Problem: Large number of observations or large dimension

Consequence: Modern regression approaches reach limits of scalability

Goal: Develop highly efficient regression approaches for massive data

Bayesian Regression
- \( \pi(\beta | X, Y) \propto \mathcal{L}(Y | X, \beta) \cdot \pi(\beta) \)
- \( \ell_p \) regression
- Hierarchical prior models

Generalised Linear Models (GLMs)
- \( h(E(Y)) = X\beta \)

Examples:
- Poisson regression for count data
- Logistic regression for binary data

Structural Constraints
- \( \min_{\xi \in \mathcal{N}} \| \xi - Y \|_R \), for some \( \mathcal{N} \subseteq \mathbb{R}^n \)
- E.g. monotonicity, unimodality
  \( \mathcal{N} = \{ \xi | \exists \xi \in [n]: \xi_1 \leq \ldots \leq \xi_t \geq \ldots \geq \xi_n \} \)
- Application: peak detection

Dimensionality reduction
- Feature selection
- Variable interactions
- SNP and genome-wide data

Data Reduction with Guaranteed Little Distortion
- Streaming
- Distributed
- Succinct representation

Sampling and projection methods
- \( \varepsilon \)-subspace embeddings
- \( \varepsilon \)-coresets
- \( \Pi \in \mathbb{R}^{k \times n}, k \ll n: \)
  \[ \forall \beta \in \mathbb{R}^d: (1 - \varepsilon)\|X\beta\| \leq \|\Pi X\beta\| \leq (1 + \varepsilon)\|X\beta\| \]

Algorithmic Principle
1. Data reduction \( [X, Y] \xrightarrow{n} [\Pi X, \Pi Y] \)
2. Statistical analysis \( \pi(\beta | X, Y) \xrightarrow{n} \pi(\beta | \Pi X, \Pi Y) \)

Spline Regression Model Structural Constraints
- \( h(E(Y)) = \beta_0 + \sum_{t=1}^T \beta_t L_t \)

Logic regression
- Interactions of (binary) variables
- Variable importance measures

Spline Regression
- Able to model unimodality
- Limited number \( B \) of splines (succinct representation)
- Peak detection in IMS spectra

Logistic Regression
- Geometric parameter captures difficulty of reduction
- \( \mu(X) = \sup_{\beta \in \mathbb{R}^n \setminus \{0\}} \frac{||X\beta||_1}{||X\beta||_2} \)
- \( \mu \)-complex data admits coreset of size \( O(\mu d \log n / \varepsilon^2) \)

Spline Regression with Guaranteed Little Distortion
- \( \varepsilon \)-subspace embeddings preserve \( \mathcal{L}(\beta | X, Y) \)
- \( \varepsilon \)-generalised linear regression, \( \rho \in [1, \infty) \)
- Hierarchical prior models

GLMs
- Lower bounds \( k \in \Omega(n / \log n) \)

Poisson regression
- Statistical model relaxation yields data summary
- Close approximation of maximum likelihood estimator

Logistic Regression
- \( \mu(X) = \sup_{\beta \in \mathbb{R}^n \setminus \{0\}} \frac{||X\beta||_2}{||X\beta||_1} \)
- \( \mu \)-complex data admits coreset of size \( O(\mu d \log n / \varepsilon^2) \)

Data Reduction with Guaranteed Little Distortion
- Streaming
- Distributed
- Succinct representation

Sampling and projection methods
- \( \varepsilon \)-subspace embeddings
- \( \varepsilon \)-coresets
- \( \Pi \in \mathbb{R}^{k \times n}, k \ll n: \)
  \[ \forall \beta \in \mathbb{R}^d: (1 - \varepsilon)\|X\beta\| \leq \|\Pi X\beta\| \leq (1 + \varepsilon)\|X\beta\| \]

Algorithmic Principle
1. Data reduction \( [X, Y] \xrightarrow{n} [\Pi X, \Pi Y] \)
2. Statistical analysis \( \pi(\beta | X, Y) \xrightarrow{n} \pi(\beta | \Pi X, \Pi Y) \)

Spline Regression Model Structural Constraints
- \( h(E(Y)) = \beta_0 + \sum_{t=1}^T \beta_t L_t \)

Logic regression
- Interactions of (binary) variables
- Variable importance measures

Spline Regression
- Able to model unimodality
- Limited number \( B \) of splines (succinct representation)
- Peak detection in IMS spectra

Logistic Regression
- Geometric parameter captures difficulty of reduction
- \( \mu(X) = \sup_{\beta \in \mathbb{R}^n \setminus \{0\}} \frac{||X\beta||_1}{||X\beta||_2} \)
- \( \mu \)-complex data admits coreset of size \( O(\mu d \log n / \varepsilon^2) \)

Data Reduction with Guaranteed Little Distortion
- Streaming
- Distributed
- Succinct representation

Sampling and projection methods
- \( \varepsilon \)-subspace embeddings
- \( \varepsilon \)-coresets
- \( \Pi \in \mathbb{R}^{k \times n}, k \ll n: \)
  \[ \forall \beta \in \mathbb{R}^d: (1 - \varepsilon)\|X\beta\| \leq \|\Pi X\beta\| \leq (1 + \varepsilon)\|X\beta\| \]

Algorithmic Principle
1. Data reduction \( [X, Y] \xrightarrow{n} [\Pi X, \Pi Y] \)
2. Statistical analysis \( \pi(\beta | X, Y) \xrightarrow{n} \pi(\beta | \Pi X, \Pi Y) \)

Spline Regression Model Structural Constraints
- \( h(E(Y)) = \beta_0 + \sum_{t=1}^T \beta_t L_t \)

Logic regression
- Interactions of (binary) variables
- Variable importance measures

Spline Regression
- Able to model unimodality
- Limited number \( B \) of splines (succinct representation)
- Peak detection in IMS spectra

Logistic Regression
- Geometric parameter captures difficulty of reduction
- \( \mu(X) = \sup_{\beta \in \mathbb{R}^n \setminus \{0\}} \frac{||X\beta||_1}{||X\beta||_2} \)
- \( \mu \)-complex data admits coreset of size \( O(\mu d \log n / \varepsilon^2) \)

Data Reduction with Guaranteed Little Distortion
- Streaming
- Distributed
- Succinct representation

Sampling and projection methods
- \( \varepsilon \)-subspace embeddings
- \( \varepsilon \)-coresets
- \( \Pi \in \mathbb{R}^{k \times n}, k \ll n: \)
  \[ \forall \beta \in \mathbb{R}^d: (1 - \varepsilon)\|X\beta\| \leq \|\Pi X\beta\| \leq (1 + \varepsilon)\|X\beta\| \]

Algorithmic Principle
1. Data reduction \( [X, Y] \xrightarrow{n} [\Pi X, \Pi Y] \)
2. Statistical analysis \( \pi(\beta | X, Y) \xrightarrow{n} \pi(\beta | \Pi X, \Pi Y) \)

Spline Regression Model Structural Constraints
- \( h(E(Y)) = \beta_0 + \sum_{t=1}^T \beta_t L_t \)

Logic regression
- Interactions of (binary) variables
- Variable importance measures

Spline Regression
- Able to model unimodality
- Limited number \( B \) of splines (succinct representation)
- Peak detection in IMS spectra

Logistic Regression
- Geometric parameter captures difficulty of reduction
- \( \mu(X) = \sup_{\beta \in \mathbb{R}^n \setminus \{0\}} \frac{||X\beta||_1}{||X\beta||_2} \)
- \( \mu \)-complex data admits coreset of size \( O(\mu d \log n / \varepsilon^2) \)

Data Reduction with Guaranteed Little Distortion
- Streaming
- Distributed
- Succinct representation

Sampling and projection methods
- \( \varepsilon \)-subspace embeddings
- \( \varepsilon \)-coresets
- \( \Pi \in \mathbb{R}^{k \times n}, k \ll n: \)
  \[ \forall \beta \in \mathbb{R}^d: (1 - \varepsilon)\|X\beta\| \leq \|\Pi X\beta\| \leq (1 + \varepsilon)\|X\beta\| \]

Algorithmic Principle
1. Data reduction \( [X, Y] \xrightarrow{n} [\Pi X, \Pi Y] \)
2. Statistical analysis \( \pi(\beta | X, Y) \xrightarrow{n} \pi(\beta | \Pi X, \Pi Y) \)

Spline Regression Model Structural Constraints
- \( h(E(Y)) = \beta_0 + \sum_{t=1}^T \beta_t L_t \)

Logic regression
- Interactions of (binary) variables
- Variable importance measures

Spline Regression
- Able to model unimodality
- Limited number \( B \) of splines (succinct representation)
- Peak detection in IMS spectra

Logistic Regression
- Geometric parameter captures difficulty of reduction
- \( \mu(X) = \sup_{\beta \in \mathbb{R}^n \setminus \{0\}} \frac{||X\beta||_1}{||X\beta||_2} \)
- \( \mu \)-complex data admits coreset of size \( O(\mu d \log n / \varepsilon^2) \)