Mathematics and Industry

How to Survive Real-World Projects as a Mathematician

Thorsten Koch

Zuse Institute Berlin

TU Berlin

CO@Work 2015

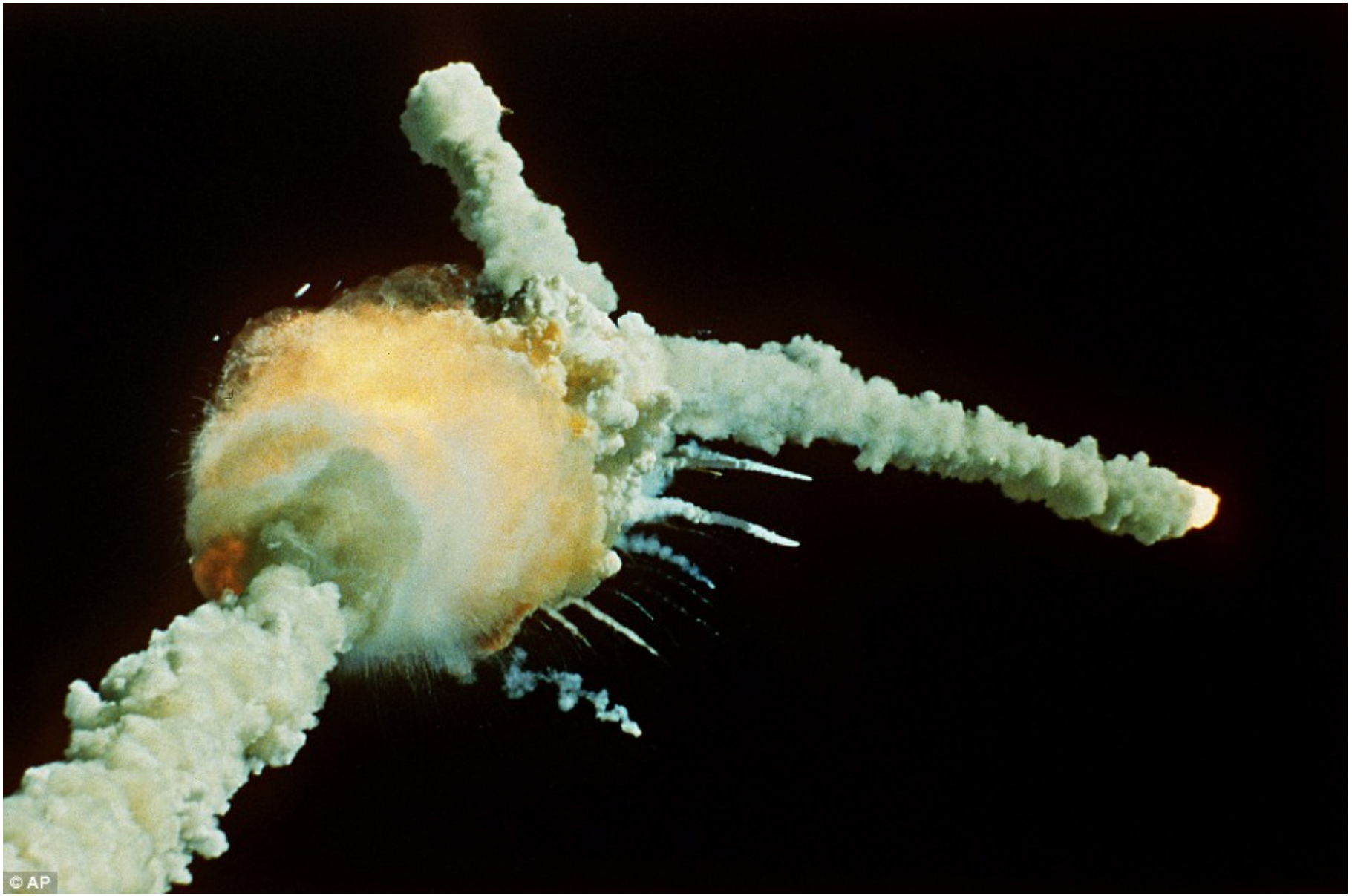
What does the R in **R&D** project mean?

What does **Research** mean?

- It means the outcome is not yet determined.
- A research project can succeed by showing that it is not possible to achieve the initial goal.
- Research is conducted by systematically trying **untested** paths and devising new methods.

Consequently, it is hard (impossible) to predict the progress of the project. For sure it will not be linear in time.

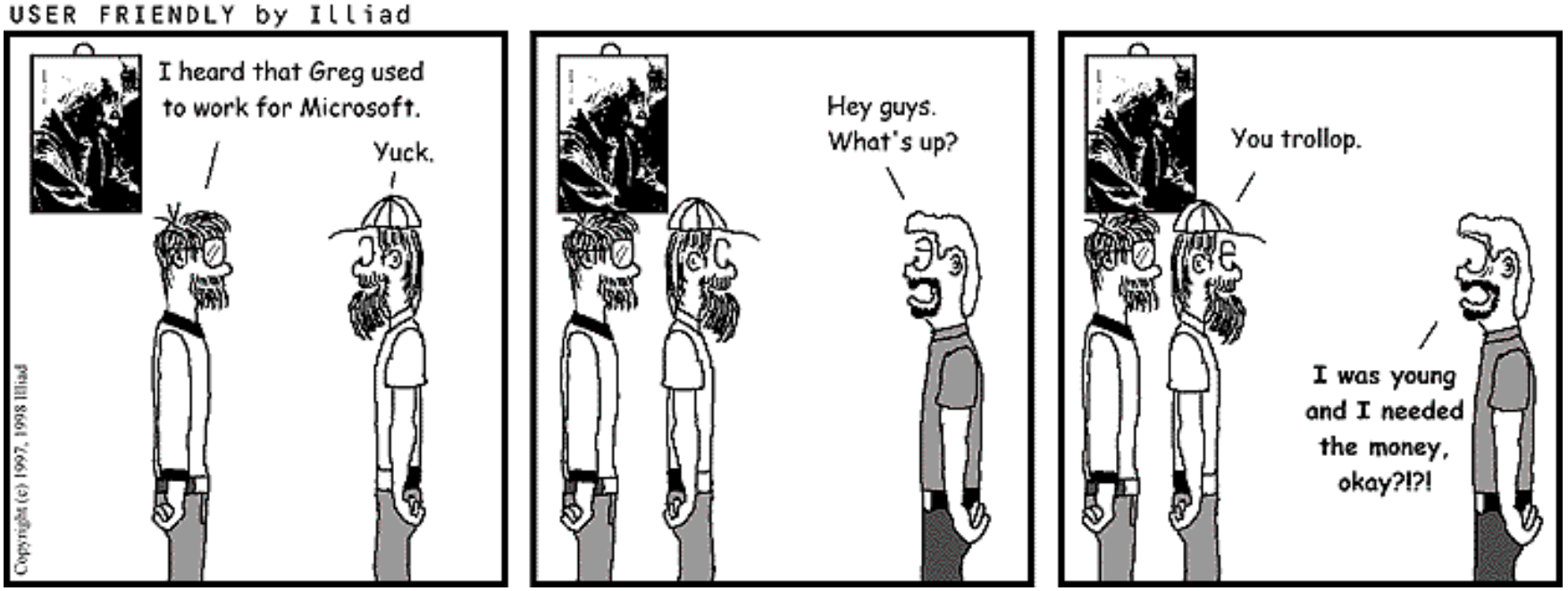
What does the R in **R&D** project mean?



Why do real-world projects at all?

- take a lot of time and effort
- can fail miserably
- often lack theoretical appeal
- results may be hard to publish

- have impact in the real-world
- challenging because rules are set
- improve something people use
- somebody actually cares about the result



WRONG ANSWER!

Getting the project

How to convince the industry people that you can help them:

- They are the specialists for the topic not you.
- Be aware they do not want a result that says:
  We can prove there exists a unique solution.
- Even if you know something about their business, regardless whom you ask, they will tell you : "We are special"
  Corollary: *Since everybody is special, they are all equal.*
- If you try to convince them by showing something similar, they might have a very narrow view with little abstraction ability.
- If asked, how much you can improve on the current solution, the correct answer is 15%

  (see G. Dueck, DMV-Mitteilungen, 2003, 44-45)

The improvement potential is always 15%

- 5% ⇨ “So much we save by simply pushing the employees.”
- 10% ⇨ “Sounds poor. We could do similar ourselves if we would get as much money as you ask for.”
- 20% ⇨ “this sounds very ambitious. You must remember: if we give you the money, we have to promise 20% to our boss. We dare not to do this.“
- 30% ⇨ “Braggart! Get out!”

From this it follows that you have to say 15%.

- I said 15% and immediately got a signature
- I said 13 % ⇨ “Why such a crooked number? How could you be so precise?”
- I said 14%, same result.

I stayed at 15 percent. Always 15 percent. Only 15 percent. All nodded, everybody satisfied. I had discovered an absolute Natural constant!

Mathematics always saves 15%. Completely regardless of the Problem!

15 Prozent. QED. Oder gibt es schon falsche Fünfzehner?! Gunter Dueck. Das Sintflutprinzip. Springer. 2000

Pilot projects

Sometimes a company will suggest to do a pilot project first:

- The unspoken expectation is that you put in more resources than what you are paid for.
- Chances for a continuation project are as good with or without a pilot project.
- You will have trouble to get up your prices again afterwards.
- If you do this, the default has to be the continuation. Just suggest the right to drop out at a certain point in case of failure.

Making a contract

- Intellectual property rights (patents)
- 3rd party code
- Don't do maintenance (could be done by a (spinoff) company)
- Right to publish
- Right to give talks
- Right to cooperate with others
- Right to continue afterwards with others (competitors)

Remember: The contract is basically useless, as you will never sue and can do little later on.

Legal stuff

censored

Administrative stuff

Is your product subject to the U.S. Export Administration Regulations (EAR), 15 CFR Parts 730-774?

Does your product fall under the U.S. Government International Traffic in Arms Regulations (ITAR) or any other local government regulations which uniquely control the product for military or encryption reasons?

Please indicate any available classification under EU Law (Council Regulation (EC) No 428/2009 of 27 August 2009 setting up a Community regime for the control of exports of dual use items and technology)

a. ECCN:

b. Does the General Software Note apply?

c. Comments:

CCATS : Commodity classification automated tracking system, alpha-numerical code assigned by US BIS.

Do you have the heart to sell software?

One option is only to provide a method that afterwards is implemented (again) by a commercial software developer.

Communication

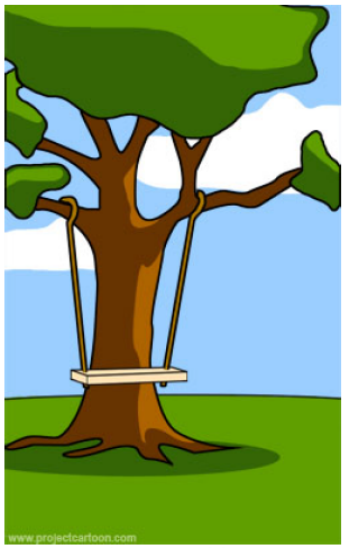
The same words may have different meanings in different Communities:

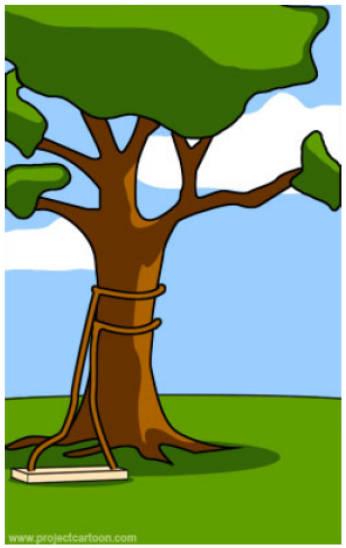
Speak the language of the problem owner

- Technical terms
- Mother tongue
- Their problem is your problem and your solution has to become their solution.
- Do not trust assumptions.
- Convince the decision makers – not only the techies.

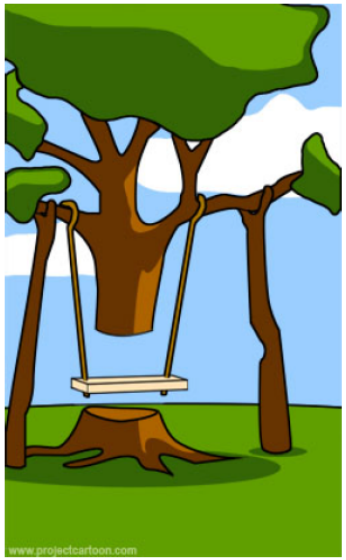
Describing the problem

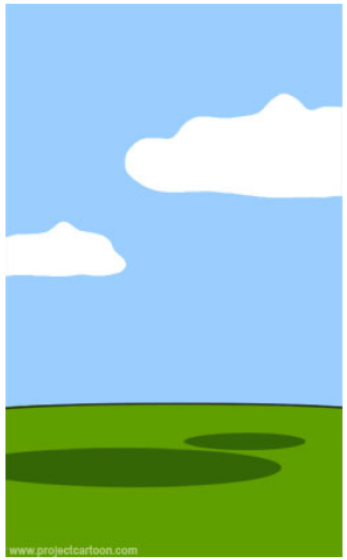
What the industry wanted

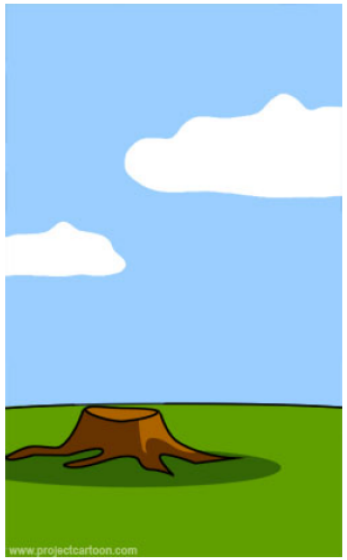

How the practitioners described it


What the mathematicians understood


How it was modelled


How it was implemented


How the project was documented


How it was supported


What was really needed

Modelling or what is the objective?

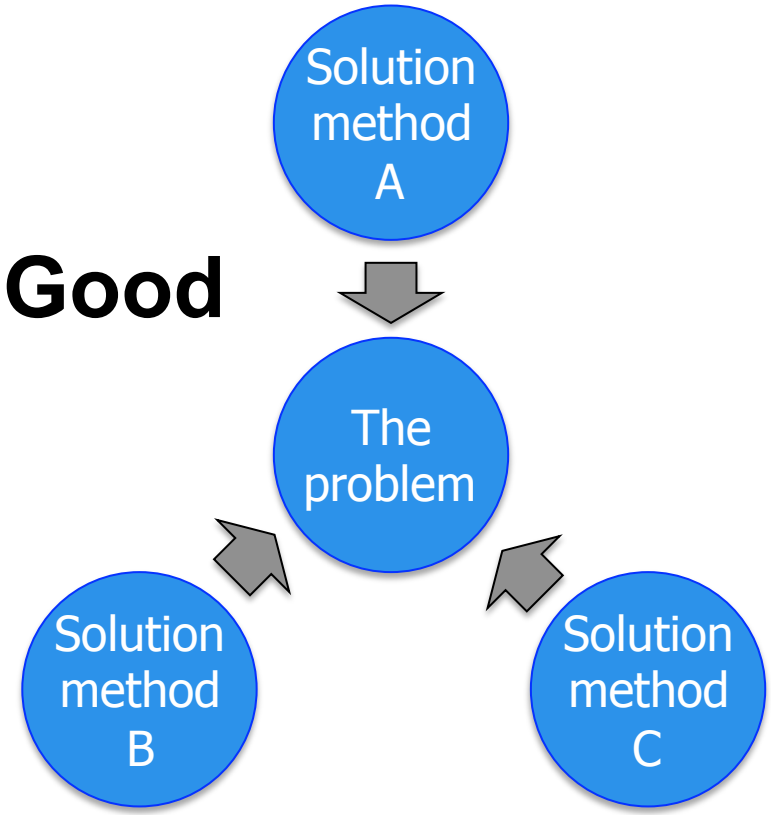
- Find out what the real objective is.
  Usually it is not what are told in the beginning.
- Sometimes it is difficult because the cost impact when optimizing virtual or already existing structures is unclear.
- Often you have to compare apples and pies.
- Preprocessing is important. Most real-world problems are rather big, at least compared to “academic toy examples”.
- Decide what you have to model and what to ignore (or fix later)

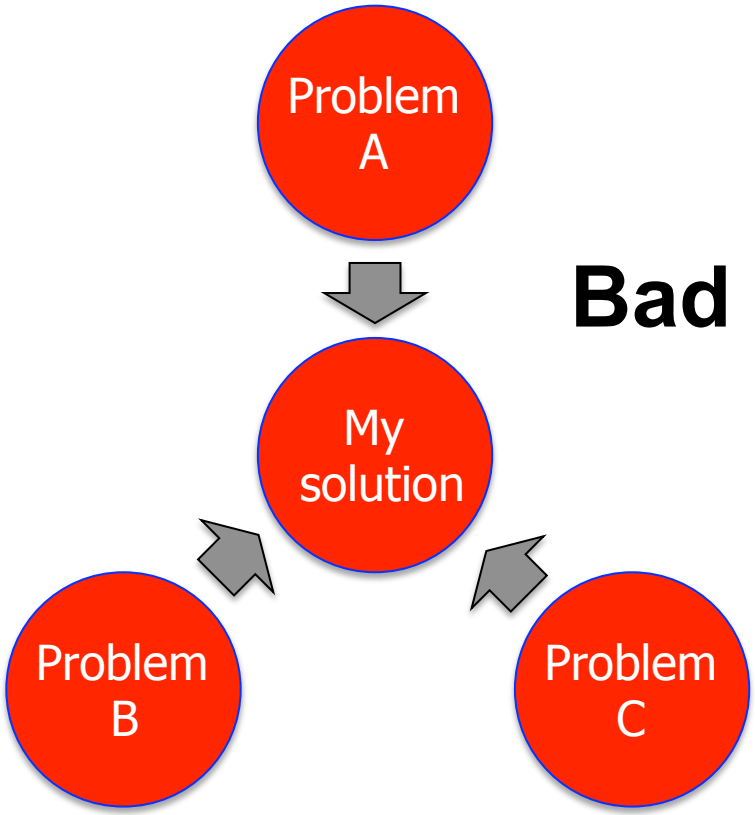
The traveling salesman problem is to mathematical programming what chess is to artificial intelligence: thoroughly useless and fiercely competitive sport that serves as a testing ground of your techniques.

—Vasek Chvatal

Think and act problem centric

- It is more important to solve the right problem than to solve the problem right.
- Identifying the problem is half of the way to the solution.





The project setup / classical approach

We have a difficult problem to solve,

The project setup / classical approach

we have a wise teacher,

The project setup / classical approach

… and we have a very determined PhD student.

Now, the student supervised by the teacher attacks the problem.

This is what we call the classical “Hero Approach”.

The project team

What if the problem is too big and you need a whole team to tackle it?

Maybe you do not have the necessary expertise and need to cooperate with other institutions.

Mathematical research usually has no suitable infrastructure to run big projects with non-disjunctive tasks.

The most important thing

DATA

The second most important thing

DATA

Data

If you start thinking about modeling the problem, you should immediately check if the necessary data is potentially available.

If you ask, there are two possible answers:

a. We do not have the data

b. We have the data

Usually, both are wrong.

Experiment: Collecting Data at CO@Work-II

Combinatorial Optimization at Work II took place at ZIB from September 21 to October 9, 2009 with 105 participants from 23 countries.

We wanted to compute the seat allocation for the lecture hall.

To do this we required ever participant to state their preferences.

Everyone should send an email with a data file.

Lets see how long it took...

File Format

ASCII text with only a LF (ASCII 10) as line separator.

Fields are separated by a single space (ASCII 32)

Line 1: ParticipantNo HasLaptop EmailAddress

e.g. `67 1 koch@zib.de`

0 = has no Laptop, 1 = has a Laptop

Lines 2-???: SeatNumber PreferenceValue

- Seat numbers start down at the low entrance, left to right, row by row.
- The hightest numbered seat is at the window side at the top.
- Count only seats that are physically there.
- The seat numbers in the file should be monotonically increasing.
- The preference values should be between 0 and 100.

e.g.
```
12 55
13 40
14 35 …
```

Rules Regarding Preference Values

Allowed values are between 0 and 100

Only seats which are not available for the participants are allowed to get a value of 0

All numbers 1-100 have to be used at least once

The average has to be between 40-60

The difference to an adjacent seat has to be < 40

The difference to a neighboring seat has to be < 20

The data should not be randomly generated

Specifying Preference Offsets

Lines ???-???: ParticipantNo PreferenceOffset
List indicating persons which you would like or not like to be your seat neighbor.
(You have to know the ParticipantNo of the person.)

- A ParticipantNo of 0 indicates an empty seat.

- The PreferenceOffset is between -20 and 20 and will be added to your PreferenceValue if the person with the given ParticipantNo is your neighbor.
  e.g.
  ```
  55 17
  27 -5
  72 8
  0 -10 …
  ```

- This list can have as many entries as you like, but there should be at least 2 entries, and the occurring participant numbers have to be unique and valid.

How To Submit

Submission of this file is required for the course

The name of the file has to be *ParticipantNo* `.txt`

It should be **attached** to an email

Send the email to koch@zib.de

The subject of the email should be
CO@Work: SeatData for *ParticipantNo*

Please, as soon as possible.

2 Days after the lecture

Mails received : 13

Different Subjects : 4 (10 1 1 1)

Wrong field spacing : 4

Seat counts : 2 (12 1)

Missing data : 1

Too much data : 1

Ok, from first view : 5 out of 13

3 Days after the lecture

Mails received : 23

Different Subjects : 6 (17 2 1 1 1 1)

Wrong field spacing : 4

Seat counts : 4 (19 1 1)

Missing data : 2

Too much data : 0

Ok, from first view : 10

Corrected : 1

Add to the specification:

A seat without a desk is not allowed for the participants
Seats with a 0 preference value are not relevant for the adjacency/ neighboring difference rules.

4 Days after the lecture

Mails received : 37

Wrong subject : 11

Wrong field spacing : 8

Strange seat counts : 5

Missing data : 2

Corrected : 3

5 Days after the lecture

Mails received	: 47
Data sets	: 41 (6 corrections)
Wrong subject	: 12
Wrong attachment name	: 2
Wrong line separator	: 29
Wrong field separator	: 10
Pref. value not used	: 11
Other Errors	: 1
Number of seats	: 153 - 181
No complains so far	: 4

7 Days after the lecture

Mails received	: 79
Data sets	: 64
Wrong subject	: 16
Wrong attachment name	: 2
Wrong line separator	: 45
Wrong field separator	: 11
Pref.value not used	: 22
Other Errors	: 2
Number of seats	: 153 - 181
No complains so far	: 8

9 Days after the lecture

Mails received	: 104
Data sets	: 76
Wrong subject	: 18
Wrong attachment name	: 2
Pref. value not used	: 19
Neighbor difference	: 21
Wrong no/seq. seats:	: 10
Wrong 0 seats	: 20
No complains so far	: 10

Overview of Errors in Data

	E7	E10	E11	E12	E13	E14	E16
5							X
6							X
12					X		X
13							X
16							X
18					X		X
19						X	X
20						X	
23					X		
24						X	
26							X
27						X	
36						X	
42					X		
45			X	X	X	X	X
47					X		
53					X		
59						X	
63			X		X	X	
64			X		X	X	X
71					X	X	

E7 bad seatno

E10 bad offset

E11 wrong seatno

E12 bad average

E13 prefval missing

E14 neigbour diff

E16 seat not 0

	E7	E10	E11	E12	E13	E14	E16
77							X
78	X		X			X	X
81				X	X		X
98					X		
99	X		X			X	
103					X	X	
107			X			X	X
108			X			X	X
111							X
121							X
128			X			X	X
129		X					
134			X	X	X	X	
135					X		
137		X			X	X	X
139					X		X
145	X		X		X	X	
160						X	
166					X	X	

Please correct and resubmit

11 Days after the lecture

Mails received	: 144
Wrong subject	: ~23
Wrong attachment name	: 4
Data sets	: 92
To be corrected	: 28
Missing	: 6
Pref. value not used	: 14
Neighbor difference	: 18
Wrong no/seq. seats	: 2

Overview of Errors in Data

	E7	E10	E11	E12	E13	E14
12					X	
18					X	
23	X		X			X
24						X
27						X
45					X	X
47					X	
63			X		X	X
71					X	X
78	X		X			X
79		X	X	X	X	
103						X
107			X			X
108			X			X
110		X				
114						X
118					X	X
128			X			X
134			X	X	X	X
135					X	
136						X
137		X			X	X
138					X	
139					X	
160						X
166					X	X

E7 bad seatno

E10 bad offset

E11 wrong seatno

E12 bad average

E13 prefval missing

E14 neigbour diff

Please correct and resubmit

13 Days after the lecture

Mails received	: 159
Wrong subject	: ~26
Wrong attachment name	: 4
Data sets	: 94
To be corrected	: 18
Missing	: 4
Preference value not used	: 9
Neighbor difference	: 14
Wrong no/sequence seats	: 3

Overview of Errors in Data

	E7	E10	E11	E12	E13	E14
18					X	
24						X
27						X
45					X	X
63					X	
71					X	X
78	X		X			X
79		X	X	X	X	
103						X
107			X			X
108			X			X
114						X
118					X	X
128			X			X
134			X	X	X	X
136						X
137		X			X	X
138					X	

E7 bad seatno

E10 bad offset

E11 wrong seatno

E12 bad average

E13 prefval missing

E14 neigbour diff

Please correct and resubmit

14 Days after the lecture

Mails received : 166
Wrong subject : ~28
Wrong attachment name : 4

Data sets : 95
To be corrected : 18
Missing : 3

Preference value not used : 7
Neighbor difference : 14
Wrong no/sequence seats : 3

Overview of Errors in Data

	E7	E10	E11	E12	E13	E14
24						X
27						X
45					X	X
71					X	X
78	X		X			X
79		X	X	X	X	
92					X	X
107			X			X
108			X			X
114						X
118					X	X
128			X			X
134			X	X	X	X
136						X
137		X			X	X

E7 bad seatno

E10 bad offset

E11 wrong seatno

E12 bad average

E13 prefval missing

E14 neigbour diff

Please correct and resubmit

15 Days after the lecture – the final day

Mails received : 172

Wrong subject : ~31

Wrong attachment name : 4

Data sets : 95

To be corrected : 13

Preference value not used : 5

Neighbor difference : 13

Wrong no/sequence seats : 2

Subject Variations

The subject of the email should be
CO@Work: SeatData for *ParticipantNo*

```
CO@Work: SeatData for 022
CO@Work:SeatData for 222
CO@Work:SeatDatafor222
CO@work: SeatData for 222
CO@Work: Seat Data for 222
Co@Work: SeatData for 222
CO@Work: SeatData for Participant222
CO@Work: SeatData for ParticipantNo
Co@Work: SeatData for Participan222
CO@WORK:  seatdata for 222
COatWork: SeatData for 222
COatWork for 222
SeatData for 222
SeatData for ParticipantNo 222
set data for participant number 222
data set participant number 222
Sitting assignment
Seats assignment
```

Overview of Errors in Data

	E7	E10	E11	E12	E13	E14
24						X
27						X
45					X	X
71					X	X
78	X		X			X
92					X	X
107			X			X
108			X			X
114						X
128			X			X
134			X	X	X	X
136						X
137		X			X	X

E7 bad seatno

E10 bad offset

E11 wrong seatno

E12 bad average

E13 prefval missing

E14 neigbour diff

Sorry,
too late to correct!

Wrong line 1: 81, 129

Final Day CO@Work 2015

- Received 73 Emails from 58 participants
- 11 wrong subjects
- 5 without an attachment
- Subjects

Subjects (names changed)

CO@Work: Data for MüllerMax

List Smith

Peter Paul Panther

Experiment List

Data for MustermannKarin

Data Experiment

CO@WORK: Daten

CO@Work: Data Experiment

CO@WORK: Data for TigerTheobald

Re: CO@Work Data for SmithPeter

Tr : CO@Work: Data for MuellerMaria

CO@work: data for KruegerFreddy

CO@Work: Data for LastnameFirstname

CO@Work: Data for Wrobel Ignaz

Final Day

- Received 77 Emails from 60 participants
- 11 wrong subjects
- 5 without an attachment
- 13 different wrong subjects
- Number of participants according to submissions
  between 1 and 106
- At least 6 attachments contained non ASCII characters,
  in at least 3 different encodings
- 2 attachment were not a text-file at all (.odt, .rtf)
- 2+ attachment had wrong extensions (.csv, .tex, .txt.txt)
- Several attachments had bad Science and Preference Indicators
- several attachments had people that did not show up in the room
- The contents of several attachments were ill formed

Data

You would think a …

- … cellular network operator knows where its base stations are located?
- … fixed network operator can tell where the parts of its network are connected?
- … chemical company knows how many plants they have?
- … 5 m long pipeline cannot have a height difference from end-to-end of 100 m?

- Many companies have their data in Excel.
  There is no formal validation or referential integrality check.
- If they did formal validation, usually they found there was information they needed which they could not input and they started to “reuse” some data fields.
- If there is not at least 1 error per 100 data sets you are not looking hard enough.
- Usually the data changes all the time.
- They might not want to give it to you.
- The data might just not exist.

The first result of an optimization project is usually to improve the quality of planning data available at the company.

Back to the past

The data might just not exist.

We all know that centrally planned economies did not do too well.

One of the (major) reasons was that the assumptions, i.e. the data they used to make their plans where faulty.

Now the big companies just try the same ☺

The first result of an optimization project is usually to improve the quality of planning data available at the company.

Preprocessing / Heuristics

- More often than not real-world problems are rather big.
- At least compared to “academic toy examples”
- In all projects I have done so far, the key to success was proper preprocessing of the data. Since it is real-world, many things will be obvious to decide. Throw them away before they make trouble.

- Very often it is quite helpful to have (quick) heuristics which find feasible solutions (fast).

- Industry is usually more interested in reasonable good solutions in short than in proven optimal solutions after a long wait.

Advertisement Break

Mathematical Programming Computation

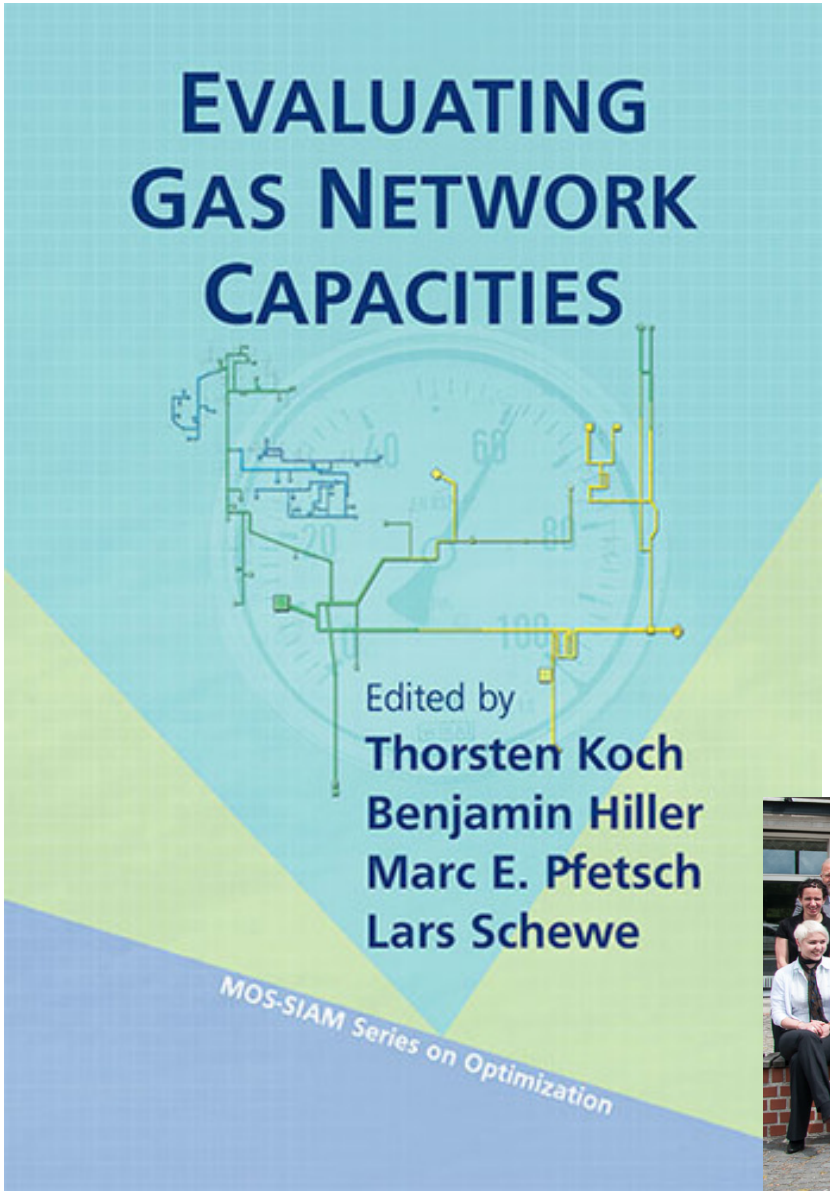
MPC will publish original research articles covering computational issues in mathematical programming.

Articles report on innovative software, comparative tests, modeling environments, libraries of data, and/or applications.

A main feature of the journal is the inclusion of accompanying software and data with submitted manuscripts. The journal's review process includes the evaluation and testing of the accompanying software. Where possible, the review will aim for verification of reported computational results.

Visit: http://mpc.zib.de and http://springer.com

About a large scale industrial optimization project

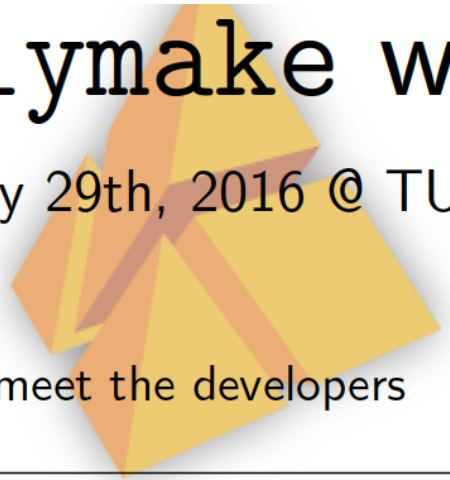


The *Research Cooperation Network Optimization* ran 6 years, involving more than 30 people from 7 research institutes and Germanys largest gas network system operator.

Here are the results.

7th polymake workshop

January 29th, 2016 @ TU Berlin



- talks
- demos
- meet the developers
- helpdesk
- tutorials

---

polymake is a software package to study:

- combinatorics and geometry of convex polytopes
- linear programs
- toric geometry
- tropical geometry

. . . and much more

---

possible topics:

beginners session, tropical geometry, . . . (suggestions welcome)

http://polymake.org

Welcome to ICMS 2016 in Berlin

ICMS 2016 is a satellite to the 7ECM Berlin

The conference will be held at Zuse Institute Berlin (ZIB) from July 11 to July 14, 2016

Chairs

- General Chair: Gert-Martin Greuel, University of Kaiserslautern
- Program Chairs:
  - Peter Paule, Johannes Kepler University Linz and RISC
  - Andrew Sommese, University of Notre Dame
- Local Chair: Thorsten Koch, Zuse Institute Berlin

Solutions

Optimality is usually not required. Industry is usually more interested in reasonable good solutions in short time than in proven optimal solutions after a long wait.

Extremal Solutions vs. 80%

It is very important that the solution is 2-optimal.

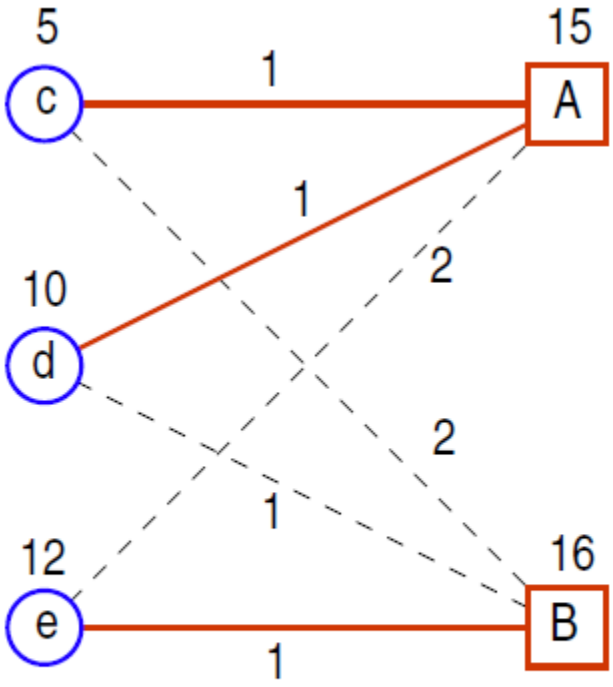
To compute a gap or prove infeasibility is important.

What does it mean? You are proveably global optimal but the former solution the company used is better.

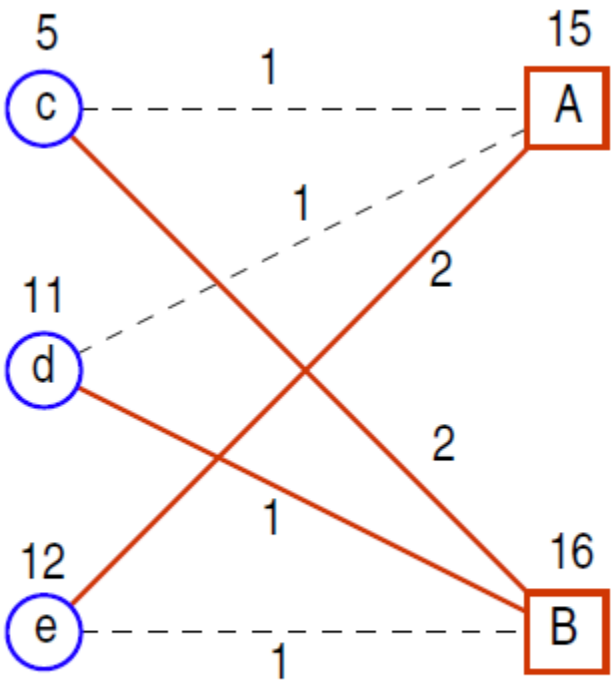
Strange solutions may have several reasons:

- Anomalies (errors) in the data
- Differences between model and reality
- Reaching of capacity or cost thresholds
- The result is just different than expected

Instable solutions



(a) Optimal solution



(b) After a small change

Strange cost functions

Cost per channel	Distance in km <45	<90	⩾90
uv to vv	3.30	4.95	6.60
vv to hv	3.50	5.25	7.00

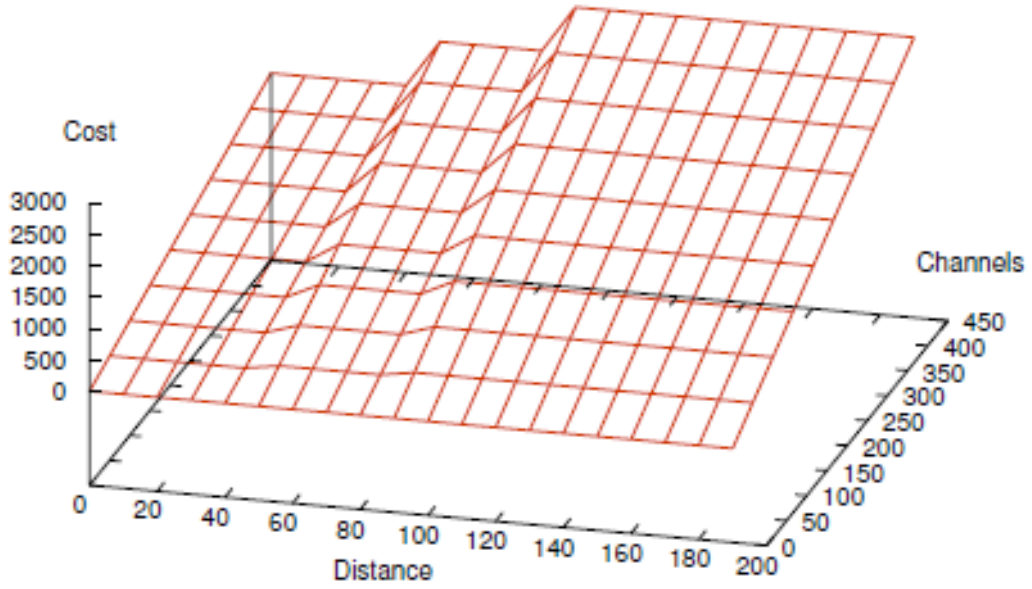
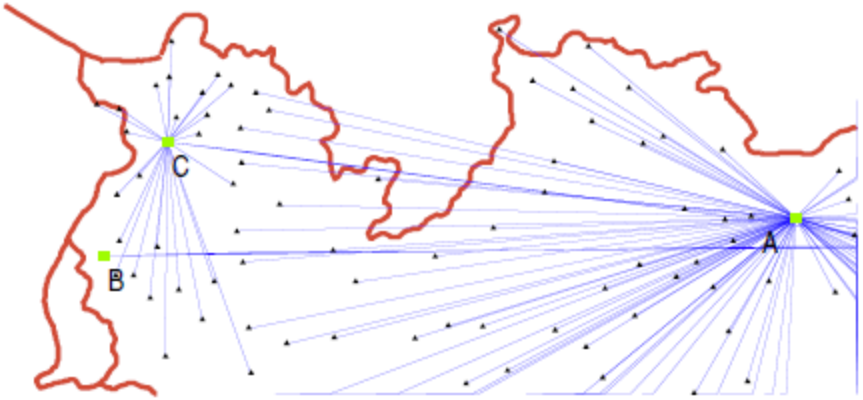


Figure 4.9: Cost function depending on distance and channels

Strange solutions



(a) Bee-line distances

(b) Transport network distances

- Why is no UV connected to B?

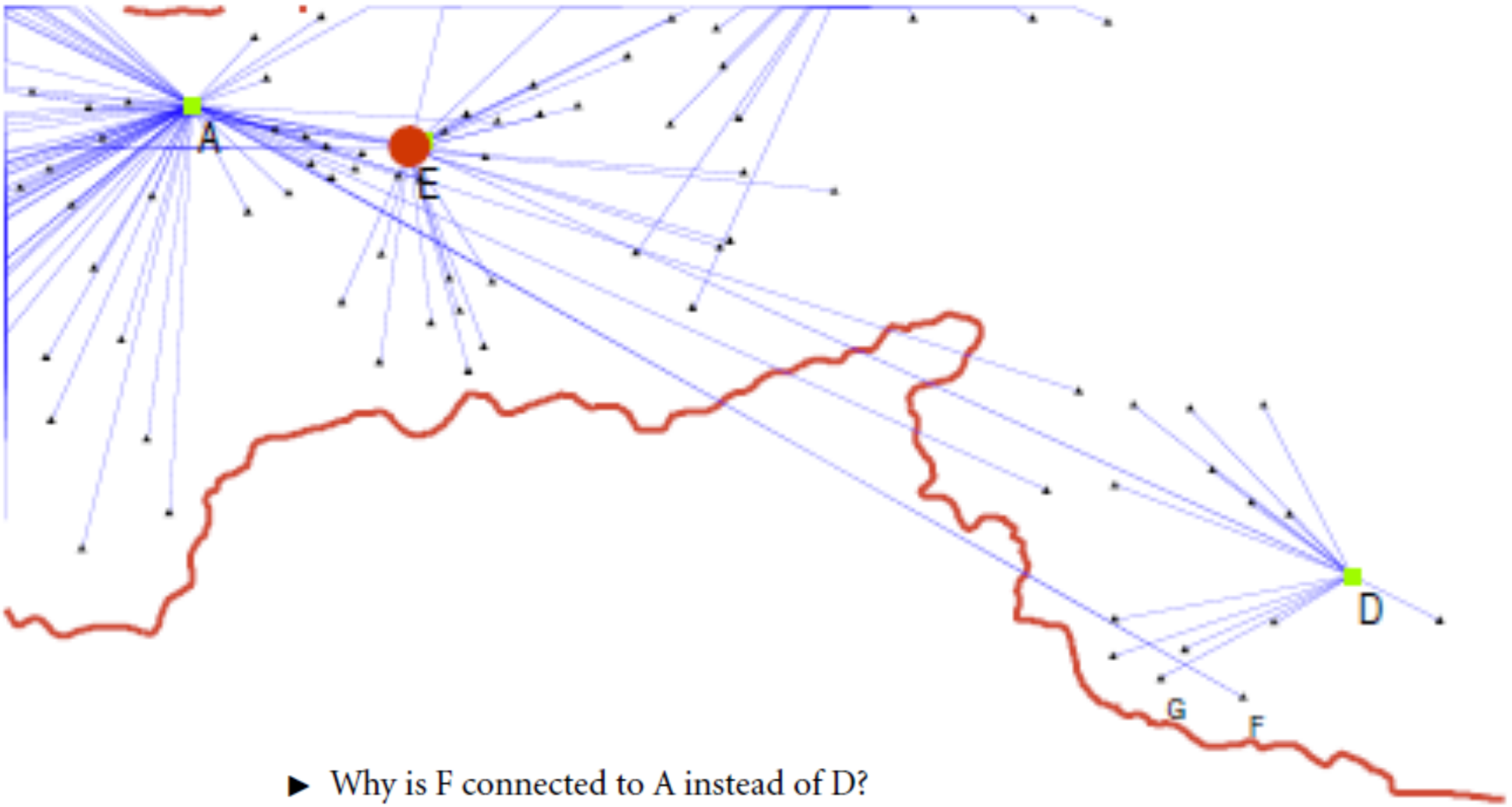
  Since all Uvs in question are less than 45km away from B and C, the connection costs are equal. Since C has enough capacity all the UV just happen to be connected to it.

- Why then are B and C both active?
  Because the input data requires it.

- Why are some UVs in the vicinity of C connected to A?

  Because connecting them to C would increase the total length of the link from the UV to the HV. VV to HV connections are only a little cheaper than UV to VV links. So the cost for first connecting to the more remote C does not pay of. This changes if instead of bee-line distances transport network distances are used.

Strange solutions



- Why is F connected to A instead of D?
- Would not E be better than A to connect F?
- Why is G connected to D, if F is not?

Checking Data and infeasibility

The data you will get will contain errors.

- Make it as easy as possible for everbody to correct it
- What do you do in case of detecting infeasibility of your model due to the data given? Providing an *Irreducible Infeasible Subsystem* (IIS) is not going to help much.

Run of the project

- The bigger the company, the more unstable is the department.
- If the project takes too long they may lose interest.
- This is research not development.
- How to report „progress“?
- Milestones in the beginning are hard to meet for research projects.
- The original solutions of the company are often infeasible.
- Check the solutions, we do errors, too.

Things to do (short list)

1. Understand problem
2. Find/decide on underlying mathematics
3. Chose a suitable frame work

Make problem specific

- I/O
- Preprocessing
- Primal heuristics
- possibly specific cuts, constraint handler

The way to real-world applications

	Problem definition	Real world constraints	Data	Code
Pure research	None	None	None	None
Applicable research	General	Unknown	Random/Simplified	Whatever
Applied research	General	Maybe	Random/Simplified	Whatever
Case study	Simplified	Some	Simplified	Whatever
Planning application	Simplified	Some more	Simplified/Real	Production
Control application	Complete	all	Real	24/7

The gap

- Research means it is unclear whether and how something it is possible.
- Once research has established a possible path into practice, implementing this path is called development.
- Between finding that something is in principle possible and having a clear path to actually do so on an industrial scale, there is a gap.
- Sometimes many things have to be combined. While for each part it is clear that it should be feasible, it is unclear whether it is possible to glue them successfully together.

In theory there is no difference between theory and practice.
In practice there is.

(In academia, there is no difference between industrial work and academic work. In industry, there is.)

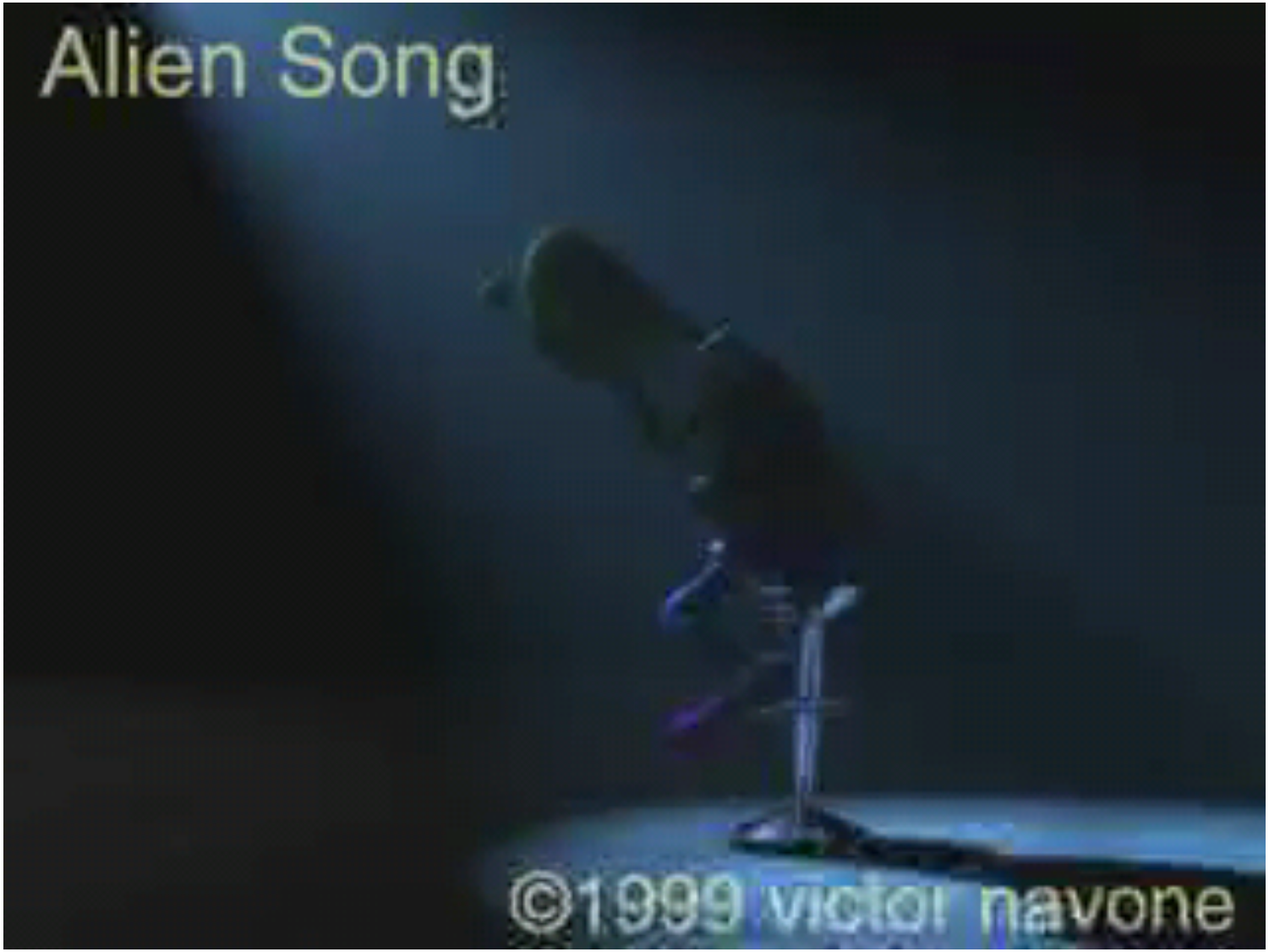
Why isn‘t it considered innovative if a solution works in industrial practise?
Guido Sands, ABB

Why do real-world projects at all?

The final test of a theory is its capacity to solve the problems which originated it.

George Dantzig (1963) in

Linear Programming and Extensions



https://www.youtube.com/watch?v=4JTLyxuylac

Thank you very much!

Questions?