

Universal laws and architecture 4:

Layering, learning, and decentralized control

John Doyle

John G Braun Professor

Control and Dynamical Systems, EE, BE

Caltech

Outline: Laws and architectures

- Motivating case studies
 - Brains
 - Computers, networks
 - Cells
 - Physiology
- Layered architecture of the cell
 - replication, transcription, translation
 - metabolism, signaling, chemotaxis
 - 2CST and cross layer control

Compute

Turing

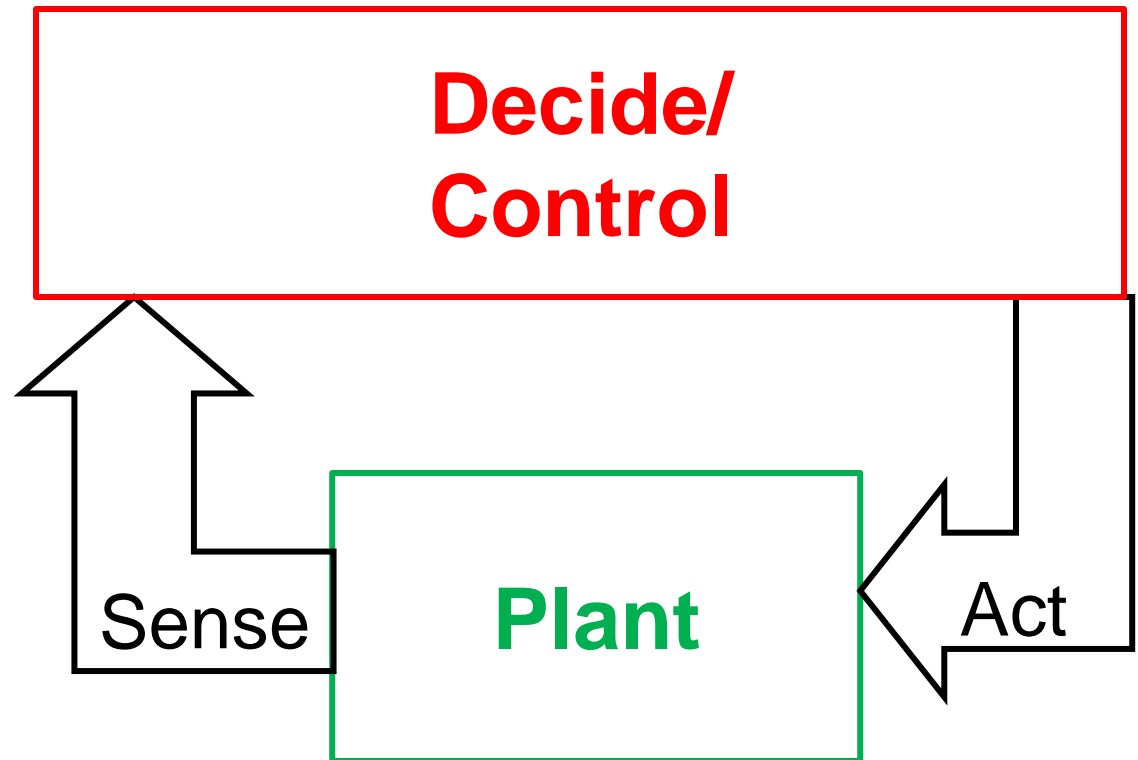
**Delay is
most
important**

Bode

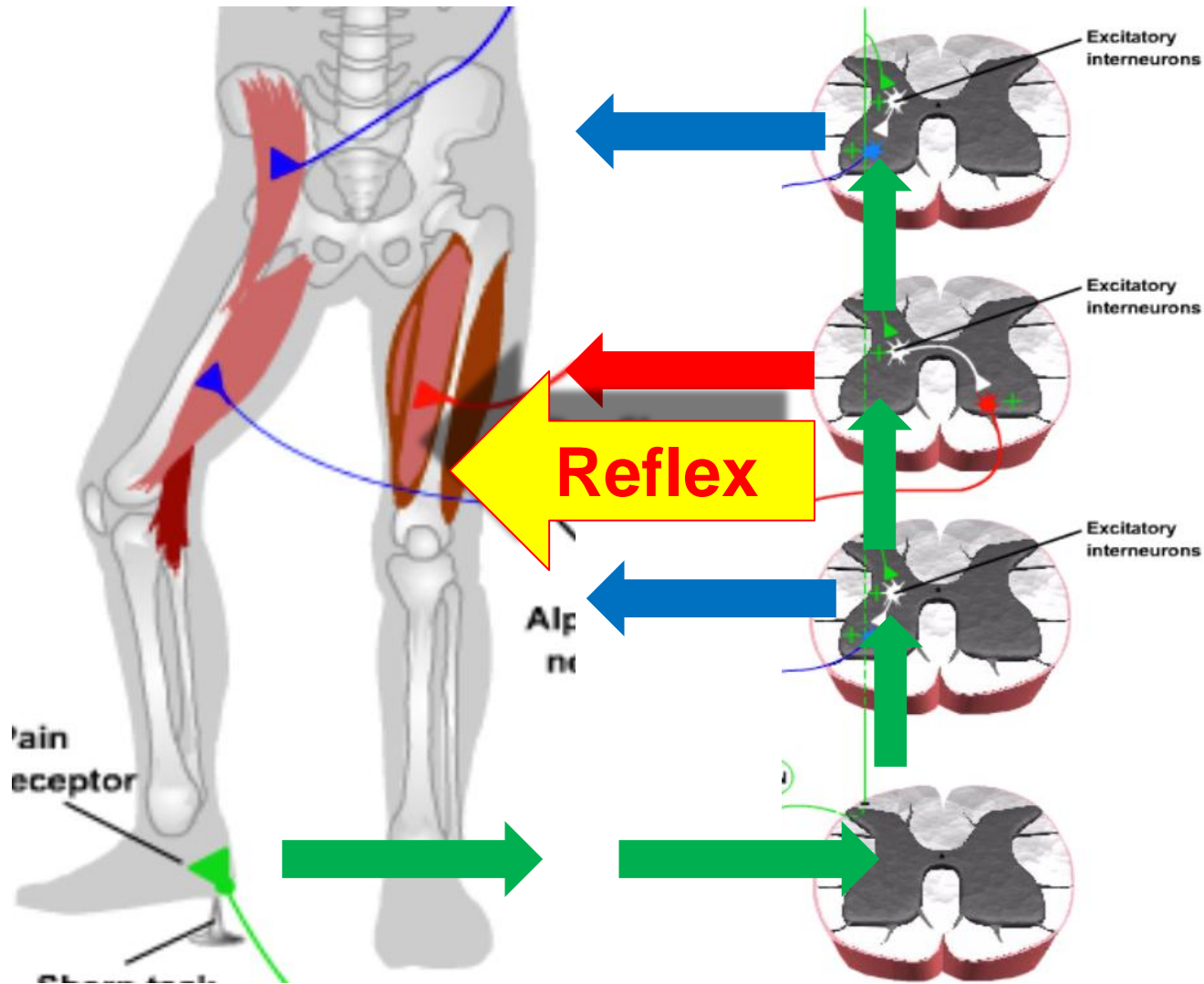
Control

Closing the loop

- On the “plant”
- On the “story”



Neuro motivation



**Fast
Inflexible**

**Slow
Flexible**

Prefrontal

Learning

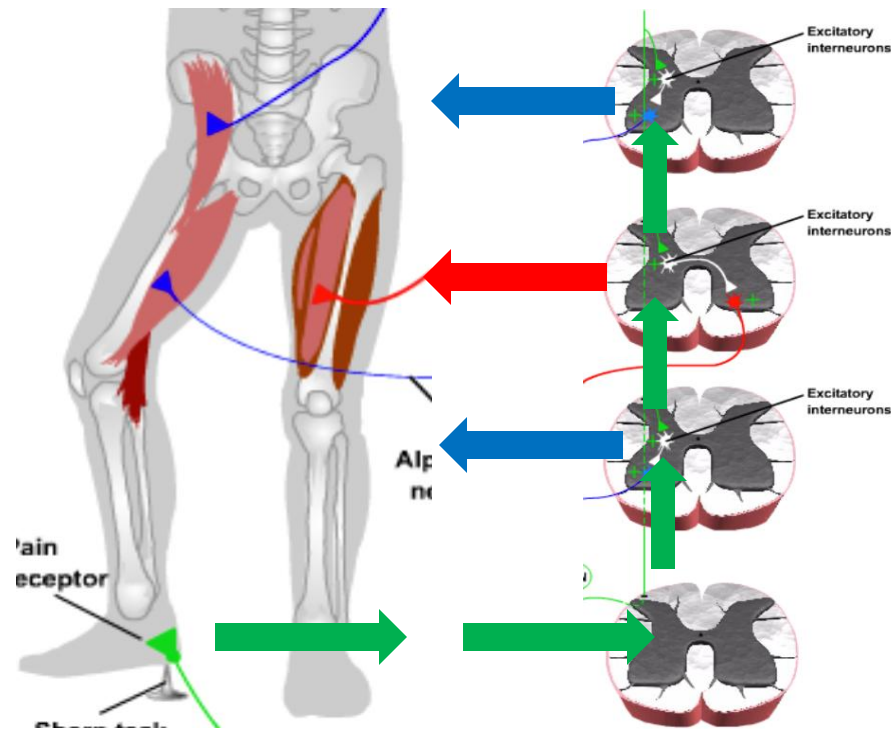
Motor

Sensory

Striatum

Ashby & Crossley

- **Acquire**
- Translate/
integrate
- Automate



Thanks to
Bassett & Grafton

**Slow
Flexible**

Prefrontal

Motor

Sensory

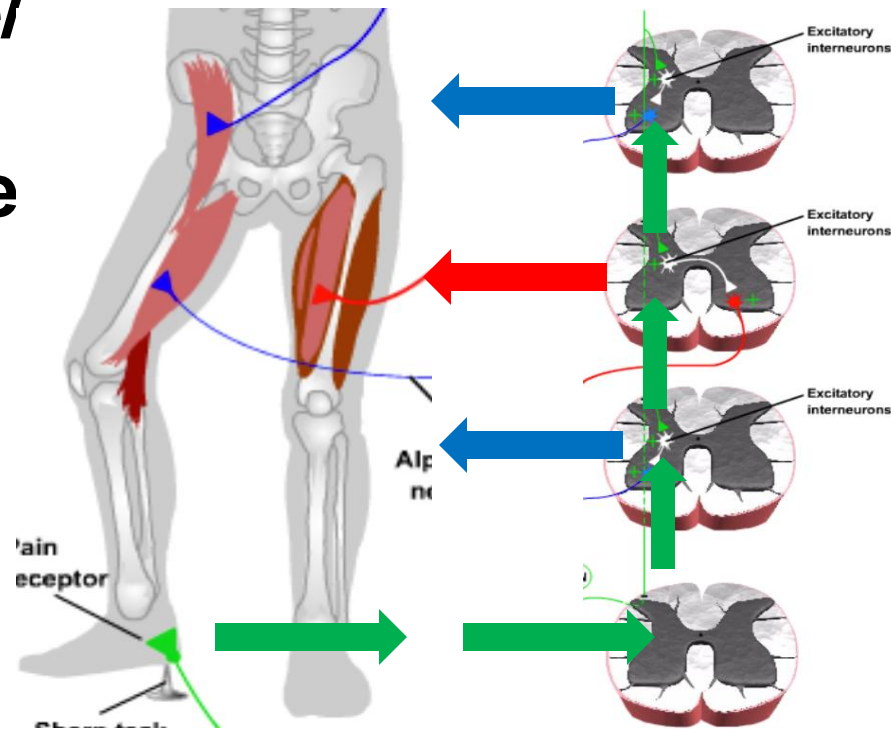
Striatum

Learning

**Fast
Inflexible**

Ashby & Crossley

- Acquire
- **Translate/
integrate**
- Automate

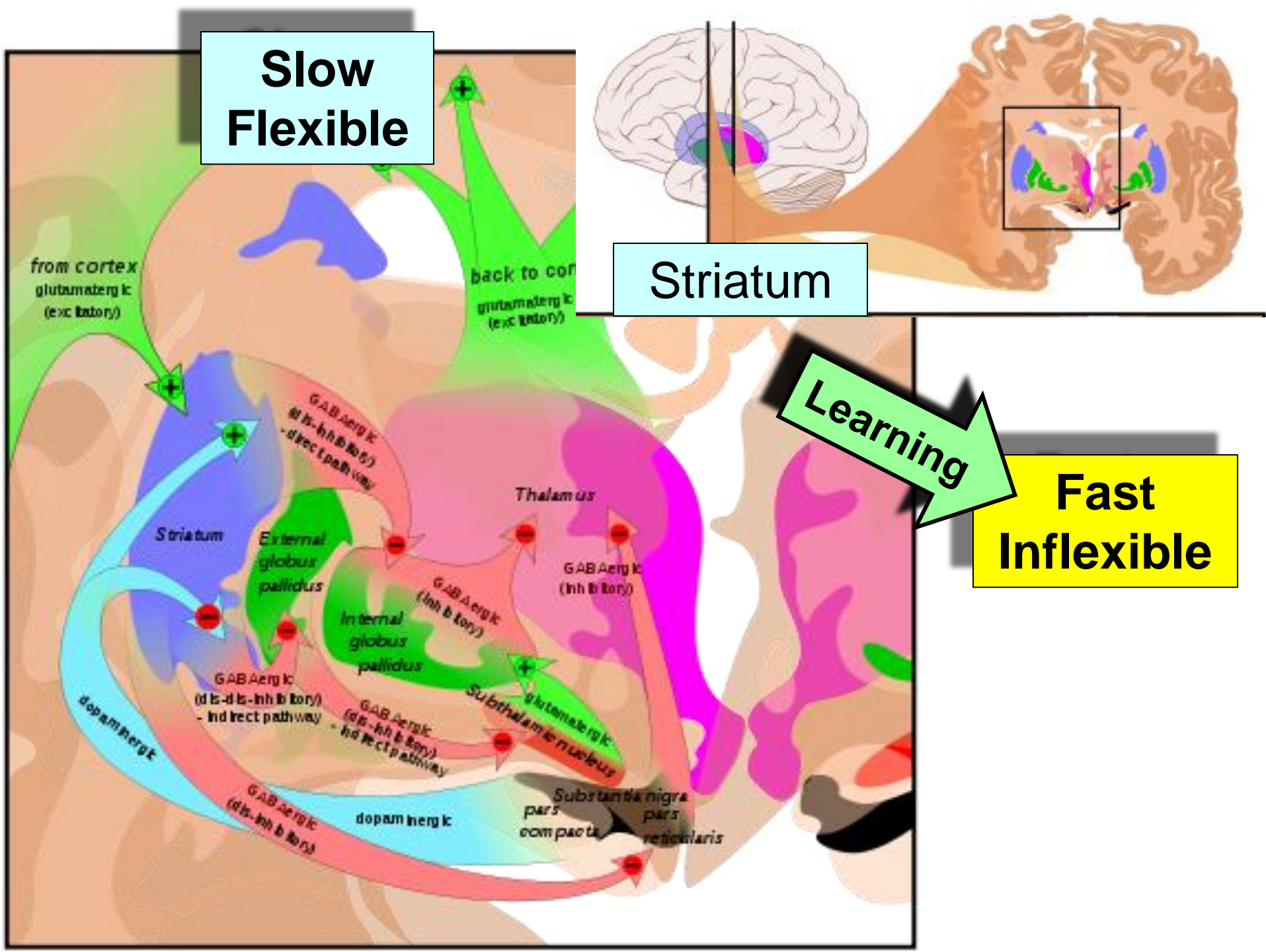


**Slow
Flexible**

Striatum

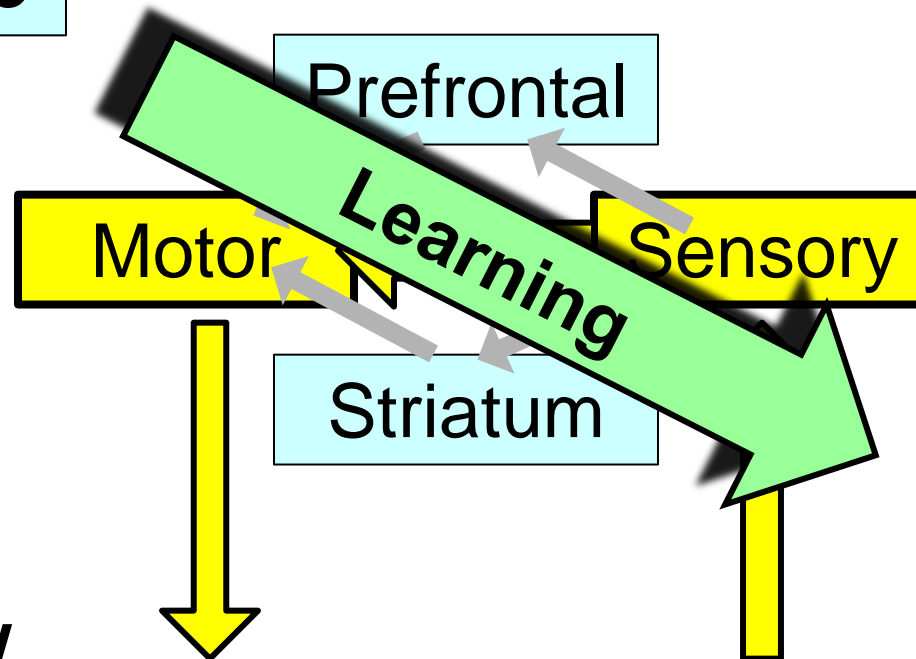
Learning

**Fast
Inflexible**



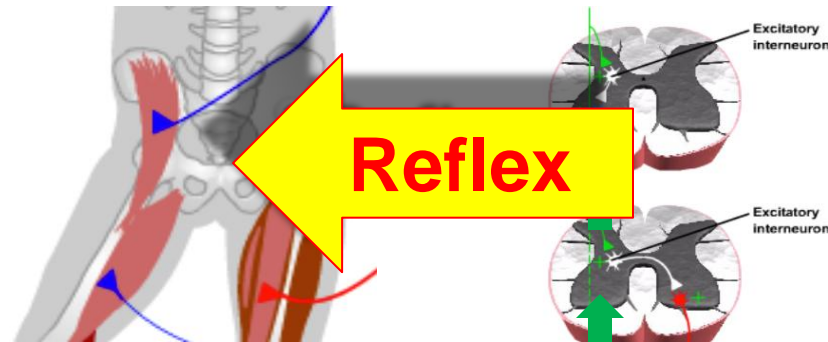
Build on Turing to show what is *necessary* to make this work.

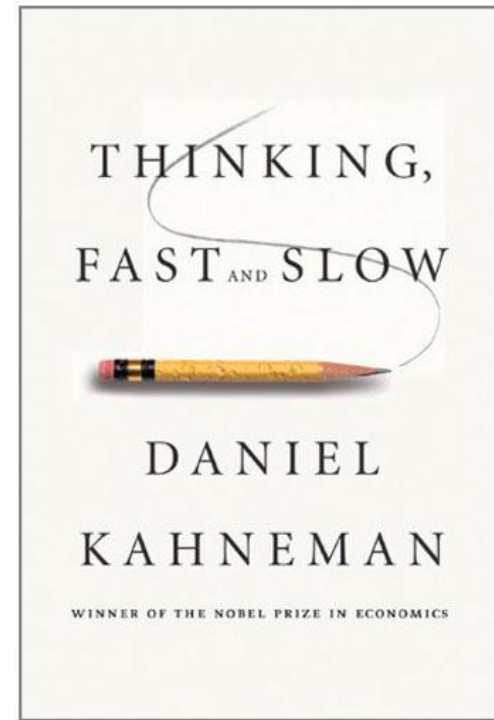
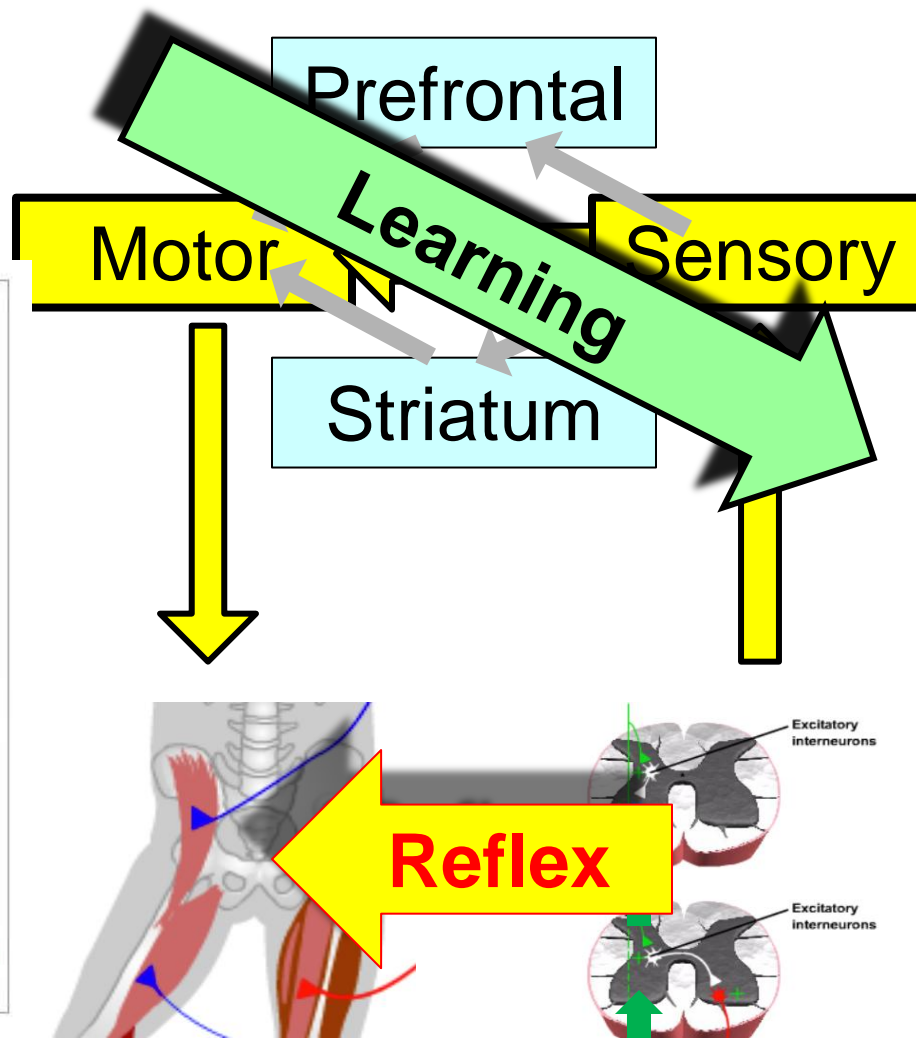
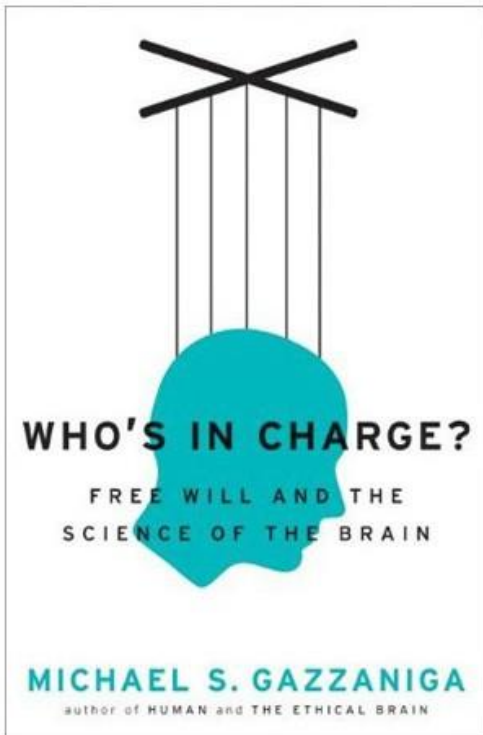
**Slow
Flexible**

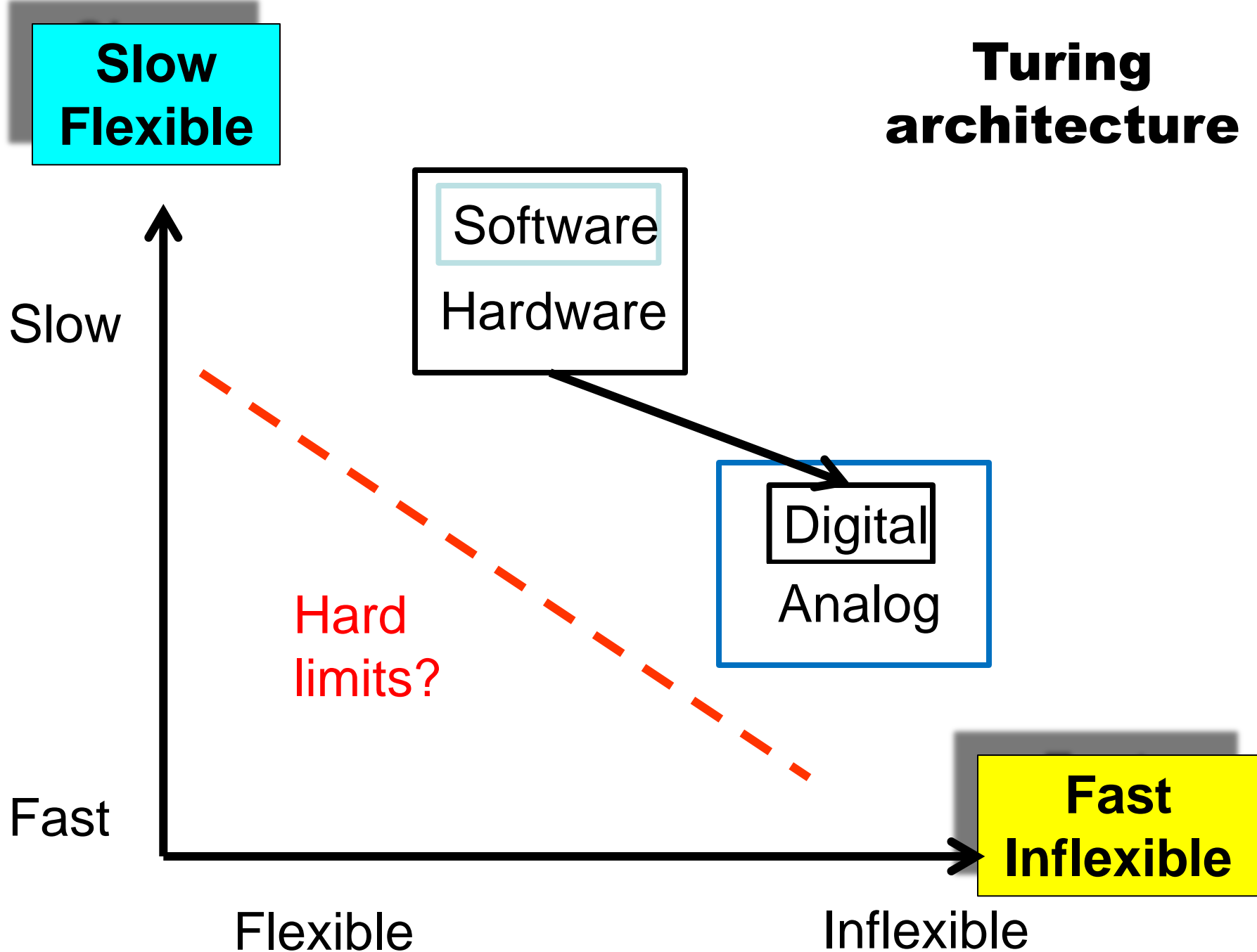


- Acquire
- Translate/
integrate
- Automate

**Fast
Inflexible**





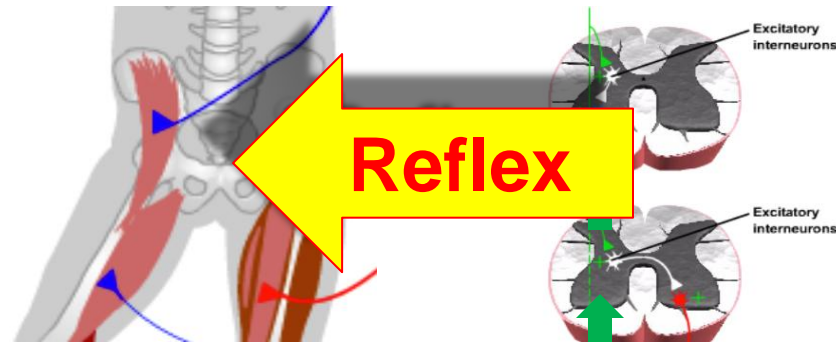


Wolpert, Grafton, etc

robust

Brain as ~~optimal~~ controller

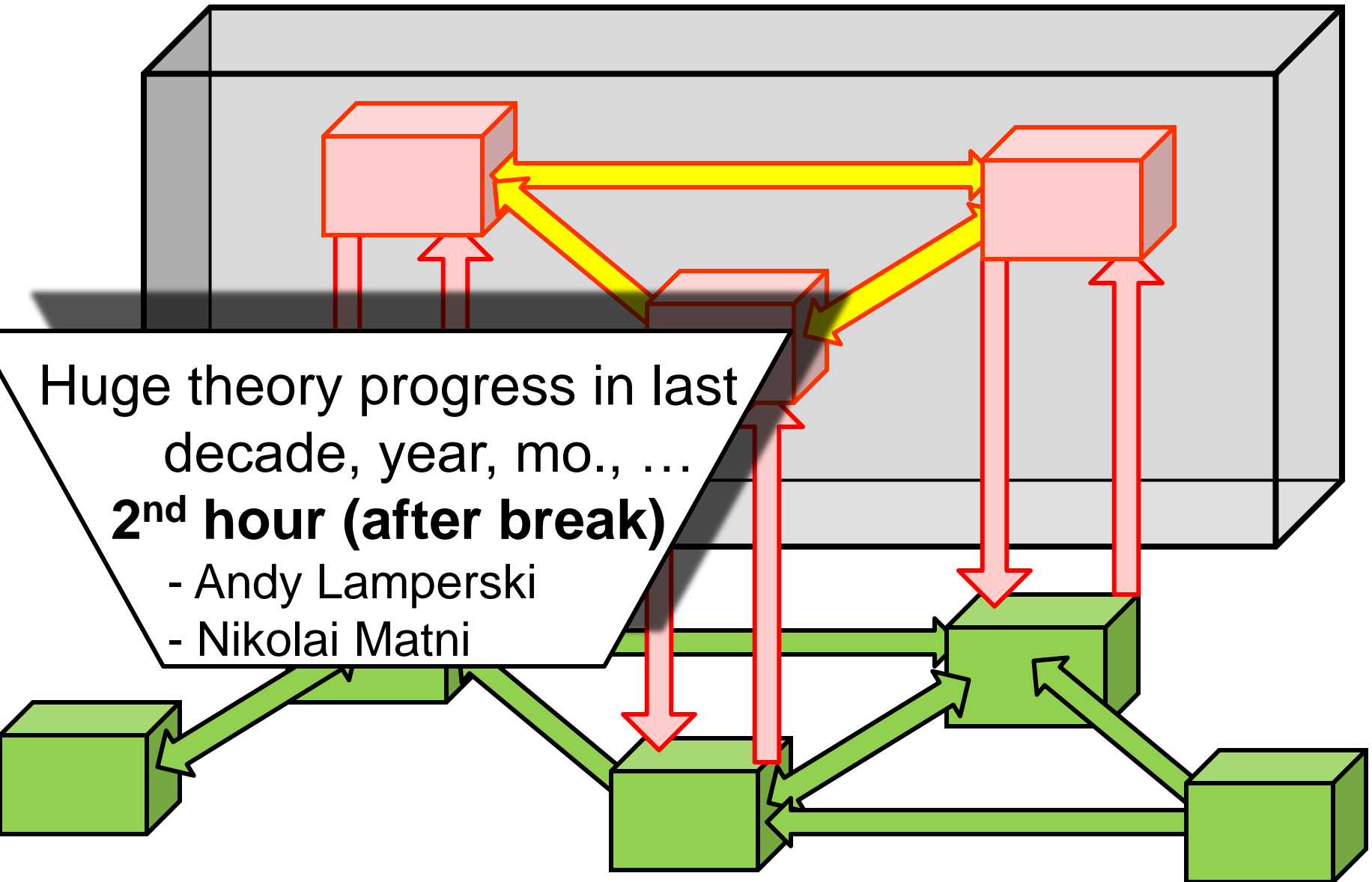
- Acquire
- Translate/
integrate
- **Automate**



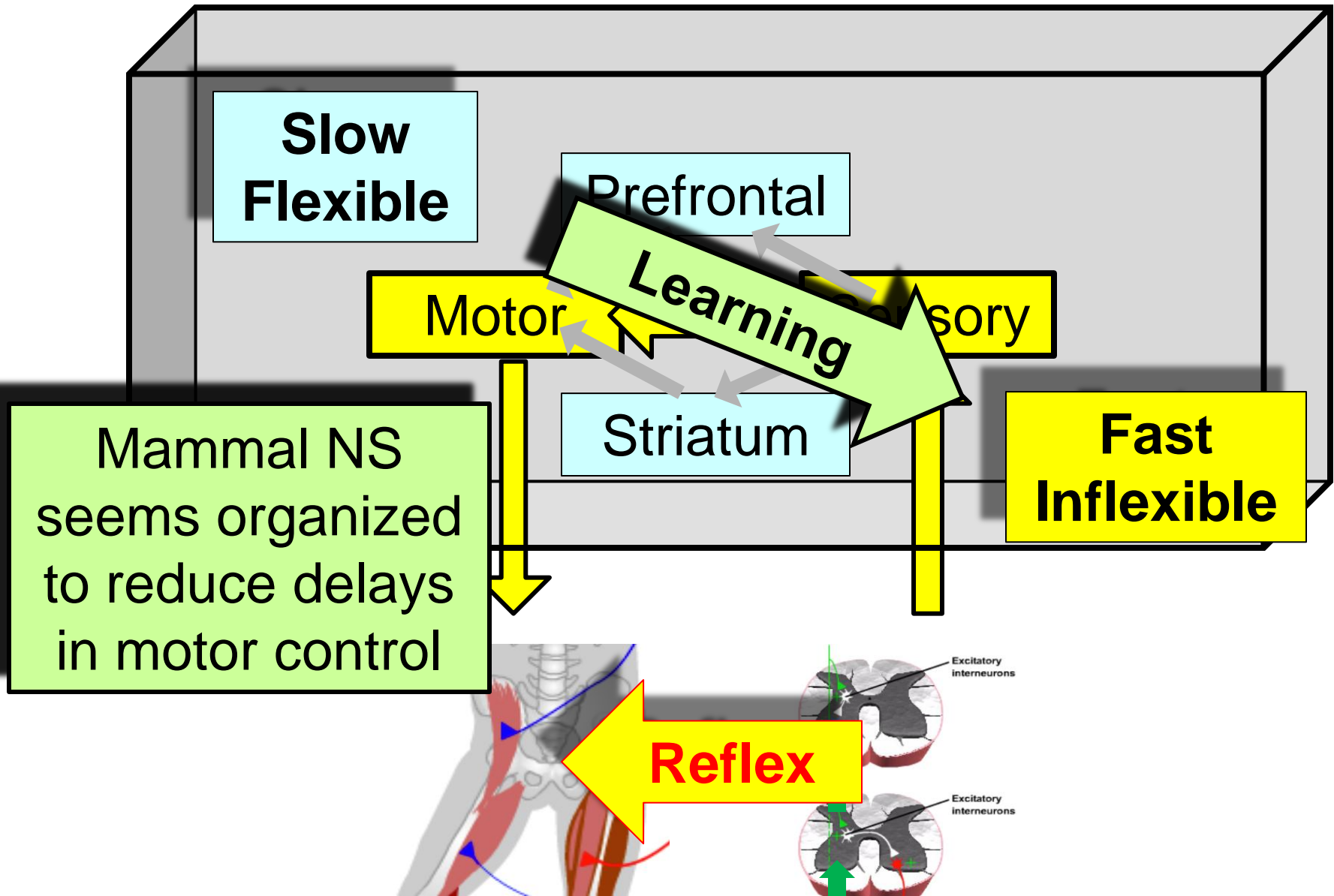
What I'm not going to talk about

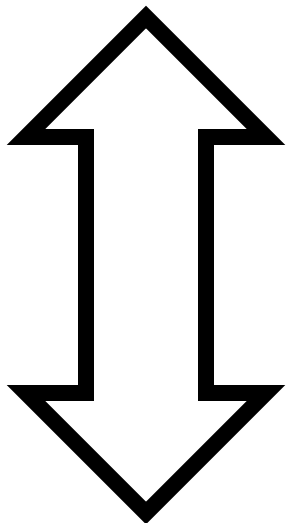
- Connections between robustness and risk sensitivity
- Asymmetry between false positives and negatives
- Risk aversion and risk seeking
- Uncertainty is more in models than in probabilities
- Life is not like a casino

Going beyond black box: control is decentralized with internal delays.

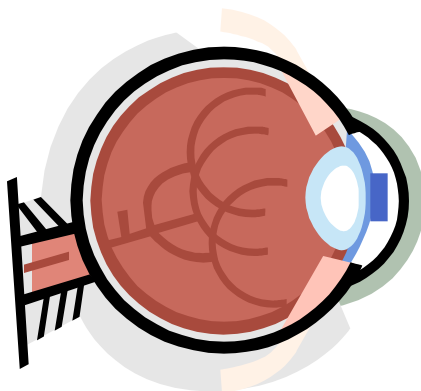
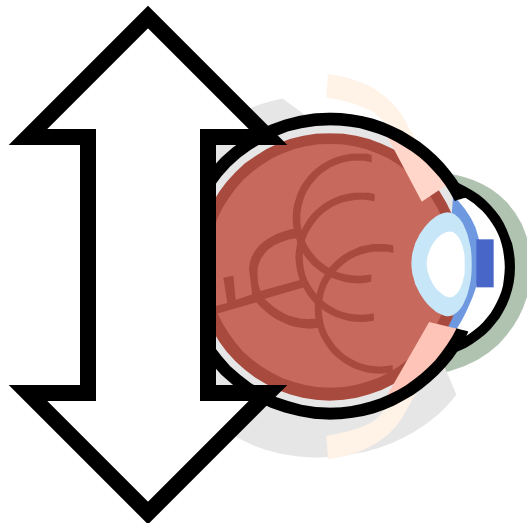


Going beyond black box: control is decentralized with internal delays.

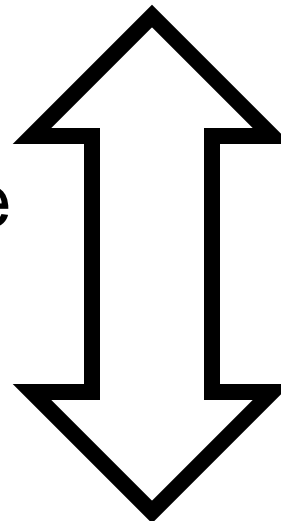




Move
head



Move
hand



Bigger error

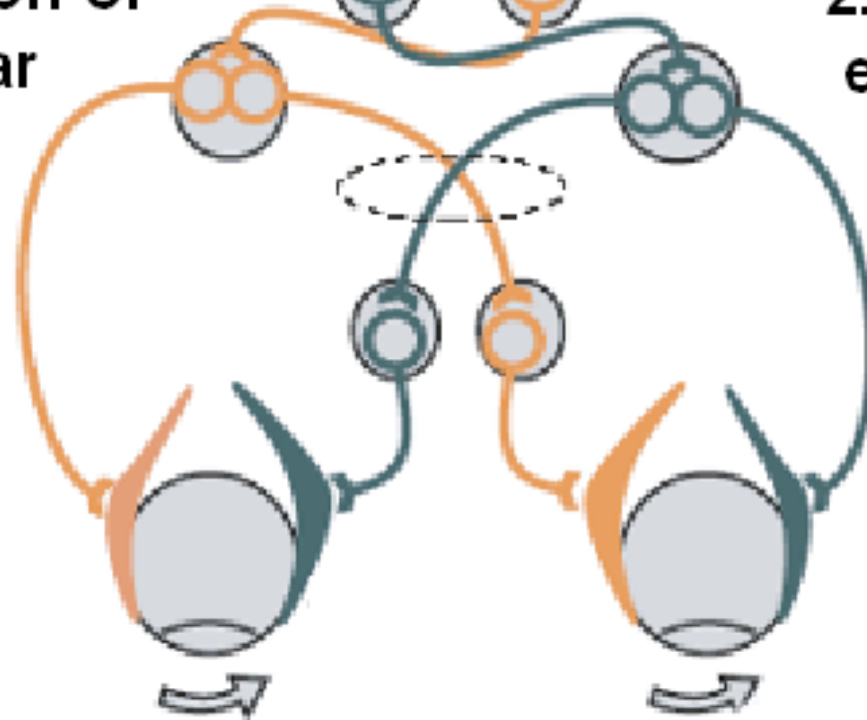
Vestibulo-ocular reflex

1. Detection of rotation

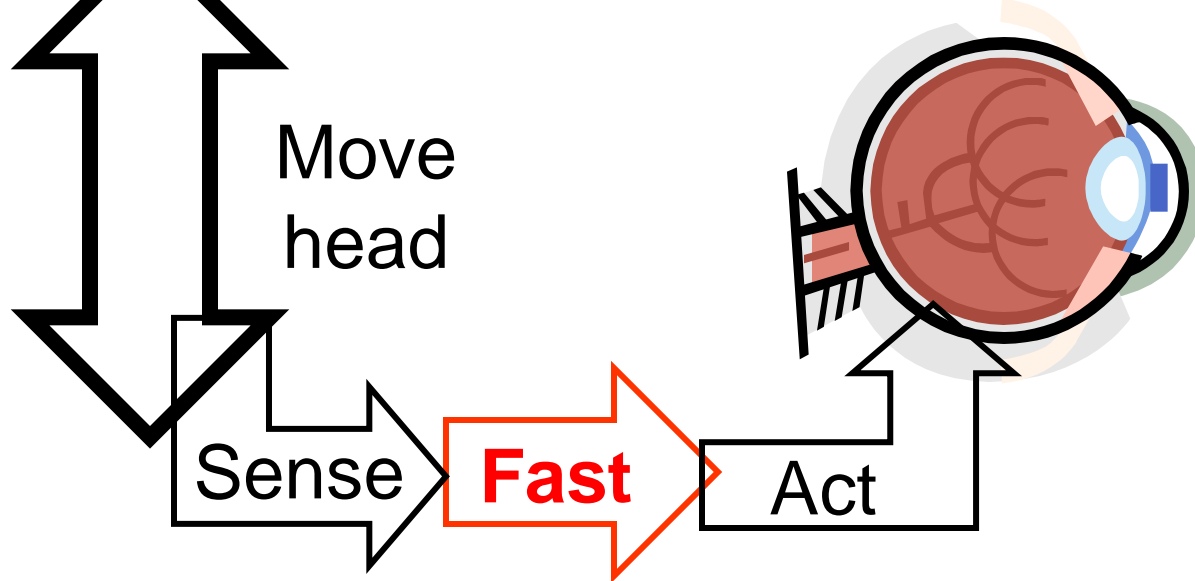


2. Inhibition of extraocular muscles on one side.

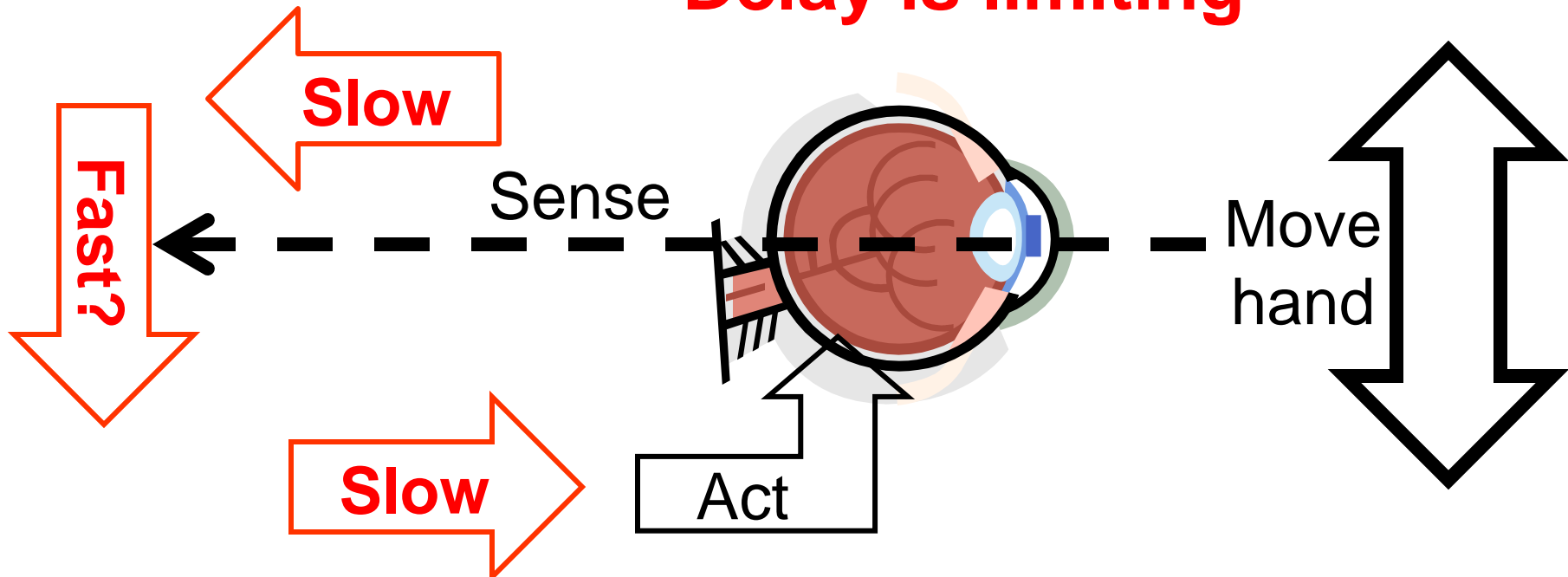
2. Excitation of extraocular muscles on the other side



3. Compensating eye movement

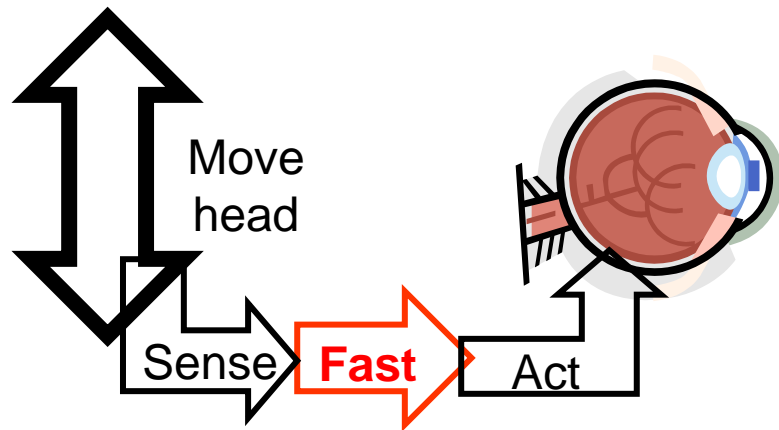


Same actuators
Delay is limiting

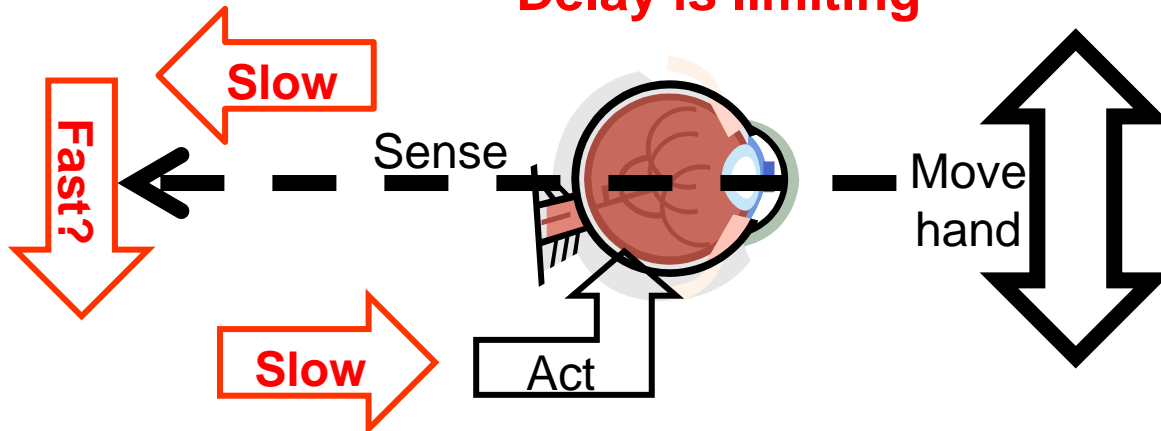


Versus standing on one leg

- Eyes open vs closed
- Contrast
 - young surfers
 - old football players



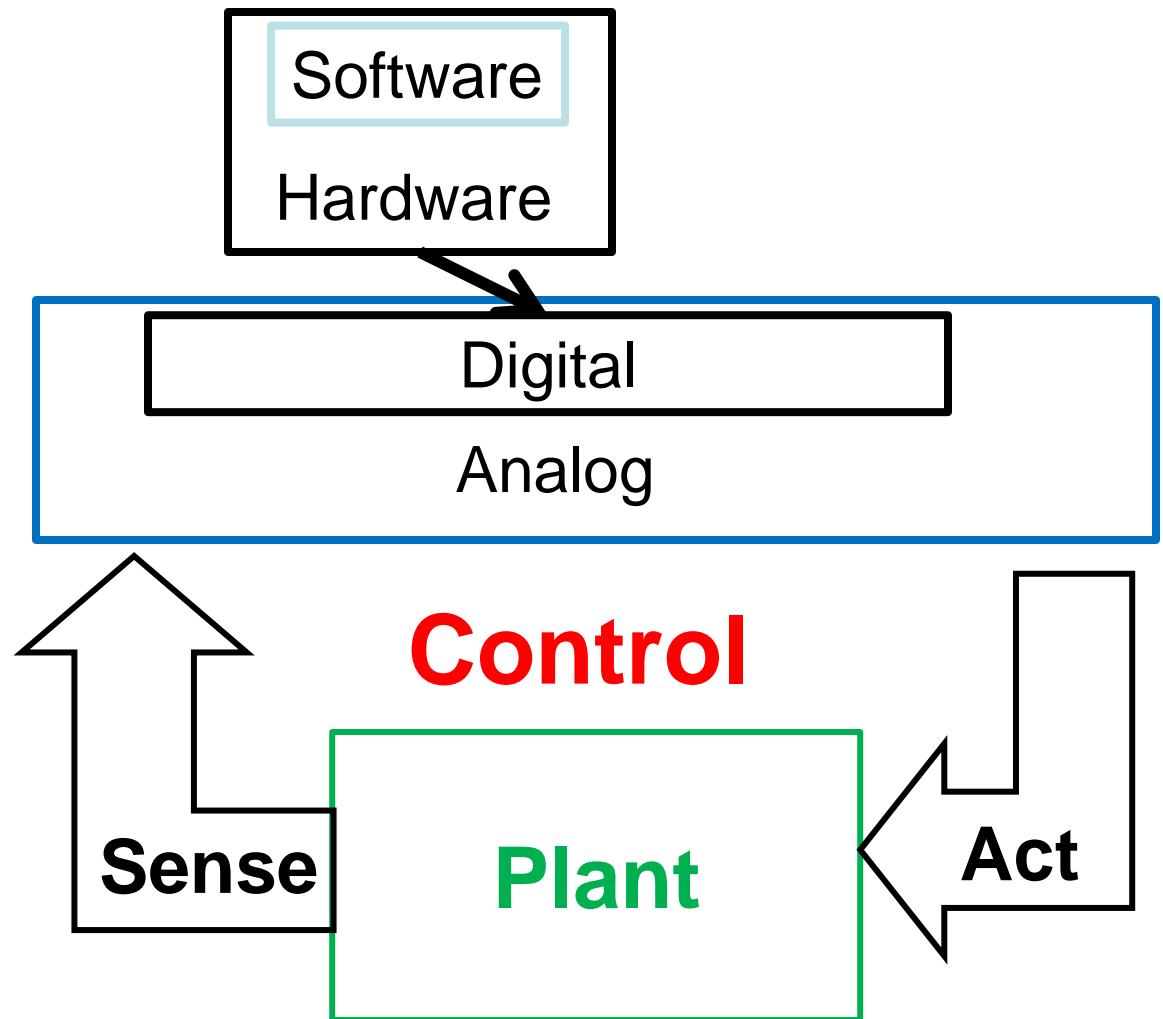
Same actuators
Delay is limiting



Compute

- Computational complexity of
- ***Designing*** control algorithms
 - ***Implementing*** control algorithms

Delay is
even more
important
in control



Control

Issues for neuroscience

- Brains and UTMs?
 - Time is most critical resource?
 - Space (memory) almost free?
- Read/write random access memory hierarchies?
- Brain >> UTM?

Conjecture

- Memory potential $\approx \infty$
- Examples
 - Insects
 - Scrub jays
 - Autistic Savants

Gallistel and King

C.R. Gallistel and
Adam Philip King

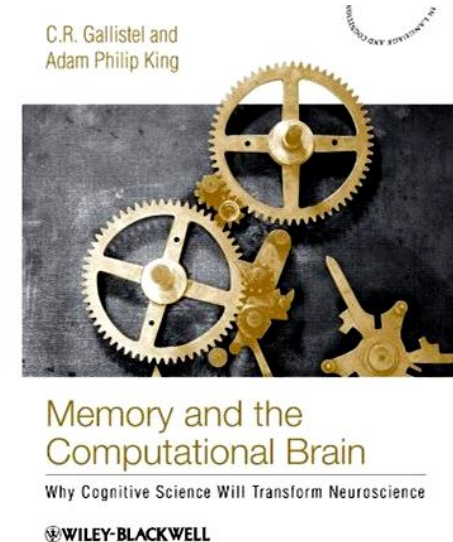
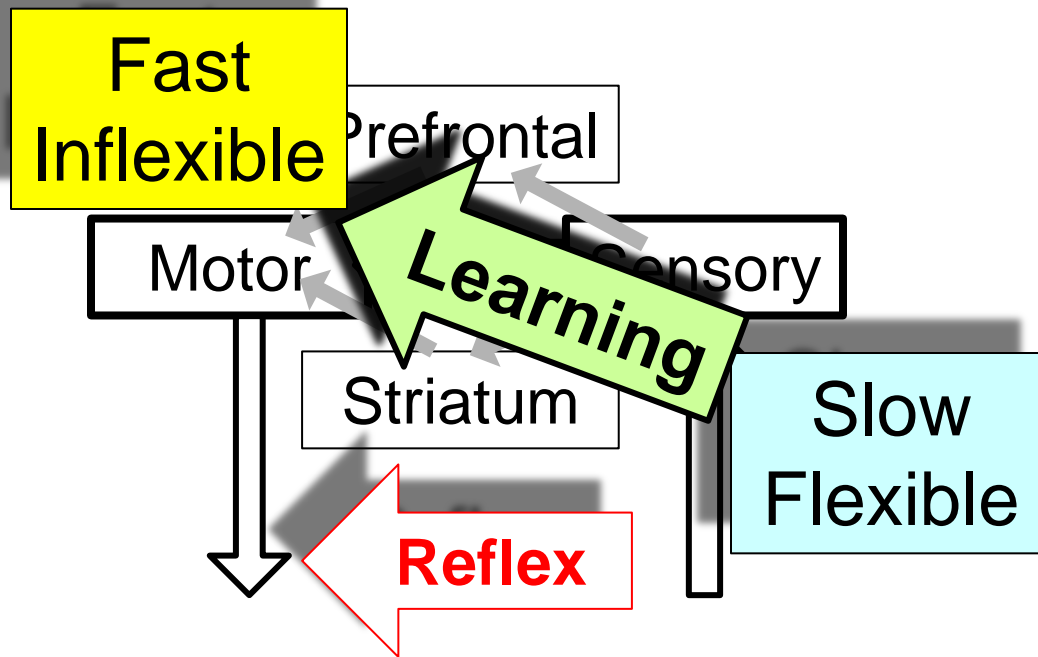


Memory and the Computational Brain

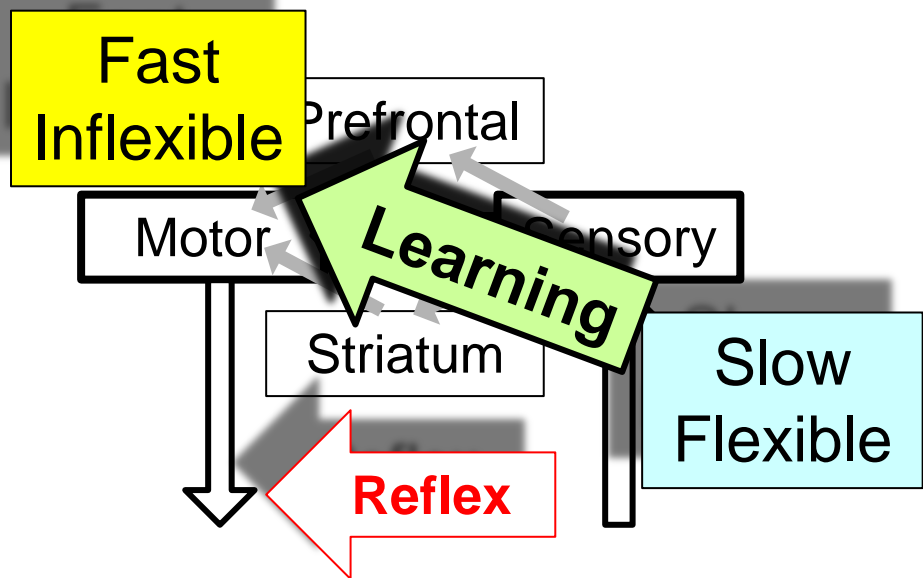
Why Cognitive Science Will Transform Neuroscience

WILEY-BLACKWELL

Gallistel and King



- Sensori-motor memory potential $\approx \infty$ (Ashby)
- Limits are on **speed** of
 - nerve propagation delays
 - learning
- But control is **never** centralized
- Is there a random access read/write memory?



The EvoPsycho
question: why?

- Sensori-motor memory potential $\approx \infty$ (Ashby)
- Limits are on **speed** of
 - nerve propagation delays (fish parts?)
 - learning ???
- I'm probably confused
- What about robust learning

**Flexible/
Adaptable/
Evolvable**

**Horizontal
Meme
Transfer**

Software

Hardware

**Horizontal
App
Transfer**

Digital
Analog

**Depends
crucially on
layered
architecture**

DNAp

Gene

Repl

D

RNAp

xRNA

transc

RN

ATP

A

AA

transl

AA

**Horizontal
Gene
Transfer**

Nucl.
AA

ATP

Precursors

Catabolism

frontal

Learning

Sensory

Striatum

Reflex

Ribosc

What I'm not going to talk about

- It's true that most “really smart scientists” think almost everything in this talk is nonsense
- Why they think this
- Why they are wrong

- Time (not space) is our problem, as usual
- Don't have enough time for what is true, so have to limit discussion of what isn't
- No one ever changes a made up mind (almost)

What I'm not going to talk about

Compute

Turing

Delay is
most
important

Bode

Control, OR

Communicate

Shannon

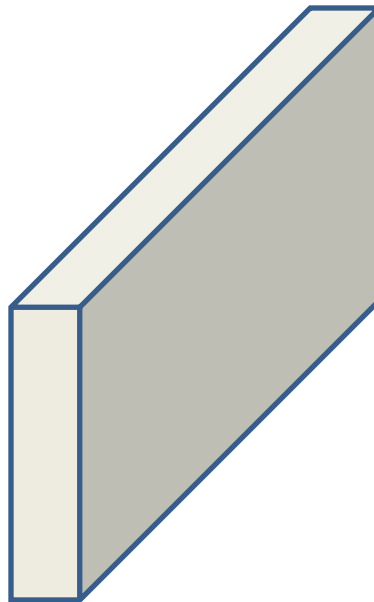
Delay is
~~*least*~~
important

Carnot

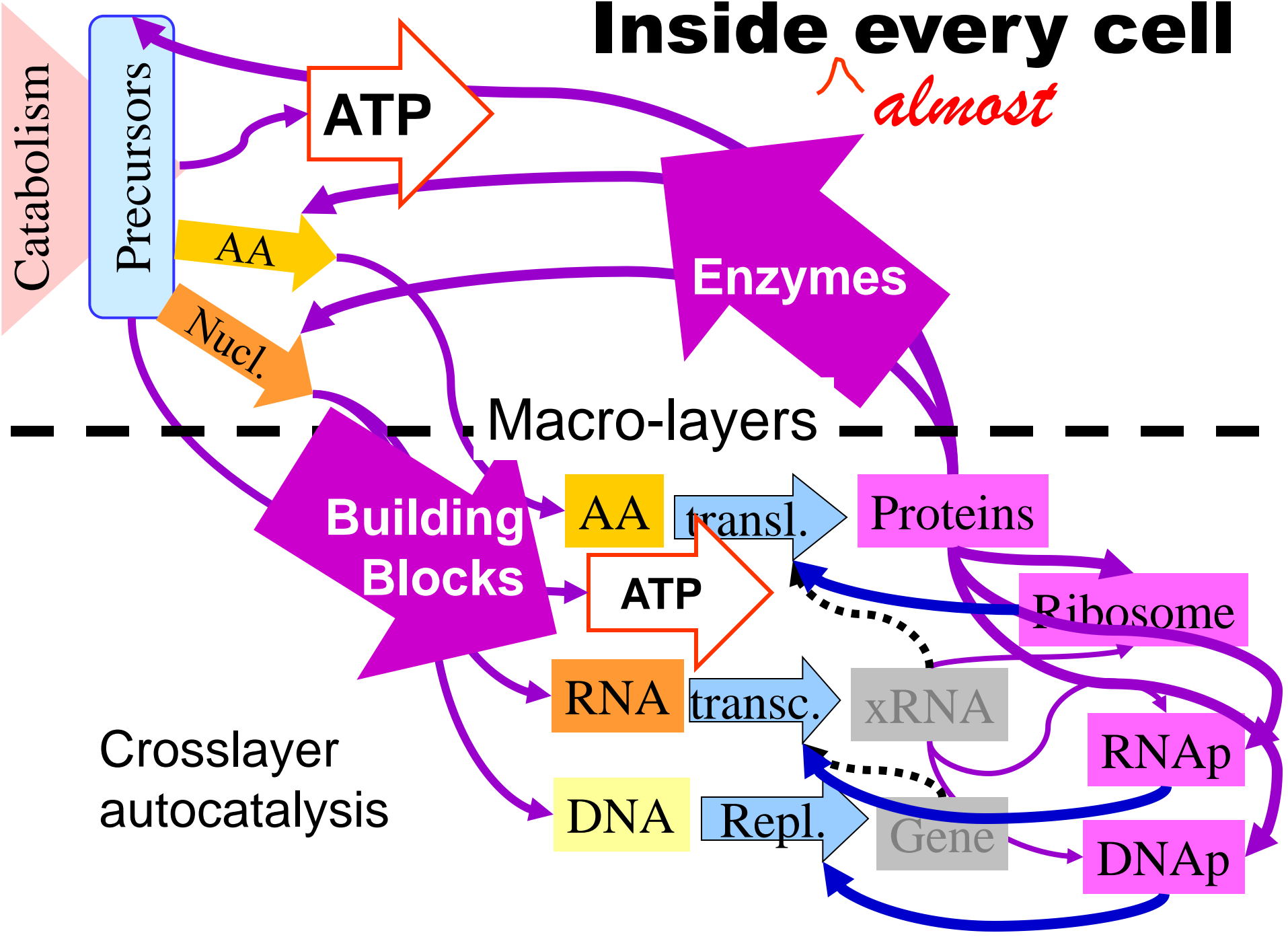
Boltzmann

Heisenberg

Physics



Inside every cell

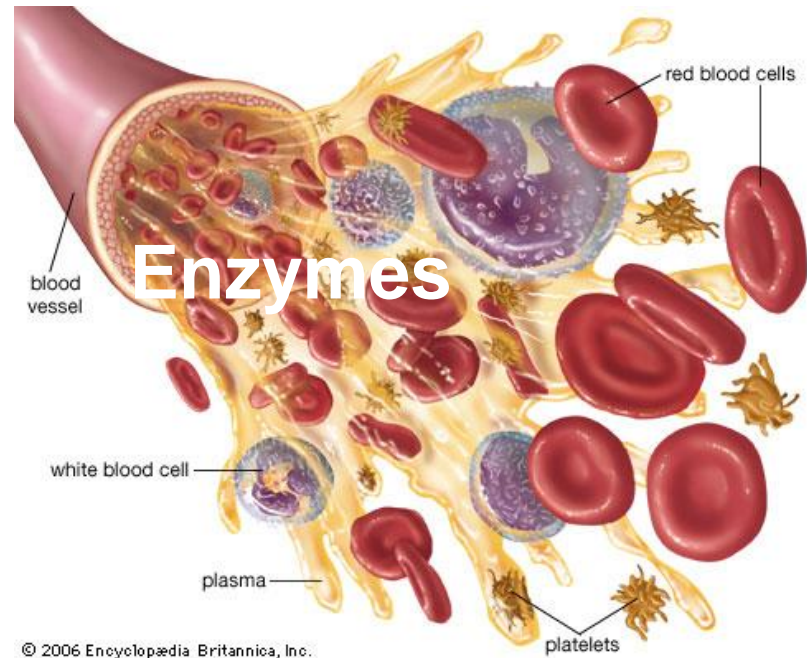


Catabolism

Precursors

ATP

Inside every cell

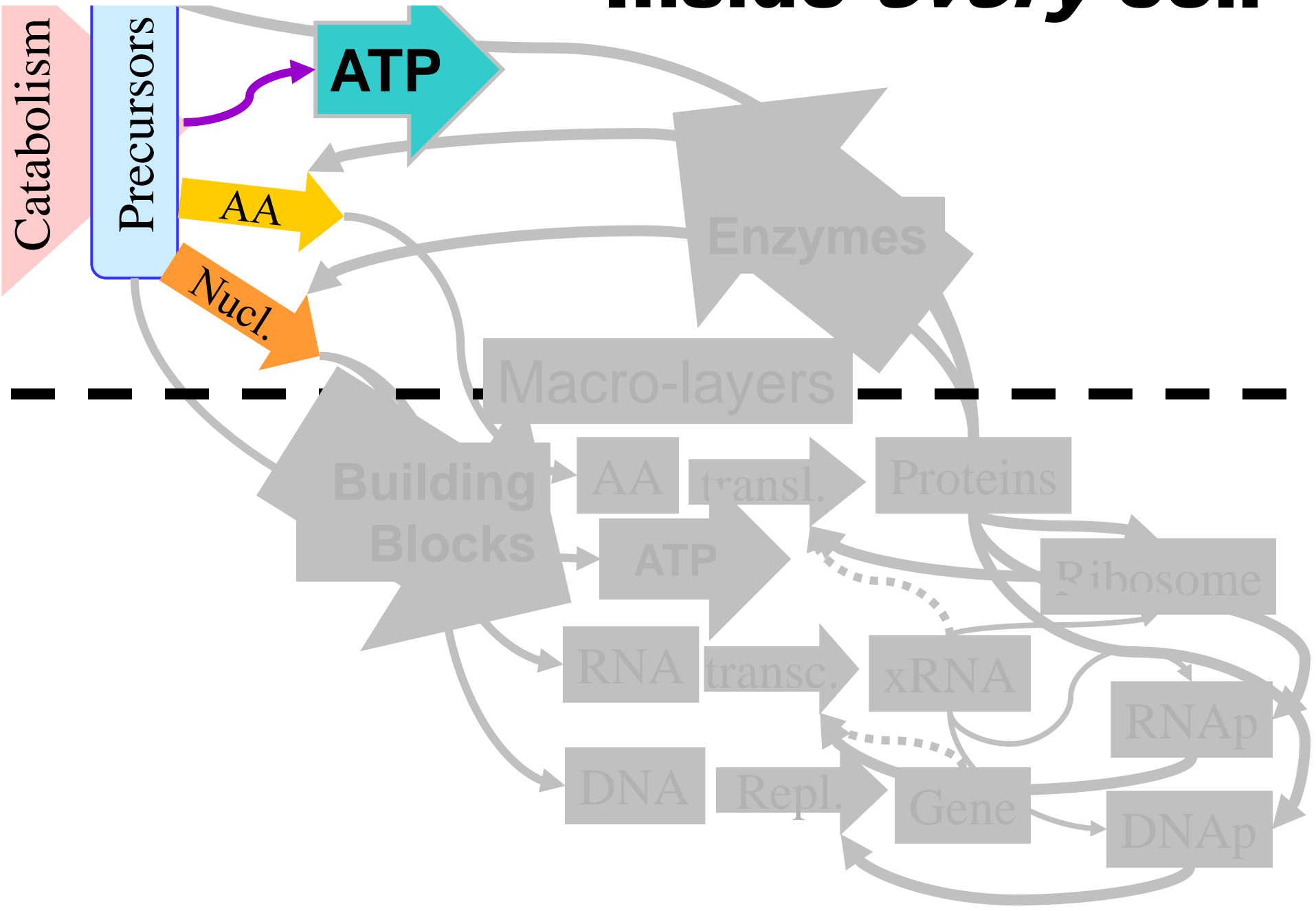


Mature red blood cells live 120 days

or “metabolism first”
origins of life?

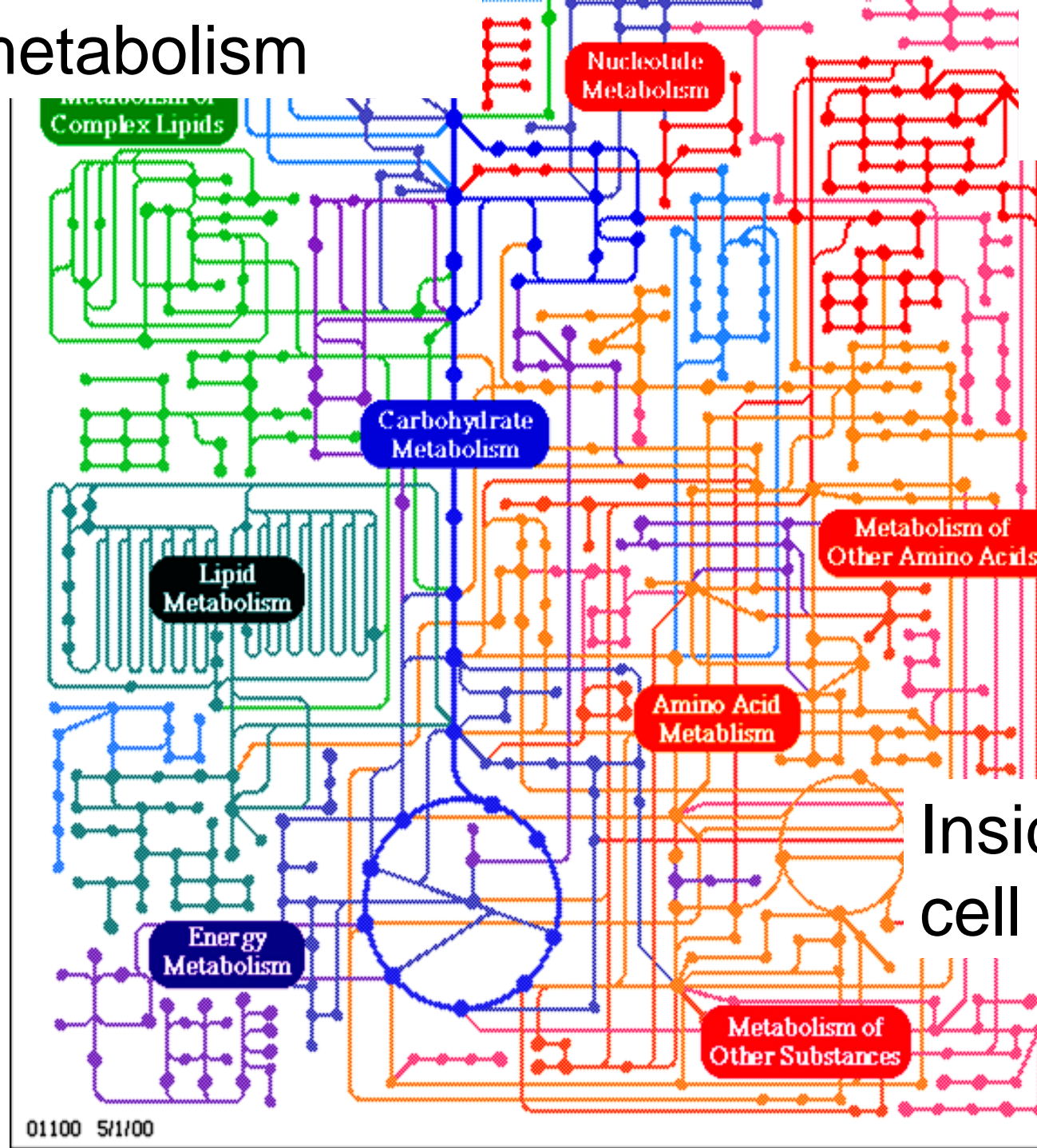
Core metabolism

Inside *every* cell



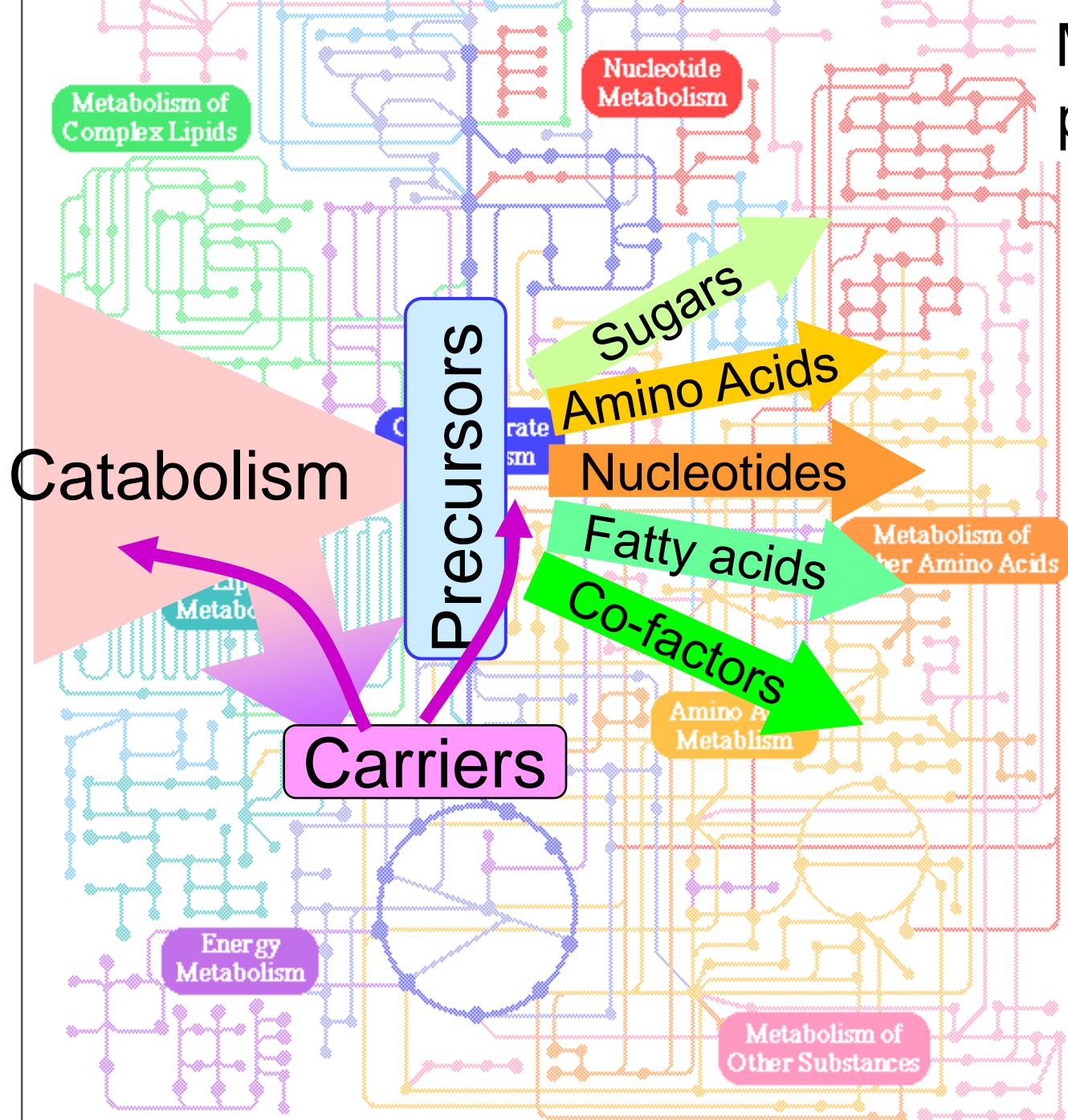
Core metabolism

Metabolic pathways



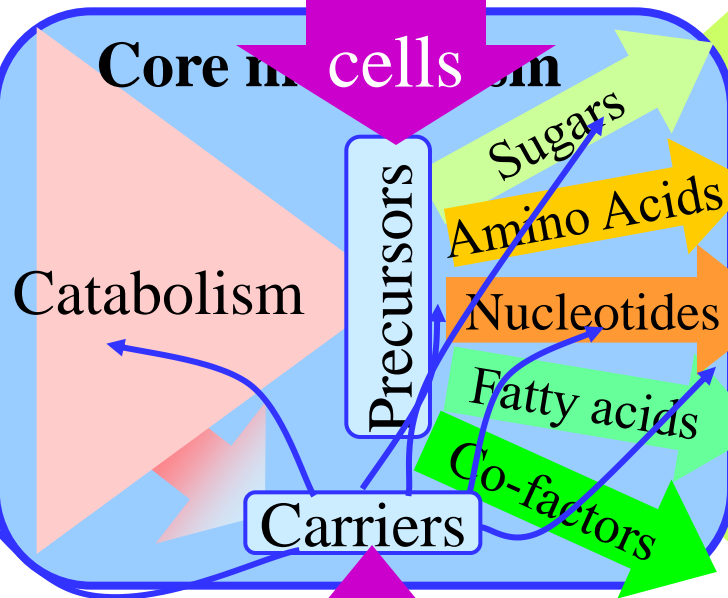
Inside every cell ($\approx 10^{30}$)

Metabolic pathways



**Taxis and
transport**

Nutrients



**Same
12
in all
cells**

**Same
8
in all
cells**

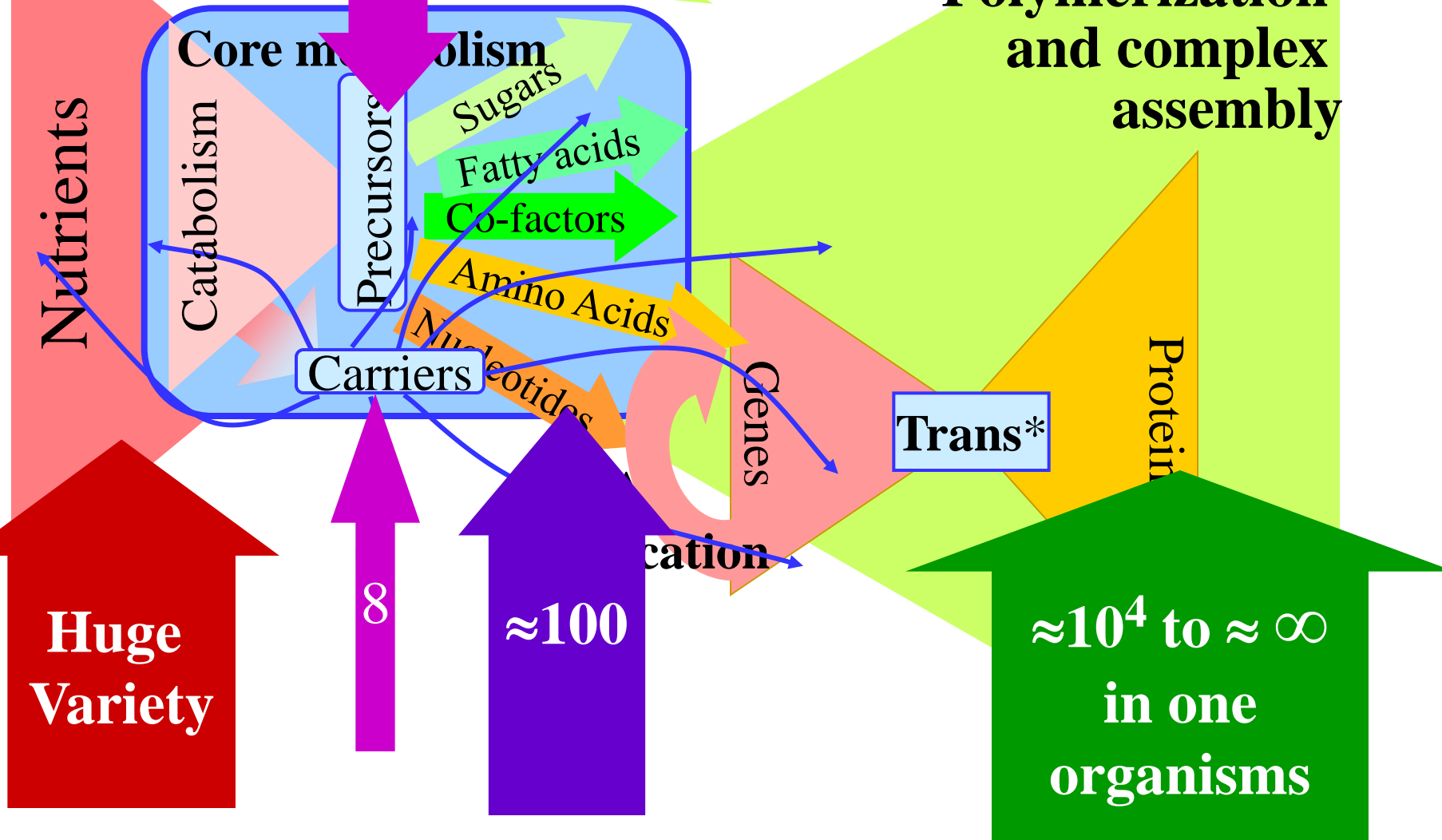
**≈100
≈same
in all
organisms**

**Huge
Variety**

Taxis and transport

Autocatalytic feedback

Polymerization and complex assembly



Universal reward systems

sports
music
dance
crafts
art
toolmaking
sex
food

VTA dopamine

Reward
Drive
Control
Memory

**Constraints
that
deconstrain**

Blood

Glucose
Oxygen

Organs
Tissues
Cells
Molecules

Universal metabolic system



Modularity 2.0

Constraints

dopamine 

Blood

Glucose

Oxygen

Modularity 2.0

sports
music
dance
crafts
art
toolmaking
sex
food

Reward
Drive
Control
Memory

**that
deconstrain**

Organs
Tissues
Cells
Molecules



Universal reward/metabolic systems

work
family
community
nature

food
sex
toolmaking
sports
music
dance
crafts
art

dopamine

Blood

Reward
Drive
Control
Memory

Organs
Tissues
Cells
Molecules

Robust and adaptive, yet ...

work
family
community
nature

sex
food
toolmaking
sports
music
dance
crafts
art

cocaine
amphetamine

dopamine

Blood

Reward
Drive
Control
Memory

Organs
Tissues
Cells
Molecules

work
family
community
nature

money

market/
consumer
culture

salt
sugar/fat
nicotine
alcohol

Vicarious

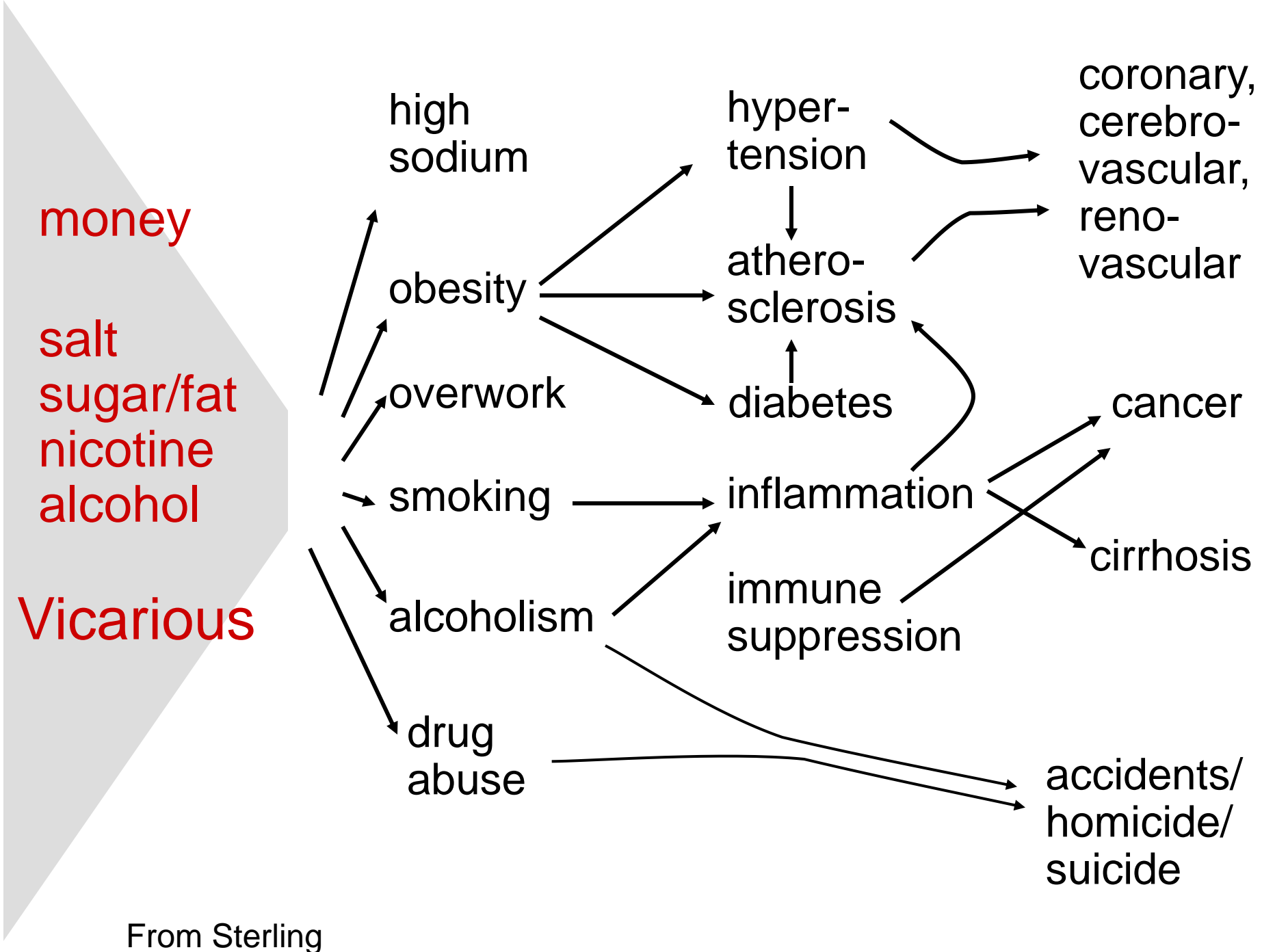
sex
toolmaking
sports
music
dance
crafts
art

industrial
agriculture

dopamine

Reward
Drive
Control
Memory

Organs
Tissues
Cells
Molecules



Universal reward systems

sports
music
dance
crafts
art
toolmaking
sex
food

ROBUST

VTA dopamine

Prefrontal
cortex

Nucleus accumbens

Blood

Glucose
Oxygen

Organs

Tissues

Cells

Molecules

Universal metabolic system

Robust

Yet Fragile

money

salt
sugar/fat
nicotine
alcohol

Vicarious

high
sodium

hyper-
tension

athero-
sclerosis

coronary,
cerebro-
vascular,
reno-
vascular

cancer

cirrhosis

alcoholism

immune
suppression

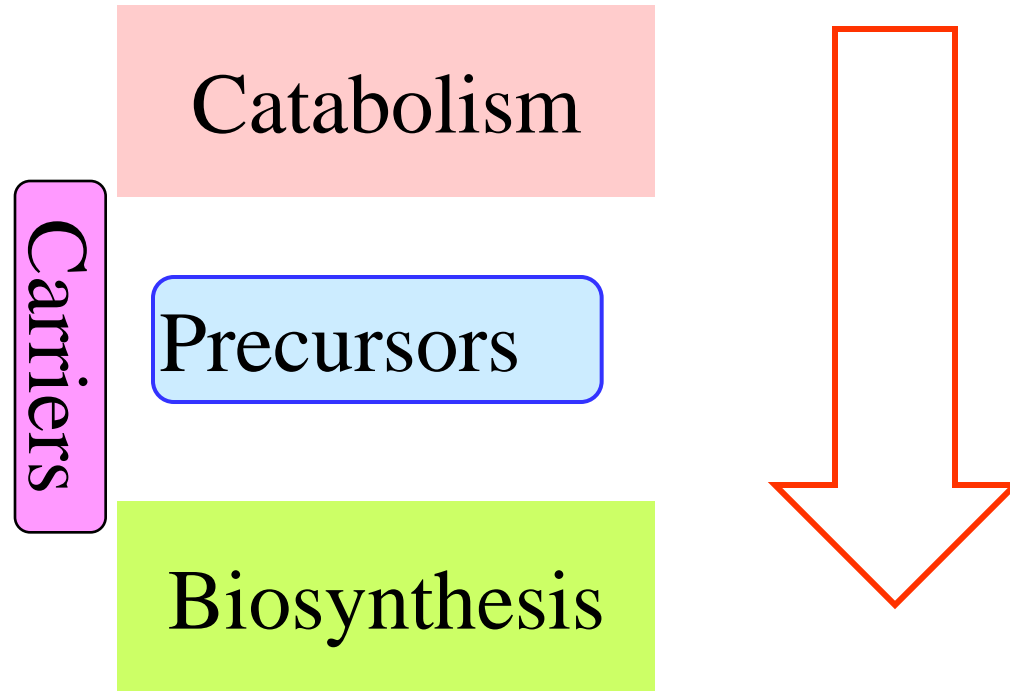
drug
abuse

accidents/
homicide/
suicide

Glucose
Oxygen

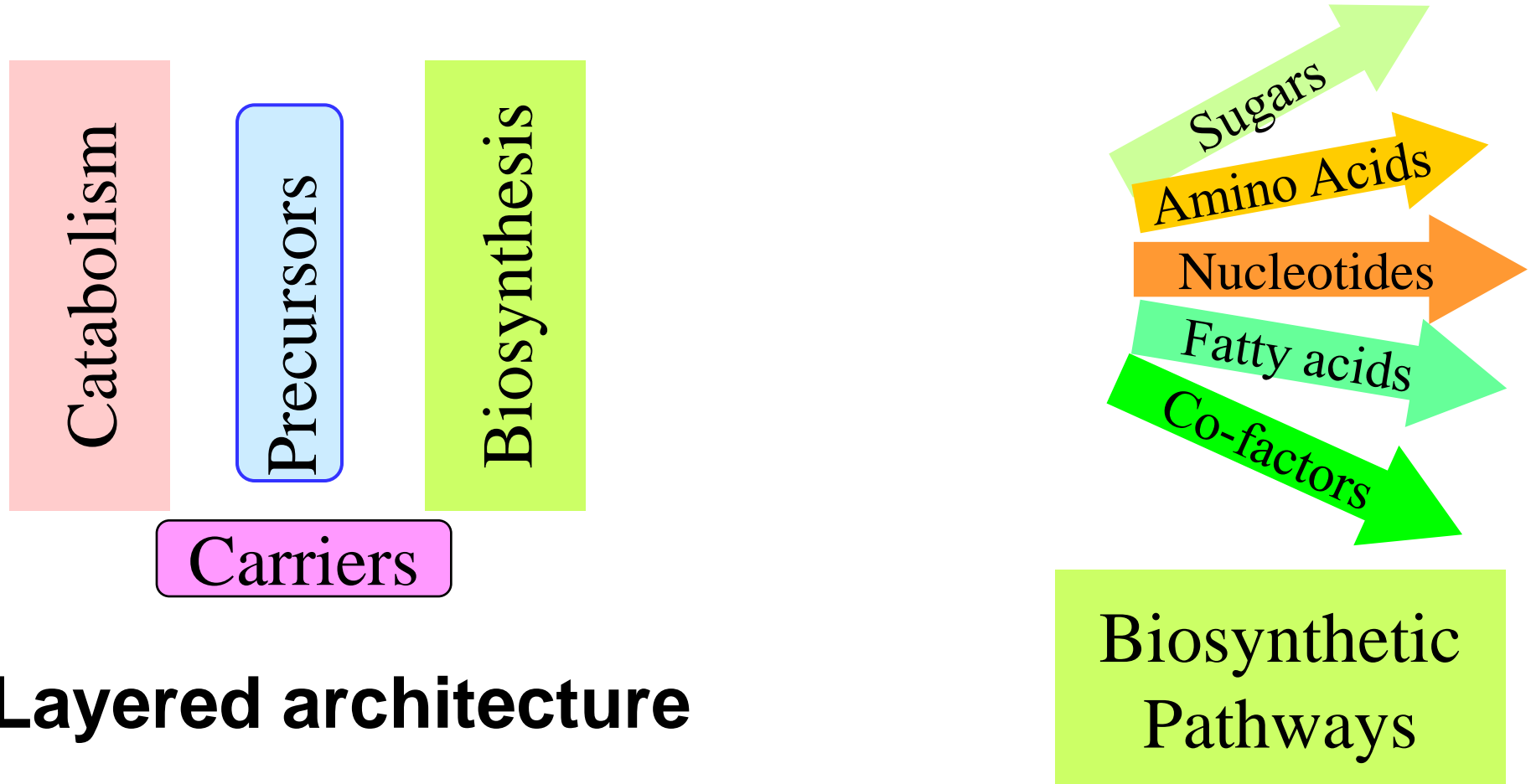
clonidine
VTA
dopamine

Inside every cell

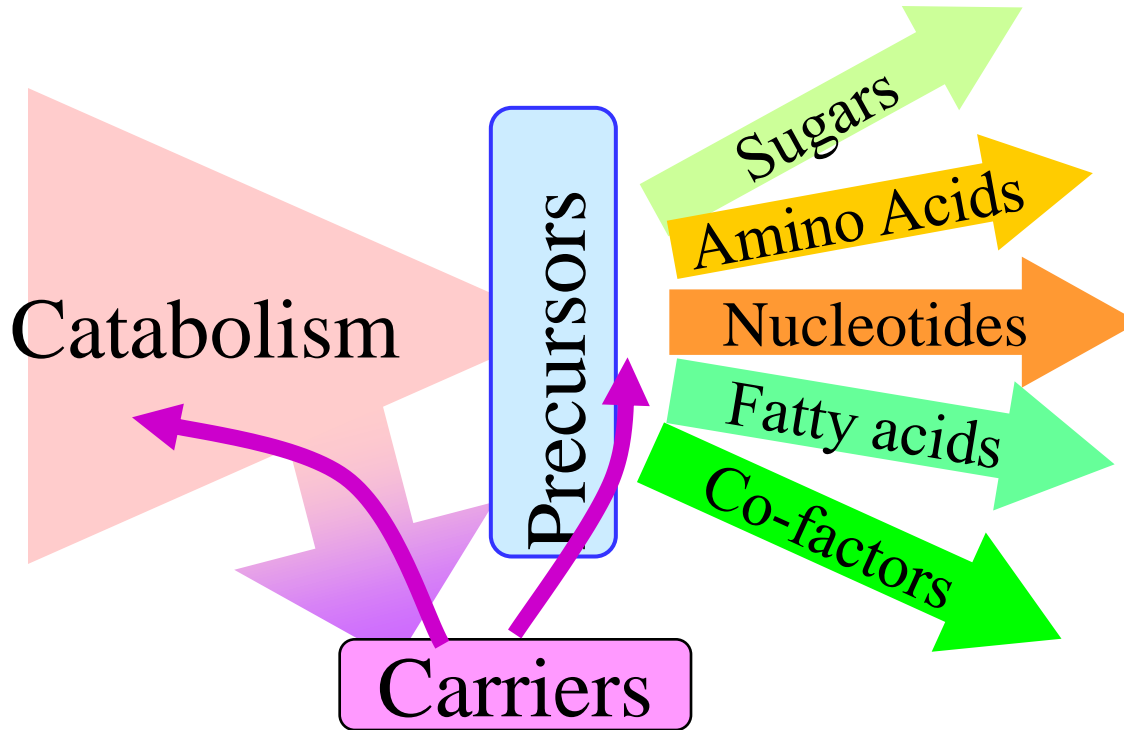


Layered architecture

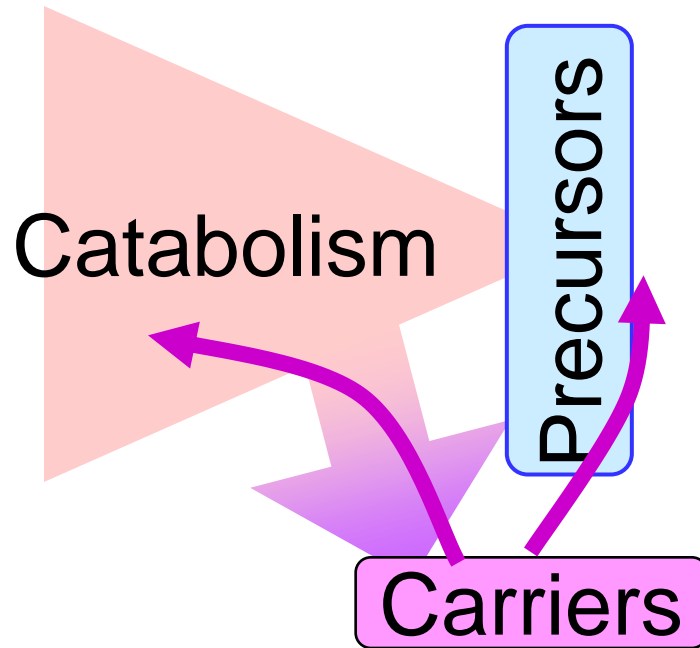
Inside every cell



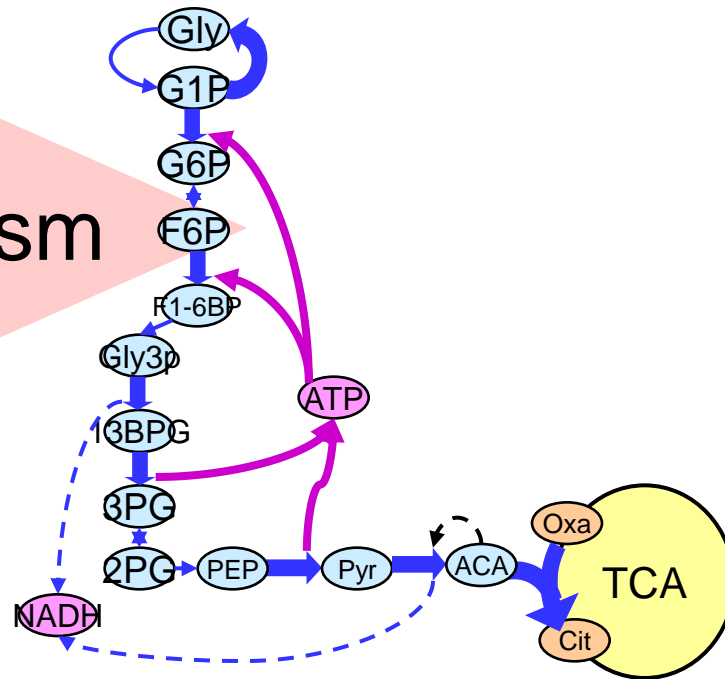
Inside every cell

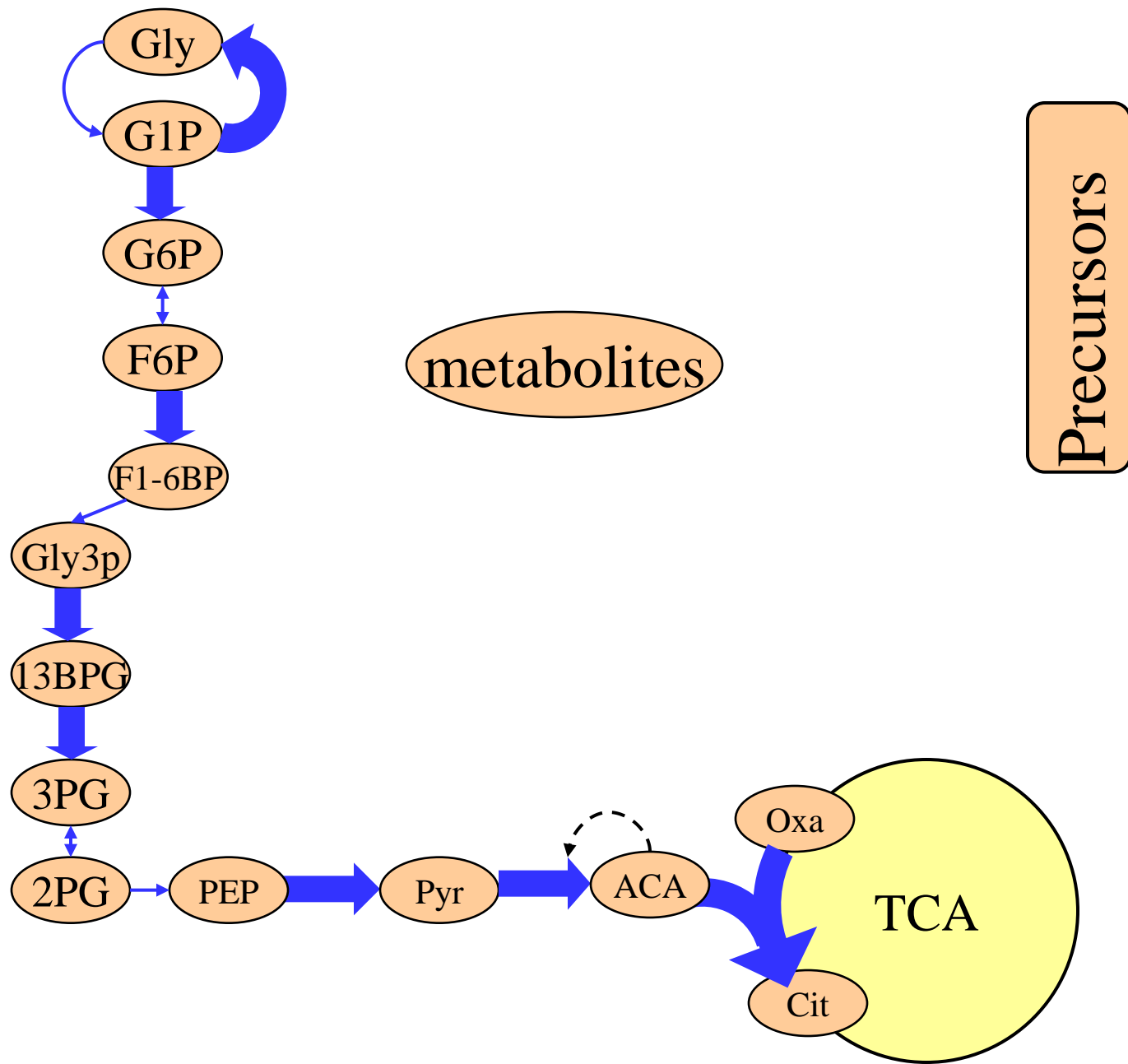


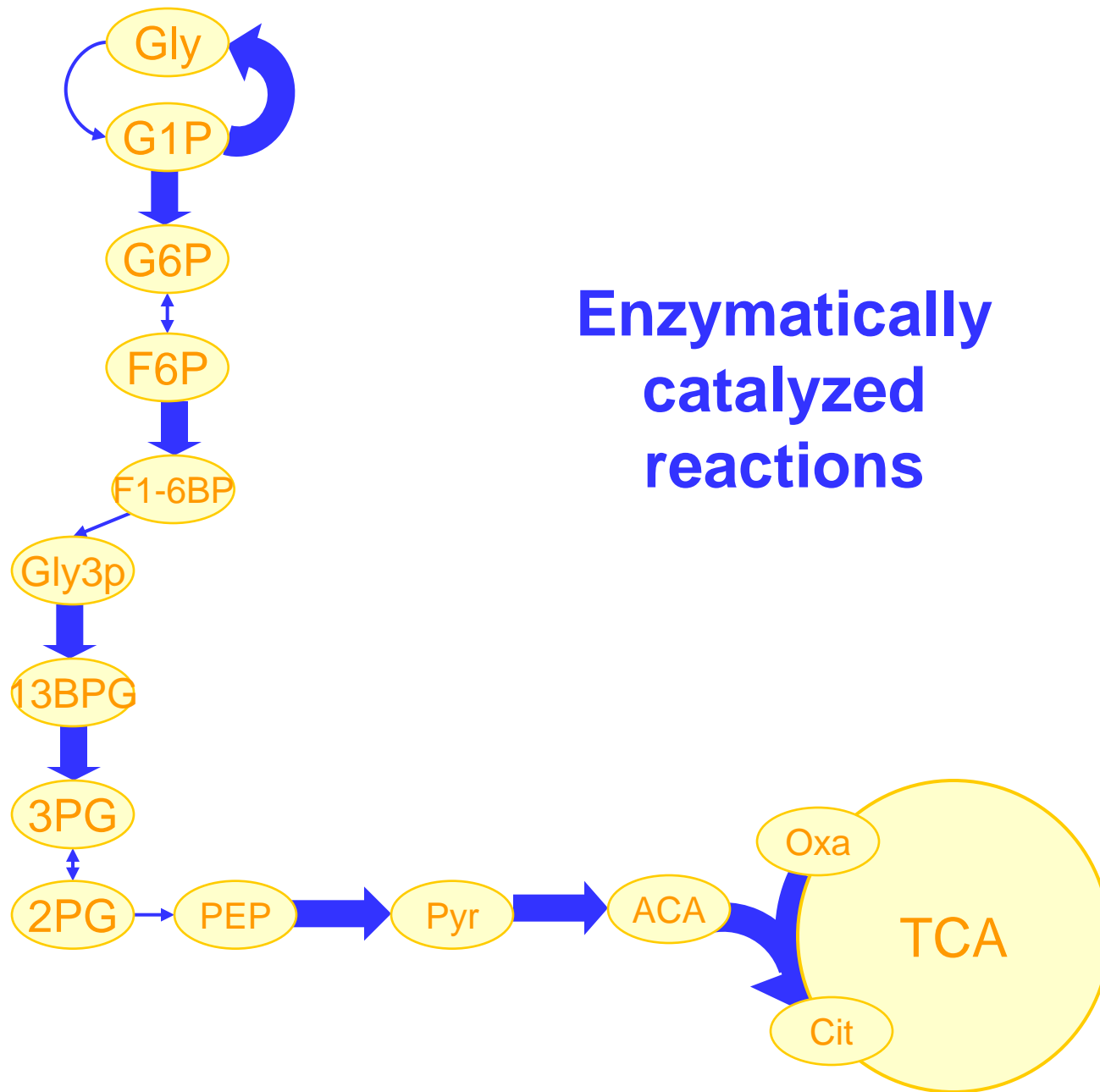
Core metabolic bowtie
Layered architecture

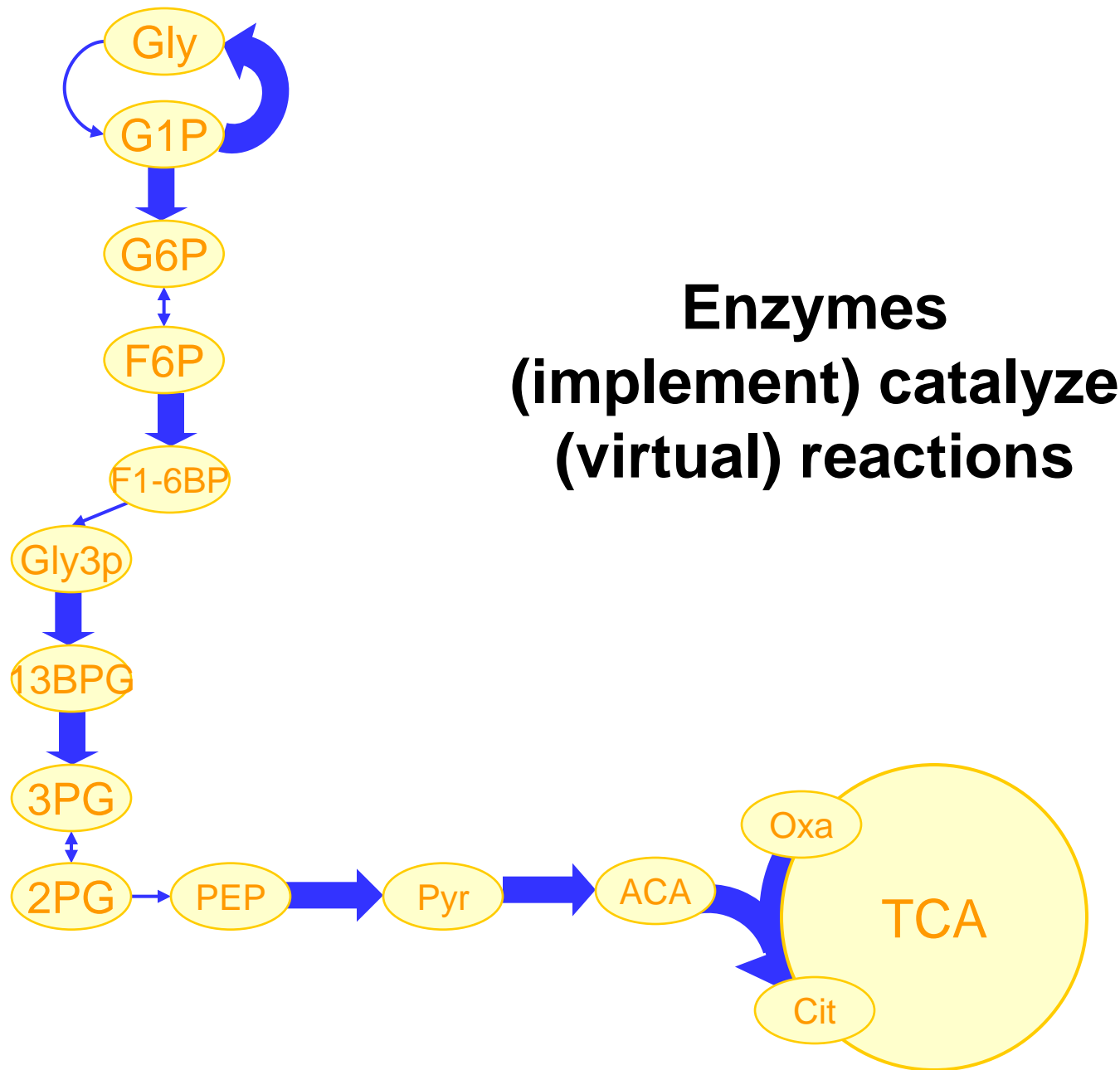


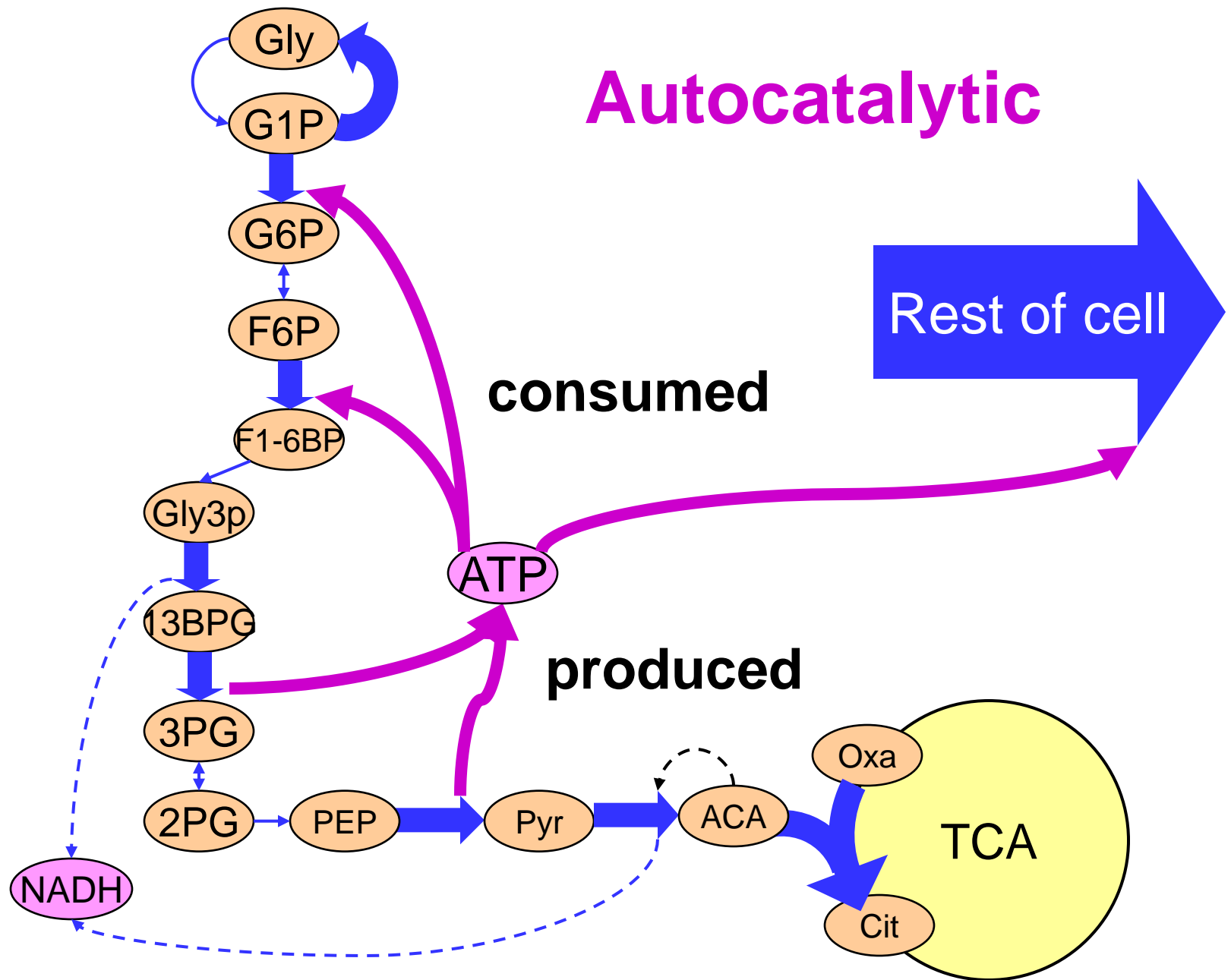
Catabolism



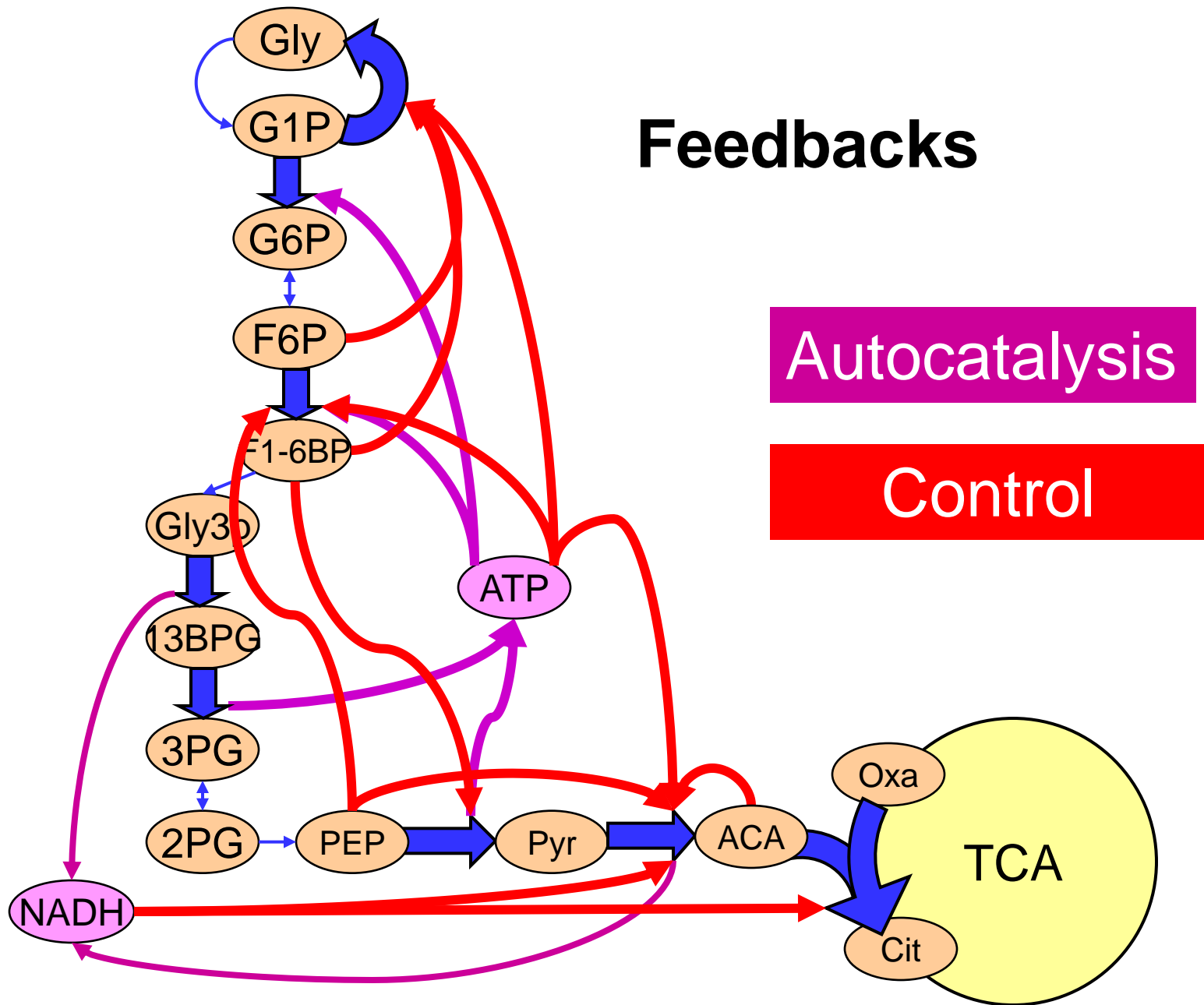


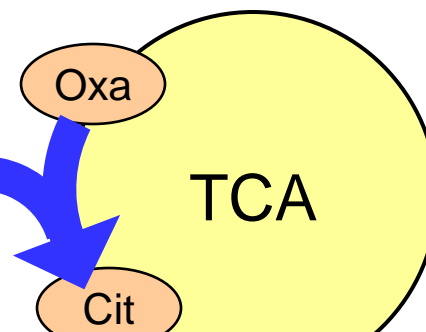
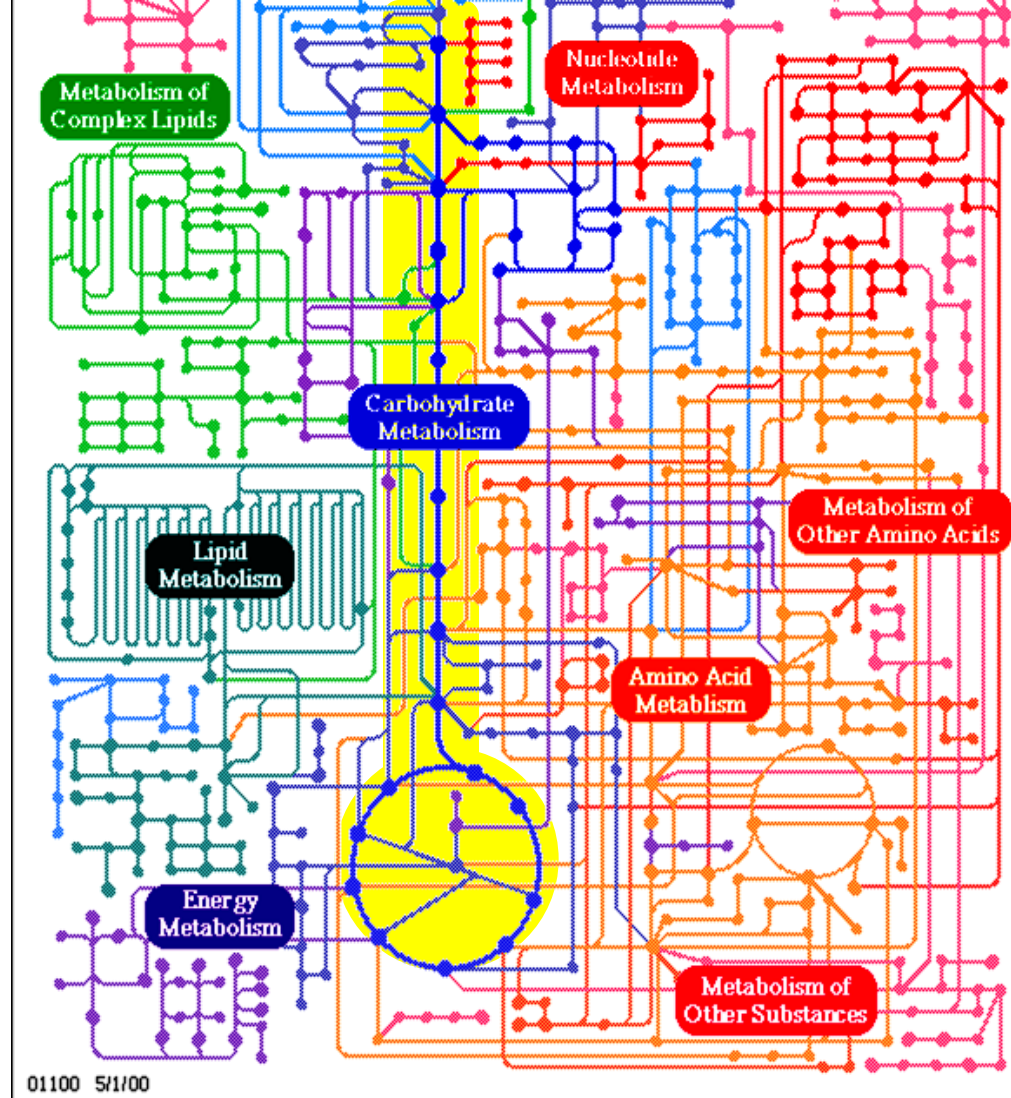
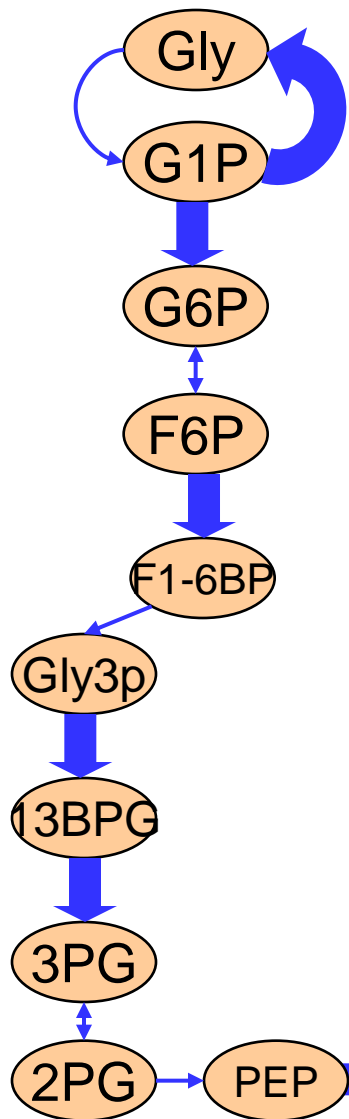


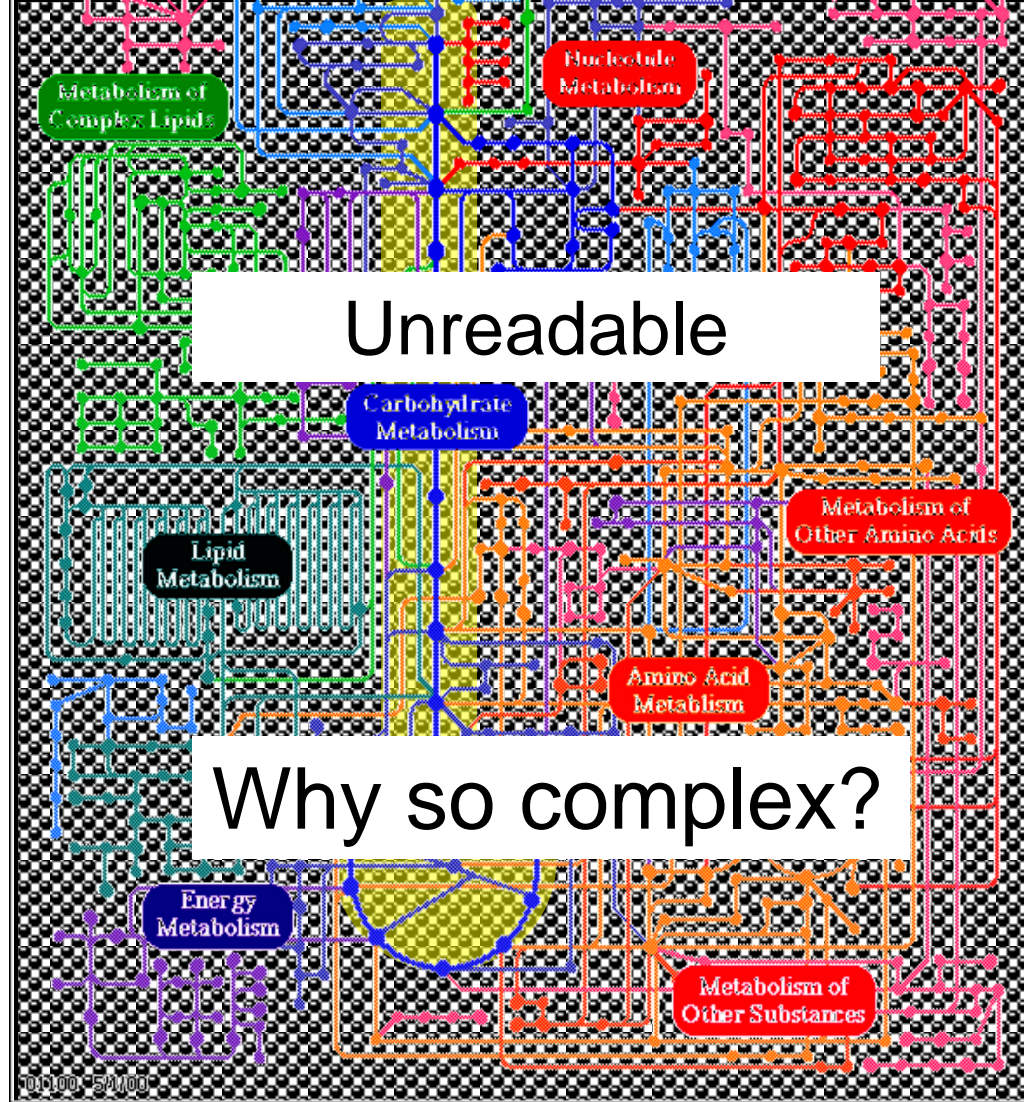
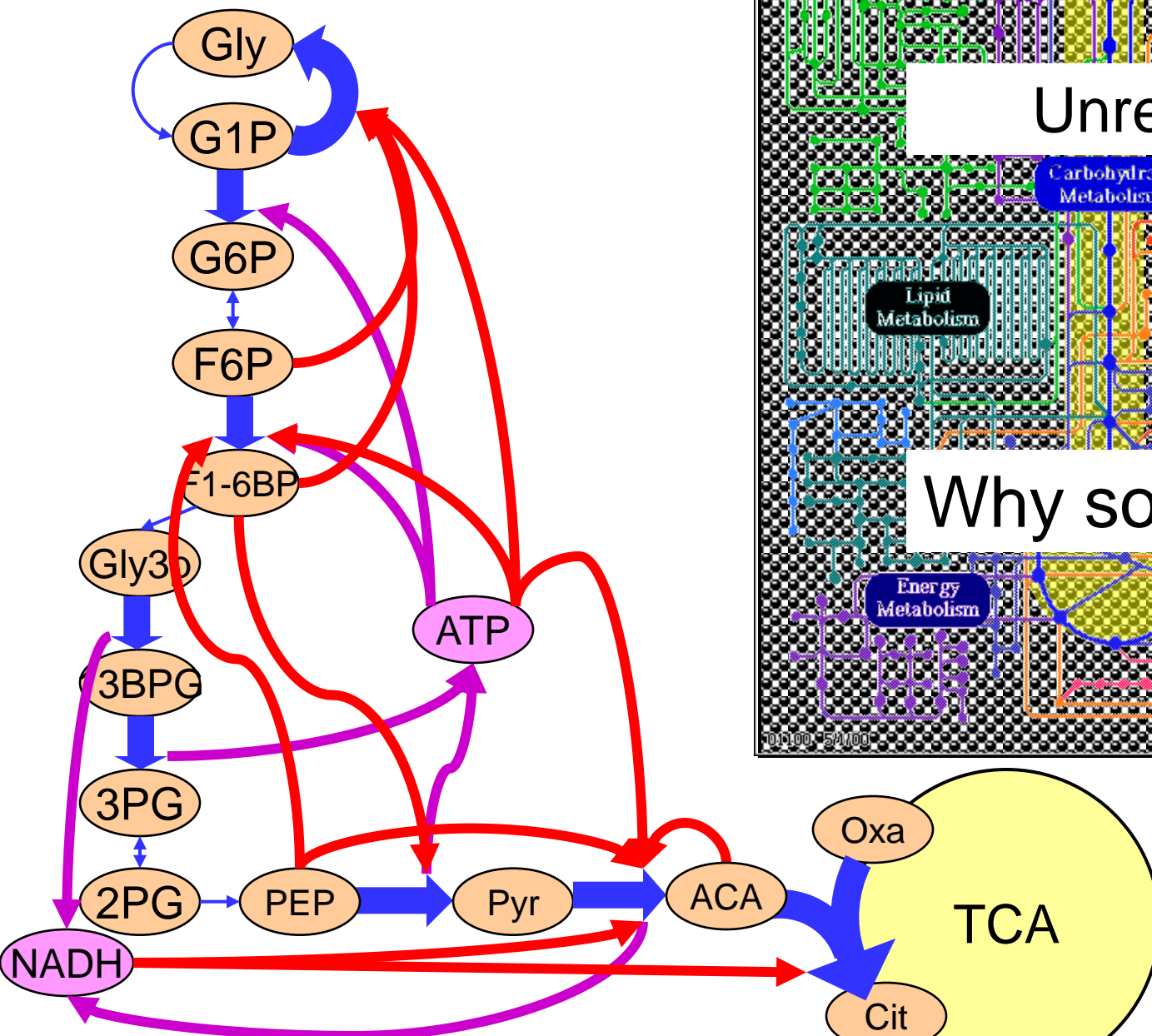




Feedbacks





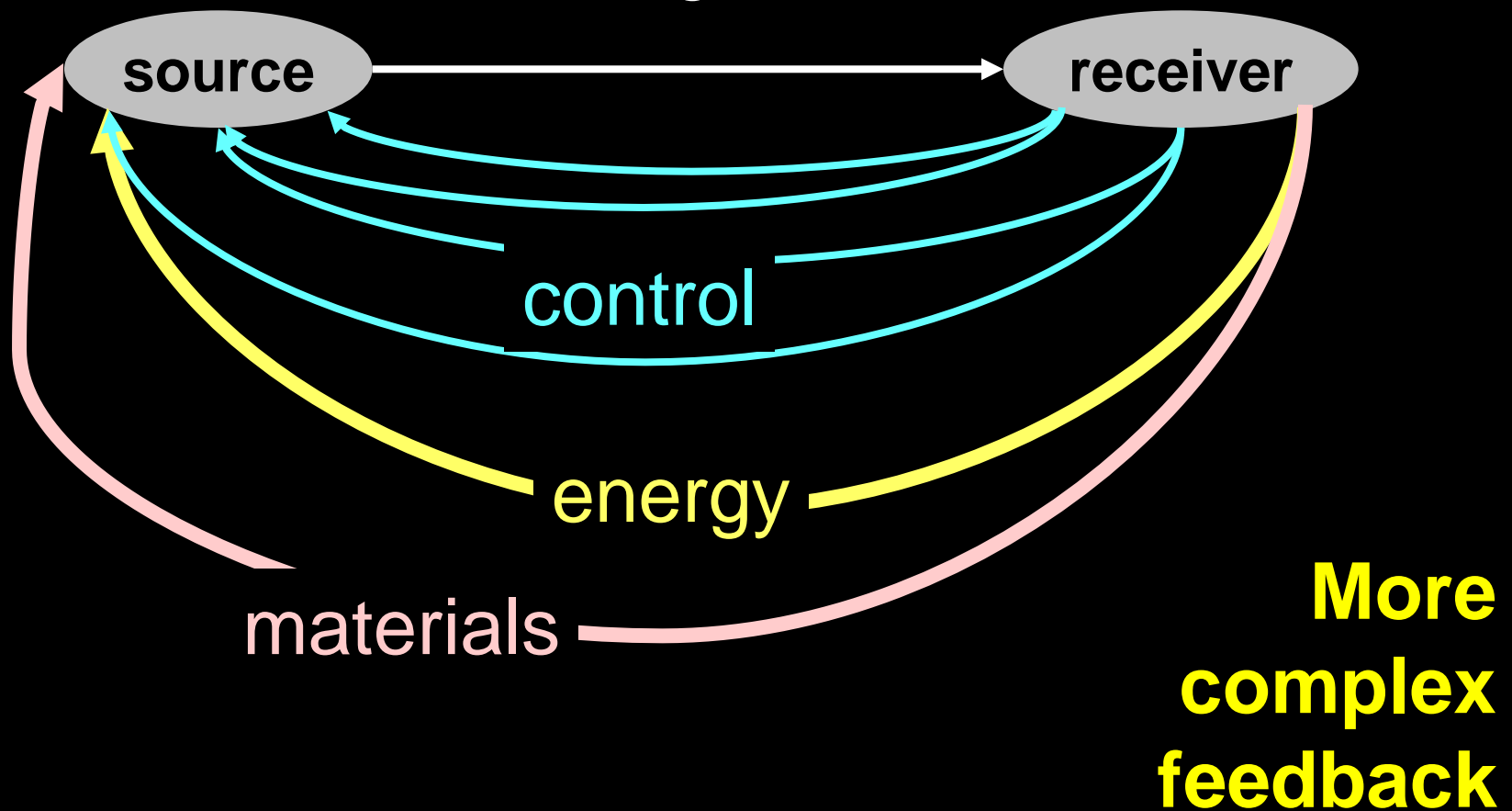


signaling
gene expression
metabolism
lineage

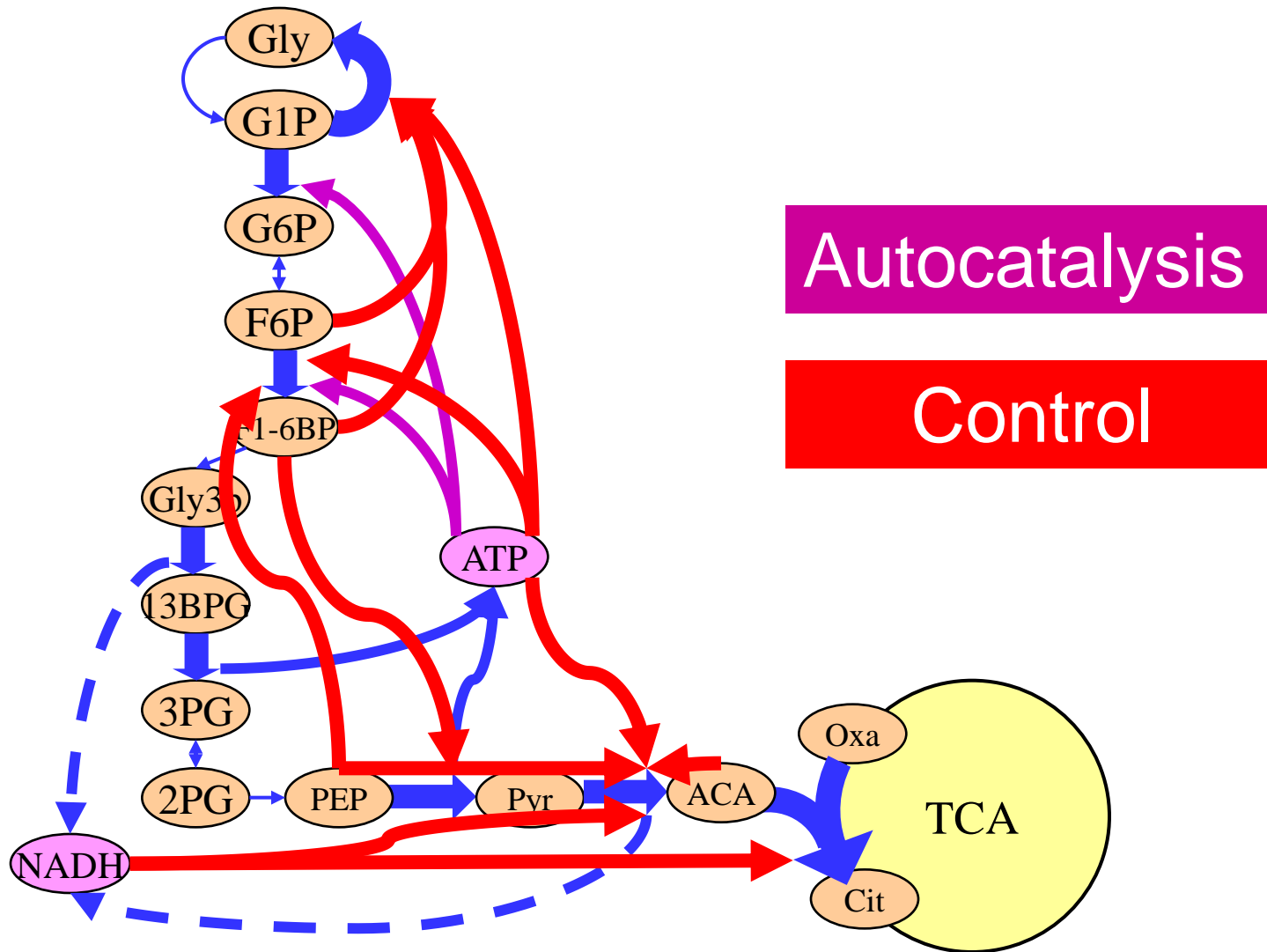


**Biological
pathways**

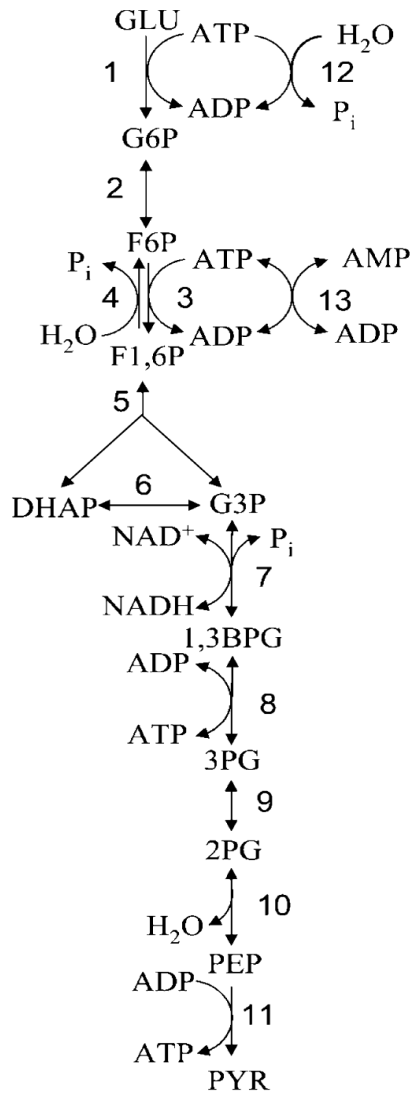
signaling
gene expression
metabolism
lineage



Feedbacks



Nikolaev, ..., Elucidation and Structural Analysis of **Conserved Pools** for Genome-Scale Metabolic Reconstructions, Biophysical Journal, Volume 88, Issue 1, January 2005, Pages 37-49



(P1) Total NAD moiety: $[\text{NAD}^+] + [\text{NADH}]$

(P2) Total adenylate moiety: $[\text{ATP}] + [\text{ADP}] + [\text{AMP}]$

(P3) Total carbon moiety: $2[\text{GLU}] + 2[\text{G6P}] + 2[\text{F6P}] + 2[\text{F1,6P}] + [\text{DHAP}] +$

(P4) Total phosphate moiety

(P5) Total oxygen moiety:

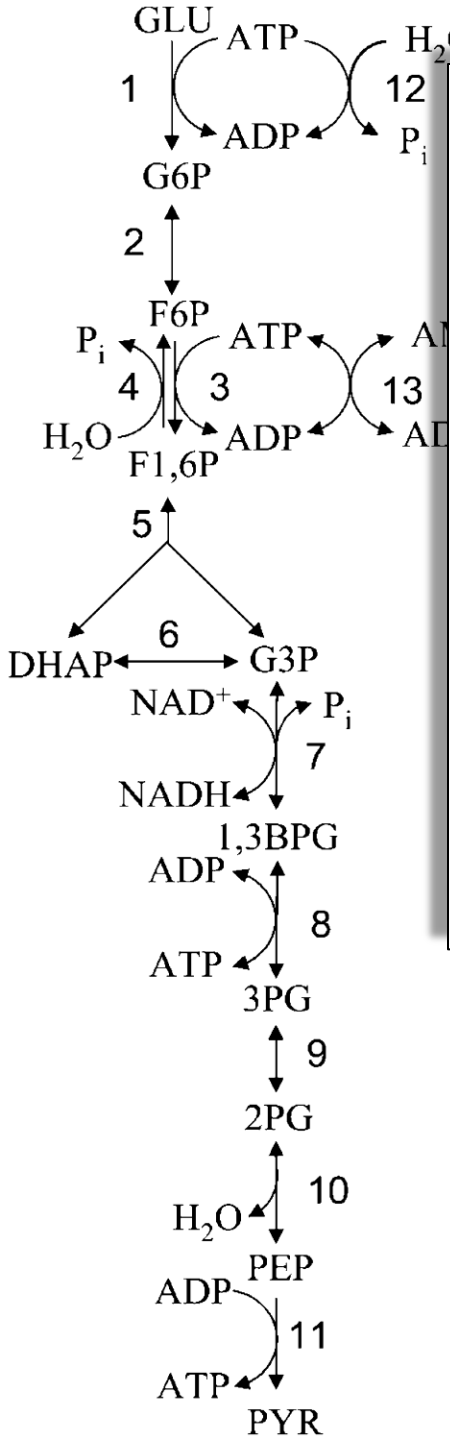
(P6) Oxidized state of meta

(P7) Reduced state of metabolites: $2[\text{GLU}] + 2[\text{G6P}] + 2[\text{F6P}] + 2[\text{F1,6P}] +$
 $+ [\text{DHAP}] + [\text{G3P}] + [\text{NADH}]$

(P8) High energy potential release: $2[\text{GLU}] + [\text{G6P}] + [\text{F6P}] + [\text{3PG}] + [\text{2PG}] +$
 $+ [\text{PYR}] + [\text{ADP}] + 2[\text{AMP}] + [\text{H}_2\text{O}]$

Constrained (conserved):

1. Total NAD moiety
2. Total Adenylate moiety
3. Total Carbon moiety
4. Total phosphate moiety
5. Total oxygen moiety
6. Oxidized state of metabolites
7. Reduced state of metabolites
8. High energy potential release



Constrained (“conserved”): Moieties

1. NAD
2. Adenylate
3. Carbon
4. phosphate
5. oxygen
6. Oxidized state of metabolites
7. Reduced state of metabolites
8. High energy potential release

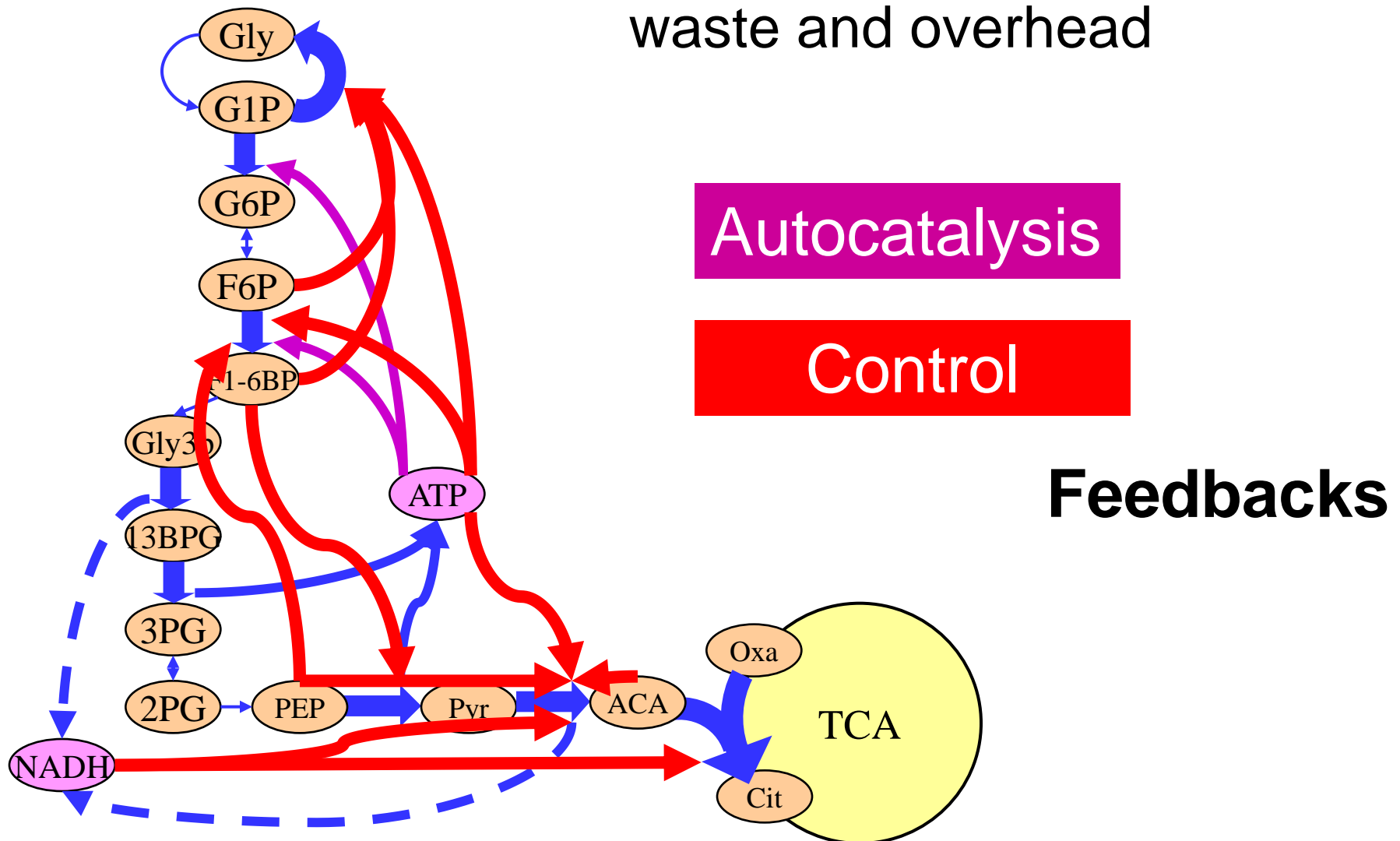
(P7) Reduced state of metabolites: $2[\text{GLU}] + 2[\text{G6P}] + 2[\text{F6P}] + 2[\text{F1,6P}] +$

$$\frac{1}{\pi} \int_0^{\infty} \ln |S(j\omega)| \left(\frac{z}{z^2 + \omega^2} \right) d\omega \geq \ln \left| \frac{z + p}{z - p} \right|$$

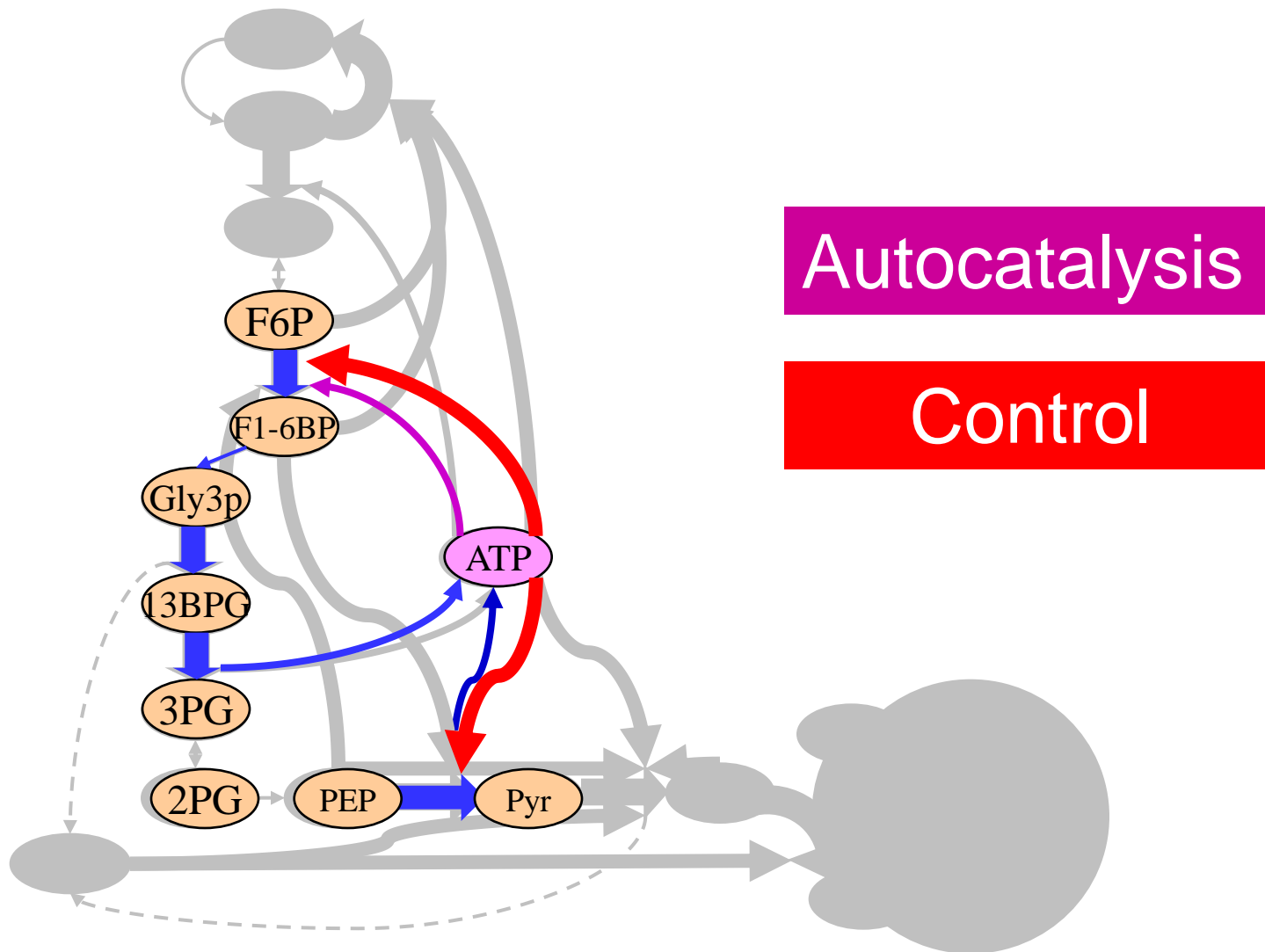
$+ [\text{PYR}] + [\text{ADP}] + 2[\text{AMP}] + [\text{H}_2\text{O}]$

Robust=maintain energy charge
w/fluctuating cell demand

Efficient=minimize metabolic
waste and overhead



Minimal model?



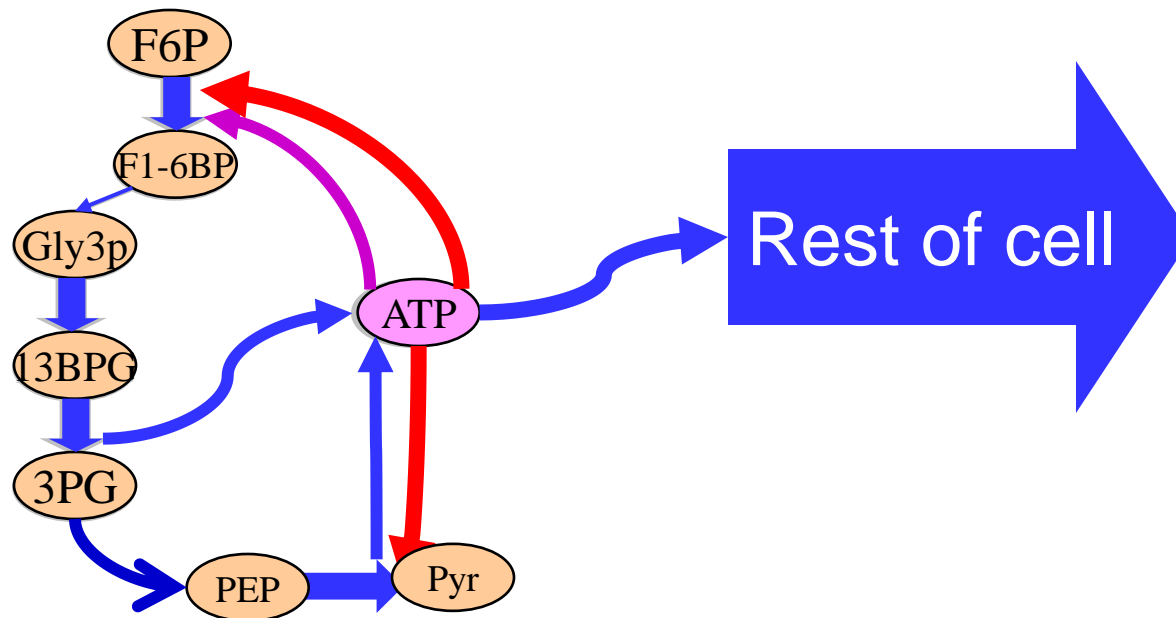
Minimal model

~1 equilibrium

2 metabolites

3 “reactions”

Control
Plus
Autocatalytic
Feedback

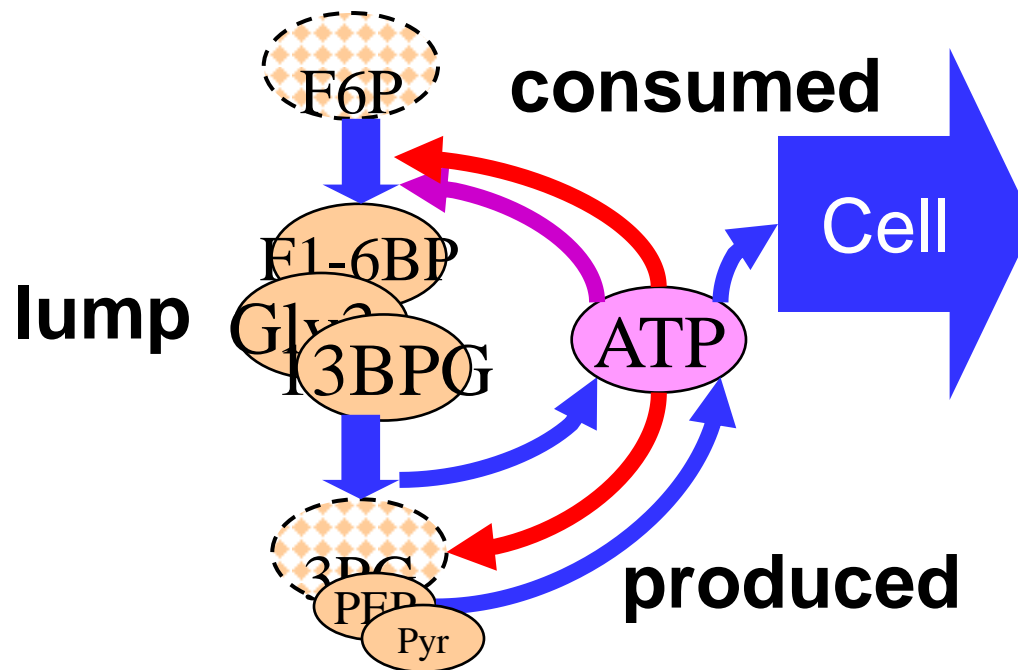


Minimal model

~1 equilibrium

2 metabolites

3 “reactions”



UG biochem, math,
control theory

Glycolytic Oscillations and Limits on Robust Efficiency

Fiona A. Chandra,^{1*} Gentian Buzi,² John C. Doyle²

Both engineering and evolution are constrained by trade-offs between efficiency and robustness, but theory that formalizes this fact is limited. For a simple two-state model of glycolysis, we explicitly derive analytic equations for hard trade-offs between robustness and efficiency with oscillations as an inevitable side effect. The model describes how the trade-offs arise from individual parameters, including the interplay of feedback control with autocatalysis of network products necessary to power and catalyze intermediate reactions. We then use control theory to prove that the essential features of these hard trade-off “laws” are universal and fundamental, in that they depend minimally on the details of this system and generalize to the robust efficiency of any autocatalytic network. The theory also suggests worst-case conditions that are consistent with initial experiments.

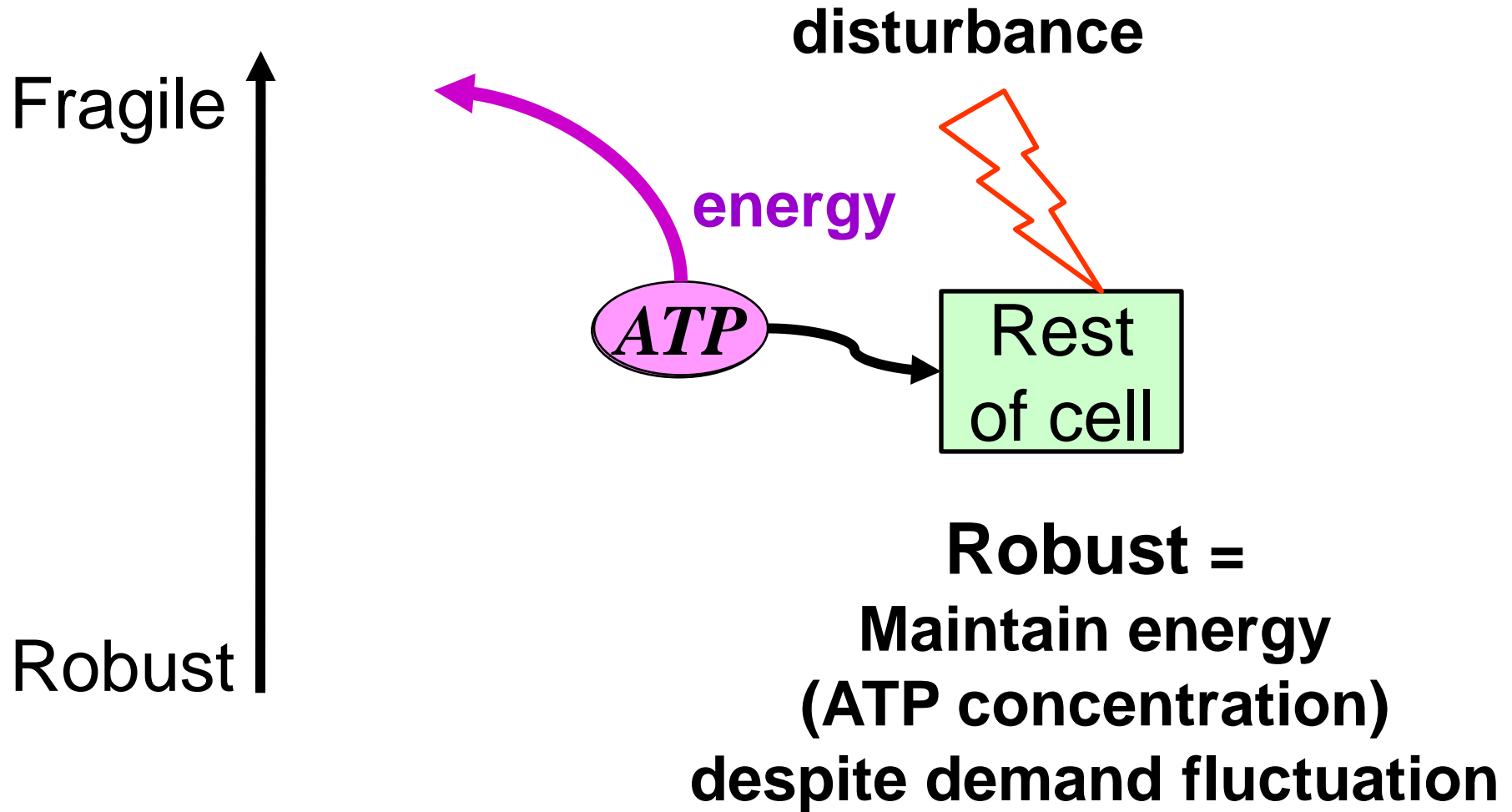
un-
fo-
w-
the cell's use of ATP. In glycolysis, two ATP molecules are consumed upstream and four are produced downstream, which normalizes to $q = 1$ (each y molecule produces two downstream) with kinetic exponent $a = 1$. To highlight essential trade-offs with the simplest possible analysis, we normalize the concentration such that the unperturbed ($\delta = 0$) steady states are $\bar{y} = 1$ and $\bar{x} = 1/k$ [the system can have one additional steady state, which is unstable when $(1, 1/k)$ is stable]. [See the supporting online material (SOM) part I]. The basal rate of the PFK reaction and the consumption rate have been normalized to 1 (the 2 in the numerator and feedback coefficients of the reactions come from these normalizations). Our results hold for more general systems as discussed below and in SOM, but the analysis

Chandra, Buzi, and Doyle

Most important paper so far.



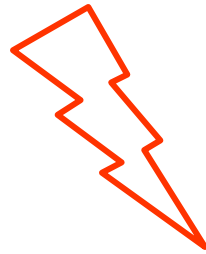
(May 21): Hard tradeoff in glycolysis



disturbance

Accurate vs
sloppy

Fragile



What makes this hard?

1. Instability (autocatalysis)
2. Delay (enzyme amount)


Robust

Robust

\approx Disturbance rejection

\approx Accurate

Fragile



Robust

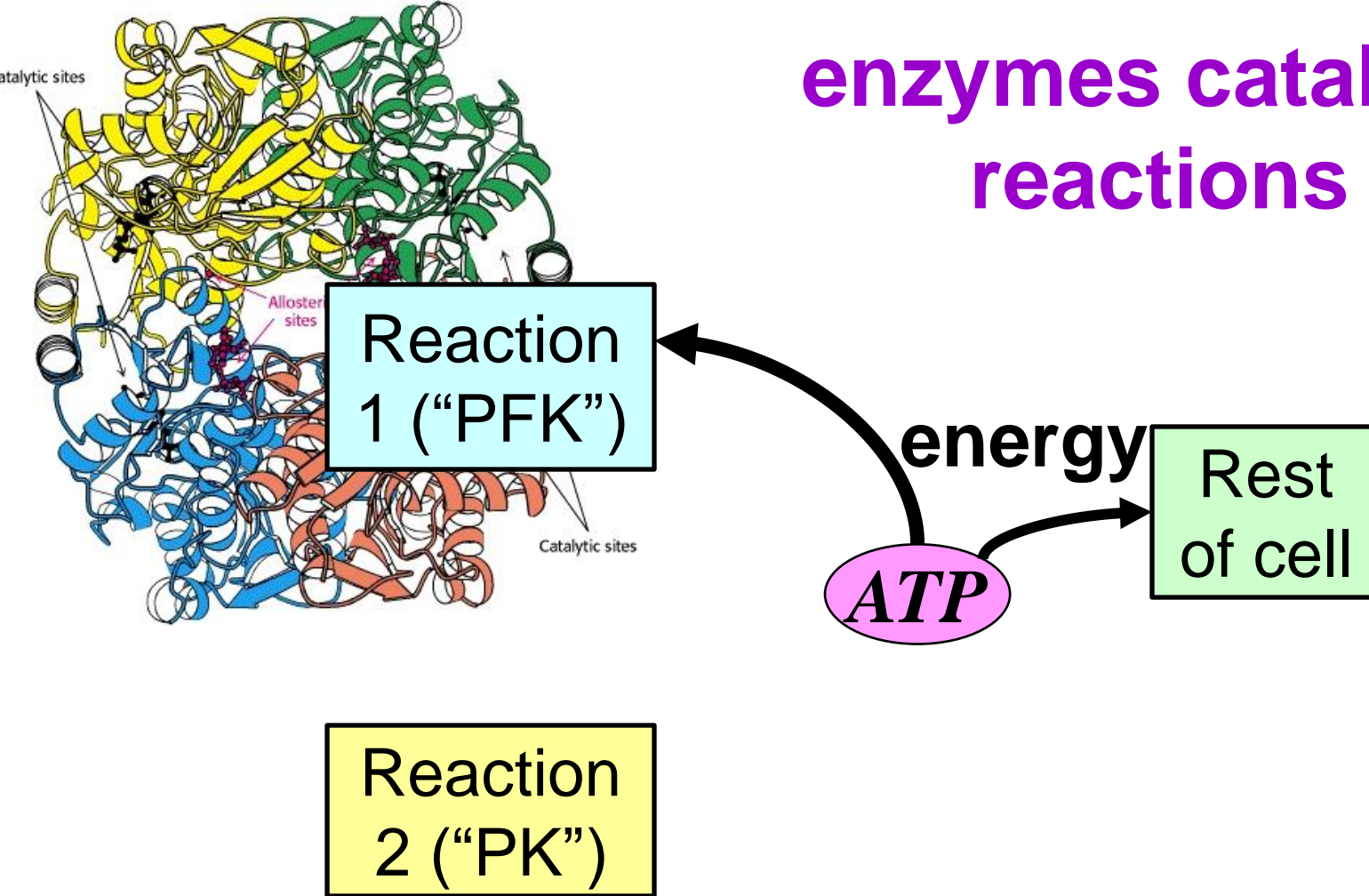
What makes this hard?

1. Instability
2. Delay

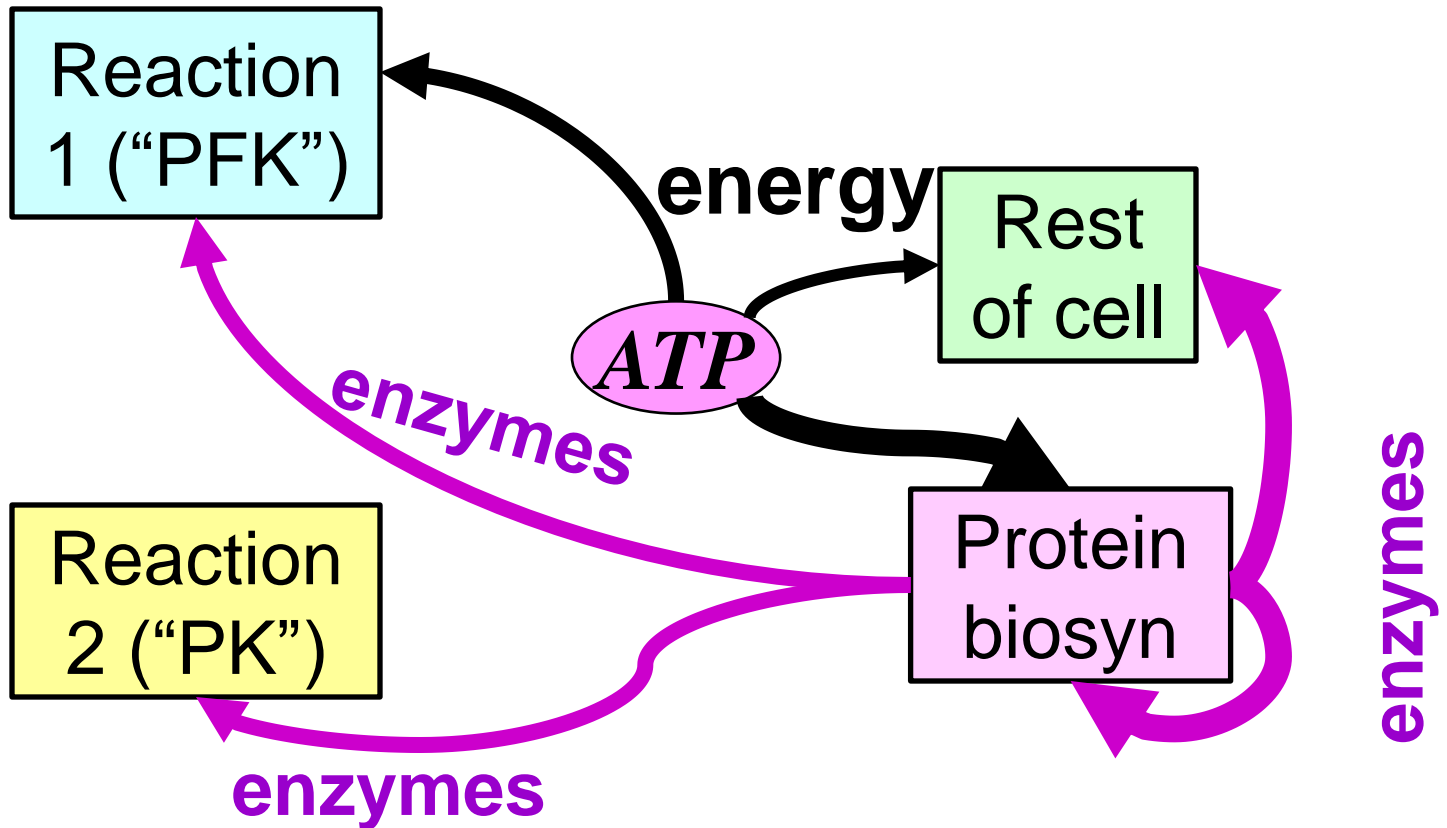
The CNS must cope with both

Today's important point

enzymes catalyze reactions



enzymes catalyze
reactions, another
source of autocatalysis



Efficient =
low metabolic overhead
 \approx low enzyme amount

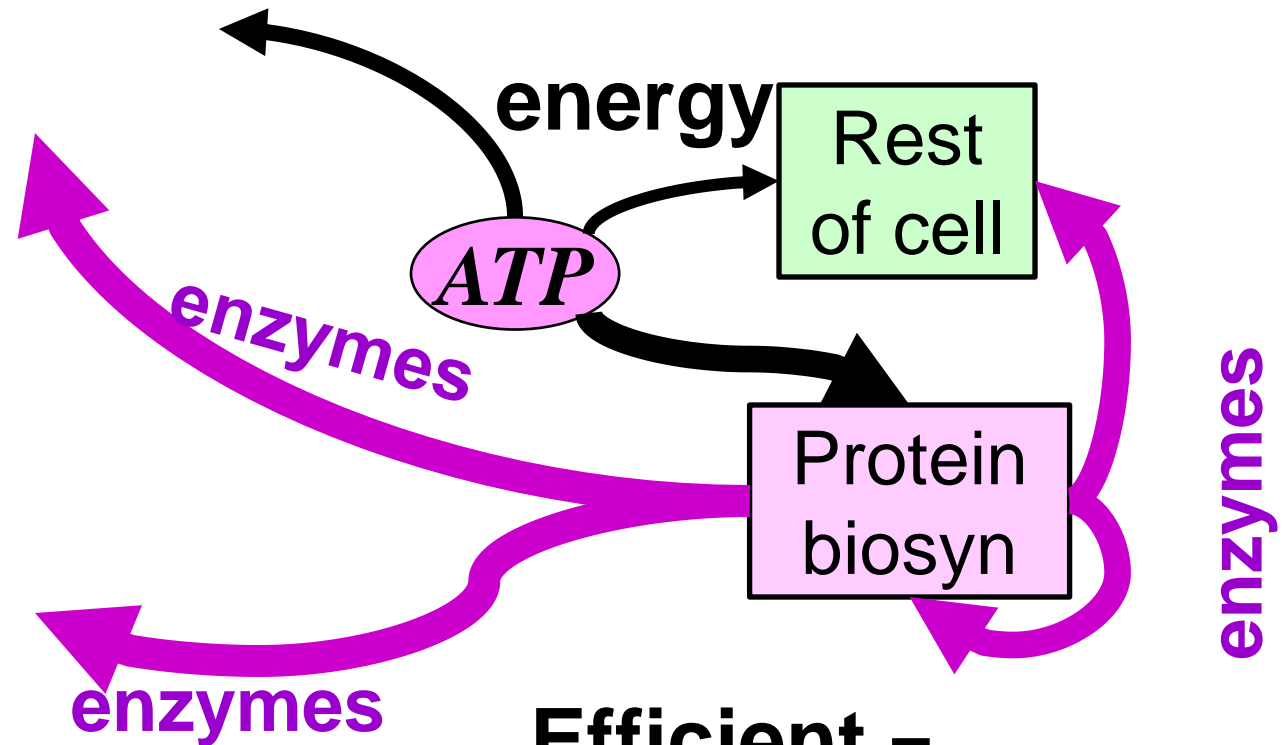
enzymes catalyze
reactions, another
source of autocatalysis

reaction
rates

\propto

enzyme
amount

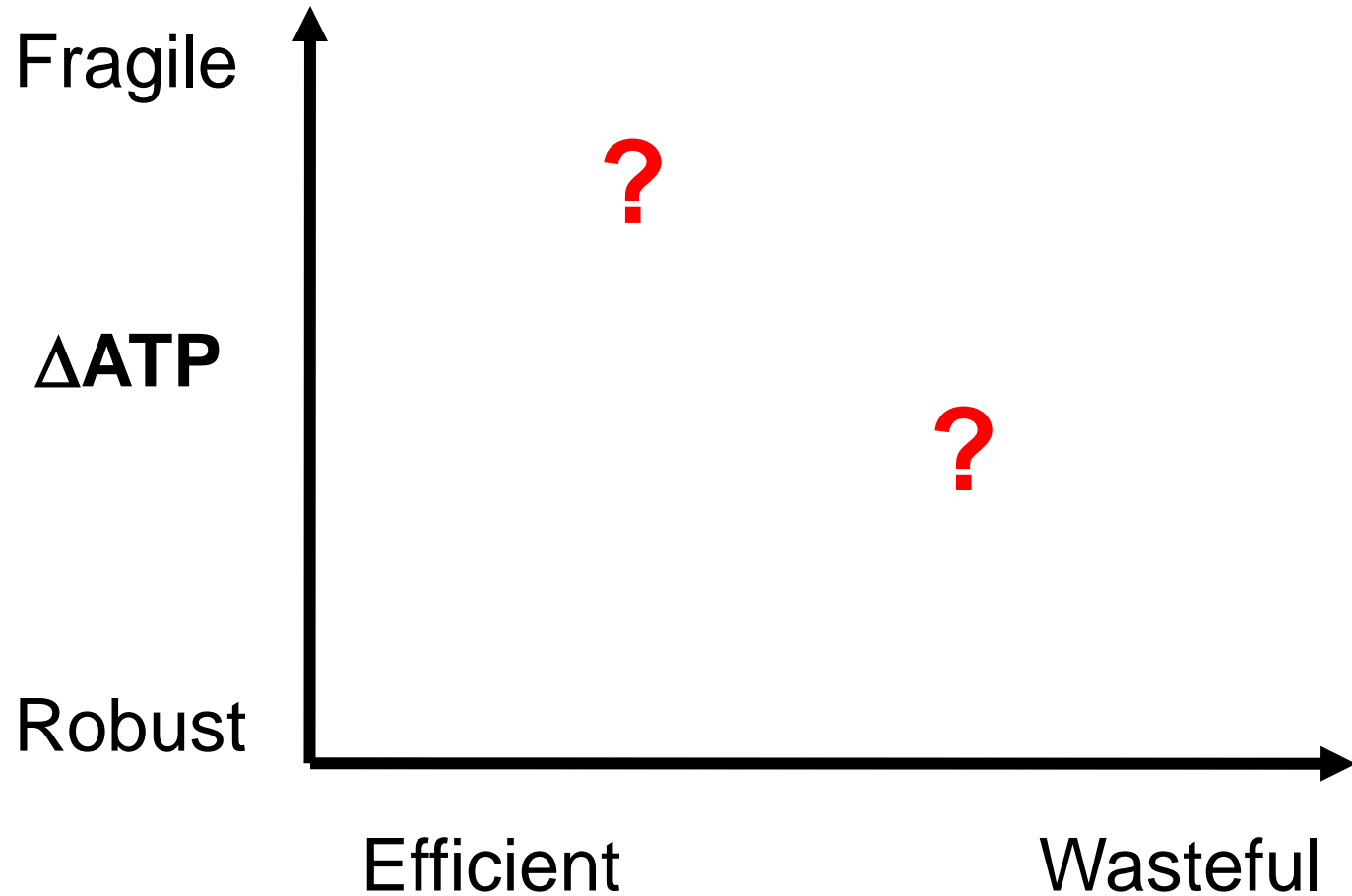
Can't make
too many
enzymes
here,
need to
supply rest
of the cell.



Efficient =

low metabolic overhead
 \approx low enzyme amount
(\Rightarrow **slow reactions**)

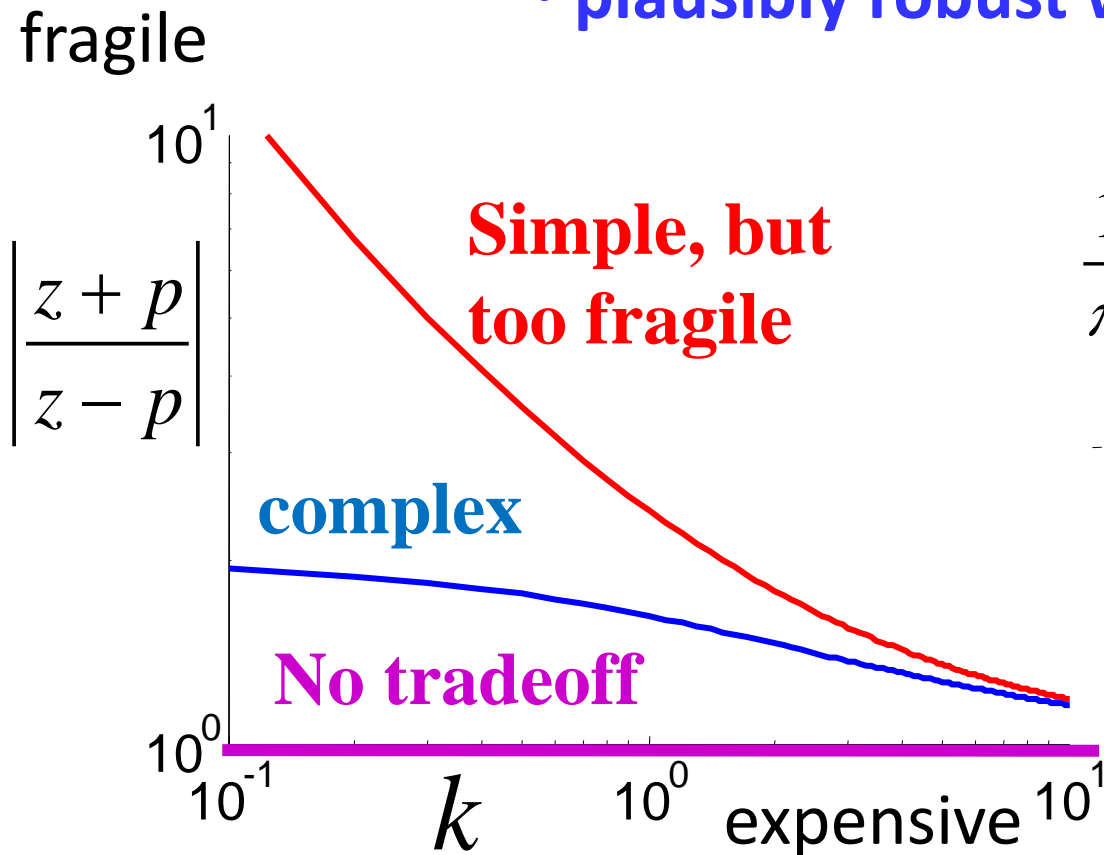
**Robust =
Maintain
ATP**



**Efficient =
low enzyme amount
(\Rightarrow slow reactions)**

(May 21): Hard tradeoff in glycolysis is

- **robustness vs efficiency**
- **absent without autocatalysis**
- **too fragile with simple control**
- **plausibly robust with complex control**

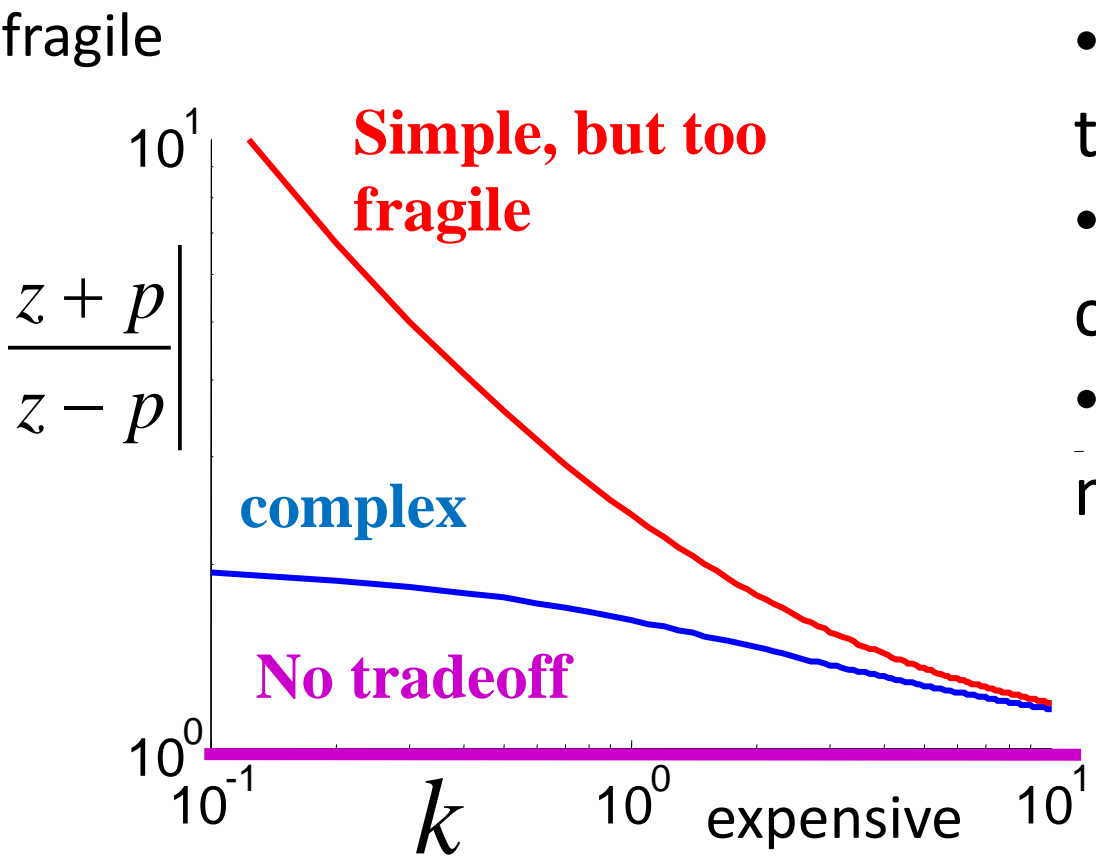


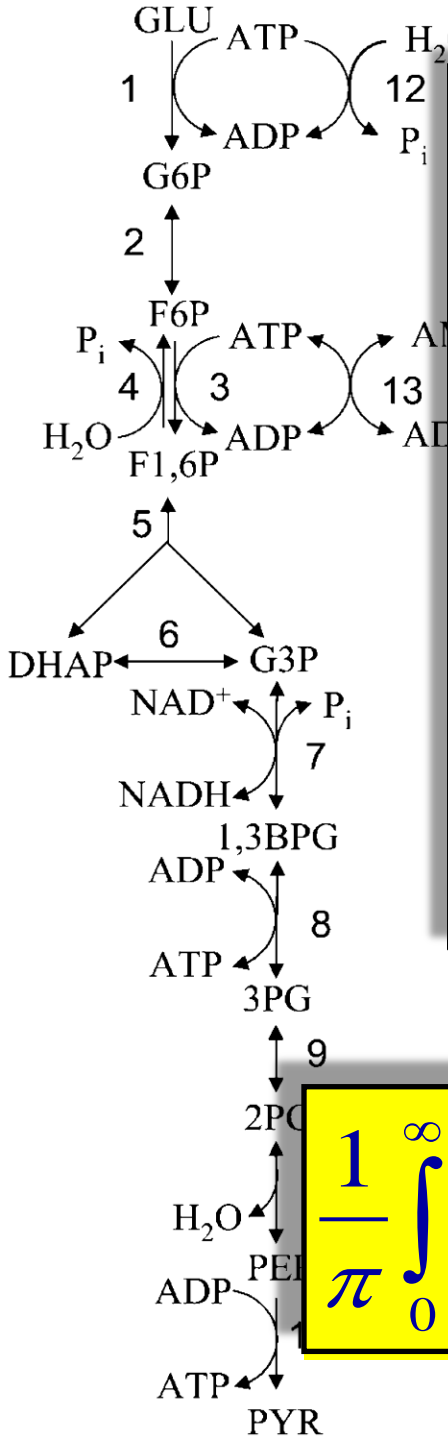
$$\frac{1}{\pi} \int_0^{\infty} \ln |S(j\omega)| \left(\frac{z}{z^2 + \omega^2} \right) d\omega$$
$$\geq \ln \left| \frac{z+p}{z-p} \right|$$

(May 21): Hard tradeoff in glycolysis is

- robustness vs efficiency
- absent without autocatalysis
- too fragile with simple control
- plausibly robust with complex control

- Evolution can
- increase complexity
- to improve robustness tradeoffs.
- But this complexity creates new fragilities
- so there is always more to this story.



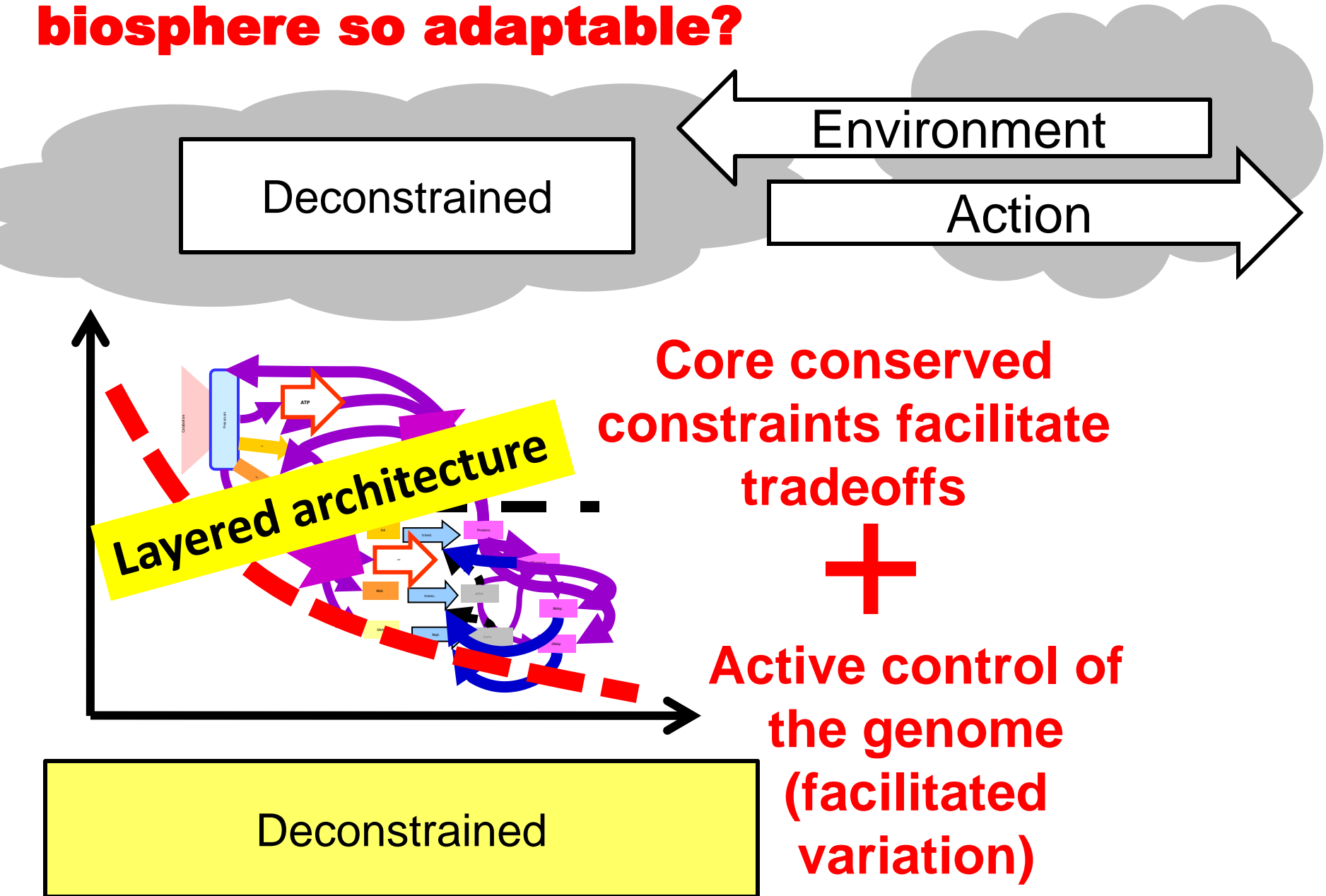


Constrained (“conserved”): Moieties

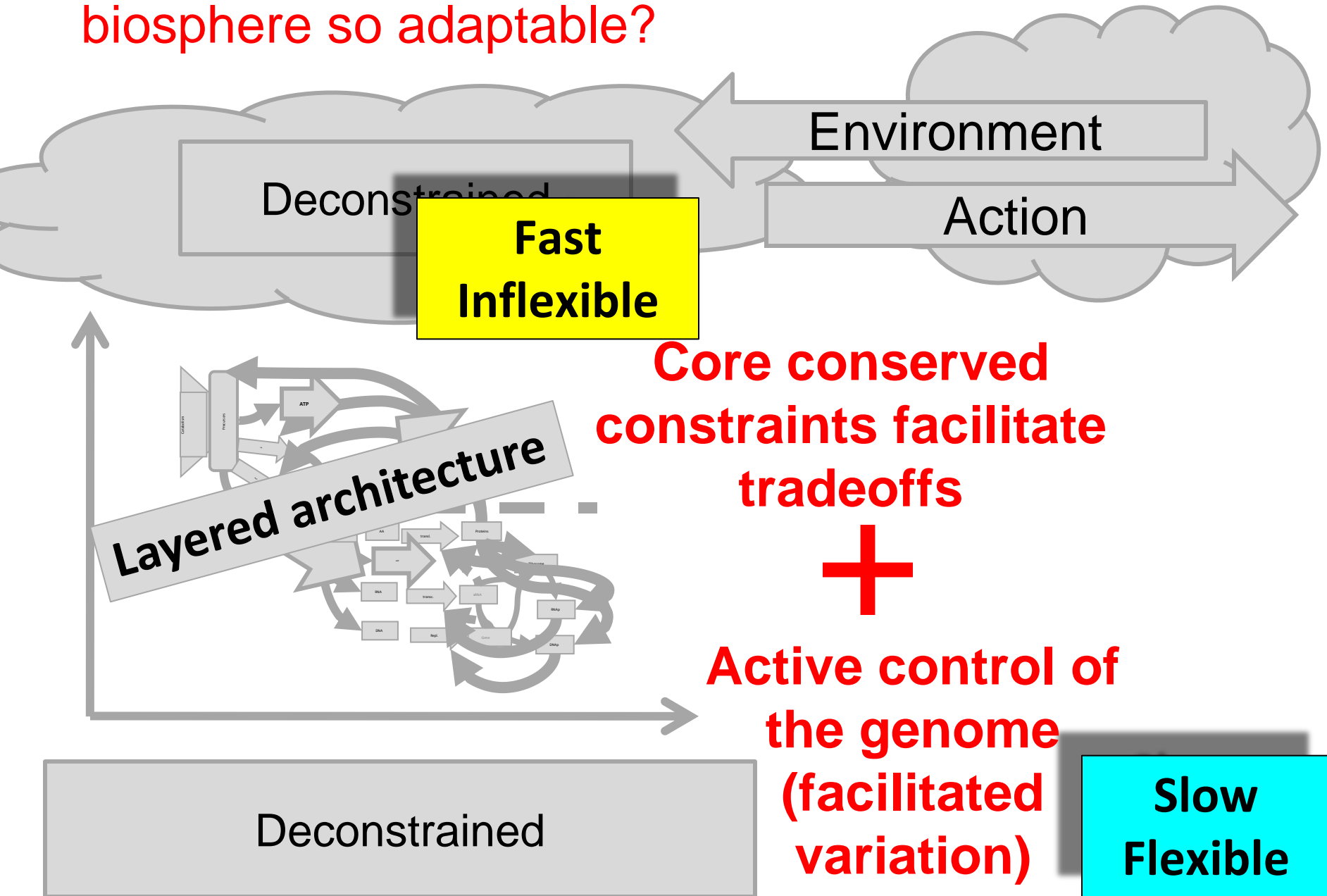
1. NAD
2. Adenylate
3. Carbon
4. phosphate
5. oxygen
6. Oxidized state of metabolites
7. Reduced state of metabolites
8. High energy potential release

$$\frac{1}{\pi} \int_0^{\infty} \ln |S(j\omega)| \left(\frac{z}{z^2 + \omega^2} \right) d\omega \geq \ln \left| \frac{z+p}{z-p} \right|$$

What makes the bacterial biosphere so adaptable?

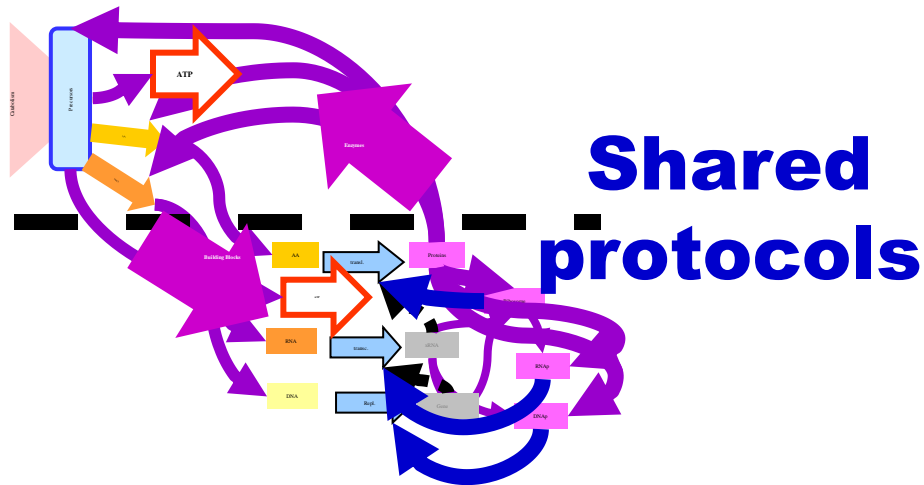


What makes the bacterial biosphere so adaptable?



Deconstrained
Environments

Bacterial biosphere



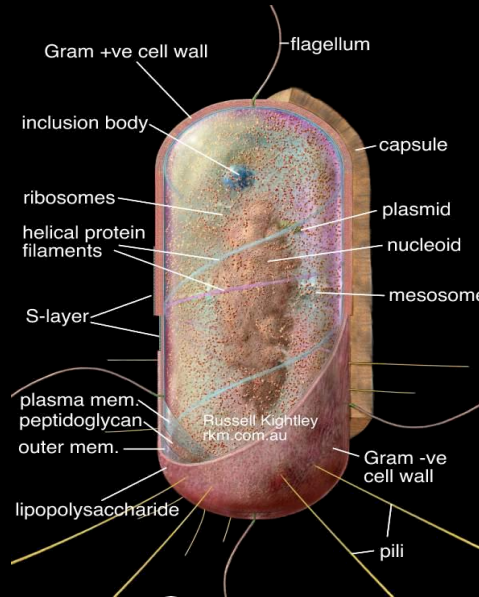
Architecture
=
Constraints
that
Deconstrain

Deconstrained Genomes

System

Architecture =Constraints

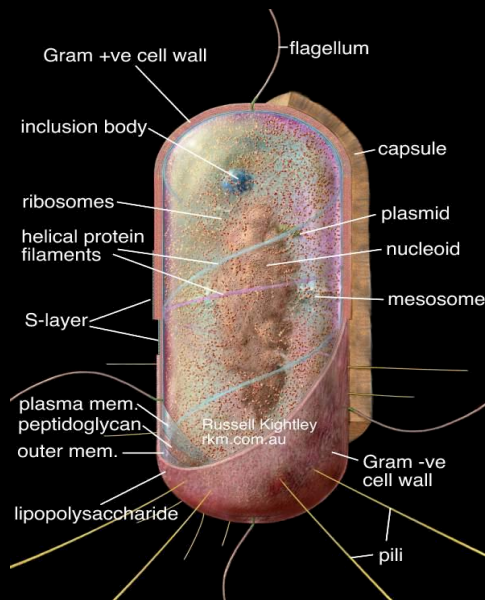
“Emergent”:
“Nontrivial”
consequences
of other
constraints



Protocols

Components

Systems requirements: Survive in hostile environments



Constraints

Components and materials: “Chemistry”

Constrained (“conserved”):

Moieties

1. NAD
2. Adenylate
3. Carbon
4. phosphate
5. oxygen
6. Oxidized state of metabolites
7. Reduced state of metabolites
8. High energy potential release

Constraints

Components and materials:
“Chemistry”

Bacterial biosphere

- carriers: ATP, NADH, etc
- Precursors, ...
- Enzymes
- Translation
- Transcription
- Replication
- ...



Protocols

Architecture = protocols
= “constraints that deconstrain”

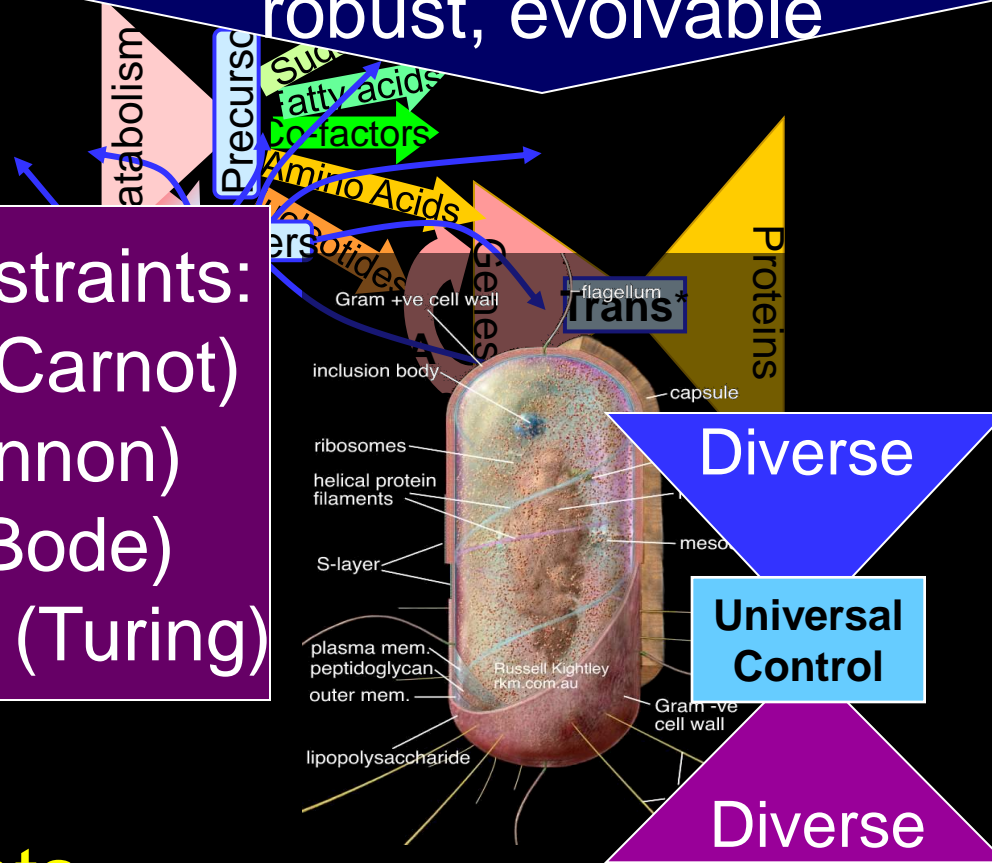
Systems requirements:
functional, efficient,
robust, evolvable

Hard constraints:
Thermo (Carnot)
Info (Shannon)
Control (Bode)
Compute (Turing)

Constraints

Components and materials:
Energy, moieties

Protocols



Systems requirements:
functional, efficient,
robust, evolvable

Hard constraints:

$$\frac{1}{\pi} \int_0^{\infty} \ln |S(j\omega)| \left(\frac{z}{z^2 + \omega^2} \right) d\omega$$

$$\geq \ln \left| \frac{z + p}{z - p} \right|$$

Constraints

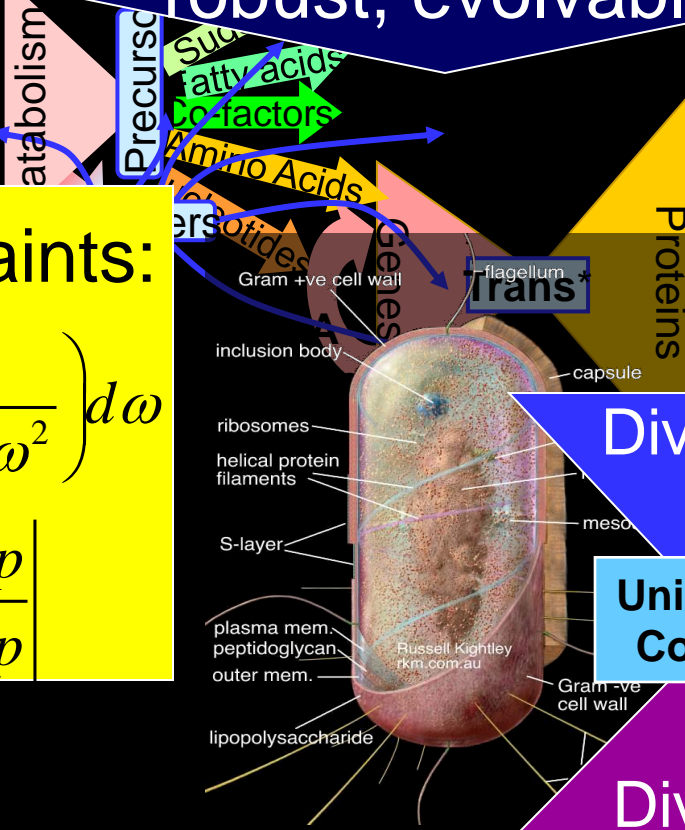
Components and materials:
Energy, moieties

Protocols

Diverse

Universal
Control

Diverse



Viruses' Life History: Towards a Mechanistic Basis of a Trade-Off between Survival and Reproduction among Phages

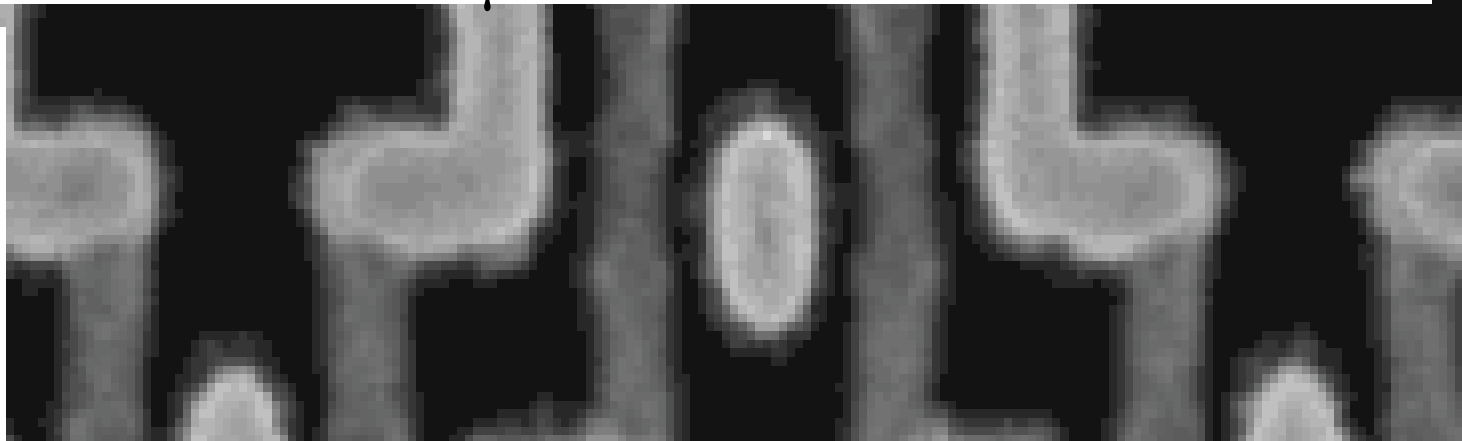
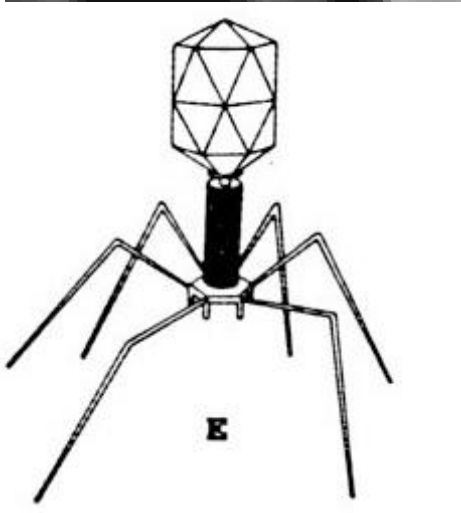
Marianne De Paepe, François Taddei*

Laboratoire de Genetique Moleculaire, Evolutive et Medicale, University of Paris 5, INSERM, Paris, France

July 2006 | Volume 4 | Issue 7 | e193

I recently found this paper, a rare example of exploring an explicit tradeoff between robustness and efficiency. This seems like an important paper but it is rarely cited.

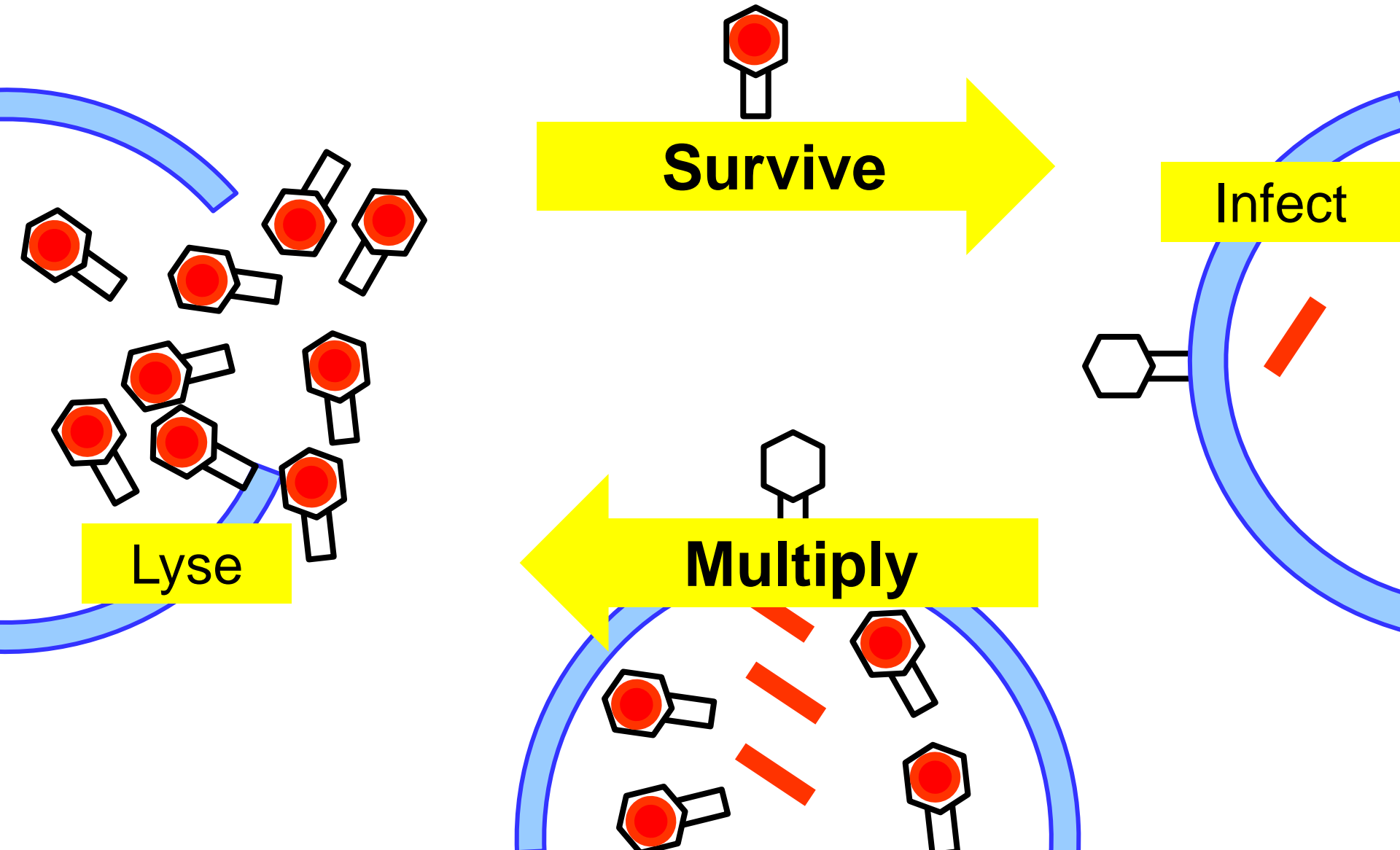
1 μm

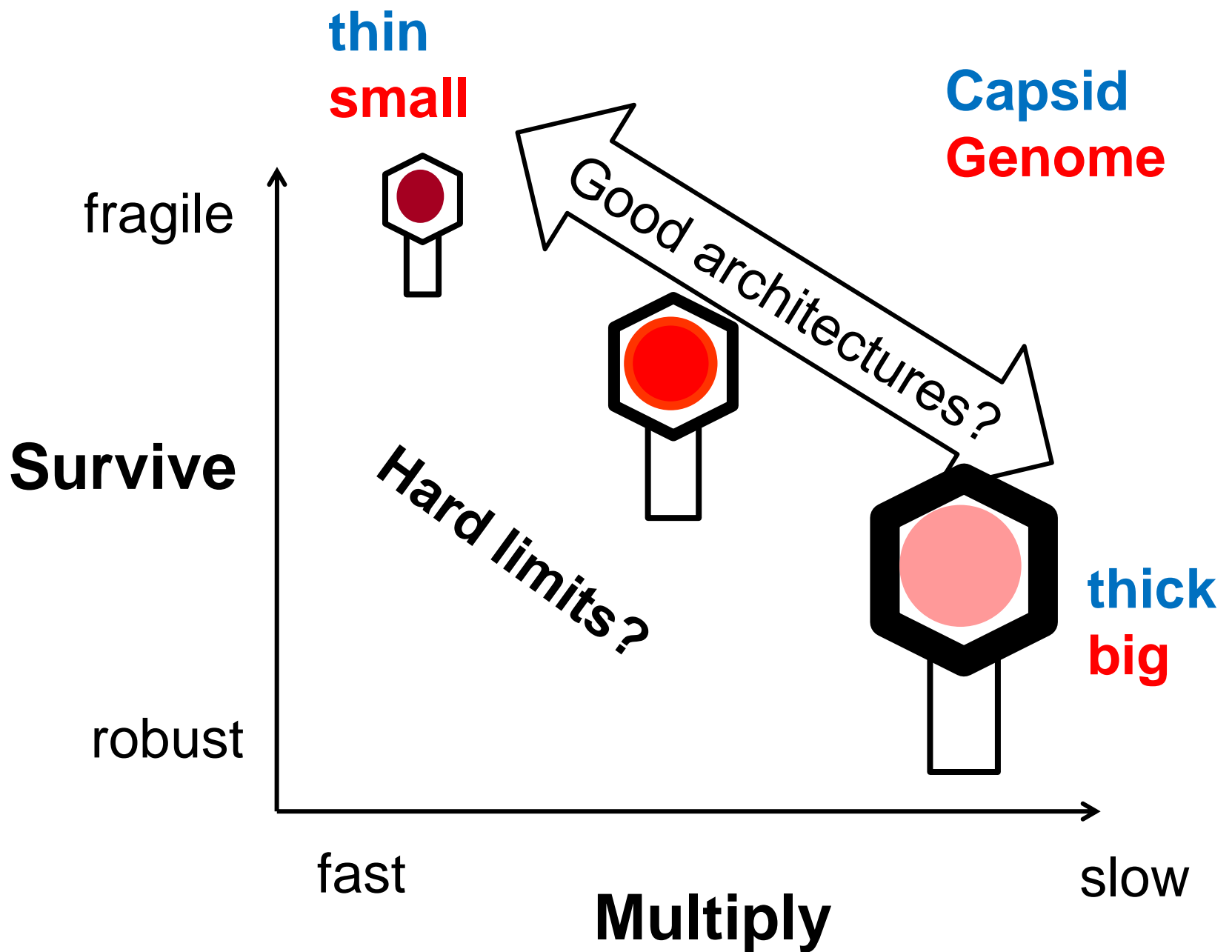


Phage

Bacteria

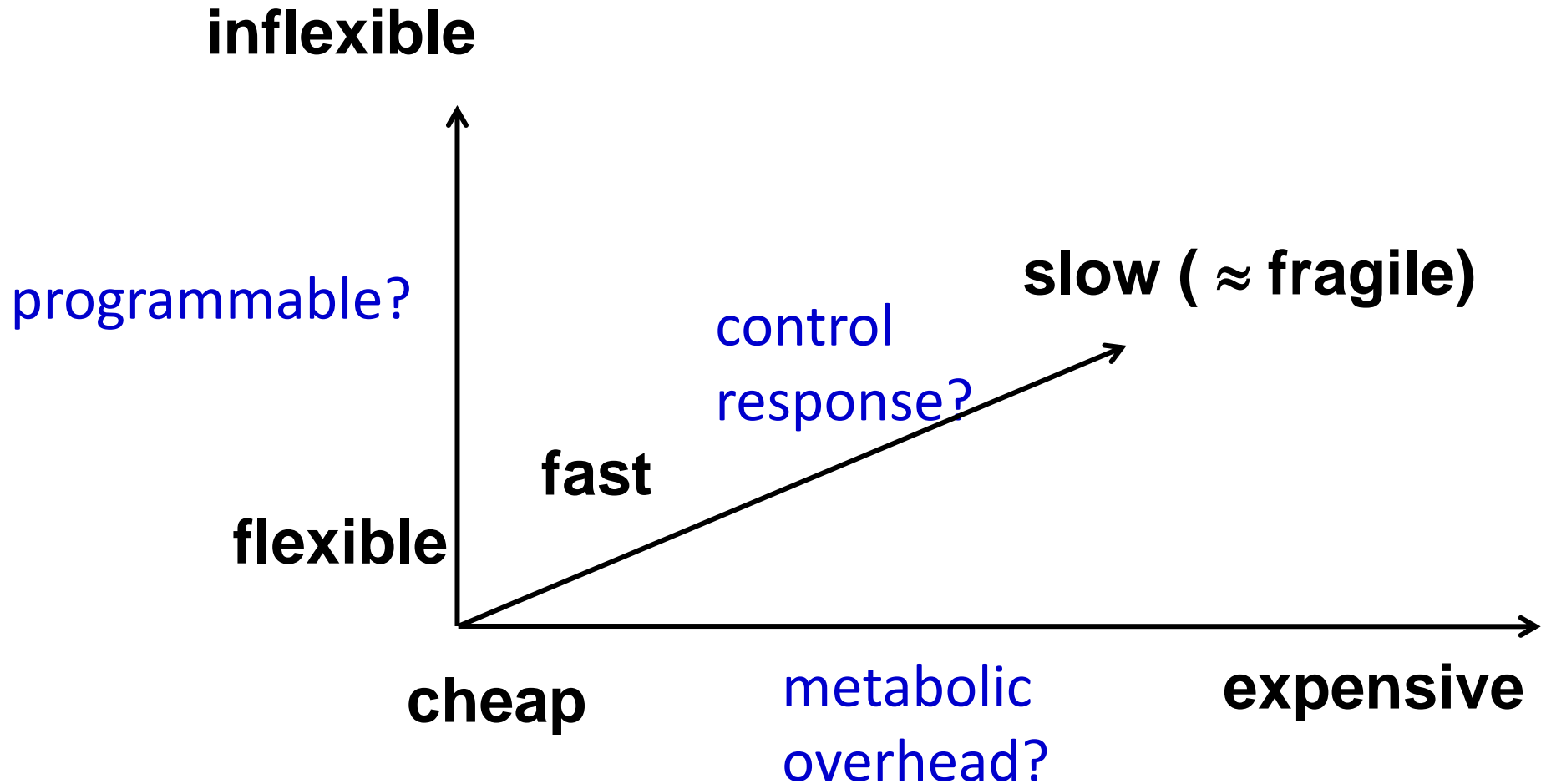
Phage lifecycle





Tradeoffs?

Accurate vs sloppy is now
an implicit dimension of
robust/fragile



Conjecture: human brain tradeoffs dominated by fast vs flexible more than robust vs cheap

1. For hunter/gatherer metabolism is far above basal, and dominated by active muscle
2. Brain homeostasis is a much greater challenge than basal metabolic demands

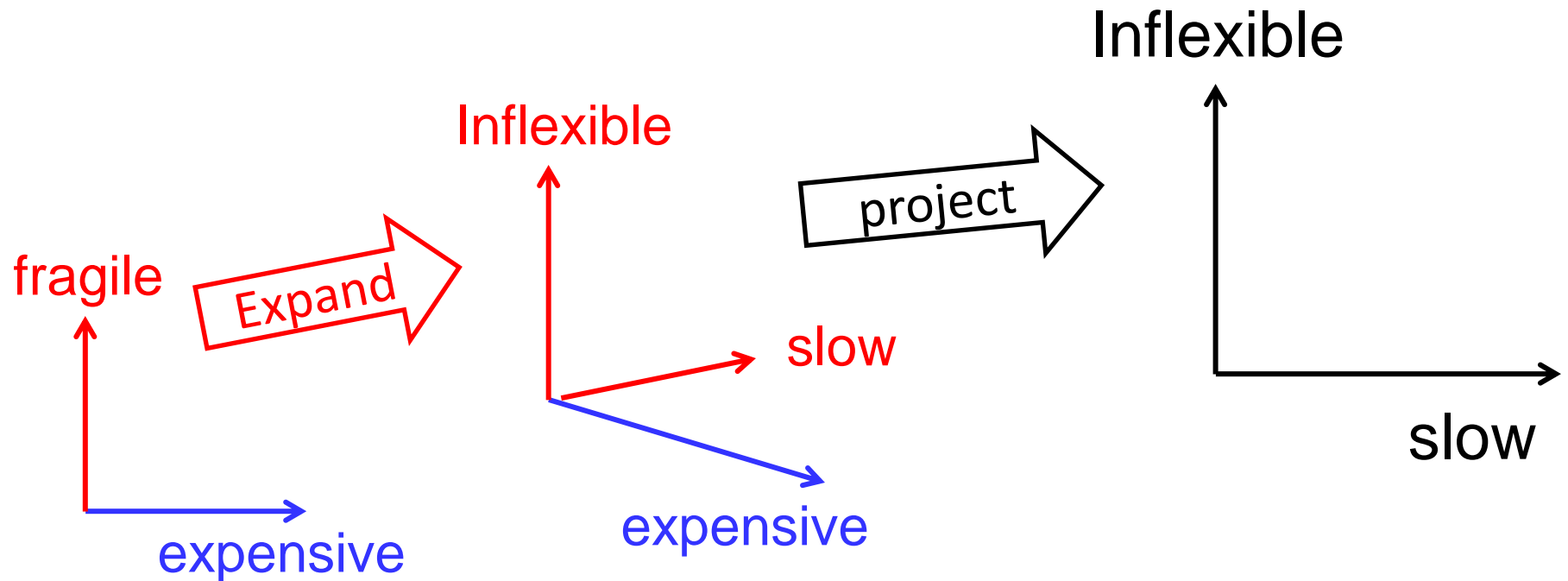
Creates new fragilities in modern lifestyle

Not true for sedentary organisms with limited nutrient diets (e.g. Koala, Panda, ...)

Conjecture: human brain tradeoffs dominated by fast vs flexible more than robust vs cheap

Fragility dimensions with most important tradeoffs:

1. latency/delay/speed of control vs.
2. flexibility/adaptability





**Slow
Flexible**

Consistent tradeoff across
very different systems :

- nervous system
- cell
- computer
(that have some shared
architecture)

Expense is
complicated tradeoff
between

- design effort
- fabrication cost
- energy use
- etc etc

Tradeoff

**Fast
Inflexible**

Slow
Flexible

Prefrontal

Learning

Motor

Sensory

Striatum

Reflex

Fast
Inflexible

Gallistel and King

C.R. Gallistel and
Adam Philip King



Memory and the
Computational Brain

Why Cognitive Science Will Transform Neuroscience

WILEY-BLACKWELL

- Sensori-motor memory potential $\approx \infty$ (Ashby)
- Limits are on **speed** of
 - nerve propagation delays
 - learning
- But control is **never** centralized
- Is there a random access read/write memory?

Horizontal
Meme
Transfer

Very Slow
Process

Slow
Flexible

“Vertical”
App Migration

Prefrontal

Motor

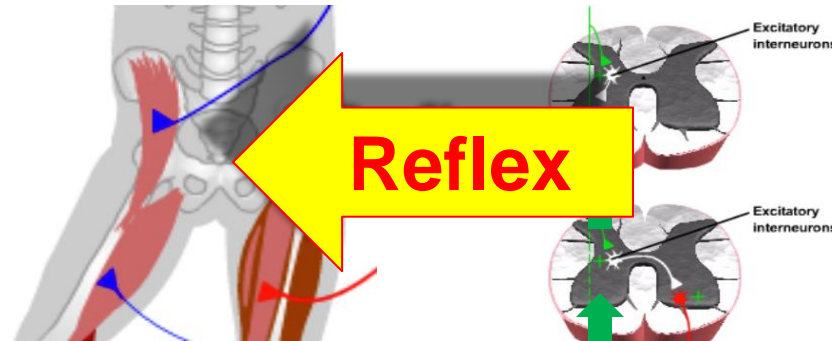
Sensory

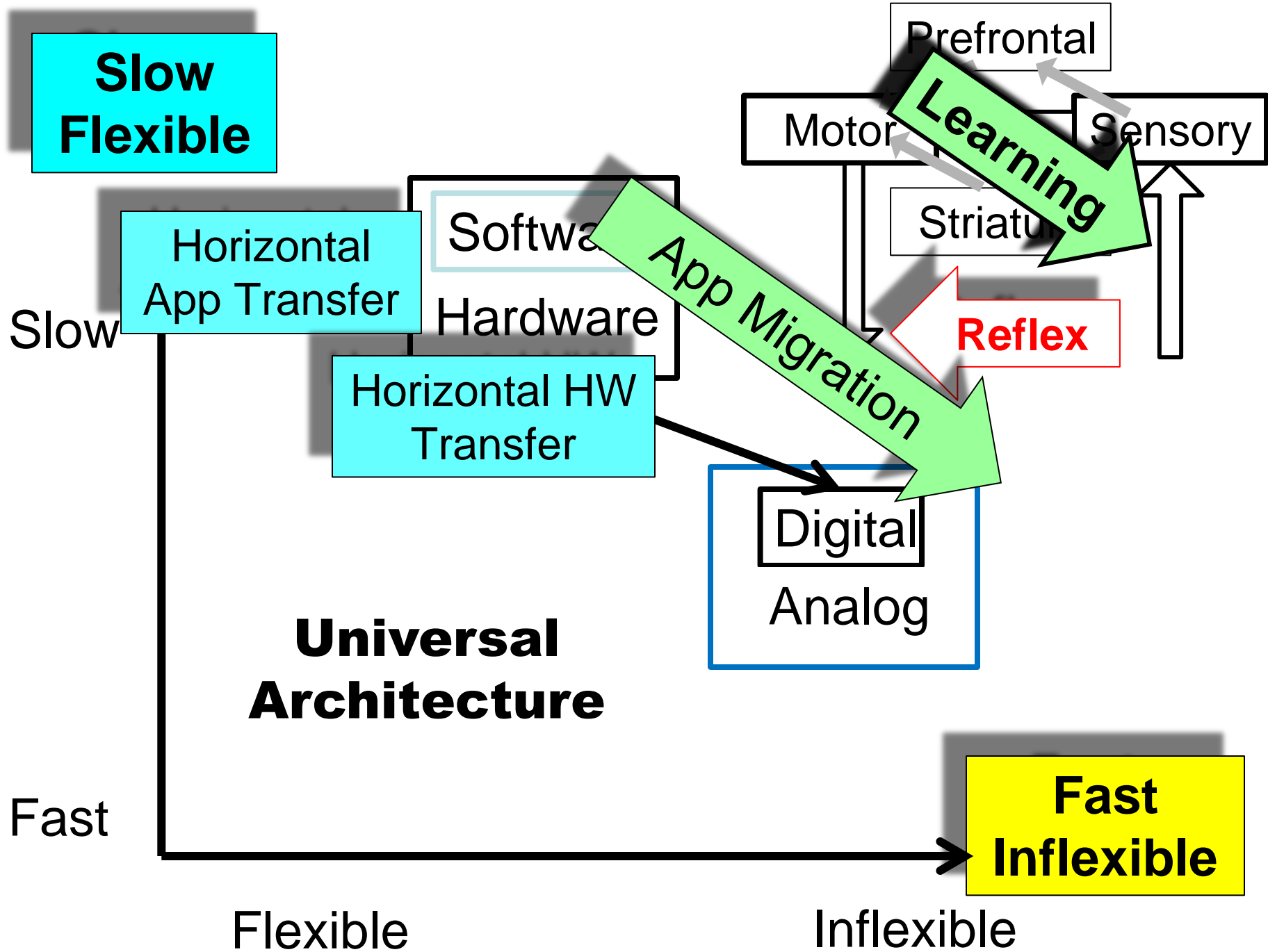
Striatum

Fast
Inflexible

- Acquire
- Translate/
integrate
- Automate

Reflex





**Slow
Flexible**

Software
Hardware

**Techno-
sphere**

Digital
Analog

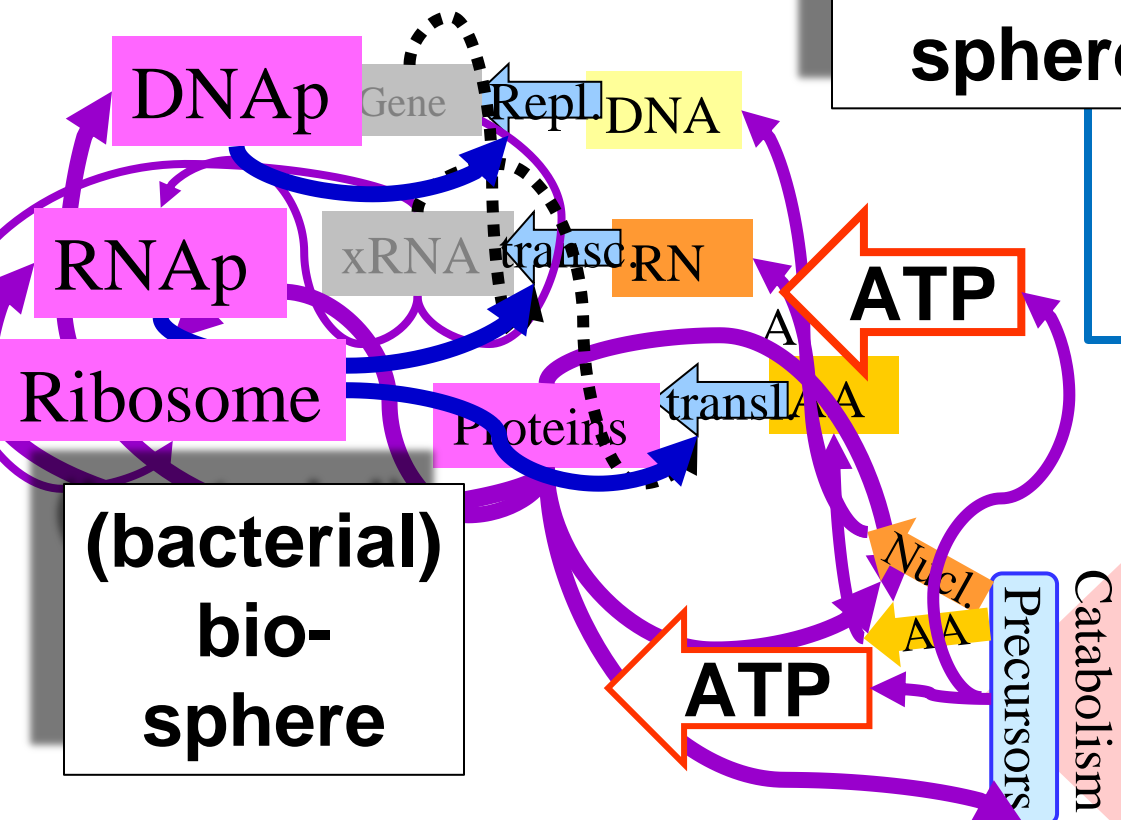
Prefrontal

**Cogni-
sphere**

ory

Motor

Reflex



**(bacterial)
bio-
sphere**

**Fast
Inflexible**

**Flexible/
Adaptable/
Evolvable**

**Horizontal
Meme
Transfer**

Software

Hardware

**Horizontal
App
Transfer**

Digital
Analog

**Depends
crucially on
layered
architecture**

DNAp

Gene

Repl

D

RNAp

xRNA

transc

RN

ATP

A

AA

transl

AA

**Horizontal
Gene
Transfer**

Nucl.
AA

ATP

Precursors

Catabolism

frontal

arning

Sensory

Striatu

Reflex

**Horizontal
Meme
Transfer**

**Horizontal
App
Transfer**

**Horizontal
Gene
Transfer**

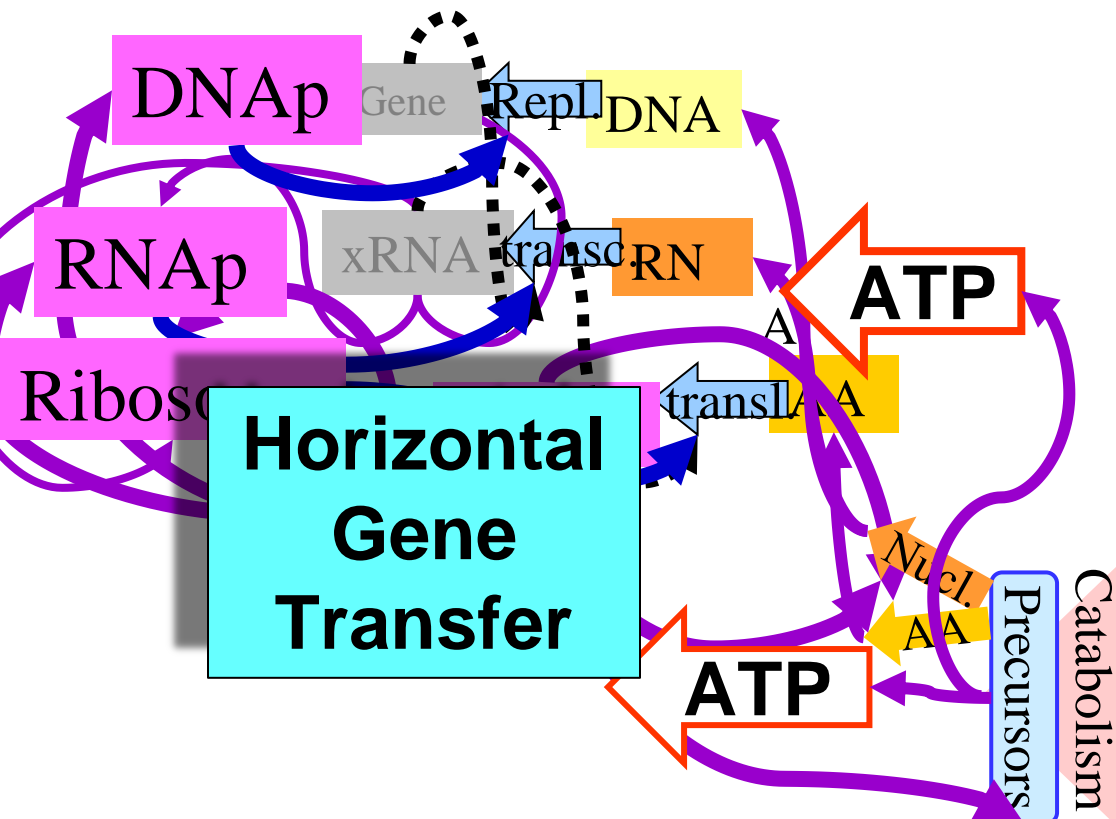
Most

- software and hardware
- new ideas (humans)
- new genes (bacteria)

is acquired by “horizontal” transfer,
though sometimes it is evolved locally

Sequence ~100 E Coli (*not* chosen randomly)

- ~ 4K genes per cell
- ~20K *different* genes in total
- ~ 1K universally shared genes



See slides on
bacterial
biosphere

**Exploiting
layered
architecture**

**Horizontal
Bad Meme
Transfer**

Virus

**Horizontal
Bad App
Transfer**

Fragility?

**Horizontal
Bad Gene
Transfer**

Virus

**Parasites &
Hijacking**

**Build on Turing to show
what is *necessary* to make
this work.**

Depends
crucially on
layered
architecture

- Acquire
- Translate/
integrate
- Automate

Horizontal
Meme
Transfer

**Horizontal
App
Transfer**

Horizontal
Gene
Transfer

Amazingly
Flexible/
Adaptable

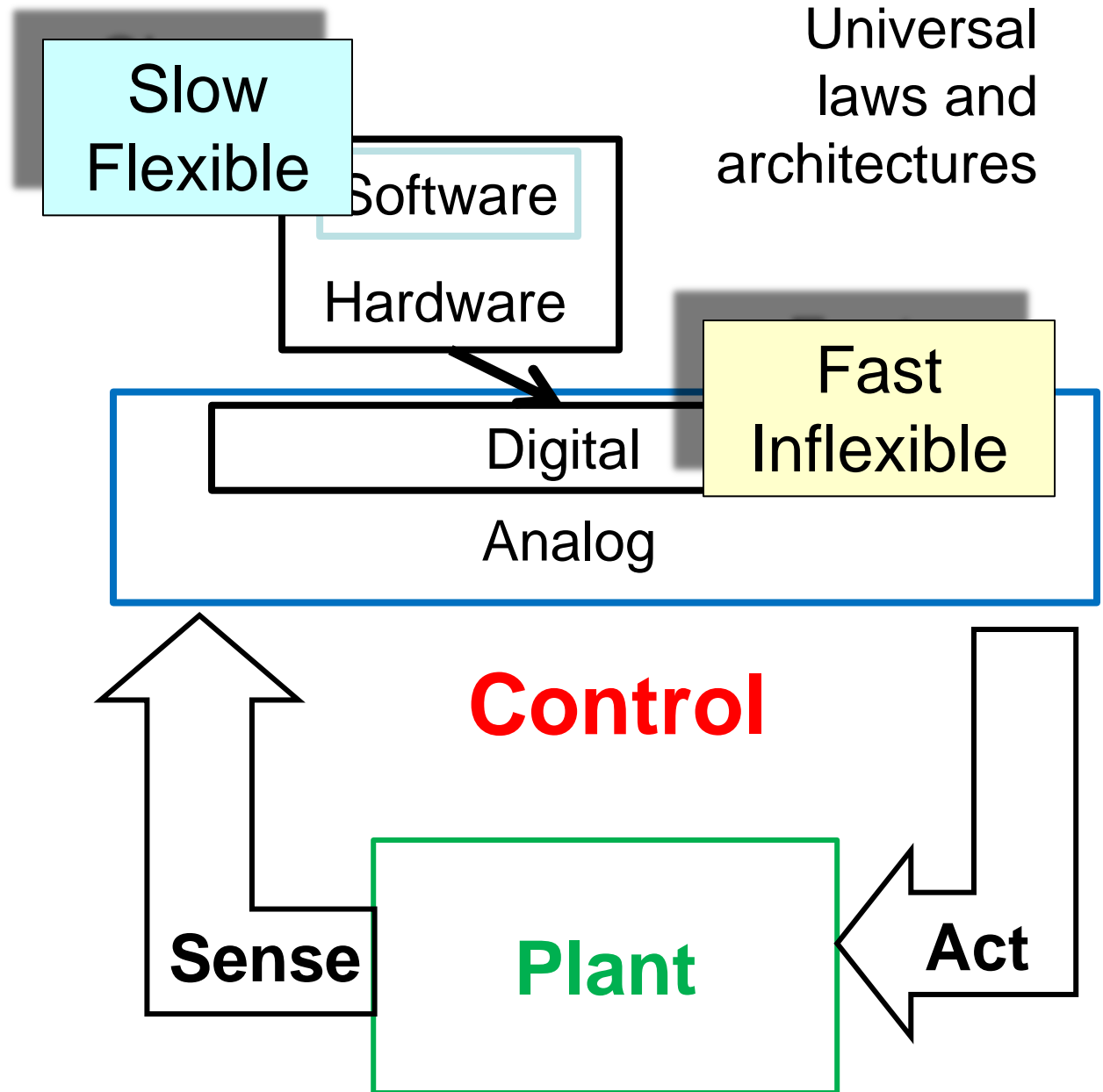
Compute

Turing

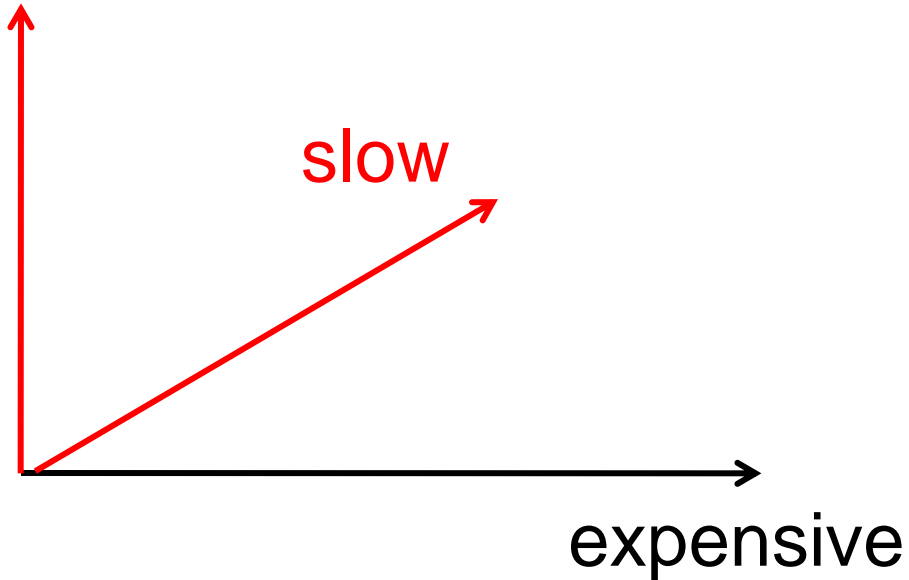
**Delay is
even more
important**

Bode

Control



Inflexible



But efficiency
tradeoffs are
different.

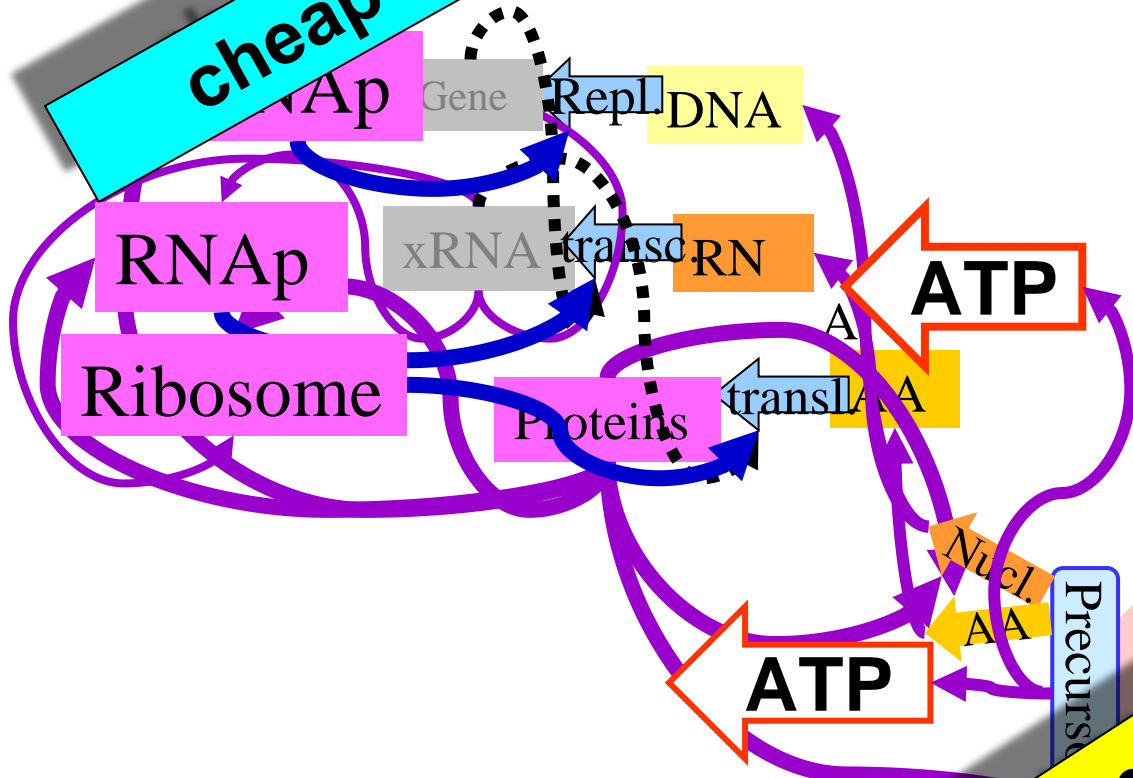
**Slow
Flexible**

ible

cheap

Ap

Gene



ATP

ATP

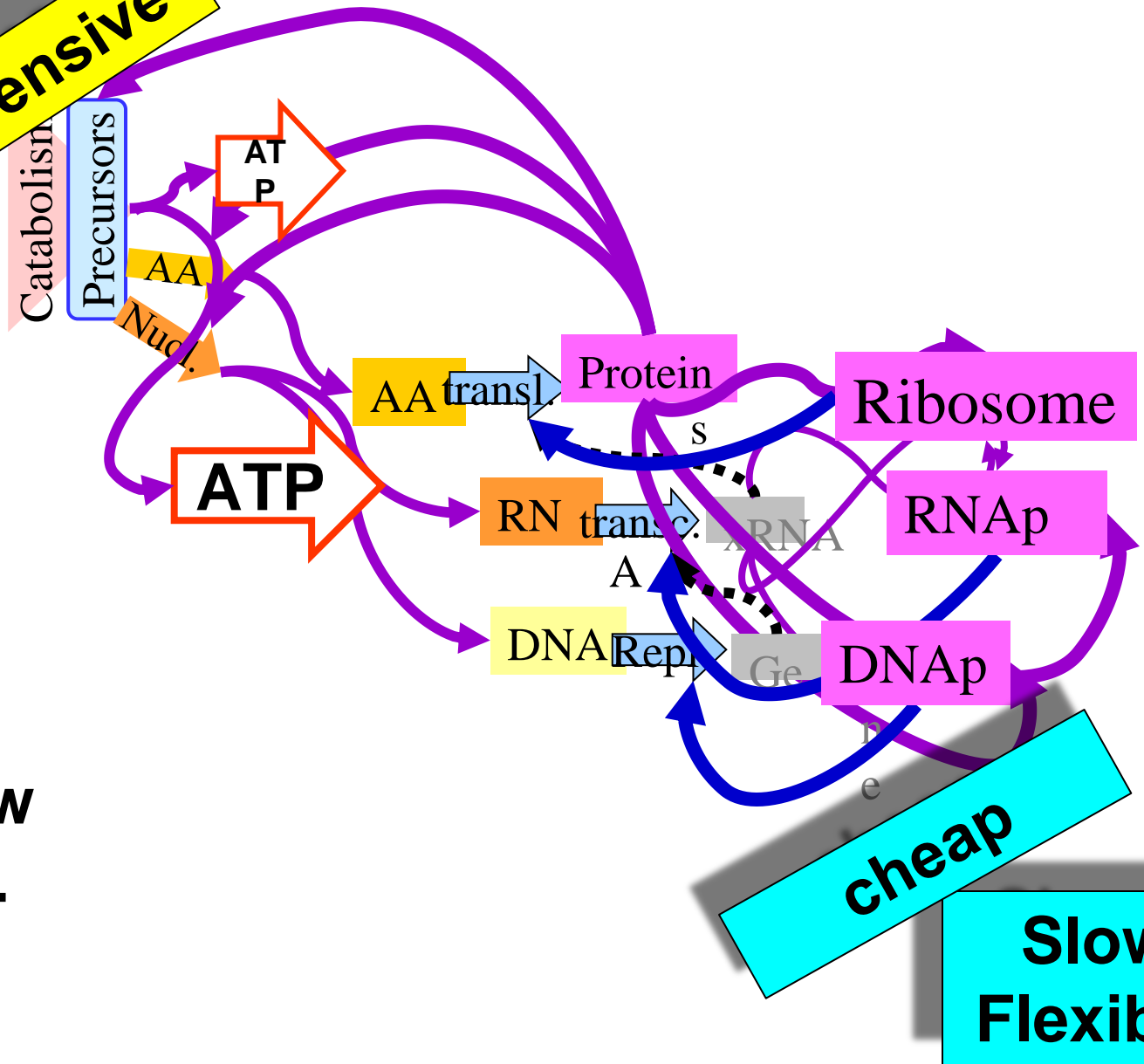
A collage of text elements. The word 'expensive' is prominently displayed in a large, bold, black font on a yellow rectangular background, tilted diagonally. To its left, the word 'Precourse' is written vertically in a blue box. Above it, 'Catabol' is partially visible. To the right, 'Fa' is visible in a yellow box. Other fragments like 'AA' and 'cl' are also present.

**Fast
Inflexible**

Cell metabolic expense lines up nicely

Fast Inflexible

expensive



Usually draw it this way...

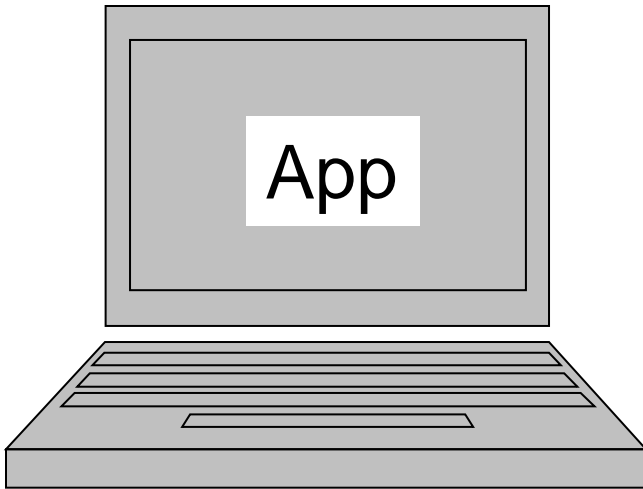
cheap

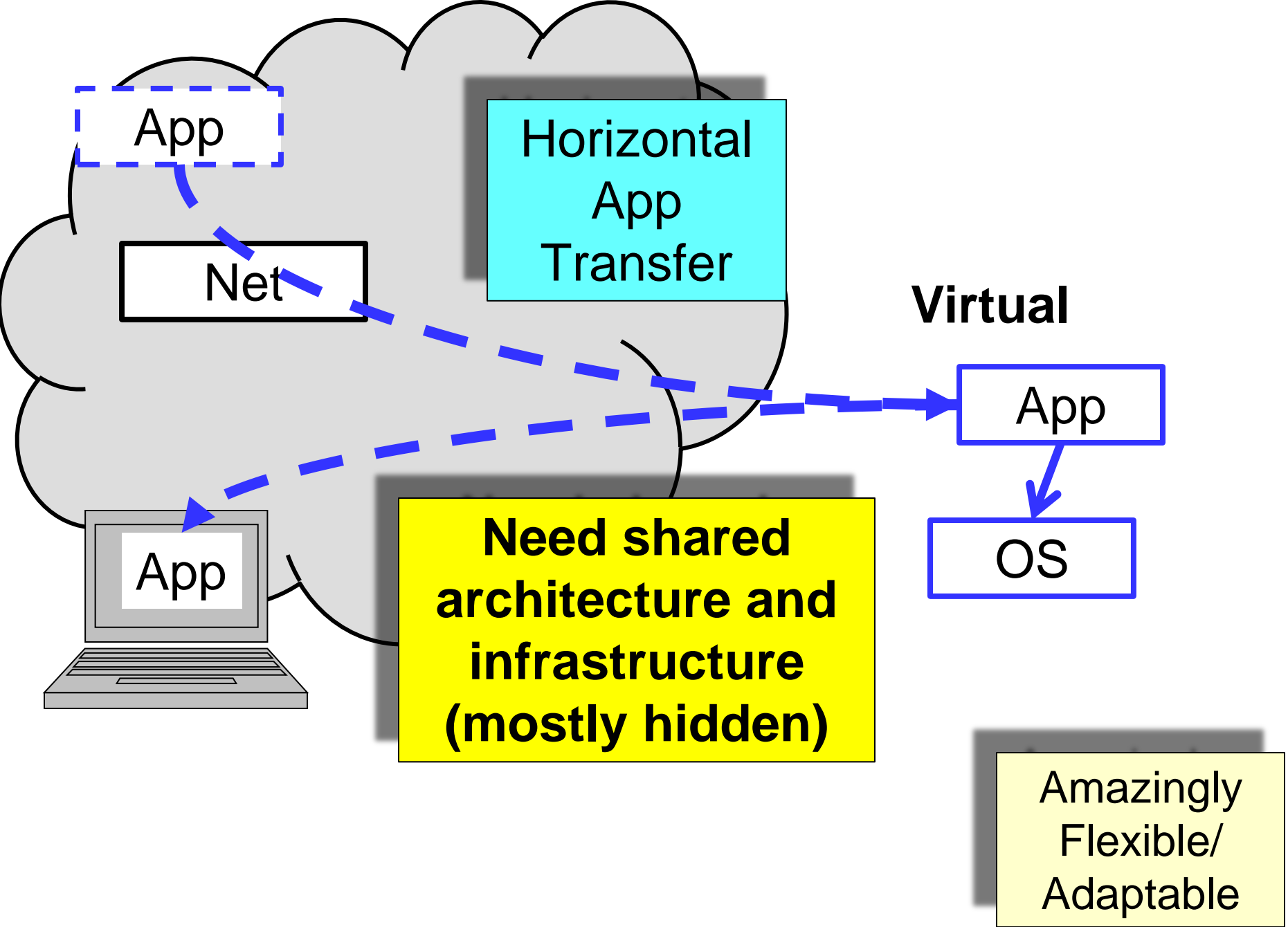
Slow Flexible

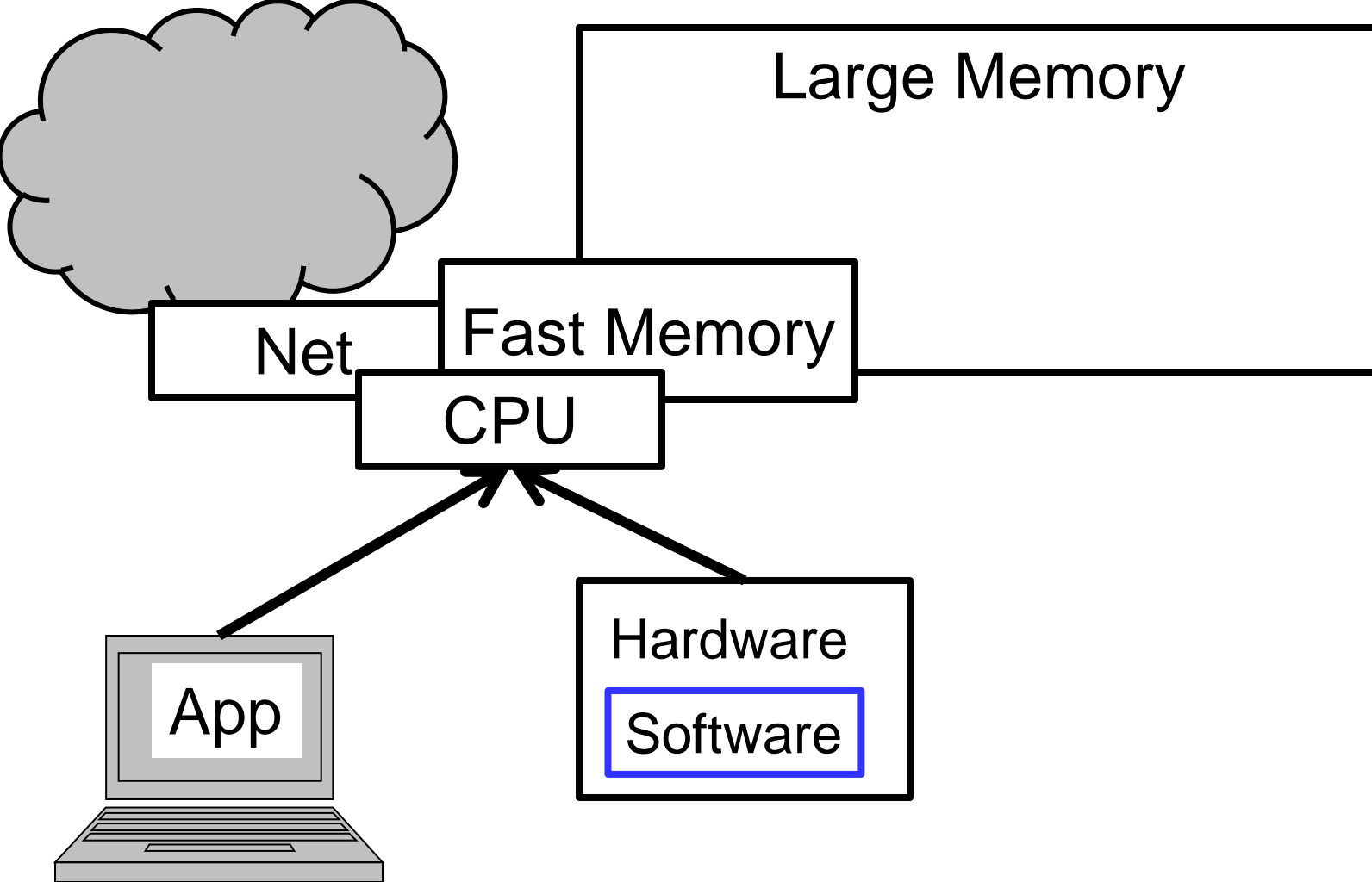
**What you see:
The hardware
interface and
the application
function**



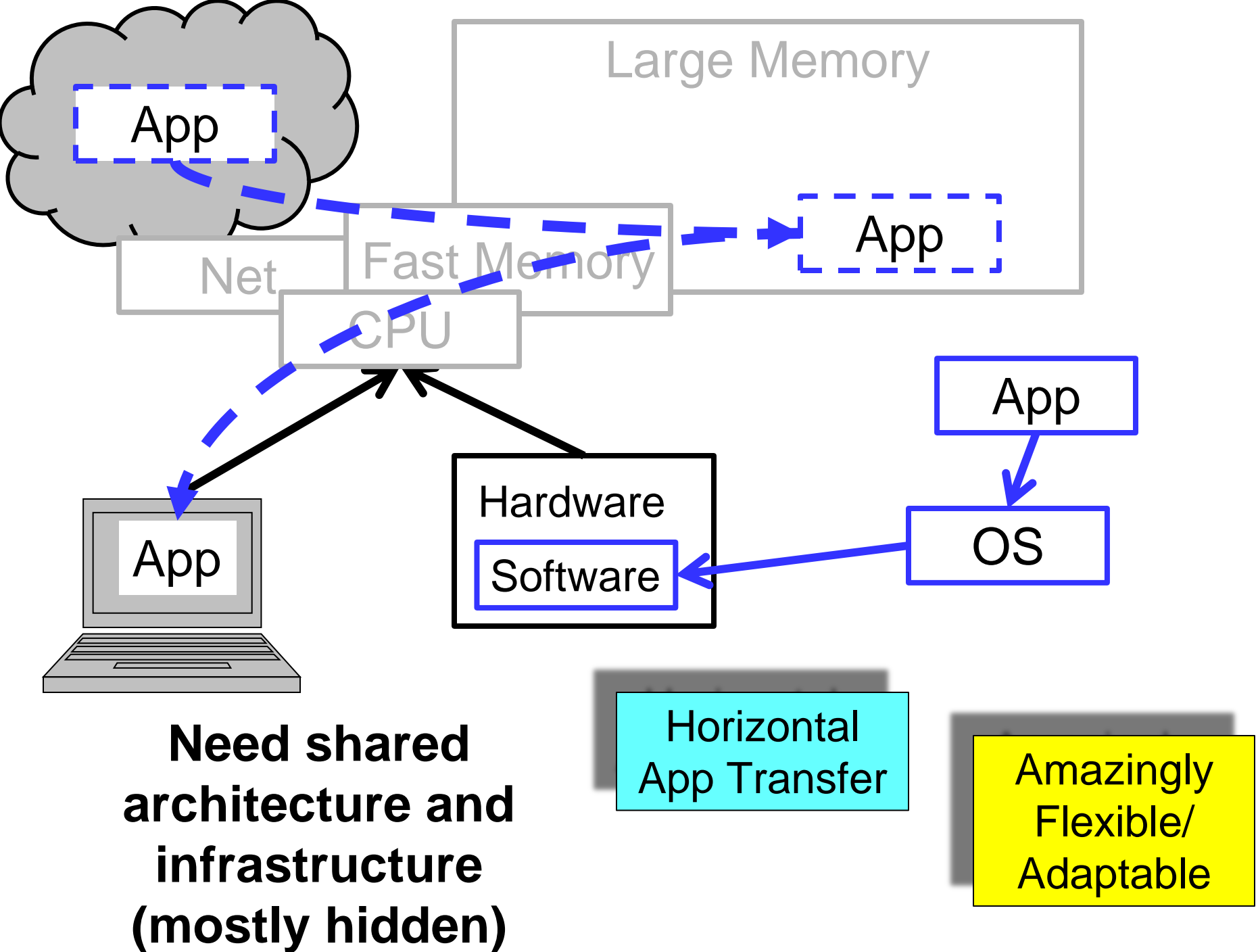
Need shared
architecture and
infrastructure
(mostly hidden)







**Need shared
architecture and
infrastructure
(mostly hidden)**



More Large Memory

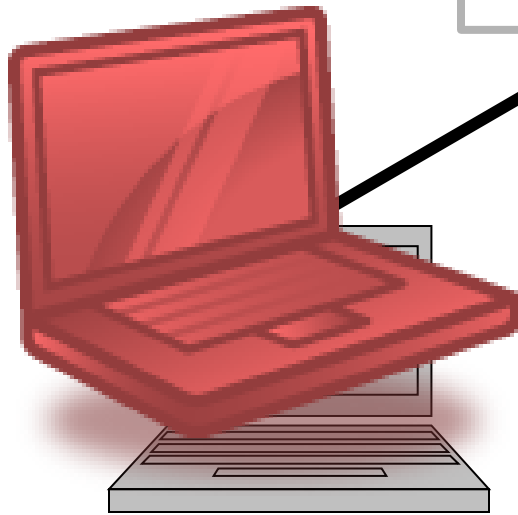
Horizontal
Hardware
Transfer

New I/O

Fast
CPU

OS

New
Hardware



**Need shared
architecture and
infrastructure
(mostly hidden)**

Amazingly
Flexible/
Adaptable

Layered architectures

Essentials

Deconstrained
(Applications)

Few global variables

Don't cross layers

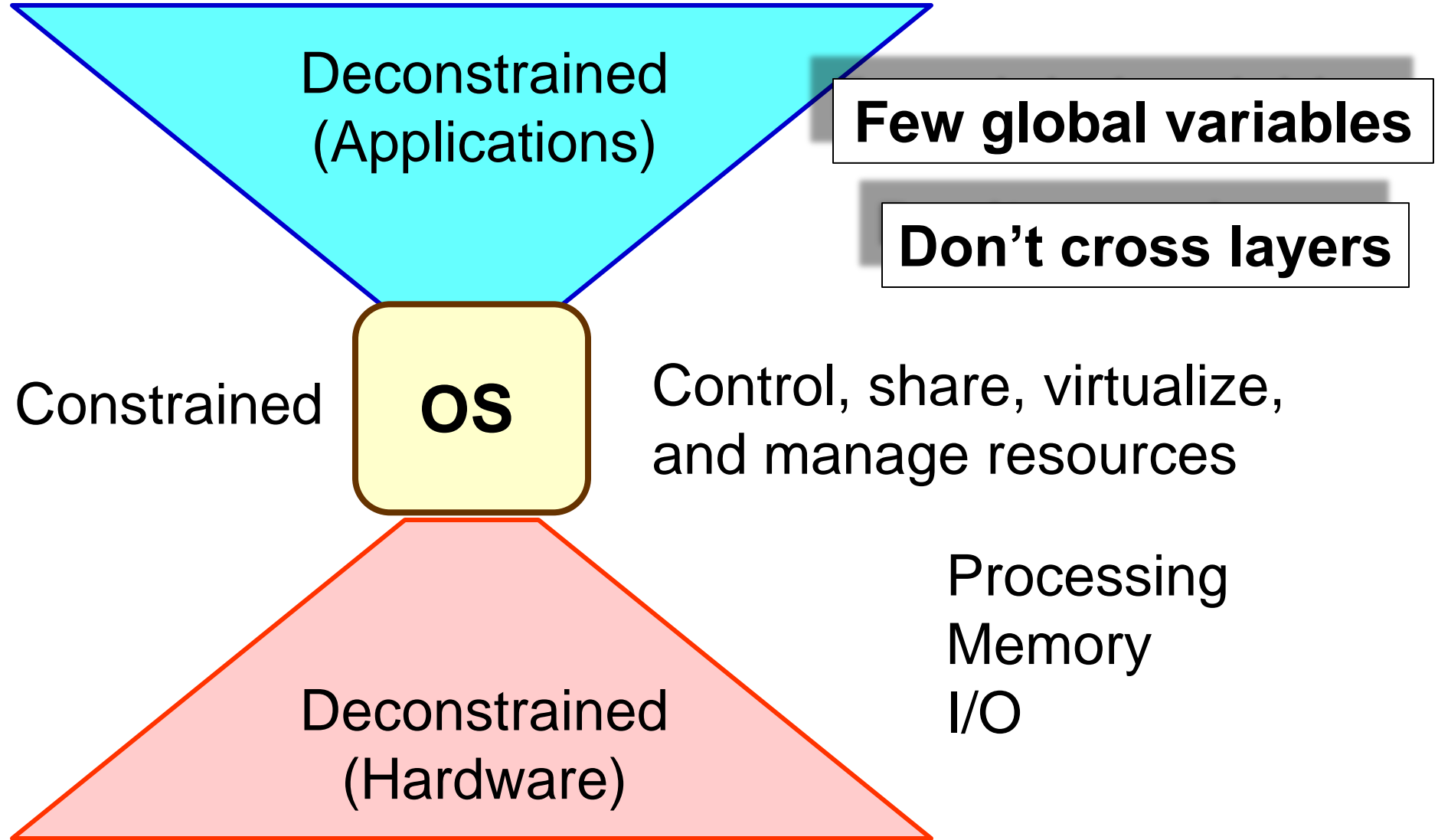
Constrained

OS

Control, share, virtualize,
and manage resources

Processing
Memory
I/O

Deconstrained
(Hardware)



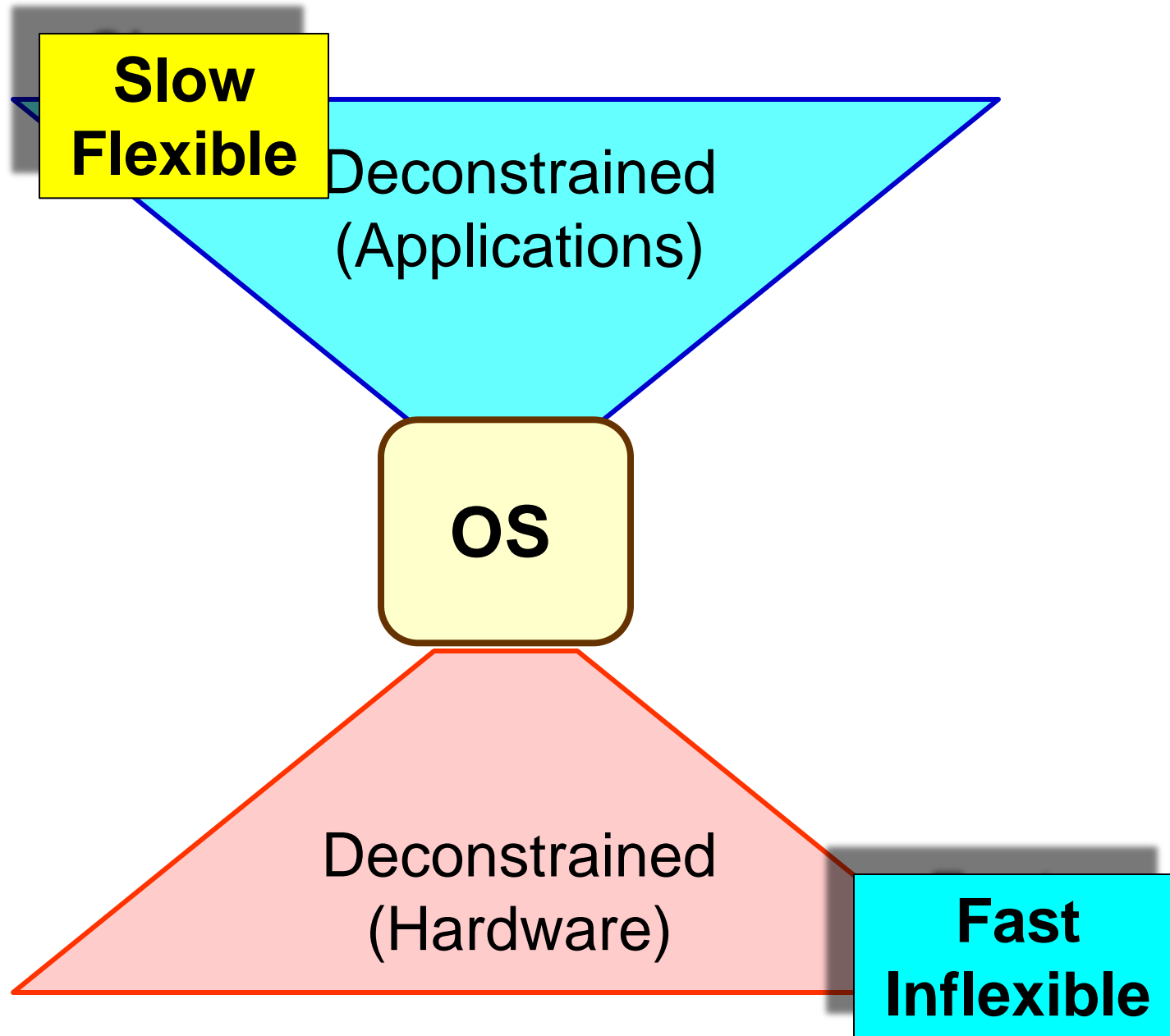
**Slow
Flexible**

Deconstrained
(Applications)

OS

Deconstrained
(Hardware)

**Fast
Inflexible**



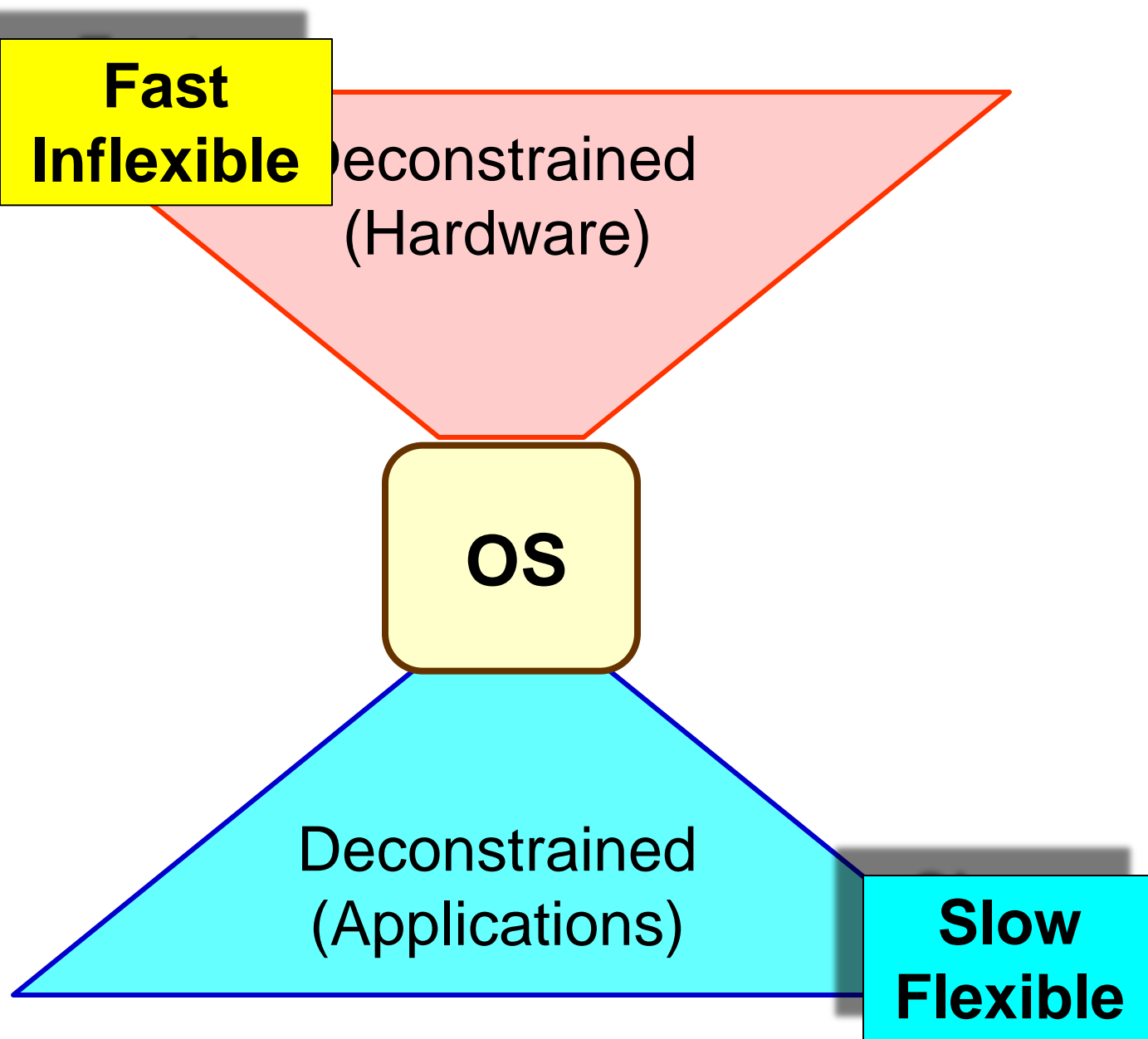
**Fast
Inflexible**

Deconstrained
(Hardware)

OS

Deconstrained
(Applications)

**Slow
Flexible**



Slow
Flexible

Very Slow
Process

Horizontal
App Transfer

Software
Hardware

“Vertical”
App Migration

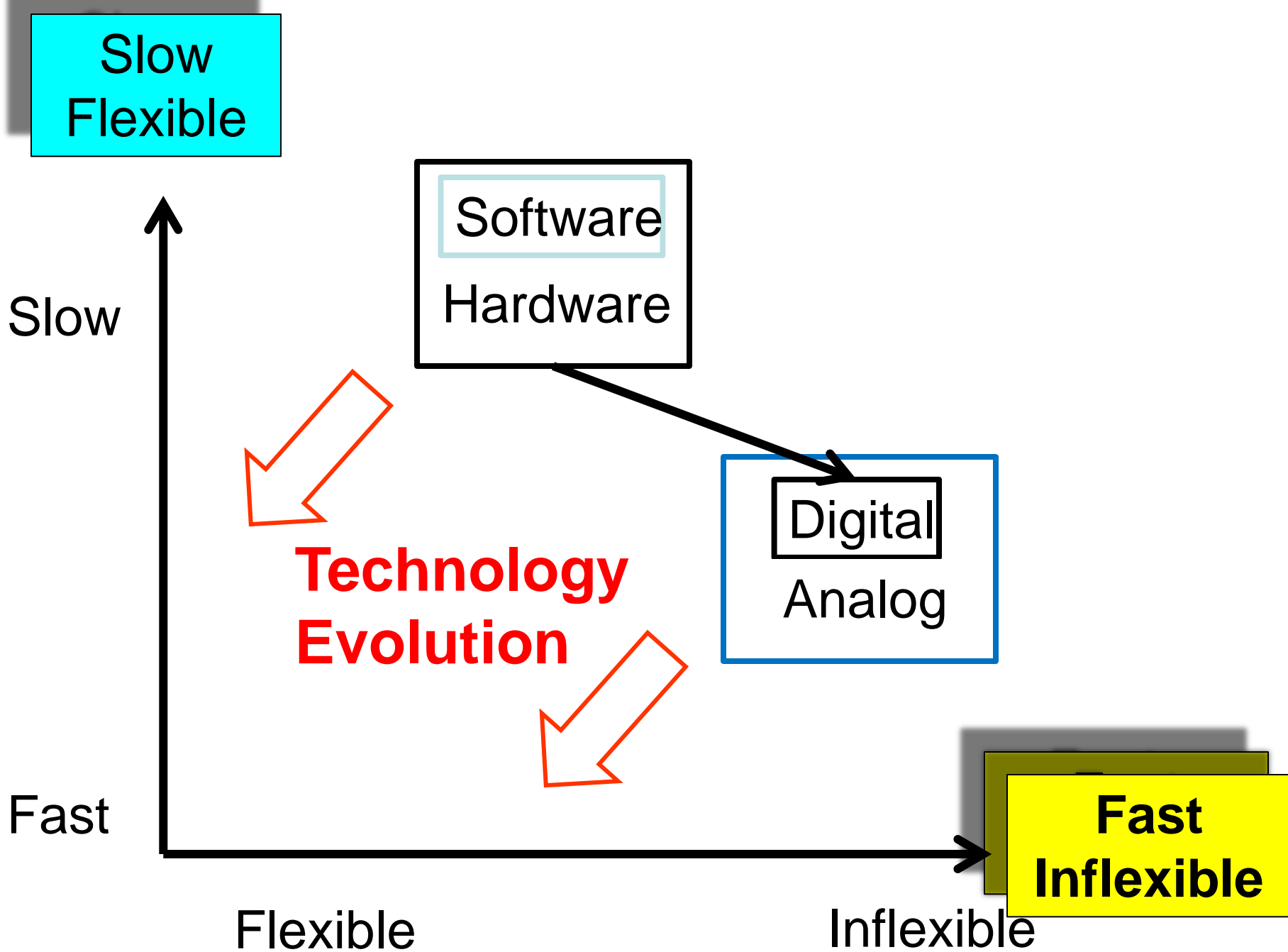
Digital
Analog

Fast
result

Fast
Inflexible

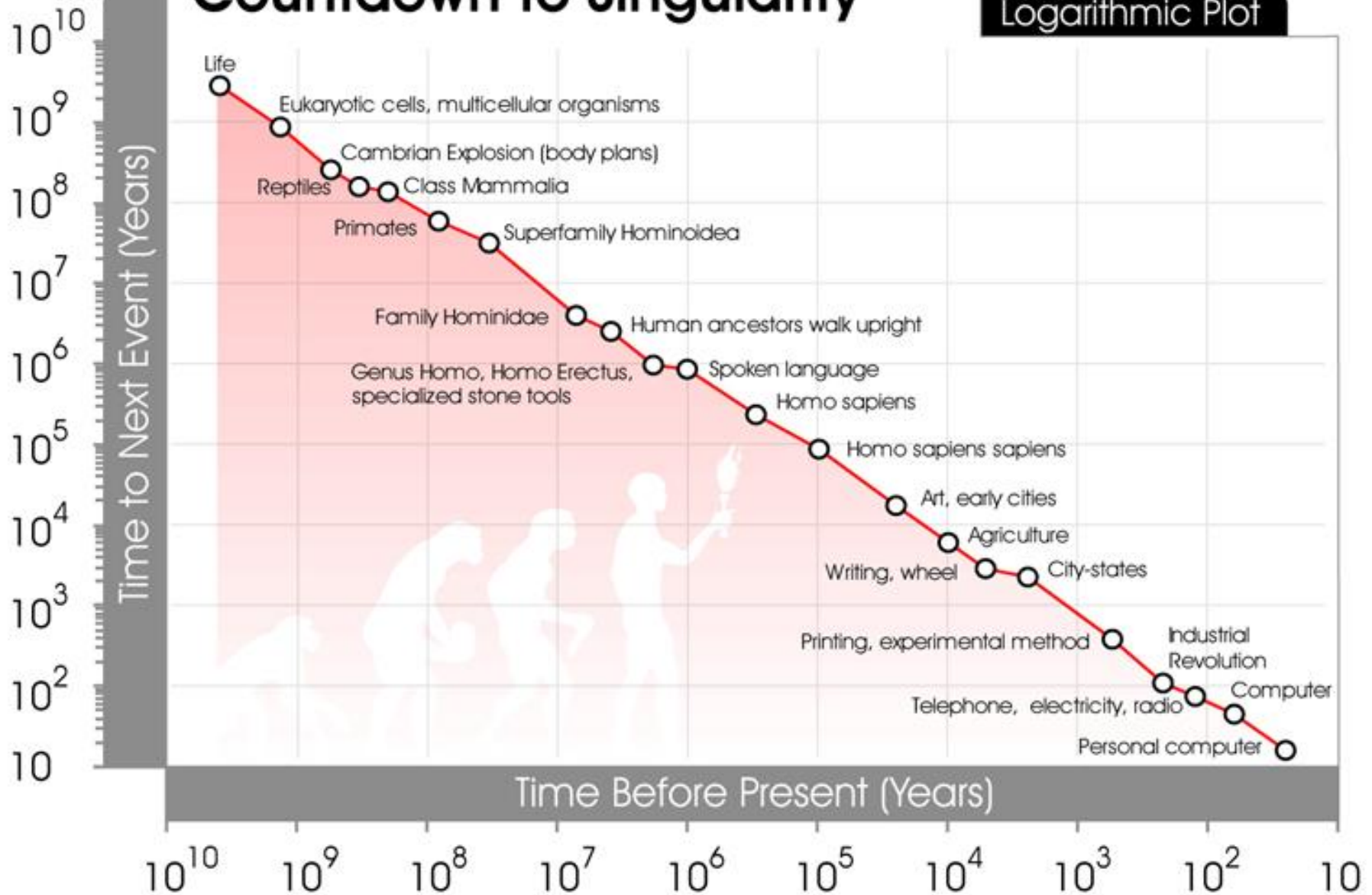
Tradeoff across
multiple layers

- Distributed
- Analog
- ASICs
- FPGA
- ...
- Compiled
- Interpreted



Countdown to Singularity

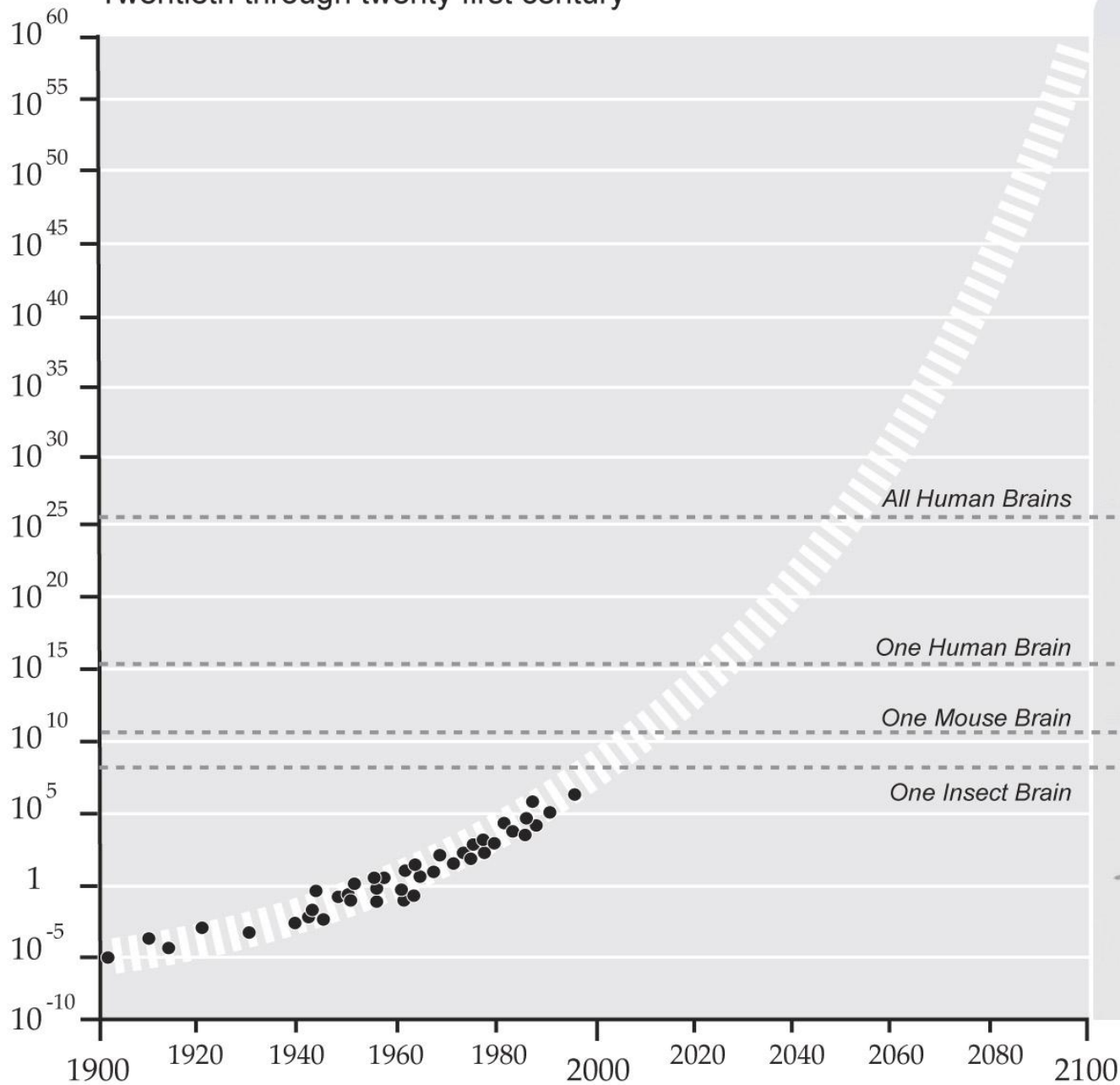
Logarithmic Plot



Exponential Growth of Computing

Twentieth through twenty first century

Calculations per Second per \$1,000



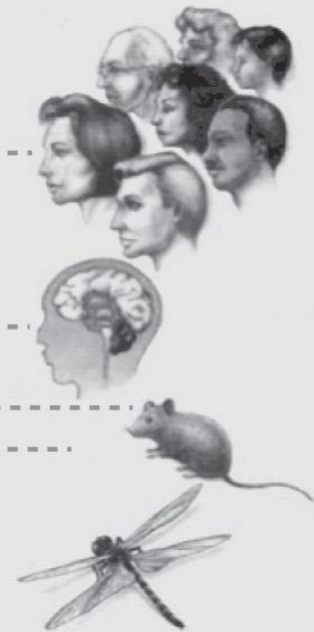
Logarithmic Plot

All Human Brains

One Human Brain

One Mouse Brain

One Insect Brain



THE END OF THEORY

Scientists have always relied on hypothesis and experimentation. Now, in the era of massive data, there's a better way.

1 TERABYTE

A \$200 HARD DRIVE
THAT HOLDS
260,000 SONGS

460 TERABYTES

ALL THE DIGITAL
WEATHER
DATA COMPILED
BY THE NATIONAL
CLIMATIC DATA
CENTER

530 TERABYTES

ALL THE VIDEOS
ON YOUTUBE

THE END OF THEORY

SCIENTISTS HAVE ALWAYS
RELIED ON HYPOTHESIS
AND EXPERIMENTATION.
NOW, IN THE ERA OF
MASSIVE DATA, THERE'S
A BETTER WAY.
BY CHRIS ANDERSON

"ALL MODELS ARE WRONG, BUT
some are useful."
So proclaimed statistician George
Box 30 years ago, and he was right. But
what choice did we have? Only models,
from cosmological equations to
theories of human behavior, seemed to
be able to consistently, if imperfectly,
explain the world around us. Until now.
Today companies like Google, which
have grown up in an era of massively

abundant data, don't have to settle for
wrong models. Indeed, they don't have
to settle for models at all.

Sixty years ago, digital computers
made information readable. Twenty
years ago, the Internet made it reach-
able. Ten years ago, the first search
engine crawlers made it a single data-
base. Now Google and like-minded
companies are sifting through the
most measured one in history, treat-

ing this massive corpus as a labora-
tory of the human condition. They are
the children of the Petabyte Age.
The Petabyte Age is different
because more is different. Kilobytes
were stored on floppy disks. Mega-
bytes were stored on hard disks.
Terabytes were stored in disk arrays.
Petabytes are stored in the cloud.

As we moved along that progression,
we went from the folder analogy to
the file cabinet analogy to the library
analogy to—well, at petabytes we
ran out of organizational analogies.

At the petabyte scale, information
is not a matter of simple three- and
four-dimensional taxonomy and order
but of dimensionally agnostic statis-
tics. It calls for a completely different
approach, one that requires us to
lose the tether of data as something

that can be visualized in its totality. It
forces us to view data mathematically
first and establish a context for it later.
For instance, Google conquered the
advertising world with nothing more
than applied mathematics. It didn't
pretend to know anything about the
culture and conventions of advertis-
ing—it just assumed that better data,
with better analytical tools, would win
the day. And Google was right.

Google's founding philosophy is
that we don't know why this page
is better than that one: If the statis-
tics of incoming links say it is, that's

good enough. No semantic or causal
analysis is required. That's why
Google can translate languages with-
out actually "knowing" them (given
equal corpus data, Google can trans-
late Klingon into Farsi as easily as it
can translate French into German).
And why it can match ads to content
without any knowledge or assump-
tions about the ads or the content.

Speaking at the O'Reilly Emerg-
ing Technology Conference this
past March, Peter Norvig, Google's
research director, offered an update

to George Box's maxim: "All models
are wrong, and increasingly you can
succeed without them."

This is a world where massive
amounts of data and applied mathe-
matics replace every other tool
that might be brought to bear. Out
with every theory of human behavior,
from linguistics to sociology. Forget
taxonomy, ontology, and psychology.
Who knows why people do what they
do? The point is they do it, and we
can track and measure it with unre-
precedented fidelity. With enough data,
the numbers speak for themselves.

The big target here isn't advertis-
ing, though. It's science. The sci-
entific method is built around testable
hypotheses. These models, for the
most part, are systems visualized in
the minds of scientists. The models are
then tested, and experiments confirm

"All models are wrong, and increasingly
you can succeed without them."

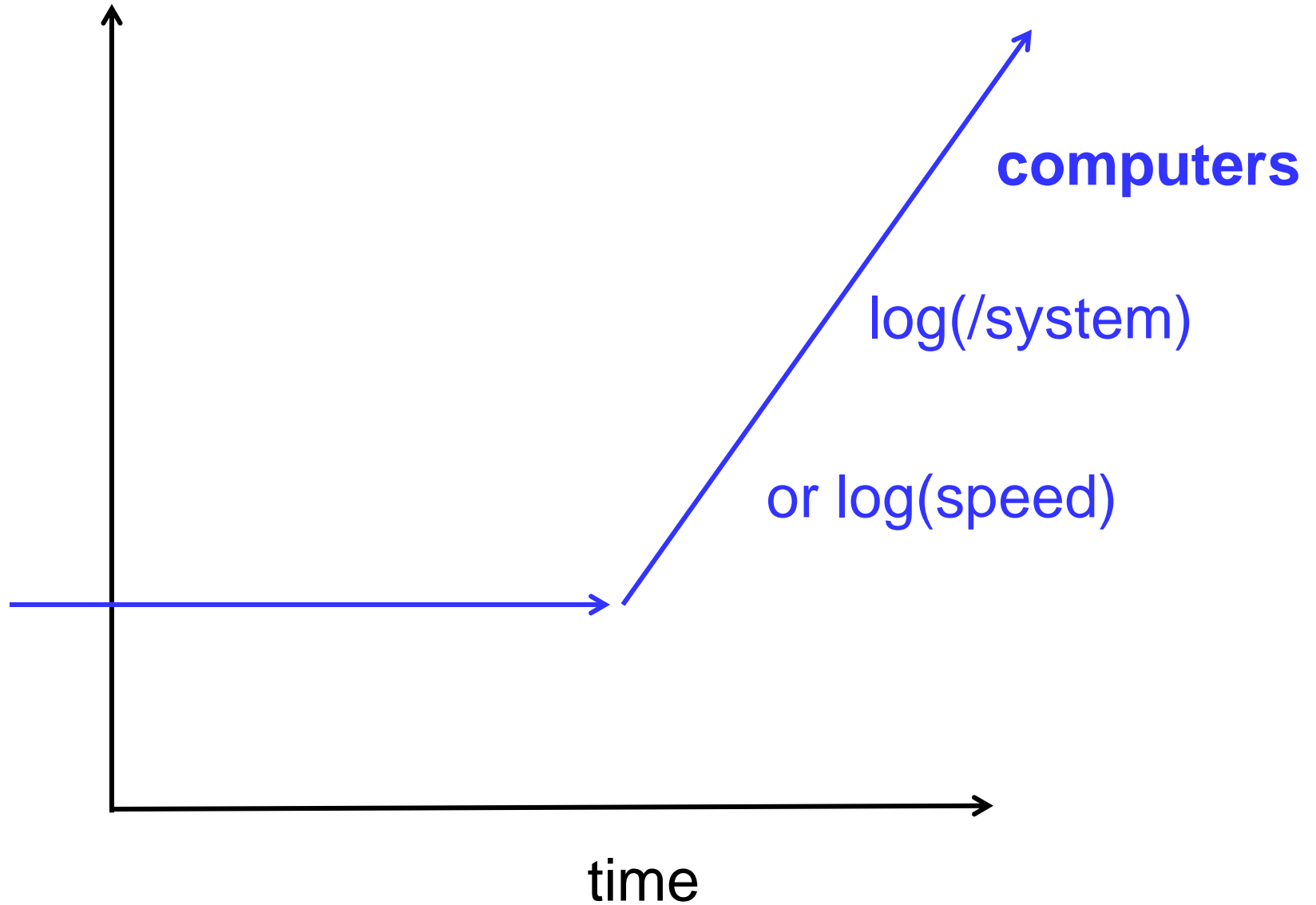
Save our
children,
stop



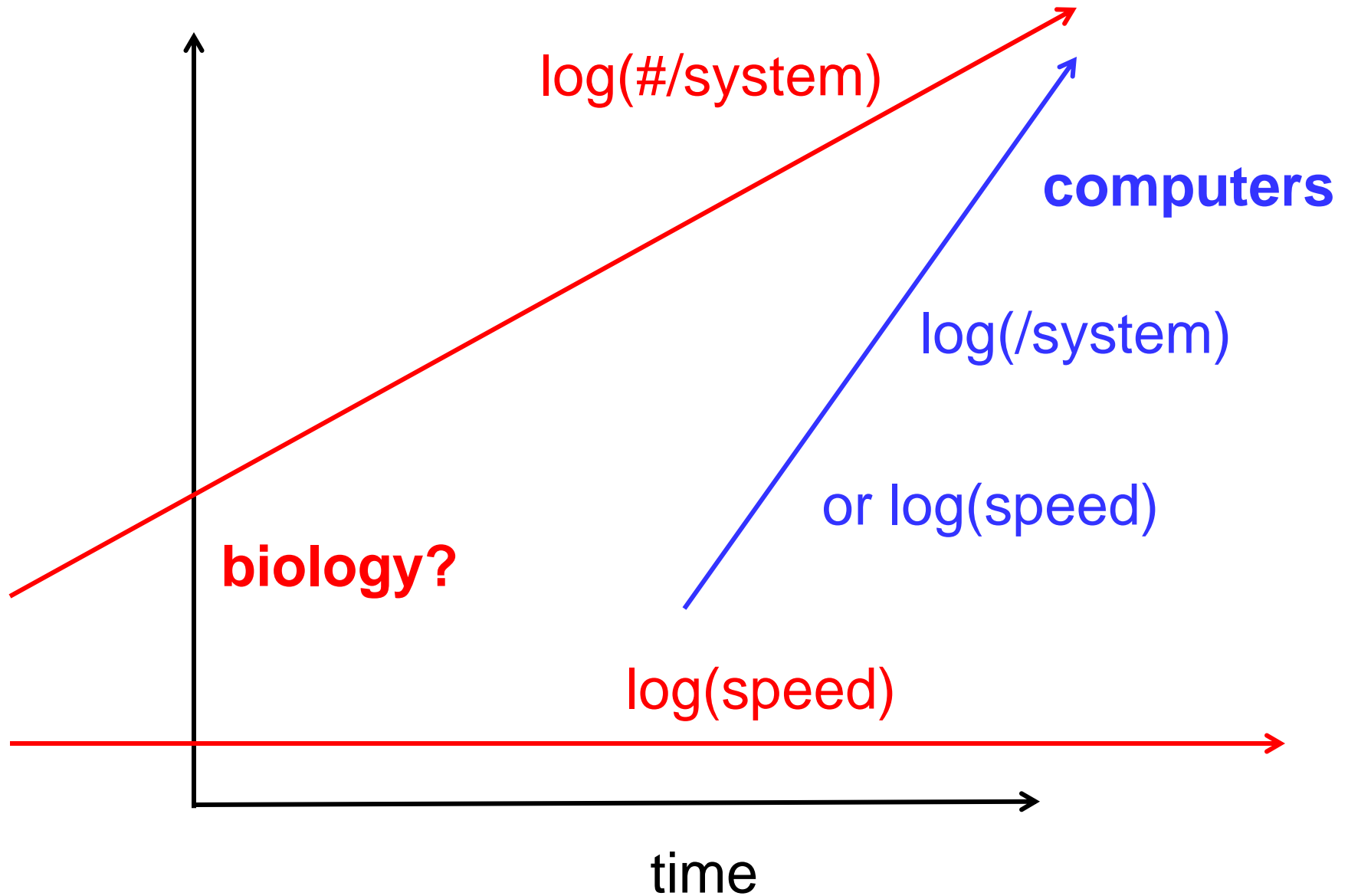
There is a
treatment.



transistors

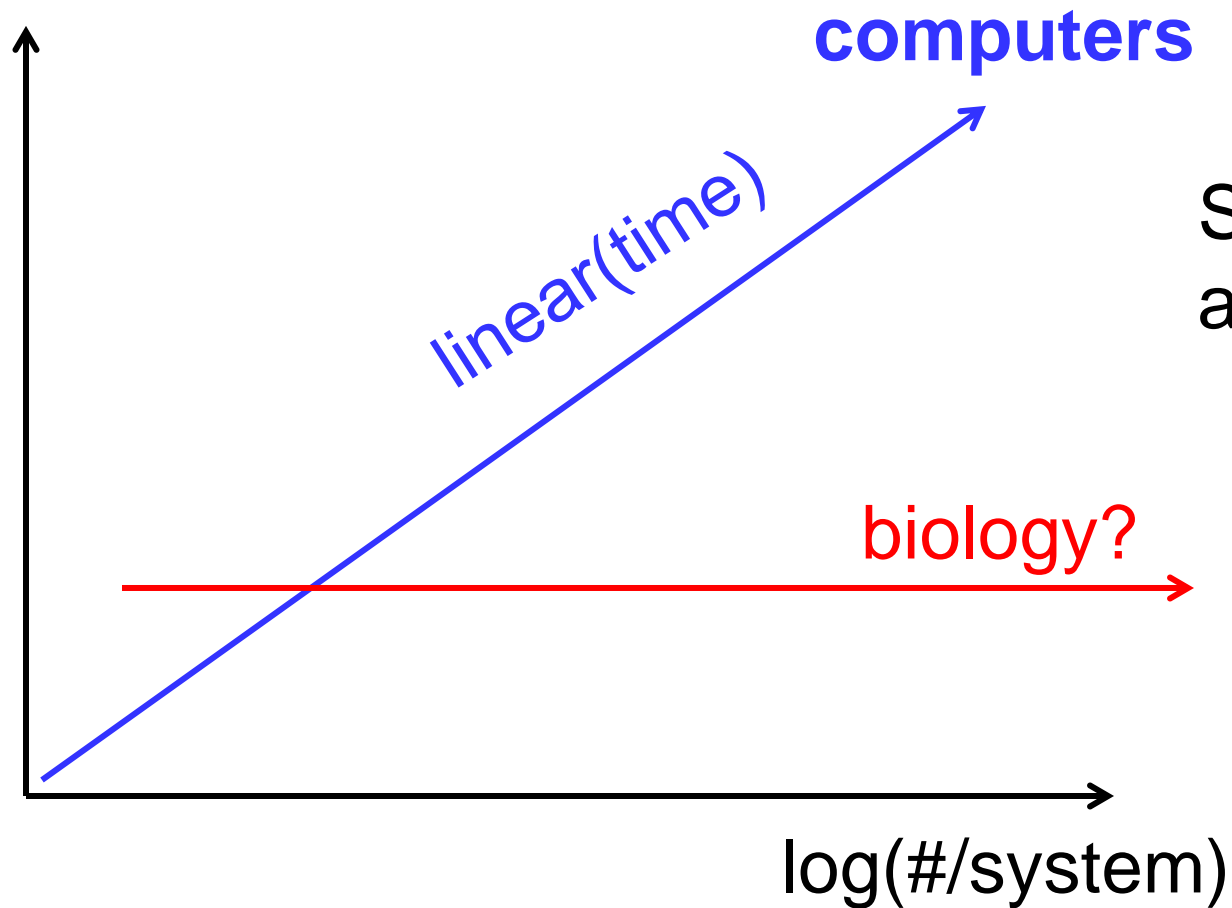


transistors or synapses*1e6



transistors or synapses

$\log(\text{speed})$



So different
architectures?

How general is this picture?

