Day 5
sampling - clustering

SAMPLE POPULATION

SAMPLING: IS ESTIMATING THE CHARACTERISTICS OF THE WHOLE POPULATION USING INFORMATION COLLECTED FROM A **SAMPLE** GROUP.

The sampling process comprises several stages:

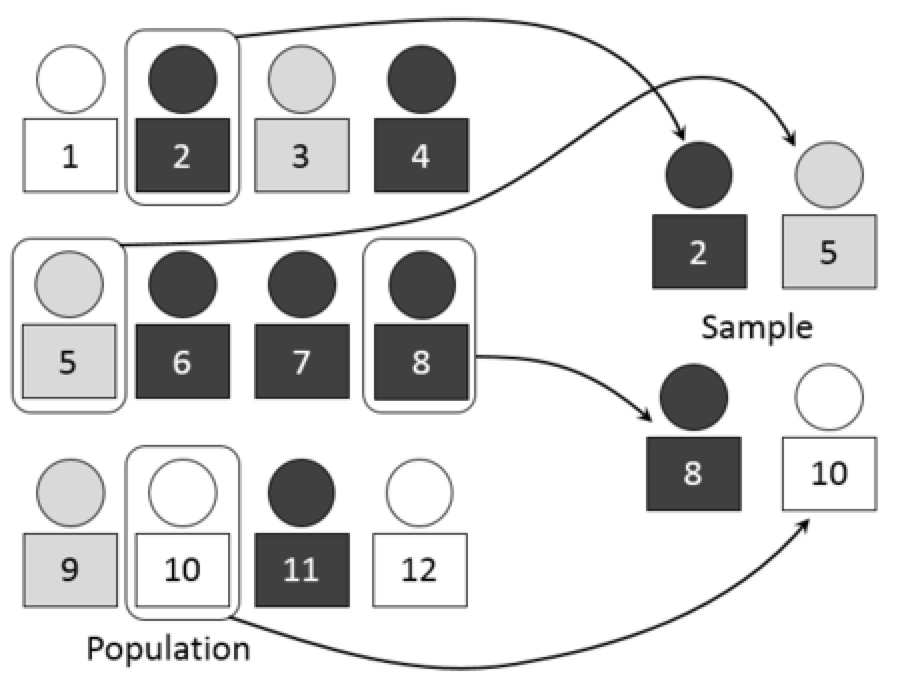
- Defining the population of concern
- Specifying a sampling frame, a set of items or events possible to measure
- Specifying a sampling method for selecting items or events from the frame
- Determining the sample size
- Implementing the sampling plan
- Sampling and data collecting

Simple random sampling

In a simple random sample (SRS) of a given size, all such subsets of the frame are given an equal probability.

In particular, the variance between individual results within the sample is a good indicator of variance in the overall population, which makes it relatively easy to estimate the accuracy of results.

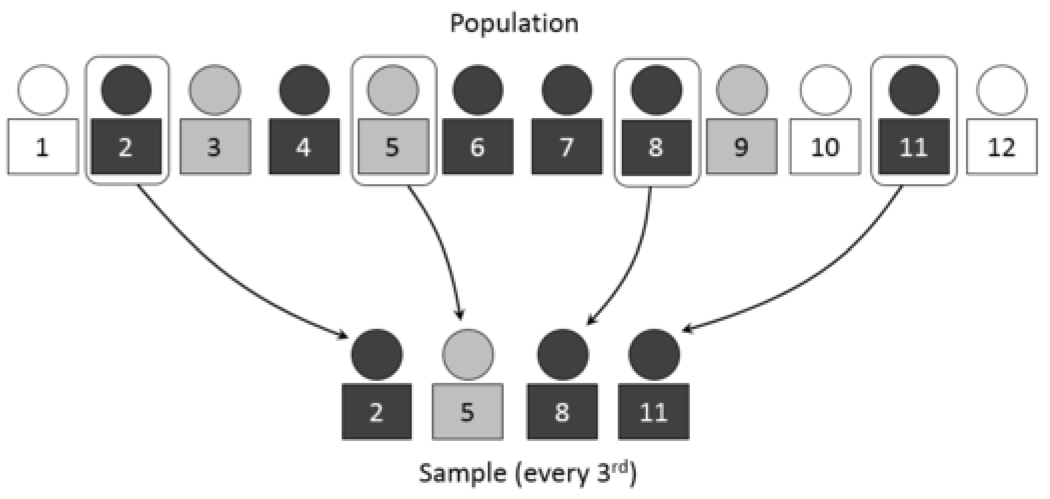
SRS can be vulnerable to sampling error because the randomness of the selection may result in a sample that doesn't reflect the makeup of the population.



Systematic sampling

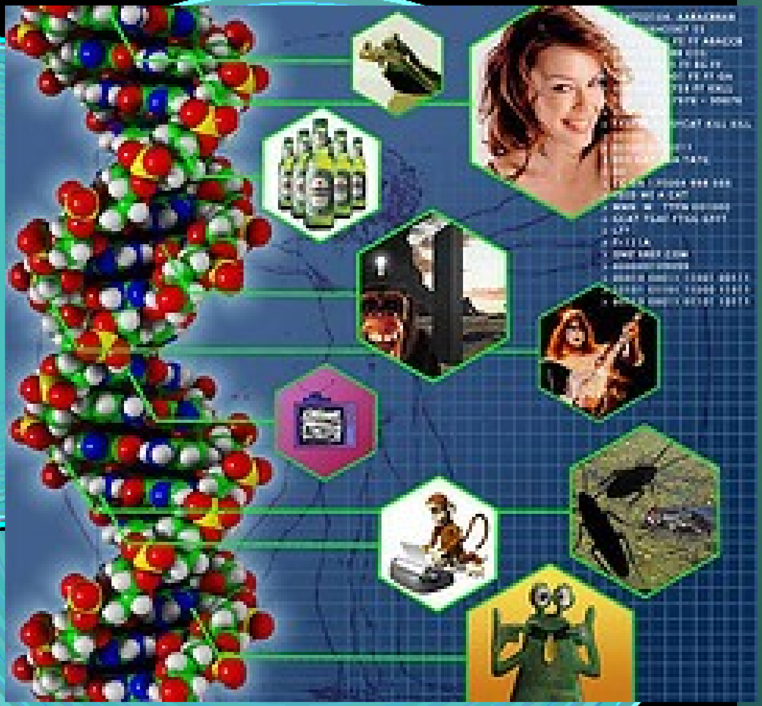
Systematic sampling (also known as interval sampling) relies on arranging the study population according to some ordering scheme and then selecting elements at regular intervals through that ordered list

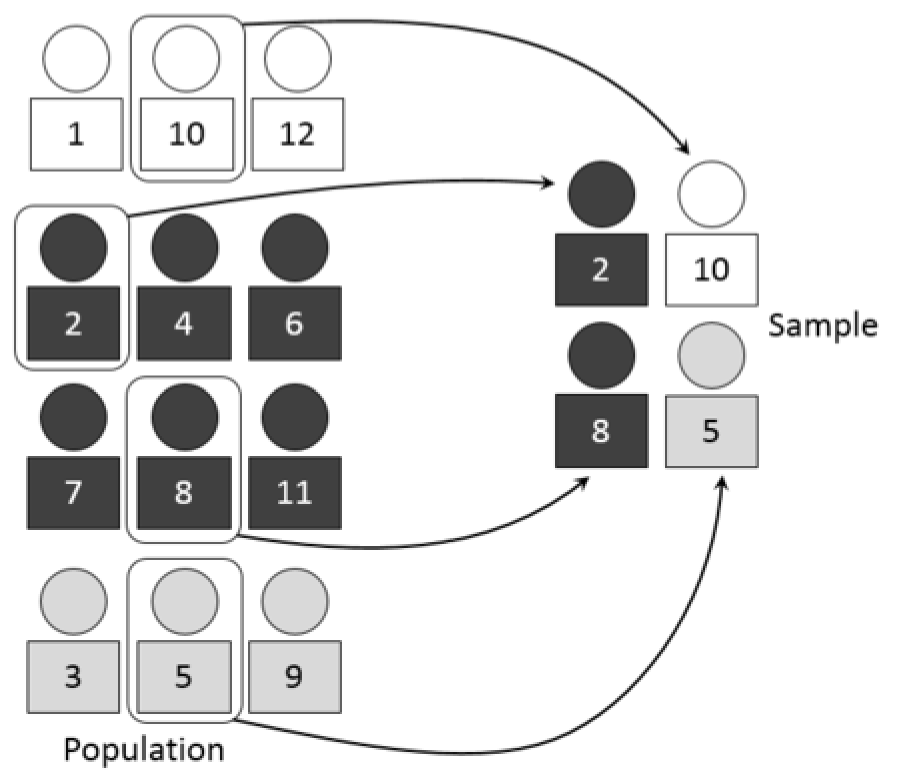
Systematic sampling involves a random start and then proceeds with the selection of every *k*th element from then onwards. In this case, *k*=(population size/sample size). It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the *k*th element in the list.



STRATIFIED SAMPLING

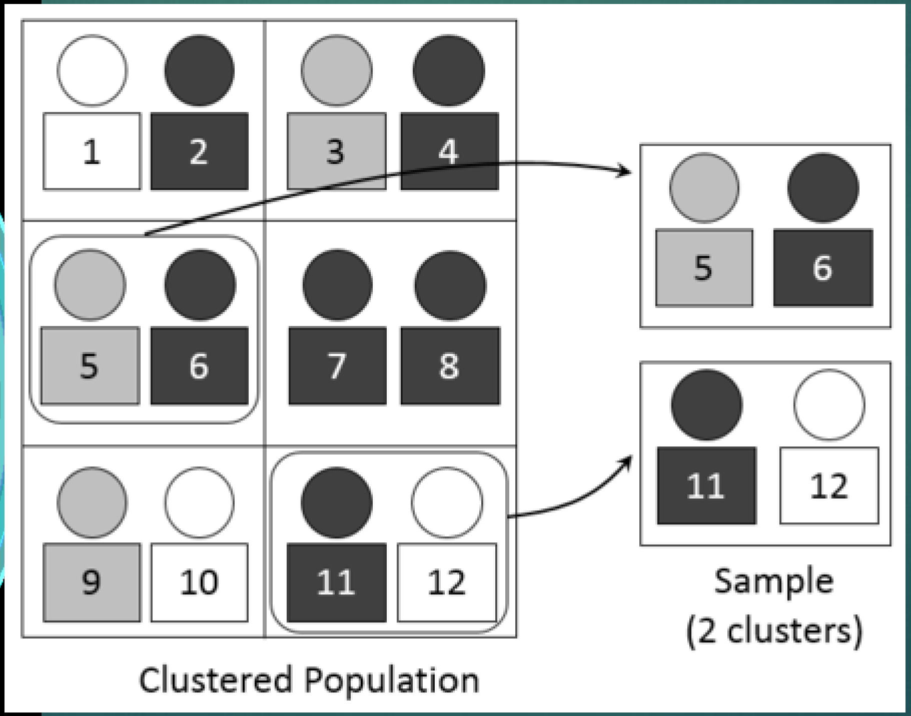
WHEN THE POPULATION EMBRACES A NUMBER OF DISTINCT CATEGORIES, THE FRAME CAN BE ORGANIZED BY THESE CATEGORIES INTO SEPARATE "STRATA." EACH STRATUM IS THEN SAMPLED AS AN INDEPENDENT SUB-POPULATION, OUT OF WHICH INDIVIDUAL ELEMENTS CAN BE RANDOMLY SELECTED





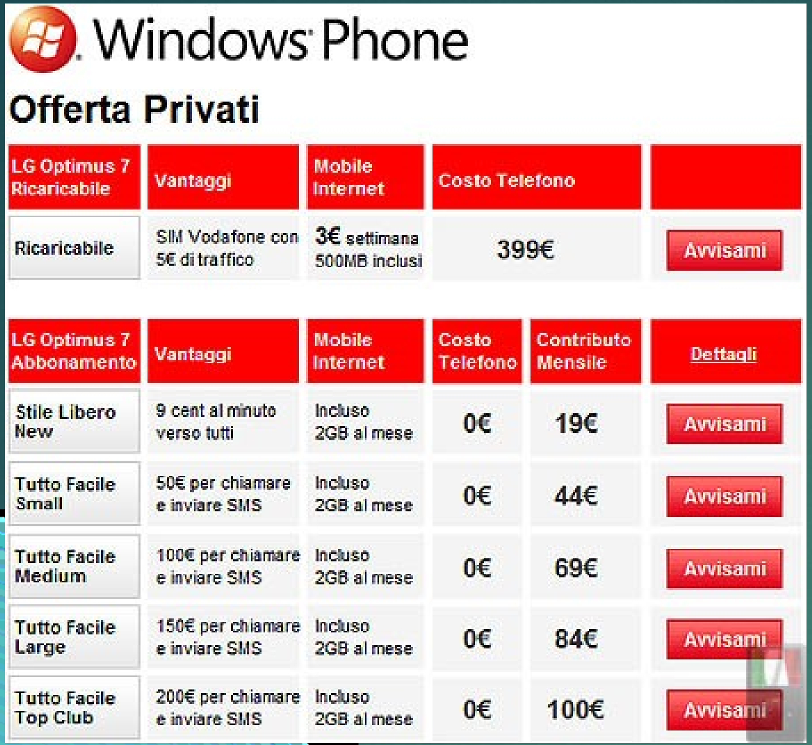
Cluster sampling

- Sometimes it is more cost-effective to select respondents in groups ('clusters')



Quota sampling
Minimax sampling
Accidental sampling
Voluntary Sampling

....



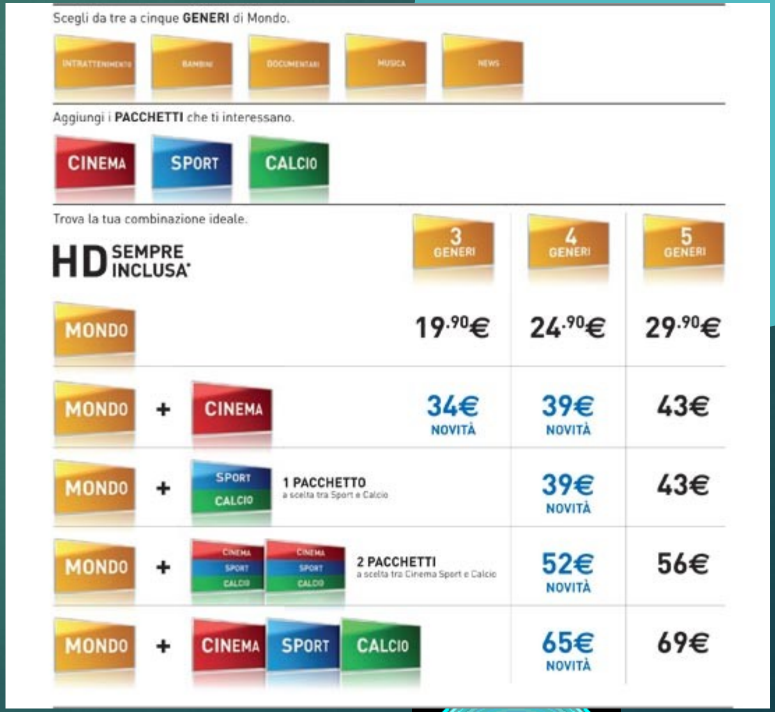
Windows Phone

Offerta Privati

LG Optimus 7 Ricaricabile	Vantaggi	Mobile Internet	Costo Telefono	
Ricaricabile	SIM Vodafone con 5€ di traffico	3€ settimana 500MB inclusi	399€	Avvisami

LG Optimus 7 Abbonamento	Vantaggi	Mobile Internet	Costo Telefono	Contributo Mensile	Dettagli
Stile Libero New	9 cent al minuto verso tutti	Incluso 2GB al mese	0€	19€	Avvisami
Tutto Facile Small	50€ per chiamare e inviare SMS	Incluso 2GB al mese	0€	44€	Avvisami
Tutto Facile Medium	100€ per chiamare e inviare SMS	Incluso 2GB al mese	0€	69€	Avvisami
Tutto Facile Large	150€ per chiamare e inviare SMS	Incluso 2GB al mese	0€	84€	Avvisami
Tutto Facile Top Club	200€ per chiamare e inviare SMS	Incluso 2GB al mese	0€	100€	Avvisami

sky



Scegli da tre a cinque **GENERI** di Mondo.

INTRATTENIMENTO BAMBINI DOCUMENTARI MUSICA NEWS

Aggiungi i **PACCHETTI** che ti interessano.

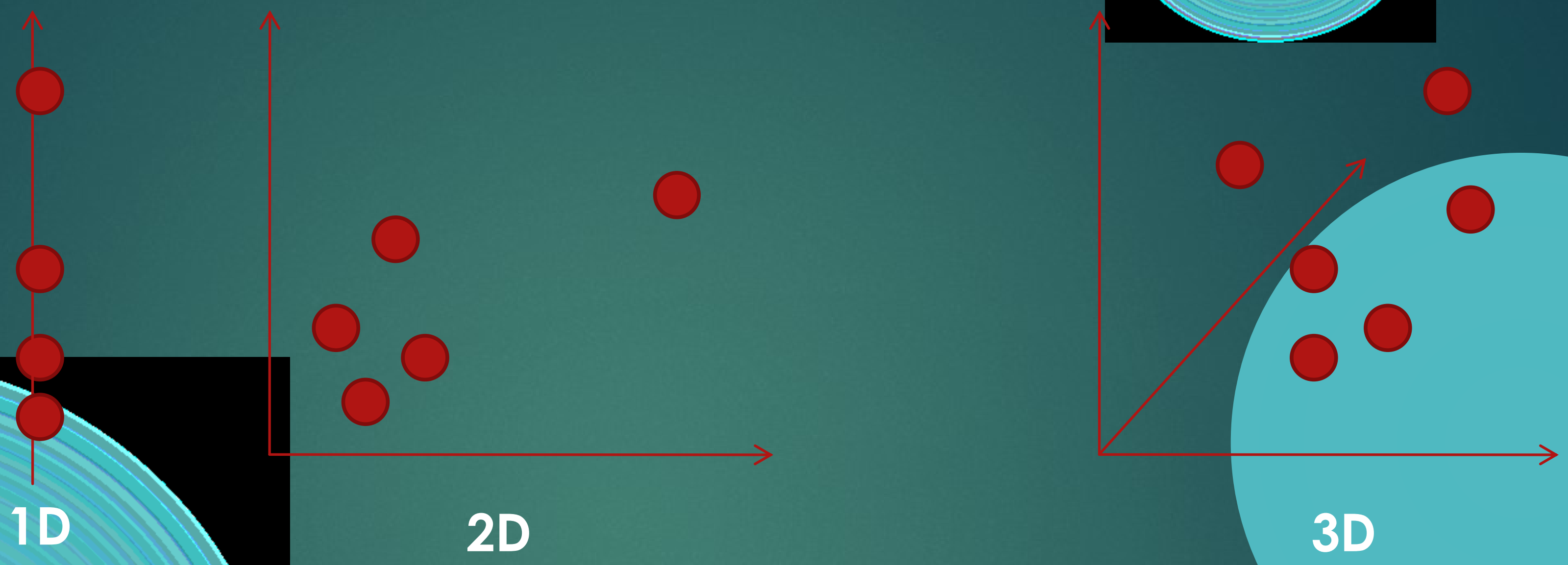
CINEMA SPORT CALCIO

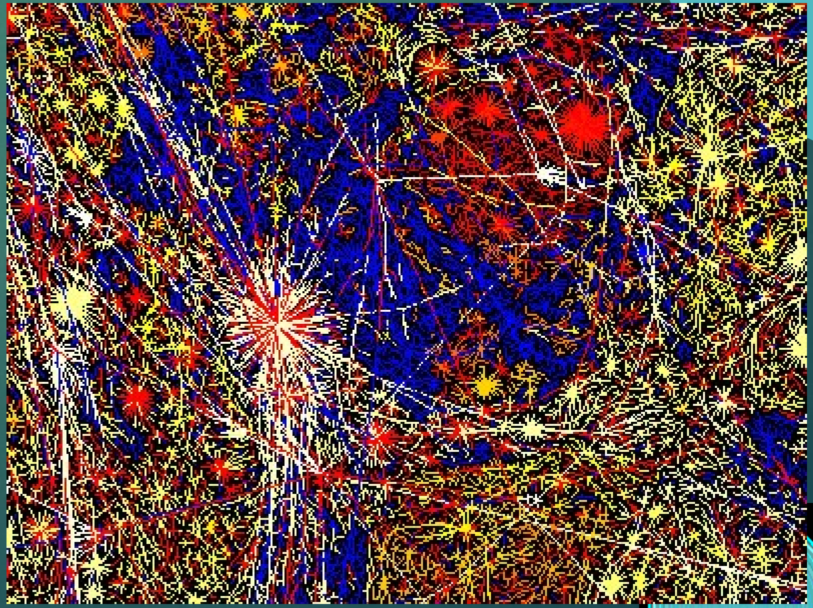
Trova la tua combinazione ideale.

HD SEMPRE INCLUSA*	3 GENERI	4 GENERI	5 GENERI
MONDO	19.90€	24.90€	29.90€
MONDO + CINEMA	34€ NOVITÀ	39€ NOVITÀ	43€
MONDO + SPORT / CALCIO 1 PACCHETTO a scelta tra Sport e Calcio		39€ NOVITÀ	43€
MONDO + CINEMA SPORT CALCIO / CINEMA SPORT CALCIO 2 PACCHETTI a scelta tra Cinema Sport e Calcio		52€ NOVITÀ	56€
MONDO + CINEMA SPORT CALCIO		65€ NOVITÀ	69€

vodafone

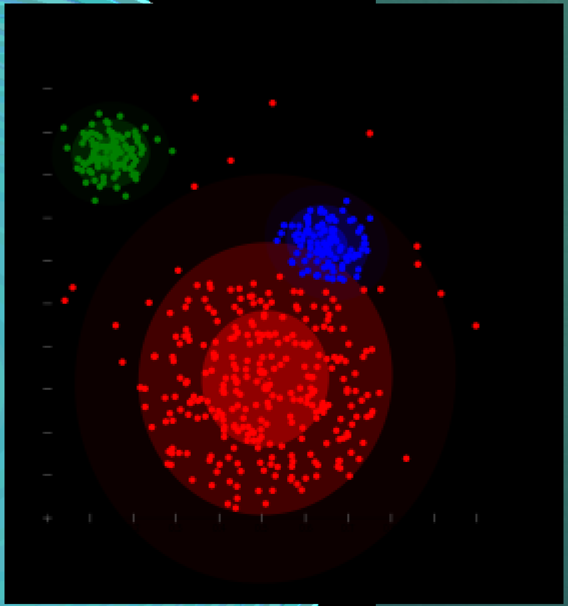
CLUSTERING

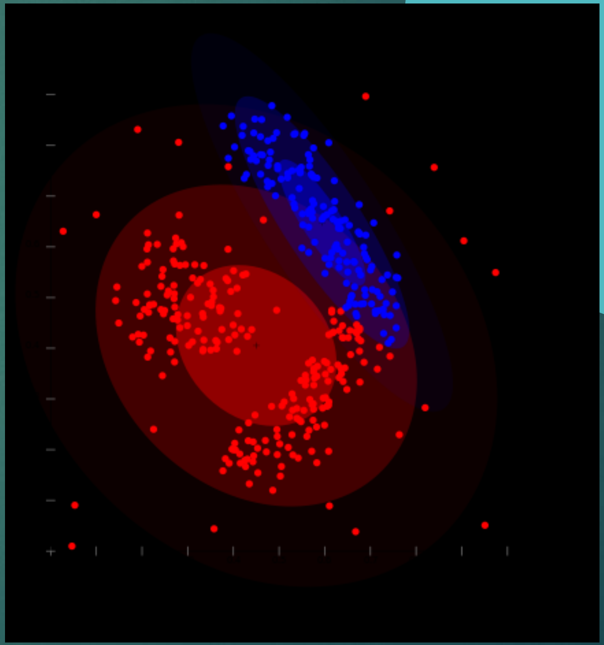




Clustering techniques are based on measures of similarity between elements. In many approaches this similarity is conceived in terms of distance in a multidimensional space.

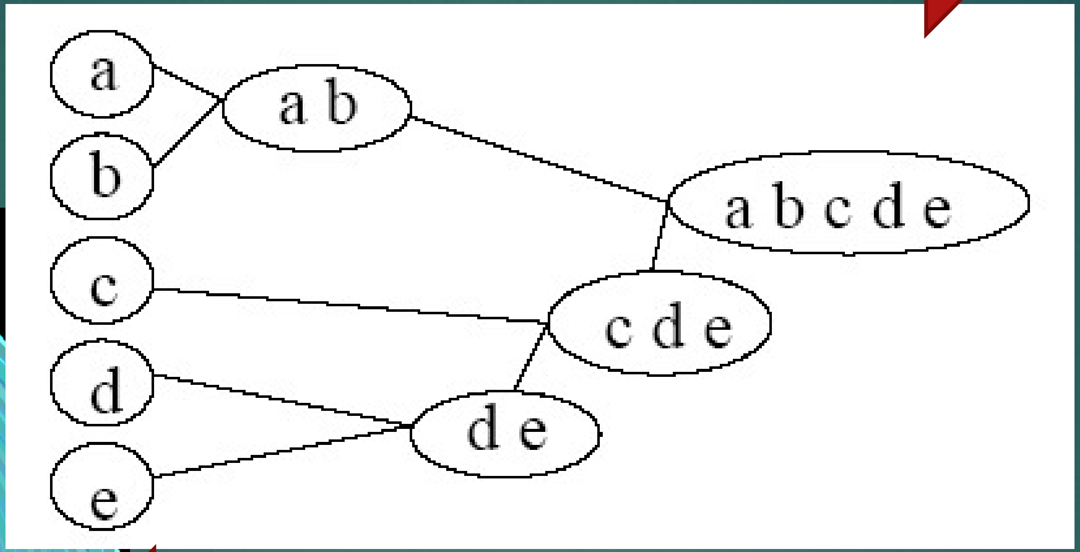
The notion of a "cluster" cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms.There is a common denominator: a group of data objects





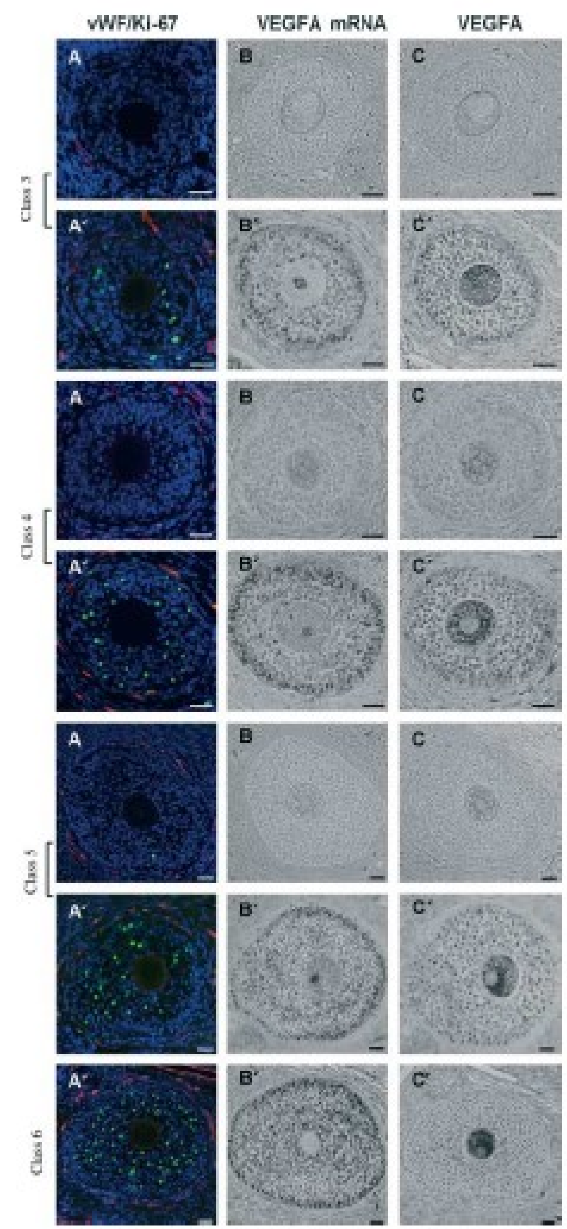
Clustering

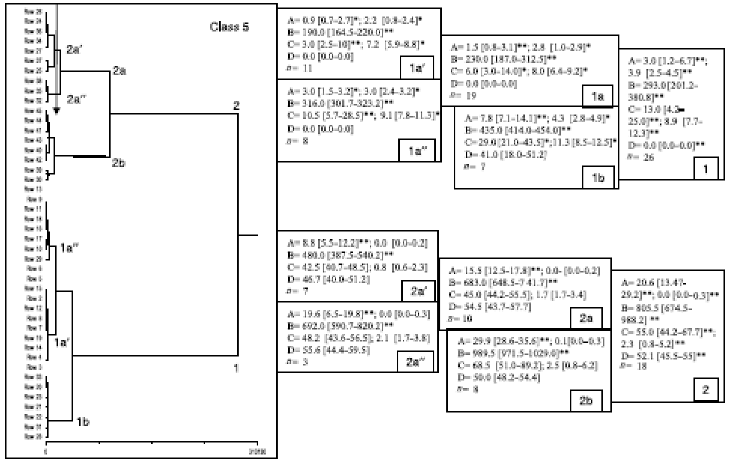
Bottom-up



Top-Down

dendogramma





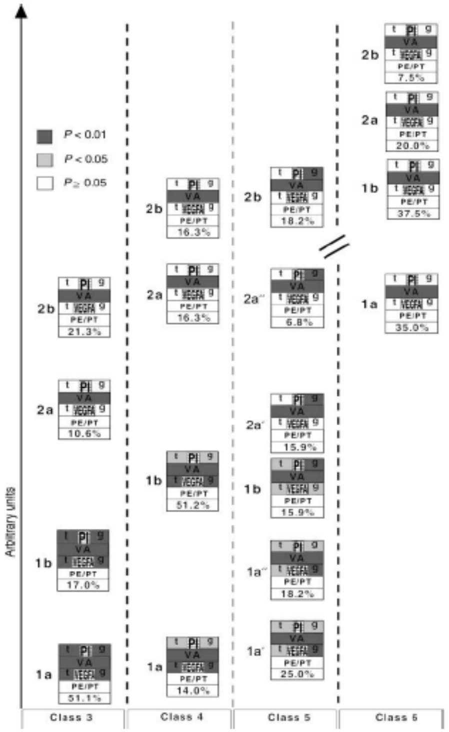


Figure 5 A general model that describes the distribution of the follicular subpopulations identified inside each follicular class by using cluster analysis. The subpopulations were represented within each class and arranged on the *y*-axis by considering the distances between the bifurcation obtained in the dendograms. The grey scale represents significant differences recovered for somatic and vascular parameters. The thickness of the dotted line represents the statistical difference among classes. t PI, theca proliferation index; g PI, granulosa proliferation index; VA, vascular area; t VEGFA, theca *VEGFA* mRNA; g VEGFA, granulosa *VEGFA* mRNA; PE/PT, proportion of proliferating endothelial cells; % value, percentage of preantral follicles belonging to the subpopulation.

CLUSTERING

Comp Clin Pathol

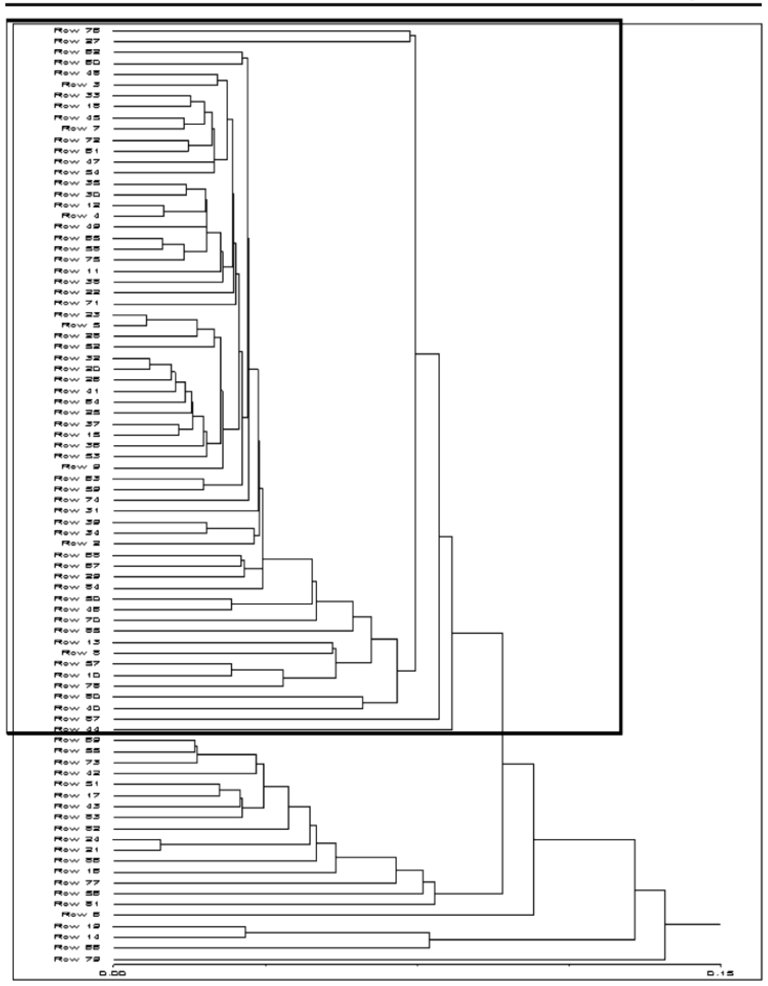


Fig. 2 Multivariate hierarchical cluster analysis of MICRO dogs. The *caption* shows the area of interest where clustering is significant

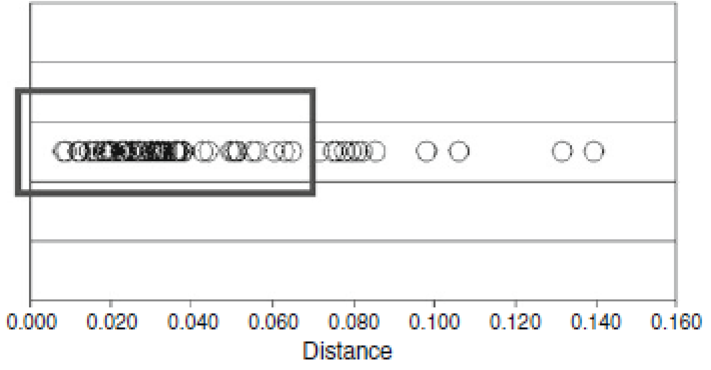
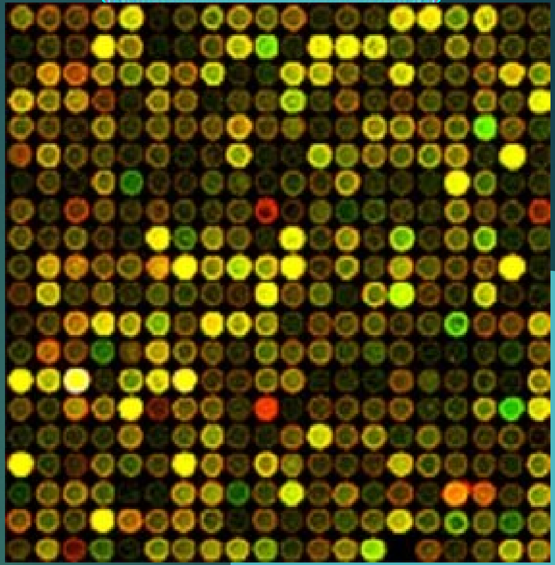
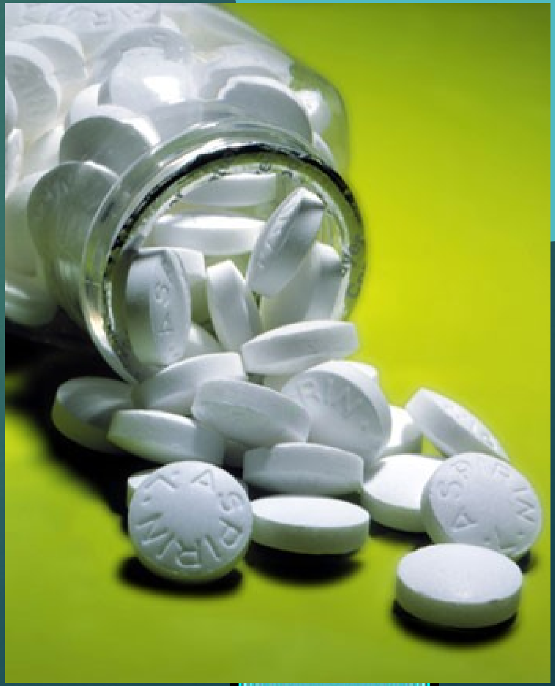


Fig. 3 Mono-dimensional distribution of distances among the dogs as measured by clustering. The *caption* shows a clear subgroup among the MICRO dogs

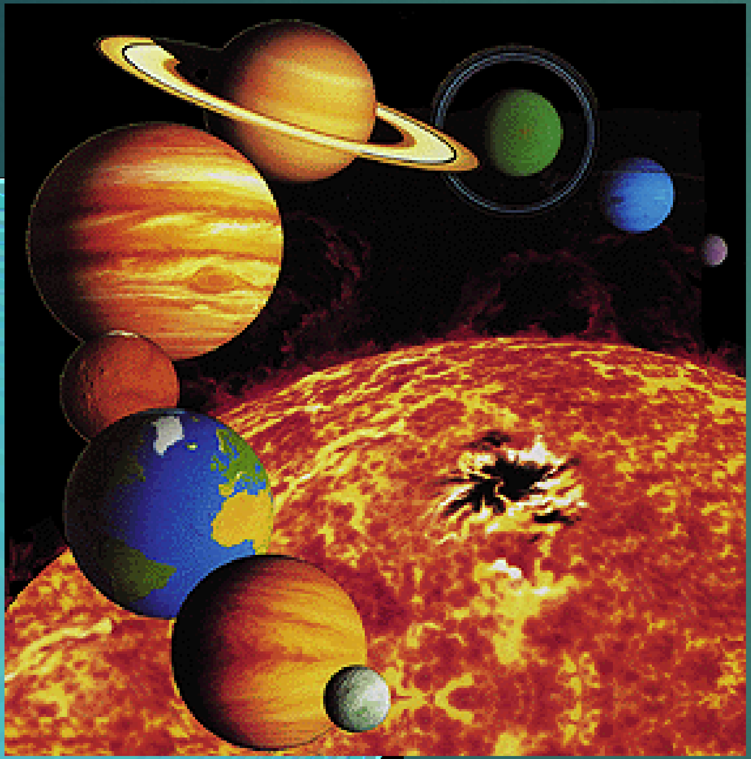
Inividuality
Complexity



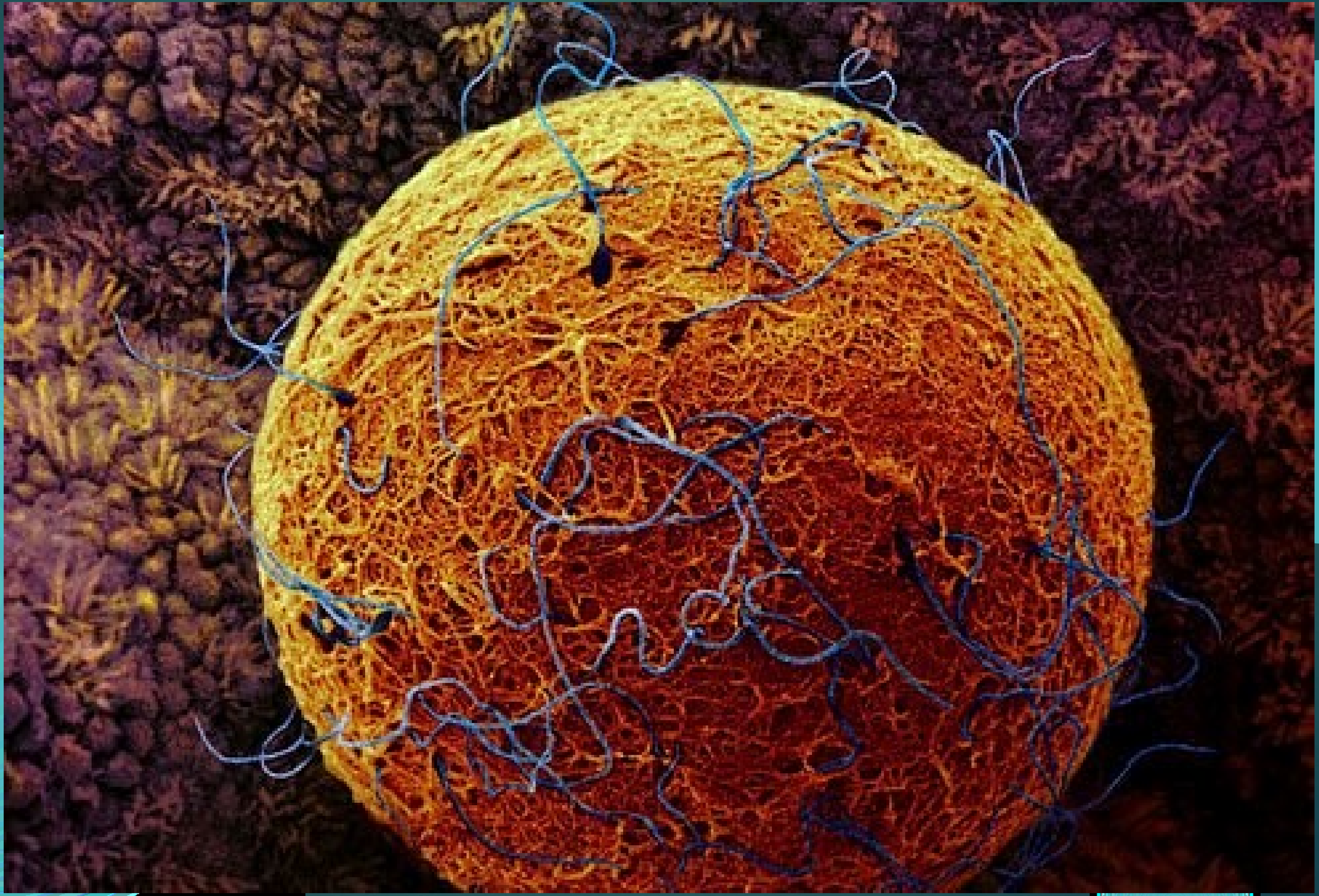
© www.matteofossati.it



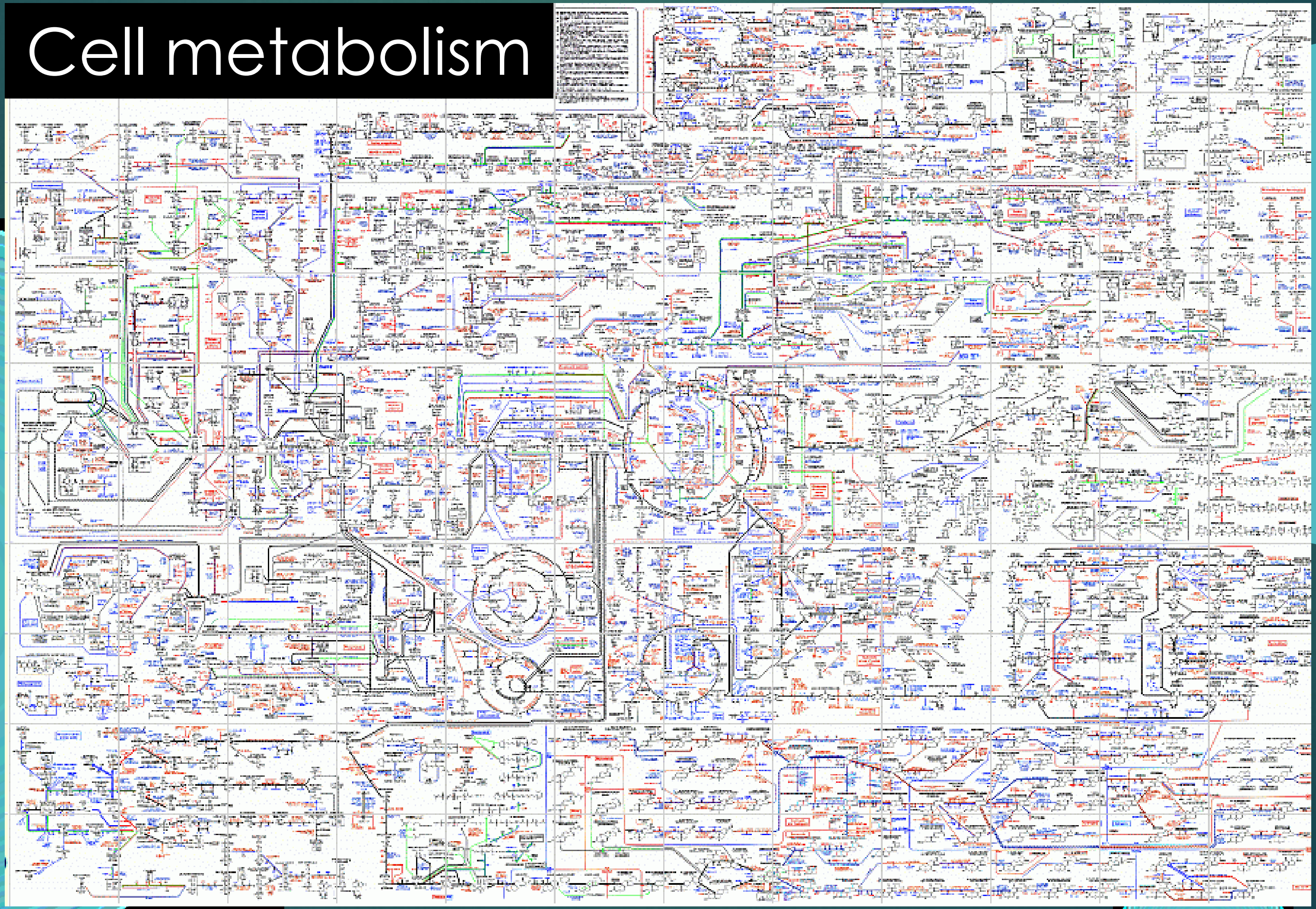
Complicated vs. complex



fertility



Biological complexity

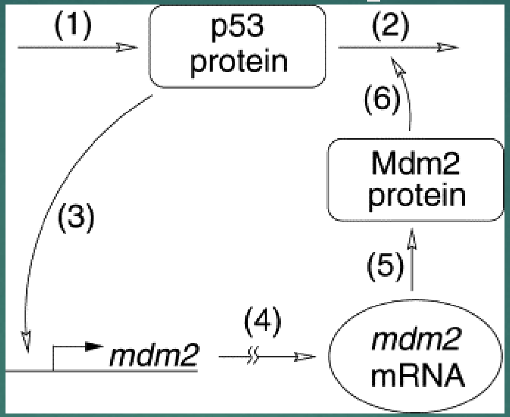


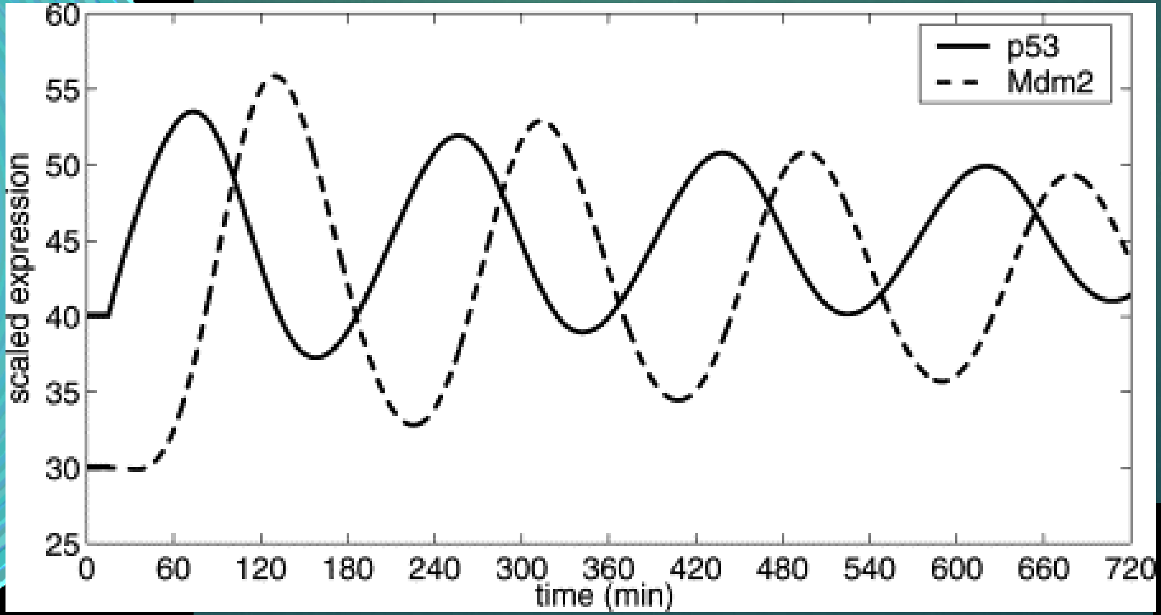
Cell metabolism

Complexity:

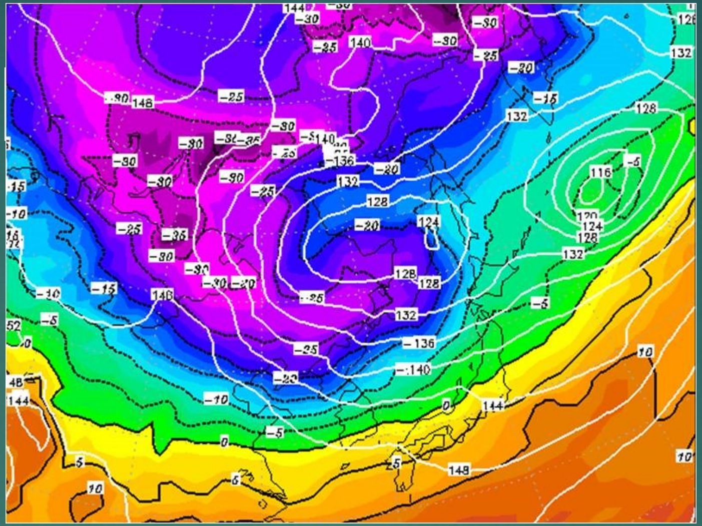
non-linearity of interactions:

An example…

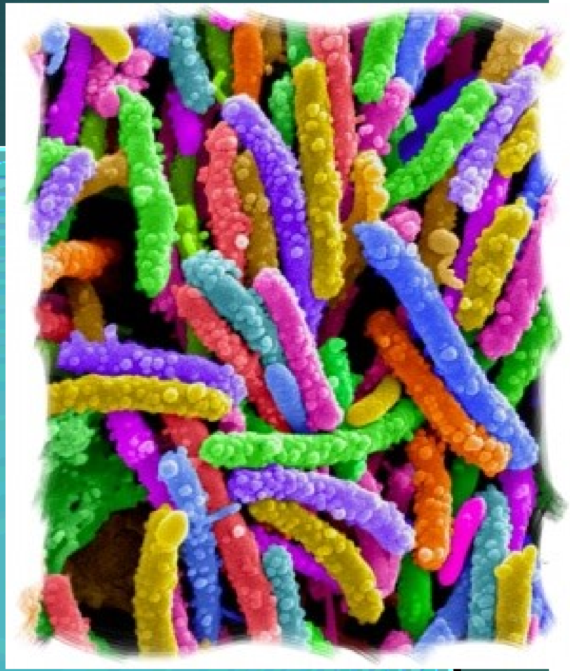

(1)
p53 protein
(2)
(6)
Mdm2 protein
(3)
(5)
(4)
mdm2
mdm2 mRNA

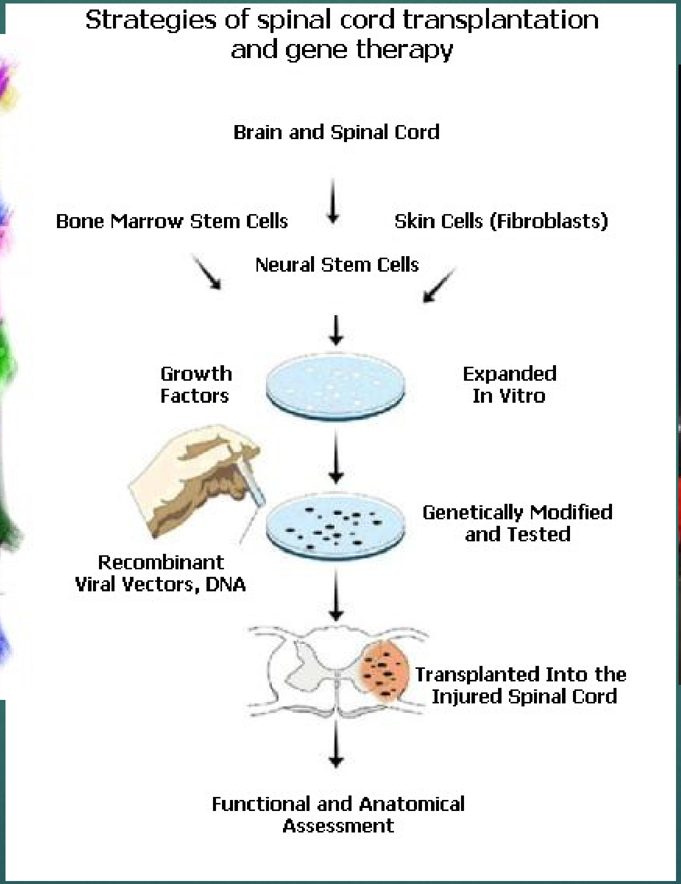

p53
Mdm2
scaled expression
60
55
50
45
40
35
30
25
0
60
120
180
240
300
360
420
480
540
600
660
720
time (min)

un



unpredictability





Strategies of spinal cord transplantation and gene therapy

Brain and Spinal Cord

Bone Marrow Stem Cells

Skin Cells (Fibroblasts)

Neural Stem Cells

Growth Factors

Expanded In Vitro

Genetically Modified and Tested

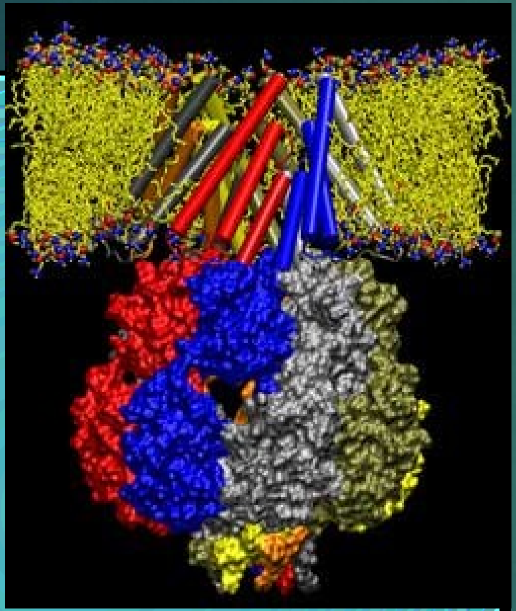
Recombinant Viral Vectors, DNA

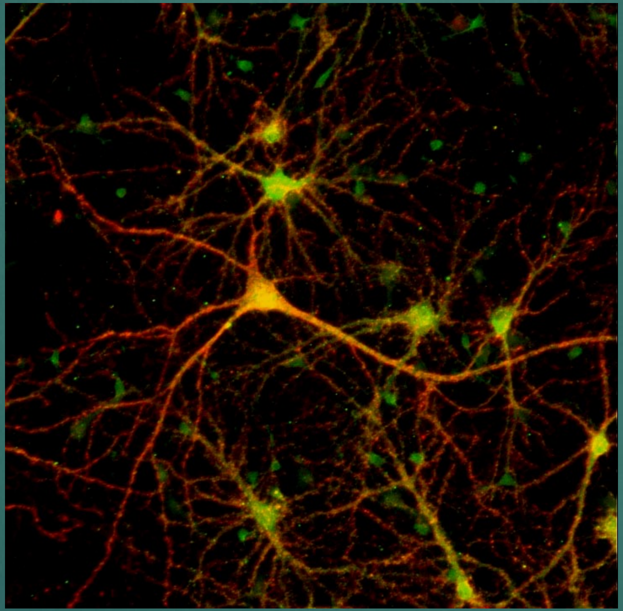
Transplanted Into the Injured Spinal Cord

Functional and Anatomical Assessment

Butterfly effect

Emergence of proprieties





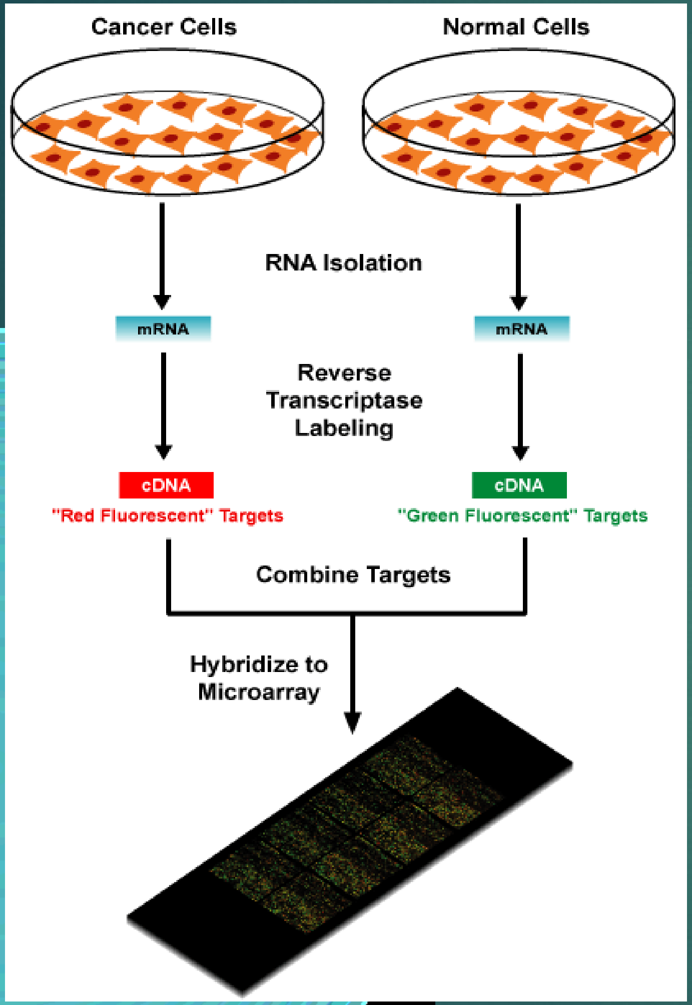
The whole is more (different) than the sum of the individual components

Human Genome Project

Begun formally in 1990, the U.S. Human Genome Project was a 13-year effort coordinated by the U.S. Department of Energy and the National Institutes of Health. The project originally was planned to last 15 years, but rapid technological advances accelerated the completion date to 2003. Project goals

- *identify* all the approximately 20,000-25,000 genes in human DNA,
- *determine* the sequences of the 3 billion chemical base pairs that make up human DNA,
- *store* this information in databases,
- *improve* tools for data analysis,
- *transfer* related technologies to the private sector, and
- *address* the ethical, legal, and social issues (ELSI) that may arise from the project.

- omics



Cancer Cells

Normal Cells

RNA Isolation

mRNA

mRNA

Reverse Transcriptase Labeling

cDNA

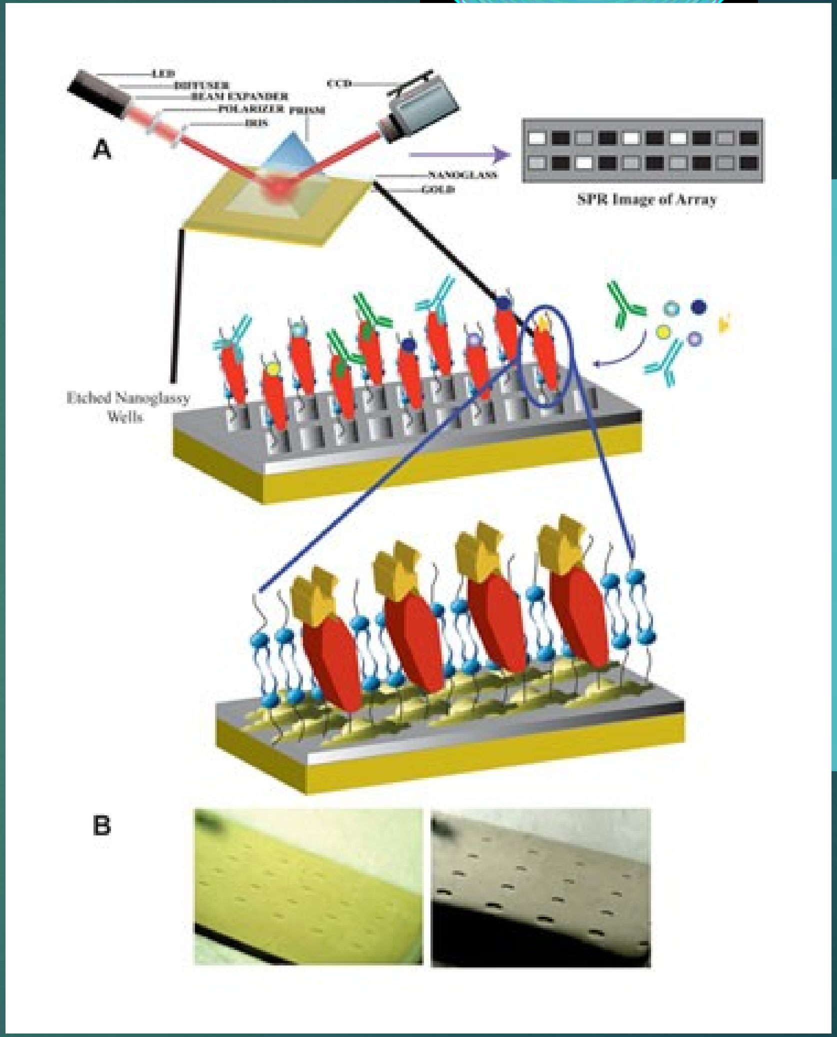
"Red Fluorescent" Targets

cDNA

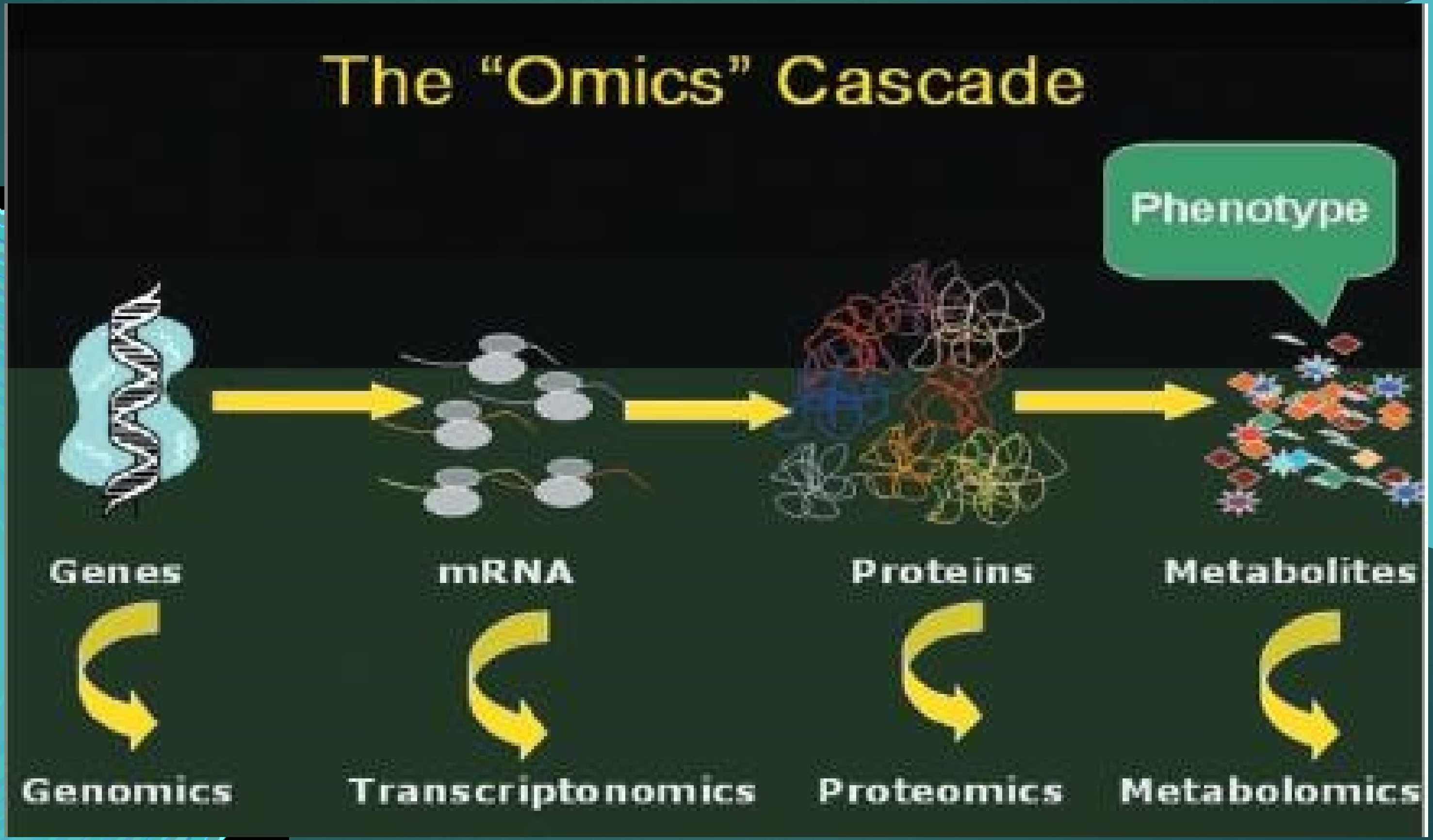
"Green Fluorescent" Targets

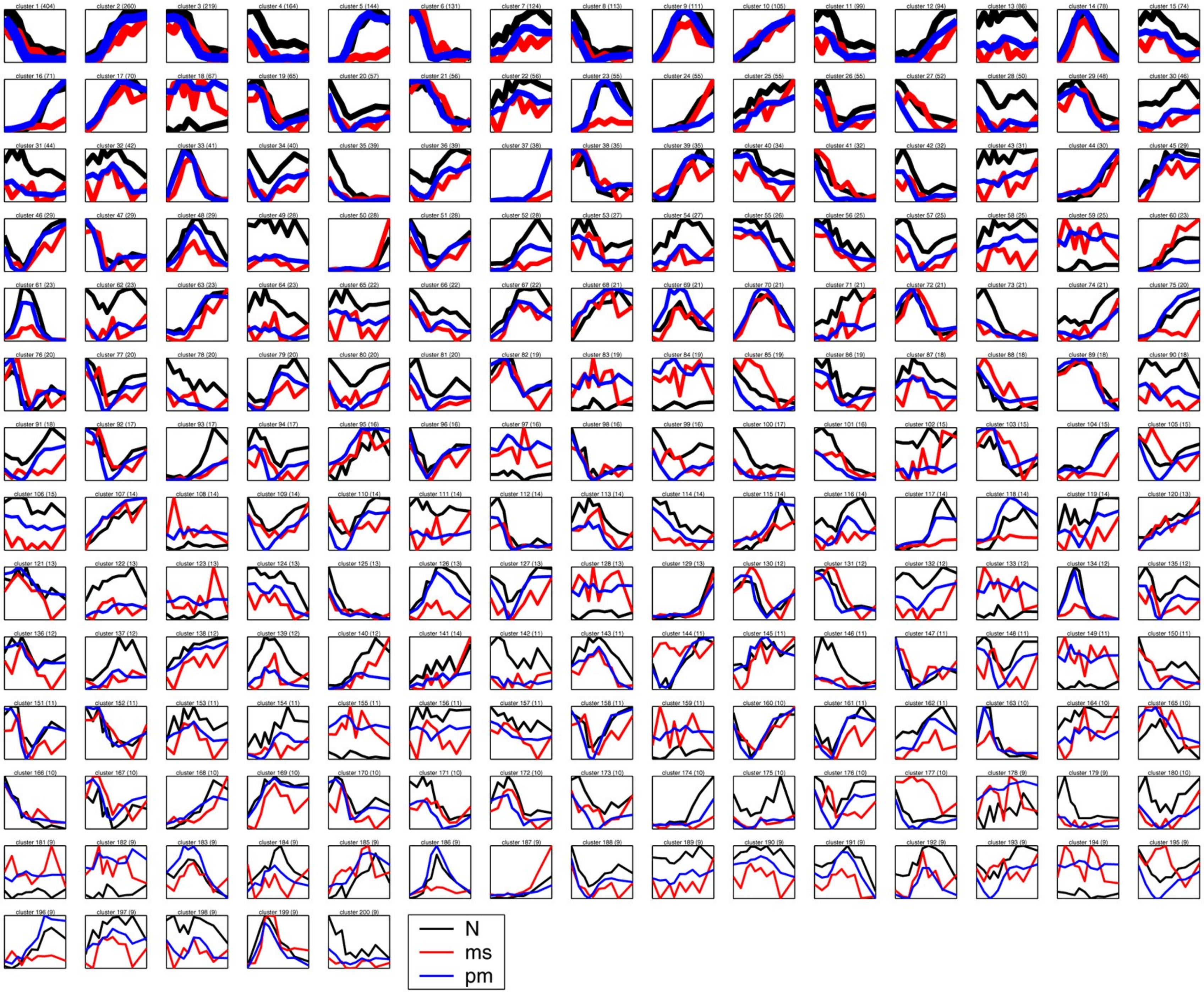
Combine Targets

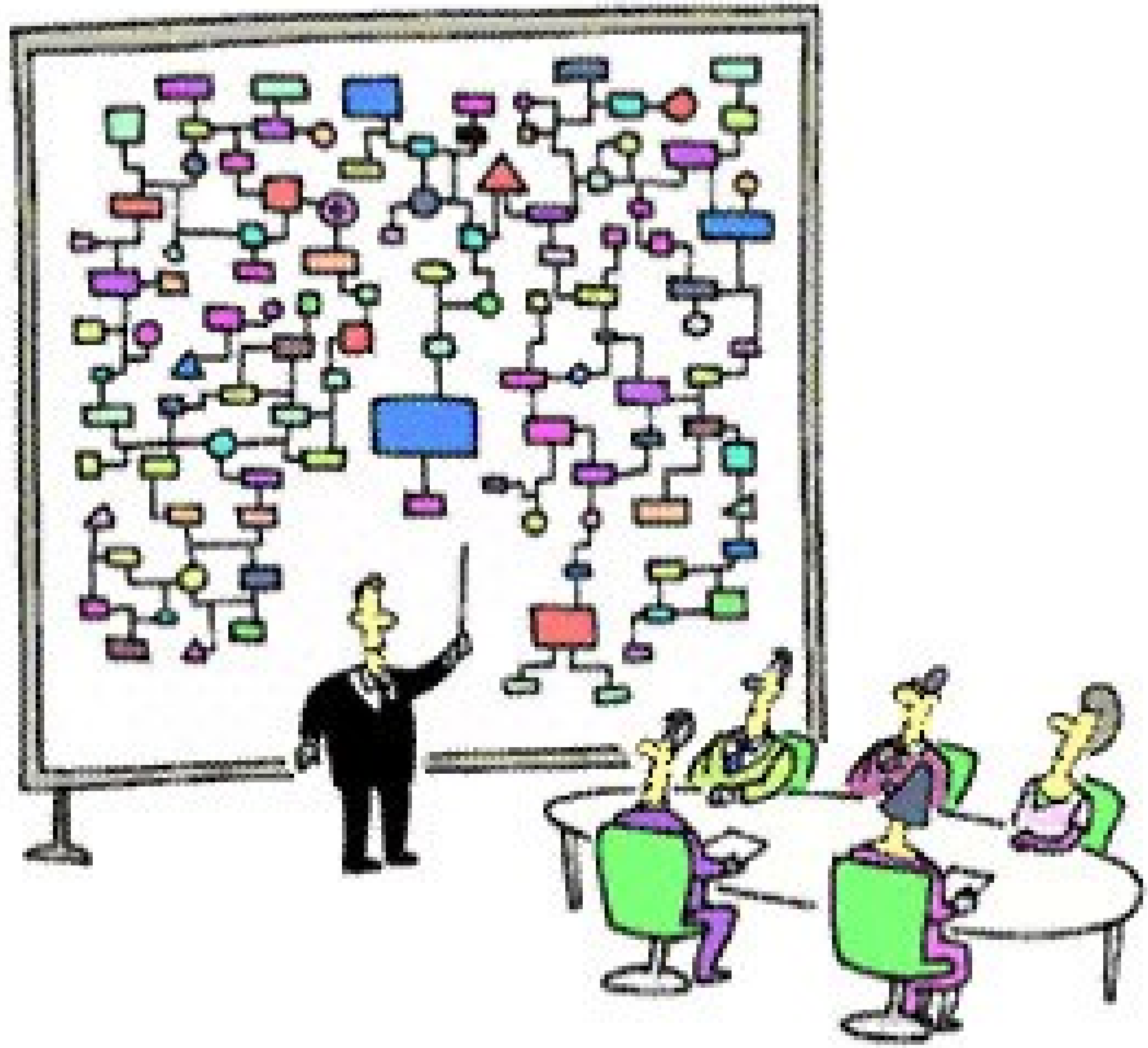
Hybridize to Microarray



- OMICS

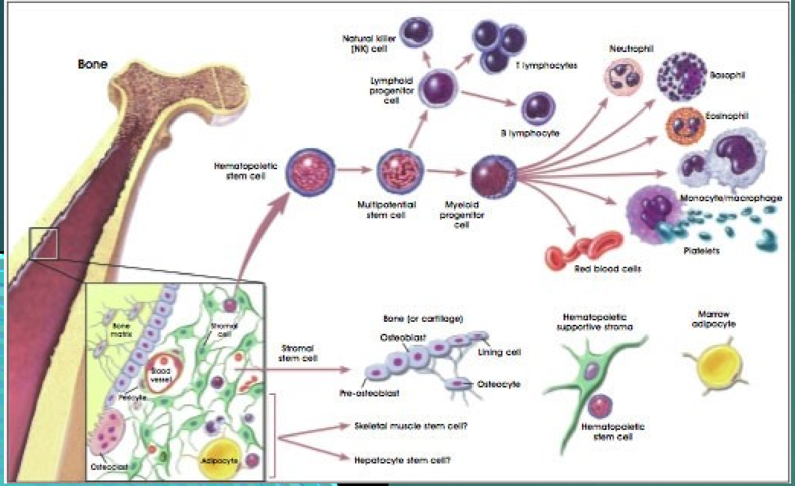

The "Omics" Cascade
Phenotype
Genes
mRNA
Proteins
Metabolites
Genomics
Transcriptonomics
Proteomics
Metabolomics

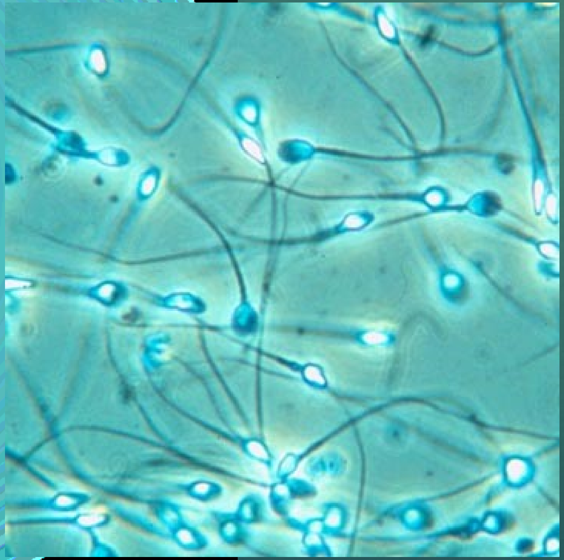

cluster 1 (404)
cluster 2 (260)
cluster 3 (219)
cluster 4 (164)
cluster 5 (144)
cluster 6 (131)
cluster 7 (124)
cluster 8 (113)
cluster 9 (111)
cluster 10 (105)
cluster 11 (99)
cluster 12 (94)
cluster 13 (86)
cluster 14 (78)
cluster 15 (74)
cluster 16 (71)
cluster 17 (70)
cluster 18 (67)
cluster 19 (65)
cluster 20 (57)
cluster 21 (56)
cluster 22 (56)
cluster 23 (55)
cluster 24 (55)
cluster 25 (55)
cluster 26 (55)
cluster 27 (52)
cluster 28 (50)
cluster 29 (48)
cluster 30 (46)
cluster 31 (44)
cluster 32 (42)
cluster 33 (41)
cluster 34 (40)
cluster 35 (39)
cluster 36 (39)
cluster 37 (38)
cluster 38 (35)
cluster 39 (35)
cluster 40 (34)
cluster 41 (32)
cluster 42 (32)
cluster 43 (31)
cluster 44 (30)
cluster 45 (29)
cluster 46 (29)
cluster 47 (29)
cluster 48 (29)
cluster 49 (28)
cluster 50 (28)
cluster 51 (28)
cluster 52 (28)
cluster 53 (27)
cluster 54 (27)
cluster 55 (26)
cluster 56 (25)
cluster 57 (25)
cluster 58 (25)
cluster 59 (25)
cluster 60 (23)
cluster 61 (23)
cluster 62 (23)
cluster 63 (23)
cluster 64 (23)
cluster 65 (22)
cluster 66 (22)
cluster 67 (22)
cluster 68 (21)
cluster 69 (21)
cluster 70 (21)
cluster 71 (21)
cluster 72 (21)
cluster 73 (21)
cluster 74 (21)
cluster 75 (20)
cluster 76 (20)
cluster 77 (20)
cluster 78 (20)
cluster 79 (20)
cluster 80 (20)
cluster 81 (20)
cluster 82 (19)
cluster 83 (19)
cluster 84 (19)
cluster 85 (19)
cluster 86 (19)
cluster 87 (18)
cluster 88 (18)
cluster 89 (18)
cluster 90 (18)
cluster 91 (18)
cluster 92 (17)
cluster 93 (17)
cluster 94 (17)
cluster 95 (16)
cluster 96 (16)
cluster 97 (16)
cluster 98 (16)
cluster 99 (16)
cluster 100 (17)
cluster 101 (16)
cluster 102 (15)
cluster 103 (15)
cluster 104 (15)
cluster 105 (15)
cluster 106 (15)
cluster 107 (14)
cluster 108 (14)
cluster 109 (14)
cluster 110 (14)
cluster 111 (14)
cluster 112 (14)
cluster 113 (14)
cluster 114 (14)
cluster 115 (14)
cluster 116 (14)
cluster 117 (14)
cluster 118 (14)
cluster 119 (14)
cluster 120 (13)
cluster 121 (13)
cluster 122 (13)
cluster 123 (13)
cluster 124 (13)
cluster 125 (13)
cluster 126 (13)
cluster 127 (13)
cluster 128 (13)
cluster 129 (13)
cluster 130 (12)
cluster 131 (12)
cluster 132 (12)
cluster 133 (12)
cluster 134 (12)
cluster 135 (12)
cluster 136 (12)
cluster 137 (12)
cluster 138 (12)
cluster 139 (12)
cluster 140 (12)
cluster 141 (14)
cluster 142 (11)
cluster 143 (11)
cluster 144 (11)
cluster 145 (11)
cluster 146 (11)
cluster 147 (11)
cluster 148 (11)
cluster 149 (11)
cluster 150 (11)
cluster 151 (11)
cluster 152 (11)
cluster 153 (11)
cluster 154 (11)
cluster 155 (11)
cluster 156 (11)
cluster 157 (11)
cluster 158 (11)
cluster 159 (11)
cluster 160 (10)
cluster 161 (11)
cluster 162 (11)
cluster 163 (10)
cluster 164 (10)
cluster 165 (10)
cluster 166 (10)
cluster 167 (10)
cluster 168 (10)
cluster 169 (10)
cluster 170 (10)
cluster 171 (10)
cluster 172 (10)
cluster 173 (10)
cluster 174 (10)
cluster 175 (10)
cluster 176 (10)
cluster 177 (10)
cluster 178 (9)
cluster 179 (9)
cluster 180 (10)
cluster 181 (9)
cluster 182 (9)
cluster 183 (9)
cluster 184 (9)
cluster 185 (9)
cluster 186 (9)
cluster 187 (9)
cluster 188 (9)
cluster 189 (9)
cluster 190 (9)
cluster 191 (9)
cluster 192 (9)
cluster 193 (9)
cluster 194 (9)
cluster 195 (9)
cluster 196 (9)
cluster 197 (9)
cluster 198 (9)
cluster 199 (9)
cluster 200 (9)
N
ms
pm

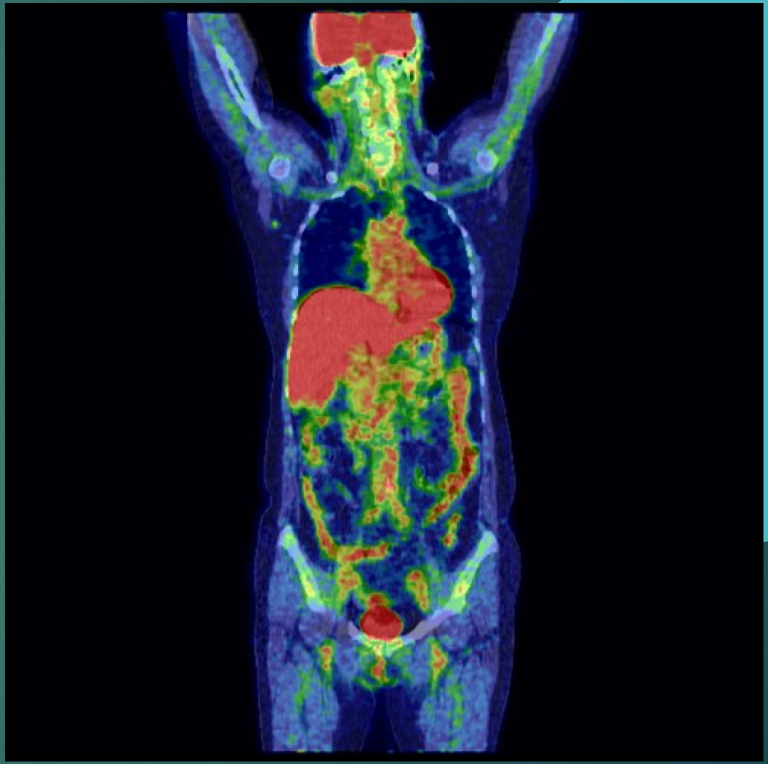


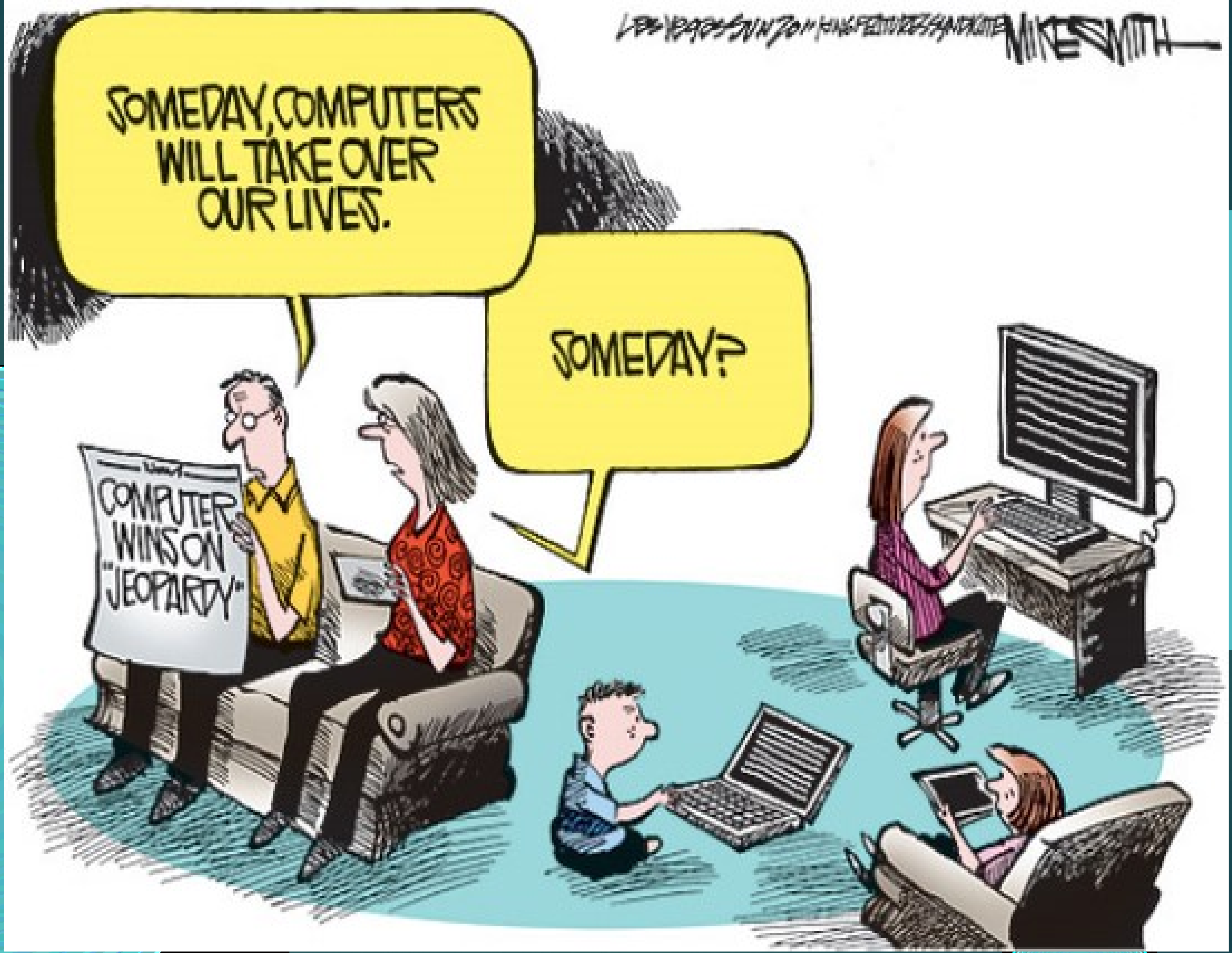
"And that's why we need a computer."

Computational models in biology and medicine








LAS VEGAS SUN 2011 KING FEATURES SYNDICATE MIKE SMITH
SOMEDAY, COMPUTERS WILL TAKE OVER OUR LIVES.
SOMEDAY?
COMPUTER WINS ON "JEOPARDY"