SWARM INTELLIGENCE METHODS FOR STATISTICAL REGRESSION

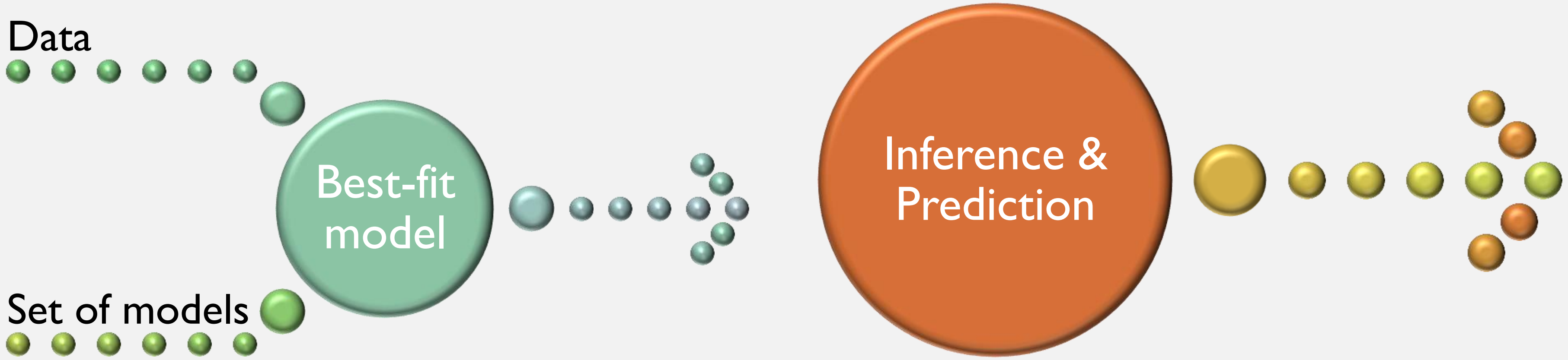
Lecture 1

Soumya D. Mohanty

University of Texas Rio Grande Valley

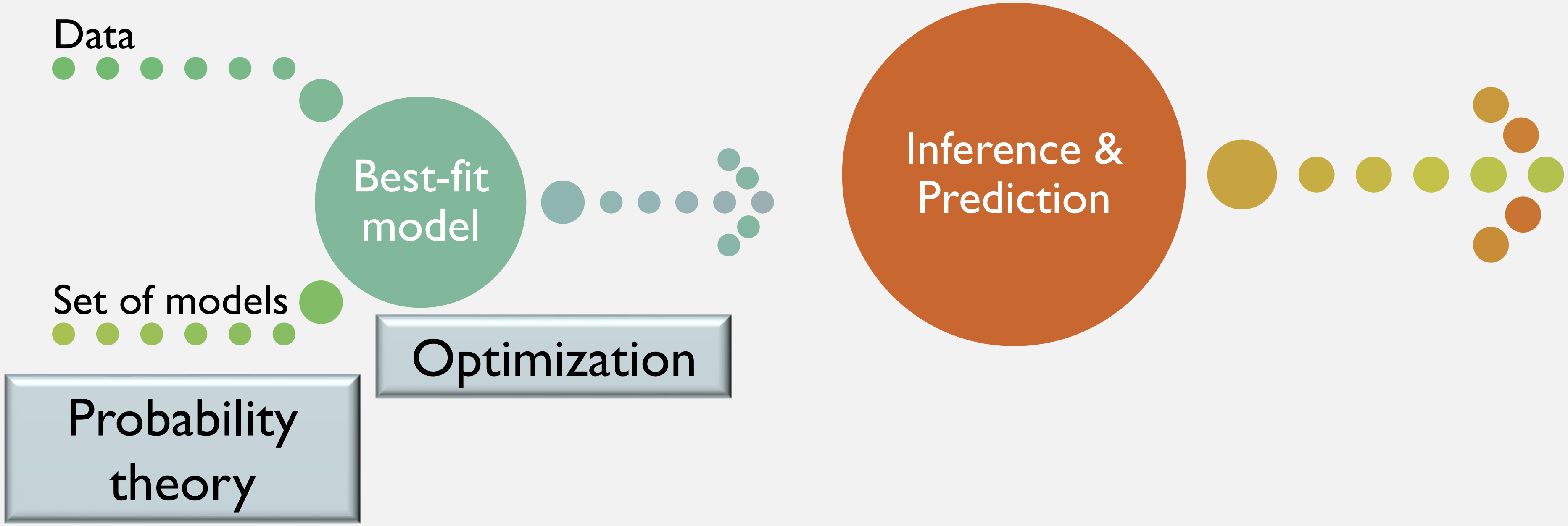
COURSE MOTIVATION

STATISTICAL ANALYSIS



MOTIVATION

Statistical data analysis stands on two legs

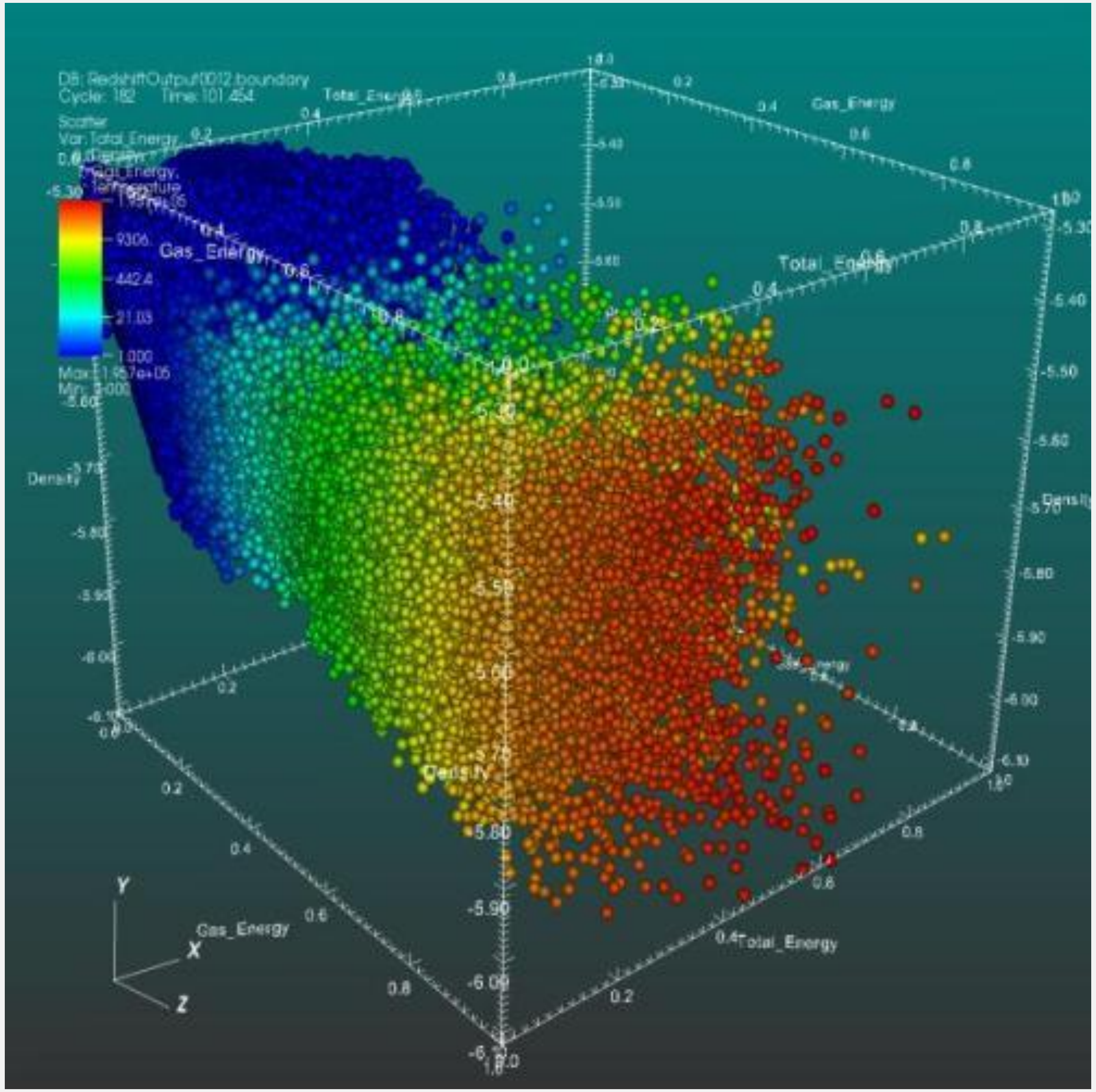


MOTIVATION

For large and complex data sets in the big data era, we need flexible models

Flexibility requires models to be

- High-dimensional and/or
- Non-linear



Wikipedia: Data visualization

MOTIVATION

Global optimization of high-dimensional, non-linear statistical models is a challenging task

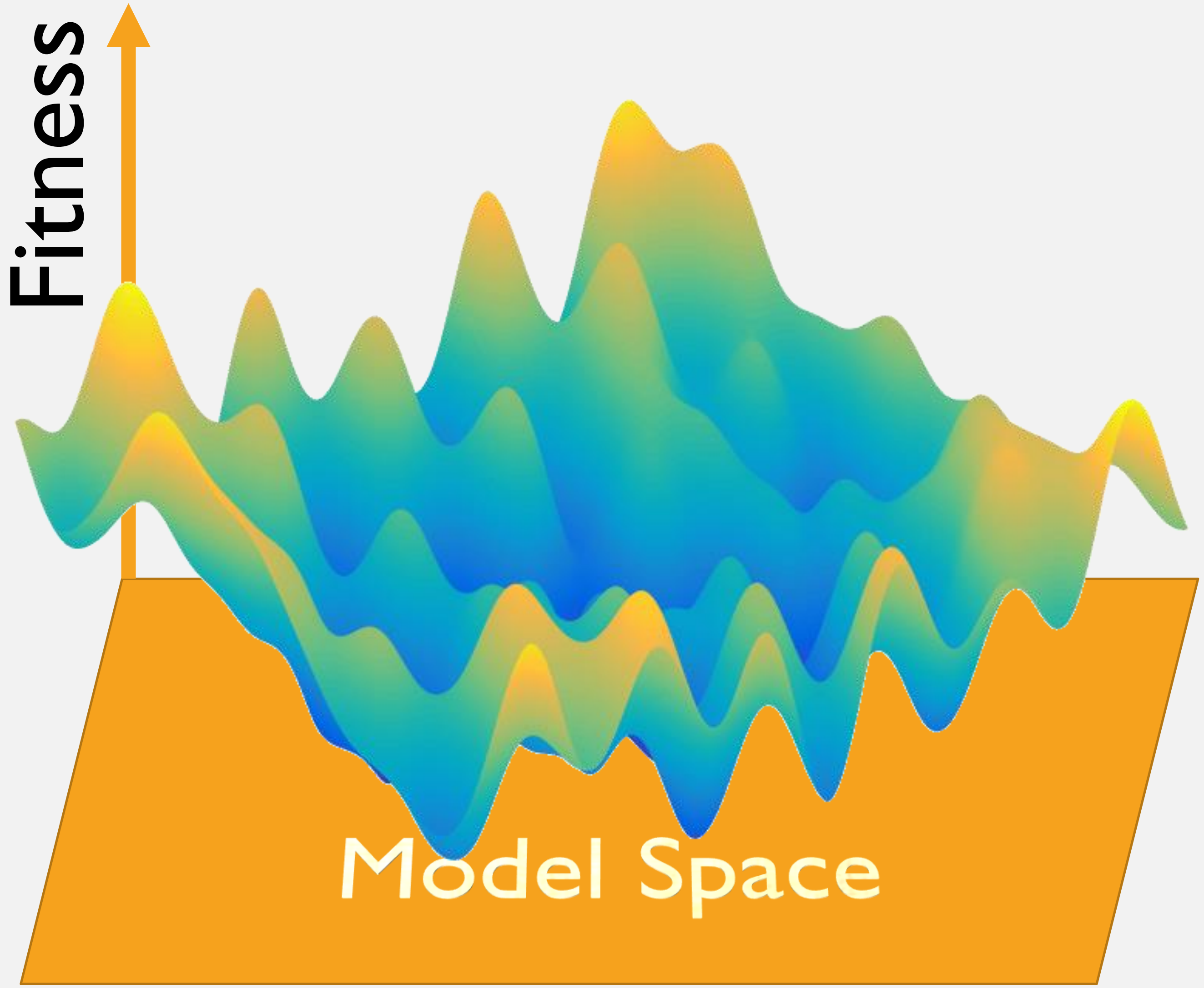
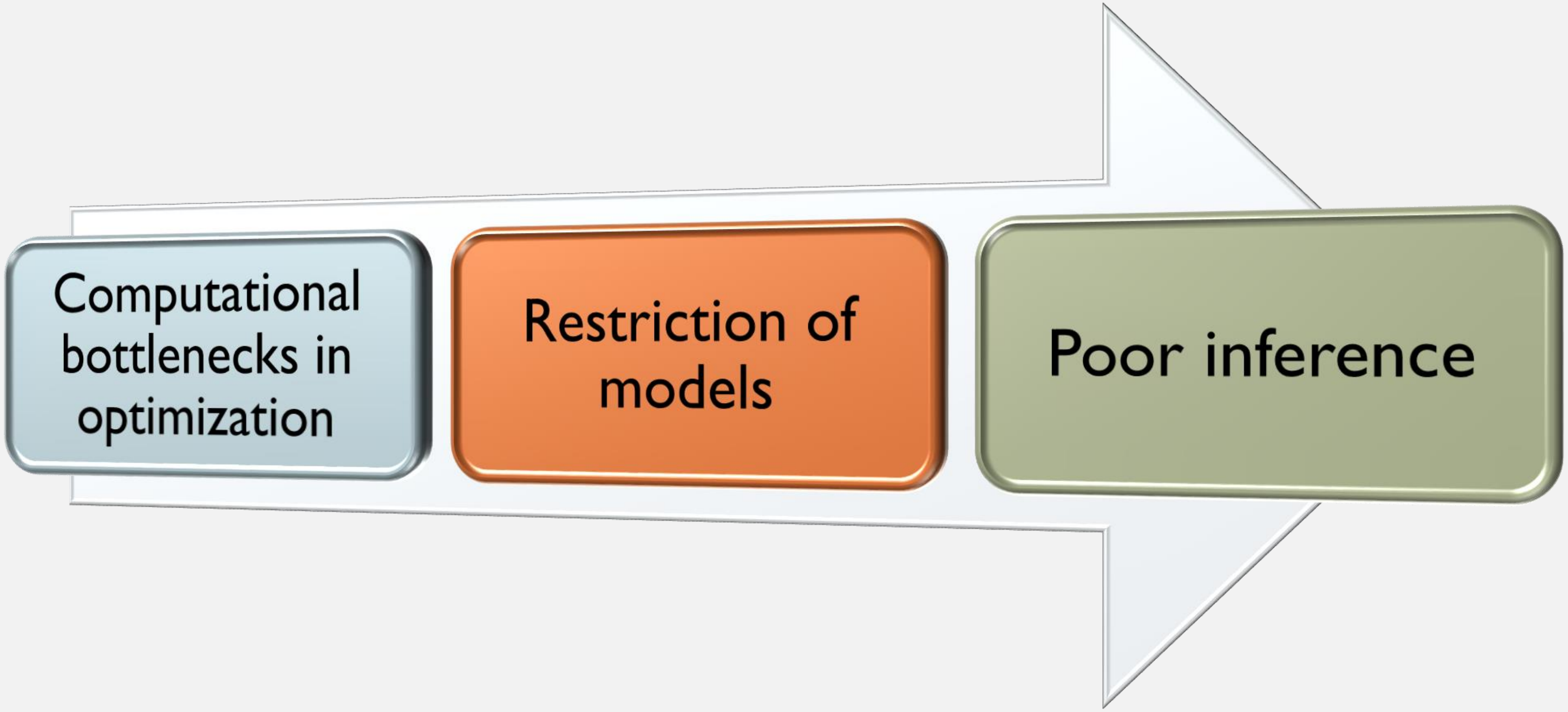


Image: https://users.ece.cmu.edu/~yuejiec/research.html

MOTIVATION


Computational bottlenecks in optimization
Restriction of models
Poor inference

MOTIVATION

Swarm intelligence (SI) methods can prove effective for optimization in statistical analysis

Success in breaking through the optimization barrier allows better modeling of data

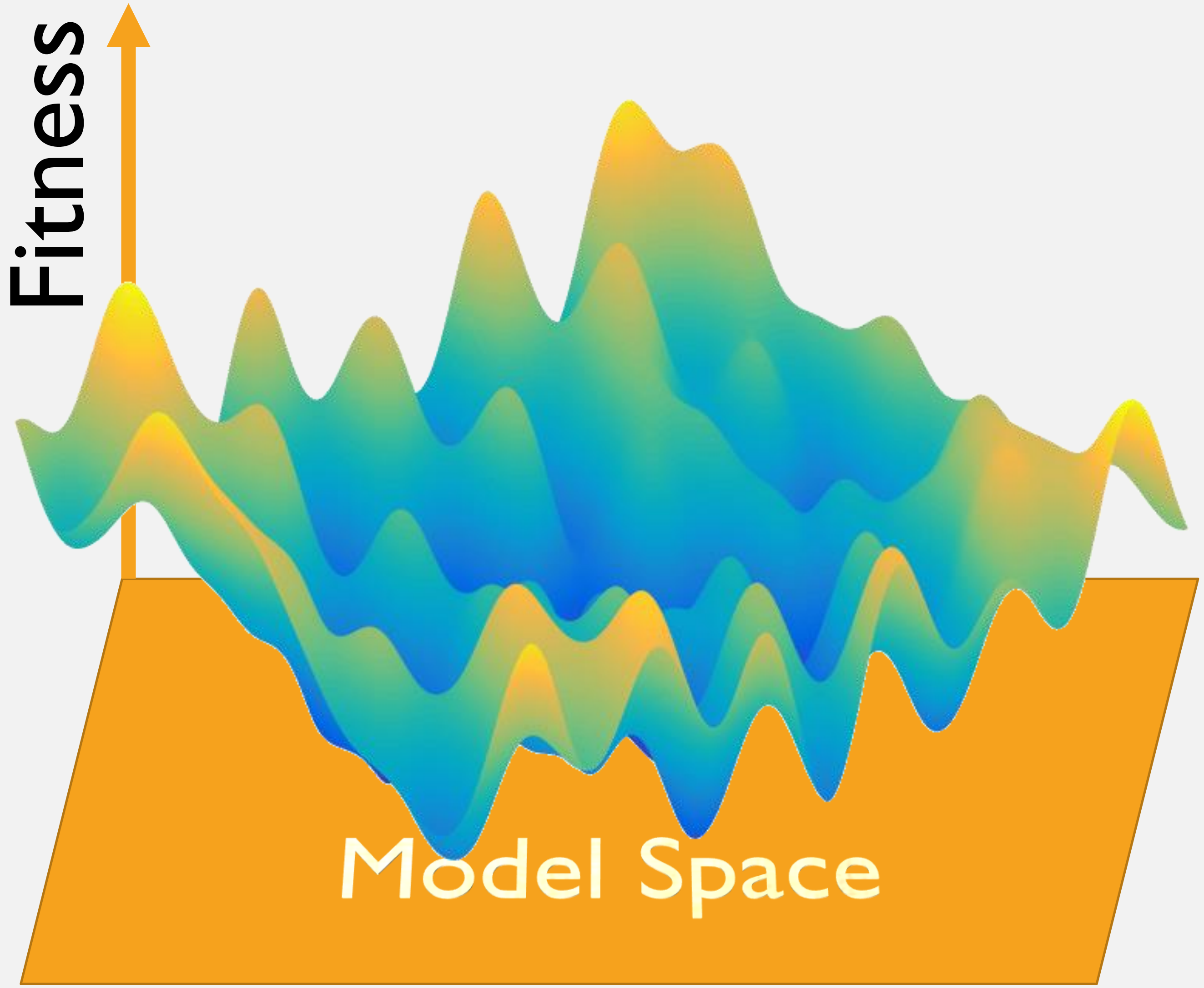
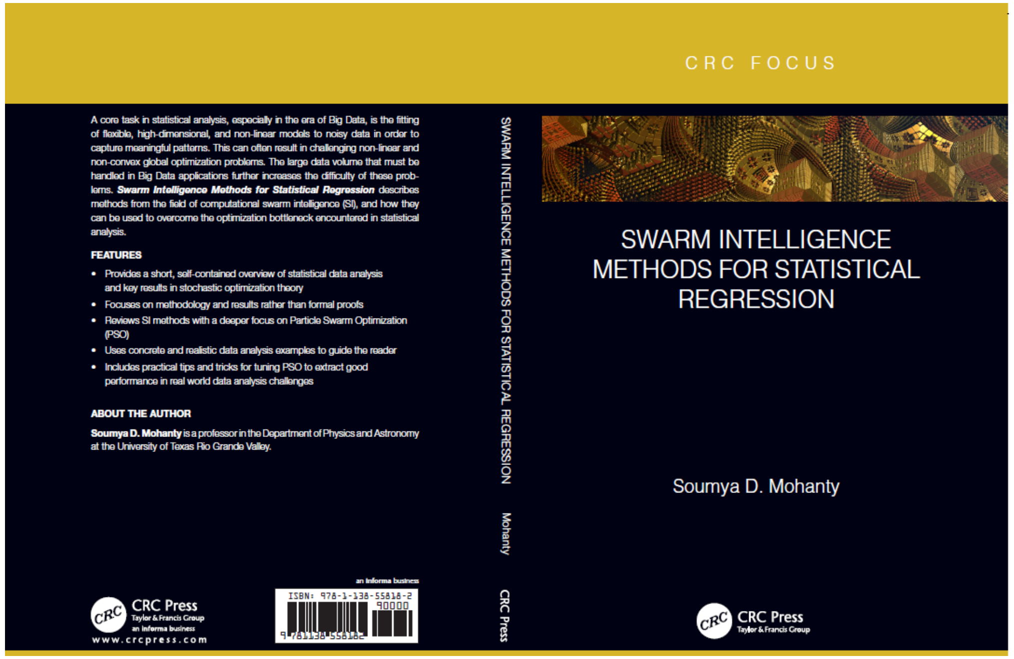


Image: https://users.ece.cmu.edu/~yuejiec/research.html

COURSE LOGISTICS & OUTLINE

LOGISTICS

- Slide contents condensed from
  - *Swarm intelligence methods for statistical regression*, Soumya D. Mohanty, Chapman Hall/ CRC Press (2018).
- <Course folder>/
  - README: Course summary
  - SLIDES: Lecture slides
  - READING: Pointers to supplementary reading
  - CODES: Examples discussed in the lectures



Lecture 1

- Review of statistical regression
- Importance of global optimization
- Successes of Swarm Intelligence in gravitational wave data analysis

Lecture 2

- Fundamental results in stochastic global optimization
- General application strategies
- Review: Evolutionary computation & swarm intelligence
- Particle swarm optimization (PSO) - I

Lecture 3

- PSO – II (Variants, settings etc)
- End-to-end applications of PSO to representative regression problems

LECTURE 1 OUTLINE

Introduction

- Statistical regression
  - Parametric: Linear / non-linear
  - Non-parametric: Linear / non-Linear
  - Example: Linear vs non-Linear
- Optimization challenges

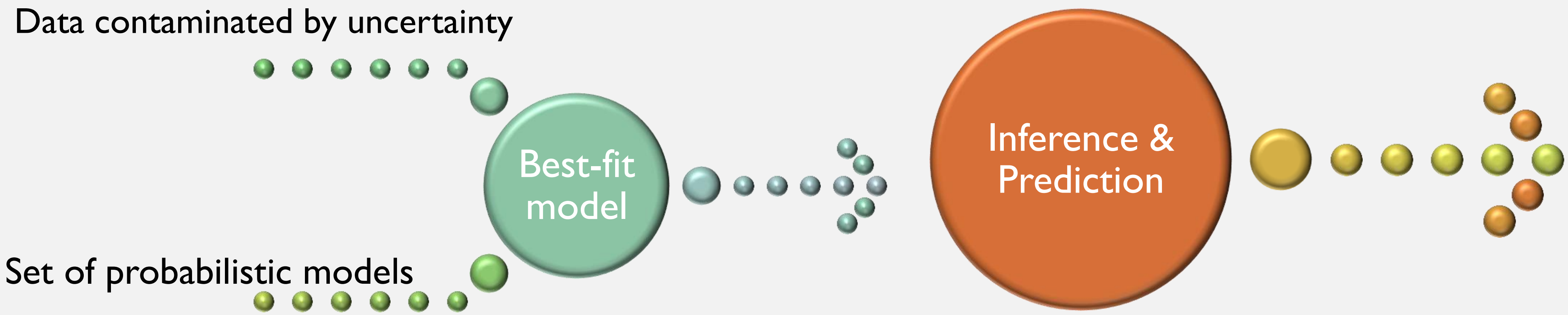
Successes of SI

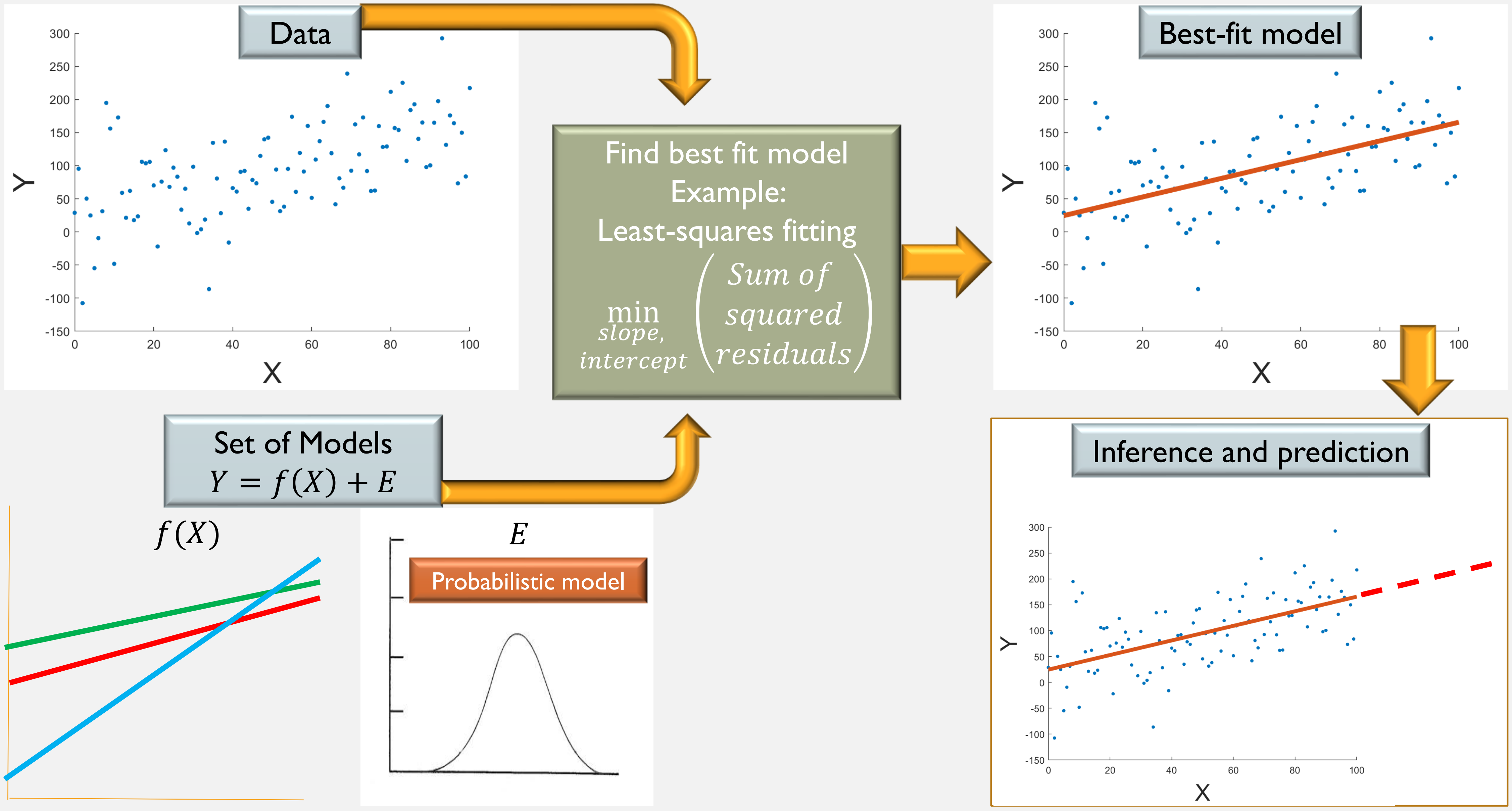
- Gravitational wave astronomy
- SI enabled methods
  - Parametric regression
  - Non-parametric regression

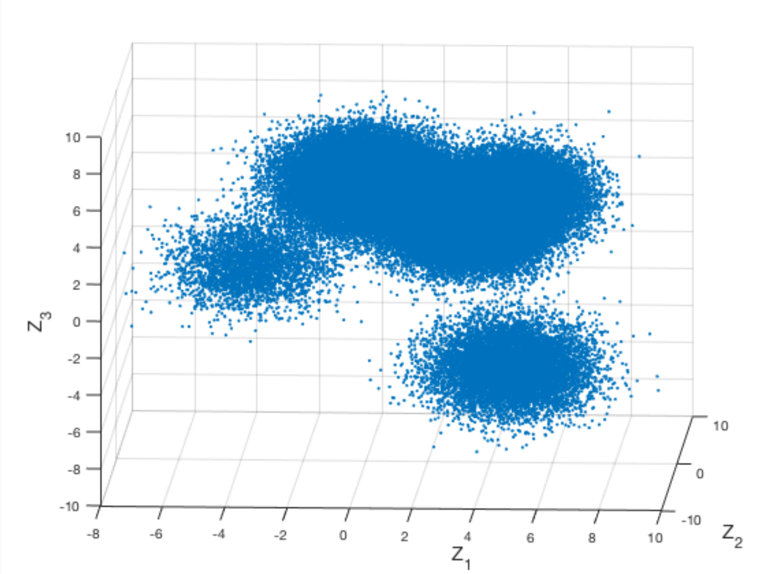
STATISTICAL REGRESSION

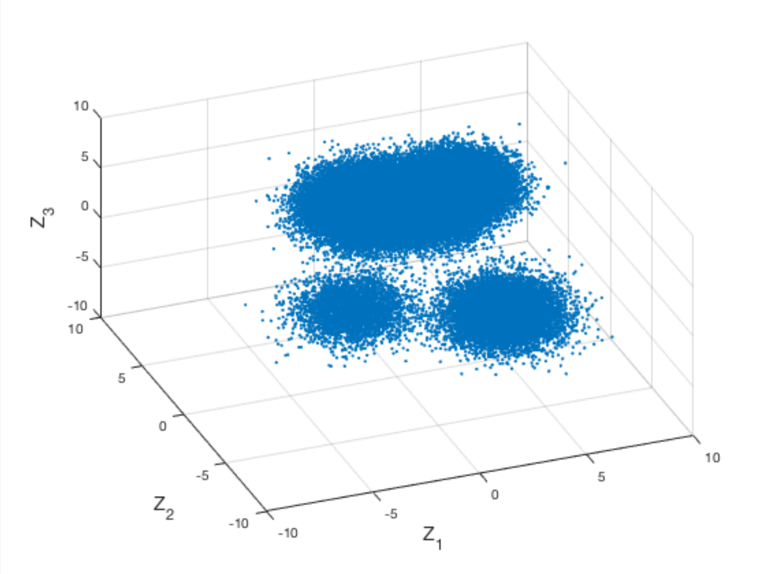
BASIC PROCESS

Statistical data analysis (a.k.a. machine learning)




Data
Find best fit model
Example:
Least-squares fitting
$\min_{slope,\ intercept} \left(\text{Sum of squared residuals} \right)$
Best-fit model
Set of Models
$Y = f(X) + E$
$f(X)$
E
Probabilistic model
Inference and prediction
X
Y





STATISTICAL ANALYSIS: GENERAL FORMULATION

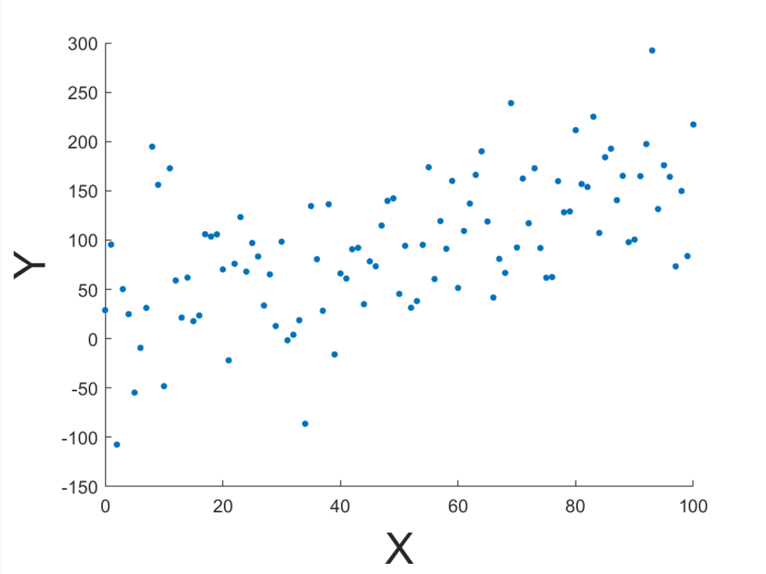
Data: Trial values $\{\bar{z}_0, \bar{z}_1, \ldots, \bar{z}_{N-1}\}$ of a vector random variable $\bar{Z} = (Z_1, Z_2, \ldots, Z_M)$

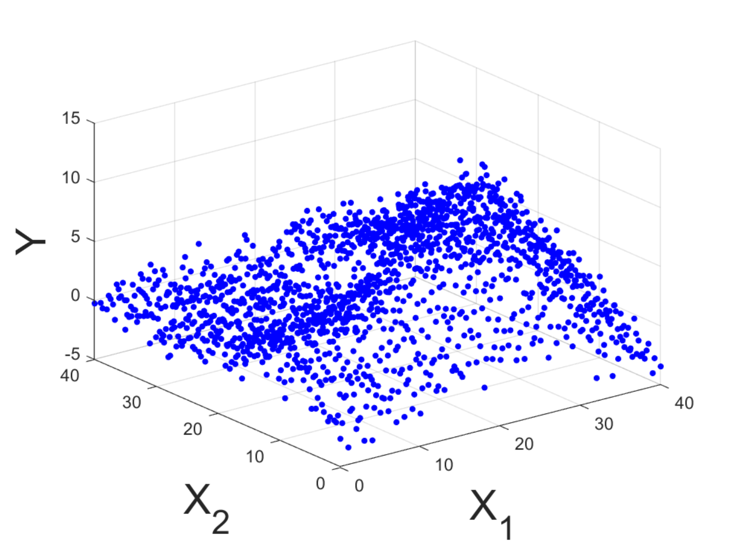
Model fitting: Find a model of the joint probability density function (pdf) of $\bar{Z}$

$$p_{\bar{Z}}(\bar{z})$$

that is best supported by the data

Density estimation is the primary goal of statistical analysis of data





STATISTICAL REGRESSION

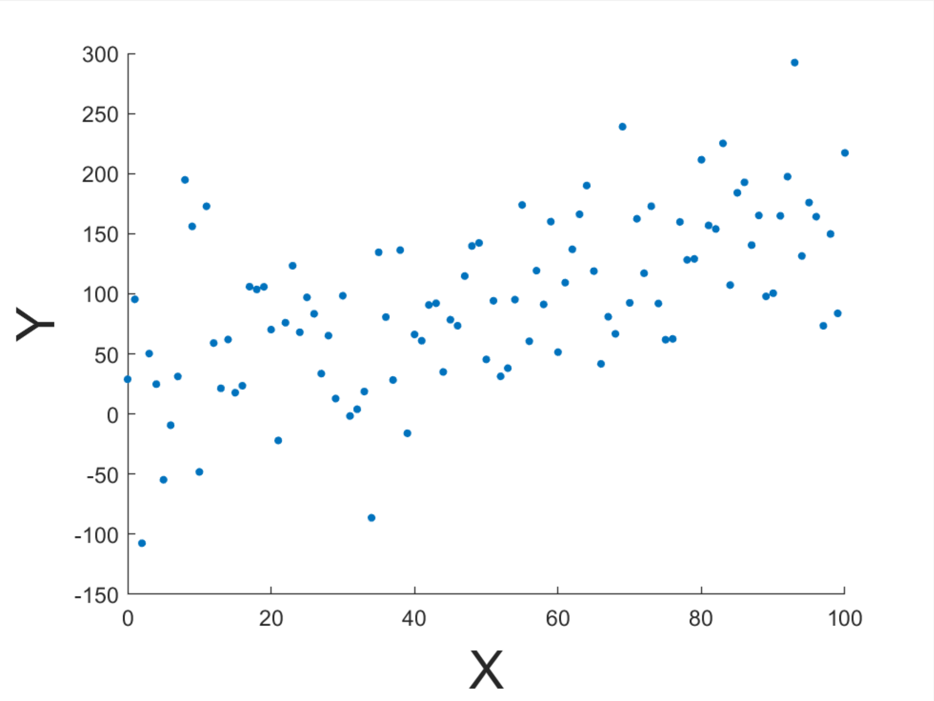
Statistical analysis on observational data in the form of pairs: $\bar{z}_i = (\bar{y}_i, \bar{x}_i), 0 \leq i \leq N-1$

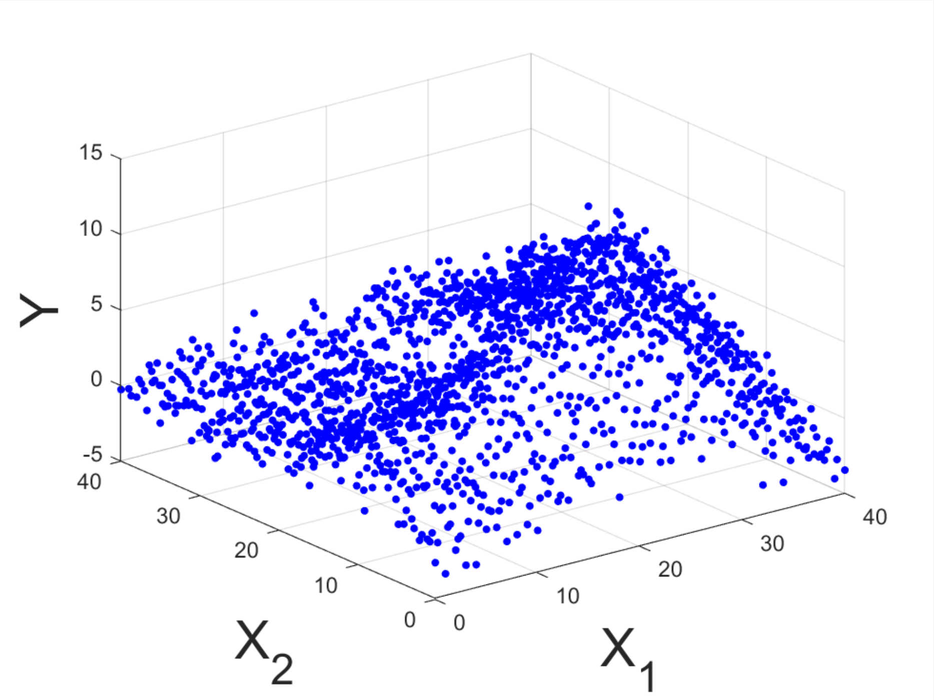
$\bar{y}_i$ is a trial value of random vector $\bar{Y}$

- $\bar{Y}$: Dependent variable
- $\bar{Y} = (\mathrm{Y}_0, \mathrm{Y}_1, \dots, Y_{K-1}) \in \mathbb{R}^K$

$\bar{x}_i$ is a trial value of random vector $\bar{X}$

- $\bar{X}$: Independent variable
- $\bar{X} = (\mathrm{X}_0, \mathrm{X}_1, \dots, X_{M-1}) \in \mathbb{R}^M$





STATISTICAL REGRESSION

Model fitting: fit a model for the conditional probability density function (pdf)

$$p_{\bar{Y}|\bar{X}}(\bar{y}|\bar{x})$$

Inference: Given any $\bar{x}$, we can make a probabilistic prediction for the value of $\bar{y}$ from the best fit model

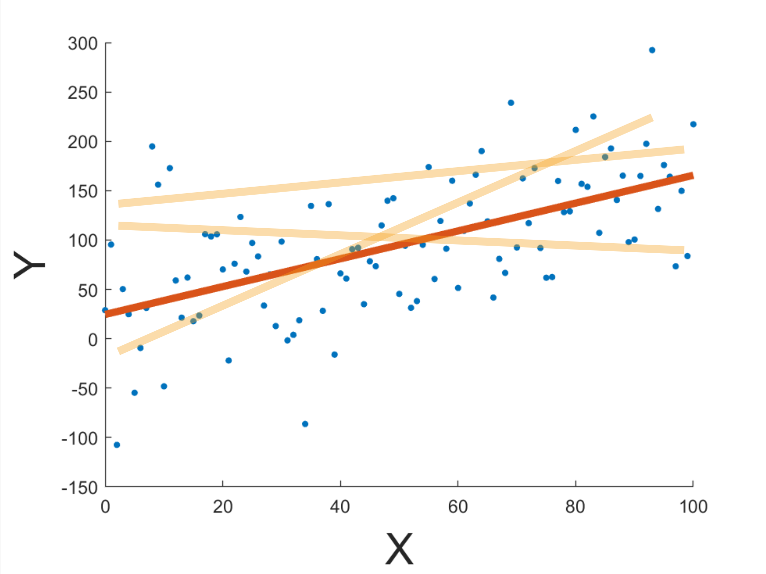
STATISTICAL ANALYSIS AND MACHINE LEARNING

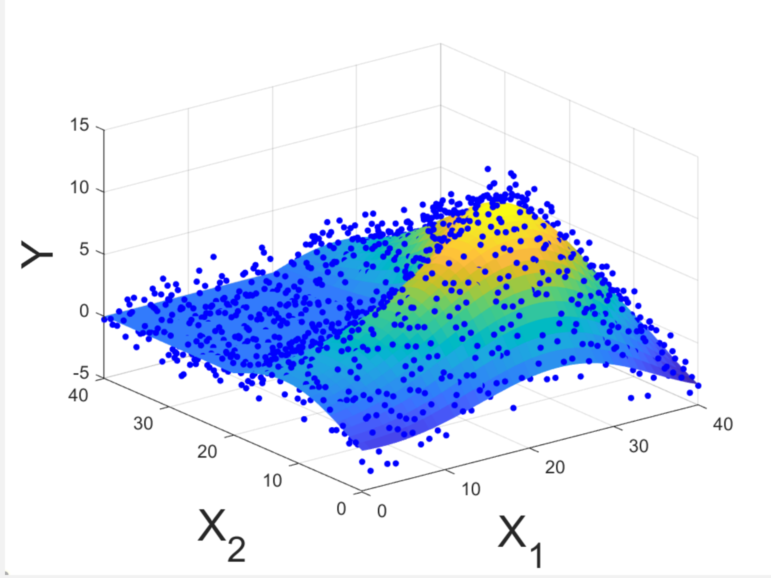
STATISTICAL ANALYSIS

- Data
- Density estimation
- Regression
- Emphasis on
  - Foundations and general results
  - Small data

MACHINE LEARNING

- = Training data
- = Unsupervised learning
- = Supervised learning
- Emphasis on
  - Computationally intensive methods
  - Big data





STATISTICAL REGRESSION

A common situation is where we assume models of the form

$$\bar{Y} = \bar{f}(\bar{X}) + \bar{E}$$

$\bar{E}$: Random vector with known joint pdf

Fitting goal: From a specified set of $\bar{f}$, pick the best one

* Only scalar X, Y from now on

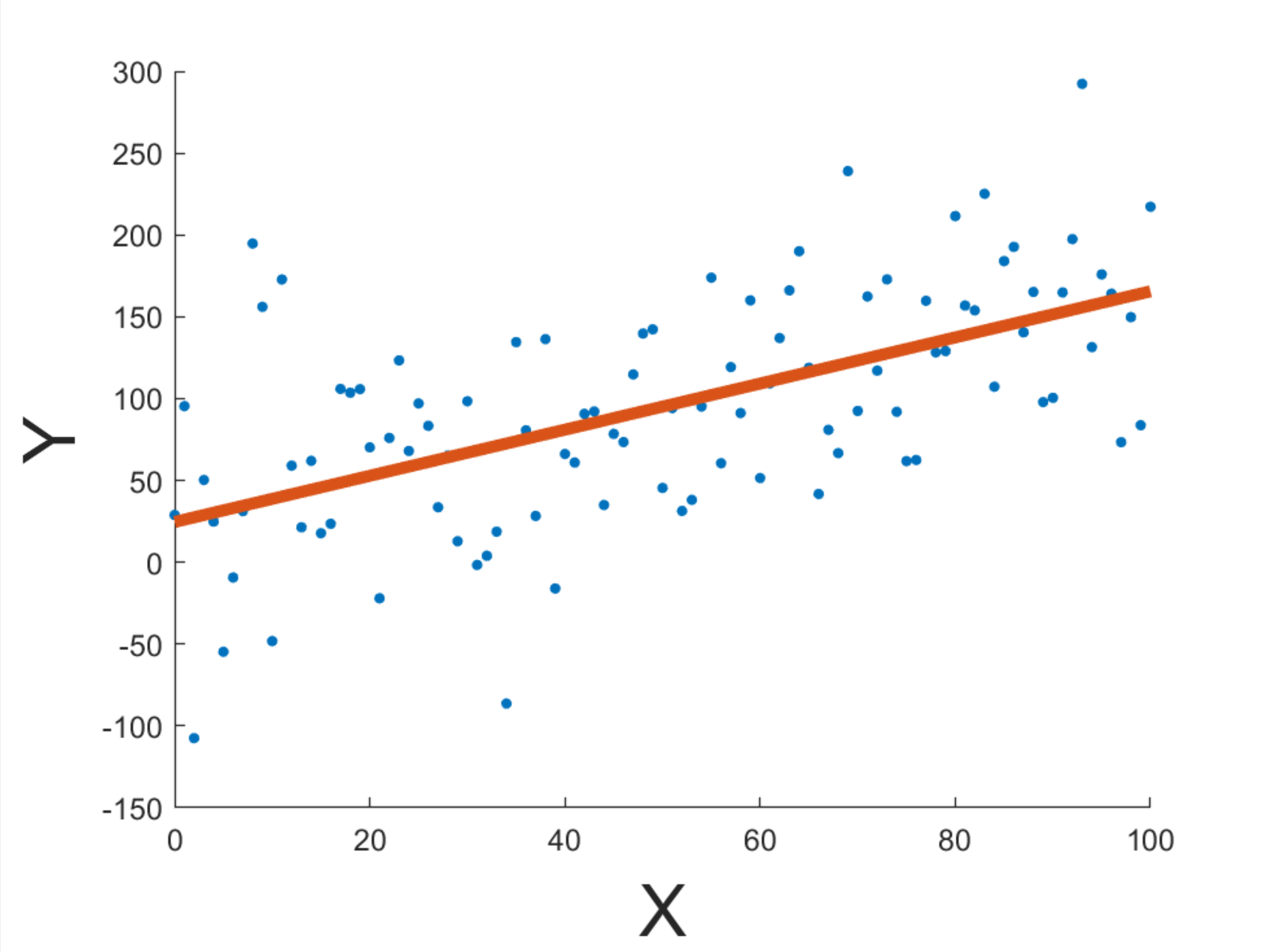
DEFINING BEST FIT

- Minimize a cost function: measures deviation of model prediction from observed data
- Example: $f(X)$ belongs to the family of straight lines

$$f(X) = aX + b$$

<u>Least-squares fit</u>

$$\min_{a,b} \sum_{i=0}^{N-1} (y_i - ax_i - b)^2$$

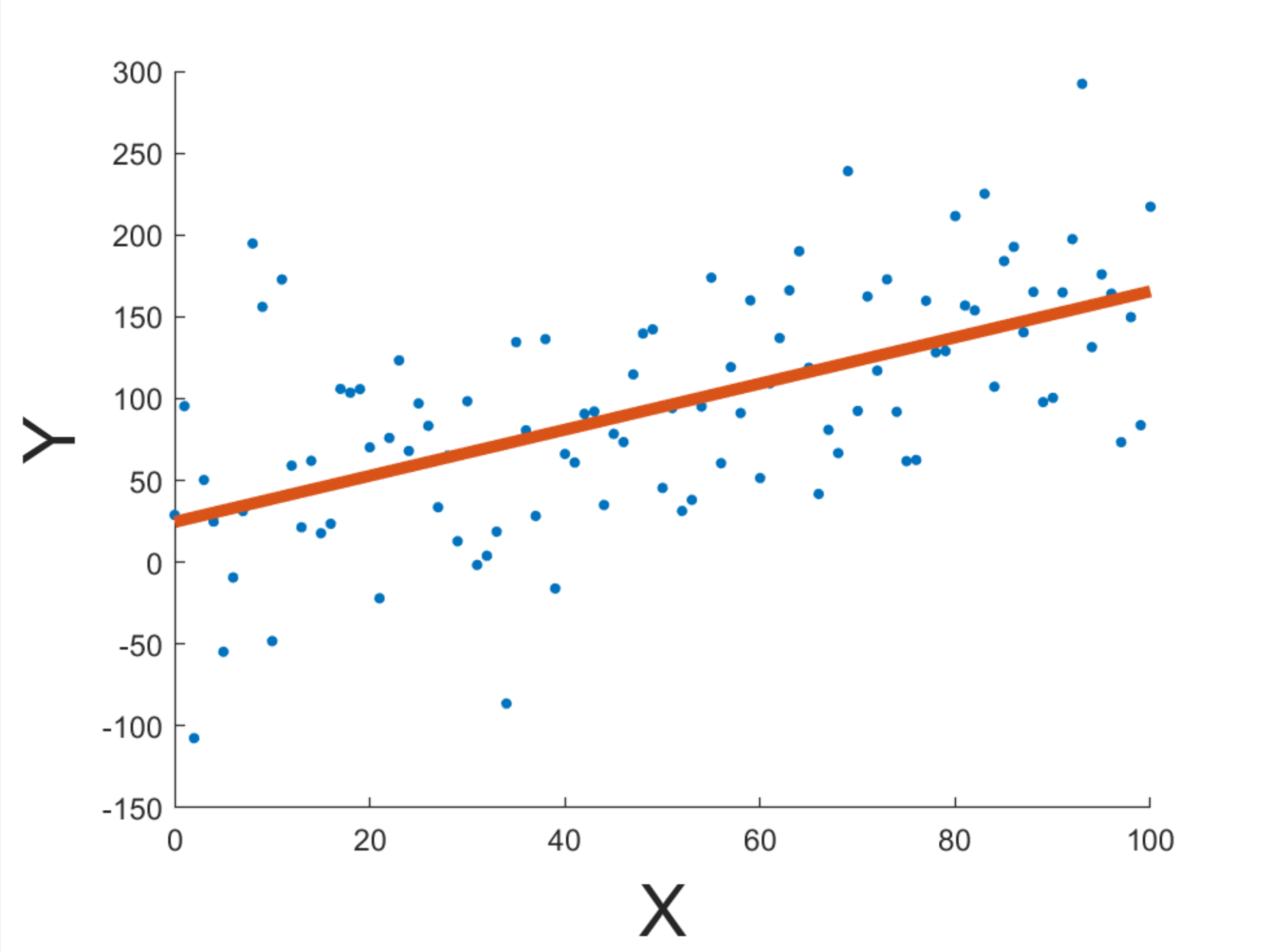


OPTIMIZATION IN STATISTICAL REGRESSION

- The least-squares procedure is grounded in probability theory
- However, its implementation requires optimization

Least-squares fit

$$\min_{a,b} \sum_{i=0}^{N-1} (y_i - ax_i - b)^2$$



LINEAR AND NON-LINEAR MODELS

Least-squares fit: general form

$$\min_{\bar{\theta}} \underbrace{\sum_{i=0}^{N-1} \left(y_i - f(x_i;\bar{\theta})\right)^2}_{\text{Sum of squared residuals}}$$

Straight line fit: $\bar{\theta} = (a, b)$ and $f\left(x;\bar{\theta}\right) = ax + b$

Linear models

$$f\left(x;\ \bar{\theta}\right) = \sum_{i=0}^{p-1} \theta_i b_i(x)$$

Straight line fit: $\theta_0 = a, \theta_1 = b, b_0(x) = x, b_1(x) = 1$

The solution to the optimization problem can be expressed algebraically

Non-linear models

(Main topic for this course)

EXAMPLE: NON-LINEAR MODEL

Quadratic chirp

$$f(x;\bar{\theta}) = A\sin(2\pi\Phi(x))$$

Instantaneous phase:

$$\Phi(x) = a_1 x + a_2 x^2 + a_3 x^3$$

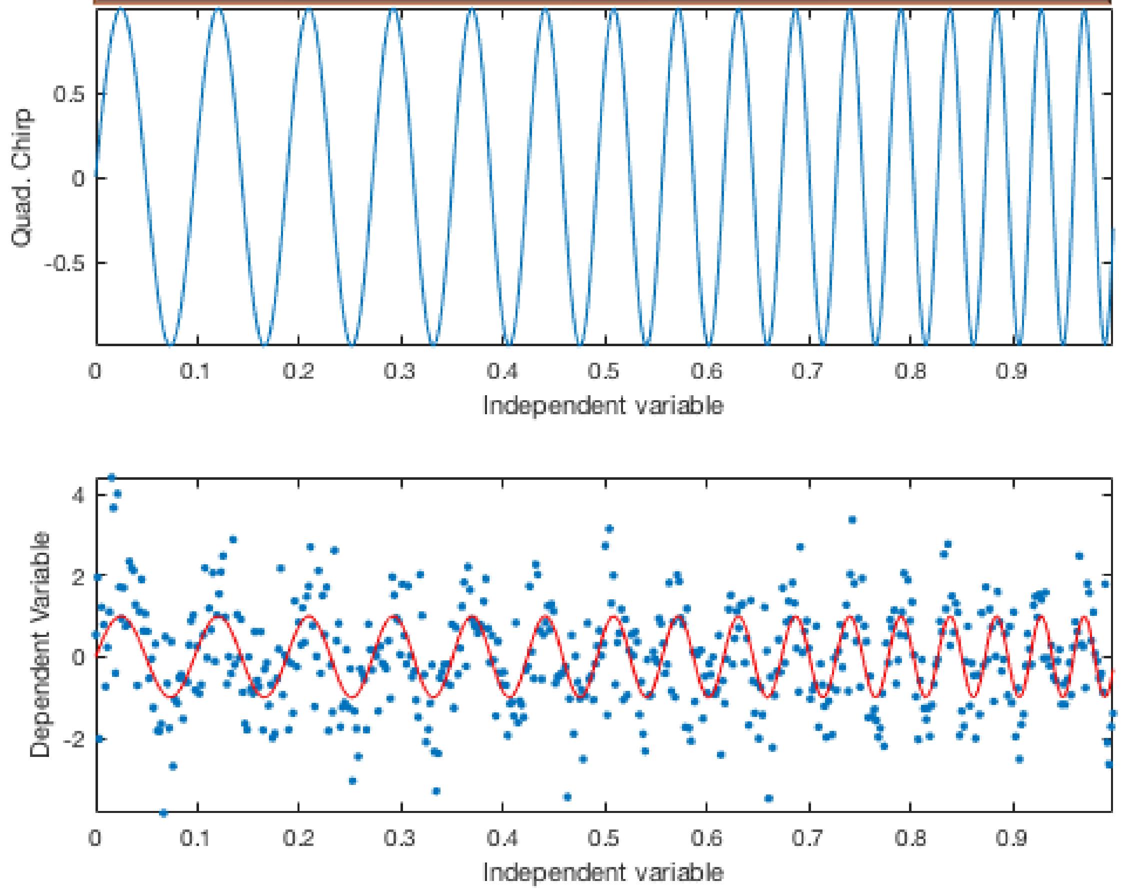
Instantaneous frequency:

$$f(x) = \frac{d\Phi}{dx}$$
$$= a_1 + 2a_2 x + 3a_3 x^2$$

(We can think of x as time t)

$f(x)$ increases with x

$1/f(x)$ (Instantaneous period) decreases with x



Data realization

PARAMETRIC REGRESSION

- Least-squares fit:

$$\min_{\bar{\theta}} \sum_{i=0}^{N-1} \left(y_i - f(x_i; \bar{\theta})\right)^2$$

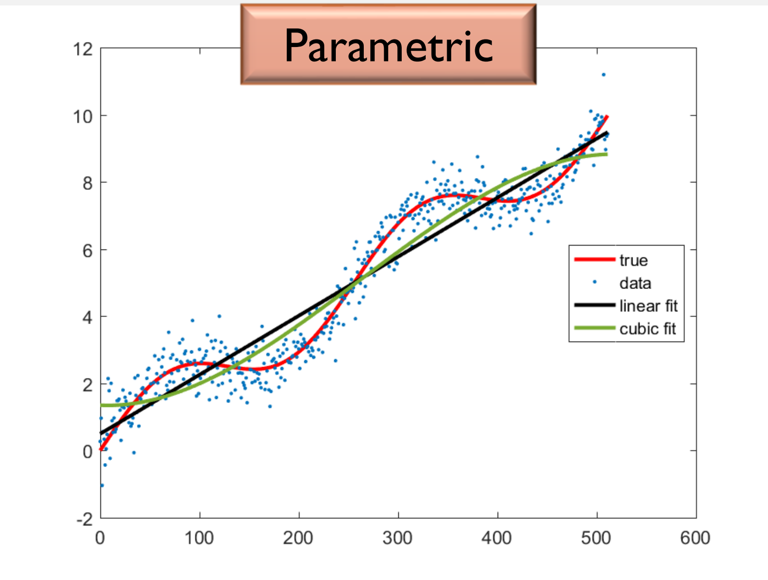
- $f(x; \bar{\theta})$ belongs to a parametric family of functions
  - Example: Straight line or quadratic chirp
- Parametric regression: Fitting parameterized models

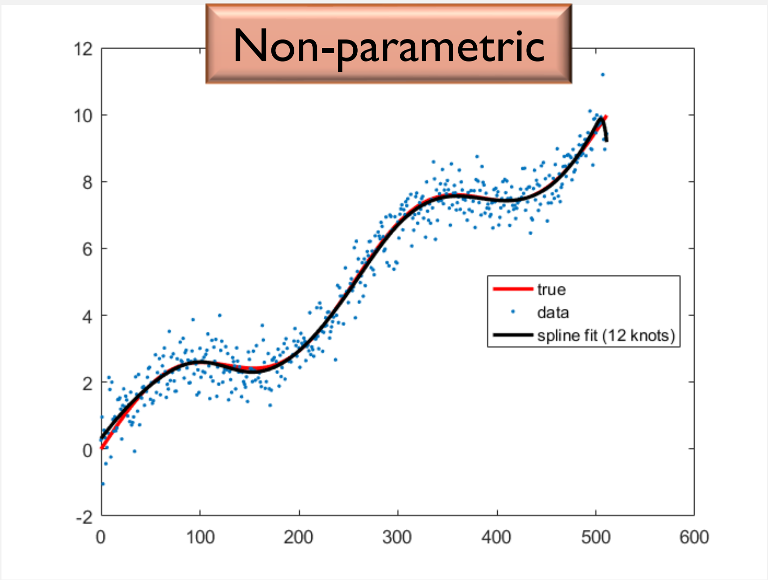
NON-PARAMETRIC REGRESSION

- Non-parametric regression: The functional form of $f(x)$ is not specified
  - No restriction: The best fit model is the data itself!
- Regularization: Broad restrictions imposed on the global properties of $f(x)$
  - Example: Smoothness
  - Regularization defines a set S of functions
  - Least-squares fit:

$$\min_{f(x)\in S} \sum_{i=0}^{N-1} (y_i - f(x_i))^2$$

- Note: Non-parametric does not mean parameter-free





REGRESSION: FIT $p_{\overline{Y}|\overline{X}}(\overline{y}|\overline{x})$

Parametric	Non-parametric
• Set of models specified in advance of data • Linear and non-linear	• Models adapt to the data • Linear and non-linear

BIG DATA AND NON-PARAMETRIC REGRESSION

Large and complex data sets in the big data era demand more flexible models

ADVANTAGES

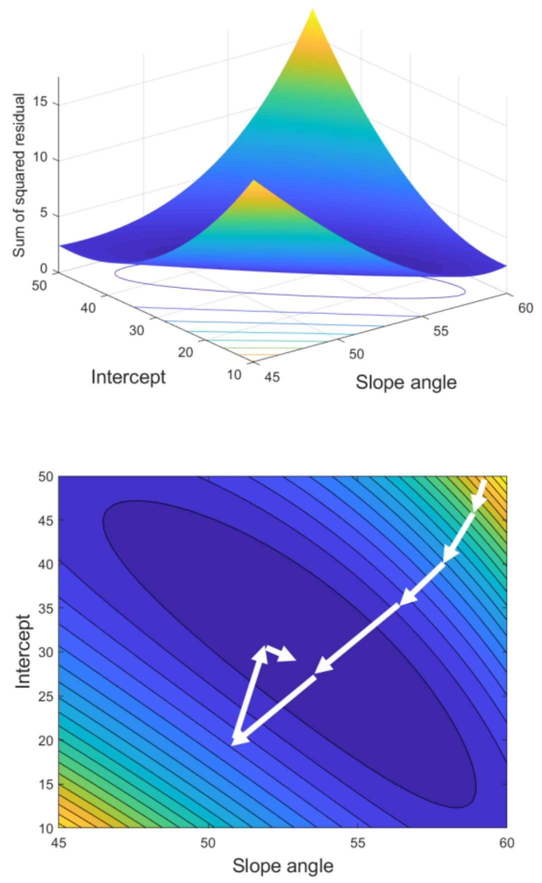
- Flexible models work better as the amount of data increases
- Growth in computing power has made non-parametric regression methods practical
  - Example: Deep artificial neural networks

DISADVANTAGES

- Large number of free parameters ⇒ Danger of overfitting ⇒ Suitable regularization needed
- Computational challenges in fitting flexible models
  - Example: Deep artificial neural networks

OPTIMIZATION: PARAMETRIC REGRESSION

Straight line fit: Sum of squared residuals



OPTIMIZATION: LINEAR MODELS

Least-squares fitting of a linear model involves minimizing a convex function

*Lecture 2

Local minimum (if it exists) is unique and is the global minimum ⇒ easy (in principle) optimization problem

Greedy methods (e.g., steepest descent) work well

OPTIMIZATION: NON-LINEAR MODEL

Quadratic chirp

$$f(x;\bar{\theta}) = A\sin(2\pi\Phi(x))$$

Instantaneous phase:

$$\Phi(x) = a_1 x + a_2 x^2 + a_3 x^3$$

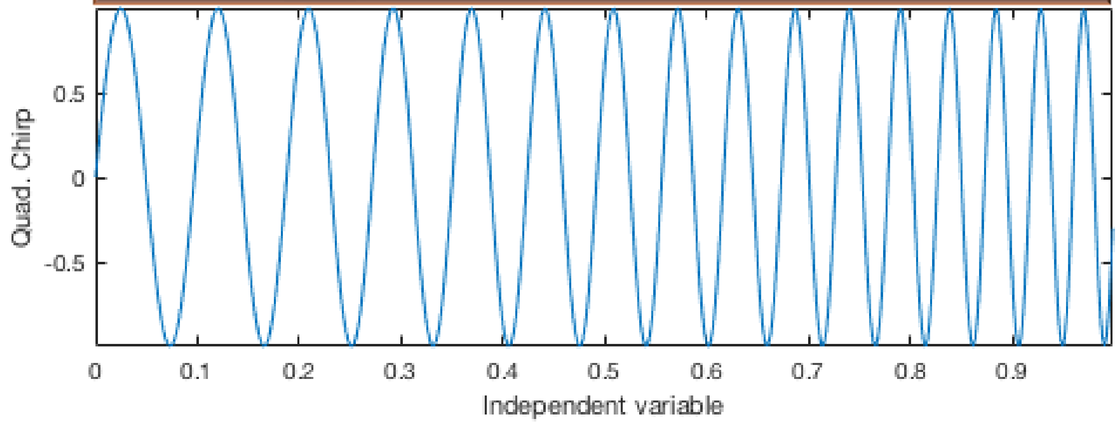
Instantaneous frequency:

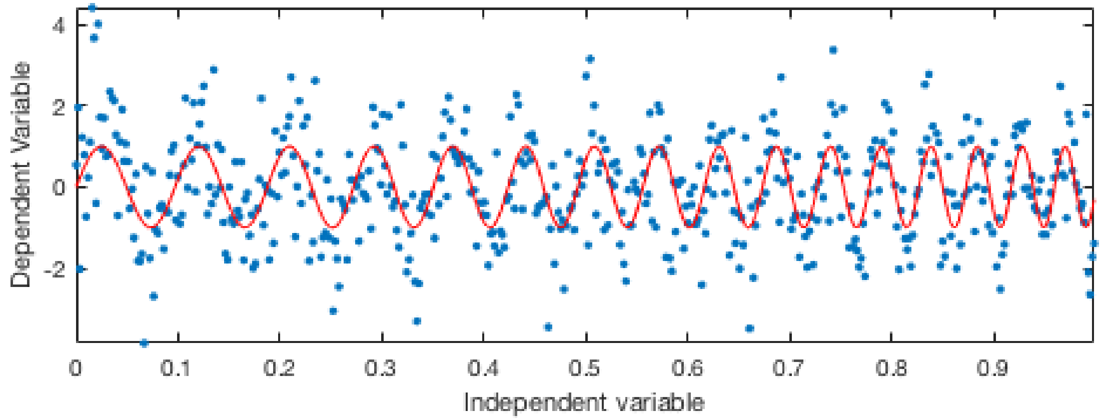
$$f(x) = \frac{d\Phi}{dx}$$
$$= a_1 + 2a_2 x + 3a_3 x^2$$

(We can think of x as time t)

$f(x)$ increases with x

$1/f(x)$ (Instantaneous period) decreases with x



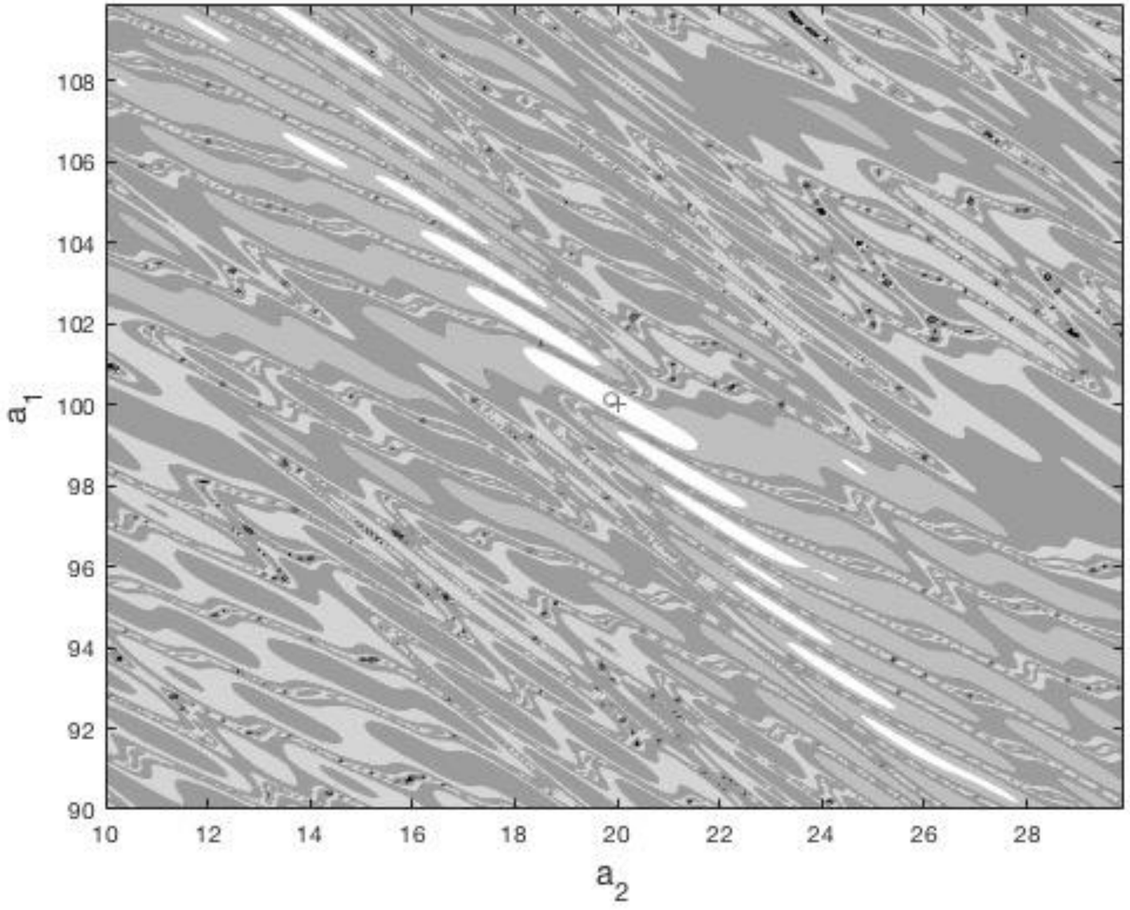


Data realization

NON-CONVEX OPTIMIZATION

- Least-squares fit of a non-linear model $\Rightarrow$ non-convex optimization problem
- Multiple local minima

Cross-sectional contours of the sum of squared residuals



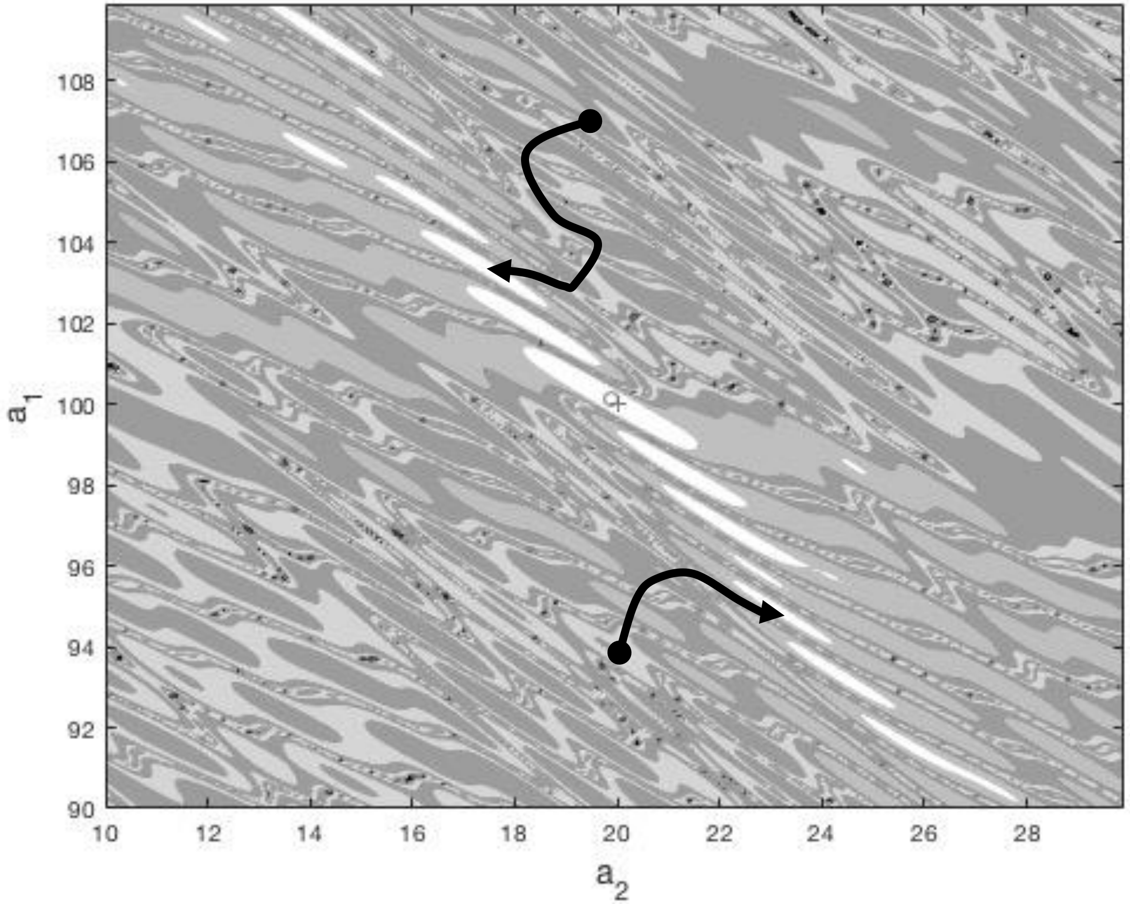
NON-CONVEX OPTIMIZATION

Local minima trap greedy algorithms

*Lecture 2

In general, deterministic algorithms for optimization must be replaced by stochastic ones

Cross-sectional contours of the sum of squared residuals



OPTIMIZATION: NON-PARAMETRIC REGRESSION

NON-PARAMETRIC REGRESSION

- Non-parametric regression: The functional form of $f(x)$ is not specified
- Regularization: Broad restrictions imposed on the global properties of $f(x)$
  - Example: Smoothness
  - Regularization defines a set S of functions
  - Least-squares fit:

$$\min_{f(x) \in S} \sum_{i=0}^{N-1} (y_i - f(x_i))^2$$

EXAMPLE: SMOOTHNESS REGULARIZATION

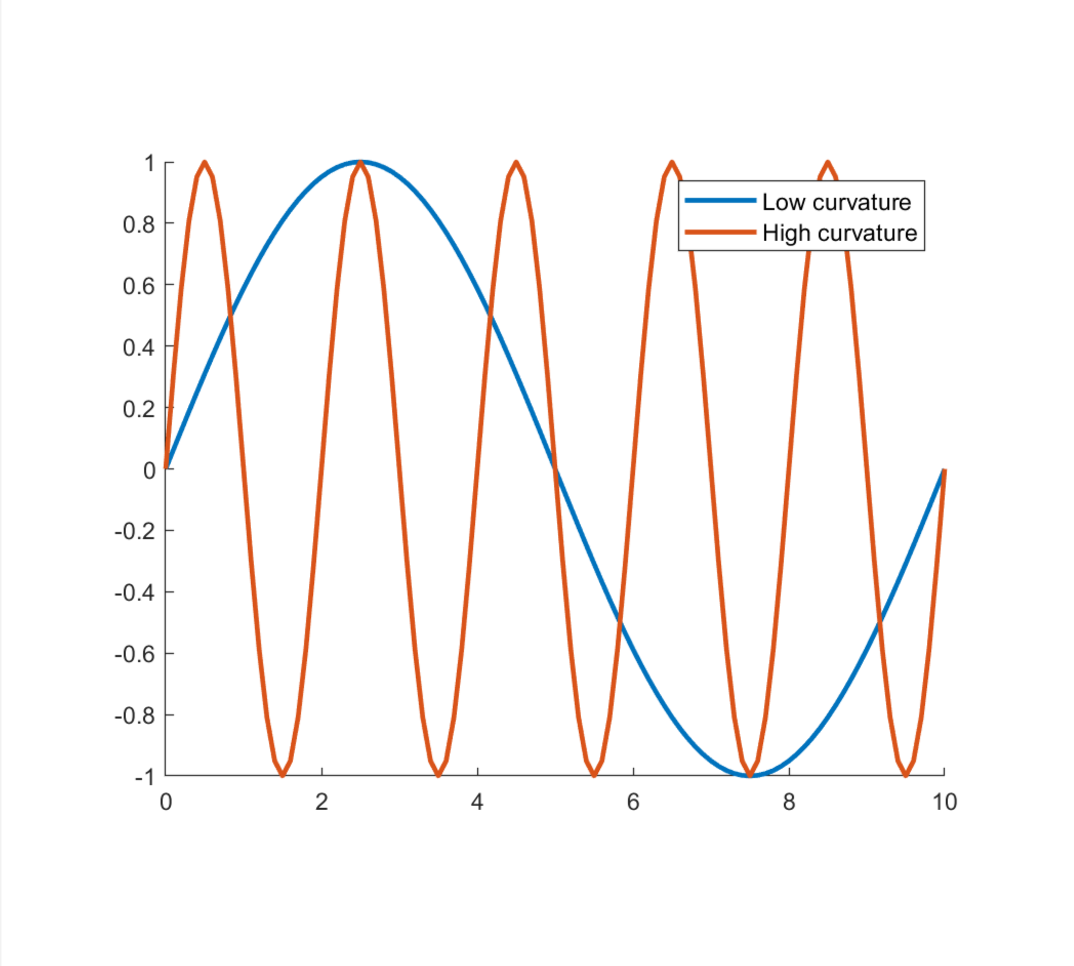
- Least-squares fit:

$$\min_{f(x) \in S} \sum_{i=0}^{N-1} (y_i - f(x_i))^2$$

- Enforce smoothness of $f(x)$
- Limit the average absolute curvature

$$\frac{1}{(b-a)} \int_a^b dx \left(\frac{d^2 f}{dx^2} \right)^2$$

- Solution: $f(x)$ must belong to the family of cubic splines

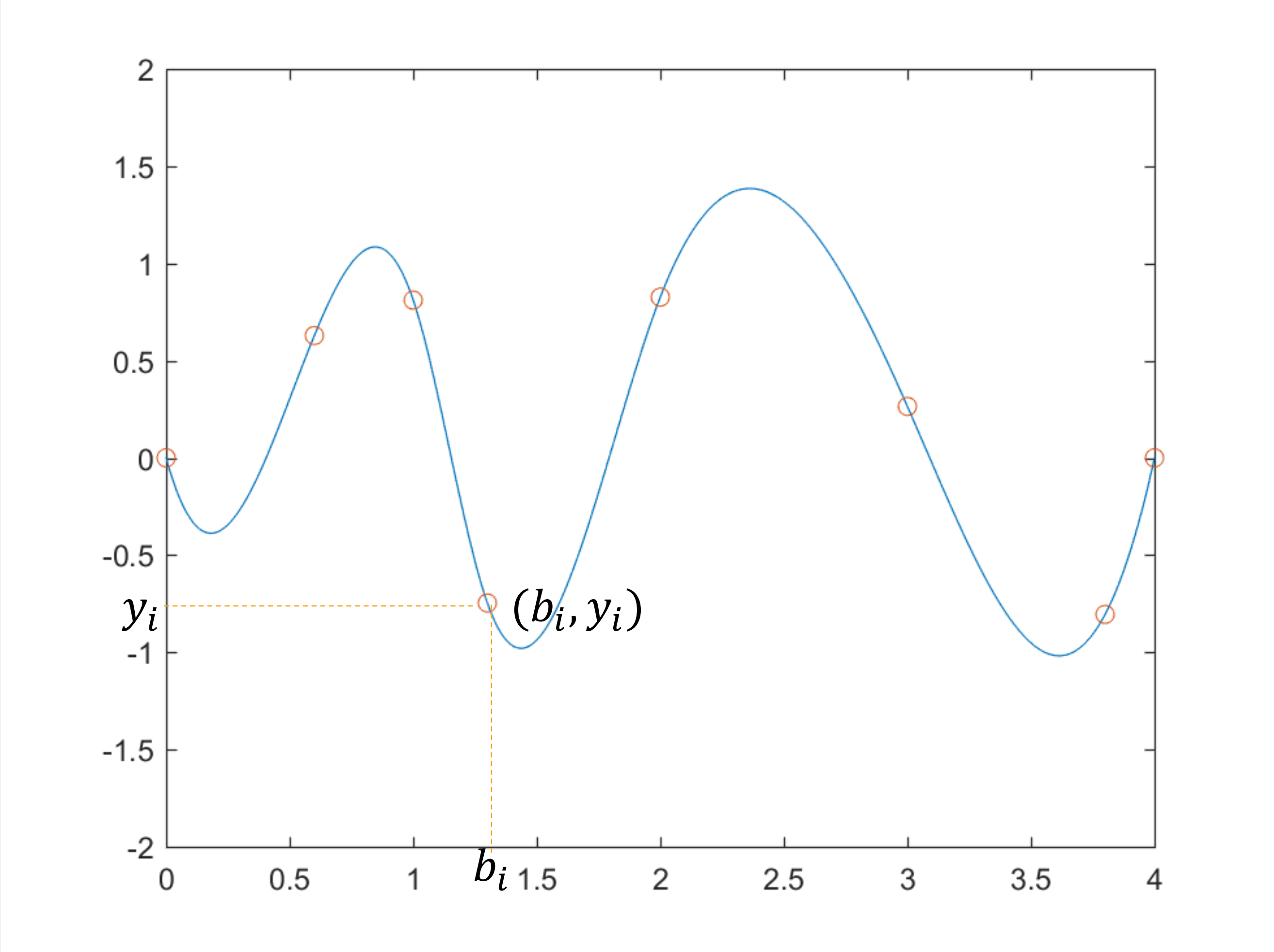


SPLINE: REFRESHER

- A spline is a piecewise polynomial function that interpolates:

  $\{(b_i, y_i)\}; i = 0, 1, \dots, M-1$

- $\{b_0, b_1, \dots, b_{M-1}\}$: Set of breakpoints
- $\{y_0, y_1, \dots, y_{M-1}\}$: Set of data values at breakpoints

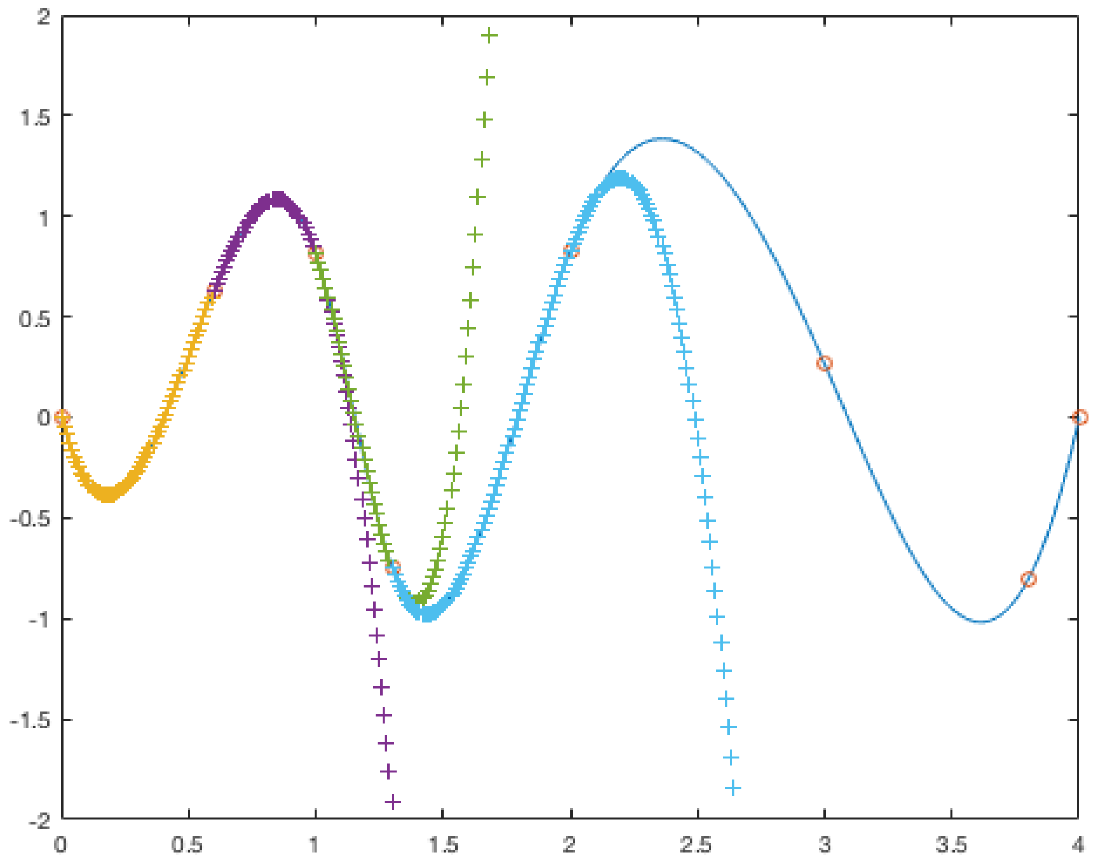


SPLINE: REFRESHER

- A spline is a piecewise polynomial function that interpolates:

$$\{(b_i, y_i)\}; i = 0, 1, \dots, M-1$$

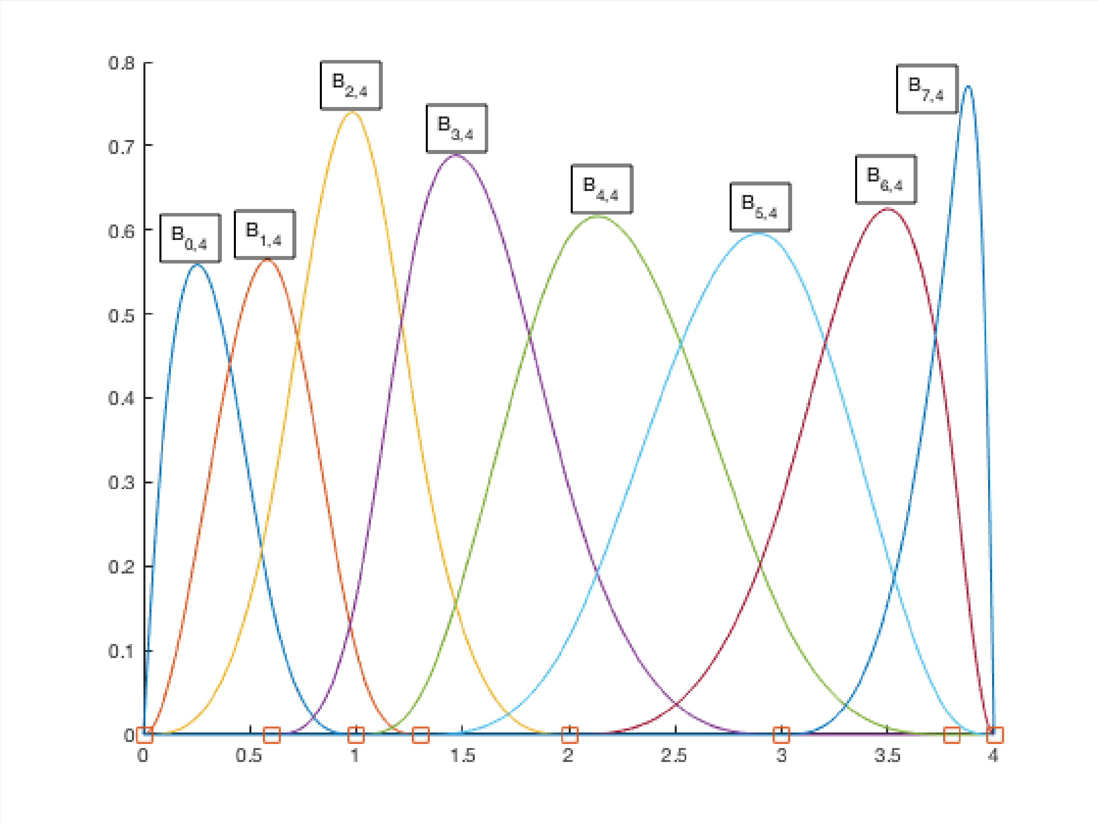
- $\{b_0, b_1, \dots, b_{M-1}\}$: Set of <u>breakpoints</u>
- $\{y_0, y_1, \dots, y_{M-1}\}$: Set of data values at breakpoints

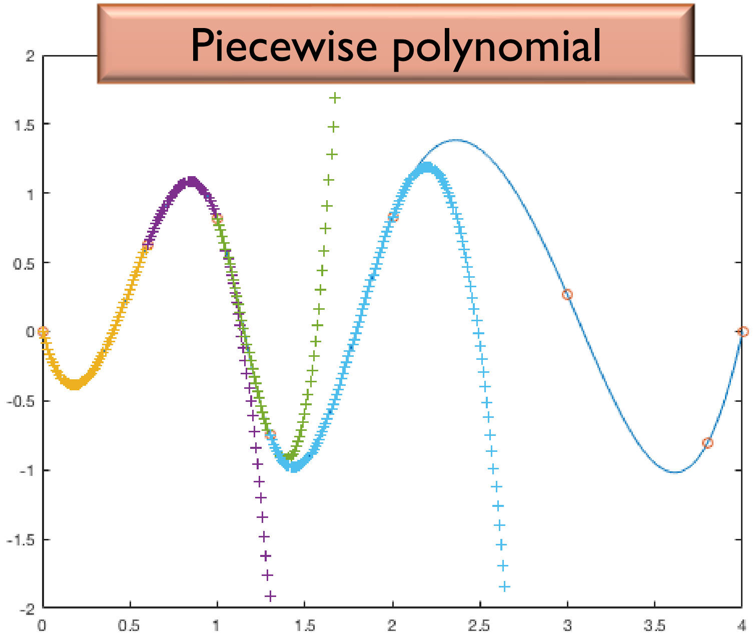


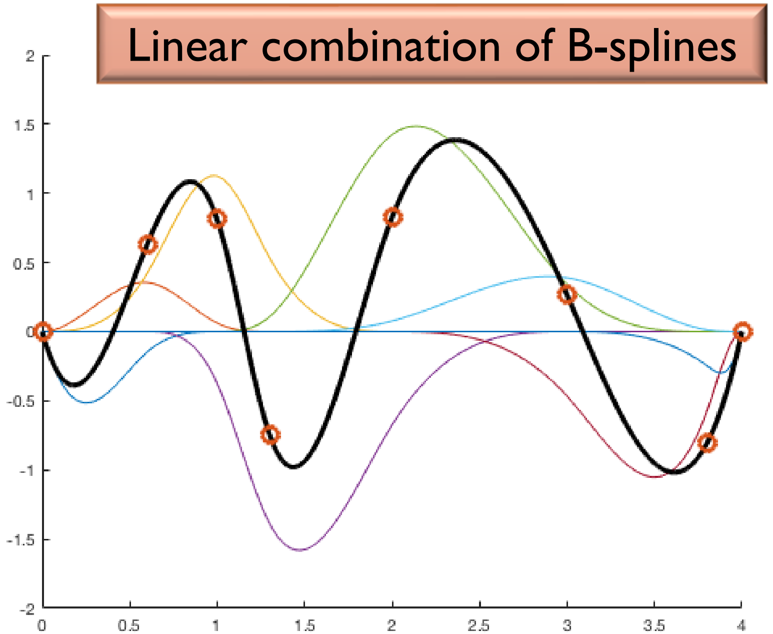
BASIS SPLINES (B-SPLINES)

- All splines of a given polynomial order, and given breakpoint sequence, form a linear vector space
- B-spline functions provide a convenient basis for this space

B-splines of order 4 for a given set of breakpoints







SPLINE REPRESENTATIONS

SPLINE SMOOTHING

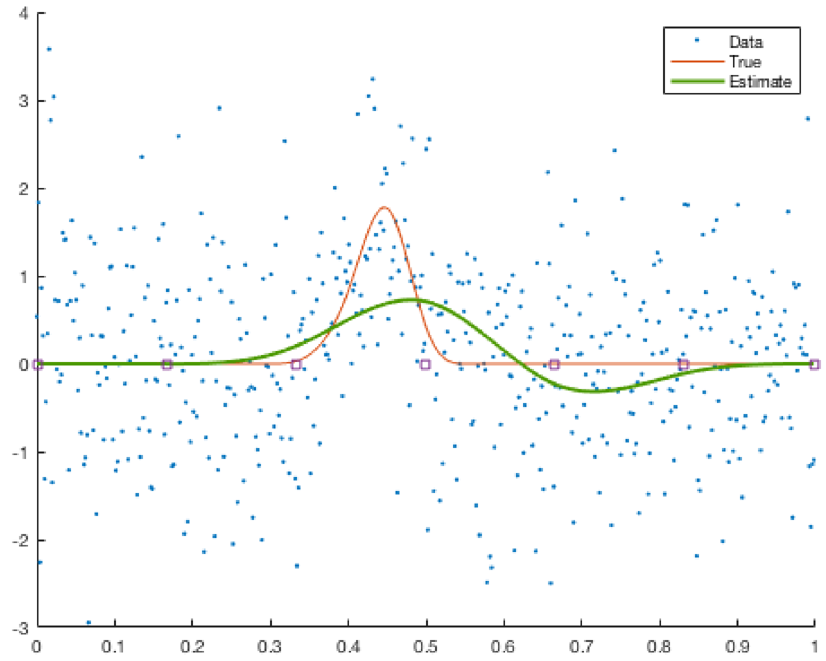
- Fixed number (M) and location of breakpoints $(\bar{b})$

$$f(x;\ \bar{\alpha}) = \sum_{j=0}^{M-1} \alpha_j B_{j,4}(x; \bar{b})$$

- Least-squares:

$$\min_{\bar{\alpha}} \sum_{i=0}^{N-1} \left(y_i - f(x_i;\ \bar{\alpha})\right)^2$$

- Optimization: Linear model



REGRESSION SPLINE

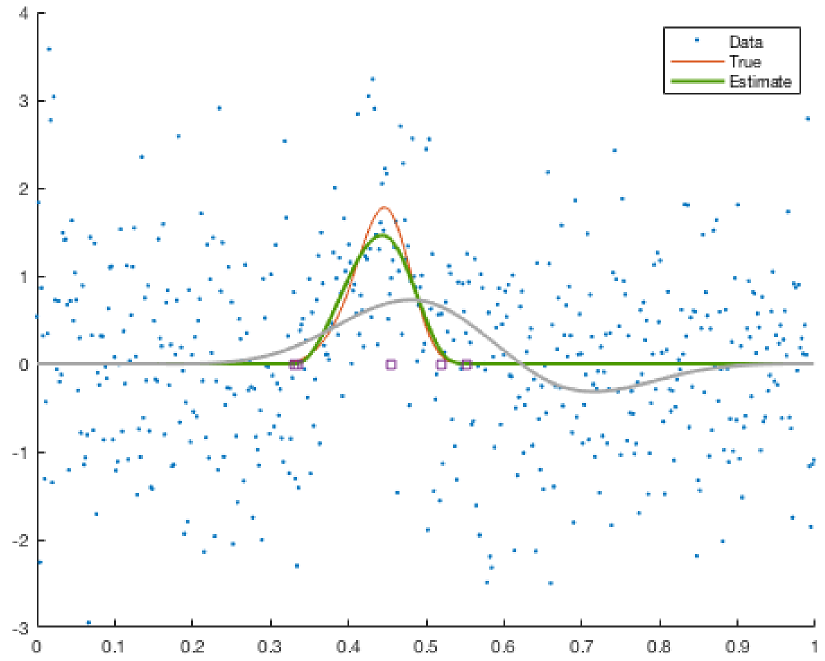
- Fixed number (M) but not fixed locations of breakpoints $(\bar{b})$

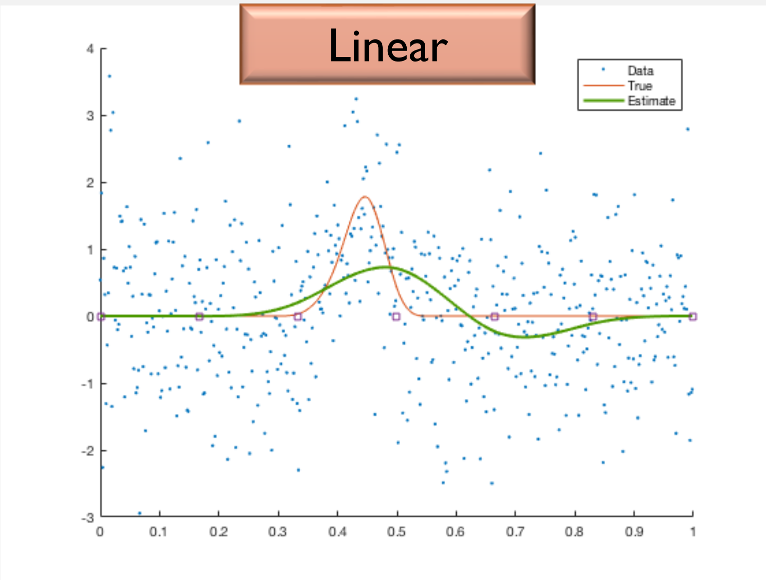
$$f(x;\ \bar{\alpha}, \bar{b}) = \sum_{j=0}^{M-1} \alpha_j B_{j,4}(x; \bar{b})$$

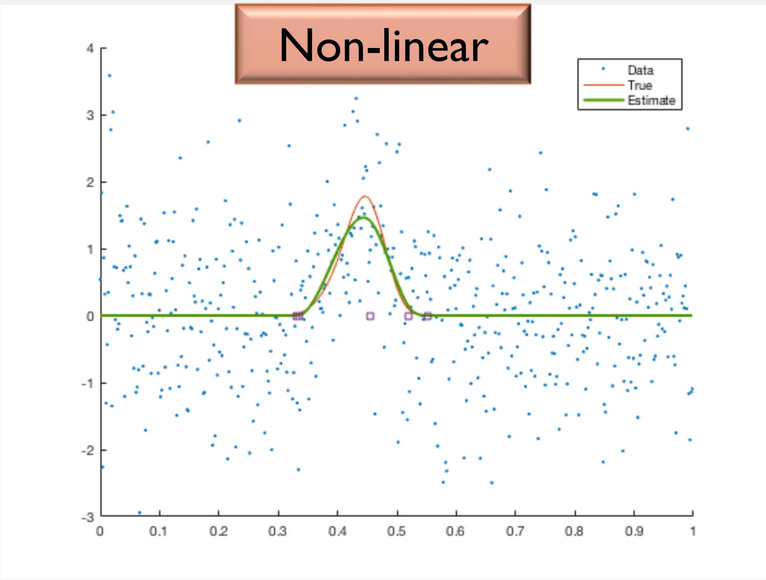
- Least-squares:

$$\min_{\bar{\alpha}, \bar{b}} \sum_{i=0}^{N-1} \left(y_i - f(x_i;\ \bar{\alpha})\right)^2$$

- Optimization: Non-linear model



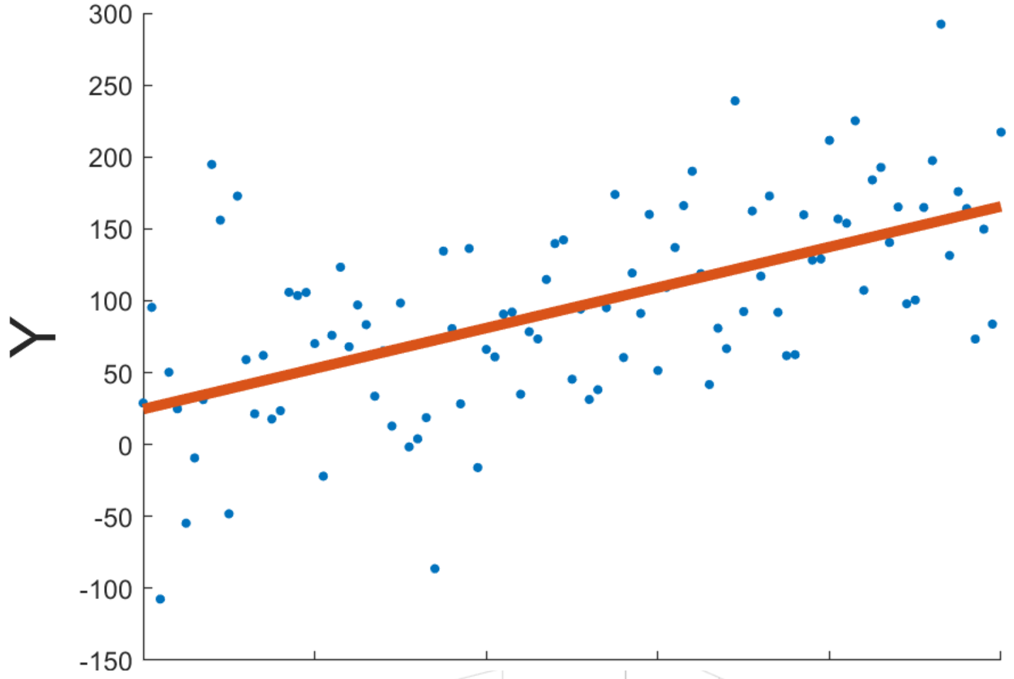


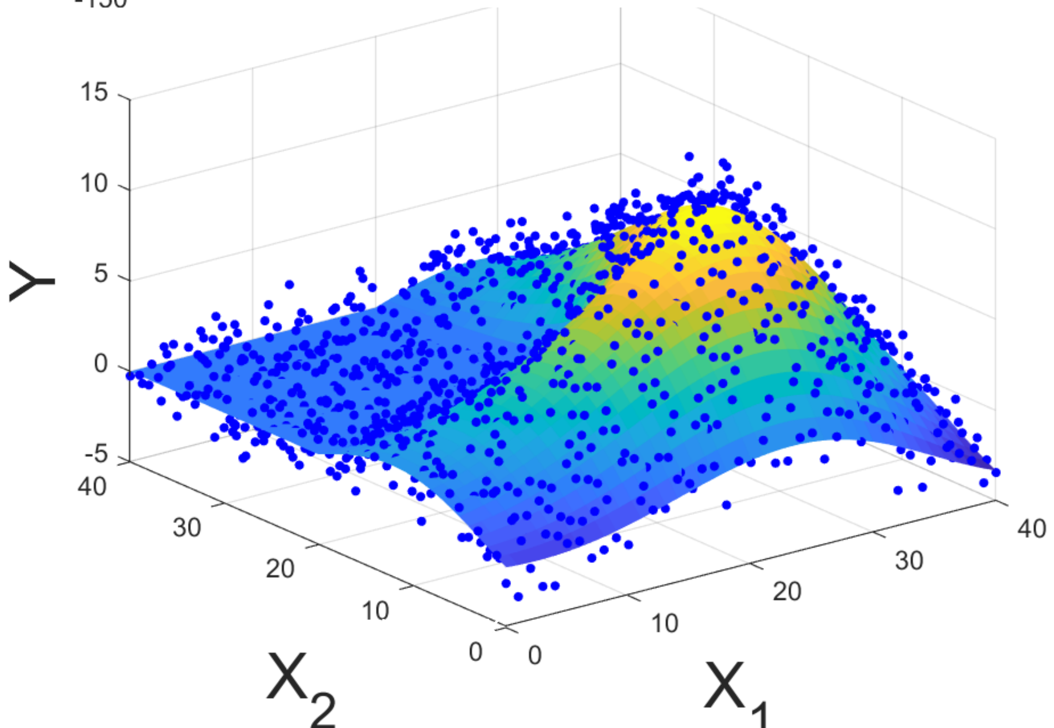


LINEAR VS NON-LINEAR MODELS

Linear	Non-linear
• Computation time: ≈ 0.1 sec • Optimization: Simple (matrix algebra)	• Computation time: ≈ 3 sec (with 4 parallel workers) • Optimization: Difficult (Swarm intelligence)

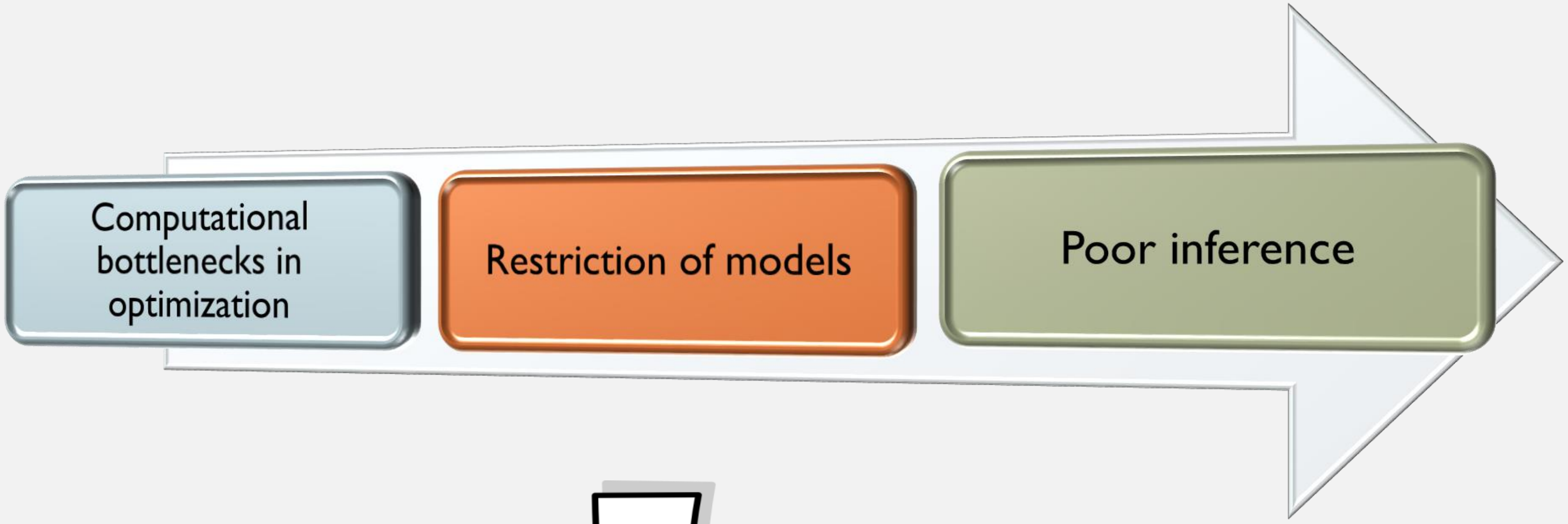
IMPORTANCE OF OPTIMIZATION IN REGRESSION

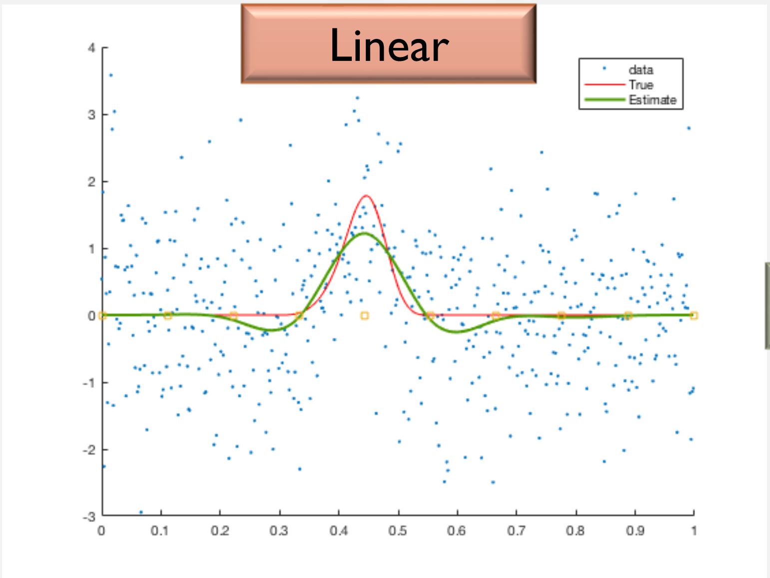


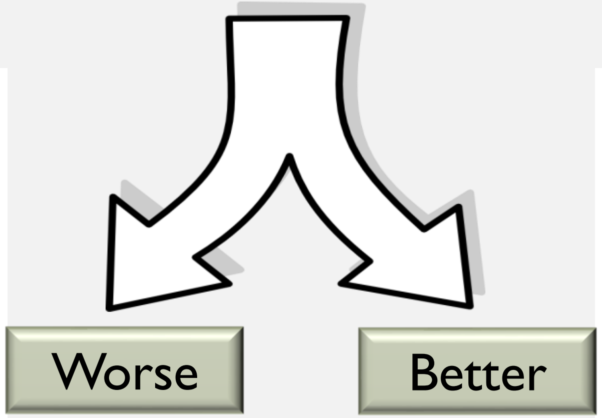


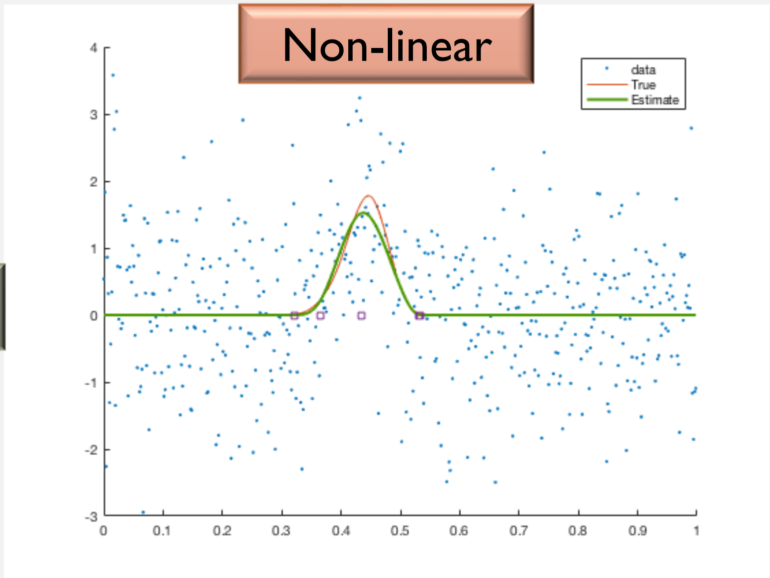
OPTIMIZATION IN REGRESSION

- Fitting requires minimization of some cost function
  - Example: Least squares
- Hence optimization is a core task in statistical regression


Computational bottlenecks in optimization
Restriction of models
Poor inference

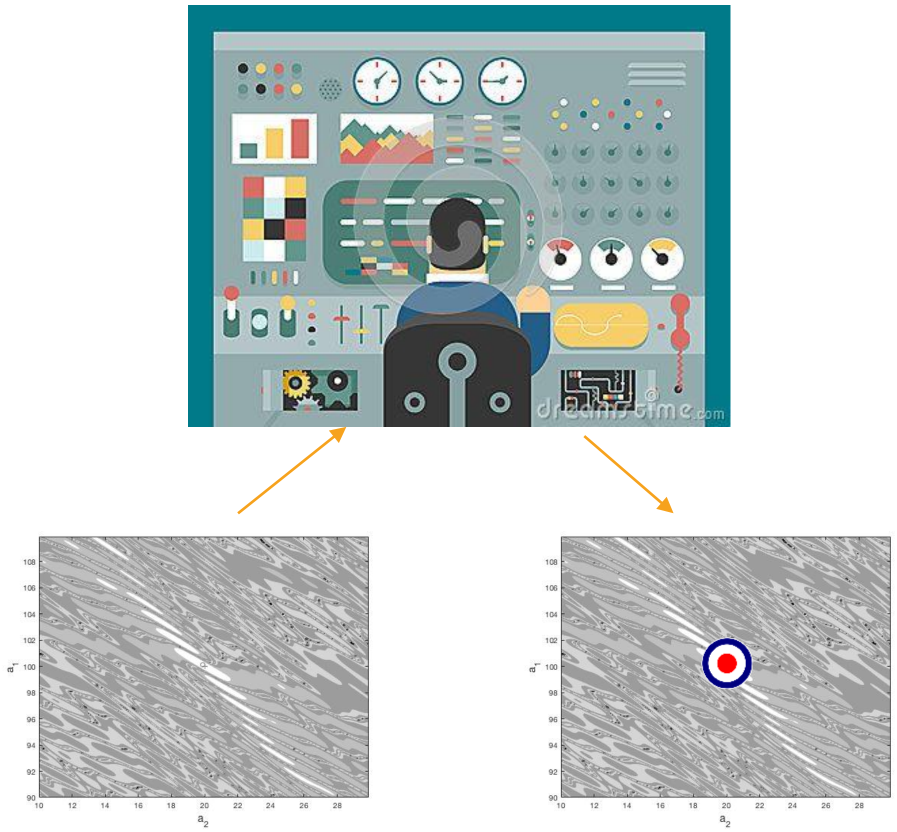

Linear
data
True
Estimate


Worse
Better


Non-linear
data
True
Estimate

BARRIERS

- Optimization methods that can handle difficult problems cannot, in general, be used as black-boxes
- Some tuning of these methods is always needed in order to extract good performance from them

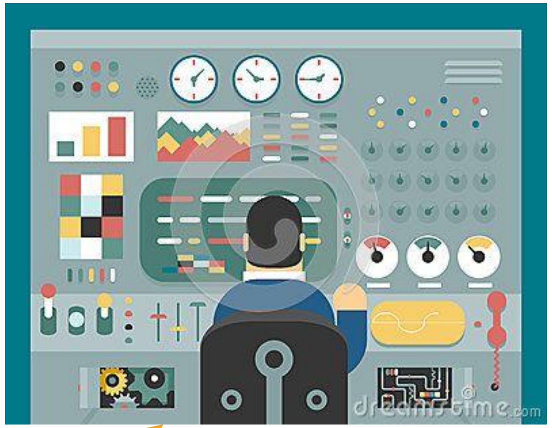


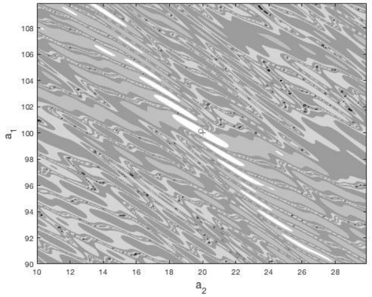
Non-convex, high-dimensional Fitness function

Global minimum

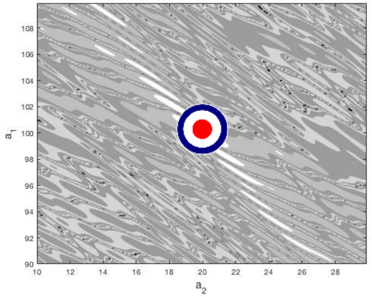
BARRIERS

- Tuning ⇒ (Sometimes considerable) expertise required in using optimization methods
- As a result, each application area often uses just a few optimization approaches
  - Example: Markov Chain Monte Carlo (MCMC) in Bayesian analysis





Non-convex, high-dimensional Fitness function

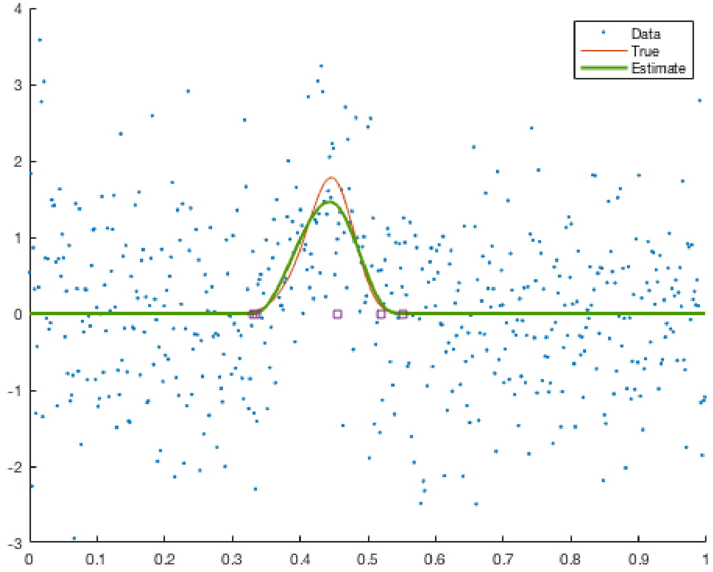


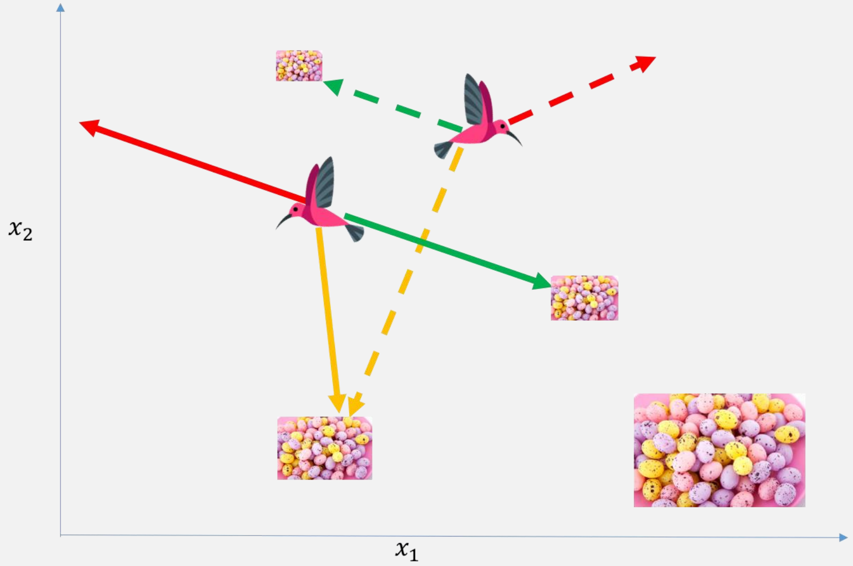
Global minimum

SWARM INTELLIGENCE

- SI: A relatively new approach in statistical analysis problems
- Example:
  - Breakpoint (Knot) optimization in regression spline is an old problem (e.g., Jupp, 1978)
  - SI method used relatively recently (Galvez, Iglesias, 2011; Mohanty, 2012)

Particle Swarm Optimization (PSO) for breakpoint optimization
Details: Lectures 2 & 3





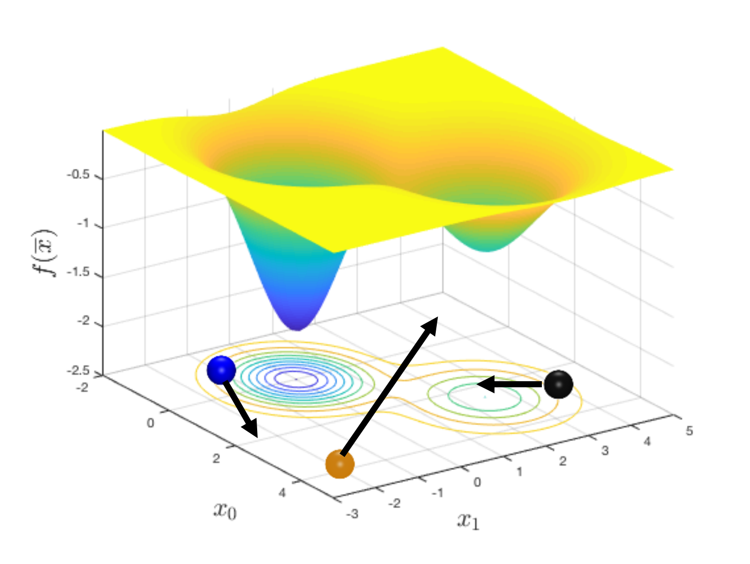
PARTICLE SWARM OPTIMIZATION

- Introduced by Kennedy & Eberhart, 1995
- A swarm intelligence method inspired by the flocking behavior of birds
  - Flocking: more efficient food search (?), predator avoidance (?)
- Model: a bird moves under random attraction towards the best food sources that it and the flock have found

PSO SCHEMATIC

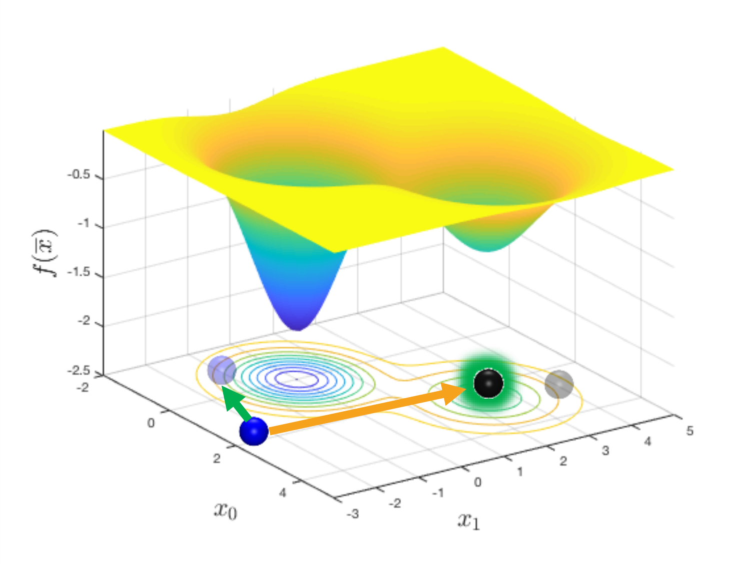
Basic setup

- Multiple agents (“particles”) moving in the search space with different “velocities”



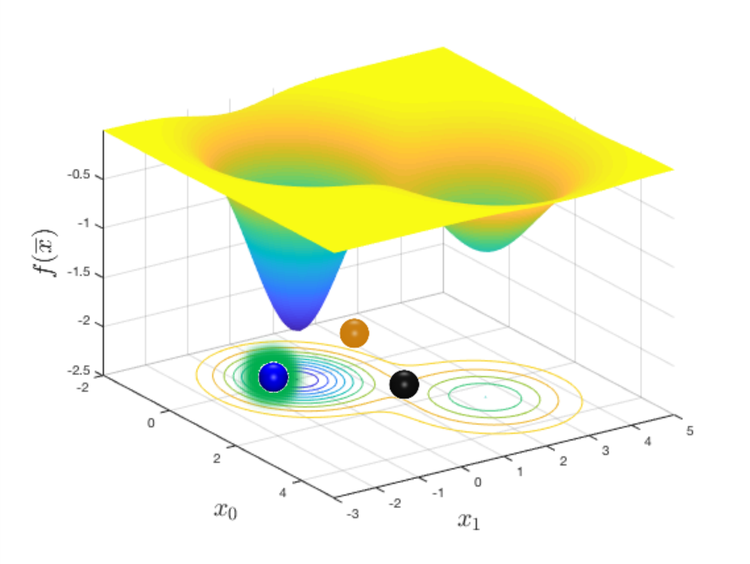
Velocity update

- Randomized acceleration towards the best agent and best location in the particle’s history + original velocity (“inertia”)



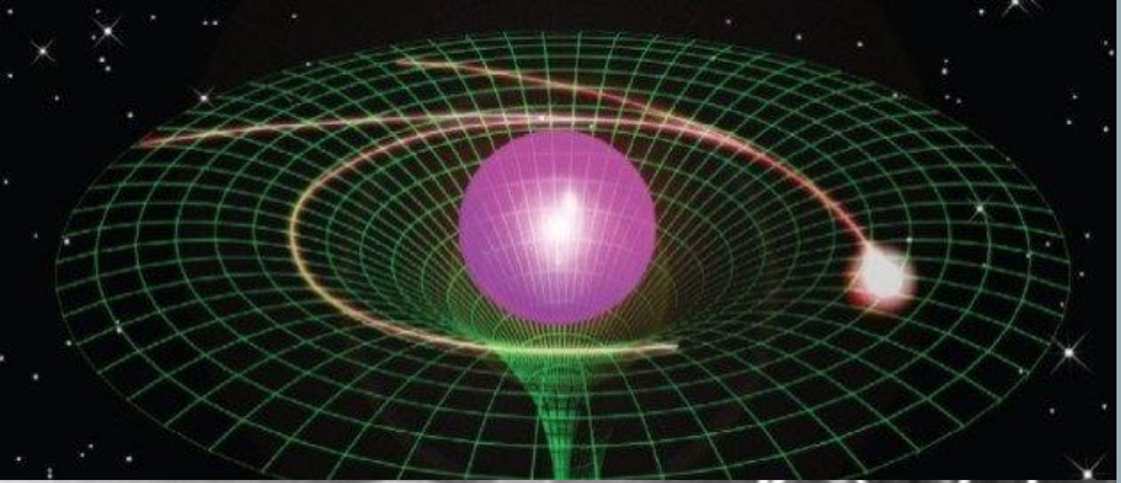
Position update

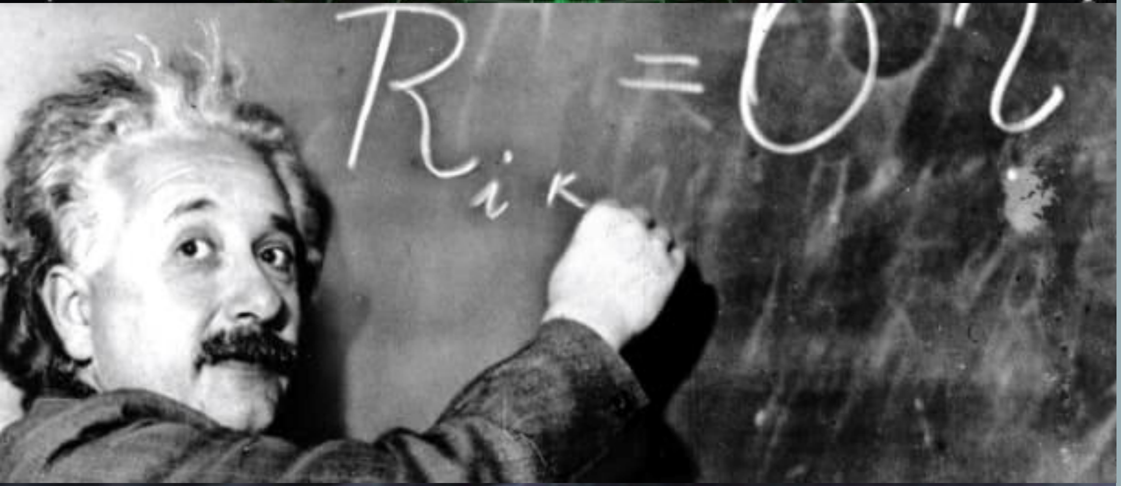
- Particles move to new positions

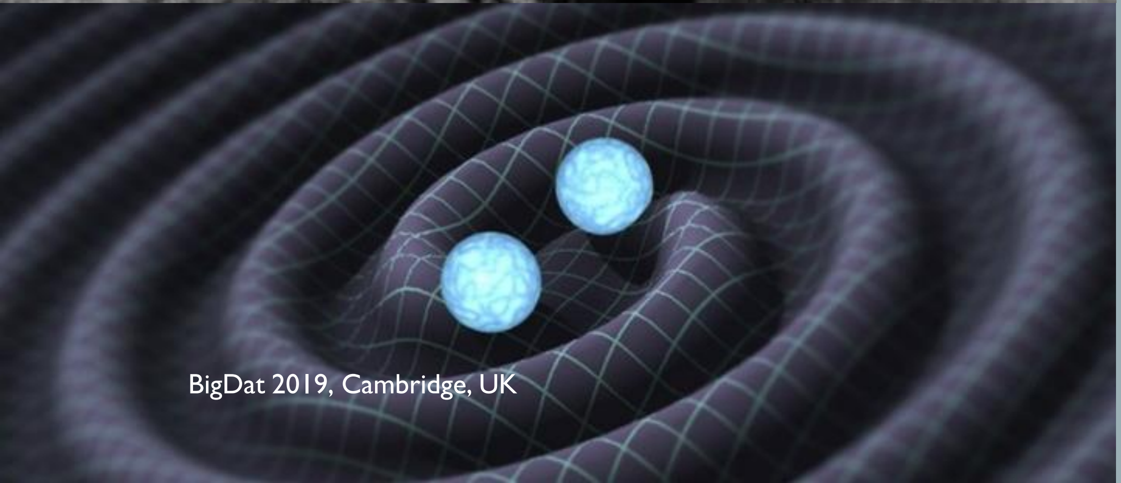


PSO SUCCESS STORIES

Applications in gravitational wave astronomy







GRAVITATIONAL WAVE ASTRONOMY

- **Einstein's general theory of relativity**: Gravitation is a manifestation of curved space-time geometry
- **Gravitational waves**: Time-dependent changes in mass-energy distributions produce ripples in space-time geometry
- **Gravitational wave astronomy**: Study of extreme systems by observing their gravitational wave emission

GW150914: FIRST DISCOVERY

2015:

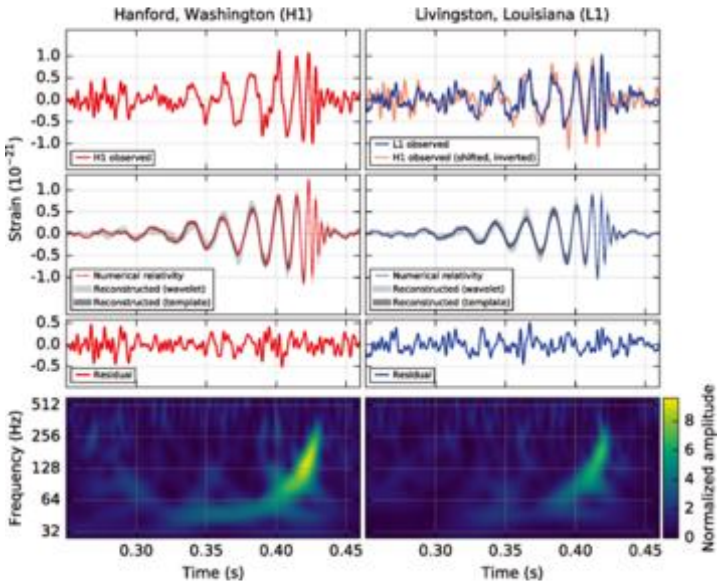
- Signal came from a binary system with two black holes
- Almost 3 times the mass of the Sun converted to energy in gravitational waves
- Outshone the entire universe in terms of power radiated!

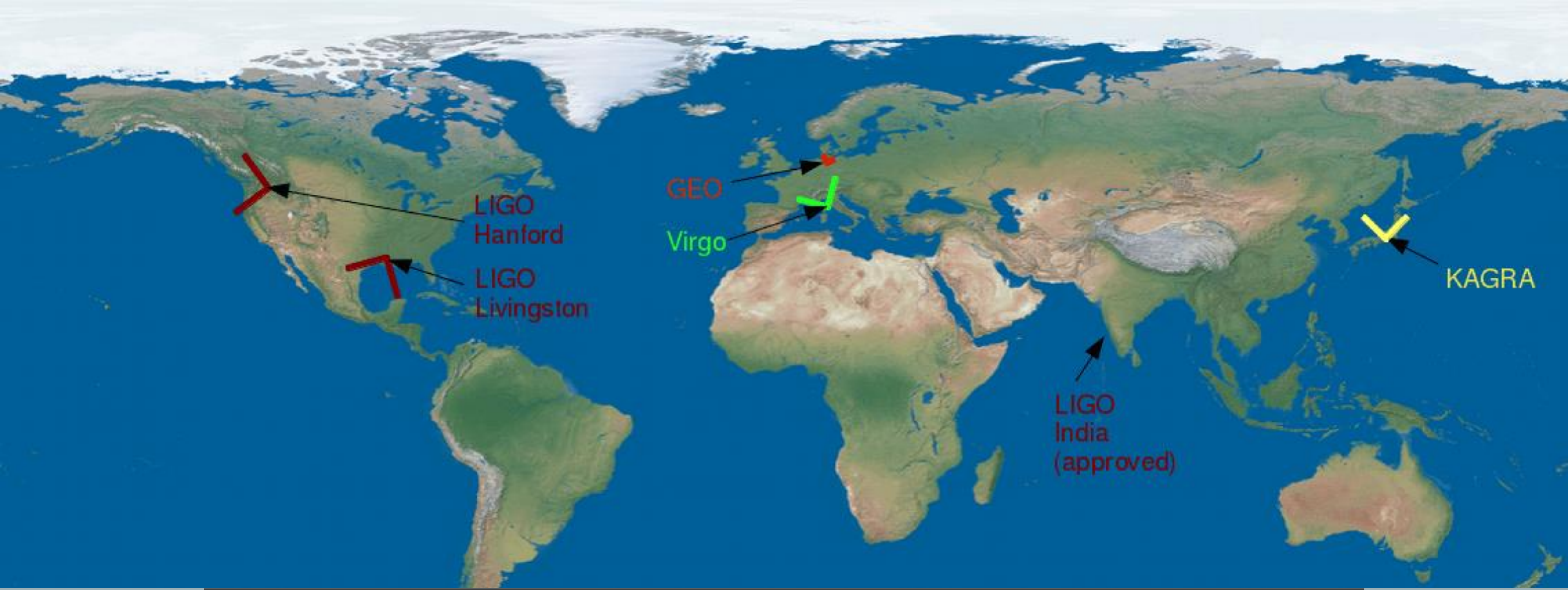
2017:

- Nobel prize in Physics (Barish, Thorne, Weiss)
- More signals (11) detected since GW150914

LIGO Hanford, WA

LIGO Livingston, LA





WORLDWIDE NETWORK OF GW DETECTORS

DATA ANALYSIS IN GW ASTRONOMY

GW astronomy is critically dependent on data analysis for extracting weak signals from instrumental noise background

PARAMETRIC REGRESSION

- Example: Signals from binary systems
- Signal shape depends on the parameters of the system
  - Sky location, masses, distance, orbital inclination,…
- Regression: non-linear model
- Optimization is computationally expensive

NON-PARAMETRIC REGRESSION

- Example: Core-collapse in supernovae
- Signal shape not known (or fundamentally unpredictable)
  - Simulations can produce plausible shapes
- Regression:
  - Linear model: Signal is a linear combination of wavelets
  - Non-linear models (less explored)

BINARY INSPIRAL SEARCH

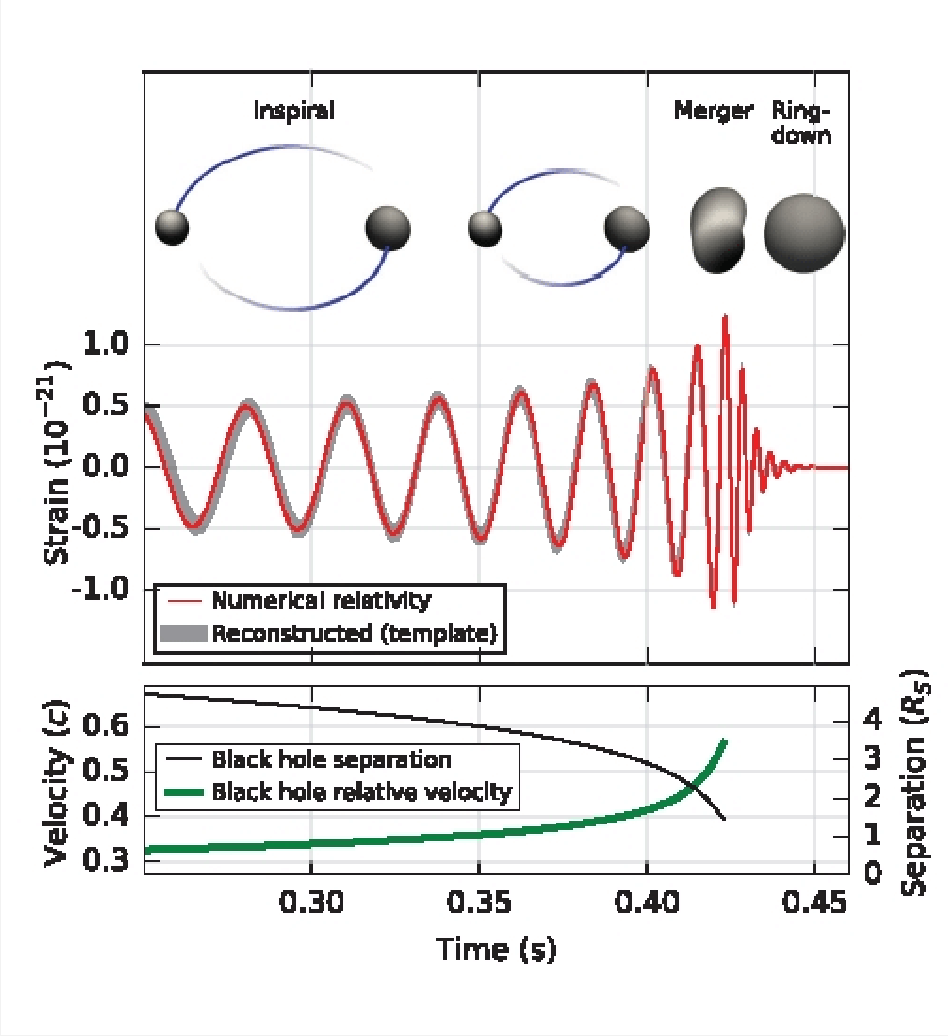
Least-squares with non-linear model:

$$\min_{\bar{\theta}} \sum_{i=0}^{N-1} \left(y_i - f(x_i;\bar{\theta})\right)^2$$

*Likelihood ratio for Gaussian iid noise

Signal ($f\left(x_i;\bar{\theta}\right)$) is predictable

$\bar{\theta}$ = [mass of each component,
sky location,
spin of each component,
orbit orientation in space, …]



BINARY INSPIRAL SEARCH

Brute force numerical optimization: $\approx 10^8$ evaluations of the sum of squared residuals (each evaluation: $\approx 10^7$ floating point operations)

Computational bottleneck $\Rightarrow$ current searches follow a sub-optimal approach $\Rightarrow$ Lower sensitivity $\Rightarrow$ Reduced rate of detections

Binary inspiral: Log-likelihood ratio (2 parameters)

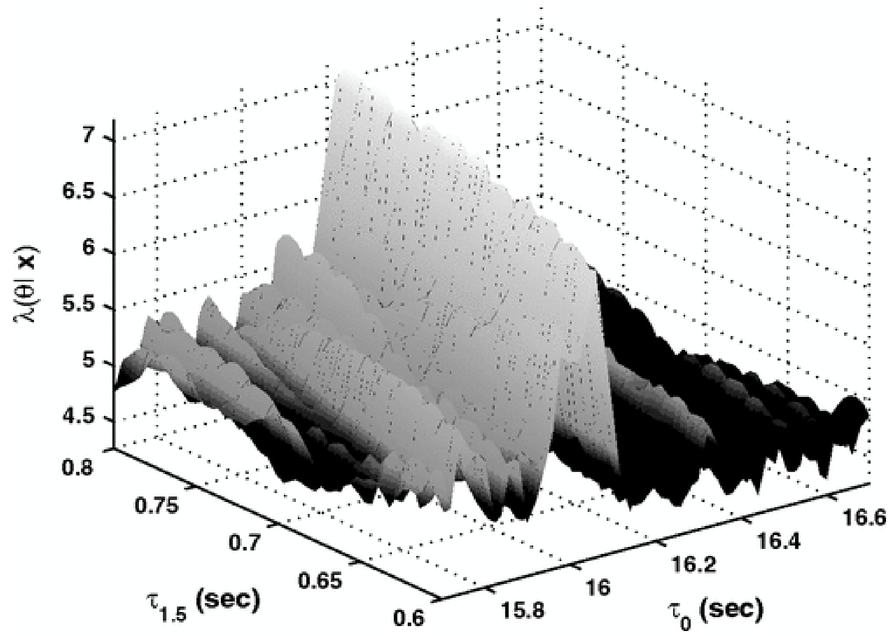
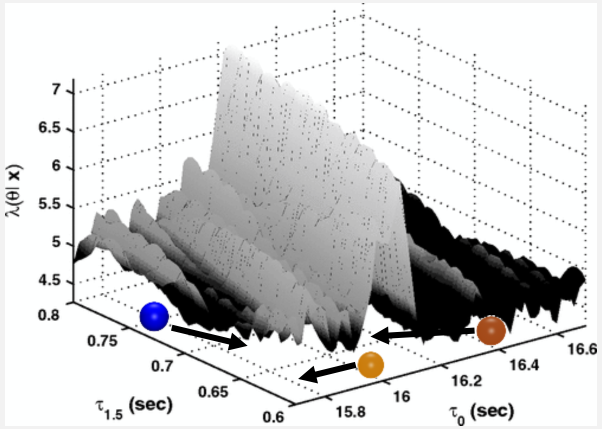
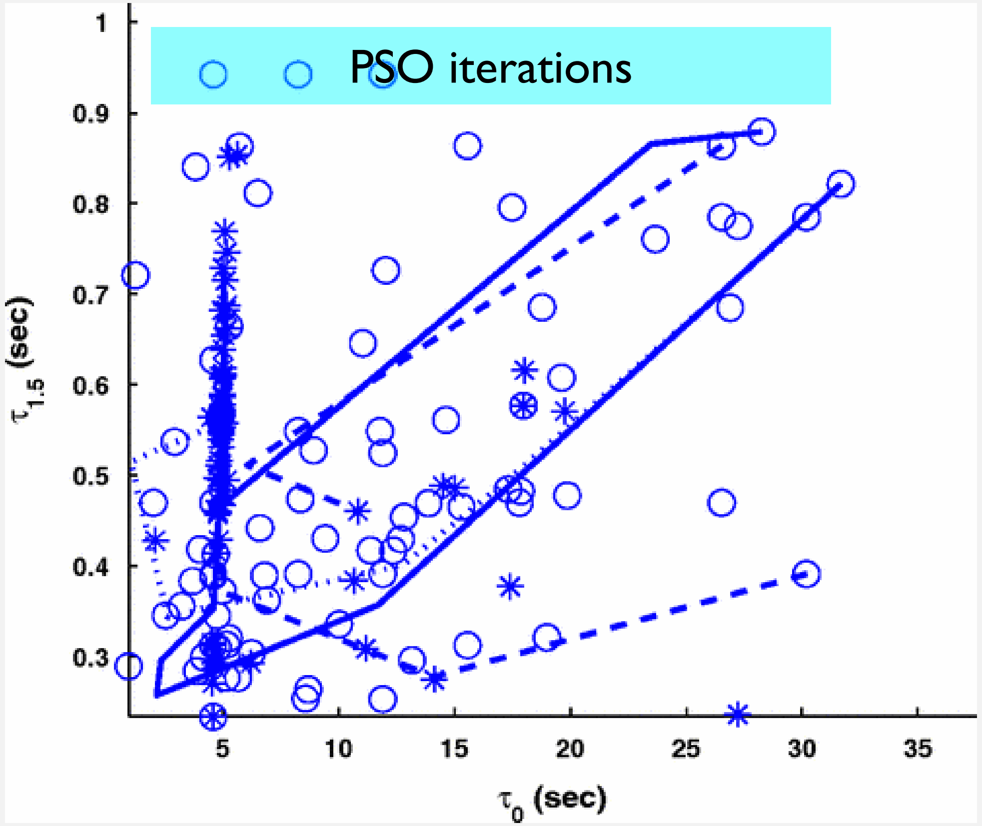


Figure from Wang, Mohanty, Physical Review D (2010)



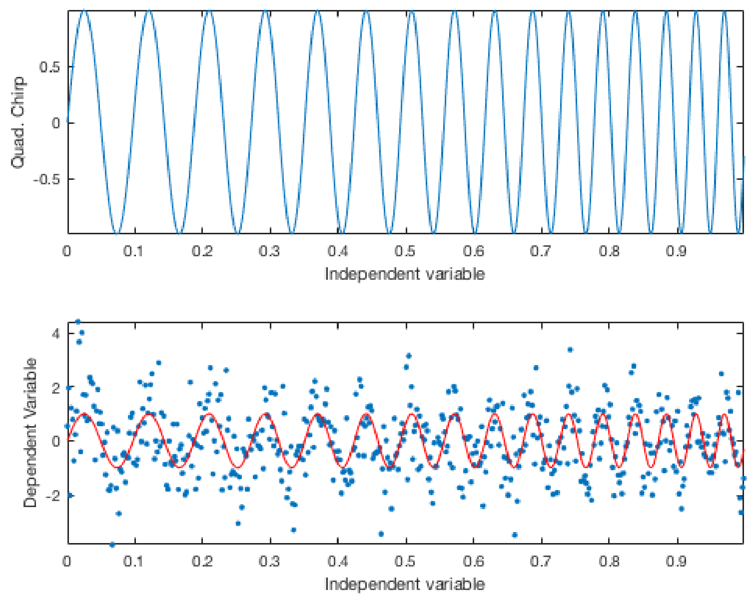


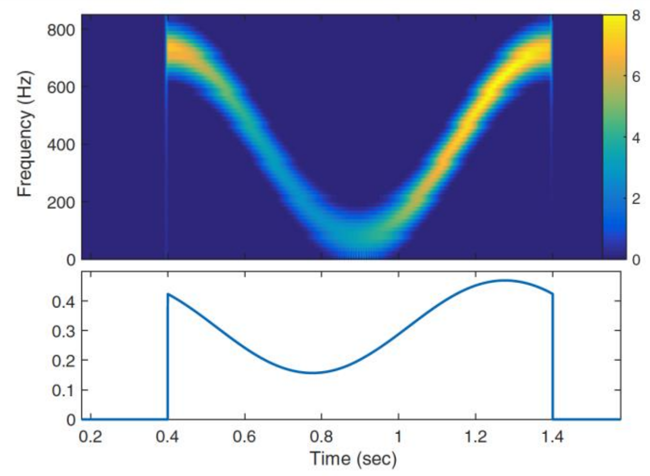
PSO-BASED BINARY INSPIRAL SEARCH

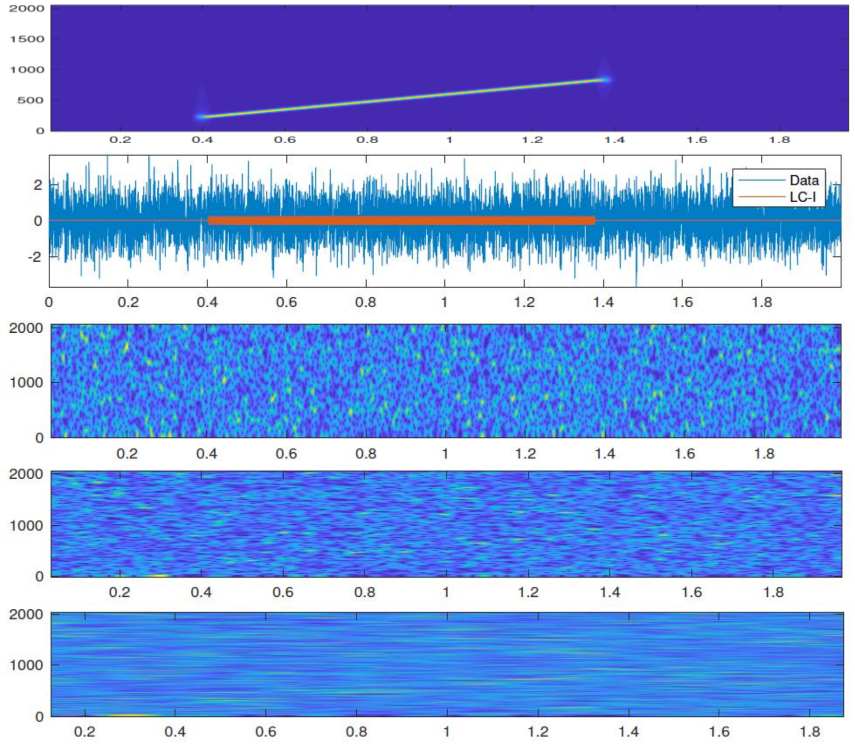
- First use in GW data analysis:
  - Wang, Mohanty, Physical Review D, 2010
- PSO: factor of ≈ 10 fewer evaluations
  - Weerathunga, Mohanty, 2017
- On the threshold of a real-time optimal search:
  - Normandin, Mohanty, Weerathunga, 2018
  - Srivastava, Nayak, Bose, 2018

SEARCH FOR UNMODELED CHIRPS

- Chirp signal: $f(x) = a(x)\sin(\Phi(x))$,
  - Where the instantaneous frequency, $\frac{d\Phi}{dx}$, changes adiabatically on timescales of the instantaneous period
  - Example: Quadratic chirp
- Unmodeled chirp signal: $a(x)$ and $\Phi(x)$ have unknown functional forms





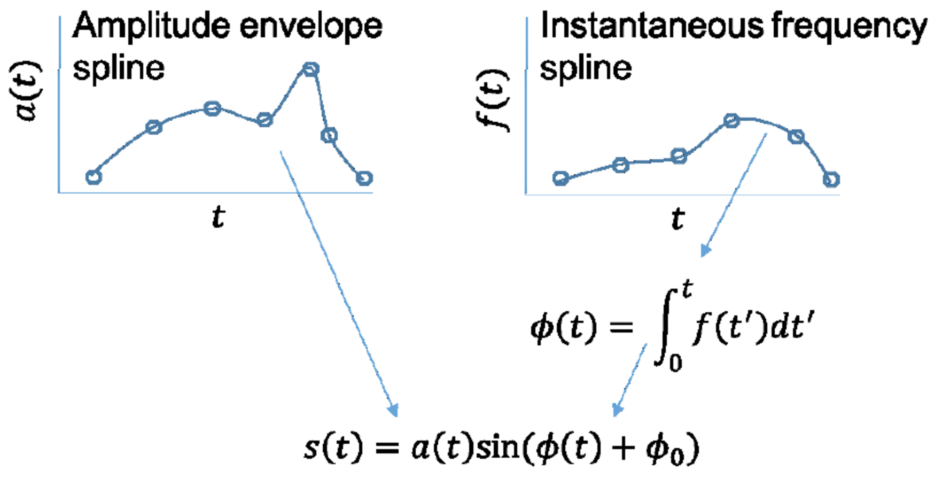


TIME-FREQUENCY ANALYSIS

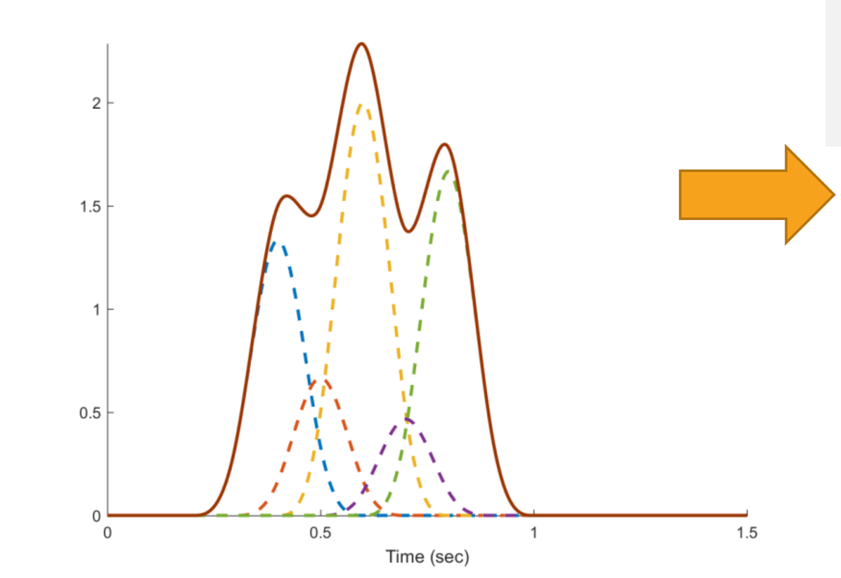
- Working definition: A chirp appears as a track in the Time-Frequency (TF) plane
- At signals strengths expected for GW signals, noise can completely mask chirp signals in a time-frequency transform
- Current searches for unmodeled GW signals are all based on some variation of time-frequency analysis

SEARCH FOR UNMODELED CHIRPS

- New approach: model the unknown functions with splines and optimize over their breakpoints
  - Soumya D. Mohanty, Physical Review D (2017).
- SEECR: Spline-Enabled Effective-Chirp Regression

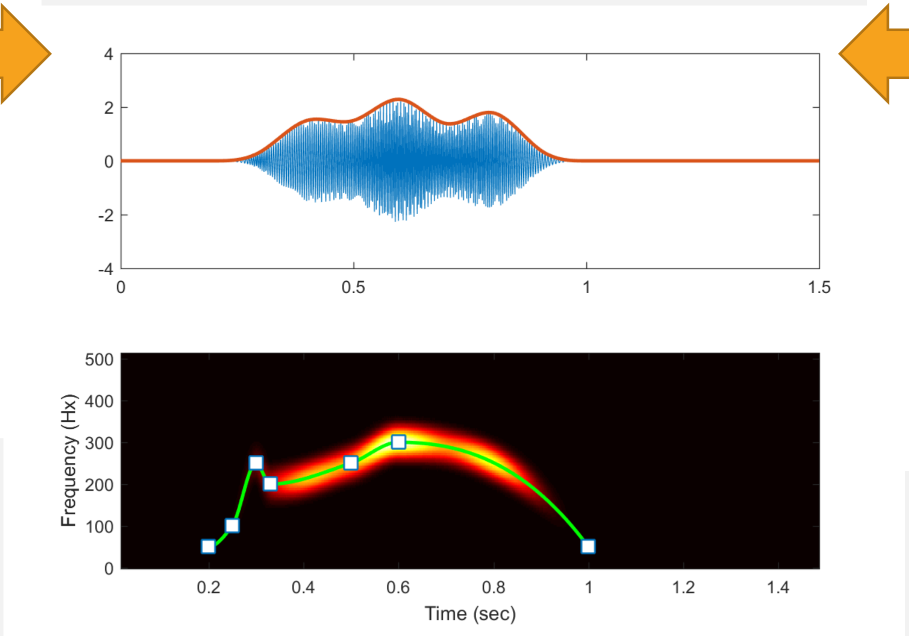


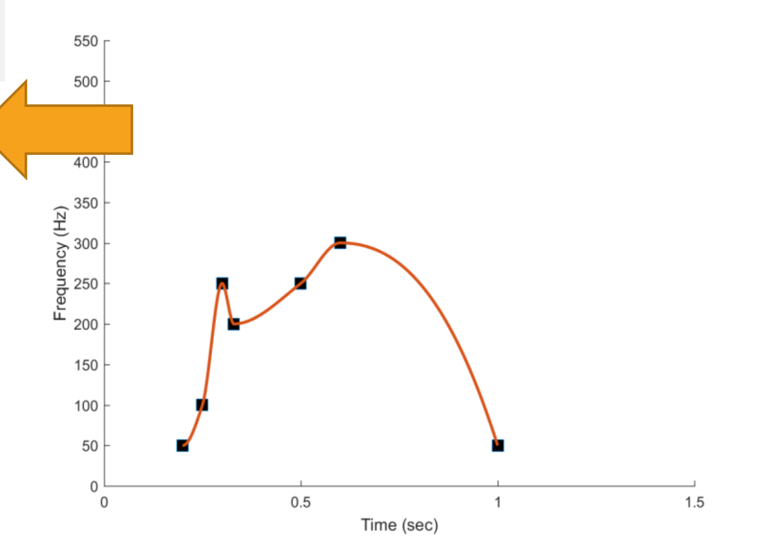
SEECR: SIGNAL MODEL



$$a(t) = \sum_{i=0}^{M-1} \alpha_i B_{i,k}(t)$$

$$s(t) = a(t)\sin(\phi(t) + \phi_0)$$



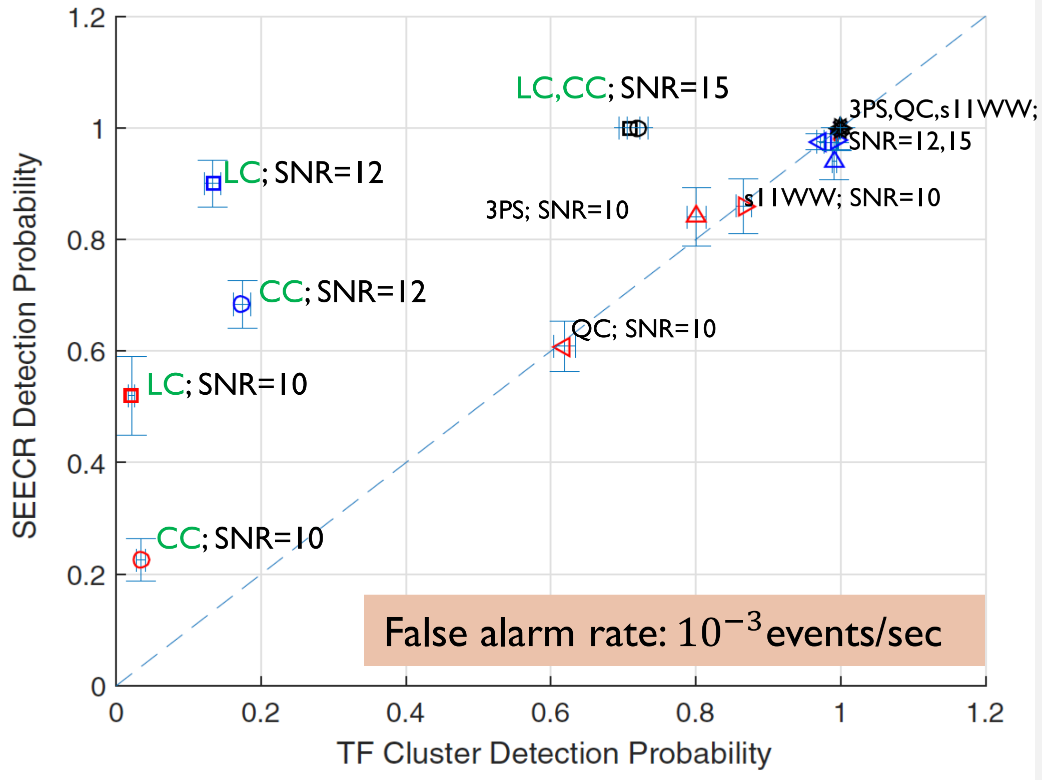


$$\phi(t) = 2\pi \int_0^t f(t')dt'$$

SEARCH FOR UNMODELED CHIRPS

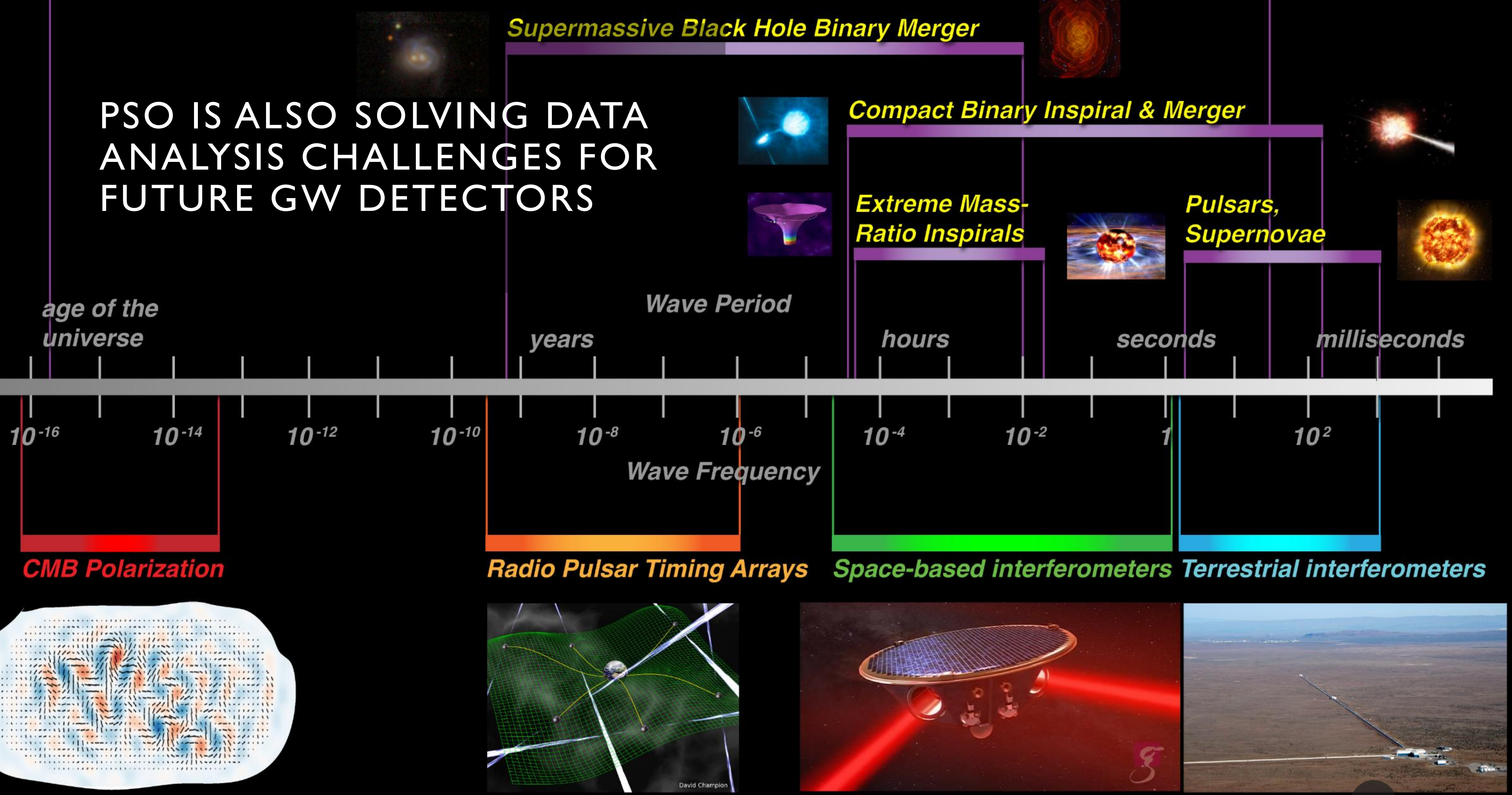
- The new approach is inconceivable without successfully solving the optimization task
  - Non-linear, high-dimensional model: up to 20 parameters used
- Reward: Significantly better performance for chirps than current approaches based on time-frequency clustering

PHYSICAL REVIEW D **96,** 102008 (2017)



LC: Linear chirp (increasing frequency with time); 1sec long CC: Cosine chirp; 1 sec QC: Quadratic chirp; 1 sec	3PS: Strongly amplitude modulated sinusoid; 1.5 sec s11WW: Acoustic supernova; 0.7 sec

PSO IS ALSO SOLVING DATA ANALYSIS CHALLENGES FOR FUTURE GW DETECTORS



SUMMARY

OPTIMIZATION IN STATISTICAL REGRESSION

- Non-linear regression models can be advantageous over linear model but involve a difficult optimization task
- Solving the optimization problem allows us to explore more flexible (and better) models
  - This may improve the predictive power of a regression model ⇒ better inferences from data
- Swarm intelligence methods can be useful tools for such optimization tasks