# Topics in Data Science = Өгөгдлийн шинжлэх ухаан

**Reproducible Research – Нөхөн сэргээгдэх судалгаа**

Монгол Улсын Үндэсний статистикийн хороо, Улаанбаатар хот, Монгол Улс, 2017 оны 5 сарын 9

Ryan Womack

Data Librarian, Rutgers University, https://ryanwomack.com

**Хэсэг 1: Танилцуулга**

**Илтгэгчийн танилцуулга**

• Data Librarian

• Эдийн засаг, номын сангийн шинжлэх ухаан, статистикийн магистрийн зэрэгтэй

- Эдгээр салбарын огтлолцол дээр ажилладаг
- Өгөгдөл хайх, ашиглахад нь хүмүүст тусалдаг
- Номын санд зориулан мэдээллийг удирддаг

• Ахисан түвшний эрдэм шинжилгээ, судалгаан дээр ажиллах

**Rutgers-ийн талаарх танилцуулга**

• Rutgers, The State University of New Jersey

• New Jersey, 8.8 сая хүн амтай, New York болон Philadelphia-ийн хооронд оршдог

• Rutgers 1766 онд байгуулагдсан, 250 гаруй жилийн түүхтэй

• Carnegie ангилал: Судалгаа – Судалгааны ажлыг маш өндөр түвшинд хийдэг, дээд түвшний ангилал (R1)

• 100 гаран судалгааны хөтөлбөртэй

• 68,000 гаруй оюутантай (120 гаруй улсын 7,500 олон улсын оюутантай)

• Дэлхийн топ 100 их сургуулийн нэг (Times Higher Education, Shanghai Ranking, CWUR болон бусад)

**Rutgers-ийн талаарх танилцуулга**

Зураг

**IASSIST-ийн талаарх танилцуулга**

• IASSIST – Нийгмийн шинжлэх ухааны мэдээллийн үйлчилгээ, технологийн олон улсын нийгэмлэг

• Мэдээллийн технологи, номын сан, өгөгдлийн үйлчилгээ, судалгаа болон дээд боловсрол, төрийн, ашгийн бус болон хувийн судалгааны салбарт ажиллаж буй 300 гаруй өгөгдлийн мэргэжилтнүүдийн нэгдэл

• IASSIST дараахь зорилгын хүрээнд ажилладаг

- Өгөгдлийн үйлчилгээг хүргэх төгс сүлжээг дэмжих сурталчлах
- Нийгмийн шинжлэх ухааны дэд бүтцийг сайжруулах

- Мэргэжлийн туршлага солилцох боломжийг бүрдүүлэх

• Жил бүрийн конференци нь харилцан туршлага судлах, харилцаа холбоо үүсгэх фоум болдог (АНУ, Канад, Европ, дараагийнх Азид?)

• IASSIST Ази дахь гишүүнчлэлээ нэмэгдүүлэхийг зорьж байгаа.

**Нөхөн сэргээгдэх байдал: Үүгээр бид юуг хэлж байгаа вэ?**

• Шинжлэх ухааны итгэл үнэмшилтэй байдал

– Шинжлэх ухааны луйвар нэмэгдэж байгаа

• Duke “starter set” болон нийтлэл

– Судалгааны сахилга бат зөрчих нь асуудал, гэхдээ хүний л алдаа

• Нэр хүнд, санхүүжилт зэрэг бүх зүйл нөлөөлж байна

– "Нээлттэй хэлбэрт тохируулах"

• Давталт (гарааны шугамнаас туршилтыг дахин хийх) нь өртөг өндөртэй, цаг хугацаа өнгөрсөн байдгаас шалтгаалан боломжгүй байж болно (Баталгаажулалтыг харна уу)

• Давталт болон нөхөн сэргээлтийн шинжлэх ухаан

• Victoria Stodden, Friedrich Leisch, and Roger D. Peng (eds.). Implementing Reproducible Research. CRC Press, 2014.

“Компьютер ашигласан бүхий л судалгаанд судлаачдын зааварчилгаа нь програм хангамжаар хадгалагддаг бөгөөд шинжлэх ухааны өгөгдөл нь тоон хэлбэрээр хадгалагддаг. Тооцооллын судалгаан дахь ердийн хэвлэмэл бүтээгдэхүүнүүд өгөгдөл дээр суурилдаг бөгөөд компьютерийн зааварчилгаа нь шинжлэх ухааны дүгнэлтийг хийх өгөгдөлд ашиглагддаг. Өгөгдлийг үүсгэх механизм болон тооцооллын зааварчилгааны нарийн төвөгтэй байдал нь уламжлалт шинжлэх ухааны хэвлэлийн хувьд хийхэд ихэнхдээ маш том, хэт том байдаг. Судалгааны процесст компьютер оролцдог хэдий ч шинжлэх ухааны хэвлэл нь шинжлэх ухааны нийтлэлээс шинжлэх ухааны бичлэг, програм хангамж болон дүгнэлтийг хийсэн өгөгдөл гэсэн гурвалсан хэлбэрт шилждэг. Энэхүү гурвалыг “судалгааны конпендиум” гэж нэрлэж ирсэн бөгөөд түүний зорилго нь салбарын бусад судлаачид өгөгдлийг програм хангамжийг ажиллуулах замаар дахин бий болгож болох судалгааны үр дүнг хувиргах явдал юм. Гэхдээ дүгнэлтийг бусад судлаачид дахин хийж болох боломжийг олгохоор өгөгдөл болон програм хангамж нь бэлэн байх ёстой.”

– Victoria Stodden - http://blog.stodden.net/2014/09/28/my-input-for-theostp-rfi-on-reproducibility/

**Нийгмийн шинжлэх ухааны загварууд**

• ICPSR сайтар боловсруулсан архивлах загвар болон codebook болон metadata-гаар өгөгдлийг баримтжуулах чиглэлээрх 50 гаруй жилийн хугацаанд ажиллаж ирсэн

• IPUMS нь микро өгөгдлийн олон улсын харьцуулалтыг хийж болох олон арван жилийн хооронд болон ялгаатай төслүүдийн хооронд харьцуулалт хийх боломжийг олгохын тулд өгөгдлийг дахин загварчлах үйл явц юм [Монгол: 1989 он болон 2000 оны өгөгдөл]

• Нийгмийн шинжлэх ухааны салбар урт хугацаандаа тогтмол биш хэдий ч сайтар зохион байгуулсан загварчлалууд хийгдсээр ирсэн

• Ижил төрлийн тооцооллын техник байгалийн шинжлэх ухаан болон дижитал хүмүүнлэгийн салбарт ашиглагдаж болох тохиолдолд шинжлэх ухааны салбар хоорондын ялгаа багассаар байна

**Агуулга**

• Бид тодорхой нөхцөл дэх зарим туршлагын талаар авч үзнэ

– хувь судлаач

– судалгааны баг

– хэрэгжиж буй томоохон хэмжээний хамтын ажиллагаа

**Хэсэг 2: Хувь хүмүүс (Бүгд)**

Өгөгдөл өөрөө

Зарим үндсэн тохиолдлууд:

• Түүхий өгөгдлөө анхны хэлбэрээр нь хадгалах, аливаа ажлын өгөгдлөөс тусгаарлах

• Хувьсагчид болон мэдээлэл цуглуулалтаа баримтжуулах

– гарсан асуултад хариу өгөхийн тулд 3 жилийн дараа төсөл рүү эргэн ороход та өөрөө мартсан байж болно

– энэ нь бусад судлаачдын хувьд мөн адил!

• Excel дээр [хэрэв чадвал] эсвэл бусад гар засвар хийдэг орчинд бүү ажилла

– үүнийг хийж байгаа бол та алхам бүрээ бичиж тэмдэглэх хэрэгтэй

– кодчилол эсвэл таны үйлдлүүдийг бүртгэх орчинг ашиглах нь илүү сайн

**DOI**

DOI, Digital Object identifier буюу дижитал объект тодорхойлогч нь маш амжилттай болсон нэг жишээ юм

• өөрчлөгдөхгүй ишлэл, зүүлт хийхэд хялбар болгосон

• ихэвчлэн чанарын өгөгдлийн сантай холбогддог

• олон сайн зүйлсийг өөртөө багтаасан

• ёс зүй: тодорхойлсон стандарт болон төвлөрсөн арга хэрэгслийг хэрэглэж ашиглахад хялбар болгосон

•DOI-д өөрийн өгөгдлийг оруулсан бол түүнтэйгээ харьцах боломжтой бөгөөд нөхөн сэргээж болно

**Код/Өгөгдөл/Баримтын нэгтгэл**

Бид R-ийн жишээн дээр ярилцах болно, гэхдээ бусад програмчлалын орчинд (Mathematica notebooks, iPython/Jupyter) үүнийг мөн дэмжиж ажилладаг

• анхлаад LaTeX + Sweave дээр ажилладаг байсан

• can embed R code-ийг суурилуулж болно, түүнийг баримт бичгийг бий болгоход ажиллуулж болно

• Текстгэй “холилдсон” кодыг гаргаж авч болох бөгөөд форматчилагдсан баримт баримт бичиг нь literate program-тай “сүлжилдсэн” байж болно.

• хамгийн сүүлийн өгөгдөл, үр дүнг авахыг ихэнхдээ зорьдог

• тухайн нөхцөл байдалд кодыг баримтжуулах, тайлбарлахад тусалдаг (literate programming)

• PDF, document, HTML форматуудыг гарган авахад хялбар

**knitr болон Markdown**

• Markdown + knitr нийтлэг ашиглагдаж байгаа бөгөөд өөрчлөлтийг хялбар хийж байна

• Энгийн syntax болон хэрэгжүүлэлт

• RStudio-той нэгдсэн

• RPubs дээрх баримт бичгийг нэг товчлуур дараад хэвлэнэ

• Илүү нарийн төвөгтэй баримт бичгийн форматын хувьд LaTeX/Sweave дээр буцааж болно

**Бусад баримт бичиг / програмын хэрэгслүүд**

• notedown нь Rmd-г ipython notebook-т хувиргаж, эргүүлж мөн хувиргадаг

• bookdown нь Rmd дээр ном бичихэд ашиглагддаг

• Zeppelin браузер дээр R-ийг интерактив байдлаар ажилладаг, мөн Apache Spark-тай ажиллаж болдог

• R notebooks нь одоо RStudio дээр хийгдэж байгаа

• Онлайн notebooks нь интерактив визуальчлалийг холбох хялбар арга зам болдог

**Богино хэмжээний Rmd файл**

— title: "A short literate programming exercise" author: "Ryan Womack" date: "October 10, 2016" output: pdf_document —

```{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE) ```

## Read in the data

Өгөгдлийг дараахь командаар уншицгаая:
```

```{r load} library(readxl)

download.file("https://ryanwomack.com/data/PharmaDemo.xls", "mydata.xls")

mydata<-read_excel("mydata.xls")

names(mydata)

attach(mydata)

```

## Describe the Data

Ингэснээр нас, Жингийн хувьсагчийн зарим хураангуй статистикийг бид гарган авах болно:

```{r summary} summary(Age)

summary(Weight) ```

Одоо өгөгдлөөр график байгуулъя:

```{r plot, echo=FALSE} library(ggplot2)

ggplot(mydata, aes(Weight, Age))+ geom_point() ```

## Regression

```{r regression} summary(lm(Age~Weight))

ggplot(mydata, aes(Weight, Age))+ geom_point()+ stat_smooth() ```

Ингээд боллоо!

**Тооцооллын орчин**

• Open source нь нөхөн сэргээгдэх байдлын чухал нөхцөл болдог

• Хэн ч гэсэн ажиллуулах програм хангамжийн хуулбарыг авч болдог

• Тохиромжтой байдлыг хангахын тулд өмнөх хувилбарыг авч болдог

• Та өөрийн тооцооллын орчны талаарх мэдээллийг мөн бүртгэж болно (session.Info() in R)

• Checkpoint package нь R дээр энэ процессыг автоматаар хийдэг

**Бусад асуудлууд**

• Output-аа хадгалж болохгүй. Энэ нь хаанаас гарч ирэх вэ? Энэ нь кодоор хийгдсэн байх ёстой.

• Сайтар форматласан өгөгдөл (tidyr) дээр засвар хийнэ, код нь (formatR) нэмэлтээр ашиглагдана

• Хэрэв “readme” файлын аргыг ашиглаж байгаа бол бүх зүйлийг баримтжуулна

**Бусад платформууд**

• Jupyter нь iPython-оос үүсэн гарсан
```

– Одоогийн байдлаар 40 орчим хэлийг дэмжиж ажилладаг

• Mathematica Notebooks, cloud support

• Cloud үйлчилгээ нь шэйр хийх үйлдлийг ихээхэн хөнгөвчилж байгаа

• Хүлээлт үүсгэж байгаа

**Хэсэг 3: Хамтын ажиллагаа**

**Хамтын ажиллагаа**

• Ижил нөхцөл (үүлэн тооцоолол, shared platforms, стандартууд) хамтын ажиллагааг илүү хялбар болгож байгаа

• Github, Bitbucket, бусад зүйлс нь програмчлал дахь хамтын ажиллагааг хялбарчилж байна

– Код нь бэлэн байснаар нөхөн сэргээлтийн үр ашигт ихээхэн чухал нөлөөг үзүүлж байна

• Open Science Framework өгөгдөлд суурилсан тусгай арга (data-specific approach)-ыг бий болгодог

• Гол онцлог нь ижил платформыг хувийн ажилдаа ашиглаад дараа нь нийтэд ашиглуулдаг явдал юм

• Сэтгэл судлалын нөхөн сэргээгдэх судалгаа нь OSF-г ашигладаг.

– Эхлэхийн өмнө шинжлэх ухааны нийтлэлүүдийг үзнэ үү

**Хамтын ажиллагааг бий болгох**

• Хамтын ажиллагааны төслүүд дараахь нөхцлийг хангасан байх ёстой/хэрэгтэй:

– data practices болон sharing platform

– coding practices болон sharing platform

• Үүнийг одоогийн байгаа дараахь платформ болон практикаар хялбархан хийдэг болсон

– ropensci

– ProjectTemplate

– Өгөгдөл болон код package.skeleton() командаар эхлэх пакетаар хуваарилагдаж болно

– Нөхөн сэргээгдэх шинжлэх ухааны Reprozip болон бусад

**Yale ISPS**

Yale-ийн Нийгмийн болон бодлогын судалгааны институт нь нөхөн сэргээгдэх байдлыг дэмжих судалгааны багийн нэг жишээ юм.

• Өгөгдөл болон баримт бичиг нь хамтдаа нэг талбарт архивлагддаг

• Өгөгдөл (DOIs биш)-ийг удирддаг

• Код болон баримтжуулалт нь архивлагддаг

• Алдаагүй ажиллагааны кодын хяналт

• Судалгааны тайлбарлах мета өгөгдлийг бий болгох сайн жишээ

• Институт тэтгэлэгийн нөхцөл байдлаар дагаж мөрдөхийг шаарддаг учраас хэрэгжих боломжтой

**Нээлттэй өгөгдөл болон хувилах үйл явц**

• Нээлттэй өгөгдөл нь олж илрүүлэх, хамтран ашиглах, дахин ашиглах боломжийг олгодог

• (Нээлттэй өгөгдлийг мөн ихэвчлэн төрийн өгөгдөл гэж тодорхойлдог)

• Судалгааны өгөгдлийг агуулах орон зай

• Many layers of Open

• Big Data World дахь Нээлттэй өгөгдөл

• Хувилах үйл явц (Нээлттэй өгөгдөлд суурилсан), шинжлэх ухааны нотолж, баталгаажуулж өгдөг

• Хувилах үйл явц Wiki

**Хэсэг 4: Багийн шинжлэх ухаан**

**Том шинжлэх ухаан**

Анхны төслийн хэрэгжилт нь цаашаа үргэлжлэх шаардлагатай болж болох олон жилийн туршид үргэлжилсэн олон байгууллага оролцсон төслүүд дахин сэргээх шаардлагатай болдог.

• Олон хүн цаг хугцаны туршид тухайн төсөл дээр шинээр ажиллаж эсвэл гарч болно

• Урьдчилан таамаглаагүй олон хэрэглээ урьдчилан төлөвлөгдсөн байдаг (“мэдэхгүйгээ мэдэх” эсвэл үүнтэй төстэй)

• Хамтын ажиллагаа, тасралтгүй байдлыг ухамсартайгаар төлөвлөсөн байх хэрэгтэй

• Шийдвэрийг одоогийн тохиромтой байдлаас илүүтэй ирээдүйн хэрэглээг харсан байдлаар гаргах хэрэгтэй

• Гэхдээ салбарын мэргэжилтнүүд энэ чиглэлд бэлтгэгдсэн байдаг

• Уургийн өгөгдлийн банк (Protein Data Bank)

• Мөн, https://www.teamsciencetoolkit.cancer.gov/

**Том шинжлэх ухаан – Дэд бүтэц**

• Том шинжлэх ухаанд гол зангилаа нь ирээдүйд дахин ашиглалтыг дэмждэг байх явдал юм

• Тэд суурь дэд бүтцийг бий болгодог (OOI - Ocean Observatories Initiative)

• Гэхдээ гол зангилаатай холбогдсон, бий болсон бусад төслийн нэгдсэн төвийг мөн бий болгонсон байдаг

• Ирээдүйн нэг гол зорилго нь илүү нийтлэг, бүх төрлийн зорилго бүхий хамтын ажиллагааны дэд бүтцийг бий болгох явдал юм

• Rutgers энэ зорилгоор Virtual Data Collaboratory-г хөгжүүлж байна

• Zenodo, Dataverse, OpenICSPR болон Open Science Data Cloud зэрэг нээлттэй дэд бүтцийг хөгжүүлсээр ирсэн

**Том шинжлэх ухаан – Практик үйл ажиллагаа**

• Standards such as DOI болон ORCID гэх мэт стандартууд сайн туршлагад оролцогчдыг өргөн хүрээнд татан оруулах боломжийг олгодог

• Өгөгдлийн агуулахууд стандартыг хөгжүүлдэг

• Re3data.org бол агуулахуудын жагсаалт юм

• Data Seal of Approval нь боломжийн өгөгдлийн практикийг ашигласан агуулахуудад олгогддог

• ISO 16363 (Итгэмжлэгдсэн дижитал агуулахууд - Trusted Digital Repositories) нь илүү хатуу стандарт юм

• Нэг чухал алхам нь төсөл зогсох үед (санаатайгаар эсвэл санамсаргүйгээр) яах ёстойг төлөвлөх явдал юм

• Тооцооллын болон нөхөн сэргээгдэх байдлын ижил түвшний ямар стандарт байна вэ?

**Том шинжлэх ухаан – Генийн өгөгдлийг хамтран ашиглах**

• Их хэмжээний хөрөнгө оруулалт их хэмжээний өгөгдөл

• Олон агуулахууд мөн хэрэглэгддэг (NIH GDS - National Institutes of Health Genomic Data Sharing)

• Одоогийн агуулахууд олон програм хангамжийн хэрэглүүрт хэрэгтэй бөгөөд тэдгээрийг нэгтгэдэг

• Гэхдээ судлаачид бүүр том өгөгдлийн санг, хүний генийн өгөгдлийн сантай байхыг хүсдэг

• Техникийн асуудлууд нь нарийн төвөгтэй хэдий ч эрх, зөвшөөрөл нь мөн адил нарийн төвөгтэй байдаг

• Хэрхэн өгөгдлийг хамгийн дээд түвшинд олж илрүүлж болохуйцаар нэгтгэж болох вэ?

**Виртуал Өгөгдлийн хамтын ажиллагаа**

• Их хэмжээний өгөгдлийг хамтран ашиглах, дүн шинжилгээ хийх, дахин ашиглах, олж илрүүлэх, хамтран ажиллах боломжийг олгох виртуал орчин [Rutgers and Penn State leads]

Зураг:

**Хэсэг 5: Монгол Улс**

**Монгол дахь нээлттэй өгөгдөл**

• Чингис хаан – 1206 онд анхны тооллогыг хийж байсан

• Монгол Улс OpenDataWatch-ийн мэдээллээр Төрийн өгөгдлийн нээлттэй байдлаар 3 дугаар байранд (125 орноос) эрэмбэлэгдэж байна.

• Монгол Улсын нээллтэй өгөгдлийн санаачлага (Open Data Initiative of Mongolia)

• МУИС болон МТҮП Нээлттэй өгөгдөл дээр хамтран ажиллаж байна

• Монгол Улсын GIS өгөгдлийн сан

• Өгөгдөл хамтран ашиглах, нөхөн сэргээгдэх судалгаа хийх суурь сайн тавигджээ!

**Хэсэг 6: Дүгнэлт**

**Дүгнэлт**

• Нээлттэй байдлыг нэмэгдүүлэх нь урт хугацааны хандлаг болж байна

– Интернэт, Төр засаг, Өгөгдөл, Програм хангамж, Үүлэн тооцоолол (Cloud Computing)

• Нөхөн сэргээгдэх судалгааны шахалт дэлхий нийтийн хандлагын дагуу нэмэгдэж байна

• Сайн мэдээ нь…

– Үр ашиг нь нийгэмд болон мэдлэг бий болгоход тодорхой харагдаж байна!

– Үүнийг хийх арга хэрэгсэл нь үргэлж хялбар болсоор байна!

– Эцэст нь энэ нь биднийг харуулах стандарт болох болно!

**Төгсгөл**

• Асуулт байна уу?

• Маш их баярлалаа!

• Намайг дагаарай…

– https://youtube.com/librarianwomack

– https://www.linkedin.com/in/ryanwomack

– https://twitter.com/ryandata

– https://ryandata.wordpress.com