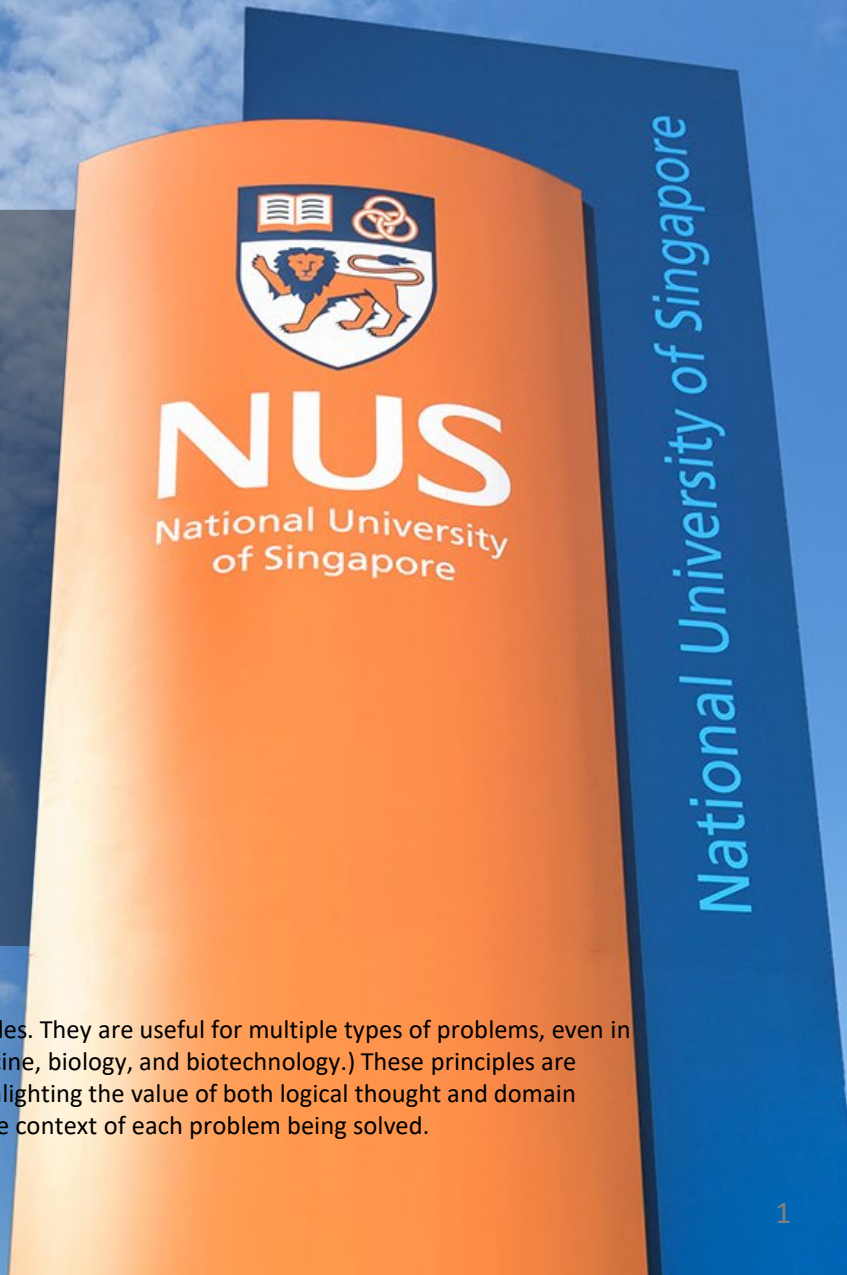
Golden thread of science

Wong Limsoon



Outline: I will present some of my favourite invariant-based problem-solving principles. They are useful for multiple types of problems, even in different disciplines (I will illustrate using different areas in computer science, medicine, biology, and biotechnology.) These principles are simple logical ways to exploit fundamental properties of each problem domain, highlighting the value of both logical thought and domain knowledge, and bringing out creative ways of applying the former to the latter in the context of each problem being solved.

© Copyright National University of Singapore. All Rights Reserved.

Golden thread of science

Science is characterized by

Observing an invariant

Proving that it is true, i.e., a law

Exploiting it to solve problems ***logically***

Three types of logical inferences

Deduction

- All men are mortal
- Socrates is a man

⇒Socrates is mortal

Abduction

- All men are mortal
- Socrates is mortal

⇒Socrates is a man,
provided there is no other explanation of Socrates' mortality

Induction

- Socrates is a man
- Socrates is mortal

⇒ All men are mortal,
provided there is no counter example

And two simple tactics

Fixing violation of invariants

Guilt by association

Invariant & science

Suppose you have a bag of x red beans and y green beans

Repeat the following:

Remove 2 beans

If both green, discard both

If both red, discard one, put back one

If one green and one red, discard red, put back green

If one bean is left behind, can you predict its colour?

Shall we bet on the color of the bean that is left behind?

You can always win

Suppose you have a bag of x red beans and y green beans

Repeat the following:

Remove 2 beans

If both green, discard both

If both red, discard one, put back one

If one green and one red, discard red, put back green

If one bean is left behind, can you predict its colour?

Shall we bet on the color of the bean that is left behind?

Intentionally left blank

What have we just seen?

Problem solving by (deductive) logical reasoning on invariants

Science is characterized by …

Observing an invariant:
Parity of green beans is invariant

Proving it:

Bet on the last red bean

• Suppose you have a bag of x red beans and y green beans	• When the parity of # of green beans (y) is even, …
• Repeat the following:	• Start with y=2n
– Remove 2 beans	
– If both green, discard both	• y=2n → y=2n-2
– If both red, discard one, put back one	• y=2n → y=2n
– If one green and one red, discard red, put back green	• y=2n → y=2n
• If one bean is left behind, can you predict its colour?	• y remains even ⇒ Last bean must be red!

Exploit it to solve problems:
Predict colour of the last bean

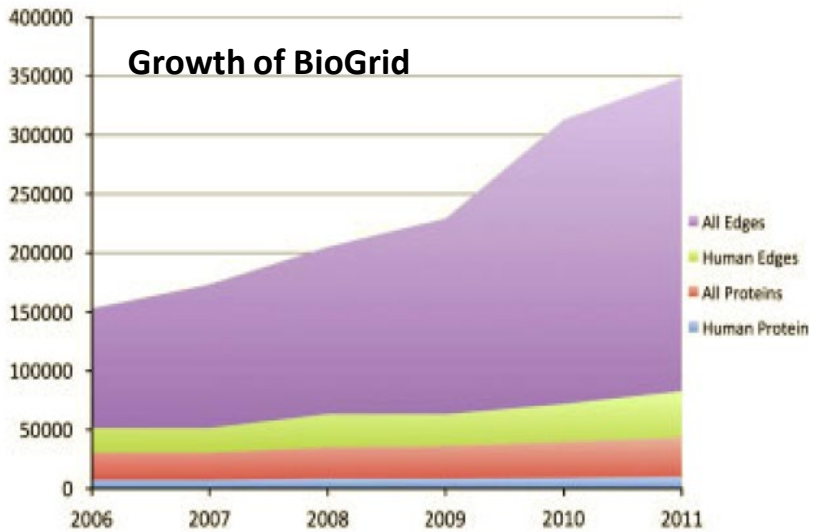
Removing noise from PPI experiments

Deduction

Protein-protein interaction detection

Many high-throughput assays for PPIs

Generating *large amounts* of expt data on PPIs can be done with ease



But …

High-throughput approaches sacrifice quality for ***quantity***:
(a) limited or biased coverage: ***false negatives***, &
(b) high error rates: ***false positives***

Noise in PPI networks

High level of noise $\Rightarrow$ Need to clean up

Experimental method category[a]	Number of interacting pairs	Co-localization[b] (%)	Co-cellular-role[b] (%)
All: All methods	9347	64	49
A: Small scale Y2H	1861	73	62
A0: GY2H Uetz *et al.* (published results)	956	66	45
A1: GY2H Uetz *et al.* (unpublished results)	516	53	33
A2: GY2H Ito *et al.* (core)	798	64	40
A3: GY2H Ito *et al.* (all)	3655	41	15
B: Physical methods	71	98	95
C: Genetic methods	1052	77	75
D1: Biochemical, *in vitro*	614	87	79
D2: Biochemical, chromatography	648	93	88
E1: Immunological, direct	1025	90	90
E2: Immunological, indirect	34	100	93
2M: Two different methods	2360	87	85
3M: Three different methods	1212	92	94
4M: Four different methods	570	95	93

Sprinzak et al., *JMB*, 327:919-923, 2003

Large disagreement betw methods

Exercise #1

Can you think of things a biologist can do to remove PPIs that are likely to be noise?

De-noising PPI networks using reproducibility

Intentionally left blank

De-noising PPI networks using localization coherence

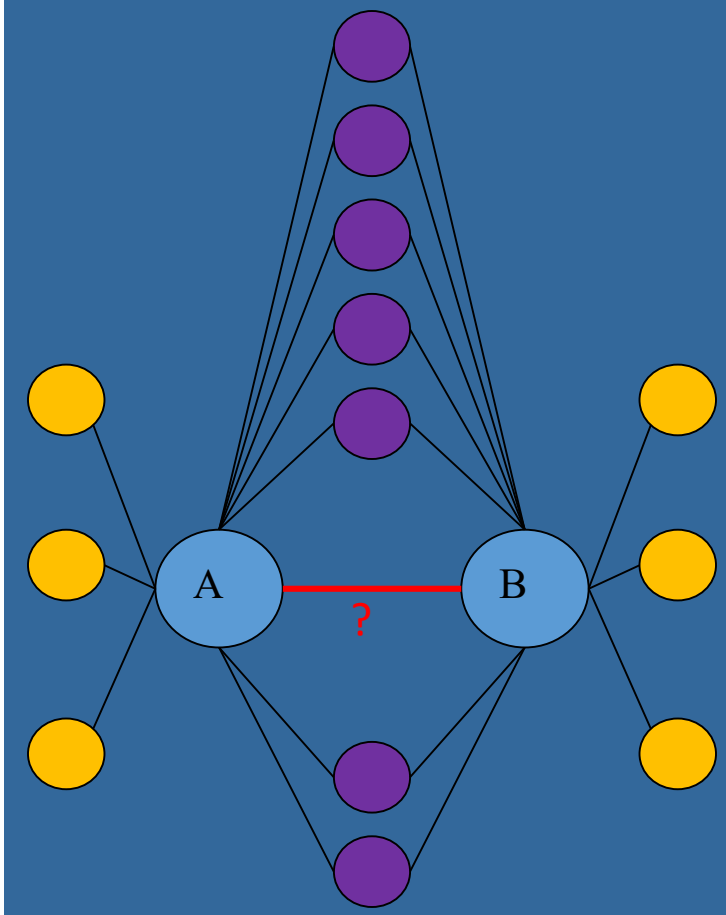
Intentionally left blank

Exercise #2

Do you really need to know where two proteins are, in order to know whether they are in the same place?

If not, how?

Topology of neighbourhood of real PPIs



Intentionally left blank

Czekanowski-Dice distance

Intentionally left blank

Adjusted CD-distance

Intentionally left blank

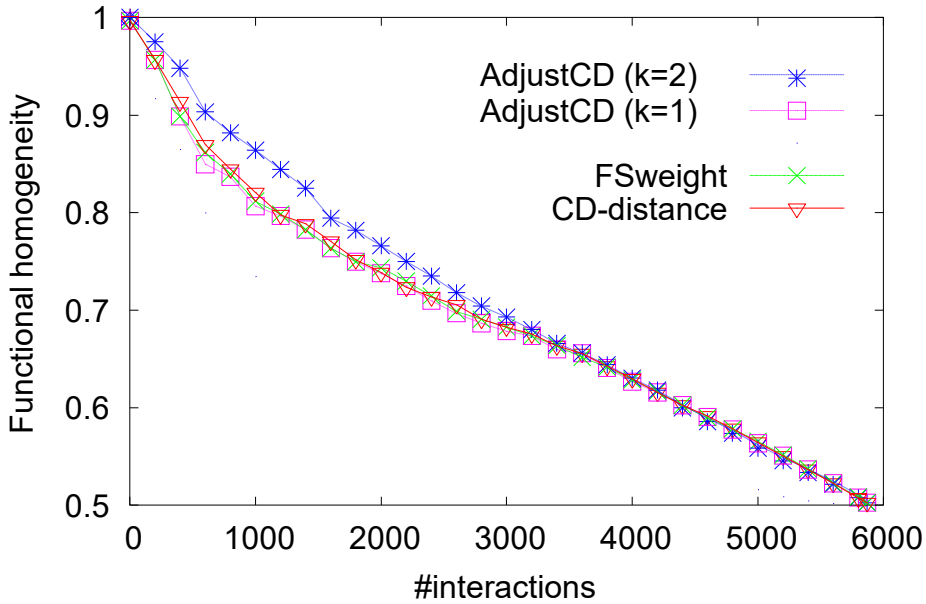
Iterated CD-distance

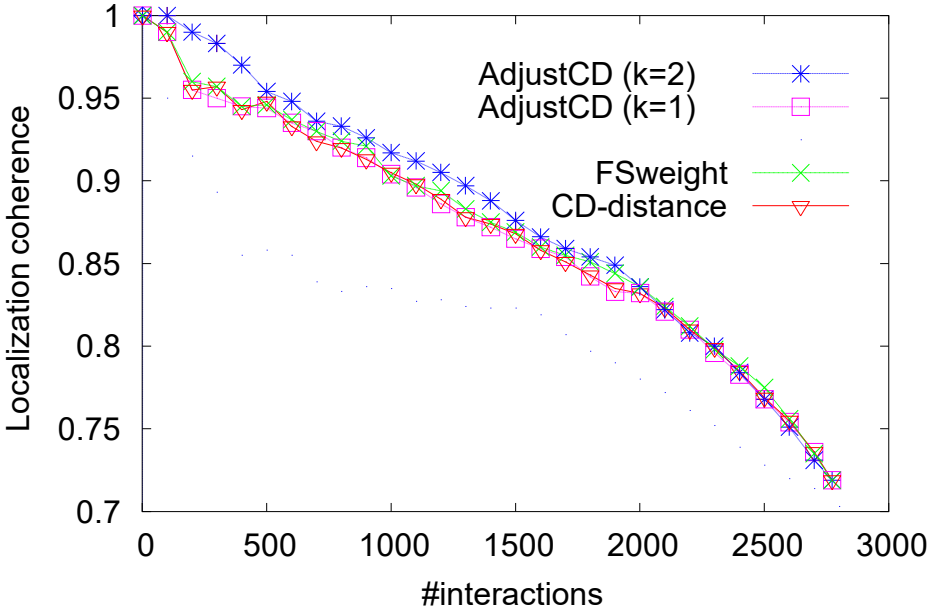
Intentionally left blank

Liu et al. Complex discovery from weighted PPI networks.
Bioinformatics, 25(15):1891-1897, 2009

Identifying false-positive PPIs

Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%

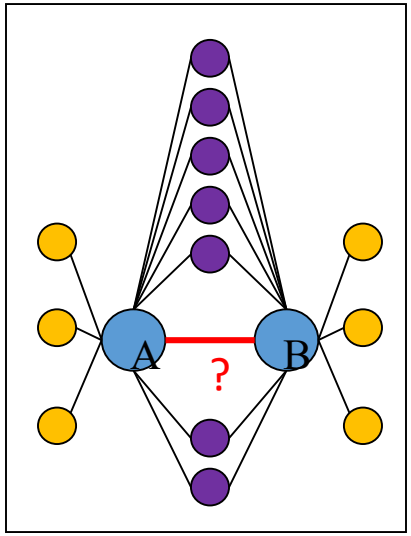




CD-distance and its variations correlate well with functional homogeneity and localization coherence

The triumph of logic

Two proteins should be in same place to interact



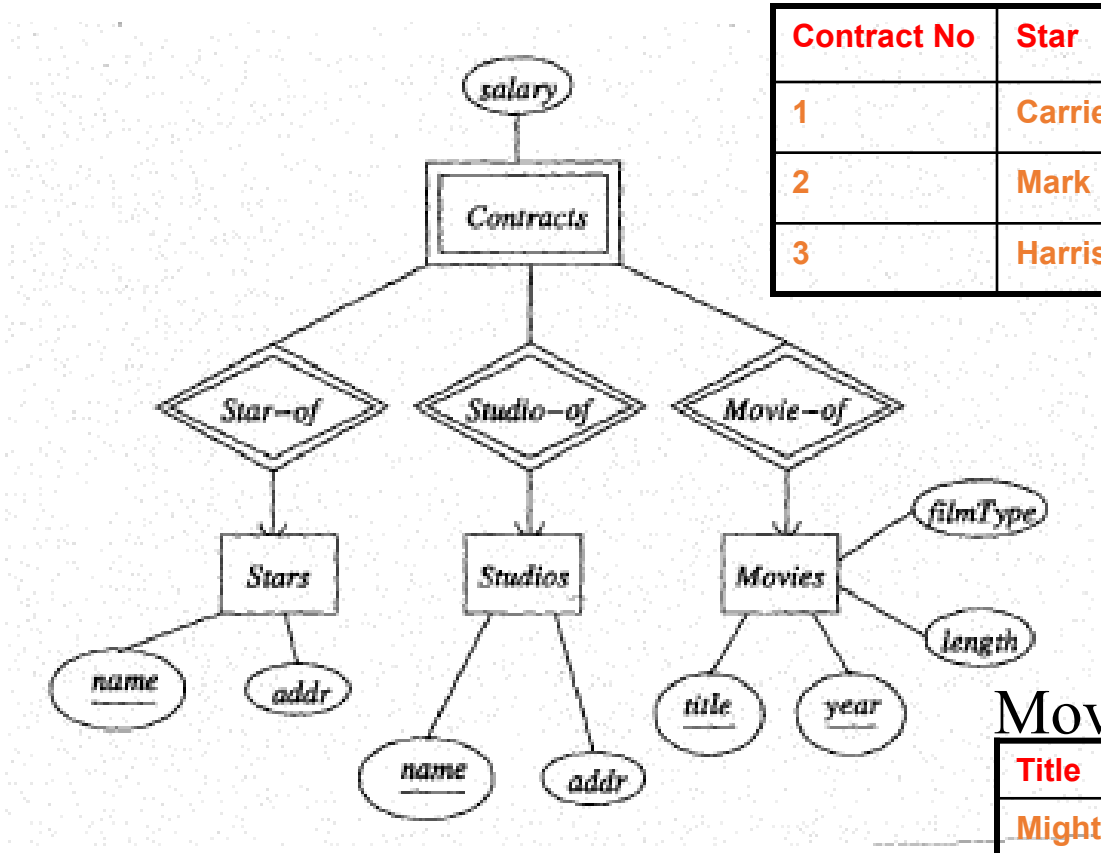
$$CD(u,v) = \frac{2\,|N_u \cap N_v|}{|N_u| + |N_v|}$$

Impact:
PPI networks can be cleansed based on topological info, w/o needing location etc info on proteins

Improving database Design

Violation of invariant

Relational data model



Contracts

Contract No	Star	Studio	Title	Salary
1	Carrie Fisher	Fox	Star Wars	$$$
2	Mark Hamill	Fox	Star Wars	$$$
3	Harrison Ford	Fox	Star Wars	$$$

Stars

Name	Address
Carrie Fisher	Hollywood
Mark Hamill	Brentwood
Harrison Ford	Beverly Hills

Movies

Title	Year	Length	Film Type
Mighty Ducks	1991	104	Color
Wayne's World	1992	95	Color
Star Wars	1977	124	Color

Design issues

How many possible alternate ways to represent movies using tables?

Why this particular set of tables to represent movies?

Indeed, why not use this alternative single table below to represent movies?

Wrong Movies

Title	Year	Length	Film Type	Studio	Star
Star Wars	1997	124	Color	Fox	Carrie Fisher
Star Wars	1997	124	Color	Fox	Mark Hamill
Star Wars	1997	124	Color	Fox	Harrison Ford
Mighty Ducks	1991	104	Color	Disney	Emilio Estevez

Anomalies

Wrong Movies

Title	Year	Length	Film Type	Studio	Star
Star Wars	1997	124	Color	Fox	Carrie Fisher
Star Wars	1997	124	Color	Fox	Mark Hamill
Star Wars	1997	124	Color	Fox	Harrison Ford
Mighty Ducks	1991	104	Color	Disney	Emilio Estevez

Redundancy: Unnecessary repetition of info

Update anomalies: If Star Wars is 125 min, we might carelessly update row 1 but not rows 2 & 3

Deletion anomalies: If Emilio Estevez is deleted from stars of Mighty Ducks, we lose all info on that movie

Some interesting questions

How to differentiate a good database design from a bad one?

How to produce a good database design automatically from a bad one?

Functional dependency

Functional dependency ($A_1, \ldots, A_n \rightarrow B_1, \ldots, B_m$)

If two rows of a table R agree on attributes $A_1, \ldots, A_n$, then they must also agree on attributes $B_1, \ldots, B_m$

FD ($A_1, \ldots, A_n \rightarrow B_1, \ldots, B_m$) is trivial if a B_i is an A_j

Wrong Movies

Title	Year	Length	Film Type	Studio	Star
Star Wars	1997	124	Color	Fox	Carrie Fisher
Star Wars	1997	124	Color	Fox	Mark Hamill
Star Wars	1997	124	Color	Fox	Harrison Ford
Mighty Ducks	1991	104	Color	Disney	Emilio Estevez

Example: Title, Year → Length, Film Type, Studio

Keys

Key is a minimal set of attributes $\{A_1, \ldots, A_n\}$ that functionally determine all other attributes of a table

Superkey is a set of attributes that contains a key

Wrong Movies

Title	Year	Length	Film Type	Studio	Star
Star Wars	1997	124	Color	Fox	Carrie Fisher
Star Wars	1997	124	Color	Fox	Mark Hamill
Star Wars	1997	124	Color	Fox	Harrison Ford
Mighty Ducks	1991	104	Color	Disney	Emilio Estevez

Example superkey: Any set of attributes that contains {Title, Year, Star} as a subset

Boyce-Codd Normal Form

A relation R is in Boyce-Codd Normal Form iff whenever there is a nontrivial FD ($A_1, \ldots, A_n \rightarrow B_1, \ldots, B_m$) for R, it is the case that $\{A_1, \ldots, A_n\}$ is a superkey for R

Theorem (Codd, 1972)

A database design has no anomalies due to FD iff all its relations are in Boyce-Codd Normal Form

How is BCNF violated here?

Wrong Movies

Title	Year	Length	Film Type	Studio	Star
Star Wars	1997	124	Color	Fox	Carrie Fisher
Star Wars	1997	124	Color	Fox	Mark Hamill
Star Wars	1997	124	Color	Fox	Harrison Ford
Mighty Ducks	1991	104	Color	Disney	Emilio Estevez

A nontrivial FD:

Title, Year → Length, Film Type, Studio

The LHS not superset of the key {Title,Year, Star}

Violate BCNF!

Anomalies due to FD's whose LHS is not superkey

Towards a better design

Use an *offending* FD ($A_1, \ldots, A_n \rightarrow B_1, \ldots, B_m$) to decompose $R(A_1, \ldots, A_n, B_1, \ldots, B_m, C_1, \ldots, C_h)$ into 2 tables

- $R_1(A_1, \ldots, A_n, B_1, \ldots, B_m)$
- $R_2(A_1, \ldots, A_n, C_1, \ldots, C_h)$

No update anomaly

No redundant info

Title	Year	Length	Film Type	Studio
Star Wars	1997	124	Color	Fox
Mighty Ducks	1991	104	Color	Disney

No deletion anomaly

Wrong Movies

Title	Year	Length	Film Type	Studio	Star
Star Wars	1997	124	Color	Fox	Carrie Fisher
Star Wars	1997	124	Color	Fox	Mark Hamill
Star Wars	1997	124	Color	Fox	Harrison Ford
Mighty Ducks	1991	104	Color	Disney	Emilio Estevez

Title	Year	Star
Star Wars	1997	Carrie Fisher
Star Wars	1997	Mark Hamill
Star Wars	1997	Harrison Ford
Mighty Ducks	1991	Emilio Estevez

The “Invariant” Perspective

The invariant:

BCNF is an invariant of a good database design

The lesson learned:

Deliver a better database design by fixing violated invariants

Inferring key mutations: Why some PTP is inefficient

Induction / fixing violated invariants

Protein tyrosine phosphatase

Sequence from a typical PTP

```
>gi|00000|PTPA-D2
EEEFKKLTSIKIQNDKMRTGNLPANMKKNRVLQIIPYEFNRVIIPVKRGEENTDYVNASF
IDGYRQKDSYIASQGPLLHTIEDFWRMIWEWKSCSIVMLTELEERGQEKCAQYWPSDGLV
SYGDITVELKKEEECESYTVRDLLVTNTRENKSRQIRQFHFHGWPEVGIPSDGKGMISII
AAVQKQQQQSGNHPITVHCSAGAGRTGTFCALSTVLERVKAEGILDVFQTVKSLRLQRPH
MVQTLEQYEFCYKVVQEYIDAFSDYANFK
```

Some PTPs are much less efficient than others

Why? And how do you figure out which mutations cause the loss of efficiency?

Exercise #5

How do you figure out which mutations cause the loss of efficiency?

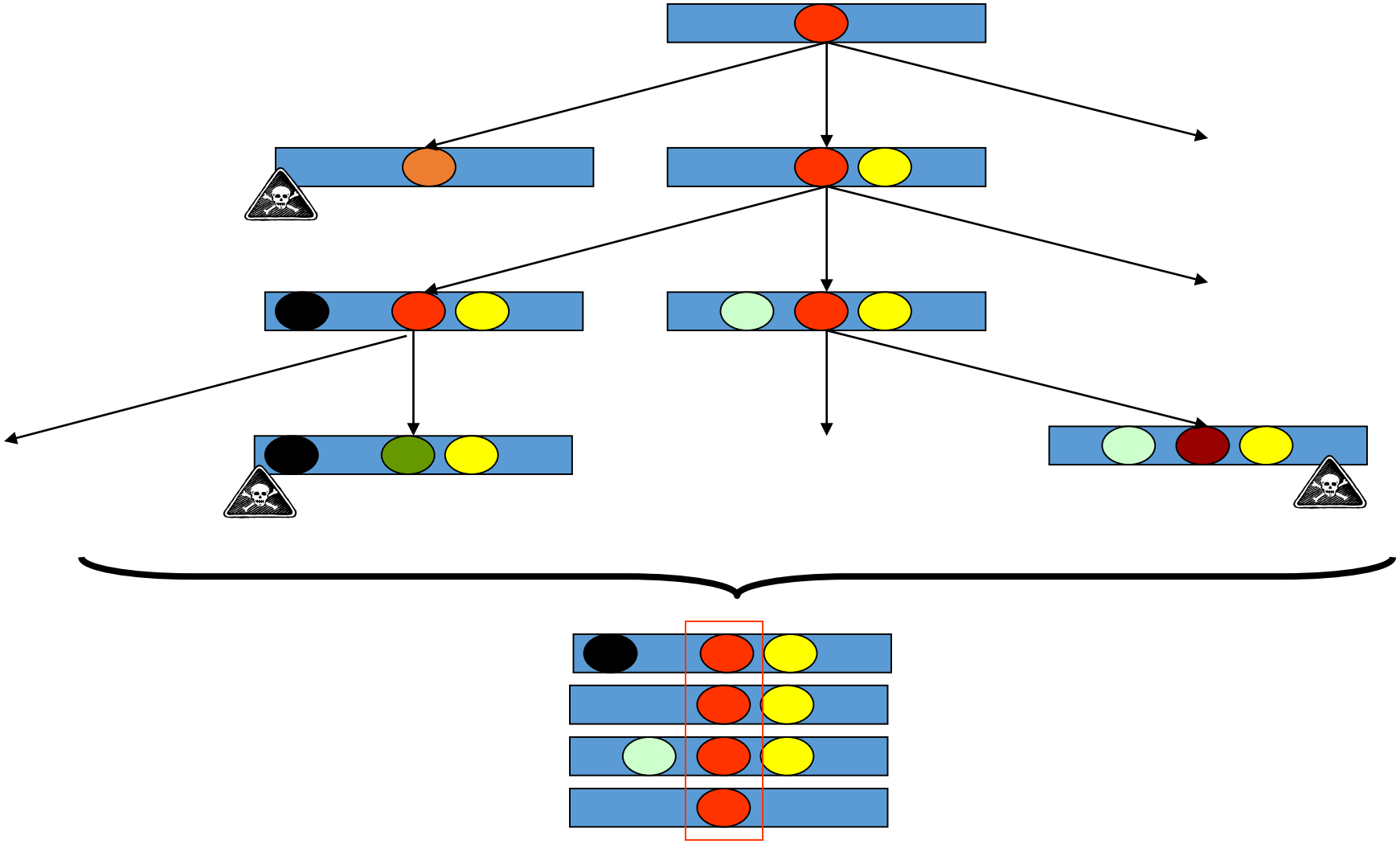
Protein tyrosine phosphatase

Sequence from a typical PTP

```
>gi|00000|PTPA-D2
EEEFKKLTSIKIQNDKMRTGNLPANMKKNRVLQIIPYEFNRVIIPVKRGEENTDYVNASF
IDGYRQKDSYIASQGPLLHTIEDFWRMIWEWKSCSIVMLTELEERGQEKCAQYWPSDGLV
SYGDITVELKKEEECESYTVRDLLVTNTRENKSRQIRQFHFHGWPEVGIPSDGKGMISII
AAVQKQQQQSGNHPITVHCSAGAGRTGTFCALSTVLERVKAEGILDVFQTVKSLRLQRPH
MVQTLEQYEFCYKVVQEYIDAFSDYANFK
```

- **Some PTPs are much less efficient than others**

Some sites are impt for PTP function



Reasoning based on an invariant…

Intentionally left blank

Key mutations: PTP D1 vs D2

```
                        ?   !  ?         ?                  ?             ? ??
gi|00000|P  D2  QFHFHGWPEVGIPSDGKGMISIIAAVQKQQQQ-SGNHPITVHCSAGAGRTGTFCALSTVL
gi|126467|      QFHFTSWPDFGVPFTPIGMLKFLKKVKACNP--QYAGAIVVHCSAGVGRTGTFVVIDAML
gi|2499753      QFHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIML
gi|462550|      QYHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGRTGTYIVIDSML
gi|2499751      QFHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLI
gi|1709906  D1  QFQFTAWPDHGVPEHPTPFLAFLRRVKTCNP--PDAGPMVVHCSAGVGRTGCFIVIDAML
gi|126471|      QLHFTSWPDFGVPFTPIGMLKFLKKVKTLNP--VHAGPIVVHCSAGVGRTGTFIVIDAMM
gi|548626|      QFHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIML
gi|131570|      QFHFTGWPDHGVPYHATGLLGFVRQVKSKSP--PNAGPLVVHCSAGAGRTGCFIVIDIML
gi|2144715      QFHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLI
                * ..  **. *.*      .                 . ****** ****..  .  ..
```

Positions marked by “!” and “?” are likely places responsible for reduced PTP activity

All PTP D1 agree on them

All PTP D2 disagree on them

Lim et al. Journal of Biological Chemistry 273:28986-28993,1998.

Confirm by mutagenesis expt

Wet experiments confirm the predictions

Mutate D → E in D1

i.e., check if D → E can cause efficiency loss

Mutate E → D in D2

i.e., show D → E is cause of efficiency loss

Impact:
Hundreds of mutagenesis expts saved by simple reasoning on (violation of) invariants!

The triumph of logic

Induction/hypothesis: A site that is critical for PTP efficiency is present in all efficient PTPs and absent in all inefficient PTPs

Observation: A site X is present in all efficient PTPs and absent in all inefficient PTPs

Abduction: Site X is critical for PTP efficiency

Bioengineering more efficient PTPs

Replace an inefficient PTP in the organism by an efficient version

Mutate E → D in D2

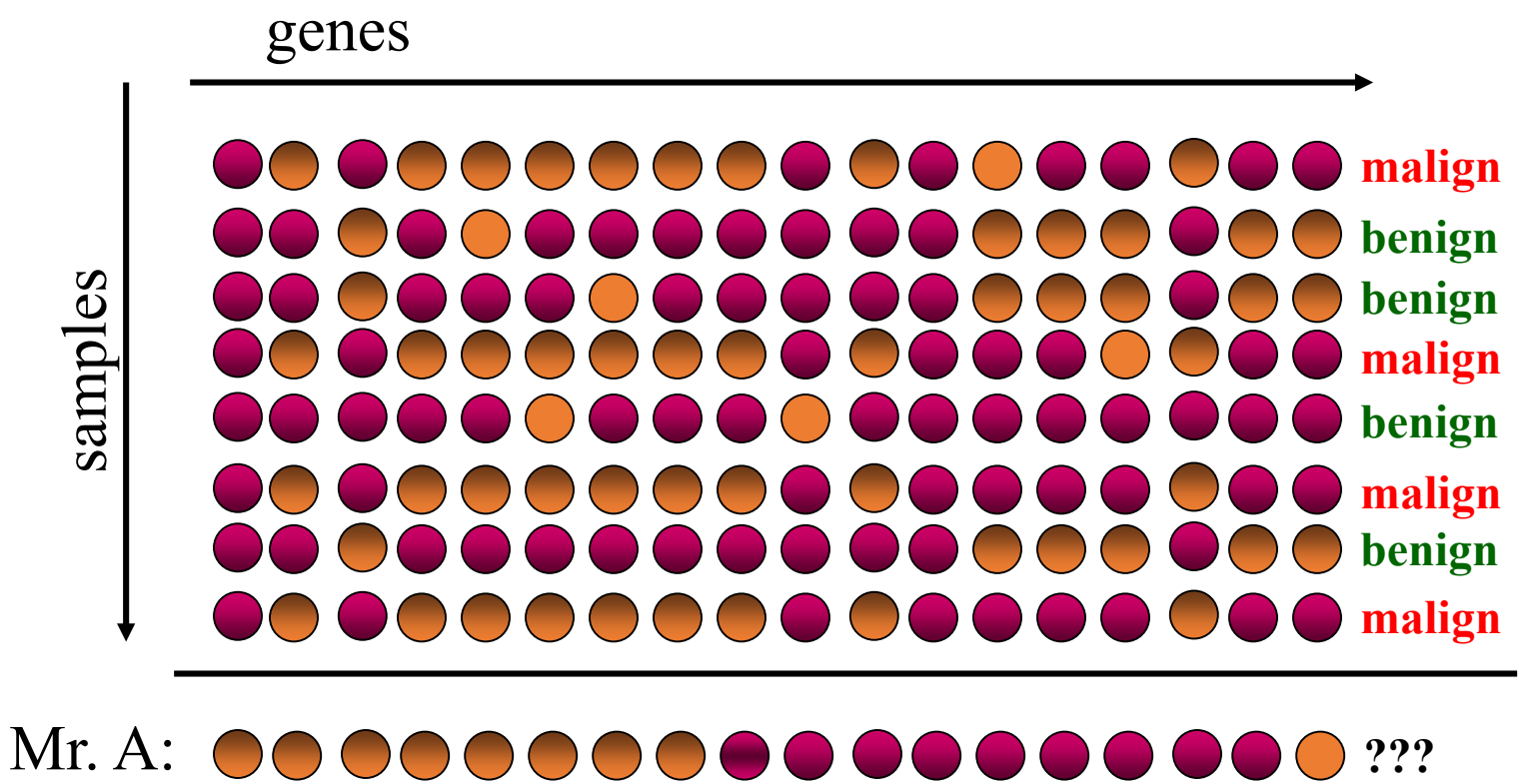
What have we just seen?

Create a more efficient PTP by fixing a violated invariant!

Diagnosing pediatric leukemias

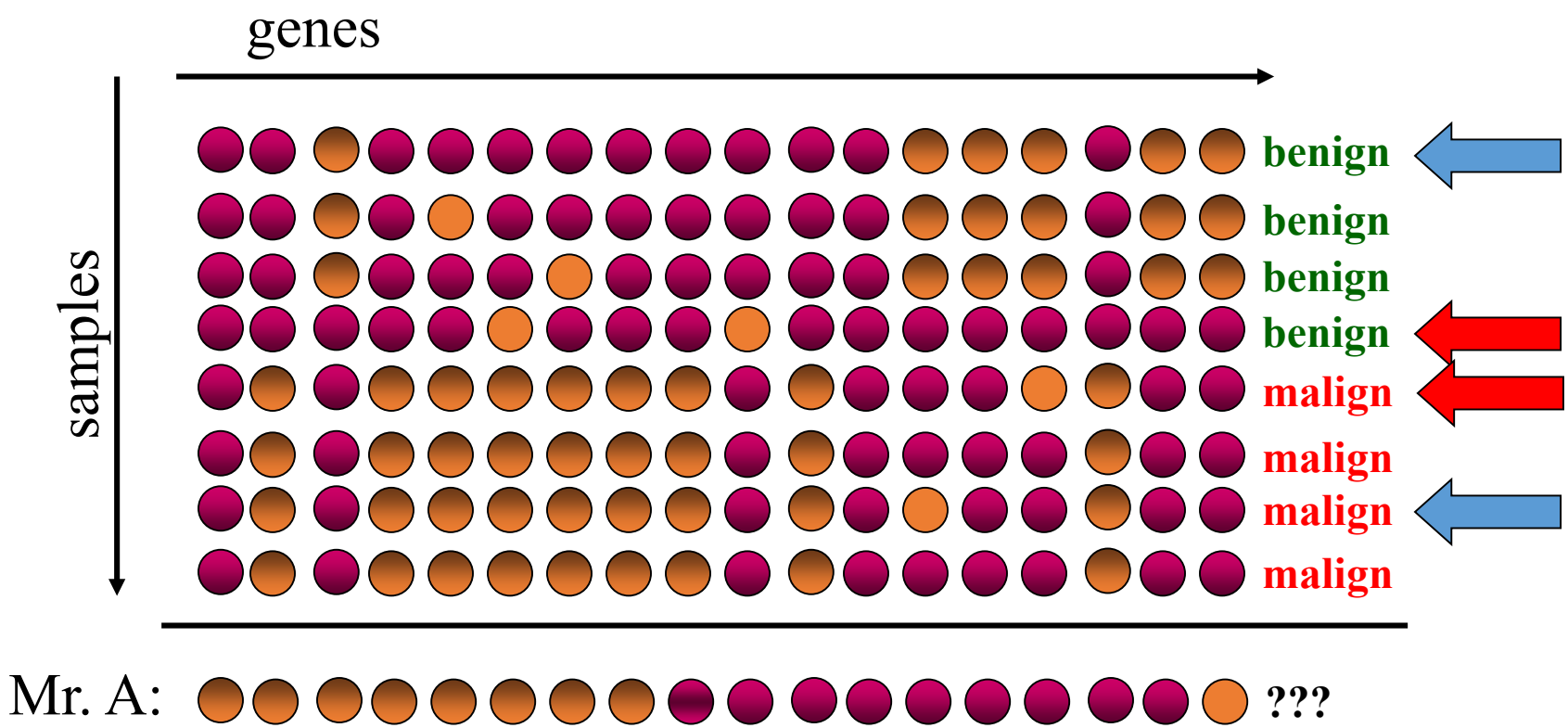
Induction & Abduction

Some Patient Samples



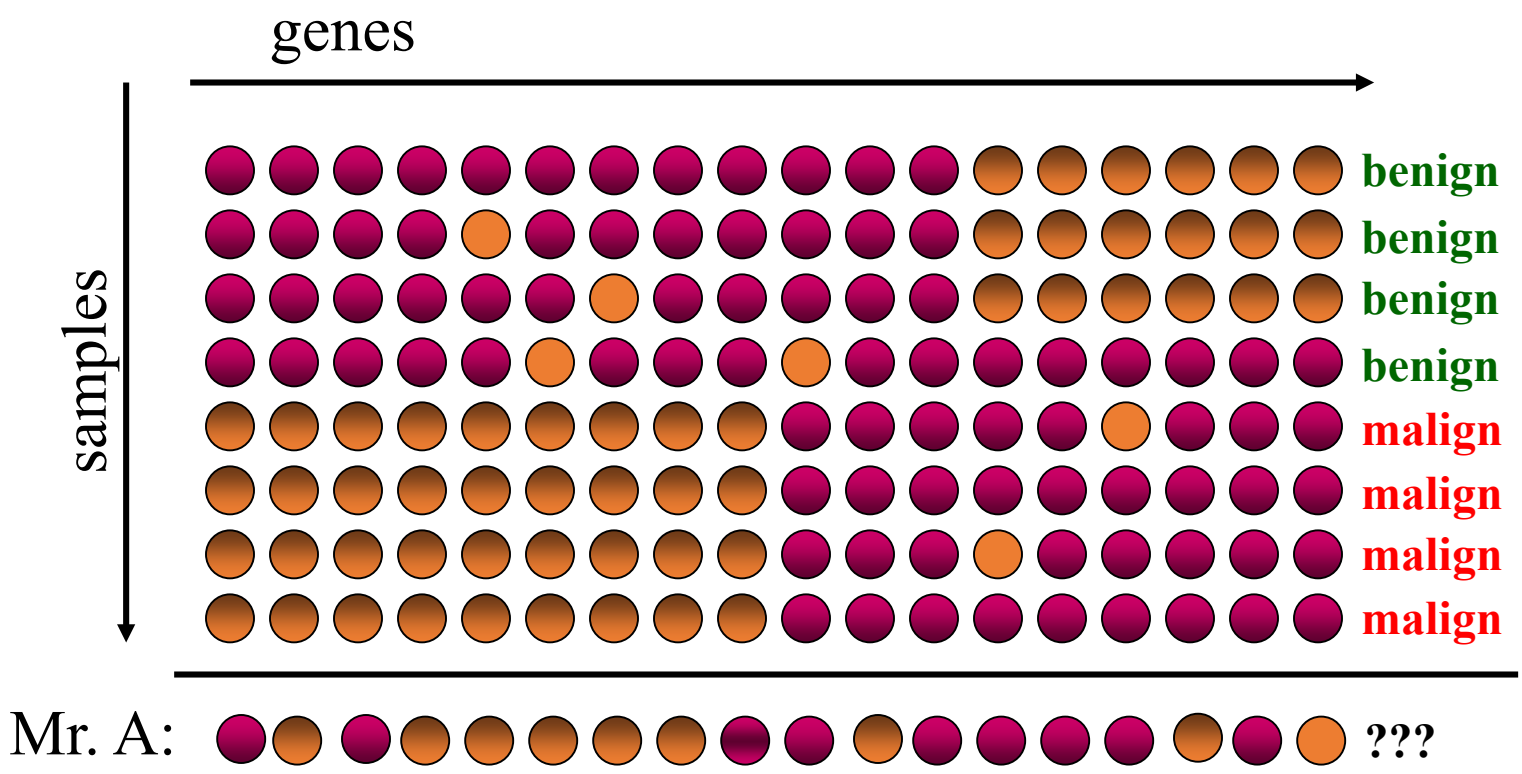
Does Mr. A have cancer?

Let's rearrange the rows …



Does Mr. A have cancer?

and the columns too …



Induction/hypothesis: Benign (malignant) tumour has lots of red (blue) genes on the left and blue (red) genes on the right

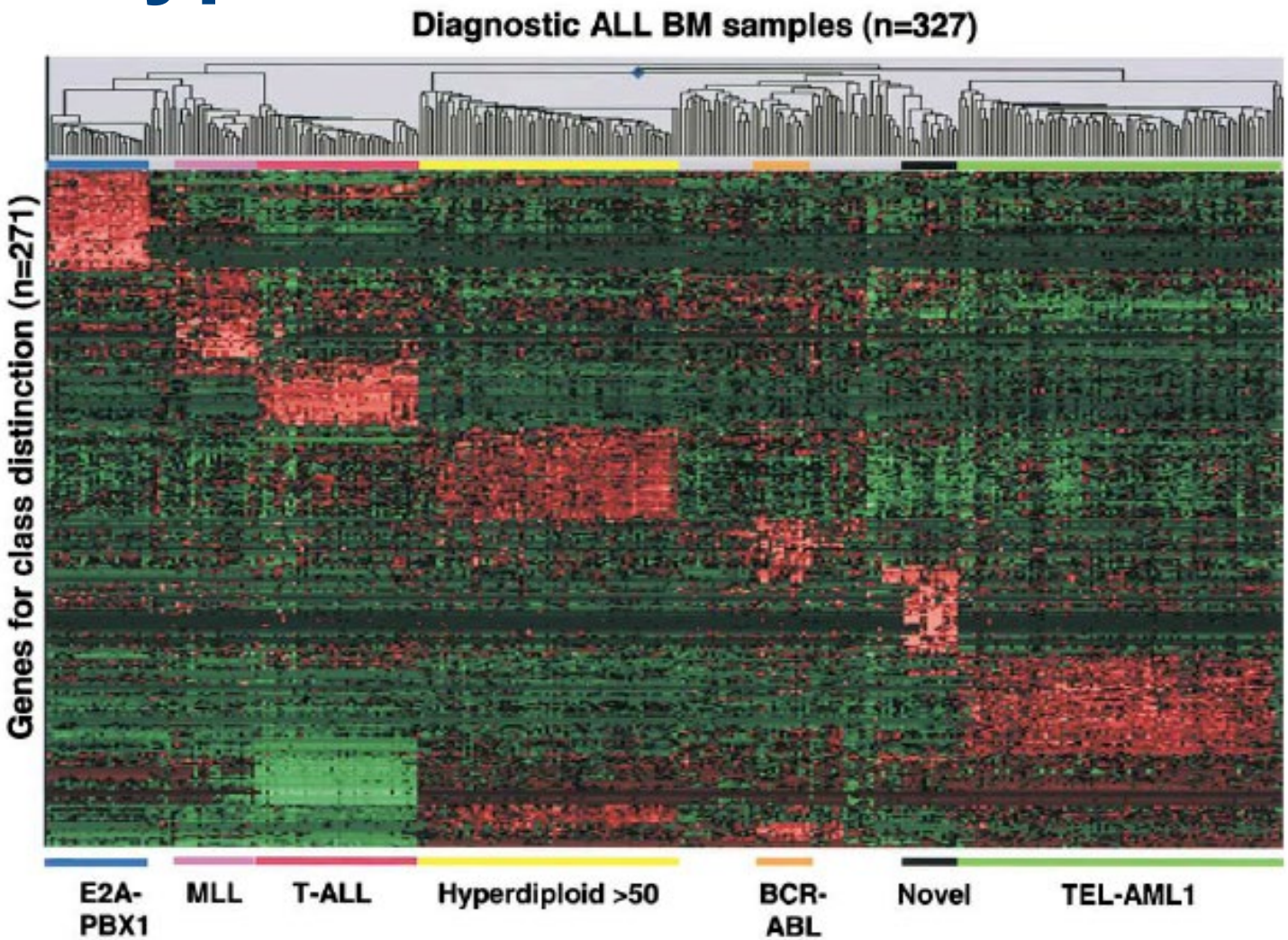
The triumph of logic

Induction/hypothesis: Benign (malignant) tumour has lots of red (blue) genes on the left and blue (red) genes on the right

Observation: Mr A's tumour has lots of blue genes on the left and red genes on the right

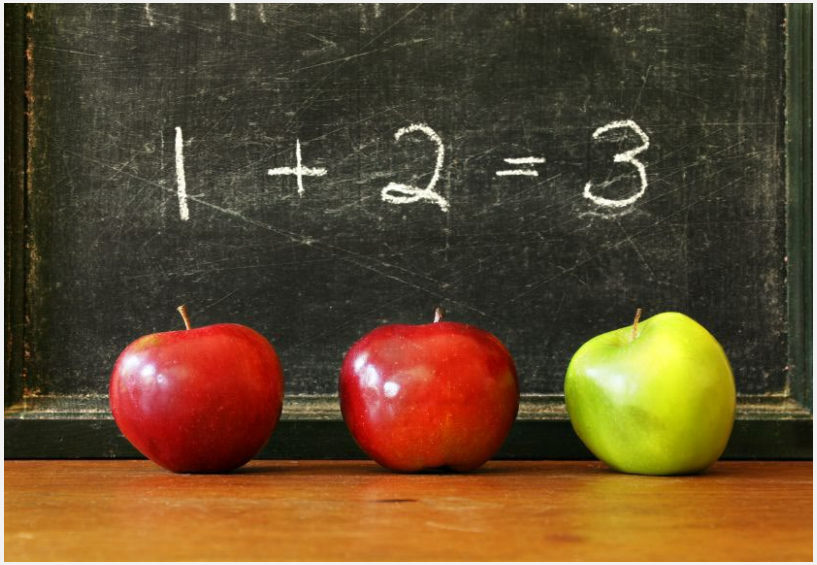
Abduction: Mr A's tumour is malignant

Invariant profile of leukemia subtypes



What have we just seen?

Guilt by association of invariants


1 + 2 = 3

What have we learned?

Three types of logical reasoning

Induction, abduction, deduction

Invariant is fundamental property of many problems

Tactics of logical problem solving

Logical reasoning on invariants

Rectifying/monitoring violation of invariants

Guilt by association of invariants