Optimization with Big Data:
An Introduction to Linear Programming with Airline Examples in Excel

Sertac Karaman
Assistant Professor of Aeronautics and Astronautics
Laboratory for Information and Decision Systems
Institute for Data, Systems, and Society
Massachusetts Institute of Technology

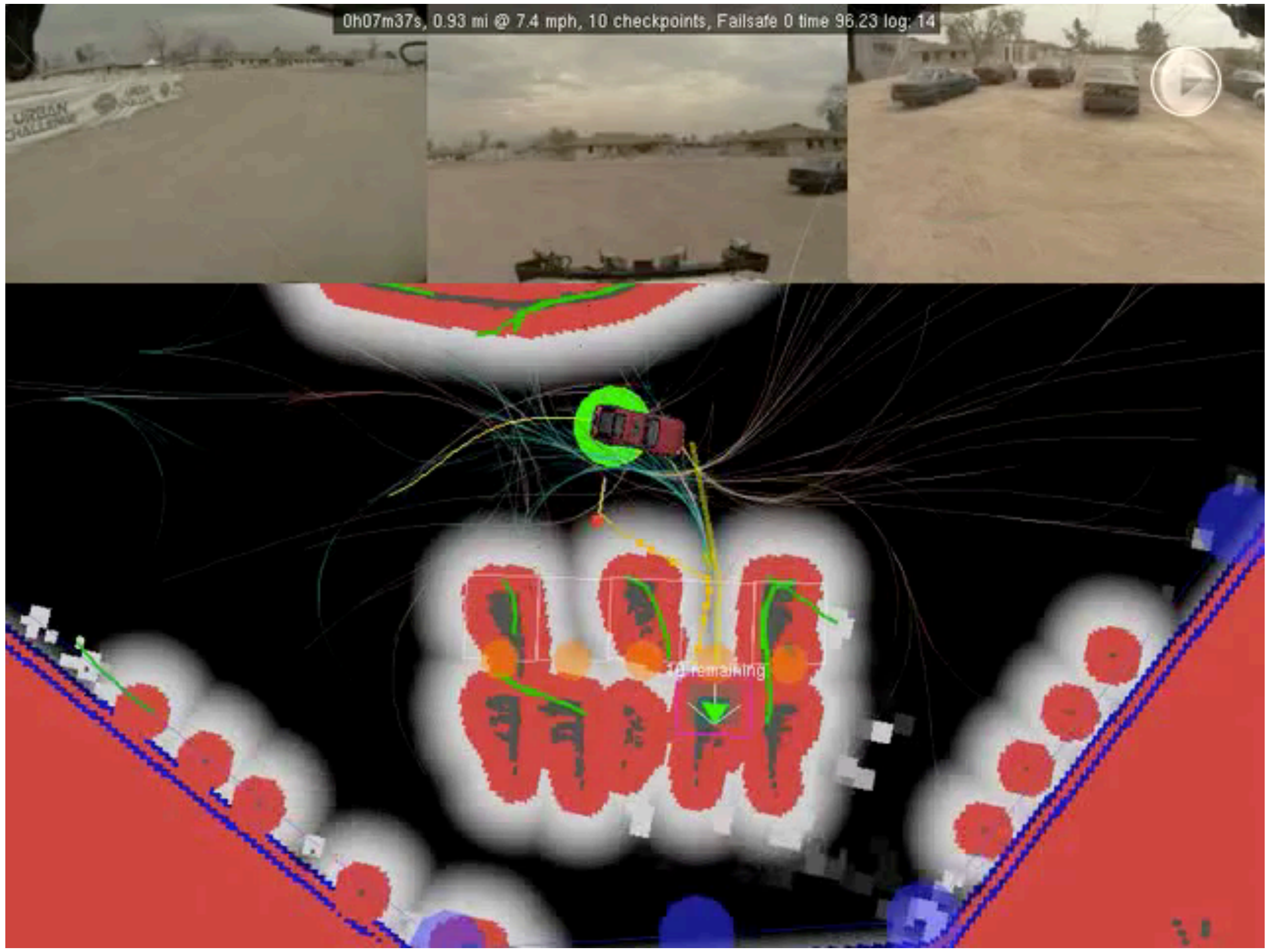
Background

- 2007, B.S., *Mechanical Engineering*
  B.S. *Computer Engineering*
  Istanbul Technical University

- 2009, M.S., *Mechanical Engineering*
  Massachusetts Institute of Technology
- 2012, Ph.D., *Electrical Engineering and Computer Science*
  Massachusetts Institute of Technology
- 2012 - present, Assistant Professor,
  Aeronautics and Astronautics,
  Massachusetts Institute of Technology

Research

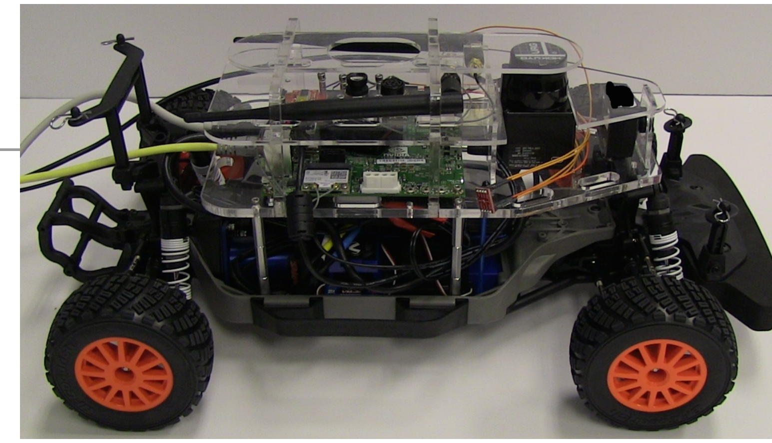
- Autonomous systems and decision making.
- Theory oriented, but build and applications as well.

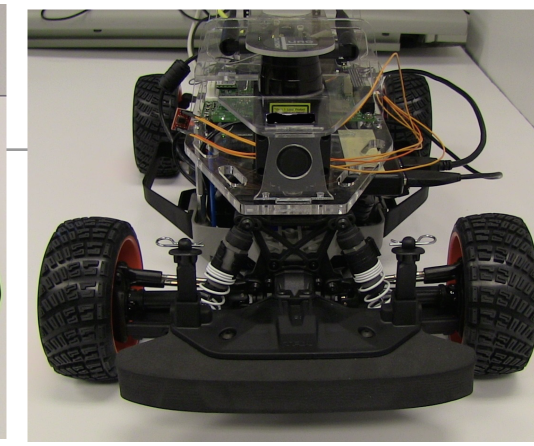


Teaching

- **Aerospace embedded systems:**

Aircraft design capstone course (with John Hansman)





Short course: Race in MIT's tunnels

- **As well as "big data", "data analytics", and "decision making" specifically *analytical methods for business professionals*.**

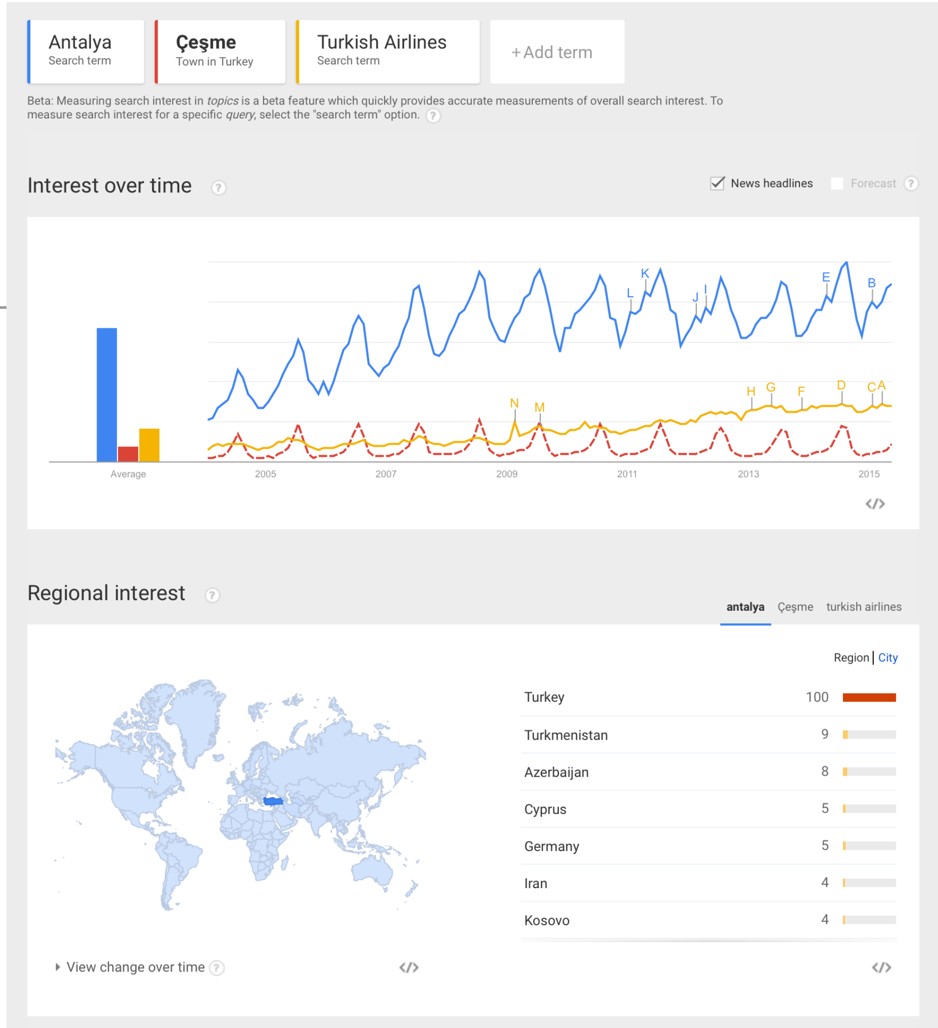
MIT AEROASTRO

Next Few "Big Data" Lectures

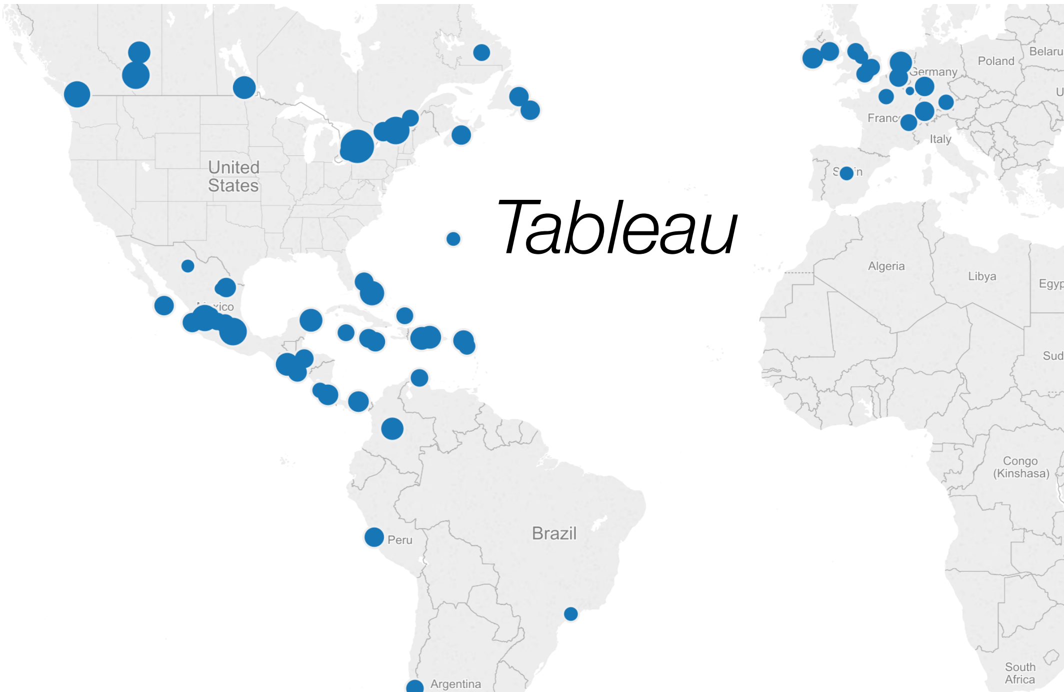
- **1. Complex decision making with big data:**
    - Understand optimization methods
    - Implement optimization using Excel
- **2. Analysis of big unstructured text:**
    - Uncovering information hidden in unstructured text (emails, social media, spoken language …)
- **3. Rapid visualization of complex big data:**
    - Tableau tutorial

- ***Methods are broadly applicable, but we present with airline examples for "Air Transport Management"***

- *<u>These classes will be MIT style! (hands on)</u>*



Google Trends



Tableau

Being Able to Utilize Big Data (and Computation) is Crucial for Almost Any Big Business Today

The New York Times

FIXES

Turning to Big, Big Data to See What Ails the World

THE WALL STREET JOURNAL

TECH

Big Data Looms as Next Battle in Europe

Lectures on Optimization Methods

- What do we want to achieve?
  - Utilize Excel to solve complex decision making problems involving big data.
  - Understand what can (and can not) be done easily (or at all?).

- What we are <u>not</u> trying to achieve?
  - Learn the complex mathematics behind decision problems.
  - Learn all possible methods for optimization methods.

Optimization Problems

- Mathematical optimization (mathematical programming) is the selection of a best option among a set of available options, defined through a mathematically-rigorous set of objectives and constraints.

- We are interested in optimization problems that involve “big data”, i.e., millions of variables to be chosen, constraints to be handled…

- Some common optimization problems:
  - Linear Programming (LP) — easy, good for big data
  - Integer Programming (IP) — hard, may be good for big data
  - Dynamic Programming (DP) — harder, rarely works with big data
  - Convex Optimization Problems — class of all ‘easy’ problems where we can work with big data
  - …

Some History

- *Fermat* & *Lagrange* found calculus based formulas (analytical solutions) to certain optimization problems.

- Algorithmic approaches developed starting from the 50s:
  - The "simplex algorithm" for "linear programming" was developed by *Dantzig* during WW II — considering logistics problems of the US military.
  - "Dynamic programming" problems are formulated and algorithms developed by *Bellman*.

- As computer technology and computation theory developed, problems that are "easier to solve" were better recognized, e.g., "convex problems".

What Are We Going to Do in This Class?

- **Linear Programming (LP)**
  - Formulation
  - Geometry & Algorithms
  - Simple applications with the Excel Solver

- **Integer Programming (IP)**
  - Formulation
  - Simple applications with the Excel Solver

- **Network Flows and Transportation Problems:**
  - Formulation as linear and integer programs
  - Applications with the Excel Solver

What We Are *NOT* Going to Do

- Dynamic programming
- Convexity and convex optimization problems
- …

The Field of Optimization is Vast!

What Kinds of Problems Can We Solve?

- **Optimal resource allocation problems:**
  - Fleet assignment
  - Crew scheduling

- **Network flow models problems:**
  - Transshipment optimization
  - Logistics (and particularly transportation)

Linear Programming Formulation

- Maximize (or minimize) an objective function with respect to constraints

$$\text{Maximize} \quad \sum_{j=1}^{n} c_j x_j$$

objective function

$$\text{subject to:} \quad \sum_{j=1}^{n} a_{ij} x_j \le b_i \quad \text{for} \quad i = 1, 2, \ldots, m$$

$$x_j \ge 0 \quad \text{for} \quad j = 1, 2, \ldots, n$$

constraints

Linear Programming Formulation

LP as a "Decision Making" Problem:

Maximize $Z = c_1x_1 + c_2x_2 + \ldots + c_nx_n$

Subject to:

$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \leq b_1$

$a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \leq b_2$

...

$a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \leq b_m$

and $x_1 \geq 0, x_2 \geq 0, \ldots, x_n \geq 0$

Value maximization subject to resource constraints

x_j Decision variable
- *we would like to determine (decide) this*

Given by the problem instance

c_j The unit value for the *j*th decision variable

b_i The available amount for the *i*th resource

a_{ij} The unit of *i*th resource required for one unit of the *j*th decision variable

Linear Programming Formulation

Maximize $Z = c_1x_1 + c_2x_2 + \ldots + c_nx_n$

Subject to:

$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \leq b_1$

$a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \leq b_2$

...

$a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \leq b_m$

and $x_1 \geq 0,\ x_2 \geq 0,\ \ldots,\ x_n \geq 0$

- **Feasible Solution:** A set of values for the decision variables (x_1, …, x_n) that satisfies all of the constraints.
- **Optimal Solution:** A feasible solution that gives the best value ($\sum_{j=1}^{n} c_i x_i$) among all feasible solutions.

Linear Programming Formulation

Maximize $Z = c_1x_1 + c_2x_2 + \ldots + c_nx_n$

Subject to:

$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \leq b_1$

$a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \leq b_2$

...

$a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \leq b_m$

and $x_1 \geq 0,\ x_2 \geq 0,\ \ldots,\ x_n \geq 0$

Minimize $Z = c_1x_1 + c_2x_2 + \ldots + c_nx_n$

Subject to:

$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \leq b_1$

$a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \leq b_2$

...

$a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \leq b_m$

and $x_1 \geq 0,\ x_2 \geq 0,\ \ldots,\ x_n \geq 0$

- The problem can be cast either as a maximization or a minimization problem. (Just multiply all the “value” constants c_i by -1.)

Linear Programming Formulation

Cost minimization subject to resource constraints

x_j Decision variable

Given by the problem instance

c_j The unit **cost** for the *j*th decision variable

b_i The available amount for the *i*th resource

a_{ij} The unit of *i*th resource required for one unit of the *j*th decision variable

$$\text{Minimize } Z = c_1x_1 + c_2x_2 + \ldots + c_nx_n$$

Subject to:

$$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \leq b_1$$

$$a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \leq b_2$$

...

$$a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \leq b_m$$

$$\text{and } x_1 \geq 0,\ x_2 \geq 0,\ \ldots,\ x_n \geq 0$$

- The problem can be cast either as a maximization or a minimization problem. (Just multiply all the “value” constants c_i by -1.)
- In the new problem, c_i can be interpreted as cost.

MIT AEROASTRO

LP Example: Maximizing Capacity with Constrained Crew and Vehicle Supply

- Freight needs to be carried with trucks.
- Two types of trucks are available in limited numbers, each with different capacity and crew requirements:

	Capacity	Crew required	Number available
Anadol	300	3	40
BMC	500	2	60

- We have a limited amount of crew: **Exactly 180 number of personnel available.** All personnel can operate either truck.

- ***How many trucks of each type would you utilize?***

LP Example: Maximizing Capacity with Constrained Crew and Vehicle Supply

- Freight needs to be carried with trucks.
- Two types of trucks are available in limited numbers, each with different capacity and crew requirements:

	Capacity	Crew required	Number available
Anadol	300	3	40
BMC	500	2	60

- We have a limited amount of crew: **Exactly 180 number of personnel available.** All personnel can operate either truck.

- ***How many trucks of each type would you utilize?***

Wait a second! - Do I really need LP for this??

LP Example: Maximizing Capacity with Constrained Crew and Vehicle Supply

1. Define the decision variables:

   x_1 : Number of Anadol trucks

   x_2 : Number of BMC trucks

2. Write down the objective function:

$$\text{Maximize} \qquad 300x_1 + 500x_2$$

Maximize capacity

3. Write down all of the constraints:

$$\text{Subject to} \qquad 3x_1 + 2x_2 \leq 180$$

Crew limitations

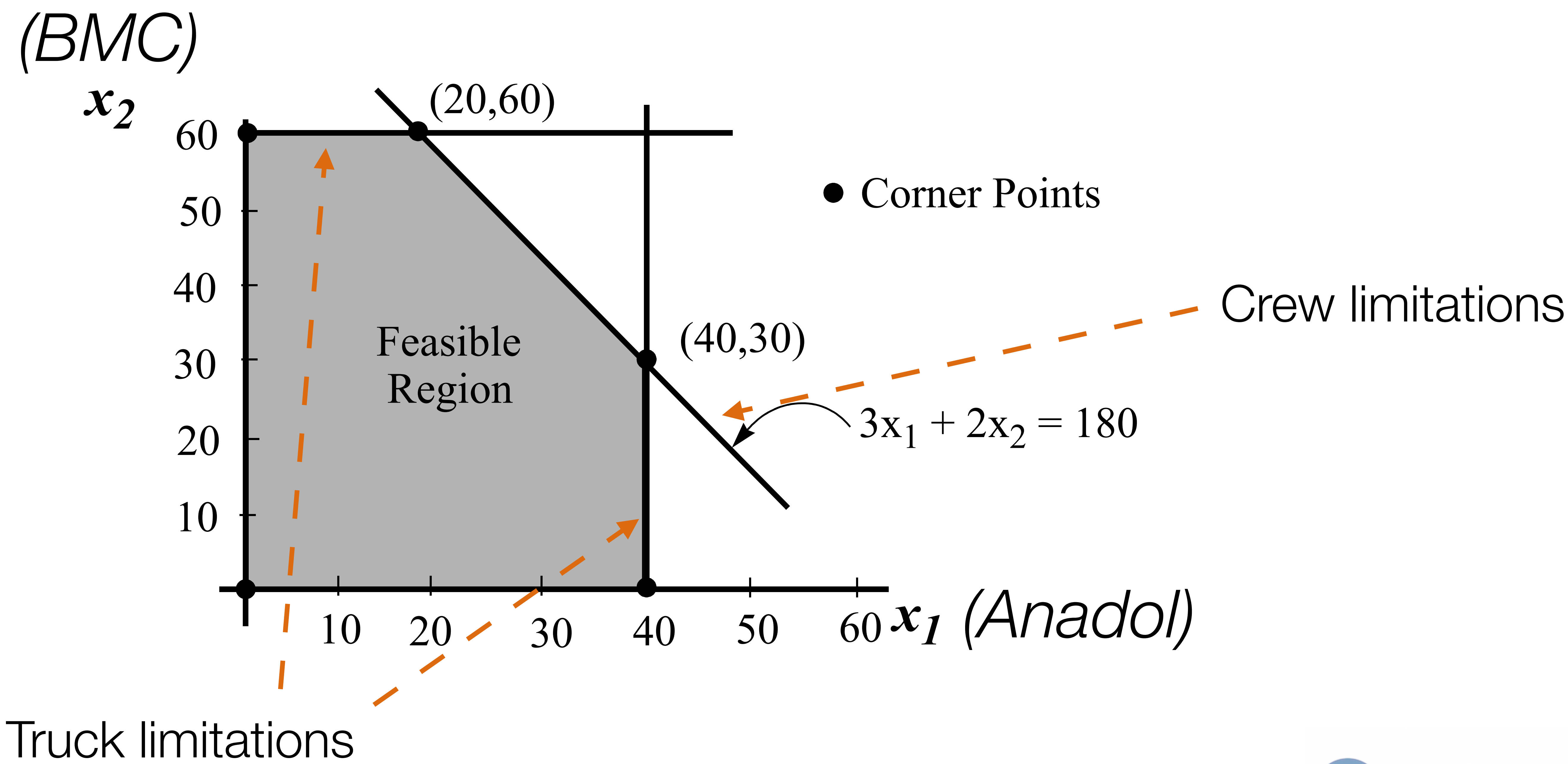
$$x_1 \leq 40$$
$$x_2 \leq 60$$

Truck limitations

$$x_1 \geq 0; x_2 \geq 0$$

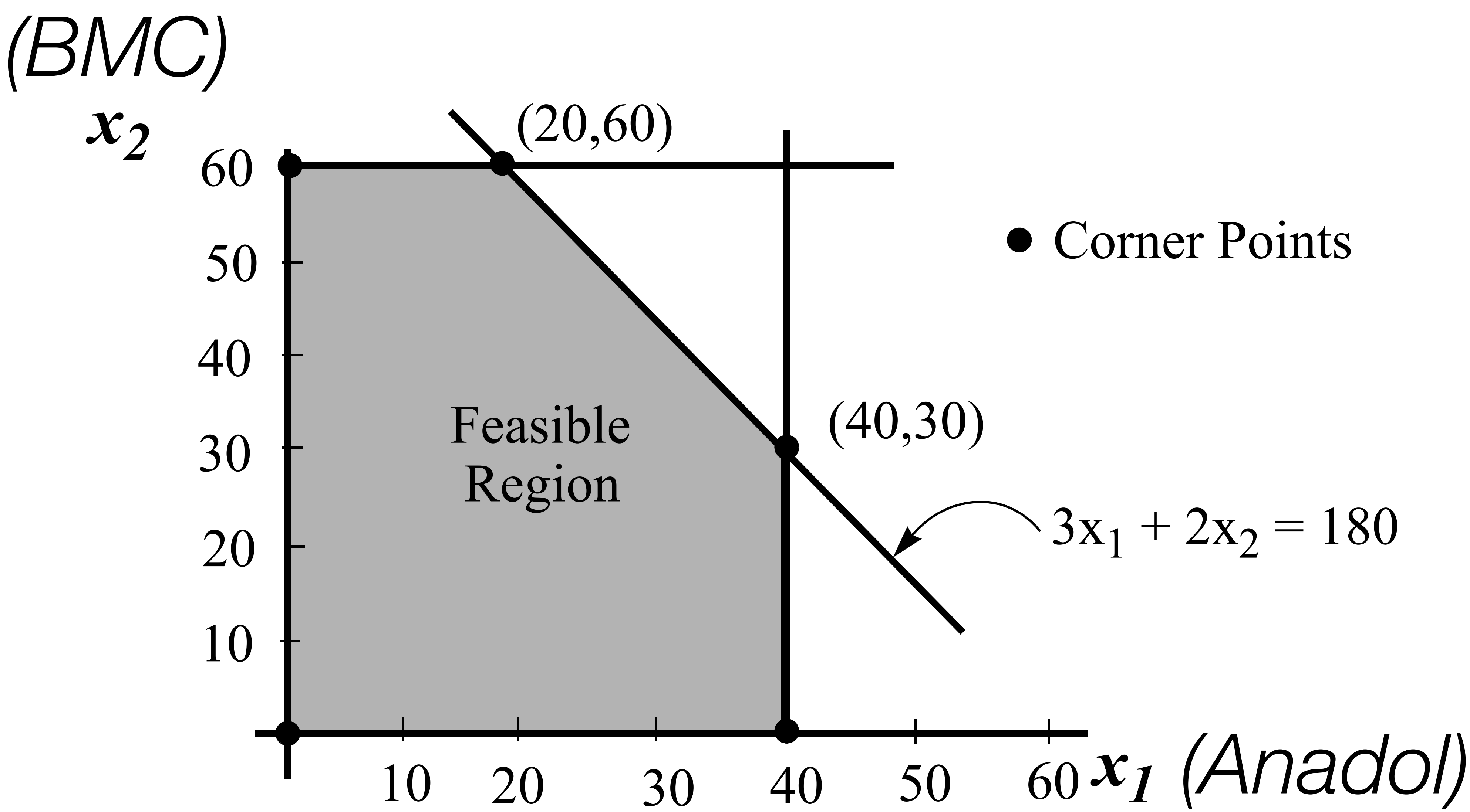
LP Geometry

- Let's look at the "decision space" - the set of all "possible decisions"



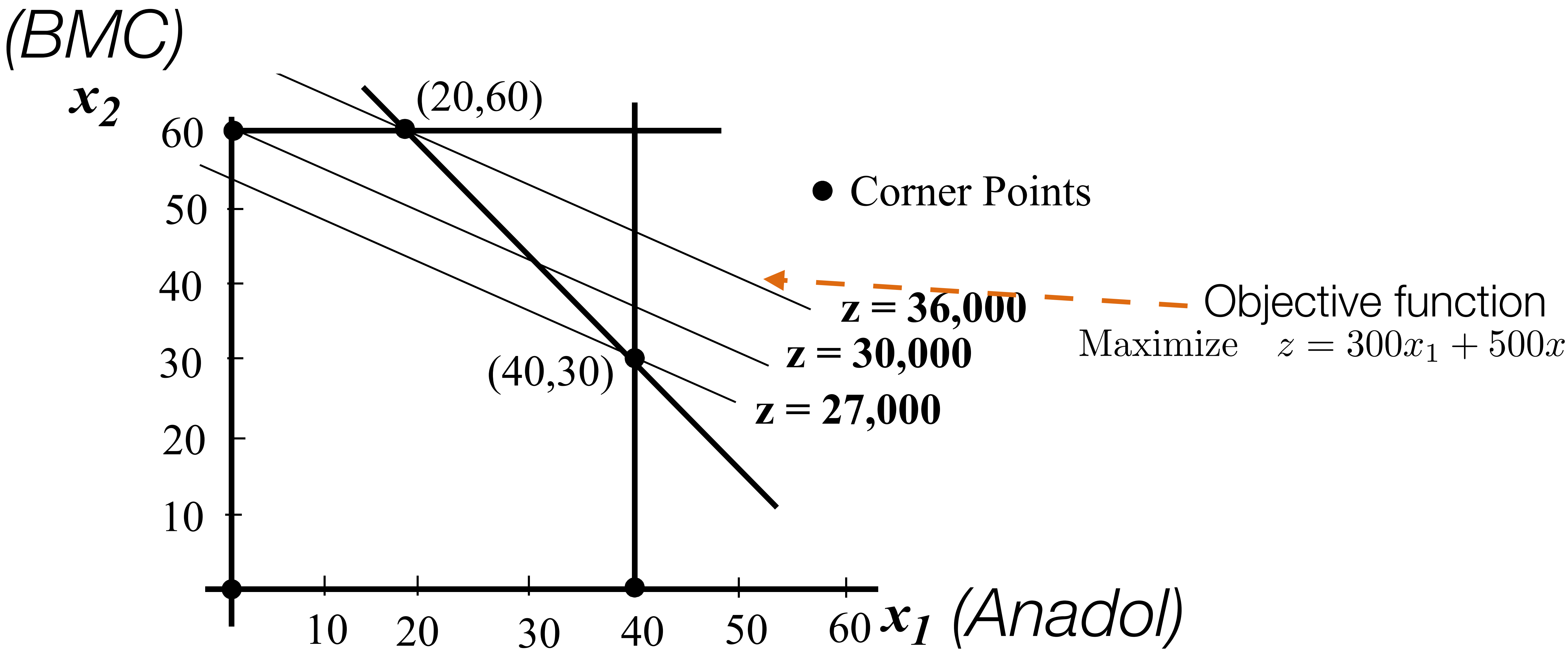
LP Geometry

- Let's look at the "decision space" - the set of all "possible decisions"



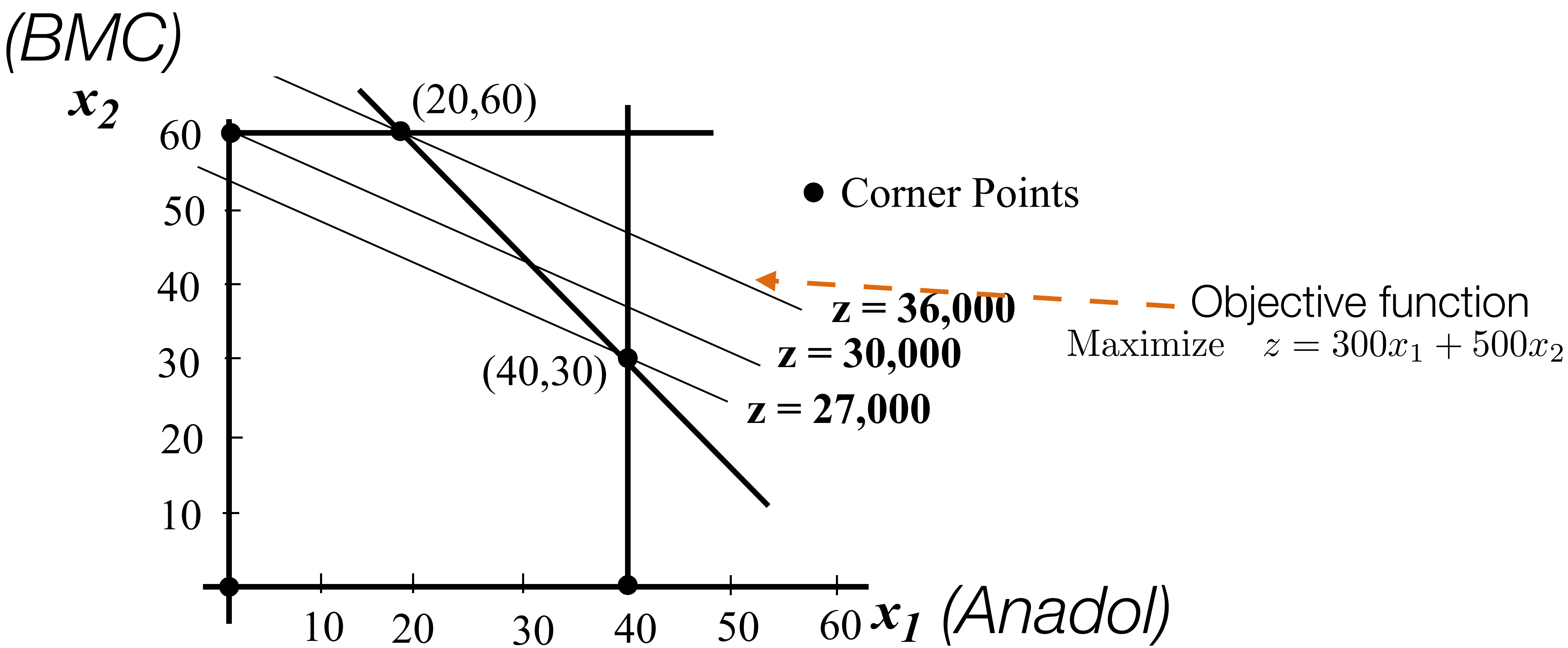
LP Geometry

- Let's look at the "decision space" - the set of all "possible decisions"



LP Geometry

- Let’s look at the “decision space” - the set of all “possible decisions”



An optimal solution can only be in corner points!!

Remark 1: The Complexity of the Problem

- Although this problem could be solved easily.
- In slight variations, it is not possible to eyeball the solution:

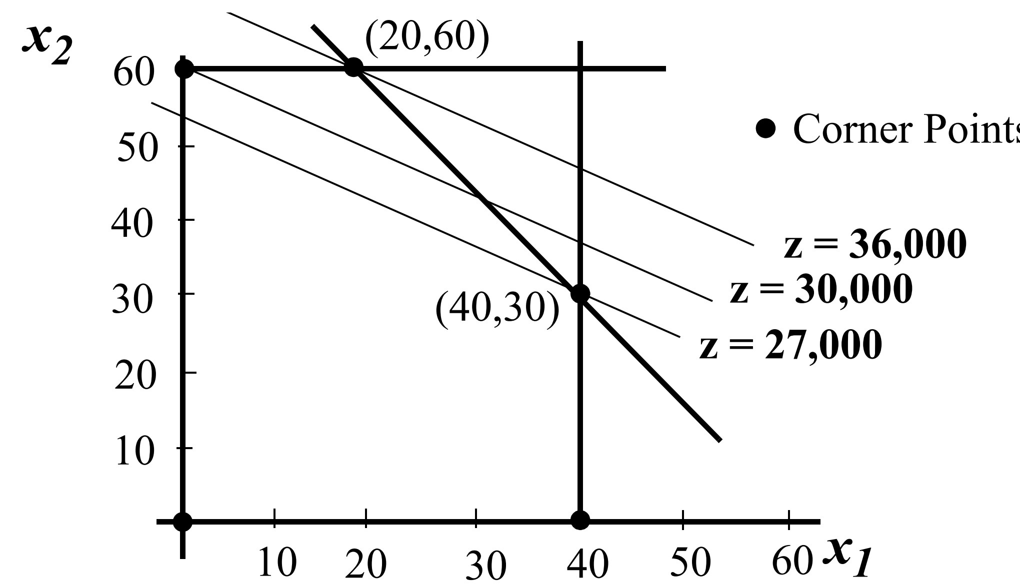
	Capacity	Crew required	Number available
A	500	3	40
B	300	2	60

- In this case, neither truck is “strictly better” than the other…

- The problem becomes much more complicated when:
  - the number of decision variables (e.g., choices for trucks) gets much larger,
  - the types of resources are many more,
  - these resources are coupled with one another in a complex manner.

Remark 2: Integer Variable

- Let's admit: I got really lucky here:



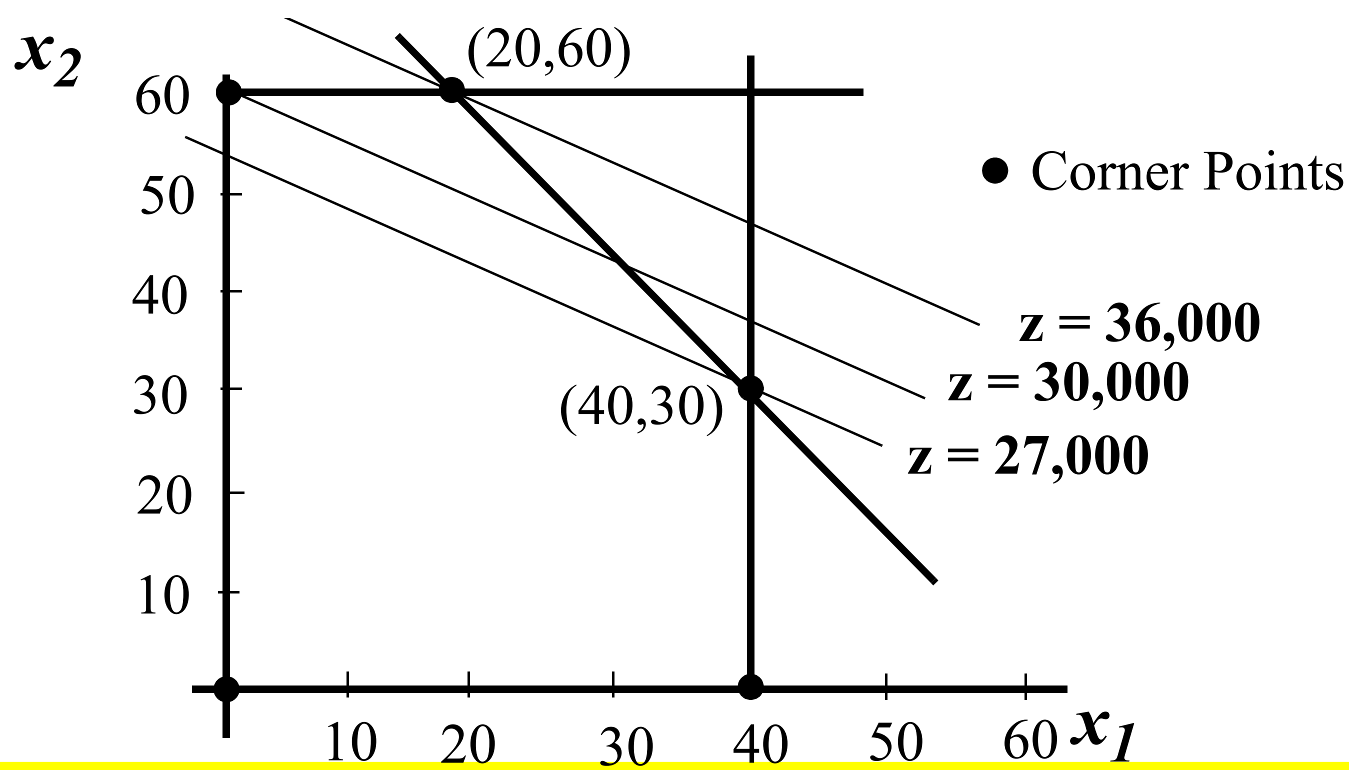
- What if I had the following problem:

	Capacity	Crew required	Number available
Anadol	300	3	40
BMC	500	2	60

- And I have only 175 crew members.
- Then the optimal solution is $x_1 = 60; \quad x_2 = 55/3 = 18.33$

Algorithms for Solving LP

- The Simplex Algorithm (by Dantzig ’48) is based on this very fact:
  All optimal solutions occur at some corner point.
- Roughly speaking, the Simplex algorithm carefully jumps through the corners to improve the value and reach an optimal solution eventually.
- *Many other (more recent) techniques are based on another fact: convexity*



This class will not look at how algorithms work.
We will focus on formulations and software!

Software for Solving LP

- Many software packages are available:
  - LINDO: Linear INteractive Discrete Optimizer
  - GAMS (also solves non-linear problems)
  - MINUS
  - Matlab Optimization Toolbox
  - QSB (LP, IP, DP, and others)
  - AMPL/CPLEX

The Excel Solver

- We will utilize the Excel Solver:
  - Solver is a "Generalized Reduced Gradient (GRG2)" implementation.
  - Developed by Leon Lasdon (UT Austin) and Allan Waren (Cleveland State University)
  - Solver allows for one function to be minimized, maximized, or set equal to a specific value.
  - Convergence criteria (convergence), integer constraint criteria (tolerance) are accessible through the OPTIONS button.
  - It also provides a simplex algorithm for solving LPs.
  - To use this tool in Excel you need to enable the Solver add-in.

By the Way, What Other Magic Can Excel Do?

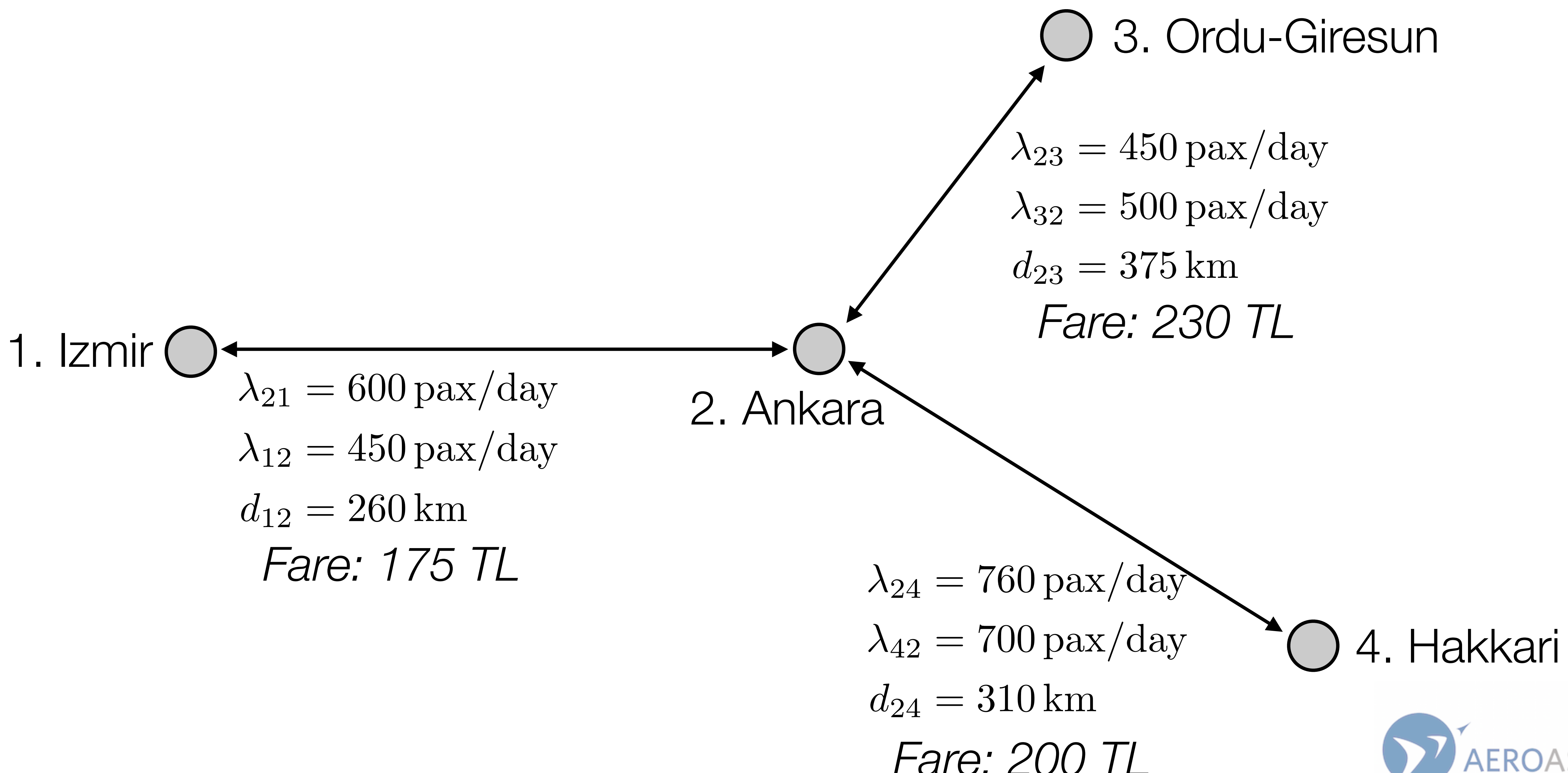
- Excel can solve simultaneous linear equations using matrix functions

- Excel can solve one nonlinear equation

- Excel does not have direct capabilities of solving n multiple nonlinear equations in *n* unknowns, but sometimes the problem can be rearranged as a minimization function

More Complex Problems: Airline Scheduling

- A small airline would like to use mathematical programming to schedule its flight to maximize profit. The following map shows the city pairs to be operated:

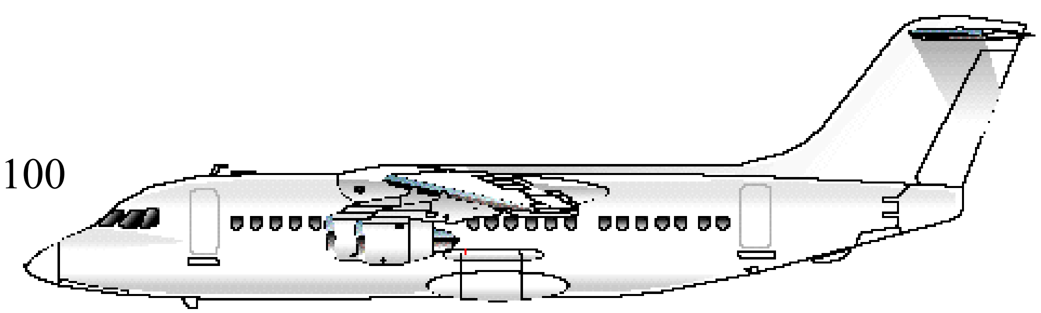


More Complex Problems: Airline Scheduling

- The Airline decided to purchase two types of aircraft:
  - Embraer 145 (45-seater)
  - Avro RJ-100 (100-seater)

EMB-145

Avro RJ-100



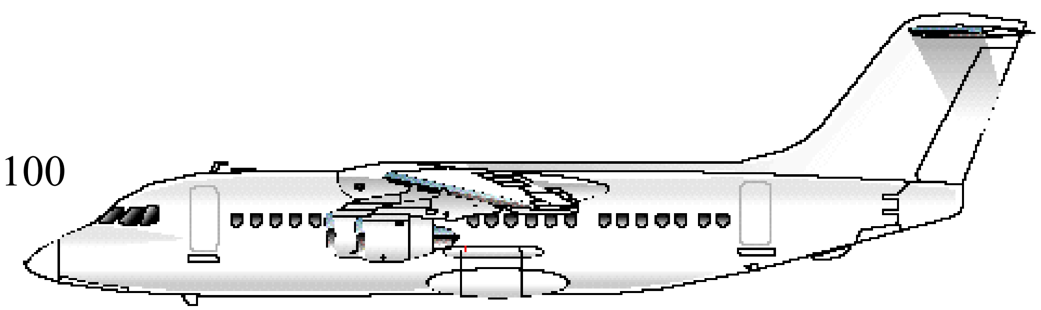
	EMB-145	Avro RJ-100
Seating capacity - n_k	50	100
Speed (km/hr) - v_k	400	425
Operating cost (TL/hr)- c_k	1,850	3,800
Max utilization (hr/day) - U_k	13	12

More Complex Problems: Airline Scheduling

- The Airline decided to purchase two types of aircraft:
  - Embraer 145 (45-seater)
  - Avro RJ-100 (100-seater)

EMB-145

Avro RJ-100



	EMB-145	Avro RJ-100
Seating capacity - n_k	50	100
Speed (km/hr) - v_k	400	425
Operating cost (TL/hr)- c_k	1,850	3,800
Max utilization (hr/day) - U_k	13	12

- ***Can you select the fleet for each route and the frequency of flights to maximize your profits?***

More Complex Problems: Airline Scheduling

- MAXIMIZE *Profit*
- SUBJECT TO:
    - Aircraft availability constraint (aircraft are limited by utilization time)
    - Demand fulfillment constraint (make sure that the demand can be served)
    - Minimum frequency constraint (good to gain and hold market share)
    - …

More Complex Problems: Airline Scheduling

- Step 1: Define the decision variables:

A_k Number of aircraft of type *k*

N_{ijk} Number of flights from *i* to *j* using aircraft *k*

$k = 1, 2$ We can use (1) EMB-145 or (2) Avro RJ-100

$i, j = 1, 2, 3, 4$ We can go to cities (1) Izmir, (2) Ankara, (3) Giresun and (4) Hakkari.

More Complex Problems: Airline Scheduling

- Step 2: Write the objective function
  - Revenue: $$\sum_{(i,j)} \lambda_{ij} fij$$
  - Costs: $$\sum_{(i,j)} \sum_{k} N_{ijk} C_{ijk}$$
  - Maximize Profit: $$\text{Maximize} \quad \sum_{(i,j)} \lambda_{ij} fij - \sum_{(i,j)} \sum_{k} N_{ijk} C_{ijk}$$

Problem-dependent Constants:

λ_{ij} Demand from *i* to *j*

f_{ij} Fare from *i* to *j*

C_{ijk} Cost of flight *i* to *j* using aircraft k

Decision Variables

A_k Number of aircraft of type *k*

N_{ijk} Number of flights from *i* to *j* using aircraft *k*

More Complex Problems: Airline Scheduling

- Step 3: Write the constraints
  3.1. Aircraft availability constraint:

Problem-dependent Constants:

(block time) (num flights)
< (utilization) (num aircraft)

$$\sum_{(i,j)} t_{ijk} N_{ijk} \leq U_k A_k \quad \text{for all } k$$

t_{ijk} Time it takes for aircraft *k* to go from *i* to *j*

U_k Maximum utilization for aircraft *k*

Decision Variables

A_k Number of aircraft of type *k*

N_{ijk} Number of flights from *i* to *j* using aircraft *k*

MIT AEROASTRO

More Complex Problems: Airline Scheduling

- Step 3: Write the constraints
  3.2. Demand fulfillment constraint

Problem-dependent Constants:

(Supply of seats offered)
> (Demand for service)

$$\sum_{k}\sum_{k} l\, n_k N_{ijk} \geq \lambda_{ij} \quad \text{for all } (i, j)$$

n_k Seating capacity of aircraft *k*

λ_{ij} Demand from *i* to *j*

l Load factor

Decision Variables

A_k Number of aircraft of type *k*

N_{ijk} Number of flights from *i* to *j* using aircraft *k*

More Complex Problems: Airline Scheduling

- Step 3: Write the constraints
  3.3. Minimum frequency constraint

Problem-dependent Constants:

(Num flights between i and j)
> (Minimum num desired flights)

$$\sum_k N_{ijk} \geq (N_{ij})_{\min} \quad \text{for all } (i, j)$$

$(N_{ij})_{\min}$ Minimum flight frequency between *i* and *j*

Decision Variables

A_k Number of aircraft of type *k*

N_{ijk} Number of flights from *i* to *j* using aircraft *k*

More Complex Problems: Airline Scheduling

Decision Variables

A_k Number of aircraft of type *k*

N_{ijk} Number of flights from *i* to *j* using aircraft *k*

Mathematical Program

$$\text{Maximize} \quad \sum_{(i,j)} \lambda_{ij} f_{ij} - \sum_{(i,j)} \sum_{k} N_{ijk} C_{ijk}$$

$$\text{Subject to} \quad \sum_{k} n_k N_{ijk} \geq \lambda_{ij}, \text{ for all city pairs } (i,j)$$

$$\sum_{(i,j)} t_{ijk} N_{ijk} \leq U_k A_k, \text{ for all aircraft } k$$

$$\sum_{k} N_{ijk} \geq (N_{ij})_{\min}, \text{ for all city pairs } (i,j)$$

Problem-dependent Constants:

λ_{ij} Demand from *i* to *j*

f_{ij} Fare from *i* to *j*

C_{ijk} Cost of flight *i* to *j* using aircraft k

t_{ijk} Time for aircraft *k* to go from *i* to *j*

U_k Maximum utilization for aircraft *k*

$(N_{ij})_{\min}$ Minimum flight frequency between *i* and *j*

n_k Seating capacity of aircraft *k*

What is Next regarding Optimization?

- Excel Solver Session this afternoon:
  - Solve the truck selection and the crew scheduling problems in Excel

- We will formalize three types of problems on June 6th
  - Transportation Problems
  - Assignment Problems
  - Network Optimization Problems