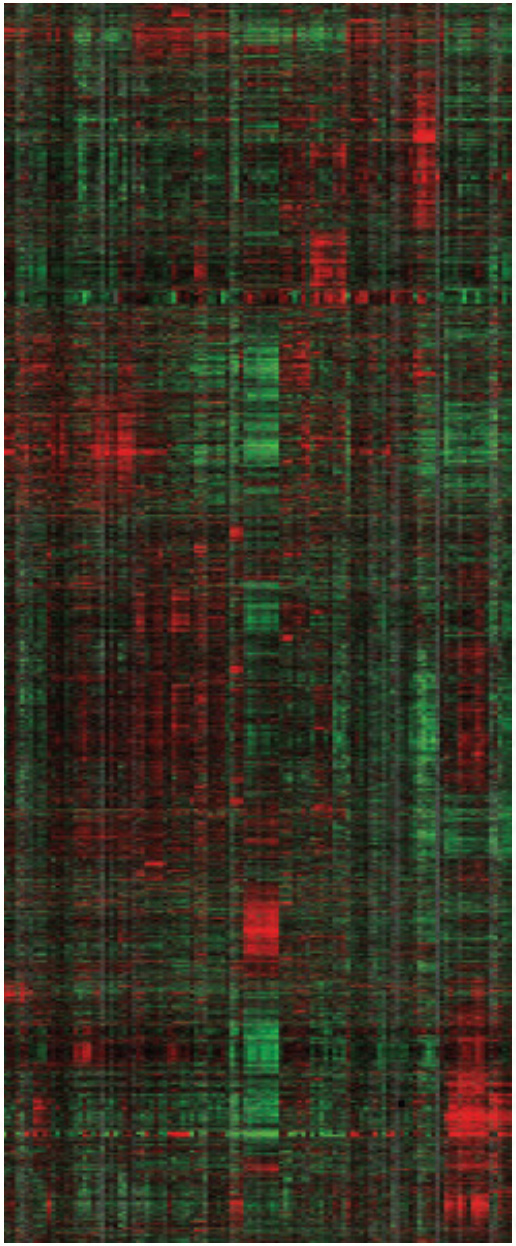


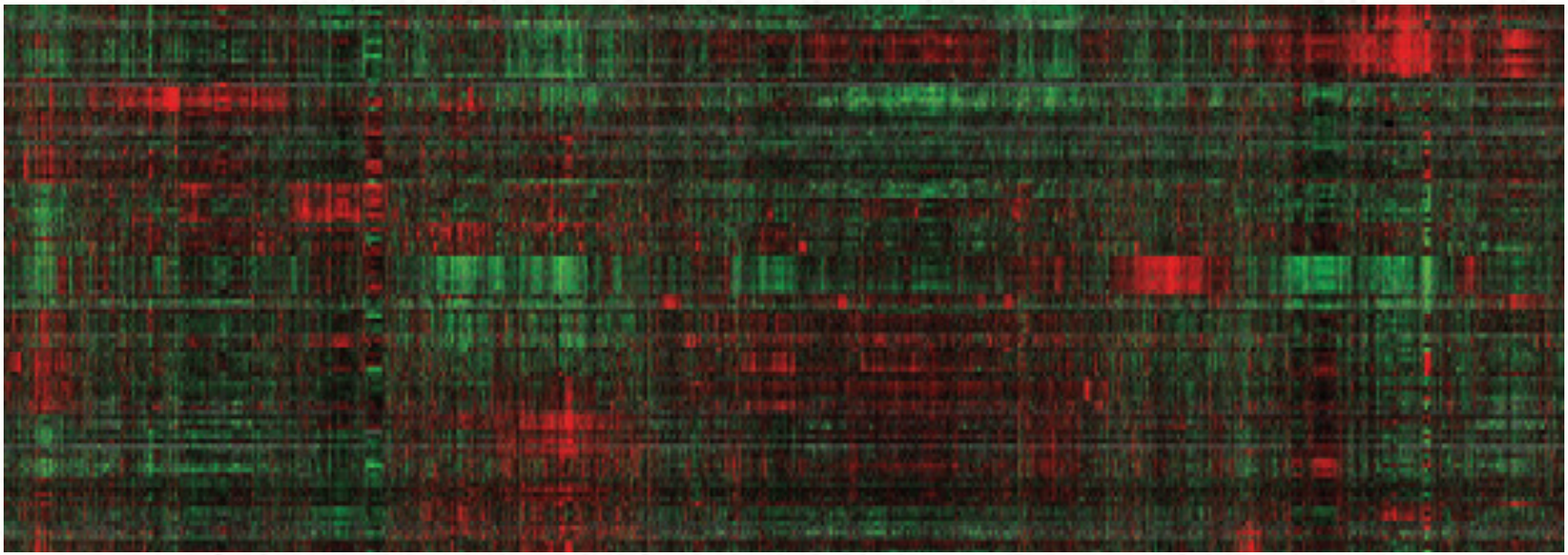
Cloud-Accelerated Data-Mining for Heteromeric Transcription Factor Complexes

Edward A. Salinas, Independent Researcher
Dr. Amitava Karmaker, Univ. WI, Stout, &
Dr. Michael Allan, US Patent Office, Acknowledged

Agenda/Overview

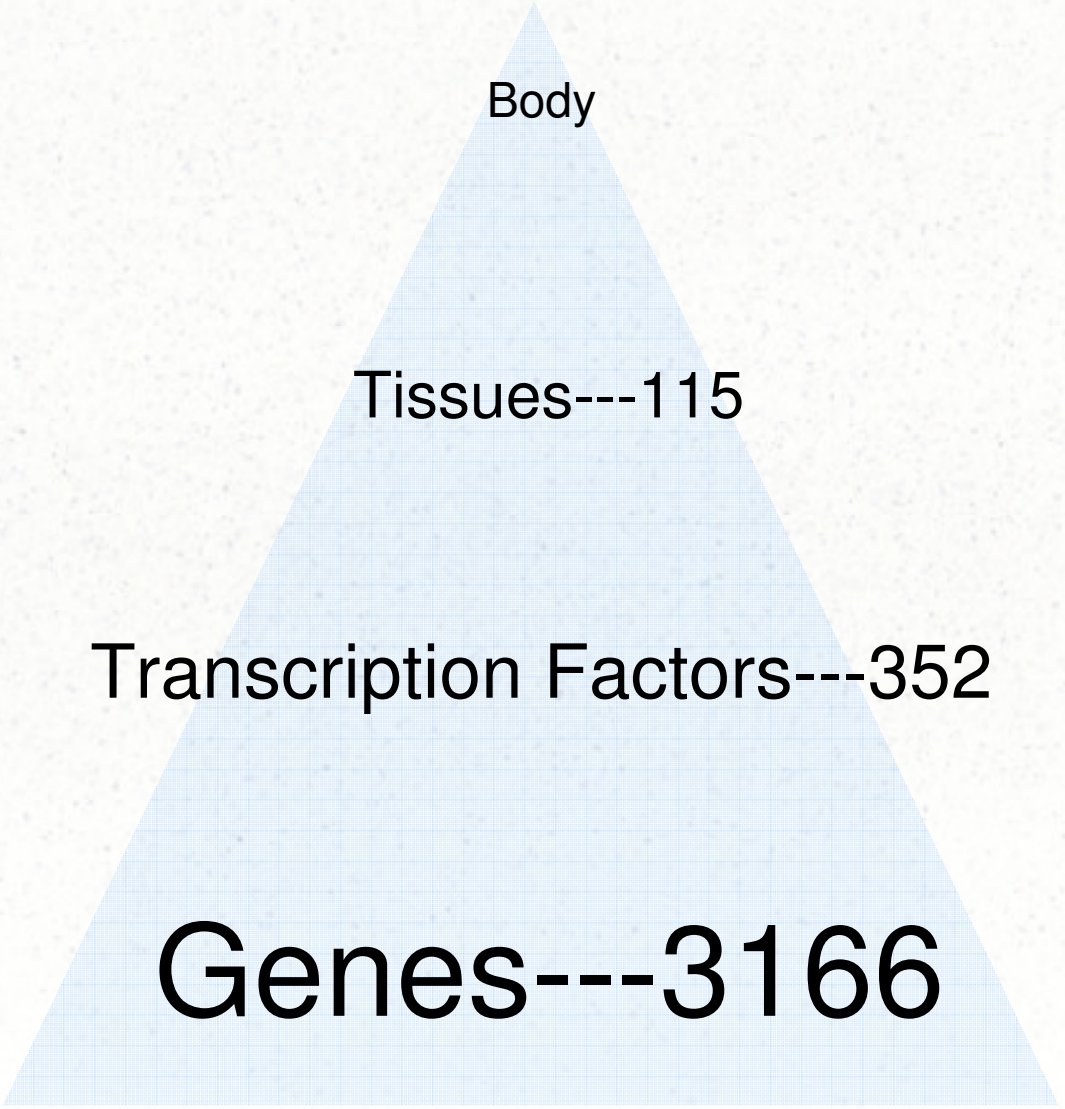
1. Project Overview
2. Technique overview
3. Exploration, validation, and Parametric Calibration analysis with cFos/cJun (AP-1)
4. “k=4” CPU-based approach (execution-time!)
5. “k=4” Cloud-based approach (Amazon EC2)
6. Execution & SpeedUp Discussion
7. Conclusion/Discussion

TF Expression Profile Data Mining May Facilitate ID of TF Complexes



MEANS: Select data, compute hypothetical expression profiles & coefficients, analyze, rank, and find hypothetical polymers.

Our algorithm calls for expression analyses with Gene and TF data.



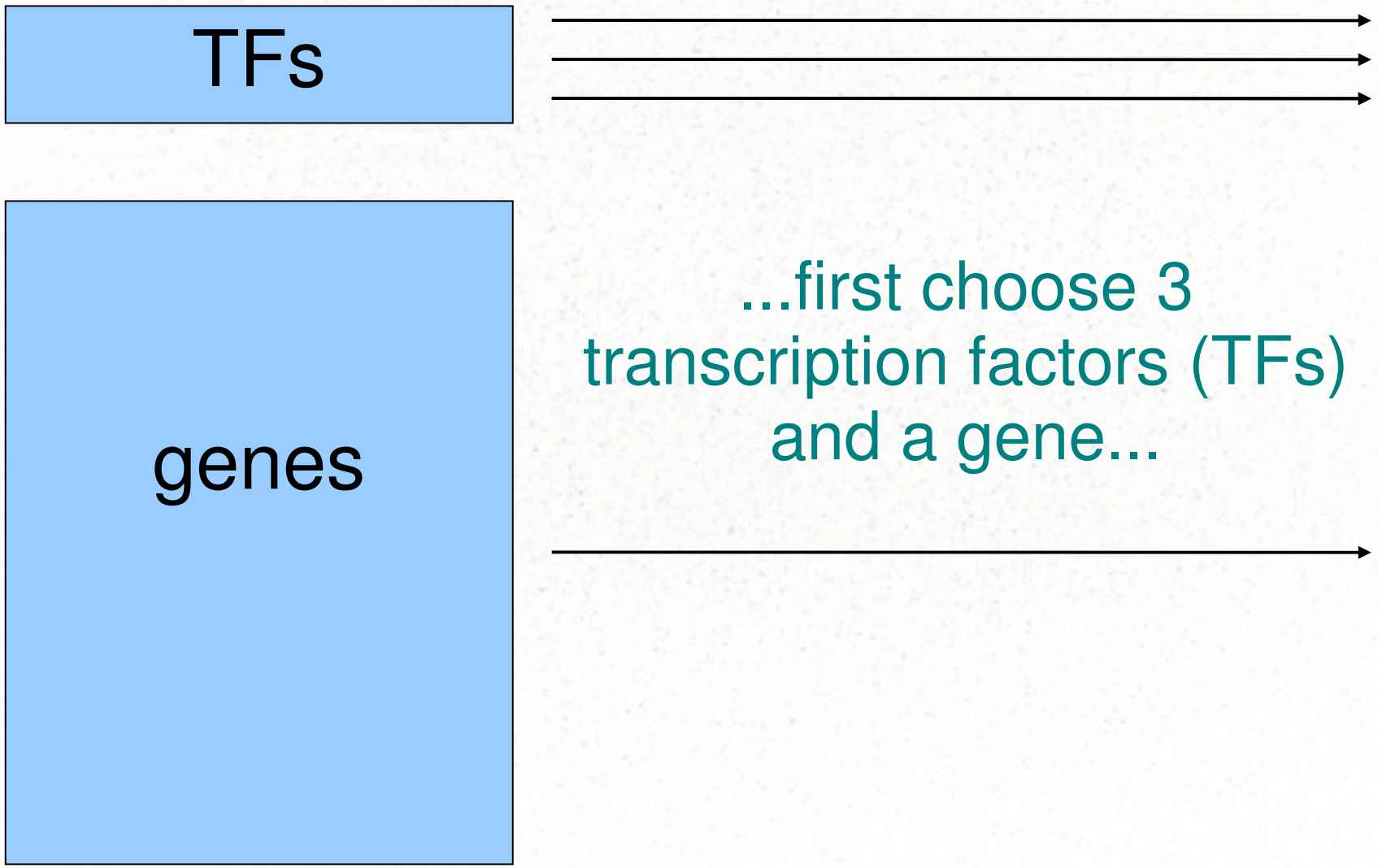
For a given gene (or Transcription Factor(TF)), the data consist of a row of expression values across the tissues. Correlation coefficients are computed between two rows (profiles) of values (a gene and a TF)

	HEART	BRAIN	LYMPH NODE	OVARY
HMGA1	0.1	-0.3	0.7	0.9
CFOS	-0.2	.4	-.6	.2

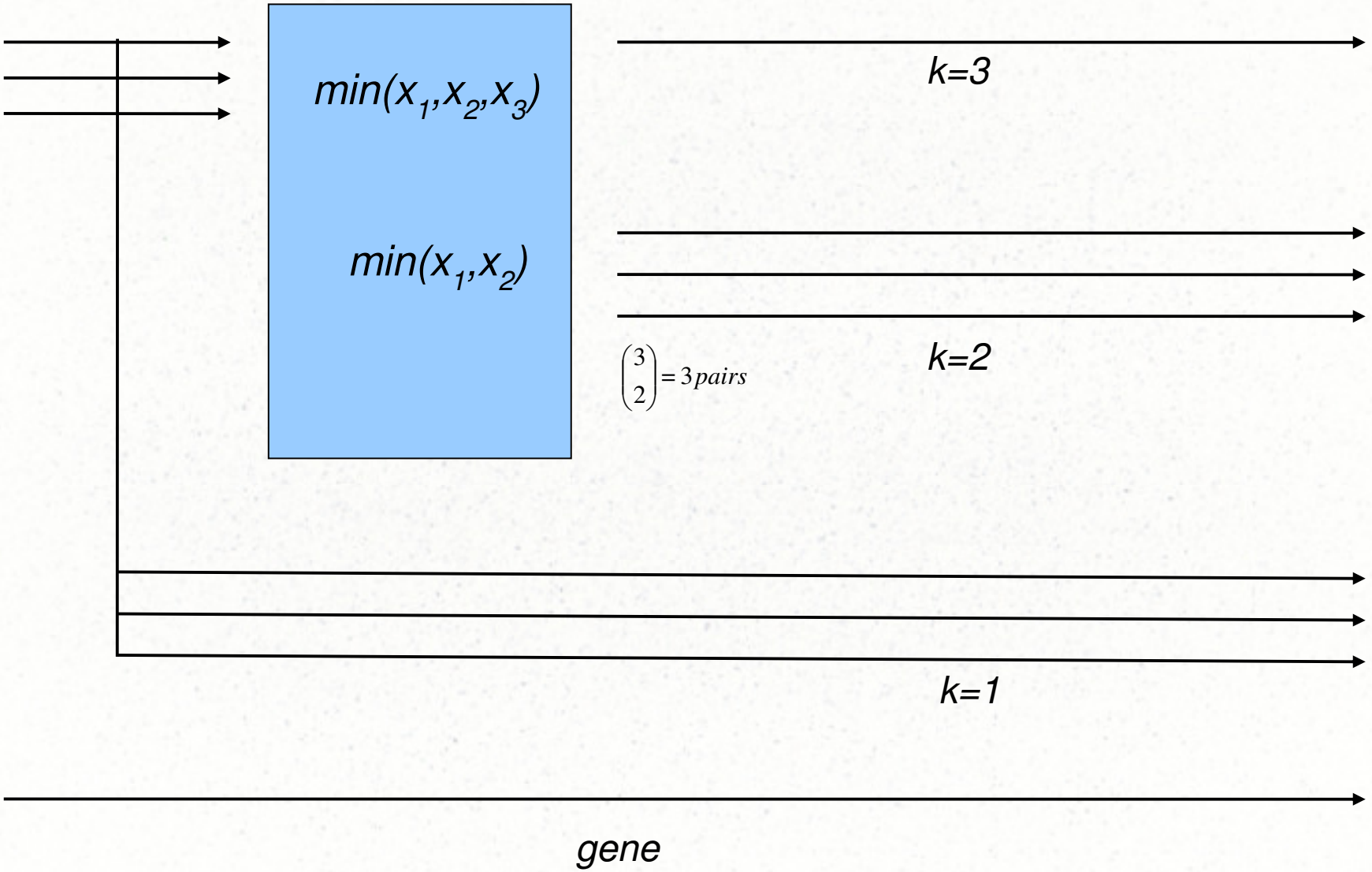
For a given set of TF profiles, a hypothetical TF composite profile is estimated by taking minima across tissues. Motivation: *limiting reactant*

	HEART	BRAIN	LYMPH NODE	OVARY
TF1	0.1	-0.3	0.7	-0.9
TF2	-0.2	0.4	-0.6	0.2
$TF(12)_H$	-0.2	-0.3	-0.6	-0.9

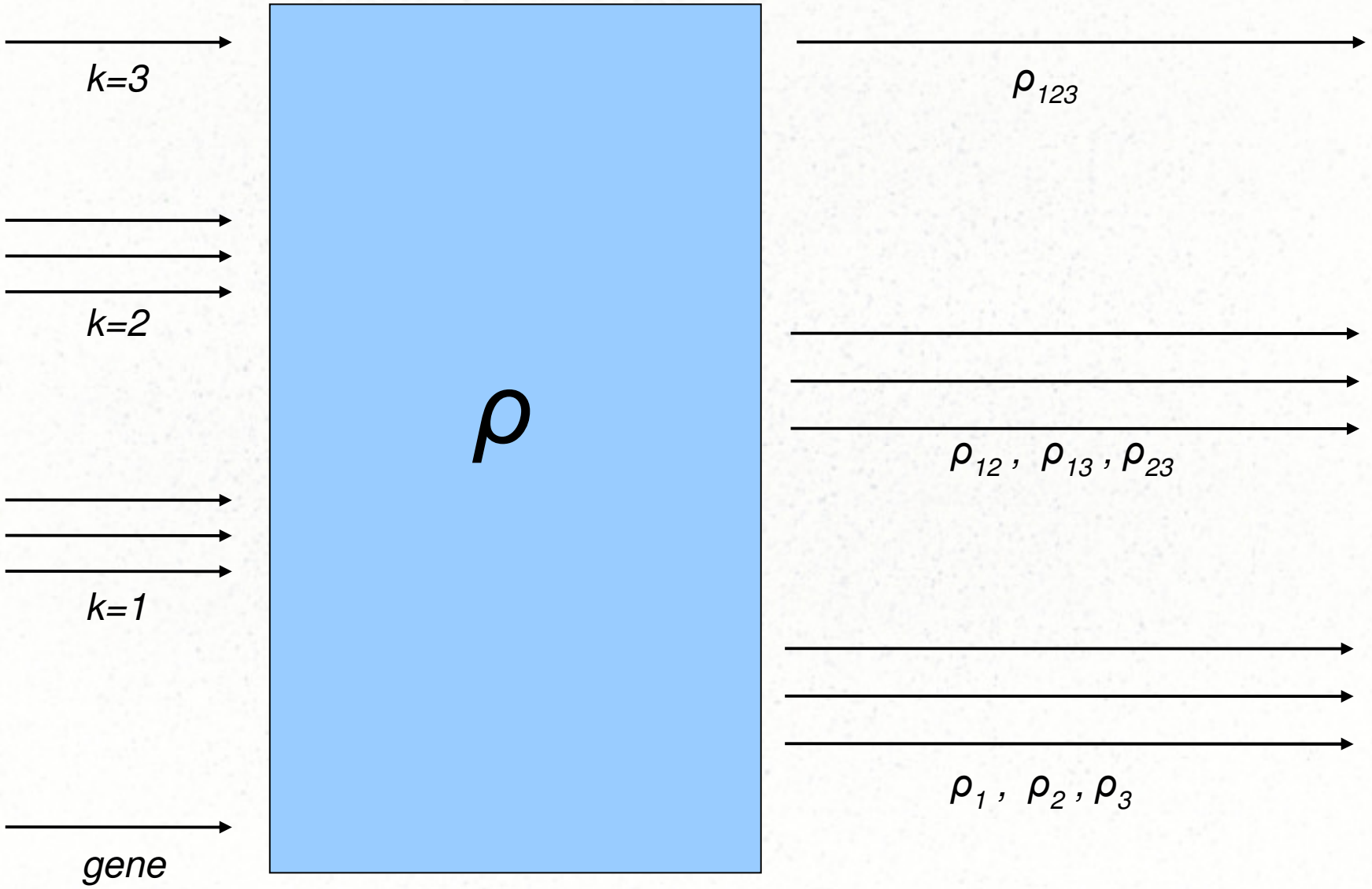
To carry out a k=3 analysis.....



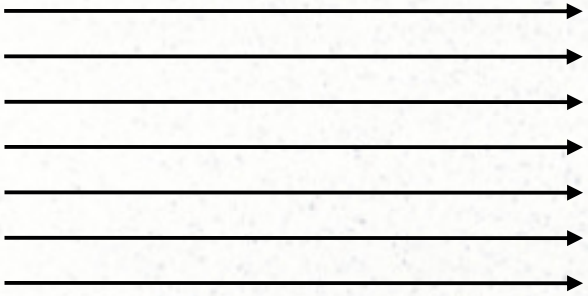
...second, compute hypothetical dimer/trimer profiles...



...third, compute coefficients between the gene profile and each of the 1+3+3=7 TF profiles....



...and finally compare the resulting coefficients to compute an absolute improvement score.



$$\min_{y \neq k_{\max}} \left(\left| \rho_{k_{\max}} - \rho_{y} \right| \right)$$

Repeat for all possible pairs of genes and combinations of selections of 3 TFs if you want to complete a full k=3 analysis!

We used a pair of TFs that are known to dimerize to explore the validity of the algorithm: CFOS & CJUN...

	g	tf1	tf2	c1	c2	cc	i
1	VARS2	FOSL1	JUN	0.43135	-0.318	0.0735	0.3578
2	RNU3IP2	FOSL1	JUN	0.50372	-0.228	0.1559	0.3479
3	ZFX	FOSL1	JUN	-0.4036	0.369	-0.0594	0.3442
4	AP2S1	FOSL1	JUN	0.46311	-0.212	0.1232	0.3355
5	LRP6	FOSL1	JUN	-0.3653	0.375	-0.0468	0.3185
6	MAP4K5	FOSL1	JUN	-0.3528	0.318	-0.0351	0.3177
7	LOC56902	FOSL1	JUN	0.41646	-0.225	0.0893	0.3141
8	EGR1	FOSL1	JUN	-0.0400	0.614	0.3041	0.3101
9	HMGA1	FOSL1	JUN	0.40596	-0.250	0.0549	0.3046
10	TAPBP	FOSL1	JUN	0.38650	-0.223	0.0802	0.3035

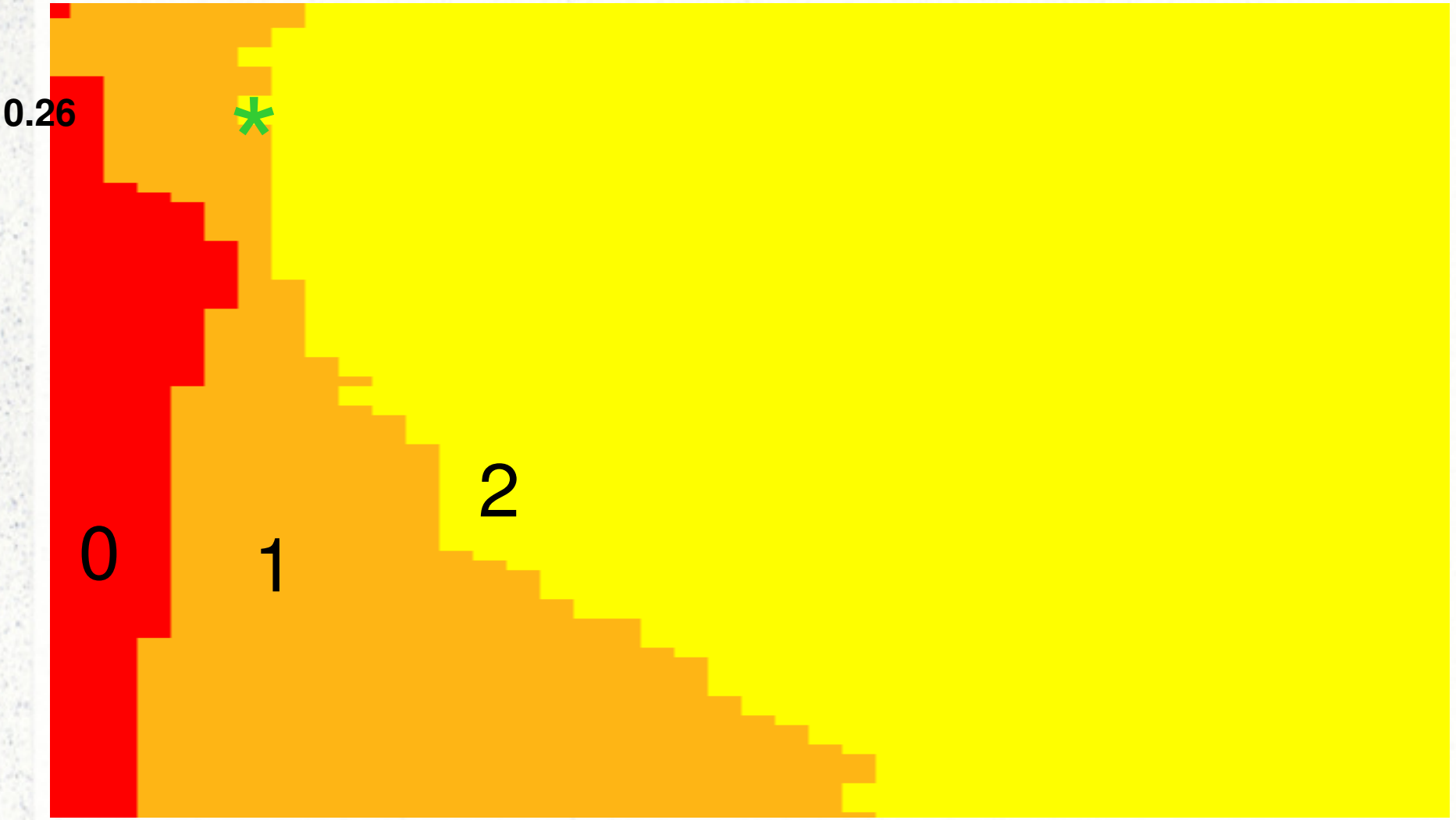
$$P(X \geq 2) = P = 8.4 * 10^{(-5)}$$

2 target genes (“true positives” out of 4 total) found in top 10 of 3166 genes

There's a *low* probability of a result at least as extreme happening by random chance!

Reject H_0 at $\alpha = 1\%$; *Here, the alpha parameter for transformation was 0.5*

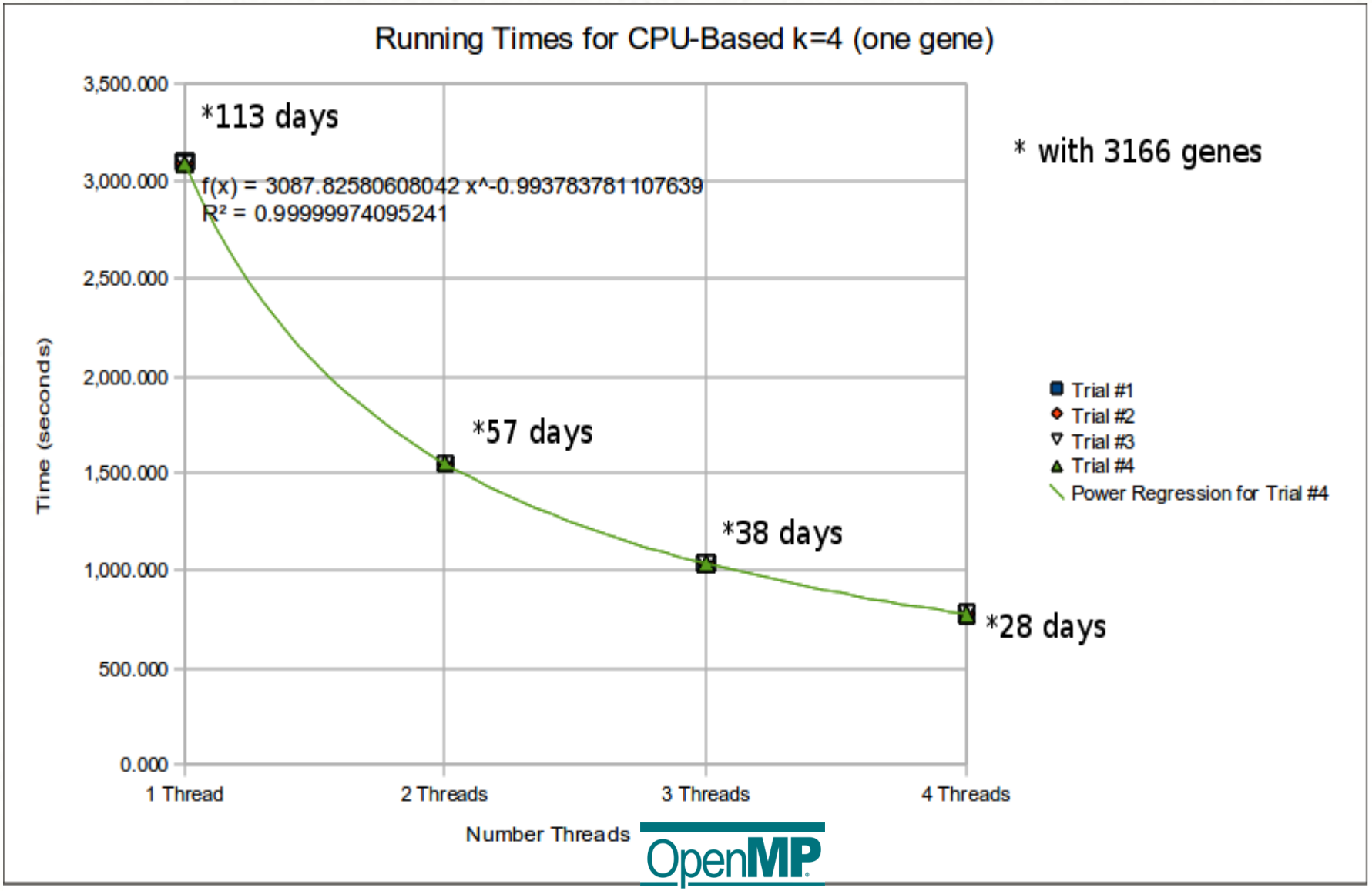
…as we calibrated alpha to be *0.26* to maximize the number of true targets in the top 10 list of target genes with the CFOS/CJUN/AP1 dataset.



Searching for higher-order polymers means “deeper data-mining”...

k	Number Coefficients to Compute
1	1,114,432
2	196,380,648
3	23,014,375,848
4	2,013,884,773,648
5	140,578,442,425,168

...a full k=4 analysis would take a long time.....



...so we decided to apply a heterogeneous set of parallel computing technologies to speed things up!

A systematic approach to deploying on the Amazon EC2 Cloud using the STAR Cluster Toolkit…

1. testing/porting & basic cloud allocation

2. parallelized code on multi-core AMIs

3. deployment with extensive resource allocation for large speedup (264 8-core nodes)

amazon webservices™

OpenMP

…And led us, in about 24 hours, to these top-10 results.

	AI	GENE	TF1	TF2	TF3	TF4
1	0.71	FGB	IRF1	MGA	PAPOLA	SNAPC3
2	0.67	FGB	E2F5	ILF3	MGA	SP110
3	0.67	FGB	EPC1	ITGB3BP	PAPOLA	SP110
4	0.67	FGB	SDCCAG33	TWISTNB	ZNF155	ZNF83
5	0.66	FGB	ILF3	MGA	NFYA	SP110
6	0.65	AFP	ILF3	MGA	SP110	ZNF83
7	0.65	EHHADH	ELL2	EWSR1	PCAF	PPARBP
8	0.65	FGB	IRF1	ITGB3BP	PAPOLA	SNAPC3
9	0.65	FGB	PRKAR1A	TWISTNB	ZNF155	ZNF83
10	0.64	FGB	E2F5	IRF1	ITGB3BP	PAPOLA

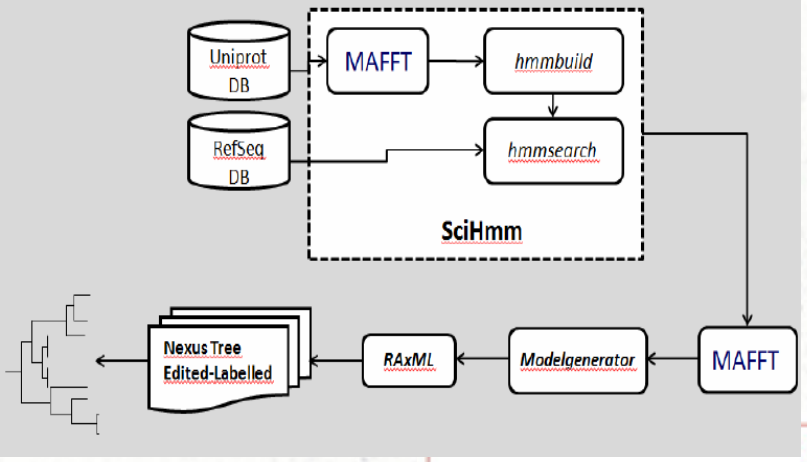
Transcription Factors and their targets were diverse

- FGB: fibrinogen beta chain (role in blood clotting)
- EPC1: part of HAT (histone acetyltransferase)
- SP110 : possible hormonal transcriptional co-activator
- IRF1 : malfunction or mutation in this gene may lead to cancer

Future possible projects and/or directions in this thread of research

- Improved handling of missing values (imputation)
- Extended algorithm for handling degrees of homopolymerity
- MAAS? (“micro-array analysis as a service”) www.minemyarray.org?

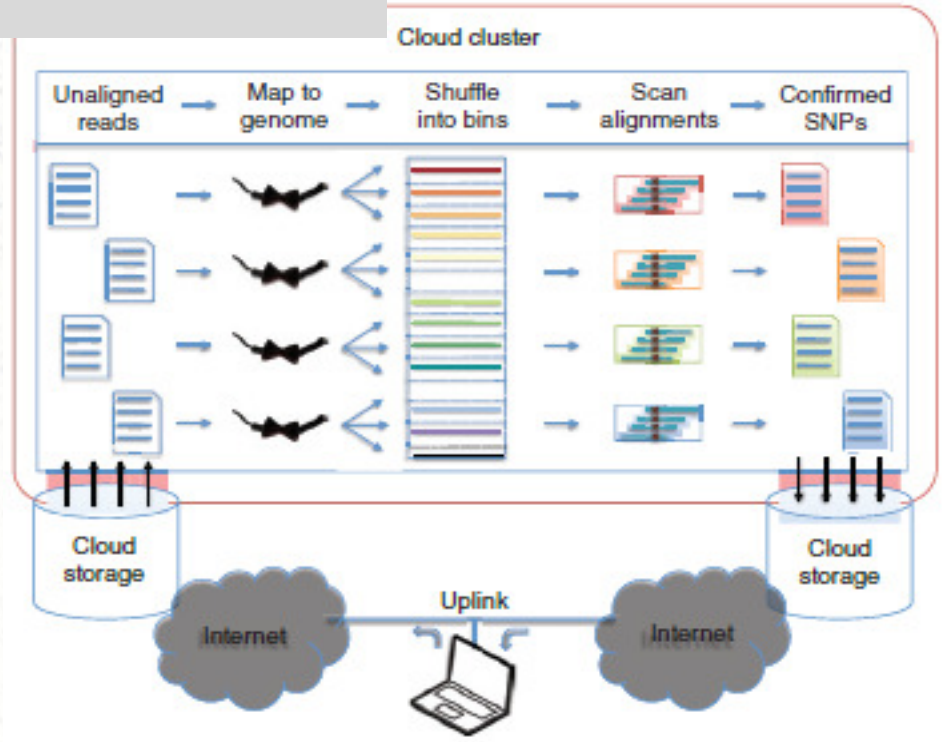
Example Cloud/Bioinformatics Applications



Phylogenetic Analysis with SciPhy

CloVR

BLAST & Assembly



Crossbow Read Mapping

THANK YOU!
QUESTIONS?