# Machine Learning in Toxicology: Towards Intelligent Access to the Content of Research

SACATM Meeting, September 2-3, 2020
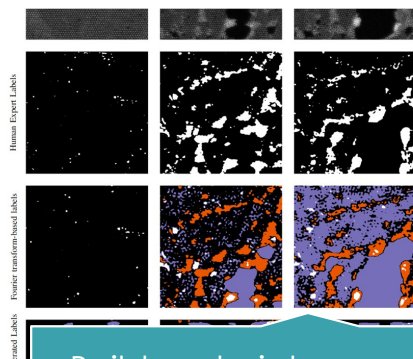
**Robert M. Patton, Ph.D.**

Computer Science and Mathematics Division

Oak Ridge National Laboratory

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

# Computational Data Analytics

**Core focus:** developing innovative analytical methods for understanding large and complex datasets



**Build materials one atom as a time**

Patton, R.M., Johnston, J.T., Young, S.R., Schuman, C.D., March, D.D., Potok, T.E., Rose, D.C., Lim, S.H., Karnowski, T.P., Ziatdinov, M.A. and Kalinin, S.V.,**167-PFlops deep learning for electron microscopy: from learning physics to atomic manipulation**. *SC18* (p. 50). IEEE Press.
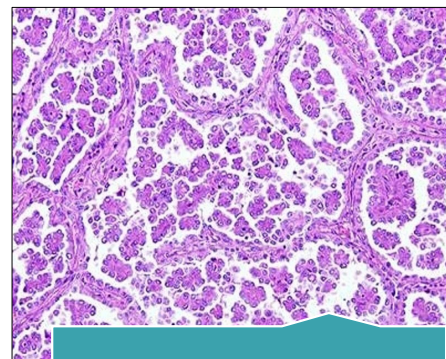


**Fight cancer**

Patton, R.M., Johnston, J.T., Young, S.R., Schuman, C.D., Potok, T.E., Rose, D.C., Lim, S.H., Chae, J., Hou, L., Abousamra, S. and Samaras, D., 2019. **Exascale Deep Learning to Accelerate Cancer Research**. *IEEE Big Data (2019)*
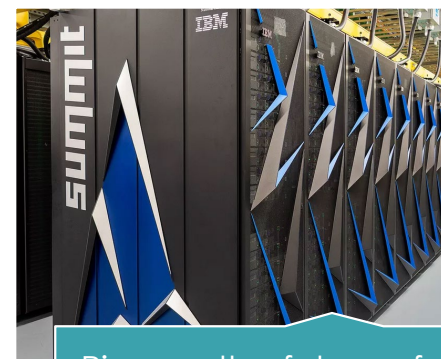


**Pioneer the future of computing**

T. Potok, C. Schuman, S. Young, R. Patton, F. Spedalieri, J. Liu, K.-T. Yao, G.S. Rose, and G. Chakma, "**A Study of Complex Deep Learning Networks on High Performance, Neuromorphic, and Quantum Computers**," ACM Journal of Emerging Technologies in Computing Systems, vol. 14, no. 2, July 2018.
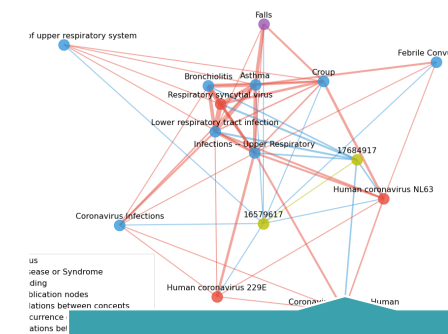


**Accelerate biomedical discovery**

R. Kannan, P. Sao, H. Lu, D. Herrmannova, R. Patton, T. Potok, V. Thakkar, R. Vuduc, "**Scalable Knowledge-Graph Analytics at 136 Petaflops/s**," Gordon Bell Submission, Under Review, April 2020.

**OAK RIDGE**
National Laboratory

# Supporting systematic review data extraction

# Goal

- Goal: automated **identification/extraction of study descriptors from research publications**
  - Information pertaining to **guideline for rodent uterotrophic bioassays** (OECD TG 440)
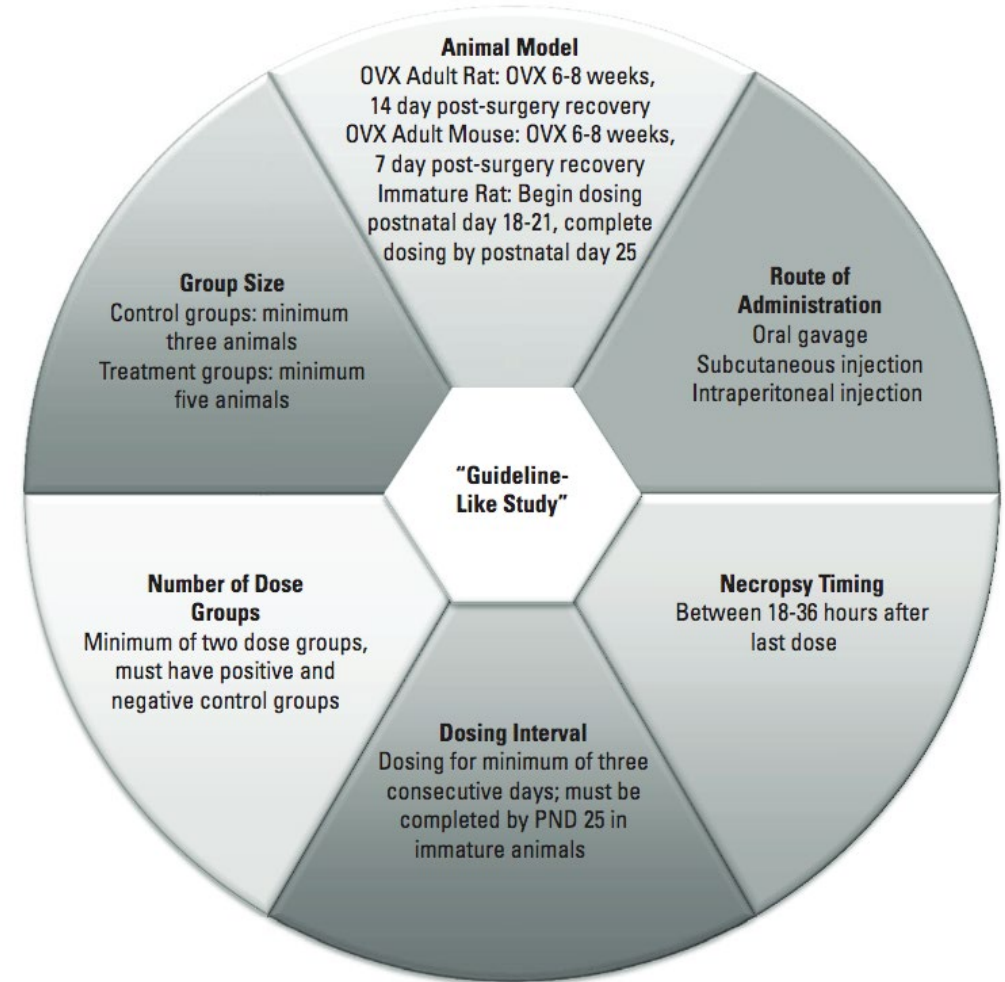
# The Data Being Extracted

- OECD Test Guideline 440: Uterotrophic Bioassay in Rodents
  - The guideline consists of **6 minimal criteria (MC)**
  - **All six MC have to be met** for a study to be **guideline-like (GL)**
  - Below: manually annotated (GL) abstract showing the six MC

The intact female weanling version in the Organization for Economic Cooperation and Development (OECD) uterotrophic assay Test Guideline (TG) 440 is proposed as an alternative to the adult ovariectomized female version, because it does not involve surgical intervention (vs the ovariectomized version) and detects direct/indirect-acting estrogenic/anti-estrogenic substances (vs the ovariectomized version which detects only direct-acting estrogenic/antiestrogenic substances binding to the estrogen receptor). This validation study followed OECD TG 440, with **six** [GS] female weanling **rats (postnatal day 21)** [GS] **per dose group** and **six treatment groups** [group size (GS)] [no. of dose groups]. [animal model] Females were weighed and dosed once daily by **oral gavage** [admin. route] for **three consecutive days** [dosing interval], with one of six doses of $17\alpha$-ethinyl estradiol in corn oil at 5 ml kg$^{-1}$ at **0 and 0.1–10 $\mu$g kg$^{-1}$ per day** [dose groups (neg. and pos. control)]. On **postnatal day 24** [necropsy timing], the juvenile females were euthanized by $CO_2$ asphyxiation, weighed, livers weighed and uteri weighed wet and blotted. The presence or absence of vaginal patency was recorded. Absolute and relative (to terminal body weight) uterine wet and blotted weights and uterine luminal fluid weights were significantly increased at 3.0 and 10.0 (both P < 0.01) $\mu$g kg$^{-1}$ per day, and increased to ~140% of control values at 1.0 $\mu$g kg$^{-1}$ per day (not statistically significantly). In vivo body weights, weight changes, feed consumption, liver weights and terminal body weights were unaffected. Vaginal patency was not acquired in any female at any dose, although vaginal puckering was observed in one female at 10.0 $\mu$g kg$^{-1}$ per day. Therefore, this intact weanling uterotrophic assay is validated in our laboratory for use under US and European endocrine toxicity testing programs/legislation.
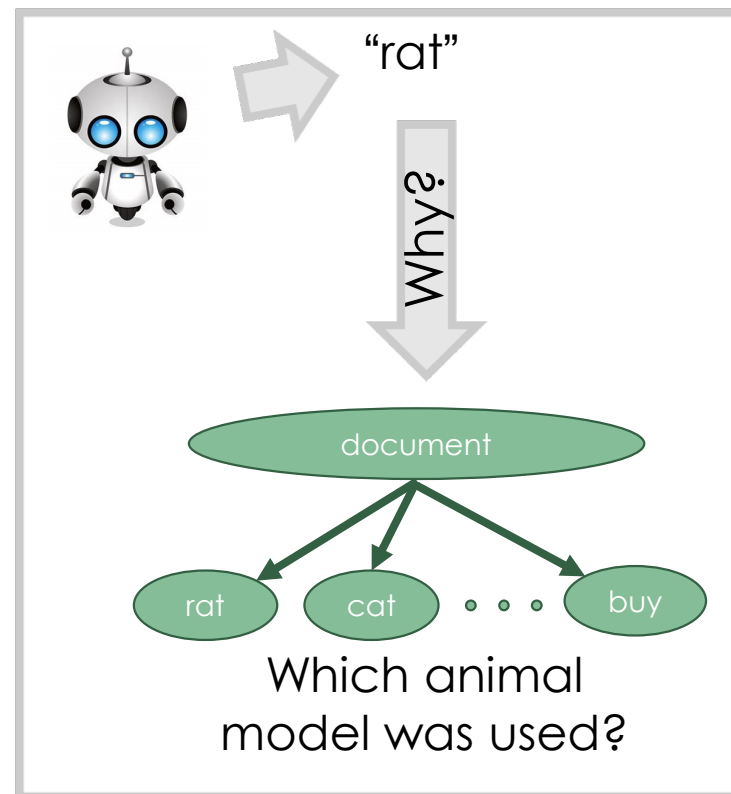
# OECD TG 440: Minimum criteria (MC)

- MC 1: Animal model
- MC 2: Group size
- MC 3: Route of administration
- MC 4: Number of dose groups
- MC 5: Dosing interval
- MC 6: Necropsy timing



**Animal Model**
OVX Adult Rat: OVX 6-8 weeks, 14 day post-surgery recovery
OVX Adult Mouse: OVX 6-8 weeks, 7 day post-surgery recovery
Immature Rat: Begin dosing postnatal day 18-21, complete dosing by postnatal day 25

**Route of Administration**
Oral gavage
Subcutaneous injection
Intraperitoneal injection

**Group Size**
Control groups: minimum three animals
Treatment groups: minimum five animals

"Guideline-Like Study"

**Necropsy Timing**
Between 18-36 hours after last dose

**Number of Dose Groups**
Minimum of two dose groups, must have positive and negative control groups

**Dosing Interval**
Dosing for minimum of three consecutive days; must be completed by PND 25 in immature animals

Source: Kleinstreuer et al. (2016). *A Curated Database of Rodent Uterotrophic Bioactivity.*

OAK RIDGE
National Laboratory

# Human Learning vs. Machine Learning

- A human non-expert can **correctly annotate criteria** given just a **description of the target information**

The intact female weanling version in the Organization for Economic Cooperation and Development (OECD) uterotrophic assay Test Guideline (TG) 440 is proposed as an alternative to the adult ovariectomized female version, because it does not involve surgical intervention (vs the ovariectomized version) and detects direct/indirect-acting estrogenic/anti-estrogenic substances (vs the ovariectomized version which detects only direct-acting estrogenic/antiestrogenic substances binding to the estrogen receptor). This validation study followed OECD TG 440, with six [GS] female weanling rats (postnatal day 21) [GS] per dose group [group size (GS)] and six treatment groups [no. of dose groups]. Females were weighed and dosed once daily by oral gavage [admin. route] for three consecutive days [dosing interval], with one of six doses of 17α-ethinyl estradiol in corn oil at 5 ml kg$^{-1}$ at 0 and 0.1–10 $\mu$g kg$^{-1}$ per day [dose groups (neg. and pos. control)]. On postnatal day 24, the juvenile females were euthanized by CO$_2$ asphyxiation, weighed, livers weighed and uteri weighed wet and blotted [necropsy timing]. The presence or absence of vaginal patency was recorded. Absolute and relative (to terminal body weight) uterine wet and blotted weights and uterine luminal fluid weights were significantly increased at 3.0 and 10.0 (both P < 0.01) $\mu$g kg$^{-1}$ per day, and increased to ~140% of control values at 1.0 $\mu$g kg$^{-1}$ per day (not statistically significantly). In vivo body weights, weight changes, feed consumption, liver weights and terminal body weights were unaffected. Vaginal patency was not acquired in any female at any dose, although vaginal puckering was observed in one female at 10.0 $\mu$g kg$^{-1}$ per day. Therefore, ...nling uterotrophic assay is validated in our laboratory for use under US and European endocrine toxicity testing programs/legislation.

# Our approach

- Goal: **extract information for the purpose of identifying GL/non-GL documents**

- Two approaches:

  1. **Classify, then extract:** Can we train a good classifier? If so, extract learned patterns (fully supervised)

  2. **Extract, then classify:** Extract text segments relevant to criteria descriptions (unsupervised), then classify the extracted segments



"rat"

Why?

document

rat   cat   • • •   buy

Which animal model was used?



**Sentence 1:**
Pre-pubertal (day 18 of life) female mice were randomized to receive placebo, 200 mg/kg CTX, or 120 mg/kg CTX.
**Similarity to MC1: 0.5478**

**Sentence 2:**
The dosages of CTX used were based on previous studies, which demonstrated a significant dose-dependent ovarian toxicity.
**Similarity to MC1: 0.3887**

OAK RIDGE
National Laboratory

# Dataset

1. Unstructured text of criteria descriptions (MC 1 – 6)

2. **670 research publications** with results for 2,615 uterotrophic bioassays

   - *A curated database of rodent uterotrophic bioactivity.* (2015). Kleinstreuer et. al. Environmental health perspectives.
   - **~120 publications** (~18%) contain GL studies (~540 out of 2,615)

OAK RIDGE
National Laboratory

# Our approach

- Break input document down to shorter segments (e.g. sentences)

- Find segments most relevant (similar) to minimum criteria descriptions

    – Leverages word embeddings

    – Correct sentence typically among the top 5 sentences

- Fully unsupervised

**Natural-language description of Minimum Criteria**

Animal model: OVX adult rats (OVX 6-8 weeks), 14 day post-surgery recovery. OVX adult mouse, ...

**Text segments from target document**

**Segment 1:** Pre-pubertal (day 18 of life) female mice were ...

**Segment 2:** The dosages of CTX used were based on previous ...

**Segment 3:** ...

**Segment 1 similarity: 0.5478**
**Segment 2 similarity: 0.3887**
**Segment 3 similarity: …**

Herrmannova et al. (2018). Unsupervised Identification of Study Descriptors in Toxicology Research: An Experimental Study. Ninth International Workshop on Health Text Mining and Information Analysis, EMNLP 2018.

OAK RIDGE
National Laboratory

# Example results

## Example 1: Necropsy timing (correct answer underlined)

The intact female weanling version in the Organization for Economic Cooperation and Development (OECD) uterotrophic assay Test Guideline (TG) 440 is proposed as an alternative to the adult ovariectomized female version, because it does not involve surgical intervention (vs the ovariectomized version) and detects direct/indirect-acting estrogenic/anti-estrogenic substances (vs the ovariectomized version which detects only direct-acting estrogenic/anti-estrogenic substances binding to the estrogen receptor). This validation study followed OECD TG 440, with six female weanling rats (postnatal day 21) per dose group and six treatment groups. Females were weighed and dosed once daily by oral gavage for three consecutive days, with one of six doses of 17alpha-ethinyl estradiol in corn oil at 5 ml kg(-)(1) at 0 and 0.1-10 microg kg(-)(1) per day. On **postnatal day 24**, the juvenile females were euthanized by CO(2) asphyxiation, weighed, livers weighed and uteri weighed wet and blotted. The presence or absence of vaginal patency was recorded. Absolute and relative (to terminal body weight) uterine wet and blotted weights and uterine luminal fluid weights were significantly increased at 3.0 and 10.0 (both P < 0.01) microg kg(-)(1) per day, and increased to ~140% of control values at 1.0 microg kg(-)(1) per day (not statistically significantly). In vivo body weights, weight changes, feed consumption, liver weights and terminal body weights were unaffected. Vaginal patency was not acquired in any female at any dose, although vaginal puckering was observed in one female at 10.0 microg kg(-)(1) per day. Therefore, this intact weanling uterotrophic assay is validated in our laboratory for use under US and European endocrine toxicity testing programs/legislation.

**Yellow: sentence** most similar to description
**Red: segment** most similar to description

## Example 2: Number of dose groups (correct answer underlined)

The intact female weanling version in the Organization for Economic Cooperation and Development (OECD) uterotrophic assay Test Guideline (TG) 440 is proposed as an alternative to the adult ovariectomized female version, because it does not involve surgical intervention (vs the ovariectomized version) and detects direct/indirect-acting estrogenic/anti-estrogenic substances (vs the ovariectomized version which detects only direct-acting estrogenic/anti-estrogenic substances binding to the estrogen receptor). This validation study followed OECD TG 440, with six female weanling rats (postnatal day 21) per dose group and **six treatment groups**. Females were weighed and dosed once daily by oral gavage for three consecutive days, with one of six doses of 17alpha-ethinyl estradiol in corn oil at 5 ml kg(-)(1) at 0 and 0.1-10 microg kg(-)(1) per day. On postnatal day 24, the juvenile females were euthanized by CO(2) asphyxiation, weighed, livers weighed and uteri weighed wet and blotted. The presence or absence of vaginal patency was recorded. Absolute and relative (to terminal body weight) uterine wet and blotted weights and uterine luminal fluid weights were significantly increased at 3.0 and 10.0 (both P < 0.01) microg kg(-)(1) per day, and increased to ~140% of control values at 1.0 microg kg(-)(1) per day (not statistically significantly). In vivo body weights, weight changes, feed consumption, liver weights and terminal body weights were unaffected. Vaginal patency was not acquired in any female at any dose, although vaginal puckering was observed in one female at 10.0 microg kg(-)(1) per day. Therefore, this intact weanling uterotrophic assay is validated in our laboratory for use under US and European endocrine toxicity testing programs/legislation.

**OAK RIDGE** National Laboratory

# Next steps

- Move beyond solely identifying occurrence of target terms in text by **incorporating context**

  - For example, distinguish cases where a chemical used as experimental treatment or as part of a laboratory procedure

- Jin et al. demonstrated biomedical contextual word vectors capable of **distinguishing different meanings of the same word**
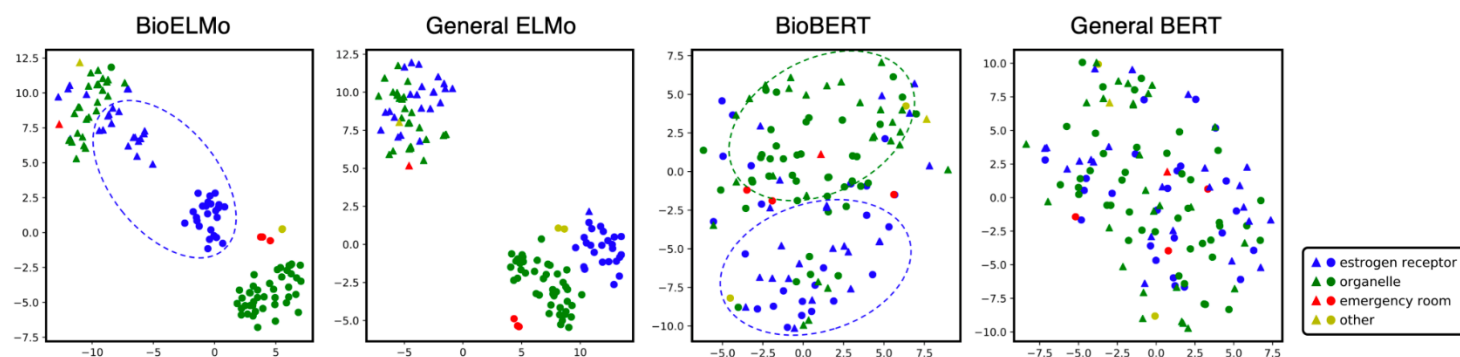


Figure 3: t-SNE visualizations of the token **ER** embeddings in different contexts by BioELMo, general ELMo, BioBERT and general BERT. ● and ▲ represent **ER** mentions within and outside of parentheses, respectively. Colors refer to different actual meanings of the **ER** mention.

Source: Jin et al. (2019). Probing Biomedical Embeddings from Language Models. 3rd Workshop on Evaluating Vector Space Representations for NLP, NAACL 2019.

OAK RIDGE
National Laboratory

# Next steps

- BioELMo word vectors for all occurrences of the word "rat" in 50 uterotrophic bioassay publications
  - Color represents paper section

- Next steps: leverage contextual vectors to **enable fast identification of relevant mentions**
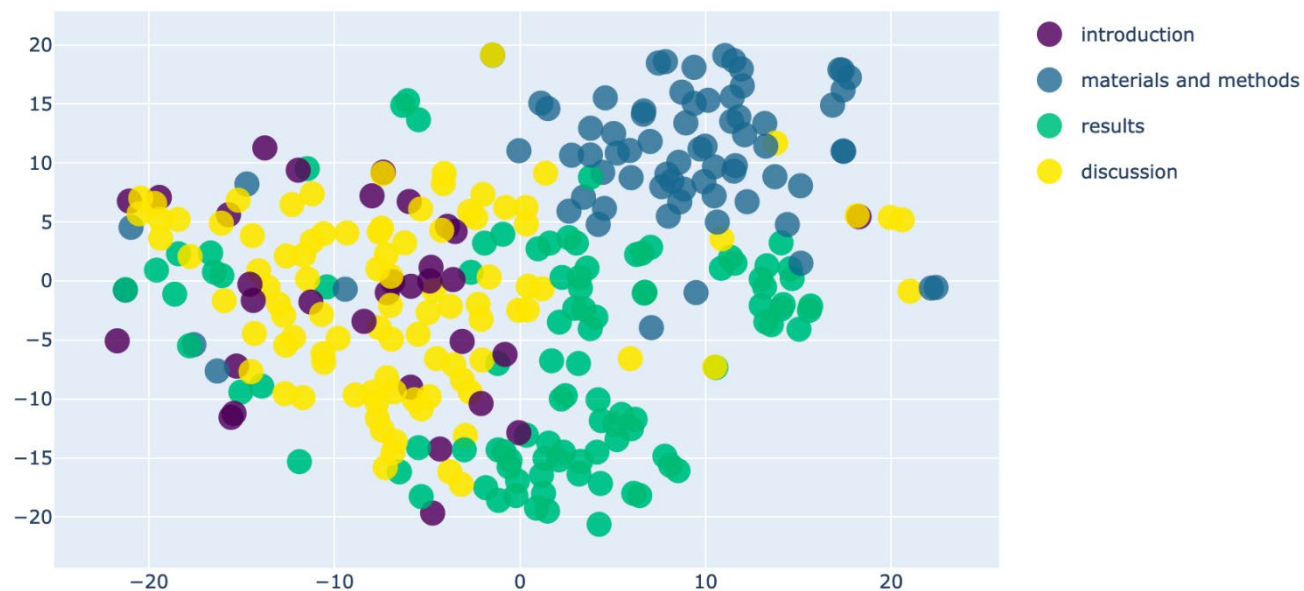
# Current limitations and future work

- Problems:
  - "Semantic implication"
  - Tables/figures
  - Multiple studies
  - Answer not in text

- Improvements
  - Making extraction more accurate
    - Applying robust evaluation metrics toward the extracted text
    - Combining both (supervised and unsupervised) approaches
  - Labeling extracted text
  - Using ensembles to improve classification / joint modelling of MC labels (e.g. using Deep Learning)

**OAK RIDGE**
National Laboratory

# Current limitations

- Tables/figures



Figure 1. Study design. The animals were divided into placebo and test groups. The effect of CTX was assessed as a function of both dose and age of exposure. Animals were euthanized and tissue was harvested and examined.
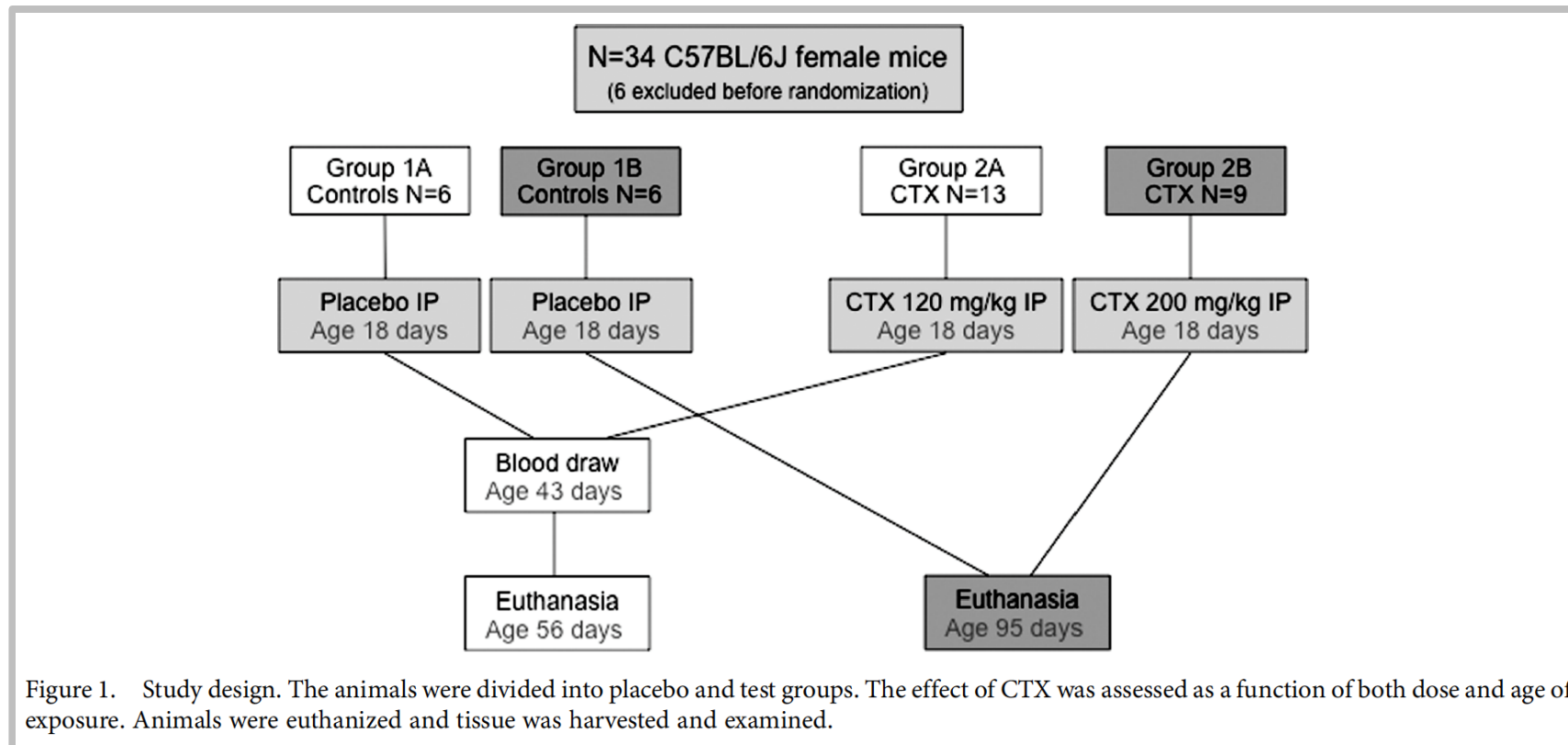
# Extracting structured data from scientific documents

# Long term goal

intubated to ensure a patent airway, and the right carotid artery was cannulated to measure arterial pressure. The right spinotrapezius muscle was then exteriorized, superfused with an electrolyte solution, and prepared for microscopic observation as previously described (Nurkiewicz et al. 2004).

The animal preparation was then transferred to the stage of an intravital microscope. Video images were displayed and videotaped for off-line analysis. During videotape replay, arteriolar inner diameters were measured and venular leukocyte adhesion was quantified.

*Experimental protocols. Protocol 1.* Arteriolar endothelium-dependent dilation was evaluated by assessing the capacity for $Ca^{2+}$-dependent endothelial nitric oxide formation in response to intraluminal infusion of the calcium ionophore A23187 (Sigma Chemical Co., St. Louis, MO). Glass micropipettes were filled with a $10^{-7}$ M solution of A23187 and inserted into the arteriolar lumen, and A23187

was then infused directly into the flow stream for 2-min periods at ejection pressures of 5, 10, 20, and 40 psi (Nurkiewicz et al. 2004). A 2-min recovery period followed each ejection. At the end of all intravital experiments, adenosine (ADO) was added to the superfusate ($10^{-4}$ M final concentration) to fully dilate the microvascular network and determine the passive diameter of each arteriole studied.

*Protocol 2.* To evaluate arteriolar responsiveness to adrenergic stimulation, phenylephrine (PHE) was iontophoretically applied to individual arterioles in rats after exposure to either saline or 0.25 mg ROFA. Micropipettes were filled with a 50-mM solution of PHE in distilled water. The pipette tip was placed in light contact with the arteriolar wall, and a current programmer delivered continuous 2-min ejection currents of 50, 100, and 200 nA (randomly). A 2-min recovery period followed each application. To exclude the possibility that adrenergic stimulation could increase NO

production and therefore attenuate the observed constrictions, these experiments were performed during NO synthase (NOS) inhibition with $N^G$-monomethyl-L-arginine ($10^{-4}$ M final superfusate concentration).

*Protocol 3.* Adhering or rolling leukocytes in first-order venules of rats after exposure to either saline, 0.25 mg ROFA, or 0.1 mg $TiO_2$ were quantified to characterize microvascular inflammation. Leukocytes that were either stationary or moving but in constant contact with the venular wall for at least 200 μm were counted for 1 min in each venule studied.

*Protocol 4.* Oxidant activity in the arteriolar wall was measured with the tetranitroblue tetrazolium (TNBT) reduction method, which provides a general index of microvascular oxidant stress (Lenda and Boegehold 2002). After 1 hr of continuous exposure to 2% TNBT superfusion, the spinotrapezius muscle was fixed with a 10% formalin solution and excised. The tissue was then viewed with bright-field microscopy, and images of microvessels were digitized and analyzed. Using a $1 \times 5$ μm photometric window, a series of average pixel intensity measurements were made along the vessel wall and in extravascular regions immediately adjacent to the wall. To assess microvascular wall levels of formazan (the reduction product of TNBT and therefore an index of oxidant activity), the measured pixel intensities were used to calculate microvascular wall light absorption ($A$): $A = \ln(I_i/I_e)$, where $I_i$ is the vessel intensity and $I_e$ is the intensity for the adjacent extravascular region. The amount of formazan formed is proportional to the level of oxidant activity, and calculated light absorption is linearly related to the amount of formazan present (Lenda and Boegehold 2002).

*Data and statistical analyses.* Arteriolar diameter ($D$, in micrometers) was sampled at 10-sec intervals. Resting vascular tone was calculated for each vessel as follows: tone = $[(D_{pass} - D_t)/D_{pass}] \times 100$, where $D_{pass}$ is

**Table 1.** Profiles of experimental animals used for intravital studies.

| Experimental group | N | Age (days) | Weight (g) | Mean arterial pressure (mm Hg) |
|---|---|---|---|---|
| Saline control | 15 | 55 ± 4 | 236 ± 4 | 98 ± 5 |
| 0.1 mg TiO₂ | 5 | 50 ± 1 | 275 ± 9* | 108 ± 6 |
| 0.1 mg ROFA | 4 | 52 ± 1 | 225 ± 3 | 102 ± 9 |
| 0.25 mg TiO₂ | 3 | 56 ± 3 | 219 ± 7 | 91 ± 15 |
| 0.25 mg ROFA | 11 | 50 ± 2 | 221 ± 8 | 94 ± 7 |

*N*, number of rats. Values are mean ± SE.
*$p < 0.05$ compared with all other groups.

**Table 2.** BAL data from saline-treated and TiO₂- and ROFA-exposed rats.

| Experimental group | Cellular content of PMNL (10⁶ cells/rat) | BAL fluid Albumin (mg/mL) | LDH (U/L) | Total AM CL |
|---|---|---|---|---|
| Saline control | 0.93 ± 0.11 | 0.13 ± 0.02 | 58 ± 10 | 7.50 ± 1.63 |
| 0.1 mg TiO₂ | 1.73 ± 0.36* | 0.14 ± 0.04 | 67 ± 13 | 5.21 ± 1.60 |
| 0.1 mg ROFA | 1.24 ± 0.28 | 0.23 ± 0.05 | 75 ± 2 | 14.47 ± 2.60*,** |
| 0.25 mg TiO₂ | 1.11 ± 0.15 | 0.19 ± 0.02 | 65 ± 8 | 3.64 ± 0.73 |
| 0.25 mg ROFA | 1.91 ± 0.20* | 0.17 ± 0.01 | 46 ± 4 | 17.42 ± 1.11*,** |

$N$ = 22 rats for saline control; $N$ = 5–7 rats for all other doses. Values are mean ± SE. CL = counts per minute × 10⁵/0.25 × 10⁶ AM/15 min.
*$p < 0.05$ compared with saline. **$p < 0.05$ compared with 0.1 mg TiO₂ and $p < 0.05$ compared with 0.25 mg TiO₂.
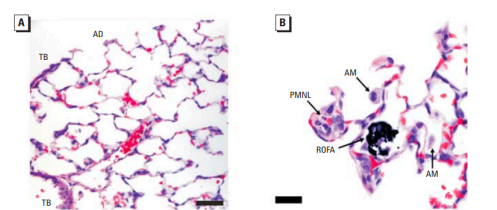
**Figure 1.** Histologic evidence of focal pulmonary alveolitis 24 hr after PM exposure. (*A*) and (*B*) are representative findings from five saline-treated rats and five rats exposed to 0.1 mg ROFA. (*A*) Saline control showing no morphologic alterations. Abbreviations: AD, alveolar duct; TB, terminal bronchioles. Bar = 50 μm. (*B*) Histopathologic alterations in a ROFA-exposed rat. Agglomerated ROFA particles in an alveolar space can be seen near an alveolar duct. The ROFA particles do not transmit light and therefore appear black when viewed in the light microscope. AMs are frequently observed in alveoli near ROFA particles and are intimately associated with the agglomerated ROFA. PMNLs are present in lesser numbers near ROFA particles, most frequently in the interstitium. Bar = 20 μm. (*C*) Mean alveolitis pathology scores.
*$p < 0.05$ compared with saline; similar results were obtained with TiO₂.

**Table 1.** Profiles of experimental animals used for intravital studies.

| Experimental group | N | Age (days) | Weight (g) | Mean arterial pressure (mm Hg) |
|---|---|---|---|---|
| Saline control | 15 | 55 ± 4 | 236 ± 4 | 98 ± 5 |
| 0.1 mg TiO₂ | 5 | 50 ± 1 | 275 ± 9* | 108 ± 6 |
| 0.1 mg ROFA | 4 | 52 ± 1 | 225 ± 3 | 102 ± 9 |
| 0.25 mg TiO₂ | 3 | 56 ± 3 | 219 ± 7 | 91 ± 15 |
| 0.25 mg ROFA | 11 | 50 ± 2 | 221 ± 8 | 94 ± 7 |

*N*, number of rats. Values are mean ± SE.
*$p < 0.05$ compared with all other groups.

## TiO₂ groups: Group size

| Dose group | Group size |
|---|---|
| Control | 15 |
| 0.1 mg | 5 |
| 0.25 mg | 3 |

## TiO₂ groups: Age

| Dose group | Age (days) |
|---|---|
| Control | Mean = 55, SE = 4 |
| 0.1 mg | Mean = 50, SE = 1 |
| 0.25 mg | Mean = 56, SE = 3 |

OAK RIDGE National Laboratory

# Short term goal



**Table 1.** Profiles of experimental animals used for intravital studies.

| Experimental group | N | Age (days) | Weight (g) | Mean arterial pressure (mm Hg) |
|---|---|---|---|---|
| Saline control | 15 | 55 ± 4 | 236 ± 4 | 98 ± 5 |
| 0.1 mg TiO$_2$ | 5 | 50 ± 1 | 275 ± 9* | 108 ± 6 |
| 0.1 mg ROFA | 4 | 52 ± 1 | 225 ± 3 | 102 ± 9 |
| 0.25 mg TiO$_2$ | 3 | 56 ± 3 | 219 ± 7 | 91 ± 15 |
| 0.25 mg ROFA | 11 | 50 ± 2 | 221 ± 8 | 94 ± 7 |

$N$, number of rats. Values are mean ± SE.
*$p < 0.05$ compared with all other groups.

**Caption:** Table 1. Profiles of experimental animals …

| Experimental group | N | Age (days) | Weight (g) | … |
|---|---|---|---|---|
| Saline control | 15 | 55 +- 4 | 236 +- 4 | … |
| 0.1 mg TiO$_2$ | 5 | 50 +- 1 | 275 +- 9 | … |
| 0.1 mg ROFA | 4 | 52 +- 1 | … | |
| 0.25 mg TiO$_2$ | 3 | … | | |
| 0.25 mg ROFA | 11 | … | | |

**Notes:** N, number of rats. Values are mean +- SE. …

OAK RIDGE National Laboratory

# DeepPDF object detection

- Deep Neural Networks highly successful in image analysis

- Apply same principle to scientific documents

- Used network which was developed for medical image segmentation
  - Performs well on small datasets

Stahl et al. (2018) DeepPDF: A Deep Learning Approach to Analyzing PDFs. 2018 Workshop on Mining Scientific Publications at the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

# DeepPDF object detection

- Deep Neural Networks highly successful in image analysis

- Apply same principle to scientific documents

- Used network which was developed for medical image segmentation
  - Performs well on small datasets

Stahl et al. (2018) DeepPDF: A Deep Learning Approach to Analyzing PDFs. 2018 Workshop on Mining Scientific Publications at the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).



OAK RIDGE National Laboratory

19

# DeepPDF object detection

## Table detection



## Paragraph detection

OAK RIDGE
National Laboratory

# Next steps – table structure prediction

- Identify rows, columns, headers, and other structure

- Example below: two successful (a, b) and two challenging (c, d) cases



(a) Row detection, no ruling lines present

(b) Column detection, no ruling lines present

(c) Nested row hierarchy

(d) Very close-by rows

Source: Schreiber et al. (2017). DeepDeSRT:Deep Learning for Detection and StructureRecognition of Tables in Document Images. ICDAR 2017.

OAK RIDGE
National Laboratory

# Next steps – text extraction

- Correctly extract text contained in tables
- Older documents particularly challenging
- Example extraction using Tabula (https://tabula.technology/)

# Towards deriving actionable insights from scientific knowledge graphs

# Research challenge: Literature-based discovery (LBD)

- LBD is the process of **knowledge extraction** and **automated hypothesis generation** from scientific literature

- Traditional approach: Swanson Linking
  - There are published relations between A and B, and B and C
  - **There may be a yet undiscovered relation between A-C**

- Demonstrated on the connection between dietary fish oil (A) and Raynaud's disease (C) through high blood viscosity (B) [1]



1. Swanson DR. *Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge*. Perspectives in Biology and Medicine. 1986. 30(1): 7-18.
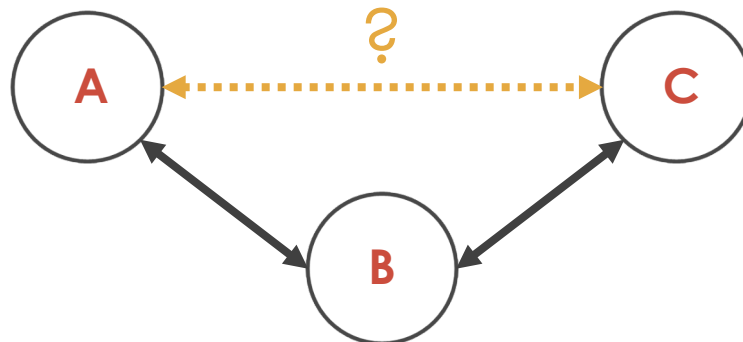
# Approach

- Previous works hypothesized **shortest path information may be used for automating literature-based discovery**

- To test this hypothesis we created a **graph of A->B->C connections** from Semantic MEDLINE + CORD-19 publications

  - **Graph nodes (18.5 million total)**
    - Biomedical concepts (UMLS)
    - Publications

  - **Graph relations (213 million total):**
    - Relations between **concepts** (extracted from publication abstracts)
    - Relations between **concepts** and **publications** (concept appears in paper)
    - Citations between **publications**

# Graph example
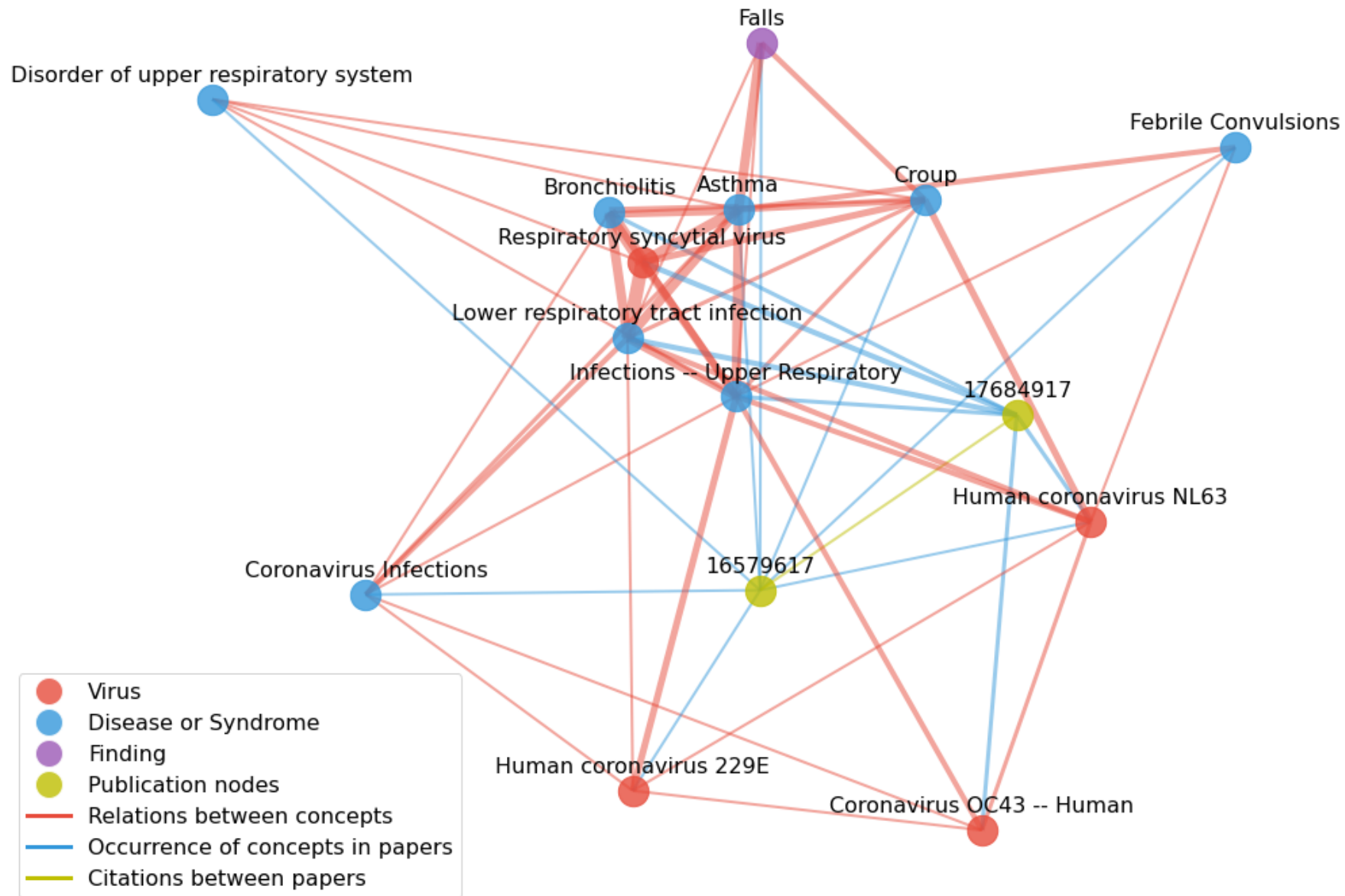
# All Pairs Shortest Path algorithm – DSNAPSHOT

- **Goal:** Find the shortest paths between all nodes within an 18 million node graph

- **Approach:** Leverage graph sparsity and algorithmic optimizations to enable GPU usage on ORNL Summit [2, 3]
  - 18 hrs to process entire graph – **10X speed up over the state of the art**
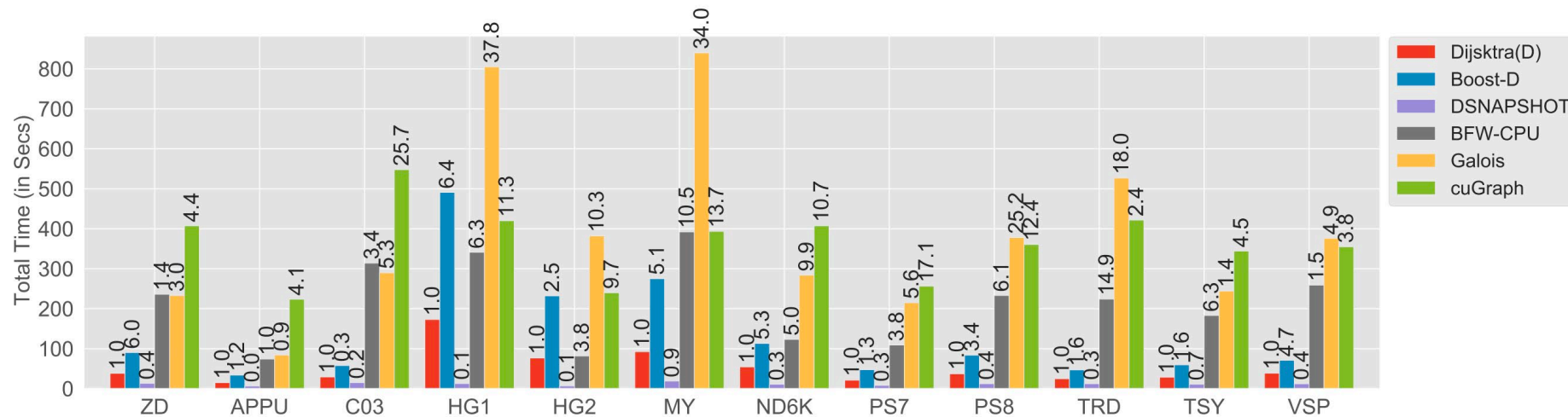
2. Sao P, Kannan R, Gera P and Vuduc R. *A supernodal all-pairs shortest path algorithm*. In Proceedings of the 25th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming. 2020. 250-261.
3. R. Kannan et al. *Scalable Knowledge-Graph Analytics at 136 Petaflops/s*. To appear in Proceedings of the ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (IEEE Supercomputing). 2020.

OAK RIDGE
National Laboratory

# DSNAPSHOT 2005 Experiment

- Use **DSNAPSHOT** on PubMed data from 1970 – 2005 (Done)

- Use **DSNAPSHOT results** to recommend new connections (A->C) for 2005 (Done)

- Validate these connections using 2020 data (in progress)

- Create a ranking method for new potential connections (in progress)

OAK RIDGE
National Laboratory
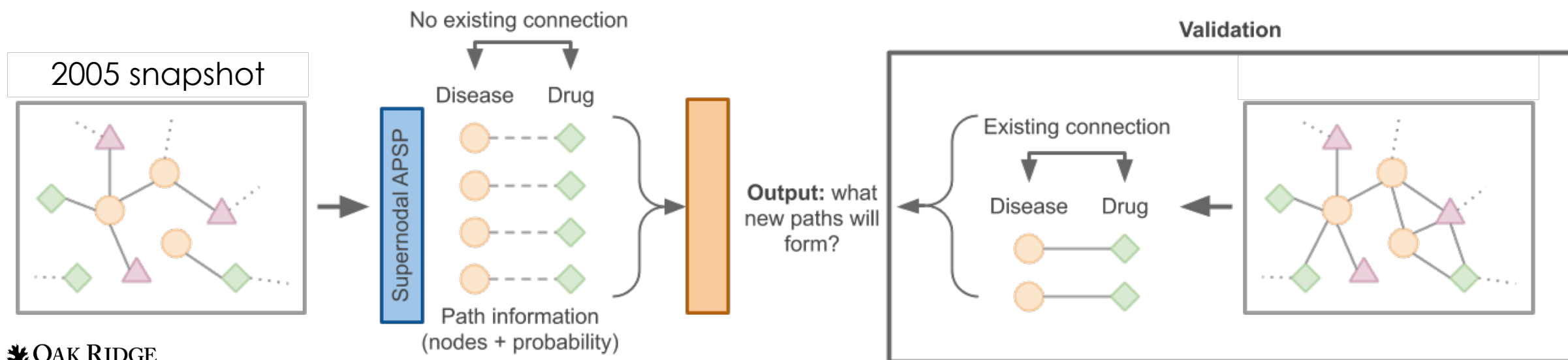
# Results of 2005 - Learning to Discover

Connections discovered by 2020
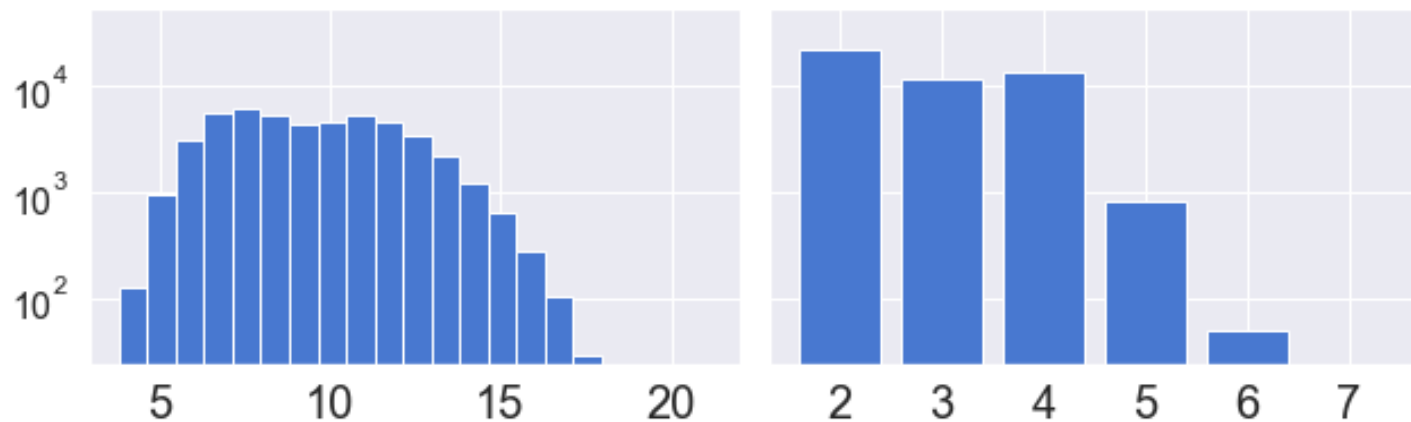


No connections by 2020

Path length

Number of edges
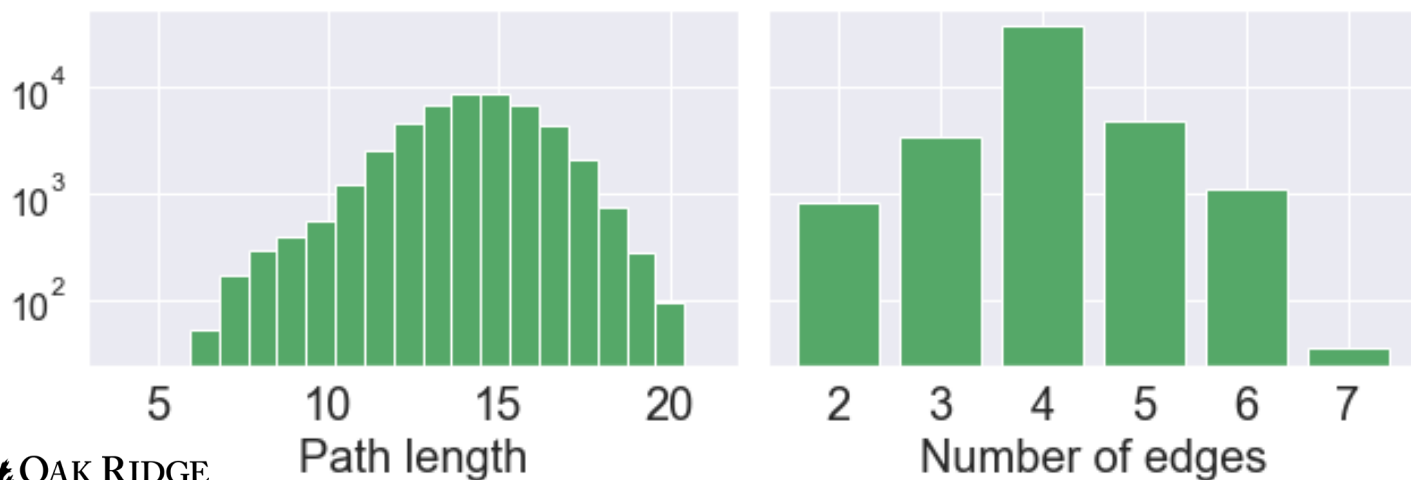
- Paths that lead to a later connection are different from paths that don't lead to a connection

- Hypothesis: This is a ranking method for predicting future discoveries

**OAK RIDGE**
National Laboratory

# Example results in **2005** snapshot

*Successful outcome of human metapneumovirus (hMPV) pneumonia in a lung transplant recipient treated with intravenous ribavirin **(2008)**:*

**hMPV** was diagnosed by polymerase chain reaction, and treated with intravenous **ribavirin** with a successful outcome.

Pharmacologic Substance

Virus

Concept:
Ribavirin

**Is there a connection?**

Concept:
Human
Metapneumovirus
(hMPV)

Virus

Concept:
Respiratory
Syncytial
Virus (RSV)
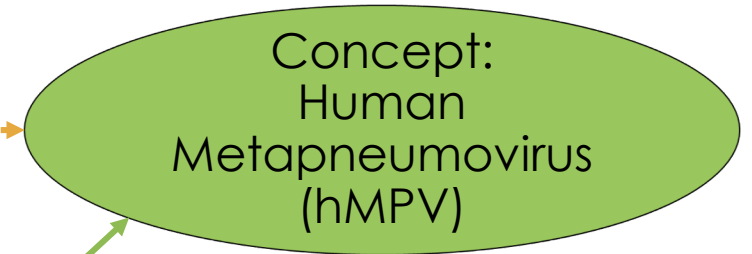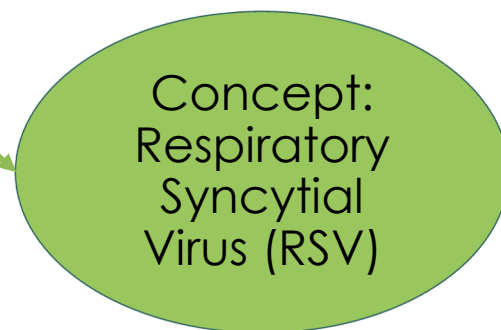
*Ribavirin and Respiratory Syncytial Virus **(1986)**:*

**Ribavirin** […] possesses an unusually broad spectrum of antiviral activity, inhibiting under laboratory conditions a wide variety of RNA and DNA viruses, including among respiratory viruses influenza types A and B, parainfluenza types 1, 2, and 3, **respiratory syncytial virus (RSV)**, and possibly coronavirus.

*The causes and diagnosis of influenza-like illness **(2004)**:*

In addition to influenza, viruses known to cause [influenza-like illness] include **respiratory syncytial virus**, rhinovirus, adenovirus, parainfluenza viruses, human coronaviruses […] and the recently recognised **human metapneumovirus**.

# Acknowledgements and contact

# Acknowledgements

- NIEHS/NTP
  - Mary Wolfe
  - Nicole Kleinstreuer
  - Charles Schmitt

- Oak Ridge National Laboratory
  - Steven Young
  - Christopher Stahl
  - Chelsey Stahl
  - Dasha Herrmannova

**NIH** National Institute of Environmental Health Sciences
*Your Environment. Your Health.*

**NTP** National Toxicology Program
U.S. Department of Health and Human Services

**OAK RIDGE** National Laboratory

OAK RIDGE
National Laboratory

# Acknowledgements

**OAK RIDGE**
National Laboratory

# Contact

**Robert M. Patton, Ph.D.**

Computer Science & Mathematics Division

Oak Ridge National Laboratory

**pattonrm@ornl.gov**

**OAK RIDGE**
National Laboratory

# Extra Content

# Results for MC 1: Animal model

The intact female weanling version in the Organization for Economic Cooperation and Development (OECD) uterotrophic assay Test Guideline (TG) 440 is proposed as an alternative to the adult ovariectomized female version, because it does not involve surgical intervention (vs the ovariectomized version) and detects direct/indirect-acting estrogenic/anti-estrogenic substances (vs the ovariectomized version which detects only direct-acting estrogenic/anti-estrogenic substances binding to the estrogen receptor). This validation study followed OECD TG 440, with six **female weanling rats (postnatal day 21)** per dose group and six treatment groups. Females were weighed and dosed once daily by oral gavage for three consecutive days, with one of six doses of 17alpha-ethinyl estradiol in corn oil at 5 ml kg(-)(1) at 0 and 0.1-10 microg kg(-)(1) per day. On postnatal day 24, the juvenile females were euthanized by $CO_2$ asphyxiation, weighed, livers weighed and uteri weighed wet and blotted. The presence or absence of vaginal patency was recorded. Absolute and relative (to terminal body weight) uterine wet and blotted weights and uterine luminal fluid weights were significantly increased at 3.0 and 10.0 (both $P < 0.01$) microg kg(-)(1) per day, and increased to ~140% of control values at 1.0 microg kg(-)(1) per day (not statistically significantly). In vivo body weights, weight changes, feed consumption, liver weights and terminal body weights were unaffected. Vaginal patency was not acquired in any female at any dose, although vaginal puckering was observed in one female at 10.0 microg kg(-)(1) per day. Therefore, this intact weanling uterotrophic assay is validated in our laboratory for use under US and European endocrine toxicity testing programs/legislation.

Correct answer (underlined)

OAK RIDGE
National Laboratory

The intact female weanling version in the Organization for Economic Cooperation and Development (OECD) uterotrophic assay Test Guideline (TG) 440 is proposed as an alternative to the adult ovariectomized female version, because it does not involve surgical intervention (vs the ovariectomized version) and detects direct/indirect-acting estrogenic/anti-estrogenic substances (vs the ovariectomized version which detects only direct-acting estrogenic/anti-estrogenic substances binding to the estrogen receptor). This validation study followed OECD TG 440, with **six** female weanling **rats** (postnatal day 21) **per dose group** and six treatment groups. Females were weighed and dosed once daily by oral gavage for three consecutive days, with one of six doses of 17alpha-ethinyl estradiol in corn oil at 5 ml kg(-)(1) at 0 and 0.1-10 microg kg(-)(1) per day. On postnatal day 24, the juvenile females were euthanized by $CO_2$ asphyxiation, weighed, livers weighed and uteri weighed wet and blotted. The presence or absence of vaginal patency was recorded. Absolute and relative (to terminal body weight) uterine wet and blotted weights and uterine luminal fluid weights were significantly increased at 3.0 and 10.0 (both $P < 0.01$) microg kg(-)(1) per day, and increased to ~140% of control values at 1.0 microg kg(-)(1) per day (not statistically significantly). In vivo body weights, weight changes, feed consumption, liver weights and terminal body weights were unaffected. Vaginal patency was not acquired in any female at any dose, although vaginal puckering was observed in one female at 10.0 microg kg(-)(1) per day. Therefore, this intact weanling uterotrophic assay is validated in our laboratory for use under US and European endocrine toxicity testing programs/legislation.

Correct answer (underlined)

OAK RIDGE
National Laboratory

# Results for MC 3: Route of administration

The intact female weanling version in the Organization for Economic Cooperation and Development (OECD) uterotrophic assay Test Guideline (TG) 440 is proposed as an alternative to the adult ovariectomized female version, because it does not involve surgical intervention (vs the ovariectomized version) and detects direct/indirect-acting estrogenic/anti-estrogenic substances (vs the ovariectomized version which detects only direct-acting estrogenic/anti-estrogenic substances binding to the estrogen receptor). This validation study followed OECD TG 440, with six female weanling rats (postnatal day 21) per dose group and six treatment groups. Females were weighed and dosed once daily by **oral gavage** for three consecutive days, with one of six doses of 17alpha-ethinyl estradiol in corn oil at 5 ml kg(-)(1) at 0 and 0.1-10 microg kg(-)(1) per day. On postnatal day 24, the juvenile females were euthanized by $CO_2$ asphyxiation, weighed, livers weighed and uteri weighed wet and blotted. The presence or absence of vaginal patency was recorded. Absolute and relative (to terminal body weight) uterine wet and blotted weights and uterine luminal fluid weights were significantly increased at 3.0 and 10.0 (both P < 0.01) microg kg(-)(1) per day, and increased to ~140% of control values at 1.0 microg kg(-)(1) per day (not statistically significantly). In vivo body weights, weight changes, feed consumption, liver weights and terminal body weights were unaffected. Vaginal patency was not acquired in any female at any dose, although vaginal puckering was observed in one female at 10.0 microg kg(-)(1) per day. Therefore, this intact weanling uterotrophic assay is validated in our laboratory for use under US and European endocrine toxicity testing programs/legislation.

Correct answer (underlined)

**OAK RIDGE**
National Laboratory

# Results for MC 4: Number of dose groups

The intact female weanling version in the Organization for Economic Cooperation and Development (OECD) uterotrophic assay Test Guideline (TG) 440 is proposed as an alternative to the adult ovariectomized female version, because it does not involve surgical intervention (vs the ovariectomized version) and detects direct/indirect-acting estrogenic/anti-estrogenic substances (vs the ovariectomized version which detects only direct-acting estrogenic/anti-estrogenic substances binding to the estrogen receptor). This validation study followed OECD TG 440, with six female weanling rats (postnatal day 21) per dose group and **six treatment** **groups**. Females were weighed and dosed once daily by oral gavage for three consecutive days, with one of six doses of 17alpha-ethinyl estradiol in corn oil at 5 ml kg(-)(1) at 0 and 0.1-10 microg kg(-)(1) per day. On postnatal day 24, the juvenile females were euthanized by $CO_2$ asphyxiation, weighed, livers weighed and uteri weighed wet and blotted. The presence or absence of vaginal patency was recorded. Absolute and relative (to terminal body weight) uterine wet and blotted weights and uterine luminal fluid weights were significantly increased at 3.0 and 10.0 (both $P < 0.01$) microg kg(-)(1) per day, and increased to ~140% of control values at 1.0 microg kg(-)(1) per day (not statistically significantly). In vivo body weights, weight changes, feed consumption, liver weights and terminal body weights were unaffected. Vaginal patency was not acquired in any female at any dose, although vaginal puckering was observed in one female at 10.0 microg kg(-)(1) per day. Therefore, this intact weanling uterotrophic assay is validated in our laboratory for use under US and European endocrine toxicity testing programs/legislation.

Correct answer (underlined)

OAK RIDGE
National Laboratory

# Results for MC 5: Dosing interval

The intact female weanling version in the Organization for Economic Cooperation and Development (OECD) uterotrophic assay Test Guideline (TG) 440 is proposed as an alternative to the adult ovariectomized female version, because it does not involve surgical intervention (vs the ovariectomized version) and detects direct/indirect-acting estrogenic/anti-estrogenic substances (vs the ovariectomized version which detects only direct-acting estrogenic/anti-estrogenic substances binding to the estrogen receptor). This validation study followed OECD TG 440, with six female weanling rats (postnatal day 21) per dose group and six treatment groups. Females were weighed and dosed once daily by oral gavage for **three consecutive days**, with one of six doses of 17alpha-ethinyl estradiol in corn oil at 5 ml kg(-)(1) at 0 and 0.1-10 microg kg(-)(1) per day. On postnatal day 24, the juvenile females were euthanized by CO(2) asphyxiation, weighed, livers weighed and uteri weighed wet and blotted. The presence or absence of vaginal patency was recorded. Absolute and relative (to terminal body weight) uterine wet and blotted weights and uterine luminal fluid weights were significantly increased at 3.0 and 10.0 (both P < 0.01) microg kg(-)(1) per day, and increased to ~140% of control values at 1.0 microg kg(-)(1) per day (not statistically significantly). In vivo body weights, weight changes, feed consumption, liver weights and terminal body weights were unaffected. Vaginal patency was not acquired in any female at any dose, although vaginal puckering was observed in one female at 10.0 microg kg(-)(1) per day. Therefore, this intact weanling uterotrophic assay is validated in our laboratory for use under US and European endocrine toxicity testing programs/legislation.

Correct answer (underlined)

OAK RIDGE
National Laboratory

# Results for MC 6: Necropsy timing

The intact female weanling version in the Organization for Economic Cooperation and Development (OECD) uterotrophic assay Test Guideline (TG) 440 is proposed as an alternative to the adult ovariectomized female version, because it does not involve surgical intervention (vs the ovariectomized version) and detects direct/indirect-acting estrogenic/anti-estrogenic substances (vs the ovariectomized version which detects only direct-acting estrogenic/anti-estrogenic substances binding to the estrogen receptor). This validation study followed OECD TG 440, with six female weanling rats (postnatal day 21) per dose group and six treatment groups. Females were weighed and dosed once daily by oral gavage for three consecutive days, with one of six doses of 17alpha-ethinyl estradiol in corn oil at 5 ml kg(-)(1) at 0 and 0.1-10 microg kg(-)(1) per day. On **postnatal day 24**, the juvenile females were euthanized by $CO_2$ asphyxiation, weighed, livers weighed and uteri weighed wet and blotted. The presence or absence of vaginal patency was recorded. Absolute and relative (to terminal body weight) uterine wet and blotted weights and uterine luminal fluid weights were significantly increased at 3.0 and 10.0 (both P < 0.01) microg kg(-)(1) per day, and increased to ~140% of control values at 1.0 microg kg(-)(1) per day (not statistically significantly). In vivo body weights, weight changes, feed consumption, liver weights and terminal body weights were unaffected. Vaginal patency was not acquired in any female at any dose, although vaginal puckering was observed in one female at 10.0 microg kg(-)(1) per day. Therefore, this intact weanling uterotrophic assay is validated in our laboratory for use under US and European endocrine toxicity testing programs/legislation.

Correct answer (underlined)

OAK RIDGE
National Laboratory