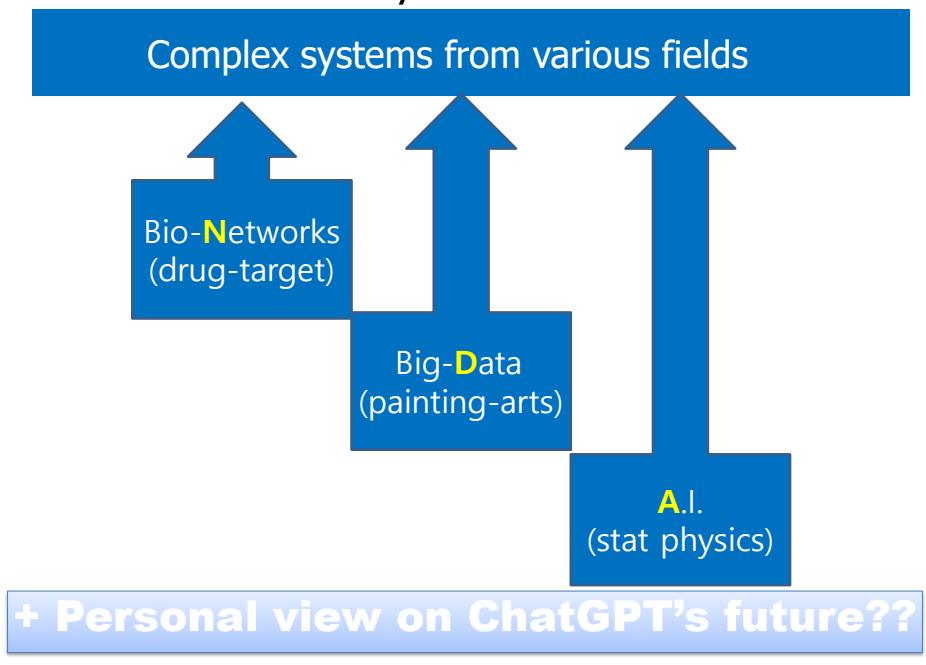
Understanding of Complex Systems

via N.N.A. (Network, Network, AI)

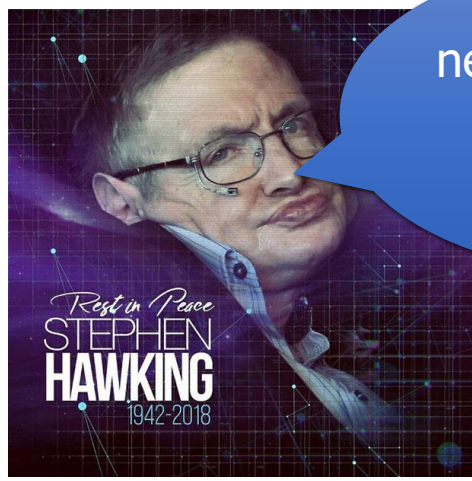


Hawoong Jeong (夏雄 鄭)
Dept. of Physics
Korea Advanced Institute of Science and Technology (KAIST)

Outline of the today's talk



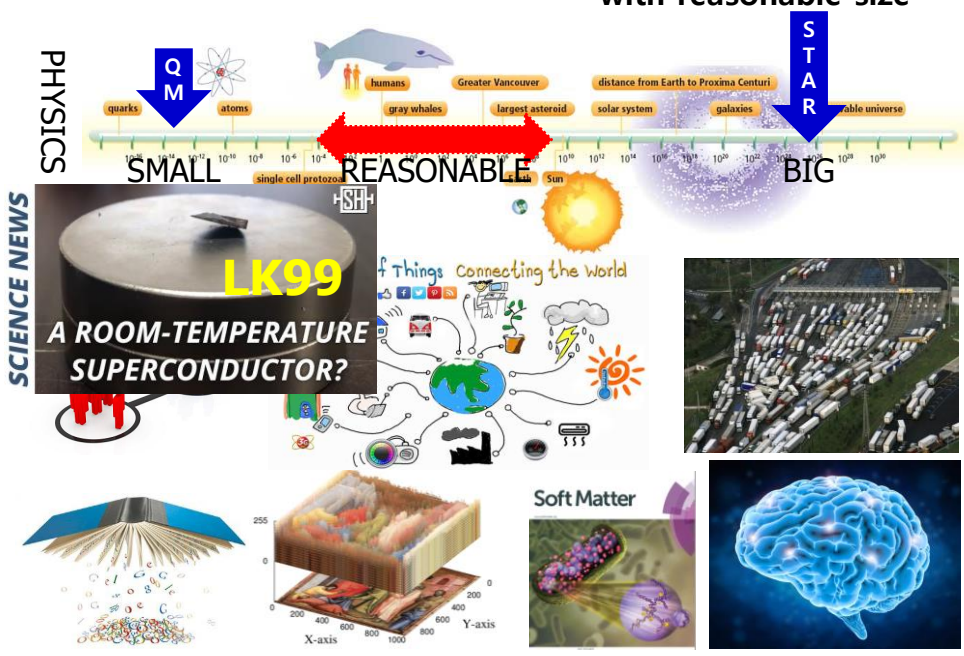
Q: Some say that while the twentieth century was the century of physics, we are now entering the century of biology. What do you think of this?



A: I think the next 21st century will be the century of complexity.



My Research Interest: **Understanding Complex Systems with reasonable size**



Networks Part 0

Why networks?

Ubiquitous networks are all around us.

NETWORKS ARE EVERYWHERE,
from micro to macro

- **Brain = Neural networ**
- **Metabolism, Cell** **etworks**
- **Food web =**
- **Society**
- **Hi** **ransportation networks**
- **Int** **= Communication networks**
- **Mon** **= Economic networks**

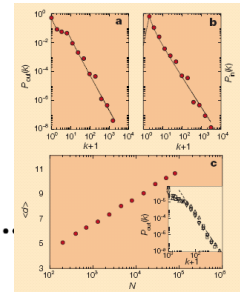
First Question:
How do they look like?

We're living in a "connected world".

Internet

Diameter of the World-Wide Web

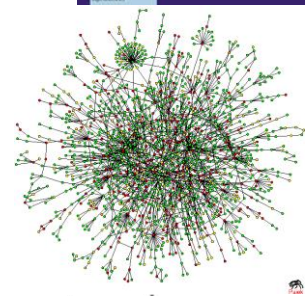
Réka Albert, Hawoong Jeong, Albert-László Barabási
Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556, USA





Error and attack tolerance of complex networks

Réka Albert, Hawoong Jeong & Albert-László Barabási



The large-scale organization of metabolic networks

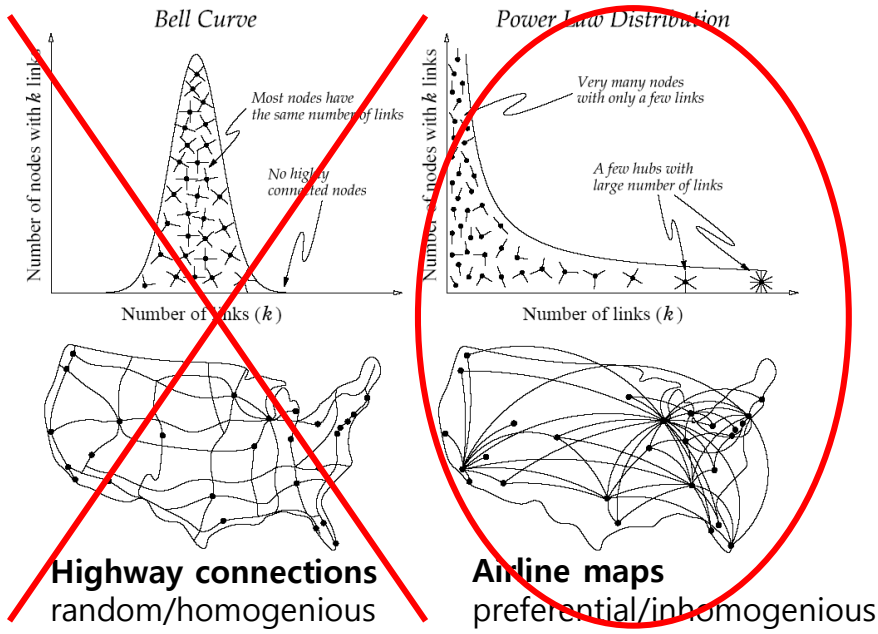
H. Jeong*, B. Tombor†, R. Albert*, Z. N. Oltvai† & A.-L. Barabási*

Lethality and centrality in protein networks

The most highly connected proteins in the cell are the most important for its survival.

H. Jeong*, S. P. Mason†, A.-L. Barabási*, Z. N. Oltvai†

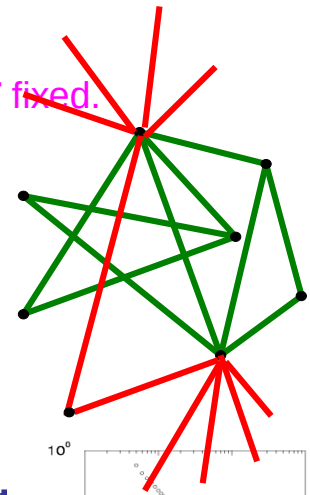
Two types of network structure/topology



Highway connections
random/homogenious

Airline maps
preferential/inhomogenious

HOW?



(1) The number of nodes (N) is **NOT** fixed.

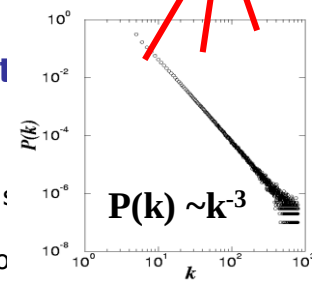
Networks continuously expand by the addition of new nodes

Examples:
WWW : addition of new documents
Citation : publication of new papers

(2) The attachment is **NOT** uniform.
(Rich gets Richer!!)

A node is linked with higher probability t already has a large number of links.

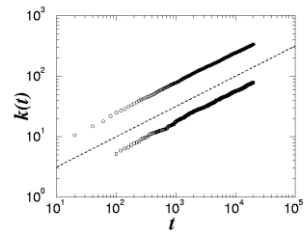
Examples :
WWW : new documents link to well known s (CNN, YAHOO, NewYork Times, etc)
Citation : well cited papers are more likely to



Continuum Theory

$$\frac{\partial k_i}{\partial t} \propto \Pi(k_i) = A\frac{k_i}{\sum_j k_j} = \frac{k_i}{2t} \quad \text{, with initial condition} \quad k_i(t_i) = m$$

$$k_i(t) = m\sqrt{\frac{t}{t_i}}$$



$$P(k_i(t) < k) = P_t(t_i > \frac{m^2 t}{k^2}) = 1 - P_t(t_i \le \frac{m^2 t}{k^2}) = 1 - \frac{m^2 t}{k^2(m_0 + t)}$$

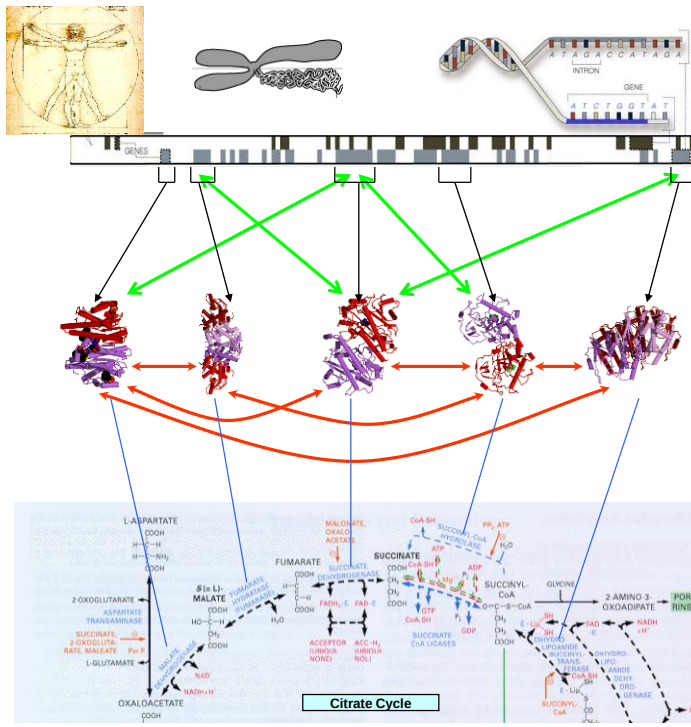
$$\therefore P(k) = \frac{\partial P(k_i(t) < k)}{\partial k} = \frac{2m^2 t}{m_o + t}\frac{1}{k^3} \sim k^{-3}$$

$$\gamma = 3$$

A.-L.Barabási, R. Albert and H. Jeong, Physica A **272,** 173 (1999)

Networks Part 1

Applications in biology



Network Biology

GENOME

protein-gene interactions

PROTEOME

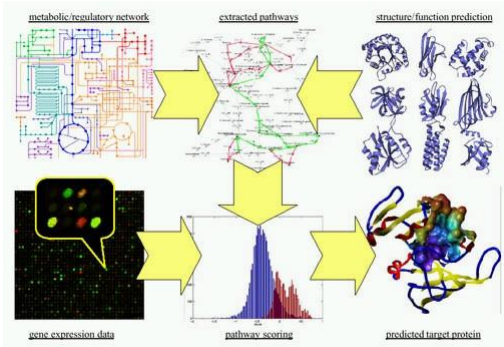
protein-protein interactions

METABOLISM

Bio-chemical reactions

Major goal of pharmaceutical bioinformatics

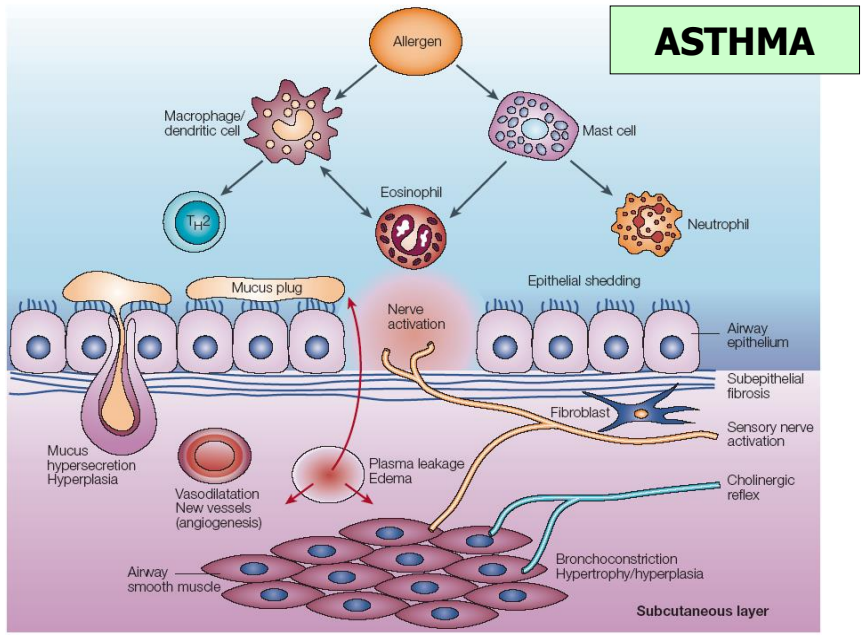
Developing computational tools for systematic *in silico* molecular **drug target identification**



In the TargId project, new bioinformatics methods combine heterogeneous information in the search for drug target proteins.

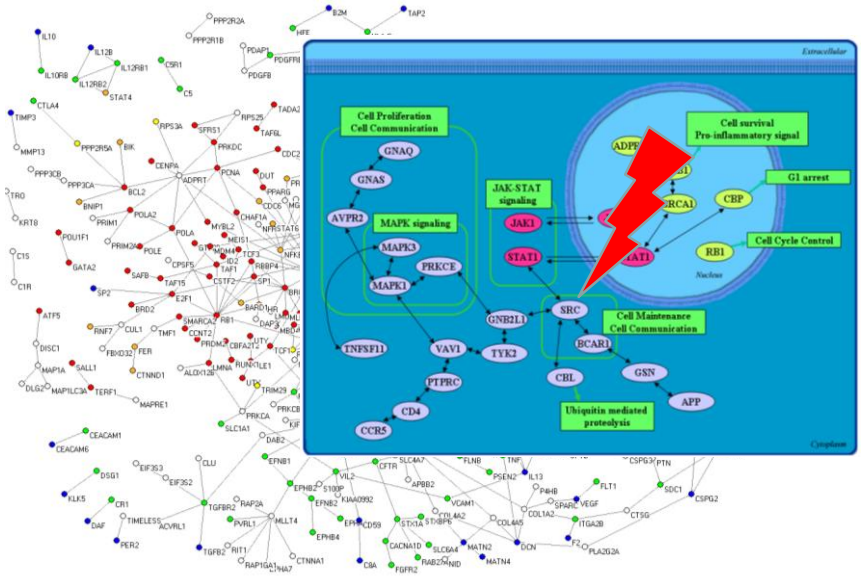
Based on

molecular biology information in database
&
unstructured information hidden in the literature



S.W. Son, **H.Jeong** et al, J. Theo. Biology **252** 722 (2008)

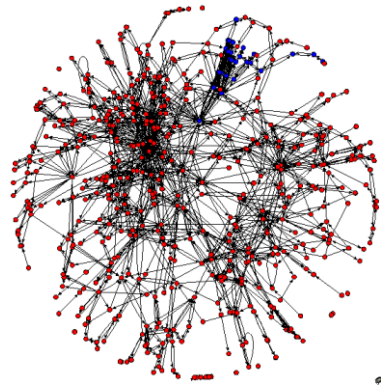
Asthma (related) network

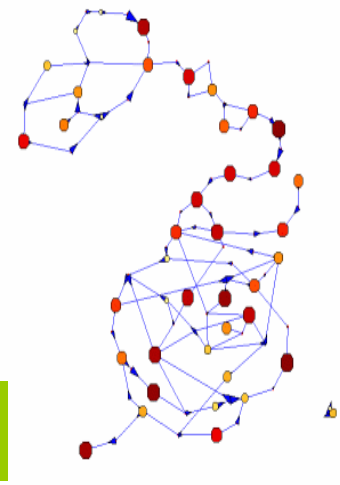


S.W. Son, **H.Jeong** et al, J. Theo. Biology **252** 722 (2008)

LITTLE PROBLEM!

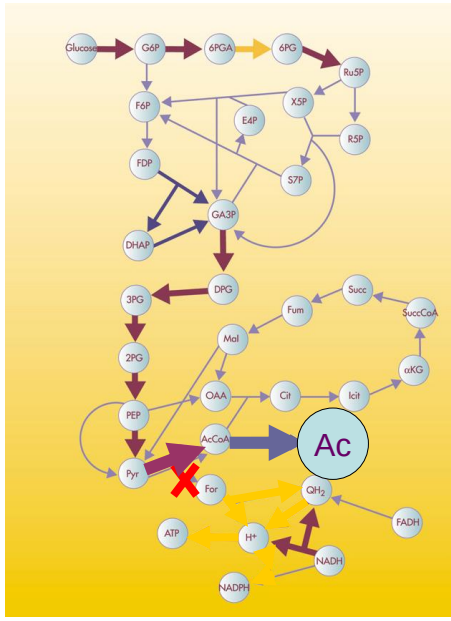
MOST NETWORKS ARE NOT STATIC, THEY'RE DYNAMIC!



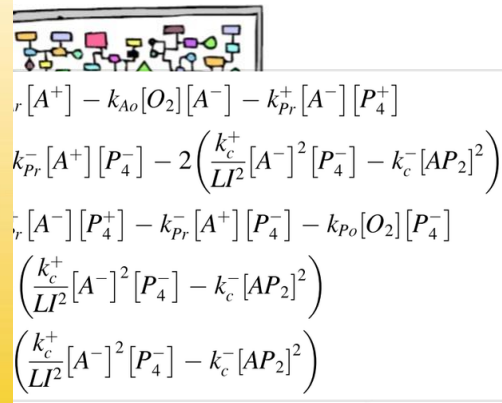


E.g. real metabolic networks are DYNAMIC!!

Metabolic Engineering



but dynamic!



$$\dot{r}\left[A^{+}\right]-k_{Ao}[O_2]\left[A^{-}\right]-k_{Pr}^{+}\left[A^{-}\right]\left[P_4^{+}\right]$$

$$k_{Pr}^{-}\left[A^{+}\right]\left[P_4^{-}\right]-2\left(\frac{k_c^{+}}{LI^2}\left[A^{-}\right]^2\left[P_4^{-}\right]-k_c^{-}\left[AP_2\right]^2\right)$$

$$\dot{r}\left[A^{-}\right]\left[P_4^{+}\right]-k_{Pr}^{-}\left[A^{+}\right]\left[P_4^{-}\right]-k_{Po}[O_2]\left[P_4^{-}\right]$$

$$\left(\frac{k_c^{+}}{LI^2}\left[A^{-}\right]^2\left[P_4^{-}\right]-k_c^{-}\left[AP_2\right]^2\right)$$

$$\left(\frac{k_c^{+}}{LI^2}\left[A^{-}\right]^2\left[P_4^{-}\right]-k_c^{-}\left[AP_2\right]^2\right)$$

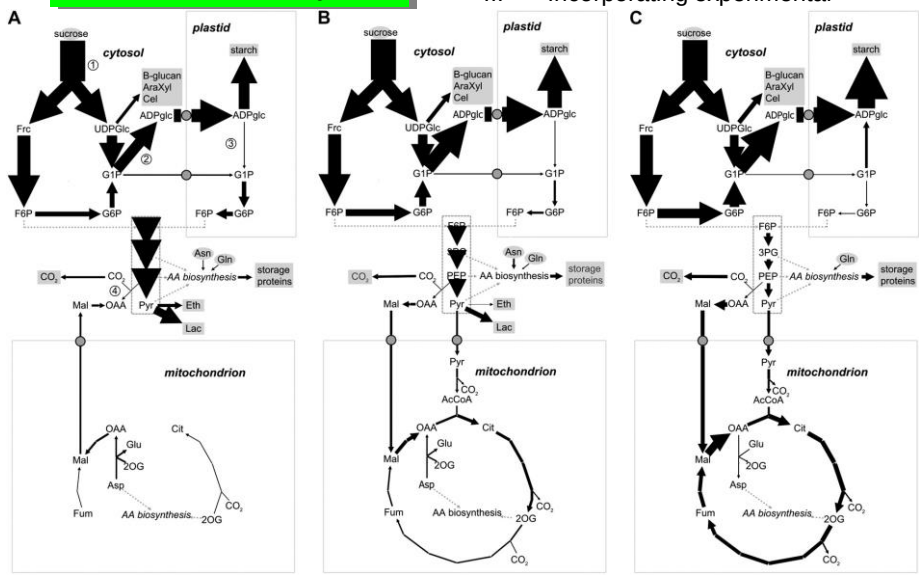
"And that's why we need a computer."

— Courtesy: Keck Graduate Institute

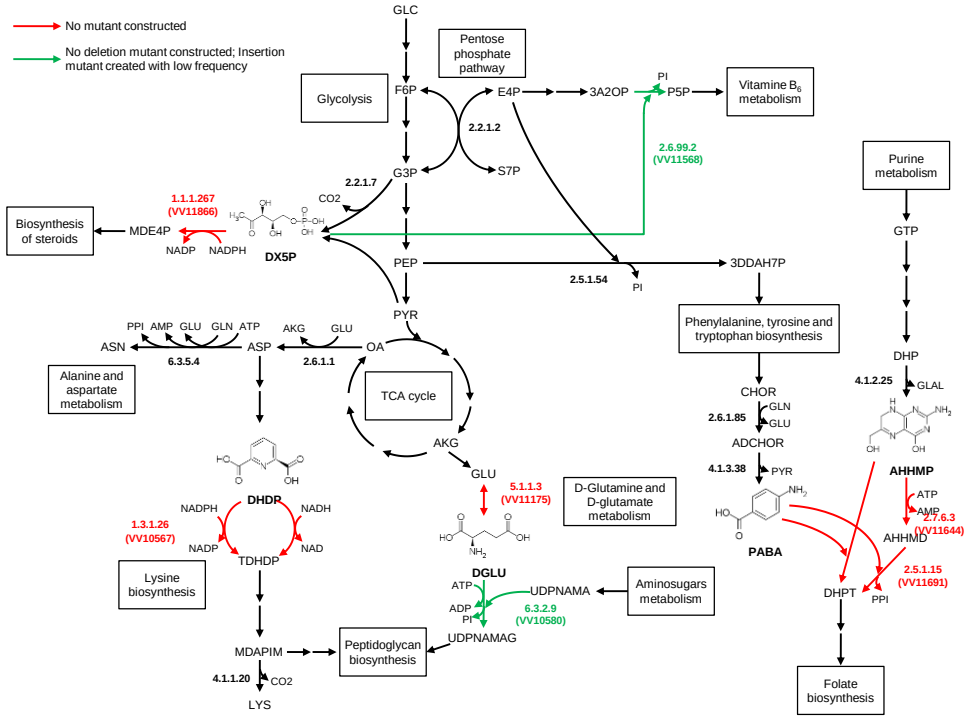
Simple approximation ~ Flux Balance Analysis

I. Stoichiometric matrix construction

II. Incorporating experimental



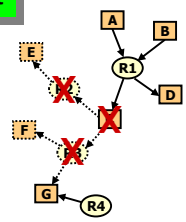
The immediate goal is to identify the steady state of the system.



in silico experiment

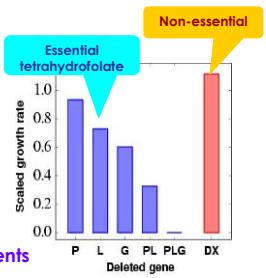
Determination of metabolite-essentiality E_M

$$E_M = \frac{\Delta \text{growth}}{\text{growth}}$$



~ 250 essential metabolites

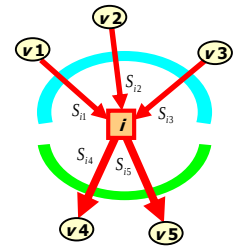
Name of metabolite	No. of participating reactions		
	total	incoming	outgoing
H_2O	407	378	226
H^+	293	82	231
ATP	176	38	174
Phosphate	149	143	35
ADP	143	141	35
Nicotinamide adenine dinucleotide	77	42	51
Nicotinamide adenine dinucleotide - reduced	72	48	40
CO_2	60	58	8
Nicotinamide adenine dinucleotide phosphate	52	45	20
Pyruvate	51	43	14
Nicotinamide adenine dinucleotide phosphate - reduced	50	44	44
L-Glutamate	46	39	14
Coenzyme A	39	38	7
ammonium	38	34	7
AMP	36	33	8
Acetyl-CoA	26	14	18
2-Oxoglutarate	25	17	20
acyl carrier protein	23	18	8
Phosphoenolpyruvate	23	4	21
L-Aspartate	20	8	15
L-Glutamine	15	2	13
Glyceraldehyde 3-phosphate	14	13	10
CMP	14	12	6
Glycerol 3-phosphate	14	10	6



Experiments

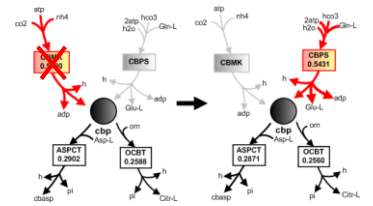
P: *purN*, L: *lpdA*, G: *glyA*, PL: *purN* + *lpdA*
PLG: *purN* + *lpdA* + *glyA*, DX: *dxs* + *xylB*

Can we test via wet-exp?



Metabolite mass-flow
= flux-sum (Φ)

$$\Phi_i = \frac{1}{2}\sum_j |S_{ij} v_j|$$

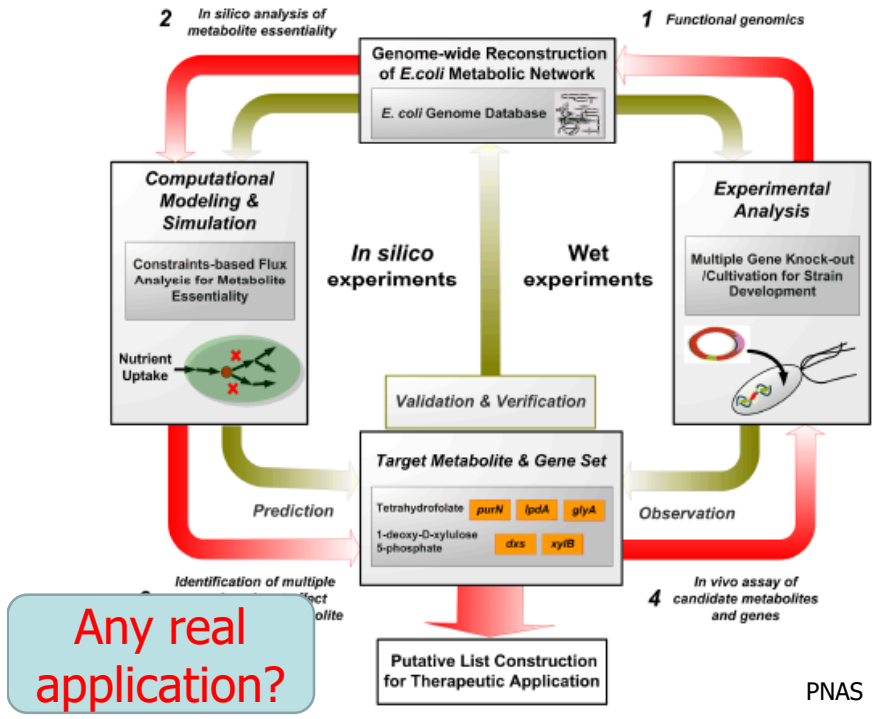


Essential metabolites maintain their flux-sums towards cellular robustness.

Attenuating such maintenance suppresses the cell growth.

=> Multiple gene knockout/knockdown around essential metabolites can be applied for antimicrobial strategy.

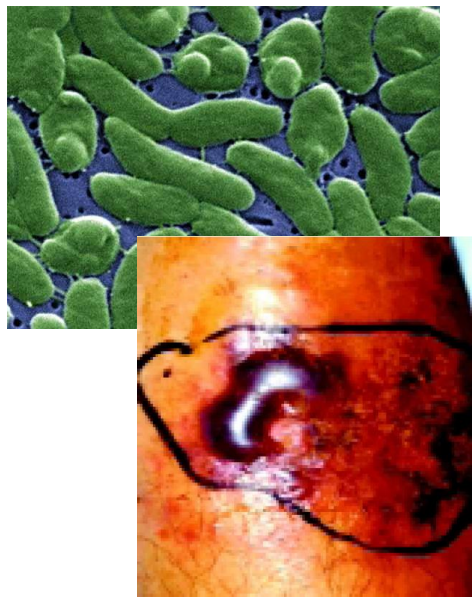
PNAS (2008)



Any real application?

PNAS (2008)

Characteristics of *Vibrio vulnificus* CMCP6



S.Y. Lee's group (KAIST)

✓ Gram-negative bacterium

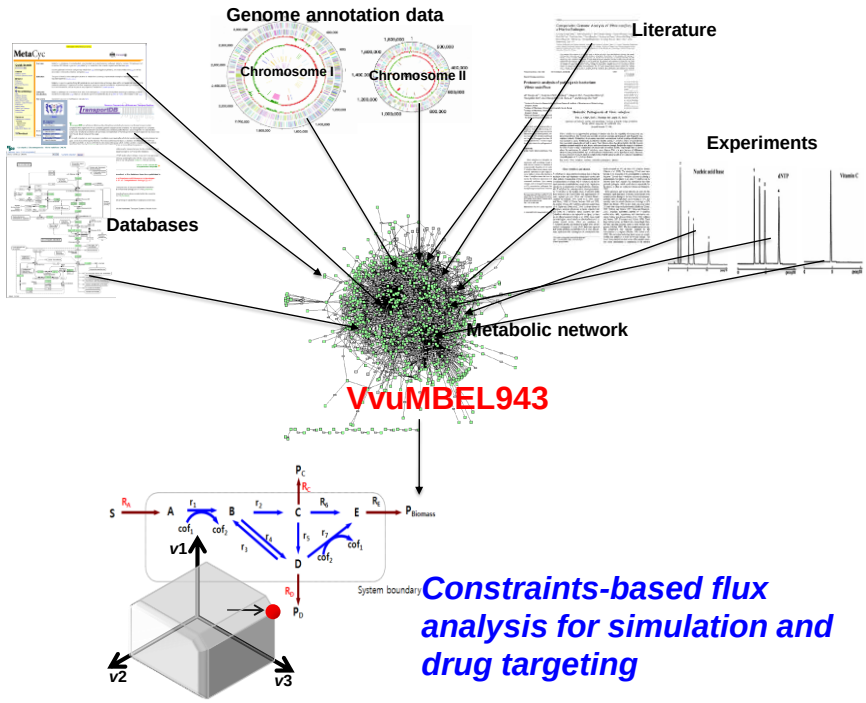
✓ Biotype 1: predominant human pathogens

✓ Typically found in estuarine waters

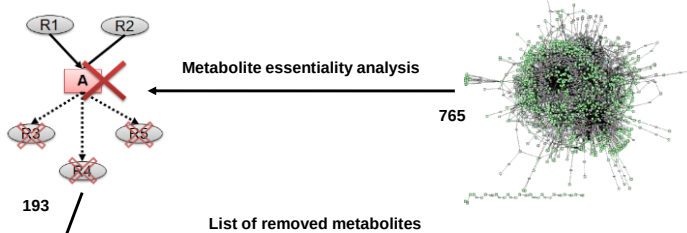
✓ Currently two strains' genomes sequenced: YJ016 (Chen *et al. Genome Res.* **13**, 2577-2587, 2003) and CMCP6 (not reported)

✓ Genome size: 5.12 Mb

✓ Causing agent of septicemia, necrotizing **wound infection**, and gastroenteritis

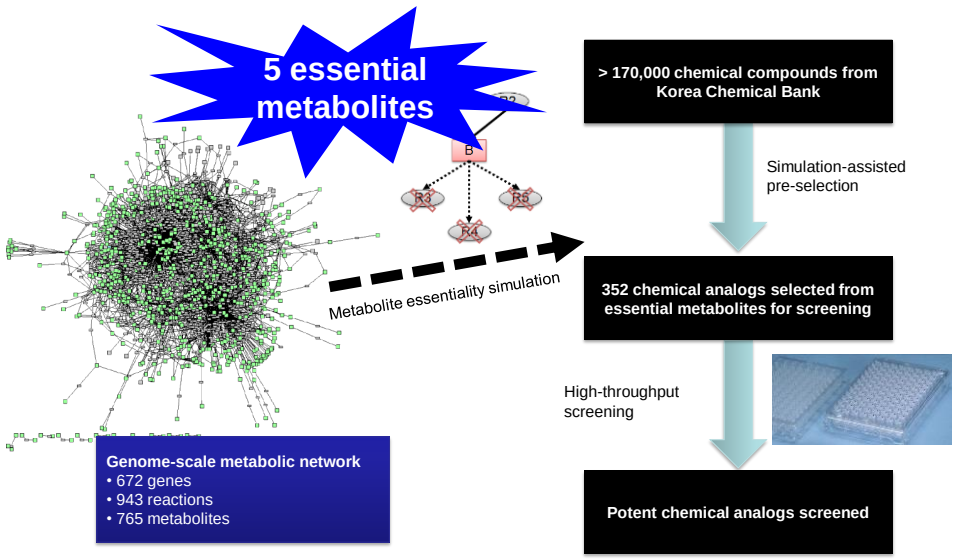


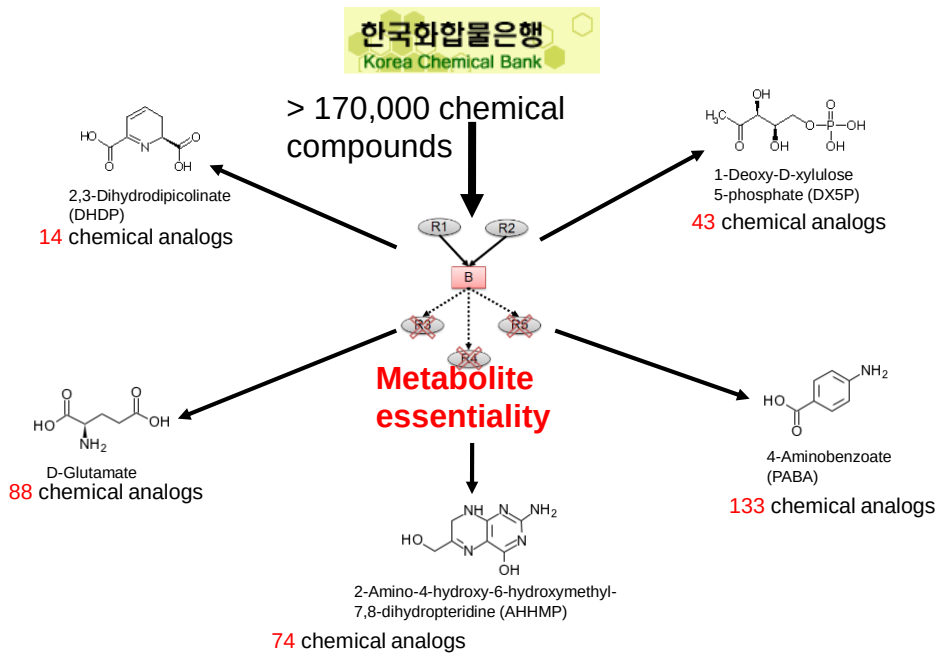
Constraints-based flux analysis for simulation and drug targeting



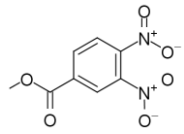
List of removed metabolites

Removing currency metabolites	ADP, AKG, ALA, AMP, ATP, CO2, COA, CTP, FAD, GDP, GLN, GLU, GMP, GTP, METTHF, MTHF, NAD, NADH, NADP, NADPH, NH3, OTHIO, PI, PYR, RTHIO, SAM, THF, UDP, UMP, UTP
163	
Removing metabolites having a single consuming reaction	2AG3PE, 2PCDPMDE, 3A2OP, 3DDAH7P, 3PSME, 4PPNCYS, 4PPNTE, 4PPNTO, A6RP, A6RP5P, A6RP5P2, ADCHOR, ADPDGDMHEP, ADPG, ADPHEP, AGL3P, ALAALA, CAV, CDPMDE, D6RP5P, D8RL, DAPIM, DB4P, DGDMH17BP, DGDMH1P, DGDMH7P, DHP, DHPANT, DHPT, DMPP, DNA, DPCOA, DQT, DTDP, ER4P, FPP, GGPP, GPP, HEPPP, HMB4PP, HPPP, IASP, ICHOR, LPS, MALCOA, MDE4P, MDECPP, NAAD, OHB, OSB, OSBCOA, PA, PANT, PEPTIDO, PGP, PHOSPHOLIPID, PNTO, PPEPTIDO, PPPP, PROTEIN, QA, RIBFLAV, RNA, SAOPIM, SDAPIM, SHCHC, SME, SME3P, TDHDP, UDCP, UDCPP, UDPMNLADGMD, UDPMNLADGMDDADA, UDPNAGEP, UDPNAM, UDPNAMA, UDPNAMAG, UPPMN(GN)LADGMDDADA, UPPMN(GN)LADGNMD(G)5DADA, UPPMN(GN)LADGNMDDADA, UPPMNLADGMDDADA
82	
Removing metabolites present in human metabolism	ACACP, ACCOA, ACP, AHTD, ARG, ASN, ASP, C120ACP, C140ACP, C150ACP, C160ACP, C161ACP, C180ACP, C181ACP, CDPDG, CYS, DALA, DATP, DCTP, DGTP, DHAP, DHF, DTMP, DTTP, E4P, F6P, FMN, FUM, G1P, G3P, G6P, GL3P, GLY, GLYCOGEN, HIS, ILE, IPP, LEU, LYS, MALACP, MET, NACN, OBUT, OIVAL, P5P, PE, PEP, PG, PHE, PPACOA, PRO, PRPP, PS, PYRDX, R5P, RL5P, S7P, SER, SUCCOA, THR, TRP, TYR, UDPG, UDPNAG, VAL, bALA
16	
Removing metabolites associated with enzymes homologous to human proteins	ASPSA, CHOR, DHN, DHSK, DMK, MDAPIM, MK, MKH2, OPP, PHT, PPAACP
5	

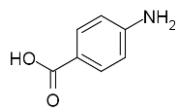

5 essential metabolites
> 170,000 chemical compounds from Korea Chemical Bank
Simulation-assisted pre-selection
Metabolite essentiality simulation
352 chemical analogs selected from essential metabolites for screening
High-throughput screening
Potent chemical analogs screened
Genome-scale metabolic network
• 672 genes
• 943 reactions
• 765 metabolites


한국화합물은행
Korea Chemical Bank
> 170,000 chemical compounds
2,3-Dihydrodipicolinate (DHDP)
14 chemical analogs
1-Deoxy-D-xylulose 5-phosphate (DX5P)
43 chemical analogs
D-Glutamate
88 chemical analogs
Metabolite essentiality
4-Aminobenzoate (PABA)
133 chemical analogs
2-Amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine (AHHMP)
74 chemical analogs

Final ‘hit’, the most potent analog



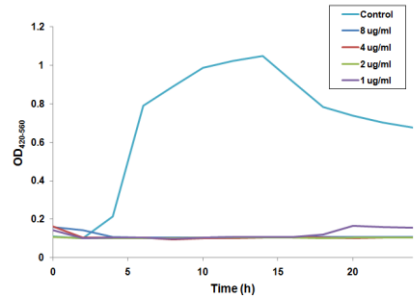
24837
(Chemical analog)



4-Aminobenzoate
(PABA, metabolite)

Compound ID	MIC (μg/ml)	MBC (μg/ml)
24837	**2**	**4**

Growth Inhibition



•MIC: minimal inhibitory concentration
•MBC: minimal bactericidal concentration

Final essential metabolites of *A. baumannii* AYE

Essential metabolites	ORFs	Metabolism	E.C. Number	Enzyme
AHHMP	ABAYE1418 OR ABAYE3176	Folate biosynthesis	2.7.6.3	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase
	ABAYE0807 OR ABAYE3568 OR ABAYE3612 OR ABAYE3616	Folate biosynthesis	2.5.1.15	dihydropteroate syn
DGLU	ABAYE0082 OR ABAYE3395	D-Glutamine and D-glutamate metabolism	5.1.1.3	gl
	ABAYE3524	Peptidoglycan biosynthesis	6.3.2.9	D-glutamate ligase
DHDP	ABAYE0036	Lysine biosynthesis		
DHP	ABAYE1417	Folate biosynthesis		
	ABAYE0811	Folate biosynthesis		
DHSK	ABAYE1539 OR ABAYE1682	Phenylalanine, tyrosine and t		ase
	ABAYE0377	Phenylalanine, t		nase
	ABAYE1685	Phenylal		te dehydrogenase
	ABAYE1683	Ph		roshikimate dehydratase
DX5P	ABAYE1581			1-deoxy-D-xylulose-5-phosphate reductoisomerase
	ABAY			pyridoxine 5-phosphate synthase
DQT			4.2.1.10	3-dehydroquinate dehydratase II; catabolic 3-dehydroquinate dehydratase (3-dehydroquinase)
		biosynthesis	1.1.99.25	quinate/shikimate dehydrogenase
KDO			2.7.7.38	3-deoxy-manno-octulosonate cytidylyltransferase
		synthesis	-	Hypothetical reaction that consumes KDO to genera to lipopolysaccharide
PABA	A A AB	te biosynthesis	2.5.1.15	dihydropteroate synthase

AHHMP, 2-Amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine; **DGLU**, D-Glutamate; **DHDP**, 2,3-Dihydrodipicolinate; **DHP**, 2-Amino-4-hydroxy-6-(D-erythro-1,2,3-trihydroxypropyl)-7,8-dihydropteridine; **DHSK**, 3-Dehydroshikimate; **DX5P**, 1-Deoxy-D-xylulose 5-phosphate; **DQT**, 3-Dehydroquinate; **KDO**, 2-Dehydro-3-deoxy-D-octonate; **PABA**, 4-Aminobenzoate.

Kim et al. *Mol. BioSyst.*, 6, 339-348 (2010)

Data Part 0

Why big-data?



















Understanding Society is difficult, because Everybody Lies…

- US Univ. of Maryland survey
  - Graduate with GPA below 2.5?
    Survey : 2%
  - Did you donate to school last year?
    Survey : 44%

- 40% Engineers said they are in top 5%
- 90% Professors say they work above-average
  (∴Don't trust your professor ☺)

(Social Desirability Bias)

"Everybody lies" by Seth SD (2017)

Everybody Lies…



- Survey in United States
  - Female sex freq = 1/week, cond… r
  - Male sex freq = 1.5/wee…
  - …ggeration of MALE??
  - Follow…ison, only 0.6B condom sales per year!!
    ➔ Both Males & Females are lying… ☺

So we have to find good reliable (big)data!!

"Everybody lies" by Seth SD (2017)

(where can we find good/reliable data?)

Two different kinds of husband?

My husband is
The best
My best friend
Amazing
The greatest

(from Social Media posting)

facebook twitter

"Everybody lies" by Seth SD (2017)

Goo Google

my wife is

my wife is **crazy**
my wife is **cheating on me**
my wife is **lazy**
my wife is **a mermaid**

Google S

my husband is

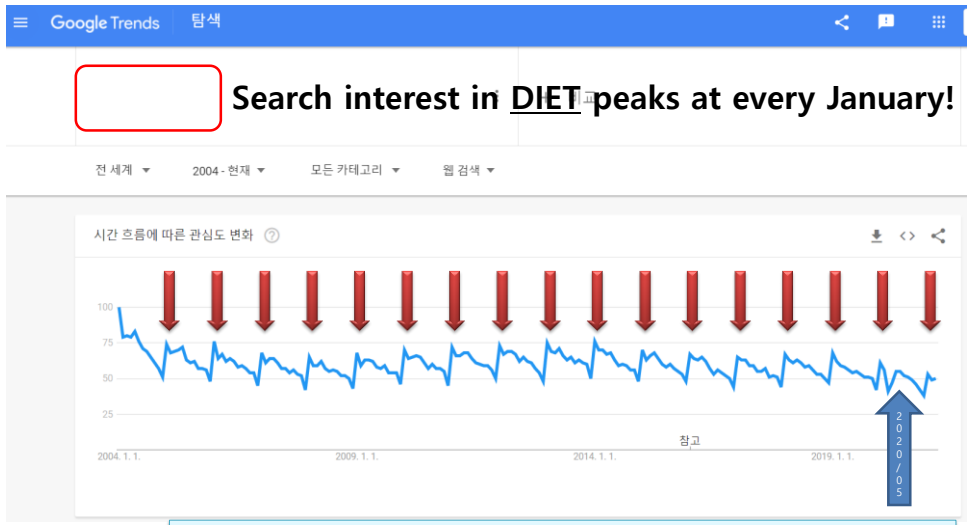
my husband is **gay**
my husband is **selfish**
my husband is **mean**
my husband is **depressed**

gle Search I'm Feeling Lucky

Repo

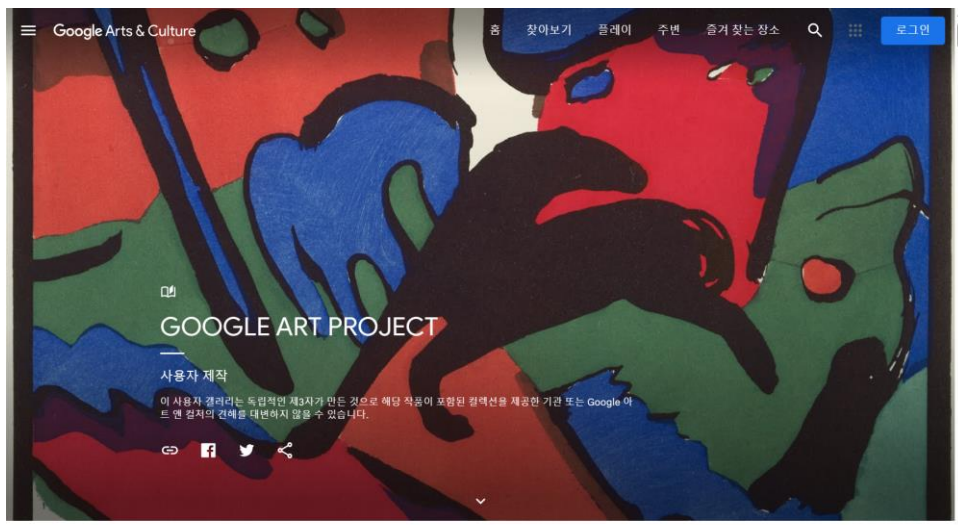


Google has lots of good & reliable data!!



People don't lie when they search… Do you?

© Google



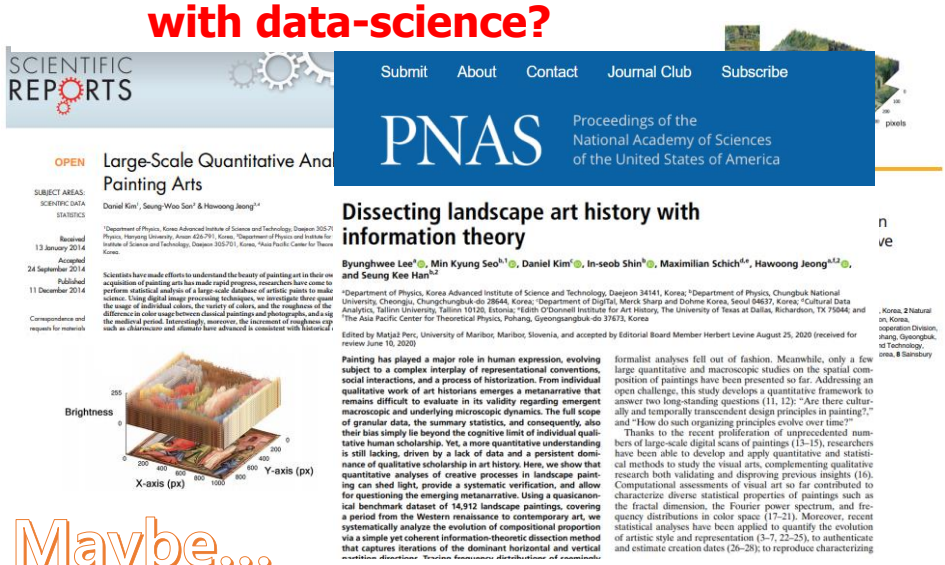
Reginald Mitchell

Google 번역

Data Part 1

Arts and big-data

Can we analyze the arts with data-science?



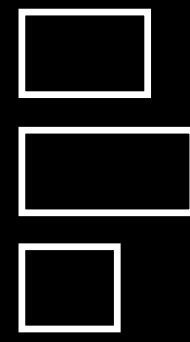
Maybe...

Not perfect, but we are trying... ;)

Simple Question

with long history…

Are there certain proportions people prefer over others?

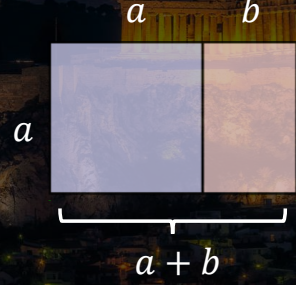


B. Lee, M.K. Seo, D. Kim, I.-s. Shin, M. Schich, **H. Jeong**, and, S. Han, "Dissecting landscape art history with information theory", Proc. Nat. Acad. Sci. USA (2020)

Debates

Aesthetic Preference on the *Golden Ratio*

the Elements, Euclid: 'Extreme and mean ratio'



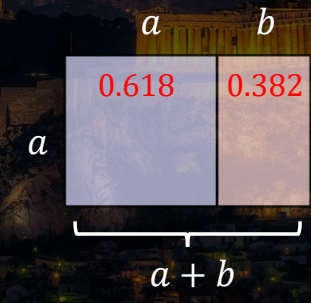
$$\frac{a+b}{a} = \frac{a}{b} \equiv \phi$$

$$1 + \frac{1}{\phi} = \phi$$

$$\phi = \frac{1+\sqrt{5}}{2} = 1.618\,033\ \ldots$$

Debates

Aesthetic Preference on the *Golden Ratio*



The Golden Section is

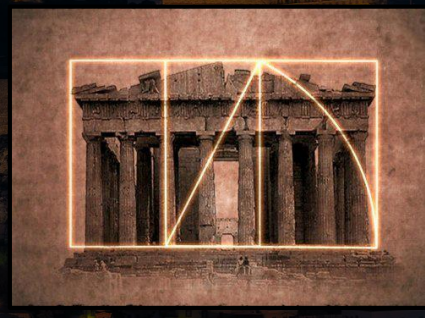
$$\Phi = \frac{a}{a+b} = \frac{b}{a} \approx 0.618$$

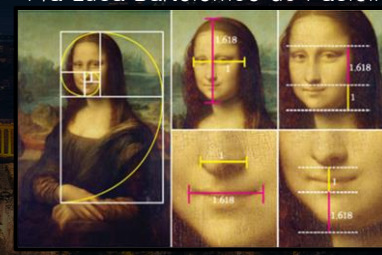
$$\phi = \frac{b}{a+b} = 1 - \Phi \approx 0.382$$

Debates: The divine proportion

the golden rectangle and golden section

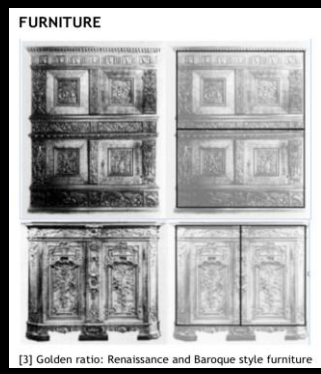
1509 Italian Mathematician.
Fra Luca Bartolomeo de Pacioli



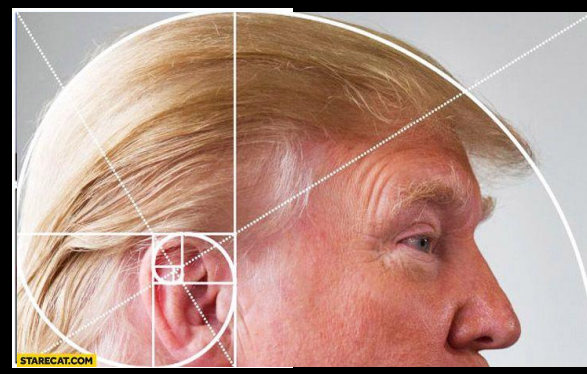




Designed golden-ratio



Jung (2007)

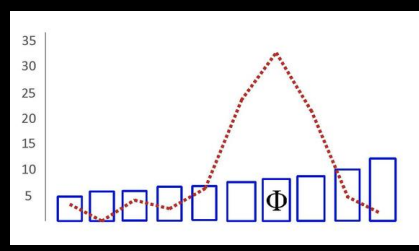


Kimberly Elam(2001)
https://www.youtube.com/watch?v=RqqErDSLtwE

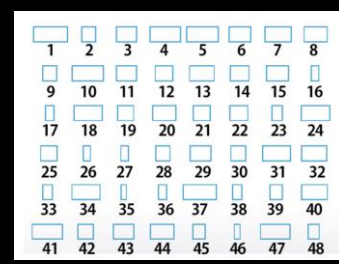
Previous attempts



Gustav Fechner
1801~1887



Fechner's experiment, 347 responses
The Rectangle Is Still Golden, Huffpost



I. C. McManus (1980)

Empirical experiments continued

Fechner (1876, 1871)
Lalo (1908)
Weber (1931)
Thompson (1946)
Nienstedt&Ross (1951)
Eysenck & Tunstall (1968)
Godkewitsch (1974)
Piehl (1978)

c.f.) Little different survey result in Asia

Limitations

1. Usually simple shapes such as rectangles or characters
   - hard to apply to real artworks in the past

2. Selective and subjective measurements in real artworks
   - often not reproducible

3. Small samples with Survey-based experiments

Our goal is

1. A systematic framework for assessing proportion in real artworks

2. Actively chosen ratio by artists not by passive respondents

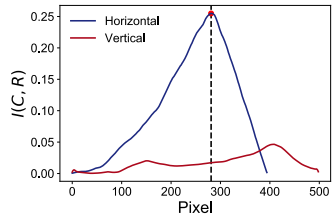
3. Apply to Large-scale real artworks (Especially landscape paintings)

Method – Information theoretic approach

Rigau et al.'s (2008) image partitioning process

$$r_c = \frac{a}{a+b} = \frac{282}{396} = \mathbf{0.712}$$





Color palette information

$$H(C) = -\sum_{c \in C} P(c) \log_2 P(c)$$

Mapping the color set C to set R regions in an image,

The conditional entropy

$$H(C|R) = -\sum_{c \in C, r \in R} P(c,r) \log_2 P(c|r)$$

$$I(C,R) = H(C) - H(C|R)$$

Compositional information

$$I(C,R) = H\left(\sum_{i=1}^{n} \pi_i C_i\right) - \sum_{i=1}^{n} \pi_i H(C_i, r_i)$$

$$H(C_i, r_i) = \sum_{c \in C_i} P(c|r_i) \log_2 P(c|r_i)$$

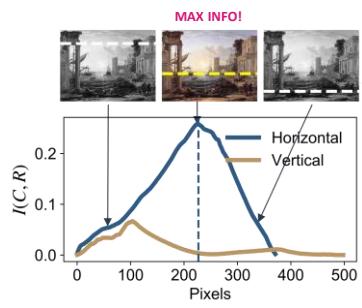
Detecting composition and compositional ratio of a painting

Rigau *et al.*'s image partitioning method

Dissection Ratio

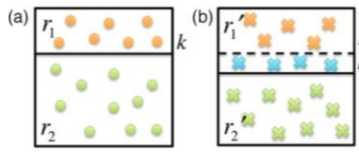
$$r_c = \frac{a}{a+b} = \frac{227}{373} = 0.608$$





Method improvement

Shin et al.'s Line-updating bi-partitioning (LUB) algorithm (2013, New Physics: Sae Mulli)



Few days → ~1 Hour
For typical(~10^6 pixels) paintings

$$
\begin{aligned}
H(C,r_1') &= \frac{k}{k+1}H(C,r_1) - \frac{k}{k+1}\log_2\frac{k}{k+1} \\
&\quad + \Bigg[-\sum_{c\epsilon C_m} \frac{N(c)+n(c)}{L_x(k+1)} \log_2 \frac{N(c)+n(c)}{L_x(k+1)} - \Bigg(-\sum_{c\epsilon C_m} \frac{N(c)}{L_x(k+1)} \log_2 \frac{N(c)}{L_x(k+1)} \Bigg) \\
&\quad - \sum_{c\epsilon C_a} \frac{n(c)}{L_x(k+1)} \log_2 \frac{n(c)}{L_x(k+1)} \Bigg] \\
H(C,r_2') &= \frac{L_y-k}{L_y-(k+1)}H(C,r_2) - \frac{L_y-k}{L_y-(k+1)}\log_2\frac{L_y-k}{L_y-(k+1)} \\
&\quad - \Bigg[-\sum_{c\epsilon C_m} \frac{N(c)-n(c)}{L_x(L_y-(k+1))} \log_2 \frac{N(c)-n(c)}{L_x(L_y-(k+1))} - \Bigg(-\sum_{c\epsilon C_m} \frac{N(c)}{L_x(L_y-(k+1))} \log_2 \frac{N(c)}{L_x(L_y-(k+1))} \Bigg) \\
&\quad - \sum_{c\epsilon C_a} \frac{n(c)}{L_x(L_y-(k+1))} \log_2 \frac{n(c)}{L_x(L_y-(k+1))} \Bigg]
\end{aligned}
\tag{8}
$$



Data: Landscape Paintings



Web Gallery of Art (WGA)

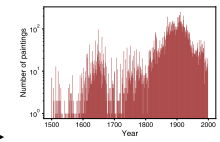


WIKIART (WA)

WEB GALLERY OF ART — 3,610 paintings

WIKIART VISUAL ART ENCYCLOPEDIA — 12,430 paintings

Removal of 949 overlapping paintings →

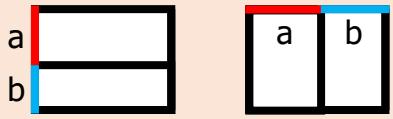


15,092 paintings

(1,483 Artists)

Definition: Partition Direction and Proportion r_c

Direction: Horizontal (H) Vertical (V)



Proportion: $r_c = \frac{a}{a+b}$

V



Proportion = 0.220



0.490

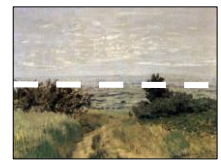


0.777

H



0.418



0.509



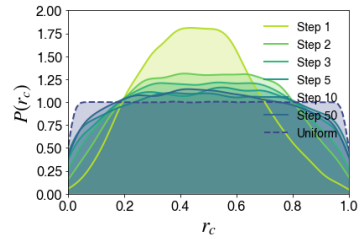
0.644

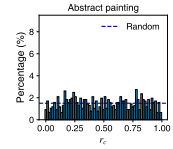
Results

Characteristics of Landscape Paintings

Early partitions are meaningful

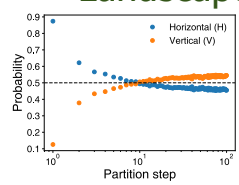




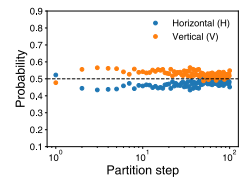




Landscape

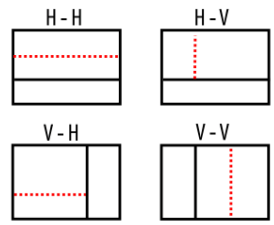


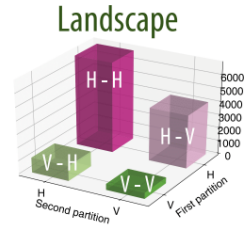
Abstract

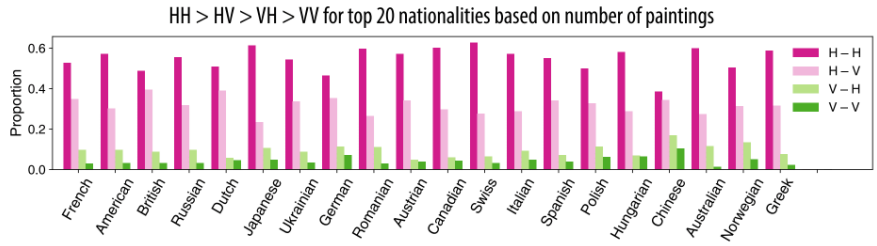


Characteristics of Landscape Paintings

Universally preferred composition in landscape paintings

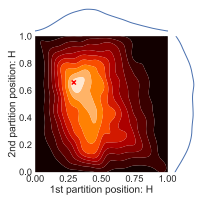




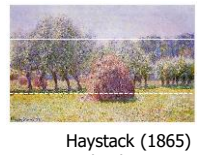


Characteristics of Landscape Paintings

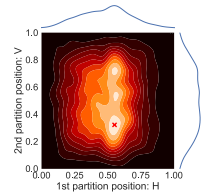
Typical compositional ratio used in landscape painting



H - H



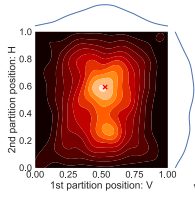
Haystack (1865)
Claude Monet



H - V



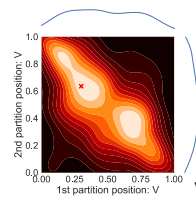
Frogner Manor (1842)
Johan Christian Dahl



V - H



View of Lake Nemi (1790)
Nicolas-Didier Bouguet



V - V

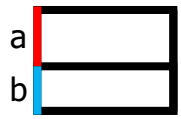


Vesuvius from Portici (1785)
Joseph Wright

Coming back to our main question

Let's focus on first horizontal partition proportion

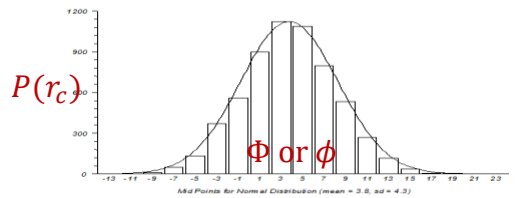
$$r_C = \frac{a}{a+b}$$



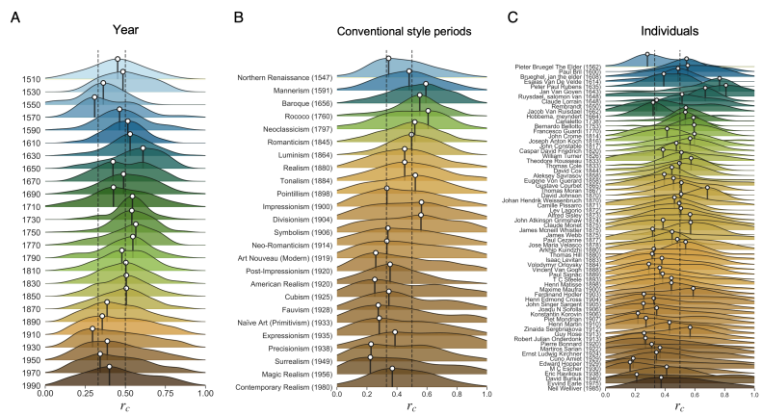
Unvarying distribution over time?

Preference on golden section?

$$r_c = \frac{a}{a+b} = \Phi \text{ or } \phi?$$

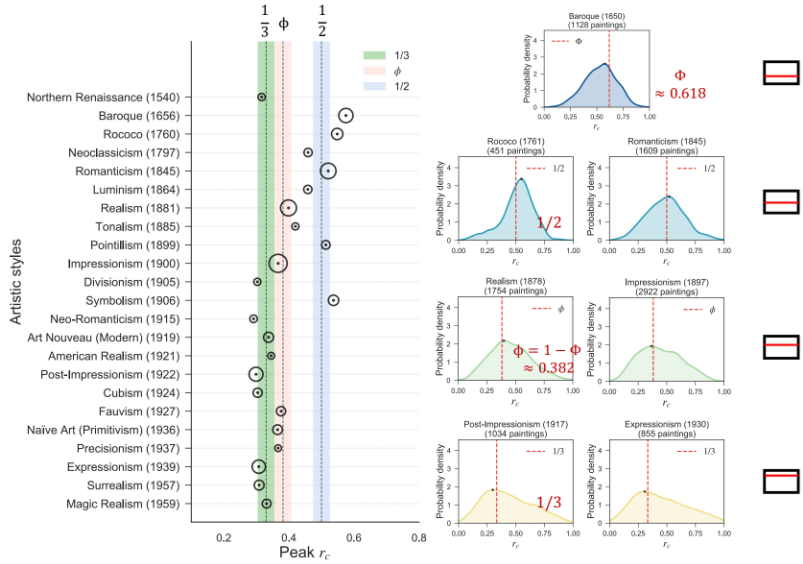


Result: The dominant horizon evolves systematically in landscape art history



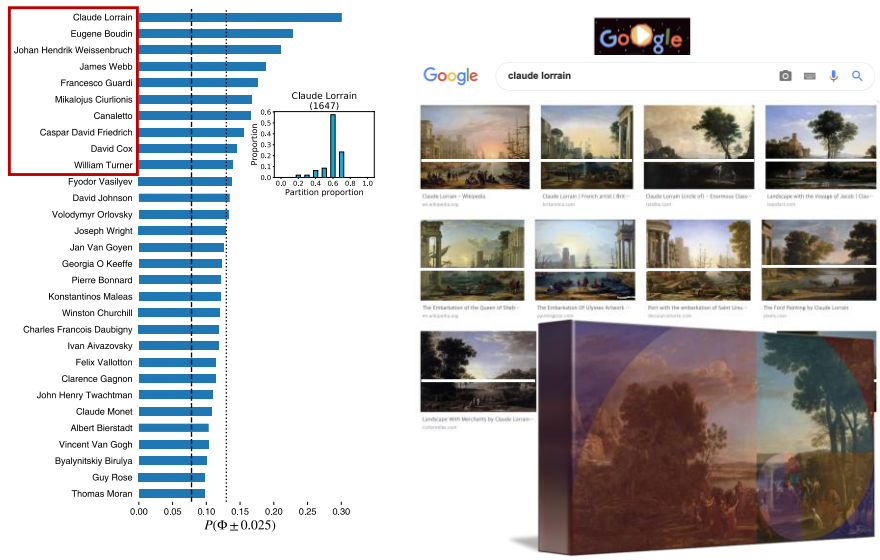
Temporal change in preferred compositional proportion

Proportion preference has been continually changed



Who loved to use the golden section the most?

Claude Lorrain (1600-1682, France) favored Φ (0.618) the most



Short summary on Golden ratio…

1. Developed a systematic framework for assessing compositional proportion
2. Found typical composition in landscape paintings
3. Proportion preference continuously changed over art history moving toward ~1/3
4. The golden ratio was found to be most preferred proportion only in "some" artistic periods and of "some" artists.

B. Lee, M.K. Seo, D. Kim, I.-s. Shin, M. Schich, **H. Jeong**, and, S. Han, "Dissecting landscape art history with information theory", Proc. Nat. Acad. Sci. USA (2020)

Combining Data together with network can do many things!



Rumor and Truth spreading pattern on SNS



(a) Bigfoot (rumor) (b) Summize (non-rumor)

Data: 100 case studies on Twitter from 2006 to 2009

Rumor and Non-Rumor show different spreading pattern ➔ can be predicted ~90%
Conclusion: Not early detection, but after some period, we can distinguish them!

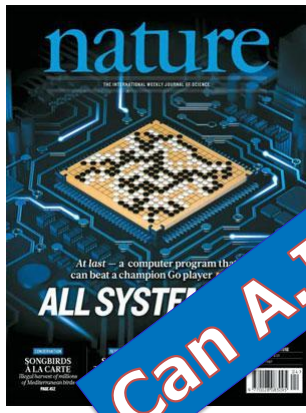
© M.Y. Cha (KAIST) IEEE International Conference on Data Mining (2013)

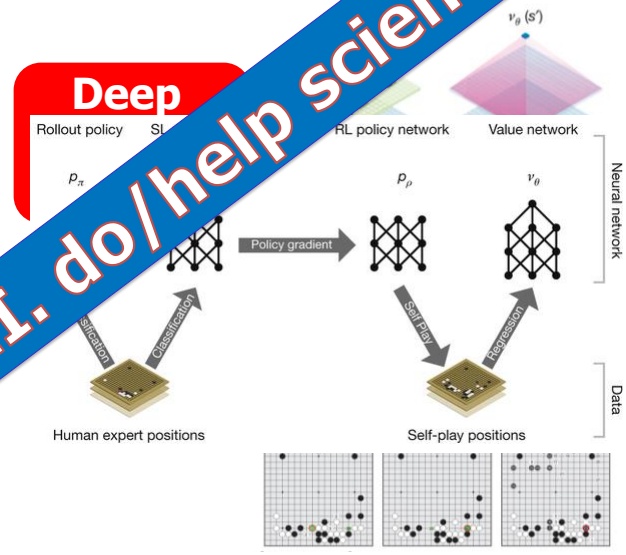
AI part 0

Why AI?
Network + Data = Deep Learning



AlphaGo vs Lee Sedol



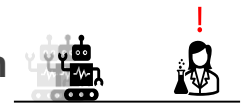


Can A.I. do/help science?

2016/03/08

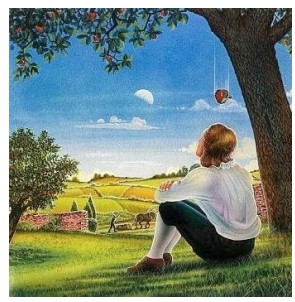
Introduction

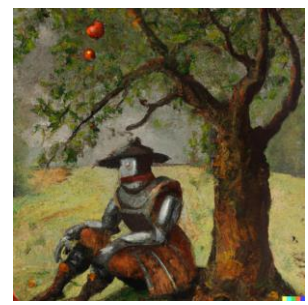
The big question



☐ **A road to machine scientists (automated science)**

- Can machine replace human components from science?
- Can machine discover scientific concepts without human?



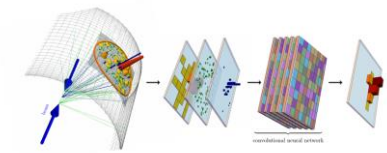


☐ **Can machine (Deep Learning) achieve scientific discovery *automatically?***

Introduction

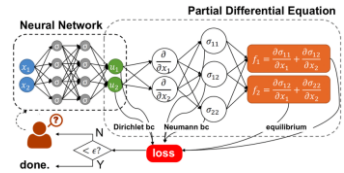
The answer is; (partially) yes!

☐ **Directly finding scientific concepts from large-scale data**



Searching high-energy particles (CNN)

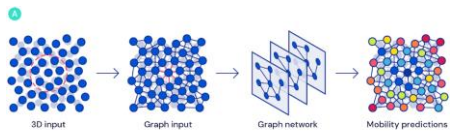
Schwartz, M. D. (2021). Modern machine learning and particle physics. *arXiv preprint arXiv:2103.12226*.



Physics-Informed NN (PINN)

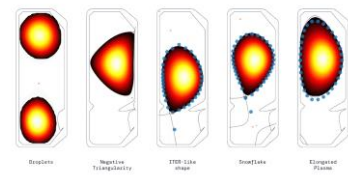
Peng, G. C., Alber, M., Buganza Tepole, A., Cannon, W. R., De, S., Dura-Bernal, S., ... & Kuhl, E. (2021). Multiscale modeling meets machine learning: What can we learn?. *Archives of Computational Methods in Engineering*, *28*(3), 1017-1037.

☐ **Providing scientific insights from the black-box result**



Understanding Glass systems(GNN)

Bapst, V., Keck, T., Grabska-Barwińska, A., Donner, C., Cubuk, E. D., Schoenholz, S. S., ... & Kohli, P. (2020). Unveiling the predictive power of static structure in glassy systems. *Nature Physics*, *16*(4), 448-454.



Plasma control in tokamak (RL)

Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., ... & Riedmiller, M. (2022). Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, *602*(7897), 414-419.

Overview

Central question

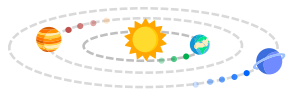
"What scientific concepts in complex systems are *automatically learnable* via deep learning?"

Overview

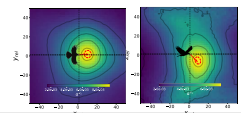
Central question

"What scientific concepts in complex systems are *automatically learnable* via deep learning?"

Conserved quantity hidden in trajectory data



Interaction function in complex systems



AI part 1

Discovering conservation laws

Phys. Rev. Res. **3** L042035 (2021)

Introduction

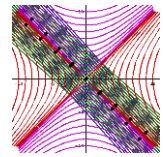
Conservation laws and Symmetries

☐ Conservation laws

- Many systems in nature possess conservation laws.
- Invariants are values that do not change during the system's transition.



Angular momentum conservation



Lorentz invariance

$$\begin{aligned} \vec{A} &\rightarrow \vec{A} - \vec{\nabla} f(\vec{r}, t) \\ \phi &\rightarrow \phi + \frac{1}{c}\frac{\partial f(\vec{r}, t)}{\partial t} \\ f(\vec{r}, t) &= \frac{\hbar c}{e}\lambda(\vec{r}, t). \end{aligned}$$

Gauge Orbit

☐ Invariants ⇔ Symmetry

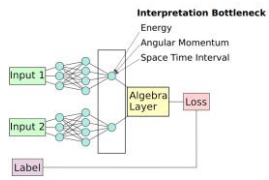
- In a physical system, Noether's theorem states that every conservation law has a corresponding symmetry.
- By investigating invariants in the system, one can get rich information about the structural symmetries and low-dimensional embedding of system dynamics.
- But… now there is too many data!

Method

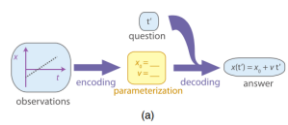
How can we find a conserved quantity?

☐ Attempts to automatically discover conservation laws

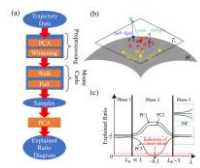
- Limitations on data conditions or limited to find # of conservation law only.



"Discovering Symmetry Invariants Interpreting SNN" (**PRR**, 2020)

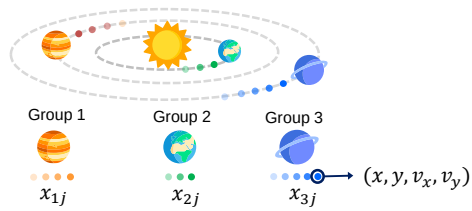


"Discovering Physical Concepts with Neural Networks" (**PRL**, 2020)



"Machine learning conservation laws from trajectories" (**PRL**, 2021)

- How can we find the angular momentum in the real Copernicus case?

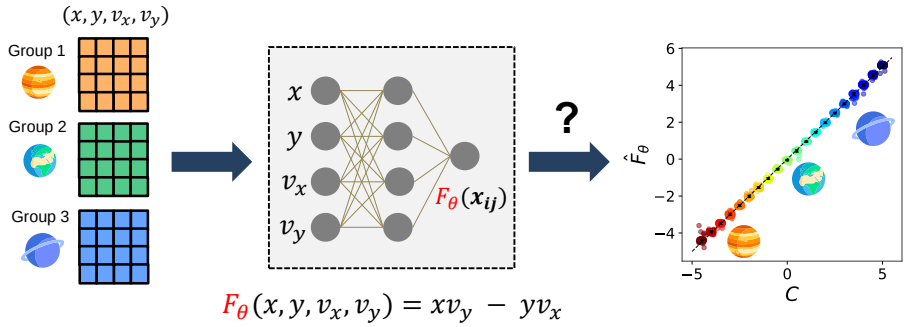


Method

Main goal

☐ **Two criteria for conserved quantity among the group**

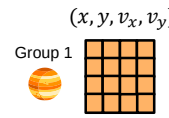
- 1) The model output from the same group should be equal in the ideal case, or at least its deviation $\mathrm{Var}\left(F_\theta(x_{ij})\right)$ should be minimized.
- 2) The model output should produce a ground-truth invariant, or at least a value strongly correlated with the true invariant C.



$$F_\theta(x, y, v_x, v_y) = xv_y - yv_x$$

Method

Intuition for NV loss



☐ **Naïve loss for ConservNet**

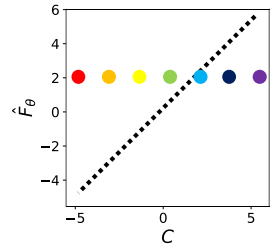
- If the invariant function V exist, then data points in each group satisfies $V(x_{ij}) = C_i$, thus, the following loss L_{naive} will find such function V.

ith group, jth element

$$L_{naive} = \sum_i \mathrm{Var}\left(F_\theta(x_{ij})\right)$$

- But this is very loose condition, since a simple **constant function** will easily satisfy it.

$$F_\theta(x_{ij}) = 2$$



$$\mathrm{Var}(F_\theta(z)) = 0 \text{ for } \forall z$$

Group Variation should decrease only for trajectory data, not for random value!

☐ **Avoid trivial convergence**

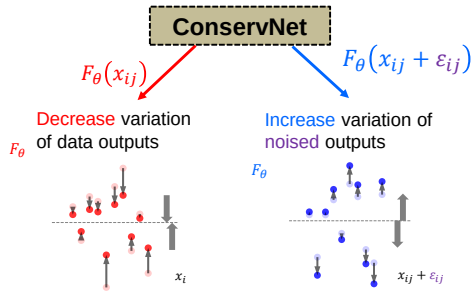
Method

Intuition for NV Loss

☐ Spreader for prohibit trivial convergence

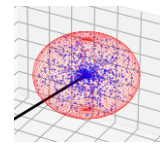
- This is the reason we should **spread** the output from **noised input**,

$F_\theta(x_{ij}+\varepsilon_{ij})$, by fixing its variance to at least Q with loss $\left|Q - \mathrm{Var}\left(F_\theta(x_{ij}+\varepsilon_{ij})\right)\right|$.



$\widehat{\varepsilon_{ij}} \sim U(0, \boldsymbol{I})$, $c \sim U(0, R^{\frac{1}{2}})$

$\varepsilon_{ij} = \frac{c\widehat{\varepsilon_{ij}}}{\|\widehat{\varepsilon_{ij}}\|_2}$, $[\![\varepsilon_{ij}]\!] \leq 1 = \mathbf{R}$



Noise-Variance (NV) Loss

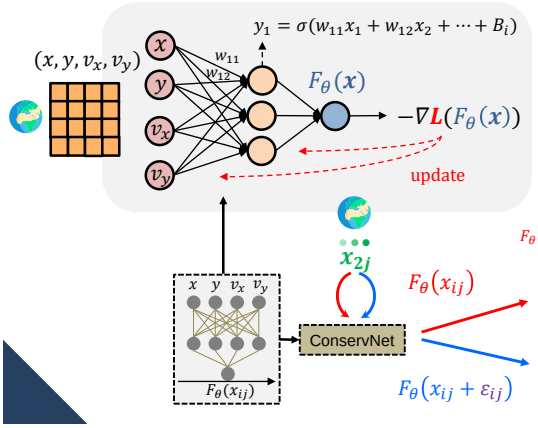
$$\boldsymbol{L} = \sum_i \mathrm{Var}\left(F_\theta(x_{ij})\right) + \left|Q - \mathrm{Var}\left(F_\theta(x_{ij}+\varepsilon_{ij})\right)\right|$$

Method

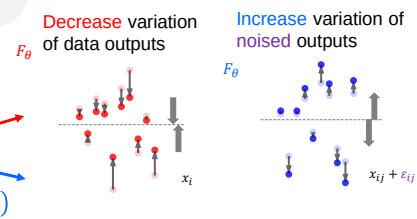
Neural architecture

☐ Simple MLP

- We employ MLP of 3 hidden layers and width of 120 (Adam, lr=1e-5)
- This (relatively) shallow network was enough for all the results.
- At every batch, we feed **a single group** of datapoints with newly sampled noise.



$$\sum_i \mathrm{Var}\left(F_\theta(x_{ij})\right) + \left|Q - \mathrm{Var}\left(F_\theta(x_{ij}+\varepsilon_{ij})\right)\right|$$



Materials

Model systems

	System	Invariant formula
3 Synthetic models	S1	$C = x_1 - 2x_2x_3 + 3x_4^2$
	S2	$C = 3x_1 + 2\sin(x_2) + \sqrt{\lvert x_1 \rvert}x_3^3$
	S3	$C = 2x_1x_2 - (\ln(\lvert x_1 + x_3 \rvert) - x_4)/x_3$
2 Physical models	Lotka–Volterra	$C = \alpha \ln(x) + \delta \ln(y) - \beta x - \gamma y$
	Kepler problem	$C_1 = xv_y - yv_x$ $C_2 = \frac{1}{2}m(v_x^2 + v_y^2) - \frac{GMm}{r}$ $C_3 = \mathbf{p} \times \mathbf{L} - mk\hat{\mathbf{r}}$
1 Real world system	Double pendulum (experiment)	$C_{\text{ideal}} = L_1^2(m_1 + m_2)\omega^2 + m_2L_2^2\omega^2$ $+2m_1m_2L_1L_2\omega_1\omega_2\cos(\theta_1 - \theta_2)$ $-2gL_1(m_1 + m_2)\cos(\theta_1) - 2gm_2L_2\cos(\theta_2)$

- We use total six systems for verify performances of our model.
- ConservNet is trained and tested for 20 batches of 100 points each. (20, 100)
- For double pendulum, the data is collected from motion tracking system.

Materials

Model systems

Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *science*, *324*(5923), 81-85.

	System	
3 Synthetic models	S1	
	S2	
	S3	
2 Physical models	Lotka–Volterra	
	Kepler problem	
1 Real world system	Double pendulum (experiment)	$C_{\text{ideal}} = L_1^2(m_1 + m_2)\omega^2 + m_2L_2^2\omega^2$ $+2m_1m_2L_1L_2\omega_1\omega_2\cos(\theta_1 - \theta_2)$ $-2gL_1(m_1 + m_2)\cos(\theta_1) - 2gm_2L_2\cos(\theta_2)$

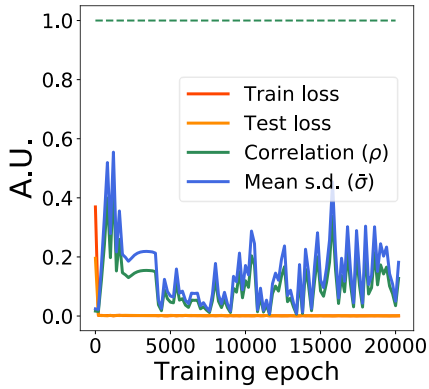


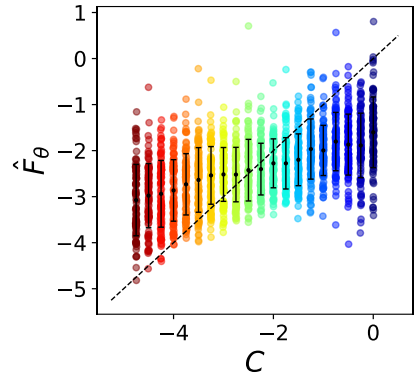
- We use total six systems for verify performances of our model.
- ConservNet is trained and tested for 20 batches of 100 points each. (20, 100)
- For double pendulum, the data is collected from motion tracking system.

Result

Loss without spreader term

$$L_{naive} = \sum_i \mathrm{Var}\left(F_\theta(x_{ij})\right) + \left|Q - \mathrm{Var}\left(F_\theta(x_{ij} + \varepsilon_{ij})\right)\right|$$

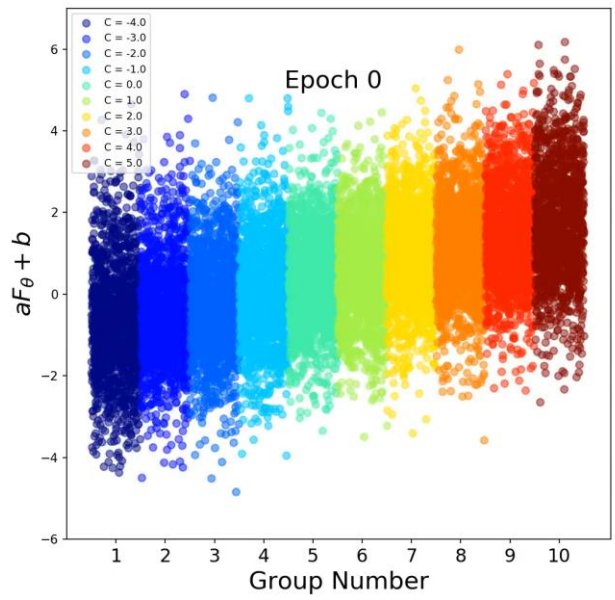




- Without spreader, naïve loss function falls into trivial convergence and shows poor performance.

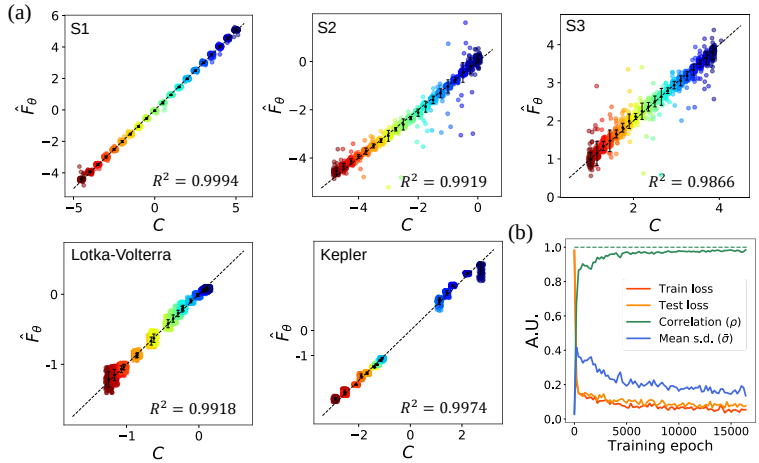
Result

Training results



Result

Correlation between invariant and model output

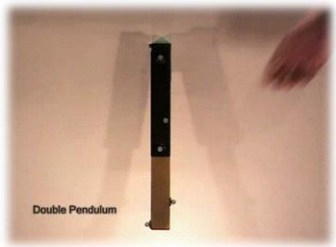


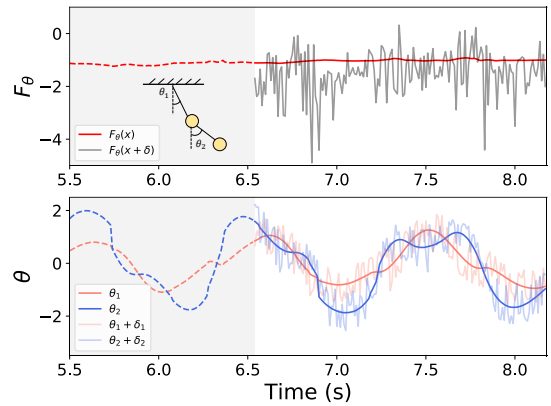
- ConservNet output shows nearly linear relationship with true invariants.
- During its training, correlation increases while standard deviation decreases.

Result

Real double pendulum

$$C_{ideal} = L_1^2(m_1 + m_2)\omega^2 + m_2L_2^2\omega^2 + 2m_1m_2L_1L_2\omega_1\omega_2\cos(\theta_1 - \theta_2) - 2gL_1(m_1 + m_2)\cos(\theta_1) - 2gm_2L_2\cos(\theta_2)$$





- In real DP case, there are only one group of data (with ~700 data points).
- After training, the model output becomes **constant** for true trajectory, but not for *perturbed trajectories.*

Conclusion

Summary

☐ **Automated discovery of scientific concepts**

- We propose a neural network to extract conserved quantity from data without any domain knowledge with high robustness.
- Its output can be further inspected with various techniques such as cross-section investigation and symbolic regression.
- Beyond classical 'invariant', one can capture any non-autonomous system dynamics by standardize its governing equation to $f\left(\frac{dx}{dt}, x, y_1, y_2, \ldots, y_n\right) = 0$.

Letter Open Access

Discovering invariants via machine learning

Seungwoong Ha and Hawoong Jeong
Phys. Rev. Research **3**, L042035 – Published 9 December 2021

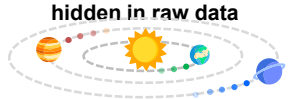
(*Physical Review Research*, *3*(4), L042035.)

Overview

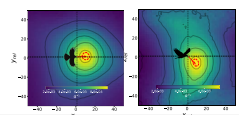
Central question

"What scientific concepts in complex systems are *automatically learnable* via deep learning?"

Conserved quantity hidden in raw data



Interaction function in complex systems



AI part 2

Unraveling hidden interactions

Sci. Rep. **11** 12804 (2021)

Introduction

Main question



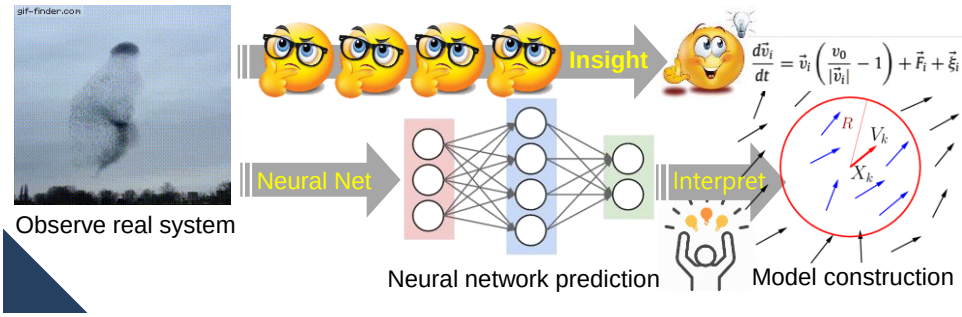
- How can one figure out the '**Individual**' reacts to other by simply observing the data from these complex collective phenomenon…
  - … with **minimum prior assumption** to the system?

Introduction

Main question

☐ **Data-driven modeling**

- If we cannot pose strict restriction to model, model itself should be *flexible* enough to represent all the possible functional form.
- Neural network excels at approximate (almost) arbitrary functions.
- Let the Neural Network learns interaction from data!



Method

Main contribution

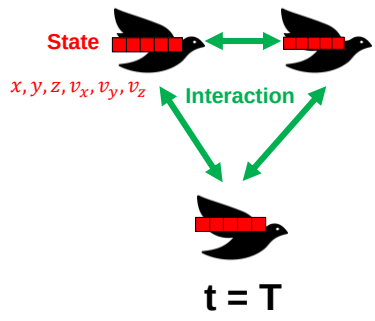
☐ We propose '**AgentNet**', unified framework of neural network to disentangle the decision rule of single agent from complex system.

☐ AgentNet approximates the transition function of single agent with objective of predicting agent's state in future.

☐ AgentNet showed great performance in terms of predicting state transition and identifying **variable-wise** interaction range/strength.

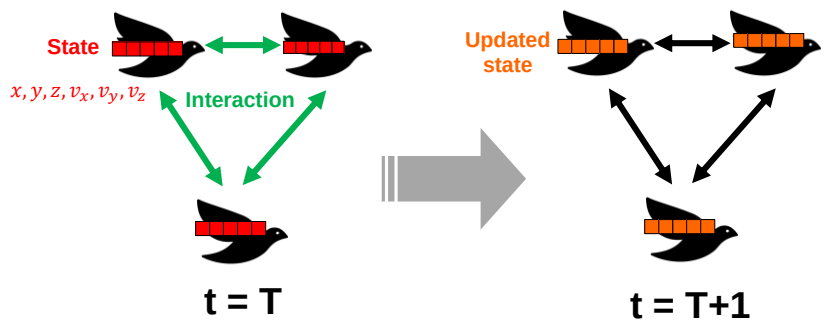
Method

Neural interaction network



Method

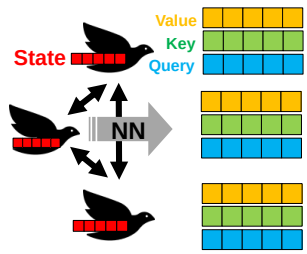
Neural interaction network



Method

Attention mechanism

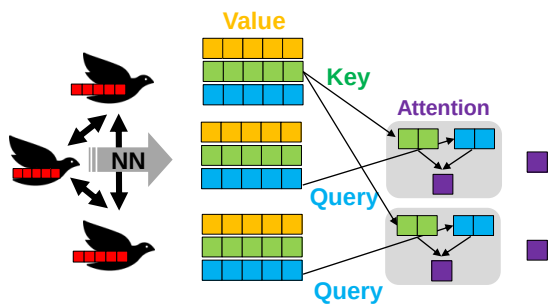
Value : My state **Key : Receiver** **Query : Sender**



Method

Attention mechanism

Attention for you = *f* (My receiver , Your sender)

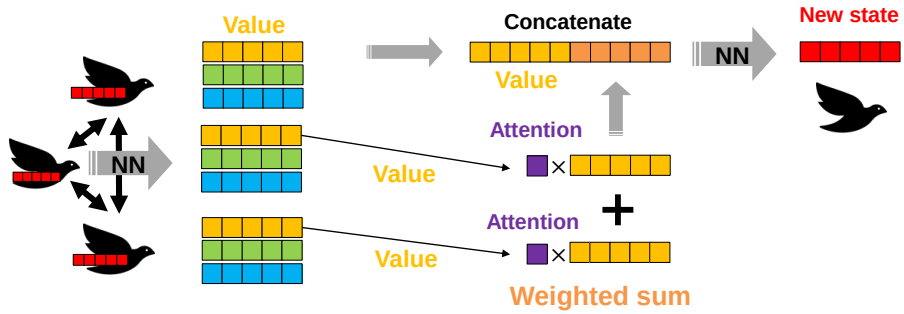


Value : My state **Key : Receiver** **Query : Sender**

Method

Attention mechanism

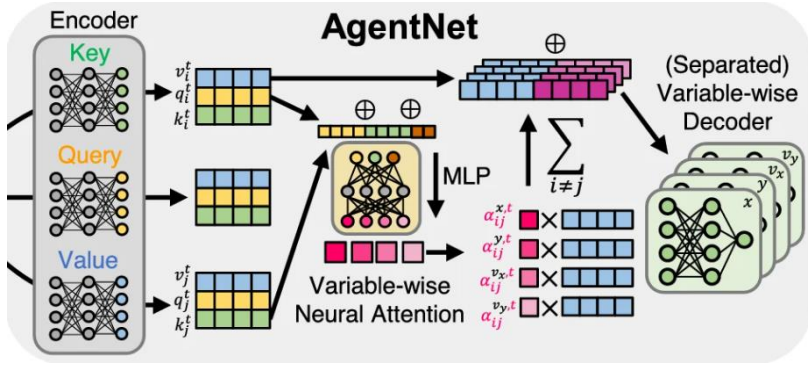
New state = f (My state, Σ(Attention × Your state))



Value : My state Key : Receiver Query : Sender

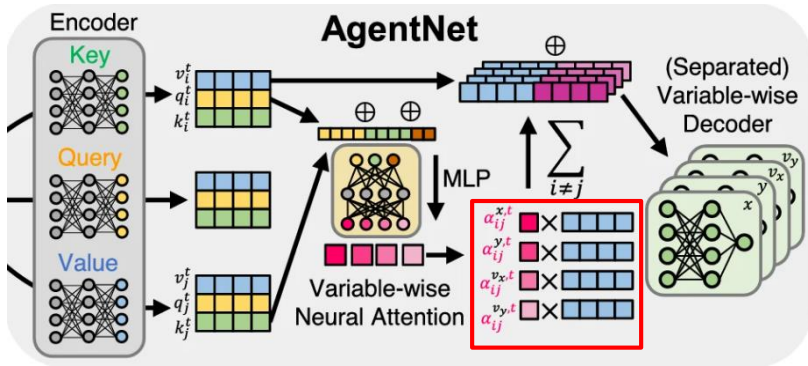
Method

Variable attention mechanism



Method

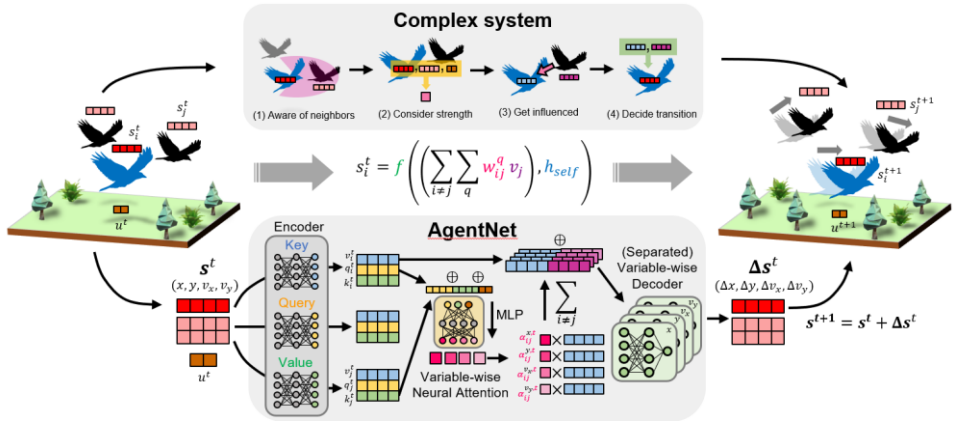
Variable attention mechanism



☐ **Separate attention for each state variable!**

Method

Problem formulation

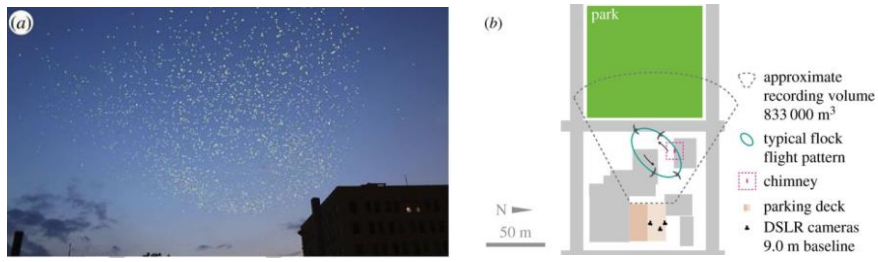


Result

Chimney swift flock

☐ **Bird flock data**

- 3d trajectories of ~1600 chimney swifts.
- Generally, turns around the chimney.
- This flock data can be modeled as dynamic graph.

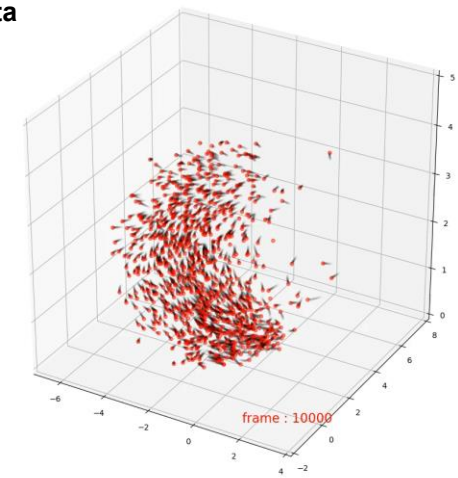


Evangelista, Dennis J., et al. "Three-dimensional trajectories and network analyses of group behaviour within chimney swift flocks during approaches to the roost." *Proceedings of the Royal Society B: Biological Sciences* 284.1849 (2017): 20162602.

Result

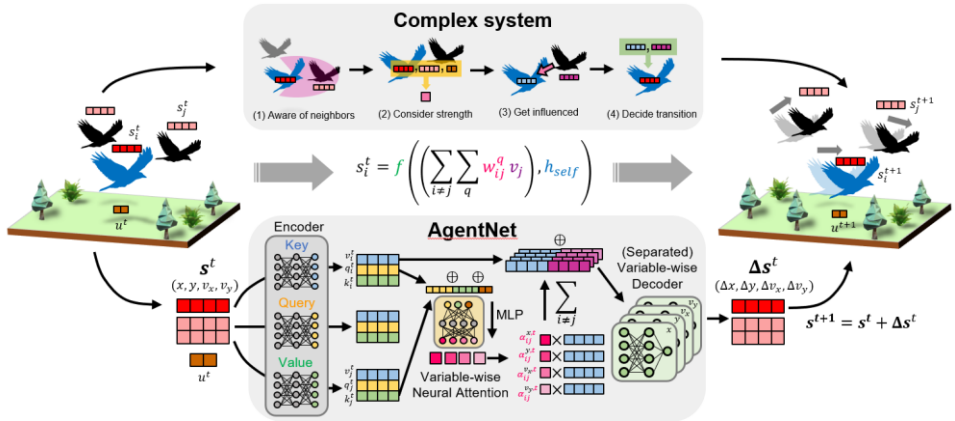
Chimney swift flock

☐ **Bird flock data**



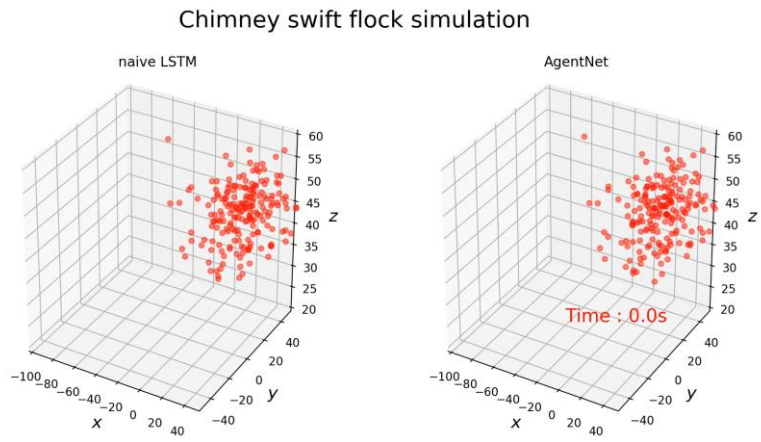
Result

Chimney swift flock



Result

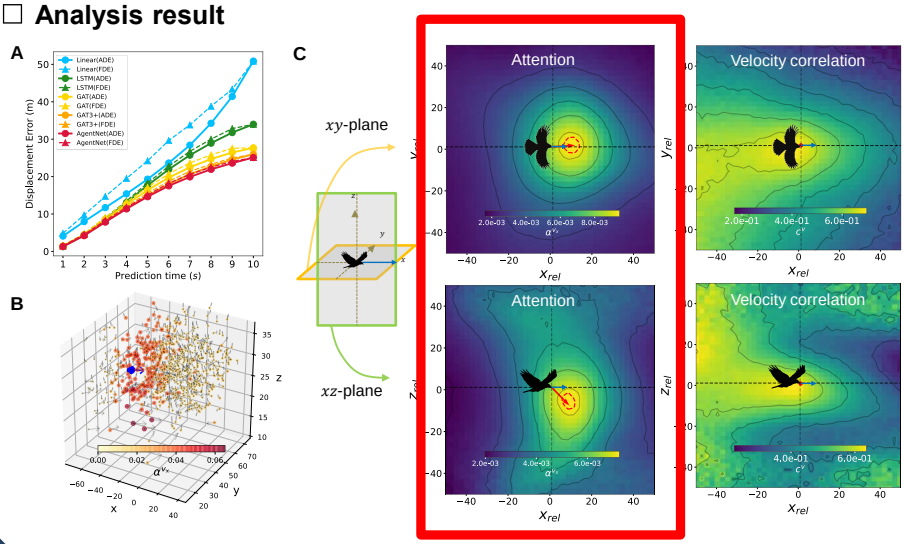
Bird Flock data learning and simulation



Result

#4. Chimney swift flock

☐ **Analysis result**



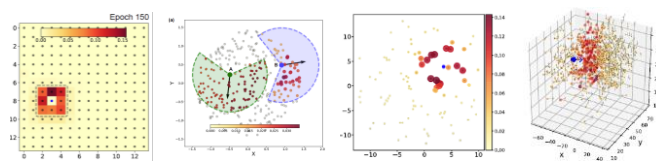
☐ **Interaction range is different from correlation!**

Discussion

Summary

☐ **Neural network to aid system modeling**

- We have demonstrated the capability of AgentNet, unified framework of neural network based on Interaction Network with attention mechanism that can be applied to virtually any agent-based system.
- This Data-driven approaches helps analytic modeling.
- See the unknown interaction via variable-wise attention visualization.



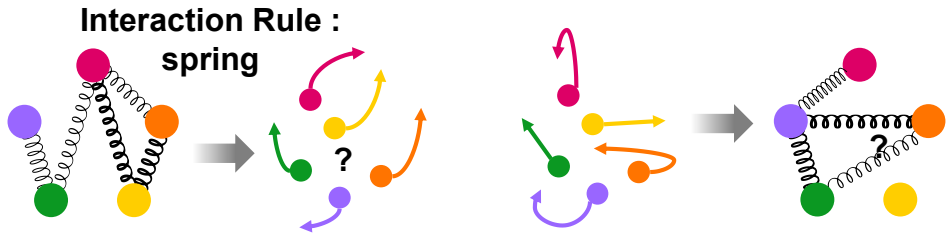
Article | Open Access | Published: 17 June 2021

Unraveling hidden interactions in complex systems with deep learning

Seungwoong Ha & Hawoong Jeong

(Scientific reports, *11*(1), 1-13)

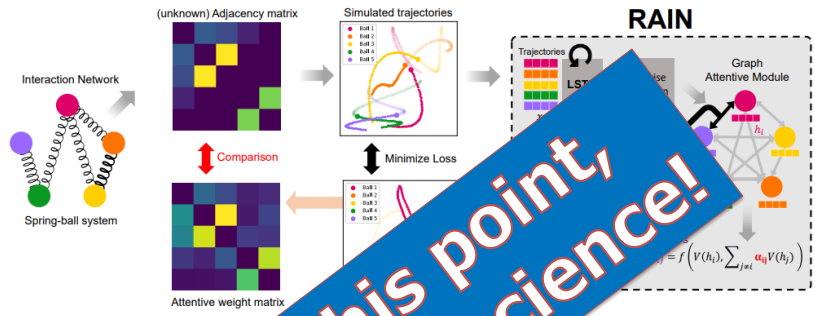
Recent works



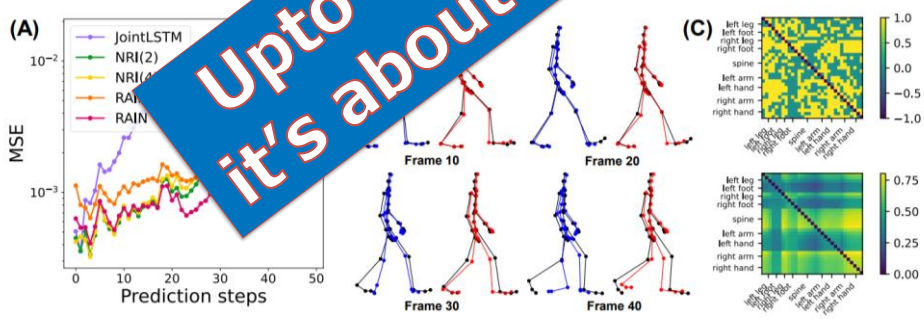
General rule depends on hidden connection

Relational Attentive Inference Network (RAIN) to infer continuously weighted interaction graphs without any ground truth interaction strengths + Pairwise Attention (PA)

S. Ha, H. Jeong "Learning Heterogeneous Interaction Strengths by Trajectory Prediction with Graph Neural Network", ICLR (2023)



Motion Capture Data: M (NRI) vs RAIN + PA



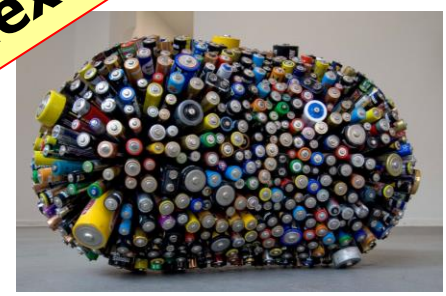
Upto this point, it's about science!

Don't forget the power of network & data!
But connect them wisely with





≠



THINK DATA & NETWORKS together with A.I. for Complex Systems

selected References [stat.kaist.ac.kr]

- G. Kim, D.-K. Kim, and **H. Jeong**, *Nature Communications* (2024)
- S. Ha, and **H. Jeong**, *ICML (Machine Learning)* (2023)
- S. Ha, and **H. Jeong**, *ICLR (Learning Representations)* (2023)
- S. Ha, and **H. Jeong**, *Phys. Rev. Res.* (2021) *; Sci. Rep.* (2021)
- D.-K. Kim, Y. Bae, S. Lee, and **H. Jeong**, *Phys. Rev. Lett.* (2020)
- B. Lee, M.K. Seo, D. Kim, I.-S. Shin, M. Schich, **H. Jeong**, and, S. Han, *PNAS* (2020)
- J. Yun, S.-H. Lee, and **H. Jeong**, *Nature Human Behaviours* (2018)
- D. Kim, D. B. Cerigo, H. Youn, and **H. Jeong**, *Euro. Phys. J. Data Science* (2016)
- Y. Baek, D. Kim, M. Ha, **H. Jeong**, *Phys. Rev. Lett.* (2012)
- S.-W. Son, B.-J. Kim, H. Hong, **H. Jeong** *Phys. Rev. Lett*. (2009)
- J. Um, S.-W. Son, S.-I. Lee, **H. Jeong**, B.J. Kim, *PNAS* (2009)
- H.-J. Youn, M. Gastner, **H. Jeong**, *Phys. Rev. Lett.* (2008)
- P.-J. Kim, D.-Y. Lee, T.Y. Kim, K.H. Lee, **H. Jeong**, S.Y. Lee, S. Park, *PNAS* (2007).
- P. Holme, **H. Jeong**, *Bioinformatics* (2003).
- S.H. Yook, **H. Jeong**, A.-L. Barabasi, *PNAS* (2002).
- **H. Jeong**, S.P. Mason, A.-L. Barabasi, Z.N. Oltvai, *Nature* **411** 41 (2001).
- **H. Jeong**, B. Tombor, R. Albert, Z.N.Oltvai, A.-L. Barabasi, *Nature* **407** 651 (2000).
- R. Albert, **H. Jeong**, A.-L. Barabasi, *Nature* **406** 378 (2000).
- R. Albert, **H. Jeong**, A.-L. Barabasi, *Nature* **401** 130 (1999).

A.I.

Data

Network