

Understanding Big Data sets

Holger Rootzén, Mathematical Sciences

A revolution

Matematiken behövs för framtidens utmaningar

Kongressbiblioteket i Washington DC, Library of Congress, är världens största bibliotek. De mer än 96 miljoner böckerna och manuskripten innehåller 100 terabyte data. I år kommer minst en miljon gånger mer data än så att skapas. Om tio år hundra gånger mer. Datamängderna inom vetenskap, i industrin och i samhället blir alltmer omfattande och ökar i en takt som är svår att förstå. Det ger oss nya utmaningar, men också möjligheter. För forskare i matematik och statistik gäller det att utveckla metoder för att klara problemen och använda de möjligheter som skapas. Det kan handla om hur man bäst planerar gigantiska försök, finner mönster och strukturer i enorma datamängder eller hittar de guldkorn som finns gömda i berg av siffror. Utmaningarna finns inom många områden. Här är åtta exempel:

**Collect
Produce
store**

Understand and use

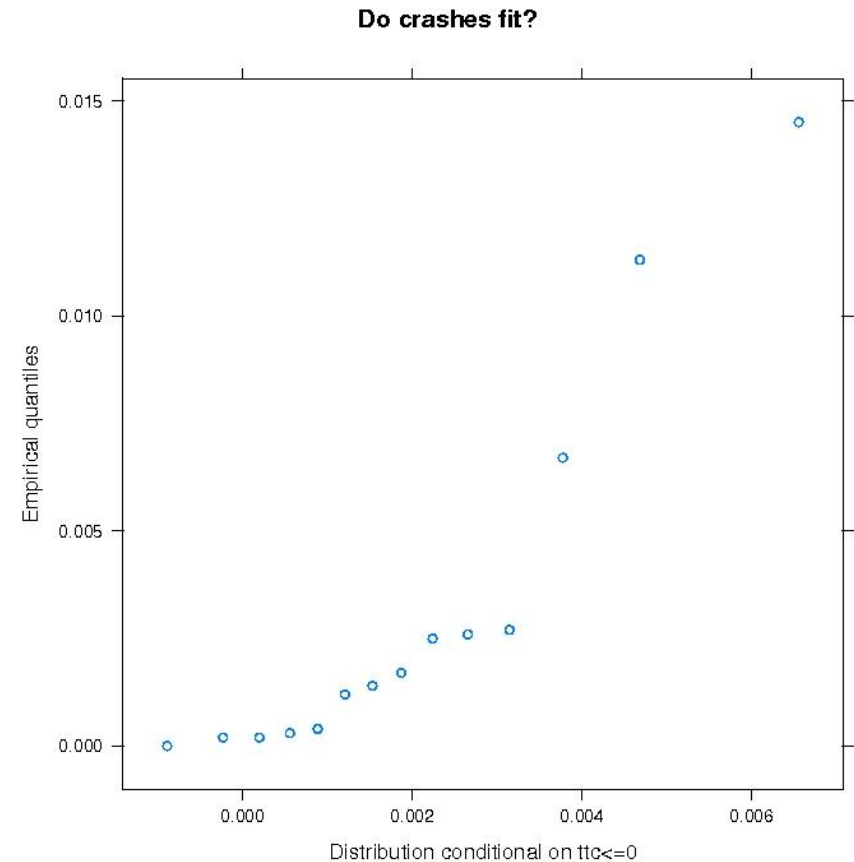
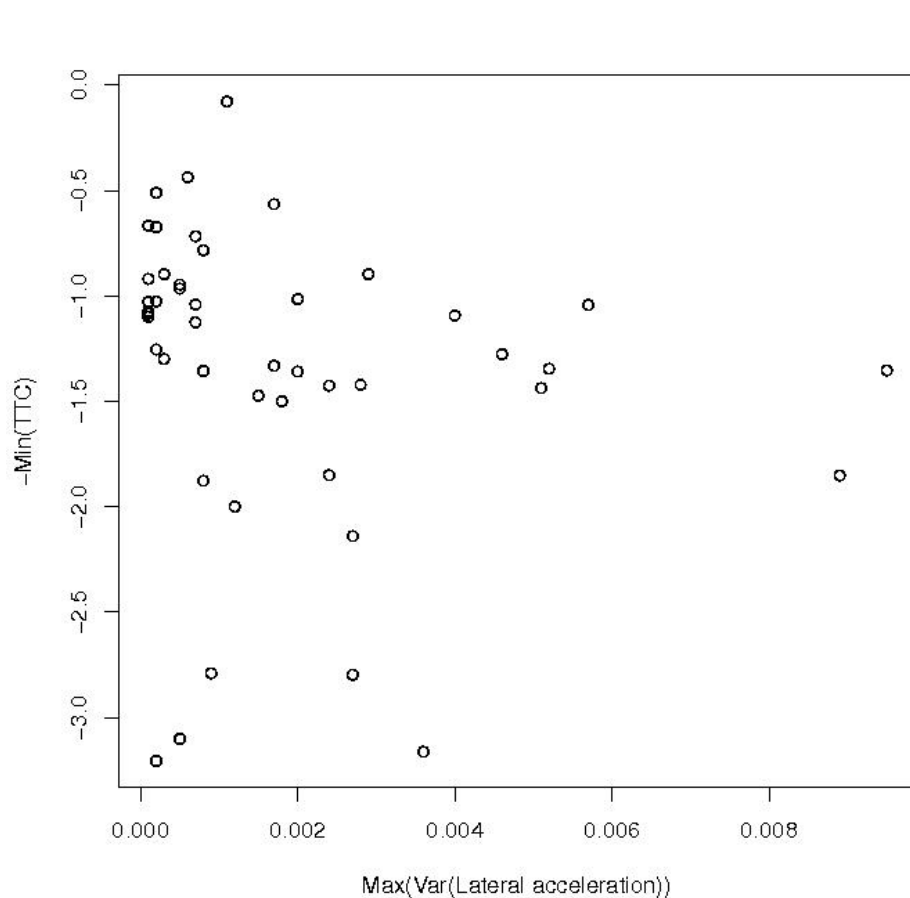
Three examples from my own research:

Car accident prevention



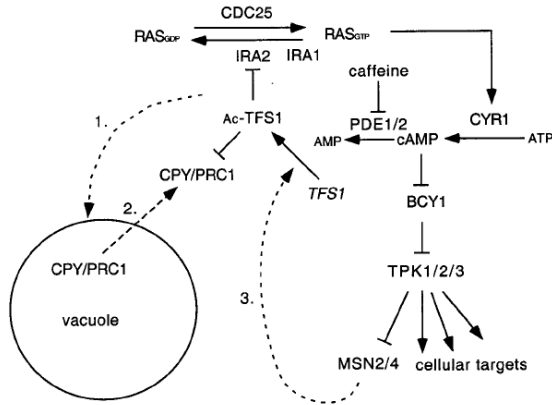
Next study: 1000 cars in 3 years (?)

Will crashes be prevented if we prevent nearcrashes?



Crash \rightarrow Time To Collision extremely small (i.e. < 0)
 \rightarrow are other variables (eye movement, speed, unstable steering, ...) extreme at the same time?

False positive test results



Yeast
understanding basic
life functions

11,904 p-values

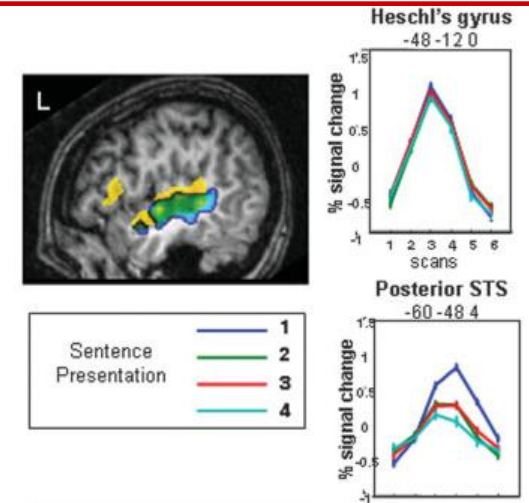
*Blomberg et al. 2003,
2010*



Arabidopsis Thaliana
association mapping

3,745 p-values

Zhao et al. 2007



fMRI brain scans
function of brain
language network

appr. 3 mill. p-values

Taylor et al. 2006



27 million p-values from study of interaction between pairs of genes in yeast, coming from 27 million experiments in 4 replications



... and this is the future

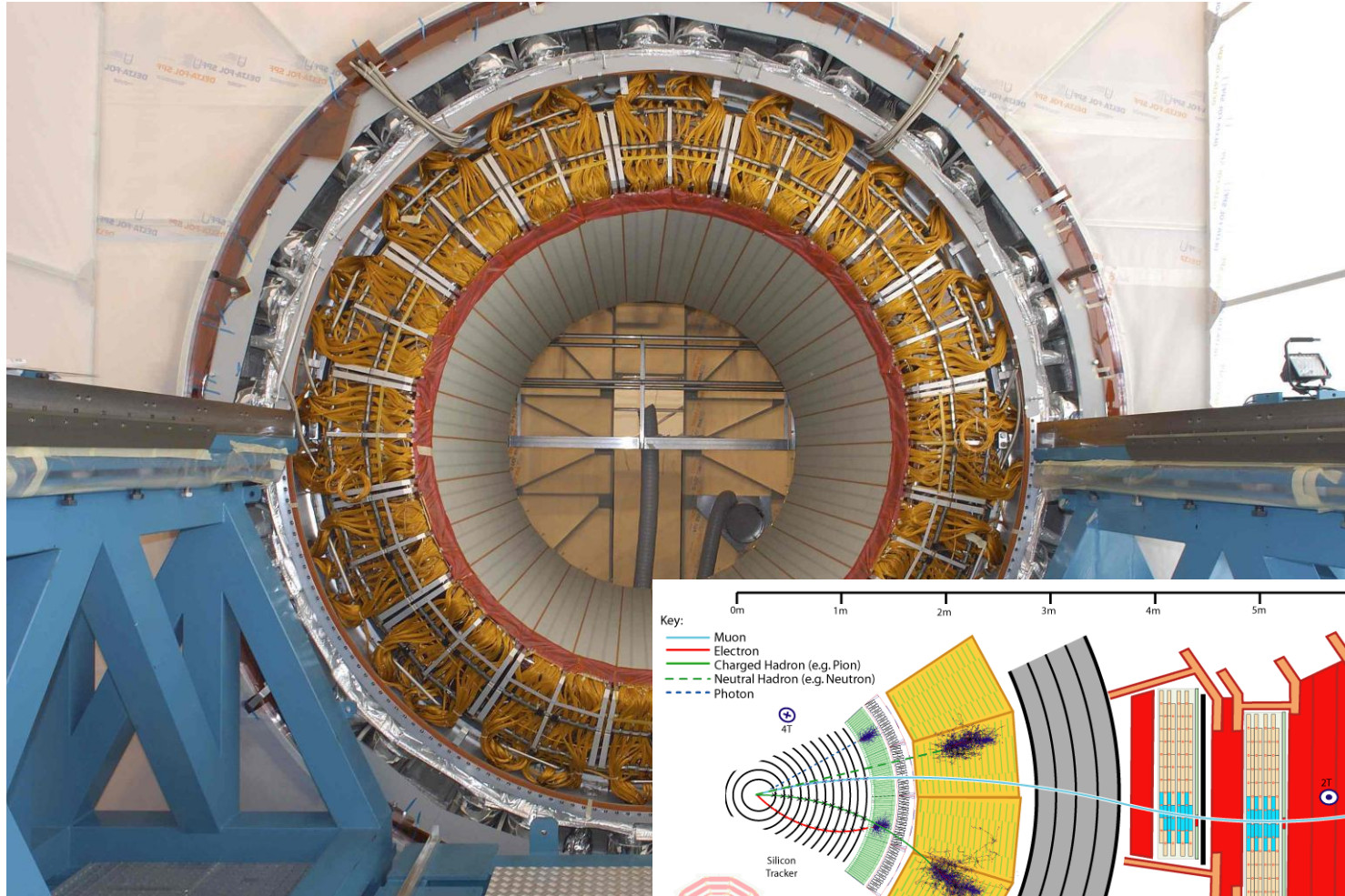


***how many of
the positive
test results are
false?***

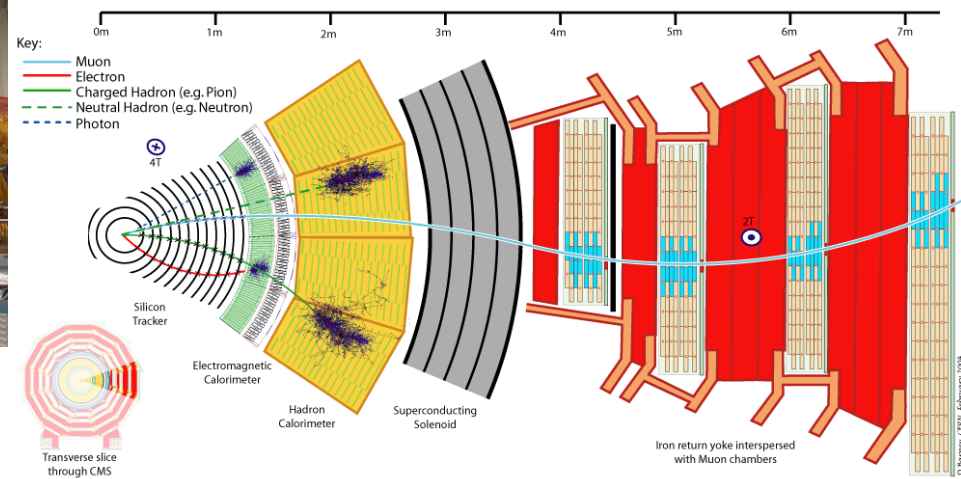
***which pre-
processing
method is
best? *)***

*) typically true null
distribution is different from
the theoretical one

(maybe) same problem when you want to separate electrons from "jets"



detector in Atlas



Pinapple express



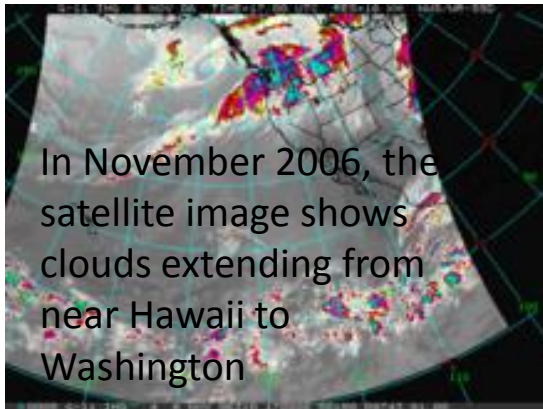
Pineapple express



Unusually high precipitation in the winter of 2005 caused an ephemeral lake to occur in the Badwater Basin of Death Valley National Park.



November 2006 flood, Granite Falls on the Stillaguamish River



In November 2006, the satellite image shows clouds extending from near Hawaii to Washington

A Pineapple Express battered Southern California from January 7 through January 11, 2005. The storm caused mud slides and flooding, with some locations receiving spectacular totals: San Marcos Pass, in Santa Barbara County, received 24.57 inches (624 mm), and Opid's Camp in the San Gabriel Mountains of Los Angeles County was deluged with 31.61 inches (803 mm) of rain in the five day period.

Extreme Episodes

Climate change: *slow down* – reduce CO₂-emissions
mitigate – bigger drains, or higher dykes,
or stronger buildings, or better handling of heat
waves, or ... ??

”map” of -- distribution of wind speed
-- distribution of precipitation
-- ...

in an extreme episode (= storm)

New models, estimation methods, limit theory, computational
methods, ...

Modellera biologin

En ny svensk simuleringsmodell använder två miljoner differentialekvationer för att beskriva hur ett hjärta fungerar. Men ännu mera detaljerad, realistisk och flexibel modellering kommer att behövas för att utveckla framtidens nya individanpassade läkemedel och behandlingar.

Göra fartyg säkrare

För att uppskatta risker för utmattningsbrott utrustas fartyg med elektronisk utrustning som mäter belastningen på många ställen och som ger massiva datamängder. Förutsägelser av framtida belastningar och risker baseras på detaljerade matematiska modeller för vindar och vågor på hela världshavet och under hela fartygets livslängd.

Reda ut näten

Internet, Facebook och proteininteraktion. En ny gren av matematiken som sysslar med modellering, informationsextraktion och förståelse av tekniska, sociala och biologiska nätverk växer fram. Den är ännu i sin linda och ny matematik kommer att behövas.

Filma i detalj

Långa filmer som följer mikroskopiska förlopp är viktigt för forskning inom många områden. Bildmängdernas komplexitet och storlek ställer extrema krav på modellering och analys. Resultaten är viktiga för materialvetenskap, samt för läkemedels- och livsmedelsindustrin.

Hantera finanser

Varje dag läggs data om många hundra miljoner finansiella transaktioner ut på nätet. Dessa kan användas för att förbättra resultat och minska risker i finansiella institutioner, och matematiska metoder utvecklas snabbt. Att inte data i denna utveckling skulle utsätta svensk ekonomi för stora risker.

Optimera strålning

NY teknik gör det möjligt att individanpassa strålbehandling för maximal effekt och minimala biverkningar. Optimering av behandlingen kräver extremt dataintensiva beräkningar. Nya behandlingar, som lättjonstrålning, innebär nya utmaningar för matematiken.

Spara fibrer

Detaljerad matematisk och statistisk modellering av dynamiken i miljontals fibrer i en pappersbana kommer att ge papper med bättre egenskaper och med mindre materialåtgång. Detta är av centralt intresse för Sverige.

Tolka klimatdata

En enda klimatsimulering genererar många gigabyte data. Mängderna som varje dag samlas in via satelliter och väderstationer är ännu större. Nya matematiska och statistiska metoder behövs för att med hjälp av dessa data förstå effekterna av klimatförändringar.

more abstract