KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML)
Workshop for Life and Medical Scientists

생명정보학 & 머신러닝 워크샵 (온라인)

Bayesian interpretation in the context of large biological data collections

이영석 _ KAIST

KSBI
KOREAN SOCIETY FOR BIOINFORMATICS
한국생명정보학회

⚠

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2024 워크샵 온라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML)
Workshop for Life and Medical Scientists

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2024에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 벌써 10년 차를 맞이하게 되었습니다. BIML 워크샵은 국내 생명정보학 분야의 최초이자 최고 수준의 교육프로그램으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되어 있습니다. 올해 인공지능 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 인공지능 기반 자료모델링 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 단일세포오믹스, 공간오믹스, 메타오믹스, 그리고 롱리드염기서열 자료 분석에 대한 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다.

올해 BIML의 가장 큰 변화는 최근 연구 수요가 급증하고 있는 의료정보자료 분석에 대한 현장 강의를 추가하였다는 것입니다. 특히 의료정보자료 분석을 많이 수행하시는 의과학자 및 의료정보 연구자들께서 본 강좌를 통해 많은 도움을 받으실 수 있기를 기대하고 있습니다. 또한 다양한 생명정보학 분야에 대한 온라인 강좌 프로그램도 점차 증가하고 있는 생명정보 분석기술의 다양화에 발맞추기 위해 작년과 비교해 5강좌 이상을 신규로 추가했습니다. 올해는 무료 강좌 5개를 포함하여 35개 이상의 온라인 강좌가 개설되어 제공되며, 연구 주제에 따른 연관된 강좌 추천 및 강연료 할인 프로그램도 제공되며, 온라인을 통한 Q&A 세션도 마련될 예정입니다. BIML-2024는 국내 주요 연구중심 대학의 전임 교원이자 각 분야 최고 전문가들의 강의로 구성되었기에 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것이라 확신합니다.

BIML-2024을 준비하기까지 너무나 많은 수고를 해주신 운영위원회의 정성원, 우현구, 백대현, 김태민, 김준일, 김상우, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 강사분들께 깊은 감사를 드립니다.

2024년 2월

한국생명정보학회장 이인석

강의개요

Bayesian interpretation in the context of large biological data collections

The advance of biotechnology has enabled the democratization of massive bio-data generation at the level of individual laboratories, thus providing a multi-faceted view of the complexity of living systems. Yet, much of this data is left under-utilized or sometimes even misinterpreted owing to the lack of appropriate computational tools and bioinformatic algorithms. In this course, we will cover the theorical basis of one computational technique called the Bayesian methodology and its success in interpreting large biological data collections. We will start by introducing the difference between Frequentist and Bayesian, and then build up to probabilistic graphical models and specialized bioinformatic algorithms for reconstructing biological networks from public data, quantifying gene expression by expectation maximization, and more. It is not required but recommended to read the following materials before this class:

1. Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics, 12(1), 1-16.
2. Lee YS, Krishnan A, Zhu Q and Troyanskaya OG (2013) "Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies." Bioinformatics 29 (23), 3036-3044

* 강의 난이도: 초급

* 강의: 이영석 교수 (한국과학기술원 바이오 및 뇌공학과)

Curriculum Vitae

Speaker Name: Young-suk Lee, Ph.D.

▶ Personal Info

Name	Young-suk Lee
Title	Assistant Professor
Affiliation	Korea Advanced Institute of Science and Technology

▶ Contact Information

Address	E16 Rm#1113, 291 Daehak-ro, Yuseong-gu, Daejeon 34141
Email	youngl@kaist.ac.kr
Phone Number	042-350-7924

Research Interest

Bioinformatics, Functional genomics, Computational molecular biology

Educational Experience

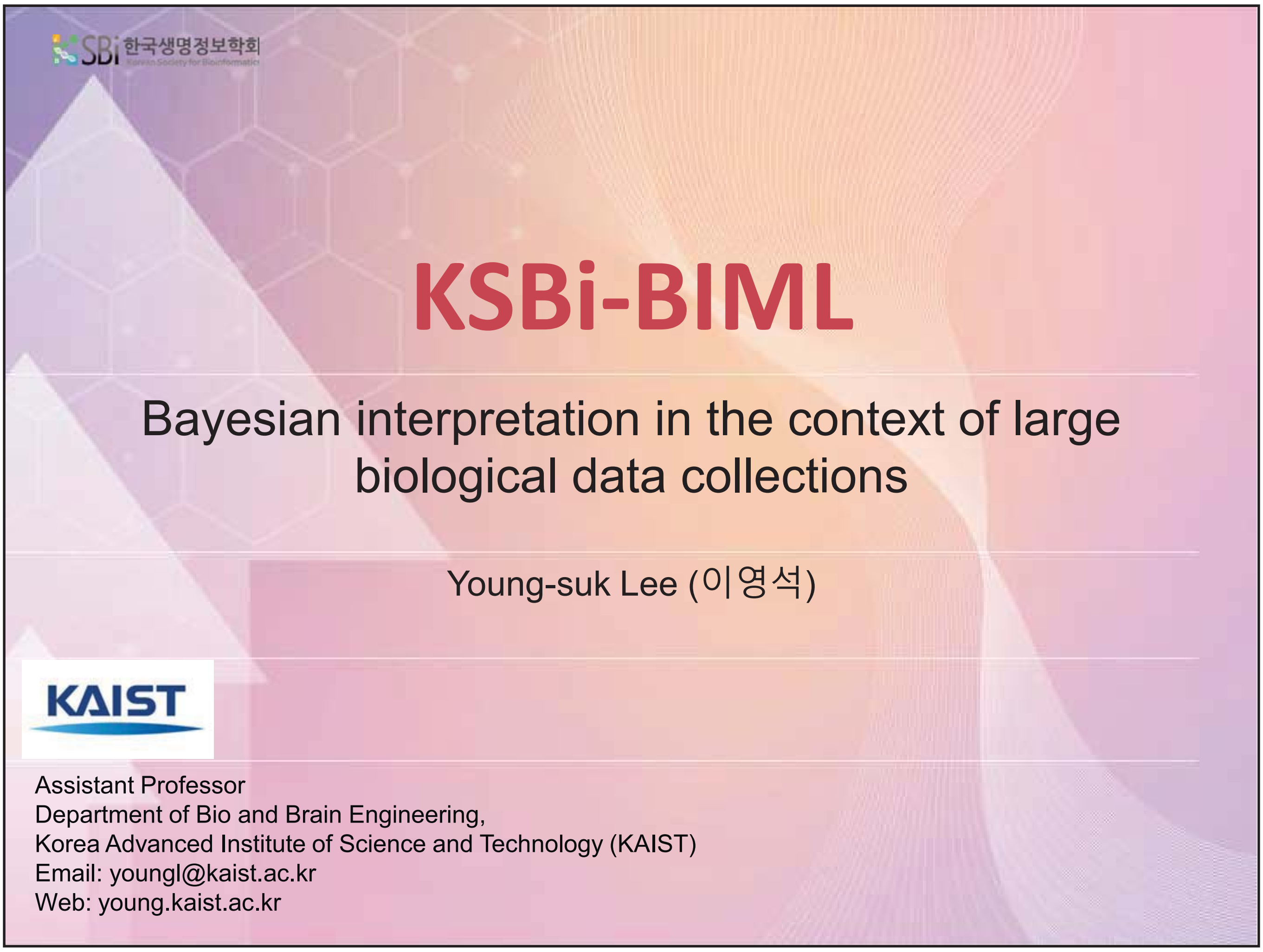
2010	B.S. Computer Science and B.S. Mathematics, The University of Texas at Austin
2014	M.S. Computer Science, Princeton University
2016	Ph.D. Computer Science, Princeton University

Professional Experience

2016-2020	Research fellow, Seoul National University and Institute for Basic Science
2020-	Assistant Professor, Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology

Selected Publications (5 maximum)

1. Lee S*, **Lee YS***, Choi Y, Son A, Park Y, Lee KM, Kim J, Kim JS, Kim VN (2021) "The SARS-CoV-2 RNA interactome." *Molecular Cell* *equal contributions
2. Kim D*, **Lee YS***, Jung SJ*, Yeo J*, Seo JJ, Lee YY, Lim J, Chang H, Song J, Yang J, Jung G, Ahn K and Kim VN (2020) "Viral hijacking of the TENT4-ZCCHC14 complex protects viral RNAs via mixed tailing." *Nature structural & molecular biology* *equal contributions
3. **Lee YS**, Krishnan A, Oughtred R, Rust R, Chang CS, Ryu J, Kristensen VN, Dolinski K, Theesfeld CL and Troyanskaya OG (2019) "A Computational Framework for Genome-wide Characterization of the Human Disease Landscape." *Cell systems* 8 (2), 152-162. e6
4. **Lee YS**, Wong AK, Tadych A, Hartmann BM, Park CY, DeJesus VA, Ramos I, Zaslavsky E, Sealfon SC and Troyanskaya OG (2018) "Interpretation of an individual functional genomics experiment guided by massive public data." *Nature methods* 15 (12), 1049
5. Lim J*, Kim D*, **Lee YS***, Ha M, Lee M, Yeo J, Chang H, Song J, Ahn K and Kim VN (2018) "Mixed tailing by TENT4A and TENT4B shields mRNA from rapid deadenylation." *Science* 361 (6403), 701-704, *equal contributions

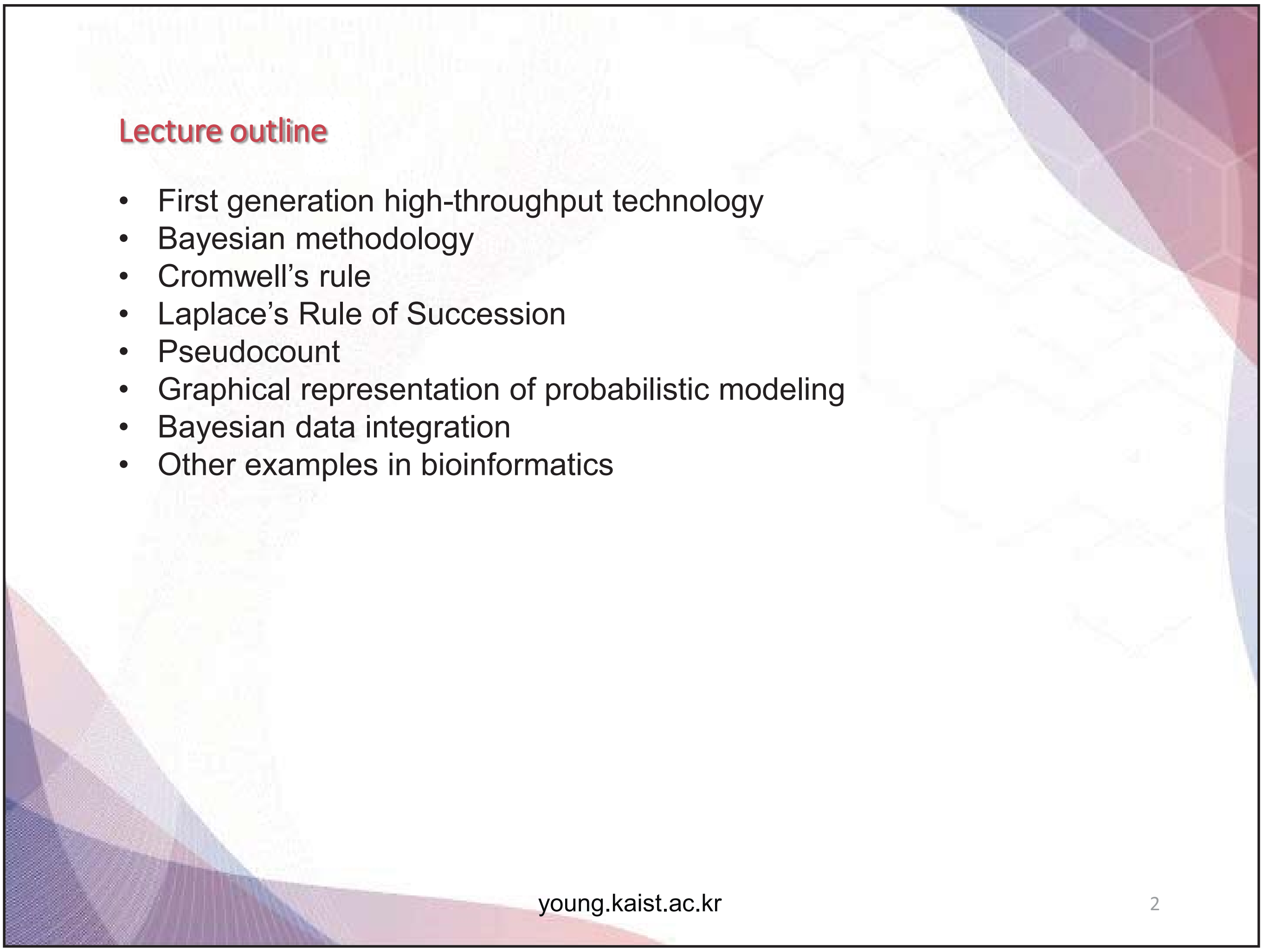


KSBi-BIML

Bayesian interpretation in the context of large biological data collections

Young-suk Lee (이영석)

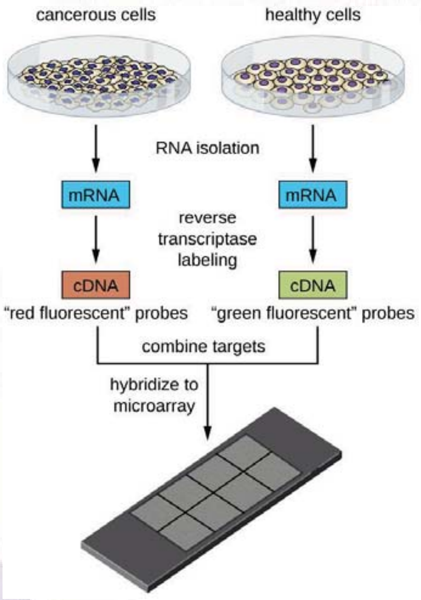
Assistant Professor
Department of Bio and Brain Engineering,
Korea Advanced Institute of Science and Technology (KAIST)
Email: youngl@kaist.ac.kr
Web: young.kaist.ac.kr



Lecture outline

- First generation high-throughput technology
- Bayesian methodology
- Cromwell's rule
- Laplace's Rule of Succession
- Pseudocount
- Graphical representation of probabilistic modeling
- Bayesian data integration
- Other examples in bioinformatics

First-generation high-throughput biotechnology


cancerous cells

healthy cells

RNA isolation

mRNA

mRNA

reverse transcriptase labeling

cDNA

cDNA

“red fluorescent” probes

“green fluorescent” probes

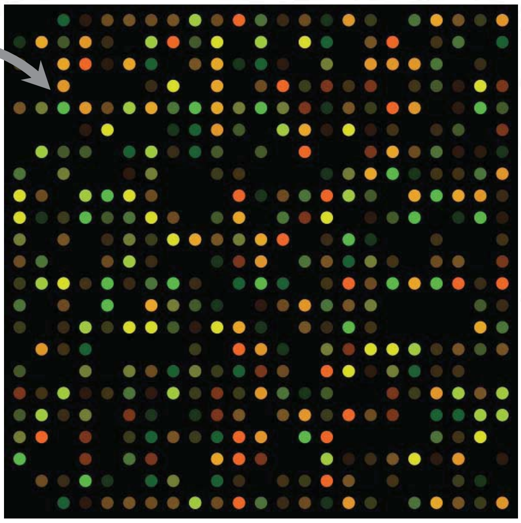
combine targets

hybridize to microarray

By OpenStax Microbiology under CC BY 4.0 (left);
Schutz under CC BY-SA 3.0 (right)

Simultaneous measurement of thousands of genes

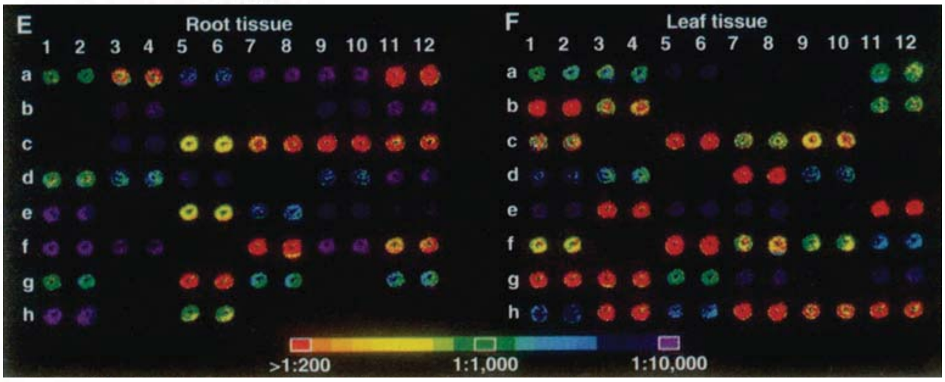
Activity of specific gene



Dual channel (green + red) version

By Guillaume Paumier under CC BY-SA 3.0

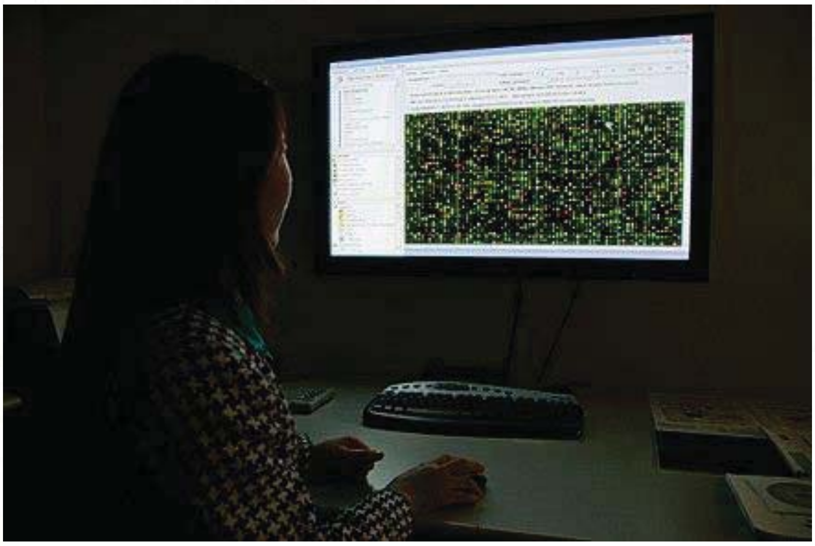
Growth of “gene chip” industry and massive bio-data generation



One of the first gene chips

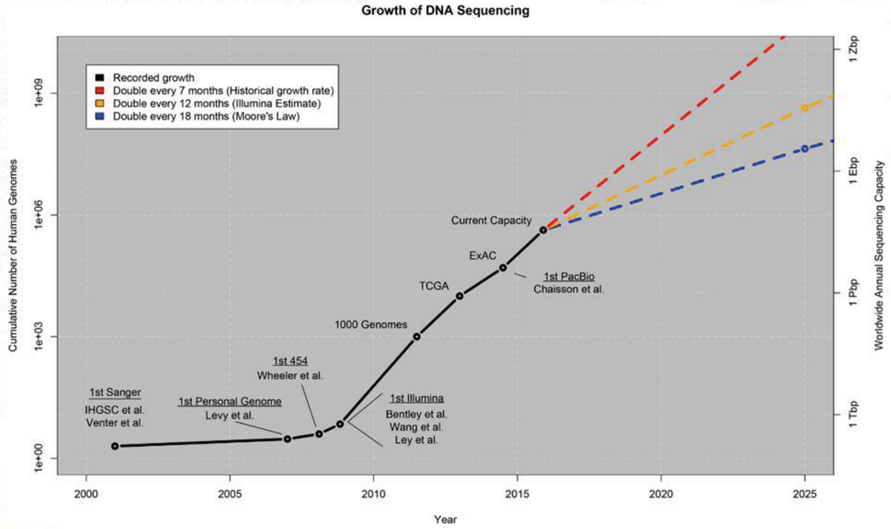
Mark Schena et al (1995); Cited by 13,441

How to manage, handle, and ultimately interpret this bio big data?



National Center for Toxicological Research scientist reviews microarray data

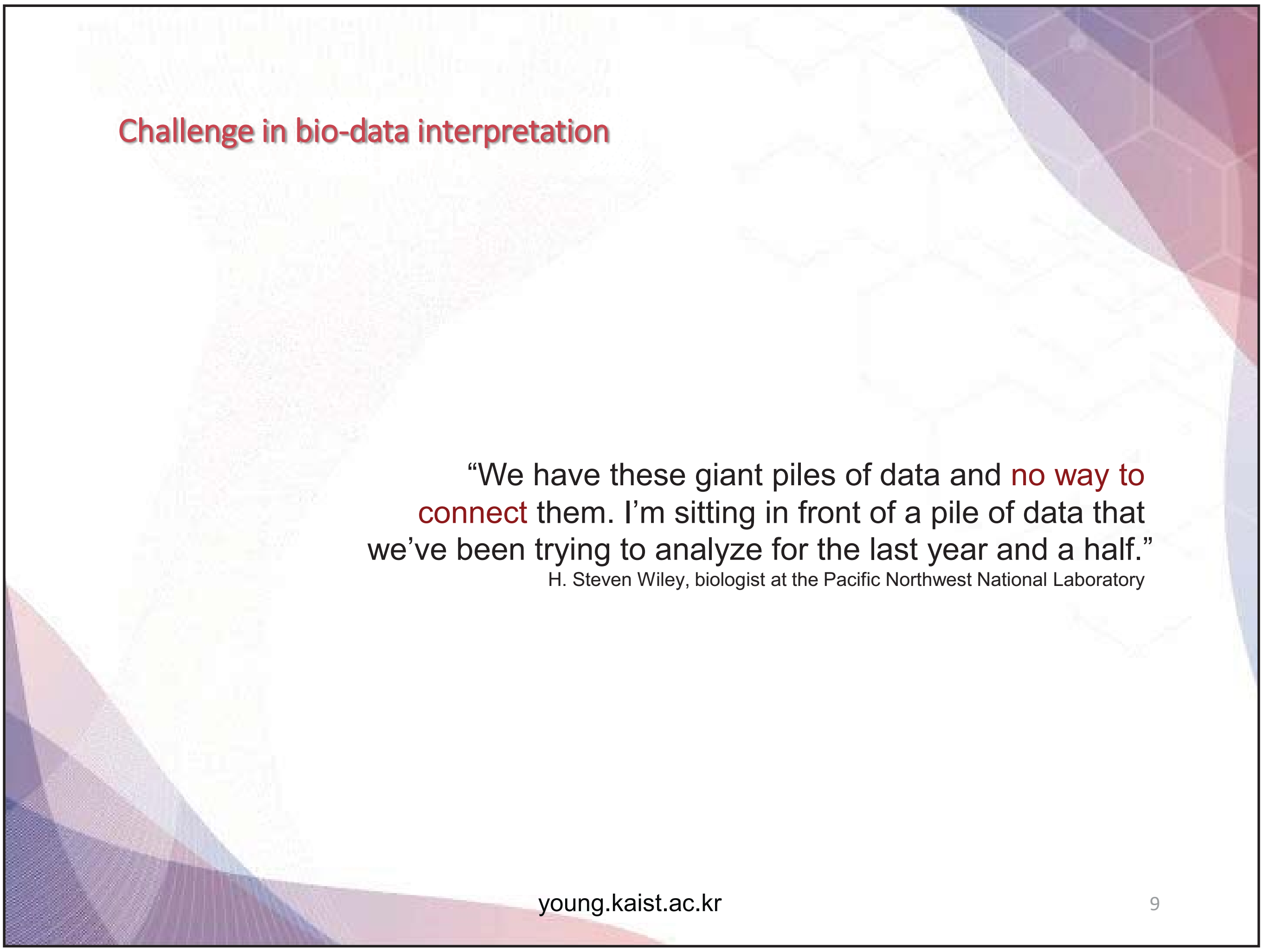
Accumulation of rich and genome-wide data



Stephens et al. 2015

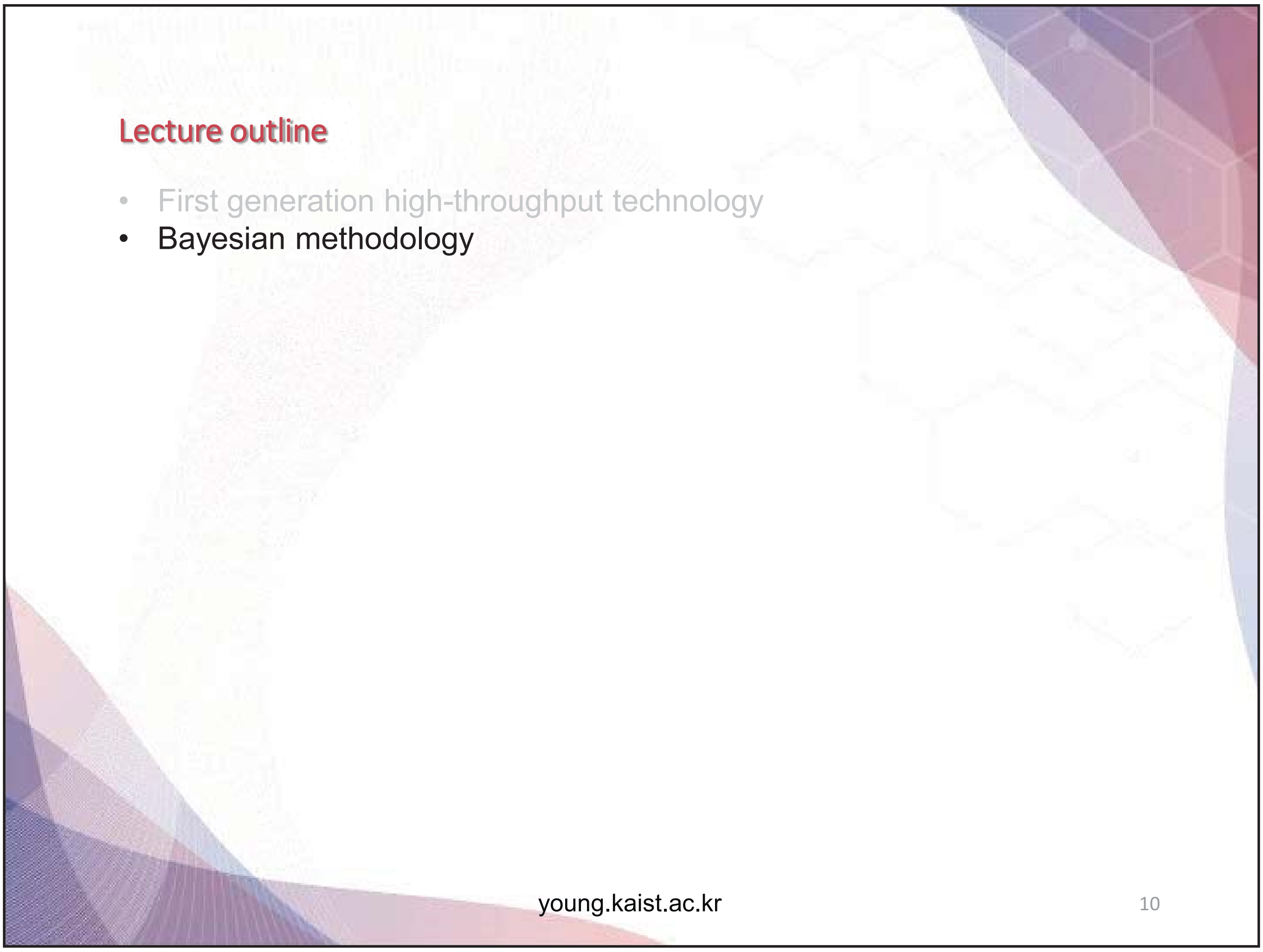
We now have over 2 Million Human Genomes!

TELOMERE-TO-TELOMERE CONSORTIUM
TOWARDS A COMPLETE REFERENCE OF HUMAN GENOME DIVERSITY
HUMAN PANGENOME



Challenge in bio-data interpretation

> “We have these giant piles of data and no way to connect them. I’m sitting in front of a pile of data that we’ve been trying to analyze for the last year and a half.”
>
> H. Steven Wiley, biologist at the Pacific Northwest National Laboratory

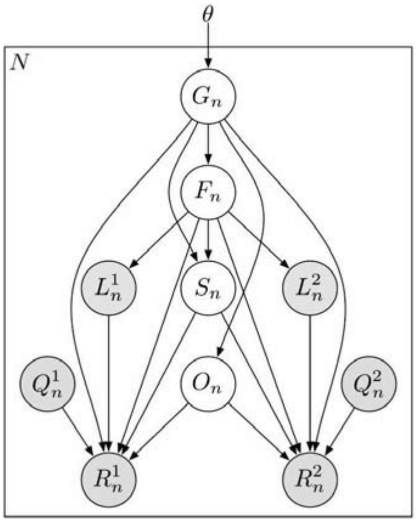


Lecture outline

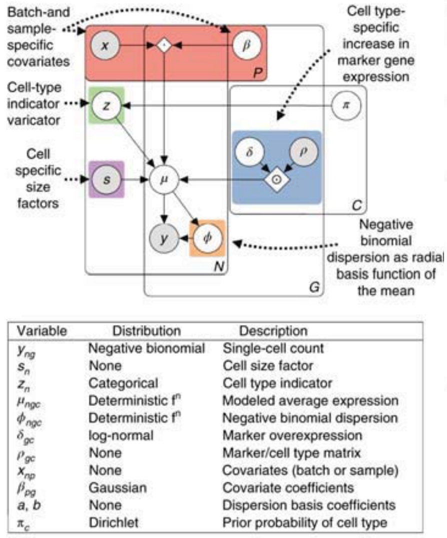
- First generation high-throughput technology
- Bayesian methodology

Bayesian methodology in bioinformatics

Examples of specialized bio-algorithms for gene quantification from RNA-seq and cell-type assignment from scRNA-seq



Graphical model used by RSEM
Cited by: 11,735



Batch-and sample- specific covariates
Cell-type indicator varicator
Cell specific size factors
Cell type- specific increase in marker gene expression
Negative binomial dispersion as radial basis function of the mean
x
β
P
z
π
δ
ρ
s
μ
C
y
ϕ
N
G

Variable	Distribution	Description
y_{ng}	Negative bionomial	Single-cell count
s_n	None	Cell size factor
z_n	Categorical	Cell type indicator
μ_{ngc}	Deterministic f^n	Modeled average expression
ϕ_{ngc}	Deterministic f^n	Negative binomial dispersion
δ_{gc}	log-normal	Marker overexpression
ρ_{gc}	None	Marker/cell type matrix
x_{np}	None	Covariates (batch or sample)
β_{pg}	Gaussian	Covariate coefficients
a, b	None	Dispersion basis coefficients
π_c	Dirichlet	Prior probability of cell type

Graphical model used by CellAssign

Li and Dewey 2011; Zhang et al. 2019

Bayesian approach for data analysis

Likelihood

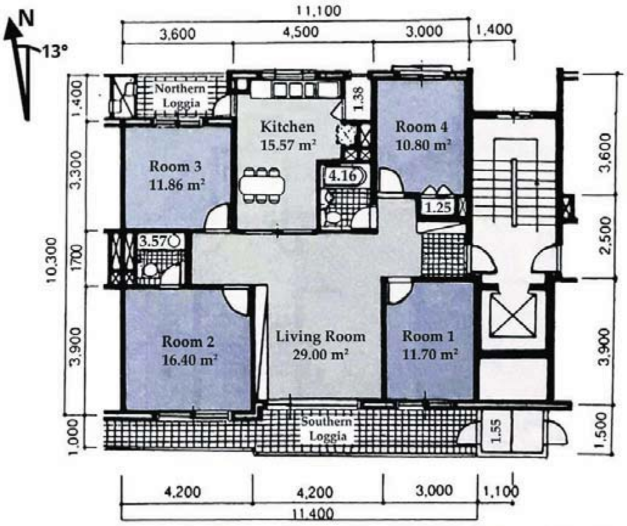
Prior

$$P(Parameter|Data) = \frac{P(Data|Parameter)P(Parameter)}{P(Data)}$$

Posterior

Bayesian vs. Frequentist reasoning

Example: Find your phone



Typical apartment floor plan in Korea

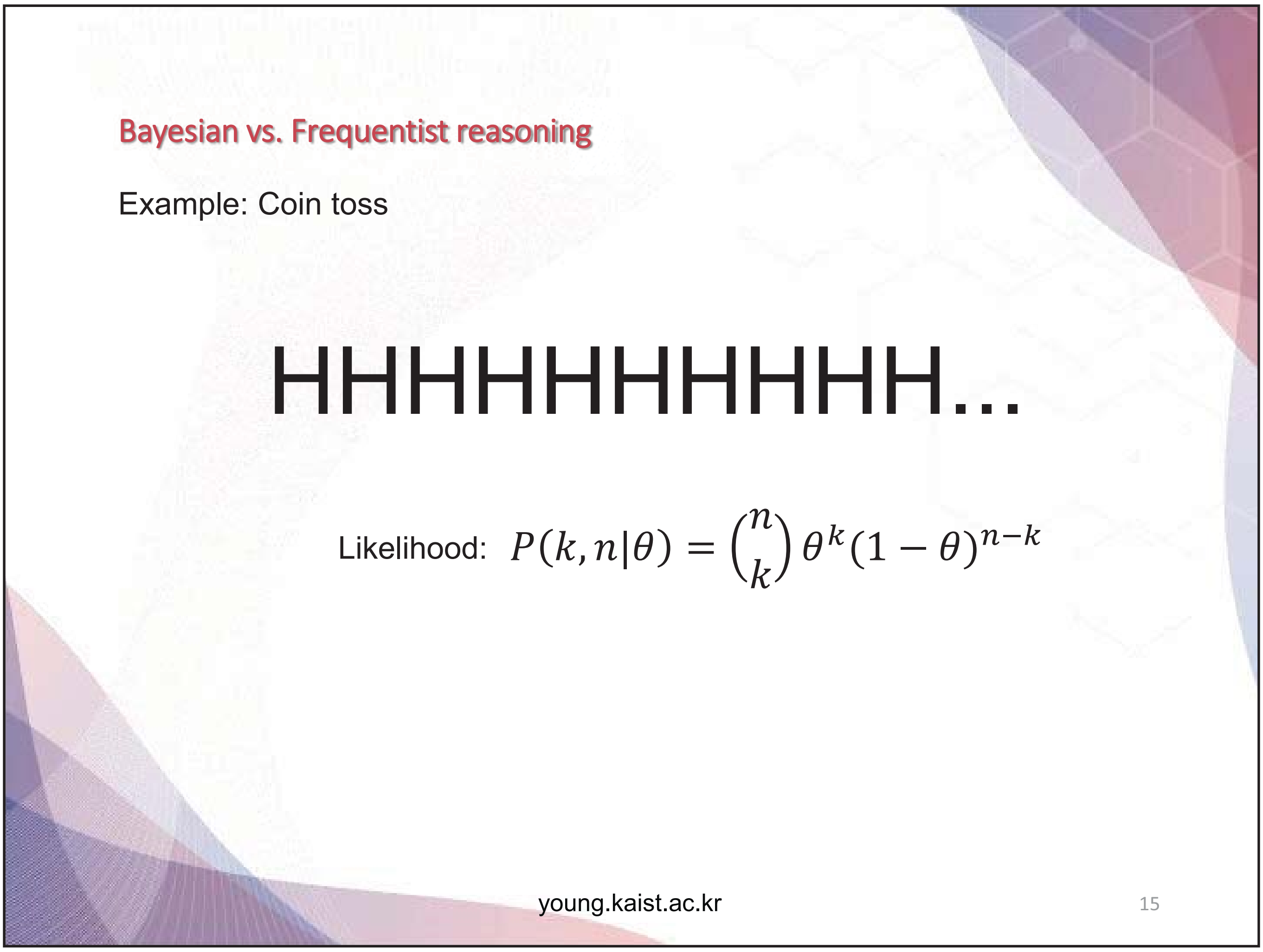
Image from Amoruso et al. 2018
under CC BY 4.0

Bayesian vs. Frequentist reasoning

Example: Coin toss

HHHHHHHHHH...

What is the probability that the next coin toss will return head?

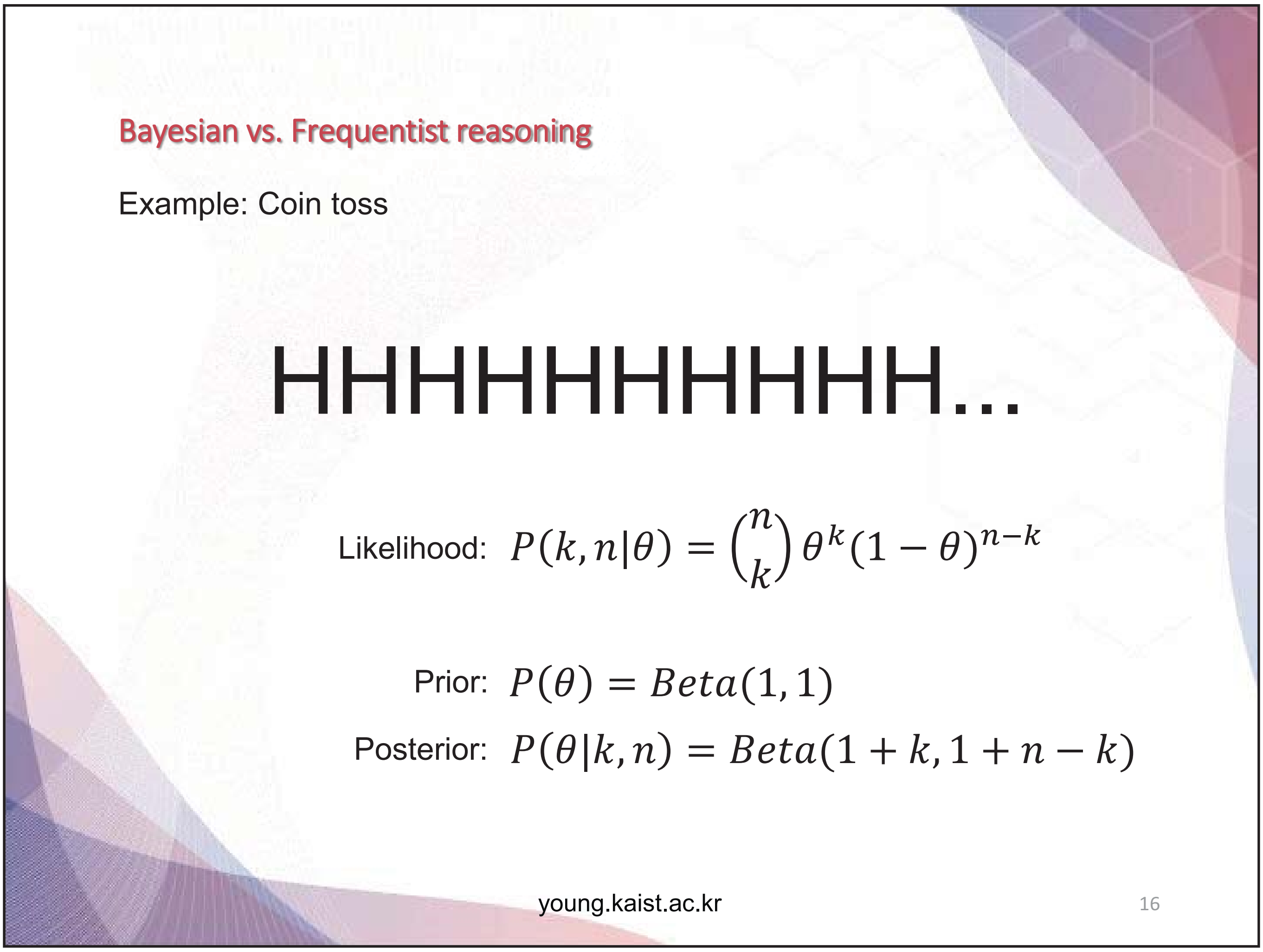


Bayesian vs. Frequentist reasoning

Example: Coin toss

HHHHHHHHHH...

Likelihood: $P(k,n|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$



Bayesian vs. Frequentist reasoning

Example: Coin toss

HHHHHHHHHH...

Likelihood: $P(k,n|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$

Prior: $P(\theta) = Beta(1,1)$

Posterior: $P(\theta|k,n) = Beta(1+k,1+n-k)$

Bayesian methodology as a generalization of Cromwell's rule

Cromwell's rule states that we should not use of probabilities of 1 or 0, except when applied to logical statements.

What is the probability that the sun will not rise tomorrow?

Photo by Klemen Tušar on Unsplash

Laplace's law of succession: define random variables

Example: Coin toss

- $X_i = \{\text{the value of the i} - \text{th coin toss; head} = 1 \text{ and tail} = 0\}$
- $S_n = \{\text{the total number of heads}\} = X_1 + \cdots + X_n$

Laplace's law of succession: observations

Example: Coin toss

- $X_i = \{\text{the value of the i} - \text{th coin toss; head} = 1 \text{ and tail} = 0\}$
- $S_n = \{\text{the total number of heads}\} = X_1 + \cdots + X_n$

HHHHHHHHHH

- $k = \{\text{number of heads}\} = 10$
- $n = \{\text{number of coin toss}\} = 10$

Laplace's law of succession: mathematical assumption

Example: Coin toss

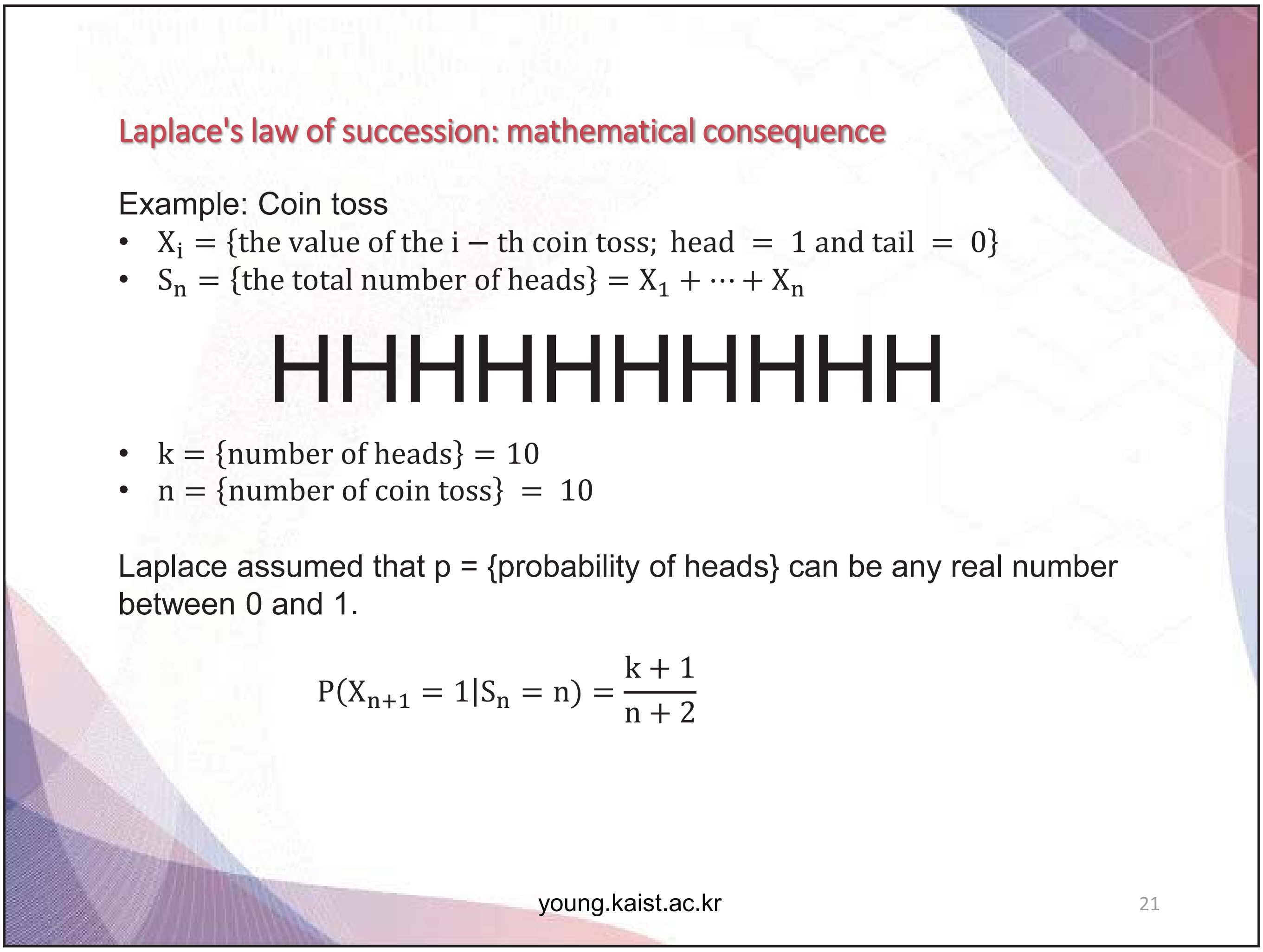
- $X_i = \{\text{the value of the i} - \text{th coin toss; head} = 1 \text{ and tail} = 0\}$
- $S_n = \{\text{the total number of heads}\} = X_1 + \cdots + X_n$

HHHHHHHHHH

- $k = \{\text{number of heads}\} = 10$
- $n = \{\text{number of coin toss}\} = 10$

Laplace assumed that p = {probability of heads} can be any real number between 0 and 1.

What is the probability that the next coin toss is heads?



Laplace's law of succession: mathematical consequence

Example: Coin toss

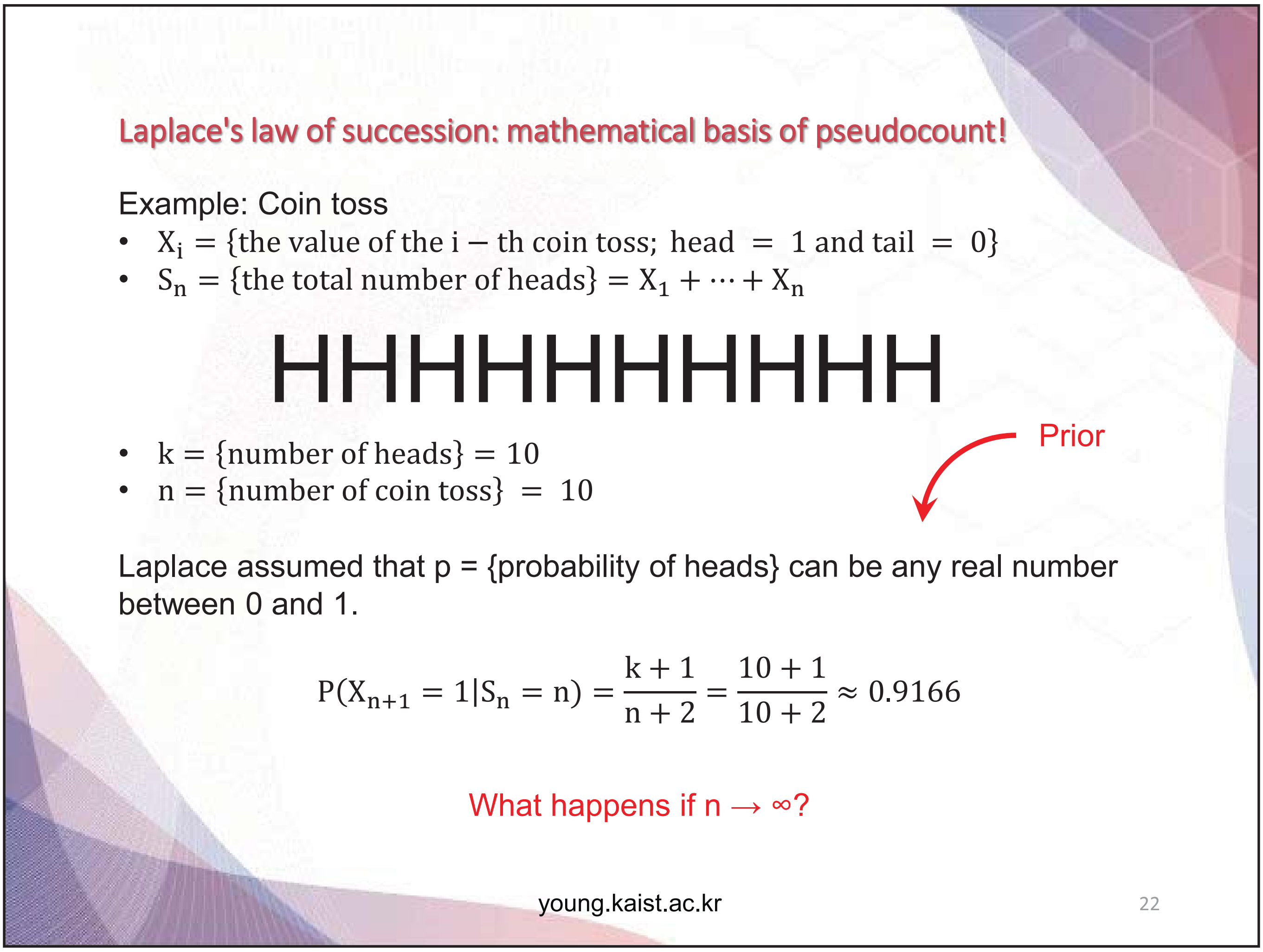
- $X_i = \{\text{the value of the i} - \text{th coin toss; head } = 1 \text{ and tail } = 0\}$
- $S_n = \{\text{the total number of heads}\} = X_1 + \cdots + X_n$

HHHHHHHHHH

- $k = \{\text{number of heads}\} = 10$
- $n = \{\text{number of coin toss}\} = 10$

Laplace assumed that p = {probability of heads} can be any real number between 0 and 1.

$$P(X_{n+1} = 1 | S_n = n) = \frac{k+1}{n+2}$$



Laplace's law of succession: mathematical basis of pseudocount!

Example: Coin toss

- $X_i = \{\text{the value of the i} - \text{th coin toss; head } = 1 \text{ and tail } = 0\}$
- $S_n = \{\text{the total number of heads}\} = X_1 + \cdots + X_n$

HHHHHHHHHH

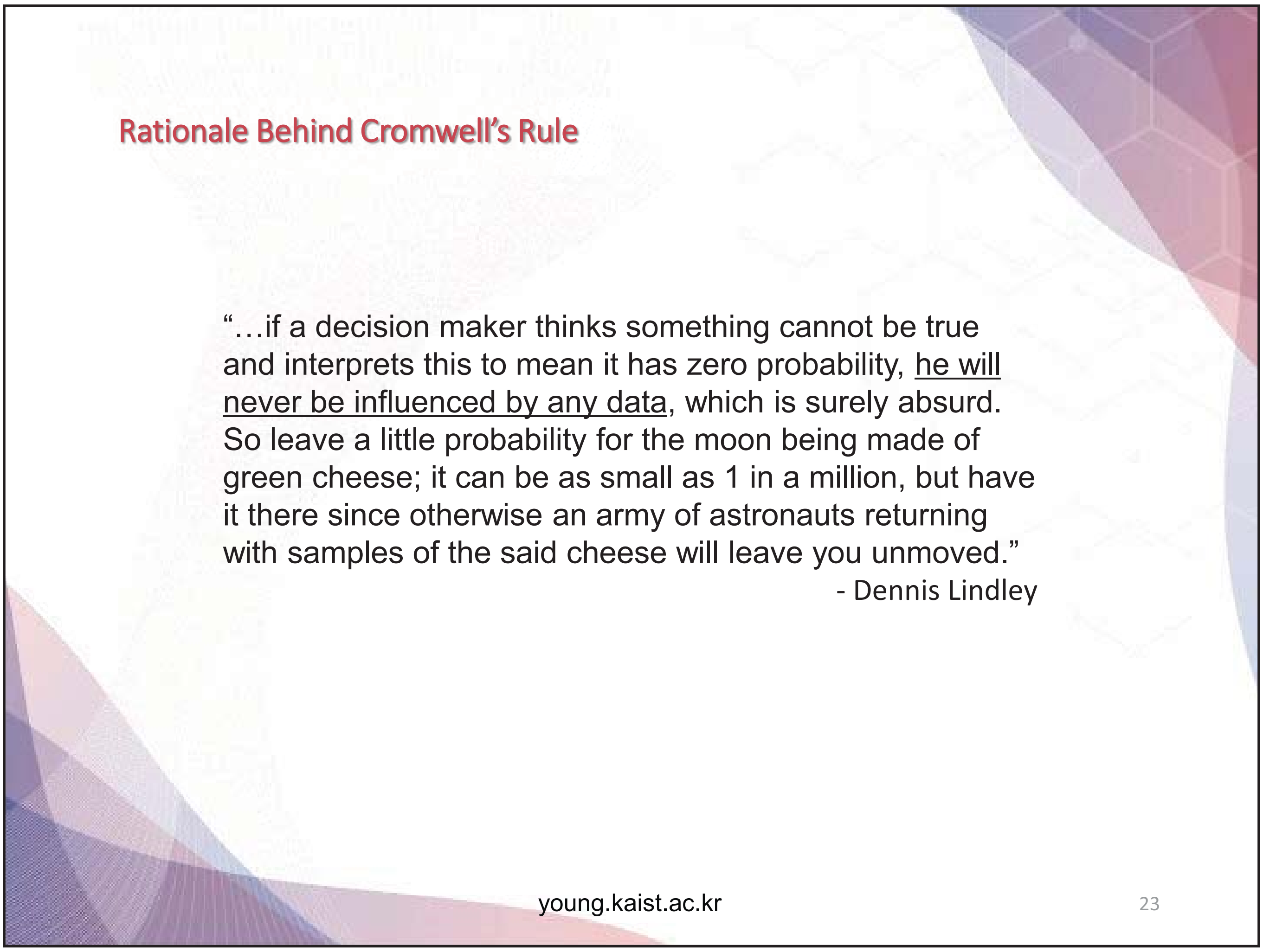
- $k = \{\text{number of heads}\} = 10$
- $n = \{\text{number of coin toss}\} = 10$

Prior

Laplace assumed that p = {probability of heads} can be any real number between 0 and 1.

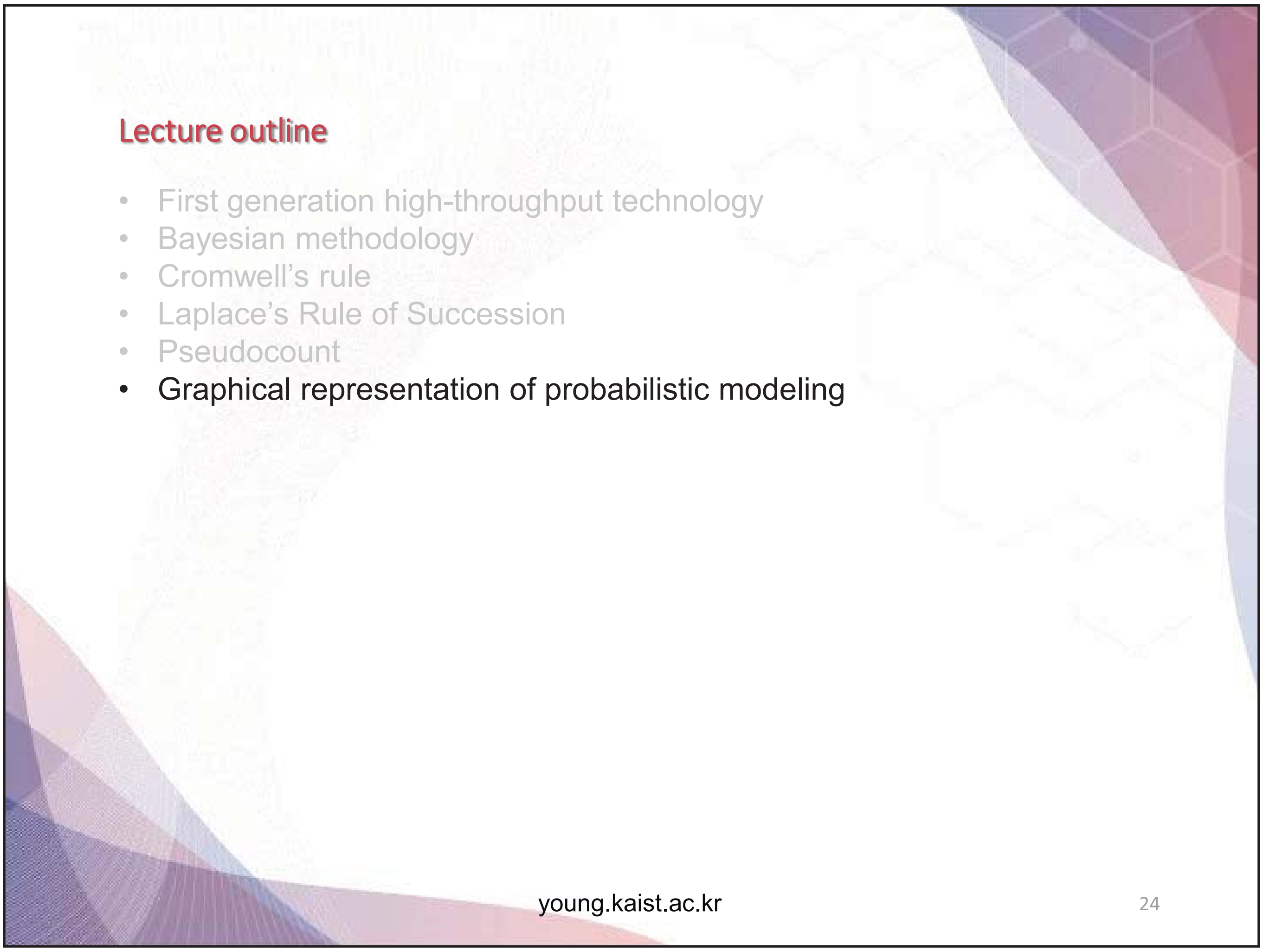
$$P(X_{n+1} = 1 | S_n = n) = \frac{k+1}{n+2} = \frac{10+1}{10+2} \approx 0.9166$$

What happens if n → ∞?



Rationale Behind Cromwell's Rule

> "…if a decision maker thinks something cannot be true and interprets this to mean it has zero probability, he will never be influenced by any data, which is surely absurd. So leave a little probability for the moon being made of green cheese; it can be as small as 1 in a million, but have it there since otherwise an army of astronauts returning with samples of the said cheese will leave you unmoved."
>
> - Dennis Lindley



Lecture outline

- First generation high-throughput technology
- Bayesian methodology
- Cromwell's rule
- Laplace's Rule of Succession
- Pseudocount
- Graphical representation of probabilistic modeling

Recall conditional probability

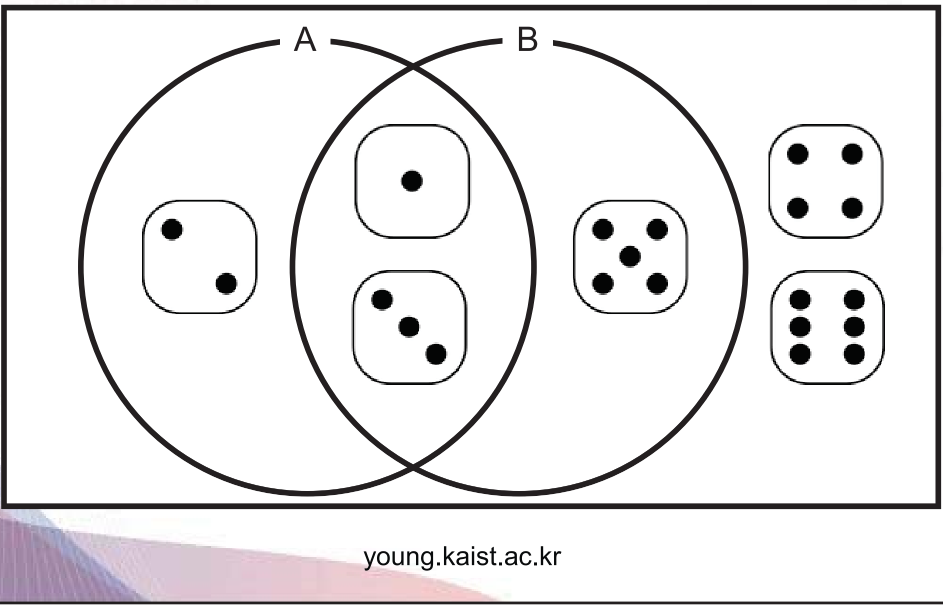
- A = { rolling a dice and it’s value is less than 4 }
- B = { rolling a dice and it’s value is an odd number }

$$\mathrm{P(\,B \mid A\,)} = \frac{\mathrm{P(\,B\,,\ A\,)}}{\mathrm{P(\,A\,)}} = ?$$

Recall conditional probability

- A = { rolling a dice and it’s value is less than 4 }
- B = { rolling a dice and it’s value is an odd number }

$$\mathrm{P(\,B \mid A\,)} = \frac{\mathrm{P(\,B\,,\ A\,)}}{\mathrm{P(\,A\,)}} = \frac{\#\{\text{rolling a 1 or 3}\}}{\#\{\text{rolling a 1, 2, or 3}\}} = \frac{2}{3}$$



Notations for probabilistic graphical models (i.e. Bayesian networks)

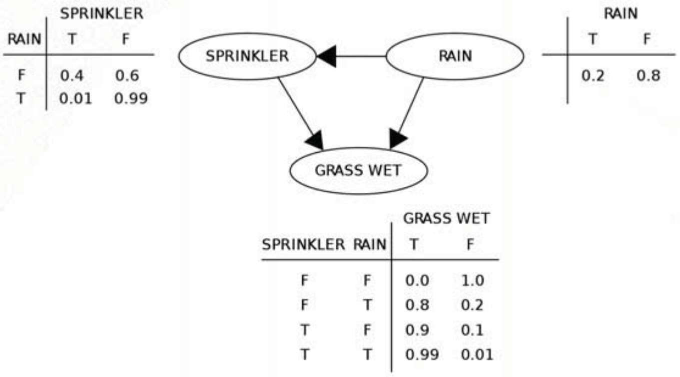
○ Random variable

◎ or ● Observation

Ⓐ → Ⓑ Conditional probability P(B | A)

Example of probabilistic graphical models

Given the following probabilistic graphical model, what is the equivalent factorization of the joint probability?



	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99

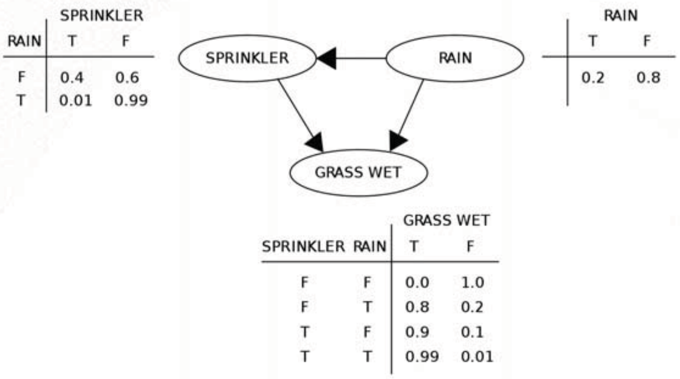
RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

$$\mathrm{P}(\mathrm{G}, \mathrm{S}, \mathrm{R}) = \ ?$$

Example of probabilistic graphical models

Given the following probabilistic graphical model, what is the equivalent factorization of the joint probability?



	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99

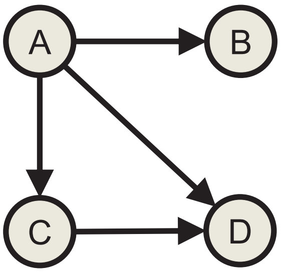
RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

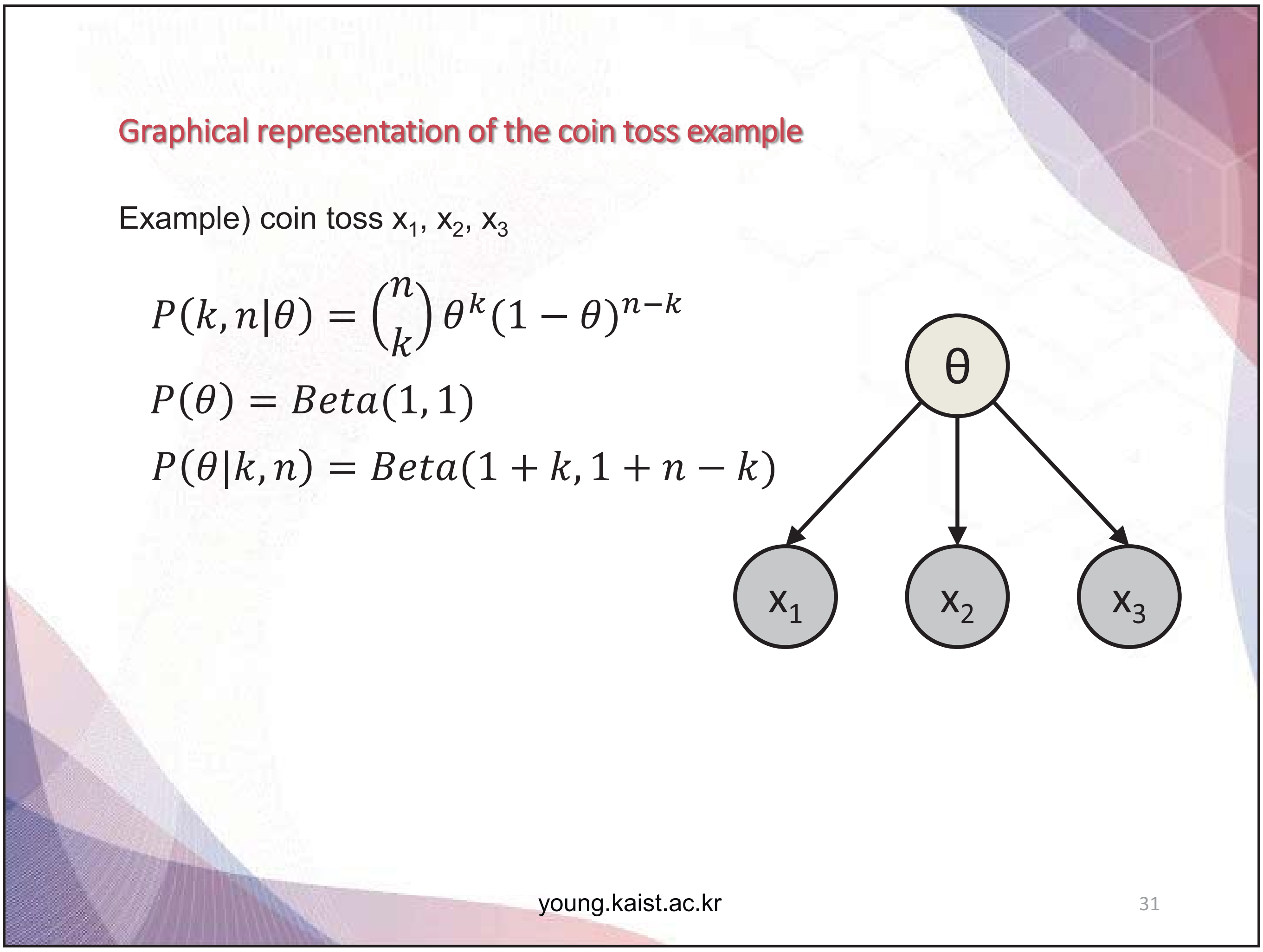
$$\mathrm{P(G, S, R) = P(R) \cdot P(S|R) \cdot P(G|S, R)}$$

Take-home exercise: reading probabilistic graphical models

Given the following probabilistic graphical model, what is the equivalent factorization of the joint probability?



$$\mathrm{P(A, B, C, D) = ?}$$



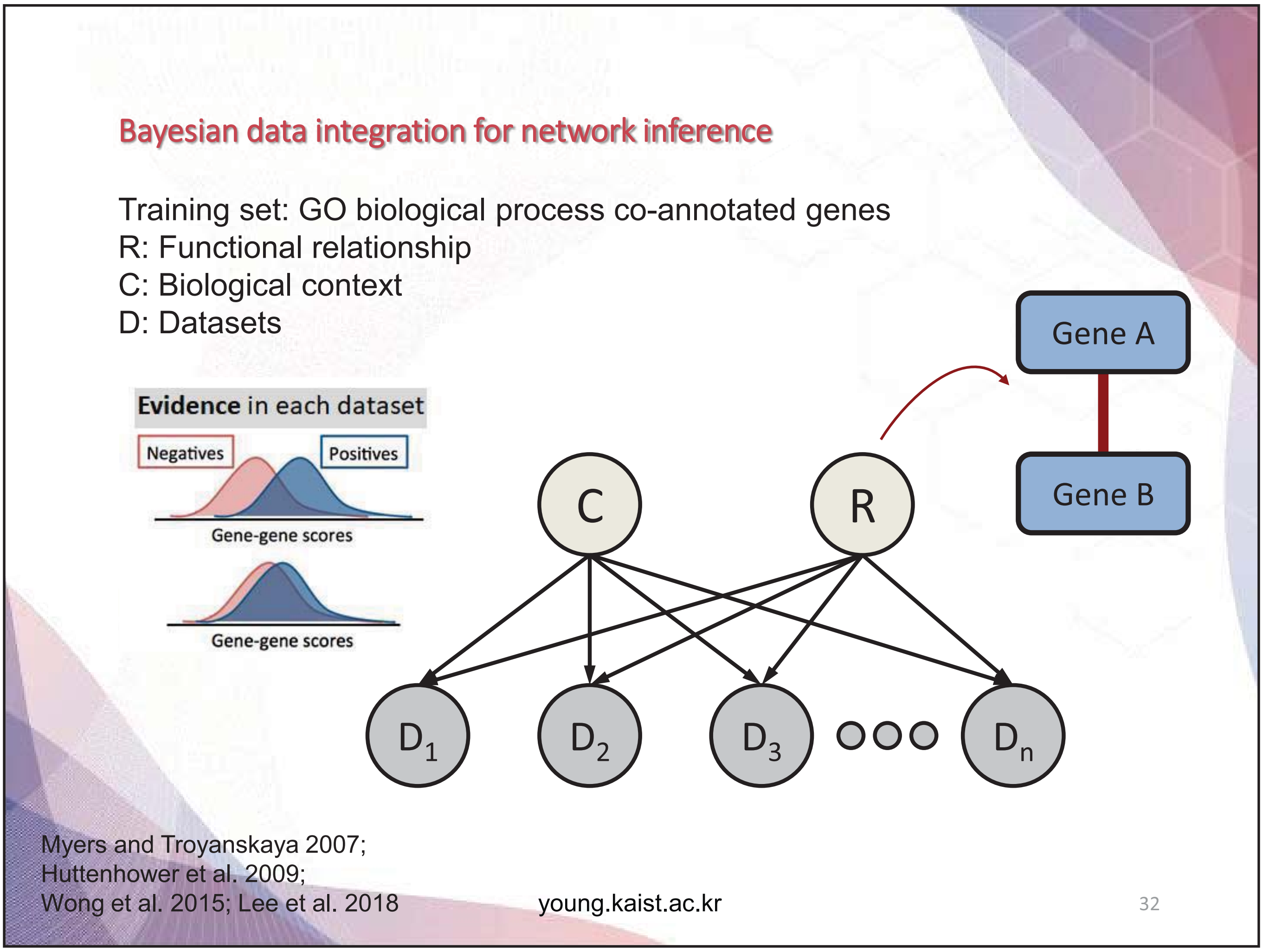
Graphical representation of the coin toss example

Example) coin toss x_1, x_2, x_3

$$P(k,n|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

$$P(\theta) = Beta(1,1)$$

$$P(\theta|k,n) = Beta(1+k, 1+n-k)$$



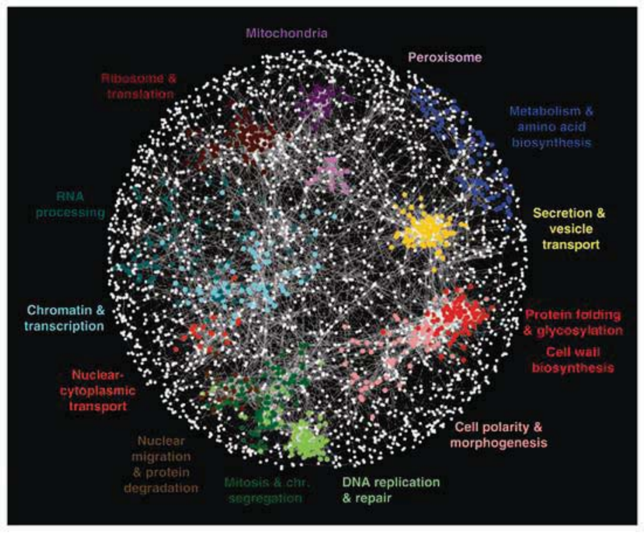
Bayesian data integration for network inference

Training set: GO biological process co-annotated genes
R: Functional relationship
C: Biological context
D: Datasets

Myers and Troyanskaya 2007;
Huttenhower et al. 2009;
Wong et al. 2015; Lee et al. 2018

Key in understanding the social network of the cell

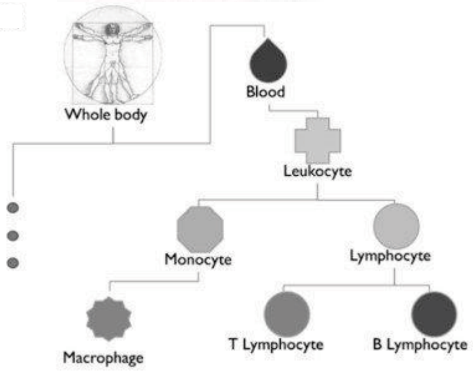
Each node represents a single gene/protein and the specific network connections are responsible for each biological process.


Mitochondria
Peroxisome
Ribosome & translation
Metabolism & amino acid biosynthesis
RNA processing
Secretion & vesicle transport
Chromatin & transcription
Protein folding & glycosylation
Cell wall biosynthesis
Nuclear-cytoplasmic transport
Cell polarity & morphogenesis
Nuclear migration & protein degradation
Mitosis & chr. segregation
DNA replication & repair

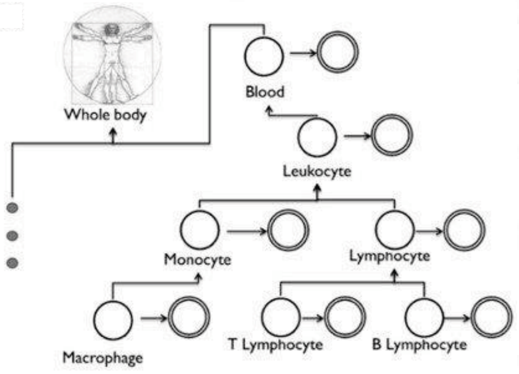
Costanzo M, Baryshnikova A, et al. 2010

Hierarchical-aware integration of computation and genome-wide experiments

Graphical modeling based on known hierarchical associations for multi-label classification



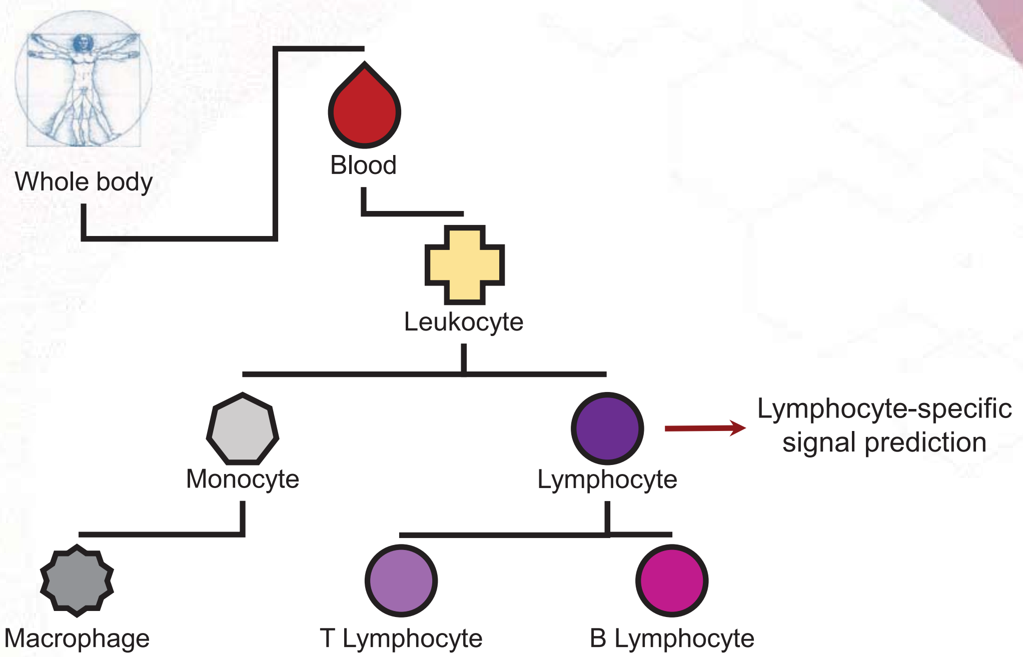
Hierarchy of tissues and cell-types



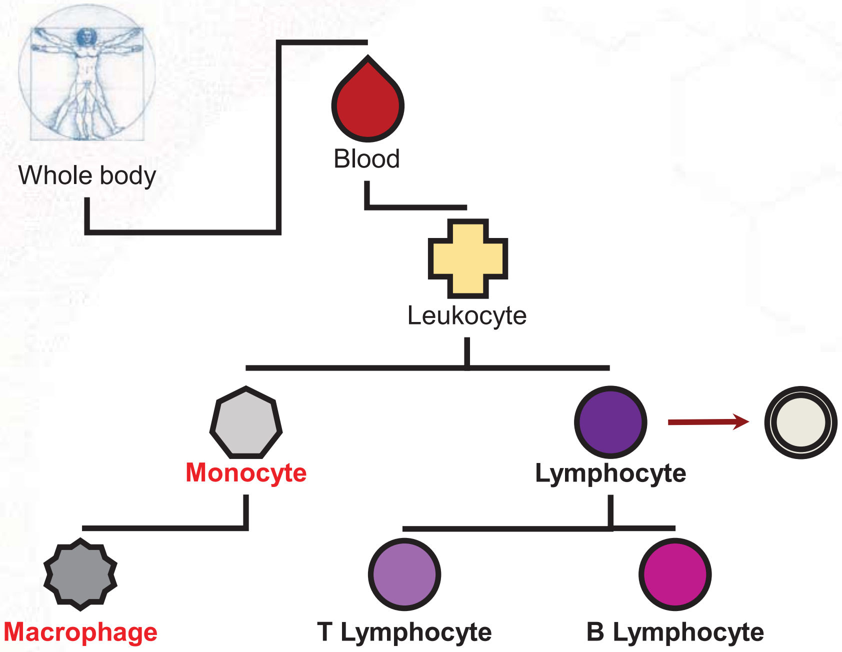
Graphical model for tissue and cell-type classification and prediction

Barutcuoglu et al. 2006;
Guan et al. 2008; Park et al. 2010;
Lee et al. 2013; Lee et al. 2019

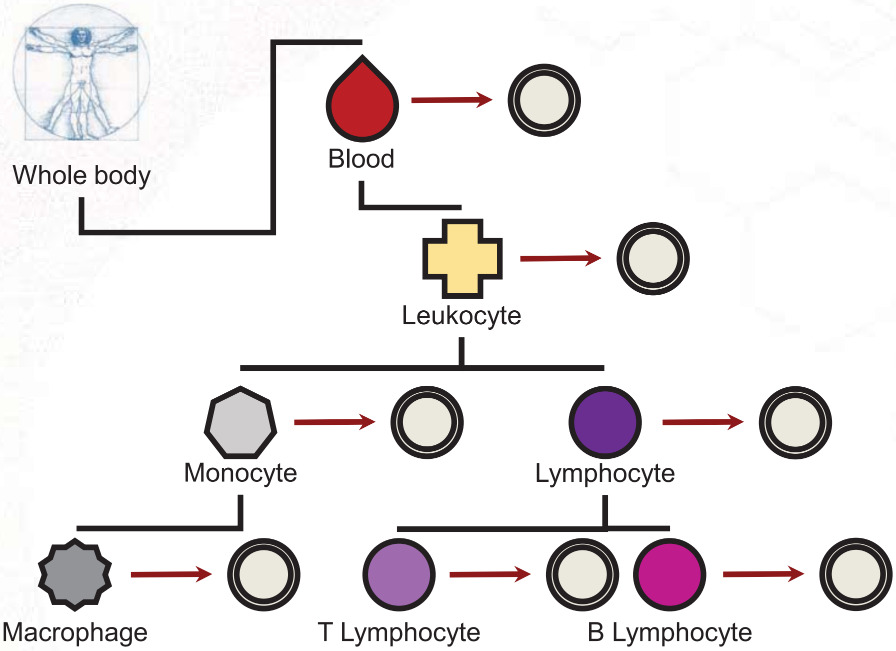
Construct individual classifier for lymphocyte



Construct individual classifier for lymphocyte



Construct individual classifier for each tissue and cell-type



Whole body

Blood

Leukocyte

Monocyte

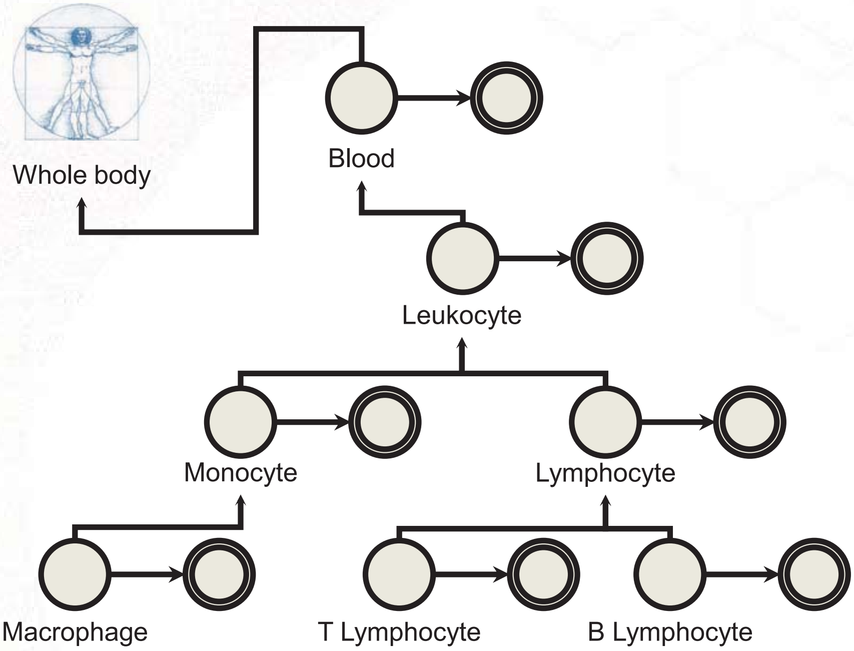
Lymphocyte

Macrophage

T Lymphocyte

B Lymphocyte

Model aggregation (Bayesian Correction)



Whole body

Blood

Leukocyte

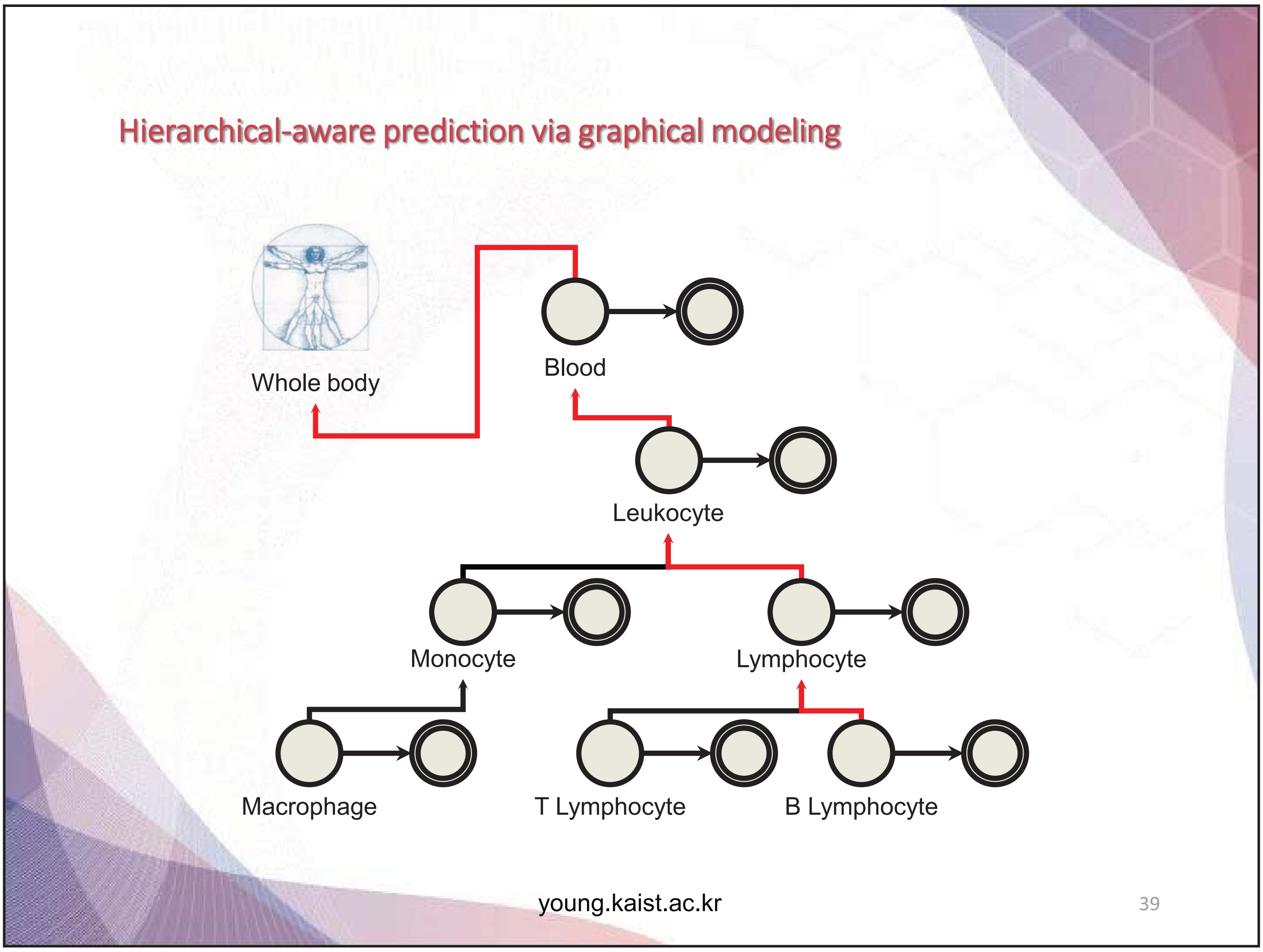
Monocyte

Lymphocyte

Macrophage

T Lymphocyte

B Lymphocyte



Hierarchical-aware prediction via graphical modeling

Whole body

Blood

Leukocyte

Monocyte

Lymphocyte

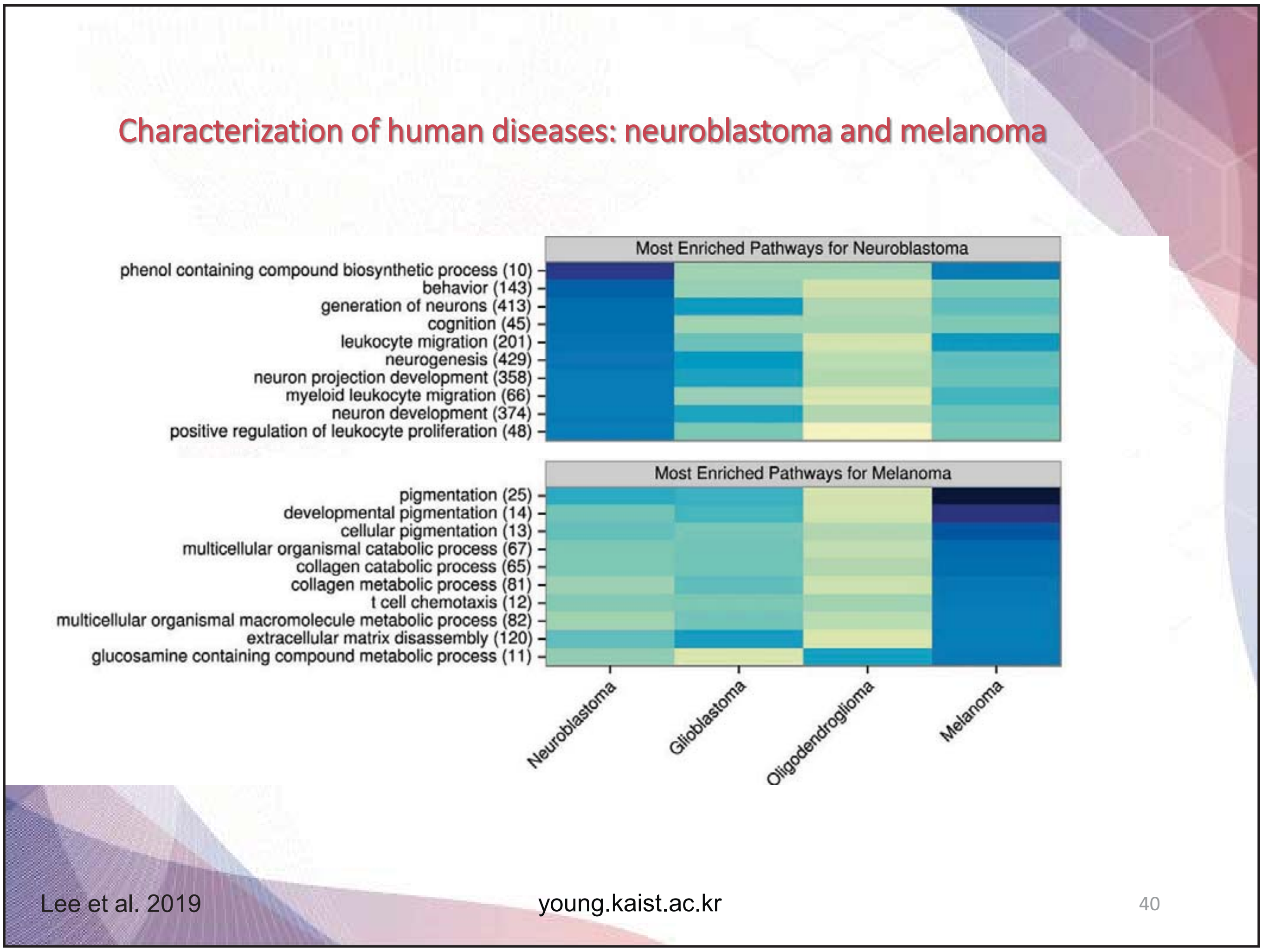
Macrophage

T Lymphocyte

B Lymphocyte

young.kaist.ac.kr

39



Characterization of human diseases: neuroblastoma and melanoma

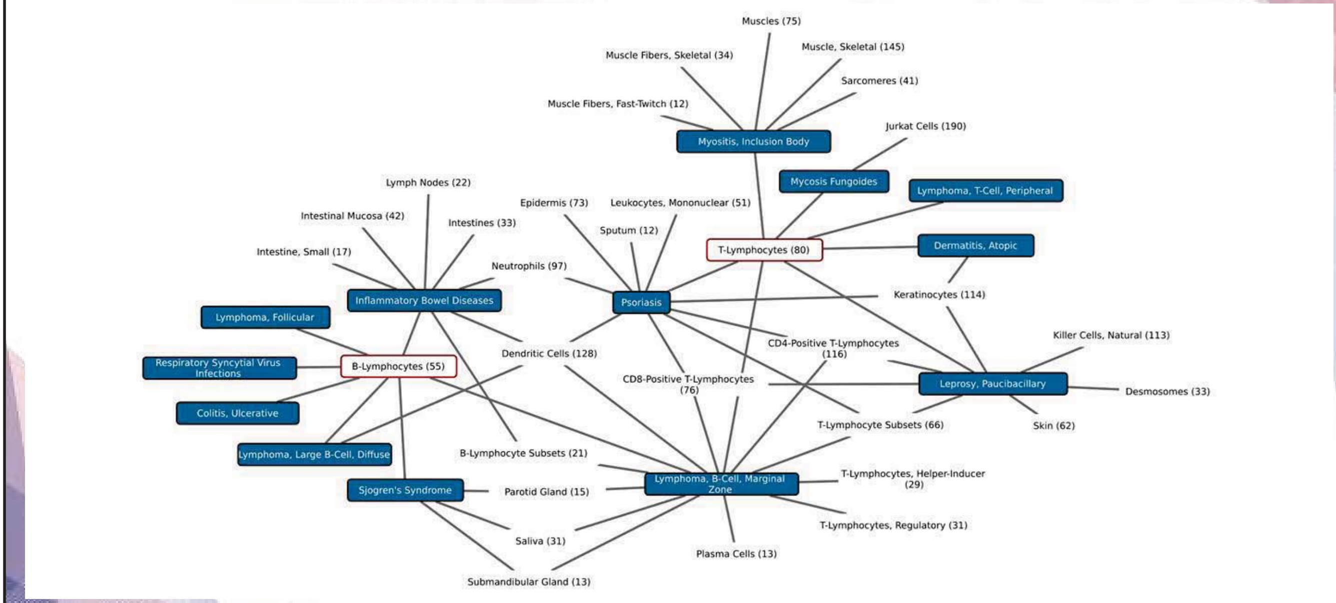
Lee et al. 2019

young.kaist.ac.kr

40

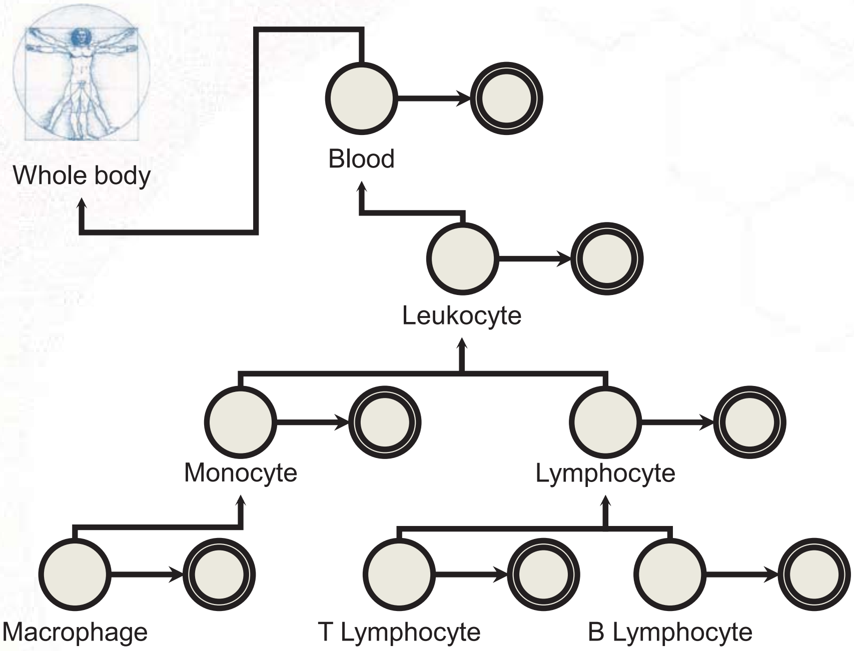
Tissue-specificity of human complex diseases

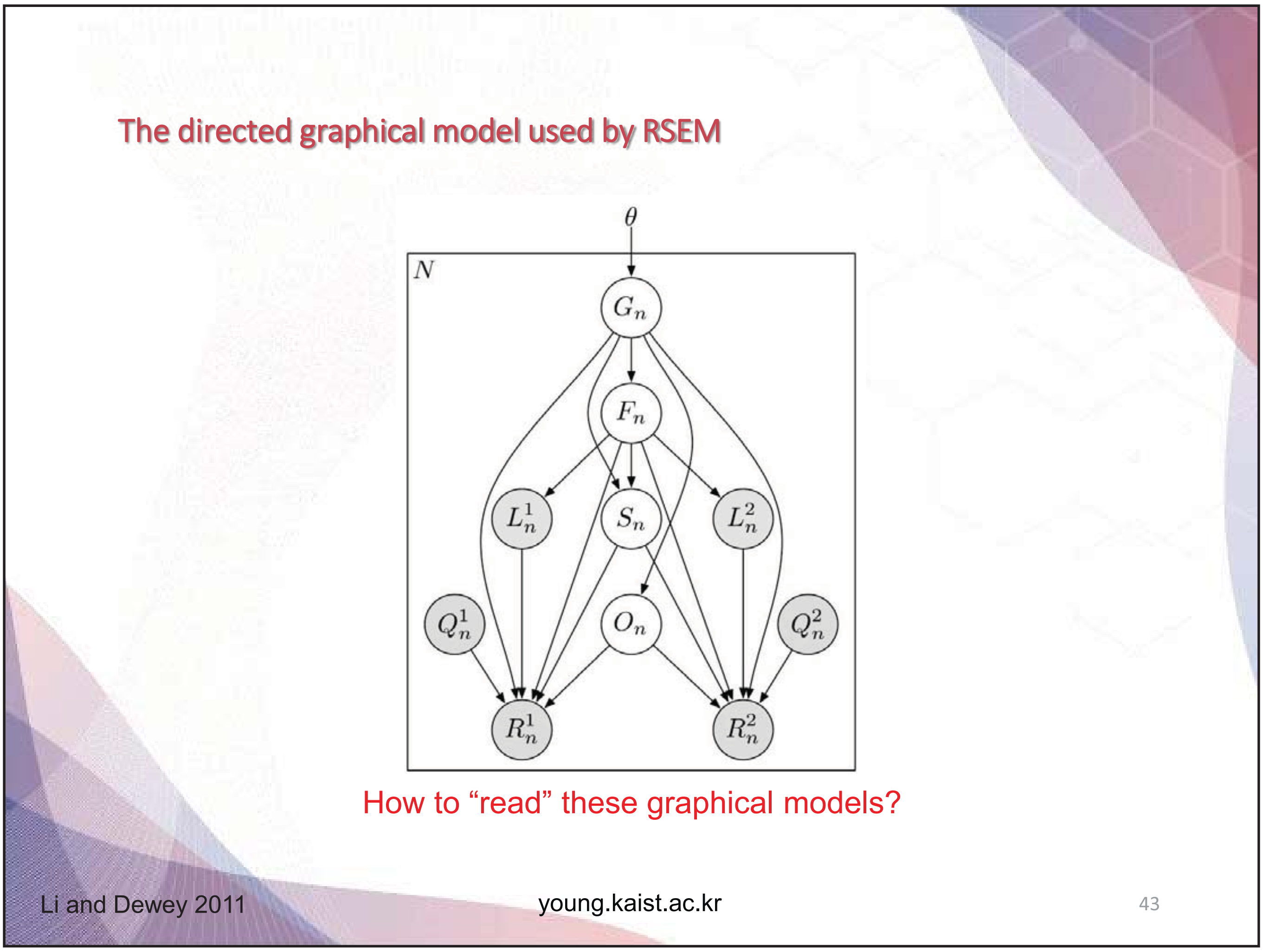
Bipartite graph of human tissues and complex diseases related to T-cells and B-cells



Lee et al. 2019

Take-home exercise: what is the equivalent factorization?

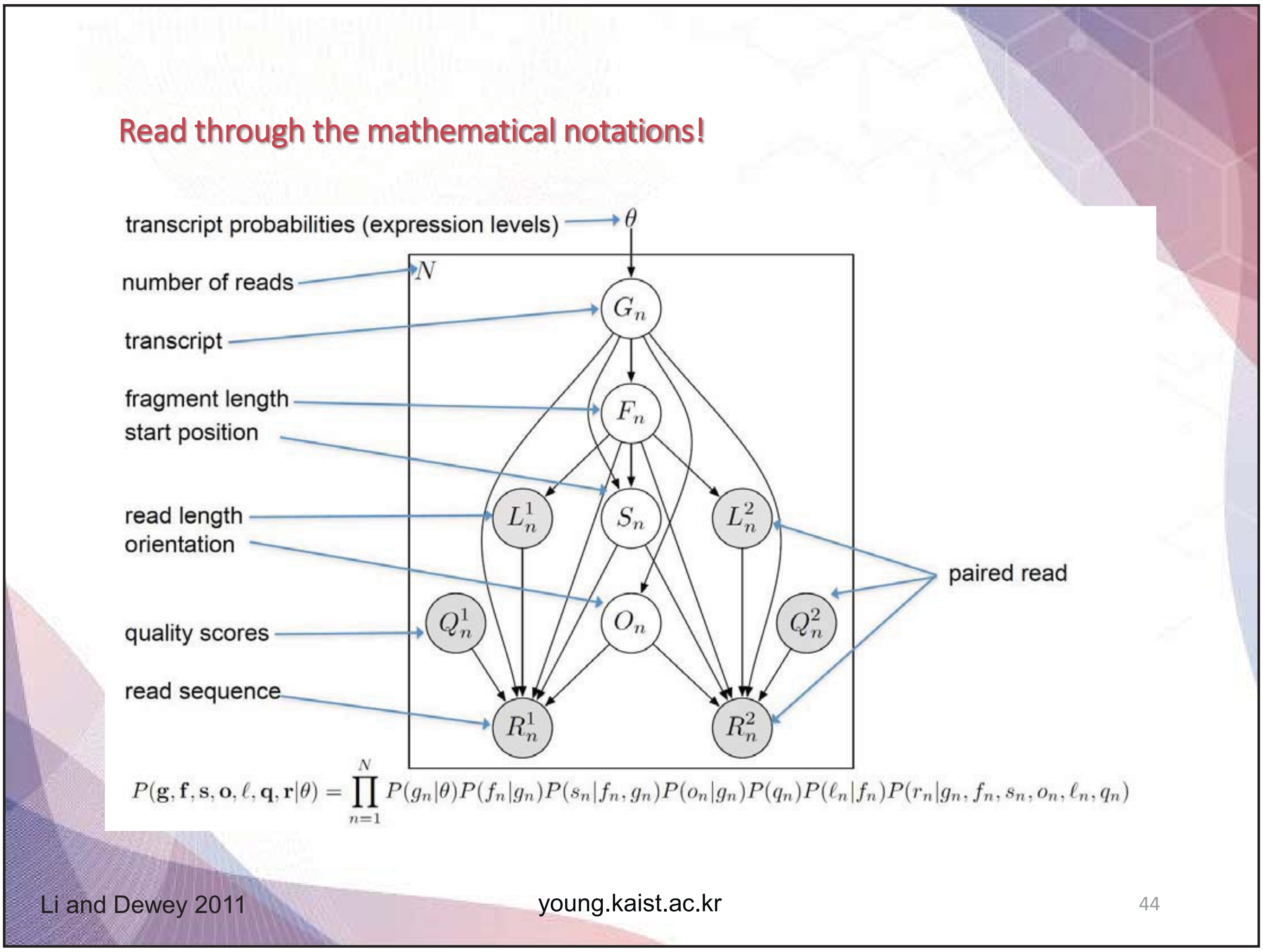




The directed graphical model used by RSEM

How to “read” these graphical models?

Li and Dewey 2011

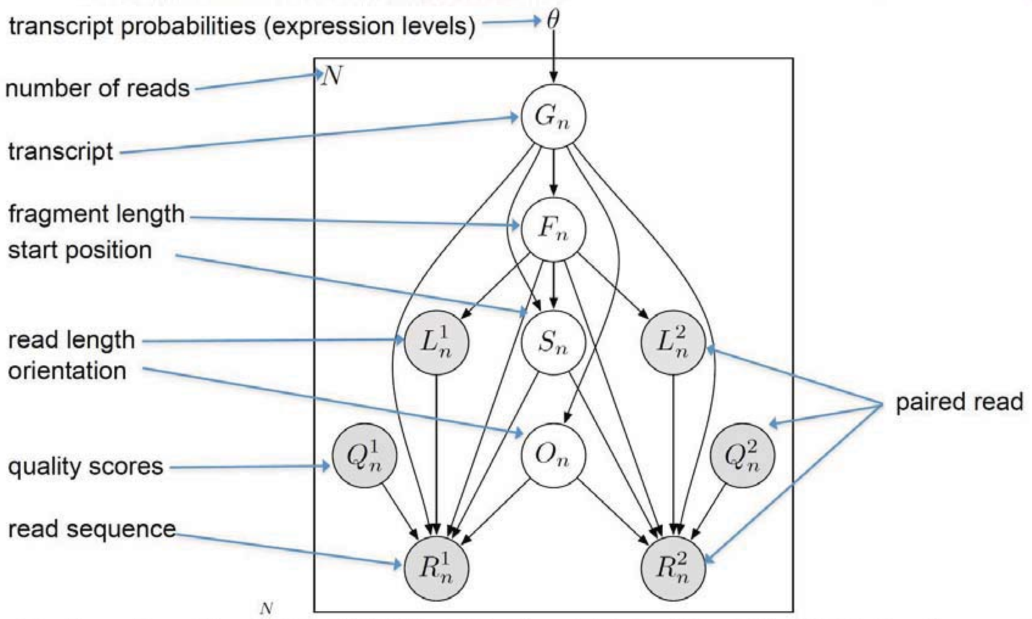


Read through the mathematical notations!

$$P(\mathbf{g},\mathbf{f},\mathbf{s},\mathbf{o},\boldsymbol{\ell},\mathbf{q},\mathbf{r}|\theta) = \prod_{n=1}^{N} P(g_n|\theta)P(f_n|g_n)P(s_n|f_n,g_n)P(o_n|g_n)P(q_n)P(\ell_n|f_n)P(r_n|g_n,f_n,s_n,o_n,\ell_n,q_n)$$

Li and Dewey 2011

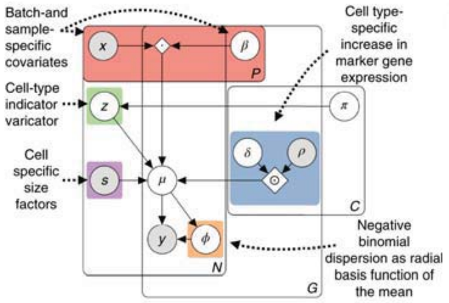
Now, what does this graphical modal say about RSEM?

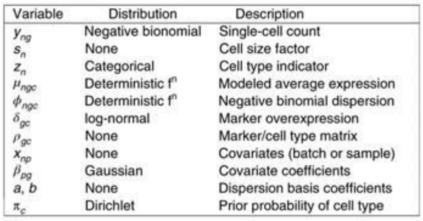


$$P(\mathbf{g},\mathbf{f},\mathbf{s},\mathbf{o},\boldsymbol{\ell},\mathbf{q},\mathbf{r}|\theta) = \prod_{n=1}^{N} P(g_n|\theta)P(f_n|g_n)P(s_n|f_n,g_n)P(o_n|g_n)P(q_n)P(\ell_n|f_n)P(r_n|g_n,f_n,s_n,o_n,\ell_n,q_n)$$

Li and Dewey 2011

Take-home exercise: what does the graphical model say about CellAssign?


Batch-and sample-specific covariates
x
β
P
Cell type-specific increase in marker gene expression
Cell-type indicator varicator
z
π
Cell specific size factors
s
μ
δ
ρ
C
y
ϕ
N
G
Negative binomial dispersion as radial basis function of the mean

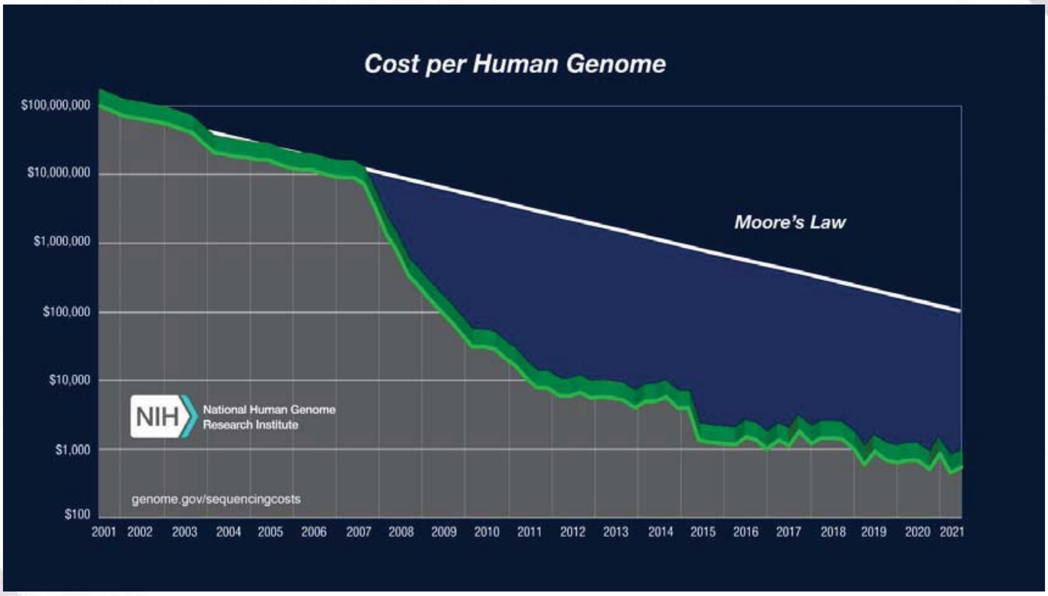


Variable	Distribution	Description
y_{ng}	Negative bionomial	Single-cell count
s_n	None	Cell size factor
z_n	Categorical	Cell type indicator
μ_{ngc}	Deterministic f^n	Modeled average expression
ϕ_{ngc}	Deterministic f^n	Negative binomial dispersion
δ_{gc}	log-normal	Marker overexpression
ρ_{gc}	None	Marker/cell type matrix
x_{np}	None	Covariates (batch or sample)
β_{pg}	Gaussian	Covariate coefficients
a, b	None	Dispersion basis coefficients
π_c	Dirichlet	Prior probability of cell type

Can you write down the factorization
for the graphical model used by CellAssign?

Zhang et al. 2019

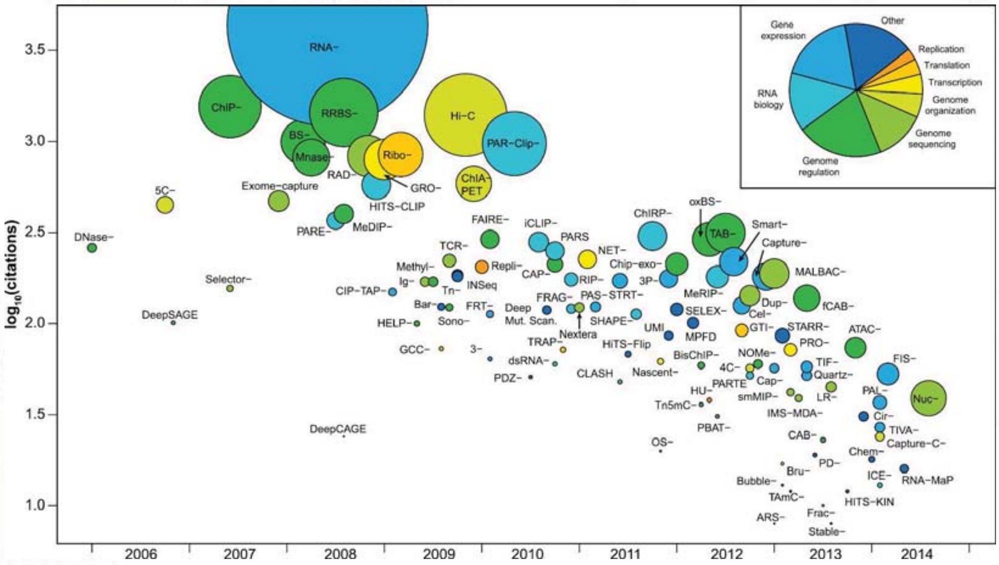
Data generation is no longer the rate limiting factor



Cost per genome data - 2021

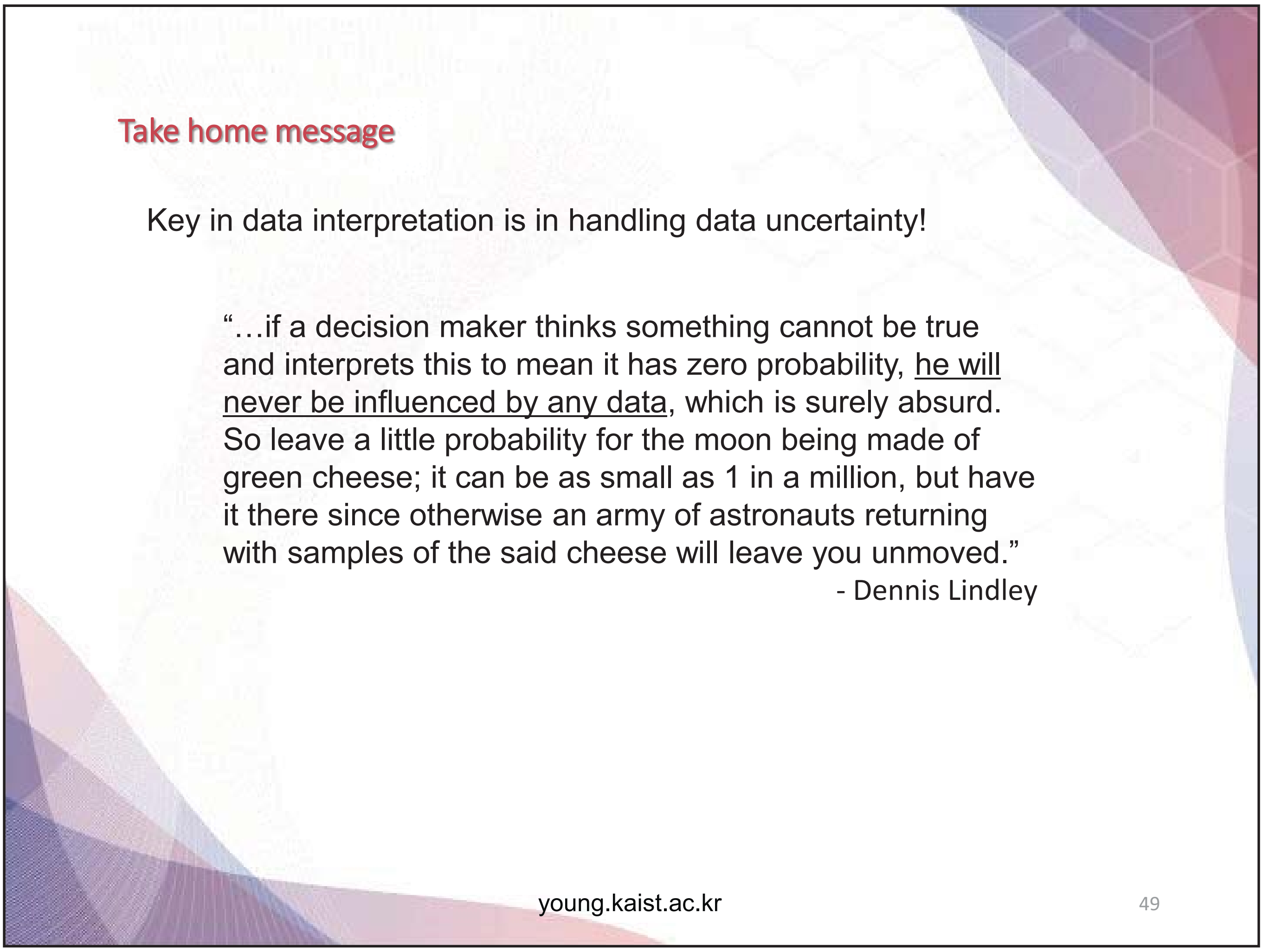
young.kaist.ac.kr

Advance in biochemical and high-throughput techniques



Reuter, Spacek and Snyder et al. 2015

young.kaist.ac.kr



Take home message

Key in data interpretation is in handling data uncertainty!

> "…if a decision maker thinks something cannot be true and interprets this to mean it has zero probability, he will never be influenced by any data, which is surely absurd. So leave a little probability for the moon being made of green cheese; it can be as small as 1 in a million, but have it there since otherwise an army of astronauts returning with samples of the said cheese will leave you unmoved."
>
> - Dennis Lindley

young.kaist.ac.kr

49

KSBi-BIML

Join me for the online Q&A session!

Young-suk Lee (이영석)

Assistant Professor
Department of Bio and Brain Engineering,
Korea Advanced Institute of Science and Technology (KAIST)
Email: youngl@kaist.ac.kr
Web: young.kaist.ac.kr