# LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations

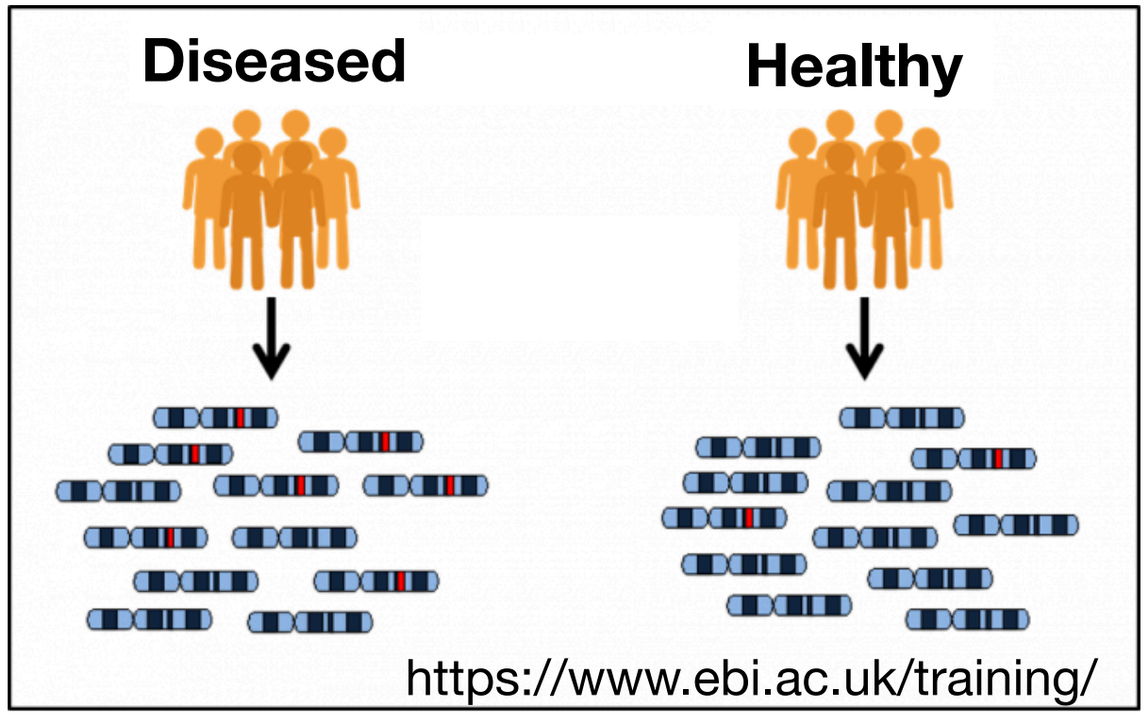
**Brian Trippe**, Jonathan Huggins, Raj Agrawal, and Tamara Broderick

# LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations

**Brian Trippe**, Jonathan Huggins, Raj Agrawal, and Tamara Broderick

**Genomic Study (motivating example)**

- **Goal:** Understand relationship between genomic variation & disease outcome
- N=20,000 samples — D=500,000 SNPs

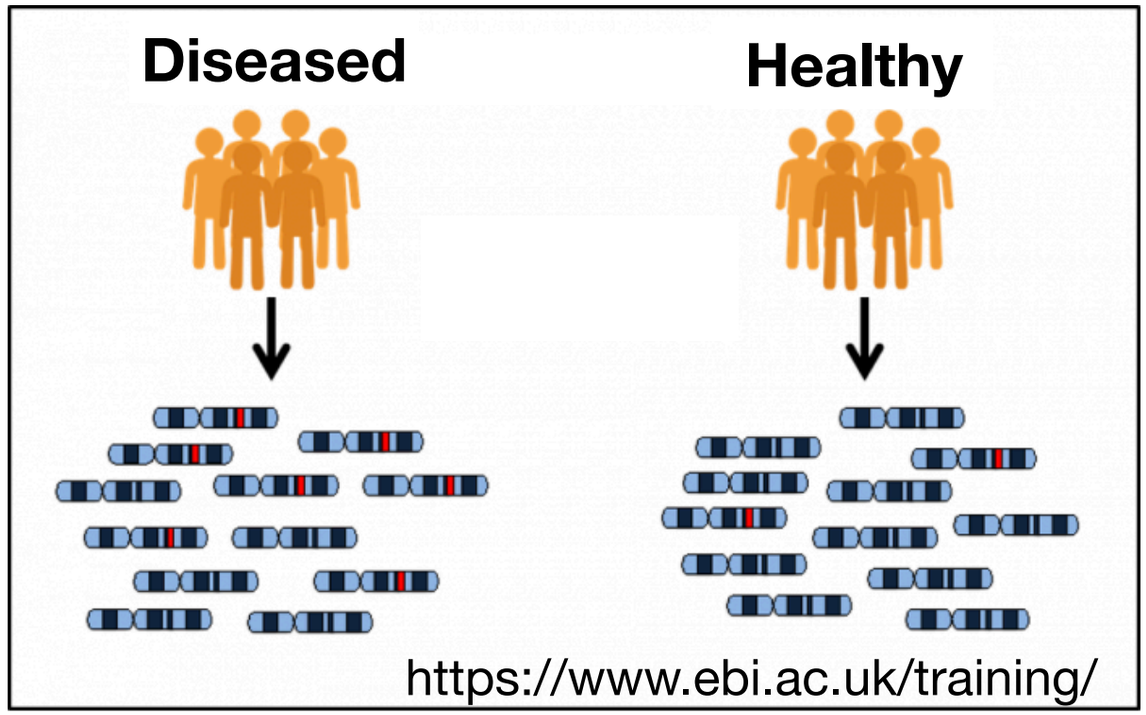


# LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations

**Brian Trippe**, Jonathan Huggins, Raj Agrawal, and Tamara Broderick

**Genomic Study (motivating example)**

- **Goal:** Understand relationship between genomic variation & disease outcome
- N=20,000 samples — D=500,000 SNPs




**Generalized Linear Models (GLMs)**

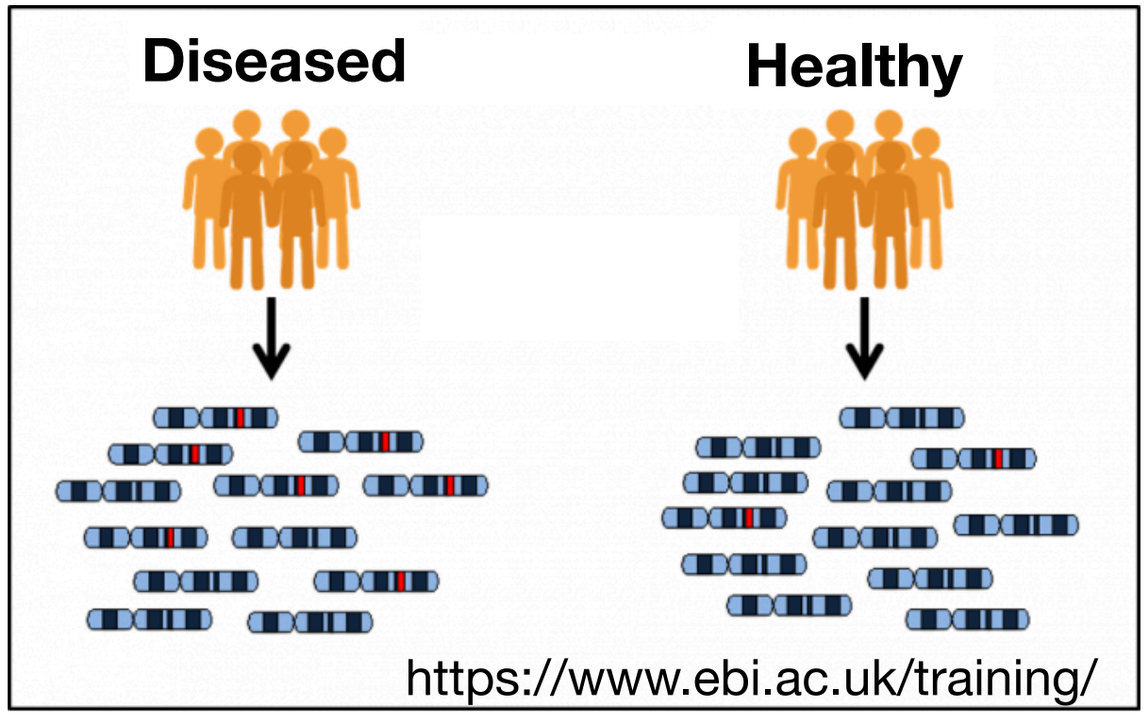
- Interpretability
- E.g. Logistic/Poisson/Negative Binomial Regression

# LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations

**Brian Trippe**, Jonathan Huggins, Raj Agrawal, and Tamara Broderick

**Genomic Study (motivating example)**

- **Goal:** Understand relationship between genomic variation & disease outcome
- N=20,000 samples — D=500,000 SNPs




**Generalized Linear Models (GLMs)**

- Interpretability
- E.g. Logistic/Poisson/Negative Binomial Regression

**Bayesian Modeling & Inference**

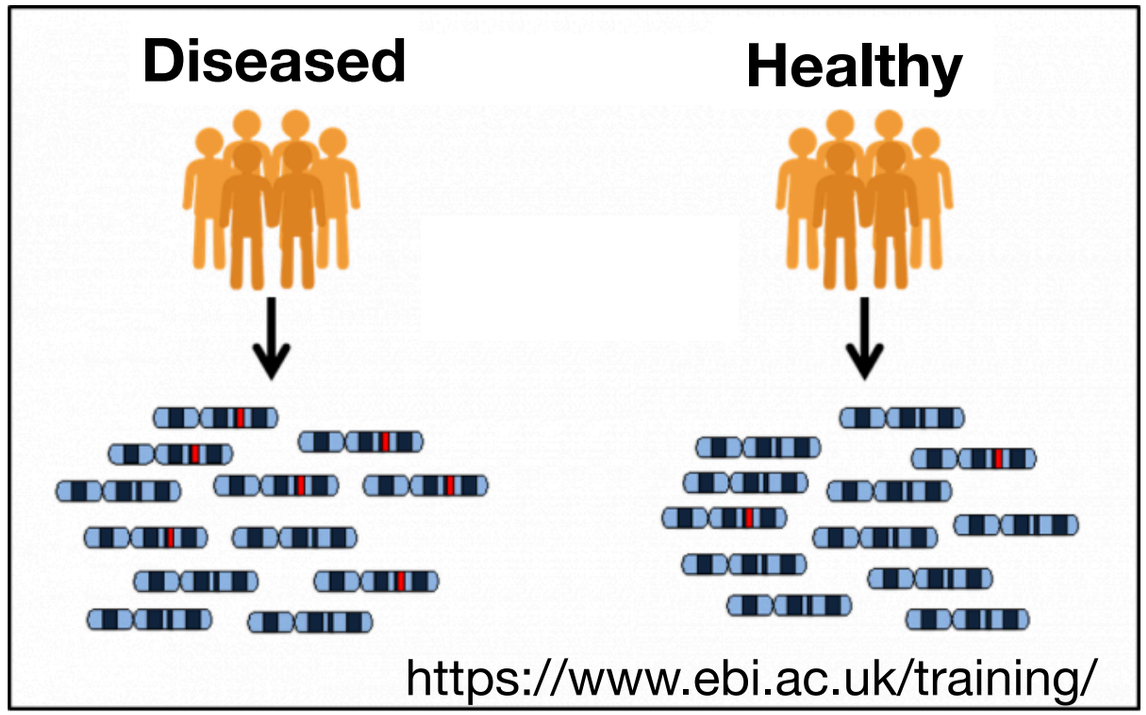
- Coherent uncertainty quantification

# LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations

**Brian Trippe**, Jonathan Huggins, Raj Agrawal, and Tamara Broderick

**Genomic Study (motivating example)**

- **Goal:** Understand relationship between genomic variation & disease outcome
- N=20,000 samples — D=500,000 SNPs




https://www.ebi.ac.uk/training/

**Generalized Linear Models (GLMs)**

- Interpretability
- E.g. Logistic/Poisson/Negative Binomial Regression

**Bayesian Modeling & Inference**

- Coherent uncertainty quantification

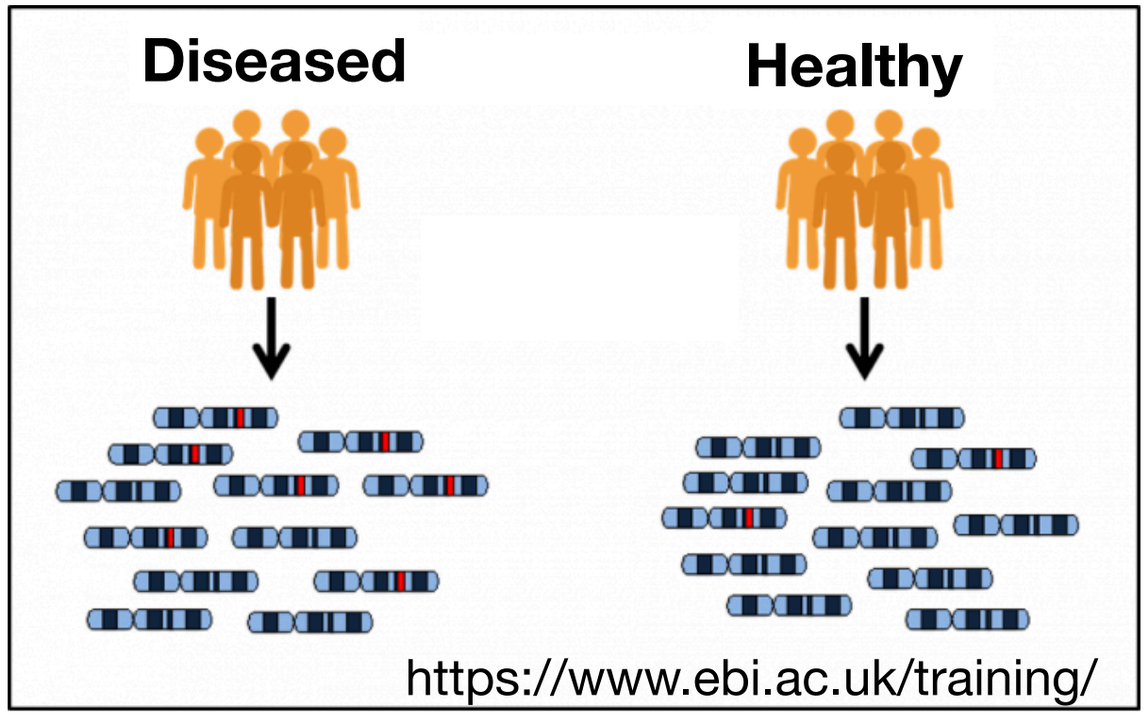
**Problem: Super-linear scaling with D**

# LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations

**Brian Trippe**, Jonathan Huggins, Raj Agrawal, and Tamara Broderick

**Genomic Study (motivating example)**

- **Goal:** Understand relationship between genomic variation & disease outcome
- N=20,000 samples — D=500,000 SNPs




**Generalized Linear Models (GLMs)**

- Interpretability
- E.g. Logistic/Poisson/Negative Binomial Regression

**Bayesian Modeling & Inference**

- Coherent uncertainty quantification

**Problem: Super-linear scaling with D**

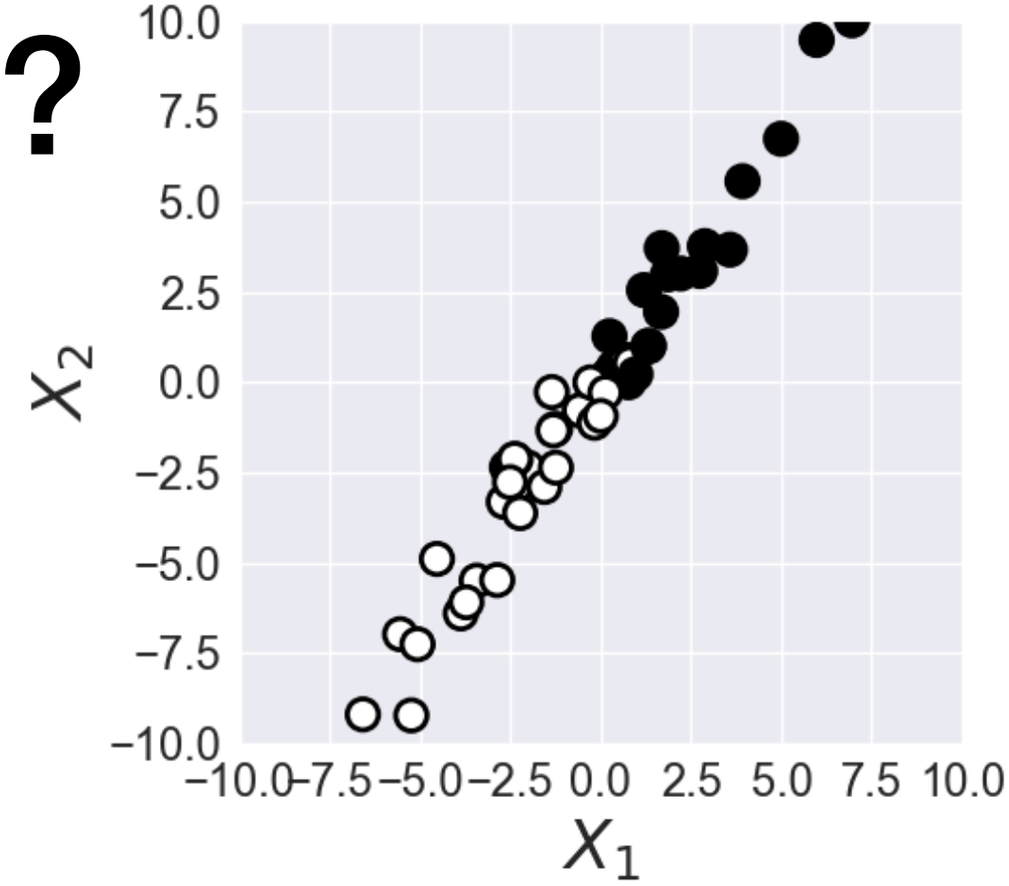
*We present **LR-GLM**, a method with linear scaling in D and theoretical guarantees on quality*

# How does it work?

# How does it work?

**Cartoon Example**

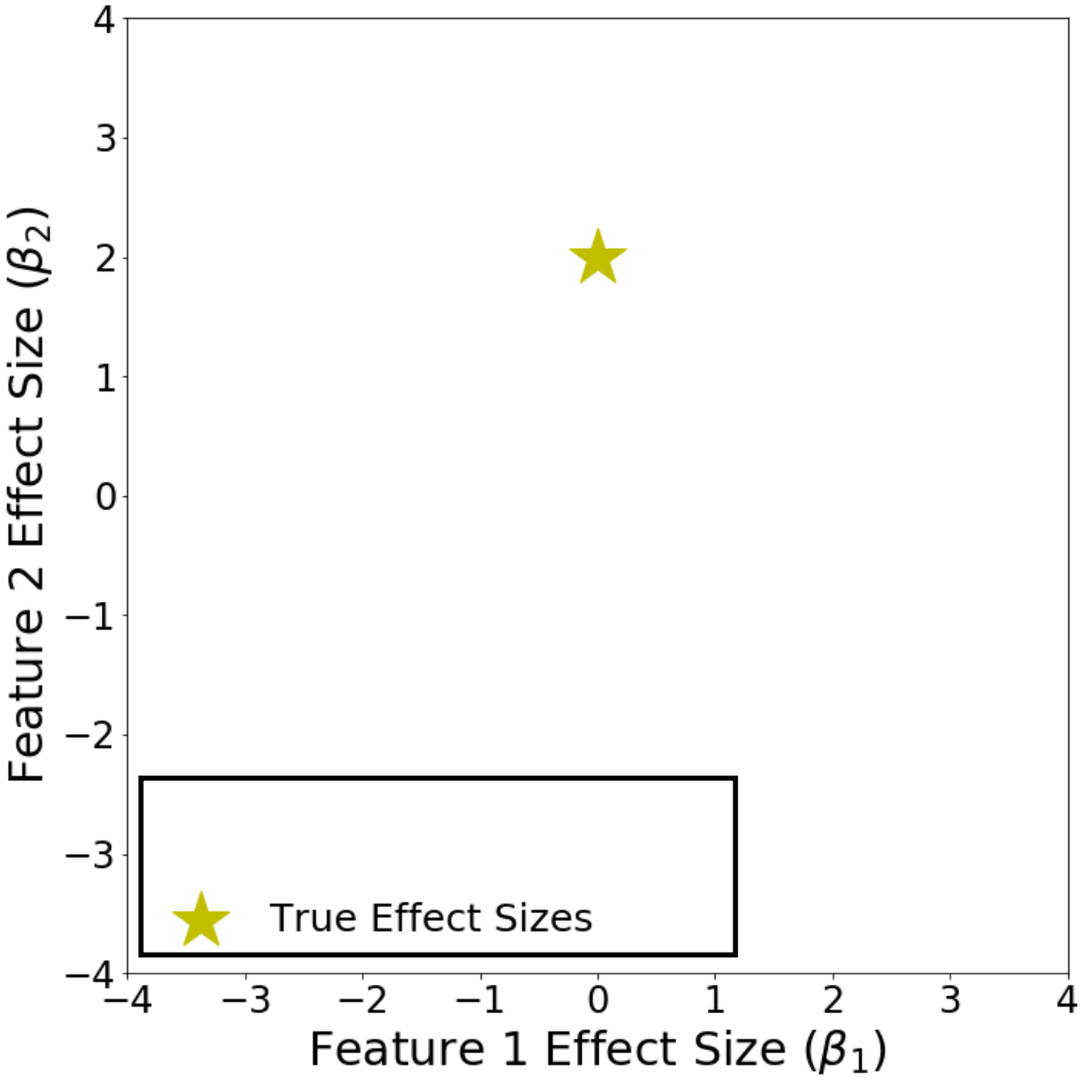
- Logistic Regression with two correlated features

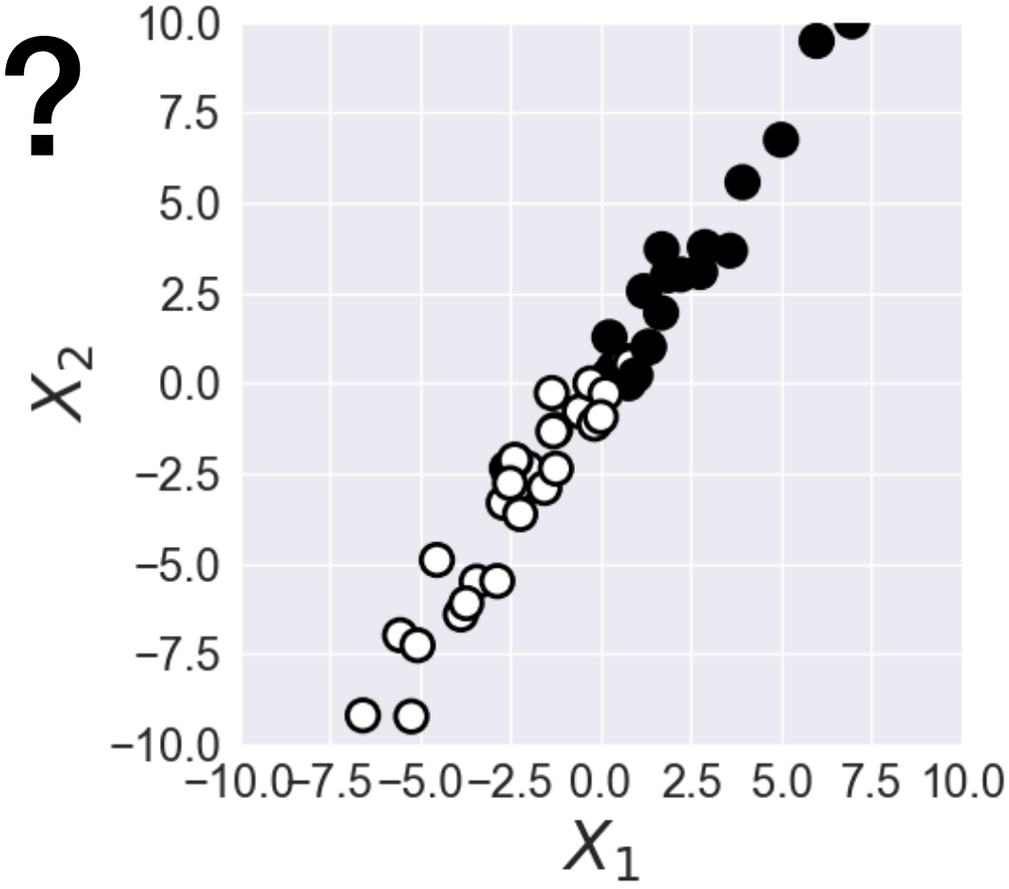


# How does it work?

**Cartoon Example**

- Logistic Regression with two correlated features

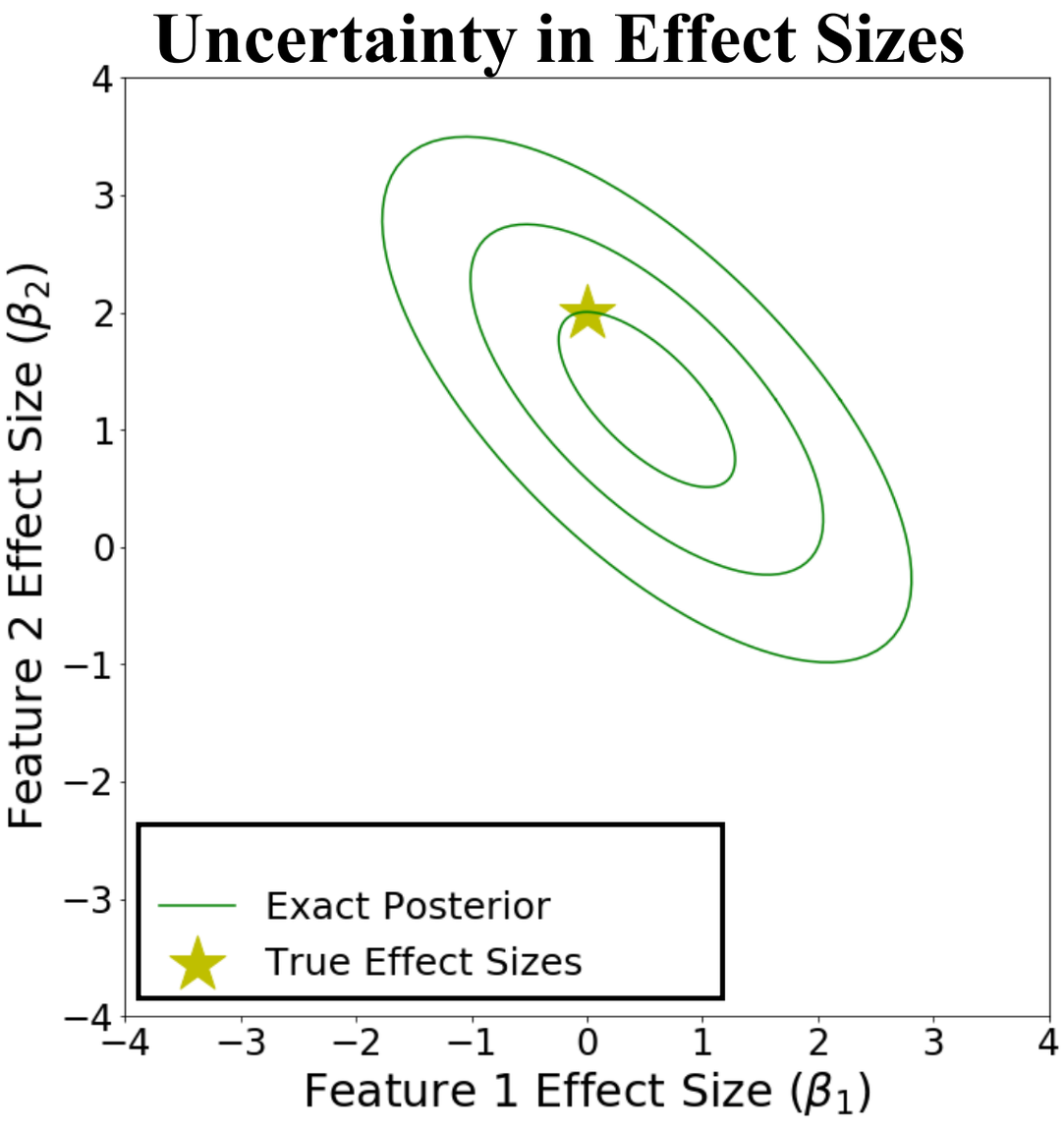


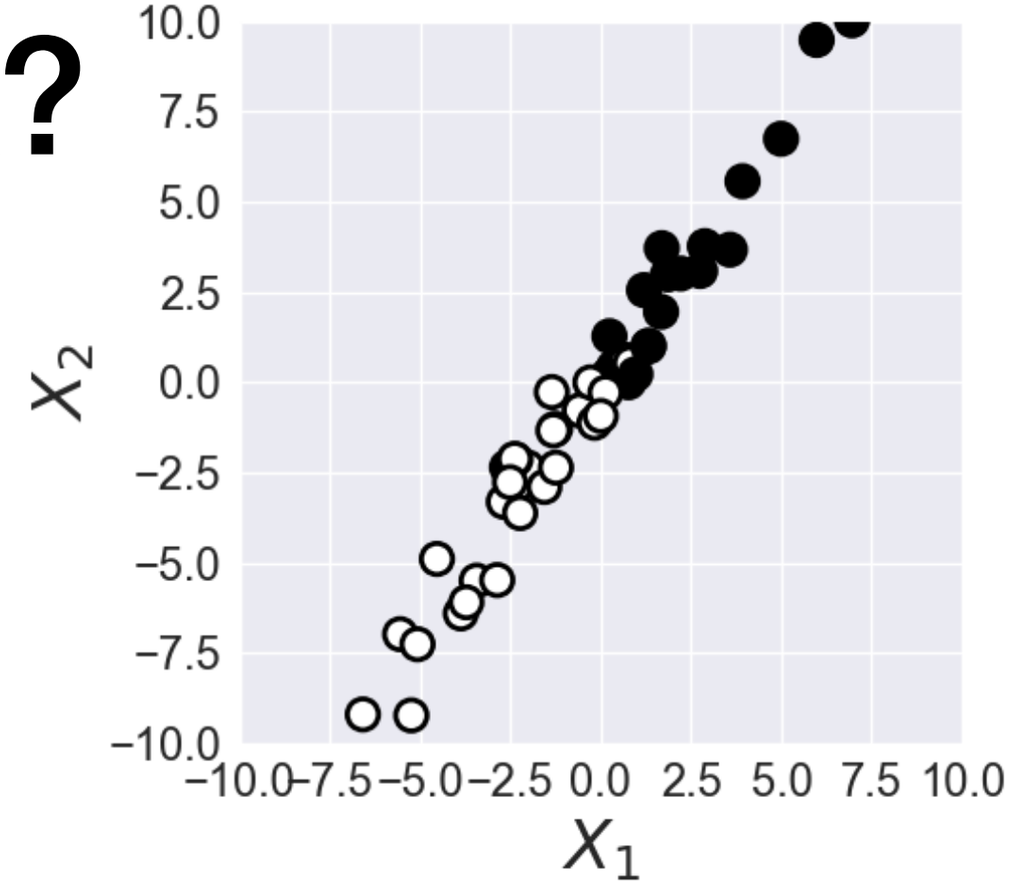



# How does it work?

**Cartoon Example**

- Logistic Regression with two correlated features

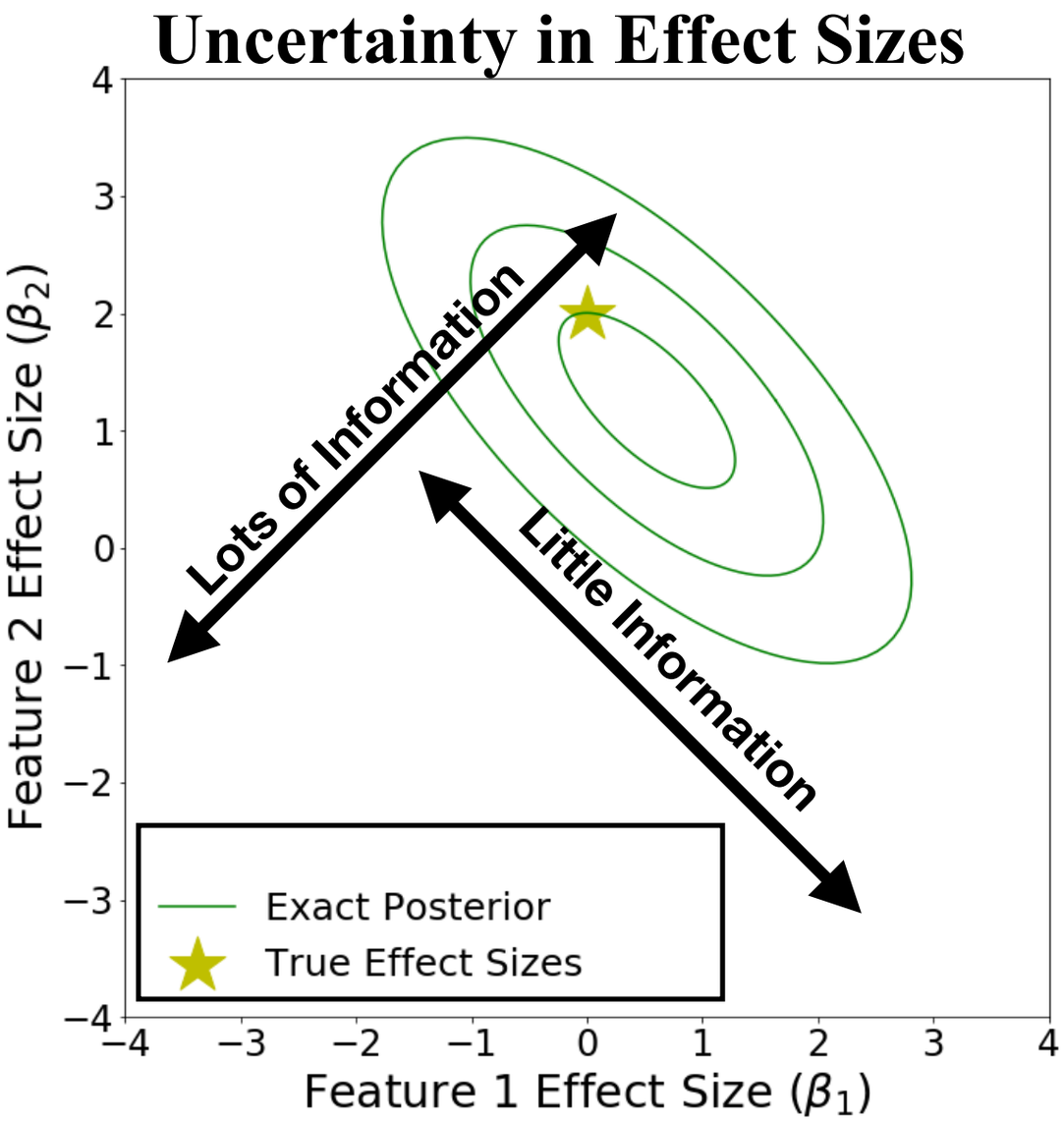


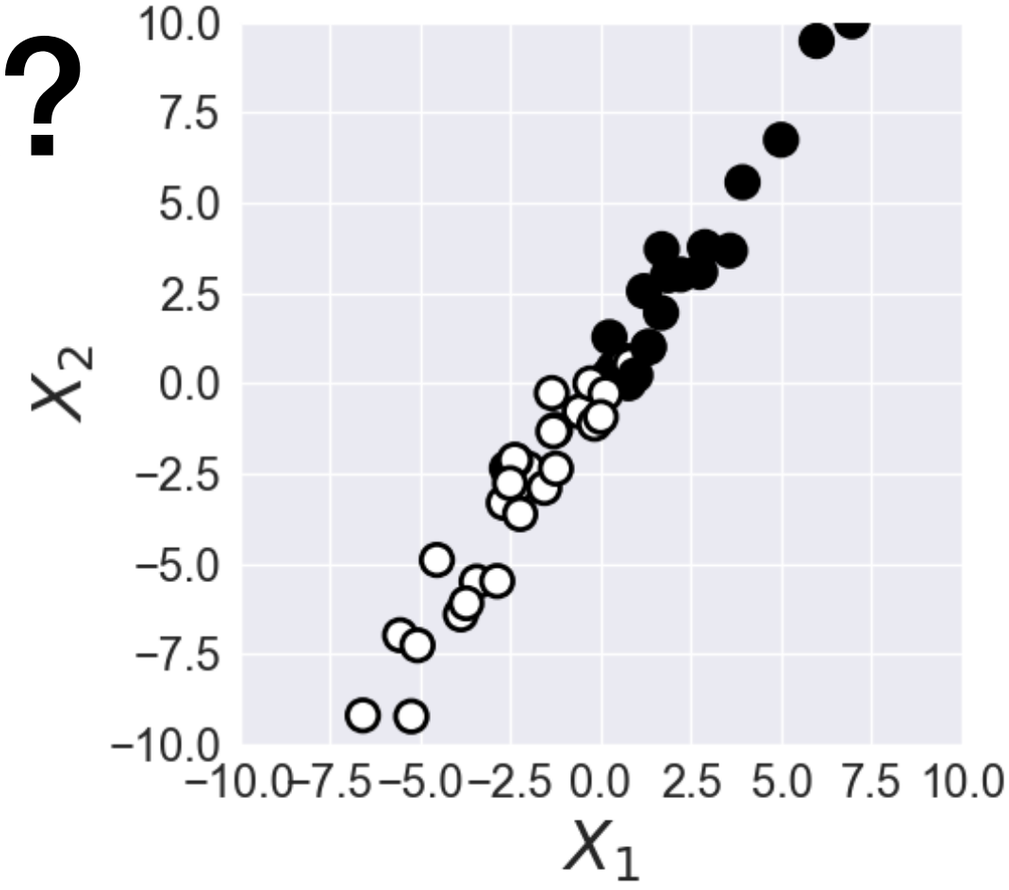



# How does it work?

**Cartoon Example**

- Logistic Regression with two correlated features

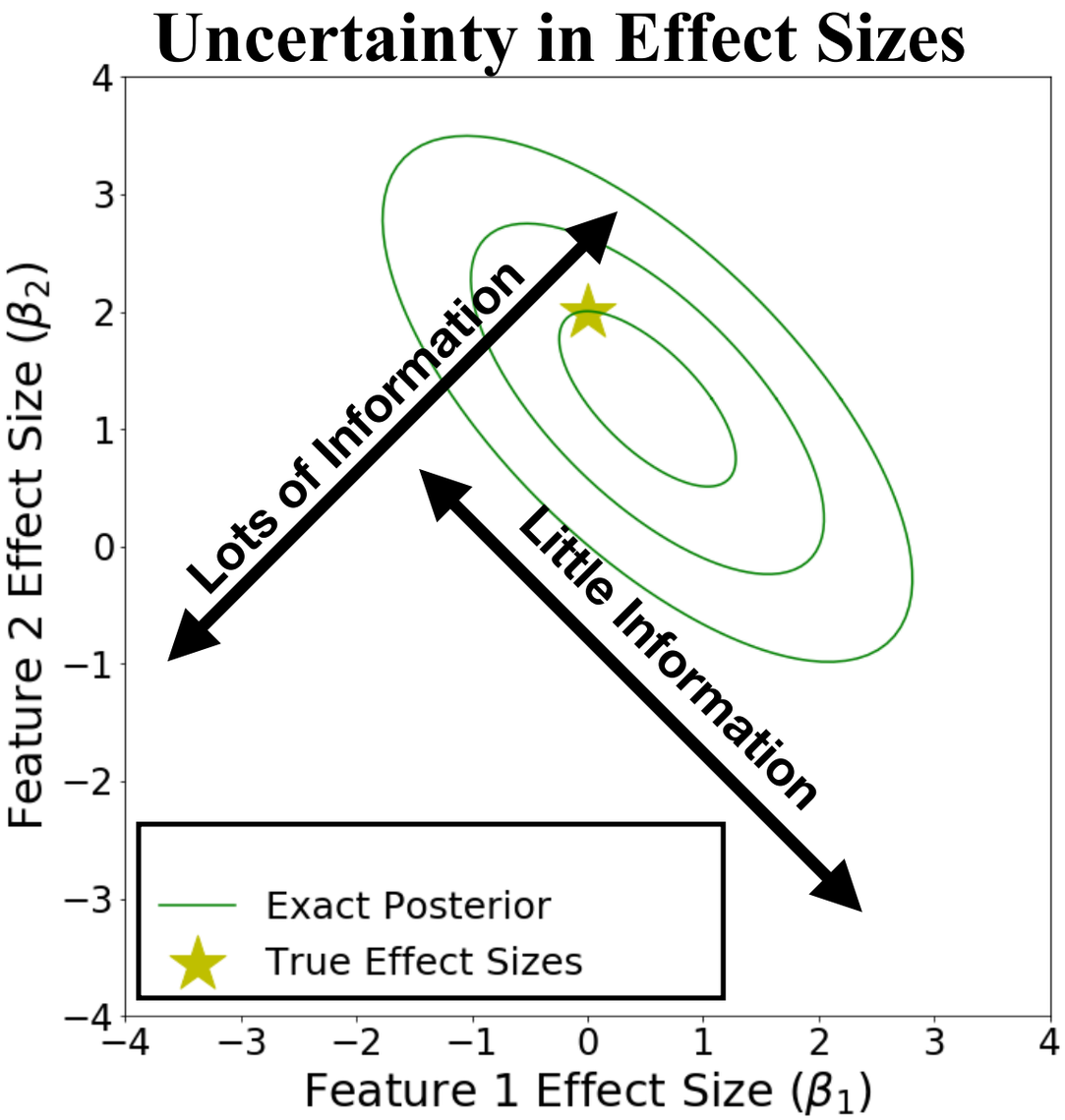


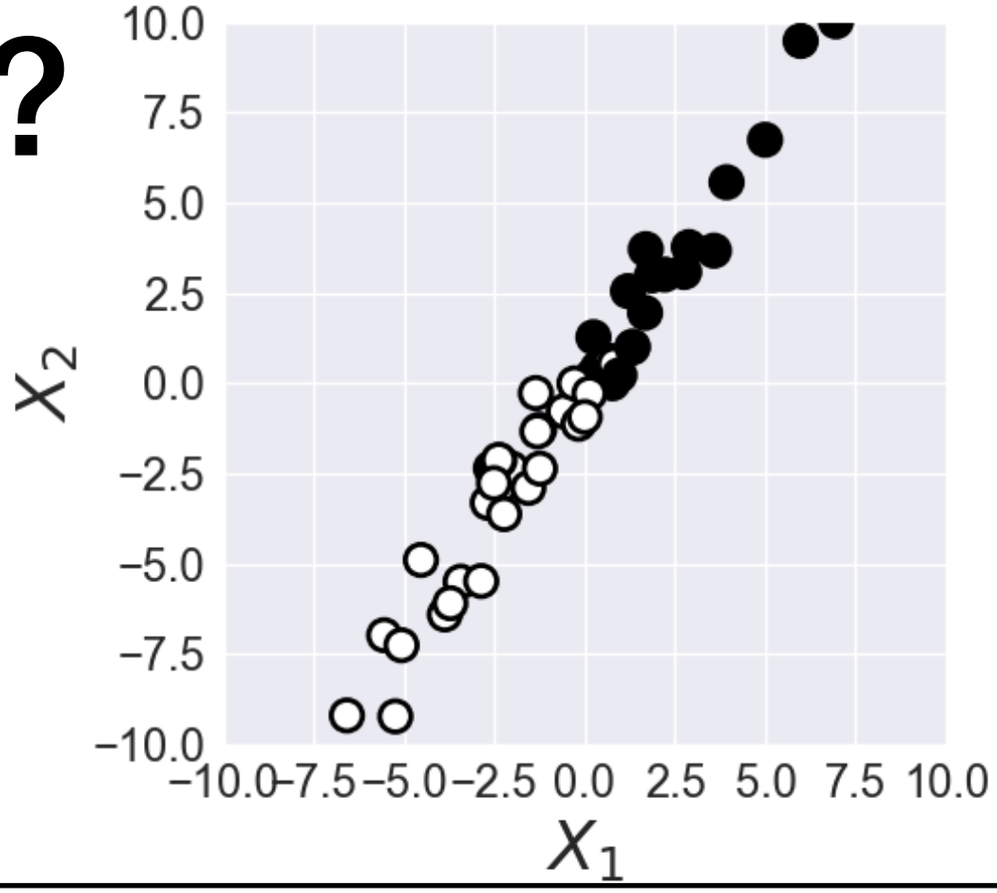



# How does it work?

**Cartoon Example**

- Logistic Regression with two correlated features




**The LR-GLM Approximation**

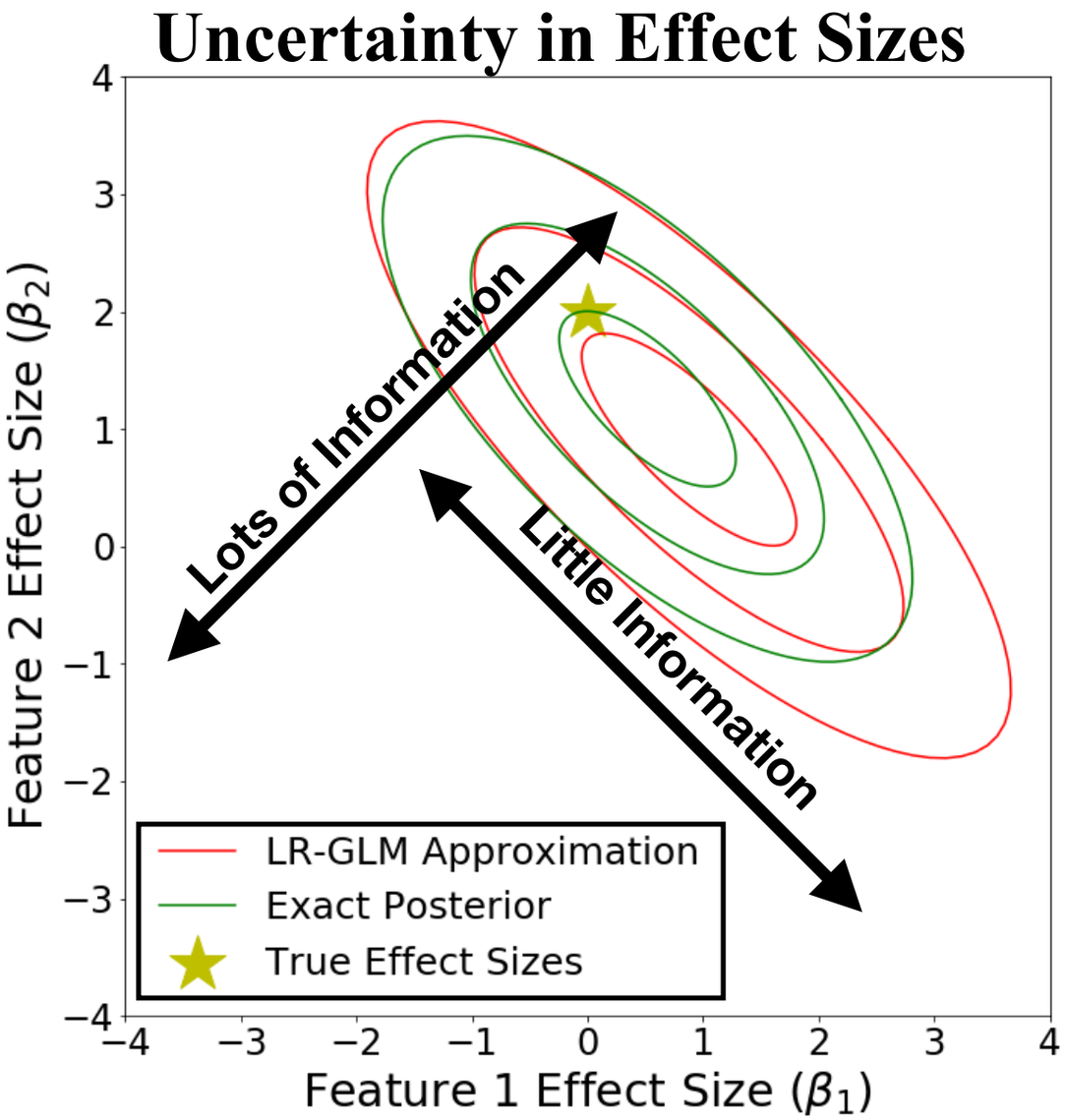
We ignore the least informative directions

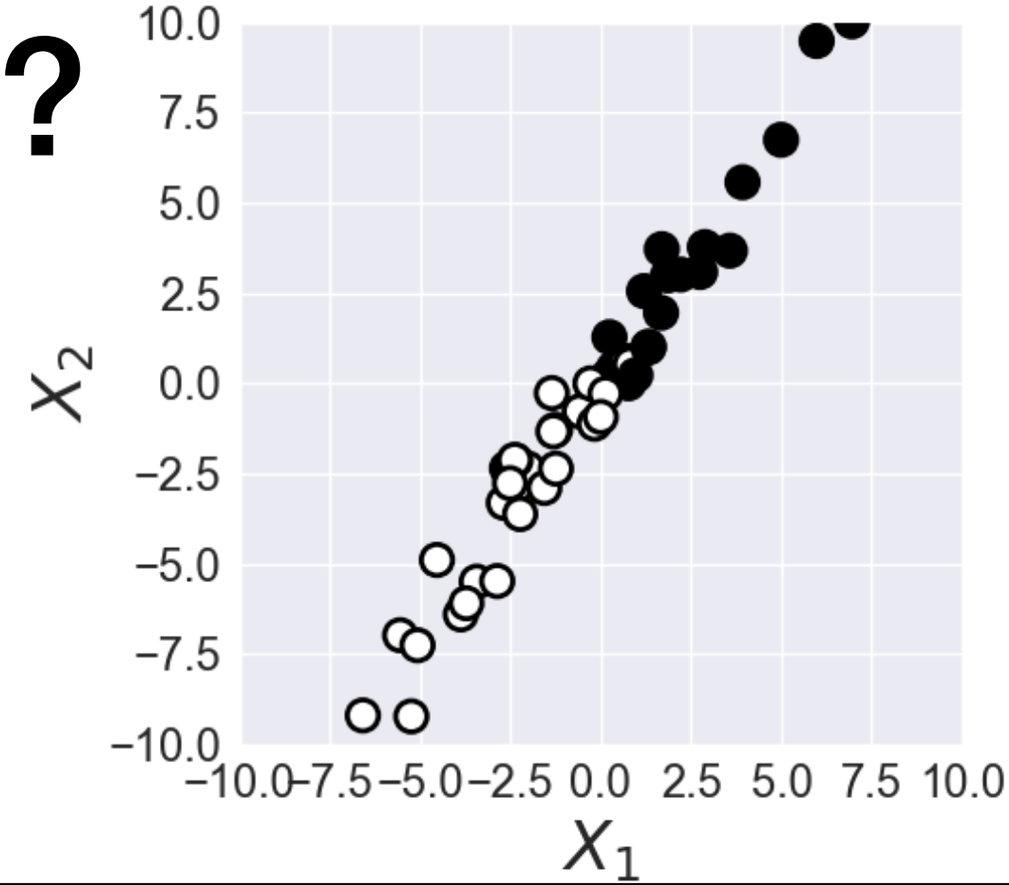
$$p(y_i|x_i^T\beta) \approx p(y_i|x_iUU^T\beta)$$

# How does it work?

**Cartoon Example**

- Logistic Regression with two correlated features




**The LR-GLM Approximation**

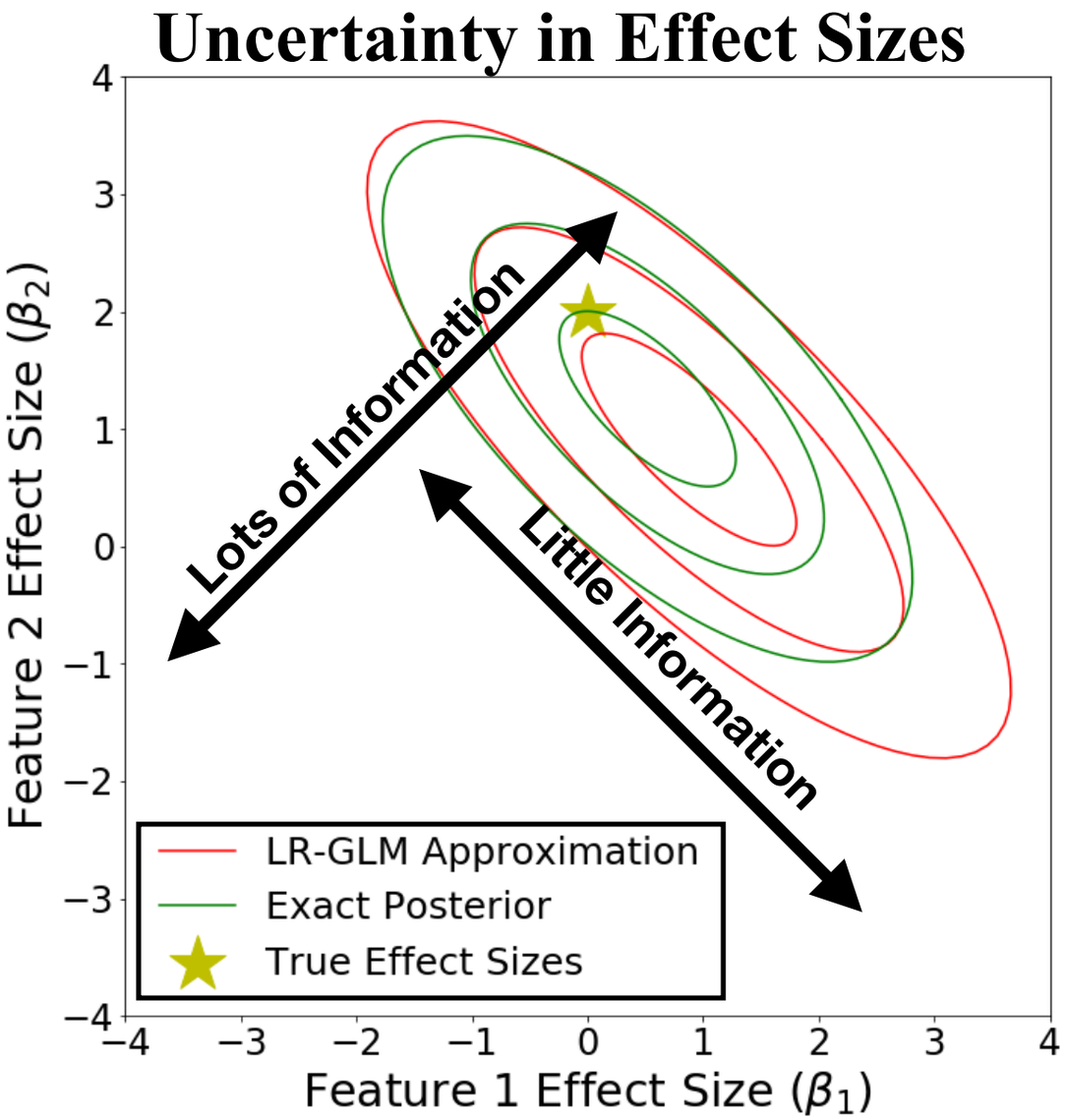
We ignore the least informative directions

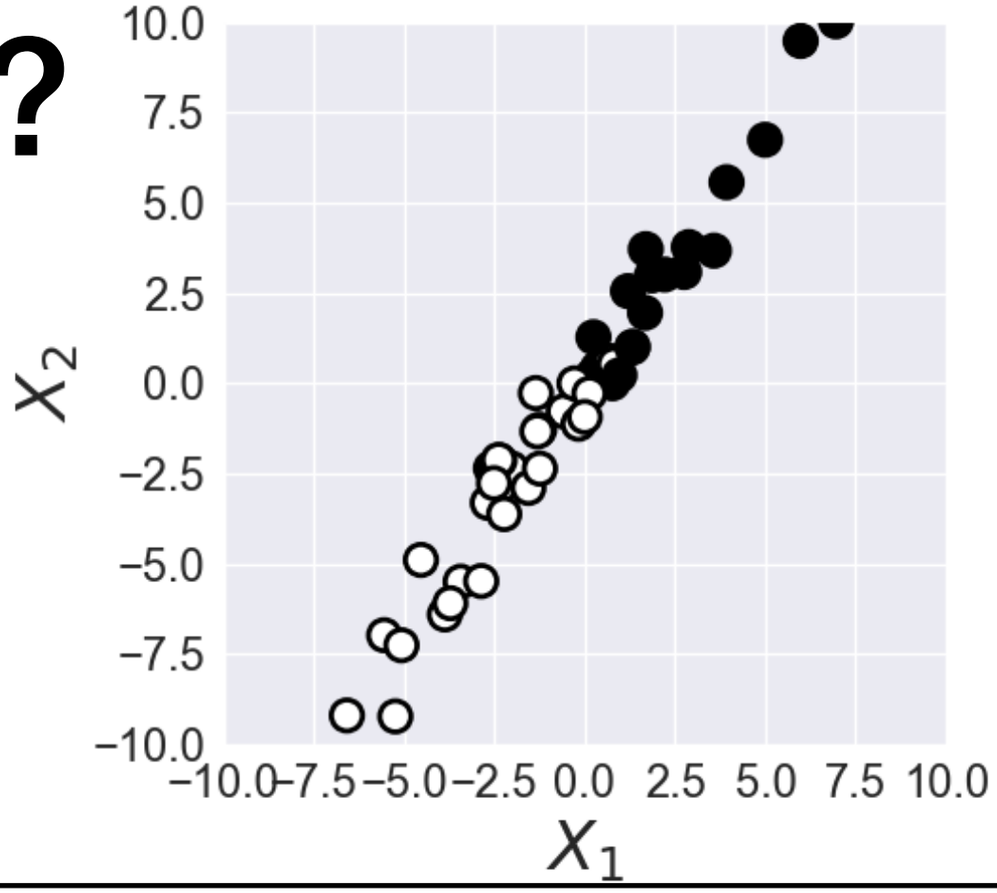
$$p(y_i|x_i^T\beta) \approx p(y_i|x_i UU^T\beta)$$

# How does it work?

**Cartoon Example**

- Logistic Regression with two correlated features




**The LR-GLM Approximation**

We ignore the least informative directions

$$p(y_i|x_i^T\beta) \approx p(y_i|x_iUU^T\beta)$$

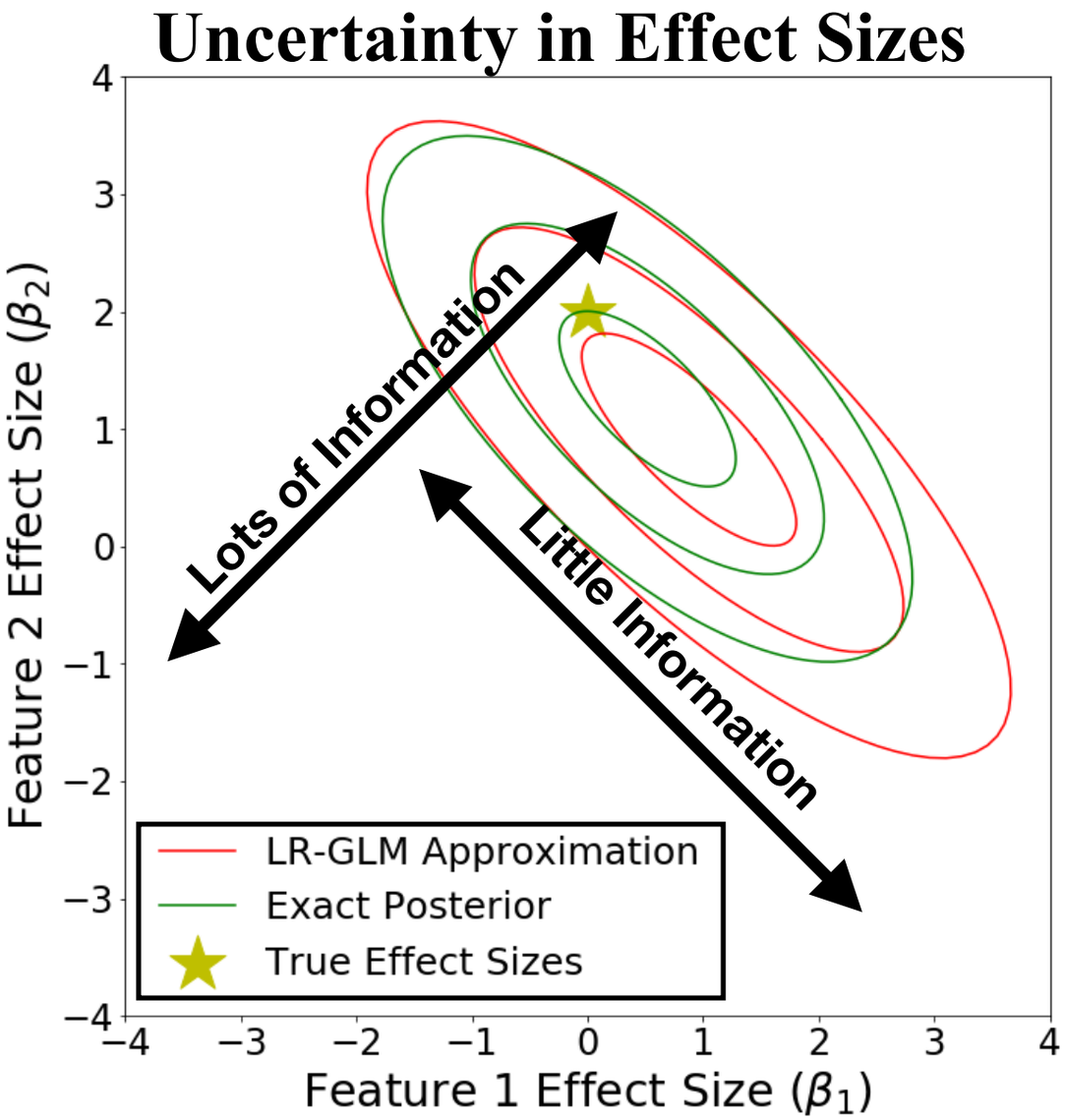
**Approximation Quality**

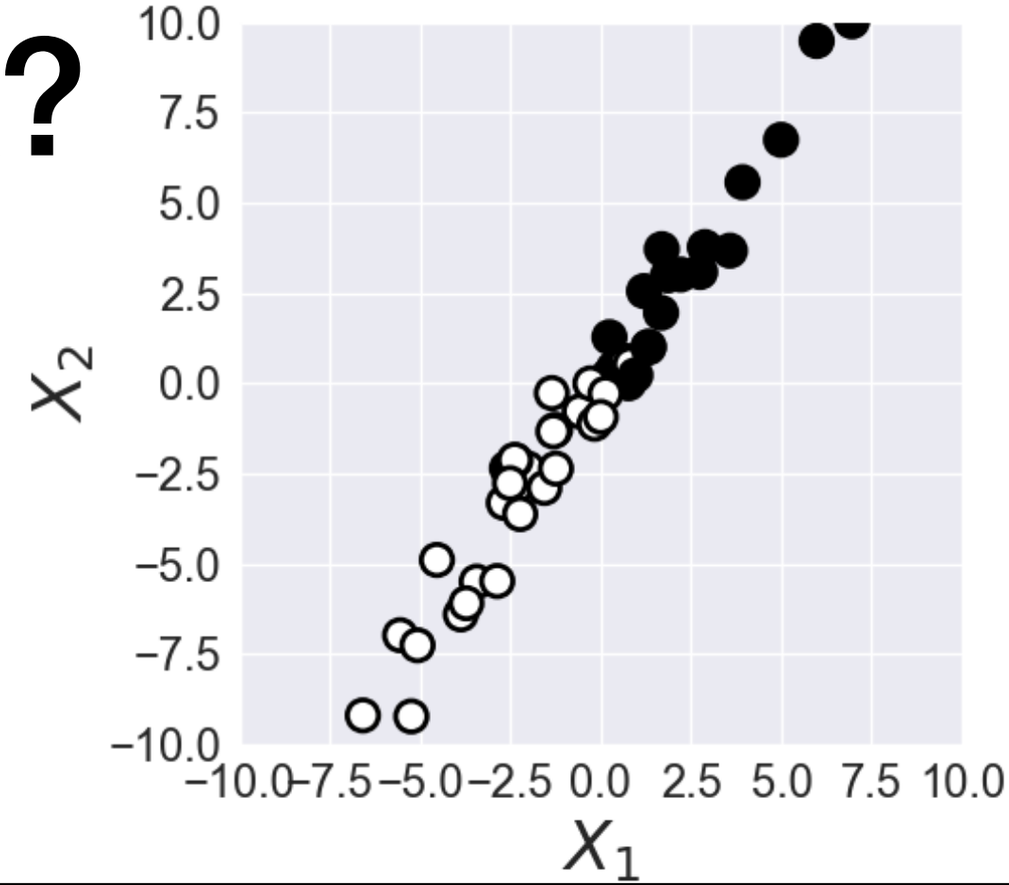
- Exact when data are low rank

# How does it work?

**Cartoon Example**

- Logistic Regression with two correlated features




**The LR-GLM Approximation**

We ignore the least informative directions

$$p(y_i|x_i^T\beta) \approx p(y_i|x_i UU^T\beta)$$

**Approximation Quality**

- Exact when data are low rank
- **We prove:** Approximation is close when the data are approximately low rank

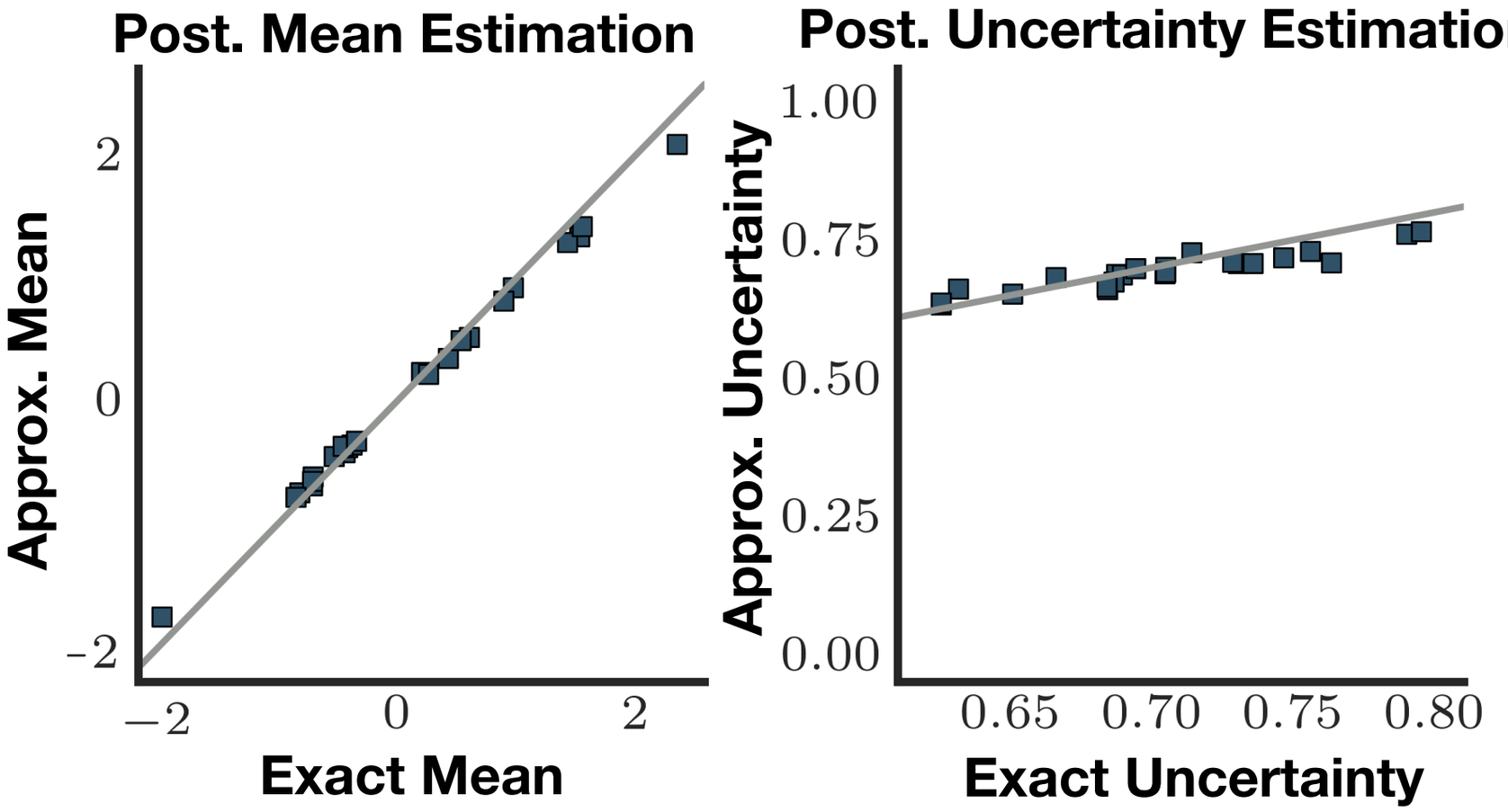
# Does it Work?

# Does it Work?

Evaluate by comparing exact means and uncertainties *(slow)* against our approximation *(fast)*

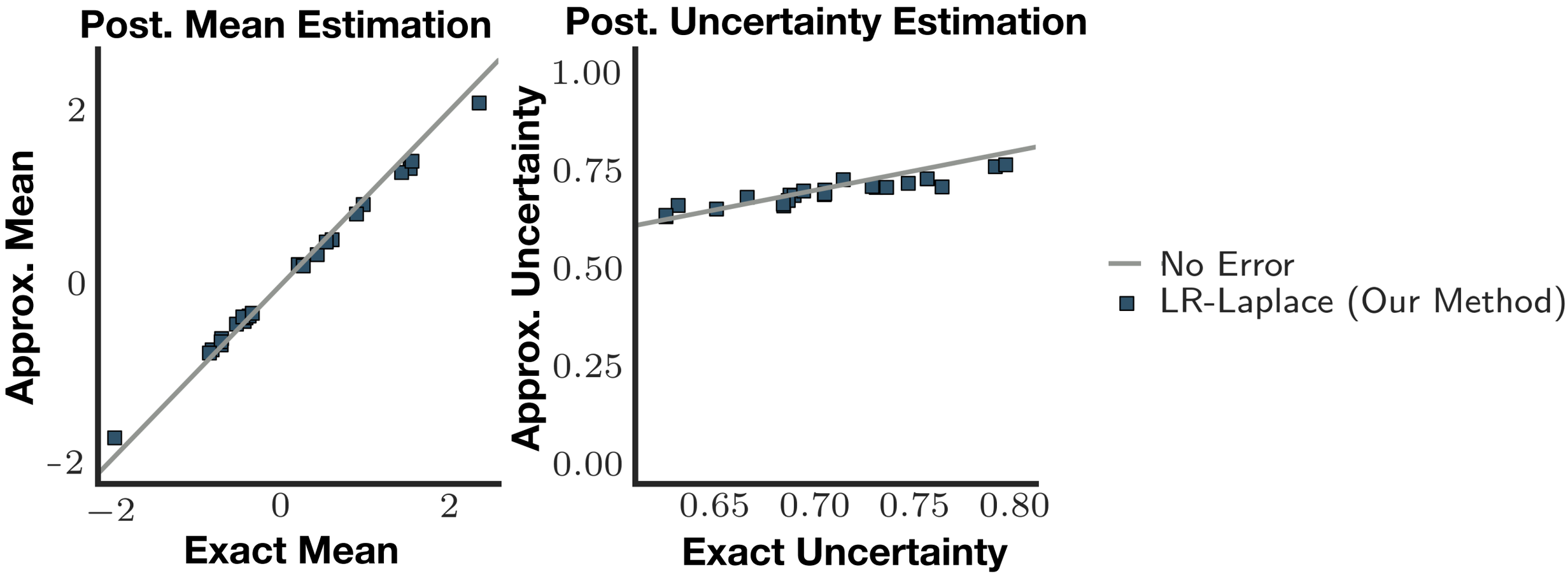
# Does it Work?

Evaluate by comparing exact means and uncertainties *(slow)* against our approximation *(fast)*



# Does it Work?

Evaluate by comparing exact means and uncertainties *(slow)* against our approximation *(fast)*

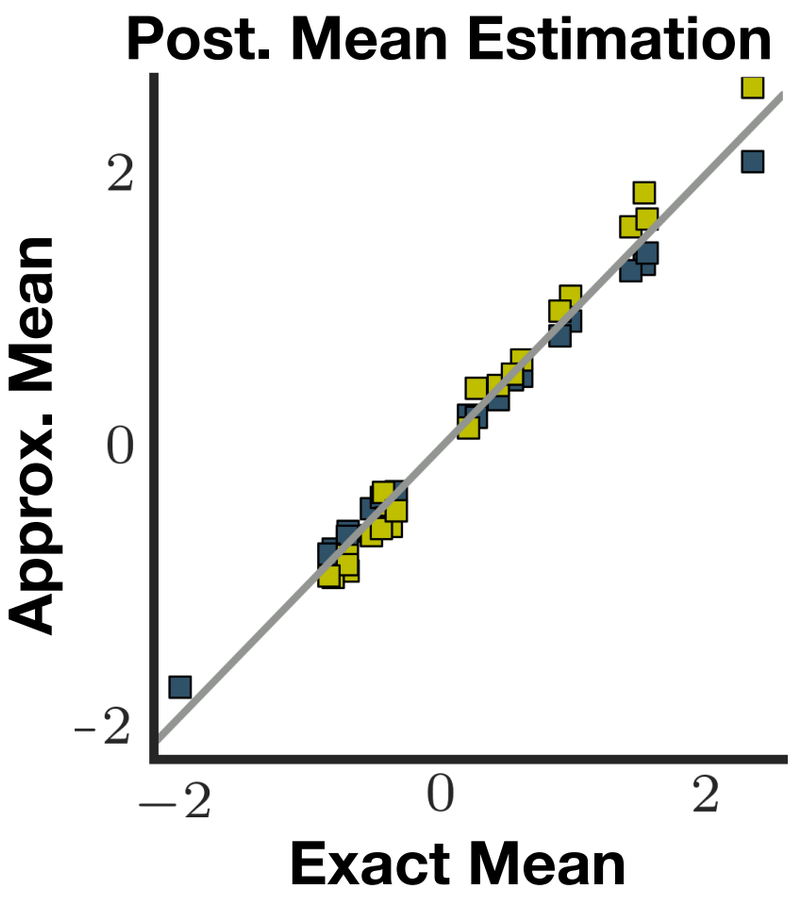


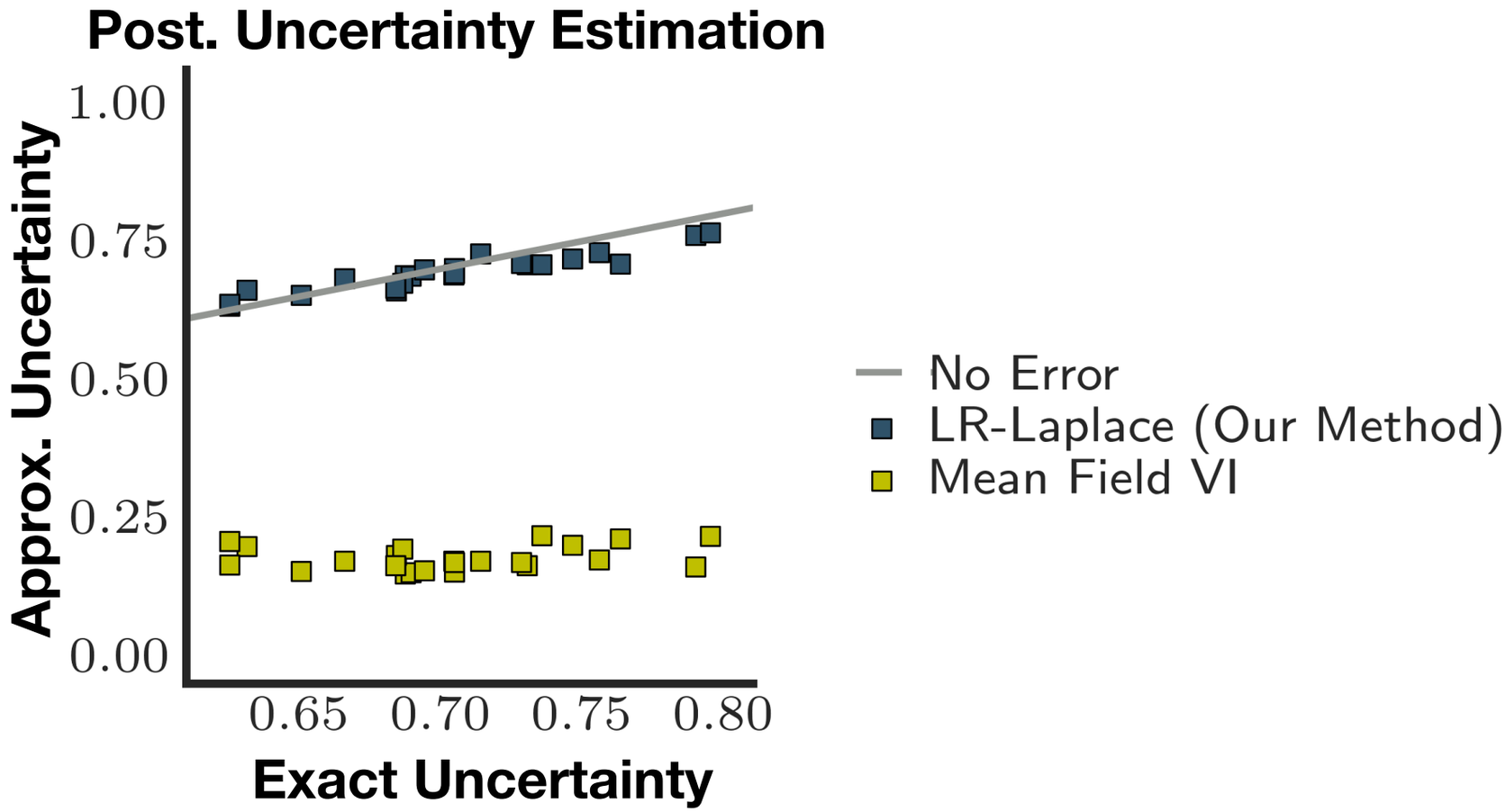

**We rigorously show…**

- Rank of approximation defines a computational-statistical trade-off

# Does it Work?

Evaluate by comparing exact means and uncertainties *(slow)* against our approximation *(fast)*



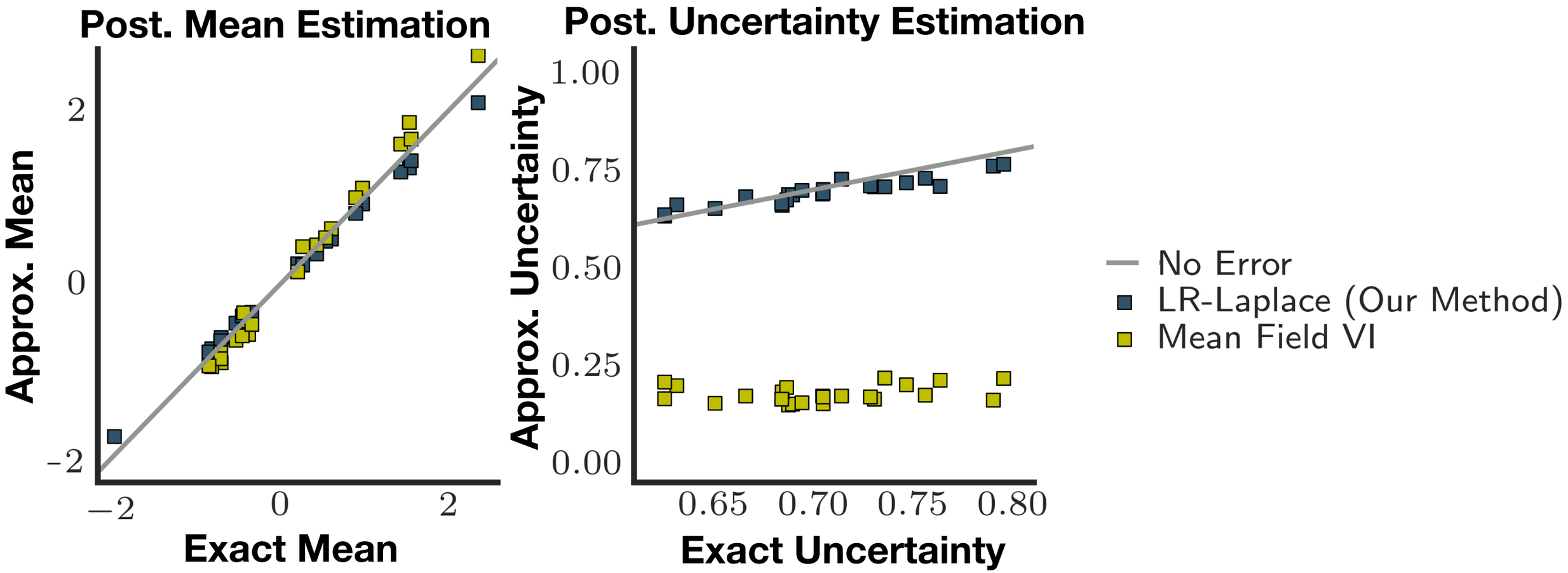




**We rigorously show…**

- Rank of approximation defines a computational-statistical trade-off

# Does it Work?

Evaluate by comparing exact means and uncertainties *(slow)* against our approximation *(fast)*

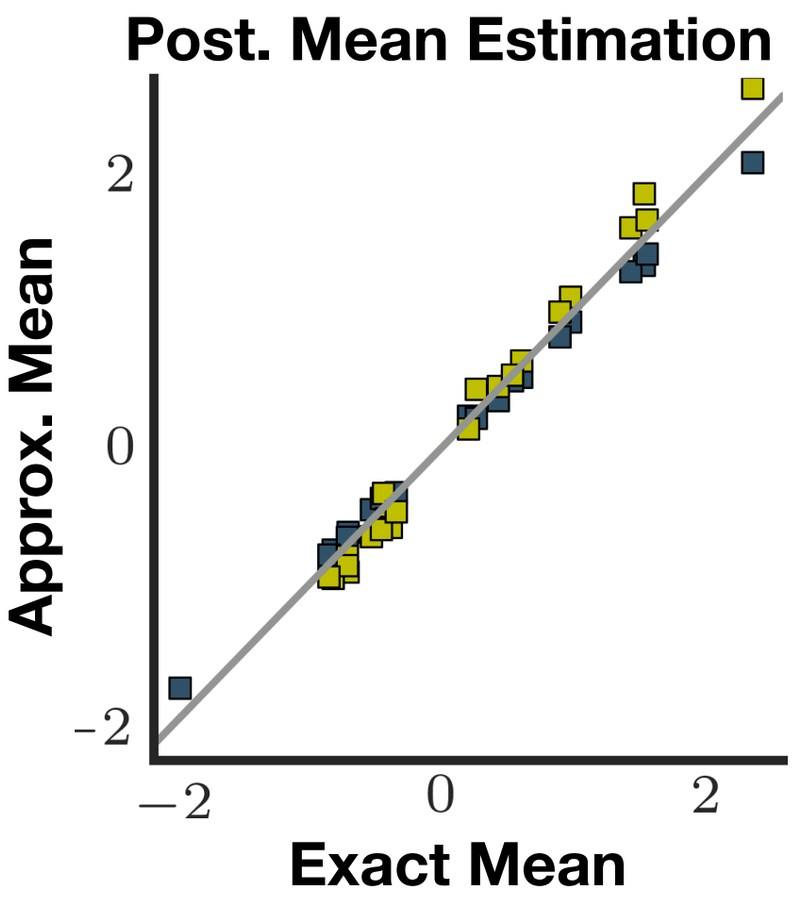


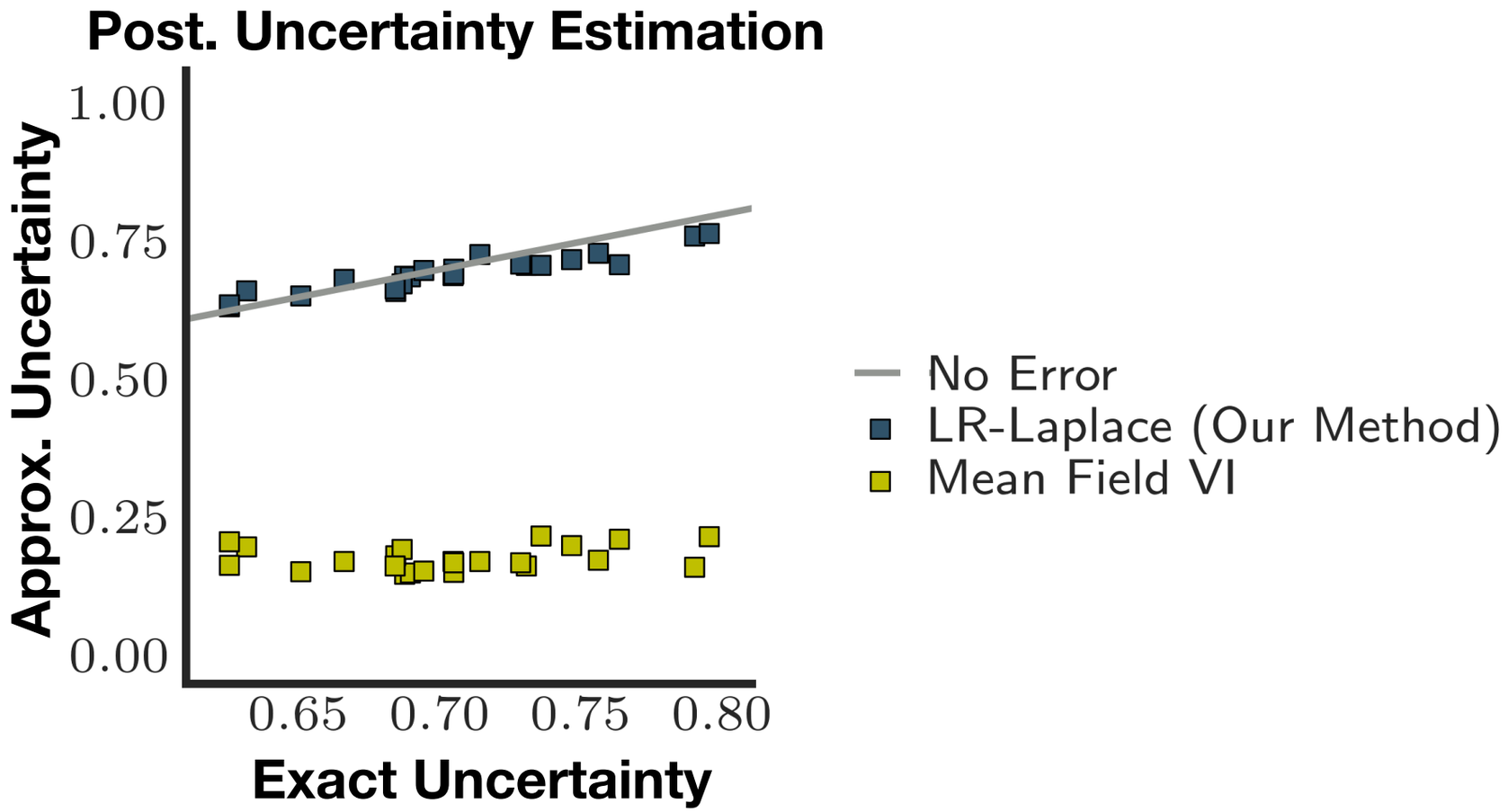

**We rigorously show…**

- Rank of approximation defines a computational-statistical trade-off
- The approximation is conservative (overestimates uncertainty)

# Does it Work?

Evaluate by comparing exact means and uncertainties *(slow)* against our approximation *(fast)*




**We rigorously show…**

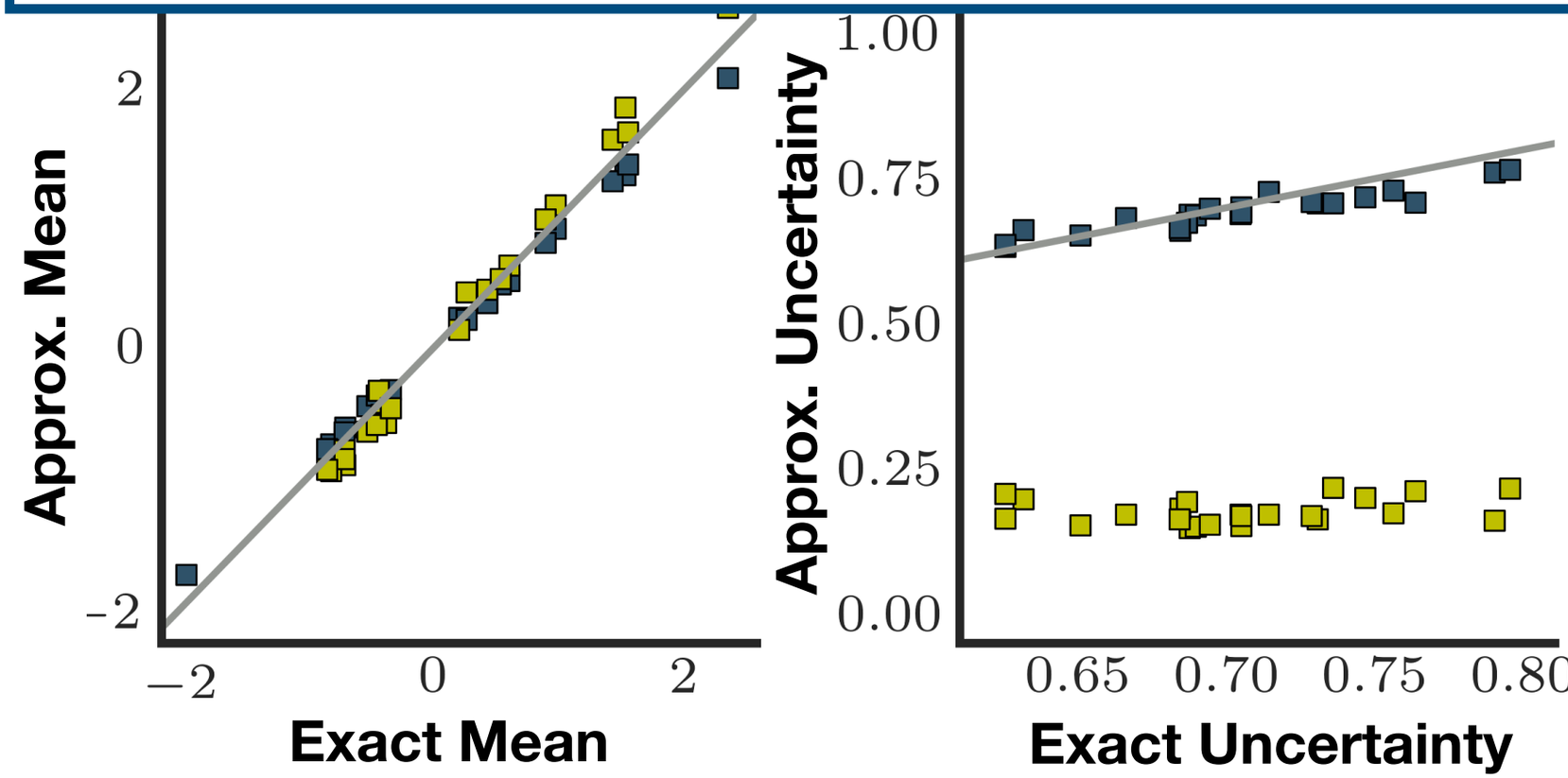
- Rank of approximation defines a computational-statistical trade-off
- The approximation is conservative (overestimates uncertainty)
- For high-dimensional, correlated data, **LR-GLM closely approximates the exact posterior up to 5X faster!**

# LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations

**Brian L. Trippe**, Jonathan H. Huggins, Raj Agrawal and Tamara Broderick

Paper: proceedings.mlr.press/v97/trippe19a

Poster: Pacific Ballroom #214




**We rigorously show…**

- Rank of approximation defines a computational-statistical trade-off
- The approximation is conservative (overestimates uncertainty)
- For high-dimensional, correlated data, **LR-GLM closely approximates the exact posterior up to 5X faster!**