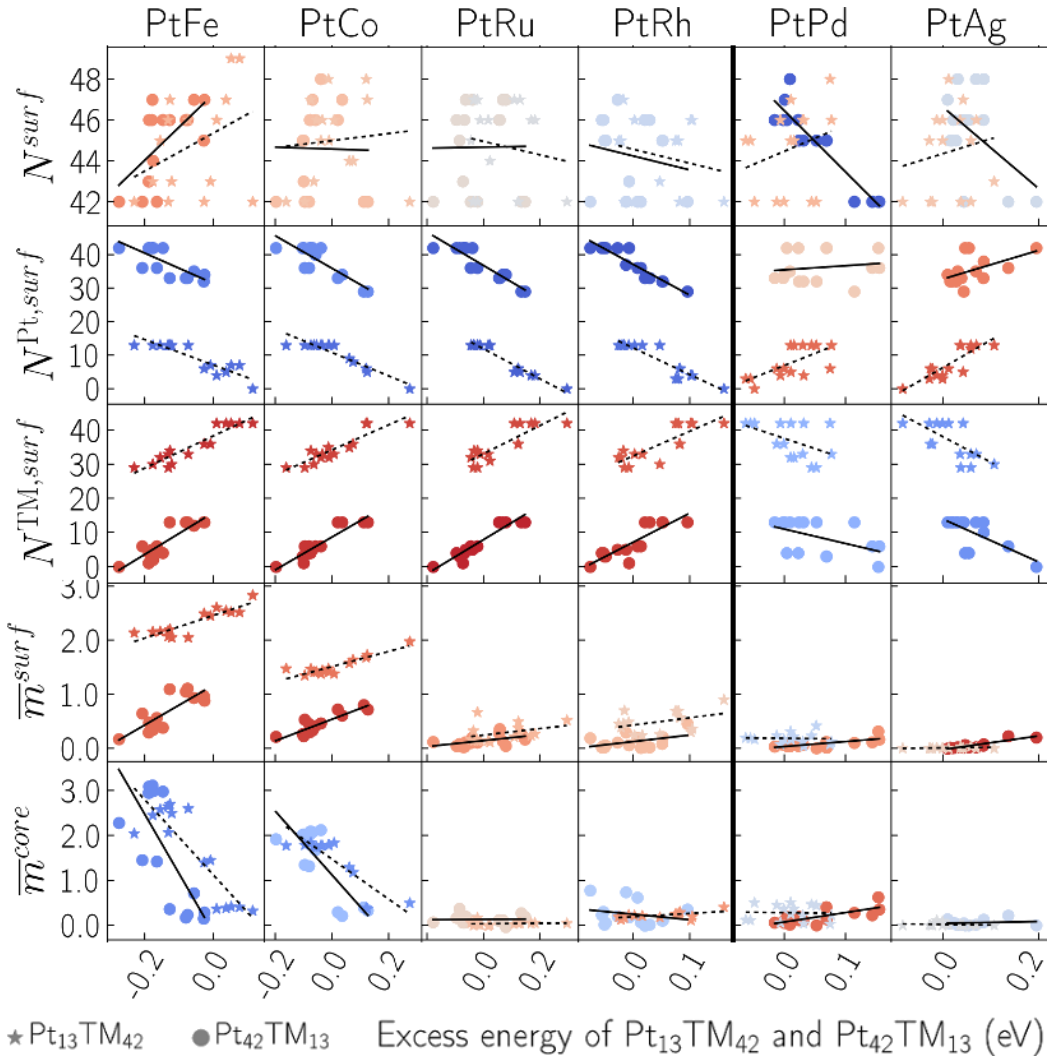
Mendes *et al.* Work:



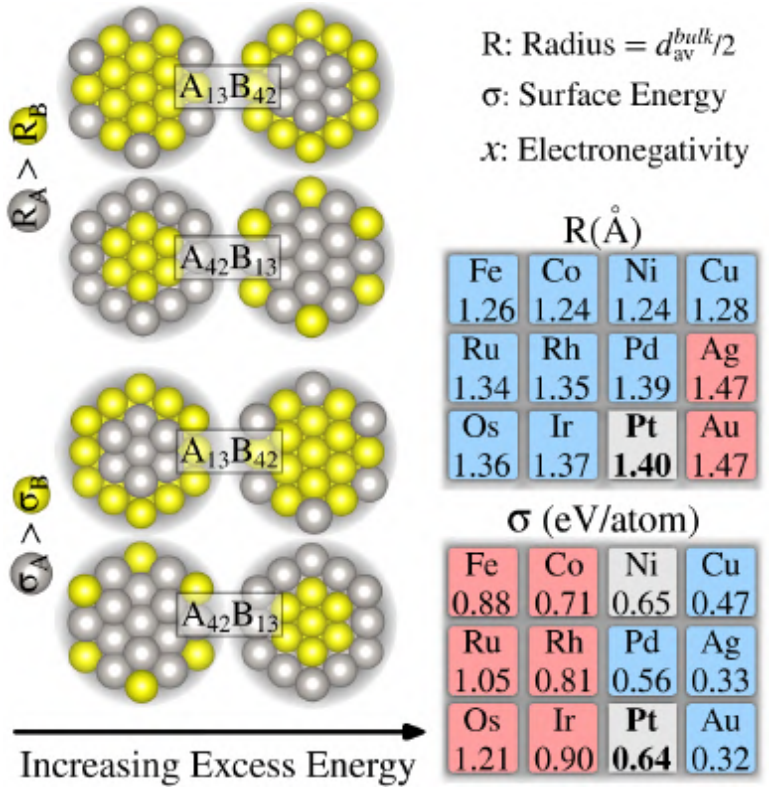
Johnatan's work:

- **Data Extraction:**
  Molecular Properties.
  Atomic Properties (bags).
- **Data Mining:**
  Obtain Classes.
  Obtain Mined Features.
  Correlation Analysis.
- **Data Presentation:**
  Select Important features.
  Build up Scatter-plot Matrix.

Results - PtTM

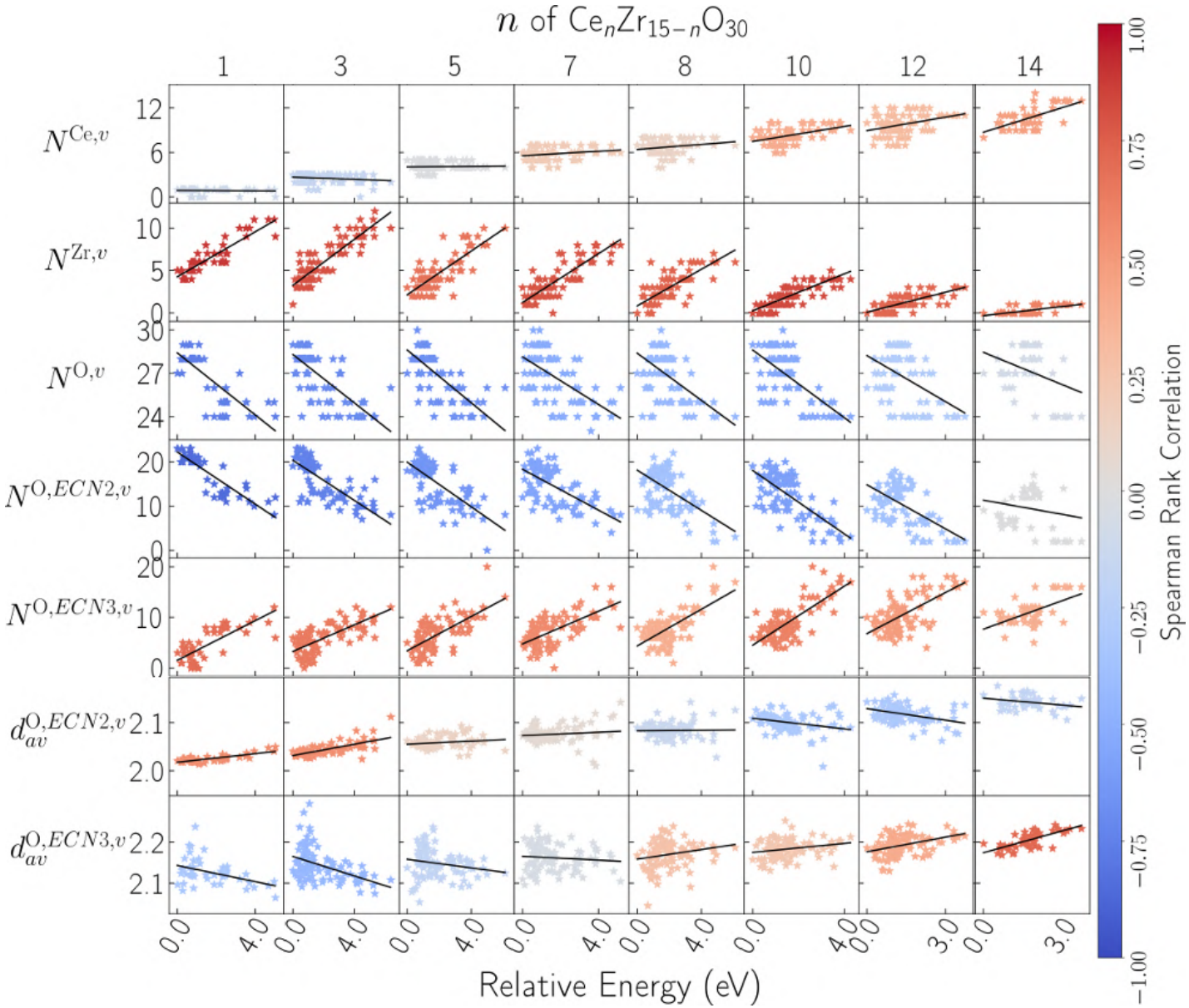


- High correlation! Few samples.
- TM *vs* Pt sites preference (N,ECN);
- Influence in other properties (m,d_{av});



Article: *J. Phys. Chem.*, **2020**, 124, 1, 1158-1164.

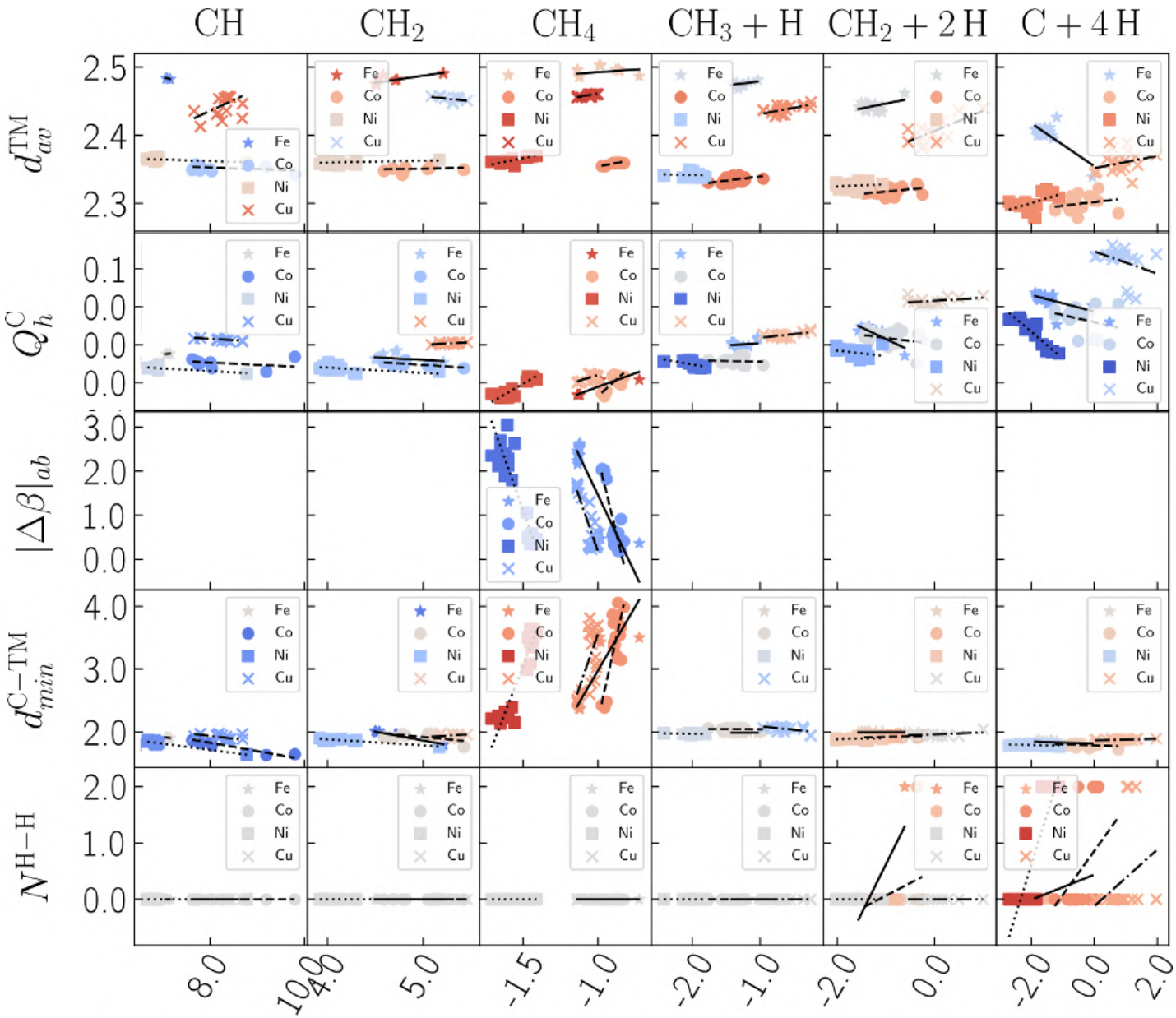
Results - $Ce_nZr_{15-n}O_{30}$



- High $n \rightarrow$ Significant correlation!
- Both Ce *vs* Zr prefer core sites;
- Zr trend is stronger;
- O prefer surface sites with ECN=2.
- Trends differ for each O species;

Article: *Phys. Chem. Chem. Phys.*, **2019**, 21, 26637-26646.

Results - $CH_n/H_m/TM_{13}$



- Data distribution problems;
- Trends change irregular with the systems;
- Few samples per many sistems.

Article submitted for the journal Fuel.

Conclusion

This Framework employed in QC datasets:

Benefits:

- Easy access to useful **chemistry information**;
- **Quantitative trends** analysis;
- Very **little explored**.

Limitation:

- Small **dataset size**;

Applicability:

- Can be applied to **any material**;
- Require small **programming skills**;
- Require some **statistical concepts**.

Acknowledgements

Thanks for your Attention!