Data and Model: Small or Big

Keunkwan Ryu
Department of Economics
Seoul National University

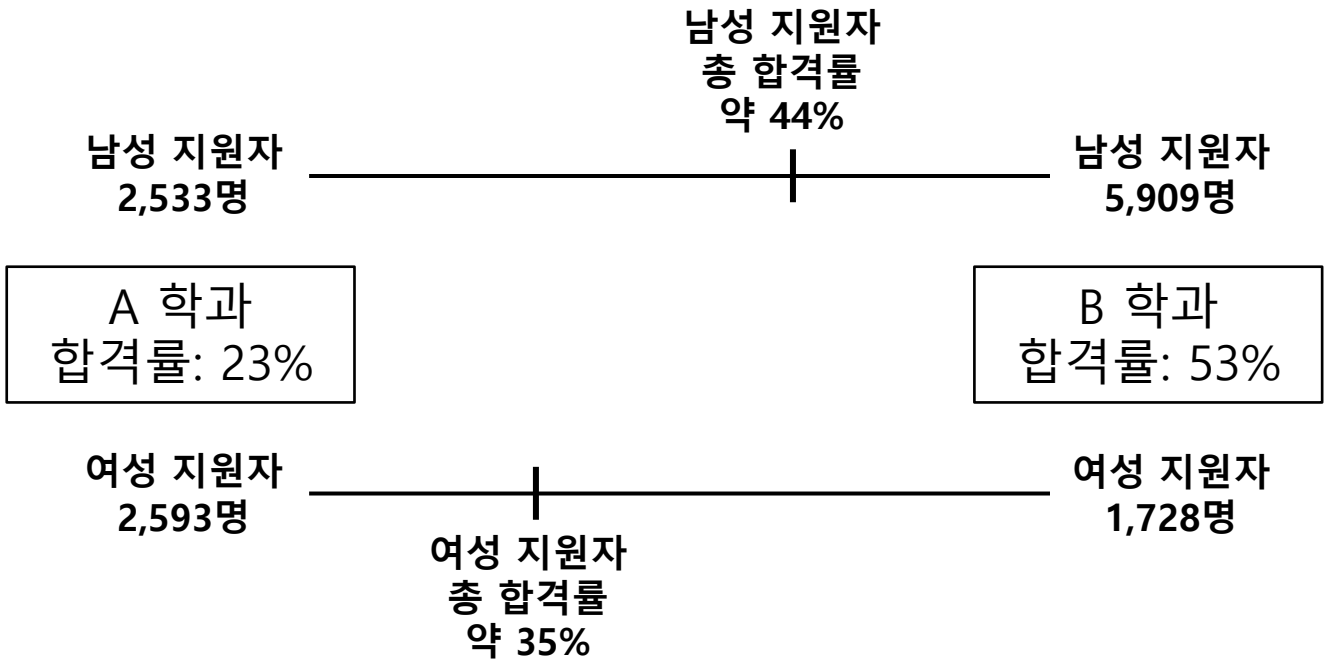
“A bit is a bit.”

-All forms of data look same in the eyes of computer.
-Anything can be represented as a sequence of bits.
-Numbers, categories, images, voice, texts, …
-Revealed preference. Data.

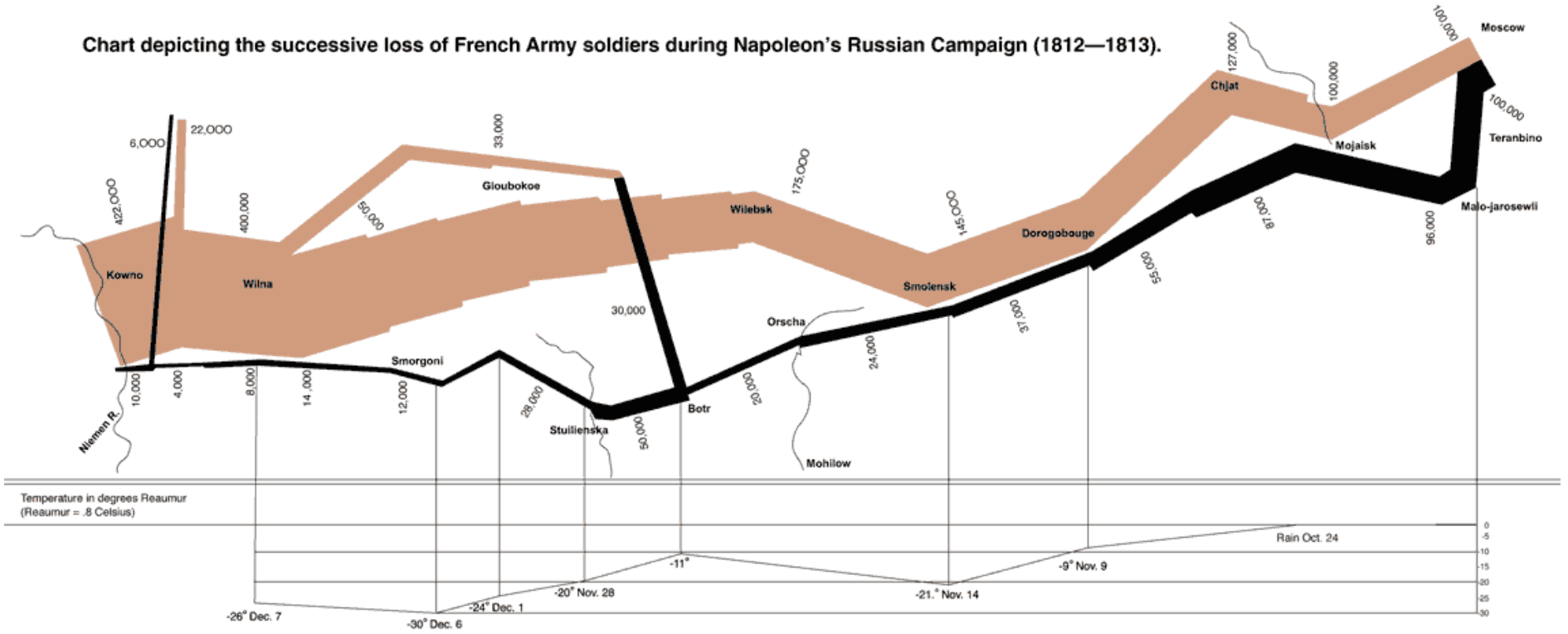
Complementarity between Data and Model

- (S, S)
- (B, B)
- In general, increase model size with data size
- Bias vs. variance trade-off

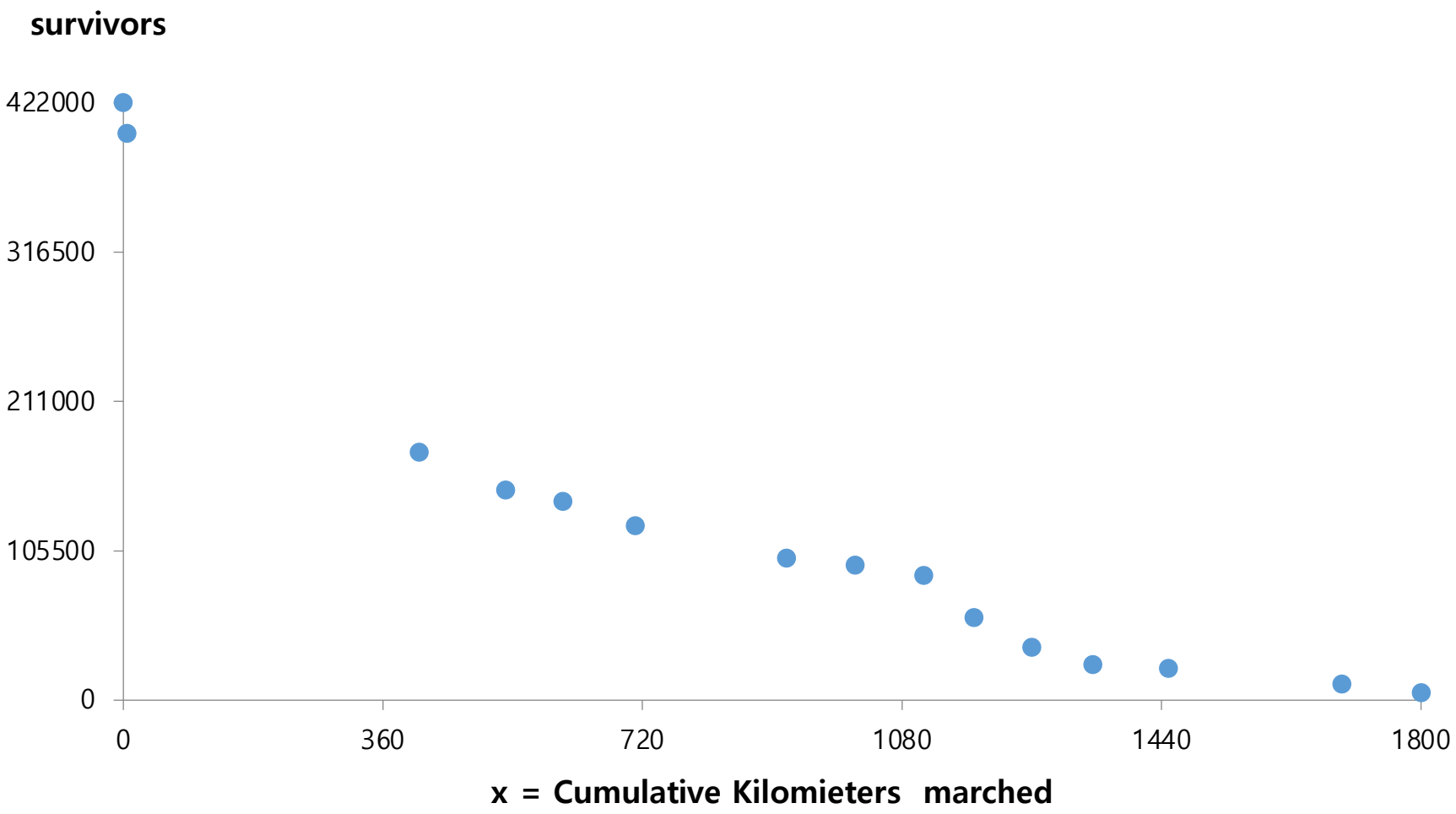
경험적 연구와 심슨의 역설



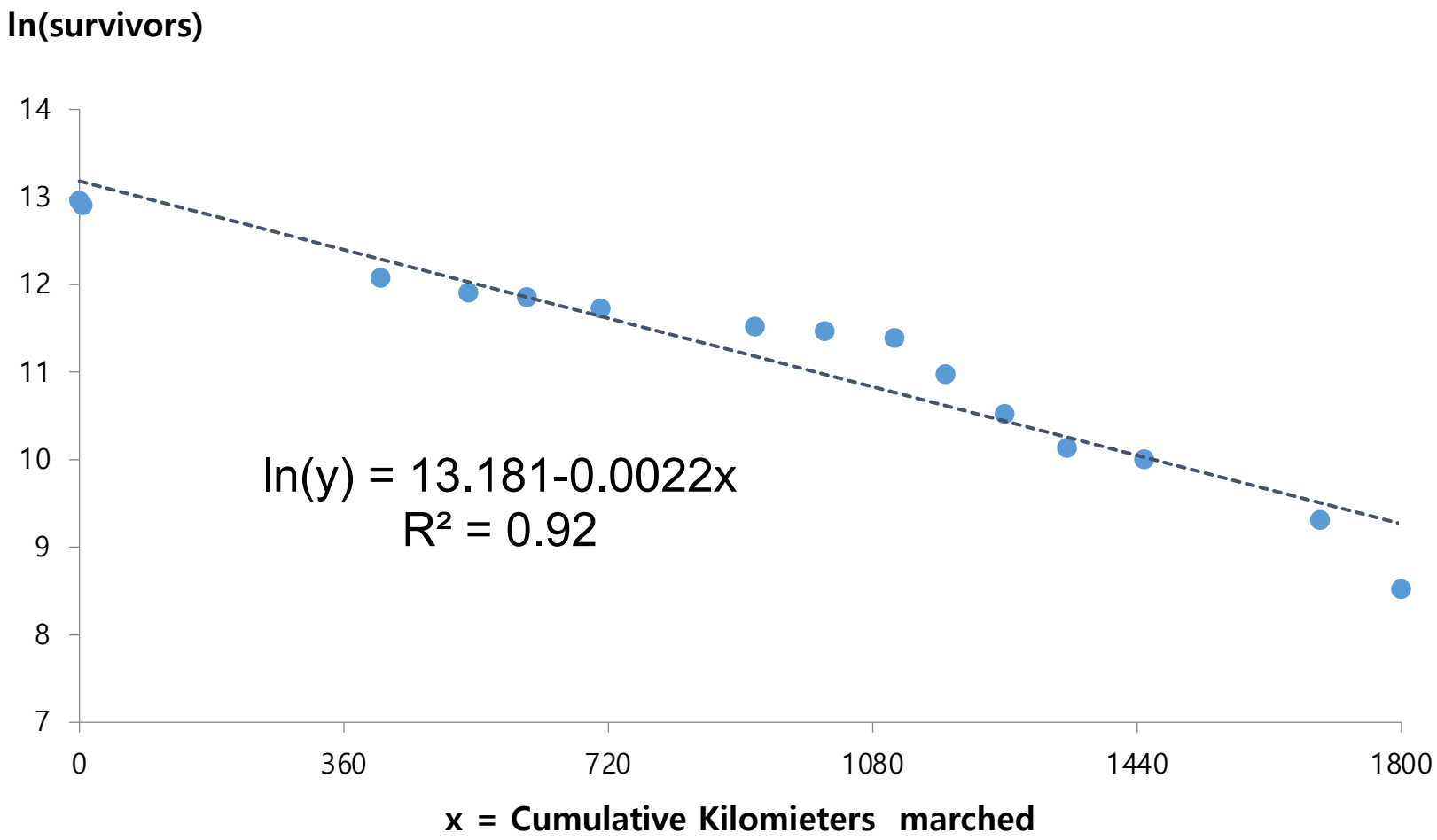
Napoleon Army's Russian Invasion in 1812



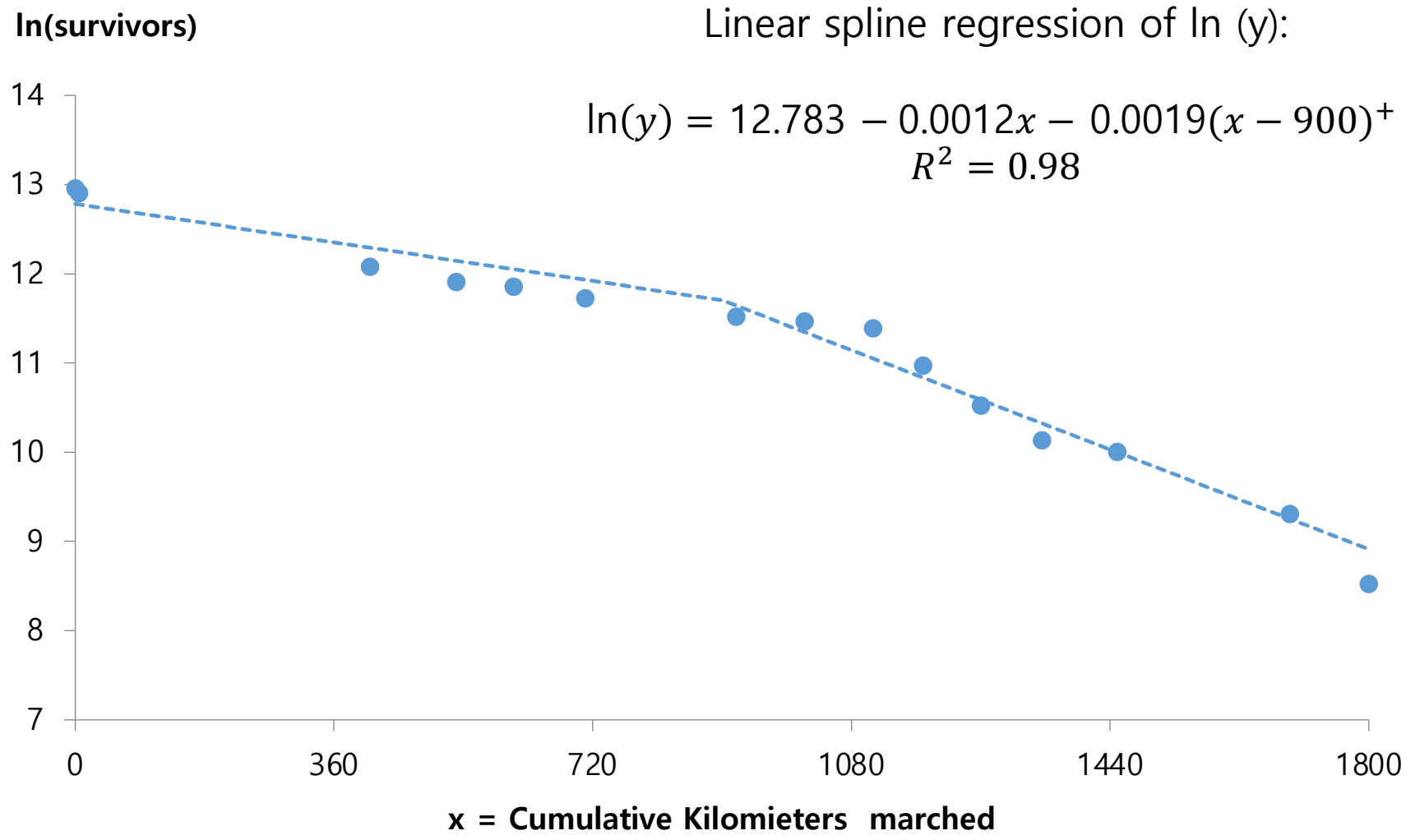
Napoleon Army's Russian Invasion in 1812



Napoleon Army's Russian Invasion in 1812



Napoleon Army's Russian Invasion in 1812



중회귀분석: 결혼시장 분석

- 국내 모 결혼정보회사의 상세한 개인 프로필 및 선택에 대한 현시 선호 데이터를 사용하여, 우리 나라 중매결혼시장에서 남녀의 배우자 선호의 차이를 비교함
- **사회 경제적인 조건과 외모 조건에 대한 선호에 있어서 남녀의 차이가 어떻게 드러나는가?**

중회귀분석: 결혼시장 분석

- 사회경제적 위세 지수 (SESI: Socio Economic Status Index)

  학력, 학벌, 직업, 소득 등을 포괄하는 지수

- 신체적 매력 지수(PAI: Physical Attractiveness Index)

  키, 체중, 인상등급 등을 포괄하는 지수

- 가정환경 지수 (FBI: Family Background Index)

  부의 학력, 직업, 재산, 양친 생존여부, 부모 이혼여부, 형제관계 등 포괄하는 지수

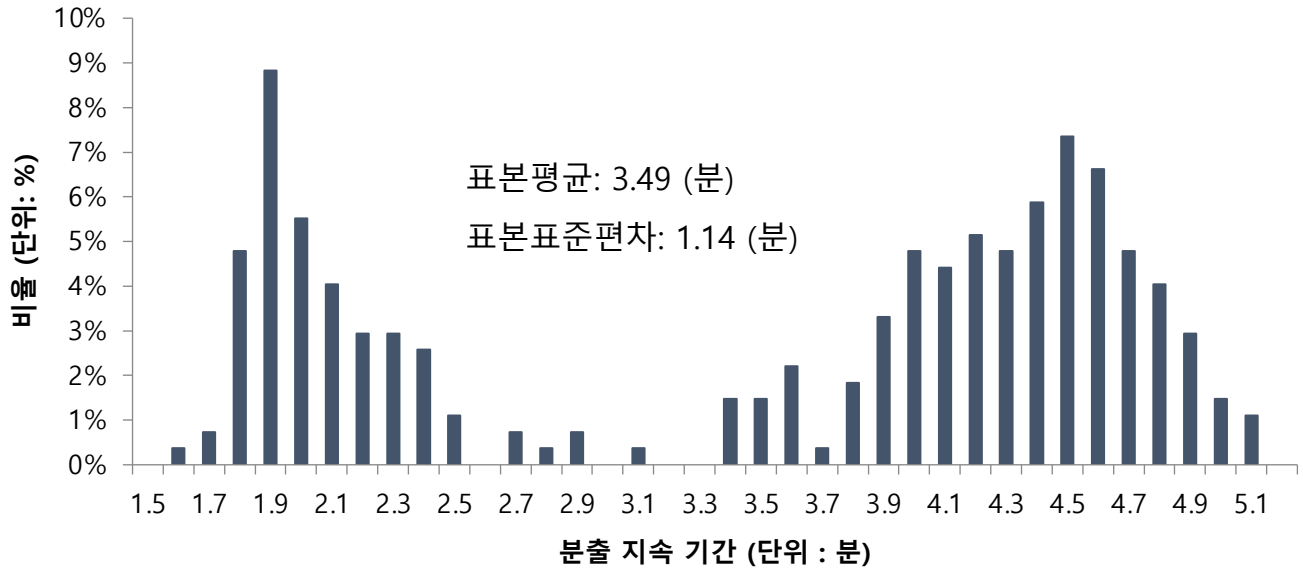
중회귀분석: 결혼시장 분석

- (반응) = α + β_1(상대의SESI) + β_2(상대의PAI) + β_3 (상대의FBI) + ε
- 반응: 좋다=1, 싫다=0

	남자의 반응		여자의 반응	
	추정치	표준오차	추정치	표준오차
SESI/100	1.23	1.30	1.91*	0.14
PAI/100	3.19*	0.34	1.18*	0.27
FBI/100	1.39*	0.53	0.29	0.87

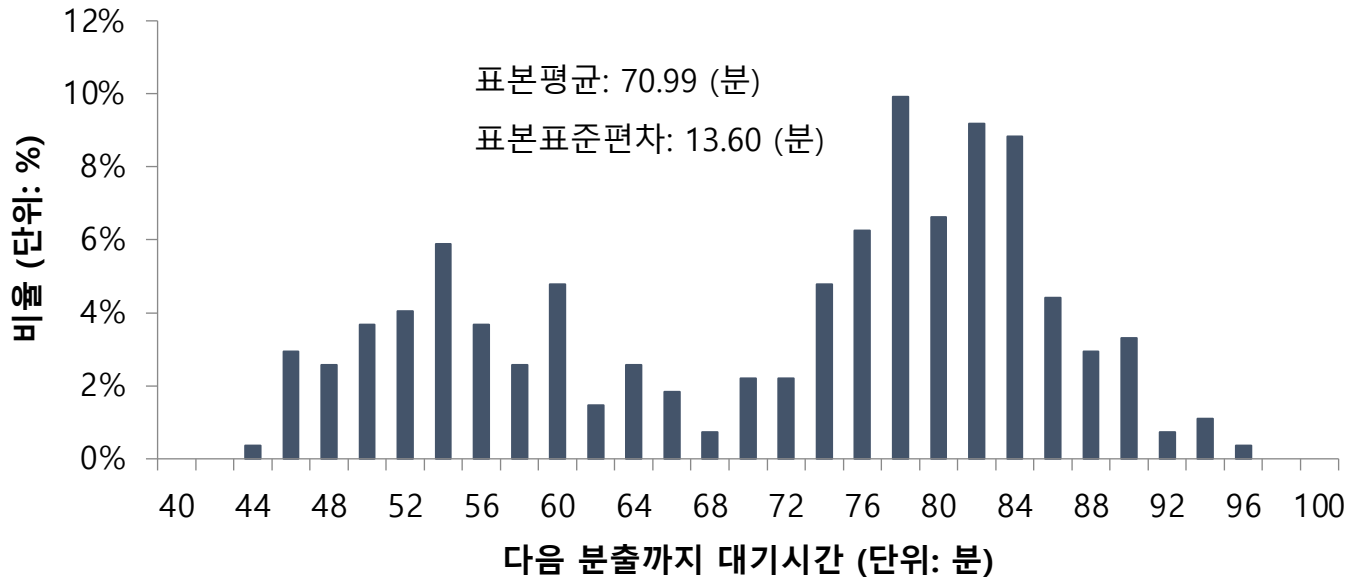
미국 Yellowstone 국립공원의 간헐천 Old Faithful

- 미국 Yellowstone 국립공원 내 간헐천 (Geyser)의 분출 지속기간(x) 분포
- 분출 지속기간의 히스토그램 : 3.2분 기준, 두 개 봉우리 갖는 쌍봉 분포



Old Faithful: One sample analysis (집단 구분 무시)

- 미국 Yellowstone 국립공원 내 간헐천 (Geyser)의 분출 대기시간 (y) 분포
- 분출 대기시간의 히스토그램 : 70분 기준, 두 개 봉우리 갖는 쌍봉 분포
- 쌍봉분포라는 사실 무시하고 단일의 정규분포로 잘못 근사하면 대기시간의 95% 예측구간은 $70.99 \pm 1.96 \times 13.60 = (44.33, 97.65)$. 무용지물의 구간임!

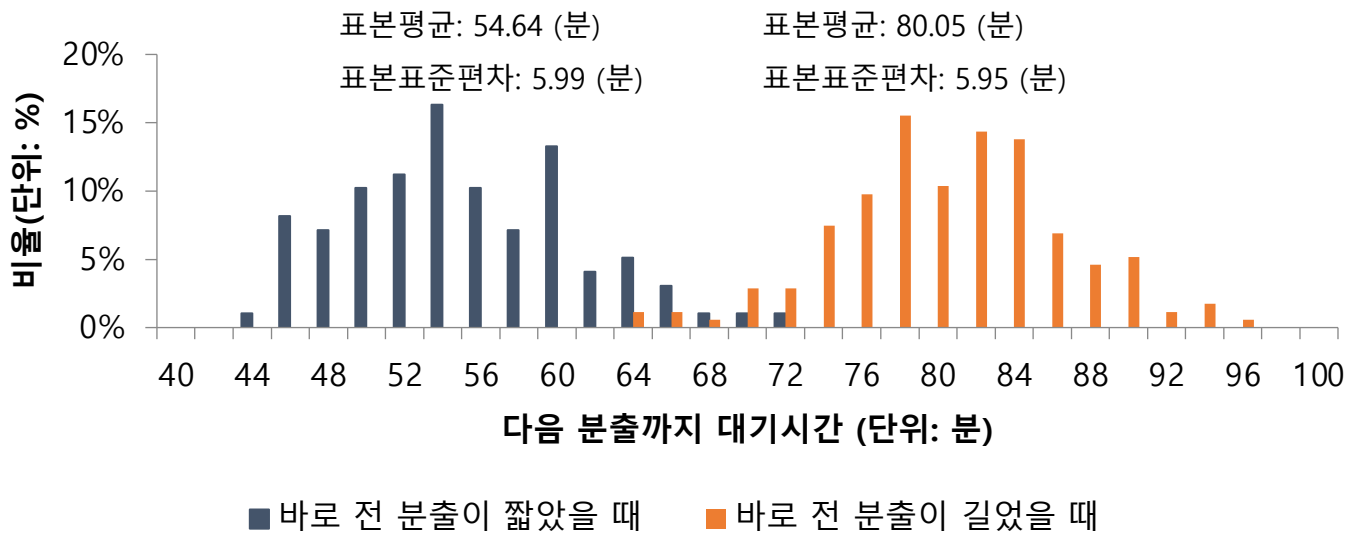


Old Faithful: Two sample analysis (집단 양분)

- 직전의 분출지속기간(x)이 길고 짧았는지에 따라 대기시간 (y) 자료를 양분
  - 직전 분출이 짧았을 때(x<3.2) 개별 y값의 95% 예측구간

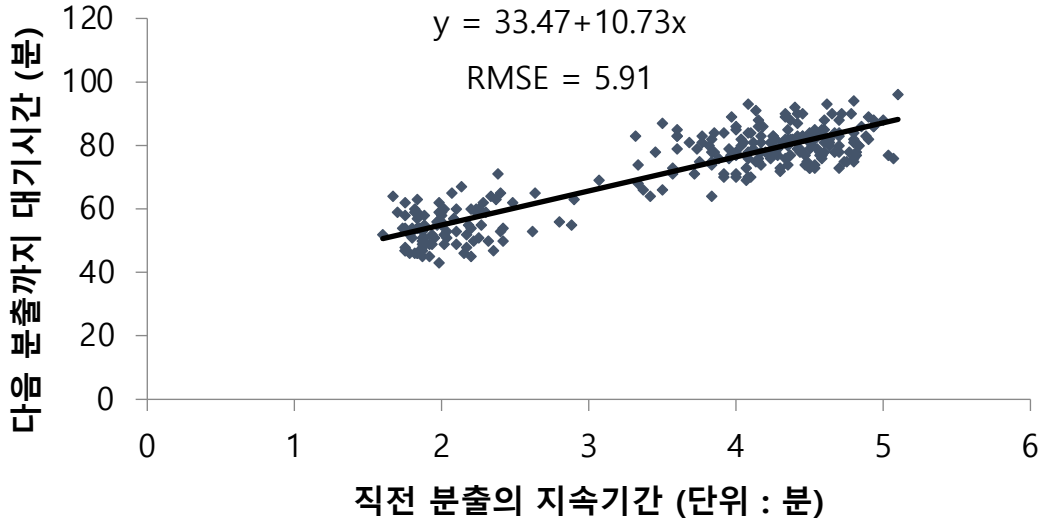
    54.64 ± 1.96 ×5.99=(42.90, 66.38)
  - 직전 분출이 길었을 때(x>3.2) 개별 y값의 95% 예측구간

    80.05 ± 1.96 ×5.95=(68.39, 91.71)



Old Faithful: Regression analysis (집단 별 분석)

- 다음 분출까지의 대기시간(y)을 직전 분출의 지속기간(x)에 회귀분석
  - 개별 y값에 대한 95% 예측구간은 33.47+10.73x ± 1.96 ×5.91
    - (43.35, 66.51) for x=2
    - (64.81, 87.97) for x=4

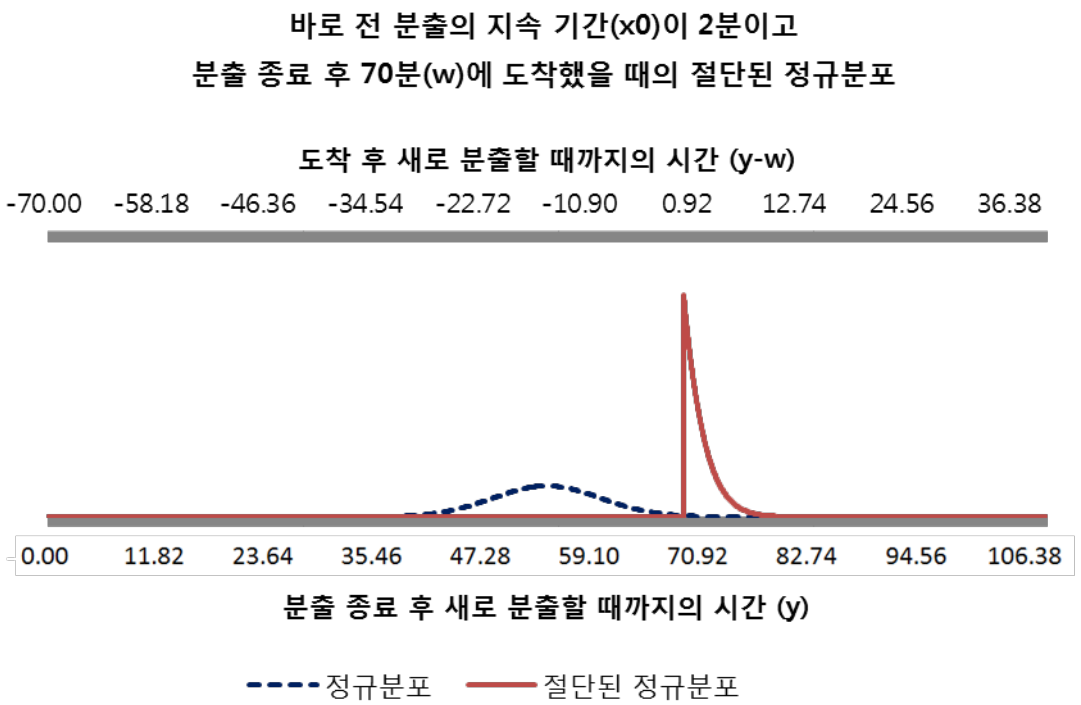


Old Faithful: Regression analysis, Real Time Updating

- 직전 분출 종료 후 70분만큼 경과한 경우

  x = 2 로 주어진 경우 y는 평균이 54.93이고 표준편차가 5.91이므로 이제껏

  70분 만큼 경과했다는 조건은 y의 분포를 의미있게 truncate시켜 업데이트함



총 대기시간: (70 , 75.38)

남은 대기시간: (0 , 5.38)

Big Data enables Bottom Up Approach

- Voice recognition in Personal Assistant such as Google Home

- Traditional: Small Data, Top Down
- Now in practice: Big Data, Bottom Up
- Same model for any language
- Newly born baby vs. DNN

Big Data

- Web data, e-commerce
- Purchases at department/grocery stores
- Bank/Credit Card transactions
- Social Network
- Telematics
- Wearables

Usage of Big Data

- AlphaGo
- FinTech
- MyData
- Autonomous driving
- Recommendation system
  - Amazon
  - NetFlix
  - Google

Big Data and Share Economy

- “Share rather than own.”
- Environmentally friendly
- Reduce waste
- Increasing utilization rate
- Better match demand and supply by ICT and Big Data

Python

- High-level programming language: Python
- Python supports TensorFlow: deep learning framework

X alone, or (X, Y)

- X=features, classifiers, characteristics, covariates
- Y=label, category, outcome
- Data only on X: unsupervised learning
- Data (X, Y): supervised learning
- Objective oriented inference: reinforcement learning
  - AlphaGo

Unsupervised Learning: X alone

- Clustering
- K-means algorithm: Repeat until convergence
  - computing centroid for each cluster
  - and assigning cluster membership
- Choose k*=point of diminishing returns

Unsupervised Learning: X alone

- Hierarchical agglomerative clustering:
- Start with each point forming its own cluster
- Repeatedly merge the clusters of the closest two points
- Represent the results using dendrogram

Similarity/dissimilarity among documents

- “bag of words” => generate X vector
- TFIDF (term frequency*inverse document frequency)
  - frequent locally (in the document)
  - rare globally (in the universe of entire documents)

The federalist problem

- Alexander Hamilton vs. James Madison
- by Frederick Mosteller (Harvard math. statistician)

Words as discriminators

- Madison words: by, also, …
- Hamilton words: to, upon, …

Model for word frequency

- “Word frequency” per 1,000 words~Poisson or negative binomial
- For each (author, word), parameters are taken from those documents whose authors are identified as either Madison or Hamilton

naïve Bayes

- Likelihood for frequencies of {word 1, word 2, …} per 1,000 words using naïve Bayes (imposing independence across words) for each author
- Posterior odds in favor of Madison for all 12 disputed papers

Supervised Learning: (X, Y)

- labeled images (cats, dogs, …)
- labeled documents (sports, news, …)

Supervised Learning: Deep Learning

- XOR problem

Deep Learning: X and Y

- Image Recognition:
- X (in ray scale)=28 by 28 matrix of pixel intensities
  X (in color) =3 of 28 by 28 (28 by 28 for each of R, G, B)
- Y=y1 (dog or not), y2 (cat or not), y3, …

Deep Learning: X and Y

- Autonomous driving=real time mapping from X to y
  - X=T of 3 by 64 by 64 tensors (time series of images), …
  - Y=wheeling angle, acceleration, deceleration (breaking), …

Applications: Credit card usage data

- Link with deposit, loan data
- Construct real time business cycle indicator
- Real-time credit rating update. Improve default prediction

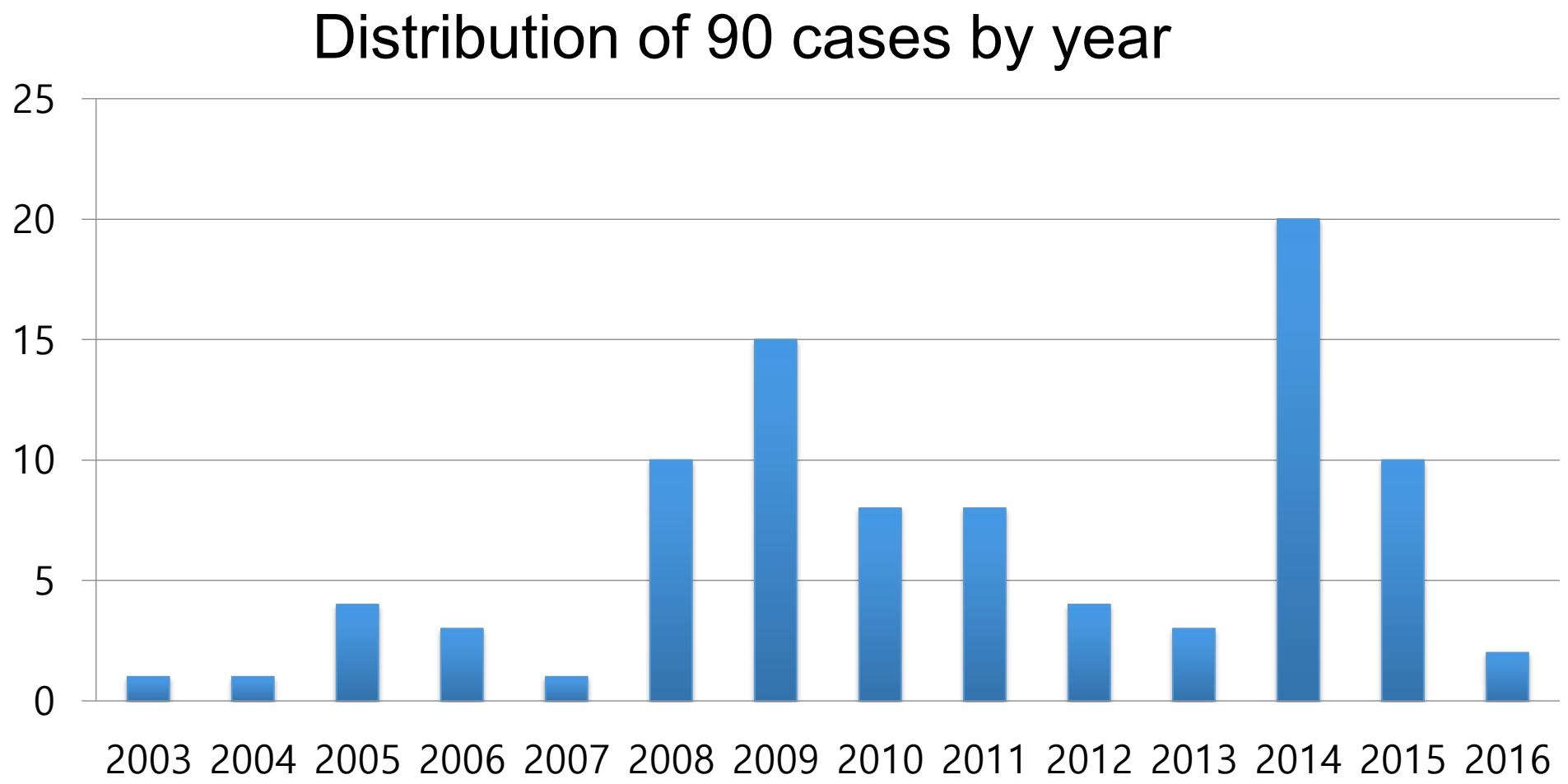
Revealed preference of judges

- Political orientation of constitutional court judges in Korea (by S. B. Kim, SNU)
- *c.f.* Clinton, Jackman, and Rivers, 2004, "The Statistical Analysis of Roll Call Data," *American Political Science Review*

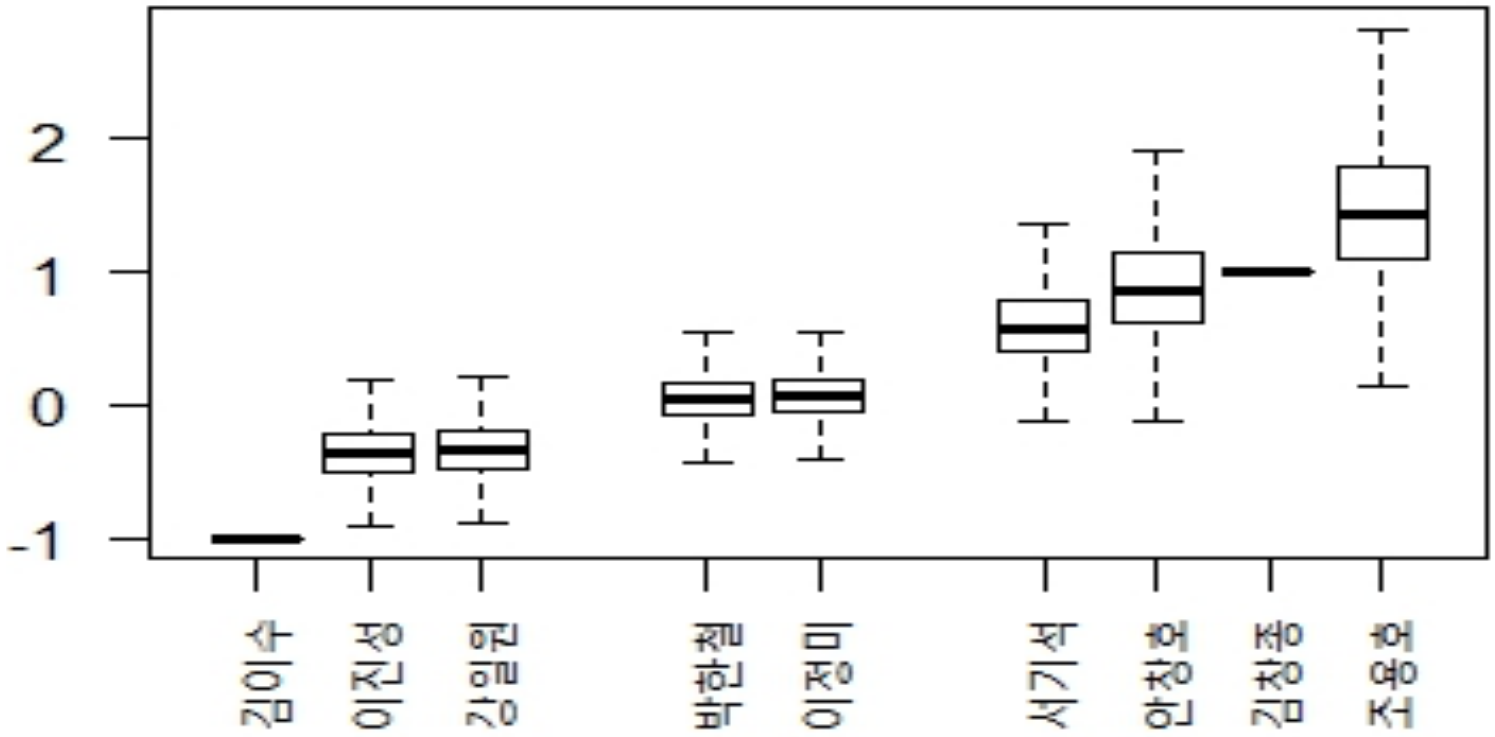
Data: 4,000 constitutional court cases over 2003~2016

-642 cases remaining after excluding rejection/unanimity

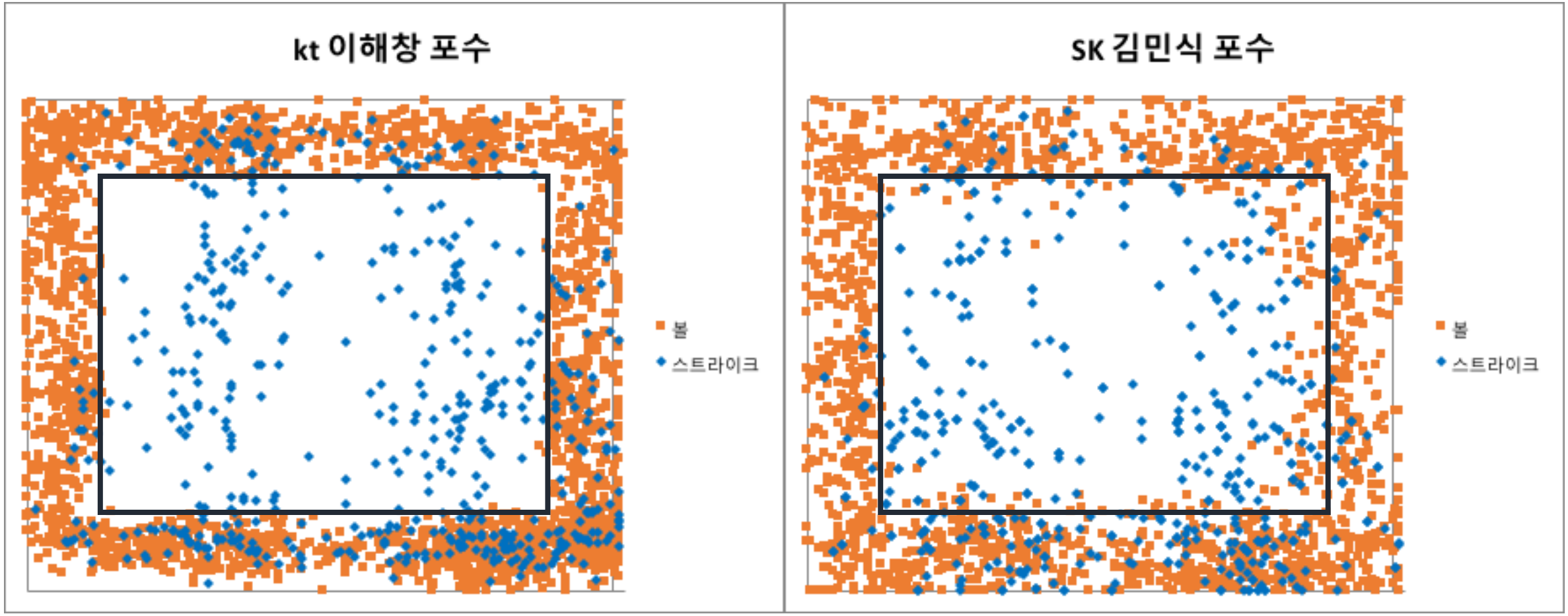
-90 cases remaining after further removing non-political cases



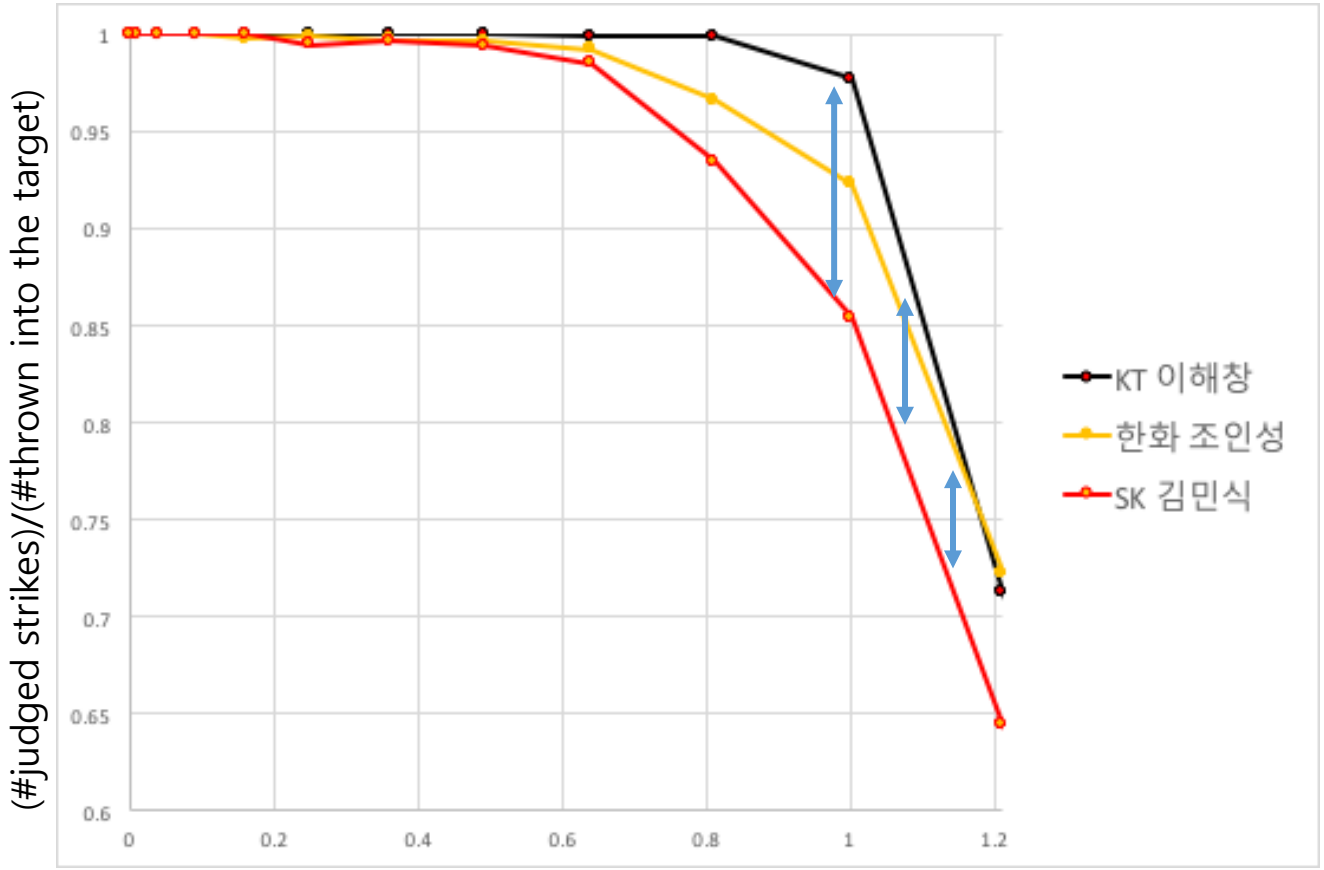
Estimated political orientation of 9 constitutional court judges



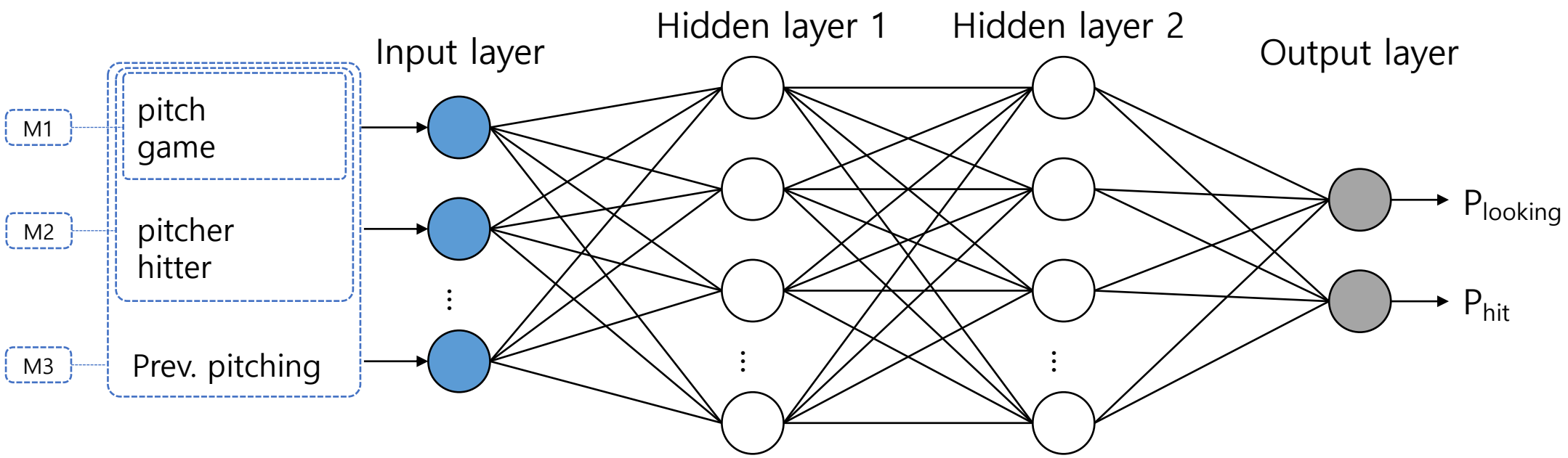
Baseball Case: Pitch by pitch data in KBL



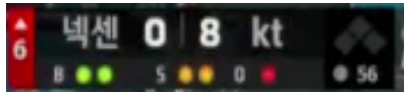
Pitch by pitch data in KBL



Pitch outcome prediction: Deep Learning
(Tensorflow, 0.7mil. pitch by pitch observations)



Pitch outcome prediction: Deep Learning



Kt (home) vs. Nexen (away) on May 27, 2016
Joo, Kwon of Kt pitching to Lim, Byungwook

Prediction from Deep Learning NN

Fast ball after fast ball

In-corner \ Out-corner				
0.070	0.149	0.197	0.152	0.073
0.082	0.441	0.524	0.449	0.235
0.039	0.531	0.642	0.432	0.249
0.014	0.481	0.535	0.275	0.231
0.002	0.229	0.159	0.162	0.036

Real pitch: fast ball and being hit

Slider after fast ball

In-corner \ Out-corner				
0.087	0.141	0.109	0.085	0.039
0.129	0.362	0.468	0.366	0.172
0.090	0.366	0.564	0.519	0.214
0.045	0.286	0.447	0.460	0.233
0.011	0.208	0.147	0.180	0.078

Slider after fast ball recommended!

Other Useful Tools

- Bayes rule
- Regularization
- Sample split
- Regression and Classification Tree, Forest
- Thompson sampling and recommendation system

Future of data science

- Genomics, Drug discovery:
  linking school, medical, military, family records using ind. ID
- Finance and marketing
- Real time business cycle indicators
- Behavioral finance/economics: interaction b/w info. and mood

Future of data science

- No returning
- In growing demand