

---

**Oleksandr Romanko, Ph.D.**

Senior Research Analyst, Risk Analytics, Watson Financial Services, IBM Canada  
Adjunct Professor, University of Toronto

# **MIE1624H – Introduction to Data Science and Analytics Lecture 5 – Modeling and Regressions**



# Modeling

---

## Models



essentially,  
all models are wrong,  
but some are useful

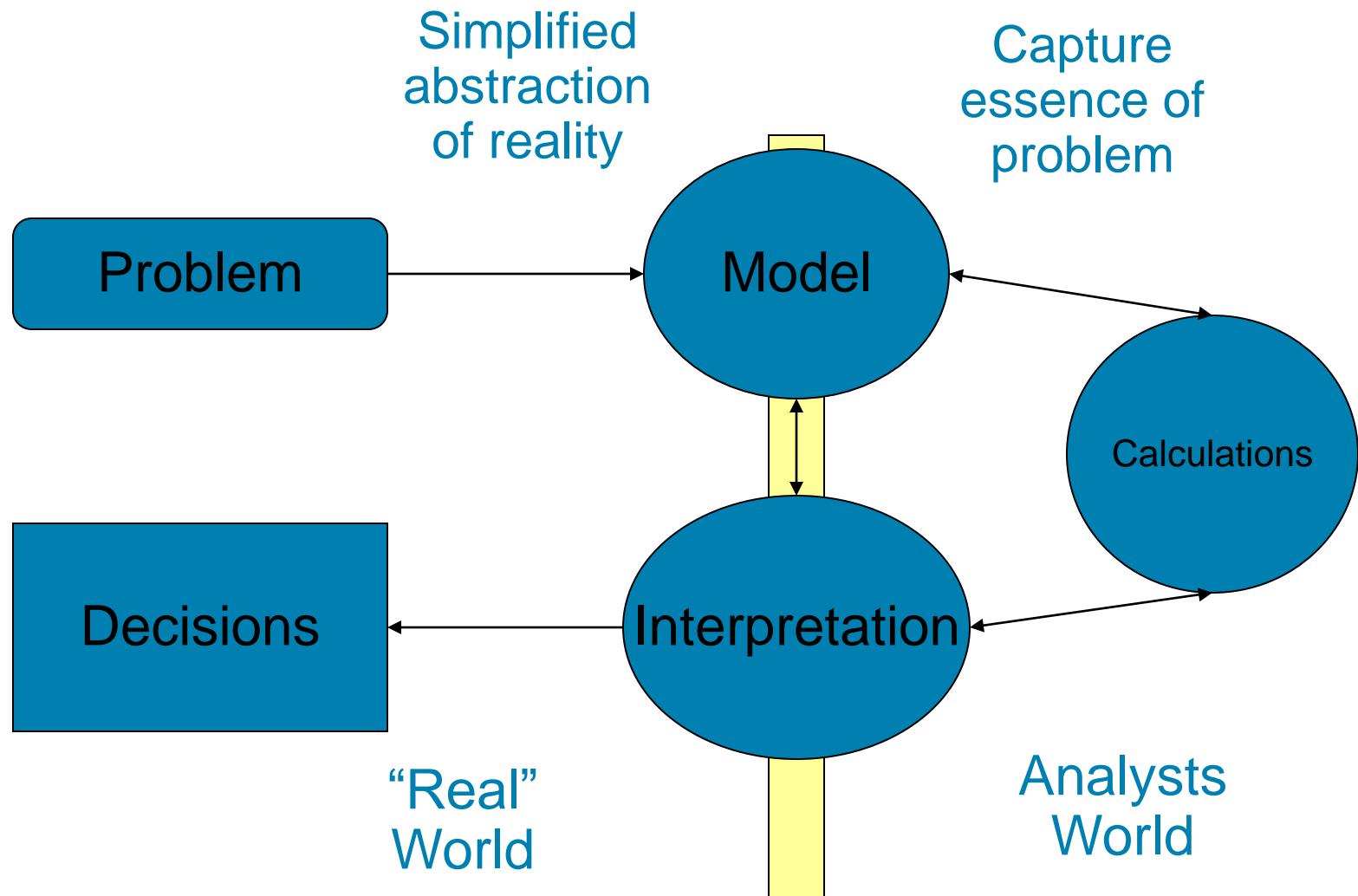
George E. P. Box

---

# Models

- Simplified representation or abstraction of reality
- Capture essence of system without unnecessary details
- Models tailored for specific types of problems
- Models help us understand the world
  - Prediction (What if?)
  - Optimization (What's best?)
- Often models much easier, faster, and cheaper to experiment with than the real system

# Models and reality





# Predictive Maintenance

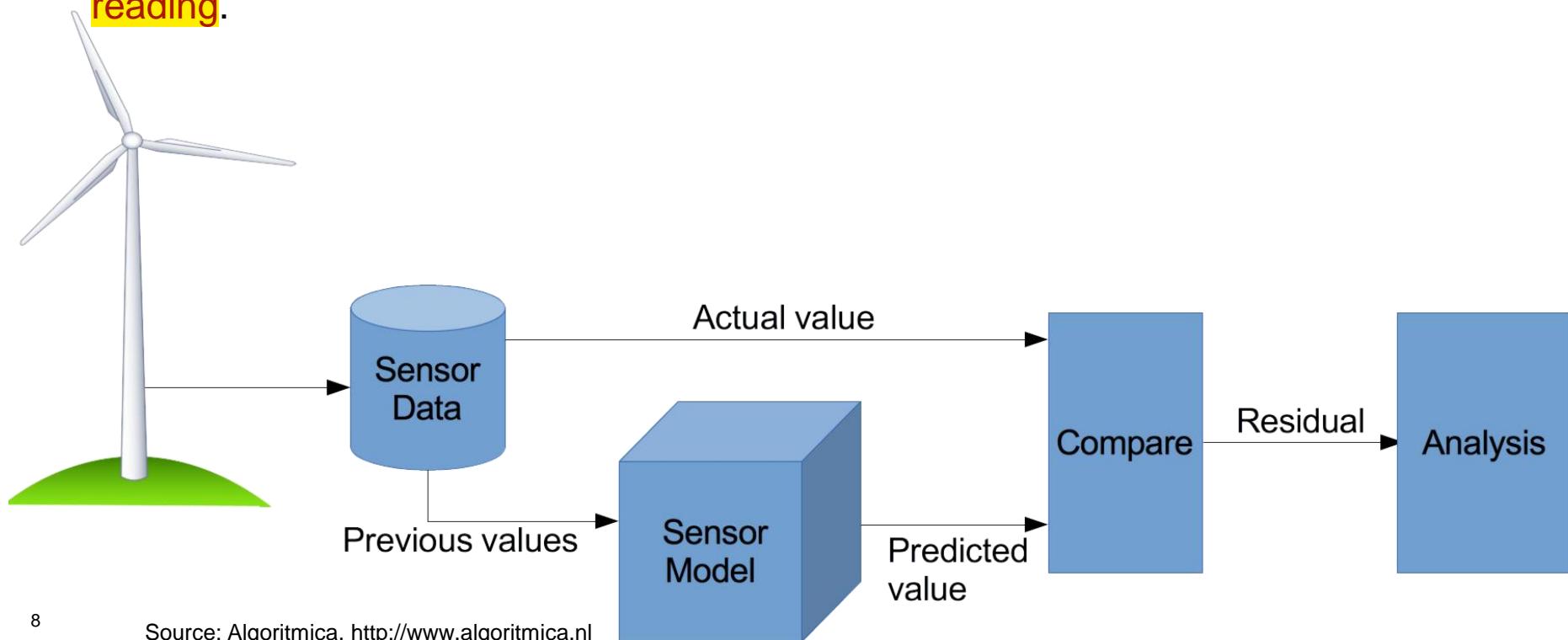
# Predictive maintenance – before big data

- **Wind turbines** are big and expensive machines, so keeping them running smoothly helps keeping their operational cost down. To do that, you need to be able to anticipate failures in heavy and expensive parts like the gearbox, generator and main shaft.
- **Physical models** explicitly describes the turbine design using detailed knowledge of its physical characteristics. A **physical model** has to be calibrated by an **expert**.
- **Preventive maintenance** saves money:
  - Shorter downtime and less lost production
  - Better planning of people and materials
  - Cheaper repairs



# Predictive maintenance – big data era

- In the big data era wind turbines have an array of sensors that measure temperatures, pressures, voltages, currents, and blade angles.
- **Data-driven approach:** a model learns the relationship between the various sensor readings based on the training data. To create such a sensor model we apply machine learning, i.e. one or more algorithms that use a set of training examples to learn a predictive model. The model can be trained by a non-turbine expert. The model then calculates its predicted value and compares it with the actual sensor reading.



# Environmental risk management



I·R·O·N·Y  
POLLUTION PRODUCED BY CLEAN ENERGY.



# Modeling Process

---

## Why do we model for decision making?

- **Building model forces detailed examination and thought about a problem**
  - structures our thinking
  - must articulate our assumptions, preconceived notions
  - model building may illuminate solution without actually using the model
- **Searching for general insights**
  - form of relationship between key variables involved in decision
  - importance of various parameters on decisions
- **Looking for specific numeric answers to a decision making problem**
  - If we add 1 lab tech between 7a-3p, how much reduction can we expect in test turnaround time?
- **Find the best way to do something**
  - Which routing schedule minimizes our delivery costs?

---

## Types of models

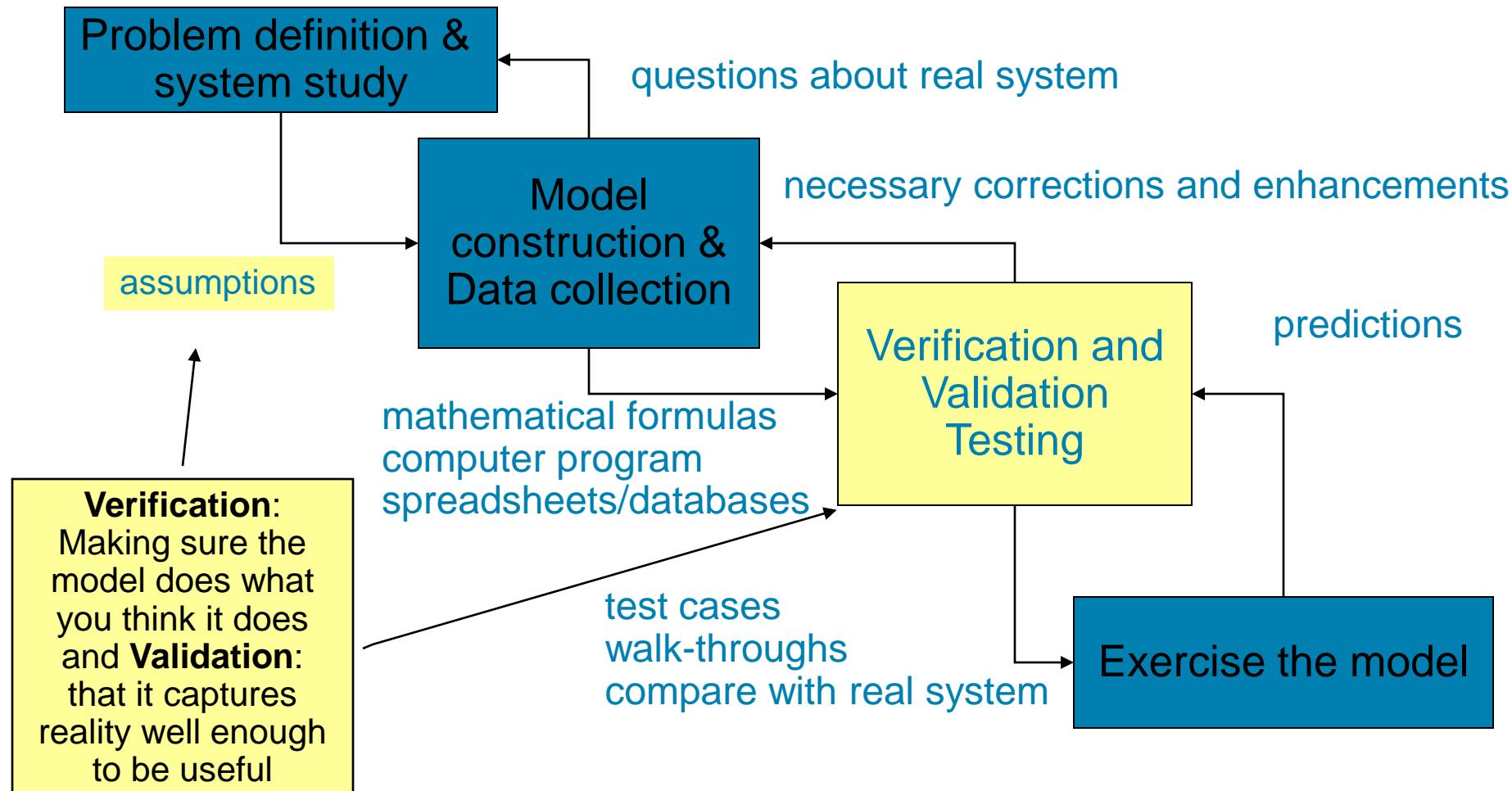
- **Physical** – cars, buildings
- **Diagrams** – flow chart, decision trees, influence diagrams
- **Statistical** – regression equation, probability distribution
- **Mathematical** – queuing model, scheduling model
- **Computer simulation**
- **Computational** – neural networks, genetic algorithms

---

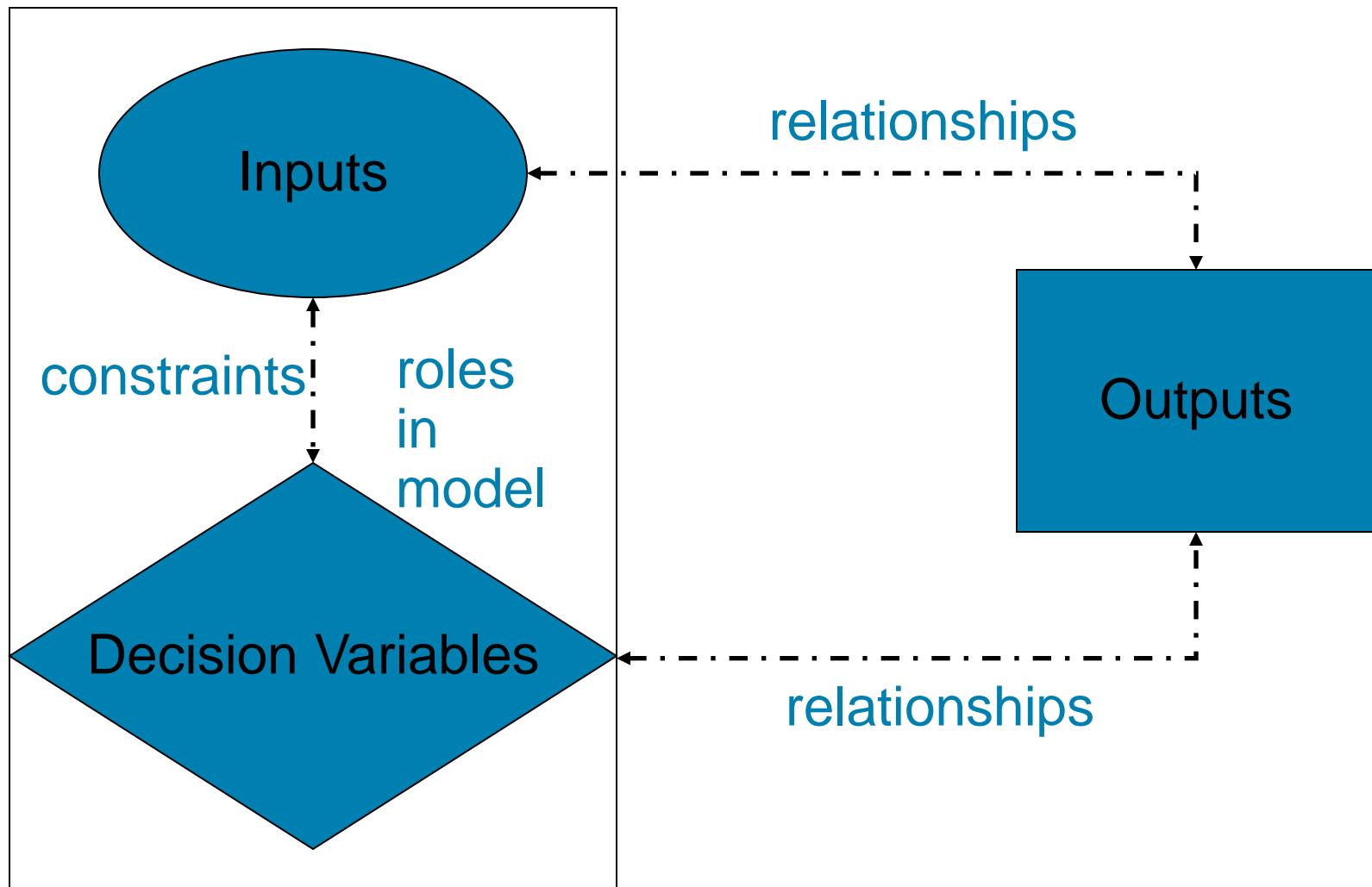
## A 7 step (idealized) modeling process

- **Define the problem**
  - “exploring the mess”
- **Observe system / collect data**
- **Formulate model(s)**
  - much “art and craft”
- **Verify/validate model and use for prediction and exploration of system being modeled**
- **Use model to help select among alternatives**
- **Present results to decision makers**
- **Implement solution and evaluate outcomes**

# A simple modeling process



## Basic modeling concepts





# Model Examples

# Pit stop analytics

F1 analytics based on past experience

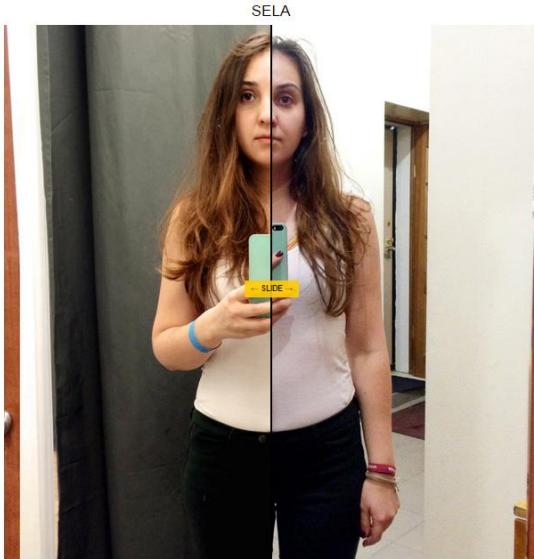


Calculations showed that time spent changing tires and refilling the tank was more than offset by the improved performance of the car on the track.

1. Softer tires stuck to the track better during turns than their harder cousins, though they wore out more quickly.
2. Less gas in the tank translated into a lighter, and therefore faster, car.

# Fitting room analytics

Good



Bad



Source: Adme

---

## Shortest path or most beautiful path?

# Forbes

---



Amit Chowdhry Contributor

*I cover noteworthy technology, startups, and gadgets*

Opinions expressed by Forbes Contributors are their own.

---

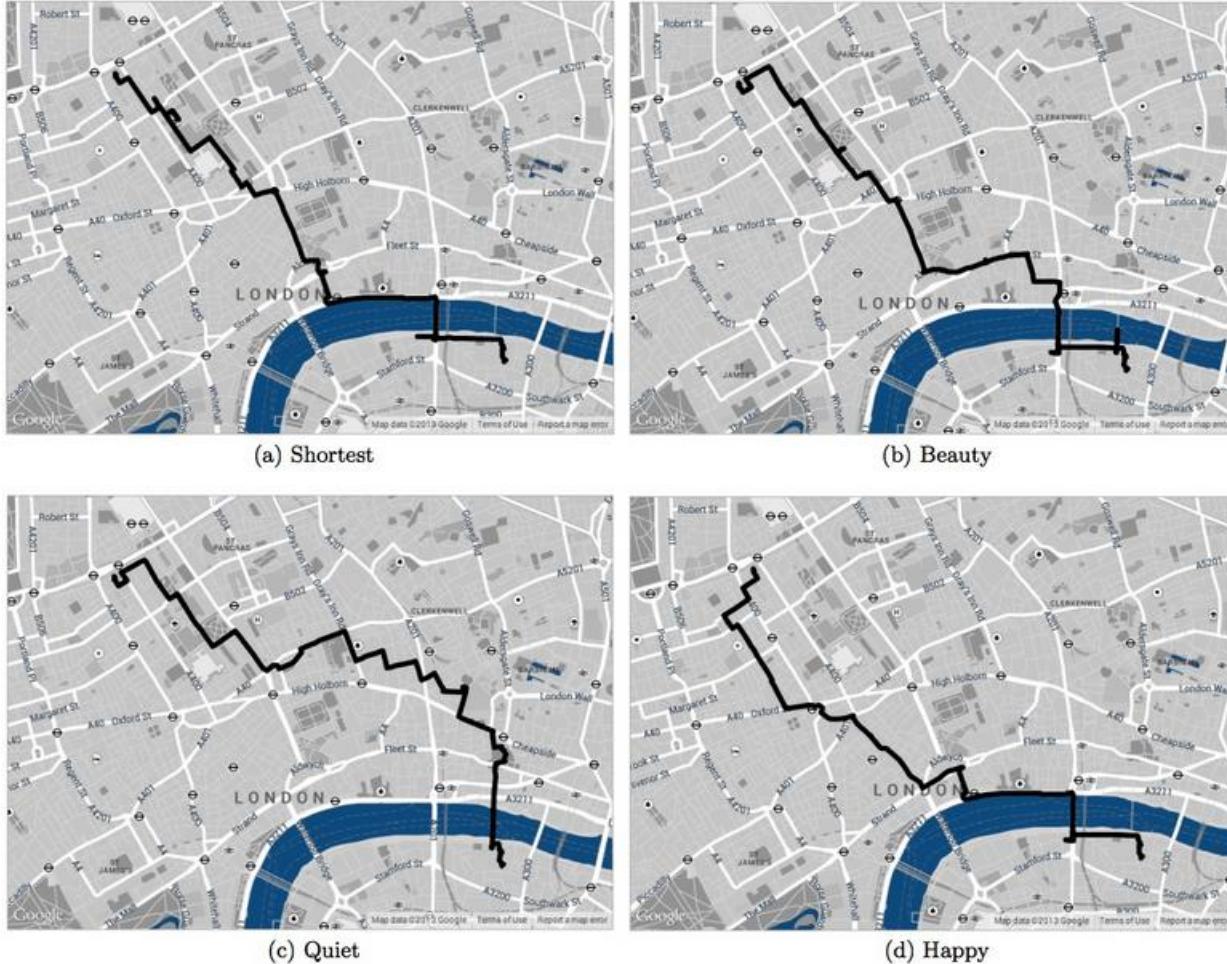
TECH 7/09/2014 @ 6:42PM | 1,273 views

# Yahoo! Researchers Have Developed A GPS Algorithm To Find 'Emotionally Pleasant' Routes

[+ Comment Now](#)

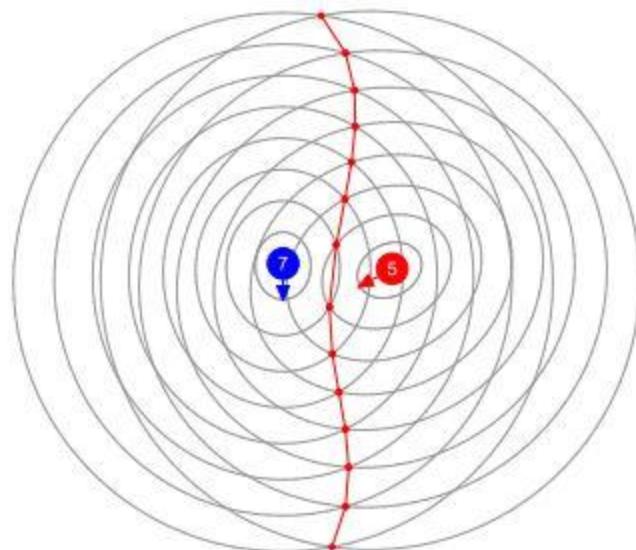
Mapping services and GPS devices are generally used for finding the shortest routes between two points while walking around or driving a car. Over the last few years, GPS technology has improved with features like the ability avoid tolls and construction. Now a team of researchers at Yahoo Labs in Barcelona, Spain want mapping services to offer pedestrians with a way to find the most scenic routes while heading towards their destination.

# Shortest path or most beautiful path?

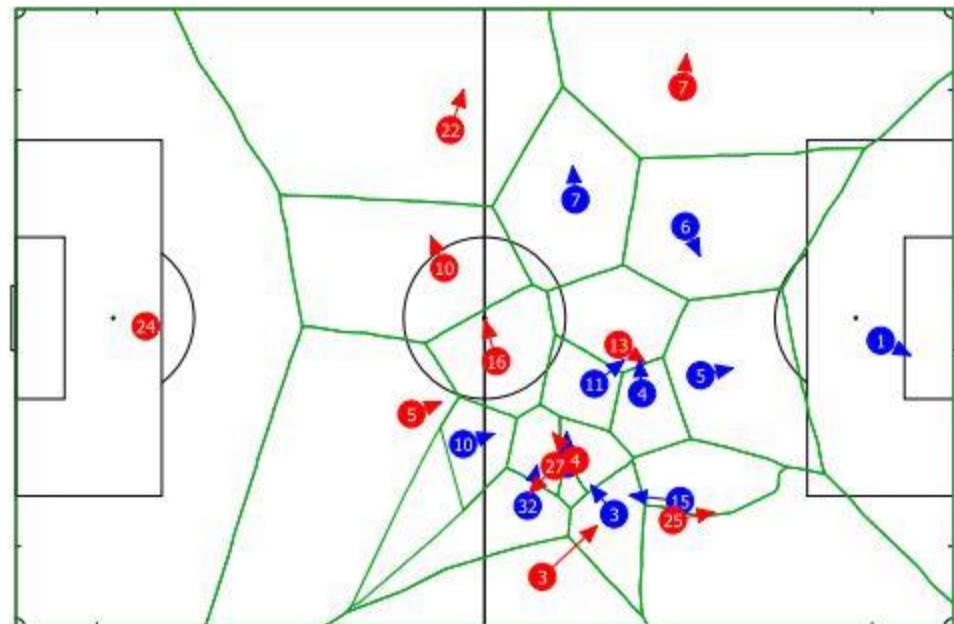


Routes plotted in London based on different algorithms. Credit: Yahoo!

# Sports analytics



(a)



(b)

# Personalized health analytics – 23andMe

The screenshot shows the 23andMe website homepage. At the top, there's a navigation bar with links for 'welcome' (which is highlighted in green), 'ancestry', 'how it works', 'buy', 'search', and 'help'. A blue banner below the navigation bar contains a warning message: '23andMe provides ancestry-related genetic reports and uninterpreted raw genetic data. We no longer offer our health-related genetic reports. If you are a current customer please go to the [health page](#) for more information. [Close alert.](#)' To the left, there's an image of a 23andMe DNA test kit box with a colorful geometric design and the text 'welcome to you®'. To the right, there's a large heading 'Find out what your DNA says about you and your family.' followed by a bulleted list of benefits: '• Learn what percent of your DNA is from populations around the world', '• Contact your DNA relatives across continents or across the street', and '• Build your family tree and enhance your experience with relatives'. Below this is a pink button labeled 'order now' and a price of '\$99'. At the bottom right, there's a circular pie chart showing estimated ancestry composition: 38.6% Sub-Saharan African (pink), 24.7% European (blue), and 20.5% East Asian (orange). A text overlay next to the chart says 'What will your Ancestry Composition look like?'. On the left side of the main content area, there's a section with the text '23 pairs of chromosomes. One unique you.' and a detailed description of the service's capabilities.

welcome ancestry how it works buy search help

!

23andMe provides ancestry-related genetic reports and uninterpreted raw genetic data. We no longer offer our health-related genetic reports. If you are a current customer please go to the [health page](#) for more information. [Close alert.](#)

welcome to you®

Find out what your DNA says about you and your family.

- Learn what percent of your DNA is from populations around the world
- Contact your DNA relatives across continents or across the street
- Build your family tree and enhance your experience with relatives

order now \$99

23 pairs of chromosomes. One unique you.

Find out what percent of your DNA comes from populations around the world, ranging from East Asia, Sub-Saharan Africa, Europe, and more. Break European ancestry down into distinct regions such as the British Isles, Scandinavia, Italy and Ashkenazi Jewish. People with mixed ancestry, African Americans, Latinos, and Native Americans will also get a detailed breakdown.

38.6% Sub-Saharan African

20.5% East Asian

24.7% European

What will your Ancestry Composition look like?

# Personalized health analytics – 23andMe

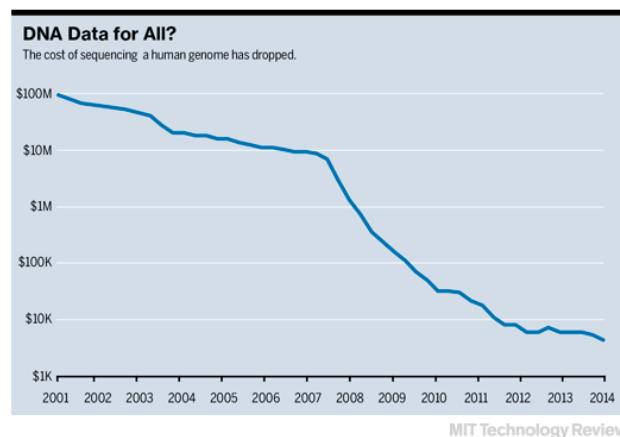
BUSINESS REPORT

Data-Driven Health Care

## 23andMe Tries to Woo the FDA

The DNA testing firm hopes a more coöperative approach with regulators will get its business back on track.

By Robert D. Hof on July 21, 2014



Anne Wojcicki bounds into a conference room in Mountain View, California, straight from a five-mile ride from home on an elliptical bike. The 40-year-old cofounder and CEO of the consumer genetic testing firm 23andMe is breathless, and not just because of the workout. On this warm day in mid-June, Wojcicki is "super-excited" about an announcement scheduled for two days hence: the Food and Drug Administration has agreed to review a health-related genetic report the company wants to make available to customers.

It's the first step out of the FDA's doghouse for 23andMe. For \$99, the company analyzes key components of a person's DNA from a vial of saliva, but last November the federal agency issued a testy [warning letter](#) barring it from marketing its service. The FDA said that by selling consumers a test and health reports that outlined their chances of getting dozens of diseases, plus their likely response to various drugs, 23andMe was effectively selling a medical device. That requires explicit approval—and the FDA said 23andMe hadn't come close to providing enough evidence that its test provides accurate, reliable health assessments.

**MIT Technology Review**  
BUSINESS REPORT

## Data-Driven Health Care

New technologies promise a flood of molecular, environmental, and behavioral patient information. Will all that data make medicine better?

**Can Technology Fix Medicine?**

Medical data is a hot spot for venture investing and product innovation. The goal: better care.

After decades as a technological laggard, medicine has entered its data age. Mobile technologies, sensors, genome sequencing, and advanced statistical software now make it possible to capture vast amounts of information about our individual makeup and our environment. Some say that much of this information could transform medicine, turning a field aimed at treating the average patient into one that tailors treatments to individuals and shifting more control and responsibility from doctors to patients.

The question is: can big data make health care better?

"There is a lot of data being gathered. That's great," says Michael J. Daniel, interim director of the Information Services Unit at the University of California, San Francisco, School of Medicine. "It's really about putting up applications that make data sensible."

### CONTENTS

[Can Technology Fix Medicine?](#)

[More Phones, Fewer Doctors](#)

[IBM Aims to Make Medical Expertise a Commodity](#)

[23andMe Tries to Woo the FDA](#)

[Mobile Health's Growing Pains](#)

# Personalized health analytics – 23andMe



23andMe genetics just got personal.

Search 23andMe

Go

[Log in](#) | [Register Your Kit](#) | [Blog](#) | [Help](#) ▾

welcome

ancestry

health

how it works

store

## Start filling in the gaps with your DNA



"Because I had given my doctor information from 23andme, he got to a diagnosis much faster. 23andme saved my life." Kirk C.

\$99\*

Our new low price for all!  
Was \$199

[Order Now »](#)

\*Requires a 1-year commitment to the [Personal Genome Service®](#) at \$9/mo. [Order for \\$399](#) without commitment.

① Get Your Kit



② Provide Saliva



③ Learn About Yourself



④ Get Monthly DNA Discoveries



Gain insight into your traits, from baldness to muscle performance. Discover risk factors for 97 diseases. Know your predicted response to drugs, from blood thinners to coffee. And uncover your ancestral origins. [start tour »](#)

Overview

Discover Health & Ancestry

Keep Your Doctor Informed

Participate In Research

# Personalized health analytics – 23andMe



# Personalized health analytics – 23andMe

 Search

Douglas Brutlag | Account ▾ | Help ▾ | Blog | Log out | Cart

My Home

Inbox (6)

My Health

Disease Risk

Carrier Status

Drug Response

Traits

Health Labs

My Ancestry

Maternal Line

Paternal Line

Relative Finder

Ancestry Painting

Global Similarity

Ancestry Labs

Sharing & Community

Compare Genes

Family Inheritance

23andMe Community

Genome Sharing

23andWe

Research Surveys (24)

Research Snippets

Research Initiatives

Research Discoveries

## disease risk

Share my health results with family and friends

Show results for

Douglas Brutlag

See new and recently updated reports »

23andWe Discoveries were made possible by 23andMe members who took surveys.

Locked Reports

Name

Confidence

Your Risk

Avg. Risk

Compared to Average

Alzheimer's Disease

★★★★



Elevated Risk

Name

Confidence

Your Risk

Avg. Risk

Compared to Average

Prostate Cancer

★★★★

22.4%

17.8%

1.26x



Colorectal Cancer

★★★★

7.1%

5.6%

1.27x



Melanoma

★★★★

6.0%

2.9%

2.10x



Restless Legs Syndrome

★★★★

2.5%

2.0%

1.25x



Exfoliation Glaucoma

★★★★

2.2%

0.7%

2.90x



Abdominal Aortic Aneurysm

★★★



Ankylosing Spondylitis

★★★



Asthma

★★★



Atopic Dermatitis

★★★



Bipolar Disorder: Preliminary Research

★★★



# Personalized health analytics – 23andMe

 Search

Douglas Brutlag | Account ▾ | Help ▾ | Blog | Log out | Cart

## drug response

Share my health results with family and friends

[My Home](#)  
Inbox (6)

[My Health](#)  
Disease Risk  
Carrier Status  
▶ Drug Response

Traits  
Health Labs

[My Ancestry](#)  
Maternal Line  
Paternal Line  
Relative Finder  
Ancestry Painting  
Global Similarity  
Ancestry Labs

[Sharing & Community](#)  
Compare Genes  
Family Inheritance  
23andMe Community  
Genome Sharing

Show results for

Douglas Brutlag

[See new and recently updated reports »](#)

23andWe Discoveries were made possible by 23andMe members who took [surveys](#).

Name	Confidence ▾	Status
Clopidogrel (Plavix®) Efficacy	★★★★	Greatly Reduced
Abacavir Hypersensitivity	★★★★	Typical
Alcohol Consumption, Smoking and Risk of Esophageal Cancer	★★★★	Typical
Fluorouracil Toxicity	★★★★	Typical
Response to Hepatitis C Treatment	★★★★	Typical
Pseudocholinesterase Deficiency	★★★★	Typical
Warfarin (Coumadin®) Sensitivity	★★★★	Typical
Oral Contraceptives, Hormone Replacement Therapy and Risk of Venous Thromboembolism	★★★★	Not Applicable
Caffeine Metabolism	★★★	Fast Metabolizer
Hepatitis C Treatment Side Effects	★★★	See Report

# Personalized health analytics – 23andMe

 Search

Douglas Brutlag | Account ▾ | Help ▾ | Blog | Log out | Cart

My Home

Inbox (6)

**My Health**

Disease Risk

Carrier Status

Drug Response

► Traits

Health Labs

**My Ancestry**

Maternal Line

Paternal Line

Relative Finder

Ancestry Painting

Global Similarity

Ancestry Labs

**Sharing & Community**

Compare Genes

Family Inheritance

23andMe Community

Genome Sharing

**23andWe**

Research Surveys (24)

Research Snippets

Research Initiatives

Research Discoveries

## traits

Share my health results with family and friends

Show results for

Douglas Brutlag

[See new and recently updated reports »](#)

23andWe Discoveries were made possible by 23andMe members who took surveys.

Name	Confidence ▲	Outcome
Alcohol Flush Reaction	★★★★	Does Not Flush
Bitter Taste Perception	★★★★	Can Taste
Earwax Type	★★★★	Wet
Eye Color	★★★★	Likely Brown
Hair Curl	★★★★	Straighter Hair on Average
Lactose Intolerance	★★★★	Likely Tolerant
Malaria Resistance (Duffy Antigen)	★★★★	Not Resistant
Male Pattern Baldness ♂	★★★★	Decreased Odds
Muscle Performance	★★★★	Likely Sprinter
Non-ABO Blood Groups	★★★★	<a href="#">See Report</a>
Norovirus Resistance	★★★★	Not Resistant
Resistance to HIV/AIDS	★★★★	Not Resistant
Smoking Behavior	★★★★	Typical
Adiponectin Levels	★★★	<a href="#">See Report</a>
Asparagus Metabolite Detection	★★★	Typical Odds of Detecting

# Personalized health analytics – 23andMe



23andMe genetics just got personal.



brutlag

genetics 101

blog

help

sign out



My Gene Journal (60)

Browse Raw Data

## family & friends

Compare Genes

Family Inheritance

## my ancestors

Maternal Line

Paternal Line

Ancestry Painting

Global Similarity

## account

My Profiles

Genome Sharing

Settings

Help/Contact Us

## maternal line

Your mitochondrial DNA determines your maternal haplogroup. [What is a haplogroup?](#)

Map

History

Haplogroup Tree

### Maternal Haplogroup: U5

Locations of haplogroup U5 circa 500 years ago, before the era of intercontinental travel.



Haplogroup U5 arose among early colonizers of Europe around 40,000 years ago; maternal descendants of those early colonizers persist in the region to this day. After the last Ice Age two subgroups of U5 expanded across Europe and into northern Africa and the Near East. Today, one subgroup, U5b1b, is shared by groups as diverse as the northern African desert-dwelling Berbers and the Scandinavian Arctic-dwelling Saami, also known as the Lapps.

Haplogroup: U5, a subgroup of [U](#)

Age: 40,000 years

Region: Europe, Near East, North Africa

Populations: Basques, Saami (Lapps) of northern Scandinavia

Highlight: Though primarily a European haplogroup, U5 was recently found in mitochondrial DNA extracted from the remains of a 6th-century AD Chinese chieftain.

### Your Family and Friends

[K1a1b1a](#) Simone Brutlag

[U5b2](#) Douglas Brutlag

[L3e](#) Nigerian Man

[D5a2](#) Chinese Man

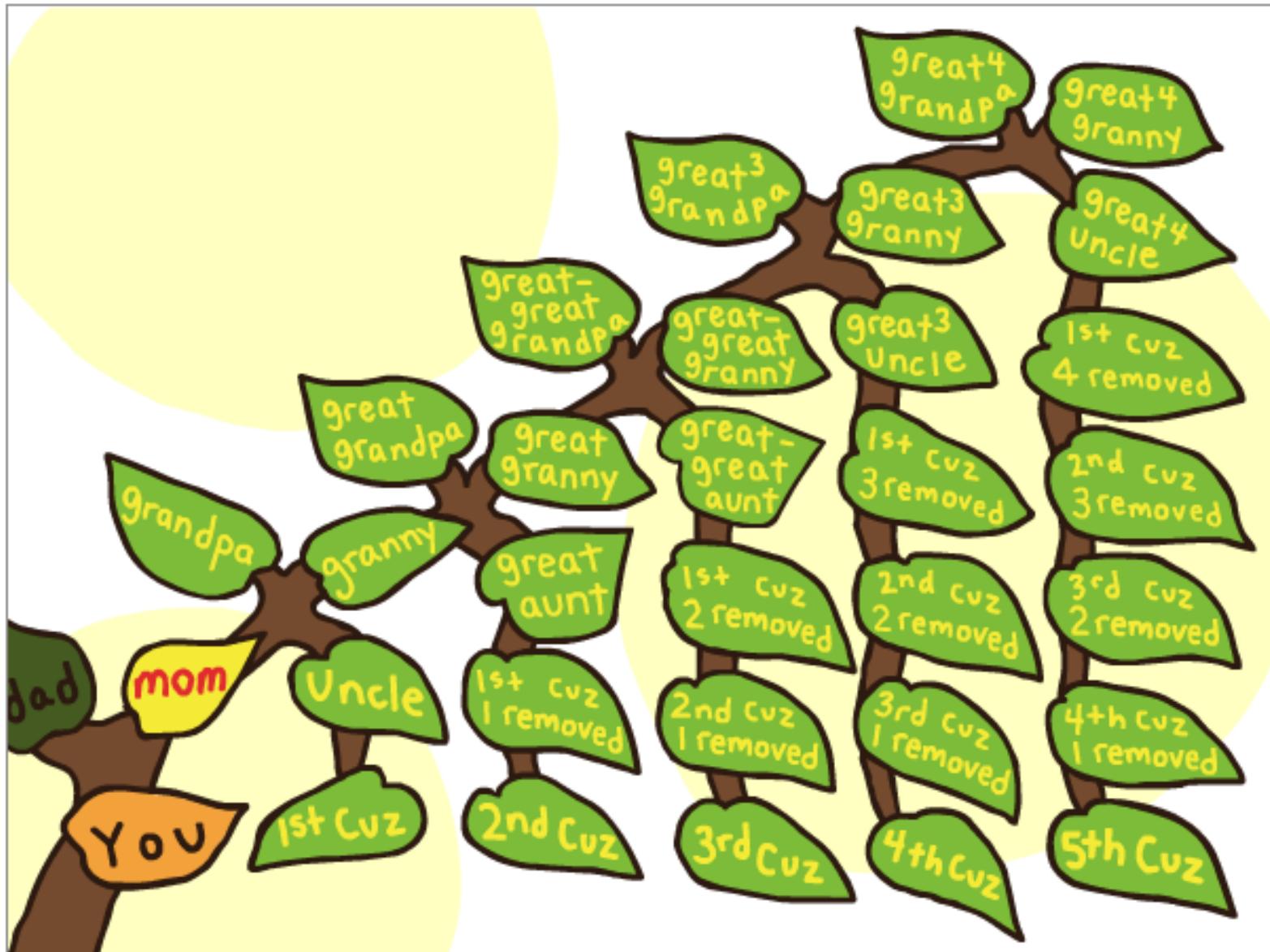
[D4e2](#) Japanese Man

# Personalized health analytics – 23andMe relative finder



		search matches	sort by relationship	25 per page	« ← 1 – 25 of 309 matches → »
	Male	You	U5b2a E1b1b1a2*		<a href="#">Update Your Profile</a>
	Benjamin Brutlag Male, b. 1980	Son 47.7% shared, 22 segments	United States Southern Europe K1a1b1a E1b1b1a2*		<a href="#">Sharing Genomes</a> <a href="#">Send a Message</a>
	Pauline Brutlag Female	Daughter 53.1% shared, 25 segments	United States Northern Europe K1a1b1a		<a href="#">Sharing Genomes</a> <a href="#">Send a Message</a>
	Male	3rd to 5th Cousin 0.47% shared, 3 segments	H3 R1a1a*		<a href="#">Send an Introduction</a>
	Larry Vongroven Male	3rd to 5th Cousin 0.54% shared, 2 segments	United States Alen, Norway Haltalen, Norway Voss, Norway 8 more Northern Europe Vongroven (Vongraven) Bakken Goodno 10 more U4b1a2 R1a1a		<a href="#">Introduction Received</a> <a href="#">Respond</a>
	Female	3rd to 5th Cousin 0.47% shared, 2 segments	K1a10		<a href="#">Send an Introduction</a>
	Male	3rd to 5th Cousin 0.34% shared, 2 segments	I2a R1a1a*		<a href="#">Send an Introduction</a>
	Male	3rd to 6th Cousin 0.39% shared, 1 segment	U5b1b1a R1b1b2a1a1		<a href="#">Send an Introduction</a>
	Male	3rd to 6th Cousin 0.36% shared, 1 segment	H2a5 R1b1b2a1a1*		<a href="#">Send an Introduction</a>
	Male	3rd to 6th Cousin 0.31% shared, 2 segments	H7a D1*		<a href="#">Send an Introduction</a>
	ivan otterness Male, b. 1938	3rd to 6th Cousin 0.30% shared, 2 segments	United States Minnesota, Wisconsin, Norway Northern Europe Otterness Brandsness Flekke 4 more K1a10 I2b1		<a href="#">Send a Message</a>
	Female	3rd to 6th Cousin 0.29% shared, 2 segments	U4b1a1		<a href="#">Introduction Received</a> <a href="#">Respond</a>

# Personalized health analytics – 23andMe



# Personalized health analytics – 23andMe



23andMe provides ancestry-related genetic reports and raw genetic data. We no longer offer health-related genetic reports at this time. If you are a current customer please go to the [health page](#) for more information. [Close alert](#).

[Developer login & signup](#) | [Docs](#) | [Help](#)

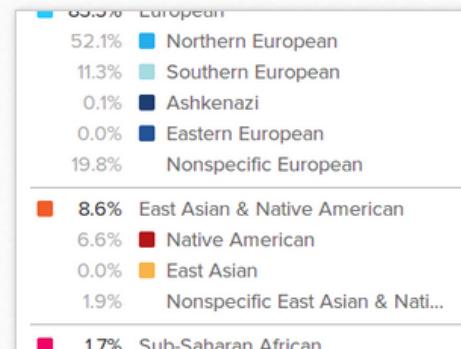
## Genetics For Your App

[Develop for free](#)

```
1  [
2    {
3      "rs2707466": "CC",
4      "rs7776725": "TT",
5      "rs9525638": "CT",
6      "id": "SP1_MOTHER_V3",
7      "rs2908004": "__"
8    },
9    {
10      "rs2707466": "CC",
11      "rs7776725": "TT",
12      "rs9525638": "CT",
13      "id": "SP1_FATHER_V3",
14      "rs2908004": "__"
15    }
16  ]
```

### RESTful genes

Our customers are genotyped for hundreds of thousands of SNPs, conveniently accessible through our free REST API. Not genotyped? We have demo endpoints.



85.5% European

Ancestry Group	Percentage
Northern European	52.1%
Southern European	11.3%
Ashkenazi	0.1%
Eastern European	0.0%
Nonspecific European	19.8%


8.6% East Asian & Native American

Ancestry Group	Percentage
Native American	6.6%
East Asian	0.0%
Nonspecific East Asian & Nati...	1.9%


1.7% Sub-Saharan African

### No need for a Ph.D.

Our scientists have analyzed [disease risk](#), calculated [ancestry](#), and found [relatives](#) for genotyped customers. You could use this data without even knowing what a gene is!

## KNEES

THEY SUPPORT US UNTIL THEY DON'T  
DID YOU KNOW YOU MIGHT HAVE A  
PROPENSITY FOR HIGHER MUSCLE ST  
SPECIFICALLY TESTED ON KNEE FLEX  
AND EXTENSION?

[LOGIN WITH 23ANDME](#)

### Build novel apps

We have [Github](#) examples in Python, Javascript, and others, to get you started quickly, as well as a [forum](#) to ask questions. Build novel apps on the human genome.



# Decision Problem

---

## Elements of decision problem

### Example: to buy a car

➤ questions

What is my major goal?

What are the decisions I can make?

Are there any limits?

➤ answers

To save as much money as possible.

Trademark, model and colour of the car.

The car must be a HYBRID car.

# Elements of decision problem

## Example: to buy a car

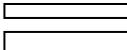
➤ questions

What is my major goal?

➤ answers

To save as much money as possible.

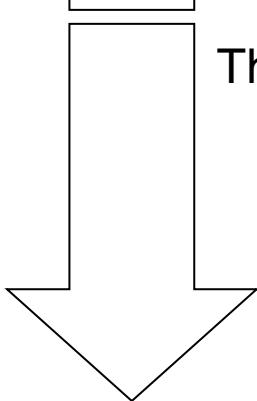
What are the decisions I can make?



Trademark, model and colour of the car.

Are there any limits?

The car must be a HYBRID car.



OBJECTIVE: **f**

---

## Elements of decision problem

### Example: to buy a car

➤ questions

What is my major goal?

What are the decisions I can make?

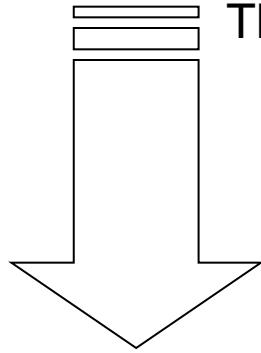
Are there any limits?

➤ answers

To save as much money as possible.

Trademark, model and colour of the car.

The car must be a HYBRID car.



DECISIONS: *x*

---

## Elements of decision problem

### Example: to buy a car

➤ questions

What is my major goal?

What are the decisions I can make?

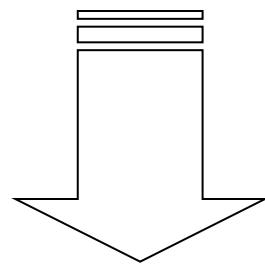
Are there any limits?

➤ answers

To save as much money as possible.

Trademark, model and colour of the car.

The car must be a HYBRID car.



CONSTRAINTS: feasible region  $\Omega$

---

# Formulation of the decision problem

## Example: to buy a car

➤ questions

What is my major goal?

What are the decisions I can make?

Are there any limits?

➤ answers

To save as much money as possible.

Trademark, model and colour of the car.

The car must be a HYBRID car.

We must transform the answers into equations:

$x$  vector of  $n$  elements (decision variables) of a function

Planning/Management Problem

---

# Formulation of the decision problem

## Example: to buy a car

➤ questions

What is my major goal?

What are the decisions I can make?

Are there any limits?

➤ answers

To save as much money as possible.

Trademark, model and colour of the car.

The car must be a HYBRID car.

We must transform the answers into equations:

$$\begin{array}{ll}\min_{\boldsymbol{x}} & f(\boldsymbol{x}) \\ \text{s.t.} & \boldsymbol{x} \in \Omega\end{array}$$

$\boldsymbol{x}$  vector of  $n$  elements (decision variables) of a function

$f(\boldsymbol{x})$  objective to optimize (max or min)

$\Omega$  set of *feasible* solutions (which satisfy all constraints)

---

## Multi-objective problems

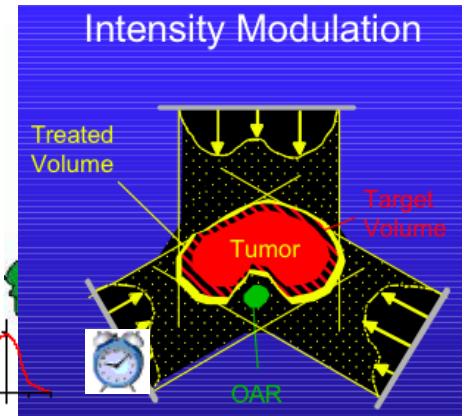
- A planning (or management) problem has very often a lot of conflicting objectives
- Sometimes we can reduce them to a single objective.  
For example, we can reduce all the objectives to a balance between costs and benefits in monetary terms (**Benefit-Cost Analysis**)
- In many cases, the objectives are incommensurable, or there are ethical reasons to refuse money transformation (e.g. human life). In this case we must use **Multi-Objective Analysis**
- In these cases, the decision must be made among a finite number of alternatives (discrete case) or within an infinite set (continuous case)

# Multi-objective optimization

- **Multi-objective optimization:** simultaneously optimizing two or more conflicting objectives subject to certain constraints

$$\begin{aligned} \text{minimize}_{\boldsymbol{x}} \quad & F(\boldsymbol{x}) = [f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \dots, f_k(\boldsymbol{x})] \\ \text{subject to} \quad & \boldsymbol{x} \in \Omega \end{aligned}$$

- Expects the objectives are typically not the same:
  - dollars, probability, units of time, ...
  - Finance: Minimize **risk** & Maximize **return**
  - Engineering: Maximize **car speed** & Maximize **safety**
- - Business: Minimize **cost** & Minimize **environmental impact**
- - Health Care: Maximize **X-ray dose to tumor** &  
Minimize **X-ray dose to healthy tissues**



---

## Multi-objective example: discrete case

Reclamation of a polluted area (planning problem)

There are  $n$  proposals

How we can choose?

There are  $m$  objectives:

- minimum cost;
- minimum realisation time;
- minimum risk;
- ...

---

## Multi-objective example: continuous case

Management of the  
water resources of a lake  
(management problem)

There are infinite release  
rules for the water

How we can choose?

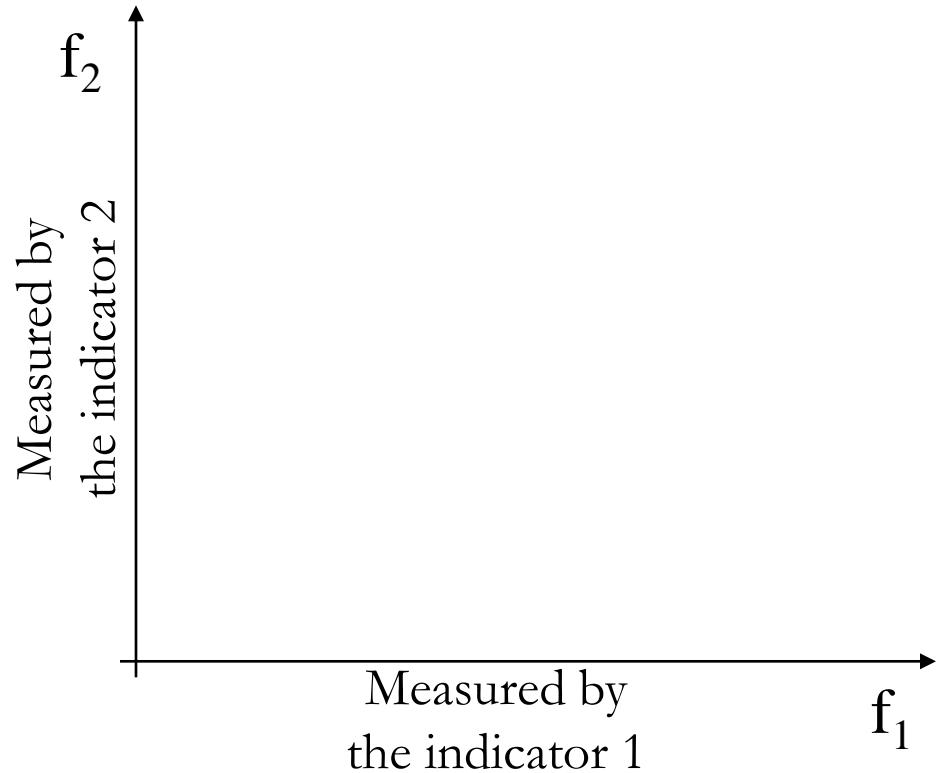
There are **3** objectives:

1. minimum flood damage
2. minimum water supply deficit
3. minimum environmental impact

# Multi-objective optimization: graphical visualization in objective space

If there are only 2 objectives, we can easily represent the solutions on a plane.

We need to define an indicator for each objective, to quantify the performances of the different solutions.



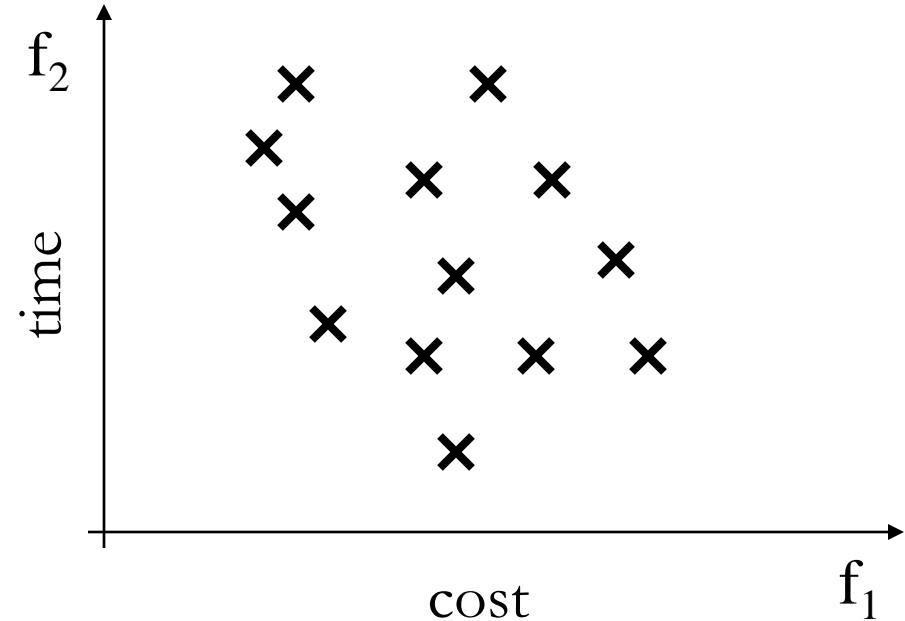
## Multi-objective example: discrete case

Reclamation of a polluted area

There are  $n$  proposals

There are  $m$  objectives:

- minimum cost;
- minimum realisation time;
- minimum risk;
- ...



Indicators (2):

- cost: millions of dollars
- time: realization years

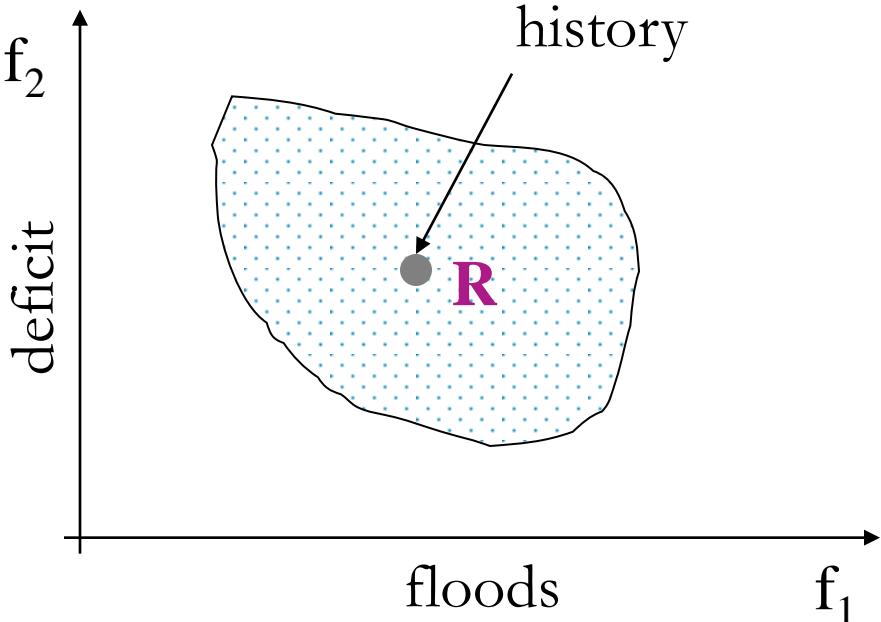
## Multi-objective example: continuous case

Management of the water resources of a lake

There are infinite release rules for the water

There are **3** objectives:

1. minimum flood damages
2. minimum water supply deficit
3. minimum environmental impact

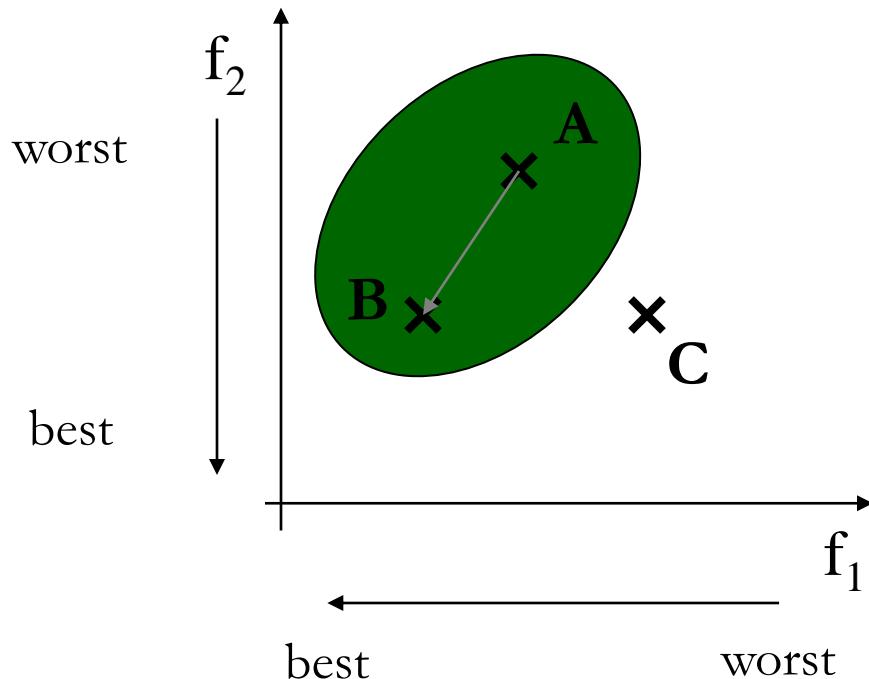


Indicators (**2**):

- days of flood
- deficit volume

## Multi-objective optimization: Pareto efficiency

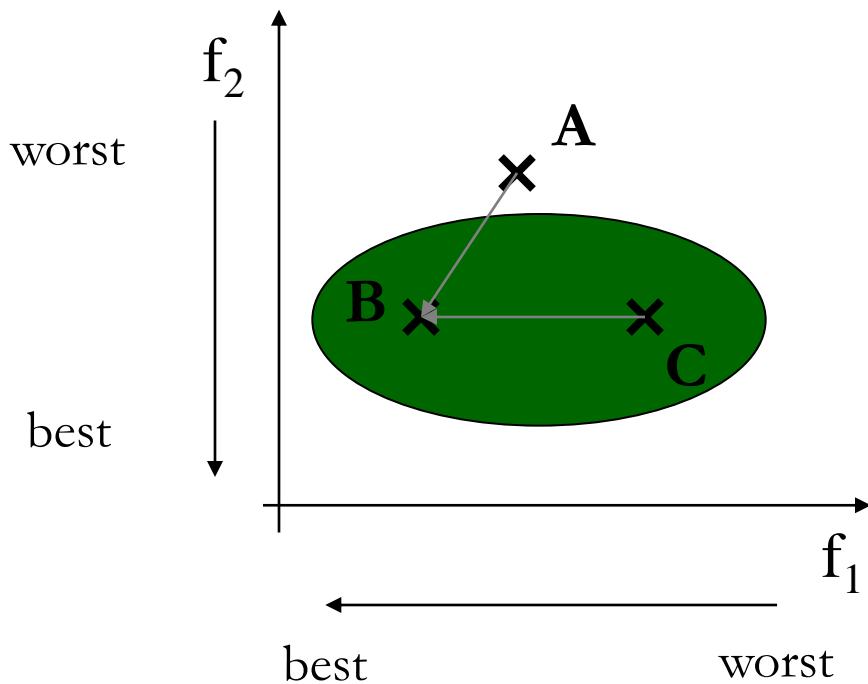
The solutions are not all equivalent for the decision-maker



- For example, if we must choose among 3 solutions...
- Given 2 objectives to minimize...  
B is preferred to (**strictly dominating**) A:  
 $f_1(B) < f_1(A)$  and  
 $f_2(B) < f_2(A)$

# Multi-objective optimization: Pareto efficiency

The solutions are not all equivalent for the decision-maker



➤ For example, if we must choose among 3 solutions...

➤ Given 2 objectives to minimize...

**B** is definitely preferred to (**strictly dominating**) **A**:

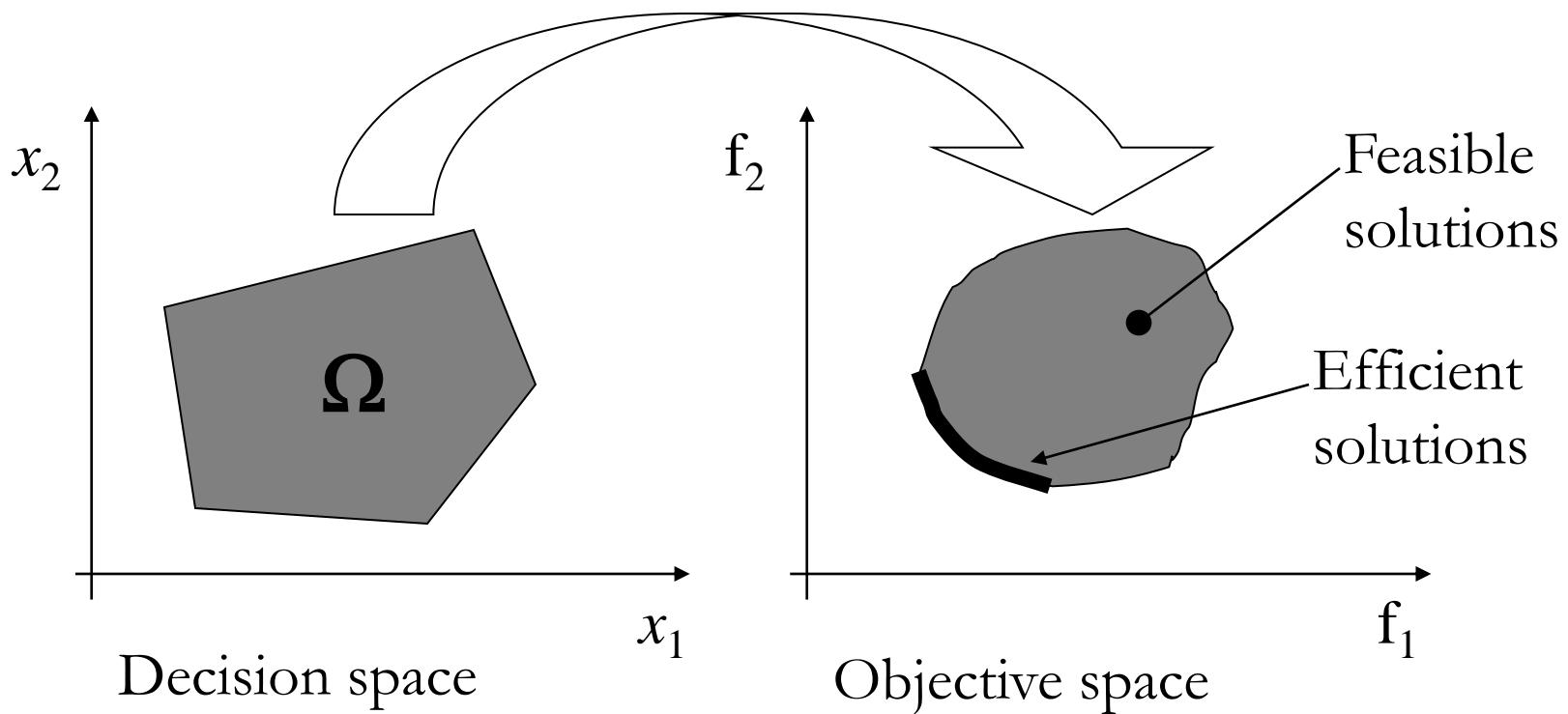
$$f_1(B) < f_1(A) \text{ and} \\ f_2(B) < f_2(A)$$

**B** is preferred to (semi dominating) **C**:

$$f_1(B) < f_1(C) \text{ and} \\ f_2(B) = f_2(C)$$

## Multi-objective optimization: Pareto efficiency

A solution is efficient whenever  
it is not dominated



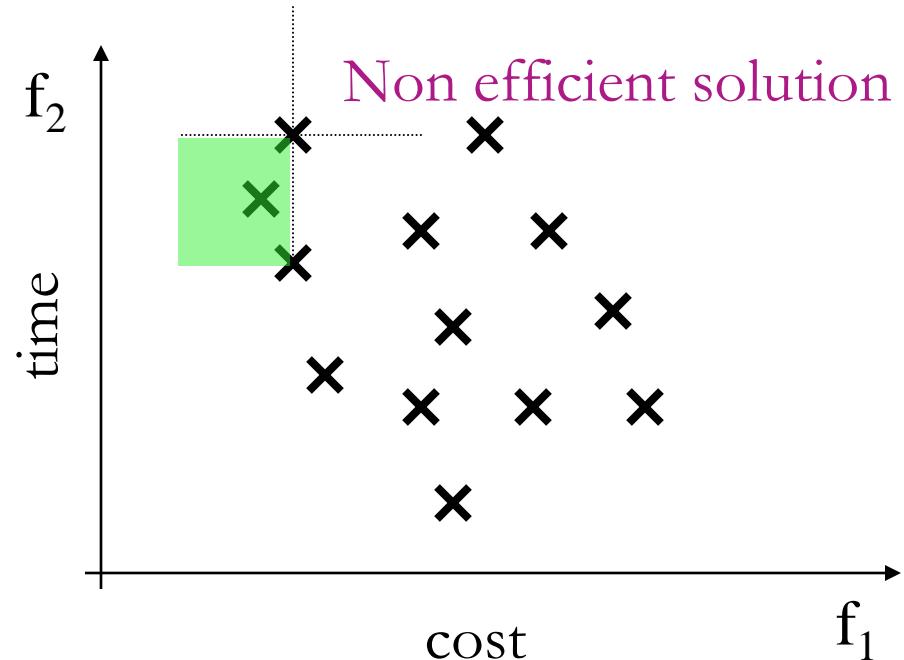
## Multi-objective example: discrete case

Reclamation of a polluted area

There are  $n$  proposals

There are  $m$  objectives:

- minimum cost;
- minimum realisation time;
- minimum risk;
- ...



Indicators (2):

- cost: millions of euros
- time: realization years

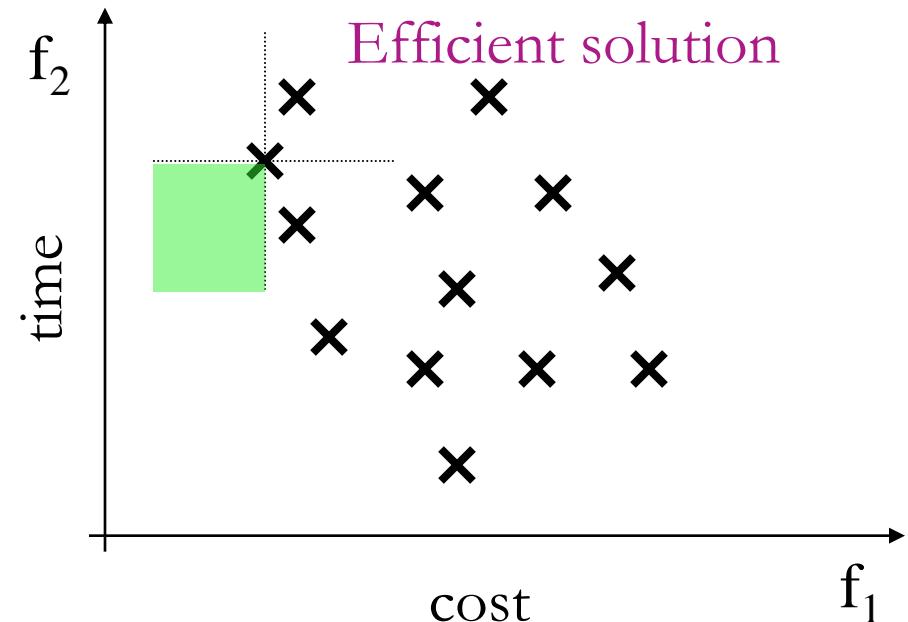
## Multi-objective example: discrete case

Reclamation of a polluted area

There are  $n$  proposals

There are  $m$  objectives:

- minimum cost;
- minimum realisation time;
- minimum risk;
- ...



Indicators (2):

- cost: millions of euros
- time: realization years

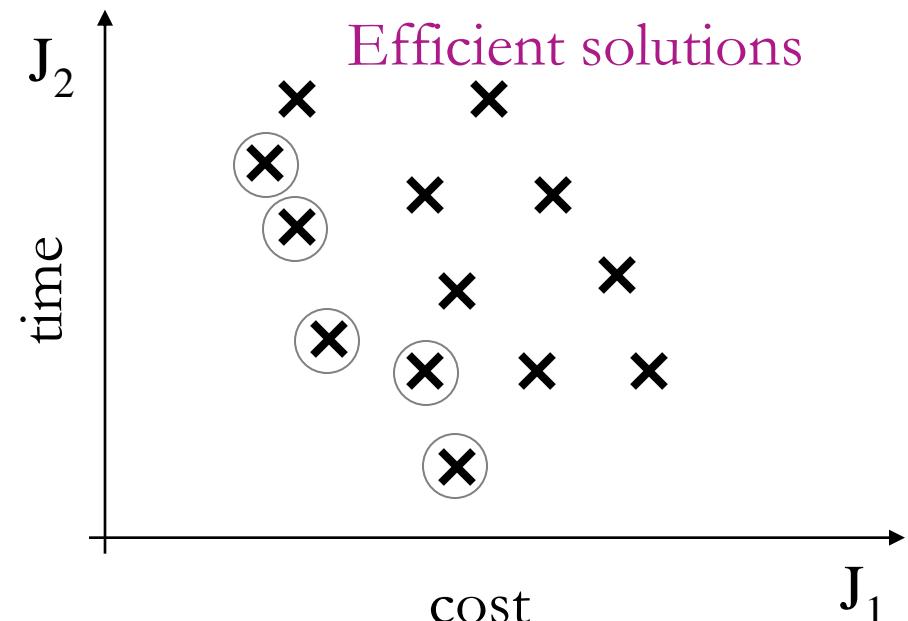
## Multi-objective example: discrete case

Reclamation of a polluted area

There are  $n$  proposals

There are  $m$  objectives:

- minimum cost;
- minimum realisation time;
- minimum risk;
- ...



Indicators (2):

- cost: millions of euros
- time: realization years

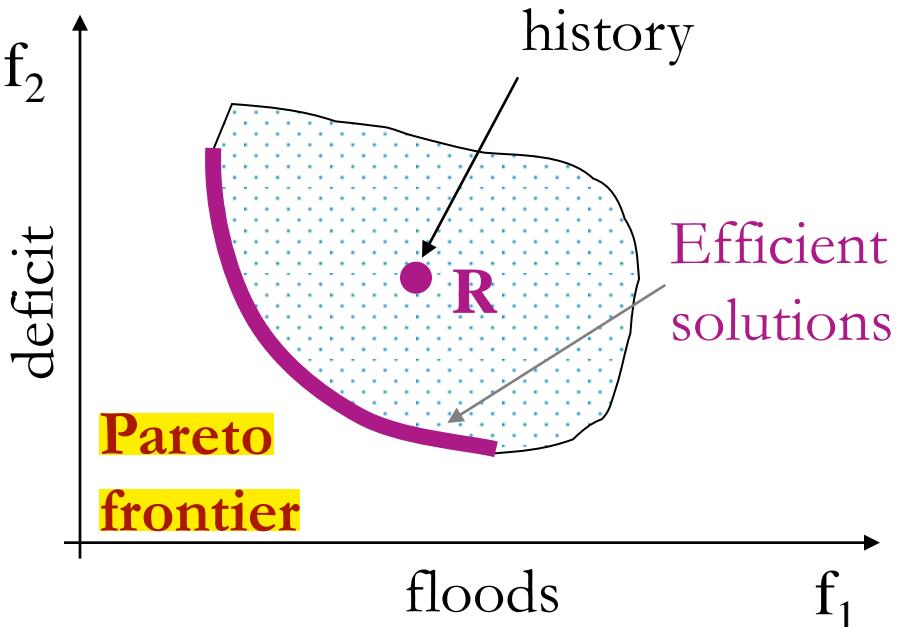
## Multi-objective example: continuous case

### Management of the water resources of a lake

There are infinite release rules for the water

There are **3** objectives:

1. minimum flood damages
2. minimum water supply deficit
3. minimum environmental impact



Indicators (**2**):

- days of flood
- deficit volume

---

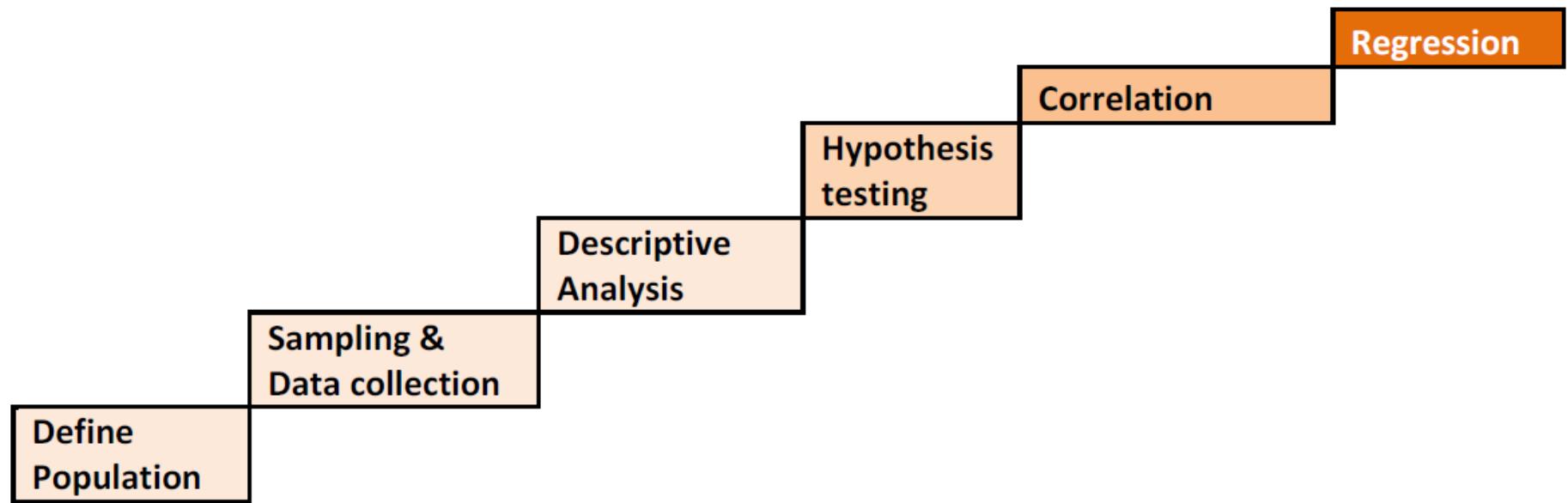
# Solutions of multi-objective problems

1. Formalization of the problem:
  - Which are the decision variables?
  - Which are the objectives?
  - Which are the constraints?
2. Research of feasible set of solutions
3. Research of efficient solutions:
  - Linear weighting method
  - Constraint method
4. Choice of best compromise



# Linear Regression – Statistical Perspective

# Linear regression – statistical perspective



# Linear regression – statistical perspective

## How is cab fare calculated?

You sit in the cab and see a base fare that is constant

- \$3.25

The fare increases with each additional 143 m at a certain rate

- \$0.25 per 143 m

The fare increases for each 29 seconds of taxi being idle

- \$0.25 per additional 29 seconds

For each passenger in excess of 4

- \$2.00

---

# Linear regression – statistical perspective

## Put it in a formula

- Fare = Constant + Dist\_Rate x Distance + Wait\_Rate x Time + FourPlus x Four or more Passengers
- Fare =  $3.25 + 0.25 \times \text{Distance} + 0.25 \times \text{Idle Time} + 2 \times \text{Four\_plus}$
- A sample trip
  - 3 people
  - 6 kms
  - 4 minutes
- Adjusting time and distance
  - $6000 \text{ m} / 143 = 41.9$  segments of 143 m
  - $4(60)/29 = 8.3$  segments of 29 seconds
- **Fare = \$3.25 + \$0.25 x 41.9 + \$0.25 x 8.3 + \$2 x 0 = \$15.8**

---

# Linear regression – statistical perspective

We can do this by hand

Fare	Distance (km)	Time (seconds)	4+ pax
\$ 15.8	6	240	0
\$ 16.9	6.7	225	0
\$ 23.1	8.5	350	1
\$ 11.4	3.8	180	0
\$ 12.1	3.25	135	1
\$ 13.5	4.5	275	0

# Linear regression – statistical perspective

## Why to regress?

What if we don't know the rates?

	<del>\$0.25 for 143-m</del>	<del>\$0.25 for 29 seconds</del>	<del>\$2 for 4+ pax</del>
Fare	Distance (km)	Time (seconds)	4+ pax
\$ 15.8	6	240	0
\$ 16.9	6.7	225	0
\$ 23.1	8.5	350	1
\$ 11.4	3.8	180	0
\$ 12.1	3.25	135	1
\$ 13.5	4.5	275	0
$y$	$x_1$	$x_2$	$x_3$

Regression estimates the unknown rates

Fare = Constant + Dist\_Rate x Distance + Wait\_Rate x Time +  
FourPlus x Four or more Passengers

$$y = \beta_0 + \beta_1 * distance + \beta_2 * time + \beta_3 * pax + \epsilon$$

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \epsilon$$

# Linear regression – statistical perspective

## Things to remember

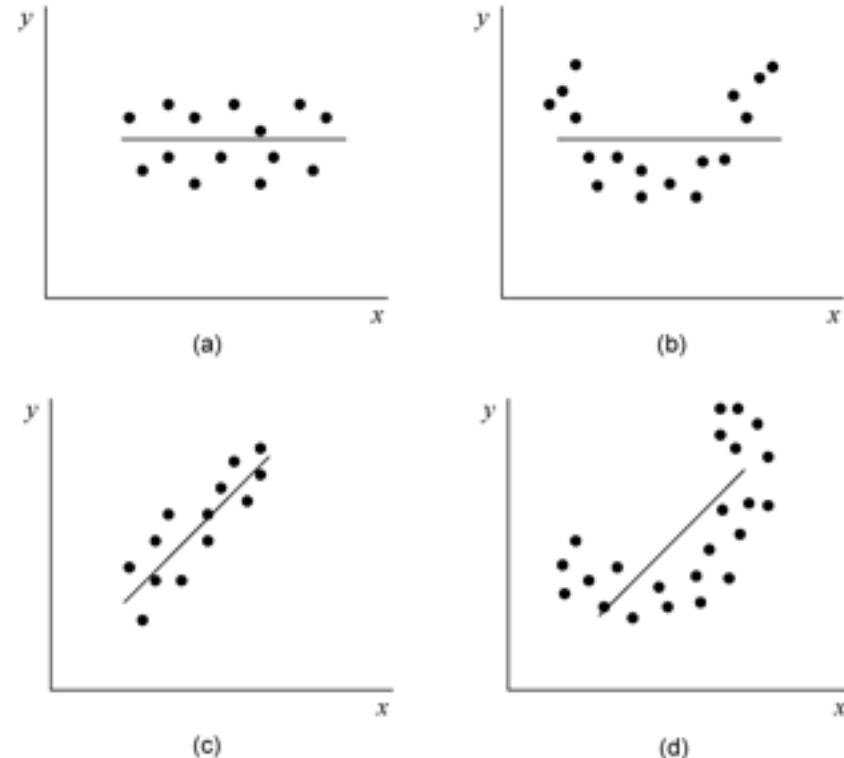
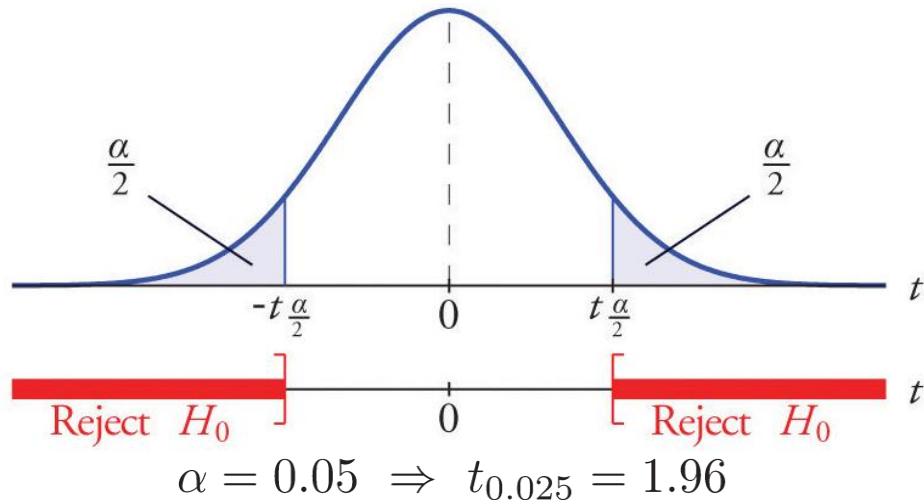
- **R-squared ( $r^2$ ):** The overall fit of the model is determined by R-squared or the adjusted R-squared. Its value varies between 0 and 1. When comparing models, a higher R-squared means a better fit.  
R-squared is a statistic that will give some information about the goodness of fit of a model. In regression, the R<sup>2</sup> coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. An R<sup>2</sup> of 1 indicates that the regression predictions perfectly fit the data.
- **P-value:** The p-value is the probability associated with a statistical test. Usually, we use 95% level as our benchmark. Hence, we conclude a statistically significant finding when the p-value associated with a test is less than 0.05.  
A small p-value (typically  $\leq 0.05$ ) indicates strong evidence against the null hypothesis, so you reject the null hypothesis. A large p-value ( $> 0.05$ ) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- **F-test:** When the p-value associated with the F-test is less than 0.05, we conclude that taken together the explanatory variables in the regression model are collectively different from 0.
- **T-statistic:** This evaluates the significance of an individual coefficient corresponding to a variable. If the t-test for a variable is greater than the critical value from the t-distribution (usually 1.96 for a two-tailed test; see Chapter 6 for details), we conclude that there exists a statistically significant relationship between the dependent and explanatory variable. Also, we can rely on the p-value, also reported in the regression output, for the corresponding t-test. If the p-value is less than 0.05, we can conclude a statistically significant relationship between the dependent and explanatory variables.

# Linear regression – statistical perspective

## Hypothesis testing

Null hypothesis: no linear relationship exists between independent variable  $x$  and dependant variable  $y$

$$H_0 : \beta_1 = 0$$



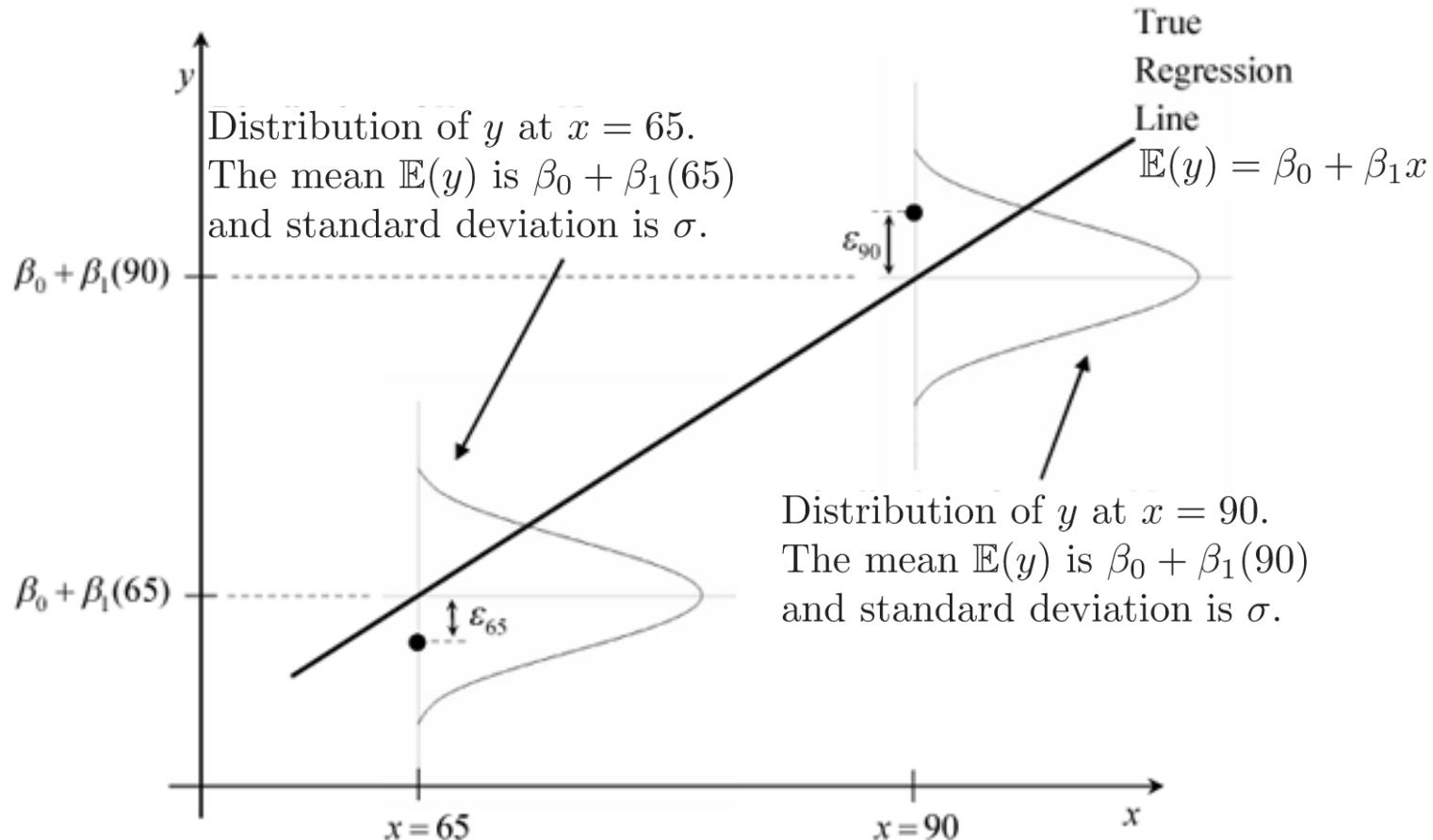
- a) no linear relationship exists ( $H_0$ )
- b) true relationship is not linear
- c) linear relationship exists ( $H_0$  rejected)
- d) a higher order model may be needed

# Linear regression – statistical perspective

## Distribution of errors

Random error term  $\epsilon$  is assumed to follow Normal distribution  $\mathcal{N}(0, \sigma^2)$

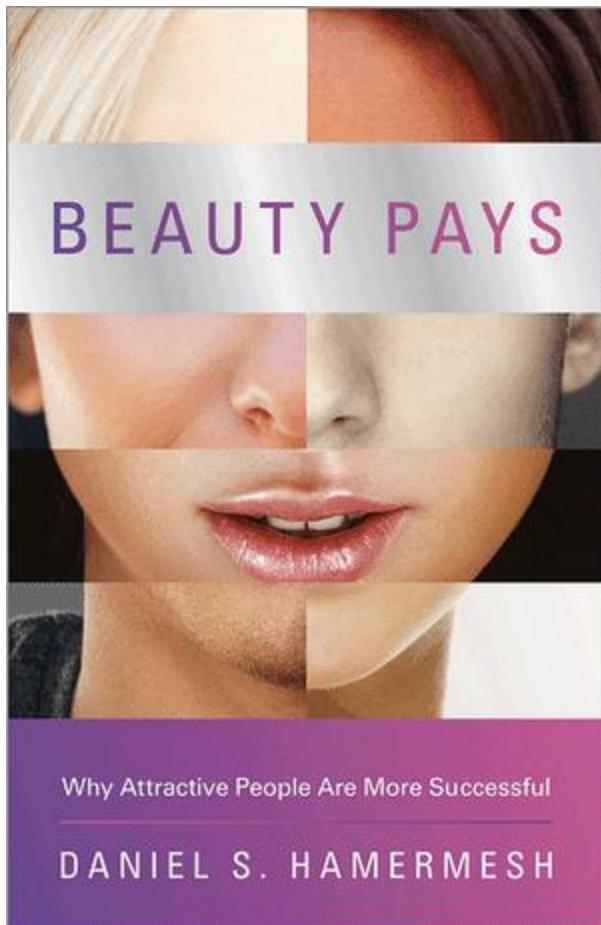
At any  $x^{(j)}$ ,  $y$  follows Normal distribution with mean  $\beta_0 + \beta_1 x^{(j)}$  and st.dev  $\sigma$



# Linear regression – statistical perspective

## Applications

Do good looking professors get higher teaching evaluations?



NBER WORKING PAPER SERIES

BEAUTY IN THE CLASSROOM:  
PROFESSORS' PULCHRITUDE AND  
PUTATIVE PEDAGOGICAL PRODUCTIVITY

Daniel S. Hamermesh  
Amy M. Parker

Working Paper 9853  
<http://www.nber.org/papers/w9853>

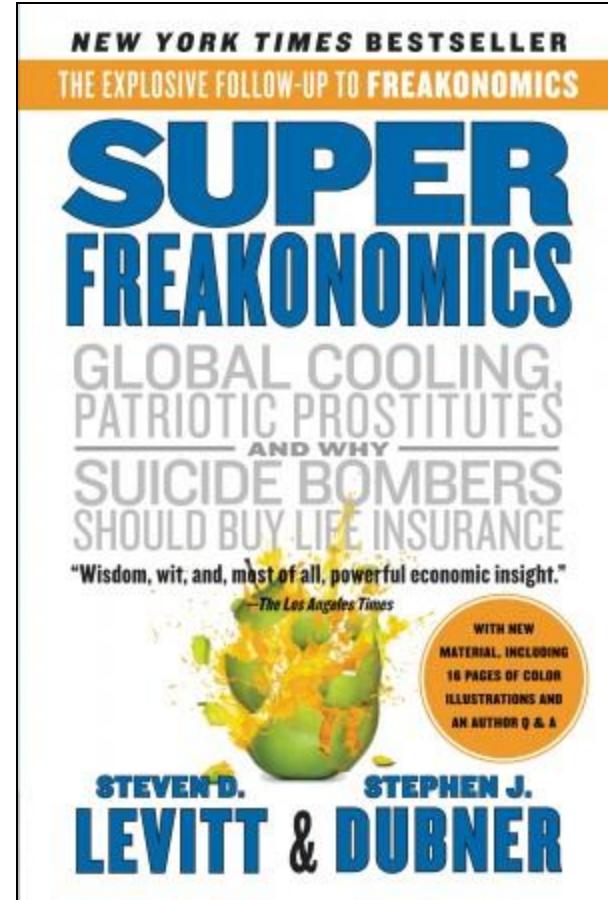
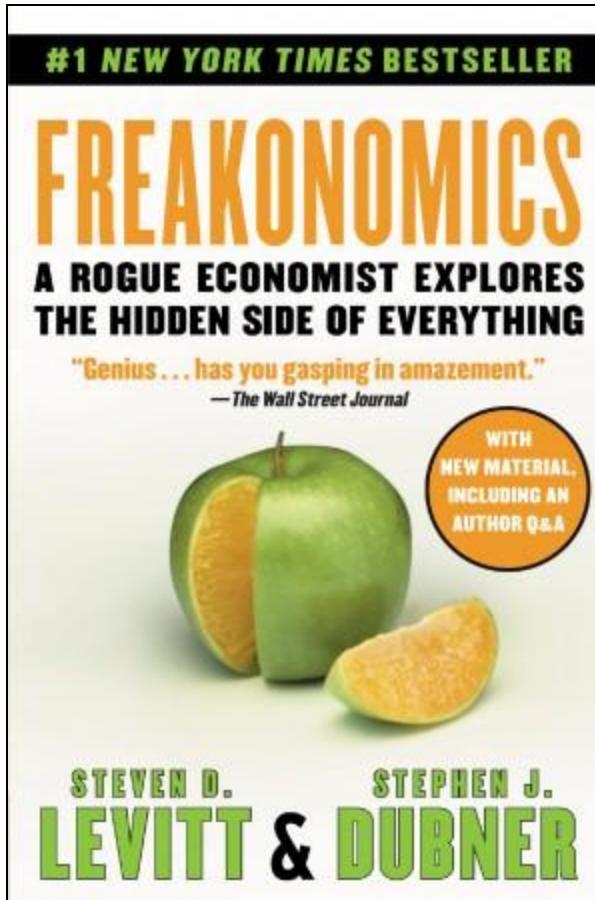
NATIONAL BUREAU OF ECONOMIC RESEARCH  
1050 Massachusetts Avenue  
Cambridge, MA 02138  
July 2003

Hamermesh, Daniel S. and Amy M. Parker (2005)  
“Beauty in the Classroom: Instructors' Pulchritude and  
Putative Pedagogical Productivity”, *Economics of  
Education Review*, August 2005, pp. 5-16

# Linear regression – statistical perspective

## Applications

### Regression analysis in Freakonomics books Econometrics



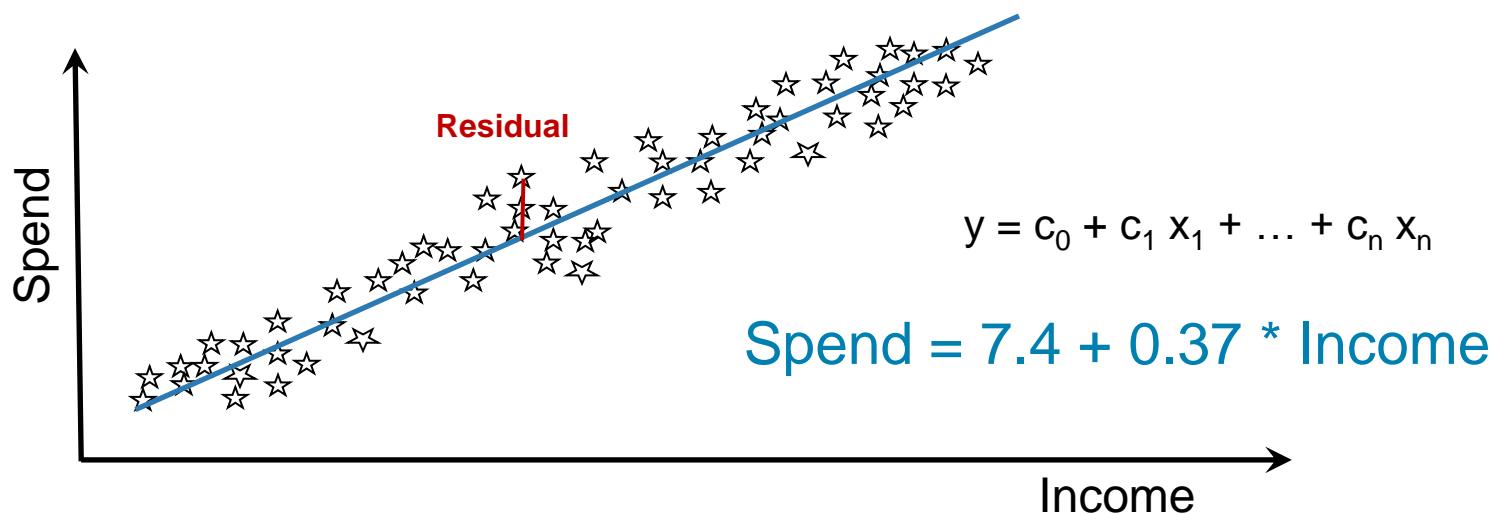
---



# **Linear Regression – Data Science Perspective**

# Linear regression

- Predict a value of a given continuous variable based on the values of other variables, assuming a linear or nonlinear model of dependency
- Virtually endless applications:
  - Election outcomes
  - Future product revenues or commodity prices
  - Wind velocity
- ✓ Both predictive and explanatory power



---

# Linear regression – usage

## Two major categories where regression analysis can be leveraged

1. To predict, estimate or forecast the values: linear regression can be used to fit a predictive model to an observed data set of  $y$  and  $x$  values
  - The model is fit to a set of known values using “training” data set and validated using a holdout “test” sample
  - Once the model is built it can be leveraged to predict the  $y$  values for the records where only  $x$  values are available
    - Estimate customer spend (scoring the universe)
    - Estimate future product demand (forecasting)
2. To quantify the strength of the relationship between  $y$  and the  $x_i$ . To assess which  $x_j$  has a strong relationship or whether a particular  $x_k$  has a statistically significant relationship with a target variable
  - The model is fit to a set of known values using “training” data set and validated using a holdout “test” sample, then the relationship between the target variable  $y$  and predictors  $x_i$  is evaluated.
    - Did the patients who received treatment (actual medication) showed statistically significant improvement vs. patients receiving a placebo?
    - What was the biggest driver in customer response (TV ad, banner, direct mail, email)

---

# Estimation / prediction

## Prediction example

- Can we predict the CO<sub>2</sub> emission of car without testing it?
- The CO<sub>2</sub> emission of a car is calculated based on the engine size, class, model, make, cylinder, fuel consumption of that car. Prediction is used to predict its expected CO<sub>2</sub> emission.

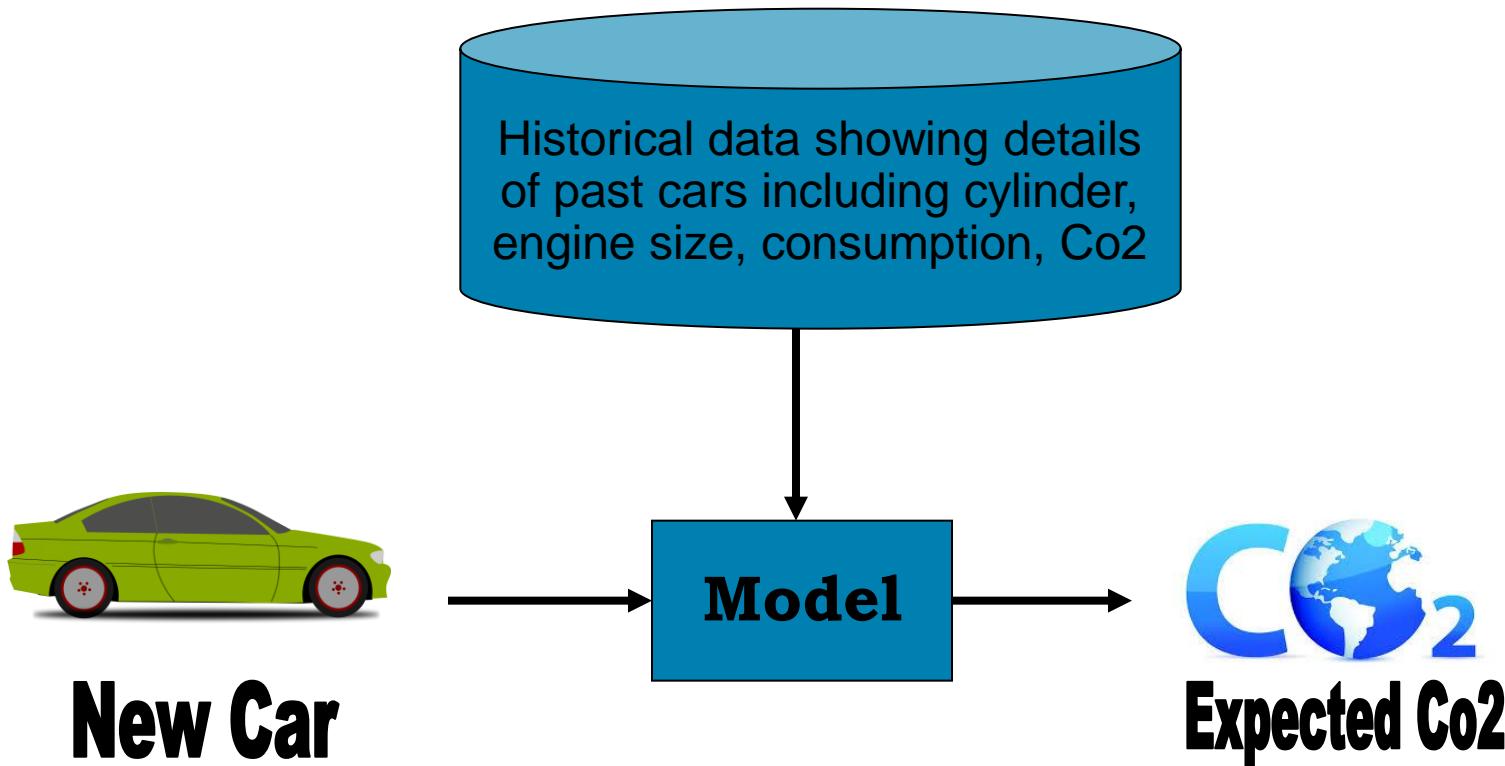
## What is a prediction?

**Prediction** is similar to classification but models are Continuous / Numerical / Ordered valued

## How does it work?

1. Split your data into **training** and **test** set
2. Construct a **model** using training set
3. **Evaluate** your model using test set
4. Use model to **predict** unknown value

## Estimation / prediction



Predict expected CO<sub>2</sub> emissions for a new car

---

```
cdf=df[['ENGINESIZE','CYLINDERS','FUELCONSUMPTION_COMB','CO2EMISSIONS']]  
cdf.head()
```

	<b>ENGINESIZE</b>	<b>CYLINDERS</b>	<b>FUELCONSUMPTION_COMB</b>	<b>CO2EMISSIONS</b>
<b>0</b>	2.0	4	8.5	196
<b>1</b>	2.4	4	9.6	221
<b>2</b>	1.5	4	5.9	136
<b>3</b>	3.5	6	11.1	255
<b>4</b>	3.5	6	10.6	244

# Dataset

Features					Target
Cylinders	Engine Size	Cons	...	CO <sub>2</sub>	
2	3	3	...	112	$y^{(2)}$
1	4	1	...	125	
1	2	2	...	101	
2	3	3	...	108	
3	4	1	...	105	
4	2	2	...	102	
2	3	3	...	121	
1	2	4	...	?	

Annotations:

- A blue bracket labeled "Features" spans the first five columns.
- A blue bracket labeled "Target" spans the last column.
- A blue brace on the left side groups the first two rows as  $x_2^{(1)}$ .
- A blue brace on the left side groups the first three rows as  $x_1^{(3)}$ .
- A blue brace on the right side groups the first four rows as the "Training set".
- A blue brace on the right side groups the last three rows as the "Test / Eval set".
- A blue brace on the right side groups the last row as the "Prediction set".

---

## Creating train and test dataset

---

```
msk = np.random.rand(len(df)) < 0.8
train = cdf[msk]
test = cdf[~msk]
```

---

---

# Prediction

- **Algorithms:**

- Regression
  - Simple regression
  - Multiple regression
  - Linear regression
  - Non-linear regression
- $k$ -nearest neighbor methods
- Neural Networks
- Support Vector Regression

- **What is regression analysis?**

- Simple regression analysis (one independent variable)
- Multiple regression analysis (multiple independent variables)

- **What are data science applications:**

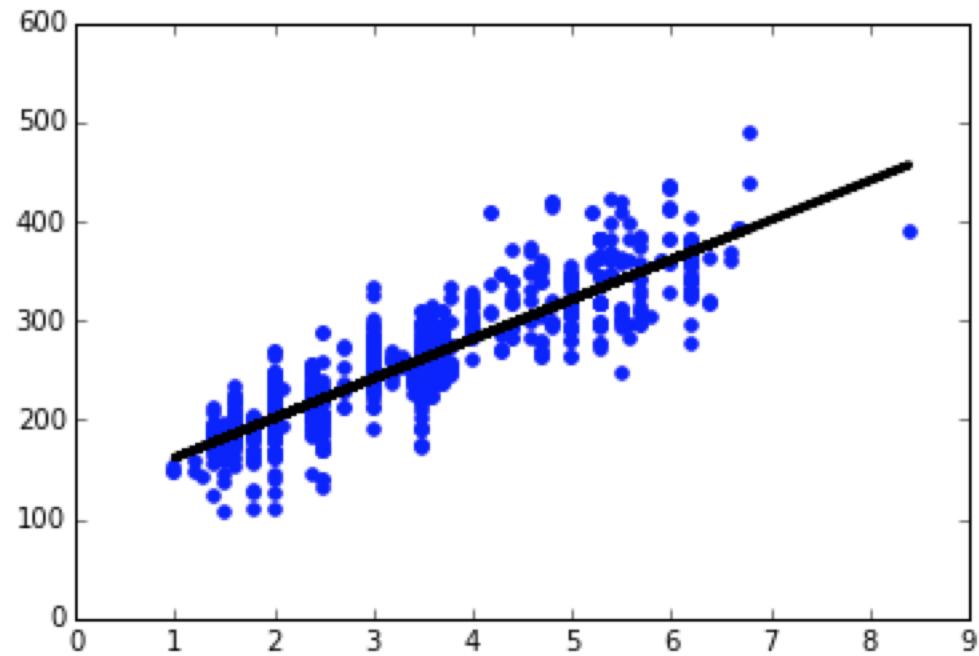
- Marketing: sales forecasting
- Psychology: satisfaction analysis
- ...

## Simple linear models

- Target value is expected to be a linear combination of the input variables (straight line)
- Regression computes  $\beta_i$  from data to minimize squared error to ‘fit’ the data

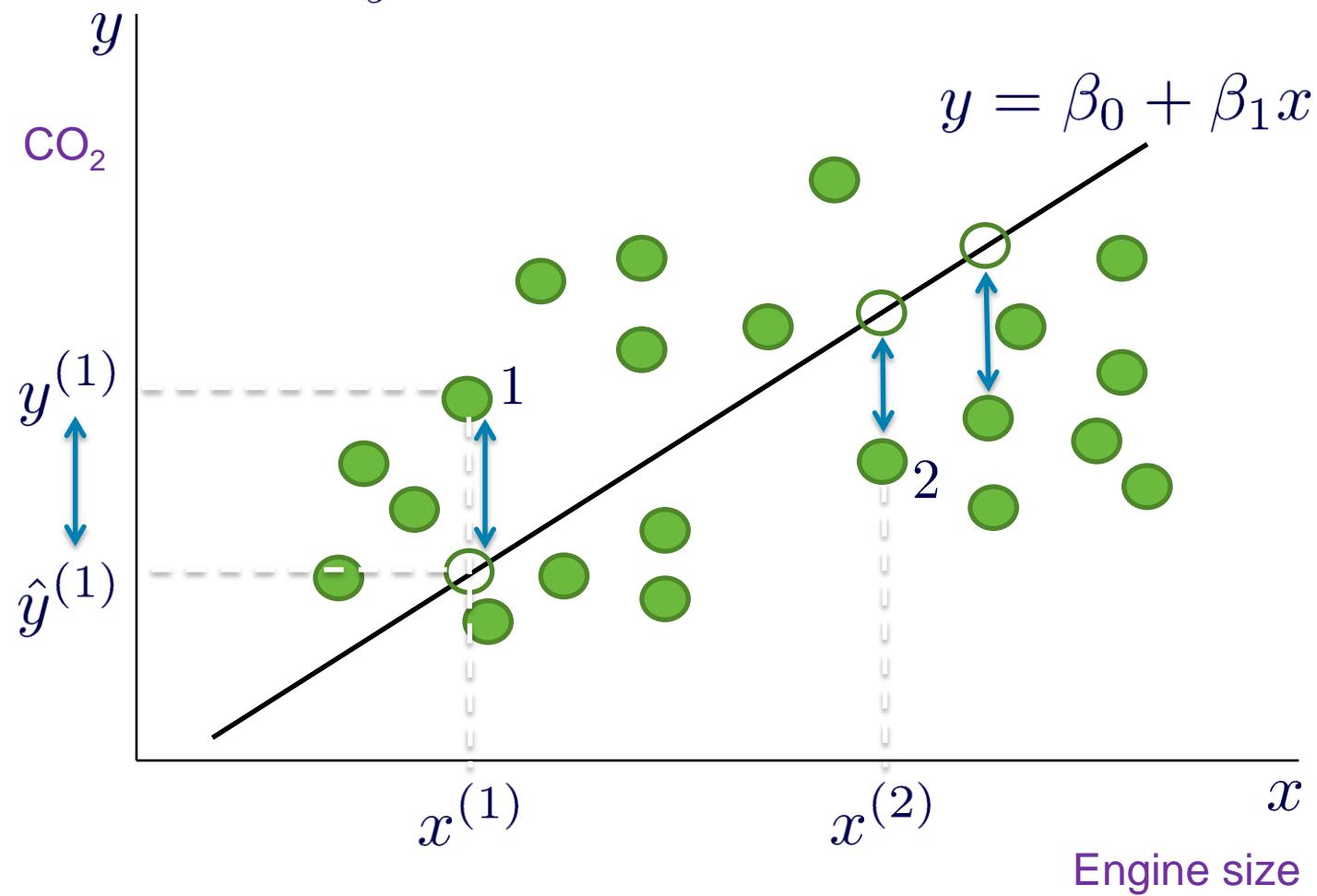
$y = \beta_0 + \beta_1 x$

↑  
↑  
a single predictor  
response variable



## Linear regression – ordinary least squares (OLS)

$$\min_{\beta_0, \beta_1} \left( y^{(1)} - \underbrace{(\beta_0 + \beta_1 x^{(1)})}_{\hat{y}^{(1)}} \right)^2 + \left( y^{(2)} - (\beta_0 + \beta_1 x^{(2)}) \right)^2 + \dots$$



---

## Linear regression – least squares

- Linear regression fits a linear model with coefficients  $\beta = (\beta_0, \beta_1, \beta_2, \dots)^T$  to minimize the residual sum of squares between the observed responses ( $y$ ) in the dataset, and the responses predicted by the linear approximation ( $\hat{y}$ )
- Mathematically it solves a problem of the form:

$$\min_{\beta} \| \mathbf{X}\beta - \mathbf{y} \|_2^2$$

- Method of least squares estimates the best-fitting straight line
- We will talk later **how** to compute  $\beta = (\beta_0, \beta_1, \beta_2, \dots)^T$  that minimizes the residual sum of squares

```
from sklearn import linear_model  
regr = linear_model.LinearRegression()  
train_x=np.asarray(train[['ENGINESIZE']])  
train_y=np.asarray(train[['CO2EMISSIONS']])  
regr.fit (train_x, train_y)  
# The coefficients  
print 'Coefficients: ', regr.coef_  
print 'Intercept: ',regr.intercept_
```

```
Coefficients: [[ 38.68021623]  
Intercept: [ 126.61208813]
```

可以用 model.coef\_ 和  
model.intercept\_ 来查看  
系数

---

## Multiple linear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

- $y \Rightarrow$  dependent / target / predicted / response variable
- $x_1, x_2, \dots \Rightarrow$  independent / input / feature / *predictor* variable
- $\beta_1, \beta_2, \dots \Rightarrow$  coefficients of dependent variables
- $\beta_0 \Rightarrow$  intercept
- $\beta_1 > 0 \Rightarrow$  positive association
- $\beta_1 < 0 \Rightarrow$  negative association
- $\beta_1 = 0 \Rightarrow$  no association

**Many nonlinear functions can be transformed into the above form**

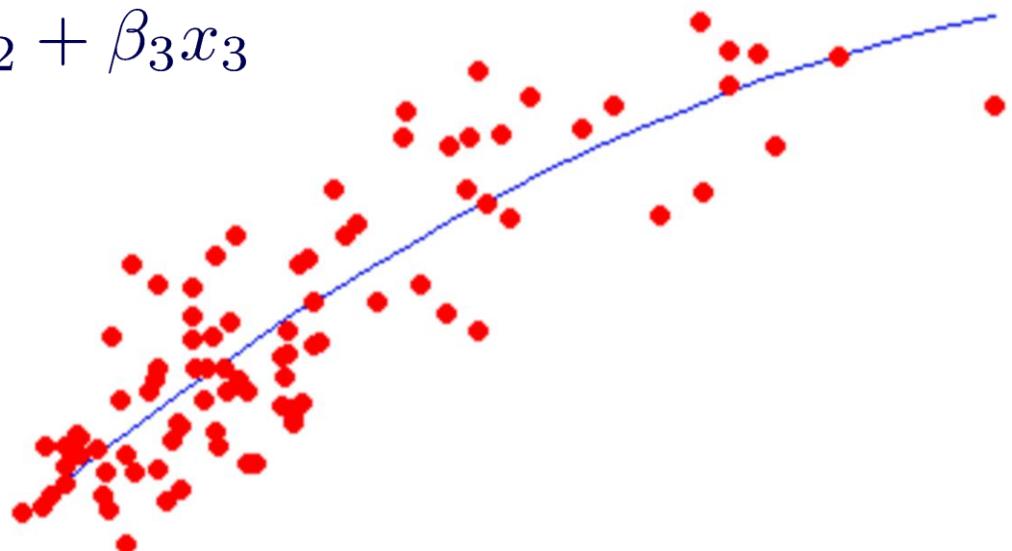
## Nonlinear regression

- Some nonlinear models can be modeled by a polynomial function
- A polynomial regression model can be transformed into linear regression model
- For example

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

is converted to linear regression with new variables  $x_2 = x^2$ ,  $x_3 = x^3$

$$y = \beta_0 + \beta_1 x + \beta_2 x_2 + \beta_3 x_3$$



# Accuracy

- Measure predictor accuracy – measure how far off the predicted value ( $\hat{y}$ ) is from the actual known value ( $y$ )
  - Loss function measures the error between  $y^{(j)}$  and the predicted value  $\hat{y}^{(j)}$ 
    - Absolute error  $|y^{(j)} - \hat{y}^{(j)}|$
    - Squared error  $(y^{(j)} - \hat{y}^{(j)})^2$
  - Test error (generalization error) – the average loss over the test set
    - Mean Absolute Error 
$$\frac{\sum_{j=1}^m |y^{(j)} - \hat{y}^{(j)}|}{m}$$
    - Mean Squared Error 
$$\frac{\sum_{j=1}^m (y^{(j)} - \hat{y}^{(j)})^2}{m}$$
  
(Residual Sum of Squares \*  $m$ )
    - Relative Absolute Error 
$$\frac{\sum_{j=1}^m |y^{(j)} - \hat{y}^{(j)}|}{\sum_{j=1}^m |y^{(j)} - \bar{y}|}$$
- `sklearn.metrics.r2_score(y_true, y_pred)`
- Relative Squared Error
  - $R^2 = (1 - \text{Relative Squared Error})$  
$$\frac{\sum_{j=1}^m (y^{(j)} - \hat{y}^{(j)})^2}{\sum_{j=1}^m (y^{(j)} - \bar{y})^2}$$

```
test_x=np.asanyarray(test[['ENGINESIZE']])
test_y=np.asanyarray(test[['CO2EMISSIONS']])
test_y_=regr.predict(test_x)

print("Residual sum of squares: %.2f"
      % np.mean((test_y_ - test_y) ** 2))
# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regr.score(test_x, test_y))
# Plot outputs
plt.scatter(test_x, test_y, color='blue')
plt.plot(test_x, test_y_, color='black', linewidth=3)
plt.xlabel("Engine size")
plt.ylabel("Emission")
plt.show()
```

Residual sum of squares: 874.93  
Variance score: 0.81

# Regression analysis: R<sup>2</sup> and variable selection

- Goodness of fit in linear regression models is generally measured by using the R<sup>2</sup>
- R<sup>2</sup> measures how well the regression line approximates the real data points, it also portrays percent of variance in the data explained by regression model
  - If the value is close to 1, the model fits perfectly and explains all variance
  - If the value is close to 0, then the model does not fit the data and doesn't explain any variance
- **Variable preparation:**
  - Interval variables can be binned or bucketed in order to capture nonlinear relationships
  - Categorical variables must be converted into binary vectors. Data sample must be large enough to accommodate all degrees of freedom
- **Variable selection** (LASSO algorithm):

$$\min_{\beta} \| \mathbf{X}\beta - \mathbf{y} \|_2^2 + \lambda \|\beta\|_1$$

↓

$$\min_{\beta} \| \mathbf{X}\beta - \mathbf{y} \|_2^2$$

subject to     $\|\beta\|_1 \leq \varepsilon$

---

## Bias-variance tradeoff

- Because  $\mathbf{X}\hat{\beta}$  fits our data well, this doesn't mean that it will be a good fit to new data
- Prediction error for new data  $\|\tilde{\mathbf{X}}\hat{\beta} - \tilde{y}\|_2^2$  should be small:

$$\text{PE}(\tilde{\mathbf{X}}) = \sigma_\epsilon^2 + \text{Bias}^2(\tilde{\mathbf{X}}\hat{\beta}) + \text{Var}(\tilde{\mathbf{X}}\hat{\beta})$$

- This decomposition is known as the **bias-variance tradeoff**:
  - As model becomes more complex (more terms included), model structure can be picked up
  - But coefficient estimates suffer from high variance as more terms are included in the model
- Introducing a little bias in our estimate for  $\beta$  might lead to a substantial decrease in variance, and hence to a substantial decrease in PE:
  - If the  $\beta_i$ 's are unconstrained they can explode and hence are susceptible to very high variance
  - To control variance, we might regularize the coefficients, i.e.,  $\min_{\beta} \|\beta\|$

# Bias-variance tradeoff

- Prediction error:

$$\text{PE}(\tilde{\mathbf{X}}) = \sigma_{\epsilon}^2 + \text{Bias}^2(\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}) + \text{Var}(\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}})$$

$$\text{PE}(\tilde{\mathbf{X}}) = \mathbb{E}[(\hat{\mathbf{y}} - \mathbf{y})^2] = \mathbb{E}[(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2]$$

**Bias - Variance Tradeoff**

$$\text{Error}(x) = \left( E[\hat{f}(x)] - f(x) \right)^2 + E[\hat{f}(x) - E[\hat{f}(x)]]^2 + \sigma_{\epsilon}^2$$

↓      ↓      ↓      ↓      ↓  
predicted    true    predicted    average predicted value    irreducible error

**Bias<sup>2</sup>**  
How much predicted values differ from true values.

**Variance**  
How predictions made on the same value vary on different realizations of the model

BY CHRIS ALBON

# Regularized regression

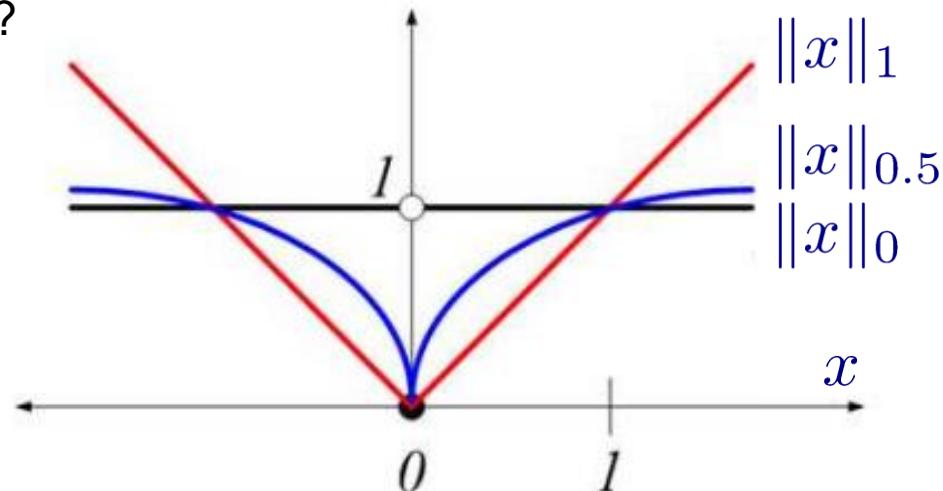
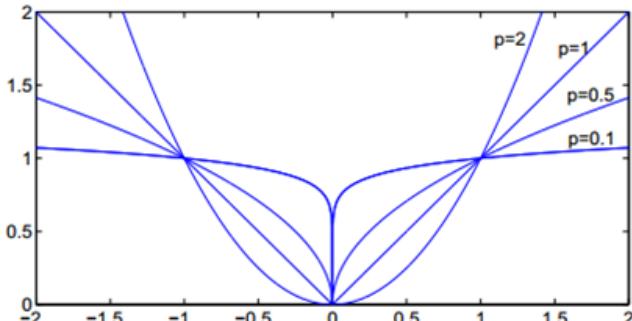
- Regularized regression is multi-objective optimization problem:

$$\begin{aligned}\min_{\beta} \quad & \|X\beta - y\|_2^2 && \leftarrow f_1(\beta) \\ \min_{\beta} \quad & \|\beta\| && \leftarrow f_2(\beta)\end{aligned}$$

- Two questions:

- How to solve this multi-objective optimization problem?
- Which norm function  $\|\cdot\|_p$  to choose?

$$\|x\|_p = \left( \sum_{i=1}^n |x_i|^p \right)^{1/p}$$



$$\text{card}(x) = \sum_i x_i^0 = \|x\|_0 \text{ with } 0^0 = 0$$

# Multi-objective optimization

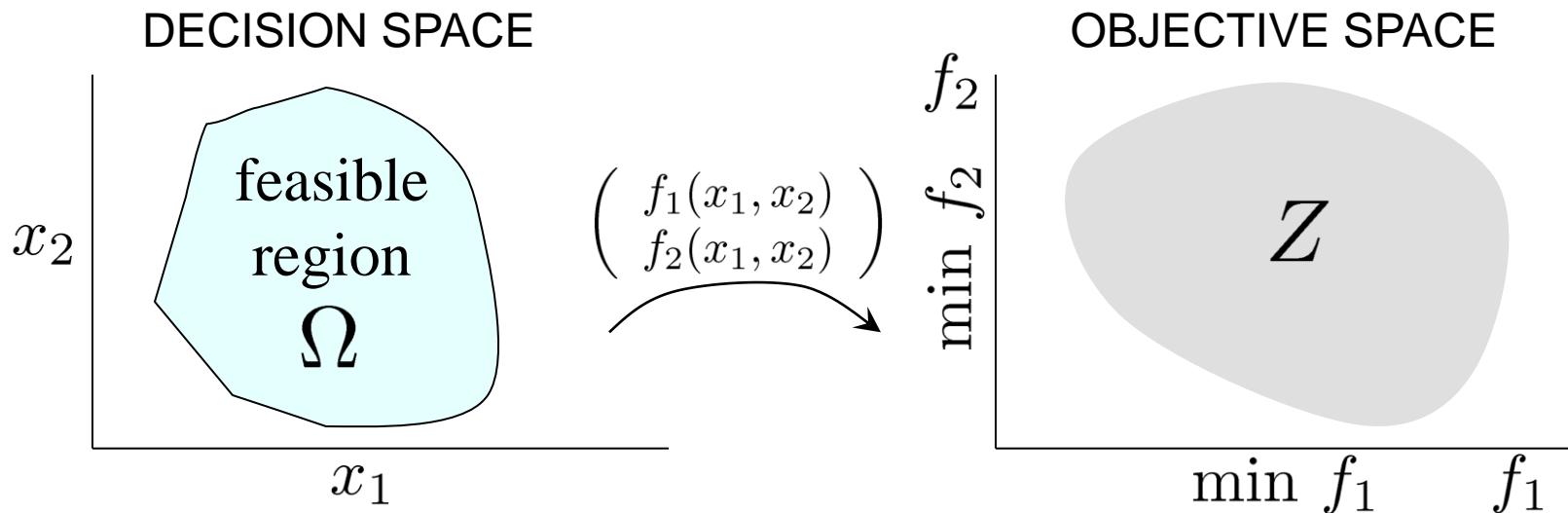
- Multi-objective optimization:

$$\begin{aligned} \min \quad & \{f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \dots, f_k(\boldsymbol{x})\} \\ \text{s.t.} \quad & \boldsymbol{x} \in \Omega \end{aligned}$$

$f_i : \mathbb{R}^n \rightarrow \mathbb{R}$  are (possibly) conflicting objectives and  $\Omega \subseteq \mathbb{R}^n$  is feasible region

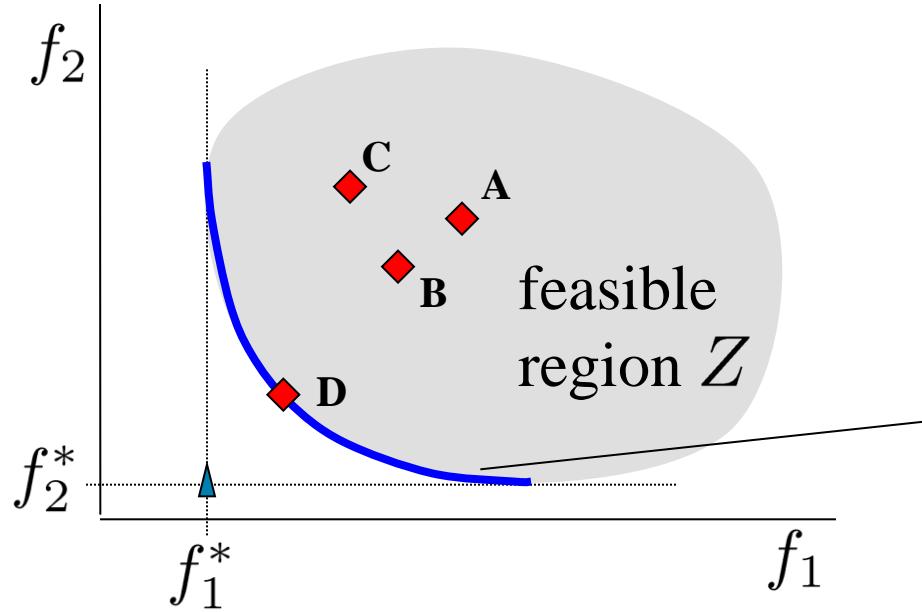
- Mapping feasible region into the objective space:

$$Z = \{\boldsymbol{z} \in \mathbb{R}^k : \boldsymbol{z} = ((f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \dots, f_k(\boldsymbol{x}))^T \forall \boldsymbol{x} \in \Omega)\}$$



## Bi-objective example

- $\min f_1 = \text{bias}$ ,  $\min f_2 = \text{variance}$  :



**Pareto frontier or efficient frontier**  
*(all non-dominated solutions)*

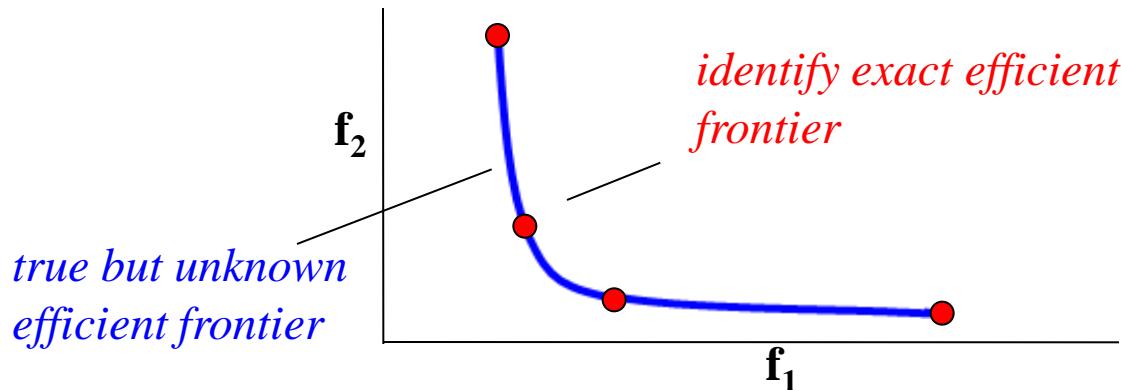
- Pareto efficiency: solutions with characteristics like **D**, are called tradeoff, Pareto optimal or non-dominated
- Multi-objective optimization goal: find solution(s) on the efficient frontier according to the decision maker preferences

## Computing efficient frontiers - possibilities

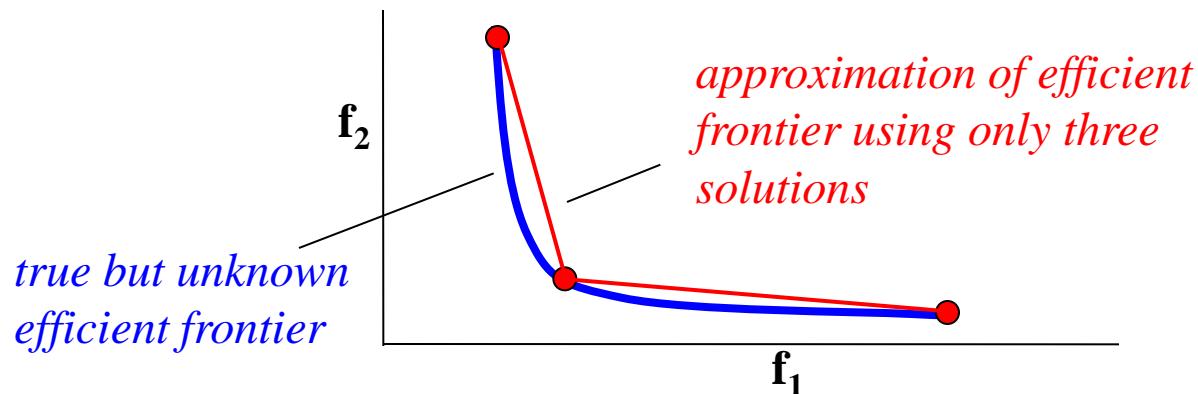
Multi-objective analysis involves **computing the efficient frontier**, evaluating it (if possible, out-of-sample) and selecting the final solution based on the decision maker preferences

### Computing efficient frontiers:

- Ideal (often unrealistic) goal: compute exact frontier



- Typical (more realistic) goal: approximate the frontier



---

# Solving multi-objective optimization problems

- Convert multi-objective optimization problem to a series of single-objective optimization problems
  
- Methods:
  - Weighting Method
  - $\varepsilon$ -Constraint (Hierarchical) Method

## Weighting method

- Assign weights to each objective
- Optimize the weighted sum of the objectives
- Multi-objective optimization with weighting method:

$$\begin{aligned} \min \quad & \omega_1 \cdot f_1(\mathbf{x}) + \omega_2 \cdot f_2(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \Omega \end{aligned}$$

$f_i$  is convex function (linear, convex quadratic)

$\Omega \subseteq \mathbb{R}^n$  (convex)

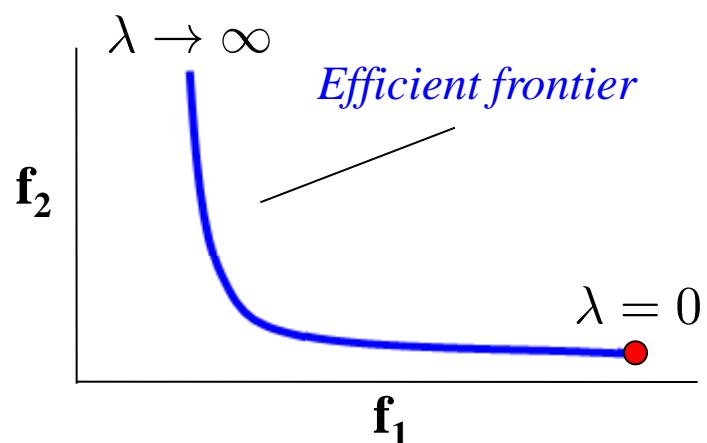
$\omega_i \in \mathbb{R}$  is the weight of the  $i$ -th objective

$\omega_i \geq 0$ ,  $i = 1, 2$  and  $\omega_1 + \omega_2 = 1$

- Easier formulation:

$$\downarrow \quad \lambda = \frac{\omega_1}{\omega_2}$$

$$\begin{aligned} \min \quad & \lambda f_1(\mathbf{x}) + f_2(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \Omega \end{aligned}$$



---

## $\varepsilon$ - constrained method

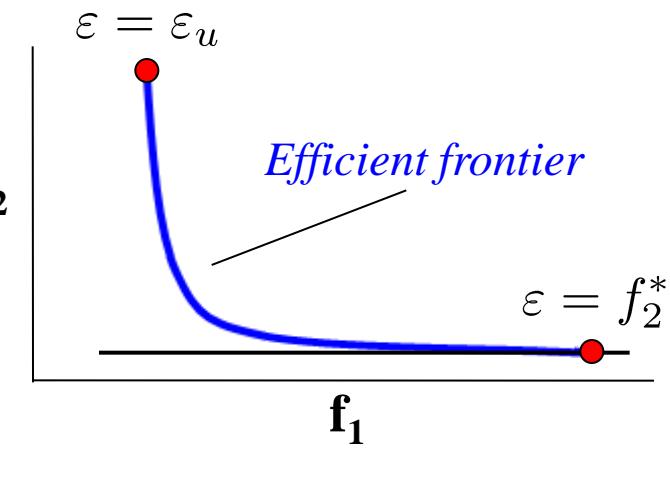
- Optimize one objective
- Convert other objectives into constraints
- Multi-objective optimization with  $\varepsilon$ -constrained method:

First step

$$\begin{array}{ll} \min & f_2(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in \Omega \\ & \downarrow \\ & f_2^* \end{array}$$

Second step

$$\begin{array}{ll} \min & f_1(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in \Omega \\ & f_2(\mathbf{x}) \leq \underbrace{(1 + \epsilon)f_2^*}_{\varepsilon} \end{array}$$



## Regularized regression

- **$L_1$  regularized regression and variable selection** (LASSO algorithm):

$$\min_{\beta} \| \mathbf{X}\beta - \mathbf{y} \|_2^2 + \lambda \|\beta\|_1$$

↓

$$\min_{\beta} \| \mathbf{X}\beta - \mathbf{y} \|_2^2$$

subject to     $\|\beta\|_1 \leq \varepsilon$

- **$L_2$  regularized regression or Tikhonov regularization** (Ridge regression):

$$\min_{\beta} \| \mathbf{X}\beta - \mathbf{y} \|_2^2 + \lambda \|\beta\|_2^2$$

↓

$$\min_{\beta} \| \mathbf{X}\beta - \mathbf{y} \|_2^2$$

subject to     $\|\beta\|_2^2 \leq \varepsilon$

---

## Regularized regression

- **$L_1$  regularized regression** (LASSO algorithm):

$$\min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda \|\beta\|_1$$

$$\min_{\beta} (\mathbf{X}\beta - \mathbf{y})^T (\mathbf{X}\beta - \mathbf{y})$$

subject to  $\sum_{i=1}^n |\beta_i| \leq \varepsilon$

- **$L_2$  regularized regression** (Ridge regression):

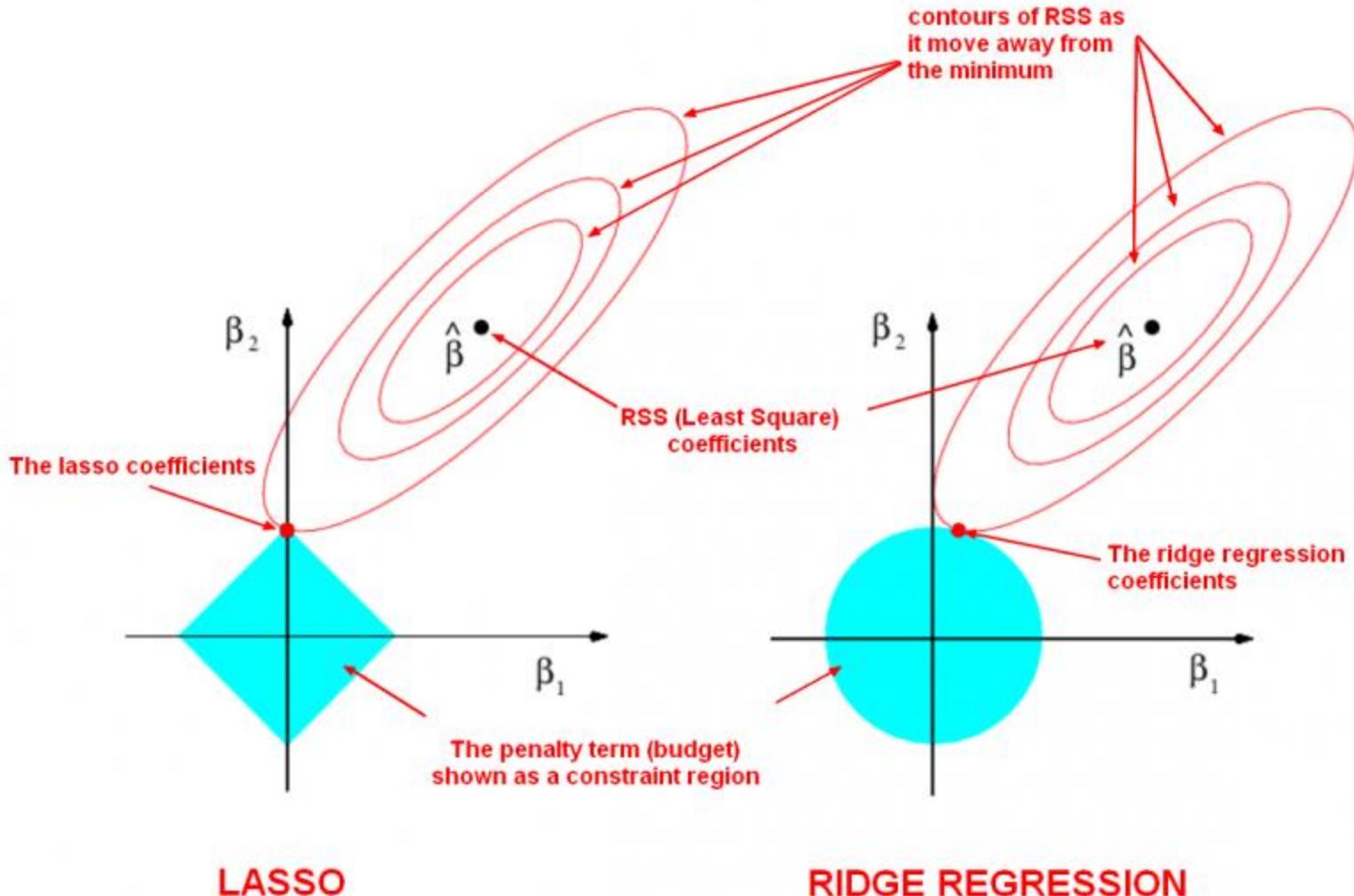
$$\|x\|_p = \left( \sum_{i=1}^n |x_i|^p \right)^{1/p}$$

$$\min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda \|\beta\|_2^2$$

$$\min_{\beta} (\mathbf{X}\beta - \mathbf{y})^T (\mathbf{X}\beta - \mathbf{y})$$

subject to  $\sum_{i=1}^n \beta_i^2 \leq \varepsilon$

# Regularized regression



## Regularized regression

- **$L_2$  regularized regression** (Ridge regression) has closed-form solution:

$$\min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda \|\beta\|_2^2$$



$$\beta = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

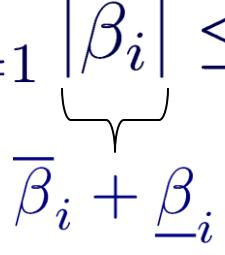
- **$L_1$  regularized regression** (LASSO algorithm) does not have a closed-form solution:

$$\min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda \|\beta\|_1$$



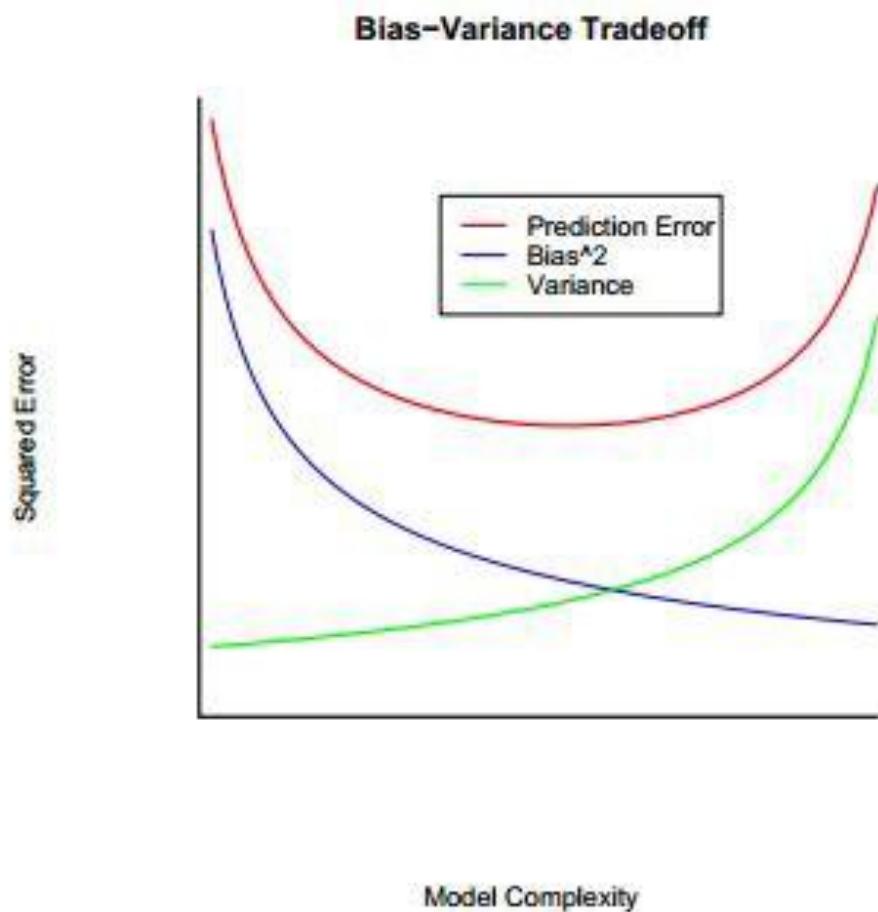
$$\min_{\beta} \beta^T (\mathbf{X}^T \mathbf{X}) \beta - 2 (\mathbf{X}^T \mathbf{y})^T \beta + \mathbf{y}^T \mathbf{y}$$

subject to  $\sum_{i=1}^n |\beta_i| \leq \varepsilon$

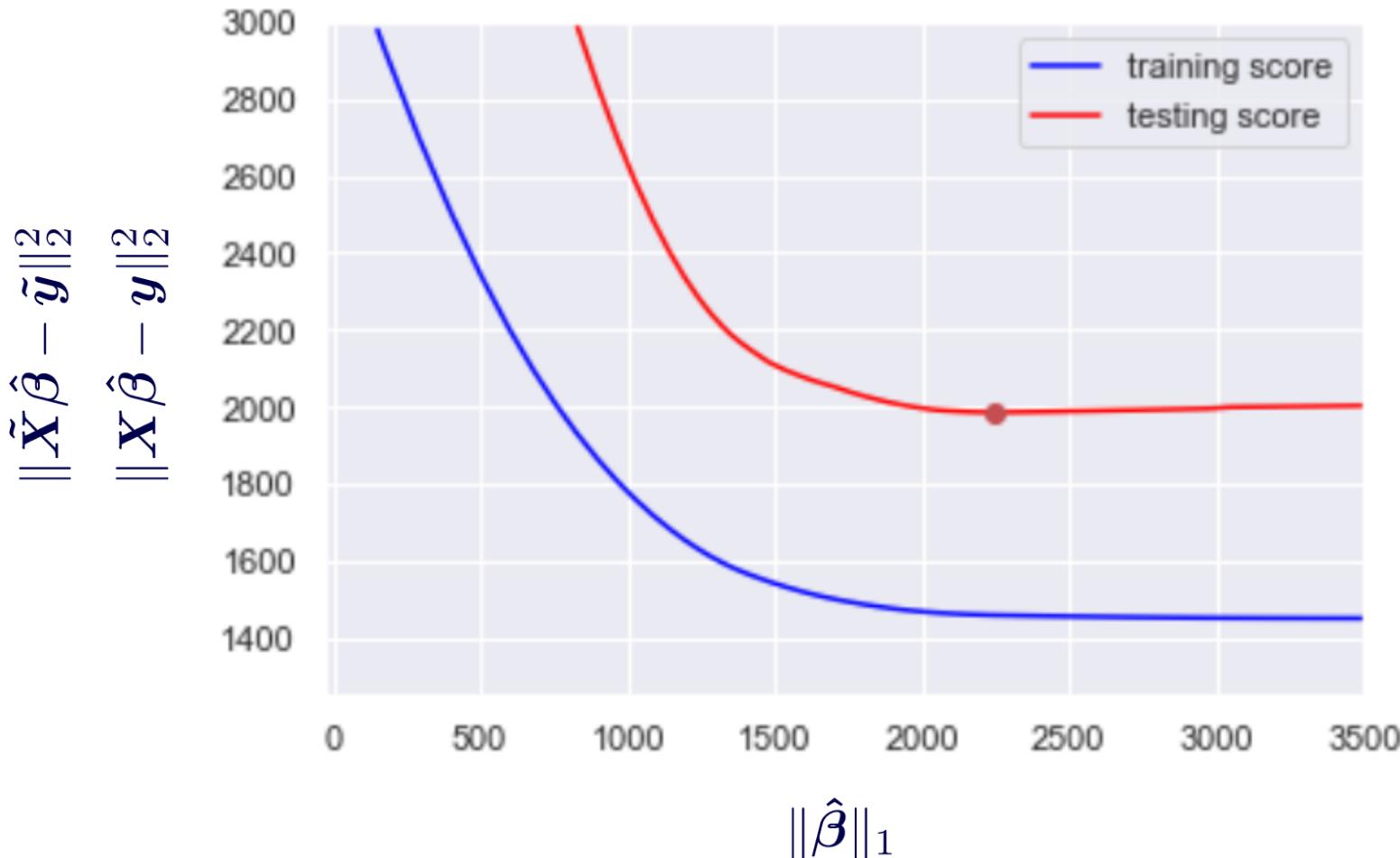


$$\left\{ \begin{array}{l} |\beta_i| = \bar{\beta}_i + \underline{\beta}_i \\ \beta_i = \bar{\beta}_i - \underline{\beta}_i \\ \bar{\beta}_i \geq 0, \underline{\beta}_i \geq 0 \end{array} \right.$$

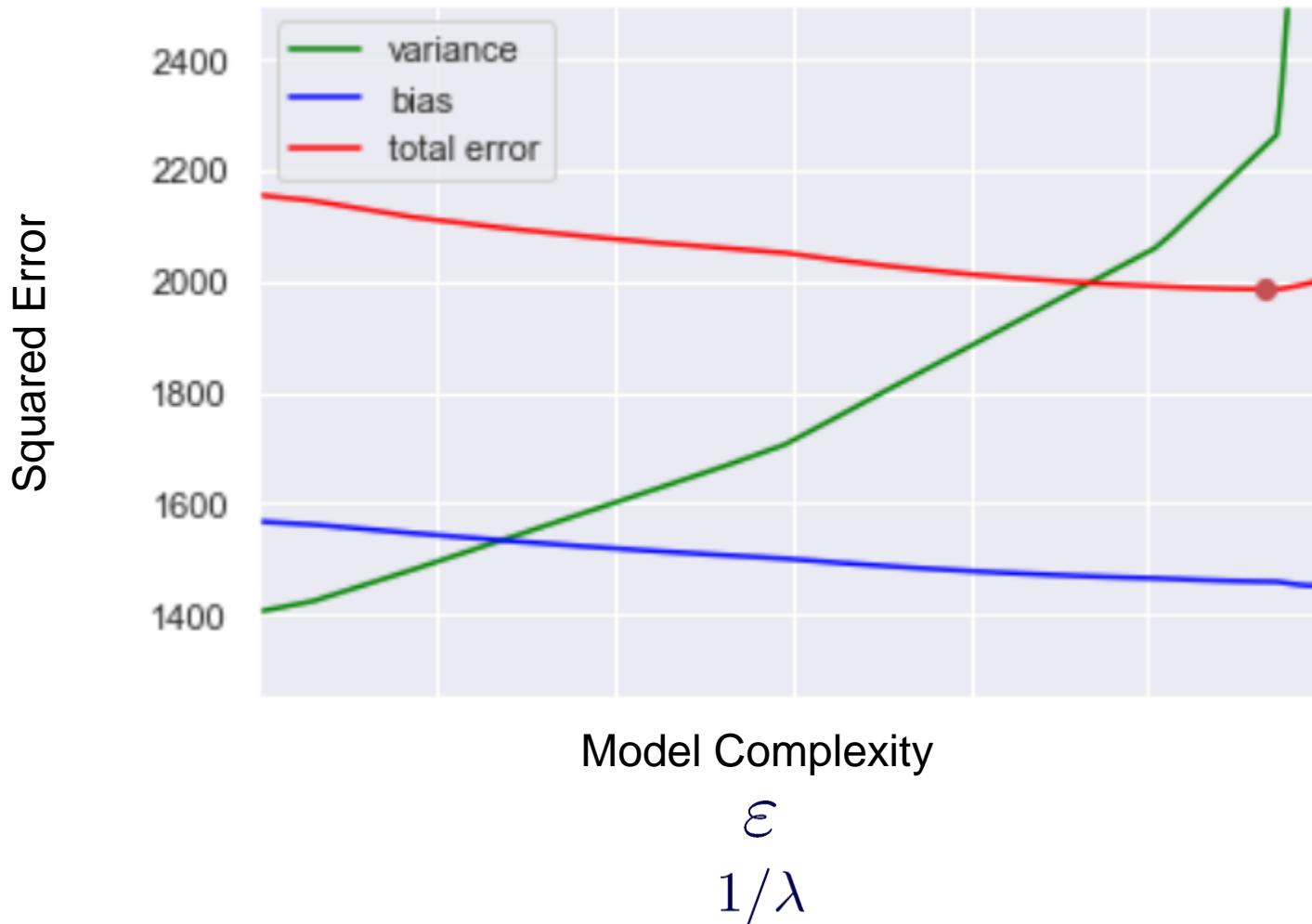
# Regularized regression – bias-variance tradeoff



## Regularized regression – bias-variance tradeoff



## Regularized regression – bias-variance tradeoff



---

## Regularized regression – hyperparameter tuning in `sklearn`

- $L_1$  regularized linear regression:

$$\min_{\beta} \frac{1}{m} \|X\beta - y\|_2^2 + \alpha \|\beta\|_1$$

- $L_1$  regularized logistic regression:

$$\min_{\theta} \|\theta\|_1 + C \cdot J(\theta)$$

hyperparameter

---

## Evaluation method

- **Holdout method**

- Given data is randomly partitioned into two independent sets
  - Training set (e.g., 2/3) for model construction
  - Test set (e.g., 1/3) for accuracy estimation
- Random sampling: a variation of holdout
  - Repeat holdout  $k$  times, accuracy = avg. of the accuracies obtained

- **Cross-validation** ( $k$ -fold, where  $k = 10$  is most popular)

- Randomly partition the data into  $k$  *mutually exclusive* subsets, each approximately equal size
- At  $i$ -th iteration, use  $D_i$  as test set and others as training set
- Leave-one-out:  $k$  folds where  $k = \#$  of tuples, for small sized data
- Stratified cross-validation: folds are stratified so that class dist. in each fold is approx. the same as that in the initial data

---



# Linear Regression – Machine Learning Perspective

---

# Linear regression – data mining and machine learning perspective

## ■ Input (features):

- $n$  - number of features
- $\mathbf{x}^{(j)}$  - input (features) of  $j$ -th training example
- $x_i^{(j)}$  - value of feature  $i$  in  $j$ -th training example

$$\mathbf{x} = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

## ■ Hypothesis:

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x} = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

## ■ Parameters:

$$\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \dots, \theta_n)^T$$

## ■ Cost function:

$$J(\theta_0, \theta_1, \theta_2, \dots, \theta_n) = \frac{1}{2m} \sum_{j=1}^m \left( h_{\boldsymbol{\theta}}(\mathbf{x}^{(j)}) - y^{(j)} \right)^2$$

## ■ Optimization:

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

---

# Linear regression – data mining and machine learning perspective

## ■ Optimization:

$$\min_{\theta} J(\theta)$$

## ■ Solution algorithms:

- Non-linear optimization methods, e.g., iterative algorithms such as gradient descent algorithms, Newton and Quasi-Newton algorithms, etc.
- Linear algebra (normal equations), i.e., solving  $\nabla J(\theta) = 0$

## ■ Solving normal equations:

$$\frac{\partial J(\theta)}{\partial \theta_i} = \frac{1}{m} \sum_{j=1}^m \left( h_{\theta}(x^{(j)}) - y^{(j)} \right) x_i^{(j)} = 0$$

$$\nabla J(\theta) = \begin{pmatrix} \frac{\partial J(\theta)}{\partial \theta_0} \\ \frac{\partial J(\theta)}{\partial \theta_1} \\ \frac{\partial J(\theta)}{\partial \theta_2} \\ \vdots \\ \frac{\partial J(\theta)}{\partial \theta_n} \end{pmatrix}$$

## ■ Solve system of linear equations for $\theta$ :

$$\sum_{j=1}^m \left( \theta^T x^{(j)} - y^{(j)} \right) x_i^{(j)} = 0$$

# Linear regression – data mining and machine learning perspective

- **Solving normal equations** ( $m$  examples,  $n$  features):

$$\sum_{j=1}^m \boldsymbol{\theta}^T \mathbf{x}^{(j)} \cdot x_i^{(j)} = \sum_{j=1}^m y^{(j)} \cdot x_i^{(j)} \quad i = 1, \dots, n$$

- **Notation:**

$$\mathbf{X} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & \cdots & x_n^{(2)} \\ 1 & x_1^{(3)} & x_2^{(3)} & x_3^{(3)} & \cdots & x_n^{(3)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & x_3^{(m)} & \cdots & x_n^{(m)} \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \\ \vdots \\ y^{(m)} \end{pmatrix}$$

- **Solution:**

$$\boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

- **Solution algorithm** (solve system of linear equations with unknown  $\boldsymbol{\theta}$ ):

$$(\mathbf{X}^T \mathbf{X}) \cdot \boldsymbol{\theta} = \mathbf{X}^T \mathbf{y}$$

---



# **Logistic Regression – Machine Learning Perspective**

---

# Other types of regression analysis

## Quantile regression

- Ordinary least squares regression approximates the conditional mean of the response variable, while quantile regression is estimating either the conditional median or other quantiles of the response variable
- This is very helpful in case of skewed data (e.g., income distribution in the US) or to deal with data without suppressing outliers

## Logistic regression

- Logistic regression is used to predict categorical target variable
- Most often a variable with a binary outcome
  - Logit and Probit regressions can also be used to predict binary outcome. While the underlying distributions are different, all three models will produce rather similar outcomes
- It is frequently used to estimate the probability of an event
  - Bank customer defaulting on the loan
  - Customer responding to a marketing promotion
  - Spam or not-spam email
  - Malignant or benign tumor

# Classification

- **Dataset:**  $D = \{(y^{(1)}, \mathbf{x}^{(1)}), \dots, (y^{(m)}, \mathbf{x}^{(m)})\}$

- **Targets:**  $y$



*Iris setosa (A)*

- **Features:**  $\mathbf{x}$

- **Parameters:**  $\theta$

- **Classifier:**  $h(\mathbf{x}|\theta) = h_{\theta}(\mathbf{x})$



*Iris versicolor (B)*

- **Predictions:**  $\hat{y} = h(\mathbf{x}|\theta)$



$y = 1$

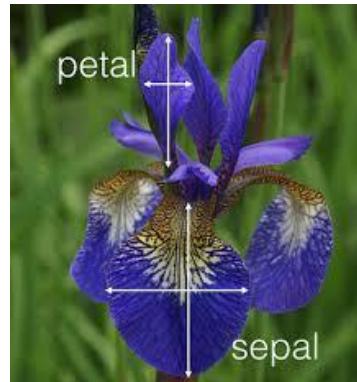
- In **classification** we are trying to **predict discrete targets**

- In the **two-class** problem  $y \in \{0, 1\}$  means that  $y$  can be equal **0** (“negative class”, e.g., spam) or **1** (“positive class”, e.g., not spam)

- Example classification problem – **classify different flowers using measurements of the flower**

# Classification

- Features are numerical attributes
- Good features can be used to predict targets (different classes)



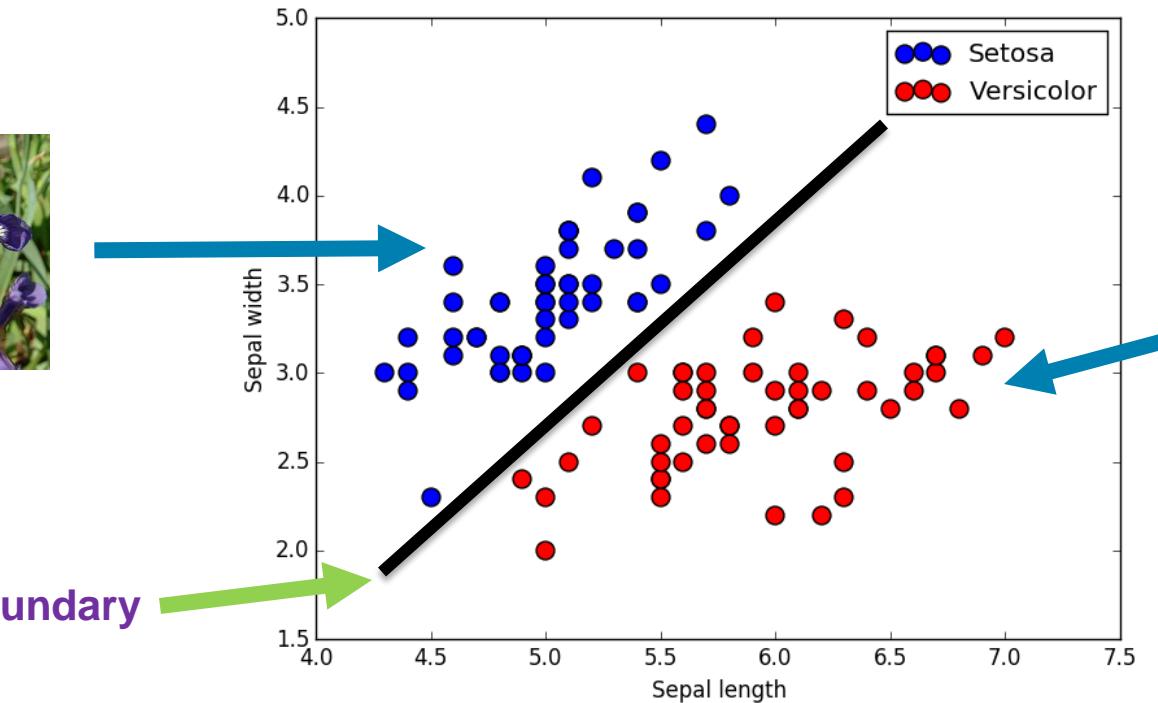
Fisher's Iris Data

Sepal length	Sepal width	Petal length	Petal width	Species
$x_1$	$x_2$	$x_3$	$x_4$	$y$

- Sometimes we can separate different classes with a linear decision boundary



A



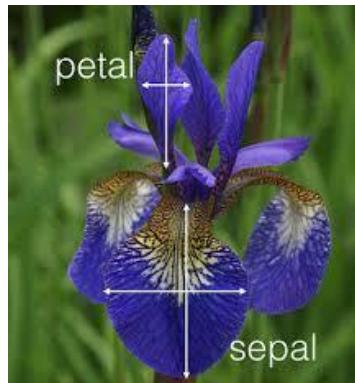
decision boundary



B

# Classification

- Features are numerical attributes
- Good features can be used to predict targets (different classes)



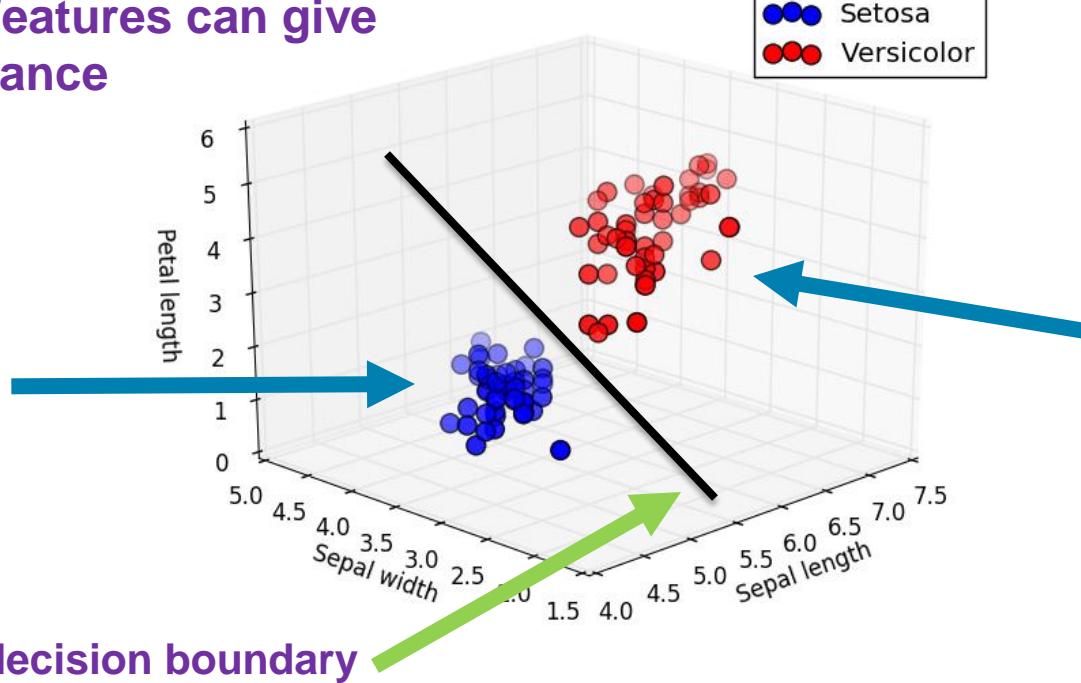
Fisher's Iris Data

Sepal length	Sepal width	Petal length	Petal width	Species
$x_1$	$x_2$	$x_3$	$x_4$	$y$

- Sometimes we can separate different classes with a linear decision boundary
- Usually more features can give better performance



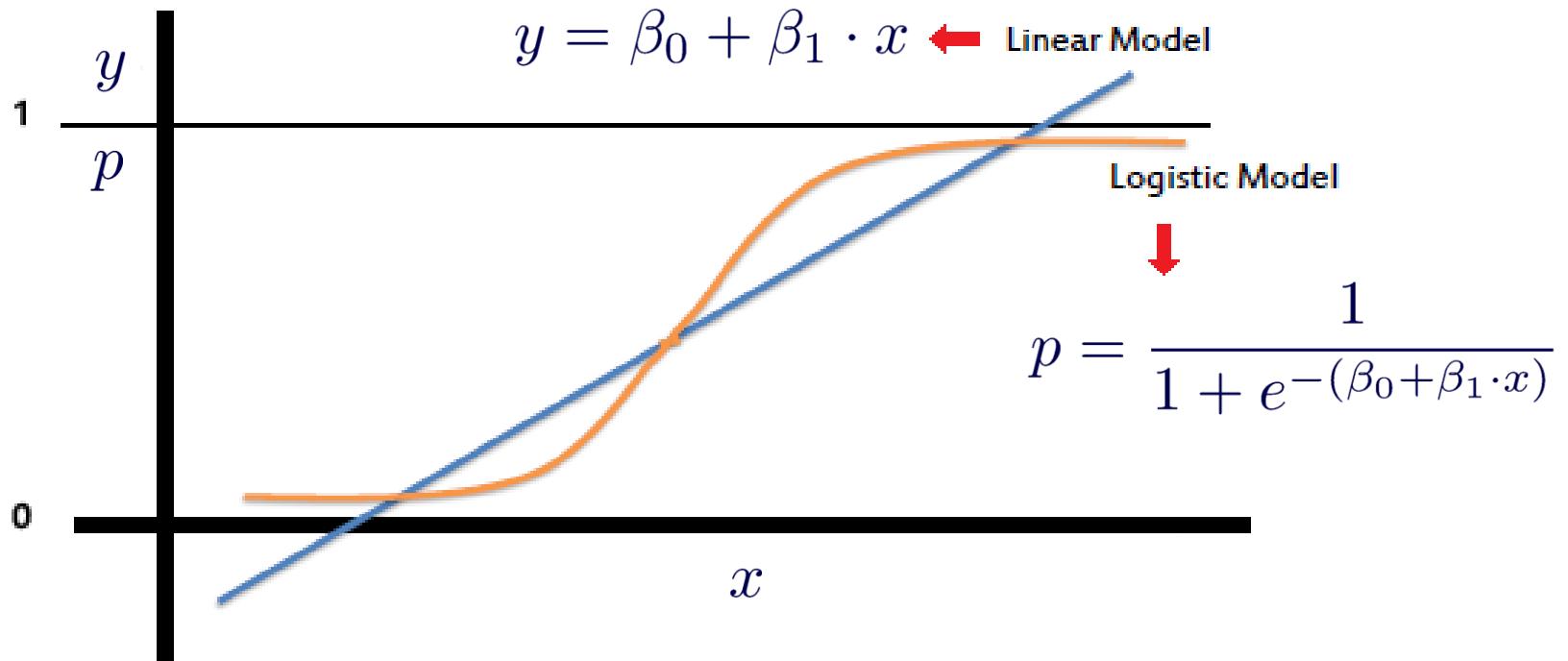
A



B

# Linear regression vs. logistic regression

- **Linear regression** predicts **real values**
- **Logistic regression** predicts **values in the range of 0 to 1**

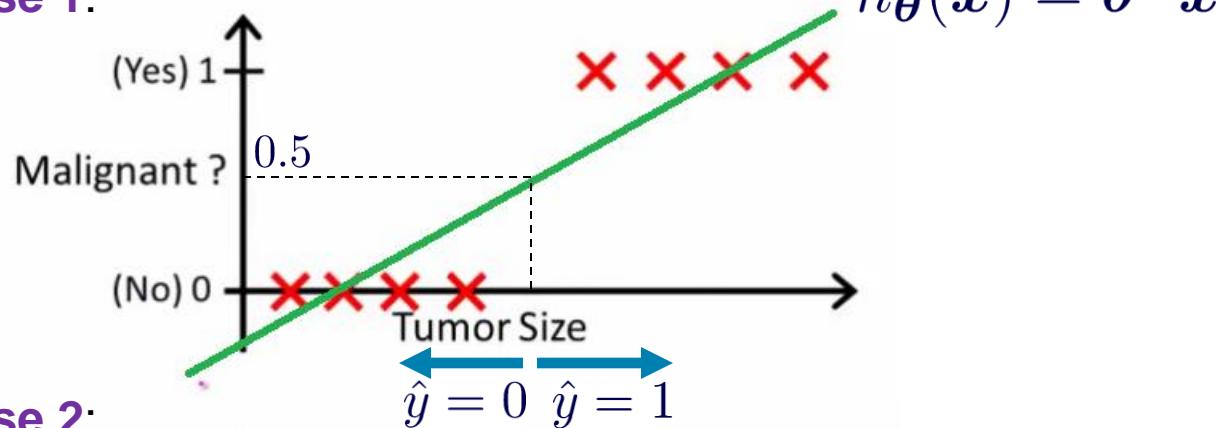


# Using linear regression for classification

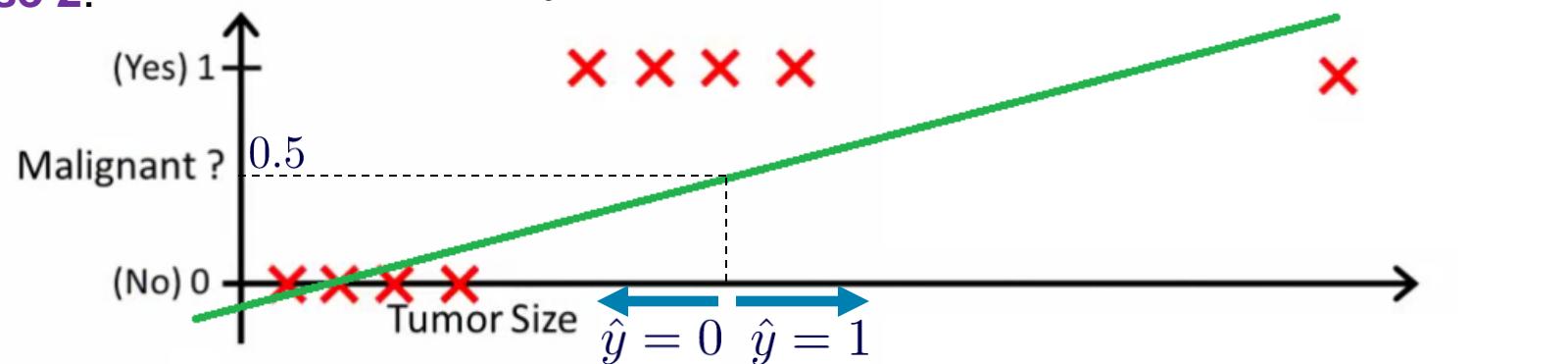
- **Threshold** classifier output  $h_{\theta}(x)$  at 0.5:

- If  $h_{\theta}(x) \geq 0.5$  predict “ $y = 1$ ”
  - If  $h_{\theta}(x) < 0.5$  predict “ $y = 0$ ”

- **Case 1:**



- **Case 2:**



- What to do if  $h_{\theta}(x) < 0$  or  $> 1$ ? **We want**  $0 \leq h_{\theta}(x) \leq 1$

# Logistic regression – hypothesis representation

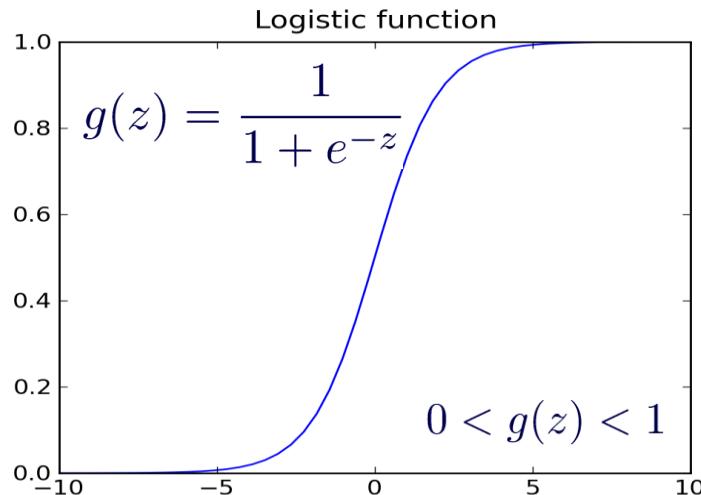
- We want  $0 \leq h_{\theta}(x) \leq 1$ :

- hypothesis for linear regression  $h_{\theta}(x) = \theta^T x$

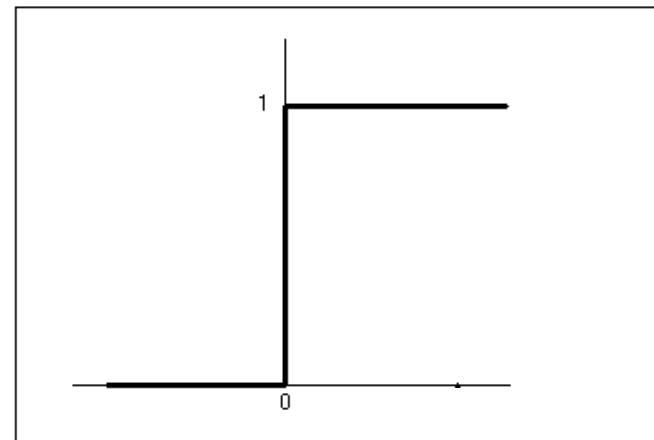
- hypothesis for logistic regression  $h_{\theta}(x) = g(\theta^T x)$

- Sigmoid function / logistic function  $g(z) = \frac{1}{1 + e^{-z}}$

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$



use logistic function to approximate threshold function



one important reason is that **logistic function is differentiable**

- We need to fit parameters  $\theta$

# Logistic regression – hypothesis representation

- $h_{\theta}(x) = \text{estimated probability that } y = 1 \text{ on input } x$ 
  - if  $x = \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} = \begin{pmatrix} 1 \\ \text{tumorSize} \end{pmatrix}$  and  $h_{\theta}(x) = 0.7$
  - tell a patient that 70% chance of tumor being malignant
- Probability that  $y = 1$ , given  $x$ , parameterized by  $\theta$ 
$$h_{\theta}(x) = g(\theta^T x) = \text{prob}(y = 1 | x; \theta) \quad \text{prob}(y = 0 | x; \theta) + \text{prob}(y = 1 | x; \theta) = 1$$
- Suppose:
  - predict " $y = 1$ " if  $h_{\theta}(x) \geq 0.5$   
→  $\theta^T x \geq 0$
  - predict " $y = 0$ " if  $h_{\theta}(x) < 0.5$   
→  $\theta^T x < 0$

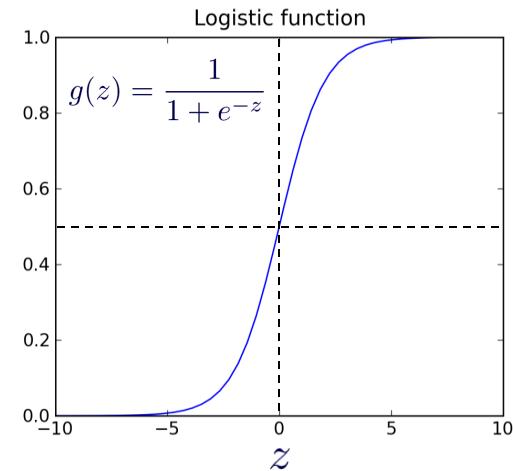
$$g(z) \geq 0.5 \text{ when } z \geq 0$$

$$h_{\theta}(x) = g(\theta^T x) \geq 0.5 \text{ when } \theta^T x \geq 0$$

$$g(z) < 0.5 \text{ when } z < 0$$

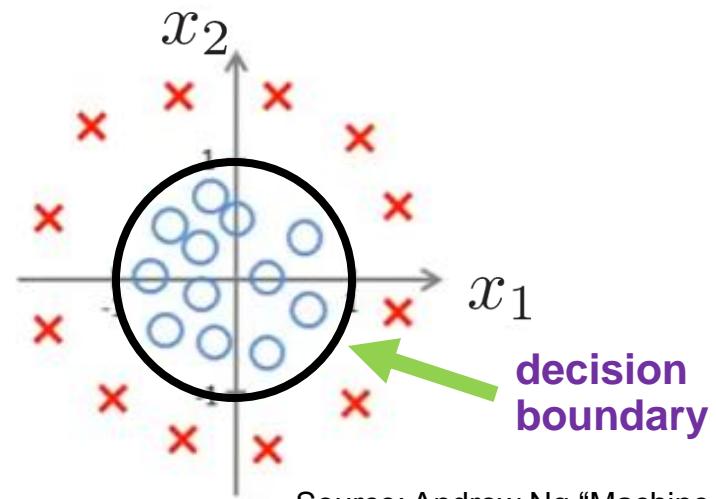
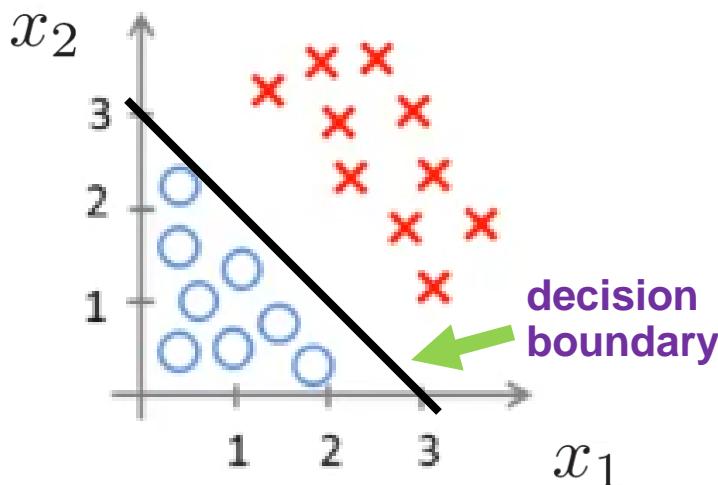
$$h_{\theta}(x) = g(\theta^T x) < 0.5 \text{ when } \theta^T x < 0$$

Source: Andrew Ng "Machine Learning"



## Logistic regression – decision boundary

- $h_{\theta}(x) = g(\theta^T x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$  fit  $\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} -3 \\ 1 \\ 1 \end{pmatrix}$ 
  - predict " $y = 1$ " if  $\theta^T x = -3 + x_1 + x_2 \geq 0$
  - predict " $y = 0$ " if  $\theta^T x = -3 + x_1 + x_2 < 0$
  
- $h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$  fit  $\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$ 
  - predict " $y = 1$ " if  $-1 + x_1^2 + x_2^2 \geq 0$
  - predict " $y = 0$ " if  $-1 + x_1^2 + x_2^2 < 0$



Source: Andrew Ng "Machine Learning"

## Logistic regression – cost function

- **Input** (training set):  $D = \{(y^{(1)}, \mathbf{x}^{(1)}), \dots, (y^{(m)}, \mathbf{x}^{(m)})\}$

- $n$  - number of features
  - $m$  - number of examples
  - $\mathbf{x}^{(j)}$  - input (features) of  $j$ -th training example
  - $x_i^{(j)}$  - value of feature  $i$  in  $j$ -th training example

$$y \in \{0, 1\} \quad \mathbf{x} = \begin{pmatrix} x_0 = 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

- **Hypothesis:**  $h_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}}}$

- **Parameters:**  $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \dots, \theta_n)^T$

- **Cost function:**  $J(\boldsymbol{\theta}) = \frac{1}{m} \sum_{j=1}^m \underbrace{\text{cost}\left(h_{\boldsymbol{\theta}}(\mathbf{x}^{(j)}), y^{(j)}\right)}_{\frac{1}{2} \left(h_{\boldsymbol{\theta}}(\mathbf{x}^{(j)}) - y^{(j)}\right)^2}$
- **Optimization:**  $\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$

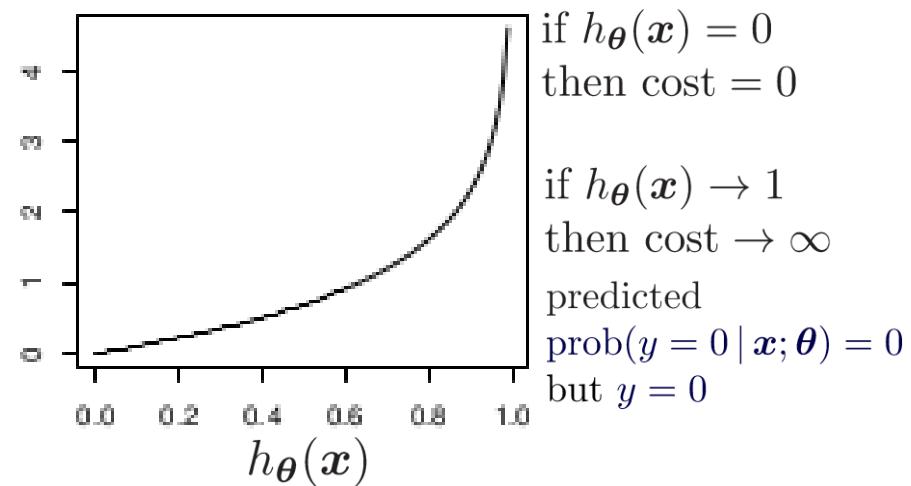
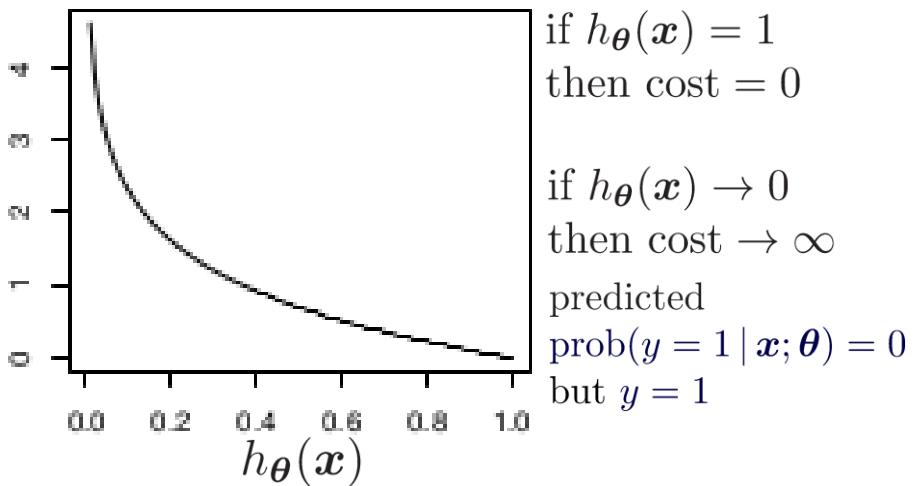
cost function that we used for linear regression is now non-convex

# Logistic regression – cost function

## ■ Cost function:

$$\text{cost}(h_{\theta}(\mathbf{x}), y) = \begin{cases} -\log(h_{\theta}(\mathbf{x})) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(\mathbf{x})) & \text{if } y = 0 \end{cases}$$

if  $y = 1$     if  $y = 0$



## ■ More compact form of cost function:

$$\text{cost}(h_{\theta}(\mathbf{x}), y) = -y \log(h_{\theta}(\mathbf{x})) - (1 - y) \log(1 - h_{\theta}(\mathbf{x}))$$

if  $y = 1$ :  $\text{cost}(h_{\theta}(\mathbf{x}), y) = -\log(h_{\theta}(\mathbf{x}))$

if  $y = 0$ :  $\text{cost}(h_{\theta}(\mathbf{x}), y) = -\log(1 - h_{\theta}(\mathbf{x}))$

# Logistic regression – optimization problem

## ■ Cost function:

$$J(\boldsymbol{\theta}) = \frac{1}{m} \sum_{j=1}^m \text{cost}\left(h_{\boldsymbol{\theta}}(\mathbf{x}^{(j)}), y^{(j)}\right)$$

$$= -\frac{1}{m} \sum_{j=1}^m \left[ y^{(j)} \log(h_{\boldsymbol{\theta}}(\mathbf{x}^{(j)})) + (1 - y^{(j)}) \log(1 - h_{\boldsymbol{\theta}}(\mathbf{x}^{(j)})) \right]$$

in statistics this **cost function** can be derived using principle of **Maximum Likelihood estimation**



## ■ Optimization to fit parameters $\boldsymbol{\theta}$ : $\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$

## ■ Solution algorithms:

- Non-linear optimization methods, e.g., iterative algorithms such as gradient descent algorithms, Newton and Quasi-Newton algorithms

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \theta_i} = \frac{1}{m} \sum_{j=1}^m \left( h_{\boldsymbol{\theta}}(\mathbf{x}^{(j)}) - y^{(j)} \right) x_i^{(j)}$$

$$\nabla J(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial J(\boldsymbol{\theta})}{\partial \theta_0} \\ \frac{\partial J(\boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial J(\boldsymbol{\theta})}{\partial \theta_2} \\ \vdots \\ \frac{\partial J(\boldsymbol{\theta})}{\partial \theta_n} \end{pmatrix}$$

- To make **predictions** given new  $\mathbf{x}$  we **output**  $h_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}}}$

# Logistic regression – multiclass classification

## ■ One-vs-all

- Train a logistic regression classifier  $h_{\theta}^{(k)}(\mathbf{x})$  for each class  $k$  to predict the probability that  $y = k$
- On a new input  $\mathbf{x}$ , to make a prediction, pick the class  $k$  that maximizes

$$\max_k h_{\theta}^{(k)}(\mathbf{x})$$

## ■ Email tagging – Work, Friends, Family, Hobby

## ■ Classify different flowers using measurements of the flower



*Iris setosa* (A)



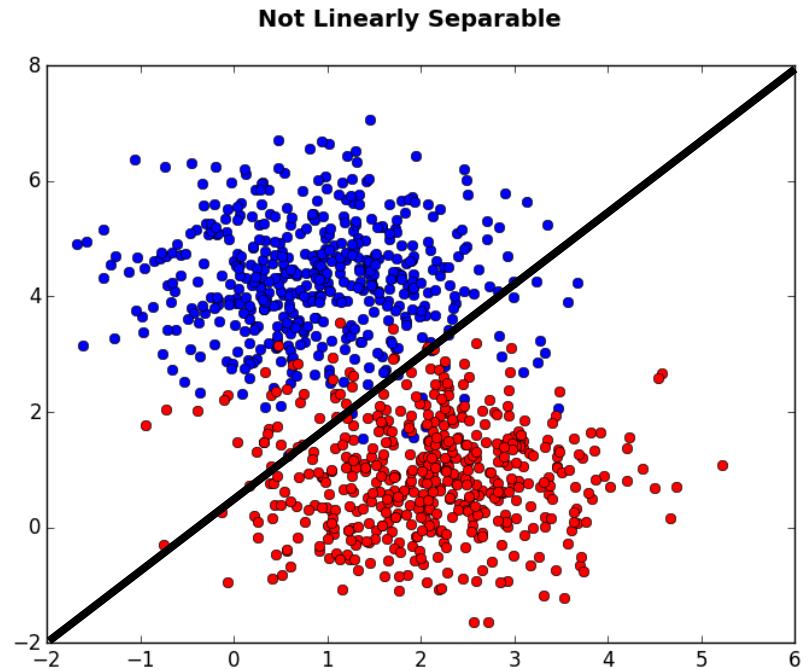
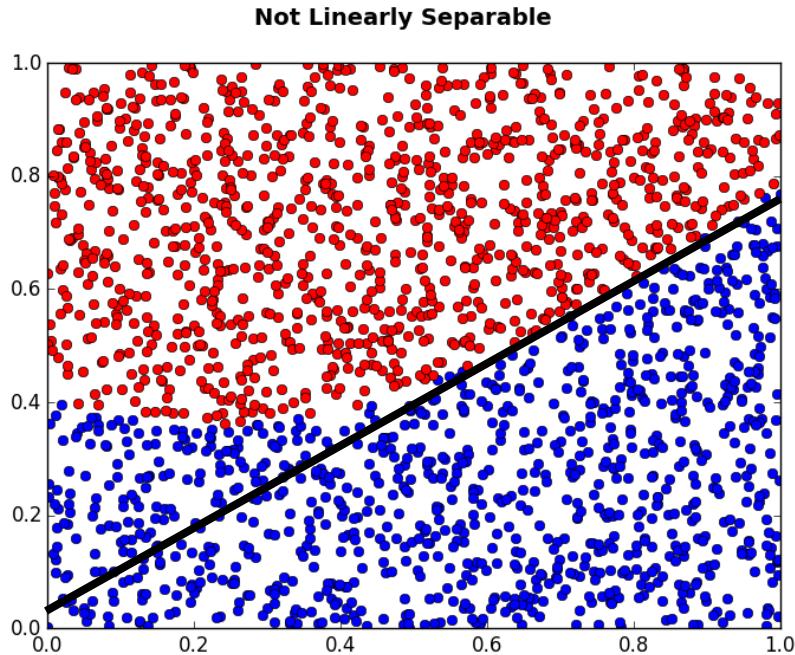
*Iris versicolor* (B)



*Iris virginica* (C)

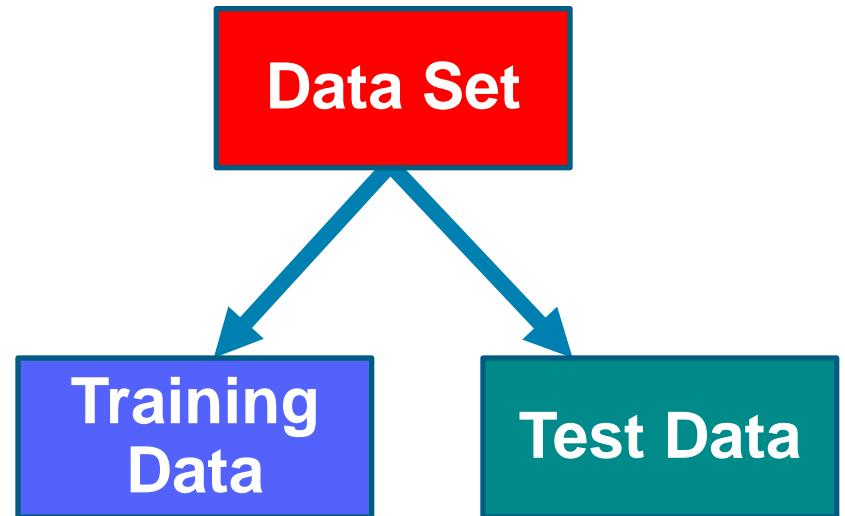
# Classification – non linearly separable classes

- Sometimes we can **not separate** data **linearly**



## Training and testing data

- First train the system using one part of the data set
- Then test the system using another part of the data set



**Data set**  
 $\{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), (y_3, \mathbf{x}_3)\}$

**Training Data:**  $\{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2)\}$

**Test Data:**  $(y_3, \mathbf{x}_3)$

# Overfitting

- In many models it can be shown that:

$$\text{Test Error} > \text{Training Error}$$

- A model generalizes well if test error is similar to training error:

$$\text{Test Error} \approx \text{Training Error}$$