Estimering av IT-prosjekter:
Hva vet vi? Hvordan bli bedre?

Magne Jørgensen
Simula Research Laboratory
www.simula.no

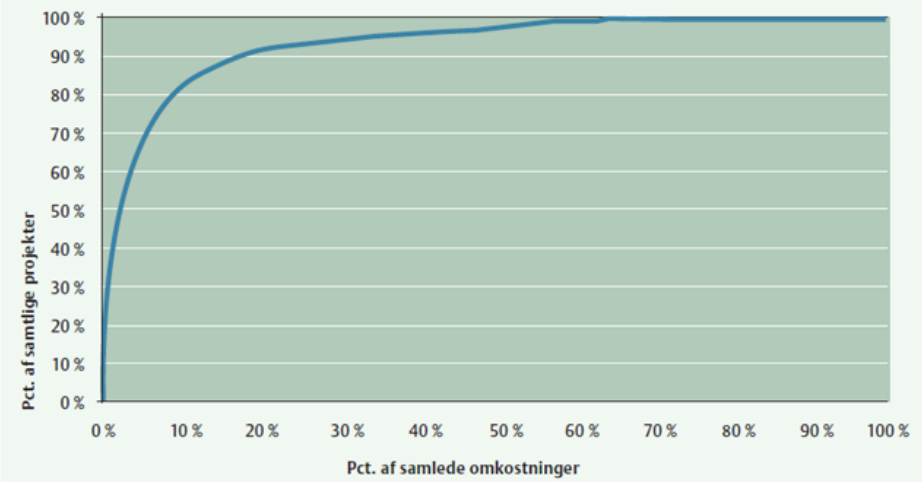
Hvor gode er vi til å estimere?

- Stort sett så er vi ikke så aller verst.
- IKKE tro på rapporter fra Standish Group og andre som ikke har et bevisst forhold til seleksjon av prosjekter.
- Trolig er gjennomsnitlig overskridelse på 20-30%.
- Men, det er noen få store prosjekter som trekker veldig opp dersom vi måler i absolutt overskridelse (se figur).
- Det er forskjell på de “skumle” og de “normale” prosjektene og vi er ikke veldig gode i å anslå usikkerhet.

FINANSMINISTERIET

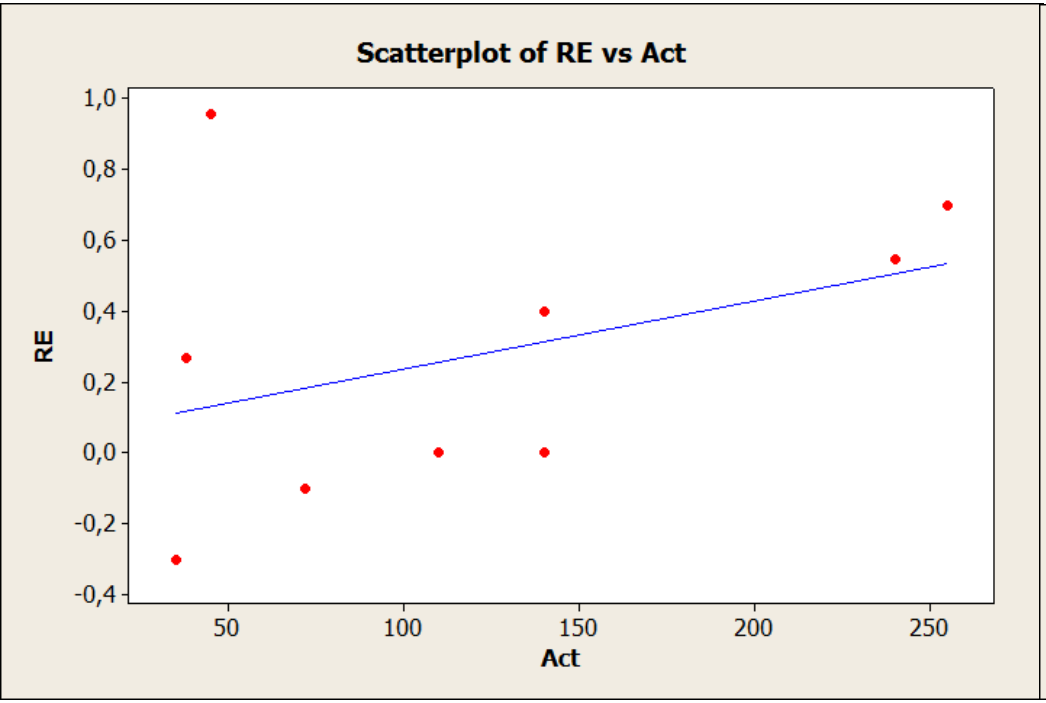
Lille gruppe projekter driver store dele af samlede omkostninger i staten

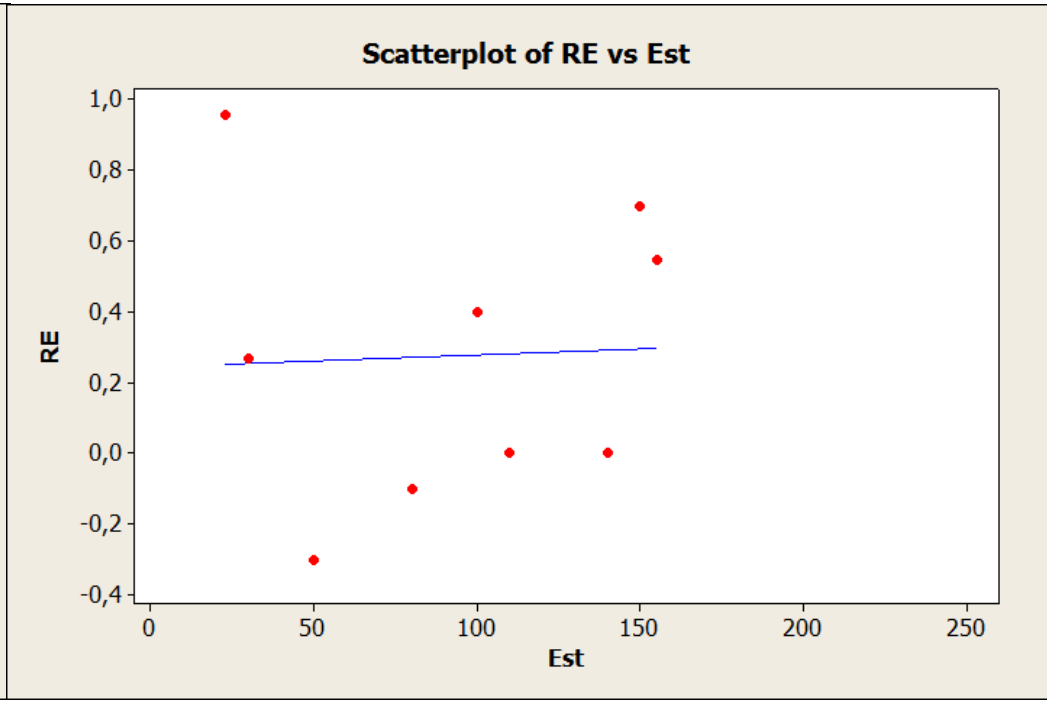
Figur 2.3 Sammenhæng mellem de kumulerede omkostninger og antallet af it-projekter, n=197



Er store prosjekter vanskeligere å estimere? (Data: Danske problemprosjekter)

RE = (Act – Est)/Est

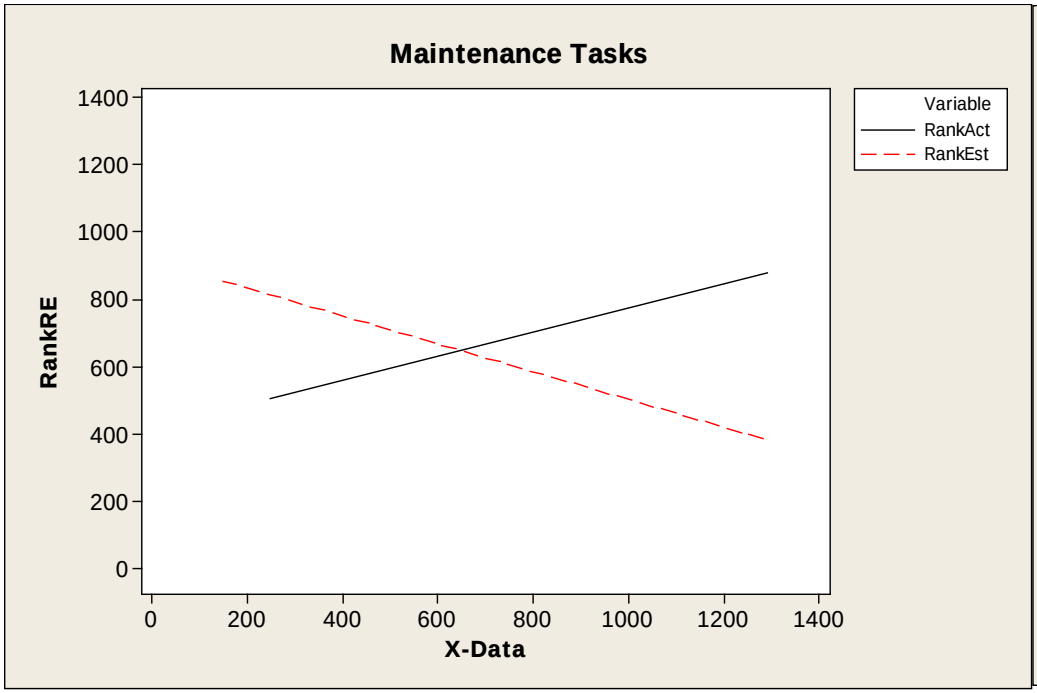


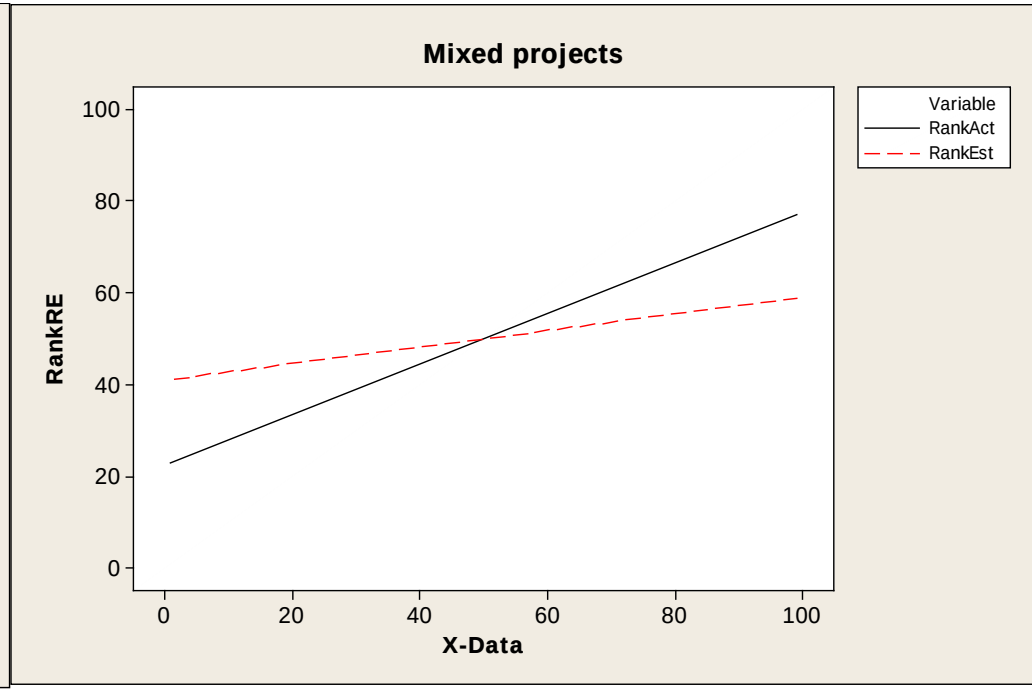


Høy korr. mellom faktisk arbeidsmengde og overskridelse

Lav korr. mellom estimert arbeidsmengde og overskridelse

Ingen klar sammenheng mellom størrelse og overskridelse





En manglende sammenheng tyder på at overskridelser skyldes heller mange små, enn et stort problem. Dette støttes av flere rapporter, f eks, rapport fra NHS (UK) som hevder at “The devil is in the details”

Indikatorer på skumle prosjekter

- Iboende faktorer:
  - Vi gjør ting som er grunnleggende forskjellig fra det vi har gjort før
  - Mange grensesnitt til andre systemer og/eller mange interessenter
  - Involverer stor grad en prosessendring
  - Problemene som skal løses er komplekse
  - Uflaks (mange små eller få store ”uflakser”)
- Faktorer vi kan gjøre noe med
  - Realisme i ambisjoner
  - Situasjoner og motivasjoner som gir over-optimisme
  - Kompetanse til kunde og leverandør
  - Oppfølging, støtte, og ledelsesforankring
  - Kommunikasjon, inkludert økt oppmerksomhet rundt kulturelle forhold
  - Anbudsprosesser (unngå ”winner’s curse” og ”adverse selection”)
  - Egnede utviklingsmodeller (særlig relatert til oppdeling av leveranse)

Noen funn i ESSU (UK):

- Many projects are over-ambitious.
- Long-term or delayed projects are often overtaken by new technology, changes in legislation and public policy.
- The private sector often overstates its ability to deliver.
- Clients are often under-resourced and/or do not have the required skills.
- The procurement process is a high-risk strategy, heavily influenced by market forces in respect to who bids, the level of competition and private sector strategies to increase market share.
- Some projects are driven by the application of the latest information and communications technology to meet ‘customer demand’ for seamless one stop contact centres combined with pressure to achieve substantial savings. However, a more incremental approach may be more desirable, effective and economical.
- Off-the-shelf is no guarantee for success. Many projects were based on off-the-shelf products, but nevertheless failed.

http://www.european-services-strategy.org.uk/news/2007/ict-contract-chaos/105-ict-contracts.pdf

Hva kommer overoptimismen av?

Overvurdering av egne evner er normalt hos mentalt friske mennesker!

Undervurdering av risiko er normalt

(Risiko for egen skilsmisse er for eksempel sterkt undervurdert selv om de fleste vet at nær 50% av ekteskap sprekker)

Biologi

- Evolusjonen synes å ha belønnet selvsikkerhet og risikoundervurdering.

- IT-ledere belønner selvsikkerhet og bruker den som indikator på dyktighet.

- En del indikasjoner på sammenheng mellom optimisme og biologi. F eks, så hjelper trolig optimisme på en del sykdomsforløp (styrking av immunforsvar), og håndtering av problematiske situasjoner (”coping”).

- Simulering indikerer at overoptimisme i noen sammenhenger gir systematisk bedre utfall enn realisme. Dette gjelder særlig når man vet lite om sannsynligheten for ulike utfall, men mer om konsekvensene av utfallene.

Grupper og optimisme

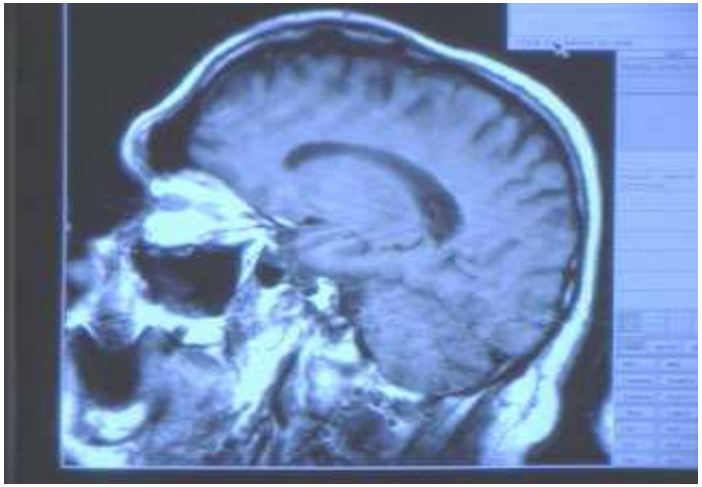
- Mange studier viser en økt risikovilje og over-sikkerhet (over-confidence), kanskje pga individenes reduserte ansvar for beslutningene.
- Våre studier på IT-prosjekter indikerer imidlertid mer realisme ved bruk av grupper til å estimere!
  - Flere hoder husker på mer (union av aktiviteter)
  - Krav om begrunnelse kan gi økt realisme
  - Noen ganger forekommer imidlertid ”group think” og dominerende individer
    - F eks at **alle** i gruppen går ut av en estimeringsdiskusjon med opplevelsen av at estimatet ble for lavt.
- Effekten av grupper på realisme trolig svært kontekstavhengig, dvs vanskelig å finne generelle mønstre.
- Bruk av Delphi-baserte metoder (strukturerte grupper med uavhengige vurderinger) gir stort sett svært bra innvirkning på realisme.
  - Som for eksempel ”Planning Poker” eller Wideband Delphi

Motivasjon

- Mange resultater viser sterk sammenheng mellom motivasjon for høy ytelse (ønsketenkning) og overoptimisme.
- Optimisme kan påvirke ytelsen positivt, MEN
  - Optimistiske estimater synes typisk å ha kun en kortvarig positiv effekt på gjennomføringen.
  - Hvor stor og hvor positiv effekten av optimisme er på ytelse overvurderes oftest sterkt.

Kognitive prosesser

- Planlegging (visualisering av fremtidsscenarier) gir ofte mindre grad av realisme enn historierefleksjoner (hvordan det har gått)!
  - Analogi-basert estimering gir ofte gode resultater
- Overvurdering av hvor mye vi kan kontrollere/håndtere er en viktig kilde til overoptimisme (the illusion of control)
  - Dyktighet vs uflaks er ofte et spørsmål om man lykkes eller mislykkes



Kunder velger over-optimistiske leverandører

- ”Winner’s curse”
  - Vektlegging av pris + mange tilbydere: ”*Vi vinner nesten bare når vi har vært over-optimististiske*”.
  - Kan gjøre gode leverandører dårlige
- Manipulerende informasjon
  - Budsjettinformasjon (Lånekassens IT-system hevdes å ha blitt estimert av stortingspolitikerne, som bestemte budsjettet).
  - Bruk av ladede ord (”lite system”)
- ”Adverse selection”
  - Jo mindre kompetanse hos leverandør, jo lavere estimat og pris
  - Jo mindre kompetanse hos kunde, jo mer vektlegging av lav pris
  - Kunde velger mindre kompetent og over-optimistisk leverandør

Ankereffekter

Eksperiment:

- HIGH (LOW) group: *"The customer has indicated that he believes that **1000 (50)** work-hours is a reasonable effort estimate for the specified system. However, the customer knows very little about the implications of his specification on the development effort and you shall not let the customer's expectations impact your estimate. Your task is to provide a realistic effort estimate of a system that meets the requirements specification and has a sufficient quality."*
- Deltakere: Erfarne systemutviklere.
- Alle (HIGH, LOW, CONTROL) fikk samme kravspesifikasjon.

Ankereffekter

- Resultater:
  - HIGH group gjennomsnitt: 555 timeverk
  - CONTROL group (uten forventinger) gjennomsnitt: 456 timeverk
  - LOW group gjennomsnitt: 99 timeverk!!!
- Ingen av utviklerne opplevde at de hadde blitt mye påvirket. De fleste mente at de ikke var påvirket i det hele tatt av kundens forventninger.

Feltstudie av det samme ….

"The preliminary budget of the new system is $10 000 [corresponding to about 100 work-hours with typical pricing in the country in which it will be built]. The preliminary budget is not built on any knowledge about the actual cost of developing the new system, and will, if needed, be extended to cover the expenses necessary to build a quality system with the desired functionality."

100 timeverk var en svært lav verdi for dette prosjektet og alle firmaene ble instruert om å IKKE bruke dette som input til estimatet.

Firmaene som deltok visste ikke at de var med i en studie, men skulle gi uavhengige estimater på et arbeid som skulle utføres **av andre (mao ingen direkte grunn til å være optimistiske)**. Firmaene ble betalt for estimeringsarbeidet.

Resultater fra feltstudien

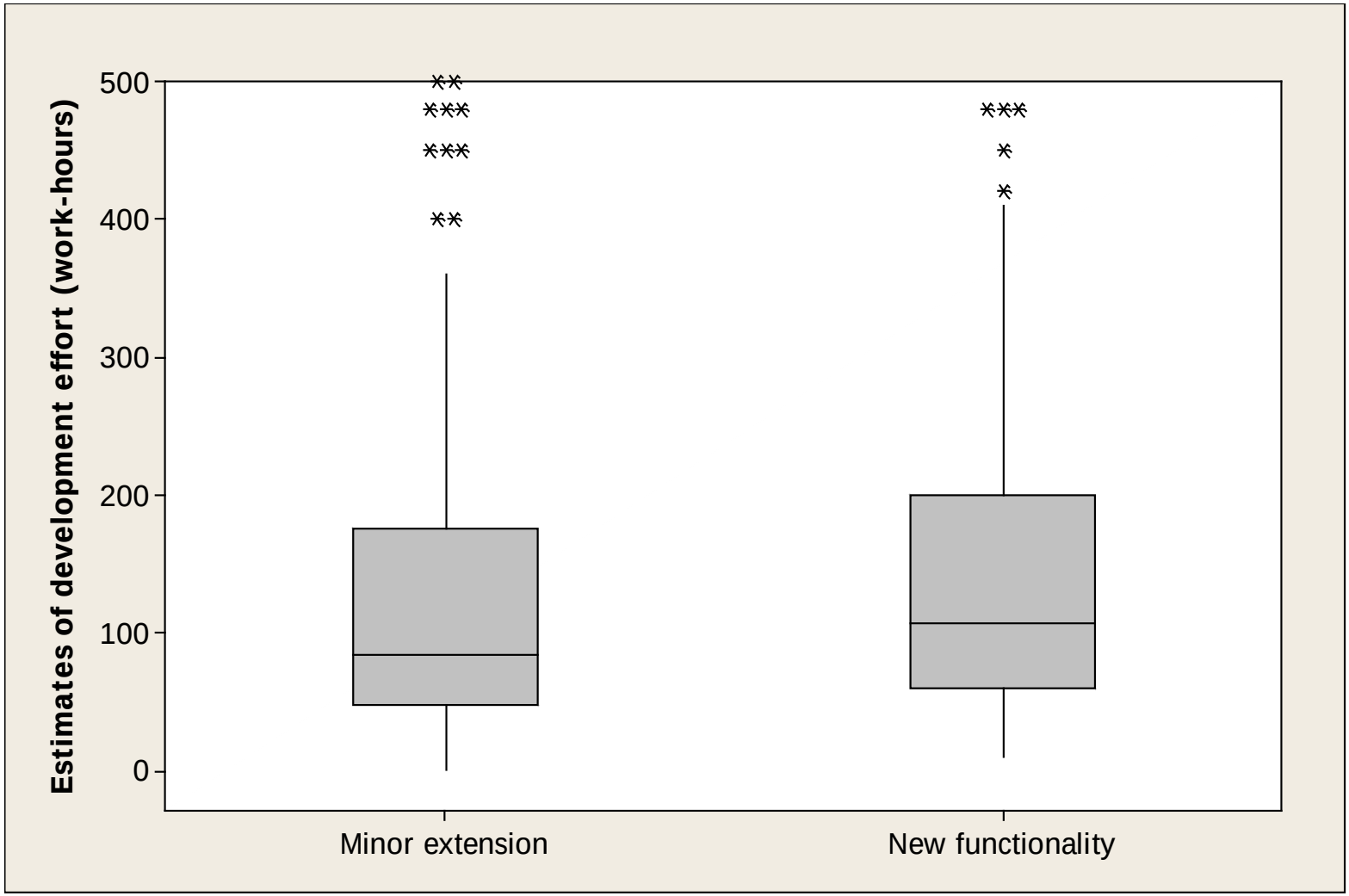
Numerical Anchor

Group	**Median estimate**
Manipulated (client's expectation, 100 work-hours)	724 work-hours (n=23)
Ordinary	956 work-hours (n=23)

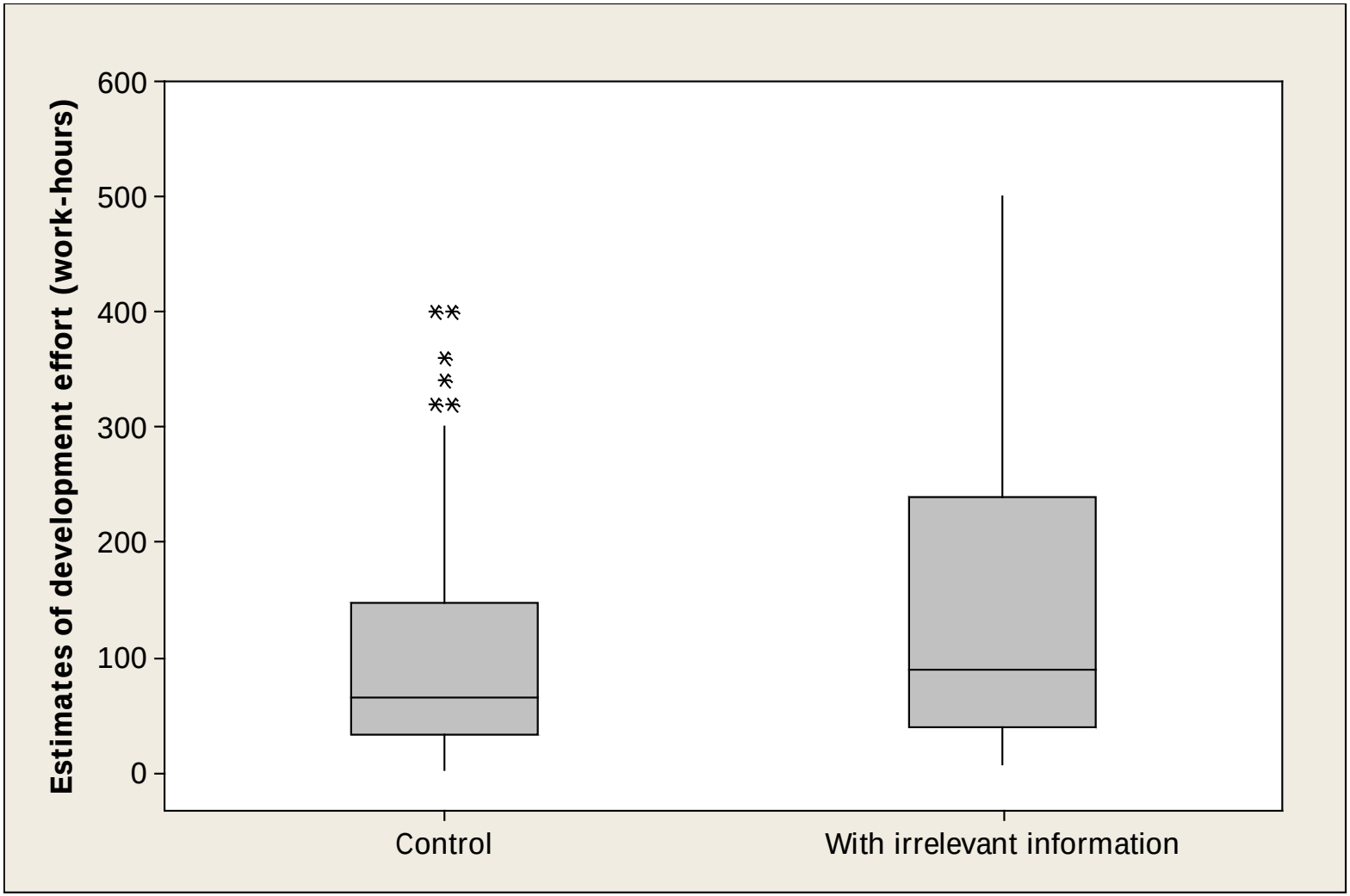
Mindre effekt enn i "laboratoriet", men fortsatt en effekt av betydning.

Trolig større effekt dersom man er i tilbuds-modus.

Ladede ord kan også fungere som anker



Irrelevant informasjon påvirker estimatet

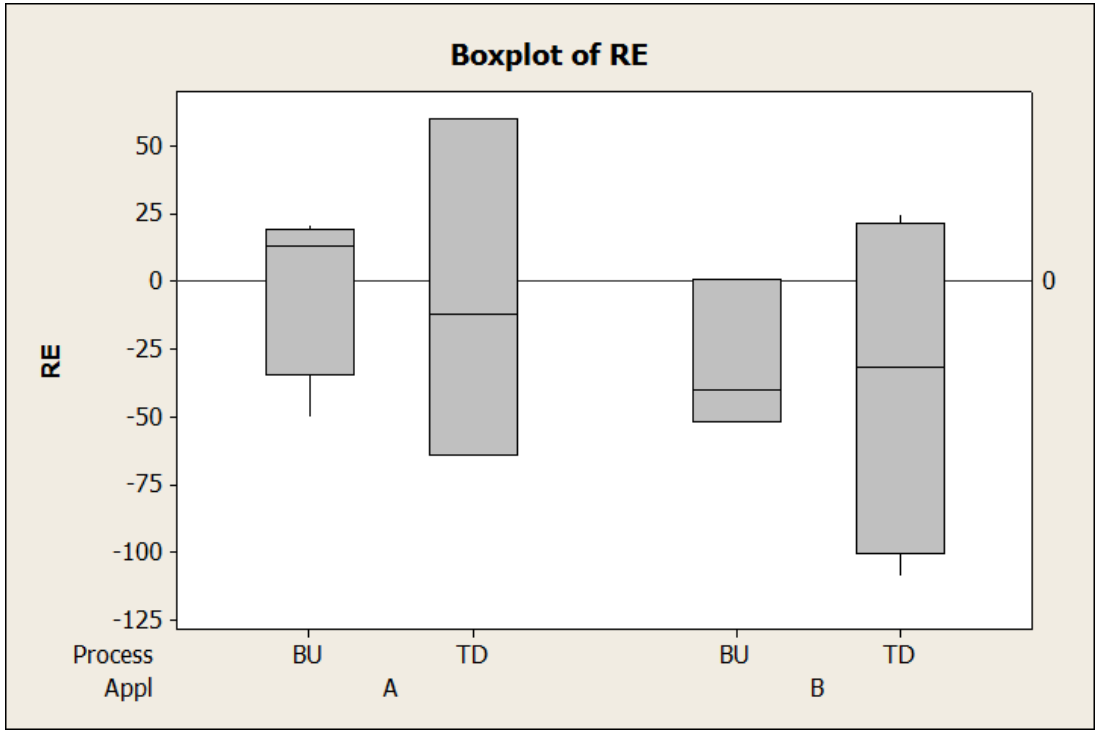


Usikkerhetsvurderinger

- Minimum-maksimum intervall: “Nesten helt sikker” = 60-70% sikkert.
- Konfidensnivå betyr minimalt for bredden på intervallene.
  - Eksperiment med fire grupper av utviklere (A: 50%, B: 75%, C: 90, D: 99% konfidensnivå)
  - Nesten samme minimum-maksimum intervaller ble angitt.
- Meningsløst å be om minimum-maksimum intervaller.
- Be i stedet om:
  - Angi en fordeling av estimeringsfeil for lignende prosjekter.
  - Bruk denne til å anslå usikkerhet.
    - Eksempel: Dersom 80% av tidligere prosjekter har overskredet estimatet med mindre enn 50%, så kan du være 80% sikker på at et budsjett som er lik 1,5*estimatet ikke vil bli overskredet.
  - Hovedfordel er fornuftig bruk av historiske data og unngåelse av ønsketenkning. Evaluert i reelle prosjekter med god virkning.

Top-down eller bottom-up

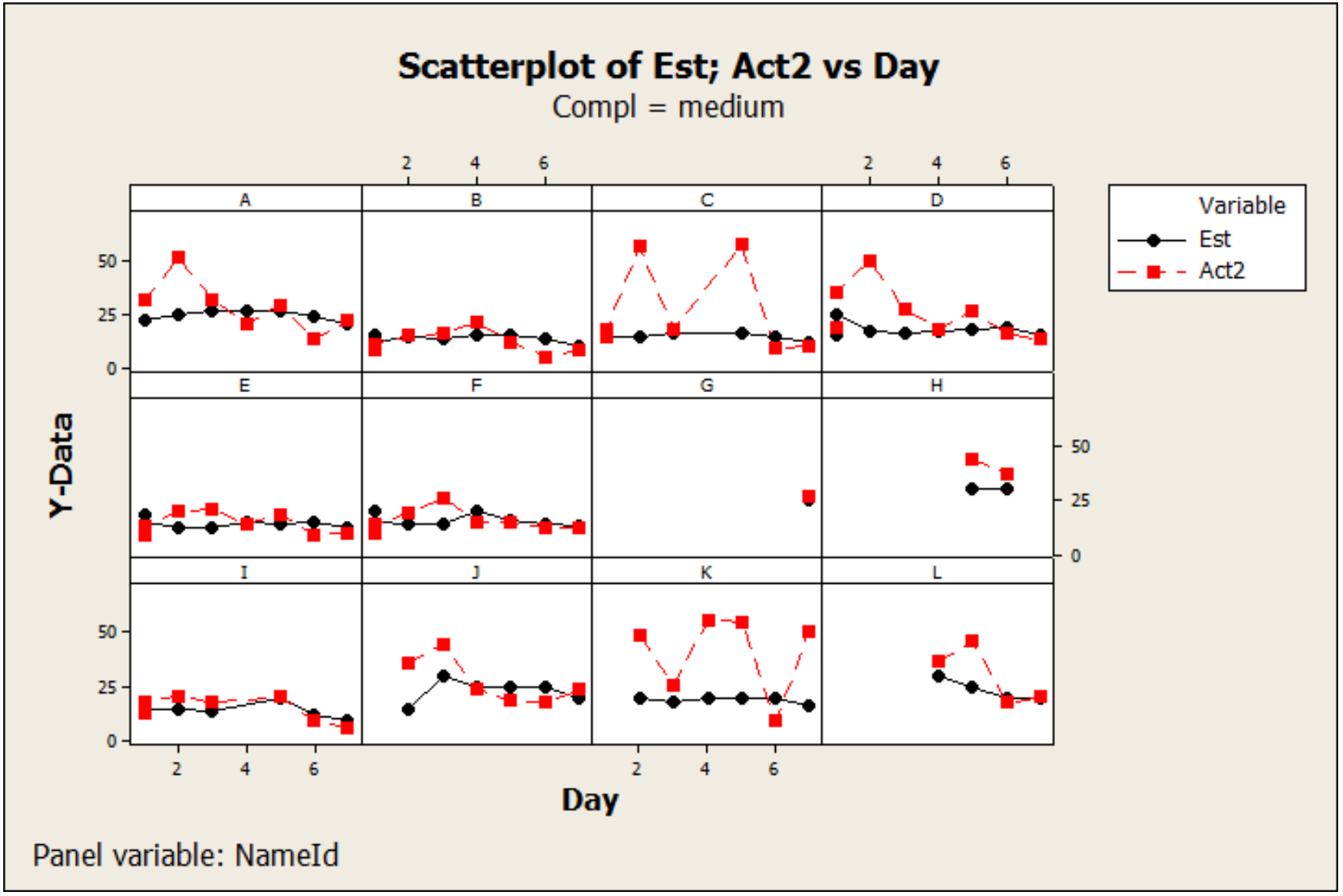
- Ja takk, begge deler.
- Studie viste at dersom gode analogier ble funnet, så var top-down den beste.
- I gjennomsnitt var imidlertid bottom-up best.
- Overraskende funn: Estimeringsteamene var relativt dårlige i å finne og bruke lignende prosjekter.
- Kan være stort potensiale for opplæring og bruk av relativ estimering på prosjektnivå – a la relativ estimering på user story nivå i agile.



Hva får vi når vi ber om et estimat?

- Planlagt arbeidsmengde (med risikobuffer)?
- Median arbeidsmengde (50% sannsynlighet for å overskride)?
- Mest sannsynlig arbeidsmengde?
- Mest sannsynlig arbeidsmengde under ideelle forutsetninger?

Estimat synes typisk å gjenspeile "Hvis ingen ting går galt"



Flere studier viser at …

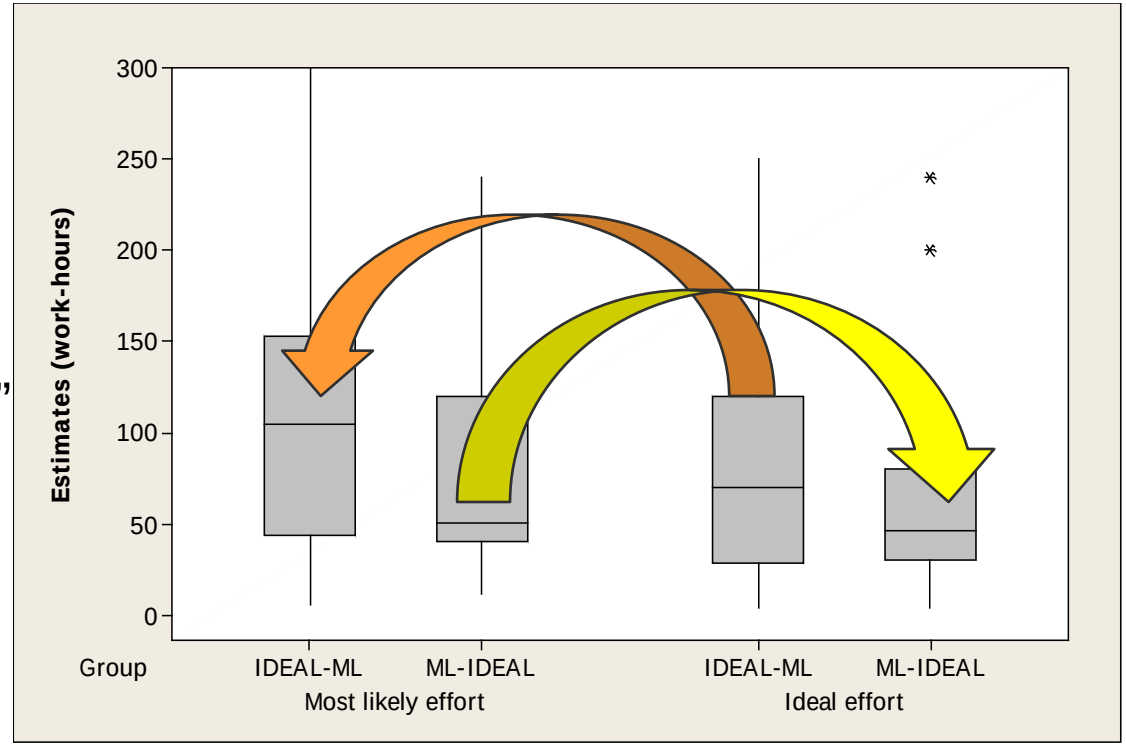
- Det er ofte liten eller ingen forskjell i prediksjoner gitt under idelle og realistiske scenarier.
- Dette har blant annet vært undersøkt innen:
  - Tidsestimering
  - Blodgiving
  - Treningsmengde
  - Sparing

Feil spørsmål, eller feil svar?

- Hva gjør man når respondentene ikke svarer på det man spør om, men på noe annet?
  - Endrer respondentenes oppførsel gjennom bedre opplæring, bedre instruksjoner, etc?
  - Endrer spørsmålet til å spørre om det som respondentene uansett gir svar på?
- Vi har evaluert en metode basert på det siste alternativet.

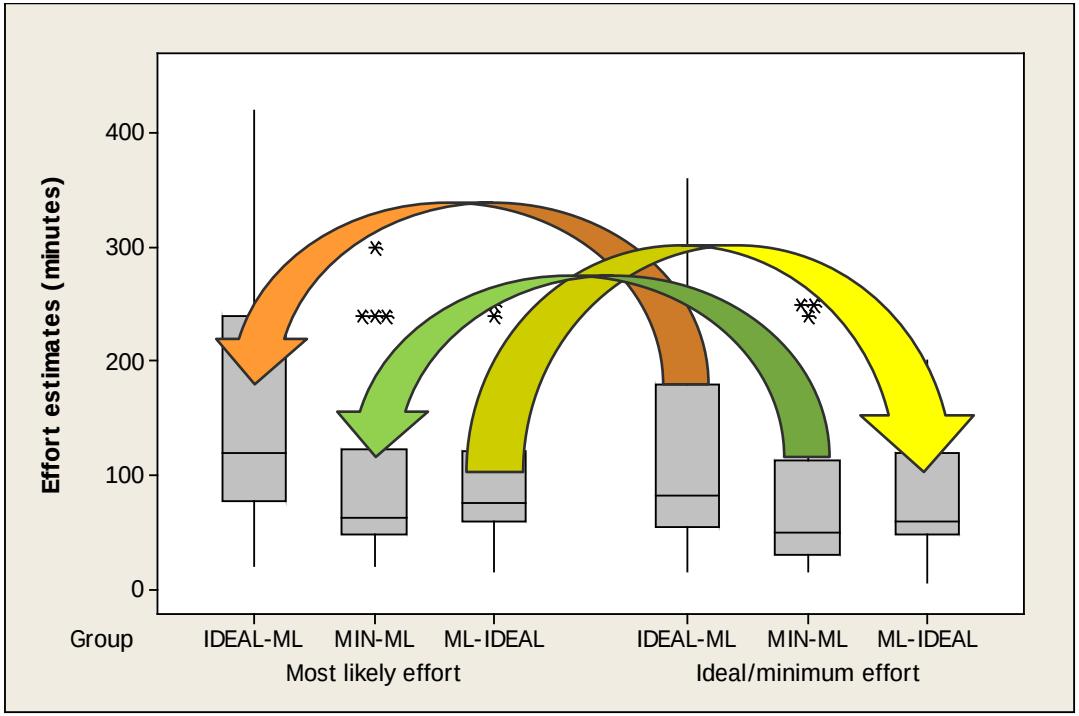
Eksperiment 1

- IDEAL-ML: Først ideell tid, deretter mest sannsynlig tid.
- ML-IDEAL: Først mest sannsynlig, så ideell tid.
- Ideell tid: *"Antall timeverk antatt full konsentrasjon, uten forstyrrelser og fullt ut produktiv."*
- Faktisk tid: Ca. 100 timeverk.
- IDEAL-ML mest realistisk.
- IDEAL samme som ML.
- 25% i ML-IDEAL gruppen økte ikke estimated fra ML til IDEAL.



Eksperiment 2

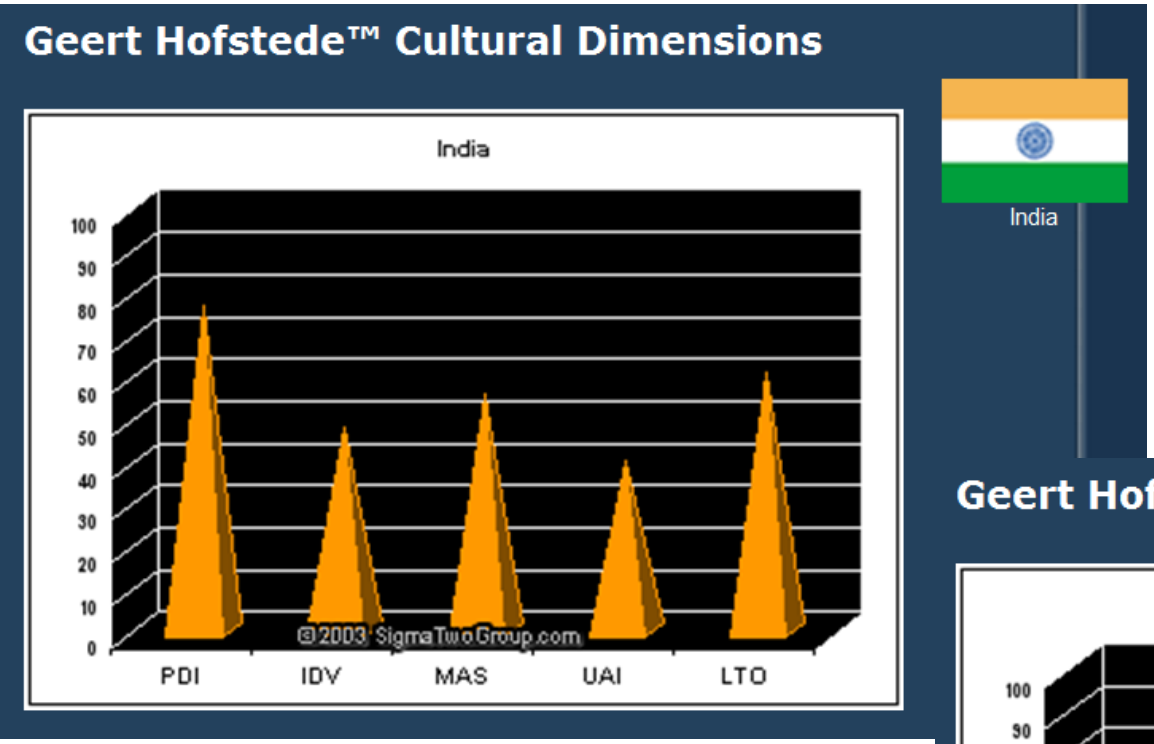
- IDEAL-ML: Først IDEAL, så ML.
- ML-IDEAL: Først ML, så IDEAL.
- MIN-ML: Først MIN, så ML.
- Realistisk tid: Ca. 140 minutter.
- IDEAL-ML mest realistisk.
- IDEAL samme som ML.
- MINIMUM først hadde IKKE samme effekt som IDEAL først.



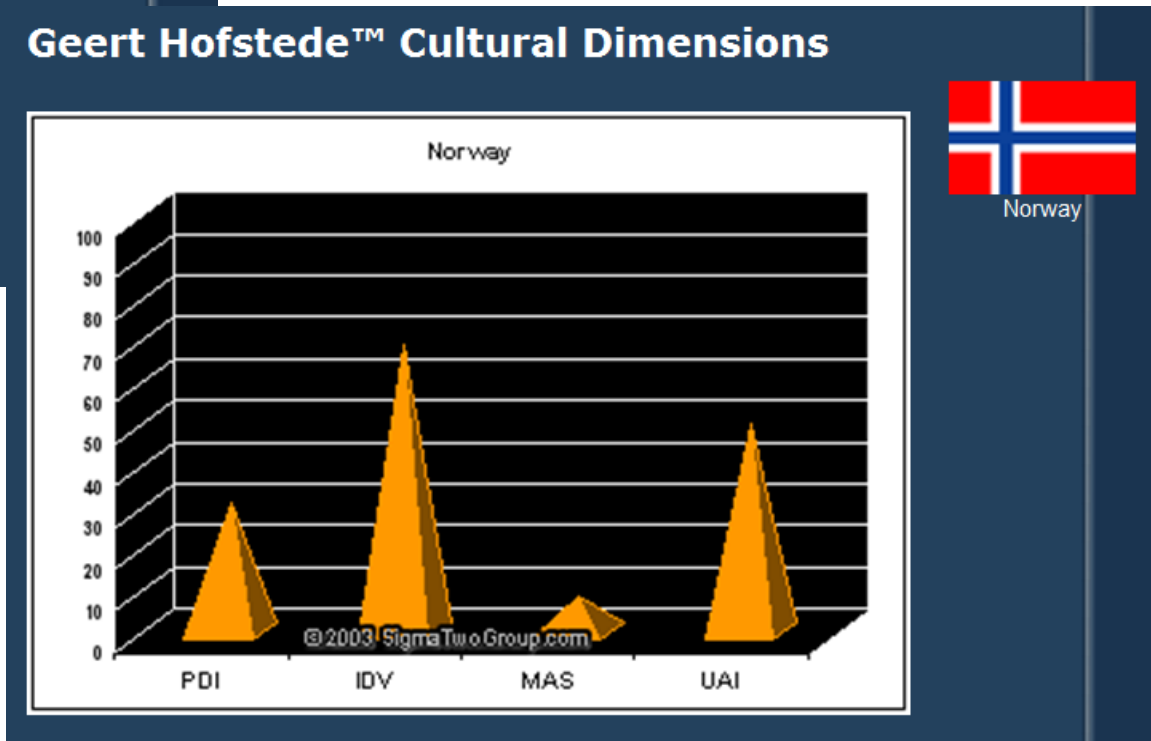
IDEAL->ML metoden

1. Be om et estimat under antagelsen av ideelle betingelser.
   a. Ulike formuleringer av hva som menes med ideelle betingelser synes å gi noenlunde samme effekt.
   b. Må skille seg tydelig fra realistiske betingelser.
2. Be deretter om et estimat av hva arbeidet realistisk sett krever, f eks i form av mest sannsynlig antall timeverk).
   a. Vi har tidligere funnet at mer risikoanalyse kan gi mer optimistiske estimater. Denne metoden er trolig mer motstandsdyktig mot denne effekten, men det kan likevel være lurt å splitte denne estimeringen i to deler:
      I. Fase 1: Estimer realistisk arbeidsmengde for kjente aktiviteter og kjent risiko.
      II. Fase 2: Estimer tillegg for ukjente aktiviteter. Baser dette på historikk over andel timeforbruk på ikke-planlagte aktiviteter og ikke-identifiserte risiko.

Kultur: Hvor like er vi egentlig ….?



PDI: Power distance index
IDV: Individualism
MAS: Masculinity
UAI: Uncertainty avoidance index
LTO: Long term orientation



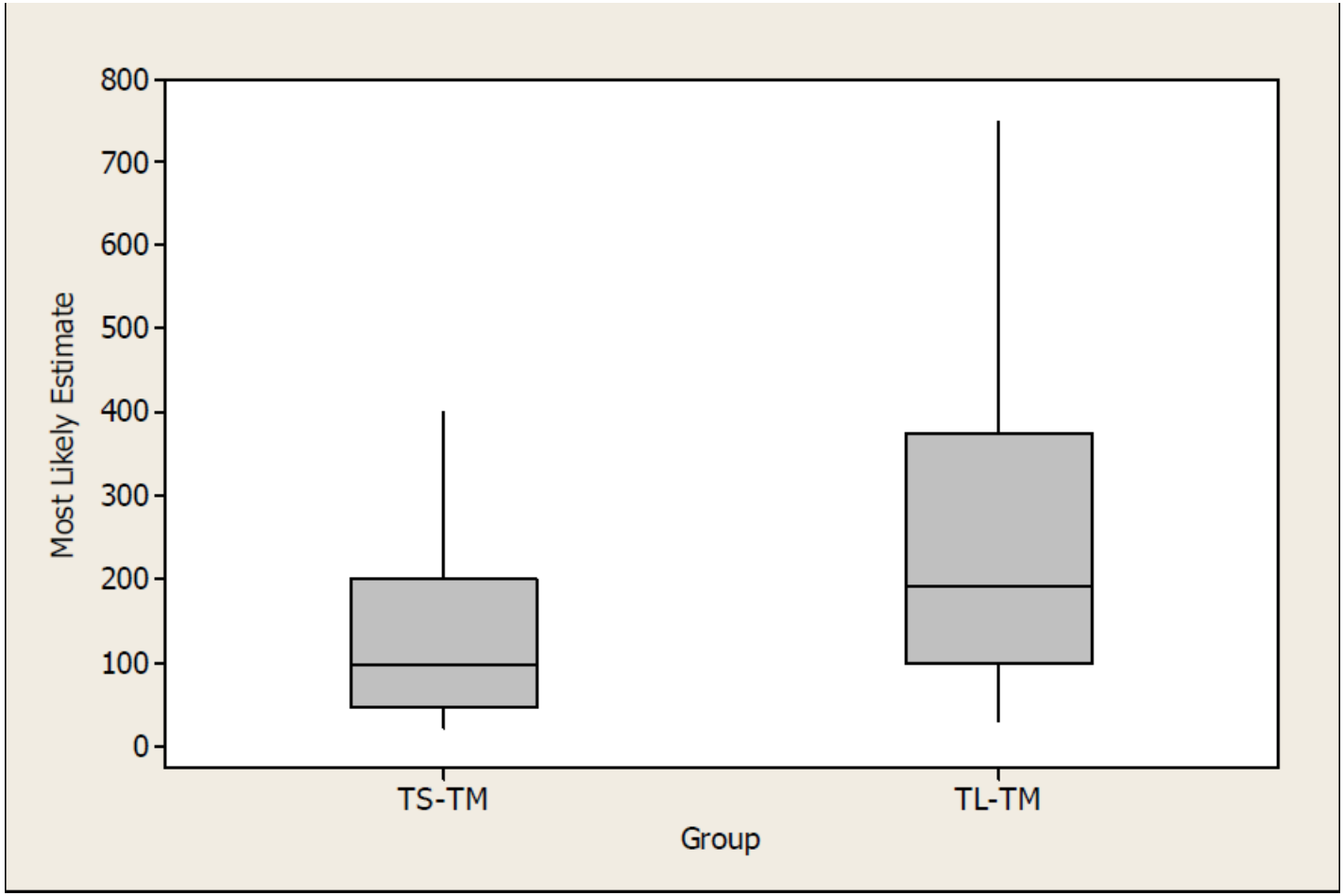
Mer om dette på:
http://www.geert-hofstede.com/

[simula . research laboratory]

Ingen systematiske forskjeller i påvirkning (men legg merke til de lave estimatene fra India)

	Estimation Task 1			Estimation Task 2			Estimation Task 3		
Group	**Low anch.**	**High anch.**	**Diff**	**"Minor ext."**	**"New func."**	**Diff**	**Control**	**Irr. Inf.**	**Diff**
India	25	150	125***	63	80	17	30	58	28*
Nepal	11	120	109***	50	152	102*	80	90	10
Poland	12	100	88***	102	110	8	80	100	20
Romania	10	70	60***	95	100	5	50	70	20
Ukraine	10	100	90***	120	120	0	60	200	140*
Vietnam	25	100	75***	90	120	30	100	100	0

Sekvenseffekt



TS-TM: Estimerte først et lite system (ca. 25 tv), så det middels store.
TL-TM: Estimerte først et relativt sett stort system (ca. 500 tv) så det middels store.

Kan vi prediktere hvem som vil være mest realistisk?

- Få robuste resultater.
- Eneste faktor som er noenlunde bra (men heller ikke den veldig bra) er tidligere estimeringsnøyaktighet
- Uklart hva estimeringsekspertise egentlig er.

Oppsummert: Hvilke godt dokumenterte forbedringsmuligheter har vi?

- Bruk av historiske data (“looking back”)
  - F eks: Analogi-basert estimering og usikkerhetsvurderinger basert på tidligere fordeling av estimeringsfeil (**NB**: Generelle estimeringsmodeller fungere dårlig)
  - Ansvarlig for innsamling og bruk av historiske data: Cost engineer? “Special Estimation Force”?
- Kombinering av UAVHENGIGE estimeringsmetoder og estimeringseksperter
- Finne relevant ekspertise (som er SMAL!)
- Gruppe-estimering (gitt at den ikke er dominert av en person)
- Unngå villedende og irrelevant informasjon, f eks ved å fjerne nøytralisere irrelevant og villedende informasjon før estimering.
- Unngå “Winner’s curse” (lettere sagt enn gjort, men noe kan gjøres)
- Identifikasjon av “skumle prosjekter” og gi disse særlig oppmerksomhet.