Ekonomiska kommentarer

SVERIGES RIKSBANK

NR 17, 2015

# Riksbankens framtida informationsförsörjning i ljuset av Big Data[1]

Jyry Hokkanen,Tor Jacobson verksamma på avdelningen för penningpolitik och vice riksbankschef Cecilia Skingsley samt Markus Tibblin verksam på avdelningen för penningpolitik[2]

**Data och statistik är en hörnsten i Riksbankens arbete. Under de senaste åren har utbudet av data ökat kraftigt och utvecklingen ser ut att fortsätta bland annat för att allt större mängd aktiviteter lagras automatiskt på olika sätt. Denna datarevolution, som mynnat ut i begrepp som Big Data, utmanar traditionellt tänkande och ställer samtidigt nya krav på bearbetning och analys. Utvecklingen av nya analysmetoder för Big Data går snabbt och idag finns flera tillämpningar som är intressanta för centralbanker. För att vara i takt med tiden arbetar Riksbanken med en informationsförsörjningsstrategi som säkerställer att relevant data och statistik finns tillgängliga för de beslut som fattas idag och i framtiden.**

Med begreppet Big Data avses de mycket stora datamängder som tack vare olika tekniska genombrott blivit möjliga att inhämta och lagra. Begreppet Big Data innefattar även strömmande data och data av icke-traditionell typ såsom text. Dessa data är ofta komplexa och i många fall ostrukturerade, vilket innebär att de vanligtvis måste bearbetas och analyseras med särskilda metoder innan de kan användas.[3]

Hantering av stora datamängder är i sig inte något nytt; det har gjorts av både forskare och statistiker sedan länge. Men idag har vi alltså bättre tekniska möjligheter att samla, lagra, strukturera och analysera allt större och mer komplexa datamängder, bland annat alla nya data som vår närvaro på internet genererar. Detta skapar i sin tur en potential att hantera, strukturera och extrahera kunskap ur stora, strömmande eller icke-traditionella datamängder på ett sätt som vi inte kunnat göra tidigare.

Big Data är en datarevolution som också kommer att påverka hur centralbanker använder och analyser data. Traditionellt används aggregerade tidsseriedata, som publiceras med en fördröjning, för att följa den ekonomiska utvecklingen. Icke-traditionell data som på olika sätt genereras i samhällsekonomin skulle ytterligare kunna förbättra förståelsen för den ekonomiska utvecklingen. Om sådana data dessutom observeras kontinuerligt kan beslutsfattarna reagera snabbare på förändringar i utvecklingen. Men för att myndigheter som till exempel Riksbanken ska kunna dra nytta av nya och allt snabbare växande datamängder krävs en väl genomtänkt strategi som utgår ifrån att data är en strategisk resurs. Datarevolutionen utmanar traditionellt tänkande inte bara kring datainhämtning och analys; den kräver dessutom nya kompetenser, ny teknik och en ändamålsenlig organisation.

## Big Data – analysmetoder och användningsområden

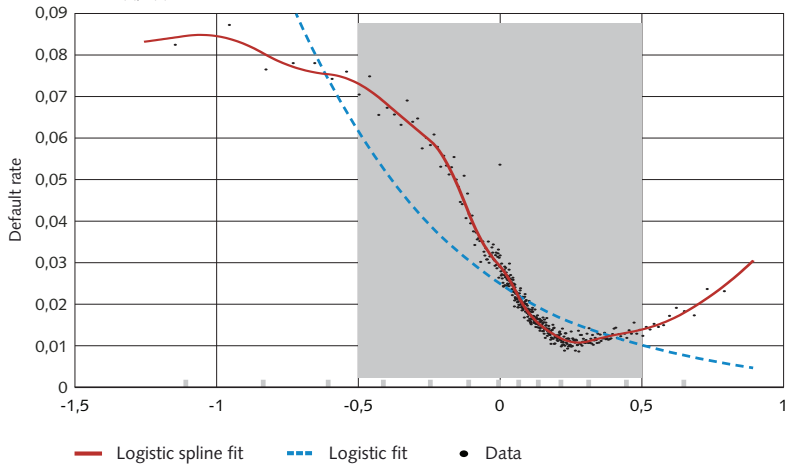
I en pedagogisk och lättillgänglig artikel från 2014 presenterar Googles chefekonom Hal Varian en uppsättning analysmetoder för hantering av stora datamängder som brukar refereras till som "machine learning", se Varian (2014). Machine learning-metoder har vuxit fram i gränssnittet mellan statistik och datavetenskap och beteckningen är en anspelning på det faktum att en del av metoderna är självspecificerande, det vill säga att algoritmerna utformar sin egen modelldesign utifrån datas utseende och därmed lär sig av data. Enkelt uttryckt är machine learning i mycket högre utsträckning än traditionell ekonomisk-statistisk analys inriktad mot förutsägelser, eller prognoser. Det finns således en större acceptans för en "black box"-ansats då analysresultaten typiskt inte är baserade på ekonomisk teori.[4] När man har stora datamängder kan data delas och användas för olika ändamål, exempelvis kan modellen

1. I september 2015 anordnade Riksbanken en workshop under rubriken "Big Data: Building data strategies for central banks in light of the data revolution". Arrangemanget samlade centralbanker, myndigheter, forskare och kommersiella användare av Big Data. Presentationerna från konferensen finns tillgängliga på Riksbankens hemsida, www.riksbank.se/sv/Press-och-publicerat/Nyheter/2015/Riksbanken-ordnar-workshop-om-big-data/.
2. Författarna tackar Mikael Apel och Marianne Sterner för värdefulla synpunkter.
3. Exempel på vad som skulle kunna kallas Big Data är textinlägg i twitter, information från de enorma transaktionsregister som byggts upp till följd av regleringar såsom MiFid eller EMIR samt realtidsdata från vägkameror (strömmande data).
4. Uttrycket "black box" syftar i det här sammanhanget på att modellernas samband eller relationer inte alltid är direkt tolkningsbara. De genererar goda prognoser, men det är inte uppenbart hur.

"träna" sig eller lära sig på en delmängd av observationerna och därefter utvärderas på en annan delmängd. I traditionell ekonometrisk analys, anpassad till och utvecklad för små datamängder och ofta aggregerade data, utgör ekonomisk teori grunden för analysen. Machine learning är i det hänseendet mindre beroende av teori och mer beroende av hur data ser ut, det vill säga datadriven.

Stora datamängder ger också större frihet i valet av funktionsform. Den stora mängden observationer gör det möjligt att upptäcka komplexa icke-linjära samband som bättre beskriver datas utseende än de påtvingade linjära samband som ofta följer av försök att applicera ekonomisk teori.[5] I figuren nedan exemplifieras detta för sambandet mellan svenska aktiebolags vinstandelar av totala tillgångar och konkurs (se Giordani, Jacobson, von Schedvin och Villani 2014). En konventionell modellansats skattar med en så kallad logistisk sannolikhetsmodell en kontinuerligt avtagande konkursrisk ju högre vinstandelen är (den streckade linjen). Ett alternativ med en flexibel icke-linjär så kallad spline-modell fångar bättre det empiriska sambandet i data och visar tydligt att "övervinster" i stället är förknippade med förhöjd konkursrisk (den heldragna linjen som ansluter till observationssvärmen). Den förstnämnda, logistiska, modellen underskattar riskerna för låg- och högvinstföretag, men den överskattar riskerna för företag som gör normalvinster. Föga förvånande finner man att den icke-linjära modellen gör betydligt bättre prognoser av konkursrisken än konventionella linjära modeller. Den här typen av spline-modeller baserade på data för hela populationen av svenska aktiebolag över lång tid kan exempelvis användas för att värdera kreditrisker i svenska bankers företagsutlåning, som ett led i Riksbankens övervakning av den finansiella stabiliteten.

**Figur 1. Sambandet mellan vinstandel och konkursrisk för svenska aktiebolag**
EBITDA/TA




Den information som fångas från internet är ofta ostrukturerad och består av text eller textfragment som behöver analyseras.[6] Källorna är många: exempelvis all typ av nyhetsrapportering, de sociala medierna, marknadsrapporter och marknadsanalyser.[7] Ett välkänt exempel är de data som olika sökverktyg på internet ger upphov till. Bholat m.fl. (2015) visar att det finns ett starkt positivt samband mellan brittisk arbetslöshetsnivå och mängden sökningar på Google efter information om villkoren för arbetslöshetsersättning. Detta visar att variationerna i sökfrekvenser på Google alltså skulle kunna användas för fånga aktuella förändringar i viktiga underliggande variabler som exempelvis total arbetslöshet.

5. Valet att i en empirisk modell ansätta linjära samband reflekterar ofta en vilja att ta en linjär teoretisk prediktion till data, men kan också vara en nödvändig eftergift då antalet observationer är otillräckligt för att beskriva de empiriska sambanden fullt ut.
6. Nätet möjliggör naturligtvis också fångst av numerära data. Ett aktuellt exempel på sådana ges i ett utvecklingsprojekt på Riksbanken som syftar till att samla in prisuppgifter på varor och tjänster genom så kallad skrapning från nätet för att i realtid uppskatta och beskriva inflationsutvecklingen.
7. Centre for Central Banking vid Bank of England har presenterat en översikt av olika metoder för textanalys som centralbanker kan tillämpa (se Bholat, Hansen, Santos och Schonhardt-Bailey 2015).

Rönnqvist och Sarlin (2015) visar i en ny studie hur avancerad textanalys kan användas för att prognosticera förhöjd stress (jämför eng. distress) för enskilda banker.[8] Enkelt uttryckt utnyttjar metoden det faktum att vissa ord från samma begreppssfär tenderar att förekomma i liknande sammanhang. Detta gör att man genom en semantisk analys av ord och av innehållet i hela artiklar kan få en bild av vilka kombinationer av ord och text som kan förutsäga förhöjd bankstress. Modellen kan användas för att generera ett stressindex för såväl enskilda bankinstitut som för hela branschgrupperingar. Även om ett förhöjt stressindex inte nödvändigtvis behöver tolkas negativt, så ger det en signal om när det är rimligt att göra en fördjupad analys av en specifik bank. Eftersom modellen baseras på textdata kan man genom utdrag ur källtexterna direkt få beskrivningar av de händelser som driver indexet. Modellen kan således göra det möjligt att samtidigt upptäcka och beskriva händelser som är relevanta för den ekonomiska utvecklingen.

## En modern centralbank behöver en informationsförsörjningsstrategi

Utöver den generella tillväxten i data har även den senaste tidens ökade reglering och övervakning av de finansiella sektorerna lett till att mängden data och statistik för analys och övervakning ökat kraftigt.[9] Denna utveckling väntas fortgå i accelererande takt och för att kunna dra nytta av utvecklingen har flera centralbanker sett över sin data- och informationsförsörjningsstrategi.

I de datastrategier som nu formuleras ses data som en strategisk resurs för hela centralbanken. Big data (tillsammans med traditionell data) ses som något som kan leda fram till en djupare förståelse av underliggande ekonomiska fenomen och ge tidiga signaler om läget i ekonomin. För att den visionen ska förverkligas krävs dock en övergripande strategi för hur insamling, bearbetning, lagring och spridning av data ska hanteras så att informationen kan komma hela organisationen till gagn. Utvecklingen av strategierna sker i nära samarbete mellan ekonomer och analytiker, vilka är de som främst analyserar data, men även jurister och ansvariga inom IT deltar.

Implementeringen av de nya strategierna kräver både mer resurser, ny kompetens och tydligare samordning mellan centralbankens olika policyområden. På Federal Reserve Board of Governors och på Bank of England har implementeringen även inneburit organisatoriska förändringar. Ansvaret för den övergripande datahanteringen har i dessa organisationer flyttats till en egen avdelning, Office for the Chief data officer. Hit har personer med kompetens inom dataarkitektur, data compliance och data governance rekryterats för att utveckla och implementera bankgemensamma system och arbetsrutiner för data. Primärt ansvarar de för att utveckla och förvalta den övergripande datamodellen[10], upprätta enkla och säkra processer för datainsamling och bearbetning samt säkerställa att data lagras på ett standardiserat och säkert sätt men samtidigt ger enkel åtkomst för olika användare. Inom Europeiska centralbanken (ECB) har ett liknande arbete skett men i annan form. En övergripande infrastruktur och datamodell utvecklas centralt inom ECB och stort fokus läggs på att strukturera insamling och lagring och på att göra data lättillgängliga för användare från flera policyområden.

## Big Data ger upphov till nya utmaningar för centralbanker

Den snabba utvecklingen och de förändringar i synen på data som följt har gett oss ett antal nya utmaningar. Trots fördelarna med att enkelt kunna få tillgång till data från andra policyområden har det varit svårt för olika avdelningar på centralbanker att utan vidare anpassa sin egen datahantering till ett övergripande ramverk. Inom Federal Reserve och Bank of England lyckades man motverka avdelningarnas motsträvighet genom att införa ett banköverskridande dataråd. Datarådet fattar beslut i över-

8. Data för deras prediktionsmodell är av två slag: dels 6,6 miljoner nyhetsartiklar i Reuters öppna arkiv från perioden 2007 till och med 2014, dels 243 identifierade "stress-händelser" som inträffade under perioden 2007 till och med juni månad 2012 och som berört någon av de 101 systemviktiga europeiska storbanker som inkluderas i studien.
9. Till exempel har nya regleringar av derivatmarknaden inom EU generat insamling av stora mängder detaljerad data. Idag beräknas dessa källor innehålla mer än 15 miljarder datapunkter.
10. En modell som beskriver hur olika data förhåller sig till varandra.

gripande frågor kopplade till insamling, bearbetning, lagring och spridning av data. I detta råd presenteras och bearbetas förslag på förändringar i datahanteringen, och banköverskridande beslut om data kan fattas med alla delar av centralbanken representerade. Datarådet har varit en viktig faktor i förändringsprocessen och säkerställt att avdelningarna på centralbanken varit trygga med att anpassa sig till de övergripande dataprinciper som fastställts.

Det har också varit utmanande att i praktiken analysera de nya datakällorna. Dels har personer med kompetens inom "data science"[11] varit svårrekryterade, dels har centralbankens IT-kapacitet inte alltid levt upp till de krav som ställts för bearbetning av de stora datamängderna. Analysmetoderna genererar dessutom ofta komplicerade modeller, vilket gör tolkningen av resultat svår och förmedlingen av policyrekommendationer än svårare.

## Riksbankens framtida informationsförsörjning

Riksbankens ansvar för penningpolitik och finansiell stabilitet fordrar tillgång till data från många olika källor för analyser och prognoser. En grundbult i dataanskaffandet har traditionellt varit, dels den statistik som Riksbanken själv reglererar i föreskrifter för de finansiella instituten, dels övrig statistik och data (antingen offentlig eller inköpt). Liksom andra centralbanker behöver även Riksbanken se över hur data hanteras internt på ett effektivt sätt och hur man tar sig an - i många fall mycket detaljerade - data från nya källor som dessutom kan innehålla ostrukturerad information och som kräver en annan form av bearbetning och analys än vad som tidigare varit gängse. För att uppnå en ändamålsenlig hantering av information behövs således en vision för informationsförsörjningen med en åtföljande strategi som vägleder dataanskaffningen och bearbetningen.

Riksbankens vision för sin informationsförsörjning kan uttryckas så enkelt som *Rätt data i rätt läge*. Strategin kräver således att relevanta data, i den mån de inte redan finns i Riksbanken, kan införskaffas på känt sätt, till känd kostnad och från rätt källa. Utöver detta krävs även att Riksbanken vet hur data ska tillgängliggöras för rätt person vid rätt tillfälle och att beslut och ansvar är klarlagda i samband med inhämtning, lagring och åtkomst av data. Behoven av att inhämta och analysera framför allt icke-strukturerade data ställer särskilt höga krav på analysverktyg och datainfrastruktur och fordrar även en aktiv omvärldsbevakning. Ambitionen att ha tillgång till relevanta data är resurskrävande och måste alltid ställas mot analysbehovet, kostnaderna, uppgiftslämnarbördan med mera.

## Slutsatser – datarevolutionen möjliggör nya arbetssätt

Mängden tillgängliga data är enorm och fortsätter att öka i allt snabbare takt. Nya analysmetoder har tillsammans med ny teknik gjort det möjligt att analysera Big Data. Flera centralbanker anser att de har möjlighet att dra nytta av utvecklingen för att skapa bättre underlag för sina policybeslut och har därför ändrat sina strategier för hantering av data och statistik. Vid dessa centralbanker har ny personal med kompetens att både behandla och analysera data rekryterats. Dels för att ordna insamling, bearbetning och lagring av data. Dels för att med hjälp av nya IT-system och analysmetoder dra nytta av innehållet i nya typer av data som textinformation eller enorma mängder mikrodata.

Big Data innebär samtidigt betydande utmaningar för centralbankerna, både tekniskt och metodologiskt. Därutöver behöver även nya avvägningar göras när det gäller strategi, organisation, kompetens och ekonomi.

Relevanta data och statistik är en hörnsten i Riksbankens arbete. För att vara i takt med tiden behöver således även Riksbanken en strategi som säkerställer att relevanta data finns tillgängliga för de beslut som fattas idag och i framtiden. Devisen *rätt data*

11. Data science är ett ämnesövergripande område där olika metoder och system används för att samla in, bearbeta och analysera Big Data. Metoderna som används baseras på en kombination av idéer från datavetenskap, programmering samt numerisk och statistisk analys. Big Data-projekt inkluderar ofta flera individer med sinsemellan kompletterande tekniska kunskaper inom dessa områden.

*i rätt läge* innebär att analysbehoven styr vilka data som ska samlas in samtidigt som anskaffning, hantering och analys av data alltid måste ställas mot kostnaderna och det potentiella mervärde en analys skulle ge.

## Referenser


Bholat, David, Stephen Hansen, Pedro Santos, and Cheryl Schonhardt-Bailey, 2015. Text mining for central banks. Centre for Central Banking Studies Handbook No. 33.

Giordani, Paolo, Tor Jacobson, Erik von Schedvin, and Mattias Villani, 2014. Taking the Twists into Account: Predicting Firm Bankruptcy Risk with Splines of Financial Ratios. *Journal of Financial and Quantitative Analysis*, 49, 1071–99.

Rönnqvist, Samuel and Peter Sarlin, 2015. Detect & Describe: Deep learning of bank stress in the news. *Proceedings of the 2015 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, forthcoming.

Sveriges riksbank, presentationsmaterial från Riksbanken workshop, "Big data: Building data strategies in light of the data revolution", www.riksbank.se/sv/Press-och-publicerat/nyheter/2015/Riksbanken-ordnar-workshop-om-big-data/

Varian, Hal R., 2014. Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28, 3–28.