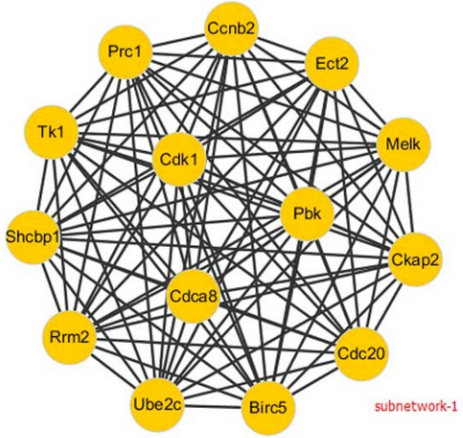
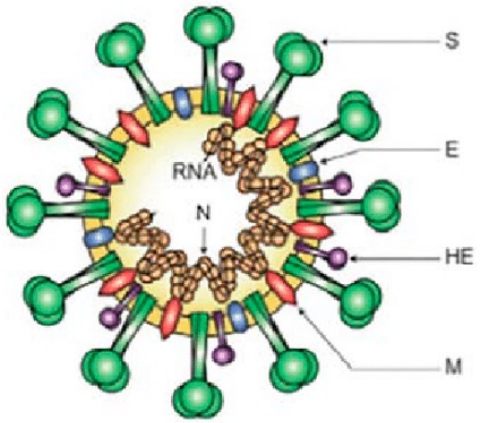
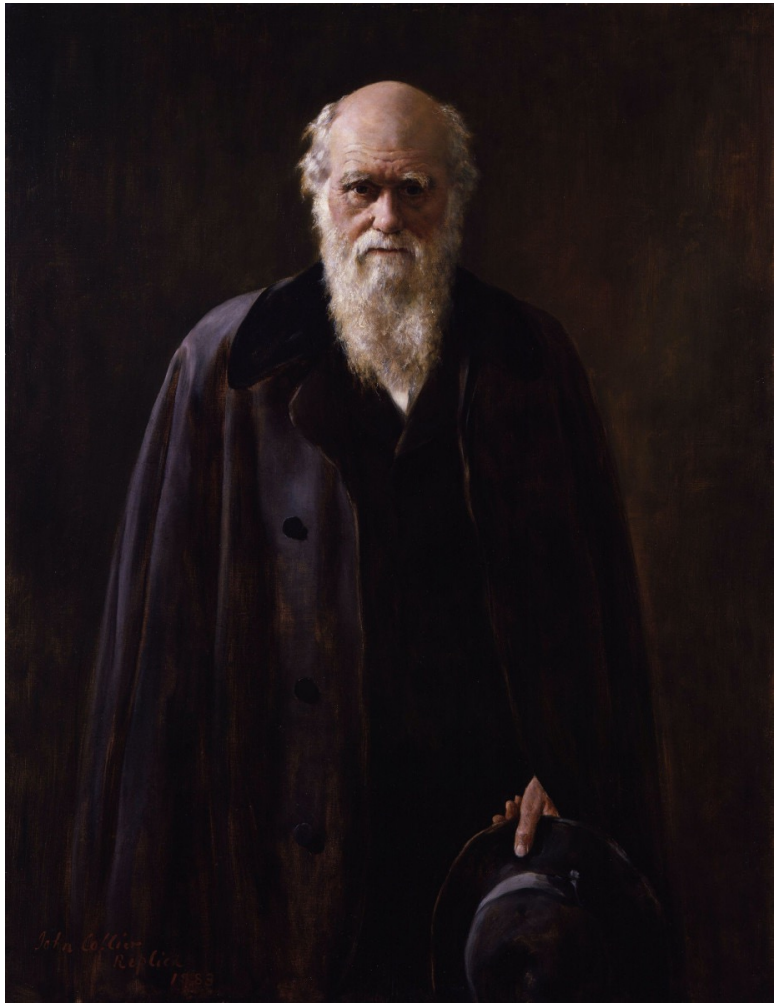
Data Science and Evolutionary Computing

Special Topics: Advanced to Programming with Data



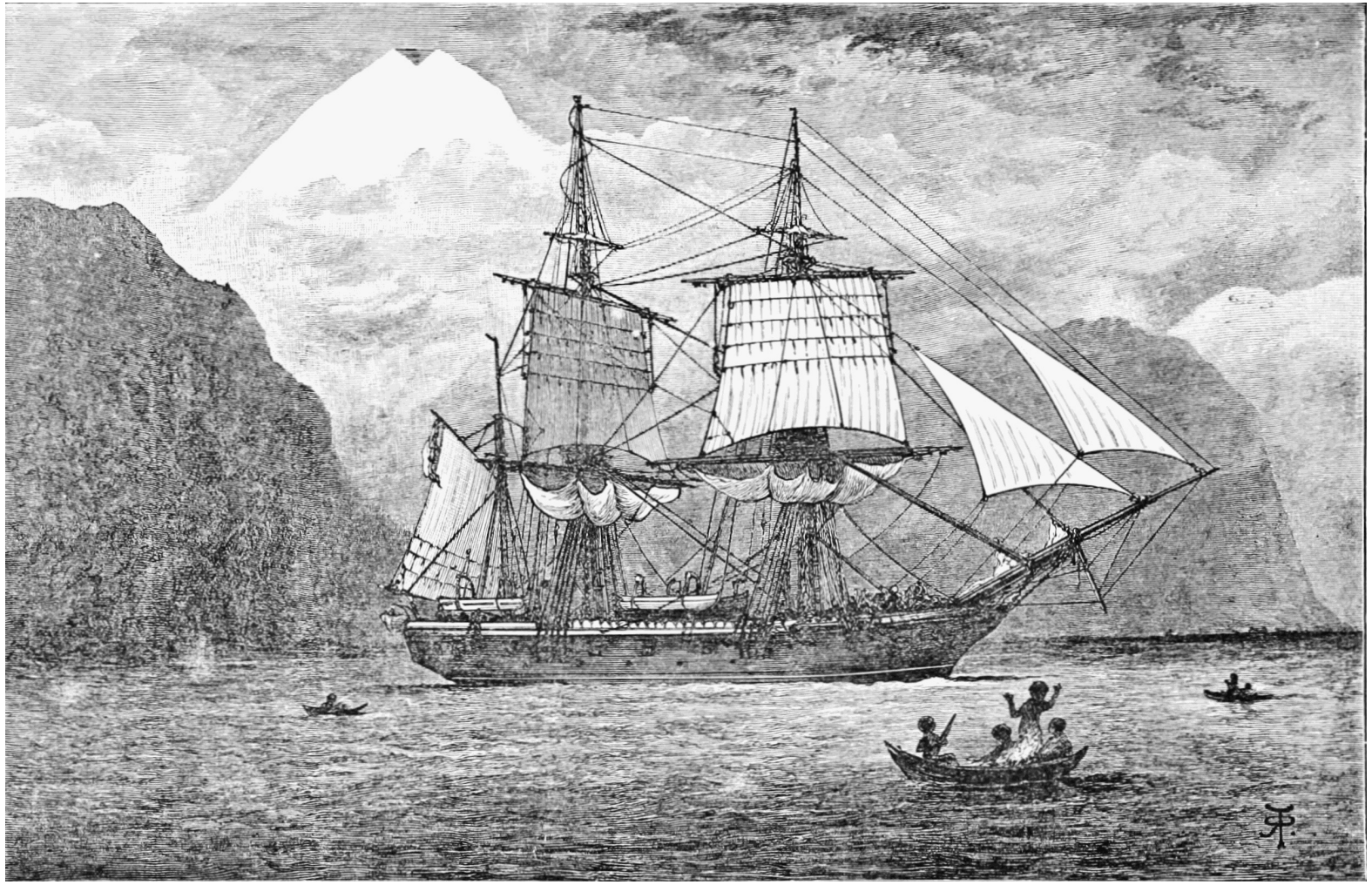


Charles Darwin



1809 - 1882

On the Origin of Species (1859)



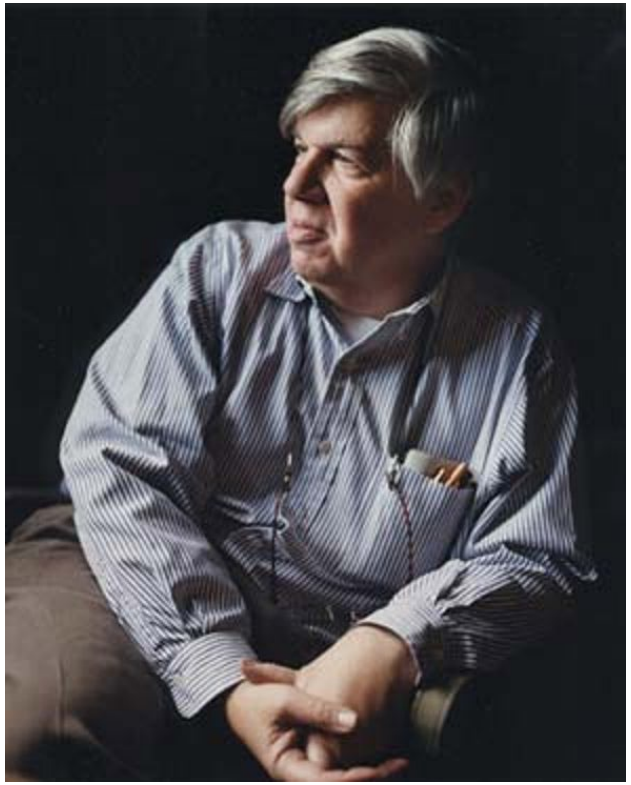
HMS Beagle

Voyage of the Beagle: 1831 – 1836

Natural Selection

[The] basis of natural selection is simplicity itself – two undeniable facts and an inescapable conclusion.

-- Ever Since Darwin (1977)



Stephen Jay Gould
Harvard Paleontologist
1941 - 2002

Variation and Inheritance

1. Organisms vary, and these variations are inherited (at least in part) by their offspring.

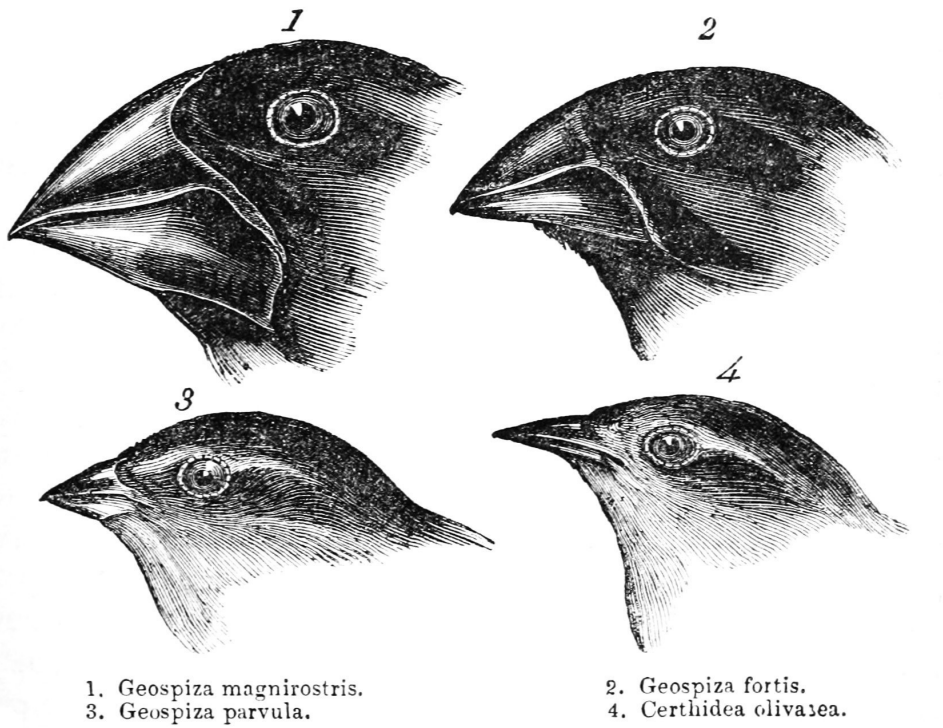
Population explosion

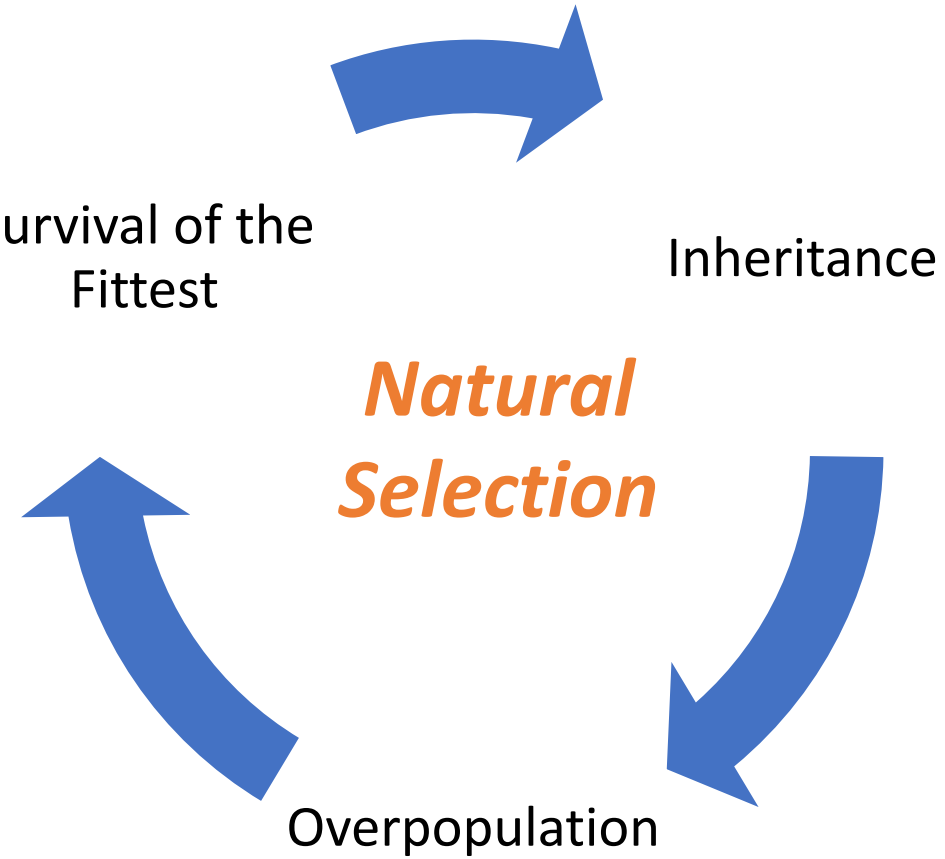
2. Organisms produce more offspring than can possibly survive.

Survival of the Fittest

3. On average, offspring that vary most strongly in directions favored by the environment will survive and propagate.

Evolution by Natural Selection

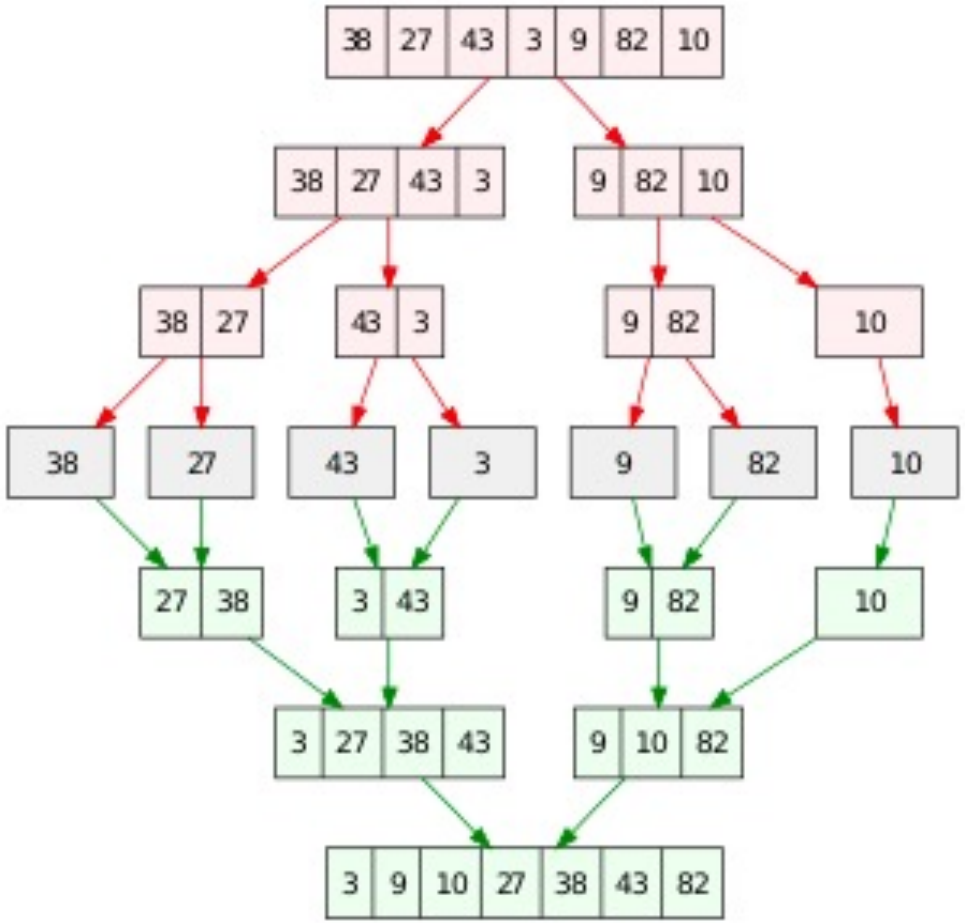




Biological systems are…

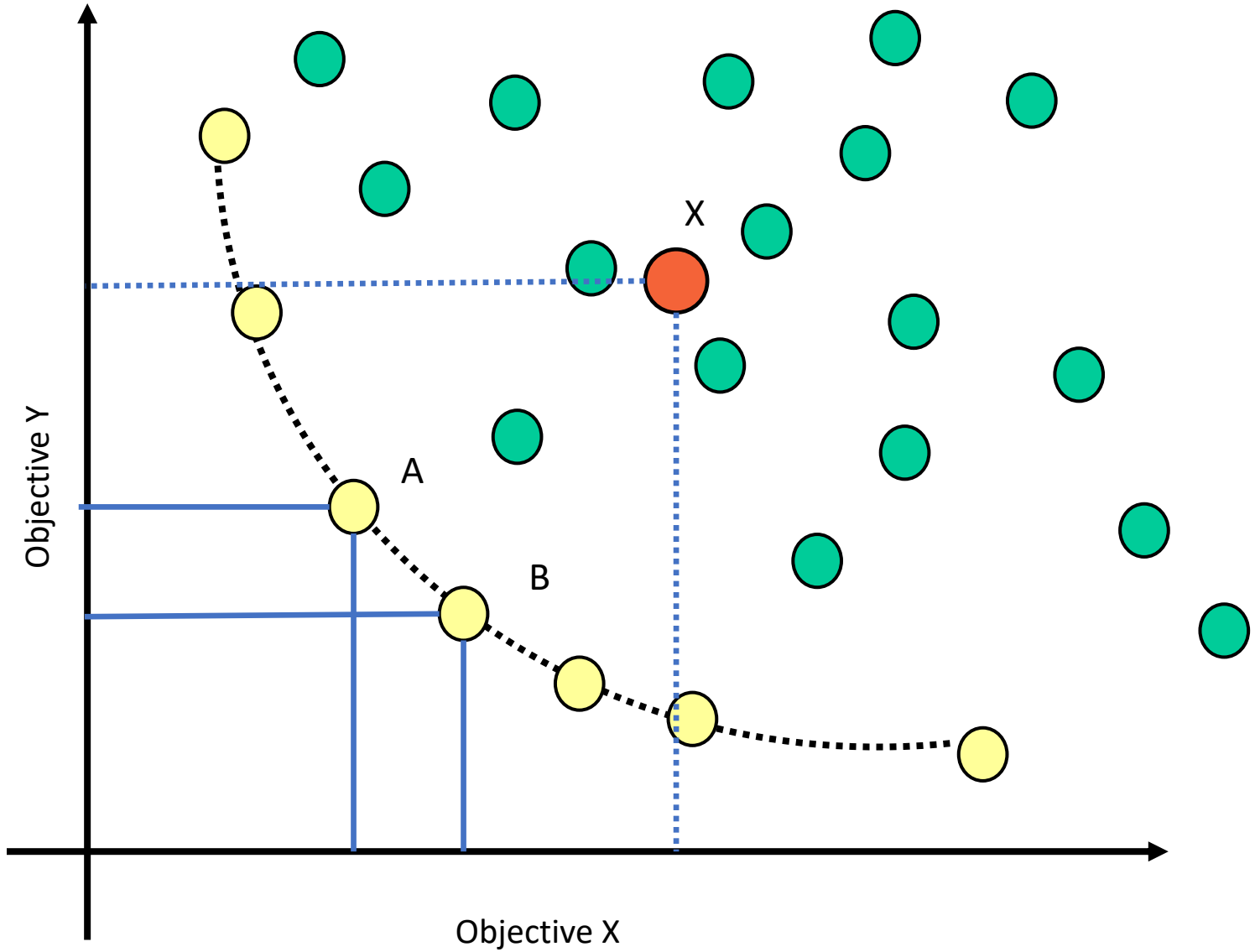
- Complex
- Diverse
- Adaptive
- Resilient
- Self-regulating (homeostatic)
- Synergistic (cooperative)
- Self-sustaining
- *CREATIVE!*

Most real-world problems are not like SORT

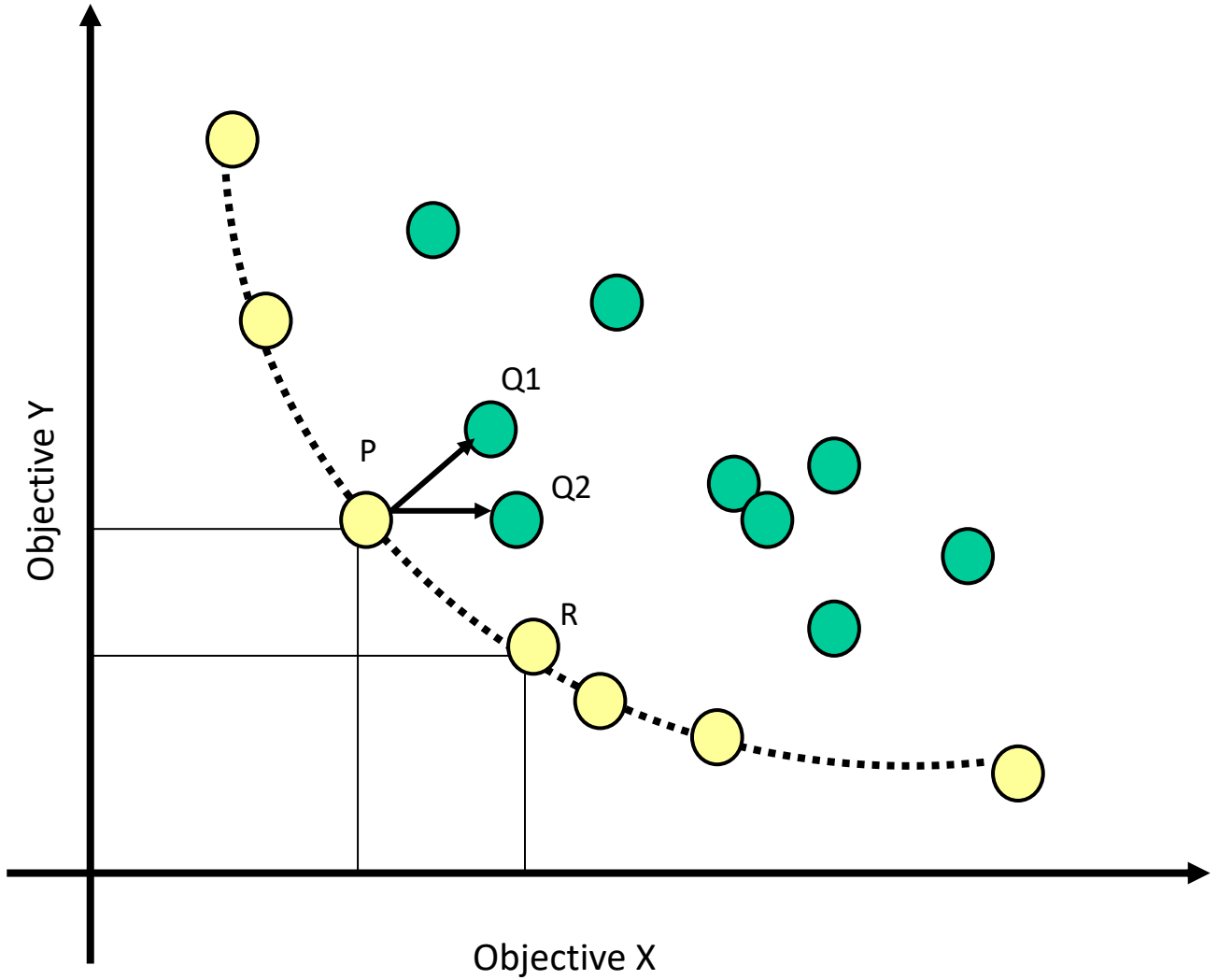


Usually, there is no single OPTIMAL solution...

Evolving Tradeoffs



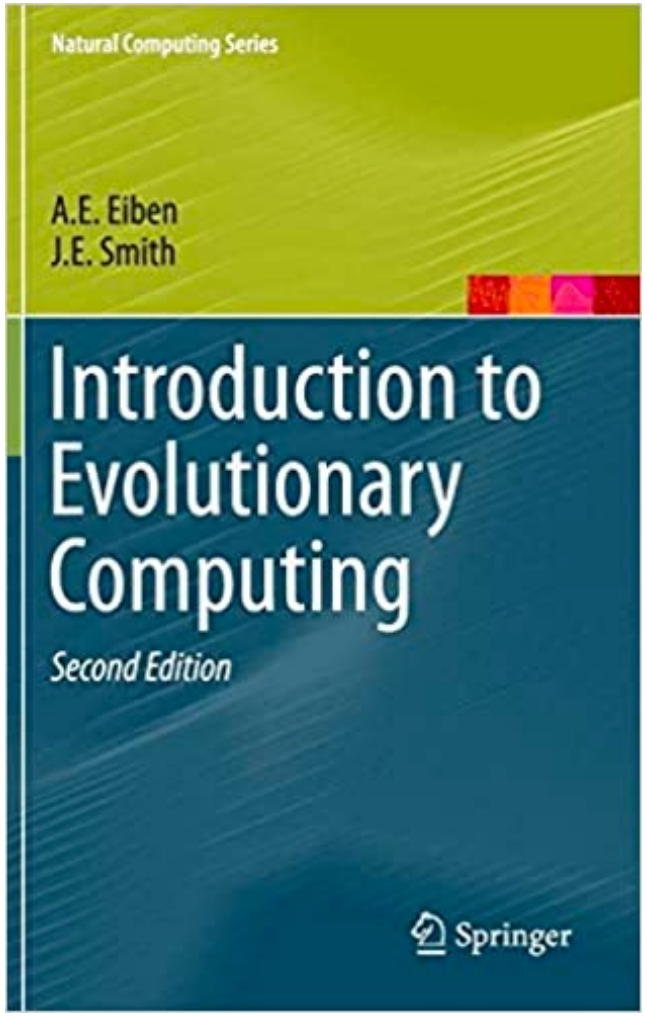
…only tradeoffs.

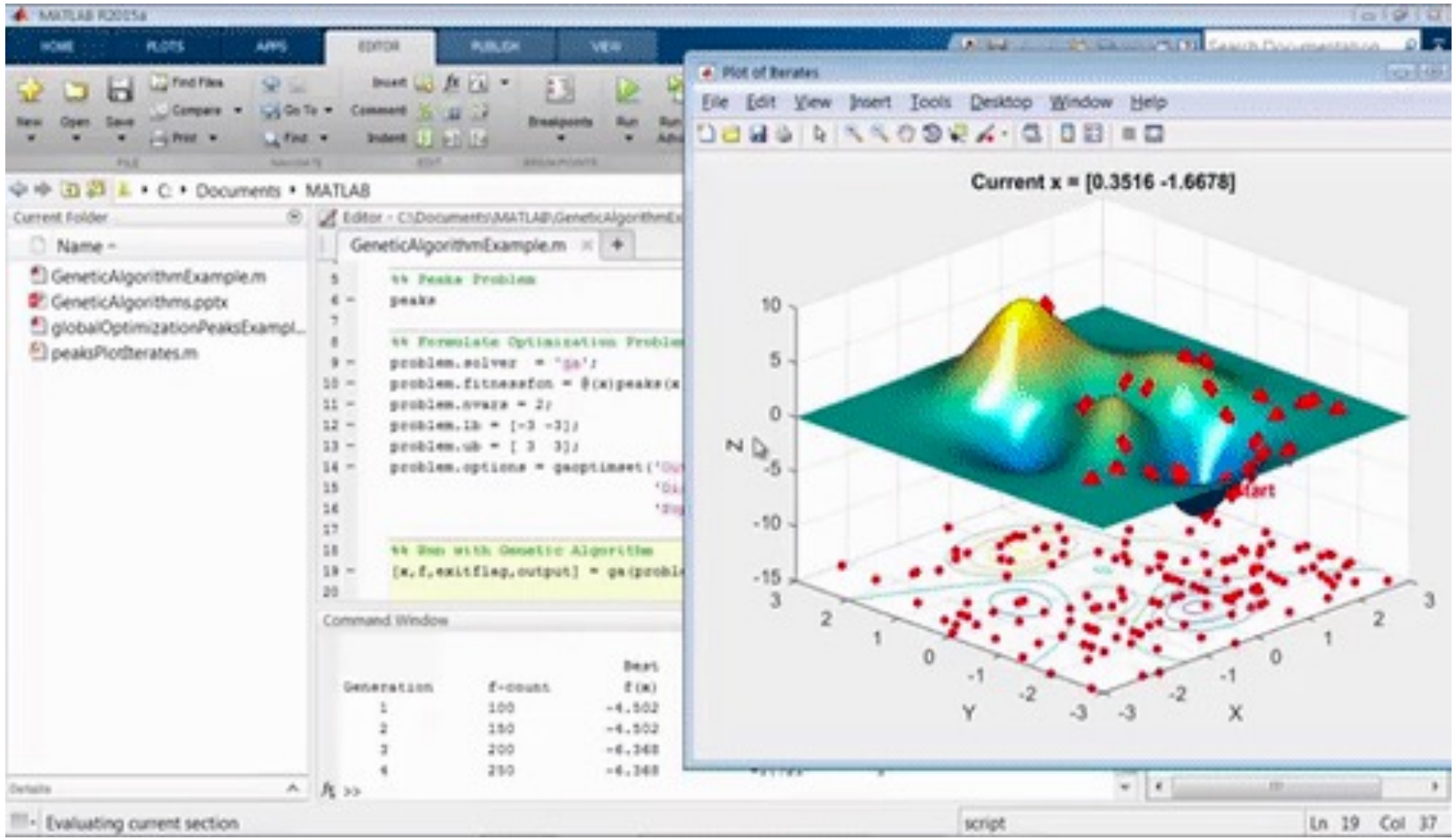


Effective management requires the careful consideration of competing alternatives.

- Peter Drucker

Algorithms that can evolve.





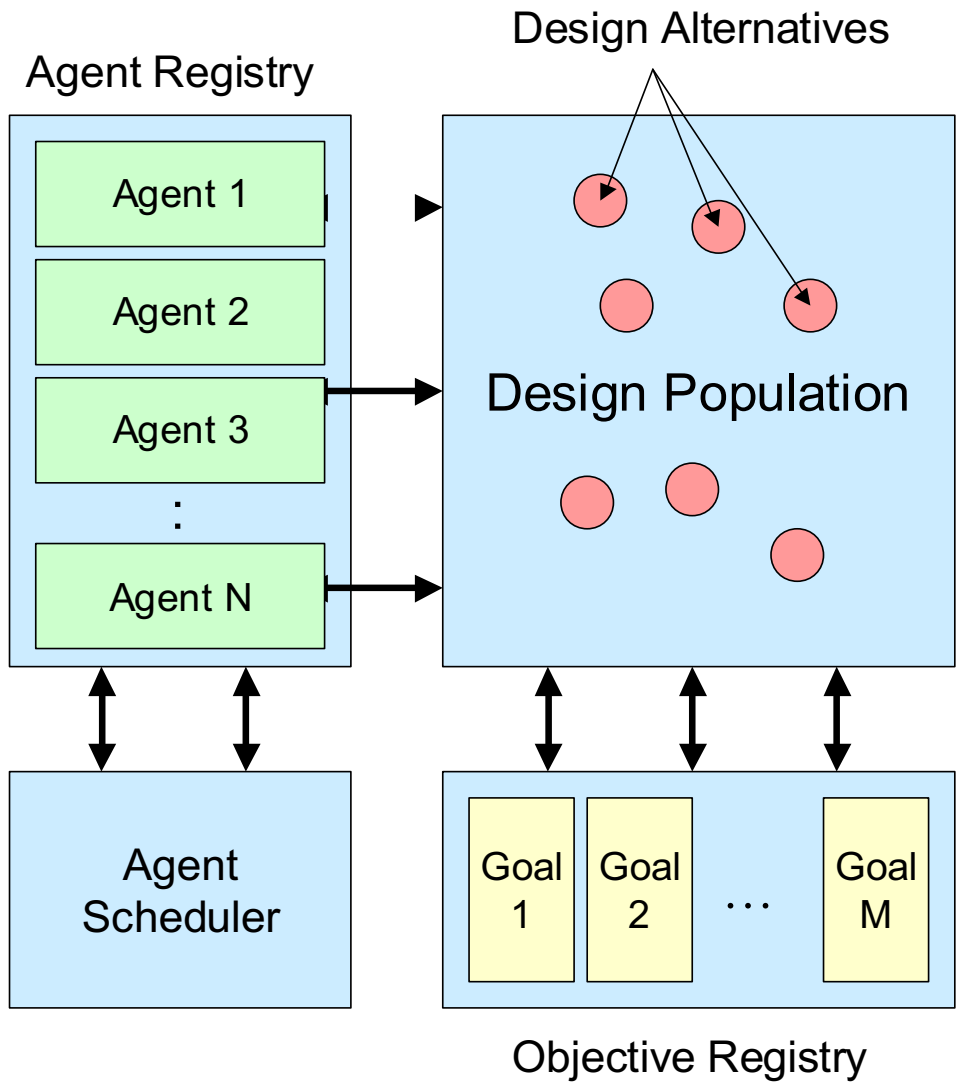
An architecture for multi-objective optimization using evolutionary computing

Classes of agents:

Creator Agent: Adds a design constructed from scratch

Improver Agent: Copies and then modifies an existing design

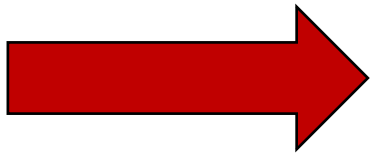
Destroyer Agent: Removes one or more unpromising designs



Sorting without **SORT**

1. Inherit Traits:

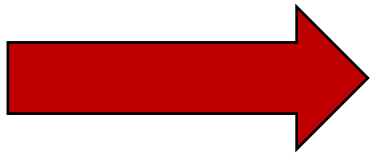
Copy a list

5, 3, 9, 2, 8, 4  **5, 3, 9, 2, 8, 4**

Sorting without **SORT**

2. Create variations:

Swap two elements (of the copy)

5, 3, 9, 2, 8, 4  5, 3, 9, 2, 8, 4

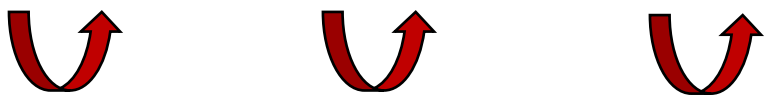
5, 3, 4, 2, 8, 9

Sorting without **SORT**

3. Measure of fitness:

How ***unsorted*** are these lists? Count the **step downs!**

5, 3, 9, 2, 8, 4



2 + 7 + 4 = **13**

5, 3, 4, 2, 8, 9

2 + 2 = **4**

Sorting without **SORT**

3. Survival of the fittest:

How ***unsorted*** are these lists? Count the **step downs!**

5, 3, 9, 2, 8, 4

2 + 7 + 4 = **13**

5, 3, 4, 2, 8, 9

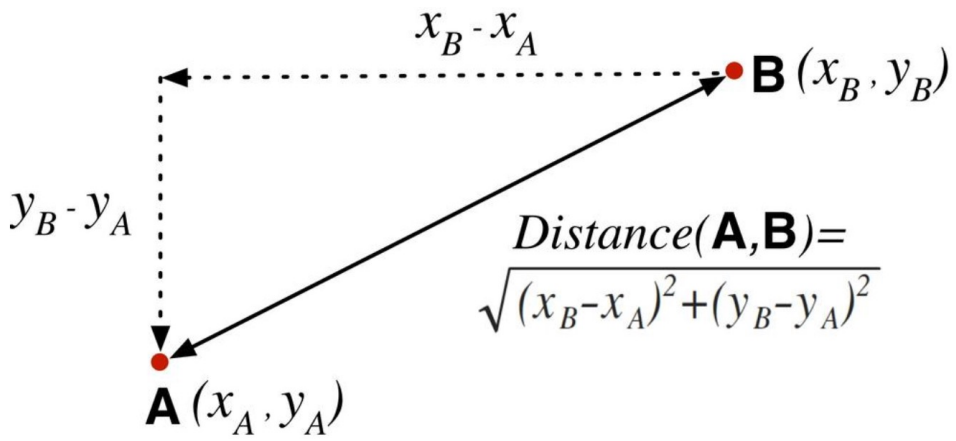
2 + 2 = **4**

k-Nearest Neighbor Learning

DS2000: Intermediate Programming with Data

Euclidean Distance

Attribute	Person A	Person B
Age	23	40
Years at current address	2	10
Residential status (1=Owner, 2=Renter, 3=Other)	2	1

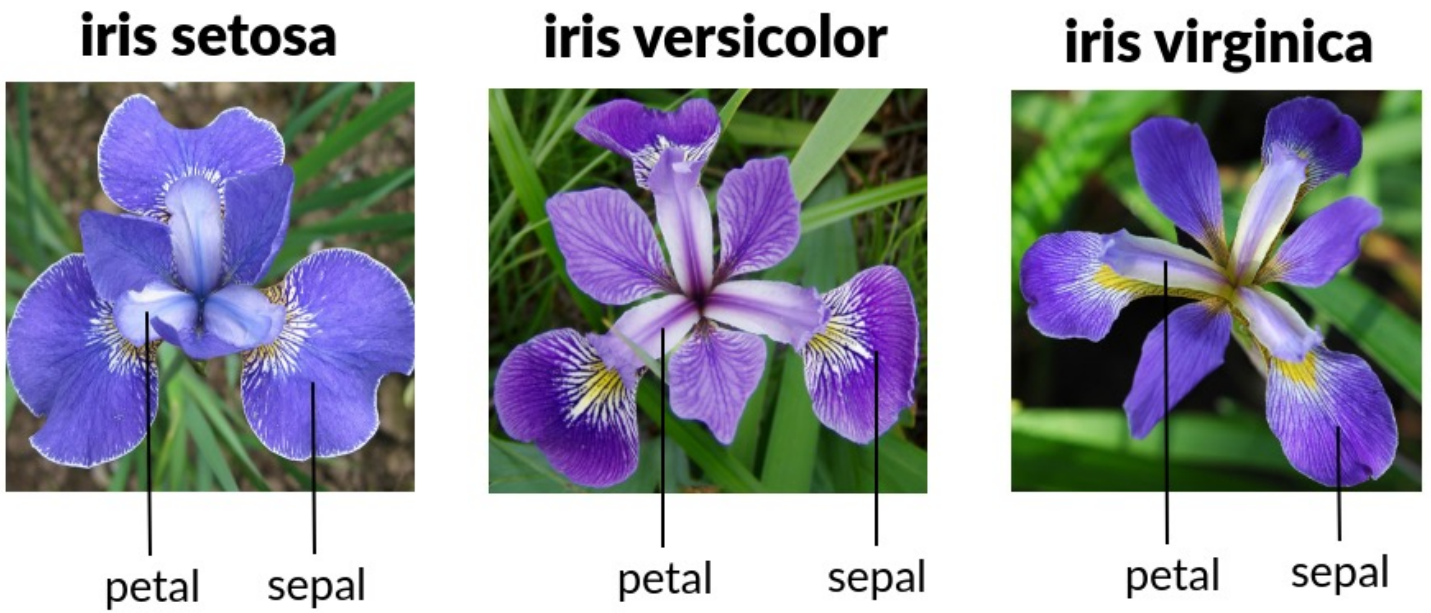


$$d(A, B) = \sqrt{(23-40)^2 + (2-10)^2 + (2-1)^2}$$
$$\approx 18.8$$

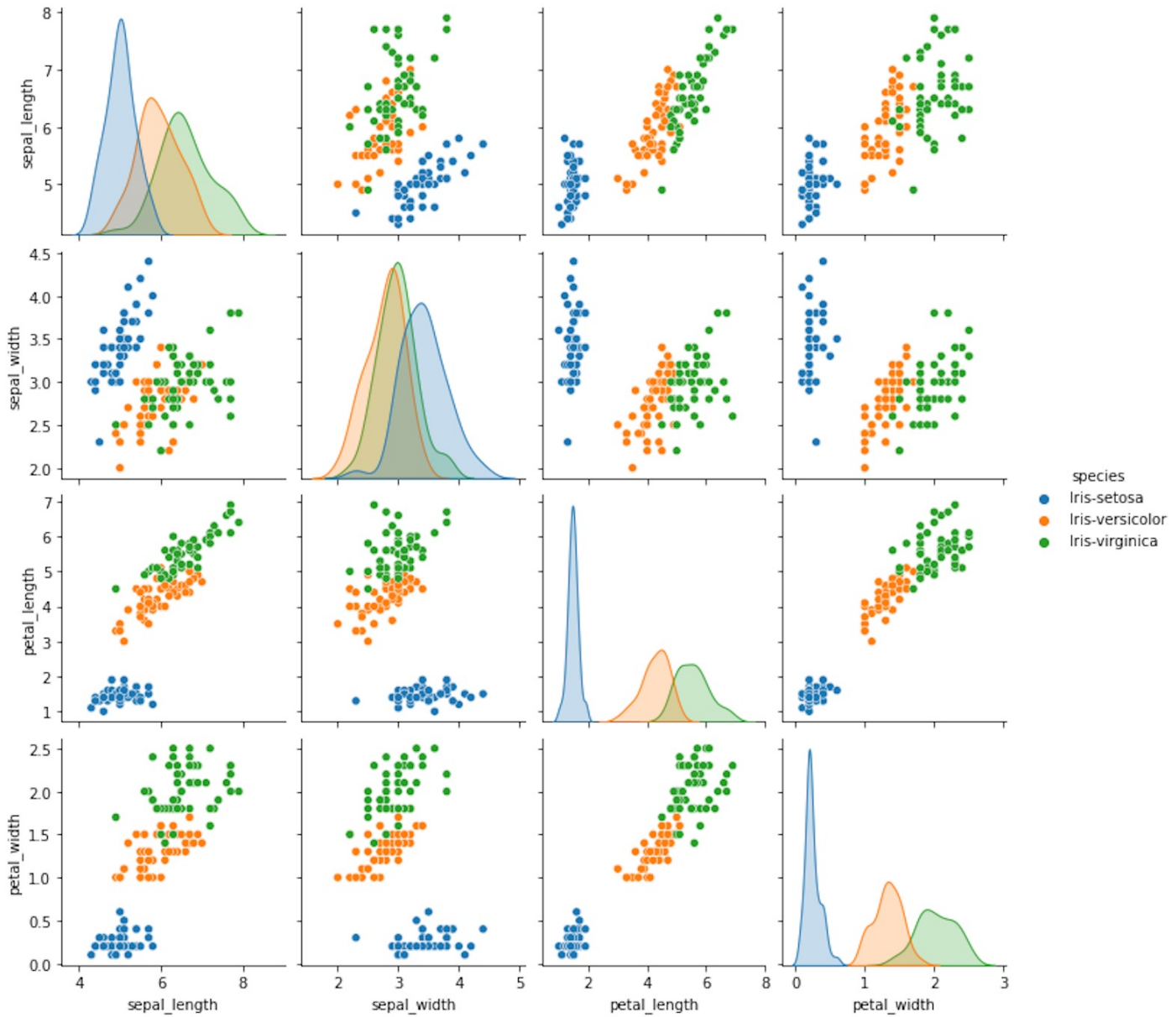
Euclidean Distance on Iris Dataset

$$\text{diff} = \sqrt{(\Delta Slen)^2 + (\Delta Swid)^2 + (\Delta Plen)^2 + (\Delta Pwid)^2}$$

Slen = Sepal Length
Swid = Sepal Width
Plen = Petal Length
Pwid = Petal Width



Why 1-NN makes mistakes on the iris dataset



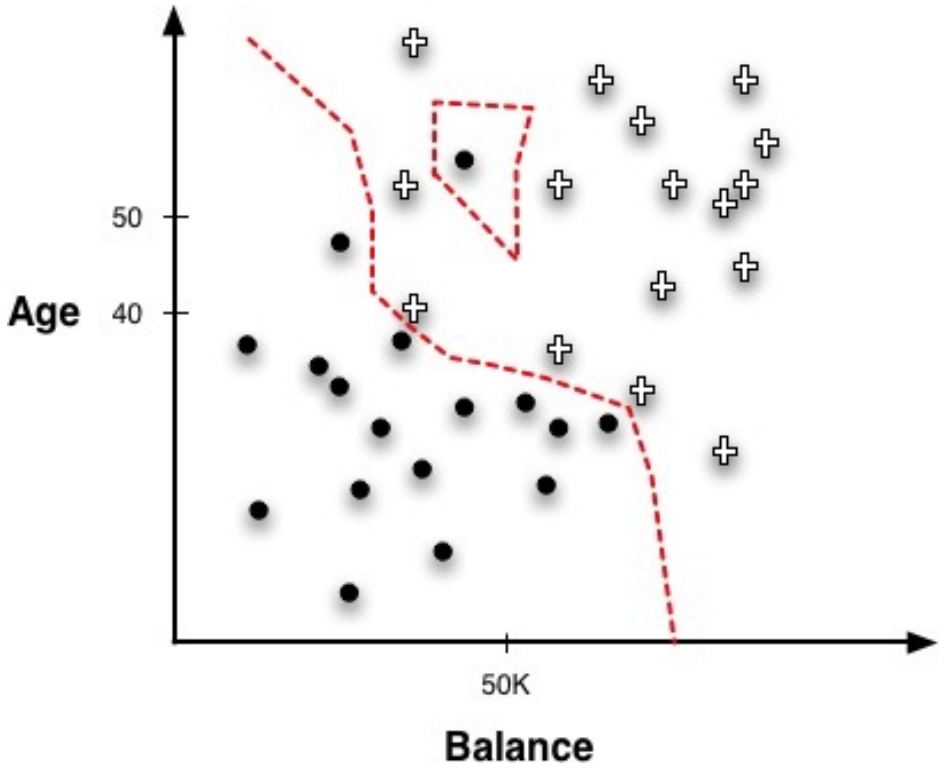
Iris Classification Report

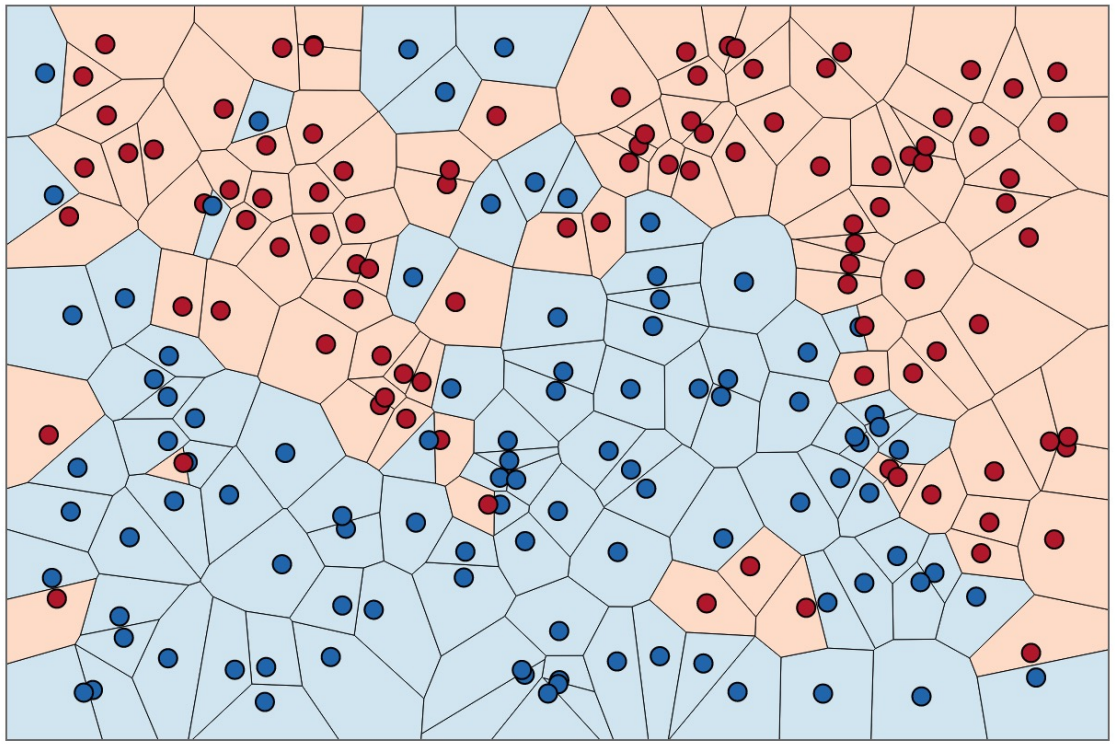
actual

predicted	**Setosa**	**Virginica**	**Versicolor**
Setosa	50	0	0
Virginica	0	47	3
Versicolor	0	3	47

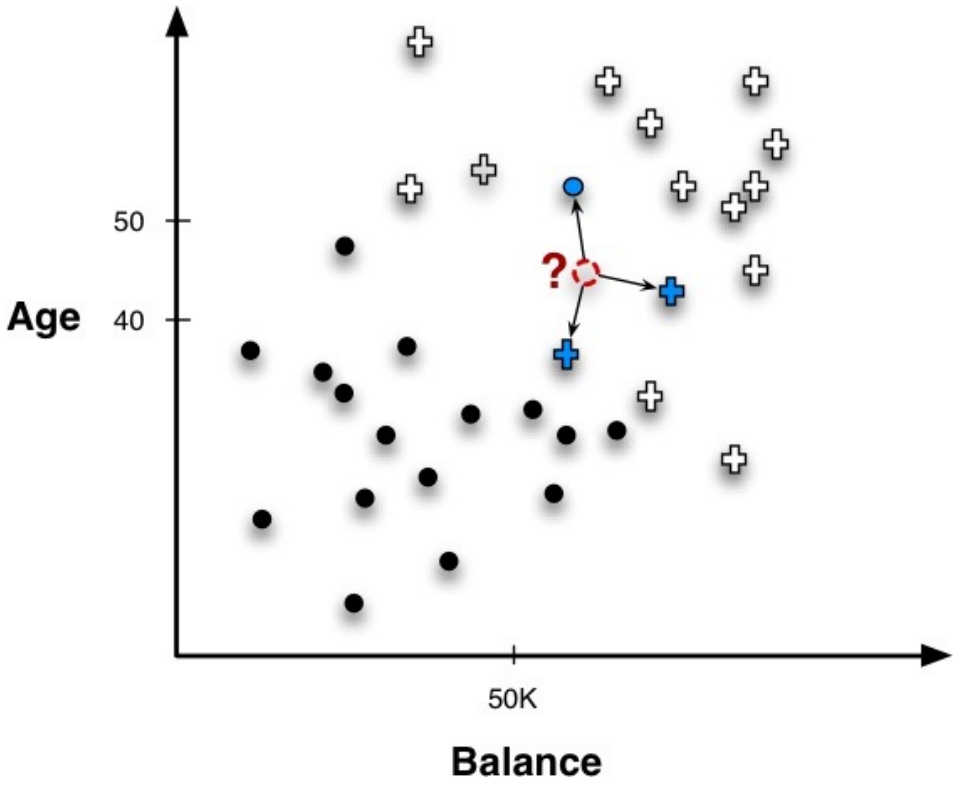
```
Classification Report:
              precision    recall  f1-score   support

      setosa       1.000     1.000     1.000        50
  versicolor       0.940     0.940     0.940        50
   virginica       0.940     0.940     0.940        50

    accuracy                          0.960       150
   macro avg       0.960     0.960     0.960       150
weighted avg       0.960     0.960     0.960       150
```

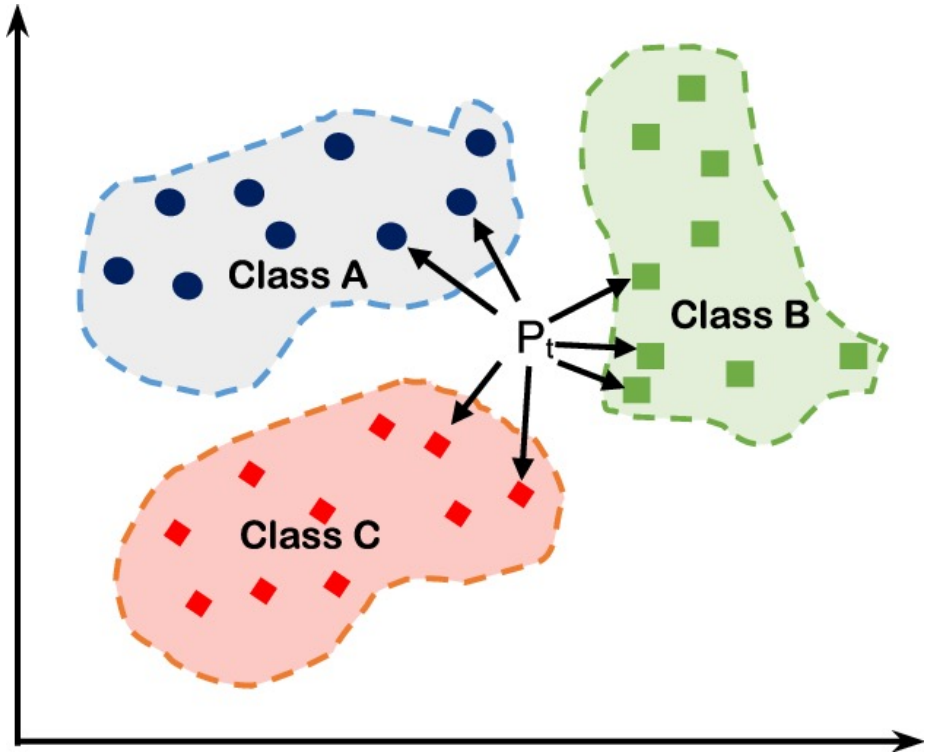
Overfitting





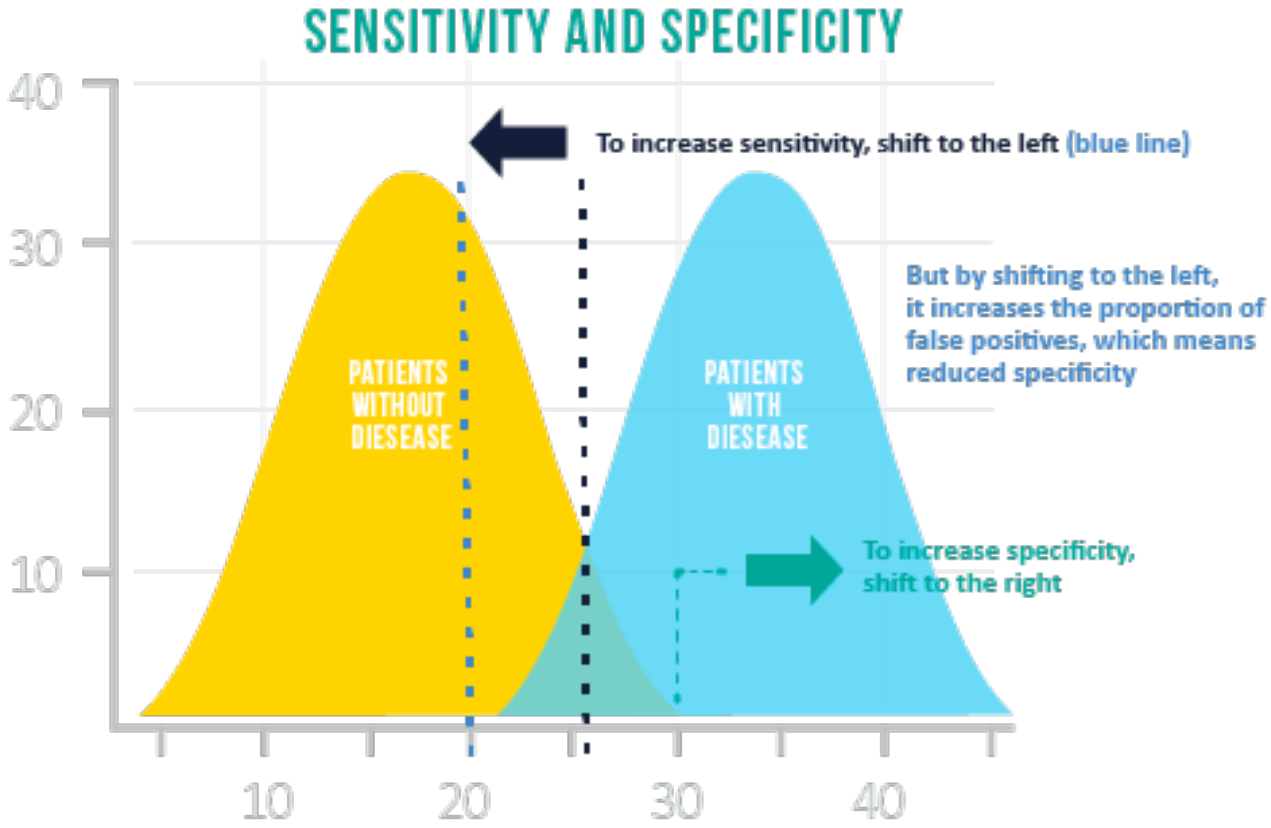
Multiple classes





Source: https://laptrinhx.com/k-nearest-neighbors-unlocked-454254569/

Eventually we'll have a readily available test …



But how will it impact public policy?

$$p(A|B) = \frac{p(B|A)\,p(A)}{p(B)}$$

Posterior probability | Likelihood | Prior probability

Measures of classification accuracy

10 photos, 7 dogs

Perfect Classification:

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	7	0
	Negative	0	3

Source: https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Imperfect Classification

		True/Actual	
		Positive ()	Negative
Predicted	Positive ()	5 (TP)	1 (FP) Type I Error
	Negative	2 (FN) Type II Error	2 (TN)

Recall: What proportion of ***actual positives*** are predicted positive?

Recall = TP / (TP + FN) = 5 / (5 + 2) = 71.4%

Recall = Sensitivity

Imperfect Classification

		True/Actual	
		Positive ()	Negative
Predicted	Positive ()	5 (TP)	1 (FP) Type I Error
	Negative	2 (FN) Type II Error	2 (TN)

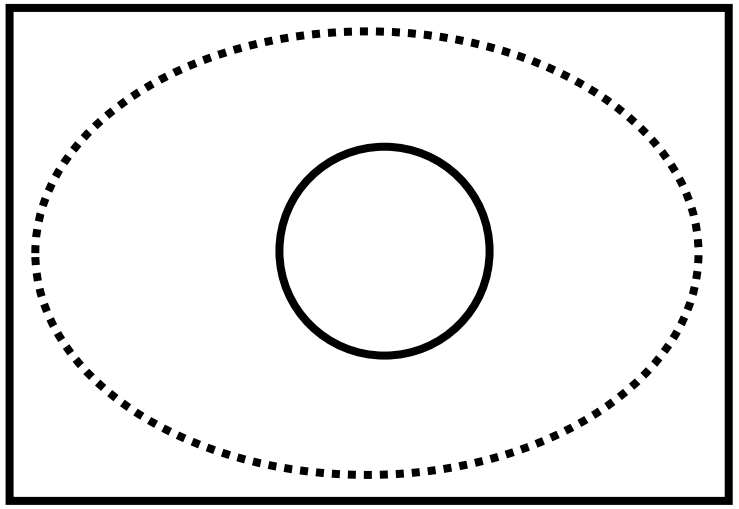
Precision: What proportion of ***predicted positives*** are truly positive?

Precision = TP / (TP + FP) = 5 / (5 + 1) = 83.3%

Specificity: What proportion of **actual negatives** are predicted negative?

Specificity = TN / (TN + FP) = 2 / (2 + 1) = 66.7%

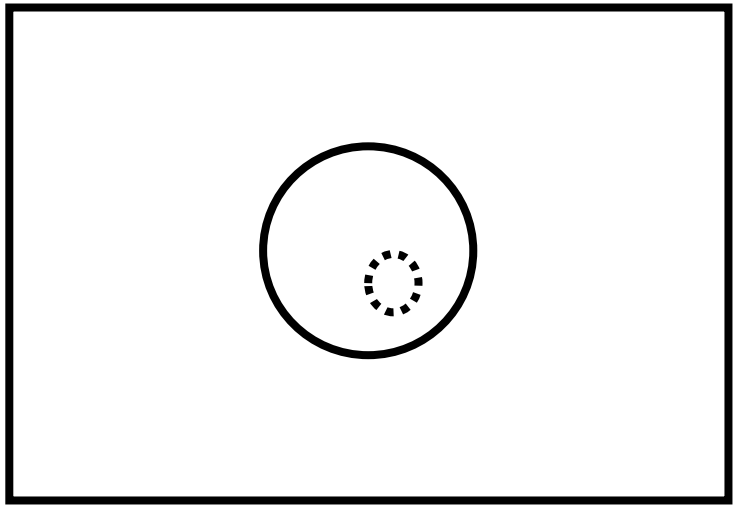
Recall/Precision and Sensitivity/Specificity



Perfect recall / sensitivity (TP/P):
Everyone truly sick is detected.

Poor precision (TP/(TP+FP)):
Few of the positives are really sick.

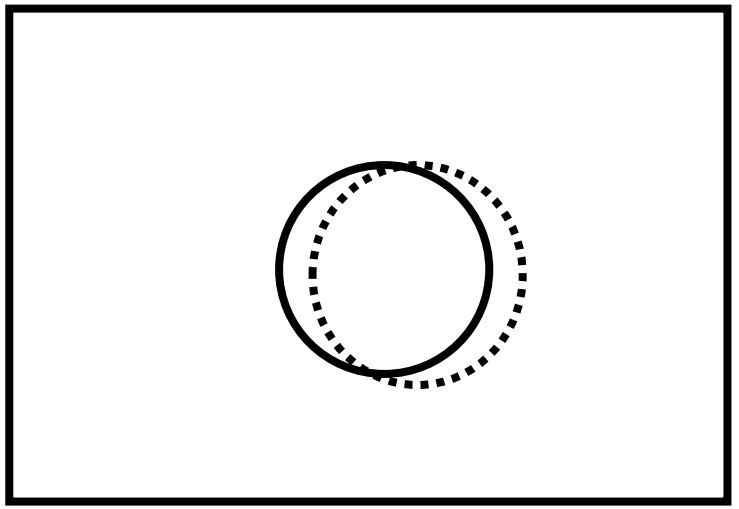
Poor specificity (TN/N):
Many healthy people are test positive.



Poor recall / sensitivity:
Many sick people aren't detected.

Perfect precision:
Every positive is truly sick.

Perfect specificity:
All healthy people test negative.



Good recall / sensitivity:
Most sick people are detected positive.

Good precision:
Most positives are *really* sick.

Good specificity:
Most healthy people test negative

General Distance Measure

$$\text{diff} = w_1|\Delta A_1|^r + w_2|\Delta A_2|^r + \ldots + w_n|\Delta A_n|^r$$

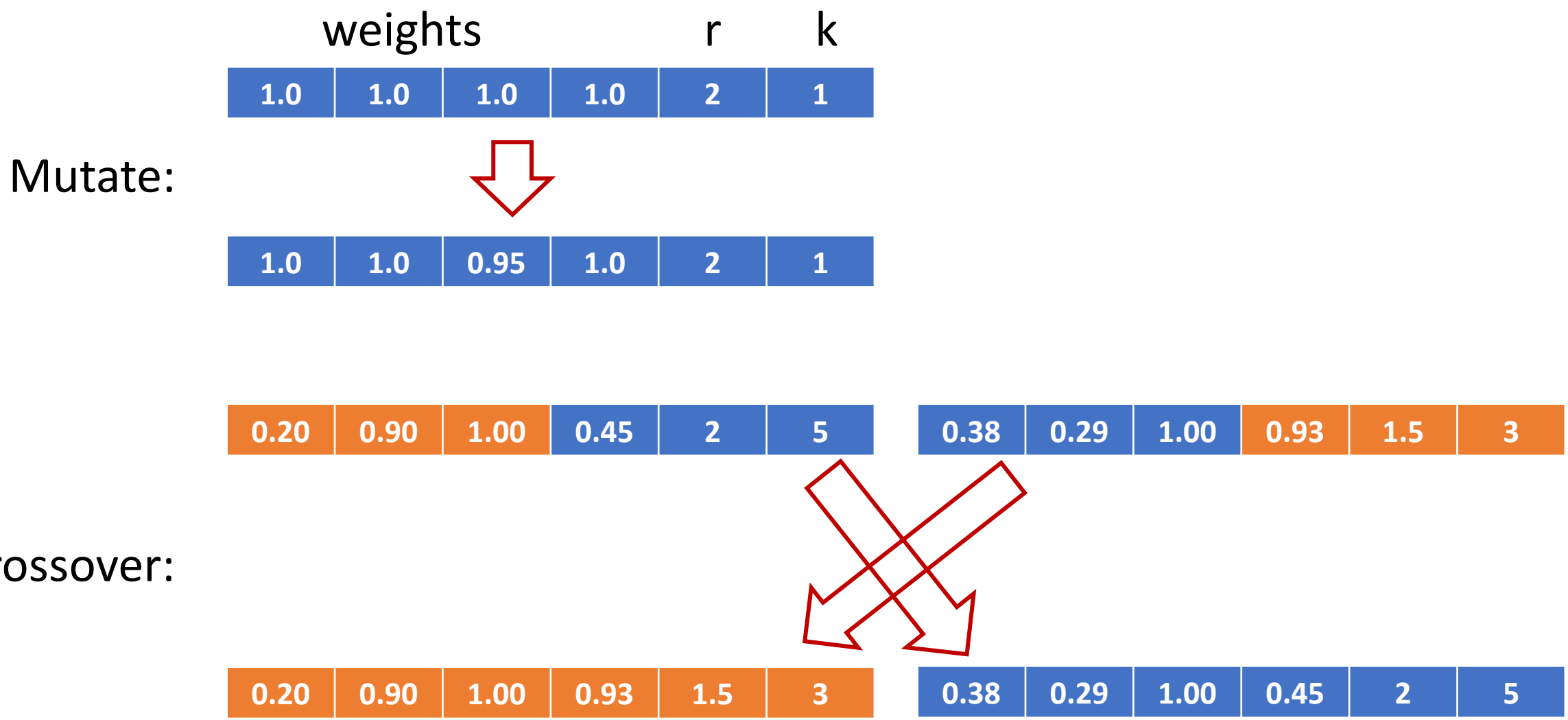
where

ΔA_i is the difference with respect to feature i,

w_k is a feature weighting factor (0.0 to 1.0), and

r is an exponent (> 0.0)

Tweaking / modifying a solution



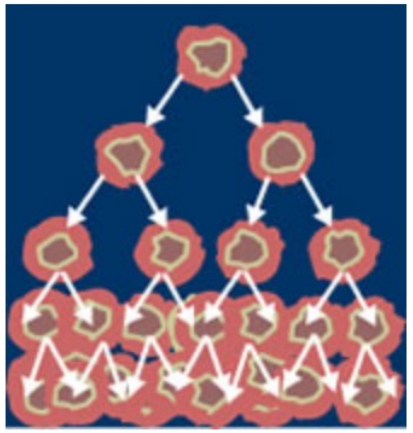
Predicting recurrence of breast cancer

UCI Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Breast Cancer Wisconsin (Prognostic) Data Set

Download: Data Folder, Data Set Description



Abstract: Prognostic Wisconsin Breast Cancer Database

Data Set Characteristics:	Multivariate	**Number of Instances:**	198	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	34	**Date Donated**	1995-12-01
Associated Tasks:	Classification, Regression	**Missing Values?**	Yes	**Number of Web Hits:**	224238

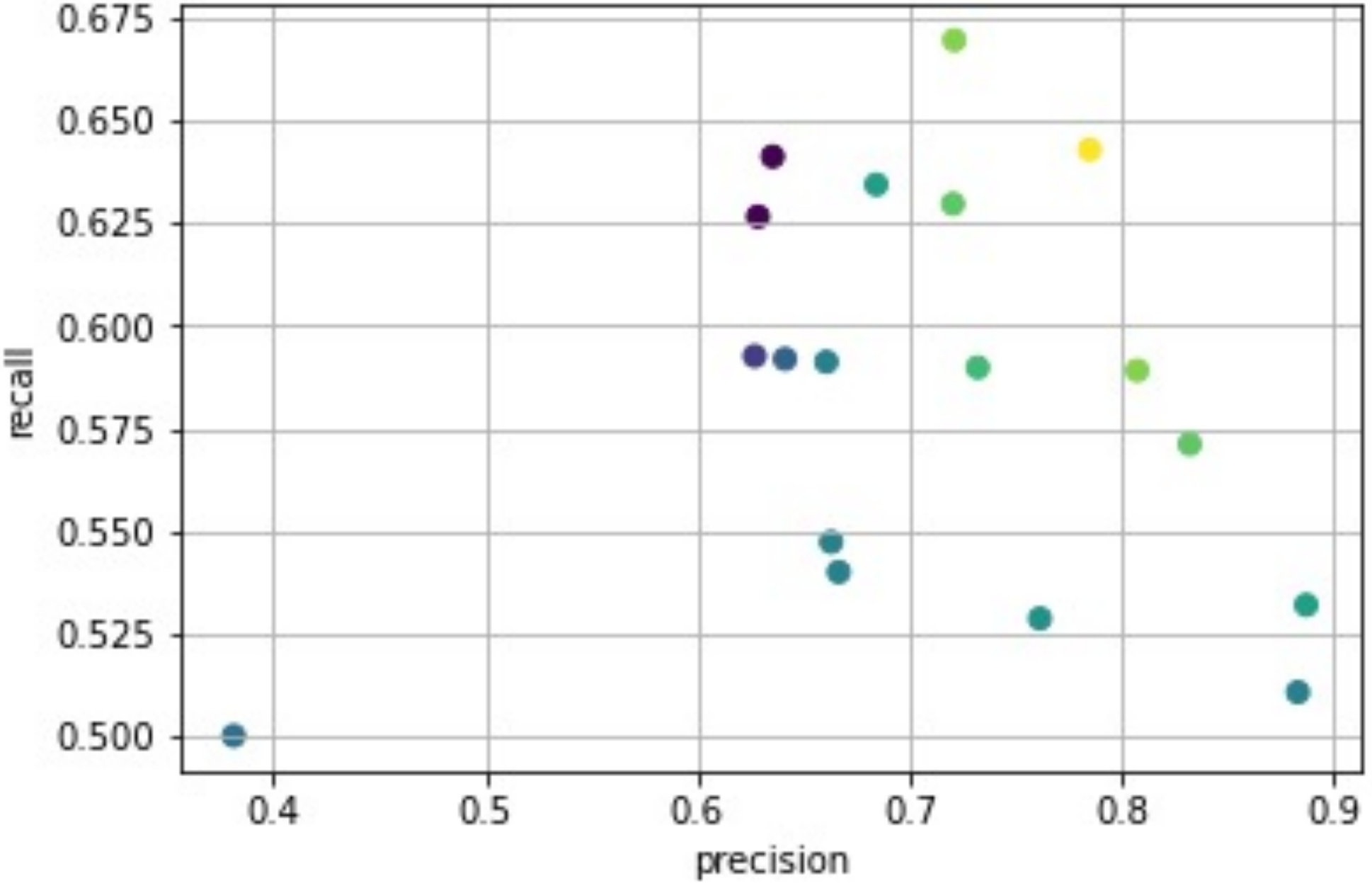
predicted

actual

	N	R
N	151	0
R	44	3

	precision	recall	f1-score	support
N	0.77436	1.00000	0.87283	151
R	1.00000	0.06383	0.12000	47
accuracy			0.77778	198
macro avg	0.88718	0.53191	0.49642	198
weighted avg	0.82792	0.77778	0.69413	198

discarded	accuracy	recall	precision	weights	R	K
32	0.76263	0.50000	0.38131	00000000000000000000000000000000	2.0085	3
31	0.74747	0.59264	0.62700	00000000000000000000000000000010	2.1085	4
31	0.73232	0.62667	0.62858	00000000000000000000000000000010	2.1085	1
30	0.73232	0.64133	0.63554	00000000100000000000000000000010	1.0264	2
31	0.75758	0.59194	0.64133	00000000000000000000001000000000	2.0085	5
31	0.76768	0.59124	0.66087	00000000000100000000000000000000	2.6851	5
31	0.76768	0.54727	0.66310	00000000000000100000000000000000	2.0085	8
31	0.76768	0.53995	0.66667	00000000000000100000000000000000	2.0085	10
30	0.77778	0.63449	0.68437	00000000100000000000000000000010	1.0537	3
30	0.79293	0.62977	0.72055	00000001000000000000000000000100	1.1662	5
29	0.79798	0.66972	0.72121	00000000100000000000001000010	1.7124	3
30	0.78788	0.58983	0.73224	00000000100000000000000000000010	1.4039	7
31	0.77273	0.52860	0.76160	00000000001000000000000000000000	2.0223	22
29	0.81313	0.64302	0.78511	00000101000000000000000000000100	1.0179	9
30	0.79798	0.58912	0.80749	00000001010000000000000000000000	2.9913	7
30	0.79293	0.57116	0.83224	00000000100000000000000000000010	2.7197	8
31	0.76768	0.51064	0.88325	00000000000000000010000000000000	2.2906	17
30	0.77778	0.53191	0.88718	00100000000000000000010000000000	2.1623	19


0.675
0.650
0.625
0.600
0.575
0.550
0.525
0.500
recall
0.4
0.5
0.6
0.7
0.8
0.9
precision

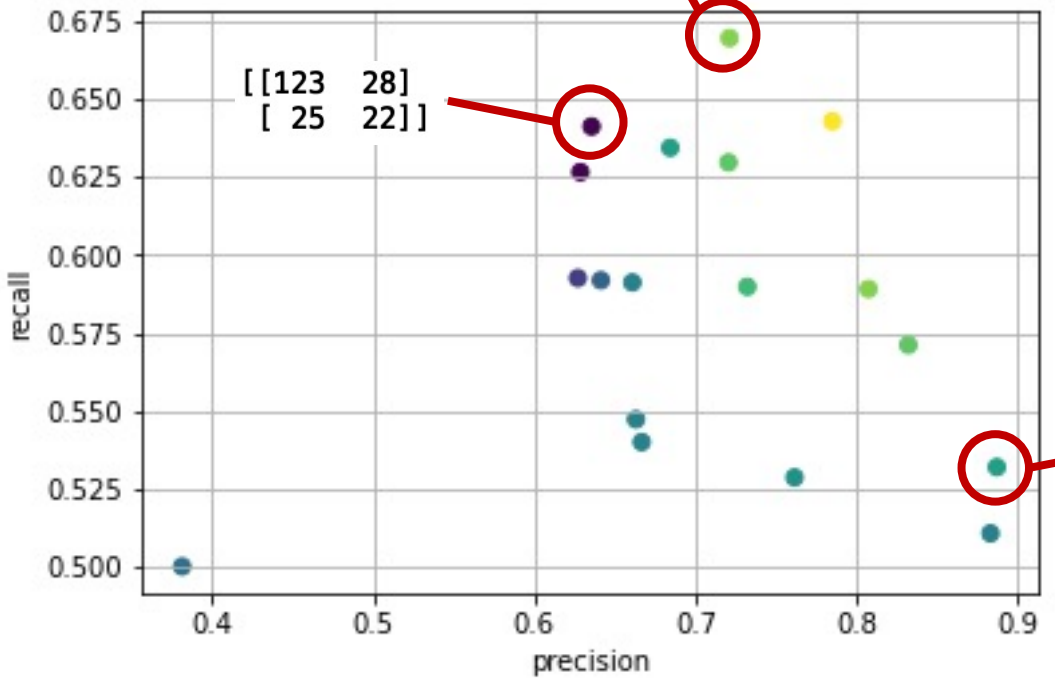
```
          predicted
            N     R
actual
N    [[138   13]
R     [ 27   20]]
```

R=1.7, K=3, Features = 3

	precision	recall	f1-score	support
N	0.83636	0.91391	0.87342	151
R	0.60606	0.42553	0.50000	47
accuracy			0.79798	198
macro avg	0.72121	0.66972	0.68671	198
weighted avg	0.78170	0.79798	0.78478	198

Recall: What fraction of the recurring cancers are we detecting (predicted to recur)?
(20 / 47 = 0.42553)

Precision: What fraction of those cancers predicted to recur actually recurred?
(20 / 33 = 0.60606)




```
          predicted
            N     R
actual
N    [[151    0]
R     [ 44    3]]
```

	precision	recall	f1-score	support
N	0.77436	1.00000	0.87283	151
R	1.00000	0.06383	0.12000	47
accuracy			0.77778	198
macro avg	0.88718	0.53191	0.49642	198
weighted avg	0.82792	0.77778	0.69413	198

R=2.3, K=17, Features = 1

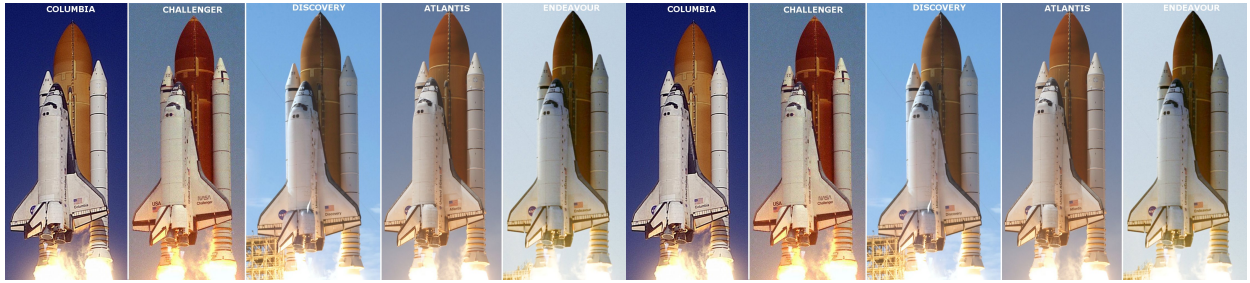
group·think

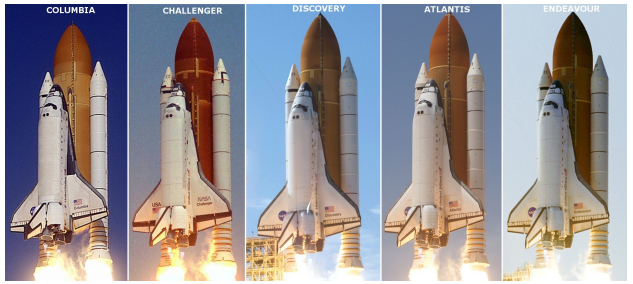
/ˈgro͞op ˌTHiNGk/

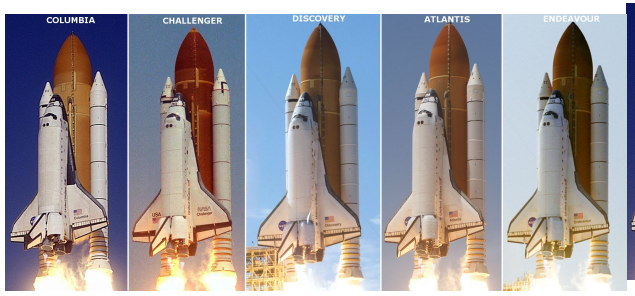
noun

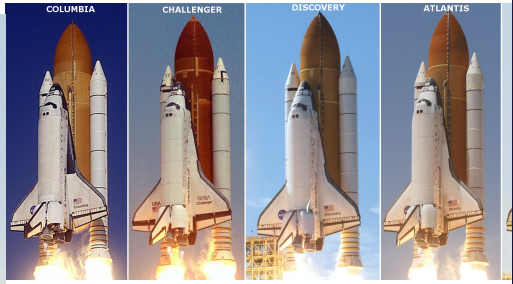
noun: **group-think**

the practice of thinking or making decisions as a group in a way that discourages creativity or individual responsibility.









Richard Feynman and the Rogers Commission

Richard Feynman
American Theoretical Physicist
1918 - 1988