Data-driven Decision Making

Introduction

João Pedro Pedroso

2024/2025

Info

- Check page in SIGARRA

Introduction

Data: powerful and ubiquitous

- How much data has been created so far?
  - according to International Data Corporation: 18 zettabytes in 2018
    `https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf`
  - 180000000000000000000000 bytes
  - data created, captured or replicated
    - vast majority, in the last few years
    - no sign of slowing down
    - at current internet speed: millions of years to download
  - by 2025: 175 zettabytes
- Produced at ever increasing rates
  - Decoding the human genome:
    - originally took more than 10 years
    - now, probably can be done in a few hours
      `https://www.sandiegouniontribune.com/business/biotech/sd-me-illumina-novaseq-20170109-story.html`

Data and Analytics are Useful

- There is a shortage of people with deep analytical skills, virtually in every country
- Critical in almost every industry
- Companies invest more and more on analytics
- Sectors: healthcare, media, sports, finance, government, ...

What is Analytics?

The science of using data to build models that lead to better decisions that add value to individuals, to companies, to institutions.

- ultimately: create value
- can use big data or "small" data

Importance:

- Organizations are becoming much more data driven
  - the number of Chief Data Officers doubled from 2012 to 2014
  - survey of 1300 companies worldwide found that over half use data analytics in everyday decisions
- MIT study on 330 public North American companies:
  - Companies in top third of industry in data-driven decision making were 5% more productive and 6% more profitable than competitors

What is Analytics?

- Descriptive: find patterns in the data
  - summary statistics
  - visualizations
  - clustering
- Predictive: predict different outcomes
  - linear regression
  - logistic regression, CART, random forests
- Prescriptive: gives advice on actions to take
  - optimization

Key Messages

- Analytics provide a competitive edge to individuals and companies
- Analytics are often critical to the success of a company
- Methodology for this course:
  - teach analytics techniques through real world examples and real data

Our Objective

- Convince you of the Analytics Edge
  - the power and importance of data
  - how analytics methods work
  - how to interpret and understand the results of analytical models
- Inspire you to use analytics in your career
  - software development in Python
  - not just hearing about analytics, but creating your own models

Some examples

eHarmony

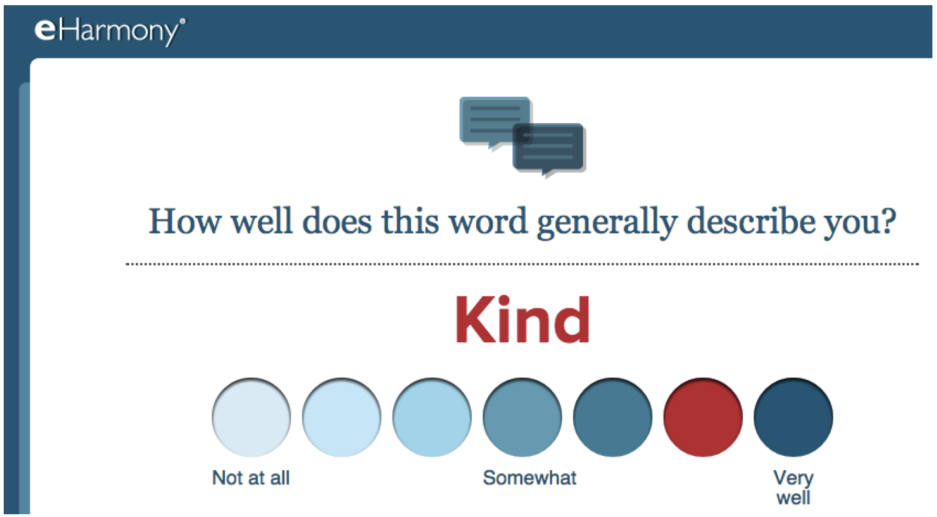
- Online dating site focused on long term relationships
- Takes a scientific approach to love and marriage
- Nearly 4% of US marriages in 2012 are a result of eHarmony
- Has generated over $1 billion in cumulative revenue

Finding Successful Matches

- First predict if users will be compatible
  - Use 29 different "dimensions of personality"
- Then need to find matches for everyone
  - Members in more than 150 countries
  - Since launching in 2000, more than 33 million members
- They use regression and optimization
  - Operates eHarmonyLabs, a relationship research facility

The Data

- Collect data through 436 questions
- About 15,000 people take the questionnaire each day



What is the Edge?

- Relies much more on data than other dating sites
- Suggests a limited number of high quality matches
  - Users don't have to search and dig through profiles
- eHarmony has successfully leveraged the power of analytics to create a successful and thriving business
  - 14% of US online dating market (2012)

KEP project

The Framingham Heart Study

- One of the most important studies of modern medicine
- Ongoing study of the residents in Framingham, MA
  - Started in 1948, now on the third generation
- Much of the now-common knowledge regarding heart disease came from this study
  - High blood pressure should be treated
  - Clogged arteries are not normal
  - Cigarette smoking can lead to heart disease

Heart Disease

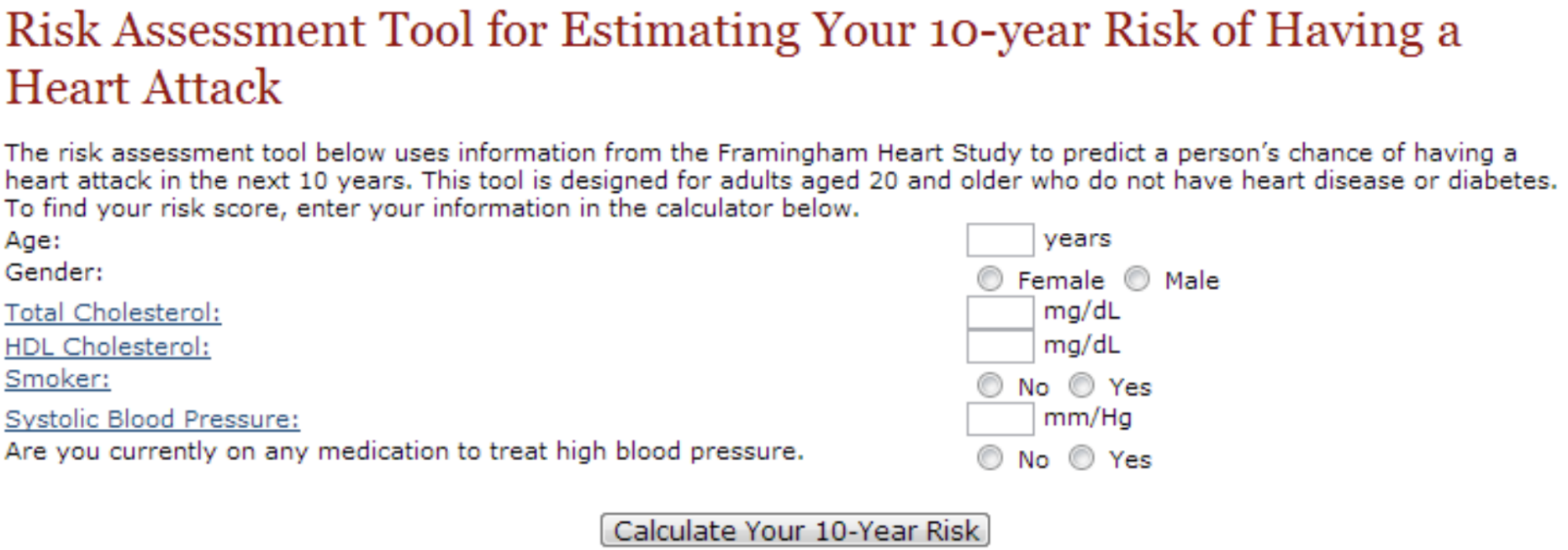
- Heart disease has been the leading cause of death worldwide since 1921
  - 7.3 million people died from CHD in 2008
- Since 1950, age-adjusted death rates have declined 60%
  - In part due to the results of the Framingham Heart Study

The Data

- 5,209 patients were enrolled in 1948
- Given a questionnaire and exam every two years
  - Physical characteristics
  - Behavioral characteristics
  - Test results
- Patient population, exam, and questions expanded over time

An Analytics Approach

- Used regression to predict whether or not a patient would develop heart disease in the next ten years
- Model tested and adjusted for different populations
- Available online



Risk Assessment Tool for Estimating Your 10-year Risk of Having a Heart Attack

The risk assessment tool below uses information from the Framingham Heart Study to predict a person's chance of having a heart attack in the next 10 years. This tool is designed for adults aged 20 and older who do not have heart disease or diabetes. To find your risk score, enter your information in the calculator below.

Age: years

Gender: Female Male

Total Cholesterol: mg/dL

HDL Cholesterol: mg/dL

Smoker: No Yes

Systolic Blood Pressure: mm/Hg

Are you currently on any medication to treat high blood pressure. No Yes

Calculate Your 10-Year Risk

What is the Edge?

- Provided necessary evidence for the development of drugs to lower blood pressure
- Paved the way for other clinical prediction rules
  - Predict clinical outcomes using patient data
- A model allows medical professionals to make predictions for patients worldwide

D2Hawkeye

- Medical software company founded in 2001
- Combined data with analytics to improve quality and cost management in healthcare
- Difficult for humans to sift through patient records
- In 2009, the company was analyzing 20 million people monthly

The Data

- Healthcare industry is data-rich, but data may be hard to access
  - often unstructured and unavailable
- Used insurance data regarding procedures, prescriptions, and diagnoses
- Doctor insight regarding risk factors
  - interactions between illnesses
- Demographic information (gender and age)

The Analytics

- Predict future health care costs
- Identify high-risk patients to be prioritized for intervention
- Created interpretable models for doctors to analyze and verify
- Significantly improved over just using historical costs

What is the Edge?

- Substantial improvement in D2Hawkeye's ability to identify patients who need more attention
- Use expert knowledge to identify new variables and refine existing variables
- Can make predictions for millions of patients without manually reading patient files

DDDM

- We'll cover these examples and many more
- Follow *The Analytics Edge* (not that closely)
- Each week will be composed of:
  - A lecture
  - A practical part in the lab
  - Quizzes in every class
- Homework assignments:
  - several projects, may be made in a group
  - individual submission in the following lab class

Goal

- Make you comfortable using analytics in your career and in your life
- Know how to work with real data, besides learning different methodologies
- Convince you of the *analytics edge*

Mathematical Optimization

Mathematical Optimization

"Nothing in the world takes place without optimization, and there is no doubt that all aspects of the world that have a rational basis can be explained by optimization methods."
Leonhard Euler, 1744 (translation found in "Optimization Stories", edited by Martin Grötschel).

Mathematical Optimization

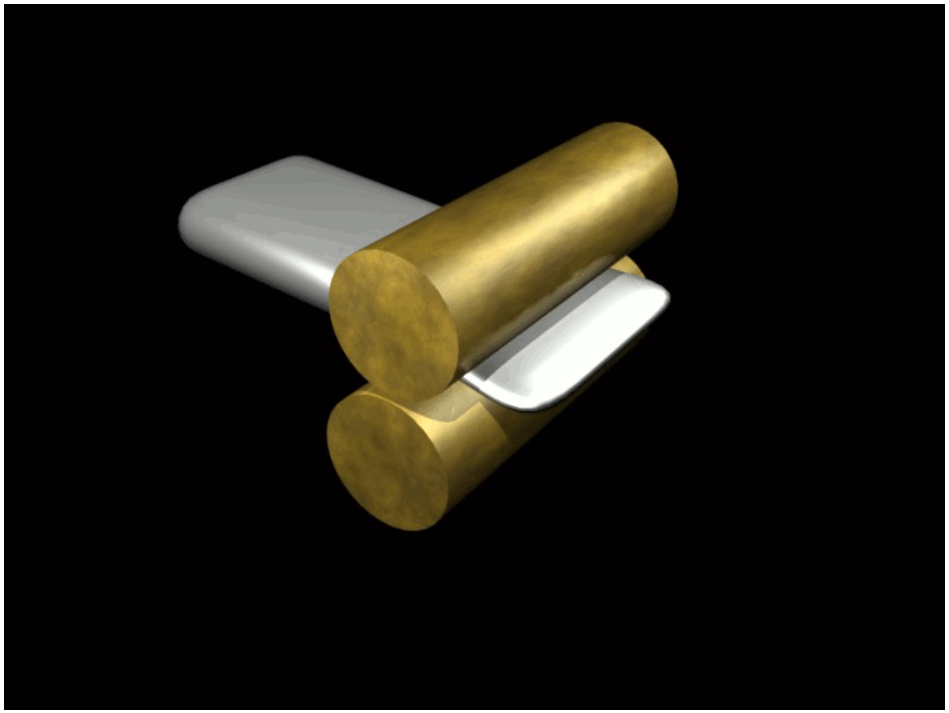
Mathematical optimization: science of finding the "best" solution

- Given an objective function, find a solution which is at least as good as any other possible solution
- Course of action:
  - describe the problem in terms of mathematical expressions
  - use some methodology to obtain an optimal solution from these formulas

Problem description

- No ambiguity allowed
- Use mathematical expressions
  - objective function (which we want to maximize of minimize);
  - conditions of the problem: constraint 1, constraint 2, ...

Example (first example from the AMPL book)



- A steel company uses a rolling mill to produce bands and coils
- It must decide how to allocate next week's mill time
- Data:

	Production (ton/h)	Profit ($/ton)	Capacity (ton/week)
Bands	200	25	6000
Coils	140	30	4000

- What is the optimum plan?

$$
\begin{aligned}
\text{maximize} \quad & 25x_B + 30x_C \\
\text{subject to} \quad & x_B/200 + x_C/140 \leq 40 \\
& 0 \leq x_B \leq 6000 \\
& 0 \leq x_C \leq 4000
\end{aligned}
$$

Try it: `https://ampl.com/try-ampl/try-ampl-online/`

```
var xb;
var xc;

maximize z: 25*xb + 30*xc;

subject to
hours: xb/200 + xc/140 <= 40;

capB: 0 <= xb <= 6000;
capC: 0 <= xc <= 4000;
```

How can we solve it?

Directly with ampl in the command line

```
# [having the previous program in file "steel.mod"]
ampl: option solver gurobi;
ampl: model "steel.mod"
ampl: solve;
Gurobi 9.1.1: optimal solution; objective 192000
1 simplex iterations
ampl: display z, xb, xc;
z = 192000
xb = 6000
xc = 1400
```

Using amplpy

- installation instructions in github page
  https://github.com/ampl/amplpy

Using Python + amplpy

In Python:

```
from amplpy import AMPL, Environment, DataFrame
ampl = AMPL()
ampl.option['solver'] = 'gurobi'
ampl.read("steel.mod")
ampl.solve()
z = ampl.obj['z']
xb = ampl.var['xb']
xc = ampl.var['xc']
print("Optimum:", z.value())
print("Solution: xb={}, xc={}".format(xb.value(), xc.value()))
```

Formalizing a mathematical optimization model

Steps: define

1. a set of variables $\rightarrow$
   unknowns to be found as a solution to the problem
2. a set of constraints $\rightarrow$
   equations or inequalities representing *requirements* as relationships between the variables
3. an objective function $\rightarrow$
   expression determining the target
   - *minimization* $\rightarrow$ smaller values are preferrable; *e.g.*, costs
   - *maximization* $\rightarrow$ larger values are better; *e.g.*, profits

Mathematical optimization: formulation

- Maximize or minimize
  - Objective function
- Subject to:
  - Constraint 1
  - Constraint 2
  - ...

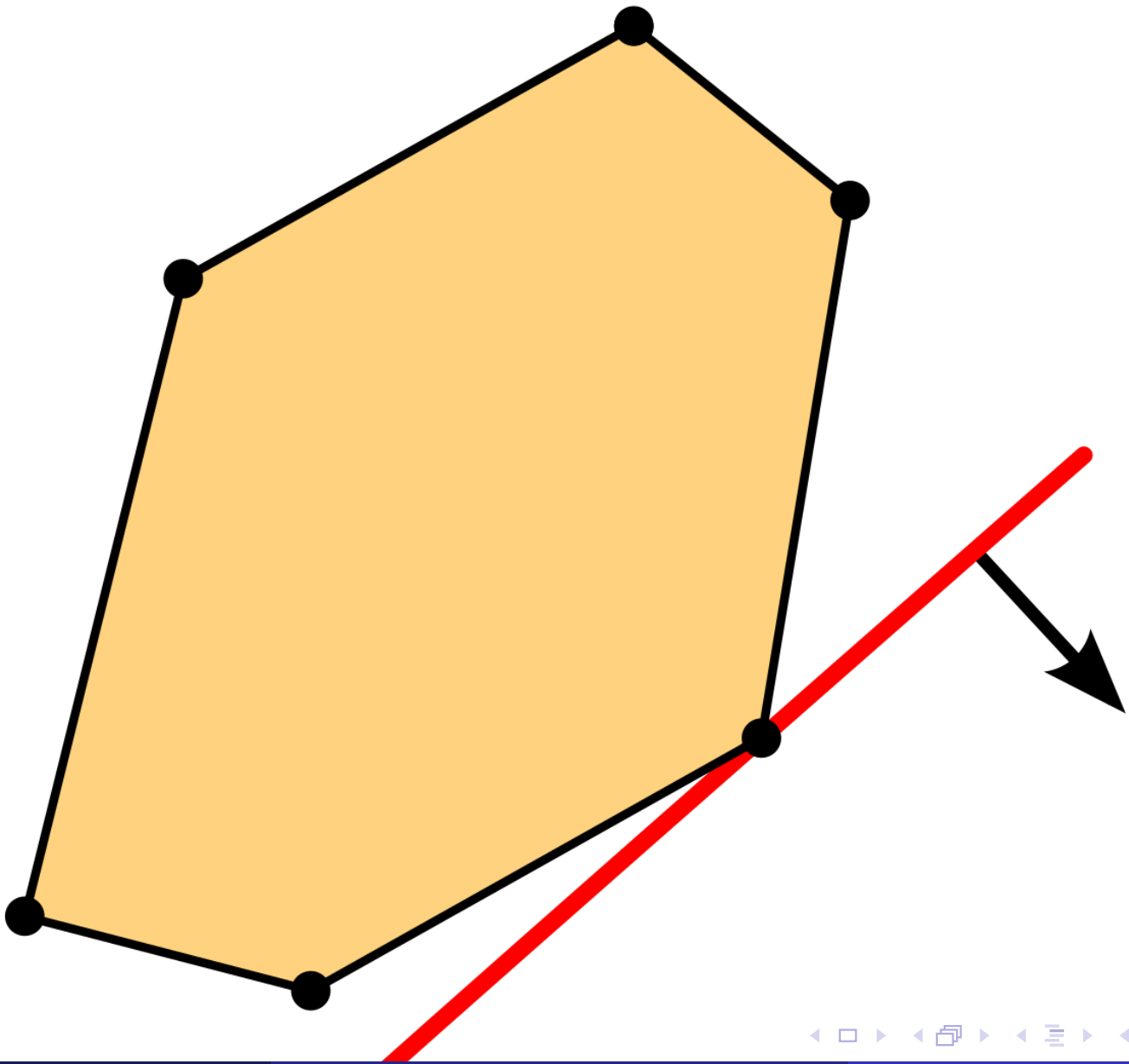
Terminology

- Optimization: seek a solution to either minimize or maximize the objective function, while satisfying all the constraints
- Maximization and minimization are essentially the same problem $\rightarrow$ convert one into the other: minus sign in the objective function
- Optimum/optimal solution $\rightarrow$ best possible from all candidate solutions, measured by the value of the objective function

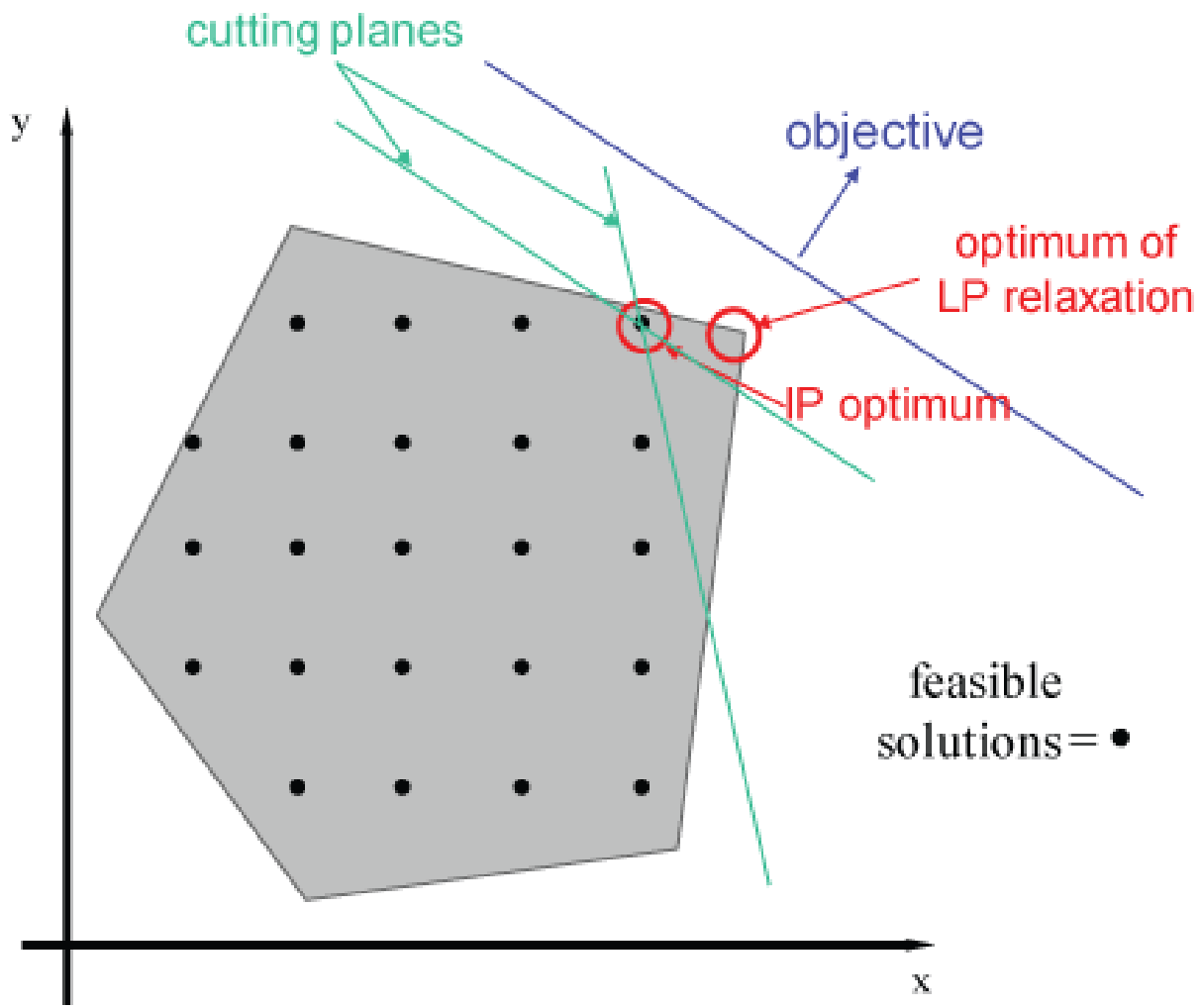
Kinds of mathematical optimization problems:

- linear optimization
  - objective function and constraints are all linear expressions
    - $a_1x_1 + a_2x_2 + \ldots + a_nx_n$
      where $x_1, x_2, \ldots, x_n$ are variables and $a_1, \ldots, a_n$ are constants
    - all variables may take on continuous (real) values
- nonlinear optimization
  - some expressions (objective and/or constraints) are not linear
- integer optimization
  - some variables must be integers
- multiobjective optimization
  - there is more than one objective to optimize simultaneously
- lexicographic goal programming
  - more than one objective, to optimize in a given order

Linear optimization

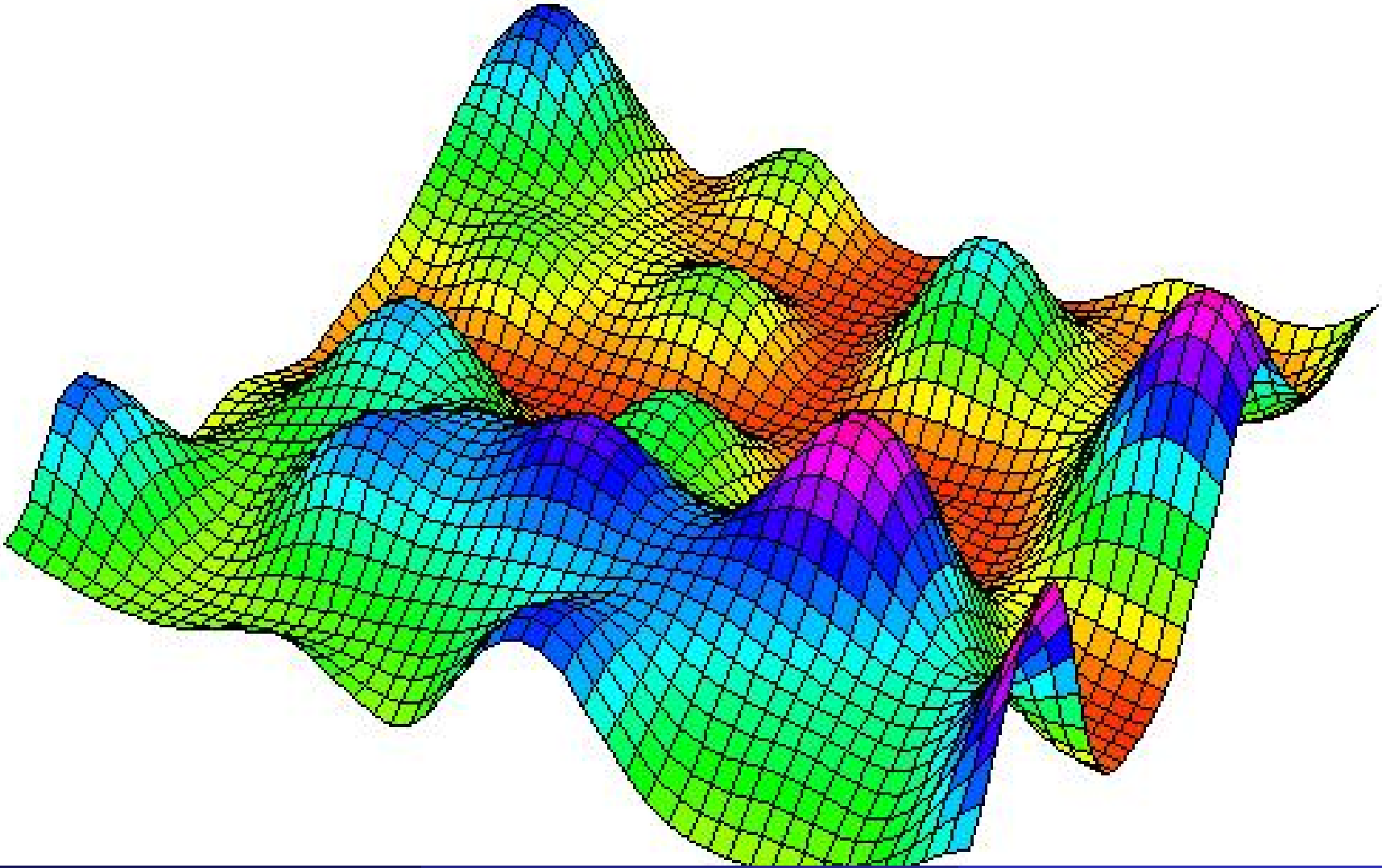


Integer optimization

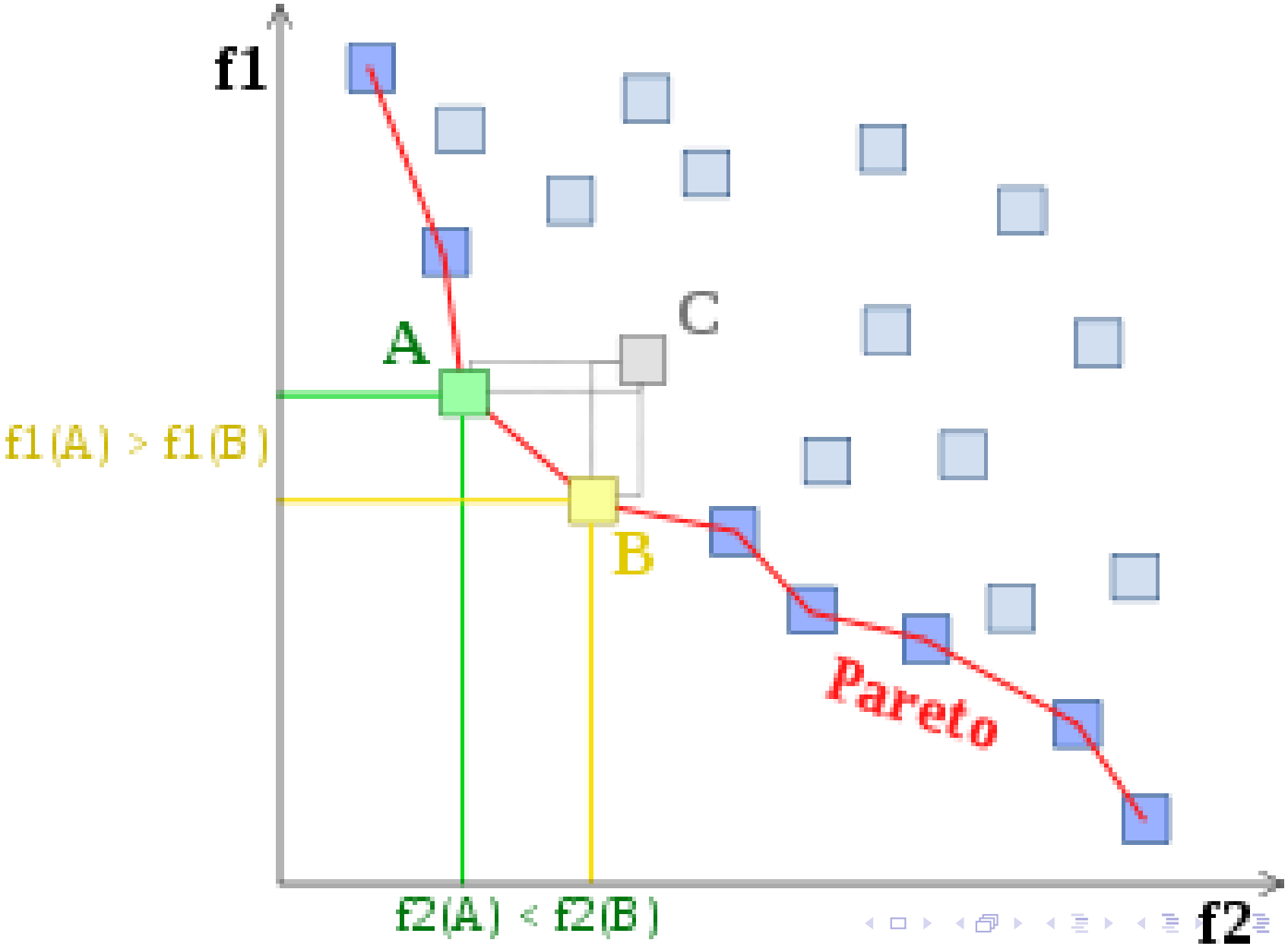


Cutting Planes

Nonlinear optimization



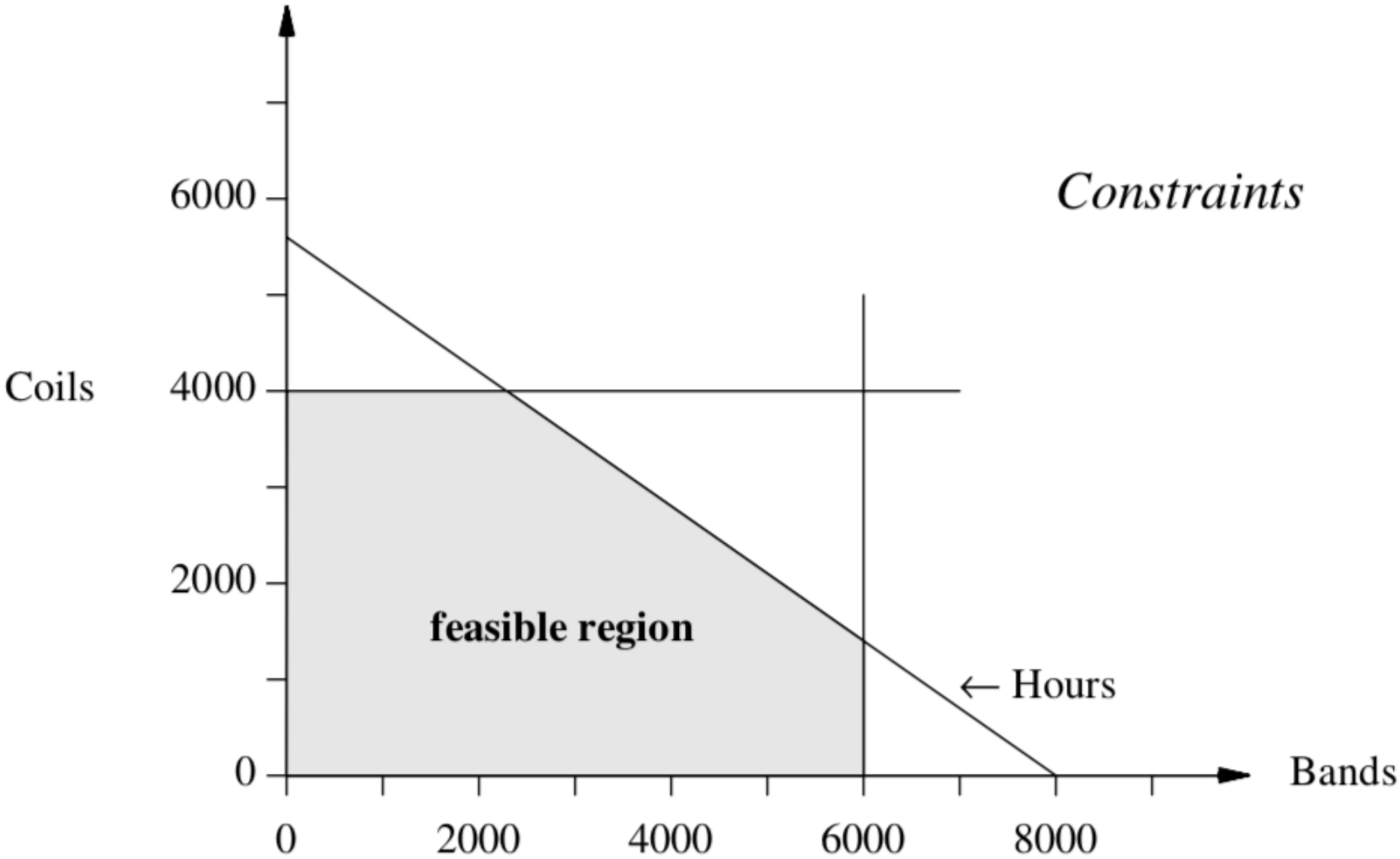
Multiobjective optimization



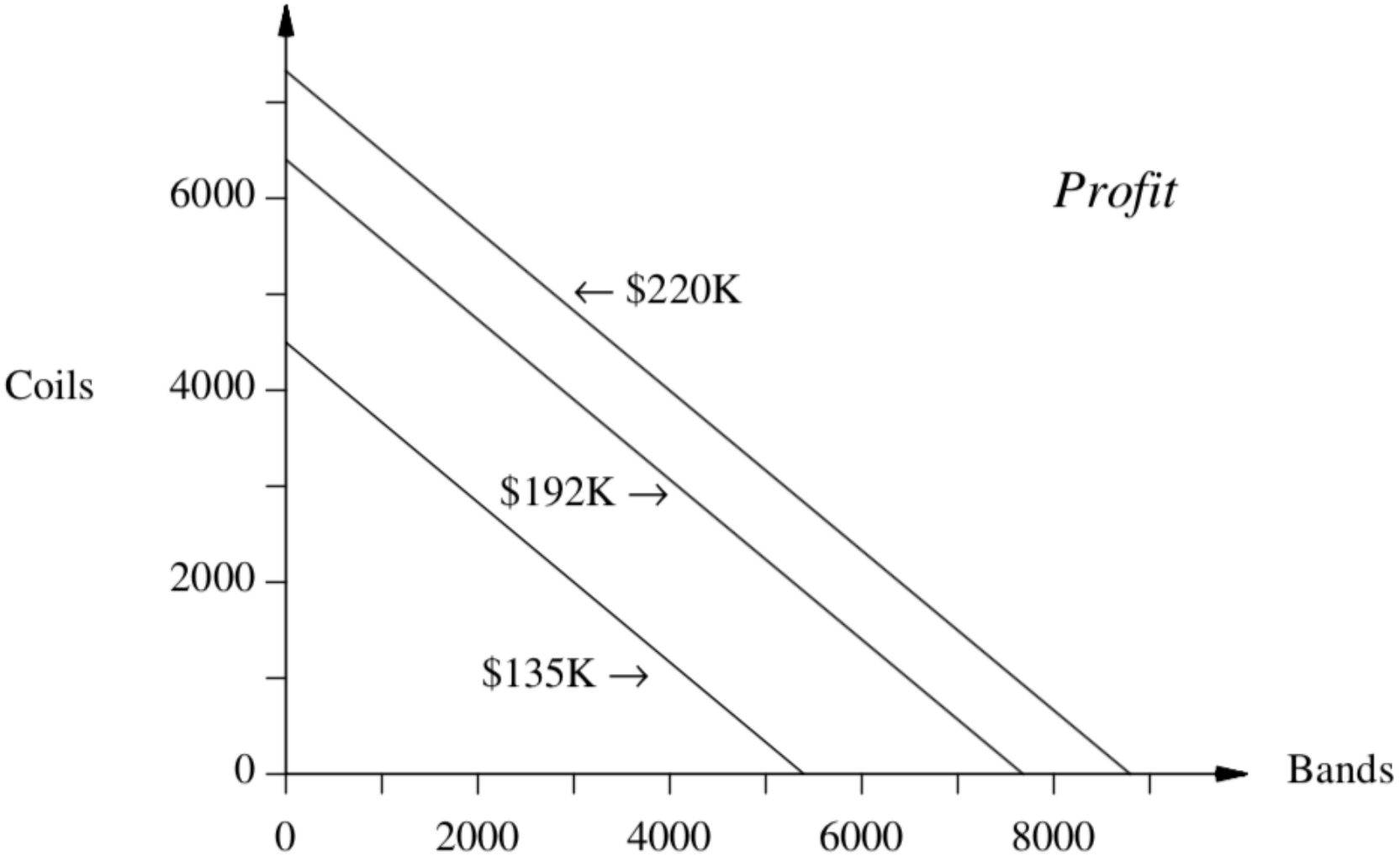
Previous linear optimization model

```
var xb;
var xc;

maximize z: 25*xb + 30*xc;

subject to
hours: xb/200 + xc/140 <= 40;

capB: 0 <= xb <= 6000;
capC: 0 <= xc <= 4000;
```

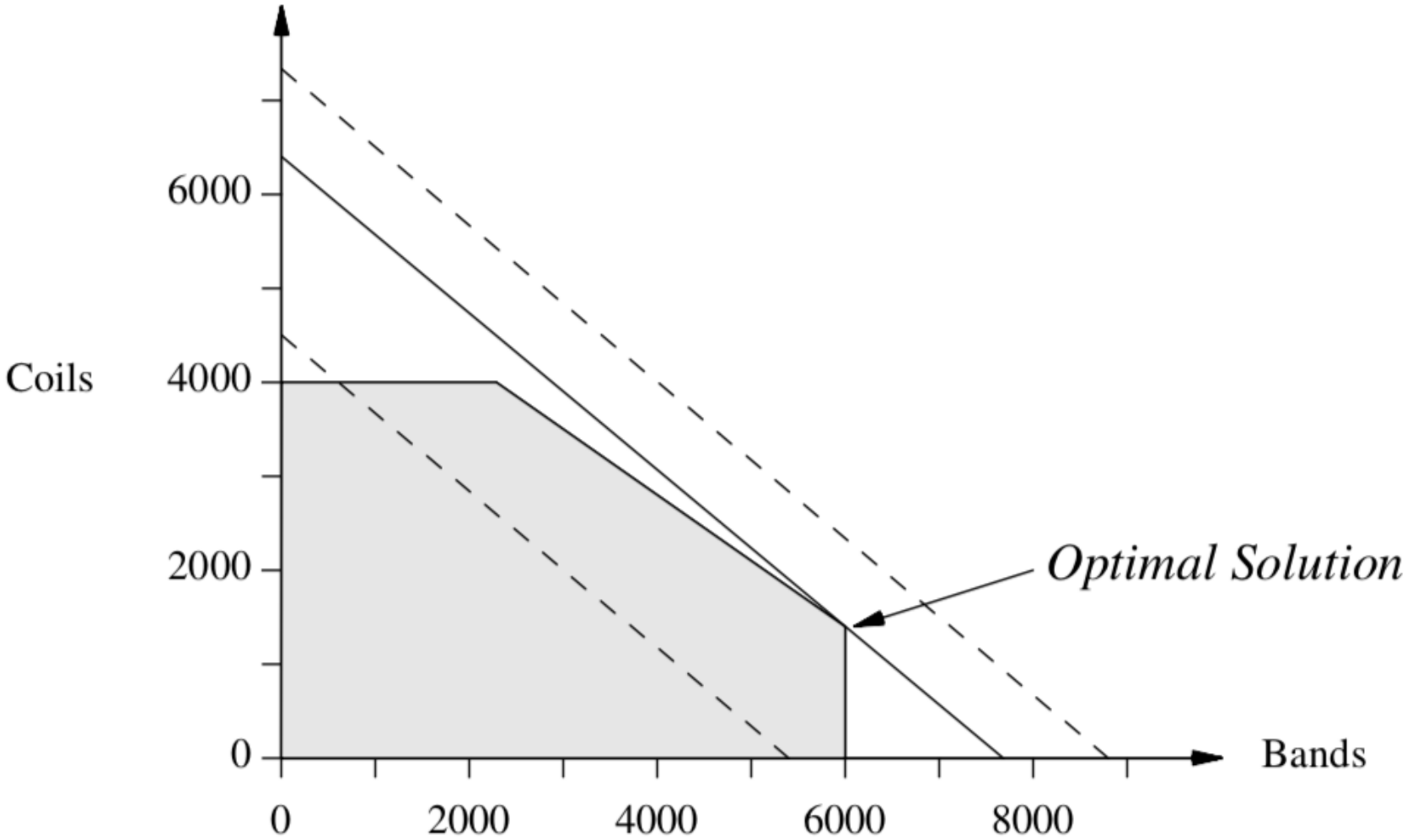
Graphical visualization: feasible region



Graphical visualization: isoprofit lines



Graphical visualization: optimum



Summary

- Linear optimization problems
  - *easy* to solve
    - efficient algorithms are known
  - supported by most software
  - given a description of the problem, an optimum solution is usually obtained in a short time
- Not all the real world problems are linear
  - in general, if not fitting in the linear optimization paradigm $\rightarrow$ nonlinear optimization problems
  - in practice, nonlinear optimization problems are often difficult to solve in a reliable manner
  - some nonlinear functions are easy; *e.g.*, convex optimization
- Integer variables
  - even if the expressions in the model are linear, the problem may be difficult
  - in general:class of difficult problems (technically: *NP-hard* class)
  - allow modeling a variety of practical situations

Bibliography

- AMPL: A Modeling Language for Mathematical Programming
  R Fourer, DM Gay, and BW Kernighan
  Second edition, ISBN 0-534-38809-4
  Available in the AMPL documentation (`http://www.ampl.com`)
- Operations research
  Wayne L. Winston
  ISBN: 9780534423629
- Integer Programming
  Laurence A. Wolsey
  Wiley-Interscience, 1998

Next class

- On labs: Software for this class
  - Python
    - data: `Pandas`
    - scientific computing: `scipy, numpy`
    - machine learning: `scikit-learn`
  - optimization: `AMPL` + `amplpy`
  - using the command line (the "shell")
- Next lecture: Linear optimization