STANDING ON THE SHOULDERS OF GIANTS

HOW MASSIVE KNOWLEDGE-BASES ARE TRANSFORMING DATA ANALYTICS IN BIOLOGY

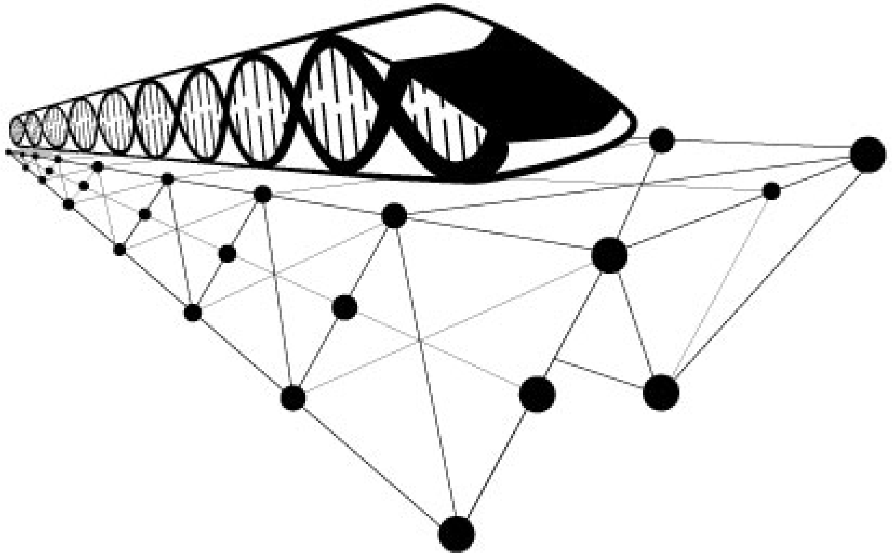
SAURABH SINHA

PROFESSOR OF COMPUTER SCIENCE

AND THE CARL R. WOESE INSTITUTE FOR GENOMIC BIOLOGY

CO-DIRECTOR & RESEACH PI, NIH BD2K CENTER OF EXCELLENCE, UIUC & MAYO CLINIC

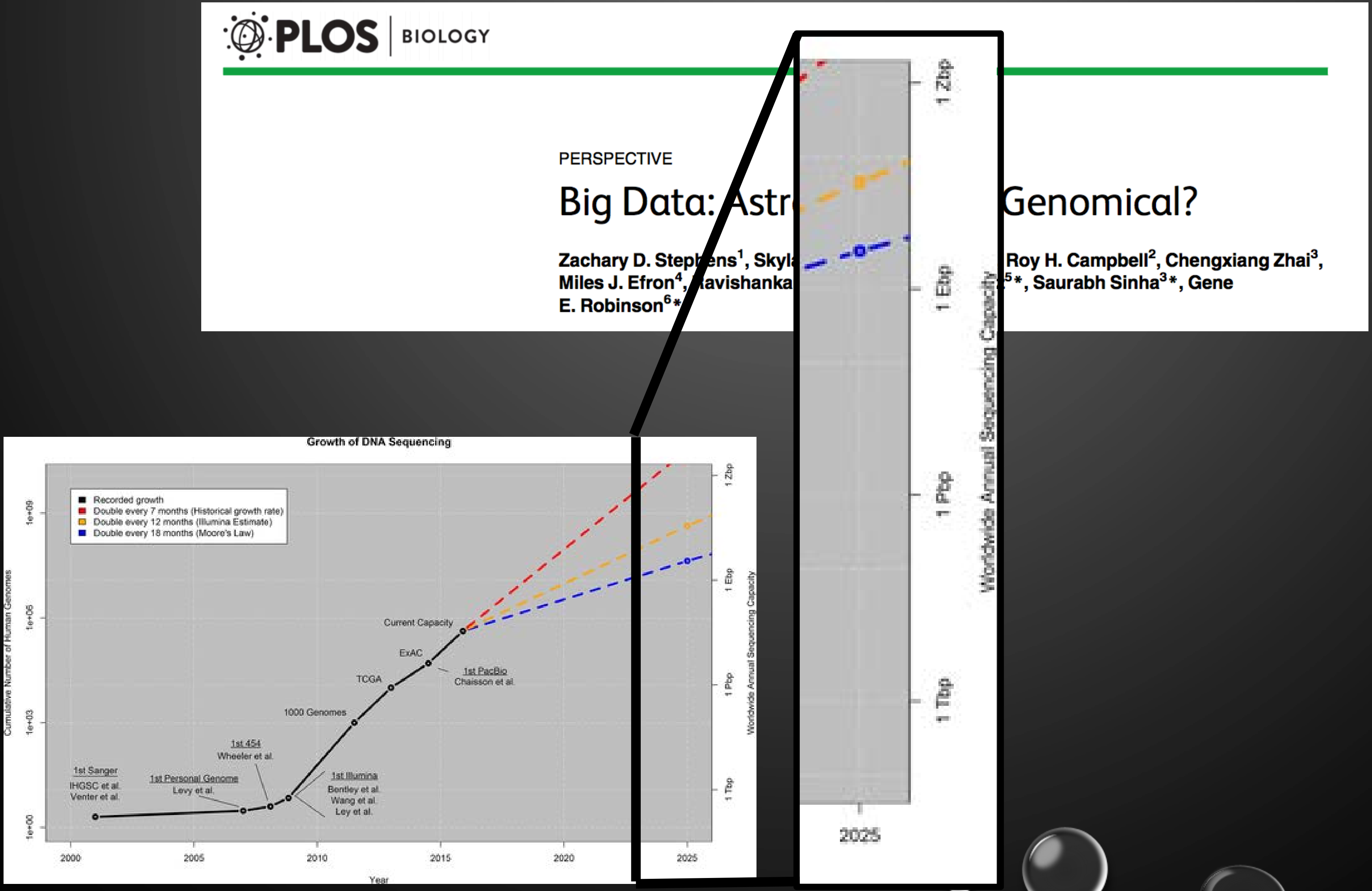
knoweng BIG DATA TO KNOWLEDGE
CENTER OF EXCELLENCE



NIH
National Institutes of Health

MAYO CLINIC

IT’S, UH, “GENOMICAL”



PLOS | BIOLOGY

PERSPECTIVE

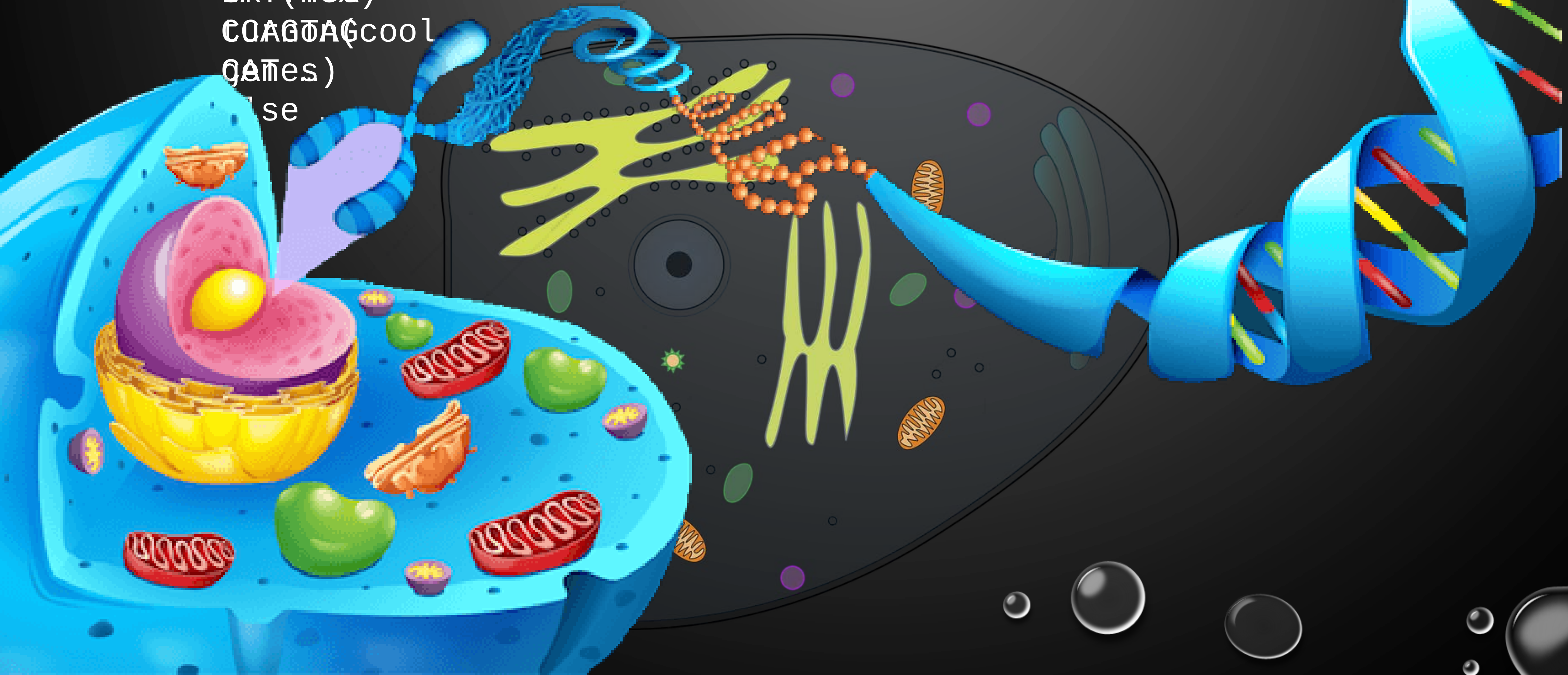
Big Data: Astr Genomical?

Zachary D. Stephens[1], Skyla Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], avishanka [5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*

IN THE BEGINNING THERE WAS THE GENOME

YOUR GENOME = YOUR CELL’S SOFTWARE

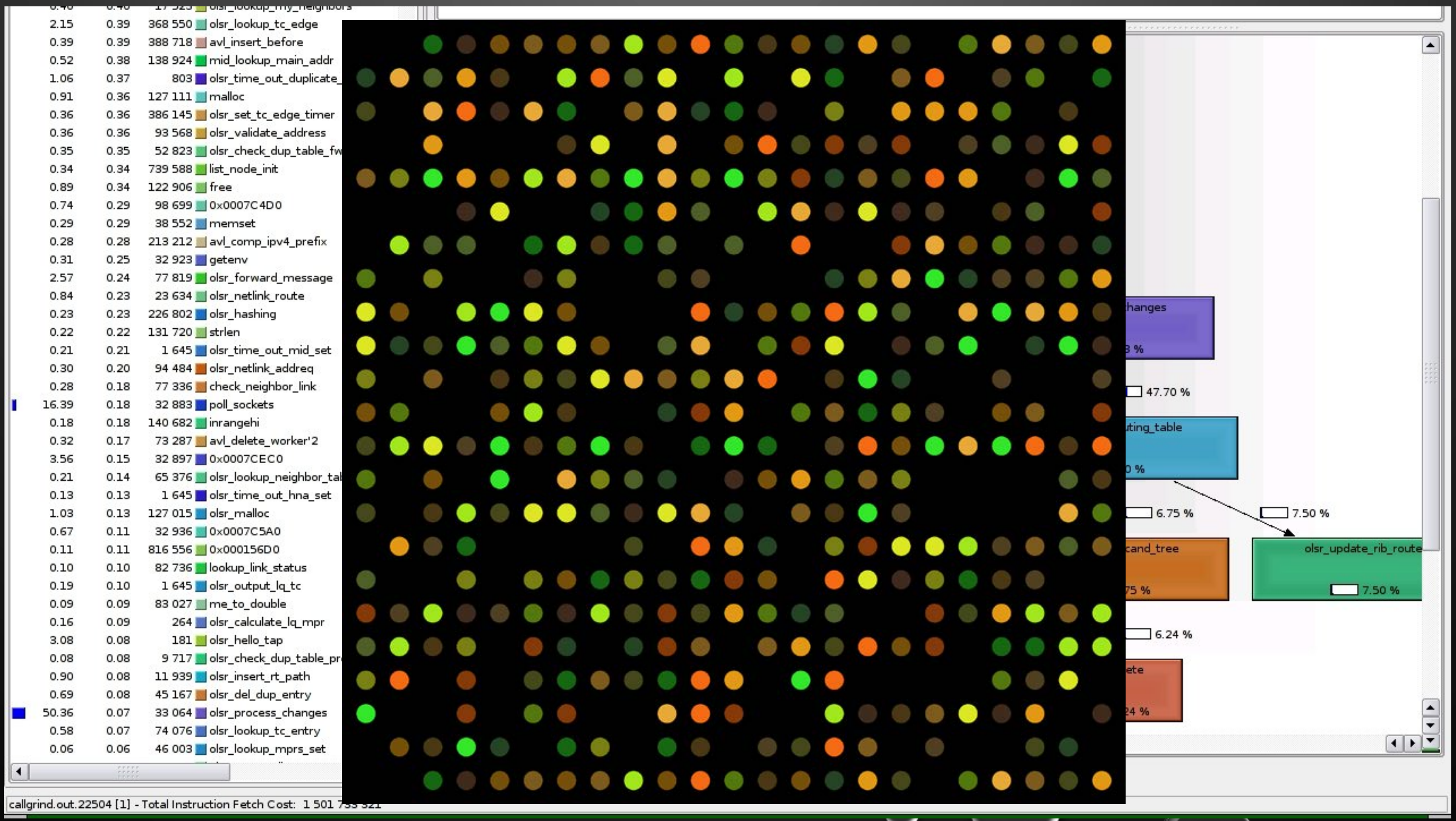
```
main()
ACGGACG
GATT(TGA)
CCAGTAGcool
games)
 se .
```



“MACHINE” CODE IS NOT VERY USEFUL

PHASE TWO : INTERPRETATION
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME
SHENEMAN The Star-Ledger

SO PEOPLE STARTED PROFILING THE CODE



```
  0.40    0.40     17 323 olsr_lookup_my_neighbors
  2.15    0.39    368 550 olsr_lookup_tc_edge
  0.39    0.39    388 718 avl_insert_before
  0.52    0.38    138 924 mid_lookup_main_addr
  1.06    0.37        803 olsr_time_out_duplicate_
  0.91    0.36    127 111 malloc
  0.36    0.36    386 145 olsr_set_tc_edge_timer
  0.36    0.36     93 568 olsr_validate_address
  0.35    0.35     52 823 olsr_check_dup_table_fw
  0.34    0.34    739 588 list_node_init
  0.89    0.34    122 906 free
  0.74    0.29     98 699 0x0007C4D0
  0.29    0.29     38 552 memset
  0.28    0.28    213 212 avl_comp_ipv4_prefix
  0.31    0.25     32 923 getenv
  2.57    0.24     77 819 olsr_forward_message
  0.84    0.23     23 634 olsr_netlink_route
  0.23    0.23    226 802 olsr_hashing
  0.22    0.22    131 720 strlen
  0.21    0.21      1 645 olsr_time_out_mid_set
  0.30    0.20     94 484 olsr_netlink_addreq
  0.28    0.18     77 336 check_neighbor_link
 16.39    0.18     32 883 poll_sockets
  0.18    0.18    140 682 inrangehi
  0.32    0.17     73 287 avl_delete_worker'2
  3.56    0.15     32 897 0x0007CEC0
  0.21    0.14     65 376 olsr_lookup_neighbor_tal
  0.13    0.13      1 645 olsr_time_out_hna_set
  1.03    0.13    127 015 olsr_malloc
  0.67    0.11     32 936 0x0007C5A0
  0.11    0.11    816 556 0x000156D0
  0.10    0.10     82 736 lookup_link_status
  0.19    0.10      1 645 olsr_output_lq_tc
  0.09    0.09     83 027 me_to_double
  0.16    0.09        264 olsr_calculate_lq_mpr
  3.08    0.08        181 olsr_hello_tap
  0.08    0.08      9 717 olsr_check_dup_table_pr
  0.90    0.08     11 939 olsr_insert_rt_path
  0.69    0.08     45 167 olsr_del_dup_entry
 50.36    0.07     33 064 olsr_process_changes
  0.58    0.07     74 076 olsr_lookup_tc_entry
  0.06    0.06     46 003 olsr_lookup_mprs_set
```

callgrind.out.22504 [1] - Total Instruction Fetch Cost: 1 501 [illegible]

ALL ROADS LEAD TO A SPREADSHEET

Conditions

Genes

	Tissue 1	Tissue 2	…	Tissue 400
Gene 1	20	5	23	37
Gene 2	10	17	201	29
…	100	102	99	84
Gene 20000	20	45	74	62

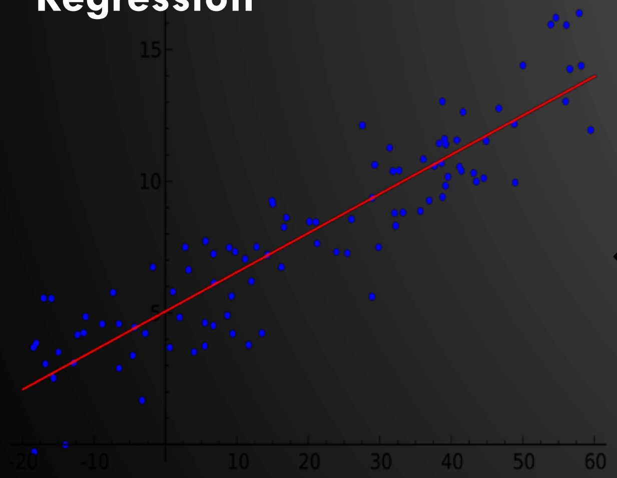
Conditions

Genes

	Patient 1	Patient 2	…	Patient 400
Gene 1	20	5	23	37
Gene 2	10	17	201	29
…	100	102	99	84
Gene 20000	20	45	74	62

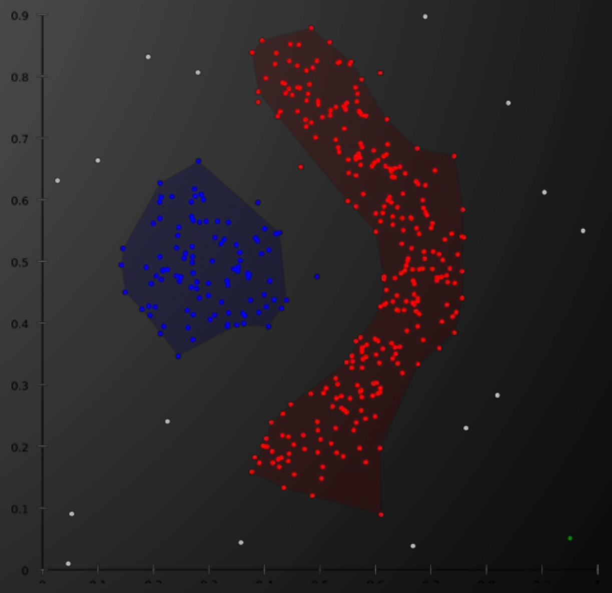
SPREADSHEET ANALYTICS (A.K.A. BIOINFORMATICS)

Regression

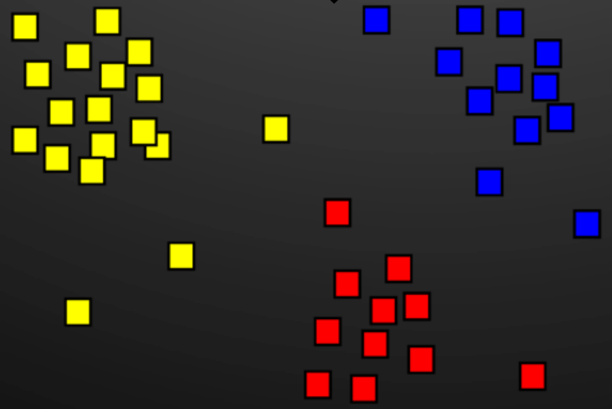


	Patient 1	Patient 2	…	Patient 400
Gene 1	20	5	23	37
Gene 2	10	17	201	29
…	100	102	99	84
Gene 20000	20	45	74	62

Classification



Clustering



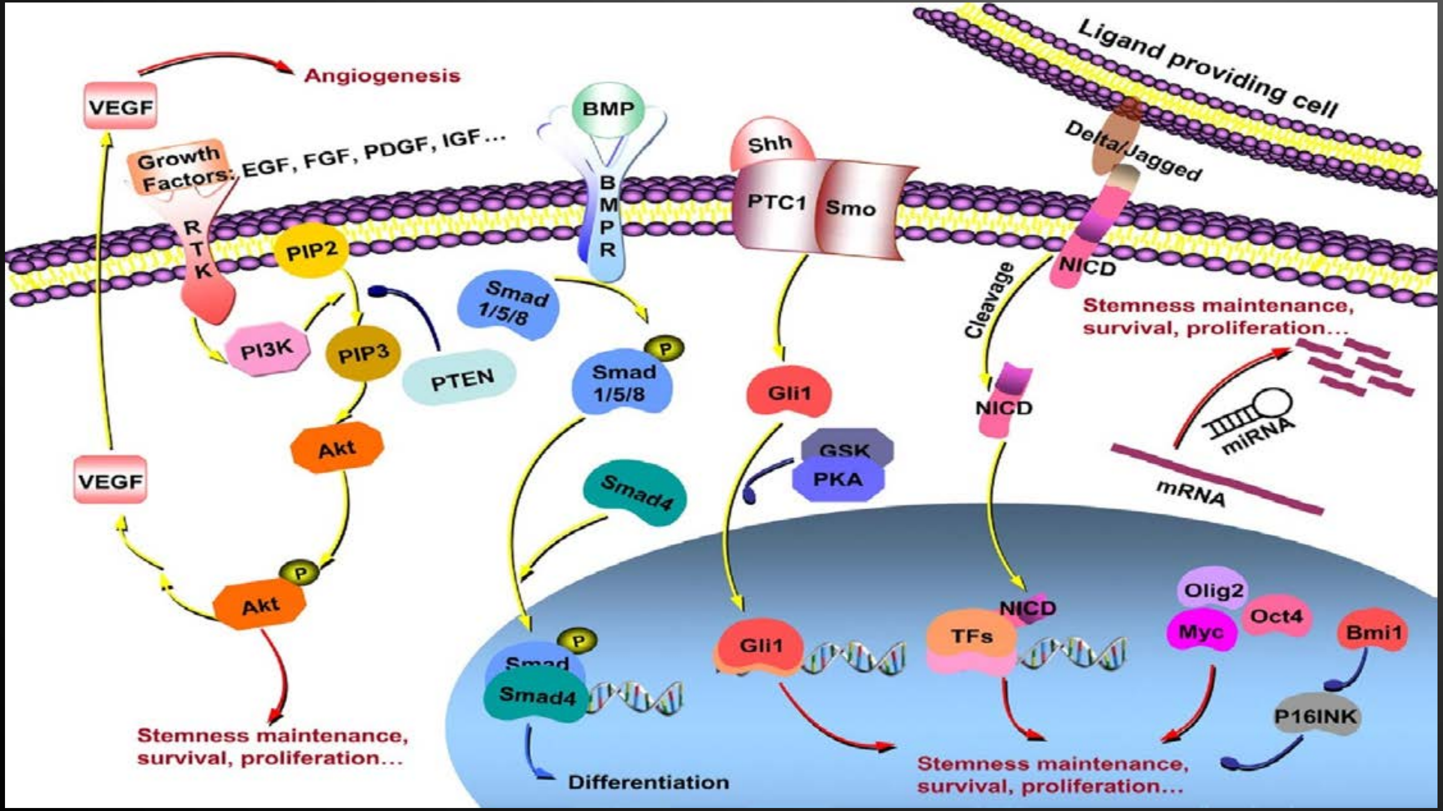
“A GOOD DECISION IS BASED ON KNOWLEDGE AND NOT ON NUMBERS”
- PLATO

THE GWAS STORY

		Disease?
I1:	AACGAGCTAGCGATCGATCGACTACGACTACGAGGT	-
I2:	AACGAGCTAGCGATCGATCGACTACGACTACGAGGT	-
I3:	AACGAGCTAGCGATCGATCGACTACGACTACGAGGT	-
I4:	AACGAGCTAGCGATCGATCGACTACGACTACGAGGT	-
I5:	AACGAGCTAGCGATCGATCGACAACGACTACGAGGT	+
I6:	AACGAGCTAGCGATCGATCGACTACGACTACGAGGT	-
I7:	AACGAGCTAGCGATCGATCGACTACGACTACGAGGT	-
I8:	AACGAGCTAGCGATCGATCGACAACGACTACGAGGT	+

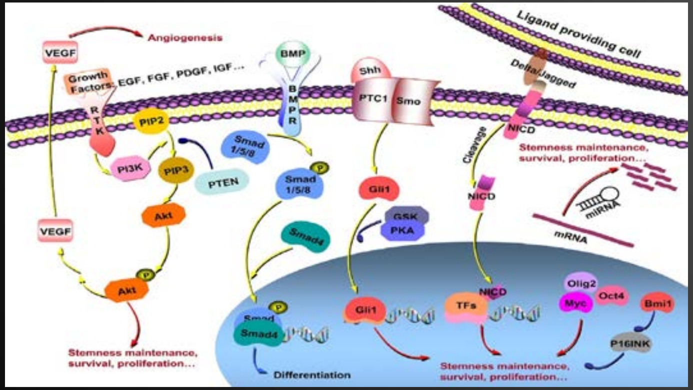
Doesn’t work as well as we’d like it to

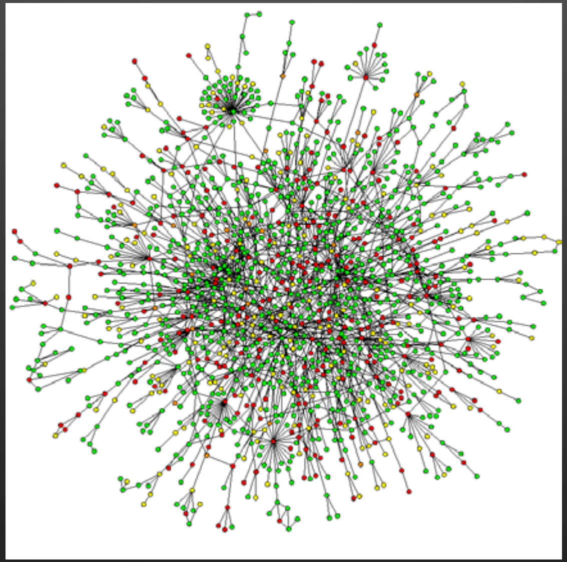
SEEK MODULES, NOT INDIVIDUAL GENES

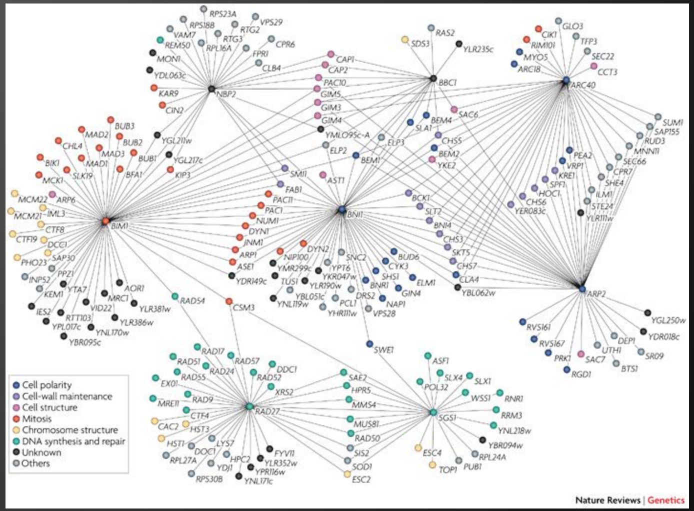


Many ways to ‘break’ the code

BIOLOGICAL NETWORKS GALORE

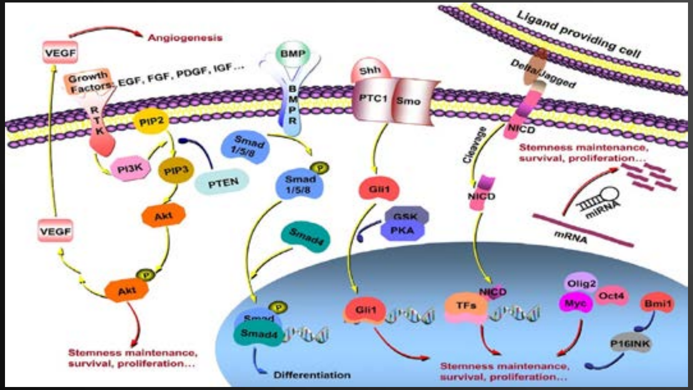


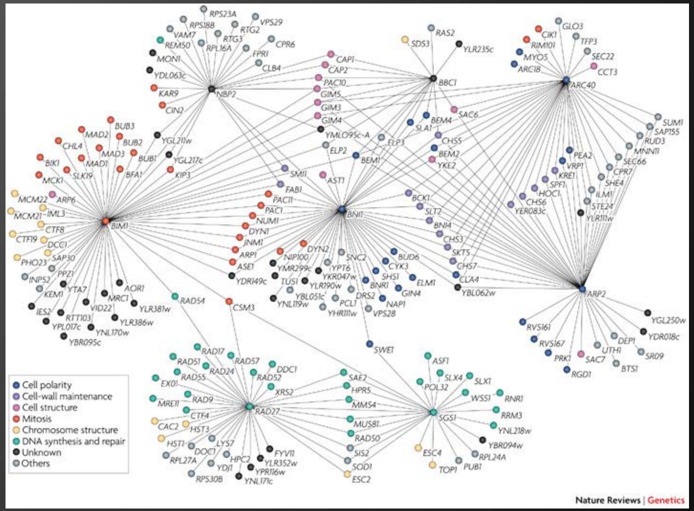




Nature Reviews | Genetics

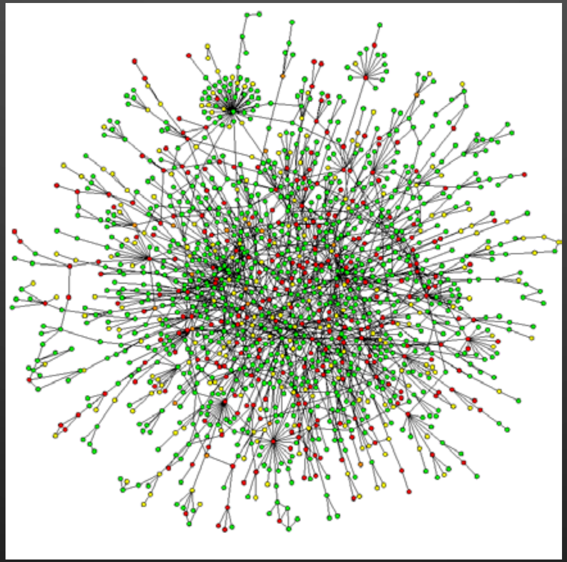
BIOLOGICAL NETWORKS GALORE





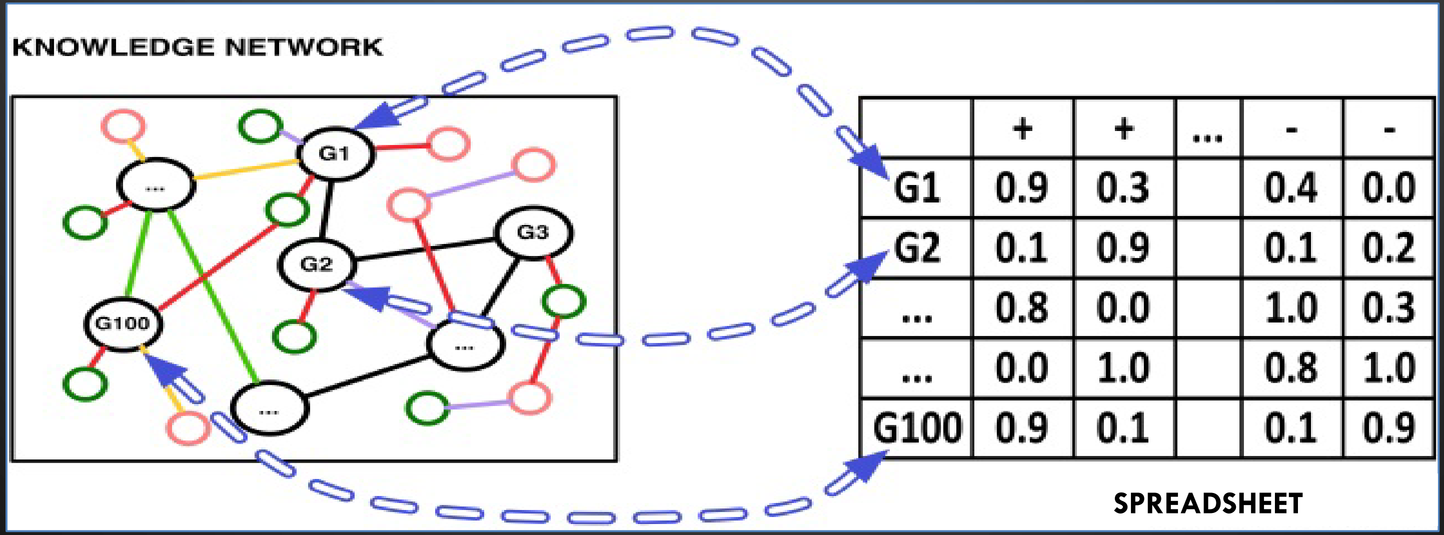
Nature Reviews | Genetics

'Knowledge Network':
3M nodes
80M edges
82 edge types



GIANT NETWORK GUIDES BIOLOGICAL ANALYSIS

Graph Mining

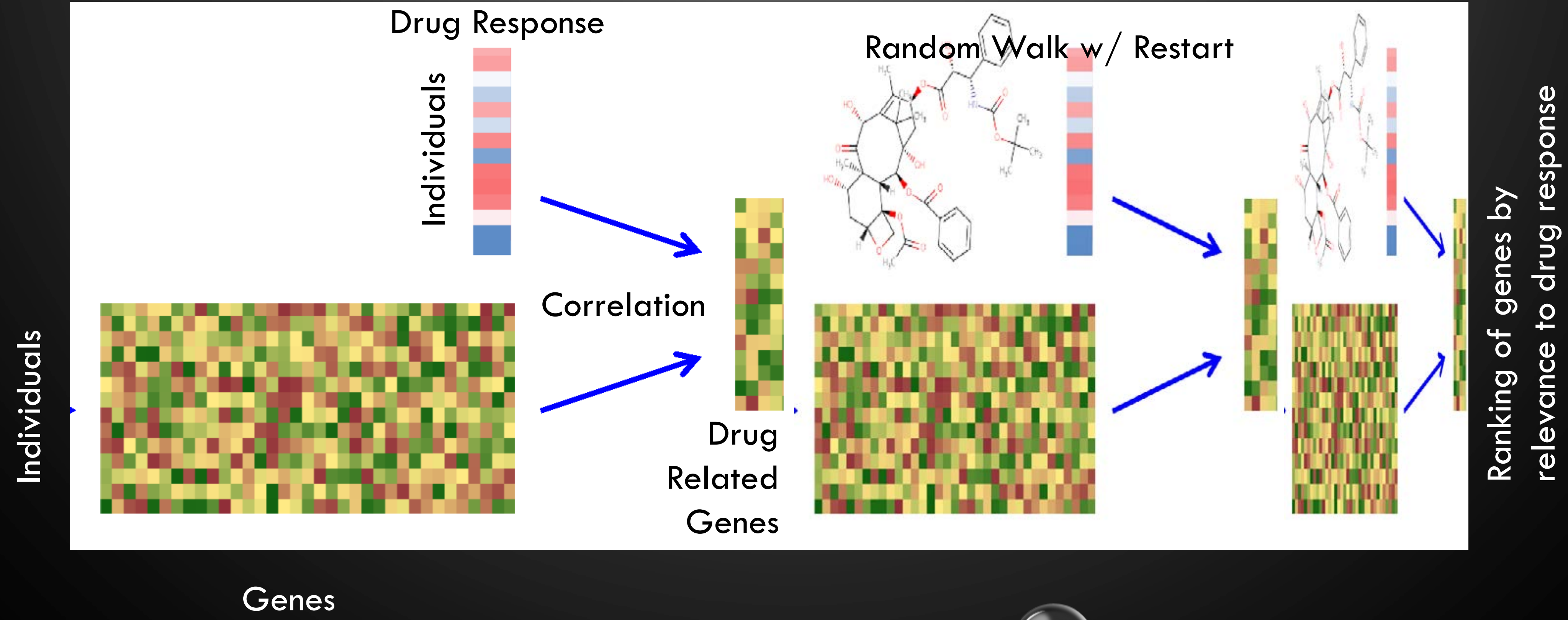


	+	+	...	-	-
G1	0.9	0.3		0.4	0.0
G2	0.1	0.9		0.1	0.2
...	0.8	0.0		1.0	0.3
...	0.0	1.0		0.8	1.0
G100	0.9	0.1		0.1	0.9

Machine learning

Knowledge network + user spreadsheet

EXAMPLE: FINDING GENES INFLUENCING DRUG RESPONSE



Drug Response

Individuals

Random Walk w/ Restart

Correlation

Individuals

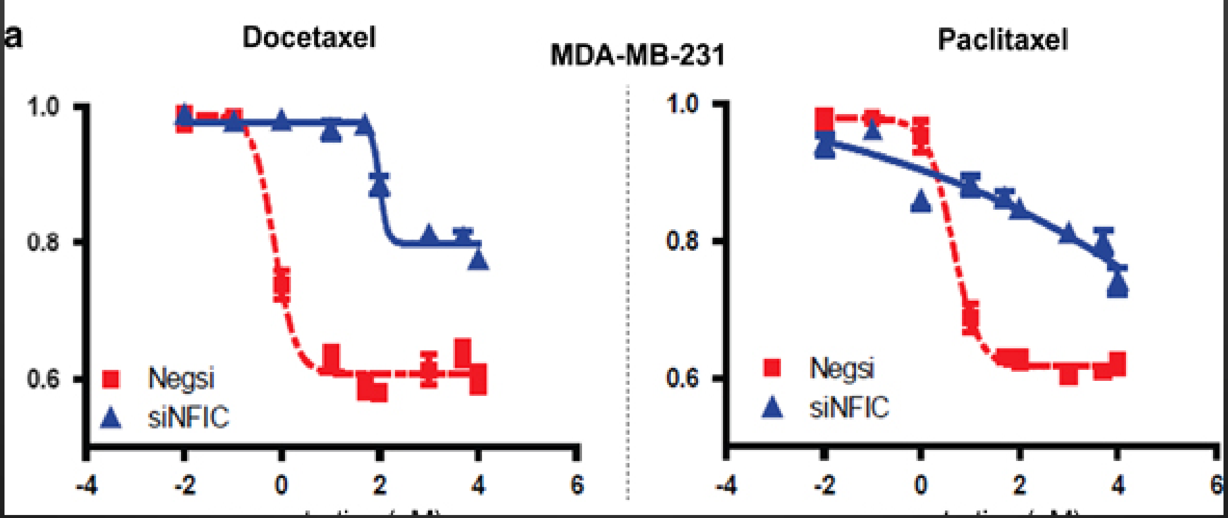
Drug Related Genes

Ranking of genes by relevance to drug response

Genes

EXAMPLE: FINDING GENES INFLUENCING DRUG RESPONSE

Validated 17 genes for several cancer drugs



CYBERINFRASTRUCTURE

HOW BIOLOGISTS DO BIOINFORMATICS TODAY

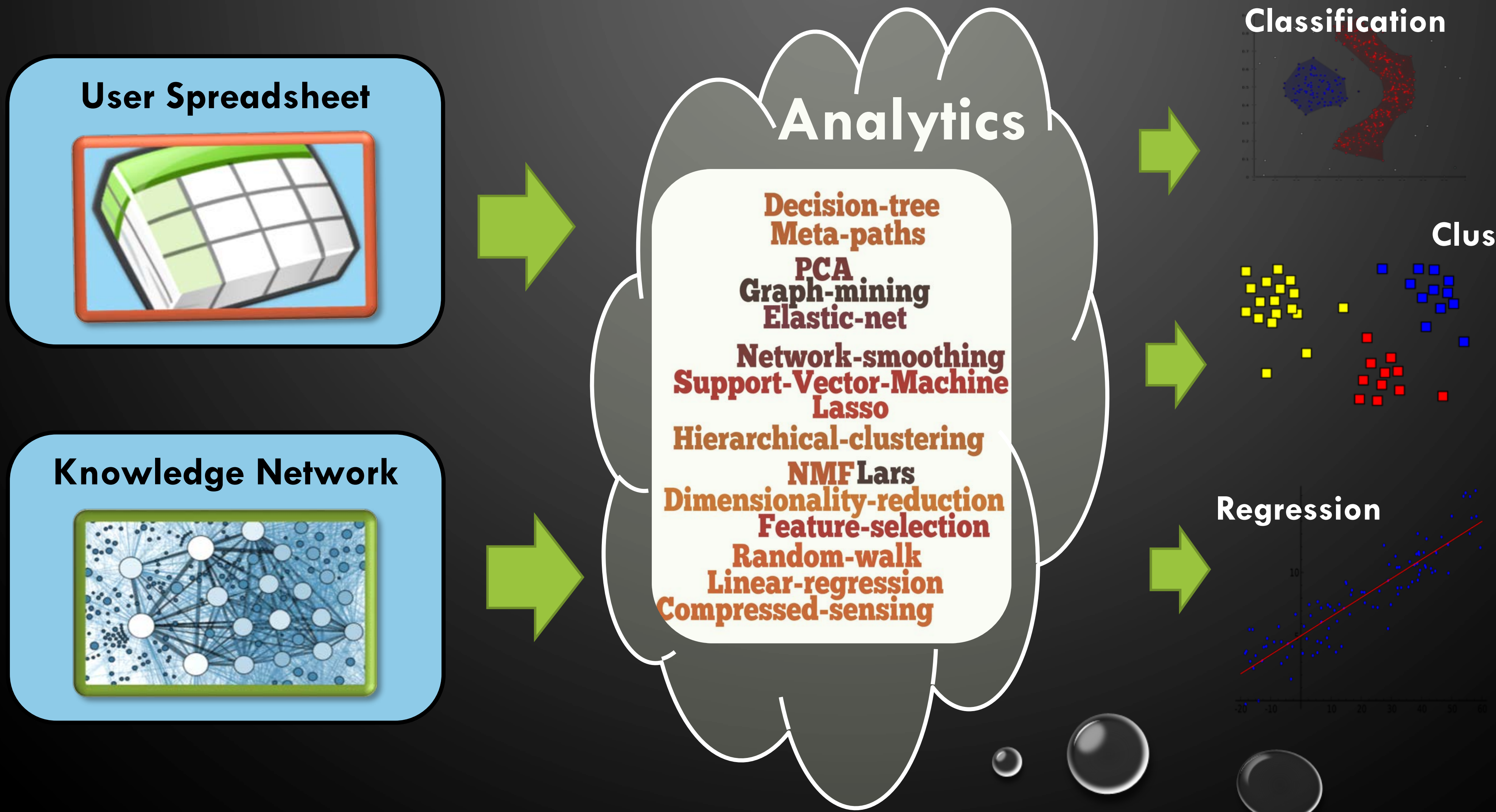
- HIRE BIOINFORMATICIAN OR SEEK BIOINFORMATICS COLLABORATOR.
- DELEGATE THE FOLLOWING:
    - DOWNLOAD AND INSTALL CODE.
    - BUY COMPUTE CLUSTERS
    - RUN CODE ON CLUSTER

IN SHORT, PAINFUL.

A RANDOM WALK *in the* CLOUDS

From the Director of "Like Water for Chocolate"
"Marvelous...
A very special film. Two thumbs up!"
—SISKEL & EBERT
DVD

CLOUD-BASED KNOWLEDGE ENGINE FOR GENOMICS



User Spreadsheet

Knowledge Network

Analytics

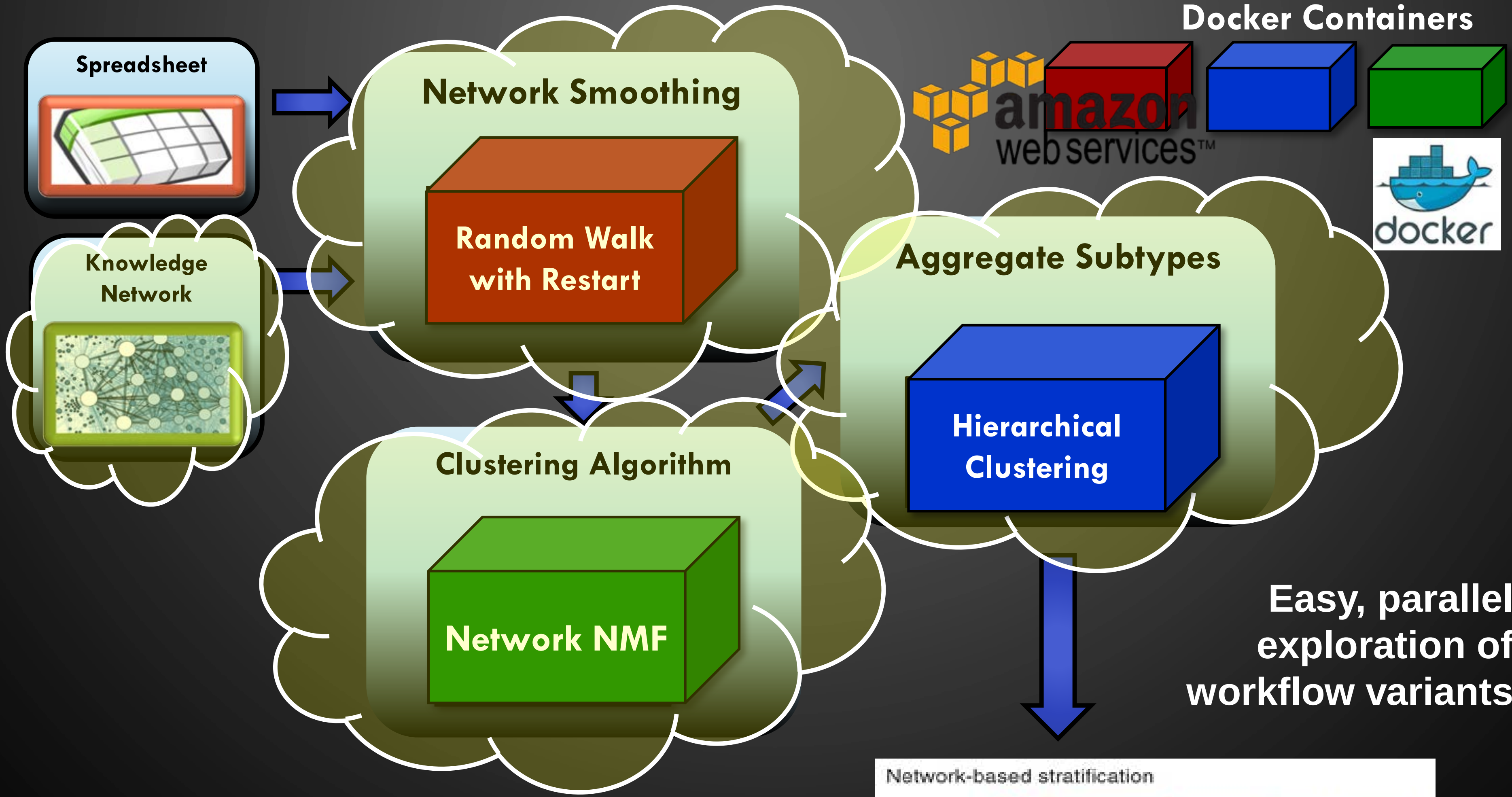
Decision-tree
Meta-paths
PCA
Graph-mining
Elastic-net
Network-smoothing
Support-Vector-Machine
Lasso
Hierarchical-clustering
NMF Lars
Dimensionality-reduction
Feature-selection
Random-walk
Linear-regression
Compressed-sensing

Classification

Clustering

Regression

COMPLEX WORKFLOWS ON THE CLOUD



Docker Containers

Spreadsheet

Knowledge Network

Network Smoothing

Random Walk with Restart

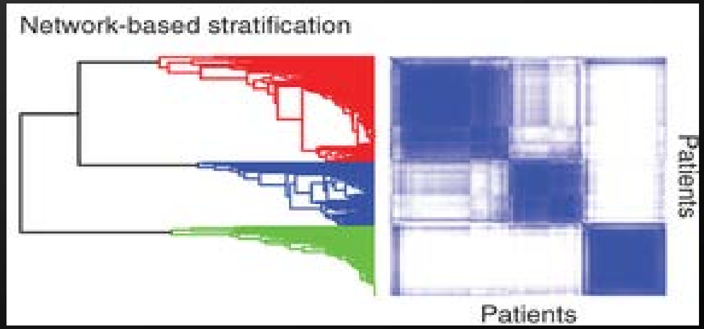
Clustering Algorithm

Network NMF

Aggregate Subtypes

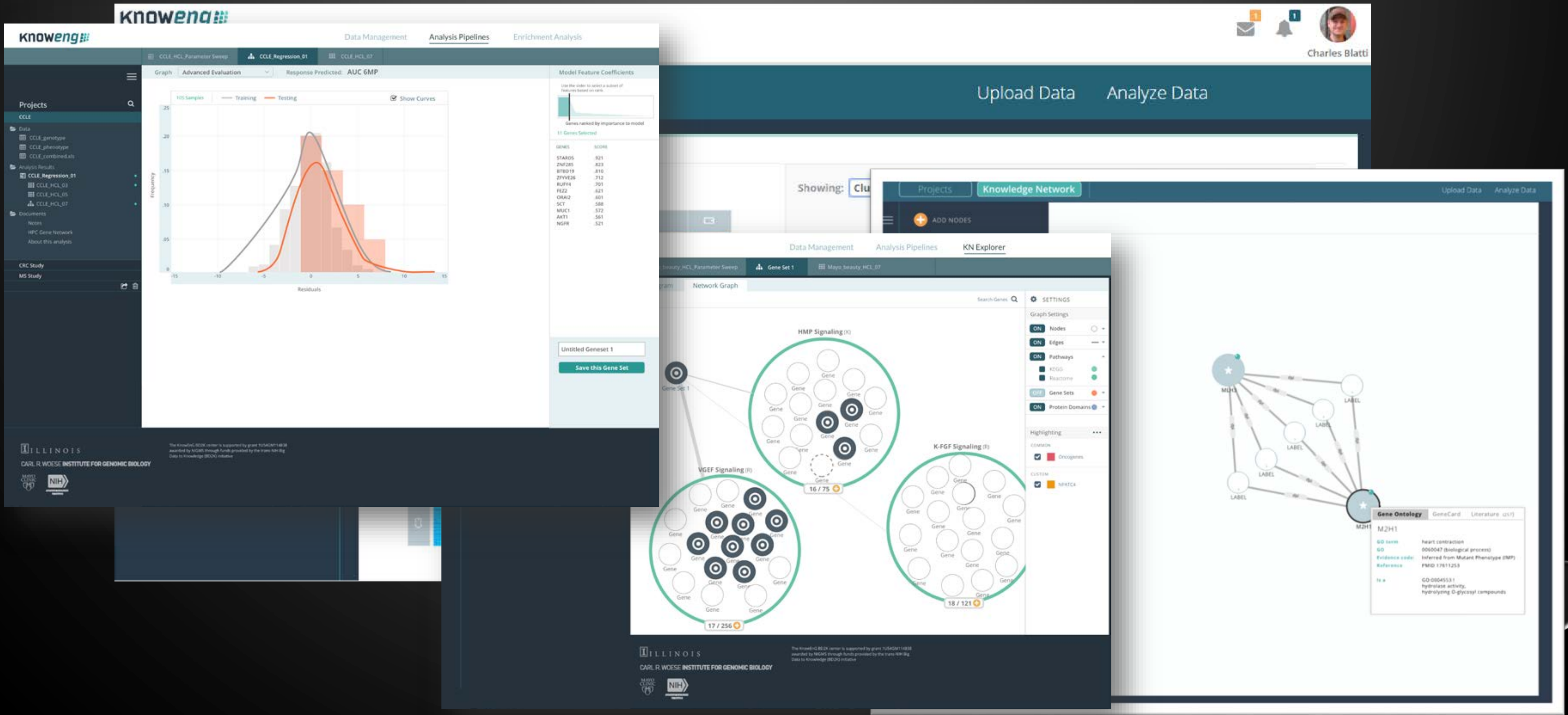
Hierarchical Clustering

Easy, parallel exploration of workflow variants



Network-based stratification of tumor mutations
Hofree et al. *Nature Methods* 2013

SOFTWARE IS ONLY AS GOOD AS ITS FRONT END



Graph: Advanced Evaluation — Response Predicted: AUC 6MP

Model Feature Coefficients

Use the slider to select a subset of features based on rank

Genes ranked by importance to model

11 Genes Selected

GENES	SCORE
STARD5	.921
ZNF285	.823
BTBD19	.810
ZFYVE26	.712
RUFY4	.701
FEZ2	.621
ORAI2	.601
SCT	.588
MUC1	.572
AKT1	.561
NGFR	.521

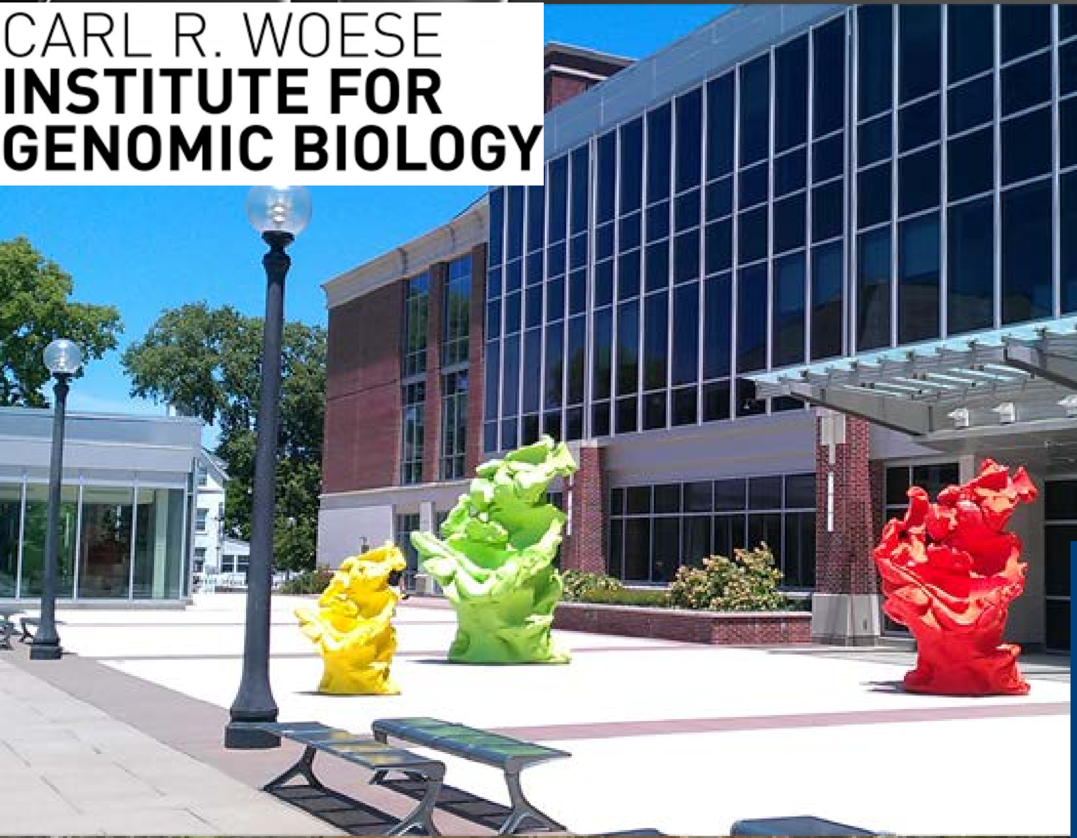
Untitled Geneset 1

Gene Ontology

M2H1

GO term	heart contraction
GO	0060047 (biological process)
Evidence code:	Inferred from Mutant Phenotype (IMP)
Reference	PMID 17611253
Is a	GO:0004553 I hydrolase activity, hydrolyzing O-glycosyl compounds

KNOWENG ET AL.


CARL R. WOESE
INSTITUTE FOR
GENOMIC BIOLOGY

MAYO
CLINIC

NIH

CS @ ILLINOIS