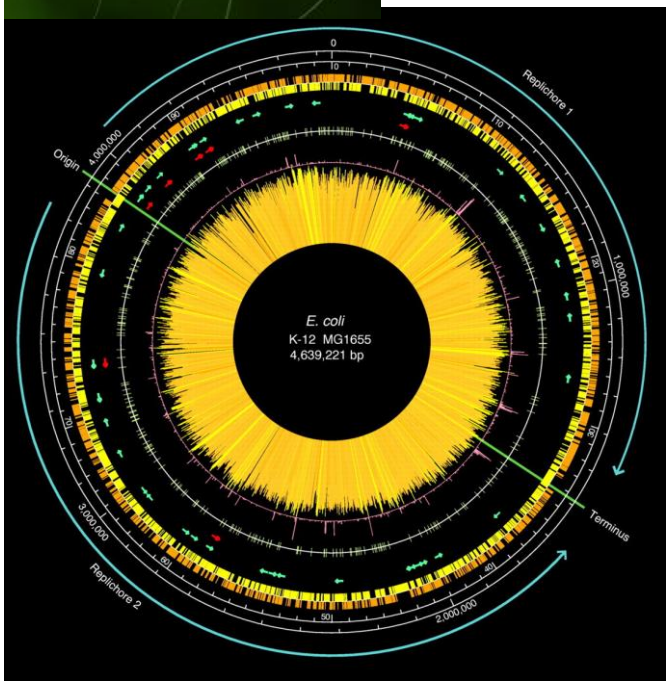
Network analysis of biological data

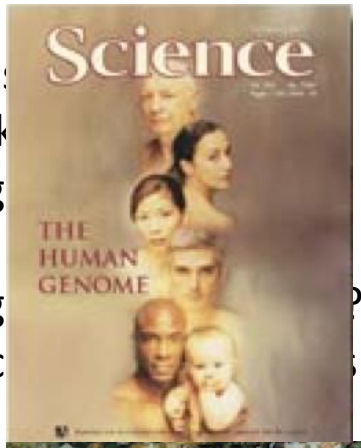
Gang Fang, PhD

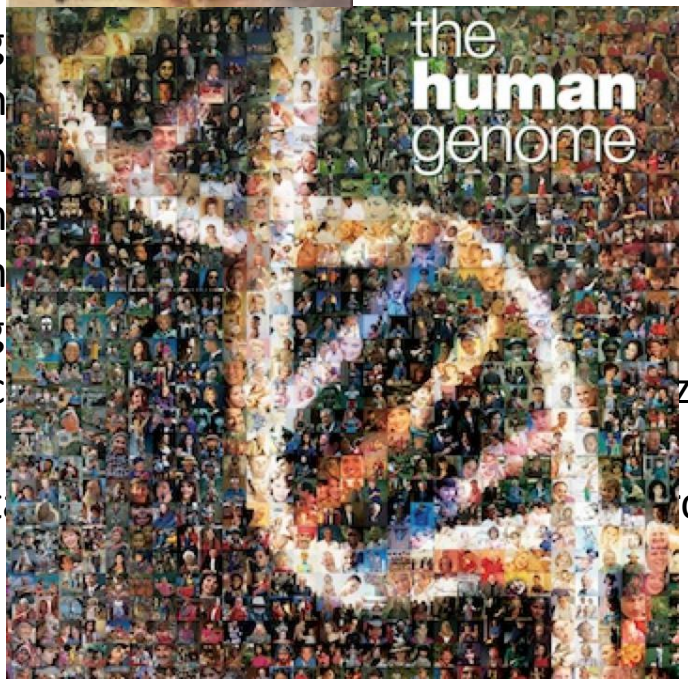
New York University Shanghai Campus

How much data is there in biology?

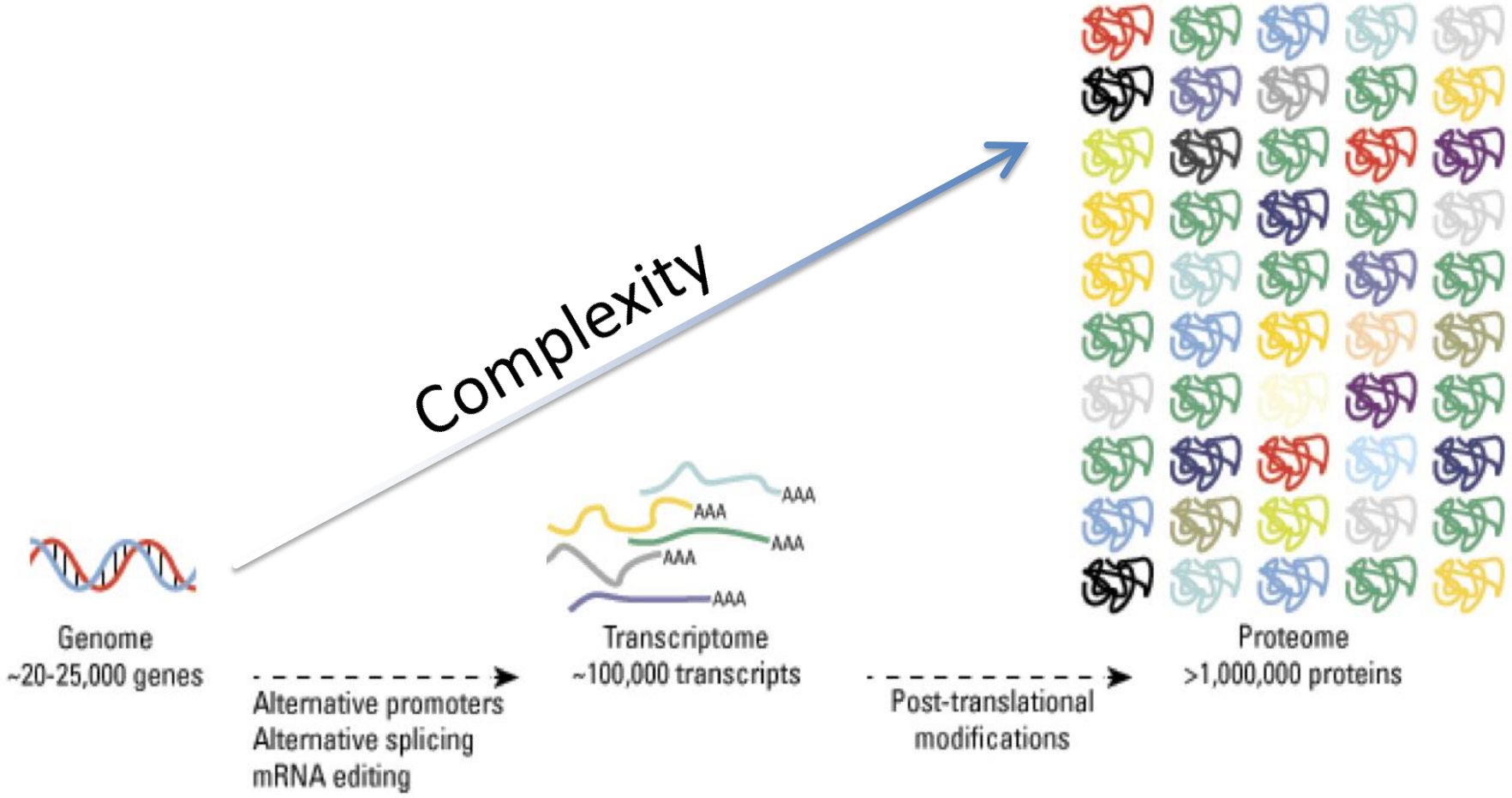


How big i
How thick
How long
…
How long ?
How muc s the cell need?
…
How long
How man
How man
How man
How man
How long
How muc ze protein?
…
Is the dat cumstances?



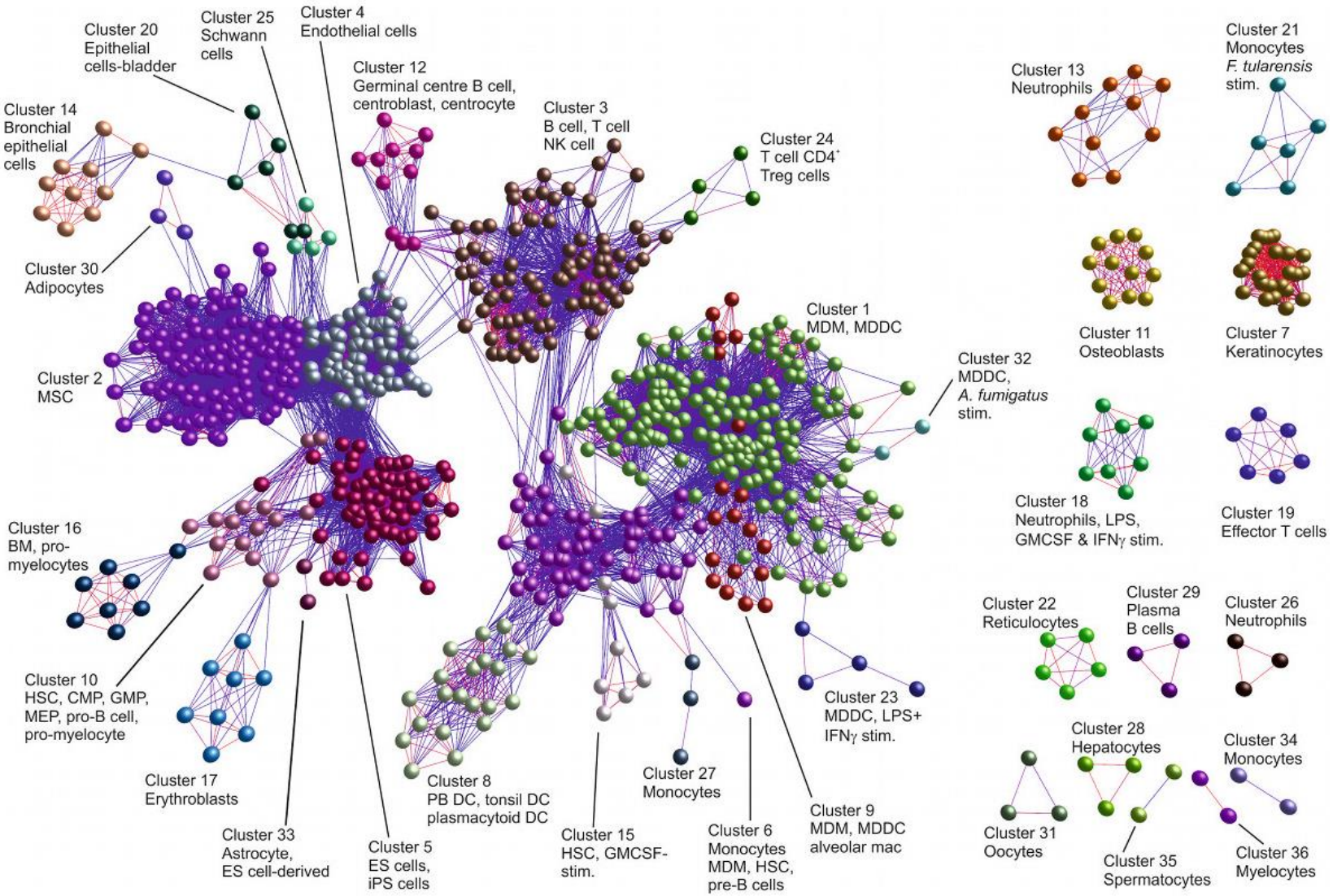


Why data matters?



"Who are they, how many, what they do, how they do it?"

What's the best way to describe biology?

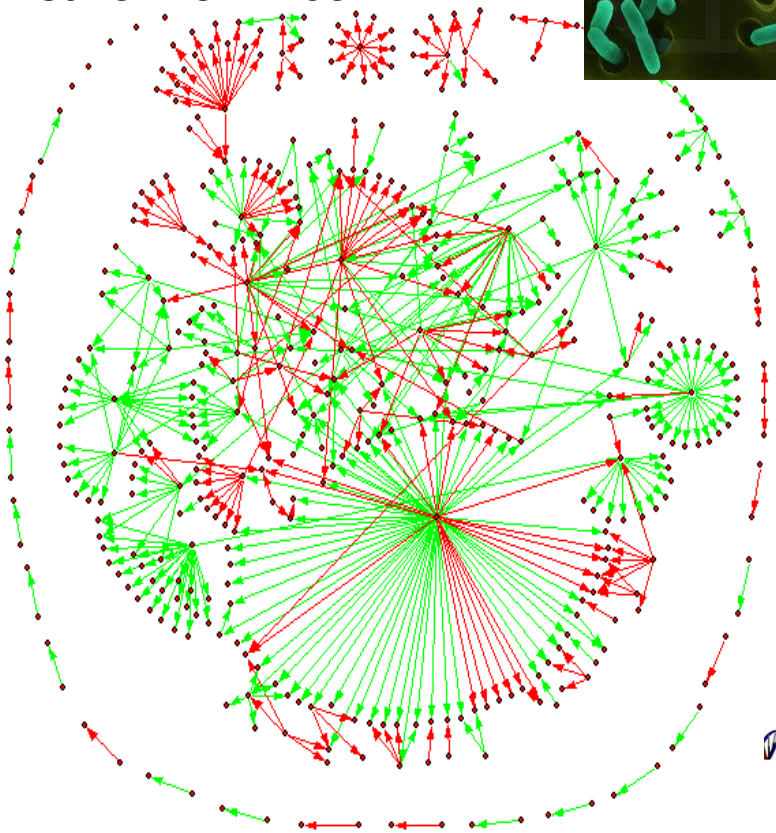


What is biological network

- Node
  Gene, protein, metabolite, any "biological object"
- Edge
  Regulation, protein-protein interaction, any kind of "similarity" or "dissimilarity" etc.
- "Weights" or features
  Conservation of gene, expression value, half time, any measurable or categorical variable.
- Network topology
  Clusters, modularity, node centrality, shortest path etc.
- Network dynamics
  Comparison of networks: time series and environmental changes
- Network rewiring and permutation
  Test your theory!

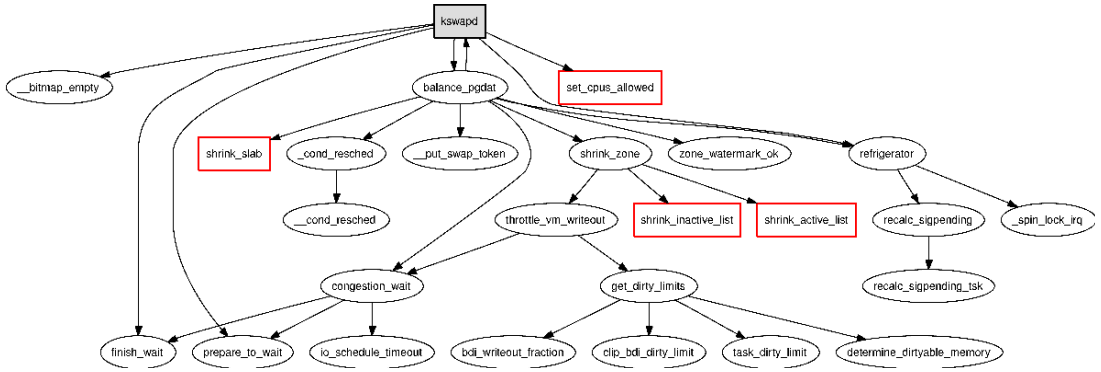
Knowledge learnt from comparing a naturally evolved biology network with a man-made network

Transcription regulatory network of *E. coli*



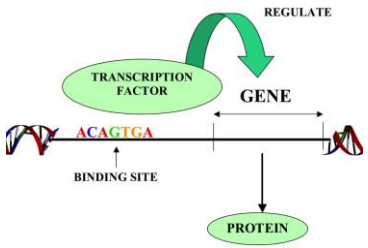
1400 nodes, 3000 edges

Call graph of the Linux kernel



From CodeViz

12000 nodes, 34000 edges

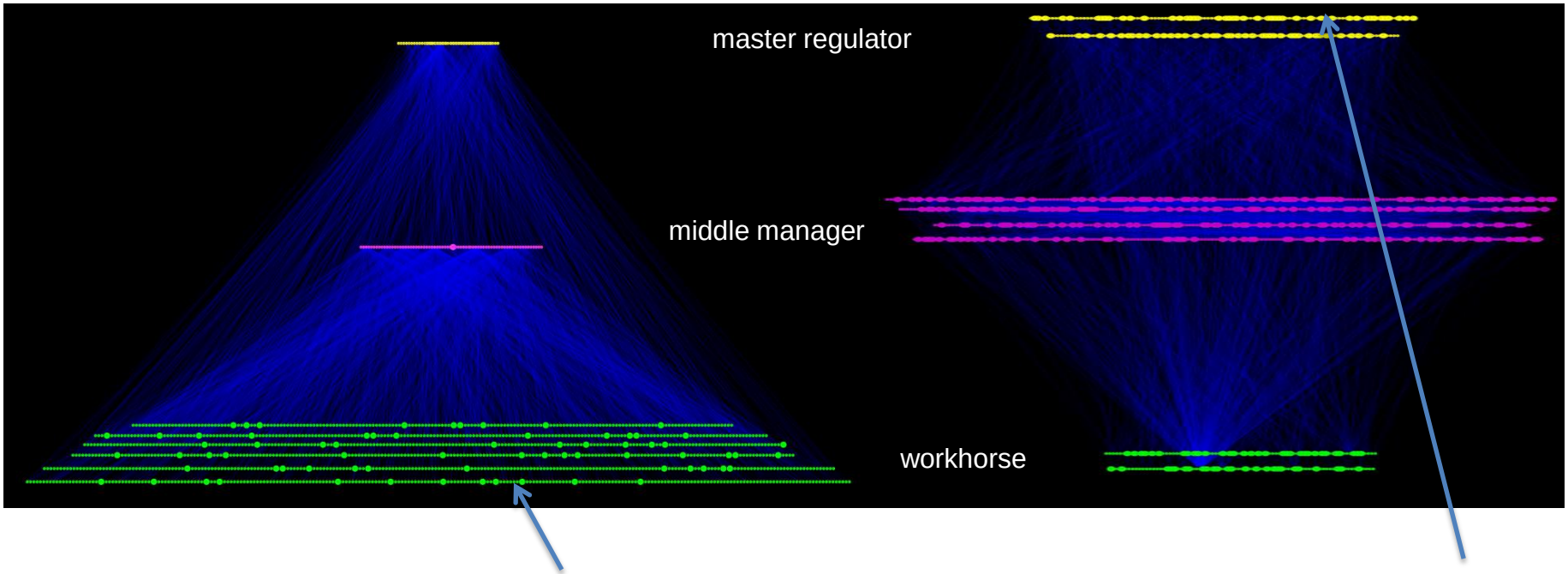


Yan KK**,** **Fang G**, Bhardwaj N, Alexander RP, Gerstein M: **Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(20):9186-9191.

Hierarchical organization: pyramidal versus top-heavy

E. coli transcriptional regulatory network

the Linux call graph



Persistent genes

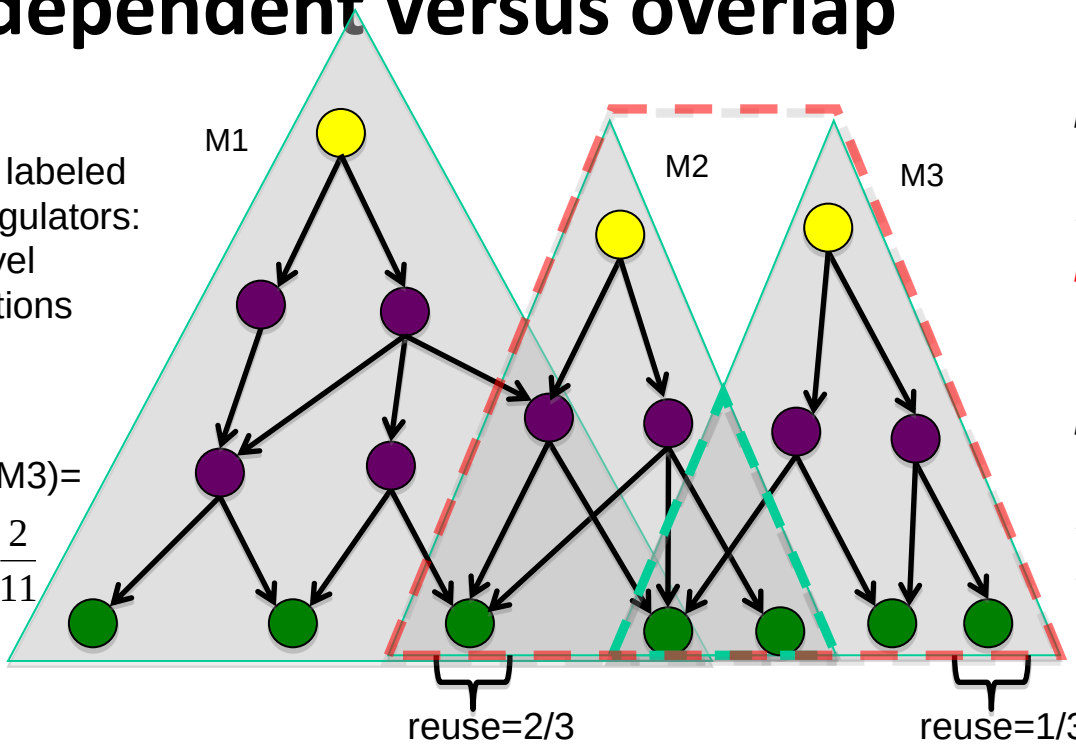
Genes subject to strong natural selective pressure

Persistent functions

Software engineers' favorite functions

Organization of Modules: independent versus overlap

Modules are labeled by master regulators: TFs, high-level starting functions



Overlap(M2,M3)= $\frac{|M2 \cap M3|}{|M2 \cup M3|} = \frac{2}{11}$

We observe opposite correlation behaviors in the two systems: Reuse and persistence are negatively correlated in the E. coli regulatory network but positively correlated in the Linux call graph.

[Spearman correlation r=−0.074 (P < 0.01) and r=0.10 (P < 10−4), respectively]

	E. Coli TRN	Linux call graph
Average overlap	4.3%	80.7%
Maximum node reuse	15.6%	87.5%
Average node reuse	3.5%	8.4%

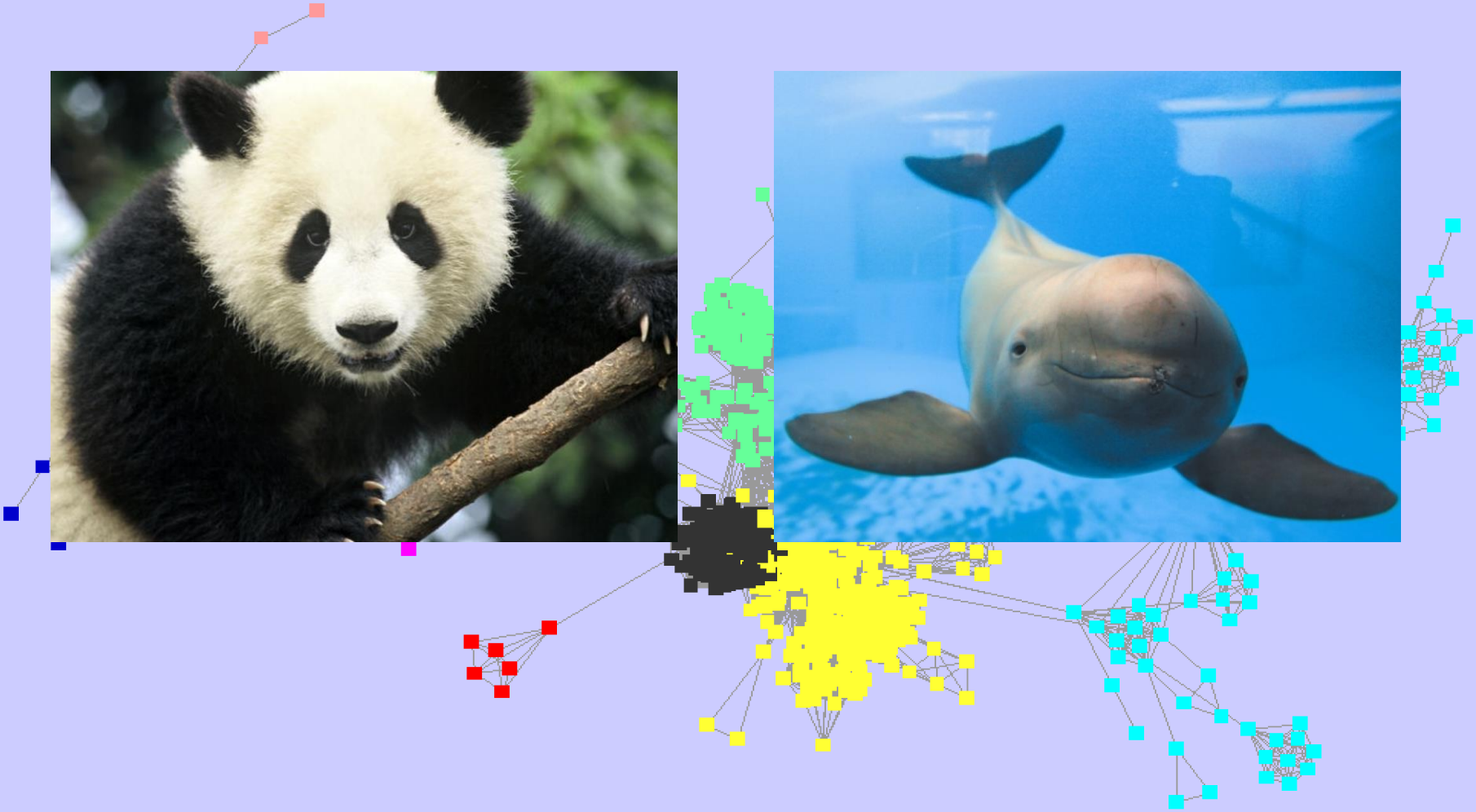
TRN: modules overlap little, components are less generic

Call graph: modules overlap, Functions are highly reused (generic): “printk”

We are more “robust” than computers!

	Biological network	Computer OS network
Modularity	High *Persistent genes are workhorse* *Low module overlap*	Low *Persistent modules are masters* *High module overlap*
Node Reuse	Low	High
Robustness	High	Low
Efficiency	Low *Billions of years*	High *20 years*

Can we use network analysis to identify protein “living fossils”?



Thank you!