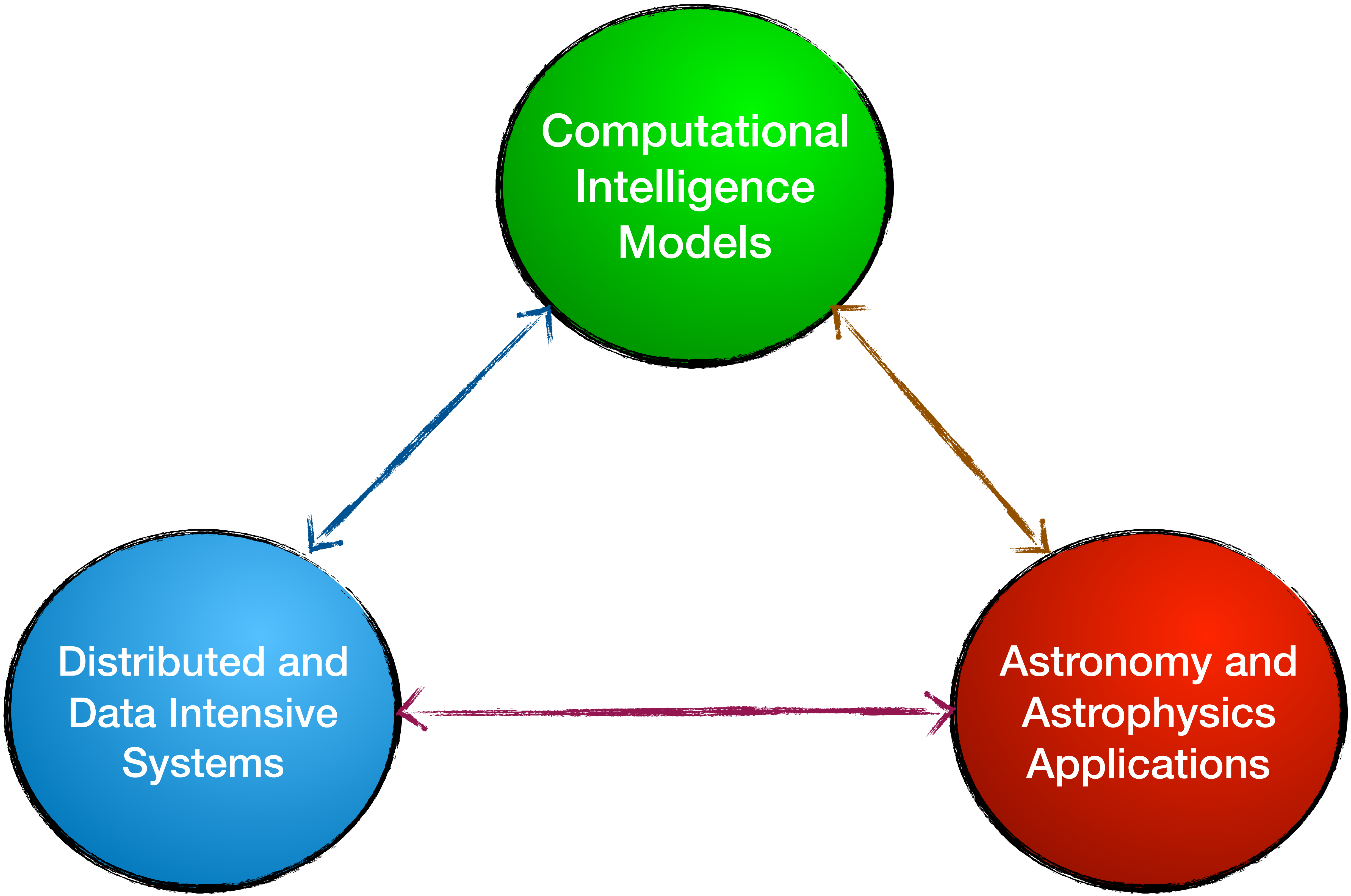
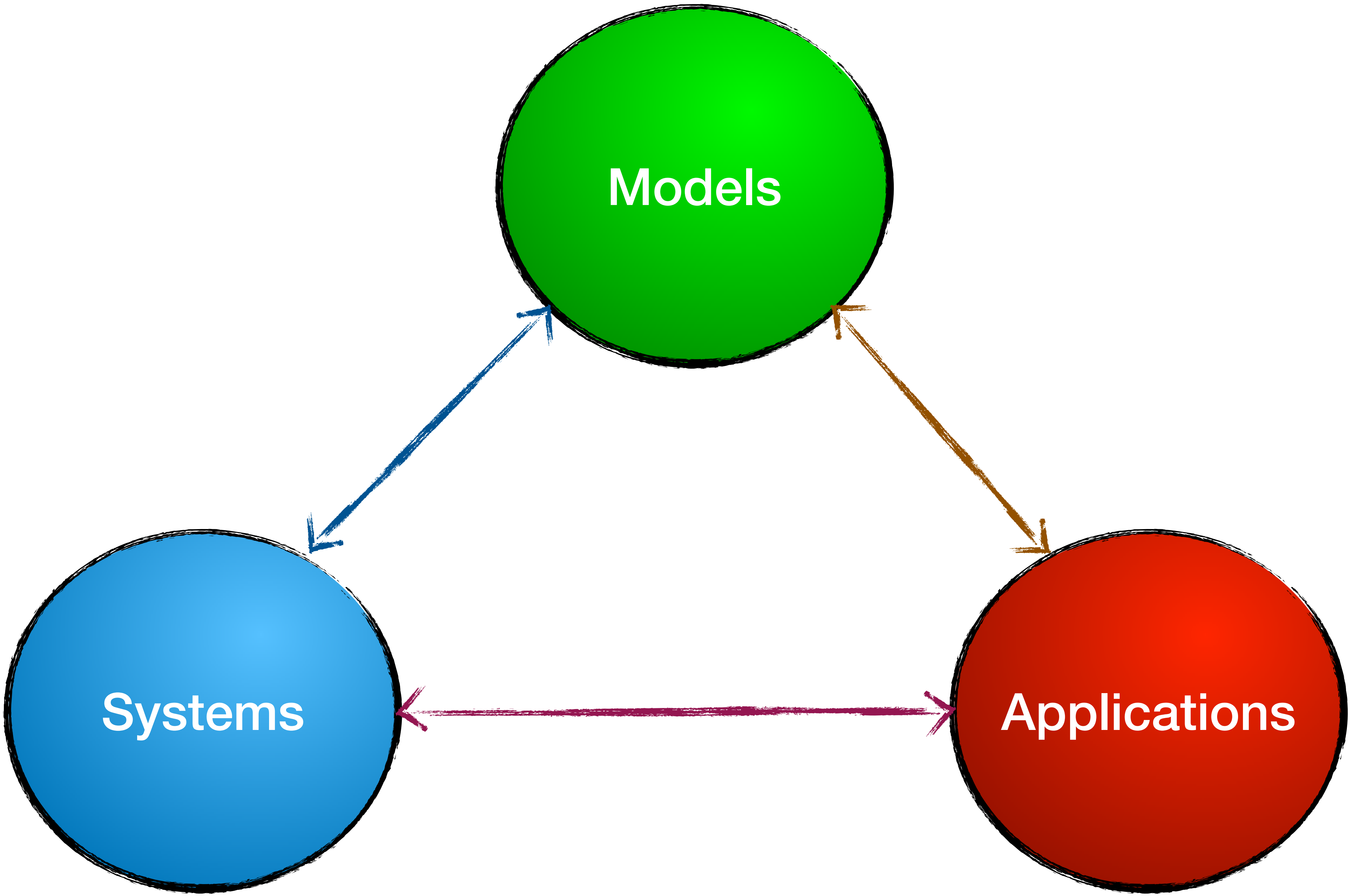
Introduction to Machine Learning

Dr. Mauricio Araya

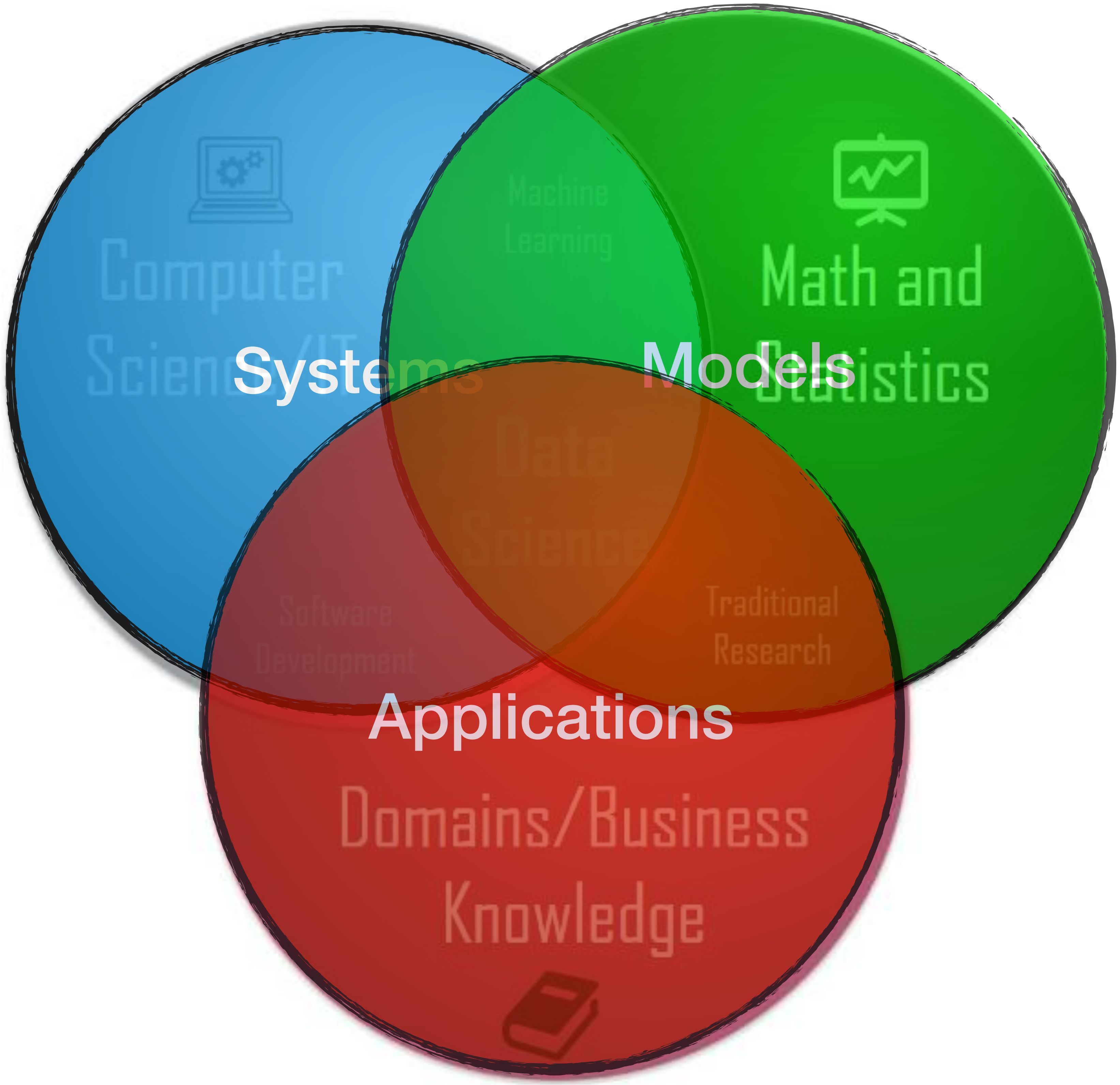
What I do


Computational Intelligence Models
Distributed and Data Intensive Systems
Astronomy and Astrophysics Applications

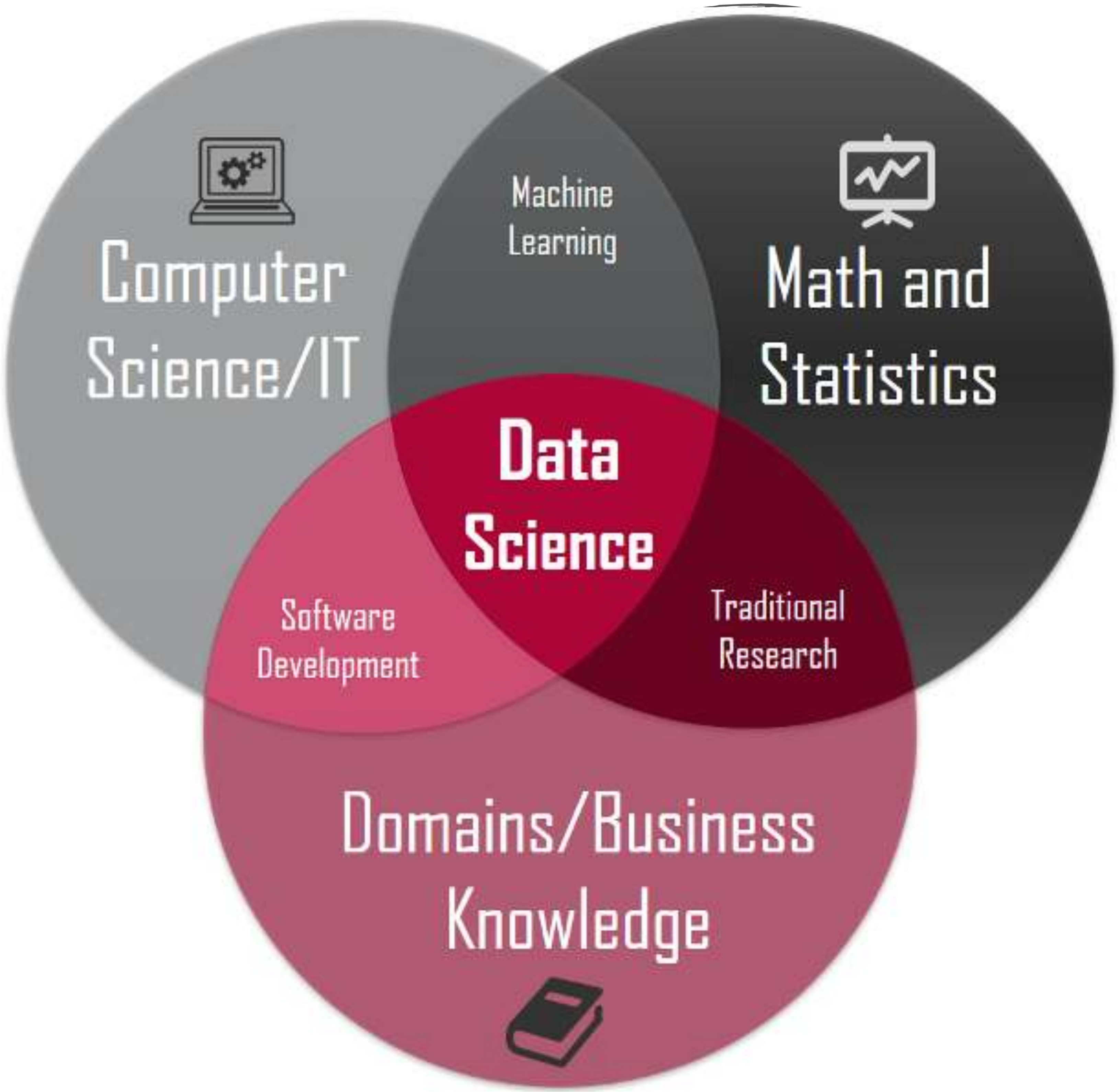
Theory, Practice and Applications

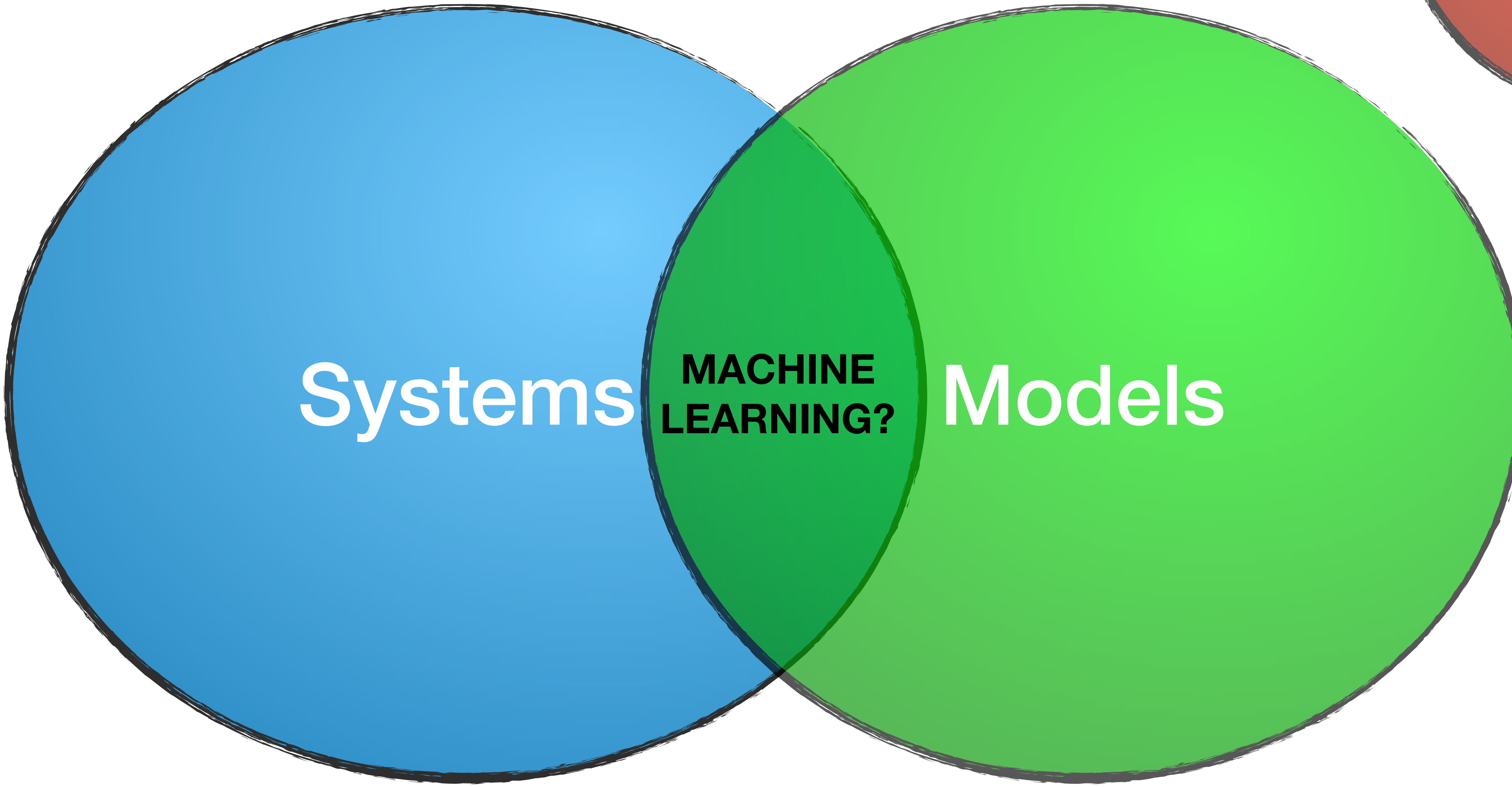


Data Science



Data Science





Machine Learning?

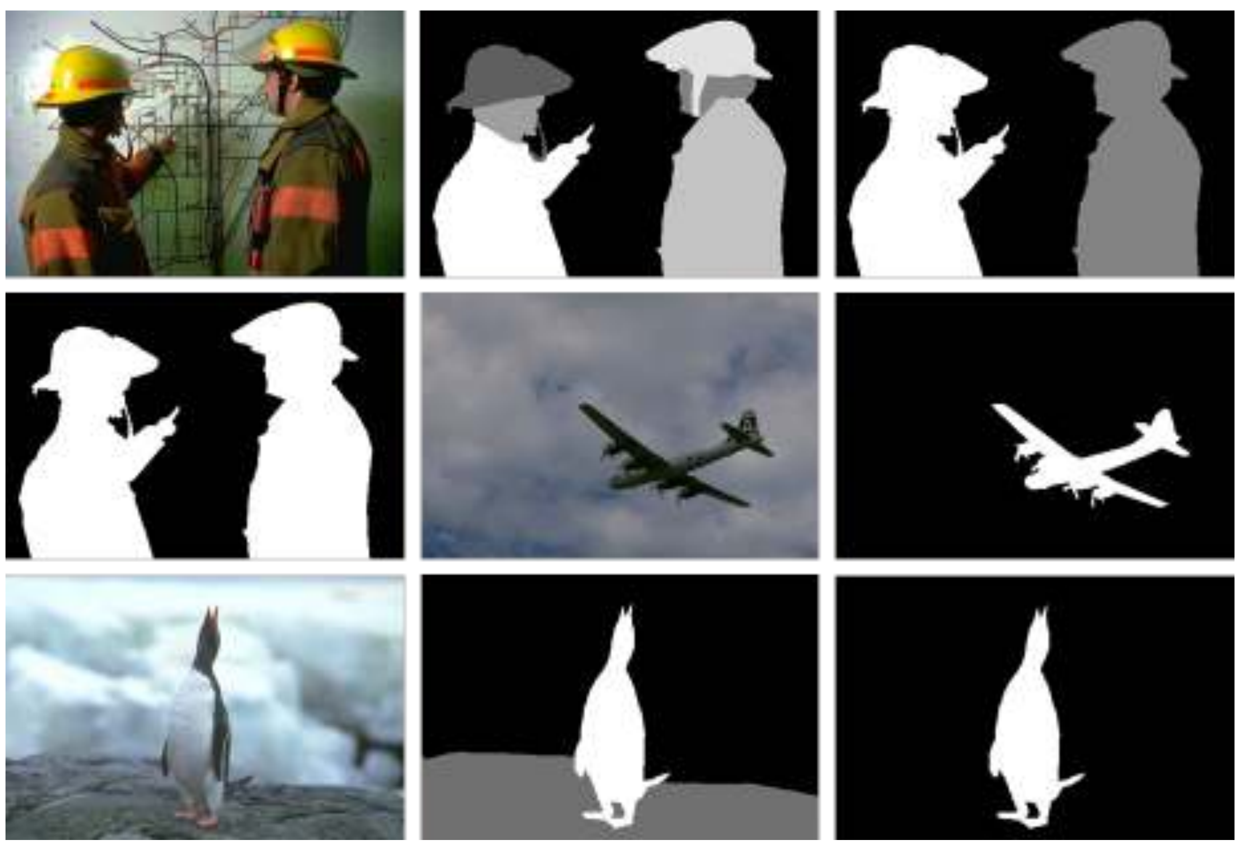
Pattern Recognition?

Artificial Intelligence?

Optimization?

Pattern Recognition

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



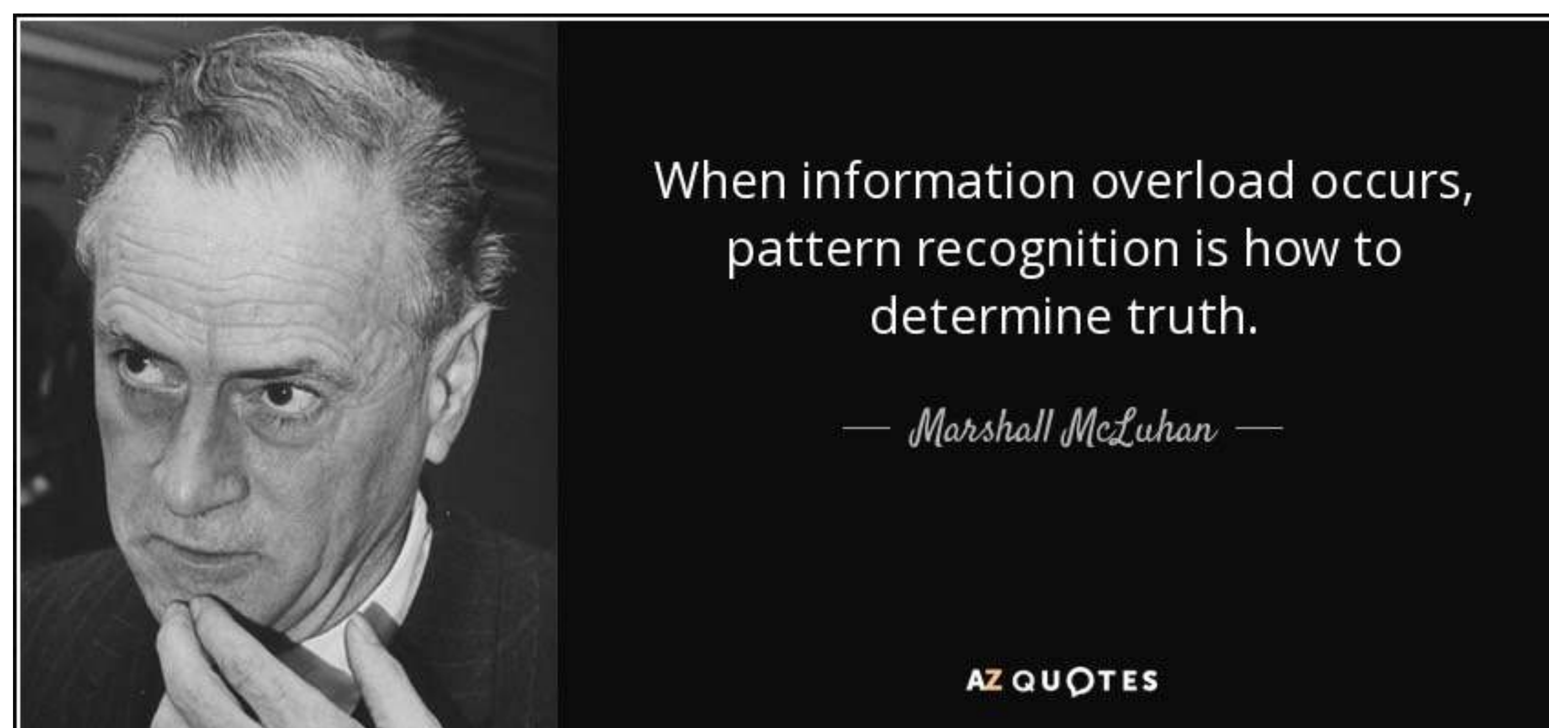
"And you are telling me
you work in pattern recognition!"

Pattern Recognition

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

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

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



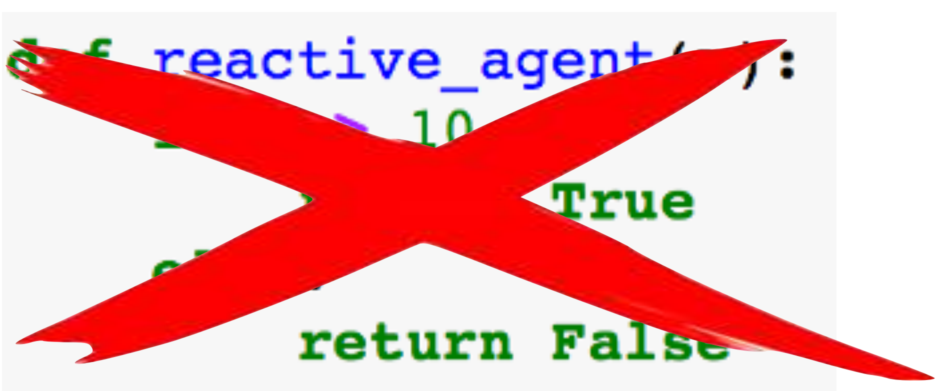
Machine Learning?

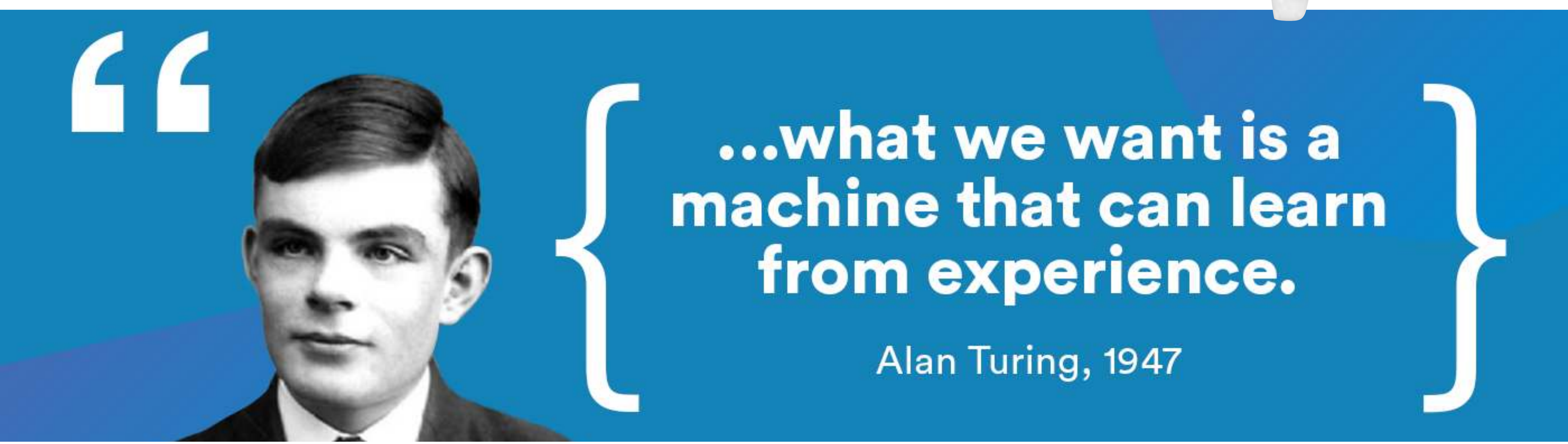
```python
def reactive_agent(x):
    if x > 10.0:
        return True
    else:
        return False
```

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

Machine Learning (ML)

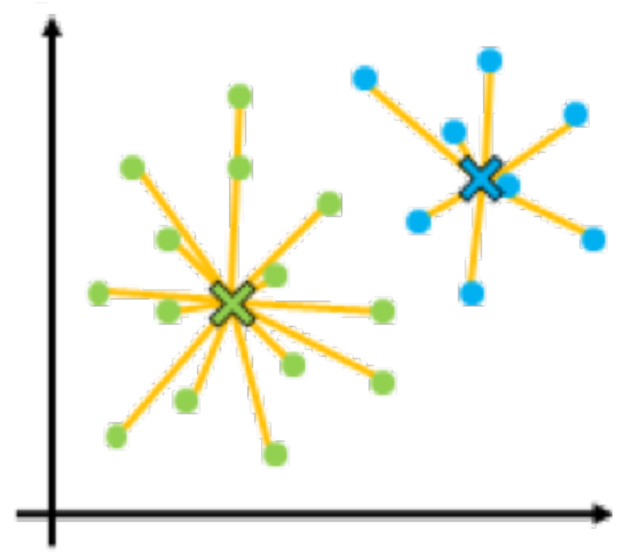
```
X = np.array([10.9, 5.34, 8.32, 12.43, 20.32, 7.24])
y = np.array([True, False, False, True, True, False])
```





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])
```



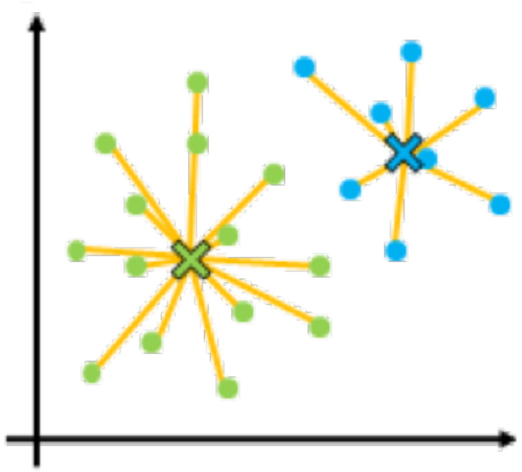
$$\mu_T = \sum_{x \in T} \frac{x}{|T|}$$

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

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

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])
```



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

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

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

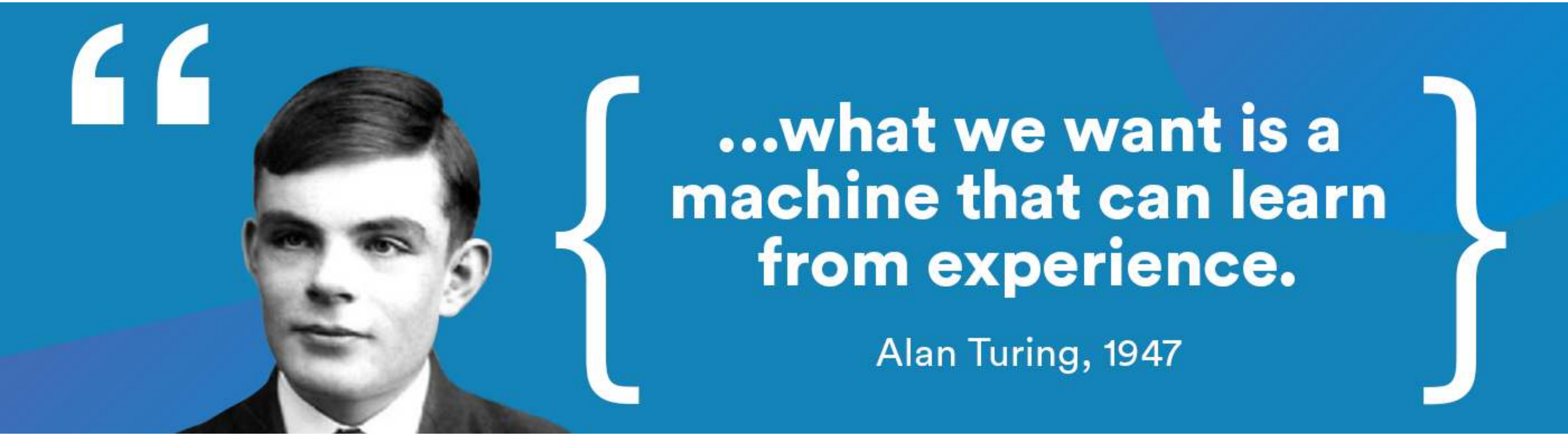
$$S = \frac{\mu_T + \mu_F}{2}$$

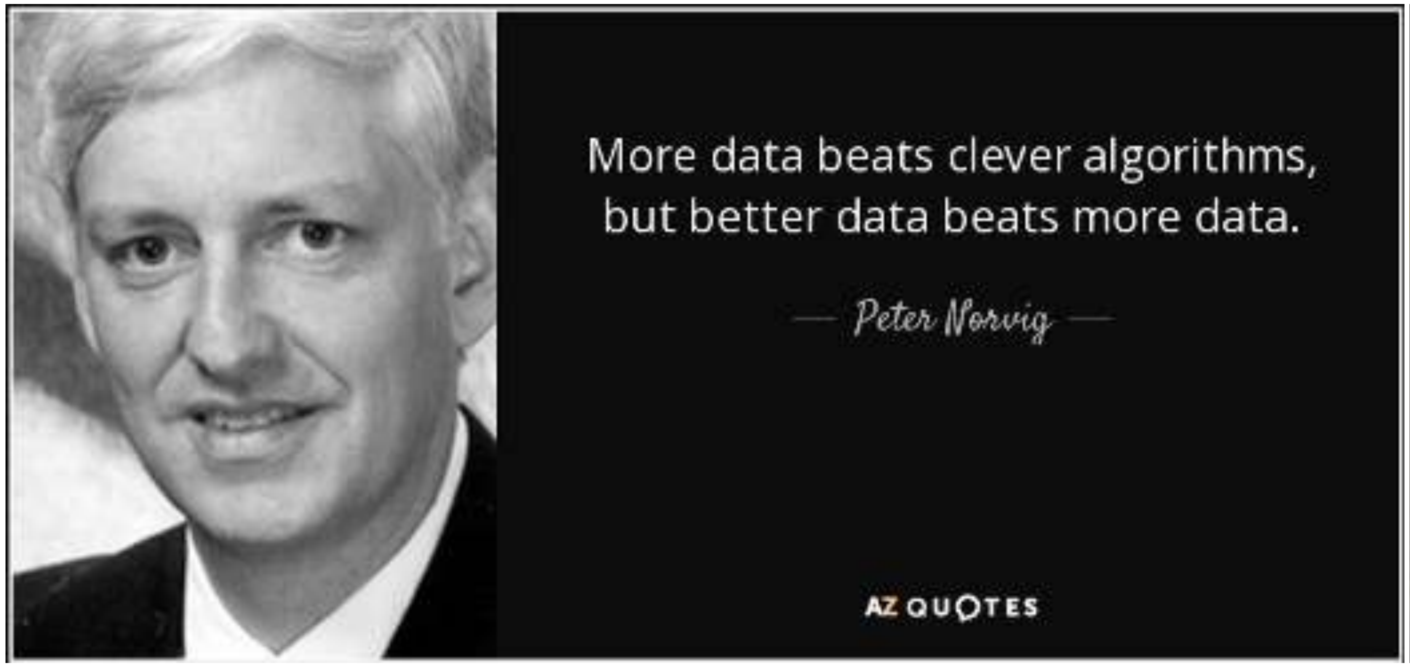
```
def learning_agent(x,Data,labels):
    v_true  = np.mean(Data[labels==True])
    v_false = np.mean(Data[labels==False])
    d_true  = np.abs(x - v_true)
    d_false  = np.abs(x - v_false)
    if d_true < d_false:
        return True
    else:
        return False
```

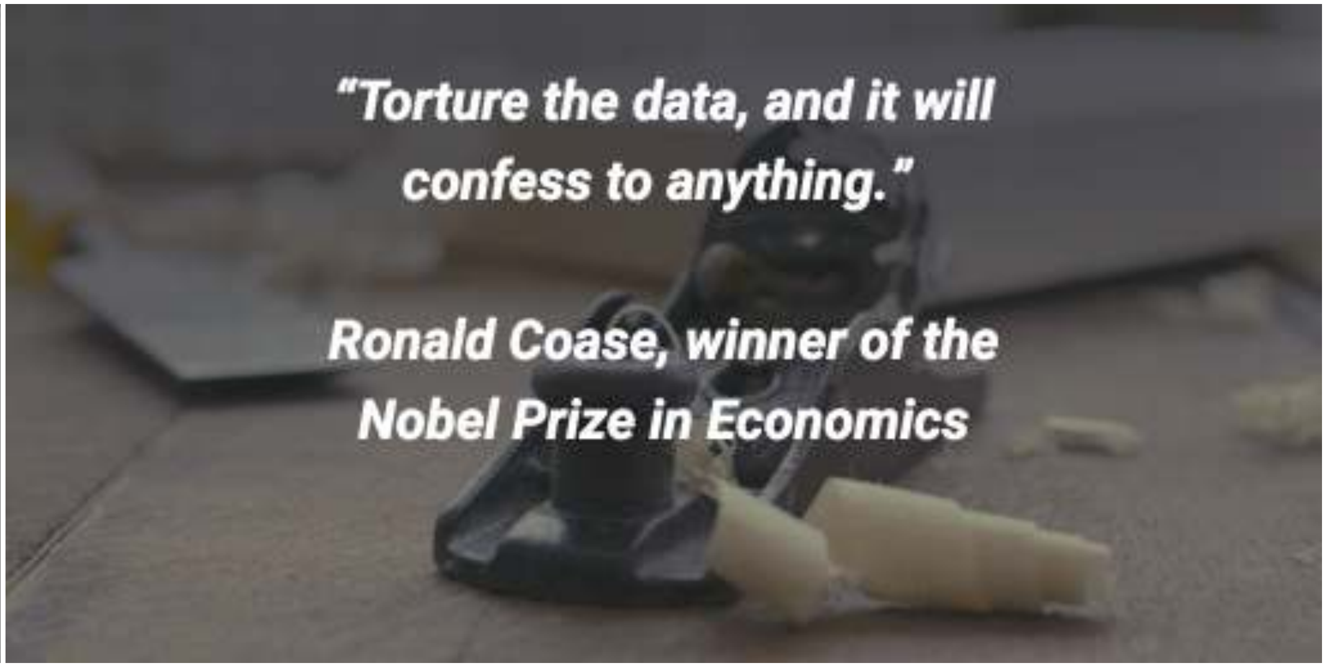
```
(v_true + v_false)/2
```

10.758333333333333

What is Machine Learning?







Is Machine Learning New?

- **Mathematician/Statistician:** of course not!, statistical learning goes back to Bayes, Legendre and Laplace!
- **Computer Scientist:** it is just another form of Artificial Intelligence. Basically smart algorithms + Moore's law
- **Astronomer/Biologist:** we have being learning from data from the dawn of the civilization.

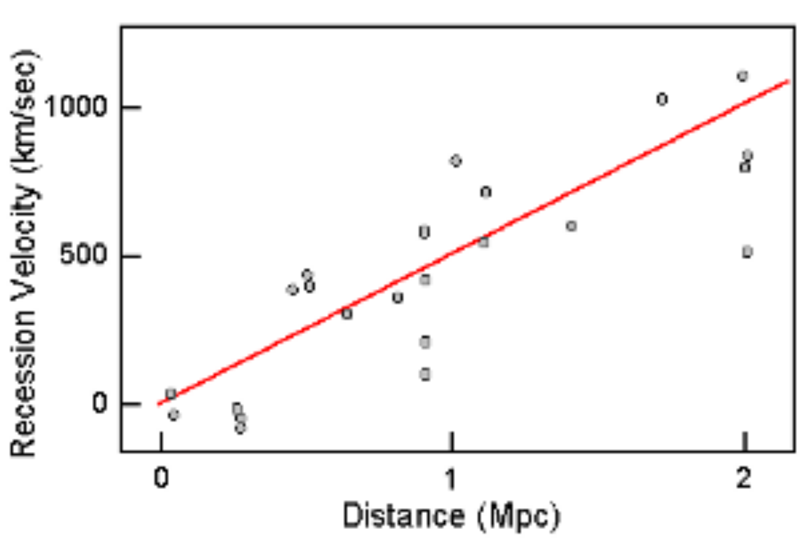
"Machine learning is applied mathematics on steroids, where the steroids are computing power and a lot of data"
—Maria Spiropulu

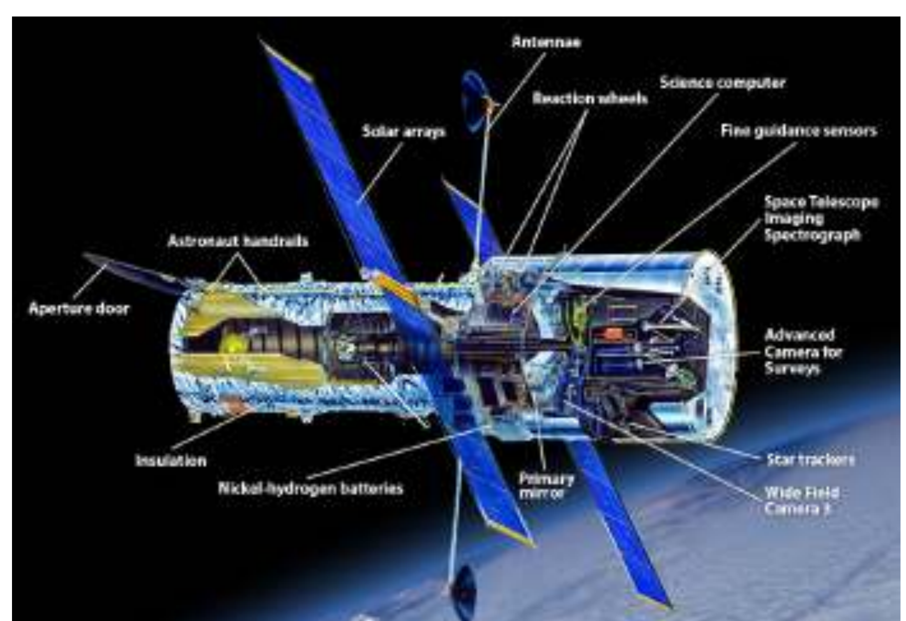
Do we need ML?

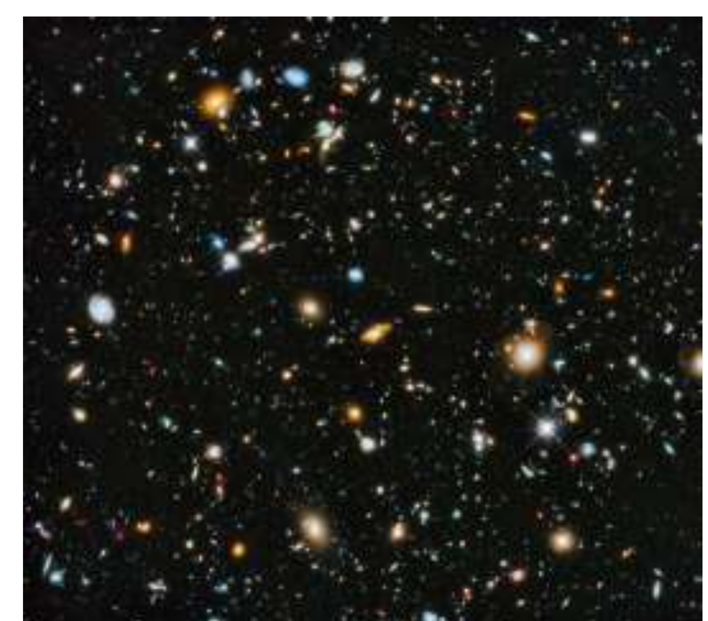
Example

Edwin Hubble started with 46 extragalactic nebulae for his famous paper. After learning that some of them were poorly measured, he selected by hand only 24. Imagine doing that for the Hubble Ultra Deep Field!

Hubble's Data (1929)







Data

Data: ML volcabulary

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

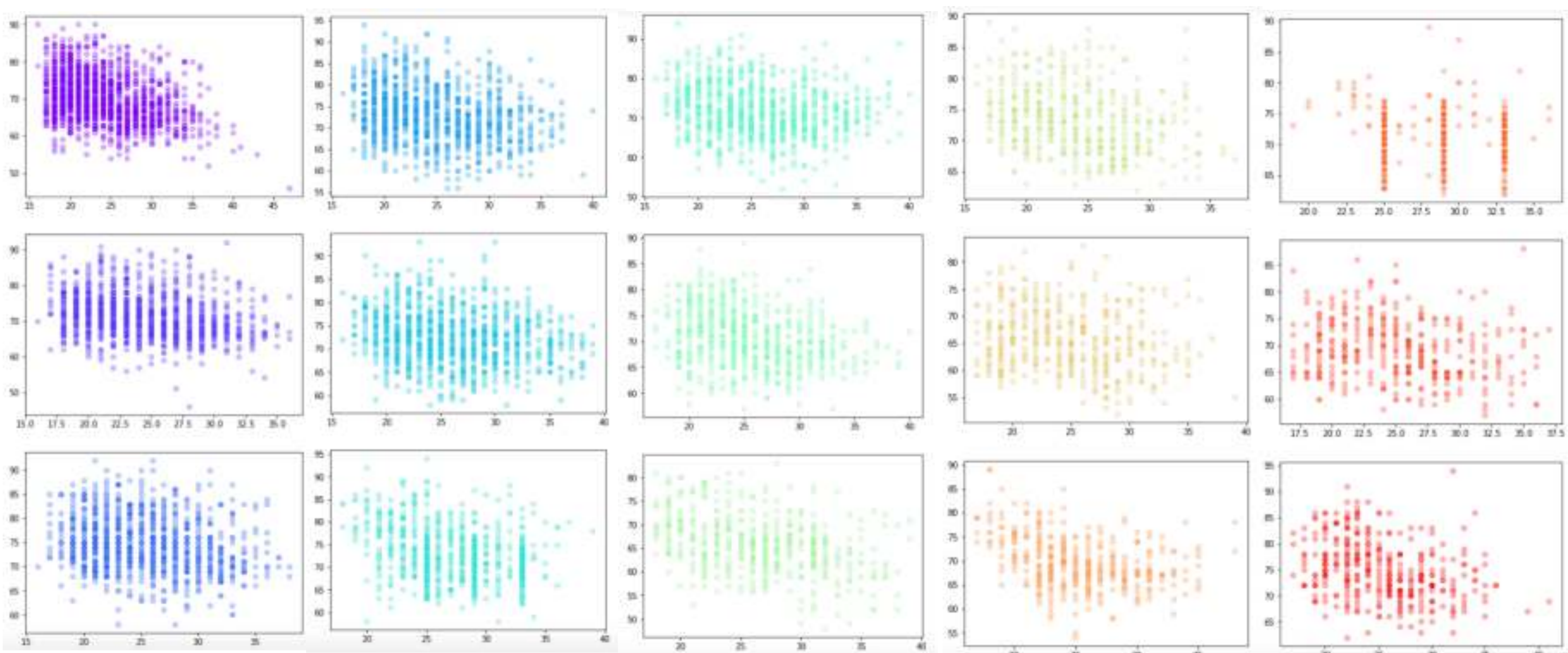
Potential vs Age

Features

Samples

	Name	Age	Nationality	Overall	Potential	Value	Photo
0	Cristiano Ronaldo	33	Portugal	94	94	[illegible]	[illegible]
1	L. Messi	30	Argentina	[illegible]	[illegible]	[illegible]	[illegible]
2	Neymar	25	Brazil	92	94	[illegible]	[illegible]
3	L. Suárez	30	Uruguay	92	92	[illegible]	[illegible]
4	M. Neuer	31	Germany	[illegible]	[illegible]	[illegible]	[illegible]
5	R. Lewandowski	28	Poland	[illegible]	[illegible]	[illegible]	[illegible]
6	De Gea	26	Spain	90	92	[illegible]	[illegible]
7	E. Hazard	26	Belgium	[illegible]	[illegible]	[illegible]	[illegible]
8	T. Kroos	27	Germany	90	90	[illegible]	[illegible]
9	G. Higuaín	29	Argentina	90	90	[illegible]	[illegible]
10	Sergio Ramos	31	Spain	[illegible]	[illegible]	[illegible]	[illegible]
11	K. De Bruyne	26	Belgium	89	92	[illegible]	[illegible]
12	T. Courtois	25	Belgium	89	92	[illegible]	[illegible]
13	A. Sánchez	28	Chile	[illegible]	[illegible]	[illegible]	[illegible]
14	L. Modrić	31	Croatia	89	89	[illegible]	[illegible]
15	G. Bale	27	Wales	89	89	[illegible]	[illegible]

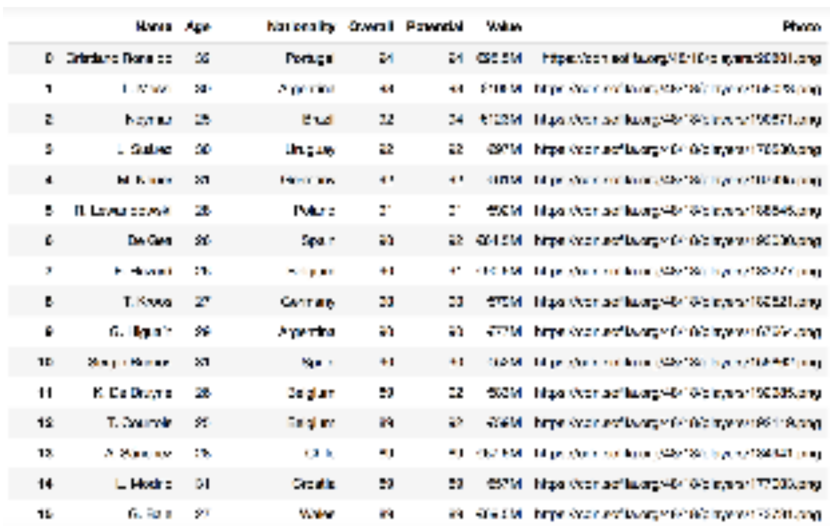
England
Germany
Spain
France
Argentina
Brazil
Italy
Colombia
Japan
Netherlands
Republic of Ireland
United States
Chile
Sweden
Portugal



Potential vs Age

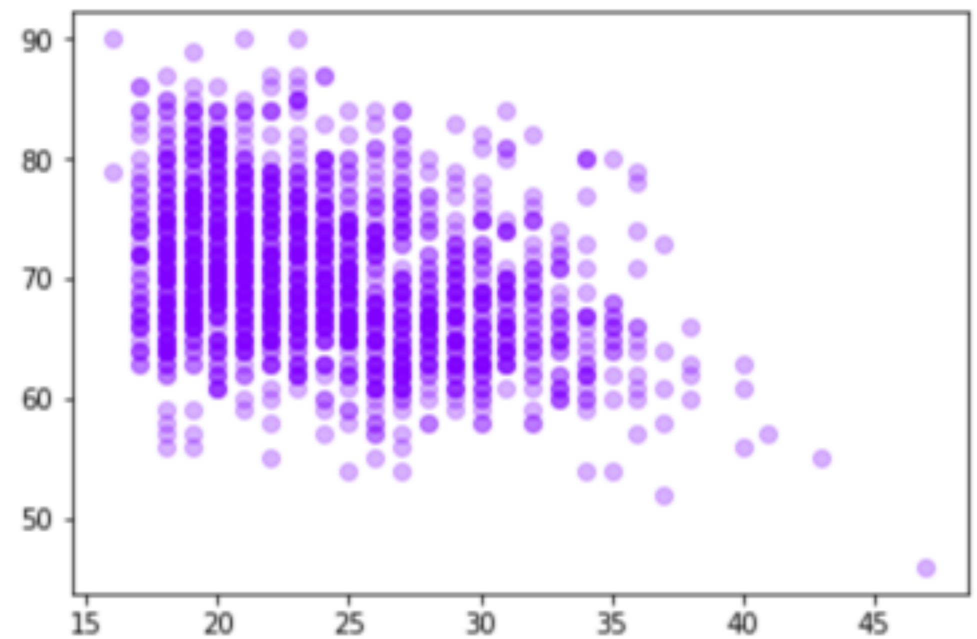
Features

Samples

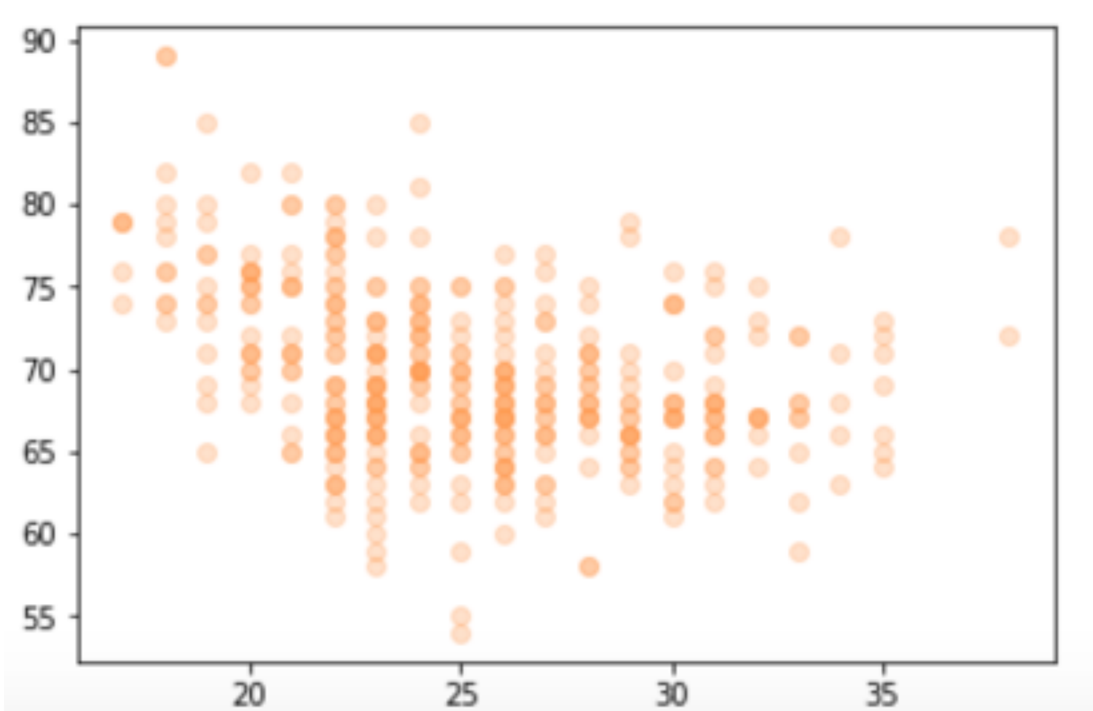


	Name	Age	Nationality	Overall	Potential	Value	Photo
0	Cristiano Ronaldo	[illegible]	Portugal	[illegible]	[illegible]	[illegible]	[illegible]
1	[illegible]	[illegible]	Argentina	[illegible]	[illegible]	[illegible]	[illegible]
2	Neymar	[illegible]	Brazil	[illegible]	[illegible]	[illegible]	[illegible]
3	L. Suárez	[illegible]	Uruguay	[illegible]	[illegible]	[illegible]	[illegible]
4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5	R. Lewandowski	[illegible]	Poland	[illegible]	[illegible]	[illegible]	[illegible]
6	De Gea	[illegible]	Spain	[illegible]	[illegible]	[illegible]	[illegible]
7	E. Hazard	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
8	T. Kroos	[illegible]	Germany	[illegible]	[illegible]	[illegible]	[illegible]
9	G. Higuaín	[illegible]	Argentina	[illegible]	[illegible]	[illegible]	[illegible]
10	[illegible]	[illegible]	Spain	[illegible]	[illegible]	[illegible]	[illegible]
11	K. De Bruyne	[illegible]	Belgium	[illegible]	[illegible]	[illegible]	[illegible]
12	T. Courtois	[illegible]	Belgium	[illegible]	[illegible]	[illegible]	[illegible]
13	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
14	L. Modrić	[illegible]	Croatia	[illegible]	[illegible]	[illegible]	[illegible]
15	G. Bale	[illegible]	Wales	[illegible]	[illegible]	[illegible]	[illegible]

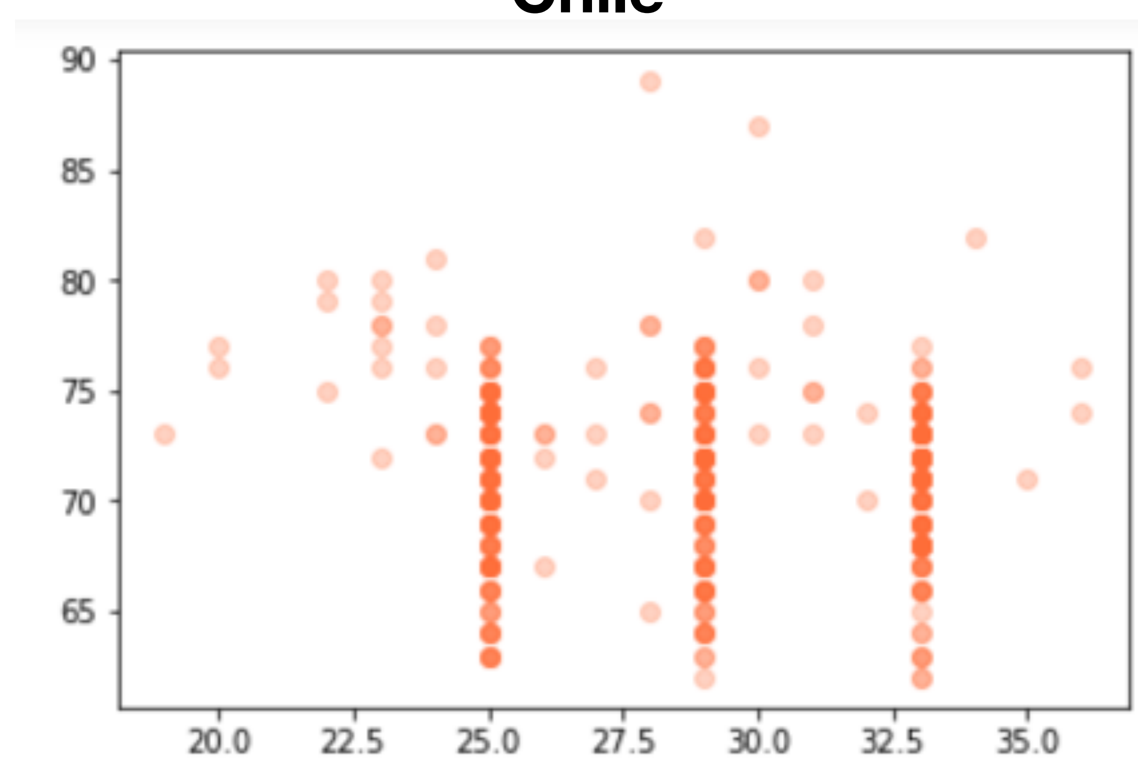
England



United States



Chile

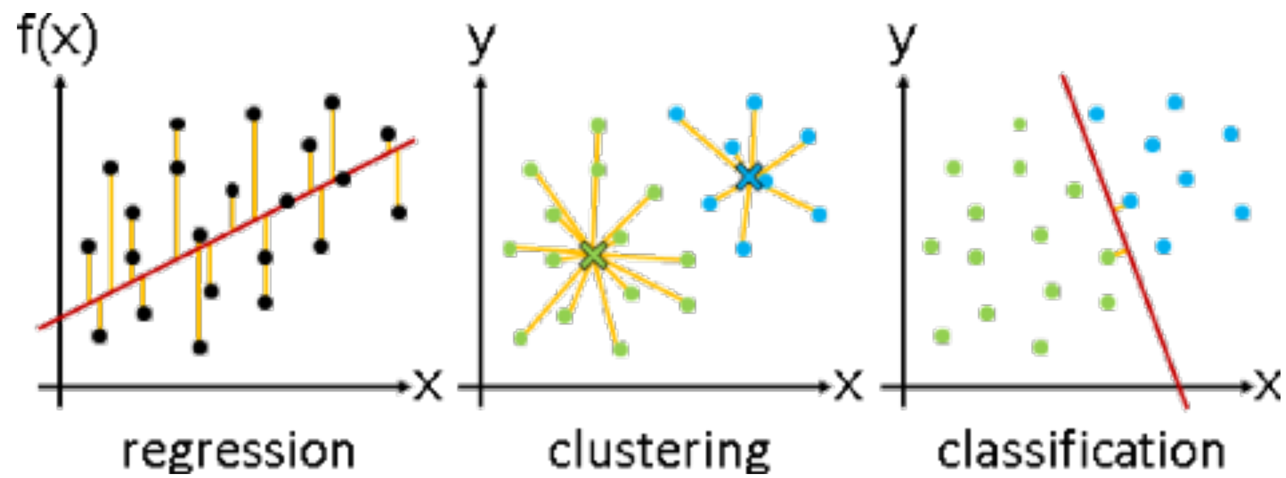


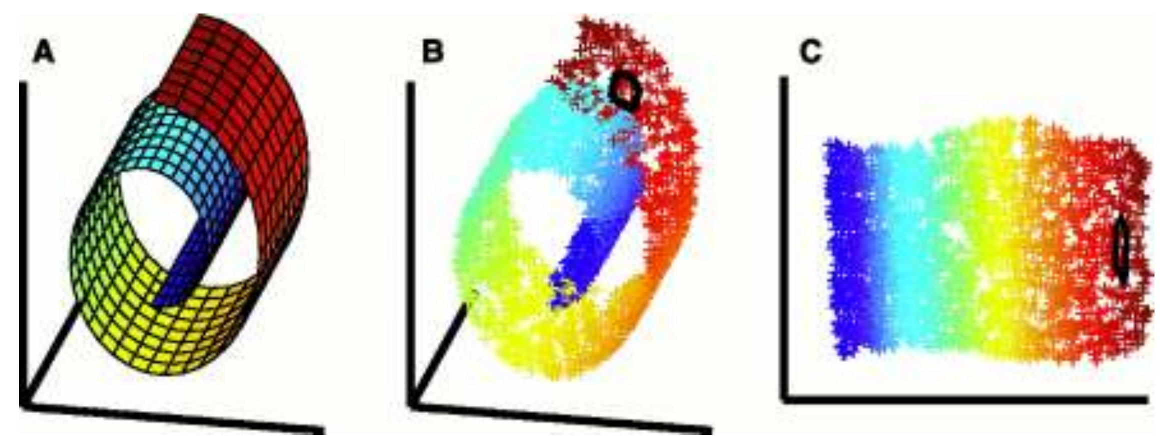
What can I do with this?

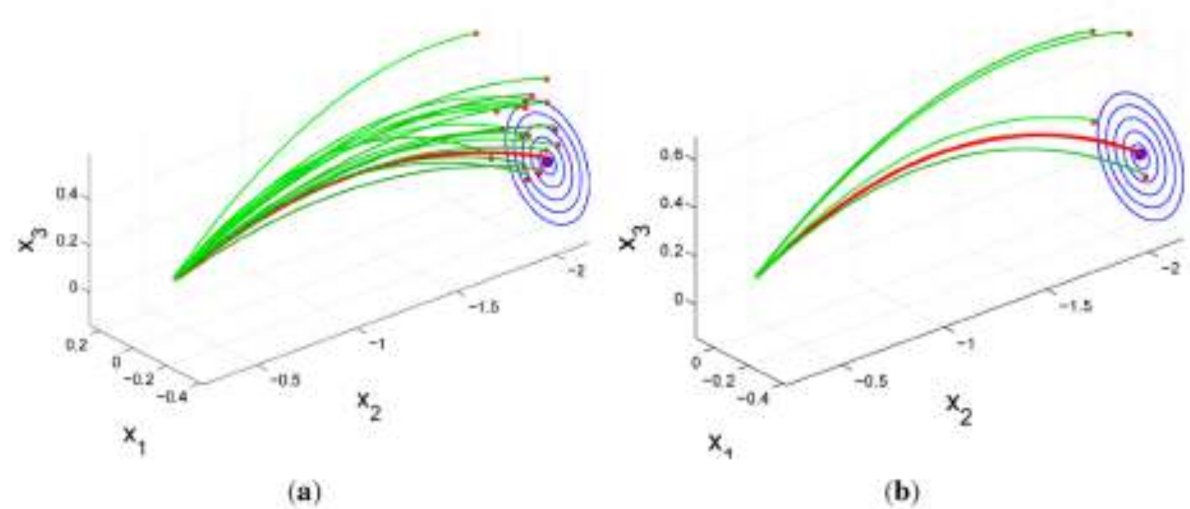
Models and Methods

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.







So ML is Optimization?

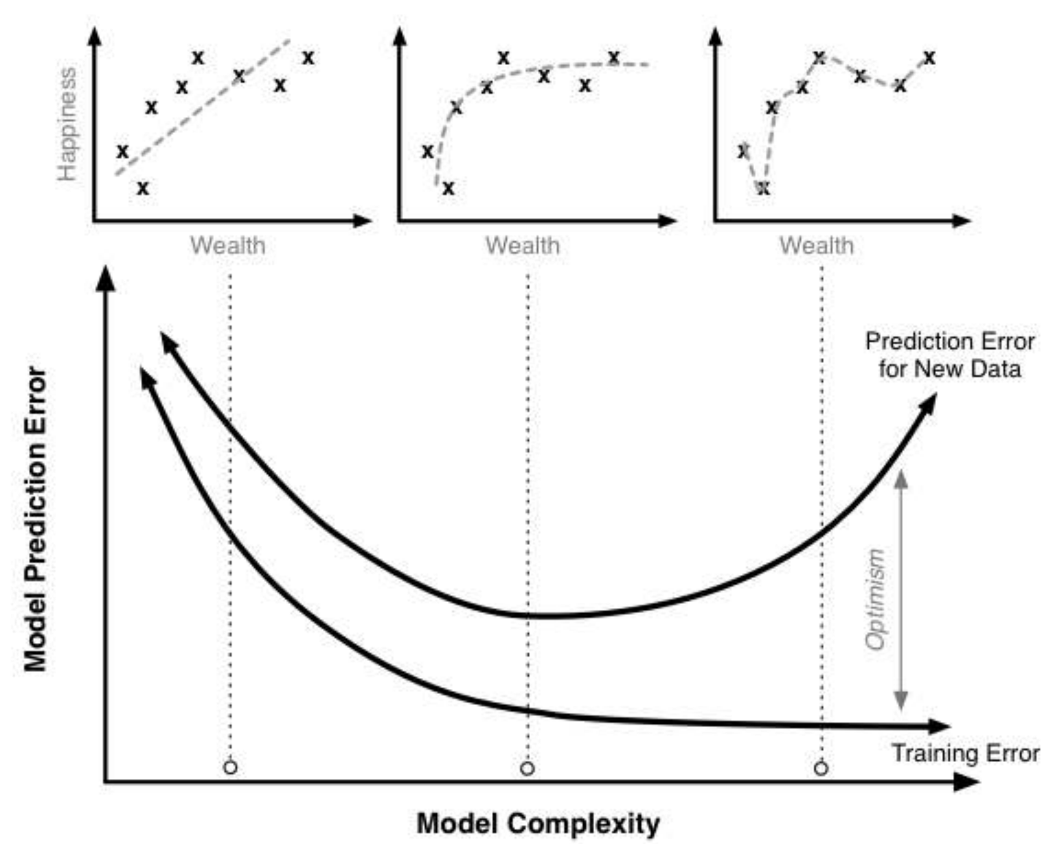
- **Yes!, but…**
  - it is optimizing for the possible future values
  - the objective is to *generalize* well
  - *prediction error* is different from *training error*
  - *do* you trust your data?

Just any optimization

$$\underset{\theta}{\operatorname{argmax}} \{f(x;\theta)\}$$

Typical ML optimization

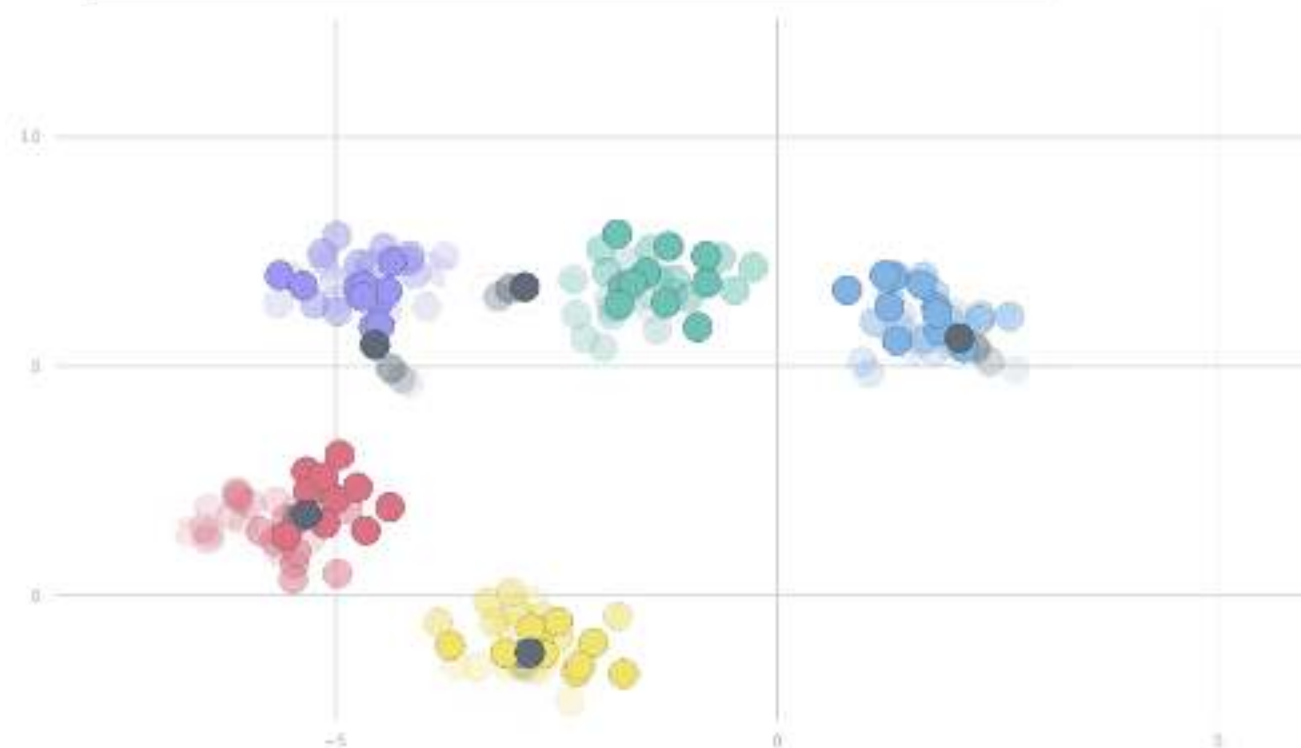
$$\underset{\theta}{\operatorname{argmax}} \{E[f(x;\theta)|X]\}$$



Machine Learning Paradigms

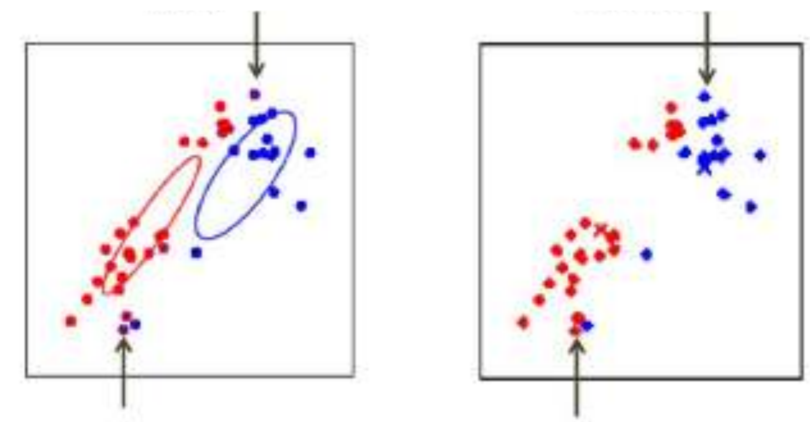
- **Task**
  - Supervised vs Unsupervised
  - Discrete vs Continuous
  - Batch vs Sequential
- **Models**
  - Parametric vs Non-Parametric
  - Linear vs Non-Linear
  - Discriminative vs Generative

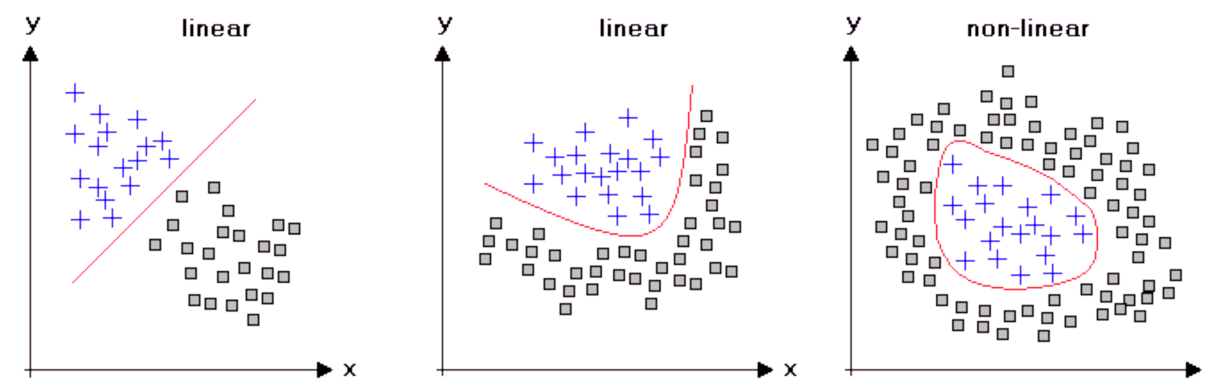
	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

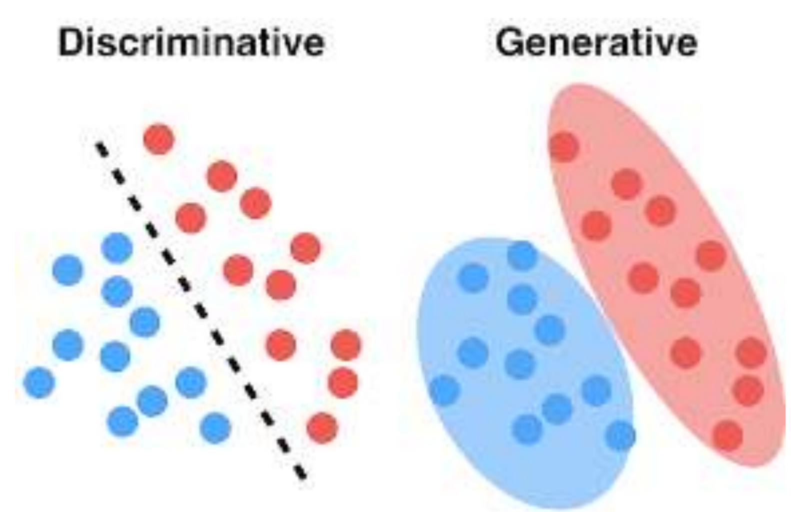


Machine Learning Paradigms

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







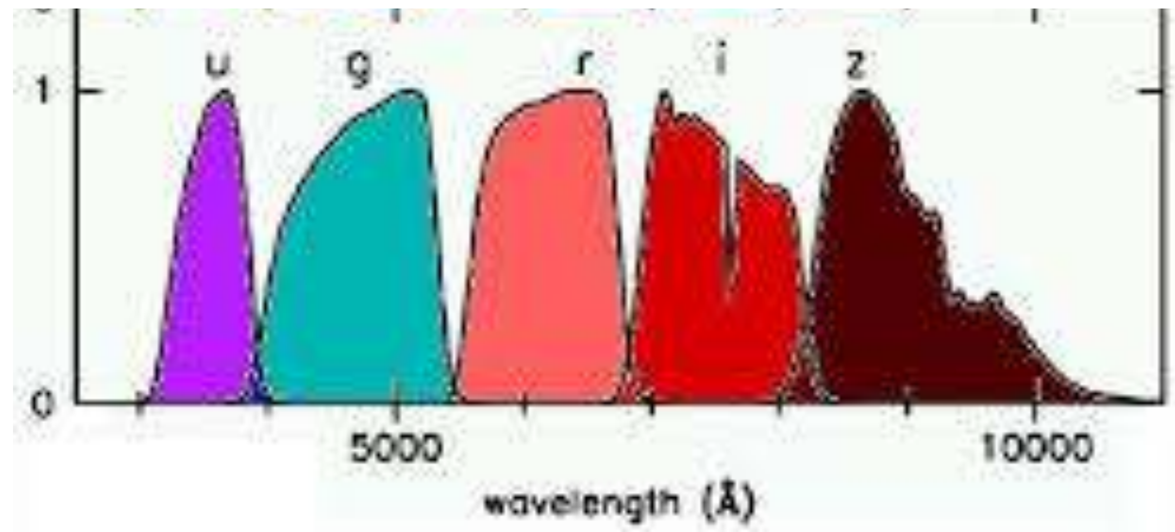
So… hands on!

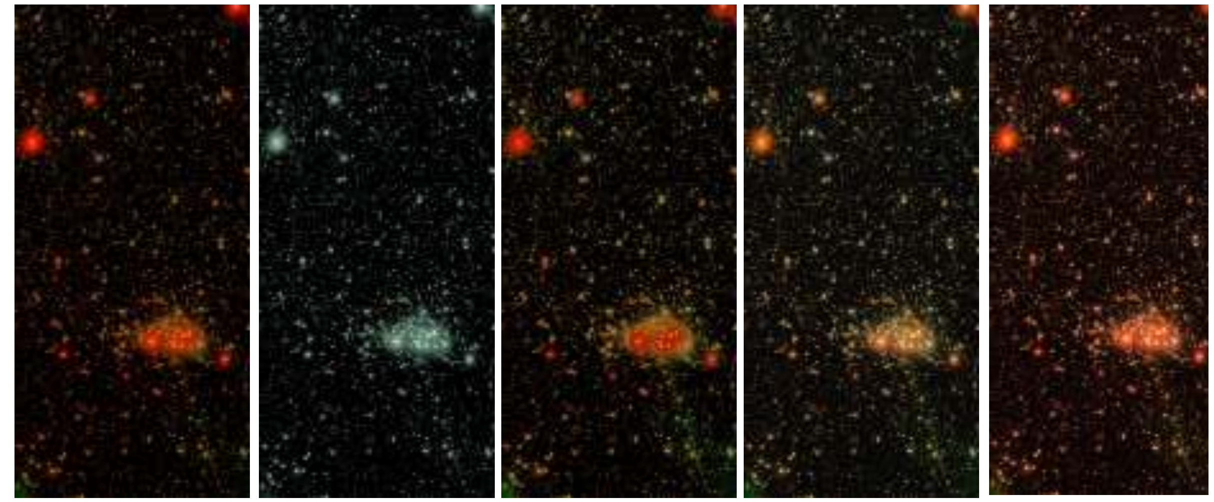
Real Stars

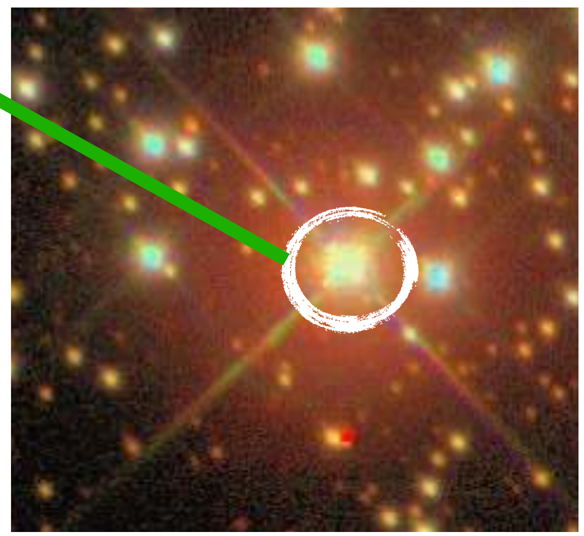
Features

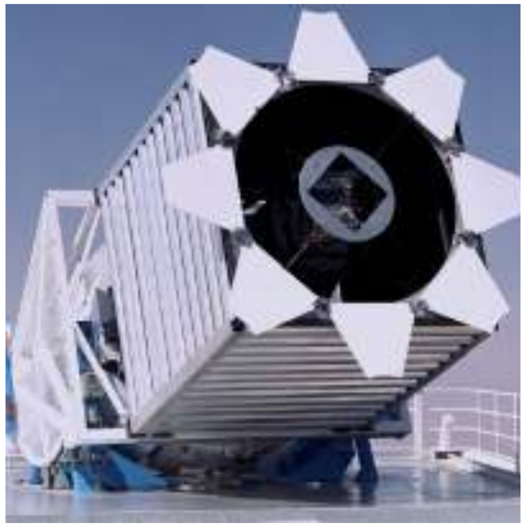
Samples

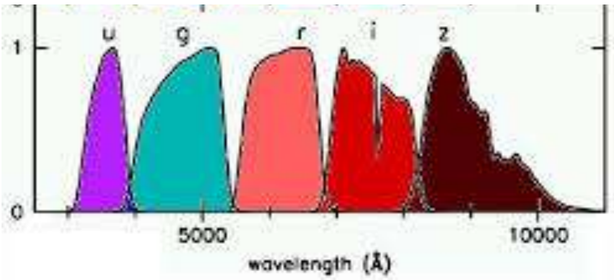
	u-g	g-r	r-i	i-z
0	1.250999	0.394000	0.137000	0.061999
1	1.048000	0.339001	0.151999	0.023001
2	1.008001	0.341999	0.129000	0.203001
3	0.965000	0.392000	0.149000	0.150000
4	1.040001	0.333000	0.125999	0.101999
5	1.154001	0.373999	0.145000	0.121000
6	0.965000	0.384001	0.118999	0.011000
7	1.015001	0.370998	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



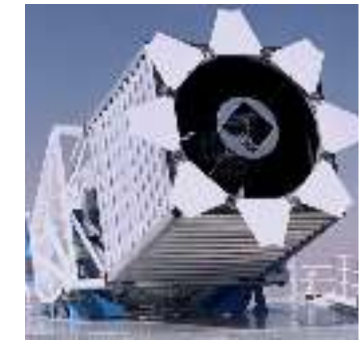








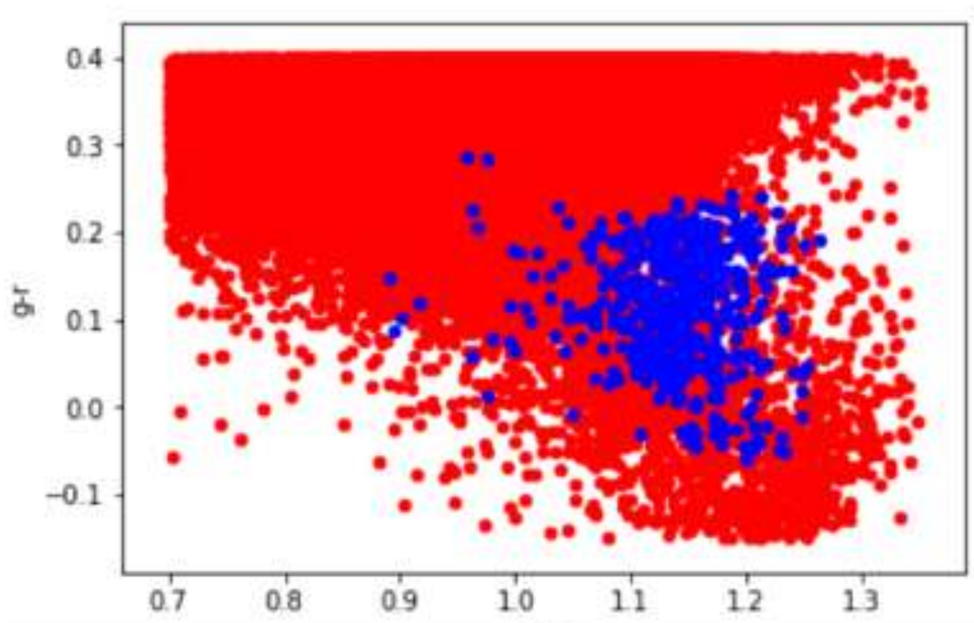
Real Stars

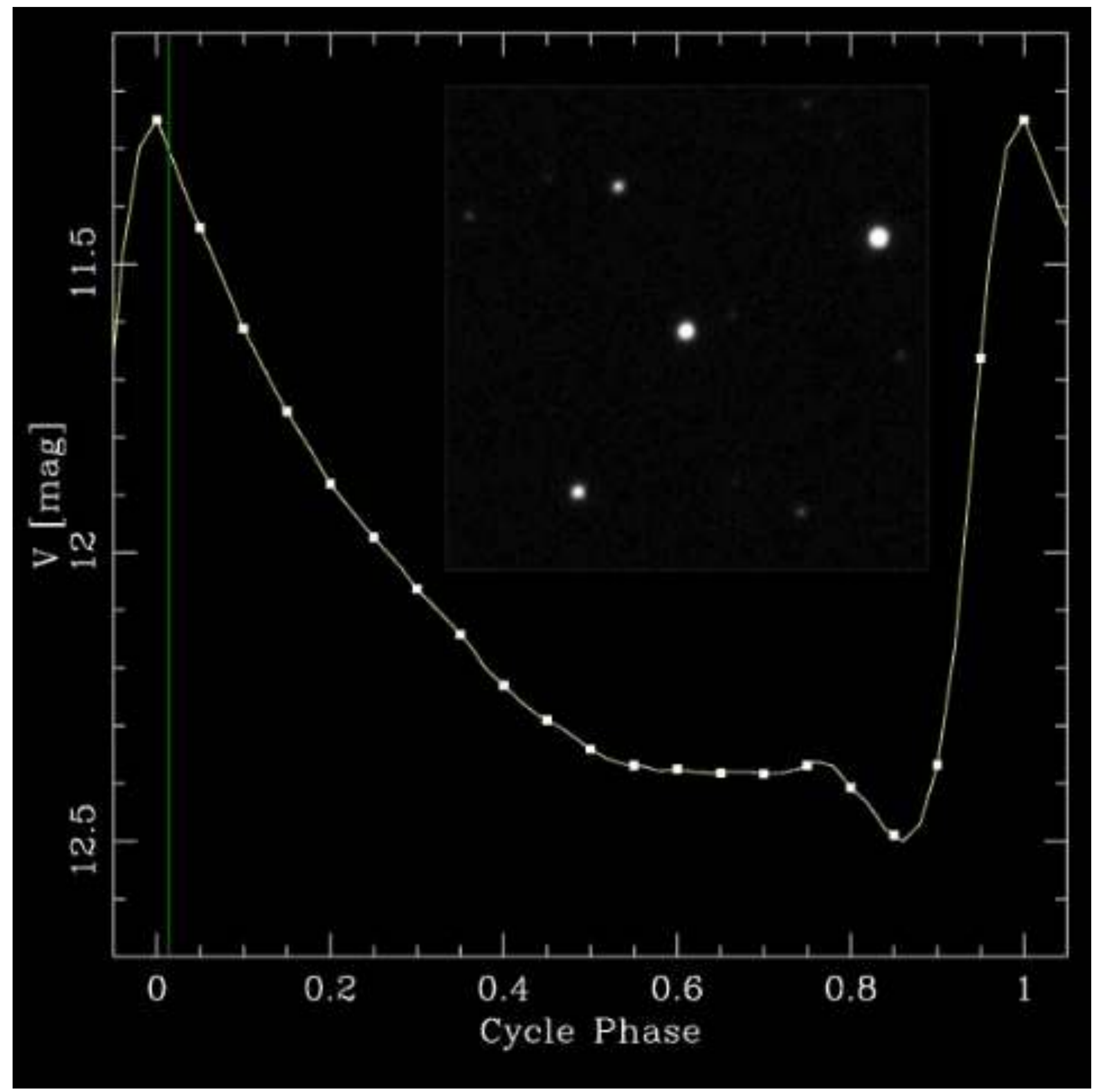


Features

Samples

	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



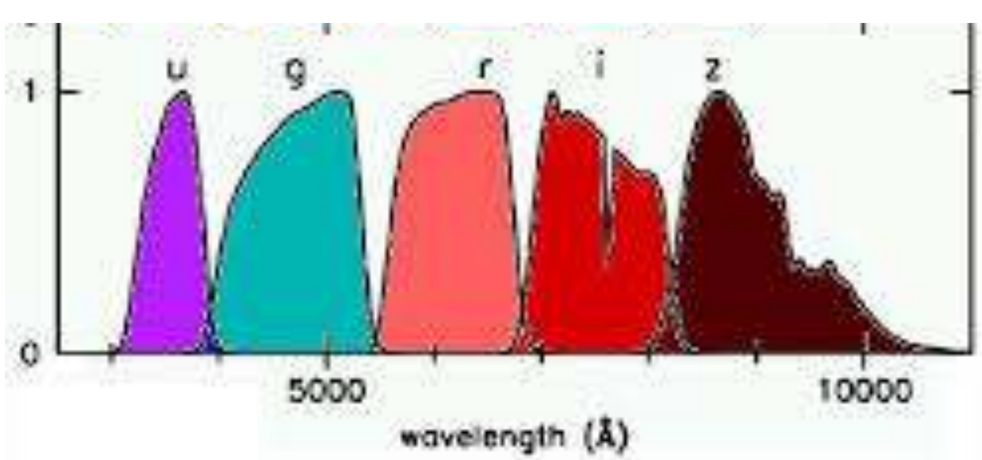


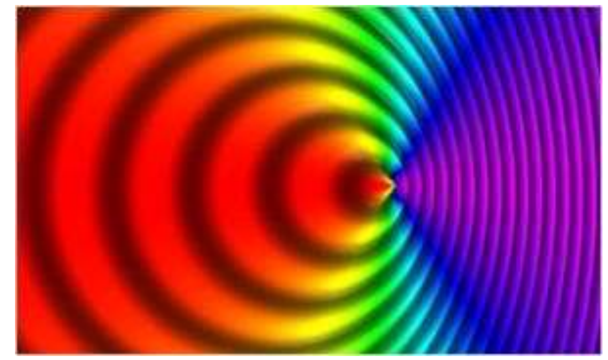
Or Galaxies

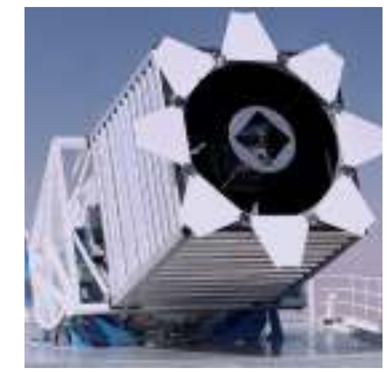
Samples

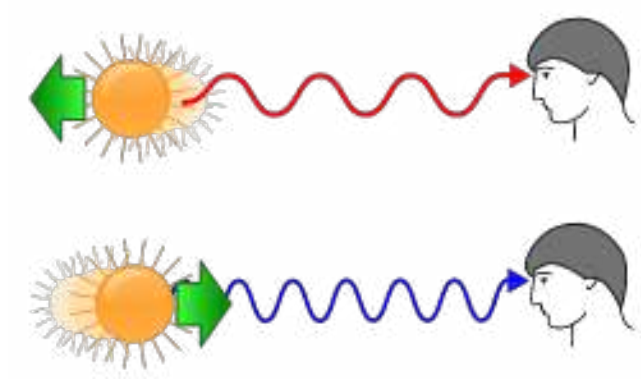
Features

	u-g	g-r	r-i	i-z	redshift
0	1.88235	0.95459	0.44631	0.32659	0.091214
1	1.97871	0.95931	0.46358	0.32285	0.117409
2	1.84007	0.92670	0.40268	0.32295	0.091852
3	1.89717	1.09666	0.47545	0.34684	0.153276
4	0.98144	0.38145	0.34404	0.04365	0.090731
...	...	...	...	...	...
1841292	1.76952	1.16660	0.57567	0.45292	0.091798
1841293	2.04845	1.07646	0.48183	0.41557	0.090638
1841294	1.19402	1.83407	1.08653	0.49640	0.567134
1841295	1.28313	1.80604	0.81906	0.51448	0.445512
1841296	1.67367	0.87776	0.47917	0.33946	0.091880

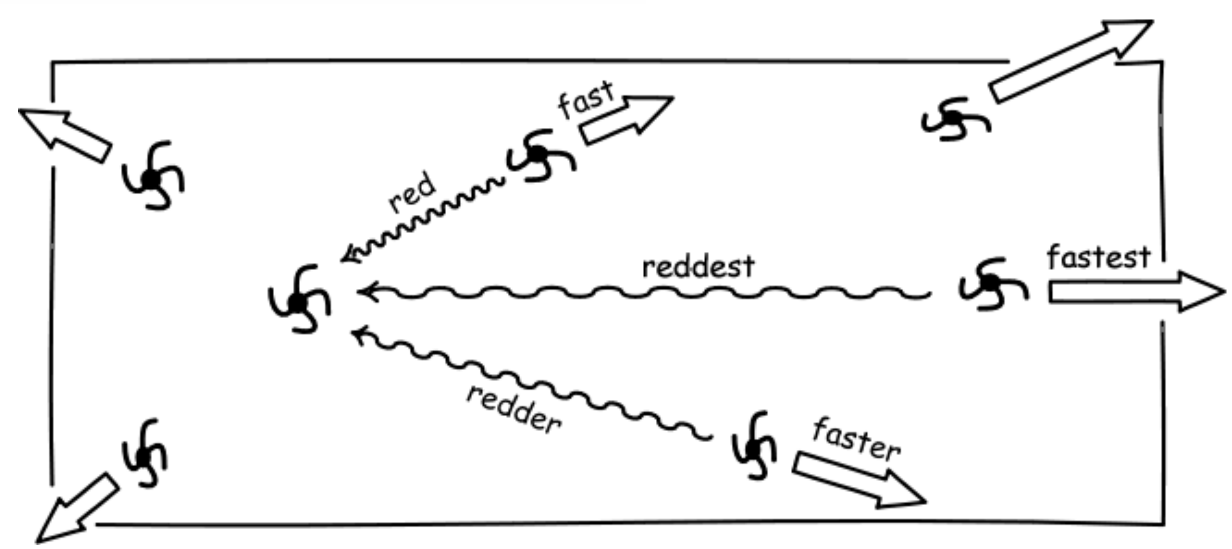








$$1 + z = \sqrt{\frac{1 + \frac{v}{c}}{1 - \frac{v}{c}}}$$



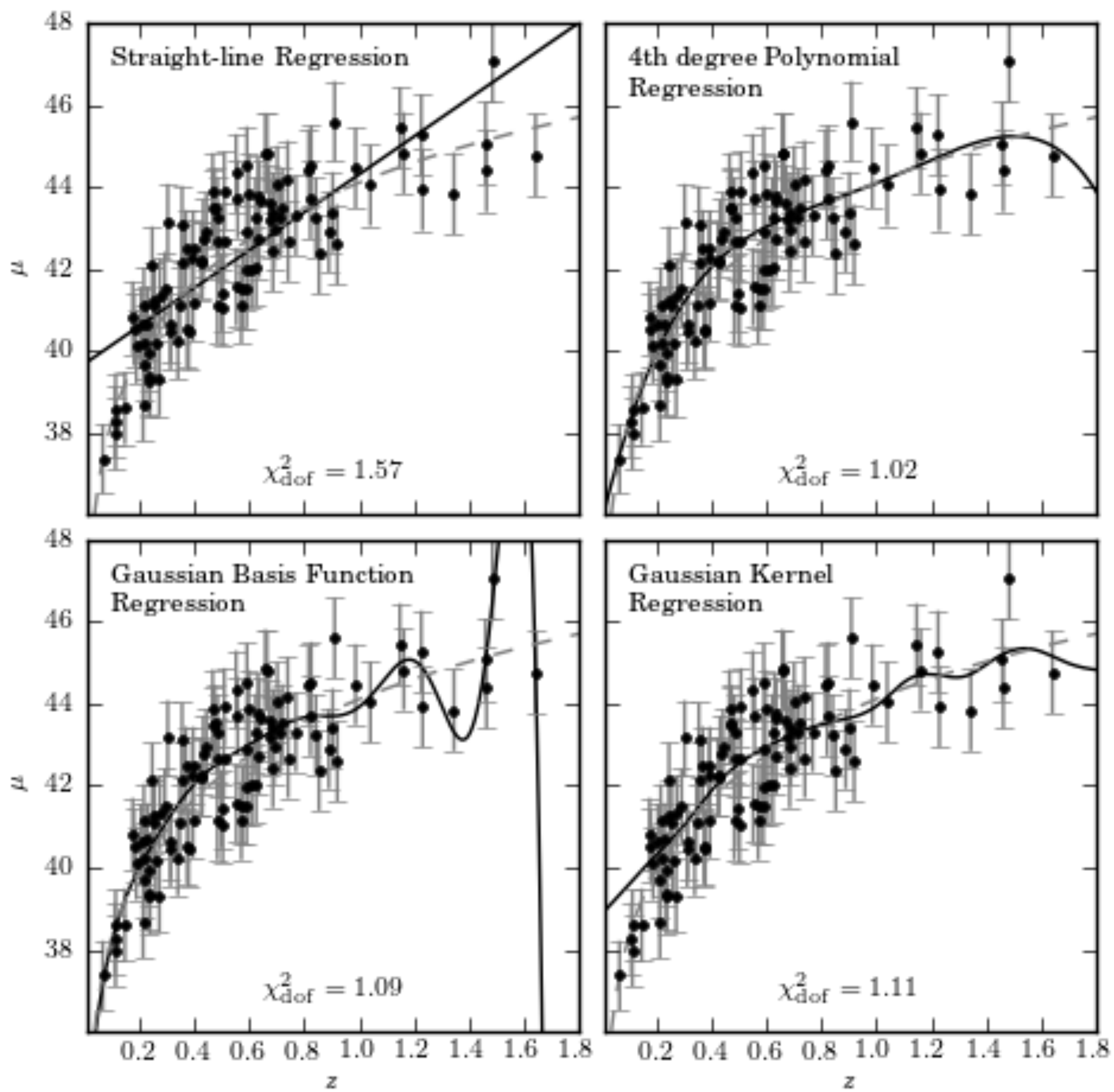
Today

Regression Models

Linear:
$y = ax + b$

Polynomial:
$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n + b$

Basis Functions:
$y = \sum_j w_j \phi_j(x)$
$\phi_j(x) = exp\left(\frac{-(x-\mu_j)^2}{2s^2}\right)$



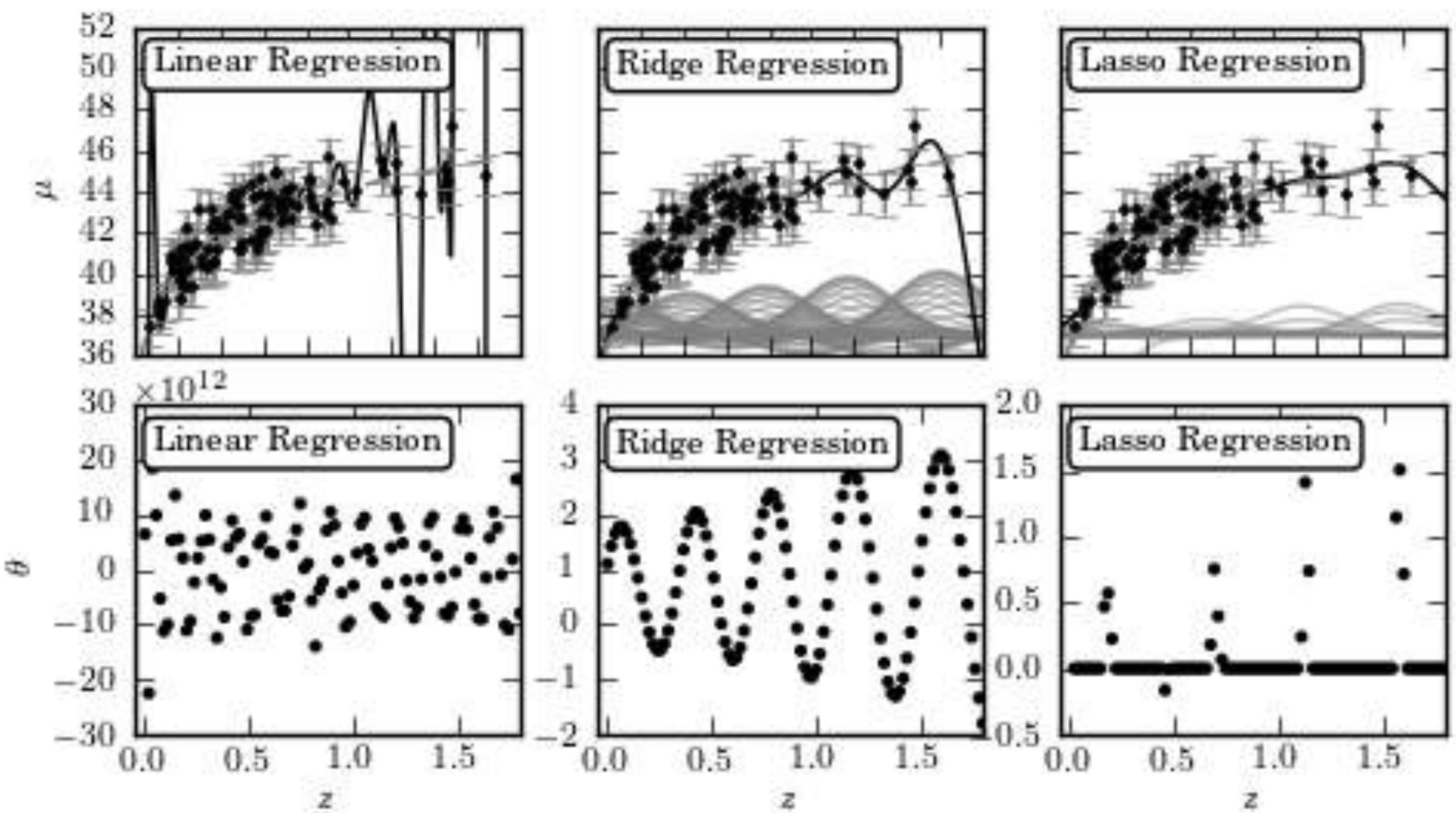
Regularized Regression

Linear:
$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\left(\sum_i (y_i - f(x_i,\theta))^2\right.$$

Ridge (L_2):
$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\left(\sum_i (y_i - f(x_i,\theta))^2 + \lambda \sum_i \theta_i^2\right.$$

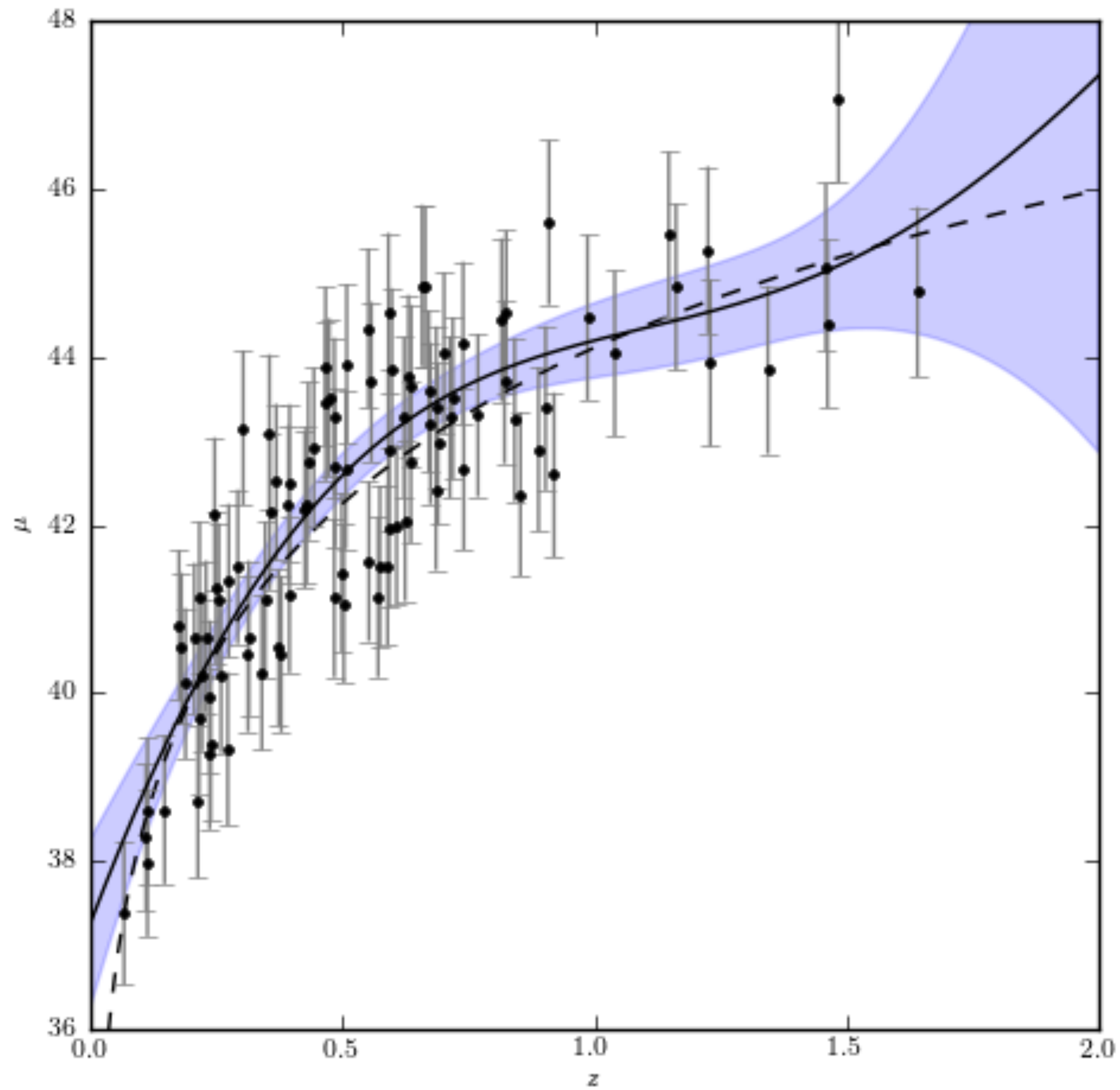
Lasso (L_1):
$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\left(\sum_i (y_i - f(x_i,\theta))^2 + \lambda \sum_i |\theta_i|\right.$$

Sparse (L_0):
$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\left(\sum_i (y_i - f(x_i,\theta))^2 + \lambda \sum_i \mathcal{I}(\theta_i \neq 0)\right)$$

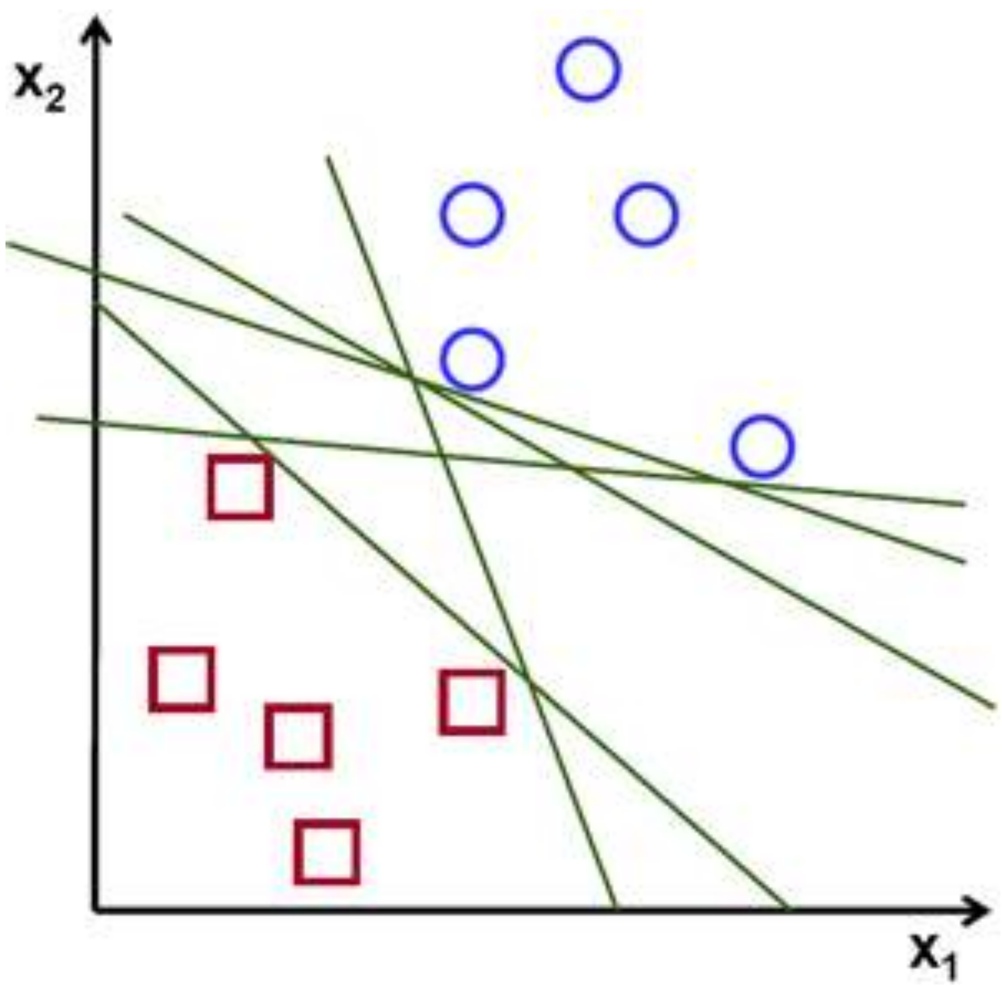


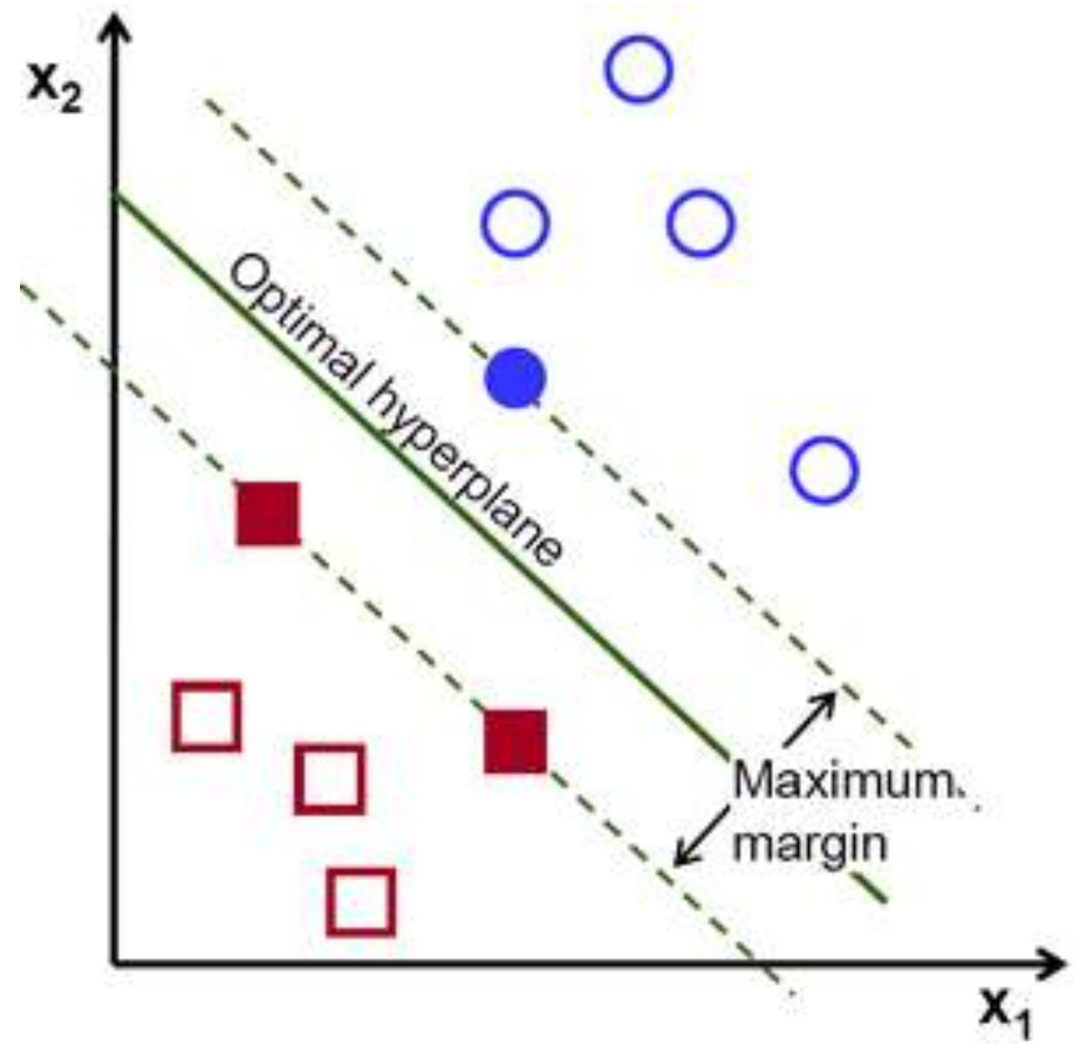
More on Regression

Gaussian Processes

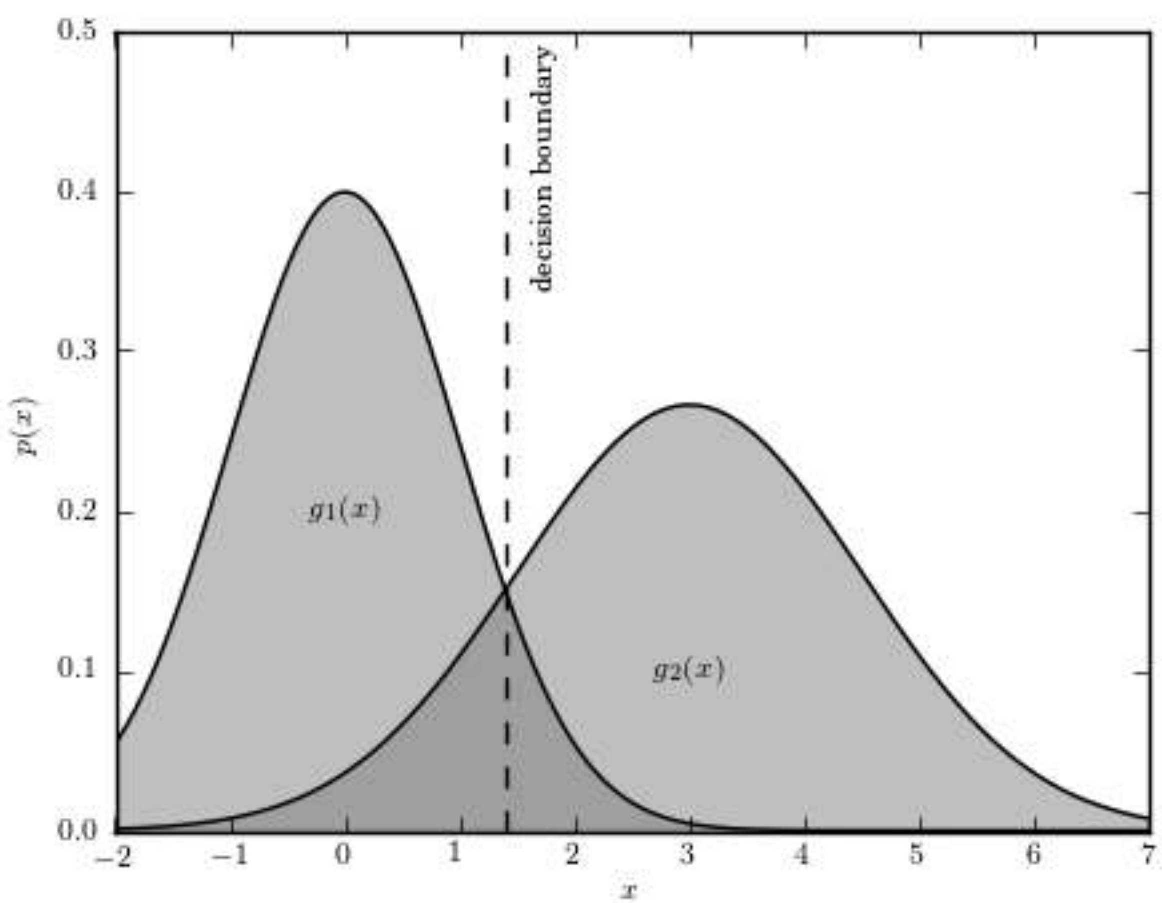


SVMs in a nutshell





Clustering with GMMs in a nutshell



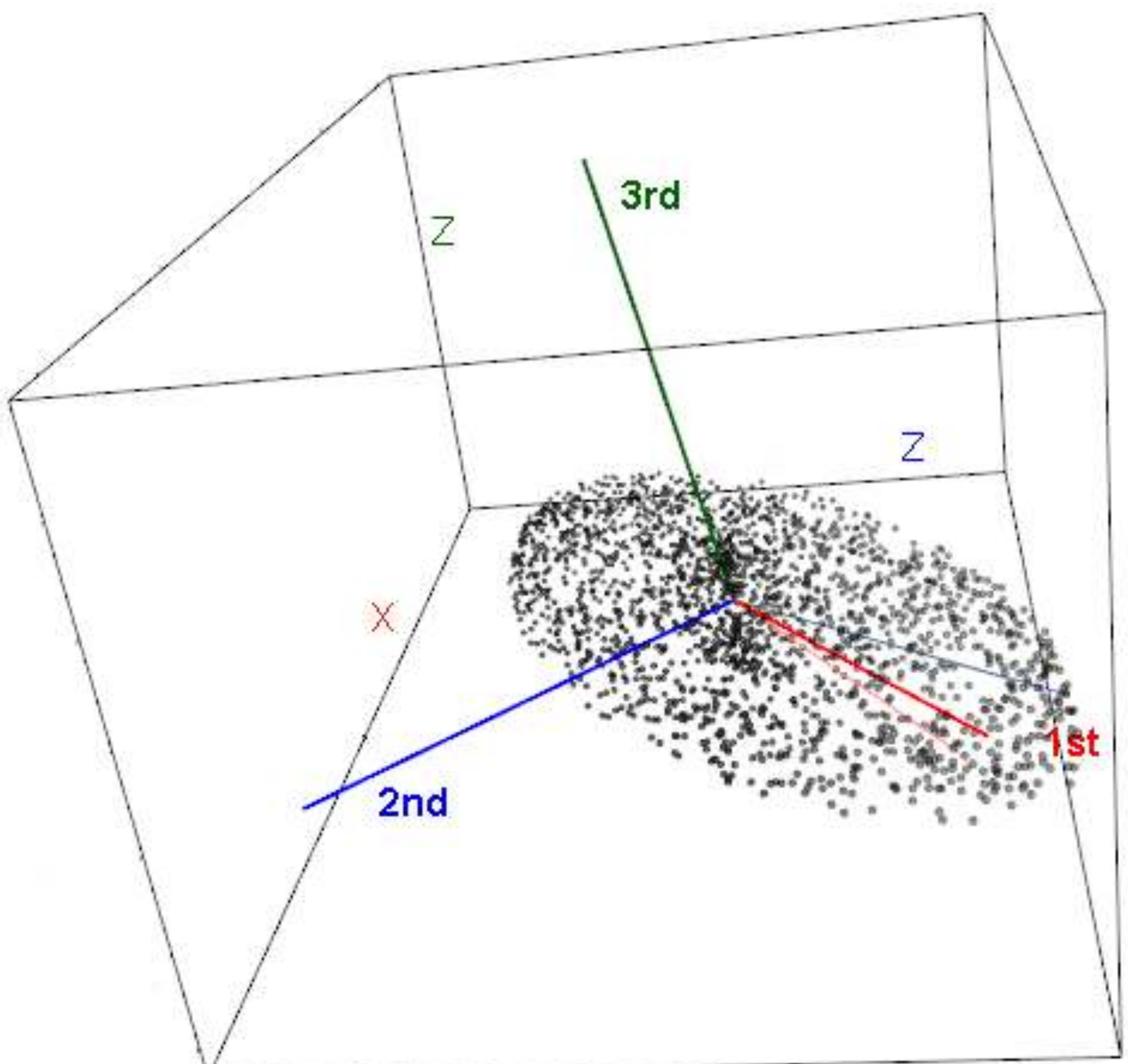
$$\mathcal{N}(x;\mu,\Sigma) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right)}{\sqrt{(2\pi)^k|\Sigma|}}$$

$$p(x) = \sum_{j=1}^{k} \phi_j \mathcal{N}(x;\mu_j,\Sigma_j)$$

$$\sum_{j=1}^{k} \phi_j = 1$$

Dimensionality Reduction using PCA in a nutshell

PCA applied to an ellipsoidically shaped point cloud



$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{W}^T$$

$$\begin{aligned}\mathbf{X}^T\mathbf{X} &= \mathbf{W}\boldsymbol{\Sigma}^T\mathbf{U}^T\mathbf{U}\boldsymbol{\Sigma}\mathbf{W}^T \\ &= \mathbf{W}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\mathbf{W}^T \\ &= \mathbf{W}\boldsymbol{\Sigma}'\mathbf{W}^T\end{aligned}$$

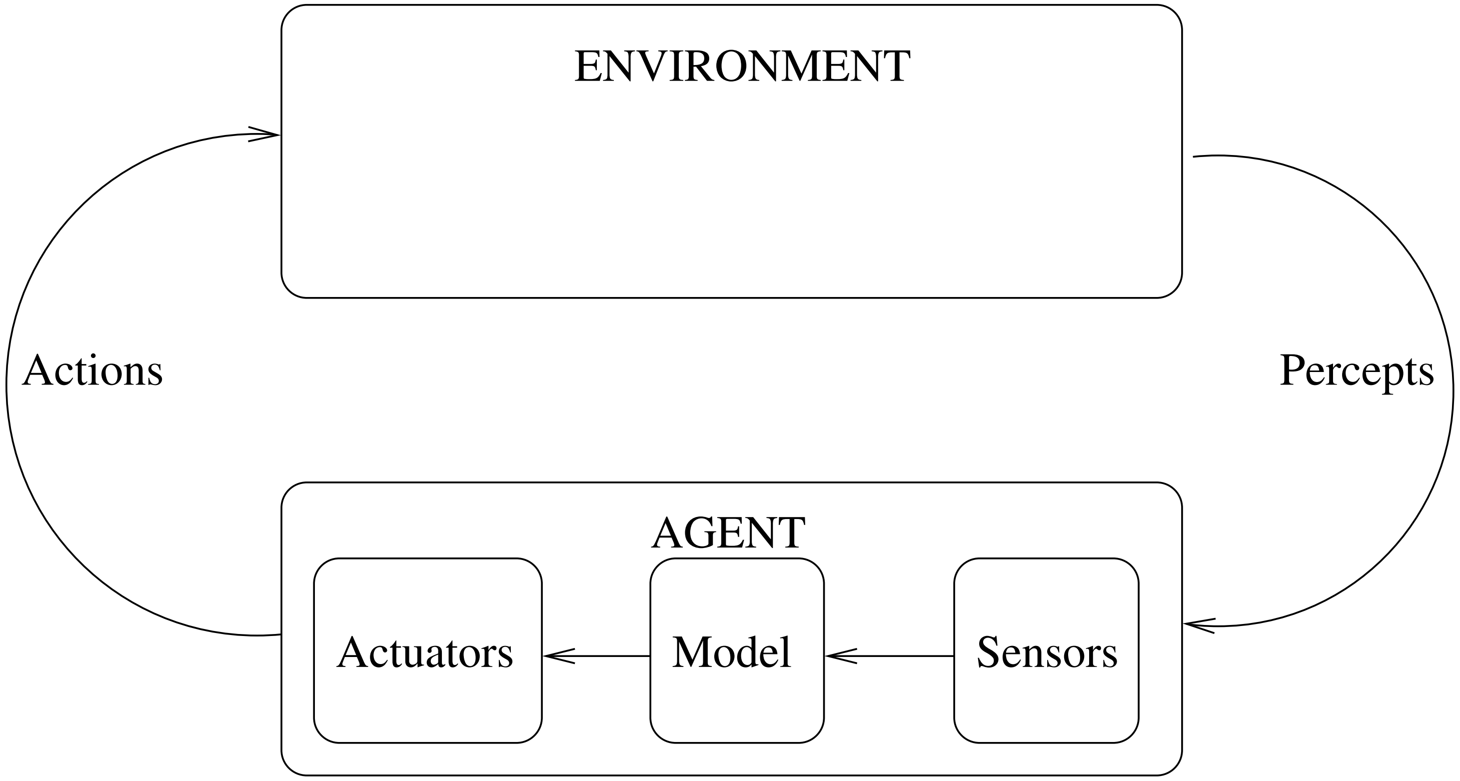
$$\begin{aligned}\mathbf{T} &= \mathbf{X}\mathbf{W} \\ &= \mathbf{U}\boldsymbol{\Sigma}\mathbf{W}^T\mathbf{W} \\ &= \mathbf{U}\boldsymbol{\Sigma}\end{aligned}$$

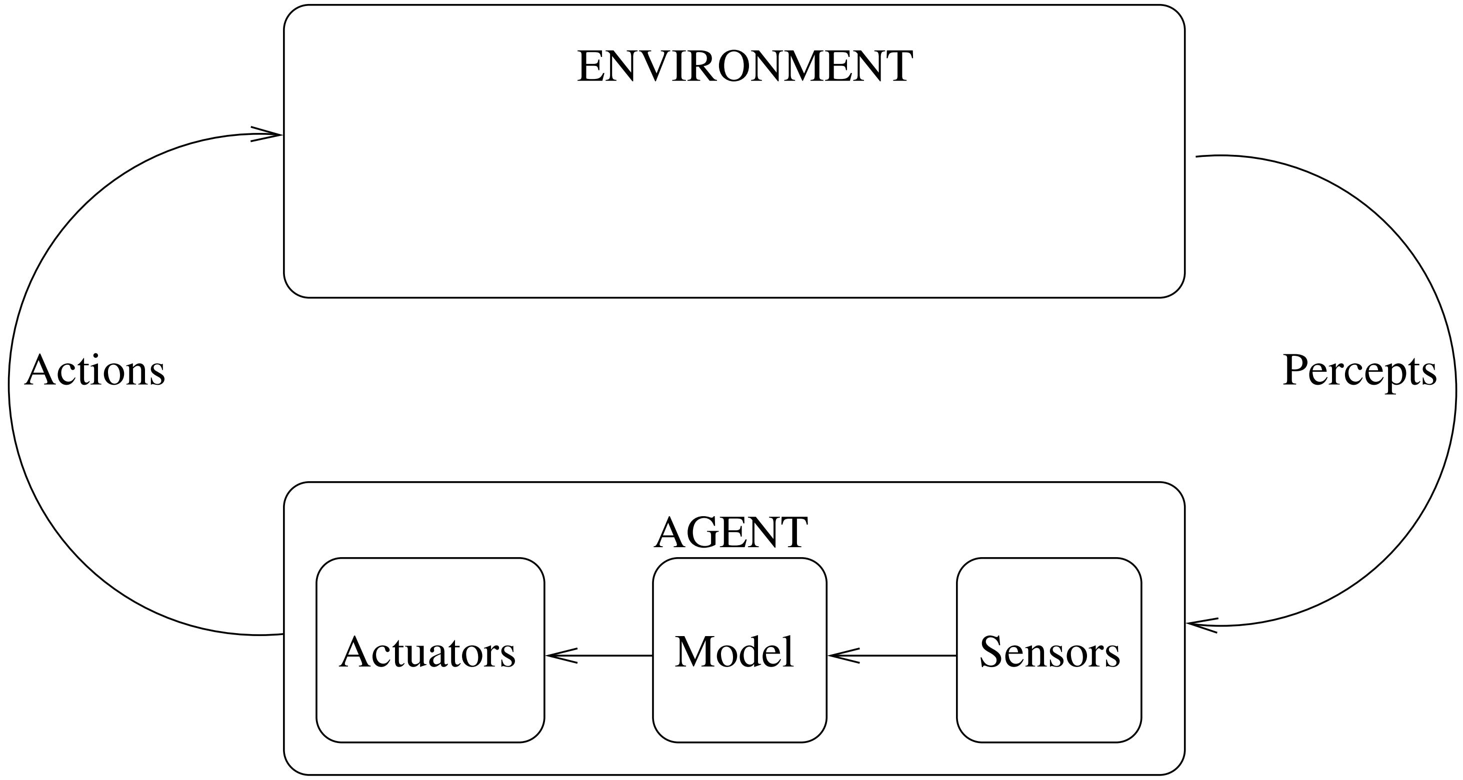
$$\mathbf{T}_L = \mathbf{U}_L\boldsymbol{\Sigma}_L = \mathbf{X}\mathbf{W}_L$$

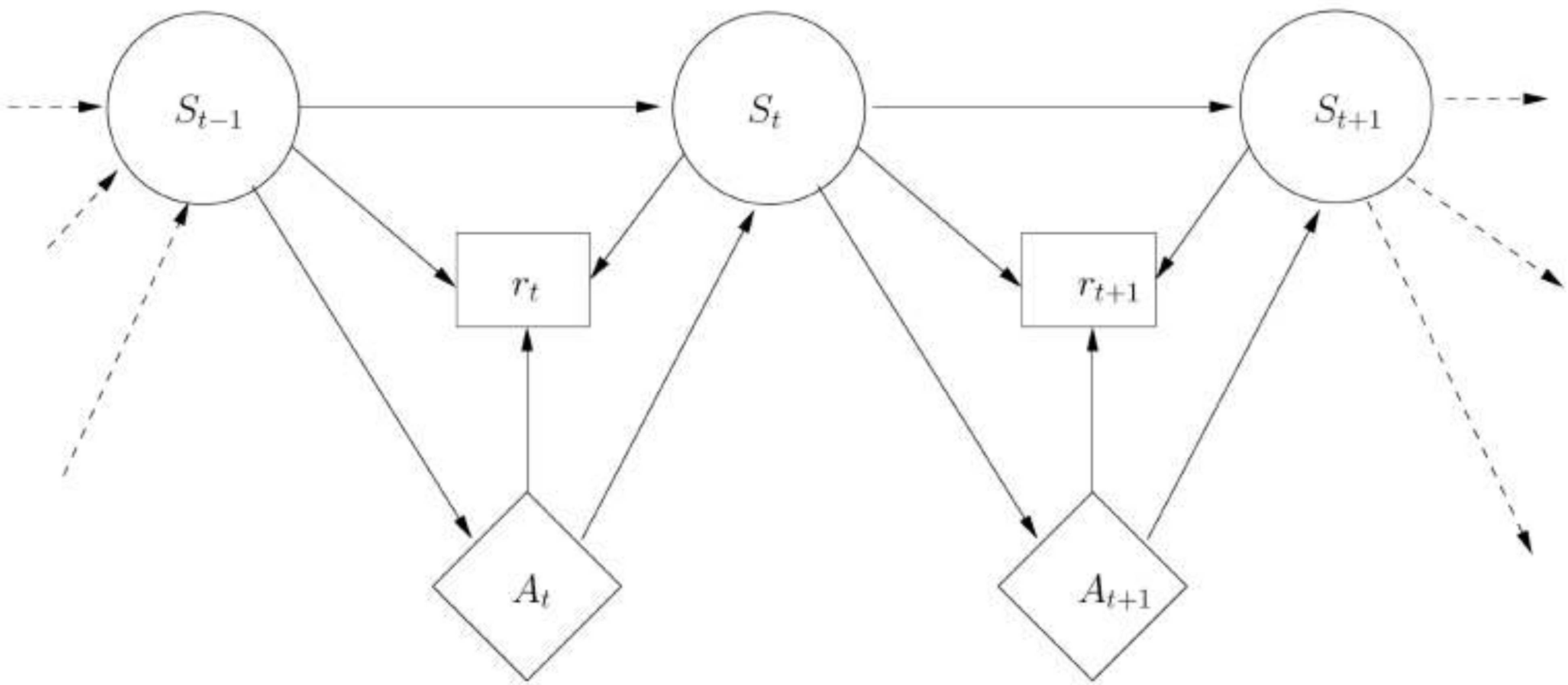
more information: www.joyofdata.de/blog/illustration-of-principal-component-analysis-pca


EXCUSE ME SIR
DO YOU HAVE A MOMENT TO
TALK ABOUT
REINFORCEMENT LEARNING

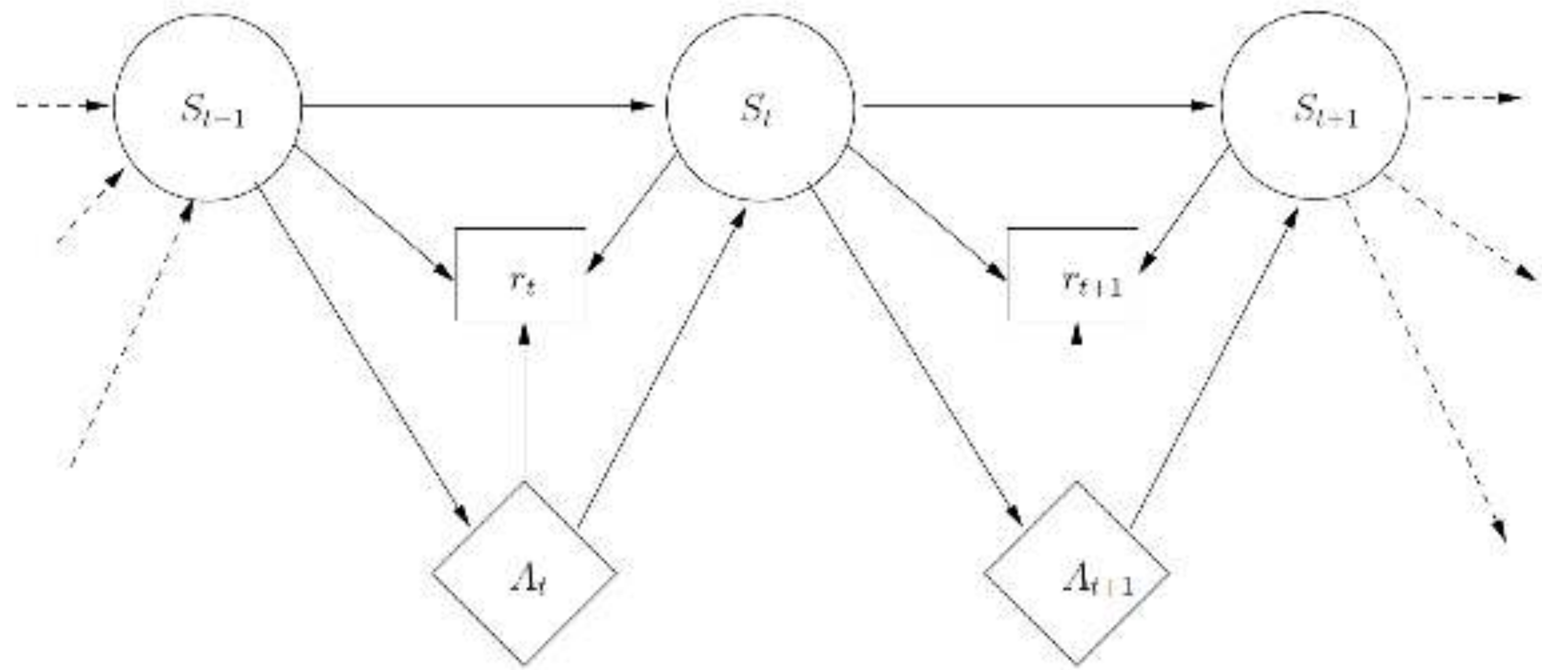
How he/she/it learn?







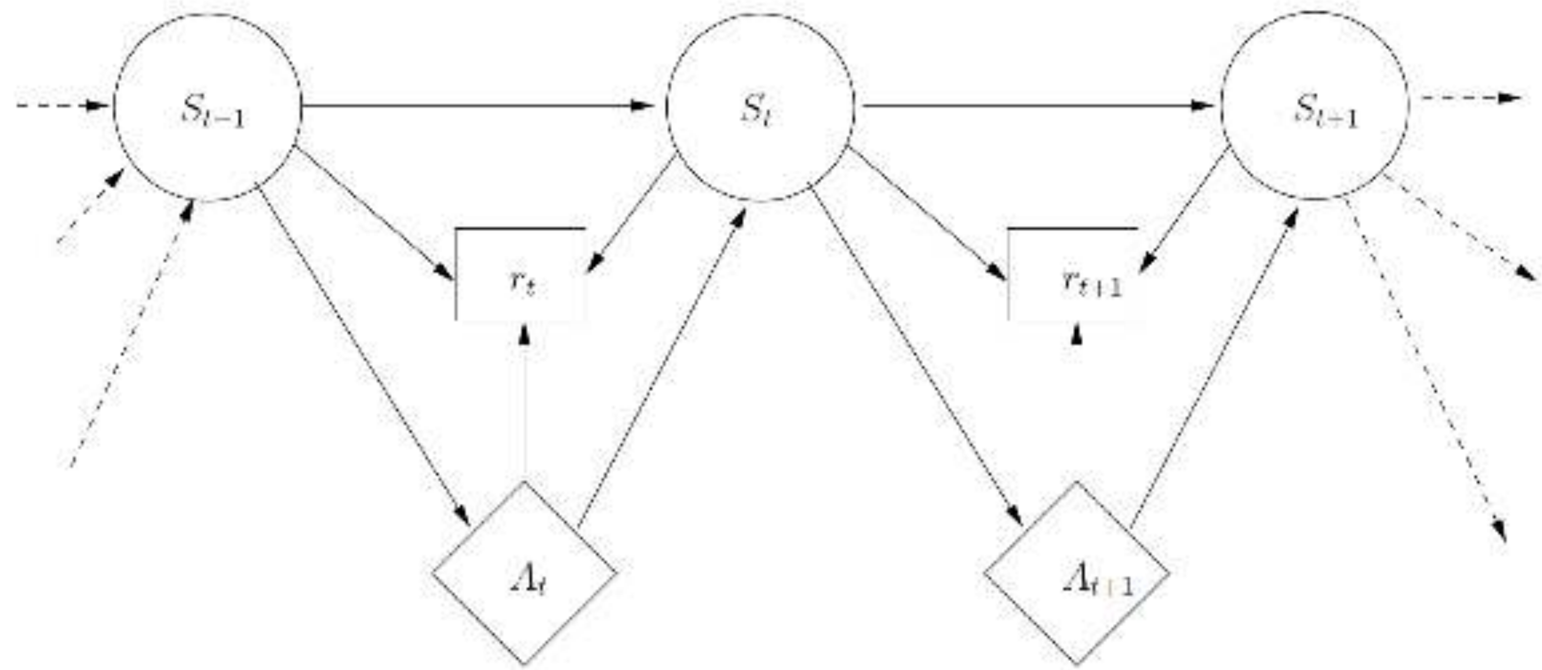
$$\langle \mathcal{S}, \mathcal{A}, T, r, s_0 \rangle$$



$$\langle \mathcal{S}, \mathcal{A}, T, r, s_0 \rangle$$

$$V_H^\pi = \mathbb{E}\left[\sum_{t=0}^{H} r(S_t, \pi_t(S_t), S_{t+1}) \middle| \pi, s_0 \right]$$

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \left\{ \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, \pi(S_t), S_t) \middle| \pi, s_0 \right] \right\}$$



$$\langle \mathcal{S}, \mathcal{A}, T, r, s_0 \rangle$$

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \left\{ \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(S_t, \pi(S_t), S_t) \middle| \pi, s_0 \right] \right\}$$

Bellman's Principle of Optimality (?)

Richard Bellman

"An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision."

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \left\{ \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(S_t, \pi(S_t), S_t) \middle| \pi, s_0 \right] \right\}$$

The Value Function (?)

Richard Bellman

$$V_t^*(s) = \max_{a \in \mathcal{A}} \left\{ \sum_{s' \in \mathcal{S}} T(s,a,s') \left[r(s,a,s') + \gamma V_{t-1}^*(s') \right] \right\}$$

Dynamic Programming, Linear Programs, Value Iteration, Policy Iteration, etc...

- Two types of uncertainty
  - **Aleatory Variability:** T or r are stochastic, e.g. $Pr(S'|S,A)$
  - **Epistemic Uncertainty:** S is latent, T is unknown, e.g. $Pr(S;b), Pr(S'|S,A;b)$

- **Epistemic Uncertainty:** S is latent, T is unknown, e.g.

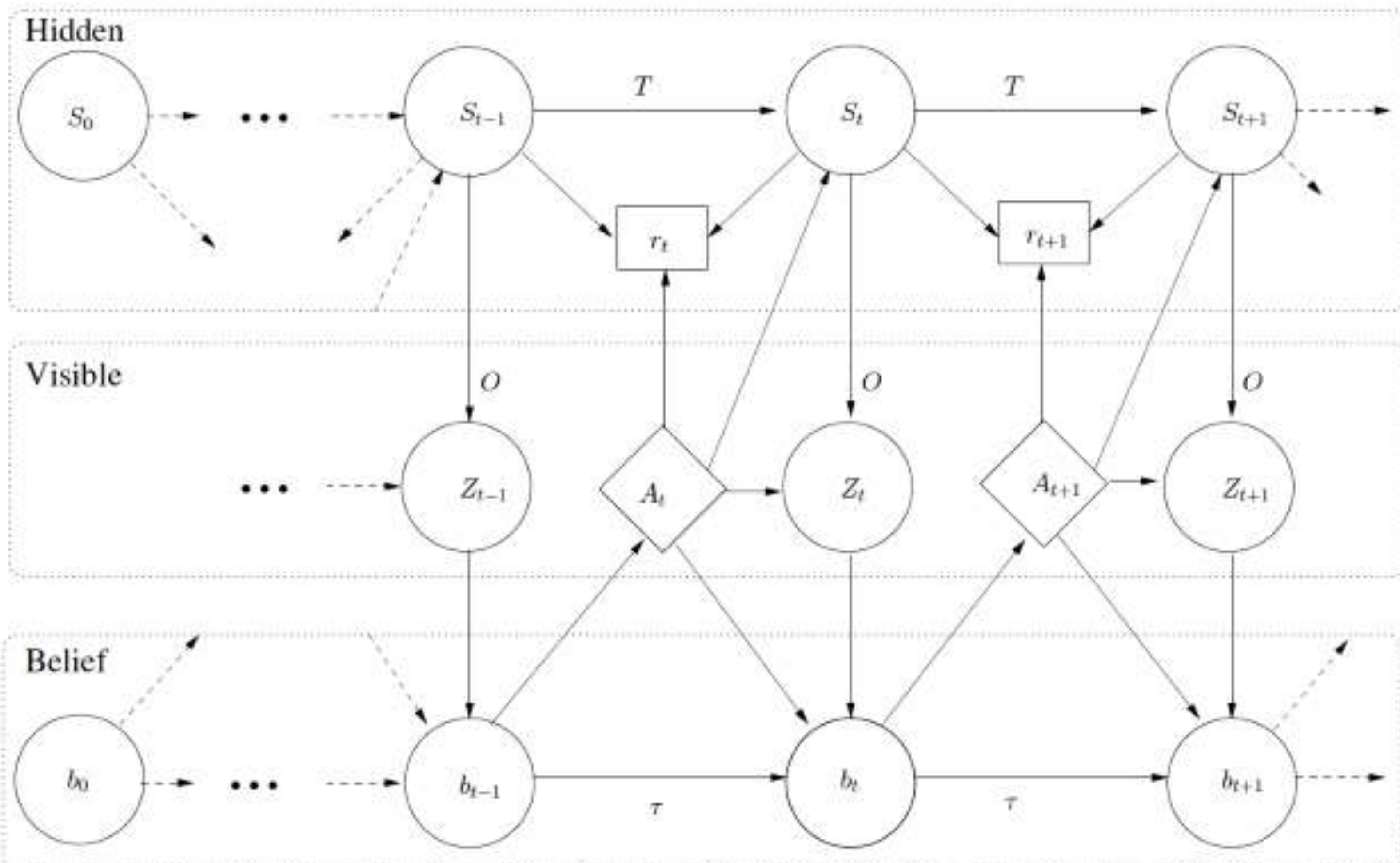
$$Pr(S; b), Pr(S'|S, A; b)$$

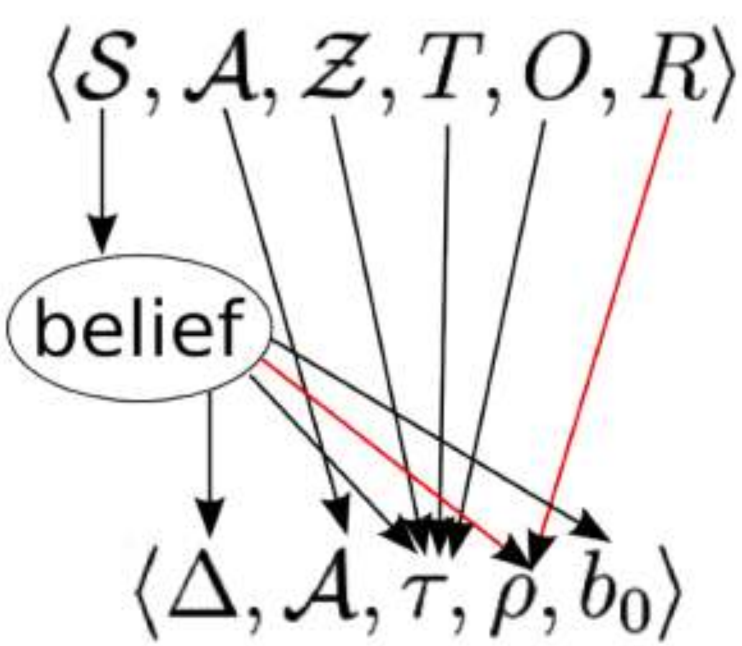
POMDP

States are hidden $\longrightarrow$ Observations

Observation Function: $O(s', a, z) = Pr(z|s', a)$

$$\langle \mathcal{S}, \mathcal{A}, \mathcal{Z}, T, O, r \rangle$$





- **Epistemic Uncertainty:** S is latent, T is unknown, e.g.

$$Pr(S; b), Pr(S'|S, A; b)$$

Reinforcement Learning (RL) (?)

- Let $M = \langle \mathcal{S}, \mathcal{A}, T, R, s_0 \rangle$ be an MDP, and π a policy.
- $V_T^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \middle| s_0 = s, T \right]$
- How to find a π that maximizes V if T is unknown?
- Dilemma: to learn or to gain rewards?

Richard Sutton

Bayesian Model-based RL (?)

- T is a random variable $\sim Pr(T|\mathbf{b})$
- The parameters $\mathbf{b}$ are the **belief over the models**
- $\mathbf{b}$ is updated using the Bayes rule
- Optimal Bayesian policy, $\underset{\pi}{\operatorname{argmax}}\, \mathbb{V}^\pi(s_t, \mathbf{b})$
- No **exploration/exploitation** dilemma

Michael O. Duff