

Data Mining

Prof. Mauricio Araya
mauricio.araya@usm.cl
Of. B339 +56 32 2654207

Un par de preguntas

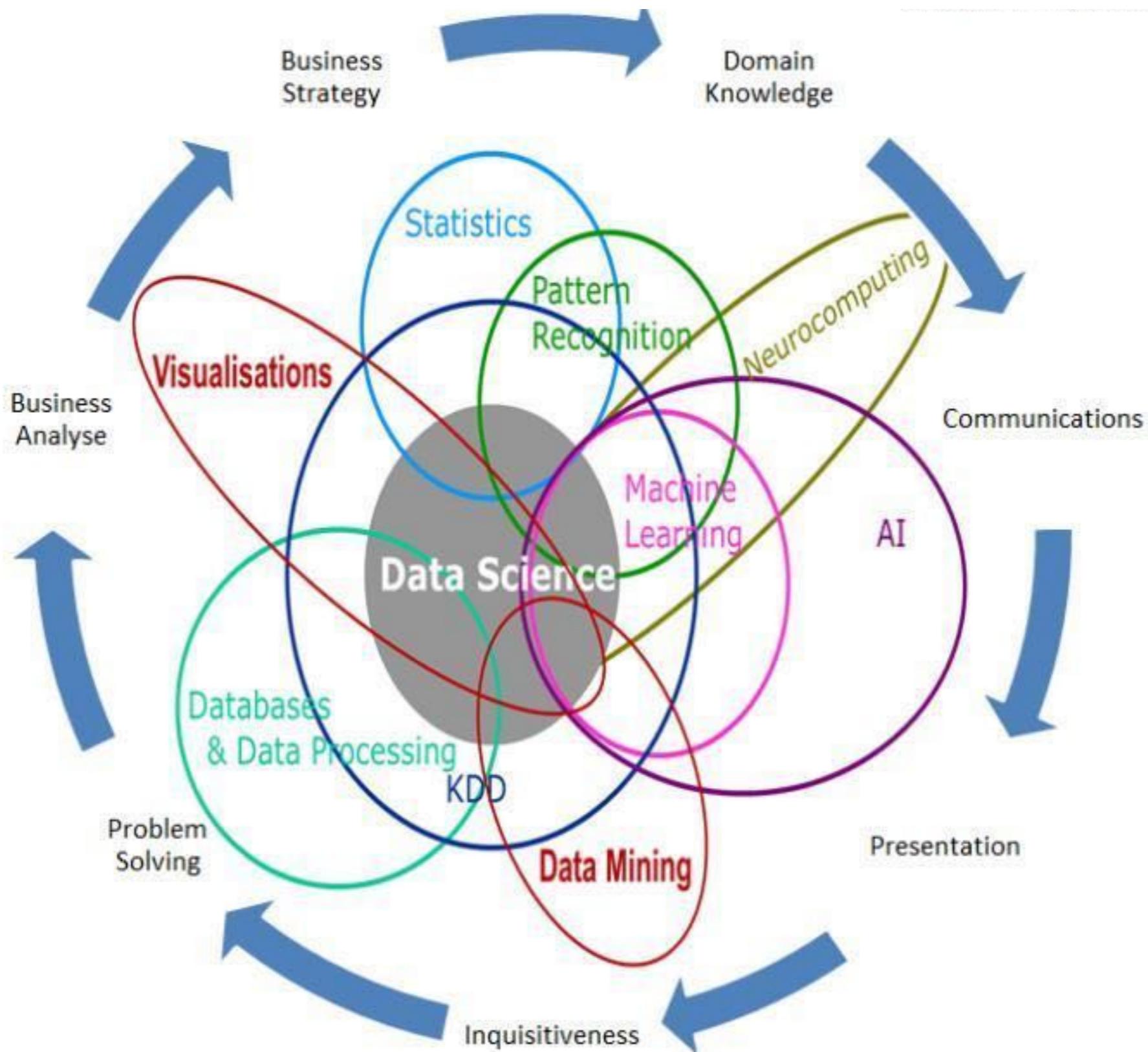


<https://www.menti.com/>

38 08 31

<https://www.menti.com/ep1guqopvx>

Data Science



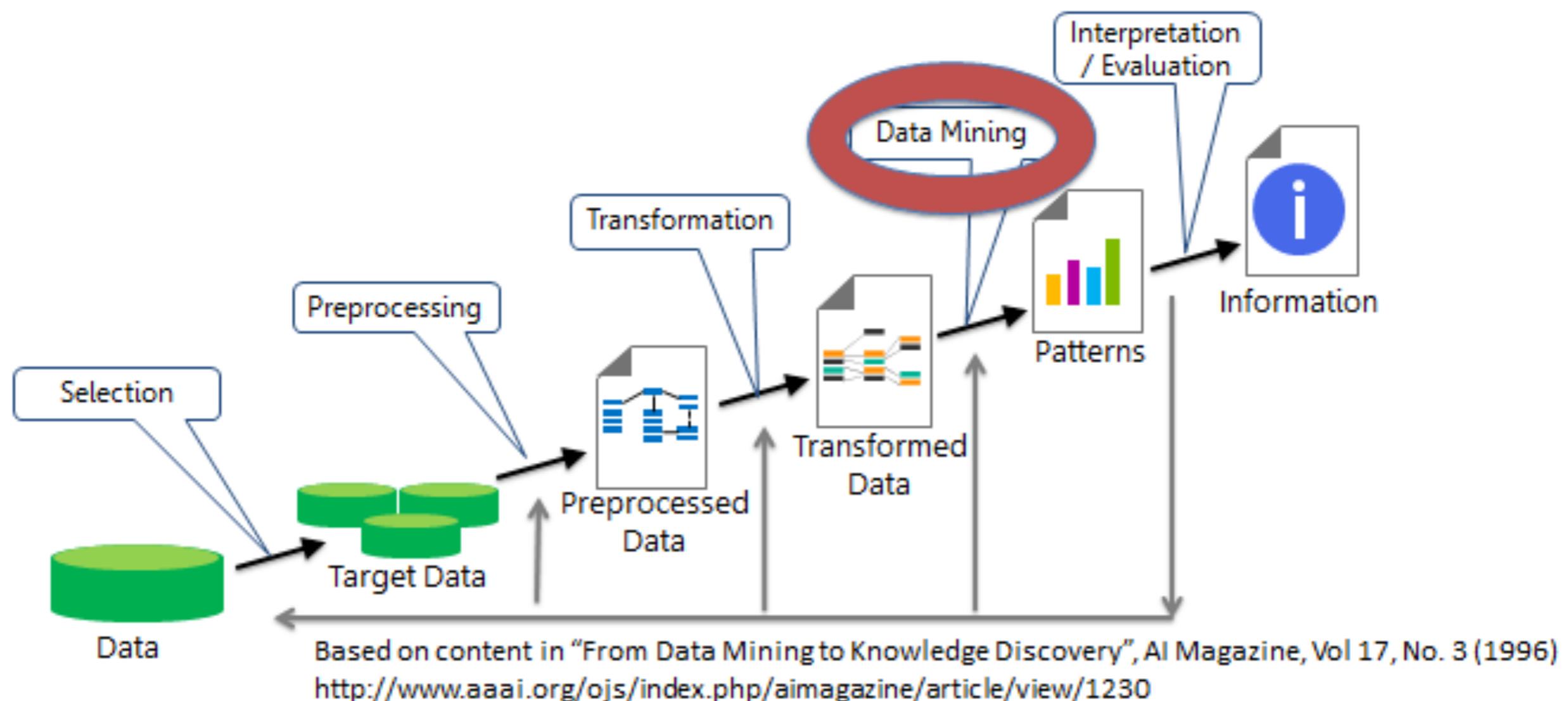
Data Mining (DM)



**Useful
knowledge
(patterns)**

```
x = np.array([10.9, 5.34, 8.32, 12.43, 20.32, 7.24])  
y = np.array([True, False, False, True, True, False])
```

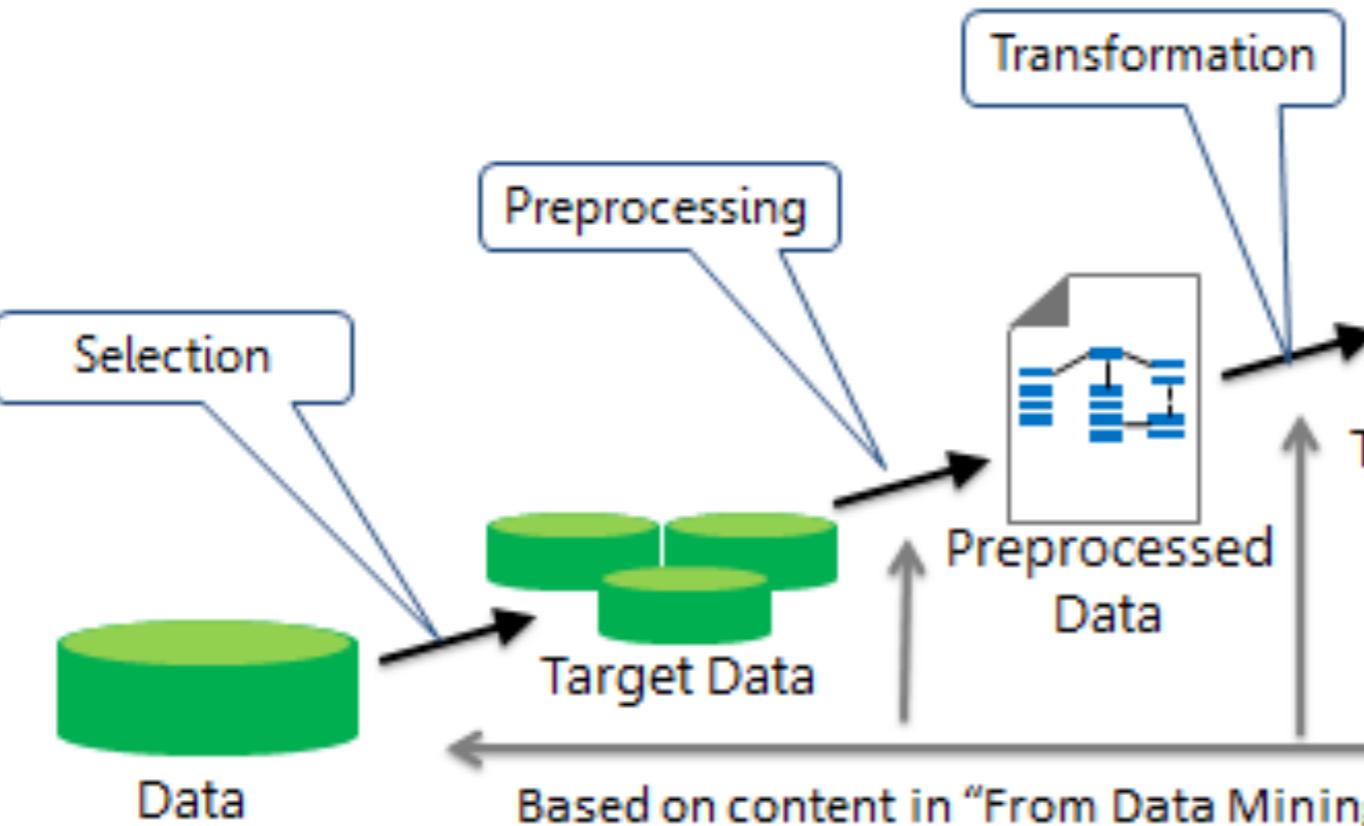
Data Mining & Data Science



Data Mining & Big Data

BIG DATA IS LIKE TEENAGE SEX: EVERYONE TALKS ABOUT IT, NOBODY REALLY KNOWS HOW TO DO IT, EVERYONE THINKS EVERYONE ELSE IS DOING IT, SO EVERYONE CLAIMS THEY ARE DOING IT...

DAN ARIELY



WE'VE DECIDED
TO TAKE BIG
DATA TO THE
NEXT LEVEL...



© D.Fletcher for CloudTweaks.com

Data Mining & Big Data

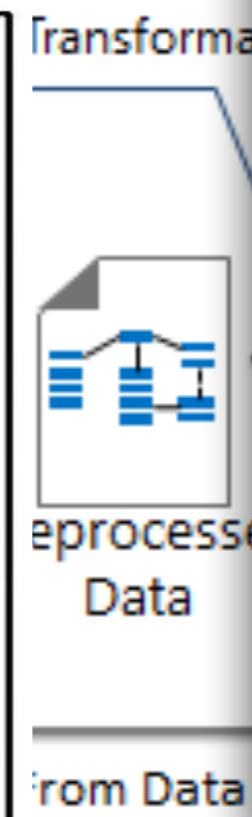
BIG DATA IS LIKE TEENAGE SEX: EVERYONE TALKS ABOUT IT, NOBODY REALLY KNOWS HOW TO DO IT, EVERYONE THINKS EVERYONE ELSE IS DOING IT, SO EVERYONE CLAIMS THEY ARE DOING IT...

DAN ARIELY

WE'VE DECIDED TO TAKE BIG DATA TO THE NEXT LEVEL...



© D.Fletcher for CloudTweaks.com



[index.php/aimagazine/article/view/1230](#)

[1. A NEW PARADIGM FOR BIG DATA](#)

PART 1 BATCH LAYER

[2. DATA MODEL FOR BIG DATA](#)

[3. DATA MODEL FOR BIG DATA: ILLUSTRATION](#)

[4. DATA STORAGE ON THE BATCH LAYER](#)

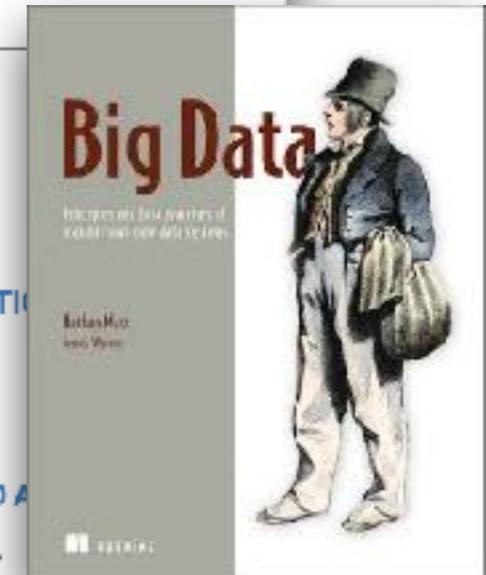
[5. DATA STORAGE ON THE BATCH LAYER: ILLUSTRATION](#)

[6. BATCH LAYER](#)

[7. BATCH LAYER: ILLUSTRATION](#)

[8. AN EXAMPLE BATCH LAYER: ARCHITECTURE AND A](#)

[9. AN EXAMPLE BATCH LAYER: IMPLEMENTATION](#)



PART 2 SERVING LAYER

[10. SERVING LAYER](#)

[11. SERVING LAYER: ILLUSTRATION](#)

PART 3 SPEED LAYER

[12. REALTIME VIEWS](#)

[13. REALTIME VIEWS: ILLUSTRATION](#)

[14. QUEUING AND STREAM PROCESSING](#)

[15. QUEUING AND STREAM PROCESSING: ILLUSTRATION](#)

[16. MICRO-BATCH STREAM PROCESSING](#)

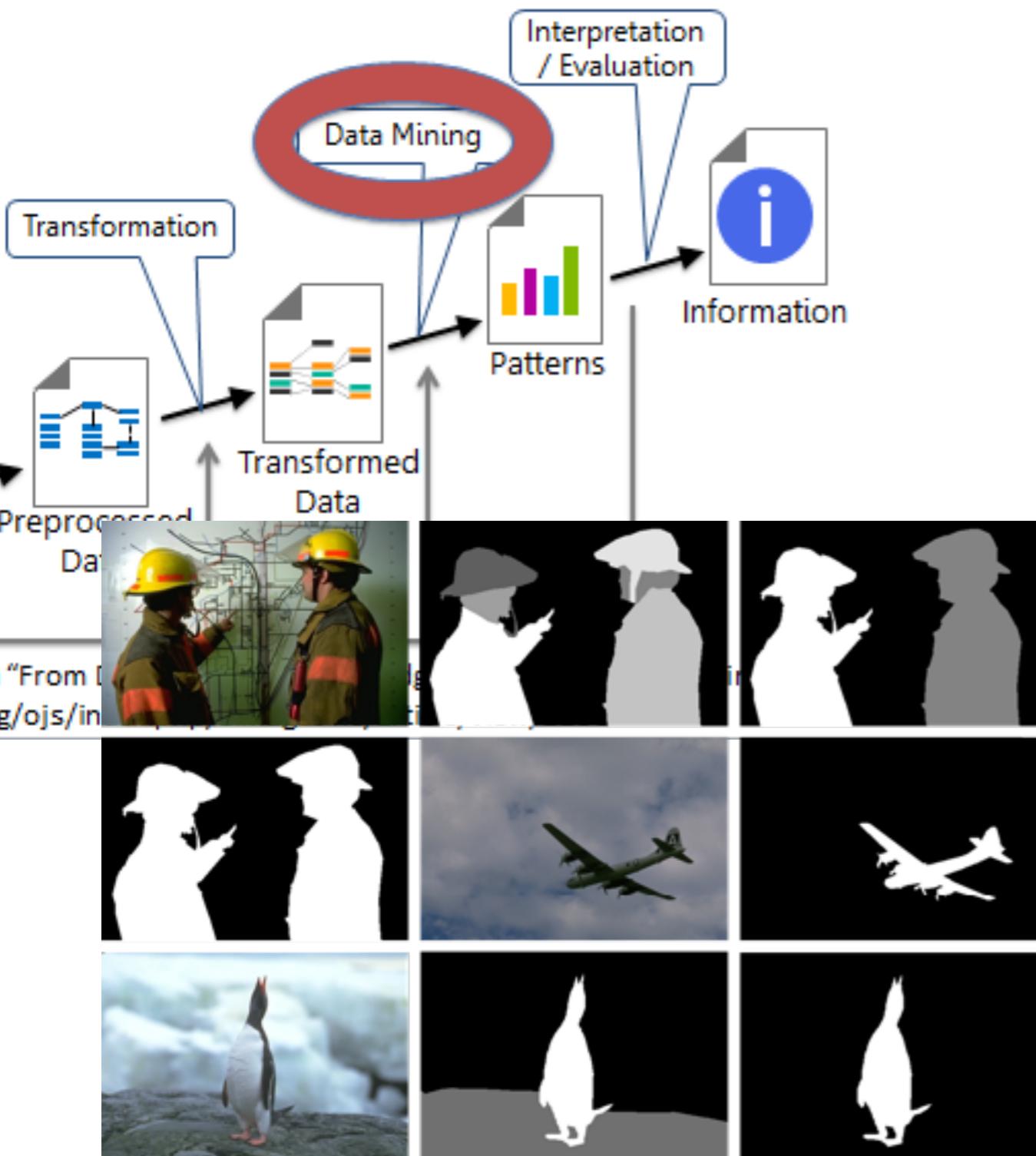
[17. MICRO-BATCH STREAM PROCESSING: ILLUSTRATION](#)

[18. LAMBDA ARCHITECTURE IN DEPTH](#)

[INDEX](#)

/vol 17, No. 3 (19

Pattern Recognition (PR)



$$f(x) = \begin{cases} True & \text{if } x > 10 \\ False & \text{else} \end{cases}$$



PR + DM

$$f(x) = \begin{cases} True & \text{if } x > 10 \\ False & \text{else} \end{cases}$$

```
def reactive_agent(x):
    if x > 10.0:
        return True
    else:
        return False
```

```
x = np.array([10.9, 5.34, 8.32, 12.43, 20.32, 7.24])
y = np.array([True, False, False, True, True, False])
```



When information overload occurs,
pattern recognition is how to
determine truth.

— Marshall McLuhan —

AZ QUOTES

Machine Learning (ML)

```
x = np.array([10.9, 5.34, 8.32, 12.43, 20.32, 7.24])  
y = np.array([True, False, False, True, True, False])
```

~~def reactive_agent():
 if x > 10:
 return True
 else:
 return False~~

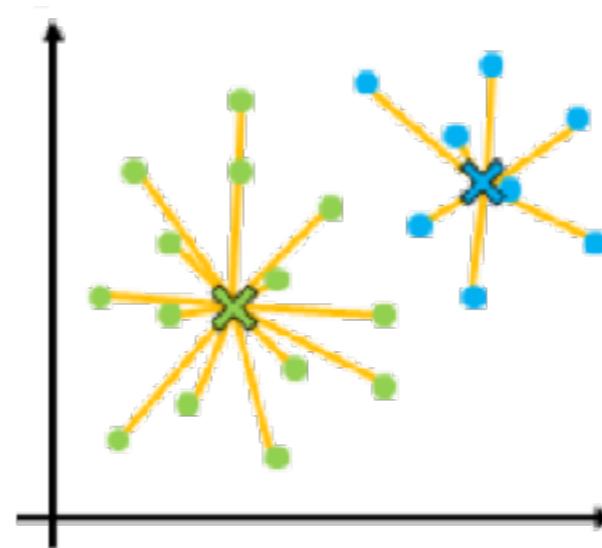
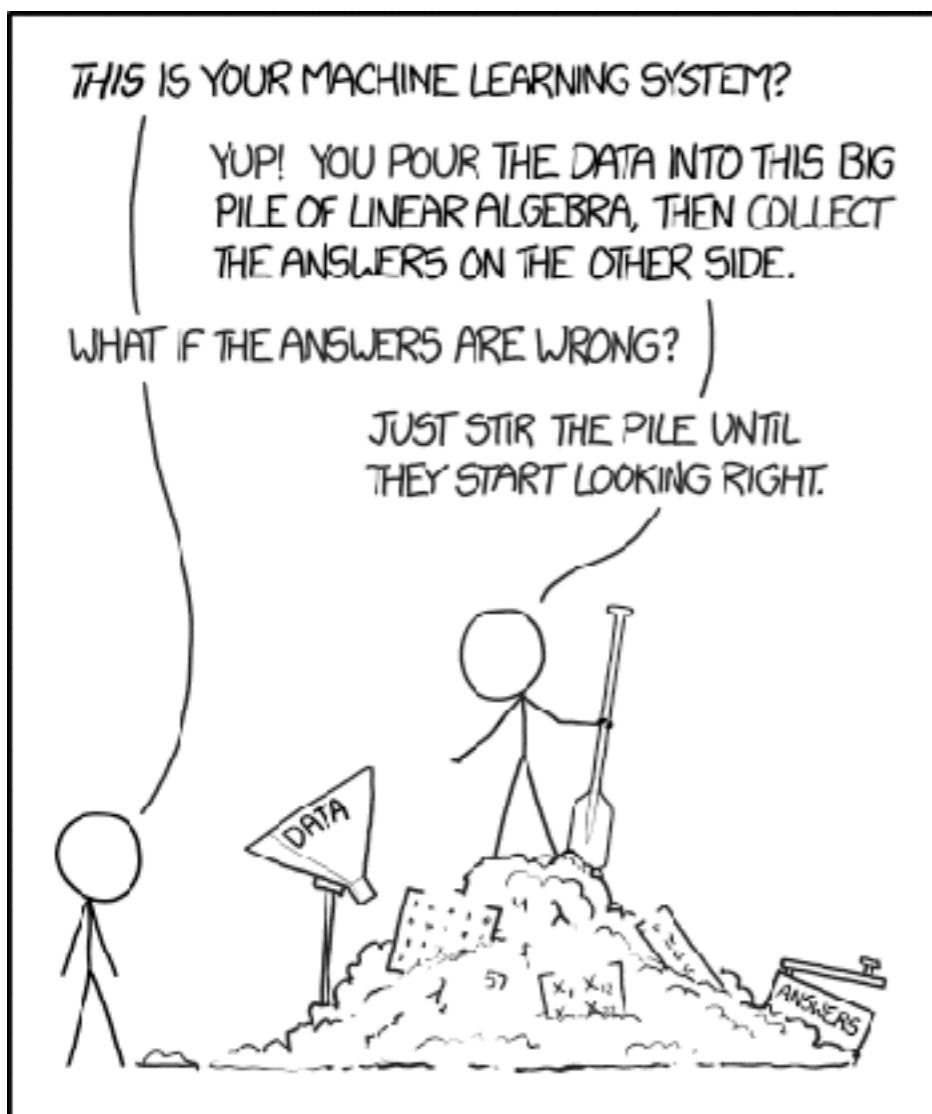


“...what we want is a machine that can learn from experience.

Alan Turing, 1947

Machine Learning (ML)

```
x = np.array([10.9, 5.34, 8.32, 12.43, 20.32, 7.24])  
y = np.array([True, False, False, True, True, False])
```



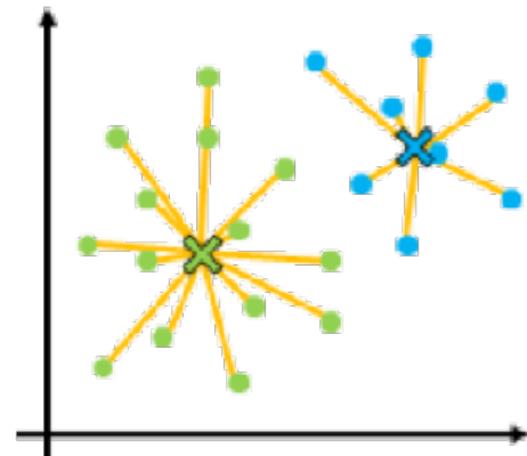
$$\mu_T = \sum_{x \in T} \frac{x}{|T|}$$

$$\mu_F = \sum_{x \in F} \frac{x}{|F|}$$

$$f(x) = \begin{cases} T & \text{if } |\mu_T - x| < |\mu_F - x| \\ F & \text{else} \end{cases}$$

Machine Learning (ML)

```
x = np.array([10.9, 5.34, 8.32, 12.43, 20.32, 7.24])  
y = np.array([True, False, False, True, True, False])
```



$$\mu_T = \sum_{x \in T} \frac{x}{|T|}$$
$$\mu_F = \sum_{x \in F} \frac{x}{|F|}$$

$$f(x) = \begin{cases} T & \text{if } |\mu_T - x| < |\mu_F - x| \\ F & \text{else} \end{cases}$$

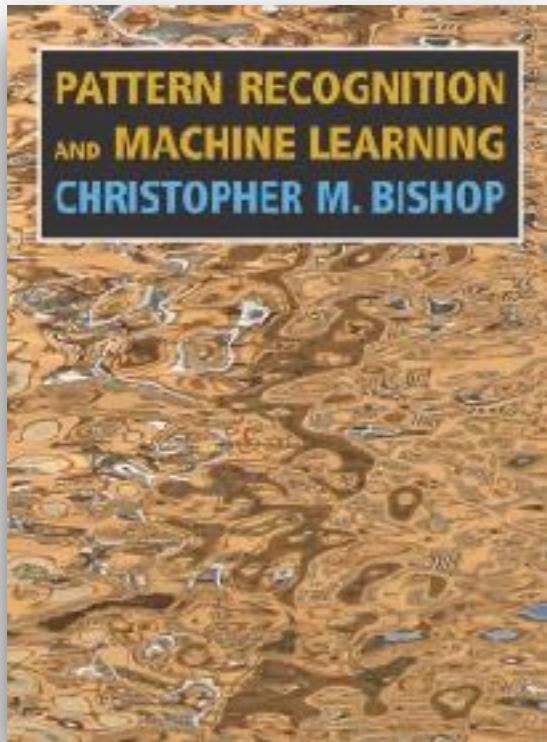
$$S = \frac{\mu_T + \mu_F}{2}$$

```
def learning_agent(x, Data, labels):  
    v_true = np.mean(Data[labels==True])  
    v_false = np.mean(Data[labels==False])  
    d_true = np.abs(x - v_true)  
    d_false = np.abs(x - v_false)  
    if d_true < d_false:  
        return True  
    else:  
        return False
```

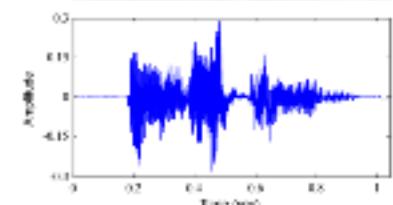
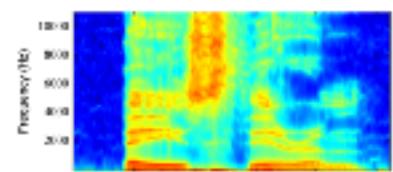
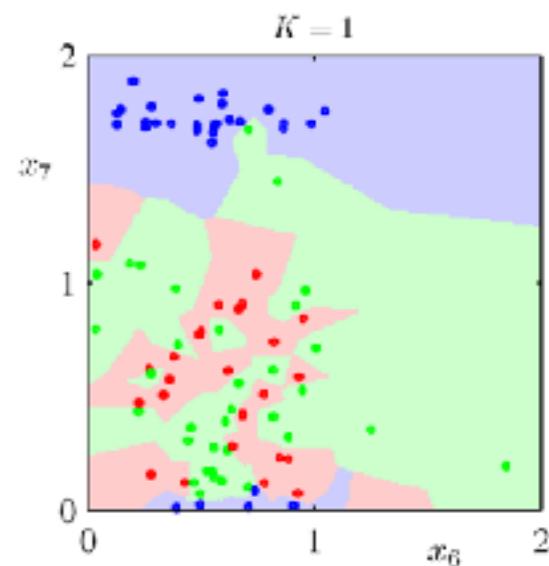
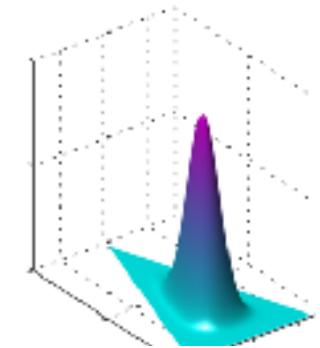
$$(v_true + v_false)/2$$

$$10.758333333333333$$

Machine Learning (ML)



Introduction.pdf	Probability Distributions.pdf	Linear Models for Regression.pdf	Linear Models for Classification.pdf
Neural Networks.pdf	Kernel Methods.pdf	Sparse Kernel Machines.pdf	Graphical Models.pdf
Mixture Models and EM.pdf	Approximate Inference.pdf	Sampling Methods.pdf	Continuous Latent Variables.pdf
Sequential Data.pdf	Combining Models.pdf	back-matter.pdf	front-matter.pdf



b	sp	z	th	it	en
Quartz				Theorem	

ML + PR + DM



More data beats clever algorithms,
but better data beats more data.

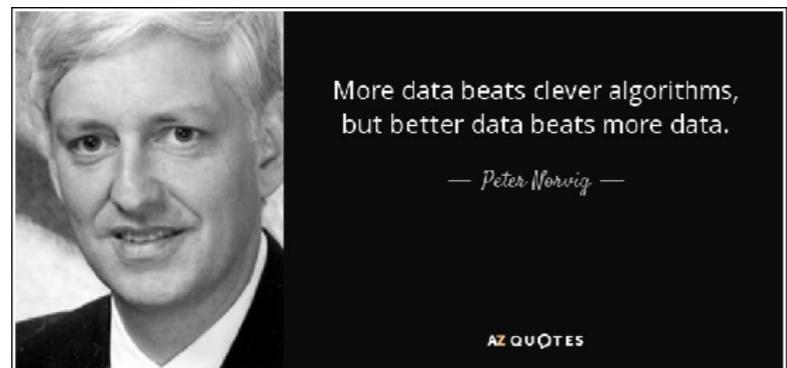
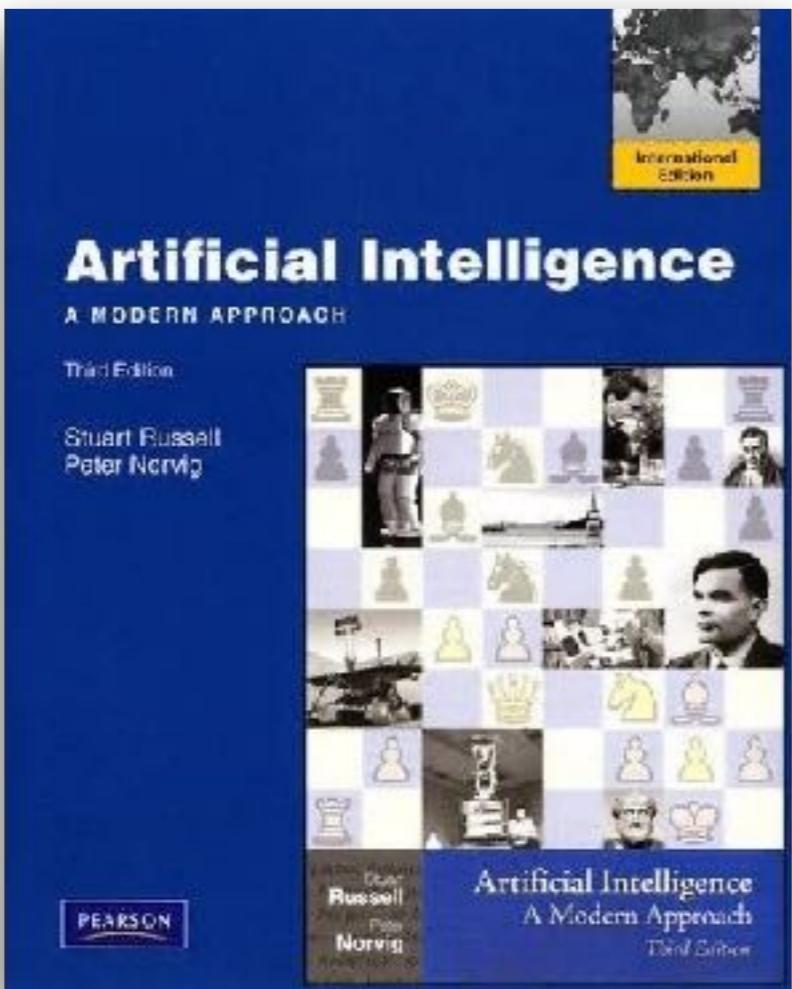
— Peter Norvig —

AZ QUOTES

*"Torture the data, and it will
confess to anything."*

**Ronald Coase, winner of the
Nobel Prize in Economics**

What About AI?



Part I Artificial Intelligence

- 1 Introduction ... 1
- 2 Intelligent Agents ... 34

Part II Problem Solving

- 3 Solving Problems by Searching ... 64
- 4 Beyond Classical Search ... 120
- 5 Adversarial Search ... 161
- 6 Constraint Satisfaction Problems ... 202

Part III Knowledge and Reasoning

- 7 Logical Agents ... 234
- 8 First-Order Logic ... 285
- 9 Inference in First-Order Logic ... 322
- 10 Classical Planning ... 366
- 11 Planning and Acting in the Real World ... 401
- 12 Knowledge Representation ... 437

Part IV Uncertain Knowledge and Reasoning

- 13 Quantifying Uncertainty ... 480
- 14 Probabilistic Reasoning ... 510
- 15 Probabilistic Reasoning over Time ... 566
- 16 Making Simple Decisions ... 610
- 17 Making Complex Decisions ... 645

Part V Learning

- 18 Learning from Examples ... 693
- 19 Knowledge in Learning ... 768
- 20 Learning Probabilistic Models ... 802
- 21 Reinforcement Learning ... 830

Part VII Communicating, Perceiving, and Acting

- 22 Natural Language Processing ... 860
- 23 Natural Language for Communication ... 888
- 24 Perception ... 928
- 25 Robotics ... 971

Data

Data



M. Nauer

Age: 31

Nat: Germany

Value: €61M

Overall: 92

L. Suárez

Age: 30

Nat: Uruguay

Value: €97M

Overall: 92

Neymar Jr.

Age: 25

Nat: Brazil

Value: €123M

Overall: 92

L. Messi

Age: 30

Nat: Argentina

Value: €105M

Overall: 93

C. Ronaldo

Age: 32

Nat: Portugal

Value: €95.5M

Overall: 94

Data: Features and Samples

Features

Samples

	Name	Age	Nationality	Overall	Potential	Value	Photo
0	Cristiano Ronaldo	32	Portugal	94	94	€95.5M	https://cdn.sofifa.org/48/18/players/20801.png
1	L. Messi	30	Argentina	93	93	€105M	https://cdn.sofifa.org/48/18/players/158023.png
2	Neymar	25	Brazil	92	94	€123M	https://cdn.sofifa.org/48/18/players/190871.png
3	L. Suárez	30	Uruguay	92	92	€97M	https://cdn.sofifa.org/48/18/players/176580.png
4	M. Neuer	31	Germany	92	92	€61M	https://cdn.sofifa.org/48/18/players/167495.png
5	R. Lewandowski	28	Poland	91	91	€92M	https://cdn.sofifa.org/48/18/players/188545.png
6	De Gea	26	Spain	90	92	€64.5M	https://cdn.sofifa.org/48/18/players/193080.png
7	E. Hazard	26	Belgium	90	91	€90.5M	https://cdn.sofifa.org/48/18/players/183277.png
8	T. Kroos	27	Germany	90	90	€79M	https://cdn.sofifa.org/48/18/players/182521.png
9	G. Higuaín	29	Argentina	90	90	€77M	https://cdn.sofifa.org/48/18/players/167664.png
10	Sergio Ramos	31	Spain	90	90	€52M	https://cdn.sofifa.org/48/18/players/155862.png
11	K. De Bruyne	26	Belgium	89	92	€83M	https://cdn.sofifa.org/48/18/players/192985.png
12	T. Courtois	25	Belgium	89	92	€59M	https://cdn.sofifa.org/48/18/players/192119.png
13	A. Sánchez	28	Chile	89	89	€67.5M	https://cdn.sofifa.org/48/18/players/184941.png
14	L. Modrić	31	Croatia	89	89	€57M	https://cdn.sofifa.org/48/18/players/177003.png
15	G. Bale	27	Wales	89	89	€69.5M	https://cdn.sofifa.org/48/18/players/173731.png

Text

Int

Cat

Per

Float

Complex

TYPES

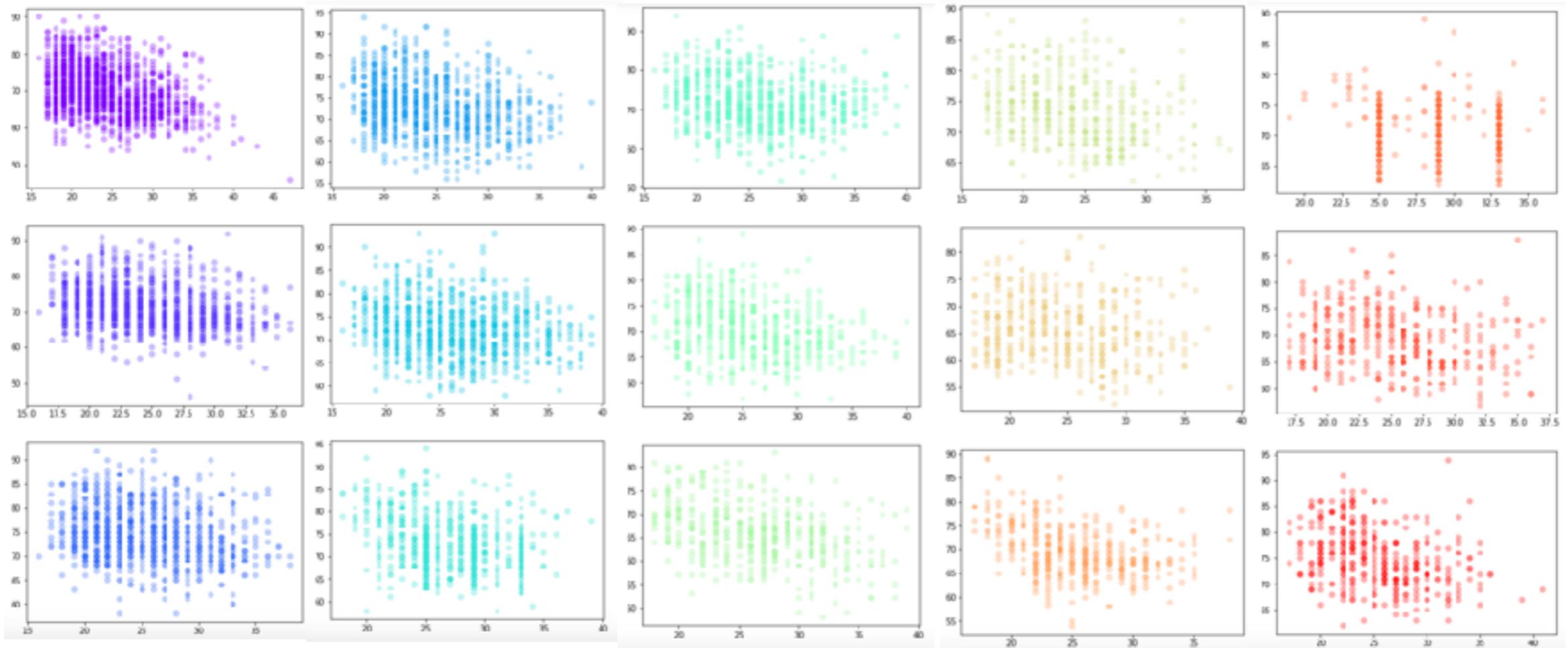
Samples

Features

	Name	Age	Nationality	Overall	Potential	Value	Photo
0	Cristiano Ronaldo	32	Portugal	94	94	€95.5M	https://cdn.sofifa.org/48/18/players/20801.png
1	L. Messi	30	Argentina	93	93	€105M	https://cdn.sofifa.org/48/18/players/158023.png
2	Neymar	25	Brazil	92	94	€123M	https://cdn.sofifa.org/48/18/players/190871.png
3	L. Suárez	30	Uruguay	92	92	€97M	https://cdn.sofifa.org/48/18/players/176580.png
4	M. Neuer	31	Germany	92	92	€61M	https://cdn.sofifa.org/48/18/players/167495.png
5	R. Lewandowski	28	Poland	91	91	€92M	https://cdn.sofifa.org/48/18/players/188545.png
6	D. De Gea	26	Spain	90	92	€64.5M	https://cdn.sofifa.org/48/18/players/193080.png
7	E. Hazard	26	Belgium	90	91	€90.5M	https://cdn.sofifa.org/48/18/players/183277.png
8	T. Kroos	27	Germany	90	90	€79M	https://cdn.sofifa.org/48/18/players/182521.png
9	G. Higuaín	29	Argentina	90	90	€77M	https://cdn.sofifa.org/48/18/players/167664.png
10	Sergio Ramos	31	Spain	90	90	€52M	https://cdn.sofifa.org/48/18/players/155862.png
11	K. De Bruyne	26	Belgium	89	92	€83M	https://cdn.sofifa.org/48/18/players/192985.png
12	T. Courtois	25	Belgium	89	92	€59M	https://cdn.sofifa.org/48/18/players/192119.png
13	A. Sánchez	28	Chile	89	89	€67.5M	https://cdn.sofifa.org/48/18/players/184941.png
14	L. Modrić	31	Croatia	89	89	€57M	https://cdn.sofifa.org/48/18/players/177003.png
15	G. Bale	27	Wales	89	89	€69.5M	https://cdn.sofifa.org/48/18/players/173731.png

England
Germany
Spain
France
Argentina
Brazil
Italy
Colombia
Japan

Netherlands
Republic of Ireland
United States
Chile
Sweden
Portugal



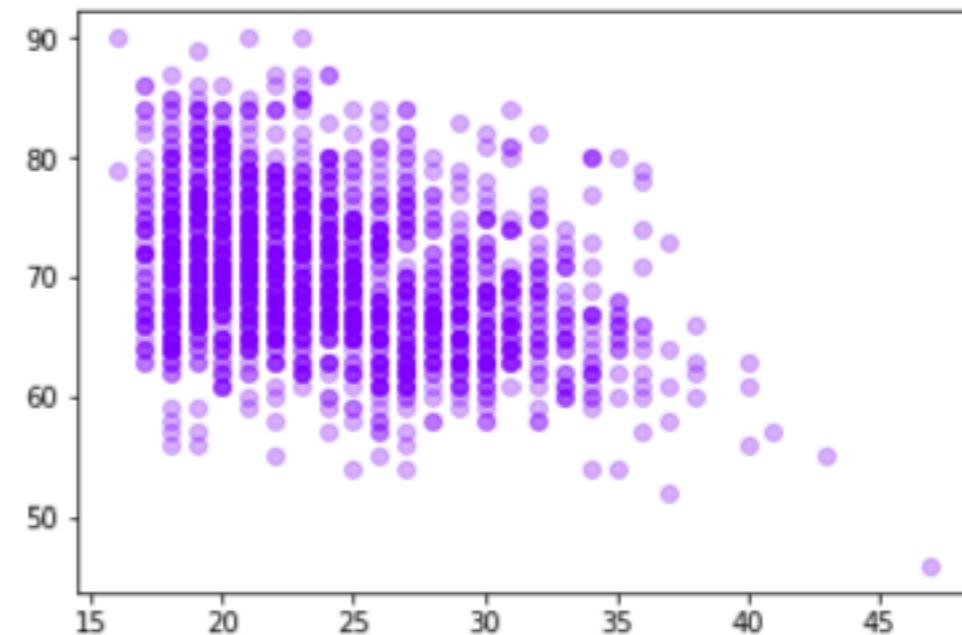
Potential vs Age

Samples

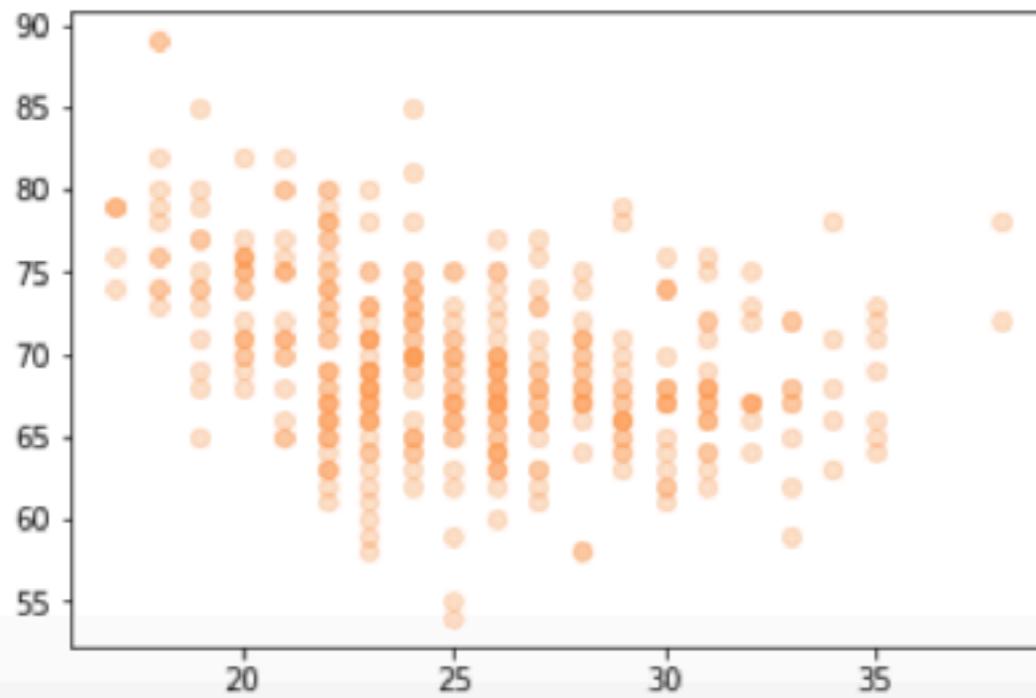
Features

	Name	Age	Nationality	Overall	Potential	Value	Photo
0	Cristiano Ronaldo	32	Portugal	94	94	€95.5M	https://cdn.sofifa.org/48/18/players/20801.png
1	L. Messi	30	Argentina	93	93	€105M	https://cdn.sofifa.org/48/18/players/158023.png
2	Neymar	25	Brazil	92	94	€123M	https://cdn.sofifa.org/48/18/players/190871.png
3	L. Suárez	30	Uruguay	92	92	€97M	https://cdn.sofifa.org/48/18/players/176580.png
4	M. Neuer	31	Germany	92	92	€61M	https://cdn.sofifa.org/48/18/players/167495.png
5	R. Lewandowski	28	Poland	91	91	€92M	https://cdn.sofifa.org/48/18/players/188545.png
6	D. De Gea	26	Spain	90	92	€64.5M	https://cdn.sofifa.org/48/18/players/193080.png
7	E. Hazard	26	Belgium	90	91	€90.5M	https://cdn.sofifa.org/48/18/players/183277.png
8	T. Kroos	27	Germany	90	90	€79M	https://cdn.sofifa.org/48/18/players/182521.png
9	G. Higuaín	29	Argentina	90	90	€77M	https://cdn.sofifa.org/48/18/players/167664.png
10	Sergio Ramos	31	Spain	90	90	€52M	https://cdn.sofifa.org/48/18/players/155862.png
11	K. De Bruyne	26	Belgium	89	92	€83M	https://cdn.sofifa.org/48/18/players/192985.png
12	T. Courtois	25	Belgium	89	92	€59M	https://cdn.sofifa.org/48/18/players/192119.png
13	A. Sánchez	28	Chile	89	89	€67.5M	https://cdn.sofifa.org/48/18/players/184941.png
14	L. Modrić	31	Croatia	89	89	€57M	https://cdn.sofifa.org/48/18/players/177003.png
15	G. Bale	27	Wales	89	89	€69.5M	https://cdn.sofifa.org/48/18/players/173731.png

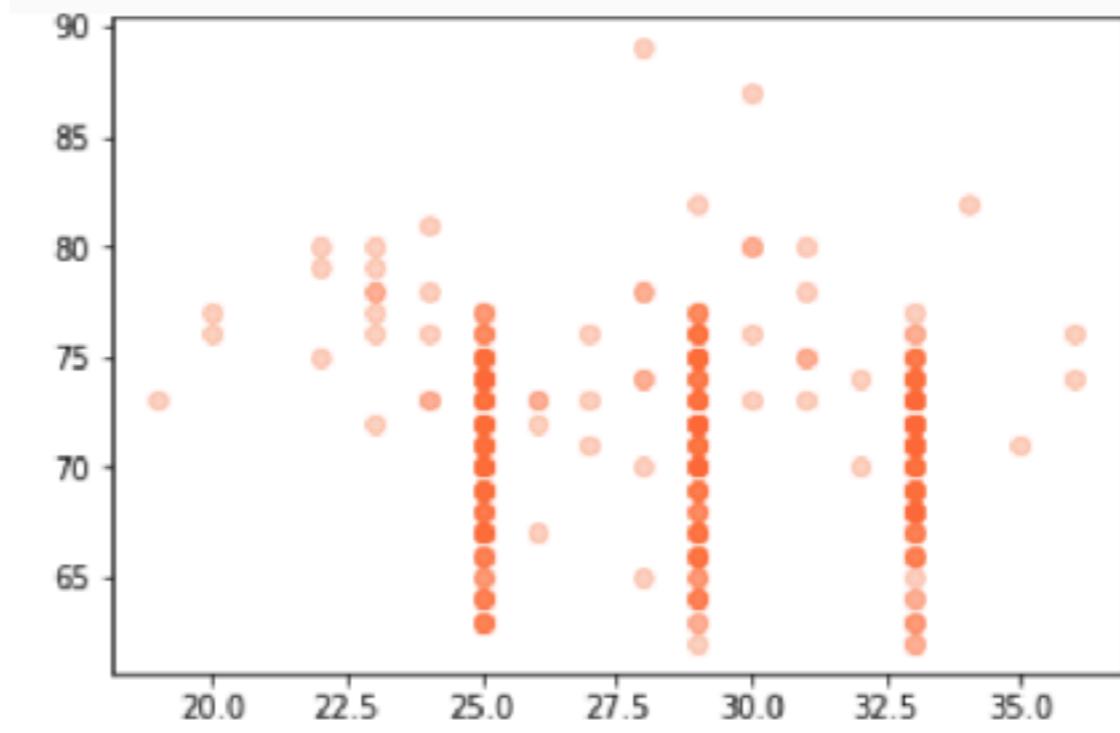
England



United States



Chile

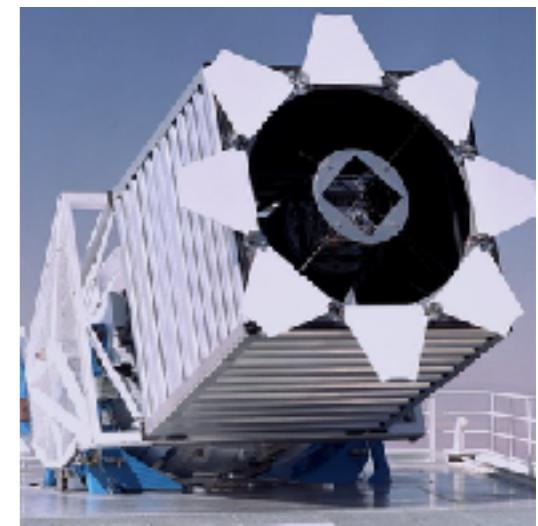
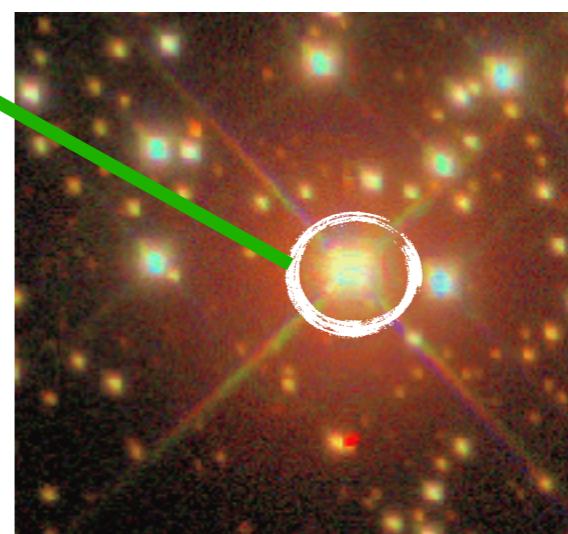
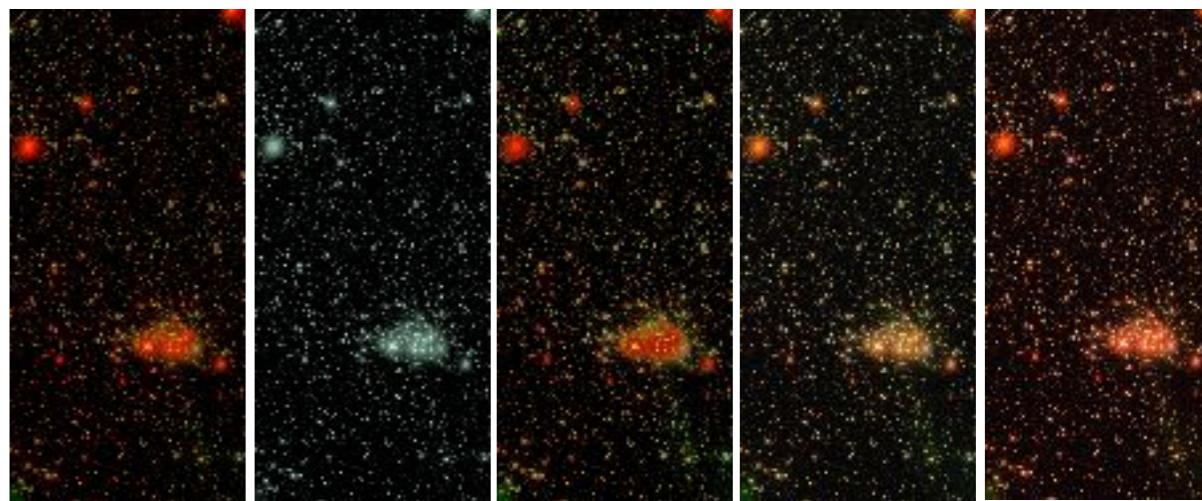
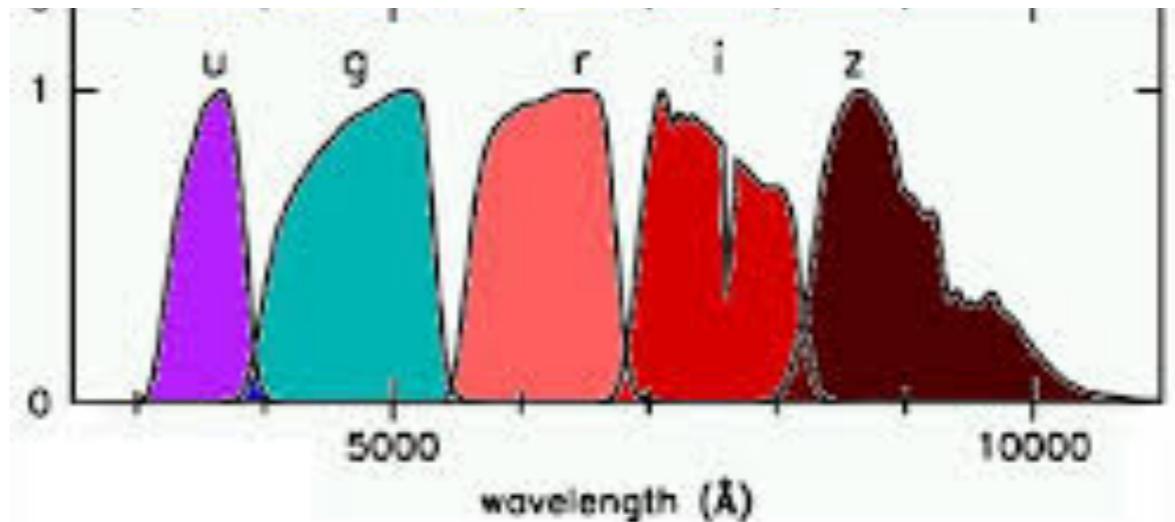


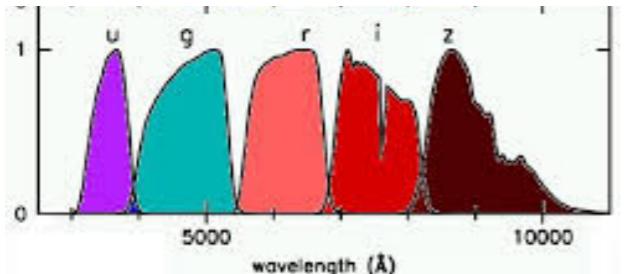
Real Stars

Samples

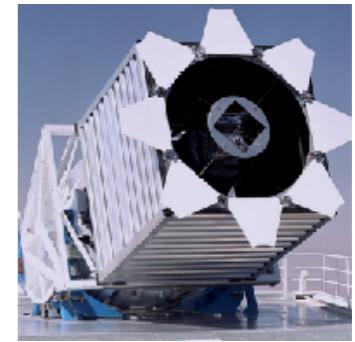
Features

	u-g	g-r	r-i	i-z
0	1.250999	0.394000	0.137000	0.061999
1	1.048000	0.339001	0.151999	0.023001
2	1.008001	0.341999	0.129000	0.203001
3	0.965000	0.392000	0.149000	0.150000
4	1.040001	0.333000	0.125999	0.101999
5	1.154001	0.373999	0.145000	0.121000
6	0.965000	0.384001	0.118999	0.011000
7	1.015001	0.370998	0.158001	0.091999
8	1.003000	0.391001	0.145000	0.074999
9	0.948000	0.330000	0.164000	0.021000
10	1.020000	0.389999	0.168001	0.070999





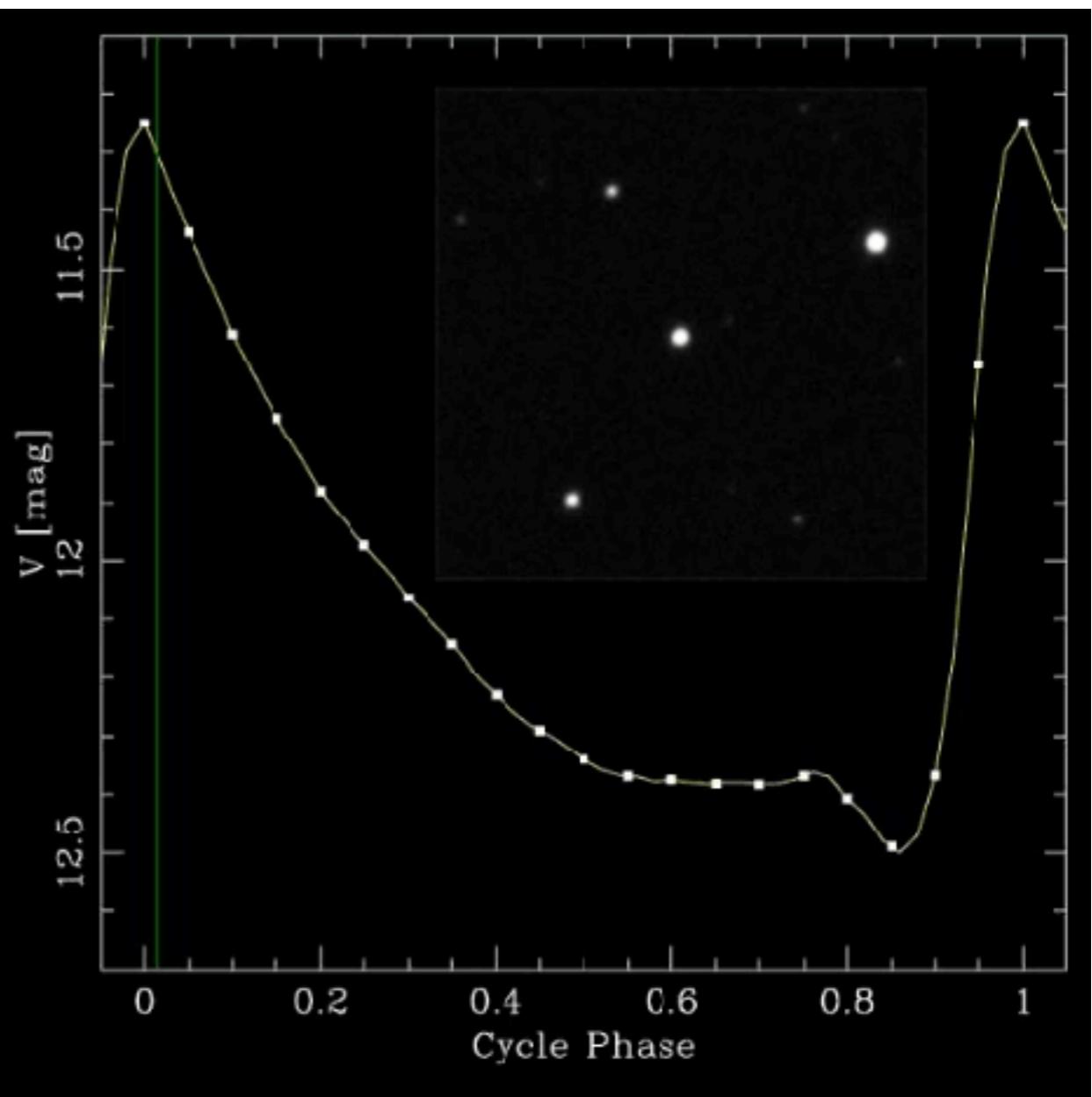
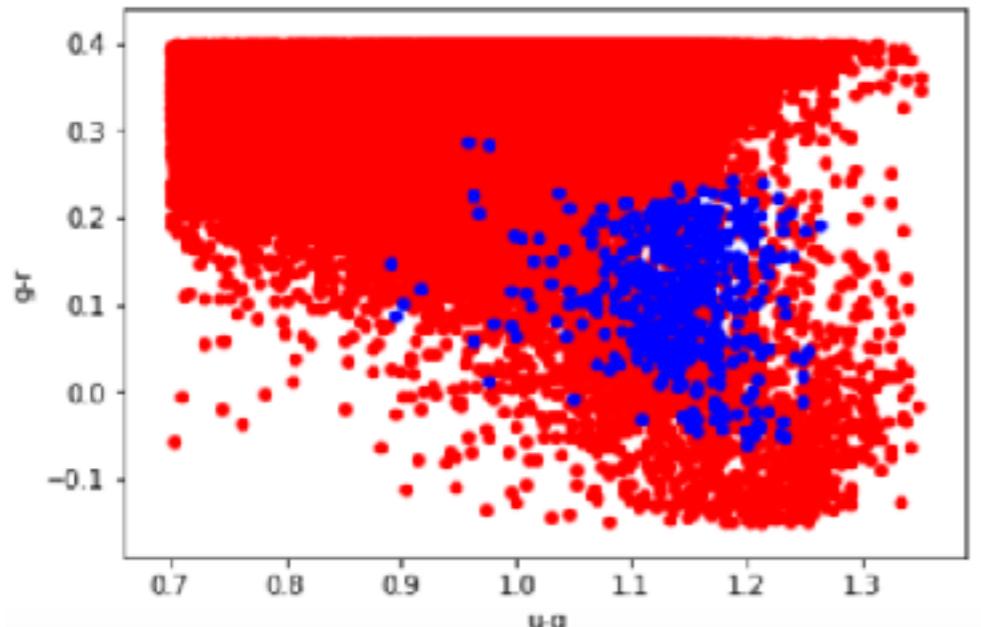
Real Stars



Samples

Features

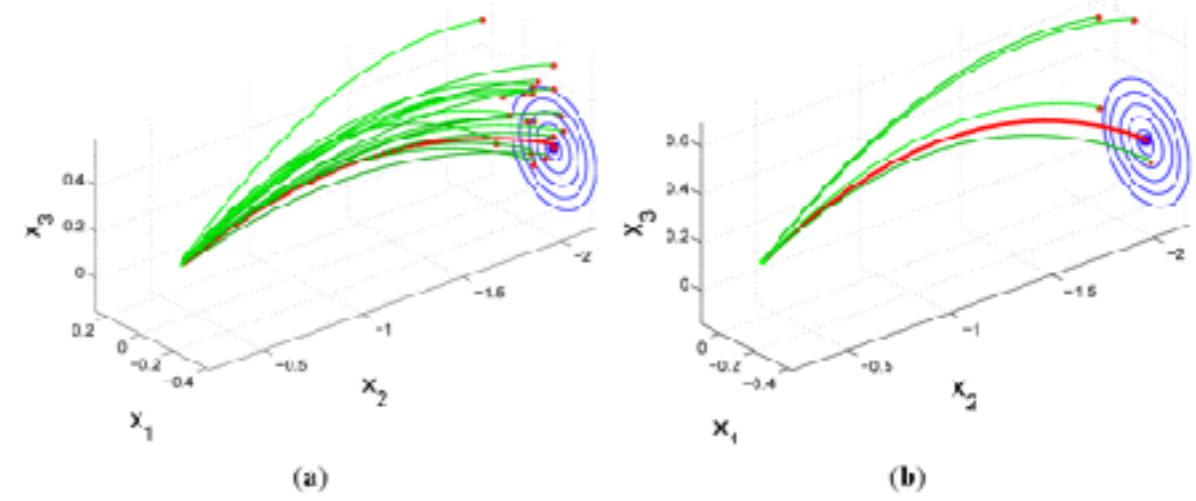
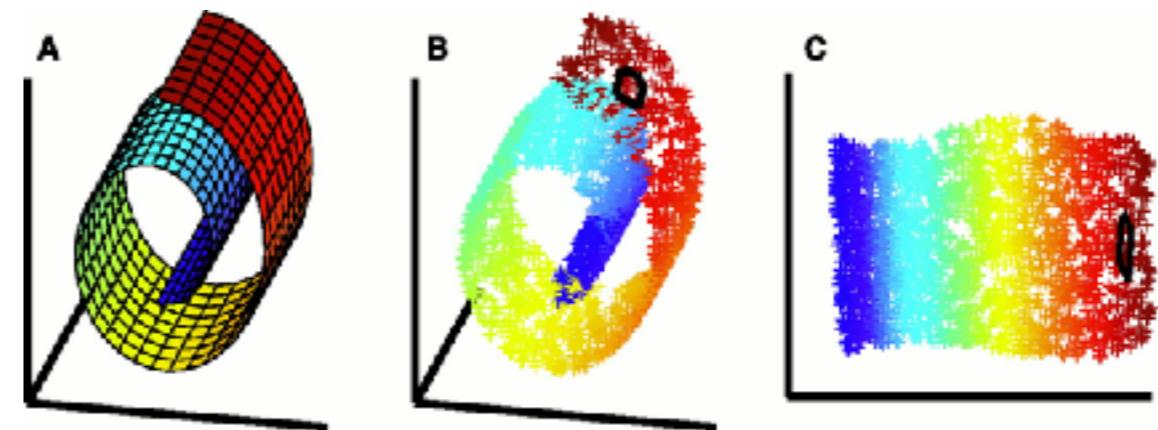
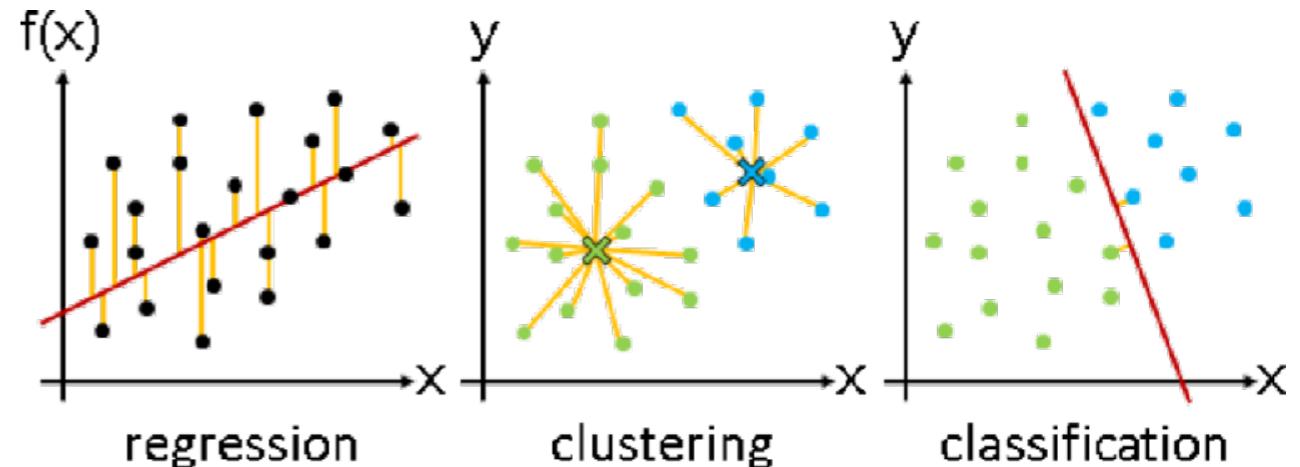
	u-g	g-r	r-i	i-z
0	1.250999	0.394000	0.137000	0.061999
1	1.048000	0.339001	0.151999	0.023001
2	1.008001	0.341999	0.129000	0.203001
3	0.965000	0.392000	0.149000	0.150000
4	1.040001	0.333000	0.125999	0.101999
5	1.154001	0.373999	0.145000	0.121000
6	0.965000	0.384001	0.118999	0.011000
7	1.015001	0.370998	0.158001	0.091999
8	1.003000	0.391001	0.145000	0.074999
9	0.948000	0.330000	0.164000	0.021000
10	1.020000	0.389999	0.168001	0.070999



Models and Methods

Machine Learning Tasks

- **Decisions** based on data
 - **Predicting**
 - regression, extrapolation, etc.
 - **Labelling**
 - classification, clustering, etc.
 - **Characterizing**
 - density estimation, model selection, dimensionality reduction, etc.
 - **Acting**
 - active learning, reinforcement learning, etc.



Machine Learning Paradigms

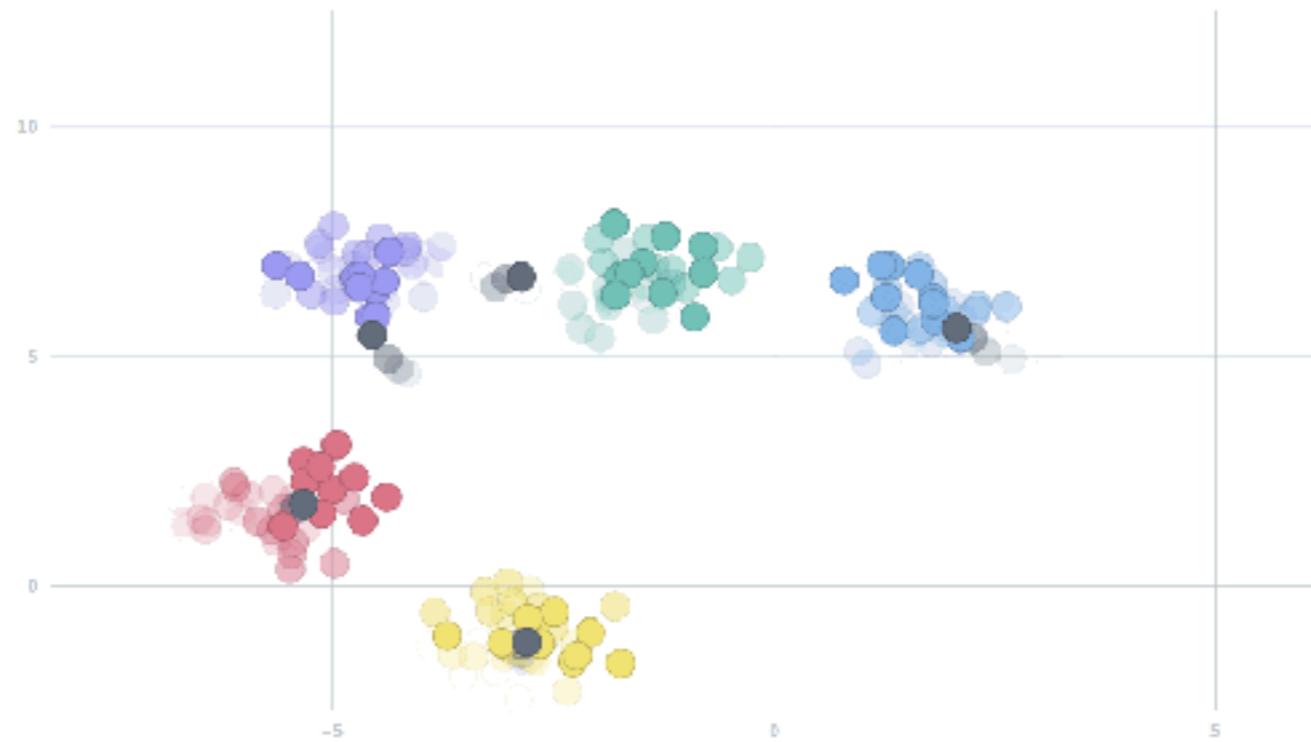
- **Task**

- Supervised vs Unsupervised
- Discrete vs Continuous
- Batch vs Sequential

	<i>Supervised Learning</i>	<i>Unsupervised Learning</i>
<i>Discrete</i>	classification or categorization	clustering
<i>Continuous</i>	regression	dimensionality reduction

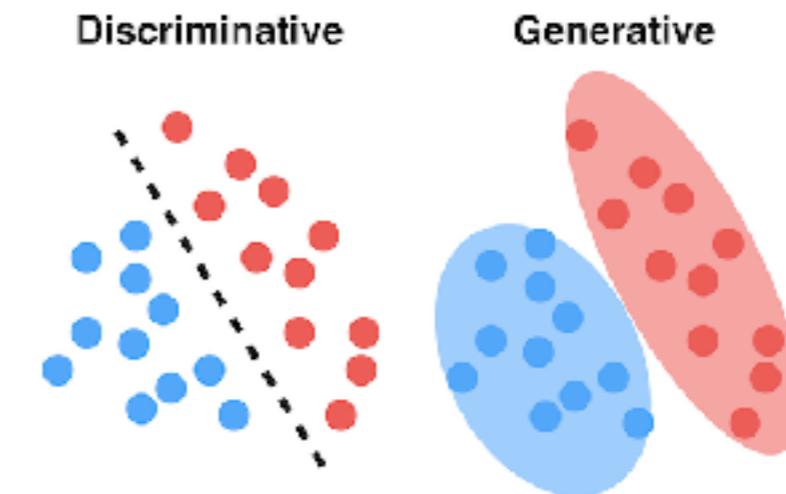
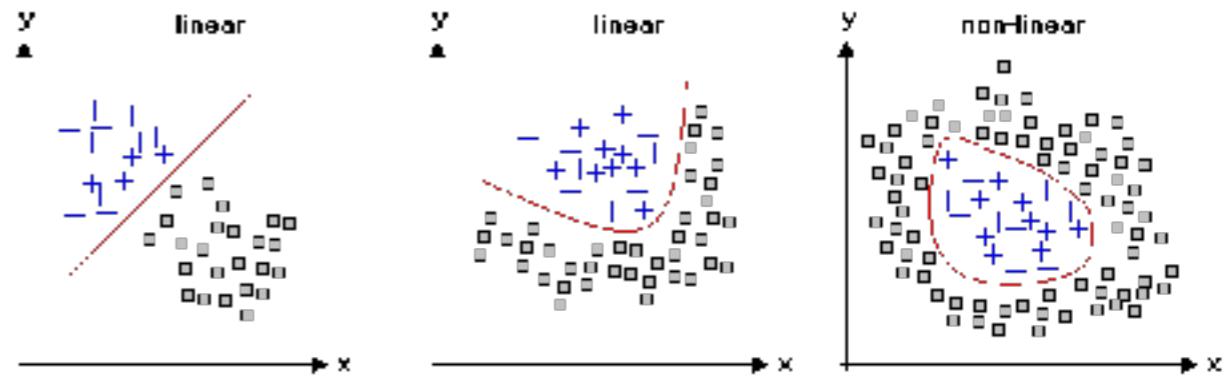
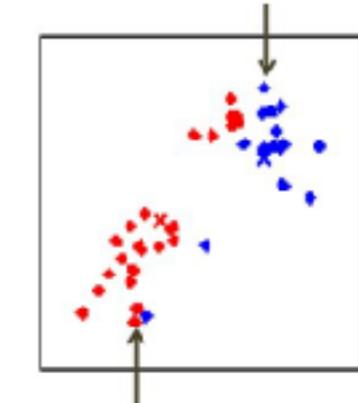
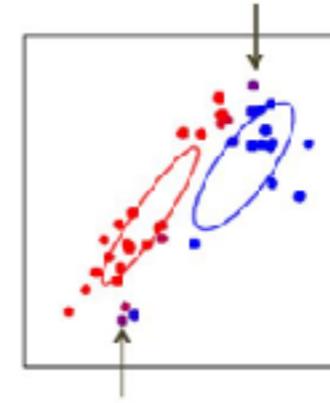
- **Models**

- Parametric vs Non-Parametric
- Linear vs Non-Linear
- Discriminative vs Generative



Machine Learning Paradigms

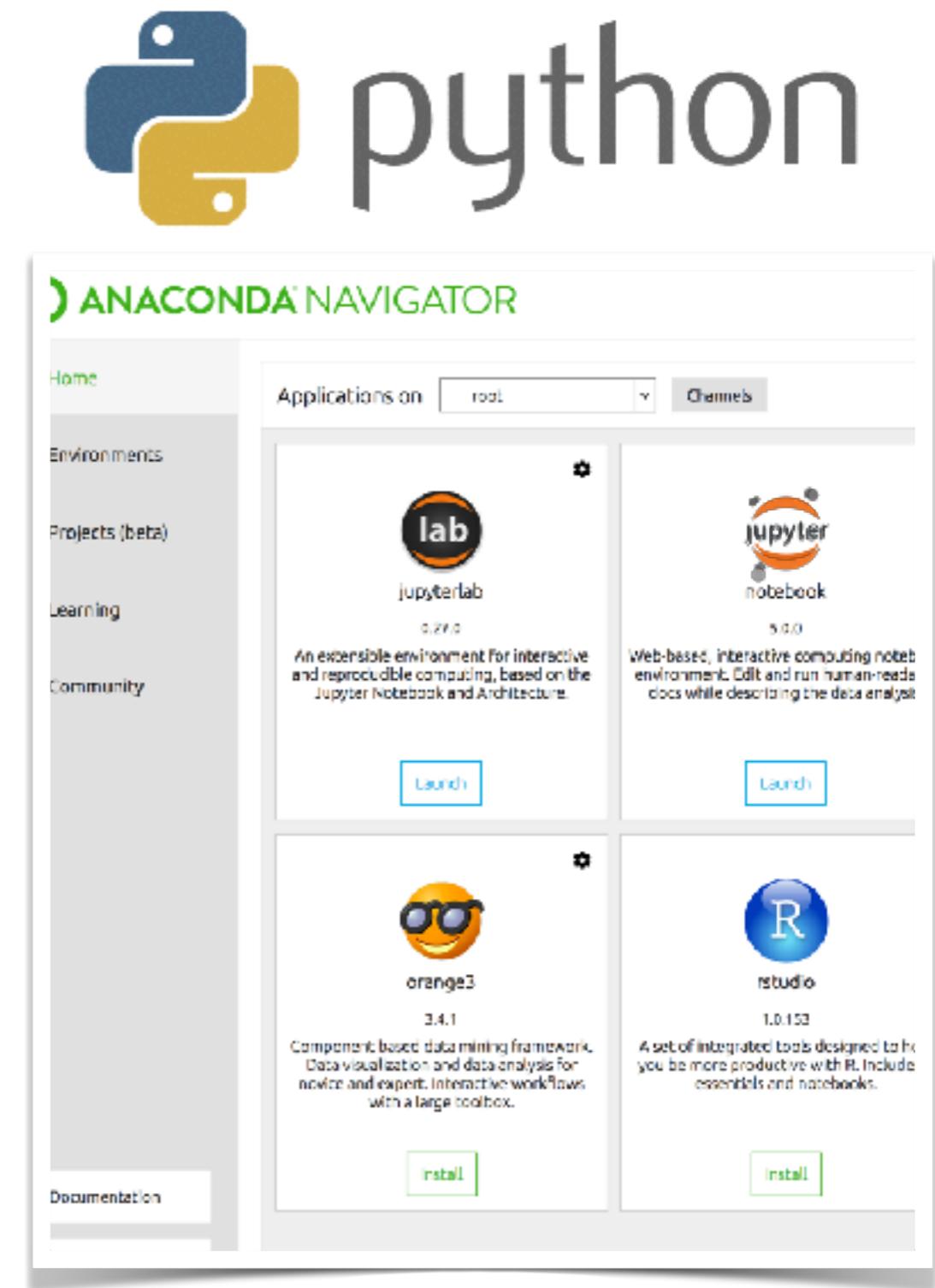
- **Task**
 - Supervised vs Unsupervised
 - Discrete vs Continuous
- Batch vs Sequential
- **Models**
 - Parametric vs Non-Parametric
 - Linear vs Non-Linear
 - Discriminative vs Generative



Jupyter

Installing Anaconda

- Jupyter runs over Python
- Anaconda is a Python Distribution
 - We will use Python 3.7
 - Environments can do the trick
 - pip at your own risk
- Over 500 MB
 - Windows, Linux and MacOS

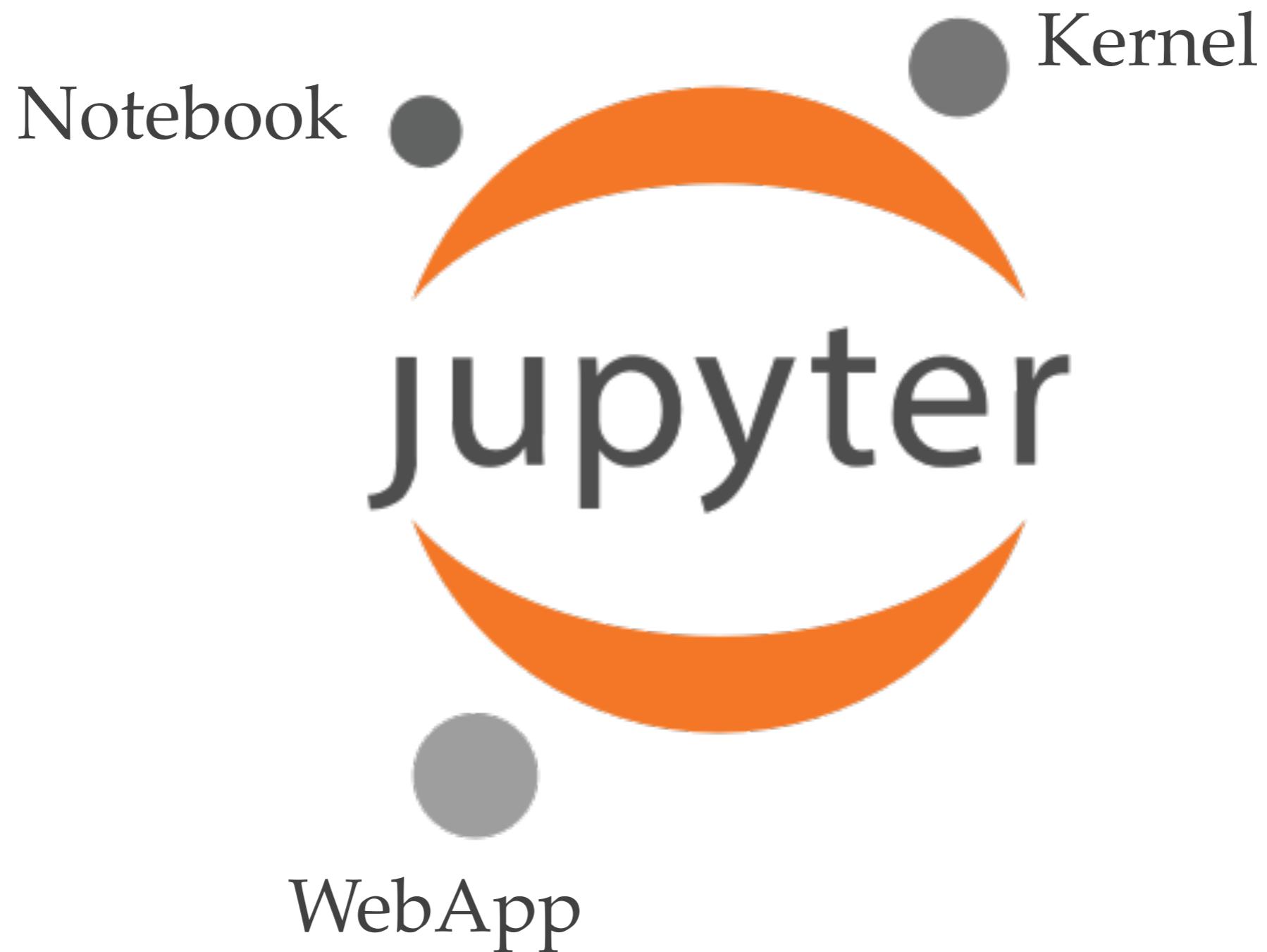


Setup an environment

- Fresh Installation:
 - `conda create --name dm21 --clone root`
- Previous Installation:
 - `conda create -n dm21 python=3.6 anaconda`
- Activate the environment
 - `source activate dm21`
- Move to a working directory
 - `mkdir directory`
 - `cd directory`
- Run Jupyter!
 - `jupyter notebook`



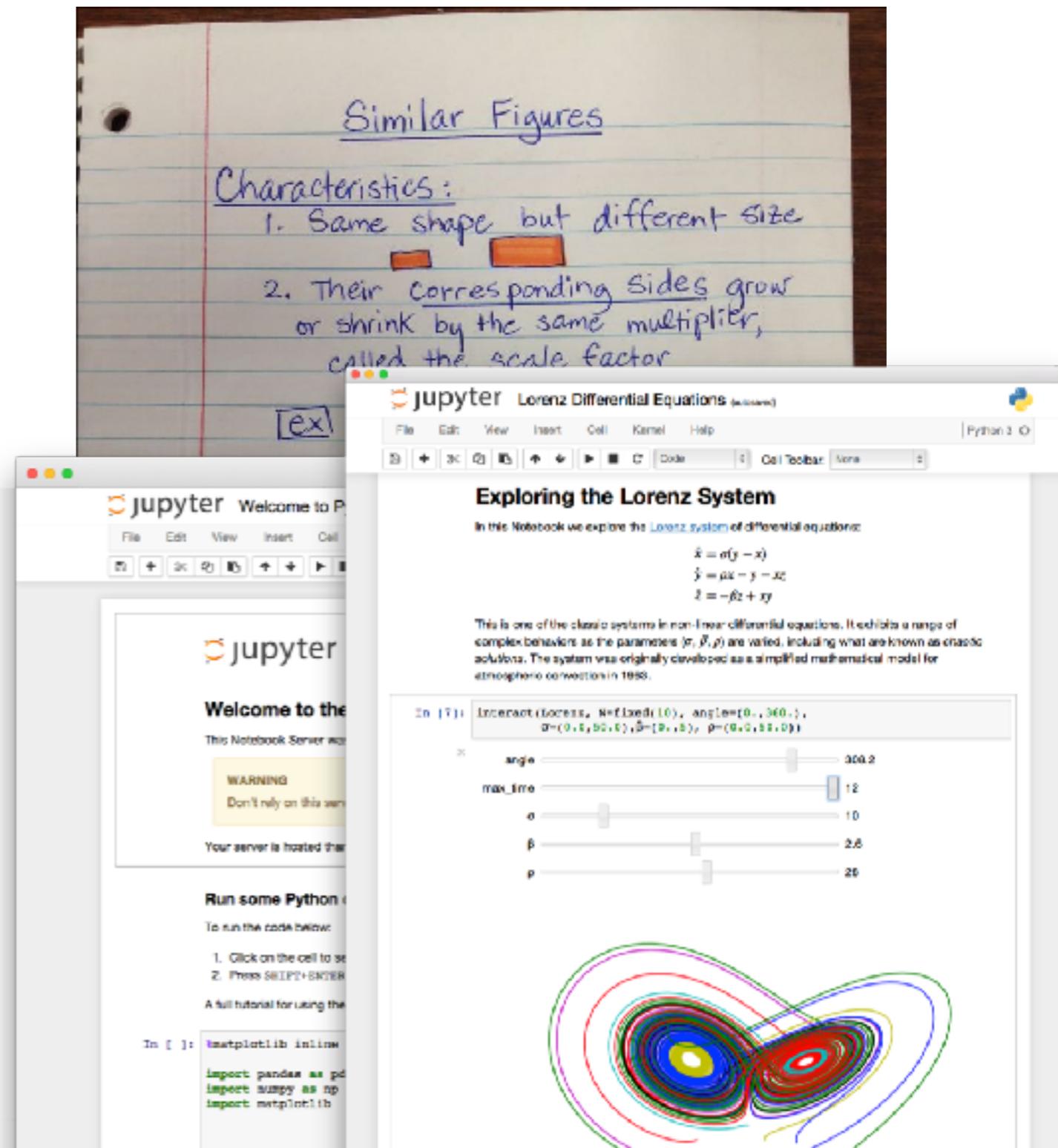
Basic Concepts of Jupyter



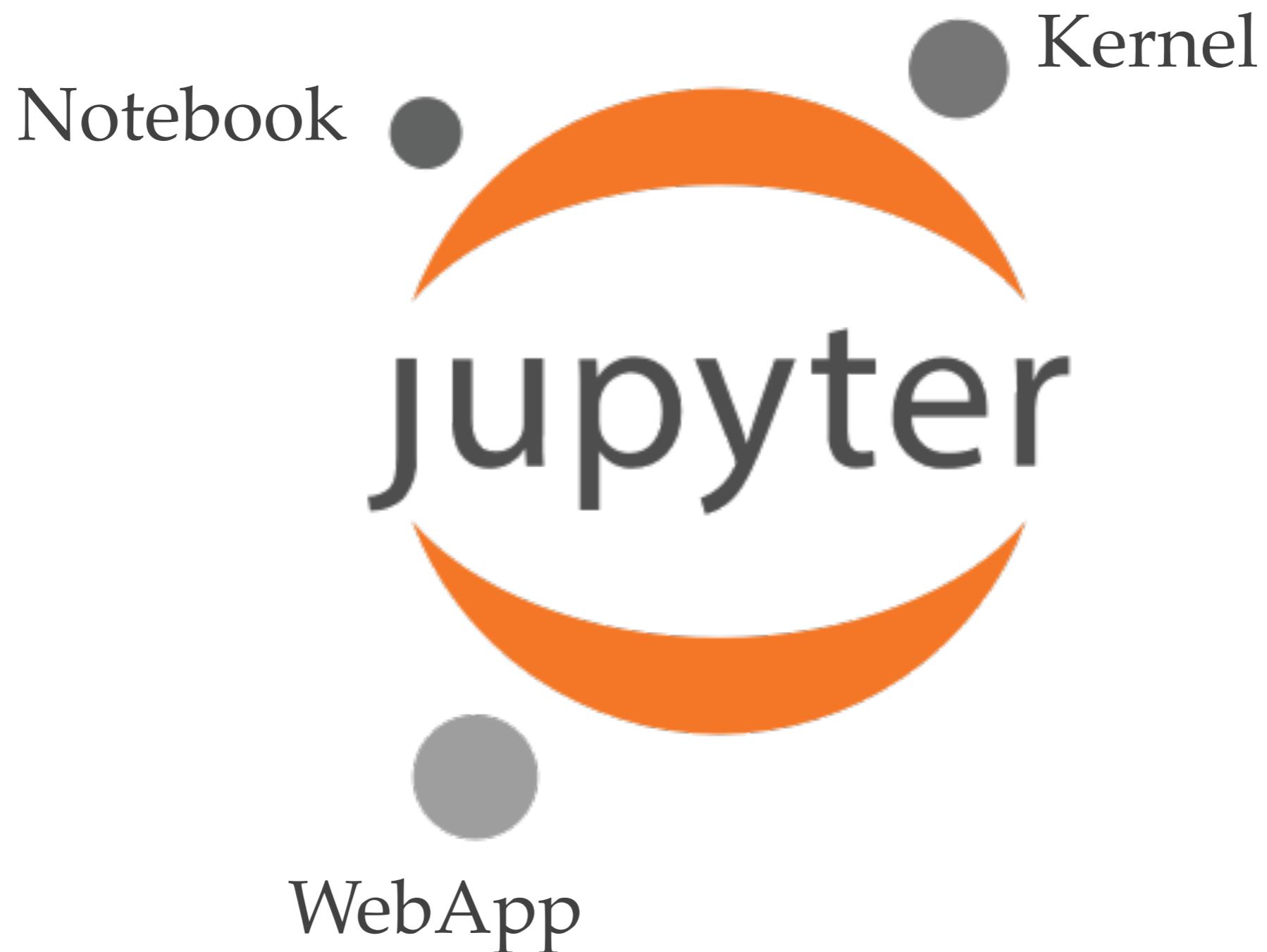
Basic Concepts of Jupyter

Notebook

- Data, text, documentation, figures in the same place
- Adds code to the same space
- Popularized by Mathematica
- Popular in data science
- Astronomy: self-documented pipelines
- Reproducible research



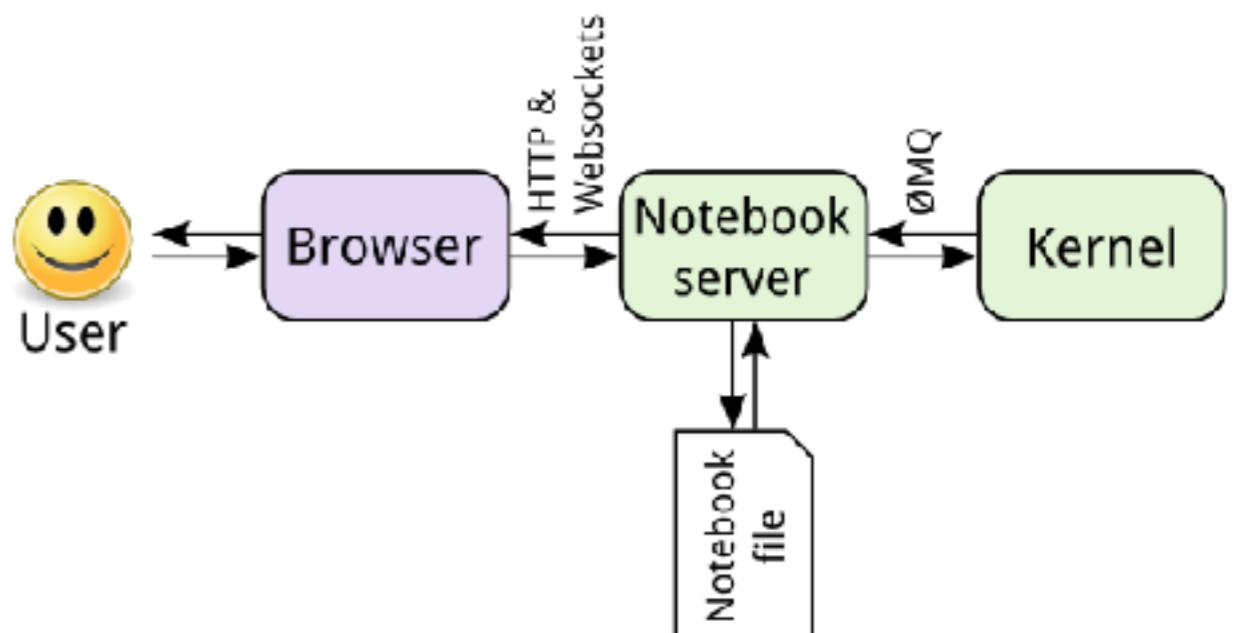
Basic Concepts of Jupyter



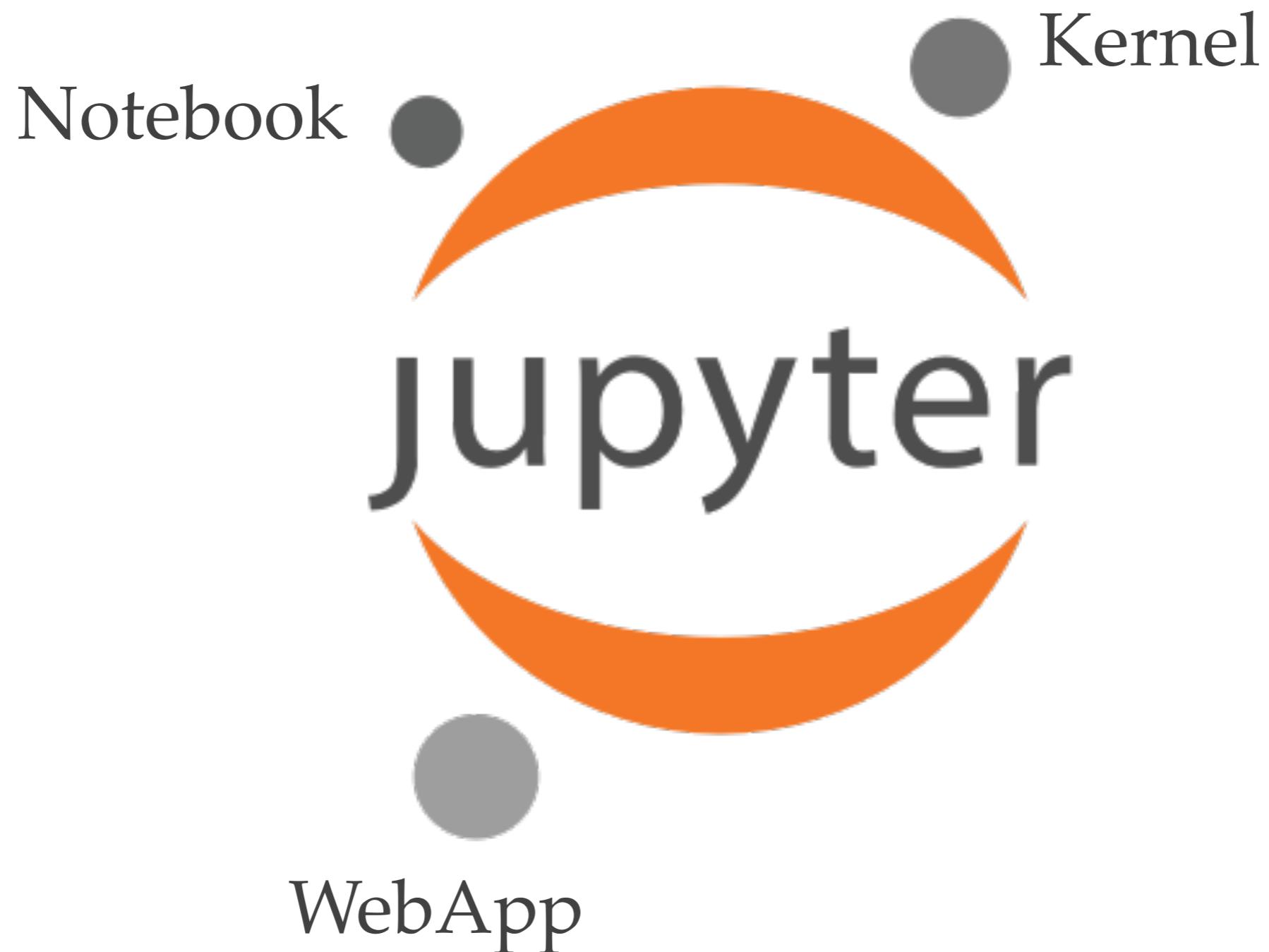
Basic Concepts of Jupyter

Kernel

- Interactive code executor
- Running in the “background” of the notebook
- Maintains the state (variables)
- Jupyter supports several
- We will use only **iPython** kernel



Basic Concepts of Jupyter



Basic Concepts of Jupyter

WebApp

- Jupyter is a **notebook server**
- Your browser is the **client**
- The state is in the server, not in the browser
- Is a Web Application, despite if you use it locally

