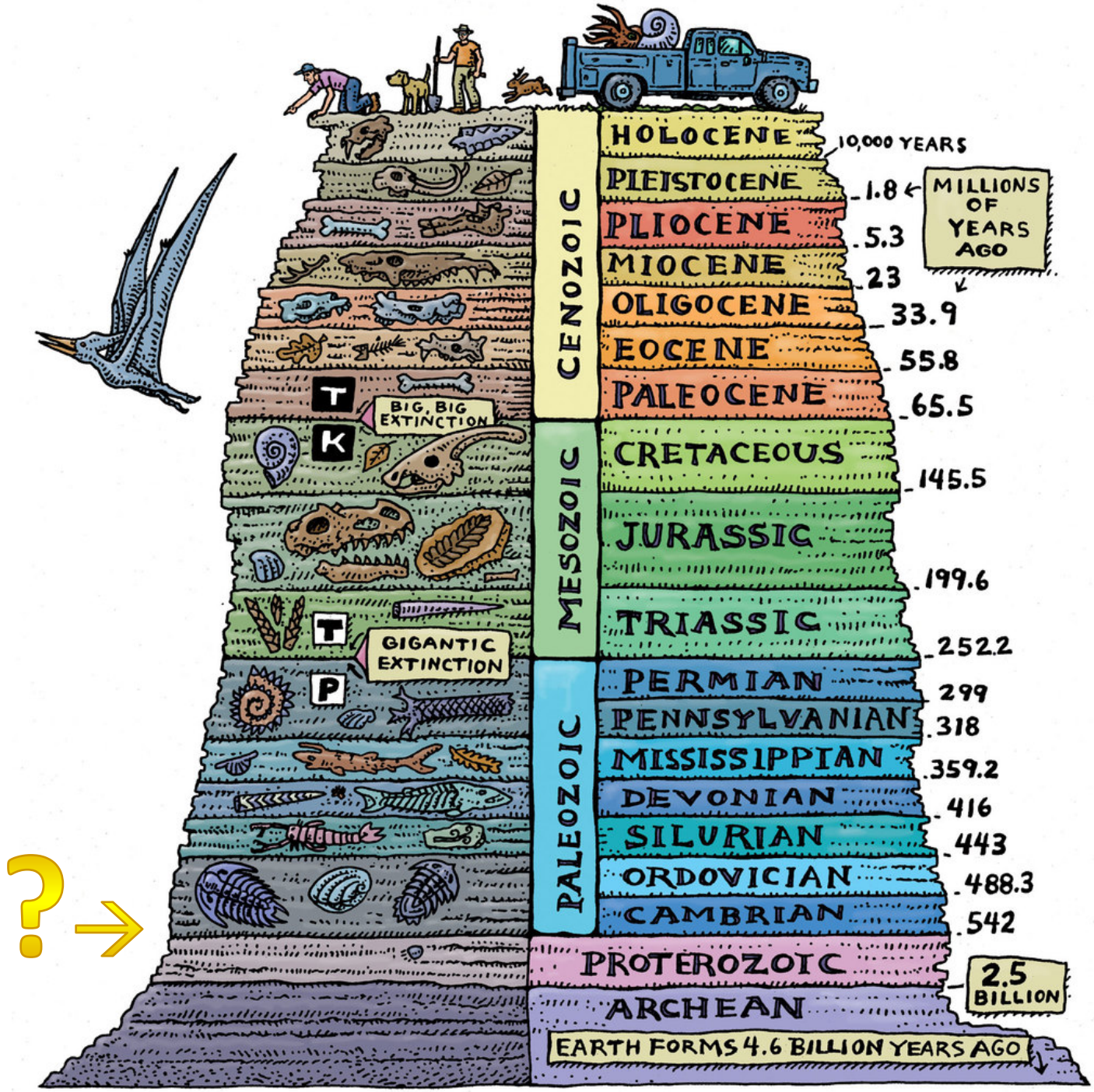
USC Viterbi
School of Engineering
Information Sciences Institute

Flipping the Light Switch

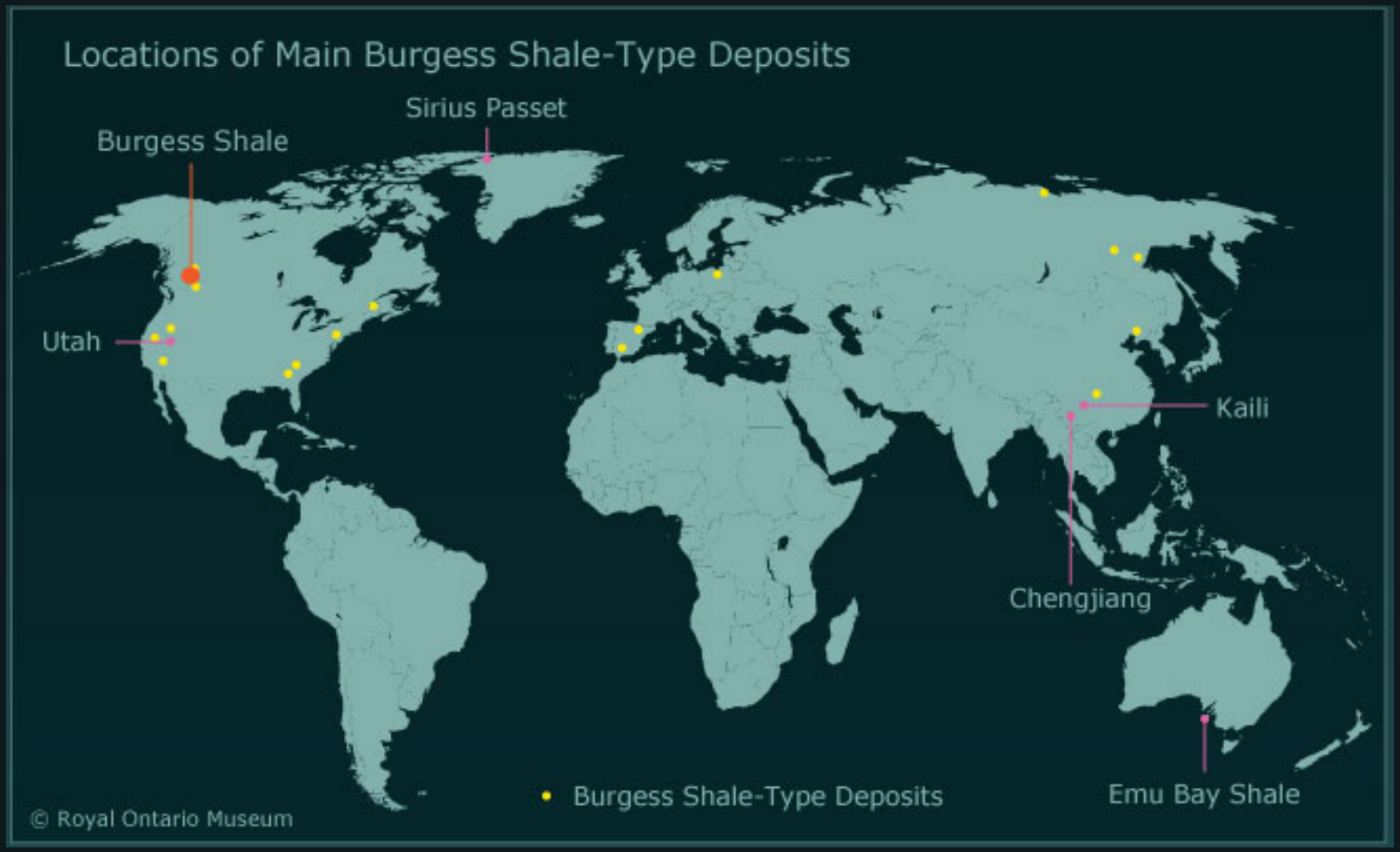
Using the Cambrian Explosion as a Metaphor for Accelerated Scientific Discovery

Gully A. Burns

Intelligent Systems Division, Information Sciences Institute


HOLOCENE
10,000 YEARS
PLEISTOCENE
1.8
MILLIONS OF YEARS AGO
PLIOCENE
5.3
MIOCENE
23
OLIGOCENE
33.9
EOCENE
55.8
PALEOCENE
65.5
CENOZOIC
T
K
BIG, BIG EXTINCTION
CRETACEOUS
145.5
JURASSIC
199.6
TRIASSIC
252.2
MESOZOIC
T
P
GIGANTIC EXTINCTION
PERMIAN
299
PENNSYLVANIAN
318
MISSISSIPPIAN
359.2
DEVONIAN
416
SILURIAN
443
ORDOVICIAN
488.3
CAMBRIAN
542
PALEOZOIC
PROTEROZOIC
2.5 BILLION
ARCHEAN
EARTH FORMS 4.6 BILLION YEARS AGO
?

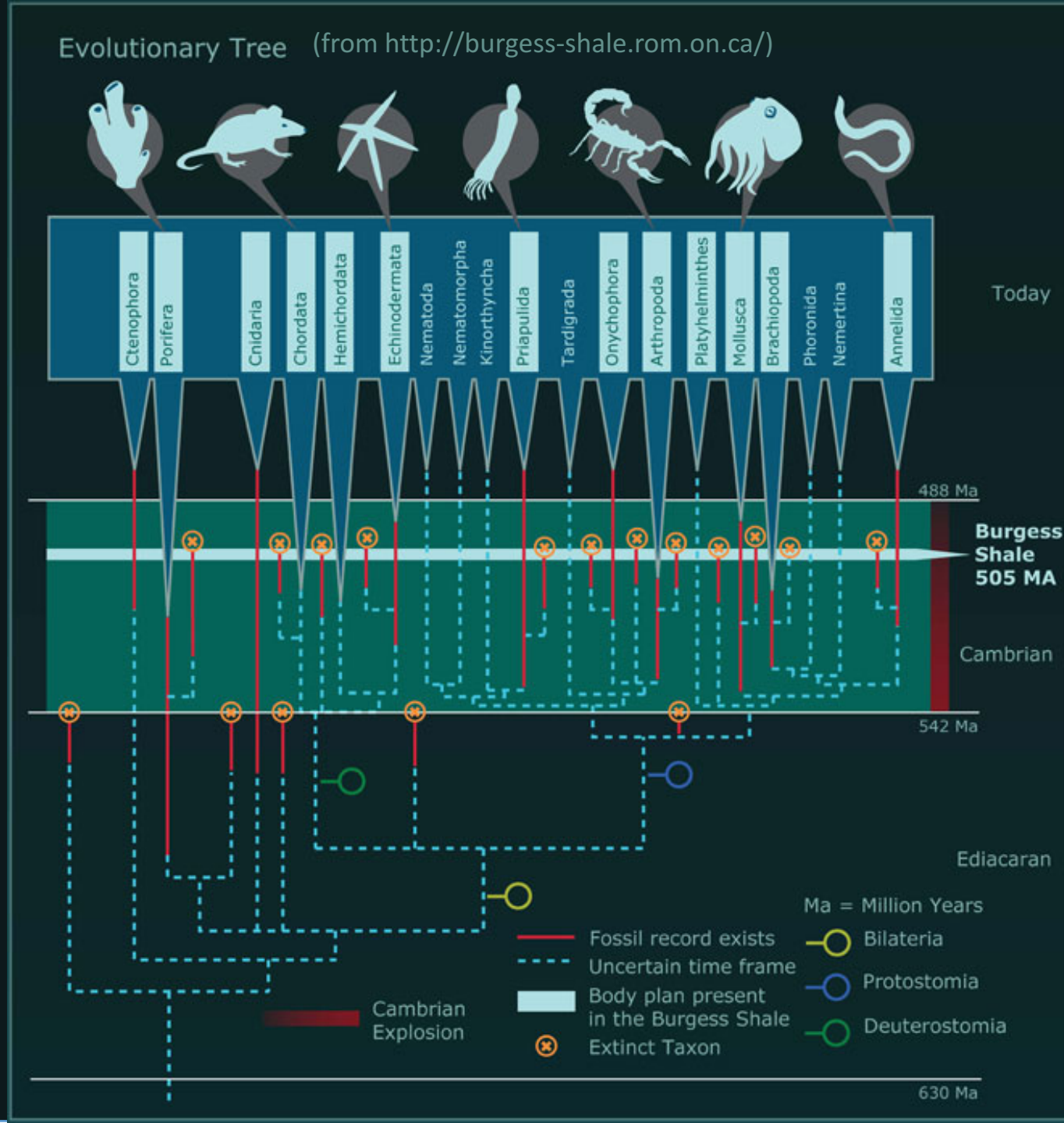
Paleontological Treasure Troves

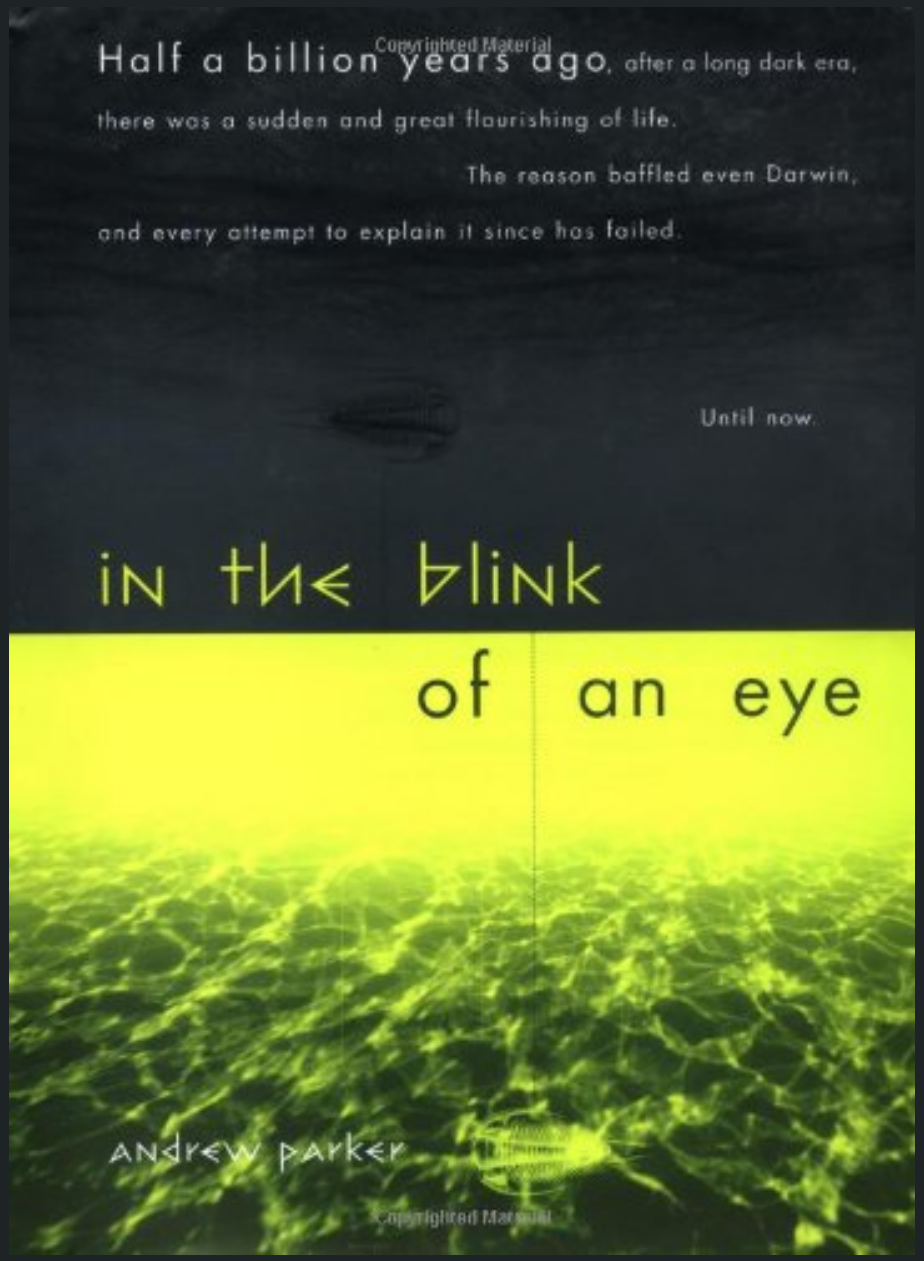


(from http://burgess-shale.rom.on.ca/)

The Cambrian Explosion

Every metazoan phylum came into existence during this period





The 'Light Switch' Hypothesis

Andrew Parker hypothesizes that the cause of the dramatic evolutionary radiation was the evolution of the early eye and visual system

This made predatory behavior much more successful, leading to defensive specializations and transformation from precambrian forms (without skeletons) to armored and more highly evolved Cambrian animals (skeletons, spikes, teeth, swimming capabilities and armor).

(not everyone agrees: see Morris (2003) 'On the First Day, God Said . . .' for a scathing book review)

Using the ‘Light Switch’ Metaphor for Scientific Discovery

Consistent with this hypothesis, we might attempt to look for:

- ‘**Light**’ – a ubiquitous data signal with rich substructure
- **‘Evolution’**– Accelerated scientific discovery
- **‘Eyes’** – Experimental methods + data gathering
- **‘Visual System’** - Data analysis and statistics
- **‘Intelligence’** – Scientific Theory

‘Cancer in the Age of Algorithms’ – a success story

Dr. Shirley Pepke had late-stage ovarian cancer and used advanced unsupervised learning to help her fight her cancer (with a colleague at ISI: Dr. Greg ver Steeg)

“It was suddenly like looking at the dictionary of tumor biology. It was suddenly pulling out all this information other algorithms couldn’t. It looked really beautiful and informative.”

Dr. Pepke switched treatments and her cancer is in remission

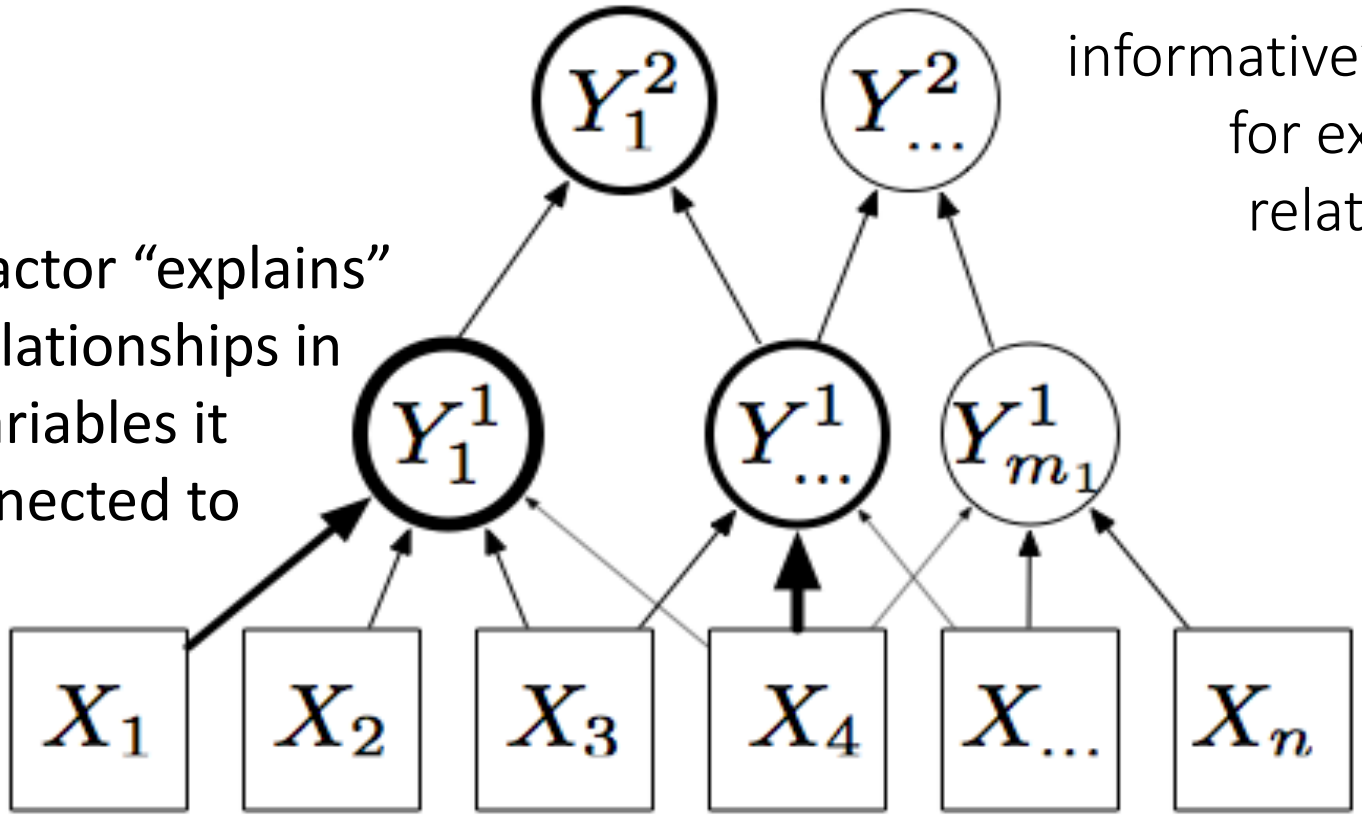
Dr. Shirley Pepke and Dr. Greg Ver Steeg

Washington Post: http://gul.ly/9u24h
Soundcloud Link: http://gul.ly/9u23o

Visual System: Greg ver Steeg's 'Correlation Explanation' (CorEx)

Search for the "most informative" factors for explaining relationships in data

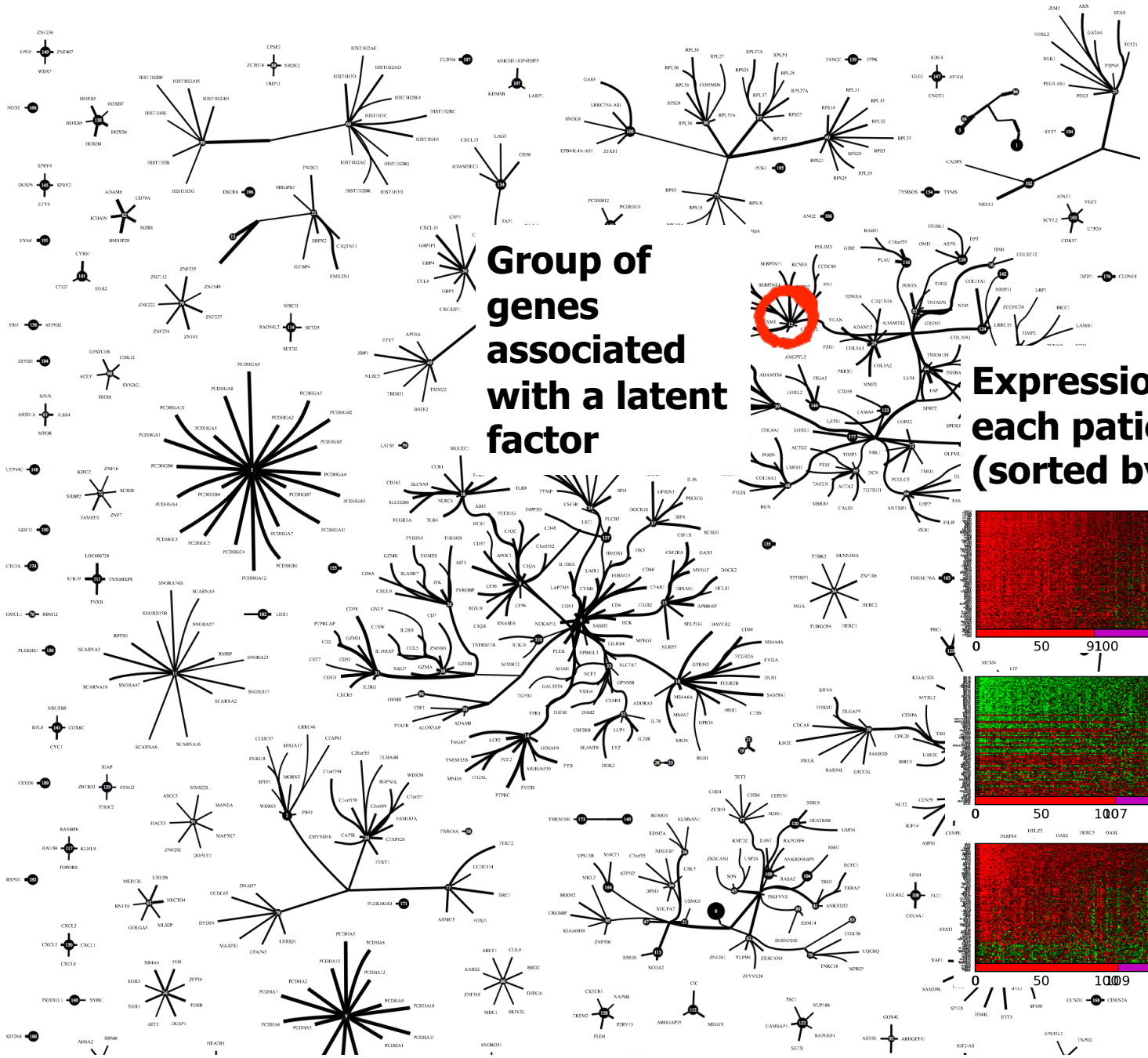
This factor "explains" the relationships in the variables it is connected to



Measurement Data Variables

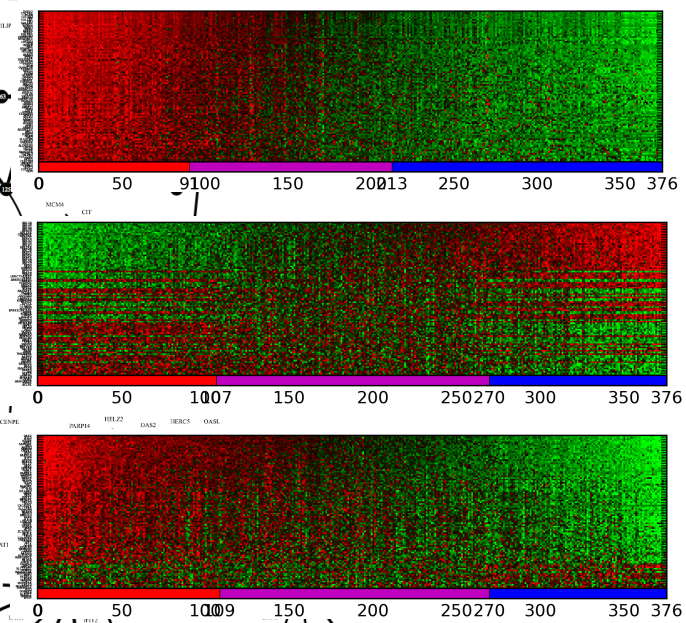
https://arxiv.org/abs/1406.1222

A fraction of the CorEX hierarchical structure

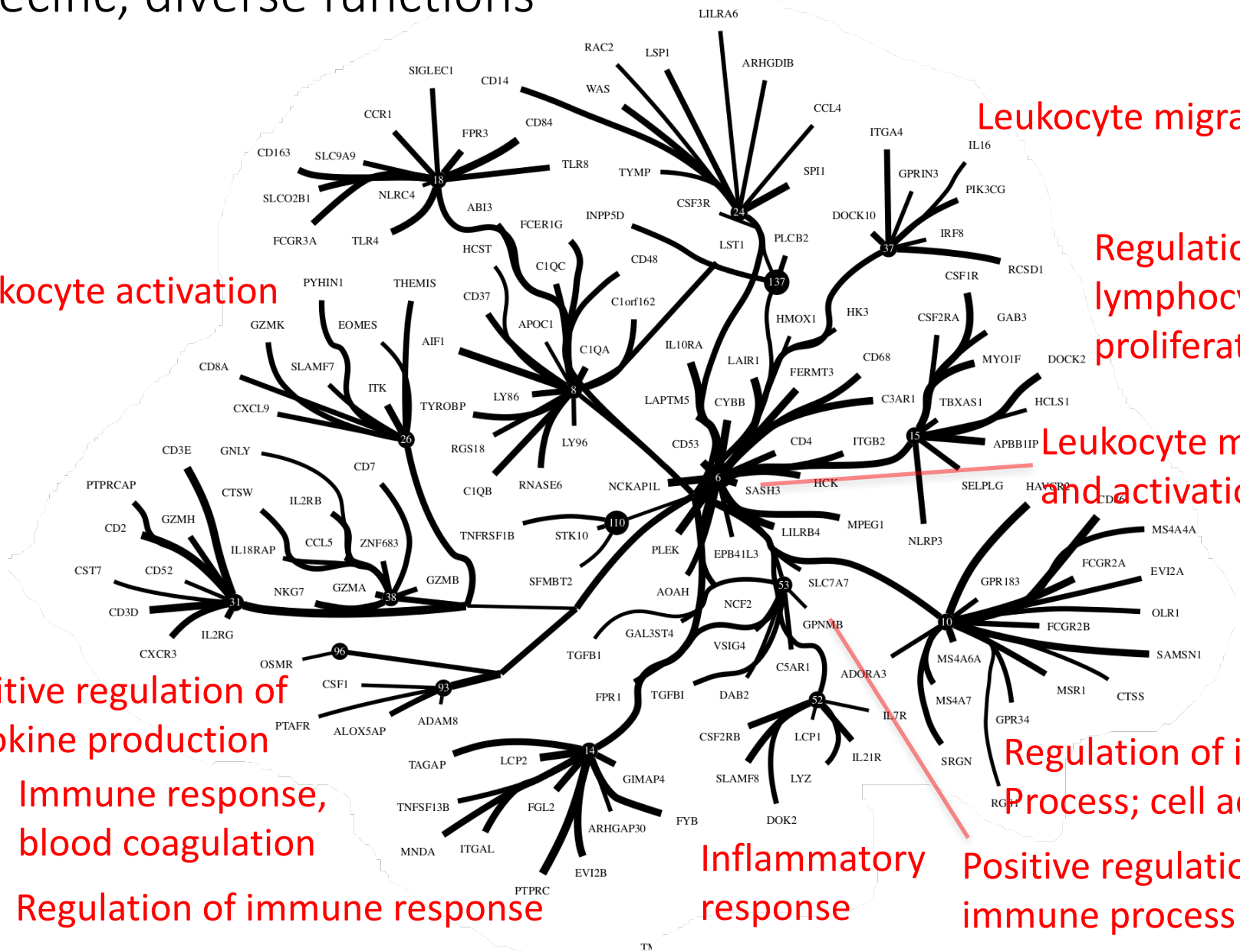


Group of genes associated with a latent factor

Expression levels for each patient (sorted by latent factor)



Gene Ontology Annotations show that groups reflect strong, specific, diverse functions



Leukocyte migration

Regulation of lymphocyte proliferation

Leukocyte activation

Leukocyte migration and activation

Positive regulation of cytokine production

Regulation of immune Process; cell activation

Immune response, blood coagulation

Inflammatory response

Positive regulation of immune process

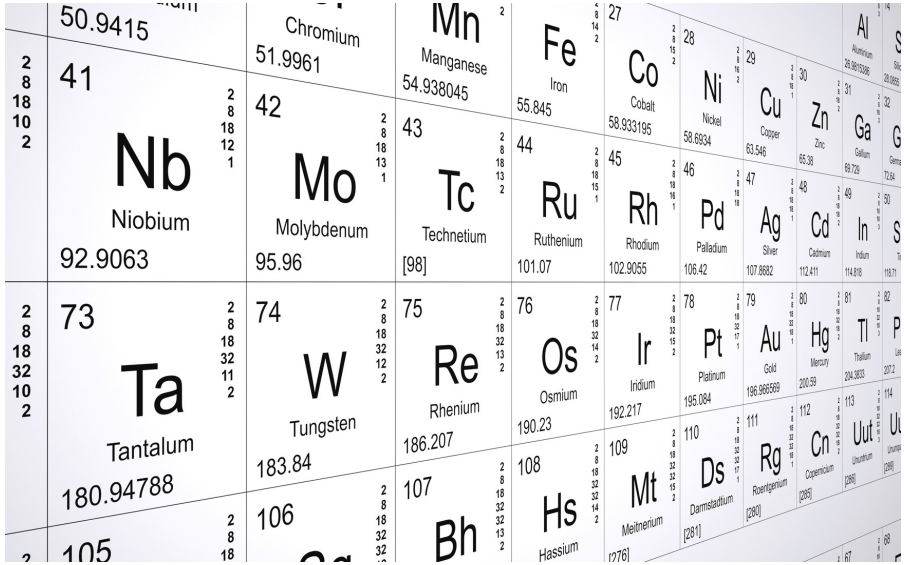
Regulation of immune response

Where's the Light Switch?

- '**Light**' => Gene Expression
- **'Evolution'** => Precision Medicine
- **'Eyes'** => Gene Chip Cancer Assays
- **'Visual System'** => CorEx (Unsupervised Learning)
- **'Intelligence'** => Linkage to Gene Ontology (requires human curated models + interpretation)

How does this idea generalize in the broader context of accelerating science?

Scientific Discovery:
From ‘Elements’ to the Elements



‘stuck’ paradigm

Observe phenomena *(‘fire’)*

Get data to characterize phenomena *(what happens when you burn things?)*

Postulate mechanism *(phlogiston)*

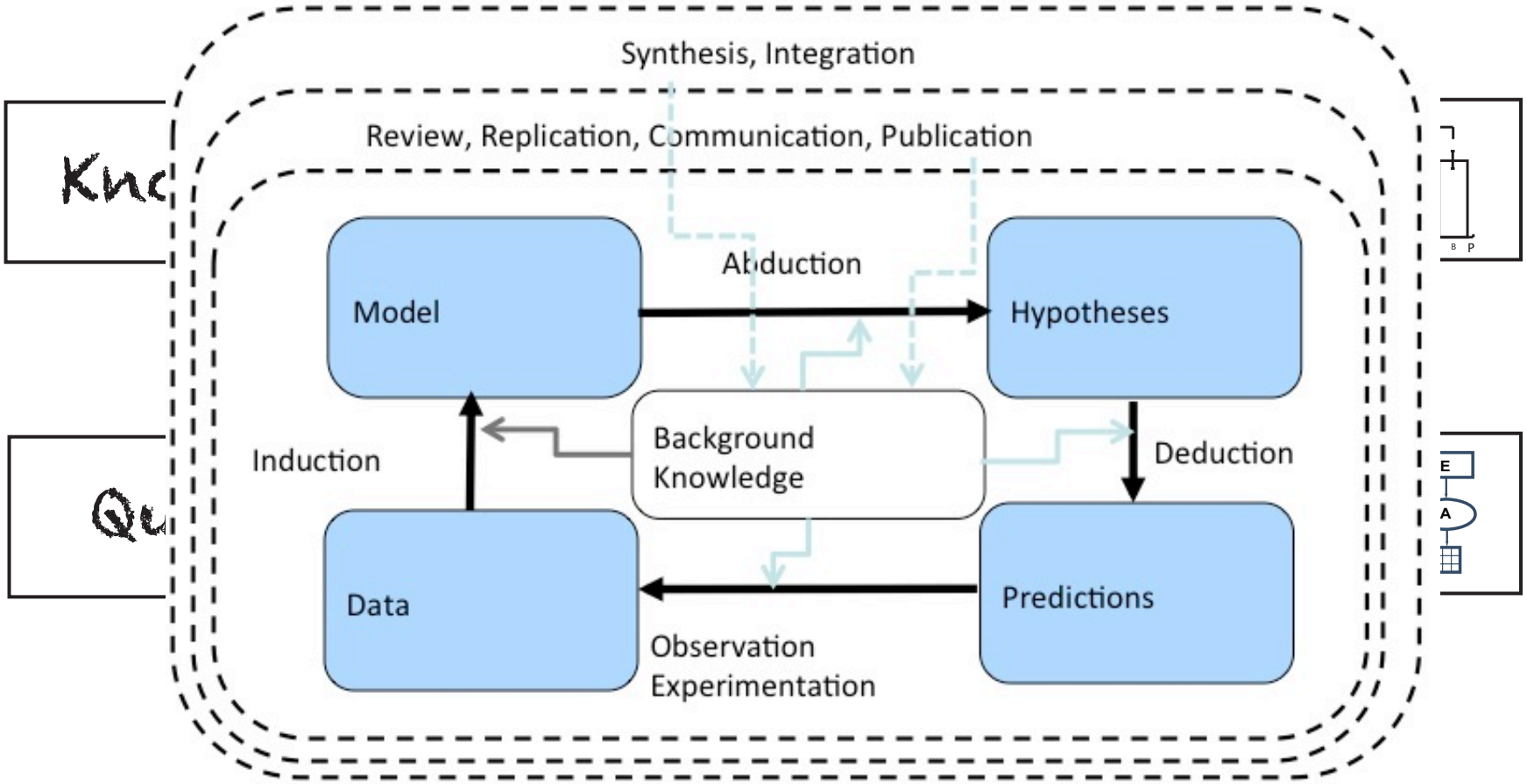
Find correct mechanism *(oxygen + combustion)*

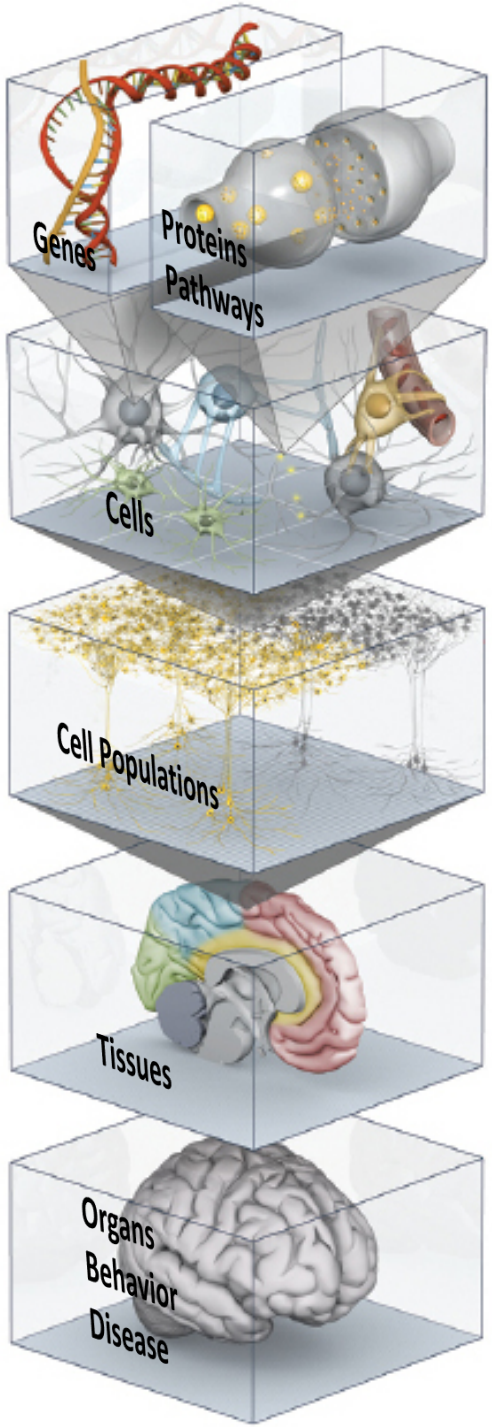
Formulate theory *(chemistry)*

Export reasoning *(periodic table)*

Technology *(e.g., combustion engine)*

Cycles of Investigation ('KQED' Model)





The Complexity of Data Spaces

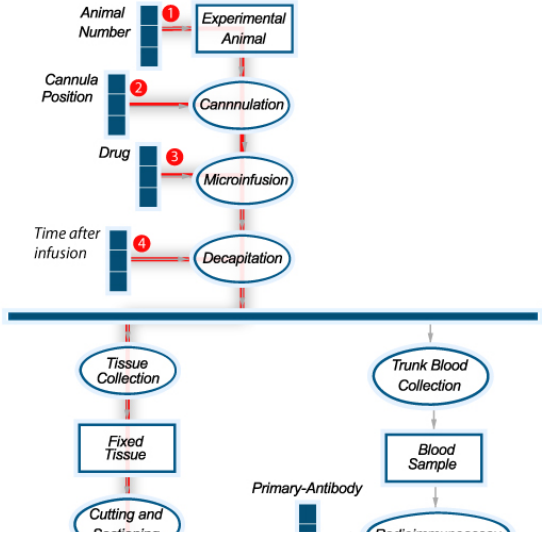
Neuroscience is an example of a complex multi-level domain

A variety of variables are needed to describe models across scales and systems.

Complex experiments have to be designed and executed by human experts to investigate and test these models.

How to apply the Cambrian Metaphor here?

Knowledge Engineering from Experimental Design

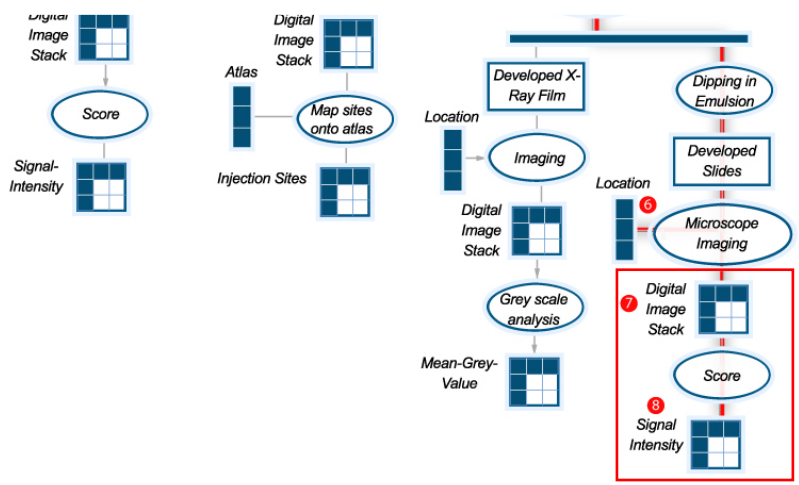


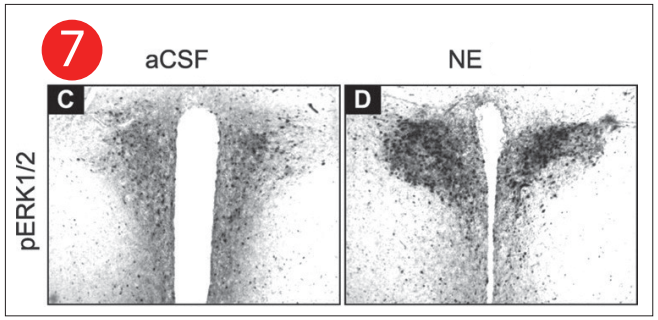
Gene Expression Lab Study
Khan et al. (2007)

Indexing Parameters ('Independent Variables')

Measurements ('Dependent Variables')

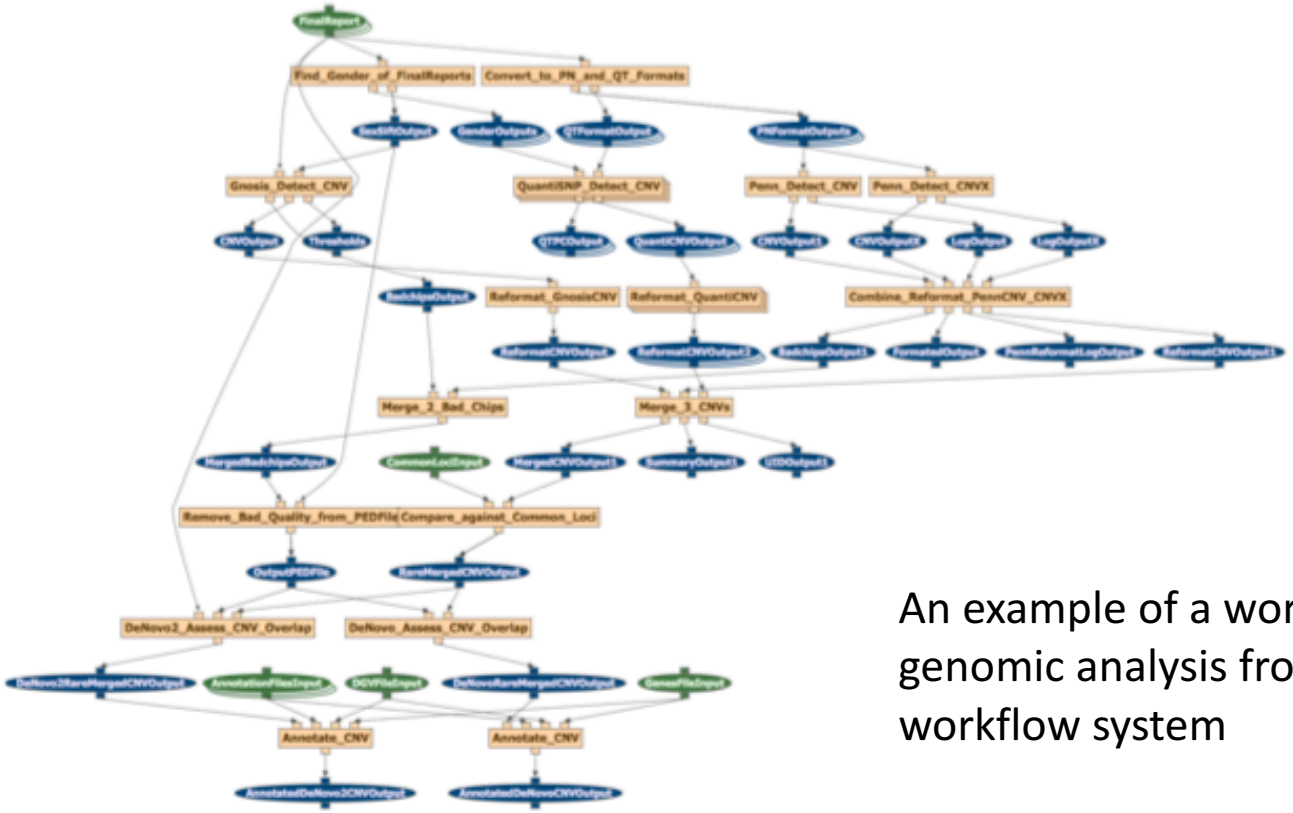
1 subject-id	2 Cannula Position	3 Drug	4 Time after infusion	5 probe	6 Location	8 Signal Intensity
RO2-164	PVH proximity	NE2	10 mins	pERK1/2	PVHmpd	high
RO2-175	PVH proximity	aCSF	10 mins	pERK1/2	PVHmpd	low





Russ et al (2011), BMC Bioinformatics

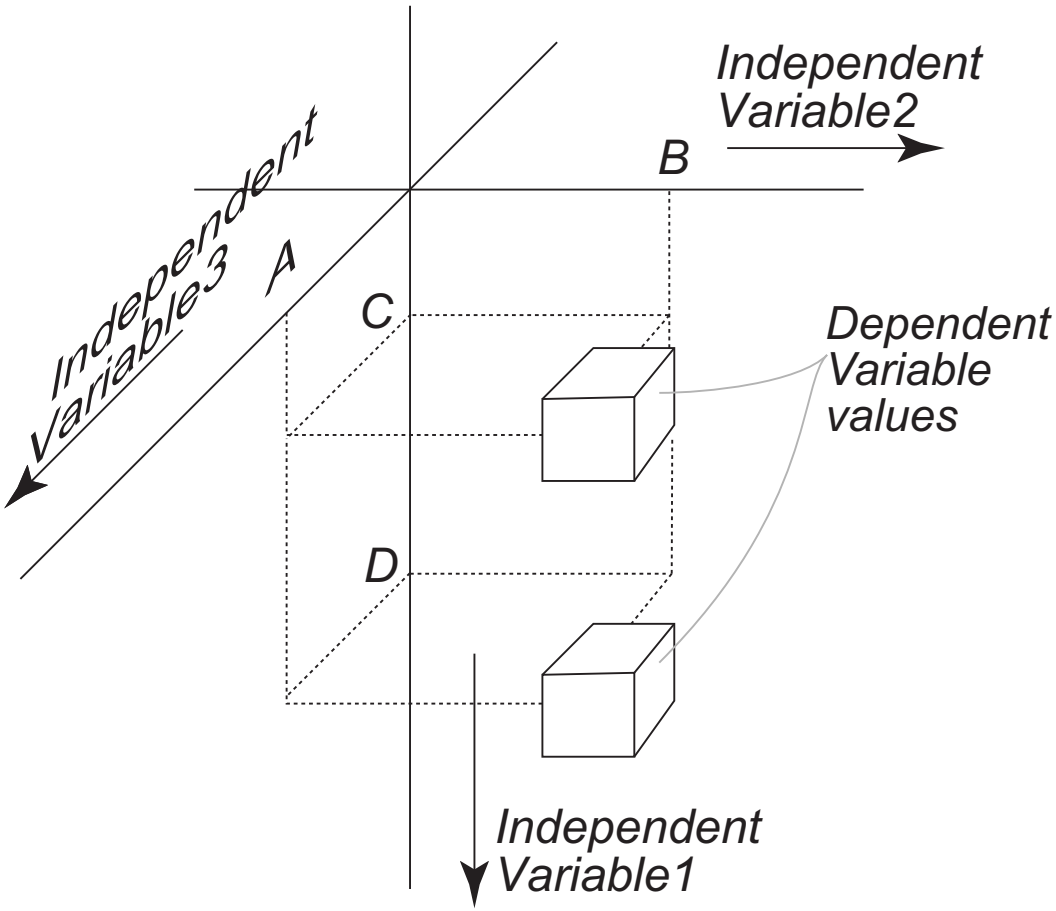
E-Science Workflows



An example of a workflow for genomic analysis from the WINGS workflow system

http://www.wings-workflows.org/

Measurements and Metadata Define the Data Space



Could we apply tools like CorEx to examine the structure of this space?

Would this be like suddenly flipping the switch and turning on the lights?

Acknowledgements

Greg ver Steeg
Oren Etzioni
Alan Watts
Prem Natarajan
Yolanda Gil
Ed Hovy

NIH, DARPA,
NSF, IARPA