150.

RIKSREVISIONEN

Datum: 2023-10-27

Sara Monaco
Ange avdelning

Intervju KTH18 230322

Från RiR: Ludvig Stendahl och Sara Monaco

Prof, datadriven i forskningen, maskininlärning, stora datamängder, genererar mkt data, jag är ju definitivt en av dem som har högre behov av infrastruktur kring data.

Den vanliga rollen som professor: söker projekt, driver projekt, drivs primärt av doktorander. Vi tar in ibland exjobbare och andra forskare i dem också då som sagt, men det är min roll att driva projekt. Jag är ansvarig för KTH strategiska samarbete med Region Stockholm och i den rollen så har jag ju sett behov på KTH av datahantering som kanske sträcker sig utanför min forskning. Jag har också varit ganska aktiv och försöka driva datafrågor på KTH varit i kontakt med Rosa bland annat för att ge input på vad jag anser behövs.

Inom kommunikation har vi ganska standardiserade datamodeller så det är mycket simulerade data. Det är väldigt lite riktig data i den meningen. Inom life science tillämpningarna så är det ju mycket mer riktig data i den mening att de uppmätt kring någon process. Jag har själv hållit mig ifrån än så länge persondata, så när det har varit data har ju varit till exempel mikroskopi data från celler som man har odlat i något experiment. I de fallen har det ju varit celler som någon gång har kommit från en människa. Man har arbetat med cellinjer som är frånkopplade från individer. Vi arbetar med genetik data nu då, men då att vi har metodutveckling så jobbar vi inte med mänsklig det mänskliga genomet. Det finns ingen poäng med det. E coli bakterier som jag sekvenseras och genererat ganska stora datamängder. Men det rör sig inte om persondata och känslig data.

Vad tillhandahåller KTH vad gäller infrastruktur, saknar du ngt?

Beräkningsinfrastruktur för beräkningar, vi har ju PD parallell dator centrum, till exempel för beräkningar. Jag utnyttjar också nationell infrastruktur i så det bland annat finns det infrastruktur som är finansierade av oss programmet eller Familjen Wallenberg. Berzelius i Linköping som en del av den nationella infrastrukturen för beräkningar och just de datorerna lämpar sig mer för den typen av algoritmer som vi testar.

Jag ser behov av den långsiktiga lagringen. Nuvarande projekt: metoder för DNA-sekvensering. Samarbete med SciLifeLab har hjälpt oss sekvensering i storleksordningen ett par 100GB terrabyte, som vi testar våra algoritmer på. Vi har behov av att säkerställa att vi inte tappar den datan som vi har lokalt på beräkningsdatorer här. Kraschar de hårddiskarna finns en risk att arbetstid och datortid går förlorad. Sådana behov går utanför vad KTH kan tillhandahålla. Vi har ju system för lagring av dokument, då är Onedrive ok. Om man vill spara flera 100 terrabyte..

Infrastruktur vid LiU är det för att göra vissa körningar under pågående projekt eller lagring under pågående projekt?

Ja för körningar men vi har egen beräkningsdator på avdelningen, så det är en video som ger grafikprocessorer kopplade så vi har en dyr dator här på avdelningen som är väldigt mycket mer kraftfull än vanliga laptops som stationära datorer. Sen finns den nationella Berzelius i Linköping som är samma typ av arkitektur, men många, många gånger större. Vi utvecklar algoritmer och testar småskaligt här först och sen så för vi över till Linköping och beräknar och det är projektform man söker beräkningstider och då har man också möjlighet. Alltså vill man ha datalagring för själva beräkningarna, men man har inte långsiktig datalagring där så vi kan inte arkivera datan och säkra den mot förlust genom att lägga det på deras datorer. När projektet är slut så förväntas vi eller förväntas de kunna ta bort datan där och det är inte en långsiktig lösning

Under pågående projekt så kan det ligga kvar på den stordatorn i Linköping berzelius?

Ja och på vår lokala dator här då som sagt.

Var ligger datan sen då? Eftersom Berzelius eller KTH inte kan hantera den?

Det är svårt. Programkoder och så, algoritmer, det kan vi hantera. Det kan vi backa upp mot KTH:s X repository, som är ordentlig backup. Men rådata ligger på vår dator på avdelningen. Vi har en kopierad till externa usb hårddiskar som vi försöker ha på andra platser. Det är lite ostrukturerat, men man får fråga en kollega om man kan ställa den i ett annat hus. Man kanske tar hem hårddisken för att skydda mot brand eller stöld. Är det bara en fråga om att skydda mot krasch skulle det går bra att ha i samma rum som datorn. Men om man vill ha ytterligare säkring mot förlust. Den ligger också på SciLifelabs dator.

Vad skulle du önska för lösning på lång sikt till lagring?

En lokal molnlösning. Vi har onedrive, det är en molnlösning som om jag förstår det ligger på sunets datorer. Men prissättningen på detta passar inte för oss. Den är till för dokument med mindre datamängder, men när det börjar bli flera 100GB terrabyte. Jag skulle vilja se motsvarigheten om jag jämför med Amazon tar för lagring. Jag skulle vilja se en lokal lagring i samma prishärad som Amazon. Jag skulle vilja lägga den någonstans och vara säker på att den inte försvinner. Jag vill vara säker på att den som tar emot kan hantera backuper. Snabba beräkningar går att ha här. Direkta snabb access för beräkningar. Det går alldeles utmärkt att ha på våra datorer för den här typen av data. Den är inte känslig, inte persondata. Jag

skulle också gärna se en betalningsmodell där man kunde förbetala för datalagringen, för vi har ju sådana krav från forskningsfinansiärer att vi ska ta hand, att vi ska spara data som vi har baserat våra publikationer på upp till 10 år till exempel. Men projekten löper ju inte så länge. Man skulle gärna under ett pågående projekt kunna betala för den långsiktiga lagringen, alltså köpa 10 års lagring.

Vem betalar för den långsiktiga lagringen? Går det på projektets anslag eller?

Det går på projektet, men problemet är att det inte finns ett smidigt sätt att köpa. Jag kan köpa hårddiskar på projektets pengar och montera dem på den och sen så kan jag lägga hårddiskarna på andra ställen, alltså sprida ut dem i KTHs lokaler eller ta hem en hårddisk eller att den doktorand som driver projektet tar hem en hårddisk. Man vill säkerställa att det finns lagrat på flera olika platser och då går det alltså för de har ju köpt en fysisk resurs som man inte fortsätter betala för. Det går ju bra. Men det är inte en riktigt önskvärd lösning. Vi ska inte ha data liggande hemma och om man behöver hitta per projekt eller om man behöver utnyttja sitt kontaktnät för att hitta andra fysiska platser på KTH så är ju inte det heller en riktigt. det är inte en bra lösning.

Hur finansiellt betungande är det inom projekten att ordna de här långsiktiga lagringslösningarna?

Att köpa hårddiskar är små kostnader. Några 10-tals tusen kr vilket är försvinnande små summor jämfört med lönekostnader och sånt. Problemet är konteringen. Om vi köper lagring som kräver årlig betalning då måste vi behålla projektet levande i 10 år. Hade varit snyggare lägga kostnaden för lagringen på det projektet som det berör.

Kan frågan lösas på KTH eller krävs nationella tag?

Det går att lösa lokalt. Men snyggare om det fanns en nationell lösning. Liknande SNIC för nationella beräkningar. Vore snyggt att göra på samma sätt. De som driver X de har ju möjlighet att sälja både beräkningar och datalagring till industrin, men någonting hindrar dem från att internsälja samma tjänster. Så jag tror att tjänsterna finns där.

Om en person går i pension, hur fungerar det vad gäller lagring?

Jag vet inte om vi har en bra plan för det.

Det kan ligga lite hårddiskar här och var...?

Ja, det är ju det som är problemet med den här lösningen som vi har när det är separata hårddiskar, att det är lätt att det faller mellan stolarna. Det är klart att man kan ha en datalagringsplan uppsatt, det överlämnades då till avdelningens ansvariga chef där när man går i pension. Men det, det är ju det är ju ett sårbart system, även om man har goda intentioner om man sätter upp det så. Ja, kan ju. Av den anledningen fungerar inte data om det är känslig data.

Infosäk, du har inte jobbat med persondata. Finns andra skyddsvärden?

Jag är inte den som har de högsta kraven kring det, men det finns DNA-sekvensering, inte lätt få fram. Vill ogärna dela innan publikationerna är framme, vill inte dela mellanstegen som är nyhetsvärdet. Risk om ngn aktör vill störa forskningen och komma åt. De beräkningsdatorer vi har – krypterade med privata nycklar. Ligger ju inte öppet för världen. Baseras på vårt kunnande att vi kan sätta upp säkra system.

Förstår en utomstående vad det handlar om?

Vi är kanske inte den grupp som haft bäst praxis kring metadata, notera data, de dataformat vi jobbar. Standardiserade dataformat, vi hittar inte på egna format. Vi jobbar inte med udda dataformat, till exempel har det varit bilddata - vanliga bilder och de kan man bara öppna och titta på. Även sekvenseringdatan är ju standardiserade format så vi hittar inte på egna dataformat för lagring. Däremot så skulle vi säkert kunna vara bättre på annotering av hela datamängderna för att förklara vad det är. Vi gör ju det när vi publicerar. När vi publicerar då annoteras det ordentligt. Icke-annoterad data är ju meningslöst att tillgängliggöra om man inte kan använda den.

Finns det guldägg i din forskning?

Tänker du att man skulle vilja komma åt det? Eller nåt man är särskilt stolt över?

Ja precis ja om man skulle kunna vilja komma åt det.

Svårt...

Det vill du inte berätta om här.

Det är svårt att komma på det för forskningen är ju inkrementell. Vi utvecklar ju en algoritm som vi gör allt vad vi kan för att den ska vara snäppet bättre än vad som existerar, för det är ju det som är nyhetsvärdet och värdet som gör att det går att publicera. Det är inte så att vi har världsomvälvande nyheter. Svårt se att man skulle kunna plocka guldägg. Man skulle kunna se allmänt att vi är bra grupp. Skulle man förstå att vår grupp var någonting på spåren som man skulle aktivt försöka ta, försöka ta sig in i våra system genom att lura oss via phishing eller så vidare och just plocka något sådant guldägg. Är nog snarare att man kan se att det är en allmänt kompetent forskargrupp på KTH. Undrar vad de gör och därmed försöka lura oss och få access till systemen. Vi sitter ju inte länge på data, görs öppet tillgängligt. När publikationen är färdig som baserar sig på den datan så är ju vår intention att göra både data och algoritmer och alla resultat öppet tillgängliga då.

Life science, finns det andra tillämpningsområden?

Det andra stora området är mobiltelefoni, signalbehandling för telekommunikation. Har haft projekt kring navigering/positionering.

Den forskning som är mer datadriven, tar ni in extern data eller hur ser det ut?

I de projekt jag har drivit har vi tagit in data som generats på andra ställen. Ex i det här projektet med sekvensering, så har data genererats på Scilifelab och KI, vi har fått logga in till deras system och tagit hem den till vår dator för beräkningar för att det är smidigare. Celler i mikroskopi... Stanfordprojekten för länge sedan, då reste doktorander med hårddiskar. När man jobbar med känsliga data, då utnyttjar man KIs hanteringsrutiner, de har bättre hanteringsrutiner för känslig data och de är medvetna om det också, så då blir lösningen att man arbetar bakom deras brandväggar... De har traditionellt större behov att skydda känslig data. Jag är bihandledare då vi försöker få in en doktorand på KI, rör ambulanssjukvården, ska baseras på riktig data. Analys av EKG osv man labbar på öppet tillgänglig data, men sen när man ska utvärdera i kontexten av den data som faktiskt finns i regionen, då får det bli så att man flyttar programkoden till bakom deras brandväggar och använder deras system för hanterar det.

Har du samarbete med externa företag?

Större projekt med analys av immunologiska prover med ett företag i Stockholm, vaccintillverkning.

Använder du deras data då? Plockar ni in den?

Vi har plockat in den. De har delat data som inte är känslig. Men antikropparna och det som är kopplat till deras kunder är känsligt för dem, så det delar de inte med oss. Men det finns generiska dataset som inte är känsligt för dem.

Inga affärshemligheter?

Nej. Den skarpa datan, där kan det finnas. Det skulle vara känsligt. Man kan ta en modellcellinje som inte är kopplad till ngn person. De är inte känsliga när det gäller data. Inte kopplad till någon person, och bara göra ett sådant test för att få karaktäristiska bilder av den här tekniken eller från den här typen av teknik. Och de är inte vare sig det känsliga vad gäller datan. Det finns inga direkta affärshemligheter kopplade till dem. Det är sådan data som ligger ute, finns öppet på nätet..

Infosäk, hur jobbar ni med det inst?

Med persondata och sånt, eller att inte förlora data?

Allt. Hur kommer ni i kontakt med frågor om informationssäkerhet om alls, diskuterar ni det? Finns det några institutionsgemensamma arbetssätt eller lärosätets övergripande?

Nej, få. Som ansvarig för projekt, så är det klart att jag har de de krav som gäller att data finns tillgänglig efter avslutat projekt, att den inte försvinner. Det tar man allvarligt på. Vi har styrdokument på KTH, det är övergripande och inte så detaljerade så att de styr arbetet.

MSB vill att man ska jobba med riskanalys och infoklassning, kommer ni i kontakt med det?

Nej, inte formellt vad gäller riskklasser. Vi gör det informellt te x hur många kopior hårddiskarna ska finnas i. Det bygger ju på en idé om riskerna. Men vi för det inte formellt.

Ni diskuterar frågorna ändå. Tillgänglighet?

Ja just tillgänglighet diskuterar vi på avd och inst. Det är drivet av förväntningar. Allt fler publikationer kräver öppen källkod.

Använder ni Onedrive eller liknande för datahantering internt? Kan ni hantera behörigheter där?

Ja vi gick från Box tidigare till Onedrive. Jag kan dela mappar. Från min syn är det undermåligt med access-rättigheter. Inte vårt önskesystem. Hade varit smidigt om jag kunde mappa upp en gemensam mapp för ett projekt mot vår beräkningsdator istället för att behöva köra dubbla system.

Vet om det finns några andra sådana program som eller lösningar som kan hantera det?

Klassiska Linux-lösningar har sån funktionalitet.

Ger du behörighet till doktorander på Onedrive?

Jag har min onedrive på KTH och sen så mig via Office funktionaliteten skulle jag dela den mappen med en doktorand. Det är ju inte skapat för de datamängder vi arbetar med i de riktigt datadrivna projekten så där blir det ju så att vi gör en egen gemensamma med access på vår beräkningsdator som de doktorander som jobbar på just det projektet får tillgång till. Via den kan vi själva hantera åtkomsten då, och det funkar ju bra och det är ju snarare backup och säkerhet. Eller att se till att datan finns kvar, men att de inte kan förlora den som är det stora problemet. I mitt fall så är det en doktorand som är ansvarig för den beräkningsdatorn. Jag meddelar henne när hon ska sätta upp ett användarkonto åt någon som vill ha access till den beräkningsdatorn

Varifrån kommer den främsta delen av forskningsmedel?

VR, SSF, EU, Wallenberg WASP (finansierat doktoranderna). Våra kostnader är lönekostnader primärt. Internt i gruppen har vi en dator som kostar en halv miljon att köpa in, men det är väl små kostnader i förhållande till lönekostnaderna. Det är typiskt projekt för en doktorand är ju på 5 miljoner.

Hantering av fodata och – är det tydligt vilka skyldigheter och förväntningar som finns på dig som forskningsledare?

Nej överlag tycker jag inte att det är tydligt. Forskningsfinansiärerna är tydligare än KTH. Vi har våra styrdokument som finns att tillgå, vi har samordnare som Rosa. Men det saknas spridning av den informationen. Det är inte tydligt vad KTH som myndighet kräver av mig, men när man får finansiering ställs krav. Som kravet att man ska se till att datan är tillgänglig i 10 år. FAIR-principerna tydliga.

Om du har frågor kring infosäk? Vet du vem du ska vända dig till?

Jag skulle vända mig till It men upplever att det kanske inte är rätt. Kraven kring informationssäkerhet är framför allt kring att inte förlora data. Men även om jag alltid tänker på projekt där det skulle kunna involveras persondata och sånt så kommer ofta de kraven från oss så att vi förstår att det finns ett behov av att hantera det på ett bättre sätt och att man sedan snarare söker lösningar på de problemen än någon som kan berätta för en vilka krav som borde ställas. Det är något jag efterfrågar på KTH. Jag upplever inte att vi har den tjänsten. Jag skulle väldigt gärna vilja att om jag sätter upp ett forskningsprojekt och som här hanterar känslig data, att det skulle finnas en person att vända sig till som kunde tala om för mig att ja den typen av data ska klassas enligt den här skalan eller ska certifieras enligt den här standarden, och att det också åtföljdes av tekniska lösningar för X lagras kopplat till utbildningsresurser som man skulle kunna sätta deltagare i projektet på för att de ska få en grundläggande förståelse för vad man får och inte får göra.

Kommer du i kontakt med KTH:s jurister?

Inte vad gäller datalagring för fo.projekt. Som programansvarig för ett utbildningsprogram kommer jag i kontakt med persondata för studenter. Det kommer påbud från juristerna. Men gäller inte forskning. Men kommer i kontakt med juristerna när vi skriver avtal.

Är det standard att ha kontakt med jurister vid avtal?

Söker man pengar från Vetenskapsrådet till exempel, så söker jag tillstånd från avdelningschefen eller prefekten att KTH är villiga, att KTH stödjer ansökan då för det. Det är ju ansvaret att ta emot pengarna. Men det är mer formalia. De fall vi har kontakt med juristerna har varit dels kring den forskningsfinansierande projektet jag nämnde tidigare. Där var vi tvungna att reda ut som vad företagen skulle bekosta, så det var en ganska utdragen process... Kopplat till industridoktorandprojekt, så har det varit en hel del frågor kring vad som gäller kring de doktoranderna då om till exempel företaget går i konkurs eller väljer att säga upp personen, vem har då ansvar för att fortsatt finansiering och sånt? Och där skrivs det ju kontrakt, där har jag varit i kontakt med juristerna.

Är du bekant med PDA? Blir det aktuellt? Tekniskt bistånd.

I liten utsträckning. Det är ju så signalbehandling, kommunikation. Det är klart att det används i krigföring också om man tar till exempel positioneringen, teknologi för att sensorfusion aktier och metrar och gyroskop och sådant används i för missiler och alltså krigsmaterial. Kopplingarna finns ju i den forskning vi bedriver. Men matematisk grundforskning sker väldigt öppet, hela forskningsområdet publicerar ju öppna journaler som är tillgängliga för alla

Det är inte riktigt så att det upplevs som hemligt. Nu är det på tal om CSC-studenter. Utbildning av studenter som går tillbaka till sina universitet och samarbetar med krigsmakten. För den här beräkningsdatorn som vi har på avdelningen så var jag tvungen att skriva på att vi inte utnyttjar den i utveckling av

krigsmaterial. Men det var ju exportkrav från X som är företaget som har sin bas i USA.

Det finns ju en samordnare eller kontaktperson på KTH för produkter med dubbla användningsområden. Har du kommit i kontakt med honom någonting?

Nej.

De krav som dyker upp kommer ofta internt sa du, eller från finansiärerna. Vad kan det vara för saker som dyker upp när det gäller infosäk?

Handlar om backup, att inte förlora data. Hårddiskkrascher. Rådata finns också på SciLife lab. Stora summor i värdet av användningen av utrustningen. Förlorad tid. Värdet av doktoranders tid.

Använder ni er av datahanteringsplaner i dina projekt?

I alla projekt haft vi diskussioner men inte formellt skrivit ngn.

VR har krav?

Jo men jag fick senast 2016 -2017.

Hur gör du bedömningen om du har känslig data?

Jag skulle vilja svara sunt förnuft. För mig har det varit persondata och var inte persondata.

Incidenter? Tex en hårddisk som kraschar, är det en incident?

Då säker jag hjälp från it. Men jag har inte rapporterat det som en incident. Sökt hjälp i forskargruppen, inköpsansvariga

Kan ni köpa vilka datorer som helst?

Vi har upphandlingsregler. Men det är så smal bransch att det finns bara en dator.

Internationellt samarbete, koppling till forskningsdata?

I samarbete har det varit muskelceller t ex med Stanford. De hade inte gillat att vi skulle göra data publik, men inga andra föreskrifter. Det hade varit pinsamt och slösaktigt av oss och partner om vi hade tagit emot datan och sen förlorat den och behövt be om den igen och inte tagit vara på den.

Utöver det du sagt, saknar du ngt för att kunna bedriva ett bra arbete?

En del infrastruktur, rutiner och utbildning. Jag har önskemål om att engagera mig i projekt med känslig data, och där ser jag behov av en infrastruktur med access rättigheter som man kan kontrollera som ansvarig forskare för projektet, men också någon form av support i termer av internutbildning av de som skulle ingå. Och så ingår prisvärd storskalig lagring.

Nåt annat du vill tillägga?

Nej.

Har vi sagt ngt känsligt som vi inte ska anteckna?

Nej. Jag har försökt vara någorlunda balanserad. Jag pratar mycket internt, jag pratar med Rosa. Jag har varit inbjuden av henne på konferenser som syftat till att framföra våra behov som forskare. Så hon har intervjuat mig som forskare. Men det är helt okej också om ni skriver forskare önskar bättre system som uppfyller de här kraven så kommer hon kunna gissa att det är jag som har sagt i alla fall. Det är inte jättehemligt, inte jättekänsligt heller.

RIKSREVISIONEN

Intervju KTH18 230322

Från RiR: Ludvig Stendahl och Sara Monaco

Prof, datadriven i forskningen, maskininlärning, stora datamängder, genererar mkt data, jag är ju definitivt en av dem som har högre behov av infrastruktur kring data.

Den vanliga rollen som professor: söker projekt, driver projekt, drivs primärt av doktorander. Vi tar in ibland exjobbare och andra forskare i dem också då som sagt, men det är min roll att driva projekt. Jag är ansvarig för KTH strategiska samarbete med Region Stockholm och i den rollen så har jag ju sett behov på KTH av datahantering som kanske sträcker sig utanför min forskning. Jag har också varit ganska aktiv och försöka driva datafrågor på KTH varit i kontakt med Rosa bland annat för att ge input på vad jag anser behövs.

Inom kommunikation har vi ganska standardiserade datamodeller så det är mycket simulerade data. Det är väldigt lite riktig data i den meningen. Inom life science tillämpningarna så är det ju mycket mer riktig data i den mening att de uppmätt kring någon process. Jag har själv hållit mig ifrån än så länge persondata, så när det har varit data har ju varit till exempel mikroskopi data från celler som man har odlat i något experiment. I de fallen har det ju varit celler som någon gång har kommit från en människa. Man har arbetat med cellinjer som är frånkopplade från individer. Vi arbetar med genetik data nu då, men då att vi har metodutveckling så jobbar vi inte med mänsklig det mänskliga genomet. Det finns ingen poäng med det. E coli bakterier som jag sekvenseras och genererat ganska stora datamängder. Men det rör sig inte om persondata och känslig data.

Vad tillhandahåller KTH vad gäller infrastruktur, saknar du ngt?

Beräkningsinfrastruktur för beräkningar, vi har ju PD parallell dator centrum, till exempel för beräkningar. Jag utnyttjar också nationell infrastruktur i så det bland annat finns det infrastruktur som är finansierade av oss programmet eller Familjen Wallenberg. Berzelius i Linköping som en del av den nationella infrastrukturen för beräkningar och just de datorerna lämpar sig mer för den typen av algoritmer som vi testar.

Jag ser behov av den långsiktiga lagringen. Nuvarande projekt: metoder för DNA-sekvensering. Samarbete med SciLifeLab har hjälpt oss sekvensering i storleksordningen ett par 100GB terrabyte, som vi testar våra algoritmer på. Vi har behov av att säkerställa att vi inte tappar den datan som vi har lokalt på beräkningsdatorer här. Kraschar de hårddiskarna finns en risk att arbetstid och datortid går förlorad. Sådana behov går utanför vad KTH kan tillhandahålla. Vi har