LECTURE 1

CEIC6789 NOTES

DATA-DRIVEN DECISION MAKING IN CHEMICAL ENGINEERING AND FOOD SCIENCE

WHAT IS DATA SCIENCE?

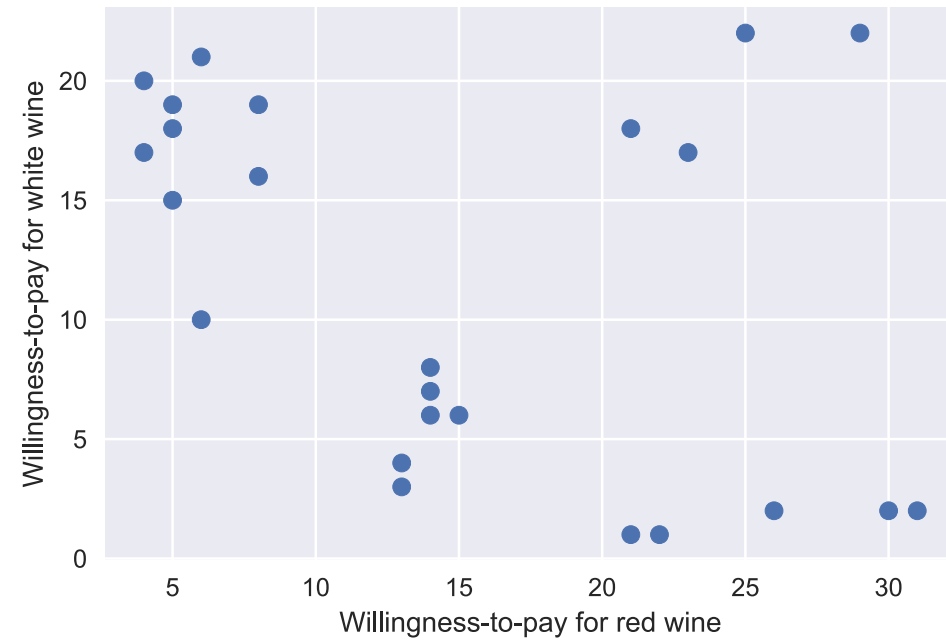DATA   ALGORITHM   INSIGHT
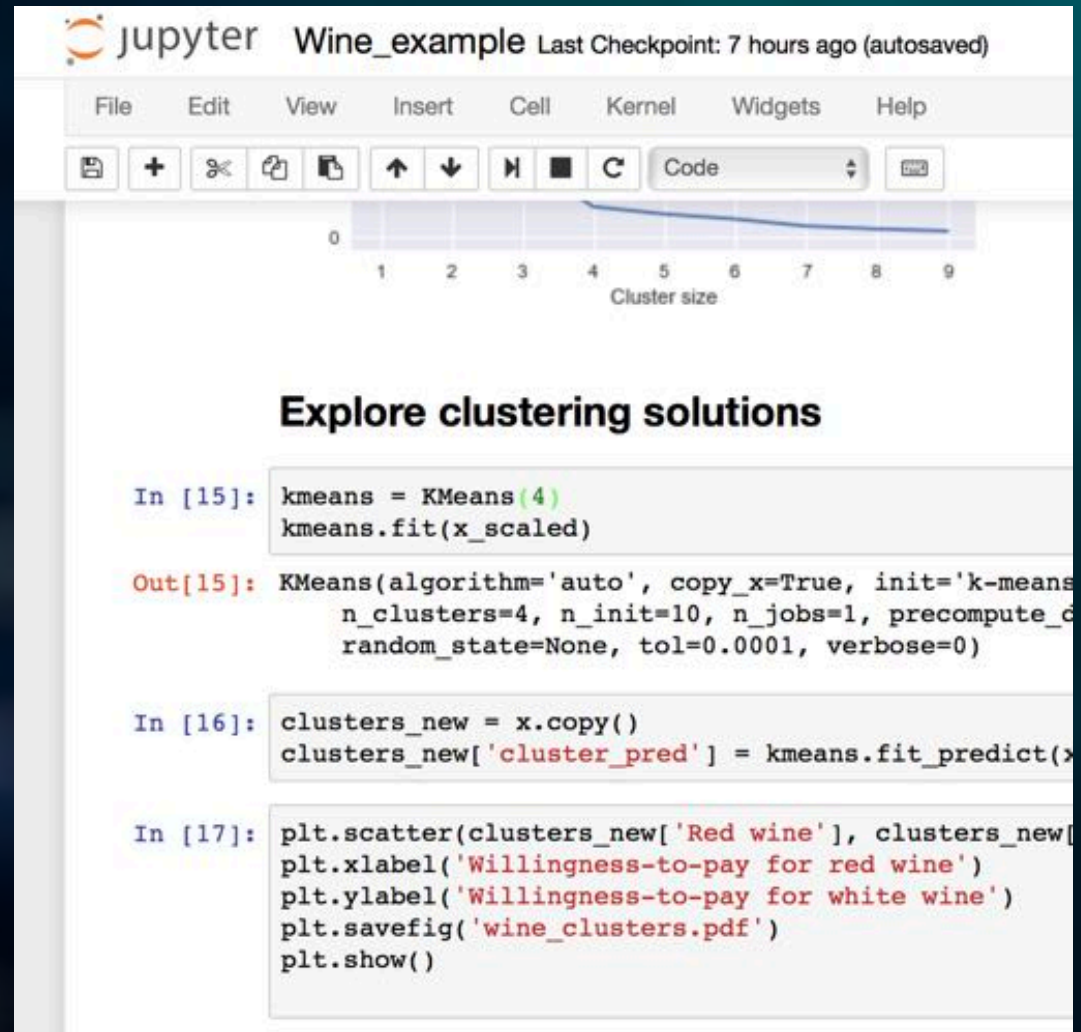
Red wine ($)	White wine ($)
21	1
15	6
5	18
31	2
25	22
6	21
14	8
8	16
21	18
26	2
4	17
23	17
14	7
5	19
13	3
4	20
5	15
29	22
22	1
8	19
13	4
30	2
14	6
6	10

DATA

ALGORITHM

Jupyter **Wine_example** Last Checkpoint: 7 hours ago (autosaved)

File   Edit   View   Insert   Cell   Kernel   Widgets   Help

Code

0

1   2   3   4   5   6   7   8   9

Cluster size

Explore clustering solutions

```
In [15]:  kmeans = KMeans(4)
          kmeans.fit(x_scaled)
```

```
Out[15]:  KMeans(algorithm='auto', copy_x=True, init='k-means
             n_clusters=4, n_init=10, n_jobs=1, precompute_d
             random_state=None, tol=0.0001, verbose=0)
```

```
In [16]:  clusters_new = x.copy()
          clusters_new['cluster_pred'] = kmeans.fit_predict(x
```

```
In [17]:  plt.scatter(clusters_new['Red wine'], clusters_new[
          plt.xlabel('Willingness-to-pay for red wine')
          plt.ylabel('Willingness-to-pay for white wine')
          plt.savefig('wine_clusters.pdf')
          plt.show()
```

ALGORITHM

INSIGHT

❖Pricing

❖Perks and Strategies

APPLICATIONS

MACHINE LEARNING & AI IN DRUG DISCOVERY
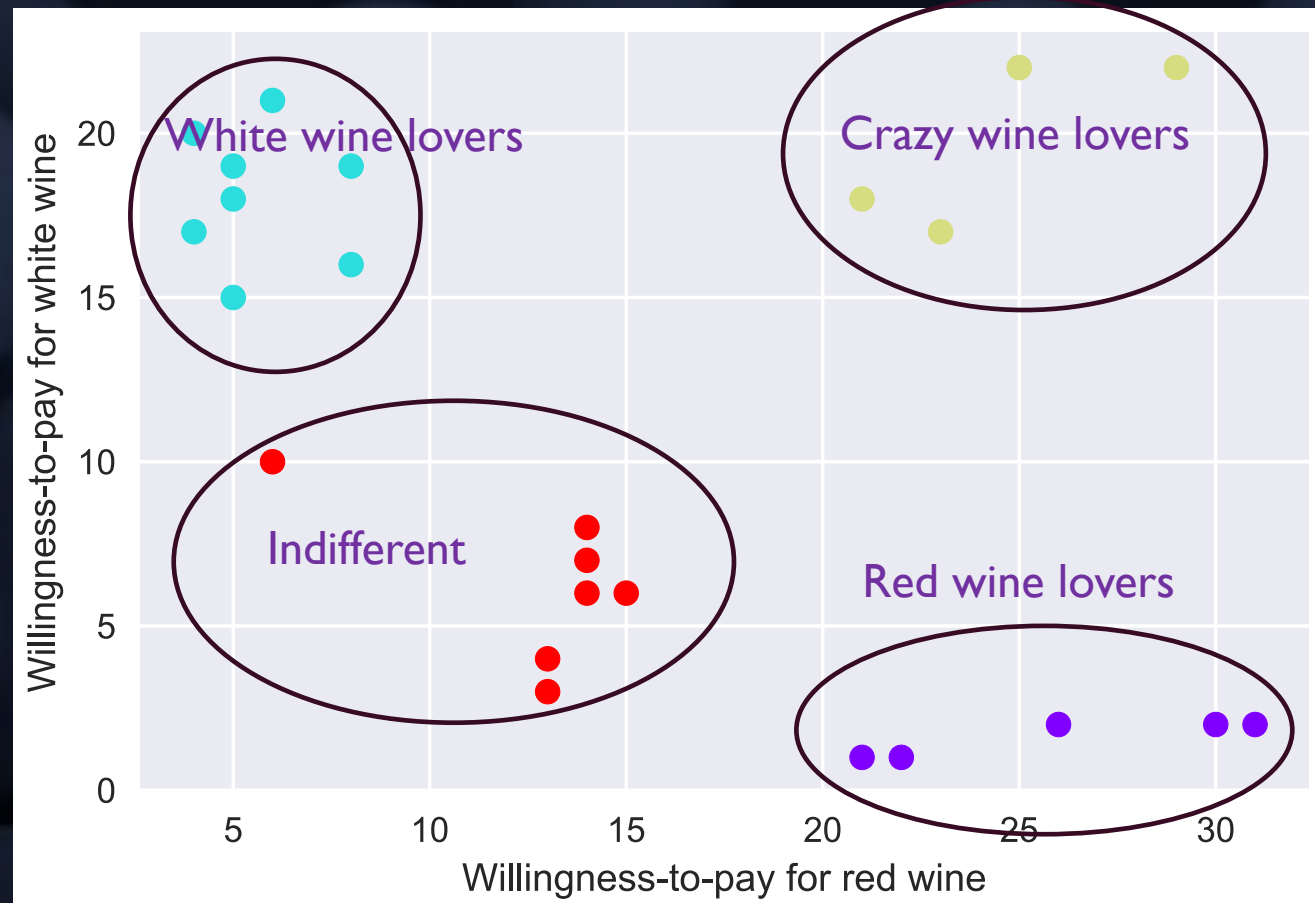
DATA SCIENCE APPLICATIONS

DATA SCIENCE TASK

DATA

Efficiency (%)	T (K)	P (bar)	composition	name
60	320	1.1	0.6	cat1
55	310	1.2	0.7	cat2
76	200	1.1	0.8	cat3
67	304	1000	0.66	cat4
68		1.3	0.82	cat5
55		0.9	0.55	25
78	1.00E+05	0.8	0.9	cat7
cat10	299	1.3	0.5	cat8

DATA SCIENTIST

Data → Analysis → Analytics

- Two scenarios
  - Given data, improve efficiency!
  - Given data, do something!

DATA SCIENCE DIAGRAM: WHAT IS INVOLVED?

DATA

Data collection
- ✧ raw data or primary data
- ✧ Cannot be analyzed

Preprocessing
- ✧ class labeling: numerical, categorical, text, video, audio etc.
- ✧ data cleaning: numbers mixed with text or vice versa
- ✧ missing values

Efficiency (%)	T (K)	P (bar)	composition	name
60	320	1.1	0.6	cat1
55	310	1.2	0.7	cat2
76	200	1.1	0.8	cat3
67	304	1000	0.66	cat4
68		1.3	0.82	cat5
55		0.9	0.55	25
78	1.00E+05	0.8	0.9	cat7
cat10	299	1.3	0.5	cat8

DATA SCIENCE DIAGRAM: WHAT IS INVOLVED?

DATA

Data collection
- ✧ raw data or primary data
- ✧ Cannot be analyzed

Preprocessing
- ✧ class labeling: numerical, categorical, text, video, audio etc.
- ✧ data cleaning: numbers mixed with text or vice versa
- ✧ missing values

ANALYSIS

Analyze the data
- ✧ What happened?
- ✧ When did it happen?
- ✧ How did it happen?
- ✧ … and many more!

Extract info and present:
- ✧ metrics
- ✧ KPIs (key performance indicators)
- ✧ Reports
- ✧ Dashboard

ANALYTICS

DATA SCIENCE DIAGRAM: WHAT IS INVOLVED?

DATA

Data collection
- ✧ raw data or primary data
- ✧ Cannot be analyzed

Preprocessing
- ✧ class labeling: numerical, categorical, text, video, audio etc.
- ✧ data cleaning: numbers mixed with text or vice versa
- ✧ missing values

ANALYSIS

Analyze the data
- ✧ What happened?
- ✧ When did it happen?
- ✧ How did it happen?
- ✧ ... and many more!

Extract info and present:
- ✧ metrics
- ✧ KPIs (key performance indicators)
- ✧ Reports
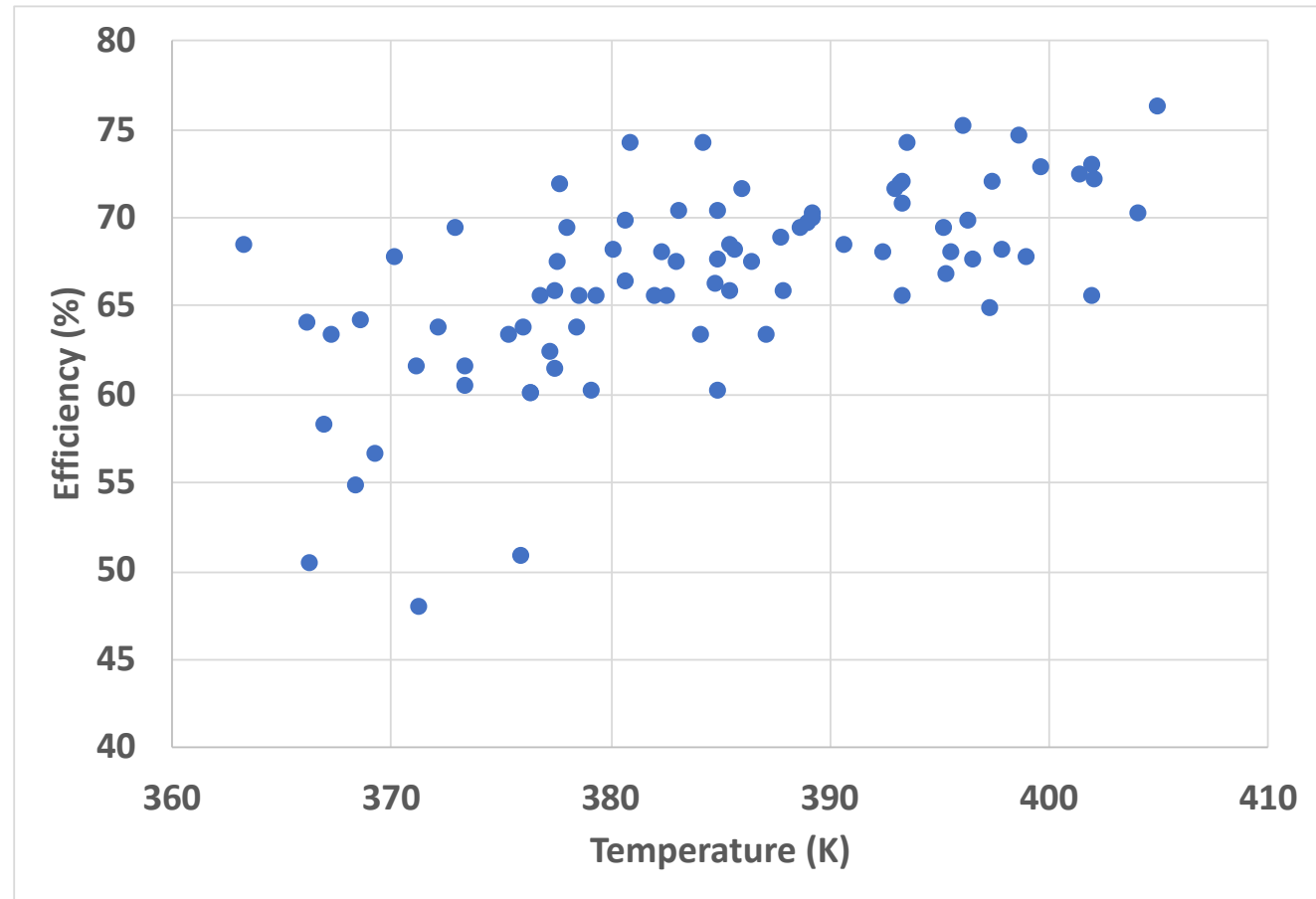- ✧ Dashboard

ANALYTICS

Branches
- ✧ Traditional methods
- ✧ Machine learning

Techniques
- ✧ Regression (linear, non-linear)
- ✧ Clustering
- ✧ PCA (Principal Component Analysis)
- ✧ Time series
- ✧ SVMs (support vector machines)
- ✧ Neural networks
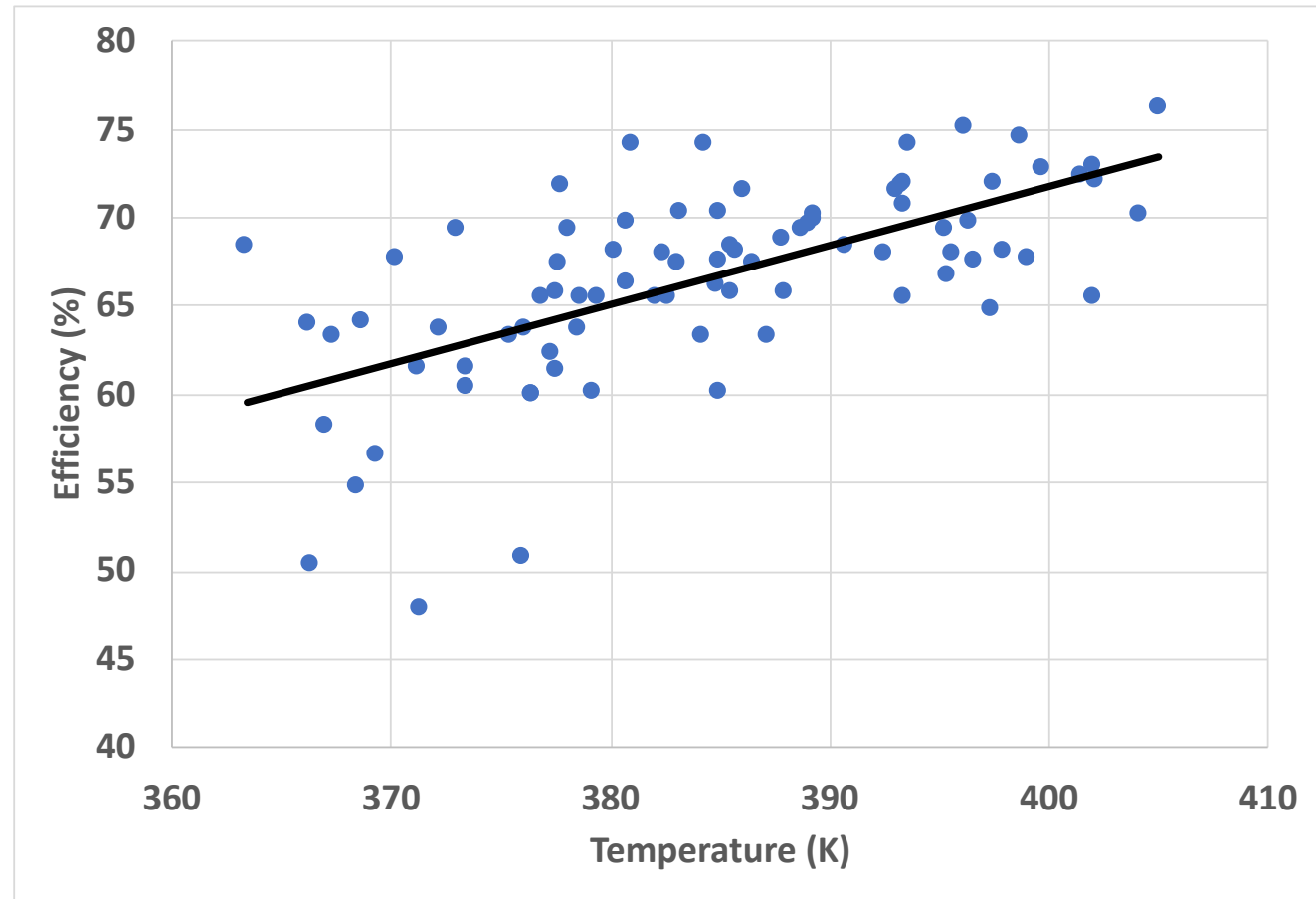- ✧ Deep learning
- ✧ Bayesian networks

REGRESSION

T (K)	efficiency (%)
371.4	48
366.4	50.4
376	50.8
368.5	54.8
369.3	56.6
367	58.2
376.4	60
376.4	60
379.2	60.2
385	60.2
373.5	60.4
377.5	61.4
373.5	61.6
371.2	61.6
377.3	62.4
387.2	63.4
375.5	63.4
367.4	63.4
384.2	63.4
378.6	63.8

REGRESSION

T (K)	efficiency (%)
371.4	48
366.4	50.4
376	50.8
368.5	54.8
369.3	56.6
367	58.2
376.4	60
376.4	60
379.2	60.2
385	60.2
373.5	60.4
377.5	61.4
373.5	61.6
371.2	61.6
377.3	62.4
387.2	63.4
375.5	63.4
367.4	63.4
384.2	63.4
378.6	63.8

DATA SCIENCE DIAGRAM: WHAT IS INVOLVED?

DATA

Data collection
✧ raw data or primary data
✧ Cannot be analyzed

Preprocessing
✧ class labeling: numerical, categorical, text, video, audio etc.
✧ data cleaning: numbers mixed with text or vice versa
✧ missing values

ANALYSIS

Analyze the data
✧ What happened?
✧ When did it happen?
✧ How did it happen?
✧ … and many more!

Extract info and present:
✧ metrics
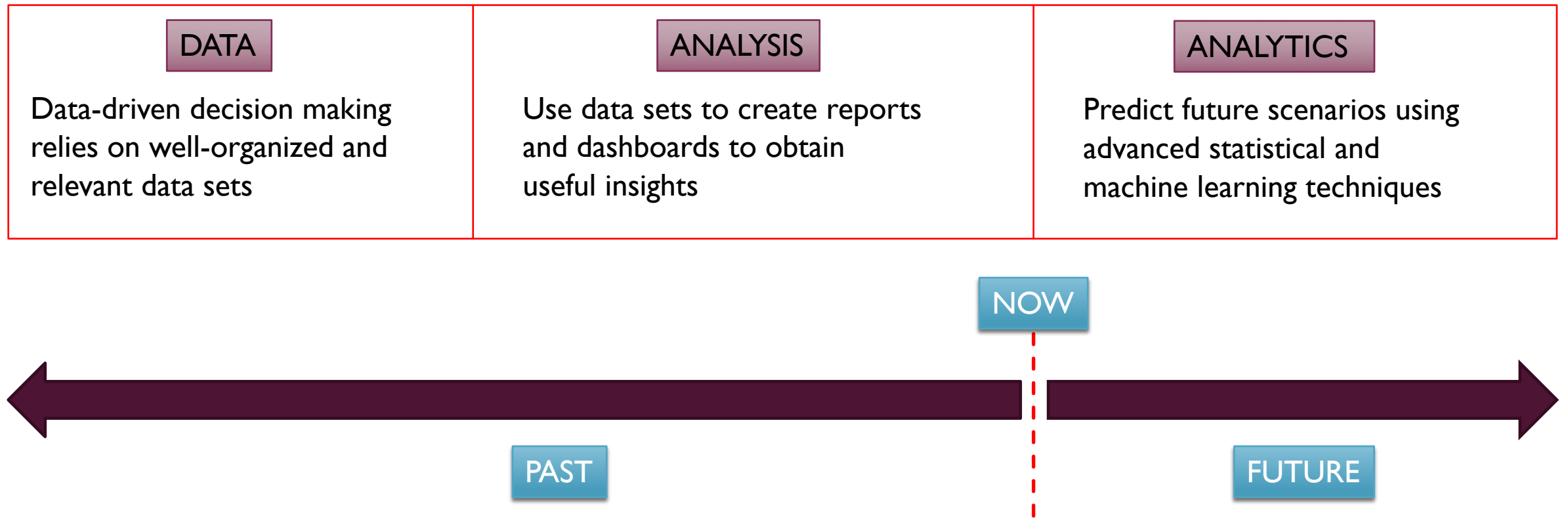✧ KPIs (key performance indicators)
✧ Reports
✧ Dashboard

ANALYTICS

Branches
✧ Traditional methods
✧ Machine learning

Techniques
✧ Regression (linear, non-linear)
✧ Clustering
✧ PCA (Principal Component Analysis)
✧ Time series
✧ SVMs (support vector machines)
✧ Neural networks
✧ Deep learning
✧ Bayesian networks

DISCUSSION
WITH EXPERTS

QUESTIONS ASKED DURING PANEL DISCUSSION

General

- Could you introduce yourself?

- What are some exciting areas where data science/machine learning (ML) is being applied?

- How important are data science/ML jobs becoming? Do you see an upward trend?

Job-specific

- What are the most important data science/ML concepts you use in your job?

- Which programming languages do you use?

- How much of learning (skills, software etc) happened before, and how much is happening during the job?

Student-specific

- What factors motivated you to pursue a data scientist role?

- Could you share your job interview experience? (Topics the interviewers expected you to know? depth or breadth? Coding skills? Software/programming language?

- Any other advice for students in their final year preparing for a data science job?